BIOINFORMATICS 1

or why biologists need computers

http://www.bioinformatics.uni-muenster.de/teaching/courses-2012/bioinf1/index.hbi

Prof. Dr. Wojciech Makałowski
Institute of Bioinformatics

1

SEQUENCING TECHNOLOGY

bioinformatic challenges

Prof. Dr. Wojciech Makałowski
Institute of Bioinformatics

2

DNA story

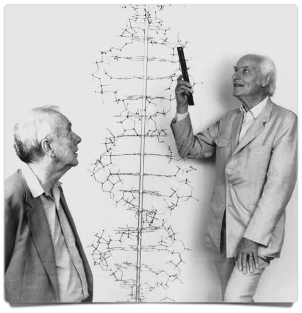
1870 Friedrich Miescher discovers DNA

1944 Oswald Avery proves that DNA is a genetic material

3

DNA story

1953 James Watson and Francis Crick discover DNA structure



4

sequencing -beginnings

1964 Robert Holley determines nucleotide sequences (77 nt) of the yeast Alanine tRNA
J. Biol. Chem. 240: 2122-2128

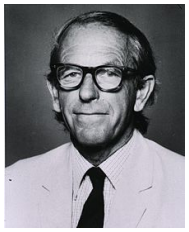
1968 Ray Wu and Dale Keiser sequenced 12 bases (!) of λ phage's 5' cohesive ends using chain termination and polyacrylamide gel electrophoresis J. Mol. Biol. 35: 523-537

5

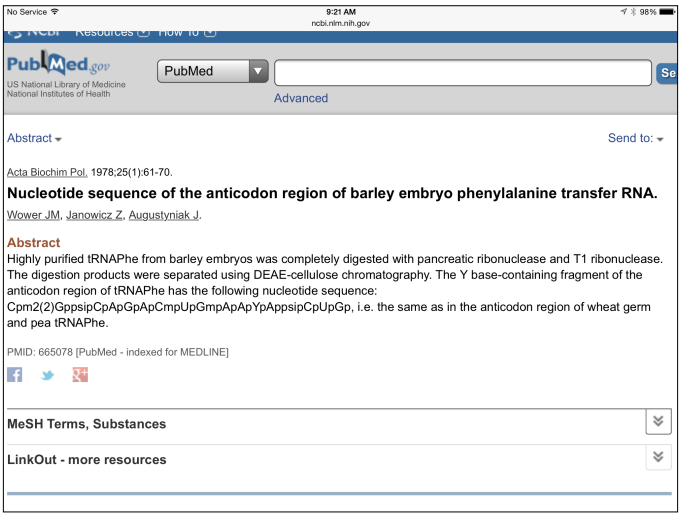
sequencing - infancy

1977 - Allan Maxam and Walter Gilbert develop DNA sequencing method by chemical degradation

1977 Fred Sanger develops 2',3'-dideoxy chain termination method



6



7

chemical degradation sequencing

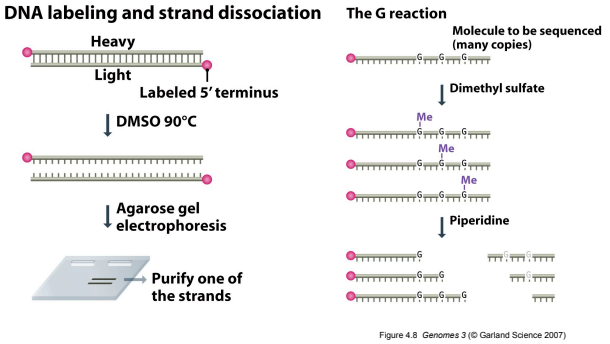
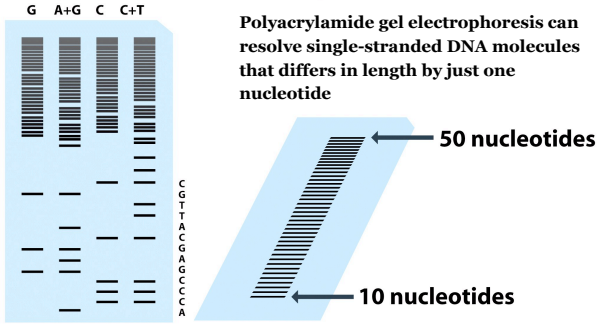


Figure 4.8 *Genomes 3* (© Garland Science 2007)

8

chemical degradation sequencing

Reading the sequence from the autoradiograph



Polyacrylamide gel electrophoresis can resolve single-stranded DNA molecules that differs in length by just one nucleotide

9

chain termination DNA sequencing

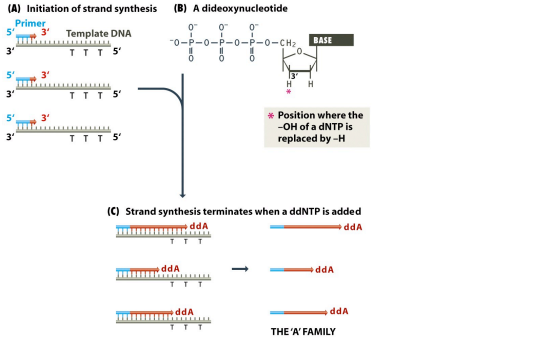


Figure 4.2 *Genomes 3* (© Garland Science 2007)

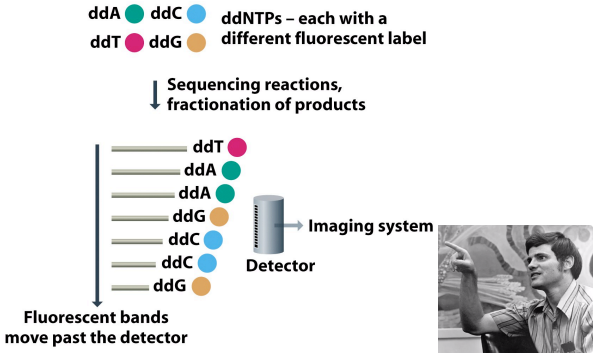
10

sequencing - maturity

- 1983 - Marvin Carruthers developed a method to construct fragments of DNA of predetermined sequence from five to about 75 base pairs long. He and Leroy Hood invented instruments that could make such fragments automatically.
- 1983 - Kary Mullis invented the polymerase chain reaction (PCR) technique
- 1987 - ABI 370; first fully automated sequencing machine
- 1995 - Craig Venter uses whole-genome shotgun sequencing technique to determine complete genome of bacterium Haemophilus influenzae
- 2005 - introduction of GS20 sequencing machine; first in the line of "Next Generation Sequencing"

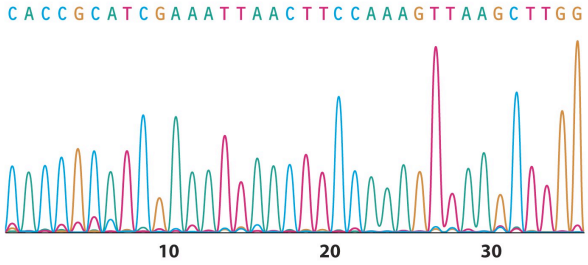
11

sequencing - maturity



12

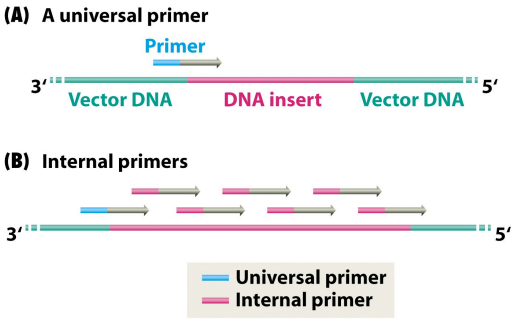
sequencing - maturity



Chromatogram of a DNA sequence generated by ABI sequencing machine

13

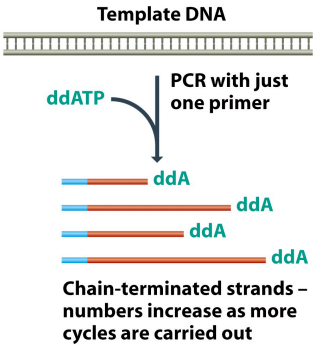
sequencing - maturity



Different types of primer for chain termination sequencing

14

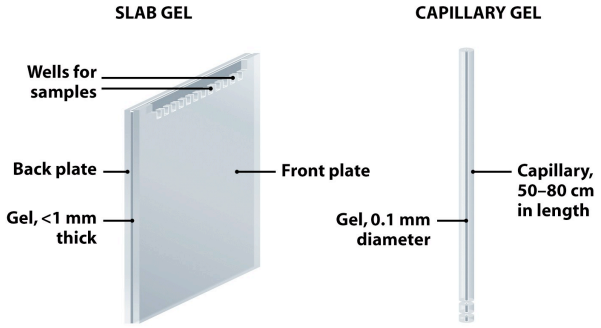
sequencing - maturity



Thermal cycle sequencing. PCR is carried out with just one primer and with the four dideoxynucleotides present in the reaction mixture. The result is a set of chain-terminated strands - the "A" family shown to the left. These strands are separated using electrophoresis methodology

15

sequencing - maturity



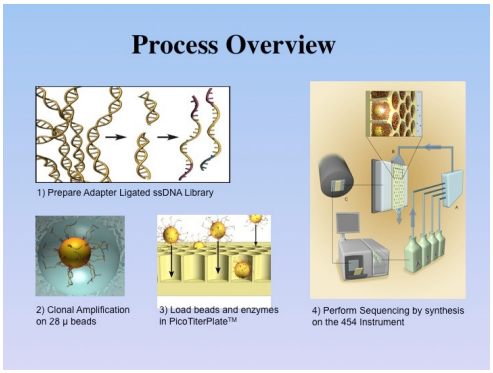
16

next generation sequencing

- Massive parallelization of the sequencing process
- Relatively short reads
- Different approaches from improving Sanger's technique to direct "observation" of DNA through a microscope
- Attempts to sequence single molecules without amplification step

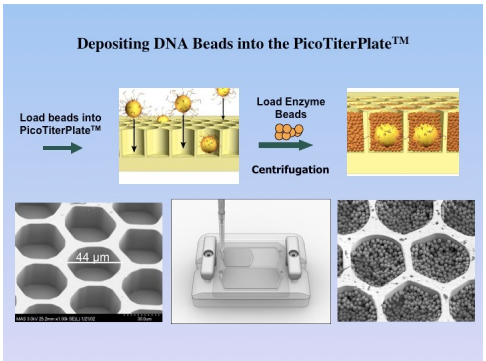
17

NGS - pyrosequencing



18

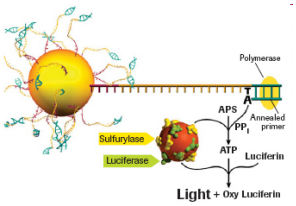
NGS - pyrosequencing



19

NGS - pyrosequencing

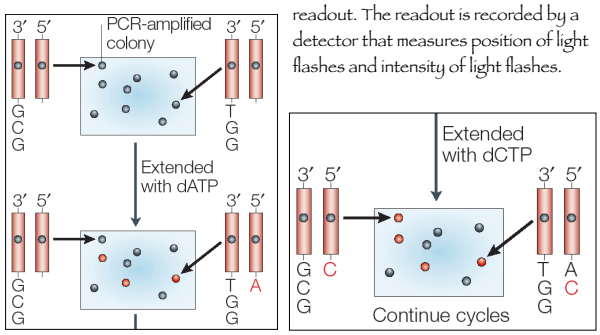
- After the emulsion PCR has been performed, the oil is removed, and the beads are put into a "picotiter" plate. Each well is just big enough to hold a single bead.
- The pyrosequencing enzymes are attached to much smaller beads, which are then added to each well.
- The plate is then repeatedly washed with the each of the four dNTPs, plus other necessary reagents, in a repeating cycle.
- The plate is coupled to a fiber optic chip. A CCD camera records the light flashes from each well.



20

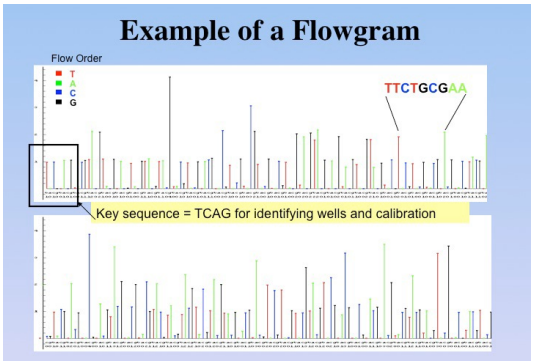
NGS - pyrosequencing

Extension with individual dNTPs gives a readout. The readout is recorded by a detector that measures position of light flashes and intensity of light flashes.



21

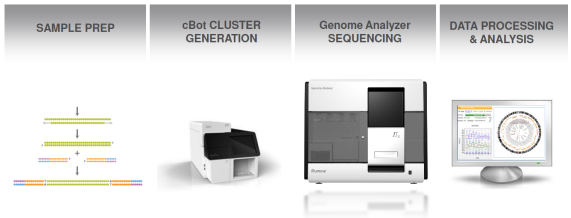
NGS - pyrosequencing



22

NGS - Illumina

Workflow



23

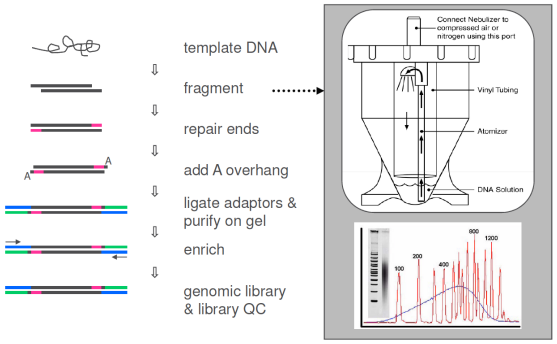
NGS - Illumina
The flow cell - a core component

EVERYTHING EXCEPT SAMPLE PREPARATION IS COMPLETED ON THE FLOW CELL

template annealing (1 - 96 samples)
template amplification
sequencing primer hybridization
Sequencing-by-synthesis reaction
generation of fluorescent signal

24

NGS - Illumina
Preparation of template

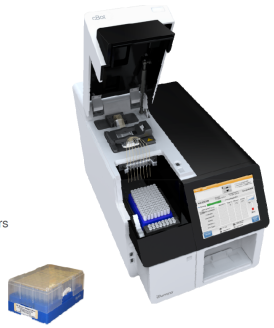


25

NGS - Illumina
The flow cell is mounted on the cBot

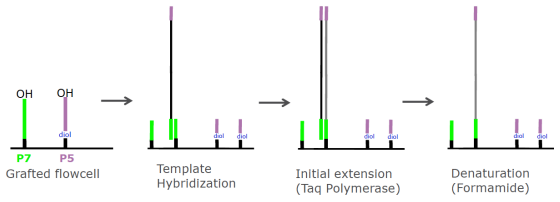
AUTOMATICALLY
loads library into the lanes of the flow cell
amplifies templates
anneals sequencing primer to templates

FEATURES
intervention-free clonal amplification in 4 hours
simple touch screen operation



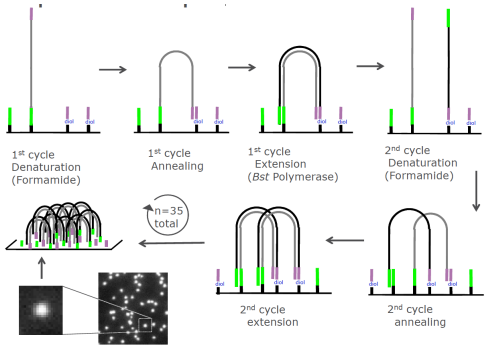
26

NGS - Illumina
Hybridization of template



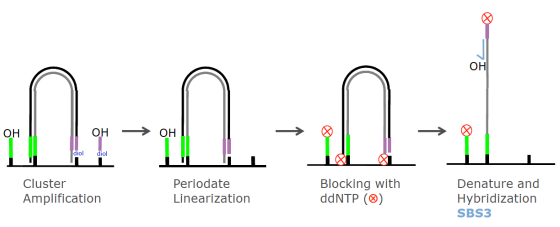
27

NGS - Illumina
Amplification of template



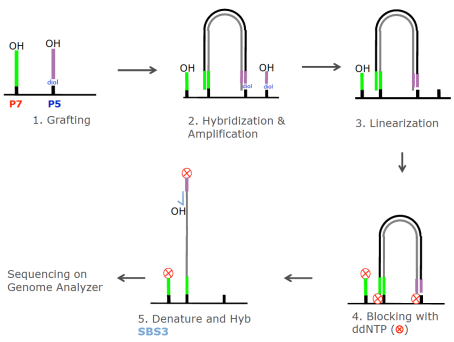
28

NGS - Illumina
Annealing of sequencing primer to template



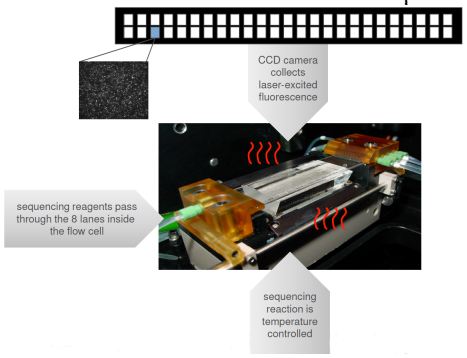
29

NGS - Illumina
Summary - "cluster generation"



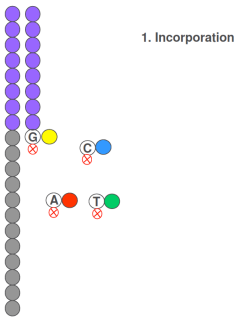
30

NGS - Illumina
The flow cell is mounted on the sequencer



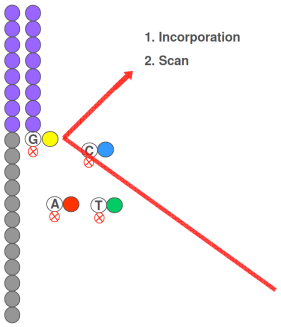
31

NGS - Illumina
Incorporation



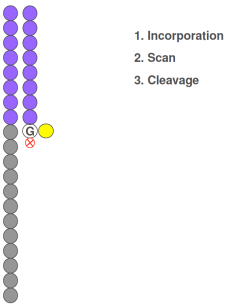
32

NGS - Illumina
Scanning



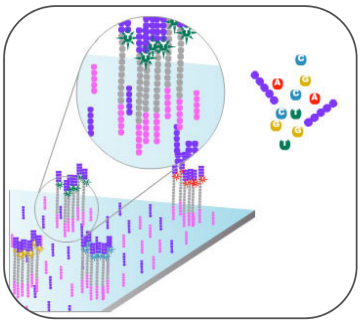
33

NGS - Illumina
Cleavage



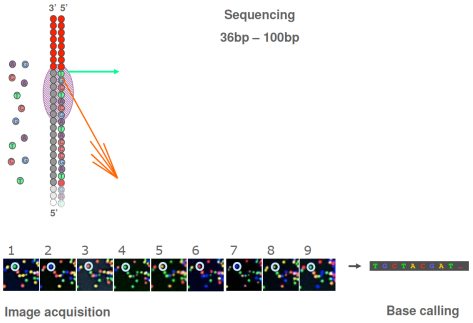
34

NGS - Illumina
Millions of clusters are sequenced in parallel



35

NGS - Illumina
A picture is taken every time a new base is added



36

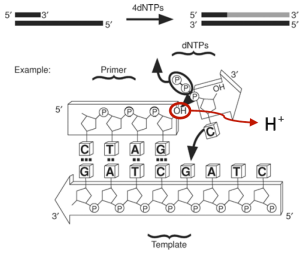
NGS -ion torrent

- Ten times faster workflow than other NGS systems
- ~2 hour sequencing runs (real-time detection of sequence extension)
- Batch sample preparation (six samples in six hours)
- Capable of six samples/day on two PGM Systems

37

NGS -ion torrent
Simple Natural Chemistry

Sequencing by synthesis

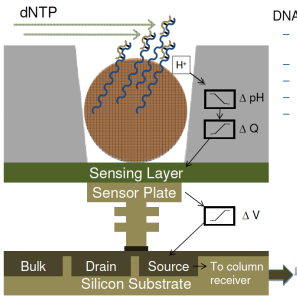


Eliminate source of sequencing errors:
- Modified bases
- Fluorescent bases
- Laser detection
- Enzymatic amplification cascades

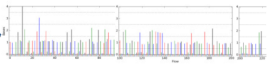
Eliminate source of read length limitations:
- Unnatural bases
- Faulty synthesis
- Slow cycle time

38

NGS -ion torrent
Fast Direct Detection



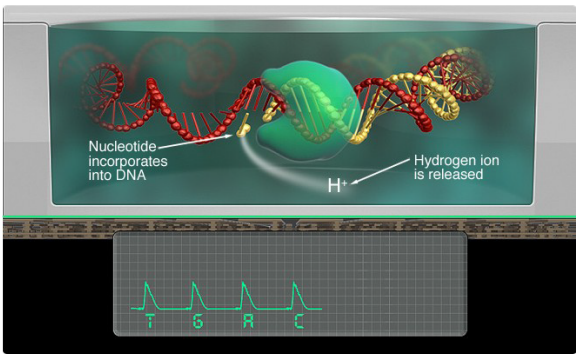
DNA → Ions → Sequence
- Nucleotides flow sequentially over Ion semiconductor chip
- One sensor per well per sequencing reaction
- Direct detection of natural DNA extension
- Millions of sequencing reactions per chip
- Fast cycle time, real time detection



39

NGS -ION TORRENT

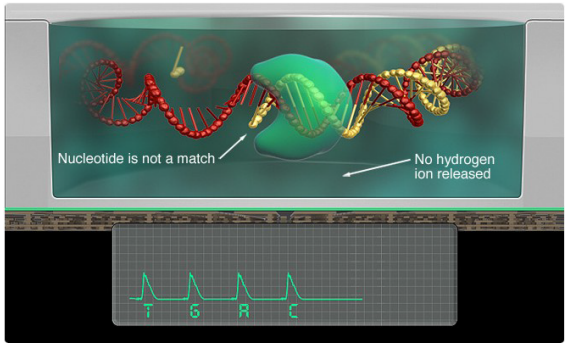
Four nucleotides flow sequentially



40

NGS -ION TORRENT

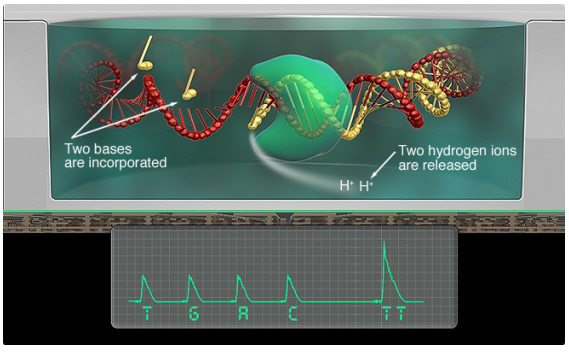
Four nucleotides flow sequentially



41

NGS -ION TORRENT

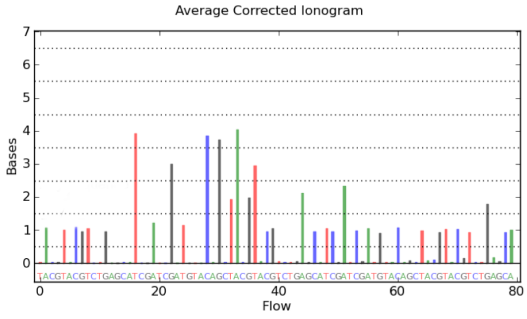
Four nucleotides flow sequentially



42

NGS -ION TORRENT

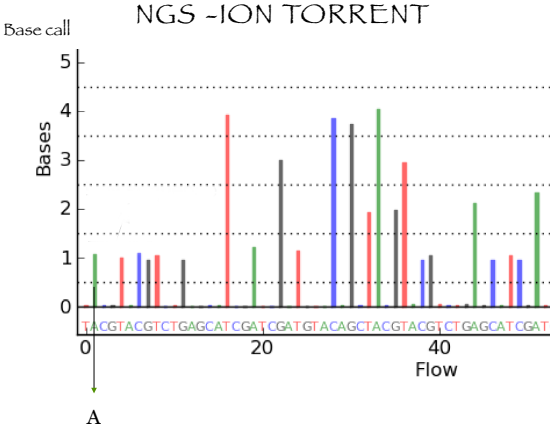
Base call



43

NGS -ION TORRENT

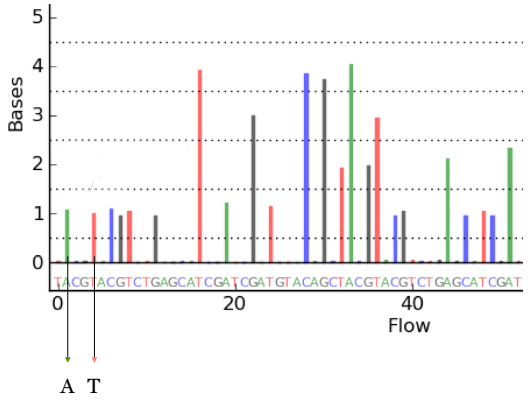
Base call



44

NGS -ION TORRENT

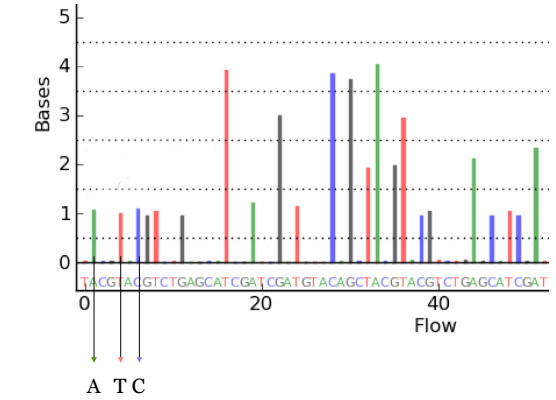
Base call



45

NGS -ION TORRENT

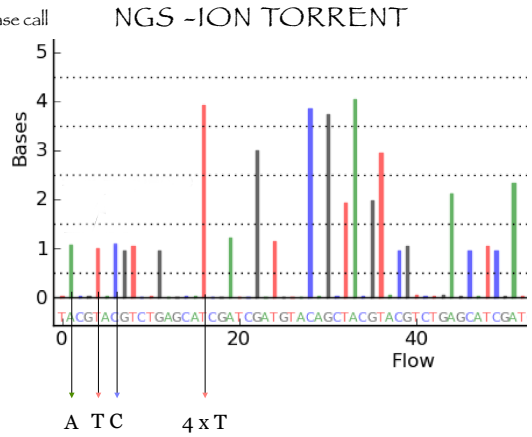
Base call



46

NGS -ION TORRENT

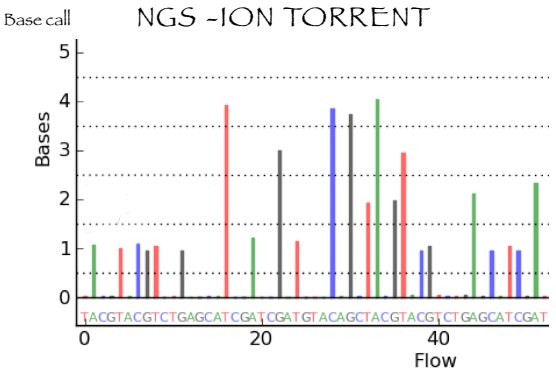
Base call



47

NGS -ION TORRENT

Base call



ATCGTGTTTTAGGGTCCCCGGGGTTAAAA...

48

Third generation sequencing

https://www.youtube.com/watch?v=_B_cUZ8hSYU

49

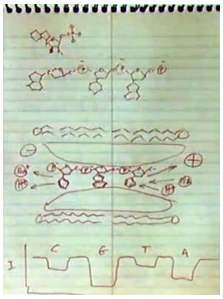
Third generation sequencing

Single molecule sequencing: MinION by Oxford Nanopore

50

Sequencing using nanopores

- Nanopores as polymer sensors.
- The idea emerged in early 1990s.
- Fundamental work done by David Deamer and Daniel Branton in collaboration with John Kasianowicz. (PNAS 1996 146:13770-13773)
- Hundreds of papers and patents since then.

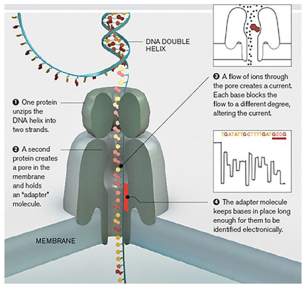


51

MinION basics

https://nanoporetech.com/science-technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing

- Synthetic membrane
- Nanopore is created by modified α-hemolysin
- Non-destructive motor protein (actually serves as a break)

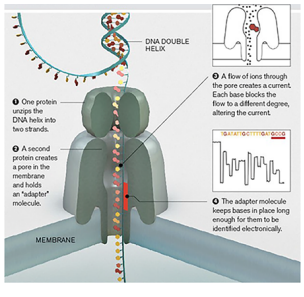


52

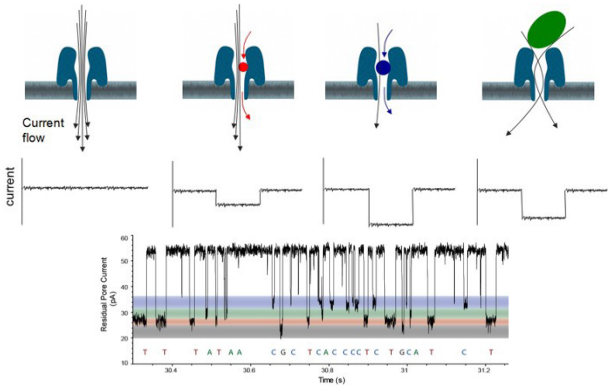
MinION basics

https://nanoporetech.com/science-technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing

- 512 channels (pores) per flow cell. Usually about 90% are working.
- Read length: > 10 Kb (Phage λ DNA, 50 Kb)
- Read speed: 8 bases to 20 bases/sec
- Run time: max 48 hours
- Error rate = 5-10 %
- Sequence yield per flow cell: 0.5 - 1.4 Gb



53

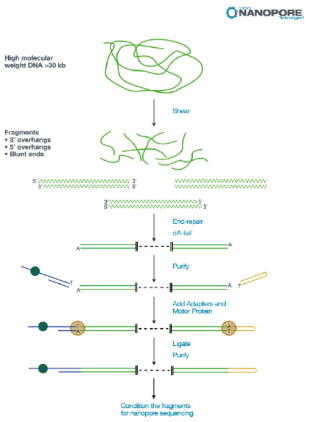


54

Easy, standard template preparation

Time of library preparation:
1D - about ten min
2D - up to two hours

Cost of a single run:
reagents $1000
flow cell $1000



55

MinION dataflow

MinION

Nanopore sensing is carried out on the sensor chip, contained in the flow cell inside the MinION device. Data is processed by an Application-Specific Integrated Circuit (ASIC) also in the flow cell and processed in real time by the MinKNOW software

MinKNOW

MinKNOW is the software that controls the MinION. It carries out several core data tasks and can be used to change experimental workflows or parameters. MinKNOW runs on the user's computer.

METRICHOR

Metrichor is an on-demand, cloud-based, bioinformatics data analysis platform. It supports Oxford Nanopore base calling software. Base calling may be made available locally .

56

MiniKNOW - Data Render

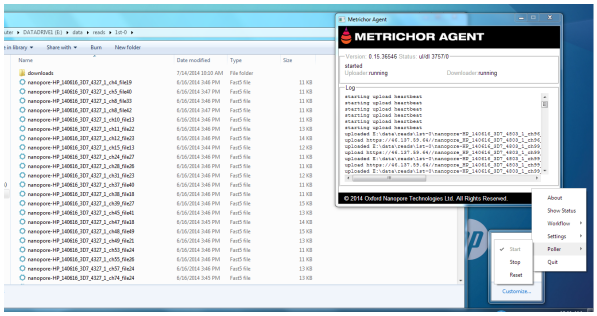
57

MiniKNOW - Channels Panel

58

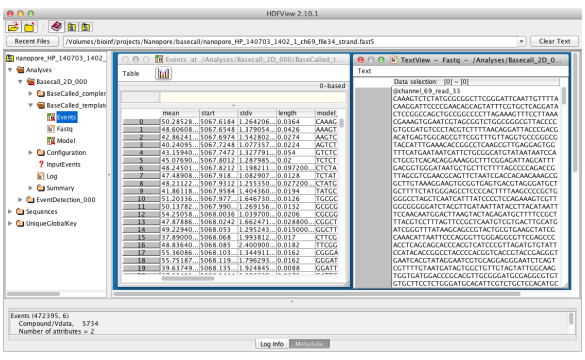
Metrichor

Raw data, BaseCall data -> .Fast5 file



59

HDFView



60

Advantage of nanopore technology

- Label-free
- Single molecule, long reads analysis
- Disposable; autoclavable after the use
- Portable; requires no pre-installation of any instruments

61

Numerous applications explored by MinION Access Program (MAP)

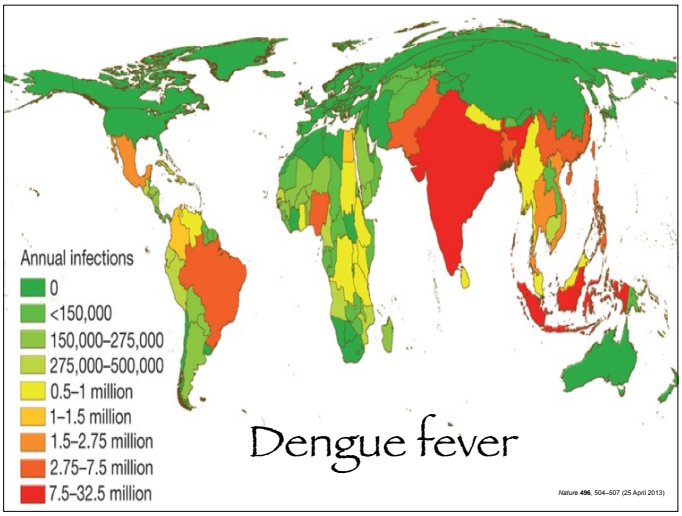
- Genomic DNA sequencing
- Metagenomic analysis
- Direct RNA sequencing
- Species identification in the field
- Splice variants identification
- Direct determination of modified nucleotides
- And many more to come…

62

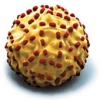
Potential for tropical diseases research and diagnostics

- In many countries where tropical diseases prevail
  - no conventional sequencer/ PCR instruments are available
  - shortage of well-trained technical staff
  - Needs for handling potentially dangerous pathogens

63



64



Dengue fever

- Transmitted by a bite of mosquito infected with dengue virus (genome size almost 11 kb)
- Febrile illness that affects infants, young children and adults with symptoms appearing 3-14 days after the infective bite.
- There are four serotypes (D1~ D4), whose genomes are about 70% identical one to each other.
- Second infection of the same serotype may cause severe symptoms; dengue hemorrhagic fever, abdominal pain, persistent vomiting, bleeding and breathing difficulty and is a potentially lethal.

65

Sample preparation

LAMP Amplification

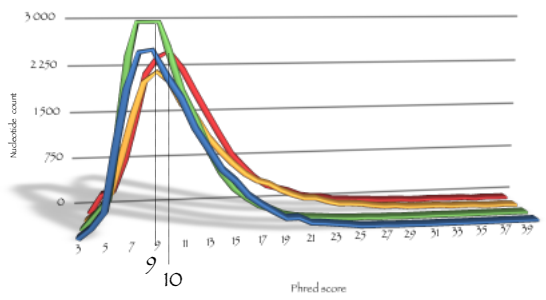
Serum (1 - 5 µL) -> Mix with Dry Lamp reagent kit -> 65 °C for 60 min -> Purification (AMPure) ->

Nonopore Sample Prep

66

C G A T

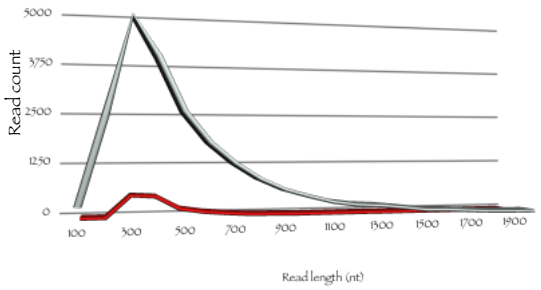
Distribution of quality scores for each nucleotide



67

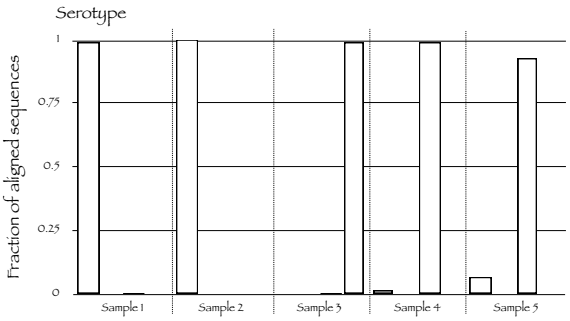
Fail Pass

Read length distribution



68

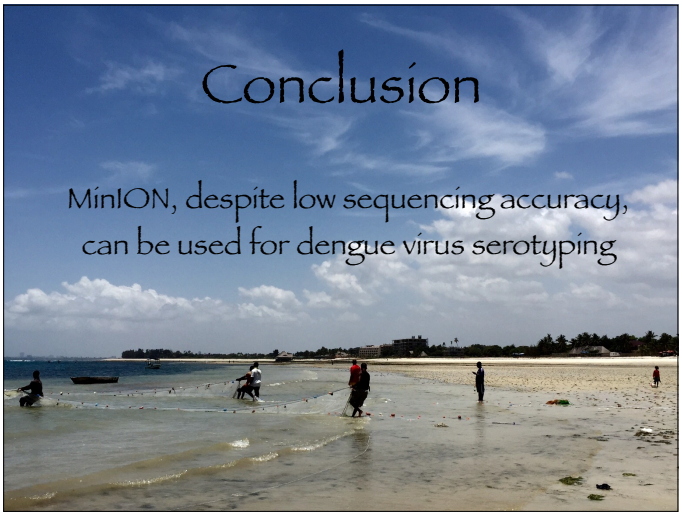
BLAST alignments against dengue genome



69

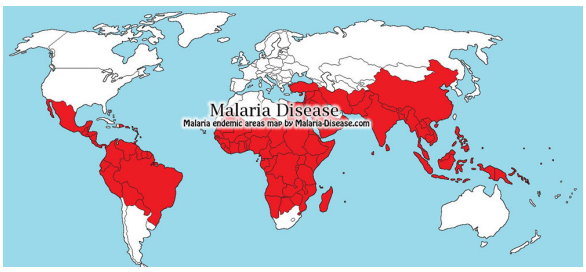
Conclusion

MinION, despite low sequencing accuracy, can be used for dengue virus serotyping



70

Malaria parasite



71

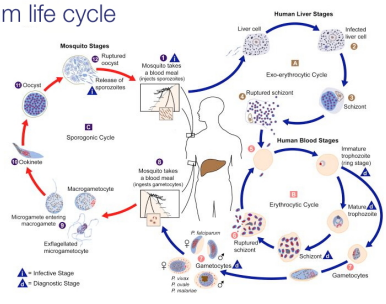
Malaria parasite

The Plasmodium life cycle

The sexual stage of the cycle is completed in the mosquito, the asexual one in human red blood cells.

Image from the Laboratory Identification of Parasites of Public Health Concern (DPDx) of the USA government of Centers for Disease Control and Prevention (CDC)

(http://www.dpd.cdc.gov/dpdx/HTML/ImageLibrary/Malaria_il.htm)



72

Spreading drug resistance

Chloroquine resistance

Sulfadoxine resistance

Dondorp et al. Nat. Microbiol (2010)

73

Drug Resistant Mutations

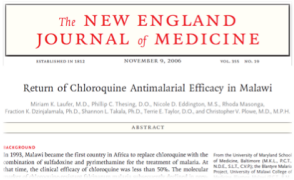
Drug	Gene	Number of known mutations
Chloroquine	CRT	1
Chloroquine/mefloquine	MDR	2
Artemisin	K13	1
Sulphadoxine-pyrimethamine	DHFR	4
Sulphadoxine-pyrimethamine	DHPS	6

Nair at al. (2014) Genome Res. 24(6):1028-38.

74

Declining of resistance parasite population

A study in Malawi, reported that population of CQ resistant P. falciparum (CQR) has decreased.

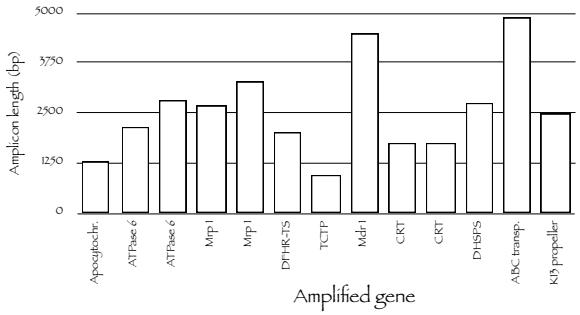

The NEW ENGLAND JOURNAL of MEDICINE

Return of Chloroquine Antimalarial Efficacy in Malawi

Declining of 76T mutation also reported in some other countries (i.e. Benin, Kenya, and Senegal) though at a slower rate.

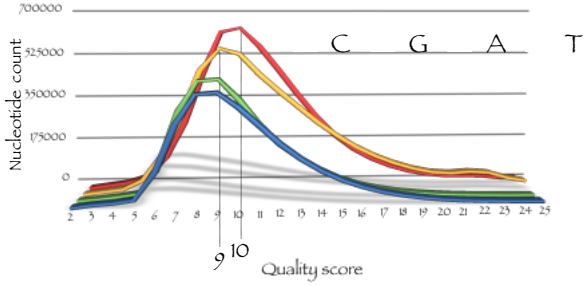
75

Targets for malaria genotyping



76

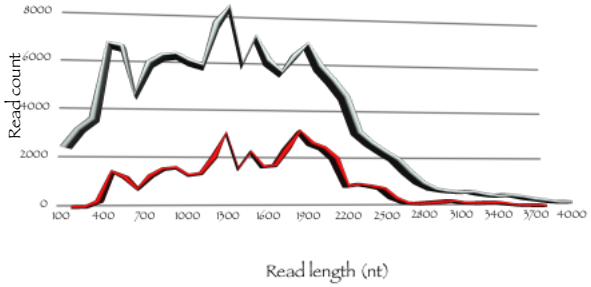
Distribution of quality scores for each nucleotide



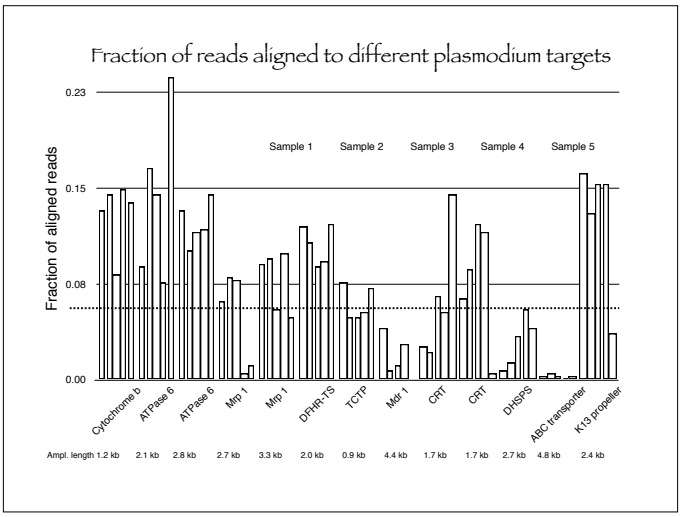
77

Fail Pass

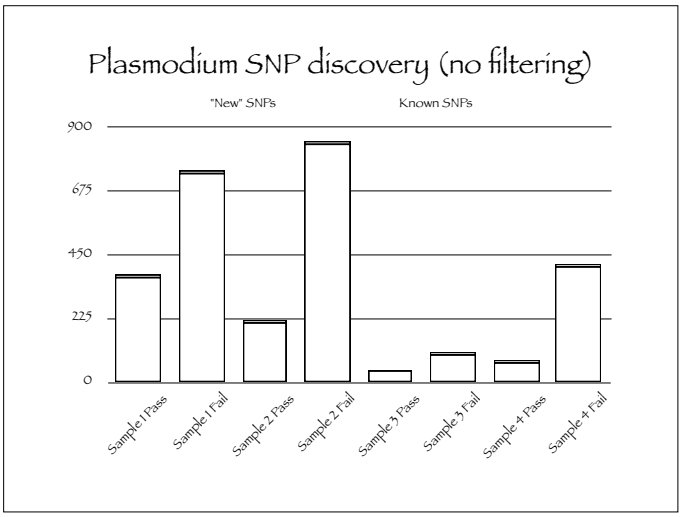
Read length distribution



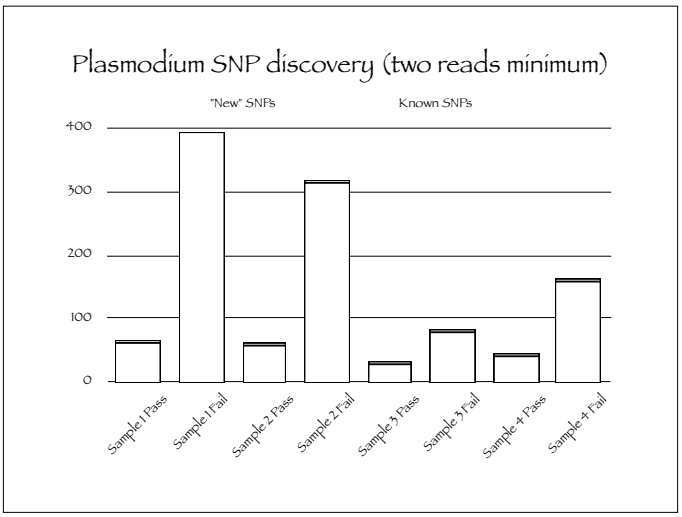
78



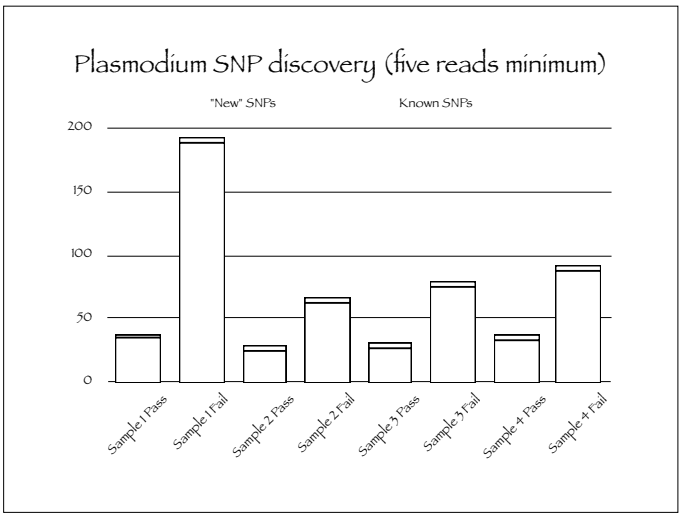
79



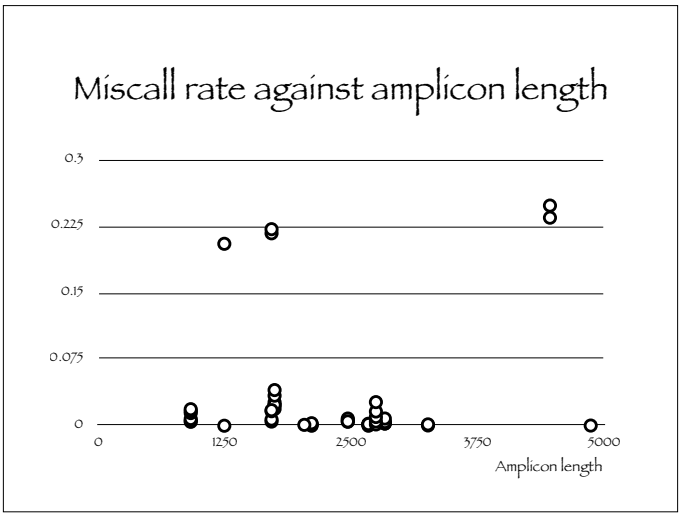
80



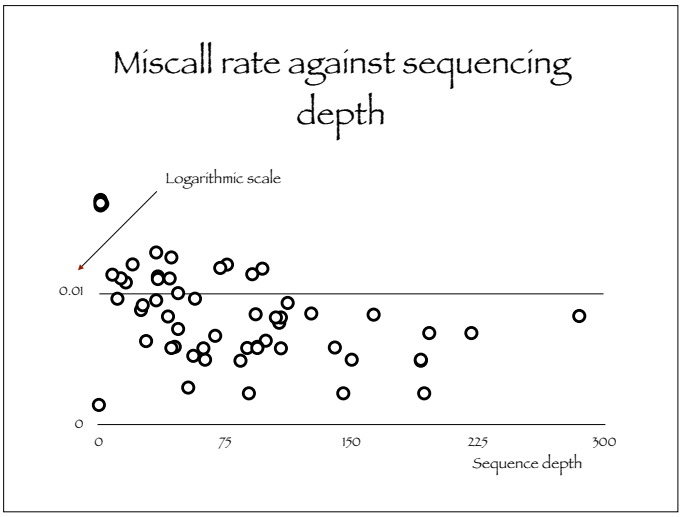
81



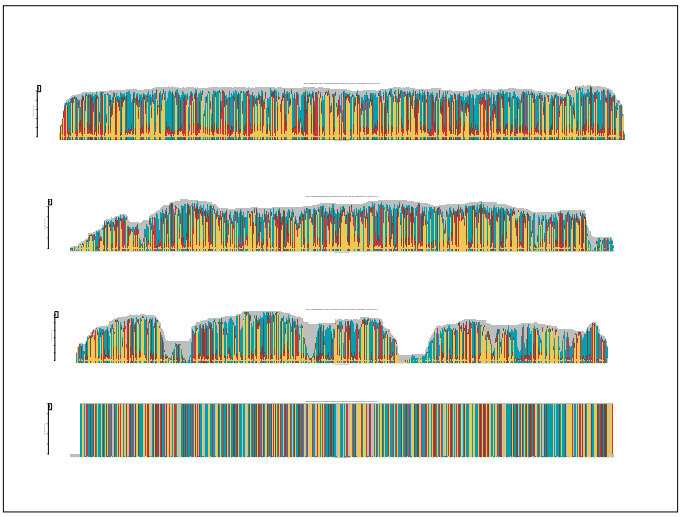
82



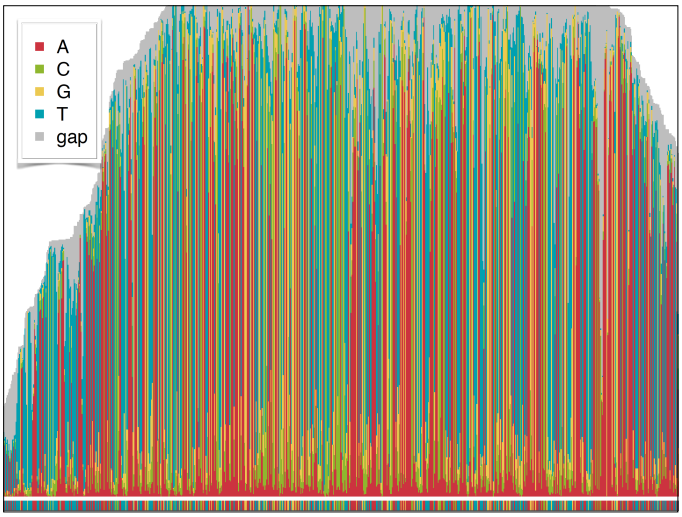
83



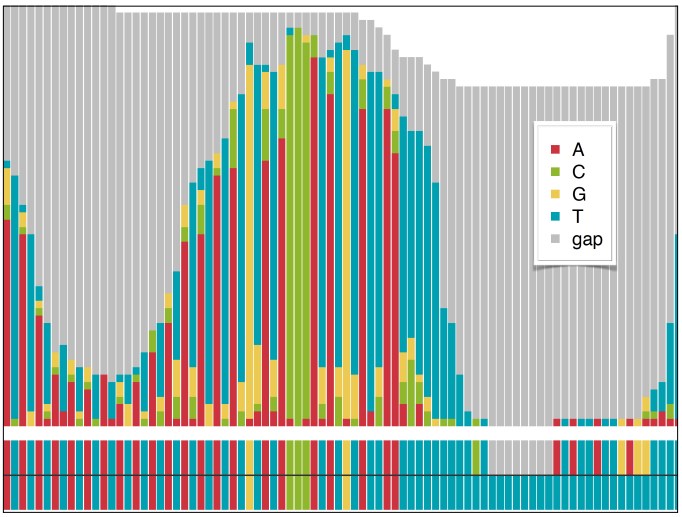
84



85



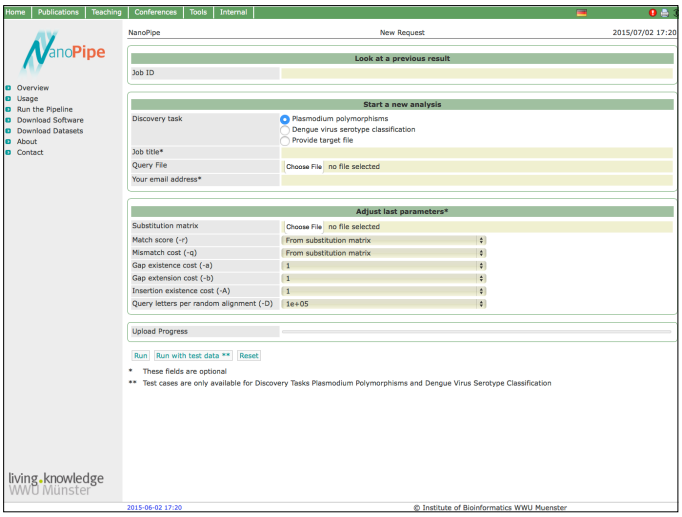
86



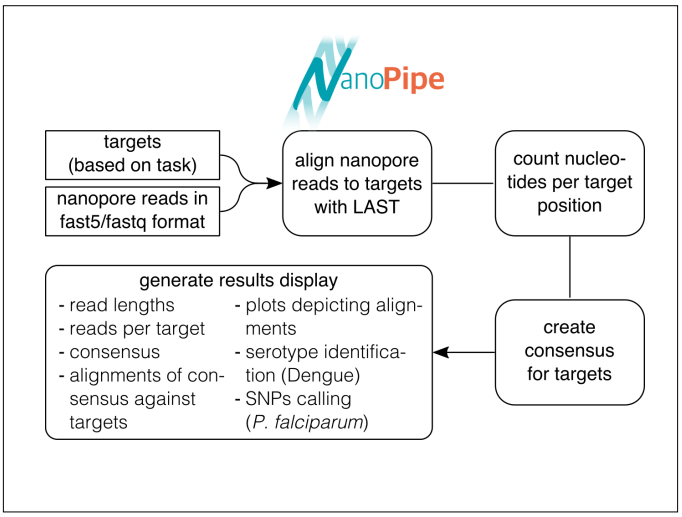
87

88

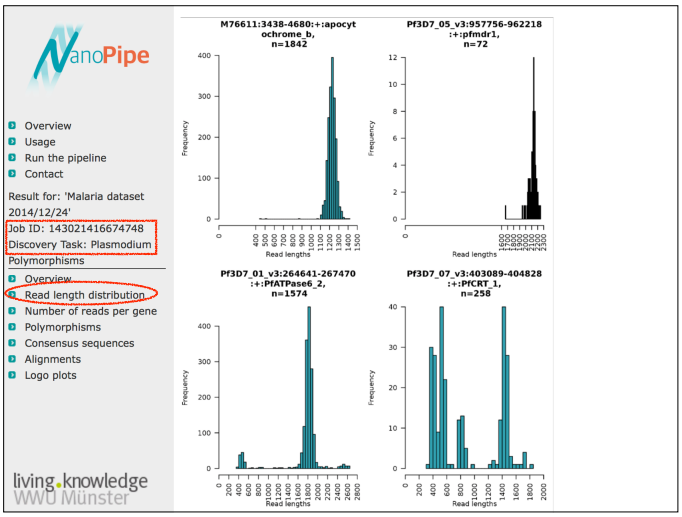
89



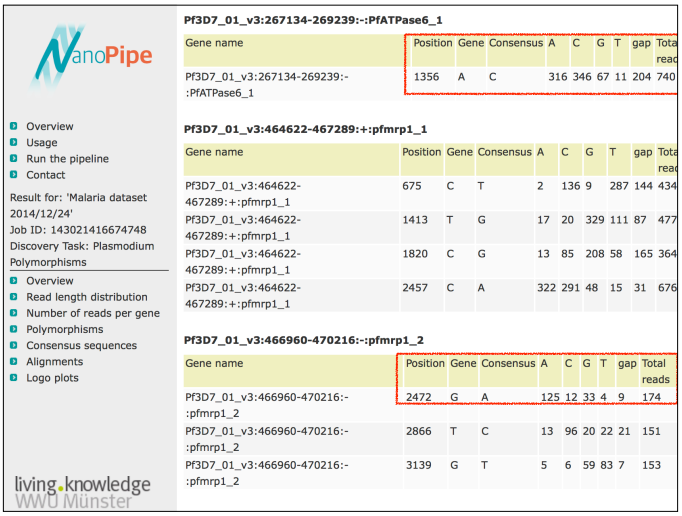
90



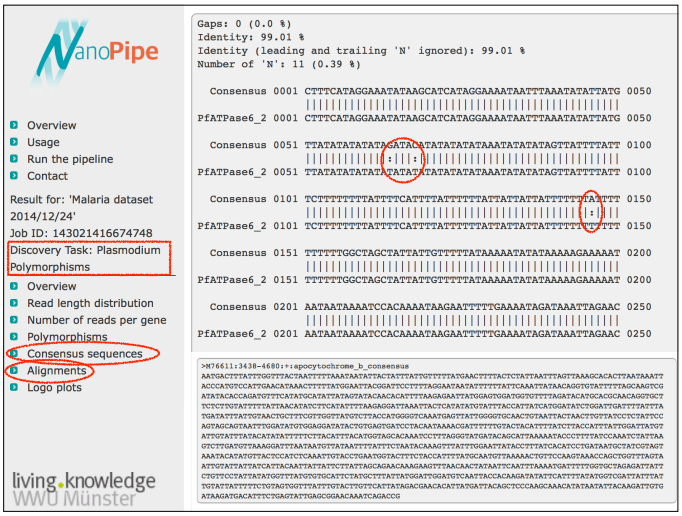
91



92



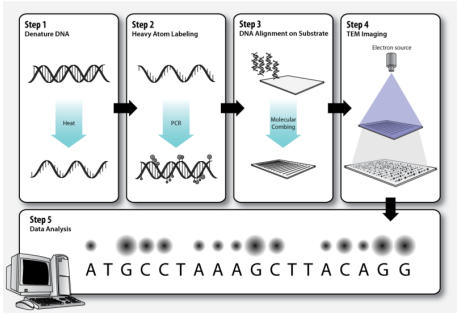
93



94

NGS - what is coming

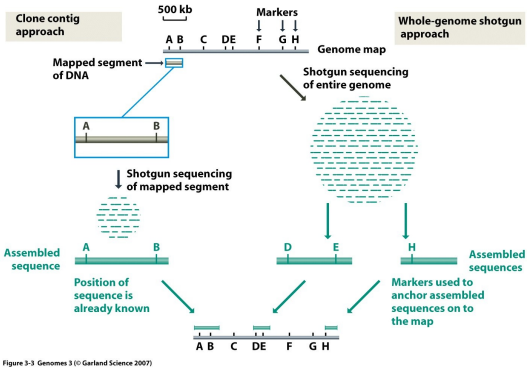
transmission electron microscopy



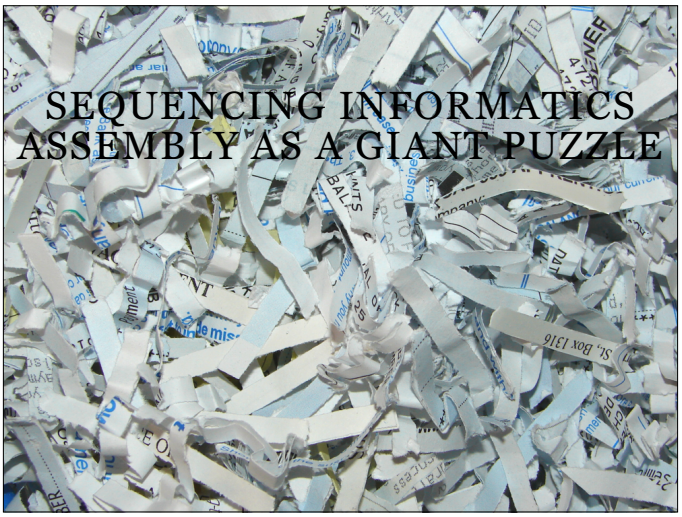
More at http://www.allseq.com/knowledgebank/

95

sequencing informatics



96

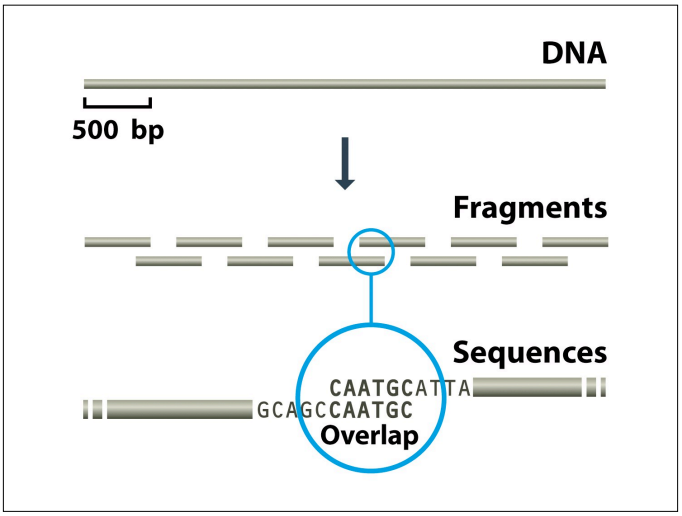


97

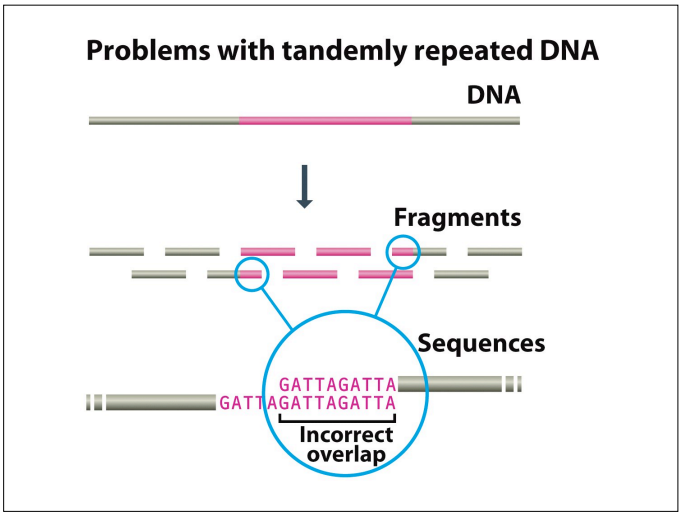
sequence assembly

- A fundamental goal of DNA sequencing has been to generate large, continuous regions of DNA sequence
- Capillary sequencing reads ~600-800 bp in length
  - Overlap based assembly algorithms (phrap, phusion, arachne)
  - Compute all overlaps of reads and then resolve the overlaps to generate the assembly
- In principle, assembling a sequence is just a matter of finding overlaps and combining them.
- In practice:
  - most genomes contain multiple copies of many sequences,
  - there are random mutations (either naturally occurring cell-to-cell variation or generated by PCR or cloning),

98



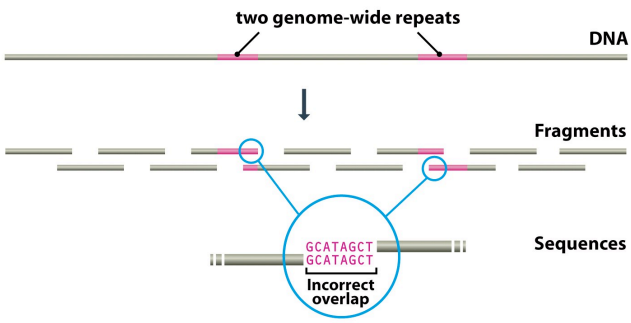
99



100

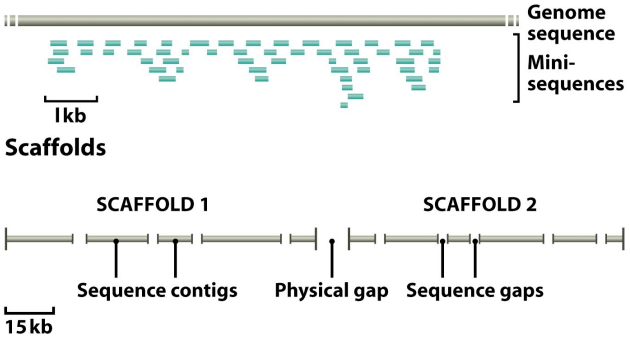
assembly problems

Problems with genome-wide repeats



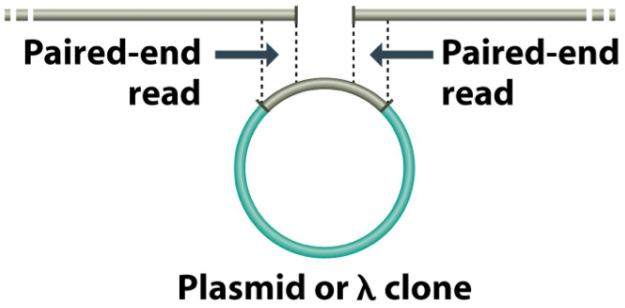
101

assembly problems
sequencing gaps



102

Assembly problems closing gaps



sequence assembly
NGS case

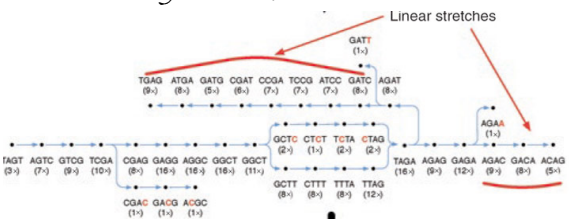
- Volume and read length of data from next-gen sequencing machines meant that the read-centric overlap approaches were not feasible
  - already in 1980's Pevzner et al. introduced an alternative assembly framework based on de Bruijn graph
  - Based on a idea of a graph with fixed-length subsequences (k-mers)
  - Key is that not storing read sequences – just k-mer abundance information in a graph structure

De Bruijn graph construction

Genome is sampled with random sequencing of for example 7 bp reads. Note errors in the reads are represented in red

Flicek & Birney (2009) Nat Meth, 6: S6-S12.

De Bruijn graph construction

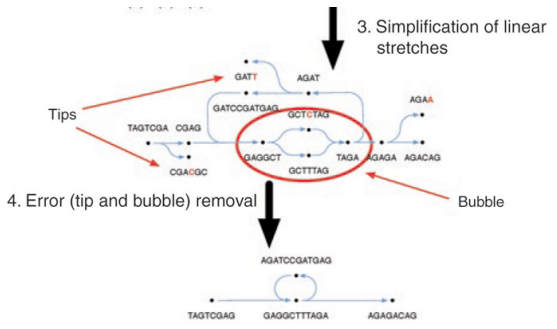


The k-mers in the reads (4-mers in this example) are collected into nodes and the coverage at each node is recorded (numbers at nodes).

Features:

- continuous linear stretches within the graph
- Sequencing errors are low frequency tips in the graph

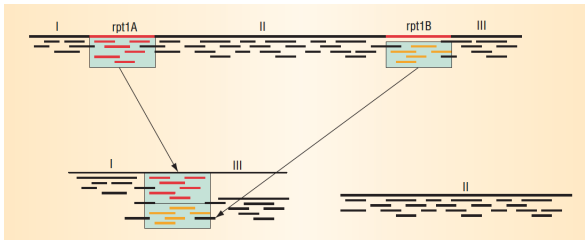
Flicek & Birney (2009) Nat Meth, 6: S6-S12.



Graph is simplified to combine nodes that are associated with the continuous linear stretches into single, larger nodes of various k-mer sizes.
Error correction removes the tips and bubbles that result from sequencing errors.
Final graph structure that accurately and completely describes in the original genome sequence

Flicek & Birney (2009) Nat Meth, 6: S6-S12.

repeats problem

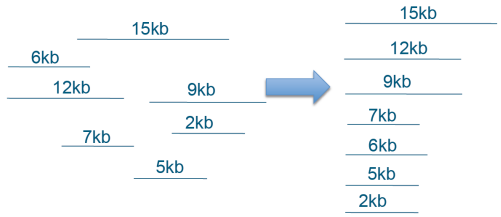


Very similar sequences may lead to false assembly, especially if the repeated region is longer than average reads length, e.g. recent tandem duplications or recent transpositions of mobile elements.

next-gen assemblers

- First de Bruijn based assembler was Newbler developed by 454 Life Sciences
  - Adapted to handle main source of error in 454 data – indels in homopolymer tracts
- Many de Bruijn assemblers subsequently developed
  - SHARCGS, VCAKE, VELVET, EULER-SR, EDENA, ABySS and ALLPATHS, SOAP
  - Most can use mate-pair information
- Slightly different approach to transcriptome assembly
  - It has to allow many discontinuous graphs representing single

assembly evaluation - N50

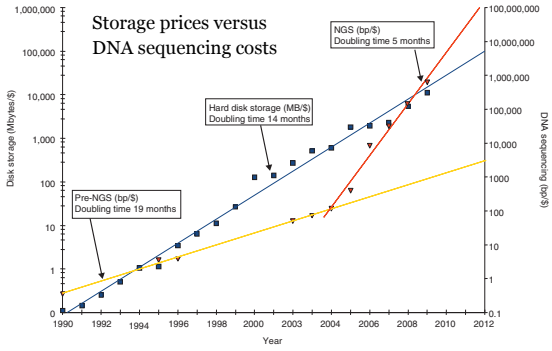


If one orders the set of contigs produced by the assembler by size, then N50 is the size of the contig such that 50% of the total bases are in contigs of equal or greater size.

15+12+9+7+6+5+2 = 56

56/2 = 28 -> N50 is 9kb (15+12 = 27 is less than 50%)

data storage problem
a consequence of sequencing technology success



BIOINFORMATICS CREED

- Remember about biology
- Do not trust the data
- Use comparative approach
- Use statistics
- Know the limits
- Remember about biology!!!