The pre-clinical evolution of haematological malignancies

Grace Collord

King's College

University of Cambridge

This dissertation was submitted for the degree of

Doctor of Philosophy

August 2019.

Declaration

I hereby declare that this dissertation is my own work and that any work done in collaboration with others is explicitly indicated in the text. This work does not contain any material substantially similar to work I have previously submitted, or am in the process of preparing, for any qualification at any institution. This dissertation does not exceed 60,000 words in length.

Grace Collord

August 2019

Summary

The pre-clinical evolution of haematological malignancies

Grace Collord

Cancer-associated somatic mutations frequently drive clonal expansions in normal ageing tissues. However, the factors governing whether pre-cancerous cells transform into cancer are poorly understood, hindering identification of clones that are clinically significant rather than benign sequelae of ageing. The main aim of this dissertation has been to explore this process in the haematopoietic system, where leukaemia-associated mutations are detectable in >10% of individuals over the age of 50. This phenomenon, termed clonal haematopoiesis (CH), is associated with an increased risk of blood cancers, though only a small minority of individuals progress.

Acute myeloid leukaemia (AML) is the commonest acute leukaemia in adults, and usually presents abruptly with complications of bone marrow failure. Using deep targeted sequencing of stored blood DNA samples from individuals who went on to develop AML and controls, we identified features of CH that predict leukaemic progression. The number, type and burden of genetic changes, as well as certain clinical variables, were predictive of AML-free survival. Examining the pre-clinical evolution of lymphoid malignancies using a similar study design and broader sequencing approach also revealed genetic and clinical features predictive of malignant transformation.

The final part of this study investigates the prevalence of clonal haematopoiesis in childhood cancer survivors treated with intensive chemo- or radiotherapy. In contrast to adult cancer patients, the prevalence of CH in children is not dramatically increased by cytotoxic treatment.

Collectively, these findings provide proof of principle that benign and pre-malignant clonal expansions in normal blood (and perhaps other tissues) may be distinguishable years prior to overt malignant transformation. This could in future enable earlier detection of those at high risk of blood cancers, and stimulate research into possible interventions to reduce the risk of progression.

Acknowledgements

I would like to thank our collaborators for decades of work collecting and curating patient samples and metadata, in particular the European Prospective Investigation into Cancer and Nutrition (EPIC) study, as well as Professor Francesco Frassoni (Instituto Giannina Gaslini, Italy), Dr Kris Bowles (Norfolk and Norwich University Hospital, UK), the Norwich Research Park Biorepository (UK), Dr Matthew Murray (University of Cambridge), Dr Anna Godfrey (University of Cambridge), Dr Amy Jones (Cambridge Haematopathology and Oncology Diagnostic Service), Kevin Jestice (Cambridge Cellular Therapy Laboratory) and Dr Joanna Baxter (Cambridge Blood and Stem Cell Biobank, UK). In particular, Robert Luben and Shabina Hayat of the EPIC-Norfolk study team were instrumental in enabling the work described in chapters 3 and 4. I am indebted to many colleagues for discussions and advice, in particular Dr Naomi Park, Dr Peter Campbell, Dr Inigo Martincorena and Dr Ignacio Varela. I would particularly like to thank Dr Moritz Gerstung for imparting most of what I know about statistics and Dr Sam Behjati, Dr Patrick Tarpey and Dr Elli Papaemmanuil for teaching me to navigate the cancer genome. I would like to thank my supervisors, Dr George Vassiliou and Professor Mike Stratton, for giving me the opportunity to pursue this work in an exceptionally collaborative research environment. I am grateful to King's College and to the funders who have generously supported this work, particularly the Wellcome Trust.

Table of Contents

Declaration 2

Acknowledgements 3

Summary 4

Table of contents 5

Abbreviations 7

Chapter 1: Introduction 9

Chapter 2: Materials and methods 39

Chapter 3: Predicting acute myeloid leukaemia risk in the general population 55

Chapter 4: The pre-clinical evolution of lymphoid neoplasms 84

Chapter 5: Clonal haematopoiesis after childhood cancer treatment 113

Chapter 6: Discussion 130

Bibliography 137

Appendices 171

Appendix 1: Discovery cohort pre-AML and control sample information A1

Appendix 2: Validation cohort pre-AML and control sample information A7

Appendix 3: Childhood cancer survivor cohort details A12

Appendix 4: Custom myeloid cancer gene panel A14

Appendix 5: Multiplex PCR primer sequences A15

Appendix 6: Custom pan-haematological cancer gene panel A16

Appendix 7: Code for AML risk prediction models A17

Appendix 8: Mutations in discovery cohort pre-AML and control samples A90

Appendix 9: Mutations in validation cohort pre-AML and control samples A94

Appendix 10: AML risk prediction model coefficients A96

Appendix 11: AML prediction model based on electronic health record data A97

Appendix 12: Discovery cohort pre-lymphoid neoplasm cases and controls metadata A98

Appendix 13: Validation cohort pre-lymphoid neoplasm cases and controls metadata A102

Appendix 14: Driver mutations in pre-lymphoid neoplasm cases and controls A104

Appendix 15: Lymphoid neoplasm risk prediction model coefficients A107

Appendix 16: First and joint first author primary research publications A108

Abbreviations

ALL	Acute lymphoblastic leukaemia
AML	Acute myeloid leukaemia
AUC	Area under the curve
bp	Base pair
BMI	Body mass index
C	Concordance
CCA	Choriocarcinoma
cDNA	Complementary deoxyribonucleic acid
CH	Clonal haematopoiesis
CH-PD	Clonal haematopoiesis with putative driver mutations
CHIP	Clonal haematopoiesis of indeterminate significance
CLL	Chronic lymphocytic leukaemia
CML	Chronic myeloid leukaemia
CNA	Copy number aberration
CVD	Cardiovascular disease
DBP	Diastolic blood pressure
DC	Discovery cohort
DNA	Deoxyribonucleic acid
ES	Ewing sarcoma
FBC	Full blood count
FFPE	Formalin-fixed paraffin-embedded
HSC	Haematopoietic stem cell
HSCT	Haematopoietic stem cell transplant
HSPC	Haematopoietic stem and progenitor cell
KM	Kaplan-Meier
GCT	Germ cell tumour
HDL	High-density lipoprotein
HL	Hodgkin lymphoma
HSC	Haematopoietic stem cell
HSCT	Haematopoietic stem cell transplant
LCH	Langerhans cell histiocytosis
LDL	Low-density lipoprotein
LL	Lymphoblastic lymphoma
LOH	Loss of heterozygosity
Mb	Megabase
MBL	Monoclonal B-cell lymphocytosis
MDS	Myelodysplastic syndrome
MGUS	Monoclonal gammopathy of undetermined significance

MM	Multiple myeloma
MPN	Myeloproliferative neoplasm
NGS	Next-generation sequencing
NHL	Non-Hodgkin lymphoma
NB	Neuroblastoma
NPC	Nasopharyngeal carcinoma
NRSTS	Non-rhabdomyosarcoma soft tissue sarcoma
PCR	Polymerase chain reaction
RBC	Red blood cell
RDW	Red cell distribution width
RNA	Ribonucleic acid
sAML	Secondary AML
SBP	Systolic blood pressure
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
RMS	Rhabdomyosarcoma
TC	Total cholesterol
T-ALL	T-cell acute lymphoblastic leukaemia
t-AML	Therapy-related AML
t-MN	Therapy-related myeloid neoplasm
VAF	Variant allele fraction
VC	Validation cohort
WBC	White blood cell
WT	Wilms tumour

Chapter 1

Introduction

Modern sequencing technologies are catalysing a revolution in our understanding of cancer genetics, developmental disorders, and ageing (Behjati et al., 2018; Martincorena and Campbell, 2015; Stratton, 2011; Yates and Campbell, 2012). Over the past decade, genomic scrutiny of over a million cancers has revealed the oncogenic mutations responsible for causing most human malignancies (Tate et al., 2019). These discoveries have enabled development of novel targeted cancer therapies and sequencing-based cancer diagnostic methods (Chang et al., 2016; Gerstung et al., 2017; Zahn, 2016). In parallel, sequencing of normal tissues has demonstrated that somatic mutations accumulate in all cells with age due to a host of extrinsic and endogenous exposures (Alexandrov et al., 2013; Hoang et al., 2016; Ju et al., 2017; Martincorena and Campbell, 2015; Yizhak et al., 2018). Somatic genetic diversity in ageing tissues provides a substrate for natural selection at the cellular level. Most somatic mutations have no discernible impact on cell function (Martincorena et al., 2017). However, recent studies have demonstrated that canonical cancer driver mutations are remarkably common in morphologically and functionally normal tissues and frequently fuel clonal expansion (Bowman et al., 2018; Martincorena et al., 2018; Martincorena et al., 2015; Moore et al., 2018; Salk et al., 2018; Yizhak et al., 2018; Yokoyama et al., 2019). The ubiquity of subclonal cancer evolutionary processes represents a daunting challenge to sequencing-based early cancer detection efforts and may also increase the toxicity of novel precision oncology drugs targeting cancer driver mutations present in a significant fraction of normal cells (Busque et al., 2018; Cohen et al., 2018; Martincorena et al., 2015). The landscape of somatic genetic diversity is currently best understood in the haematopoietic system, largely due to ease of representative sampling. Clonal haematopoiesis (CH) becomes increasingly common with age and is associated with an increased risk of haematological malignancies, though only a small minority of individuals with CH ever develop a blood cancer (Busque et al., 2018). The main aim of this dissertation has been to explore the premalignant mutational landscape of haematological cancers and the extent to which indolent clones can be distinguished from CH at high risk of malignant transformation. The general introduction to this thesis provides an overview of somatic evolution in cancer and normal tissues, with an emphasis on the haematopoietic system.

1. Somatic evolution in cancer

"At last gleams of light have come, & I am almost convinced (quite contrary to opinion I started with) that species are not (it is like confessing a murder) immutable."

- Charles Darwin to Joseph Hooker, 11 January 1844

"*One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die…. Natural Selection, as we shall hereafter see, is a power incessantly ready for action*"

- Charles Darwin, *The Origin of Species*, 1959

"If, as I believe that my theory is true & if it be accepted even by one competent judge, it will be a considerable step in science."

- Charles Darwin to Emma Darwin 5 July 1844

As presciently anticipated by Darwin, natural selection is relevant to much more than the evolution of free-living species. The cells that make up multicellular organisms possess the requisite features for natural selection according to Darwin: heritable variation that impacts fitness. Cells, like species, are mutable, inevitably accumulating changes in their genomes due to extrinsic factors (e.g., radiation) and endogenous processes (e.g., errors in DNA replication and repair) (Alexandrov et al., 2013; Martincorena and Campbell, 2015). According to current estimates, most cells accumulate one to two mutations per cell division (Yizhak et al., 2018), though this rate may vary considerably (Hoang et al., 2016). Somatic mutations generate variety and starting from early embryogenesis, multicellular organisms become mosaics of genetically distinct cells (Behjati et al., 2014; Blokzijl et al., 2016; Ju et al., 2017). This variety creates a substrate for natural selection. Although few somatic mutations impact cell function (Martincorena et al., 2017), occasionally a mutation confers a fitness advantage, favouring clonal expansion of the cell harbouring it (Martincorena and Campbell, 2015; Yates and Campbell, 2012). The competitive advantage conferred by a given mutation may be context-dependent, varying with environmental exposures (Bondar and Medzhitov, 2010; Wong et al., 2015b; Yates and Campbell, 2012; Yokoyama et al., 2019). Cell competition has been most extensively studied in simpler model organisms, where it is often a beneficial physiological process that helps ensure that tissues are made up of the healthiest cellular constituents (Amoyel and Bach, 2014; Baker and Li, 2008). In humans, somatic evolution has primarily been studied in the context of cancer, where the process produces a cell with a complement of mutations enabling it to escape normal constraints on proliferation and to invade other tissues (Hanahan and Weinberg, 2000, 2011). However, recent studies of somatic mutation in the context of human development, ageing, pre-cancer, cancer and non-malignant disease have indicated that the border between normal age-related somatic evolution and malignancy can be indistinct (Martincorena et al., 2018; Martincorena et al., 2015; Moore et al., 2018; Salk et al., 2018; Yizhak et al., 2018; Yokoyama et al., 2019). This introduction will provide an overview of somatic evolution in cancer and ageing with a focus on the haematopoietic system, which has been particularly well characterised due to ease of representative tissue sampling.

1.1 Cancer is a genetic disease

"…a malignant cell is a cell with an irreparable defect, located in the nucleus. There is a permanent change in the condition of the chromatin which forces the cell to divide."

- Theodore Boveri, '*The Origin of Malignant Tumours*', 1914 (Manchester, 1995)

"I got sort of amused tolerance at the beginning."

- Janet Rowley recalling the response of the scientific community to her 1972 discovery that chromosomal translocations could cause cancer. (Fox, 2013)

The history of the mutational theory of cancer is a reminder of the power of simple experiments interpreted well and of the amount of time it can take for pivotal discoveries to elicit follow-up work and acceptance. Theodore Boveri is generally credited with being the first biologist to recognise that abnormal genetic content is responsible for malignant transformation (Rowley, 2001). His observations stemmed from meticulous light microscope scrutiny of sea urchin embryo divisions and the observation that aberrant mitoses seemed to trigger developmental defects.

"Experiments on sea urchin embryos have led to the result that most chromosome combinations that vary from the normal lead to the death of the cell; however, other combinations occur, in which the cell, while it remains viable, does not function in a typical way."

- Theodore Boveri, '*The Origin of Malignant Tumours*', 1914 (Manchester, 1995)

Boveri concluded that chromosomal content guides embryogenesis and further speculated that the entities responsible for Mendelian traits must reside within chromosomes:

"I feel beyond any doubt that the individual chromosomes must be endowed with different qualities and that only certain combinations permit normal development."

- Boveri, 1901 (Hardy and Zacharias, 2005)

''The probability is extraordinarily high that the traits examined in the Mendelian experiments are linked to individual chromosomes''

- Boveri, 1914 (Hardy and Zacharias, 2005)

These conclusions led Boveri to revisit observations made over twenty years previously by David Hansemann (1858–1920), a German pathologist who had documented asymmetrical nuclear segregation in a host of human cancers (Hardy and Zacharias, 2005). Hansemann maintained that nuclear abnormalities were most likely to represent characteristic sequelae of the malignant process (Hardy and Zacharias, 2005). Boveri, reinterpreting Hansemann's findings in the context of the sea urchin experiments, posited that cancers are the progeny of a single cell that acquired uncontrolled growth potential due to abnormal chromosomal content (Hardy and Zacharias, 2005; Manchester, 1995). Boveri’s hypothesis that chromosomes contained the material of inheritance was confirmed by the experiments of Avery, MacLeod and McCarty in 1944 (Avery et al., 1944). Further evidence that tumours often contain wildly bizarre chromosomes accumulated over the ensuing decades as cytogenetic methods improved. In the 1950s, Hauschka, Levan, Makino and others documented that most cancer cell lines contain aberrant chromosome numbers, as well as dicentric and ring chromosomes (Rowley, 2001). However, there was no apparent trend between particular abnormalities and cancer type, leading to further scepticism of any role in carcinogenesis (Rowley, 2001).

In the 1960s and 1970s, a clear association emerged between specific chromosomal abnormalities and particular leukaemias. In 1960, Nowell and Hungerford reported the Philadelphia (Ph) chromosome in almost all cases of chronic myeloid leukaemia (CML) (Nowell and Hungerford, 1960). Aided by improved chromosome banding techniques, Janet Rowley was able to establish that the Ph chromosome represented an interchange between chromosomes 9 and 22 (Rowley, 1973). Several other recurrent translocations were discovered in the 1970s by Rowley, Zech and others, notably the AML-associated t(8;21), t(8;14) in Burkitt lymphoma and t(15;17) in acute promyelocytic leukaemia (Rowley, 2001; Zech et al., 1976). It took until the early 1980s for the diagnostic and prognostic utility of these findings to be incorporated into clinical guidance (Rowley, 2001).

The advent of clinical cytogenetics coincided with further definitive proof that somatic mutations in DNA cause cancer. Weinberg, Cooper and colleagues demonstrated that human tumour DNA introduced into a mouse fibroblast cell caused malignant transformation (Krontiris and Cooper, 1981; Shih et al., 1981). Retrieval of the human sequence from the murine malignant cells ruled out spontaneous in vitro transformation, as can occur in many putatively normal cell lines (Krontiris and Cooper, 1981; Shih et al., 1981). Isolation of the oncogenic DNA fragment led to the discovery of an activating substitution mutation in *HRAS*, thus demonstrating for the first time that simple missense mutations, in addition to chromosomal rearrangements, can cause cancer (Reddy et al., 1982; Tabin et al., 1982). This discovery stimulated widespread concerted efforts to systematically identify genetic mutations capable of causing cancer.

Cancer gene discovery efforts further accelerated following the release of the first draft human genome sequence in 2000 (Lander et al., 2001; Venter et al., 2001) and the advent of massively parallel sequencing a few years later (Stratton, 2011; Stratton et al., 2009). The ensuing revolution in genomics has yielded unprecedented insights into the pathogenesis of cancer, as well as the inextricably related processes of human development and ageing. The next section will give an overview of some important concepts that have emerged from the study of the cancer genome.

1.1.1 Classifying mutations according to selection: 'driver' and 'passenger' mutations

To date, over 1.4 million tumour samples have been sequenced, including tens of thousands of whole genomes (Sondka et al., 2018). The ability to scrutinise whole genomes from diverse cancer types has revealed dramatic variation in somatic mutation burden, ranging from over 100 per megabase (Mb) in some melanomas and mismatch-repair deficient tumours to fewer than 0.01 mutations/Mb in some childhood cancers and leukaemias (Alexandrov et al., 2013; Shlien et al., 2015; Stratton, 2011).

A key focus of cancer genomics has been to classify somatic mutations according to whether or not they are under positive, neutral or negative selective pressure. Identifying the minority of mutations that are under positive selection and playing a causative role in oncogenesis (hereafter referred to as 'driver mutations') from mutations that do not confer a fitness advantage ('passenger mutations') is an ongoing and complex task (Lawrence et al., 2013; Martincorena et al., 2017; Stratton et al., 2009). The phenotypic features under positive selection in cancers have been conceptualised as the "hallmarks" of cancer and all, in essence, promote survival and/or growth (Hanahan and Weinberg, 2000, 2011). The most recent release of the Cancer Gene Census included 719 genes implicated in driving human cancers (Tate et al., 2019), although this list is constantly being amended and expanded to accommodate new genomic and functional evidence. The extent to which negative selection shapes somatic evolution in cancers and normal tissues is contentious, though at present most evidence suggests that positive selection plays a much more important role in governing clonal dynamics (Martincorena et al., 2017; Zapata et al., 2018).

1.1.2 Classifying cancer genes: tumour suppressors and oncogenes

Although often an oversimplification, it has proven conceptually useful to broadly classify cancer genes as either tumour suppressor genes or oncogenes. Tumour suppressor genes are implicated in oncogenesis through loss-of-function mutations (Stratton et al., 2009). Tumour suppressor genes frequently encode negative regulators of cell cycle progression (e.g., *RB1, PTEN*), suppressors of cell growth (e.g., *NF1*), pro-apoptotic signalling molecules (e.g., *DAXX*), proteins linking the DNA damage response to the cell cycle (e.g., *ATM*, *TP53*), cell-adhesion mediators (e.g., *APC*), DNA damage repair proteins (e.g., *BRCA1*) and epigenetic regulators (e.g., *KDM6A, SETD2, DNMT3A, TET2*) (Martincorena et al., 2017; Stratton, 2011). Many tumour suppressors, like the prototypical *RB1* that gave rise to Knudson's 'two-hit' hypothesis (Knudson, 1971), function in a recessive manner (Stratton, 2011). However, for many tumour suppressors, haploinsufficiency alone promotes cancer development (e.g., *TP53, RUNX1, PTEN, TET2, DNMT3A*)(Döhner et al., 2015; Inoue and Fry, 2017). Many types of mutations can inactivate tumour suppressor genes, including truncating mutations (e.g., nonsense, frameshift, disruptive rearrangements, essential splice site mutations, gene deletions) as well as variants that disrupt key functional domains (Inoue and Fry, 2017).

Oncogenes are implicated in cancer through activating mutations and often encode growth factors or cytokine receptors (e.g., *EGFR, JAK2, KIT, PDGFRA*), their downstream signalling mediators (e.g., *PIK3CA, BRAF, NRAS, KRAS*) or negative regulators of tumour suppressors (e.g., *PPM1D*) (Nangalia et al., 2016; Ruark et al., 2013; Stratton, 2011). The types of mutations that result in activation or upregulation of oncogenes are diverse and include canonical hotspot missense mutations (e.g. *JAK2* V617F, *BRAF* V600E), chromosomal translocations or gene amplifications as well as deletions or truncating mutations that disrupt inhibitory regulatory domains (e.g., truncating mutations in *PPM1D* exon 6, intragenic *BRAF* deletions)(Forbes et al., 2011; Ruark et al., 2013; Stratton, 2011; Wegert et al., 2018).

It is increasingly recognised that many cancer genes, particularly those implicated in epigenetic regulation, do not fit tidily into this classification scheme. Many function as either tumour suppressors or oncogenes in different cancer types or even at different stages of the same cancer type (e.g., *EZH2*), reflecting the influence of cell-type, developmental context and epistasis on the functional significance of many cancer driver mutations (Feinberg et al., 2016; Kim and Roberts, 2016; Shen et al., 2018; Van Vlierberghe and Ferrando, 2012).

Haematological cancers, and acute myeloid leukaemia in particular, are among the most extensively sequenced and genomically well-characterised of all cancer types (Medinger and Passweg, 2017; TCGA et al., 2013). Hence, the landscape of tumour suppressor and oncogenes relevant to these conditions has been well charted and the types of mutations that appear to be under positive selection in these genes is reasonably well defined, with concordance between many large studies (Bahr et al., 2018; Chen et al., 2018; Medinger and Passweg, 2017; Petti et al., 2018; TCGA et al., 2013; Tyner et al., 2018). The experiments described in this dissertation have taken a conservative approach to driver curation based on the criteria described in the largest relevant cancer genomics to date (Chapter 2).

1.1.3 Germline contributions to cancer risk

Studies of familial cancer predisposition and rare childhood cancer syndromes identified some of the first known cancer genes (Knudson, 1971; Maris, 2015). Germline variation plays an increasingly recognised role in cancer development, though its impact likely remains underestimated (Frick et al., 2018; Hermouet and Vilaine, 2011; Hinds et al., 2016; Huang et al., 2018; Loh et al., 2018; Parsons et al., 2016; Zhang et al., 2015). According to current estimates, overall approximately 1-2.7% of individuals without cancer have a putatively deleterious germline mutation in a cancer-associated gene, compared with 8.5 – 12.6% of cancer patients (Pritchard et al., 2016; Schrader et al., 2016; Zhang et al., 2015), though this rate appears considerably higher for some rare cancer types (Ballinger et al., 2016; Lu et al., 2015). Germline variants can influence cancer development by diverse mechanisms, including by directly driving clonal growth (Loh et al., 2018; Lu et al., 2015), increasing global mutation rate (Nik-Zainal, 2014; Shlien et al., 2015), increasing the likelihood of acquiring particular somatic driver events (Hermouet and Vilaine, 2011; Hinds et al., 2016; Loh et al., 2018) or altering carcinogen metabolism (Ding et al., 2010).

Studies of cancer predisposition syndromes have also demonstrated that the biological and clinical significance of germline and somatic variants in a given gene are often dramatically different (Maris, 2015; Maris and Knudson, 2015). For example, childhood myeloproliferative disease with germline mutations in *PTPN11* may follow an indolent, selfresolving course, whereas somatic *PTPN11* mutations presage rapid progression and warrant prompt haematopoietic stem cell transplantation (HSCT)(Hasle, 2016). Furthermore, germline and somatic mutations in several cancer genes, notably *TP53* and *RB1,* drive a distinct spectrums of cancer types with predilections for different tissues and age groups (Maris and Knudson, 2015). The distinction between germline and somatic drivers is particularly relevant when interpreting the results of unmatched sequencing experiments such as those described in this thesis, and will be discussed further later on.

1.1.4 Mutational signatures

The entire complement of somatic mutations in a genome constitutes a record of the types of mutational processes operative during the lifetime of the organism. Certain patterns of mutation are characteristic of particular mutagenic exposures. For example, ultraviolet light-induced pyrimidine dimers are typically repaired by transcription-coupled nucleotide excision repair, which tends to result in C>T mutations on the untranscribed strand (Alexandrov et al., 2013). Substitutions, small insertions and deletions (indels) and complex structural events can be classified according to sequence context, thus allowing formal mathematical extraction of mutational signatures (Alexandrov et al., 2013; Li et al., 2017; Petljak et al., 2014).

Substitution mutational signatures have been most extensively studied. The six types of substitution mutation (C>A, C>G, C>T, T>A, T>C and T>G) can be classified into 96 subtypes based on their trinucleotide context. Various statistical approaches, predominantly based on non-negative matrix factorisation, can discern distinct patterns of co-occurrence of substitution types (Alexandrov et al., 2018; Alexandrov et al., 2013). At present, only a minority of putative mutational signatures have a known cause (Alexandrov et al., 2018; Alexandrov et al., 2013). Nevertheless, mutational signature analysis has yielded compelling insights into the causes and epidemiology of several cancer types, and are increasingly being used clinically to guide diagnosis, prognostication and therapeutic strategy (Behjati et al., 2016; Hoang et al., 2013; Ma et al., 2018; Petljak and Alexandrov, 2016; Poon et al., 2015).

All cancers harbour a significant number of mutations attributed to ageing-associated single base substitution signatures 1 (SBS1) and 5 (SBS5) (Alexandrov et al., 2018; Alexandrov et al., 2013). SBS1 is dominated by C>T mutations attributed to spontaneous deamination of

5-methylcytosine, whilst SBS5 is of unknown aetiology (Alexandrov et al., 2018; Alexandrov et al., 2013). Myeloid malignancies are characterised by very low mutation burdens, similar to those observed in normal haematopoietic stem cells from age-matched individuals (Welch et al., 2012). Consistent with this finding, most of these mutations are attributable to SBS1 and SBS5 (Alexandrov et al., 2018; Alexandrov et al., 2013). A significant proportion of AML demonstrate evidence of SBS18, attributed to reactive oxygen species-mediated DNA damage (Alexandrov et al., 2018). A small proportion of myelodysplasia and myeloproliferative disease specimens harbour mutations attributable to SBS32, a signature thought to be caused by azathioprine treatment (Alexandrov et al., 2018). Although lymphoid neoplasms are also generally dominated by age-related SBS1 and SBS5 (Alexandrov et al., 2018; Alexandrov et al., 2013), they tend to have higher mutation burdens than myeloid cancers and a more complex mutational signature complement, with some specimens harbouring evidence of defective DNA repair mechanisms or APOBEC activity (Alexandrov et al., 2018; Alexandrov et al., 2013).

1.2 Cancer is an evolutionary process

The notion that cancer development is a clonal (originating from a single ancestral cell) evolutionary process can be traced back to Boveri and was further advanced in the 1950s based on histological observation of the natural history of precancerous lesions and their response to extrinsic irritants (Denoix, 1954; Foulds, 1958). Following the acceptance of the mutational theory of cancer, Peter Nowell and John Cairns conceptualised the modern understanding of cancer evolution in their seminal 1970s reviews (Cairns, 1975; Nowell, 1976).

“The acquired genetic instability and associated selection process, most readily recognized cytogenetically, results in advanced human malignancies being highly individual karyotypically and biologically. Hence, each patient's cancer may require individual specific therapy, and even this may be thwarted by emergence of a genetically variant subline resistant to the treatment. More research should be directed toward understanding and controlling the evolutionary process in tumors before it reaches the late stage usually seen in clinical cancer.”

- Peter Nowell, 1976 (Nowell, 1976)

Cairns spoke more explicitly in terms of natural selection acting on inevitable mutations arising in stem cells throughout the lifespan of an organism:

"Survival of the rapidly renewing tissues of long-lived animals like man requires that they be protected against the natural selection of fitter variant cells (that is, the spontaneous appearance of cancer)."

- Cairns 1975 (Cairns, 1975)

The ability to sequence many specimens of the same tumour type demonstrated remarkable genetic diversity within the same histopathological diagnosis (Yates and Campbell, 2012). Phylogenetic inference, multi-region tumour sequencing and single cell methods revealed striking intra-tumour heterogeneity (Anderson et al., 2011; Gerlinger et al., 2012; Greaves, 2015; Navin et al., 2011). These observations established that the evolutionary routes to cancer are diverse and that malignant clones continue to acquire mutations, compete and evolve (Ding et al., 2012; Greaves and Maley, 2012; Nik-Zainal et al., 2012). It became possible to construct phylogenetic trees at unprecedented resolution. Consistent features of these trees illustrate key principles of cancer pathogenesis. At their base, all cancer phylogenetic trees have the ancestral cell with the initial complement of driver mutations, along with all other mutations previously acquired by that cell and captured as the clone expanded (Yates and Campbell, 2012). Each cell within the expanded clone continues to acquire mutations, which are subclonal. With a few exceptions (e.g., chromothripsis causing multiple simultaneous driver mutations (Stephens et al., 2011)), in almost all cases cancer phylogenies support the gradual, multi-step model of carcinogenesis (Greaves, 2015; Yates and Campbell, 2012). Tumour cells continually diversify through acquisition of additional mutations and clonal architecture may follow branching, parallel or convergent evolutionary trajectories (Greaves, 2015; Yates and Campbell, 2012). The relative influence of mutation-induced cell-intrinsic growth advantage, selective pressures and genetic drift in cancer evolution remains contentious (Martincorena and Campbell, 2015; Martincorena et al., 2017; Sun et al., 2017; Zink et al., 2017). Phylogenetic trees constructed from multi-region or serial sampling have yielded insights into some of the selection pressures implicated in cancer clonal competition, discussed briefly in the next section.

1.2.1 Selection pressures shaping cancer evolution

1.2.1.1 The tumour microenvironment

The idea that the tumour microenvironment influences cancer development was first put forward in the late 19th century by Ernst Fuchs and Stephen Paget based on detailed anatomical studies of tumour metastases (Fuchs, 1882; Paget, 1889). Paget likened tumour cells to 'seeds' that required a favourable microenvironment, or 'soil' to survive and grow (Paget, 1889). The factors underpinning the predilection of metastases for certain organs are still incompletely understood (Hunter et al., 2018). However, several studies that used multi-region sampling or tumour organoids have elucidated the phylogenetic relationships between primary tumour lesions and metastases and provided insight into the interplay between genetic diversification and organ-specific selection pressures (Altorki et al., 2019; Campbell et al., 2010; Gundem et al., 2015; Hunter et al., 2018; Makohon-Moore and Iacobuzio-Donahue, 2016; Roerink et al., 2018; Yachida et al., 2010). It is now clear that interactions between cancer cells and tissue microenvironment are relevant far beyond metastasis, exerting selective pressures important at all stages of solid and haematological cancer development (Medyouf, 2017; Scott and Gascoyne, 2014; Yates and Campbell, 2012; Yokoyama et al., 2019).

1.2.1.2 Cancer therapies

Anticancer therapy is often one of the most potent selective pressures governing cancer evolution (Yates and Campbell, 2012). Resistance mechanisms are diverse (Holohan et al., 2013), however, as sequencing technologies become more sensitive, it is increasingly clear that resistance mutations to both conventional cytotoxic agents and targeted therapies frequently predate treatment at extremely low subclonal levels (Karoulia et al., 2017; Kennedy et al., 2014; Schmitt et al., 2016; Wong et al., 2015a; Wong et al., 2015b). As presciently anticipated by Nowell (Nowell, 1976), the extensive genetic diversity present in fully fledged cancers represents a formidable arsenal of potential adaptive strategies and has greatly undermined targeted therapy efforts (Holohan et al., 2013).

Scrutiny of cancer genomes has yielded profound insight into the genetic drivers and evolutionary dynamics of most human cancer types. However, it is now evident that this work did not adequately capture the somatic genetic diversity and selective pressures shaping the pre-cancerous phases of oncogenesis. Recent studies of somatic evolution in morphologically normal tissues have yielded compelling biological insights into normal ageing and its relationship with cancer development. The next section will give a broad overview of these advances with a focus on the haematopoietic system.

2. Somatic evolution in normal ageing tissues and its relationship to cancer

"*Cancer is a chronic disease with a long history extending back for many years before clinical signs are evident*."

- Leslie Foulds, 1958 (Foulds, 1958)

"…the whole body is seeded with tumor cells whose evolutionary potential is revealed at unpredictable times thereafter."

- Foulds's summary of a hypothesis proposed by Pierre Denoix in his 1954 paper 'De la diversité de certains cancers' (Denoix, 1954; Foulds, 1958)

The molecular basis of multi-step carcinogenesis was meticulously dissected in childhood leukaemia and colon cancer in the 1980s and 1990s and gave preliminary insights into the ambiguous boundary between normal tissue, pre-cancer and fully-fledged malignancy (Fearon and Vogelstein, 1990; Greaves et al., 2003). Studies of monozygotic twins concordant for leukaemia demonstrated that the initiating event, typically a fusion gene, arises in a single cell *in utero*, which transfers to the second twin via a monochorionic placenta (Greaves and Wiemels, 2003). For most childhood leukaemia, the latency to disease onset suggested that the initiating translocation (most commonly the *TEL–AML1* fusion gene), requires a second hit to trigger malignant transformation (Greaves and Wiemels, 2003). In support of this hypothesis, several studies screened healthy newborns for leukaemogenic fusions and found their prevalence to be considerably higher than the cumulative incidence of childhood leukaemia (Greaves et al., 2011; Lausten-Thomsen et al., 2011; Mori et al., 2002; Zuna et al., 2011). Furthermore, not all twins concordant for the initiating event are concordant for leukaemia (Bateman et al., 2015). Collectively, these findings provided genetic evidence to support Foulds and Denoix's hypothesis that pre-cancer is considerably more common than cancer and that malignant progression is not readily predictable. These conclusions were also supported by the natural history and molecular features of the adenoma-carcinoma sequence in the colon (Fearon and Vogelstein, 1990).

The advent of sensitive sequencing methods has recently revealed that potentially pre-malignant clonal expansions are remarkably common in many normal ageing tissues (Bowman et al., 2018; Martincorena et al., 2018; Martincorena et al., 2015; Moore et al., 2018; Salk et al., 2018; Suda et al., 2018; Yizhak et al., 2018; Yokoyama et al., 2019). This phenomenon has been most extensively explored in skin (Martincorena et al., 2015), oesophagus (Martincorena et al., 2018; Yokoyama et al., 2019), endometrium (Moore et al., 2018; Salk et al., 2018; Suda et al., 2018) and blood (Bowman et al., 2018), though preliminary evidence from bulk RNA sequencing of diverse normal tissues suggests that clonal expansions harbouring canonical cancer driver mutations may be ubiquitous in most organs (Yizhak et al., 2018).

Several common themes are beginning to emerge from these findings. Firstly, there is generally a clear association between age and prevalence of readily detectable clonal expansions, with that latter apparently trending towards inevitability by midlife in many tissues (Martincorena et al., 2018; Martincorena et al., 2015; Suda et al., 2018; Young et al., 2016). However, it is not yet clear to what extent age-related mutation acquisition is a rate-limiting step in clonal expansion. Potent cancer driver mutations, including hotspot *TP53* mutations, may be dated to early infancy or childhood in several tissues and may never contribute to cancer even in high risk individuals (Greaves et al., 2011; Moore et al., 2018; Yokoyama et al., 2019). It is increasingly apparent that selective pressures, some correlated with ageing, impact the fitness advantage of particular mutations and hence modulate clonal dynamics (Hsu et al., 2018; McKerrell and Vassiliou, 2015; Murai et al., 2018; Wong et al., 2015b; Yokoyama et al., 2019). For example, exposure to smoking and alcohol accelerates clonal growth in normal oesophagus (Yokoyama et al., 2019) and ultraviolet radiation exposure influences the fitness advantage of epidermal *TP53* mutations (Murai et al., 2018). The proliferation of clonal expansions with age may reflect both mutation accrual and ageing-associated changes in tissue microenvironments that confer increasing fitness advantage on oncogenic mutations (Armitage and Doll, 1954; Nordling, 1953; Rozhok and DeGregori, 2015).

A second observation that has been made in several tissue types is that the mutational spectrum of age-associated clonal expansions may differ from that seen in cancer (Busque et al., 2018; Martincorena and Campbell, 2015; Martincorena et al., 2018; Xie et al., 2014; Yokoyama et al., 2019). For example, putative driver mutations in *NOTCH1* are more frequently seen in clonal expansions in histologically normal skin and oesophagus than in cancers arising from these tissues (Martincorena et al., 2018; Martincorena et al., 2015; Yokoyama et al., 2019). Similarly, activating mutations in *PPM1D*, which encodes a negative regulator of TP53, are more frequent in normal blood and oesophagus than in malignancy (Bowman et al., 2018; Xie et al., 2014; Yokoyama et al., 2019). Most relevant experiments have employed targeted sequencing of known cancer-associated genes, thus hindering an unbiased comparison between the mutational landscape of cancer and normal ageing. Equally, the ubiquity of certain mutations in normal tissues, and by extension their recurrence in the trunks of tumour phylogenetic trees, could lead to overestimates of their importance in cancer pathogenesis (Ciccarelli, 2019).

How mutations and selective pressures interact to determine the likelihood of malignant transformation is an important biological question with compelling clinical implications. As predicted by Cairns (Cairns, 1975), emerging evidence suggests that some epithelial tissues have evolved mechanisms for restraining growth of clones harbouring oncogenic mutations (Murai et al., 2018; Ying et al., 2018). Senescence and immune surveillance are also involved in policing mutated clones (Collado et al., 2005; Schreiber et al., 2011). However, understanding of the factors governing physiological cell competition and tissue homeostasis in humans and their relationship with carcinogenesis remains very limited. A significant obstacle to studying these questions in most organs is the inability to obtain representative tissue samples. The haematopoietic system has proven a privileged setting in which to explore somatic evolution and its relationship with ageing and ageing-associated pathologies (Bowman et al., 2018; Geiger et al., 2013; Latchney and Calvi, 2017; Lee-Six et al., 2018). The next section will summarise current understanding of clonal haematopoiesis and its clinical relevance.

3. Clonal haematopoiesis

3.1 Prevalence and mutational landscape of clonal haematopoiesis

Blood has one of the highest turn-over rates of any tissue, necessitating the production of trillions of cells per day by a population of haematopoietic stem cells (HSCs) estimated to number between 50,000 and 200,000 (Carrelha et al., 2018; Doulatov et al., 2012; Lee-Six et al., 2018). Replicative mutagenesis and other sources of genotoxic stress cause HSCs to accumulate DNA damage with age, with an estimated 14 mutations accumulating per cell per year (Flach et al., 2014; Osorio et al., 2018; Rossi et al., 2007; Welch et al., 2012; Yahata et al., 2011). Clonal haematopoiesis (CH) refers to the disproportionate expansion of one somatically mutated HSC clone relative to others. Many reports have now identified this phenomenon in a significant proportion of individuals without a haematological cancer (Acuna-Hidalgo et al., 2017; Akbari et al., 2014; Artomov et al., 2017; Bonnefond et al., 2013; Buscarlet et al., 2017; Busque et al., 1996; Busque et al., 2012; Coombs et al., 2017; Forsberg et al., 2012; Frick et al., 2018; Genovese et al., 2014; Gibson et al., 2017; Gillis et al., 2017; Jacobs et al., 2012; Jaiswal et al., 2014; Jaiswal et al., 2017; Laurie et al., 2012; Loftfield et al., 2018b; Loh et al., 2018; Machiela et al., 2015; McKerrell et al., 2015; Rodriguez-Santiago et al., 2010; Savola et al., 2017; Schick et al., 2013; Takahashi et al., 2017; Thompson et al., 2019; Vattathil and Scheet, 2016; Xie et al., 2014; Young et al., 2016; Zhou et al., 2016; Zink et al., 2017). Clonal haematopoiesis was first recognised in the 1990s when Busque and colleagues demonstrated that ageing was associated with increasingly skewed X-inactivation in blood cells (Busque et al., 1996). Busque et al. applied a PCR-based X-inactivation clonality assay to peripheral blood samples from a cohort of 295 healthy females spanning a broad age range (Busque et al., 1996). Using stringent criteria for skewing (allele ratios >= 10:1), this approach identified imbalanced X-inactivation in 22.7%, 4.5% and 1.9% of women aged >=60 years, 28-32 years and <1 month, respectively (Busque et al., 1996).

The advent of molecular karyotyping using SNP arrays demonstrated that a significant proportion of the general population harbours clonal, somatic chromosomal abnormalities in blood cells (Artomov et al., 2017; Bonnefond et al., 2013; Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012; Loftfield et al., 2018a; Loh et al., 2018; Machiela et al., 2015;

Rodriguez-Santiago et al., 2010; Schick et al., 2013; Vattathil and Scheet, 2016; Zhou et al., 2016). These studies identified a clear correlation between age and frequency of clonal mosaic aneuploidy or copy-neutral loss of heterozygosity (LOH) events, with prevalence varying from <0.5% in individuals under age 50 years to 1.9-3.4% in persons aged >60 (Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012). The most recurrent abnormalities included del(13q), trisomy 8, del(20q), del(5q) and del(7q), chromosomal changes characteristic of haematological malignancies (Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012). Mosaic chromosomal changes were associated with a five- to ten-fold higher risk of subsequently developing haematological cancers (Jacobs et al., 2012; Laurie et al., 2012; Schick et al., 2013). Longitudinal tracking of clonal chromosomal abnormalities has yielded variable results, with one study suggesting that aberrant clones may become undetectable over time (Forsberg et al., 2017), while another series of 47 individuals sampled several years apart found that most clones expanded with age (Machiela et al., 2015).

Next-generation sequencing technologies enabled higher resolution scrutiny of the genetic changes driving clonal haematopoiesis. Sequencing of healthy women with skewed X-inactivation identified mutations in the epigenetic regulator *TET2* in 5.5% (10/182 individuals) (Busque et al., 2012). In 2014, three large exome sequencing studies identified leukaemia-associated point mutations in the blood of >2% of individuals unselected for haematological phenotypes (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). All three studies reported a steep rise in CH prevalence with age, ranging from <1% under age 50 years to around 10% in individuals over age 70 (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). The majority of candidate driver mutations occurred in *TET2*, *DNMT3A* and *ASXL1*, epigenetic regulators commonly mutated in myeloid malignancies (Arber et al., 2016; Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). Jaiswal et al. interrogated a predefined set of 160 cancer-associated genes, whereas Genovese et al. and Xie et al. screened for CH in an unbiased manner on the basis of unusual allele frequencies (Genovese et al., 2014; Xie et al., 2014). The latter approach identified a broader spectrum of putative CH drivers, most notably a remarkably high frequency of mutations in *PPM1D*, a negative regulator of TP53 that is infrequently mutated in haematological or solid cancers (Genovese et al., 2014; Ruark et al., 2013; Xie et al., 2014). Other recurrently mutated genes included *JAK2, TP53*, spliceosome genes (*SF3B1, SRSF2* and *U2AF1*), *CBL, BCORL1, ATM, MYD88* and *GNAS* (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014).

Later studies of CH in the general population used more sensitive targeted sequencing approaches and demonstrated that CH prevalence increases dramatically with assay sensitivity (Acuna-Hidalgo et al., 2017; Buscarlet et al., 2017; McKerrell et al., 2015; Young et al., 2016; Zink et al., 2017). Young et al. used molecular barcoding to enable detection of mutations at a variant allele frequency (VAF) as low as 0.0003 and found CH to be ubiquitous in otherwise healthy individuals aged >50 years (Young et al., 2016). The genes recurrently implicated in CH were broadly consistent across these studies. However, whilst the prevalence of mutations in all genes increased with age, certain mutations were found to be particularly enriched in older individuals (McKerrell et al., 2015). In particular, spliceosome gene mutations were seen almost exclusively in individuals aged >70 (Acuna-Hidalgo et al., 2017; McKerrell et al., 2015), whereas the frequency of mutations in *DNMT3A* and *JAK2* increased more linearly with age (Acuna-Hidalgo et al., 2017; Buscarlet et al., 2017; McKerrell et al., 2015). A less dramatic age-dependence has been observed for *TET2* mutations (Buscarlet et al., 2017).

Ageing is just one example of how the mutational landscape of CH varies according to clinical context. CH is extremely common in aplastic anaemia patients and displays a distinct spectrum of somatic mutations (Stanley et al., 2017; Yoshizato et al., 2015). Similarly, CH enriched in *TP53* and *PPM1D* mutations is prevalent in individuals who have been exposed to chemo- and/or radiotherapy (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017). Further discussion of the interplay between somatic mutations and dynamic selection pressures is discussed in section 3.4.

Zink et al. conducted a broader, though less sensitive, screen for CH by interrogating 11,262 whole genomes (median coverage 35x) for unusual SNV allele frequency distribution, similar to the variant calling strategies applied by Xie et al. and Genovese et al. (Genovese et al., 2014; Xie et al., 2014; Zink et al., 2017). Consistent with previous data and predictions, CH was almost universally detectable in individuals >85 years of age (McKerrell et al., 2015; Young et al., 2016; Zink et al., 2017). The overall prevalence of CH (identified on the basis of having > 20 putative mosaic point mutations) was 12.5%, higher than that observed in previous studies (Zink et al., 2017). Presumptive driver mutations were most frequent in *DNMT3A*, *TET2*, *ASXL1* and *PPM1D* (Zink et al., 2017). However, candidate driver mutations were only identified in a minority of individuals with CH (Zink et al., 2017). The authors suggest genetic drift as a likely explanation for this result. However, numerical and structural chromosomal changes were not systematically identified and may account for a significant proportion of the CH cases without an apparent point mutation driver (Artomov et al., 2017; Bonnefond et al., 2013; Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012; Loftfield et al., 2018a; Loftfield et al., 2018b; Loh et al., 2018; Machiela et al., 2015; Rodriguez-Santiago et al., 2010; Schick et al., 2013; Vattathil and Scheet, 2016; Zhou et al., 2016). Contiguous gene deletions and rearrangements are common initiating driver events in many haematological cancers. It is possible that structural variants under positive selection underpinned a significant proportion of the CH cases attributed to drift. It is also conceivable that there is only partial overlap between cancer drivers and the mutations that are under positive selection in somatic evolution in normal ageing blood. The preponderance of *PPM1D* and *NOTCH1* mutations in clonal expansions in normal tissues compared to cancers may support this hypothesis (Bowman et al., 2018; Martincorena et al., 2018; Martincorena et al., 2015; Yokoyama et al., 2019). Zink et al. did perform an unbiased search for novel driver genes, but did not identify many candidates (Zink et al., 2017).

Mutations in certain common myeloid cancer genes, notably *FLT3* and *NPM1*, were consistently absent in even the most sensitive CH screens, supporting their role as late cooperating/transforming mutations rather than initiating events (Acuna-Hidalgo et al., 2017; Genovese et al., 2014; Jaiswal et al., 2014; McKerrell et al., 2015; Xie et al., 2014).

3.2 Germline influences on CH

Extensive evidence demonstrates that germline variation is an important determinant of clonal haematopoiesis risk and clinical outcome (Buscarlet et al., 2017; Frick et al., 2018; Hinds et al., 2016; Jones et al., 2009; Kilpivaara et al., 2009; Koren et al., 2014; Loftfield et al., 2018a; Loh et al., 2018; Olcaydu et al., 2009; Thompson et al., 2019; Wright et al., 2017; Zhou et al., 2016; Zink et al., 2017). Heritable polymorphisms can influence CH development by increasing susceptibility to somatic mutagenesis (Hinds et al., 2016; Jones et al., 2009; Kilpivaara et al., 2009; Koren et al., 2014; Loh et al., 2018; Olcaydu et al., 2009; Zhou et al., 2016) or by modulating positive or negative clonal selection (Hinds et al., 2016; Loh et al., 2018). For example, the *JAK2* 46/1 haplotype is a well-recognised risk factor for acquiring *JAK2* V617F-positive CH and progressing to a myeloid neoplasm (Jones et al., 2009; Kilpivaara et al., 2009; Olcaydu et al., 2009). Polymorphisms in several other genes, including *TERT, TET2, ATM* and *CHEK2*, are also associated with *JAK2* V617-driven myeloproliferative neoplasms and hence perhaps also antecedent clonal haematopoiesis (Hinds et al., 2016). Over 150 loci have now been strongly linked to overall CH risk, or risk of particular chromosomal losses or likelihood of specific LOH events amplifying the selective advantage conferred by inherited or somatic driver events (Loh et al., 2018; Thompson et al., 2019; Wright et al., 2017; Zink et al., 2017). Additionally, several germline polymorphisms have been shown to impact leucocyte DNA replication timing, and by consequence, the susceptibility of nearby sequence to somatic mutagenesis (Koren et al., 2014). In a recent large survey of mosaic chromosomal changes in peripheral blood, Loh et al. identified several highly penetrant heritable variants associated with increasing mutability of nearby DNA sequence, including in the myeloid oncogene *MPL* (Loh et al., 2018). Several of the variants were also subject to clonal selection and impacted risk of progression to haematological cancer (Loh et al., 2018).

A main emerging message from these studies is the increasingly blurry distinction between heritable and somatically acquired determinants of clonal haematopoiesis development and natural history. Furthermore, the influence of germline variation on CH incidence and outcome probably remains underestimated. Several studies report familial or ethnic clustering of CH suggesting yet to be discovered heritable risk factors (Buscarlet et al., 2017; Frick et al., 2018; Loftfield et al., 2018a). Moreover, a large number of uncommon germline variants have emerged as important determinants of haematological phenotypes in the general population, and it is plausible that these exert epistatic, lineage biased effects on CH evolution (Astle et al., 2016).

3.3 Clinical significance of clonal haematopoiesis

3.3.1 Impact of clonal haematopoiesis on blood indices

Mutations common in CH are implicated in ineffective haematopoiesis, impaired differentiation and cytopenias when they occur in individuals with MDS or AML (Papaemmanuil et al., 2016; Steensma et al., 2015). However, CH harbouring putative driver mutations (CH-PD) is not generally associated with any abnormalities in blood cell counts (Buscarlet et al., 2017; Jaiswal et al., 2014; McKerrell et al., 2015). Jaiswal et al. analysed blood indices data available for 3107 individuals, 4.5% of whom had CH-PD and found no significant differences in haemoglobin levels, platelet counts or white-cell differential counts (Jaiswal et al., 2014). The only blood index that differed significantly was red cell distribution width (RDW), which was higher in individuals with CH-PD and correlated with mutation VAF (Jaiswal et al., 2014). Moreover, although the prevalence of a single cytopenia was not influenced by CH status, individuals with multiple cytopenias were more likely to have CH (odds ratio 3.0)(Jaiswal et al., 2014).

While CH may rarely cause haematological indices to deviate to a clinically significant degree, Loh et al. recently demonstrated that some acquired mutations correlate with trends in blood counts, though generally within the reference range (Loh et al., 2018). Their findings suggest lineage-specific clonal selection pressures mirroring those observed in blood cancers (Loh et al., 2018). For example, chromosome 9p LOH (encompassing *JAK2*) and trisomy 12 (highly recurrent in CLL) were associated with higher granulocyte and lymphocyte counts, respectively (Loh et al., 2018).

3.3.2 Clonal haematopoiesis and haematological malignancy

Numerous studies have reported a clear association between CH in haematologically normal individuals and risk of developing a haematological malignancy (Coombs et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Gillis et al., 2017; Greaves and Wiemels, 2003; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Loh et al., 2018; Schick et al., 2013; Takahashi et al., 2017; Zink et al., 2017). This is perhaps unsurprising given that the multi-step model of cancer implies a premalignant phase in cancer evolution (Yates and Campbell, 2012). Furthermore, several studies of haematological cancer evolution have demonstrated that myeloid malignancies evolve from a population of preleukaemic stem cells harbouring initiating driver mutations, and that such preleukaemic HSCs can persist during long-term remission and serve as a reservoir for relapse (Greaves et al., 2003; Jan et al., 2012; Shlush et al., 2017; Shlush et al., 2014). Similar observations hold true for the commonest lymphoid malignancies (Landgren et al., 2009; Ojha et al., 2014; Rawstron et al., 2008). However, the prevalence of preleukaemic HSC clones and the rate and determinants of progression to leukaemia remain unknown. The studies cited above demonstrate that the rate of CH in the general population, and in particular CH harbouring putative driver mutations (CH-PD), vastly exceeds the cumulative incidence of blood cancers (Bowman et al., 2018). Given the variation in cohort characteristics, follow-up time and CH detection sensitivity, it is unsurprising that the strength of the association reported between CH and haematological cancer risk has varied between studies (Coombs et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Gillis et al., 2017; Greaves and Wiemels, 2003; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Loh et al., 2018; Schick et al., 2013; Takahashi et al., 2017; Zink et al., 2017). Notably, Zink et al. and Genovese et al. found that the risk of malignant progression was the same regardless of whether a point mutation driver (versus no driver) was identified (Genovese et al., 2014; Zink et al., 2017). However, as discussed previously, it is possible that CH without such mutations may reflect unsought structural driver events.

Most studies of cohorts unselected for cancer or haematological phenotype have reported an approximately ten-fold increased risk of blood cancer among individuals with CH (Genovese et al., 2014; Jaiswal et al., 2014). However, this still reflects a low absolute risk for malignant progression. Jaiswal et al. found that individuals with CH-PD (assay sensitivity limit 3.5% and 7.0% for SNVs and indels, respectively) had a 4% risk of blood cancer diagnoses over a median follow-up period of 7.9 years (Jaiswal et al., 2014). This translates into an overall annual progression rate of 0.5%, rising to 1% per year among individuals with driver mutations present at VAF > 0.1 (Jaiswal et al., 2014). Similarly, Genovese et al. reported similar findings, and in addition were able to demonstrate a clonal relationship between CH and blood cancer in the two individuals for whom diagnostic bone marrow specimens were available (Genovese et al., 2014). In both of these cases, the interval between blood sampling and cancer diagnosis was modest (2 and 34 months) (Genovese et al., 2014). Both Jaiswal et al. and Genovese et al. found that only a minority of the blood cancers arising during follow-up were diagnosed in individuals with antecedent CH: 5/16 (31%) and 13/31 (42%), respectively (Genovese et al., 2014; Jaiswal et al., 2014). This finding, in conjunction with the ubiquity of CH relative to blood cancer incidence, raises clinically and biologically compelling questions about the natural history of haematological cancers and the pathophysiological relevance of CH.

From a clinical perspective, it is sobering that the main cause of mortality from many of the commonest adult haematological cancers remains treatment resistance, despite a growing arsenal of novel targeted therapies (Abdi et al., 2013; Döhner et al., 2015; Woyach and Johnson, 2015). There is hence a compelling rationale for identifying and treating a genomically simpler antecedent of the disease. In this context, reduction of clonal size rather than complete clonal extinction may be sufficient to significantly reduce the risk of malignant progression. Such an approach has proven very effective in CML, which has been transformed into a chronic condition by targeted therapy, whereas CML blast crisis remains very challenging to treat (Gore et al., 2018; Hunger, 2017; O'Brien et al., 2003). The eventual feasibility of earlier detection and intervention for nascent blood cancers will invariably be hampered by the high prevalence of benign CH, given the relative rarity of the former. However, CH is associated with and may play a causal role in several much commoner conditions, which may broaden indications for its use as a clinical biomarker or a therapeutic target for non-haematological pathologies. The broader clinical significance of CH is summarised in the following sections.

3.3.3 Clonal haematopoiesis and non-haematological cancers

Clonal haematopoiesis has been associated with both a higher risk of solid cancers (Akbari et al., 2014; Artomov et al., 2017; Bowman et al., 2018; Ruark et al., 2013; Thompson et al., 2019) and with higher mortality among solid tumour and lymphoma patients (Coombs et al., 2017; Gibson et al., 2017). However, it is challenging to study the relationship between CH and solid cancer risk given that cancer treatments dramatically increase CH incidence and many study participants were not chemotherapy/radiotherapy naïve (Akbari et al., 2014; Artomov et al., 2017; Ruark et al., 2013). It is also possible that germline cancer predisposition is a confounding risk factor for both CH and overall cancer risk.

The association between CH and mortality among cancer patients has been consistently observed across diverse cohorts (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017), though may also be subject to some confounding factors, e.g., germline cancer predisposition. Furthermore, cancer treatment intensity correlates with CH risk (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017) and toxicity-related mortality, and may be higher in individuals with more advanced malignancies. These potential confounders are hard to control for across retrospective cohorts comprising individuals with diverse solid cancer types.

Any mechanistic link between CH and solid tumour pathogenesis remains speculative. It is possible that clonal haematopoiesis may promote solid tumour growth by fostering hospitable tissue microenvironments (Bowman et al., 2018). The term 'tumour-associated macrophage' (TAM) encompasses phenotypically diverse cells that can play both oncogenic or tumour-suppressive roles (Mantovani et al., 2017). It is intriguing that the cytokine profile of the *TET2*-mutated macrophages implicated in atherosclerosis (Jaiswal et al., 2017) shares key features with that seen in oncogenic TAMs (Storr et al., 2017; Wang et al., 2018).

3.3.4 Clonal haematopoiesis and non-malignant conditions

Several studies have found that clonal haematopoiesis is associated with a higher overall mortality rate that is only partially due to cancer deaths (Coombs et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Jaiswal et al., 2014; Loftfield et al., 2018a; Loh et al., 2018; Zink et al., 2017). The majority of excess mortality has been attributed to cardiovascular disease (CVD), ischaemic stroke and diabetes (Bonnefond et al., 2013; Coombs et al., 2017; Fuster et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Jaiswal et al., 2014; Jaiswal et al., 2017; Loftfield et al., 2018a; Sano et al., 2018a; Sano et al., 2018b). Preliminary evidence also links CH with rarer inflammatory conditions, such as rheumatoid arthritis (Savola et al., 2017).

It has long been recognised that known cardiovascular risk factors - namely hypertension, lipid profile, smoking and obesity – only partially account for atherosclerotic diseases burden and that other poorly characterised pro-inflammatory processes likely contribute (Ross, 1999). A large prospective case-control study recently confirmed the association between CH and risk of coronary heart disease, independent of age and other known risk factors (Jaiswal et al., 2017). This association held regardless of whether CH harboured mutations in *DNMT3A, TET2, JAK2* or *ASXL1* (Jaiswal et al., 2017). Individuals with CH had significantly more coronary artery calcification, a surrogate marker of atherosclerosis severity (Jaiswal et al., 2017). Moreover, compelling evidence now supports a causal role for CH in atherosclerosis and cardiometabolic disease (Fuster et al., 2017; Jaiswal et al., 2017; Sano et al., 2018a). Jaiswal et al. engrafted *TET2*-mutated cells into hypercholesterolaemia-prone mice and found that the TET2-deficient animals developed accelerated atherosclerotic disease (Jaiswal et al., 2017). Transcriptional profiling of *TET2*-mutant macrophages from arterial plaques revealed increased expression of pro-inflammatory mediators implicated in atherosclerosis, including CXCL1, CXCL2, IL-1b and IL-6 (Jaiswal et al., 2017). These findings were corroborated by a similar mouse model study by Fuster et al., which further demonstrated that inhibition of IL-1b secretion was more effective in slowing atherosclerosis in mice engrafted with TET2-deficient bone marrow than in controls (Fuster et al., 2017). Sano et al. found that *TET2*-mutant CH increases IL-1b levels, accelerates cardiac failure in mice, and can be mitigated with anti-inflammatory therapy targeting IL-1b production (Sano et al., 2018a). A recent randomised, double blind trial of canakinumab, a therapeutic monoclonal antibody targeting IL-1b, reduced cardiovascular morbidity and mortality in humans independent of lipid profile (Ridker et al., 2017). Trial participants were not screened for CH, so it remains to be investigated whether CH could serve as a useful human biomarker or therapeutic target in its own right.

Myeloproliferative diseases are associated with increased cardiovascular morbidity and mortality mediated by multiple mechanisms (Deininger et al., 2017). In a retrospective nested case-control study including 10,000 individuals without a known myeloid neoplasm, *JAK2*-mutant CH was associated with an increased thrombosis risk (Wolach et al., 2018). This association appears at least partially attributable to a mutant *JAK2*-mediated increase in pro-thrombotic neutrophil extracellular trap (NET) formation (Wolach et al., 2018). In a mouse model of *JAK2*-mutant CH, NET formation and thrombosis was reduced upon administration of ruxolitinib, a *JAK2* inhibitor (Wolach et al., 2018).

It is not yet known whether CH with mutations in other genes plays a causative role in atherosclerosis, though the strong association between *DNMT3A*- and *ASXL1*-mutant CH and CVD (Jaiswal et al., 2017) warrants further investigation. It is intriguing that atherogenic haemodynamic stress appears to reprogram endothelial gene expression via a DNA methyl-transferase (DNMT)-dependent mechanism and that DNMT inhibition with siRNA or decitabine can reduce vascular endothelial inflammation and atherosclerosis formation in multiple mouse models (Dunn et al., 2014; Zhou et al., 2014). It is therefore possible that *DNMT3A*-mutant CH promotes endothelial dysfunction by epigenetic mechanisms, and might conceivably be amenable to nucleoside analogue treatment.

The hypothesis that CH can contribute to inflammatory conditions is further substantiated by a recent study investigating the impact of donor CH on allogeneic haematopoietic stem cell transplantation (HSCT) outcomes (Frick et al., 2018). Frick et al. found that recipients of CH-positive transplants had a significantly higher rate of chronic graft versus host disease and lower rate of relapse (Frick et al., 2018).

Collectively, these studies suggest a causal link between CH and non-malignant conditions, including leading causes of morbidity and mortality in the general population. It is therefore possible that CH may prove to be a useful biomarker and/or modifiable risk factor in a range of clinical contexts.

3.4 Selection pressures influencing clonal haematopoiesis

Which selective pressures influence somatic evolution in the haematopoietic system? Do certain driver events confer strong enough cell-intrinsic growth advantage that they render clonal expansion inevitable? To what extent do environmental selection pressures determine the fitness advantage conferred by mutations and the pathophysiological outcome of CH? Are any of these selective pressures clinically modifiable? Although these questions remain largely unanswered, it is clear that the incidence and natural history of CH is influenced by clinical context.

3.4.1 Ageing

CH prevalence consistently rises with age, which is itself the dominant risk factor for most haematological malignancies (Busque et al., 2018). Haematopoietic ageing is characterised by HSC functional decline and myeloid bias reflected in a tendency towards anaemia and innate and adaptive immune senescence (Pang et al., 2011; Rossi et al., 2007; Rossi et al., 2005). Although HSCs accumulate mutations throughout life, ageing is associated with accelerated accrual of DNA damage (Flach et al., 2014; Osorio et al., 2018; Rossi et al., 2007; Welch et al., 2012). Age-associated genotoxic stress can induce apoptosis or differentiation, thus potentially depleting the functional HSC pool (Adams et al., 2015; Flach et al., 2014; Geiger et al., 2013; Rossi et al., 2007; Yahata et al., 2011). These factors may create an environment where HSCs with greater proliferative capacity or resistance to DNA-damage induced apoptosis and/or terminal differentiation contribute disproportionately to haematopoiesis (Latchney and Calvi, 2017; Pang et al., 2017). Mutations in many recurrent CH drivers, notably *DNMT3A, ASXL1* and *TET2*, may confer a competitive advantage through their ability to increase HSC self-renewal and inhibit differentiation (Abdel-Wahab et al., 2012; Challen et al., 2011; Dominguez et al., 2018; Jeong et al., 2018; Ko et al., 2011; Moran-Crusio, 2011). Similarly, HSC harbouring mutations in *TP53* or *PPM1D* are likely to have a particular competitive advantage in the context of genotoxic stress (Bondar and Medzhitov, 2010; Hsu et al., 2018; Kahn et al., 2018; Wong et al., 2015b).

3.4.2 Cytotoxic therapies

Studies of CH in cohorts of cancer patients who have received intensive chemo and/or radiotherapy have demonstrated an elevated prevalence of CH with marked enrichment for *PPM1D* and *TP53* mutated clones (Akbari et al., 2014; Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Ruark et al., 2013; Takahashi et al., 2017). These findings suggest that exogenous genotoxic stress confers a strong competitive advantage on HSCs harbouring mutations that interfere with the DNA-damage response and apoptosis. In vivo studies of murine HSC competition have demonstrated that cells with *TP53* or *PPM1D* mutations outcompete their wild-type peers in the context of ionising radiation and chemotherapy, respectively (Bondar and Medzhitov, 2010; Hsu et al., 2018; Kahn et al., 2018). CH arising in the context of cancer treatment and its relationship with therapy-related myeloid neoplasms is further discussed in the introduction to chapter 5.

3.4.3 Immune-mediated selection

CH is particularly common in the context of bone marrow failure syndromes (Mehta et al., 2010; Reina-Castillon et al., 2017; Stanley et al., 2017; Yoshizato et al., 2015), corroborating the notion that HSC functional decline and depletion promotes cell competition. CH arising in the context of autoimmune-mediated acquired aplastic anaemia (AA) is another example of environmental context influencing HSC somatic evolution (McKerrell and Vassiliou, 2015; Yoshizato et al., 2015). CH is present in the majority of AA patients, and the mutational spectrum reflects the selective pressure exerted by immune attack on HSCs (Stanley et al., 2017; Yoshizato et al., 2015). For example, mutations in *PIGA* are highly recurrent and result in reduced cell surface expression of glycophosphotidylinositol-anchored autoantigens (McKerrell and Vassiliou, 2015; Yoshizato et al., 2015). Deletion of chromosome 6p, which encompasses human leucocyte antigen alleles, is likely to further aid immune escape (Stanley et al., 2017).

4. Sequencing strategies for studying somatic evolution

High resolution insight into somatic evolution in normal ageing tissues requires detection of rare mutations and represents a considerable technical challenge. The Illumina sequencing platform currently has the lowest error rate, though this varies considerably across different genomic regions according to the GC content and other base composition features (Hoang et al., 2016; Ross et al., 2013). With sophisticated post-sequencing analysis techniques, mutations in less error-prone genomic regions can be detected with a sensitivity >0.1%, though this is still inadequate for detecting rare mutations in cells that have not undergone appreciable clonal expansion (Gerstung et al., 2014; Hoang et al., 2016; Martincorena et al., 2015; Ross et al., 2013).

Strategies for overcoming this challenge include growing single-cell derived colonies (Lee-Six et al., 2018) or organoids (Blokzijl et al., 2016; Roerink et al., 2018), laser capture microdissection of clonal units from tissue sections (Moore et al., 2018), single cell sequencing (Navin et al., 2011; Potter et al., 2013; Zong et al., 2012) and error-corrected sequencing using molecular barcodes (Kennedy et al., 2014; Kinde et al., 2011; Mattox et al., 2017). The latter method involves using barcoded adaptors to label both strands from a single DNA molecule. This manoeuvre greatly helps distinguish artefacts (which will almost always be called on one strand only) from real mutations (apparent in both strands from the same DNA molecule)(Kennedy et al., 2014; Schmitt et al., 2012). However, error-corrected sequencing is tractable only for very limited target regions and can be insensitive, in part due to inefficient pull-down of target regions (Kennedy et al., 2014; Schmitt et al., 2012). It is also more labour-intensive and expensive due to the need to sequence each individual molecule sufficiently deeply to generate consensus sequences (Kennedy and Ebert, 2017; Kennedy et al., 2014). A main emphasis of the work in this thesis is to better define pathophysiologically significant clonal haematopoiesis, ideally using clinically tractable sampling and sequencing approaches that might eventually be applied in a ‘real world’ setting. The experiments described here have primarily used bulk peripheral blood and bone marrow samples. For a subset of this work (Chapter 3), we compared the performance of consensus sequencing with molecular barcodes and ultradeep targeted sequencing, which is now routinely available in clinical diagnostic laboratories.

5. Thesis Aims

In summary, cancer is a clonal genetic disease adept at evolving resistance to both conventional and targeted therapies (Stratton, 2011; Stratton et al., 2009; Yates and Campbell, 2012). Knowledge of the genetic basis of cancers has galvanised research into early detection using increasingly sensitive sequencing technologies (Cohen et al., 2018; Etzioni et al., 2003; Newman et al., 2016). It is conceivable that earlier detection of asymptomatic, genetically simpler pre-cancerous lesions might enable therapeutic intervention, including targeted therapies for single oncogene addictions, analogous to treatment of chronic phase CML (O'Brien et al., 2003) or therapies to mitigate selection pressures that favour clonal expansion. The success of early cancer detection efforts will hinge upon the ability to distinguish pre-cancer from ubiquitous benign clonal expansions in normal ageing tissues. In the blood system, CH harbouring canonical leukaemia-associated mutations is a risk factor for haematological malignancy (Bowman et al., 2018). However, only a small minority of affected individuals progress, and determinants of evolutionary trajectories remain poorly understood (Figure 1.1). This dissertation investigates the pre-malignant landscape of several common haematological neoplasms and the feasibility of identifying individuals with CH at high risk of developing a blood cancer. The main aims of this project are as follows.

1. Describe the premalignant mutational landscape of the commonest haematological neoplasms and compare this with age-related CH in the general population.
2. Investigate the extent to which benign clonal haematopoiesis can be distinguished from clones at high risk of malignant transformation.
3. Investigate the prevalence of CH in childhood cancer survivors and the natural history of childhood therapy-related myeloid neoplasms.

Figure 1.1

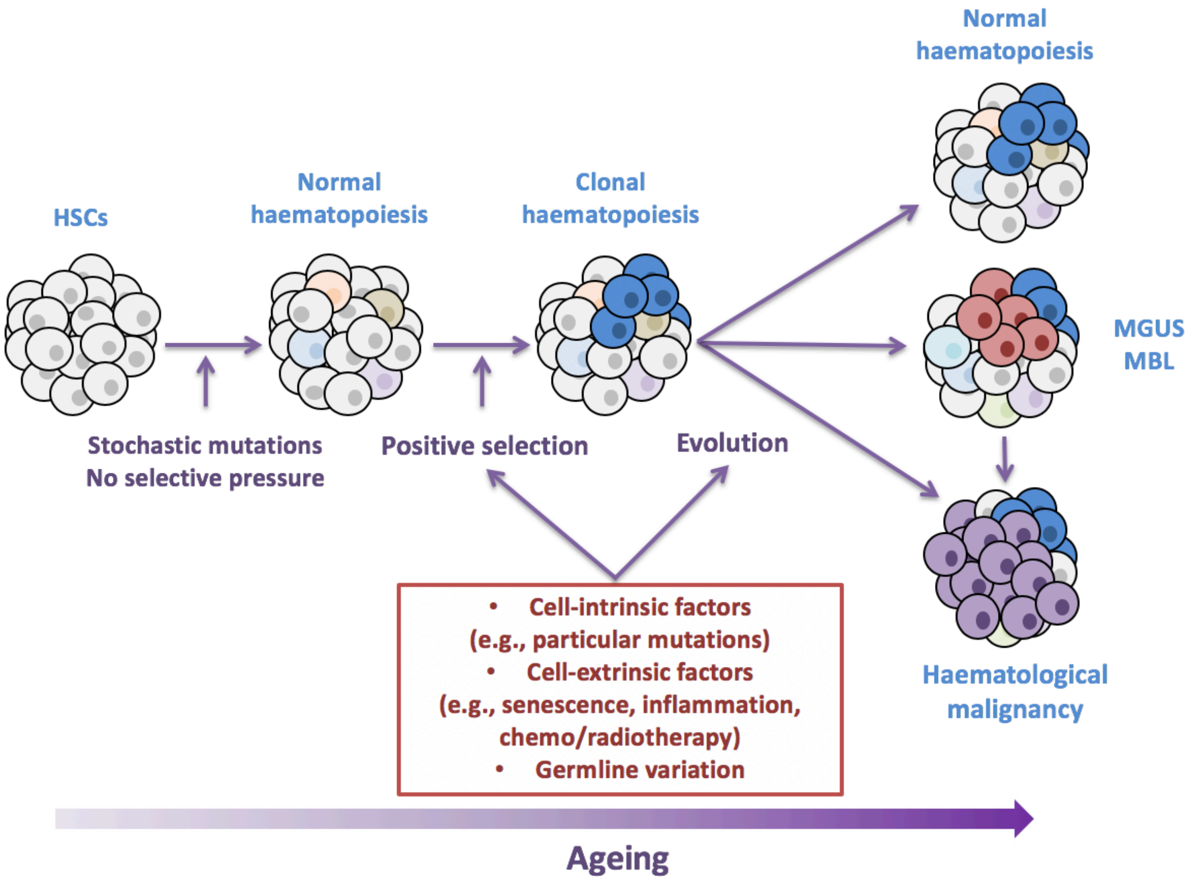


Figure 1.1 | Initiation and evolution of clonal haematopoiesis
Shown is a model illustrating the process of somatic mutation accumulation in HSCs and different clonal trajectories, with known and hypothetical influences on mutation acquisition and/or positive selection highlighted in red. As yet poorly-defined mutational processes acting on HSCs generate somatic genetic diversity in the HSC pool with time, represented here as a mosaic of distinctly coloured cells. Cells with a relative fitness advantage under the selective pressures prevailing in the haematopoietic microenvironment undergo clonal expansion. Clonal haematopoiesis is a nearly inevitable consequence of ageing, and may play a role in maintaining adequate haematopoiesis in a senescing haemopoietic niche. A minority of individuals may progress to a neoplastic disorder. MGUS, monoclonal gammopathy of unknown significance; MBL, monoclonal B-cell lymphocytosis.

Chapter 2

Materials and Methods

1. Patient samples

1.1 Pre-AML and control peripheral blood samples (Chapter 3)

For the study of the pre-clinical evolution of AML described in Chapter 3, samples from pre-AML cases and age- and sex-matched controls were collected by collaborators at the European Prospective Investigation into Cancer and nutrition (EPIC) study (Riboli et al., 2002). Samples were divided into discovery and validation cohorts and sequenced at the Wellcome Sanger Institute and the University of Toronto, respectively (see section Methods section 2.1 and 2.2).

Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, the Weizmann institute of science Ethics board approval #60-1 and East of England - Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). *De novo* AML patients were identified based on the following ICD9 codes: 9861/3 9860/3 9801/3 9866/3 9891/3 9867/3 9874/3 9840/3 9872/3 9895/3 9873/3. All patients provided peripheral blood samples from which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen prior to DNA extraction.

1.1.1 Discovery cohort samples

A total of 509 DNA samples were collected from individuals upon enrolment into the EPIC study between 1993 and 1998 across 17 different centres (Riboli et al., 2002). The pre-AML group comprised 95 individuals who developed *de novo* AML an average of 6.37 years

(IQR=4.88 years) after the sample was collected. The control group included 414 age and gender matched individuals with no record of haematological disorders (mean follow-up period 11.9 years). The median age at recruitment was 56.75 years (range 36.08 to 74.42). In order to minimize any possible demographic biases, an approximate 1:4.5 pre-AML to control ratio was maintained across the different centres. Discovery cohort (DC) sample metadata is detailed in Appendix 1.

1.1.2 Validation cohort samples

Samples were ascertained from individuals enrolled in the EPIC-Norfolk longitudinal cohort study between 1994 and 2010 (Day et al., 1999). Samples and clinical metadata were available from 37 AML patients (of which 8 were already included in the discovery cohort) and 262 age- and gender-matched controls without a history of cancer or any haematological condition. The median time between the first blood sampling and AML diagnosis was 12.3 years (IQR 8.3 years). The median follow-up period for the control cohort was 17.5 years (IQR 3.8). For 12 individuals in the pre-AML cohort, 2-3 blood specimens were available, taken a median of 3.4 years apart. Of the 262 controls, 141 had multiple blood samples available, spanning a median of 10.5 years. Blood counts and other clinical parameters were available for all study participants (Appendix 2).

1.2 Childhood cancer survivor cohort samples (Chapter 5)

We obtained peripheral blood DNA samples from patients enrolled on long-term follow-up after treatment for a paediatric malignancy and from 3 age-matched controls with no cancer history. Written informed consent was obtained for sample collection and DNA sequencing from all patients or their guardian in accordance with the Declaration of Helsinki and protocols approved by the relevant institutional ethics committees (approval numbers 09REG2015, 1-09/12/2015). The median age at cancer diagnosis was 4.5 years, and the commonest malignancies were acute lymphoblastic leukaemia (n=21), neuroblastoma (n=17) and non-Hodgkin lymphoma (n=10). Nineteen patients had received a hematopoietic stem cell transplant (8 allogeneic and 11 autologous). The median interval between completion of cancer treatment and blood sampling was 6 years (range 2 – 25). Patient characteristics are summarized in Table 4.1 and individual sample details are shown in Appendix 3.

1.3 Paediatric therapy-related myeloid neoplasm samples (Chapter 5)

We obtained bilateral bone marrow biopsies and serial peripheral blood DNA samples from a paediatric neuroblastoma patient who developed a therapy-related myeloid neoplasm. Written informed consent was obtained for sample collection and DNA sequencing from the guardian in accordance with the Declaration of Helsinki and protocols approved by the relevant institutional ethics committees (REC reference 16/EE/0394).

1.4 Pre-lymphoid neoplasm cohort and controls

For the study of the pre-clinical evolution of lymphoid neoplasms (LN) described in Chapter 4, samples from pre-LN cases and age- and sex-matched controls were collected by collaborators at the European Prospective Investigation into Cancer and nutrition (EPIC)-Norfolk study (Day et al., 1999; Riboli et al., 2002).

Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, East of England - Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). Pre-LN cases were identified based on the following ICD10 codes: C81*, C82*, C83*, C84*, C85*, C86*, C87*, C88*, C89*, C90*, C91*. All patients provided peripheral blood samples from which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen prior to DNA extraction.

2. Library preparation and sequencing

2.1 Targeted sequencing of discovery cohort pre-AML and control samples (Chapter 3)

Library preparation and sequencing of discovery cohort samples was performed by Sagi Abelson and colleagues (Princess Margaret Cancer Centre, University Health Network, Toronto). Targeted deep sequencing was performed using error-corrected sequencing (ECS) as detailed below.

Ligation of sequencing adaptors. Shearing of genomic DNA, preparation of pre-capture sequencing libraries, hybridization-based enrichment, assessment of the libraries quality and enrichment following hybridization were performed as previously described (Newman et al., 2014). Briefly, 100ng of genomic DNA was sheared before library construction (KAPA Hyper Prep Kit #KK8504, Kapa Biosystems) with a Covaris E220 instrument using the recommended settings for 250bp fragments. Following end repair and A-tailing, adapter ligation was performed using 100-fold molar excess of Molecular Index Adapter. Library clean-up was performed with Agencourt AMPure XP beads (Beckman-Coulter) and the ligated fragments were then amplified for 8 cycles using 0.5µM Illumina universal and indexing primers.

Target capture. Targeted capture was carried out on pools containing 3 indexed libraries. Each pool of adaptor-ligated DNA was combined with 5µl of 1mg/ml Cot-I DNA (Invitrogen), and 1 nmol each of xGEN Universal Blocking Oligo – TS-p5, and xGen Universal Blocking Oligo – TS-p7 (8nt). The mixture was dried using a SpeedVac and then re-suspended in 1.1µl water, 8.5µl NimbleGen 2× hybridization buffer and 3.4µl NimbleGen hybridization component A. The mixture was heat denatured at 95°C for 10 minutes before addition of 4µL of xGen Lockdown Probes (xGen® AML Cancer Panel v1.0, 3pmol). The panel was designed to include all genes recurrently mutated in the 2013 TCGA study of AML (TCGA et al., 2013). Each pool was then hybridized at 47°C for 72 hr. Washing and recovery of the captured DNA was performed according to the manufacturer's specifications. Briefly, 100µl of clean streptavidin beads was added to each capture. Following separation and removal of supernatant on a magnet, 200µL 1X Stringent Wash Buffer was added and the reaction was incubated at 65°C for 5 min. Supernatant containing unbound DNA was removed before repeating the high stringency wash one additional time. Next, the bound DNA was washed as follows: 1) 200µl 1X Wash Buffer I and separation of the supernatants by magnetic separation, 2) 200µl 1X Wash Buffer II following magnetic separation, 3) 200µl 1X Wash Buffer III and removal of the supernatants using magnetic separation. The captured DNA on beads was resuspended in 40µl of Nuclease-Free water before dividing the total volume into 2 PCR tubes and subjecting the libraries to 10 cycles of post-capture amplification (manufacturer recommended conditions; Kapa Biosystems). Prior to sequencing, libraries were spiked in with 2% PhiX.

2.2 Targeted sequencing of validation cohort pre-AML and control samples and AML diagnostic specimens (Chapter 3)

This section describes the sequencing methods for the validation cohort (VC) pre-AML and control samples discussed in Chapter 3.

Targeted sequencing was performed using a custom cRNA bait set (SureSelect, Agilent, UK, ELID #0537771, contributed by Dr Elli Papaemmanuil and Dr Peter Campbell) designed complementary to all coding exons of 111 genes implicated in myeloid leukaemogenesis (Appendix 4). Genomic DNA was sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75 base paired-end sequencing as per Illumina and Agilent SureSelect protocols.

2.3 Multiplex PCR design and sequencing (Chapter 5)

This section describes the sequencing strategy used to screen peripheral blood samples from childhood cancer survivors for clonal haematopoiesis (Chapter 5). The multiplex PCR panel was designed by Dr Naomi Park and Dr George Vassiliou as detailed in a published protocol (Park and Vassiliou, 2017) and I performed PCR experiments with supervision from Dr Park. Primers were designed using mprimer software (Shen et al., 2010) and checked for specificity using MFEprimer (Qu and Zhang, 2015). In order to minimise primer dimer formation, primers were synthesised to include a single 2'-O-Methyl base substitution, one base from the 3'-end. The multiplex PCR amplifies 32 regions of 14 genes frequently mutated in CH or t-MN (Table 4.2) (Bowman et al., 2018; McNerney et al., 2017). This is an extension of a previously validated assay (McKerrell et al., 2015) to include all coding exons of *TP53* and *PPM1D*, genes implicated in t-MN pathogenesis (Gibson et al., 2017; Hsu et al., 2018; McNerney et al., 2017). Primer sequences are detailed in Appendix 5. Amplicon libraries were sequenced on the Illumina MiSeq platform as detailed in Park et al. (Park and Vassiliou, 2017).

2.4 Targeted sequencing using a custom pan-haematological cancer panel

This section describes the sequencing methods for the diagnostic AML bone marrow samples discussed in Chapter 3, the pre-lymphoid cancer specimens and controls discussed in Chapter 4 and the paediatric therapy-related myeloid neoplasm described in Chapter 5. Targeted sequencing was performed using a custom cRNA bait set (SureSelect, Agilent, UK, ELID ID: 3129591) designed complementary to all coding exons of 95 genes recurrently mutated in myeloid and lymphoid haematological cancers, including the genes most frequently implicated in paediatric MPN/MDS (Appendix 6). Genes implicated in lymphoid neoplasms were selected with input from Dr Philip Beer. Genomic DNA was sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75 base paired-end sequencing as per Illumina and Agilent SureSelect protocols.

2.5 Whole genome sequencing

Whole genome sequencing of peripheral blood DNA (Chapter 5) was performed by 150-bp- paired-end sequencing on the Illumina Hiseq X10 platform. The Illumina no-PCR protocol was followed to construct short insert libraries, prepare flow cells and generate clusters (Kozarewa et al., 2009).

3. Variant calling

3.1 Variant calling in pre-AML and control samples

Variant filtering and annotation for the discovery cohort (section 3.1.1) and validation cohort (section 3.1.2) was performed by Dr Sagi Abelson and myself, respectively. After filtering and annotation, both datasets were combined and driver mutation calling and additional artefact filtering was performed by me as detailed in sections 3.1.3 and 3.1.4.

Figure 2.1

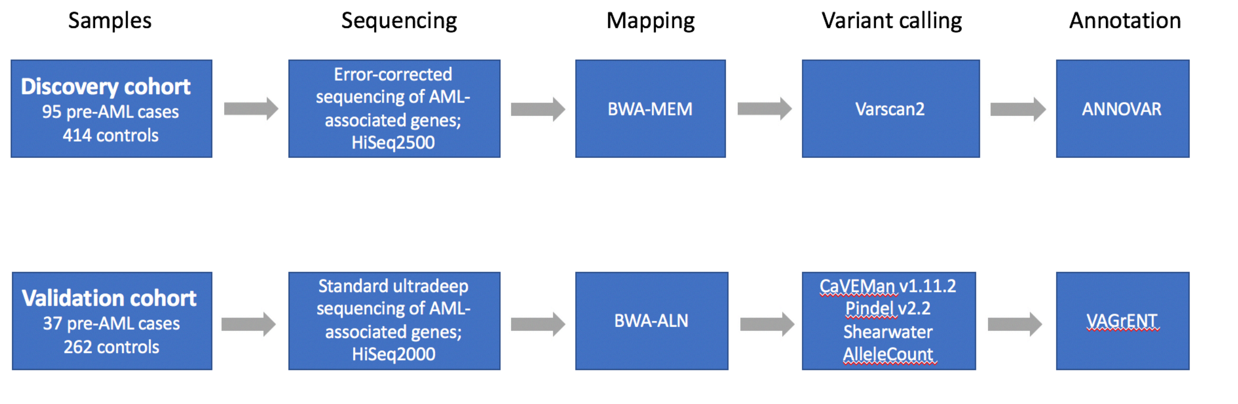


Figure 2.1 | Overview of Chapter 3 experimental design. Discovery and validation cohort pre-AML and control samples were processed using different sequencing and bioinformatic pipelines, summarised above.

3.1.1 Discover cohort variant calling and error correction

126bp paired-end read sequencing data from the Illumina HiSeq2500 platform was converted to fastq format. The 2bp molecular barcode information of each read was trimmed and incorporated into the read name. The thymine nucleotide required for ligation was removed from the sequences. The processed FASTQ files were then aligned to the hg19 reference genome using the Burroughs-Wheeler Aligner (BWA-MEM) (Li and Durbin, 2010). Indel-re-alignment was performed using GATK (McKenna et al., 2010). An in-house algorithm was written to collapse read families that share the same molecular barcode sequence, the left most genomic position of where each read of the pair maps to the reference and the CIGAR string. Families comprised of at least 2 reads were used to generate consensus reads (CR) and a consensus base was called when there was at least 70% agreement. When a consensus base was called, it was assigned with the maximum base quality score observed in its corresponding pre-collapsed reads. Furthermore, when possible, duplex reads (DR) were generated from two CR, from a singleton read (SR) and a CR, or from two SR (Kennedy et al., 2014). For each sequenced sample, we generated two BAM files, called bam1 and bam2. Bam1 consists of DR, CR and singleton reads, thereby including some error corrected and non-error corrected reads. Bam2 consists of DR and CR but not singleton reads. Both files were then analysed to detect single nucleotide variants (SNVs) and small insertions and deletions (indels) using Varscan2 (Koboldt et al., 2012). In order to further remove sequencing artefacts and improve sensitivity, we applied a two-step statistical polishing approach that models the error rate at each sequenced genomic position. For both steps, bam1 was used and all the samples except the sample being investigated were included for error rate modelling. At step one, as previously described (Newman et al., 2014), the error rates were modelled by fitting weibull distribution curves to the non-reference allele fractions. SNVs with allele fractions that were statistically distinguishable from the background error rates were further analysed. At Step 2, the coverage of the non-reference allele fractions was considered by using linear line fitting that describes the negative correlation that exist between the log (non-reference allele fraction) and the corresponding log(coverage) values. This allowed us to estimate different error rates at different coverage depths. Indel errors were filtered using barcode mediated error correction alone. At least 10 CR, 5 supporting reads on the forward strand, 5 supporting reads on the reverse strand, and 2 DR were required to call an indel. Variants were annotated using Annovar (Yang and Wang, 2015). Additional post-processing steps applied to data from both the discovery and validation cohorts are detailed in section 3.1.3.

3.1.2 Validation cohort variant calling

Sequencing reads were aligned to the reference genome (GRCh37d5) using the Burrows-Wheeler aligner (BWA-ALN)(Li and Durbin, 2009). Unmapped reads, PCR duplicates and reads mapping to regions outside the target regions (merged exonic regions + 10bp either side of each exon) were excluded from analysis. Sequencing depth at each base was assessed using Bedtools coverage v2.24.0 (Quinlan and Hall, 2010).

Substitutions

Somatic single nucleotide variants (SNVs) were called using Shearwater, an algorithm developed for detecting subclonal mutations in deep sequencing experiments (https://github.com/gerstung-lab/deepSNV v1.21.5) (Gerstung et al., 2012; Gerstung et al., 2014; Martincorena et al., 2015) considering only reads with minimum nucleotide and mapping quality of 25 and 40, respectively. This algorithm models the error rate at individual loci using information from multiple unrelated samples. Additionally, allele counts at the recurrent AML mutation hotspots listed in section 3.1.4 were generated using an in-house script (https://github.com/cancerit/alleleCount) and manually inspected in the Jbrowse genome browser (Buels et al., 2016). To further complement our SNV calling approach, we applied an extensively validated in-house version of CaVEMan v1.11.2 (Cancer Variants through Expectation Maximization)(Stephens et al., 2012). CaVEMan compares sequencing reads between study and nominated normal samples and uses a naïve Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan). Post-processing filters required that the following criteria were met for CaVEMan to call a somatic substitution:

1) If coverage of the mutant allele was less than 8, at least one mutant allele was detected in the first 2/3 of the read.
2) Less than 3% of the mutant alleles with base quality ≥ 15 were found in the nominated normal sample.
3) Mean mapping quality of the mutant allele reads was ≥ 21.

4) Mutation does not fall in a simple repeat or centromeric region.

5) Fewer than 10% of the reads covering the position contained an indel according to mapping.

6) Less than 80% of the reads report the mutant allele at the same read position.

7) At least a third of the reads calling the variant had a base quality of 25 or higher.

8) Not all mutant alleles reported in the second half of the read.

9) Position does not fall within a germline insertion or deletion.

The following additional post-processing criteria were applied to all SNV calls:

1) Minimum VAF 0.5% with a minimum of 5 bidirectional reads reporting the mutant allele (with at least 2 reads in forward and reverse directions).

2) No indel called within a read length (75bp) of the putative substitution.

Small insertions and deletions

Small insertions and deletions were sought using two complementary approaches. Firstly, an in-house version of Pindel v2.2 (Raine et al., 2015) (https://github.com/cancerit/cgpPindel) was applied. We additionally used the aforementioned Shearwater algorithm (Gerstung et al., 2012; Gerstung et al., 2014; Martincorena et al., 2015) in order to increase sensitivity for indels present at low VAF. VAF correction was performed using an in-house script (https://github.com/cancerit/vafCorrect). Post-processing filters required that the following criteria were met for a variant to be called:

1) Minimum of 5 reads supporting the variant with minimum of 2 reads in each direction. For Pindel, the total read count was based on the union of BWA and Pindel reads reporting the mutant allele.

2) Minimum VAF 0.5%

3) Variant not present within an unmatched normal panel of approximately 400 samples.

4) No reads supporting the variant identified in the nominated normal sample.

Mutations were annotated according to ENSEMBL version 58 using VAGrENT (Menzies et al., 2002) for transcript and protein effects (https://github.com/cancerit/VAGrENT) and Annovar (Yang and Wang, 2015) for additional functional annotation.

3.1.3 Additional post-processing filters applied to all data

The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event:

1) Any mutant allele reported within 75bp of another variant.
2) Any mutant allele with a population allele frequency > 1 in 1000 according to any of five large polymorphism databases: ExAC, 1000 Genomes Project, ESP6500, CG46, Kaviar that is not a canonical hotspot driver mutation with COSMIC recurrence > 100.
3) Mutations that were present in > 10% of the control cohort but not recurrent in COSMIC were flagged as potential germline variants or sequencing artefact.
4) As artefactual indels tend to be recurrent, any indels occurring in >2 samples were flagged for additional inspection.

3.1.4 Curation of oncogenic variants

Putative oncogenic variants were identified according to evidence for functional relevance in AML as previously described and used to define CH-PD (Gerstung et al., 2017; Papaemmanuil et al., 2016).

Variants were annotated as likely driver events if they fulfilled any of the following criteria:

1) Truncating mutations (nonsense, essential splice site or frameshift indel) in the following genes implicated in AML pathogenesis by loss-of-function: *NF1, DNMT3A, TET2, IKZF1, RAD21, WT1, KMT2D, SH2B3, TP53, CEBPA, ASXL1, RUNX1, BCOR, KDM6A, STAG2, PHF6, KMT2C*.
2) Truncating variants in *CALR* exon 9.
3) *JAK2* V617F
4) *FLT3* ITD

5) Non-synonymous variants at the following hotspot residues:

   a. *CBL* E366, L380, C384, C404, R420, C396
   b. *DNMT3A* R882
   c. *FLT3* D835
   d. *IDH1* R132
   e. *IDH2* R172, R140
   f. *KIT* W557, V559, D816
   g. *KRAS* A146, Q61, G13, G12
   h. *MPL* W515
   i. *NRAS* Q61, G12, G13
   j. *SF3B1* K700, K666
   k. *SRSF2* P95
   l. *U2AF1* Q157, R156, S34

6) Non-synonymous variants reported at least 10 times in COSMIC with VAF < 42% and population allele frequency < 0.003.

7) Non-synonymous variants clustering within a functionally validated domain or within 4 amino acids of a hotspot variant with population allele frequency < 0.003 and VAF < 42%.

8) Non-synonymous variants reported in COSMIC > 100 times with population allele frequency < 0.003 regardless of VAF.

This driver curation strategy inevitably runs a small risk of including germline variants in familial AML genes, e.g., *RUNX1*. However, in most settings, where a matched constitutional DNA sample is likely to be unavailable, this seems the best approach.

Of note, the entire validation cohort included 37 pre-AMLs, 8 of these were also included in the original discovery cohort and therefore were excluded from the validation cohort for downstream analysis. Both the discovery and the validation cohorts sourced samples from different centres participating in the EPIC study, hence the overlap. However, discovery and validation cohorts were sequenced by two independent research groups using different methods, as described above. Putative driver mutations detected for the duplicated samples by the two different methods were highly similar. All 9 driver mutations detected in the discovery cohort with VAF>0.015 were detected in the validation cohort samples, while 8 other mutations (7 in TET2 or DNMT3A) with lower VAFs escaped validation. The latter is probably due to the higher VAF cut-off applied to the validation cohort sequencing method and the stochastic failure to sample a small clone in two independent experiments.

3.2 Variant calling from multiplex PCR sequencing

Reads were aligned to human genome build GRCh37d5 using the Burrows-Wheeler Aligner (Li and Durbin, 2010) and analysed for somatic single nucleotide variants and indels. Allele counts across target hotspots were generated using an in-house script (https://github.com/cancerit/alleleCount), considering only loci with ≥1000 reads and minimum base and mapping quality of 25 and 35, respectively. In order to identify SNV and indels in *TP53* and *PPM1D*, 3 variant callers were applied: Shearwater (https://github.com/gerstung-lab/deepSNV v1.21.5)(Gerstung et al., 2012; Gerstung et al., 2014; Martincorena et al., 2015), cgpPindel v2.2 (Raine et al., 2015) and CaVEMan v1.11.2 (Cancer Variants through Expectation Maximization, https://github.com/cancerit/CaVEMan)(Stephens et al., 2012) as describe in section 3.1.2 above.

3.3 Variant calling for non-AML pre-malignant samples and controls

SNV and indel calling was performed as described in 3.1.2 and 3.1.3. The strategy for curating putative driver variants was adjusted to account for the greater number of genes included in the larger bait panel (Appendix 6). Specifically, variants were flagged as candidate driver events if they fulfilled any of the following criteria:

1) Nonsense or frameshift mutations in the following genes: *ARID1A, ASXL1, ATM, B2M, BCOR, BCORL1, CALR, CDKN2A, CEBPA, CREBBP, CSF1R, CSF3R, CUX1, DNMT3A, EP300, FBXW7, KDM6A, KMT2C, KMT2D, NF1, NOTCH2, NPM1, PAX5, PHF6, POT1, PPM1D, PRDM1, PTEN, RAD21, SETD2, SOCS1, STAG2, TET2, TNFAIP3, TNFRSF14, TP53, WT1, ZRSR2*
2) Splice site mutations in the following genes: *ARID1A, ATM, BCOR, CBL, CD79B, CDKN2A, CUX1, DNMT3A, KDM6A, NF1, PAX5, PHF6, PRDM1, PTEN, SETD2, STAG2, WT1, ZRSR2*

3) Missense mutations in the following genes were considered if they passed SNP and artefact filters and had support as candidate drivers based on relevant literature (Tate et al., 2019): *ARID1A, ASXL1, ATM, B2M, BCL6, BCORL1, BRAF, CALR, CARD11, CBL, CD79B, CDKN2A, CEBPA, CREBBP, CSF1R, CSF3R, CUX1, DNMT3A, EP300, ETNK1, EZH2, FBXW7, FLT3, GATA2, GNAS, H3F3A, IDH1, IDH2, IL7R, JAK2, KIT, KMT2D, KRAS, MPL, MYD88, NF1, NOTCH1, NOTCH2, NRAS, PAX5, PDGFRA, PHF6, PIM1, POT1, PPM1D* (exon 6), *PRDM1, PTEN, PTPN11, RAD21, SETBP1, SETD2, SF3B1, SRSF2, STAG2, STAT3, TET2, TNFRSF14, TP53, U2AF1, WT1, XPO1, ZEB1, ZRSR2*

4) Non-synonymous variants reported at least 10 times in COSMIC with VAF < 35% and population allele frequency < 0.003.

5) Non-synonymous variants clustering within a functionally domain or within 4 amino acids of a hotspot variant with population allele frequency < 0.003 and VAF < 35%.

6) Non-synonymous variants reported in COSMIC > 150 times with population allele frequency < 0.003 regardless of VAF.

3.4 Screening for pathogenic germline variants

All mutations flagged by SNP filters (VAF > 0.42 and present in ExAC, 1000 Genomes Project, ESP6500, CG46 or Kaviar databases) were screened against the ClinVar database (Landrum et al., 2016) and Human Gene Mutation Database (HGMD) (Stenson et al., 2003) to identify potential cancer predisposition germline variants.

3.5 Variant calling from whole genome sequences (Chapter 5)

Whole genome sequences were mapped to the GRCh37d5 reference genome using the Burroughs-Wheeler Aligner (BWA-mem) (Li and Durbin, 2010). The Cancer Genome Project (Wellcome Trust Sanger Institute) variant calling pipeline was used to call somatic mutations which includes the following algorithms: CaVEMan (1.11.0)(Jones et al., 2016) for substitutions; an in-house version of Pindel (2.2.2; github.com/cancerit/cgpPindel)(Raine et al., 2015) for indels; BRASS (5.3.3; github.com/cancerit/BRASS) for rearrangements (Li et al., 2017), and ASCAT NGS (4.0.0) for copy number aberrations (Van Loo et al., 2010). In addition to filters inherent to the CaVEMan algorithm, the following post-processing filtering criteria were applied for substitutions: a minimum two reads in each direction reporting the mutant allele; at least ten fold coverage at the mutant allele locus; minimum variant allele fraction 5%; no insertion or deletion called within a read length (150bp) of the putative substitution; no soft-clipped reads reporting the mutant allele; median BWA alignment score of the reads reporting the mutant allele ≥ 140. The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event: any mutant allele reported within 150bp of another variant; any mutant allele with a population allele frequency > 1 in 1000 according to any of five large polymorphism databases: ExAC, 1000 Genomes Project, ESP6500, CG46, Kaviar.

To identify potential driver events in whole genome data, I considered variants presenting in established cancer genes (Tate et al., 2019). Tumour suppressor coding variants were considered if they were annotated as functionally deleterious by an in-house version of VAGrENT (http://cancerit.github.io/VAGrENT/) (Menzies et al., 2002), or alternatively if they were disruptive rearrangement breakpoints or homozygous deletions. Additionally, homozygous deletions were required to be focal (<1 Mb in size) or constitute a known contiguous gene syndrome implicated in t-MN (McNerney et al., 2017). Mutations in oncogenes were considered driver events if they were located at previously reported canonical hot spots (point mutations) or amplified the intact gene. Amplifications also had to be focal (<1 Mb) and increase the copy number of oncogenes to a minimum of 5 copies.

3.6 Copy number variation in targeted sequencing data

To detect copy number aberrations in the paediatric t-MN case discussed in Chapter 5, I applied FACETS (Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing), an allele-specific copy number analysis (ASCN) method (Shen and Seshan, 2016).

4. Predictive modelling

Regularised logistic and Cox proportional hazards regression approaches were tested in generating the predictive models described in Chapters 3 and 4.

Dr Moritz Gerstung wrote the initial version of the code for Chapter 3 and closely supervised all further iterations of the models described in Chapter 3. The code for the models described in Chapter 4 was written by me using a very similar analysis framework and methods as in Chapter 3.

4.1 Cox proportional hazards model with random effects

We used a Cox proportional hazards regression to model haematological malignancy-free survival as previously described (Gerstung et al., 2017). We used random effects for the Cox proportional hazards model in the CoxHD R package developed by Dr Gerstung (http://github.com/gerstung-lab/CoxHD). A key strength of this approach is the ability to include many variables in one model while shrinking estimated effects for parameters with weak support in the data, thus controlling for overfitting. We used weighting to minimise the biases introduced by the artificial case-control ratio (Antoniou et al., 2005) and calculated hazard ratios relative to the (approximate) true cumulative incidence of either AML (Chapter 3) or all lymphoid malignancies (Chapter 4) in the given age range over a follow up of 10-20 years. Full details of model derivation and comparisons with alternative methods are included in the accompanying code (Appendix 7). In brief, variables comprised age, gender, the variant allele fraction of putative driver mutations and selected clinical variables when available. We performed agnostic imputation of missing variables by mean and linear rescaling of gene variables by a power of 10 to a magnitude of 1.

All blood samples taken within 6 months of cancer diagnosis were excluded from model training. Among the pre-AML samples (Chapter 3), 4 individuals were thus removed from the discovery cohort. For one individual in the validation cohort who provided 3 pre-diagnostic samples, the 3rd sample was taken within this time frame and was also excluded (though their older samples allowed this individual to remain in the modelling analysis).

For each model, the following measures of predictive accuracy were evaluated before and after leave-one-out cross-validation (LOOCV): (i) concordance (C)(Harrell et al., 1996), (ii) time-dependent area under the receiver-operating characteristic curve (AUC)(O'Quigley et al., 2005) and (iii) Uno's estimator of cumulative/dynamic AUC (Uno et al., 2007). Coefficient confidence intervals were calculated using 100 bootstrap samples.

Concordance measures were obtained using the survConcordance() function implemented in the survival R package (Therneau and Grambsch, 2000). Dynamic AUC was calculated with AUC.uno() implemented in the survAUC package (Heagerty et al., 2000). Time-independent AUC was calculated by the performance function implemented in the ROCR package (Sing et al., 2005). The expected incidence of each haematological malignancy was calculated from the UK office of national statistics, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/. All-cause mortality data was obtained from the office of national statistics (https://www.ons.gov.uk/).

4.2 Ridge regularised logistic regression

Using the same covariates as in the Cox proportional hazard models, we fitted a ridge regularised logistic regression model to dichotomised outcome data. While logistic regression is a common choice for case-control analyses, a downside of this approach is the inability to explicitly use time-dependent covariates. The penalty parameter was chosen using LOOCV on the full cohort; this value was then used on the discovery and validation cohorts to yield the same scaling of coefficients. Confidence intervals were calculated using 100 bootstrap samples. Fitting was performed using the glmnet R package (Simon et al., 2011). AUC as the primary performance metric was calculated using the ROCR R package (Sing et al., 2005).

Chapter 3

Predicting acute myeloid leukaemia risk in the general population

1. Introduction

As discussed in Chapter 1, CH harbouring canonical leukaemia-associated mutations is a risk factor for haematological malignancy, though only a small minority of affected individuals progress (Bowman et al., 2018). Acute myeloid leukaemia (AML) is the commonest acute leukaemia in adults and typically presents suddenly as a fulminant disease with a poor prognosis (Döhner et al., 2015). This chapter describes an experiment to distinguish individuals at high risk of developing *de novo* acute myeloid leukaemia (AML) from those with indolent CH at low risk of malignant transformation. The introduction provides background on AML and reviews existing literature on its pre-clinical evolution and relationship to clonal haematopoiesis.

1.1 Acute myeloid leukaemia

1.1.1 Definition and epidemiology

AML is an aggressive haematopoietic stem cell disorder characterized by clonal proliferation of poorly differentiated myeloid cells (Döhner et al., 2015). It is the commonest acute leukaemia among adults, and comprises around 20% of all paediatric leukaemia (Döhner et al., 2015). The incidence of AML increases dramatically with age, and exceeds 100 cases per 100,000 in those over the age of 60, with a higher risk among men (CRUK, 2018;

SEER, 2018). There are around 3,100 new AML cases and 2,500 AML-related deaths each year in the UK (CRUK, 2018).

1.1.2 Aetiology and risk factors

The dominant AML risk factor is age, though the role ageing plays in the aetiology of AML is incompletely understood (Döhner et al., 2015). The somatic mutation burden seen in AML correlates with age at diagnosis and is similar to that observed in normal HSCs from age-matched individuals without a haematological disorder (Welch et al., 2012). Unlike many common adult epithelial cancers, the role of extrinsic mutational processes appears to be minor, with the age-related mutational SBS11 and SBS5 accounting for the vast majority of AML mutations (Alexandrov et al., 2018; Alexandrov et al., 2013).

Environmental or occupational chemical exposures, notably to benzene and other industrial solvents, may play a role in a minority of AML cases, though evidence for a causal link is weak (Austin et al., 1988).

Germline variants in a growing number of genes have been implicated in myeloid malignancies, including *RUNX1, GATA2, TERT, ATG2B, TP53* and *CEBPA* (Hinds et al., 2016; Saliba et al., 2015; Smith et al., 2004; Zhang et al., 2015). As discussed in the general introduction, germline and somatic mutations in the same cancer gene generally carry different biological and clinical significance and merit distinction (Arber et al., 2016; Döhner et al., 2015). Furthermore, recent evidence has suggested that the distinction between germline and somatic mutation is less clear than previously thought, with a growing catalogue of highly penetrant germline variants strongly predisposing to acquisition or clonal selection of particular somatic mutations (Hinds et al., 2016; Loh et al., 2018).

Other myeloid neoplasms, most commonly myeloproliferative neoplasms and myelodysplastic syndromes, may transform into AML, termed secondary AML (sAML) (Deininger et al., 2017; Sperling et al., 2017).

The most prevalent extrinsic risk factor for AML is previous exposure to chemotherapy or radiotherapy, in particular alkylating agents and topoisomerase II inhibitors (McNerney et al., 2017). Any AML that arises after cytotoxic treatment is termed therapy-related AML (t-AML) and is discussed further in the introduction to Chapter 5.

AML that presents suddenly with manifestations of bone marrow failure is termed *de novo* AML to distinguish it from sAML and t-AML, although, as discussed later on, these distinctions are not always straight-forward or biologically meaningful.

1.1.3 AML genetics

The genetic diversity of AML was first revealed by cytogenetic analyses in the 1970s (Rowley, 2008), and has since been well characterised by several large genomic studies (Arber et al., 2016; Gerstung et al., 2017; Papaemmanuil et al., 2016; TCGA et al., 2013). According to the classic “two-hit” model of AML leukaemogenesis proposed by Gilliland and Griffin, two types of mutations are required to produce AML: type II mutations that impair differentiation and subsequent apoptosis and are typically initiating events, and type I mutations that endow pre-leukaemic clones with a proliferative advantage (Gilliland and Griffin, 2002). Genomic studies have corroborated the main concepts of this model, providing further evidence that the block in differentiation is the initiating event for *de novo* AML. Many of the commonest mutations in AML founding clones target epigenetic regulators (Kronke et al., 2013; Shlush et al., 2014; Welch, 2014), which play central roles in haematopoietic stem cell differentiation (Abdel-Wahab et al., 2012; Challen et al., 2011; Figueroa et al., 2010a; Figueroa et al., 2010b). Furthermore, leukaemia-associated mutations in epigenetic regulators are common drivers of CH, whereas ‘type I’ mutations are very rarely observed in association with CH, consistent with this class of genetic events occurring later in leukaemogenesis after differentiation arrest has been established (Genovese et al., 2014; McKerrell et al., 2015; Xie et al., 2014).

Although this model remains conceptually useful, sequencing studies have revealed diverse genetic routes to AML, with recurrent mutations identified in over 70 genes (Papaemmanuil et al., 2016; TCGA et al., 2013). The majority of patients harbour multiple driver events, and both individual mutations and co-occurrence patterns are powerful determinants of clinical outcome (Gerstung et al., 2017; Huet et al., 2018; Papaemmanuil et al., 2016). The most recurrent structural and numerical chromosomal abnormalities include t(8;21), inv(16), t(15;17), 11q (MLL) fusions, inv(3), t(6;9), -7/7q, +8/8q, -5/5q and -17/17p (Papaemmanuil et al., 2016; TCGA et al., 2013). The majority of driver events in adult AML, however, are point mutations (single nucleotide variants and indels)(Papaemmanuil et al., 2016; TCGA et al., 2013). Frequently mutated genes include epigenetic regulators (*DNMT3A,* *TET2, IDH1, IDH2*), genes involved in the RNA splicing machinery (*SF3B1, SRSF2, U2AF1, ZRSR2*), chromatin regulators (*ASXL1, BCOR, STAG2, MLL-PTD, EZH2, PHF6*), transcription factors (*RUNX1, GAT2, CEBPA*), *NPM1*, and genes involved in RAS and/or STAT signalling (*NRAS, KRAS, PTPN11, NF1, FLT3, CBL, KIT*)(Papaemmanuil et al., 2016; TCGA et al., 2013).

1.1.4 AML classification schemes

The World Health Organisation (WHO) Classification of Haematopoietic and Lymphoid Tissues subdivides AML into four categories: AML with recurrent genetic abnormalities, AML with myelodysplasia-related changes, therapy-related AML and AML not otherwise specified (NOS)(Arber et al., 2016). The latter group is further subdivided by morphological features. The WHO classification scheme was updated in 2016 to include several new disease categories within the section of AML with recurrent genetic abnormalities (Arber et al., 2016). However, several studies suggest that WHO subgroups still do not adequately capture the molecular heterogeneity of AML, which underpins its biological and prognostic features (Gerstung et al., 2017; Metzeler et al., 2016; Papaemmanuil et al., 2016). The largest genomic study of AML to date included 1540 patients enrolled in three prospective clinical trials and identified eleven prognostically relevant molecular-genetic subgroups (Gerstung et al., 2017; Papaemmanuil et al., 2016). This study added considerable nuance to our understanding of AML biological mechanisms and genetic classification. For example, mutations affecting different loci in the same gene, e.g., *IDH2* p.R140 and *IDH2* p.R172, had divergent co-occurrence patterns and impacts on clinical outcome.

1.1.5 Treatment challenges

Despite much progress in understanding AML genetics and pathogenesis, standard AML therapy has changed very little over the past three decades (Döhner et al., 2015; Yates et al., 1973). The backbone of therapy remains the combination of two drugs developed in the 1950s, namely daunorubicin and cytarabine, compounds serendipitously derived from soil microbes and marine sponges, respectively (Schwartsmann et al., 2001; Stutzman-Engwall and Hutchinson, 1989). Improvements in patient outcomes are primarily attributable to better supportive care during periods of myelosuppression (Döhner et al., 2015). Although most patients capable of tolerating intensive chemotherapy achieve remission, the majority succumb to relapse (Döhner et al., 2015; Rubnitz et al., 2014). Overall survival rates are 35% to 40% for younger patients and 5% to 15% for patients over the age of 60 (Dohner et al., 2010; Rubnitz et al., 2014). Efforts to target recurrently mutated oncogenes, notably the tyrosine kinases FLT3 and KIT, have been met with rapid emergence of disease resistance and little improvement in overall survival (Döhner et al., 2015; Stein, 2015; Wander et al., 2014).

1.2 The relationship between CH and AML

As discussed in Chapter 1, the two largest studies of clonal haematopoiesis in the general population demonstrated an increased risk of haematological cancers in general (not specifically AML) in those with CH, which was higher in those with mutations at high VAFs (Genovese et al., 2014; Jaiswal et al., 2014). Genovese et al. identified thirty-one participants diagnosed with a hematologic cancer more than 6 months after DNA sampling, of whom thirteen (42%) had antecedent CH (Genovese et al., 2014). Of these, two developed AML and one developed "acute leukemia of unspecified origin". Of the remaining ten, three developed CLL, two MPN (both *JAK2* V617F mutated), one B-cell lymphoma, one multiple myeloma, one monoclonal gammopathy of unknown significance, one CMML and one MDS (Genovese et al., 2014). Two of the three MDS/AMLs in this paper were diagnosed within two months after DNA sampling (Genovese et al., 2014). Furthermore, Genovese et al. found that CH with putative drivers (CH-PD) afforded the same risk of haematological cancers as CH without known drivers, potentially alluding to indirect risks associated with CH (Jaiswal et al., 2014). Similarly, Jaiswal et al. reported sixteen haematological cancers during a median 95-month follow-up period, of which only five (31%) had CH detected in their pre-diagnosis sample (Jaiswal et al., 2014). Of these, two developed lymphoma, one "cancer of the spleen" (*JAK2* V617F mutated), one "myeloid leukaemia" and one "leukaemia" not otherwise specified (Jaiswal et al., 2014). Together, these two studies captured up to five possible AMLs amongst 29,652 study participants (Genovese et al., 2014; Jaiswal et al., 2014). Collectively, only a minority of blood cancers arising during follow-up were diagnosed in individuals with antecedent CH, and several of these were indolent myeloproliferative or chronic lymphoid conditions. It therefore remained unclear whether or not CH could be used to predict the subsequent development of blood cancers, let alone of *de novo* AML, with any degree of sensitivity or specificity.

2. Results

To investigate whether individuals at high risk of developing *de novo* AML can be distinguished from those with benign CH, genes recurrently mutated in AML or CH were deep-sequenced in peripheral blood cell DNA from a total of 125 individuals sampled before AML diagnosis (pre-AML group), together with 676 unselected age- and gender-matched individuals (control group). To detect somatic mutations with maximum sensitivity, deep error-corrected targeted sequencing was first applied to a discovery cohort of 95 pre-AML cases sampled on average 6.3 years before AML diagnosis and 414 age- and gender-matched controls (Appendix 1). Error-corrected sequencing was performed by Dr Sagi Abelson as detailed in Methods section 2.1. A validation cohort comprising 29 pre-AML cases and 262 controls (Appendix 2) was analysed using conventional deep sequencing with an overlapping gene panel (Methods section 2.2).

2.1 Prevalence of CH-PD in pre-AML versus controls

Taking both cohorts together, CH, defined by the presence of mutations in putative driver genes (CH-PD), was found in 73.4% of the pre-AML cases at a median of 7.6 years before diagnosis (Appendices 8 and 9). By contrast, CH-PD was observed in 36.7% of controls ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Figure 3.1a). This CH-PD prevalence in the controls is consistent with data from a study of more than 2,000 healthy individuals assayed using a similarly sensitive error-corrected sequencing method (Acuna-Hidalgo et al., 2017). Additionally, 39% of pre-AML cases over age 50 had a driver mutation with a VAF exceeding 10%, compared to only 4% of controls, a prevalence that is in line with the largest studies of CH-PD in the general population (Genovese et al., 2014) ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Figure 3.1b). The median number of driver mutations per individual increased with age and was significantly higher in the pre-AML group relative to controls ($P < 2.2 \times 10^{-16}$, two-sided Wilcoxon rank-sum test; Figure 3.1c). Furthermore, examination of VAF distribution revealed significantly larger clones among the pre-AML cases ($P = 1.2 \times 10^{-13}$, two-sided Wilcoxon rank-sum test; Figure 3.1d).

Figure 3.1

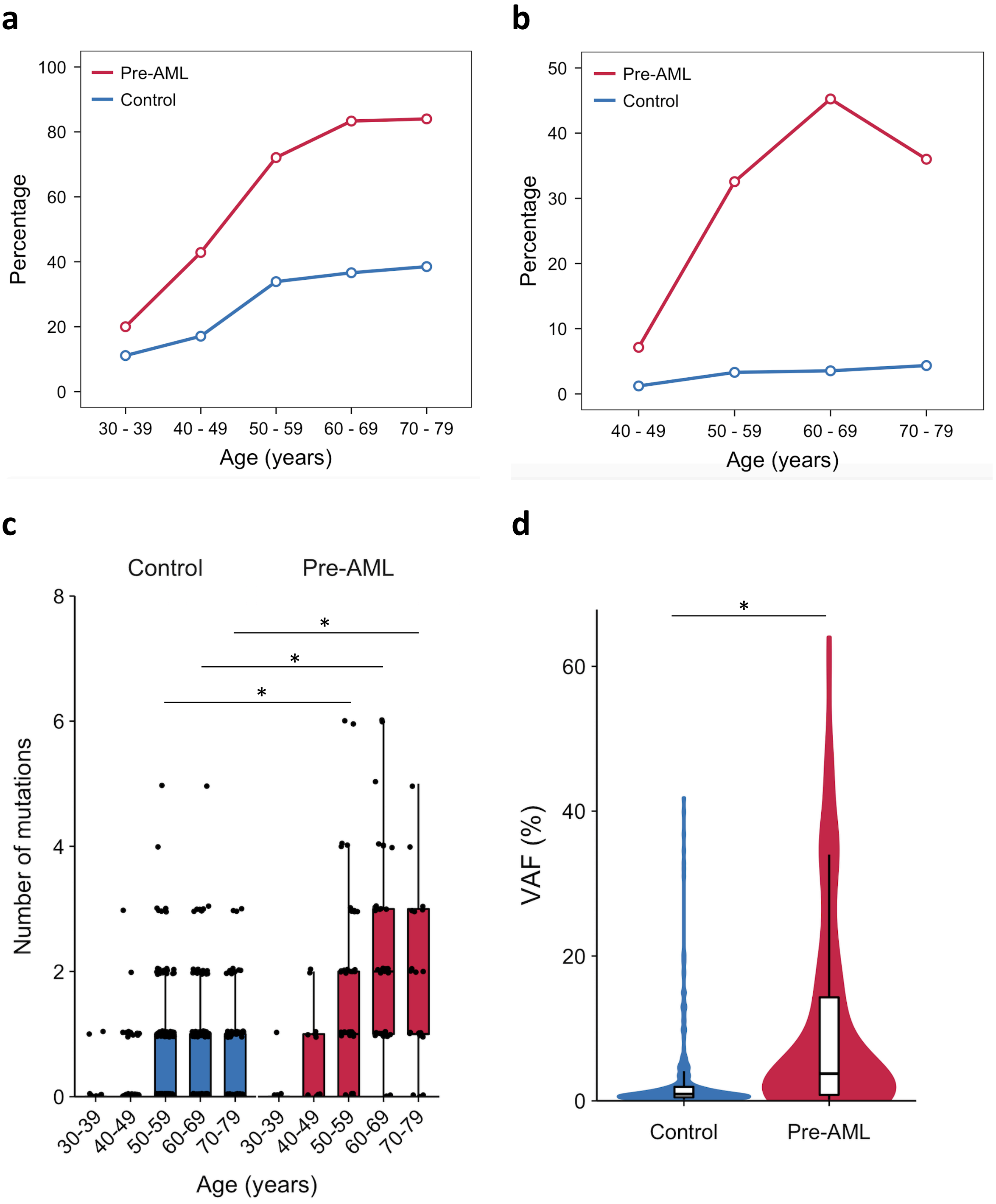


Figure 3.1 | Prevalence of CH-PD, number of mutations and clone size in pre-AML and control cohorts. a, Prevalence of CH-PD among pre-AML cases (red) and controls (blue). **b,** Prevalence of CH-PD clones with VAF > 10% among pre-AML cases (red) and controls (blue). **c**, The number of CH-PD mutations detected in cases and controls according to age. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. **d**, VAF of CH-PD mutations. All panels show data for n = 800 biologically independent samples. *P < 0.0005, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction.

2.2 Clonal dynamics over time and evolution to AML

In order to explore the mechanisms underpinning the higher mutation burden in pre-AMLs and the relationship between CH-PD and future leukaemia, I sequenced serially collected samples available for a subset of the VC (12 pre-AMLs and 141 controls) as well as three FFPE-fixed bone marrow biopsy samples available from AML diagnosis (PD29962, PD30054, PD30089). Comparison of the pre-AML mutations to the mutations detected in the diagnostic specimen demonstrated that most, though not all, drivers persisted and of these only a subset expanded to become clonal in the future AML (Figure 3.2a-c). The sensitivity of sequencing for the AML diagnostic samples was limited by the low quality of the FFPE-derived DNA and variable sequencing coverage. For PD29962, no putative drivers with VAF exceeding 9% were detected at diagnosis. In this individual, a clone harbouring a *TET2* p.E852* variant persisted for over 14 years, but decreased in size. A *KRAS* p.G12D variant also detected pre-diagnosis became undetectable, though with only 79 reads covering this locus in the diagnosis DNA, it is possible that it persisted at a subclonal level. Both PD30054 and PD30089 show evidence of persistent clones that became clonal in the AML, as well as new drivers present at diagnosis. PD30089 also developed a *JAK2* p.V617F-mutated clone, which persisted but decreased in size. For an additional case (PD29918), a third blood sample was taken very close to AML diagnosis (~1 month prior), demonstrating an *SRSF2* p.P95R mutation detected at all three time points (Figure 3.2d), which almost certainly contributed to the AML, while the second mutation detected (*TET2* p.S354*) persisted at declining VAF. Furthermore, data from individuals for whom blood sampling was done less than a year before AML diagnosis (n=9) show that the majority of these cases have driver mutations at high VAF (Figure 3.2e-f, Appendix 9), again suggesting that the pre-AML clones detected are likely to include those that later evolved into AML in most cases. Collectively these findings suggest that the driver mutations identified in pre-AML cases may represent a combination of pre-leukaemic clones as well as additional ‘bystander’ clones which do not transform. Several studies suggest that such independent clones may be common in AML patients at diagnosis (Parkin et al., 2017; Wong et al., 2015a). For example, a recent study of patients undergoing induction therapy found that five out of fifteen had marked expansion of clones unrelated to the founding AML clone but detectable in diagnostic specimens using error-corrected sequencing (Wong et al., 2015a).

Figure 3.2

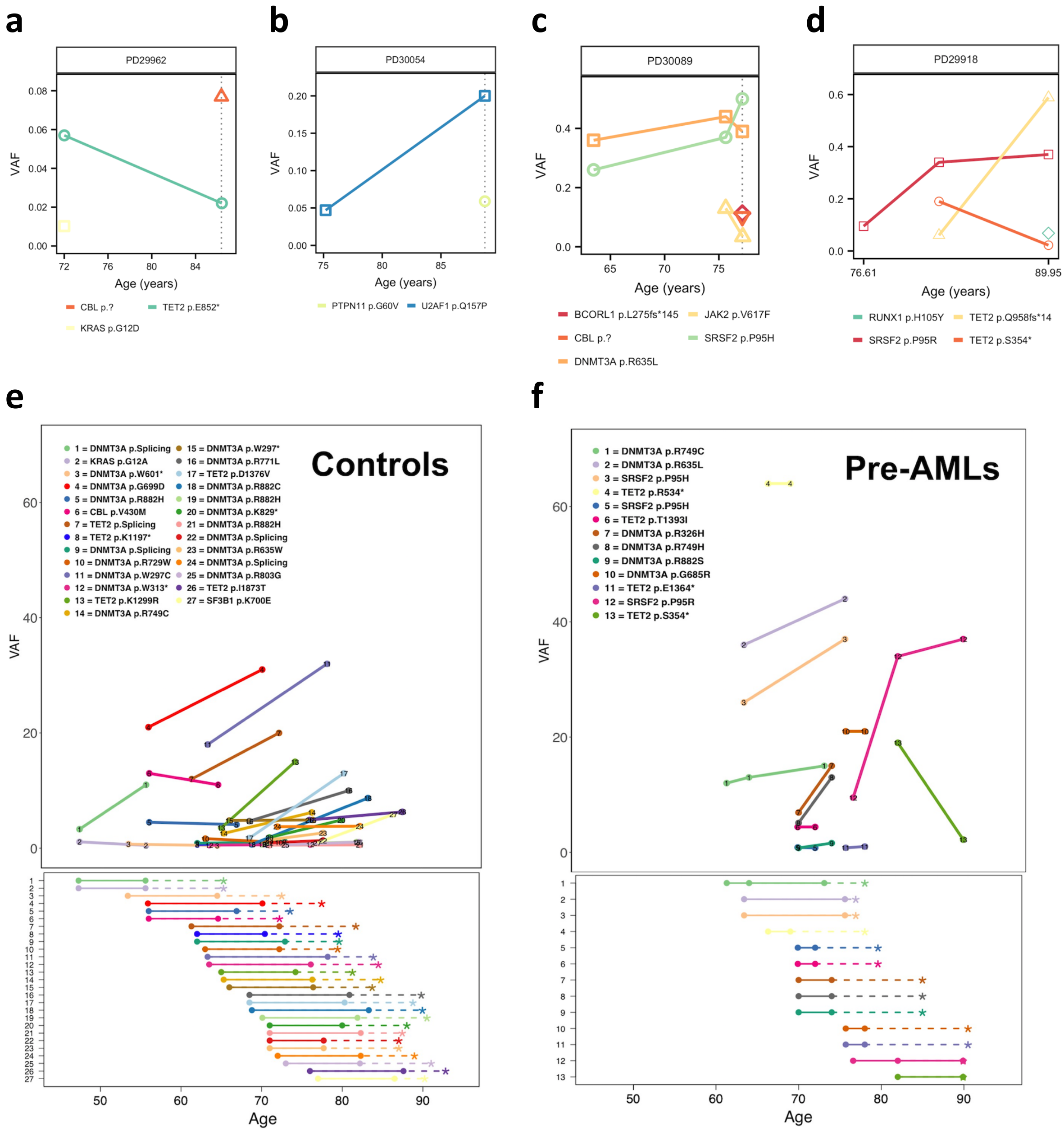


Figure 3.2 | Evolution of clonal haematopoiesis and relationship with future AML. **a-c,** VAF trajectories of putative driver mutations in three individuals for whom bone marrow biopsy specimens taken at time of AML diagnosis (dashed black vertical line) were available for sequencing. Note that coverage for the diagnostic sample of PD30089 was insufficient to meaningfully compare the relative VAFs of the drivers in *DNMT3A* and *SRSF2*. **d,** VAF trajectories of driver mutations in an individual sampled three times, with last sample taken one month before AML diagnosis.

e,f, VAF trajectory of persistent clones carrying putative driver mutations in controls (**e**) and pre-AML cases (**f**). Upper plots: Circles denote individual serial samples and solid lines representing the growth trajectory between serial samples. Lower plots: dashed lines indicate the time interval between the last sampling and the end of follow-up (controls) or AML diagnosis (cases). Code for panels e and f by Dr Sagi Abelson.

We sought to formally assess whether the clonal expansion rate was significantly different for the serial samples taken from controls versus pre-AMLs. However, this measurement is confounded by multiple factors, not least the inability to determine whether or not co-occurring mutations reside in the same clone. Hence, this experiment is inadequate to draw any conclusions. Studying the impact of mutation on AML development at the clonal level, for example by culturing and sequencing single-cell derived colonies, would help to address this question (Nangalia et al., 2019).

2.3 The genetic landscape of pre-AML versus CH

In line with previous studies of CH in the general population (Jaiswal et al., 2014; Xie et al., 2014), *DNMT3A* and *TET2* were the most commonly mutated genes in both groups (Figure 3.3a). No canonical *NPM1* mutations nor any *FLT3*-internal tandem duplication mutations were detectable, consistent with these arising late in leukaemogenesis (Kronke et al., 2013; McKerrell et al., 2015). Recurrent *CEBPA* mutations, which are implicated in around 10% of *de novo* AML (Papaemmanuil et al., 2016), were also absent, suggesting that driver events in this gene may also be late events in *de novo* AML evolution, despite their involvement in familial AML. Notably, mutations in splicing factor genes (*SF3B1, SRSF2* and *U2AF1*) were significantly enriched among the pre-AML cases relative to the controls (odds ratio, 17.5; 95% confidence interval, 8.1–40.4; *P* = 5.2×10^{-16}, two-sided Fisher's exact test) and were present in significantly younger individuals (median age 60.3 compared to 77.3 years, *P* = 1.7 × 10−4, two-sided Wilcoxon rank-sum test; Figure 3.3b). Screening all SNPs for potential pathogenic germline variants relevant to cancer or blood disorders (Methods section 3.4) identified only one likely pathogenic lesion, *MPL* p.Q186K (ClinVar accession RCV000015217.22). This SNP has been implicated in congenital amegakaryocytic thrombocytopenia (Ihara et al., 1999), though the participant carrying it (PD30060) had normal pre-diagnosis blood counts and developed AML aged 91.

Figure 3.3

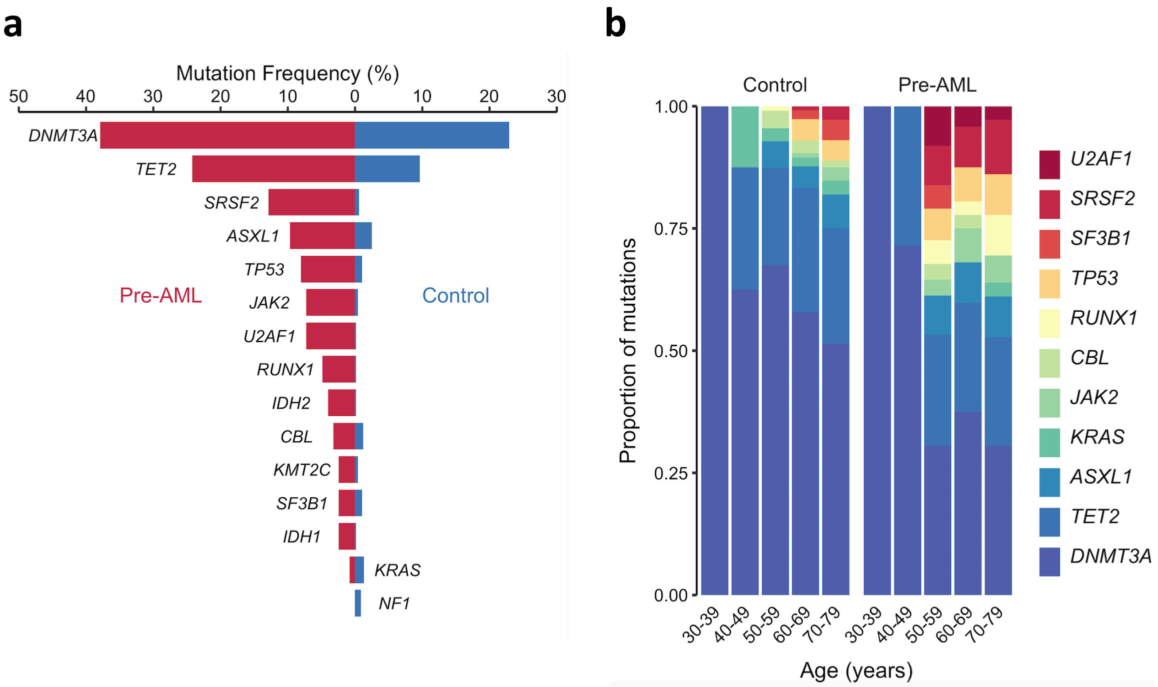


Figure 3.3 | The mutational landscape of clonal haematopoiesis in pre-AML and controls. a, Proportion of pre-AML cases (red) and controls (blue) who had CH-PD mutations in recurrently mutated genes. **b**, Relative frequency of mutations in the indicated genes according to age group for pre-AML cases and controls. **P* < 0.05, Fisher's exact test with Bonferroni multiple testing correction.

2.4 Genetic AML risk prediction model

These findings demonstrate marked differences in both mutation burden and driver landscape between CH-PD observed in controls and pre-AML. Moreover, these results, in conjunction with recent insights into the origins of AML relapse (Shlush et al., 2017), suggests that AML progression typically occurs over many years through clonal evolution of pre-leukaemic haematopoietic stem and progenitor cells (HSPCs) before acquisition of late mutations leads to overt malignant transformation. In order to quantify the relative contributions of driver mutations and clone sizes to the risk of progressing to AML, we applied a Cox proportional hazards regression approach, which achieved similar performance in both the discovery cohort (concordance (C) = 0.77 ± 0.03) and the validation cohort (C = 0.84 ± 0.05; Figure 3.4a-f and Table 3.1). A ridge regularised logistic regression model trained using the same variables produced very similar results (Table 3.2) As discussed in Methods section 4.1, we used weighting to minimise the biases introduced by the artificial case-control ratio (Antoniou et al., 2005; Therneau and Grambsch, 2000) and calculated hazard ratios relative to the (approximate) true cumulative incidence of about 1-3/1,000 in the given age range over a follow up of 10-20 years. The observed driver mutation frequency and VAF in pre-malignant samples closely resembled values expected based on the estimated risks, indicating that risk model and driver prevalence are well aligned (Figure 3.4g-h).

Table 3.1 Cox proportional hazard model performance

Cox proportional hazards model	Concordance	Standard error	Time-dependent AUC
VC data and fit	0.84	0.05	0.74
DC data and fit	0.77	0.03	0.78
VC fit DC data	0.72	0.03	0.7
DC fit VC data	0.82	0.05	0.79
Combined cohorts	0.77	0.05	0.79*

*Derived from 100 bootstraps out-of-bag validation
DC, discovery cohort; VC, validation cohort

Figure 3.4

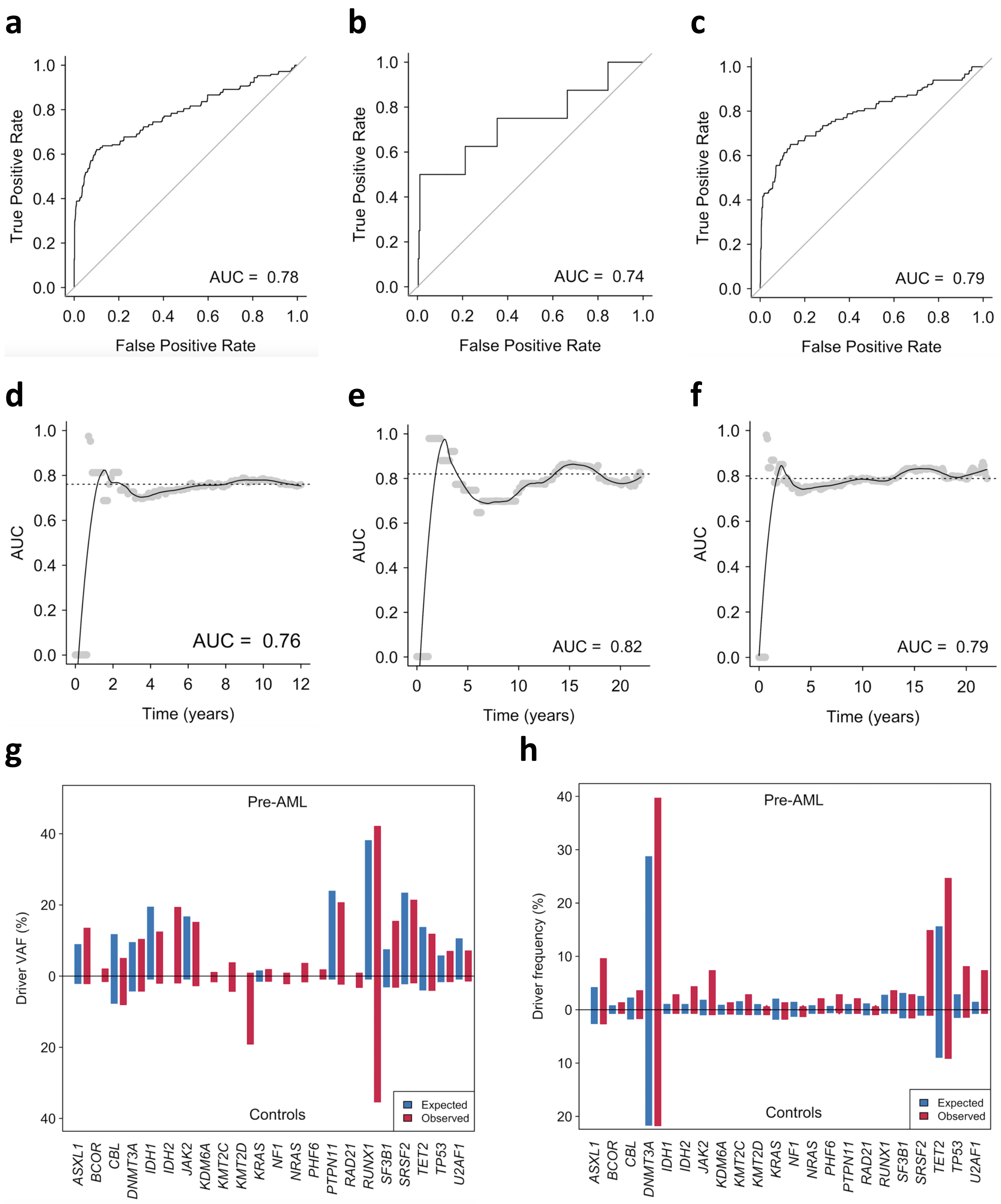


Figure 3.4 | AML predictive model performance. a–c, Time-dependent receiver operating characteristic curve for Cox proportional hazards model of AML-free survival trained on the discovery cohort (n = 505 unique individuals, 91 pre-AML and 414 controls) (**a**), validation cohort (n = 291 unique individuals, 29 pre-AML and 262 controls) (**b**) and combined cohorts (**c**). **d–f**, Dynamic AUC for Cox proportional hazards models trained on the discovery cohort (**d**), validation cohort (**e**) or combined cohort (**f**). **g, h**, Red and blue bars indicate the observed and expected VAF (**g**) and driver frequency (**h**) of pre-AML cases and controls for each gene indicated on the *x* axis. One can speculate that the discrepancies between expected and observed driver VAF for RUNX1 and KMT2D relate to the relatively high prevalence of pathogenic germline mutations seen in these genes and the challenge in distinguishing the latter from somatic drivers.

Table 3.2 Ridge regularised logistic regression model performance

Ridge regularised logistic regression	AUC
VC data and fit	0.85
DC data and fit	0.76
VC fit DC data	0.69
DC fit VC data	0.81
Combined	0.81*

*Derived from 100 bootstraps out-of-bag validation

DC, discovery cohort; VC, validation cohort

Models that were only trained on data from the discovery or validation cohort had similar coefficients (Figure 3.5, Appendix 10). We therefore combined the datasets for a more accurate analysis of the contributions of mutations in individual genes to risk (C = 0.77 ± 0.05; area under curve, 0.79; Figure 3.4c,f and Table 3.1).

Quantitatively, we found that driver mutations in most genes conferred an approximately twofold increased risk of developing AML per 5% increase in clone size (Figure 3.5). Notable exceptions to this trend were the most frequently mutated CH genes, *DNMT3A* and *TET2*, which conferred a relatively lower risk of progression to AML (Figure 3.5, Fig 3.6a,c,e). By contrast, a larger effect size was apparent for *TP53* (hazard ratio, 12.5; 95% confidence interval, 5.0–160.5) and *U2AF1* (hazard ratio, 7.9; 95% confidence interval, 4.1–192.2) mutations (Figure 3.5, Figure 3.6a,b,d). However, other CH-PD genes, such as *SRSF2*, contributed a similar relative risk owing to their presence at a higher VAF in pre-AML cases (Figure 3.5, Figure 3.6a). Because the effect of each driver mutation is deleterious and the effect of multiple mutations that are present in the same individual is multiplicative, a higher number of mutations is predicted to increase the risk of progression to AML (Figure 3.7a). Similarly, the size of the largest driver clone was also strongly associated with the risk of progression to AML, in agreement with the risk of individual mutations generally being proportional to VAF (Figure 3.7b).

Estimates of model sensitivity and specificity necessitate arbitrary age-cut-offs which dramatically impact the interpretation of predictions. Is it most relevant to know whether or not an individual will develop AML before age 100 or before age 60 and which estimate should sensitivity/specificity be determined for? The Cox proportional hazards model illustrated in figure 3.5 facilitate a more tangible interpretation of excess risk on an individual level, harnessing the genomic snapshot from a blood sample to estimate the risk of developing AML over the next 10 years in a manner which accounts both for a person's age and the incidence of AML in their given age bracket.

Comparing AML risk prediction models based on the VAF of mutations in individual genes versus mutation burden alone demonstrated that the gene-level model performed best (Figure 3.7c,d). Concordance and AUC were both 3-4% improved for the models incorporating gene-level risk, which is a considerable margin, particularly for a rare disease. Moreover, the disparities in gene-level hazard ratios (HR) were significant (Figure 3.5), despite the fact that the genes with the highest HR are not mutated frequently enough to have a very dramatic effect on overall model AUC. Collectively, although the VAF and the number of mutations confer much of the predictive value, the gene-level analysis (Figure 3.5) does demonstrate distinct gene-level risks, and is able to quantify the cumulative impact of multiple mutations and clonal size on the likelihood of progression to AML. Furthermore, in order to examine whether the genetic model can distinguish between CH-PD and pre-AML even when individuals without mutations were excluded, we retrained the model using only cases and controls with CH-PD. We found that performance was if anything marginally improved by this manoeuvre (Concordance > 0.8 on both discovery and validation cohorts, Appendix 7).

Figure 3.5

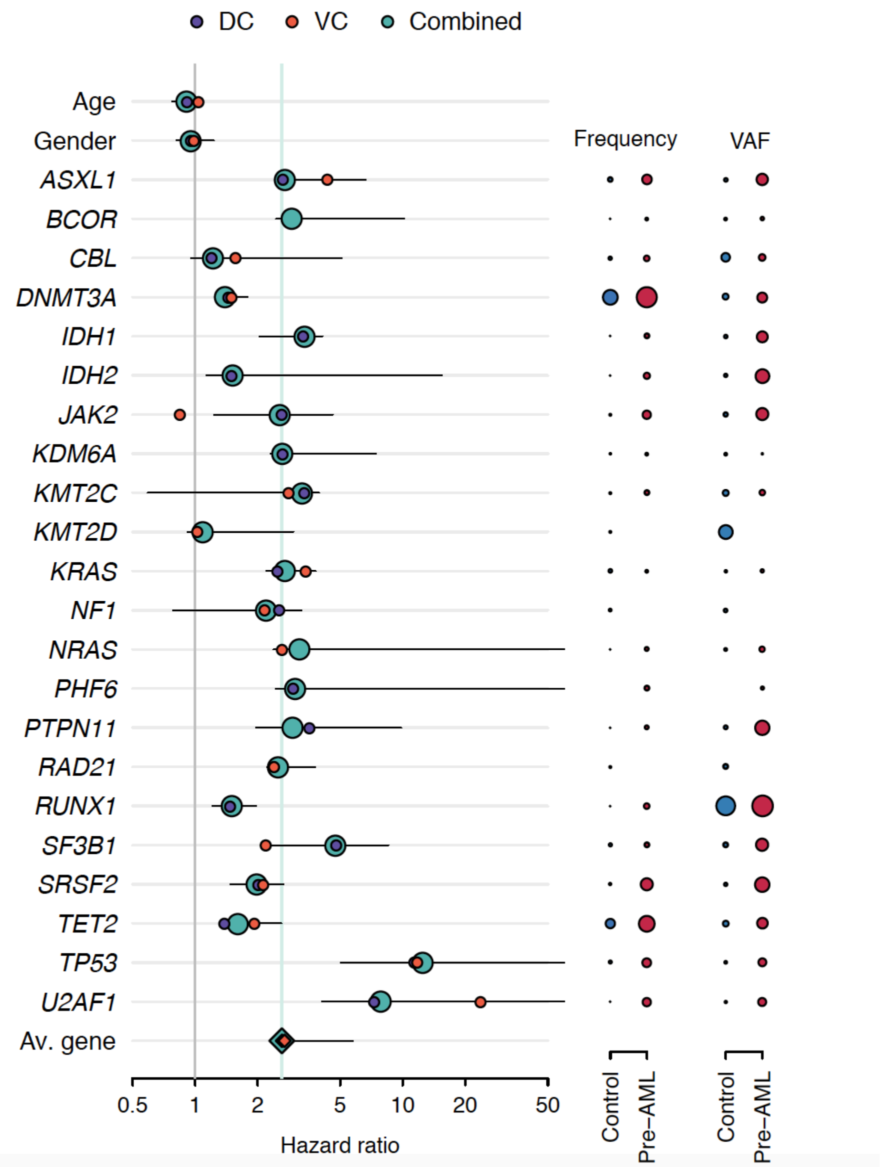


Figure 3.5 | Forest plot indicating gene-level hazard ratios for risk of developing AML. Purple, orange and green circles indicate hazard ratios (HR) for the discovery (DC), validation (VC) and combined cohort, respectively. Horizontal lines denote 95% confidence intervals for the combined cohort. For each gene, the indicated HR applies to the 10-year risk of AML conferred by each 5% increase in mutation VAF. The green vertical line indicates the mean HR across all genes. The HR for *RUNX1* must be interpreted with caution owing to the relatively high prevalence of deleterious germline variants in this gene, which may not be readily distinguishable from somatic mutations in unmatched sequencing assays. The proportion of individuals with mutations in each gene and the average VAF are indicated to the right of the forest plot.

Figure 3.6

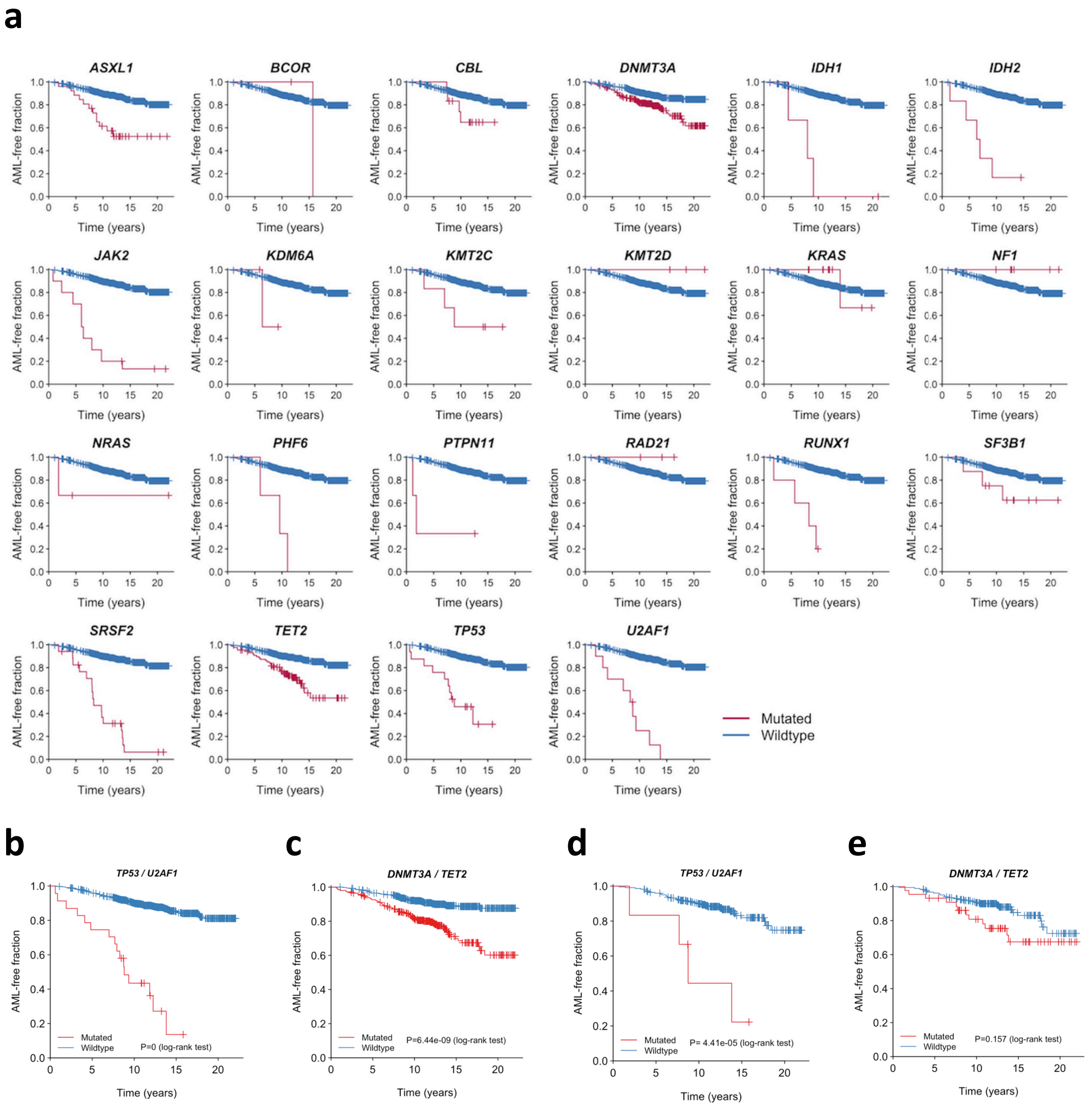


Figure 3.6 | Gene-level impact on AML-free survival. a, Kaplan–Meier (KM) curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status in genes mutated in at least three samples across the combined validation and discovery cohorts. *n* = 796 unique individuals. **b-c** For illustrative purposes, KM curves according to co-mutation status in *DNMT3A/TET2* and *TP32/U2AF1* are shown. All patients harbouring any mutation in *TP53* or *U2AF1* (**b**) or *DNMT3A* or *TET2* (**c**). **d,e** The same relationship between mutation status and AML-free survival persists when considering only individuals with a total of one driver mutation. KM curves for participants with their only driver mutation in either *DNMT3A* or *TET2* (**d**) or *U2AF1* or *TP53* (**e**). Red and blue lines indicate mutated and wildtype, respectively. *P*-values for significance of survival differences by mutation status calculated by the log-rank test. AML, acute myeloid leukaemia; KM, Kaplan-Meier.

Figure 3.7

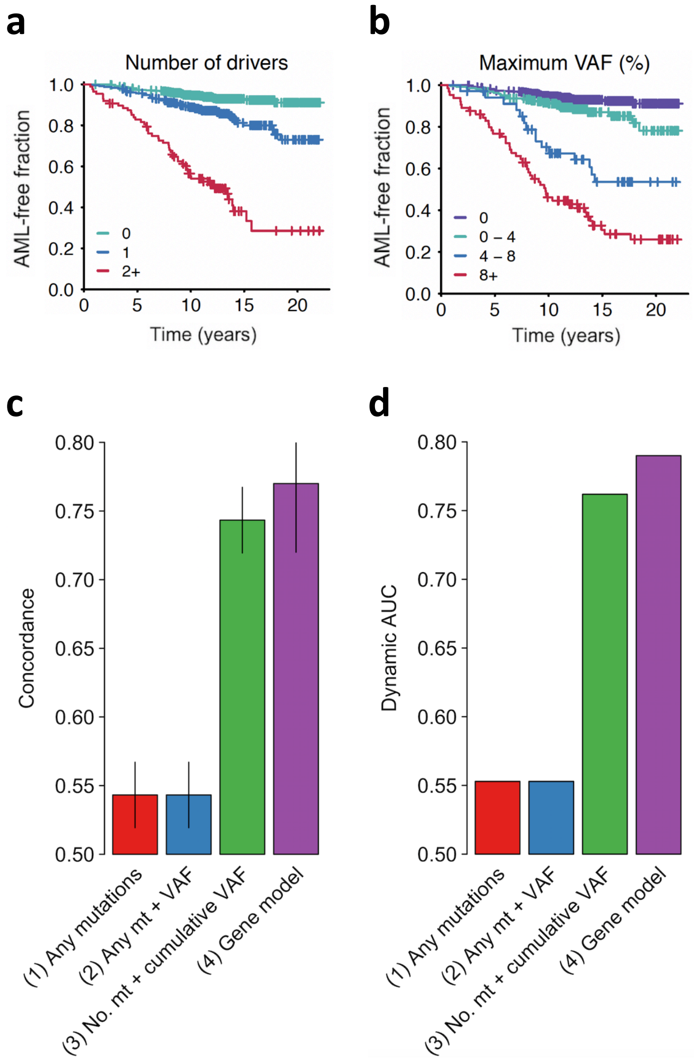


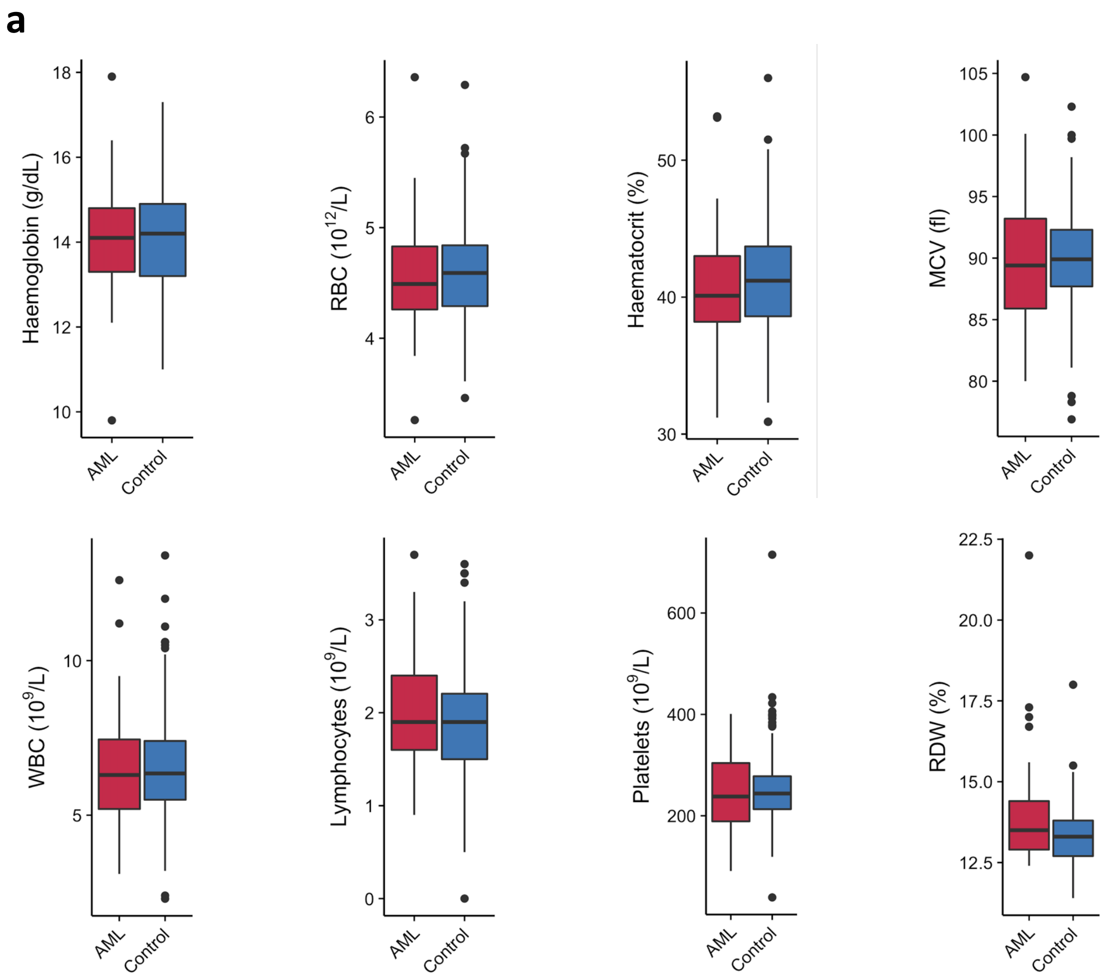
Figure 3.7 | Performance of AML risk prediction models based on gene-level factors versus mutation burden.

a-b, Kaplan–Meier curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to number of driver mutations per individual (**a**) and largest clone detected (**b**). VAF bins of 4% are shown in (**b**) to illustrate the consistency of the trend towards lower AML-free survival with larger clone size. **c**, Leave-one-out crossvalidated concordance C of different risk models based on (1) the presence of any mutation, (2) the presenced of any mutation and the cumulative VAF of different clones, (3) the number of different driver mutations and cumulative VAF as predictors and (4) a model incorporating the effects of individual genes. **d,** Same models as in (**c**), but using Uno’s dynamic AUC as a measure of model performance. VAF, variant allele fraction; mt, mutation; No. mt, number of mutations; AUC, area under the curve.

2.5 Clinical factors associated with AML risk

Although genetic features alone are capable of identifying many individuals at risk of developing AML in these experimental cohorts, AML incidence rates in the general population are low (4:100,000) (Deschler and Lubbert, 2006), and thus millions of individuals would need to be screened to identify the few pre-AML cases, with many false positives. To determine whether routinely available clinical information could improve prediction accuracy or identify a high-risk population for targeted genetic screening, I initially reviewed full blood count and biochemistry data that were available for 37 of the pre-AML cases and 262 controls. These data also permitted a screen for any potentially undiagnosed cases of MDS, a known risk factor for (secondary) AML (Arber et al., 2016). The diagnosis of MDS based on the WHO criteria relies not only on the presence of dysplasia in at least one lineage, but also on the presence of at least one significant cytopenia (haemoglobin (Hb) <10g/dL; platelet count<100 x10^{9}/L and absolute neutrophil count<1.8 x 10^{9}/L)(Arber et al., 2016). The latest WHO criteria state verbatim that "Cytopenia is a 'sine qua non' for any MDS diagnosis…", hence enabling exclusion of MDS based on normal blood counts alone (Arber et al., 2016). Out of the 37 pre-AMLs only one had Hb<10g/dL at recruitment (PD30116, Hb 9.8g/dL); however, three years later Hb had normalised to 13.7g/dL, thus excluding MDS. The only other cytopenia in a pre-AML was a sample with platelets of 91 x 10^{9}/L at baseline (PD30010); however, 3.7 years later the platelet count had risen above the WHO guideline threshold (106 x 10^{9}/L), suggesting that MDS was not the diagnosis. CH-PD was also overwhelmingly associated with normal blood counts in the controls, even in individuals harbouring multiple mutations at high VAF (e.g., PD35659c, PD35733b and PD35788b with leukaemia-free follow-up of 20.3, 20.4 and 17 years, respectively). The presence of normal blood counts in association with large clones corroborates the findings of previous studies of CH in the general population (Buscarlet et al., 2017; Jaiswal et al., 2014; McKerrell et al., 2015). Overall, full blood count data between controls and pre-AMLs did not differ, with the notable exception of red cell distribution width (RDW) (Figure 3.8a,b) Despite the limited sample size, there was a significant association between higher RDW and risk of progression to AML (P = 0.0016, Wald test with Bonferroni multiple-testing correction). Although traditionally used in the evaluation of anaemias, raised RDW has been correlated with inflammation, ineffective erythropoiesis, CVD and adverse outcomes in several inflammatory and malignant conditions (Hu et al., 2017). The correlation between RDW and risk of AML development remained highly significant when only controls with CH-PD were compared to pre-AMLs ($P = 3.5 \times 10^{-6}$, Wald test with Bonferroni multiple testing correction). Higher RDW has previously been associated with CH and overall mortality (Jaiswal et al., 2014; Salvagno et al., 2015), but has never been shown to distinguish CH from pre-leukaemia.

Figure 3.8



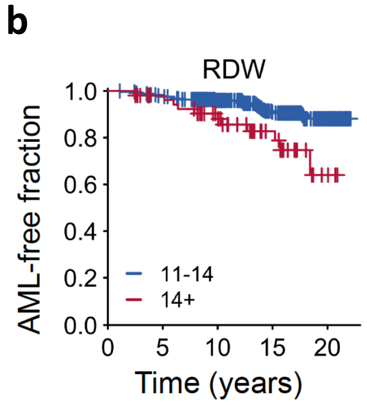


Figure 3.8 | Full blood count indices in pre-AMLs and controls. a, Box plots of full blood count parameters. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. **b**, Kaplan–Meier curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curve is stratified according to RDW measurement data for n = 299 unique individuals for whom full blood count measurements were available. Among the blood indices shown, only RDW was significantly different between pre-AML cases and controls (P = 0.0016, Wald test with Bonferroni multiple-testing correction).

In order to verify RDW as a predictive factor and determine whether additional clinical parameters are associated with risk of AML development, we collaborated with Dr Netta Mendelson Cohen, Dr Elisabeth Niemeyer and Dr Noam Barda, who analysed the Clalit electronic health record (EHR) database (Balicer and Afek, 2017). This resource contains EHRs for an average of 3.45 million individuals per year collected over a 15-year period. Stringent criteria based on diagnostic codes and treatment records identified 875 AML cases (Appendix 11). Consistent with case ascertainment strategy for the genetic model, all cases of secondary AML following another myeloid malignancy were excluded. Analysis of RDW trends revealed significantly raised measurements several years before AML diagnosis relative to age and sex-matched controls (Figure 3.9a). The most pronounced increase in RDW was observed at 6-12 months before diagnosis, with ~10% of pre-AMLs having RDW values which were greater than the 99th centile of the controls. Many other blood indices, including several full blood count (FBC) parameters, changed six months to a year before diagnosis. Additional parameters that correlated with risk of AML development included reductions in monocyte, platelet, red blood cell and white blood cell counts (Figure 3.9a). However, in the majority of cases measurements did not fall outside the normal reference ranges. Nevertheless, these values were statistically distinct from those seen in large numbers of age and sex-matched controls. This is important, as it shows that these individuals did not have undiagnosed MDS/MPN, and suggests instead that evolving *de novo* AML may sometimes have a considerable prodrome with subtle but discernible clinical manifestations, potentially reflecting large pre-leukaemic clones.

Our collaborators next applied a machine-learning approach to construct an AML prediction model based entirely on variables that are routinely documented in electronic health records (Appendix 11). This model predicted AML 6–12 months before diagnosis with a sensitivity of 25.7% and overall specificity of 98.2%. The model performed consistently across different age groups with an increased relative risk of 28 for males and 24 for females between the age of 60 and 70 years (Figure 3.9b). To our knowledge this represents the first analysis of its kind in AML prediction from routinely collected clinical records. In order to better understand which patients are most likely to be accurately classified by this model, our collaborators compared absolute laboratory values for true positives and false negatives. This revealed that 35.5% of false-negative predictions were for patients for whom infrequent blood count data were available. Some of the true-positive cases had mildly abnormal blood counts that would not initiate a diagnostic work-up (Figure 3.9c), whilst cytopenias that would be compatible with undiagnosed myelodysplastic syndrome (Arber et al., 2016) were uncommon. Other non-haematological variables associated with progression to AML included higher triglyceride levels and lower high- and low-density lipoprotein levels (Figure 3.9d).

Figure 3.9

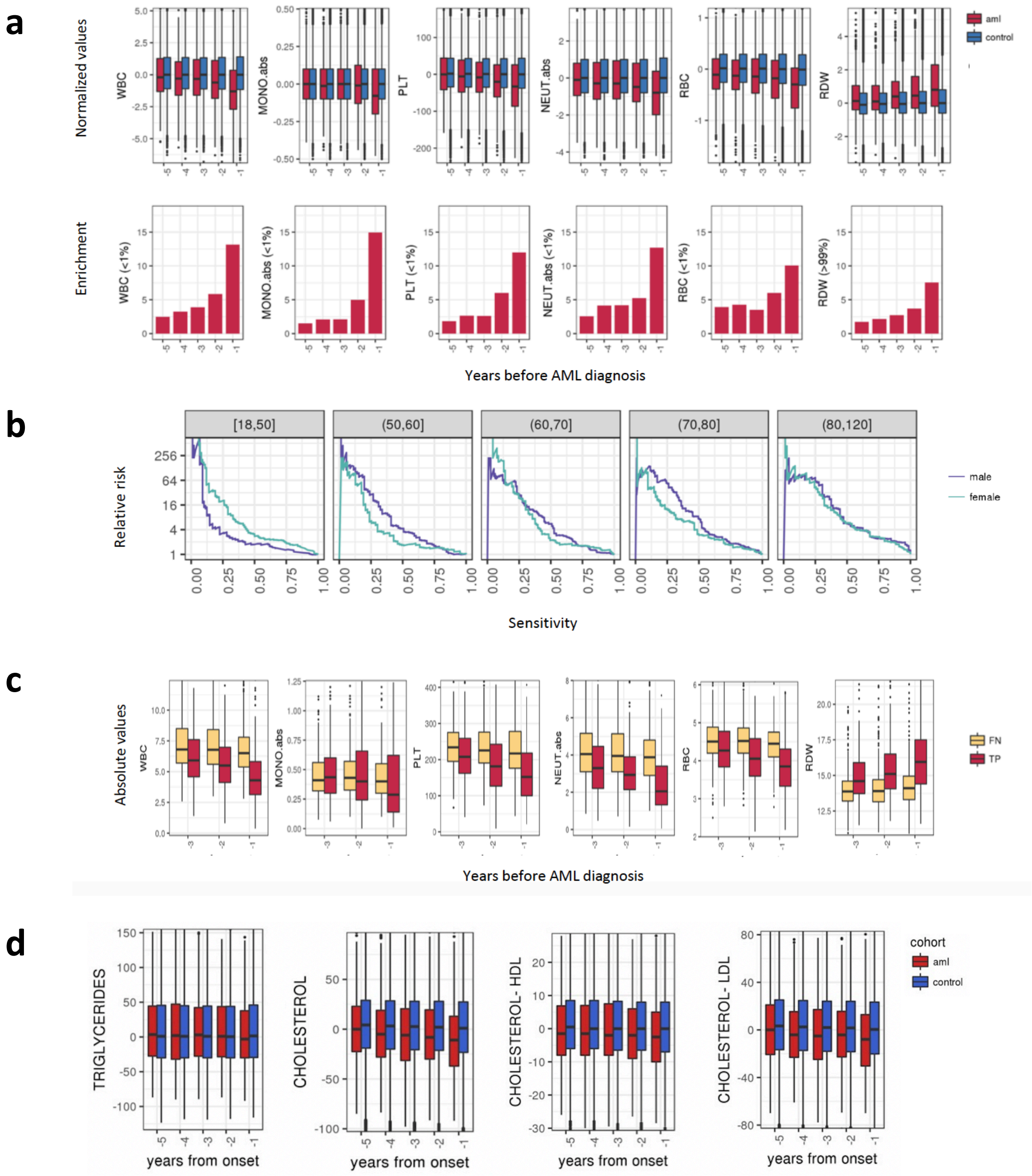


Figure 3.9 | Increased risk of AML development inferred from electronic health records. a, Box plots of normalized laboratory measurements. Increased RDW, reduction in monocyte, platelet, red blood cell (RBC) and white blood cell (WBC) counts (top) show a high association (bottom) with a higher risk of AML development and differed at least a year before AML diagnosis. **b**, Model performance stratification by age and gender. Age ranges are indicated above each graph. **c**, Absolute laboratory values for true positive (TP) and false negative (FN) predictions. **d,** Box plots of lipid levels. Box plots indicate median, first and third quartiles and 1.5× interquartile range. WBC, white blood cell count; MONO.abs, absolute monocyte count; PLT, platelet; NEUT, neutrophil; RBC, red blood cell; RDW, red cell distribution width; FN, false positive; TP, true positive; AML, acute myeloid leukaemia; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

3. Discussion

This study sought to explore the natural history and genetic landscape of nascent AML and the extent to which the latter is distinct from CH in the general population. Collectively, these findings provide new insights into the pre-clinical evolution of AML and the feasibility of identifying CH at high risk of malignant transformation.

3.1 A long latency period is the rule rather than the exception in AML

This work demonstrates for the first time that pre-leukaemic clones can be detected in the majority of individuals who develop AML 6 or more years before clinical disease manifestations, even when interrogating for point mutations alone. This long latency has now also been reported by Desai et al, who performed a very similar nested case-control study (Desai et al., 2018). Desai and colleagues sequenced 67 AML-associated genes in peripheral blood samples from 212 women diagnosed with AML a median of 9.6 years later alongside the same number of controls (Desai et al., 2018). Consistent with our results, pre-leukaemic clones (VAF>1%) were present in 68.6% and 30.9% of pre-AML cases and controls, respectively (Desai et al., 2018). This long pre-clinical evolution highlights important aspects of AML biology and reveals that the window for potential intervention is measured in years for the majority of individuals who develop AML.

3.2 The distinct driver landscape of pre-AML

This work also reveals that the mutational landscape, and not simply the mutation burden, differs between CH in controls versus pre-AML. The differences in the mutational spectrum observed between pre-AML cases and controls may arise through cell-intrinsic or -extrinsic factors. As discussed in Chapter 1, previous studies of clonal haematopoiesis have demonstrated that clones with particular mutations dominate in the context of specific environmental pressures (Gibson et al., 2017; Hsu et al., 2018; McKerrell et al., 2015; Takahashi et al., 2017; Wong et al., 2015b), suggesting an important role for cell-extrinsic factors in haematopoietic somatic evolution. Although such factors in CH remain poorly understood, it is intriguing that mutations in splicing factor genes and *TP53* were significantly enriched among the pre-AMLs relative to the controls, with the former presenting in significantly younger individuals than in benign CH. Spliceosome mutations appear to confer a competitive advantage in the context of ageing, and were almost exclusively observed in the general population in individuals over age 70 years (McKerrell et al., 2015). Similarly, clones harbouring *TP53* mutations expand dramatically with exposure to intensive chemo- and/or radiotherapy (Bondar and Medzhitov, 2010; Wong et al., 2015b). However, *TP53*-mutated HSC clones are very common at extremely low VAF in the elderly, but tend to remain stable in size over time, suggesting only a modest selective advantage in the absence of increased genotoxic stress (Wong et al., 2015b). Therefore, it is possible that the significantly higher prevalence of clones with *TP53* and spliceosome gene mutations in pre-AML cases may reflect distinct microenvironmental selection pressures rather than earlier mutation acquisition.

3.3 The significance of the higher mutation burden in pre-AML

The observation of the higher burden of putatively oncogenic mutations (driver mutations) in the pre-AML cases across all age groups raised two main related questions. Firstly, what is the mechanism underpinning the discrepancy in mutation burden between controls and pre-AMLs? Secondly, do driver mutations detected in pre-AML cases reflect the presence of an AML ancestor, or do these mutations behave as surrogate markers of factors predisposing to leukaemogenesis?

Although speculative, several mechanisms may account for the higher mutation burden and clone size observed in the pre-AMLs. It could reflect a higher mutation rate in the pre-AML cases, for example due to higher HSC turnover, potentially secondary to depletion of the functional HSC pool. Alternatively, chance may play a dominant role, with stochastic driver mutation acquisition triggering clonal expansion, thus increasing the odds of further driver events on a pre-malignant background leading to selection for progressively more mutated clones. However, this multistage cancer evolution paradigm does not account for the relationship between the fitness advantage conferred by a driver mutation and the environmental context of the mutated cell (Rozhok et al., 2014). Clones with drivers could be under stronger selective pressure in certain bone marrow environments, as is seen in particular clinical contexts such as aplastic anaemia or after intensive cytotoxic therapy (Hsu et al., 2018; Wong et al., 2015b; Yoshizato et al., 2015). As discussed in the introduction, the presence of selective pressure favouring clonal expansions, rather than mutation acquisition, may thus be an important determinant of the number of mutations detectable by bulk sample sequencing.

Our time series experiment and sequencing of diagnostic specimens helped partially address the second question, demonstrating that clones in pre-AML cases represent a combination of leukaemia ancestors and ‘bystander’ clones that likely are not related to the future AML. However, our experiment using bulk cell populations was too small and hindered by confounding factors to enable strong conclusions about clonal growth kinetics or mutation rates. We hope that future experiments using single cell and/or highly purified cell population studies on viable cells at serial time points will shed light on these questions.

3.4 Rationale for AML risk prediction and future directions

Cancer predictive models have enabled successful early detection and intervention programmes for several solid tumours (Vickers, 2011; Wang et al., 2014). However, screening tests are unavailable for the sub-clinical stages of most haematological malignancies. Given that the main cause of mortality in AML is treatment resistance/relapse (Döhner et al., 2015), there is a rationale for identifying and treating a genomically simpler antecedent of the disease. In this context, reduction of clonal size rather than complete clonal extinction may be sufficient to significantly reduce the risk or slow AML progression. Such an approach has proven very effective in CML, which has been transformed by targeted therapy into a chronic condition with a dramatically reduced incidence of progression to CML blast crisis (Kalmanti et al., 2015). Furthermore, CH is associated with and may play a causal role in common non-malignant conditions (Fuster et al., 2017; Jaiswal et al., 2017), which may strengthen the case for screening and intervention.

3.4.1 Further development of genetic AML prediction methods

This study provides proof-of-concept for the feasibility of early detection of healthy individuals at high risk of developing AML. The models presented here demonstrate that somatic genetic features are predictive of AML progression and that the presence of mutations in certain genes confers a greater risk. Desai et al have since identified similar gene-level risk factors (Desai et al., 2018). Consistent with our results, *TP53* mutations conferred the highest odds ratio of progression from CH to AML, followed by drivers in *IDH1/2* and spliceosome genes (Desai et al., 2018). Although Kaplan-Meier analysis (Figure 3.6) is consistent with a trend towards shorter AML-free survival with *IDH1/2* mutations, we chose not to group functionally-related genes in our analysis in order to reach significance, as their mechanistic consequences may differ (e.g., *IDH2* p.R140 and *IDH2* p.R172 (Papaemmanuil et al., 2016)). In addition to improving model performance, the identification of highly significant disparities in gene-level HR offers compelling biological insights into the determinants of clonal progression, which warrant further investigation.

Given that most of the genetic model's predictive power stems from mutations with VAFs >0.005, our data suggests that conventional deep targeted sequencing, as used for the validation cohort, is adequate for future screens when combined with stringent variant calling and driver mutation curation. Thus, the additional cost of error correcting sequencing is unlikely to be justified. However, it is possible that future studies may show that specific mutations may have predictive value when detected accurately even at low VAF (e.g. *U2AF1* hotspot variants).

As recurrent chromosomal translocations are likely to be initiating events in approximately 20% of AML (Papaemmanuil et al., 2016), incorporating these into the genetic model is likely to further increase predictive accuracy. McKerrell et al. have shown that it is feasible to simultaneously capture several recurrent translocations/inversions with targeted panels only slightly larger than the ones used in the current study (McKerrell et al., 2016). Additionally, expanding this dataset will make it possible to investigate whether co-mutation patterns carry prognostic significance, as is the case in AML (Gerstung et al., 2017; Papaemmanuil et al., 2016).

3.4.2 Combining clinical and genetic information to risk-stratify clonal haematopoiesis

The predictive model based on mutations and demographic features partially overcomes the limitations imposed by the low overall incidence of AML, but does not eliminate them. We have shown that commonly recorded clinical parameters, notably RDW and other FBC indices, may identify a smaller population with higher pre-test AML risk for screening. Although clinical parameters were predictive relatively close to the time of AML diagnosis, pre-AML clones can be of significant size many years before diagnosis and it is entirely plausible that surrogate laboratory markers of their presence may be identifiable much earlier, as we found for RDW in the validation cohort. Analysis of the 37 individuals for whom both genomic and clinical information were available found that 6% of the relative risk contribution was attributable to clinical variables, suggesting that combining routinely available clinical data with genomic variables may strengthen AML prediction models. Extending this analysis in a large EHR database further revealed that pre-AML has additional subtle clinical manifestations which in themselves had considerable predictive power 6-12 months prior AML diagnosis. This further supports a role for clinical variables in strengthening genomic prediction models and/or in targeting the population most likely to benefit from screening for CH.

Defining the population most likely to benefit from genetic screening will also depend on improved understanding of the role of CH in common non-malignant conditions. If, as several recent studies strongly suggest, some pre-leukaemic clones are pro-inflammatory and actively promote atherosclerosis and cerebro/cardiovascular adverse events (Fuster et al., 2017; Jaiswal et al., 2017), then a significantly larger proportion of the population might benefit from screening for CH and could thus be considered for possible interventions to suppress pre-leukaemic clones and/or mitigate established cardiovascular risk factors (blood pressure, dyslipidaemia, etc). Our analysis of a large EHR database reveals that subtle clinical manifestations, including trends in triglycerides and RDW that are established risk factors for cardio/cerebrovascular disease also correlated with risk of AML. It is conceivable that there are unifying characteristics of high-risk CH emblematic of the emerging links between ageing and dysregulated inflammation or immune senescence (Green et al., 2011; Shaw et al., 2013).

Clearly these findings cannot address the challenging question of how genomic screening methods should be implemented in a real-world setting, and a combined clinical and genetic screening approach requires validation in large prospective cohort studies. Promisingly, the infrastructure for performing such studies is increasingly available, for example the UK Biobank (Bycroft et al., 2018). These resources should help stimulate large prospective studies that take account of all health outcomes associated with CH.

Chapter 4

The pre-clinical evolution of lymphoid neoplasms

1. Introduction

As discussed in Chapter 1, the initial exome-based screens for CH in the general population established that most somatic mutations occur in a limited number of genes most frequently implicated in myeloid neoplasms (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). However, two of these studies screened broadly for candidate driver events and revealed a broader mutational spectrum, including rare oncogenic mutations in several genes closely associated with lymphoid malignancies, such as *ATM*, *CREBBP* and *MYD88* (Genovese et al., 2014; Xie et al., 2014). The majority of the sensitive, targeted surveys of CH-PD in the general population have since been biased towards detecting mutations in myeloid cancer genes (Acuna-Hidalgo et al., 2017; Coombs et al., 2017; McKerrell et al., 2015; Young et al., 2016). Collectively, these studies have yielded several important insights into CH that were inaccessible to the initial exome screens, for example the high prevalence of small clones harbouring spliceosome gene mutations in older individuals (discussed in Chapter 1, section 3.4.1)(McKerrell et al., 2015; McKerrell and Vassiliou, 2015). Although there is considerable overlap between the cancer genes involved in the commonest lymphoid and myeloid malignancies, the former are generally characterised by more diverse genetic landscapes, with a significant proportion of driver events occurring in infrequently mutated cancer genes (Bolli et al., 2014; Landau et al., 2015; Landau and Wu, 2013; Reddy et al., 2017; Sabarinathan et al., 2017). Given the current literature on CH, it is unclear whether or not a similar spectrum of mutations affecting these less recurrent cancer genes is mirrored in the general ageing population at very low VAF. This is relevant to understanding the selective pressures operative in the ageing haematopoietic niche and to understanding the relationship between CH-PD and lymphoid neoplasms.

As discussed in the introduction to Chapter 3, the studies reporting an association between CH and haematological malignancies were not powered to study distinct classes of blood cancer (Genovese et al., 2014; Jaiswal et al., 2014). The work described in Chapter 3 delineates notable differences in the prevalence and mutational landscape of CH-PD in individuals who later develop *de novo* AML versus that seen in controls, and demonstrates that these genetic features have predictive value for future AML development. However, the extent to which the same is true for other blood cancers remains poorly understood.

The work described in this chapter aims to explore this question by undertaking a broader survey of candidate CH-PD driver genes (Appendix 6) in a cohort of individuals later diagnosed with a lymphoid neoplasm and healthy controls, using a nested case-control experimental design similar to that described in Chapter 3 for AML.

Aims:

1) Compare the prevalence and mutational landscape of CH-PD in the general population with that observed in individuals who go on to develop a lymphoid neoplasm.
2) Correlate genetic features and routinely collected clinical variables with risk of progression to lymphoid malignancy
3) Investigate the combined predictive power of genetic, clinical and demographic features to identify individuals at high risk of developing a lymphoid neoplasm.

2. Results

2.1 Cohort overview

Our EPIC-Norfolk (Day et al., 1999) collaborators (Nick Wareham, Robert Luben, Shabina Hayat and Abigail Britten) identified a discovery cohort comprising 118 study participants diagnosed with a lymphoid neoplasm a mean of 8.0 years (IQR 4.3 - 11.1) after peripheral blood sampling and 118 age- and sex-matched controls with no record of any cancer or haematological disorder (Appendix 12). Individuals were excluded if they were sampled less than 6 months before diagnosis or had a lymphocyte count of 5 x 10^9/L or above, which might be high enough to trigger a clinical work-up for monoclonal B-cell lymphocytosis (MBL) according to current diagnostic criteria (Swerdlow et al., 2016). Given that MBL is a known risk factor for chronic lymphocytic leukaemia (Strati and Shanafelt, 2015), the commonest chronic leukaemia in adults (Dores et al., 2007), we focussed on individuals with lymphocyte counts that would not, in isolation, elicit clinical suspicion of an underlying neoplasm (Swerdlow et al., 2016). The mean age at blood sampling for discovery cohort cases was 64.6 years (IQR 57.0 - 71.8). A validation cohort was also sourced from EPIC-Norfolk and included 71 pre-lymphoid neoplasm (pre-LN) cases and 71 controls (Appendix 13). The mean interval between blood sampling and diagnosis for the validation cohort cases was 8.4 years (IQR 4.1 – 12.3) and mean age at sampling was 64.0 years (IQR 59.4 – 69.8). For the controls, the mean duration of follow-up was 15.4 and 16.4 years for the discovery and validation cohorts, respectively. Serial premalignant samples were available for a subset of the discovery cohort cases and controls. Clinical metadata including full blood count, lipid profile, blood pressure and anthropomorphic measurements were available for the majority of cases and controls. Moreover, out of the 262 controls with clinical metadata described in Chapter 3, 189 were adequately age-and sex-matched to the pre-LN cases, providing a case:control ratio of 1:2 for analysis of clinical factors associated with progression to lymphoid malignancy. These controls were also used to compare mutation frequency in genes that overlapped across the gene panels (Appendices 4 and 6).

The spectrum of future LN diagnoses was similar between the discovery and validation cohorts and is summarised in Table 4.1 with complete metadata for both cohorts detailed in Appendices 12 and 13. For many cases, particularly individuals later diagnosed with a non-

Hodgkin lymphoma, histopathological subtype is unknown. Furthermore, disease classification schemes have evolved dramatically over the course of the recruitment period (Campo et al., 2011; Chapuy et al., 2018; Swerdlow et al., 2016), which would complicate translating historical diagnoses into currently recognised disease entities, and is not essential for the aforementioned aims of this study.

Table 4.1 | Pre-LN cohort summary

Diagnosis	Diagnosis abbreviation	Number of individuals	Mean interval between sample and diagnosis (years)	Mean age at sampling (years)
Peripheral T-cell lymphoma NOS	PTCL NOS	6	8	65.0
Mycosis fungoides	MF	1	3.1	69.9
Non-Hodgkin lymphoma NOS	NHL NOS	37	6.5	65.2
Acute lymphoblastic leukemia	ALL	1	13.3	50.2
Lymphoblastic lymphoma	LL	1	18.5	60.0
Multiple myeloma	MM	43	7.9	63.7
B-cell non-Hodgkin lymphoma	B-NHL	26	7.2	63.2
Diffuse large B-cell lymphoma	DLBCL	25	10.9	64.5
Chronic lymphocytic leukemia	CLL	20	9.1	67.3
Monoclonal gammopathy of undetermined significance	MGUS	12	8.2	65.3
Hodgkin lymphoma	HL NOS	4	14.1	56.1
Small cell B-cell lymphoma	SLL	4	8.4	61.5
Waldenstrom macroglobulinaemia	WM	3	3.9	67.9
Hairy-cell leukemia	HCL	2	4.8	72.8
Nodular sclerosis Hodgkin lymphoma	NScHL	2	5.1	61.3
Extramedullary plasmacytoma	EP	2	6.4	64.2

2.2 Prevalence of CH-PD and driver mutation burden

Peripheral blood samples were deep sequenced with a custom panel comprising 95 genes implicated in haematological malignancies (Methods section 2.4 and Appendix 6). Average sequencing coverage was >5,000 (IQR 4,750 – 5,800). The prevalence of CH-PD was significantly higher in pre-LN cases than in controls (*P* = 0.0019, two-sided Fisher's exact test), though the difference was less dramatic than that observed for pre-AML (Figure 4.1a). Overall the prevalence of CH-PD in pre-LN cases and controls was 35.4% and 20.6%, respectively (Figure 4.1a,b). These proportions were similar across the discovery cohort (CH-PD prevalence of 33.9% in cases and 17.8% in controls) and validation cohort (38% and 25.4% for cases and controls, respectively). The average number of driver mutations identified in pre-LN cases was 0.43 compared to 0.25 for controls (*P*=0.0016, two-sided Wilcoxon rank-sum test), with a significant trend towards increasing driver mutation burden with age (Figure 4.1c). Moreover, as seen for pre-AMLs, the VAF of driver mutations was significantly higher in pre-LN cases versus controls (median VAF 6.9% and 2.8%, respectively; *P* = 0.00036, Wilcoxon rank-sum test; Figure 4.1d).

Figure 4.1

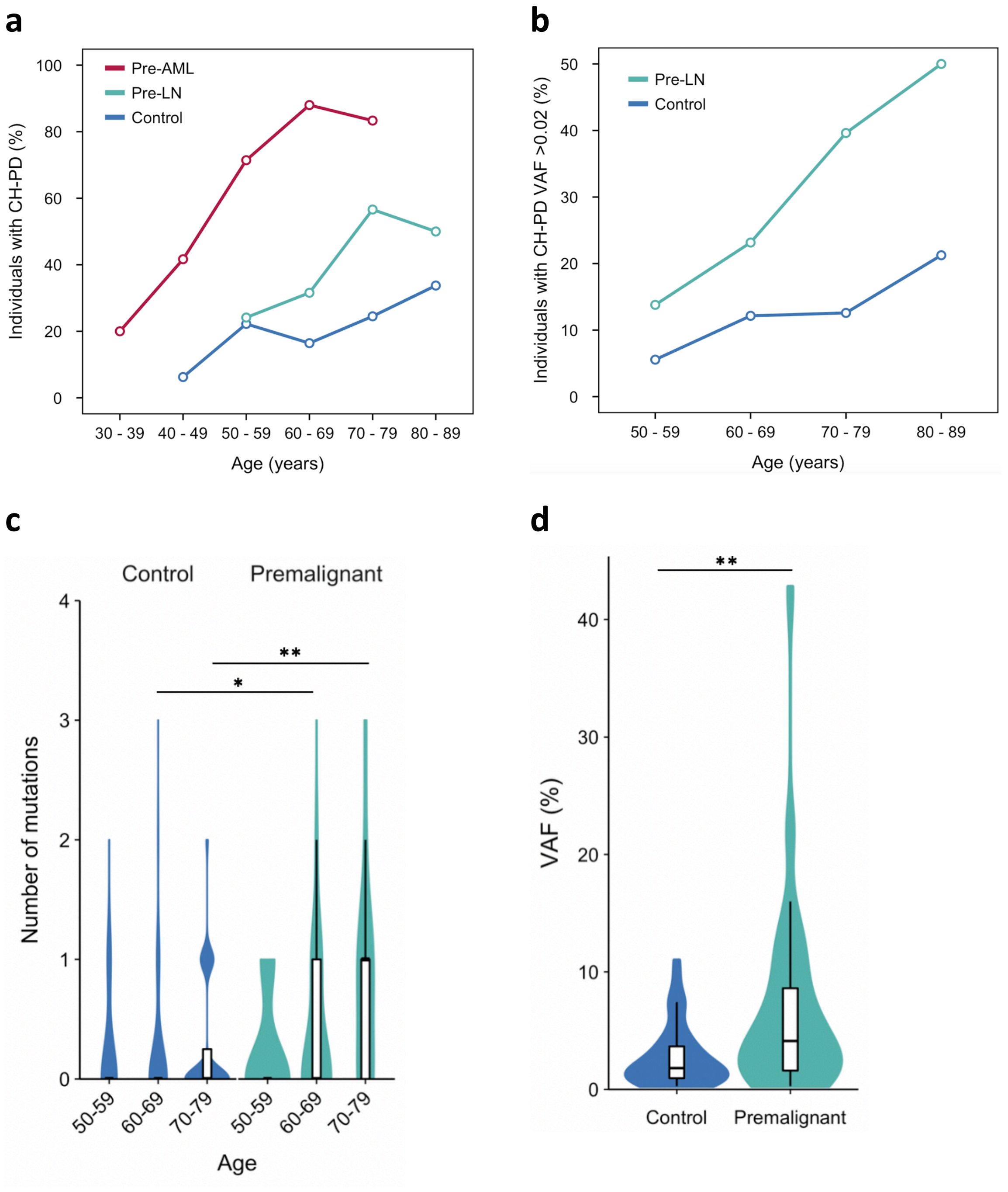


Figure 4.1 | Prevalence of CH-PD, number of mutations and clone size in pre-LN and control cohorts. a, Prevalence of CH-PD among pre-LN cases (green), controls (blue) and pre-AML (red; data from chapter 3). **b,** Prevalence of CH-PD clones with VAF > 2% among pre-LN cases (green) and controls (blue) is shown to put the data in the context of the historical definition of 'clonal haematopoiesis of indeterminate potential' (CHIP). **c**, The number of CH-PD mutations detected in pre-LN cases and controls according to age. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. **d**, VAF of CH-PD mutations in pre-LN cases (green) and controls (blue). * indicates P < 0.1; ** indicates P<0.001, two-sided Wilcoxon rank-sum test with Benjamini-Hochberg multiple testing correction.

2.3 Mutational spectrum of CH-PD in individuals who later developed a lymphoid neoplasm

Among the 189 discovery and validation cohort controls, the top three most frequently mutated genes were *DNMT3A, TET2* and *ASXL1* (Figure 4.2a-c, Appendix 14), consistent with the findings of other studies of CH-PD in the general population (Bowman et al., 2018). By contrast, among individuals who later developed a lymphoid blood cancer, the most recurrently mutated genes were *DNMT3A* (16.4% of cases versus 14.4% of controls), *TET2* (6.9% of cases vs 2.7% of controls), *ATM* (2.7% of cases vs 0.53% of controls) and *TP53* (2.7% of cases and 1.1% of controls). Among the genes recurrently mutated in both cases and controls, the mean mutation VAF was consistently higher in cases, though this difference only reached statistical significance on an individual gene level for *DNMT3A* (mean VAF in cases and controls 5.9% and 2.8%, respectively; $P = 0.029$, two-sided Wilcoxon rank-sum test with BH multiple testing correction). Furthermore, CH-PD in the pre-LN cases demonstrated a remarkably diverse spectrum of mutations, with putative driver variants identified in a total of 24 genes, compared to 11 genes among the controls (Figure 4.2a,b). Although there is broad overlap between the cancer genes implicated in myeloid and lymphoid malignancies (Arber et al., 2016; Sabarinathan et al., 2017; Swerdlow et al., 2016), several of the genes mutated among the cases are predominantly implicated in the latter, including *POT1*, *XPO1*, *HIST1H1E*, *NOTCH1*, *NOTCH2*, *ATM* and *CCND3* (Arber et al., 2016; Hing et al., 2016; Lunning and Green, 2015; Sabarinathan et al., 2017; Swerdlow et al., 2016).

Although data were too sparse to discern significant changes in the mutational spectrum with age, it is noteworthy that mutations in spliceosome genes (*SF3B1*, *SRSF2* and *U2AF1*) were only observed in controls over the age of 70, consistent with previous studies strongly associating these mutations with CH-PD in older individuals (Figure 4.2d)(McKerrell et al., 2015). Among the cases, the splicing gene mutation with the highest VAF (*SF3B1* p. K700E, VAF 2.1%) occurred in a 54-year-old man (PD00315) sampled 8 years before diagnosis with chronic lymphocytic leukaemia (CLL).

Figure 4.2

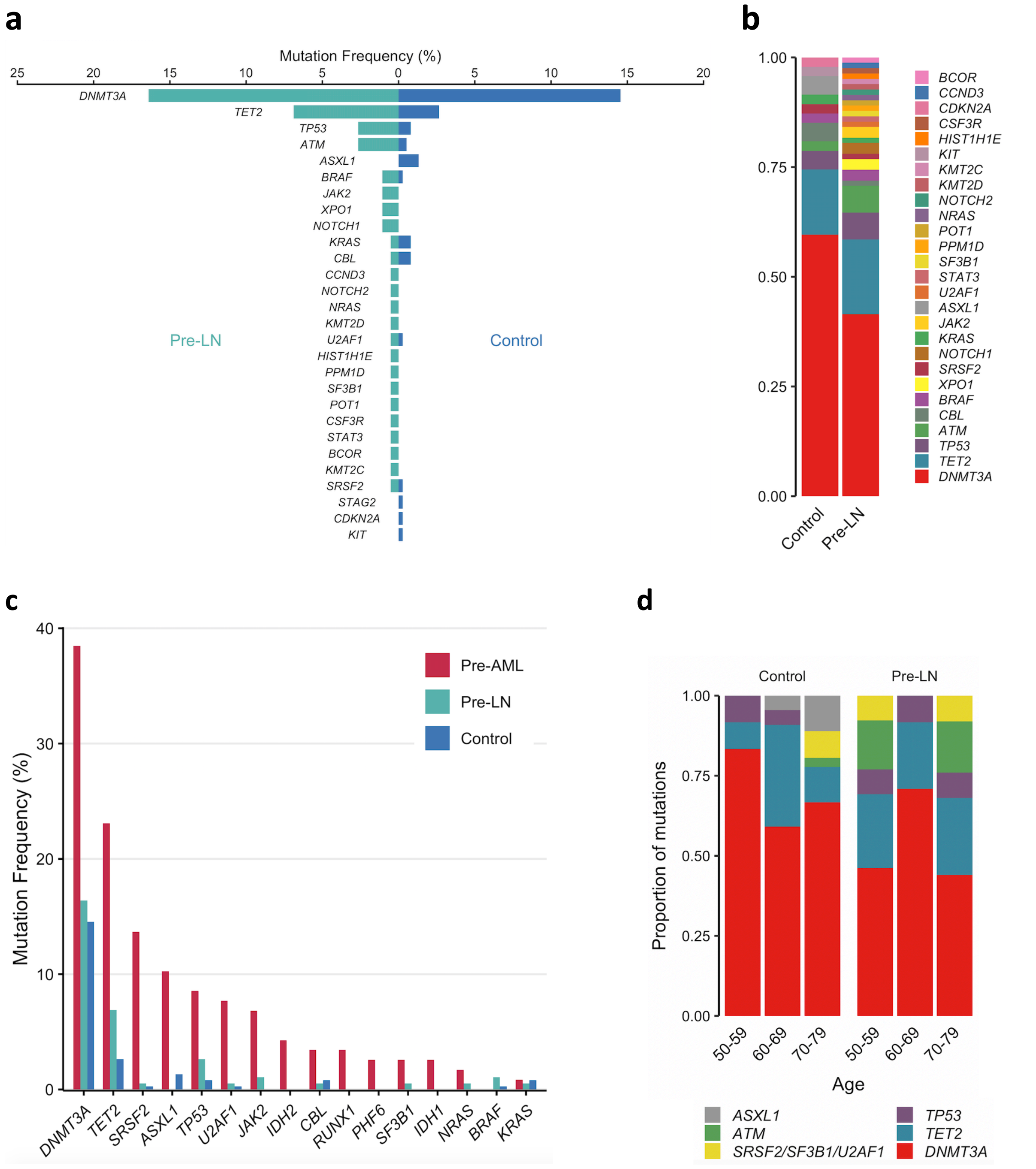


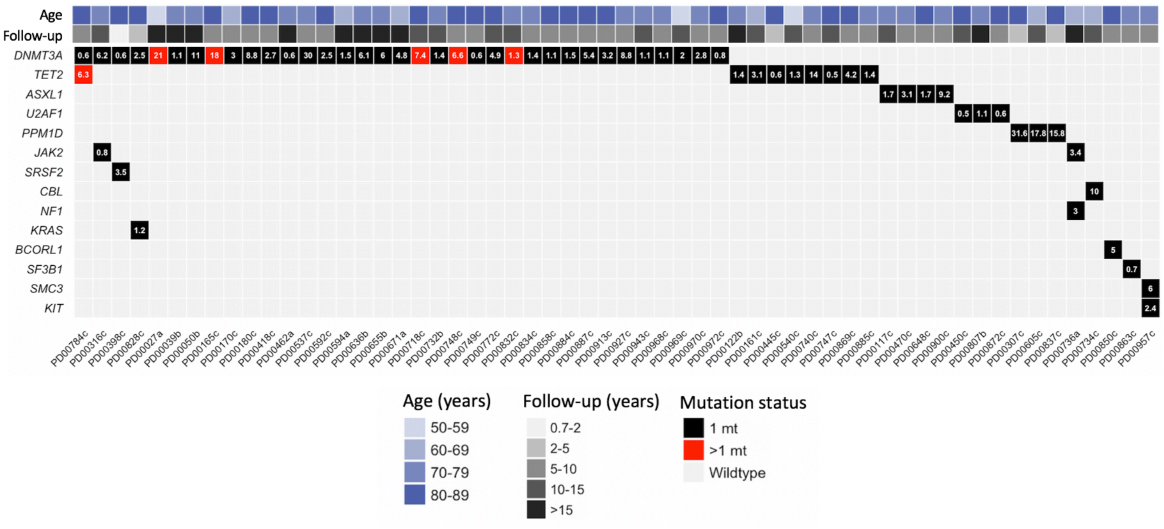
Figure 4.2 | The mutational spectrum of clonal haematopoiesis in individuals who developed a lymphoid neoplasm years later versus controls. a, Proportion of pre-LN cases (green) and controls (blue) with driver mutations each given gene. **b**, Relative frequency of mutations in the indicated genes according to age group for pre-LN cases and controls. **c**, Proportion of pre-AML (red), pre-LN (green) and control (blue) individuals with driver mutations in genes sequenced for both the pre-AML (chapter 3) and pre-LN cohorts. **d,** Relative frequency of mutations in the indicated genes according to age group for pre-LN cases and controls; only genes mutated at least 5 times included in panel, with spliceosome genes *SRSF2, SF3B1* and *U2AF1* aggregated.

2.4 Mutational spectrum in an extension cohort of older individuals with no record of cancer or a blood disorder

The more diverse genetic landscape of CH-PD in the pre-LN cases is intriguing, though the limited sample sizes and 1:1 case:control ratio warrant cautious interpretation. Although collectively a significant proportion of the mutations observed in the pre-LN cases occur in genes never or rarely reported in CH-PD in the general population, individual genes were infrequently mutated. Hence, despite the notable differences in mutational spectra between pre-LN cases and controls, considering all genes mutated more than 5 times across both cohorts on an individual basis, only *TET2* mutations approached significance for enrichment among the pre-LN cases (6.9% vs 2.7% mutated) (P = 0.05, one-sided Fisher's exact test with BH multiple testing correction). Is the absence of recurrent LN-drivers in the 189 age-and sex-matched controls included in the discovery and validation cohorts truly representative of the frequency of such mutations in the general ageing population? As mentioned in the introduction, most of the sensitive targeted surveys of CH-PD have used gene panels restricted to the most recurrent CH-PD driver genes and have not included the aforementioned LN-associated cancer genes (Acuna-Hidalgo et al., 2017; Coombs et al., 2017; Gibson et al., 2017; McKerrell et al., 2017; McKerrell et al., 2015; Young et al., 2016). The cumulative incidence of both common adult lymphoid malignancies and of CH-PD increases dramatically with age (Howlader et al., 2011), and it is conceivable that a more diverse CH-PD genetic landscape enriched for recurrent LN drivers emerges at higher rates in older age groups, analogous to the trend observed for spliceosome gene mutations (McKerrell et al., 2015). To investigate this possibility, we sequenced an extension cohort of 234 individuals (n=238 samples) with no record of any prior or subsequent cancer diagnosis or known blood disorder. The mean age at blood sampling was 74.4 years (IQR 67.5-81.6), more than ten years older on average than the control cohort. The mean follow-up was 11.9 years (IQR 8.0-16.4). Out of the 234 individuals, 58 (24.8%) had CH-PD (Appendix 14). Despite high coverage (median >5,000X) and sensitivity to detect small clones down to VAF 0.5%, the genetic landscape was consistent with that observed in previous studies of CH-PD in the general ageing population. In particular, no canonical drivers associated with lymphoid malignancies were identified (Figure 4.3a,b), in contrast to the pre-LN cohort.

Figure 4.3

a



b

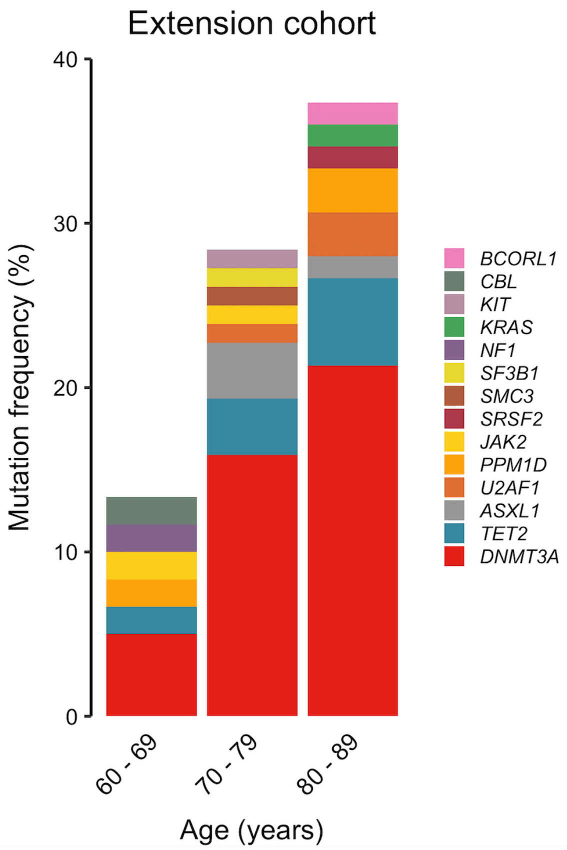


Figure 4.3 | The mutational spectrum of clonal haematopoiesis in an extension control cohort of older individuals with no history of cancer or haematological disorder. a, Co-mutation plot including only individuals with CH-PD (58 out of 234 individuals in the older extension cohort). The top two rows indicate age at sampling and follow-up period in years. Tiles are coloured according to mutation status for each given gene and number of drivers identified: pale grey, wild type; black, one driver mutation; red, two driver mutations. The mutation VAF (%) is indicated in white text within each tile. Where two mutations were identified in a given gene and sample (red tiles), the highest VAF is shown. **b,** Proportion of individuals with driver mutations in each given gene according to age group.

2.5 Clonal dynamics over time and relationship with future lymphoid neoplasm

Examining co-mutation patterns in those with a future LN diagnosis (Figure 4.4a-b) invites some initial speculation regarding the relationship between CH-PD and future LN. For many cases, the only CH-PD mutations detected occur in genes that are seldom implicated as drivers in the lymphoid cancer type diagnosed years later. The most notable example is *DNMT3A*, the most frequently mutated gene among both cases and controls (Figure 4.2a). Although *DNMT3A* does play a role in some lymphoid malignancies, particularly T-cell leukaemia/lymphoma (Couronne et al., 2012; Haney et al., 2016a; Haney et al., 2016b), it is not among the most recurrently mutated genes in these disorders (Brunetti et al., 2017; Sabarinathan et al., 2017). By contrast, the *BRAF* p.V600E, *POT1* p.K90E and *XPO1* p.E571 hotspot mutations preceding diagnoses of hairy cell leukaemia (HCL), small cell B-cell lymphoma (SLL) and CLL, respectively, are highly plausible drivers of the respective latent malignancies, but are rarely if ever associated with CH in the general population (Landau et al., 2015; Pinzaru et al., 2016; Tiacci et al., 2011).

In order to further investigate the relationship between CH-PD detected years before LN diagnosis and the future malignancy, serial peripheral blood DNA samples were sequenced from 104 individuals, including 69 pre-LN cases and 35 controls. The mean interval between earliest and latest sample was 7.3 years. No diagnostic specimens were available; however, for 16 of the pre-LN cases, at least one peripheral blood sample taken less than 6 months before diagnosis (n = 5 individuals) or after diagnosis (n = 11 individuals) was sequenced.

Of the 69 serially sampled pre-LN cases, 22 had at least one driver detected in an earlier time point sample. Out of the 26 distinct mutations identified, 25 persisted in the later sample and 1 became undetectable. The only non-persistent clone harboured a *KRAS* p.G13D mutation present at 1% VAF in PD00003 at age 62.4 and no longer detectable in a sample taken 8.5 years later. Among the 35 controls with serial samples, 7 had mutations detected in their earlier samples. Of the 10 distinct variants, 5 persisted and 5 were no longer detected in the subsequent sample. The latter group comprised low VAF mutations in *DNMT3A* (n=4) and *KRAS* (n=1). Consistent with the patterns seen in pre-AML cases and controls, examining the VAF trajectories of the persistent mutations over time demonstrated variable behaviour, including for clones with mutations in the same gene (Figure 4.5a,b). However, the numbers of cases and controls with mutations were insufficient to infer any significant overall difference in clonal growth rates between pre-LN cases and controls.

Examining the sequence of mutation acquisition and VAF trajectories among the pre-LN cases revealed several notable findings (Figure 4.6a-k). Among the 16 pre-LN cases with peri- or post-diagnosis samples available, 7 harboured antecedent CH-PD. All 7 individuals harboured at least one driver in *DNMT3A* (Figure 4.6a-g), all of which persisted across serial samples. In 4/7 cases, the size of the *DNMT3A* clone(s) diminished over time (Figure 4.6a,d,f,g), and in 2 of these cases this decline coincided with late acquisition of at least one driver mutation in a canonical lymphoid cancer gene, specifically *CCND3* and *CREBBP* in an NHL and *SF3B1* in a CLL case (Figure 4.6d,g)(Chapuy et al., 2018; Lunning and Green, 2015; Mullighan, 2014; Okosun et al., 2014; Sabarinathan et al., 2017). The same phenomenon is observed in two other cases, with the appearance of a relatively LN-specific driver mutation (e.g., in *NOTCH1, POT1* and *HIST1H1E*)(Sabarinathan et al., 2017; Swerdlow et al., 2016) years before diagnosis also coinciding with stable or falling VAF of mutations in the canonical CH/myeloid neoplasm drivers *DNMT3A* and *U2AF1*, respectively (Figure 4.6i,j). These observations strongly suggest the presence of distinct, potentially competing clones and supports the hypothesis that a significant proportion of the CH-PD in the pre-LN cases is not phylogenetically related to the future malignancy, despite large clone sizes in most instances. Four serially-sampled pre-LN cases harboured drivers in genes more frequently mutated in LN than in CH-PD, namely *CCND3, ATM, BRAF* and *TP53* (Figure 4.4a), and in each of these cases VAF increased over time. Hence, despite limited data, this time series experiment suggests that CH-PD in pre-LN cases represents a combination of pre-malignant clones and 'bystander' clones, analogous to the situation observed in pre-AML.

Figure 4.4

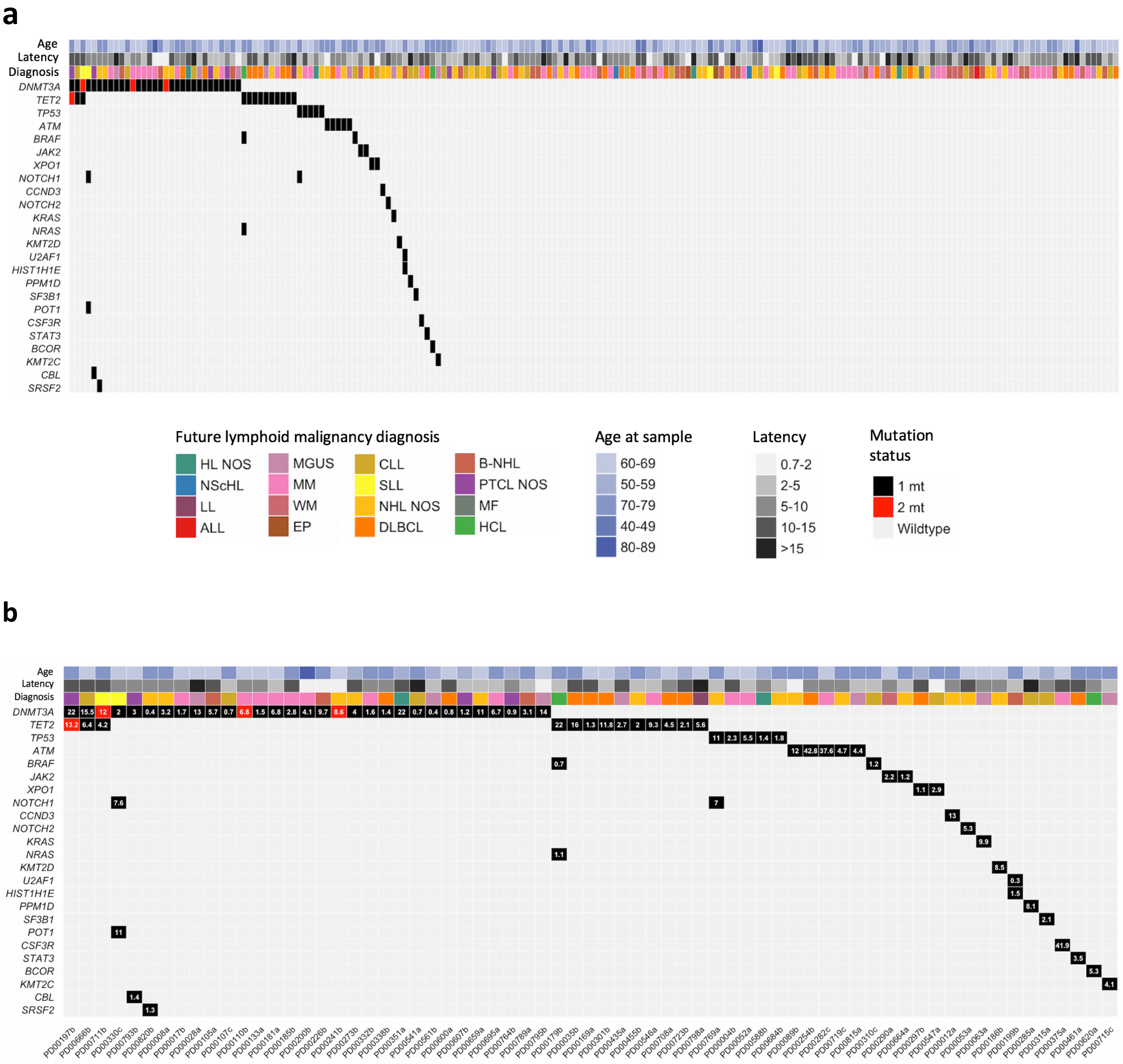


Figure 4.4 | Mutation co-occurrence in pre-LN cases according to diagnosis, latency and age at sampling. a, Co-mutation plot for all 189 pre-LN cases. Top three rows indicate age at sampling, latency and sample and future LN diagnosis. Tiles are coloured according to mutation status for each given gene and number of drivers identified: pale grey, wild type; black, one driver mutation; red, two driver mutations. **b**, Co-mutation plot including only cases with CH-PD. The mutation VAF percentage is indicated in white text within each tile. Where two mutations were identified in a given gene and sample (red tiles), the highest VAF is shown. MM, multiple myeloma; NHL NOS, non-Hodgkin lymphoma not otherwise specified; MGUS, monoclonal gammopathy of undetermined significance; DLBCL, diffuse large B-cell lymphoma; B-NHL, B-cell non-Hodgkin lymphoma; CLL, chronic lymphocytic leukemia; HCL, hairy-cell leukemia; PTCL NOS, peripheral T-cell lymphoma NOS; WM, Waldenstrom macroglobulinaemia; SLL, small cell B-cell lymphoma; HL, Hodgkin lymphoma; LL, lymphoblastic lymphoma; NScHL, nodular sclerosis Hodgkin lymphoma; EP, extramedullary plasmacytoma; MF, mycosis fungoides; ALL, acute lymphoblastic leukemia.

Figure 4.5

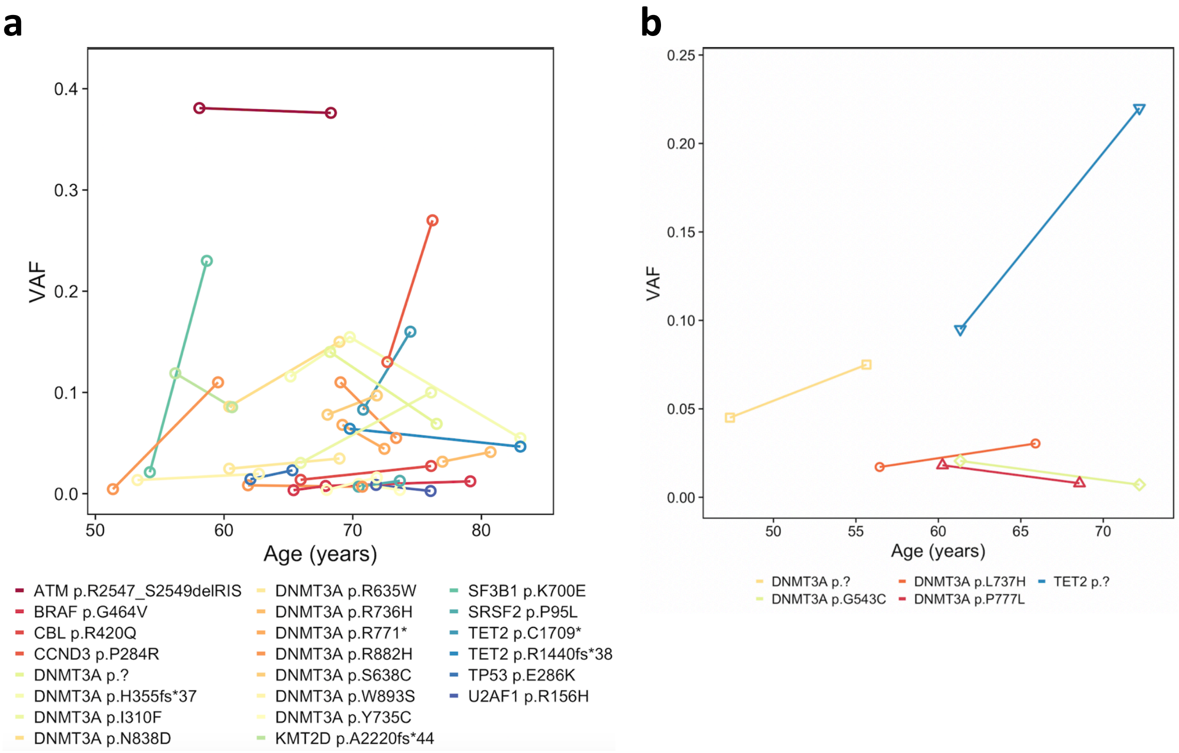


Figure 4.5 | VAF trajectories of persistent mutations in serially sampled pre-LN cases and controls. **a-b,** VAF trajectories of CH-PD driver mutations persisting across serial samples from cases sampled years before diagnosis of a lymphoid neoplasm **(a)** and controls **(b)**. X-axis denotes age at sampling and y-axis mutation VAF.

Figure 4.6

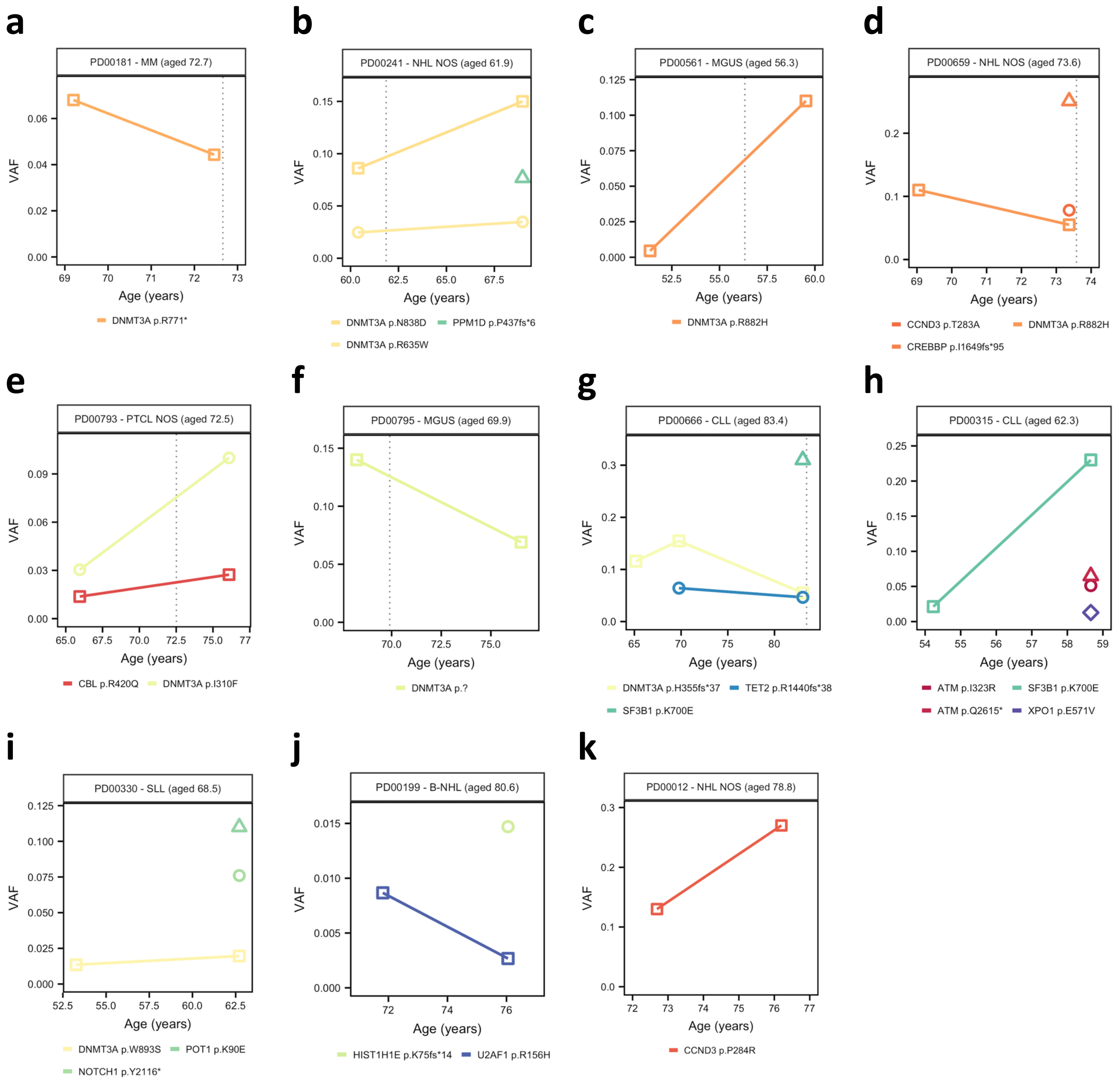


Figure 4.6 | Evolution of clonal haematopoiesis and relationship with future lymphoid neoplasm. a-h, VAF trajectories of putative driver mutations in 7 individuals for whom peripheral blood taken near or after cancer diagnosis was available for sequencing. Future LN diagnosis and age at diagnosis are indicated in parentheses above the plot. Vertical dotted lines demarcate pre- and post-diagnosis periods. **i-k,** VAF trajectories of putative driver mutations in an additional 5 cases sampled multiple times years before cancer diagnosis. Age at sampling and mutation VAF are shown on the x- and y-axis, respectively. LN, lymphoid neoplasm; VAF, variant allele fraction; MM, multiple myeloma; NHL NOS, non-Hodgkin lymphoma not otherwise specified; MGUS, monoclonal gammopathy of undetermined significance; B-NHL, B-cell non-Hodgkin lymphoma; CLL, chronic lymphocytic leukemia; PTCL NOS, peripheral T-cell lymphoma NOS; SLL, small cell B-cell lymphoma.

2.6 Clinical factors associated with future development of a lymphoid malignancy

Full blood count parameters, lipid profile, C-reactive protein, blood pressure and anthropomorphic measurements were available for most of the pre-LN cases and controls (Figure 4.7). The case:control ratio for this analysis was 1:2 due to inclusion of 189 age-and sex-matched controls from the validation cohort described in Chapter 3. Consistent with the observations in the pre-AML cases and controls and previous studies of CH-PD (Jaiswal et al., 2014; McKerrell and Vassiliou, 2015), blood counts did not differ significantly between pre-malignant cases and controls or between individuals with and without CH-PD (Figure 4.7). Assessing all clinical parameters available for the majority of pre-LN cases and controls revealed significantly lower levels of high-density lipoprotein (HDL) in pre-LN cases (*P*=0.048, two-sided Wilcoxon rank-sum test with BH multiple testing correction). No other trends in clinical variables remained significant after multiple testing correction. There were no significant differences in clinical parameters when only cases and controls with CH-PD were compared to each other or when all individuals (cases and controls) with CH-PD were compared to individuals with no detectable mutations. Kaplan-Meier analysis of the impact of clinical variables on LN-free survival showed trends towards shorter time to cancer progression with higher RDW, though this correlation did not reach significance (Figure 4.8).

Figure 4.7

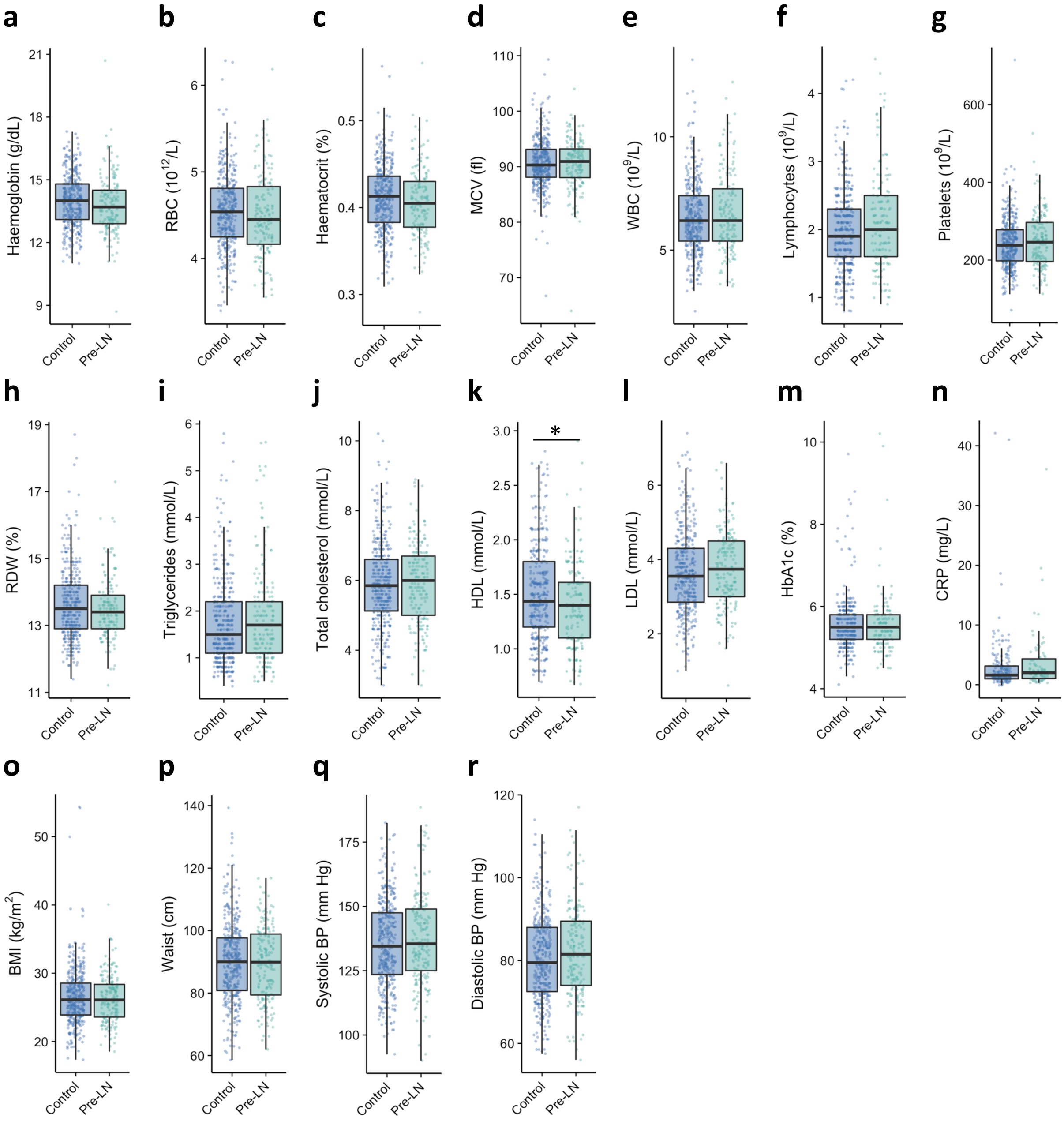


Figure 4.7 | Full blood count and metabolic parameters in pre-LN cases and controls. Box plots of full blood count parameters **(a-h)**, biochemistry measurements **(i-n)**, body mass index **(o)**, waist circumference **(p),** and blood pressure **(q-r)** available for a subset of cases and pre-LN controls. Boxplot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. RBC, red blood cell; MCV, mean corpuscular volume; WBC, white blood cell; RDW, red cell distribution width; HDL, high density lipoprotein; LDL, low density lipoprotein; HbA1c, haemoglobin A1c; CRP, C-reactive protein; BMI, body mass index; BP, blood pressure. * *P*=0.048, two-sided Wilcoxon rank-sum test with BH multiple testing correction

Figure 4.8

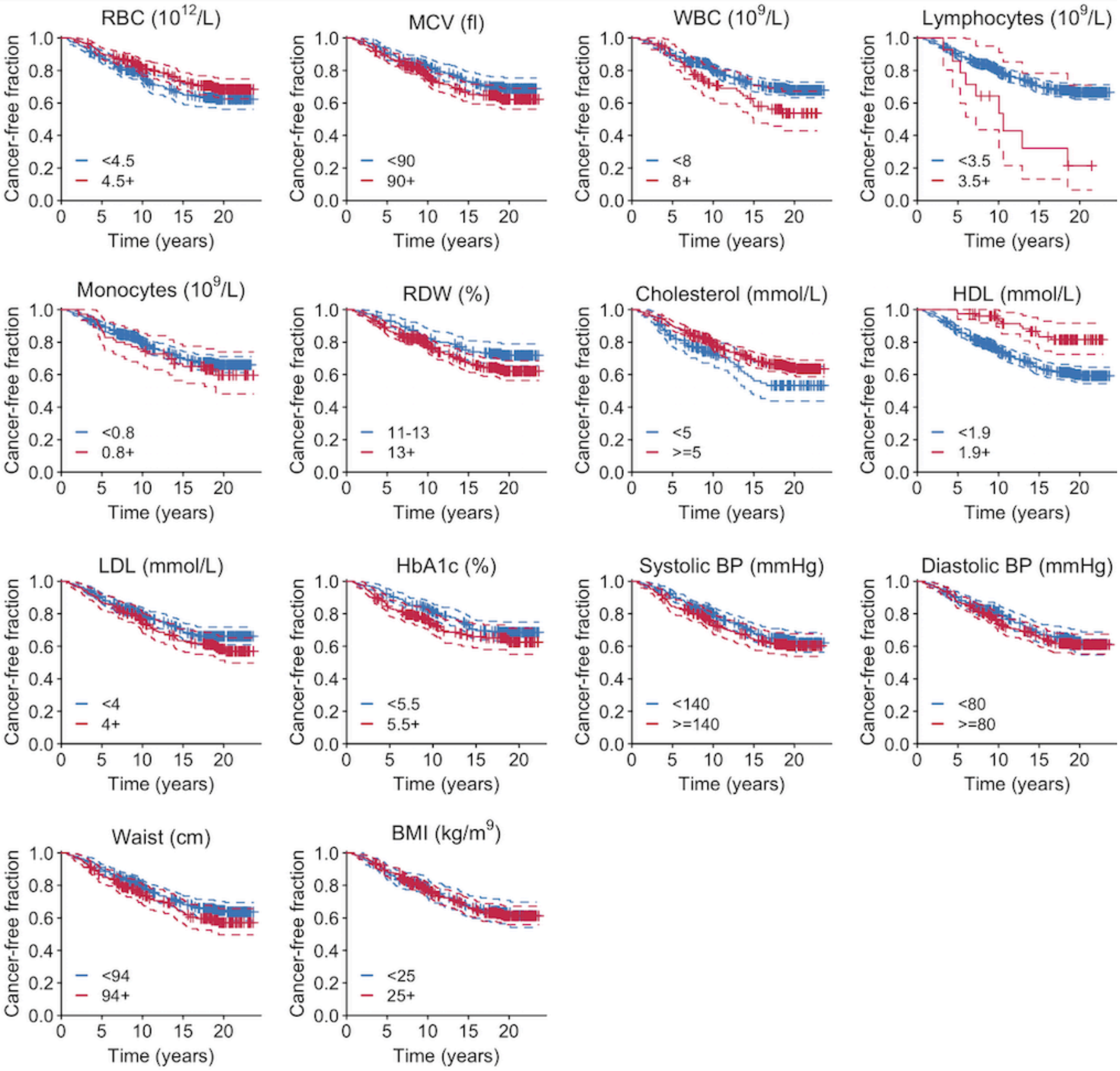


Figure 4.8 | Impact of clinical variables on lymphoid neoplasm-free survival. Kaplan–Meier curves of LN-free survival, defined as the time between sample collection and LN diagnosis, death or last follow-up. Survival curves are stratified according to cutoffs indicated in the lower left corner of each plot. *n* = 567 unique individuals, including 189 pre-LN cases and 378 age- and sex-matched controls. 95% confidence intervals indicated by dashed lines. RBC, red blood cell; MCV, mean corpuscular volume; WBC, white blood cell; RDW, red cell distribution width; HDL, high density lipoprotein; LDL, low density lipoprotein; HbA1c, haemoglobin A1c; BMI, body mass index; BP, blood pressure.

2.7 Predicting progression to lymphoid malignancy

On the basis of these findings, an approach similar to that described in Chapter 3 was developed to quantify the relative contributions of driver mutations, clone sizes and clinical factors to the risk of progressing to a lymphoid malignancy. In keeping with results from Chapter 3, Kaplan-Meier analysis of the impact of the number of drivers and mutation VAF demonstrated consistent correlation between mutation burden and progression-free survival, though these trends did not reach significance (Figure 4.9a). This correlation held even when the additional set of controls was incorporated and analysis was restricted to genes included in the myeloid panel used in Chapter 3 (Figure 4.9b). Although the relative infrequency of CH among pre-LN cases limited the power of KM analysis, a trend towards shorter LN-free survival was observed with larger *DNMT3A* clones (Figure 4.9c) or the presence of mutations in any of the LN-associated genes *XPO1, POT1, CCND3, HIST1H1E, NOTCH1* or *NOTCH2* (Figure 4.9d). KM curves for individual genes are shown in Figure 4.10.

Figure 4.9

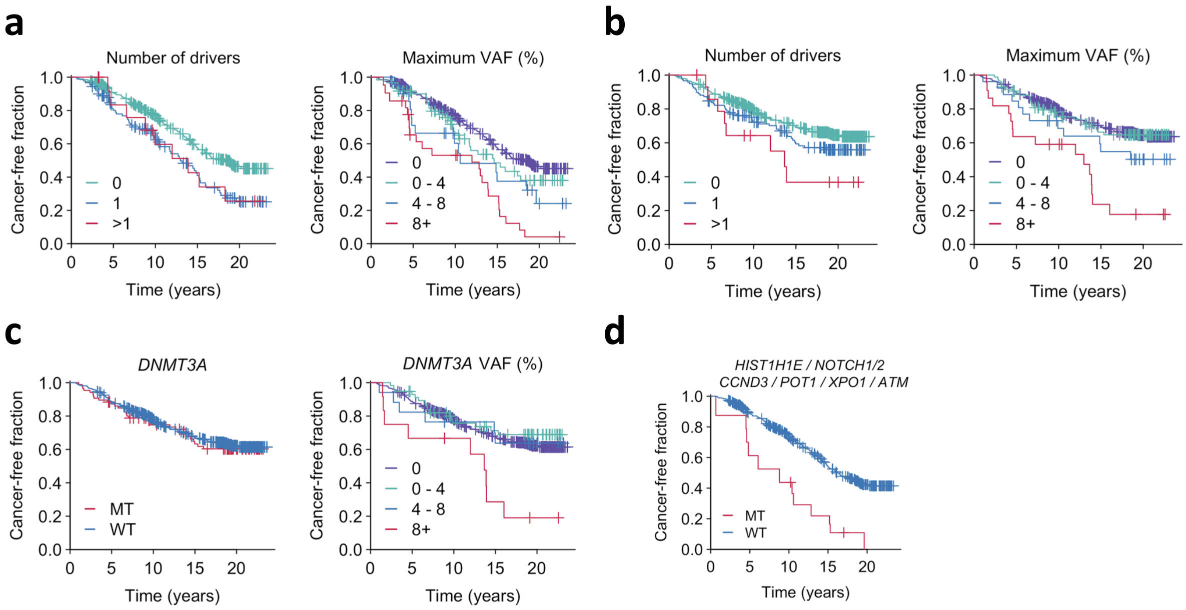


Figure 4.9 | Impact of mutation burden on lymphoid neoplasm-free survival. a,b Kaplan–Meier (KM) curves of LN-free survival, defined as the time between sample collection and LN diagnosis, death or last follow-up. Survival curves are stratified according to number of driver mutations per individual and largest clone detected. Panel **(a)** includes all genes sequenced across the 189 pre-LN cases and 189 age- and sex-matched controls. The same trends, albeit not reaching significance, persist when only mutations in genes sequenced by the myeloid panel are included in the analysis (189 pre-LN cases and 378 controls) **(b)**. **c,** KM curves of LN-free survival stratified by *DNMT3A* mutation status and VAF of *DNMT3A* mutations. **d,** KM curve of LN-free survival stratified according to mutation status in any of six infrequently mutated lymphoid neoplasm-associated driver genes. VAF, variant allele fraction.

Figure 4.10

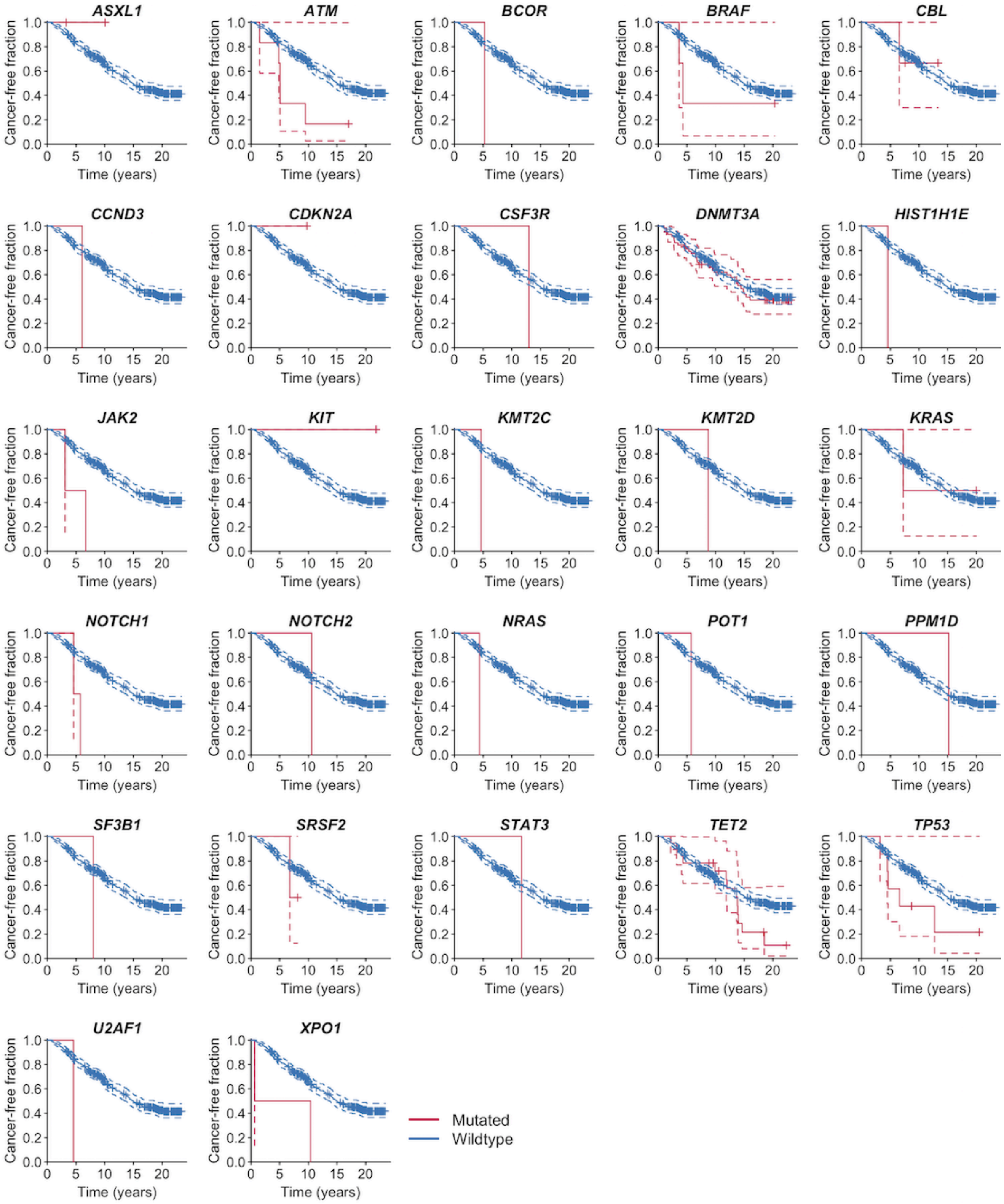


Figure 4.10 | Gene-level impact on LN-free survival. Kaplan–Meier (KM) curves of LN-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status. *n* = 378 unique individuals (189 pre-LN cases and 189 controls). LN, lymphoid neoplasm; VAF, variant allele fraction. Dashed lines indicate 95% confidence intervals.

However, the high proportion of infrequently mutated genes dominating the genetic landscape of CH-PD among pre-LN cases and lower prevalence of CH-PD among pre-LN relative to pre-AML hindered robust identification of gene-level risk factors for malignant progression. Regularised logistic and Cox proportional hazards regression approaches were applied as described in Chapter 3 (see Methods section 4). Excluding infrequently mutated genes from model training eliminated a significant proportion of CH-PD mutations from analysis and yielded fairly homogenous gene-level hazard ratios with wide confidence intervals for most genes (Figure 4.11). Notable exceptions were *DNMT3A* and *TET2*, which were the most recurrently mutated genes across both cohorts and were thus amenable to more accurate analysis of the mutation contribution to LN progression risk (Figure 4.11a). Quantitatively, driver mutations in *DNMT3A* and *TET2* conferred a 1.5 to twofold increased 10-year risk of LN per 5% increase in clone size (Figure 4.11a and Appendix 15). Remarkably, these hazard ratios are virtually identical to the effect sizes observed for these genes in the AML prediction model (Figure 3.5). In order to achieve more accurate estimates of HRs for clinical variables and the subset of genes sequenced across both gene panels, the model was retrained using an additional set of 189 controls sequenced with the myeloid panel used in Chapter 3 for a case:control ratio of 1:2. The genes analysed were restricted to those overlapping between both panels and mutated at least twice in either discovery or validation cohort. Hazard ratios for overlapping variables were concordant, albeit with narrower confidence intervals (Figure 4.11b).

Figure 4.11

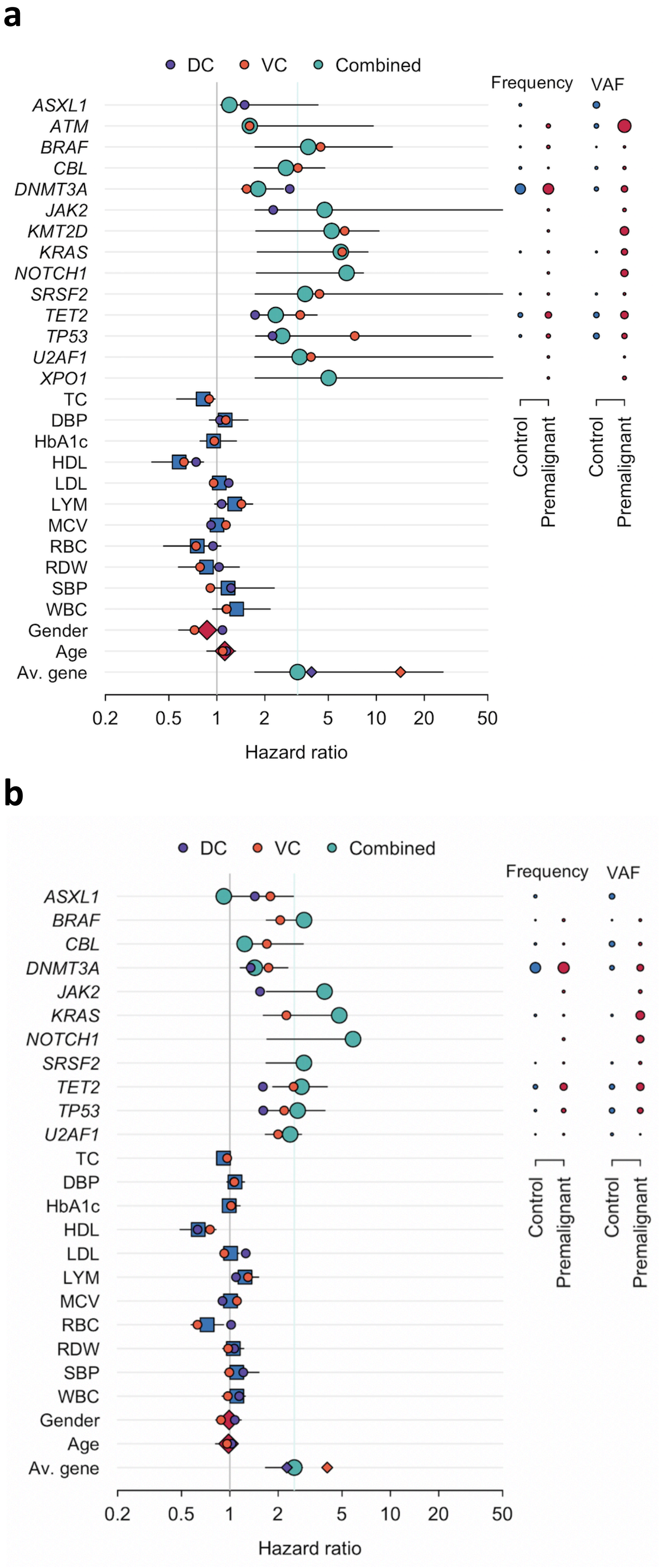


Figure 4.11 | Forest plots of hazard ratios for risk progression to lymphoid malignancy. a, Forest plot for Cox proportional hazards model using a 1:1 case control ratio and including all myeloid and lymphoid cancer genes. **b,** Model restricted to myeloid panel genes and incorporating an additional 189 age-and sex-matched controls for a 1:2 case:control ratio and hence more accurate estimates of risk associated with clinical factors and genes sequenced across both panels. Purple, orange and green circles indicate hazard ratios (HR) for the discovery (DC), validation (VC) and combined cohort, respectively. Horizontal lines denote 95% confidence intervals for the combined cohort. For each gene, the indicated HR applies to the 10-year risk of lymphoid blood cancer conferred by each 5% increase in mutation VAF. The green vertical line indicates the mean HR across all genes. Blue (controls) and red (pre-LN) circles to the right of the forest plot indicate the proportion of individuals with mutations in each gene and the average mutation VAF, which aids in the interpretation of hazard ratios. For example, *ATM*, a recurrent driver gene in several lymphoid malignancies, is almost exclusively mutated in pre-LN cases but at relatively high VAF, which translates into a modest HR for each 5% increase in clone size.

Overall, genetic and clinical parameters explained approximately 45% and 12% of the absolute variance in LN-free survival between individuals, respectively. Notably, clinical factors explained a comparable proportion of the variance. The coefficients for clinical variables were consistent between models trained on the discovery and validation cohorts (Figure 4.11a,b). Interestingly, lower HDL was associated with a modest but significant increase in risk of LN progression (Figure 4.11a,b). Consistent with this finding, lower total cholesterol was also associated with a smaller but still significantly increase in risk.

Unsurprisingly, models did not achieve anywhere near the predictive power observed for AML, with concordance and AUC both ≤0.7 for models trained on either cohort (Table 4.2). Nevertheless, this analysis yielded robust estimates of the risk conferred by lower HDL levels and mutations in *DNMT3A* and *TET2*, findings with compelling biological implications that warrant further investigation.

Table 4.2 Cox proportional hazard model performance

Cox proportional hazards model	Concordance	Standard error	Time-dependent AUC
VC data and fit	0.60	0.035	0.67
DC data and fit	0.70	0.029	0.64
VC fit DC data	0.58	0.035	0.60
DC fit VC data	0.60	0.027	0.67
Combined cohorts	0.67	0.022	0.67

*Derived from 100 bootstraps out-of-bag validation
DC, discovery cohort; VC, validation cohort

3. Discussion

The main aim of this experiment was to characterise the prevalence and genetic landscape of CH-PD in individuals who go on to develop a lymphoid neoplasm. To this end, I have deep sequenced peripheral blood specimens from 189 pre-LN cases and 189 age- and sex-matched controls using a much broader gene panel than has been applied in previous similarly sensitive assays for CH. To investigate potential enrichment for LN-associated mutations in older age, this study was extended to include samples from a further 234 healthy older individuals. Serial samples, including peri- and post diagnosis blood samples, provided insight into clonal dynamics and the relationship between CH-PD and future malignancy. Clinical metadata, including full blood count parameters and lipid profile, were analysed for any association with CH-PD or future LN risk. Genetic and clinical variables were then incorporated into predictive models to seek any significant risk factors for LN progression and assess their collective power to identify individuals at high risk of future LN development.

3.1 CH-PD frequently precedes LN diagnosis and is characterised by a diverse mutational spectrum

This work demonstrates that CH-PD becomes more prevalent among individuals who develop a lymphoid malignancy years before diagnosis and is characterised by a more diverse genetic landscape than that observed in pre-AML cases or in the general population. The experiment described in Chapter 3 demonstrated that pre-AML exhibits a mutational spectrum that closely overlaps with that seen in the general population but is enriched for mutations in particular genes. By contrast, the pre-LN cohort harboured rare events in a number of genes highly associated with LN pathogenesis and rarely if ever reported in the current CH literature, including *ATM, CCND3, POT1, HIST1H1E, XPO1, NOTCH1* and *NOTCH2* (Arber et al., 2016; Kandoth et al., 2013; Martincorena et al., 2017; Sabarinathan et al., 2017; Swerdlow et al., 2016). Among these, *ATM* was the most recurrently mutated in pre-LN cases, ranking third after *DNMT3A* and *TET2*. The genetic heterogeneity observed in the pre-LN cohort is reminiscent of the genomic landscapes of the most common lymphoid blood cancers in adults, which tend to be characterised by a large number of infrequently mutated putative cancer genes (Landau and Wu, 2013; Reddy et al., 2017; Sabarinathan et al., 2017; Swerdlow et al., 2016).

3.2 CH-PD as a biomarker for lymphoid blood cancer risk

Despite an overall more varied mutational spectrum in pre-LN CH-PD, the two top genes remained *DNMT3A* and *TET2*. Mutations in both of these genes, and in particular *TET2,* are implicated in both B- and T-cell lymphoid malignancies (Couronne et al., 2012; Dominguez et al., 2018; Haney et al., 2016a; Haney et al., 2016b; Mouly et al., 2018; Quivoron et al., 2011). TET2 deficiency in particular has been shown to increase HSC mutation rate and predispose to lymphoid and myeloid malignancies (Pan et al., 2017). However, the high frequency of *TET2/DNMT3A* mutations in in pre-LN CH-PD relative to lymphoid cancers, in conjunction with the results of the time series experiment, suggests that *DNMT3A*-mutated clones in particular often do not represent ancestors of the future cancer. Nevertheless, mutations in *DNMT3A* and *TET2* confer a significantly increased risk for progression to LN, with hazard ratios comparable to those observed in the AML prediction model (Figure 3.5 and Figure 4.11). Although speculative, there are several possible explanations for this observation. As alluded to in Chapters 1 and 3, it is possible that clones that are not phylogenetically related to the future malignancy are surrogate markers of selective pressures that impart a strong growth advantage on pre-malignant HSCs. There is increasing precedent for this hypothesis in the haematopoietic system and other tissues. For example, as discussed in depth in Chapter 5, activating mutations in *PPM1D*, a negative regulator of TP53, confer a selective advantage on HSCs in the context of cytotoxic therapy (Gibson et al., 2017; Hsu et al., 2018; Takahashi et al., 2017). *PPM1D*-mutated CH-PD is a biomarker of therapy-related AML risk, despite that the *PPM1D*-mutations often persist at low VAF alongside the evolving AML (Gibson et al., 2017; Gillis et al., 2017). Remarkably, a similar scenario has recently been described in oesophageal epithelium, which is increasingly populated by *PPM1D*-mutated clonal expansions with age (Yokoyama et al., 2019). Exposure to alcohol and smoking, strong risk factors for oesophageal cancer, were associated with expansion of *PPM1D*-mutated epithelial clones, though *PPM1D* is not a recurrent driver in oesophageal malignancies (Yokoyama et al., 2019).

Current understanding of the selective pressures influencing somatic evolution in the haematopoietic system remains limited. However, age-related increases in endogenous genotoxic stress and reduced HSC self-renewal capacity may be important factors (Pang et al., 2017; Yahata et al., 2011). It is plausible that inter-individual variation in the pace and nature of age-related processes may influence the spectrum of mutations that confer selective advantage on HSCs. In this context it is noteworthy that *TP53* and *ATM*, both critical mediators of DNA damage response and cell cycle checkpoint control (Roos et al., 2016), constituted the third and fourth most frequently mutated genes in this pre-LN cohort. Whilst this result warrants confirmation in larger studies, it is conceivable that some individuals experience more severe/earlier DNA-damage associated HSC senescence and that this favours expansion of clones with mutations that repress DNA-damage-induced apoptosis and cell cycle arrest. By extension, such individuals would likely be at higher risk of stochastic driver mutation acquisition and clonal evolution of any one of numerous pre-malignant clones. As mentioned in Chapter 3, Wong et al. recently reported a high prevalence of 'bystander' pre-leukaemic clones in AML patients at diagnosis, suggesting that their leukaemia arose from one of many candidate pre-malignant HSCs (Wong et al., 2015a).

3.3 RDW and lymphoid neoplasm risk

Notably, RDW was not significantly increased among pre-LN cases, in contrast to the scenario observed for pre-AML. As discussed in Chapter 3, higher RDW has previously been associated with CH in the general population (Jaiswal et al., 2014). However, we have shown that comparing pre-AML cases and controls with CH-PD revealed that RDW could help distinguish pre-AML (including cases without detectable CH-PD) from CH in individuals who did not develop a blood cancer during follow-up. The association between higher RDW and risk of developing AML was validated in a large electronic medical records dataset. It is possible that a weaker correlation does exists between pre-LN and RDW that this study was underpowered to detect, as hinted by the subtle trend discernible on KM analysis (Figure 4.8). However, this result nevertheless suggests that RDW is not a universally strong discriminator between indolent and pre-malignant CH-PD. This experiment may mask lymphoid cancer subtype-specific associations between RDW and warrants further investigation.

3.4 Lower high-density lipoprotein levels and lymphoid cancer risk

Among all clinical variables analysed, only HDL levels differed significantly between pre-LN cases and controls. The association between lower HDL and future LN was corroborated by Cox proportional hazards modelling, which identified a modestly increased risk of LN with lower HDL and total cholesterol (Figure 4.11a,b). Hypocholesterolaemia is a common finding in lymphoma and leukaemia patients, and has also been reported in association with some solid tumour types (Lim et al., 2007; Pirro et al., 2018). Lower HDL in particular has been previously identified as a preclinical feature of non-Hodgkin lymphoma discernible years before diagnosis (Lim et al., 2007). Low HDL at lymphoma diagnosis has also been correlated with poorer prognosis (Matsuo et al., 2017). The mechanisms underlying these observations are unclear with no compelling evidence of a causative link between low cholesterol and haematological malignancies (Pirro et al., 2018). However, numerous studies report that lymphoma cells and leukaemia blasts have higher HDL and/or LDL uptake receptor activity (Goncalves et al., 2005; Vitols et al., 1990; Vitols et al., 1985) and that cholesterol metabolism may represent a viable therapeutic target for several mature B-cell malignancies (McMahon et al., 2017). It is therefore possible that pre-malignant CH displays similar behaviour, leading to reductions in circulating levels of HDL even years prior to overt malignant transformation. This is a particularly intriguing hypothesis in view of the emerging causal role of CH-PD in atherosclerosis (Fuster et al., 2017; Jaiswal et al., 2017; Sano et al., 2018a). It is even conceivable that plaque-resident clonal haematopoietic cells may accelerate atheroma progression in part by increasing lipid accumulation at sites of inflamed endothelium.

3.5 Experiment limitations and future directions

This experiment has several important limitations. Firstly, the pre-LN cohort encompasses diverse diseases presenting over a long period during which histopathological classification schemes and diagnostic guidelines evolved considerably (Campo et al., 2011; Swerdlow et al., 2016). This limited the scope to investigate the natural history of or distinct genetic/clinical risk factors for individual cancer types. Furthermore, structural events, particularly translocations involving the immunoglobulin heavy chain (IGH) genes and numerical chromosomal aberrations, are frequent initiating events of lymphoid malignancies and their detection requires a much broader and more costly sequencing approach (Bolli et al., 2014; Landau et al., 2015). While the main aim of this experiment was to characterise the point mutation spectrum of CH-PD in pre-LN and investigate the predictive value of both putative ancestral and 'bystander' clones in assessing risk of progression, the power of predictive models would likely be increased by screening for subclonal large copy number changes and recurrent translocations.

Moreover, these results provide further evidence that malignant and cardiovascular adverse outcomes associated with CH might be linked. The association of lower HDL with LN progression risk, in conjunction with the clinical AML prediction model described in Chapter 3, hint that there may be unifying features of 'high risk' CH that could eventually help define a useful biomarker and/or therapeutic target. Hence this experiment reinforces the need for future studies of CH to correlate genetics with detailed clinical and phenotypic metadata and to try to move beyond investigating malignant and cardiometabolic disease associations in isolation.

Chapter 5

Clonal haematopoiesis after childhood cancer treatment

1. Introduction

The findings of the preceding chapters demonstrate that pre-malignant CH is associated with clinical and genetic features that can help distinguish individuals at highest risk of developing certain blood cancers, particularly *de novo* AML. These experiments studied individuals from the general population without a known history of cancer or haematological disorder. Further work will be necessary to adapt AML predictive models to patient groups prone to CH with distinct genetic features. As discussed in the general introduction, CH is particularly common in certain clinical contexts, notably aplastic anaemia and following cytotoxic treatment for an unrelated malignancy (Bowman et al., 2018). CH in adult cancer patients has recently become an active area of research due to the increasing numbers of cancer survivors at elevated risk of CH-associated pathology, including therapy-related myeloid neoplasms (t-MN) and earlier onset of common non-malignant conditions, particularly cardiovascular disease (Bowman et al., 2018; Carver et al., 2007; Morton et al., 2018). CH has emerged as a potentially promising biomarker for the risk of t-MN and other late effects of cancer treatment (Bolton et al., 2019; Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017).

Childhood cancer survivors display an earlier onset of ageing-associated cardiometabolic conditions (Armstrong et al., 2016; Bhakta et al., 2017; Rowland and Bellizzi, 2014) and an elevated risk of t-MN and other secondary malignancies (Bhatia et al., 2007; Pui et al., 1991; Turcotte et al., 2018). Predicting and mitigating long-term complications of treatment is emerging as a dominant challenge in an era where a large proportion of children with cancer can be cured of their primary malignancy (Oeffinger et al., 2006). However, the prevalence, genetic landscape and clinical significance of CH in this population is largely unknown.

The aims of the experiments described in this chapter were the following:

1) Evaluate whether CH is prevalent in childhood cancer survivors who have received intensive cytotoxic treatment and/or radiotherapy.
2) Investigate the natural history of a case of paediatric t-MN lacking an MLL rearrangement.

The following introduction provides an overview of existing literature on cytotoxic therapy related CH and the pathogenesis of t-MN.

1.1 Therapy-related myeloid neoplasms

Epidemiology and risk factors

Therapy-related myeloid neoplasms comprise any AML or MDS arising after chemo and/or radiotherapy for a primary cancer, organ transplant or auto-immune condition (Arber et al., 2016). It constitutes one of the most challenging long-term complications of cancer treatment, with survival measured in months for most patients (Bhatia, 2013). Cytotoxic agents associated with the highest risk of t-MN are alkylating agents, topoisomerase II inhibitors and platinum-based drugs (Morton et al., 2018). The incidence and risk factors for t-MN have fluctuated as chemotherapy regimens for the commonest solid cancers have evolved (Bhatia, 2013; Morton et al., 2018). Over the past several decades, t-MN has accounted for a rising proportion of all newly diagnosed AML/MDS cases (Morton et al., 2018; Morton et al., 2014). Currently, t-MN constitutes around 10-20% of AML and MDS diagnoses, with an annual incidence of approximately 0.62/100,000 (De Roos et al., 2010; Hulegardh et al., 2015; Morton et al., 2018). A recent survey of all t-MN cases entered in the US SEER cancer registry between 2000 and 2014 found that nearly all solid tumour types were associated with t-MN, with the highest risk seen in patients treated for malignant bone tumours, followed by soft tissue sarcoma, testicular cancer, ovarian carcinoma and CNS malignancies (Morton et al., 2018). These findings represent a modest departure from previous epidemiological trends showing highest t-MN risk among breast cancer and lymphoma patients (Morton et al., 2010; Morton et al., 2018). Several solid tumour types were newly associated with t-MN risk, most likely reflecting recent introduction or increase in use of platinum agents to treatment protocols (Morton et al., 2018). Younger age at chemo/radiotherapy exposure correlated with higher t-MN risk, with high cumulative incidence of t-MN observed in children treated for solid tumours (5% to 11%) (Bhatia et al., 2007; Kushner et al., 1998; Le Deley et al., 2003; Morton et al., 2018).

Around 16-20% of t-MN patients harbour penetrant germline variants implicated in cancer susceptibility (Churpek et al., 2016; Felix et al., 1996; Schulz et al., 2012; Voso et al., 2015), compared with 9.5-12.6% of cancer patients overall and 1-2.7% of individuals without cancer (Pritchard et al., 2016; Schrader et al., 2016; Zhang et al., 2015). Cancer-predisposing germline mutations in t-MN patients are frequently reported in genes involved in mediating cellular responses to DNA damage, such as *BRCA1, BRCA2, BARD1* and *TP53* (Felix et al., 1996; Felix et al., 1998; Schulz et al., 2012). This observation may help explain the notorious chemo-resistance of t-MNs (Bhatia, 2013; McNerney et al., 2017). Germline factors may constitute a particularly powerful risk factor in children at highest risk of t-MN. For example, children with soft tissue or bone malignancies have an 11% cumulative 5-year risk of t-MN (Bhatia et al., 2007). This patient group appears to have an exceptionally high burden of germline variants predisposing to cancer, identified in nearly 50% of individuals in the most recent survey (Ballinger et al., 2016).

Genomic landscape and classification

The somatic genomic features of t-MN are similar overall to those seen in non-therapy related myeloid neoplasms, but with dramatic enrichment for high-risk changes, notably rearrangements involving *KMT2A* (*MLL*) and *RUNX1*, *TP53* mutations and chromosome 5 and/or 7 losses (Bhatia, 2013; Smith et al., 2003). In adults, two subtypes of t-MN are delineated based on chemotherapy exposure, genomic features and clinical behaviour (McNerney et al., 2017). The alkylating agent-related class of t-MN constitutes around 70% of cases and is characterised by the high-risk cytogenetic changes del(5q) and -7/del(7q) and *TP53* mutations (in around 33%) (Heuser, 2016), a relatively long latency (5-7 years from cytotoxic exposure) and a tendency to initially present as MDS progressing towards AML (McNerney et al., 2017). In addition to alkylating agents (e.g., cyclophosphamide, melphalan), this class of t-MN is associated with exposure to platinum-based agents (e.g., cisplatin, carboplatin) and purine analogues (e.g., azathioprine, fludarabine) (McNerney et al., 2017; Offman et al., 2004; Waterman et al., 2012). The second broad category of t-MN is associated with topoisomerase II inhibitor exposure (e.g., anthracyclines and etoposide)(McNerney et al., 2017). The topoisomerase II (TOP2) inhibitor class of t-MN typically presents as frank AML and has a shorter latency (median 2-3 years) (Heuser, 2016; Smith et al., 2003). This may in part be driven by translocations that are common in these t-MN involving *KMT2A* (*MLL*), *RUNX1* or *PML-RARA*, powerful oncogenic rearrangements that tend to require few cooperating events to trigger leukaemic transformation (Andersson et al., 2015; McNerney et al., 2017; Papaemmanuil et al., 2016; TCGA et al., 2013).

t-MN pathogenesis: chemotherapy-induced DNA damage or clonal selection?

Until recently, the conventional model of t-MN pathogenesis proposed that most cases were attributable to somatic driver events directly induced by cytotoxic agents (Bhatia, 2013). Many chemotherapy drugs associated with t-MN are mutagenic, and some are associated with particular patterns of genomic damage. For example, TOP2 inhibitors may increase the likelihood of reciprocal translocations by delaying ligation of double-strand breaks, thus prolonging the opportunity for recombination with DNA from another chromosome (Cowell and Austin, 2012). In keeping with this model, fusion oncogenes in t-MN arising post TOP2 inhibitor treatment tend to have breakpoints consistent with processing of 4-base staggered double-strand breaks from TOP2-mediated cleavage (Felix, 2001; Hasan et al., 2008; Mistry et al., 2005).

The alkylating agent class of t-MN is characterised by complex karyotypes, high numbers of copy number aberrations and *TP53* mutations in over a third of cases (Itzhar et al., 2011; Smith et al., 2003). Alkylating agents covalently modify DNA and promote DNA cross-linking double-strand breaks (Fu et al., 2012). It was thought that this genotoxicity induced structural changes and occasionally *TP53* mutations, with the latter contributing to genomic instability (Bhatia, 2013).

However, this model of t-MN pathogenesis was refuted by the work of Wong et al, who investigated the natural history of *TP53*-mutated t-MN (Wong et al., 2015b). Ultra-sensitive duplex sequencing demonstrated that the *TP53* driver mutation present (at clonal VAF) in the t-MN was usually detectable at very low levels (VAF 0.003-0.7%) in bone marrow samples taken prior to commencing cytotoxic treatment for the primary malignancy (Wong et al., 2015b). Furthermore, the point mutation burden and patterns did not differ between t-MN and *de novo* AML (Wong et al., 2015b). These findings suggested that cytotoxic treatment selected for pre-existing *TP53*-mutated HSCs, and that the cytogenetic complexity observed in the t-MNs reflected abrogation of the TP53-mediated DNA damage response and survival of cells that would otherwise have undergone apoptosis (Wong et al., 2015b). The clonal selection model was corroborated by a follow-up experiment in which *TP53*-mutated clones transplanted into mice only expanded if the animals were exposed to cytotoxic therapy (Wong et al., 2015b). Moreover, screening peripheral blood samples from a cohort of otherwise healthy elderly individuals (n=20) identified *TP53* mutations at very low VAF (<0.1%) in 37% (Wong et al., 2015b). These mutations persisted over time with little or no clonal expansion, suggesting that they conferred minimal selective advantage in the absence of unusual levels of genotoxic stress (Wong et al., 2015b). A contemporaneous study by Ok et al. compared *TP53* mutations in t-AML and *de novo* AML and found no evidence suggesting that *TP53* drivers in the former were induced by distinct chemotherapy-related mutational processes: there were no differences in mutation distribution, sequence context or proportion of transitions versus transversions (Ok et al., 2015). The finding that t-MN *TP53* drivers predate chemotherapy exposure has since been reproduced by other experiments (Schulz et al., 2015; Takahashi et al., 2017).

Clonal haematopoiesis as a biomarker for t-MN risk

An important role for clonal selection in t-MN pathogenesis was corroborated by recent studies investigating CH in cancer patients. CH is dramatically more prevalent in cancer survivors compared to individuals of the same age who have not been exposed to cytotoxic agents/radiotherapy and is enriched for mutations in *TP53* and its negative regulator *PPM1D* (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017). Numerous elegant studies in both mouse and human have demonstrated that cytotoxic agents and radiotherapy promote expansion of HSCs harbouring *TP53* or *PPM1D* mutations (Bondar and Medzhitov, 2010; Hsu et al., 2018; Kahn et al., 2018; Wong et al., 2015b). In keeping with the clinical significance of CH in the general population, CH in cancer survivors is associated with higher risk of t-MN as well as with non-malignant adverse outcomes (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017).

Childhood t-MN

Although t-MN is a leading cause of death in paediatric cancer patients surviving their primary cancers, relatively little is known about its pathogenesis in children (Bhatia et al., 2007; Heuser, 2016; Kushner et al., 1998; Le Deley et al., 2003; Pui et al., 1991). The relative contributions of germline risk factors, chemotherapy-induced driver mutations and clonal selection are unclear. The genomic landscape of paediatric t-MNs has not been well characterised, complicating efforts to trace their clonal evolution. However, it is conceivable that the genetic basis overlaps with that of paediatric AML/MDS/MPN arising in the absence of cytotoxic therapy, possibly with enrichment for high-risk features as seen in adults. Compared to adult MDS, paediatric myeloid neoplasms are enriched for mutations in the RAS oncogenes as well as *RUNX1, SETBP1* and *ASXL1* (Locatelli and Strahm, 2018; Pastor et al., 2017). Furthermore, more than 30% of paediatric MDS patients have an inherited cancer predisposition or bone marrow failure syndrome compared to <5% of adults (Hasle, 2016). Deletions affecting chromosome 7 (-7/7q-) or chromosome 5 (-5/-5q) are present in around 25% and 1% of paediatric MDS cases, respectively (Hasle, 2016).

Allogeneic HSCT remains the only potential cure for paediatric t-MN (Locatelli and Strahm, 2018) and unlike their adult counterparts, most children with t-MN are HSCT candidates (Hasle, 2016; Locatelli and Strahm, 2018). Importantly, the only factor associated with improved overall survival in paediatric t-MN patients is shorter delay between t-MN diagnosis and transplant (Locatelli and Strahm, 2018; Maher et al., 2017). It is therefore conceivable that early detection and monitoring of patients at highest risk of progressing to t-MN could improve outcomes by minimising the interval between t-MN manifestations and allogeneic HSCT.

Current knowledge of paediatric t-MN natural history is limited to four case reports of children with MLL-rearranged (MLLr) t-MN after TOP2 inhibitor treatment (Blanco et al., 2001; Megonigal et al., 2000; Ng et al., 2004; Robinson et al., 2008). As discussed above, there is some evidence that reciprocal fusions involving MLL may be directly induced by TOP2 inhibitors (McNerney et al., 2017). Consistent with this view, in each of these four cases, sensitive methods failed to detect the MLL fusion in blood or bone marrow samples taken before chemotherapy exposure (Blanco et al., 2001; Megonigal et al., 2000; Ng et al., 2004; Robinson et al., 2008). However, in three of the four case reports, the MLL fusion was detectable in blood and/or bone marrow over a year before t-MN presented clinically (17, 15.5, and 37 months latency in Blanco et al, Megonigal et al, and Robinson et al, respectively) (Blanco et al., 2001; Megonigal et al., 2000; Robinson et al., 2008). The shortest interval between MLLr detection and t-MN diagnosis (3 months) was reported by Ng et al in a child who developed t-MN only six months after diagnosis with hemophagocytic lymphohistiocytosis (Ng et al., 2004). These case reports offer some hope that even chemotherapy-induced fusion oncogenes generally associated with shorter latency to t-MN may be detectable early enough in disease evolution to enable monitoring and expedite definitive treatment. However, I could not identify any studies investigating the natural history of paediatric t-MN lacking an oncogenic fusion.

2. Results

2.1 Prevalence of CH-PD in childhood cancer survivors

To determine whether CH prevalence is elevated in children who have undergone intensive chemo/radiotherapy, we performed targeted deep sequencing of peripheral blood DNA from 84 paediatric cancer survivors to search for candidate driver mutations. The median age at cancer diagnosis was 4.5 years, and the commonest malignancies were acute lymphoblastic leukaemia (n=21), neuroblastoma (n=17) and non-Hodgkin lymphoma (n=10). Nineteen children had received a hematopoietic stem cell transplant (8 allogeneic and 11 autologous). The median interval between completion of cancer treatment and blood sampling was 6 years (range 2 – 25). Patient characteristics are summarised in Table 5.1 with details for each individual shown in Appendix 3.

Table 5.1 | Cohort summary

Diagnosis	Number of individuals	Mean age at diagnosis (years)	Mean time since last chemo/radiotherapy (years)
Neuroblastoma	17	3.0	11.4
Rhabdomyosarcoma	7	5.5	6.7
Acute lymphoblastic leukaemia	21	4.2	6.8
Non-Hodgkin lymphoma	10	6.9	8.3
Germ cell tumour	4	10.0	5.1
Lymphoblastic lymphoma	3	6.1	7.8
Hodgkin lymphoma	6	14.6	5.6
Nephroblastoma	5	3.5	7.6
Hepatoblastoma	1	0.3	9.4
Ewing sarcoma	4	8.0	8.0
Non-rhabdomyosarcoma soft tissue sarcoma	2	7.7	6.0
Choriocarcinoma	1	12.8	3.5
Nasopharyngeal carcinoma	1	15.9	3.0
Langerhans cell histiocytosis	2	3.4	6.6

Multiplex PCR was used to amplify 32 selected regions of 14 genes frequently mutated in CH and t-MN, including hotspots in the RAS oncogenes *NRAS* and *KRAS* (recurrently mutated in paediatric MDS/MPN) and all exons of *TP53* and *PPM1D* (Table 5.2; Methods section 2.3)(Coombs et al., 2017; Gibson et al., 2017; Locatelli and Strahm, 2018).

Table 5.2 | Genomic regions sequenced by multiplex PCR

Gene	Chromosome	Target codon/exon
NRAS	1	p.G12D
SF3B1	2	p.K666N; p.K700E
DNMT3A	2	p.R882/p.R693C
IDH1	2	p.R132H
KIT	4	exon 17
NPM1	5	p.L287fs*13
JAK2	9	p.V617F
KRAS	12	p.G12R
IDH2	15	p.R140Q; p.R172K
PPM1D	17	exons 1 - 6
TP53	17	exons 1 - 12
SRSF2	17	p.P95L
ASXL1	20	exon 12
U2AF1	21	p.S34F; p.Q157R

The median sequencing depth achieved across all regions of interest was 5,295X. No somatic mutations above the assay sensitivity threshold (VAF ≥ 0.008) were observed in any of the 84 long-term paediatric oncology follow-up patients nor in 3 children with no history of cancer (Methods section 3.2).

2.2 Tracing the evolution of a paediatric t-MN with driver mutations in *PTPN11* and *SETBP1* to emergence in early neuroblastoma treatment

As discussed in the introduction, studies of paediatric t-MN evolution have thus far been limited to case reports of children presenting with MLLr t-MN (Blanco et al., 2001; Megonigal et al., 2000; Ng et al., 2004; Robinson et al., 2008). The aim of this experiment was to retrace the emergence of a paediatric t-MN with genetic features akin to the alkylating agent class of adult t-MN described earlier.

Case Report

A 4-year old girl presented with high-risk, metastatic (stage 4) neuroblastoma with bone marrow involvement. Apart from focal neuroblastoma involvement, the initial bilateral staging trephines and aspirates showed normal trilineage haematopoiesis. Pre-treatment blood counts were normal. She was enrolled on the high-risk neuroblastoma SIOPEN trial protocol (HR-NBL-1.7/SIOPEN, NCT01704716) and underwent Rapid COJEC induction chemotherapy consisting of ten weeks of treatment with a total of five chemotherapy agents: carboplatin, etoposide, vincristine, cisplatin and cyclophosphamide at cumulative doses of 1.5g/m2, 1.4g/m^2, 12g/m^2, 320mg/m^2, 4.2g/m^2, respectively. Bilateral restaging bone marrow biopsies performed following completion of induction chemotherapy and count recovery (day 120 of treatment) remained positive for neuroblastoma infiltration. She therefore received additional induction chemotherapy to achieve metastatic remission, i.e., two cycles of TVD: Topotecan, Vincristine, Doxorubicin at cumulative doses of 15mg/m^2, 4mg/m^2 and 90mg/m^2, respectively. Platelet and neutrophil count recovery were unusually slow (3 months), though the child remained well with no infectious complications. Bone marrow examination following count recovery was normal, with cytomorphological examination negative for metastatic disease. Peripheral blood CD34+ stem cells (PBSC) were therefore harvested and she completed treatment, which included surgery, myeloablative therapy with busulfan and melphalan (BuMel), autologous PBSC rescue, irradiation of the site of primary disease (21 Gy), differentiation therapy (isotretinoin) and anti-GD2 immunotherapy. She remained well throughout, despite slow platelet and neutrophil count recovery after high-dose BuMel. Eight months after finishing treatment (32 months after diagnosis), she was incidentally noted on routine follow-up to have developed moderate peripheral cytopenia with Hb 102 g/dL, white cell count 2.3 x 10^9/L, neutrophils 1.29 x 10^9/L and platelets 91 x 10^9/L. Bone marrow examination revealed <5% blasts and no evidence of neuroblastoma recurrence. G-banded bone marrow karyotyping revealed monosomy 7 in keeping with a developing t-MN. Two months later the patient suffered local neuroblastoma relapse and succumbed to disease progression soon thereafter.

Retracing molecular emergence of t-MN

We applied whole genome and deep targeted sequencing to identify driver events in the peripheral blood at the time of t-MN diagnosis (32 months after first chemotherapy). Sequences were analysed against the reference genome in order to call deleterious germline variants and to achieve maximum sensitivity for somatic changes (Methods section 3.5). In parallel, a matched analysis was performed using whole genome sequencing of parental blood samples. Median coverage of t-MN, maternal and paternal blood samples was 74X, 111X and 100X, respectively. Whole genome sequencing identified somatic complex changes in chromosome 7 (a major clone with 7q- and a subclone with complete monosomy 7) and canonical hotspot mutations in *PTPN11* and *SETBP1* (Figure 5.1 and Table 5.3). Both copy number and point mutation drivers variants were validated by deep targeted sequencing (Methods 3.3-3.6, Figure 5.1). Moreover, unmatched analysis identified a deleterious germline *BARD1* p.E652fs*69 mutation strongly associated with hereditary cancer predisposition (ClinVar accession numbers RCV000115621.5, RCV000200198.2) (De Brakeleer et al., 2010; Ramus et al., 2015; Schrader et al., 2016; Smith et al., 2016). Although this variant had not been detected by routine clinical genetics targeted screening for cancer predisposition during neuroblastoma work-up, it was also present at SNP VAF in the maternal blood sample.

In order to retrace the emergence of the t-MN clone, we performed ultradeep targeted sequencing (median coverage 25,000X) of bilateral bone marrow biopsies taken at the end of Rapid COJEC induction, 4.5 months into treatment and 29 months prior to t-MN presentation. Unfortunately, these were the earliest samples able to be sequenced, with no pre-treatment specimens available. The *PTPN11* p.G503E mutation was present in both left and right bone marrow biopsies at VAF of 0.12% and 0.09%, respectively. The *SETBP1* p.D868G mutation was detected in the left bone marrow biopsy at a lower VAF of 0.074%. These variants were detected by two algorithms, including shearwater, which accounts for the local error rate when calling subclonal mutations (Methods section 3.3) (Gerstung et al., 2012; Gerstung et al., 2014). Although sequencing of PBSC harvest is underway to further validate this finding, the depth of the sequencing and presence of the *PTPN11* in both marrow samples gives a reasonable degree of confidence in its validity. Furthermore, several reads supporting the bone marrow *PTPN11* mutation was subsequently identified by the clinical diagnostic service using targeted sequencing on an orthogonal platform (Ion Torrent)(data not shown; personal communication from Dr Sam Behjati). The *SETBP1* mutation may be genuine in the left bone marrow and have escaped detection in the contralateral specimen due to rarity and stochastic molecule sampling, but nonetheless warrants additional validation. Copy number analysis of the targeted bone marrow sequencing revealed concordant changes consistent with recurrent copy number aberrations (CNAs) observed in neuroblastoma (Figure 5.1d,e) (Matthay et al., 2016), though lacked sensitivity to confidently call any chromosome 7 losses.

Table 5.3 | Summary of samples and genetic abnormalities

Sample ID	Month since NBL diagnosis	Clinical context	Sample type	Somatic driver events	
				Mutation	**VAF (%)**
PD31013c	32	t-MN diagnosis	Peripheral blood	*PTPN11* G503E	51.0
				SETBP1 D868G	50.0
				-7/7q-	-
PD31013d	4.5	Staging post rapid COJEC induction	Bone marrow (right iliac crest)	*PTPN11* G503E	0.09
				SETBP1 D868G	-
PD31013e	4.5		Bone marrow (left iliac crest)	*PTPN11* G503E	0.12
				SETBP1 D868G	0.074

Figure 5.1

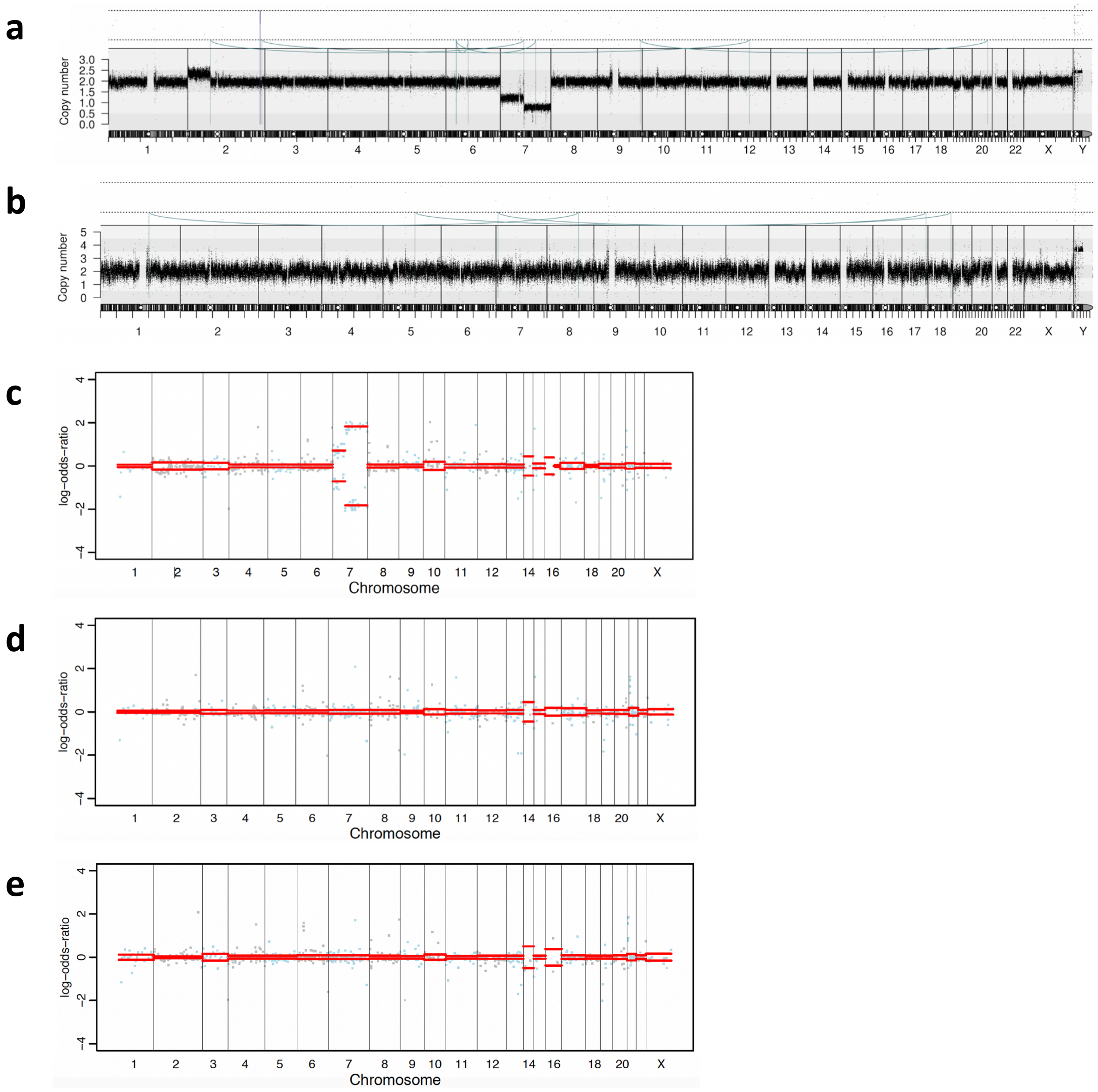


Figure 5.1 | Copy number profiles. a,b, Copy number changes and rearrangements detected from whole genome sequences of PD31013c (t-MN peripheral blood sample) (**a**) and PD31013d (right post-induction bone marrow biopsy) (**b**). The x axis shows chromosomal position and the y axis shows absolute copy number. Each dot in the plot represents the copy number of a particular genomic position (10 mega base bins). Coloured lines indicate breakpoints with rearrangements: brown, tandem duplication; blue, deletion; green and turquoise, inversion; grey, interchromosomal rearrangement. **c-e,** Copy number profiles derived from deep (>20,000x) targeted sequencing of t-MN (**c**) and bilateral bone marrow biopsies taken after induction chemotherapy, 15 months before t-MN emergence: PD31013d (right bone marrow) and PD31013e (left bone marrow) represented in panels (**d**) and (**e**), respectively. X-axis represents chromosome position. Y-axes represents allele-specific log-odds-ratio data with chromosomes alternating in blue and gray.

3. Discussion

The absence of any CH in the 84 heavily treated childhood cancer survivors screened stands in stark contrast to the situation recently observed in adults, where CH with candidate driver mutations is dramatically more common post chemo/radiotherapy than in the general population (Bowman et al., 2018; Gibson et al., 2017). Gibson et al. identified CH in over 25% of lymphoma survivors aged 30-39 and in over 40% of those aged 60-69 years (Gibson et al., 2017). Only 10 patients aged 20-29 were included in this study, none of whom had detectable CH, albeit using a less sensitive assay (detection threshold >2%)(Gibson et al., 2017). The most commonly mutated gene was *PPM1D*, which was captured in its entirety in our assay, followed by *DNMT3A* (most recurrent hotspot captured), *TET2* (not captured) and *TP53* (all exons captured) (Table 5.2)(Gibson et al., 2017). These findings have three plausible explanations. Firstly, somatic driver mutations may be extremely uncommon in the young even after exposure to chemotherapy, and hence the substrates for clonal selection are lacking. Secondly, it is possible that accrual of recognized 'driver' mutations is usually insufficient to trigger clonal expansion in the context of a very young haematopoietic niche. This hypothesis is supported by the fact that HSC mutations do begin accumulating early in life (Welch et al., 2012) and that the selective advantage of some CH drivers (most notably spliceosome gene mutations) appears to be age-dependent, implicating age-related changes in HSCs and/or their environment as key determinants of relative fitness (Link and Walter, 2016; McKerrell et al., 2015). This potential explanation is further supported by evidence that cancer-associated mutations are less able to drive clonal expansion in young compared to old stem cells (Zhu et al., 2016). Moreover, a recent study using ultra-sensitive sequencing of serially collected peripheral blood samples demonstrated that bona-fide driver mutations do not always lead to clonal expansion, even after several years (Young et al., 2016). Similar findings have been reported in other tissues, notably oesophagus and kidney, where oncogenic mutation acquisition has been timed to early childhood and adolescence, respectively (Mitchell et al., 2018; Yokoyama et al., 2019). The third potential explanation for our results is that the mutations under positive selection in paediatric cancer patients are so distinct from those observed in adult counterparts that this assay simply does not capture them. Our assay did include targets that are preferentially mutated in paediatric myeloid neoplasms – namely hotspots two RAS oncogenes and *ASXL1* exon 12 – but lacked other genes and hotspots that are likely to be enriched in paediatric t-MN or CH, notably *SETBP1, PTPN11* and *RUNX1* (Hasle, 2016; Tartaglia et al., 2003). In summary, these results should not necessarily be taken to reflect absence of potentially oncogenic HSC mutations in young cancer survivors. Rather, it is possible that even canonical CH driver mutations may not commonly drive clonal outgrowth in children and young adults despite exposure to cytotoxic drugs. More sensitive DNA sequencing methods may enable detection of very rare mutated cells in this patient group, which would lend support to this hypothesis. Equally, future sequencing studies assessing larger cohorts with a broader gene panel are warranted to explore the genetic landscape of paediatric CH. Ideally such work would be informed by a comprehensive understanding of the genomic features of paediatric t-MN, which is currently lacking.

The second experiment described in this chapter traced the emergence of t-MN during treatment for high-risk neuroblastoma. We applied deep targeted sequencing to track missense driver mutations in *PTPN11* to bone marrow samples taken at the end of induction chemotherapy. This case adds to the limited existing knowledge of paediatric t-MN evolution in several ways. Firstly, these findings contribute a fifth case to the literature suggesting that paediatric t-MN evolution typically becomes detectable very early in the treatment for the primary malignancy (Blanco et al., 2001; Megonigal et al., 2000; Ng et al., 2004; Robinson et al., 2008). In particular, this appears to be the first case reporting early molecular emergence of a non-MLLr case of paediatric t-MN. Although this patient was exposed to high doses of TOP2 inhibitors as well as platinum and alkylating agents, the clinical presentation and genomic features of this t-MN are reminiscent of the so-called alkylating agent class of adult t-MN with chromosome 7 loss, no fusion oncogene and an indolent clinical presentation with MDS rather than overt AML. In retrospect, the slow platelet and neutrophil count recovery following high-dose chemotherapy suggests early clinical manifestations of t-MN. This is in keeping with the tendency for paediatric MDS and MPN/MDS to present with neutropenia and/or thrombocytopaenia (Hasle, 2016; Kardos et al., 2003; Niemeyer and Baumann, 2011), whereas adult MDS most frequently manifests with isolated anaemia (Locatelli and Strahm, 2018; Raza and Galili, 2012). As mentioned earlier, the most frequent cooperating point mutation drivers in adult t-MN occur in *TP53*, whereas drivers in *PTPN11* and *SETBP1* are relatively rare (observed in 3-9% and 3% of adult t-MN cases, respectively) (McNerney et al., 2017). However, mutations in these genes are enriched in paediatric MPN/MDS (Hasle, 2016; Locatelli and Strahm, 2018). Somatic *PTPN11* mutations in particular are a feature of high-risk paediatric MDS warranting prompt allogeneic HSCT (Locatelli and Strahm, 2018).

Moreover, the incidental discovery of a deleterious germline *BARD1* mutation by whole genome sequencing provides further evidence that the contribution of germline predisposition to t-MN (and childhood cancer in general) may be underestimated. *BARD1* is a tumour suppressor involved in regulating the DNA damage response and TP53-mediated apoptosis (Irminger-Finger and Jefford, 2006). Loss-of-function germline mutations have been implicated in susceptibility to a variety of cancers, including t-MN (Irminger-Finger and Jefford, 2006; Schulz et al., 2012).

All discussion of clinical ramifications of these findings remains highly speculative at this point. However, current evidence indicates that the only factor clearly associated with improved childhood t-MN survival is shorter interval between t-MN diagnosis and allogeneic HSCT (Locatelli and Strahm, 2018; Maher et al., 2017). Hence it is possible that earlier detection of early t-MN clones could help address a major cause of mortality in children with cancer (Bhatia et al., 2007; Heuser, 2016; Kushner et al., 1998; Le Deley et al., 2003; Pui et al., 1991). With specific regard to neuroblastoma patients, it is conceivable that early identification of patients who may later require an allogeneic HSCT for t-MN could alter the risk/benefit analysis vis a vis proceeding with myeloablative treatment and autologous PBSC rescue (Fish and Grupp, 2008; Yalcin et al., 2015).

Collectively these findings propose several follow-up experiments. Firstly, scant knowledge of the genomic landscape of paediatric t-MN warrants collaborative efforts to whole genome sequence a sizeable cohort exposed to a range of treatment protocols. This in turn will inform future studies of the prevalence and prognostic significance of CH in childhood cancer patients. In the first instance, targeted sequencing assays should include genes preferentially mutated in paediatric myeloid neoplasms, enough heterozygous SNPs to call subclonal chromosomal arm-level copy number changes and sufficient intron tiling to detect recurrent AML-associated rearrangements. These are tractable goals even with panels small enough for routine clinical use (McKerrell et al., 2016). In the first instance, a retrospective case-control study could help assess the utility of CH-PD as a biomarker of t-MN risk. However, given the high cumulative incidence of paediatric t-MN (Bhatia et al., 2007;

Heuser, 2016; Kushner et al., 1998; Le Deley et al., 2003; Pui et al., 1991), a prospective approach in the context of clinical trial also warrants consideration, particularly for neuroblastoma and sarcoma protocols associated with the highest risk (Bhatia, 2013; Bhatia et al., 2007; Kushner et al., 1998; Morton et al., 2018).

Chapter 6

Discussion

Collectively, this work has shed light on the landscape of clonal haematopoiesis in three distinct settings: in the years preceding a diagnosis of either AML (Chapter 3) or a lymphoid malignancy (Chapter 4) and following intensive cytotoxic therapy for a childhood cancer (Chapter 5). In this discussion I will highlight common themes emerging from the results of the preceding chapters and provide an overview of further questions and areas for methods development.

1. Overview of emerging concepts

1.1 Key points:

- CH in individuals who years later develop a haematological malignancy is characterised by a different genetic landscape compared to CH in the general population, not merely by a higher mutation burden.
- Predictive models incorporating genetic and demographic variables identify most individuals with CH at high risk of progression to AML. Mutations in *TP53* and *U2AF1* are associated with a higher risk of AML progression than somatic events in the most frequently mutated CH genes.
- Clones harbouring *DNMT3A* or *TET2* mutations confer similar risks of progressing to AML versus a lymphoid neoplasm.
- Readily available clinical information improves CH risk-stratification. Higher RDW helps discriminate indolent CH from pre-AML. Lower cholesterol is reaffirmed as a likely biomarker of both lymphoid and myeloid malignancy risk.

- This work adds to the preliminary evidence suggesting that evolution of childhood t-MN may frequently be detectable early in treatment for the primary malignancy. For childhood cancer patients, the relative rarity of CH, heavy burden of t-MN and survival advantage of prompt HSCT highlight this patient group as a top priority for further study of the clinical utility of CH screening.

1.2 The mutational spectrum of premalignant CH

The prevalence, number of driver mutations and clone sizes all tended, unsurprisingly (Genovese et al., 2014; Jaiswal et al., 2014), to be markedly higher among individuals who later developed a blood cancer. However, there were also significant differences in the genetic landscape of CH in these different contexts. Within the pre-AML cohort, the spectrum of CH drivers overlapped with that seen in the general population, but was enriched for spliceosome mutations in younger individuals. By contrast, the mutational landscape preceding lymphoid cancer diagnosis was remarkably diverse, with a long 'tail' of driver mutations in genes seldom if ever implicated in CH in the general population but highly associated with lymphoid neoplasms.

1.3 CH as a biomarker of blood cancer risk irrespective of phylogenetic relationship with future malignancy

Several findings reported here add to the growing evidence that CH is a risk factor for haematological malignancy even when not related to the future neoplastic clone. Models estimating future AML or LN risk demonstrated that the number, clone size and specific genes mutated all carried predictive value. Although the power to discern gene-level risk for the pre-LN cohort was limited by the large number of infrequently mutated genes, a key finding from the LN predictive models was that *DNMT3A* and *TET2* mutations were robustly predictive of future LN risk, and that hazard ratios were equivalent to those observed for AML progression. Given that *DNMT3A* and *TET2* are much less frequently implicated as drivers in lymphoid compared to myeloid cancers, this finding suggests that CH can be a biomarker of blood cancer risk independent of the relationship between the CH clone and future malignancy. This is in keeping with observations that CH is a biomarker for t-MN risk in adult cancer patients, despite that the antecedent CH and future t-MN are often phylogenetically unrelated (Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017). Equally, a recent study suggests that *de novo* AML frequently arises from one out of many co-existing independent CH clones detectable pre-treatment (Wong et al., 2015a).

The time-course experiment data in both Chapters 3 and 4 provide further insight into the relationship between CH and future malignancy risk. Variable clonal growth trajectories were observed in premalignant cases and controls. Many clones regressed over time, including some harbouring high VAF canonical hotspot mutations, e.g., *DNMT3A* p.R882H. Hence the cell-intrinsic self-renewal advantage conferred by such mutations (Brunetti et al., 2017) does not necessarily induce inexorable clonal expansion over time, despite that they collectively confer higher leukaemia risk. Among the few pre-AML for whom diagnostic or peri-diagnostic specimens were available, most clones, though not all, expanded and appeared likely to contribute to the AML. The pre-LN serial sampling data offers even more compelling evidence that mutations unrelated to the future cancer are *bona fide* biomarkers of malignant transformation risk. Comparing pre-LN cases to controls revealed that *DNMT3A* mutations were present at significantly higher VAF in pre-LN cases. Nevertheless, even large (VAF>5%) *DNMT3A*-mutated clones often declined in size leading up to cancer diagnosis, frequently coinciding with the appearance of new, LN-associated drivers. Hence it is likely that most of the predictive power of *DNMT3A* mutations does not stem from their direct contribution to LN evolution.

Collectively, these experiments, in conjunction with the aforementioned t-MN studies, strongly suggests that CH unrelated to the future malignant clones is nevertheless a biomarker of malignant transformation risk. There are several non-mutually exclusive potential explanations for this observation. It is possible that the HSC mutation rate, and hence the likelihood of serial acquisition of drivers in any given clone, tends to be higher among individuals who develop a cancer, and the presence of multiple detectable clones is a surrogate marker of the higher mutation rate. However, the mutation burden and signatures in AML compared to normal HSCs of the same age argue against this as a universally active mechanism (Alexandrov et al., 2013; Welch et al., 2012). Alternatively, CH may be a surrogate marker of the presence/intensity of selection pressures that influence the fitness advantage conferred by particular driver mutations. Studies of CH in the context of aplastic anaemia (Yoshizato et al., 2015) and cytotoxic therapies (Gibson et al., 2017; Hsu et al., 2018; Kahn et al., 2018) provide strong evidence that extrinsic selective pressures can dramatically increase the prevalence of CH, shape the genetic landscape, and increase the malignant transformation risk. By extension, it is conceivable that the same may be true for diverse subtler extrinsic selection pressures, e.g., arising from variable ageing processes, environmental exposures, or inter-individual genetic variation. For example, physiological ageing processes occur at different rates in different individuals (Andersen et al., 2012; Finkel et al., 2007; Lopez-Otin et al., 2013). It is conceivable that age-associated increases in endogenous genotoxic stress (Rossi et al., 2007) and declines in HSC self-renewal capacity (Flach et al., 2014; Geiger et al., 2013) occur earlier or more severely in some individuals. This in turn could confer selective advantage on many mutated HSCs, increasing the number of detectable clones in younger age groups and the probability of any one of the clones acquiring additional oncogenic hits. These questions warrant further investigation, as discussed below.

2. Further questions and methodological challenges

2.1 To what extent is mutation acquisition a rate-limiting step in CH evolution?

Understanding the relative importance of mutation acquisition and extrinsic selective pressures in CH pathogenesis is an important gap in knowledge, not least for informing any future intervention strategies. For certain genes, e.g., *TP53*, very sensitive sequencing assays have demonstrated that driver mutations are common in older individuals at extremely low VAF and tend to be stable over time in the absence of any environmental selective pressures which increase mutated HSC fitness advantage (Wong et al., 2015b). By contrast, the exponential increase in the prevalence of CH harbouring spliceosome gene mutations observed in individuals aged >70 years (McKerrell et al., 2015) is poorly understood. It is possible that this phenomenon reflects ageing-associated changes in the haematopoietic niche (McKerrell and Vassiliou, 2015). For instance, spliceosome mutations may generate neoantigens that elicit a stronger immune response in younger individuals (McKerrell and Vassiliou, 2015). However, this speculation has yet to be supported by experiments demonstrating low-level persistence of rare HSCs carrying spliceosome mutations in younger individuals. The sensitivity of error-corrected sequencing assays has been a major obstacle to this type of experiment (Kennedy et al., 2014; Schmitt et al., 2012). In particular, sensitivity is hindered by target pulldown efficiency and stochastic molecular sampling, issues which can be partially circumvented by multiple target enrichment steps and using a limited number of cells as starting material (Schmitt et al., 2015). However, novel methods of increasing sensitive, accurate detection of specific mutations (Nachmanson et al., 2018; Newman et al., 2016) could be applied to the detection of canonical spliceosome gene hotspot driver variants in younger cohorts.

2.2 Haematopoiesis and ageing in health and disease

To what extent do the number, mutation rate, and clonal dynamics of haematopoietic stem and progenitor cells vary between individuals? These are pertinent questions for understanding the increased CH burden seen in individuals who later develop a haematological malignancy, as discussed above. Two recent studies have used somatic mutations to study clonal dynamics in native human haematopoiesis (Lee-Six et al., 2018; Osorio et al., 2018). Based on this work, it is likely that there are circa 50,000-200,000 HSCs contributing to haematopoiesis, dividing roughly every 2-20 months with around 14 mutations introduced per cell division (Lee-Six et al., 2018; Osorio et al., 2018). Applying similar approaches, potentially with superimposed phenotypic information, to many individuals across the age range and in disease/cancer-predisposition states will likely give valuable insights into haematopoietic ageing and CH pathogenesis.

2.3 Refining CH detection methods

The definition and terminology used to describe CH has evolved rapidly and sometimes included VAF cut-offs (Bejar, 2017). However, the latter have been decided based on technical limitations rather than mature understanding of what constitutes clinically significant CH (Bejar, 2017; Steensma et al., 2015). In future, cheaper sequencing should enable comprehensive assays to detect subclonal cancer-associated structural events in addition to point mutations. Novel sequencing methods for detecting rare somatic mutations, notably bottleneck sequencing (BotSeq), may enable broader screens for genes under positive selection in CH (Hoang et al., 2016). Briefly, BotSeq combines molecular barcoding with a subsequent dilution step, permitting highly accurate detection of rare mutations across the entire genome without the need to achieve prohibitively expensive sequencing depth. It is conceivable, though currently entirely speculative, that transcriptional or methylation-based signal may also be amenable to identifying and characterising CH and may warrant exploration in tandem with future studies of genomically-defined CH.

2.4 Prospective longitudinal studies of CH and potential intervention strategies

An important next step will be to establish large prospective longitudinal studies enabling validation and refinement of combined genomic-clinical CH risk prediction models. Ideally such studies will examine multiple clinically relevant sequelae of CH and permit identification of high-risk groups that might benefit from intervention. The nature of potential interventions is speculative at present. An increasing arsenal of targeted therapies active against recurrent cancer-associated CH mutations, including those in splicing genes (Lee et al., 2016), *JAK2* (Van den Neste et al., 2018; Vannucchi and Harrison, 2016) and *IDH1/2* (Döhner et al., 2015), may warrant investigation in high-risk CH. Moreover, two recent studies suggest that a much less costly option, ascorbic acid (vitamin C), helps restore *TET2* function in HSCs and stall leukaemia progression (Agathocleous et al., 2017; Cimmino et al., 2017). Lastly, this work further corroborates a long-recognised connection between hypocholesterolaemia and haematological malignancies. Lower HDL and LDL were both risk factors for AML in the clinical risk prediction model discussed in Chapter 3. Lower HDL was associated with a higher risk of developing a lymphoid neoplasm (Chapter 4). The latter result corroborates previous work identifying low HDL as a biomarker of future lymphoma risk years prior to diagnosis (Matsuo et al., 2017). Hypocholesterolaemia is common among blood and solid cancer patients and is inversely correlated with cancer cell LDL-/HLD-receptor activity (Ho et al., 1978; Vitols et al., 1985; Vitols et al., 1984; Vitols et al., 1992). A mendelian randomisation study by Benn et al. found that the correlation between low LDL and cancer was absent in individuals with genetic predisposition to hypocholesterolaemia, suggesting a causal link (Benn et al., 2011), though this remains contentious (Pirro et al., 2018). Pharmacologic agents targeting HDL uptake receptors and other targets involved in cholesterol metabolism have shown early evidence of therapeutic potential in several haematological malignancies (Crusz and Balkwill, 2015; McMahon et al., 2017; Pandyra et al., 2014). Interestingly, statin treatment is associated with a significant relative risk reduction for several solid tumours as well as cardiovascular disease (Demierre et al., 2005; Poynter et al., 2005). The molecular mechanisms underpinning these observations are poorly understood and may involve pleiotropic effects on multiple processes relevant to oncogenesis, including angiogenesis and inflammation (Crusz and Balkwill, 2015; Demierre et al., 2005; Hanahan and Weinberg, 2011). Collectively, these observations suggest that existing agents targeting cholesterol metabolism (Pandyra et al., 2014) warrant investigation as potential strategies for mitigating cardiovascular disease and cancer risks associated with CH.

In summary, the degree to which clones at high risk of malignant transformation - in blood and other tissues - can be reliably distinguished from their indolent counterparts is an important biological question with compelling clinical ramifications. This dissertation has explored the ability of genetic and clinical factors to identify individuals at high risk of AML and other haematological malignancies. Understanding the selective pressures and cell-intrinsic mechanisms governing clonal fate is the next important step in developing strategies to predict and prevent progression to overt malignancy.

Bibliography

Abdel-Wahab, O., Adli, M., LaFave, L.M., Gao, J., Hricik, T., Shih, A.H., Pandey, S., Patel, J.P., Chung, Y.R., Koche, R., *et al.* (2012). ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. Cancer cell *22*, 180-193.

Abdi, J., Chen, G., and Chang, H. (2013). Drug resistance in multiple myeloma: latest findings and new concepts on molecular mechanisms. Oncotarget *4*, 2186-2207.

Acuna-Hidalgo, R., Sengul, H., Steehouwer, M., van de Vorst, M., Vermeulen, S.H., Kiemeney, L., Veltman, J.A., Gilissen, C., and Hoischen, A. (2017). Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. American journal of human genetics.

Adams, P.D., Jasper, H., and Rudolph, K.L. (2015). Aging-Induced Stem Cell Mutations as Drivers for Disease and Cancer. Cell stem cell *16*, 601-612.

Agathocleous, M., Meacham, C.E., Burgess, R.J., Piskounova, E., Zhao, Z., Crane, G.M., Cowin, B.L., Bruner, E., Murphy, M.M., Chen, W., *et al.* (2017). Ascorbate regulates haematopoietic stem cell function and leukaemogenesis. Nature *advance online publication*.

Akbari, M.R., Lepage, P., Rosen, B., McLaughlin, J., Risch, H., Minden, M., and Narod, S.A. (2014). PPM1D mutations in circulating white blood cells and the risk for ovarian cancer. Journal of the National Cancer Institute *106*, djt323.

Alexandrov, L., Kim, J., Haradhvala, N.J., Huang, M.N., Ng, A.W.T., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E., Lopez-Bigas, N., *et al.* (2018). The Repertoire of Mutational Signatures in Human Cancer. 322859.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.-L., *et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Altorki, N.K., Markowitz, G.J., Gao, D., Port, J.L., Saxena, A., Stiles, B., McGraw, T., and Mittal, V. (2019). The lung microenvironment: an important regulator of tumour growth and metastasis. Nature reviews Cancer *19*, 9-31.

Amoyel, M., and Bach, E.A. (2014). Cell competition: how to eliminate your neighbours. Development (Cambridge, England) *141*, 988-1000.

Andersen, S.L., Sebastiani, P., Dworkis, D.A., Feldman, L., and Perls, T.T. (2012). Health span approximates life span among many supercentenarians: compression of morbidity at the approximate limit of life span. J Gerontol A Biol Sci Med Sci *67*, 395-405.

Anderson, K., Lutz, C., van Delft, F.W., Bateman, C.M., Guo, Y., Colman, S.M., Kempski, H., Moorman, A.V., Titley, I., Swansbury, J.*, et al.* (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature *469*, 356-361.

Andersson, A.K., Ma, J., Wang, J., Chen, X., Gedman, A.L., Dang, J., Nakitandwe, J., Holmfeldt, L., Parker, M., Easton, J.*, et al.* (2015). The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. Nature genetics *47*, 330-337.

Antoniou, A.C., Goldgar, D.E., Andrieu, N., Chang-Claude, J., Brohet, R., Rookus, M.A., and Easton, D.F. (2005). A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. Genetic epidemiology *29*, 1-11.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391-2405.

Armitage, P., and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. British journal of cancer *8*, 1-12.

Armstrong, G.T., Chen, Y., Yasui, Y., Leisenring, W., Gibson, T.M., Mertens, A.C., Stovall, M., Oeffinger, K.C., Bhatia, S., Krull, K.R.*, et al.* (2016). Reduction in Late Mortality among 5-Year Survivors of Childhood Cancer. The New England journal of medicine *374*, 833-842.

Artomov, M., Rivas, M.A., Genovese, G., and Daly, M.J. (2017). Mosaic mutations in blood DNA sequence are associated with solid tumor cancers. NPJ genomic medicine *2*, 22.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A.*, et al.* (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415-1429.e1419.

Austin, H., Delzell, E., and Cole, P. (1988). Benzene and leukemia. A review of the literature and a risk assessment. American journal of epidemiology *127*, 419-439.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. The Journal of experimental medicine *79*, 137-158.

Bahr, C., von Paleske, L., Uslu, V.V., Remeseiro, S., Takayama, N., Ng, S.W., Murison, A., Langenfeld, K., Petretich, M., Scognamiglio, R.*, et al.* (2018). A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature.

Baker, N.E., and Li, W. (2008). Cell competition and its possible relation to cancer. Cancer research *68*, 5505-5507.

Balicer, R.D., and Afek, A. (2017). Digital health nation: Israel's global big data innovation hub. Lancet (London, England) *389*, 2451-2453.

Ballinger, M.L., Goode, D.L., Ray-Coquard, I., James, P.A., Mitchell, G., Niedermayr, E., Puri, A., Schiffman, J.D., Dite, G.S., Cipponi, A*., et al.* (2016). Monogenic and polygenic determinants of sarcoma risk: an international genetic study. The Lancet Oncology *17*, 1261-1271.

Bateman, C.M., Alpar, D., Ford, A.M., Colman, S.M., Wren, D., Morgan, M., Kearney, L., and Greaves, M. (2015). Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. Leukemia *29*, 58-65.

Behjati, S., Gundem, G., Wedge, D.C., Roberts, N.D., Tarpey, P.S., Cooke, S.L., Van Loo, P., Alexandrov, L.B., Ramakrishna, M., Davies, H*., et al.* (2016). Mutational signatures of ionizing radiation in second malignancies. Nature communications *7*, 12605.

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G*., et al.* (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *513*, 422-425.

Behjati, S., Lindsay, S., Teichmann, S.A., and Haniffa, M. (2018). Mapping human development at single-cell resolution. Development (Cambridge, England) *145*.

Bejar, R. (2017). CHIP, ICUS, CCUS and other four-letter words. Leukemia *31*, 1869-1871.

Benn, M., Tybjaerg-Hansen, A., Stender, S., Frikke-Schmidt, R., and Nordestgaard, B.G. (2011). Low-density lipoprotein cholesterol and the risk of cancer: a mendelian randomization study. Journal of the National Cancer Institute *103*, 508-519.

Bhakta, N., Liu, Q., Ness, K.K., Baassiri, M., Eissa, H., Yeo, F., Chemaitilly, W., Ehrhardt, M.J., Bass, J., Bishop, M.W*., et al.* (2017). The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). Lancet (London, England) *390*, 2569-2582.

Bhatia, S. (2013). Therapy-related myelodysplasia and acute myeloid leukemia. Seminars in oncology *40*, 666-675.

Bhatia, S., Krailo, M.D., Chen, Z., Burden, L., Askin, F.B., Dickman, P.S., Grier, H.E., Link, M.P., Meyers, P.A., Perlman, E.J*., et al.* (2007). Therapy-related myelodysplasia and acute myeloid leukemia after Ewing sarcoma and primitive neuroectodermal tumor of bone: A report from the Children's Oncology Group. Blood *109*, 46-51.

Blanco, J.G., Dervieux, T., Edick, M.J., Mehta, P.K., Rubnitz, J.E., Shurtleff, S., Raimondi, S.C., Behm, F.G., Pui, C.H., and Relling, M.V. (2001). Molecular emergence of acute myeloid leukemia during treatment for acute lymphoblastic leukemia. Proceedings of the National Academy of Sciences of the United States of America *98*, 10338-10343.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P*., et al.* (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature *538*, 260-264.

Bolli, N., Avet-Loiseau, H., Wedge, D.C., Van Loo, P., Alexandrov, L.B., Martincorena, I., Dawson, K.J., Iorio, F., Nik-Zainal, S., Bignell, G.R*., et al.* (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nature communications *5*, 2997.

Bolton, K.L., Gillis, N.K., Coombs, C.C., Takahashi, K., Zehir, A., Bejar, R., Garcia-Manero, G., Futreal, A., Jensen, B.C., Diaz, L.A., Jr*., et al.* (2019). Managing Clonal Hematopoiesis in Patients With Solid Tumors. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *37*, 7-11.

Bondar, T., and Medzhitov, R. (2010). p53-mediated hematopoietic stem and progenitor cell competition. Cell stem cell *6*, 309-322.

Bonnefond, A., Skrobek, B., Lobbens, S., Eury, E., Thuillier, D., Cauchi, S., Lantieri, O., Balkau, B., Riboli, E., Marre, M*., et al.* (2013). Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. Nature genetics *45*, 1040-1043.

Bowman, R.L., Busque, L., and Levine, R.L. (2018). Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell stem cell *22*, 157-170.

Brunetti, L., Gundry, M.C., and Goodell, M.A. (2017). DNMT3A in Leukemia. Cold Spring Harbor perspectives in medicine *7*.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L*., et al.* (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome biology *17*, 66.

Buscarlet, M., Provost, S., Feroz Zada, Y., Barhdadi, A., Bourgoin, V., Lepine, G., Mollica, L., Szuber, N., Dube, M.P., and Busque, L. (2017). DNMT3A and TET2 dominate clonal hematopoiesis, demonstrate benign phenotypes and different genetic predisposition. Blood.

Busque, L., Buscarlet, M., Mollica, L., and Levine, R.L. (2018). Concise Review: Age-Related Clonal Hematopoiesis: Stem Cells Tempting the Devil. Stem cells (Dayton, Ohio) *36*, 1287-1294.

Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M., and Gilliland, D.G. (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood *88*, 59-65.

Busque, L., Patel, J.P., Figueroa, M.E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A*., et al.* (2012). Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nature genetics *44*, 1179-1181.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J*., et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203-209.

Cairns, J. (1975). Mutation selection and the natural history of cancer. Nature *255*, 197-200.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.L.*, et al.* (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109-1113.

Campo, E., Swerdlow, S.H., Harris, N.L., Pileri, S., Stein, H., and Jaffe, E.S. (2011). The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood *117*, 5019-5032.

Carrelha, J., Meng, Y., Kettyle, L.M., Luis, T.C., Norfo, R., Alcolea, V., Boukarabila, H., Grasso, F., Gambardella, A., Grover, A.*, et al.* (2018). Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. Nature *554*, 106-111.

Carver, J.R., Shapiro, C.L., Ng, A., Jacobs, L., Schwartz, C., Virgo, K.S., Hagerty, K.L., Somerfield, M.R., and Vaughn, D.J. (2007). American Society of Clinical Oncology clinical evidence review on the ongoing care of adult cancer survivors: cardiac and pulmonary late effects. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *25*, 3991-4008.

Challen, G.A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J.S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y.*, et al.* (2011). Dnmt3a is essential for hematopoietic stem cell differentiation. Nature genetics *44*, 23-31.

Chang, W., Brohl, A.S., Patidar, R., Sindiri, S., Shern, J.F., Wei, J.S., Song, Y.K., Yohe, M.E., Gryder, B., Zhang, S.*, et al.* (2016). MultiDimensional ClinOmics for Precision Therapy of Children and Adolescent Young Adults with Relapsed and Refractory Cancer: A Report from the Center for Cancer Research. Clinical cancer research : an official journal of the American Association for Cancer Research *22*, 3810-3820.

Chapuy, B., Stewart, C., Dunford, A.J., Kim, J., Kamburov, A., Redd, R.A., Lawrence, M.S., Roemer, M.G.M., Li, A.J., Ziepert, M.*, et al.* (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nature medicine *24*, 679-690.

Chen, J., Kao, Y.R., Sun, D., Todorova, T.I., Reynolds, D., Narayanagari, S.R., Montagna, C., Will, B., Verma, A., and Steidl, A.U. (2018). Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level. Nature medicine.

Churpek, J.E., Marquez, R., Neistadt, B., Claussen, K., Lee, M.K., Churpek, M.M., Huo, D., Weiner, H., Bannerjee, M., Godley, L.A.*, et al.* (2016). Inherited mutations in cancer susceptibility genes are common among survivors of breast cancer who develop therapy-related leukemia. Cancer *122*, 304-311.

Ciccarelli, F.D. (2019). Mutations differ in normal and cancer cells of the oesophagus. Nature *565*, 301-303.

Cimmino, L., Dolgalev, I., Wang, Y., Yoshimi, A., Martin, G.H., Wang, J., Ng, V., Xia, B., Witkowski, M.T., Mitchell-Flack, M.*, et al.* (2017). Restoration of TET2 Function Blocks Aberrant Self-Renewal and Leukemia Progression. Cell *170*, 1079-1095.e1020.

Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A.*, et al.* (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science (New York, NY) *359*, 926-930.

Collado, M., Gil, J., Efeyan, A., Guerra, C., Schuhmacher, A.J., Barradas, M., Benguria, A., Zaballos, A., Flores, J.M., Barbacid, M.*, et al.* (2005). Tumour biology: senescence in premalignant tumours. Nature *436*, 642.

Coombs, C.C., Zehir, A., Devlin, S.M., Kishtagari, A., Syed, A., Jonsson, P., Hyman, D.M., Solit, D.B., Robson, M.E., Baselga, J.*, et al.* (2017). Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. Cell stem cell *21*, 374-382 e374.

Couronne, L., Bastard, C., and Bernard, O.A. (2012). TET2 and DNMT3A mutations in human T-cell lymphoma. The New England journal of medicine *366*, 95-96.

Cowell, I.G., and Austin, C.A. (2012). Mechanism of generation of therapy related leukemia in response to anti-topoisomerase II agents. International journal of environmental research and public health *9*, 2075-2091.

CRUK (2018). Acute myeloid leukaemia (AML) incidence statistics.

Crusz, S.M., and Balkwill, F.R. (2015). Inflammation and cancer: advances and new agents. Nature reviews Clinical oncology *12*, 584-596.

Day, N., Oakes, S., Luben, R., Khaw, K.T., Bingham, S., Welch, A., and Wareham, N. (1999). EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. British journal of cancer *80 Suppl 1*, 95-103.

De Brakeleer, S., De Greve, J., Loris, R., Janin, N., Lissens, W., Sermijn, E., and Teugels, E. (2010). Cancer predisposing missense and protein truncating BARD1 mutations in non-BRCA1 or BRCA2 breast cancer families. Hum Mutat *31*, E1175-1185.

De Roos, A.J., Deeg, H.J., Onstad, L., Kopecky, K.J., Bowles, E.J., Yong, M., Fryzek, J., and Davis, S. (2010). Incidence of myelodysplastic syndromes within a nonprofit healthcare system in western Washington state, 2005-2006. American journal of hematology *85*, 765-770.

Deininger, M.W.N., Tyner, J.W., and Solary, E. (2017). Turning the tide in myelodysplastic/myeloproliferative neoplasms. Nature reviews Cancer *17*, 425-440.

Demierre, M.F., Higgins, P.D., Gruber, S.B., Hawk, E., and Lippman, S.M. (2005). Statins and cancer prevention. Nature reviews Cancer *5*, 930-942.

Denoix, P.F. (1954). De la diversité de certains cancers. Monographie de l'Institut national d'hygiène *5*.

Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M.S., Cheang, G., Lee, S., Samuel, M., Ritchie, E.K., Guzman, M.L., Ballman, K.V.*, et al.* (2018). Somatic mutations precede acute myeloid leukemia years before diagnosis. Nature medicine *24*, 1015-1023.

Deschler, B., and Lubbert, M. (2006). Acute myeloid leukemia: epidemiology and etiology. Cancer *107*, 2099-2107.

Ding, J.H., Li, S.P., Cao, H.X., Wu, J.Z., Gao, C.M., Liu, Y.T., Zhou, J.N., Chang, J., and Yao, G.H. (2010). Alcohol dehydrogenase-2 and aldehyde dehydrogenase-2 genotypes, alcohol drinking and the risk for esophageal cancer in a Chinese population. Journal of human genetics *55*, 97-102.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D.*, et al.* (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506-510.

Dohner, H., Estey, E.H., Amadori, S., Appelbaum, F.R., Buchner, T., Burnett, A.K., Dombret, H., Fenaux, P., Grimwade, D., Larson, R.A.*, et al.* (2010). Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood *115*, 453-474.

Döhner, H., Weisdorf, D.J., and Bloomfield, C.D. (2015). Acute Myeloid Leukemia. New England Journal of Medicine *373*, 1136-1152.

Dominguez, P.M., Ghamlouch, H., Rosikiewicz, W., Kumar, P., Beguelin, W., Fontan, L., Rivas, M.A., Pawlikowska, P., Armand, M., Mouly, E.*, et al.* (2018). TET2 Deficiency Causes Germinal Center Hyperplasia, Impairs Plasma Cell Differentiation, and Promotes B-cell Lymphomagenesis. Cancer discovery *8*, 1632-1653.

Dores, G.M., Anderson, W.F., Curtis, R.E., Landgren, O., Ostroumova, E., Bluhm, E.C., Rabkin, C.S., Devesa, S.S., and Linet, M.S. (2007). Chronic lymphocytic leukaemia and small lymphocytic lymphoma: overview of the descriptive epidemiology. British journal of haematology *139*, 809-819.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: a human perspective. Cell stem cell *10*, 120-136.

Dunn, J., Qiu, H., Kim, S., Jjingo, D., Hoffman, R., Kim, C.W., Jang, I., Son, D.J., Kim, D., Pan, C.*, et al.* (2014). Flow-dependent epigenetic DNA methylation regulates endothelial gene expression and atherosclerosis. The Journal of clinical investigation *124*, 3187-3199.

Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G., and Hartwell, L. (2003). The case for early detection. Nature reviews Cancer *3*, 243-252.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759-767.

Feinberg, A.P., Koldobskiy, M.A., and Gondor, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nature reviews Genetics *17*, 284-299.

Felix, C.A. (2001). Leukemias related to treatment with DNA topoisomerase II inhibitors. Medical and pediatric oncology *36*, 525-535.

Felix, C.A., Hosler, M.R., Provisor, D., Salhany, K., Sexsmith, E.A., Slater, D.J., Cheung, N.K., Winick, N.J., Strauss, E.A., Heyn, R.*, et al.* (1996). The p53 gene in pediatric therapy-related leukemia and myelodysplasia. Blood *87*, 4376-4381.

Felix, C.A., Megonigal, M.D., Chervinsky, D.S., Leonard, D.G., Tsuchida, N., Kakati, S., Block, A.M., Fisher, J., Grossi, M., Salhany, K.I.*, et al.* (1998). Association of germline p53 mutation with MLL segmental jumping translocation in treatment-related leukemia. Blood *91*, 4451-4456.

Figueroa, M.E., Abdel-Wahab, O., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F.*, et al.* (2010a). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer cell *18*, 553-567.

Figueroa, M.E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P.J., Schifano, E., Booth, J., van Putten, W., Skrabanek, L.*, et al.* (2010b). DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer cell *17*, 13-27.

Finkel, T., Serrano, M., and Blasco, M.A. (2007). The common biology of cancer and ageing. Nature *448*, 767-774.

Fish, J.D., and Grupp, S.A. (2008). Stem cell transplantation for neuroblastoma. Bone Marrow Transplant *41*, 159-165.

Flach, J., Bakker, S.T., Mohrin, M., Conroy, P.C., Pietras, E.M., Reynaud, D., Alvarez, S., Diolaiti, M.E., Ugarte, F., Forsberg, E.C.*, et al.* (2014). Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. Nature *512*, 198-202.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research *39*, D945-950.

Forsberg, L.A., Gisselsson, D., and Dumanski, J.P. (2017). Mosaicism in health and disease - clones picking up speed. Nature reviews Genetics *18*, 128-142.

Forsberg, L.A., Rasi, C., Razzaghian, H.R., Pakalapati, G., Waite, L., Thilbeault, K.S., Ronowicz, A., Wineinger, N.E., Tiwari, H.K., Boomsma, D.*, et al.* (2012). Age-related somatic structural changes in the nuclear genome of human blood cells. American journal of human genetics *90*, 217-228.

Foulds, L. (1958). The natural history of cancer. Journal of chronic diseases *8*, 2-37.

Fox, M. (2013). Janet D. Rowley, Physician, Dies at 88; Discovered That Cancer Can Be Genetic. In New York Times (New York).

Frick, M., Chan, W., Arends, C.M., Hablesreiter, R., Halik, A., Heuser, M., Michonneau, D., Blau, O., Hoyer, K., Christen, F.*, et al.* (2018). Role of Donor Clonal Hematopoiesis in

Allogeneic Hematopoietic Stem-Cell Transplantation. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, JCO2018792184.

Fu, D., Calvo, J.A., and Samson, L.D. (2012). Balancing repair and tolerance of DNA damage caused by alkylating agents. Nature reviews Cancer *12*, 104-120.

Fuchs, E. (1882). Das Sarkom des Uvealtractus. Graefe's Archiv für Ophthalmologie *12*, 233.

Fuster, J.J., MacLauchlan, S., Zuriaga, M.A., Polackal, M.N., Ostriker, A.C., Chakraborty, R., Wu, C.L., Sano, S., Muralidharan, S., Rius, C.*, et al.* (2017). Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. Science (New York, NY) *355*, 842-847.

Geiger, H., de Haan, G., and Florian, M.C. (2013). The ageing haematopoietic stem cell compartment. Nature reviews Immunology *13*, 376-389.

Genovese, G., Kahler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M.*, et al.* (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. The New England journal of medicine *371*, 2477-2487.

Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A.*, et al.* (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine *366*, 883-892.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nature communications *3*, 811.

Gerstung, M., Papaemmanuil, E., and Campbell, P.J. (2014). Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics (Oxford, England) *30*, 1198-1204.

Gerstung, M., Papaemmanuil, E., Martincorena, I., Bullinger, L., Gaidzik, V.I., Paschka, P., Heuser, M., Thol, F., Bolli, N., Ganly, P.*, et al.* (2017). Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nature genetics *49*, 332-340.

Gibson, C.J., Lindsley, R.C., Tchekmedyian, V., Mar, B.G., Shi, J., Jaiswal, S., Bosworth, A., Francisco, L., He, J., Bansal, A.*, et al.* (2017). Clonal Hematopoiesis Associated With Adverse Outcomes After Autologous Stem-Cell Transplantation for Lymphoma. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, Jco2016716712.

Gilliland, D.G., and Griffin, J.D. (2002). The roles of FLT3 in hematopoiesis and leukemia. Blood *100*, 1532-1542.

Gillis, N.K., Ball, M., Zhang, Q., Ma, Z., Zhao, Y., Yoder, S.J., Balasis, M.E., Mesa, T.E., Sallman, D.A., Lancet, J.E.*, et al.* (2017). Clonal haemopoiesis and therapy-related myeloid malignancies in elderly patients: a proof-of-concept, case-control study. The Lancet Oncology *18*, 112-121.

Goncalves, R.P., Rodrigues, D.G., and Maranhao, R.C. (2005). Uptake of high density lipoprotein (HDL) cholesteryl esters by human acute leukemia cells. Leukemia research *29*, 955-959.

Gore, L., Kearns, P.R., de Martino Lee, M.L., De Souza, C.A., Bertrand, Y., Hijiya, N., Stork, L.C., Chung, N.G., Cardos, R.C., Saikia, T.*, et al.* (2018). Dasatinib in Pediatric Patients With Chronic Myeloid Leukemia in Chronic Phase: Results From a Phase II Trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, Jco2017759597.

Greaves, M. (2015). Evolutionary determinants of cancer. Cancer discovery *5*, 806-820.

Greaves, M., Colman, S.M., Kearney, L., and Ford, A.M. (2011). Fusion genes in cord blood. Blood *117*, 369-370; author reply 370-361.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306-313.

Greaves, M.F., Maia, A.T., Wiemels, J.L., and Ford, A.M. (2003). Leukemia in twins: lessons in natural history. Blood *102*, 2321-2333.

Greaves, M.F., and Wiemels, J. (2003). Origins of chromosome translocations in childhood leukaemia. Nature reviews Cancer *3*, 639-649.

Green, D.R., Galluzzi, L., and Kroemer, G. (2011). Mitochondria and the autophagy-inflammation-cell death axis in organismal aging. Science (New York, NY) *333*, 1109-1112.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M., Papaemmanuil, E., Brewer, D.S., Kallio, H.M., Hognas, G., Annala, M.*, et al.* (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353-357.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Haney, S.L., Upchurch, G.M., Opavska, J., Klinkebiel, D., Appiah, A.K., Smith, L.M., Heavican, T.B., Iqbal, J., Joshi, S., and Opavsky, R. (2016a). Loss of Dnmt3a induces CLL and PTCL with distinct methylomes and transcriptomes in mice. Scientific reports *6*, 34222.

Haney, S.L., Upchurch, G.M., Opavska, J., Klinkebiel, D., Hlady, R.A., Roy, S., Dutta, S., Datta, K., and Opavsky, R. (2016b). Dnmt3a Is a Haploinsufficient Tumor Suppressor in CD8+ Peripheral T Cell Lymphoma. PLoS Genet *12*, e1006334.

Hardy, P.A., and Zacharias, H. (2005). Reappraisal of the Hansemann-Boveri hypothesis on the origin of tumors. Cell biology international *29*, 983-992.

Harrell, F.E., Jr., Lee, K.L., and Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine *15*, 361-387.

Hasan, S.K., Mays, A.N., Ottone, T., Ledda, A., La Nasa, G., Cattaneo, C., Borlenghi, E., Melillo, L., Montefusco, E., Cervera, J.*, et al.* (2008). Molecular analysis of t(15;17) genomic breakpoints in secondary acute promyelocytic leukemia arising after treatment of multiple sclerosis. Blood *112*, 3383-3390.

Hasle, H. (2016). Myelodysplastic and myeloproliferative disorders of childhood. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program *2016*, 598-604.

Heagerty, P.J., Lumley, T., and Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics *56*, 337-344.

Hermouet, S., and Vilaine, M. (2011). The JAK2 46/1 haplotype: a marker of inappropriate myelomonocytic response to cytokine stimulation, leading to increased risk of inflammation, myeloid neoplasm, and impaired defense against infection? Haematologica *96*, 1575-1579.

Heuser, M. (2016). Therapy-related myeloid neoplasms: does knowing the origin help to guide treatment? Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program *2016*, 24-32.

Hinds, D.A., Barnholt, K.E., Mesa, R.A., Kiefer, A.K., Do, C.B., Eriksson, N., Mountain, J.L., Francke, U., Tung, J.Y., Nguyen, H.M.*, et al.* (2016). Germline variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood.

Hing, Z.A., Blachly, J.S., Goettl, V.M., Singh, G., Byrd, J.C., and Lapalombella, R. (2016). Exploring the Role of the Recurrent Exportin 1 (XPO1/CRM1) Mutations E571G and E571K in Chronic Lymphocytic Leukemia. *128*, 972-972.

Ho, Y.K., Smith, R.G., Brown, M.S., and Goldstein, J.L. (1978). Low-density lipoprotein (LDL) receptor activity in human acute myelogenous leukemia cells. Blood *52*, 1099-1114.

Hoang, M.L., Chen, C.H., Sidorenko, V.S., He, J., Dickman, K.G., Yun, B.H., Moriya, M., Niknafs, N., Douville, C., Karchin, R.*, et al.* (2013). Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Science translational medicine *5*, 197ra102.

Hoang, M.L., Kinde, I., Tomasetti, C., McMahon, K.W., Rosenquist, T.A., Grollman, A.P., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2016). Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America *113*, 9846-9851.

Holohan, C., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2013). Cancer drug resistance: an evolving paradigm. Nature reviews Cancer *13*, 714-726.

Howlader, N., Noone, A., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., Altekruse, S., Kosary, C., Ruhl, J., and Tatalovich, Z. (2011). SEER cancer statistics review, 1975–2008. Bethesda, MD: National Cancer Institute *19*.

Hsu, J.I., Dayaram, T., Tovy, A., De Braekeleer, E., Jeong, M., Wang, F., Zhang, J., Heffernan, T.P., Gera, S., Kovacs, J.J.*, et al.* (2018). PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. Cell stem cell *23*, 700-713 e706.

Hu, L., Li, M., Ding, Y., Pu, L., Liu, J., Xie, J., Cabanero, M., Li, J., Xiang, R., and Xiong, S. (2017). Prognostic value of RDW in cancers: a systematic review and meta-analysis. Oncotarget *8*, 16027-16035.

Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N.*, et al.* (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355-370.e314.

Huet, S., Paubelle, E., Lours, C., Grange, B., Courtois, L., Chabane, K., Charlot, C., Mosnier, I., Simonet, T., Hayette, S.*, et al.* (2018). Validation of the prognostic value of the knowledge bank approach to determine AML prognosis in real life. Blood *132*, 865-867.

Hulegardh, E., Nilsson, C., Lazarevic, V., Garelius, H., Antunovic, P., Rangert Derolf, A., Mollgard, L., Uggla, B., Wennstrom, L., Wahlin, A.*, et al.* (2015). Characterization and prognostic features of secondary acute myeloid leukemia in a population-based setting: a report from the Swedish Acute Leukemia Registry. American journal of hematology *90*, 208-214.

Hunger, S.P. (2017). CML in blast crisis: more common than we think? Blood *129*, 2713-2714.

Hunter, K.W., Amin, R., Deasy, S., Ha, N.H., and Wakefield, L. (2018). Genetic insights into the morass of metastatic heterogeneity. Nature reviews Cancer *18*, 211-223.

Ihara, K., Ishii, E., Eguchi, M., Takada, H., Suminoe, A., Good, R.A., and Hara, T. (1999). Identification of mutations in the c-mpl gene in congenital amegakaryocytic thrombocytopenia. Proceedings of the National Academy of Sciences of the United States of America *96*, 3132-3136.

Inoue, K., and Fry, E.A. (2017). Haploinsufficient tumor suppressor genes. Advances in medicine and biology *118*, 83-122.

Irminger-Finger, I., and Jefford, C.E. (2006). Is there more to BARD1 than BRCA1? Nature reviews Cancer *6*, 382-391.

Itzhar, N., Dessen, P., Toujani, S., Auger, N., Preudhomme, C., Richon, C., Lazar, V., Saada, V., Bennaceur, A., Bourhis, J.H.*, et al.* (2011). Chromosomal minimal critical regions in therapy-related leukemia appear different from those of de novo leukemia by high-resolution aCGH. PloS one *6*, e16623.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J.*, et al.* (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nature genetics *44*, 651-658.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., *et al.* (2014). Age-related clonal hematopoiesis associated with adverse outcomes. The New England journal of medicine *371*, 2488-2498.

Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., *et al.* (2017). Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. The New England journal of medicine *377*, 111-121.

Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L., Quake, S.R., and Majeti, R. (2012). Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Science translational medicine *4*, 149ra118.

Jeong, M., Park, H.J., Celik, H., Ostrander, E.L., Reyes, J.M., Guzman, A., Rodriguez, B., Lei, Y., Lee, Y., Ding, L., *et al.* (2018). Loss of Dnmt3a Immortalizes Hematopoietic Stem Cells In Vivo. Cell reports *23*, 1-10.

Jones, A.V., Chase, A., Silver, R.T., Oscier, D., Zoi, K., Wang, Y.L., Cario, H., Pahl, H.L., Collins, A., Reiter, A., *et al.* (2009). JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nature genetics *41*, 446-449.

Jones, D., Raine, K.M., Davies, H., Tarpey, P.S., Butler, A.P., Teague, J.W., Nik-Zainal, S., and Campbell, P.J. (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. Current protocols in bioinformatics *56*, 15.10.11-15.10.18.

Ju, Y.S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L.B., Rahbari, R., Wedge, D.C., Davies, H.R., Ramakrishna, M., Fullam, A., *et al.* (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature *543*, 714-718.

Kahn, J.D., Miller, P.G., Silver, A.J., Sellar, R.S., Bhatt, S., Gibson, C., McConkey, M., Adams, D., Mar, B., Mertins, P., *et al.* (2018). PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. Blood *132*, 1095-1105.

Kalmanti, L., Saussele, S., Lauseker, M., Muller, M.C., Dietz, C.T., Heinrich, L., Hanfstein, B., Proetel, U., Fabarius, A., Krause, S.W., *et al.* (2015). Safety and efficacy of imatinib in CML over a period of 10 years: data from the randomized CML-study IV. Leukemia *29*, 1123-1132.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., *et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Kardos, G., Baumann, I., Passmore, S.J., Locatelli, F., Hasle, H., Schultz, K.R., Stary, J., Schmitt-Graeff, A., Fischer, A., Harbott, J., *et al.* (2003). Refractory anemia in childhood: a retrospective analysis of 67 patients with particular reference to monosomy 7. Blood *102*, 1997-2003.

Karoulia, Z., Gavathiotis, E., and Poulikakos, P.I. (2017). New perspectives for targeting RAF kinase in human cancer. Nature reviews Cancer *17*, 676-691.

Kennedy, J.A., and Ebert, B.L. (2017). Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *35*, 968-974.

Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J., Ahn, E.H., Prindle, M.J., Kuong, K.J., Shen, J.-C., Risques, R.-A*., et al.* (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protocols *9*, 2586-2606.

Kilpivaara, O., Mukherjee, S., Schram, A.M., Wadleigh, M., Mullally, A., Ebert, B.L., Bass, A., Marubayashi, S., Heguy, A., Garcia-Manero, G*., et al.* (2009). A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. Nature genetics *41*, 455-459.

Kim, K.H., and Roberts, C.W. (2016). Targeting EZH2 in cancer. Nature medicine *22*, 128-134.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., and Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America *108*, 9530-9535.

Knudson, A.G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America *68*, 820-823.

Ko, M., Bandukwala, H.S., An, J., Lamperti, E.D., Thompson, E.C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S.B., and Rao, A. (2011). Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. Proceedings of the National Academy of Sciences of the United States of America *108*, 14566-14571.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research *22*, 568-576.

Koren, A., Handsaker, R.E., Kamitaki, N., Karlic, R., Ghosh, S., Polak, P., Eggan, K., and McCarroll, S.A. (2014). Genetic variation in human DNA replication timing. Cell *159*, 1015-1026.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature methods *6*, 291-295.

Kronke, J., Bullinger, L., Teleanu, V., Tschurtz, F., Gaidzik, V.I., Kuhn, M.W., Rucker, F.G., Holzmann, K., Paschka, P., Kapp-Schworer, S*., et al.* (2013). Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. Blood *122*, 100-108.

Krontiris, T.G., and Cooper, G.M. (1981). Transforming activity of human tumor DNAs. Proceedings of the National Academy of Sciences of the United States of America *78*, 1181-1184.

Kushner, B.H., Cheung, N.K., Kramer, K., Heller, G., and Jhanwar, S.C. (1998). Neuroblastoma and treatment-related myelodysplasia/leukemia: the Memorial Sloan-Kettering experience and a literature review. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *16*, 3880-3889.

Landau, D.A., Tausch, E., Taylor-Weiner, A.N., Stewart, C., Reiter, J.G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Bottcher, S*., et al.* (2015). Mutations driving CLL and their evolution in progression and relapse. Nature *526*, 525-530.

Landau, D.A., and Wu, C.J. (2013). Chronic lymphocytic leukemia: molecular heterogeneity revealed by high-throughput genomics. Genome medicine *5*, 47.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W*., et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landgren, O., Albitar, M., Ma, W., Abbasi, F., Hayes, R.B., Ghia, P., Marti, G.E., and Caporaso, N.E. (2009). B-cell clones as early markers for chronic lymphocytic leukemia. The New England journal of medicine *360*, 659-667.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J*., et al.* (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research *44*, D862-868.

Latchney, S.E., and Calvi, L.M. (2017). The aging hematopoietic stem cell niche: Phenotypic and functional changes and mechanisms that contribute to hematopoietic aging. Seminars in hematology *54*, 25-32.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C*., et al.* (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nature genetics *44*, 642-650.

Lausten-Thomsen, U., Madsen, H.O., Vestergaard, T.R., Hjalgrim, H., Nersting, J., and Schmiegelow, K. (2011). Prevalence of t(12;21)[ETV6-RUNX1]-positive cells in healthy neonates. Blood *117*, 186-189.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A*., et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Le Deley, M.C., Leblanc, T., Shamsaldin, A., Raquin, M.A., Lacour, B., Sommelet, D., Chompret, A., Cayuela, J.M., Bayle, C., Bernheim, A*., et al.* (2003). Risk of secondary leukemia after a solid tumor in childhood according to the dose of epipodophyllotoxins and anthracyclines: a case-control study by the Societe Francaise d'Oncologie Pediatrique.

Journal of clinical oncology : official journal of the American Society of Clinical Oncology *21*, 1074-1081.

Lee, S.C., Dvinge, H., Kim, E., Cho, H., Micol, J.B., Chung, Y.R., Durham, B.H., Yoshimi, A., Kim, Y.J., Thomas, M.*, et al.* (2016). Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins. Nature medicine *22*, 672-678.

Lee-Six, H., Obro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E.*, et al.* (2018). Population dynamics of normal human blood inferred from somatic mutations. Nature *561*, 473-478.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754-1760.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *26*, 589-595.

Li, Y., Roberts, N., Weischenfeldt, J., Wala, J.A., Shapira, O., Schumacher, S., Khurana, E., Korbel, J.O., Imielinski, M., Beroukhim, R.*, et al.* (2017). Patterns of structural variation in human cancer. bioRxiv.

Lim, U., Gayles, T., Katki, H.A., Stolzenberg-Solomon, R., Weinstein, S.J., Pietinen, P., Taylor, P.R., Virtamo, J., and Albanes, D. (2007). Serum high-density lipoprotein cholesterol and risk of non-hodgkin lymphoma. Cancer research *67*, 5569-5574.

Link, D.C., and Walter, M.J. (2016). /`CHIP/'ping away at clonal hematopoiesis. Leukemia *30*, 1633-1635.

Locatelli, F., and Strahm, B. (2018). How I treat myelodysplastic syndromes of childhood. Blood *131*, 1406-1414.

Loftfield, E., Zhou, W., Graubard, B.I., Yeager, M., Chanock, S.J., Freedman, N.D., and Machiela, M.J. (2018a). Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. Scientific reports *8*, 12316.

Loftfield, E., Zhou, W., Yeager, M., Chanock, S.J., Freedman, N.D., and Machiela, M.J. (2018b). Mosaic Y loss is moderately associated with solid tumor risk. Cancer research.

Loh, P.R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F., Birmann, B.M., Talkowski, M.E., Bakhoum, S.F., McCarroll, S.A.*, et al.* (2018). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature *559*, 350-355.

Lopez-Otin, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194-1217.

Lu, C., Xie, M., Wendl, M.C., Wang, J., McLellan, M.D., Leiserson, M.D.M., Huang, K.-l., Wyczalkowski, M.A., Jayasinghe, R., Banerjee, T.*, et al.* (2015). Patterns and functional implications of rare germline variants across 12 cancer types. Nature communications *6*.

Lunning, M.A., and Green, M.R. (2015). Mutation of chromatin modifiers; an emerging hallmark of germinal center B-cell lymphomas. Blood cancer journal *5*, e361.

Ma, J., Setton, J., Lee, N.Y., Riaz, N., and Powell, S.N. (2018). The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. Nature communications *9*, 3292.

Machiela, M.J., Zhou, W., Sampson, J.N., Dean, M.C., Jacobs, K.B., Black, A., Brinton, L.A., Chang, I.S., Chen, C., Chen, C.*, et al.* (2015). Characterization of large structural genetic mosaicism in human autosomes. American journal of human genetics *96*, 487-497.

Maher, O.M., Silva, J.G., Wu, J., Liu, D., Cooper, L.J., Tarek, N., Worth, L., Lee, D.A., Petropoulos, D., Franklin, A.R.*, et al.* (2017). Outcomes of children, adolescents, and young adults following allogeneic stem cell transplantation for secondary acute myeloid leukemia and myelodysplastic syndromes-The MD Anderson Cancer Center experience. Pediatric transplantation *21*.

Makohon-Moore, A., and Iacobuzio-Donahue, C.A. (2016). Pancreatic cancer biology and genetics from an evolutionary perspective. Nature reviews Cancer *16*, 553-565.

Manchester, K.L. (1995). Theodor Boveri and the origin of malignant tumours. Trends in cell biology *5*, 384-387.

Mantovani, A., Marchesi, F., Malesci, A., Laghi, L., and Allavena, P. (2017). Tumour-associated macrophages as treatment targets in oncology. Nature reviews Clinical oncology *14*, 399-416.

Maris, J.M. (2015). Defining Why Cancer Develops in Children. New England Journal of Medicine *373*, 2373-2375.

Maris, J.M., and Knudson, A.G. (2015). Revisiting tissue specificity of germline cancer predisposing mutations. Nature reviews Cancer *15*, 65-66.

Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science (New York, NY) *349*, 1483-1489.

Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R.*, et al.* (2018). Somatic mutant clones colonize the human esophagus with age. Science (New York, NY) *362*, 911-917.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041 e1021.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M.*, et al.* (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science (New York, NY) *348*, 880-886.

Matsuo, T., Tashiro, H., Sumiyoshi, R., Yamamoto, T., Oka-Miura, Y., Matsumoto, K., Ooi, J., and Shirafuji, N. (2017). Low High-Density Lipoprotein Cholesterol (HDL) Is a Significant Poor Prognostic Factor in Malignant Lymphoma. *130*, 5148-5148.

Matthay, K.K., Maris, J.M., Schleiermacher, G., Nakagawara, A., Mackall, C.L., Diller, L., and Weiss, W.A. (2016). Neuroblastoma. Nature reviews Disease primers *2*, 16078.

Mattox, A.K., Wang, Y., Springer, S., Cohen, J.D., Yegnasubramanian, S., Nelson, W.G., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2017). Bisulfite-converted duplexes for the strand-specific detection and quantification of rare mutations. Proceedings of the National Academy of Sciences of the United States of America *114*, 4733-4738.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.

McKerrell, T., Moreno, T., Ponstingl, H., Bolli, N., Dias, J.M., Tischler, G., Colonna, V., Manasse, B., Bench, A., Bloxham, D.*, et al.* (2016). Development and validation of a comprehensive genomic diagnostic tool for myeloid malignancies. Blood.

McKerrell, T., Park, N., Chi, J., Collord, G., Moreno, T., Ponstingl, H., Dias, J., Gerasimou, P., Melanthiou, K., Prokopiou, C.*, et al.* (2017). JAK2 V617F hematopoietic clones are present several years prior to MPN diagnosis and follow different expansion kinetics. Blood Adv *1*, 968-971.

McKerrell, T., Park, N., Moreno, T., Grove, C.S., Ponstingl, H., Stephens, J., Crawley, C., Craig, J., Scott, M.A., Hodkinson, C.*, et al.* (2015). Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. Cell reports *10*, 1239-1245.

McKerrell, T., and Vassiliou, G.S. (2015). Aging as a driver of leukemogenesis. Science translational medicine *7*, 306fs338.

McMahon, K.M., Scielzo, C., Angeloni, N.L., Deiss-Yehiely, E., Scarfo, L., Ranghetti, P., Ma, S., Kaplan, J., Barbaglio, F., Gordon, L.I.*, et al.* (2017). Synthetic high-density lipoproteins as targeted monotherapy for chronic lymphocytic leukemia. Oncotarget *8*, 11219-11227.

McNerney, M.E., Godley, L.A., and Le Beau, M.M. (2017). Therapy-related myeloid neoplasms: when genetics and environment collide. Nature reviews Cancer *17*, 513-527.

Medinger, M., and Passweg, J.R. (2017). Acute myeloid leukaemia genomics. British journal of haematology *179*, 530-542.

Medyouf, H. (2017). The microenvironment in human myeloid malignancies: emerging concepts and therapeutic implications. Blood *129*, 1617-1626.

Megonigal, M.D., Cheung, N.K., Rappaport, E.F., Nowell, P.C., Wilson, R.B., Jones, D.H., Addya, K., Leonard, D.G., Kushner, B.H., Williams, T.M.*, et al.* (2000). Detection of leukemia-associated MLL-GAS7 translocation early during chemotherapy with DNA topoisomerase II inhibitors. Proceedings of the National Academy of Sciences of the United States of America *97*, 2814-2819.

Mehta, P.A., Harris, R.E., Davies, S.M., Kim, M.O., Mueller, R., Lampkin, B., Mo, J., Myers, K., and Smolarek, T.A. (2010). Numerical chromosomal changes and risk of development of myelodysplastic syndrome--acute myeloid leukemia in patients with Fanconi anemia. Cancer genetics and cytogenetics *203*, 180-186.

Menzies, A., Teague, J.W., Butler, A.P., Davies, H., Tarpey, P., Nik-Zainal, S., and Campbell, P.J. (2002). VAGrENT: Variation Annotation Generator. In Current protocols in bioinformatics (John Wiley & Sons, Inc.).

Metzeler, K.H., Herold, T., Rothenberg-Thurley, M., Amler, S., Sauerland, M.C., Gorlich, D., Schneider, S., Konstandin, N.P., Dufour, A., Braundl, K., *et al.* (2016). Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. Blood *128*, 686-698.

Mistry, A.R., Felix, C.A., Whitmarsh, R.J., Mason, A., Reiter, A., Cassinat, B., Parry, A., Walz, C., Wiemels, J.L., Segal, M.R., *et al.* (2005). DNA topoisomerase II in therapy-related acute promyelocytic leukemia. The New England journal of medicine *352*, 1529-1538.

Mitchell, T.J., Turajlic, S., Rowan, A., Nicol, D., Farmery, J.H.R., O'Brien, T., Martincorena, I., Tarpey, P., Angelopoulos, N., Yates, L.R., *et al.* (2018). Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell *173*, 611-623.e617.

Moore, L., Leongamornlert, D., Coorens, T.H., Sanders, M.A., Ellis, P., Dawson, K., Maura, F., Nangalia, J., Tarpey, P.S., Brunner, S.F., *et al.* (2018). The mutational landscape of normal human endometrial epithelium. 505685.

Moran-Crusio, K. (2011). Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer cell *20*, 11-24.

Mori, H., Colman, S.M., Xiao, Z., Ford, A.M., Healy, L.E., Donaldson, C., Hows, J.M., Navarrete, C., and Greaves, M. (2002). Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proceedings of the National Academy of Sciences of the United States of America *99*, 8242-8247.

Morton, L.M., Curtis, R.E., Linet, M.S., Bluhm, E.C., Tucker, M.A., Caporaso, N., Ries, L.A., and Fraumeni, J.F., Jr. (2010). Second malignancy risks after non-Hodgkin's lymphoma and chronic lymphocytic leukemia: differences by lymphoma subtype. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *28*, 4935-4944.

Morton, L.M., Dores, G.M., Schonfeld, S.J., Linet, M.S., Sigel, B.S., Lam, C.J.K., Tucker, M.A., and Curtis, R.E. (2018). Association of Chemotherapy for Solid Tumors With Development of Therapy-Related Myelodysplastic Syndrome or Acute Myeloid Leukemia in the Modern Era. JAMA Oncol.

Morton, L.M., Gibson, T.M., Clarke, C.A., Lynch, C.F., Anderson, L.A., Pfeiffer, R., Landgren, O., Weisenburger, D.D., and Engels, E.A. (2014). Risk of myeloid neoplasms after solid organ transplantation. Leukemia *28*, 2317-2323.

Mouly, E., Ghamlouch, H., Della-Valle, V., Scourzic, L., Quivoron, C., Roos-Weil, D., Pawlikowska, P., Saada, V., Diop, M.K., Lopez, C.K.*, et al.* (2018). B-cell tumor development in Tet2-deficient mice. Blood Adv *2*, 703-714.

Mullighan, C.G. (2014). The genomic landscape of acute lymphoblastic leukemia in children and young adults. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program *2014*, 174-180.

Murai, K., Skrupskelyte, G., Piedrafita, G., Hall, M., Kostiou, V., Ong, S.H., Nagy, T., Cagan, A., Goulding, D., Klein, A.M.*, et al.* (2018). Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations. Cell stem cell *23*, 687-699.e688.

Nachmanson, D., Lian, S., Schmidt, E.K., Hipp, M.J., Baker, K.T., Zhang, Y., Tretiakova, M., Loubet-Senear, K., Kohrn, B.F., Salk, J.J.*, et al.* (2018). Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). Genome research *28*, 1589-1599.

Nangalia, J., Grinfeld, J., and Green, A.R. (2016). Pathogenesis of Myeloproliferative Disorders. Annual review of pathology *11*, 101-126.

Nangalia, J., Mitchell, E., and Green, A.R. (2019). Clonal approaches to understanding the impact of mutations on hematologic disease development. Blood.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D.*, et al.* (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90-94.

Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C., Modlin, L.A., Liu, C.L., Neal, J.W., Wakelee, H.A., Merritt, R.E.*, et al.* (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nature medicine *20*, 548-554.

Newman, A.M., Lovejoy, A.F., Klass, D.M., Kurtz, D.M., Chabon, J.J., Scherer, F., Stehr, H., Liu, C.L., Bratman, S.V., Say, C.*, et al.* (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotech *34*, 547-555.

Ng, A., Ravetto, P.F., Taylor, G.M., Wynn, R.F., and Eden, O.B. (2004). Coexistence of treatment-related MLL cleavage and rearrangement in a child with haemophagocytic lymphohistiocytosis. British journal of cancer *91*, 1990-1992.

Niemeyer, C.M., and Baumann, I. (2011). Classification of childhood aplastic anemia and myelodysplastic syndrome. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program *2011*, 84-89.

Nik-Zainal, S. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat Genet *46*, 487-491.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* (2012). The life history of 21 breast cancers. Cell *149*, 994-1007.

Nordling, C.O. (1953). A new theory on cancer-inducing mechanism. British journal of cancer *7*, 68-72.

Nowell, P., and Hungerford, D. (1960). A minute chromosome in human granulocytic leukemia. Science (New York, NY) *132*, 1497.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science (New York, NY) *194*, 23-28.

O'Brien, S.G., Guilhot, F., Larson, R.A., Gathmann, I., Baccarani, M., Cervantes, F., Cornelissen, J.J., Fischer, T., Hochhaus, A., Hughes, T.*, et al.* (2003). Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. The New England journal of medicine *348*, 994-1004.

O'Quigley, J., Xu, R., and Stare, J. (2005). Explained randomness in proportional hazards models. Statistics in medicine *24*, 479-489.

Oeffinger, K.C., Mertens, A.C., Sklar, C.A., Kawashima, T., Hudson, M.M., Meadows, A.T., Friedman, D.L., Marina, N., Hobbie, W., Kadan-Lottick, N.S.*, et al.* (2006). Chronic health conditions in adult survivors of childhood cancer. The New England journal of medicine *355*, 1572-1582.

Offman, J., Opelz, G., Doehler, B., Cummins, D., Halil, O., Banner, N.R., Burke, M.M., Sullivan, D., Macpherson, P., and Karran, P. (2004). Defective DNA mismatch repair in acute myeloid leukemia/myelodysplastic syndrome after organ transplantation. Blood *104*, 822-828.

Ojha, J., Secreto, C., Rabe, K., Ayres-Silva, J., Tschumper, R., Dyke, D.V., Slager, S., Fonseca, R., Shanafelt, T., Kay, N.*, et al.* (2014). Monoclonal B-cell lymphocytosis is characterized by mutations in CLL putative driver genes and clonal heterogeneity many years before disease progression. Leukemia *28*, 2395-2398.

Ok, C.Y., Patel, K.P., Garcia-Manero, G., Routbort, M.J., Peng, J., Tang, G., Goswami, M., Young, K.H., Singh, R., Medeiros, L.J.*, et al.* (2015). TP53 mutation characteristics in therapy-related myelodysplastic syndromes and acute myeloid leukemia is similar to de novo diseases. Journal of hematology & oncology *8*, 45.

Okosun, J., Bodor, C., Wang, J., Araf, S., Yang, C.Y., Pan, C., Boller, S., Cittaro, D., Bozek, M., Iqbal, S.*, et al.* (2014). Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. Nature genetics *46*, 176-181.

Olcaydu, D., Harutyunyan, A., Jager, R., Berg, T., Gisslinger, B., Pabinger, I., Gisslinger, H., and Kralovics, R. (2009). A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. Nature genetics *41*, 450-454.

Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H., Hasaart, K., de la Fonteijne, L., Varela, I., Camargo, F.D., and van Boxtel, R. (2018). Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell reports *25*, 2308-2316.e2304.

Paget, S. (1889). The distribution of secondary growths in cancer of the breast. The Lancet *133*, 571-573.

Pan, F., Wingo, T.S., Zhao, Z., Gao, R., Makishima, H., Qu, G., Lin, L., Yu, M., Ortega, J.R., Wang, J*., et al.* (2017). Tet2 loss leads to hypermutagenicity in haematopoietic stem/progenitor cells. Nature communications *8*, 15102.

Pandyra, A., Mullen, P.J., Kalkat, M., Yu, R., Pong, J.T., Li, Z., Trudel, S., Lang, K.S., Minden, M.D., Schimmer, A.D*., et al.* (2014). Immediate utility of two approved agents to target both the metabolic mevalonate pathway and its restorative feedback loop. Cancer research *74*, 4772-4782.

Pang, W.W., Price, E.A., Sahoo, D., Beerman, I., Maloney, W.J., Rossi, D.J., Schrier, S.L., and Weissman, I.L. (2011). Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. Proceedings of the National Academy of Sciences of the United States of America *108*, 20012-20017.

Pang, W.W., Schrier, S.L., and Weissman, I.L. (2017). Age-associated changes in human hematopoietic stem cells. Seminars in hematology *54*, 39-42.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N*., et al.* (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. New England Journal of Medicine *374*, 2209-2221.

Park, N., and Vassiliou, G. (2017). Design and Application of Multiplex PCR Seq for the Detection of Somatic Mutations Associated with Myeloid Malignancies. Methods in molecular biology *1633*, 87-99.

Parkin, B., Londono-Joshi, A., Kang, Q., Tewari, M., Rhim, A.D., and Malek, S.N. (2017). Ultrasensitive mutation detection identifies rare residual cells causing acute myelogenous leukemia relapse. The Journal of clinical investigation *127*, 3484-3495.

Parsons, D.W., Roy, A., Yang, Y., Wang, T., Scollon, S., Bergstrom, K., Kerstein, R.A., Gutierrez, S., Petersen, A.K., Bavle, A*., et al.* (2016). Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. JAMA Oncol *2*.

Pastor, V., Hirabayashi, S., Karow, A., Wehrle, J., Kozyra, E.J., Nienhold, R., Ruzaike, G., Lebrecht, D., Yoshimi, A., Niewisch, M*., et al.* (2017). Mutational landscape in children with myelodysplastic syndromes is distinct from adults: specific somatic drivers and novel germline variants. Leukemia *31*, 759-762.

Petljak, M., and Alexandrov, L.B. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis *37*, 531-540.

Petljak, M., Butler, A.P., Bolli, N., Davies, H.R., Knappskog, S., Martin, S., Papaemmanuil, E., Ramakrishna, M., Shlien, A., Simonic, I*., et al.* (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Nature genetics *24*, 52-60.

Petti, A.A., Williams, S.R., Miller, C.A., Fiddes, I.T., Srivatsan, S.N., Chen, D.Y., Fronick, C.C., Fulton, R.S., Church, D.M., and Ley, T.J. (2018). Mutation detection in thousands of acute myeloid leukemia cells using single cell RNA-sequencing.

Pinzaru, A.M., Hom, R.A., Beal, A., Phillips, A.F., Ni, E., Cardozo, T., Nair, N., Choi, J., Wuttke, D.S., Sfeir, A*., et al.* (2016). Telomere Replication Stress Induced by POT1 Inactivation Accelerates Tumorigenesis. Cell reports *15*, 2170-2184.

Pirro, M., Ricciuti, B., Rader, D.J., Catapano, A.L., Sahebkar, A., and Banach, M. (2018). High density lipoprotein cholesterol and cancer: Marker or causative? Progress in lipid research *71*, 54-69.

Poon, S.L., Huang, M.N., Choo, Y., McPherson, J.R., Yu, W., Heng, H.L., Gan, A., Myint, S.S., Siew, E.Y., Ler, L.D*., et al.* (2015). Mutation signatures implicate aristolochic acid in bladder cancer development. Genome medicine *7*, 38.

Potter, N.E., Ermini, L., Papaemmanuil, E., Cazzaniga, G., Vijayaraghavan, G., Titley, I., Ford, A., Campbell, P., Kearney, L., and Greaves, M. (2013). Single-cell mutational profiling and clonal phylogeny in cancer. Genome research *23*, 2115-2125.

Poynter, J.N., Gruber, S.B., Higgins, P.D., Almog, R., Bonner, J.D., Rennert, H.S., Low, M., Greenson, J.K., and Rennert, G. (2005). Statins and the risk of colorectal cancer. The New England journal of medicine *352*, 2184-2192.

Pritchard, C.C., Mateo, J., Walsh, M.F., De Sarkar, N., Abida, W., Beltran, H., Garofalo, A., Gulati, R., Carreira, S., Eeles, R*., et al.* (2016). Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. New England Journal of Medicine *375*, 443-453.

Pui, C.H., Ribeiro, R.C., Hancock, M.L., Rivera, G.K., Evans, W.E., Raimondi, S.C., Head, D.R., Behm, F.G., Mahmoud, M.H., Sandlund, J.T*., et al.* (1991). Acute myeloid leukemia in children treated with epipodophyllotoxins for acute lymphoblastic leukemia. The New England journal of medicine *325*, 1682-1687.

Qu, W., and Zhang, C. (2015). Selecting specific PCR primers with MFEprimer. Methods in molecular biology *1275*, 201-213.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841-842.

Quivoron, C., Couronne, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.H*., et al.* (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer cell *20*, 25-38.

Raine, K.M., Hinton, J., Butler, A.P., Teague, J.W., Davies, H., Tarpey, P., Nik-Zainal, S., and Campbell, P.J. (2015). cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. Current protocols in bioinformatics *52*, 15.17.11-12.

Ramus, S.J., Song, H., Dicks, E., Tyrer, J.P., Rosenthal, A.N., Intermaggio, M.P., Fraser, L., Gentry-Maharaj, A., Hayward, J., Philpott, S.*, et al.* (2015). Germline Mutations in the BRIP1, BARD1, PALB2, and NBN Genes in Women With Ovarian Cancer. Journal of the National Cancer Institute *107*.

Rawstron, A.C., Bennett, F.L., O'Connor, S.J., Kwok, M., Fenton, J.A., Plummer, M., de Tute, R., Owen, R.G., Richards, S.J., Jack, A.S.*, et al.* (2008). Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. The New England journal of medicine *359*, 575-583.

Raza, A., and Galili, N. (2012). The genetic basis of phenotypic heterogeneity in myelodysplastic syndromes. Nature reviews Cancer *12*, 849-859.

Reddy, A., Zhang, J., Davis, N.S., Moffitt, A.B., Love, C.L., Waldrop, A., Leppa, S., Pasanen, A., Meriranta, L., Karjalainen-Lindsberg, M.L.*, et al.* (2017). Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell *171*, 481-494.e415.

Reddy, E.P., Reynolds, R.K., Santos, E., and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature *300*, 149-152.

Reina-Castillon, J., Pujol, R., Lopez-Sanchez, M., Rodriguez-Santiago, B., Aza-Carmona, M., Gonzalez, J.R., Casado, J.A., Bueren, J.A., Sevilla, J., Badel, I.*, et al.* (2017). Detectable clonal mosaicism in blood as a biomarker of cancer risk in Fanconi anemia. Blood Adv *1*, 319-329.

Riboli, E., Hunt, K.J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondiere, U.R., Hemon, B., Casagrande, C., Vignat, J.*, et al.* (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public health nutrition *5*, 1113-1124.

Ridker, P.M., Everett, B.M., Thuren, T., MacFadyen, J.G., Chang, W.H., Ballantyne, C., Fonseca, F., Nicolau, J., Koenig, W., Anker, S.D.*, et al.* (2017). Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. The New England journal of medicine *377*, 1119-1131.

Robinson, B.W., Cheung, N.K., Kolaris, C.P., Jhanwar, S.C., Choi, J.K., Osheroff, N., and Felix, C.A. (2008). Prospective tracing of MLL-FRYL clone with low MEIS1 expression from emergence during neuroblastoma treatment to diagnosis of myelodysplastic syndrome. Blood *111*, 3802-3812.

Rodriguez-Santiago, B., Malats, N., Rothman, N., Armengol, L., Garcia-Closas, M., Kogevinas, M., Villa, O., Hutchinson, A., Earl, J., Marenne, G.*, et al.* (2010). Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. American journal of human genetics *87*, 129-138.

Roerink, S.F., Sasaki, N., Lee-Six, H., Young, M.D., Alexandrov, L.B., Behjati, S., Mitchell, T.J., Grossmann, S., Lightfoot, H., Egan, D.A*., et al.* (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. Nature *556*, 457-462.

Roos, W.P., Thomas, A.D., and Kaina, B. (2016). DNA damage and the balance between survival and death in cancer biology. Nature reviews Cancer *16*, 20-33.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome biology *14*, R51.

Ross, R. (1999). Atherosclerosis--an inflammatory disease. The New England journal of medicine *340*, 115-126.

Rossi, D.J., Bryder, D., Seita, J., Nussenzweig, A., Hoeijmakers, J., and Weissman, I.L. (2007). Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. Nature *447*, 725-729.

Rossi, D.J., Bryder, D., Zahn, J.M., Ahlenius, H., Sonu, R., Wagers, A.J., and Weissman, I.L. (2005). Cell intrinsic alterations underlie hematopoietic stem cell aging. Proceedings of the National Academy of Sciences of the United States of America *102*, 9194-9199.

Rowland, J.H., and Bellizzi, K.M. (2014). Cancer survivorship issues: life after treatment and implications for an aging population. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *32*, 2662-2668.

Rowley, J.D. (1973). Identification of a Translocation with Quinacrine Fluorescence in a Patient with Acute Leukemia. Annales De Genetique *16*, 109-112.

Rowley, J.D. (2001). Chromosome translocations: dangerous liaisons revisited. Nature reviews Cancer *1*, 245-250.

Rowley, J.D. (2008). Chromosomal translocations: revisited yet again. Blood *112*, 2183-2189.

Rozhok, A.I., and DeGregori, J. (2015). Toward an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations. Proceedings of the National Academy of Sciences of the United States of America *112*, 8914-8921.

Rozhok, A.I., Salstrom, J.L., and DeGregori, J. (2014). Stochastic modeling indicates that aging and somatic evolution in the hematopoetic system are driven by non-cell-autonomous processes. Aging (Albany NY) *6*, 1033-1048.

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D.N., Renwick, A., Seal, S., Ramsay, E*., et al.* (2013). Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. Nature *493*, 406-410.

Rubnitz, J.E., Inaba, H., Leung, W.H., Pounds, S., Cao, X., Campana, D., Ribeiro, R.C., and Pui, C.H. (2014). Definition of cure in childhood acute myeloid leukemia. Cancer *120*, 2490-2496.

Sabarinathan, R., Pich, O., Martincorena, I., Rubio-Perez, C., Juul, M., Wala, J., Schumacher, S., Shapira, O., Sidiropoulos, N., Waszak, S.*, et al.* (2017). The whole-genome panorama of cancer drivers. bioRxiv.

Saliba, J., Saint-Martin, C., Di Stefano, A., Lenglet, G., Marty, C., Keren, B., Pasquier, F., Valle, V.D., Secardin, L., Leroy, G.*, et al.* (2015). Germline duplication of ATG2B and GSKIP predisposes to familial myeloid malignancies. Nature genetics *47*, 1131-1140.

Salk, J.J., Loubet-Senear, K., Maritschnegg, E., Valentine, C.C., Williams, L.N., Horvat, R., Vanderstichele, A., Nachmanson, D., Baker, K.T., Emond, M.J.*, et al.* (2018). Ultra-sensitive sequencing for cancer detection reveals progressive clonal selection in normal tissue over a century of human lifespan.

Salvagno, G.L., Sanchis-Gomar, F., Picanza, A., and Lippi, G. (2015). Red blood cell distribution width: A simple parameter with multiple clinical applications. Critical reviews in clinical laboratory sciences *52*, 86-105.

Sano, S., Oshima, K., Wang, Y., MacLauchlan, S., Katanasaka, Y., Sano, M., Zuriaga, M.A., Yoshiyama, M., Goukassian, D., Cooper, M.A.*, et al.* (2018a). Tet2-Mediated Clonal Hematopoiesis Accelerates Heart Failure Through a Mechanism Involving the IL-1beta/NLRP3 Inflammasome. J Am Coll Cardiol *71*, 875-886.

Sano, S., Wang, Y., and Walsh, K. (2018b). Clonal Hematopoiesis and Its Impact on Cardiovascular Disease. Circulation journal : official journal of the Japanese Circulation Society *83*, 2-11.

Savola, P., Kelkka, T., Rajala, H.L., Kuuliala, A., Kuuliala, K., Eldfors, S., Ellonen, P., Lagstrom, S., Lepisto, M., Hannunen, T.*, et al.* (2017). Somatic mutations in clonally expanded cytotoxic T lymphocytes in patients with newly diagnosed rheumatoid arthritis. Nature communications *8*, 15869.

Schick, U.M., McDavid, A., Crane, P.K., Weston, N., Ehrlich, K., Newton, K.M., Wallace, R., Bookman, E., Harrison, T., Aragaki, A.*, et al.* (2013). Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. PloS one *8*, e59823.

Schmitt, M.W., Fox, E.J., Prindle, M.J., Reid-Bayliss, K.S., True, L.D., Radich, J.P., and Loeb, L.A. (2015). Sequencing small genomic targets with high efficiency and extreme accuracy. Nature methods *12*, 423-425.

Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences of the United States of America *109*, 14508-14513.

Schmitt, M.W., Loeb, L.A., and Salk, J.J. (2016). The influence of subclonal resistance mutations on targeted cancer therapy. Nature reviews Clinical oncology *13*, 335-347.

Schrader, K.A., Cheng, D.T., Joseph, V., Prasad, M., Walsh, M., Zehir, A., Ni, A., Thomas, T., Benayed, R., Ashraf, A.*, et al.* (2016). Germline Variants in Targeted Tumor Sequencing Using Matched Normal DNA. JAMA Oncol *2*, 104-111.

Schreiber, R.D., Old, L.J., and Smyth, M.J. (2011). Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science (New York, NY) *331*, 1565-1570.

Schulz, E., Kashofer, K., Heitzer, E., Mhatre, K.N., Speicher, M.R., Hoefler, G., and Sill, H. (2015). Preexisting TP53 mutation in therapy-related acute myeloid leukemia. Annals of hematology *94*, 527-529.

Schulz, E., Valentin, A., Ulz, P., Beham-Schmid, C., Lind, K., Rupp, V., Lackner, H., Wolfler, A., Zebisch, A., Olipitz, W.*, et al.* (2012). Germline mutations in the DNA damage response genes BRCA1, BRCA2, BARD1 and TP53 in patients with therapy related myeloid neoplasms. Journal of medical genetics *49*, 422-428.

Schwartsmann, G., Brondani da Rocha, A., Berlinck, R.G., and Jimeno, J. (2001). Marine organisms as a source of new anticancer agents. The Lancet Oncology *2*, 221-225.

Scott, D.W., and Gascoyne, R.D. (2014). The tumour microenvironment in B cell lymphomas. Nature reviews Cancer *14*, 517-534.

SEER (2018). Cancer Stat Facts: Leukemia - Acute Myeloid Leukemia (AML).

Shaw, A.C., Goldstein, D.R., and Montgomery, R.R. (2013). Age-dependent dysregulation of innate immunity. Nature reviews Immunology *13*, 875-887.

Shen, L., Shi, Q., and Wang, W. (2018). Double agents: genes with both oncogenic and tumor-suppressor functions. Oncogenesis *7*, 25.

Shen, R., and Seshan, V.E. (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic acids research *44*, e131.

Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., Hang, X., Wang, X., Zhao, D., and Zhang, C. (2010). MPprimer: a program for reliable multiplex PCR primer design. BMC bioinformatics *11*, 143.

Shih, C., Padhy, L.C., Murray, M., and Weinberg, R.A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. Nature *290*, 261-264.

Shlien, A., Campbell, B.B., de Borja, R., Alexandrov, L.B., Merico, D., Wedge, D., Van Loo, P., Tarpey, P.S., Coupland, P., Behjati, S.*, et al.* (2015). Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. Nature genetics *47*, 257-262.

Shlush, L.I., Mitchell, A., Heisler, L., Abelson, S., Ng, S.W.K., Trotman-Grant, A., Medeiros, J.J.F., Rao-Bhatia, A., Jaciw-Zurakowsky, I., Marke, R.*, et al.* (2017). Tracing the origins of relapse in acute myeloid leukaemia to stem cells. Nature *547*, 104-108.

Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., Kennedy, J.A., Schimmer, A.D., Schuh, A.C., Yee, K.W*., et al.* (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328-333.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw *39*, 1-13.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics (Oxford, England) *21*, 3940-3941.

Smith, A.L., Alirezaie, N., Connor, A., Chan-Seng-Yue, M., Grant, R., Selander, I., Bascunana, C., Borgida, A., Hall, A., Whelan, T*., et al.* (2016). Candidate DNA repair susceptibility genes identified by exome sequencing in high-risk pancreatic cancer. Cancer letters *370*, 302-312.

Smith, M.L., Cavenagh, J.D., Lister, T.A., and Fitzgibbon, J. (2004). Mutation of CEBPA in familial acute myeloid leukemia. The New England journal of medicine *351*, 2403-2407.

Smith, S.M., Le Beau, M.M., Huo, D., Karrison, T., Sobecks, R.M., Anastasi, J., Vardiman, J.W., Rowley, J.D., and Larson, R.A. (2003). Clinical-cytogenetic associations in 306 patients with therapy-related myelodysplasia and myeloid leukemia: the University of Chicago series. Blood *102*, 43-52.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature reviews Cancer *18*, 696-705.

Sperling, A.S., Gibson, C.J., and Ebert, B.L. (2017). The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. Nature reviews Cancer *17*, 5-19.

Stanley, N., Olson, T.S., and Babushok, D.V. (2017). Recent advances in understanding clonal haematopoiesis in aplastic anaemia. British journal of haematology *177*, 509-525.

Steensma, D.P., Bejar, R., Jaiswal, S., Lindsley, R.C., Sekeres, M.A., Hasserjian, R.P., and Ebert, B.L. (2015). Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood *126*, 9-16.

Stein, E.M. (2015). Molecularly targeted therapies for acute myeloid leukemia. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program *2015*, 579-583.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat *21*, 577-581.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A*., et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R.*, et al.* (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404.

Storr, S.J., Safuan, S., Ahmad, N., El-Refaee, M., Jackson, A.M., and Martin, S.G. (2017). Macrophage-derived interleukin-1beta promotes human breast cancer cell migration and lymphatic adhesion in vitro. Cancer immunology, immunotherapy : CII *66*, 1287-1294.

Strati, P., and Shanafelt, T.D. (2015). Monoclonal B-cell lymphocytosis and early-stage chronic lymphocytic leukemia: diagnosis, natural history, and risk stratification. Blood *126*, 454-462.

Stratton, M.R. (2011). Exploring the genomes of cancer cells: progress and promise. Science (New York, NY) *331*, 1553-1558.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724.

Stutzman-Engwall, K.J., and Hutchinson, C.R. (1989). Multigene families for anthracycline antibiotic production in Streptomyces peucetius. Proceedings of the National Academy of Sciences of the United States of America *86*, 3135-3139.

Suda, K., Nakaoka, H., Yoshihara, K., Ishiguro, T., Tamura, R., Mori, Y., Yamawaki, K., Adachi, S., Takahashi, T., Kase, H.*, et al.* (2018). Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. Cell reports *24*, 1777-1789.

Sun, R., Hu, Z., Sottoriva, A., Graham, T.A., Harpak, A., Ma, Z., Fischer, J.M., Shibata, D., and Curtis, C. (2017). Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nature genetics *49*, 1015-1024.

Swerdlow, S.H., Campo, E., Pileri, S.A., Harris, N.L., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G.A., Zelenetz, A.D.*, et al.* (2016). The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood *127*, 2375-2390.

Tabin, C.J., Bradley, S.M., Bargmann, C.I., Weinberg, R.A., Papageorge, A.G., Scolnick, E.M., Dhar, R., Lowy, D.R., and Chang, E.H. (1982). Mechanism of activation of a human oncogene. Nature *300*, 143-149.

Takahashi, K., Wang, F., Kantarjian, H., Doss, D., Khanna, K., Thompson, E., Zhao, L., Patel, K., Neelapu, S., Gumbs, C.*, et al.* (2017). Preleukaemic clonal haemopoiesis and risk of therapy-related myeloid neoplasms: a case-control study. The Lancet Oncology *18*, 100-111.

Tartaglia, M., Niemeyer, C.M., Fragale, A., Song, X., Buechner, J., Jung, A., Hahlen, K., Hasle, H., Licht, J.D., and Gelb, B.D. (2003). Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. Nature genetics *34*, 148-150.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E.*, et al.* (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic acids research *47*, D941-D947.

TCGA, Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A., Hoadley, K., Triche, T.J., Jr., Laird, P.W.*, et al.* (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England journal of medicine *368*, 2059-2074.

Therneau, T., and Grambsch, P.M. (2000). Modeling Survival Data: Extending the Cox Model, 1st edn edn (New York: Springer-Verlag).

Thompson, D., Genovese, G., Halvardson, J., Ulirsch, J., Wright, D., Terao, C., Davidsson, O., Day, F., Sulem, P., Jiang, Y.*, et al.* (2019). Genetic predisposition to mosaic Y chromosome loss in blood is associated with genomic instability in other tissues and susceptibility to non-haematological cancers. 514026.

Tiacci, E., Trifonov, V., Schiavoni, G., Holmes, A., Kern, W., Martelli, M.P., Pucciarini, A., Bigerna, B., Pacini, R., Wells, V.A.*, et al.* (2011). BRAF mutations in hairy-cell leukemia. The New England journal of medicine *364*, 2305-2315.

Turcotte, L.M., Neglia, J.P., Reulen, R.C., Ronckers, C.M., van Leeuwen, F.E., Morton, L.M., Hodgson, D.C., Yasui, Y., Oeffinger, K.C., and Henderson, T.O. (2018). Risk, Risk Factors, and Surveillance of Subsequent Malignant Neoplasms in Survivors of Childhood Cancer: A Review. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *36*, 2145-2152.

Tyner, J.W., Tognon, C.E., Bottomly, D., Wilmot, B., Kurtz, S.E., Savage, S.L., Long, N., Schultz, A.R., Traer, E., Abel, M.*, et al.* (2018). Functional genomic landscape of acute myeloid leukaemia. Nature *562*, 526-531.

Uno, H., Cai, T., Tian, L., and Wei, L.J. (2007). Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models. Journal of the American Statistical Association *102*, 527-537.

Van den Neste, E., Andre, M., Gastinne, T., Stamatoullas, A., Haioun, C., Belhabri, A., Reman, O., Casasnovas, O., Guesquieres, H., Verhoef, G.*, et al.* (2018). Phase II study of oral JAK1/JAK2 inhibitor ruxolitinib in advanced relapsed/refractory Hodgkin lymphoma. Haematologica.

Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B.*, et al.* (2010). Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America *107*, 16910-16915.

Van Vlierberghe, P., and Ferrando, A. (2012). The molecular basis of T cell acute lymphoblastic leukemia. The Journal of clinical investigation *122*, 3398-3406.

Vannucchi, A.M., and Harrison, C. (2016). Emerging treatments for classical myeloproliferative neoplasms. Blood.

Vattathil, S., and Scheet, P. (2016). Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. American journal of human genetics *98*, 571-578.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science (New York, NY) *291*, 1304-1351.

Vickers, A.J. (2011). Prediction models in cancer care. CA: a cancer journal for clinicians *61*, 315-326.

Vitols, S., Angelin, B., Ericsson, S., Gahrton, G., Juliusson, G., Masquelier, M., Paul, C., Peterson, C., Rudling, M., Soderberg-Reid, K.*, et al.* (1990). Uptake of low density lipoproteins by human leukemic cells in vivo: relation to plasma lipoprotein levels and possible relevance for selective chemotherapy. Proceedings of the National Academy of Sciences of the United States of America *87*, 2598-2602.

Vitols, S., Gahrton, G., Bjorkholm, M., and Peterson, C. (1985). Hypocholesterolaemia in malignancy due to elevated low-density-lipoprotein-receptor activity in tumour cells: evidence from studies in patients with leukaemia. Lancet (London, England) *2*, 1150-1154.

Vitols, S., Gahrton, G., Ost, A., and Peterson, C. (1984). Elevated low density lipoprotein receptor activity in leukemic cells with monocytic differentiation. Blood *63*, 1186-1193.

Vitols, S., Peterson, C., Larsson, O., Holm, P., and Aberg, B. (1992). Elevated uptake of low density lipoproteins by human lung cancer tissue in vivo. Cancer research *52*, 6244-6247.

Voso, M.T., Fabiani, E., Zang, Z., Fianchi, L., Falconi, G., Padella, A., Martini, M., Li Zhang, S., Santangelo, R., Larocca, L.M.*, et al.* (2015). Fanconi anemia gene variants in therapy-related myeloid neoplasms. Blood cancer journal *5*, e323.

Wander, S.A., Levis, M.J., and Fathi, A.T. (2014). The evolving role of FLT3 inhibitors in acute myeloid leukemia: quizartinib and beyond. Therapeutic advances in hematology *5*, 65-77.

Wang, Q., He, Z., Huang, M., Liu, T., Wang, Y., Xu, H., Duan, H., Ma, P., Zhang, L., Zamvil, S.S.*, et al.* (2018). Vascular niche IL-6 induces alternative macrophage activation in glioblastoma through HIF-2alpha. Nature communications *9*, 559.

Wang, X., Oldani, M.J., Zhao, X., Huang, X., and Qian, D. (2014). A review of cancer risk prediction models with genetic variants. Cancer informatics *13*, 19-28.

Waterman, J., Rybicki, L., Bolwell, B., Copelan, E., Pohlman, B., Sweetenham, J., Dean, R., Sobecks, R., Andresen, S., and Kalaycio, M. (2012). Fludarabine as a risk factor for poor stem cell harvest, treatment-related MDS and AML in follicular lymphoma patients after autologous hematopoietic cell transplantation. Bone Marrow Transplant *47*, 488-493.

Wegert, J., Vokuhl, C., Collord, G., Del Castillo Velasco-Herrera, M., Farndon, S.J., Guzzo, C., Jorgensen, M., Anderson, J., Slater, O., Duncan, C.*, et al.* (2018). Recurrent intragenic rearrangements of EGFR and BRAF in soft tissue tumors of infants. Nature communications *9*, 2378.

Welch, J.S. (2014). Mutation position within evolutionary subclonal architecture in AML. Seminars in hematology *51*, 273-281.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J.*, et al.* (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264-278.

Wolach, O., Sellar, R.S., Martinod, K., Cherpokova, D., McConkey, M., Chappell, R.J., Silver, A.J., Adams, D., Castellano, C.A., Schneider, R.K.*, et al.* (2018). Increased neutrophil extracellular trap formation promotes thrombosis in myeloproliferative neoplasms. Science translational medicine *10*.

Wong, T.N., Miller, C.A., Klco, J.M., Petti, A., Demeter, R., Helton, N.M., Li, T., Fulton, R.S., Heath, S.E., Mardis, E.R.*, et al.* (2015a). Rapid expansion of pre-existing non-leukemic hematopoietic clones frequently follows induction therapy for de novo AML. Blood.

Wong, T.N., Ramsingh, G., Young, A.L., Miller, C.A., Touma, W., Welch, J.S., Lamprecht, T.L., Shen, D., Hundal, J., Fulton, R.S.*, et al.* (2015b). Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. Nature *518*, 552-555.

Woyach, J.A., and Johnson, A.J. (2015). Targeted therapies in CLL: mechanisms of resistance and strategies for management. Blood *126*, 471-477.

Wright, D.J., Day, F.R., Kerrison, N.D., Zink, F., Cardona, A., Sulem, P., Thompson, D.J., Sigurjonsdottir, S., Gudbjartsson, D.F., Helgason, A.*, et al.* (2017). Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. Nature genetics *49*, 674-679.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A.*, et al.* (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nature medicine *20*, 1472-1478.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A.*, et al.* (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114-1117.

Yahata, T., Takanashi, T., Muguruma, Y., Ibrahim, A.A., Matsuzawa, H., Uno, T., Sheng, Y., Onizuka, M., Ito, M., Kato, S.*, et al.* (2011). Accumulation of oxidative DNA damage restricts the self-renewal capacity of human hematopoietic stem cells. Blood *118*, 2941-2950.

Yalcin, B., Kremer, L.C., and van Dalen, E.C. (2015). High-dose chemotherapy and autologous haematopoietic stem cell rescue for children with high-risk neuroblastoma. The Cochrane database of systematic reviews, Cd006301.

Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protocols *10*, 1556-1566.

Yates, J.W., Wallace, H.J., Jr., Ellison, R.R., and Holland, J.F. (1973). Cytosine arabinoside (NSC-63878) and daunorubicin (NSC-83142) therapy in acute nonlymphocytic leukemia. Cancer chemotherapy reports *57*, 485-488.

Yates, L.R., and Campbell, P.J. (2012). Evolution of the cancer genome. Nature reviews Genetics *13*, 795-806.

Ying, Z., Sandoval, M., and Beronja, S. (2018). Oncogenic activation of PI3K induces progenitor cell differentiation to suppress epidermal growth. Nature cell biology *20*, 1256-1266.

Yizhak, K., Aguet, F., Kim, J., Hess, J., Kubler, K., Grimsby, J., Frazer, R., Zhang, H., Haradhvala, N., Rosebrock, D.*, et al.* (2018). A comprehensive analysis of RNA sequences reveals macroscopic somatic clonal expansion across normal tissues. 416339.

Yokoyama, A., Kakiuchi, N., Yoshizato, T., Nannya, Y., Suzuki, H., Takeuchi, Y., Shiozawa, Y., Sato, Y., Aoki, K., Kim, S.K.*, et al.* (2019). Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature *565*, 312-317.

Yoshizato, T., Dumitriu, B., Hosokawa, K., Makishima, H., Yoshida, K., Townsley, D., Sato-Otsubo, A., Sato, Y., Liu, D., Suzuki, H.*, et al.* (2015). Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia. The New England journal of medicine *373*, 35-47.

Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nature communications *7*, 12484.

Zahn, L.M. (2016). Unleashing the power of precision medicine. Science (New York, NY) *354*, 1546-1548.

Zapata, L., Pich, O., Serrano, L., Kondrashov, F.A., Ossowski, S., and Schaefer, M.H. (2018). Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. Genome biology *19*, 67.

Zech, L., Haglund, U., Nilsson, K., and Klein, G. (1976). Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. International journal of cancer Journal international du cancer *17*, 47-56.

Zhang, J., Walsh, M.F., Wu, G., Edmonson, M.N., Gruber, T.A., Easton, J., Hedges, D., Ma, X., Zhou, X., Yergeau, D.A.*, et al.* (2015). Germline Mutations in Predisposition Genes in Pediatric Cancer. The New England journal of medicine *373*, 2336-2346.

Zhou, J., Li, Y.S., Wang, K.C., and Chien, S. (2014). Epigenetic Mechanism in Regulation of Endothelial Function by Disturbed Flow: Induction of DNA Hypermethylation by DNMT1. Cellular and molecular bioengineering *7*, 218-224.

Zhou, W., Machiela, M.J., Freedman, N.D., Rothman, N., Malats, N., Dagnall, C., Caporaso, N., Teras, L.T., Gaudet, M.M., Gapstur, S.M.*, et al.* (2016). Mosaic loss of chromosome Y is associated with common variation near TCL1A. Nature genetics *48*, 563-568.

Zhu, L., Finkelstein, D., Gao, C., Shi, L., Wang, Y., Lopez-Terrada, D., Wang, K., Utley, S., Pounds, S., Neale, G.*, et al.* (2016). Multi-organ Mapping of Cancer Risk. Cell *166*, 1132-1146.e1137.

Zink, F., Stacey, S.N., Norddahl, G.L., Frigge, M.L., Magnusson, O.T., Jonsdottir, I., Thorgeirsson, T.E., Sigurdsson, A., Gudjonsson, S.A., Gudmundsson, J.*, et al.* (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood *130*, 742-752.

Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science (New York, NY) *338*, 1622-1626.

Zuna, J., Madzo, J., Krejci, O., Zemanova, Z., Kalinova, M., Muzikova, K., Zapotocky, M., Starkova, J., Hrusak, O., Horak, J.*, et al.* (2011). ETV6/RUNX1 (TEL/AML1) is a frequent prenatal first hit in childhood leukemia. Blood *117*, 368-369; author reply 370-361.

Appendices

Appendix 1: Discovery cohort pre-AML and control sample information

Sample	Group	Age at sampling (years)	Follow-up (years)	Gender
EPIC_0001	Control	58.6	14.9	male
EPIC_0002	Control	55.2	14.7	male
EPIC_0003	Control	60.8	14.1	female
EPIC_0004	Control	62.2	14.1	female
EPIC_0005	pre-AML	62.9	0.6	female
EPIC_0006	pre-AML	60.8	10.9	female
EPIC_0007	Control	62.6	14.1	female
EPIC_0008	Control	62.4	14	female
EPIC_0009	Control	62.4	14	female
EPIC_0010	Control	55.3	14.5	female
EPIC_0011	Control	55	14.4	female
EPIC_0012	Control	51.4	13	male
EPIC_0013	Control	52	13	male
EPIC_0014	pre-AML	55.8	12.4	female
EPIC_0015	pre-AML	46.5	12.1	female
EPIC_0016	Control	49.1	13.8	female
EPIC_0017	Control	46	13.7	female
EPIC_0018	Control	46.8	13.7	female
EPIC_0020	Control	46.2	13.7	female
EPIC_0021	Control	56.1	14.7	male
EPIC_0022	Control	57.1	13.6	female
EPIC_0023	Control	41.1	13.1	female
EPIC_0024	Control	41.6	9.1	female
EPIC_0025	Control	41.7	12.9	female
EPIC_0026	Control	41.6	12.9	female
EPIC_0027	Control	63.7	8.2	female
EPIC_0028	Control	63.7	12.7	female
EPIC_0029	Control	50	13.5	female
EPIC_0030	Control	49.8	13.3	female
EPIC_0031	Control	57.3	12	male
EPIC_0032	Control	57.9	12	male
EPIC_0033	Control	62.3	11.8	female
EPIC_0034	Control	55	14.7	male
EPIC_0035	Control	55.4	14.7	male
EPIC_0036	Control	55.7	14.7	male
EPIC_0037	pre-AML	55.8	3.2	male
EPIC_0038	Control	49.5	13.7	female
EPIC_0039	Control	58.2	14.2	female
EPIC_0040	pre-AML	58.5	10	female
EPIC_0041	Control	58.7	14.1	female
EPIC_0042	Control	59.3	14.1	male
EPIC_0043	Control	58.1	14.1	female
EPIC_0044	pre-AML	58.3	8.3	male
EPIC_0045	Control	58.7	12.8	male
EPIC_0046	Control	54.3	14	male
EPIC_0047	pre-AML	54	2.8	male
EPIC_0048	Control	55	13.9	male
EPIC_0049	Control	50.4	13.8	male
EPIC_0050	Control	50.2	13.7	male
EPIC_0051	pre-AML	50.2	6	male
EPIC_0052	Control	50.6	13.5	male
EPIC_0053	Control	63.8	7.6	female
EPIC_0054	Control	51.1	12.6	male
EPIC_0055	Control	48.1	12.5	male
EPIC_0056	Control	55.6	14.2	female
EPIC_0057	Control	55.5	14.2	female
EPIC_0058	Control	58.6	14	male
EPIC_0059	Control	64.2	9.3	male
EPIC_0060	Control	64.3	8.6	male
EPIC_0061	Control	64.8	9.9	male
EPIC_0062	pre-AML	64.9	1.8	male
EPIC_0063	Control	49.1	13.7	female
EPIC_0064	pre-AML	57.2	3.9	female
EPIC_0065	Control	57.8	13.4	female
EPIC_0066	Control	57.7	10.4	female
EPIC_0067	pre-AML	66.5	10.7	female
EPIC_0068	Control	60.8	13.8	female
EPIC_0069	Control	73.8	13.7	female
EPIC_0070	Control	60.4	13.6	female
EPIC_0071	Control	49.5	13.2	male
EPIC_0072	Control	48.8	12.5	male
EPIC_0073	Control	55.9	12.1	female
EPIC_0074	Control	55.1	12.1	female
EPIC_0075	pre-AML	56	5.8	female
EPIC_0076	Control	63.4	12.5	female
EPIC_0077	Control	56.9	14.3	male
EPIC_0078	Control	56.5	14.3	male
EPIC_0079	Control	56.6	14.2	male
EPIC_0080	Control	52.6	14.2	male
EPIC_0081	pre-AML	52.6	8.8	male
EPIC_0082	Control	52.2	14.1	male
EPIC_0083	Control	52.8	14.1	male

EPIC_0084	Control	55.7	12	female
EPIC_0085	Control	48.5	11.9	female
EPIC_0086	Control	59.6	11.8	female
EPIC_0087	pre-AML	48.9	8.4	female
EPIC_0088	Control	59.3	11.8	female
EPIC_0089	Control	59.4	11.7	female
EPIC_0090	Control	48.6	11.7	female
EPIC_0091	Control	48.9	11.7	female
EPIC_0092	Control	64.6	11.8	male
EPIC_0093	Control	57	12.9	male
EPIC_0094	Control	56.3	12.9	male
EPIC_0095	Control	52.9	12.9	female
EPIC_0096	pre-AML	56.7	12	male
EPIC_0097	Control	56.9	12.6	male
EPIC_0098	Control	55.4	12.6	male
EPIC_0099	pre-AML	56.2	7.7	female
EPIC_0100	Control	56.2	12.8	female
EPIC_0101	Control	52.7	12.8	female
EPIC_0102	Control	52.7	12.7	female
EPIC_0103	Control	53	12.4	female
EPIC_0104	Control	52.2	12.4	female
EPIC_0105	Control	56	12.5	male
EPIC_0106	Control	55.4	12.4	male
EPIC_0107	Control	73.9	10.7	female
EPIC_0108	Control	66.1	13.6	female
EPIC_0109	Control	66	13.6	female
EPIC_0110	Control	66.2	11.4	female
EPIC_0111	Control	70.1	13.5	male
EPIC_0112	Control	60.5	13.4	female
EPIC_0113	Control	49.9	12.6	male
EPIC_0114	Control	67.1	13.3	female
EPIC_0115	Control	67.1	13.2	female
EPIC_0116	Control	55.4	12.5	female
EPIC_0117	Control	67.5	13.1	female
EPIC_0118	Control	68	13	female
EPIC_0119	Control	68.7	12.8	female
EPIC_0120	Control	44.9	10	male
EPIC_0121	Control	44.5	9.8	male
EPIC_0122	Control	44.2	9.8	male
EPIC_0123	Control	63.2	9.8	male
EPIC_0124	Control	63.7	9.7	male
EPIC_0125	Control	55.8	12.9	female
EPIC_0126	Control	55.3	12.8	female
EPIC_0127	Control	55.5	12.5	female
EPIC_0128	Control	43.5	10.7	male
EPIC_0129	Control	56	11	male
EPIC_0130	Control	56.5	11.1	male
EPIC_0131	Control	56.3	11.5	male
EPIC_0132	pre-AML	56.1	9.6	male
EPIC_0133	Control	56.5	10.9	male
EPIC_0134	Control	43.2	11.3	male
EPIC_0135	Control	43.2	11.1	male
EPIC_0136	Control	61.1	8.1	male
EPIC_0137	Control	56.2	8.1	female
EPIC_0138	Control	56.8	8.1	female
EPIC_0139	Control	61.5	8.1	male
EPIC_0140	Control	61.6	8.1	male
EPIC_0141	pre-AML	60.5	4.5	male
EPIC_0142	Control	60.5	8	male
EPIC_0143	Control	56.5	8.2	female
EPIC_0144	Control	60	7.9	male
EPIC_0145	Control	60.2	8	male
EPIC_0146	Control	53.8	8.2	male
EPIC_0147	pre-AML	53	8.1	male
EPIC_0148	Control	43.3	10.9	male
EPIC_0149	Control	61.6	10.8	male
EPIC_0150	Control	50.6	12.7	female
EPIC_0151	Control	54.4	12.8	female
EPIC_0152	Control	54.9	12.7	female
EPIC_0153	Control	50.3	12.3	female
EPIC_0154	Control	46.4	12.3	male
EPIC_0155	Control	46.4	12.3	male
EPIC_0156	Control	50.6	12.3	female
EPIC_0157	Control	50.6	12.2	female
EPIC_0158	Control	62.6	12.1	male
EPIC_0159	Control	62.4	11.8	male
EPIC_0160	Control	36.7	11.7	female
EPIC_0161	Control	36.6	11.6	female
EPIC_0162	pre-AML	36.8	2.9	female
EPIC_0163	Control	36.1	11.4	female
EPIC_0164	Control	36.2	11.4	female
EPIC_0165	pre-AML	58.9	11.1	male
EPIC_0166	Control	58.4	12.7	male
EPIC_0167	Control	58.2	12.6	male
EPIC_0168	Control	60.6	7.9	female
EPIC_0169	Control	60.5	11.6	female
EPIC_0170	Control	58.4	11.9	female

EPIC_0171	pre-AML	54	8.7	male
EPIC_0172	Control	54.9	12.5	male
EPIC_0174	Control	54.1	12.4	male
EPIC_0175	Control	58.6	12.6	male
EPIC_0176	pre-AML	64.5	4.1	male
EPIC_0177	Control	64.2	13.6	male
EPIC_0178	Control	64.5	13.4	male
EPIC_0179	Control	59	13.2	female
EPIC_0180	Control	59.6	13.1	female
EPIC_0181	Control	40	13	female
EPIC_0182	Control	39.2	12.9	female
EPIC_0183	Control	50.7	12.9	female
EPIC_0184	Control	59.4	12.8	female
EPIC_0185	Control	56.3	10.7	female
EPIC_0186	Control	56.4	10.6	female
EPIC_0187	Control	56.3	10.6	female
EPIC_0188	Control	56.3	10.6	female
EPIC_0189	Control	50.7	12.8	female
EPIC_0190	Control	50.1	12.8	female
EPIC_0191	Control	39.2	12.7	female
EPIC_0192	Control	50.2	12.6	female
EPIC_0193	Control	56.4	12.7	female
EPIC_0194	pre-AML	56.1	8	female
EPIC_0195	Control	52.2	12.7	female
EPIC_0196	Control	55.4	12.3	female
EPIC_0197	Control	55.8	12.1	female
EPIC_0198	Control	48.2	12.1	female
EPIC_0199	Control	68.6	12.6	female
EPIC_0200	Control	57	12.6	female
EPIC_0201	Control	69	12.6	female
EPIC_0202	Control	52.8	12.6	female
EPIC_0203	Control	56.2	12.5	female
EPIC_0204	Control	52.8	12.5	female
EPIC_0205	Control	55.4	12.1	female
EPIC_0206	Control	48.4	12.1	female
EPIC_0207	Control	69.5	11.9	female
EPIC_0208	Control	67.7	11.8	female
EPIC_0209	Control	48.9	11.8	female
EPIC_0210	Control	58.5	12.1	female
EPIC_0211	Control	58.8	11.8	female
EPIC_0212	pre-AML	64.2	11	male
EPIC_0213	Control	64.8	11.8	male
EPIC_0214	Control	46.9	12.1	male
EPIC_0215	Control	46.7	12	male
EPIC_0216	Control	46.9	12	male
EPIC_0217	Control	46.6	13.6	male
EPIC_0218	Control	55.3	11.7	male
EPIC_0219	pre-AML	67.8	9.5	female
EPIC_0220	Control	67.3	11.5	female
EPIC_0221	Control	69.1	11.5	female
EPIC_0222	Control	58.4	12.3	male
EPIC_0223	pre-AML	74.3	1.8	female
EPIC_0224	Control	69.6	11.5	female
EPIC_0225	Control	69.6	9.7	female
EPIC_0226	Control	64.4	12.6	female
EPIC_0227	Control	74.4	10.3	female
EPIC_0228	Control	55.1	11.7	male
EPIC_0229	Control	37	10.9	female
EPIC_0230	Control	69.8	13.3	female
EPIC_0231	Control	70	11.7	female
EPIC_0232	Control	70.8	11.8	female
EPIC_0233	Control	64.3	11.9	female
EPIC_0234	pre-AML	69.9	9.2	female
EPIC_0235	Control	74.2	8	female
EPIC_0236	Control	58.2	12.4	male
EPIC_0237	Control	58.6	12.1	male
EPIC_0238	Control	52.1	12.1	female
EPIC_0239	Control	67.4	12.1	female
EPIC_0240	Control	52.8	12.1	female
EPIC_0241	Control	67.1	12	female
EPIC_0242	Control	68.9	11.6	female
EPIC_0243	Control	38.4	11.4	female
EPIC_0244	Control	38.9	11.2	female
EPIC_0245	Control	38.6	11.2	female
EPIC_0246	pre-AML	39	4.9	female
EPIC_0247	Control	68.2	11.4	female
EPIC_0248	Control	68.4	11.4	female
EPIC_0249	pre-AML	69	4.7	female
EPIC_0250	Control	43.6	8.8	female
EPIC_0251	Control	70.5	13.3	female
EPIC_0252	Control	36.5	12.1	female
EPIC_0253	Control	46	11.8	male
EPIC_0254	Control	70.6	12	female
EPIC_0255	pre-AML	36.2	8.1	female
EPIC_0256	Control	43.3	11.8	female
EPIC_0258	Control	36.2	13.5	female
EPIC_0259	Control	43.2	11.3	female

EPIC_0260	Control	36.7	8.1	female
EPIC_0261	pre-AML	58.3	7	male
EPIC_0262	Control	66.1	11.2	female
EPIC_0263	Control	66.4	11.2	female
EPIC_0264	Control	55.8	11.2	female
EPIC_0265	Control	71.1	11	male
EPIC_0266	Control	55.6	11	female
EPIC_0267	Control	54.3	10.3	female
EPIC_0268	Control	54.1	10.3	female
EPIC_0269	pre-AML	54.5	9.4	female
EPIC_0270	Control	54.7	10.1	female
EPIC_0271	pre-AML	56.4	8.5	female
EPIC_0272	Control	54.7	10	female
EPIC_0273	Control	56.2	9.9	female
EPIC_0274	Control	56.1	9.9	female
EPIC_0275	Control	56.2	9.9	female
EPIC_0276	Control	43.5	10.3	female
EPIC_0277	Control	42.6	12.2	female
EPIC_0278	Control	42.8	12	female
EPIC_0279	pre-AML	42.4	9.8	female
EPIC_0280	Control	42.2	11.9	female
EPIC_0281	Control	57.9	12.3	female
EPIC_0282	Control	57.2	12.1	female
EPIC_0283	Control	36.9	10.9	female
EPIC_0284	Control	36.5	10.9	female
EPIC_0285	Control	68.6	4	female
EPIC_0286	Control	51	10.2	male
EPIC_0287	Control	51.1	10.1	male
EPIC_0288	Control	51.1	9.9	male
EPIC_0289	Control	72.6	12.7	male
EPIC_0290	Control	72.6	11.9	male
EPIC_0291	Control	72.8	8.5	male
EPIC_0292	Control	68.4	10.9	female
EPIC_0293	Control	68.7	10.8	female
EPIC_0294	Control	63.4	10.8	female
EPIC_0295	Control	63.2	10.8	female
EPIC_0296	Control	55.2	10.9	female
EPIC_0297	Control	71.8	10.9	male
EPIC_0298	Control	71.8	10.8	male
EPIC_0299	Control	55.6	10.7	female
EPIC_0300	pre-AML	64.9	6	male
EPIC_0301	Control	71.4	10.6	female
EPIC_0302	Control	43.4	10.6	female
EPIC_0303	Control	64	10.6	male
EPIC_0304	Control	66.7	6.8	female
EPIC_0305	Control	66	10.5	female
EPIC_0306	Control	58.9	11.5	male
EPIC_0307	Control	58.4	11.4	male
EPIC_0308	Control	67.3	11.3	male
EPIC_0309	pre-AML	69.9	4.8	male
EPIC_0310	Control	56.6	11.2	male
EPIC_0311	pre-AML	56.1	4.4	male
EPIC_0312	Control	56.5	11.1	male
EPIC_0313	Control	56.4	11.9	male
EPIC_0314	Control	64.2	9.6	female
EPIC_0315	Control	64.3	9.5	female
EPIC_0316	Control	64.3	9.5	female
EPIC_0317	pre-AML	64.3	7.9	female
EPIC_0318	Control	64.7	9.5	female
EPIC_0319	Control	42.1	11.7	female
EPIC_0320	Control	56	11.7	male
EPIC_0321	Control	60.8	8.1	male
EPIC_0322	Control	56.7	3.4	female
EPIC_0323	pre-AML	36.3	9.3	female
EPIC_0324	Control	36.8	10.6	female
EPIC_0325	Control	68.5	10.5	female
EPIC_0326	Control	48.3	10.9	female
EPIC_0327	pre-AML	43.6	5.3	female
EPIC_0328	Control	71.5	10.4	female
EPIC_0329	Control	43.6	10.4	female
EPIC_0330	Control	43.9	10.4	female
EPIC_0331	Control	43.9	10.3	female
EPIC_0332	Control	71.9	10.3	female
EPIC_0333	Control	66.9	8	female
EPIC_0334	Control	66.5	7.9	female
EPIC_0336	pre-AML	50.9	3.2	female
EPIC_0337	pre-AML	63.1	5.6	female
EPIC_0338	pre-AML	59.1	4.6	female
EPIC_0339	pre-AML	60.2	5.5	female
EPIC_0340	pre-AML	43.5	2.9	female
EPIC_0341	pre-AML	66.6	1.9	female
EPIC_0342	pre-AML	51.4	6.4	male
EPIC_0343	pre-AML	50.3	4.7	female
EPIC_0344	pre-AML	55.8	4.7	female
EPIC_0346	pre-AML	58.9	0.8	male
EPIC_0347	pre-AML	64.1	3.9	female
EPIC_0348	pre-AML	70.3	0.9	female

EPIC_0349	pre-AML	61.5	7.6	male
EPIC_0350	Control	61.9	12.2	male
EPIC_0351	Control	63.3	10.4	female
EPIC_0352	Control	51.5	10.9	male
EPIC_0353	Control	51.5	8.1	male
EPIC_0354	Control	56.6	9.3	female
EPIC_0355	Control	56.4	9.2	female
EPIC_0356	Control	50.5	9.1	female
EPIC_0357	Control	50.4	8.9	female
EPIC_0358	Control	50.9	9.3	female
EPIC_0359	pre-AML	48.2	7.9	male
EPIC_0360	Control	51.8	11.5	male
EPIC_0361	Control	51.6	11	male
EPIC_0362	Control	61.4	10.9	male
EPIC_0363	Control	71.3	7.8	male
EPIC_0364	Control	49.5	7.4	female
EPIC_0365	Control	49.9	7.5	female
EPIC_0366	Control	49.4	8.2	female
EPIC_0367	Control	71.6	11.2	male
EPIC_0368	Control	71.5	10.8	male
EPIC_0369	Control	61.8	11.1	male
EPIC_0370	Control	61.3	10.6	male
EPIC_0371	Control	61.9	10.7	male
EPIC_0372	Control	71.6	10.3	male
EPIC_0373	pre-AML	55.6	4.4	male
EPIC_0374	pre-AML	49.9	3.4	female
EPIC_0375	Control	56.2	11.6	male
EPIC_0376	Control	56.8	10.6	male
EPIC_0377	pre-AML	66.4	0.7	female
EPIC_0378	pre-AML	56.6	6.5	male
EPIC_0379	Control	56.5	6.7	male
EPIC_0380	Control	57	10.5	male
EPIC_0381	pre-AML	49.4	4	female
EPIC_0382	Control	56.7	9.1	female
EPIC_0383	Control	49.8	7.5	female
EPIC_0384	Control	51.5	12.9	female
EPIC_0385	Control	51.7	13.1	female
EPIC_0386	Control	51.4	12.9	female
EPIC_0388	Control	61.6	10.9	male
EPIC_0389	Control	51.6	12.8	female
EPIC_0390	Control	61.1	8	male
EPIC_0391	pre-AML	67	9.1	female
EPIC_0392	pre-AML	56.3	7.4	female
EPIC_0393	Control	73.8	7.2	male
EPIC_0394	Control	73.9	10.9	male
EPIC_0395	Control	66.5	11.1	female
EPIC_0396	Control	71.4	10.9	male
EPIC_0397	pre-AML	69.2	9.7	female
EPIC_0398	Control	48.9	12.5	male
EPIC_0399	Control	64.5	13.8	male
EPIC_0400	Control	56.3	12.9	female
EPIC_0401	pre-AML	55.8	10	male
EPIC_0402	Control	56.6	9.9	female
EPIC_0403	Control	73.6	13.7	female
EPIC_0404	Control	73.7	13.7	female
EPIC_0405	Control	66.7	13.6	female
EPIC_0406	pre-AML	70.3	7.2	male
EPIC_0407	Control	70.3	13.4	male
EPIC_0408	Control	70.8	13.4	male
EPIC_0409	Control	73.9	12.7	male
EPIC_0410	Control	73.2	5.9	male
EPIC_0411	Control	58.1	11.6	male
EPIC_0412	Control	70	11	male
EPIC_0413	Control	70	13.2	female
EPIC_0414	Control	59.2	14.1	male
EPIC_0415	Control	66.9	11.9	female
EPIC_0416	Control	60.4	14.1	female
EPIC_0417	Control	60.6	12.9	female
EPIC_0418	Control	57.6	13.5	female
EPIC_0419	Control	54.1	12.7	male
EPIC_0420	Control	56.7	12.9	female
EPIC_0421	Control	55.8	12.9	male
EPIC_0422	Control	68.8	10.9	female
EPIC_0423	Control	69.2	11.5	male
EPIC_0424	pre-AML	63.8	7.3	female
EPIC_0425	Control	63.3	10.4	female
EPIC_0426	Control	58.1	12.3	male
EPIC_0427	Control	64.5	11.9	male
EPIC_0428	Control	64.2	11.8	male
EPIC_0429	Control	57.9	4.6	female
EPIC_0430	Control	57.3	11.5	female
EPIC_0431	Control	61.1	10.8	male
EPIC_0432	Control	53.4	10.4	male
EPIC_0433	Control	51.4	12.7	male
EPIC_0434	Control	52.9	14.2	male
EPIC_0435	Control	59.5	11.8	female
EPIC_0436	Control	59.7	13.2	female

EPIC_0437	Control	39.2	13	female
EPIC_0438	Control	49.2	13.5	female
EPIC_0439	pre-AML	64.8	9.1	female
EPIC_0440	Control	64.9	11.5	female
EPIC_0441	Control	55.9	14.7	male
EPIC_0442	Control	59.4	14.1	male
EPIC_0443	Control	54.7	14	male
EPIC_0444	Control	50.7	13.6	male
EPIC_0445	Control	46.9	12.8	male
EPIC_0446	pre-AML	47	7.6	male
EPIC_0447	Control	69.6	11.7	female
EPIC_0448	pre-AML	71	8.8	male
EPIC_0449	Control	64.9	10.5	male
EPIC_0450	pre-AML	51.5	2.6	female
EPIC_0451	Control	55.1	12.7	female
EPIC_0452	Control	62.5	11.7	male
EPIC_0453	Control	67.9	11.4	male
EPIC_0454	pre-AML	41.4	6.2	female
EPIC_0455	pre-AML	49.6	9.2	male
EPIC_0456	Control	67.4	11.8	female
EPIC_0457	Control	64.9	11.6	female
EPIC_0458	pre-AML	52.7	0.4	female
EPIC_0459	Control	67.9	12.1	female
EPIC_0460	Control	68.4	10.7	female
EPIC_0461	pre-AML	73.6	2.4	female
EPIC_0462	Control	52.4	13	female
EPIC_0463	Control	63.5	7.9	male
EPIC_0464	pre-AML	61.7	6	male
EPIC_0465	Control	58.1	11.6	female
EPIC_0466	Control	55	11.8	male
EPIC_0467	Control	58.9	8.9	male
EPIC_0468	Control	64.4	12.1	male
EPIC_0469	pre-AML	68	6.4	female
EPIC_0470	pre-AML	71.9	5.6	male
EPIC_0471	Control	58.8	12	male
EPIC_0472	pre-AML	39.5	11.1	female
EPIC_0473	pre-AML	59	11.8	female
EPIC_0474	Control	60.8	11.8	female
EPIC_0475	Control	67.8	11.2	male
EPIC_0476	Control	70.3	13.7	male
EPIC_0477	pre-AML	59.4	8.3	male
EPIC_0478	Control	53.8	10.1	male
EPIC_0479	pre-AML	56.6	0.2	female
EPIC_0480	Control	58.9	14.8	male
EPIC_0481	Control	49.1	13.3	female
EPIC_0482	pre-AML	57.8	11.2	male
EPIC_0483	Control	58.5	12.9	male
EPIC_0484	Control	58	12.8	male
EPIC_0485	Control	54.5	12.9	female
EPIC_0486	Control	48.7	12.6	male
EPIC_0487	Control	61	13.8	female
EPIC_0488	Control	46.3	13.7	female
EPIC_0490	pre-AML	52.8	0.3	female
EPIC_0491	Control	49.1	13	male
EPIC_0492	Control	67.2	13	female
EPIC_0493	pre-AML	44.9	0	male
EPIC_0494	Control	64.9	11.6	female
EPIC_0495	Control	62.3	12.1	male
EPIC_0496	pre-AML	62.6	1.4	male
EPIC_0497	Control	55.4	13.5	male
EPIC_0498	pre-AML	55.2	8	male
EPIC_0499	Control	67.8	11.5	female
EPIC_0500	Control	66.7	11.4	female
EPIC_0501	Control	72.8	9	male
EPIC_0502	Control	64.3	9.9	female
EPIC_0503	Control	56.2	12.6	male
EPIC_0504	pre-AML	68.5	7	female
EPIC_0505	Control	68.4	10.5	female
EPIC_0506	Control	71.2	10.5	female
EPIC_0507	pre-AML	46.6	1.8	male
EPIC_0508	pre-AML	68.4	6.2	female
EPIC_0509	pre-AML	66.8	4.1	female
EPIC_0510	pre-AML	58.2	9.1	male
EPIC_0511	Control	56.9	9.3	female
EPIC_0512	pre-AML	58.8	3.7	female
EPIC_0513	Control	48.2	10.8	female
EPIC_0514	pre-AML	48.4	4.7	female
EPIC_0516	pre-AML	72.9	7.8	male
EPIC_0517	Control	48.2	10.9	female

Appendix 2: Validation cohort pre-AML and control sample information

Sample ID	Group	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	BMI	Total cholesterol (mmol/L)	HDL (mmol/L)	LDL (mmol/L)	Triglycerides (mmol/L)	Lymphocytes (10^9/L)	MCV (fL)	RDW	WBC (10^9/L)	RBC (10^9/L)	Haematocrit (%)	Platelets (10^9/L)	Haemoglobin (g/dL)	Age at sample	Follow-up (years)
PD35595b	Control	Male	181	108	25.8	6.8	1.5	4.3	2.1	2.2	86.5	14	7	4.9	42.6	239	14.8	68.2	21.3
PD35724b	Control	Male	124	74	26.8	4.3	0.7	3.1	1	-	-	-	-	-	-	-	-	63.8	21.5
PD35520b	Control	Female	109	72	26.7	5.5	1.3	3.5	1.6	1.6	87.2	12.5	4.5	5	43.5	222	14.5	47.4	18
PD35651b	Control	Female	154	96	32	7.8	1.1	5.9	1.9	2.9	88.7	12.5	7	4	35.8	287	12.5	54.7	19.1
PD35622b	Control	Female	124	78	23.6	5.9	1.7	3.5	1.6	2	81.1	15.5	5.1	5	40.3	270	13.3	51.8	21.4
PD35518b	Control	Male	131	76	25	6.8	1.4	4.8	1.2	2.2	90.2	12.7	5.1	4.2	38	232	12.8	62	19.8
PD35626b	Control	Female	171	108	28.9	6.4	1.8	4	1.2	1.8	91	12.7	5.9	4.2	37.8	193	13.2	72	19.9
PD35711b	Control	Female	138	76	28.5	5.7	1.4	3.6	1.6	3.5	88.3	13.6	9.4	4.7	41.4	209	13.9	75	19.8
PD35786b	Control	Male	142	90	27.9	6.8	0.8	5	2.2	-	-	-	-	-	-	-	-	55.9	21.5
PD30073b	pre-AML	Female	143	82	33.3	7.2	1.2	5.3	1.7	3.3	87.2	12.9	9.2	5	43.8	149	14.6	71.8	1.2
PD35526b	Control	Male	149	78	25.9	6.6	1.2	4.5	1.9	2.1	87.1	13.2	8.1	4.9	42.5	224	14.6	68.6	20.7
PD35716b	Control	Male	123	83	23.8	6.2	1.7	3.6	2	2.4	89	12.8	6.3	4.3	38.8	267	13.2	49.8	19.2
PD35685b	Control	Male	134	87	26.2	5.3	0.9	3.2	2.5	2.1	88.5	13.4	7.4	5.2	45.6	318	16.2	56.8	20.1
PD35758b	Control	Male	156	99	33.9	7.3	1.6	5	1.6	1.6	84.7	14	6.2	5.3	45	158	15.7	65.1	21
PD35605b	Control	Female	127	85	23.5	6.4	1.5	4.3	1.2	-	-	-	-	-	-	-	-	61.2	21.7
PD35708b	Control	Male	160	84	25.7	5.9	-	-	6.3	1.9	88.2	12.5	6.2	4.4	39.1	225	13.7	67.2	19.8
PD35705b	Control	Male	163	92	24.3	5.9	1	3.7	2.7	1.3	95	12.4	6	4.6	43.7	191	14.8	69	20.7
PD35528b	Control	Female	158	95	31.1	8.3	1.9	5.4	2.2	2.3	78.3	14.2	7.1	4.9	38	231	12.5	53.3	18.8
PD35615b	Control	Female	128	68	23.2	8.3	-	-	6.3	2.1	91.5	12.1	5.5	4.2	38.2	161	13.6	68.8	19
PD35678b	Control	Male	135	78	25.4	6.1	1.2	4.1	1.8	-	-	-	-	-	-	-	-	68.2	21.9
PD35586b	Control	Male	147	91	19.4	5.6	2.2	2.9	1	-	-	-	-	-	-	-	-	67.1	17.3
PD35673b	Control	Male	110	67	26.4	6.2	1.6	4	1.4	1.9	89.7	12.3	7	4.9	43.8	268	14.8	48.2	19.8
PD35659b	Control	Male	153	94	25.2	5.8	1.2	3.6	2.1	2	91	13.4	7.6	5.7	51.5	278	17.3	68.6	20.3
PD35536b	Control	Female	106	66	23.9	5	1.8	2.7	1.1	2.2	91.6	11.4	6.2	3.7	33.6	272	11.6	49.7	20.4
PD35543b	Control	Male	154	86	20.9	5.8	1.5	3.7	1.3	-	-	-	-	-	-	-	-	65.1	18.8
PD29856c	pre-AML	Male	116	77	26.4	4.2	1	2.7	1.1	-	-	-	-	-	-	-	-	57.2	17.8
PD35572b	Control	Male	140	94	30.6	9.4	1.7	6.5	2.7	2.1	92.7	12.2	5.9	5.3	48.9	269	16.1	48.9	20.2
PD35631b	Control	Female	170	104	38.4	6.1	2.1	3	2.1	-	-	-	-	-	-	-	-	57.6	18.7
PD35599b	Control	Female	150	100	27.1	6.1	1	4.5	1.3	-	-	-	-	-	-	-	-	54.5	22
PD29810c	Control	Male	114	72	21.5	4.9	0.6	3.1	2.8	1.5	88.6	15.3	8	4.1	36.6	136	12	45.9	18.6
PD35522b	Control	Female	142	90	27.3	7.3	1.2	4.6	3.2	2	85.1	14	7.1	4.4	37.9	422	13.3	67.3	20.5
PD29804c	Control	Female	146	95	26.1	6.2	1.7	4.3	0.6	2	89.9	13.3	4.5	5.2	46.9	218	14.8	45.7	5.1
PD35625b	Control	Male	162	106	27.5	6	0.7	3.8	3.2	-	-	-	-	-	-	-	-	48.8	22.1
PD35589b	Control	Female	116	79	25.9	5.5	1.2	3.8	1	1.2	84.8	14.9	6.1	4.6	39.2	305	13.7	62.8	20.9
PD29792b	pre-AML	Female	142	89	30.2	6	1.5	4.2	0.6	-	-	-	-	-	-	-	-	64.9	14.1
PD30060c	pre-AML	Female	148	85	22.4	7.1	1.7	4.9	1.2	1.7	90.1	14.2	4.8	4.2	37.7	252	12.1	75.8	15.2
PD35519b	Control	Female	134	85	29.5	6.5	1.4	4.4	1.5	2.2	90.5	13	12	4.5	40.4	384	14	65.7	20.2
PD35763b	Control	Male	151	90	22.9	3.8	1.3	2	1.2	2.3	85.2	14.4	7.6	4.8	40.7	268	14.2	64	20.5
PD35725b	Control	Female	148	84	31	8.4	1.8	5.7	1.9	1.2	86.1	14.5	3.7	4.3	37.2	234	12.9	70	20.9
PD35507b	Control	Male	174	104	23.6	6.3	1.8	4.1	0.9	-	-	-	-	-	-	-	-	55.4	21.9
PD29836c	pre-AML	Female	152	92	28.5	6.4	1.4	4.1	2.1	1.7	88.4	13	5.4	4.2	37.2	175	12.6	70	10
PD35556b	Control	Female	138	77	25.9	5.6	1.5	3.4	1.5	2.2	89.9	12.6	6.7	4.8	42.7	223	14.5	64.5	20.8
PD35616b	Control	Male	138	87	31	6.5	1.1	4.4	2.2	1.9	94.1	13.3	6	4.7	44.6	203	15.5	68.4	21
PD35787b	Control	Male	107	61	25.7	5.8	1.1	4.2	1.2	2.2	91.2	13.7	6	4.5	41.2	144	14.1	68.2	20.7
PD35775b	Control	Female	122	80	29.3	6.4	2.1	4	0.7	1.5	86.9	13.8	5.3	4.2	36.6	227	12.2	64	20
PD35665b	Control	Male	115	78	25.3	5.9	1.4	3.8	1.5	-	-	-	-	-	-	-	-	65.5	22
PD35760b	Control	Male	128	74	30.5	6	1	4.3	1.5	2.2	85.2	14.9	6.3	5.4	46.3	145	15.6	67.5	19.8
PD35764b	Control	Male	118	72	26.5	4.3	0.8	2.8	1.5	1.6	86.6	13.2	8	4.9	42.7	304	14.8	61.7	21.4
PD35660b	Control	Female	136	82	27.6	6	1.2	4.1	1.6	2.3	88	12.4	6.1	4.4	38.8	272	13.2	59.2	21.1
PD30010c	pre-AML	Male	168	108	27.2	6.3	1.1	3.4	3.9	2.4	100	15.6	3.7	3.9	38.7	91	13.3	66.3	12.7
PD35777b	Control	Male	143	92	27.9	5.8	-	-	5.3	2.8	82.3	14	7.7	5.4	44.8	274	15	61.4	19.6
PD35694b	Control	Male	168	99	33.7	5.8	1.9	3.2	1.7	2	95.4	14.1	4.8	4.7	44.7	235	14.6	72.7	18.7
PD35781b	Control	Male	128	83	28.2	4.1	0.9	2.6	1.3	2.2	90.1	12.8	7.1	4.5	40.4	219	14.1	59	21.6
PD35552b	Control	Male	120	76	26.3	6.3	1.4	4.1	1.9	2.1	93.3	13.2	7	4.8	45.1	280	14.8	61.7	18.9
PD35757b	Control	Female	122	74	24.2	6.9	1.3	4.7	2	2	93.8	12.7	6.4	4.2	38.9	255	12.3	65.2	18.5
PD35587b	Control	Female	134	82	27.5	6.7	1.8	4.4	1.1	2.3	89.3	13	6.5	3.8	33.7	198	11.4	69.3	21.3
PD30116c	pre-AML	Male	143	82	26.2	5.5	1.1	3.8	1.4	1.7	80	17.3	5.2	3.9	31.2	207	9.8	69.9	5.1
PD29858b	pre-AML	Female	150	90	25.2	7.6	1.6	5.2	1.8	1	88.9	12.7	5.5	4.5	39.9	243	13.7	73.6	2.4
PD35676b	Control	Female	150	82	25.8	8.3	1.6	6	1.6	-	-	-	-	-	-	-	-	64.9	22
PD30008c	Control	Male	122	78	26.2	4.9	1.1	3.2	1.4	1.3	87.2	14.1	5.6	5.2	45	275	14.7	56.6	20
PD35684b	Control	Female	113	74	22.9	4.7	2	2.3	0.8	2.6	96.6	11.8	9	4.3	41.5	284	14.4	46.4	19.7
PD30111c	pre-AML	Female	116	76	21.1	9	1.9	6.1	2.3	1.6	89.1	13.2	6	4.3	38.7	201	12.9	48.4	4.6
PD30159c	Control	Female	108	66	22.4	8.2	1.4	5.5	3	3	92.5	13.3	6.5	4.3	39.5	229	13.4	69	18.7
PD29948b	pre-AML	Female	156	82	28	7.9	1.2	5.5	2.5	2.4	84.6	13.2	6.3	4.7	39.6	374	14.1	72.2	17.8
PD30086b	pre-AML	Male	150	87	31	4.2	0.7	2.3	2.7	1.3	95.9	13.7	5.1	4.1	39.5	185	13.7	66.4	13.6
PD35702b	Control	Male	112	68	29.1	7	0.8	5.6	1.4	-	-	-	-	-	-	-	-	67.7	22
PD35768b	Control	Female	157	91	34.3	5.3	1.1	3.2	2.2	3.5	94.3	12.8	8.7	4.8	45.2	209	15.5	68.9	21.1
PD35573b	Control	Female	128	62	23.9	6.2	0.8	4.6	1.8	2.2	87	12.9	5.8	3.8	33.4	245	11.5	71.5	19.3
PD35525b	Control	Male	122	72	26.1	5.1	1	3.1	2.2	2.6	88.7	13.3	7.9	5	44.5	268	15	66.7	18.8
PD30154c	pre-AML	Female	124	82	25.3	7.3	1.5	4.4	3.1	2.1	84.7	13.1	5.5	4.9	41.9	225	13.9	61.3	15.7
PD35569b	Control	Male	124	84	23.1	6.4	1.6	4.2	1.5	2.7	86	12.2	7.6	4.8	41	283	14.8	53.5	19.1
PD35640b	Control	Female	140	80	33	6.2	1.7	3.5	2.1	2.3	89.9	13.7	8.3	4.2	38	203	13	68.4	19.4
PD35612b	Control	Female	138	80	36.6	6.1	1.5	4.1	1.1	-	-	-	-	-	-	-	-	56.7	21.9
PD35667b	Control	Female	110	68	20.5	7.6	1.6	5.6	0.9	-	-	-	-	-	-	-	-	68.9	21.7
PD29935c	pre-AML	Male	137	94	27.7	8.4	1.7	5.7	2.1	1.9	87	14.1	6.6	5.2	45.4	268	15.5	61.3	17.7
PD35740b	Control	Male	126	74	24	7.5	1.6	5.2	1.5	-	-	-	-	-	-	-	-	69.8	21.8
PD29933c	pre-AML	Male	176	97	25.1	5.6	1.4	3.6	1.4	1.6	92.2	12.7	5.4	4.6	42.6	191	14.8	73.2	5.8
PD35545b	Control	Male	145	88	34.2	5	1.5	2.9	1.4	1.5	91.6	13.3	5.8	4.8	44.4	258	15.7	70.1	20.4
PD29951b	pre-AML	Female	110	74	27.6	5.8	1.6	3.3	1.9	1.6	89.4	14.7	5.4	4.2	37.9	312	12.4	58.6	18.4
PD35782b	Control	Male	118	70	20.7	5.4	2.2	2.7	1	1.4	85.7	13.2	5.9	4.7	40.4	220	14.3	48.1	20.5
PD35549b	Control	Female	116	78	29.4	5.6	1.9	2.9	1.7	1.9	88	13.5	7.2	4	35.5	293	12	49.5	20.2
PD35637b	Control	Male	121	80	25.8	6	1.2	3.9	2.1	2.6	93.2	13.3	8.3	4.7	43.5	190	14.8	60.2	15.6
PD29762b	pre-AML	Female	180	96	25.4	6.7	1.7	3.4	3.6	3.1	98	12.9	8.2	4	39.4	235	14	60.2	9.8
PD35733b	Control	Female	140	84	26.1	5.9	2.1	3.3	1	1.7	85.9	13.5	4.2	4.7	40.4	243	14.1	61.3	20.4
PD30089b	pre-AML	Female	112	70	27.3	7.6	1	6.1	1.1	1.9	94.8	12.9	4	4	38.2	336	12.4	63.4	13.5
PD30058c	Control	Female	148	95	25.5	7.2	1.3	4.8	2.5	2.8	85.1	12.7	7.5	4.7	39.9	302	13.6	56.2	19.3
PD35650b	Control	Female	120	76	24.8	7.2	2	4.8	0.9	2	85.6	14.2	6.3	4.8	40.7	306	13	50.5	20.7
PD29851c	pre-AML	Female	126	76	27.7	7.3	1.3	4.8	2.5	-	-	-	-	-	-	-	-	55.8	12.2
PD35691b	Control	Male	158	102	21.7	7.5	1.6	5.3	1.4	2.9	88.5	13.6	8.2	5.2	46	288	16.5	64.8	19.6
PD35722b	Control	Male	143	98	27.9	7.1	1.3	4.7	2.5	-	-	-	-	-	-	-	-	57.1	21.5
PD35610b	Control	Female	110	70	22.5	7.2	1.6	4.9	1.5	2.8	93.8	12.4	7.1	4.3	40.7	313	14.1	47.7	20.9
PD35580b	Control	Female	146	88	21.2	6.9	1.3	5.1	1.1	1.9	90.4	12.7	6	5.4	48.8	250	16.7	68.5	21.3
PD29929c	pre-AML	Female	155	92	26.2	8.5	1.8	5.8	2	2.2	92.8	12.4	6.2	4.4	40.9	332	14.2	68.4	6.6
PD35613b	Control	Female	123	70	28	6.2	1.6	4.3	0.7	2	88	13.1	5	4.4	39	220	13.6	63.5	21
PD35509b	Control	Male	172	104	28	9	1.1	6.2	3.7	-	-	-	-	-	-	-	-	69.4	21.8
PD35609b	Control	Female	138	82	22.2	8.2	1.8	5.9	1.1	-	-	-	-	-	-	-	-	75	21.5
PD35550b	Control	Female	178	102	31.4	7.1	1.6	4.5	2.1	1.8	85.6	13.7	5.8	4.8	41.2	301	14.1	67.5	21.2

PD29946c	pre-AML	Female	158	97	30.6	6.9	1.3	4.4	2.6	2.3	85.1	12.6	7.9	4.7	40.1	332	13.6	70.1	14.9
PD30031b	pre-AML	Male	158	96	28.8	5	1	3.4	1.4	2	85.9	13.5	7.4	5.3	46	300	15.3	71.7	14.3
PD35647b	Control	Female	150	83	30	7.3	1.1	5.4	1.7	-	-	-	-	-	-	-	-	73.3	21.8
PD35624b	Control	Female	133	84	26.8	7.1	1	5.3	1.7	2.3	91.8	15	5.4	4.9	45	212	15	70.3	21.7
PD35601b	Control	Female	136	85	23	5.3	1.1	3.9	0.6	-	-	-	-	-	-	-	-	57.4	21.7
PD35564b	Control	Male	156	100	31.8	7.8	-	-	4.6	2.9	92.7	13.3	8.5	5.3	49.5	397	17.1	71.7	13.7
PD35508b	Control	Female	157	79	25.4	6.8	1.2	4.6	2.1	1.3	82.7	13.9	5.9	4.4	36.4	315	12.6	66.9	21.4
PD30120c	pre-AML	Male	135	84	29.9	5.7	1.3	3.8	1.4	1.8	90.2	14.2	6.1	4.8	43.6	210	14.4	69.7	12.3
PD35664b	Control	Male	107	64	25.2	7.7	1.1	5.3	2.9	3.6	89.7	14.2	10	5.1	45.8	226	15.3	44.4	19.5
PD29993b	pre-AML	Female	142	83	28.3	7.1	2.3	4.1	1.6	2	82.6	14.2	7	4.6	37.9	337	13.7	71.6	2.4
PD35652b	Control	Female	134	80	32.8	5.1	1.4	3.1	1.4	-	-	-	-	-	-	-	-	57.8	19.6
PD29989c	Control	Male	132	84	26.1	7.1	1.3	5.2	1.3	1.5	92	12	4.2	4.9	44.7	240	15.1	47.8	20.3
PD29962b	pre-AML	Male	140	86	26.6	5.1	1	3.7	0.9	2.5	90.3	13.2	7.4	4.8	43.1	238	14.8	72	14
PD35688b	Control	Female	138	76	25.4	7.8	1.2	5.5	2.3	-	-	-	-	-	-	-	-	68.8	21.5
PD35780b	Control	Male	158	90	25.3	6.4	1.2	4.4	1.8	1.4	94.4	12.5	5.2	4.3	40.4	202	13.6	65.3	19.5
PD35514b	Control	Female	127	71	21.8	6.2	1.6	4.4	0.4	-	-	-	-	-	-	-	-	72	20.7
PD35636b	Control	Female	146	91	30.8	7	1.3	4.9	1.8	-	-	-	-	-	-	-	-	64.9	21.9
PD29978c	pre-AML	Male	171	103	26.7	5.6	1.2	3.2	2.6	3.3	86.2	14.8	7.2	5.1	43.6	122	14.9	61.7	12.3
PD35707b	Control	Male	163	98	25.6	7.3	1.1	4.9	2.8	-	-	-	-	-	-	-	-	70.3	21.7
PD35596b	Control	Male	104	64	17.6	5	1.5	3.1	0.8	0.8	90.8	12.7	2.3	4.3	38.9	182	13.7	48.3	19.8
PD35720b	Control	Female	128	83	22.3	6.7	2.1	4.3	0.8	2.8	89.4	11.9	6.8	3.5	30.9	218	11.3	60.1	19.4
PD35579b	Control	Female	169	98	31.2	7.4	1.2	4.8	3	2.9	93.8	12.4	8.2	4.3	40.4	276	13.3	63.4	20.6
PD35565b	Control	Male	137	88	30.1	5.4	1.1	3.1	2.5	2.2	90.3	12.6	5.8	4.6	41.8	132	14.6	57.3	21.2
PD35723b	Control	Male	122	78	30.9	5.8	1.1	2.9	3.9	2	88.6	12.8	7.9	4.7	42	216	15.3	58.6	20.5
PD29918c	pre-AML	Male	158	92	27.3	5.5	0.9	3.1	3.2	1.8	93.2	12.7	5.7	4.4	41.2	173	14.2	76.6	13.4
PD35645b	Control	Male	124	68	24.8	5.3	1.2	3.4	1.6	-	-	-	-	-	-	-	-	73.3	21.6
PD29960c	pre-AML	Female	124	81	21.5	6.8	1.5	4.8	1.1	1.9	91	12.5	12.6	4.6	41.8	306	15	56.1	7.9
PD35515b	Control	Female	156	90	30.8	5.9	1.3	3.6	2.1	1.9	86.1	12.7	5.7	4.2	36.6	376	12.5	70.4	20.1
PD35717b	Control	Female	116	76	17.3	6.3	2.4	3.3	1.3	1.5	87.6	13.2	9.4	4.3	37.9	279	13.3	65.7	20.6
PD35690b	Control	Male	115	72	26.1	6.2	1.7	3.9	1.3	2.1	92.3	12.8	5.3	4.6	42.6	243	14.3	74.8	20
PD35623b	Control	Female	166	110	24.1	7.8	1.7	5.3	1.8	2.5	90.1	13.8	8	4.8	43.6	351	15.1	65.4	21.1
PD29897b	pre-AML	Female	123	82	27	4.8	2	2.2	1.4	1.8	91.7	12.9	4.8	4.3	39.5	278	13.8	60.2	5.8
PD35738b	Control	Female	124	78	24.8	6	1.1	4.6	0.6	-	-	-	-	-	-	-	-	60.3	22.1
PD35553b	Control	Male	144	94	24.9	4.9	1	3.1	1.8	-	-	-	-	-	-	-	-	67	21.5
PD35697b	Control	Female	120	66	25.5	6.3	1.4	3.7	2.6	2.4	89.5	13.3	7.7	4	36	247	12.3	63.7	20.7
PD35608b	Control	Male	142	80	23.6	6	2.6	3	0.8	1.7	93.1	13.1	8.1	4.4	40.9	349	14.2	64.8	20.5
PD35773b	Control	Female	118	79	30	7	1.6	4.5	1.9	1.4	92.5	12.4	7	3.9	36.5	210	12.8	72.3	19.3
PD29867b	pre-AML	Male	144	92	26.7	6.7	1	4.2	3.3	-	-	-	-	-	-	-	-	68	15
PD29996b	pre-AML	Female	109	66	34.4	5.9	1.1	4	1.7	1.6	97.5	12.5	5.2	4.3	41.9	255	14.6	52.4	4.6
PD35721b	Control	Male	126	78	29.4	4.9	1.4	2.7	1.8	1.5	87	12.9	7.4	5	43.9	300	15.1	53	20.1
PD29907c	pre-AML	Female	118	70	32.1	7.5	0.8	6.2	1.1	3.7	80.7	16.7	11.2	5.1	41.2	380	14	68	6
PD35512b	Control	Female	112	68	26.1	5.9	1.9	3.4	1.5	1.7	93.3	12.5	4.9	4.2	39.6	238	14.1	49.4	16.6
PD35646b	Control	Female	104	65	23.8	6.9	1.9	4.8	0.6	2	99.7	12.1	5	4.3	43.3	261	14.3	47.9	15.6
PD35686b	Control	Male	128	72	27.4	5.7	1.2	2.8	3.8	2.6	96.5	12.4	6.1	4.6	44.6	172	14.2	70.6	17.6
PD35642b	Control	Female	114	66	22.8	4.7	1.4	3	0.7	1	87.8	12.8	4.6	4.6	40	298	13.2	50	17.4
PD35710b	Control	Female	112	65	27.1	4.6	1.1	3	1.3	1.6	82.5	14.5	8.5	3.9	32.3	339	11.5	69.9	17.2
PD35620b	Control	Female	152	96	19.9	4.6	2.6	1.7	0.7	1.4	92.6	12.9	5.6	4.3	40.2	138	12.9	56.4	17.6
PD35670b	Control	Male	122	79	27.8	5.5	1.1	3.6	1.8	1.2	91.7	13.8	6.3	5.2	47.8	174	17.2	62.6	16.4
PD35540b	Control	Male	106	74	26.6	6.2	1.2	4.4	1.4	1.7	88.5	13.3	5.6	4.3	38.2	178	14.2	65.8	11.3
PD35627b	Control	Male	154	87	29.1	7.4	1	4.3	4.8	2	92	13	5.8	4.9	45.2	197	16.4	69.3	16.9
PD35661b	Control	Male	146	99	29.1	8.3	1.2	6	2.5	1.8	98.2	12.2	8.1	4.6	45	231	14.7	71.4	17.9
PD35641b	Control	Male	146	76	26.6	5.1	1	3.7	0.9	2.3	90.3	13.2	6.2	4.8	43.3	119	14.7	72.9	18.1
PD35731b	Control	Male	134	92	28.4	6.7	1.1	4.3	3	2.5	93	12.4	6.4	4.4	41.3	284	13.2	68.9	17.8
PD35638b	Control	Female	118	66	21.2	5.9	1.8	3.8	0.7	2.4	93.8	12.4	7.7	4.6	43.5	193	14.2	63.2	17.5
PD35712b	Control	Male	115	68	22.7	4.4	2.1	1.9	1	2	94	13.7	6	4.4	41.2	222	15	59.9	16.2
PD35558b	Control	Male	106	62	22.8	5.2	1.5	3.3	0.9	1.5	92.6	13.4	9.7	4.8	44.1	272	15.3	74.5	15.9
PD35598b	Control	Male	139	87	29.4	7.7	1.3	5.3	2.6	1.6	94.4	12.9	4.9	5	47.4	125	15.6	56.2	17.7
PD35769b	Control	Female	145	84	25.3	6.7	2.2	4.2	0.7	2.5	93.6	12.9	6.5	5.2	48.7	228	14.9	65.2	15.7
PD35511b	Control	Female	144	78	26.4	8.7	1.1	5.5	4.8	2.6	95.1	14.6	5.5	4.3	41.3	331	14.1	73.5	13.9
PD35693b	Control	Male	144	75	24.7	9.3	1.6	6.5	2.8	1.9	88.3	14.6	6.3	4.6	40.3	400	14	73.5	16.4
PD35700b	Control	Female	134	80	24.9	6.1	1.4	3.3	3.1	1.8	91.5	14	7.1	4.3	39.2	261	14.1	77.4	16.9
PD35674b	Control	Female	158	93	23.8	6.3	1.5	4.3	1.1	1.8	87.1	13	6.1	4.2	36.8	271	12.4	66	17.5
PD35632b	Control	Female	164	89	29.2	7	1.4	4.3	2.9	1.7	95.9	11.8	6.9	4.3	41.8	310	13.5	76.2	13.3
PD35657b	Control	Male	160	114	31.1	6.1	0.8	4.1	2.8	2.5	89.4	12.9	8.7	4.8	42.5	224	14.8	61	16.4
PD35706b	Control	Male	128	85	25.4	8.1	1.4	5.8	2	1.5	90.1	12.3	5.8	4.7	42.3	392	14.7	52.3	16.6
PD35524b	Control	Female	104	61	19.6	4.4	1.4	2.7	0.7	1.6	85	14	9.1	4.2	36	185	12.4	52.6	16.3
PD35756b	Control	Male	130	78	22.8	5.1	0.7	3.1	2.9	1.3	87.9	13.3	5.5	4.3	37.8	244	13.8	76.5	16.8
PD29931b	pre-AML	Female	160	94	32.4	6	1.1	3.6	3	2.7	86.8	13.7	9.5	4.6	40	276	14.2	71.1	13.9
PD35633b	Control	Male	150	86	26.4	5.7	1.1	4	1.4	2.1	93	13	6.2	4.2	39	275	14.2	76.3	12.4
PD35715b	Control	Male	140	96	27.1	6.3	1.3	4.5	1.3	3.4	96.1	15.3	8.4	4.6	43.8	268	15.4	67.7	15.9
PD35529b	Control	Female	128	82	27.5	5.2	1.8	3	1.2	1.9	83	14.1	6.9	4.4	36.6	325	13	70	15.8
PD35732b	Control	Female	125	78	27.6	5.5	1.7	3.5	0.7	2.7	87.3	12.9	7.8	4.6	40.1	223	14.7	56.4	17.1
PD35571b	Control	Female	142	74	23.9	5.2	1.3	2.5	3.1	1.9	89.2	13.6	6.9	4.3	38.6	269	12.7	52.7	17.3
PD35611b	Control	Female	148	98	26.2	7	2.1	4.6	0.8	1.8	94.6	14.5	6.5	4.3	40.5	293	14.9	73.7	17.2
PD35703b	Control	Male	142	86	28.2	5.7	1.2	3.7	1.8	1.9	88.5	14.7	7	4.9	43.6	276	15.2	77.1	16.4
PD35654b	Control	Male	144	88	22.4	5.3	1.2	3.1	2.3	3	91	14.4	7.5	5.3	48.4	153	15.9	71.3	13.1
PD35639b	Control	Female	132	78	25.1	5.4	1.2	3.6	1.5	2.6	90.2	13.1	5.2	4.6	41.7	351	13.2	67.4	17.3
PD35534b	Control	Female	105	66	21.9	7	2.4	3.7	2.1	1.7	92.8	12.5	4.8	4.1	37.6	233	13.1	66.3	16.8
PD35581b	Control	Male	126	74	22.8	5.2	1.2	3.7	0.8	2.5	90.5	13.4	7.5	4.3	39.2	192	13.6	68.6	17.3
PD35542b	Control	Female	146	88	30.2	5.2	1.2	3.5	1.3	2.1	87.4	14.6	6.4	4.8	41.9	434	14.5	65.7	16.3
PD35594b	Control	Female	126	82	28.1	6.1	1.5	4.1	1.2	2.6	94.8	13	7.4	4.5	42.9	266	13.6	68.2	17.6
PD29907b	pre-AML	Female	141	76	32.2	7.9	0.8	6.5	1.5	2.4	85	17	8.6	5.1	43	400	14.8	71.9	6
PD35591b	Control	Female	128	80	21.2	5.8	1.9	3.7	0.5	1.6	85	13.9	4.9	4.4	37.3	245	13.2	56.9	17.2
PD30023b	pre-AML	Male	190	116	27.3	4.3	1.6	2.2	1.3	2	92.4	13.3	5.3	5	45.9	176	16.4	61.8	3.2
PD35762b	Control	Female	118	66	25.9	5.5	1.6	3.5	0.9	2.6	89.9	14	6.5	4.7	42.2	274	14.1	76.1	16.4
PD35582b	Control	Female	163	83	26.5	7.7	1.8	5	2.1	-	-	-	-	-	-	-	-	76.4	16.3
PD35583b	Control	Male	127	58	24.9	4.2	1.2	2.6	1	1.2	89.3	13.7	5.7	4.6	41.2	125	14.6	76.3	11.5
PD35619b	Control	Female	150	87	31	7.3	1.2	5.4	1.6	1.5	96.2	12.2	4.8	4.3	41.7	215	13.6	66.4	17.8
PD35541b	Control	Male	132	70	30.1	5.7	1.5	3.7	1.2	1.7	90.2	14.3	5.4	4.6	41.4	222	14.4	72.4	16
PD35662b	Control	Female	114	70	23.7	8.4	1.3	6.6	1.1	1.4	90	14.7	4.9	4.7	42.1	243	13.3	72.9	17.7
PD35672b	Control	Male	132	80	26.1	5.4	1.1	2.8	3.3	1.3	92.3	12.8	6.5	5.1	47	177	15.5	66.3	12.4
PD35682b	Control	Male	134	88	27.3	6.6	1.3	4.5	1.9	2.2	91.2	13.7	6.3	4.8	43.3	349	14	61.8	15.5
PD35704b	Control	Male	131	86	29.6	9.6	1.1	6.4	4.7	2.8	84.1	14.4	9.7	5.7	47.6	317	16.5	56.8	16.2
PD35671b	Control	Female	152	84	22.8	6.5	1.7	4.2	1.4	0.9	88.1	13.1	5	4.8	42.2	220	13.9	71.6	18.1
PD30054b	pre-AML	Male	128	74	24.4	6.7	1.2	4.6	2	0.9	95.2	13.1	3.1	4.3	41	134	14.4	75.2	13.8
PD35759b	Control	Female	122	74	23.7	5.3	1.5	3.1	1.7	2.4	89.6	12.7	7.2	4	35.7	251	12.8	52.3	16.3
PD35523b	Control	Male	132	73	28.6	5.7	0.9	3.5	3	2.7	91.1	13.7	5.6	4.2	38.4	268	13.5	63.9	16.5
PD35547b	Control	Male	140	73	24.4	5.2	0.9	3.2	2.5	1.6	88.8	14.4	7.5	5.7	50.8	296	16.5	69.1	12.6
PD35699b	Control	Male	130	74	26.7	4.8	1.4	3	0.9	1.7	97.7	12.1	6.2	3.9	38.3	194	14.3	57	17.1
PD30111b	pre-AML	Female	112	64	20.1	8	1.7	5.6	1.7	1.3	88.9	12.4	5	4.4	39.1	392	14.6	51	4.6
PD35709b	Control	Female	126	72	21.2	4.5	1.5	2.6	0.9	3.1	95.8	13.9	11.1	4.5	43.4	249	14	73.3	17.4
PD29836b	pre-AML	Female	146	82	29.6	7.1	1.3	5.2	1.5	1.6	85.7	14.1	5.2	4.6	39.3	229	13.2	72.4	10
PD35635b	Control	Female	148	87	30.7	6.5	1.6	3.9	2.4	3.2	87.7	12.2	7.5	4.5	39.2	256	13	62.5	17.8
PD29978b	pre-AML	Male	146	86	27.4	6.7	1.2	4.5	2.3	-	-	-	-	-	-	-	-	65.4	12.3

PD35677b	Control	Female	158	86	24	6.6	1.3	4.6	1.6	1.5	88.7	13.5	6	4	35.7	234	12.3	71.9	16.5
PD35784b	Control	Female	156	92	27.1	7.4	2	4.5	2	2.4	94.4	13.2	5.6	3.8	35.6	285	12.5	68.5	16.4
PD35544b	Control	Male	144	82	27.9	6.5	0.8	4.1	3.7	2	87.7	12.7	5.2	5.1	44.9	363	15.4	52.9	17.2
PD35771b	Control	Male	140	88	27.3	6.6	2.1	3.9	1.5	2.3	94.4	13.1	6.1	4.9	46.5	216	14.8	63.6	17.7
PD35726b	Control	Male	152	90	26.2	6.5	2	4.2	0.8	1.7	97.7	12.9	5.3	4.3	41.9	234	14	79.3	15.9
PD35785b	Control	Male	142	90	27.6	5.3	1.9	3	1.3	1.2	91	13.5	5.8	5	45.5	331	15.2	56	15.8
PD35701b	Control	Male	118	76	25.8	6.8	1.5	4.6	1.7	1.2	93.3	12.7	5.3	5.2	48.6	274	15.7	73.4	17.5
PD35776b	Control	Male	122	66	27.2	6.6	1.3	4.3	2.4	2.2	90.4	13.4	7.8	4.5	40.7	196	14.1	71.6	8.2
PD29764b	pre-AML	Female	132	70	27.1	6.1	2.1	3.4	1.4	2.8	80.7	22	7.9	4.5	36.2	280	12.1	78.6	10.4
PD35683b	Control	Female	116	71	26.5	5.4	2	2.8	1.4	1.3	91	12.9	5.8	3.9	35.5	193	12.6	69.6	16.1
PD35607b	Control	Male	153	90	25.9	5.6	1.1	3.5	2.3	1.3	90.5	13.6	6.2	4.2	37.8	255	13.9	77.6	17.2
PD35533b	Control	Male	135	88	26.2	5.2	1.7	3	1.2	2	89	13.5	6.3	4.8	43.2	293	14.1	58.3	16.2
PD30154b	pre-AML	Female	132	88	25.4	8.4	1.2	4.8	5.4	2.3	84.9	14.4	8.4	4.9	42	296	14.1	63.9	15.7
PD35555b	Control	Female	132	77	20.7	6.9	2.7	3.8	1	1.2	87	13.5	5.2	4.6	40.2	258	13.9	72.8	16.1
PD35614b	Control	Male	122	76	27.6	4.1	1.1	2.6	0.9	1.8	90.4	14.3	6.4	5.4	49	268	15.9	71.1	18
PD35517b	Control	Female	122	74	27.7	6.2	2.7	3.1	1.3	2.1	89.2	13.4	7.7	4.5	40	406	13.7	53.9	16.2
PD29896b	pre-AML	Female	148	98	27.8	8.2	1.2	5.4	3.6	3	93.9	15.4	8.4	4.3	40.7	325	13.7	70.6	6.4
PD29946b	pre-AML	Female	141	86	29.9	5.6	1.2	3.8	1.5	2.2	85.9	13.1	6.9	4.4	37.9	287	13.5	74	14.9
PD35597b	Control	Female	130	72	19.5	6.2	2.3	3.5	0.9	1.4	88.4	12.5	5.3	4.3	38.1	264	14.2	45.6	16.6
PD35789b	Control	Male	111	78	24.6	6.1	1.3	3.7	2.6	1.9	92.2	13.5	5.3	4.8	43.8	315	14.8	51.6	17.1
PD35539b	Control	Female	120	74	23.6	7.4	1.7	5.3	1	1.9	94	11.9	5.3	4.3	40.6	255	13.6	63.8	16.4
PD35679b	Control	Female	148	88	22	5.3	1.9	2.5	2	1.4	89.5	12.6	7	3.9	35.2	332	12	69.7	17.4
PD30060b	pre-AML	Female	160	92	24.1	5.3	1.7	3	1.5	2.4	87.3	14.7	7.1	4.3	37.2	401	12.4	78.5	15.2
PD35681b	Control	Male	131	72	21.2	7.1	2	4.7	1	1.6	84.2	14	5	5	42.3	209	14.8	57.7	17.2
PD29933b	pre-AML	Male	148	92	24.9	5.5	1.7	3.3	1.1	1.1	95.2	13.2	3.7	4.2	40	161	14.5	77	5.8
PD35590b	Control	Female	124	70	26.7	7.1	1.6	4.7	1.8	1.5	86.6	12.5	5	4.6	39.5	278	13.8	70.2	17
PD35546b	Control	Female	112	68	21.3	7.5	1.3	5.7	1.3	0.8	91.2	14.3	3.7	4	36.5	243	11.9	52.5	15.6
PD35521b	Control	Female	182	106	28.7	6.4	2	3.8	1.5	1.3	87	14	6.6	4.3	37	180	13.3	79.5	16
PD35570b	Control	Male	146	86	32.6	6.6	1.3	4.3	2.2	2.2	89.3	13.7	5.4	5	45	223	16.2	60.2	16.9
PD35696b	Control	Male	146	78	25.8	7.2	2	4.9	0.8	1.5	90.7	12.9	4.7	4.5	40.5	208	14.5	65.7	16.3
PD35551b	Control	Female	148	79	29.2	3.5	1.2	1	2.9	3.1	76.9	18	9.8	5.2	40.1	312	11.8	53.5	15.7
PD35554b	Control	Male	152	88	29.2	6.1	0.9	3.7	3.4	2	92.3	13.5	6.9	5.2	47.6	264	15.5	73.6	18
PD35527b	Control	Male	110	76	26	5.2	1.8	2.9	1.1	2.2	91.1	13.7	5.9	4.6	42	321	14.8	50.1	16.2
PD30120b	pre-AML	Male	120	74	27.9	6.2	1.7	4.1	1	1.3	90	13.1	4.9	4.6	41.6	205	14.6	72.2	12.3
PD35560b	Control	Female	152	97	31.7	5.7	2	3.2	1.1	2.7	89	13	7.4	4.1	36.3	39	12.8	69.6	17
PD35566b	Control	Female	120	84	18.6	5.8	2.9	2.4	1.1	2.2	88.3	13.3	5.6	4.3	38.3	253	12.7	57	17.7
PD35663b	Control	Male	128	86	28.7	5.5	0.9	3.8	1.9	2.2	90.8	12.6	6.2	4.5	40.7	220	13	54	17.8
PD35617b	Control	Male	174	100	26.6	4.4	2	1.8	1.4	2.2	94.8	13.8	7.2	4.2	39.8	263	13.5	79.6	17.3
PD35698b	Control	Female	148	90	28	5.1	1.1	3.5	1.2	2	82	14.4	5.4	4	33.1	289	11.1	71.3	17.4
PD35510b	Control	Male	144	82	23.1	5.2	0.8	3.8	1.5	1.8	92.9	14.3	7.9	3.7	34.6	715	11.4	74.4	13.6
PD35746b	Control	Male	147	88	23.8	7.5	1.5	4.8	2.8	1.4	86.3	14.3	4.8	4.4	38.3	220	14.2	65.3	17.2
PD35561b	Control	Male	154	94	31.2	5.3	1	3.2	2.6	1.5	88.6	13.4	7.7	4.9	43.3	262	15.2	77.5	15.5
PD35538b	Control	Female	118	74	26.2	5.7	2.3	3	0.9	2.8	91.6	14.7	7.6	4.2	38.7	215	12.7	62.9	17.6
PD35718b	Control	Male	132	75	24.8	5.5	1.1	3.4	2.3	1.2	93	12.6	4	4.6	43.2	186	15.6	63.7	16.5
PD35767b	Control	Male	139	90	29.3	6.7	1.2	4.7	1.8	-	-	-	-	-	-	-	-	51.7	12.3
PD35761b	Control	Male	146	80	29.9	4.4	0.9	2.5	2.2	2.4	97.9	13.8	10.6	4.3	42.6	210	14.5	71	12.8
PD35562b	Control	Male	113	66	26.5	6	1.1	3.5	3.1	1.5	90	13	4.7	4.8	42.8	201	13.8	69.2	10.4
PD35714b	Control	Male	129	84	24.5	6.5	1	3.8	3.8	1.2	87.6	14	4.9	5	43.5	186	14.9	59.3	16.4
PD35648b	Control	Female	130	79	27.4	6	2	3.5	1.2	1.8	89.3	12.8	4.4	4.2	38	213	13.1	76	14.2
PD35516b	Control	Female	125	70	22.9	6.1	2.1	3.6	0.9	2.5	88.5	12.6	8.2	4.7	41.4	261	14.2	66.4	16.5
PD35778b	Control	Male	171	100	29.8	5.8	1.6	3.9	0.8	2	88.5	13.3	6.3	5.2	45.7	185	15.8	66.4	17
PD35621b	Control	Male	138	84	23.5	5.7	2.4	2.5	1.9	2	98.1	12.4	4.3	4.5	44.6	176	14.6	61.1	17.7
PD35530b	Control	Female	123	74	23.3	5.9	2.8	2.8	0.8	1.5	82.8	13	5.5	3.7	30.9	267	11	50.3	16.1
PD29851b	pre-AML	Female	130	80	27.7	6.8	1.2	4.2	3.1	3	91.8	12.9	8.7	4.7	43	238	15.1	60.4	12.2
PD29874b	pre-AML	Male	110	68	25.4	5.5	1.6	3.4	1.1	1.6	86.7	14	6.7	5.4	47.2	228	16.1	74.2	3.8
PD35788b	Control	Female	147	80	20.9	10	2.3	6.9	1.8	2.6	95.9	12.3	8.5	4.1	38.9	282	14.1	72.6	17
PD35675b	Control	Female	139	84	29.7	6.1	1.6	3.4	2.5	2.5	83.4	14.1	7.3	4.5	37.9	319	13	58.3	17
PD30116b	pre-AML	Male	152	88	26.7	5.5	1.1	3.6	1.8	1.7	90.2	14.8	6.3	4.2	38	183	13.7	72.8	5.1
PD35719b	Control	Male	136	98	29	7.6	1	6.1	1.2	1.8	91.7	12.2	6	5.2	47.4	206	16	56	17.9
PD35531b	Control	Female	144	93	28.3	5.7	1.4	3.5	1.8	1.9	88.4	13.3	6.6	4.8	42.5	229	13.8	62.1	17.6
PD35774b	Control	Female	110	70	28.1	4.9	2.1	2.4	0.9	1.1	102	13.6	3.2	3.6	37	227	11.9	65.9	15.5
PD35644b	Control	Female	137	79	32.8	8	1.4	5.7	2	1.9	90.4	14.9	6.4	4.4	39.4	157	12.7	62.8	17.7
PD35765b	Control	Female	132	82	19.2	5.5	2.1	2.5	2	2.5	90.1	13.4	9.8	4.8	42.8	322	15.5	73.8	14.6
PD35783b	Control	Female	146	88	27.7	5.2	1.6	2.9	1.7	1.7	87.2	12.3	4.3	4.1	35.7	253	12.6	52.3	16.7
PD35628b	Control	Male	124	86	30.1	7.3	1.3	5.2	1.9	1.9	90.4	15.2	5.3	4.6	42	225	14.7	79	16.7
PD35766b	Control	Female	155	88	27.2	6.6	1.7	3.5	3.1	2	92.2	11.8	5.2	4.7	43.1	148	15.2	76.3	13.9
PD35629b	Control	Female	152	86	27.6	5.7	1.3	4	1	1.8	89.2	13.3	6.2	4.1	36.5	275	12.8	78.8	16.6
PD35585b	Control	Female	104	64	20.9	6.9	1.6	4.9	1	1.6	88.3	12.8	4.3	4.8	42.4	217	14.2	71.8	17
PD35592b	Control	Male	117	76	24.8	6	1.5	3.3	2.7	2.5	84.9	14.4	7.3	4.9	41.9	178	15	58.3	16.9
PD35588b	Control	Female	134	80	27.7	5.1	1.5	2.7	2.1	2.1	88.4	13.6	5.4	5	44.5	207	14.3	57.7	17.7
PD35713b	Control	Male	102	64	21	5.3	1.4	3.7	0.6	1.2	93.1	14.3	4.3	4.7	43.3	159	14.2	59.6	15.6
PD35568b	Control	Male	145	88	25.4	5.5	1.2	3.3	2.3	1.7	91.9	13.5	5.4	4.8	44.2	321	15.4	71.8	16.3
PD29856b	pre-AML	Male	130	82	30.3	4	0.9	2.1	2.3	2.6	83.4	13.6	7	6.4	53.1	238	17.9	61.5	17.8
PD35557b	Control	Male	162	82	25.8	5.7	1.6	3.6	1.2	1	95.2	13.5	6.3	4.7	44.8	229	14.6	78	15.3
PD35603b	Control	Male	123	76	33.9	5.4	1.2	3.4	1.8	2.3	89.1	13.5	8.2	6.3	56	379	16.6	43.9	15.6
PD35669b	Control	Male	176	94	23.9	5.7	1.3	3.4	2.3	1.6	90	13.3	4.2	4.1	37.1	174	14.3	73.1	16.8
PD29935b	pre-AML	Male	140	92	28	6	1.2	4.3	1.2	2.1	81.5	14.6	6.7	4.4	36	304	12.4	65	17.7
PD29960b	pre-AML	Female	106	64	20.8	3.9	1.6	1.6	1.6	1.8	97	14.4	7.1	3.8	37.3	120	13.2	59.6	7.9
PD35602b	Control	Male	139	82	28.4	7.9	1.6	5.3	2.4	2.3	78.8	12.6	8.1	5.6	44.4	141	15.3	66.7	17.5
PD35535b	Control	Male	150	90	28.8	5.6	1.6	3.6	0.9	2.1	88.8	13.5	7.1	5.6	49.5	272	16.5	61.8	15.9
PD35584b	Control	Male	126	76	29.5	4.8	0.9	3.4	1.3	2.4	95.2	12.6	7.7	4.5	43	254	14.1	59	17.9
PD35532b	Control	Female	142	82	26.9	5.4	1.2	3.3	2	2.3	89.1	13	7.4	4.5	40.3	255	14.4	65.8	16.2
PD30010b	pre-AML	Male	138	78	28.5	6.6	1	3.1	5.6	1.9	105	15.1	3.9	3.3	34.2	106	12.4	70	12.7
PD35513b	Control	Male	152	88	26.8	3.6	1.8	1.1	1.6	1.2	91.8	12.5	5	4.5	41.1	217	15	70.5	16.2
PD35772b	Control	Male	156	90	35.9	6.5	1.3	4.3	2.1	2.5	87.2	12.7	8.7	4.7	40.5	269	14.8	60.5	16.8
PD35604b	Control	Male	129	72	27.4	3.6	1.4	1.9	0.8	2.2	88.1	12.7	6.3	5.2	45.7	229	15.9	58.4	17.4
PD35606b	Control	Male	130	86	28.4	7.3	1.7	5.1	1.3	2.5	100	13.2	5.8	4.6	46.3	213	15	66.1	17.7
PD35618b	Control	Male	158	92	27.7	5.2	1.6	3.1	1.1	2	89.4	13.2	5.1	4.8	42.8	241	15.4	77.9	13.2
PD35755b	Control	Female	100	60	23.9	6.9	1.5	4.5	2	1.9	91.4	13.6	5.3	4.1	37.4	265	12.9	71.4	16
PD35575b	Control	Female	134	81	24.2	5.9	3	2.2	1.6	1.3	86.9	13.3	5.1	4.6	39.6	243	13.4	79.3	14
PD35655b	Control	Male	158	91	25.2	6.2	1	4.1	2.6	1.8	90.2	14.1	6.5	4.2	38.2	347	13.1	66.2	16.4
PD35630b	Control	Female	156	90	24.7	6.4	2.8	3.4	0.6	2.1	88.8	13.3	6.5	4.2	36.8	236	13.2	70.3	17.4
PD35680b	Control	Female	149	88	22.3	6.4	1.5	4.2	1.7	2.4	93.7	12.6	6.6	4.1	38.6	222	12.2	76.4	17.5
PD29929b	pre-AML	Female	154	84	26.8	9.1	1.9	5.7	3.3	3.1	94.7	12.7	8.4	4.6	43.4	226	15.4	72	6.6
PD35559b	Control	Male	133	93	34.7	6	0.9	4.4	1.7	1.8	91	13.8	5.8	5.2	47.7	214	16	71.5	16.7
PD35649b	Control	Female	127	74	25.8	8.7	2	5.6	2.5	1.2	89.5	14.1	6.3	4.2	37.8	170	13.3	69.6	17.1
PD35537b	Control	Female	157	96	31.7	6.5	1.7	3.3	3.5	2.1	87.2	12.1	6	4.8	41.5	324	14.2	71.6	16.5
PD35563b	Control	Male	170	99	28.5	6.7	1.2	4.5	2.3	1.8	94	13.2	6.3	4.6	43.1	255	15.7	71.6	17
PD35666b	Control	Female	152	90	24.7	6.1	1.5	3.2	3.2	2.3	87.7	12.6	8	4.6	40.1	178	14.2	76.8	17.4
PD35577b	Control	Female	148	98	24.8	5.1	1.7	2.8	1.4	2	92	12.8	8.5	4.6	42.6	245	15.7	55.7	16.9
PD35687b	Control	Male	128	94	22.5	5.5	1.6	3.4	1.1	2.1	86.7	12.3	6.5	4.6	39.8	338	14	57.2	16.4
PD35695b	Control	Male	165	91	24.2	7.3	1.3	5	2.3	1.7	91.9	12.8	6	4.8	44.1	216	15.7	77.2	15.9
PD29918b	pre-AML	Male	135	82	28.7	5.3	0.9	3.7	1.7	2.1	97.5	13.9	6.3	5.4	53.2	189	16.1	82	13.4
PD35653b	Control	Male	140	83	25.2	4.3	1	2.9	0.9	2.5	81.2	14.3	7.8	5.2	42.5	136	13.8	74.1	13.1

PD35574b	Control	Male	130	72	27.8	6.1	1.4	3.7	2.4	2.2	86.1	12.7	8	4.8	41.2	267	13.3	59.3	17.5
PD35597c	Control	Female	130	78	22.9	6.3	1.8	4.2	0.8	-	-	-	5.8	-	-	-	-	54.6	16.6
PD35510c	Control	Male	114	70	24.1	4.7	1.1	2.8	1.8	1.9	-	-	7.2	-	-	-	-	82.5	13.6
PD35540c	Control	Male	102	70	-	4.2	1.4	2.3	1.3	1.5	-	-	6.8	-	-	-	-	74.9	11.3
PD35731c	Control	Male	141	93	29	8.2	1.2	5	4.5	2.4	-	-	6.9	-	-	-	-	79.3	17.8
PD35762c	Control	Female	152	73	25.5	5.2	1.7	3.2	0.7	1.7	-	-	7.3	-	-	-	-	84.4	16.4
PD35553c	Control	Male	122	56	25.4	4.9	1.1	2.3	3.5	1.8	-	-	5.2	-	-	-	-	79.7	21.5
PD35660c	Control	Female	144	84	26.7	4.9	1.1	3.1	1.5	1.8	-	-	4.4	-	-	-	-	74.2	21.1
PD35533c	Control	Male	149	88	27.5	4.3	1.6	2.2	1.3	2.2	-	-	7.8	-	-	-	-	66.4	16.2
PD35558c	Control	Male	140	71	23.1	3.4	1.5	1.5	0.9	1.2	-	-	6.5	-	-	-	-	82.4	15.9
PD35733c	Control	Female	154	85	29.2	5.7	1.7	3.3	1.6	2.2	-	-	5	-	-	-	-	72.2	20.4
PD35585c	Control	Female	121	68	21.6	6.5	1.6	4.4	1.1	1.5	-	-	4.6	-	-	-	-	79.8	17
PD35768c	Control	Female	142	70	31.8	4.8	1.8	2.6	0.9	1.7	-	-	5.2	-	-	-	-	83.3	21.1
PD35777c	Control	Male	146	84	28.1	3.3	1	0.8	3.4	2.9	-	-	8.4	-	-	-	-	75.4	19.6
PD35787c	Control	Male	132	82	28.5	2.8	1.3	1.3	0.5	1.2	-	-	7.2	-	-	-	-	80.5	20.7
PD35606c	Control	Male	142	89	29.9	5.4	1.7	3.3	1	2	-	-	5.3	-	-	-	-	76.4	17.7
PD35548c	Control	Male	150	82	26.8	5.3	1.4	2.9	2.4	0.8	-	-	5.7	-	-	-	-	88.1	8.7
PD35759c	Control	Female	112	82	23.7	6.5	1.5	3.6	3.1	2.5	-	-	7.6	-	-	-	-	63	16.3
PD35633c	Control	Male	86	40	23.2	4.4	1.2	2.9	0.7	1.1	-	-	7.2	-	-	-	-	84.8	12.4
PD35771c	Control	Male	156	98	25.9	4.9	2	2.5	0.9	1.9	-	-	6	-	-	-	-	73.4	17.7
PD35677c	Control	Female	137	74	25.5	4.2	1.8	2.1	0.8	1.3	-	-	5.7	-	-	-	-	82.4	16.5
PD35584c	Control	Male	108	74	28.3	3.8	1	2.4	1.2	1.7	-	-	6.8	-	-	-	-	68	17.9
PD35582c	Control	Female	146	81	25.8	6.8	1.5	4.3	2.4	2.7	-	-	8.2	-	-	-	-	83.2	16.3
PD35595c	Control	Male	130	68	25.6	5.9	1.7	3.7	1.3	2.1	-	-	6.5	-	-	-	-	80.6	21.3
PD35613c	Control	Female	148	88	31.8	5.8	1.9	2.9	2.3	1.8	-	-	4.3	-	-	-	-	76.2	21
PD35552c	Control	Male	112	74	24.2	4.9	1.4	3.1	1	1.5	-	-	9.1	-	-	-	-	71.9	18.9
PD35652c	Control	Female	120	70	34.4	3.9	1	2.6	0.7	1.2	-	-	3.7	-	-	-	-	72.8	19.6
PD35586c	Control	Male	142	88	19.8	5.4	1.9	3.2	0.7	1.8	-	-	7.4	-	-	-	-	79.5	17.3
PD35516c	Control	Female	147	83	23.6	4.3	1.9	2	1	1.7	-	-	6.7	-	-	-	-	74.6	16.5
PD35575c	Control	Female	151	82	23.2	6.5	2.8	3	1.8	1.4	-	-	5.7	-	-	-	-	87	14
PD35644c	Control	Female	130	70	31.8	5.1	1.5	3.1	1.3	1.6	-	-	6.7	-	-	-	-	73	17.7
PD35756c	Control	Male	142	80	23.4	4.6	0.8	2.9	2.1	1.2	-	-	6.3	-	-	-	-	87.6	16.8
PD35579c	Control	Female	138	66	30.9	4.2	1.6	1.9	1.6	2.2	-	-	7.1	-	-	-	-	78.2	20.6
PD35732c	Control	Female	116	72	25.7	4.7	1.8	2.6	0.8	-	-	-	3.6	-	-	-	-	65.9	17.1
PD35719c	Control	Male	139	94	28.6	6.9	1.1	5.3	1.2	1.3	-	-	5.1	-	-	-	-	66.2	17.9
PD35564c	Control	Male	127	58	33.1	3.7	1.1	1.8	2	1.9	-	-	9.2	-	-	-	-	84.2	13.7
PD35779c	Control	Male	142	80	28.2	5.1	0.9	2.6	3.6	2.7	-	-	7.4	-	-	-	-	69	10
PD35600c	Control	Female	154	67	31.4	8.8	1.9	5.6	3	1	-	-	6.1	-	-	-	-	86.7	7.6
PD35778c	Control	Male	138	83	29.8	5.2	1.1	2.7	3.1	2.3	-	-	5.9	-	-	-	-	76.6	17
PD35758c	Control	Male	142	90	35.2	4.4	1.5	1.9	2.2	1.6	-	-	7.3	-	-	-	-	75.8	21
PD35630c	Control	Female	138	73	23	5.5	2.2	3.1	0.6	1.4	-	-	7.7	-	-	-	-	80.3	17.4
PD35592c	Control	Male	149	84	23.7	5.8	1.6	3.6	1.3	2.3	-	-	7	-	-	-	-	66.7	16.9
PD35738c	Control	Female	152	90	23.9	4	1.6	2	0.9	1.6	-	-	6.6	-	-	-	-	74.6	22.1
PD35545c	Control	Male	106	72	31.7	3.8	1.3	1.8	1.6	1.1	-	-	5.1	-	-	-	-	82	20.4
PD35568c	Control	Male	136	82	25.4	5.4	1.4	3	2.4	1.6	-	-	6.6	-	-	-	-	79.9	16.3
PD35684c	Control	Female	126	76	22.8	5.9	2.1	3.1	1.6	2.4	-	-	9.4	-	-	-	-	56.5	19.7
PD35574c	Control	Male	126	70	29	5.7	1.3	3.9	1.1	2.8	-	-	7	-	-	-	-	68.2	17.5
PD35559c	Control	Male	110	70	31.4	4.2	1.6	2.2	0.9	-	-	-	6.8	-	-	-	-	80.5	16.7
PD35561c	Control	Male	160	80	32.2	4.5	1.2	2.3	2.4	1.4	-	-	7	-	-	-	-	85.9	15.5
PD35665c	Control	Male	128	73	25.4	4.5	1.6	2.3	1.4	1.7	-	-	6.4	-	-	-	-	80.2	22
PD35724c	Control	Male	137	71	26.3	4.8	0.9	-	4.7	2.3	-	-	9.2	-	-	-	-	77.9	21.5
PD35534c	Control	Female	111	66	22.2	7.2	2.2	4.5	1.1	1.6	-	-	4.5	-	-	-	-	74.5	16.8
PD35669c	Control	Male	139	69	23.9	4.4	1.2	2.4	1.8	1.4	-	-	5.9	-	-	-	-	81.5	16.8
PD35624c	Control	Female	160	91	27.3	6.5	0.7	-	4.6	1.1	-	-	4.7	-	-	-	-	84.7	21.7
PD35647c	Control	Female	138	70	29.1	4.3	1.5	2.3	1.2	1.8	-	-	6.9	-	-	-	-	87.3	21.8
PD35544c	Control	Male	165	92	26.4	6.3	1.3	3.5	3.3	1.6	-	-	4.2	-	-	-	-	61.2	17.2
PD35616c	Control	Male	129	72	33.8	5.1	1	3.5	1.5	1.9	-	-	7	-	-	-	-	79.6	21
PD35520c	Control	Female	114	68	29.3	4.7	1.4	1.5	4.1	2.4	-	-	7.4	-	-	-	-	55.6	18
PD35634c	Control	Female	133	63	24.7	7.7	2.7	4.4	1.4	2.1	-	-	5.6	-	-	-	-	61.4	7.9
PD29914c	Control	Male	121	60	33.3	3	0.9	1.6	1.2	1	-	-	5.6	-	-	-	-	66.8	1.1
PD35538c	Control	Female	130	84	26.8	4.9	1.9	2.8	0.5	2	-	-	6.4	-	-	-	-	72.9	17.6
PD35709c	Control	Female	135	70	22.1	5.1	2.8	1.9	1	1.8	-	-	7.9	-	-	-	-	83.7	17.4
PD35620c	Control	Female	156	93	19.3	6	2.8	2.8	0.9	-	-	-	4.6	-	-	-	-	66.9	17.6
PD35560c	Control	Female	137	72	31.8	3.7	1.7	1.3	1.6	0	-	-	6.6	-	-	-	-	77.7	17
PD35770c	Control	Male	140	85	24.6	5.5	1.6	3.5	0.9	1	-	-	4.4	-	-	-	-	81	8
PD35556c	Control	Female	129	76	24.3	4.9	1.8	2.5	1.4	2.8	-	-	8.6	-	-	-	-	76.4	20.8
PD35635c	Control	Female	144	92	33.8	6.4	1.3	4	2.6	2.7	-	-	6.1	-	-	-	-	72.9	17.8
PD35661c	Control	Male	122	72	29.3	4.4	1.3	2.5	1.4	1.2	-	-	7.2	-	-	-	-	81.9	17.9
PD35773c	Control	Female	111	70	32.3	4	1.8	1.7	1.1	1.4	-	-	6.8	-	-	-	-	85.5	19.3
PD35508c	Control	Female	144	74	27.2	4.5	1.9	2.1	1.3	1.1	-	-	5.3	-	-	-	-	79.4	21.4
PD35578c	Control	Female	208	102	22.8	6.7	1.8	4.2	1.7	2.1	-	-	7.2	-	-	-	-	86.1	7.7
PD35761c	Control	Male	144	76	31.7	3.1	1.1	1.5	1.1	1	-	-	10.4	-	-	-	-	81.2	12.8
PD35594c	Control	Female	163	94	27.6	6.4	1.4	4.5	1.3	2.4	-	-	5.2	-	-	-	-	76.7	17.6
PD35637c	Control	Male	123	80	25.8	5.7	1.7	3.7	0.7	1.5	-	-	5.8	-	-	-	-	69.5	15.6
PD29935d	pre-AML	Male	102	72	27.9	3.4	1	1.7	1.6	1.6	-	-	7.6	-	-	-	-	73.1	17.7
PD35612c	Control	Female	140	72	39.1	4.2	1.8	1.8	1.4	1.6	-	-	7.6	-	-	-	-	71.2	21.9
PD35699c	Control	Male	142	90	27.7	5.3	1.6	3.2	1.1	2	-	-	5.7	-	-	-	-	66.7	17.1
PD35570c	Control	Male	150	88	33.6	5.9	1.3	3.7	2.1	1.4	-	-	5.9	-	-	-	-	68.6	16.9
PD35656c	Control	Female	150	82	27.6	5.9	1.2	3.3	3.1	2.6	-	-	3.6	-	-	-	-	75.4	10.2
PD35526c	Control	Male	139	69	27	4.1	1.4	2.2	1.3	1.8	-	-	7.2	-	-	-	-	80.4	20.7
PD35581c	Control	Male	136	77	22.5	5.1	1.1	3.5	1.2	2.1	-	-	10.2	-	-	-	-	77.2	17.3
PD35788c	Control	Female	146	70	22.8	4	1.9	1.8	0.8	1.7	-	-	6.3	-	-	-	-	82.4	17
PD35722c	Control	Male	136	84	28.3	4.6	1.2	2.4	2.3	1.5	-	-	5.1	-	-	-	-	70.9	21.5
PD35590c	Control	Female	142	62	26.8	4.4	1.7	2.1	1.4	1.8	-	-	6.9	-	-	-	-	78.3	17
PD35532c	Control	Female	124	66	27.4	4.6	1.4	2.6	1.4	1.7	-	-	6.4	-	-	-	-	73.9	16.2
PD35760c	Control	Male	126	62	34.9	3.9	1	2.4	1.1	1.5	-	-	7.6	-	-	-	-	80.5	19.8
PD35748c	Control	Female	146	72	26.5	3.4	1.4	1.5	1.1	1.6	-	-	8.1	-	-	-	-	76.4	8.3
PD35740c	Control	Male	128	66	25.2	4.4	2.1	1.8	1.1	1.2	-	-	4.9	-	-	-	-	84.1	21.8
PD35563c	Control	Male	151	82	30.2	4.4	2	2	1	0.8	-	-	10.5	-	-	-	-	80	17
PD30058d	Control	Female	147	77	26.5	5.9	1.5	3.6	1.8	-	-	-	-	-	-	-	-	67.4	19.3
PD35785c	Control	Male	139	90	28	6.3	1.7	4.1	1.2	1.1	-	-	7.6	-	-	-	-	64.6	15.8
PD35695c	Control	Male	138	72	25.3	6.8	1.2	4.2	3.2	1.4	-	-	7.1	-	-	-	-	85	15.9
PD35659c	Control	Male	126	68	23.6	5.2	1	2.9	2.9	2.1	-	-	13.4	-	-	-	-	80.4	20.3
PD35576c	Control	Male	141	77	32.7	4.7	1.4	2.4	2.1	2.1	-	-	8.8	-	-	-	-	68.4	8.7
PD35583c	Control	Male	112	61	24.9	2.7	1.1	1.2	0.8	1.4	-	-	9.3	-	-	-	-	83.9	11.5
PD35702c	Control	Male	124	76	27.7	3.7	1	1.4	2.9	1.3	-	-	4.7	-	-	-	-	81.2	22
PD35772c	Control	Male	161	88	39.4	5.9	1.2	3.5	2.7	2.8	-	-	10.6	-	-	-	-	68.8	16.8
PD35765c	Control	Female	151	82	18.6	6.2	3.2	2.5	1.3	0.5	-	-	10	-	-	-	-	81.9	14.6
PD35673c	Control	Male	116	71	27.6	6.6	1.8	4.3	1.2	1.1	-	-	5.3	-	-	-	-	61.9	19.8
PD35573c	Control	Female	196	86	20.3	5	1.1	3.5	1	2.3	-	-	6.6	-	-	-	-	84.1	19.3
PD35618c	Control	Male	130	70	27.8	3.7	1.2	2	1.2	1.2	-	-	5.4	-	-	-	-	86.5	13.2
PD35591c	Control	Female	152	70	23.1	6	1.6	3.8	1.4	2.3	-	-	8.1	-	-	-	-	66.5	17.2
PD35519c	Control	Female	137	74	32.6	7.7	1.9	4.6	2.7	3.1	-	-	8.2	-	-	-	-	77.7	20.2
PD35782c	Control	Male	144	86	20.1	5.5	2.7	2.5	0.7	-	-	-	4.7	-	-	-	-	61.5	20.5

PD35580c	Control	Female	110	70	22.7	3.7	1.4	1.8	1.2	2	-	-	6.4	-	-	-	-	81	21.3
PD35639c	Control	Female	143	84	25.5	6	1.4	3.8	2	2.1	-	-	5.1	-	-	-	-	74.4	17.3
PD35767c	Control	Male	148	82	29.5	6	1.6	3.9	1.2	2.2	-	-	8.1	-	-	-	-	60.4	12.3
PD35514c	Control	Female	131	70	24.2	7.7	2	5	1.7	1	-	-	8.7	-	-	-	-	84.7	20.7
PD35555c	Control	Female	140	76	19.2	5.3	2.6	2.2	1.2	0.8	-	-	6.6	-	-	-	-	80.8	16.1
PD35607c	Control	Male	160	88	26.4	4.9	1.3	3.1	1.1	1.2	-	-	6.6	-	-	-	-	87.3	17.2
PD35755c	Control	Female	118	74	23.3	5.5	1.3	3.4	1.8	1.6	-	-	4.7	-	-	-	-	77.8	16
PD35698c	Control	Female	152	76	27.4	5.2	1	3.6	1.4	1.8	-	-	5.9	-	-	-	-	81.1	17.4
PD35648c	Control	Female	121	74	27.1	5.7	1.7	3.4	1.4	1.7	-	-	4.2	-	-	-	-	85.8	14.2
PD35746c	Control	Male	166	87	23.6	5.2	1.8	2.3	2.6	1.9	-	-	6.1	-	-	-	-	74.9	17.2
PD35596c	Control	Male	106	62	18.2	4.1	1.4	2.4	0.7	0.6	-	-	2.4	-	-	-	-	58.6	19.8
PD35577c	Control	Female	142	94	27.9	5	1.8	2.6	1.4	2.1	-	-	6.2	-	-	-	-	63.8	16.9
PD35571c	Control	Female	131	61	23.8	4.5	1.6	2.3	1.4	1.7	-	-	5.4	-	-	-	-	61.1	17.3
PD35710c	Control	Female	128	66	29.1	5.2	1.6	3.1	1.2	-	-	-	4.2	-	-	-	-	80	17.2
PD35554c	Control	Male	129	64	29.9	5	0.9	2.6	3.3	1.9	-	-	7.5	-	-	-	-	82.3	18
PD29918d	pre-AML	Male	146	79	28.7	4.3	0.9	2.5	2	1.5	-	-	3.5	-	-	-	-	89.9	13.4
PD35766c	Control	Female	148	68	30.5	5.3	1.8	2.5	2.3	1.9	-	-	7.4	-	-	-	-	85.9	13.9
PD35565c	Control	Male	132	72	32.9	3.7	0.8	-	5	2	-	-	5.9	-	-	-	-	69.7	21.2
PD35562c	Control	Male	127	69	24.6	4.1	1.4	2	1.7	1.3	-	-	6.6	-	-	-	-	77.4	10.4
PD35623c	Control	Female	148	72	25.3	4.6	2.1	2	1.1	1.8	-	-	7.5	-	-	-	-	78.9	21.1
PD35569c	Control	Male	134	83	22.9	6	1.7	3.8	1.2	2.1	-	-	6.1	-	-	-	-	64.6	19.1
PD35789c	Control	Male	124	76	25.2	6.7	1.1	4.3	3	2.3	-	-	5.4	-	-	-	-	60.9	17.1
PD35786c	Control	Male	140	94	27.1	6.7	1	4.2	3.3	1.5	-	-	6.3	-	-	-	-	70.2	21.5
PD35550c	Control	Female	136	60	33	4.7	1.6	2.4	1.6	1.8	-	-	7.2	-	-	-	-	79.8	21.2
PD35622c	Control	Female	141	78	25.4	5.7	1.9	3.2	1.5	2.3	-	-	5.5	-	-	-	-	65.8	21.4
PD35780c	Control	Male	143	86	26	4	1.5	2.3	0.6	-	-	-	-	-	-	-	-	76.4	19.5
PD35546c	Control	Female	120	71	22.3	4.3	1.4	2.6	0.8	1.2	-	-	7	-	-	-	-	61.3	15.6
PD35763c	Control	Male	138	82	27.8	3.5	1	1.8	1.6	1.9	-	-	5.9	-	-	-	-	77.4	20.5
PD35783c	Control	Female	180	92	27.6	5.4	1.7	3.1	1.5	1.7	-	-	5.9	-	-	-	-	60.5	16.7
PD35566c	Control	Female	109	72	19.9	5.3	2.2	2.8	0.7	2	-	-	6.4	-	-	-	-	66	17.7
PD35757c	Control	Female	132	80	28.6	7.3	1.4	-	4.8	1.5	-	-	6.2	-	-	-	-	75.2	18.5
PD35542c	Control	Female	150	86	30.4	5.6	1.4	3.3	2.1	2.2	-	-	7.2	-	-	-	-	74.3	16.3
PD35605c	Control	Female	153	88	24.1	4	1.4	2.2	1	-	-	-	3.3	-	-	-	-	75.8	21.7
PD35528c	Control	Female	156	83	31.4	7.8	1.8	5.3	1.6	2	-	-	6.5	-	-	-	-	64.2	18.8
PD35589c	Control	Female	121	68	21.9	4.8	1.5	2.8	1.3	0.9	-	-	5.7	-	-	-	-	76.1	20.9
PD35557c	Control	Male	148	74	26.4	4.5	1.7	2.7	0.4	0.7	-	-	5.2	-	-	-	-	86.9	15.3
PD35531c	Control	Female	158	95	26.1	4.6	1.5	2.4	1.6	1.6	-	-	7.4	-	-	-	-	70.5	17.6
PD35507c	Control	Male	178	117	24.8	6.5	2	4.3	0.6	1.8	-	-	5.6	-	-	-	-	68.4	21.9
PD35704c	Control	Male	130	84	27.3	3.7	1.2	1.7	1.8	3	-	-	7.2	-	-	-	-	64.6	16.2
PD35764c	Control	Male	133	76	27.1	4	1.1	2.4	1.1	1.7	-	-	8.9	-	-	-	-	76.1	21.4
PD35628c	Control	Male	139	86	28.5	4.1	1.4	2.3	1	1.4	-	-	6.1	-	-	-	-	89.9	16.7
PD35781c	Control	Male	140	78	30	4.1	1	2.6	1.3	1.6	-	-	7.4	-	-	-	-	73.3	21.6
PD35588c	Control	Female	110	71	29	4.8	1.4	2	2.9	2.1	-	-	6.7	-	-	-	-	67.6	17.7
PD35662c	Control	Female	140	82	22.5	6.9	1.6	5	0.8	1.2	-	-	6.2	-	-	-	-	83.4	17.7
PD35587c	Control	Female	106	64	30	5.2	1.8	2.9	1.3	3.1	-	-	9.6	-	-	-	-	82	21.3
PD35726c	Control	Male	152	80	25.4	6	1.6	4.1	0.8	1	-	-	6.2	-	-	-	-	85.6	15.9
PD35539c	Control	Female	123	72	24.5	4.9	1.6	2.5	1.8	2.1	-	-	7.5	-	-	-	-	72.2	16.4
PD35572c	Control	Male	134	90	31.6	5.3	1.4	2.8	2.5	2.7	-	-	7.9	-	-	-	-	60.5	20.2
PD30089c	pre-AML	Female	142	60	28.5	4.6	1.4	2.9	0.7	1.4	-	-	4	-	-	-	-	75.6	13.5
PD35697c	Control	Female	138	68	23.2	5.7	1.5	3.6	1.5	2.3	-	-	6.9	-	-	-	-	78.6	20.7
PD35769c	Control	Female	162	89	24.6	6.3	2.4	3.4	1.3	1.6	-	-	5.7	-	-	-	-	74.2	15.7

Appendix 3: Childhood cancer survivor cohort details

Study ID	Sex	Diagnosis	Age at diagnosis	Months since cytotoxic treatment
1	female	NB	15.4	64.3
2	male	RMS	11.1	21.7
3	female	ALL	5.7	132.4
4	NA	ALL	NA	NA
5*	female	ALL	1.1	106.4
6	female	ALL	6.1	80.3
7	male	NB	6.3	231.9
8§	male	NHL	4.7	176.2
9	female	ALL	1.7	52.6
10§	male	ALL	6.9	298.2
11	female	GCT	9.3	25.9
12	male	RMS	6	102.9
13	female	NHL	7.1	103.9
14	male	ALL	6.9	177.4
15	female	NHL	9.4	80.1
16	male	NB	0.6	94
17*	male	LL	5.8	55.4
18	male	HL	14.8	136.6
19	male	WT	0.8	57.3
20	male	RMS	3.1	47.6
21	female	ALL	9.1	35.7
22	male	HL	10.9	43.5
23	male	ALL	4	49.5
24	male	HL	14.2	42.5
25	male	HB	0.3	112.9
26*	male	ALL	0.6	81.1
27	male	HL	7.1	86.2
28	male	GCT	15.4	26.7
29	male	RMS	5.8	76.2
30§	male	NHL	15.5	46.6
31	male	HL	25.4	48.5
32§	male	ES	4.6	141.5
33	male	LL	9.3	112.9
34	female	ES	3.3	74.3
35*	male	NB	2.3	102.9
36	male	NHL	2	46.4
37	male	NB	3.4	166.4
38	female	NB	1.7	124.8
39*	male	LL	3.2	112.9
40§	male	NB	0.5	289.3
41	female	WT	3.1	105.9
42	male	NB	0.9	268.4
43	female	NB	0.6	238.8
44	male	NHL	5.8	183.2

45	male	RMS	8.4	192.2
46	male	NRSTS	4.3	105.9
47	male	ALL	3	58.3
48	male	ALL	3.9	35.7
49	male	NB	5.5	NA
50	female	ES	13.4	69.2
51*	male	ALL	4.7	89.1
52	female	CCA	12.8	41.5
53	male	NB	4	73.3
54	female	WT	4.8	63.4
55	male	HL	15.3	46.4
56*	male	ALL	1.5	44.5
57	male	NPC	15.9	35.4
58	female	NHL	8.7	25.7
59	male	ALL	4.5	59.4
60	male	ALL	3.6	35.9
61	male	NB	5.8	34.5
62	male	NHL	2.6	59.3
63	male	NHL	9.1	62.4
64	female	RMS	3	80.1
65	female	NB	0.3	138.6
66	female	RMS	1.1	45.4
67	female	ALL	2.4	54.4
68	male	NHL	3.7	212.9
69	female	NRSTS	11	38.1
70	male	NB	0.4	45.4
71	female	LCH	3.7	88.1
72*	female	LCH	3.1	69.2
73	female	WT	3.8	142.7
74	female	GCT	0	131.8
75	male	GCT	15.4	NA
76§	female	WT	4.9	96.1
77	female	ALL	8.2	45.4
78§	female	NB	1.1	39.6
79	male	ALL	4.5	77.2
80§	male	NB	1.3	194.1
81	male	ALL	3.3	48.5
82	female	NB	0.3	75.2
83	male	ALL	3	75.2
84	male	ES	10.7	100

RMS, rhabdomyosarcoma; ALL, acute lympoblastic leukaemia; NB, neuroblastoma; NHL, non-Hodgkin lymphoma; GCT, germ cell tumour; LL, lymphoblastic lymphoma; HL, Hodgkin lymphoma; WT, Wilms tumour; ES, Ewing sarcoma; NRSTS, non-rhabdomyosarcoma soft tissue sarcoma; NPC, nasopharyngeal sarcoma; CCA, choriocarcinoma; LCH, Langerhans cell histiocytosis; NA, no data. Patients who received a haematopoietic stem cell transplant (HSCT) are indicated with the symbols * (allogeneic HSCT) or § (autologous HSCT).

Appendix 4: Custom myeloid cancer gene panel

ABL1	*CSF2RB*	*FBXW7*	*MLL2*	*PPFIA2*	*SMG1*
ASXL1	*CSF3R*	*FLT3*	*MLL3*	*PRPF40B*	*SMPD3*
ASXL2	*CTCF*	*FNDC1*	*MLL5*	*PRPF8*	*SRSF2*
ASXL3	*CUL1*	*GATA1*	*MPL*	*PTEN*	*STAG1*
ATRX	*CUL2*	*GATA2*	*MYB*	*PTPN11*	*STAG2*
BCOR	*CUL3*	*GNAS*	*MYC*	*PTPRT*	*STAT5B*
BRAF	*CUX1*	*GNB1*	*MYH11*	*RAD21*	*SUZ12*
CACNA1E	*DAXX*	*HRAS*	*NF1*	*RAD51*	*TERT*
CBFB	*DCAF7*	*IDH1*	*NOTCH1*	*RARA*	*TET2*
CBL	*DCLK1*	*IDH2*	*NOTCH2*	*RB1*	*TP53*
CBLB	*DIAPH2*	*IRF1*	*NPM1*	*RIT1*	*U2AF1*
CBLC	*DNMT1*	*JAK2*	*NRAS*	*RPS6KA6*	*U2AF2*
CBX7	*DNMT3A*	*JAK3*	*PDS5B*	*RUNX1*	*UGT2A3*
CDH23	*EED*	*KDM6A*	*PHACTR1*	*SETBP1*	*WT1*
CDKN2A	*EP300*	*KIT*	*PHF6*	*SF1*	*ZFP36*
CEBPA	*EPOR*	*KRAS*	*PHF8*	*SF3B1*	*ZRSR2*
CNTN5	*ETV6*	*LUC7L2*	*PHIP*	*SH2B3*	
CREBBP	*EZH2*	*MED12*	*PIK3CA*	*SMC1A*	
CSF1R	*FAM5C*	*MLL*	*PML*	*SMC3*	

Appendix 5: Multiplex PCR primer sequences

PLEX	PRIMER NAME	GENE	TARGETED EXON/CODON	PRIMER SEQUENCE3
1	ASXL1_exon12_a_F	ASXL1	exon12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGACCCTCGCAGACATTAmAA
1	ASXL1_exon12_a_R	ASXL1	exon12	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCTGTAGATCTGACGTACACmUT
1	ASXL1_exon12_b_F	ASXL1	exon12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTGGTGATGGTGGTGmAG
1	ASXL1_exon12_b_R	ASXL1	exon12	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGCATCTCCTAGCCCATmCT
1	ASXL1_exon12_c_F	ASXL1	exon12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTACTACAGAGGGCTACAGTmUG
1	ASXL1_exon12_c_R	ASXL1	exon12	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTTGCTCCTCATCATCACTTmUC
1	DNMT3A_p.R693C_F	DNMT3A	p.R693C	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTCATGTTCTTGGTGTTTTAT
1	DNMT3A_p.R693C_R	DNMT3A	p.R693C	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTTTTCTCCCCCAGGGTATTTG
1	IDH1_p.R132H_F	IDH1	p.R132H	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAAATGTGTGTAAATATACAGTTAT
1	IDH1_p.R132H_R	IDH1	p.R132H	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTATTATCTGCAAAAATATCCCCC
1	IDH2_p.R172K_IDH2_p.R140Q_F	IDH2	p.R172K, p.R140Q	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGAGGATGGCTAGGCGAGGA
1	IDH2_p.R172K_IDH2_p.R140Q_R	IDH2	p.R172K, p.R140Q	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTCACAGAGTTCAAGCTGAAG
1	JAK2_p.V617F_F	JAK2	p.V617F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCTTTCTTTGAAGCAGCAAG
1	JAK2_p.V617F_R	JAK2	p.V617F	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTAGTTTACACTGACACCTAGCTG
1	KIT_exon17_F	KIT	exon17	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTTTTCTTTTCTCCTCCAAC
1	KIT_exon17_R	KIT	exon17	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTCCTTTGCAGGACTGTCAAG
1	KRAS_p.G12R_F	KRAS	p.G12R	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTTGGATCATATTCGTCCACA
1	KRAS_p.G12R_R	KRAS	p.G12R	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTAAGGTACTGGTGGAGTATTTGA
1	NPM1_p.L287fs*13_F	NPM1	p.L287fs*13	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTTTGGAATTAAATTACATCTGA
1	NPM1_p.L287fs*13_R	NPM1	p.L287fs*13	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTAAAATTTTTTAACAAATTGTTTAAACT
1	NRAS_p.G12D_F	NRAS	p.G12D	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGGTAAAGATGATCCGACAA
1	NRAS_p.G12D_R	NRAS	p.G12D	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGCCAATTAACCCTGATTACTG
1	SF3B1_p.K666N_F	SF3B1	p.K666N	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCCTGTCTCCTAAAGAAAAAA
1	SF3B1_p.K666N_R	SF3B1	p.K666N	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTAGAGCTTTTGCTGTTGTAGC
1	SF3B1_p.K700E_F	SF3B1	p.K700E	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGTAATTTAGATTTATGTCGCC
1	SF3B1_p.K700E_R	SF3B1	p.K700E	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGCATAGTTAAAACCTGTGTTT
1	SRSF2_p.P95L_F	SRSF2	p.P95L	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCTTCGCCGCGGACCTTTGT
1	SRSF2_p.P95L_R	SRSF2	p.P95L	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAGGACGCTATGGATGCCATG
1	U2AF1_p.Q157R_F	U2AF1	p.Q157R	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGGTTGGAAGGAGACATTTAmCT
1	U2AF1_p.Q157R_R	U2AF1	p.Q157R	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAAAAGGCTGTGATTGACTTmGA
1	U2AF1_p.S34F_F	U2AF1	p.S34F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATCACCTGCCTCACTATTmAT
1	U2AF1_p.S34F_R	U2AF1	p.S34F	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTTTCAAAATTGGAGCATGTCmGT
2	PPM1D_exon1_a_F	PPM1D	exon1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGCGCCTAGTGTGTCmUC
2	PPM1D_exon1_a_R	PPM1D	exon1	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCCTTTCCCCGAGACTmUC
2	PPM1D_exon1_c_F	PPM1D	exon1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTCCTCCGTGGCCTTmUT
2	PPM1D_exon1_c_R	PPM1D	exon1	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCAAACAAGCCAGGGAACTTmAC
2	PPM1D_exon3_F	PPM1D	exon3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGAGCTATCTTAGTTGTTmGT
2	PPM1D_exon3_R	PPM1D	exon3	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGCCAAGTAAGGGTTTAGTTmCT
2	PPM1D_exon5_a_F	PPM1D	exon5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTAGTGGCAGCTAAmAT
2	PPM1D_exon5_a_R	PPM1D	exon5	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGTCATCACACAGGTTTCTTGmAC
2	PPM1D_exon6_a_F	PPM1D	exon6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCATAGATTTGTTGAGTTCTmGG
2	PPM1D_exon6_a_R	PPM1D	exon6	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGGAAGGCTATTATTCAAAGAATmCA
2	PPM1D_exon6_c_F	PPM1D	exon6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTAGAAGAGTCCAATTCTGGmCC
2	PPM1D_exon6_c_R	PPM1D	exon6	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTCAACATCGGCACCAAATTTmAA
2	TP53_exon1_F	TP53	exon1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCAAAGACCCAAAACCCAAmAA
2	TP53_exon1_R	TP53	exon1	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTTGATTTGAATTCCCGTTGTmCC
2	TP53_exon10_a_F	TP53	exon10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTGAAGTCTCATGGAAGCCmAG
2	TP53_exon10_a_R	TP53	exon10	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGGACGATATTGAACAATGGmUT
2	TP53_exon10_b_F	TP53	exon10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAGGGACAGAAGATGACAmGG
2	TP53_exon10_b_R	TP53	exon10	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGACTGCTCTTTTCACCCATCmUA
2	TP53_exon11_F	TP53	exon11	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGACTGTAGATGGGTGAAAAmGA
2	TP53_exon11_R	TP53	exon11	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTAGACCTATGGAAACTGTGAGmUG
2	TP53_exon12_F	TP53	exon12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACGTTGTTTTCAGGAAGTCmUG
2	TP53_exon2_F	TP53	exon2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAGAATGGAATCCTATGGCmUT
2	TP53_exon2_R	TP53	exon2	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCATGTTGCTTTTGTACCGTCmAT
2	TP53_exon3_F	TP53	exon3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTAGGCTAAGCTATGATGmUT
2	TP53_exon3_R	TP53	exon3	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCTCCTGGTTGTAGCTAACTmAA
2	TP53_exon5_F	TP53	exon5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTCCACTTGATAAGAGGTCmCC
2	TP53_exon5_R	TP53	exon5	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAAGAGAATCTCCGCAAGAAmAG
2	TP53_exon7_F	TP53	exon7	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGAGGTGGATGGGTAGTAGmUA
2	TP53_exon7_R	TP53	exon7	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTATCTTGGGCCTGTGTTATCmUC
2	TP53_exon9_F	TP53	exon9	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATCAGTGAGGAATCAGAGmGC
2	TP53_exon9_R	TP53	exon9	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTTTCAACTCTGTCTCCTTCCmUC
3	PPM1D_exon1_b_F	PPM1D	exon1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACCGACGGCTGAAGAAmAA
3	PPM1D_exon1_b_R	PPM1D	exon1	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTTCTTGATGAAACCCCACAmAG
3	PPM1D_exon2_F	PPM1D	exon2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTGCAAGAGTGAAATATTmUT
3	PPM1D_exon2_R	PPM1D	exon2	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAAAGAGAAAACGACAGAATmGT
3	PPM1D_exon4_F	PPM1D	exon4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTTCCAACTAATACTTCTTGmCT
3	PPM1D_exon4_R	PPM1D	exon4	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTACCAAAACAATGTTTAGACAmAC
3	PPM1D_exon5_b_F	PPM1D	exon5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGCCATAGTAATCTGCATmCT
3	PPM1D_exon5_b_R	PPM1D	exon5	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTCGAGTTCAAATCCAAAATCCmUG
3	PPM1D_exon6_b_F	PPM1D	exon6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACCCTCAAAAGATCCAGAAmCC
3	PPM1D_exon6_b_R	PPM1D	exon6	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTCGACTTAAGCCATTTCGTCmUA
3	TP53_exon12_R	TP53	exon12	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGATCCCCACTTTTCCTCTmUG
3	TP53_exon4_F	TP53	exon4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAGGCAAAGTCATAGAACCmAT
3	TP53_exon4_R	TP53	exon4	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGACTGTTTTACCTGCAATTmGG
3	TP53_exon6_F	TP53	exon6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAGGCAAGGAAAGGTGATAmAA
3	TP53_exon6_R	TP53	exon6	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTAGGACCTGATTTCCTTACTmGC
3	TP53_exon8_F	TP53	exon8	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGCACATCTCATGGGGTTAmUA
3	TP53_exon8_R	TP53	exon8	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGATTCCTCACTGATTGCTCmUT

Nucleotide sequences for multiplexed primers used in plexes 1 - 3.

* Consecutive primers constitute forward (F) and reverse (R) primer pairs for the indicated loci

† Forward primers format: 5’ ACACTCTTTCCCTACACGACGCTCTTCCGATCT-[gene-specific forward] 3’,
Reverse primerformat:5’ TCGGCATTCCTGCTGAACCGCTCTTCCGATCT-[gene-specific reverse] 3’

‡ "m" denotes a single 2'-O-Methyl base in place of the DNA base, used in order to minimise potential primer dimers

Appendix 6: Custom pan-haematological cancer gene panel

ARID1A	*CREBBP*	*HIST1H1D*	*NOTCH2*	*SOCS1*
ASXL1	*CSF1R*	*HIST1H1E*	*NPM1*	*SRSF2*
ATM	*CSF3R*	*IDH1*	*NRAS*	*STAG2*
ATP6AP1	*CUX1*	*IDH2*	*PAX5*	*STAT3*
ATP6V1B2	*DNMT3A*	*IKZF3*	*PDGFRA*	*STAT6*
B2M	*EBF1*	*IL7R*	*PHF6*	*TCF3*
BCL10	*EP300*	*IRF8*	*PIM1*	*TET2*
BCL2	*ETNK1*	*JAK2*	*POT1*	*TNFAIP3*
BCL6	*ETV6*	*KDM6A*	*POU2F2*	*TNFRSF14*
BCOR	*EZH2*	*KIT*	*PPM1D*	*TP53*
BCORL1	*FBXW7*	*KMT2C*	*PRDM1*	*U2AF1*
BRAF	*FLT3*	*KMT2D*	*PTEN*	*WT1*
CALR	*FOXO1*	*KRAS*	*PTPN11*	*XPO1*
CARD11	*GATA2*	*MBD1*	*RAD21*	*ZEB1*
CBL	*GNA13*	*MEF2B*	*RRAGC*	*ZRSR2*
CCND3	*GNAS*	*MPL*	*RUNX1*	
CD58	*GNB1*	*MYC*	*SETBP1*	
CD79B	*H3F3A*	*MYD88*	*SETD2*	
CDKN2A	*HIST1H1B*	*NF1*	*SF3B1*	
CEBPA	*HIST1H1C*	*NOTCH1*	*SMC3*	

Appendix 7

Code for the derivation of the genetic AML prediction model

Discriminating evolution of acute myeloid leukaemia from age-related clonal haematopoiesis

Grace Collord & Moritz Gerstung

Tue Jul 24 16:38:48 2018

- 1 Preliminaries
  - 1.1 Libraries
- 2 AML incidence data
- 3 Discovery cohort
  - 3.1 Data
- 4 Validation cohort
  - 4.1 Data
- 5 Expected AML incidence
  - 5.1 Validation cohort
  - 5.2 Discovery cohort
- 6 Combined data
- 7 Coxph model fits
  - 7.1 Discovery cohort
    - 7.1.1 Non-adjusted
    - 7.1.2 Adjusted
  - 7.2 Validation cohort
    - 7.2.1 Non-adjusted
    - 7.2.2 Adjusted
  - 7.3 Cross-validation
    - 7.3.1 Non-adjusted
    - 7.3.2 Adjusted
  - 7.4 Combined
    - 7.4.1 Non-adjusted
    - 7.4.2 Adjusted
    - 7.4.3 Bootstrap
    - 7.4.4 Forest plot
    - 7.4.5 Dichotomous variables
    - 7.4.6 Bootstrap adjustment
    - 7.4.7 LOOCV
      - 7.4.7.1 Individual Predictions (non-adjusted)
      - 7.4.7.2 Jackknife variance
    - 7.4.8 Multiple bootstraps
    - 7.4.9 Individual Predictions with corrected baseline
    - 7.4.10 Some simulations
    - 7.4.11 Simple models
      - 7.4.11.1 Presence of any mutation
      - 7.4.11.2 Number of mutations + vaf
      - 7.4.11.3 Number of mutations + cumulative vaf
- 8 Logistic regression
  - 8.1 Combined
  - 8.2 Discovery cohort
  - 8.3 Validation cohort
  - 8.4 Bootstrap CIs
  - 8.5 Forest plot
  - 8.6 AUC
- 9 Tabulate results
- 10 Clinical/Demographic model
  - 10.1 Validation cohort
  - 10.2 Expected AML incidence
  - 10.3 Combined data
  - 10.4 Coxph model fits
    - 10.4.1 Discovery cohort
      - 10.4.1.1 Raw
    - 10.4.2 Validation cohort
      - 10.4.2.1 Raw
      - 10.4.2.2 Adjusted
- 11 Model excluding controls without mutations
  - 11.1 Validation cohort
  - 11.2 Expected AML incidence
  - 11.3 Combined data
  - 11.4 Coxph model fits
    - 11.4.1 DC
      - 11.4.1.1 Raw
      - 11.4.1.2 Adjusted

        - 11.4.2 Validation cohort
            - 11.4.2.1 Raw
            - 11.4.2.2 Adjusted
- 12 CoxPH model excluding all samples without ARCH-PD
    - 12.1 Discovery cohort
    - 12.2 Validation cohort
    - 12.3 Expected AML incidence
    - 12.4 Combined data
    - 12.5 Coxph model fits
        - 12.5.1 Toronto
            - 12.5.1.1 Raw
            - 12.5.1.2 Adjusted
        - 12.5.2 Validation cohort
            - 12.5.2.1 Raw
            - 12.5.2.2 Adjusted
- 13 Session

1 Preliminaries

1.1 Libraries

```
library(CoxHD)
library(survAUC)
library(survivalROC)
library(glmnet)
library(RColorBrewer)
library(stringr)
library(dplyr)
library(readr)

set1 <- RColorBrewer::brewer.pal(8, "Set1")
```

Helper functions

```
superSet <- function(x, s, fill=NA){
    i <- intersect(colnames(x), s)
    n <- setdiff(s, colnames(x))
    y <- x[,i]
    if(length(n) > 0)
        y <- cbind(y,  matrix(fill, ncol=length(n), dimnames=list(NULL, n)) )[,s]
    return(y)
}
```

2 AML incidence data

Use known AML incidence to correct bias using weighted controls. The expected incidence of AML was calculated from the UK office of national statistics, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence (http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence). Spline function to interpolate Male denoted by 1 and female by 0

```
age_incidence <- read.table("data/aml_age_incidence.txt", header=TRUE, sep="\t")
head(age_incidence)
```

	Age.Range <fctr>	Male.Cases <int>	Female.Cases <int>	Male.Rates <dbl>	Female.Rates <dbl>
1	0 to 04	18	12	0.9	0.6
2	05 to 09	10	10	0.5	0.5
3	10 to 14	8	10	0.4	0.6
4	15 to 19	15	14	0.7	0.8
5	20 to 24	21	18	1.0	0.8
6	25 to 29	22	20	1.0	0.9

6 rows

```
tail(age_incidence)
```

	Age.Range <fctr>	Male.Cases <int>	Female.Cases <int>	Male.Rates <dbl>	Female.Rates <dbl>
14	65 to 69	205	140	12.2	7.9

15	70 to 74	256	162	21.2	12.0
16	75 to 79	270	179	28.3	15.7
17	80 to 84	235	165	36.1	18.4
18	85 to 89	139	122	40.4	20.7
19	90+	53	85	35.6	22.2

6 rows

```
str(age_incidence)
```

```
## 'data.frame':    19 obs. of  5 variables:
##  $ Age.Range   : Factor w/ 19 levels "0 to 04","05 to 09",..: 1 2 3 4 5 6 7 8 9
10 ...
##  $ Male.Cases  : int  18 10 8 15 21 22 21 34 39 51 ...
##  $ Female.Cases: int  12 10 10 14 18 20 20 23 39 53 ...
##  $ Male.Rates  : num  0.9 0.5 0.4 0.7 1 1 1 1.7 1.8 2.2 ...
##  $ Female.Rates: num  0.6 0.5 0.6 0.8 0.8 0.9 0.9 1.2 1.7 2.2 ...
```

```
aml_inc <- function(gender, x){
    if(gender==1)
        splinefun(x=c(seq(0,90,5)), y=c(cumsum(age_incidence$Male.Rates/100000)*5)
, method="mono")(x)
    else
        splinefun(x=c(seq(0,90,5)), y=c(cumsum(age_incidence$Female.Rates/100000)*
5), method="mono")(x)
}
```

All cause mortality from the office of national statistics (https://www.ons.gov.uk/ (https://www.ons.gov.uk/)).

```
all_cause_mortality <- read.table("data/all_cause_mortality.txt", sep="\t", skip=1
, header=TRUE)
head(all_cause_mortality)
```

	x <int>	mx <dbl>	qx <dbl>	lx <dbl>	dx <dbl>	ex <dbl>	X <lgl>	mx.1 <dbl>	qx.1 <dbl>
1	0	0.004234	0.004225	100000.0	422.5	79.17	NA	0.003521	0.003515
2	1	0.000306	0.000306	99577.5	30.5	78.51	NA	0.000246	0.000246
3	2	0.000163	0.000163	99547.1	16.2	77.53	NA	0.000137	0.000137
4	3	0.000127	0.000127	99530.8	12.6	76.54	NA	0.000105	0.000105
5	4	0.000090	0.000090	99518.2	8.9	75.55	NA	0.000081	0.000081
6	5	0.000092	0.000092	99509.3	9.2	74.56	NA	0.000067	0.000067

6 rows | 1-10 of 13 columns

```
all_surv <- function(gender, age1, age2){
    if(gender==1)
        s <- all_cause_mortality$lx
    else
        s <- all_cause_mortality$lx.1
    f <- function(x) exp(splinefun(all_cause_mortality$x, log(s), method="mono")(x
))
    f(age2) / f(age1)
}
```

Function combining both

```
aml_inc_cr <- Vectorize(function(gender, age1, age2) sum(diff(aml_inc(gender, seq(
age1,age2,1) ))*all_surv(gender, age1, seq(age1,age2-1,1)) ), c("gender","age1","a
ge2"))
```

3 Discovery cohort

3.1 Data

4 (of 95) cases that were sampled within 6 months of AML diagnosis are excluded to avoid skewing model towards significance

```
f = "data/DC_vaf_matrix_414ctrl_91aml.csv"
```

```
torontoData <- read.csv(f)
torontoData$gender <- ifelse(torontoData$Sex == "male", 1, 0)
torontoData$gender <- as.numeric(torontoData$gender)
colnames(torontoData)
```

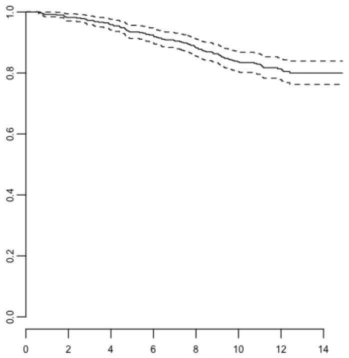
```
##  [1] "Sample"     "ASXL1"      "BCOR"       "CALR"       "CBL"        "DNMT3A"
"IDH1"       "IDH2"
##  [9] "JAK2"       "KDM6A"      "KIT"        "KMT2C"      "KRAS"       "NF1"
"NRAS"       "PHF6"
## [17] "PTPN11"     "RUNX1"      "SF3B1"      "SRSF2"      "TET2"       "TP53"
"U2AF1"      "Diagnosis"
## [25] "fu_years"   "age"        "Sex"        "no_drivers" "gender"
```

Manually standardize

```
torontoData <- torontoData[!duplicated(torontoData),]

gene_vars <- c("CALR", "NRAS", "DNMT3A", "SF3B1", "IDH1", "KIT", "TET2", "RAD21",
"JAK2", "CBL", "KRAS", "PTPN11", "IDH2", "TP53", "NF1", "SRSF2", "CEBPA", "ASXL1",
"RUNX1", "U2AF1", "BCOR", "KDM6A", "PHF6", "KMT2C", "KMT2D")

torontoX <- torontoData[, colnames(torontoData) %in% c(gene_vars, "age", "gender")
]

torontoX <- as.data.frame(torontoX)
```

Only include genes in model if mutated in >2 samples

```
thr <- 2
torontoX <- torontoX[,colSums(torontoX != 0)>=thr]

torontoGroups <- factor(names(torontoX) %in% c("age","gender")+1, level=1:2, label
s=c("Genes","Demographics"))

torontoX$age <- torontoX$age/10
names(torontoX)[which(names(torontoX)=="age")] <- "age_10"
g <- torontoGroups == "Genes"
torontoX[,g] <- torontoX[,g]*10
names(torontoX)[g] <- paste(names(torontoX)[g], "0.1",sep="_")

torontoSurv <- Surv(time = torontoData$fu_years, event = torontoData$Diagnosis=="A
ML")
plot(survfit(torontoSurv~ 1))
```



4 Validation cohort

4.1 Data

```
f = "data/VC_vaf_matrix_no_duplicates_262ctrl_29aml_nodates.csv"
sangerData <- read.csv(f)
colnames(sangerData)
```

```
##  [1] "X"          "Sample"     "ASXL1"      "BCOR"       "CBL"        "CEBPA"
"DNMT3A"     "IDH1"
##  [9] "IDH2"       "JAK2"       "KMT2C"      "KMT2D"      "KRAS"       "NF1"
```

```
"NRAS"       "PTPN11"
## [17] "RAD21"      "SF3B1"      "SRSF2"      "TET2"       "TP53"       "U2AF1"
"Individual" "hcdate"
## [25] "Diagnosis"  "age"        "gender"     "systol"     "diastol"    "bmi"
"cholestl"   "triglyc"
## [33] "hdl"        "ldl"        "lym"        "mcv"        "rdw"        "wbc"
"rbc"        "hct"
## [41] "plt"        "hgb"        "dodx"
```

```
head(sangerData[, c("Sample", "gender")]) #male=1, female=0
```

	Sample <fctr>	gender <int>
1	PD29762b	0
2	PD29764b	0
3	PD29792b	0
4	PD29804c	0
5	PD29810c	1
6	PD29836c	0

6 rows

NB all dates are jittered

```
sangerData$hcdate <- as.Date(sangerData$hcdate)
sangerData$dodx <- as.Date(sangerData$dodx)

sangerPatients <- sub("[a-z]+$","", sangerData$Sample)
o <- order(sangerPatients, as.numeric(sangerData$hcdate))

sangerData <- sangerData[o,]
sangerPatients <- sangerPatients[o]

clinical_vars <- c("systol", "diastol", "bmi", "cholestl", "triglyc", "hdl", "ldl"
, "lym", "mcv", "rdw", "wbc", "plt", "hgb")
sangerX <- sangerData[, colnames(sangerData) %in% c(gene_vars, "age","gender",clin
ical_vars)]
sangerX <- as.data.frame(sangerX)

sangerX <- sangerX[,colSums(sangerX != 0,na.rm=TRUE)>=thr]
sangerGroups <- factor(grepl("^[a-z]", colnames(sangerX))*2, levels=0:2, labels=c(
"Genes", "Demographics", "Blood"))
sangerGroups[names(sangerX) %in% c("age","gender")] <- "Demographics"
table(sangerGroups)
```

```
## sangerGroups
##        Genes Demographics        Blood
##           15            2           13
```

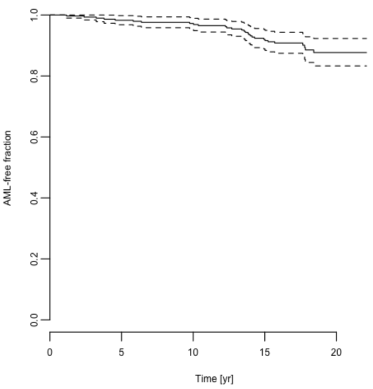
```
g <- sangerGroups=="Genes"
sangerX[g] <- sangerX[g] * 10
names(sangerX)[g] <- paste(names(sangerX[g]),"0.1", sep="_")
y <- StandardizeMagnitude(sangerX[!g])
sangerX <- cbind(sangerX[g],y)

poorMansImpute <- function(x) {x[is.na(x)] <- mean(x, na.rm=TRUE); return(x)}
sangerX <- as.data.frame(sapply(sangerX, poorMansImpute))

foo <- split(sangerData[,c("Diagnosis","hcdate","dodx")], sangerPatients)
```

```
bar <- do.call("rbind",lapply(foo, function(x){
                    y <- x
                    n <- nrow(y)
                    y[-n,"Diagnosis"] <- "Control"
                    start <- as.numeric(y$hcdate - y$hcdate[1])/365.25
                    end <- c(as.numeric(y$hcdate - y$hcdate[1])[-1]/365.25, as.num
eric(y$dodx[n] - y$hcdate[1])/365.25)
                    return(data.frame(Diagnosis=y[,"Diagnosis"], start=start, end=
end))
                }))

bar[1:6, ]
```

	Diagnosis <fctr>	start <dbl>	end <dbl>
PD29762	AML	0	9.754962
PD29764	AML	0	10.360027
PD29792	AML	0	14.108145
PD29804	Control	0	5.138946
PD29810	Control	0	18.573580
PD29836.1	Control	0	2.414784

6 rows

```
sangerPatientsSplit <- unlist(sapply(names(foo), function(n) rep(n, nrow(foo[[n]])
)))

sangerSurv <- Surv(time = bar$start, time2 = bar$end, event = bar$Diagnosis!="Cont
rol", origin = 0)
plot(survfit(sangerSurv ~ 1), ylab="AML-free fraction", xlab="Time [yr]")
```



5 Expected AML incidence

5.1 Validation cohort

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
head(sangerSurv[w,])
```

```
## [1] (0.000000, 9.754962]  (0.000000,10.360027]  (0.000000,14.108145]  (0.000000
, 5.138946+] (0.000000,18.573580+]
## [6] (2.414784,10.023272]
```

```
sangerSurv2 <- Surv(sangerSurv[w,2], sangerSurv[w,3])

expected_rate_sanger_cr <- mean(aml_inc_cr(sangerX[w,"gender"],sangerX[w,"age_10"]
*10, sangerX[w,"age_10"]*10+ pmax(1,sangerSurv2[,1]))[!sangerSurv2[,2]])

n_total_sanger <- sum(sangerSurv2[,2])/expected_rate_sanger_cr
n_total_sanger
```

```
## [1] 10406.64
```

5.2 Discovery cohort

```
expected_rate_toronto_cr <- mean(aml_inc_cr(torontoX[,"gender"],torontoX[,"age_10"
]*10, torontoX[,"age_10"]*10+ pmax(1,torontoSurv[,1]))[!torontoSurv[,2]])

n_total_toronto <- sum(torontoSurv[,2])/expected_rate_toronto_cr
n_total_toronto
```

```
## [1] 72377.73
```

6 Combined data

Survival

```
allSurv <- rbind(sangerSurv, Surv(rep(0, nrow(torontoSurv)), torontoSurv[,1], toro
ntoSurv[,2]))
allSurv <- Surv(allSurv[,1], allSurv[,2], allSurv[,3])
```

Data matrix

```
cohort <- c(rep("Sanger", nrow(sangerX)), rep("Toronto", nrow(torontoX)))
i <- c(sort(setdiff(gene_vars,"CALR")),"age","gender")
allX <- rbind(superSet(sangerData,i,fill=0), superSet(torontoData,i,fill=0))
colnames(allX)
```

```
##  [1] "ASXL1"  "BCOR"   "CBL"    "CEBPA"  "DNMT3A" "IDH1"   "IDH2"   "JAK2"   "K
DM6A"  "KIT"    "KMT2C"  "KMT2D"
## [13] "KRAS"   "NF1"    "NRAS"   "PHF6"   "PTPN11" "RAD21"  "RUNX1"  "SF3B1"  "S
RSF2"  "TET2"   "TP53"   "U2AF1"
## [25] "age"    "gender"
```

```
allX <- allX[,colSums(allX>0)>=thr]
allX <- cbind(allX, cohort=cohort=="Sanger") + 0
allGroups <- factor(grepl("^[A-Z]",colnames(allX))+0, levels=1:0, labels=c("Genes"
,"Demographics"))

g <- allGroups=="Genes"
allX <- cbind(10*allX[,g], StandardizeMagnitude(allX[,!g]))
colnames(allX)[g] <- paste(colnames(allX)[g],"0.1",sep="_")
control <- c(sangerData$Diagnosis=="Control", torontoData$Diagnosis=="Control")
```

Weights

```
weights <- rep(1, nrow(allX))
weights[cohort=="Sanger" & control] <- n_total_sanger/sum(cohort=="Sanger" & contr
ol & allSurv[,1]==0)
weights[cohort=="Toronto" & control] <- n_total_toronto/sum(cohort=="Toronto" & co
ntrol)

n_total <- n_total_sanger + n_total_toronto
n_total
```

```
## [1] 82784.38
```

Kaplan-Meier analysis

```
X = allX
surv = allSurv
pal1 <- c("#C32B4A", "#3F76B4", "#57B2AB", "#5E4FA2", "#EB6046")

colnames(X)
```

```
##  [1] "ASXL1_0.1"  "BCOR_0.1"   "CBL_0.1"    "DNMT3A_0.1" "IDH1_0.1"   "IDH2_0.1
"   "JAK2_0.1"   "KDM6A_0.1"
##  [9] "KMT2C_0.1"  "KMT2D_0.1"  "KRAS_0.1"   "NF1_0.1"    "NRAS_0.1"   "PHF6_0.1
"   "PTPN11_0.1" "RAD21_0.1"
## [17] "RUNX1_0.1"  "SF3B1_0.1"  "SRSF2_0.1"  "TET2_0.1"   "TP53_0.1"   "U2AF1_0.
1"  "age_10"     "gender"
## [25] "cohort"
```

```
names(X) <- str_replace(names(X), "[_]{1}[0-9]{1,}[\\.]{0,1}[0-9]{0,2}", "")
X$no_drivers <- rowSums((X[, colnames(X) %in% gene_vars]>0))
summary(X$no_drivers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.5263  1.0000  5.0000
```

```
X$max_vaf <- apply(X[, intersect(gene_vars, colnames(X))], 1, max, na.rm = TRUE)

genes <- c("DNMT3A", "TET2", "TP53", "U2AF1")

n_drivers <- cut(X$no_drivers, c( -1, 0, 1,  max(X$no_drivers)))
levels(n_drivers) <- c(0,1,"2+")

mvaf <- cut(X$max_vaf*10, c( -1, 0, 4, 8, max(X$max_vaf*10)))  #multiply by 10 to
reverse VAF standardisation
levels(mvaf) <- c("0", "0 - 4", "4 - 8", "8+")

par(mfrow=c(2,4), mar = c(1.8, 1.9, 1.7, 0.1) + 0.1, mgp=c(2.2,0.4,0), bty="L", xp
d=TRUE, las=1, tcl=-0.15, cex.axis=1, cex.lab = 1)
for (i in 1:length(genes)) {
  #i <- 1
  gene <- genes[i]
  plot(survfit(surv ~ X[[gene]] == 0), col= pal1, bty='L', yaxs='i', ylim=c(0,1.01
), mark.time = T, conf.int = F)
  mtext(gene, font=3, side = 3, line = 0.1, cex = 0.7)
  legend("bottomleft", col=pal1[1:2], lty=1, c("MT","WT"), lwd = 1.1, bty="n", nco
l = 1, cex = 0.9)
}
plot(survfit(surv ~ n_drivers), col=rev(pal1[1:3]), conf.int = F, mark.time = T, b
ty='L', yaxs='i', ylim=c(0,1.01))
mtext("Number of drivers", font=1, side = 3, line = 0.4, cex = 0.7)
legend("bottomleft", legend = levels(n_drivers), col= rev(pal1[1:3]), lty=1, lwd =
1.1, bty='n', title="", cex = 0.9)
plot(survfit(surv ~ mvaf), col= rev(pal1[1:4]), conf.int = F, mark.time = T, bty='
L', yaxs='i', ylim=c(0,1.01))
mtext("Maximum VAF (%)", font=1, side = 3, line = 0.4, cex = 0.7)
legend("bottomleft", levels(mvaf), col=rev(pal1[1:4]), lty=1, lwd = 1.1, bty='n',
title="", cex = 0.9)

genes <- intersect(colnames(X), gene_vars)
length(genes)
```

```
## [1] 22
```

```
png("./figures/CombinedCohorts.KM.curves.png", width = 35, height = 20, units = "c
m", res = 300)
par(mfrow=c(4,7), mar = c(3.7, 3.5, 1.6, 1) + 0.1, mgp=c(1.9,0.4,0), bty="L", xpd=
TRUE, las=1, tcl=-0.2, cex.axis=1, cex.lab = 1.2)
for (i in 1:length(genes)) {
  #i <- 1
  gene <- genes[i]
  plot(survfit(surv ~ X[[gene]] == 0), col= pal1, xlab='Time (years)', ylab = 'AML
-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.time = T, conf.int = F)
  mtext(gene, font=4, side = 3, cex = 0.9, line = 0.35)
}
plot.new(); par(xpd=NA)
legend(x = -0.5, y = 0.5, col=pal1[1:2], lty=1, c("Mutated","Wildtype"), cex=1.4,
lwd = 2, bty="n", ncol = 1)
dev.off()
```

```
## pdf
##   2
```

7 Coxph model fits

```
sigma0 <- 0.1
nu <- 1
which.mu <- "Genes"
```

7.1 Discovery cohort

7.1.1 Non-adjusted

```
fitToronto <- CoxRFX(torontoX, torontoSurv, groups=torontoGroups, which.mu=which.m
u, nu=nu, sigma0=sigma0)
waldToronto <- WaldTest(fitToronto)
```

```
##                    group    coef   coef-mu     sd      z df  p.value sig
## ASXL1_0.1          Genes  0.6715  3.40e-02 0.1169  5.745  1 9.19e-09 ***
## CALR_0.1           Genes  0.6168 -2.07e-02 0.0717  8.603  1 7.76e-18 ***
## CBL_0.1            Genes  0.5158 -1.22e-01 0.1311  3.935  1 8.30e-05 ***
## DNMT3A_0.1         Genes  0.5860 -5.15e-02 0.1017  5.761  1 8.36e-09 ***
## IDH1_0.1           Genes  0.6818  4.43e-02 0.1269  5.373  1 7.74e-08 ***
## IDH2_0.1           Genes  0.5153 -1.22e-01 0.1159  4.446  1 8.74e-06 ***
## JAK2_0.1           Genes  0.6967  5.92e-02 0.1249  5.580  1 2.40e-08 ***
## KDM6A_0.1          Genes  0.6375  2.36e-05 0.0581 10.982  1 4.67e-28 ***
## KMT2C_0.1          Genes  0.6602  2.27e-02 0.0618 10.689  1 1.14e-26 ***
## KRAS_0.1           Genes  0.6350 -2.46e-03 0.0581 10.932  1 8.12e-28 ***
## NF1_0.1            Genes  0.6359 -1.61e-03 0.0581 10.947  1 6.86e-28 ***
## PHF6_0.1           Genes  0.6429  5.40e-03 0.0586 10.978  1 4.87e-28 ***
## PTPN11_0.1         Genes  0.6546  1.71e-02 0.0583 11.224  1 3.11e-29 ***
## RUNX1_0.1          Genes  0.3926 -2.45e-01 0.0927  4.236  1 2.27e-05 ***
## SF3B1_0.1          Genes  0.7605  1.23e-01 0.1045  7.274  1 3.49e-13 ***
## SRSF2_0.1          Genes  0.4847 -1.53e-01 0.0944  5.134  1 2.83e-07 ***
## TET2_0.1           Genes  0.6127 -2.48e-02 0.1300  4.712  1 2.46e-06 ***
## TP53_0.1           Genes  0.8595  2.22e-01 0.0875  9.823  1 8.99e-23 ***
## U2AF1_0.1          Genes  0.8524  2.15e-01 0.0785 10.860  1 1.79e-27 ***
## age_10     Demographics -0.0387 -3.87e-02 0.0943 -0.410  1 6.82e-01
## gender     Demographics -0.0434 -4.34e-02 0.1069 -0.406  1 6.85e-01
```

```
survConcordance(fitToronto$surv ~ fitToronto$linear.predictors)
```

```
## Call:
## survConcordance(formula = fitToronto$surv ~ fitToronto$linear.predictors)
##
##   n= 505
## Concordance= 0.7426378 se= 0.03079247
## concordant discordant  tied.risk  tied.time   std(c-d)
##  28925.000  10024.000      0.000      1.000   2398.672
```

7.1.2 Adjusted

```
fitWeightedToronto <- CoxRFX(torontoX, torontoSurv, torontoGroups, which.mu=which.
mu, sigma0=sigma0, nu=nu, weights=weights[cohort=="Toronto"])
waldWeightedToronto <- WaldTest(fitWeightedToronto)
```

```
##                   group    coef coef-mu     sd      z df  p.value sig
## ASXL1_0.1         Genes  1.9481  0.0184 0.1452 13.415  1 4.92e-41 ***
## CALR_0.1          Genes  0.8664 -1.0633 0.7205  1.202  1 2.29e-01
## CBL_0.1           Genes  0.3846 -1.5451 0.3618  1.063  1 2.88e-01
## DNMT3A_0.1        Genes  0.7091 -1.2206 0.1236  5.736  1 9.70e-09 ***
## IDH1_0.1          Genes  2.3976  0.4679 0.3353  7.151  1 8.63e-13 ***
## IDH2_0.1          Genes  0.8112 -1.1185 0.2286  3.548  1 3.88e-04 ***
## JAK2_0.1          Genes  1.9253 -0.0044 0.1819 10.586  1 3.45e-26 ***
## KDM6A_0.1         Genes  1.9404  0.0107 0.1355 14.323  1 1.56e-46 ***
## KMT2C_0.1         Genes  2.4139  0.4841 0.6457  3.739  1 1.85e-04 ***
## KRAS_0.1          Genes  1.8253 -0.1044 0.1565 11.665  1 1.93e-31 ***
## NF1_0.1           Genes  1.8627 -0.0670 0.1522 12.238  1 1.94e-34 ***
## PHF6_0.1          Genes  2.1738  0.2441 0.1301 16.706  1 1.19e-62 ***
## PTPN11_0.1        Genes  2.5509  0.6212 0.2150 11.867  1 1.76e-32 ***
## RUNX1_0.1         Genes  0.7839 -1.1458 0.1361  5.761  1 8.38e-09 ***
## SF3B1_0.1         Genes  3.1354  1.2057 0.3087 10.156  1 3.11e-24 ***
## SRSF2_0.1         Genes  1.3985 -0.5312 0.1706  8.196  1 2.49e-16 ***
## TET2_0.1          Genes  0.6793 -1.2504 0.2014  3.373  1 7.43e-04 ***
## TP53_0.1          Genes  4.8882  2.9585 0.4224 11.572  1 5.69e-31 ***
## U2AF1_0.1         Genes  3.9699  2.0402 0.3601 11.024  1 2.94e-28 ***
## age_10     Demographics -0.0869 -0.0869 0.0996 -0.872  1 3.83e-01
## gender     Demographics -0.0443 -0.0443 0.1112 -0.399  1 6.90e-01
```

```
survConcordance(fitWeightedToronto$surv ~ fitWeightedToronto$linear.predictors, we
ights=weights[cohort=="Toronto"])
```

```
## Call:
## survConcordance(formula = fitWeightedToronto$surv ~ fitWeightedToronto$linear.p
redictors,
##     weights = weights[cohort == "Toronto"])
##
##   n= 505
## Concordance= 0.7739557 se= 0.03055735
## concordant discordant  tied.risk  tied.time   std(c-d)
##  4719299.0  1378335.7        0.0        1.0   372655.1
```

Uno’s estimator of cumulative/dynamic AUC

```
a <- AUC.uno(torontoSurv, torontoSurv, fitWeightedToronto$linear.predictors, times
= seq(0,12, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.761
```

```
png("./figures/DC.adj.coxph.auc.uno.png", width = 9, height = 10, units = "cm", re
s = 800)
par(mar = c(3.2, 3.2, 4, 2) + 0.1, mgp=c(2,0.5,0), bty="L",  tcl =-0.2, las = 1, c
ex=1)
plot(a$times, a$auc, xlab="Time (years)", ylab="AUC", pch=16, col="grey80", ylim =
c(0,1.0))
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(a$iauc,2
)))
dev.off()
```

```
## pdf
##   2
```

Time-dependent ROC AUC

```
r <- survivalROC(Stime = torontoSurv[,1], status=torontoSurv[,2], marker=fitWeight
edToronto$linear.predictors-colMeans(fitWeightedToronto$Z) %*% fitWeightedToronto$
coefficients, predict.time = 10, method="NNE", span=0.001)
round(r$AUC, digits = 3)
```

```
## [1] 0.783
```

```
png("./figures/DC.adj.coxph.roct.png", width = 9, height = 10, units = "cm", res =
500)
par(mar = c(3.2, 3.2, 4, 2) + 0.1, mgp=c(2,0.5,0), bty="L",  tcl =-0.2, las = 1, c
ex = 1)
plot(r$FP, r$TP, type='s',
     xlab="False Positive Rate", ylab="True Positive Rate",
     col = "black")
abline(a = 0, b = 1, col = "grey70", lty = 1, lwd = 1)
legend("bottomright", bty = "n", legend = paste("AUC = ",round(r$AUC,2)))
dev.off()
```

```
## pdf
##   2
```

7.2 Validation cohort

7.2.1 Non-adjusted

```
fitSanger <- CoxRFX(sangerX, sangerSurv, groups=sangerGroups, which.mu=which.mu, n
u=nu, sigma0=sigma0)
waldSanger <- WaldTest(fitSanger)
```

```
##                     group     coef   coef-mu      sd       z df  p.value sig
## ASXL1_0.1           Genes  0.76929  0.138331 0.11468  6.7084  1 1.97e-11 ***
## CBL_0.1             Genes  0.62044 -0.010519 0.09149  6.7814  1 1.19e-11 ***
## DNMT3A_0.1          Genes  0.51590 -0.115058 0.11678  4.4176  1 9.98e-06 ***
## JAK2_0.1            Genes  0.58502 -0.045941 0.10315  5.6716  1 1.42e-08 ***
## KMT2C_0.1           Genes  0.64589  0.014930 0.08616  7.4961  1 6.57e-14 ***
## KMT2D_0.1           Genes  0.50507 -0.125896 0.15209  3.3209  1 8.97e-04 ***
## KRAS_0.1            Genes  0.63604  0.005083 0.08495  7.4876  1 7.02e-14 ***
## NF1_0.1             Genes  0.62556 -0.005397 0.08610  7.2657  1 3.71e-13 ***
## NRAS_0.1            Genes  0.63025 -0.000712 0.08492  7.4214  1 1.16e-13 ***
## RAD21_0.1           Genes  0.62875 -0.002212 0.08524  7.3763  1 1.63e-13 ***
## SF3B1_0.1           Genes  0.62728 -0.003678 0.08572  7.3181  1 2.52e-13 ***
## SRSF2_0.1           Genes  0.58180 -0.049163 0.12680  4.5883  1 4.47e-06 ***
## TET2_0.1            Genes  0.69969  0.068723 0.11185  6.2555  1 3.96e-10 ***
## TP53_0.1            Genes  0.69326  0.062294 0.08559  8.0998  1 5.51e-16 ***
## U2AF1_0.1           Genes  0.70018  0.069214 0.08556  8.1832  1 2.76e-16 ***
## age_10       Demographics  0.10777  0.107774 0.12063  0.8934  1 3.72e-01
## gender       Demographics  0.00589  0.005894 0.10667  0.0553  1 9.56e-01
## systol_100          Blood  0.03002  0.030016 0.04429  0.6777  1 4.98e-01
## diastol_100         Blood  0.04718  0.047181 0.02863  1.6478  1 9.94e-02   .
## bmi_10              Blood  0.14183  0.141832 0.07973  1.7790  1 7.52e-02   .
## cholestl_10         Blood  0.00525  0.005246 0.01501  0.3496  1 7.27e-01
## triglyc             Blood  0.00450  0.004496 0.10599  0.0424  1 9.66e-01
## hdl                 Blood -0.09452 -0.094522 0.08059 -1.1729  1 2.41e-01
## ldl                 Blood  0.11424  0.114236 0.11019  1.0367  1 3.00e-01
## lym                 Blood  0.10961  0.109610 0.10081  1.0872  1 2.77e-01
## mcv_100             Blood -0.01645 -0.016447 0.00817 -2.0136  1 4.41e-02   *
## rdw_10              Blood  0.06116  0.061157 0.01972  3.1015  1 1.93e-03  **
## wbc_10              Blood  0.01499  0.014994 0.04138  0.3623  1 7.17e-01
## plt_100             Blood  0.06837  0.068369 0.09739  0.7020  1 4.83e-01
## hgb_10              Blood  0.04890  0.048900 0.02466  1.9826  1 4.74e-02   *
```

```
survConcordance(sangerSurv ~ fitSanger$linear.predictors)
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitSanger$linear.predictors)
##
##   n= 445
## Concordance= 0.793915 se= 0.05514512
## concordant discordant  tied.risk  tied.time   std(c-d)
##  5532.0000  1436.0000     0.0000     0.0000   768.5024
```

7.2.2 Adjusted

```
fitWeightedSanger <- CoxRFX(sangerX, sangerSurv, sangerGroups, which.mu=which.mu,
sigma0=sigma0, nu=nu, weights=weights[cohort=="Sanger"])
```

```
waldWeightedSanger <- waldTest(fitWeightedSanger)
```

```
##                      group     coef  coef-mu      sd       z df  p.value sig
## ASXL1_0.1            Genes  2.93589  0.95179 0.45155  6.5018  1 7.93e-11 ***
## CBL_0.1              Genes  0.89451 -1.08959 1.25454  0.7130  1 4.76e-01
## DNMT3A_0.1           Genes  0.80635 -1.17775 0.22686  3.5544  1 3.79e-04 ***
## JAK2_0.1             Genes -0.33650 -2.32060 0.95076 -0.3539  1 7.23e-01
## KMT2C_0.1            Genes  2.07422  0.09012 1.10633  1.8749  1 6.08e-02   .
## KMT2D_0.1            Genes  0.05067 -1.93343 0.81191  0.0624  1 9.50e-01
## KRAS_0.1             Genes  2.45194  0.46784 0.41069  5.9702  1 2.37e-09 ***
## NF1_0.1              Genes  1.54402 -0.44008 0.90581  1.7046  1 8.83e-02   .
## NRAS_0.1             Genes  1.92976 -0.05434 0.37569  5.1366  1 2.80e-07 ***
## RAD21_0.1            Genes  1.75445 -0.22966 0.66215  2.6496  1 8.06e-03  **
## SF3B1_0.1            Genes  1.56640 -0.41770 0.99531  1.5738  1 1.16e-01
## SRSF2_0.1            Genes  1.51230 -0.47181 0.27893  5.4217  1 5.90e-08 ***
## TET2_0.1             Genes  1.31638 -0.66772 0.13659  9.6374  1 5.56e-22 ***
## TP53_0.1             Genes  4.92658  2.94248 0.92037  5.3528  1 8.66e-08 ***
## U2AF1_0.1            Genes  6.33456  4.35046 0.76145  8.3191  1 8.86e-17 ***
## age_10        Demographics  0.03788  0.03788 0.11866  0.3193  1 7.50e-01
## gender        Demographics -0.01411 -0.01411 0.10079 -0.1400  1 8.89e-01
## systol_100           Blood  0.01712  0.01712 0.04486  0.3816  1 7.03e-01
## diastol_100          Blood  0.03900  0.03900 0.02964  1.3156  1 1.88e-01
## bmi_10               Blood  0.15297  0.15297 0.08406  1.8198  1 6.88e-02   .
## cholestl_10          Blood  0.00238  0.00238 0.01544  0.1542  1 8.77e-01
## triglyc              Blood -0.03451 -0.03451 0.11758 -0.2935  1 7.69e-01
## hdl                  Blood -0.12128 -0.12128 0.08447 -1.4357  1 1.51e-01
## ldl                  Blood  0.13215  0.13215 0.11436  1.1555  1 2.48e-01
## lym                  Blood  0.07976  0.07976 0.10326  0.7724  1 4.40e-01
## mcv_100              Blood -0.02401 -0.02401 0.00786 -3.0529  1 2.27e-03  **
## rdw_10               Blood  0.06721  0.06721 0.01666  4.0355  1 5.45e-05 ***
## wbc_10               Blood  0.00757  0.00757 0.04834  0.1567  1 8.76e-01
## plt_100              Blood  0.08415  0.08415 0.09986  0.8427  1 3.99e-01
## hgb_10               Blood  0.03718  0.03718 0.02437  1.5255  1 1.27e-01
```

```
waldWeightedSanger$p.adj <- p.adjust(p=waldWeightedSanger$p.value, method = "bonfe
rroni")
#View(waldWeightedSanger)

survConcordance(sangerSurv ~ fitWeightedSanger$linear.predictors, weights=weights[
cohort=="Sanger"])
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitWeightedSanger$linear.predictors,
##     weights = weights[cohort == "Sanger"])
##
##   n= 445
## Concordance= 0.8351691 se= 0.05475847
## concordant discordant  tied.risk  tied.time   std(c-d)
##  218019.86   43028.90       0.00       0.00   28589.26
```

Uno's estimator of cumulative/dynamic AUC

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))  #get right censored surv
ival data for each individual
s <- Surv(sangerSurv[w,2], sangerSurv[w,3])  ##Adjust according to dimensions of s
urvival object
a <- AUC.uno(s, s, fitWeightedSanger$linear.predictors[w], times= seq(0, 22, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.82
```

```
png("./figures/VC.ajd.coxph.auc.uno.png", width = 9, height = 10, units = "cm", re
s = 500)
par(mar = c(3.2, 3.2, 4, 2) + 0.1, mgp=c(2,0.5,0), bty="L",  tcl =-0.2, las = 1, c
ex=1)
plot(a$times, a$auc, xlab="Time (years)", ylab="AUC", pch=16, col="grey80", ylim =
c(0,1.0))
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", legend = paste("AUC = ",round(a$iauc,2)))
dev.off()
```

```
## pdf
##   2
```

Time-dependent ROC AUC

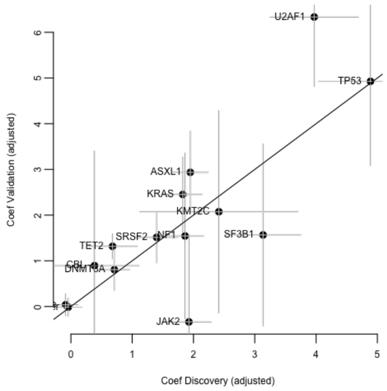
```
r <- survivalROC(Stime = s[,1], status=s[,2], marker=fitWeightedSanger$linear.pred
ictors[w]-colMeans(fitWeightedSanger$Z[w,]) %*% fitWeightedSanger$coefficients, pr
edict.time = 10, method="NNE", span=0.001)
round(r$AUC, digits = 3)
```

```
## [1] 0.737
```

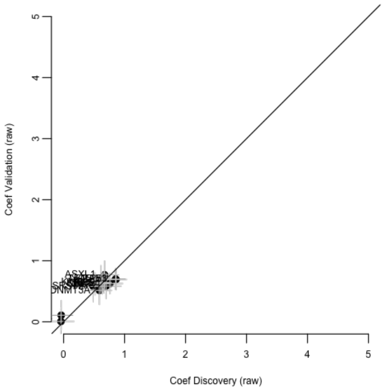
```
png("./figures/VC.ajd.coxph.roct.png", width = 9, height = 10, units = "cm", res =
500)
par(mar = c(3.2, 3.2, 4, 2) + 0.1, mgp=c(2,0.5,0), bty="L",  tcl =-0.2, las = 1, c
ex = 1)
plot(r$FP, r$TP, type='s',
     xlab="False Positive Rate", ylab="True Positive Rate",
     col = "black")
abline(a = 0, b = 1, col = "grey70", lty = 1, lwd = 1)
legend("bottomright", bty = "n", legend = paste("AUC = ",round(r$AUC,2)))
dev.off()
```

```
## pdf
##   2
```

```
i <- intersect(rownames(waldWeightedSanger), rownames(waldWeightedToronto))
plot( waldWeightedToronto[i,"coef"], waldWeightedSanger[i, "coef"], xlab="Coef Dis
covery (adjusted)", ylab="Coef Validation (adjusted)", pch=19, cex=1)
segments(waldWeightedToronto[i,"coef"]  - 2*waldWeightedToronto[i,"sd"], waldWeigh
tedSanger[i, "coef"], waldWeightedToronto[i,"coef"]  + 2*waldWeightedToronto[i,"sd
"], waldWeightedSanger[i, "coef"], col="grey" )
segments(waldWeightedToronto[i,"coef"]  , waldWeightedSanger[i, "coef"]-  2*waldWe
ightedSanger[i,"sd"], waldWeightedToronto[i,"coef"] , waldWeightedSanger[i, "coef"
] +2*waldWeightedSanger[i,"sd"], col="grey")
text(labels=sub("_.+","", i), waldWeightedToronto[i,"coef"], waldWeightedSanger[i,
"coef"], pos=2, adj=c(0,1))
abline(0,1)
```



```
plot( waldToronto[i,"coef"], waldSanger[i, "coef"], xlab="Coef Discovery (raw)", y
lab="Coef Validation (raw)", pch=19, cex=1, ylim=c(0,5),xlim=c(0,5))
segments(waldToronto[i,"coef"]  - 2*waldToronto[i,"sd"], waldSanger[i, "coef"], wa
ldToronto[i,"coef"]  + 2*waldToronto[i,"sd"], waldSanger[i, "coef"], col="grey" )
segments(waldToronto[i,"coef"]  , waldSanger[i, "coef"]-  2*waldSanger[i,"sd"], wa
ldToronto[i,"coef"] , waldSanger[i, "coef"] +2*waldSanger[i,"sd"], col="grey")
text(labels=sub("_.+","", i), waldToronto[i,"coef"], waldSanger[i, "coef"], pos=2,
adj=c(0,1))
abline(0,1)
```



7.3 Cross-validation

7.3.1 Non-adjusted

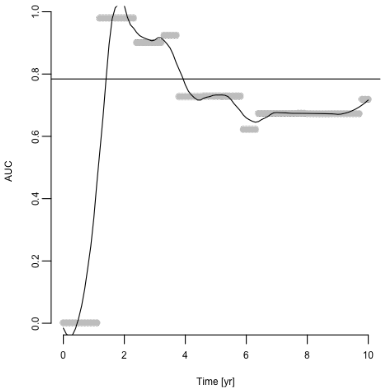
```
sangerImp <- torontoX[1:nrow(sangerX),]
sangerImp[,] <- NA
i <- intersect(names(sangerX),colnames(torontoX))
sangerImp[,i] <- sangerX[,i]
j <- setdiff(names(torontoX)[torontoGroups=="Genes"], names(sangerX))
sangerImp[,j] <- 0
```

DC fit, VC data

```
pS <- PredictRiskMissing(fitToronto, sangerImp)
survConcordance(sangerSurv ~ pS[,1])
```

```
## Call:
## survConcordance(formula = sangerSurv ~ pS[, 1])
##
##   n= 445
## Concordance= 0.7963548 se= 0.05514445
## concordant discordant  tied.risk  tied.time   std(c-d)
##   5545.000   1415.000      8.000      0.000    768.493
```

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
s <- Surv(sangerSurv[w,2], sangerSurv[w,3])
t <- seq(0,10,0.1)
a <- AUC.uno(torontoSurv, s, pS[w,1], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```



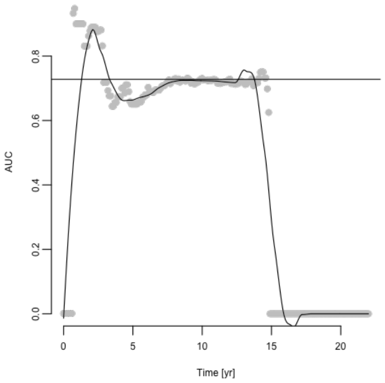
```
torontoImp <- sangerX[1:nrow(torontoX),]
torontoImp[,] <- NA
i <- intersect(names(sangerX),colnames(torontoX))
torontoImp[,i] <- torontoX[,i]
j <- setdiff(names(sangerX)[sangerGroups=="Genes"], names(torontoX))
torontoImp[,j] <- 0
```

VC fit, DC data

```
pT <- PredictRiskMissing(fitSanger, torontoImp)
survConcordance(torontoSurv ~ pT[,1])
```

```
## Call:
## survConcordance(formula = torontoSurv ~ pT[, 1])
##
##   n= 505
## Concordance= 0.6992477 se= 0.03079247
## concordant discordant  tied.risk  tied.time   std(c-d)
##  27235.000  11714.000      0.000      1.000   2398.672
```

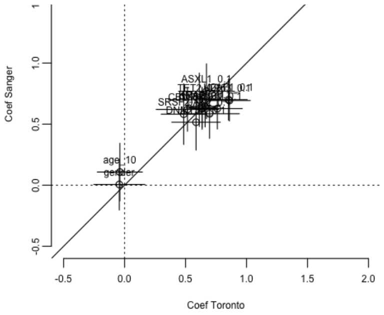
```
t <- seq(0,22,0.1)
a <- AUC.uno(s, torontoSurv, pT[,1], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```



```
sangerM <- sangerX
sangerM[,sangerGroups=="Blood"] <- NA
p <- PredictRiskMissing(fitSanger, sangerM)
survConcordance(sangerSurv ~ p[,1])
```

```
## Call:
## survConcordance(formula = sangerSurv ~ p[, 1])
##
##   n= 445
## Concordance= 0.8069747 se= 0.05514449
## concordant discordant  tied.risk  tied.time   std(c-d)
##  5619.0000  1341.0000     8.0000     0.0000   768.4936
```

```
plot(waldToronto[i,"coef"], waldSanger[i,"coef"], xlab="Coef Toronto", ylab="Coef
Sanger", xlim=c(-0.5,2), ylim=c(-0.5,2))
text(labels=i,waldToronto[i,"coef"], waldSanger[i,"coef"], pos=3)
segments(x0=waldToronto[i,"coef"], x1=waldToronto[i,"coef"], y0= waldSanger[i,"coe
f"]-1.96*waldSanger[i,"sd"], y1=waldSanger[i,"coef"]+1.96*waldSanger[i,"sd"])
segments(x0=waldToronto[i,"coef"]-1.96*waldToronto[i,"sd"], x1=waldToronto[i,"coef
"]+1.96*waldToronto[i,"sd"], y0= waldSanger[i,"coef"], y1=waldSanger[i,"coef"])
abline(0,1)
abline(h=0, lty=3)
abline(v=0, lty=3)
```



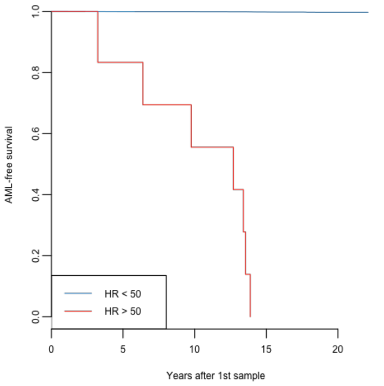
7.3.2 Adjusted

DC fit, VC data

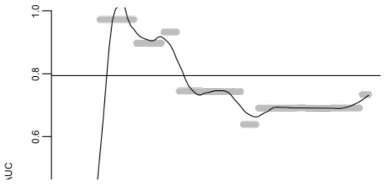
```
pS <- PredictRiskMissing(fitWeightedToronto, sangerImp)
survConcordance(sangerSurv ~ pS[,1], weights=weights[cohort=="Sanger"])
```

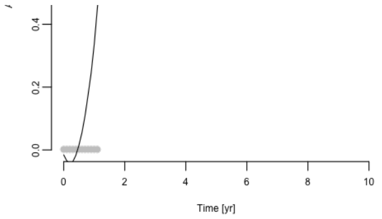
```
## Call:
## survConcordance(formula = sangerSurv ~ pS[, 1], weights = weights[cohort ==
##     "Sanger"])
##
##   n= 445
## Concordance= 0.821456 se= 0.05475772
##  concordant  discordant   tied.risk   tied.time    std(c-d)
## 214281.1753  46449.8206    317.7601      0.0000  28588.8682
```

```
m <- as.numeric(colSums(fitWeightedToronto$Z * weights[cohort=="Toronto"])/sum(wei
ghts[cohort=="Toronto"])) %*% coef(fitWeightedToronto)
plot(survfit(sangerSurv ~ exp(pS[,1]-as.numeric(m))>50, weights=weights[cohort=="S
anger"]), col=set1[2:1], ylab="AML-free survival", xlab='Years after 1st sample')
legend("bottomleft", c("HR < 50", "HR > 50"), lty=1, col=set1[2:1])
```



```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
s <- Surv(sangerSurv[w,2], sangerSurv[w,3])
t <- seq(0,10,0.1)
a <- AUC.uno(torontoSurv, s, pS[w,1], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```





```
png("./figures/DCfit.VCdata.adj.coxph.auc.uno.png", width = 14, height = 14, units
= "cm", res = 500)
par(mar = c(4, 4, 4, 2) + 0.1, mgp=c(2.7,0.7,0), bty="L",  tcl =-0.2, las = 1, cex
.lab = 1.1)
plot(a$times, a$auc, xlab="Time (years)", ylab="AUC", pch=16, col="grey80", ylim =
c(0,1.0))
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc, lty = 3, lwd = 1)
mtext("DC fit, VC data", font= 2, side = 3, cex = 1, line = 0.5)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(a$iauc,2
)))
dev.off()
```

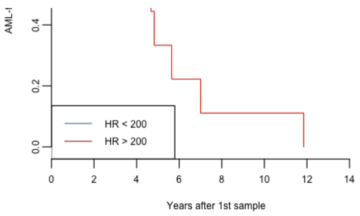
```
## pdf
##   2
```

VC fit, DC data

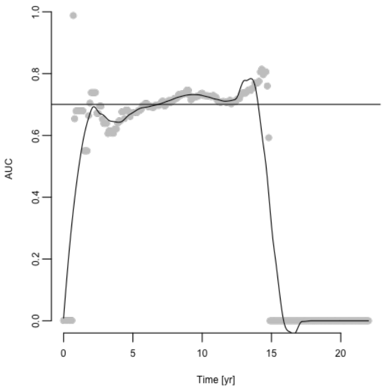
```
pT <- PredictRiskMissing(fitWeightedSanger, torontoImp)
survConcordance(torontoSurv ~ pT[,1], weights=weights[cohort=="Toronto"])
```

```
## Call:
## survConcordance(formula = torontoSurv ~ pT[, 1], weights = weights[cohort ==
##     "Toronto"])
##
##   n= 505
## Concordance= 0.7202544 se= 0.03055735
## concordant discordant  tied.risk  tied.time   std(c-d)
##  4391848.0  1705786.7        0.0        1.0   372655.1
```

```
m <- as.numeric(colSums(fitWeightedSanger$Z * weights[cohort=="Sanger"])/sum(weigh
ts[cohort=="Sanger"])) %*% coef(fitWeightedSanger)
plot(survfit(torontoSurv ~ exp(pT[,1]-as.numeric(m))>200, weights=weights[cohort==
"Toronto"]), col=set1[2:1], ylab="AML-free survival", xlab='Years after 1st sample
')
legend("bottomleft", c("HR < 200", "HR > 200"), lty=1, col=set1[2:1])
```



```
t <- seq(0,22,0.1)
a <- AUC.uno(s, torontoSurv, pT[,1], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```



```
png("./figures/VCfit.DCdata.adj.coxph.auc.uno.png", width = 14, height = 14, units
= "cm", res = 500)
par(mar = c(4, 4, 4, 2) + 0.1, mgp=c(2.7,0.7,0), bty="L",  tcl =-0.2, las = 1, cex
.lab = 1.1)
plot(a$times, a$auc, xlab="Time (years)", ylab="AUC", pch=16, col="grey80", ylim =
c(0,1.0))
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc, lty = 3, lwd = 1)
mtext("VC fit, DC data", font= 2, side = 3, cex = 1, line = 0.5)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(a$iauc,2
)))#dev.off()
dev.off()
```

```
## pdf
##   2
```

7.4 Combined

7.4.1 Non-adjusted

```
fitAll <- CoxRFX(allX, allSurv, allGroups, which.mu=which.mu, sigma0=sigma0, nu=nu
)
fitAll
```

```
## Means:
##               mean    sd  z   p.val sig
## Genes         0.79 0.068 12 3.9e-31 ***
## Demographics 0.00 0.000  0      NA
## 
## Variances - p-values only indicative:
##               sigma2 chisq  df   p.val sig
## Genes           0.19    25 9.2 2.7e-03  **
## Demographics   0.48    25 2.7 1.2e-05 ***
## 
## Partial log hazard:
##               Cov[g,g] Sum(Cov[,g])   MSE
## Genes             0.40         0.41 0.012
## Demographics     0.45         0.46 0.032
## TOTAL              NaN         0.88 0.044
```

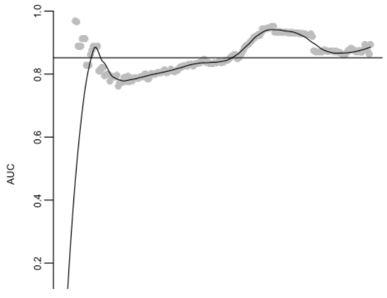
```
WaldTest(fitAll, uncentered=FALSE)
```

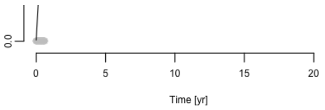
```
##                        group      coef coef-mu      sd       z df  p.value sig
## ASXL1_0.1              Genes -0.042129 -0.8326 0.12580 -0.3349  1 7.38e-01    
## BCOR_0.1               Genes  0.018602 -0.7719 0.00792  2.3484  1 1.89e-02   *
## CBL_0.1                Genes -0.313214 -1.1037 0.20346 -1.5394  1 1.24e-01    
## DNMT3A_0.1             Genes -0.233727 -1.0242 0.10840 -2.1561  1 3.11e-02   *
## IDH1_0.1               Genes  0.021937 -0.7685 0.20020  0.1096  1 9.13e-01    
## IDH2_0.1               Genes -0.278283 -1.0687 0.15309 -1.8177  1 6.91e-02   .
## JAK2_0.1               Genes -0.030573 -0.8210 0.14841 -0.2060  1 8.37e-01    
## KDM6A_0.1              Genes  0.000538 -0.7899 0.00638  0.0843  1 9.33e-01    
## KMT2C_0.1              Genes  0.068877 -0.7216 0.08598  0.8011  1 4.23e-01    
## KMT2D_0.1              Genes -0.391241 -1.1817 0.20457 -1.9125  1 5.58e-02   .
## KRAS_0.1               Genes  0.006235 -0.7842 0.01271  0.4907  1 6.24e-01    
## NF1_0.1                Genes -0.020208 -0.8107 0.03223 -0.6270  1 5.31e-01    
## NRAS_0.1               Genes  0.034555 -0.7559 0.01285  2.6887  1 7.17e-03  **
## PHF6_0.1               Genes  0.016466 -0.7740 0.01532  1.0749  1 2.82e-01    
## PTPN11_0.1             Genes  0.360022 -0.4304 0.20817  1.7295  1 8.37e-02   .
## RAD21_0.1              Genes -0.006662 -0.7971 0.01823 -0.3654  1 7.15e-01    
## RUNX1_0.1              Genes -0.399568 -1.1900 0.11410 -3.5019  1 4.62e-04 ***
## SF3B1_0.1              Genes  0.239576 -0.5509 0.20922  1.1451  1 2.52e-01    
## SRSF2_0.1              Genes -0.290822 -1.0813 0.13577 -2.1420  1 3.22e-02   *
## TET2_0.1               Genes -0.158347 -0.9488 0.10442 -1.5165  1 1.29e-01    
## TP53_0.1               Genes  0.686128 -0.1043 0.19933  3.4423  1 5.77e-04 ***
## U2AF1_0.1              Genes  0.711837 -0.0786 0.19998  3.5595  1 3.72e-04 ***
## age_10          Demographics -0.034319 -0.0343 0.10560 -0.3250  1 7.45e-01    
## gender          Demographics -0.096757 -0.0968 0.18251 -0.5302  1 5.96e-01    
## cohort          Demographics -1.297202 -1.2972 0.24120 -5.3781  1 7.53e-08 ***
## mu.Genes                  NA  0.790457      NA      NA      NA  1       NA    
## mu.Demographics           NA  0.000000      NA      NA      NA  1       NA    
```

```
survConcordance(allSurv ~ fitAll$linear.predictors)
```

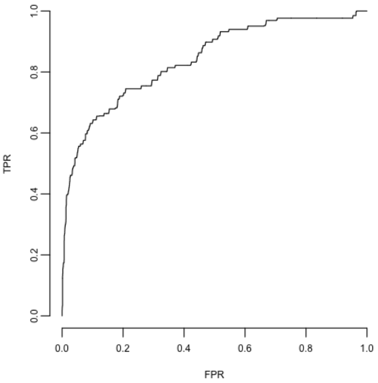
```
## Call:
## survConcordance(formula = allSurv ~ fitAll$linear.predictors)
## 
##   n= 950 
## Concordance= 0.8059859 se= 0.02746324
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  61799.000  14873.000      8.000      1.000   4211.763 
```

```
w <- c(which(allSurv[,1]==0)[-1]-1, nrow(allSurv))
s <- Surv(allSurv[w,2], allSurv[w,3])
t <- seq(0,22,0.1)
a <- AUC.uno(s, s, fitAll$linear.predictors[w], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```





```
r <- survivalROC(Stime = s[,1], status=s[,2], marker=fitAll$linear.predictors[w]-c
olMeans(fitAll$Z[w,]) %*% fitAll$coefficients, predict.time = 10, method="NNE", sp
an=0.001)
plot(r$FP, r$TP, type='s', xlab="FPR", ylab="TPR")
```



```
round(r$AUC, 3)
```

```
## [1] 0.84
```

7.4.2 Adjusted

```
fitWeighted <- CoxRFX(allX, allSurv, allGroups, which.mu=which.mu, sigma0=sigma0,
nu=nu, weights=weights)
waldWeighted <- WaldTest(fitWeighted)
```

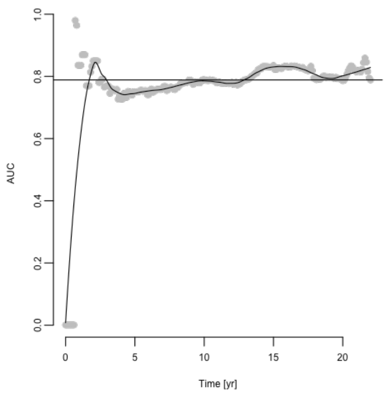
```
##                    group    coef coef-mu     sd      z df  p.value sig
## ASXL1_0.1          Genes  1.9907  0.0666 0.1328 14.985  1 9.18e-51 ***
## BCOR_0.1           Genes  2.1375  0.2134 0.1144 18.677  1 7.57e-78 ***
## CBL_0.1            Genes  0.3984 -1.5256 0.3634  1.096  1 2.73e-01
## DNMT3A_0.1         Genes  0.6589 -1.2652 0.1112  5.926  1 3.10e-09 ***
## IDH1_0.1           Genes  2.4306  0.5065 0.3313  7.337  1 2.18e-13 ***
## IDH2_0.1           Genes  0.8422 -1.0818 0.2181  3.862  1 1.13e-04 ***
## JAK2_0.1           Genes  1.8770 -0.0471 0.1954  9.607  1 7.44e-22 ***
## KDM6A_0.1          Genes  1.9370  0.0129 0.1241 15.607  1 6.51e-55 ***
## KMT2C_0.1          Genes  2.3674  0.4434 0.7114  3.328  1 8.75e-04 ***
## KMT2D_0.1          Genes  0.1632 -1.7609 0.4835  0.338  1 7.36e-01
## KRAS_0.1           Genes  1.9831  0.0590 0.1706 11.622  1 3.20e-31 ***
## NF1_0.1            Genes  1.5839 -0.3402 0.4410  3.592  1 3.29e-04 ***
## NRAS_0.1           Genes  2.3167  0.3926 0.1248 18.569  1 5.76e-77 ***
## PHF6_0.1           Genes  2.2266  0.3025 0.1241 17.937  1 6.04e-72 ***
## PTPN11_0.1         Genes  2.1631  0.2390 0.3107  6.962  1 3.35e-12 ***
## RAD21_0.1          Genes  1.8365 -0.0876 0.2512  7.311  1 2.65e-13 ***
## RUNX1_0.1          Genes  0.8106 -1.1134 0.1329  6.098  1 1.08e-09 ***
## SF3B1_0.1          Genes  3.1070  1.1829 0.3114  9.977  1 1.92e-23 ***
## SRSF2_0.1          Genes  1.3684 -0.5557 0.1491  9.176  1 4.47e-20 ***
## TET2_0.1           Genes  0.9527 -0.9714 0.1172  8.126  1 4.45e-16 ***
## TP53_0.1           Genes  5.0534  3.1293 0.3907 12.934  1 2.88e-38 ***
## U2AF1_0.1          Genes  4.1247  2.2006 0.3300 12.498  1 7.67e-36 ***
## age_10      Demographics -0.0962 -0.0962 0.0863 -1.114  1 2.65e-01
## gender      Demographics -0.0522 -0.0522 0.1044 -0.499  1 6.17e-01
## cohort      Demographics  0.0499  0.0499 0.0973  0.512  1 6.08e-01
```

```
survConcordance(fitWeighted$surv ~ fitWeighted$linear.predictor, weights=weights)
```

```
## Call:
## survConcordance(formula = fitWeighted$surv ~ fitWeighted$linear.predictor,
##     weights = weights)
##
##   n= 950
## Concordance= 0.7778849 se= 0.02802535
##   concordant   discordant     tied.risk     tied.time     std(c-d)
## 6313552.2348 1802641.1313      317.7601        1.0000  454936.0746
```

Dynamic/cumulative AUC

```
w <- c(which(allSurv[,1]==0)[-1]-1, nrow(allSurv))
survAll2 <- Surv(allSurv[w,2], allSurv[w,3])
t <- seq(0,22,0.1)
a <- AUC.uno(survAll2, survAll2, fitWeighted$linear.predictor[w], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```



```
round(a$iauc, 3)
```

```
## [1] 0.789
```

```
png("./figures/combined.ajd.coxph.auc.uno.png", width = 9, height = 10, units = "c
m", res = 500)
par(mar = c(3.2, 3.2, 4, 2) + 0.1, mgp=c(2,0.5,0), bty="L",  tcl =-0.2, las = 1, c
ex=1)
plot(a$times, a$auc, xlab="Time (years)", ylab="AUC", pch=16, col="grey80", ylim =
c(0,1.0))
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc, lty = 3, lwd = 1)
#mtext("Combined adjusted Cox PH", font= 2, side = 3, line = 0.5)
legend("bottomright", bty = "n", legend = paste("AUC = ",round(a$iauc,2)))
dev.off()
```

```
## pdf
##   2
```

Time-depenent ROC

```
r <- survivalROC(Stime = survAll2[,1], status=survAll2[,2], marker=fitWeighted$lin
ear.predictors[w]-colMeans(fitWeighted$Z[w,]) %*% fitWeighted$coefficients, predic
t.time = 10, method="NNE", span=0.001)
round(r$AUC, 3)
```

```
## [1] 0.791
```

```
png("./figures/Combined.adj.coxph.roct.png", width = 9, height = 10, units = "cm",
res = 500)
par(mar = c(3.2, 3.2, 4, 2) + 0.1, mgp=c(2,0.5,0), bty="L",  tcl =-0.2, las = 1, c
ex = 1)
plot(r$FP, r$TP, type='s',
     xlab="False Positive Rate", ylab="True Positive Rate",
     col = "black")
abline(a = 0, b = 1, col = "grey70", lty = 1, lwd = 1)
legend("bottomright", bty = "n", legend = paste("AUC = ",round(r$AUC,2)))
dev.off()
```

```
## pdf
##   2
```

7.4.3 Bootstrap

```
coefWeightedBoot <- sapply(1:100, function(foo){
            set.seed(foo)
            b <- unique(sample(1:nrow(allX), replace=TRUE))
            fitWeighted <- CoxRFX(allX[b,], allSurv[b,], allGroups, which.mu=which
.mu, sigma0=sigma0, nu=5, weights=weights[b])
            c(coef(fitWeighted), 'mu.Genes'=fitWeighted$mu["Genes"])
        })
```

```
concBoots <- sapply(1:100, function(foo){
            set.seed(foo)
            b <- unique(sample(1:nrow(allX), replace=TRUE))
            oob <- !1:nrow(allX) %in% b
            c(inb=as.numeric(survConcordance(allSurv[b,]~ as.matrix(allX)[b,] %*%
coefWeightedBoot[-26,foo], weights=weights[b])$concordance),
                    oob=as.numeric(survConcordance(allSurv[oob,]~ as.matrix(allX)[
oob,] %*% coefWeightedBoot[-26,foo],weights=weights[oob])$concordance),
                    auc = AUC.uno(survAll2[oob[w],], survAll2[oob[w],], as.matrix(
allX)[w,][oob[w],] %*% coefWeightedBoot[-26,foo], times=t)$iauc
            )
        })

apply(concBoots,1,quantile)
```

```
##            inb       oob       auc
## 0%   0.7127155 0.6414249 0.6163769
## 25%  0.7623231 0.7268340 0.7333587
## 50%  0.7757864 0.7643297 0.7833229
## 75%  0.7985773 0.7875492 0.8223659
## 100% 0.8519811 0.8713292 0.8805585
```

7.4.4 Forest plot

Figure 3

```
pal1 <- c("#C32B4A", "#3F76B4", "#57B2AB", "#5E4FA2", "#EB6046")
rownames(waldWeighted)
```

```
##  [1] "ASXL1_0.1"  "BCOR_0.1"   "CBL_0.1"    "DNMT3A_0.1" "IDH1_0.1"   "IDH2_0.1
"   "JAK2_0.1"   "KDM6A_0.1"
##  [9] "KMT2C_0.1"  "KMT2D_0.1"  "KRAS_0.1"   "NF1_0.1"    "NRAS_0.1"   "PHF6_0.1
"   "PTPN11_0.1" "RAD21_0.1"
## [17] "RUNX1_0.1"  "SF3B1_0.1"  "SRSF2_0.1"  "TET2_0.1"   "TP53_0.1"   "U2AF1_0.
1"  "age_10"     "gender"
## [25] "cohort"
```

```
png("./figures/Combined.adj.coxph.boostrapped.forest.png", width = 15.5, height =
17, units = "cm", res = 800)
par(bty="n", mar=c(3,6,3,15)+.5, mgp=c(2,0.5,0), xpd=FALSE, tcl=-.25, cex = 0.9)
c <- c(waldWeighted[-25,"coef"], "mu"=fitWeighted$mu["Genes"]); names(c)[1:24] <-
rownames(waldWeighted)[-25] #-25 removes 'cohort' variable
o <- c(23:24,1:22,25)
s <- c(rep(1,2), rep(.5, 23))
c <- exp(c*c(rep(0.5,22), c(1,1),0.5))
ci <- apply(coefWeightedBoot,1,quantile, c(0.025,0.975))[,-25] * rep(c(rep(0.5,22)
, c(1,1),0.5), each=2)
y <- rev(seq_along(c))
plot(c[o], y, xlab="Hazard ratio", log='x', ylab='', xaxt = "n", yaxt="n", pch=NA,
xlim=c(0.5,50))
atx <- axTicks(1)
axis(1,at=atx,labels=atx)
segments(x0=0.5, x1 = 50, y0=y, y1=y, col="#EEEEEE", lty=1)
abline(v=1, lty=1, col="grey")
abline(v=c["mu.Genes"], col=mg14::colTrans("#57B2AB"), lty=1)
segments(exp(ci[1,o]), y, exp(ci[2,o]),y)
points(c[o], y, xlab="",  bg=pal1[3], cex=2, pch=c(rep(21,24), 23))
m1 <- match(names(c)[o],rownames(waldWeightedToronto))[-25]
points(exp(c(waldWeightedToronto$coef[m1], fitWeightedToronto$mu["Genes"])*s), y,b
g=pal1[4], pch=c(rep(21,24), 23), cex=1)
m2 <- match(names(c)[o],rownames(waldWeightedSanger))[-25]
points(exp(c(waldWeightedSanger$coef[m2], fitWeightedSanger$mu["Genes"])*s), y,bg=
pal1[5], pch=c(rep(21,24), 23), cex=1)
mtext(side=2, sub("mu.Genes","Av. gene", sub("_.+","", sub("age", "Age", sub("gend
er", "Gender", names(c)[o])))), at=y, las=2, cex=0.85, font=c(1,1,rep(3,22),1))
r <- sapply(split(as.data.frame(allX>0), control), colMeans)
f <- sapply(split(allX, control), apply, 2, function(x) mean(x[x>0]))
par(xpd=NA)
points(rep(100,22),y[3:24], cex=sqrt(r[o[3:24],2]*10), pch=21, bg=pal1[2])
points(rep(100*1.5,22), y[3:24], cex=sqrt(r[o[3:24],1]*10), pch=21, bg=pal1[1])
points(rep(360,22),y[3:24], cex=sqrt(f[o[3:24],2]), pch=21, bg=set1[2])
points(rep(360*1.5,22), y[3:24], cex=sqrt(f[o[3:24],1]), pch=21, bg=pal1[1])
legend(x=0.8, y=27.8, pch=21, pt.bg=pal1[c(4,5,3)], c("DC","VC","Combined"), bty="
n", ncol=3, text.width=0.25)
text(y=24, x=100, "     Frequency", cex = 0.92)
text(y=24, x=360*1.5, "VAF    ", cex = 0.92)
axis(1, at=c(100,100*1.5), c("Control ","Pre-AML "), las=2, line=-1, cex = 0.89)
axis(1, at=c(360,360*1.5), c("Control ","Pre-AML "), las=2, line=-1, cex = 0.89)
dev.off()
```

```
## pdf
##   2
```

```
Fig3Data1 <- data.frame(Parameter = sapply(strsplit(names(c[o]), "_"), "[", 1),
                        CombinedModel.HR = round(c[o], 1),
                        CombinedModel.HR.CI2.5 = round(exp(ci[1,o]), 1),
                        CombinedModel.HR.CI97.5 = round(exp(ci[2,o]),1),
                        DC.HR = round(exp(c(waldWeightedToronto$coef[m1], fitWeigh
tedToronto$mu["Genes"])*s),1),
                        VC.HR = round(exp(c(waldWeightedSanger$coef[m2], fitWeight
edSanger$mu["Genes"])*s),1)
                        )
rownames(Fig3Data1) <- NULL
head(Fig3Data1)
```

	Parameter	CombinedModel....	CombinedModel.HR.CI2.5	CombinedModel.HR.CI97.5	D(
	<fctr>	<dbl>	<dbl>	<dbl>	<c
1	age	0.9	0.8	1.0	
2	gender	0.9	0.8	1.2	
3	ASXL1	2.7	2.5	6.6	

3 ASXL1	2.7	2.5	[illegible]
4 BCOR	2.9	2.5	11.1
5 CBL	1.2	1.0	5.1
6 DNMT3A	1.4	1.2	1.8

6 rows

```
table(rownames(r)==rownames(f))
```

```
##
## TRUE
##   25
```

```
Fig3Data2 <- data.frame(Parameter = sapply(strsplit(rownames(r), "_"), "[", 1)[1:2
2],
                        Frequency_PreAML = round(r[1:22, 1],3),
                        Frequency_Controls = round(r[1:22, 2],3),
                        MeanVAF_PreAML = round(f[1:22, 1],3),
                        MeanVAF_Control = round(f[1:22, 2],3))
head(Fig3Data2)
```

	Parameter <fctr>	Frequency_PreAML <dbl>	Frequency_Controls <dbl>	MeanVAF_Pre... <dbl>	MeanV
ASXL1_0.1	ASXL1	0.090	0.021	1.262	
BCOR_0.1	BCOR	0.008	0.001	0.117	
CBL_0.1	CBL	0.030	0.011	0.414	
DNMT3A_0.1	DNMT3A	0.391	0.212	0.950	
IDH1_0.1	IDH1	0.023	0.001	1.156	
IDH2_0.1	IDH2	0.038	0.001	1.848	

6 rows

```
rownames(Fig3Data2) <- NULL
Fig3Data <- left_join(x = Fig3Data1, y = Fig3Data2, by = 'Parameter')
```

```
## Warning: Column `Parameter` joining factors with different levels, coercing to
character vector
```

```
Fig3Data$Parameter <- ifelse(Fig3Data$Parameter == "mu.Genes", "Av.gene", Fig3Data
$Parameter)
#View(Fig3Data)
write_csv(Fig3Data, "./figures/Figure3_Data.csv")
```

7.4.5 Dichotomous variables

```
allXDich <- allX
allXDich[allGroups=="Genes"] <- (allXDich[allGroups=="Genes"] > 0) + 0
fitWeightedDich <- CoxRFX(allXDich, allSurv, allGroups, which.mu=which.mu, sigma0=
sigma0, nu=nu, weights=weights)

WaldTest(fitWeightedDich)
```

```
##                    group    coef coef-mu     sd       z df  p.value sig
## ASXL1_0.1          Genes  1.3797 -0.3942 0.3175  4.3456  1 1.39e-05 ***
## BCOR_0.1           Genes  2.5308  0.7570 0.8406  3.0106  1 2.61e-03  **
## CBL_0.1            Genes  0.3932 -1.3806 0.4991  0.7879  1 4.31e-01    
## DNMT3A_0.1         Genes  0.7794 -0.9944 0.2049  3.8048  1 1.42e-04 ***
## IDH1_0.1           Genes  2.0403  0.2665 0.5817  3.5073  1 4.53e-04 ***
## IDH2_0.1           Genes  3.9907  2.2169 0.5363  7.4414  1 9.96e-14 ***
## JAK2_0.1           Genes  3.2315  1.4577 0.3911  8.2629  1 1.42e-16 ***
## KDM6A_0.1          Genes  0.7396 -1.0343 0.7822  0.9456  1 3.44e-01    
## KMT2C_0.1          Genes -0.4630 -2.2368 0.5910 -0.7834  1 4.33e-01    
## KMT2D_0.1          Genes  0.8142 -0.9597 0.9409  0.8653  1 3.87e-01    
## KRAS_0.1           Genes -0.0209 -1.7948 0.7030 -0.0298  1 9.76e-01    
## NF1_0.1            Genes -1.1385 -2.9124 0.8236 -1.3824  1 1.67e-01    
## NRAS_0.1           Genes  1.6320 -0.1419 0.7812  2.0891  1 3.67e-02   *
## PHF6_0.1           Genes  4.0915  2.3176 0.7069  5.7883  1 7.11e-09 ***
## PTPN11_0.1         Genes  2.2597  0.4859 0.6548  3.4510  1 5.59e-04 ***
## RAD21_0.1          Genes  1.0923 -0.6816 0.9283  1.1767  1 2.39e-01    
## RUNX1_0.1          Genes  2.6557  0.8818 0.5738  4.6284  1 3.69e-06 ***
## SF3B1_0.1          Genes  0.0815 -1.6924 0.6027  0.1352  1 8.92e-01    
## SRSF2_0.1          Genes  4.2431  2.4693 0.3084 13.7566  1 4.65e-43 ***
## TET2_0.1           Genes  0.9715 -0.8023 0.2351  4.1328  1 3.58e-05 ***
## TP53_0.1           Genes  2.0033  0.2295 0.4168  4.8067  1 1.53e-06 ***
## U2AF1_0.1          Genes  5.7172  3.9433 0.4178 13.6831  1 1.28e-42 ***
## age_10     Demographics -0.3024 -0.3024 0.0958 -3.1571  1 1.59e-03  **
## gender     Demographics -0.0512 -0.0512 0.1362 -0.3759  1 7.07e-01    
## cohort     Demographics  0.2569  0.2569 0.1435  1.7896  1 7.35e-02   .
```

```
survConcordance(allSurv ~ fitWeightedDich$linear.predictors, weights=weights)
```

```
## Call:
## survConcordance(formula = allSurv ~ fitWeightedDich$linear.predictors, 
##     weights = weights)
## 
##   n= 950 
## Concordance= 0.764251 se= 0.02802535
##   concordant   discordant    tied.risk    tied.time     std(c-d) 
## 6202805.3608 1913213.1798     492.5856       1.0000  454936.0734
```

7.4.6 Bootstrap adjustment

To compare to the weighted CoxRFX models

```
set.seed(42)

p <- c(rep(n_total_sanger, sum(cohort=="Sanger" & control)), rep(n_total_toronto, 
sum(cohort=="Toronto" & control)))
b42 <- c(sample(which(control), size=round(n_total) - sum(!control), prob=p, repla
ce=TRUE), which(!control))

fitBoot <- CoxRFX(allX[b42,], allSurv[b42,], allGroups, which.mu=which.mu, sigma0=
sigma0, nu=nu)

set.seed(42)
b <- c(sample(which( sangerData$Diagnosis=="Control"), size=round(n_total_sanger) 
- sum(sangerData$Diagnosis!="Control"), replace=TRUE), which(sangerData$Diagnosis!
="Control"))

fitBootSanger <- CoxRFX(sangerX[b,], sangerSurv[b,], sangerGroups, which.mu=which.
mu, sigma0=sigma0, nu=nu)

survConcordance(fitBootSanger$surv ~ fitBootSanger$linear.predictors)
```

```
## Call:
## survConcordance(formula = fitBootSanger$surv ~ fitBootSanger$linear.predictors)
## 
```

```
##   n= 10407
## Concordance= 0.8334695 se= 0.05475909
## concordant discordant  tied.risk  tied.time   std(c-d)
##   140833.0    28139.0        0.0        0.0    18505.5
```

```
waldBootSanger <- WaldTest(fitBootSanger)
```

```
##                    group     coef  coef-mu      sd       z df  p.value sig
## ASXL1_0.1          Genes  2.75130  0.85036 0.44987  6.1157  1 9.61e-10 ***
## CBL_0.1            Genes  0.90179 -0.99914 1.17452  0.7678  1 4.43e-01    
## DNMT3A_0.1         Genes  0.75840 -1.14254 0.22408  3.3845  1 7.13e-04 ***
## JAK2_0.1           Genes -0.20568 -2.10662 0.92220 -0.2230  1 8.24e-01    
## KMT2C_0.1          Genes  2.16912  0.26819 0.96833  2.2401  1 2.51e-02   *
## KMT2D_0.1          Genes  0.06618 -1.83475 0.76576  0.0864  1 9.31e-01    
## KRAS_0.1           Genes  2.31066  0.40972 0.38106  6.0638  1 1.33e-09 ***
## NF1_0.1            Genes  1.57512 -0.32581 0.77819  2.0241  1 4.30e-02   *
## NRAS_0.1           Genes  1.84937 -0.05157 0.35761  5.1715  1 2.32e-07 ***
## RAD21_0.1          Genes  1.70593 -0.19501 0.58727  2.9049  1 3.67e-03  **
## SF3B1_0.1          Genes  1.54550 -0.35544 0.87032  1.7758  1 7.58e-02   .
## SRSF2_0.1          Genes  1.40565 -0.49529 0.27962  5.0271  1 4.98e-07 ***
## TET2_0.1           Genes  1.25279 -0.64815 0.13571  9.2317  1 2.66e-20 ***
## TP53_0.1           Genes  4.63845  2.73751 0.89272  5.1959  1 2.04e-07 ***
## U2AF1_0.1          Genes  5.78946  3.88853 0.73724  7.8528  1 4.07e-15 ***
## age_10      Demographics  0.04278  0.04278 0.11873  0.3603  1 7.19e-01    
## gender      Demographics -0.01852 -0.01852 0.10088 -0.1836  1 8.54e-01    
## systol_100         Blood  0.02344  0.02344 0.04556  0.5145  1 6.07e-01    
## diastol_100        Blood  0.04133  0.04133 0.03020  1.3686  1 1.71e-01    
## bmi_10             Blood  0.14916  0.14916 0.08426  1.7702  1 7.67e-02   .
## cholestl_10        Blood  0.00303  0.00303 0.01547  0.1958  1 8.45e-01    
## triglyc            Blood -0.02770 -0.02770 0.11803 -0.2347  1 8.14e-01    
## hdl                Blood -0.12117 -0.12117 0.08479 -1.4291  1 1.53e-01    
## ldl                Blood  0.13479  0.13479 0.11448  1.1775  1 2.39e-01    
## lym                Blood  0.08408  0.08408 0.10435  0.8057  1 4.20e-01    
## mcv_100            Blood -0.02485 -0.02485 0.00798 -3.1160  1 1.83e-03  **
## rdw_10             Blood  0.06629  0.06629 0.01703  3.8934  1 9.88e-05 ***
## wbc_10             Blood  0.01199  0.01199 0.04735  0.2532  1 8.00e-01    
## plt_100            Blood  0.09163  0.09163 0.10006  0.9158  1 3.60e-01    
## hgb_10             Blood  0.03986  0.03986 0.02497  1.5960  1 1.10e-01    
```

```
set.seed(42)
b <- c(sample(which( torontoData$Diagnosis=="Control"), size=round(n_total_toronto
) - sum(torontoData$Diagnosis!="Control"), replace=TRUE), which(torontoData$Diagno
sis!="Control"))

fitBootToronto <- CoxRFX(torontoX[b,], torontoSurv[b,], torontoGroups, which.mu=wh
ich.mu, sigma0=sigma0, nu=nu)
survConcordance(fitBootToronto$surv ~ fitBootToronto$linear.predictors)
```

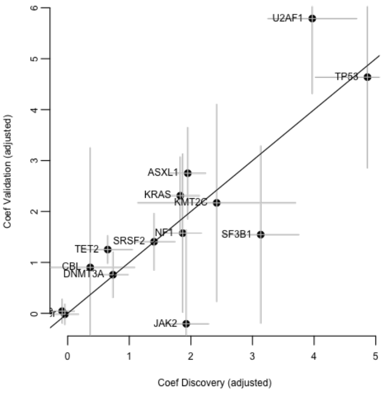
```
## Call:
## survConcordance(formula = fitBootToronto$surv ~ fitBootToronto$linear.predictor
s)
## 
##   n= 72378
## Concordance= 0.7750173 se= 0.03055346
## concordant discordant  tied.risk  tied.time   std(c-d)
##  4722585.0  1370937.0        0.0        1.0   372356.4
```

```
waldWeightedToronto <- WaldTest(fitBootToronto)
```

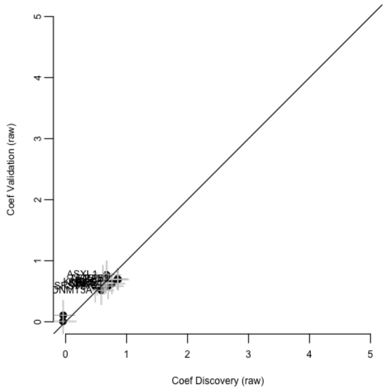
```
##                   group    coef  coef-mu     sd      z df  p.value sig
## ASXL1_0.1         Genes  1.9494  0.01801 0.1451 13.430  1 4.03e-41 ***
## CALR_0.1          Genes  0.9415 -0.98990 0.7233  1.302  1 1.93e-01    
## CBL_0.1           Genes  0.3663 -1.56509 0.3604  1.016  1 3.09e-01    
## DNMT3A_0.1        Genes  0.7358 -1.19559 0.1243  5.921  1 3.20e-09 ***
## IDH1_0.1          Genes  2.3973  0.46594 0.3355  7.145  1 8.98e-13 ***
## IDH2_0.1          Genes  0.8078 -1.12360 0.2283  3.538  1 4.03e-04 ***
## JAK2_0.1          Genes  1.9240 -0.00738 0.1822 10.562  1 4.49e-26 ***
## KDM6A_0.1         Genes  1.9436  0.01219 0.1340 14.506  1 1.12e-47 ***
## KMT2C_0.1         Genes  2.4194  0.48806 0.6410  3.774  1 1.60e-04 ***
## KRAS_0.1          Genes  1.8282 -0.10316 0.1559 11.725  1 9.46e-32 ***
## NF1_0.1           Genes  1.8677 -0.06366 0.1512 12.353  1 4.69e-35 ***
## PHF6_0.1          Genes  2.1755  0.24415 0.1302 16.711  1 1.08e-62 ***
## PTPN11_0.1        Genes  2.5369  0.60555 0.2217 11.445  1 2.49e-30 ***
## RUNX1_0.1         Genes  0.7795 -1.15181 0.1359  5.738  1 9.57e-09 ***
## SF3B1_0.1         Genes  3.1337  1.20231 0.3091 10.138  1 3.76e-24 ***
## SRSF2_0.1         Genes  1.4023 -0.52910 0.1703  8.235  1 1.80e-16 ***
## TET2_0.1          Genes  0.6503 -1.28104 0.2012  3.232  1 1.23e-03  **
## TP53_0.1          Genes  4.8664  2.93502 0.4220 11.532  1 9.14e-31 ***
## U2AF1_0.1         Genes  3.9705  2.03910 0.3601 11.025  1 2.89e-28 ***
## age_10     Demographics -0.0891 -0.08907 0.0998 -0.892  1 3.72e-01    
## gender     Demographics -0.0449 -0.04493 0.1114 -0.403  1 6.87e-01    
```

Compare results

```
i <- intersect(rownames(waldBootSanger), rownames(waldWeightedToronto))
plot( waldWeightedToronto[i,"coef"], waldBootSanger[i, "coef"], xlab="Coef Discove
ry (adjusted)", ylab="Coef Validation (adjusted)", pch=19, cex=1)#sqrt(colMeans(rb
ind(sangerX[,i], torontoX[,i])>0)*100))
segments(waldWeightedToronto[i,"coef"]  - 2*waldWeightedToronto[i,"sd"], waldBootS
anger[i, "coef"], waldWeightedToronto[i,"coef"]  + 2*waldWeightedToronto[i,"sd"],
waldBootSanger[i, "coef"], col="grey" )
segments(waldWeightedToronto[i,"coef"]  , waldBootSanger[i, "coef"]-  2*waldBootSa
nger[i,"sd"], waldWeightedToronto[i,"coef"] , waldBootSanger[i, "coef"] +2*waldBoo
tSanger[i,"sd"], col="grey")
text(labels=sub("_.+","", i), waldWeightedToronto[i,"coef"], waldBootSanger[i, "co
ef"], pos=2, adj=c(0,1))
abline(0,1)
```



```
plot( waldToronto[i,"coef"], waldSanger[i, "coef"], xlab="Coef Discovery (raw)", y
lab="Coef Validation (raw)", pch=19, cex=1, ylim=c(0,5),xlim=c(0,5))#sqrt(colMeans
(rbind(sangerX[,i], torontoX[,i])>0)*100))
segments(waldToronto[i,"coef"]  - 2*waldToronto[i,"sd"], waldSanger[i, "coef"], wa
ldToronto[i,"coef"]  + 2*waldToronto[i,"sd"], waldSanger[i, "coef"], col="grey" )
segments(waldToronto[i,"coef"]  , waldSanger[i, "coef"]-  2*waldSanger[i,"sd"], wa
ldToronto[i,"coef"] , waldSanger[i, "coef"] +2*waldSanger[i,"sd"], col="grey")
text(labels=sub("_.+","", i), waldToronto[i,"coef"], waldSanger[i, "coef"], pos=2,
adj=c(0,1))
abline(0,1)
```



7.4.7 LOOCV

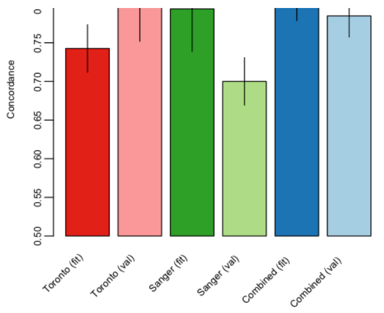
```
samples <- factor(c(as.character(sangerData$Individual), as.character(torontoData$
Sample)))
```

```
looAll <- do.call("rbind",mclapply(levels(samples), function(l){
                    i <- samples!=l
                    f <<- CoxRFX(allX[i,], allSurv[i,], allGroups, which.mu=which.
mu, sigma0=sigma0, nu=nu)
                    p <- as.matrix(allX[!i,,drop=FALSE]) %*% f$coefficients
                    r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(
f$coefficients), byrow=TRUE), linear.predictor=p)
                    colnames(r) <- c(names(f$coefficients), "linear.predictor")
                    as.data.frame(r)
                }, mc.cores=4))
looAll <- looAll[order(order(samples)),]
pp <- looAll$linear.predictor

c <- rbind(
         `Toronto (fit)`=as.data.frame(survConcordance(torontoSurv ~ fitToronto$lin
ear.predictors)[c("concordance","std.err")]),
         `Toronto (val)`=as.data.frame(survConcordance(sangerSurv ~ pS[,1])[c("conc
ordance","std.err")]),
         `Sanger (fit)`=as.data.frame(survConcordance(sangerSurv ~ fitSanger$linear
.predictors)[c("concordance","std.err")]),
         `Sanger (val)`=as.data.frame(survConcordance(torontoSurv ~ pT[,1])[c("conc
ordance","std.err")]),
         `Combined (fit)`=as.data.frame(survConcordance(allSurv ~ fitAll$linear.pre
dictors)[c("concordance","std.err")]),
         `Combined (val)`=as.data.frame(survConcordance(allSurv ~ pp)[c("concordanc
e","std.err")]))

c
```

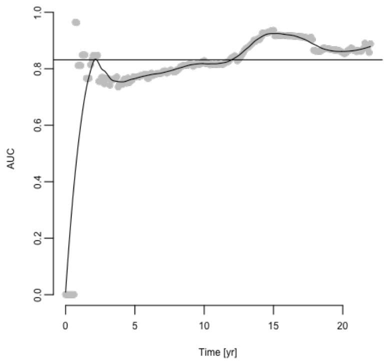
	concordance <dbl>	std.err <dbl>
Toronto (fit)	0.7426378	0.03079247
Toronto (val)	0.8069747	0.05514445
Sanger (fit)	0.7939150	0.05514512
Sanger (val)	0.7000180	0.03079247
Combined (fit)	0.8059859	0.02746324
Combined (val)	0.7847548	0.02746328

6 rows

```
par(mar=c(5,3,1,1), mgp=c(2,.5,0))
b <- barplot(c$concordance-0.5, ylab="Concordance", col=rev(RColorBrewer::brewer.p
al(6,"Paired")), ylim=c(0.5,0.88), offset=0.5)
mg14::rotatedLabel(x=b, labels=rownames(c))
segments(b,c$concordance+c$std.err,b,c$concordance-c$std.err)
```



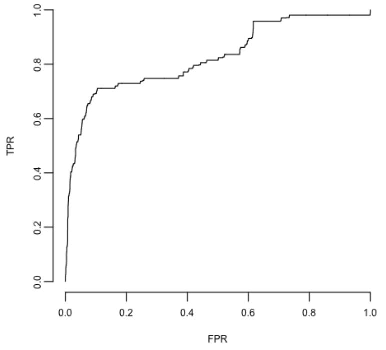
```
w <- c(which(allSurv[,1]==0)[-1]-1, nrow(allSurv))
survAll2 <- Surv(allSurv[w,2], allSurv[w,3])
t <- seq(0,22,0.1)
a <- AUC.uno(survAll2, survAll2, looAll$linear.predictor[w], times=t)
plot(a$times, a$auc, xlab="Time [yr]", ylab="AUC", pch=16, col='grey')
lines(a$times, predict(loess(a$auc ~ a$times, span=0.25)))
abline(h=a$iauc)
```



```
round(a$iauc, 3)
```

```
## [1] 0.832
```

```
r <- survivalROC(Stime = survAll2[,1], status=survAll2[,2], marker=looAll$linear.p
redictor[w], predict.time = 10, method="NNE", span=0.001)#0.25*nrow(s)^(-0.20))
plot(r$FP, r$TP, type='s', xlab="FPR", ylab="TPR")
```

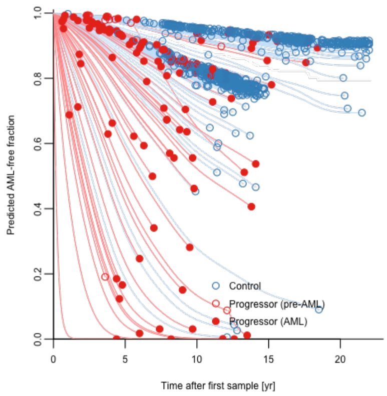


```
round(r$AUC, 3)
```

```
## [1] 0.825
```

7.4.7.1 Individual Predictions (non-adjusted)

```
plot(survfit(allSurv~1), conf.int=FALSE, xlab='Time after first sample [yr]', ylab
='Predicted AML-free fraction', col='white', bty='L', yaxs='i', ylim=c(0,1.01))
d <- data.frame(t=NULL, s=NULL, pch=NULL, col=character())
for(i in unique(samples)){
    km <- exp(predict(smooth.spline(log(summary(survfit(allSurv[samples!=i,]~1), t
imes=t)$surv), df=10))$y)
    l0 <- colMeans(fitAll$Z[samples!=i,,drop=FALSE]) %*% as.numeric(looAll[samples
==i,][1,colnames(fitAll$Z)])
    kmi <- function(km, s, lp, l0){
        .kmi <- function(km, sj, lpj, l0) km[t >= sj[,1] & t <= sj[,2]]^exp(lpj-l0
)
        k0 <- 1
        for(j in 1:nrow(s)) {
            k <- .kmi(km, s[j,], lp[j], l0)
            k <- k * k0/k[1]
            w <- t >= s[j,1] & t <= s[j,2]
            k0 <- k[length(k)]
            c <- if(s[nrow(s),3]==1) set1[1] else set1[2]
            #if(c==set1[1]) next
            lines(t[w], k, col=mg14:::colTrans(c), type='l')
            p <- if(s[j,3]==1) 19 else 1
            #points(t[w][length(k)], k[length(k)], col=c, pch=p)
            d <<- rbind(d, data.frame(t=t[w][length(k)], s=k[length(k)], pch=p, co
l=c))
        }
    }
    kmi(km, allSurv[samples==i,], looAll$linear.predictor[samples==i], l0)
}
points(d$t, d$s, pch=d$pch, col=as.character(d$col))
legend("bottomright", pch=c(1,1,19), col=c(set1[2], set1[1], set1[1]), legend=c("C
ontrol", "Progressor (pre-AML)", "Progressor (AML)"), bty='n')
```



7.4.7.2 Jackknife variance

```
i <- !duplicated(samples)
coef.jack <- colMeans(looAll[i,-ncol(looAll)])
var.jack <- rowSums((t(looAll[i,-ncol(looAll)]) - coef.jack)^2) * (sum(i)-1)/sum(i
)

p.jack <- pchisq(coef.jack^2/var.jack,1, lower.tail=FALSE)
```

```
data.frame(coef.jack, p.jack, sig=mg14::sig2star(p.jack), n=colSums(allX[i,]>0))
```

	coef.jack <dbl>	p.jack <dbl>	sig <fctr>	n <dbl>
ASXL1_0.1	0.74835623	1.277998e-05	***	26
BCOR_0.1	0.80859507	2.311062e-04	***	1
CBL_0.1	0.47795378	3.123703e-01		12
DNMT3A_0.1	0.55685260	7.358773e-06	***	194
IDH1_0.1	0.81211760	5.586147e-10	***	3
IDH2_0.1	0.51251777	1.351015e-01		6
JAK2_0.1	0.75979214	3.181470e-08	***	10
KDM6A_0.1	0.79059980	7.666406e-05	***	3
KMT2C_0.1	0.85878619	5.304616e-04	***	6
KMT2D_0.1	0.40005469	3.584861e-01		1

1-10 of 25 rows Previous **1** 2 3 Next

7.4.8 Multiple bootstraps

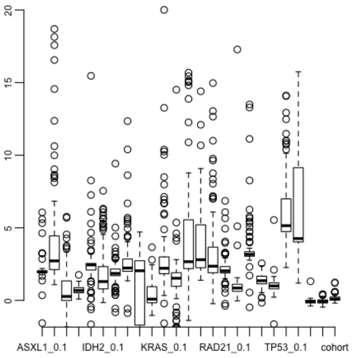
```
save(file="boot.RData", control, allX, allSurv, sigma0, nu, which.mu, allGroups, n
_total, cohort, p)
```

```
fitBoots <- simplify2array(mclapply(1:100, function(foo){
                    set.seed(foo)
                    w <- which(control)
                    s <- sample(seq_along(which(control)), replace=TRUE)
                    b <- c(sample(which(control)[s], size=round(n_total) - sum(!co
ntrol), prob=p[s], replace=TRUE), sample(which(!control), replace=TRUE))
                    fitBoot <- CoxRFX(allX[b,], allSurv[b,], allGroups, which.mu=w
hich.mu, sigma0=sigma0, nu=nu)
                    fitBoot$coefficients
                }, mc.cores=4))
save(fitBoots, file="fitBoots.RData")
```

```
load('fitBoots.RData')

WaldTest(fitBoot)
```

```
##                   group    coef coef-mu     sd      z df  p.value sig
## ASXL1_0.1         Genes  1.9782  0.0682 0.1330 14.873  1 4.90e-50 ***
## BCOR_0.1          Genes  2.1204  0.2104 0.1157 18.319  1 5.81e-75 ***
## CBL_0.1           Genes  0.3747 -1.5352 0.3614  1.037  1 3.00e-01
## DNMT3A_0.1        Genes  0.6499 -1.2600 0.1133  5.735  1 9.77e-09 ***
## IDH1_0.1          Genes  2.4215  0.5116 0.3299  7.341  1 2.12e-13 ***
## IDH2_0.1          Genes  0.8614 -1.0486 0.2191  3.931  1 8.47e-05 ***
## JAK2_0.1          Genes  1.8708 -0.0391 0.1956  9.562  1 1.15e-21 ***
## KDM6A_0.1         Genes  1.9211  0.0112 0.1251 15.363  1 2.92e-53 ***
## KMT2C_0.1         Genes  2.3935  0.4836 0.7067  3.387  1 7.07e-04 ***
## KMT2D_0.1         Genes  0.1309 -1.7790 0.4810  0.272  1 7.86e-01
## KRAS_0.1          Genes  1.9602  0.0503 0.1717 11.415  1 3.53e-30 ***
## NF1_0.1           Genes  1.5704 -0.3396 0.4386  3.580  1 3.43e-04 ***
## NRAS_0.1          Genes  2.3060  0.3960 0.1213 19.014  1 1.31e-80 ***
## PHF6_0.1          Genes  2.2127  0.3028 0.1241 17.835  1 3.80e-71 ***
## PTPN11_0.1        Genes  2.1333  0.2233 0.3110  6.860  1 6.86e-12 ***
## RAD21_0.1         Genes  1.8285 -0.0815 0.2524  7.244  1 4.36e-13 ***
## RUNX1_0.1         Genes  0.8075 -1.1025 0.1325  6.095  1 1.10e-09 ***
## SF3B1_0.1         Genes  3.0963  1.1863 0.3107  9.967  1 2.13e-23 ***
## SRSF2_0.1         Genes  1.3408 -0.5692 0.1503  8.923  1 4.55e-19 ***
## TET2_0.1          Genes  0.9202 -0.9897 0.1179  7.807  1 5.85e-15 ***
## TP53_0.1          Genes  5.0203  3.1104 0.3921 12.803  1 1.57e-37 ***
## U2AF1_0.1         Genes  4.0999  2.1900 0.3306 12.402  1 2.54e-35 ***
## age_10     Demographics -0.0761 -0.0761 0.0912 -0.835  1 4.04e-01
## gender     Demographics -0.0530 -0.0530 0.1157 -0.458  1 6.47e-01
## cohort     Demographics  0.1992  0.1992 0.1103  1.806  1 7.09e-02   .
```

```
boxplot(t(fitBoots), ylim=c(-1,20))
points(fitBoot$coefficiencts, pch="*", col='red')
```



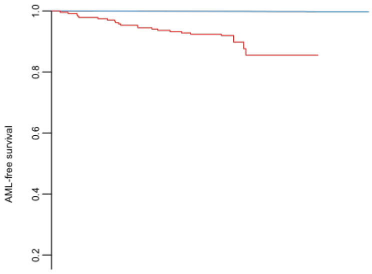
Concordance on out of bag samples

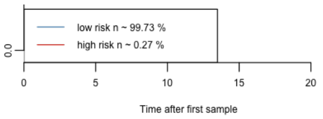
```
concBoots <- sapply(1:100, function(foo){
            set.seed(foo)
            w <- which(control)
            s <- sample(seq_along(which(control)), replace=TRUE)
            b <- c(sample(which(control)[s], size=round(n_total) - sum(!control),
prob=p[s], replace=TRUE), sample(which(!control), replace=TRUE))
            oob <- !1:nrow(allX) %in% b
            oos <- c(sample(which(oob & control), size=round(n_total) - sum(!contr
ol), replace=TRUE), sample(which(oob&!control), size=sum(!control), replace=TRUE))
            c(inb=as.numeric(survConcordance(allSurv[b,]~ as.matrix(allX)[b,] %*%
fitBoots[,foo])$concordance),
                    oob=as.numeric(survConcordance(allSurv[oob,]~ as.matrix(allX)[
oob,] %*% fitBoots[,foo])$concordance),
                    oos=as.numeric(survConcordance(allSurv[oos,]~ as.matrix(allX)[
oos,] %*% fitBoots[,foo])$concordance)
            )
        })
```

```
looAllWeighted <- do.call("rbind",mclapply(levels(samples), function(l){
                    i <- samples!=l
                    f <<- CoxRFX(allX[i,], allSurv[i,], allGroups, which.mu=which.
mu, sigma0=sigma0, nu=nu, weights=weights[i])
                    p <- as.matrix(allX[!i,,drop=FALSE]) %*% f$coefficients
                    r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(
f$coefficients), byrow=TRUE), linear.predictor=p)
                    colnames(r) <- c(names(f$coefficients), "linear.predictor")
                    as.data.frame(r)
                }, mc.cores=4))
looAllWeighted <- looAllWeighted[order(order(samples)),]
pp <- looAllWeighted$linear.predictor
survConcordance(allSurv ~ pp, weights=weights)
```

```
## Call:
## survConcordance(formula = allSurv ~ pp, weights = weights)
##
##   n= 950
## Concordance= 0.7561883 se= 0.02802535
## concordant discordant  tied.risk  tied.time   std(c-d)
##  6137610.4  1978900.7        0.0        1.0   454936.2
```

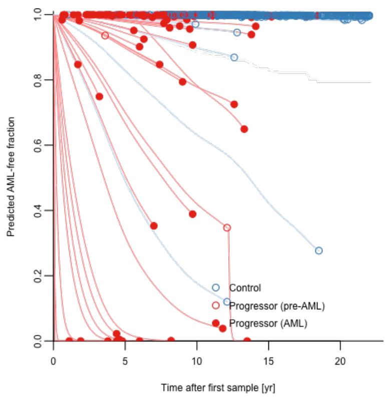
```
h <- exp(looAllWeighted$linear.predictor) > 100
plot(survfit(allSurv ~ h, weights=weights), col=set1[2:1], ylab="AML-free survival
", xlab="Time after first sample")
f <- sum(h*weights)/sum(weights) *100
legend("bottomleft", lty=1, col=set1[2:1], paste(c("low risk", "high risk"), "n ~"
, round(c( 100-f,f), 2),"%"))
```





7.4.9 Individual Predictions with corrected baseline

```
plot(survfit(allSurv~1), conf.int=FALSE, xlab='Time after first sample [yr]', ylab
='Predicted AML-free fraction', col='white', bty='L', yaxs='i', ylim=c(0,1.01))
d <- data.frame(t=NULL, s=NULL, pch=NULL, col=character())
for(i in unique(samples)){
    km <- exp(predict(smooth.spline(log(summary(survfit(allSurv[samples!=i,]~1, we
ights=weights[samples!=i]), times=t)$surv), df=10))$y)
    l0 <- colSums(fitAll$Z[samples!=i,,drop=FALSE] * weights[samples!=i]) %*% as.n
umeric(looAllWeighted[samples==i,][1,colnames(fitAll$Z)]) / sum(weights[samples!=i
])
    kmi <- function(km, s, lp, l0){
        .kmi <- function(km, sj, lpj, l0) km[t >= sj[,1] & t <= sj[,2]]^exp(lpj-l0
)
        k0 <- 1
        for(j in 1:nrow(s)) {
            k <- .kmi(km, s[j,], lp[j], l0)
            k <- k * k0/k[1]
            w <- t >= s[j,1] & t <= s[j,2]
            k0 <- k[length(k)]
            c <- if(s[nrow(s),3]==1) set1[1] else set1[2]
            lines(t[w], k, col=mg14:::colTrans(c), type='l')
            p <- if(s[j,3]==1) 19 else 1
            d <<- rbind(d, data.frame(t=t[w][length(k)], s=k[length(k)], pch=p, co
l=c))
        }
    }
    kmi(km, allSurv[samples==i,], looAllWeighted$linear.predictor[samples==i], l0)
}
points(d$t, d$s, pch=d$pch, col=as.character(d$col))
legend("bottomright", pch=c(1,1,19), col=c(set1[2], set1[1], set1[1]), legend=c("C
ontrol", "Progressor (pre-AML)", "Progressor (AML)"), bty='n')
```

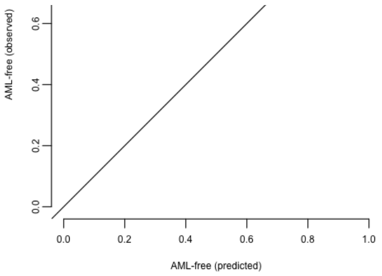


Callibration

```
p10 <- km[t==10]^exp(looAllWeighted$linear.predictor)
c <- cut(p10, c(0,0.4,0.95,0.99,1))
table(c)
```

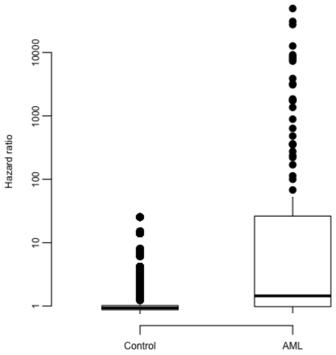
```
## c
##     (0,0.4]  (0.4,0.95] (0.95,0.99]    (0.99,1]
##          11          16          12         908
```

```
s <- summary(survfit(allSurv~c, weights=weights), times=10)
m <- sapply(split(p10,c), mean)
plot(m, s$surv, xlab="AML-free (predicted)", ylab="AML-free (observed)", xlim=c(0,
1), ylim=c(0,1))
segments(m,s$lower,m,s$upper)
abline(0,1)
```



Hazard

```
boxplot(exp(fitBoot$linear.predictors) ~ factor(1-control[b42], labels=c("Control"
,"AML")), log='y', ylab="Hazard ratio", pch=19, staplewex=0, lty=1, boxwex=0.5)
```



7.4.10 Some simulations

```
bX <- sapply(1:50, function(foo){
            set.seed(foo)
            X <- rbind(apply(allX[control,], 2, sample, n_total-sum(!control), rep
lace=TRUE), apply(allX[!control,], 2, sample) )
            lambda0 <- 5e-4
            r <- X%*%coef(fitBoot)
            t <- rexp(n_total, lambda0 * exp(r))
            tmax <- 13 + runif(n_total, 0,1)
            s <- Surv(pmin(t,tmax), t < tmax)
            cases <- which(s[,2]==1)
            controls1 <- sample(which(s[,2]==0), size=1*length(cases))
            controls4 <- sample(which(s[,2]==0), size=sum(control))
            cbind(controls_inc=colMeans(X[controls4,allGroups=="Genes"]>0), AML_in
c=colMeans(X[cases,allGroups=="Genes"]>0), controls_vaf=apply(X[controls4,allGroup
s=="Genes"], 2, function(x) mean(x[x>0])),AML_vaf=apply(X[cases,allGroups=="Genes"
], 2, function(x) mean(x[x>0])))
        }, simplify='array')
```

Expected vs observed driver frequency

```
graphics.off()
png("./figures/driver.freq.simulation.png", width = 15, height = 14, units = "cm",
res = 500)
par(mar = c(5, 4, 1.5, 0.5) + 0.1, mgp=c(2,0.4,0), las=1, tcl=-0.2, cex = 1)
plot(-rowMeans(bX[,'controls_inc',]), type='h', lend = 2, ylim=c(-.5,1)/2.5, lwd=8
, xaxt='n', yaxt = 'n',  ylab="Driver frequency (%)", xlab="", col=pal1[2])
atx <- axTicks(2)
axis(2,at=atx,labels= c(20, 10, 0, 10, 20, 30, 40))
points(x=1:22+.5,-colMeans(allX[control,allGroups=="Genes"]>0), type='h', lend = 2
, lwd=8, col=pal1[1])
points(rowMeans(bX[,"AML_inc",]), type='h', lend = 2, lwd=8, col=pal1[2])
points(x=1:22+.5,colMeans(allX[!control,allGroups=="Genes"]>0), type='h', lend = 2
, lwd=8, col=pal1[1])
mtext(side=1, at=1:22,sub("_.+","",colnames(allX)[allGroups=="Genes"]), las=2, fon
t=3, line=0.7)
```

```
mtext(text = "Pre-AML", side=3, at = 12, las=1, font=1, line=-1.5, cex = 1.1)
mtext(text = "Controls", side=1, at = 12, las=1, font=1, line=-1.5, cex = 1.1)
legend("bottomright", fill=pal1[2:1], c("Expected","Observed"), cex = 0.8)
abline(h=0)
dev.off()
```

```
## null device 
##           1
```

Expected vs observed driver VAF

```
avgVaf <- function(x) mean(x[x>0])

png("./figures/driver.vaf.simulation.png", width = 15, height = 14, units = "cm", 
res = 500)
par(mar = c(5, 4, 1.5, 0.5) + 0.1, mgp=c(2,0.4,0), las=1, tcl=-0.2, cex=1)
plot(-apply(bX[,'controls_vaf',],1,avgVaf)*10, type='h', lend = 2, ylim=c(-40,50), 
lwd=8, xaxt='n', yaxt = 'n', ylab="Driver VAF (%)", xlab="", col=pal1[2])
atx <- axTicks(2)
axis(2,at=atx,labels= c(40, 20,0, 20, 40))
points(x=1:22+.5,-apply(allX[control,allGroups=="Genes"],2,avgVaf)*10, type='h', l
end = 2, lwd=8, col=pal1[1])
points(apply(bX[,"AML_vaf",],1,avgVaf)*10, type='h', lend = 2, lwd=8, col=pal1[2])
points(x=1:22+.5,apply(allX[!control,allGroups=="Genes"],2,avgVaf)*10, type='h', l
end = 2, lwd=8, col=pal1[1])
mtext(side=1, at=1:22,sub("_.+","",colnames(allX)[allGroups=="Genes"]), las=2, fon
t=3, line = 0.7)
mtext(text = "Pre-AML", side=3, at = 12, las=1, font=1, line=-1.5, cex = 1.1)
mtext(text = "Controls", side=1, at = 12, las=1, font=1, line=-1.5, cex = 1.1)
legend("bottomright", fill=pal1[2:1], c("Expected","Observed"), cex = 0.8)
abline(h=0)
dev.off()
```

```
## pdf 
##   2
```

7.4.11 Simple models

```
samples <- factor(c(as.character(sangerData$Individual), as.character(torontoData$
Sample)))
```

max vaf:

```
v <- apply(allX[,allGroups=="Genes"], 1, max)*10
```

cumulative vaf

```
c <- apply(allX[,allGroups=="Genes"], 1, sum)*10
```

number of mutations

```
m <- rowSums(allX[,allGroups=="Genes"]>0)
```

any mutation

```
a <- as.integer(m>0)
```

7.4.11.1 Presence of any mutation

```
d <- data.frame(a)
summary(f <- coxph(allSurv ~ ., data=d ))
```

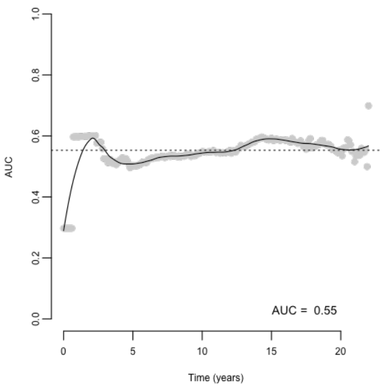
```
## Call:
## coxph(formula = allSurv ~ ., data = d)
## 
##   n= 950, number of events= 120 
## 
##     coef exp(coef) se(coef)     z Pr(>|z|)    
## a 1.5144    4.5468   0.2046 7.402 1.35e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##   exp(coef) exp(-coef) lower .95 upper .95
## a     4.547     0.2199     3.045      6.79
## 
## Concordance= 0.672  (se = 0.023 )
## Rsquare= 0.064   (max possible= 0.801 )
## Likelihood ratio test= 63.31  on 1 df,   p=2e-15
## Likelihood ratio test= 65.51  on 1 df,   p=2e-15
## Wald test            = 54.78  on 1 df,   p=1e-13
## Score (logrank) test = 66.02  on 1 df,   p=4e-16
```

```
los <- do.call("rbind",mclapply(levels(samples), function(l){
  i <- samples!=l
  f <<- coxph(allSurv ~ ., data=d, subset=i)
  p <- as.matrix(d[!i,]) %*% f$coefficients
  r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(f$coefficients), b
yrow=TRUE), linear.predictor=p)
  colnames(r) <- c(names(f$coefficients), "linear.predictor")
  as.data.frame(r)
}, mc.cores=4))
psAnyMt <- los[order(order(samples)),]

survConcordance(allSurv ~ psAnyMt$linear.predictor)
```

```
## Call:
## survConcordance(formula = allSurv ~ psAnyMt$linear.predictor)
##
##   n= 950
## Concordance= 0.5431925 se= 0.02388586
## concordant discordant  tied.risk  tied.time   std(c-d)
##  34829.000  28205.000  13646.000      1.000   3663.136
```

Dynamic/cumulative AUC

```
w <- c(which(allSurv[,1]==0)[-1]-1, nrow(allSurv))
survAll2 <- Surv(allSurv[w,2], allSurv[w,3])
t <- seq(0,22,0.1)
allX2 <- allX[w, ]

auc.uno <- AUC.uno(survAll2, survAll2, psAnyMt$linear.predictor[w], times=t)

plot(auc.uno$times, auc.uno$auc, xlab="Time (years)", ylab="AUC", pch=16, col="gre
y80", ylim = c(0,1.0))
lines(auc.uno$times, predict(loess(auc.uno$auc ~ auc.uno$times, span=0.25)))
abline(h=auc.uno$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(auc.uno$
iauc,2)))
```



```
AnyMt.a <- auc.uno
```

Presence of any mutation + vaf

Presence of any mutation + vaf

```
d <- data.frame(a,v)
summary(f <- coxph(allSurv ~ ., data=d ))
```

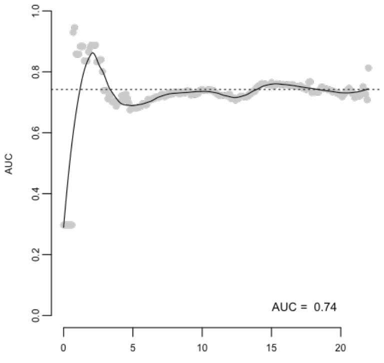
```
## Call:
## coxph(formula = allSurv ~ ., data = d)
## 
##   n= 950, number of events= 120 
## 
##       coef exp(coef) se(coef)     z Pr(>|z|)    
## a 1.025548  2.788622 0.223677 4.585 4.54e-06 ***
## v 0.050613  1.051915 0.005605 9.030  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##   exp(coef) exp(-coef) lower .95 upper .95
## a     2.789     0.3586     1.799     4.323
## v     1.052     0.9506     1.040     1.064
## 
## Concordance= 0.737  (se = 0.024 )
## Rsquare= 0.119   (max possible= 0.801 )
## Likelihood ratio test= 120.5  on 2 df,   p=<2e-16
## Wald test            = 161.8  on 2 df,   p=<2e-16
## Score (logrank) test = 263.9  on 2 df,   p=<2e-16
```

```
los <- do.call("rbind",mclapply(levels(samples), function(l){
  i <- samples!=l
  f <<- coxph(allSurv ~ ., data=d, subset=i)
  p <- as.matrix(d[!i,]) %*% f$coefficients
  r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(f$coefficients), b
yrow=TRUE), linear.predictor=p)
  colnames(r) <- c(names(f$coefficients), "linear.predictor")
  as.data.frame(r)
}, mc.cores=4))
psAnyMtVaf <- los[order(order(samples)),]

survConcordance(allSurv ~ psAnyMtVaf$linear.predictor)
```

```
## Call:
## survConcordance(formula = allSurv ~ psAnyMtVaf$linear.predictor)
## 
##   n= 950 
## Concordance= 0.7287559 se= 0.0238873
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  49091.000  14009.000  13580.000      1.000   3663.356
```

Dynamic/cumulative AUC

```
auc.uno <- AUC.uno(survAll2, survAll2, psAnyMtVaf$linear.predictor[w], times=t)

plot(auc.uno$times, auc.uno$auc, xlab="Time (years)", ylab="AUC", pch=16, col="gre
y80", ylim = c(0,1.0))
lines(auc.uno$times, predict(loess(auc.uno$auc ~ auc.uno$times, span=0.25)))
abline(h=auc.uno$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(auc.uno$
iauc,2)))
```



Time (years)

```
AnyMtVaf.a <- auc.uno
```

7.4.11.2 Number of mutations + vaf

```
d <- data.frame(m,v)
summary(f <- coxph(allSurv ~ ., data=d ))
```

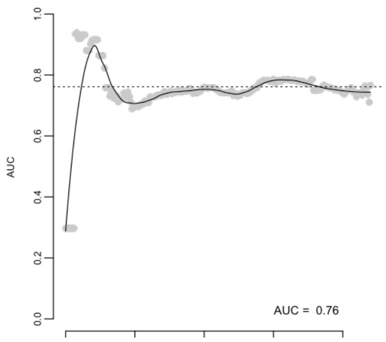
```
## Call:
## coxph(formula = allSurv ~ ., data = d)
## 
##   n= 950, number of events= 120 
## 
##       coef exp(coef) se(coef)     z Pr(>|z|)    
## m 0.653487  1.922231 0.088287 7.402 1.34e-13 ***
## v 0.040976  1.041827 0.006562 6.245 4.25e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##   exp(coef) exp(-coef) lower .95 upper .95
## m     1.922     0.5202     1.617     2.285
## v     1.042     0.9599     1.029     1.055
## 
## Concordance= 0.744  (se = 0.024 )
## Rsquare= 0.142   (max possible= 0.801 )
## Likelihood ratio test= 145.3  on 2 df,   p=<2e-16
## Wald test            = 213.3  on 2 df,   p=<2e-16
## Score (logrank) test = 302.9  on 2 df,   p=<2e-16
```

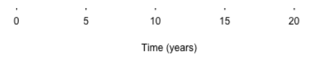
```
los <- do.call("rbind",mclapply(levels(samples), function(l){
  i <- samples!=l
  f <<- coxph(allSurv ~ ., data=d, subset=i)
  p <- as.matrix(d[!i,]) %*% f$coefficients
  r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(f$coefficients), b
yrow=TRUE), linear.predictor=p)
  colnames(r) <- c(names(f$coefficients), "linear.predictor")
  as.data.frame(r)
}, mc.cores=4))
psNMtVaf <- los[order(order(samples)),]

survConcordance(allSurv ~ psNMtVaf$linear.predictor)
```

```
## Call:
## survConcordance(formula = allSurv ~ psNMtVaf$linear.predictor)
## 
##   n= 950 
## Concordance= 0.7431403 se= 0.0238873
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  50194.000  12906.000  13580.000      1.000   3663.356
```

Dynamic/cumulative AUC

```
auc.uno <- AUC.uno(survAll2, survAll2, psNMtVaf$linear.predictor[w], times=t)

plot(auc.uno$times, auc.uno$auc, xlab="Time (years)", ylab="AUC", pch=16, col="gre
y80", ylim = c(0,1.0))
lines(auc.uno$times, predict(loess(auc.uno$auc ~ auc.uno$times, span=0.25)))
abline(h=auc.uno$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(auc.uno$
iauc,2)))
```





```
NMtVaf.a <- auc.uno
```

7.4.11.3 Number of mutations + cumulative vaf

```
d <- data.frame(m,c)
summary(f <- coxph(allSurv ~ ., data=d ))
```

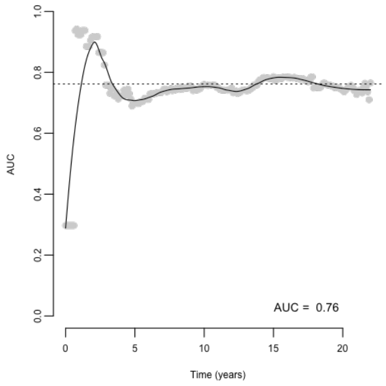
```
## Call:
## coxph(formula = allSurv ~ ., data = d)
## 
##   n= 950, number of events= 120 
## 
##       coef exp(coef) se(coef)     z Pr(>|z|)    
## m 0.613264  1.846449 0.090393 6.784 1.17e-11 ***
## c 0.033648  1.034220 0.005036 6.681 2.38e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##   exp(coef) exp(-coef) lower .95 upper .95
## m     1.846     0.5416     1.547     2.204
## c     1.034     0.9669     1.024     1.044
## 
## Concordance= 0.744  (se = 0.024 )
## Rsquare= 0.144   (max possible= 0.801 )
## Likelihood ratio test= 148.2  on 2 df,   p=<2e-16
## Wald test            = 223.3  on 2 df,   p=<2e-16
## Score (logrank) test = 350.7  on 2 df,   p=<2e-16
```

```
los <- do.call("rbind",mclapply(levels(samples), function(l){
  i <- samples!=l
  f <<- coxph(allSurv ~ ., data=d, subset=i)
  p <- as.matrix(d[!i,]) %*% f$coefficients
  r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(f$coefficients), b
yrow=TRUE), linear.predictor=p)
  colnames(r) <- c(names(f$coefficients), "linear.predictor")
  as.data.frame(r)
}, mc.cores=4))
psNMtCumVaf <- los[order(order(samples)),]

survConcordance(allSurv ~ psNMtCumVaf$linear.predictor)
```

```
## Call:
## survConcordance(formula = allSurv ~ psNMtCumVaf$linear.predictor)
## 
##   n= 950 
## Concordance= 0.743362 se= 0.0238873
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  50211.000  12889.000  13580.000      1.000   3663.356
```

Dynamic/cumulative AUC

```
auc.uno <- AUC.uno(survAll2, survAll2, psNMtCumVaf$linear.predictor[w], times=t)

plot(auc.uno$times, auc.uno$auc, xlab="Time (years)", ylab="AUC", pch=16, col="gre
y80", ylim = c(0,1.0))
lines(auc.uno$times, predict(loess(auc.uno$auc ~ auc.uno$times, span=0.25)))
abline(h=auc.uno$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(auc.uno$
iauc,2)))
```



```
NMtCumVaf.a <- auc.uno
```

Gene-level risks

```
d <- allX
summary(f <- coxph(allSurv ~ ., data=d))
```

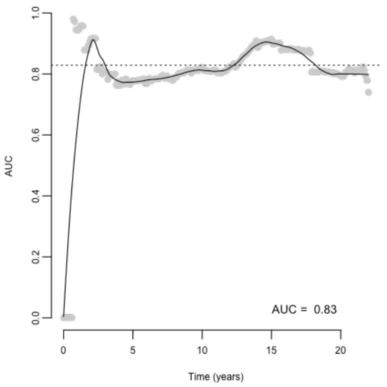
```
## Call:
## coxph(formula = allSurv ~ ., data = d)
## 
##   n= 950, number of events= 120 
## 
##                  coef  exp(coef)   se(coef)      z Pr(>|z|)    
## ASXL1_0.1     0.45410    1.57475    0.25483  1.782   0.0748 .  
## BCOR_0.1      4.53517   93.23942   15.29850  0.296   0.7669    
## CBL_0.1       0.02418    1.02448    0.74288  0.033   0.9740    
## DNMT3A_0.1    0.13468    1.14417    0.18286  0.737   0.4614    
## IDH1_0.1      0.39412    1.48307    0.63231  0.623   0.5331    
## IDH2_0.1      0.51163    1.66800    0.29079  1.759   0.0785 .  
## JAK2_0.1      0.59064    1.80514    0.39331  1.502   0.1332    
## KDM6A_0.1     0.15988    1.17337   32.12704  0.005   0.9960    
## KMT2C_0.1    -0.50258    0.60497    1.77003 -0.284   0.7765    
## KMT2D_0.1    -0.01333    0.98676    0.58364 -0.023   0.9818    
## KRAS_0.1      0.54336    1.72178   12.36468  0.044   0.9649    
## NF1_0.1      -0.76668    0.46455    5.94275 -0.129   0.8973    
## NRAS_0.1      7.40428 1643.00852    6.01855  1.230   0.2186    
## PHF6_0.1      4.31340   74.69375   15.42773  0.280   0.7798    
## PTPN11_0.1    4.49429   89.50474    6.18432  0.727   0.4674    
## RAD21_0.1     0.07319    1.07594    6.89358  0.011   0.9915    
## RUNX1_0.1     0.17980    1.19698    0.24611  0.731   0.4650    
## SF3B1_0.1     1.10331    3.01414    0.52063  2.119   0.0341 *  
## SRSF2_0.1     0.34535    1.41248    0.21771  1.586   0.1127    
## TET2_0.1      0.17179    1.18743    0.20206  0.850   0.3952    
## TP53_0.1      2.17381    8.79176    0.55321  3.929 8.51e-05 ***
## U2AF1_0.1     2.74012   15.48884    0.35246  7.774 7.58e-15 ***
## age_10       -0.01189    0.98818    0.10907 -0.109   0.9132    
## gender       -0.01138    0.98868    0.19862 -0.057   0.9543    
## cohort       -0.13561    0.87318    0.23791 -0.570   0.5687    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##            exp(coef) exp(-coef) lower .95 upper .95
## ASXL1_0.1     1.5747  0.6350222 9.557e-01 2.595e+00
## BCOR_0.1     93.2394  0.0107251 8.861e-12 9.811e+14
## CBL_0.1       1.0245  0.9761095 2.389e-01 4.394e+00
## DNMT3A_0.1    1.1442  0.8739972 7.995e-01 1.637e+00
## IDH1_0.1      1.4831  0.6742750 4.295e-01 5.121e+00
## IDH2_0.1      1.6680  0.5995195 9.434e-01 2.949e+00
## JAK2_0.1      1.8051  0.5539734 8.351e-01 3.902e+00
## KDM6A_0.1     1.1734  0.8522477 5.283e-28 2.606e+27
## KMT2C_0.1     0.6050  1.6529815 1.884e-02 1.943e+01
## KMT2D_0.1     0.9868  1.0134221 3.144e-01 3.097e+00
## KRAS_0.1      1.7218  0.5807959 5.142e-11 5.765e+10
## NF1_0.1       0.4646  2.1526020 4.060e-06 5.315e+04
## NRAS_0.1   1643.0085  0.0006086 1.238e-02 2.181e+08
## PHF6_0.1     74.6937  0.0133880 5.510e-12 1.012e+15
## PTPN11_0.1   89.5047  0.0111726 4.872e-04 1.644e+07
## RAD21_0.1     1.0759  0.9294227 1.459e-06 7.936e+05
## RUNX1_0.1     1.1970  0.8354364 7.389e-01 1.939e+00
## SF3B1_0.1     3.0141  0.3317696 1.086e+00 8.362e+00
## SRSF2_0.1     1.4125  0.7079756 9.219e-01 2.164e+00
## TET2_0.1      1.1874  0.8421566 7.991e-01 1.764e+00
## TP53_0.1      8.7918  0.1137429 2.973e+00 2.600e+01
## U2AF1_0.1    15.4888  0.0645626 7.763e+00 3.091e+01
## age_10        0.9882  1.0119578 7.980e-01 1.224e+00
## gender        0.9887  1.0114489 6.699e-01 1.459e+00
## cohort        0.8732  1.1452345 5.478e-01 1.392e+00
##
## Concordance= 0.81  (se = 0.027 )
## Rsquare= 0.069   (max possible= 0.801 )
## Likelihood ratio test= 67.53  on 25 df,   p=9e-06
## Wald test            = 110.8  on 25 df,   p=9e-13
## Score (logrank) test = 782.6  on 25 df,   p=<2e-16
```

```
los <- do.call("rbind",mclapply(levels(samples), function(l){
  i <- samples!=l
  f <<- coxph(allSurv ~ ., data=d, subset=i)
  p <- as.matrix(d[!i,]) %*% f$coefficients
  r <- cbind(matrix(f$coefficients, nrow=length(p), ncol=length(f$coefficients), b
yrow=TRUE), linear.predictor=p)
  colnames(r) <- c(names(f$coefficients), "linear.predictor")
  as.data.frame(r)
}, mc.cores=4))
psGenes <- los[order(order(samples)),]

survConcordance(allSurv ~ psGenes$linear.predictor)
```

```
## Call:
## survConcordance(formula = allSurv ~ psGenes$linear.predictor)
##
##   n= 950
## Concordance= 0.7799296 se= 0.02746327
## concordant discordant  tied.risk  tied.time   std(c-d)
##  59805.000  16875.000      0.000      1.000   4211.768
```

Dynamic/cumulative AUC

```
auc.uno <- AUC.uno(survAll2, survAll2, psGenes$linear.predictor[w], times=t)

plot(auc.uno$times, auc.uno$auc, xlab="Time (years)", ylab="AUC", pch=16, col="gre
y80", ylim = c(0,1.0))
lines(auc.uno$times, predict(loess(auc.uno$auc ~ auc.uno$times, span=0.25)))
abline(h=auc.uno$iauc, lty = 3, lwd = 1)
legend("bottomright", bty = "n", cex = 1.2, legend = paste("AUC = ",round(auc.uno$
iauc,2)))
```



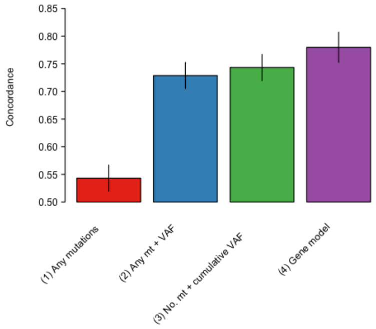
```
Genes.a <- auc.uno
```

```
# Concordance summary
c <- rbind(
  `(1) Any mutations`=as.data.frame(survConcordance(allSurv ~ psAnyMt$linear.predi
ctor)[c("concordance","std.err")]),
  `(2) Any mt + VAF`=as.data.frame(survConcordance(allSurv ~ psAnyMtVaf$linear.pre
dictor)[c("concordance","std.err")]),
  `(3) No. mt + cumulative VAF`=as.data.frame(survConcordance(allSurv ~ psNMtCumVa
f$linear.predictor)[c("concordance","std.err")]),
  `(4) Gene model`=as.data.frame(survConcordance(allSurv ~ psGenes$linear.predicto
r)[c("concordance","std.err")]))

c
```

	concordance <dbl>	std.err <dbl>
(1) Any mutations	0.5431925	0.02388586
(2) Any mt + VAF	0.7287559	0.02388730
(3) No. mt + cumulative VAF	0.7433620	0.02388730
(4) Gene model	0.7799296	0.02746327

4 rows

```
set1 <- RColorBrewer::brewer.pal(6,"Set1")

par(mar = c(9, 4, 1.5, 0.5) + 0.1, mgp=c(2.7,0.4,0), las=1, tcl=-0.2)
b <- barplot(c$concordance-0.5, ylab="Concordance", col=set1, ylim=c(0.5,0.88), of
fset=0.5)
mg14::rotatedLabel(x=b, labels=rownames(c))
segments(b,c$concordance+c$std.err,b,c$concordance-c$std.err)
```



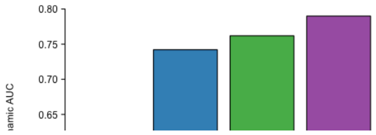
Dynamic/cumulative AUC summary

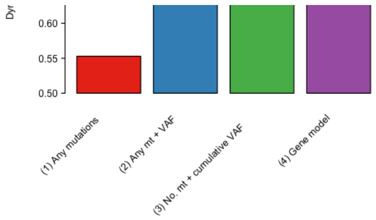
```
d.auc <- data.frame(iauc = c(AnyMt.a$iauc, AnyMtVaf.a$iauc, NMtCumVaf.a$iauc, 0.79
))
rownames(d.auc) <- c("(1) Any mutations", "(2) Any mt + VAF", "(3) No. mt + cumula
tive VAF", "(4) Gene model")

d.auc
```

	iauc <dbl>
(1) Any mutations	0.5528776
(2) Any mt + VAF	0.7420613
(3) No. mt + cumulative VAF	0.7618961
(4) Gene model	0.7900000

4 rows

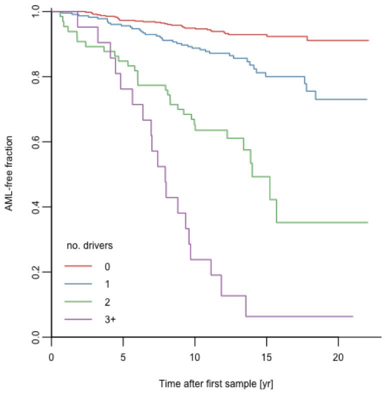
```
par(mar = c(9, 4, 1.5, 0.5) + 0.1, mgp=c(2.7,0.4,0), las=1, tcl=-0.2)
b <- barplot(d.auc$iauc-0.5, ylab="Dynamic AUC", col=set1, ylim=c(0.5,0.80), offse
t=0.5)
mg14::rotatedLabel(x=b, labels=rownames(d.auc))
```





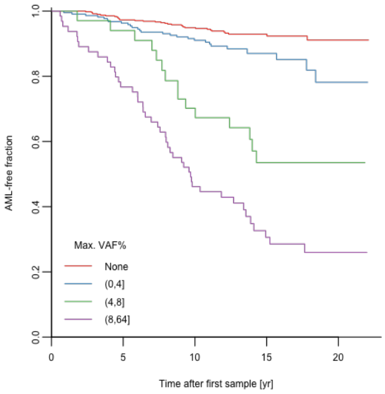
AML-free survival by number of drivers

```
nonc <- rowSums(allX[,allGroups=="Genes"]>0)
nonc <- cut(nonc, c(-1,0,1,2,max(nonc)))
plot(survfit(allSurv~nonc), col=set1, xlab='Time after first sample [yr]', ylab='A
ML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01))
legend("bottomleft", c(0,1,2,"3+"), col=set1, lty=1, bty='n', title="no. drivers")
```



AML-free survival by max VAF

```
mvaf <- apply(allX[,allGroups=="Genes"], 1, max)*10
mvaf <- cut(mvaf, c(-1,0,4,8,max(mvaf)))
plot(survfit(allSurv~mvaf), col=set1, xlab='Time after first sample [yr]', ylab='A
ML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01))
levels(mvaf)[1] <- "None"
legend("bottomleft", levels(mvaf), col=set1, lty=1, bty='n', title="Max. VAF%")
```



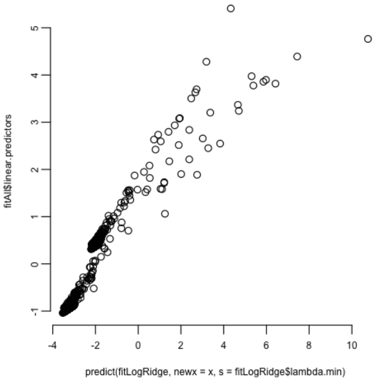
8 Logistic regression

```
library(glmnet)
library(ROCR)
```

8.1 Combined

```
set.seed(42)
y <- allSurv[,3]
x <- allX
x <- as.matrix(cbind(x, mu.Genes=rowSums(x[,allGroups=="Genes"])))
fitLogRidge <- cv.glmnet(x, y, alpha=0, standardize=FALSE, penalty.factor=c(allGro
ups=="Genes",FALSE), family="binomial", lambda=10^seq(-5,5,0.1)/nrow(x))
```

```
fitLog <- glm(y ~ x[,-ncol(x)], family="binomial")
coefLogRidge <- coef(fitLogRidge, s=fitLogRidge$lambda.min)[-1,1]
w <- names(coefLogRidge) %in% colnames(allX)[allGroups=="Genes"]
coefLogRidge[w] <- coefLogRidge[w] + coefLogRidge["mu.Genes"]
names(coefLogRidge) <- colnames(x)
s <- summary(survfit(allSurv ~1))

plot(predict(fitLogRidge, newx=x, s=fitLogRidge$lambda.min),fitAll$linear.predicto
rs)
```



```
cor(predict(fitLogRidge, newx=x, s=fitLogRidge$lambda.min),fitAll$linear.predictor
s)
```

```
##        [,1]
## 1 0.9325608
```

8.2 Discovery cohort

```
set.seed(42)
x <- cbind(as.matrix(torontoX), mu.Genes=rowSums(torontoX[torontoGroups=="Genes"])
)
fitLogRidgeToronto <- cv.glmnet(x, torontoSurv[,2], alpha=0, standardize=FALSE, pe
nalty.factor=c(torontoGroups=="Genes",FALSE), family="binomial", lambda=10^seq(-5,
5,0.1)/nrow(x))
l <- max(which(abs(fitLogRidgeToronto$cvm- min(fitLogRidgeToronto$cvm)) < 0.01))
coefFitLogRidgeToronto <- coef(fitLogRidgeToronto, s=fitLogRidge$lambda.min *nrow(
allX)/nrow(torontoX))[-1,1]
w <- names(coefFitLogRidgeToronto) %in% colnames(torontoX)[torontoGroups=="Genes"]
coefFitLogRidgeToronto[w] <- coefFitLogRidgeToronto[w] + coefFitLogRidgeToronto["m
u.Genes"]
```

8.3 Validation cohort

```
set.seed(42)
x <- cbind(as.matrix(sangerX), mu.Genes=rowSums(sangerX[sangerGroups=="Genes"]))
y <- sangerSurv[,3]
fitLogRidgeSanger <- glmnet(x, y, alpha=0, standardize=FALSE, penalty.factor=c(san
gerGroups%in%c("Genes","Blood"),1e-2) , family="binomial",lambda=10^seq(-5,5,0.1)/
nrow(x))
coefFitLogRidgeSanger <- coef(fitLogRidgeSanger, s=fitLogRidge$lambda.min*nrow(all
X)/nrow(sangerX)/4)[-1,1]
w <- names(coefFitLogRidgeSanger) %in% colnames(sangerX)[sangerGroups=="Genes"]
coefFitLogRidgeSanger[w] <- coefFitLogRidgeSanger[w] + coefFitLogRidgeSanger["mu.G
enes"]
coefFitLogRidgeSanger
```

```
##    ASXL1_0.1     CBL_0.1  DNMT3A_0.1    JAK2_0.1   KMT2C_0.1   KMT2D_0.1     KRAS
_0.1      NF1_0.1     NRAS_0.1   RAD21_0.1
##   1.61735484  0.62402794  0.60690505  1.21223108  1.28664688  0.38990853  1.3057
9768  1.05008349  1.12131863  1.08384807
##    SF3B1_0.1   SRSF2_0.1    TET2_0.1    TP53_0.1   U2AF1_0.1      age_10       ge
```

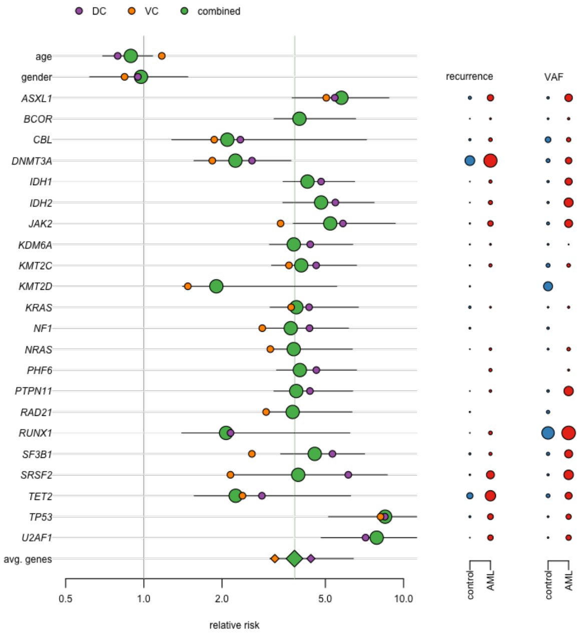
```
nder  systol_100 diastol_100      bmi_10
##  0.95795153  0.76775960  0.87432787  2.09849607  2.46513749  0.15915519 -0.1710
4884 -0.26674155  0.40623412  0.78151214
## cholestl_10     triglyc         hdl         ldl         lym     mcv_100      rd
w_10      wbc_10     plt_100      hgb_10
##  0.02221735 -0.02231645 -0.60655423  0.08051073  0.02388812 -0.48424380  1.4392
5261 -0.13343432  0.28531137  0.80105113
##    mu.Genes
##  1.16143798
```

8.4 Bootstrap CIs

```
coefLogRidgeBoot <- sapply(1:100, function(foo){
            set.seed(foo)
            y <- allSurv[,3]
            x <- allX
            x <- as.matrix(cbind(x, mu.Genes=rowSums(x[,allGroups=="Genes"])))
            b <- sample(1:nrow(x), replace=TRUE)
            fitLogRidgeBoot <- glmnet(x[b,], y[b], alpha=0, standardize=FALSE, pen
alty.factor=c(allGroups=="Genes",FALSE, FALSE), family="binomial", lambda=10^seq(-
5,5,0.1)/nrow(x))
            coefLogRidgeBoot <- coef(fitLogRidgeBoot, s=fitLogRidge$lambda.min)[-1
,1]
            w <- names(coefLogRidgeBoot) %in% colnames(allX)[allGroups=="Genes"]
            coefLogRidgeBoot[w] <- coefLogRidgeBoot[w] + coefLogRidgeBoot["mu.Gene
s"]
            names(coefLogRidgeBoot) <- colnames(x)
            coefLogRidgeBoot
        })
```

8.5 Forest plot

```
par(bty="n", mar=c(3,6,3,10)+.5, mgp=c(2,0.5,0), xpd=FALSE)
c <- exp(coefLogRidge[-25])
o <- c(23:24,1:22,25)
ci <- apply(coefLogRidgeBoot,1,quantile, c(0.025,0.975))[,-25]
y <- rev(seq_along(c))
plot(c[o], y, xlab="relative risk", log='x', ylab='', yaxt="n", pch=NA, xlim=c(0.5
,10))
abline(h=y, col="#EEEEEE", lty=1)
abline(v=1, lty=1, col="grey")
abline(v=c["mu.Genes"], col=mg14::colTrans(set1[3]), lty=1)
segments(exp(ci[1,o]), y, exp(ci[2,o]), y)
segments(exp(ci[1,o]), y, exp(ci[2,o]),y)
points(c[o], y, xlab="relative risk",  bg=set1[3], cex=2, pch=c(rep(21,24), 23))
m <- match(names(c)[o],names(coefFitLogRidgeToronto))
points(exp(coefFitLogRidgeToronto[m]), y,bg=set1[4], pch=c(rep(21,24), 23), cex=1)
m <- match(names(c)[o],names(coefFitLogRidgeSanger))
points(exp(coefFitLogRidgeSanger[m]), y,bg=set1[5], pch=c(rep(21,24), 23), cex=1)
mtext(side=2, sub("mu.Genes","avg. genes",sub("_.+","",names(c)[o])), at=y, las=2,
font=c(1,1,rep(3,22),1))

r <- sapply(split(as.data.frame(allX>0), control), colMeans)
f <- sapply(split(allX, control), apply, 2, function(x) mean(x[x>0]))
par(xpd=NA)
points(rep(18,22),y[3:24], cex=sqrt(r[o[3:24],2]*10), pch=21, bg=set1[2])
points(rep(18*1.2,22), y[3:24], cex=sqrt(r[o[3:24],1]*10), pch=21, bg=set1[1])
points(rep(36,22),y[3:24], cex=sqrt(f[o[3:24],2]), pch=21, bg=set1[2])
points(rep(36*1.2,22), y[3:24], cex=sqrt(f[o[3:24],1]), pch=21, bg=set1[1])
legend(x=0.5, y=28, pch=21, pt.bg=set1[c(4,5,3)], c("DC","VC","combined"), bty="n"
, ncol=3, text.width=0.1)

text(y=24, x=18, "recurrence")
text(y=24, x=38, "VAF")

axis(1, at=c(18,18*1.2), c("control","AML"), las=2, line=-1)
axis(1, at=c(36,36*1.2), c("control","AML"), las=2, line=-1)
```



8.6 AUC

```
aucLogRidgeBoot <- t(sapply(1:100, function(foo){
                    set.seed(foo)
                    y <- allSurv[,3]
                    x <- allX
                    x <- as.matrix(cbind(x, mu.Genes=rowSums(x[,allGroups=="Genes"
])))
                    b <- sample(1:nrow(x), replace=TRUE)
                    oob <- setdiff(1:nrow(x),b)
                    c(inb=performance(prediction(x[b,] %*% coefLogRidgeBoot[,foo],
y[b]), "auc")@y.values[[1]]
y[b]), "auc")@y.values[[1]],
                            oob=performance(prediction(x[oob,] %*% coefLogRidgeBoo
t[,foo], y[oob]),"auc")@y.values[[1]])
                }))

apply(aucLogRidgeBoot, 2, quantile)
```

```
##            inb       oob
## 0%   0.7600825 0.7331746
## 25%  0.7981192 0.7814137
## 50%  0.8107881 0.8058353
## 75%  0.8228798 0.8254089
## 100% 0.8616209 0.8650056
```

```
performance(prediction(as.matrix(torontoX) %*% coefFitLogRidgeToronto[-22], toront
oSurv[,2]),"auc")@y.values[[1]]
```

```
## [1] 0.7649573
```

```
performance(prediction(as.matrix(sangerImp) %*% coefFitLogRidgeToronto[-22], sange
rSurv[,3]),"auc")@y.values[[1]]
```

```
## [1] 0.806366
```

```
performance(prediction(as.matrix(sangerX) %*% coefFitLogRidgeSanger[-31], sangerSu
rv[,3]),"auc")@y.values[[1]]
```

```
## [1] 0.8479775
```

```
performance(prediction(ImputeMissing(sangerX, as.matrix(torontoImp)) %*% coefFitLo
gRidgeSanger[-31], torontoSurv[,2]),"auc")@y.values[[1]]
```

```
## [1] 0.6885916
```

9 Tabulate results

```
# library(xlsx)
# wb <- createWorkbook("xlsx")
# sheet  <- createSheet(wb, sheetName="Cox PH adjusted (combined)")
# addDataFrame(waldWeighted,
#       sheet,
#       colnamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE) + Border(),
#       rownamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE)
# )
# sheet  <- createSheet(wb, sheetName="Cox PH adjusted (DC)")
# addDataFrame(waldWeightedToronto,
#       sheet,
#       colnamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE) + Border(),
#       rownamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE)
# )
#
```

```
# sheet  <- createSheet(wb, sheetName="Cox PH adjusted (VC)")
# addDataFrame(waldWeightedSanger,
#       sheet,
#       colnamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE) + Border(),
#       rownamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE)
# )
#
# sheet  <- createSheet(wb, sheetName="Logistic regression (combined)")
# addDataFrame(data.frame(`Coef combined`=coefLogRidge, CI=t(apply(coefLogRidgeBoo
t, 1, quantile, c(0.025,0.975))),
#               check.names=FALSE),
#       sheet,
#       colnamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE) + Border(),
#       rownamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE)
# )
#
# sheet  <- createSheet(wb, sheetName="Logistic regression (DC)")
# addDataFrame(data.frame(`Coef combined`=coefFitLogRidgeToronto,
#               check.names=FALSE),
#       sheet,
#       colnamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE) + Border(),
#       rownamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE)
# )
# sheet  <- createSheet(wb, sheetName="Logistic regression (Sanger)")
# addDataFrame(data.frame(`Coef combined`=coefFitLogRidgeSanger,
#               check.names=FALSE),
#       sheet,
#       colnamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE) + Border(),
#       rownamesStyle = CellStyle(wb) + Font(wb, isBold=TRUE)
# )
# saveWorkbook(wb, file="SupplementaryTables.xlsx")
```

10 Clinical/Demographic model

Necessary to reconstruct matrices and survival objects to use data from VC for all 8 samples sequenced in both cohorts ## Discovery cohort Data 83 pre-AML (keeping duplicates with validation cohort)

```
f = "data/DC_vaf_matrix_no_duplicates_414ctrl_83aml.csv"
torontoData <- read.csv(f)

torontoData$gender <- ifelse(torontoData$Sex == "male", 1,
                             ifelse(torontoData$Sex == "female", 0, torontoData$Se
x))
table(torontoData$gender)
```

```
##
##   0   1
## 293 204
```

```
torontoData$gender <- as.numeric(torontoData$gender)
colnames(torontoData)
```

```
##  [1] "Sample"     "ASXL1"      "BCOR"       "CALR"       "CBL"        "DNMT3A"
"IDH1"       "IDH2"
##  [9] "JAK2"       "KDM6A"      "KIT"        "KMT2C"      "KRAS"       "NF1"
"NRAS"       "PHF6"
## [17] "PTPN11"     "RUNX1"      "SF3B1"      "SRSF2"      "TET2"       "TP53"
"U2AF1"      "Diagnosis"
## [25] "fu_years"   "age"        "Sex"        "no_drivers" "gender"
```

Manually standardize magnitudes

```
torontoData <- torontoData[!duplicated(torontoData),]

gene_vars <- c("CALR", "NRAS", "DNMT3A", "SF3B1", "IDH1", "KIT", "TET2", "RAD21",
"JAK2", "CBL", "KRAS", "PTPN11", "IDH2", "TP53", "NF1", "SRSF2", "CEBPA", "ASXL1",
"RUNX1", "U2AF1", "BCOR", "KDM6A", "PHF6", "KMT2C", "KMT2D")

torontoX <- torontoData[, colnames(torontoData) %in% c(gene_vars, "age", "gender")
]

torontoX <- as.data.frame(torontoX)
```

Only include genes in model if mutated in >2 samples

```
thr <- 2
torontoX <- torontoX[,colSums(torontoX != 0)>=thr]
```

```
torontoGroups <- factor(names(torontoX) %in% c("age","gender")+1, level=1:2, label
s=c("Genes","Demographics"))
colnames(torontoX)
```

```
##  [1] "ASXL1"  "CALR"   "CBL"    "DNMT3A" "IDH1"   "IDH2"   "JAK2"   "KDM6A"  "K
MT2C"  "KRAS"   "NF1"    "PHF6"  
## [13] "PTPN11" "RUNX1"  "SF3B1"  "SRSF2"  "TET2"   "TP53"   "U2AF1"  "age"    "g
ender"
```

```
torontoGroups
```

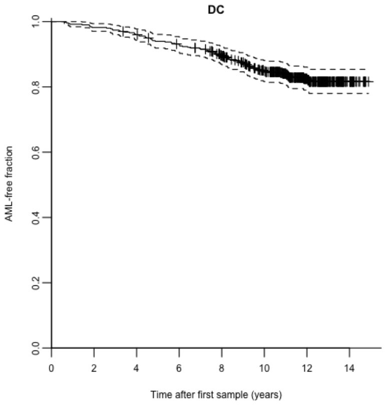
```
##  [1] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
##  [9] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
## [17] Genes        Genes        Genes        Demographics Demographics
## Levels: Genes Demographics
```

Manually standardize age and mutation VAFs

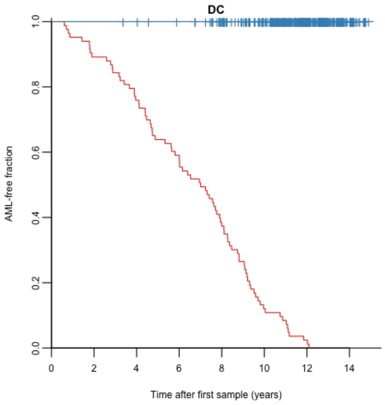
```
torontoX$age <- torontoX$age/10
names(torontoX)[which(names(torontoX)=="age")] <- "age_10"
g <- torontoGroups == "Genes"
torontoX[,g] <- torontoX[,g]*10
names(torontoX)[g] <- paste(names(torontoX)[g], "0.1",sep="_")
colnames(torontoX)
```

```
##  [1] "ASXL1_0.1"  "CALR_0.1"   "CBL_0.1"    "DNMT3A_0.1" "IDH1_0.1"   "IDH2_0.1
"   "JAK2_0.1"   "KDM6A_0.1" 
##  [9] "KMT2C_0.1"  "KRAS_0.1"   "NF1_0.1"    "PHF6_0.1"   "PTPN11_0.1" "RUNX1_0.
1"  "SF3B1_0.1"  "SRSF2_0.1" 
## [17] "TET2_0.1"   "TP53_0.1"   "U2AF1_0.1"  "age_10"     "gender"
```

```
torontoSurv <- Surv(torontoData$fu_years, torontoData$Diagnosis=="AML")
plot(survfit(torontoSurv~ 1), col= "black", main = "DC", xlab='Time after first sa
mple (years)', ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.t
ime = T)
```



```
plot(survfit(torontoSurv ~ torontoData$Diagnosis), xlab='Time after first sample (
years)', main = "DC", ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01),
mark.time = T, col = set1[1:2])
```



10.1 Validation cohort

all 37 pre-AML samples including overlap with DC

```
f = "data/VC_vaf_matrix_262ctrl_37aml_nodates.csv"
sangerData <- read.csv(f)

sangerData$hcdate <- as.Date(sangerData$hcdate)
sangerData$dodx <- as.Date(sangerData$dodx)

sangerPatients <- sub("[a-z]+$","", sangerData$Sample)
o <- order(sangerPatients, as.numeric(sangerData$hcdate))

sangerData <- sangerData[o,]
sangerPatients <- sangerPatients[o]

clinical_vars <- c("systol", "diastol", "bmi", "cholestl", "triglyc", "hdl", "ldl"
, "lym", "mcv", "rdw", "wbc", "plt", "hgb")
sangerX <- sangerData[, colnames(sangerData) %in% c(gene_vars, "age","gender",clin
ical_vars)]
sangerX <- as.data.frame(sangerX)

sangerX <- sangerX[,colSums(sangerX != 0,na.rm=TRUE)>=thr]
sangerGroups <- factor(grepl("^[a-z]", colnames(sangerX))*2, levels=0:2, labels=c(
"Genes", "Demographics", "Blood"))
sangerGroups[names(sangerX) %in% c("age","gender")] <- "Demographics"
table(sangerGroups)
```

```
## sangerGroups
##        Genes Demographics        Blood 
##           15            2           13
```

```
colnames(sangerX)
```

```
##  [1] "ASXL1"    "CBL"      "DNMT3A"   "JAK2"     "KMT2C"    "KMT2D"    "KRAS"    
"NF1"      "NRAS"     "RAD21"   
## [11] "SF3B1"    "SRSF2"    "TET2"     "TP53"     "U2AF1"    "age"      "gender"  
"systol"   "diastol"  "bmi"     
## [21] "cholestl" "triglyc"  "hdl"      "ldl"      "lym"      "mcv"      "rdw"     
"wbc"      "plt"      "hgb"
```

```
sangerGroups
```

```
##  [1] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
##  [9] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Demographics
## [17] Demographics Blood        Blood        Blood        Blood        Blood       
Blood        Blood       
## [25] Blood        Blood        Blood        Blood        Blood        Blood       
## Levels: Genes Demographics Blood
```

```
poorMansImpute <- function(x) {x[is.na(x)] <- mean(x, na.rm=TRUE); return(x)}
sangerX <- as.data.frame(sapply(sangerX, poorMansImpute))

foo <- split(sangerData[,c("Diagnosis","hcdate","dodx")], sangerPatients)

bar <- do.call("rbind",lapply(foo, function(x){
  y <- x
  n <- nrow(y)
  y[-n,"Diagnosis"] <- "Control"
  start <- as.numeric(y$hcdate - y$hcdate[1])/365.25
  end <- c(as.numeric(y$hcdate - y$hcdate[1])[-1]/365.25, as.numeric(y$dodx[n] - y
$hcdate[1])/365.25)
  return(data.frame(Diagnosis=y[,"Diagnosis"], start=start, end=end))
}))

bar[1:10, ]
```

	Diagnosis	start	end
	<fctr>	<dbl>	<dbl>
PD29762	AML	0.000000	9.754962
PD29764	AML	0.000000	10.360027
PD29792	AML	0.000000	14.108145
PD29804	Control	0.000000	5.138946
PD29810	Control	0.000000	18.573580

PD29810	Control	0.000000	18.573580
PD29836.1	Control	0.000000	2.414784
PD29836.2	AML	2.414784	10.023272
PD29851.1	Control	0.000000	4.599589
PD29851.2	AML	4.599589	12.205339
PD29856.1	Control	0.000000	4.331280

1-10 of 10 rows

```
sangerPatientsSplit <- unlist(sapply(names(foo), function(n) rep(n, nrow(foo[[n]])
)))

sangerSurv <- Surv(time = bar$start, time2 = bar$end, event = bar$Diagnosis!="Cont
rol", origin = 0)

plot(survfit(sangerSurv~ 1), col= "black", main = "VC", xlab='Time after first sam
ple (years)', ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.ti
me = T) #mark = 1
```

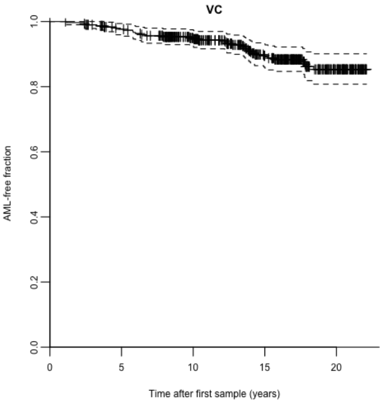


Figure 3 c-e

```
summary(sangerX$rdw)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   11.40   13.10   13.42   13.42   13.42   22.00
```

```
rdw <- cut(sangerX$rdw, c(11, 14, max(sangerX$rdw)))
levels(rdw) <- c("11-14", "14+")
table(rdw)
```

```
## rdw
## 11-14   14+ 
##   400    59
```

```
selected_genes <- c("DNMT3A", "TET2", "TP53", "U2AF1")

png("./figures/CombinedCohorts.KM.selected.genes.png", width = 8.5, height = 17.5,
units = "cm", res = 800)
par(mfrow=c(4,2), mar = c(1.9, 1.9, 1.7, 0.7) + 0.1, mgp=c(2.2,0.4,0), bty="L", xp
d=TRUE, las=1, tcl=-0.15, cex.axis=1.15, cex.lab = 1)
for (i in 1:length(selected_genes)) {
  #i <- 1
  gene <- selected_genes[i]
  plot(survfit(surv ~ X[[gene]] == 0), col= pal1, bty='L', yaxs='i', ylim=c(0,1.01
), mark.time = T, conf.int = F)
  mtext(gene, font=3, side = 3, line = 0.2, cex = 0.83)
  legend("bottomleft", col=pal1[1:2], lty=1, c("MT","WT"), lwd = 1.5, bty="n", nco
l = 1, cex = 0.9, seg.len=0.7)
}
plot(survfit(surv ~ n_drivers), col=rev(pal1[1:3]), conf.int = F, mark.time = T, b
ty='L', yaxs='i', ylim=c(0,1.01))
mtext("Number of drivers", font=1, side = 3, line = 0.7, cex = 0.83)
legend("bottomleft", legend = levels(n_drivers), col= rev(pal1[1:3]), lty=1, lwd =
1.5, bty='n', title="", cex = 1, seg.len=0.7)
plot(survfit(surv ~ mvaf), col= rev(pal1[1:4]), conf.int = F, mark.time = T, bty='
L', yaxs='i', ylim=c(0,1.01))
mtext("Maximum VAF (%)", font=1, side = 3, line = 0.7, cex = 0.83)
```

```
legend("bottomleft", levels(mvaf), col=rev(pal1[1:4]), lty=1, lwd = 1.5, bty='n',
title="", cex = 1, seg.len=0.7)
plot(survfit(sangerSurv ~ rdw), col= rev(pal1[1:2]), conf.int = F, mark.time = T,
bty='L', yaxs='i', ylim=c(0,1.01))
mtext("RDW", font=1, side = 3, line = 0.2, cex = 0.83)
legend("bottomleft", levels(rdw), col=rev(pal1[1:2]), lty=1, lwd = 1.5, bty='n', t
itle="", cex = 1, seg.len=0.7)
dev.off()
```

```
## pdf
##   2
```

Standardise magnitudes

```
g <- sangerGroups=="Genes"
sangerX[g] <- sangerX[g] * 10
names(sangerX)[g] <- paste(names(sangerX[g]),"0.1", sep="_")
y <- StandardizeMagnitude(sangerX[!g])
sangerX <- cbind(sangerX[g],y)
```

10.2 Expected AML incidence

Validation cohort

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
sangerSurv2 <- Surv(sangerSurv[w,2], sangerSurv[w,3])

expected_rate_sanger_cr <- mean(aml_inc_cr(sangerX[w,"gender"],sangerX[w,"age_10"]
*10, sangerX[w,"age_10"]*10+ pmax(1,sangerSurv2[,1]))[!sangerSurv2[,2]])

n_total_sanger <- sum(sangerSurv2[,2])/expected_rate_sanger_cr
n_total_sanger
```

```
## [1] 13277.44
```

Discovery cohort only

```
expected_rate_toronto_cr <- mean(aml_inc_cr(torontoX[,"gender"],torontoX[,"age_10"
]*10, torontoX[,"age_10"]*10+ pmax(1,torontoSurv[,1]))[!torontoSurv[,2]])

n_total_toronto <- sum(torontoSurv[,2])/expected_rate_toronto_cr
n_total_toronto
```

```
## [1] 66014.85
```

10.3 Combined data

Survival

```
allSurv <- rbind(sangerSurv, Surv(rep(0, nrow(torontoSurv)), torontoSurv[,1], toro
ntoSurv[,2]))
allSurv <- Surv(allSurv[,1], allSurv[,2], allSurv[,3])
```

Data matrix

```
cohort <- c(rep("Sanger", nrow(sangerX)), rep("Toronto", nrow(torontoX)))
i <- c(sort(setdiff(gene_vars,"CALR")),"age","gender")
allX <- rbind(superSet(sangerData,i,fill=0), superSet(torontoData,i,fill=0))
allX <- allX[,colSums(allX>0)>=thr]
allX <- cbind(allX, cohort=cohort=="Sanger") + 0
allGroups <- factor(grepl("^[A-Z]",colnames(allX))+0, levels=1:0, labels=c("Genes"
,"Demographics"))

g <- allGroups=="Genes"
allX <- cbind(10*allX[,g], StandardizeMagnitude(allX[,!g]))
colnames(allX)[g] <- paste(colnames(allX)[g],"0.1",sep="_")
control <- c(sangerData$Diagnosis=="Control", torontoData$Diagnosis=="Control")
```

Weights

```
weights <- rep(1, nrow(allX))
weights[cohort=="Sanger" & control] <- n_total_sanger/sum(cohort=="Sanger" & contr
ol & allSurv[,1]==0)
weights[cohort=="Toronto" & control] <- n_total_toronto/sum(cohort=="Toronto" & co
ntrol)

n_total <- n_total_sanger + n_total_toronto
n_total
```

```
## [1] 79292.3
```

10.4 Coxph model fits

```
sigma0 <- 0.1
nu <- 1
which.mu <- "Genes"
```

10.4.1 Discovery cohort

10.4.1.1 Raw

```
fitToronto <- CoxRFX(torontoX, torontoSurv, groups=torontoGroups, which.mu=which.m
u, nu=nu, sigma0=sigma0)
waldToronto <- WaldTest(fitToronto)
```

```
##                   group    coef   coef-mu     sd      z df  p.value sig
## ASXL1_0.1          Genes  0.6922  0.049613 0.1172  5.908  1 3.47e-09 ***
## CALR_0.1           Genes  0.6239 -0.018696 0.0710  8.784  1 1.58e-18 ***
## CBL_0.1            Genes  0.5335 -0.109028 0.1293  4.126  1 3.70e-05 ***
## DNMT3A_0.1         Genes  0.5843 -0.058207 0.1059  5.517  1 3.44e-08 ***
## IDH1_0.1           Genes  0.6912  0.048657 0.1245  5.550  1 2.86e-08 ***
## IDH2_0.1           Genes  0.5136 -0.128999 0.1151  4.460  1 8.19e-06 ***
## JAK2_0.1           Genes  0.7120  0.069470 0.1243  5.730  1 1.00e-08 ***
## KDM6A_0.1          Genes  0.6419 -0.000647 0.0590 10.887  1 1.32e-27 ***
## KMT2C_0.1          Genes  0.6658  0.023265 0.0621 10.725  1 7.79e-27 ***
## KRAS_0.1           Genes  0.6403 -0.002210 0.0590 10.855  1 1.89e-27 ***
## NF1_0.1            Genes  0.6412 -0.001393 0.0590 10.870  1 1.61e-27 ***
## PHF6_0.1           Genes  0.6475  0.004993 0.0595 10.891  1 1.27e-27 ***
## PTPN11_0.1         Genes  0.6595  0.016950 0.0592 11.145  1 7.57e-29 ***
## RUNX1_0.1          Genes  0.4100 -0.232587 0.0923  4.443  1 8.89e-06 ***
## SF3B1_0.1          Genes  0.7728  0.130235 0.1019  7.585  1 3.33e-14 ***
## SRSF2_0.1          Genes  0.4783 -0.164235 0.0945  5.062  1 4.16e-07 ***
## TET2_0.1           Genes  0.6389 -0.003667 0.1295  4.932  1 8.13e-07 ***
## TP53_0.1           Genes  0.8079  0.165351 0.0673 12.009  1 3.19e-33 ***
## U2AF1_0.1          Genes  0.8537  0.211135 0.0773 11.048  1 2.23e-28 ***
## age_10      Demographics -0.0836 -0.083628 0.0975 -0.858  1 3.91e-01
## gender      Demographics  0.0113  0.011327 0.1091  0.104  1 9.17e-01
```

```
survConcordance(fitToronto$surv ~ fitToronto$linear.predictors)
```

```
## Call:
## survConcordance(formula = fitToronto$surv ~ fitToronto$linear.predictors)
##
##   n= 497
## Concordance= 0.7538671 se= 0.03218546
## concordant discordant  tied.risk  tied.time   std(c-d)
##   26561.00    8672.00       0.00       1.00    2267.98
```

10.4.2 Validation cohort

10.4.2.1 Raw

```
fitSanger <- CoxRFX(sangerX, sangerSurv, groups=sangerGroups, which.mu=which.mu, n
u=nu, sigma0=sigma0)
waldSanger <- WaldTest(fitSanger)
```

```
##                    group     coef   coef-mu      sd      z df  p.value sig
## ASXL1_0.1           Genes  0.64051  0.105357 0.11285  5.676  1 1.38e-08 ***
## CBL_0.1             Genes  0.52291 -0.012246 0.08720  5.997  1 2.01e-09 ***
## DNMT3A_0.1          Genes  0.43301 -0.102144 0.11026  3.927  1 8.60e-05 ***
## JAK2_0.1            Genes  0.52046 -0.014699 0.09655  5.391  1 7.02e-08 ***
## KMT2C_0.1           Genes  0.54634  0.011184 0.08151  6.703  1 2.05e-11 ***
## KMT2D_0.1           Genes  0.42573 -0.109421 0.14122  3.015  1 2.57e-03  **
## KRAS_0.1           Genes  0.53897  0.003816 0.08013  6.726  1 1.74e-11 ***
## NF1_0.1            Genes  0.52911 -0.006044 0.08135  6.504  1 7.80e-11 ***
## NRAS_0.1           Genes  0.53431 -0.000849 0.08011  6.670  1 2.56e-11 ***
## RAD21_0.1          Genes  0.53226 -0.002897 0.08049  6.613  1 3.77e-11 ***
## SF3B1_0.1          Genes  0.53076 -0.004391 0.08104  6.550  1 5.76e-11 ***
## SRSF2_0.1          Genes  0.50357 -0.031583 0.11851  4.249  1 2.14e-05 ***
## TET2_0.1           Genes  0.58716  0.052000 0.10482  5.602  1 2.12e-08 ***
## TP53_0.1           Genes  0.58827  0.053119 0.08077  7.283  1 3.25e-13 ***
## U2AF1_0.1          Genes  0.59395  0.058796 0.08084  7.347  1 2.03e-13 ***
## age_10      Demographics  0.08031  0.080306 0.11847  0.678  1 4.98e-01
## gender      Demographics -0.11803 -0.118029 0.11360 -1.039  1 2.99e-01
## systol_100         Blood  0.01074  0.010736 0.04230  0.254  1 8.00e-01
## diastol_100        Blood  0.02297  0.022974 0.02697  0.852  1 3.94e-01
## bmi_10             Blood  0.09128  0.091285 0.07510  1.215  1 2.24e-01
## cholestl_10        Blood  0.00934  0.009343 0.01381  0.676  1 4.99e-01
## triglyc            Blood  0.02435  0.024354 0.09637  0.253  1 8.00e-01
## hdl                Blood -0.07521 -0.075205 0.07691 -0.978  1 3.28e-01
## ldl                Blood  0.12764  0.127641 0.09931  1.285  1 1.99e-01
## lym                Blood  0.07714  0.077135 0.09427  0.818  1 4.13e-01
## mcv_100            Blood -0.00987 -0.009867 0.00826 -1.195  1 2.32e-01
## rdw_10             Blood  0.06196  0.061956 0.02072  2.990  1 2.79e-03  **
## wbc_10             Blood  0.01894  0.018939 0.03734  0.507  1 6.12e-01
## plt_100            Blood  0.05344  0.053435 0.09405  0.568  1 5.70e-01
## hgb_10             Blood  0.05198  0.051979 0.02446  2.125  1 3.36e-02   *
```

```
survConcordance(sangerSurv ~ fitSanger$linear.predictors)
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitSanger$linear.predictors)
##
##   n= 459
## Concordance= 0.7224015 se= 0.04865039
## concordant discordant  tied.risk  tied.time   std(c-d)
##  6714.0000  2580.0000     0.0000     0.0000   904.3134
```

10.4.2.2 Adjusted

```
fitWeightedSanger <- CoxRFX(sangerX, sangerSurv, sangerGroups, which.mu=which.mu,
sigma0=sigma0, nu=nu, weights=weights[cohort=="Sanger"])
waldWeightedSanger <- WaldTest(fitWeightedSanger)
```

```
##                    group      coef   coef-mu      sd        z df  p.value sig
## ASXL1_0.1          Genes  2.634306  0.838861 0.43502  6.05558  1 1.40e-09 ***
## CBL_0.1            Genes  0.630557 -1.164888 1.13502  0.55555  1 5.79e-01
## DNMT3A_0.1         Genes  0.698827 -1.096619 0.22597  3.09251  1 1.98e-03  **
## JAK2_0.1           Genes  0.049363 -1.746082 0.90486  0.05455  1 9.56e-01
## KMT2C_0.1          Genes  1.829655  0.034210 1.05055  1.74162  1 8.16e-02   .
## KMT2D_0.1          Genes -0.004783 -1.800228 0.75790 -0.00631  1 9.95e-01
## KRAS_0.1           Genes  2.139544  0.344099 0.40749  5.25049  1 1.52e-07 ***
## NF1_0.1            Genes  1.252510 -0.542935 0.89204  1.40410  1 1.60e-01
## NRAS_0.1           Genes  1.730987 -0.064459 0.36379  4.75820  1 1.95e-06 ***
## RAD21_0.1          Genes  1.487062 -0.308383 0.68933  2.15726  1 3.10e-02   *
## SF3B1_0.1          Genes  1.309652 -0.485793 0.96376  1.35890  1 1.74e-01
## SRSF2_0.1          Genes  1.451418 -0.344027 0.27015  5.37269  1 7.76e-08 ***
## TET2_0.1           Genes  1.222954 -0.572491 0.12864  9.50695  1 1.96e-21 ***
## TP53_0.1           Genes  4.699561  2.904116 0.91319  5.14632  1 2.66e-07 ***
## U2AF1_0.1          Genes  5.800067  4.004622 0.74776  7.75664  1 8.72e-15 ***
## age_10      Demographics  0.024711  0.024711 0.12062  0.20487  1 8.38e-01
## gender      Demographics -0.140352 -0.140352 0.11358 -1.23575  1 2.17e-01
## systol_100         Blood -0.000324 -0.000324 0.04456 -0.00726  1 9.94e-01
## diastol_100        Blood  0.019654  0.019654 0.02894  0.67907  1 4.97e-01
## bmi_10             Blood  0.101555  0.101555 0.08137  1.24811  1 2.12e-01
## cholestl_10        Blood  0.007469  0.007469 0.01457  0.51275  1 6.08e-01
## triglyc            Blood  0.007316  0.007316 0.10707  0.06832  1 9.46e-01
## hdl                Blood -0.108973 -0.108973 0.08295 -1.31365  1 1.89e-01
## ldl                Blood  0.149658  0.149658 0.10397  1.43938  1 1.50e-01
## lym                Blood  0.066987  0.066987 0.09901  0.67660  1 4.99e-01
## mcv_100            Blood -0.015964 -0.015964 0.00832 -1.91787  1 5.51e-02   .
## rdw_10             Blood  0.073201  0.073201 0.01789  4.09058  1 4.30e-05 ***
## wbc_10             Blood  0.020190  0.020190 0.04345  0.46465  1 6.42e-01
## plt_100            Blood  0.077199  0.077199 0.10027  0.76987  1 4.41e-01
## hgb_10             Blood  0.044376  0.044376 0.02513  1.76558  1 7.75e-02   .
```

```
survConcordance(sangerSurv ~ fitWeightedSanger$linear.predictors, weights=weights[
cohort=="Sanger"])
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitWeightedSanger$linear.predictors,
##     weights = weights[cohort == "Sanger"])
##
##   n= 459
## Concordance= 0.7639423 se= 0.04828991
## concordant discordant  tied.risk  tied.time   std(c-d)
##  334537.56  103371.88       0.00       0.00   42293.22
```

Uno’s estimator of cumulative/dynamic AUC

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
s <- Surv(sangerSurv[w,2], sangerSurv[w,3])
a <- AUC.uno(s, s, fitWeightedSanger$linear.predictors[w], times= c(0, 22, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.761
```

11 Model excluding controls without mutations

Include only controls with ARCH & all pre-AML (regardless of mutation status) ## Discovery cohort (Toronto) Data

```
f = "data/DC_vaf_matrix_no_duplicates_414ctrl_83aml.csv"
torontoData <- read.csv(f)

gene_vars <- c("CALR", "NRAS", "DNMT3A", "SF3B1", "IDH1", "KIT", "TET2", "RAD21",
"JAK2", "CBL", "KRAS", "PTPN11", "IDH2", "TP53", "NF1", "SRSF2", "CEBPA", "ASXL1",
"RUNX1", "U2AF1", "BCOR", "KDM6A", "PHF6", "KMT2C", "KMT2D")

table(torontoData$Diagnosis)
```

```
##
##     AML Control
##      83     414
```

```
torontoData$gender <- ifelse(torontoData$Sex == "male", 1,
                             ifelse(torontoData$Sex == "female", 0, torontoData$Se
x))
dim(torontoData)
```

```
## [1] 497  29
```

```
torontoData <- torontoData[rowSums(torontoData[, colnames(torontoData) %in% gene_v
ars])>0 | torontoData$Diagnosis == "AML", ]
dim(torontoData)
```

```
## [1] 240  29
```

```
table(torontoData$gender)
```

```
##
##   0   1
## 135 105
```

```
torontoData$gender <- as.numeric(torontoData$gender)
colnames(torontoData)
```

```
##  [1] "Sample"     "ASXL1"      "BCOR"       "CALR"       "CBL"        "DNMT3A"
"IDH1"       "IDH2"
##  [9] "JAK2"       "KDM6A"      "KIT"        "KMT2C"      "KRAS"       "NF1"
"NRAS"       "PHF6"
## [17] "PTPN11"     "RUNX1"      "SF3B1"      "SRSF2"      "TET2"       "TP53"
"U2AF1"      "Diagnosis"
## [25] "fu_years"   "age"        "Sex"        "no_drivers" "gender"
```

Manually standardize magnitudes

```
torontoData <- torontoData[!duplicated(torontoData),]
```

```
torontoX <- torontoData[, colnames(torontoData) %in% c(gene_vars, "age", "gender")
]

torontoX <- as.data.frame(torontoX)
thr <- 2
torontoX <- torontoX[,colSums(torontoX != 0)>=thr]

torontoGroups <- factor(names(torontoX) %in% c("age","gender")+1, level=1:2, label
s=c("Genes","Demographics"))
colnames(torontoX)
```

```
##  [1] "ASXL1"  "CALR"   "CBL"    "DNMT3A" "IDH1"   "IDH2"   "JAK2"   "KDM6A"  "K
MT2C"  "KRAS"   "NF1"    "PHF6"  
## [13] "PTPN11" "RUNX1"  "SF3B1"  "SRSF2"  "TET2"   "TP53"   "U2AF1"  "age"    "g
ender"
```

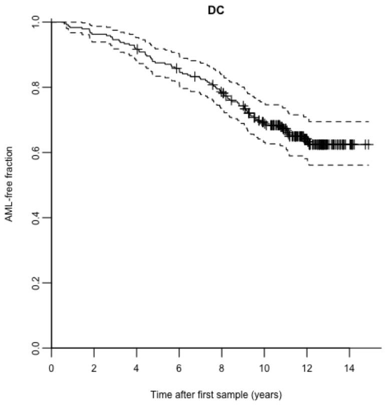
```
torontoGroups
```

```
##  [1] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
##  [9] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
## [17] Genes        Genes        Genes        Demographics Demographics
## Levels: Genes Demographics
```

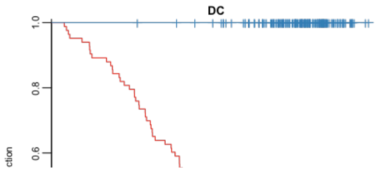
```
# Manually standardize age and mutation VAFs
torontoX$age <- torontoX$age/10
names(torontoX)[which(names(torontoX)=="age")] <- "age_10"
g <- torontoGroups == "Genes"
torontoX[,g] <- torontoX[,g]*10
names(torontoX)[g] <- paste(names(torontoX)[g], "0.1",sep="_")
colnames(torontoX)
```

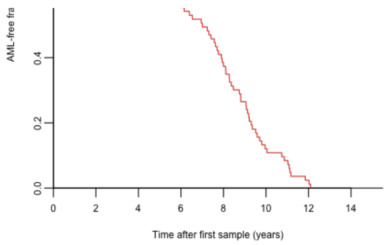
```
##  [1] "ASXL1_0.1"  "CALR_0.1"   "CBL_0.1"    "DNMT3A_0.1" "IDH1_0.1"   "IDH2_0.1
"   "JAK2_0.1"   "KDM6A_0.1" 
##  [9] "KMT2C_0.1"  "KRAS_0.1"   "NF1_0.1"    "PHF6_0.1"   "PTPN11_0.1" "RUNX1_0.
1"  "SF3B1_0.1"  "SRSF2_0.1" 
## [17] "TET2_0.1"   "TP53_0.1"   "U2AF1_0.1"  "age_10"     "gender"
```

```
torontoSurv <- Surv(torontoData$fu_years, torontoData$Diagnosis=="AML")
plot(survfit(torontoSurv~ 1), col= "black", main = "DC", xlab='Time after first sa
mple (years)', ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.t
ime = T)
```



```
plot(survfit(torontoSurv ~ torontoData$Diagnosis), xlab='Time after first sample (
years)', main = "DC", ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01),
mark.time = T, col = set1[1:2])
```





11.1 Validation cohort

```
f = "data/VC_vaf_matrix_262ctrl_37aml_nodates.csv"
sangerData <- read.csv(f)
dim(sangerData)
```

```
## [1] 459  43
```

```
sangerData <- sangerData[rowSums(sangerData[, colnames(sangerData) %in% gene_vars]
)>0 | sangerData$Diagnosis == "AML", ]
dim(sangerData)
```

```
## [1] 173  43
```

```
length(unique(sangerData$Individual))
```

```
## [1] 128
```

```
sangerData$hcdate <- as.Date(sangerData$hcdate)
sangerData$dodx <- as.Date(sangerData$dodx)

sangerPatients <- sub("[a-z]+$","", sangerData$Sample)
o <- order(sangerPatients, as.numeric(sangerData$hcdate))

sangerData <- sangerData[o,]
sangerPatients <- sangerPatients[o]

clinical_vars <- c("systol", "diastol", "bmi", "cholestl", "triglyc", "hdl", "ldl"
, "lym", "mcv", "rdw", "wbc", "plt", "hgb")
sangerX <- sangerData[, colnames(sangerData) %in% c(gene_vars, "age","gender",clin
ical_vars)]
sangerX <- as.data.frame(sangerX)

sangerX <- sangerX[,colSums(sangerX != 0,na.rm=TRUE)>=thr]
sangerGroups <- factor(grepl("^[a-z]", colnames(sangerX))*2, levels=0:2, labels=c(
"Genes", "Demographics", "Blood"))
sangerGroups[names(sangerX) %in% c("age","gender")] <- "Demographics"
table(sangerGroups)
```

```
## sangerGroups
##        Genes Demographics        Blood 
##           15            2           13
```

```
colnames(sangerX)
```

```
##  [1] "ASXL1"    "CBL"      "DNMT3A"   "JAK2"     "KMT2C"    "KMT2D"    "KRAS"    
"NF1"      "NRAS"     "RAD21"   
## [11] "SF3B1"    "SRSF2"    "TET2"     "TP53"     "U2AF1"    "age"      "gender"  
"systol"   "diastol"  "bmi"     
## [21] "cholestl" "triglyc"  "hdl"      "ldl"      "lym"      "mcv"      "rdw"     
"wbc"      "plt"      "hgb"
```

```
sangerGroups
```

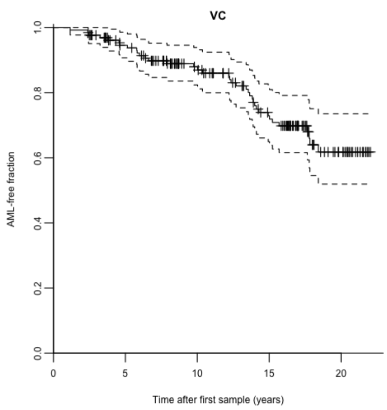
```
##  [1] Genes        Genes        Genes        Genes        Genes        Genes
Genes        Genes
##  [9] Genes        Genes        Genes        Genes        Genes        Genes
Genes        Demographics
## [17] Demographics Blood        Blood        Blood        Blood        Blood
Blood        Blood
## [25] Blood        Blood        Blood        Blood        Blood        Blood
## Levels: Genes Demographics Blood
```

```
g <- sangerGroups=="Genes"
sangerX[g] <- sangerX[g] * 10
names(sangerX)[g] <- paste(names(sangerX[g]),"0.1", sep="_")
y <- StandardizeMagnitude(sangerX[!g])
sangerX <- cbind(sangerX[g],y)

poorMansImpute <- function(x) {x[is.na(x)] <- mean(x, na.rm=TRUE); return(x)}
sangerX <- as.data.frame(sapply(sangerX, poorMansImpute))

foo <- split(sangerData[,c("Diagnosis","hcdate","dodx")], sangerPatients)

bar <- do.call("rbind",lapply(foo, function(x){
  y <- x
  n <- nrow(y)
  y[-n,"Diagnosis"] <- "Control"
  start <- as.numeric(y$hcdate - y$hcdate[1])/365.25
  end <- c(as.numeric(y$hcdate - y$hcdate[1])[-1]/365.25, as.numeric(y$dodx[n] - y
$hcdate[1])/365.25)
  return(data.frame(Diagnosis=y[,"Diagnosis"], start=start, end=end))
}))

bar[1:10, ]
```

	Diagnosis <fctr>	start <dbl>	end <dbl>
PD29762	AML	0.000000	9.754962
PD29764	AML	0.000000	10.360027
PD29792	AML	0.000000	14.108145
PD29810	Control	0.000000	18.573580
PD29836.1	Control	0.000000	2.414784
PD29836.2	AML	2.414784	10.023272
PD29851.1	Control	0.000000	4.599589
PD29851.2	AML	4.599589	12.205339
PD29856.1	Control	0.000000	4.331280
PD29856.2	AML	4.331280	17.828884

1-10 of 10 rows

```
sangerPatientsSplit <- unlist(sapply(names(foo), function(n) rep(n, nrow(foo[[n]])
)))

sangerSurv <- Surv(time = bar$start, time2 = bar$end, event = bar$Diagnosis!="Cont
rol", origin = 0)

plot(survfit(sangerSurv~ 1), col= "black", main = "VC", xlab='Time after first sam
ple (years)', ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.ti
me = T) #mark = 1
```



11.2 Expected AML incidence

Validation cohort

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
sangerSurv2 <- Surv(sangerSurv[w,2], sangerSurv[w,3]) ## Unique individuals

expected_rate_sanger_cr <- mean(aml_inc_cr(sangerX[w,"gender"],sangerX[w,"age_10"]
*10, sangerX[w,"age_10"]*10+ pmax(1,sangerSurv2[,1]))[!sangerSurv2[,2]])

n_total_sanger <- sum(sangerSurv2[,2])/expected_rate_sanger_cr
n_total_sanger
```

```
## [1] 14208.3
```

Discovery cohort

```
expected_rate_toronto_cr <- mean(aml_inc_cr(torontoX[,"gender"],torontoX[,"age_10"
]*10, torontoX[,"age_10"]*10+ pmax(1,torontoSurv[,1]))[!torontoSurv[,2]])

n_total_toronto <- sum(torontoSurv[,2])/expected_rate_toronto_cr
n_total_toronto
```

```
## [1] 55688.66
```

11.3 Combined data

Survival

```
allSurv <- rbind(sangerSurv, Surv(rep(0, nrow(torontoSurv)), torontoSurv[,1], toro
ntoSurv[,2]))
allSurv <- Surv(allSurv[,1], allSurv[,2], allSurv[,3])
```

Data matrix

```
cohort <- c(rep("Sanger", nrow(sangerX)), rep("Toronto", nrow(torontoX)))
i <- c(sort(setdiff(gene_vars,"CALR")),"age","gender")
allX <- rbind(superSet(sangerData,i,fill=0), superSet(torontoData,i,fill=0))
allX <- allX[,colSums(allX>0)>=thr]
allX <- cbind(allX, cohort=cohort=="Sanger") + 0
allGroups <- factor(grepl("^[A-Z]",colnames(allX))+0, levels=1:0, labels=c("Genes"
,"Demographics"))

g <- allGroups=="Genes"
allX <- cbind(10*allX[,g], StandardizeMagnitude(allX[,!g]))
colnames(allX)[g] <- paste(colnames(allX)[g],"0.1",sep="_")
control <- c(sangerData$Diagnosis=="Control", torontoData$Diagnosis=="Control")
```

Weights

```
weights <- rep(1, nrow(allX))
weights[cohort=="Sanger" & control] <- n_total_sanger/sum(cohort=="Sanger" & contr
ol & allSurv[,1]==0)
weights[cohort=="Toronto" & control] <- n_total_toronto/sum(cohort=="Toronto" & co
ntrol)

n_total <- n_total_sanger + n_total_toronto
n_total
```

```
## [1] 69896.97
## [1] 69896.97
```

11.4 Coxph model fits

```
sigma0 <- 0.1
nu <- 1
which.mu <- "Genes"
```

11.4.1 DC

11.4.1.1 Raw

```
fitToronto <- CoxRFX(torontoX, torontoSurv, groups=torontoGroups, which.mu=which.m
u, nu=nu, sigma0=sigma0)
waldToronto <- WaldTest(fitToronto)
```

```
##                    group    coef   coef-mu     sd      z df  p.value sig
## ASXL1_0.1          Genes  0.4801  0.050389 0.1108  4.335  1 1.46e-05 ***
## CALR_0.1           Genes  0.4076 -0.022055 0.0700  5.824  1 5.76e-09 ***
## CBL_0.1            Genes  0.3119 -0.117817 0.1151  2.710  1 6.72e-03  **
## DNMT3A_0.1         Genes  0.3010 -0.128687 0.1054  2.857  1 4.28e-03  **
## IDH1_0.1           Genes  0.4535  0.023828 0.1092  4.152  1 3.29e-05 ***
## IDH2_0.1           Genes  0.3789 -0.050806 0.1052  3.602  1 3.15e-04 ***
## JAK2_0.1           Genes  0.4956  0.065922 0.1136  4.364  1 1.28e-05 ***
## KDM6A_0.1          Genes  0.4288 -0.000932 0.0594  7.214  1 5.45e-13 ***
## KMT2C_0.1          Genes  0.4450  0.015284 0.0619  7.194  1 6.28e-13 ***
## KRAS_0.1           Genes  0.4257 -0.004039 0.0595  7.156  1 8.31e-13 ***
## NF1_0.1            Genes  0.4272 -0.002451 0.0595  7.183  1 6.80e-13 ***
## PHF6_0.1           Genes  0.4321  0.002404 0.0598  7.230  1 4.83e-13 ***
## PTPN11_0.1         Genes  0.4414  0.011735 0.0596  7.407  1 1.29e-13 ***
## RUNX1_0.1          Genes  0.2761 -0.153642 0.0890  3.102  1 1.92e-03  **
## SF3B1_0.1          Genes  0.5346  0.104912 0.0892  5.993  1 2.06e-09 ***
## SRSF2_0.1          Genes  0.3772 -0.052539 0.0883  4.274  1 1.92e-05 ***
## TET2_0.1           Genes  0.4247 -0.005040 0.1174  3.617  1 2.98e-04 ***
## TP53_0.1           Genes  0.5441  0.114421 0.0665  8.181  1 2.81e-16 ***
## U2AF1_0.1          Genes  0.5788  0.149112 0.0722  8.015  1 1.10e-15 ***
## age_10      Demographics -0.3093 -0.309301 0.1116 -2.771  1 5.59e-03  **
## gender      Demographics -0.0253 -0.025329 0.1385 -0.183  1 8.55e-01
```

```
survConcordance(fitToronto$surv ~ fitToronto$linear.predictors, weights = weights[
cohort=="Toronto"])
```

```
## Call:
## survConcordance(formula = fitToronto$surv ~ fitToronto$linear.predictors,
##     weights = weights[cohort == "Toronto"])
##
##   n= 240
## Concordance= 0.7539084 se= 0.03193557
## concordant discordant  tied.risk  tied.time   std(c-d)
##  3255935.4  1062805.9        0.0        1.0   275842.9
```

11.4.1.2 Adjusted

```
fitWeightedToronto <- CoxRFX(torontoX, torontoSurv, torontoGroups, which.mu=which.
mu, sigma0=sigma0, nu=nu, weights=weights[cohort=="Toronto"])
waldWeightedToronto <- WaldTest(fitWeightedToronto)
```

```
##                    group    coef coef-mu    sd       z df  p.value sig
## ASXL1_0.1          Genes  1.9719  0.1365 0.150 13.1816  1 1.12e-39 ***
## CALR_0.1           Genes -0.0794 -1.9147 1.174 -0.0676  1 9.46e-01
## CBL_0.1            Genes  0.0165 -1.8188 0.426  0.0388  1 9.69e-01
## DNMT3A_0.1         Genes  0.3722 -1.4631 0.153  2.4301  1 1.51e-02   *
## IDH1_0.1           Genes  2.3375  0.5022 0.350  6.6815  1 2.36e-11 ***
## IDH2_0.1           Genes  0.5915 -1.2438 0.240  2.4621  1 1.38e-02   *
## JAK2_0.1           Genes  1.7762 -0.0592 0.193  9.2213  1 2.94e-20 ***
## KDM6A_0.1          Genes  1.6689 -0.1664 0.362  4.6081  1 4.06e-06 ***
## KMT2C_0.1          Genes -1.2330 -3.0683 1.191 -1.0356  1 3.00e-01
## KRAS_0.1           Genes  0.9875 -0.8478 0.555  1.7785  1 7.53e-02   .
## NF1_0.1            Genes  1.3623 -0.4730 0.501  2.7193  1 6.54e-03  **
## PHF6_0.1           Genes  2.6990  0.8636 0.255 10.5887  1 3.36e-26 ***
## PTPN11_0.1         Genes  3.6339  1.7986 0.723  5.0228  1 5.09e-07 ***
## RUNX1_0.1         Genes  0.6233 -1.2120 0.136  4.5906  1 4.42e-06 ***
## SF3B1_0.1         Genes  3.1088  1.2735 0.305 10.1981  1 2.02e-24 ***
## SRSF2_0.1         Genes  1.4956 -0.3397 0.172  8.6791  1 3.99e-18 ***
## TET2_0.1          Genes  0.5772 -1.2581 0.232  2.4920  1 1.27e-02   *
## TP53_0.1          Genes  8.9422  7.1069 0.823 10.8665  1 1.66e-27 ***
## U2AF1_0.1         Genes  4.0190  2.1836 0.384 10.4738  1 1.14e-25 ***
## age_10     Demographics -0.5274 -0.5274 0.135 -3.9171  1 8.96e-05 ***
## gender     Demographics  0.0323  0.0323 0.175  0.1842  1 8.54e-01
```

```
survConcordance(fitWeightedToronto$surv ~ fitWeightedToronto$linear.predictors, we
ights=weights[cohort=="Toronto"])
```

```
## Call:
## survConcordance(formula = fitWeightedToronto$surv ~ fitWeightedToronto$linear.p
redictors,
##     weights = weights[cohort == "Toronto"])
##
##   n= 240
## Concordance= 0.7701663 se= 0.03193557
## concordant discordant  tied.risk  tied.time   std(c-d)
##  3326148.9   992592.4        0.0        1.0   275842.9
```

```
#Uno's estimator of cumulative/dynamic AUC
a <- AUC.uno(torontoSurv, torontoSurv, fitWeightedToronto$linear.predictors, times
= seq(0,12, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.756
```

11.4.2 Validation cohort

11.4.2.1 Raw

```
fitSanger <- CoxRFX(sangerX, sangerSurv, groups=sangerGroups, which.mu=which.mu, n
u=nu, sigma0=sigma0)
waldSanger <- WaldTest(fitSanger)
```

```
##                    group     coef   coef-mu      sd       z df  p.value sig
## ASXL1_0.1          Genes  0.41389  1.04e-01 0.13253  3.1229  1 1.79e-03  **
## CBL_0.1            Genes  0.27978 -3.01e-02 0.10678  2.6202  1 8.79e-03  **
## DNMT3A_0.1         Genes  0.15476 -1.55e-01 0.12703  1.2183  1 2.23e-01
## JAK2_0.1           Genes  0.33012  2.02e-02 0.10874  3.0359  1 2.40e-03  **
## KMT2C_0.1          Genes  0.30175 -8.17e-03 0.09722  3.1037  1 1.91e-03  **
## KMT2D_0.1          Genes  0.14350 -1.66e-01 0.15722  0.9127  1 3.61e-01
## KRAS_0.1           Genes  0.30998  5.67e-05 0.09168  3.3811  1 7.22e-04 ***
## NF1_0.1            Genes  0.29225 -1.77e-02 0.09499  3.0768  1 2.09e-03  **
## NRAS_0.1           Genes  0.30685 -3.07e-03 0.09158  3.3507  1 8.06e-04 ***
## RAD21_0.1          Genes  0.29301 -1.69e-02 0.09373  3.1261  1 1.77e-03  **
## SF3B1_0.1          Genes  0.29894 -1.10e-02 0.09393  3.1825  1 1.46e-03  **
## SRSF2_0.1          Genes  0.40493  9.50e-02 0.13441  3.0125  1 2.59e-03  **
## TET2_0.1           Genes  0.37910  6.92e-02 0.11275  3.3624  1 7.73e-04 ***
## TP53_0.1           Genes  0.36746  5.75e-02 0.09308  3.9479  1 7.88e-05 ***
## U2AF1_0.1          Genes  0.37254  6.26e-02 0.09357  3.9813  1 6.85e-05 ***
## age_10      Demographics -0.01773 -1.77e-02 0.11451 -0.1548  1 8.77e-01
## gender      Demographics -0.03369 -3.37e-02 0.10501 -0.3208  1 7.48e-01
## systol_100         Blood  0.00145  1.45e-03 0.03839  0.0377  1 9.70e-01
## diastol_100        Blood  0.00773  7.73e-03 0.02329  0.3321  1 7.40e-01
## bmi_10             Blood  0.06828  6.83e-02 0.07091  0.9628  1 3.36e-01
## cholestl_10        Blood  0.01797  1.80e-02 0.01274  1.4109  1 1.58e-01
## triglyc            Blood  0.00471  4.71e-03 0.09569  0.0492  1 9.61e-01
## hdl                Blood -0.00891 -8.91e-03 0.07257 -0.1227  1 9.02e-01
## ldl                Blood  0.16056  1.61e-01 0.09725  1.6510  1 9.87e-02   .
## lym                Blood -0.02015 -2.01e-02 0.08835 -0.2280  1 8.20e-01
## mcv_100            Blood -0.00369 -3.69e-03 0.00786 -0.4694  1 6.39e-01
## rdw_10             Blood  0.05420  5.42e-02 0.02080  2.6056  1 9.17e-03  **
## wbc_10             Blood  0.00379  3.79e-03 0.03521  0.1077  1 9.14e-01
## plt_100            Blood  0.03410  3.41e-02 0.09166  0.3720  1 7.10e-01
## hgb_10             Blood  0.03314  3.31e-02 0.02245  1.4763  1 1.40e-01
```

```
survConcordance(sangerSurv ~ fitSanger$linear.predictors)
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitSanger$linear.predictors)
##
```

```
##   n= 173
## Concordance= 0.6611972 se= 0.05025086
## concordant discordant  tied.risk  tied.time   std(c-d)
##  2176.0000  1115.0000     0.0000     0.0000   330.7512
```

11.4.2.2 Adjusted

```
fitWeightedSanger <- CoxRFX(sangerX, sangerSurv, sangerGroups, which.mu=which.mu,
sigma0=sigma0, nu=nu, weights=weights[cohort=="Sanger"])
waldWeightedSanger <- WaldTest(fitWeightedSanger)
```

```
##                     group      coef   coef-mu      sd        z df  p.value sig
## ASXL1_0.1          Genes  2.580959  1.414558 0.47618  5.42008  1 5.96e-08 ***
## CBL_0.1            Genes -0.660213 -1.826614 1.39628 -0.47284  1 6.36e-01
## DNMT3A_0.1         Genes  0.223151 -0.943251 0.24504  0.91066  1 3.62e-01
## JAK2_0.1           Genes  0.705927 -0.460474 1.04486  0.67562  1 4.99e-01
## KMT2C_0.1          Genes -0.385529 -1.551931 1.44435 -0.26692  1 7.90e-01
## KMT2D_0.1          Genes -0.627231 -1.793633 1.03607 -0.60539  1 5.45e-01
## KRAS_0.1           Genes  1.299133  0.132731 0.78999  1.64450  1 1.00e-01
## NF1_0.1            Genes -0.815764 -1.982166 1.46470 -0.55695  1 5.78e-01
## NRAS_0.1           Genes  0.728314 -0.438088 0.64251  1.13355  1 2.57e-01
## RAD21_0.1          Genes -0.678392 -1.844793 1.44210 -0.47042  1 6.38e-01
## SF3B1_0.1          Genes  0.072745 -1.093657 1.47708  0.04925  1 9.61e-01
## SRSF2_0.1          Genes  1.726024  0.559622 0.23912  7.21826  1 5.27e-13 ***
## TET2_0.1           Genes  1.101278 -0.065124 0.15079  7.30320  1 2.81e-13 ***
## TP53_0.1           Genes  4.694801  3.528400 1.13074  4.15198  1 3.30e-05 ***
## U2AF1_0.1          Genes  7.530821  6.364419 1.06931  7.04270  1 1.89e-12 ***
## age_10      Demographics -0.190256 -0.190256 0.13151 -1.44666  1 1.48e-01
## gender      Demographics -0.029742 -0.029742 0.12174 -0.24430  1 8.07e-01
## systol_100         Blood -0.032537 -0.032537 0.04764 -0.68293  1 4.95e-01
## diastol_100        Blood  0.000105  0.000105 0.02958  0.00356  1 9.97e-01
## bmi_10             Blood  0.098774  0.098774 0.08970  1.10111  1 2.71e-01
## cholestl_10        Blood  0.024226  0.024226 0.01553  1.55989  1 1.19e-01
## triglyc            Blood  0.051097  0.051097 0.11392  0.44854  1 6.54e-01
## hdl                Blood -0.082426 -0.082426 0.09326 -0.88380  1 3.77e-01
## ldl                Blood  0.248075  0.248075 0.11127  2.22950  1 2.58e-02   *
## lym                Blood -0.054414 -0.054414 0.10621 -0.51234  1 6.08e-01
## mcv_100            Blood -0.010783 -0.010783 0.00915 -1.17903  1 2.38e-01
## rdw_10             Blood  0.095279  0.095279 0.01797  5.30078  1 1.15e-07 ***
## wbc_10             Blood  0.011314  0.011314 0.04898  0.23099  1 8.17e-01
## plt_100            Blood  0.057755  0.057755 0.11248  0.51347  1 6.08e-01
## hgb_10             Blood  0.016212  0.016212 0.02615  0.62004  1 5.35e-01
```

```
waldWeightedSanger$p.adj <- p.adjust(p = waldWeightedSanger$p.value, method = "bon
ferroni")
#View(waldWeightedSanger)

survConcordance(sangerSurv ~ fitWeightedSanger$linear.predictors, weights=weights[
cohort=="Sanger"])
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitWeightedSanger$linear.predictors,
##     weights = weights[cohort == "Sanger"])
##
##   n= 173
## Concordance= 0.7231124 se= 0.0489519
## concordant discordant  tied.risk  tied.time   std(c-d)
##  296852.77  113668.16       0.00       0.00   40191.56
```

```
#Uno's estimator of cumulative/dynamic AUC
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
s <- Surv(sangerSurv[w,2], sangerSurv[w,3])
a <- AUC.uno(s, s, fitWeightedSanger$linear.predictors[w], times= c(0, 22, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.403
```

12 CoxPH model excluding all samples without ARCH-PD

12.1 Discovery cohort

Data

```
f = "data/DC_vaf_matrix_414ctrl_91aml.csv"
torontoData <- read.csv(f)

gene_vars <- c("CALR", "NRAS", "DNMT3A", "SF3B1", "IDH1", "KIT", "TET2", "RAD21",
"JAK2", "CBL", "KRAS", "PTPN11", "IDH2", "TP53", "NF1", "SRSF2", "CEBPA", "ASXL1",
"RUNX1", "U2AF1", "BCOR", "KDM6A", "PHF6", "KMT2C", "KMT2D")

table(torontoData$Diagnosis)
```

```
## 
##     AML Control 
##      91     414
```

```
torontoData$gender <- ifelse(torontoData$Sex == "male", 1,
                             ifelse(torontoData$Sex == "female", 0, torontoData$Se
x))
dim(torontoData)
```

```
## [1] 505  29
```

```
torontoData <- torontoData[rowSums(torontoData[, colnames(torontoData) %in% gene_v
ars])>0, ]
dim(torontoData)
```

```
## [1] 221  29
```

```
table(torontoData$gender)
```

```
## 
##   0   1 
## 126  95
```

```
torontoData$gender <- as.numeric(torontoData$gender)
colnames(torontoData)
```

```
##  [1] "Sample"     "ASXL1"      "BCOR"       "CALR"       "CBL"        "DNMT3A"    
"IDH1"       "IDH2"      
##  [9] "JAK2"       "KDM6A"      "KIT"        "KMT2C"      "KRAS"       "NF1"       
"NRAS"       "PHF6"      
## [17] "PTPN11"     "RUNX1"      "SF3B1"      "SRSF2"      "TET2"       "TP53"      
"U2AF1"      "Diagnosis" 
## [25] "fu_years"   "age"        "Sex"        "no_drivers" "gender"
```

Manually standardize magnitudes

```
torontoData <- torontoData[!duplicated(torontoData),]

torontoX <- torontoData[, colnames(torontoData) %in% c(gene_vars, "age", "gender")
]

torontoX <- as.data.frame(torontoX)
thr <- 2
torontoX <- torontoX[,colSums(torontoX != 0)>=thr]

torontoGroups <- factor(names(torontoX) %in% c("age","gender")+1, level=1:2, label
s=c("Genes","Demographics"))
colnames(torontoX)
```

```
##  [1] "ASXL1"  "CALR"   "CBL"    "DNMT3A" "IDH1"   "IDH2"   "JAK2"   "KDM6A"  "K
MT2C"  "KRAS"   "NF1"    "PHF6"  
## [13] "PTPN11" "RUNX1"  "SF3B1"  "SRSF2"  "TET2"   "TP53"   "U2AF1"  "age"    "g
ender"
```

```
torontoGroups
```

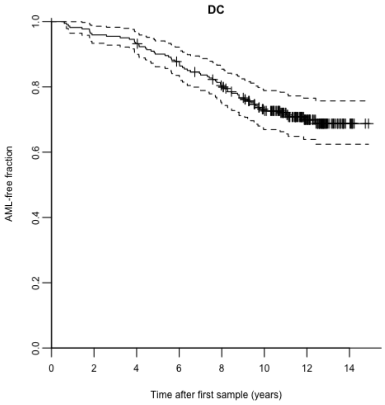
```
##  [1] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
##  [9] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
## [17] Genes        Genes        Genes        Demographics Demographics
## Levels: Genes Demographics
```

Manually standardize age and mutation VAFs

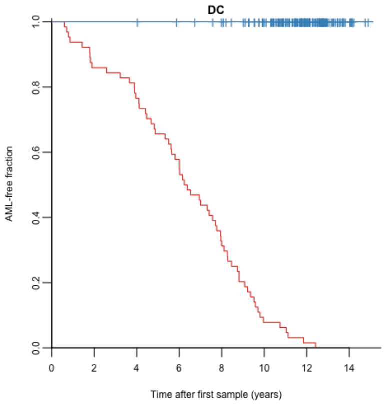
```
torontoX$age <- torontoX$age/10
names(torontoX)[which(names(torontoX)=="age")] <- "age_10"
g <- torontoGroups == "Genes"
torontoX[,g] <- torontoX[,g]*10
names(torontoX)[g] <- paste(names(torontoX)[g], "0.1",sep="_")
colnames(torontoX)
```

```
##  [1] "ASXL1_0.1"  "CALR_0.1"   "CBL_0.1"    "DNMT3A_0.1" "IDH1_0.1"   "IDH2_0.1
"   "JAK2_0.1"   "KDM6A_0.1" 
##  [9] "KMT2C_0.1"  "KRAS_0.1"   "NF1_0.1"    "PHF6_0.1"   "PTPN11_0.1" "RUNX1_0.
1"  "SF3B1_0.1"  "SRSF2_0.1" 
## [17] "TET2_0.1"   "TP53_0.1"   "U2AF1_0.1"  "age_10"     "gender"
```

```
torontoSurv <- Surv(torontoData$fu_years, torontoData$Diagnosis=="AML")
plot(survfit(torontoSurv~ 1), col= "black", main = "DC", xlab='Time after first sa
mple (years)', ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.t
ime = T)
```



```
plot(survfit(torontoSurv ~ torontoData$Diagnosis), xlab='Time after first sample (
years)', main = "DC", ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01),
mark.time = T, col = set1[1:2])
```



12.2 Validation cohort

```
f = "data/VC_vaf_matrix_no_duplicates_262ctrl_29aml_nodates.csv"
sangerData <- read.csv(f)
dim(sangerData)
```

```
## [1] 445  43
```

```
sangerData <- sangerData[rowSums(sangerData[, colnames(sangerData) %in% gene_vars]
)>0, ]
dim(sangerData)
```

```
## [1] 149  43
```

```
sangerData$hcdate <- as.Date(sangerData$hcdate)
sangerData$dodx <- as.Date(sangerData$dodx)

sangerPatients <- sub("[a-z]+$","", sangerData$Sample)
o <- order(sangerPatients, as.numeric(sangerData$hcdate))

sangerData <- sangerData[o,]
sangerPatients <- sangerPatients[o]

clinical_vars <- c("systol", "diastol", "bmi", "cholestl", "triglyc", "hdl", "ldl"
, "lym", "mcv", "rdw", "wbc", "plt", "hgb")
sangerX <- sangerData[, colnames(sangerData) %in% c(gene_vars, "age","gender",clin
ical_vars)]
sangerX <- as.data.frame(sangerX)

sangerX <- sangerX[,colSums(sangerX != 0,na.rm=TRUE)>=thr]
sangerGroups <- factor(grepl("^[a-z]", colnames(sangerX))*2, levels=0:2, labels=c(
"Genes", "Demographics", "Blood"))
sangerGroups[names(sangerX) %in% c("age","gender")] <- "Demographics"
table(sangerGroups)
```

```
## sangerGroups
##        Genes Demographics        Blood 
##           15            2           13
```

```
colnames(sangerX)
```

```
##  [1] "ASXL1"    "CBL"      "DNMT3A"   "JAK2"     "KMT2C"    "KMT2D"    "KRAS"    
"NF1"      "NRAS"     "RAD21"   
## [11] "SF3B1"    "SRSF2"    "TET2"     "TP53"     "U2AF1"    "age"      "gender"  
"systol"   "diastol"  "bmi"     
## [21] "cholestl" "triglyc"  "hdl"      "ldl"      "lym"      "mcv"      "rdw"     
"wbc"      "plt"      "hgb"
```

```
sangerGroups
```

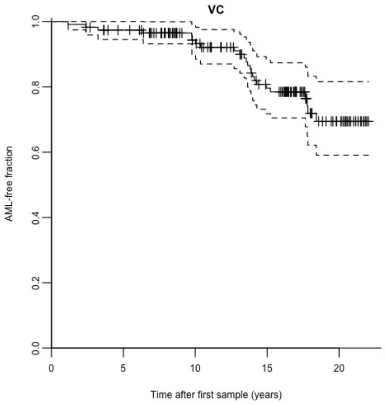
```
##  [1] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Genes       
##  [9] Genes        Genes        Genes        Genes        Genes        Genes       
Genes        Demographics
## [17] Demographics Blood        Blood        Blood        Blood        Blood       
Blood        Blood       
## [25] Blood        Blood        Blood        Blood        Blood        Blood       
## Levels: Genes Demographics Blood
```

```r
g <- sangerGroups=="Genes"
sangerX[g] <- sangerX[g] * 10
names(sangerX)[g] <- paste(names(sangerX[g]),"0.1", sep="_")
y <- StandardizeMagnitude(sangerX[!g])
sangerX <- cbind(sangerX[g],y)

poorMansImpute <- function(x) {x[is.na(x)] <- mean(x, na.rm=TRUE); return(x)}
sangerX <- as.data.frame(sapply(sangerX, poorMansImpute))

foo <- split(sangerData[,c("Diagnosis","hcdate","dodx")], sangerPatients)

bar <- do.call("rbind",lapply(foo, function(x){
  y <- x
  n <- nrow(y)
  y[-n,"Diagnosis"] <- "Control"
  start <- as.numeric(y$hcdate - y$hcdate[1])/365.25
  end <- c(as.numeric(y$hcdate - y$hcdate[1])[-1]/365.25, as.numeric(y$dodx[n] - y
$hcdate[1])/365.25)
  return(data.frame(Diagnosis=y[,"Diagnosis"], start=start, end=end))
}))

bar[1:10, ]
```

	Diagnosis <fctr>	start <dbl>	end <dbl>
PD29762	AML	0.000000	9.754962
PD29764	AML	0.000000	10.360027
PD29792	AML	0.000000	14.108145
PD29810	Control	0.000000	18.573580
PD29836.1	Control	0.000000	2.414784
PD29836.2	AML	2.414784	10.023272
PD29856	AML	0.000000	17.828884
PD29896	AML	0.000000	6.387406
PD29918.1	Control	0.000000	5.442847
PD29918.2	AML	5.442847	13.396304

1-10 of 10 rows

```r
sangerPatientsSplit <- unlist(sapply(names(foo), function(n) rep(n, nrow(foo[[n]])
)))

sangerSurv <- Surv(time = bar$start, time2 = bar$end, event = bar$Diagnosis!="Cont
rol", origin = 0)

plot(survfit(sangerSurv~ 1), col= "black", main = "VC", xlab='Time after first sam
ple (years)', ylab='AML-free fraction', bty='L', yaxs='i', ylim=c(0,1.01), mark.ti
me = T) #mark = 1
```



12.3 Expected AML incidence

Validation cohort

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
sangerSurv2 <- Surv(sangerSurv[w,2], sangerSurv[w,3])

expected_rate_sanger_cr <- mean(aml_inc_cr(sangerX[w,"gender"],sangerX[w,"age_10"]
*10, sangerX[w,"age_10"]*10+ pmax(1,sangerSurv2[,1]))[!sangerSurv2[,2]])

n_total_sanger <- sum(sangerSurv2[,2])/expected_rate_sanger_cr
n_total_sanger
```

```
## [1] 9216.197
```

Discovery cohort

```
expected_rate_toronto_cr <- mean(aml_inc_cr(torontoX[,"gender"],torontoX[,"age_10"
]*10, torontoX[,"age_10"]*10+ pmax(1,torontoSurv[,1]))[!torontoSurv[,2]])

n_total_toronto <- sum(torontoSurv[,2])/expected_rate_toronto_cr
n_total_toronto
```

```
## [1] 42940.66
```

12.4 Combined data

Survival

```
allSurv <- rbind(sangerSurv, Surv(rep(0, nrow(torontoSurv)), torontoSurv[,1], toro
ntoSurv[,2]))
allSurv <- Surv(allSurv[,1], allSurv[,2], allSurv[,3])
```

Data matrix

```
cohort <- c(rep("Sanger", nrow(sangerX)), rep("Toronto", nrow(torontoX)))
i <- c(sort(setdiff(gene_vars,"CALR")),"age","gender")
allX <- rbind(superSet(sangerData,i,fill=0), superSet(torontoData,i,fill=0))
allX <- allX[,colSums(allX>0)>=thr]
allX <- cbind(allX, cohort=cohort=="Sanger") + 0
allGroups <- factor(grepl("^[A-Z]",colnames(allX))+0, levels=1:0, labels=c("Genes"
,"Demographics"))

g <- allGroups=="Genes"
allX <- cbind(10*allX[,g], StandardizeMagnitude(allX[,!g]))
colnames(allX)[g] <- paste(colnames(allX)[g],"0.1",sep="_")
control <- c(sangerData$Diagnosis=="Control", torontoData$Diagnosis=="Control")
```

Weights

```
weights <- rep(1, nrow(allX))
weights[cohort=="Sanger" & control] <- n_total_sanger/sum(cohort=="Sanger" & contr
ol & allSurv[,1]==0)
weights[cohort=="Toronto" & control] <- n_total_toronto/sum(cohort=="Toronto" & co
ntrol)

n_total <- n_total_sanger + n_total_toronto
n_total
```

```
## [1] 52156.85
```

12.5 Coxph model fits

```
sigma0 <- 0.1
nu <- 1
which.mu <- "Genes"
```

12.5.1 Toronto

12.5.1.1 Raw

```
fitToronto <- CoxRFX(torontoX, torontoSurv, groups=torontoGroups, which.mu=which.m
u, nu=nu, sigma0=sigma0)
```

```
u, nu=nu, sigma0=sigma0)
waldToronto <- WaldTest(fitToronto)
```

```
##                   group    coef   coef-mu     sd      z df  p.value sig
## ASXL1_0.1         Genes  0.5750  0.032700 0.1158  4.964  1 6.91e-07 ***
## CALR_0.1          Genes  0.5200 -0.022339 0.0744  6.990  1 2.74e-12 ***
## CBL_0.1           Genes  0.4268 -0.115522 0.1231  3.469  1 5.23e-04 ***
## DNMT3A_0.1        Genes  0.4724 -0.069936 0.1062  4.448  1 8.66e-06 ***
## IDH1_0.1          Genes  0.5730  0.030722 0.1188  4.822  1 1.42e-06 ***
## IDH2_0.1          Genes  0.4711 -0.071177 0.1126  4.184  1 2.86e-05 ***
## JAK2_0.1          Genes  0.6084  0.066072 0.1214  5.011  1 5.43e-07 ***
## KDM6A_0.1         Genes  0.5420 -0.000284 0.0628  8.629  1 6.17e-18 ***
## KMT2C_0.1         Genes  0.5603  0.017953 0.0656  8.545  1 1.29e-17 ***
## KRAS_0.1          Genes  0.5394 -0.002952 0.0628  8.583  1 9.20e-18 ***
## NF1_0.1           Genes  0.5404 -0.001954 0.0628  8.599  1 8.07e-18 ***
## PHF6_0.1          Genes  0.5469  0.004542 0.0632  8.655  1 4.91e-18 ***
## PTPN11_0.1        Genes  0.5556  0.013243 0.0631  8.810  1 1.25e-18 ***
## RUNX1_0.1         Genes  0.3347 -0.207621 0.0917  3.650  1 2.62e-04 ***
## SF3B1_0.1         Genes  0.6532  0.110858 0.0963  6.781  1 1.19e-11 ***
## SRSF2_0.1         Genes  0.4370 -0.105330 0.0920  4.750  1 2.03e-06 ***
## TET2_0.1          Genes  0.5053 -0.037059 0.1248  4.050  1 5.12e-05 ***
## TP53_0.1          Genes  0.7280  0.185639 0.0825  8.828  1 1.07e-18 ***
## U2AF1_0.1         Genes  0.7148  0.172443 0.0805  8.879  1 6.76e-19 ***
## age_10     Demographics -0.0236 -0.023625 0.1092 -0.216  1 8.29e-01
## gender     Demographics -0.0832 -0.083228 0.1113 -0.748  1 4.55e-01
```

```
survConcordance(fitToronto$surv ~ fitToronto$linear.predictors)
```

```
## Call:
## survConcordance(formula = fitToronto$surv ~ fitToronto$linear.predictors)
## 
##   n= 221 
## Concordance= 0.7806171 se= 0.03687602
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  8981.0000  2524.0000     0.0000     1.0000   848.5173
```

12.5.1.2 Adjusted

```
fitWeightedToronto <- CoxRFX(torontoX, torontoSurv, torontoGroups, which.mu=which.
mu, sigma0=sigma0, nu=nu, weights=weights[cohort=="Toronto"])
waldWeightedToronto <- WaldTest(fitWeightedToronto)
```

```
##                   group    coef  coef-mu    sd      z df  p.value sig
## ASXL1_0.1         Genes  1.9878  0.06756 0.150 13.267  1 3.60e-40 ***
## CALR_0.1          Genes  0.6189 -1.30126 0.758  0.817  1 4.14e-01    
## CBL_0.1           Genes  0.2531 -1.66705 0.379  0.668  1 5.04e-01    
## DNMT3A_0.1        Genes  0.5859 -1.33434 0.136  4.313  1 1.61e-05 ***
## IDH1_0.1          Genes  2.4124  0.49218 0.341  7.083  1 1.41e-12 ***
## IDH2_0.1          Genes  0.8067 -1.11352 0.231  3.498  1 4.70e-04 ***
## JAK2_0.1          Genes  1.9535  0.03333 0.193 10.131  1 4.01e-24 ***
## KDM6A_0.1         Genes  1.9181 -0.00209 0.163 11.792  1 4.31e-32 ***
## KMT2C_0.1         Genes  2.3735  0.45328 0.730  3.250  1 1.16e-03  **
## KRAS_0.1          Genes  1.7434 -0.17684 0.195  8.955  1 3.38e-19 ***
## NF1_0.1           Genes  1.8059 -0.11434 0.190  9.518  1 1.77e-21 ***
## PHF6_0.1          Genes  2.2276  0.30741 0.144 15.462  1 6.24e-54 ***
## PTPN11_0.1        Genes  2.5970  0.67679 0.277  9.366  1 7.52e-21 ***
## RUNX1_0.1         Genes  0.7172 -1.20303 0.137  5.235  1 1.65e-07 ***
## SF3B1_0.1         Genes  3.2528  1.33260 0.321 10.149  1 3.36e-24 ***
## SRSF2_0.1         Genes  1.4698 -0.45035 0.170  8.656  1 4.91e-18 ***
## TET2_0.1          Genes  0.5707 -1.34952 0.211  2.699  1 6.96e-03  **
## TP53_0.1          Genes  5.2413  3.32111 0.440 11.916  1 9.82e-33 ***
## U2AF1_0.1         Genes  3.9483  2.02809 0.365 10.817  1 2.87e-27 ***
## age_10     Demographics -0.0820 -0.08201 0.117 -0.700  1 4.84e-01    
## gender     Demographics -0.0899 -0.08989 0.117 -0.771  1 4.41e-01
```

```
survConcordance(fitWeightedToronto$surv ~ fitWeightedToronto$linear.predictors, we
ights=weights[cohort=="Toronto"])
```

```
## Call:
## survConcordance(formula = fitWeightedToronto$surv ~ fitWeightedToronto$linear.p
redictors,
##     weights = weights[cohort == "Toronto"])
## 
##   n= 221 
## Concordance= 0.8454794 se= 0.03633541
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  2196217.1   401382.8        0.0        1.0   188769.7
```

Uno’s estimator of cumulative/dynamic AUC

```
a <- AUC.uno(torontoSurv, torontoSurv, fitWeightedToronto$linear.predictors, times
= seq(0,12, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.791
```

12.5.2 Validation cohort

12.5.2.1 Raw

```
fitSanger <- CoxRFX(sangerX, sangerSurv, groups=sangerGroups, which.mu=which.mu, n
u=nu, sigma0=sigma0)
waldSanger <- WaldTest(fitSanger)
```

```
##                      group      coef   coef-mu      sd        z df  p.value sig
## ASXL1_0.1            Genes  0.673478  0.158950 0.12882  5.22794  1 1.71e-07 ***
## CBL_0.1              Genes  0.495353 -0.019175 0.10735  4.61426  1 3.94e-06 ***
## DNMT3A_0.1           Genes  0.328415 -0.186113 0.13178  2.49210  1 1.27e-02   *
## JAK2_0.1             Genes  0.493355 -0.021173 0.11739  4.20278  1 2.64e-05 ***
## KMT2C_0.1            Genes  0.519077  0.004549 0.10042  5.16888  1 2.36e-07 ***
## KMT2D_0.1            Genes  0.341708 -0.172820 0.16670  2.04989  1 4.04e-02   *
## KRAS_0.1             Genes  0.517799  0.003272 0.09650  5.36592  1 8.05e-08 ***
## NF1_0.1              Genes  0.501902 -0.012625 0.09919  5.06022  1 4.19e-07 ***
## NRAS_0.1             Genes  0.534425  0.019897 0.09703  5.50790  1 3.63e-08 ***
## RAD21_0.1            Genes  0.503868 -0.010660 0.09793  5.14544  1 2.67e-07 ***
## SF3B1_0.1            Genes  0.507855 -0.006673 0.09801  5.18184  1 2.20e-07 ***
## SRSF2_0.1            Genes  0.529928  0.015400 0.14168  3.74021  1 1.84e-04 ***
## TET2_0.1             Genes  0.593720  0.079192 0.12273  4.83743  1 1.32e-06 ***
## TP53_0.1             Genes  0.584538  0.070010 0.09773  5.98121  1 2.21e-09 ***
## U2AF1_0.1            Genes  0.592496  0.077968 0.09770  6.06442  1 1.32e-09 ***
## age_10      Demographics  0.084731  0.084731 0.12166  0.69645  1 4.86e-01    
## gender      Demographics -0.007960 -0.007960 0.10340 -0.07698  1 9.39e-01    
## systol_100           Blood  0.033564  0.033564 0.03644  0.92111  1 3.57e-01    
## diastol_100          Blood  0.032432  0.032432 0.02299  1.41095  1 1.58e-01    
## bmi_10               Blood  0.081752  0.081752 0.06892  1.18610  1 2.36e-01    
## cholestl_10          Blood  0.014082  0.014082 0.01344  1.04742  1 2.95e-01    
## triglyc              Blood -0.000827 -0.000827 0.10813 -0.00765  1 9.94e-01    
## hdl                  Blood -0.007587 -0.007587 0.06927 -0.10952  1 9.13e-01    
## ldl                  Blood  0.134372  0.134372 0.11043  1.21684  1 2.24e-01    
## lym                  Blood  0.076500  0.076500 0.08867  0.86278  1 3.88e-01    
## mcv_100              Blood -0.012801 -0.012801 0.00713 -1.79436  1 7.28e-02   .
## rdw_10               Blood  0.058557  0.058557 0.01828  3.20254  1 1.36e-03  **
## wbc_10               Blood  0.016691  0.016691 0.03908  0.42707  1 6.69e-01    
## plt_100              Blood  0.095820  0.095820 0.09229  1.03821  1 2.99e-01    
## hgb_10               Blood  0.006904  0.006904 0.01981  0.34856  1 7.27e-01
```

```
survConcordance(sangerSurv ~ fitSanger$linear.predictors)
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitSanger$linear.predictors)
## 
##   n= 149 
## Concordance= 0.7918502 se= 0.06247796
## concordant discordant  tied.risk  tied.time   std(c-d) 
##    1438.00     378.00       0.00       0.00     226.92
```

12.5.2.2 Adjusted

```
fitWeightedSanger <- CoxRFX(sangerX, sangerSurv, sangerGroups, which.mu=which.mu,
sigma0=sigma0, nu=nu, weights=weights[cohort=="Sanger"])
waldWeightedSanger <- WaldTest(fitWeightedSanger)
```

```
##                      group    coef coef-mu     sd       z df  p.value sig
## ASXL1_0.1          Genes  3.2736  1.1639 0.5035  6.5016  1 7.95e-11 ***
## CBL_0.1            Genes  0.4415 -1.6682 1.4885  0.2966  1 7.67e-01
## DNMT3A_0.1         Genes  0.5963 -1.5134 0.2434  2.4497  1 1.43e-02   *
## JAK2_0.1           Genes -0.0225 -2.1322 1.0506 -0.0214  1 9.83e-01
## KMT2C_0.1          Genes  0.8233 -1.2864 1.4975  0.5498  1 5.82e-01
## KMT2D_0.1          Genes -0.1936 -2.3033 0.9186 -0.2108  1 8.33e-01
## KRAS_0.1           Genes  2.6546  0.5449 0.6402  4.1468  1 3.37e-05 ***
## NF1_0.1            Genes  0.8839 -1.2258 1.4275  0.6192  1 5.36e-01
## NRAS_0.1           Genes  4.8796  2.7699 0.6294  7.7532  1 8.96e-15 ***
## RAD21_0.1          Genes  0.8665 -1.2432 1.4103  0.6144  1 5.39e-01
## SF3B1_0.1          Genes  1.2701 -0.8396 1.4768  0.8601  1 3.90e-01
## SRSF2_0.1          Genes  1.6909 -0.4188 0.2626  6.4399  1 1.20e-10 ***
## TET2_0.1           Genes  1.3640 -0.7457 0.1595  8.5534  1 1.19e-17 ***
## TP53_0.1           Genes  5.1102  3.0005 1.0728  4.7634  1 1.90e-06 ***
## U2AF1_0.1          Genes  8.0069  5.8972 0.9739  8.2214  1 2.01e-16 ***
## age_10      Demographics -0.0522 -0.0522 0.1212 -0.4306  1 6.67e-01
## gender      Demographics -0.0216 -0.0216 0.0988 -0.2185  1 8.27e-01
## systol_100         Blood  0.0064  0.0064 0.0409  0.1566  1 8.76e-01
## diastol_100        Blood  0.0251  0.0251 0.0269  0.9320  1 3.51e-01
## bmi_10             Blood  0.0956  0.0956 0.0826  1.1574  1 2.47e-01
## cholestl_10        Blood  0.0143  0.0143 0.0155  0.9246  1 3.55e-01
## triglyc            Blood -0.0533 -0.0533 0.1279 -0.4169  1 6.77e-01
## hdl                Blood -0.0505 -0.0505 0.0839 -0.6015  1 5.48e-01
## ldl                Blood  0.2011  0.2011 0.1239  1.6229  1 1.05e-01
## lym                Blood  0.0499  0.0499 0.0996  0.5009  1 6.16e-01
## mcv_100            Blood -0.0238 -0.0238 0.0075 -3.1777  1 1.48e-03  **
## rdw_10             Blood  0.0832  0.0832 0.0142  5.8698  1 4.36e-09 ***
## wbc_10             Blood  0.0108  0.0108 0.0544  0.1988  1 8.42e-01
## plt_100            Blood  0.1509  0.1509 0.1056  1.4297  1 1.53e-01
## hgb_10             Blood -0.0224 -0.0224 0.0217 -1.0308  1 3.03e-01
```

```
survConcordance(sangerSurv ~ fitWeightedSanger$linear.predictors, weights=weights[
cohort=="Sanger"])
```

```
## Call:
## survConcordance(formula = sangerSurv ~ fitWeightedSanger$linear.predictors, 
##     weights = weights[cohort == "Sanger"])
## 
##   n= 149 
## Concordance= 0.8671072 se= 0.06105924
## concordant discordant  tied.risk  tied.time   std(c-d) 
##  135478.93   20763.49       0.00       0.00   19080.09
```

Uno’s estimator of cumulative/dynamic AUC

```
w <- c(which(sangerSurv[,1]==0)[-1]-1, nrow(sangerSurv))
s <- Surv(sangerSurv[w,2], sangerSurv[w,3])
a <- AUC.uno(s, s, fitWeightedSanger$linear.predictors[w], times= c(0, 22, 0.1))
round(a$iauc, digits = 3)
```

```
## [1] 0.587
```

13 Session

```
devtools::session_info()
```

```
## Session info ------------------------------------------------------------------
----------------------------------------
```

```
##  setting  value                       
##  version  R version 3.5.1 (2018-07-02)
##  system   x86_64, darwin17.6.0        
##  ui       X11                         
##  language (EN)                        
##  collate  C                           
##  tz       Europe/London               
##  date     2018-07-24
```

```
## Packages ----------------------------------------------------------------------
----------------------------------------
```

```
##  package      * version  date       source        
##  abind          1.4-5    2016-07-21 CRAN (R 3.5.1)
##  assertthat     0.2.0    2017-04-11 CRAN (R 3.5.1)
##  backports      1.1.2    2017-12-13 cran (@1.1.2)
##  base         * 3.5.1    2018-07-09 local
##  bindr          0.1.1    2018-03-13 CRAN (R 3.5.1)
##  bindrcpp       0.2.2    2018-03-29 CRAN (R 3.5.1)
##  bitops         1.0-6    2013-08-17 CRAN (R 3.5.1)
##  broom          0.5.0    2018-07-17 cran (@0.5.0)
##  car            3.0-0    2018-04-02 CRAN (R 3.5.1)
##  carData        3.0-1    2018-03-28 CRAN (R 3.5.1)
##  caTools        1.17.1.1 2018-07-20 CRAN (R 3.5.1)
##  cellranger     1.1.0    2016-07-27 CRAN (R 3.5.1)
##  codetools      0.2-15   2016-10-05 CRAN (R 3.5.1)
##  compiler       3.5.1    2018-07-09 local
##  CoxHD        * 0.0.73   2018-07-23 Github (gerstung-lab/CoxHD@bc60c16)
##  crayon         1.3.4    2017-09-16 CRAN (R 3.5.1)
##  curl           3.2      2018-03-28 CRAN (R 3.5.1)
##  data.table     1.11.4   2018-05-27 CRAN (R 3.5.1)
##  datasets     * 3.5.1    2018-07-09 local
##  devtools       1.13.6   2018-06-27 CRAN (R 3.5.1)
##  digest         0.6.15   2018-01-28 CRAN (R 3.5.1)
##  dplyr        * 0.7.6    2018-06-29 CRAN (R 3.5.1)
##  evaluate       0.11     2018-07-17 CRAN (R 3.5.1)
##  forcats        0.3.0    2018-02-19 cran (@0.3.0)
##  foreach      * 1.4.4    2017-12-12 CRAN (R 3.5.1)
##  foreign        0.8-71   2018-07-20 CRAN (R 3.5.1)
##  gdata          2.18.0   2017-06-06 CRAN (R 3.5.1)
##  glmnet       * 2.0-16   2018-04-02 CRAN (R 3.5.1)
##  glue           1.3.0    2018-07-17 CRAN (R 3.5.1)
##  gplots       * 3.0.1    2016-03-30 CRAN (R 3.5.1)
##  graphics     * 3.5.1    2018-07-09 local
##  grDevices    * 3.5.1    2018-07-09 local
##  grid           3.5.1    2018-07-09 local
##  gtools         3.8.1    2018-06-26 CRAN (R 3.5.1)
##  haven          1.1.2    2018-06-27 cran (@1.1.2)
##  hms            0.4.2    2018-03-10 CRAN (R 3.5.1)
##  htmltools      0.3.6    2017-04-28 CRAN (R 3.5.1)
##  iterators      1.0.10   2018-07-13 CRAN (R 3.5.1)
##  jomo           2.6-2    2018-04-26 cran (@2.6-2)
##  jsonlite       1.5      2017-06-01 CRAN (R 3.5.1)
##  KernSmooth     2.23-15  2015-06-29 CRAN (R 3.5.1)
##  knitr        * 1.20     2018-02-20 CRAN (R 3.5.1)
##  lattice        0.20-35  2017-03-25 CRAN (R 3.5.1)
##  lme4           1.1-17   2018-04-03 cran (@1.1-17)
##  magrittr       1.5      2014-11-22 CRAN (R 3.5.1)
##  MASS           7.3-50   2018-04-30 cran (@7.3-50)
##  Matrix       * 1.2-14   2018-04-09 CRAN (R 3.5.1)
##  memoise        1.1.0    2017-04-21 CRAN (R 3.5.1)
##  methods      * 3.5.1    2018-07-09 local
##  mg14           0.0.5    2018-07-23 Github (mg14/mg14@6a63283)
##  mice           3.1.0    2018-06-20 cran (@3.1.0)
##  minqa          1.2.4    2014-10-09 cran (@1.2.4)
##  mitml          0.3-6    2018-07-10 cran (@0.3-6)
##  mvtnorm        1.0-8    2018-05-31 cran (@1.0-8)
##  nlme           3.1-137  2018-04-07 cran (@3.1-137)
##  nloptr         1.0.4    2017-08-22 cran (@1.0.4)
##  nnet           7.3-12   2016-02-02 cran (@7.3-12)
##  openxlsx       4.1.0    2018-05-26 CRAN (R 3.5.1)
##  pan            1.6      2018-06-29 cran (@1.6)
##  parallel     * 3.5.1    2018-07-09 local
##  pillar         1.3.0    2018-07-14 CRAN (R 3.5.1)
##  pkgconfig      2.0.1    2017-03-21 CRAN (R 3.5.1)
##  purrr          0.2.5    2018-05-29 CRAN (R 3.5.1)
##  R6             2.2.2    2017-06-17 CRAN (R 3.5.1)
##  RColorBrewer * 1.1-2    2014-12-07 CRAN (R 3.5.1)
##  Rcpp           0.12.18  2018-07-23 CRAN (R 3.5.1)
##  readr        * 1.1.1    2017-05-16 CRAN (R 3.5.1)
##  readxl         1.1.0    2018-04-20 CRAN (R 3.5.1)
##  rio            0.5.10   2018-03-29 CRAN (R 3.5.1)
##  rj           * 2.0.5-2  2018-07-23 local
##  rj.gd          2.0.0-1  2018-07-23 local
##  rlang          0.2.1    2018-05-30 CRAN (R 3.5.1)
##  rmarkdown      1.10     2018-06-11 CRAN (R 3.5.1)
##  ROCR         * 1.0-7    2015-03-26 CRAN (R 3.5.1)
##  rpart          4.1-13   2018-02-23 cran (@4.1-13)
##  rprojroot      1.3-2    2018-01-03 CRAN (R 3.5.1)
##  splines        3.5.1    2018-07-09 local
##  stats        * 3.5.1    2018-07-09 local
##  stringi        1.2.4    2018-07-20 CRAN (R 3.5.1)
##  stringr      * 1.3.1    2018-05-10 CRAN (R 3.5.1)
##  survAUC      * 1.0-5    2012-09-04 CRAN (R 3.5.1)
##  survival     * 2.42-6   2018-07-13 CRAN (R 3.5.1)
##  survivalROC  * 1.0.3    2013-01-13 CRAN (R 3.5.1)
##  tibble         1.4.2    2018-01-22 CRAN (R 3.5.1)
##  tidyr          0.8.1    2018-05-18 cran (@0.8.1)
##  tidyselect     0.2.4    2018-02-26 CRAN (R 3.5.1)
##  tools          3.5.1    2018-07-09 local
##  utils        * 3.5.1    2018-07-09 local
##  withr          2.1.2    2018-03-15 CRAN (R 3.5.1)
##  yaml           2.1.19   2018-05-01 CRAN (R 3.5.1)
##  zip            1.0.0    2017-04-25 CRAN (R 3.5.1)
```

This code and all data necessary to execute it is available from http://www.github.com/gerstung-lab/ (http://www.github.com/gerstung-lab/)

Appendix 8: Mutations in discovery cohort pre-AML and control samples

Sample ID	Type	Chromosome	Position	WT	MT	VAF	Gene	Protein	Effect	Group
EPIC_0001	indel	2	25463314	TGCCCTC	-	0.0119	DNMT3A	p.?	Essential splice	Control
EPIC_0001	sub	2	25463541	G	C	0.0058	DNMT3A	p.S714C	Missense	Control
EPIC_0003	sub	2	25469038	G	C	0.0091	DNMT3A	p.R474G	Missense	Control
EPIC_0003	sub	2	25470581	C	T	0.0048	DNMT3A	p.G298E	Missense	Control
EPIC_0005	sub	17	7578394	T	C	0.1298	TP53	p.H179R	Missense	Pre-AML
EPIC_0005	sub	2	25469542	C	T	0.0105	DNMT3A	p.W409*	Nonsense	Pre-AML
EPIC_0007	sub	2	25467408	C	T	0.0139	DNMT3A	p.?	Essential splice	Control
EPIC_0007	sub	4	106197285	T	C	0.0076	TET2	p.I1873T	Missense	Control
EPIC_0014	sub	2	25467408	C	T	0.0479	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0020	sub	2	25469632	C	T	0.0043	DNMT3A	p.R379H	Missense	Control
EPIC_0022	sub	2	25467448	C	A	0.0177	DNMT3A	p.G543V	Missense	Control
EPIC_0024	sub	2	25466797	C	A	0.0271	DNMT3A	p.V636L	Missense	Control
EPIC_0027	sub	2	25459806	T	G	0.0039	DNMT3A	p.K826T	Missense	Control
EPIC_0028	sub	4	106190775	T	A	0.0123	TET2	p.Y1351*	Nonsense	Control
EPIC_0032	sub	2	25457231	G	A	0.0955	DNMT3A	p.Q886*	Nonsense	Control
EPIC_0034	sub	20	31024116	C	T	0.0032	ASXL1	p.Q1201*	Nonsense	Control
EPIC_0034	indel	4	106196981	ATGTTCA	-	0.0100	TET2	1772_F1773de	Inframe	Control
EPIC_0039	sub	2	25464433	G	A	0.0049	DNMT3A	p.H694Y	Missense	Control
EPIC_0039	sub	20	31022592	C	T	0.0039	ASXL1	p.R693*	Nonsense	Control
EPIC_0040	sub	11	119148930	T	C	0.0035	CBL	p.C384R	Missense	Pre-AML
EPIC_0040	sub	2	25463286	C	T	0.0144	DNMT3A	p.R736H	Missense	Pre-AML
EPIC_0043	sub	2	25469539	G	A	0.0092	DNMT3A	p.A410V	Missense	Control
EPIC_0044	indel	17	7578390	GTGGGGGCAGCGCCTCACAAC	-	0.0099	TP53	p.T170fs*5	Frameshift	Pre-AML
EPIC_0044	sub	21	44524456	G	A	0.0056	U2AF1	p.S34F	Missense	Pre-AML
EPIC_0049	sub	2	25457176	G	A	0.0096	DNMT3A	p.P904L	Missense	Control
EPIC_0051	sub	9	5073770	G	T	0.4345	JAK2	p.V617F	Missense	Pre-AML
EPIC_0051	sub	X	133551305	T	C	0.0101	PHF6	p.I314T	Missense	Pre-AML
EPIC_0053	sub	2	25467023	C	T	0.0410	DNMT3A	p.?	Essential splice	Control
EPIC_0054	sub	12	25398281	C	T	0.0062	KRAS	p.G13D	Missense	Control
EPIC_0056	sub	2	25464576	C	T	0.0087	DNMT3A	p.G646E	Missense	Control
EPIC_0056	sub	2	25470011	A	T	0.0047	DNMT3A	p.L344Q	Missense	Control
EPIC_0058	sub	11	119149287	A	G	0.0102	CBL	p.D432G	Missense	Control
EPIC_0059	sub	2	25463596	G	A	0.0030	DNMT3A	p.Q696*	Nonsense	Control
EPIC_0059	sub	X	44918491	G	A	0.0097	KDM6A	p.?	Essential splice	Control
EPIC_0062	indel	20	31022403	ACCACTGCCATAGAGAGGCGG	-	0.1784	ASXL1	p.H630fs*66	Frameshift	Pre-AML
EPIC_0062	sub	21	36164601	G	A	0.5874	RUNX1	p.P425L	Missense	Pre-AML
EPIC_0062	indel	21	36252852	-	CCT	0.0198	RUNX1	p.?	Essential splice	Pre-AML
EPIC_0064	sub	2	198266834	T	C	0.2949	SF3B1	p.K700E	Missense	Pre-AML
EPIC_0065	sub	2	25463563	C	G	0.0113	DNMT3A	p.G707R	Missense	Control
EPIC_0065	sub	4	106190882	A	T	0.0322	TET2	p.N1387I	Missense	Control
EPIC_0066	sub	2	25463239	A	G	0.0099	DNMT3A	p.F752L	Missense	Control
EPIC_0067	indel	20	31022403	ACCACTGCCATAGAGAGGCGG	-	0.0048	ASXL1	p.H630fs*66	Frameshift	Pre-AML
EPIC_0067	sub	20	31022838	T	A	0.0054	ASXL1	p.L775I	Missense	Pre-AML
EPIC_0067	sub	20	31022839	T	A	0.0021	ASXL1	p.L775*	Nonsense	Pre-AML
EPIC_0069	sub	4	106162529	A	C	0.0967	TET2	p.Y1148S	Missense	Control
EPIC_0071	sub	4	106193748	C	T	0.0063	TET2	p.R1404*	Nonsense	Control
EPIC_0073	sub	2	25462025	G	C	0.0048	DNMT3A	p.F794L	Missense	Control
EPIC_0074	indel	11	119149355	-	ATG	0.3287	CBL	p.Y455fs*16	Frameshift	Control
EPIC_0074	sub	2	25467442	T	C	0.0071	DNMT3A	p.E545G	Missense	Control
EPIC_0074	sub	2	25469647	T	C	0.0039	DNMT3A	p.?	Essential splice	Control
EPIC_0075	sub	2	25466799	C	T	0.0579	DNMT3A	p.R635Q	Missense	Pre-AML
EPIC_0075	sub	2	25470947	T	A	0.0398	DNMT3A	p.K272*	Nonsense	Pre-AML
EPIC_0075	sub	4	106180899	T	G	0.0055	TET2	p.F1309L	Missense	Pre-AML
EPIC_0076	sub	2	25462068	A	C	0.0023	DNMT3A	p.I780S	Missense	Control
EPIC_0076	sub	2	25463182	G	A	0.0131	DNMT3A	p.R771*	Nonsense	Control
EPIC_0076	sub	2	25470549	G	C	0.0048	DNMT3A	p.R309G	Missense	Control
EPIC_0081	sub	2	25469965	G	T	0.0570	DNMT3A	p.Y359*	Nonsense	Pre-AML
EPIC_0081	sub	20	31023395	G	A	0.0026	ASXL1	p.W960*	Nonsense	Pre-AML
EPIC_0082	indel	2	25463316	CC	-	0.1900	DNMT3A	p.G726fs*53	Frameshift	Control
EPIC_0084	sub	12	25398255	G	T	0.0059	KRAS	p.Q22K	Missense	Control
EPIC_0084	indel	19	13054605	GAG	-	0.0025	CALR	p.E378fs*10	Frameshift	Control
EPIC_0084	sub	4	106196306	C	T	0.0076	TET2	p.Q1547*	Nonsense	Control
EPIC_0090	sub	2	25463562	C	G	0.0042	DNMT3A	p.G707A	Missense	Control
EPIC_0090	sub	2	25467198	G	T	0.0028	DNMT3A	p.C559*	Nonsense	Control
EPIC_0090	sub	2	25470533	C	T	0.0277	DNMT3A	p.W314*	Nonsense	Control
EPIC_0095	sub	20	31023504	G	T	0.0031	ASXL1	p.E997*	Nonsense	Control
EPIC_0098	sub	2	25462086	T	G	0.0095	DNMT3A	p.?	Essential splice	Control
EPIC_0099	sub	17	7577580	T	C	0.0078	TP53	p.Y234C	Missense	Pre-AML
EPIC_0099	sub	17	7578555	C	T	0.0739	TP53	p.?	Essential splice	Pre-AML
EPIC_0100	sub	11	119148912	T	G	0.0069	CBL	p.F378V	Missense	Control
EPIC_0106	sub	20	31022853	C	T	0.0037	ASXL1	p.Q780*	Nonsense	Control
EPIC_0106	sub	4	106155612	C	A	0.0028	TET2	p.C171*	Nonsense	Control
EPIC_0111	sub	2	25467204	G	T	0.0145	DNMT3A	p.C557*	Nonsense	Control
EPIC_0112	sub	11	119148930	T	C	0.0022	CBL	p.C384R	Missense	Control
EPIC_0116	sub	4	106180849	A	T	0.0040	TET2	p.M1293L	Missense	Control
EPIC_0119	sub	17	29683508	C	G	0.0029	NF1	p.S2549*	Nonsense	Control
EPIC_0119	sub	4	106190798	G	C	0.0109	TET2	p.R1359P	Missense	Control
EPIC_0120	sub	4	106162529	A	G	0.0150	TET2	p.Y1148C	Missense	Control
EPIC_0123	indel	19	13054627	-	TTGTC	0.1380	CALR	p.K385fs*5	Frameshift	Control
EPIC_0125	sub	2	25466834	G	T	0.0110	DNMT3A	p.Y623*	Nonsense	Control
EPIC_0126	sub	2	25462021	C	A	0.0061	DNMT3A	p.G796C	Missense	Control
EPIC_0127	sub	2	25459829	A	T	0.0055	DNMT3A	p.C818*	Nonsense	Control
EPIC_0127	sub	2	25467504	A	T	0.0026	DNMT3A	p.C524*	Nonsense	Control
EPIC_0129	sub	2	25464531	A	T	0.0102	DNMT3A	p.I661N	Missense	Control
EPIC_0132	sub	2	25463179	A	G	0.0051	DNMT3A	p.F772L	Missense	Pre-AML
EPIC_0132	sub	21	36206716	G	A	0.4918	RUNX1	p.Q266*	Nonsense	Pre-AML
EPIC_0132	sub	4	106197287	G	C	0.0224	TET2	p.E1874Q	Missense	Pre-AML
EPIC_0132	sub	X	133551203	G	A	0.0063	PHF6	p.C280Y	Missense	Pre-AML
EPIC_0135	sub	4	106197296	A	G	0.0050	TET2	p.K1877E	Missense	Control
EPIC_0137	sub	4	106197296	A	G	0.0103	TET2	p.K1877E	Missense	Control
EPIC_0138	sub	4	106190827	T	C	0.0541	TET2	p.S1369P	Missense	Control
EPIC_0141	sub	17	74732959	G	C	0.3732	SRSF2	p.P95R	Missense	Pre-AML
EPIC_0141	sub	2	209113113	G	A	0.0318	IDH1	p.R132C	Missense	Pre-AML
EPIC_0141	sub	9	5073770	G	T	0.1759	JAK2	p.V617F	Missense	Pre-AML
EPIC_0142	sub	2	198267342	G	A	0.0023	SF3B1	p.A672V	Missense	Control
EPIC_0147	sub	17	74732959	G	C	0.2108	SRSF2	p.P95R	Missense	Pre-AML

EPIC_0147	indel	20	31022536	ACCCTGAG	-	0.0622	ASXL1	p.E676fs*25	Frameshift	Pre-AML
EPIC_0149	sub	4	106156747	C	T	0.0088	TET2	p.R550*	Nonsense	Control
EPIC_0152	sub	2	25463169	A	G	0.0020	DNMT3A	p.?	Essential splice	Control
EPIC_0152	sub	2	25466797	C	T	0.0131	DNMT3A	p.V636M	Missense	Control
EPIC_0156	sub	2	25463568	A	G	0.0113	DNMT3A	p.I705T	Missense	Control
EPIC_0158	sub	4	106156975	C	T	0.0019	TET2	p.Q626*	Nonsense	Control
EPIC_0165	indel	2	25458595	AT	-	0.0110	DNMT3A	p.L859fs*22	Frameshift	Pre-AML
EPIC_0165	sub	2	198267484	G	A	0.0377	SF3B1	p.R625C	Missense	Pre-AML
EPIC_0165	sub	4	106164020	T	G	0.0144	TET2	p.I1177S	Missense	Pre-AML
EPIC_0166	sub	4	106164084	G	T	0.0225	TET2	p.W1198C	Missense	Control
EPIC_0166	sub	4	106193801	C	G	0.0058	TET2	p.Y1421*	Nonsense	Control
EPIC_0169	sub	11	119148958	T	A	0.0018	CBL	p.I393N	Missense	Control
EPIC_0169	sub	2	25458579	T	A	0.1495	DNMT3A	p.E865V	Missense	Control
EPIC_0170	sub	2	25463298	A	C	0.0040	DNMT3A	p.F732C	Missense	Control
EPIC_0171	sub	21	44524456	G	A	0.0078	U2AF1	p.S34F	Missense	Pre-AML
EPIC_0174	sub	4	106164752	A	G	0.0029	TET2	p.E1207G	Missense	Control
EPIC_0175	sub	2	25467411	G	T	0.0059	DNMT3A	p.C555*	Nonsense	Control
EPIC_0176	sub	2	25463307	C	T	0.0039	DNMT3A	p.R729Q	Missense	Pre-AML
EPIC_0176	sub	2	25470516	G	A	0.0581	DNMT3A	p.R320*	Nonsense	Pre-AML
EPIC_0176	sub	20	31021187	C	T	0.0043	ASXL1	p.Q396*	Nonsense	Pre-AML
EPIC_0176	sub	21	44514777	T	C	0.0540	U2AF1	p.Q157R	Missense	Pre-AML
EPIC_0177	sub	20	31022839	T	G	0.0094	ASXL1	p.L775*	Nonsense	Control
EPIC_0177	sub	4	106193850	A	T	0.0033	TET2	p.K1438*	Nonsense	Control
EPIC_0181	sub	2	25470498	G	A	0.0048	DNMT3A	p.R326C	Missense	Control
EPIC_0184	sub	4	106180870	T	G	0.0061	TET2	p.F1300V	Missense	Control
EPIC_0184	sub	4	106190855	G	A	0.0237	TET2	p.C1378Y	Missense	Control
EPIC_0184	sub	4	106193751	G	T	0.0071	TET2	p.E1405*	Nonsense	Control
EPIC_0185	sub	4	106196627	C	T	0.0341	TET2	p.Q1654*	Nonsense	Control
EPIC_0186	sub	2	25459806	T	C	0.0037	DNMT3A	p.K826R	Missense	Control
EPIC_0186	sub	2	25463247	C	T	0.0199	DNMT3A	p.R749H	Missense	Control
EPIC_0186	sub	2	25464433	G	A	0.0044	DNMT3A	p.H694Y	Missense	Control
EPIC_0186	sub	2	25466812	T	C	0.0148	DNMT3A	p.R631G	Missense	Control
EPIC_0186	sub	2	25467059	G	A	0.0100	DNMT3A	p.Q606*	Nonsense	Control
EPIC_0191	sub	2	25459804	C	T	0.0999	DNMT3A	p.?	Essential splice	Control
EPIC_0194	indel	17	74732959	G	GGGC	0.2175	SRSF2	p.R94_P95insF	Inframe	Pre-AML
EPIC_0194	sub	4	106156747	C	T	0.0027	TET2	p.R550*	Nonsense	Pre-AML
EPIC_0194	sub	4	106164914	G	A	0.0051	TET2	p.R1261H	Missense	Pre-AML
EPIC_0194	sub	4	106193995	C	G	0.0039	TET2	p.S1486*	Nonsense	Pre-AML
EPIC_0195	sub	2	25469028	C	T	0.0184	DNMT3A	p.?	Essential splice	Control
EPIC_0196	indel	2	25457160	AA	-	0.0174	DNMT3A	p.F909fs*13	Frameshift	Control
EPIC_0196	sub	2	25464498	A	C	0.0080	DNMT3A	p.V672G	Missense	Control
EPIC_0196	sub	2	25468888	C	T	0.0078	DNMT3A	p.?	Essential splice	Control
EPIC_0197	sub	2	25470005	G	A	0.0087	DNMT3A	p.P346L	Missense	Control
EPIC_0202	sub	17	29663350	G	T	0.0050	NF1	p.?	Essential splice	Control
EPIC_0203	sub	2	25457243	G	A	0.0221	DNMT3A	p.R882C	Missense	Control
EPIC_0203	sub	2	25463284	G	A	0.0021	DNMT3A	p.L737F	Missense	Control
EPIC_0203	sub	2	25463579	G	C	0.0038	DNMT3A	p.F701L	Missense	Control
EPIC_0203	sub	2	25467523	T	C	0.0034	DNMT3A	p.?	Essential splice	Control
EPIC_0205	sub	4	106164824	T	C	0.0022	TET2	p.L1231P	Missense	Control
EPIC_0205	sub	4	106196213	C	T	0.0072	TET2	p.R1516*	Nonsense	Control
EPIC_0208	sub	11	119149238	T	A	0.0057	CBL	p.C416S	Missense	Control
EPIC_0208	sub	2	25470583	C	G	0.0025	DNMT3A	p.W297C	Missense	Control
EPIC_0208	sub	4	106180817	G	C	0.0040	TET2	p.G1282A	Missense	Control
EPIC_0209	sub	2	25462086	T	C	0.0084	DNMT3A	p.?	Essential splice	Control
EPIC_0212	sub	X	133559301	C	T	0.0119	PHF6	p.R347*	Nonsense	Pre-AML
EPIC_0213	indel	2	25463298	AAG	-	0.0041	DNMT3A	p.F732fs*1	Frameshift	Control
EPIC_0213	sub	2	25464463	C	A	0.0053	DNMT3A	p.V684F	Missense	Control
EPIC_0213	sub	20	31022592	C	T	0.0032	ASXL1	p.R693*	Nonsense	Control
EPIC_0215	sub	2	25466787	A	C	0.0050	DNMT3A	p.L639R	Missense	Control
EPIC_0218	sub	2	25469032	T	A	0.0018	DNMT3A	p.R476*	Nonsense	Control
EPIC_0219	sub	2	25459804	C	T	0.0037	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0220	sub	20	31024242	C	T	0.0022	ASXL1	p.Q1243*	Nonsense	Control
EPIC_0221	sub	2	25464483	T	C	0.0049	DNMT3A	p.H677R	Missense	Control
EPIC_0223	sub	1	115256535	G	T	0.0209	NRAS	p.A59D	Missense	Pre-AML
EPIC_0223	sub	12	112888148	A	G	0.0360	PTPN11	p.K55R	Missense	Pre-AML
EPIC_0223	sub	17	74732959	G	T	0.3172	SRSF2	p.P95H	Missense	Pre-AML
EPIC_0223	sub	4	106156725	G	C	0.0124	TET2	p.K542N	Missense	Pre-AML
EPIC_0224	sub	2	25467432	C	T	0.2043	DNMT3A	p.M548I	Missense	Control
EPIC_0225	sub	2	25468192	A	T	0.0030	DNMT3A	p.I495N	Missense	Control
EPIC_0226	sub	12	112924336	G	A	0.0143	PTPN11	p.V428M	Missense	Control
EPIC_0226	sub	2	25458574	A	T	0.0486	DNMT3A	p.?	Essential splice	Control
EPIC_0230	indel	2	25459845	AGCT	-	0.1946	DNMT3A	812_L813delin	Inframe	Control
EPIC_0230	indel	2	25469513	GGCCAGAAGGCTGGAA	-	0.0063	DNMT3A	14_G418delFC	Inframe	Control
EPIC_0234	sub	15	90631934	C	T	0.0375	IDH2	p.R140Q	Missense	Pre-AML
EPIC_0234	sub	2	25463287	G	A	0.1087	DNMT3A	p.R736C	Missense	Pre-AML
EPIC_0236	sub	4	106155530	T	A	0.0021	TET2	p.L144*	Nonsense	Control
EPIC_0241	sub	20	31021472	C	T	0.0234	ASXL1	p.Q491*	Nonsense	Control
EPIC_0246	sub	2	25469161	T	A	0.0058	DNMT3A	p.K433*	Nonsense	Pre-AML
EPIC_0248	sub	4	106156057	G	T	0.0034	TET2	p.E320*	Nonsense	Control
EPIC_0249	sub	20	31022418	G	T	0.4608	ASXL1	p.E635*	Nonsense	Pre-AML
EPIC_0249	sub	4	106156852	T	G	0.0036	TET2	p.S585A	Missense	Pre-AML
EPIC_0254	sub	2	25467477	G	C	0.0038	DNMT3A	p.Y533*	Nonsense	Control
EPIC_0261	sub	17	7576852	C	T	0.0740	TP53	p.?	Essential splice	Pre-AML
EPIC_0261	sub	2	25470015	T	A	0.0144	DNMT3A	p.K343*	Nonsense	Pre-AML
EPIC_0261	sub	21	44514777	T	G	0.0607	U2AF1	p.Q157P	Missense	Pre-AML
EPIC_0261	sub	4	106180784	G	C	0.0029	TET2	p.C1271S	Missense	Pre-AML
EPIC_0261	sub	7	151875055	G	A	0.0543	KMT2C	p.Q2495*	Nonsense	Pre-AML
EPIC_0261	sub	7	151878286	T	C	0.0024	KMT2C	p.Q2220R	Missense	Pre-AML
EPIC_0263	sub	17	7579358	C	G	0.0044	TP53	p.R110P	Missense	Control
EPIC_0263	sub	2	25464568	C	T	0.0037	DNMT3A	p.V649M	Missense	Control
EPIC_0269	sub	2	25458649	G	A	0.0081	DNMT3A	p.Q842*	Nonsense	Pre-AML
EPIC_0269	sub	20	31023717	C	T	0.0038	ASXL1	p.R1068*	Nonsense	Pre-AML
EPIC_0269	sub	21	44514777	T	G	0.0658	U2AF1	p.Q157P	Missense	Pre-AML
EPIC_0270	sub	2	25463170	C	T	0.0473	DNMT3A	p.?	Essential splice	Control
EPIC_0270	sub	2	25463235	C	T	0.0030	DNMT3A	p.W753*	Nonsense	Control
EPIC_0271	sub	4	106156875	T	A	0.1784	TET2	p.Y592*	Nonsense	Pre-AML
EPIC_0271	sub	4	106180795	G	T	0.2218	TET2	p.G1275W	Missense	Pre-AML
EPIC_0272	sub	2	25458625	C	T	0.0032	DNMT3A	p.V850I	Missense	Control
EPIC_0272	sub	4	106197149	C	T	0.0051	TET2	p.Q1828*	Nonsense	Control
EPIC_0274	indel	21	36164771	-	ATGCCG	0.3451	RUNX1	p.M368fs*228	Frameshift	Control

EPIC_0275	sub	2	25464451	G	T	0.0082	DNMT3A	p.R688S	Missense	Control
EPIC_0279	sub	2	25468919	C	A	0.2569	DNMT3A	p.E482*	Nonsense	Pre-AML
EPIC_0280	sub	2	25457173	A	C	0.0066	DNMT3A	p.L905R	Missense	Control
EPIC_0281	sub	2	25463568	A	G	0.0043	DNMT3A	p.I705T	Missense	Control
EPIC_0281	sub	4	106180795	G	C	0.0081	TET2	p.G1275R	Missense	Control
EPIC_0285	sub	2	25469548	A	C	0.0030	DNMT3A	p.I407S	Missense	Control
EPIC_0289	sub	2	25462011	G	C	0.0047	DNMT3A	p.P799R	Missense	Control
EPIC_0289	sub	2	25464456	T	A	0.0054	DNMT3A	p.D686V	Missense	Control
EPIC_0290	sub	2	25466770	T	C	0.0139	DNMT3A	p.T645A	Missense	Control
EPIC_0290	sub	4	106156255	G	C	0.0022	TET2	p.V386L	Missense	Control
EPIC_0291	sub	11	119148976	T	A	0.0491	CBL	p.L399H	Missense	Control
EPIC_0291	sub	17	7578478	G	C	0.0030	TP53	p.P151R	Missense	Control
EPIC_0291	sub	2	25466791	A	T	0.0032	DNMT3A	p.S638T	Missense	Control
EPIC_0292	sub	12	25398284	C	A	0.0030	KRAS	p.G12V	Missense	Control
EPIC_0292	sub	17	7577149	A	C	0.0040	TP53	p.N263K	Missense	Control
EPIC_0292	sub	17	7578413	C	A	0.0055	TP53	p.V173L	Missense	Control
EPIC_0295	sub	4	106196621	C	T	0.0041	TET2	p.Q1652*	Nonsense	Control
EPIC_0297	sub	2	25463297	A	C	0.0152	DNMT3A	p.F732L	Missense	Control
EPIC_0300	sub	2	25458696	T	C	0.0181	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0300	sub	4	106155430	A	T	0.0112	TET2	p.K111*	Nonsense	Pre-AML
EPIC_0300	sub	4	106156069	C	T	0.0343	TET2	p.Q324*	Nonsense	Pre-AML
EPIC_0303	sub	2	25464451	G	A	0.0055	DNMT3A	p.R688C	Missense	Control
EPIC_0303	sub	2	25467190	C	T	0.0031	DNMT3A	p.C562Y	Missense	Control
EPIC_0305	sub	2	25457243	G	T	0.0069	DNMT3A	p.R882S	Missense	Control
EPIC_0305	sub	4	55599321	A	T	0.0081	KIT	p.D816V	Missense	Control
EPIC_0306	sub	2	25459805	C	G	0.0033	DNMT3A	p.K826N	Missense	Control
EPIC_0307	sub	2	25468163	C	A	0.1741	DNMT3A	p.E505*	Nonsense	Control
EPIC_0308	sub	2	25467482	C	T	0.0152	DNMT3A	p.G532S	Missense	Control
EPIC_0309	sub	17	7577121	G	A	0.1051	TP53	p.R273C	Missense	Pre-AML
EPIC_0309	sub	17	7578524	G	C	0.1643	TP53	p.Q136E	Missense	Pre-AML
EPIC_0309	sub	2	25463229	A	C	0.1641	DNMT3A	p.F755C	Missense	Pre-AML
EPIC_0309	sub	2	25463532	T	A	0.0034	DNMT3A	p.N717I	Missense	Pre-AML
EPIC_0309	sub	2	25467023	C	A	0.0523	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0309	sub	4	106156741	C	T	0.0050	TET2	p.Q548*	Nonsense	Pre-AML
EPIC_0311	sub	15	90631934	C	T	0.4299	IDH2	p.R140Q	Missense	Pre-AML
EPIC_0311	sub	17	74732959	G	T	0.4382	SRSF2	p.P95H	Missense	Pre-AML
EPIC_0312	sub	4	106190867	A	G	0.0056	TET2	p.H1382R	Missense	Control
EPIC_0315	sub	2	25466823	G	C	0.0130	DNMT3A	p.P627R	Missense	Control
EPIC_0315	indel	4	106180830	TT	-	0.0146	TET2	p.F1287fs*76	Frameshift	Control
EPIC_0315	indel	4	106196766	AT	-	0.0049	TET2	p.N1700fs*19	Frameshift	Control
EPIC_0317	sub	17	7577539	G	C	0.0030	TP53	p.R248G	Missense	Pre-AML
EPIC_0317	sub	2	25464534	T	C	0.0570	DNMT3A	p.Y660C	Missense	Pre-AML
EPIC_0317	sub	9	5073770	G	T	0.0149	JAK2	p.V617F	Missense	Pre-AML
EPIC_0318	sub	2	25463287	G	A	0.0271	DNMT3A	p.R736C	Missense	Control
EPIC_0325	sub	4	106164769	G	A	0.0030	TET2	p.W1182*	Nonsense	Control
EPIC_0327	sub	2	25457176	G	A	0.0242	DNMT3A	p.P904L	Missense	Pre-AML
EPIC_0327	indel	4	106156316	TT	-	0.0117	TET2	p.S407fs*20	Frameshift	Pre-AML
EPIC_0329	sub	2	25466790	G	T	0.0068	DNMT3A	p.S638Y	Missense	Control
EPIC_0332	sub	2	25467023	C	T	0.0051	DNMT3A	p.?	Essential splice	Control
EPIC_0336	sub	2	25463289	T	G	0.0110	DNMT3A	p.Y735S	Missense	Pre-AML
EPIC_0337	sub	2	25463170	C	T	0.0131	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0337	sub	2	25469150	G	T	0.0075	DNMT3A	p.Y436*	Nonsense	Pre-AML
EPIC_0339	sub	2	25467436	A	T	0.0131	DNMT3A	p.L547H	Missense	Pre-AML
EPIC_0341	sub	21	44524456	G	T	0.2561	U2AF1	p.S34Y	Missense	Pre-AML
EPIC_0346	sub	2	25464429	A	G	0.0477	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0346	sub	9	5073784	G	C	0.1247	JAK2	p.E621D	Missense	Pre-AML
EPIC_0347	sub	4	106155781	A	T	0.0015	TET2	p.K228*	Nonsense	Pre-AML
EPIC_0348	sub	17	7579538	A	G	0.0019	TP53	p.I50T	Missense	Pre-AML
EPIC_0348	sub	2	25467496	T	G	0.0020	DNMT3A	p.Q527P	Missense	Pre-AML
EPIC_0349	sub	11	119148891	T	C	0.1271	CBL	p.Y371H	Missense	Pre-AML
EPIC_0350	sub	4	106164793	T	G	0.0034	TET2	p.C1221G	Missense	Control
EPIC_0354	sub	2	25467190	C	A	0.0095	DNMT3A	p.C562F	Missense	Control
EPIC_0362	sub	2	25469641	G	T	0.1486	DNMT3A	p.A376D	Missense	Control
EPIC_0367	sub	2	25470545	A	G	0.0048	DNMT3A	p.I310T	Missense	Control
EPIC_0367	sub	4	106155439	C	T	0.0037	TET2	p.Q114*	Nonsense	Control
EPIC_0367	sub	4	106197248	G	A	0.0052	TET2	p.G1861R	Missense	Control
EPIC_0368	sub	12	25398281	C	T	0.0064	KRAS	p.G13D	Missense	Control
EPIC_0368	sub	17	7577124	C	T	0.0064	TP53	p.V272M	Missense	Control
EPIC_0371	indel	4	106196282	CAG	-	0.0067	TET2	p.Q1539fs*38	Frameshift	Control
EPIC_0372	sub	4	106163989	A	T	0.0034	TET2	p.?	Essential splice	Control
EPIC_0377	sub	4	106190860	C	G	0.3476	TET2	p.H1380D	Missense	Pre-AML
EPIC_0377	indel	4	106196430	-	TATGGAAGCACCAG	0.1272	TET2	p.Y1589fs*30	Frameshift	Pre-AML
EPIC_0378	sub	2	25457242	C	T	0.1671	DNMT3A	p.R882H	Missense	Pre-AML
EPIC_0379	sub	2	25464578	T	C	0.0047	DNMT3A	p.?	Essential splice	Control
EPIC_0381	sub	2	25466800	G	A	0.0278	DNMT3A	p.R635W	Missense	Pre-AML
EPIC_0382	sub	4	106156348	C	T	0.0044	TET2	p.Q417*	Nonsense	Control
EPIC_0389	sub	2	25467073	C	A	0.0189	DNMT3A	p.W601L	Missense	Control
EPIC_0389	sub	2	25468122	C	A	0.0054	DNMT3A	p.K518N	Missense	Control
EPIC_0392	sub	11	119148892	A	G	0.0034	CBL	p.Y371C	Missense	Pre-AML
EPIC_0392	sub	2	25457242	C	T	0.3685	DNMT3A	p.R882H	Missense	Pre-AML
EPIC_0392	sub	2	198267371	G	C	0.1042	SF3B1	p.H662Q	Missense	Pre-AML
EPIC_0392	sub	20	31021319	A	C	0.0031	ASXL1	p.K440Q	Missense	Pre-AML
EPIC_0395	indel	4	106180798	CTGGATCC	-	0.0046	TET2	p.L1276fs*85	Frameshift	Control
EPIC_0396	sub	4	106164020	T	G	0.0050	TET2	p.I1177S	Missense	Control
EPIC_0397	sub	11	119148537	C	T	0.0317	CBL	p.H360Y	Missense	Pre-AML
EPIC_0397	sub	17	74732959	G	T	0.2987	SRSF2	p.P95H	Missense	Pre-AML
EPIC_0397	sub	4	106155354	T	G	0.0067	TET2	p.Y85*	Nonsense	Pre-AML
EPIC_0397	sub	4	106197248	G	T	0.0049	TET2	p.G1861*	Nonsense	Pre-AML
EPIC_0397	sub	9	5073770	G	T	0.1488	JAK2	p.V617F	Missense	Pre-AML
EPIC_0399	indel	2	25462073	-	AGGGTTGGACTACA	0.0040	DNMT3A	p.M779fs*2	Frameshift	Control
EPIC_0400	sub	2	25463247	C	T	0.4181	DNMT3A	p.R749H	Missense	Control
EPIC_0402	sub	17	29527461	C	T	0.0094	NF1	p.R304*	Nonsense	Control
EPIC_0402	sub	2	25466793	A	T	0.0369	DNMT3A	p.L637Q	Missense	Control
EPIC_0404	sub	4	106164778	C	T	0.0049	TET2	p.R1216*	Nonsense	Control
EPIC_0408	sub	9	5073770	G	T	0.0126	JAK2	p.V617F	Missense	Control
EPIC_0409	sub	2	25459851	T	A	0.0174	DNMT3A	p.D811V	Missense	Control
EPIC_0409	sub	4	106193892	C	T	0.3680	TET2	p.R1452*	Nonsense	Control
EPIC_0410	sub	2	25463227	C	T	0.0036	DNMT3A	p.E756K	Missense	Control
EPIC_0410	sub	X	44969323	G	A	0.0054	KDM6A	p.?	Essential splice	Control
EPIC_0411	sub	2	25467099	G	C	0.0024	DNMT3A	p.Y592*	Nonsense	Control

EPIC_0412	indel	4	106155605	AT	-	0.0345	TET2	p.H169fs*14	Frameshift	Control
EPIC_0413	sub	17	7577545	T	C	0.0056	TP53	p.M246V	Missense	Control
EPIC_0413	sub	2	25463286	C	T	0.0163	DNMT3A	p.R736H	Missense	Control
EPIC_0413	sub	2	25467428	C	T	0.0044	DNMT3A	p.G550R	Missense	Control
EPIC_0415	sub	17	7578404	A	T	0.0033	TP53	p.C176S	Missense	Control
EPIC_0415	sub	2	25458595	A	G	0.0341	DNMT3A	p.W860R	Missense	Control
EPIC_0415	sub	2	25463182	G	A	0.0171	DNMT3A	p.R771*	Nonsense	Control
EPIC_0415	sub	2	198267370	T	G	0.0190	SF3B1	p.T663P	Missense	Control
EPIC_0415	sub	4	106190905	G	A	0.0131	TET2	p.?	Essential splice	Control
EPIC_0421	sub	2	25462014	A	G	0.0180	DNMT3A	p.L798P	Missense	Control
EPIC_0422	sub	2	25457242	C	T	0.0432	DNMT3A	p.R882H	Missense	Control
EPIC_0422	sub	2	25463316	C	T	0.0106	DNMT3A	p.G726D	Missense	Control
EPIC_0422	sub	2	25464456	T	A	0.0223	DNMT3A	p.D686V	Missense	Control
EPIC_0423	sub	2	25463170	C	T	0.0215	DNMT3A	p.?	Essential splice	Control
EPIC_0424	sub	4	106156211	T	A	0.0468	TET2	p.L371*	Nonsense	Pre-AML
EPIC_0426	sub	2	25466802	A	C	0.0063	DNMT3A	p.I634S	Missense	Control
EPIC_0427	indel	2	25463243	GGGGCG	-	0.0040	DNMT3A	p.R749fs*6	Frameshift	Control
EPIC_0428	sub	2	25464490	C	G	0.0040	DNMT3A	p.V675L	Missense	Control
EPIC_0431	sub	2	25463568	A	G	0.0501	DNMT3A	p.I705T	Missense	Control
EPIC_0431	sub	4	106156468	G	A	0.0036	TET2	p.A457T	Missense	Control
EPIC_0433	sub	4	106182926	T	A	0.0045	TET2	p.L1322Q	Missense	Control
EPIC_0435	sub	12	25380276	T	C	0.0030	KRAS	p.Q61R	Missense	Control
EPIC_0436	sub	2	25466799	C	T	0.0140	DNMT3A	p.R635Q	Missense	Control
EPIC_0436	sub	2	25467485	C	T	0.0040	DNMT3A	p.D531N	Missense	Control
EPIC_0436	sub	20	31022382	C	T	0.0075	ASXL1	p.Q623*	Nonsense	Control
EPIC_0445	sub	4	106162559	C	T	0.0088	TET2	p.A1158V	Missense	Control
EPIC_0447	sub	X	39933843	G	T	0.0069	BCOR	p.Y252*	Nonsense	Control
EPIC_0448	sub	17	7578259	A	T	0.0579	TP53	p.V197E	Missense	Pre-AML
EPIC_0448	indel	20	31022403	ACCACTGCCATAGAGAGGCGG	-	0.0483	ASXL1	p.H630fs*66	Frameshift	Pre-AML
EPIC_0448	sub	7	151884437	C	A	0.0039	KMT2C	p.E1640*	Nonsense	Pre-AML
EPIC_0449	sub	4	106197437	A	G	0.0041	TET2	p.K1924E	Missense	Control
EPIC_0450	sub	2	25463169	A	C	0.0943	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0452	sub	17	74732960	G	C	0.0042	SRSF2	p.P95A	Missense	Control
EPIC_0453	sub	4	106164772	C	T	0.0130	TET2	p.R1214W	Missense	Control
EPIC_0454	sub	2	25457282	C	A	0.0103	DNMT3A	p.G869C	Missense	Pre-AML
EPIC_0459	sub	2	25463566	C	T	0.0038	DNMT3A	p.G706R	Missense	Control
EPIC_0459	sub	2	25464451	G	A	0.0044	DNMT3A	p.R688C	Missense	Control
EPIC_0459	indel	4	106196515	CCCTTACC	-	0.0049	TET2	p.P1617fs*4	Frameshift	Control
EPIC_0460	indel	2	25467145	TTAATGGCTGCCTGGGCAG	-	0.0054	DNMT3A	571_K577deli	Inframe	Control
EPIC_0462	sub	2	25464460	C	T	0.0197	DNMT3A	p.G685R	Missense	Control
EPIC_0462	indel	20	31017747	CAG	-	0.0049	ASXL1	p.S204fs*49	Frameshift	Control
EPIC_0464	sub	2	25458696	T	G	0.0041	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0464	sub	2	25463184	G	A	0.1909	DNMT3A	p.S770L	Missense	Pre-AML
EPIC_0464	sub	9	5073770	G	T	0.2352	JAK2	p.V617F	Missense	Pre-AML
EPIC_0466	sub	2	25463290	A	G	0.0061	DNMT3A	p.Y735H	Missense	Control
EPIC_0468	sub	4	106164788	A	T	0.0347	TET2	p.H1219L	Missense	Control
EPIC_0469	sub	15	90631934	C	T	0.1137	IDH2	p.R140Q	Missense	Pre-AML
EPIC_0469	sub	2	25463184	G	A	0.1850	DNMT3A	p.S770L	Missense	Pre-AML
EPIC_0469	sub	2	25463536	C	T	0.0516	DNMT3A	p.V716I	Missense	Pre-AML
EPIC_0469	sub	2	25470584	C	T	0.0025	DNMT3A	p.W297*	Nonsense	Pre-AML
EPIC_0469	sub	9	5073770	G	T	0.0151	JAK2	p.V617F	Missense	Pre-AML
EPIC_0469	sub	X	44929280	A	T	0.0020	KDM6A	p.T794S	Missense	Pre-AML
EPIC_0470	sub	17	74732959	G	T	0.1429	SRSF2	p.P95H	Missense	Pre-AML
EPIC_0470	sub	20	31022288	C	A	0.1162	ASXL1	p.Y591*	Nonsense	Pre-AML
EPIC_0470	sub	21	36164601	G	A	0.0042	RUNX1	p.P425L	Missense	Pre-AML
EPIC_0470	sub	21	36252882	G	T	0.0795	RUNX1	p.D160E	Missense	Pre-AML
EPIC_0470	sub	21	36259171	C	T	0.0076	RUNX1	p.R107H	Missense	Pre-AML
EPIC_0473	sub	20	31022902	G	A	0.3710	ASXL1	p.W796*	Nonsense	Pre-AML
EPIC_0473	sub	21	44514777	T	G	0.0049	U2AF1	p.Q157P	Missense	Pre-AML
EPIC_0473	indel	4	106196992	CT	-	0.0093	TET2	p.S1776fs*44	Frameshift	Pre-AML
EPIC_0474	sub	2	25464462	A	T	0.0034	DNMT3A	p.V684D	Missense	Control
EPIC_0474	indel	2	25471033	CTGGCCTCCT	-	0.0130	DNMT3A	240_S243delir	Inframe	Control
EPIC_0476	indel	4	106196958	ATAACTACAG	-	0.0080	TET2	1765_S1767de	Inframe	Control
EPIC_0477	indel	17	74732962	-	GAG	0.3155	SRSF2	p.R94fs*151	Frameshift	Pre-AML
EPIC_0477	sub	21	36171607	G	A	0.4910	RUNX1	p.R320*	Nonsense	Pre-AML
EPIC_0479	sub	21	36252940	G	A	0.0073	RUNX1	p.S141L	Missense	Pre-AML
EPIC_0479	sub	3	128200730	A	C	0.0678	GATA2	p.L359V	Missense	Pre-AML
EPIC_0486	sub	12	25398248	A	T	0.0040	KRAS	p.I24N	Missense	Control
EPIC_0490	indel	4	106164025	AG	-	0.0560	TET2	p.R1179fs*47	Frameshift	Pre-AML
EPIC_0493	sub	13	28592642	C	G	0.0778	FLT3	p.D835H	Missense	Pre-AML
EPIC_0496	sub	15	90631839	T	A	0.0036	IDH2	p.R172W	Missense	Pre-AML
EPIC_0497	sub	2	25463508	C	T	0.0046	DNMT3A	p.?	Essential splice	Control
EPIC_0498	sub	17	74732959	G	T	0.2079	SRSF2	p.P95H	Missense	Pre-AML
EPIC_0498	sub	2	209113112	C	A	0.0109	IDH1	p.R132L	Missense	Pre-AML
EPIC_0498	sub	4	106190843	G	A	0.0126	TET2	p.C1374Y	Missense	Pre-AML
EPIC_0501	sub	2	25463563	C	T	0.0105	DNMT3A	p.G707S	Missense	Control
EPIC_0501	sub	2	25469548	A	C	0.0030	DNMT3A	p.I407S	Missense	Control
EPIC_0503	sub	2	25458669	G	T	0.0027	DNMT3A	p.T835K	Missense	Control
EPIC_0504	sub	15	90631934	C	T	0.3390	IDH2	p.R140Q	Missense	Pre-AML
EPIC_0504	sub	17	74732959	G	A	0.0097	SRSF2	p.P95L	Missense	Pre-AML
EPIC_0504	sub	2	25467467	A	T	0.0603	DNMT3A	p.C537S	Missense	Pre-AML
EPIC_0504	sub	2	25469085	C	G	0.0067	DNMT3A	p.R458P	Missense	Pre-AML
EPIC_0507	sub	2	25457242	C	T	0.0635	DNMT3A	p.R882H	Missense	Pre-AML
EPIC_0507	sub	4	106182914	A	G	0.0084	TET2	p.?	Essential splice	Pre-AML
EPIC_0508	sub	2	25468120	A	G	0.0035	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0509	sub	2	25457171	T	C	0.0052	DNMT3A	p.K906E	Missense	Pre-AML
EPIC_0509	sub	2	25466800	G	A	0.3882	DNMT3A	p.R635W	Missense	Pre-AML
EPIC_0510	sub	2	209113112	C	T	0.3042	IDH1	p.R132H	Missense	Pre-AML
EPIC_0511	sub	2	25462012	G	T	0.0352	DNMT3A	p.P799T	Missense	Control
EPIC_0512	sub	2	25463289	T	G	0.0031	DNMT3A	p.Y735S	Missense	Pre-AML
EPIC_0512	sub	2	25467132	C	T	0.0053	DNMT3A	p.W581*	Nonsense	Pre-AML
EPIC_0512	sub	2	25467408	C	T	0.0154	DNMT3A	p.?	Essential splice	Pre-AML
EPIC_0512	sub	2	25469946	G	T	0.0076	DNMT3A	p.R366S	Missense	Pre-AML
EPIC_0512	sub	4	106155652	C	T	0.0111	TET2	p.Q185*	Nonsense	Pre-AML
EPIC_0512	sub	4	106164726	G	A	0.0112	TET2	p.?	Essential splice	Pre-AML
EPIC_0516	sub	4	106164764	G	A	0.0121	TET2	p.C1211Y	Missense	Pre-AML

Appendix 9: Mutations in validation cohort pre-AML, control and AML diagnostic samples

Sample ID	Type	Chromosome	Position	WT	MT	VAF	Gene	Protein	Effect	Group
PD29762b	sub	17	74732959	G	T	0.1500	SRSF2	p.P95H	Missense	Pre-AML
PD29762b	sub	4	106164913	C	A	0.0840	TET2	p.R1261S	Missense	Pre-AML
PD29762b	indel	4	106193849	G	GA	0.2857	TET2	p.R1440fs*38	Frameshift	Pre-AML
PD29762b	indel	4	106197311	GC	G	0.1362	TET2	p.T1883fs*4	Frameshift	Pre-AML
PD29764b	sub	4	106157827	C	T	0.0980	TET2	p.Q910*	Nonsense	Pre-AML
PD29792b	indel	4	106157182	AT	A	0.3197	TET2	p.M695fs*5	Frameshift	Pre-AML
PD29792b	sub	4	106158509	G	C	0.3500	TET2	p.?	Essential splice	Pre-AML
PD29810c	indel	12	49418417	C	CA	0.3988	KMT2D	p.M5332fs*13	Frameshift	Control
PD29836b	sub	17	74732959	G	T	0.0077	SRSF2	p.P95H	Missense	Pre-AML
PD29836b	sub	4	106190900	C	T	0.0440	TET2	p.T1393I	Missense	Pre-AML
PD29836c	sub	17	74732959	G	T	0.0083	SRSF2	p.P95H	Missense	Pre-AML
PD29836c	sub	4	106190900	C	T	0.0440	TET2	p.T1393I	Missense	Pre-AML
PD29856c	sub	1	115256521	A	C	0.0340	NRAS	p.Y64D	Missense	Pre-AML
PD29896b	indel	20	31022837	AT	A	0.2587	ASXL1	p.L775fs*1	Frameshift	Pre-AML
PD29918b	sub	17	74732959	G	C	0.3400	SRSF2	p.P95R	Missense	Pre-AML
PD29918b	sub	19	33792753	A	G	0.0868	CEBPA	p.S190P	Missense	Pre-AML
PD29918b	sub	4	106156160	C	G	0.1900	TET2	p.S354*	Nonsense	Pre-AML
PD29918c	sub	17	74732959	G	C	0.0950	SRSF2	p.P95R	Missense	Pre-AML
PD29918d	sub	17	74732959	G	C	0.3700	SRSF2	p.P95R	Missense	Pre-AML
PD29918d	sub	21	36259178	G	A	0.0680	RUNX1	p.H105Y	Missense	Pre-AML
PD29918d	sub	4	106156160	C	G	0.0220	TET2	p.S354*	Nonsense	Pre-AML
PD29931b	sub	17	74732959	G	C	0.1100	SRSF2	p.P95R	Missense	Pre-AML
PD29931b	sub	2	25457242	C	T	0.3700	DNMT3A	p.R882H	Missense	Pre-AML
PD29935b	sub	2	25463248	G	A	0.1300	DNMT3A	p.R749C	Missense	Pre-AML
PD29935c	sub	2	25463248	G	A	0.1200	DNMT3A	p.R749C	Missense	Pre-AML
PD29935d	sub	2	25463248	G	A	0.1500	DNMT3A	p.R749C	Missense	Pre-AML
PD29946b	sub	2	25457243	G	T	0.0159	DNMT3A	p.R882S	Missense	Pre-AML
PD29946b	sub	2	25463247	C	T	0.1300	DNMT3A	p.R749H	Missense	Pre-AML
PD29946b	sub	2	25470497	C	T	0.1500	DNMT3A	p.R326H	Missense	Pre-AML
PD29946c	sub	2	25457243	G	T	0.0074	DNMT3A	p.R882S	Missense	Pre-AML
PD29946c	sub	2	25463247	C	T	0.0510	DNMT3A	p.R749H	Missense	Pre-AML
PD29946c	sub	2	25470497	C	T	0.0690	DNMT3A	p.R326H	Missense	Pre-AML
PD29948b	indel	2	25469083	TC	T	0.0181	DNMT3A	p.K459fs*192	Frameshift	Pre-AML
PD29951b	sub	2	25467479	A	T	0.0340	DNMT3A	p.Y533N	Missense	Pre-AML
PD29962b	sub	12	25398284	C	T	0.0102	KRAS	p.G12D	Missense	Pre-AML
PD29962b	sub	4	106157653	G	T	0.0570	TET2	p.E852*	Nonsense	Pre-AML
PD29993b	sub	2	25463286	C	T	0.0217	DNMT3A	p.R736H	Missense	Pre-AML
PD29993b	sub	2	25469139	C	T	0.0140	DNMT3A	p.W440*	Nonsense	Pre-AML
PD29993b	sub	9	5073770	G	T	0.0051	JAK2	p.V617F	Missense	Pre-AML
PD30010b	sub	4	106156699	A	T	0.6400	TET2	p.R534*	Nonsense	Pre-AML
PD30010c	sub	4	106156699	A	T	0.6400	TET2	p.R534*	Nonsense	Pre-AML
PD30023b	sub	17	7576852	C	T	0.0830	TP53	p.?	Essential splice	Pre-AML
PD30023b	sub	2	25470015	T	A	0.0140	DNMT3A	p.K343*	Nonsense	Pre-AML
PD30023b	sub	21	44514777	T	G	0.0600	U2AF1	p.Q157P	Missense	Pre-AML
PD30023b	sub	7	151875055	G	A	0.0297	KMT2C	p.Q2495*	Nonsense	Pre-AML
PD30031b	sub	2	25467139	T	C	0.0420	DNMT3A	p.D579G	Missense	Pre-AML
PD30054b	sub	21	44514777	T	G	0.0470	U2AF1	p.Q157P	Missense	Pre-AML
PD30060b	sub	2	25464460	C	T	0.2100	DNMT3A	p.G685R	Missense	Pre-AML
PD30060b	sub	4	106190812	G	T	0.0099	TET2	p.E1364*	Nonsense	Pre-AML
PD30060c	sub	2	25464460	C	T	0.2100	DNMT3A	p.G685R	Missense	Pre-AML
PD30060c	sub	4	106190812	G	T	0.0077	TET2	p.E1364*	Nonsense	Pre-AML
PD30073b	sub	12	112924336	G	A	0.3600	PTPN11	p.V428M	Missense	Pre-AML
PD30073b	sub	4	106182914	A	G	0.3400	TET2	p.?	Essential splice	Pre-AML
PD30073b	sub	4	106196213	C	T	0.3400	TET2	p.R1516*	Nonsense	Pre-AML
PD30086b	sub	17	74732959	G	A	0.0220	SRSF2	p.P95L	Missense	Pre-AML
PD30089b	sub	17	74732959	G	T	0.2600	SRSF2	p.P95H	Missense	Pre-AML
PD30089b	sub	2	25466799	C	A	0.3600	DNMT3A	p.R635L	Missense	Pre-AML
PD30089c	sub	17	74732959	G	T	0.3700	SRSF2	p.P95H	Missense	Pre-AML
PD30089c	sub	2	25466799	C	A	0.4400	DNMT3A	p.R635L	Missense	Pre-AML
PD30089c	sub	9	5073770	G	T	0.1300	JAK2	p.V617F	Missense	Pre-AML
PD30120b	sub	17	7577099	C	T	0.0135	TP53	p.R280K	Missense	Pre-AML
PD30120b	sub	2	25464573	A	C	0.0078	DNMT3A	p.L647R	Missense	Pre-AML
PD30154b	sub	2	25470551	C	T	0.0082	DNMT3A	p.G308D	Missense	Pre-AML
PD30154b	sub	X	39922984	G	A	0.0117	BCOR	p.Q1242*	Nonsense	Pre-AML
PD35511b	sub	2	25457242	C	T	0.0056	DNMT3A	p.R882H	Missense	Control
PD35515b	indel	4	106193849	G	GA	0.0443	TET2	p.R1440fs*38	Frameshift	Control
PD35518b	sub	2	25457209	C	T	0.0110	DNMT3A	p.W893*	Nonsense	Control
PD35519c	sub	17	74732959	G	A	0.0178	SRSF2	p.P95L	Missense	Control
PD35520b	sub	12	25398284	C	G	0.0109	KRAS	p.G12A	Missense	Control
PD35520b	sub	2	25468935	T	A	0.0330	DNMT3A	p.?	Essential splice	Control
PD35520c	sub	12	25398284	C	G	0.0048	KRAS	p.G12A	Missense	Control
PD35520c	sub	2	25468935	T	A	0.1100	DNMT3A	p.?	Essential splice	Control
PD35525b	sub	20	31021295	C	T	0.0326	ASXL1	p.Q432*	Nonsense	Control
PD35529b	sub	17	7576865	A	T	0.0216	TP53	p.Y327*	Nonsense	Control
PD35531b	sub	4	106164079	A	T	0.0064	TET2	p.K1197*	Nonsense	Control
PD35531c	sub	4	106164079	A	T	0.0075	TET2	p.K1197*	Nonsense	Control
PD35534b	sub	12	25380275	T	G	0.0070	KRAS	p.Q61H	Missense	Control
PD35537b	sub	2	25467158	G	A	0.0074	DNMT3A	p.Q573*	Nonsense	Control
PD35538b	sub	2	25467407	A	G	0.0085	DNMT3A	p.?	Essential splice	Control
PD35538c	sub	2	25467407	A	G	0.0112	DNMT3A	p.?	Essential splice	Control
PD35539b	sub	2	25463308	G	A	0.0165	DNMT3A	p.R729W	Missense	Control
PD35539c	sub	2	25463308	G	A	0.0100	DNMT3A	p.R729W	Missense	Control
PD35539c	sub	2	25470535	C	T	0.0420	DNMT3A	p.W313*	Nonsense	Control
PD35542b	sub	4	106180868	A	G	0.0354	TET2	p.K1299R	Missense	Control
PD35542c	sub	4	106180868	A	G	0.1500	TET2	p.K1299R	Missense	Control
PD35545b	sub	2	25457242	C	T	0.0066	DNMT3A	p.R882H	Missense	Control
PD35545c	sub	2	25457242	C	T	0.0105	DNMT3A	p.R882H	Missense	Control
PD35548c	sub	21	44514780	C	T	0.0054	U2AF1	p.R156H	Missense	Control
PD35553c	indel	4	106164861	ACT	A	0.0444	TET2	p.Y1245fs*22	Frameshift	Control
PD35553c	sub	4	106182983	C	G	0.0165	TET2	p.A1341G	Missense	Control
PD35554b	sub	2	2.55E+07	T	C	0.0059	DNMT3A	p.R803G	Missense	Control
PD35554c	sub	2	2.55E+07	T	C	0.0106	DNMT3A	p.R803G	Missense	Control
PD35556b	sub	2	25459806	T	C	0.0233	DNMT3A	p.K826R	Missense	Control
PD35558b	sub	20	31021176	C	G	0.0184	ASXL1	p.S392*	Nonsense	Control
PD35558c	sub	2	198267369	G	A	0.0116	SF3B1	p.T663I	Missense	Control
PD35559b	sub	2	25466800	G	A	0.0178	DNMT3A	p.R635W	Missense	Control
PD35560b	sub	4	106180852	T	A	0.0102	TET2	p.Y1294N	Missense	Control
PD35563b	sub	2	25458688	T	A	0.0184	DNMT3A	p.K829*	Nonsense	Control
PD35563b	indel	2	25464450	CG	C	0.0100	DNMT3A	p.R688fs*17	Frameshift	Control
PD35563c	sub	2	25458688	T	A	0.0490	DNMT3A	p.K829*	Nonsense	Control
PD35568c	sub	20	31022903	G	A	0.0168	ASXL1	p.W796*	Nonsense	Control
PD35569b	sub	2	25467073	C	T	0.0070	DNMT3A	p.W601*	Nonsense	Control
PD35569c	sub	2	25467073	C	T	0.0044	DNMT3A	p.W601*	Nonsense	Control
PD35576c	indel	2	25467447	G	GC	0.1172	DNMT3A	p.R544fs*2	Frameshift	Control
PD35578c	sub	2	25462075	C	T	0.0186	DNMT3A	p.V778M	Missense	Control
PD35579b	sub	2	25470583	C	A	0.1800	DNMT3A	p.W297C	Missense	Control
PD35579c	sub	2	25470583	C	A	0.3200	DNMT3A	p.W297C	Missense	Control

PD35580b	sub	2	25463181	C	A	0.0470	DNMT3A	p.R771L	Missense	Control
PD35580c	sub	2	25463181	C	A	0.1000	DNMT3A	p.R771L	Missense	Control
PD35580c	sub	2	25470569	C	T	0.0127	DNMT3A	p.G302D	Missense	Control
PD35582b	sub	2	25464538	G	C	0.0069	DNMT3A	p.R659G	Missense	Control
PD35587c	sub	2	198267484	G	A	0.0121	SF3B1	p.R625C	Missense	Control
PD35588b	sub	2	25467466	C	G	0.0054	DNMT3A	p.C537S	Missense	Control
PD35592c	sub	4	106190898	C	G	0.0430	TET2	p.S1392R	Missense	Control
PD35594c	indel	4	106158496	T	TG	0.0720	TET2	p.C1133fs*9	Frameshift	Control
PD35599b	sub	1	115256530	G	T	0.0077	NRAS	p.Q61K	Missense	Control
PD35599b	sub	2	25470545	A	C	0.0147	DNMT3A	p.I310S	Missense	Control
PD35600c	sub	2	25462018	T	C	0.1800	DNMT3A	p.N797D	Missense	Control
PD35600c	sub	2	25463287	G	A	0.0125	DNMT3A	p.R736C	Missense	Control
PD35600c	sub	2	25466796	A	C	0.0167	DNMT3A	p.V636G	Missense	Control
PD35601b	sub	2	25469646	C	T	0.0180	DNMT3A	p.?	Essential splice	Control
PD35606b	sub	2	25470583	C	T	0.0480	DNMT3A	p.W297*	Nonsense	Control
PD35606c	sub	2	25470583	C	T	0.0490	DNMT3A	p.W297*	Nonsense	Control
PD35612b	sub	15	90631934	C	T	0.0109	IDH2	p.R140Q	Missense	Control
PD35612b	sub	7	151970884	A	C	0.0499	KMT2C	p.Y306*	Nonsense	Control
PD35613b	sub	2	25470535	C	T	0.0052	DNMT3A	p.W313*	Nonsense	Control
PD35613c	sub	2	25470535	C	T	0.0063	DNMT3A	p.W313*	Nonsense	Control
PD35613c	sub	2	209113112	C	T	0.0115	IDH1	p.R132H	Missense	Control
PD35613c	sub	4	106156975	C	T	0.1100	TET2	p.Q626*	Nonsense	Control
PD35616c	sub	2	25467134	A	T	0.0073	DNMT3A	p.W581R	Missense	Control
PD35617b	sub	2	198266834	T	C	0.0084	SF3B1	p.K700E	Missense	Control
PD35618b	sub	2	198266834	T	C	0.0091	SF3B1	p.K700E	Missense	Control
PD35618c	sub	17	29576135	C	T	0.0070	NF1	p.Q1370*	Nonsense	Control
PD35618c	sub	17	74732959	G	A	0.0138	SRSF2	p.P95L	Missense	Control
PD35618c	sub	2	198266834	T	C	0.0590	SF3B1	p.K700E	Missense	Control
PD35618c	sub	4	106164778	C	T	0.0133	TET2	p.R1216*	Nonsense	Control
PD35620b	sub	2	25457242	C	T	0.0450	DNMT3A	p.R882H	Missense	Control
PD35620c	sub	2	25457242	C	T	0.0410	DNMT3A	p.R882H	Missense	Control
PD35621b	sub	7	151970855	G	T	0.0475	KMT2C	p.T316N	Missense	Control
PD35629b	sub	2	25457243	G	A	0.0052	DNMT3A	p.R882C	Missense	Control
PD35636b	sub	2	25467497	G	A	0.0450	DNMT3A	p.Q527*	Nonsense	Control
PD35637c	indel	12	49441815	GC	G	0.0262	KMT2D	p.A1390fs*27	Frameshift	Control
PD35638b	sub	2	25464451	G	T	0.0086	DNMT3A	p.R688S	Missense	Control
PD35639b	indel	2	25464469	TG	T	0.0105	DNMT3A	p.M682fs*23	Frameshift	Control
PD35647c	indel	20	31021175	TC	T	0.0053	ASXL1	p.S392fs*1	Frameshift	Control
PD35652c	sub	2	25462005	A	G	0.0095	DNMT3A	p.M801T	Missense	Control
PD35652c	sub	2	25467478	T	C	0.0076	DNMT3A	p.Y533C	Missense	Control
PD35653b	sub	2	25467099	G	C	0.0055	DNMT3A	p.Y592*	Nonsense	Control
PD35654b	sub	2	198266834	T	C	0.0600	SF3B1	p.K700E	Missense	Control
PD35659b	sub	4	106190849	A	T	0.0175	TET2	p.D1376V	Missense	Control
PD35659c	indel	2	25468168	G	GT	0.1286	DNMT3A	p.T503fs*43	Frameshift	Control
PD35659c	sub	4	106190849	A	T	0.1300	TET2	p.D1376V	Missense	Control
PD35660c	sub	17	74732959	G	T	0.0063	SRSF2	p.P95H	Missense	Control
PD35665c	indel	12	49434957	TA	T	0.1224	KMT2D	p.Y2199fs*65	Frameshift	Control
PD35666b	sub	2	25463290	A	T	0.0179	DNMT3A	p.Y735N	Missense	Control
PD35667b	sub	2	25458696	T	G	0.0077	DNMT3A	p.?	Essential splice	Control
PD35671b	sub	20	31024492	C	T	0.0110	ASXL1	p.P1326L	Missense	Control
PD35675b	sub	2	25457285	A	G	0.0154	DNMT3A	p.F868L	Missense	Control
PD35677b	sub	2	25457242	C	T	0.0051	DNMT3A	p.R882H	Missense	Control
PD35677c	sub	2	25457242	C	T	0.0057	DNMT3A	p.R882H	Missense	Control
PD35677c	indel	2	25467039	G	GT	0.0539	DNMT3A	p.N612fs*7	Frameshift	Control
PD35678b	sub	2	25463248	G	T	0.0145	DNMT3A	p.R749S	Missense	Control
PD35683b	sub	2	25470579	T	A	0.0082	DNMT3A	p.K299*	Nonsense	Control
PD35685b	sub	2	25463584	G	C	0.0102	DNMT3A	p.P700A	Missense	Control
PD35686b	sub	2	25469528	A	C	0.0330	DNMT3A	p.F414V	Missense	Control
PD35687b	sub	2	25457242	C	T	0.0079	DNMT3A	p.R882H	Missense	Control
PD35688b	sub	17	29562934	A	G	0.0383	NF1	p.?	Essential splice	Control
PD35688b	sub	9	5073770	G	T	0.0352	JAK2	p.V617F	Missense	Control
PD35693b	sub	8	117875485	A	T	0.0158	RAD21	p.L53*	Nonsense	Control
PD35700b	sub	2	25466852	C	T	0.0253	DNMT3A	p.?	Essential splice	Control
PD35704b	sub	11	119149280	G	A	0.1300	CBL	p.V430M	Missense	Control
PD35704c	sub	11	119149280	G	A	0.1100	CBL	p.V430M	Missense	Control
PD35705b	sub	2	25458580	C	T	0.0203	DNMT3A	p.E865K	Missense	Control
PD35709c	sub	2	25469632	C	T	0.0570	DNMT3A	p.R379H	Missense	Control
PD35711b	sub	12	25378562	C	T	0.0093	KRAS	p.A146T	Missense	Control
PD35719c	sub	4	106182972	T	A	0.0078	TET2	p.Y1337*	Nonsense	Control
PD35723b	sub	2	25467467	A	G	0.0156	DNMT3A	p.C537R	Missense	Control
PD35724b	sub	7	151873585	G	A	0.0054	KMT2C	p.Q2985*	Nonsense	Control
PD35724b	sub	8	117859932	T	A	0.0127	RAD21	p.?	Essential splice	Control
PD35732c	sub	2	25463283	A	T	0.0272	DNMT3A	p.L737H	Missense	Control
PD35733b	sub	2	25467449	C	A	0.0230	DNMT3A	p.G543C	Missense	Control
PD35733b	sub	4	106180931	G	A	0.1200	TET2	p.?	Essential splice	Control
PD35733c	sub	4	106180931	G	A	0.2000	TET2	p.?	Essential splice	Control
PD35755b	sub	2	25461994	C	T	0.0093	DNMT3A	p.?	Essential splice	Control
PD35755b	sub	2	25466800	G	A	0.0144	DNMT3A	p.R635W	Missense	Control
PD35755c	sub	2	25461994	C	T	0.0132	DNMT3A	p.?	Essential splice	Control
PD35755c	sub	2	25466800	G	A	0.0265	DNMT3A	p.R635W	Missense	Control
PD35756b	sub	2	25470498	G	A	0.0144	DNMT3A	p.R326C	Missense	Control
PD35756b	sub	4	106197285	T	C	0.0490	TET2	p.I1873T	Missense	Control
PD35756c	sub	4	106197285	T	C	0.0630	TET2	p.I1873T	Missense	Control
PD35760c	sub	17	29562957	C	T	0.0144	NF1	p.Q1298*	Nonsense	Control
PD35762c	sub	2	25467059	G	A	0.0085	DNMT3A	p.Q606*	Nonsense	Control
PD35763c	indel	20	31022951	TC	T	0.0324	ASXL1	p.I814fs*4	Frameshift	Control
PD35768b	sub	2	25457243	G	A	0.0065	DNMT3A	p.R882C	Missense	Control
PD35768c	sub	2	25457243	G	A	0.0870	DNMT3A	p.R882C	Missense	Control
PD35769c	indel	4	106190781	CA	C	0.0147	TET2	p.R1354fs*9	Frameshift	Control
PD35769c	sub	4	106197255	C	A	0.1300	TET2	p.A1863D	Missense	Control
PD35777b	sub	2	25464531	A	G	0.0114	DNMT3A	p.I661T	Missense	Control
PD35778b	sub	8	117874079	C	T	0.0411	RAD21	p.?	Essential splice	Control
PD35780b	sub	2	25463248	G	A	0.0258	DNMT3A	p.R749C	Missense	Control
PD35780c	sub	2	25457155	C	A	0.0133	DNMT3A	p.C911F	Missense	Control
PD35780c	sub	2	25463248	G	A	0.0620	DNMT3A	p.R749C	Missense	Control
PD35780c	sub	9	5073770	G	T	0.0082	JAK2	p.V617F	Missense	Control
PD35786b	sub	2	25457243	G	A	0.0093	DNMT3A	p.R882C	Missense	Control
PD35786b	sub	2	25463586	C	T	0.2100	DNMT3A	p.G699D	Missense	Control
PD35786c	sub	2	25463586	C	T	0.3100	DNMT3A	p.G699D	Missense	Control
PD35788b	sub	2	25458695	C	T	0.0373	DNMT3A	p.?	Essential splice	Control
PD35788b	sub	2	25466790	G	A	0.0550	DNMT3A	p.S638F	Missense	Control
PD35788b	sub	20	31023963	G	T	0.0353	ASXL1	p.G1150*	Nonsense	Control
PD35788c	sub	2	25458695	C	T	0.0381	DNMT3A	p.?	Essential splice	Control
PD29962a2	sub	4	106157653	G	T	0.022	TET2	p.E852*	Missense	AML diagnosis
PD29962a2	sub	11	119158556	GAATAGCAGC	T	0.076923	CBL	p.?	Missense	AML diagnosis
PD30054a2	sub	12	112888163	G	T	0.059	PTPN11	p.G60V	Missense	AML diagnosis
PD30054a2	sub	21	44514777	T	G	0.2	U2AF1	p.Q157P	Missense	AML diagnosis
PD30089d2	sub	9	5073770	G	T	0.034	JAK2	p.V617F	Missense	AML diagnosis
PD30089d2	sub	11	119167619	GGGGAGCAAT	A	0.101695	CBL	p.?	Missense	AML diagnosis
PD30089d2	sub	X	129147566	A	AC	0.113208	BCORL1	p.L275fs*145	Missense	AML diagnosis

Appendix 10: AML risk prediction model coefficients

Cox proportional hazards model trained on the discovery cohort

Variable	Coefficient*	*P*- value
ASXL1	0.964	2.97E-40
CALR	0.465	1.94E-01
CBL	0.178	3.21E-01
DNMT3A	0.370	2.64E-09
IDH1	1.185	1.41E-12
IDH2	0.403	4.22E-04
JAK2	0.953	8.25E-26
KDM6A	0.962	1.98E-48
KMT2C	1.193	1.54E-04
KRAS	0.905	3.75E-32
NF1	0.924	6.25E-35
PHF6	1.073	4.50E-62
PTPN11	1.251	1.10E-30
RUNX1	0.389	1.09E-08
SF3B1	1.550	1.21E-23
SRSF2	0.692	5.53E-16
TET2	0.323	1.33E-03
TP53	2.403	4.42E-30
U2AF1	1.966	9.67E-28
age	-0.090	3.68E-01
gender	-0.046	6.78E-01

Cox proportional hazards model trained on validation cohort

Variable	Coefficient*	*P*- value
ASXL1	0.735	7.54E-11
CBL	0.224	4.77E-01
DNMT3A	0.202	3.75E-04
JAK2	-0.085	7.22E-01
KMT2C	0.519	6.13E-02
KMT2D	0.013	9.51E-01
KRAS	0.614	2.37E-09
NF1	0.386	8.88E-02
NRAS	0.483	2.81E-07
RAD21	0.439	8.16E-03
SF3B1	0.392	1.16E-01
SRSF2	0.379	5.58E-08
TET2	0.329	5.11E-22
TP53	1.233	8.49E-08
U2AF1	1.587	8.08E-17
age	0.019	7.50E-01
gender	-0.014	8.88E-01
systolic_BP_100	0.017	7.04E-01
diastolic_BP_100	0.039	1.89E-01
BMI_10	0.153	6.88E-02
Total_cholesterol_10	0.002	8.77E-01
Triglycerides	-0.034	7.69E-01
HDL	-0.121	1.51E-01
LDL	0.132	2.48E-01
Lymphocytes	0.080	4.40E-01
MCV_100	-0.024	2.27E-03
RDW_10	0.067	5.41E-05
WBC_10	0.008	8.76E-01
PLT_100	0.084	3.99E-01
HGB_10	0.037	1.28E-01

Cox proportional hazards model trained on combined cohort

Variable	Coefficient*	*P*- value
ASXL1	0.986	7.20E-50
BCOR	1.058	8.00E-78
CBL	0.200	2.69E-01
DNMT3A	0.331	2.31E-09
IDH1	1.203	3.60E-13
IDH2	0.418	1.24E-04
JAK2	0.930	1.24E-21
KDM6A	0.960	2.67E-55
KMT2C	1.166	9.17E-04
KMT2D	0.079	7.41E-01
KRAS	0.982	2.13E-31
NF1	0.785	3.10E-04
NRAS	1.145	5.03E-76
PHF6	1.101	2.07E-71
PTPN11	1.074	4.45E-12
RAD21	0.909	4.59E-13
RUNX1	0.403	1.36E-09
SF3B1	1.539	5.35E-23
SRSF2	0.678	8.33E-20
TET2	0.477	3.08E-16
TP53	2.502	1.35E-37
U2AF1	2.047	2.60E-35
age	-0.101	2.40E-01
gender	-0.053	6.07E-01
cohort	0.020	8.35E-01

* Gene coefficients indicate risk per 10% increase in VAF; P-values for the coefficients are calculated by Wald test

Ridge regularised logistic regression model trained on discovery cohort

Variable	Coefficient
ASXL1	0.846
CALR	0.626
CBL	0.428
DNMT3A	0.479
IDH1	0.786
IDH2	0.849
JAK2	0.882
KDM6A	0.738
KMT2C	0.764
KRAS	0.733
NF1	0.735
PHF6	0.765
PTPN11	0.736
RUNX1	0.384
SF3B1	0.836
SRSF2	0.906
TET2	0.523
TP53	1.068
U2AF1	0.983
age_10	-0.116
gender	-0.026
Av. Genes	0.740

Ridge regularised logistic regression model trained on validation cohort

Variable	Coefficient*
ASXL1	0.809
CBL	0.312
DNMT3A	0.303
JAK2	0.606
KMT2C	0.643
KMT2D	0.195
KRAS	0.653
NF1	0.525
NRAS	0.561
RAD21	0.542
SF3B1	0.479
SRSF2	0.384
TET2	0.437
TP53	1.049
U2AF1	1.233
age_10	0.080
gender	-0.086
systol_100	-0.133
diastol_100	0.203
bmi_10	0.391
cholestl_10	0.011
triglyc	-0.011
hdl	-0.303
ldl	0.040
lym	0.012
mcv_100	-0.242
rdw_10	0.720
wbc_10	-0.067
plt_100	0.143
hgb_10	0.401
Av. Genes	0.581

Ridge regularised logistic regression model trained on combined cohort

Variable	Coefficient*	CI.2.5%	CI.97.5%
ASXL1	0.876	0.657	1.087
BCOR	0.690	0.577	0.939
CBL	0.370	0.123	0.988
DNMT3A	0.406	0.222	0.652
IDH1	0.725	0.617	0.935
IDH2	0.786	0.616	1.021
JAK2	0.826	0.662	1.115
KDM6A	0.665	0.556	0.927
KMT2C	0.698	0.566	0.944
KMT2D	0.321	0.171	0.856
KRAS	0.676	0.559	0.951
NF1	0.651	0.539	0.908
NRAS	0.664	0.558	0.925
PHF6	0.691	0.588	0.943
PTPN11	0.676	0.576	0.926
RAD21	0.660	0.554	0.923
RUNX1	0.364	0.168	0.914
SF3B1	0.758	0.606	0.979
SRSF2	0.684	0.385	1.080
TET2	0.407	0.223	0.917
TP53	1.070	0.818	1.314
U2AF1	1.032	0.786	1.321
age_10	-0.058	-0.183	0.039
gender	-0.013	-0.241	0.196
cohort	-0.573	-0.853	-0.293
Av. Genes	0.668	0.558	0.929

* Gene coefficients indicate risk per 10% increase in VAF

Appendix 11: AML prediction model based on electronic health record data

AML case ascertainment from Clalit database

Cases included with diagnosis 205.0*	1696

Exclusion criteria	Number of retained cases
Prior diagnosis among the following: •ESSENTIAL THROMBOCYTHEMIA •HIGH/LOW GRADE MYELODYSPLASTIC SYNDROME LESIONS •MYELODYSPLASTIC SYNDROME WITH 5Q DELETION •MYELODYSPLASTIC SYNDROME, UNSPECIDIED •POLYCYTHEMIA VERA •MYELOFIBROSIS •OPERATIONS ON BONE MARROW AND SPLEEN •CHRONIC MYELOMONOCYTIC LEUKEMIA •CHRONIC MYELOID LEUKEMIA	1431
Received medications suggesting alternative diagnosis: •IMATINIB •DASATINIB •METHOTREXATE •TRETINOIN •DASATINIB •ANAGRELIDE •HYDROXYCARBAMIDE •ASPARAGINASE •PEGASPARGASE •ARSENIC TRIOXIDE	1210
No record of hospitalisation near time of diagnosis	1042
Age < 18	960
Received 6-mercaptopurine post diagnosis •Multiple doses •Combined with ALL diagnosis	929
Filter on onset year >=2003	875
Total number of AML cases retained	**875**

Laboratory test result variables

Parameters included in clinical model	
Haematocrit (HCT)	SPECIFIC GRAVITY
Mean corpuscular volume (MCV)	CK-CREAT.KINASE(CPK)
Red blood cell count (RBC)	PT-INR
Haemoglobin (HGB)	MICRO%/HYPO%
mean corpuscular hemoglobin (MCH)	VITAMIN B12
mean corpuscular hemoglobin concentration (MCHC)	IRON
White blood cell count (WBC)	PT %
Platelet count (PLT)	Prothrombine time (PT- SEC)
Lymphocyte percentage (LYM%)	Chloride (Cl)
Neutrophil percentage (NEUT%)	LIPEMIC
Eosinophil percentage (EOS %)	ICTERIC
Monocyte percentage (MON%)	HEMOLYTIC
Basophil percentage (BASO %)	HEMOGLOBIN A1C CALCULATED
Absolute lymphocyte count (LYMP.abs)	CH
Absolute neutrophil count (NEUT.abs)	GLOBULIN
Absolute eosinophil count (EOS.abs)	FERRITIN
Absolute monocyte count (MONO.abs)	T4- FREE
BASOPHILES (abs)	APTT-sec
Mean platelet volume (MPV)	FOLIC ACID
Red cell distribution width (RDW)	PDW
CREATININE- BLOOD	Myeloperoxidase index (MPXI)
GLUCOSE- BLOOD	TRANSFERRIN
UREA- BLOOD	PCT
SODIUM	CHOLESTEROL HDL RATIO
POTASSIUM	BILIRUBIN INDIRECT
GLUTAMIC OXALOACETIC TRANSAMINASE	HCT/HGB RATIO
GLUTAMIC PYRUVIC TRANSAMINASE	CREATININE URINE SAMPLE
MICR %	SEDIMENTATION RATE
HYPO %	ERYTHROCYTES
MACRO%	LEUCOCYTES
PHOSPHATASE- ALKALINE	C-REACTIVE PROTEIN (CRP)
CHOLESTEROL	RDW-CV
TRIGLYCERIDES	M.ALBUM/CREAT RATIO
LUC%	AMYLASE- BLOOD
LUC	MICROALBU U SAMP
CHOLESTEROL- HDL	PROTEIN
CALCIUM- BLOOD	MAGNESIUM- BLOOD
HYPER%	Hemoglobin distribution width
URIC ACID- BLOOD	FIBRINOGEN
CHOLESTEROL- LDL	SODIUM- BLOOD
BILIRUBIN TOTAL	VITAMIN D3- 25-0H- RIA
ALBUMIN	POTASSIUM- BLOOD
PROTEIN-TOTAL-BLOOD	RDW-SD
PHOSPHORUS- BLOOD	Prostate specific antigen (PSA)
TSH (THYROID STIMULATING HORMONE)	T3- FREE
LACTIC DEHYDROGENASE (LDH) -BLOOD	Activated partial thromboplastin time
GAMMA-GLUTAMYL TRANSPEPTIDASE	NORMOBLAST.%
BILIRUBIN- DIRECT	ESTRADIOL (E-2)
NON-HDL_CHOLESTEROL	Absolute normoblast count
PH-u	Leutinising hormone (LH)

Diagnostic code variables

Diagnoses included in clinical model
ACUTE BRONCHITIS
ACUTE NASOPHARYNGITIS (COMMON COLD)
ANEMIA OTHER/UNSPECIFIED
ANEMIA, UNSPECIFIED
BACK SYMPTOMS/COMPLAINTS
CELLULITIS AND ABSCESS OF UNSPECIFIED SITES
CHRONIC RENAL FAILURE
COLITIS,ENTERITIS,GASTROENTERITIS PRESUMED INFECTIOUS ORIGIN
CONGESTIVE HEART FAILURE
CONTACT DERMATITIS AND OTHER ECZEMA, UNSPECIFIED CAUSE
DEBILITY, UNSPECIFIED
DERMATOPHYTOSIS OF FOOT
DISEASES AND CONDITIONS OF THE TEETH AND SUPPORT.STRUCTURES
DISTURBANCE OF SKIN SENSATION
ESSENTIAL HYPERTENSION
FEVER
INFERTILITY, FEMALE
IRON DEFICIENCY ANEMIA, UNSPECIFIED
MIXED DISORDERS OF CONDUCT EMOTIONS
OSTEOARTHROSIS AND ALLIED DISORDERS
PAIN IN LIMB
PNEUMONIA, ORGANISM UNSPECIFIED
VARICOSE VEINS OF LOWER EXTREMITIES

Appendix 12: Discovery cohort pre-lymphoid neoplasm cases and controls metadata

Individual ID	Sample ID	Group	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	BMI	Total cholesterol (mmol/L)	HDL (mmol/L)	LDL (mmol/L)	Triglycerides (mmol/L)	Lymphocytes (10^9/L)	MCV (fL)	RDW	WBC (10^9/L)	RBC (10^9/L)	Haematocrit (%)	Platelets (10^9/L)	Haemoglobin (g/dL)	HbA1c (%)	Age at first sample	Age at sample	Follow-up (years)
PD00001	PD00001a	Control	Female	138	80	36.6	6.1	1.5	4.1	1.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	56.7	56.7	23.9
PD00001	PD00001c	Control	Female	140	72	39.1	4.2	1.8	1.8	1.4	1.6	95.8	14.3	7.6	4.4	0.4	243	14.3	5.5	56.7	71.2	23.9
PD00003	PD00003b	Pre-LN	Female	147	92	27.1	5.2	1.7	2.7	1.9	2	89.1	12	6.6	4.1	0.4	246	13.3	5.2	62.4	62.4	15.7
PD00003	PD00003c	Pre-LN	Female	150	76	26.6	4.9	1.6	2.7	1.5	2.6	90	NA	11	4.5	0.4	250	13.7	5.5	62.4	70.9	15.7
PD00004	PD00004a	Pre-LN	Female	125	76	23.1	6.2	1.8	4.1	0.7	1.6	96.1	13	4.4	4.1	0.4	260	12.7	4.4	62	62	16
PD00004	PD00004b	Pre-LN	Female	122	70	23.5	7.1	2.2	4.7	0.5	1.8	98.7	13.2	4.6	4.5	0.4	192	15.5	4.9	62	65.3	16
PD00005	PD00005b	Control	Male	130	72	27.8	6.1	1.4	3.7	2.4	2.2	86.1	12.7	8	4.8	0.4	267	13.3	5.6	59.3	59.3	19.5
PD00005	PD00005c	Control	Male	126	70	29	5.7	1.3	3.9	1.1	2.8	87	NA	7	4.8	0.4	262	14.3	5.4	59.3	68.2	19.5
PD00011	PD00011b	Control	Female	158	93	23.8	6.3	1.5	4.3	1.1	1.8	87.1	13	6.1	4.2	0.4	271	12.4	5.7	66	66	19.5
PD00017	PD00017b	Pre-LN	Female	139	87	31.4	7.4	1.1	4.8	3.4	2	92.6	13.8	6.7	4.7	0.4	189	14	6.2	66.6	66.6	6.8
PD00021	PD00021a	Control	Male	129	84	28	7.6	1.6	5.5	1.1	1.5	88.8	13	5.7	4.7	0.4	311	14.6	4.9	57.8	57.8	13.7
PD00022	PD00022a	Control	Male	134	78	22.6	5.6	1.1	4	1	1.2	76.2	16.4	5	4.6	0.3	441	11.7	NA	71.8	71.8	8.7
PD00023	PD00023b	Control	Male	126	76	29.5	4.8	0.9	3.4	1.3	2.4	95.2	12.6	7.7	4.5	0.4	254	14.1	5.6	59	59	19.9
PD00023	PD00023c	Control	Male	108	74	28.3	3.8	1	2.4	1.2	1.7	92	NA	6.8	4.3	0.4	256	13.5	6	59	68	19.9
PD00026	PD00026b	Pre-LN	Male	149	91	23.7	5.9	0.9	4	2.2	1.6	92.7	12.6	5.2	4.4	0.4	206	14.4	5.4	64.7	64.7	12.8
PD00026	PD00026c	Pre-LN	Male	164	86	25.6	6.2	1	4.5	1.7	1.1	97.7	13.8	5.8	4.4	0.4	231	14	6	64.7	75	12.8
PD00031	PD00031a	Control	Male	107	61	25.7	5.8	1.1	4.2	1.2	2.2	91.2	13.7	6	4.5	0.4	144	14.1	NA	68.2	68.2	22.7
PD00031	PD00031c	Control	Male	132	82	28.5	2.8	1.3	1.3	0.5	1.2	95.9	14.7	7.2	4.1	0.4	118	13.3	5.8	68.2	80.5	22.7
PD00034	PD00034b	Control	Female	146	88	27.7	5.2	1.6	2.9	1.7	1.7	87.2	12.3	4.3	4.1	0.4	253	12.6	5.1	52.3	52.3	18.7
PD00034	PD00034c	Control	Female	180	92	27.6	5.4	1.7	3.1	1.5	1.7	88.9	13.2	5.9	4.4	0.4	291	13.2	5.3	52.3	60.5	18.7
PD00035	PD00035a	Pre-LN	Male	129	78	26.2	6.1	1.7	3.9	1	1.3	95.7	13.6	3.7	4.1	0.4	166	13.5	4.4	70.8	70.8	17.7
PD00035	PD00035b	Pre-LN	Male	132	82	27.8	6.6	1.5	4.2	2	1.2	95.3	13.2	5.3	4.1	0.4	185	14	5.2	70.8	74.5	17.7
PD00036	PD00036a	Control	Male	128	88	27	4.1	0.8	2.3	2.1	2.5	93.1	13	6.7	4.9	0.5	258	16.1	5.9	58.9	58.9	8.2
PD00038	PD00038a	Pre-LN	Female	122	78	23.6	5.5	0.9	3.8	1.8	2.3	89.8	12.5	6.7	4.1	0.4	377	12.1	5.3	49.4	49.4	18.4
PD00038	PD00038b	Pre-LN	Female	108	70	23.2	3.7	1.5	1.6	1.4	2	92.9	13.4	5.1	4.1	0.4	298	12.8	5.1	49.4	53	18.4
PD00038	PD00038c	Pre-LN	Female	115	68	22.1	6	1.6	3.4	2.2	1.6	95.5	13.8	3.7	4	0.4	294	13	5.8	49.4	63.8	18.4
PD00041	PD00041b	Control	Male	112	70	24.5	3.7	1.3	2.1	0.8	1.5	86.5	13.9	5.9	4.9	0.4	213	14.8	5.4	51.5	51.5	18.4
PD00042	PD00042a	Control	Female	146	84	19.9	6.6	1.6	4.5	1.1	1	90.5	13.4	3.5	4.6	0.4	332	14.3	NA	62.6	62.6	23.2
PD00043	PD00043b	Pre-LN	Male	142	81	26.6	6.1	1.1	4.1	2.1	1.7	92.2	13	5	4.8	0.4	214	14.2	6	68.9	68.9	11.4
PD00049	PD00049a	Pre-LN	Female	132	85	28.5	5.1	1.1	3.3	1.5	3.5	88.1	12.9	8.3	4.3	0.4	255	12.9	NA	61.5	61.5	16.7
PD00049	PD00049b	Pre-LN	Female	153	84	31.5	4.5	1.3	2.4	1.9	4	87.6	14.7	8.5	4.6	0.4	217	14.2	5.8	61.5	65.2	16.7
PD00051	PD00051b	Control	Female	142	82	26.9	5.4	1.2	3.3	2	2.3	89.1	13	7.4	4.5	0.4	255	14.4	5.6	65.8	65.8	18.2
PD00051	PD00051c	Control	Female	124	66	27.4	4.6	1.4	2.6	1.4	1.7	91.6	13.3	6.4	4.3	0.4	434	13.5	5.9	65.8	73.9	18.2
PD00063	PD00063a	Pre-LN	Female	136	74	24.8	5.8	2.9	2.5	0.8	4.3	95.4	12.7	8.8	3.9	0.4	360	12.7	NA	62.8	62.8	7.2
PD00065	PD00065b	Pre-LN	Female	128	76	31	6.5	1.9	4.2	0.9	2.2	85.8	13.4	7.4	4.1	0.3	287	12.2	5.7	54.2	54.2	2.5
PD00068	PD00068a	Control	Female	109	72	26.7	5.5	1.3	3.5	1.6	1.6	87.2	12.5	4.5	5	0.4	222	14.5	NA	47.4	47.4	20
PD00068	PD00068c	Control	Female	114	68	29.3	4.7	1.4	1.5	4.1	2.4	83	NA	7.4	4.3	0.4	212	12.3	5.4	47.4	55.6	20
PD00069	PD00069b	Pre-LN	Male	163	117	26.9	5.8	1.8	3.5	1.1	NA	93.7	13.3	NA	4.2	0.4	177	12.9	5.7	72.8	72.8	2.9
PD00070	PD00070b	Control	Female	156	90	24.7	6.4	2.8	3.4	0.6	2.1	88.8	13.3	6.5	4.2	0.4	236	13.2	5.2	70.3	70.3	19.4
PD00070	PD00070c	Control	Female	138	73	23	5.5	2.2	3.1	0.6	1.4	94.9	15.2	7.7	4	0.4	262	12.8	5.5	70.3	80.3	19.4
PD00071	PD00071b	Control	Female	157	84	27.5	5.7	2.7	1.6	3.2	2.1	90	12.4	6.2	3.7	0.3	162	12.2	5.7	71.5	71.5	17
PD00073	PD00073b	Control	Female	129	80	26	4.6	1.5	2.6	1.2	1.8	92.5	14	7	4	0.4	301	13	5.6	70.8	70.8	17.8
PD00074	PD00074b	Control	Male	129	84	24.5	6.5	1	3.8	3.8	1.2	87.6	14	4.9	5	0.4	186	14.9	5.2	59.3	59.3	18.4
PD00075	PD00075b	Control	Male	122	66	27.2	6.6	1.3	4.3	2.4	2.2	90.4	13.4	7.8	4.5	0.4	196	14.1	5.5	71.6	71.6	8.2
PD00076	PD00076b	Pre-LN	Male	153	92	27.4	4.3	0.7	2.9	1.6	2.5	94.6	12.8	5.4	3.6	0.3	147	12.6	5.3	52.6	52.6	12.6
PD00077	PD00077b	Control	Female	140	82	27.1	5.3	2.1	2.8	1	1.8	86.1	14.2	4.8	4.5	0.4	167	13.7	5.1	65.2	65.2	18.4
PD00079	PD00079b	Pre-LN	Male	178	106	25.9	6	0.9	3.5	3.6	4	92.9	12.6	10.2	4.8	0.4	135	14.4	5.5	69.5	69.5	5.3
PD00080	PD00080a	Control	Male	138	79	24.6	7.1	1.3	4.2	3.6	2.5	97	13.3	6.8	4.8	0.5	251	15	4.6	61	61	21
PD00084	PD00084b	Control	Female	150	92	24.9	7.2	1.6	5	1.5	3	93.8	13.3	7.3	3.5	0.3	282	11.8	5.5	57.7	57.7	19
PD00089	PD00089a	Pre-LN	Male	168	95	25	5.9	1.5	3.9	1.2	2	91.6	13.3	5	5.1	0.5	251	16.1	9.1	69	69	4.5
PD00089	PD00089b	Pre-LN	Male	154	83	26.2	5	1.5	3.1	0.9	2.9	92.1	14.5	7	6.2	0.6	196	20.7	9.9	69	72	4.5
PD00094	PD00094b	Control	Male	132	74	27.4	6.3	2.1	3.8	0.9	1.9	93.5	13	5.8	4.5	0.4	247	14.7	4.9	73.8	73.8	18.6
PD00095	PD00095b	Pre-LN	Male	145	100	28.9	7.6	1	4.3	5.1	2.4	92.2	16.1	5.4	5.1	0.5	251	15.1	5.2	58	58	9.4
PD00097	PD00097a	Pre-LN	Female	121	75	29	6.8	1.4	4.8	1.4	1.6	89.4	15.2	4.9	4.2	0.4	273	12	5.3	64.5	64.5	13.7
PD00097	PD00097b	Pre-LN	Female	141	79	32.5	6.7	1.5	4.7	1.1	1.5	86.5	15.1	6.7	4	0.3	252	12.2	5.7	64.5	67	13.7
PD00099	PD00099b	Control	Male	155	98	24.9	5.2	1.5	3	1.7	2.3	92.1	14.2	8.1	5.1	0.5	323	16.2	5.3	63.3	63.3	18.6
PD00100	PD00100a	Control	Female	122	80	29.3	6.4	2.1	4	0.7	1.5	86.9	13.8	5.3	4.2	0.4	227	12.2	NA	64	64	21.9
PD00100	PD00100c	Control	Female	103	58	30.9	5.7	1.6	3.4	1.6	1.9	91.1	14.2	7.3	4	0.4	230	12.4	5.9	64	75.4	21.9
PD00103	PD00103a	Control	Male	104	64	17.6	5	1.5	3.1	0.8	0.8	90.8	12.7	2.3	4.3	0.4	182	13.7	5.2	48.3	48.3	21.8
PD00103	PD00103c	Control	Male	106	62	18.2	4.1	1.4	2.4	0.7	0.6	92	NA	2.4	3.8	0.3	185	12.2	5.4	48.3	58.6	21.8
PD00106	PD00106a	Pre-LN	Female	124	79	23	5.2	1.7	3.1	0.8	1.3	88.6	13.9	5.4	4.2	0.4	303	12.7	NA	52.2	52.2	13.7
PD00106	PD00106b	Pre-LN	Female	140	89	24.9	4.7	2.1	2.3	0.8	1.5	91.3	13.1	4.1	4	0.4	274	12.7	4.5	52.2	56.2	13.7
PD00107	PD00107b	Pre-LN	Male	120	76	24.2	4.7	1.2	3.1	0.9	1.6	95.8	14	4.3	4.2	0.4	250	14.1	7.6	61.9	61.9	11.7
PD00107	PD00107c	Pre-LN	Male	136	70	23.7	3.8	1.6	1.8	1.1	1.4	96.5	14.5	4.7	4.2	0.4	260	13.8	10.2	61.9	70.7	11.7
PD00110	PD00110a	Pre-LN	Female	150	90	26.6	6.1	1.6	3.9	1.3	1.9	92	12.9	9.7	4.3	0.4	228	14	NA	59.6	59.6	10.4
PD00110	PD00110b	Pre-LN	Female	152	94	29.3	7.1	1.3	4.6	2.8	2.1	94.7	13.6	6.1	3.7	0.4	224	12.2	5.1	59.6	63.5	10.4
PD00111	PD00111b	Pre-LN	Female	148	93	27.8	6.5	1.7	3.9	2.1	3.5	92.4	12.2	8.3	4.5	0.4	212	13.3	5.8	55.9	55.9	7.9
PD00112	PD00112a	Control	Female	122	74	24.2	6.9	1.3	4.7	2	2	93.8	12.7	6.4	4.2	0.4	255	12.3	4.8	65.2	65.2	20.1
PD00112	PD00112c	Control	Female	132	80	28.6	7.3	1.4	NA	4.8	1.5	95.5	13.6	6.2	4.3	0.4	241	13.8	5.8	65.2	75.2	20.1
PD00113	PD00113b	Pre-LN	Female	131	78	25.6	6.8	1.8	4	2.3	NA	NA	NA	NA	NA	NA	NA	NA	5.3	52	52	7.2
PD00115	PD00115a	Control	Male	144	92	25.6	6.1	1.3	4	1.9	2	92.3	12.3	7.7	4.1	0.4	273	13	5.8	69.4	69.4	15.6
PD00116	PD00116b	Control	Female	152	86	27.6	5.7	1.3	4	1	1.8	89.2	13.3	6.2	4.1	0.4	275	12.8	5.8	78.8	78.8	18.6
PD00117	PD00117b	Control	Female	125	70	22.9	6.1	2.1	3.6	0.9	2.5	88.5	12.6	8.2	4.7	0.4	261	14.2	5.2	66.4	66.4	18.5
PD00117	PD00117c	Control	Female	147	83	23.6	4.3	1.9	2	1	1.7	92.2	13.4	6.7	4.6	0.4	197	13.9	5.5	66.4	74.6	18.5
PD00121	PD00121b	Control	Female	173	101	50	5.5	0.7	3.1	3.9	3.4	87.7	13.8	9.1	4.2	0.4	191	12.7	8.8	54.9	54.9	19.1
PD00125	PD00125a	Pre-LN	Female	113	72	22.7	5.1	2	2.8	0.6	1.7	97.4	13	4	3.8	0.4	127	12.5	NA	49.4	49.4	16.5
PD00127	PD00127b	Control	Male	156	95	26.6	9.2	1.4	6.8	2.3	1.4	93.9	15.2	4.6	5.2	0.5	290	16.3	5.1	67.5	67.5	17.8
PD00129	PD00129b	Control	Male	132	77	25.8	4.6	1.3	2.2	2.6	1.8	93.1	13.7	9.1	4.8	0.4	285	14.2	5.7	72	72	10.4
PD00130	PD00130a	Control	Female	116	76	23	5.6	1.6	3.5	1.1	1	88	14	4	4.1	0.4	238	12.1	5	50.9	50.9	20.9
PD00132	PD00132a	Pre-LN	Female	142	76	23.7	6.8	1.8	4.5	1.2	3.2	92	14.3	7.6	4.3	0.4	172	13.4	5.6	72.1	72.1	13.3
PD00132	PD00132b	Pre-LN	Female	142	78	24.3	7.1	1.8	4.8	1.3	3.7	96	13.8	8.4	4.2	0.4	164	13.2	5.1	72.1	75.3	13.3
PD00135	PD00135b	Pre-LN	Female	156	84	23.6	6.1	1.5	3.9	1.6	1.6	87.1	15.2	8.2	4.7	0.4	324	13.7	5.1	67.9	67.9	6.1
PD00140	PD00140b	Control	Male	130	79	26.5	5.7	1.6	3.6	1.3	1.7	96.8	13.1	5.8	4.7	0.4	279	14.6	5.3	72.8	72.8	11.4
PD00142	PD00142b	Control	Female	137	79	32.8	8	1.4	5.7	2	1.9	90.4	14.9	6.4	4.4	0.4	157	12.7	5.9	62.8	62.8	19.7
PD00142	PD00142c	Control	Female	130	70	31.8	5.1	1.5	3.1	1.3	1.6	94.4	14.8	6.7	4.2	0.4	152	13.3	6.2	62.8	73	19.7
PD00148	PD00148b	Control	Male	152	88	29.2	6.1	0.9	3.7	3.4	2	92.3	13.5	6.9	5.2	0.5	264	15.5	5.5	73.6	73.6	20
PD00148	PD00148c	Control	Male	129	64	29.9	5	0.9	2.6	3.3	1.9	86	NA	7.5	5.1	0.4	241	15.4	5.4	73.6	82.3	20
PD00152	PD00152b	Control	Female	126	72	30	4.9	1.4	2.7	1.9	2.1	92.1	14.8	5.1	4.3	0.4	214	13.2	5.3	71.3	71.3	17.8
PD00153	PD00153a	Pre-LN	Male	120	76	25.6	7.9	2	5.6	0.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	50.3	50.3	15.1
PD00153	PD00153b	Pre-LN	Male	133	86	26.5	5.8	1.9	3.6	0.8	1	92.1	13.1	3.5	4.9	0.4	215	14.3	4.9	50.3	54.5	15.1
PD00159	PD00159b	Control	Female	164	89	29.2	7	1.4	4.3	2.9	1.7	95.9	11.8	6.9	4.3	0.4	310	13.5	5.7	76.2	76.2	13.3
PD00160	PD00160b	Control	Male	140	84	29.6	7.2	0.9	3.8	5.6	2.9	89.1	12.3	6.9	4.5	0.4	257	14.6	5.5	68.3	68.3	18.8
PD00163	PD00163b	Control	Female	114	66	22.8	4.7	1.4	3	0.7	1	87.8	12.8	4.6	4.6	0.4	298	13.2	5.6	50	50	19.4
PD00164	PD00164b	Control	Male	130	82	24.3	5.4	0.9	3.4	2.6	4.2	102.1	13.6	9.5	4.5	0.5	216	15.5	5.4	55.6	55.6	18.6
PD00166	PD00166b	Control	Female	126	72	21.2	4.5	1.5	2.6	0.9	3.1	95.8	13.9	11.1	4.5	0.4	249	14	5.5	73.3	73.3	19.4
PD00166	PD00166c	Control	Female	135	70	22.1	5.1	2.8	1.9	1	1.8	92.7	14.5	7.9	3.9	0.4	265	11.9	5.8	73.3	83.7	19.4
PD00170	PD00170b	Control	Female	152	96	19.9	4.6	2.6	1.7	0.7	1.4	92.6	12.9	5.6	4.3	0.4	138	12.9	5.2	56.4	56.4	19.6
PD00170	PD00170c	Control	Female	156	93	19.3	6	2.8	2.8	0.9	NA	92.5	13.8	4.6	4.3	0.4	186	13.1	5.4	56.4	66.9	19.6
PD00171	PD00171b	Control	Female	145	84	25.3	6.7	2.2	4.2	0.7	2.5	93.6	12.9	6.5	5.2	0.5	228	14.9	5.2	65.2	65.2	17.7
PD00171	PD00171c	Control	Female	162	89	24.6	6.3	2.4	3.4	1.3	1.6	92.7	13.9	5.7	4.6	0.4	183	14.5	5.8	65.2	74.2	17.7
PD00172	PD00172b	Control	Female	114	70	23.7	8.4	1.3	6.6	1.1	1.4	90	14.7	4.9	4.7	0.4	243	13.3	5.7	72.9	72.9	19.8
PD00172	PD00172c	Control	Female	140	82	22.5	6.9	1.6	5	0.8	1.2	89.1	13.9	6.2	4.5	0.4	242	13.5	6	72.9	83.4	19.8
PD00176	PD00176a	Control	Female	140	83	21	6	1.5	3.5	2.2	1.8	88.3	12.6	8.6	4.3	0.4	241	12.9	5.9	57.5	57.5	15.6
PD00176	PD00176c	Control	Female	116	68	20.9	5.3	1.4	3	2.1	1	93.9	14.6	4.9	3.8	0.4	212	11.8	5.4	57.5	71.7	15.6
PD00177	PD00177a	Control	Female	165	90	31	5.4	1.5	3.4	1.2	1.7	89.2	12.8	5.6	3.9	0.3	246	12.1	NA	67.5	67.5	10.9
PD00179	PD00179b	Pre-LN	Male	174	104	33.4	4.6	0.9	3.2	1.1	2.1	94.9	12.9	6.2	4.9	0.5	185	15.5	5	71.7	71.7	4.3
PD00185	PD00185b	Pre-LN	Male	154	98	25.8	5.9	0.7	2.7	5.6	2	85.2	16.2	5.4	4.9	0.4	195	15.2	4.9	71.8	71.8	14.5
PD00186	PD00186a	Pre-LN	Female	131	76	27.5	7.4	1.2	4.5	3.7	1.2	89.8	11.9	7.5	4.5	0.4	333	14.2	NA	56.2	56.2	13.2
PD00186	PD00186b	Pre-LN	Female	146	84	31.2	8.8	1.4	5.5	4.2	1.3	91.7	12.7	6.4	5.4	0.5	320	17.1	5.6	56.2	60.6	13.2
PD00192	PD00192a	Control	Female	106	66	23.9	5	1.8	2.7	1.1	2.2	91.6	11.4	6.2	3.7	0.3	272	11.6	NA	49.7	49.7	22.4
PD00195	PD00195a	Pre-LN	Female	157	87	26.6	6.9	1.7	4.7	1.2	2.2	85	14.1	6.4	4.5	0.4	257	13.2	5	68.7	68.7	14.7
PD00195	PD00195b	Pre-LN	Female	128	71	26.4	5.9	1.8	3.8	0.7	1.6	87.7	14.8	6.1	4.3	0.4	259	13	5.4	68.7	71.5	14.7
PD00197	PD00197b	Pre-LN	Female	126	70	24.2	7.2	1.5	5	1.7	1.7	88.5	13.5	4.6	4.7	0.4	258	13.8	5.4	75.8	75.8	12
PD00198	PD00198a	Control	Female	142	90	27.3	7.3	1.2	4.6	3.2	2	85.1	14	7.1	4.5	0.4	422	13.3	NA	67.3	67.3	22.5
PD00199	PD00199a	Pre-LN	Female	160	98	28.2	6	1.4	4	1.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	71.8	71.8	8.8
PD00199	PD00199b	Pre-LN	Female	130	76	27.2	5.8	1.8	3.4	1.5	2.2	92.8	13.6	6.3	4.1	0.4	187	12.8	5.9	71.8	76	8.8
PD00200	PD00200a	Pre-LN	Female	149	88	25.9	5.4	1.5	3.4	1.1	2.1	93.3	13.8	5.7	4.1	0.4	176	13.2	NA	77	77	4.8
PD00200	PD00200b	Pre-LN	Female	174	94	26.2	7.2	1.9	5	0.8	1.7	97.2	14.4	5	3.9	0.4	212	13.5	5.9	77	80.7	4.8
PD00203	PD00203a	Control	Female	152	92	25.4	7.4	0.8	5	3.4	1.5	86.2	12.2	3.9	4.8	0.4	259	14.4	NA	67.4	67.4	22.3
PD00205	PD00205b	Control	Male	142	90	27.6	5.3	1.9	3	1.3	1.2	91	13.5	5.8	5	0.5	331	15.2	5.2	56	56	17.8
PD00205	PD00205c	Control	Male	139	90	28	6.3	1.7	4.1	1.2	1.1	89.9	13.1	7.6	5.2	0.5	293	16.4	5.7	56	64.6	17.8
PD00206	PD00206b	Control	Female	126	72	26.4	5.6	2.5	2.6	1.2	1.2	95.8	14.3	5	4.1	0.4	200	13.5	5.7	70.7	70.7	18.4
PD00213	PD00213b	Control	Male	132	73	28.6	5.7	0.9	3.5	3	2.7	91.1	13.7	5.6	4.2	0.4	268	13.5	5.5	63.9	63.9	18.5
PD00214	PD00214b	Control	Female	118	66	21.2	5.9	1.8	3.8	0.7	2.4	93.8	12.4	7.7	4.6	0.4	193	14.2	5.3	63.2	63.2	19.5
PD00214	PD00214c	Control	Female	118	76	20	5.5	2.3	3	0.5	1.1	84.5	18	5.6	4.2	0.4	258	11.6	5.9	63.2	73.2	19.5
PD00217	PD00217a	Control	Male	127	69	21.3	9.4	1.9	6.8	1.5	1.5	84.9	13.7	4.8	4.4	0.4	224	13.1	NA	71.8	71.8	20.6
PD00218	PD00218b	Control	Female	139	84	29.7	6.1	1.6	3.4	2.5	2.5	83.4	14.1	7.3	4.5	0.4	319	13	5.5	58.3	58.3	19
PD00222	PD00222b	Control	Female	130	72	19.5	6.2	2.3	3.5	0.9	1.4	88.4	12.5	5.3	4.3	0.4	264	14.2	4.7	45.6	45.6	18.6
PD00222	PD00222c	Control	Female	130	78	22.9	6.3	1.8	4.2	0.8	NA	89.3	13.8	5.8	5.1	0.5	296	15.5	5.4	45.6	54.6	18.6
PD00225	PD00225b	Pre-LN	Female	127	68	22.3	5.5	1.5	3.2	1.9	2.3	91.5	11.7	6.1	4.3	0.4	187	14	4.9	66.6	66.6	5.9
PD00226	PD00226a	Pre-LN	Male	148	78	25.8	6.5	1	4.5	2.2	2.2	90.8	12.6	5.7	4.6	0.4	243	14.4	NA	68	68	5.4
PD00226	PD00226b	Pre-LN	Male	132	78	26.7	5.2	0.9	3.6	1.7	2	92.2	12.4	6.3	4.2	0.4	172	14.2	5.3	68	71.9	5.4
PD00227	PD00227b	Control	Female	136	76	29.2	8.3	1.4	5.8	2.6	2.5	83.5	13.8	6.5	4.4	0.4	311	13.1	6	62.4	62.4	19.1
PD00230	PD00230b	Control	Female	124	70	26.7	7.1	1.6	4.7	1.8	1.5	86.6	12.5	5	4.6	0.4	278	13.8	6	70.2	70.2	19

PD00230	PD00230c	Control	Female	142	62	26.8	4.4	1.7	2.1	1.4	1.8	88	NA	6.9	4.8	0.4	308	14.1	6	70.2	78.3	19
PD00239	PD00239b	Control	Female	132	77	20.7	6.9	2.7	3.8	1	1.2	87	13.5	5.2	4.6	0.4	258	13.9	5.3	72.8	72.8	17.9
PD00239	PD00239c	Control	Female	140	76	19.2	5.3	2.6	2.2	1.2	0.8	87.3	14.5	6.6	4.4	0.4	221	13.3	5.4	72.8	80.8	17.9
PD00241	PD00241b	Pre-LN	Female	140	78	25	6.3	1.4	4.3	1.4	1.5	89.5	13.2	5.2	4.2	0.4	259	12.7	4.7	60.4	60.4	1.5
PD00241	PD00241c	Pre-LN	Female	135	72	26.3	5.3	1.6	3.2	1.1	1.9	93	NA	5.6	4.3	0.4	219	13.6	5.6	60.4	69	1.5
PD00243	PD00243a	Control	Female	124	78	23.6	5.9	1.7	3.5	1.6	2	81.1	15.5	5.1	5	0.4	270	13.3	NA	51.8	51.8	23.4
PD00243	PD00243c	Control	Female	141	78	25.4	5.7	1.9	3.2	1.5	2.3	91.7	14.4	5.5	4.6	0.4	224	14	5.7	51.8	65.8	23.4
PD00247	PD00247a	Control	Female	156	90	30.8	5.9	1.3	3.6	2.1	1.9	86.1	12.7	5.7	4.2	0.4	376	12.5	6.1	70.4	70.4	21.2
PD00251	PD00251b	Pre-LN	Male	139	90	25.7	6.7	1.1	4.5	2.6	2.9	88.8	13.4	7.7	4.8	0.4	302	14.7	6	68.4	68.4	4.6
PD00253	PD00253a	Control	Male	134	87	26.2	5.3	0.9	3.2	2.5	2.1	88.5	13.4	7.4	5.2	0.5	318	16.2	5.6	56.8	56.8	22.1
PD00254	PD00254a	Pre-LN	Male	140	84	26.7	5.6	1.2	3.7	1.6	1.5	90.4	13.2	5	4.6	0.4	247	13.4	4.8	67	67	12.8
PD00254	PD00254b	Pre-LN	Male	130	80	29	6	1.3	3.6	2.5	1.8	92.5	12.6	5.4	4.2	0.4	176	14	5.1	67	70.4	12.8
PD00257	PD00257a	Control	Male	138	72	24.6	6.7	1.7	4.6	1	1.7	89.2	12.9	4.8	4.5	0.4	177	13.6	5	74.8	74.8	20.8
PD00259	PD00259b	Control	Female	125	78	27.6	5.5	1.7	3.5	0.7	2.7	87.3	12.9	7.8	4.6	0.4	223	14.7	5.1	56.4	56.4	19.1
PD00259	PD00259c	Control	Female	116	72	25.7	4.7	1.8	2.6	0.8	NA	87.2	15.9	3.6	4.8	0.4	188	14	6	56.4	65.9	19.1
PD00263	PD00263a	Pre-LN	Male	146	88	31.5	4.8	1	3.2	1.2	3	87	12.9	8.8	5.3	0.5	305	16.5	NA	48.7	48.7	15
PD00266	PD00266a	Control	Male	138	87	31	6.5	1.1	4.4	2.2	1.9	94.1	13.3	6	4.7	0.4	203	15.5	NA	68.4	68.4	23
PD00266	PD00266c	Control	Male	129	72	33.8	5.1	1	3.5	1.5	1.9	94	NA	7	4.1	0.4	179	13	NA	68.4	79.6	23
PD00270	PD00270a	Control	Male	124	84	23.1	6.4	1.6	4.2	1.5	2.7	86	12.2	7.6	4.8	0.4	283	14.8	5.5	53.5	53.5	21.1
PD00270	PD00270c	Control	Male	134	83	22.9	6	1.7	3.8	1.2	2.1	91.2	13.6	6.1	4.4	0.4	287	13.5	5.3	53.5	64.6	21.1
PD00272	PD00272b	Control	Male	128	86	28.7	5.5	0.9	3.8	1.9	2.2	90.8	12.6	6.2	4.5	0.4	220	13	5.9	54	54	19.8
PD00273	PD00273b	Pre-LN	Female	132	86	25.3	5.3	2	2.6	1.6	3	93	13.1	8.4	4.7	0.4	305	14.7	5.3	50	50	14.8
PD00275	PD00275b	Control	Male	136	98	29	7.6	1	6.1	1.2	1.8	91.7	12.2	6	5.2	0.5	206	16	5	56	56	19.9
PD00275	PD00275c	Control	Male	139	94	28.6	6.9	1.1	5.3	1.2	1.3	86.5	15.7	5.1	5.2	0.4	240	15.2	5.5	56	66.2	19.9
PD00276	PD00276a	Pre-LN	Female	170	103	25.8	6.5	1.4	3.8	2.8	1.7	90.1	13.6	6.8	4.6	0.4	327	14.1	6.3	75	75	8.6
PD00276	PD00276b	Pre-LN	Female	188	106	27.3	6.8	1.2	4.7	2.1	1.8	89.4	13.7	6.8	4.6	0.4	266	14.5	5.7	75	78.8	8.6
PD00277	PD00277b	Control	Male	111	78	24.6	6.1	1.3	3.7	2.6	1.9	92.2	13.5	5.3	4.8	0.4	315	14.8	5.4	51.6	51.6	19.1
PD00277	PD00277c	Control	Male	124	76	25.2	6.7	1.1	4.3	3	2.3	98.4	14.5	5.4	4.2	0.4	210	13.8	5.6	51.6	60.9	19.1
PD00281	PD00281b	Pre-LN	Female	116	69	29.2	5.7	1.5	3.4	1.9	1.5	93	12.8	4.4	4	0.4	315	12.5	5.4	65.8	65.8	13
PD00282	PD00282b	Pre-LN	Male	123	70	30.8	5	0.8	2.9	2.9	2.3	94.8	12.2	7	4.5	0.4	278	14.6	5.1	58.1	58.1	15.3
PD00282	PD00282c	Pre-LN	Male	131	71	33.6	5.4	0.8	3.1	3.3	2.7	97.3	13.6	6.5	4.3	0.4	269	14.3	5.6	58.1	68.3	15.3
PD00285	PD00285a	Pre-LN	Male	150	86	29.6	6.6	1	4.8	1.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	68	68	15.2
PD00287	PD00287a	Control	Female	110	70	22.5	7.2	1.6	4.9	1.5	2.8	93.8	12.4	7.1	4.3	0.4	313	14.1	NA	47.7	47.7	22.9
PD00289	PD00289b	Control	Male	147	88	23.8	7.5	1.5	4.8	2.8	1.4	86.3	14.3	4.8	4.4	0.4	220	14.2	5.6	65.3	65.3	19.2
PD00289	PD00289c	Control	Male	166	87	23.6	5.2	1.8	2.3	2.6	1.9	84.8	15.9	6.1	4.5	0.4	240	12.8	6.5	65.3	74.9	19.2
PD00292	PD00292b	Control	Female	147	80	20.9	10	2.3	6.9	1.8	2.6	95.9	12.3	8.5	4.1	0.4	282	14.1	5.9	72.6	72.6	19
PD00292	PD00292c	Control	Female	146	70	22.8	4	1.9	1.8	0.8	1.7	93.8	13	6.3	4.1	0.4	216	13.3	6.3	72.6	82.4	19
PD00294	PD00294b	Control	Male	153	90	25.9	5.6	1.1	3.5	2.3	1.3	90.5	13.6	6.2	4.2	0.4	255	13.9	5.8	77.5	77.5	18.4
PD00294	PD00294c	Control	Male	160	88	26.4	4.9	1.3	3.1	1.1	1.2	89.6	15.7	6.6	4.3	0.4	305	13.3	5.9	77.5	87.3	18.4
PD00297	PD00297b	Pre-LN	Female	136	82	21.9	6.7	1.9	4.1	1.7	2	92.3	13.9	5.9	4.4	0.4	240	13.3	6	55.9	55.9	10.4
PD00299	PD00299b	Pre-LN	Male	120	72	30.2	6.4	1.6	4.2	1.5	2.3	90.3	14.8	5.1	5.1	0.5	263	14.9	5.7	54.4	54.4	6.7
PD00301	PD00301b	Pre-LN	Male	144	92	29.4	5.2	1.5	2.7	2.3	1.6	90.8	13	6.5	5	0.5	171	15.3	5.8	66.9	66.9	2.2
PD00302	PD00302b	Control	Female	157	96	31.7	6.5	1.7	3.3	3.5	2.1	87.2	12.1	6	4.8	0.4	324	14.2	7.7	71.6	71.6	18.5
PD00304	PD00304a	Pre-LN	Female	160	95	28.5	5.4	1.2	2.5	3.6	2.3	85.9	12.6	7	4.3	0.4	294	13.3	5.7	62.6	62.6	8.2
PD00304	PD00304b	Pre-LN	Female	162	94	27	5	1.2	2.5	3	3.9	88	12.6	6.6	5.5	0.5	315	16.6	6.1	62.6	66.5	8.2
PD00304	PD00304c	Pre-LN	Female	142	79	26.5	4.3	1.2	1.9	2.7	1.7	90.2	13.9	8	3.8	0.3	309	11.8	6.7	62.6	75.5	8.2
PD00310	PD00310a	Pre-LN	Male	148	93	26	5.9	1	3.6	3	2.5	85.8	14.2	7.1	5.5	0.5	182	15.2	5.4	65.4	65.4	17.4
PD00310	PD00310b	Pre-LN	Male	174	102	26.6	6.8	1.1	3.5	5	2.2	82.2	14.6	8.3	5.6	0.5	192	16.3	5.5	65.4	67.9	17.4
PD00310	PD00310c	Pre-LN	Male	146	72	27.9	4	1.1	1.9	2.2	2.6	89.3	13.7	7.4	4.8	0.4	136	14.6	6.1	65.4	79.1	17.4
PD00312	PD00312b	Control	Male	130	74	26.7	4.8	1.4	3	0.9	1.7	97.7	12.1	6.2	3.9	0.4	194	14.3	5.2	57	57	19.1
PD00312	PD00312c	Control	Male	142	90	27.7	5.3	1.6	3.2	1.1	2	101.6	14.2	5.7	4	0.4	162	14.1	5.6	57	66.7	19.1
PD00318	PD00318a	Pre-LN	Female	120	76	23.3	6.8	1.3	4.7	1.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	53	53	20.2
PD00322	PD00322a	Pre-LN	Male	139	82	22	4.9	1.2	3	1.6	2	87.8	13.4	7.4	4.9	0.4	284	14.8	6	60	60	8.5
PD00322	PD00322b	Pre-LN	Male	146	88	22.9	5.1	1.2	3.2	1.6	1.9	89.2	13.2	8.5	4.9	0.4	264	15.1	5.7	60	63.8	8.5
PD00330	PD00330b	Pre-LN	Male	134	84	21.5	4.9	1	3.1	1.8	1.8	88.7	12.7	6.3	4.2	0.4	260	13.8	5.3	53.3	53.3	15.2
PD00330	PD00330c	Pre-LN	Male	116	61	22.3	4	0.8	2.8	1.3	2.1	89.7	14.5	4.9	3.7	0.3	142	11.1	5.3	53.3	62.7	15.2
PD00332	PD00332a	Pre-LN	Male	118	70	20.1	6.6	2	4	1.4	3.8	93.4	13	12.6	4.2	0.4	279	12.9	5.7	68	68	9.2
PD00332	PD00332b	Pre-LN	Male	92	60	19.5	6.6	1.9	4.2	1.1	2.4	94.7	13.6	8.3	4.1	0.4	341	13.1	6.3	68	71.9	9.2
PD00334	PD00334b	Control	Female	104	65	23.8	6.9	1.9	4.8	0.6	2	99.7	12.1	5	4.3	0.4	261	14.3	5	47.9	47.9	17.6
PD00336	PD00336b	Control	Male	106	74	26.6	6.2	1.2	4.4	1.4	1.7	88.5	13.3	5.6	4.3	0.4	178	14.2	5.7	65.8	65.8	11.3
PD00336	PD00336c	Control	Male	102	70	NA	4.2	1.4	2.3	1.3	1.5	90.4	14.2	6.8	4.1	0.4	218	12.8	5.9	65.8	74.9	11.3
PD00337	PD00337a	Control	Female	116	78	29.4	5.6	1.9	2.9	1.7	1.9	88	13.5	7.2	4	0.4	293	12	NA	49.5	49.5	22.2
PD00338	PD00338b	Pre-LN	Female	131	75	24.9	6.5	1.6	4.3	1.4	2.1	89.7	13.5	6.4	4.3	0.4	251	12.7	5.6	72	72	9.5
PD00341	PD00341b	Control	Female	148	98	24.8	5.1	1.7	2.8	1.4	2	92	12.8	8.5	4.6	0.4	245	15.7	4.8	55.7	55.7	18.2
PD00341	PD00341c	Control	Female	142	94	27.9	5	1.8	2.6	1.4	2.1	90	NA	6.2	5.1	0.5	306	15.2	5.3	55.7	63.8	18.2
PD00345	PD00345a	Pre-LN	Male	145	80	28.3	5.2	1.3	3.1	1.9	2.6	101	13	9.2	4.6	0.5	227	14.8	5.2	61.6	61.6	12.8
PD00345	PD00345c	Pre-LN	Male	145	89	25.4	4.5	1.7	2.6	0.6	1.9	104	14.7	6.9	4.2	0.4	191	14.5	5.6	61.6	73.7	12.8
PD00350	PD00350b	Control	Male	150	90	28.8	5.6	1.6	3.6	0.9	2.1	88.8	13.5	7.1	5.6	0.5	272	16.5	5.9	61.8	61.8	16.2
PD00351	PD00351a	Pre-LN	Female	120	76	28.4	6.6	1.8	4.4	0.9	2.5	86.5	13.5	6	4.4	0.4	336	13.5	NA	54.7	54.7	13.9
PD00353	PD00353b	Control	Male	133	93	34.7	6	0.9	4.4	1.7	1.8	91	13.8	5.8	5.2	0.5	214	16	5.4	71.5	71.5	18.7
PD00353	PD00353c	Control	Male	110	70	31.4	4.2	1.6	2.2	0.9	NA	95	16.9	6.8	5	0.5	184	15.8	6	71.5	80.5	18.7
PD00355	PD00355a	Control	Male	140	94	30.6	9.4	1.7	6.5	2.7	2.1	92.7	12.2	5.9	5.3	0.5	269	16.1	5.4	48.9	48.9	22.2
PD00355	PD00355c	Control	Male	134	90	31.6	5.3	1.4	2.8	2.5	2.7	93	NA	7.9	4.7	0.4	253	14.7	6	48.9	60.5	22.2
PD00356	PD00356b	Control	Female	155	88	27.2	6.6	1.7	3.5	3.1	2	92.2	11.8	5.2	4.7	0.4	148	15.2	9.7	76.3	76.3	13.9
PD00356	PD00356c	Control	Female	148	68	30.5	5.3	1.8	2.5	2.3	1.9	92.7	14	7.4	4.3	0.4	194	13.8	7.5	76.3	85.9	13.9
PD00360	PD00360b	Control	Female	122	73	26.6	6.1	1.4	4.2	1.1	1.7	88.2	12	5.7	4.5	0.4	190	13.8	5.2	56.6	56.6	18.1
PD00360	PD00360c	Control	Female	132	66	28.2	6.1	1.3	4.4	0.9	NA	93.7	12.7	3.7	4.2	0.4	182	13.2	5.5	56.6	66.6	18.1
PD00361	PD00361b	Control	Male	110	76	26	5.2	1.8	2.9	1.1	2.2	91.1	13.7	5.9	4.6	0.4	321	14.8	5.1	50.1	50.1	18.2
PD00363	PD00363b	Control	Male	154	92	31.4	5.8	1.1	3	3.9	2.6	91.8	13	7.8	4.5	0.4	340	14.1	6.5	74.5	74.5	18.5
PD00365	PD00365b	Control	Female	122	74	23.7	5.3	1.5	3.1	1.7	2.4	89.6	12.7	7.2	4	0.4	251	12.8	5.4	52.3	52.3	18.3
PD00365	PD00365c	Control	Female	112	82	23.7	6.5	1.5	3.6	3.1	2.5	93	14.4	7.6	4	0.4	246	12.9	5.9	52.3	63	18.3
PD00367	PD00367a	Control	Female	128	83	22.3	6.7	2.1	4.3	0.8	2.8	89.4	11.9	6.8	3.5	0.3	218	11.3	5	60.1	60.1	21.4
PD00369	PD00369b	Pre-LN	Female	125	74	22.7	6.3	1.6	3.9	1.8	1.4	88.9	15.1	3.8	5	0.4	214	14.4	5.6	68.8	68.8	12.1
PD00369	PD00369c	Pre-LN	Female	122	62	24.9	4	1.5	0.6	4.3	1.2	94.3	14.1	4.8	4.4	0.4	196	13.5	6	68.8	76.8	12.1
PD00371	PD00371b	Pre-LN	Male	142	67	21.6	4.6	0.8	3	1.9	1.6	93.4	14.3	5.5	3.6	0.3	259	11.4	5.9	66.7	66.7	2.2
PD00377	PD00377a	Pre-LN	Male	144	94	22.9	4.5	1.5	2.5	1.1	1.7	82.1	12.8	4.9	4.8	0.4	229	13.9	4.8	53.4	53.4	6.8
PD00377	PD00377b	Pre-LN	Male	132	79	22	4.4	1.7	2.3	1	1.5	86	13.3	7.5	4.9	0.4	228	14.5	5.2	53.4	56	6.8
PD00377	PD00377c	Pre-LN	Male	136	78	23.6	4.1	1.4	2.1	1.3	1.5	88.8	13	5.8	4.7	0.4	191	14.1	5.4	53.4	66.1	6.8
PD00379	PD00379b	Control	Male	140	83	25.2	4.3	1	2.9	0.9	2.5	81.2	14.3	7.8	5.2	0.4	136	13.8	6	74.1	74.1	13.1
PD00380	PD00380c	Control	Female	133	63	24.7	7.7	2.7	4.4	1.4	2.1	87.8	13.7	5.6	4.7	0.4	234	13.9	5.2	61.4	61.4	9.9
PD00385	PD00385b	Control	Male	117	76	24.8	6	1.5	3.3	2.7	2.5	84.9	14.4	7.3	4.9	0.4	178	15	5.6	58.3	58.3	18.9
PD00385	PD00385c	Control	Male	149	84	23.7	5.8	1.6	3.6	1.3	2.3	86.5	15.2	7	5.1	0.4	170	14.9	5.4	58.3	66.7	18.9
PD00386	PD00386a	Pre-LN	Female	110	62	22.4	3.4	1	1.9	1.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	47.3	47.3	8
PD00388	PD00388b	Control	Male	118	76	25.8	6.8	1.5	4.6	1.7	1.2	93.3	12.7	5.3	5.2	0.5	274	15.7	5.5	73.4	73.4	19.5
PD00389	PD00389b	Control	Female	150	87	31	7.3	1.2	5.4	1.6	1.5	96.2	12.2	4.8	4.3	0.4	215	13.6	5.4	66.4	66.4	19.8
PD00390	PD00390b	Control	Female	137	80	23.6	7.7	1.6	5.7	1	3.1	89.2	13	6.5	4.4	0.4	202	13.5	5.8	60	60	18.5
PD00394	PD00394b	Pre-LN	Male	142	82	21.1	4.2	1.3	2.5	0.9	2	90.8	12.5	8.7	4.4	0.4	453	12.9	5.9	80.1	80.1	8.6
PD00399	PD00399a	Pre-LN	Male	130	86	25.2	6.4	1.5	4.3	1.3	NA	NA	NA	NA	NA	NA	NA	NA	NA	56.7	56.7	6
PD00399	PD00399b	Pre-LN	Male	136	90	26.2	6.2	1	4.7	1.2	1.2	89.5	13.5	3.4	4.5	0.4	241	13.1	6	56.7	61.5	6
PD00403	PD00403b	Control	Male	140	73	24.4	5.2	0.9	3.2	2.5	1.6	88.8	14.4	7.5	5.7	0.5	296	16.5	5.8	69.1	69.1	12.6
PD00410	PD00410b	Control	Female	142	86	24.4	8.6	1.2	6.4	2.2	2	90.1	14.5	6.2	4.5	0.4	352	13.6	5.2	61.2	61.2	18.7
PD00414	PD00414b	Control	Female	100	58	19.8	6.9	2.7	4	0.6	2	87.5	12.3	5.5	4.3	0.4	197	13.3	5.3	63.5	63.5	18.8
PD00415	PD00415a	Control	Female	172	108	27.7	7.6	1.4	5.1	2.4	2.1	83.4	13.1	6.6	4.4	0.4	158	12.4	5.7	58	58	21.6
PD00417	PD00417a	Pre-LN	Male	134	84	20.7	6.1	1.4	4.2	1.1	0.9	84.5	13.8	3.5	4.5	0.4	138	12.5	5.1	65.3	65.3	12.3
PD00417	PD00417b	Pre-LN	Male	143	84	21.8	7	1.3	4.3	3.1	1.3	85.1	14.1	4.9	5.3	0.5	207	15.5	5.3	65.3	67.7	12.3
PD00421	PD00421c	Control	Female	100	58	24.6	3.5	1	1.9	1.4	1.3	88.2	13.5	4.8	4.2	0.4	194	12.8	7	59.4	59.4	4.7
PD00425	PD00425a	Pre-LN	Female	149	75	23.4	8.7	1.4	6.2	2.6	2.5	89.4	12.4	6.2	4.3	0.4	160	13.2	5	72.5	72.5	17.7
PD00426	PD00426b	Control	Female	150	80	25.7	5.5	1.9	2.8	1.9	2.3	87.8	14.2	8.7	4.4	0.4	295	13.5	6.1	70.2	70.2	19
PD00427	PD00427b	Pre-LN	Female	117	72	28.9	6.6	1.5	4.3	1.9	2.1	94.5	13.9	6.4	4.5	0.4	269	13.7	5.1	48	48	13.3
PD00429	PD00429a	Control	Male	170	92	28.4	7	1.4	4.8	1.8	2	66.7	14.4	5.9	6.3	0.4	293	13.1	5.2	64.1	64.1	20.5
PD00431	PD00431b	Control	Female	164	84	28.3	7.2	1.9	3.9	3.2	1.8	87.8	13.4	5.5	4.8	0.4	227	14.6	5.5	77.3	77.3	18.6
PD00448	PD00448a	Pre-LN	Female	134	76	23.3	5.3	1.5	3.4	0.8	1.4	90.4	13.4	6.4	4.5	0.4	299	14	NA	65.7	65.7	14.2
PD00448	PD00448b	Pre-LN	Female	127	67	23.9	4.9	1.1	2.9	2	1.4	94.1	13.6	5.3	4.4	0.4	232	14.4	6.2	65.7	69.8	14.2
PD00449	PD00449a	Pre-LN	Female	122	86	30.4	4.6	1	2.7	2	1.7	90.6	12.9	6.3	4.7	0.4	162	14.2	5.9	45.9	45.9	15.9
PD00449	PD00449b	Pre-LN	Female	124	82	30.9	4.7	1	3	1.7	1.9	88.8	12.9	6.8	4.8	0.4	220	13.6	5.4	45.9	48.3	15.9
PD00451	PD00451a	Control	Female	120	78	24.2	5.3	2.3	2.6	0.8	1.8	90.2	12	6.8	3.9	0.4	169	12.5	NA	51.5	51.5	22.5
PD00452	PD00452b	Control	Male	114	86	25.9	6	1	3.6	3.2	1.8	95.8	12.1	5.3	4.4	0.4	165	14.4	5.6	68.8	68.8	17.9
PD00454	PD00454b	Control	Female	126	60	25.3	8.5	0.9	5	5.8	2	92.7	12.8	5.5	3.7	0.3	200	12	5.4	75.8	75.8	18.5
PD00455	PD00455a	Pre-LN	Female	168	94	24.7	7.9	1.9	5.5	1.1	2.6	98.3	14.3	6.9	3.9	0.4	344	13.3	NA	68.3	68.3	6.9
PD00455	PD00455b	Pre-LN	Female	179	102	25.4	6.3	1.7	4.2	0.9	2	99.3	15.3	6.6	3.6	0.4	327	12.8	5.9	68.3	71.9	6.9
PD00462	PD00462a	Control	Male	145	88	34.2	5	1.5	2.9	1.4	1.5	91.6	13.3	5.8	4.8	0.4	258	15.7	NA	70.2	70.2	20.7
PD00462	PD00462c	Control	Male	106	72	31.7	3.8	1.3	1.8	1.6	1.1	99	15.9	5.1	4.3	0.4	267	13.9	8.3	70.2	82	20.7
PD00464	PD00464b	Pre-LN	Female	159	98	23.8	4.7	1.6	2.8	0.8	2.1	97.3	13.3	6	4.4	0.4	198	13.1	5.4	64.3	64.3	5.4
PD00465	PD00465b	Control	Female	155	96	28.3	7.6	2	5.2	0.9	1.5	89.1	12.9	4.8	5	0.4	330	15.8	5.1	70	70	19.1
PD00478	PD00478a	Control	Female	138	76	28.5	5.7	1.4	3.6	1.6	3.5	88.3	13.6	9.4	4.7	0.4	209	13.9	5.3	75	75	20.9
PD00482	PD00482a	Control	Female	132	82	31.5	6.4	1.6	3.5	2.8	2.4	90	13.3	6	4.3	0.4	209	12.9	5.7	61.6	61.6	21.8
PD00483	PD00483a	Pre-LN	Female	134	90	27.6	6.2	1.6	3.7	1.9	NA	NA	NA	NA	NA	NA	NA	NA	NA	59.7	59.7	16
PD00484	PD00484b	Control	Male	138	84	25.4	4.3	1.5	2.2	1.4	1.5	94.1	12.6	5	4.3	0.4	152	14.2	5.5	68.9	68.9	18.6
PD00485	PD00485b	Pre-LN	Male	144	88	25.4	5.2	0.8	3.6	1.9	1.8	87.2	13.5	5.5	4.7	0.4	227	14.5	5.1	69.9	69.9	14.2
PD00491	PD00491b	Control	Male	123	72	27.5	4.6	1.2	2.5	2	2.6	87.7	14	5.2	5.4	0.5	320	15.9	4.7	68.6	68.6	18.5
PD00494	PD00494a	Control	Female	178	102	31.4	7.1	1.6	4.5	2.1	1.8	85.6	13.7	5.8	4.8	0.4	301	14.1	NA	67.5	67.5	23.2
PD00494	PD00494c	Control	Female	136	60	33	4.7	1.6	2.4	1.6	1.8	88	NA	7.2	4.6	0.4	303	13.3	6.8	67.5	79.8	23.2
PD00496	PD00496b	Control	Male	128	85	25.4	8.1	1.4	5.8	2	1.5	90.1	12.3	5.8	4.7	0.4	392	14.7	5.4	52.3	52.3	18.6
PD00497	PD00497b	Pre-LN	Female	138	79	24.8	6.8	1.3	4.5	2.3	3.8	90.4	13.2	8.6	4.4	0.4	341	12.5	5.8	58.3	58.3	3.2
PD00506	PD00506a	Pre-LN	Female	146	88	25.8	6.4	2.1	4	0.8	2.1	93.4	13.1	6.6	4.1	0.4	304	13.6	4.7	67.5	67.5	14.8
PD00506	PD00506b	Pre-LN	Female	150	88	24.1	4.9	1.5	2.8	1.4	NA	NA	NA	NA	NA	NA	NA	NA	NA	67.5	70.2	14.8
PD00510	PD00510a	Pre-LN	Male	131	78	26.9	6.3	1.4	4.1	1.7	1.8	92.2	12.4	6	4.6	0.4	218	14.8	NA	53.7	53.7	13.9
PD00510	PD00510b	Pre-LN	Male	130	77	25.8	6.3	1.7	3.9	1.6	1.6	91.6	12.5	4.9	4.8	0.4	188	15.3	5.4	53.7	57	13.9
PD00510	PD00510c	Pre-LN	Male	154	76	24.2	5.4	1.2	3.3	2	1.1	93	14	4.8	4.4	0.4	185	13.9	5.3	53.7	65.5	13.9
PD00512	PD00512b	Pre-LN	Female	132	80	22.2	5.1	1.8	2.8	1.2	2.3	85.7	13.3	6.9	4.3	0.4	208	12.6	5.6	69.4	69.4	14.8

PD00512	PD00512c	Pre-LN	Female	132	76	20.3	5.7	1.8	3.4	1.3	2.3	91	NA	9.2	4.2	0.4	264	12.4	5.6	69.4	77.8	14.8
PD00514	PD00514b	Control	Male	122	79	27.8	5.5	1.1	3.6	1.8	1.2	91.7	13.8	6.3	5.2	0.5	174	17.2	5.4	62.6	62.6	18.4
PD00515	PD00515a	Control	Male	118	72	26.5	4.3	0.8	2.8	1.5	1.6	86.6	13.2	8	4.9	0.4	304	14.8	NA	61.7	61.7	23.4
PD00515	PD00515c	Control	Male	133	76	27.1	4	1.1	2.4	1.1	1.7	89.3	14.6	8.9	4.9	0.4	291	14.3	5.7	61.7	76.1	23.4
PD00516	PD00516a	Pre-LN	Female	110	66	23	4.6	2	2.3	0.6	1.4	94.5	12.2	5.1	3.8	0.4	224	12.3	NA	49.6	49.6	10.4
PD00516	PD00516b	Pre-LN	Female	90	56	22.4	4	1.8	1.9	0.7	2.1	93.3	11.2	6.6	3.7	0.3	206	12.6	4.8	49.6	53.2	10.4
PD00516	PD00516c	Pre-LN	Female	108	64	21.3	4.1	1.9	2	0.5	0.9	99.9	13.3	3.3	3.7	0.4	176	12.4	5	49.6	63.5	10.4
PD00517	PD00517a	Pre-LN	Female	116	80	24.8	4.5	1.1	2.4	2.3	1.8	86.4	12.5	4.9	3.7	0.3	237	11.4	4.7	54.8	54.8	4.1
PD00517	PD00517b	Pre-LN	Female	128	76	25.5	5.2	1.3	3.3	1.5	2	89	13.1	5.5	4.1	0.4	328	12.4	5.3	54.8	57.8	4.1
PD00517	PD00517c	Pre-LN	Female	136	82	25.9	5.2	1.2	3.4	1.5	1.8	91.7	13.5	5.3	3.8	0.3	252	12.1	5.4	54.8	67.4	4.1
PD00518	PD00518a	Control	Female	128	68	20.8	5.5	1.4	3.3	1.7	1.7	84	13.2	5.6	4.4	0.4	277	12.4	NA	56.9	56.9	22.9
PD00519	PD00519a	Control	Male	121	80	25.8	6	1.2	3.9	2.1	2.6	93.2	13.3	8.3	4.7	0.4	190	14.8	4.1	60.2	60.2	15.6
PD00519	PD00519c	Control	Male	123	80	25.8	5.7	1.7	3.7	0.7	1.5	98	NA	5.8	4.6	0.4	NA	15.6	6.1	60.2	69.5	15.6
PD00521	PD00521a	Pre-LN	Female	120	81	25.9	4.5	1.1	2.7	1.5	NA	NA	NA	NA	NA	NA	NA	NA	NA	59.8	59.8	16.6
PD00525	PD00525a	Control	Male	142	80	23.6	6	2.6	3	0.8	1.7	93.1	13.1	8.1	4.4	0.4	349	14.2	NA	64.8	64.8	22.1
PD00528	PD00528a	Pre-LN	Female	140	79	24.5	6.1	2	3.7	0.9	2.6	91.8	12.8	8.3	4.2	0.4	369	12.4	4.6	62	62	11
PD00528	PD00528b	Pre-LN	Female	149	92	25.6	5.6	1.9	3.3	1	2.8	88.2	12.5	12.4	4.4	0.4	404	13.6	5.4	62	64.3	11
PD00530	PD00530b	Pre-LN	Male	130	86	27.4	4.1	1	2.6	1.2	2.4	84.7	13.6	5.6	5.4	0.5	180	15.7	5.6	57.9	57.9	15.9
PD00532	PD00532a	Pre-LN	Female	174	110	27.9	7.3	1.6	4.7	2.4	2.2	92.5	13.4	7.9	4.9	0.4	315	14	5.8	72	72	9.1
PD00539	PD00539a	Pre-LN	Female	124	74	22.5	6.8	1.8	4.3	1.6	1.9	88	13.7	7.8	4.2	0.4	315	13.1	5.5	51.2	51.2	16.4
PD00543	PD00543b	Control	Female	112	68	21.3	7.5	1.3	5.7	1.3	0.8	91.2	14.3	3.7	4	0.4	243	11.9	5.3	52.5	52.5	17.6
PD00543	PD00543c	Control	Female	120	71	22.3	4.3	1.4	2.6	0.8	1.2	91.8	15	7	3.8	0.3	221	11.8	5.8	52.5	61.3	17.6
PD00551	PD00551b	Pre-LN	Female	128	76	25.4	6.7	1.5	4.3	2	2	91.8	14.8	6.2	4	0.4	230	12.8	5.3	61.6	61.6	11.4
PD00551	PD00551c	Pre-LN	Female	122	72	24.1	4.8	1.4	2.8	1.4	2.2	96	NA	6.1	3.9	0.4	255	12.4	5.5	61.6	69.9	11.4
PD00553	PD00553b	Control	Male	115	68	22.7	4.4	2.1	1.9	1	2	94	13.7	6	4.4	0.4	222	15	5.3	59.9	59.9	18.2
PD00559	PD00559a	Control	Male	104	68	22.5	5.2	2	2.7	1.1	1.4	91.9	13.9	5.4	5	0.5	272	15.4	5	52.6	52.6	16.9
PD00561	PD00561b	Pre-LN	Female	134	90	20.4	6.4	2.7	3.2	1.1	1.8	90.4	14.7	6.1	3.9	0.3	245	12.1	4.8	51.4	51.4	5
PD00561	PD00561c	Pre-LN	Female	140	98	18.7	7.7	1.5	5.7	1.3	1.1	95.7	15.6	4.6	3.8	0.4	277	12.7	5.3	51.4	59.5	5
PD00565	PD00565b	Pre-LN	Female	128	74	27.9	7.1	1.4	4.6	2.6	1.4	85.5	14.2	5.2	4.4	0.4	343	12.9	5.7	73.3	73.3	11
PD00569	PD00569a	Pre-LN	Male	141	88	29.7	5.2	1	3.7	1.1	2.3	89.8	13.2	6.4	5.1	0.5	240	15.8	4.8	64.3	64.3	11.8
PD00569	PD00569b	Pre-LN	Male	148	92	30.4	4.7	0.9	3.2	1.5	2.5	91.8	13.6	7	5.1	0.5	219	16	5	64.3	67.5	11.8
PD00571	PD00571a	Control	Female	134	80	32.8	5.1	1.4	3.1	1.4	NA	NA	NA	NA	NA	NA	NA	NA	NA	57.8	57.8	19.6
PD00571	PD00571c	Control	Female	120	70	34.4	3.9	1	2.6	0.7	1.2	109.3	17	3.7	3.4	0.4	112	12.7	4.9	57.8	72.8	19.6
PD00576	PD00576b	Control	Male	144	82	23.1	5.2	0.8	3.8	1.5	1.8	92.9	14.3	7.9	3.7	0.3	715	11.4	5.5	74.3	74.3	13.6
PD00576	PD00576c	Control	Male	114	70	24.1	4.7	1.1	2.8	1.8	1.9	92	13.6	7.2	3.7	0.3	232	11.5	5.8	74.3	82.5	13.6
PD00578	PD00578a	Pre-LN	Female	114	72	31.2	5.1	1	3.4	1.5	NA	NA	NA	NA	NA	NA	NA	NA	NA	58.6	58.6	11.8
PD00578	PD00578b	Pre-LN	Female	110	62	30.3	5.7	0.9	4.1	1.7	2.3	92	12.8	6.1	4.5	0.4	205	13.6	5.7	58.6	62.5	11.8
PD00581	PD00581a	Control	Male	118	70	20.7	5.4	2.2	2.7	1	1.4	85.7	13.2	5.9	4.7	0.4	220	14.3	NA	48.1	48.1	22.5
PD00581	PD00581c	Control	Male	144	86	20.1	5.5	2.7	2.5	0.7	NA	90	14.2	4.7	4.7	0.4	228	14.2	5.6	48.1	61.5	22.5
PD00584	PD00584b	Control	Male	102	64	21	5.3	1.4	3.7	0.6	1.2	93.1	14.3	4.3	4.7	0.4	159	14.2	4.8	59.6	59.6	17.6
PD00585	PD00585b	Pre-LN	Female	148	90	25.4	8.9	1.8	6.6	1.2	1.2	84.5	13.1	3.6	4.4	0.4	136	13	5.4	75	75	10.6
PD00588	PD00588b	Pre-LN	Female	120	83	25.4	5.6	1.5	3.6	1.3	1.8	93.2	13.4	5.7	3.8	0.4	204	13	5.6	58.9	58.9	6.6
PD00590	PD00590b	Control	Female	122	72	24.3	7.1	1.4	4.7	2.2	2.2	90.8	12.9	5.9	3.9	0.4	205	12.5	5.2	58.9	58.9	18.4
PD00591	PD00591b	Control	Male	106	62	22.8	5.2	1.5	3.3	0.9	1.5	92.6	13.4	9.7	4.8	0.4	272	15.3	6	74.5	74.5	18
PD00591	PD00591c	Control	Male	140	71	23.1	3.4	1.5	1.5	0.9	1.2	94.5	13.9	6.5	4.8	0.5	187	15.3	7	74.5	82.5	18
PD00604	PD00604a	Pre-LN	Female	119	70	23.9	6.4	1.4	4.5	1.2	2.1	96.7	13	4.8	4.4	0.4	191	13.7	5.2	73.7	73.7	4.6
PD00604	PD00604b	Pre-LN	Female	126	73	24	6.1	1.5	3.8	1.9	2.2	95.7	13.7	5.9	4.3	0.4	304	13.7	5.8	73.7	76.2	4.6
PD00605	PD00605a	Control	Male	137	88	30.1	5.4	1.1	3.1	2.5	2.2	90.3	12.6	5.8	4.6	0.4	132	14.6	NA	57.3	57.3	23.2
PD00605	PD00605c	Control	Male	132	72	32.9	3.7	0.8	NA	5	2	95	NA	5.9	4.4	0.4	131	15.1	5.6	57.3	69.7	23.2
PD00606	PD00606a	Pre-LN	Male	132	82	26.3	8.9	1.2	5.6	4.8	2.5	95.2	12	7.4	4.3	0.4	149	13.9	5.2	54.3	54.3	8.2
PD00606	PD00606b	Pre-LN	Male	118	82	26.6	8	1.4	5.3	3	2.6	93.6	13.1	7.2	4.4	0.4	156	14.4	5.6	54.3	56.8	8.2
PD00606	PD00606c	Pre-LN	Male	106	64	28.7	4.1	2.1	1.6	1	1	95.1	14	9.1	4.2	0.4	116	13.4	NA	54.3	67.7	8.2
PD00607	PD00607b	Pre-LN	Male	132	94	24.7	7.4	1.4	5.3	1.7	1.5	91	13.3	5.4	4.9	0.4	196	15.9	5.7	55.9	55.9	2.9
PD00610	PD00610b	Control	Male	154	83	26.3	6.6	1.9	4.1	1.4	1.3	86.3	13.7	6.4	4.9	0.4	295	15.1	5.3	72.5	72.5	16
PD00611	PD00611b	Control	Male	139	90	29.3	6.7	1.2	4.7	1.8	NA	NA	NA	NA	NA	NA	NA	NA	5.4	51.8	51.8	12.3
PD00611	PD00611c	Control	Male	148	82	29.5	6	1.6	3.9	1.2	2.2	100.3	12.9	8.1	4	0.4	221	13.6	5.8	51.8	60.4	12.3
PD00613	PD00613b	Pre-LN	Female	120	74	22.1	5.6	1.6	3.5	1.1	3.5	91.9	14.4	6.9	4.1	0.4	210	12.5	5.5	54.6	54.6	15.1
PD00618	PD00618a	Control	Female	127	85	23.5	6.4	1.5	4.3	1.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	61.2	61.2	23.7
PD00618	PD00618c	Control	Female	153	88	24.1	4	1.4	2.2	1	NA	97.6	16	3.3	3.7	0.4	189	12.3	5.2	61.2	75.8	23.7
PD00623	PD00623a	Control	Female	176	108	30.9	6.2	2.1	3.4	1.6	2.2	84.6	14.4	6.1	4.9	0.4	198	13.9	5.3	67	67	21
PD00627	PD00627b	Pre-LN	Male	144	94	24.6	5.7	1.7	3.5	1.1	1.4	93.4	12.7	5	4.7	0.4	281	14	5.1	50.7	50.7	9
PD00627	PD00627c	Pre-LN	Male	146	90	23.4	6.1	2.1	3.4	1.4	1.2	97.6	15.8	4.7	4.5	0.4	212	14	5.6	50.7	59.4	9
PD00628	PD00628b	Control	Male	156	90	35.9	6.5	1.3	4.3	2.1	2.5	87.2	12.7	8.7	4.7	0.4	269	14.8	5.4	60.5	60.5	18.8
PD00628	PD00628c	Control	Male	161	88	39.4	5.9	1.2	3.5	2.7	2.8	89.8	14.1	10.6	5.1	0.5	308	14.8	5.9	60.5	68.8	18.8
PD00631	PD00631b	Control	Female	144	81	32.5	5.9	1.1	3.9	2.1	2.2	89.5	13.1	7	5	0.4	201	14.6	6	65.3	65.3	19.7
PD00632	PD00632a	Pre-LN	Female	121	79	24	6.7	2.4	4	0.7	1.9	91.9	13	6.2	4.2	0.4	280	12.5	5.3	56.6	56.6	18.9
PD00632	PD00632b	Pre-LN	Female	119	71	22.6	5.7	1.9	3.4	0.9	1.7	96.2	13.9	4.2	4	0.4	237	12.9	5	56.6	60.3	18.9
PD00632	PD00632c	Pre-LN	Female	122	73	22.9	6.6	2.5	3.8	0.8	1.8	96.6	14	6.6	4	0.4	240	12.8	5.4	56.6	70.7	18.9
PD00636	PD00636b	Control	Male	146	76	26.6	5.1	1	3.7	0.9	2.3	90.3	13.2	6.2	4.8	0.4	119	14.7	6.1	72.9	72.9	20.1
PD00638	PD00638a	Control	Female	NA	NA	26.4	5.4	2	2.8	1.4	2.4	90.5	12.3	6.8	4.2	0.4	248	13.5	NA	71.1	71.1	22.6
PD00639	PD00639b	Control	Male	136	74	23.5	6.2	1.3	3.1	4.1	2.3	91.8	13.2	6.7	4.5	0.4	166	13.3	5.4	64.7	64.7	19.5
PD00640	PD00640a	Control	Female	146	88	21.2	6.9	1.3	5.1	1.1	1.9	90.4	12.7	6	5.4	0.5	250	16.7	NA	68.5	68.5	22.5
PD00640	PD00640c	Control	Female	110	70	22.7	3.7	1.4	1.8	1.2	2	92	NA	6.4	5	0.5	262	15.5	6	68.5	81	22.5
PD00642	PD00642a	Pre-LN	Female	116	73	24.9	5.7	0.8	4.4	1.2	1.8	85.8	15	8.4	4.3	0.4	297	12	4.7	48.5	48.5	13.5
PD00642	PD00642b	Pre-LN	Female	118	73	27.1	6.3	1	4.9	0.9	2	89.9	15.3	5.6	4.6	0.4	310	13.1	5.4	48.5	50.9	13.5
PD00644	PD00644a	Pre-LN	Male	130	82	29.1	6.8	1.1	5	1.5	2.5	91.3	12.3	7.1	4.7	0.4	248	15.8	NA	60	60	7
PD00644	PD00644b	Pre-LN	Male	122	70	28.8	7.2	1.4	5.5	0.8	3	94.2	12.8	7.6	4.6	0.4	211	14.1	5.3	60	63.7	7
PD00645	PD00645b	Control	Male	140	96	27.1	6.3	1.3	4.5	1.3	3.4	96.1	15.3	8.4	4.6	0.4	268	15.4	5.4	67.7	67.7	17.9
PD00647	PD00647b	Control	Male	154	87	29.1	7.4	1	4.3	4.8	2	92	13	5.8	4.9	0.5	197	16.4	8.5	69.3	69.3	18.9
PD00651	PD00651b	Control	Female	152	84	22.8	6.5	1.7	4.2	1.4	0.9	88.1	13.1	5	4.8	0.4	220	13.9	5.4	71.6	71.6	20.1
PD00654	PD00654b	Control	Female	128	80	21.2	5.8	1.9	3.7	0.5	1.6	85	13.9	4.9	4.4	0.4	245	13.2	5.8	56.9	56.9	19.2
PD00654	PD00654c	Control	Female	152	70	23.1	6	1.6	3.8	1.4	2.3	89.6	14.1	8.1	4.5	0.4	274	13.6	5.8	56.9	66.5	19.2
PD00657	PD00657b	Control	Female	114	67	24.9	5.5	1.4	3.4	1.7	1.3	85.4	13.7	4.7	4.3	0.4	257	12.9	4.9	64.5	64.5	18.5
PD00662	PD00662b	Control	Male	145	88	25.4	5.5	1.2	3.3	2.3	1.7	91.9	13.5	5.4	4.8	0.4	321	15.4	5.7	71.8	71.8	18.3
PD00662	PD00662c	Control	Male	136	82	25.4	5.4	1.4	3	2.4	1.6	93.8	15.2	6.6	5	0.5	321	16.1	5.9	71.8	79.9	18.3
PD00666	PD00666a	Pre-LN	Male	134	80	21	6.3	1.6	4	1.6	NA	NA	NA	NA	NA	NA	NA	NA	NA	65.2	65.2	18.3
PD00666	PD00666b	Pre-LN	Male	147	76	21.6	5.5	1.3	3.7	1.1	1.4	91.4	12.3	8.6	4.1	0.4	211	12.9	5.2	65.2	69.8	18.3
PD00666	PD00666c	Pre-LN	Male	142	70	20.1	5.1	1.5	3.2	0.9	17.5	90.8	15.8	23.3	3.8	0.3	112	11.2	6.2	65.2	83	18.3
PD00668	PD00668a	Control	Male	130	82	24.1	6.2	1.7	3.8	1.7	2.8	90.7	14.7	6.2	4.3	0.4	219	13.4	5.8	67.6	67.6	9.8
PD00672	PD00672b	Control	Male	156	85	26.1	6.4	1.5	4.3	1.5	1	95.3	12.8	4.1	4.3	0.4	183	14.5	5.3	71.5	71.5	18.9
PD00676	PD00676a	Control	Female	128	62	23.9	6.2	0.8	4.6	1.8	2.2	87	12.9	5.8	3.8	0.3	245	11.5	NA	71.5	71.5	19.3
PD00676	PD00676c	Control	Female	196	86	20.3	5	1.1	3.5	1	2.3	90.6	15.7	6.6	3.9	0.4	283	11.6	6.1	71.5	84.1	19.3
PD00677	PD00677a	Control	Female	166	110	24.1	7.8	1.7	5.3	1.8	2.5	90.1	13.8	8	4.8	0.4	351	15.1	NA	65.4	65.4	23.1
PD00677	PD00677c	Control	Female	148	72	25.3	4.6	2.1	2	1.1	1.8	91	15.6	7.5	4.9	0.4	253	14.8	6	65.4	78.9	23.1
PD00678	PD00678a	Control	Female	110	70	26.7	10.2	1.7	7.4	2.5	1.7	91.4	13	6	4.5	0.4	359	13.4	5.4	58.3	58.3	20.8
PD00682	PD00682b	Control	Male	126	76	29.9	5.6	1	3	3.7	1.8	84	14.9	6	5.5	0.5	234	15.8	5.5	61.4	61.4	18.5
PD00683	PD00683a	Control	Male	128	72	24.8	4	0.8	2.7	1	2.2	87.9	18.7	6.3	3.9	0.3	341	11.5	NA	63.8	63.8	23.2
PD00684	PD00684b	Pre-LN	Female	155	92	40.1	4.7	1.4	2.3	2.4	2.4	90.8	13.5	6.5	4.2	0.4	147	13.4	5.6	76.8	76.8	3.2
PD00687	PD00687b	Control	Male	132	70	30.1	5.7	1.5	3.7	1.2	1.7	90.2	14.3	5.4	4.6	0.4	222	14.4	4.6	72.5	72.5	18
PD00688	PD00688a	Control	Female	118	72	19.2	5.4	2.5	2.5	0.8	1.8	93.5	13	6.5	4.3	0.4	240	14.3	5	50.7	50.7	21.5
PD00691	PD00691a	Control	Female	120	76	24.8	7.2	2	4.8	0.9	2	85.6	14.2	6.3	4.8	0.4	306	13	NA	50.5	50.5	22.7
PD00693	PD00693a	Control	Female	146	90	NA	6.1	1.2	4	1.9	2.6	84.6	14.2	7.8	5.1	0.4	310	14.7	NA	59	59	18
PD00698	PD00698a	Control	Male	154	100	27.4	4.7	0.9	2.6	2.7	2.4	87.6	12.8	6.4	4.8	0.4	200	14.5	5.5	66.8	66.8	19.2
PD00698	PD00698c	Control	Male	144	70	29.6	3.9	1.1	2.3	1.3	2	93.3	15.2	5.9	4	0.4	147	12.5	5.7	66.8	81.2	19.2
PD00705	PD00705b	Pre-LN	Female	146	82	22.1	8.7	1.3	6.4	2.2	1.4	92.8	13.7	3.5	4.1	0.4	245	13.3	5.7	78.6	78.6	9.9
PD00706	PD00706a	Pre-LN	Female	128	80	19.9	6.9	1.6	4.8	1.2	1.7	92	13.4	7.9	4.4	0.4	347	13.5	5.4	47.8	47.8	5
PD00706	PD00706b	Pre-LN	Female	140	82	18.5	7.2	1.6	5.1	1.3	1.8	91.9	12.6	11.7	4.4	0.4	341	14.4	7	47.8	50.7	5
PD00711	PD00711b	Pre-LN	Male	129	81	22.8	6.5	1.4	4.5	1.4	2.4	85.7	13.9	8	4.9	0.4	177	14.7	5.5	75.6	75.6	13.9
PD00713	PD00713b	Pre-LN	Male	133	87	24.1	4.3	1.7	2.3	0.7	2.8	87.2	13.1	7.4	4.6	0.4	147	13.6	5.8	65.4	65.4	4.8
PD00714	PD00714b	Pre-LN	Male	123	75	29.8	6.5	0.9	4.9	1.7	3.1	89.2	13.6	10.3	5.1	0.5	253	15.1	7.4	53.7	53.7	6.3
PD00715	PD00715c	Pre-LN	Female	134	81	29.2	6.2	1.1	3.3	4	2.4	90	NA	8.3	4	0.4	282	12.6	6.5	77.3	77.3	4.6
PD00719	PD00719a	Pre-LN	Female	120	70	24.2	5.9	1	4.5	0.9	NA	NA	NA	NA	NA	NA	NA	NA	NA	58	58	19.6
PD00719	PD00719c	Pre-LN	Female	134	79	25.8	6.7	1.4	4.7	1.5	1.9	91.7	14.7	5.1	4.7	0.4	211	14.1	5.5	58	72.5	19.6
PD00720	PD00720a	Control	Male	168	99	33.7	5.8	1.9	3.2	1.7	2	95.4	14.1	4.8	4.7	0.4	235	14.6	5.1	72.7	72.7	20.7
PD00723	PD00723a	Pre-LN	Female	106	66	22.2	4.2	2.4	1.4	0.9	1.8	92	13.1	10.4	3.8	0.3	341	11.8	5.3	50.2	50.2	17.8
PD00723	PD00723b	Pre-LN	Female	96	58	22.1	4.6	2.4	1.8	1	1.8	94.8	13.6	7.1	3.8	0.4	297	12.1	5	50.2	54.2	17.8
PD00724	PD00724a	Control	Male	126	78	29.4	4.9	1.4	2.7	1.8	1.5	87	12.9	7.4	5	0.4	300	15.1	4.7	53	53	22.1
PD00727	PD00727a	Control	Female	157	79	25.4	6.8	1.2	4.6	2.1	1.3	82.7	13.9	5.9	4.4	0.4	315	12.6	NA	66.9	66.9	23.2
PD00727	PD00727c	Control	Female	144	74	27.2	4.5	1.9	2.1	1.3	1.1	87	NA	5.3	4.4	0.4	280	12.9	6.5	66.9	79.4	23.2
PD00728	PD00728a	Pre-LN	Male	145	90	26.5	6.1	1.1	3.7	2.9	2.3	93.2	11.7	5.5	4.4	0.4	302	14.2	5.5	57	57	10.4
PD00728	PD00728b	Pre-LN	Male	137	88	28	4.4	1.1	2.5	1.9	1.9	91.7	11.7	5.1	4.4	0.4	283	14.8	5.2	57	61	10.4
PD00730	PD00730a	Pre-LN	Male	114	78	27.5	7.8	0.9	5	4.2	1.9	92.1	13.8	5.3	5	0.5	202	15.4	5.6	68.6	68.6	12.6
PD00730	PD00730b	Pre-LN	Male	112	74	27.8	6.8	1.1	3.4	5.1	1.5	93.2	12.9	4.6	5.2	0.5	259	15.2	5.6	68.6	71.6	12.6
PD00731	PD00731a	Pre-LN	Male	122	82	30.2	5.9	1.6	3.8	1	1.9	88.1	13.2	5.8	5	0.4	241	15.7	NA	50	50	7
PD00731	PD00731b	Pre-LN	Male	123	72	28.2	4.7	1.8	2.7	0.5	1.7	90.1	12.7	5.7	5	0.5	222	15.7	4.9	50	53.6	7
PD00732	PD00732b	Control	Female	144	78	26.4	8.7	1.1	5.5	4.8	2.6	95.1	14.6	5.5	4.3	0.4	331	14.1	5.4	73.5	73.5	13.9
PD00734	PD00734b	Control	Male	131	86	29.6	9.6	1.1	6.4	4.7	2.8	84.1	14.4	9.7	5.7	0.5	317	16.5	4.8	56.8	56.8	18.2
PD00734	PD00734c	Control	Male	130	84	27.3	3.7	1.2	1.7	1.8	3	91.4	14.6	7.2	4.9	0.4	359	14.6	5.2	56.8	64.6	18.2
PD00737	PD00737b	Control	Male	170	99	28.5	6.7	1.2	4.5	2.3	1.8	94	13.2	6.3	4.6	0.4	255	15.7	5.1	71.6	71.6	19
PD00737	PD00737c	Control	Male	151	82	30.2	4.4	2	2	1	0.8	95.3	14	10.5	3.9	0.4	320	12.4	6	71.6	80	19
PD00738	PD00738b	Control	Female	142	82	23.7	8	2.2	5.2	1.4	2.6	88	12.7	8.4	5.4	0.5	355	14.9	5.3	52.6	52.6	18.1
PD00739	PD00739a	Pre-LN	Female	150	94	29.7	7	1.1	4.9	2.4	2.1	92.2	12.6	6.8	4.6	0.4	359	13.7	5.2	60.1	60.1	14.3
PD00740	PD00740b	Control	Female	146	88	30.2	5.2	1.2	3.5	1.3	2.1	87.4	14.6	6.4	4.8	0.4	434	14.5	5.5	65.7	65.7	18.3
PD00740	PD00740c	Control	Female	150	86	30.4	5.6	1.4	3.3	2.1	2.2	89.4	15.2	7.2	4.6	0.4	379	13.8	5.9	65.7	74.3	18.3
PD00744	PD00744b	Pre-LN	Male	135	84	26.8	4.9	1.4	3.2	0.8	2.1	86.6	14.2	4.7	4.2	0.4	334	12.8	5.3	61.1	61.1	10.7
PD00745	PD00745a	Control	Male	127	82	27	5.7	1.9	3.3	1	1.7	95.1	11.8	5.6	4.4	0.4	181	14.6	5.6	63.6	63.6	22.2
PD00746	PD00746b	Control	Female	154	89	23.3	8.8	1.8	6.7	0.8	1.6	91.2	13.5	4.3	4.6	0.4	264	14.1	5.2	55.8	55.8	18.5
PD00748	PD00748b	Control	Female	158	86	24	6.6	1.3	4.6	1.6	1.5	88.7	13.5	6	4	0.4	234	12.3	5.6	71.9	71.9	16.5
PD00748	PD00748c	Control	Female	137	74	25.5	4.2	1.8	2.1	0.8	1.3	89.2	14.3	5.7	4	0.4	284	11.8	6.1	71.9	82.4	16.5
PD00749	PD00749a	Control	Female	133	84	26.8	7.1	1	5.3	1.7	2.3	91.8	15	5.4	4.9	0.4	212	15	NA	70.3	70.3	23.7

PD00749	PD00749c	Control	Female	160	91	27.3	6.5	0.7	NA	4.6	1.1	82.4	15.4	4.7	4.3	0.4	241	11.3	6.7	70.3	84.7	23.7
PD00751	PD00751a	Control	Male	122	78	30.9	5.8	1.1	2.9	3.9	2	88.6	12.8	7.9	4.7	0.4	216	15.3	NA	58.6	58.6	22.5
PD00754	PD00754b	Control	Female	182	106	28.7	6.4	2	3.8	1.5	1.3	87	14	6.6	4.3	0.4	180	13.3	5.2	79.5	79.5	16.1
PD00756	PD00756a	Control	Male	128	74	30.5	6	1	4.3	1.5	2.2	85.2	14.9	6.3	5.4	0.5	145	15.6	5.7	67.5	67.5	21.6
PD00756	PD00756c	Control	Male	126	62	34.9	3.9	1	2.4	1.1	1.5	91.7	16	7.6	4.9	0.4	118	14.6	6.6	67.5	80.5	21.6
PD00761	PD00761a	Control	Female	146	86	23.9	6.7	2	4.3	0.9	1.6	90.1	13	5.8	4.1	0.4	269	13.2	NA	49.6	49.6	22.5
PD00763	PD00763b	Control	Female	116	72	30	7.2	1.5	4.9	1.9	2.4	95	15.3	7.3	4.6	0.4	243	14.1	5.1	50	50	17.8
PD00764	PD00764b	Pre-LN	Female	137	84	25.4	7.7	1.7	5.2	1.8	2	87.1	13	5.7	4.1	0.4	284	12.7	5.4	56.7	56.7	9.8
PD00765	PD00765a	Control	Male	124	76	29.9	5.4	1	3.6	1.7	1.1	91	14	3.9	4.5	0.4	198	14.3	NA	75	75	23
PD00772	PD00772b	Control	Female	134	80	24.9	6.1	1.4	3.3	3.1	1.8	91.5	14	7.1	4.3	0.4	261	14.1	5.1	77.4	77.4	18.9
PD00772	PD00772c	Control	Female	129	68	22.9	3.6	1	2	1.4	1.9	90.6	14.9	14.9	3.5	0.3	541	10.5	6.1	77.4	85.9	18.9
PD00773	PD00773a	Pre-LN	Female	100	60	21.8	5.7	2.2	3	1.1	1.7	95.2	12.2	5.6	4.3	0.4	190	13.7	NA	60.7	60.7	10.4
PD00774	PD00774b	Control	Male	129	72	27.4	3.6	1.4	1.9	0.8	2.2	88.1	12.7	6.3	5.2	0.5	229	15.9	5.3	58.4	58.4	19.4
PD00775	PD00775a	Control	Male	153	88	25.2	4.7	1.3	3	0.8	1.8	93.7	12.9	6.5	4.8	0.5	205	15.6	NA	47.8	47.8	22.9
PD00776	PD00776b	Control	Female	156	92	27.1	7.4	2	4.5	2	2.4	94.4	13.2	5.6	3.8	0.4	285	12.5	6	68.5	68.5	18.4
PD00776	PD00776c	Control	Female	148	84	22	8.1	1.8	5.5	2.1	NA	100.6	13.8	5.2	3.9	0.4	311	12.8	6.2	68.5	77.8	18.4
PD00780	PD00780b	Control	Male	134	92	28.4	6.7	1.1	4.3	3	2.5	93	12.4	6.4	4.4	0.4	284	13.2	6.1	68.9	68.9	19.8
PD00780	PD00780c	Control	Male	141	93	29	8.2	1.2	5	4.5	2.4	90.6	14.8	6.9	4.4	0.4	317	13.4	6.3	68.9	79.3	19.8
PD00781	PD00781b	Control	Male	134	88	27.3	6.6	1.3	4.5	1.9	2.2	91.2	13.7	6.3	4.8	0.4	349	14	5.4	61.8	61.8	17.5
PD00783	PD00783a	Control	Female	158	95	31.1	8.3	1.9	5.4	2.2	2.3	78.3	14.2	7.1	4.9	0.4	231	12.5	5.7	53.3	53.3	20.8
PD00783	PD00783c	Control	Female	156	83	31.4	7.8	1.8	5.3	1.6	2	87.8	13.2	6.5	4.4	0.4	240	13.3	5.6	53.3	64.2	20.8
PD00786	PD00786b	Control	Male	118	72	25.4	5.6	1.2	3.9	1.3	2.3	88.5	14	6.6	5.3	0.5	236	16	4.8	60.5	60.5	19.3
PD00787	PD00787b	Pre-LN	Female	118	77	28.1	6.5	1.2	4.6	1.7	2	89.2	14.1	5.4	4.7	0.4	309	13.8	5.1	63.4	63.4	4.2
PD00790	PD00790a	Control	Male	128	88	23.8	5.8	1.2	3.7	2.1	2	89.3	13.5	6.4	4.9	0.4	262	14.3	5.5	50.6	50.6	20.9
PD00791	PD00791a	Control	Male	128	83	28.2	4.1	0.9	2.6	1.3	2.2	90.1	12.8	7.1	4.5	0.4	219	14.1	NA	59	59	23.6
PD00791	PD00791c	Control	Male	140	78	30	4.1	1	2.6	1.3	1.6	94.1	14	7.4	4.5	0.4	202	14.2	6.7	59	73.3	23.6
PD00792	PD00792a	Control	Male	122	72	26.1	5.1	1	3.1	2.2	2.6	88.7	13.3	7.9	5	0.4	268	15	5.3	66.7	66.7	20.8
PD00793	PD00793b	Pre-LN	Female	116	78	27.2	6.9	1.7	3.9	2.9	2.6	87.1	12.9	6.8	3.7	0.3	304	11.1	5.1	66	66	6.6
PD00793	PD00793c	Pre-LN	Female	111	72	25.5	7	1.2	4.8	2.2	1.7	88.5	14.3	5.8	4.3	0.4	228	12.9	5.7	66	76.1	6.6
PD00794	PD00794a	Control	Male	110	67	26.4	6.2	1.6	4	1.4	1.9	89.7	12.3	7	4.9	0.4	268	14.8	5.1	48.2	48.2	21.8
PD00794	PD00794c	Control	Male	116	71	27.6	6.6	1.8	4.3	1.2	1.1	92.3	12.8	5.3	4.9	0.5	217	15.1	5.7	48.2	61.9	21.8
PD00795	PD00795b	Pre-LN	Male	128	76	23.8	6.4	0.8	4.2	3.2	1.2	92.5	12.5	5.1	4.5	0.4	195	14.2	5.4	68.2	68.2	1.7
PD00795	PD00795c	Pre-LN	Male	135	81	24.2	3.2	0.9	1.5	1.9	1.6	97	NA	5.6	3.6	0.3	300	11.8	5.8	68.2	76.5	1.7
PD00799	PD00799a	Control	Male	120	76	26.3	6.3	1.4	4.1	1.9	2.1	93.3	13.2	7	4.8	0.5	280	14.8	7.6	61.7	61.7	19.8
PD00799	PD00799c	Control	Male	112	74	24.2	4.9	1.4	3.1	1	1.5	92	NA	9.1	4.1	0.4	241	12.7	7.7	61.7	71.9	19.8
PD00800	PD00800a	Control	Male	140	81	23.2	5.1	1.6	2.8	1.5	2.2	89.3	12.8	6.3	4.9	0.4	232	15.2	5.3	65.6	65.6	21.8
PD00802	PD00802a	Control	Male	107	64	25.2	7.7	1.1	5.3	2.9	3.6	89.7	14.2	10	5.1	0.5	226	15.3	5	44.4	44.4	21.5
PD00804	PD00804b	Control	Female	124	72	23.9	5.3	2.1	2.7	1.2	1.9	93	13	5.5	4	0.4	154	13.6	5.3	58.1	58.1	19.1
PD00806	PD00806b	Control	Male	135	88	26.2	5.2	1.7	3	1.2	2	89	13.5	6.3	4.8	0.4	293	14.1	5.3	58.3	58.3	18.2
PD00806	PD00806c	Control	Male	149	88	27.5	4.3	1.6	2.2	1.3	2.2	90.3	13.9	7.8	4.6	0.4	273	14.5	5.4	58.3	66.4	18.2
PD00807	PD00807b	Control	Male	121	68	32.3	4.7	1	2.2	3.5	2	92.8	12.9	6.6	4.7	0.4	114	15.4	5.4	70.9	70.9	17.2
PD00812	PD00812b	Control	Female	122	74	27.7	6.2	2.7	3.1	1.3	2.1	89.2	13.4	7.7	4.5	0.4	406	13.7	5.1	53.9	53.9	18.2
PD00813	PD00813b	Control	Female	152	90	24.7	6.1	1.5	3.2	3.2	2.3	87.7	12.6	8	4.6	0.4	178	14.2	5.4	76.8	76.8	19.4
PD00814	PD00814a	Control	Female	171	108	28.9	6.4	1.8	4	1.2	1.8	91	12.7	5.9	4.2	0.4	193	13.2	NA	72	72	19.9
PD00819	PD00819b	Pre-LN	Male	144	88	27.8	4.8	0.8	3	2.2	2.7	85.2	15.3	7.9	4.6	0.4	337	12.9	5.2	61.7	61.7	3.7
PD00820	PD00820a	Pre-LN	Male	134	84	27.2	4.4	1.3	2.8	0.8	3.8	96.2	13.4	9.8	4.7	0.5	273	15.3	4.4	70.5	70.5	10
PD00820	PD00820b	Pre-LN	Male	125	74	26.4	4.2	1.4	2.7	0.5	2	97.9	14	7	4	0.4	285	12.6	4.8	70.5	73.7	10
PD00821	PD00821a	Control	Female	154	96	32	7.8	1.1	5.9	1.9	2.9	88.7	12.5	7	4	0.4	287	12.5	5.2	54.7	54.7	21.1
PD00827	PD00827c	Control	Male	126	80	26.3	5.4	1.1	3.3	2.2	1.2	96.9	14.2	4.2	4.6	0.4	179	15.1	5.8	75	75	7.2
PD00831	PD00831c	Control	Male	159	84	31.6	3.5	0.8	1.8	2.1	2.3	93.1	14.3	8.1	4.5	0.4	151	14.4	5.7	72.9	72.9	6.5
PD00833	PD00833c	Control	Male	138	74	24.2	3.5	1.5	1.5	1.1	1.4	93.3	13.7	6.6	4.4	0.4	172	13.6	6	75.2	75.2	8.9
PD00835	PD00835c	Control	Female	125	74	30.3	5.6	1.9	2.9	1.8	NA	106.6	17.8	6.1	4.1	0.4	203	13.2	6.4	61.8	61.8	8.9
PD00836	PD00836c	Control	Male	139	94	28.6	6.9	1.1	5.3	1.2	1.3	86.5	15.7	5.1	5.2	0.4	240	15.2	5.5	66.2	66.2	9.7
PD00844	PD00844c	Control	Female	129	80	24.9	5.1	1.8	2.8	1.1	1.5	93.7	15.3	4.5	3.9	0.4	216	12.7	4.9	58.1	58.1	9.9
PD00849	PD00849c	Control	Female	108	68	20.5	5	1.7	2.8	1.1	2.3	95.5	13.1	6.4	3.9	0.4	190	12.8	5.6	55.1	55.1	7.4
PD00852	PD00852c	Control	Female	142	78	22.8	5.9	2.9	2.8	0.5	2.6	101.2	15	5.2	4.2	0.4	182	13.2	5.9	75	75	9.1
PD00856	PD00856c	Control	Female	123	72	24.5	4.9	1.6	2.5	1.8	2.1	92.3	13.6	7.5	4.2	0.4	258	13.1	5.5	72.2	72.2	10
PD00860	PD00860c	Control	Male	126	76	26.8	5.2	1.2	3	2.2	1.9	89.5	13.8	9.1	4.9	0.4	200	14.6	NA	61.2	61.2	8.7
PD00861	PD00861c	Control	Female	158	77	54.2	3	1	1.3	1.7	1.8	86.7	14.9	8	4	0.4	238	11.7	7.4	67	67	7
PD00864	PD00864c	Control	Male	164	88	24	3.8	2.1	1.6	0.4	2	94.1	13.3	7.9	4.6	0.4	176	14.6	5.8	75.5	75.5	6.5
PD00865	PD00865c	Control	Male	128	77	30.1	6.9	1.3	4.6	2.2	1.6	83.6	13.1	8.6	5.7	0.5	135	16.1	6	78.5	78.5	7.7
PD00866	PD00866c	Control	Male	149	84	23.7	5.8	1.6	3.6	1.3	2.3	86.5	15.2	7	5.1	0.4	170	14.9	5.4	66.7	66.7	10.5
PD00870	PD00870c	Control	Female	108	68	32.3	7	1	5	2.2	1.5	90.2	14.1	5.2	4.8	0.4	221	14.5	5.8	60.7	60.7	7.1
PD00889	PD00889c	Control	Female	136	70	24.1	5.2	2.3	2.6	0.7	2.4	96	14.4	5.6	4.3	0.4	161	14.1	5.9	70.1	70.1	7.1
PD00890	PD00890c	Control	Male	151	82	30.2	4.4	2	2	1	0.8	95.3	14	10.5	3.9	0.4	320	12.4	6	80	80	10.6
PD00893	PD00893c	Control	Male	124	61	25.8	3.8	1.5	2.1	0.6	1.1	91.4	14.5	5.3	4.2	0.4	216	13.1	7.9	75.9	75.9	3.5
PD00894	PD00894c	Control	Female	103	78	28.1	7.2	2.5	4.3	1	2.2	97.9	13.9	4.1	3.9	0.4	239	12.9	5.4	55	55	10.5
PD00895	PD00895c	Control	Male	147	86	31	3.2	1.1	1.1	2.3	1.1	91	15.6	6.2	4.8	0.4	200	14.3	6.2	72.3	72.3	7.6
PD00899	PD00899c	Control	Male	129	70	22.8	5	1.9	2.8	0.7	2.3	99.8	13.8	5.4	4.4	0.4	174	14.8	5.7	70.6	70.6	7.4
PD00901	PD00901c	Control	Female	120	71	22.3	4.3	1.4	2.6	0.8	1.2	91.8	15	7	3.8	0.3	221	11.8	5.8	61.3	61.3	8.8
PD00904	PD00904c	Control	Female	147	83	23.6	4.3	1.9	2	1	1.7	92.2	13.4	6.7	4.6	0.4	197	13.9	5.5	74.6	74.6	10.3
PD00909	PD00909c	Control	Male	157	89	27.6	4.5	1.4	2.4	1.6	1	94.7	14.9	4.5	4.6	0.4	214	14.5	5.5	76.5	76.5	8.9
PD00927	PD00927c	Control	Female	140	82	28.8	4.6	1.4	2.4	1.9	4.2	92.2	13.7	10.6	4.6	0.4	245	14	5.9	77.1	77.1	8.9
PD00928	PD00928c	Control	Female	125	72	21.2	6.7	1.6	4.6	1.1	0.9	89.9	13.7	3.9	4.4	0.4	187	13.2	5.4	72.2	72.2	10.4
PD00929	PD00929c	Control	Male	149	88	27.5	4.3	1.6	2.2	1.3	2.2	90.3	13.9	7.8	4.6	0.4	273	14.5	5.4	66.4	66.4	10
PD00930	PD00930c	Control	Male	144	87	34.5	4.7	1.2	2.5	2.2	1.9	95.6	13.9	7.3	5	0.5	184	16	5.6	72.2	72.2	8.9
PD00944	PD00944c	Control	Female	119	71	29	5.4	1.6	3	1.9	1.8	91.8	14.4	6.1	4.2	0.4	262	13	5.9	70.3	70.3	8.8
PD00945	PD00945c	Control	Male	157	89	27.6	4.5	1.4	2.4	1.6	1	94.7	14.9	4.5	4.6	0.4	214	14.5	5.5	76.5	76.5	8.9
PD00946	PD00946c	Control	Female	158	100	23.4	5	1.8	2.4	1.8	1.6	91.6	14.9	6.4	4.5	0.4	254	13.5	6.1	71	71	8.9
PD00950	PD00950c	Control	Male	160	87	28.9	5.4	1.4	3.5	1.3	2.2	94.1	14.3	7.2	4.8	0.4	206	15.2	5.8	72.8	72.8	8.1
PD00952	PD00952c	Control	Female	112	70	20.3	5.1	1.3	3.3	1.3	1.6	89.9	15.7	4.7	4.2	0.4	131	12.4	5.8	63	63	8.4
PD00955	PD00955c	Control	Female	150	86	32.4	7.7	1.6	5.5	1.5	2.2	91.7	15.2	6.2	4.1	0.4	252	12.4	6.3	67	67	8.7
PD00958	PD00958c	Control	Female	152	86	28.6	6.9	1.9	4.6	0.9	1.2	92.5	14.4	4.7	4.6	0.4	356	14.2	5.5	82.3	82.3	6.8
PD00960	PD00960c	Control	Female	130	70	31.8	5.1	1.5	3.1	1.3	1.6	94.4	14.8	6.7	4.2	0.4	152	13.3	6.2	73	73	9.4
PD00961	PD00961c	Control	Male	158	96	33.8	4.7	1.1	2.7	2.1	1.5	98.7	14.2	6.4	4.5	0.4	193	14.8	5.7	69.7	69.7	8
PD00962	PD00962c	Control	Male	126	78	25.8	6.1	1	3.1	4.4	4.1	91.2	13.9	8.8	5.2	0.5	181	16.2	5.9	66.7	66.7	7.8
PD00970	PD00970c	Control	Male	178	84	28.1	4	1.1	2.4	1.3	1.2	103.2	14.7	5.5	3.8	0.4	166	13.6	6.1	75.6	75.6	8.2

Appendix 13: Validation cohort pre-lymphoid neoplasm cases and controls metadata

Individual ID	Sample ID	Group	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	BMI	Total cholesterol (mmol/L)	HDL (mmol/L)	LDL (mmol/L)	Triglycerides (mmol/L)	Lymphocytes (10^9/L)	MCV (fL)	RDW	WBC (10^9/L)	RBC (10^9/L)	Haematocrit (%)	Platelets (10^9/L)	Haemoglobin (g/dL)	HbA1c (%)	Age at first sample	Age at sample	Follow-up (years)
PD00006	PD00006b	Control	Female	110	66	33.2	4.9	1	3.2	1.7	2.1	90.9	13.8	8.9	4.1	0.4	207	12.7	5.5	44.6	44.6	17.8
PD00007	PD00007b	Control	Male	116	74	27.3	5.9	1	3.7	2.8	2.7	91.7	13	13.4	5.1	0.5	278	16.1	5.3	51.9	51.9	19.3
PD00008	PD00008a	Pre-LN	Female	162	96	20.6	6.8	1.8	4.5	1	1.6	84.5	14.3	6.4	4.9	0.4	232	14.3	NA	71.7	71.7	9.7
PD00012	PD00012a	Pre-LN	Male	182	106	26	5.5	1.3	3.7	1	3.1	82.9	14.6	7.1	4.9	0.4	187	14.7	NA	72.7	72.7	6.1
PD00012	PD00012b	Pre-LN	Male	170	98	25.9	5.9	1	4.4	1.3	16.7	83	14.6	22.8	5	0.4	193	14.7	5.6	72.7	76.2	6.1
PD00013	PD00013b	Control	Female	120	84	18.6	5.8	2.9	2.4	1.1	2.2	88.3	13.3	5.6	4.3	0.4	253	12.7	5.9	57	57	19.7
PD00013	PD00013c	Control	Female	109	72	19.9	5.3	2.2	2.8	0.7	2	92	NA	6.4	4.3	0.4	313	13.2	6	57	66	19.7
PD00018	PD00018a	Control	Female	148	84	31	8.4	1.8	5.7	1.9	1.2	86.1	14.5	3.7	4.3	0.4	234	12.9	NA	70	70	21.5
PD00020	PD00020b	Control	Male	126	82	25.3	6.7	0.8	4.4	3.5	3.5	88.5	12.6	7.4	5.3	0.5	225	16.8	5.3	55.8	55.8	18.7
PD00028	PD00028a	Pre-LN	Female	144	84	27.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	66.6	66.6	16
PD00030	PD00030b	Control	Male	140	88	27.3	6.6	2.1	3.9	1.5	2.3	94.4	13.1	6.1	4.9	0.5	216	14.8	5.7	63.6	63.6	19.7
PD00030	PD00030c	Control	Male	156	98	25.9	4.9	2	2.5	0.9	1.9	94.1	14.5	6	4.4	0.4	202	14	5.5	63.6	73.4	19.7
PD00033	PD00033a	Pre-LN	Male	149	89	22.9	6.4	1	4.5	2	2.3	83.2	12.9	7.5	5.4	0.5	333	15.5	NA	51.4	51.4	3.8
PD00033	PD00033b	Pre-LN	Male	128	74	19.1	4.5	0.9	3.2	0.9	1.1	68.9	15.6	7.7	5.1	0.4	737	10.5	6.4	51.4	55.2	3.8
PD00033	PD00033c	Pre-LN	Male	120	82	23.6	3.5	0.8	2	1.6	1.1	87	NA	5.1	4.7	0.4	189	13.4	6.1	51.4	64	3.8
PD00045	PD00045b	Control	Female	130	79	27.4	6	2	3.5	1.2	1.8	89.3	12.8	4.4	4.2	0.4	213	13.1	5	76	76	14.2
PD00045	PD00045c	Control	Female	121	74	27.1	5.7	1.7	3.4	1.4	1.7	91.2	14.1	4.2	4.5	0.4	213	13.5	5.6	76	85.8	14.2
PD00046	PD00046a	Pre-LN	Female	139	88	31.3	6.7	1.4	4.5	1.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	64.3	64.3	7.5
PD00050	PD00050b	Control	Female	146	74	22.5	7.2	2.3	4.6	0.8	1.8	86.8	13.9	5.5	3.6	0.3	254	11.7	5.8	78.2	78.2	19.1
PD00052	PD00052a	Pre-LN	Female	152	98	30.7	7.1	1.6	4.7	1.8	2.8	96.3	13.6	9.8	4.2	0.4	402	13.1	6.3	62.6	62.6	4.5
PD00053	PD00053a	Pre-LN	Female	106	70	28.4	6.2	1.2	3.2	4	3.6	92.2	12.5	7.9	4	0.4	322	13	NA	63.7	63.7	10.6
PD00059	PD00059b	Control	Female	151	90	30.7	6.9	1.6	3.1	4.9	2	93.3	12.4	8	3.9	0.4	267	13	5.7	76.2	76.2	18.1
PD00061	PD00061a	Control	Female	110	67	26.2	6	1	4.3	1.6	2.2	85.7	14.9	6.1	4.2	0.4	293	11.8	NA	47.6	47.6	22.7
PD00064	PD00064b	Control	Male	146	99	29.1	8.3	1.2	6	2.5	1.8	98.2	12.2	8.1	4.6	0.4	231	14.7	5.8	71.4	71.4	19.9
PD00064	PD00064c	Control	Male	122	72	29.3	4.4	1.3	2.5	1.4	1.2	101.5	13.9	7.2	4.3	0.4	196	14.6	6.4	71.4	81.9	19.9
PD00072	PD00072a	Control	Male	158	102	21.7	7.5	1.6	5.3	1.4	2.9	88.5	13.6	8.2	5.2	0.5	288	16.5	5.3	64.8	64.8	21.6
PD00081	PD00081a	Control	Female	116	76	17.3	6.3	2.4	3.3	1.3	1.5	87.6	13.2	9.4	4.3	0.4	279	13.3	NA	65.7	65.7	22.6
PD00082	PD00082b	Control	Female	139	85	25	6.9	1.9	4.5	1.3	2.5	87.2	14.2	6.4	4.7	0.4	288	13.8	4.7	59.3	59.3	18.4
PD00083	PD00083a	Pre-LN	Female	148	93	22.7	5.9	1	4.1	1.8	2.6	86	13.5	10.6	4.1	0.4	233	12.7	4.9	68.3	68.3	0.9
PD00085	PD00085b	Control	Male	146	78	25.8	7.2	2	4.9	0.8	1.5	90.7	12.9	4.7	4.5	0.4	208	14.5	5.2	65.7	65.7	18.3
PD00087	PD00087a	Control	Female	136	83	26.1	8	1.5	4.9	3.7	1.7	88.7	12.9	5.4	4.3	0.4	341	12.4	5.6	63.3	63.3	20.7
PD00090	PD00090b	Control	Female	148	90	28	5.1	1.1	3.5	1.2	2	82	14.4	5.4	4	0.3	289	11.1	5.8	71.3	71.3	19.4
PD00090	PD00090c	Control	Female	152	76	27.4	5.2	1	3.6	1.4	1.8	90.9	16.3	5.9	4	0.4	234	12.4	6.1	71.3	81.1	19.4
PD00096	PD00096a	Control	Female	110	70	28.1	5	1.2	3.4	0.9	2.4	94.4	13.4	7.9	4.1	0.4	222	12.6	4.8	59.7	59.7	20.9
PD00104	PD00104b	Control	Male	123	76	33.9	5.4	1.2	3.4	1.8	2.3	89.1	13.5	8.2	6.3	0.6	379	16.6	5.1	44	44	17.6
PD00104	PD00104c	Control	Male	126	79	34.2	5	1.1	3	2.1	2.3	88	NA	7.8	4.8	0.4	260	14.7	5.5	44	49.5	17.6
PD00105	PD00105a	Pre-LN	Female	132	78	23.2	5.3	2	3	0.7	2.4	92	12.9	8.6	4.3	0.4	339	13.5	NA	66.3	66.3	15
PD00119	PD00119a	Pre-LN	Female	150	88	27.3	7.5	0.8	5.9	1.9	1.6	88.1	13.5	5.9	4.4	0.4	181	12.3	4.6	65.1	65.1	7
PD00128	PD00128b	Control	Female	128	82	27.5	5.2	1.8	3	1.2	1.9	83	14.1	6.9	4.4	0.4	325	13	8.2	70	70	17.8
PD00133	PD00133a	Pre-LN	Female	144	93	35	7.6	1.3	5.5	1.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	69.7	69.7	7
PD00137	PD00137a	Pre-LN	Male	146	93	29.7	7.3	0.9	4.8	3.7	3.4	87.1	13.1	9.7	5.1	0.4	268	14.5	5.3	76.1	76.1	3.9
PD00137	PD00137b	Pre-LN	Male	126	82	31.7	7.7	0.7	4.1	6.4	5.6	86.3	13.6	11.4	4.9	0.4	269	14.5	5.8	76.1	78.3	3.9
PD00138	PD00138a	Control	Female	140	80	33	6.2	1.7	3.5	2.1	2.3	89.9	13.7	8.3	4.2	0.4	203	13	NA	68.4	68.4	19.5
PD00139	PD00139a	Pre-LN	Female	155	90	28.6	7.2	1.6	5	1.2	2.1	94.2	13	6	4.5	0.4	313	14.1	NA	62.5	62.5	19
PD00144	PD00144b	Control	Female	148	88	22	5.3	1.9	2.5	2	1.4	89.5	12.6	7	3.9	0.4	332	12	7.9	69.7	69.7	19.4
PD00145	PD00145a	Pre-LN	Male	180	102	31.3	4.5	1.1	2.2	2.8	2.8	90.1	13.7	9.7	5.6	0.5	176	17.4	4.5	71.4	71.4	4.5
PD00146	PD00146b	Control	Male	127	68	24.9	6.3	1	3	5.2	2.9	85.5	14.2	8.6	4.7	0.4	199	14	5.4	74.1	74.1	15.9
PD00149	PD00149b	Control	Female	105	66	21.9	7	2.4	3.7	2.1	1.7	92.8	12.5	4.8	4.1	0.4	233	13.1	5.3	66.3	66.3	18.8
PD00149	PD00149c	Control	Female	111	66	22.2	7.2	2.2	4.5	1.1	1.6	95	NA	4.5	4.4	0.4	247	13.7	5.5	66.3	74.5	18.8
PD00151	PD00151a	Control	Male	148	90	28.6	5.9	1	3.9	2.2	2.5	87.6	12.9	7.7	4.8	0.4	259	15.3	NA	63.5	63.5	22.5
PD00155	PD00155a	Control	Female	113	72	24.2	4.9	1.4	3	1.3	2	91.9	12.2	6.5	4.7	0.4	333	13.9	4.6	66.1	66.1	21.5
PD00158	PD00158c	Control	Female	146	72	26.5	3.4	1.4	1.5	1.1	1.6	98.7	16.3	8.1	4	0.4	71	13	5.5	76.4	76.4	8.5
PD00167	PD00167a	Control	Male	156	99	33.9	7.3	1.6	5	1.6	1.6	84.7	14	6.2	5.3	0.4	158	15.7	NA	65.1	65.1	21.9
PD00167	PD00167c	Control	Male	142	90	35.2	4.4	1.5	1.9	2.2	1.6	86	NA	7.3	5.6	0.5	157	16.2	8.1	65.1	75.8	21.9
PD00169	PD00169a	Pre-LN	Female	113	69	24.5	4.6	1.4	2.2	2.1	0.9	93.7	13.1	4.6	4.5	0.4	285	13.2	NA	62.9	62.9	11.8
PD00173	PD00173a	Control	Female	116	79	25.9	5.5	1.2	3.8	1	1.2	84.8	14.9	6.1	4.6	0.4	305	13.7	NA	62.8	62.8	22.9
PD00173	PD00173c	Control	Female	121	68	21.9	4.8	1.5	2.8	1.3	0.9	88.1	17.8	5.7	4.3	0.4	272	12.2	5.9	62.8	76.1	22.9
PD00180	PD00180a	Control	Female	157	91	34.3	5.3	1.1	3.2	2.2	3.5	94.3	12.8	8.7	4.8	0.5	209	15.5	NA	68.9	68.9	21.2
PD00180	PD00180c	Control	Female	142	70	31.8	4.8	1.8	2.6	0.9	1.7	100.4	13.6	5.2	3.8	0.4	127	13	6	68.9	83.3	21.2
PD00181	PD00181a	Pre-LN	Male	120	72	27.3	6.8	1.3	4.8	1.7	1.3	97.6	13.1	5.4	4.3	0.4	174	14.2	5.8	69.2	69.2	3.5
PD00181	PD00181b	Pre-LN	Male	126	68	25.4	7	1.5	5	1.1	1.7	102.5	15	5.6	3.9	0.4	188	13	5.5	69.2	72.5	3.5
PD00187	PD00187b	Control	Female	110	70	28.1	4.9	2.1	2.4	0.9	1.1	102.3	13.6	3.2	3.6	0.4	227	11.9	5.2	65.9	65.9	17.5
PD00188	PD00188a	Control	Female	118	79	30	7	1.6	4.5	1.9	1.4	92.5	12.4	7	3.9	0.4	210	12.8	NA	72.3	72.3	19.2
PD00188	PD00188c	Control	Female	111	70	32.3	4	1.8	1.7	1.1	1.4	95.3	13.6	6.8	4	0.4	211	13.1	6.4	72.3	85.5	19.2
PD00189	PD00189b	Control	Male	118	79	33.2	8.5	1	5.6	4.3	2.3	94.6	13.5	7.7	4.9	0.5	323	16	6	67.9	67.9	17.5
PD00191	PD00191a	Pre-LN	Female	132	83	19.8	4.7	1.8	2.5	0.9	1.1	88	13.7	5.6	4.2	0.4	161	12.2	NA	64.9	64.9	6.5
PD00193	PD00193a	Pre-LN	Female	124	82	30.3	6.1	1.2	3.9	2.2	2.6	87.6	13	10.1	4.5	0.4	335	14.2	5.4	48.6	48.6	7.3
PD00196	PD00196a	Control	Male	153	94	25.2	5.8	1.2	3.6	2.1	2	91	13.4	7.6	5.7	0.5	278	17.3	4.3	68.6	68.6	22.3
PD00196	PD00196c	Control	Male	126	68	23.6	5.2	1	2.9	2.9	2.1	91.5	13.2	13.4	5.4	0.5	324	16.7	5.3	68.6	80.4	22.3
PD00201	PD00201a	Pre-LN	Male	114	69	23.6	5.5	1.1	3.3	2.4	1.5	90.4	11.8	6.5	4.1	0.4	288	12.2	5.9	62.5	62.5	0.9
PD00212	PD00212a	Pre-LN	Male	107	69	27.5	5	1.1	3.5	1	2.1	90.4	13	9.3	4.6	0.4	182	14.6	5	59.4	59.4	14.3
PD00215	PD00215b	Control	Male	139	87	29.4	7.7	1.3	5.3	2.6	1.6	94.4	12.9	4.9	5	0.5	125	15.6	5.7	56.2	56.2	19.7
PD00219	PD00219a	Control	Female	113	72	23	5.8	2.1	3.2	1.2	1.9	93.9	13	7.1	3.9	0.4	238	11.9	4.5	47.4	47.4	20.7
PD00223	PD00223a	Pre-LN	Male	149	84	28.8	6.1	NA	NA	4.8	NA	NA	NA	NA	NA	NA	NA	NA	NA	64.2	64.2	7.4
PD00224	PD00224a	Control	Male	123	80	25.7	6.7	1.2	4.9	1.3	2	87.5	14.2	7.4	5	0.4	316	14.6	6.3	48.4	48.4	22.2
PD00231	PD00231b	Control	Female	123	74	23.3	5.9	2.8	2.8	0.8	1.5	82.8	13	5.5	3.7	0.3	267	11	5	50.3	50.3	18.1
PD00232	PD00232a	Pre-LN	Female	132	78	28.7	6.8	1.6	4.2	2.4	2.3	91.6	16.2	6.1	4.6	0.4	201	13.2	5.6	68.4	68.4	9.6
PD00237	PD00237b	Control	Female	134	81	24.2	5.9	3	2.2	1.6	1.3	86.9	13.3	5.1	4.6	0.4	243	13.4	5.6	79.3	79.3	14
PD00237	PD00237c	Control	Female	151	82	23.2	6.5	2.8	3	1.8	1.4	91	NA	5.7	4.4	0.4	229	13	5.6	79.3	87	14
PD00238	PD00238a	Pre-LN	Male	130	74	23.6	5.3	1.5	3	1.7	3.7	88.7	12.6	9	4.4	0.4	184	13.5	NA	65.4	65.4	6
PD00242	PD00242b	Control	Female	126	82	28.1	6.1	1.5	4.1	1.2	2.6	94.8	13	7.4	4.5	0.4	266	13.6	5.6	68.2	68.2	19.5
PD00242	PD00242c	Control	Female	163	94	27.6	6.4	1.4	4.5	1.3	2.4	91	NA	5.2	4.2	0.4	291	13.1	5.5	68.2	76.7	19.5
PD00245	PD00245b	Control	Male	138	84	23.5	5.7	2.4	2.5	1.9	2	98.1	12.4	4.3	4.5	0.4	176	14.6	5.2	61.1	61.1	19.7
PD00248	PD00248b	Control	Male	160	114	31.1	6.1	0.8	4.1	2.8	2.5	89.4	12.9	8.7	4.8	0.4	224	14.8	5.1	61	61	18.4
PD00252	PD00252b	Control	Female	158	97	27.2	6.1	1.8	3.9	1	3.5	89.4	11.9	7.9	4.3	0.4	312	13.8	5	62.8	62.8	18.5
PD00255	PD00255b	Control	Male	146	86	32.6	6.6	1.3	4.3	2.2	2.2	89.3	13.7	5.4	5	0.4	223	16.2	5.1	60.2	60.2	18.9
PD00255	PD00255c	Control	Male	150	88	33.6	5.9	1.3	3.7	2.1	1.4	92.4	14.5	5.9	5	0.5	273	15.7	5.3	60.2	68.6	18.9
PD00256	PD00256a	Control	Male	145	90	24.4	5.4	1	3.4	2.3	3	92.3	12.8	7.7	4.5	0.4	314	13.9	5.9	61.4	61.4	21.5
PD00260	PD00260a	Pre-LN	Female	123	72	28.8	5.8	NA	NA	1.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	47.1	47.1	10.1
PD00261	PD00261b	Control	Male	166	110	38.4	6.7	0.9	3.9	4.2	3.3	87.8	13.3	11.9	4.9	0.4	207	14.8	6.1	58.7	58.7	19.5
PD00265	PD00265a	Pre-LN	Female	137	69	23.5	5.6	1.3	3.8	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	62.6	62.6	1
PD00278	PD00278a	Control	Female	134	85	29.5	6.5	1.4	4.4	1.5	2.2	90.5	13	12	4.5	0.4	384	14	5.9	65.7	65.7	22.2
PD00278	PD00278c	Control	Female	137	74	32.6	7.7	1.9	4.6	2.7	3.1	90	14.9	8.2	4.8	0.4	315	15	5.6	65.7	77.7	22.2
PD00279	PD00279a	Pre-LN	Male	151	86	30.2	6.5	1	4	3.4	2.3	89.7	12.5	7.5	4.5	0.4	219	14.1	4.9	53.6	53.6	4.3
PD00286	PD00286a	Control	Male	129	80	24.2	6.5	1.1	4.6	1.8	1.6	90.8	13.1	5.5	5.2	0.5	250	15.9	NA	61.8	61.8	22.5
PD00290	PD00290a	Pre-LN	Male	154	92	27.2	5.2	1.6	2.9	1.7	1.1	92.9	12.6	4.5	5	0.5	278	16.6	4.5	65.5	65.5	6.7
PD00295	PD00295b	Control	Male	128	94	22.5	5.5	1.6	3.4	1.1	2.1	86.7	12.3	6.5	4.6	0.4	338	14	5.3	57.2	57.2	18.4
PD00296	PD00296a	Pre-LN	Female	114	78	31.3	5.9	1.6	3.3	2.2	2.2	91.9	13.5	6.9	4.5	0.4	314	13.3	NA	63.4	63.4	10.1
PD00300	PD00300b	Control	Male	132	70	24.7	6.9	1.2	5.1	1.5	1.1	87.6	14.1	4.1	4.9	0.4	172	14.9	5	73.2	73.2	17.6
PD00303	PD00303a	Control	Male	116	74	26.4	5	1.3	3.4	0.7	2.2	88.4	13.3	6.8	4.8	0.4	247	14.2	5.5	64.9	64.9	21.2
PD00307	PD00307b	Control	Male	127	58	24.9	4.2	1.2	2.6	1	1.2	89.3	13.7	5.7	4.6	0.4	125	14.6	5.1	76.3	76.3	11.5
PD00307	PD00307c	Control	Male	112	61	24.9	2.7	1.1	1.2	0.8	1.4	91	NA	9.3	4	0.4	150	12.3	5.6	76.3	83.9	11.5
PD00308	PD00308a	Pre-LN	Female	176	92	29.7	6.4	1.8	4.1	1.2	2.3	80.9	13.2	5.9	4.6	0.4	196	13.4	5.8	70	70	1.9
PD00315	PD00315a	Pre-LN	Male	134	86	26.3	6.4	0.9	4.4	2.4	NA	NA	NA	NA	NA	NA	NA	NA	NA	54.2	54.2	8.1
PD00315	PD00315b	Pre-LN	Male	130	80	24.9	6.3	1	4.2	2.6	9.1	93.5	12.8	14.3	4.6	0.4	233	14.5	5.6	54.2	58.7	8.1
PD00316	PD00316a	Control	Male	158	90	25.3	6.4	1.2	4.4	1.8	1.4	94.4	12.5	5.2	4.3	0.4	202	13.6	6	65.3	65.3	21.5
PD00316	PD00316c	Control	Male	143	86	26	4	1.5	2.3	0.6	NA	NA	NA	NA	NA	NA	NA	NA	5.6	65.3	76.4	21.5
PD00324	PD00324a	Pre-LN	Male	178	104	29.5	6.3	NA	NA	4.9	1.8	88.2	12.8	6.3	5.5	0.5	252	16.1	5.2	65.9	65.9	3.9
PD00325	PD00325a	Pre-LN	Female	117	73	27.5	5.8	1.1	4.1	1.4	NA	NA	NA	NA	NA	NA	NA	NA	NA	46.4	46.4	11.5
PD00325	PD00325c	Pre-LN	Female	138	86	26.2	6.8	1.4	4.6	1.9	3.5	93.3	14.2	6.8	4.4	0.4	126	13.7	5.8	46.4	63.7	11.5
PD00327	PD00327b	Control	Female	104	61	19.6	4.4	1.4	2.7	0.7	1.6	85	14	9.1	4.2	0.4	185	12.4	5.1	52.6	52.6	18.3
PD00333	PD00333c	Control	Female	144	76	33.9	3.2	1	1.4	1.8	1.1	95.7	14.2	3.9	4	0.4	195	12.6	8.6	67.6	67.6	8.7
PD00342	PD00342a	Pre-LN	Male	146	92	30.3	5.5	1.5	3.2	1.9	2.5	91.1	12.9	9.6	4.3	0.4	254	13.6	6.5	58.3	58.3	10.8
PD00344	PD00344a	Control	Female	138	77	25.9	5.6	1.5	3.4	1.5	2.2	89.9	12.6	6.7	4.8	0.4	223	14.5	NA	64.5	64.5	22.8
PD00344	PD00344c	Control	Female	129	76	24.3	4.9	1.8	2.5	1.4	2.8	96	NA	8.6	4.7	0.4	241	14.5	5.9	64.5	76.4	22.8
PD00347	PD00347b	Control	Male	132	75	24.8	5.5	1.1	3.4	2.3	1.2	93	12.6	4	4.6	0.4	186	15.6	5.1	63.7	63.7	18.5
PD00349	PD00349a	Pre-LN	Female	162	112	26	6.4	1.3	3.9	2.6	1.5	81.8	13.1	5.6	4.9	0.4	207	13.9	5.5	64.9	64.9	9.2
PD00352	PD00352a	Control	Male	166	98	23.2	6.1	2.4	3.4	0.8	1.7	92.5	13.1	5.9	4.7	0.4	128	15.1	5.7	61.4	61.4	21.3
PD00364	PD00364a	Control	Male	134	84	25.2	6.5	1.8	4.2	1.1	1.7	89.6	13.6	5.7	5.1	0.5	280	15.9	5.3	53.4	53.4	22
PD00366	PD00366a	Pre-LN	Male	141	92	32.1	5.8	1	3.3	3.3	3.1	85.9	13.7	8.8	5.2	0.4	247	15.6	NA	57.8	57.8	16.8
PD00372	PD00372b	Control	Male	140	82	31.6	6.3	1.8	4	1.3	1.9	91.6	12.4	6.7	5.1	0.5	342	16.1	5	61.1	61.1	17.8
PD00375	PD00375a	Pre-LN	Female	131	73	23.5	6.2	2.3	3.4	1.1	2	91.3	13.6	4.4	4	0.4	212	11.8	5.2	61.3	61.3	13
PD00376	PD00376b	Control	Female	112	68	26.1	5.9	1.9	3.4	1.5	1.7	93.3	12.5	4.9	4.2	0.4	238	14.1	4.7	49.4	49.4	18.6
PD00387	PD00387a	Control	Male	142	105	28.1	8.1	0.9	5.9	3	2.6	92.6	14	5.9	6.1	0.6	268	17	4.9	59.7	59.7	20.7
PD00397	PD00397b	Control	Female	132	82	19.2	5.5	2.1	2.5	2	2.5	90.1	13.4	9.8	4.8	0.4	322	15.5	5.5	73.8	73.8	14.6
PD00397	PD00397c	Control	Female	151	82	18.6	6.2	3.2	2.5	1.3	0.5	105	15.6	10	4.4	0.5	260	14	NA	73.8	81.9	14.6
PD00404	PD00404a	Pre-LN	Female	117	72	22.8	4.9	1.4	2.8	1.5	2.2	90.3	13.3	5.7	4.6	0.4	202	14.8	NA	66.3	66.3	2.1
PD00406	PD00406a	Pre-LN	Male	150	90	28	7.7	1.4	4.6	3.8	1	81.6	14.2	3.6	5	0.4	263	13.1	5.4	61.1	61.1	2.6
PD00408	PD00408b	Control	Male	163	106	27.8	5.5	1.4	3.6	1.3	1.6	102.6	12.6	7.4	4.2	0.4	277	15	5.5	65.9	65.9	18.4
PD00416	PD00416b	Control	Female	120	74	23.6	7.4	1.7	5.3	1	1.9	94	11.9	5.3	4.3	0.4	255	13.6	5.2	63.8	63.8	18.4
PD00416	PD00416c	Control	Female	123	72	24.5	4.9	1.6	2.5	1.8	2.1	92.3	13.6	7.5	4.2	0.4	258	13.1	5.5	63.8	72.2	18.4
PD00430	PD00430a	Pre-LN	Male	118	79	21	7	1.3	4.5	2.5	2.3	94.1	13.3	6.3	4.9	0.5	297	15.3	5.7	51.8	51.8	18.8

PD00434	PD00434b	Pre-LN	Female	137	81	23.1	6.1	2	3.8	0.8	2	93.6	12.8	7.5	3.8	0.4	255	12.4	4.9	53.1	53.1	10.1
PD00435	PD00435a	Pre-LN	Female	172	84	23.2	5.5	1.8	3.2	1.1	1.1	64	17.3	4	4.4	0.3	526	8.7	5.6	57.6	57.6	14.6
PD00437	PD00437a	Control	Female	123	70	28	6.2	1.6	4.3	0.7	2	88	13.1	5	4.4	0.4	220	13.6	NA	63.5	63.5	23
PD00437	PD00437c	Control	Female	148	88	31.8	5.8	1.9	2.9	2.3	1.8	91.7	13.7	4.3	4.2	0.4	214	12.7	5.7	63.5	76.2	23
PD00440	PD00440a	Control	Female	132	74	28.1	8.2	1.6	5.9	1.6	1.7	92.9	12.1	5	4.8	0.4	239	14.5	6.1	65.6	65.6	20.9
PD00441	PD00441a	Control	Male	148	90	24.1	4.8	0.8	3.3	1.6	2.9	91.5	12.6	6.4	4.3	0.4	164	13.4	NA	68.5	68.5	23.4
PD00442	PD00442a	Control	Female	136	82	27.6	6	1.2	4.1	1.6	2.3	88	12.4	6.1	4.4	0.4	272	13.2	NA	59.2	59.2	23.1
PD00442	PD00442c	Control	Female	144	84	26.7	4.9	1.1	3.1	1.5	1.8	88	13.7	4.4	4.6	0.4	195	13.4	5.7	59.2	74.2	23.1
PD00457	PD00457a	Control	Female	127	86	19	6.6	2.3	4	0.8	1.3	89.8	16	3.5	4.6	0.4	127	13	5	61.2	61.2	20.6
PD00460	PD00460b	Control	Female	156	88	25.1	5.8	2.4	2.7	1.7	1.8	95	12.7	6.2	3.5	0.3	287	12.1	5.5	72.7	72.7	17.8
PD00461	PD00461a	Pre-LN	Male	141	77	22.9	6	1.5	4.2	0.6	NA	NA	NA	NA	NA	NA	NA	NA	NA	73.8	73.8	11.7
PD00470	PD00470a	Control	Male	151	90	22.9	3.8	1.3	2	1.2	2.3	85.2	14.4	7.6	4.8	0.4	268	14.2	NA	64	64	22.5
PD00470	PD00470c	Control	Male	138	82	27.8	3.5	1	1.8	1.6	1.9	90.4	15.5	5.9	4.6	0.4	203	13.4	5.7	64	77.4	22.5
PD00471	PD00471a	Control	Female	140	84	26.1	5.9	2.1	3.3	1	1.7	85.9	13.5	4.2	4.7	0.4	243	14.1	NA	61.3	61.3	22.4
PD00471	PD00471c	Control	Female	154	85	29.2	5.7	1.7	3.3	1.6	2.2	84	NA	5	4.5	0.4	187	13.1	NA	61.3	72.2	22.4
PD00476	PD00476a	Control	Male	114	68	30.4	6.7	1.1	4.5	2.3	2.2	89	12.9	6.3	4.7	0.4	225	13.6	4.8	65.5	65.5	20.3
PD00476	PD00476c	Control	Male	110	67	30.3	3.2	0.8	2	1	8.6	87.9	14.4	12.8	4.3	0.4	124	12.8	6.3	65.5	80.6	20.3
PD00493	PD00493b	Pre-LN	Female	174	103	26.4	7	1.9	4.4	1.7	2.7	81.4	13.7	6.8	5.2	0.4	340	14.7	5.6	68	68	4.2
PD00499	PD00499a	Pre-LN	Male	126	81	28.4	7.4	1	5.3	2.3	3.1	91.5	12.6	9.1	4.8	0.4	292	15.3	NA	50.8	50.8	9.1
PD00501	PD00501b	Control	Female	158	96	26.8	6.6	1.5	4.5	1.5	2.3	81	13.3	6	4.5	0.4	256	12.7	5.5	60.7	60.7	19.4
PD00502	PD00502a	Pre-LN	Male	165	91	27.6	3	0.9	1.6	1.1	1.3	94.8	14	4.5	3.9	0.4	253	11.8	5.4	75	75	0.8
PD00529	PD00529a	Control	Female	108	68	23	4.4	2	2.1	0.7	2	88.5	13	5.6	4.4	0.4	167	13.4	NA	65.3	65.3	22.8
PD00537	PD00537a	Control	Female	169	98	31.2	7.4	1.2	4.8	3	2.9	93.8	12.4	8.2	4.3	0.4	276	13.3	NA	63.4	63.4	22.6
PD00537	PD00537c	Control	Female	138	66	30.9	4.2	1.6	1.9	1.6	2.2	94.9	13.8	7.1	4	0.4	210	12.4	5.9	63.4	78.2	22.6
PD00540	PD00540b	Control	Female	148	79	29.2	3.5	1.2	1	2.9	3.1	76.9	18	9.8	5.2	0.4	312	11.8	5.4	53.5	53.5	17.7
PD00540	PD00540c	Control	Female	152	90	27.1	3.9	1	2	2.4	2.5	87	NA	6.9	5.3	0.5	197	15.7	5.1	53.5	59.4	17.7
PD00541	PD00541a	Pre-LN	Male	122	76	25.4	5.6	1.1	3.2	2.8	NA	NA	NA	NA	NA	NA	NA	NA	NA	73.2	73.2	15.3
PD00546	PD00546a	Pre-LN	Female	129	78	23.4	6.1	1.7	3.8	1.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	59.3	59.3	4.1
PD00547	PD00547a	Pre-LN	Female	144	84	26.1	6.5	1.1	4.5	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	70.8	70.8	0.7
PD00548	PD00548a	Control	Female	111	70	19.2	4.2	1.3	2.5	1	1.9	87.6	13	6.5	4.1	0.4	225	12.2	5.4	46.4	46.4	5.4
PD00550	PD00550a	Pre-LN	Male	150	90	25.8	5.6	0.8	4.2	1.4	1.5	92.4	13.5	7.2	5	0.5	397	16.3	NA	73.3	73.3	10
PD00552	PD00552a	Pre-LN	Female	144	76	22.3	4.8	0.7	3.2	2.1	1	93.5	13.9	5.3	3.5	0.3	329	11.5	5	67.6	67.6	3.8
PD00554	PD00554a	Control	Female	142	88	27.2	5.1	1	2.9	2.7	1.6	85.6	14.6	5.5	4.6	0.4	284	12.1	5.4	58.5	58.5	20.5
PD00556	PD00556a	Control	Male	145	84	25.3	4.8	1.6	2.9	0.7	2.6	83.4	14.5	6.8	4.7	0.4	238	13	5.4	64.8	64.8	20
PD00562	PD00562b	Control	Male	144	82	27.9	6.5	0.8	4.1	3.7	2	87.7	12.7	5.2	5.1	0.4	363	15.4	4.9	52.9	52.9	19.2
PD00562	PD00562c	Control	Male	165	92	26.4	6.3	1.3	3.5	3.3	1.6	91	NA	4.2	5.2	0.5	262	15.4	5	52.9	61.2	19.2
PD00566	PD00566b	Control	Female	134	80	27.7	5.1	1.5	2.7	2.1	2.1	88.4	13.6	5.4	5	0.4	207	14.3	5.7	57.7	57.7	19.7
PD00566	PD00566c	Control	Female	110	71	29	4.8	1.4	2	2.9	2.1	86.9	13.9	6.7	5	0.4	200	14.8	6	57.7	67.6	19.7
PD00580	PD00580b	Control	Female	132	78	25.1	5.4	1.2	3.6	1.5	2.6	90.2	13.1	5.2	4.6	0.4	351	13.2	5.3	67.4	67.4	19.3
PD00580	PD00580c	Control	Female	143	84	25.5	6	1.4	3.8	2	2.1	92	NA	5.1	3.9	0.4	317	11.8	5.4	67.4	74.4	19.3
PD00583	PD00583b	Control	Male	131	72	21.2	7.1	2	4.7	1	1.6	84.2	14	5	5	0.4	209	14.8	5	57.7	57.7	19.2
PD00593	PD00593b	Control	Female	128	77	29.1	7.7	2.3	4.9	1.3	1.2	85.7	13.5	3.6	4.6	0.4	166	14	5.1	66.4	66.4	18.9
PD00595	PD00595b	Control	Male	144	75	24.7	9.3	1.6	6.5	2.8	1.9	88.3	14.6	6.3	4.6	0.4	400	14	6	73.5	73.5	18.4
PD00597	PD00597a	Control	Female	130	89	26.6	5.4	1.4	2.8	2.6	3	85.4	13	8.3	4.7	0.4	232	14.2	NA	63.2	63.2	22.8
PD00599	PD00599a	Control	Male	163	92	24.3	5.9	1	3.7	2.7	1.3	95	12.4	6	4.6	0.4	191	14.8	NA	69	69	22.7
PD00600	PD00600a	Pre-LN	Male	156	108	27.4	7.7	1	5.5	2.7	2.4	88.2	13.4	7.7	4.6	0.4	244	14.3	NA	68.2	68.2	12.7
PD00620	PD00620a	Pre-LN	Female	134	68	19.3	5.1	1.4	3.1	1.3	NA	NA	NA	NA	NA	NA	NA	NA	NA	73.9	73.9	5.2
PD00621	PD00621b	Control	Male	113	66	26.5	6	1.1	3.5	3.1	1.5	90	13	4.7	4.8	0.4	201	13.8	5.7	69.2	69.2	10.4
PD00621	PD00621c	Control	Male	127	69	24.6	4.1	1.4	2	1.7	1.3	87	NA	6.6	4.4	0.4	201	13.1	5.7	69.2	77.4	10.4
PD00622	PD00622b	Control	Female	102	64	19.7	4.4	1.6	2.4	1	1.4	92.3	15.2	4.8	4.1	0.4	187	13.2	5	52.7	52.7	18.7
PD00624	PD00624a	Pre-LN	Male	122	76	20.8	6.3	1.4	4.6	0.7	2.2	88.2	12.4	5.5	4.5	0.4	420	13.1	4.8	44.5	44.5	14.3
PD00624	PD00624c	Pre-LN	Male	118	72	26.8	4.8	1.4	3	0.9	1.5	92.8	13.9	4.8	4.4	0.4	246	13.9	5.6	44.5	58.4	14.3
PD00626	PD00626a	Pre-LN	Female	124	76	35.1	4.6	1.1	3.1	1	2	80.1	12.9	8.6	4.5	0.4	310	12	5.8	50.2	50.2	13.3
PD00646	PD00646b	Control	Female	116	65	23	4.6	2.1	2.3	0.6	1.5	93.9	13.3	5.3	4.5	0.4	160	14.4	5.7	64.6	64.6	19.4
PD00652	PD00652b	Control	Male	148	92	21	4.5	0.8	3.2	1.3	1.7	89.6	13.4	7.4	4.8	0.4	255	15.1	4.6	64.3	64.3	19.2
PD00656	PD00656b	Control	Female	104	58	22.4	6.3	0.8	5	1.2	1.9	99.1	15.4	5.2	4.2	0.4	194	12.9	4.8	62.8	62.8	17.6
PD00659	PD00659a	Pre-LN	Male	155	86	26.2	8.2	1.6	5.6	2.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	69	69	4.5
PD00659	PD00659b	Pre-LN	Male	169	89	26.3	5	0.9	3	2.5	26.6	86.9	15.6	33.4	4.6	0.4	104	13	7.2	69	73.4	4.5
PD00663	PD00663a	Pre-LN	Female	106	70	24.3	7.4	1.8	5.1	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	67.6	67.6	14.1
PD00664	PD00664a	Pre-LN	Male	155	80	27.4	4.6	1.1	2.5	2.2	2.1	89	13.3	6.4	4.1	0.4	153	13.1	8.6	75.9	75.9	3.1
PD00665	PD00665b	Control	Male	116	65	27.9	6.3	1.6	3.7	2.2	3	89.7	15.6	8.4	5.1	0.5	224	13.9	5.2	74.9	74.9	17.3
PD00673	PD00673a	Pre-LN	Male	128	76	27.6	6	1	4.5	1	2.5	89.4	13.4	6.3	5	0.4	208	15.4	NA	65.3	65.3	19.4
PD00690	PD00690a	Control	Male	130	80	30.3	5.7	0.8	3.6	2.9	1.9	90.8	12.7	6.1	4.7	0.4	246	14.6	5.7	54.9	54.9	21.6
PD00692	PD00692b	Control	Female	142	88	20.9	6.7	2.6	3.7	0.9	1.7	85.8	14.3	5.1	4.5	0.4	322	12.6	5.8	66.3	66.3	19.6
PD00695	PD00695a	Pre-LN	Male	154	90	28.3	4.2	1.4	2	1.8	NA	NA	NA	NA	NA	NA	NA	NA	NA	73.5	73.5	2.8
PD00703	PD00703b	Control	Female	118	74	26.2	5.7	2.3	3	0.9	2.8	91.6	14.7	7.6	4.2	0.4	215	12.7	5.8	62.9	62.9	19.6
PD00703	PD00703c	Control	Female	130	84	26.8	4.9	1.9	2.8	0.5	2	93.4	16.7	6.4	4	0.4	268	12.3	6.2	62.9	72.9	19.6
PD00708	PD00708a	Pre-LN	Male	132	82	31	6.2	1	3.6	3.4	1.7	83	15.2	5.6	4.7	0.4	294	14.5	NA	68.5	68.5	9.9
PD00712	PD00712b	Control	Female	144	93	28.3	5.7	1.4	3.5	1.8	1.9	88.4	13.3	6.6	4.8	0.4	229	13.8	5.3	62.1	62.1	19.5
PD00712	PD00712c	Control	Female	158	95	26.1	4.6	1.5	2.4	1.6	1.6	85	NA	7.4	4.7	0.4	215	13.9	5.2	62.1	70.5	19.5
PD00717	PD00717a	Control	Female	113	74	22.9	4.7	2	2.3	0.8	2.6	96.6	11.8	9	4.3	0.4	284	14.4	4.9	46.4	46.4	21.7
PD00717	PD00717c	Control	Female	126	76	22.8	5.9	2.1	3.1	1.6	2.4	95	NA	9.4	4	0.4	322	13.6	5.2	46.4	56.5	21.7
PD00721	PD00721a	Control	Male	123	83	23.8	6.2	1.7	3.6	2	2.4	89	12.8	6.3	4.3	0.4	267	13.2	5.5	49.8	49.8	21.2
PD00722	PD00722a	Pre-LN	Male	130	86	31	6.7	1.3	4.4	2.1	2	88.8	14	5	4.8	0.4	205	14.8	NA	54.4	54.4	7
PD00725	PD00725b	Control	Female	149	88	22.3	6.4	1.5	4.2	1.7	2.4	93.7	12.6	6.6	4.1	0.4	222	12.2	5.8	76.4	76.4	17.5
PD00726	PD00726b	Control	Male	124	62	23.1	6.1	1.9	3.8	1	1.9	89.1	12.9	7.2	4.9	0.4	279	15.2	7.1	64.8	64.8	14.7
PD00729	PD00729a	Pre-LN	Male	126	73	25.3	4.9	1.3	3	1.3	1.7	84.7	15.1	5.5	4.6	0.4	170	13.3	NA	74.9	74.9	5.1
PD00735	PD00735a	Pre-LN	Male	155	85	25.7	5.5	NA	NA	5.6	2	89.9	13.4	6.3	4.6	0.4	178	15.1	NA	62.7	62.7	16.4
PD00741	PD00741a	Pre-LN	Male	127	90	35	6.2	1	4.6	1.4	2.1	86.4	13.4	7.2	5.3	0.5	336	15.1	NA	49.8	49.8	15.9
PD00742	PD00742a	Pre-LN	Male	152	85	26.1	3.3	0.8	2.3	0.5	NA	NA	NA	NA	NA	NA	NA	NA	NA	62	62	2.8
PD00747	PD00747b	Control	Male	176	94	23.9	5.7	1.3	3.4	2.3	1.6	90	13.3	4.2	4.1	0.4	174	14.3	5.7	73.2	73.2	18.8
PD00747	PD00747c	Control	Male	139	69	23.9	4.4	1.2	2.4	1.8	1.4	92.1	13.4	5.9	3.9	0.4	206	12.2	6.1	73.2	81.5	18.8
PD00755	PD00755a	Control	Male	115	72	26.1	6.2	1.7	3.9	1.3	2.1	92.3	12.8	5.3	4.6	0.4	243	14.3	NA	74.8	74.8	22
PD00757	PD00757a	Pre-LN	Male	123	74	25.1	4.9	1.1	3.2	1.4	NA	NA	NA	NA	NA	NA	NA	NA	NA	58.8	58.8	14.7
PD00758	PD00758a	Control	Female	131	79	21.6	5.5	1.9	3.1	1.2	2.1	89.5	12.4	6.8	4	0.4	215	12.8	5	49.7	49.7	21.1
PD00759	PD00759b	Control	Male	143	76	25	5.5	1	4	1.2	1.8	90.5	13.8	5.7	4.5	0.4	147	14.4	4.8	76	76	17.9
PD00762	PD00762b	Control	Female	142	74	23.9	5.2	1.3	2.5	3.1	1.9	89.2	13.6	6.9	4.3	0.4	269	12.7	5.4	52.6	52.6	19.4
PD00762	PD00762c	Control	Female	131	61	23.8	4.5	1.6	2.3	1.4	1.7	92	NA	5.4	4.6	0.4	217	14	5.6	52.6	61.1	19.4
PD00767	PD00767b	Control	Male	142	89	22.7	5.8	1.4	3.9	1.2	2.7	89.7	12.8	6.4	4.2	0.4	201	13.2	5.8	50.8	50.8	19.5
PD00769	PD00769a	Pre-LN	Female	149	84	23.4	7.7	1.5	5.3	1.9	NA	NA	NA	NA	NA	NA	NA	NA	NA	75	75	4.6
PD00770	PD00770a	Control	Female	120	66	25.5	6.3	1.4	3.7	2.6	2.4	89.5	13.3	7.7	4	0.4	247	12.3	NA	63.7	63.7	22.7
PD00770	PD00770c	Control	Female	138	68	23.2	5.7	1.5	3.6	1.5	2.3	93.1	14.8	6.9	4.1	0.4	224	12.6	5.8	63.7	78.6	22.7
PD00779	PD00779b	Control	Female	148	87	30.7	6.5	1.6	3.9	2.4	3.2	87.7	12.2	7.5	4.5	0.4	256	13	6.2	62.5	62.5	19.8
PD00779	PD00779c	Control	Female	144	92	33.8	6.4	1.3	4	2.6	2.7	90.3	14.6	6.1	4	0.4	262	12	6.5	62.5	72.9	19.8
PD00789	PD00789a	Pre-LN	Male	126	72	24.1	4.9	1.4	2.8	1.5	2	82.5	13.4	9.6	5.2	0.4	113	14.1	5.6	76.2	76.2	2.4
PD00798	PD00798a	Pre-LN	Male	130	77	29.3	7.5	NA	NA	5	4.5	90.9	13.3	9.7	5.2	0.5	232	15.9	5.9	60	60	18.5
PD00805	PD00805b	Control	Female	128	76	26.1	5.8	1.3	3.6	2.1	2.3	89	12.3	6.4	3.9	0.3	239	12.2	5.3	69.1	69.1	18.4
PD00809	PD00809a	Control	Male	131	76	25	6.8	1.4	4.8	1.2	2.2	90.2	12.7	5.1	4.2	0.4	232	12.8	6	62	62	21.8
PD00810	PD00810c	Control	Male	120	62	23.9	3.8	1.2	2.3	0.8	3.1	89	15.6	11.5	4.7	0.4	209	14	7.4	69.6	69.6	3.3
PD00815	PD00815a	Pre-LN	Male	148	94	21.8	5.8	0.9	4.3	1.3	NA	NA	NA	NA	NA	NA	NA	NA	NA	72	72	4.8
PD00816	PD00816b	Control	Male	148	98	27.6	6.1	1.1	4.2	1.8	2.7	89.5	13.5	6.7	5	0.4	331	15.9	5	63	63	18
PD00818	PD00818a	Pre-LN	Female	134	72	27.9	6.3	1.4	4.3	1.4	2.8	86	14.3	8.9	4.3	0.4	255	12.5	NA	69.9	69.9	7.6
PD00830	PD00830c	Control	Female	92	58	27.6	6	1.4	4.1	1.3	2	93.9	13	7	4	0.4	331	12.5	6	64.9	64.9	7.2
PD00839	PD00839c	Control	Male	126	80	26.3	5.4	1.1	3.3	2.2	1.2	96.9	14.2	4.2	4.6	0.4	179	15.1	5.8	75	75	7.2
PD00840	PD00840c	Control	Male	153	76	21.1	5.9	1.7	3.6	1.4	1.4	91.7	15	4.9	4.5	0.4	246	13.6	5.9	69.2	69.2	6.5
PD00843	PD00843c	Control	Female	138	78	24.7	6.9	1.6	4.8	1.2	1.8	92.7	13.6	5	4.1	0.4	170	13.3	5.2	67.3	67.3	10.1
PD00845	PD00845c	Control	Female	142	78	54.4	4	1	2.5	1.3	2	88.1	16.2	8.1	4.5	0.4	115	13.1	7.4	72.9	72.9	4.1
PD00853	PD00853c	Control	Female	158	100	23.4	5	1.8	2.4	1.8	1.6	91.6	14.9	6.4	4.5	0.4	254	13.5	6.1	71	71	8.9
PD00855	PD00855c	Control	Female	137	74	22.6	5.3	1.2	3.8	0.7	1.9	89.6	13.2	5.6	4.2	0.4	188	12.9	5.5	71.6	71.6	8.8
PD00867	PD00867c	Control	Female	130	68	31.8	4.7	0.9	2.7	2.6	2.1	88.4	14.9	8.7	4.8	0.4	235	14.1	6	75.6	75.6	9.4
PD00868	PD00868c	Control	Male	128	78	26.2	4.1	1.4	2.2	1.3	1.4	95.5	14.3	6.2	4.8	0.5	274	15	6.1	63	63	7.9
PD00871	PD00871c	Control	Female	148	81	31.8	4.7	2	2.2	1.2	1.3	86.1	16.3	4.1	4.7	0.4	159	13.3	5.5	76.3	76.3	8.9
PD00879	PD00879c	Control	Male	161	88	39.4	5.9	1.2	3.5	2.7	2.8	89.8	14.1	10.6	5.1	0.5	308	14.8	5.9	68.8	68.8	10.5
PD00882	PD00882c	Control	Male	156	90	25.9	3.5	1.1	1.8	1.1	1.8	101.5	13.5	9.5	4.5	0.5	175	15.3	5.6	71.5	71.5	8.1
PD00888	PD00888c	Control	Male	127	70	26.2	5.5	2	3.1	0.9	1.4	95.7	14.2	5	4.7	0.5	280	15.1	5.8	61.5	61.5	6.8
PD00891	PD00891c	Control	Male	140	88	30.5	5.7	1.4	3.9	1	1.4	93.5	14.3	5.6	5	0.5	167	15.4	5.9	73.5	73.5	7.6
PD00905	PD00905c	Control	Male	126	78	25.8	6.1	1	3.1	4.4	4.1	91.2	13.9	8.8	5.2	0.5	181	16.2	5.9	66.7	66.7	7.8
PD00908	PD00908c	Control	Male	139	90	28	6.3	1.7	4.1	1.2	1.1	89.9	13.1	7.6	5.2	0.5	293	16.4	5.7	64.6	64.6	9.2
PD00911	PD00911c	Control	Female	124	66	27.4	4.6	1.4	2.6	1.4	1.7	91.6	13.3	6.4	4.3	0.4	434	13.5	5.9	73.9	73.9	10.1
PD00915	PD00915c	Control	Male	128	80	21.5	4.8	1.4	3.1	0.7	1	94.6	14.6	3.9	4.1	0.4	170	13.4	5.2	68.4	68.4	8.8
PD00918	PD00918c	Control	Female	166	94	24.8	7	2.2	4.3	1.3	2	90.4	15.4	8.2	5	0.4	242	14.7	5.1	70.1	70.1	7.6
PD00919	PD00919c	Control	Female	134	74	24.6	4.1	1.6	2.1	1	2.4	83.9	16	6.3	4.2	0.4	222	11.9	7.2	73.5	73.5	8
PD00923	PD00923c	Control	Male	124	69	26.9	3.6	1.6	1.7	0.8	3.2	92.7	14.9	7.1	4.5	0.4	241	13.4	6.2	70.6	70.6	8.8
PD00934	PD00934c	Control	Female	142	78	22.8	5.9	2.9	2.8	0.5	2.6	101.2	15	5.2	4.2	0.4	182	13.2	5.9	75	75	9.1
PD00935	PD00935c	Control	Male	138	74	24.2	3.5	1.5	1.5	1.1	1.4	93.3	13.7	6.6	4.4	0.4	172	13.6	6	75.2	75.2	8.9
PD00949	PD00949c	Control	Female	130	78	22.9	6.3	1.8	4.2	0.8	NA	89.3	13.8	5.8	5.1	0.5	296	15.5	5.4	54.6	54.6	9.6
PD00959	PD00959c	Control	Male	150	88	33.6	5.9	1.3	3.7	2.1	1.4	92.4	14.5	5.9	5	0.5	273	15.7	5.3	68.6	68.6	10.6
PD00967	PD00967c	Control	Female	137	74	22.6	5.3	1.2	3.8	0.7	1.9	89.6	13.2	5.6	4.2	0.4	188	12.9	5.5	71.6	71.6	8.8
PD00973	PD00973c	Control	Female	132	66	28.2	6.1	1.3	4.4	0.9	NA	93.7	12.7	3.7	4.2	0.4	182	13.2	5.5	66.6	66.6	8.1

Appendix 14: Driver mutations in pre-lymphoid neoplasm cases and controls

Cohort	Individual ID	Sample ID	Group	Type	Chromosome	Position	WT	MT	VAF	Gene	Protein	Effect
Discovery	PD00004	PD00004b	Case	sub	17	7577082	C	T	0.0231	TP53	p.E286K	Missense
Discovery	PD00017	PD00017b	Case	sub	2	25457242	C	T	0.0167	DNMT3A	p.R882H	Missense
Discovery	PD00035	PD00035b	Case	sub	4	106196794	T	A	0.16	TET2	p.C1709*	Nonsense
Discovery	PD00063	PD00063a	Case	sub	12	25378561	G	A	0.099	KRAS	p.A146V	Missense
Discovery	PD00089	PD00089b	Case	sub	11	108216546	G	T	0.12	ATM	p.R2832L	Missense
Discovery	PD00107	PD00107c	Case	sub	2	25457242	C	T	0.0069	DNMT3A	p.R882H	Missense
Discovery	PD00110	PD00110b	Case	indel	2	25459847	C	CATAA	0.0682	DNMT3A	p.K812fs*44	Frameshift
Discovery	PD00110	PD00110b	Case	sub	2	25467091	A	G	0.042	DNMT3A	p.L595P	Missense
Discovery	PD00179	PD00179b	Case	sub	1	115258747	C	G	0.0108	NRAS	p.G12A	Missense
Discovery	PD00179	PD00179b	Case	sub	4	106196551	T	G	0.22	TET2	p.Y1628*	Nonsense
Discovery	PD00179	PD00179b	Case	sub	7	140453136	A	T	0.0071	BRAF	p.V600E	Missense
Discovery	PD00185	PD00185b	Case	sub	2	25463289	T	C	0.0282	DNMT3A	p.Y735C	Missense
Discovery	PD00186	PD00186b	Case	indel	12	49434894	GC	G	0.0855	KMT2D	p.A2220fs*44	Frameshift
Discovery	PD00197	PD00197b	Case	sub	2	25457242	C	T	0.22	DNMT3A	p.R882H	Missense
Discovery	PD00197	PD00197b	Case	indel	4	106156452	AG	A	0.0213	TET2	p.E452fs*34	Frameshift
Discovery	PD00197	PD00197b	Case	indel	4	106197132	C	CA	0.132	TET2	p.N1823fs*1	Frameshift
Discovery	PD00199	PD00199b	Case	sub	21	44514780	C	T	0.0027	U2AF1	p.R156H	Missense
Discovery	PD00199	PD00199b	Case	indel	6	26156839	AG	A	0.0147	HIST1H1E	p.K75fs*14	Frameshift
Discovery	PD00200	PD00200b	Case	sub	2	25463286	C	T	0.0412	DNMT3A	p.R736H	Missense
Discovery	PD00226	PD00226b	Case	sub	2	25466790	G	C	0.097	DNMT3A	p.S638C	Missense
Discovery	PD00241	PD00241b	Case	sub	2	25458661	T	C	0.086	DNMT3A	p.N838D	Missense
Discovery	PD00241	PD00241b	Case	sub	2	25466800	G	A	0.0247	DNMT3A	p.R635W	Missense
Discovery	PD00254	PD00254b	Case	indel	11	108121593	CA	C	0.428	ATM	p.K468fs*5	Frameshift
Discovery	PD00273	PD00273b	Case	indel	2	25463206	C	CGTTA	0.04	DNMT3A	p.V763fs*1	Frameshift
Discovery	PD00282	PD00282c	Case	indel	11	108202611	CTCTAGAATT	C	0.3761	ATM	p.R2547_S2549delRIS	Inframe
Discovery	PD00285	PD00285a	Case	indel	17	58740541	GACTTT	G	0.0815	PPM1D	p.T483fs*4	Frameshift
Discovery	PD00297	PD00297b	Case	sub	2	61719472	C	T	0.0105	XPO1	p.E571K	Missense
Discovery	PD00301	PD00301b	Case	indel	4	106193849	G	GA	0.1179	TET2	p.R1440fs*38	Frameshift
Discovery	PD00310	PD00310c	Case	sub	7	140481417	C	A	0.0123	BRAF	p.G464V	Missense
Discovery	PD00330	PD00330c	Case	sub	2	25457209	C	G	0.0196	DNMT3A	p.W893S	Missense
Discovery	PD00330	PD00330c	Case	sub	7	124503682	T	C	0.11	POT1	p.K90E	Missense
Discovery	PD00330	PD00330c	Case	sub	9	139391843	G	C	0.076	NOTCH1	p.Y2116*	Nonsense
Discovery	PD00332	PD00332b	Case	sub	2	25463289	T	C	0.016	DNMT3A	p.Y735C	Missense
Discovery	PD00338	PD00338b	Case	sub	2	25457242	C	T	0.0136	DNMT3A	p.R882H	Missense
Discovery	PD00351	PD00351a	Case	sub	2	25467134	A	T	0.22	DNMT3A	p.W581R	Missense
Discovery	PD00455	PD00455b	Case	sub	4	106164829	T	G	0.0204	TET2	p.W1233G	Missense
Discovery	PD00561	PD00561b	Case	sub	2	25457242	C	T	0.0045	DNMT3A	p.R882H	Missense
Discovery	PD00588	PD00588b	Case	sub	17	7577120	C	T	0.0138	TP53	p.R273H	Missense
Discovery	PD00607	PD00607b	Case	sub	2	25466799	C	T	0.0121	DNMT3A	p.R635Q	Missense
Discovery	PD00666	PD00666b	Case	indel	2	25469976	GGT	G	0.1547	DNMT3A	p.H355fs*37	Frameshift
Discovery	PD00666	PD00666b	Case	indel	4	106193849	G	GA	0.0642	TET2	p.R1440fs*38	Frameshift
Discovery	PD00684	PD00684b	Case	sub	17	7578394	T	C	0.018	TP53	p.H179R	Missense
Discovery	PD00711	PD00711b	Case	sub	2	25467073	C	T	0.12	DNMT3A	p.W601*	Nonsense
Discovery	PD00711	PD00711b	Case	indel	2	25468894	ATGTTCCGG	A	0.0609	DNMT3A	p.R488fs*1	Frameshift
Discovery	PD00711	PD00711b	Case	indel	4	106194058	AG	A	0.0417	TET2	p.A1508fs*63	Frameshift
Discovery	PD00715	PD00715c	Case	indel	7	151882659	TC	T	0.041	KMT2C	p.E1689fs*28	Frameshift
Discovery	PD00719	PD00719c	Case	sub	11	108196083	A	T	0.047	ATM	p.K2207*	Nonsense
Discovery	PD00723	PD00723b	Case	sub	4	106196546	C	T	0.0215	TET2	p.Q1627*	Nonsense
Discovery	PD00764	PD00764b	Case	sub	2	25463289	T	C	0.0089	DNMT3A	p.Y735C	Missense
Discovery	PD00793	PD00793b	Case	sub	11	119149251	G	A	0.0137	CBL	p.R420Q	Missense
Discovery	PD00793	PD00793b	Case	sub	2	25470546	T	A	0.0304	DNMT3A	p.I310F	Missense
Discovery	PD00795	PD00795b	Case	sub	2	25468202	C	G	0.14	DNMT3A	p.?	Essential splice
Discovery	PD00820	PD00820b	Case	sub	17	74732959	G	A	0.0127	SRSF2	p.P95L	Missense
Discovery	PD00820	PD00820b	Case	sub	2	25463289	T	C	0.0037	DNMT3A	p.Y735C	Missense
Discovery	PD00021	PD00021a	Control	sub	2	25457243	G	A	0.0078	DNMT3A	p.R882C	Missense
Discovery	PD00068	PD00068a	Control	sub	12	25398284	C	G	0.0051	KRAS	p.G12A	Missense
Discovery	PD00068	PD00068a	Control	sub	2	25468935	T	A	0.045	DNMT3A	p.?	Essential splice
Discovery	PD00070	PD00070c	Control	sub	2	25457176	G	A	0.0125	DNMT3A	p.P904L	Missense
Discovery	PD00071	PD00071b	Control	sub	11	108186841	G	A	0.028	ATM	p.?	Essential splice
Discovery	PD00159	PD00159b	Control	sub	11	119148991	G	A	0.0181	CBL	p.C404Y	Missense
Discovery	PD00259	PD00259c	Control	sub	2	25463283	A	T	0.0304	DNMT3A	p.L737H	Missense
Discovery	PD00259	PD00259c	Control	indel	4	106156403	AC	A	0.0188	TET2	p.H436fs*11	Frameshift
Discovery	PD00385	PD00385c	Control	sub	4	106190898	C	G	0.038	TET2	p.S1392R	Missense
Discovery	PD00421	PD00421c	Control	sub	2	25463182	G	A	0.0077	DNMT3A	p.R771*	Nonsense
Discovery	PD00431	PD00431b	Control	sub	2	25463234	C	T	0.049	DNMT3A	p.W753*	Nonsense
Discovery	PD00465	PD00465b	Control	sub	2	25463566	C	T	0.0689	DNMT3A	p.G706R	Missense
Discovery	PD00571	PD00571c	Control	sub	2	25467478	T	C	0.0095	DNMT3A	p.Y533C	Missense
Discovery	PD00651	PD00651b	Control	sub	2	25457176	G	A	0.0169	DNMT3A	p.P904L	Missense
Discovery	PD00683	PD00683a	Control	sub	2	25463289	T	C	0.0736	DNMT3A	p.Y735C	Missense
Discovery	PD00688	PD00688a	Control	sub	2	25463289	T	C	0.0149	DNMT3A	p.Y735C	Missense
Discovery	PD00745	PD00745a	Control	sub	2	25457242	C	T	0.0108	DNMT3A	p.R882H	Missense
Discovery	PD00751	PD00751a	Control	sub	2	25467467	A	G	0.0109	DNMT3A	p.C537R	Missense
Discovery	PD00776	PD00776c	Control	sub	2	25463601	T	C	0.0378	DNMT3A	p.?	Essential splice
Discovery	PD00895	PD00895c	Control	sub	11	119148919	T	C	0.0057	CBL	p.L380P	Missense
Discovery	PD00928	PD00928c	Control	indel	2	25469539	GC	G	0.033	DNMT3A	p.A410fs*241	Frameshift
Discovery	PD00930	PD00930c	Control	sub	2	25470575	A	C	0.0547	DNMT3A	p.L300R	Missense
Discovery	PD00930	PD00930c	Control	sub	4	106158563	T	C	0.031	TET2	p.L1155S	Missense
Extension	PD00027	PD00027a	NA	sub	2	25463586	C	T	0.21	DNMT3A	p.G699D	Missense
Extension	PD00039	PD00039b	NA	sub	2	25457243	G	T	0.011	DNMT3A	p.R882S	Missense
Extension	PD00050	PD00050b	NA	sub	2	25467448	C	G	0.11	DNMT3A	p.G543A	Missense
Extension	PD00117	PD00117c	NA	sub	20	31024770	A	T	0.0168	ASXL1	p.K1419*	Nonsense
Extension	PD00122	PD00122b	NA	indel	4	106180853	AC	A	0.0138	TET2	p.Y1295fs*68	Frameshift
Extension	PD00161	PD00161c	NA	sub	4	106196491	T	A	0.0312	TET2	p.Y1608*	Nonsense
Extension	PD00165	PD00165c	NA	sub	2	25462018	T	C	0.18	DNMT3A	p.N797D	Missense
Extension	PD00165	PD00165c	NA	sub	2	25466796	A	C	0.0139	DNMT3A	p.V636G	Missense
Extension	PD00170	PD00170c	NA	sub	2	25457242	C	T	0.0301	DNMT3A	p.R882H	Missense
Extension	PD00180	PD00180c	NA	sub	2	25457243	G	A	0.0876	DNMT3A	p.R882C	Missense

Extension	PD00307	PD00307c	NA	indel	17	58740653	CA	C	0.3164	PPM1D	p.M521fs*1	Frameshift
Extension	PD00398	PD00398c	NA	sub	17	74732959	G	A	0.0354	SRSF2	p.P95L	Missense
Extension	PD00398	PD00398c	NA	sub	2	25467448	C	T	0.0061	DNMT3A	p.G543D	Missense
Extension	PD00418	PD00418c	NA	sub	2	25462075	C	T	0.0272	DNMT3A	p.V778M	Missense
Extension	PD00462	PD00462a	NA	sub	2	25457242	C	T	0.0064	DNMT3A	p.R882H	Missense
Extension	PD00470	PD00470c	NA	indel	20	31022951	TC	T	0.0306	ASXL1	p.I814fs*4	Frameshift
Extension	PD00537	PD00537c	NA	sub	2	25470583	C	A	0.3	DNMT3A	p.W297C	Missense
Extension	PD00540	PD00540c	NA	sub	4	106196823	G	A	0.0133	TET2	p.G1719E	Missense
Extension	PD00592	PD00592c	NA	sub	2	25463182	G	A	0.0254	DNMT3A	p.R771*	Nonsense
Extension	PD00605	PD00605c	NA	indel	17	58740684	CT	C	0.1781	PPM1D	p.P531fs*8	Frameshift
Extension	PD00636	PD00636b	NA	indel	2	25469967	GCCTGGTGGAAC	A	0.0613	DNMT3A	p.S352fs*48	Frameshift
Extension	PD00648	PD00648c	NA	indel	20	31021175	TC	T	0.0169	ASXL1	p.S392fs*1	Frameshift
Extension	PD00655	PD00655b	NA	indel	2	25466846	AG	A	0.0604	DNMT3A	p.P619fs*32	Frameshift
Extension	PD00671	PD00671a	NA	sub	2	25467497	G	A	0.0476	DNMT3A	p.Q527*	Nonsense
Extension	PD00718	PD00718c	NA	indel	2	25463566	CA	C	0.0739	DNMT3A	p.I705fs*74	Frameshift
Extension	PD00732	PD00732b	NA	sub	2	25457242	C	T	0.0139	DNMT3A	p.R882H	Missense
Extension	PD00734	PD00734c	NA	sub	11	119149280	G	A	0.1	CBL	p.V430M	Missense
Extension	PD00736	PD00736a	NA	sub	17	29562934	A	G	0.0305	NF1	p.?	Essential splice
Extension	PD00736	PD00736a	NA	sub	9	5073770	G	T	0.0338	JAK2	p.V617F	Missense
Extension	PD00740	PD00740c	NA	sub	4	106180868	A	G	0.14	TET2	p.K1299R	Missense
Extension	PD00748	PD00748c	NA	sub	2	25457242	C	T	0.0078	DNMT3A	p.R882H	Missense
Extension	PD00748	PD00748c	NA	indel	2	25467039	G	GT	0.0656	DNMT3A	p.N612fs*7	Frameshift
Extension	PD00772	PD00772c	NA	sub	2	25466852	C	T	0.0494	DNMT3A	p.?	Essential splice
Extension	PD00784	PD00784c	NA	sub	4	106197374	C	T	0.048	TET2	p.Q1903*	Nonsense
Extension	PD00807	PD00807b	NA	sub	21	44524456	G	A	0.0107	U2AF1	p.S34F	Missense
Extension	PD00828	PD00828c	NA	sub	12	25398285	C	T	0.0118	KRAS	p.G12S	Missense
Extension	PD00828	PD00828c	NA	sub	2	25467484	T	C	0.0251	DNMT3A	p.D531G	Missense
Extension	PD00832	PD00832c	NA	sub	2	25463170	C	T	0.0071	DNMT3A	p.?	Essential splice
Extension	PD00832	PD00832c	NA	sub	2	25470579	T	A	0.0129	DNMT3A	p.K299*	Nonsense
Extension	PD00834	PD00834c	NA	sub	2	25457243	G	T	0.014	DNMT3A	p.R882S	Missense
Extension	PD00837	PD00837c	NA	indel	17	58740653	CA	C	0.1576	PPM1D	p.M521fs*1	Frameshift
Extension	PD00850	PD00850c	NA	sub	X	129148664	G	T	0.0496	BCORL1	p.R639L	Missense
Extension	PD00858	PD00858c	NA	sub	2	25463289	T	C	0.0109	DNMT3A	p.Y735C	Missense
Extension	PD00863	PD00863c	NA	sub	2	198267359	C	A	0.0067	SF3B1	p.K666N	Missense
Extension	PD00869	PD00869c	NA	indel	4	106156933	TGGGGGGCTCC	C	0.0425	TET2	p.P612fs*21	Frameshift
Extension	PD00872	PD00872c	NA	sub	21	44524456	G	A	0.0065	U2AF1	p.S34F	Missense
Extension	PD00884	PD00884c	NA	sub	2	25463182	G	A	0.0146	DNMT3A	p.R771*	Nonsense
Extension	PD00885	PD00885c	NA	sub	4	106193977	C	G	0.0141	TET2	p.S1480C	Missense
Extension	PD00887	PD00887c	NA	indel	2	25471082	CA	C	0.0536	DNMT3A	p.V227fs*89	Frameshift
Extension	PD00900	PD00900c	NA	sub	20	31021295	C	T	0.092	ASXL1	p.Q432*	Nonsense
Extension	PD00913	PD00913c	NA	sub	2	25457163	A	C	0.0321	DNMT3A	p.Y908*	Nonsense
Extension	PD00927	PD00927c	NA	indel	2	25505536	CCACCTGCAAATG	C	0.0879	DNMT3A	p.?	Essential splice
Extension	PD00943	PD00943c	NA	sub	2	25466800	G	A	0.0105	DNMT3A	p.R635W	Missense
Extension	PD00957	PD00957c	NA	sub	10	112333508	G	T	0.0601	SMC3	p.?	Essential splice
Extension	PD00957	PD00957c	NA	sub	4	55604646	G	C	0.0237	KIT	p.D952H	Missense
Extension	PD00968	PD00968c	NA	sub	2	25457242	C	T	0.011	DNMT3A	p.R882H	Missense
Extension	PD00969	PD00969c	NA	sub	2	25463182	G	A	0.0201	DNMT3A	p.R771*	Nonsense
Extension	PD00970	PD00970c	NA	sub	2	25457209	C	T	0.0277	DNMT3A	p.W893*	Nonsense
Extension	PD00972	PD00972c	NA	sub	2	25457242	C	T	0.0078	DNMT3A	p.R882H	Missense
Validation	PD00008	PD00008a	Case	sub	2	25457243	G	A	0.0319	DNMT3A	p.R882C	Missense
Validation	PD00012	PD00012a	Case	sub	6	41903706	G	C	0.13	CCND3	p.P284R	Missense
Validation	PD00028	PD00028a	Case	sub	2	25457243	G	A	0.13	DNMT3A	p.R882C	Missense
Validation	PD00052	PD00052a	Case	sub	17	7578190	T	C	0.055	TP53	p.Y220C	Missense
Validation	PD00053	PD00053a	Case	indel	1	120458435	T	TG	0.0534	NOTCH2	p.I2304fs*9	Frameshift
Validation	PD00105	PD00105a	Case	sub	2	25457176	G	A	0.057	DNMT3A	p.P904L	Missense
Validation	PD00133	PD00133a	Case	indel	2	25468914	CA	C	0.0154	DNMT3A	p.V483fs*168	Frameshift
Validation	PD00169	PD00169a	Case	sub	4	106196580	C	G	0.0132	TET2	p.S1638*	Nonsense
Validation	PD00181	PD00181a	Case	sub	2	25463182	G	A	0.068	DNMT3A	p.R771*	Nonsense
Validation	PD00290	PD00290a	Case	sub	9	5073770	G	T	0.0218	JAK2	p.V617F	Missense
Validation	PD00315	PD00315a	Case	sub	2	198266834	T	C	0.0213	SF3B1	p.K700E	Missense
Validation	PD00375	PD00375a	Case	sub	1	36937219	C	T	0.419	CSF3R	p.R367Q	Missense
Validation	PD00435	PD00435a	Case	sub	4	106164778	C	T	0.0269	TET2	p.R1216*	Nonsense
Validation	PD00461	PD00461a	Case	sub	17	40474482	T	A	0.035	STAT3	p.Y640F	Missense
Validation	PD00541	PD00541a	Case	sub	2	25457176	G	A	0.0072	DNMT3A	p.P904L	Missense
Validation	PD00546	PD00546a	Case	sub	4	106197318	C	T	0.093	TET2	p.T1884I	Missense
Validation	PD00547	PD00547a	Case	sub	2	61719471	T	C	0.0285	XPO1	p.E571G	Missense
Validation	PD00600	PD00600a	Case	sub	2	25463289	T	C	0.0077	DNMT3A	p.Y735C	Missense
Validation	PD00620	PD00620a	Case	sub	X	39921510	G	C	0.053	BCOR	p.S1437*	Nonsense
Validation	PD00659	PD00659a	Case	sub	2	25457242	C	T	0.11	DNMT3A	p.R882H	Missense
Validation	PD00664	PD00664a	Case	sub	9	5073770	G	T	0.0123	JAK2	p.V617F	Missense
Validation	PD00695	PD00695a	Case	indel	2	25463554	AG	A	0.0669	DNMT3A	p.C710fs*69	Frameshift
Validation	PD00708	PD00708a	Case	sub	4	106180865	G	A	0.045	TET2	p.C1298Y	Missense
Validation	PD00769	PD00769a	Case	sub	17	7574003	G	A	0.11	TP53	p.R342*	Nonsense
Validation	PD00769	PD00769a	Case	indel	9	139390648	CAG	C	0.0696	NOTCH1	p.P2514fs*4	Frameshift
Validation	PD00789	PD00789a	Case	indel	2	25464532	TG	T	0.0306	DNMT3A	p.Y660fs*1	Frameshift
Validation	PD00798	PD00798a	Case	sub	4	106197377	C	T	0.056	TET2	p.H1904Y	Missense
Validation	PD00815	PD00815a	Case	sub	11	108216597	G	C	0.044	ATM	p.R2849P	Missense
Validation	PD00255	PD00255b	Control	sub	2	25462077	G	A	0.0181	DNMT3A	p.P777L	Missense
Validation	PD00261	PD00261b	Control	sub	2	25457243	G	T	0.0065	DNMT3A	p.R882S	Missense
Validation	PD00333	PD00333c	Control	sub	17	7578268	A	C	0.078	TP53	p.L194R	Missense
Validation	PD00387	PD00387a	Control	sub	2	25457242	C	T	0.0133	DNMT3A	p.R882H	Missense
Validation	PD00408	PD00408b	Control	sub	4	106180794	G	C	0.0283	TET2	p.Q1274H	Missense
Validation	PD00471	PD00471a	Control	sub	2	25467449	C	A	0.0205	DNMT3A	p.G543C	Missense
Validation	PD00471	PD00471a	Control	sub	4	106180931	G	A	0.095	TET2	p.?	Essential splice
Validation	PD00476	PD00476a	Control	sub	7	140453155	C	T	0.0027	BRAF	p.D594N	Missense
Validation	PD00554	PD00554a	Control	sub	17	7577117	A	T	0.0145	TP53	p.V274D	Missense
Validation	PD00566	PD00566c	Control	indel	9	21974794	CATGCTGCTCCC	A	0.0743	CDKN2A	p.A4_P11delAAGSSMEP	Inframe
Validation	PD00597	PD00597a	Control	indel	2	25505536	CCACCTGCAAATG	C	0.0413	DNMT3A	p.?	Essential splice
Validation	PD00809	PD00809a	Control	sub	2	25457209	C	T	0.0143	DNMT3A	p.W893*	Nonsense
Validation	PD00809	PD00809a	Control	sub	4	55593639	G	T	0.0059	KIT	p.V569F	Missense
Validation	PD00810	PD00810c	Control	sub	20	31022592	C	T	0.0087	ASXL1	p.R693*	Nonsense
Validation	PD00810	PD00810c	Control	indel	4	106155537	TA	T	0.0281	TET2	p.K147fs*5	Frameshift
Validation	PD00810	PD00810c	Control	indel	4	106197288	AG	A	0.0389	TET2	p.E1874fs*13	Frameshift

Validation	PD00830	PD00830c	Control	sub	2	25463170	C	T	0.0075	DNMT3A	p.?	Essential splice
Validation	PD00911	PD00911c	Control	sub	20	31021295	C	T	0.11	ASXL1	p.Q432*	Nonsense
Validation	PD00918	PD00918c	Control	sub	2	25466800	G	A	0.0102	DNMT3A	p.R635W	Missense
Serial sample	PD00003	PD00003b	NA	sub	12	25398281	C	T	0.0104	KRAS	p.G13D	Missense
Serial sample	PD00004	PD00004a	NA	sub	17	7577082	C	T	0.0143	TP53	p.E286K	Missense
Serial sample	PD00012	PD00012b	NA	sub	6	41903706	G	C	0.27	CCND3	p.P284R	Missense
Serial sample	PD00035	PD00035a	NA	sub	4	106196794	T	A	0.083	TET2	p.C1709*	Nonsense
Serial sample	PD00068	PD00068c	NA	sub	2	25468935	T	A	0.075	DNMT3A	p.?	Essential splice
Serial sample	PD00107	PD00107b	NA	sub	2	25457242	C	T	0.0083	DNMT3A	p.R882H	Missense
Serial sample	PD00166	PD00166c	NA	sub	2	25469632	C	T	0.0271	DNMT3A	p.R379H	Missense
Serial sample	PD00181	PD00181b	NA	sub	2	25463182	G	A	0.0443	DNMT3A	p.R771*	Nonsense
Serial sample	PD00186	PD00186a	NA	indel	12	49434894	GC	G	0.1189	KMT2D	p.A2220fs*44	Frameshift
Serial sample	PD00199	PD00199a	NA	sub	21	44514780	C	T	0.0087	U2AF1	p.R156H	Missense
Serial sample	PD00200	PD00200a	NA	sub	2	25463286	C	T	0.0316	DNMT3A	p.R736H	Missense
Serial sample	PD00226	PD00226a	NA	sub	2	25466790	G	C	0.078	DNMT3A	p.S638C	Missense
Serial sample	PD00241	PD00241c	NA	indel	17	58740401	A	AT	0.0768	PPM1D	p.P437fs*6	Frameshift
Serial sample	PD00241	PD00241c	NA	sub	2	25458661	T	C	0.15	DNMT3A	p.N838D	Missense
Serial sample	PD00241	PD00241c	NA	sub	2	25466800	G	A	0.0347	DNMT3A	p.R635W	Missense
Serial sample	PD00282	PD00282b	NA	indel	11	108202611	CTCTAGAATT	C	0.3809	ATM	p.R2547_S2549delRIS	Inframe
Serial sample	PD00310	PD00310a	NA	sub	7	140481417	C	A	0.0035	BRAF	p.G464V	Missense
Serial sample	PD00310	PD00310b	NA	sub	7	140481417	C	A	0.0077	BRAF	p.G464V	Missense
Serial sample	PD00315	PD00315b	NA	sub	11	108117757	T	G	0.0512	ATM	p.I323R	Missense
Serial sample	PD00315	PD00315b	NA	sub	11	108203543	C	T	0.0649	ATM	p.Q2615*	Nonsense
Serial sample	PD00315	PD00315b	NA	sub	2	61719471	T	A	0.0128	XPO1	p.E571V	Missense
Serial sample	PD00315	PD00315b	NA	sub	2	198266834	T	C	0.23	SF3B1	p.K700E	Missense
Serial sample	PD00330	PD00330b	NA	sub	2	25457209	C	G	0.0135	DNMT3A	p.W893S	Missense
Serial sample	PD00332	PD00332a	NA	sub	2	25463289	T	C	0.0038	DNMT3A	p.Y735C	Missense
Serial sample	PD00471	PD00471c	NA	sub	2	25467449	C	A	0.0071	DNMT3A	p.G543C	Missense
Serial sample	PD00471	PD00471c	NA	sub	4	106180931	G	A	0.22	TET2	p.?	Essential splice
Serial sample	PD00476	PD00476c	NA	sub	17	7577538	C	G	0.19	TP53	p.R248P	Missense
Serial sample	PD00476	PD00476c	NA	sub	6	41903688	A	G	0.21	CCND3	p.I290T	Missense
Serial sample	PD00476	PD00476c	NA	sub	7	140453155	C	T	0.24	BRAF	p.D594N	Missense
Serial sample	PD00561	PD00561c	NA	sub	2	25457242	C	T	0.11	DNMT3A	p.R882H	Missense
Serial sample	PD00659	PD00659b	NA	indel	16	3781420	TG	T	0.2509	CREBBP	p.I1649fs*95	Frameshift
Serial sample	PD00659	PD00659b	NA	sub	2	25457242	C	T	0.055	DNMT3A	p.R882H	Missense
Serial sample	PD00659	PD00659b	NA	sub	6	41903710	T	C	0.078	CCND3	p.T283A	Missense
Serial sample	PD00666	PD00666a	NA	indel	2	25469976	GGT	G	0.1158	DNMT3A	p.H355fs*37	Frameshift
Serial sample	PD00666	PD00666c	NA	indel	2	25469976	GGT	G	0.0549	DNMT3A	p.H355fs*37	Frameshift
Serial sample	PD00666	PD00666c	NA	sub	2	198266834	T	C	0.31	SF3B1	p.K700E	Missense
Serial sample	PD00666	PD00666c	NA	indel	4	106193849	G	GA	0.0465	TET2	p.R1440fs*38	Frameshift
Serial sample	PD00793	PD00793c	NA	sub	11	119149251	G	A	0.0274	CBL	p.R420Q	Missense
Serial sample	PD00793	PD00793c	NA	sub	2	25470546	T	A	0.1	DNMT3A	p.I310F	Missense
Serial sample	PD00795	PD00795c	NA	sub	2	25468202	C	G	0.069	DNMT3A	p.?	Essential splice
Serial sample	PD00820	PD00820a	NA	sub	17	74732959	G	A	0.0069	SRSF2	p.P95L	Missense
Serial sample	PD00820	PD00820a	NA	sub	2	25463289	T	C	0.0097	DNMT3A	p.Y735C	Missense

Appendix 15: Lymphoid neoplasm risk prediction model coefficients

Cox proportional hazards model trained on the discovery cohort

Variable	Coefficient	P value	Adjusted P value
ATM	0.946	5.45E-09	1.25E-07
BRAF	2.996	8.01E-19	1.84E-17
CBL	2.341	1.57E-04	3.61E-03
DNMT3A	0.861	2.26E-03	5.20E-02
KMT2D	3.691	3.15E-05	7.23E-04
KRAS	3.621	3.45E-05	7.93E-04
SRSF2	2.962	4.94E-21	1.14E-19
TET2	2.408	8.83E-12	2.03E-10
TP53	3.982	1.16E-29	2.68E-28
U2AF1	2.718	2.16E-18	4.97E-17
TC	-0.222	1.51E-04	3.48E-03
Diastolic BP	0.129	2.17E-01	1.00E+00
HbA1c	-0.037	6.62E-01	1.00E+00
HDL	-0.475	2.91E-03	6.69E-02
LDL	-0.047	5.66E-01	1.00E+00
LYM	0.355	5.14E-03	1.18E-01
MCV	0.131	2.15E-02	4.94E-01
RBC	-0.301	6.28E-02	1.00E+00
RDW	-0.243	9.85E-02	1.00E+00
Systolic BP	-0.094	5.48E-01	1.00E+00
WBC	0.144	2.83E-01	1.00E+00
Gender	-0.323	1.79E-02	4.12E-01
Age	0.086	3.86E-01	1.00E+00

Cox proportional hazards model trained on validation cohort

Variable	Coefficient	P value	Adjusted P value
ASXL1	0.472	4.49E-01	1.00E+00
DNMT3A	2.214	8.38E-05	1.51E-03
JAK2	1.651	1.68E-04	3.03E-03
TET2	0.857	1.01E-01	1.00E+00
TP53	1.642	6.90E-03	1.24E-01
TC	-0.092	3.27E-01	1.00E+00
Diastolic BP	0.031	5.45E-01	1.00E+00
HbA1c	-0.044	7.17E-01	1.00E+00
HDL	-0.284	1.47E-02	2.65E-01
LDL	0.165	9.53E-02	1.00E+00
LYM	0.097	4.28E-01	1.00E+00
MCV	-0.084	4.57E-05	8.23E-04
RBC	-0.031	7.96E-01	1.00E+00
RDW	0.042	3.70E-01	1.00E+00
Systolic BP	0.180	1.97E-02	3.55E-01
WBC	0.151	4.03E-02	7.26E-01
Gender	0.143	2.13E-01	1.00E+00
Age	0.078	5.03E-01	1.00E+00

TC, total cholesterol; BP, blood pressure; HDL, high-density lipoprotein; LDL, low density lipoprotein; LYM, lymphocytes; width; WBC, white blood cells MCV, mean corpuscular volume; RBC, red cell distribution

Cox proportional hazards model trained on combined cohort

Variable	Coefficient	P value	Adjusted P value
ASXL1	0.362	6.32E-01	1.00E+00
ATM	0.951	1.33E-09	3.72E-08
BRAF	2.639	2.31E-18	6.46E-17
CBL	1.995	3.99E-04	1.12E-02
DNMT3A	1.192	8.74E-06	2.45E-04
JAK2	3.112	1.26E-28	3.53E-27
KMT2D	3.315	6.36E-05	1.78E-03
KRAS	3.579	1.59E-05	4.46E-04
NOTCH1	3.747	1.71E-06	4.78E-05
SRSF2	2.550	1.39E-17	3.90E-16
TET2	1.700	1.23E-06	3.43E-05
TP53	1.888	2.25E-03	6.31E-02
U2AF1	2.392	3.12E-18	8.74E-17
XPO1	3.228	1.92E-31	5.38E-30
TC	-0.198	1.12E-03	3.14E-02
Diastolic BP	0.120	2.51E-01	1.00E+00
HbA1c	-0.047	5.71E-01	1.00E+00
HDL	-0.544	9.67E-05	2.71E-03
LDL	0.035	6.04E-01	1.00E+00
LYM	0.258	1.61E-02	4.52E-01
MCV	0.002	9.72E-01	1.00E+00
RBC	-0.283	4.19E-02	1.00E+00
RDW	-0.150	2.90E-01	1.00E+00
Systolic BP	0.162	2.76E-01	1.00E+00
WBC	0.286	5.14E-02	1.00E+00
Gender	-0.142	1.90E-01	1.00E+00
Age	0.115	1.58E-01	1.00E+00

Appendix 16

First and joint first author primary research publications

1) Abelson, S., Collord G., et al. (2018). "Prediction of acute myeloid leukaemia risk in healthy individuals." *Nature* 559 (7714): 400-404. [PMID: 29988082]

2) Collord, G, et al. (2018). "An integrated genomic analysis of anaplastic meningioma identifies prognostic molecular signatures." *Sci Rep* 8(1): 13537. [PMID: 30202034]

3) Caesar R, Collord G, et al. (2018). "Targeting MEK in vemurafenib-resistant hairy cell leukemia." *Leukemia*. [PMID: 30341394]

4) Wegert J, Vokuhl C, Collord G, et al. (2018). "Recurrent intragenic rearrangements of EGFR and BRAF in soft tissue tumors of infants." *Nat Commun* 9(1): 2378. [PMID: 29915264]

5) Collord G, et al. (2018). "Recurrent histone mutations in T-cell acute lymphoblastic leukaemia." *Br J Haematol*. [PMID: 29602208]

6) Collord G, et al. (2017). "Clonal haematopoiesis is not prevalent in survivors of childhood cancer." *Br J Haematol.* [PMID: 28369776]

LETTER

https://doi.org/10.1038/s41586-018-0317-6

Prediction of acute myeloid leukaemia risk in healthy individuals

Sagi Abelson[1,46], Grace Collord[2,3,46], Stanley W. K. Ng[4], Omer Weissbrod[5], Netta Mendelson Cohen[5], Elisabeth Niemeyer[6], Noam Barda[7], Philip C. Zuzarte[8], Lawrence Heisler[8], Yogi Sundaravadanam[8], Robert Luben[9], Shabina Hayat[9], Ting Ting Wang[1,10], Zhen Zhao[1], Iulia Cirlan[1], Trevor J. Pugh[1,8,10], David Soave[8], Karen Ng[8], Calli Latimer[2], Claire Hardy[2], Keiran Raine[2], David Jones[2], Diana Hoult[11], Abigail Britten[11], John D. McPherson[8], Mattias Johansson[12], Faridah Mbabaali[8], Jenna Eagles[8], Jessica K. Miller[8], Danielle Pasternack[8], Lee Timms[8], Paul Krzyzanowski[8], Philip Awadalla[8], Rui Costa[13], Eran Segal[5], Scott V. Bratman[1,8,14], Philip Beer[2], Sam Behjati[2,3], Inigo Martincorena[2], Jean C. Y. Wang[1,15,16], Kristian M. Bowles[17,18], J. Ramón Quirós[19], Anna Karakatsani[20,21], Carlo La Vecchia[20,22], Antonia Trichopoulou[20], Elena Salamanca-Fernández[23,24], José M. Huerta[24,25], Aurelio Barricarte[24,26,27], Ruth C. Travis[28], Rosario Tumino[29], Giovanna Masala[30], Heiner Boeing[31], Salvatore Panico[32], Rudolf Kaaks[33], Alwin Krämer[34], Sabina Sieri[35], Elio Riboli[36], Paolo Vineis[36], Matthieu Foll[12], James McKay[12], Silvia Polidoro[37], Núria Sala[38], Kay-Tee Khaw[39], Roel Vermeulen[40], Peter J. Campbell[2,41], Elli Papaemmanuil[2,42], Mark D. Minden[1,10,15,16], Amos Tanay[5], Ran D. Balicer[7], Nicholas J. Wareham[11], Moritz Gerstung[2,13,47]*, John E. Dick[1,43,47]*, Paul Brennan[12,47]*, George S. Vassiliou[2,41,44,47]* & Liran I. Shlush[1,6,45,47]*

The incidence of acute myeloid leukaemia (AML) increases with age and mortality exceeds 90% when diagnosed after age 65. Most cases arise without any detectable early symptoms and patients usually present with the acute complications of bone marrow failure[1]. The onset of such de novo AML cases is typically preceded by the accumulation of somatic mutations in preleukaemic haematopoietic stem and progenitor cells (HSPCs) that undergo clonal expansion[2,3]. However, recurrent AML mutations also accumulate in HSPCs during ageing of healthy individuals who do not develop AML, a phenomenon referred to as age-related clonal haematopoiesis (ARCH)[4–8]. Here we use deep sequencing to analyse genes that are recurrently mutated in AML to distinguish between individuals who have a high risk of developing AML and those with benign ARCH. We analysed peripheral blood cells from 95 individuals that were obtained on average 6.3 years before AML diagnosis (pre-AML group), together with 414 unselected age- and gender-matched individuals (control group). Pre-AML cases were distinct from controls and had more mutations per sample, higher variant allele frequencies, indicating greater clonal expansion, and showed enrichment of mutations in specific genes. Genetic parameters were used to derive a model that accurately predicted AML-free survival; this model was validated in an independent cohort of 29 pre-AML cases and 262 controls. Because AML is rare, we also developed an AML predictive model using a large electronic health record database that identified individuals at greater risk. Collectively our findings provide proof-of-concept that it is possible to discriminate ARCH from pre-AML many years before malignant transformation. This could in future enable earlier detection and monitoring, and may help to inform intervention.

To examine the occurrence of somatic mutations before the development of AML, we carried out deep error-corrected targeted sequencing of AML-associated genes in a discovery cohort of 95 pre-AML cases and 414 age- and gender-matched controls (Supplementary Table 1). A validation cohort comprising 29 pre-AML cases and 262 controls (Supplementary Table 1) was analysed using deep sequencing with an overlapping gene panel. Taking both cohorts together, ARCH, defined on the basis of putative driver mutations (ARCH-PD), was found in 73.4% of the pre-AML cases at a median of 7.6 years before diagnosis. By contrast, ARCH-PD was observed in 36.7% of controls ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Fig. 1a), consistent with data from a study of more than 2,000 unselected individuals assayed using a similarly sensitive method[9,10]. Additionally, 39% of pre-AML cases above the age of 50 had a driver mutation with a variant allele frequency (VAF) of more than 10%, compared to only 4% of controls,

[1]Princess Margaret Cancer Centre, University Health Network (UHN), Toronto, Ontario, Canada. [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [3]Department of Paediatrics, University of Cambridge, Cambridge, UK. [4]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada. [5]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [6]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [7]Clalit Research Institute, Tel Aviv, Israel. [8]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [9]Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge School of Clinical Medicine, Cambridge, UK. [10]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [11]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [12]International Agency for Research on Cancer, World Health Organization, Lyon, France. [13]European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Wellcome Genome Campus, Hinxton, UK. [14]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [15]Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [16]Division of Medical Oncology and Hematology, University Health Network, Toronto, Ontario, Canada. [17]Department of Molecular Haematology, Norwich Medical School, The University of East Anglia, Norwich, UK. [18]Department of Haematology, Norfolk and Norwich University Hospitals NHS Trust, Norwich, UK. [19]Public Health Directorate, Asturias, Spain. [20]Hellenic Health Foundation, Athens, Greece. [21]2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Athens, Greece. [22]Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy. [23]Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibs.GRANADA, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain. [24]CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain. [25]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [26]Navarra Public Health Institute, Pamplona, Spain. [27]Navarra Institute for Health Research, Pamplona, Spain. [28]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [29]Cancer Registry and Histopathology Department, Civic-M. P. Arezzo Hospital, Azienda Sanitaria Provinciale, Ragusa, Italy. [30]Cancer Risk Factors and Life-Style Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy. [31]Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany. [32]Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy. [33]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [34]Clinical Cooperation Unit Molecular Hematology/Oncology, German Cancer Research Center (DKFZ) and Department of Internal Medicine V, University of Heidelberg, Heidelberg, Germany. [35]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy. [36]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [37]Italian Institute for Genomic Medicine, Torino, Italy. [38]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program and Translational Research Laboratory, Catalan Institute of Oncology, ICO-IDIBELL, Barcelona, Spain. [39]University of Cambridge, Cambridge, UK. [40]Division of Environmental Epidemiology and Veterinary Public Health, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands. [41]Department of Haematology, University of Cambridge, Cambridge, UK. [42]Center for Molecular Oncology and Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [43]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [44]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [45]Division of Hematology, Rambam Healthcare Campus, Haifa, Israel. [46]These authors contributed equally: Sagi Abelson, Grace Collord. [47]These authors jointly supervised this work: Moritz Gerstung, John E. Dick, Paul Brennan, George S. Vassiliou, Liran I. Shlush. *e-mail: moritz.gerstung@ebi.ac.uk; John.Dick@uhnresearch.ca; BrennanP@iarc.fr; gsv20@sanger.ac.uk; liranshlush3@gmail.com

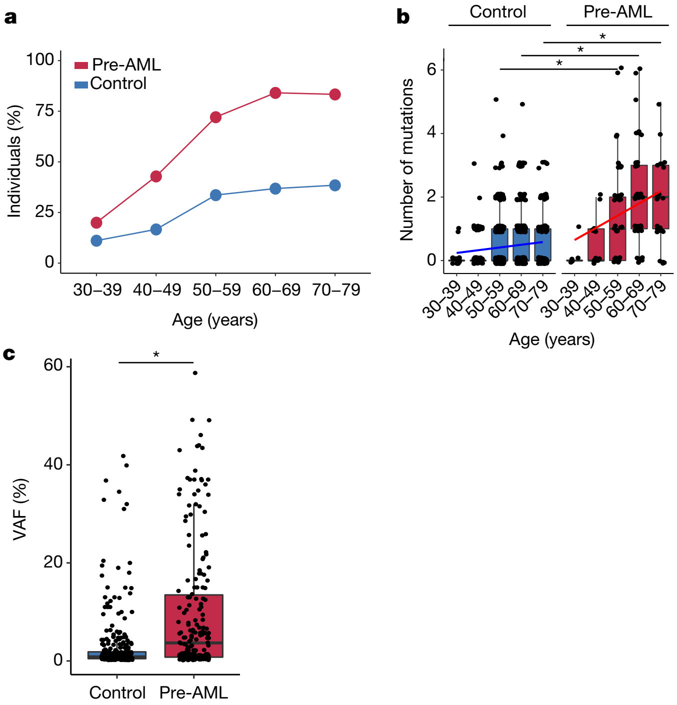


Fig. 1 | Prevalence of ARCH, number of mutations and clone size in individuals who developed AML. a, Prevalence of ARCH-PD among pre-AML cases (red) and controls (blue). **b**, The number of ARCH-PD mutations detected in cases and controls according to age. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. Individual values are indicated as dots. **c**, VAF of ARCH-PD mutations. $*P < 0.0005$, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. All panels show data for $n = 800$ biologically independent samples.

a prevalence that is in line with the largest studies of ARCH in the general population[4] ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Extended Data Fig. 1).

The median number of ARCH-PD mutations per individual increased with age and was significantly higher in the pre-AML group relative to controls (Fig. 1b and Supplementary Table 2). Furthermore, examination of ARCH-PD VAF distribution revealed significantly larger clones among the pre-AML cases ($P = 1.2 \times 10^{-13}$, two-sided Wilcoxon rank-sum test; Fig. 1c). To gain insight into clonal growth dynamics, we examined serially collected samples that were available for a subset of the validation cohort. We did not find significant differences in clonal expansion rates between pre-AML cases and controls (Extended Data Fig. 2a, b), although this may in part reflect the shorter follow-up of pre-AML cases, small sample size and large variance in growth rates (Extended Data Fig. 2c). The observed differences between pre-AML cases and controls may arise through cell-intrinsic or -extrinsic factors. Although these variables have not been adequately studied in ARCH, a number of observations in different contexts, such as aplasia, advanced age and after chemotherapy, have shown that increased clonal fitness is associated with distinct mutations depending on context[10–12]. Notably, mutations in splicing factor genes were significantly enriched among the pre-AML cases relative to the controls (odds ratio, 17.5; 95% confidence interval, 8.1–40.4; $P = 5.2 \times 10^{-16}$, two-sided Fisher's exact test) and were present in significantly younger individuals (median age 60.3 compared to 77.3 years, $P = 1.7 \times 10^{-4}$, two-sided Wilcoxon rank-sum test; Fig. 2a). Previous work suggests that spliceosome mutations appear to confer a competitive advantage in the context of ageing[10]. Therefore, it is possible that the significantly higher prevalence of such clones in younger pre-AML cases may reflect extrinsic selection pressures rather than earlier mutation acquisition.

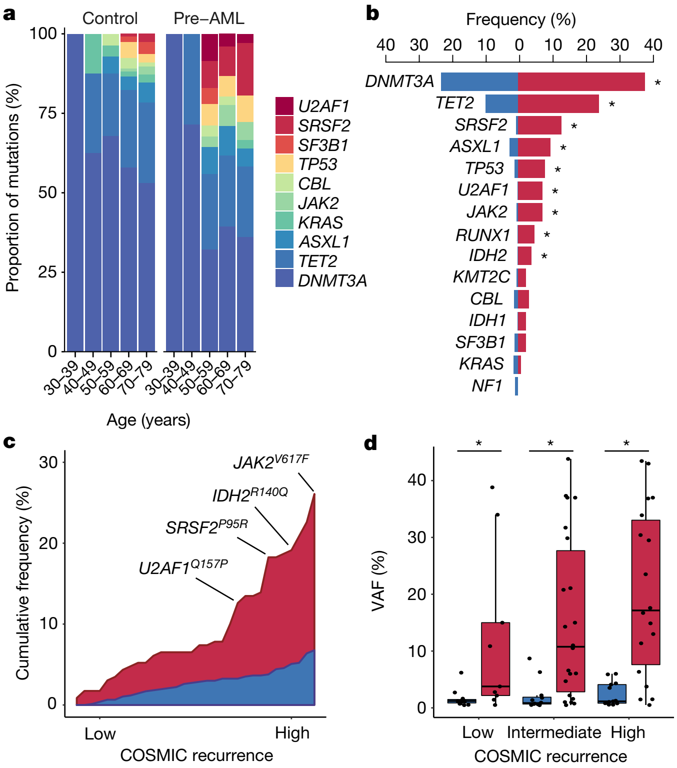


Fig. 2 | Accumulation of specific recurrent AML mutations in healthy individuals at a young age is associated with progression to AML. a, Relative frequency of mutations in the indicated genes according to age group for pre-AML cases and controls. **b**, Proportion of pre-AML cases (red) and controls (blue) who had ARCH-PD mutations in recurrently mutated genes. $*P < 0.05$, Fisher's exact test with Bonferroni multiple testing correction. **c**, The cumulative frequency of recurrent AML mutations (reported in >5 specimens in COSMIC) in pre-AML cases and controls. ARCH-PD mutations are ranked from left to right along the *x* axis from low to high recurrence. **d**, VAF of recurrent mutations in pre-AML cases and controls. Low, intermediate and highly recurrent COSMIC mutations are defined as those reported in 5–19 samples, 20–300 samples and >300 samples, respectively. Box plots indicate median, first and third quartiles and 1.5× interquartile range. $*P < 0.05$, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. All panels show data for $n = 800$ unique individuals.

In line with previous reports[5,6], we found that *DNMT3A* and *TET2* were the most commonly mutated genes in both groups (Fig. 2b). We could not identify any canonical *NPM1* mutations nor any *FLT3*-internal tandem duplication mutations, consistent with these arising late in leukaemogenesis[10,13]. Recurrent *CEBPA* mutations, which are implicated in around 10% of de novo AML[14], were also absent, suggesting that driver events in this gene may also be late events in AML evolution. In order to quantify the effect of different mutations on the likelihood of progression to AML, we ranked ARCH-PD mutations based on the number of times that they have been reported in Catalogue of Somatic Mutations in Cancer (COSMIC) database among individuals with haematological malignancies[15]. We found that mutations that are highly recurrent in cancer specimens were more common in pre-AML cases than in controls with ARCH-PD, whereas driver events in the controls tended to affect loci that are less frequently mutated in haematological malignancies and occurred at significantly lower VAF (Fig. 2c, d). Overall, these findings demonstrate notable differences in the mutational landscape of ARCH and pre-AML. Moreover, this work, in conjunction with recent insights into the origins of AML relapse[16], suggests that AML progression typically occurs over many years through clonal evolution of preleukaemic HSPCs before acquisition of late mutations leads to overt malignant transformation.

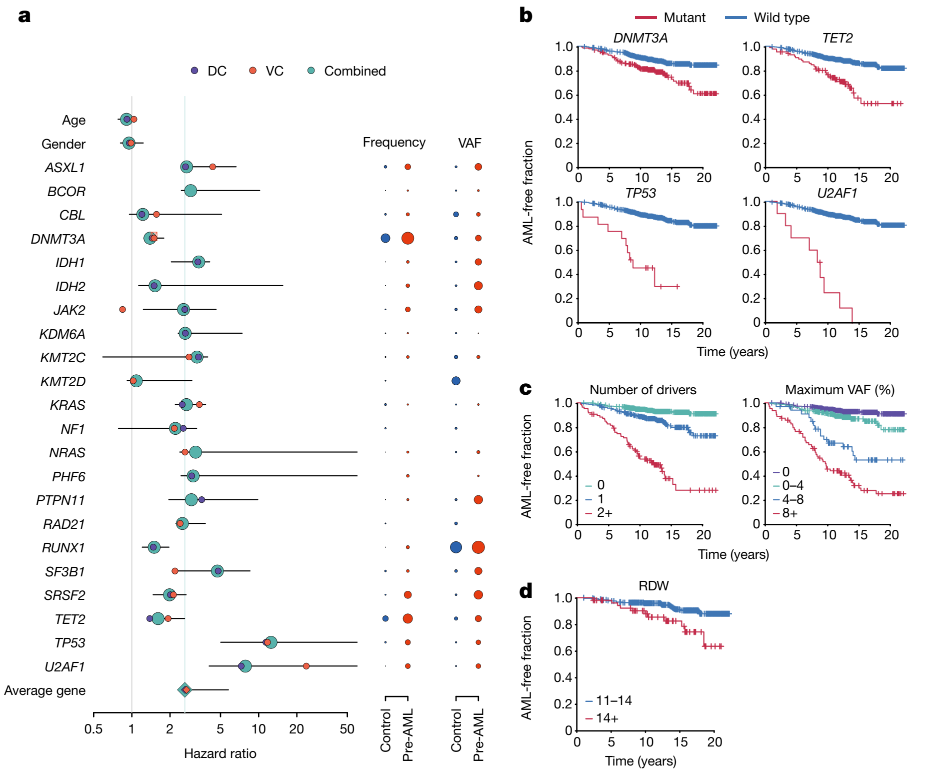


Fig. 3 | Model of future risk of AML. a, Forest plot of the risk of AML. Purple, orange and green circles indicate hazard ratios for the discovery (DC), validation (VC) and combined cohort, respectively. The horizontal lines denote 95% confidence intervals for the combined cohort. For each gene, the indicated hazard ratio applies to the 10-year risk of AML development conferred by each 5% increase in mutation VAF. The green vertical line indicates the mean hazard ratio across all genes. The hazard ratio for *RUNX1* must be interpreted with caution owing to the relatively high prevalence of deleterious germline variants in this gene, which may not be readily distinguishable from somatic mutations in unmatched sequencing assays (see Methods). The proportion of individuals with mutations in each gene and the average VAF are indicated to the right of the forest plot; red and blue circles represent pre-AML cases and controls, respectively, with circle sizes scaled to reflect mutation frequency and VAF. **b**–**d**, Kaplan–Meier curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status for selected genes (**b**), number of driver mutations per individual and largest clone detected (**c**) and RDW (**d**). Data for $n = 796$ unique individuals (**a**–**c**); $n = 299$ individuals for whom RDW measurements were available (**d**).

On the basis of these findings, we next developed an approach to quantify the relative contributions of driver mutations and clone sizes to the risk of progressing to AML. We tested different regularised logistic and Cox proportional hazards regression approaches, which achieved similar performance in both the discovery cohort (concordance (C) = 0.77 ± 0.03) and the validation cohort ($C = 0.84 \pm 0.05$; Extended Data Figs. 3, 4 and Supplementary Table 3). Models that were only trained on data from the discovery or validation cohort had similar coefficients (Fig. 3a). We therefore combined the datasets for a more accurate analysis of the contributions of mutations in individual genes to risk ($C = 0.77 \pm 0.05$; area under curve, 0.79; Supplementary Table 3). Quantitatively, we found that driver mutations in most genes conferred an approximately twofold increased risk of developing AML per 5% increase in clone size (Fig. 3a and Supplementary Table 3). Notable exceptions to this trend are the most frequently mutated ARCH genes, *DNMT3A* and *TET2*, which confer a lower risk of progression to AML (Fig. 3a, b and Supplementary Table 3). By contrast, a larger effect size was apparent for *TP53* (hazard ratio, 12.5; 95% confidence interval, 5.0–160.5) and *U2AF1* (hazard ratio, 7.9; 95% confidence interval, 4.1–192.2) mutations (Fig. 3a, b). However, we note that other ARCH-PD genes, such as *SRSF2*, can contribute a similar relative risk owing to their presence at a higher VAF in pre-AML cases (Fig. 3a, Extended Data Fig. 5a and Supplementary Note). Of note, mutations in *TP53* and spliceosome genes (including *U2AF1*) are also associated with a poorer prognosis in AML[14]. Because the effect of each ARCH-PD mutation is deleterious and the effect of multiple mutations that are present in the same individual is multiplicative, a higher number of mutations is predicted to increase the risk of progression to AML (Fig. 3c). Similarly, the size of the largest driver clone was also strongly associated with the risk of progression to AML, in agreement with the risk of individual mutations generally being proportional to VAF (Fig. 3c). Collectively, although the VAF and the number of mutations confer much of the predictive value, this model does demonstrate distinct gene-level risk factors, and is able to quantify the cumulative impact of multiple mutations and clonal size on the likelihood of progression to AML.

Although our predictive model performs well in identifying those at risk of developing AML in our experimental cohorts, AML incidence rates in the general population are low (4:100,000)[1], and thus millions of individuals would need to be screened to identify the few pre-AML cases, with many false positives. We therefore sought to determine whether routinely available clinical information could improve prediction accuracy or identify a high-risk population for targeted genetic screening. We first analysed complete blood count and biochemistry data that were available for 37 of the pre-AML cases and 262 controls. As reported previously[5,10,17], ARCH-PD was overwhelmingly associated with normal blood counts and this was also the case for pre-AML cases, indicating that these did not represent undiagnosed myelodysplastic syndrome[18]. We identified a significant association between higher red blood cell distribution width (RDW) and risk of progression to AML ($P = 0.0016$, Wald test with Bonferroni multiple-testing correction, Fig. 3d). Although traditionally used in the evaluation of anaemia, raised RDW has been correlated with inflammation, ineffective erythropoiesis, cardiovascular disease and adverse outcomes in several inflammatory and malignant conditions[19]. The correlation between RDW and risk of AML development remained highly significant when controls without ARCH-PD were excluded

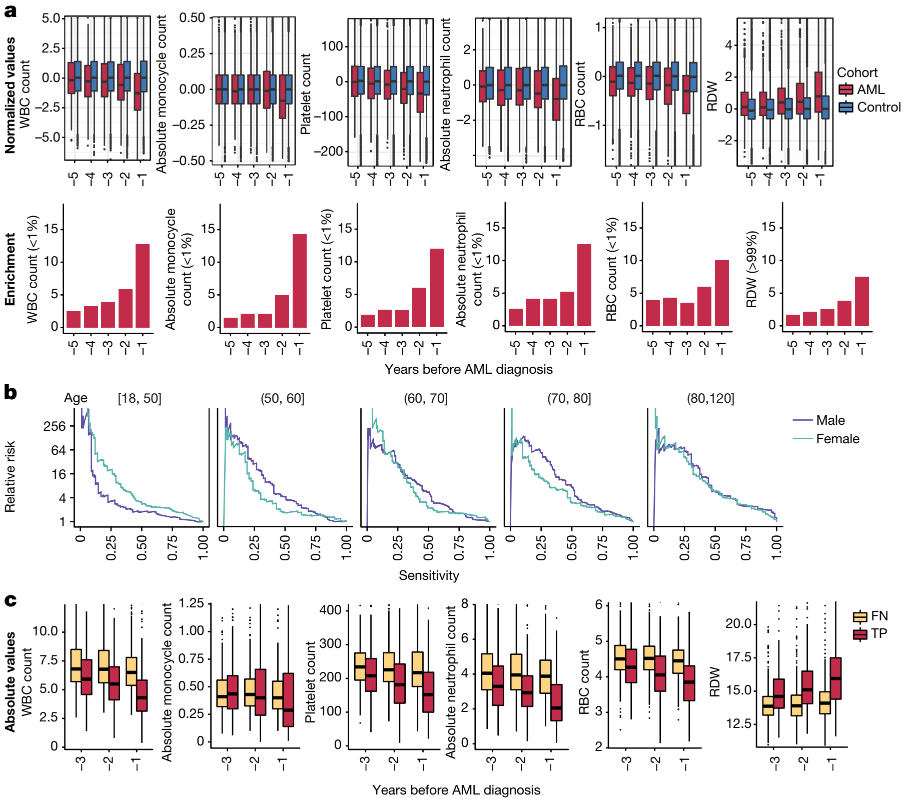


Fig. 4 | Increased risk of AML development inferred from electronic health records. a, Box plot of normalized laboratory measurements. Increased RDW, reduction in monocyte, platelet, red blood cell (RBC) and white blood cell (WBC) counts (top) show a high association (bottom) with a higher risk of AML development and differed at least a year before AML diagnosis. **b**, Model performance stratification by age and gender. Age ranges are indicated above each graph. **c**, Absolute laboratory values for true positive (TP) and false negative (FN) predictions. Box plots indicate median, first and third quartiles and 1.5× interquartile range.

from the analysis ($P = 3.5 \times 10^{-6}$, Wald test with Bonferroni multiple testing correction; Extended Data Fig. 5b). Higher RDW has previously been associated with ARCH and overall mortality[5], but has never been shown to distinguish ARCH from pre-leukaemia. In order to verify RDW as a predictive factor and determine whether additional clinical parameters are associated with risk of AML development, we studied the Clalit database[20], which contains electronic health records that include an average of 3.45 million individuals per year and data that were collected over a 15-year period[21]. We identified 875 cases with AML using stringent criteria based on diagnostic codes and treatment records (Extended Data Fig. 6 and Supplementary Table 4). Analysis of RDW trends revealed significantly raised measurements several years before AML diagnosis relative to age and sex-matched controls (Fig. 4a). Additional parameters that correlated with risk of AML development included reductions in monocyte, platelet, red blood cell and white blood cell counts, albeit usually remaining above the thresholds for clinically relevant cytopenias[18] (Fig. 4a and Extended Data Fig. 7). These findings suggest that evolving de novo AML may sometimes have a considerable prodrome with subtle but discernible clinical manifestations. We next applied a machine-learning approach to construct an AML prediction model based entirely on variables that are routinely documented in electronic health records (Extended Data Fig. 8 and Supplementary Table 4). This model was able to predict AML 6–12 months before diagnosis with a sensitivity of 25.7% and overall specificity of 98.2%. The model performed consistently across different age groups with an increased relative risk of 28 and 24 for males and females, respectively, between the age of 60 and 70 years (Fig. 4b). To better understand which patients are most likely to be accurately classified by this model, we compared absolute laboratory values for true positives and false negatives. We found that 35.5% of false-negative predictions were for patients for whom infrequent blood count data were available (Extended Data Fig. 9). Some of the true-positive cases had mildly abnormal blood counts that would not initiate a diagnostic work-up (Fig. 4c), and cytopenias that would be compatible with undiagnosed myelodysplastic syndrome[18] were uncommon.

Collectively, our findings provide new insights into the pre-clinical evolution of AML and support the hypothesis that individuals at high risk of AML development can be identified years before they develop overt disease. To this end, we present two distinct models for the prediction of de novo AML: one based on somatic point mutations and the other on routinely documented clinical information. We find that basic clinical and laboratory data can identify a high-risk subgroup 6–12 months before AML presentation, while genetic information can identify a substantial fraction of cases several years to more than a decade before diagnosis. By characterizing features that distinguish benign ARCH from pre-leukaemia, our models give valuable insights into leukaemogenesis. It is evident from the current study, together with our recent analysis of mutation acquisition from pre-leukaemic development through to relapse[16], that long-term pre-leukaemic HSPCs frequently carry mutations and undergo considerable clonal expansion while retaining differentiation capacity for years before AML diagnosis. Furthermore, it is clear that some mutations, particularly those affecting *TP53* and *U2AF1*, impart a relatively high risk of subsequent AML, whereas mutations in other genes, for example *DNMT3A* and *TET2*, confer a lesser risk of malignant transformation. Previous studies suggest that oncogenic mutations in *TP53* and spliceosome genes confer little or no competitive advantage in the absence of particular selective pressures[11,22], indicating that cell-extrinsic factors may be important determinants of clonal trajectory.

Cancer predictive models have enabled successful early detection and intervention programmes for several solid tumours[23–25]. However, screening tests are unavailable for the sub-clinical stages of most haematological malignancies. Our study provides proof-of-concept for the feasibility of early detection of healthy individuals at high risk of developing AML, and is a first step in the design of future clinical studies to investigate the potential benefits of early interventions in this deadly disease. However, the infrequency of AML necessitates that future screening tests provide high sensitivity and specificity. Our findings suggest that basic clinical data may identify a higher risk population that might benefit from targeted genetic screening. Equally, combining clinical and genetic information in a single model and including structural driver events is likely to improve model accuracy further. Nevertheless, establishing the utility of such a tandem approach will require extensive clinical and genetic analysis on the same population cohort, in a prospective setting. Furthermore, ARCH is associated with several non-malignant conditions[4,5], and may have a causal role in cardiovascular disease[26,27]. Therefore, genetic testing for ARCH may also prove useful in the management of common age-related diseases. Moreover, this study has broader implications for cancer screening and early intervention beyond AML. Advances in sequencing technologies have revealed a remarkable degree of somatic genetic diversity in normal ageing tissues, often characterized by the presence of clones that have canonical oncogenic mutations[28]. The degree to which clones at high risk of malignant transformation can be reliably distinguished from their indolent counterparts is an important biological question with compelling clinical ramifications. Understanding the selective pressures and cell-intrinsic mechanisms governing clonal fate is the next important step in developing strategies to predict and prevent progression to overt malignancy.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0317-6

Received: 9 July 2017; Accepted: 3 May 2018;
Published online 9 July 2018.

1. Deschler, B. & Lübbert, M. Acute myeloid leukemia: epidemiology and etiology. *Cancer* **107**, 2099–2107 (2006).
2. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl Acad. Sci. USA* **111**, 2548–2553 (2014).
3. Shlush, L. I. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
4. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
5. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
6. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
7. Busque, L. et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).
8. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
9. Acuna-Hidalgo, R. et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am. J. Hum. Genet.* **101**, 50–64 (2017).
10. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
11. Wong, T. N., et al. Role of *TP53* mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518**, 552–555 (2015).
12. Yoshizato, T. et al. Somatic mutations and clonal hematopoiesis in aplastic anemia. *N. Engl. J. Med.* **373**, 35–47 (2015).
13. Krönke, J. et al. Clonal evolution in relapsed *NPM1*-mutated acute myeloid leukemia. *Blood* **122**, 100–108 (2013).
14. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
15. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
16. Shlush, L. I. et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
17. Buscarlet, M. et al. *DNMT3A* and *TET2* dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* **130**, 753–762 (2017).
18. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
19. Hu, L. et al. Prognostic value of RDW in cancers: a systematic review and meta-analysis. *Oncotarget* **8**, 16027–16035 (2017).
20. Balicer, R. D. & Afek, A. Digital health nation: Israel's global big data innovation hub. *Lancet* **389**, 2451–2453 (2017).
21. Dagan, N., Cohen-Stavi, C., Leventer-Roberts, M. & Balicer, R. D. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. *Br. Med. J.* **356**, i6755 (2017).
22. McKerrell, T. & Vassiliou, G. S. Aging as a driver of leukemogenesis. *Sci. Transl. Med.* **7**, 306fs38 (2015).
23. Vickers, A. J. Prediction models in cancer care. *CA Cancer J. Clin.* **61**, 315–326 (2011).
24. Cassidy, A. et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br. J. Cancer* **98**, 270–276 (2008).
25. Wang, X., Oldani, M. J., Zhao, X., Huang, X. & Qian, D. A review of cancer risk prediction models with genetic variants. *Cancer Inform.* **13**, 19–28 (2014).
26. Fuster, J. J. et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
27. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
28. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).

Acknowledgements This work was supported by a Quest for Cure grant to L.I.S., J.C.Y.W. and M.D.M. from the Leukemia and Lymphoma Society, and the following grants to L.I.S from: ERC Horizon 2020 MAMLE, Abisch-Frenkel foundation and an American Society of Hematology Scholar Award. Further funding to J.E.D. was provided by the Canada Research Chair Program, Ontario Institute for Cancer Research, the province of Ontario, Canadian Cancer Society, the Canadian Institutes for Health Research and the Ontario Ministry of Health and Long Term Care to UHN, whose views are not expressed here. Work conducted at the Sanger Institute was supported by the Wellcome Trust and UK Medical Research Council. S.A. was personally funded by the Benjamin Pearl fellowship from the McEwen Centre for Regenerative Medicine, G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); G.S.V. by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA) and a Cancer Research UK Senior Cancer Research Fellowship (C22324/A23015). G.S.V.'s laboratory is also funded by the Kay Kendall Leukaemia Fund and Bloodwise. We thank A. Mitchell and all members of the Dick and Shlush laboratories for comments and T. Hudson for early study planning; G. Barabash for organising the Clalit dataset collaboration. The EPIC study centres were supported by the Hellenic Health Foundation, Regional Government of Asturias, the Regional Government of Murcia (no. 6236), the Spanish Ministry of Health network RTICCC (ISCIII RD12/0036/0018), FEDER funds/European Regional Development Fund (ERDF), "a way to build Europe", Generalitat de Catalunya, AGAUR 2014SGR726; EPIC Ragusa in Italy-Aire-Onlus Ragusa; Epic Italy-Associazione Italiana per la Ricerca sul Cancro (AIRC) Milan, Italy. S.V.B. and T.J.P. are supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at the Princess Margaret Cancer Centre.

Reviewer information *Nature* thanks R. Levine, P. Van Loo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.W.K.N., O.W., N.M.C. and E.N. contributed equally to the work. S.A. performed error-corrected sequencing, analysed sequencing data, performed statistical analyses, contributed to genetic predictive model derivation and wrote the manuscript. G.C. performed variant calling, statistical analyses, derived genetic predictive models and wrote the manuscript. M.G., S.W.K.N., O.W. and R.C. derived genetic predictive models. N.M.C., E.N. and N.B. derived the clinical prediction model. P.C.Z., Z.Z., I.C., K.N., C.L., C.H., D.H., F.M., J.E., J.K.M., D.P., L.T., P.K., S.V.B. and A.Br. and A.Ba. provided sequencing and technical support and enabled sample acquisition. L.H., Y.S., T.T.W., T.J.P., K.R. and D.J. provided bioinformatics support. R.L., S.H., M.J., K.M.B., A.Kr. and N.J.W. enabled sample acquisition, clinical data curation and/or provided clinical expertise. D.S., J.D.M., P.A., E.S., S.B., P.Be., M.D.M and I.M. contributed to data analysis and interpretation. P.J.C. and E.P. contributed to data interpretation and designed the targeted sequencing assay for the validation cohort. J.C.Y.W. revised the manuscript. J.R.Q., A.Ka., C.L.V., A.T., E.S.-F., J.M.H., R.C.T., R.T., G.M., H.B., S.Pa., R.K., S.S., S.Po., N.J.W., N.S., K.-T.K., M.F., J.M.K., E.R., P.V. and R.V. enabled sample acquisition (EPIC). A.T. and R.D.B. analysed Clalit data and derived the clinical prediction model. M.G. derived predictive genetic models, contributed to sequencing data analysis and manuscript writing. J.E.D. contributed to funding applications, study supervision and manuscript writing. P.Br. supervised sample acquisition from all EPIC centres. G.S.V. and L.I.S. designed and supervised all aspects of the study and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0317-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0317-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.G., J.E.D., P.B., G.S.V. or L.I.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study participants. Samples for both the discovery and validation cohort were obtained from participants in the EPIC study[29]. All relevant ethical regulations were followed. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols were approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, the Weizmann Institute of Science Ethics board approval #60-1 and East of England–Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). Patients with AML were identified based on the following ICD9 codes: 9861/3, 9860/3, 9801/3, 9866/3, 9891/3, 9867/3, 9874/3, 9840/3, 9872/3, 9895/3, 9873/3, which included only cases of de novo AML, and no secondary AML. All patients provided peripheral blood samples for which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen before DNA extraction.

Discovery cohort. In total, 509 DNA samples were collected from individuals upon enrolment into the EPIC study between 1993 and 1998 across 17 different centres[29] (Supplementary Table 1). Altogether, 95 individuals who developed AML an average of 6.3 years (interquartile range (IQR) = 4.8 years) after the sample was collected were included in the pre-AML group. For the control group, 414 age- and gender-matched individuals were selected, as they did not develop any haematological disorders during the average follow-up period of 11.6 years (IQR = 2.1 years). The median age at recruitment was 56.7 years (range, 36.08–74.42). In order to minimize any possible demographic biases, an approximate 1:4.5 pre-AML to control ratio was maintained across the different centres.

Validation cohort. Samples were obtained from individuals enrolled in the EPIC-Norfolk longitudinal cohort study between 1994 and 2010. Samples and clinical metadata were available from 37 patients with AML (of which 8 were already included in the discovery cohort) and 262 age- and gender-matched controls without a history of cancer or any haematological conditions. The average time between the first blood sampling and AML diagnosis was 10.5 years (IQR = 8.3 years). The average follow-up period for the control cohort was 17.5 years (IQR = 3.8). For 12 individuals in the pre-AML cohort, 2–3 blood specimens were available, taken a median of 3.4 years apart. Of the 262 controls, 141 had multiple blood samples available, spanning a median of 10.5 years. Blood counts and other clinical parameters were available for all study participants (Supplementary Table 1).

Targeted sequencing. *Discovery cohort sequencing.* Targeted deep sequencing was performed using error-corrected sequencing as follows.

Shearing of genomic DNA, preparation of pre-capture sequencing libraries, hybridization-based enrichment, assessment of the libraries quality and enrichment following hybridization were performed as previously described[30]. In brief, 100 ng of genomic DNA was sheared before library construction (KAPA Hyper Prep Kit KK8504, Kapa Biosystems) with a Covaris E220 instrument using the recommended settings for 250-bp fragments. Following end repair and A-tailing, adaptor ligation was performed using 100-fold molar excess of Molecular Index Adaptor. Library clean-up was performed with Agencourt AMPure XP beads (Beckman-Coulter) and the ligated fragments were then amplified for eight cycles using 0.5 μM Illumina universal and indexing primers.

Targeted capture was carried out on pools containing three indexed libraries. Each pool of adaptor-ligated DNA was combined with 5 μl of 1 mg ml^{-1} Cot-I DNA (Invitrogen), and 1 nmol each of xGEN Universal Blocking Oligo, TS-p5, and xGen Universal Blocking Oligo, TS-p7 (8 nucleotides). The mixture was dried using a SpeedVac and then re-suspended in 1.1 μl water, 8.5 μl NimbleGen 2× hybridization buffer and 3.4 μl NimbleGen hybridization component A. The mixture was heat denatured at 95 °C for 10 min before addition of 4 μl of xGen Lockdown Probes (xGen AML Cancer Panel v.1.0, 3 pmol). Each pool was then hybridized at 47 °C for 72 h. Washing and recovery of the captured DNA was performed according to the manufacturer's specifications. In brief, 100 μl of clean streptavidin beads was added to each capture. Following separation and removal of the supernatant using a magnet, 200 μl 1× Stringent Wash Buffer was added and the reaction was incubated at 65 °C for 5 min. The supernatant containing unbound DNA was removed before repeating the high stringency wash one additional time. Then, the bound DNA was washed as follows: (1) 200 μl 1× Wash Buffer I and separation of the supernatants by magnetic separation; (2) 200 μl 1× Wash Buffer II after magnetic separation; (3) 200 μl 1× Wash Buffer III and removal of the supernatants using magnetic separation. The captured DNA on beads was resuspended in 40 μl of Nuclease-Free water before dividing the total volume into two PCR tubes and subjecting the libraries to 10 cycles of post-capture amplification (manufacturer-recommended conditions; Kapa Biosystems). Before sequencing, libraries were spiked with 2% PhiX.

Validation cohort sequencing. Targeted sequencing was performed using a custom complementary RNA bait set (SureSelect, Agilent, ELID 0537771) designed complementary to all coding exons of 111 genes that have been implicated in myeloid leukaemogenesis (Extended Data Table 1). Genomic DNA was extracted from peripheral whole blood and sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75-bp paired-end sequencing as per Illumina and Agilent SureSelect protocols.

Variant calling. *Discovery cohort variant calling and error correction.* The 126-bp paired-end reads sequencing data from the Illumina platform were converted to FASTQ format, the 2-bp molecular barcode information at each read of the pair was trimmed and was written in the reads' name. The thymine nucleotide required for ligation was removed from the sequences. Burrows–Wheeler aligner (BWA-mem)[31] was used for alignment of the processed FASTQ files to the reference hg19 genome, after realignment of insertions and deletions (indels) using GATK[32]. An in-house algorithm was written to collapse read families that share the same molecular barcode sequence, the left-most genomic position of where each read of the pair maps to the reference and the CIGAR string. Families that consisted of at least two reads were used to generate consensus reads and a consensus base was called when there was at least 70% agreement. When a consensus base was called, it was assigned with the maximum base quality score observed in its corresponding pre-collapsed reads. Furthermore, when possible, duplex reads[33] were generated from two consensus reads, from a singleton read and a consensus read, or from two singleton reads. For each sequenced sample, we generated two BAM files, called BAM1 and BAM2. BAM1 consisted of duplex reads, consensus reads and singleton reads, thereby including some error-corrected and non-error corrected reads, while still containing all the genomic information encoded in the data in the form of unique DNA molecules. BAM2 consisted of duplex reads and consensus reads but not singleton reads. Both files were then analysed to detect single nucleotide variants (SNVs) and small indels using Varscan2[34]. To further remove sequencing artefacts and improve sensitivity, we applied a two-step polishing statistical approach that models the error rate for each sequenced genomic position. For both steps, BAM1 was used and all samples except the sample that was investigated were included for error rate modelling. At step one, as previously described[30], the error rates were modelled by fitting Weibull distribution curves to the non-reference allele fractions. SNVs with allele fractions that were statistically distinguishable from the background error rates ($P = 0$) were further analysed. At step 2, the coverage of the non-reference allele fractions was considered using linear line fitting that describes the negative correlation that exist between the log(non-reference allele fraction) and the corresponding log(coverage) values. This allowed us to estimate different error rates at different coverage depths. Because indel errors are rare and cannot be appropriately modelled by the same statistical framework, they were called using barcode-mediated error correction alone. At least 10 consensus reads, 5 supporting reads on the forward strand, 5 supporting reads on the reverse strand and 2 duplex reads were required to call an indel. Additional post-processing steps applied to data from both the discovery cohort and validation cohort are detailed in 'Additional post-processing filters applied to discovery and validation cohort data'. Variants were annotated using Annovar[35].

Validation cohort variant calling. Sequencing reads were aligned to the reference genome (GRCh37d5) using the Burrows–Wheeler aligner (BWA-aln)[31]. Unmapped reads, PCR duplicates and reads mapping to regions outside the target regions (merged exonic regions and 10 bp either side of each exon) were excluded from analysis. Sequencing depth at each base was assessed using Bedtools coverage v.2.24.0[36].

Somatic SNVs were called using shearwater, an algorithm developed for detecting subclonal mutations in deep-sequencing experiments (https://github.com/gerstung-lab/deepSNV v.1.21.5)[37–39] considering only reads with minimum nucleotide and mapping quality of 25 and 40, respectively. This algorithm models the error rate at individual loci using information from multiple unrelated samples. Additionally, allele counts at the recurrent AML mutation hotspots listed in 'Curation of oncogenic variants' were generated using an in-house script (https://github.com/cancerit/alleleCount) and manually inspected in the Jbrowse genome browser[40]. To further complement our SNV calling approach, we applied an extensively validated in-house version of CaVEMan v.1.11.2 (Cancer variants through expectation maximization)[41]. CaVEMan compares sequencing reads between study and nominated normal samples and uses a naive Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan).

Post-processing filters required that the following criteria were met for CaVEMan to call a somatic substitution. (1) If coverage of the mutant allele was less than 8, at least one mutant allele was detected in the first two-thirds of the read. (2) Less than 3% of the mutant alleles with base quality ≥15 were found in the nominated normal sample. (3) Mean mapping quality of the mutant allele reads was ≥21. (4) The mutation does not fall in a simple repeat or centromeric region. (5) Fewer than 10% of the reads covering the position contained an indel according to mapping. (6) Less than 80% of the reads report the mutant allele at the same read position. (7) At least a third of the reads calling the variant had a base quality of 25 or higher. (8) Not all mutant alleles reported in the second half of the read. (9) Position does not fall within a germline insertion or deletion.

The following additional post-processing criteria were applied to all SNV calls. (1) Minimum VAF = 0.5% with a minimum of five bidirectional calls reporting the mutant allele (with at least two reads in forward and reverse directions). (2) No indel called within a read length (75 bp) of the putative substitution.

Small indels were sought using two complementary bioinformatics approaches. First, an in-house version of Pindel v.2.2[42] (https://github.com/cancerit/cgpPindel) was applied. We additionally used the aforementioned deepSNV algorithm in order to increase sensitivity for indels present at low VAF. VAF correction was performed using an in-house script (https://github.com/cancerit/vafCorrect).

Post-processing filters required that the following criteria were met for a variant to be called. (1) A minimum of five reads supporting the variant with a minimum of two reads in each direction. For Pindel, the total read count was based on the union of the BWA and Pindel reads reporting the mutant allele. (2) VAF $\geq$ 0.5%. (3) Variant not present within an unmatched normal panel of approximately 400 samples. (4) No reads supporting the variant identified in the nominated normal sample.

Mutations were annotated according to ENSEMBL v.58 using VAGrENT[43] for transcript and protein effects (https://github.com/cancerit/VAGrENT) and Annovar[35] for additional functional annotation.

Additional post-processing filters applied to discovery and validation cohort data. The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event. (1) Any mutant allele reported within 75 bp of another variant. (2) Any mutant allele with a population allele frequency >1 in 1,000 according to any of five large polymorphism databases (ExAC, 1000 Genomes Project, ESP6500, CG46 and Kaviar) that is not a canonical hotspot driver mutation with COSMIC recurrence >100. (3) Mutations that were present in >10% of the control cohort but not recurrent in COSMIC were flagged as potential germline variants or sequencing artefacts. (4) As artefactual indels tend to be recurrent, any indels occurring in >2 samples were flagged as for additional inspection.

Curation of oncogenic variants. Putative oncogenic variants were identified according to evidence for functional relevance in AML as previously described and used to define ARCH-PD[14].

Variants were annotated as likely driver events if they fulfilled any of the following criteria. (1) Truncating mutations (nonsense, essential splice site or frameshift indel) in the following genes implicated in AML pathogenesis by loss-of-function: *NF1*, *DNMT3A*, *TET2*, *IKZF1*, *RAD21*, *WT1*, *KMT2D*, *SH2B3*, *TP53*, *CEBPA*, *ASXL1*, *RUNX1*, *BCOR*, *KDM6A*, *STAG2*, *PHF6* and *KMT2C*. (2) Truncating variants in *CALR* exon 9. (3) *JAK2*^{V617F}. (4) *FLT3* internal tandem duplication. (5) Non-synonymous variants at the following hotspot residues: *CBL* E366, L380, C384, C404, R420 and C396; *DNMT3A* R882; *FLT3* D835; *IDH1* R132; *IDH2* R172 and R140; *KIT* W557, V559 and D816; *KRAS* A146, Q61, G13 and G12; *MPL* W515; *NRAS* Q61, G12 and G13; *SF3B1* K700 and K666; *SRSF2* P95; *U2AF1* Q157, R156 and S34. (6) Non-synonymous variants reported at least 10 times in COSMIC with VAF <42% and population allele frequency <0.003. (7) Non-synonymous variants clustering within a functionally validated locus or within four amino acids of a hotspot variant with population allele frequency <0.003 and VAF <42%. (8) Non-synonymous variants reported in COSMIC >100 times with population allele frequency <0.003 regardless of VAF.

Our driver curation strategy inevitably runs a small risk of including germline variants in familial AML genes. We feel that in the real world, where a matched constitutional DNA sample would be unavailable, this is the best approach.

Statistical analysis. All statistical analyses were performed in the R statistical programming environment. A two-sided Wilcoxon rank-sum test was used to assign significance level for differences in the median number of somatic mutations among the pre-AML and control groups, the median VAF of mutations among groups. and the age of individuals with spliceosome mutations. Fisher's exact test was used to assess the significance of differences in the prevalence of ARCH among the groups and spliceosome mutations in the pre-AML group.

Predictive modelling. *Cox proportional hazards model with random effects.* We used a Cox proportional hazards regression to model AML progression-free survival as previously described[14,38]. We used random effects for the Cox proportional hazards model in the CoxHD R package (http://github.com/gerstung-lab/CoxHD). A key strength of this approach is the ability to include many variables in one model while shrinking estimated effects for parameters with weak support in the data, thus controlling for overfitting. We used weighting to minimize the biases introduced by the artificial case–control ratio[44,45] and calculated hazard ratios relative to the (approximate) true cumulative incidence of about 1–3/1,000 in the given age range over a follow up of 10–20 years. The observed driver mutation frequency and VAF in pre-AML cases closely resembled values expected based on the estimated risks, indicating that risk model and driver prevalence are well aligned (Extended Data Fig. 4). Full details of model derivation and comparisons with alternative methods are included in the accompanying code (Supplementary Note, also available at https://github.com/gerstung-lab/preAML). In brief, variables comprised age, gender and the VAF of putative driver mutations (see 'Curation of oncogenic variants' for details of variant curation). We performed agnostic imputation of missing variables by mean and linear rescaling of gene variables by a power of 10 to a magnitude of 1. The model was first trained separately on the discovery cohort and validation cohort. For each of these two models, we evaluated the following measures of predictive accuracy before and after leave-one-out cross-validation (LOOCV): concordance (C)[46] and time-dependent area under the receiver-operating characteristic curve (AUC)[47]. The models trained on the validation and discovery cohorts were then cross-validated using the data from the other cohort. In view of the cross-validation results and close correlation between coefficients (Supplementary Table 3), we derived a model on the combined cohorts using both cohorts in order to achieve greater accuracy on the individual effects. Confidence intervals were calculated using 100 bootstrap samples. The coefficients and performance metrics for each iteration of the model are included in Supplementary Table 3.

Concordance measures were obtained using the survConcordance() function implemented in the survival R package[45]. Dynamic AUC was calculated with AUC.uno() implemented in the survAUC package. Time-independent AUCs were calculated using the performance function implemented in the ROCR package. The expected incidence of AML was calculated from the UK office of national statistics, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence. All-cause mortality data was obtained from the office of national statistics (https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables).

Ridge-regularized logistic regression. Using the same covariates as in 'Cox proportional hazards model with random effects', we fitted a ridge-regularized logistic regression model to dichotomised outcome data. While logistic regression is a common choice for case–control analyses, a downside of this approach is the inability to explicitly use time-dependent covariates. The penalty parameter was chosen using LOOCV on the full cohort; this value was then used on the discovery cohort and validation cohort to yield the same scaling of coefficients. Confidence intervals were calculated using 100 bootstrap samples. Fitting was performed using the glmnet R package. AUC as the primary performance metric was calculated using the ROCR R package.

Additional regression models. Two alternative predictive models were developed. Model 1 performs logistic-regression-based predictions using four types of features: gender, age at blood sampling, the sum of the VAFs ARCH-PD reported in COSMIC v.80 to be recurrent (at least two case reports in haematopoietic and lymphoid tissues) and somatic mutation burden of selected genes, where each gene was represented by the sum of the VAFs corresponding to ARCH-PD mutations in that gene. We measured the predictive performance of each gene via the AUC obtained in a fivefold cross-validation when using only the gene as a predictive feature, and only retained genes with AUC > 55% in the final model.

For model 2 we applied LASSO regression as implemented in the glmnet R package, while enabling LOOCV to fit a Cox regression model. A minimal subset of ARCH-PD variants was selected for which the respective weighted combined VAFs were highly predictive of AML development in the training set. Scores were calculated for each patient as a linear combination of VAF of mutations weighted by regression coefficients that were estimated from the training data. As most scores were zero in the training subset, non-zero scores were discretized to take on a value of 1 that corresponds to AML prediction.

Models 1 and 2 were trained on the discovery cohort and tested for their association with AML development using the validation cohort data. Survival analysis was performed using the Kaplan–Meier and Cox proportional hazards models. Wald's test was used to evaluate the significance of hazard ratios. Logistic regression models were used with the positive predictive value metric to determine the ability of various mutations and other patient parameters to predict AML development. The rms R package was used for logistic regression analysis, and the pROC 1.8 R package was used for receiver-operating characteristic curve analysis.

AML-predictive model based on electronic health records. *Clalit database.* The Clalit database includes information from patients covered by the Clalit health services in Israel[20] during the years 2002–2017. The Clalit training-set data, contains the electronic health records (EHR) of 3.45 million individuals per year on average. All data was anonymized through hashing of personal identifiers and addresses and randomization of dates by sampling a random number of weeks for each patient and adding it to all dates in the patient diagnoses, laboratory and medication records. This approach maintained differential data analysis per patient. Diagnoses codes were acquired from both primary care and hospitalization records, and were mapped to the ICD-9 coding system for historical reasons, with few exceptions that used a partial ICD-10 coding system. Laboratory records were normalized for age and gender by subtracting raw test values from the median levels observed among all test values with matching gender and age (using a bin size of five years). We observed some chronological biases in laboratory ranges, but avoid normalizing these and instead insured case and controls are matched for chronological distributions.

Defining AML cases. We screened for all active patients ($18 < \text{age} < 100$) who were diagnosed with AML (ICD-9 code 205.0*) between the years 2003 and 2016. We then excluded cases based on the following criteria. (1) We excluded patients with prior myeloid malignancies to omit secondary AML, consistent with the case selection for the genetic model. The following diagnosis were excluded if documented within five years before the diagnosis of AML: essential thrombocythemia (ICD-9 238.71), low-grade myelodysplastic syndrome (MDS) (ICD-9 238.72); high-grade MDS lesions (ICD-9 238.73); MDS with 5q deletion (ICD-9 238.74); MDS, unspecified (ICD-9 238.75); polycythemia vera (ICD-9 238.4); myelofibrosis (ICD-9 289.83); chronic myelomonocytic leukaemia (ICD-9 206.10-206.22).

(2) Patients that had any procedures performed on bone marrow or spleen (ICD-10 code Z41) in the five-year period before first mention of AML diagnosis code in their record. These patients were presumed to have an inaccurate AML diagnosis date or misdiagnosis recorded.

(3) Patients that received medications suggestive of an alternative diagnosis of chronic myeloid leukaemia, lymphoid malignancy or acute promyelocytic leukaemia (APL). At any time before diagnosis: imatinib, dasatinib, anagrelide, hydroxycarbamide, asparaginase, pegaspargase or arsenic trioxide. At any time after diagnosis: imatinib, dasatinib, methotrexate, tretinoin or arsenic trioxide. At any time after diagnosis, along with any acute lymphoblastic leukaemia diagnosis (ICD-9 204) or more than single dose: mercaptopurine. APL cases were excluded as early diagnosis of APL will most probably not change its outcome, as treatment is successful already.

(4) Patients without a hospitalization record within three months before or after the onset diagnosis. This parameter was used as it is unlikely that a patient with AML will not be hospitalized close to diagnosis. This filter reduced false-positive cases and better defined the onset date.

We refined the estimated time of onset using the earliest time at which any of the following diagnosis appeared in the patient's history: amyloidosis (ICD-9 277.3), lymphoid leukaemia (ICD-9 204), myeloid leukaemia (ICD-9 205), leukaemia of unspecified cell type (ICD-9 208).

This strategy retained 875 AML cases in the training set for further analysis. These were further validated by manual expert inspection of the complete records of 8% of the cases.

To define the control set, we included all Clalit individuals that were not cases. Since our analysis was aggregating data from a historical time window of 15 years, we associated each control with a randomized time point for evaluation. Using this approach, both cases and controls represented a specific time point in the historical record of a patient, with matching calendric, age and gender distributions. Through this strategy 5,238,528 controls were used.

Defining features for construction of a predictive a score. We extracted the following features for discriminative analysis of cases and controls (this procedure was applied repeatedly in cross-validation as discussed below). (1) Age (in years) at time point. (2) Gender. (3) Laboratory features. Out of 2,770 different types of laboratory tests, we selected the top 50 most frequent laboratory tests (Supplementary Table 4). For each laboratory measurement, we used median age- and gender-normalized test values per patient in three time windows for 6–12 months before onset, 1–2 years before onset and 2–3 years before onset. In addition, we compute the slope of the normalized laboratory measurements for the 6–12 month time window using a linear regression model. (4) Diagnosis features. Of the 1780 different major ICD-9 diagnosis codes, we selected only diagnoses that were previously observed in at least 10 different cases and have an increased relative risk for AML >twofold (as observed in the training set, Supplementary Table 4). For each diagnosis code, we mark whether it appeared in each of the patients in time intervals of 6 months to 3 years, and 3–5 years before onset. (5) BMI features. For each patient in the cohort, we extracted median BMI, weight and height as measured in time intervals of 6 months to 2 years, and 2–3 years before onset.

Gradient boosting. We used the R package xgboost to infer parameters for a classifier given cases and controls. Objective was set to binary:logistic, the evaluation metric to AUC. We set nrounds = 5000, eta = 0.001, gamma = 0.1, lambda = 0.01, alpha = 0.01, max_depth = 6, min_child_weight = 2, subsample = 0.7 and colsample_bytree = 0.7. The boosting algorithm reports a function f that computes a predictive score given the features. Given a threshold T the expression $f(\text{patient features}) > T$ defines a classifier. To standardise thresholds we estimate quantiles for the scores on the training set $T(p) = \text{quantile}(f(\text{train}),p)$ and define the classifier for specificity level p as $f(\text{patient features}) > T(p)$ (Supplementary Table 4).

Cross-validation and relative risk evaluation. To evaluate the predictive value of the classification scheme while considering the strong age and gender biases in the incidence of AML, we performed fivefold cross-validation after splitting the cases and controls into five age- and gender-matched groups. For each fold, we sampled 100,000 controls and combined with the cases, constructed the feature set and trained the model. The model was then tested on the fold cases along with 200,000 sampled controls. We used standardized classifier parameters and standardized thresholds that were inferred based on each training set to generate a series of classifications on each test set and merged these based on the control quantiles in the test as described above. Given a threshold p to define high and low prediction score, we counted for each bin b that defines a patient in a specific age (<40, 40–50, 50–60, 60–70, 70–80, >80) and gender group: the number of cases in bin b ($N^b{}_{\text{case}}$) and the number of controls in bin b ($N^b{}_{\text{control}}$) where N^b is the number of patients in bin b (entire database minus recall controls that are only a sample of the cohort). $N^b(\text{case, high score}) = N^b{}_{\text{TP}}$ indicates the number of true positives (TP); $N^b(\text{case, low score}) = N^b{}_{\text{FN}}$ indicates the number of false negatives (FN); $N^b(\text{control, high score}) = N^b{}_{\text{FP}}$ indicates the number of false positives (FP); $N^b(\text{control, low score}) = N^b{}_{\text{TN}}$ indicates number of true negatives (TN).

For each age and gender group, the absolute risk for AML in the bin is computed by $r^b{}_{\text{abs}} = N^b{}_{\text{case}}/N^b$. The absolute risk given a high score is estimated as $r^b{}_{\text{abs,high}} = N^b{}_{\text{TP}}/(N^b{}_{\text{FP}}+ N^b{}_{\text{TP}})$. The relative risk in the bin is defined by $\text{rr}^b = r^b{}_{\text{abs,high}}/r^b{}_{\text{abs}}$ where the sensitivity level for the classifier threshold level is defined as $\text{sense}^b = N^b{}_{\text{TP}}/N^b{}_{\text{case}}$.

$$\text{rr} = \frac{\dfrac{\dfrac{\text{TP} \times \text{cases}}{(\text{TP} + \text{FN})}}{\dfrac{\text{TP} \times \text{cases}}{(\text{TP} + \text{FN})} + \dfrac{\text{FP} \times \text{controls}}{(\text{FP} + \text{TN})}}}{\dfrac{\text{cases}}{\text{cases}+\text{controls}}}$$

Clonal growth rate calculation. Individual clones were defined by different mutations in different study participants. Per clone we calculated α according to the following equation:

$$a = \log(V/V_0)\ /(T - T_0)$$

where T and T_0 indicate the age of the individual at the two measurement time points. V and V_0 correspond to the VAF at T and T_0, respectively.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code for derivation of the prediction model is publically available on Github (https://github.com/gerstung-lab/preAML). Code for the analysis of error-corrected sequencing is available from the Shlush lab upon request.

Data availability. Targeted sequencing data for the discovery cohort are deposited as BAM files at the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAD00001003583. All other data are available from the corresponding authors upon reasonable request. Sequencing data for the validation cohort are deposited at the European Genome-phenome Archive with accession number EGAD00001003703.

29. Riboli, E. et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
30. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
31. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
32. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
33. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
34. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
35. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10**, 1556–1566 (2015).
36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
37. Gerstung, M. et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
38. Gerstung, M. et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49**, 332–340 (2017).
39. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
40. Buels, R. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
41. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
42. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.17.1–15.7.12 (2015).
43. Menzies, A. et al. VAGrENT: Variation Annotation Generator. *Curr. Protoc. Bioinformatics* **52**, 15.18.1–15.18.11 (2015).

44. Antoniou, A. C. et al. A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. *Genet. Epidemiol.* **29**, 1–11 (2005).
45. Therneau, T. & Grambsch P. M. *Modeling Survival Data: Extending the Cox Model* 1st edn (Springer-Verlag, New York, 2000).
46. Harrell, F. E. Jr, Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
47. O'Quigley, J., Xu, R. & Stare, J. Explained randomness in proportional hazards models. *Stat. Med.* **24**, 479–489 (2005).

SCIENTIFIC REPORTS

OPEN

An integrated genomic analysis of anaplastic meningioma identifies prognostic molecular signatures

Received: 30 April 2018
Accepted: 16 August 2018
Published online: 10 September 2018

Grace Collord[1,2], Patrick Tarpey[1], Natalja Kurbatova[3], Inigo Martincorena[1], Sebastian Moran[4], Manuel Castro[4], Tibor Nagy[1], Graham Bignell[1], Francesco Maura[1,5,6], Matthew D. Young[1], Jorge Berna[7], Jose M. C. Tubio[7], Chris E. McMurran[8], Adam M. H. Young[8], Mathijs Sanders[1,21], Imran Noorani[1,8], Stephen J. Price[8], Colin Watts[9], Elke Leipnitz[10], Matthias Kirsch[10], Gabriele Schackert[10], Danita Pearson[11], Abel Devadass[11], Zvi Ram[17], V. Peter Collins[11], Kieren Allinson[11], Michael D. Jenkinson[12,20], Rasheed Zakaria[12,13], Khaja Syed[12,13], C. Oliver Hanemann[14], Jemma Dunn[14], Michael W. McDermott[15], Ramez W. Kirollos[8], George S. Vassiliou[1,16], Manel Esteller[4,18,19], Sam Behjati[1,2], Alvis Brazma[3], Thomas Santarius[8] & Ultan McDermott[1,20,22]

Anaplastic meningioma is a rare and aggressive brain tumor characterised by intractable recurrences and dismal outcomes. Here, we present an integrated analysis of the whole genome, transcriptome and methylation profiles of primary and recurrent anaplastic meningioma. A key finding was the delineation of distinct molecular subgroups that were associated with diametrically opposed survival outcomes. Relative to lower grade meningiomas, anaplastic tumors harbored frequent driver mutations in SWI/SNF complex genes, which were confined to the poor prognosis subgroup. Aggressive disease was further characterised by transcriptional evidence of increased PRC2 activity, stemness and epithelial-to-mesenchymal transition. Our analyses discern biologically distinct variants of anaplastic meningioma with prognostic and therapeutic significance.

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK. [2]Department of Paediatrics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK. [4]Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Catalonia, Spain. [5]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [6]Department of Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. [7]Mobile Genomes and Disease, Molecular Medicine and Chronic diseases Centre (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, 15706, Spain. [8]Department of Neurosurgery, Department of Clinical Neuroscience, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK. [9]Department of Neurosurgery, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. [10]Klinik und Poliklink für Neurochirurgie, "Carl Gustav Carus" Universitätsklinikum, Technische Universität Dresden, Fetscherstrasse 74, 01307, Dresden, Germany. [11]Department of Pathology, Cambridge University Hospital, CB2 0QQ, Cambridge, UK. [12]Department of Neurosurgery, The Walton Centre, Liverpool, L9 7LJ, UK. [13]Institute of Integrative Biology, University of Liverpool, Liverpool, L9 7LJ, UK. [14]Institute of Translational and Stratified Medicine, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth University, Plymouth, Devon, PL4 8AA, UK. [15]Department of Neurosurgery, UCSF Medical Center, San Francisco, CA, 94143-0112, USA. [16]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge, CB2 0QQ, UK. [17]Department of Neurosurgery, Tel-Aviv Medical Center, Tel-Aviv, Israel. [18]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Catalonia, Spain. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. [20]Institute of Translational Medicine, University of Liverpool, Liverpool, L9 7LJ, UK. [21]Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands. [22]Present address: AstraZeneca, CRUK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, UK. Grace Collord and Patrick Tarpey contributed equally. Correspondence and requests for materials should be addressed to T.S. (email: ts381@cam.ac.uk) or U.M. (email: um1@sanger.ac.uk)

Meningiomas arise from arachnoidal cells of the meninges and are classified as grade I (80% of cases), grade II (10–20%) or grade III (1–3%). Grade III meningiomas comprise papillary, rhabdoid and anaplastic histological subtypes, with anaplastic tumors accounting for the vast majority of grade III diagnoses[1,2]. Nearly half of anaplastic meningiomas represent progression of a previously resected lower grade tumor, whereas the remainder arise *de novo*[3,4]. Recurrence rates are 5–20% and 20–40%, respectively, for grade I and II tumors[2,5]. By contrast, the majority of anaplastic meningioma patients suffer from inexorable recurrences with progressively diminishing benefit from repeated surgery and radiotherapy and 5-year overall survival of 30–60%[4,6].

A recent study of 775 grade I and grade II meningiomas identified five molecular subgroups defined by driver mutation profile[7]. In keeping with previous smaller studies, mutually exclusive mutations in *NF2* and *TRAF7* were the most frequent driver events, followed by mutations affecting key mediators of PI3K and Hedgehog signalling[7,8]. Recurrent hotspot mutations were also identified in the catalytic unit of RNA polymerase II (*POLR2A*) in 6% of grade I tumors[7]. More recently, a study comparing benign versus *de novo* atypical (grade II) meningiomas found the latter to be significantly associated with *NF2* and *SMARCB1* mutations[9]. Atypical meningiomas were further defined by DNA and chromatin methylation patterns consistent with upregulated PRC2 activity, aberrant Homeobox domain methylation and transcriptional dysregulation of pathways involved in proliferation and differentiation[9].

Despite the high mortality rate of anaplastic meningiomas, efforts to identify adjuvant treatment strategies have been hampered by a limited understanding of the distinctive molecular features of this aggressive subtype. A recent analysis of meningioma methylation profiles identified distinct subgroups within Grade III tumors predictive of survival outcomes, though the biology underpinning these differences and any therapeutic implications remain unknown[10]. Here, we present an analysis of the genomic, transcriptional and DNA methylation patterns defining anaplastic meningioma. Our results reveal molecular hallmarks of aggressive disease and suggest novel approaches to risk stratification and targeted therapy.

Results

Overview of the genomic landscape of primary and recurrent anaplastic meningioma. We performed whole genome sequencing (WGS) on a discovery set of 19 anaplastic meningiomas resected at first presentation ('primary'). A subsequent validation cohort comprised 31 primary tumors characterised by targeted sequencing of 366 cancer genes. We integrated genomic findings with RNA sequencing and methylation array profiling in a subset of samples (Supplementary Table S1). Somatic copy number alterations and rearrangements were derived from whole genome sequencing reads, with RNA sequences providing corroborating evidence for gene fusions. Given the propensity of anaplastic meningioma to recur, we studied by whole genome sequencing 13 recurrences from 7 patients.

Excluding a hypermutated tumor (PD23359a, see Supplementary Discussion), the somatic point mutation burden of primary anaplastic meningioma was low with a median of 28 somatic coding mutations per tumor (range 11 to 71; mean sequencing coverage 66X) (Supplementary Fig. S1). Mutational signatures analysis of substitutions identified in whole genome sequences revealed the age-related, ubiquitous processes 1 and 5 as the predominant source of substitutions (Supplementary Fig. S2)[11]. The rearrangement landscape was also relatively quiet, with a median of 12 structural rearrangements (range 0–79) in the 18 primary tumor genomes (Supplementary Fig. S3, Table S3). Somatic retrotransposition events, a significant source of structural variants in over half of human cancers, were scarce (Supplementary Fig. S4, Table S4)[12]. Analysis of expressed gene fusions did not reveal any recurrent events involving putative cancer genes (Supplementary Table S5).

Recurrent large copy number changes were in keeping with known patterns in aggressive meningiomas, notably frequent deletions affecting chromosomes 1p, 6q, 14 and 22q (Fig. 1b, Supplementary Table S6)[7,9,13].

Driver genes do not delineate subgroups of anaplastic meningioma. Over 80% of low grade meningiomas segregate into 5 distinct subgroups based on driver mutation profile[7,9]. In anaplastic meningioma, however, we found a more uniform driver landscape dominated by deleterious mutations in *NF2* (Fig. 1a). A key feature distinguishing anaplastic meningioma from its lower grade counterparts were driver events in genes of the SWI/SNF chromatin regulatory complex (Fig. 1a; Supplementary Fig. S7). The SWI/SNF (mSWI/SNF or BAF) complex is the most commonly mutated chromatin-regulatory complex in cancer[14,15], and acts as a tumor suppressor in many cell types by antagonising the chromatin modifying PRC2[16–18]. The most frequently mutated SWI/SNF component was *ARID1A*, which harbored at least one deleterious somatic change in 12% of our cohort of 50 primary tumors (Supplementary Table S1). *ARID1A* has not been implicated as a driver in grade I or grade II meningiomas[7,9]. Single variants in *SMARCB1*, *SMARCA4* and *PBRM1* were also detected in three tumors (Supplementary Fig. S7). In total, 16% of anaplastic meningiomas contained a damaging SWI/SNF gene mutation. By contrast, SWI/SNF genes are mutated in <5% of benign and atypical meningiomas[7,9].

In the combined cohort of 50 primary tumors, we found at least one driver mutation in *NF2* in 70%, similar to the prevalence reported in atypical meningiomas and more than twice that found in grade I tumors[7,9]. As observed in other cancer types, it is possible that non-mutational mechanisms may contribute to *NF2* loss of function in a proportion of anaplastic meningiomas[19,20]. We considered promoter hypermethylation as a source of additional *NF2* inactivation, but found no evidence of this (Supplementary Table S7). There was no significant difference in NF2 expression between *NF2* mutant and wild-type tumors (*p*-value 0.960; Supplementary Fig. S8), suggesting that a truncated dysfunctional protein may be expressed.

Other driver genes commonly implicated in low grade tumors were not mutated, or very infrequently (Fig. 1a). Furthermore, and consistent with the most recent reports[7,9], we did not observe an increased frequency of *TERT* promoter mutations, previously associated with progressive or high grade tumors[21]. Notably[13], methylation analysis revealed *CDKN2A* and *PTEN* promoter hypermethylation in 17% and 11% of primary tumors, respectively (Fig. 1a). We did not find evidence of novel cancer genes in our cohort, applying established methods

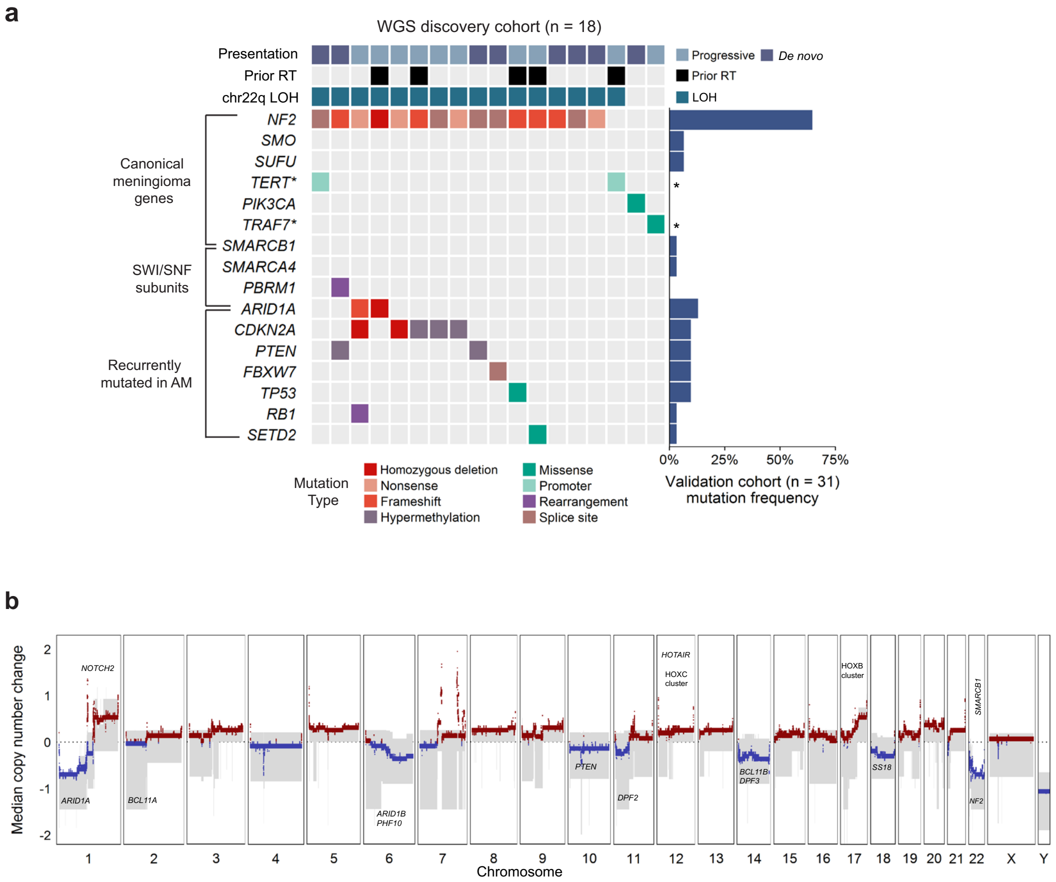


Figure 1. The landscape of driver mutations and copy number alterations in anaplastic meningioma. (**a**) The landscape of somatic driver variants in primary anaplastic meningioma. Somatic mutation and promoter methylation data is shown for a discovery cohort of 18 primary tumors characterised by whole genome sequencing. Mutations in recurrently altered genes, established meningioma genes and SWI/SNF complex subunits are included. Samples are annotated for chromosome 22q LOH, prior radiotherapy exposure, and clinical presentation (*de novo* verus progression from a lower grade meningioma). The bar plot to the right indicates mutation frequency in a validation cohort of 31 primary tumors sequenced with a 366 cancer gene panel. Asterisks indicate genes not included in the targeted sequencing assay. (**b**) Aggregate copy number profile of primary anaplastic meningioma. For the 18 tumors characterized by whole genome sequencing, the median relative copy number change was calculated across the genome in 10 kilobase segments, adjusting for ploidy. The grey shaded area indicates the first and third quantile of copy number for each genomic segment. The solid red and blue lines represent the median relative copy number gain and loss, respectively, with zero indicating no copy number change. X-axis: Chromosomal position. Y-axis: median relative copy number change. Potential target genes are noted. AM, anaplastic meningioma; LOH, loss of heterozygosity; RT, radiotherapy.

to search for enrichment of non-synonymous mutations[22]. The full driver landscape of anaplastic meningioma, considering point mutations, structural variants with resulting copy number changes and promoter hypermethylation is presented in Supplementary Fig. S7.

The genomic landscape of recurrent tumors was largely static both with respect to driver mutations and structural variation. Driver mutations differed between primary and recurrent tumors for only two of eleven patients with serial resections available. For seven sets of recurrent tumors studied by whole genome sequencing, only two demonstrated any discrepancies in large copy number variants (PD23344 and PD23346; Supplementary Fig. S5). Similarly, matched primary and recurrent samples clustered closely together by PCA of transcriptome data, suggesting minimal phenotypic evolution (Supplementary Fig. S6).

Differential gene expression defines anaplastic meningioma subgroups with prognostic and biological significance. We performed messenger RNA (mRNA) sequencing of 31 anaplastic meningioma samples from a total of 28 patients (26 primary tumors and 5 recurrences). Gene expression variability within the cohort did not correlate with clinical parameters including prior radiotherapy, anatomical location or clinical presentation (*de novo* versus progressive tumor) (Supplementary Fig. S6). However, unsupervised hierarchical clustering demonstrated segregation of tumors into two main groups, hereafter referred to as C1 and C2 (Fig. 2a). These groups were recapitulated by principal component analysis (PCA) of normalised transcript counts (Fig. 2b), which delineated C1 as a well-demarcated cluster clearly defined by the first two principal components

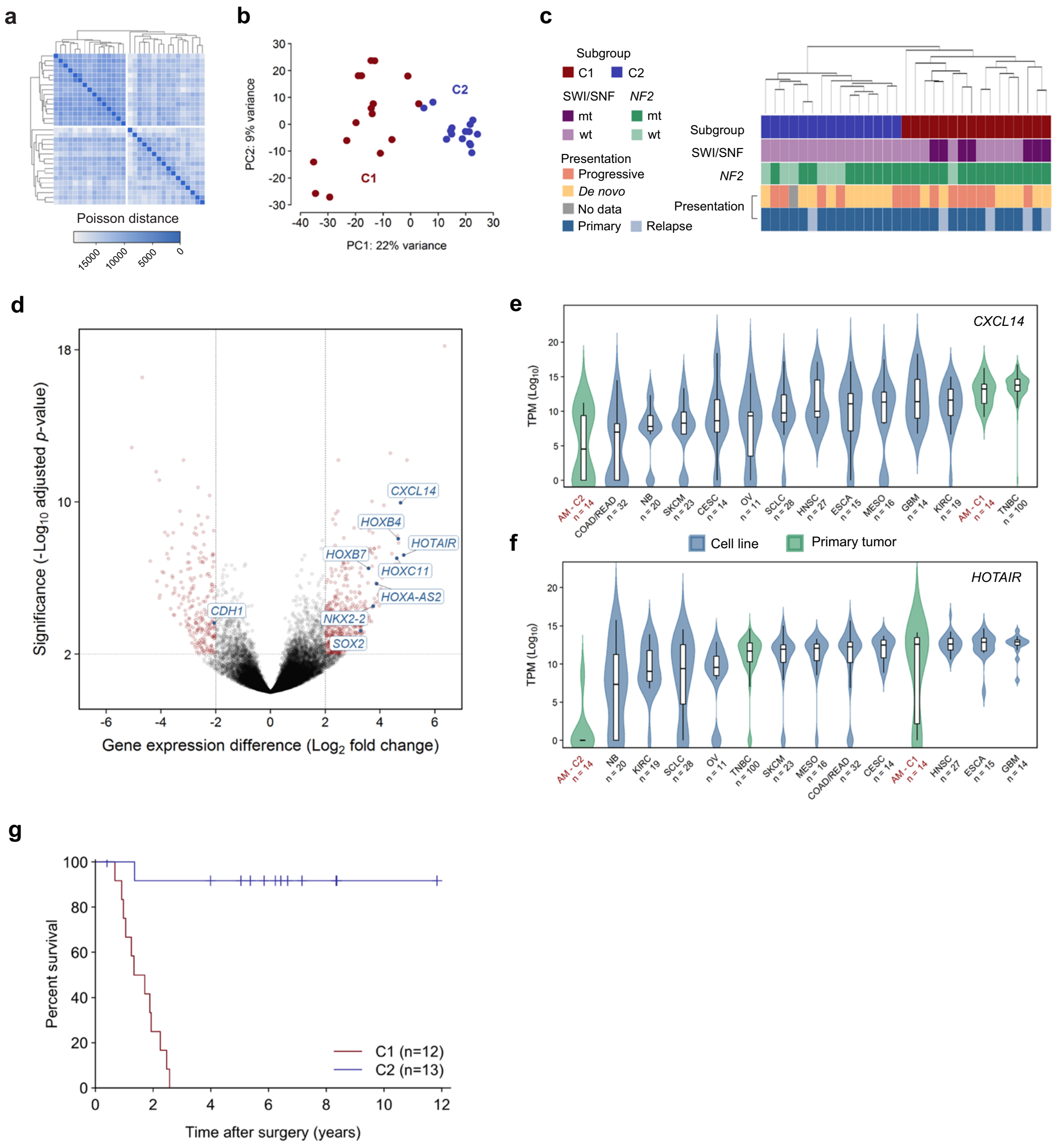


Figure 2. Transcriptomic classification of anaplastic meningioma. (**a**) Unsupervised hierarchical clustering and (**b**) principal component analysis of anaplastic meningioma gene expression revealed two subgroups (denoted C1 and C2). (**c**) Dendrogram obtained by unsupervised clustering annotated with clinical and genomic features. (**d**) Volcano plot depicting genes differentially expressed between C1 versus C2 anaplastic meningioma samples. X-axis, $\log_2$ fold change; y-axis, $-\log_{10}$ adjusted *P*-value. Genes with an adjusted *P*-value < 0.01 and absolute $\log_2$ fold change >2 are highlighted in red. (**e**,**f**) Box plots of (**e**) CXLC14 and (**f**) HOTAIR expression across 31 anaplastic meningomas classified into C1 and C2 subgroups, 100 primary breast tumors, and 219 cancer cell lines from 11 tumor types. Upper and lower box hinges correspond to first and third quartiles, horizontal line and whiskers indicate the median and 1.5-fold the interquartile range, respectively. Underlying violin plots show data distribution and are color-coded according to specimen source (blue, cell line; green, primary tumor). X-axis indicates tumor type and number of samples; y-axis shows $\log_{10}$ TPM values. (**g**) Kaplan-Meier curves showing overall survival for 25 anaplastic meningioma patients in C1 and C2 subgroups for whom follow-up data was available. Dashes indicate timepoints at which subjects were censored at time of last follow-up. TPM, transcripts per kilobase million; AM, anaplastic meningioma; TNBC, triple negative breast carcinoma; wt, wild-type; mt, mutated; PC, principal component.

(PC). Of note, all SWI/SNF mutations were confined to the poor prognosis (C1) subgroup (Fig. 2c). C1 constituted a more diffuse group on PCA, distinguished from C2 mainly along the first principal component. We next retrospectively sought follow-up survival data from the time of first surgery, which was available for 25 of the 28 patients included in the transcriptome analysis (12 patients in C1, 13 in C2; mean follow-up of 1,403 days from surgery). We observed a significantly worse overall survival outcome in C1 compared to C2 ($P < 0.0001$; hazard ratio 17.0, 95% CI 5.2–56.0) (Fig. 2g; Supplementary Table S8). The subgroups were well balanced with respect to potential confounding features such as gender, age, radiotherapy, anatomical location and amount of residual tumor remaining after surgery (Supplementary Table S9).

Recent work has demonstrated that anaplastic meningiomas segregate into 2–3 prognostically significant subgroups on the basis of methylation profile[10]. Unsupervised hierarchical clustering using methylation data available for a subset of the cohort (n = 19) demonstrated segregation into two main groups largely overlapping the subgroups delineated on the basis of gene expression profile, though correlation with survival outcomes was less marked (Supplementary Fig. S8).

Transcriptional programs segregating indolent and aggressive anaplastic meningioma. Nineteen hundred genes underpinned the differentiation of anaplastic meningioma into subgroups C1 and C2, which could be reduced to only 6 transcripts selected on the basis of PCA coefficient and differential expression analysis (see Methods; Supplementary Tables S10 and S11, Fig. S9). Pathway enrichment analysis was most significant for evidence of epithelial-mesenchymal transition (EMT) in the C1 tumors, with concordant loss of E-cadherin (*CDH1*) and upregulation of *CXCL14*, both prognostic biomarkers in diverse other cancers (Supplementary Table S12, Fig. 2d–f)[23–25]. EMT, which involves reprogramming of adherent epithelial cells into migratory mesenchymal cells, is critical for embryogenesis and tissue plasticity, and can play an important role in malignant progression, metastasis and therapy resistance[24,26]. Interestingly, NF2 and the closely related cytoskeletal protein ezrin normally help maintain E-cadherin expression at adherence junctions, whereas *HOXB7* and *HOXB9*, both overexpressed in C1 tumors, suppress *CDH1* expression[27–29]. It is increasingly recognised that CXCL14 and other EMT mediators are often derived from cancer-associated fibroblasts (CAFs) and function in a paracrine manner[25,30,31]. It is hence possible that some of the gene expression patterns we observed may reflect differences in the tumor stromal compartment, itself an increasingly recognised therapeutic target[30,32,33].

The C1 tumors were further characterised by upregulation of transcriptional programs associated with increased proliferation, PRC2 activity and stem cell phenotype (Supplementary Table S13). Hox genes constituted a notable proportion of the transcripts distinguishing the two anaplastic meningioma subgroups, largely underpinning the significance of pathways involved in tissue morphogenesis. Furthermore, differentially methylated genes were also significantly enriched for Hox genes, with pathway analysis results corroborating the main biological themes apparent from the transcriptome (Supplementary Tables S14 and S15). Given the transcriptional evidence of increased PRC2 activity in the C1 subgroup, is noteworthy that SWI/SNF gene mutations occurred exclusively in C1 tumors ($P = 0.016$, Fisher's exact test).

Comparison of the anaplastic and benign meningioma transcriptome. Previous studies investigating the relationship between meningioma WHO grade and gene expression profiles have included few anaplastic tumors[34,35]. We therefore extended our analysis to include published RNA sequences from 19 benign grade I meningiomas. External data was processed using our in-house pipeline with additional measures taken to minimise batch effects (Methods, Supplementary Tables S16 and S17). Unsupervised hierarchical clustering and principal component analysis demonstrated clear tumor segregation by histological grade (Fig. 3a,b). In keeping with previous reports, the anaplastic tumors demonstrated marked upregulation of major growth factor receptor and kinase circuits implicated in meningioma pathogenesis, notably epidermal growth factor receptor (EGFR), insulin-like growth factor (IGFR), vascular endothelial growth factor receptor (VEGFR) and mTOR complex 1 (mTORC1) kinase complex[36–41].

Consistent with there being a coherent biological trend across histological grades and anaplastic meningioma subgroups, we noted significant overlap between genes differentially expressed between grades and between C1 and C2 tumors (hypergeometric distribution $P = 5.08 \times 10^{-9}$). In keeping with this finding, formal pathway analysis identified significant dysregulation of stemness, proliferation, EMT and PRC2 activity (Supplementary Tables S18 and S19). The most significantly dysregulated pathways also included TGF-beta, Wnt and integrin signalling, mediators of invasion and mesenchymal differentiation that are normally in part controlled by NF2 and other Hippo pathway members[20,24,42]. Yes-associated protein 1 (Yap1), a cornerstone of oncogenic Hippo signalling, is frequently overexpressed in cancer and synergises with Wnt signalling to induce EMT[43,44]. *YAP1* was upregulated in anaplastic tumors along with *MYL9*, a key downstream effector essential for Yap1-mediated stromal reprogramming (Fig. 3c)[43].

Discussion

Meningiomas constitute a common, yet diverse tumor type with few therapeutic options[6,7,9,45]. Efforts to improve clinical outcomes have been hampered by limited understanding of the molecular determinants of aggressive disease. Here, we explored genomic, epigenetic and transcriptional features of anaplastic meningioma, the most lethal meningioma subtype[4].

Frequent somatic changes in SWI/SNF complex genes, predominantly *ARID1A*, constitute the main genomic distinction between anaplastic and lower grade meningiomas[7,9]. SWI/SNF inactivation is associated with aberrant PRC2 activation, stem cell-like phenotype and poor outcomes in diverse cancer types[46–48].

Although anaplastic tumors resist comprehensive classification based on driver mutation patterns, transcriptional profiling revealed two biologically distinct subgroups with dramatically divergent survival outcomes. This finding is emblematic of the limitations of histopathological grading as a risk stratification system for meningioma[2,4,10,45,49]. All SWI/SNF mutations were confined to the poor prognosis (C1) subgroup, which was further characterised by transcriptional signatures of PRC2 target activation, stemness, proliferation and mesenchymal differentiation. These findings were in part underpinned by differential expression of Hox genes. Acquisition of invasive capacity and stem cell traits are frequently co-ordinately dysregulated in cancer, often through subversion of Hox gene programs integral to normal tissue morphogenesis[50–52]. Hox genes have a central role in orchestrating vertebrate development and act as highly context-dependent oncogenes and tumor suppressors in cancer[51,53].

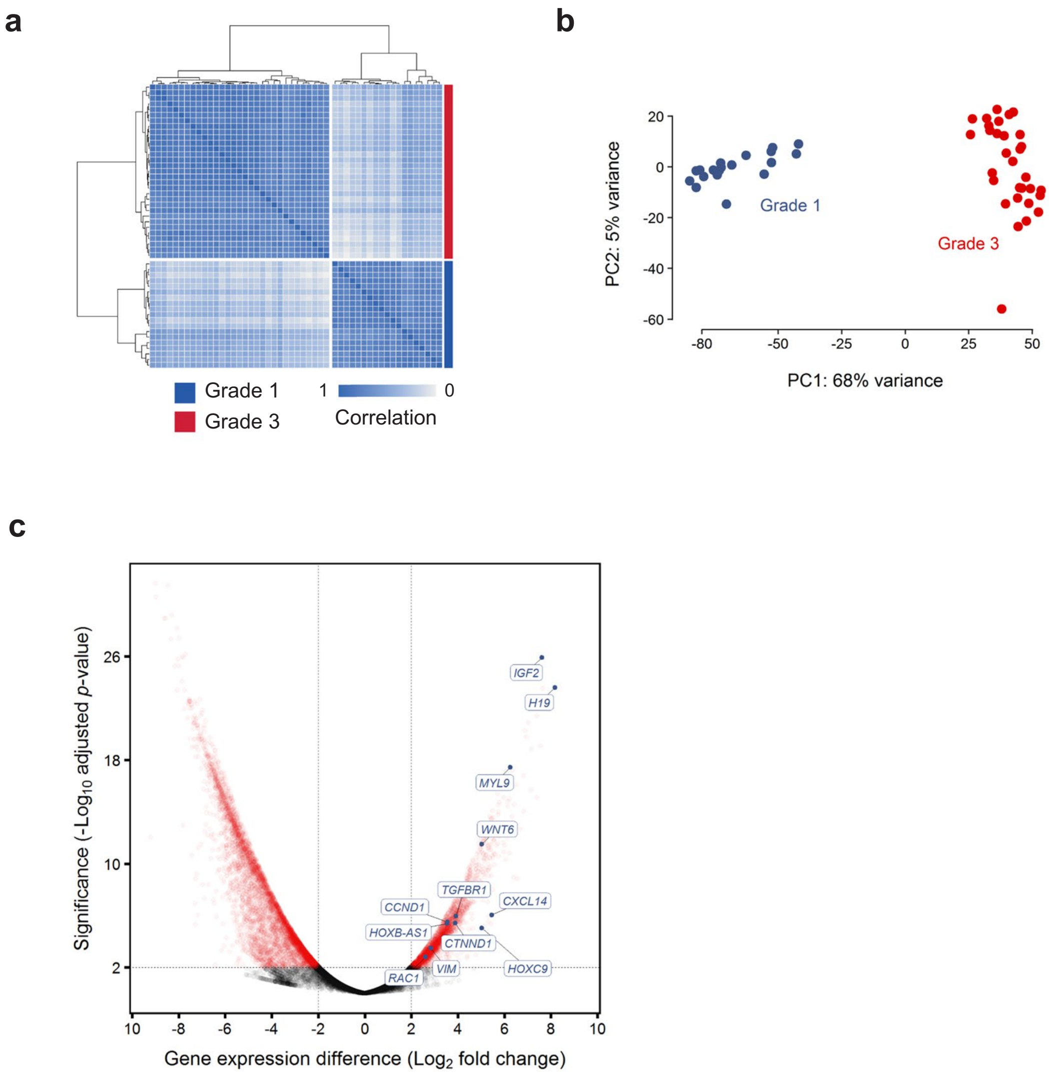


Figure 3. Differences in gene expression profile between grade I and anaplastic meningomas. (**a**,**b**) Normalised transcript counts from grade I and anaplastic meningioma samples clustered by (**a**) Pearson's correlation coefficient and (**b**) principal component analysis. (**c**) Volcano plot illustrating differences in gene expression between anaplastic versus grade I meningiomas with selected genes indicated. The horizontal axis shows the $\log_2$ fold change and the vertical axis indicates the $-\log_{10}$ adjusted *P*-value. Genes with an adjusted *P*-value < 0.01 and absolute $\log_2$ fold change >2 are highlighted in red. PC, principal component.

Several of the most starkly upregulated Hox genes in the C1 tumors consistently function as oncogenes across a range of solid and haematological malignancies, including *HOTAIR*, *HOXB7*, *HOXA4*, *HOXA-AS2*, *HOXC11*, and *NKX2-2*[28,29,51,54–62]. Like many other long non-coding RNAs (lncRNA), *HOTAIR* and *HOXA-AS2* modulate gene expression primarily by interacting directly with chromatin remodelling complexes, exerting oncogenic activity by recruiting PRC2 to target genes[54,56,61–65]. *HOXA-AS2* has been shown to mediate transcriptional repression of the tumor suppressor gene *CDKN2A* (p16$^{\mathrm{INK4A}}$), deletion of which is associated with poor meningioma survival[54,61,62,66,67]. Given the antagonistic relationship between the SWI/SNF and PRC2 chromatin regulators, deleterious SWI/SNF mutations and overexpression of lncRNAs known to mediate PRC2 activity emerge as potentially convergent mechanisms underpinning the differences between C1 and C2 tumors[68]. Further endorsing a link between transcriptional subgroups and chromatin dysregulation, 15 of the differentially expressed transcripts delineating C1 and C2 subgroups (absolute $\log_2$ fold change >2 and FDR < 0.01) are among the 50 genes most often associated with frequently bivalent chromatin segments (FBS) in cancer, including 11 transcripts from the HOXB cluster on chromosome 17[69]. This overlap was highly statistically significant (hypergeometric distribution $P = 1.98 \times 10^{-11}$). Bivalent, or epigenetically 'poised', chromatin is characterised by finely balanced activating (H3K4me1/H3K4me3) and repressive (H3K27me3) histone marks and pre-loaded DNA polymerase II poised to transcribe in response to modest epigenetic changes[70]. Bivalent chromatin most often marks genes involved in developmental reprogramming, in particular Hox cluster genes and homeotic non-coding transcripts, and is a frequent target of aberrant chromatin modification in cancer[65,69,71].

In the context of recent studies of lower grade meningiomas, our findings raise the possibility that the balance between PRC2 and SWI/SNF activity may have broader relevance to meningioma pathogenesis. Compared to grade I tumors, atypical meningiomas are more likely to harbor *SMARCB1* mutations and large deletions encompassing chromosomes 1q, 6q and 14q. Notably, these genomic regions encompass *ARID1A* and several other SWI/SNF subunit genes. Both *SMARCB1* mutations and the aforementioned copy number changes were associated with epigenetic evidence of increased PRC2 activity, differential Homeobox domain methylation, and upregulation of proliferation and stemness programs in atypical grade II meningiomas[9].

The extent to which SWI/SNF depletion plays a role in meningioma development may be therapeutically relevant. Diverse SWI/SNF mutated cancers exhibit dependence on both catalytic and non-catalytic functions of EZH2, a core subunit of PRC2[72–74]. Several EZH2 inhibitors are in development with promising initial clinical results[75]. Other modulators of PRC2 activity, including *HOTAIR*, may also be relevant therapeutic targets[76,77]. Furthermore, growing recognition of the relationship between EMT and resistance to conventional and targeted anti-cancer agents has profound implications for rational integration of treatment approaches[32,33]. Notably, EGFR inhibition has yielded disappointing response rates in meningioma despite high EGFR expression[37,78]. A mesenchymal phenotype is strongly associated with resistance to EGFR inhibitors in lung and colorectal cancer[32,33,79–81]. Combining agents that abrogate EMT with other therapies is a promising strategy for addressing cell-autonomous and extrinsic determinants of disease progression and may warrant further investigation in meningioma[32,33].

This study has revealed biologically and prognostically significant anaplastic meningioma subgroups and identified potentially actionable alternations in SWI/SNF genes, PRC2 activity and EMT regulatory networks. However, a substantially larger series of tumors, ideally nested in a prospective multicentre observational study, will be required to expand upon our main findings and explore mechanistic and therapeutic ramifications of meningioma diversity.

Methods

Sample selection.

DNA was extracted from 70 anaplastic meningiomas; 51 samples at first resection ('primary') and 19 from subsequent recurrences. Matched normal DNA was derived from peripheral blood lymphocytes. Written informed consent was obtained for sample collection and DNA sequencing from all patients in accordance with the Declaration of Helsinki and protocols approved by the NREC/Health Research Authority (REC reference 7/YH/0101) and Ethics Committee at University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany (EK 323122008). Samples underwent independent specialist pathology review (V.P.C and K.A). DNA extracted from fresh-frozen material was submitted for whole genome sequencing whereas that derived from formalin-fixed paraffin-embedded (FFPE) material underwent deep targeted sequencing of 366 cancer genes.

One tumor sample PD23348 (and two subsequent recurrences) separated from the main study samples in a principal components analysis of transcriptomic data (Supplementary Fig. S10). Analysis of WGS and RNA sequencing data identified an expressed gene fusion, *NAB2-STAT6*. This fusion is pathognomonic of meningeal hemangiopericytoma, now classified as a separate entity, solitary fibrous tumors[82–84]. We therefore excluded three samples from this tumor from further study. A second sample (PD23354a), diagnosed as an anaplastic meningioma with papillary features, was found to have a strong APOBEC mutational signature as well as an *EML4-ALK* gene fusion (exon 6 EML4, exon 19 ALK) (Supplementary Fig. S11)[85]. Therefore this sample was also removed as a likely metastasis from a primary lung adenocarcinoma. The hypermutator sample PD23359a underwent additional pathological review to confirm the diagnosis of anaplastic meningioma (K.A., Department of Histopathology, Cambridge University Hospital, Cambridge, UK).

RNA was extracted from fresh-frozen material from 34 primary and recurrent tumors, 3 of which were from PD23348 and were subsequently excluded from final analyses (Supplementary Table S1).

Whole genome sequencing. Short insert 500 bp genomic libraries were constructed, flowcells prepared and sequencing clusters generated according to Illumina library protocols[86]. 108 base/100 base (genomic), or 75 base (transcriptomic) paired-end sequencing were performed on Illumina X10 genome analyzers in accordance with the Illumina Genome Analyzer operating manual. The average sequence coverage was 65.8X for tumor samples and 33.8X for matched normal samples (Supplementary Table S1).

Targeted genomic sequencing.

For targeted sequencing we used a custom cRNA bait set (Agilent) to enrich for all coding exons of 366 cancer genes (Supplementary Table S20). Short insert libraries (150 bp) were prepared and sequenced on the Illumina HiSeq 2000 using 75 base paired-end sequencing as per Illumina protocol. The average sequence coverage was 469X for the tumor samples.

RNA sequencing and data processing.

For transcriptome sequencing, 350 bp poly-A selected RNA libraries were prepared on the Agilent Bravo platform using the Stranded mRNA library prep kit from KAPA Biosystems. Processing steps were unchanged from those specified in the KAPA manual except for use of an in-house indexing set. Reads were mapped to the GRCh37 reference genome using STAR (v2.5.0c)[87]. Mean sequence coverage was 128X. Read counts per gene, based on the union of all exons from all possible transcripts, were then extracted BAM files using HTseq (v0.6.1)[88]. Transcripts Per kilobase per Million reads (TPM) were generated using an in-house python script (https://github.com/TravisCG/SI_scripts/blob/master/tpm.py)[87,88]. We downloaded archived RNA sequencing FASTQ files for 19 grade I meningioma samples representing the major mutational groups (*NF2*/chr22 loss, *POLR2A*, *KLF4*/*TRAF7*, *PI3K* mutant) (ArrayExpress: GSE85133)[7]. Reads were then processed using STAR and HTseq as described above. Cancer cell line (n = 252) and triple-negative breast cancer (n = 100) RNA sequencing data was generated in-house by the aforementioned sequencing and bioinformatic pipeline.

Expressed gene fusions were sought using an in-house pipeline incorporating three algorithms: TopHat-Fusion (v2.1.0), STAR-Fusion (v0.1.1) and deFuse (v0.7.0) (https://github.com/cancerit/cgpRna)[87,89,90]. Fusions identified by one or two algorithms or also detected in the matched normal sample were flagged as likely artefacts. Fusions were further annotated according to whether they involved a kinase or known oncogene and whether they occurred near known fragile sites or rearrangement break points[91] (Supplementary Table S5).

The C1 and C2 subgroups were defined by unsupervised hierarchical clustering using Poisson distance between samples[92,93]. Poisson distance was calculated using the PoissonDistance function implemented in the 'PoiClaClu' R package[92] and unsupervised hierarchical clustering performed with the stats::hclust() function using the 250 transcripts with the most variable expression across tumors. Silhouette information was computed using the cluster::silhouette() function. The highest mean silhouette score was consistently achieved with two clusters.

Differential gene expression and pathway enrichment analysis. The DESeq2 R package was used for all differential gene expression analyses[94,95]. DESeq2 uses shrinkage estimation of dispersion for the sample-specific count normalization and subsequently applies a linear regression method to identify differentially expressed genes (DEGs)[94,95].

Preliminary comparison of anaplastic and externally-generated grade I meningioma data revealed evidence of laboratory batch effects, which we mitigated with two batch-correction methods: RUVg and PEER[96,97]. RUVg estimates the factor attributed to spurious variation using control genes that are assumed to have constant expression across samples[98–100]. We selected control genes (*RPL37A*, *EIF2B1*, *CASC3*, *IPO8*, *MRPL19*, *PGK1* and *POP4*) on the basis of previous studies of suitable control genes for transcript-based assays in meningioma[101]. PEER ('probabilistic estimation of expression residuals') is based on factor analysis methods that infer broad variance components in the measurements. PEER can find hidden factors that are orthogonal to the known covariates. We applied this feature of PEER to remove additional hidden effect biases. The final fitted linear regression model consists of the factor identified by RUVg method that represents the unwanted laboratory batch effect and 13 additional hidden factors found by PEER that are orthogonal to the estimated laboratory batch effect. Using this approach we were able to reduce the number of DEGs from more than 18000 to 8930, of which <4,000 are predicted to be protein-coding.

To identify biological pathways differentially expressed between meningioma grades and anaplastic meningioma subgroups we applied a functional class scoring algorithm using a collection of 461 published gene sets mapped to 10 canonical cancer hallmarks (Supplementary Table S21)[50,102–106]. We further corroborated these findings with a more general Gene Ontology (GO) pathway analysis[107].

Identification of 6 transcripts recapitulating anaplastic meningioma clusters. Mapped RNA sequencing reads were normalised using the regularised logarithm (rlog) function implemented by the DESeq2 package[94,95]. PCA was performed using the top 500 most variably expressed transcripts and the R stats::prcomp function[108]. Given that primary component 1 (PC1) was the vector most clearly distinguishing the closely clustered C2 subgroup from the more diffusely clustered C1 (Fig. 3a), we extracted the top 50 transcripts with the highest absolute PC1 coefficients. We then identified the subset that overlapped with the most significantly differentially expressed genes (absolute $\log_2$ fold change >4 and adjust p-value < 0.0001) between i) the C1 and C2 anaplastic meningioma subgroups and ii) the C1 anaplastic meningiomas and the 19 grade I tumors (Supplementary Tables S10 and S17). Iteratively reducing the number of PC1 components identified the minimum number of transcripts that recapitulated segregation of C1 and C2 tumors upon unsupervised hierarchical clustering and PCA (Supplementary Table S11, Fig. S9).

Processing of genomic sequencing data. Genomic reads were aligned to the reference human genome (GRCh37) using the Burrows-Wheeler Aligner, BWA (v0.5.9)[109]. CaVEMan (Cancer Variants Through Expectation Maximization: http://cancerit.github.io/CaVEMan/) was used for calling somatic substitutions. Small insertions and deletions (indels) in tumor and normal reads were called using a modified Pindel version 2.0. (http://cancerit.github.io/cgpPindel/) on the NCBI37 genome build[110,111]. Annotation was according to ENSEMBL version 58. Structural variants were called using a bespoke algorithm, BRASS (BReakpoint AnalySiS) (https://github.com/cancerit/BRASS) as previously described[112].

The ascatNGS algorithm was used to estimate tumor purity and ploidy and to construct copy number profiles from whole genome data[113].

Identification of cancer genes based on the impact of coding mutations. To identify recurrently mutated driver genes, we applied an established dN/dS method that considers the mutation spectrum, the sequence of each gene, the impact of coding substitutions (synonymous, missense, nonsense, splice site) and the variation of the mutation rate across genes[22].

Identification of driver mutations in known cancer genes. Non-synonymous coding variants detected by Caveman and Pindel algorithms were flagged as putative driver mutations according to set criteria and further curated following manual inspection in the Jbrowse genome browser[114]. Variants were screened against lists of somatic mutations identified by a recent study of 11,119 human tumors encompassing 41 cancer types and also against a database of validated somatic drivers identified in cancer sequencing studies at the Wellcome Trust Sanger Institute (Supplementary Tables S22 and S23)[115].

Copy number data was analysed for homozygous deletions encompassing tumor suppressor genes and for oncogene amplifications exceeding 5 or 9 copies for diploid and tetraploid genomes, respectively. Only focal (<1 Mb) copy number variants meeting these criteria were considered potential drivers. Additional truncating events (disruptive rearrangement break points, nonsense point mutations, essential splice site mutations and frameshift indels) in established tumor suppressors were also flagged as potential drivers. Only rearrangements with breakpoints able to be reassembled at base pair resolution are included in this dataset.

TraFiC pipeline for retrotransposon integration detection. For the identification of putative solo-L1 and L1-transduction integration sites, we used the TraFiC (Transposome Finder in Cancer) algorithm[12]. TraFiC uses paired-end sequencing data for the detection of somatic insertions of transposable elements (TEs) and exogenous viruses. The identification of somatic TEs (solo-L1, Alu, SINE, and ERV) is performed in three steps: (i) selection of candidate reads, (ii) transposable element masking, (iii) clustering and prediction of TE integration sites and (iv) filtering of germline events[12].

Methylation arrays and analysis. We performed quantitative methylation analysis of 850,000 CpG sites in 25 anaplastic meningiomas. Bisulfite-converted DNA (bs-DNA) was hybridized on the Ilumina Infinium HumanMethylationEPIC BeadChip array following the manufacturer's instructions. All patient DNA samples were assessed for integrity, quantity and purity by electrophoresis in a 1.3% agarose gel, picogreen quantification and Nanodrop measurements. Bisulfite conversion of 500 ng of genomic DNA was done using the EZ DNA Methylation Kit (Zymo Research), following the manufacturer's instructions. Resulting raw intensity data (IDATs) were normalized using the Illumina normalization method developed under the minfi R package (v1.19.10). Normalized intensities were then used to calculate DNA methylation levels (beta values). We then excluded from the analysis the positions with background signal levels in methylated and unmethylated channels ($p > 0.01$). Finally we removed probes with one or more single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) >1% in the first 10 bp of the interrogated CpG, as well as the probes related to X and Y chromosomes. From the filtered positions, we selected only CpG sites present both in promoter regions (TSS, 5′UTR and 1st exon) and CpG islands (UCSC database, genome version hg19).

For the supervised analysis of the probes, CpG sites were selected by applying an ANOVA test to identify statistically significant CpG positions (FDR adjusted p-value < 0.01) that were differentially methylated among the compared groups ($\Delta\beta > 0.2$). Selected CpG sites were later clustered based on the Manhattan distances aggregated by ward's linkage. Finally, the genes corresponding to the selected CpGs were used to perform a Gene Set Enrichment Analysis (GSEA) with curated gene sets in the Molecular Signatures Database[116]. The gene sets used were: H: hallmark gene sets, BP: GO biological process, CC: GO cellular component, MF: GO molecular function and C3: motif gene sets (http://software.broadinstitute.org/gsea/msigdb/collections.jsp). The gene clusters resulting from the hypergeometric test with a FDR adjusted p-value < 0.05 were finally considered. We observed high levels of methylation for *CREBBP* in the majority of tumor samples, however, similar patterns were manifest in normal tissue controls, hence *CREBBP* hypermethyation does not appear to be a feature of oncogenesis in these samples.

For principal component analysis, we used the R function prcomp to calculate the Singular Value Decomposition of the beta value matrix after removing the CpGs without methylation information. We plotted the first two principal components which contain most variation by using the ggbiplot R package (http://github.com/vqv/ggbiplot). For each group we plotted a normal data ellipse with size defined as a normal probability equal to 0.68. Unsupervised hierarchical clustering was performed with the stats::hclust() function using the 75 probes with the highest variance in methylation beta values.

Mutational signature analysis. Mutational signature extraction was performed using the nonnegative matrix factorization (NNMF) algorithm[11]. Briefly, the algorithm identifies a minimal set of mutational signatures that optimally explains the proportions of mutation types found across a given mutational catalogue and then estimates the contribution of each identified signature to the mutation spectra of each sample.

Patient survival analysis. The Kaplan-Meier method was used to analyze survival outcomes by the log-rank Mantel-Cox test, with hazard ratio and two-sided 95% confidence intervals calculated using the Mantel_Haenszel test (GraphPad Prism, ver 7.02). Overall survival data from time of first surgery for each anaplastic meningioma within gene-expression defined subgroups C1 and C2 was collected and used to plot a Kaplan-Meier survival curve.

Supplementary Discussion

A hypermutator anaplastic meningioma with a haploid genome. One primary anaplastic meningioma resected from an 85-year old female (PD23359a) had a hypermutator phenotype, with 27,332 point mutations and LOH across nearly its entire genome (Supplementary Fig. S12, Table S24). Independent pathological review confirmed the original diagnosis of anaplastic meningioma, and transcriptome analysis demonstrated that this tumor clustered appropriately with the rest of the cohort (Fig. 3a,b). The majority of the mutations were substitutions, 72% of which were C > T transitions. We identified two deleterious mutations in DNA damage repair mediators: a *TP53* p.R248Q missense mutation and a homozygous truncating variant in the mismatch repair gene *MSH6* (p.L1330Vfs*9). Despite the latter finding, mutational signatures analysis was dominated by signature 1, with no evidence of signatures typically associated with defects in homologous recombination, mismatch repair or *POLE* activity (signatures 3, 6, 10, 15, 20 or 26). The copy number profile is most consistent with this tumor having first undergone haploidization of its genome, with the exception of chromosomes 7, 19 and 20, followed by whole genome duplication (Supplementary Fig. S12). Of note, several important oncogenes are located on chromosome 7, including *EGFR*, *MET* and *BRAF*. Widespread LOH has been described in a significant proportion of oncocytic follicular thyroid cancers where preservation of chromosome 7 heterozygosity has also been observed[117].

Data Availability

All sequencing data that support the findings of this study have been deposited in the European Genome-Phenome Archive and are accessible through the accession numbers EGAS00001000377, EGAS00001000828, EGAS00001000859, EGAS00001001155 and EGAS00001001873. All other relevant data are available from the corresponding author on request.

References

1. Mawrin, C. & Perry, A. Pathological classification and molecular genetics of meningiomas. *J Neurooncol* **99**, 379–391, https://doi.org/10.1007/s11060-010-0342-2 (2010).
2. Rogers, C. L. *et al.* Pathology concordance levels for meningioma classification and grading in NRG Oncology RTOGTrial 0539. *Neuro Oncol* **18**, 565–574, https://doi.org/10.1093/neuonc/nov247 (2016).
3. Moliterno, J. *et al.* Survival in patients treated for anaplastic meningioma. *Journal of neurosurgery* **123**, 23–30, https://doi.org/10.3171/2014.10.JNS14502 (2015).
4. Champeaux, C., Wilson, E., Brandner, S., Shieff, C. & Thorne, L., World Health Organization. grade III meningiomas. A retrospective study for outcome and prognostic factors assessment. *Br J Neurosurg* **29**, 693–698, https://doi.org/10.3109/02688697.2015.1054350 (2015).
5. Durand, A. *et al.* WHO grade II and III meningiomas: a study of prognostic factors. *J Neurooncol* **95**, 367–375, https://doi.org/10.1007/s11060-009-9934-0 (2009).
6. Buttrick, S., Shah, A. H., Komotar, R. J. & Ivan, M. E. Management of Atypical and Anaplastic Meningiomas. *Neurosurg Clin N Am* **27**, 239–247, https://doi.org/10.1016/j.nec.2015.11.003 (2016).
7. Clark, V. E. *et al.* Recurrent somatic mutations in POLR2A define a distinct subset of meningiomas. *Nat Genet* **48**, 1253–1259, https://doi.org/10.1038/ng.3651 http://www.nature.com/ng/journal/vaop/ncurrent/abs/ng.3651.html - supplementary-information (2016).
8. Clark, V. E. *et al.* Genomic analysis of non-NF2 meningiomas reveals mutations in TRAF7, KLF4, AKT1, and SMO. *Science* **339**, 1077–1080, https://doi.org/10.1126/science.1233009 (2013).
9. Harmanci, A. S. *et al.* Integrated genomic analyses of de novo pathways underlying atypical meningiomas. *Nat Commun* **8**, 14433, https://doi.org/10.1038/ncomms14433 (2017).
10. Sahm, F. *et al.* DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. *Lancet Oncol* **18**, 682–694, https://doi.org/10.1016/S1470-2045(17)30155-9 (2017).
11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421, https://doi.org/10.1038/nature12477 (2013).
12. Tubio, J. M. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343, https://doi.org/10.1126/science.1251343 (2014).
13. Galani, V. *et al.* Genetic and epigenetic alterations in meningiomas. *Clinical neurology and neurosurgery* **158**, 119–125, https://doi.org/10.1016/j.clineuro.2017.05.002 (2017).
14. Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* **45**, 592–601, https://doi.org/10.1038/ng.2628 (2013).
15. Shain, A. H. & Pollack, J. R. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. *PLoS One* **8**, e55119, https://doi.org/10.1371/journal.pone.0055119 (2013).
16. Wu, J. I., Lessard, J. & Crabtree, G. R. Understanding the words of chromatin regulation. *Cell* **136**, 200–206, https://doi.org/10.1016/j.cell.2009.01.009 (2009).
17. Kia, S. K., Gorski, M. M., Giannakopoulos, S. & Verrijzer, C. P. SWI/SNF mediates polycomb eviction and epigenetic reprogramming of the INK4b-ARF-INK4a locus. *Mol Cell Biol* **28**, 3457–3464, https://doi.org/10.1128/MCB.02019-07 (2008).
18. Wilson, B. G. & Roberts, C. W. SWI/SNF nucleosome remodellers and cancer. *Nat Rev Cancer* **11**, 481–492, https://doi.org/10.1038/nrc3068 (2011).
19. Morales, F. C., Molina, J. R., Hayashi, Y. & Georgescu, M. M. Overexpression of ezrin inactivates NF2 tumor suppressor in glioblastoma. *Neuro Oncol* **12**, 528–539, https://doi.org/10.1093/neuonc/nop060 (2010).
20. Petrilli, A. M. & Fernandez-Valle, C. Role of Merlin/NF2 inactivation in tumor biology. *Oncogene* **35**, 537–548, https://doi.org/10.1038/onc.2015.125 (2016).
21. Goutagny, S. *et al.* High incidence of activating TERT promoter mutations in meningiomas undergoing malignant progression. *Brain Pathol* **24**, 184–189, https://doi.org/10.1111/bpa.12110 (2014).
22. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e1021, https://doi.org/10.1016/j.cell.2017.09.042 (2017).
23. Berx, G. & van Roy, F. Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harbor perspectives in biology* **1**, a003129, https://doi.org/10.1101/cshperspect.a003129 (2009).
24. De Craene, B. & Berx, G. Regulatory networks defining EMT during cancer initiation and progression. *Nat Rev Cancer* **13**, 97–110, https://doi.org/10.1038/nrc3447 (2013).
25. Benarafa, C. & Wolf, M. CXCL14: the Swiss army knife chemokine. *Oncotarget* **6**, 34065–34066, https://doi.org/10.18632/oncotarget.6040 (2015).
26. Polyak, K. & Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer* **9**, 265–273, https://doi.org/10.1038/nrc2620 (2009).
27. Pujuguet, P., Del Maestro, L., Gautreau, A., Louvard, D. & Arpin, M. Ezrin regulates E-cadherin-dependent adherens junction assembly through Rac1 activation. *Mol Biol Cell* **14**, 2181–2191, https://doi.org/10.1091/mbc.E02-07-0410 (2003).
28. Hayashida, T. *et al.* HOXB9, a gene overexpressed in breast cancer, promotes tumorigenicity and lung metastasis. *Proc Natl Acad Sci USA* **107**, 1100–1105, https://doi.org/10.1073/pnas.0912710107 (2010).
29. Wu, X. *et al.* HOXB7, a homeodomain protein, is overexpressed in breast cancer and confers epithelial-mesenchymal transition. *Cancer Res* **66**, 9527–9534, https://doi.org/10.1158/0008-5472.CAN-05-4470 (2006).
30. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* **16**, 582–598, https://doi.org/10.1038/nrc.2016.73 (2016).
31. Sjoberg, E., Augsten, M., Bergh, J., Jirstrom, K. & Ostman, A. Expression of the chemokine CXCL14 in the tumour stroma is an independent marker of survival in breast cancer. *Br J Cancer* **114**, 1117–1124, https://doi.org/10.1038/bjc.2016.104 (2016).
32. Gotwals, P. *et al.* Prospects for combining targeted and conventional cancer therapy with immunotherapy. *Nat Rev Cancer* advance online publication, https://doi.org/10.1038/nrc.2017.17 (2017).
33. Marcucci, F., Stassi, G. & De Maria, R. Epithelial-mesenchymal transition: a new target in anticancer drug discovery. *Nat Rev Drug Discov* **15**, 311–325, https://doi.org/10.1038/nrd.2015.13 (2016).
34. Watson, M. A. *et al.* Molecular characterization of human meningiomas by gene expression profiling using high-density oligonucleotide microarrays. *Am J Pathol* **161**, 665–672, https://doi.org/10.1016/s0002-9440(10)64222-8 (2002).
35. Wrobel, G. *et al.* Microarray-based gene expression profiling of benign, atypical and anaplastic meningiomas identifies novel genes associated with meningioma progression. *Int J Cancer* **114**, 249–256, https://doi.org/10.1002/ijc.20733 (2005).

36. Mawrin, C. *et al*. Different activation of mitogen-activated protein kinase and Akt signaling is associated with aggressive phenotype of human meningiomas. *Clin Cancer Res* **11**, 4074–4082, https://doi.org/10.1158/1078-0432.ccr-04-2550 (2005).
37. Mawrin, C., Chung, C. & Preusser, M. Biology and clinical management challenges in meningioma. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Meeting*, e106–115, https://doi.org/10.14694/EdBook_AM.2015.35.e106 (2015).
38. Lopez-Lago, M. A., Okada, T., Murillo, M. M., Socci, N. & Giancotti, F. G. Loss of the tumor suppressor gene NF2, encoding merlin, constitutively activates integrin-dependent mTORC1 signaling. *Mol Cell Biol* **29**, 4235–4249, https://doi.org/10.1128/mcb.01578-08 (2009).
39. James, M. F. *et al*. NF2/merlin is a novel negative regulator of mTOR complex 1, and activation of mTORC1 is associated with meningioma and schwannoma growth. *Mol Cell Biol* **29**, 4250–4261, https://doi.org/10.1128/mcb.01581-08 (2009).
40. Johnson, M. D., Okedli, E., Woodard, A., Toms, S. A. & Allen, G. S. Evidence for phosphatidylinositol 3-kinase-Akt-p7S6K pathway activation and transduction of mitogenic signals by platelet-derived growth factor in meningioma cells. *Journal of neurosurgery* **97**, 668–675, https://doi.org/10.3171/jns.2002.97.3.0668 (2002).
41. Weisman, A. S., Raguet, S. S. & Kelly, P. A. Characterization of the epidermal growth factor receptor in human meningioma. *Cancer Res* **47**, 2172–2176 (1987).
42. Harvey, K. F., Zhang, X. & Thomas, D. M. The Hippo pathway and human cancer. *Nat Rev Cancer* **13**, 246–257, https://doi.org/10.1038/nrc3458 (2013).
43. Calvo, F. *et al*. Mechanotransduction and YAP-dependent matrix remodelling is required for the generation and maintenance of cancer-associated fibroblasts. *Nat Cell Biol* **15**, 637–646, https://doi.org/10.1038/ncb2756 (2013).
44. Rosenbluh, J. *et al*. beta-Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. *Cell* **151**, 1457–1473, https://doi.org/10.1016/j.cell.2012.11.026 (2012).
45. Louis, D. N. *et al*. The2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803–820, https://doi.org/10.1007/s00401-016-1545-1 (2016).
46. Le Loarer, F. *et al*. SMARCA4 inactivation defines a group of undifferentiated thoracic malignancies transcriptionally related to BAF-deficient sarcomas. *Nat Genet* **47**, 1200–1205, https://doi.org/10.1038/ng.3399 (2015).
47. Luchini, C. *et al*. Prognostic role and implications of mutation status of tumor suppressor gene ARID1A in cancer: a systematic review and meta-analysis. *Oncotarget* **6**, 39088–39097, https://doi.org/10.18632/oncotarget.5142 (2015).
48. Lu, C. & Allis, C. D. SWI/SNF complex in cancer. *Nat Genet* **49**, 178–179, https://doi.org/10.1038/ng.3779 (2017).
49. Goldbrunner, R. *et al*. EANO guidelines for the diagnosis and treatment of meningiomas. *Lancet Oncol* **17**, e383–391, https://doi.org/10.1016/S1470-2045(16)30321-7 (2016).
50. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674, https://doi.org/10.1016/j.cell.2011.02.013 (2011).
51. Shah, N. & Sukumar, S. The Hox genes and their roles in oncogenesis. *Nat Rev Cancer* **10**, 361–371, https://doi.org/10.1038/nrc2826 (2010).
52. Xu, Q. *et al*. Long non-coding RNA regulation of epithelial-mesenchymal transition in cancer metastasis. *Cell Death Dis* **7**, e2254, https://doi.org/10.1038/cddis.2016.149 (2016).
53. Krumlauf, R. Hox genes in vertebrate development. *Cell* **78**, 191–201 (1994).
54. Xie, M. *et al*. Long noncoding RNA HOXA-AS2 promotes gastric cancer proliferation by epigenetically silencing P21/PLK3/DDIT3 expression. *Oncotarget* **6**, 33587–33601, https://doi.org/10.18632/oncotarget.5599 (2015).
55. Bao, X. *et al*. Knockdown of long non-coding RNA HOTAIR increases miR-454-3p by targeting Stat3 and Atg12 to inhibit chondrosarcoma growth. *Cell Death Dis* **8**, e2605, https://doi.org/10.1038/cddis.2017.31 (2017).
56. Gupta, R. A. *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076, https://doi.org/10.1038/nature08975 (2010).
57. Kim, K. *et al*. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* **32**, 1616–1625, http://www.nature.com/onc/journal/v32/n13/suppinfo/onc2012193s1.html (2013).
58. Li, X. *et al*. Long non-coding RNA HOTAIR, a driver of malignancy, predicts negative prognosis and exhibits oncogenic activity in oesophageal squamous cell carcinoma. *Br J Cancer* **109**, 2266–2278, https://doi.org/10.1038/bjc.2013.548 (2013).
59. Ozes, A. R. *et al*. NF-kappaB-HOTAIR axis links DNA damage response, chemoresistance and cellular senescence in ovarian cancer. *Oncogene* **35**, 5350–5361, https://doi.org/10.1038/onc.2016.75 (2016).
60. Shi, J. *et al*. Long non-coding RNA in glioma: signaling pathways. *Oncotarget*. https://doi.org/10.18632/oncotarget.15175 (2017).
61. Ding, J. *et al*. Long noncoding RNA HOXA-AS2 represses P21 and KLF2 expression transcription by binding with EZH2, LSD1 in colorectal cancer. *Oncogenesis* **6**, e288, https://doi.org/10.1038/oncsis.2016.84 (2017).
62. Zhao, H., Zhang, X., Frazao, J. B., Condino-Neto, A. & Newburger, P. E. HOX antisense lincRNA HOXA-AS2 is an apoptosis repressor in all trans retinoic acid treated NB4 promyelocytic leukemia cells. *J Cell Biochem* **114**, 2375–2383, https://doi.org/10.1002/jcb.24586 (2013).
63. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641, https://doi.org/10.1016/j.cell.2009.02.006 (2009).
64. Rinn, J. L. *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323, https://doi.org/10.1016/j.cell.2007.05.022 (2007).
65. Khalil, A. M. *et al*. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* **106**, 11667–11672, https://doi.org/10.1073/pnas.0904715106 (2009).
66. Goutagny, S. *et al*. Genomic profiling reveals alternative genetic pathways of meningioma malignant progression dependent on the underlying NF2 status. *Clin Cancer Res* **16**, 4155–4164, https://doi.org/10.1158/1078-0432.CCR-10-0891 (2010).
67. Bostrom, J. *et al*. Alterations of the tumor suppressor genes CDKN2A (p16(INK4a)), p14(ARF), CDKN2B (p15(INK4b)), and CDKN2C (p18(INK4c)) in atypical and anaplastic meningiomas. *Am J Pathol* **159**, 661–669, https://doi.org/10.1016/S0002-9440(10)61737-3 (2001).
68. Kadoch, C. & Crabtree, G. R. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Sci Adv* **1**, e1500447, https://doi.org/10.1126/sciadv.1500447 (2015).
69. Bernhart, S. H. *et al*. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Sci Rep* **6**, 37393, https://doi.org/10.1038/srep37393 (2016).
70. Voigt, P., Tee, W. W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev* **27**, 1318–1338, https://doi.org/10.1101/gad.219626.113 (2013).
71. Bernstein, B. E. *et al*. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326, https://doi.org/10.1016/j.cell.2006.02.041 (2006).
72. Helming, K. C., Wang, X. & Roberts, C. W. Vulnerabilities of mutant SWI/SNF complexes in cancer. *Cancer Cell* **26**, 309–317, https://doi.org/10.1016/j.ccr.2014.07.018 (2014).
73. Kim, K. H. *et al*. SWI/SNF-mutant cancers depend on catalytic and non-catalytic activity of EZH2. *Nat Med* **21**, 1491–1496, https://doi.org/10.1038/nm.3968 (2015).
74. Bitler, B. G. *et al*. Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. *Nat Med* **21**, 231–238, https://doi.org/10.1038/nm.3799 http://www.nature.com/nm/journal/v21/n3/abs/nm.3799.html - supplementary-information (2015).

75. Kim, K. H. & Roberts, C. W. Targeting EZH2 in cancer. *Nat Med* **22**, 128–134, https://doi.org/10.1038/nm.4036 (2016).
76. Ozes, A. R. *et al.* Therapeutic targeting using tumor specific peptides inhibits long non-coding RNA HOTAIR activity in ovarian and breast cancer. *Scientific reports* **7**, 894, https://doi.org/10.1038/s41598-017-00966-3 (2017).
77. Pfister, S. X. & Ashworth, A. Marked for death: targeting epigenetic changes in cancer. *Nature reviews. Drug discovery* **16**, 241–263, https://doi.org/10.1038/nrd.2016.256 (2017).
78. Norden, A. D. *et al.* Phase II trials of erlotinib or gefitinib in patients with recurrent meningioma. *J Neurooncol* **96**, 211–217, https://doi.org/10.1007/s11060-009-9948-7 (2010).
79. Byers, L. A. *et al.* An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical cancer research: an official journal of the American Association for Cancer Research* **19**, 279–290, https://doi.org/10.1158/1078-0432.ccr-12-1558 (2013).
80. Buonato, J. M. & Lazzara, M. J. ERK1/2 blockade prevents epithelial-mesenchymal transition in lung cancer cells and promotes their sensitivity to EGFR inhibition. *Cancer Res* **74**, 309–319, https://doi.org/10.1158/0008-5472.can-12-4721 (2014).
81. Thomson, S., Petti, F., Sujka-Kwok, I., Epstein, D. & Haley, J. D. Kinase switching in mesenchymal-like non-small cell lung cancer lines contributes to EGFR inhibitor resistance through pathway redundancy. *Clinical & experimental metastasis* **25**, 843–854, https://doi.org/10.1007/s10585-008-9200-4 (2008).
82. Chmielecki, J. *et al.* Whole-exome sequencing identifies a recurrent NAB2-STAT6 fusion in solitary fibrous tumors. *Nat Genet* **45**, 131–132, https://doi.org/10.1038/ng.2522 (2013).
83. Gao, F. *et al.* Inversion-mediated gene fusions involving NAB2-STAT6 in an unusual malignant meningioma. *Br J Cancer* **109**, 1051–1055, https://doi.org/10.1038/bjc.2013.395 (2013).
84. Schweizer, L. *et al.* Meningeal hemangiopericytoma and solitary fibrous tumors carry the NAB2-STAT6 fusion and can be diagnosed by nuclear expression of STAT6 protein. *Acta Neuropathol* **125**, 651–658, https://doi.org/10.1007/s00401-013-1117-6 (2013).
85. Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566, https://doi.org/10.1038/nature05945 (2007).
86. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods* **6**, 291–295, https://doi.org/10.1038/nmeth.1311 (2009).
87. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).
88. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–169, https://doi.org/10.1093/bioinformatics/btu638 (2015).
89. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS computational biology* **7**, e1001138, https://doi.org/10.1371/journal.pcbi.1001138 (2011).
90. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* **12**, R72, https://doi.org/10.1186/gb-2011-12-8-r72 (2011).
91. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898, https://doi.org/10.1038/nature08768 (2010).
92. M. Witten, D. Witten, D. M.: *Classification and clustering of sequencing data using a Poisson model. Ann. Appl. Stat. 5*(4), 2493–2518 Vol. 5 (2012).
93. Reeb, P. D., Bramardi, S. J. & Steibel, J. P. Assessing Dissimilarity Measures for Sample-Based Hierarchical Clustering of RNA Sequencing Data Using Plasmode Datasets. *PLoS One* **10**, e0132310, https://doi.org/10.1371/journal.pone.0132310 (2015).
94. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, https://doi.org/10.1186/gb-2010-11-10-r106 (2010).
95. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. 2. *Genome biology* **15**, 550, https://doi.org/10.1186/s13059-014-0550-8 (2014).
96. Peixoto, L. *et al.* How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic acids research* **43**, 7664–7674, https://doi.org/10.1093/nar/gkv736 (2015).
97. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500–507, https://doi.org/10.1038/nprot.2011.457 (2012).
98. Wang, X. *et al.* Analysis of gene expression profiling in meningioma: deregulated signaling pathways associated with meningioma and EGFL6 overexpression in benign meningioma tissue and serum. *PLoS One* **7**, e52707, https://doi.org/10.1371/journal.pone.0052707 (2012).
99. Savvidis, C. & Koutsilieris, M. Circadian rhythm disruption in cancer biology. *Mol Med* **18**, 1249–1260, https://doi.org/10.2119/molmed.2012.00077 (2012).
100. Sharma, S., Ray, S., Moiyadi, A., Sridhar, E. & Srivastava, S. Quantitative proteomic analysis of meningiomas for the identification of surrogate protein markers. *Sci Rep* **4**, 7140, https://doi.org/10.1038/srep07140 (2014).
101. Pfister, C., Tatabiga, M. S. & Roser, F. Selection of suitable reference genes for quantitative real-time polymerase chain reaction in human meningiomas and arachnoidea. *BMC Res Notes* **4**, 275, https://doi.org/10.1186/1756-0500-4-275 (2011).
102. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161, https://doi.org/10.1186/1471-2105-10-161 (2009).
103. Iorio, F. *et al.* Population-level characterization of pathway alterations with SLAPenrich dissects heterogeneity of cancer hallmark acquisition. *bioRxiv*. https://doi.org/10.1101/077701 (2016).
104. Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40**, 499–507, https://doi.org/10.1038/ng.127 (2008).
105. Sarrio, D. *et al.* Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res* **68**, 989–997, https://doi.org/10.1158/0008-5472.can-07-2017 (2008).
106. Wong, D. J. *et al.* Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell stem cell* **2**, 333–344, https://doi.org/10.1016/j.stem.2008.02.009 (2008).
107. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**, D1049–D1056, https://doi.org/10.1093/nar/gku1179 (2015).
108. R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2016).
109. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
110. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)* **25**, 2865–2871, https://doi.org/10.1093/bioinformatics/btp394 (2009).
111. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Current protocols in bioinformatics* **52**, 15.17.11–12, https://doi.org/10.1002/0471250953.bi1507s52 (2015).
112. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54, https://doi.org/10.1038/nature17676 (2016).

113. Raine, K. M. *et al*. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics* **56**, 15.19.11–15.19.17, https://doi.org/10.1002/cpbi.17 (2016).
114. Buels, R. *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology* **17**, 66, https://doi.org/10.1186/s13059-016-0924-1 (2016).
115. Chang, M. T. *et al*. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**, 155–163, https://doi.org/10.1038/nbt.3391 (2016).
116. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550, https://doi.org/10.1073/pnas.0506580102 (2005).
117. Corver, W. E. *et al*. Genome haploidisation with chromosome 7 retention in oncocytic follicular thyroid carcinoma. *PLoS One* **7**, e38287, https://doi.org/10.1371/journal.pone.0038287 (2012).

Acknowledgements

This work was supported by the Wellcome Trust, Cancer Research UK, Meningioma UK and Tadhg and Marie-Louise Flood. U.M. was personally supported by a Cancer Research UK Clinician Scientist Fellowship; G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); F.M. by A.I.L. (Associazione Italiana Contro le Leucemie-Linfomi e Mieloma ONLUS) and by S.I.E.S. (Società Italiana di Ematologia Sperimentale); S.B. was funded by a Wellcome Trust Intermediate Clinical Research Fellowship and a St. Baldrick's Foundation Robert J. Arceci Innovation Award. J.M.C.T. is supported by ERC Starting Grant StG-2016_716290_SCUBA CANCERS and by MINECO Grants RYC 2014 14999 and SAF2015-66368-P. J.B. was funded by the charity Brain Tumour Research. The samples were received from the tissue banks from Cambridge (UK), Dresden (Germany), Liverpool (UK), Plymouth (UK) and Tel Aviv (Israel). The Human Research Tissue Bank is supported by the NIHR Cambridge Biomedical Research Centre. We are grateful to the patients who enabled this study and to the clinical teams coordinating their care.

Author Contributions

G.C. and N.K. performed mRNA expression analysis. G.C. and P.T. analysed whole genome and targeted sequencing data. I.M. performed statistical analyses to detect novel driver mutations. S.M. analysed methylation array data. F.M. generated mutational signatures analysis. J.M.C.T. and M.C. performed retrotransposon analysis. C.O.H. and J.D. performed protein expression analysis. A.B., S.B. and M.Y. contributed to data analysis strategy. A.Y., T.N., G.R.B. and J.T. provided informatic support. T.S., R.W.K., M.K., G.S., D.P., A.D., C.E.M., A.Y., I.N., S.J.P., C.W., Z.R., M.D.J., R.Z. and K. S. provided samples and clinical data. S.B., G.S.V., I.N. and M.W.M. provided conceptual advice. V.P.C. and K.A. carried out central pathology review. U.M. and T.S. devised and supervised the project. G.C. wrote the manuscript with input from U.M., S.B., T.S., P.T., and G.S.V. All authors approved the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-31659-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04650-6 OPEN

Recurrent intragenic rearrangements of *EGFR* and *BRAF* in soft tissue tumors of infants

Jenny Wegert[1], Christian Vokuhl[2], Grace Collord[3,4], Martin Del Castillo Velasco-Herrera[3], Sarah J. Farndon[3,5], Charlotte Guzzo[3], Mette Jorgensen[6], John Anderson[5,6], Olga Slater[6], Catriona Duncan[6], Sabrina Bausenwein[1], Heike Streitenberger[1], Barbara Ziegler[1], Rhoikos Furtwängler[7], Norbert Graf[7], Michael R. Stratton[3], Peter J. Campbell[3], David TW Jones[8,9], Christian Koelsche[10,11,12], Stefan M. Pfister[8,9,13], William Mifsud[6], Neil Sebire[5,6], Monika Sparber-Sauer[14], Ewa Koscielniak[14,15], Andreas Rosenwald[16,17], Manfred Gessler[1,17] & Sam Behjati[3,4]

Soft tissue tumors of infancy encompass an overlapping spectrum of diseases that pose unique diagnostic and clinical challenges. We studied genomes and transcriptomes of cryptogenic congenital mesoblastic nephroma (CMN), and extended our findings to five anatomically or histologically related soft tissue tumors: infantile fibrosarcoma (IFS), nephroblastomatosis, Wilms tumor, malignant rhabdoid tumor, and clear cell sarcoma of the kidney. A key finding is recurrent mutation of *EGFR* in CMN by internal tandem duplication of the kinase domain, thus delineating CMN from other childhood renal tumors. Furthermore, we identify *BRAF* intragenic rearrangements in CMN and IFS. Collectively these findings reveal novel diagnostic markers and therapeutic strategies and highlight a prominent role of isolated intragenic rearrangements as drivers of infant tumors.

[1] Theodor-Boveri-Institute/Biocenter, Developmental Biochemistry, University of Wuerzburg, 97074 Wuerzburg, Germany. [2] Kiel Pediatric Tumor Registry, Section of Pediatric Pathology, Department of Pathology, Christian Albrechts University, 24105 Kiel, Germany. [3] Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK. [4] Department of Paediatrics, University of Cambridge, Cambridge, CB2 0QQ, UK. [5] UCL Great Ormond Street Institute of Child Health, London, WC1N 1EH, UK. [6] Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, UK. [7] Department of Pediatric Oncology and Hematology, Saarland University Hospital, 66421 Homburg, Germany. [8] Hopp Children's Cancer Center at the NCT Heidelberg (KiTZ), 69120 Heidelberg, Germany. [9] Department of Pediatric Neurooncology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), 69120 Heidelberg, Germany. [10] Clinical Cooperation Unit Neuropathology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [11] Department of Neuropathology, Institute of Pathology, Heidelberg University Hospital, 69120 Heidelberg, Germany. [12] Department of General Pathology, Institute of Pathology, Heidelberg University Hospital, 69120 Heidelberg, Germany. [13] Department of Pediatric Hematology and Oncology, Heidelberg University Hospital, 69120 Heidelberg, Germany. [14] Klinikum Stuttgart—Olgahospital, Stuttgart Cancer Center, Zentrum für Kinder-, Jugend- und Frauenmedizin, Pediatrics 5 (Oncology, Hematology, Immunology), 70174 Stuttgart, Germany. [15] Department of Pediatric Hematology and Oncology, Children's Hospital, 72076 Tübingen, Germany. [16] Institute of Pathology, University of Wuerzburg, 97080 Wuerzburg, Germany. [17] Comprehensive Cancer Center Mainfranken, University of Wuerzburg, 97078 Wuerzburg, Germany. These authors contributed equally: Jenny Wegert, Christian Vokuhl, Grace Collord, Martin Del Castillo Velasco-Herrera. These authors jointly supervised this work: Manfred Gessler, Sam Behjati. Correspondence and requests for materials should be addressed to M.G. (email: gessler@biozentrum.uni-wuerzburg.de) or to S.B. (email: sb31@sanger.ac.uk)

Many childhood tumors show a predilection for specific developmental stages. Tumors that predominantly occur in infancy include congenital mesoblastic nephroma (CMN), which accounts for 4% of all childhood renal malignancies and the majority of those diagnosed in children under 6 months of age[1,2]. CMN is classified histologically into classical, cellular, and mixed subtypes based primarily on degree of cellularity and mitotic activity[3]. The cellular variant is characterized by a sarcoma-like diffuse hypercellular morphology, whereas classical CMN is composed of less proliferative spindle cells[3]. Cellular CMN is driven by rearrangements involving the tropomyosin receptor kinase (TRK) gene *NTRK3*, most commonly a t(12;15)(p13;q25) reciprocal translocation with the *ETV6* transcription factor[4,5]. Less frequent somatic aberrations include trisomies of chromosomes 8, 11, 17, and 20[6,7] and rarer TRK fusions, involving *NTRK1*, *NTRK2*, or *NTRK3*[8]. By contrast, the genetic changes underpinning the classical variant, accounting for >30% of cases, are unknown[9]. Cellular CMN shares its genetic and morphological hallmarks with infantile fibrosarcoma (IFS), a spindle cell tumor typically arising in the soft tissues of the extremities or abdomen[5,9,10].

Standard treatment for CMN and IFS is complete surgical resection[9–11]. In the case of IFS, local control frequently requires cytotoxic chemotherapy[10,11]. The role for up-front chemotherapy in CMN is less clear[9]. Recently, a phase I/II clinical trial of a selective TRK inhibitor, larotrectinib, reported high response rates in diverse tumor types harboring TRK gene fusions, including IFS and other soft tissue tumors of infancy[12]. Morbidity and infrequent death result from tumor recurrence or from treatment-related complications[9–11].

Here, we investigated the genetic basis of CMN and IFS lacking the canonical *NTRK3-ETV6* fusion gene. We identify oncogenic rearrangements in MAPK signaling genes across all cases interrogated by unbiased sequencing, notably therapeutically tractable intragenic rearrangements in *EGFR* and *BRAF*.

Results

Overview of the genomic landscape of CMN. To identify the genetic basis of cryptogenic CMN, we first applied whole genome and transcriptome sequencing to a discovery cohort of ten classical CMN lacking an *NTRK3* fusion (Supplementary Data 1). Somatic variants were identified by comparing tumor and matched peripheral blood sequences (see Methods). The genomic landscape was universally quiet, with a low burden of point mutations (median of 45 substitutions and 9 insertions or deletions per genome; Supplementary Data 2). The predominant mutational signatures, as defined by the trinucleotide context of substitutions, were the ubiquitous signatures 1 and 5[13]

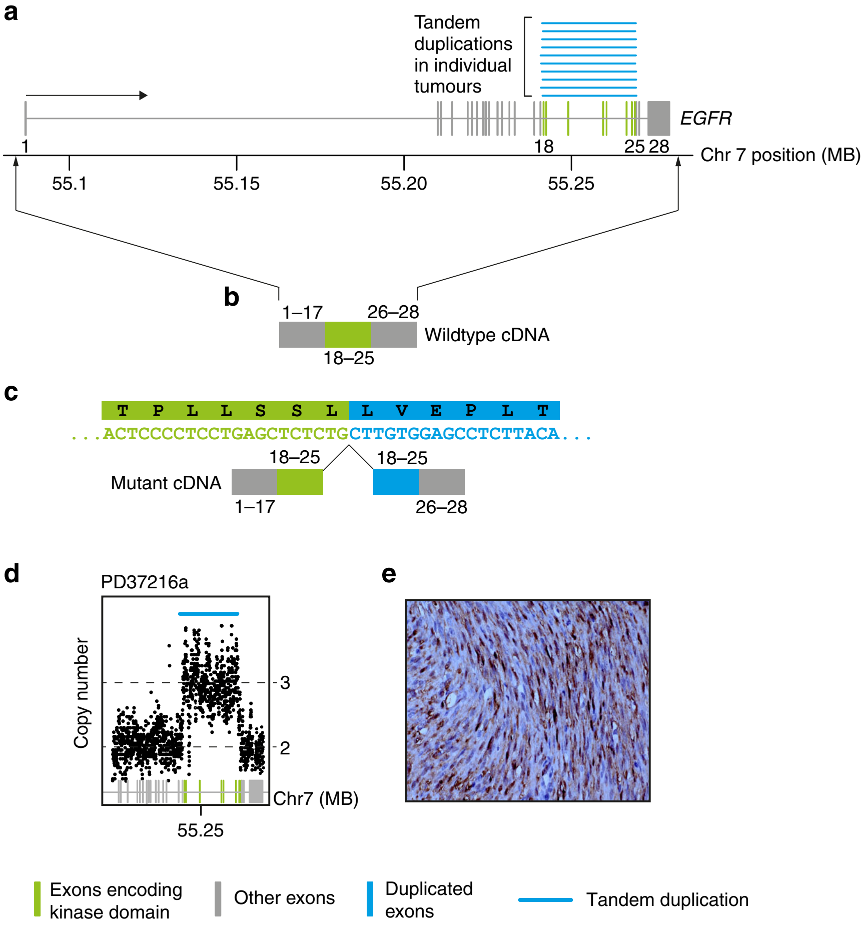


Fig. 1 *EGFR* internal tandem duplication. **a** The genomic footprint of *EGFR* is depicted with exons represented by gray and green vertical lines. Green exons encode the kinase domain. Blue lines superiorly show the tandem duplications found in the discovery cohort of ten congenital mesoblastic nephroma of classical histology. **b** Schematic of the wild-type transcript. **c** Schematic of the fusion transcript annotated with cDNA sequence of rearrangements (sense orientation) and protein translation. **d** Intragenic copy number of *EGFR* showing focal amplification over the kinase domain (*x*-axis: genomic coordinate; *y*-axis: copy number derived from coverage). **e** Representative phospo-ERK immunohistochemistry

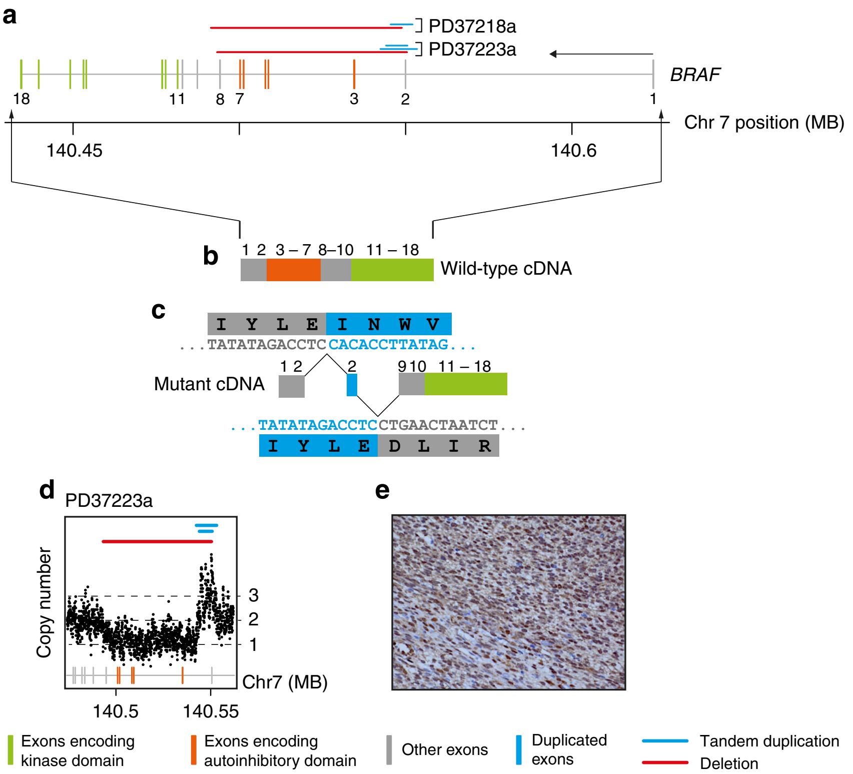


Fig. 2 Internal *BRAF* deletion. **a** The genomic footprint of *BRAF* is depicted with exons represented by gray, green, and orange vertical lines. Green and orange exons encode the kinase domain and conserved region 1, respectively. Horizontal lines above exons demarcate rearrangements (blue: tandem duplication; red: deletion). **b** Outline of wild-type transcript. **c** Outline of fusion transcript with cDNA sequence of rearrangements (sense orientation) with translation. **d** Intragenic copy number of *BRAF* (*x*-axis: genomic coordinate; *y*-axis: copy number derived from coverage). **e** Representative phospho-ERK immunohistochemistry

(Supplementary Fig. 1). Copy number changes and structural rearrangements were likewise scarce (Supplementary Fig. 2).

Internal tandem duplication of the *EGFR* kinase domain in CMN. Annotating all cases for potential oncogenic variants revealed a single intragenic, in-frame internal tandem duplication (ITD) of the *EGFR* kinase domain in all ten tumors (Table 1; Fig. 1; Supplementary Data 3). The breakpoints clustered in a narrow genomic window around the kinase domain of *EGFR* encoded in exons 18—25 (Fig. 1a). This rearrangement is rarely observed in several other tumor types including in glioma and in lung adenocarcinoma, and confers sensitivity to a targeted EGFR inhibitor, afatinib[14]. We validated all rearrangements by genomic copy number analysis and reconstruction of cDNA reads spanning the breakpoint junction (Fig. 1; see Methods). Of note, the same mutant cDNA junction sequence was found in every case, irrespective of the genomic location of breakpoints. A search for additional known or novel driver variants revealed no further plausible candidates in any of the *EGFR*-mutant tumors. We next extended this investigation to seven non-classical CMN lacking an *NTRK3* fusion, including four mixed cellularity cases and three cellular tumors (Table 1; Supplementary Data 1). Two of the four mixed cellularity tumors surveyed also harbored an *EGFR*-ITD. Of note, for one child with *EGFR*-ITD-positive mixed cellularity CMN (PD37214), both primary tumor and recurrence were studied, with no additional driver events apparent at relapse.

***BRAF* rearrangements in CMN and IFS**. A further striking finding was the discovery of mutations in the *BRAF* oncogene in 2/3 cellular histology CMNs. *BRAF* fusions have been implicated in a minority of IFS but not in CMN[15]. In both cases the *BRAF* rearrangement involved a compound deletion of conserved region 1 (CR1) and tandem duplication of exon 2 (Fig. 2; Table 1; Supplementary Data 3). CR1 encompasses the negative regulatory Ras-binding domain (RBD), loss of which is predicted to generate a constitutively active form of BRAF[16,17]. Mutated tumors displayed intense staining of phosphorylated ERK by immunohistochemistry, consistent with activated signaling downstream of BRAF (Figs. 1e and 2e). A further tumor harbored the *KIAA1549-BRAF* fusion, a molecular hallmark of a childhood brain tumor, pilocytic astrocytoma[18,19]. This fusion likewise results in loss of the N-terminal portion of the BRAF protein containing the RBD[17,18].

Other TRK fusions in CMN. The remaining two cases of CMN interrogated by whole genome and transcriptome sequencing were accounted for by gene fusions involving *NTRK1*, an alternate kinase of the TRK family of protein kinases: *TPR-NTRK1* and *LMNA-NTRK1*. Both of these fusions have been observed in IFS and rarely in adult cancers, but not, to our knowledge, in CMN[20–23] (Table 1). Hence, every cryptogenic CMN interrogated by whole-genome sequencing contained an oncogenic rearrangement in *BRAF*, *EGFR*, or *NTRK1*, all of which encode kinases involved in MAPK signaling and are amenable to inhibition with existing drugs[9,12,14,17,24].

***EGFR-ITD* distinguishes CMN from other childhood renal tumors**. To validate and extend our findings, we screened IFS and a range of childhood renal tumors for *EGFR*-ITD, *BRAF*-ID, and *ETV6-NTRK3* using PCR. Tumor types included additional cases of CMN ($n = 63$), IFS ($n = 26$), Wilms tumor ($n = 208$), clear cell sarcoma of the kidney without *BCOR* rearrangements ($n = 20$), malignant rhabdoid tumor ($n = 3$), and nephroblastomatosis

Table 1 Rearrangements in infant soft tissue tumors

Assay	Tumor type	Subtype	Total	*EGFR*-ITD	*BRAF*-ID	*BRAF*-ID + *ETV6-NTRK3*	*ETV6-NTRK3*	*KIAA1549-BRAF*	*LMNA-NTRK1*	*EML4-NTRK3*	*TPR-NTRK1*
WGS + mRNA sequencing	CMN	Cellular	3	0	2	0	0	0	1	0	0
		Classical	10	10	0	0	0	0	0	0	0
		Mixed	4	2	0	0	0	1	0	0	1
	IFS	–	1	0	0	0	0	0	0	1	0
PCR for *EGFR*-ITD, *BRAF*-ID and *ETV6-NTRK3*	CMN	Cellular	17	2	0	0	13	-	-	-	-
		Classical	35	20	0	0	0	-	-	-	-
		Mixed	11	9	0	0	0	-	-	-	-
	IFS	-	26	0	1	2	16	-	-	-	-
	WT	-	208	0	0	0	0	-	-	-	-
	CCSK[a]	-	20	0	0	0	0	-	-	-	-
	MRT	-	3	0	0	0	0	-	-	-	-
	NB	-	12	0	0	0	0	-	-	-	-

CMN congenital mesoblastic nephroma, *IFS* infantile fibrosarcoma, *WT* Wilms tumor, *CCSK* clear cell sarcoma of the kidney, *MRT* malignant rhabdoid tumor, *NB* nephroblastomatosis, *WGS* whole genome sequencing, *mRNA* messenger RNA, *PCR* polymerase chain reaction

[a] Negative for BCOR rearrangement

($n = 12$; Table 1; Supplementary Data 1). *EGFR*-ITD was most prevalent in classical and mixed cellularity CMN, though was also found in cellular CMN (2/17 cases). The frequency of *EGFR* rearrangement in classical tumors was lower in the validation cohort (20/35 cases) than in the initial discovery cohort (10/10 cases). None of the IFS cases, nor other childhood kidney tumors, harbored *EGFR*-ITD. However, we encountered three cases of IFS with intragenic *BRAF* deletions. Remarkably, in two cases *BRAF*-ID co-occurred with *NTRK3* fusions, the disease-defining mutation of IFS. We were unable to accurately estimate relative allele frequencies by nested PCR (see Methods). Hence, it is possible that both fusions co-exist within the same clone or represent independent clones that evolved in parallel within the same tumor.

Discussion

In this exploration of infant tumors we identify ITD of the *EGFR* kinase domain that delineates a genetic subgroup of CMN transcending histological subtypes. Additionally, we report a novel rearrangement of *BRAF* present in both cellular CMN and IFS. These mutations represent diagnostic markers that can be readily integrated into routine clinical practice. Furthermore, EGFR and BRAF emerge as therapeutic targets, which may be exploited in certain clinical situations, e.g., large surgically intractable tumors, disease recurrence or metastases.

It is noteworthy that an oncogenic mutation was identified in every tumor that we studied by whole-genome sequencing. Of these, 78% harbored either *EGFR*-ITD or *BRAF*-ID, while the remaining 22% presented with non-canonical mutations involving *BRAF*, *NTRK1*, or *NTRK3*. This suggests that less recurrent rearrangement variants, albeit implicated in the same signaling circuitry, may elude detection by targeted diagnostic assays. Moreover, our results indicate that a subset of tumors harbor multiple drivers with important implications for targeted therapy efforts. The finding of co-mutation of *NTRK3* and *BRAF* in IFS raises the possibility of intrinsic resistance of some tumors to TRK inhibition, regardless of whether these mutations occur in the same clone or in independent competing clones. This finding is pertinent to clinical trials of TRK inhibitors in CMN and IFS[12]. In this vein a structurally similar *BRAF* fusion transcript, albeit without duplication of exon 2, has recently been implicated as a mechanism of resistance to certain BRAF/MEK inhibitors[16,17]. These considerations underscore the need for adequate genomic profiling in order to match patients to the most appropriate basket studies and to enable meaningful interpretation of treatment responses. Therefore, we would advocate extending the diagnostic work-up of refractory or relapsed CMN and IFS to whole genome sequencing, particularly in the context of clinical trials.

Biologically our findings draw further parallels between CMN and IFS. We identify *BRAF* and *NTRK1* as additional cancer genes operative in both malignancies, substantiating the view that these diagnoses represent variants on the same disease spectrum converging on aberrant RAS-RAF-MEK-ERK signaling[5,8,9]. Furthermore, in the wider context of the childhood cancer genome, our findings add to the growing body of studies that identify short distance intragenic rearrangements as a dominant source of oncogenic mutations in otherwise quiet genomes. We note the parallel between CMN, clear cell sarcoma of the kidney and low-grade glioma that are in large part driven by ITDs often involving kinase domains, mostly as isolated driver events[18,25–29]. Furthermore, even in acute myeloid leukemia, where *FLT3*-ITD is a recurrent driver event in adult disease, childhood AML demonstrates a distinct structural variant profile enriched for focal chromosomal gains and losses[30]. We can only speculate on the biological significance of this parallel which may allude to specific mutational mechanisms operative during discrete stages of human development.

Methods

Patient samples. All tissue samples were obtained after gaining written informed consent for tumor banking and future research from the patient (or their guardian) in accordance with the Declaration of Helsinki and appropriate national and local ethical review processes. German tissue samples were obtained from the following studies: SIOP93-01/GPOH and SIOP2001/GPOH (Ethikkommission der Ärztekammer des Saarlandes reference numbers 23.4.93/Ls and 136/01), the PTT2.0 study (Medical Faculty Heidelberg ethics reference number S-546/2016), the CWS trials CWS-96 and CWS-2002P (Universitätsklinikum Tübingen Medizinische Fakultät ethics approval, reference numbers 105/95 and 51/2003) and the SoTiSaR registry (ethics approval reference 158/2009B02). UK patients were enrolled under ethics approval from National Research Ethics Service Committee East of England, Cambridge Central (reference 16/EE/0394). Use of UK archival material was approved by the National Research Ethics Service Committee London Brent (reference 16/LO/0960). Additional tissue was obtained from the UK Children's Cancer and Leukaemia Group tissue bank.

Sequencing. Tumor DNA and RNA were extracted from fresh frozen tissue that had been reviewed by reference pathologists. Normal tissue DNA was derived from blood samples. Whole genome sequencing was performed by 150-bp paired-end sequencing on the Illumina HiSeq X platform. We followed the Illumina no-PCR library protocol to construct short insert libraries, prepare flowcells, and generate clusters. Coverage was at least 30×. Messenger RNA was enriched by polyAselection and sequenced on an Illumina HiSeq 2000 (paired end, 75-bp read length). DNA and RNA sequencing reads were aligned to the GRCh 37d5 reference genome using the Burrows−Wheeler transform (BWA-MEM)[31] and STAR (2.0.42)[32], respectively.

Variant detection. The Cancer Genome Project (Wellcome Trust Sanger Institute) variant calling pipeline was used to call somatic mutation and includes the following algorithms: CaVEMan (1.11.0)[33] for substitutions, an in-house version of Pindel (2.2.2; github.com/cancerit/cgpPindel)[34] for indels, BRASS (5.3.3; github.com/cancerit/BRASS) for rearrangements, and ASCAT NGS (4.0.0) for copy number aberrations[35]. RNA sequences were analyzed with an in-house pipeline (github.com/cancerit/cgpRna/wiki) which uses HTSeq[36] for gene feature counts, and a combination of TopHat-Fusion (v2.1.0)[37], STAR-fusion (v0.1.1)[32] and DeFuse (v0.7.0)[38] to detect expressed gene fusions. In addition to filters inherent to the CaVEMan algorithm, we used the following post-processing filtering criteria for substitutions: a minimum of two reads in each direction reporting the mutant allele, at least tenfold coverage at the mutant allele locus, minimum variant allele fraction 5%; no insertion or deletion called within a read length (150 bp) of the putative substitution, no soft-clipped reads reporting the mutant allele, and a median BWA alignment score of the reads reporting the mutant allele ≥140. The following variants were flagged for additional inspection for potential artifacts, germline contamination or index-jumping event: any mutant allele reported within 150 bp of another variant, any mutant allele with a population allele frequency >1 in 1000 according to any of five large polymorphism databases (ExAC, 1000 Genomes Project, ESP6500, CG46, Kaviar), variant reported in more than 10% of the tumor samples and mutant allele reported in >1% of the matched normal reads. For indels, the inbuilt filters of the Pindel algorithm, as implemented in our pipeline, were used. In addition, recurrent indels occurring in >2 samples were flagged for additional inspection.

Mutational signatures were derived using principal component analysis and non-negative matrix factorization as implemented in the SomaticSignatures R package[39].

Variant validation. The Cancer Genome Project (Wellcome Trust Sanger Institute) variant calling pipeline has been continually validated and bench-marked[40,41]. We confirmed variant calling quality through manual visual inspection of raw sequencing read for 8% of all variants called. All rearrangements reported were validated by reconstruction at base pair resolution and by cDNA reads spanning the breakpoint junction.

Analysis of mutations in cancer genes. We considered variants as potential drivers if they presented in established cancer genes[42]. Tumor suppressor coding variants were considered if they were annotated as functionally deleterious by an in-house version of VAGrENT (http://cancerit.github.io/VAGrENT/)[43] or were disruptive rearrangement breakpoints or focal (<1 Mb) homozygous deletions. Mutations in oncogenes were considered driver events if they were located at previously reported canonical hot spots (point mutations) or amplified the intact gene. Amplifications also had to be focal (<1 Mb) and increase the copy number of oncogenes to a minimum of five copies for a diploid genome. To search for driver variants in novel cancer genes or in non-coding regions, we employed previously developed statistical methods that identify significant enrichment of mutations, taking into account various confounders such as overall mutation burden and local variation in the mutability of the genomic region[44].

Targeted mutation screening. RNA from frozen tumors (1 µg) or corresponding to approximately 5 cm^2 of 10 µm FFPE sections was reverse transcribed using oligo-dT or random hexamer primers (RevertAid first strand cDNA synthesis kit, ThermoFisher). PCR screening was performed using primer combinations that allow amplification of candidate alterations as well as additional control fragments from the unaffected allele to assess cDNA quality. Amplified fragments were sequenced by Sanger sequencing (GATC, Konstanz, Germany) using primers detailed in Supplementary Table 1.

Immunohistochemistry. Immunohistochemical staining for phospho-ERK1/2 (Cell Signaling Technology, clone D13.14.4E) was performed according to standard protocol (dilution 1:800, pre-treatment with target retrieval TR6.1, Dako). Results were scored in a semi-quantitative fashion (negative, weak, moderate, strong).

Code availability. The algorithms used to analyze sequencing data are available at http://cancerit.github.io/.

Data availability. All data supporting the findings of this study are available within the article and its supplementary files or from the corresponding author on reasonable request. Sequencing data have been deposited at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) that is hosted by the European Bioinformatics Institute (accession numbers EGAS00001002534 and EGAS00001002171).

Received: 15 December 2017 Accepted: 14 May 2018
Published online: 18 June 2018

References

1. Marsden, H. B. & Lawler, W. Primary renal tumours in the first year of life. A population based review. *Virchows Arch. A. Pathol. Anat. Histopathol.* **399**, 1–9 (1983).
2. Glick, R. D. et al. Renal tumors in infants less than 6 months of age. *J. Pediatr. Surg.* **39**, 522–525 (2004).
3. Charles, A. K., Vujanic, G. M. & Berry, P. J. Renal tumours of childhood. *Histopathology* **32**, 293–309 (1998).
4. Rubin, B. P. et al. Congenital mesoblastic nephroma t(12;15) is associated with ETV6-NTRK3 gene fusion: cytogenetic and molecular relationship to congenital (infantile) fibrosarcoma. *Am. J. Pathol.* **153**, 1451–1458 (1998).
5. Knezevich, S. R. et al. ETV6-NTRK3 gene fusions and trisomy 11 establish a histogenetic link between mesoblastic nephroma and congenital fibrosarcoma. *Cancer Res.* **58**, 5046–5048 (1998).
6. Adam, L. R., Davison, E. V., Malcolm, A. J., Pearson, A. D. & Craft, A. W. Cytogenetic analysis of a congenital fibrosarcoma. *Cancer Genet. Cytogenet.* **52**, 37–41 (1991).
7. Schofield, D. E., Yunis, E. J. & Fletcher, J. A. Chromosome aberrations in mesoblastic nephroma. *Am. J. Pathol.* **143**, 714–724 (1993).
8. Church, A. J. et al. Recurrent EML4-NTRK3 fusions in infantile fibrosarcoma and congenital mesoblastic nephroma suggest a revised testing strategy. *Mod. Pathol.* **31**, 463–473 (2018).
9. Gooskens, S. L. et al. Congenital mesoblastic nephroma 50 years after its recognition: a narrative review. *Pediatr. Blood Cancer* **64**, e26437 (2017).
10. Orbach, D. et al. Infantile fibrosarcoma: management based on the European experience. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **28**, 318–323 (2010).
11. Soule, E. H. & Pritchard, D. J. Fibrosarcoma in infants and children: a review of 110 cases. *Cancer* **40**, 1711–1721 (1977).
12. Drilon, A. et al. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. *N. Engl. J. Med.* **378**, 731–739 (2018).
13. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
14. Gallant, J. N. et al. EGFR kinase domain duplication (EGFR-KDD) is a novel oncogenic driver in lung cancer that is clinically responsive to afatinib. *Cancer Discov.* **5**, 1155–1163 (2015).
15. Kao, Y. C. et al. Recurrent BRAF gene fusions in a subset of pediatric spindle cell sarcomas: expanding the genetic spectrum of tumors with overlapping features with infantile fibrosarcoma. *Am. J. Surg. Pathol.* **42**, 28-38 (2018).
16. Johnson, D. B. et al. BRAF internal deletions and resistance to BRAF/MEK inhibitor therapy. *Pigment Cell Melanoma Res.* **31**, 432-436 (2018).
17. Karoulia, Z., Gavathiotis, E. & Poulikakos, P. I. New perspectives for targeting RAF kinase in human cancer. *Nat. Rev. Cancer* **17**, 676–691 (2017).
18. Jones, D. T. et al. Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res.* **68**, 8673–8677 (2008).
19. Ross, J. S. et al. The distribution of BRAF gene fusions in solid tumors and response to targeted therapy. *Int. J. Cancer* **138**, 881–890 (2016).
20. Wong V. et al. Evaluation of a congenital infantile fibrosarcoma by comprehensive genomic profiling reveals an LMNA-NTRK1 gene fusion responsive to crizotinib. *J. Natl Cancer Inst.* **108**, djv307 (2016).
21. Davis, J. L. et al. Infantile NTRK-associated Mesenchymal Tumors. *Pediatr. Dev. Pathol.* **21**, 68–78 (2018).
22. Sartore-Bianchi, A. et al. Sensitivity to entrectinib associated with a novel LMNA-NTRK1 gene fusion in metastatic colorectal cancer. *J. Natl Cancer Inst.* **108**, djv306 (2016).
23. Doebele, R. C. et al. An oncogenic NTRK fusion in a patient with soft-tissue sarcoma with response to the tropomyosin-related kinase inhibitor LOXO-101. *Cancer Discov.* **5**, 1049–1057 (2015).
24. Cook, P. J. et al. Somatic chromosomal engineering identifies BCAN-NTRK1 as a potent glioma driver and therapeutic target. *Nat. Commun.* **8**, 15987 (2017).
25. Roy, A. et al. Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. *Nat. Commun.* **6**, 8891 (2015).
26. Zhang, J. et al. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nat. Genet.* **45**, 602–612 (2013).
27. Jones, D. T. et al. Oncogenic RAF1 rearrangement and a novel BRAF mutation as alternatives to KIAA1549:BRAF fusion in activating the MAPK pathway in pilocytic astrocytoma. *Oncogene* **28**, 2119–2123 (2009).
28. Jones, D. T. et al. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.* **45**, 927–932 (2013).
29. Paugh, B. S. et al. Genome-wide analyses identify recurrent amplifications of receptor tyrosine kinases and cell-cycle regulatory genes in diffuse intrinsic

pontine glioma. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **29**, 3999–4006 (2011).

30. Bolouri, H. et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24**, 103–112 (2018).
31. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows—Wheeler transform. *Bioinformatics (Oxford, England)* **26**, 589–595 (2010).
32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
33. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic sngle nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.11–15.10.18 (2016).
34. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and dletion events from paired end sequencing. *Curr. Protoc. Bioinform.* **52**, 15.17.11–15.17.12 (2015).
35. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinforma.* **56**, 15.19.11–15.19.17 (2016).
36. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–169 (2015).
37. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
38. McPherson, A. et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7**, e1001138 (2011).
39. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics (Oxford, England)* **31**, 3673–3675 (2015).
40. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
41. Behjati, S. et al. Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. *Nat. Commun.* **8**, 15936 (2017).
42. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–d783 (2017).
43. Menzies, A. et al. VAGrENT: Variation Annotation Generator. *Curr. Protoc. Bioinform.* **52**, 15.8.1–15.8.11 (2015).
44. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e1021 (2017).

Acknowledgements

This work was supported by funding by the Wellcome Trust, St. Baldrick's Foundation, the Deutsche Forschungsgemeinschaft (GE 539/13-1), the Deutsche Krebshilfe (50-2709-Gr2, T9/96/TrI, 50-2721-Tr2) and NIHR GOSH BRC. G.C., S.B., C.G., P.J.C., and M.R.S. received personal fellowships from the Wellcome Trust. The Cooperative Weichteilsarkom Studiengruppe (CWS) was additionally supported by the Deutsche Kinderkrebsstiftung (SoTiSaR, A2007/13DKS2009.08) and by the Förderkreis Krebskranke Kinder e.V. Stuttgart, Germany. The SIOP-RTSG/GPOH-nephroblastoma study group is supported by the charity "Elterninitiative krebskranker Kinder im Saarland e.V.". We thank children and their families for participating in our research and the clinical teams involved in their care. We thank Sabine Roth and Sharna Lunn for expert technical assistance.

Author contributions

J.W., G.C., M.D.C.V.H., and C.G. analyzed sequencing data. C.V. performed histological analyses. S.Ba., H.S., and B.Z. provided technical assistance. S.J.F., M.J., J.A., O.S., C.D., R.F., N.G., D.T.W.J., C.K., S.M.P., W.M., E.K., N.S., A.R. and M.S.-S. curated and reviewed the samples, clinical data, and/or provided clinical expertise. M.R.S. and P.J.C. contributed to discussions. M.G. and S.B. directed this research and wrote the manuscript, with contributions from G.C., J.W., and M.D.C.V.H.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04650-6.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Recurrent histone mutations in T-cell acute lymphoblastic leukaemia

Mutations affecting key modifiable histone type 3 (H3; Table SI) residues are frequent oncogenic events in certain solid tumours (Feinberg *et al*, 2016), and have also recently been implicated in a subset of acute myeloid leukaemia (AML) (Lehnertz *et al*, 2017). Here, we systematically reviewed the somatic mutations in >20 000 cancer specimens to identify tumours harbouring H3 mutations. In a subset of T-cell acute lymphoblastic leukaemia (T-ALL) we identified non-methionine mutations of the key modifiable H3 residues, lysine (K) 27 and 36.

The starting point of our investigation was a search for H3 hotspot mutations in 1020 human cancer cell lines (Table SII). In two cell lines, both derived from T-ALL, we found lysine-to-arginine mutations at H3K27 and H3K36 (Table I). One of the cell lines, LOUCY, is derived from a *NOTCH1* wild-type adult T-ALL (Ben-Bassat *et al*, 1990). The second, CML-T1, was derived from the T-lymphoblastic blast crisis of chronic myeloid leukaemia (Kuriyama *et al*, 1989). Ten further T-ALL cell lines lacked coding H3 mutations (Table SIII). In solid tumours, H3K27 and H3K36 are typically mutated to methionine (Fig 1) (Feinberg *et al*, 2016). However, recent functional studies of H3 lysine-to-isoleucine mutations in AML demonstrate that the latter also dramatically alter global H3 methylation and acetylation patterns (Lehnertz *et al*, 2017). Therefore, we speculated that lysine-to-non-methionine mutations may also be drivers of a subset of T-ALL.

We next searched for canonical H3 mutations in a published targeted sequencing study of 633 epigenetic regulator genes in >1000 childhood tumours encompassing 21 cancer subtypes (Huether *et al*, 2014). Amongst 91 T-ALL specimens, there were two cases with canonical H3 mutations: *H3F3A* p.K27R and *H3F3A* p.K36R (Table I). Both mutations were clonal, with a variant allele fraction (VAF) of 38% and 55%, respectively. Among the 37 tumours with H3K mutations, lysine-to-arginine mutations were restricted to T-ALL (P = 0·001502; Fisher's exact test).

We then extended our screen for H3 mutations to 18 704 tumours, encompassing >60 cancer types other than T-ALL (Tables SIV and SV). This dataset comprised 8764 internally sequenced specimens and 9940 TCGA samples re-analysed using an in-house variant calling pipeline as previously described (Martincorena *et al*, 2017). We identified only one neomorphic H3 mutation in an acute leukaemia specimen: a previously reported *HIST1H3D* p.K27M mutation in an adult AML case (TCGA-AB2927-03) (Lehnertz *et al*, 2017).

Finally, we examined an additional T-ALL cohort by capillary sequencing of recurrently mutated modifiable residues K27, G34, and K36 across four frequently mutated H3 genes (Tables SVI and SVII). The cohort comprised 38 T-ALL cases described in detail previously (Maser *et al*, 2007). One specimen from a 30-year-old patient harboured a *H3F3A* p.K27N mutation (Figure S1). Interestingly, a *H3F3A* p.K27N mutation and a *H3F3A* p.K27T variant were previously identified in a T-ALL RNA sequencing study (n = 31) (Atak *et al*, 2013). Collectively, our findings indicate that H3K27 and H3K36 mutations are recurrent in T-ALL, a result we were able to reproduce across multiple different cohorts encompassing adult and paediatric cases.

This finding is congruent with the fact that mutations in *SETD2* and *EZH2*, methyltransferases that catalyse trimethylation (me3) of H3K36 and H3K27, respectively, are frequent T-ALL drivers (Belver & Ferrando, 2016). Disruptive *SETD2* alterations occur in 7·8% of early T cell precursor acute lymphoblastic leukaemia (ETP-ALL), an aggressive subtype with stem cell-like features (Belver & Ferrando, 2016). Interestingly, both T-ALL specimens with H3K36R mutations originated from ETP-ALL (Table I). Notably, mutually exclusive *SETD2* and H3K36/H3K34 mutations are reported in paediatric high grade glioma, where both result in reduced H3K36me3 mediated by *SETD2* (Feinberg *et al*, 2016). It is unclear whether a similar co-mutation pattern exists in T-ALL, as H3 genes have not been included in targeted sequencing panels used by the largest T-ALL genomic studies (Belver & Ferrando, 2016).

The role of H3K27 modifications in T-ALL pathogenesis is complex (Belver & Ferrando, 2016). It is plausible that mutations affecting this residue could impact the activity of several histone modifiers with established roles in T-ALL pathogenesis. Loss-of-function mutations in *EZH2* or other core components of Polycomb repressive complex 2 (PRC2) are found in 42% of ETP-ALL and 25% of T-ALL overall (Belver & Ferrando, 2016). Impaired PRC2 catalytic activity in T-ALL is associated with reduced H3K27me3, stemness and poor prognosis (Belver & Ferrando, 2016). *H3F3A* p.K27M mutations appear to act predominantly by blocking H3K27 di- and trimethylation and increasing H3K27 acetylation (Feinberg *et al*, 2016). Recent work demonstrates that H3K27I mutations in AML are associated with similar changes in H3 modification patterns (Lehnertz *et al*, 2017), suggesting that other non-methionine mutations at modifiable H3 residues may influence the activity of PRC2 and

doi: 10.1111/bjh.15155

Table I. Type 3 histone mutations in T cell leukaemia.

Sample name	Sample type	Donor age (years)	Donor sex	H3 mutation
LOUCY	Cell line derived from ETP-ALL	38	Female	*HIST1H3G* p.K36R
CML-T1	Cell line derived from the acute T-lympoblastic blast crisis of CML	36	Female	*H3F3A* p.K27R
SJTALL174	Primary ETP-ALL specimen	Unknown (paediatric)	Unknown	*H3F3A* p.K36R
SJTALL080	Primary T-ALL specimen	Unknown (paediatric)	Unknown	*H3F3A* p.K27R
PD2752a	Primary T-ALL specimen	30	Male	*H3F3A* p.K27N

Out of 141 T cell leukaemia specimens screened (12 cell lines and 129 primary samples), 5 (3·5%) harboured a missense mutation at a modifiable lysine residues K27 or K36. CML, chronic myeloid leukaemia; ETP-ALL, early T cell precursor acute lymphoblastic leukaemia; T-ALL, T cell acute lymphoblastic leukaemia.

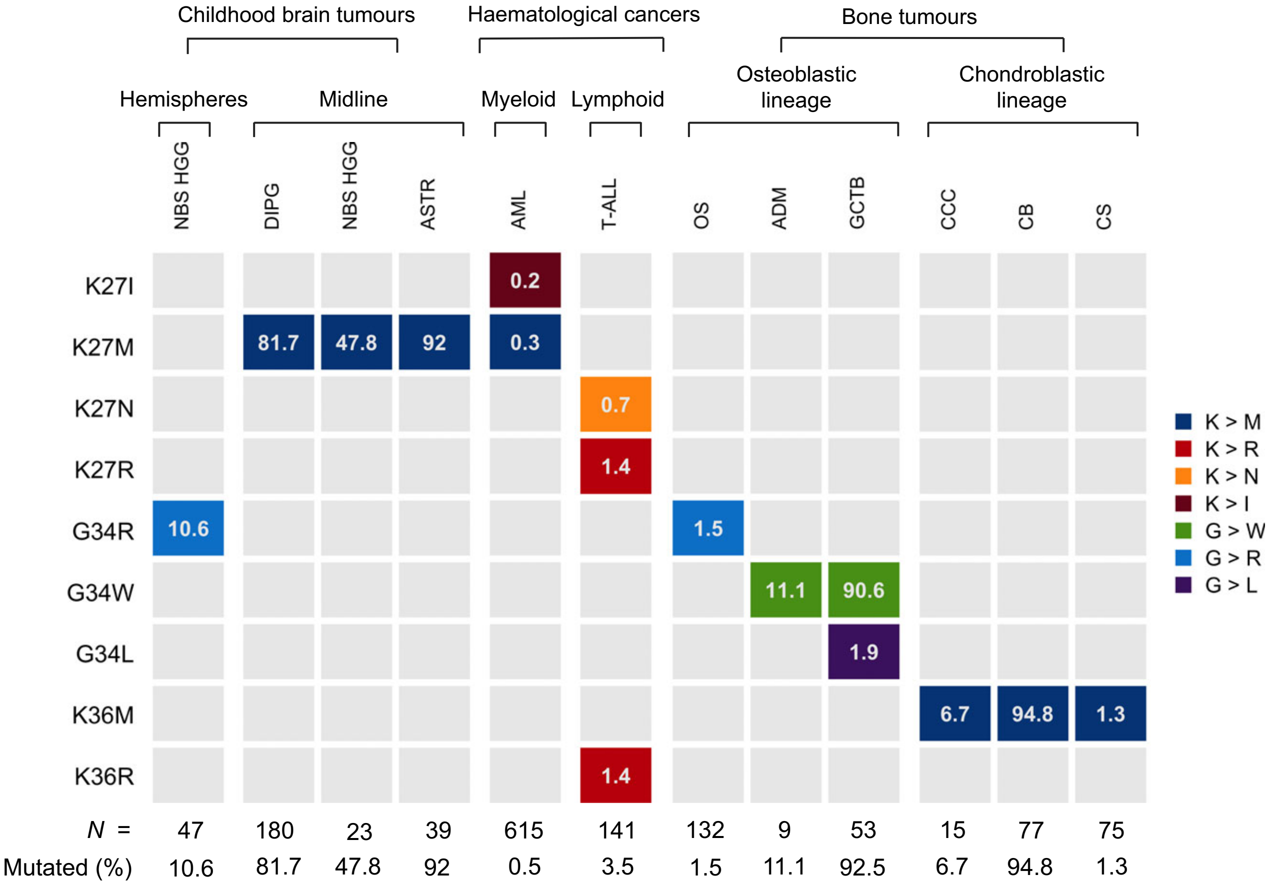


Fig 1. Prevalence and amino acid specificity of type 3 histone mutations in different cancer types. Columns indicate cancer types and rows show key histone type 3 regulatory residues. Tiles are coloured according to amino acid substitution. The percentage of each tumour type affected by the given class of histone mutation is indicated within the tiles and the overall prevalence of histone mutations is summarised at the bottom of each column. NBS HGG, non-brain stem high grade glioma; DIPG, diffuse intrinsic pontine glioma; ASTR, astrocytoma; AML, acute myeloid leukaemia; T-ALL, T cell acute lymphoblastic leukaemia; OS, osteosarcoma; ADM, adamantinoma; GCTB, giant cell tumour of bone; CCC, clear cell chondrosarcoma; CB, chondroblastoma; CS, chondrosarcoma.

other histone modifying enzymes. The lysine-specific demethylases *JMJD3* and *UTX* are further important regulators of H3K27me3 distribution in T-ALL (Belver & Ferrando, 2016), and it is conceivable that these enzymes may also be affected by H3K27 or H3K36 mutations.

A feature of H3 mutations in solid cancers is their exquisite tumour type specificity (Fig 1) (Feinberg *et al*, 2016). In this context, it is notable that 5/5 H3 mutations in T-ALL identified by this survey are lysine-to-non-methionine mutations, and 4/5 are lysine-to-arginine mutations. Out of the >20 000 tumour specimens screened for H3 variants, only two other samples harboured H3 lysine-to-arginine mutations, both at low VAF and in tumours with relatively high coding mutation burdens (TCGA-BT-A20Q-01 and TCGA-

AN-A0FW-01). Hence, it is possible that lysine-to-arginine mutations confer particular selective advantage in the context of T cell leukaemogenesis.

In summary, ~3% of T-ALL harbour non-methionine variants in H3 genes at key modifiable lysine residues. Given the role of dysregulated H3K27/H3K36 modification in T-ALL pathogenesis and the established prognostic significance of mutations in lysine-specific histone modifiers (Belver & Ferrando, 2016), this finding warrants further investigation of the prevalence, clinical and functional significance of H3 mutations in T-ALL. In light of the recent discovery of oncogenic H3K37 mutations in AML (Lehnertz *et al*, 2017), our findings suggest a broader role for histone mutations in acute leukaemias and clearly justify incorporation of H3 genes into haematological cancer sequencing panels.

Acknowledgments

This work was supported by the Wellcome Trust. S.B. was funded by a Wellcome Trust Intermediate Clinical Research Fellowship and a St. Baldrick's Foundation Robert J. Arceci Innovation Award; G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); N.B. by AIRC (Associazione Italiana per la Ricerca sul Cancro) through a MFAG (n.17658); G.S.V by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA) and P.J.C. by a Wellcome Trust Senior Clinical Research Fellowship (WT088340MA). We thank Professor Adele Fielding for providing samples.

Authorship

S.B., M.R.S. and P.J.C. conceived and designed the study. G.C. and S.B. performed analysis with input from M.Y., I.M. and N.B. L.F. contributed materials. G.C. and S.B. wrote the manuscript with contributions from G.S.V. and P.J.C.

Conflict of interest

The authors have no competing financial interests to declare.

Grace Collord[1,2]
Inigo Martincorena[1]
Matthew D. Young[1]
Letizia Foroni[3,4]
Niccolo Bolli[5,6]
Michael R. Stratton[1]
George S. Vassiliou[1,7]
Peter J. Campbell[1,7]
Sam Behjati[1,2]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, [2]Department of Paediatrics, University of Cambridge, Cambridge, [3]Centre for Haematology, Faculty of Medicine, Imperial College London, [4]Clinical Haematology, Imperial College Healthcare NHS Trust, London,UK, [5]Department of Oncology and Haemato-Oncology, University of Milan, [6]Department of Oncology and Haematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan,Italy and [7]Department of Haematology, University of Cambridge, Cambridge, UK.

E-mails: pc8@sanger.ac.uk; sb31@sanger.ac.uk

Keywords: acute leukaemia, cancer genetics, aetiology, haematological malignancy

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Histone 3 mutation in T-ALL validation cohort.

Table SI. Type 3 histone genes.

Table SII. COSMIC version 81 cell lines screened for type 3 histone mutations.

Table SIII. T-cell leukaemia lines screened for type 3 histone mutations.

Table SIV. Internal database screened for histone 3 mutations.

Table SV. TCGA cohort screened for histone 3 mutations.

Table SVI. Validation cohort of 38 primary human T-ALL specimens screened by Sanger sequencing of histone 3 genes.

Table SVII. Primers used to Sanger sequence hotspot residues in histone 3 genes.

References

Atak, Z.K., Gianfelici, V., Hulselmans, G., De Keersmaecker, K., Devasia, A.G., Geerdens, E., Mentens, N., Chiaretti, S., Durinck, K., Uyttebroeck, A., Vandenberghe, P., Wlodarska, I., Cloos, J., Foa, R., Speleman, F., Cools, J. & Aerts, S. (2013) Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genetics*, **9**, e1003997.

Belver, L. & Ferrando, A. (2016) The genetics and mechanisms of T cell acute lymphoblastic leukaemia. *Nature Reviews Cancer*, **16**, 494–507.

Ben-Bassat, H., Shlomai, Z., Kohn, G. & Prokocimer, M. (1990) Establishment of a human T-acute lymphoblastic leukemia cell line with a (16;20) chromosome translocation. *Cancer Genetics and Cytogenetics*, **49**, 241–248.

Feinberg, A.P., Koldobskiy, M.A. & Gondor, A. (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, **17**, 284–299.

Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C., Shurtleff, S.A., Mulder, H.L., Boggs, K., Vadordaria, B., Cheng, J., Yergeau, D., Song, G., Becksfort, J., Lemmon, G., Weber, C., Cai, Z., Dang, J., Walsh, M., Gedman, A.L., Faber, Z., Easton, J., Gruber, T., Kriwacki, R.W., Partridge, J.F., Ding, L., Wilson, R.K., Mardis, E.R., Mullighan, C.G., Gilbertson, R.J., Baker, S.J., Zambetti, G., Ellison, D.W., Zhang, J. & Downing, J.R. (2014) The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nature Communications*, **5**, 3630.

Kuriyama, K., Gale, R.P., Tomonaga, M., Ikeda, S., Yao, E., Klisak, I., Whelan, K., Yakir, H., Ichimaru, M., Sparkes, R.S. (1989) CML-T1: a cell line derived from T-lymphocyte acute phase of chronic myelogenous leukemia. *Blood*, **74**, 1381–1387.

Lehnertz, B., Zhang, Y.W., Boivin, I., Mayotte, N., Tomellini, E., Chagraoui, J., Lavallee, V.P., Hebert, J. & Sauvageau, G. (2017) H3(K27M/I) mutations promote context-dependent transformation in acute myeloid leukemia with RUNX1 alterations. *Blood*, **130**, 2204–2214.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R. & Campbell, P.J. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, e1021.

Maser, R.S., Choudhury, B., Campbell, P.J., Feng, B., Wong, K.K., Protopopov, A., O'Neil, J., Gutierrez, A., Ivanova, E., Perna, I., Lin, E., Mani, V., Jiang, S., McNamara, K., Zaghlul, S., Edkins, S., Stevens, C., Brennan, C., Martin, E.S., Wiedemeyer, R., Kabbarah, O., Nogueira, C., Histen, G., Aster, J., Mansour, M., Duke, V., Foroni, L., Fielding, A.K., Goldstone, A.H., Rowe, J.M., Wang, Y.A., Look, A.T., Stratton, M.R., Chin, L., Futreal, P.A. & DePinho, R.A. (2007) Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature*, **447**, 966–971.

Leukemia
https://doi.org/10.1038/s41375-018-0270-2

BRIEF COMMUNICATION

Mechanisms of resistance

Targeting MEK in vemurafenib-resistant hairy cell leukemia

Rebecca Caeser[1,2] · Grace Collord[3,4] · Wen-Qing Yao[5] · Zi Chen[5] · George S. Vassiliou[1,3] · Philip A. Beer[3] · Ming-Qing Du[5] · Mike A. Scott[6] · George A. Follows[7] · Daniel J. Hodson[1,2,7]

Received: 17 July 2018 / Revised: 13 August 2018 / Accepted: 20 August 2018

Hairy cell leukemia (HCL) is a chronic, incurable B cell malignancy that presents with splenomegaly, bone marrow infiltration, and cytopenias [1]. Long remissions are typically achieved with purine analogs such as cladribine, but most cases will relapse and require further therapy. The discovery of the *BRAF* V600E mutation in almost all cases of HCL [2] has led to the widespread adoption of the BRAF inhibitor vemurafenib for treatment of patients relapsing after cladribine [3–5]. Impressive responses are reported; however, relapse is inevitable [5, 6] and hematologists are now faced with a growing number of patients with vemurafenib-resistant HCL. The optimal management of these patients remains unclear.

The molecular basis of vemurafenib resistance has been extensively investigated in recent years in patients with *BRAF* mutant solid organ malignancies such as melanoma and colorectal cancer [7]. Resistance to vemurafenib in melanoma frequently results from reactivation of ERK pathway signaling by a variety of genetic mechanisms that include activating mutations of *NRAS* or *KRAS*, amplification of mutant *BRAF*, aberrant splicing of *BRAF*, and activating mutation of *MAP2K1*, which encodes the MEK1 protein [7, 8]. ERK-independent mechanisms are less frequent and include activation of PI3K signaling [7]. The predominance of genetic mechanisms converging on ERK reactivation has led to the successful use of dual MEK/BRAF inhibition in melanoma [9]. In colorectal cancer however, different mechanisms apply; here primary resistance usually results from reduced feedback inhibition upon upstream receptor tyrosine kinase activity leading to restoration of ERK activity [10]. In this scenario, combined BRAF and MEK inhibition has not proved effective [11].

In contrast to our comprehensive understanding in solid organ cancer, very little is known about the mechanistic basis of vemurafenib resistance in HCL. Given the diversity of resistance mechanisms established in other cancers, it is unclear which, if any, of these might predominate in HCL. Two acquired subclonal, activating *KRAS* mutations were previously reported in a single patient with vemurafenib resistance [5]. Deletions of *NF1* and *NF2* have been proposed as an alternative mechanism in another case of primary resistance [12]. The use of MEK inhibition has been suggested as a logical therapeutic strategy in patients who have reactivated ERK signaling. However, the use of MEK inhibition has never previously been reported in a patient with HCL and at present there is no consensus on the optimal management of patients relapsing on vemurafenib.

A 74-year-old patient with HCL had been treated at our institution with splenectomy, cladribine, and pentostatin. We previously reported his initial response to vemurafenib at a dose of 240 mg twice daily [4]. This dose was lower than used in the initial phase II trial [5], but has since been shown in several reports to be an effective dosing strategy for HCL [3, 13, 14]. Vemurafenib was initially stopped after 58 days; however, this was associated with rapid return of marrow infiltration and thrombocytopenia. Vemurafenib was restarted at the same dose and cytopenias rapidly

These authors contributed equally: Rebecca Caeser, Grace Collord.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41375-018-0270-2) contains supplementary material, which is available to authorized users.

✉ Daniel J. Hodson
djh1002@cam.ac.uk

[1] Department of Haematology, University of Cambridge, Cambridge, UK

[2] Wellcome-MRC Cambridge Stem Cell Institute, Cambridge, UK

[3] Wellcome Sanger Institute, Hinxton, UK

[4] Department of Paediatrics, University of Cambridge, Cambridge, UK

[5] Division of Molecular Histopathology, University of Cambridge, Cambridge, UK

[6] Haematopathology & Oncology Diagnostic Service, Cambridge University Hospitals, Cambridge, UK

[7] Department of Haematology, Cambridge University Hospitals, Cambridge, UK

Published online: 19 October 2018

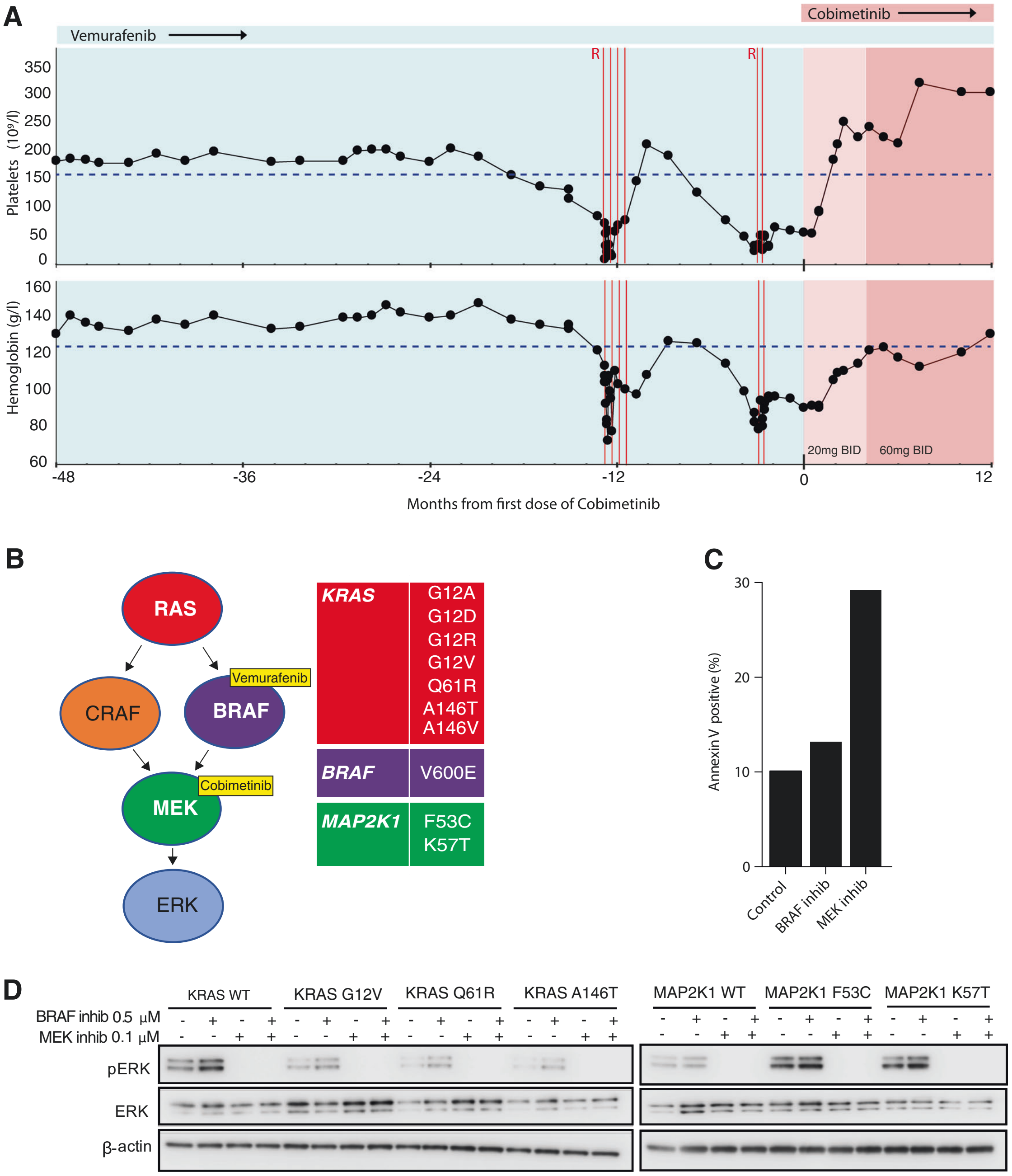


Fig. 1 a The patient's peripheral blood indices are shown over time relative to the first dose of the MEK inhibitor cobimetinib. Vertical red lines indicate the timing of rituximab dosing. Blue shading indicates vemurafenib monotherapy 240 mg twice daily (vem mono). Pale pink shading indicates vemurafenib with cobimetinib 20 mg daily (cobi-20). Darker pink indicates vemurafenib with cobimetinib 60 mg daily (21/28 days) (cobi-60). The lower limits of normal reference values are indicated by horizontal dashed lines. **b** Schematic of the MEK-ERK signaling pathway with mutations identified in purified tumor cells after emergence of resistance to vemurafenib. **c** Annexin V staining was used to quantify the induction of apoptosis in tumor cells purified from the patient and incubated for 48 h ex vivo with inhibitors of BRAF (vemurafenib) or MEK (trametinib). Apoptosis is induced by MEK inhibition but not by BRAF inhibition. **d** Immunoblots of a lymphoma cell line transduced with the indicated *KRAS* or *MAP2K1* constructs and incubated with inhibitors of BRAF or MEK. Complete suppression of ERK activity is seen with MEK inhibition but not with BRAF inhibition

resolved. Continuous low-dose vemurafenib continued to sustain his remission for over 3 years, attesting to the efficacy of this dosing schedule. However, 38 months after restarting vemurafenib, his blood indices deteriorated, and he required platelet transfusion (Fig. 1a). Bone marrow trephine biopsy confirmed relapse of HCL. A trial of rituximab with continued vemurafenib led to transient recovery of hematological indices. However, bone marrow infiltration did not improve over the next 4 months, and the patient became anemic, thrombocytopenic, and required

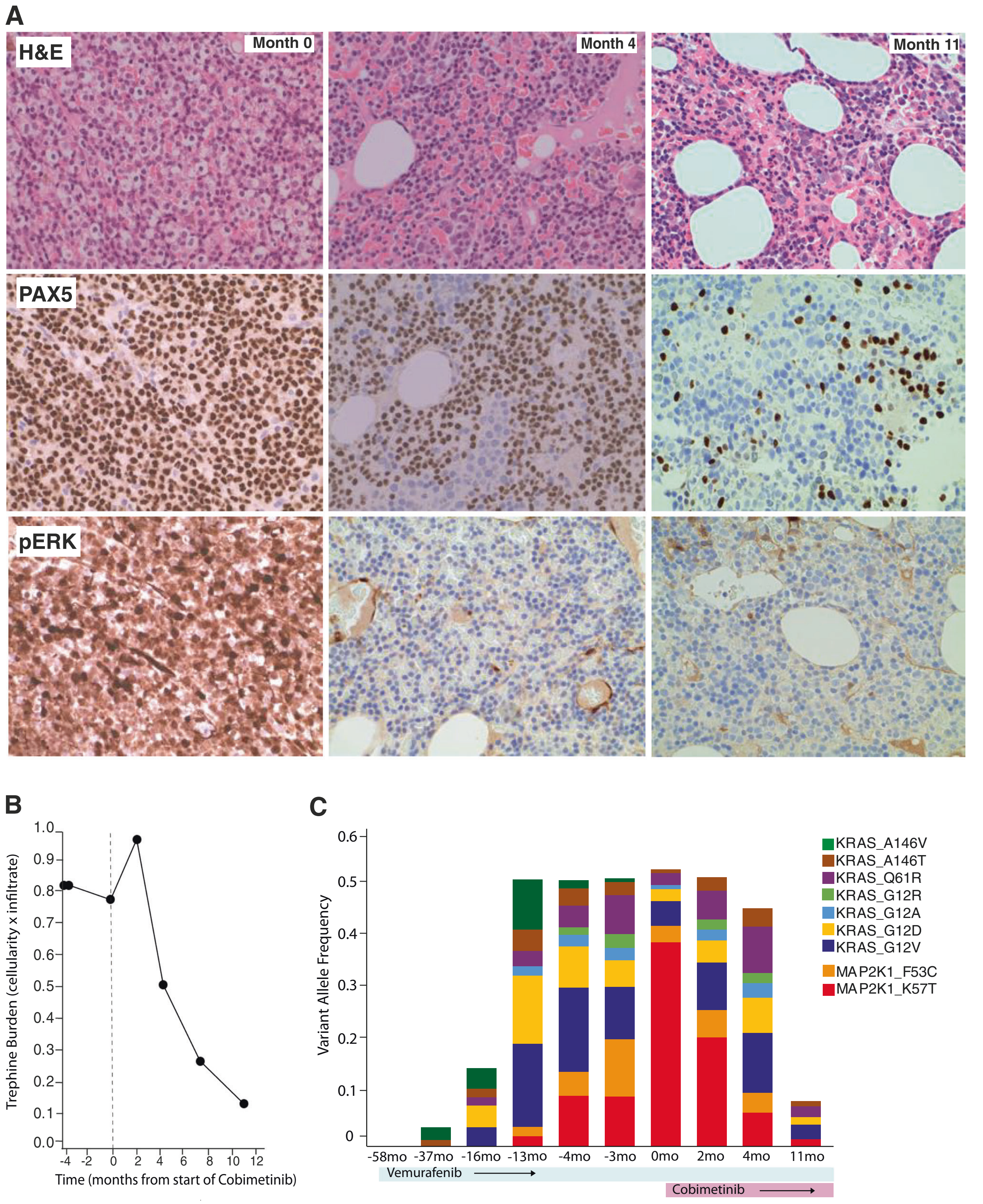


Fig. 2 **a** Bone marrow trephine biopsies stained with H&E (top) or PAX5 antibody (middle) or pERK (lower) taken at the indicated time points relative to start of cobimetinib. **b** Leukemic burden prior to and after starting cobimetinib therapy was calculated as the product of bone marrow trephine cellularity and leukemic cell infiltrate. **c** Mutant allele frequency for the indicated *KRAS* and *MAP2K1* mutations quantified by targeted amplicon sequencing at multiple time point relative to treatment

further platelet transfusion. A second trial of two doses of rituximab produced a minimal improvement of platelet count to 30×10^9/l. The patient became systemically unwell with B symptoms. Bone marrow trephine biopsy confirmed 99% infiltration with HCL.

To elucidate the mechanism of his resistance we performed whole-genome and deep-targeted sequencing of 292 genes (Supplementary Table 1) of DNA from purified tumor cells collected prior to starting vemurafenib and again at relapse. Samples were used with informed written patient consent in accordance with the Declaration of Helsinki and appropriate institutional ethical approvals. Sequencing studies revealed the presence of the known *BRAF* V600E mutation and chromosome 7q deletion. Remarkably, we also identified seven distinct activating mutations in *KRAS* and two mutations in *MAP2K1* (encoding MEK1) (Fig. 1b and Supplementary Table 2). These were detectable at relapse but were not detectable prior to vemurafenib exposure. Allele frequencies were consistent with the parallel, convergent evolution of multiple clones. Deep-targeted amplicon sequencing at multiple time points showed how *KRAS* mutations developed early, initially with codon 146 mutations, followed by emergence of the more classical activating mutations of codons 12 and 61 [15]. *MAP2K1* mutations appeared later with *MAP2K1* p.K57T expanding to become the dominant clone (Fig. 2c and Supplementary Table 2). The convergent nature of these mutations strongly implicated reactivation of MEK-ERK signaling as the likely mechanism of resistance. Indeed, immunohistochemistry confirmed strong pERK activity in all tumor cells (Fig. 2a). We looked for other mechanisms of resistance reported in melanoma. Specifically, we found no genetic or protein evidence of either increased pAKT activity or altered BRAF splicing (Supplementary Figure 1A & B).

We concluded that reactivation of MEK/ERK activity was the most likely driver of relapse and hypothesized that MEK inhibition might be a successful therapeutic strategy. Expression of the KRAS and MAP2K1 mutants in a lymphoid cell line showed that while each mutation was able to activate ERK in the presence of vemurafenib, all mutations remained sensitive to MEK inhibition (Fig. 1d). Exposure of the patient's purified tumor cells to vemurafenib ex vivo failed to completely suppress ERK activity and did not induce apoptosis. In contrast, ERK activity was completely suppressed and apoptosis induced by exposure to a MEK inhibitor (Supplementary Figure 1C and Fig. 1c).

Based on our in vitro experiments, we treated the patient with the MEK inhibitor cobimetinib, initially at 20 mg daily combined with vemurafenib 240 mg twice daily. Remarkably, B symptoms resolved within 1 week, followed by rapid platelet count recovery. A bone marrow biopsy at 4 months showed complete suppression of ERK activity (Fig. 2a). However, HCL marrow infiltration persisted, and therefore cobimetinib dose was increased to 60 mg daily (taken 21 out of 28 days). The dose was well tolerated and was associated with further resolution of cytopenias, a substantial reduction in bone marrow cellularity and HCL infiltration, ongoing suppression of ERK activity and restoration of normal hematopoiesis (Fig. 2a, b). Ongoing treatment was also associated with suppression of mutant allele frequencies for *BRAF*, *KRAS*, and *MAP2K1* mutations (Fig. 2c). At 12 months, the patient remains well and asymptomatic with continued combination therapy.

The finding of nine activating mutations, all converging upon the activation of RAS/RAF/MEK/ERK signaling, underscores the centrality of this pathway in HCL and its reactivation as a potent mechanism of resistance to vemurafenib in this disease. This report provides a detailed analysis of the molecular basis for acquired vemurafenib resistance in HCL. It is the first reported use of a MEK inhibitor to treat vemurafenib-resistant HCL. It proposes a new therapeutic option for such patients and provides impetus for testing in a formal trial setting.

Acknowledgements DJH was personally supported by a Clinician Scientist Fellowship from the Medical Research Council (MR/M008584/1), GC by a Wellcome Clinical PhD Fellowship (WT098051). W-QY was supported by an International Collaboration Award from the Pathological Society of UK and Ireland. Research in M.D. lab was supported by grants from Bloodwise. Core support was received from the Cancer Research UK, Cambridge Cancer Centre. We thank Joanna Baxter and Cambridge Blood and Stem Cell Bank for sample collection and storage, and Calli Latimer and Claire Hardy for expert technical assistance.

Author contributions RC designed and performed the experiments. GC analyzed the whole-genome sequencing and cRNA bait pulldown data. PAB designed the targeted gene pulldown panel. W-QY and ZC designed and conducted the targeted amplicon sequencing and analyzed the results. MSA, GAF, M-QD, GSV, and PAB provided clinical and diagnostic expertise and contributed to data interpretation. DJH designed the experiments, provided clinical expertise, directed the research, and wrote the manuscript with contributions from RC and GC.

Compliance with ethical standards

Conflict of interest DJH research funding: Gilead Sciences. GAF honoraria: Bayer AG, Roche, Gilead Sciences, Janssen Pharmaceuticals, and AbbVie. Consulting or advisory role: Bayer AG, Roche, Gilead Sciences, Janssen Pharmaceuticals, AbbVie. Speakers' bureau: Bayer AG, Roche, Gilead Sciences, Janssen Pharmaceuticals, and AbbVie. PAB employment: Karus.

References

1. Falini B, Martelli MP, Tiacci E. BRAF V600E mutation in hairy cell leukemia: from bench to bedside. Blood. 2016;128:1918–27.
2. Tiacci E, Trifonov V, Schiavoni G, Holmes A, Kern W, Martelli MP, et al. BRAF mutations in hairy-cell leukemia. N Engl J Med. 2011;364:2305–15.
3. Dietrich S, Glimm H, Andrulis M, von Kalle C, Ho AD, Zenz T. BRAF inhibition in refractory hairy-cell leukemia. N Engl J Med. 2012;366:2038–40.
4. Follows GA, Sims H, Bloxham DM, Zenz T, Hopper MA, Liu H, et al. Rapid response of biallelic BRAF V600E mutated hairy cell leukaemia to low dose vemurafenib. Br J Haematol. 2013;161:150–3.
5. Tiacci E, Park JH, De Carolis L, Chung SS, Broccoli A, Scott S, et al. Targeting mutant BRAF in relapsed or refractory hairy-cell leukemia. N Engl J Med. 2015;373:1733–47.
6. Holderfield M, Deuker MM, McCormick F, McMahon M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. Nat Rev Cancer. 2014;14:455–67.
7. Shi H, Hugo W, Kong X, Hong A, Koya RC, Moriceau G, et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. Cancer Discov. 2014;4:80–93.
8. Poulikakos PI, Persaud Y, Janakiraman M, Kong X, Ng C, Moriceau G, et al. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). Nature. 2011;480:
387–90.
9. Larkin J, Ascierto PA, Dreno B, Atkinson V, Liszkay G, Maio M, et al. Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. N Engl J Med. 2014;371:1867–76.
10. Corcoran RB, Ebi H, Turke AB, Coffee EM, Nishino M, Cogdill AP, et al. EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. Cancer Discov. 2012;2:227–35.
11. Kopetz S, Desai J, Chan E, Hecht JR, O'Dwyer PJ, Maru D, et al. Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. J Clin Oncol. 2015;33:4032–8.
12. Durham BH, Getta B, Dietrich S, Taylor J, Won H, Bogenberger JM, et al. Genomic analysis of hairy cell leukemia identifies novel recurrent genetic alterations. Blood. 2017;130:1644–8.
13. Peyrade F, Re D, Ginet C, Gastaud L, Allegra M, Ballotti R, et al. Low-dose vemurafenib induces complete remission in a case of hairy-cell leukemia with a V600E mutation. Haematologica. 2013;98:e20–2.
14. Dietrich S, Pircher A, Endris V, Peyrade F, Wendtner CM, Follows GA, et al. BRAF inhibition in hairy cell leukemia with low-dose vemurafenib. Blood. 2016;127:2847–55.
15. Janakiraman M, Vakiani E, Zeng Z, Pratilas CA, Taylor BS, Chitale D, et al. Genomic and biological characterization of exon 4 KRAS mutations in human cancer. Cancer Res. 2010;70:5901–11.

Clonal haematopoiesis is not prevalent in survivors of childhood cancer

Clonal haematopoiesis driven by leukaemia-associated somatic mutations is a common feature of ageing (Link & Walter, 2016). This phenomenon, termed clonal haematopoiesis of indeterminate potential (CHIP), is associated with an increased risk of haematological malignancies and all-cause mortality (Link & Walter, 2016). Recent empirical evidence and computational models suggest that mutation acquisition may not be the major rate-limiting factor in the emergence of CHIP (Altrock *et al*, 2015; McKerrell *et al*, 2015; Link & Walter, 2016; Young *et al*, 2016). Instead, clonal expansion of mutant haematopoietic stem cells (HSCs) probably reflects the interaction between the effects of driver mutations and selection pressures prevailing in the bone marrow microenvironment (Link & Walter, 2016). Notably, cytotoxic therapies have been shown to favour expansion of pre-malignant haematopoietic clones (Link & Walter, 2016). Furthermore, both adult and paediatric cancer patients treated with intensive chemoradiotherapy display an earlier onset of ageing-associated morbidities and an elevated risk of therapy-related myeloid neoplasms (t-MN) and other secondary malignancies (Rowland & Bellizzi, 2014). A recent study in adult cancer patients found that CHIP was more prevalent than in the general population and was strongly associated with t-MN and overall mortality (Gibson *et al*, 2017). Although CHIP is extremely rare in healthy young individuals, its prevalence and prognostic significance in paediatric cancer patients has not been studied. We therefore performed targeted deep sequencing of peripheral blood DNA from 84 childhood cancer survivors to search for subclonal mutations common in t-MN and adult clonal haematopoiesis. No individuals with somatic variants at these loci were identified. Whilst our findings could be explained by a rarity of driver mutations, the fact that human HSCs accrue somatic variants from the first decade of life (Welch *et al*, 2012) proposes the alternative possibility that such mutations may not confer clonal advantage in the young.

We obtained peripheral blood DNA samples from patients enrolled on long-term follow-up after treatment for a paediatric malignancy and from three age-matched controls with no cancer history. Written informed consent was obtained for sample collection and DNA sequencing from all patients or their guardian in accordance with the Declaration of Helsinki and protocols approved by the relevant institutional ethics committees (approval numbers 09REG2015, 1-09/12/2015). The median age at cancer diagnosis was 4·5 years, and the commonest malignancies were acute lymphoblastic leukaemia (*n* = 21), neuroblastoma (*n* = 17) and non-Hodgkin lymphoma (*n* = 10). Nineteen patients had received a HSC transplant (8 allogeneic and 11 autologous). The median interval between completion of cancer treatment and blood sampling was 6 years (range 2–25). Patient characteristics are summarized in Table SI.

We performed targeted next generation sequencing (NGS) using multiplex polymerase chain reaction to amplify 32 regions of 14 genes frequently mutated in CHIP or t-MN (Table I) (McKerrell *et al*, 2015; Link & Walter, 2016; Gibson *et al*, 2017). For this we extended a previously validated assay that detected clonal haemopoiesis in 2·6% of unselected adults (McKerrell *et al*, 2015), to include all coding exons of *TP53* and *PPM1D*, genes implicated in t-MN pathogenesis (Rowland & Bellizzi, 2014; Link & Walter, 2016; Gibson *et al*, 2017). Primer design and sequencing was performed as described previously (McKerrell *et al*, 2015); see Table SII for primer sequences. Reads were aligned to human genome build GRCh37 using the Burrows-Wheeler Aligner (Li & Durbin, 2010) and analysed for somatic single nucleotide variants. Allele counts were generated using an in-house script (https://github.com/cancerit/alleleCount), considering only loci with ≥1000 reads and minimum base and mapping quality of 25 and 35, respectively. Somatic mutations with

Table I. Genomic regions sequenced.

Gene	Chromosome	Target codon/exon
NRAS	1	p.G12
SF3B1	2	p.K666; p.K700
DNMT3A	2	p.R882
IDH1	2	p.R132
KIT	4	exon 17
NPM1	5	exon 12
JAK2	9	p.V617
KRAS	12	p.G12
IDH2	15	p.R140; p.R172
PPM1D	17	exons 1–6
TP53	17	exons 1–12
SRSF2	17	p.P95
ASXL1	20	exon 12
U2AF1	21	p.S34; p.Q157

doi: 10.1111/bjh.14630

variant allele frequency (VAF) ≥0·008 (McKerrell *et al*, 2015) were sought and examined visually and by interrogation with the Shearwater algorithm (https://github.com/mg14/deepSNV) (Gerstung *et al*, 2014).

The median sequencing depth across regions of interest was 5·3 × 10^3. No somatic mutations with VAF ≥ 0·008 were observed in any of our patients or controls, demonstrating that CHIP driven by mutations at these loci is not prevalent in young individuals who have received cytotoxic treatment. By contrast, Gibson *et al* (2017) identified post-chemotherapy CHIP (VAF > 0·02) in 29·9% of 401 adult lymphoma patients. Notably, mutations in *PPM1D*, a regulator of *TP53*, were the commonest CHIP drivers (Gibson *et al*, 2017). Similarly, several smaller studies have demonstrated clonal expansion in older patients undergoing chemoradiotherapy for other cancers (Link & Walter, 2016). An investigation of haematopoietic clonal dynamics in 15 adult acute myeloid leukaemia patients found that, after induction chemotherapy, five had marked expansion of clones unrelated to their leukaemia (Link & Walter, 2016). Most clones carried canonical leukaemia mutations and continued to expand years after remission (Link & Walter, 2016). In a study exploring the clonal origins of t-MN, *TP53*-mutated clones expanded dramatically after cytotoxic treatment, whereas the same mutations demonstrated very modest clonal advantage in healthy individuals (Link & Walter, 2016). In light of the above, our findings have two plausible explanations: (i) that somatic driver mutations are very uncommon in young individuals even after exposure to chemotherapy or (ii) that accrual of such mutations is insufficient to trigger clonal expansion in this age group. The latter is supported by findings that oncogenic mutations begin accumulating early in life (Welch *et al*, 2012) and that cancer-associated mutations are less able to drive clonal expansion in young compared to old stem cells (Zhu *et al*, 2016). The fact that bona-fide driver mutations do not always lead to haematopoietic clonal expansion, even after several years, was highlighted by Young *et al* (2016), using ultra-sensitive sequencing. Therefore our results should not be taken to reflect absence of potentially oncogenic HSC mutations in young cancer survivors. Rather, it is possible that even canonical leukaemogenic mutations may not commonly drive clonal outgrowth in children and young adults despite exposure to extreme haematopoietic stress, implicating age-related changes in HSCs and/or their microenvironment as key determinants of relative fitness. More sensitive DNA sequencing methods may enable detection of very rare cells harbouring known CHIP drivers mutations in similar patient cohorts, which would lend support to this hypothesis. Studies of larger numbers of paediatric cancer survivors are needed to identify rare individuals with CHIP after chemoradiotherapy, whose particular characteristics may offer insights into factors facilitating clonal outgrowth of mutated HSCs. Furthermore, in view of the shifting patterns of mutations driving CHIP in different adult age groups (McKerrell *et al*, 2015), selective pressures particular to a less mature bone marrow environment may confer clonal advantage on a distinct spectrum of somatic variants in the very young. Although a much broader screening approach is required to identify such mutations, the potential role for CHIP as a biomarker for patient risk-stratification (Gibson *et al*, 2017) may render this a worthwhile endeavour.

Acknowledgements

This project was funded by the Wellcome Trust Sanger Institute (grant number WT098051). G.S.V. is funded by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA). F.F. is funded by Compagnia di San Paolo Grant: “Le cellule staminali del sangue nei guariti di leucemia” Codice SIME 2013-0958 (codice ROL 4201). I.V is funded by the Spanish Ministerio de Economía y Competitividad, Programa Ramón y Cajal.

Author contributions

GSV, GC and FF conceived and designed the study. NH designed sequencing assays. GC performed experiments and bioinformatics analysis. GC and GSV wrote the manuscript with input from FF. DJ and IV wrote scripts and contributed to analysis strategy. FF, MP, MD and DC contributed to sample acquisition and patient recruitment.

Grace Collord[1,2]
Naomi Park[1]
Marina Podestà[3]
Monica Dagnino[3]
Daniela Cilloni[4]
David Jones[1]
Ignacio Varela[5]
Francesco Frassoni[3,*]
George S. Vassiliou[1,6,7,*]

[1]Wellcome Trust Sanger Institute, [2]Department of Paediatrics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, [3]Laboratorio Cellule Staminali e Terapie Cellulari, Istituto Giannina Gaslini IRCCS, Genova, [4]Department of Clinical and Biological Sciences, University of Turin, Turin, Italy, [5]Instituto de Biomedicina y Biotecnología de Cantabria, Cantabria, Spain, [6]Department of Haematology, University of Cambridge, and [7]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

E-mails: gsv20@sanger.ac.uk, francesco.l.frassoni@gmail.com

**Contributed equally.*

Keywords: haematopoiesis, late effects of therapy, haematopoietic stem cells, paediatric cancer, clonal evolution

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table SI. Patient characteristics
Table SII. Primer sequences

References

Altrock, P.M., Liu, L.L. & Michor, F. (2015) The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, **15**, 730–745.

Gerstung, M., Papaemmanuil, E. & Campbell, P.J. (2014) Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics (Oxford, England)*, **30**, 1198–1204.

Gibson, C.J., Lindsley, R.C., Tchekmedyian, V., Mar, B.G., Shi, J., Jaiswal, S., Bosworth, A., Francisco, L., He, J., Bansal, A., Morgan, E.A., Lacasce, A.S., Freedman, A.S., Fisher, D.C., Jacobsen, E., Armand, P., Alyea, E.P., Koreth, J., Ho, V., Soiffer, R.J., Antin, J.H., Ritz, J., Nikiforow, S., Forman, S.J., Michor, F., Neuberg, D., Bhatia, R., Bhatia, S. & Ebert, B.L. (2017) Clonal hematopoiesis associated with adverse outcomes after autologous stem-cell transplantation for lymphoma. *Journal of Clinical Oncology*, JCO2016716712. [Epub ahead of print]

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **26**, 589–595.

Link, D.C. & Walter, M.J. (2016) 'CHIP'ping away at clonal hematopoiesis. *Leukemia*, **30**, 1633–1635.

McKerrell, T., Park, N., Moreno, T., Grove, C.S., Ponstingl, H., Stephens, J., Understanding Society Scientific Group, Crawley, C., Craig, J., Scott, M.A., Hodkinson, C., Baxter, J., Rad, R., Forsyth, D.R., Quail, M.A., Zeggini, E., Ouwehand, W., Varela, I. & Vassiliou, G.S. (2015) Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Reports*, **10**, 1239–1245.

Rowland, J.H. & Bellizzi, K.M. (2014) Cancer survivorship issues: life after treatment and implications for an aging population. *Journal of Clinical Oncology*, **32**, 2662–2668.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., Kandoth, C., Fulton, R.S., McLellan, M.D., Dooling, D.J., Wallis, J.W., Chen, K., Harris, C.C., Schmidt, H.K., Kalicki-Veizer, J.M., Lu, C., Zhang, Q., Lin, L., O'Laughlin, M.D., McMichael, J.F., Delehaunty, K.D., Fulton, L.A., Magrini, V.J., McGrath, S.D., Demeter, R.T., Vickery, T.L., Hundal, J., Cook, L.L., Swift, G.W., Reed, J.P., Alldredge, P.A., Wylie, T.N., Walker, J.R., Watson, M.A., Heath, S.E., Shannon, W.D., Varghese, N., Nagarajan, R., Payton, J.E., Baty, J.D., Kulkarni, S., Klco, J.M., Tomasson, M.H., Westervelt, P., Walter, M.J., Graubert, T.A., DiPersio, J.F., Ding, L., Mardis, E.R. & Wilson, R.K. (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell*, **150**, 264–278.

Young, A.L., Challen, G.A., Birmann, B.M. & Druley, T.E. (2016) Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications*, **7**, 12484.

Zhu, L., Finkelstein, D., Gao, C., Shi, L., Wang, Y., Lopez-Terrada, D., Wang, K., Utley, S., Pounds, S., Neale, G., Ellison, D., Onar-Thomas, A. & Gilbertson, R.J. (2016) Multi-organ mapping of cancer risk. *Cell*, **166**, 1132–1146.e7.