



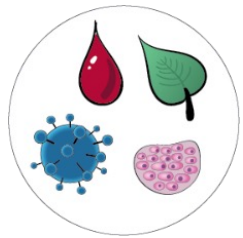
# Unlock the promise of genomics through PacBio sequencing

Single Molecule Real-time Sequencing Analysis Overview

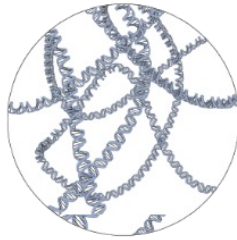
27 June 2023

彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

# From Sample to SMRT Sequencing

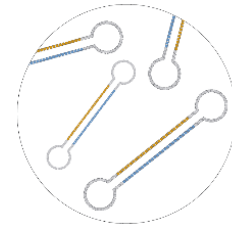


From viruses to vertebrates



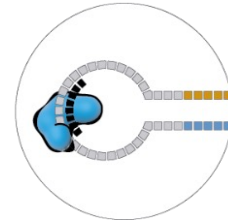
Isolate DNA or RNA

Ligate  
adapters  
+



Generate SMRTbell libraries

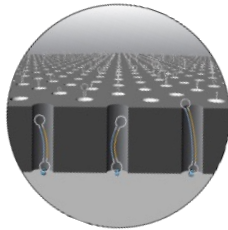
+  
Primer &  
Polymerase



Prepare sequencing reaction

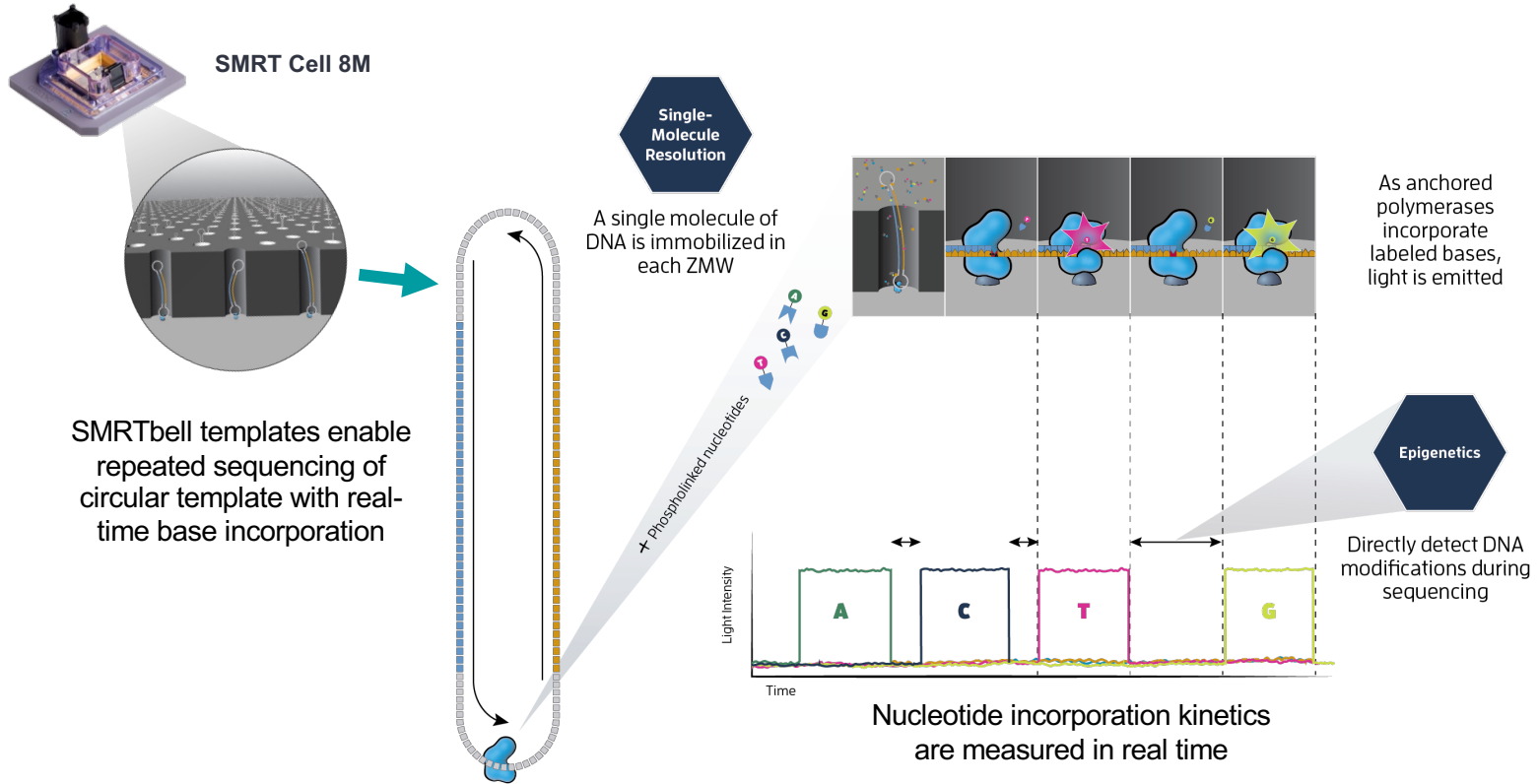


Use PacBio Sequel Systems to  
sequence genomes, transcriptomes,  
and epigenomes

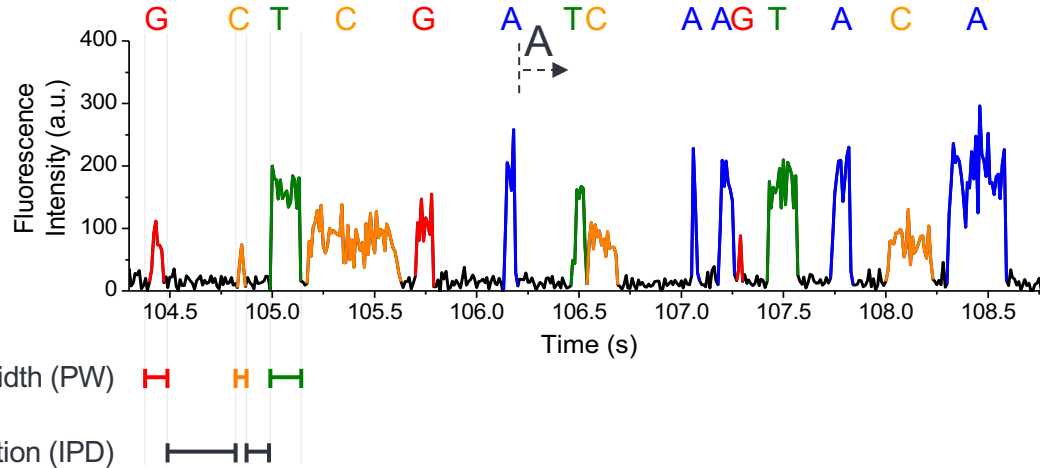
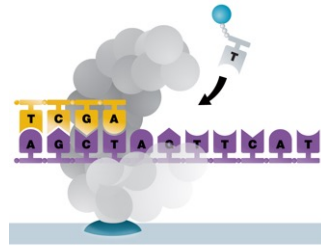


SMRT Cells contain millions  
of zero-mode waveguides  
(ZMWs)

# Single Molecule, Real-Time (SMRT) Sequencing



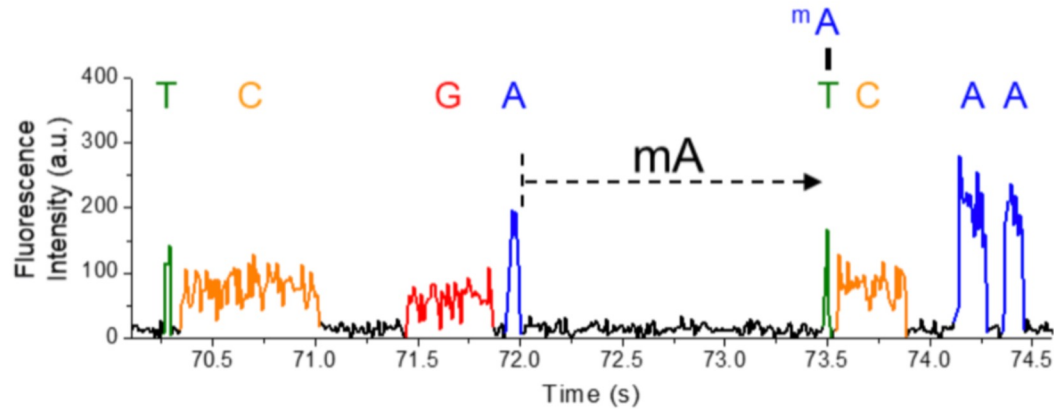
# PacBio sequencing detects methylation in native DNA



Fluorescent label indicates base identity

Kinetics impacted by base context and epigenetics

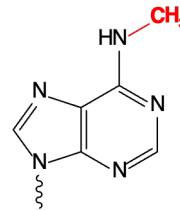
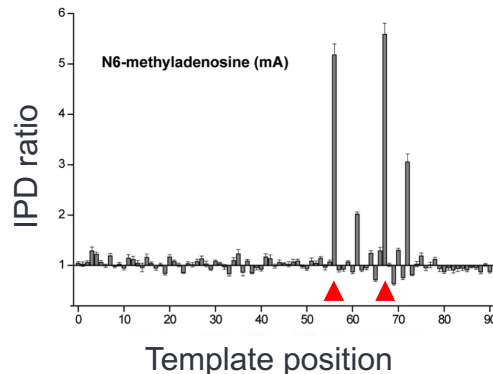
# PacBio sequencing detects methylation in native DNA



Fluorescent label indicates base identity

Kinetics impacted by base context and epigenetics

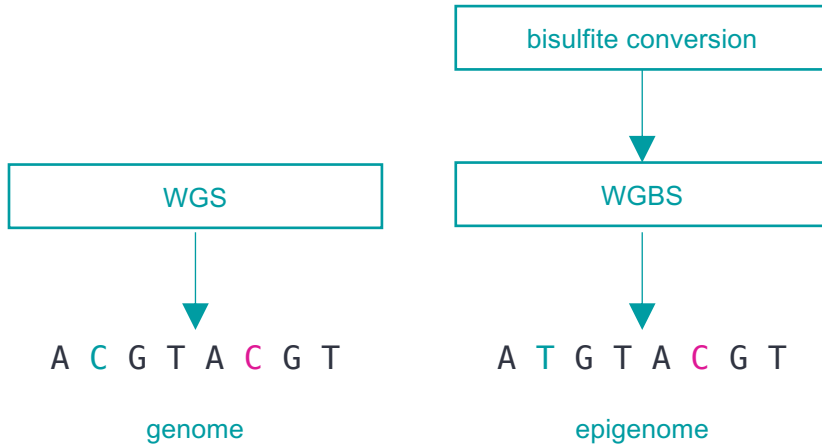
Interpulse duration (IPD)



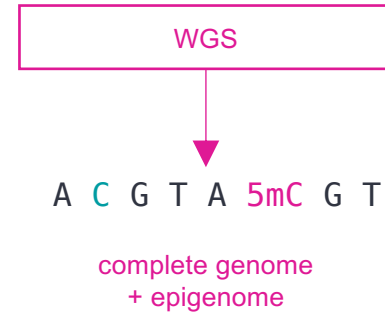
6mA  
strong, localized signal

# Measuring 5mC methylation with DNA sequencing

short read sequencing – 2 libraries



HiFi sequencing – 1 library

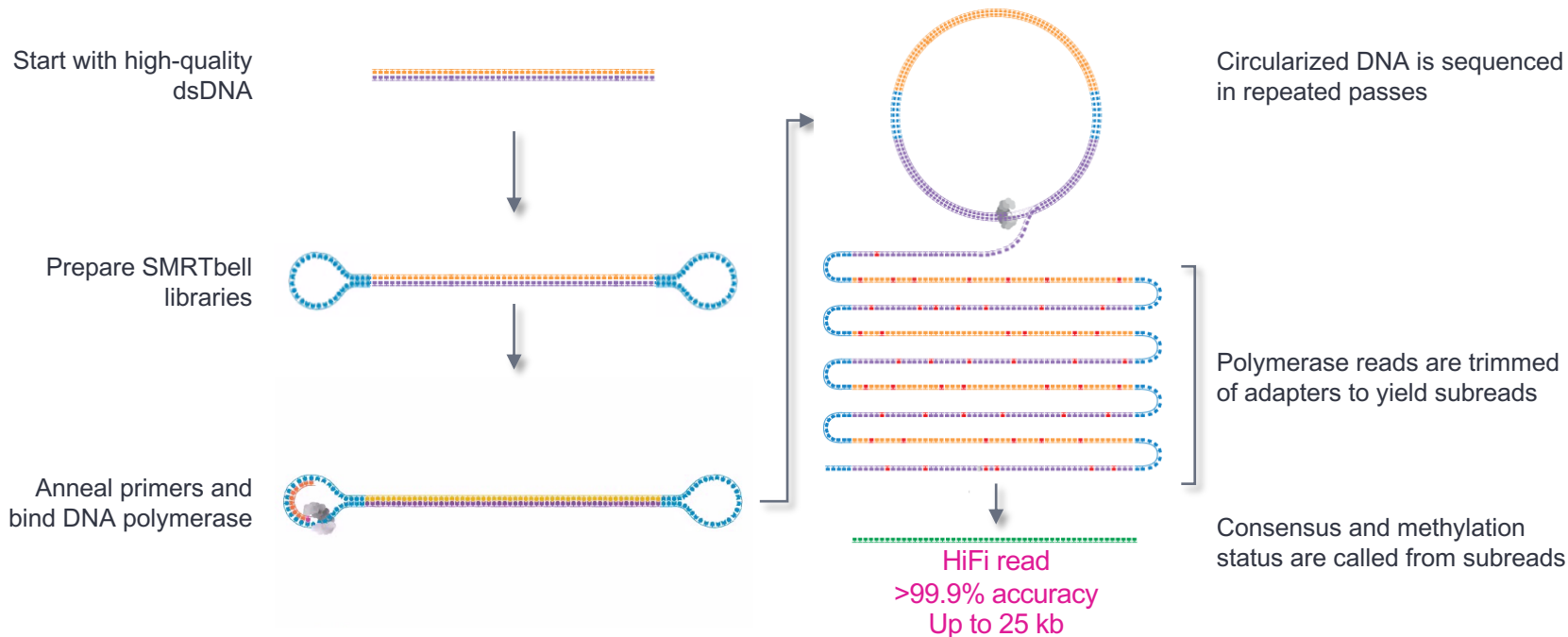


# What are HiFi reads?

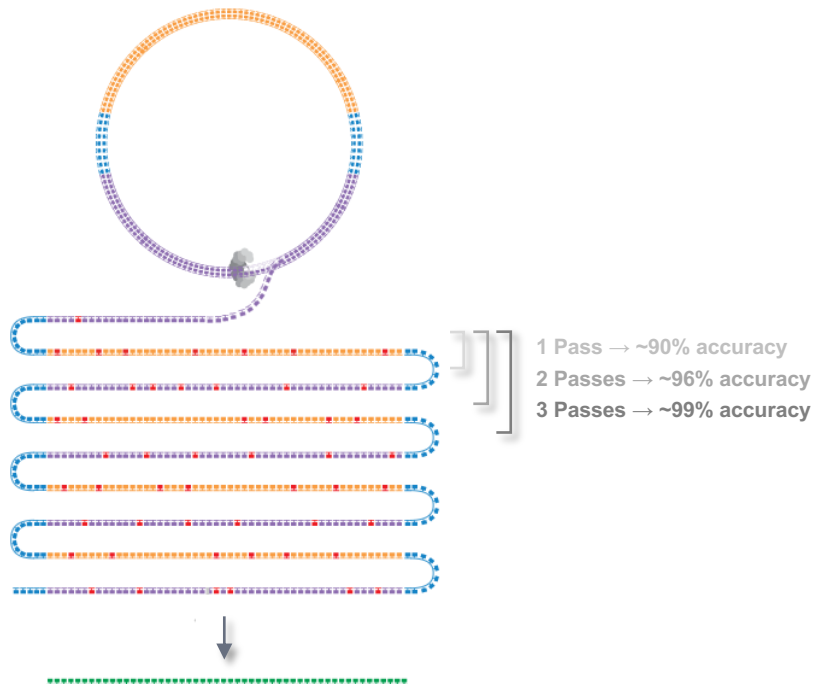
HiFi reads are produced using circular consensus sequencing (CCS) on PacBio long-read systems.

HiFi reads provide base-level resolution with 99.9% single-molecule read accuracy.

HiFi reads are unbiased, no DNA amplification, least GC content and sequence complexity bias

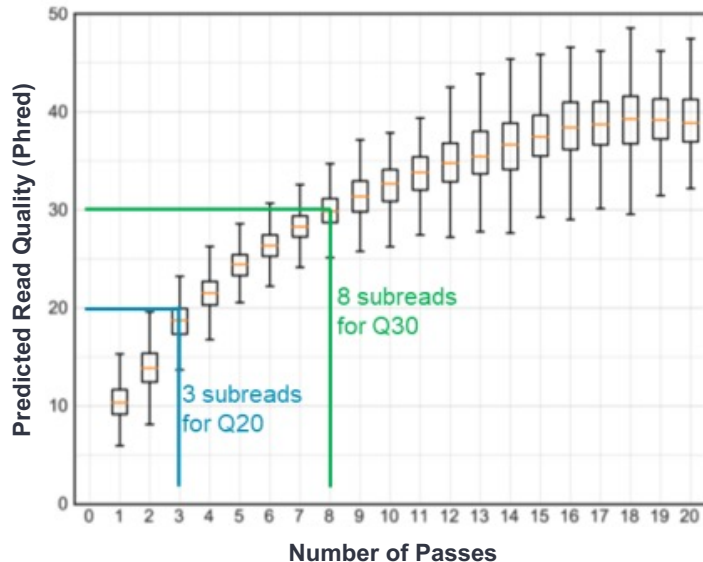


# PacBio HiFi reads combine long read lengths with high accuracy



HiFi read  
(99.9% accuracy)

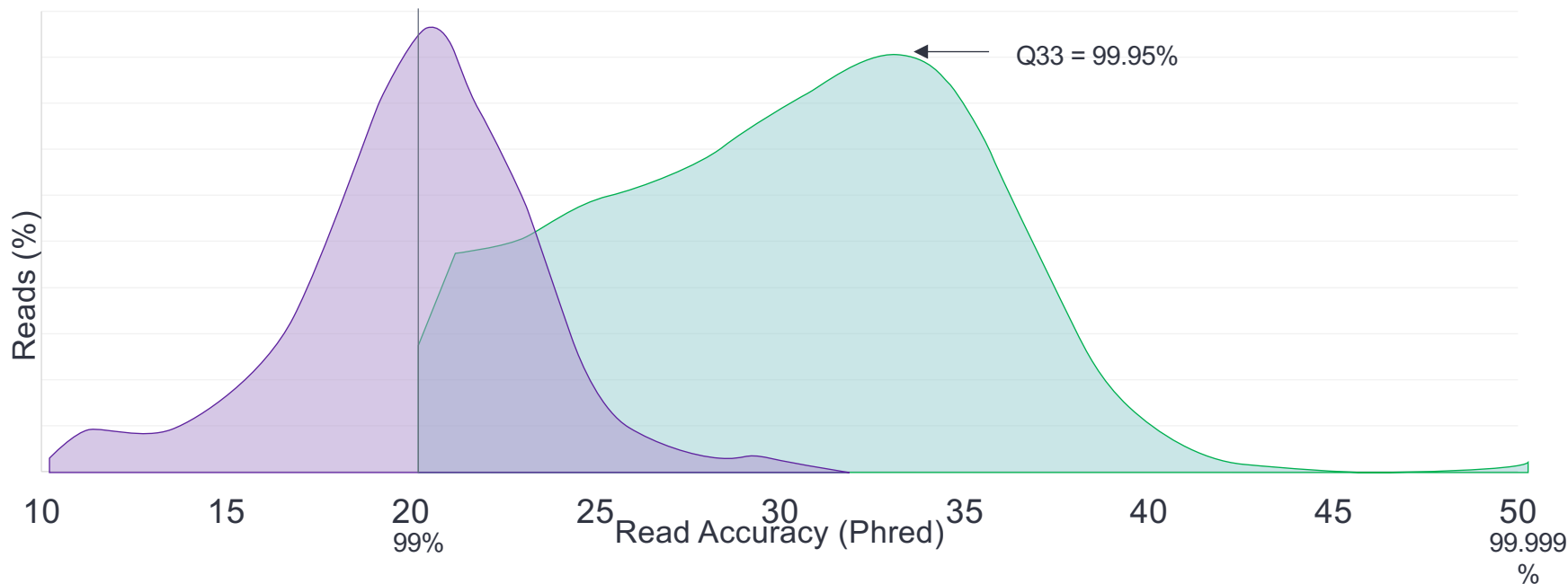
Cutoff Q20 – 99%  
Mean accuracy Q30 – 99.9%



Predicted read quality for CCS reads with different numbers of passes for a GIAB HG002 human sample. (CCS data generated using SMRT Link v8.0 CCS analysis application.)



## ONT's Q20+ chemistry vs HiFi

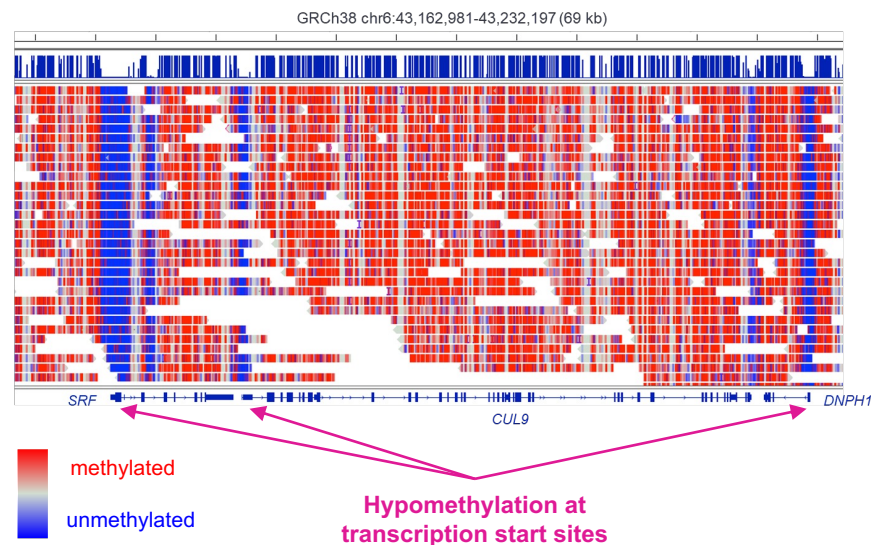
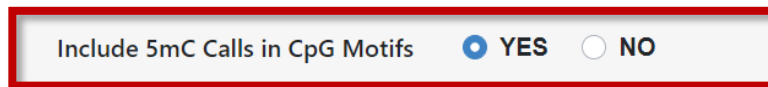


“Q20+” overlaid with typical HiFi results: No comparison on quality.

# Include 5mC calls in CpG motifs

If selected, kinetic signatures of cytosine bases in CpG motifs will be automatically analyzed to identify the presence of 5mC during on-instrument CCS (Sequel IIe system only) or during CCS analysis in SMRT Link

- **Default setting = YES** when specifying 'HiFi Reads' or 'Custom' application types
- 5mC detection is automatically performed on-instrument with the Sequel IIe system and in SMRT Link with the Sequel II system (data outputs are the **same** for both methods)
- 5mC calls are output in `hifi_reads.bam` as BAM standard MM and ML tags and can be easily visualized in [IGV](#)
- Processing and storage requirements are **minimal**:
  - File size increase is ~5%
  - On-instrument processing time for Sequel IIe systems is ~10 minutes
- Kinetics are not retained in the CCS analysis output by default, but they can **optionally** be retained as well.
- 5mC calls require a **CpG context and symmetrical methylation** (i.e., does not detect hemi-methylated sites)
- Though trained on human data, 5mC detection has been demonstrated to work on non-human data (e.g., plants (Maize)).
- 5mC consensus calling and other tools planned for a future SMRT Link version.
  - For guidance on command line tool options for 5mC analysis, please contact your local PacBio support team or [PacBio Technical Support](#)



Example IGV plot demonstrating 5mC detection in HiFi reads for a human HG002 sample. Hypomethylation at active transcription start sites can be easily visualized (unpublished data).

# Representation of 5mC CpG data uses BAM format standard

Standard library prep, no extra compute, negligible data footprint, and standardized representation

Sequel IIe system



hifi\_reads.bam

30 GB for  
SEQ + QUAL

+5% for  
methylation

Sequence Alignment/Map Optional Fields Specification

The SAM/BAM Format Specification Working Group

14 Jul 2021

2.1 Base modifications

MM:Z:C+m,5,12,0

ML:B:C,204,89,26

Supported in IGV 2.10

## Also available as off-instrument analysis

Output kinetics tags from Sequel IIe system and run primrose software or SMRT Link workflow



on-instrument

Run design

CCS Analysis Output – Include Kinetics Information  YES  NO

TGTTTAGACTCCGTAATTACTCGCCTAGGAATTCTCAAGGGCACAATCAG

Kinetic tags

fi:B:C,22,70,24,12,10,21,16,8,45,5,31,16,12,9,12,2

fp:B:C,18,45,21,10,22,33,9,9,62,13,9,57,23,6,29,15

ri:B:C,17,30,9,12,9,7,16,17,26,16,8,19,94,5,14,14

rp:B:C,28,26,10,24,20,25,18,11,16,18,4,9,39,34,17

primrose <bamIn> <bamOut>



Mm:Z:C+m,4,12,16,4,16,19,44,10,11,4

Ml:B:C,249,4,247,177,210,228,245,244,100,246

# IGV supports coloring reads by methylation annotation

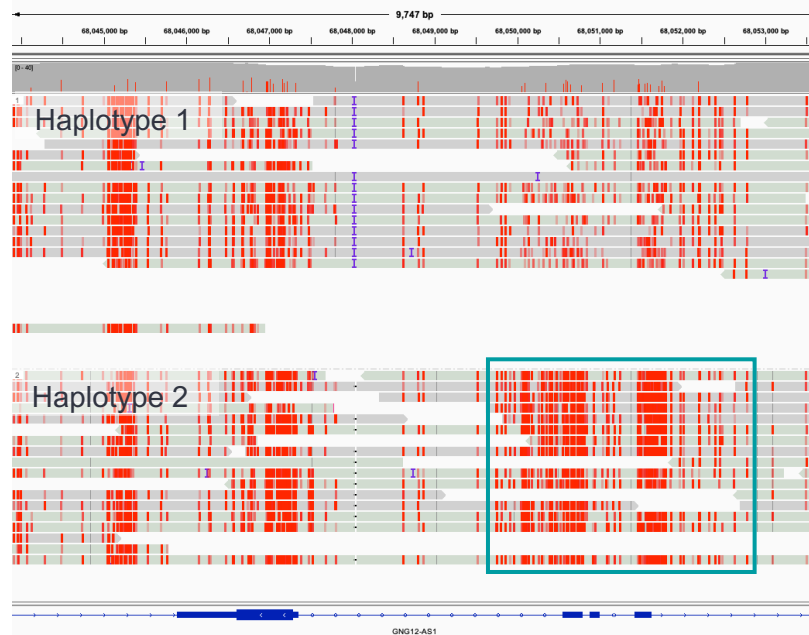


HG002.GRCh38.haplotagged.bam

- Rename Track...
- Copy read details to clipboard
- Change Track Color...
- Experiment Type
- Cluster (phase) alignments
- Linked read view (BX)
- Linked read view (MI)
- Link supplementary alignments
- Link by tag...
- Group alignments by
- Sort alignments by
- Color alignments by**
- Re-pack alignments
- ✓ Shade base by quality
- ✓ Show mismatched bases
- Show all bases
- ✓ Quick consensus mode
- View as pairs
- Set insert size options ...

- none
- read strand
- read group
- sample
- library
- movie
- ZMW
- base modification**
- tag
- bisulfite mode

Supported in IGV 2.10



Allele-specific methylation  
(imprinting)

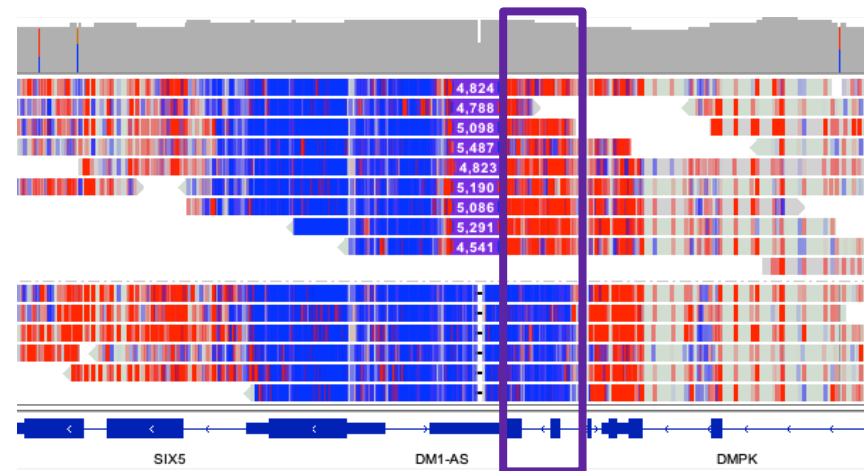
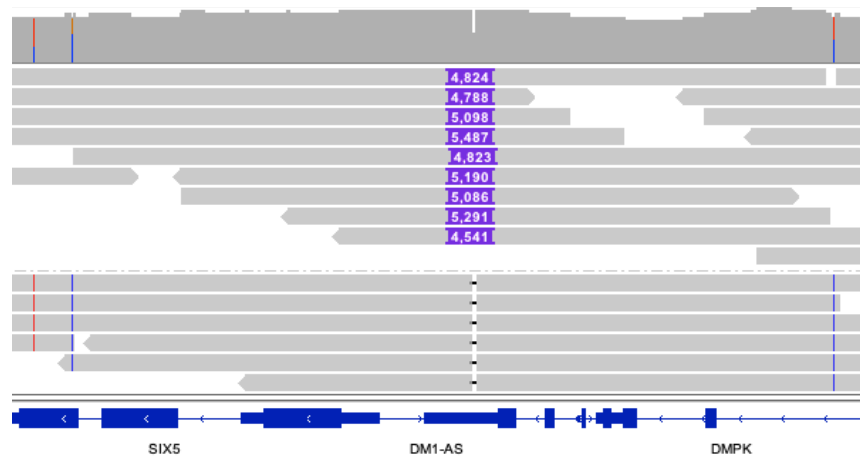
# Hypermethylation induced by pathogenic repeat expansion

## Myotonic dystrophy due to *DMPK* repeat expansion



Tomi Pastinen,  
Children's Mercy  
Kansas City

chr19:45,765,480-45,774,126 (8.6 kb)



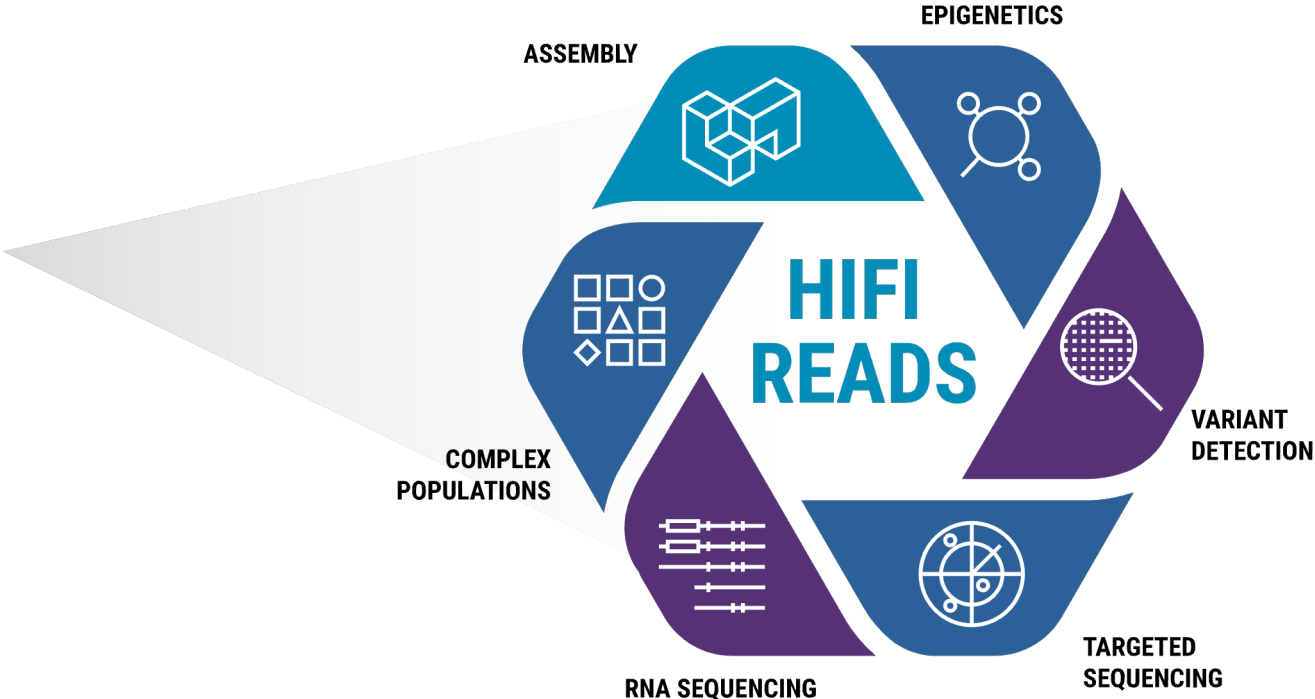
# HiFi 5-base sequencing: a complete genome & epigenome

- ✓ Genome-wide
- ✓ 1 library + 1 sequencing run
- ✓ Long reads = **phasing**
- ✓ Uniform coverage
- ✓ High mappability



A  
C  
G  
T  
+ 5mC

# Complete and accurate long-read sequencing

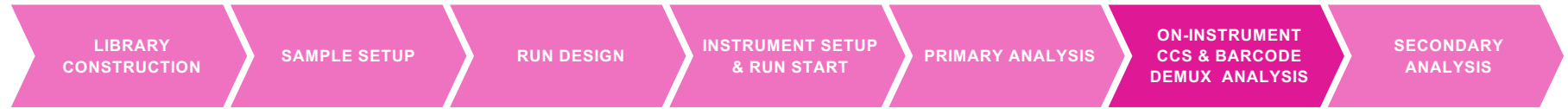






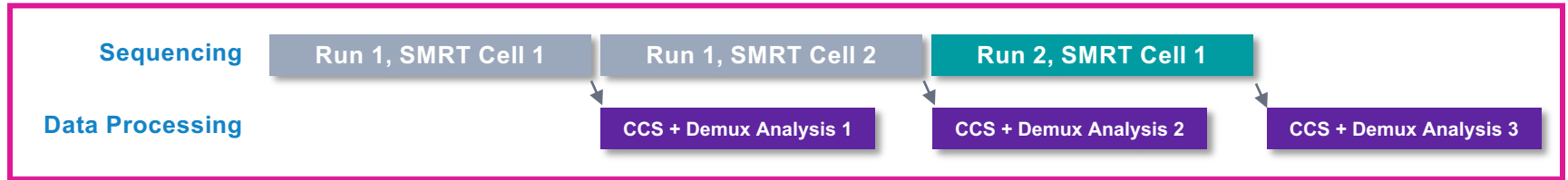
# Sequel IIe system output files & data structure

# Sequel Ie System on-instrument CCS and demultiplex analysis



Sequel Ie System computer hardware enables on-instrument circular consensus sequencing (CCS) read generation and barcode demultiplexing analysis during runs

- On-instrument CCS and barcode demultiplexing analysis occurs **in parallel** with sequencing data collection to minimize overall sequencing run times



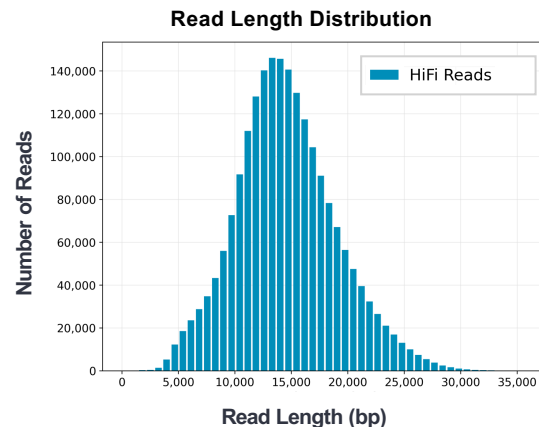
- A new sequencing run can be started while CCS data analysis from the preceding sequencing run is ongoing
- For a SMRT Cell movie collection yielding 30 Gb of HiFi ( $\geq Q20$  CCS) data, the typical on-instrument CCS analysis time is approximately 8 hours

# hifi\_reads.bam data file properties

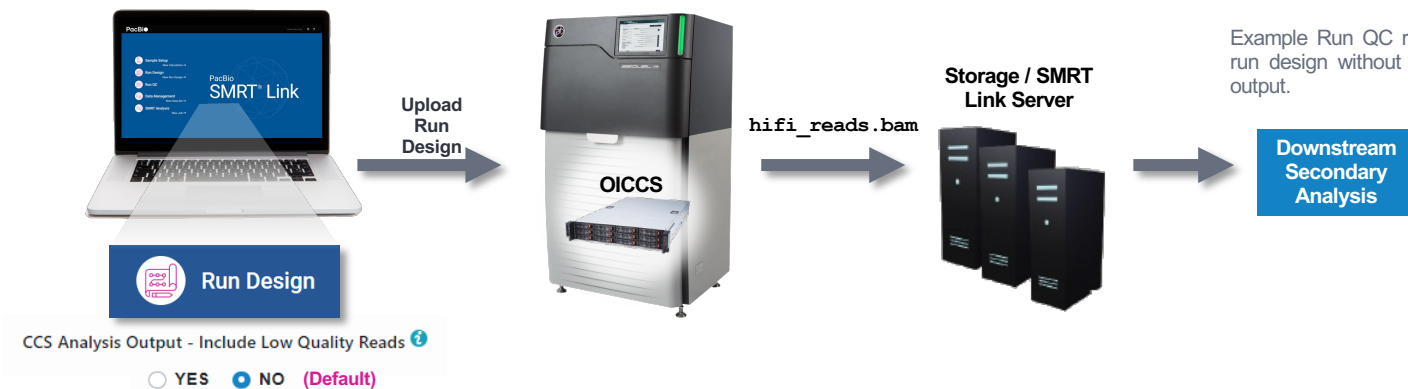
By default, the Sequel IIe System on-instrument CCS (OICCS)\* analysis workflow outputs a hifi\_reads.bam

## On-instrument CCS analysis workflow for standard (default) run designs

- A standard Run Design performs on-instrument CCS analysis **without** including low-quality reads and generates a hifi\_reads.bam output file and transfers it to the network server.
- hifi\_reads.bam contains **only** HiFi reads ( $\geq$ QV 20 CCS).
- After transferring to the storage server, the hifi\_reads.bam file can be used directly as input for downstream analysis using SMRT Analysis or other third-party analysis tools that expect  $\geq$ QV 20 data
- Refer to **Sequel II and IIe Systems: Data Files** ([102-144-100](#)) for further technical details regarding the contents of the hifi\_reads.bam file



Example Run QC read length distribution plot for a standard run design without including low-quality reads in the OICCS output.



# File and directory structure output by the Sequel Ii System

Example default file and directory structure output by the Sequel Ii system for each SMRT Cell transferred to network storage

```
<your_specified_output_directory>/r64012e_211206_183753/1_A01/  
|--m64012e_211206_183753.baz2bam_1.log  
|--m64012e_211206_183753.ccs.log  
|--m64012e_211206_183753.ccs_reports.json  
|--m64012e_211206_183753.ccs_reports.txt  
|--m64012e_211206_183753.consensusreadset.xml  
|--m64012e_211206_183753.hifi.reads.bam  
|--m64012e_211206_183753.hifi.reads.bam.pbi  
|--m64012e_211206_183753.sts.xml  
|--m64012e_211206_183753.zmw_metrics.json.gz  
|--m64012e_211206_183753.transferdone
```

**Minimum list of files** needed to analyze SMRT sequencing data in SMRT link

- .consensusreadset.xml file
- .hifi\_reads.bam file
- .hifi\_reads.bam.pbi file

- In this example, r64012e\_211206\_183753 is a directory containing the output files associated with one sequencing run
  - 64012e is the instrument ID number
  - 211206\_183753 is the run date, in YYYYMMDD format, and time, in UTC format
- The main run directory includes a subdirectory for each movie collection (SMRT Cell) associated with a sample well – in this case, 1\_A01. Each subdirectory contains data output files of interest.

## File and directory structure output by the Sequel Ite System (cont.)

If 5mC CpG Detection is performed, the following additional files are output:

```
|-- m64012e_211206_183753.5mc_report.json
|-- m64012e_211206_183753.primrose.log
```

If on-instrument barcode demultiplexing is performed, the following additional files are output:

```
|-- bc1001--bc1001/m64012e_211206_183753.bc1001--bc1001.consensusreadset.xml
|-- bc1001--bc1001/m64012e_211206_183753.hifi_reads.bc1001--bc1001.bam
|-- bc1001--bc1001/m64012e_211206_183753.hifi_reads.bc1001--bc1001.bam.pbi
|-- m64012e_211206_183753.barcodes.fasta
|-- m64012e_211206_183753.lima.log
|-- m64012e_211206_183753.lima_counts.txt
|-- m64012e_211206_183753.lima_guess.json
|-- m64012e_211206_183753.lima_guess.txt
|-- m64012e_211206_183753.lima_reports.txt
|-- m64012e_211206_183753.lima_summary.txt
|-- m64012e_211206_183753.unbarcoded.consensusreadset.xml
|-- m64012e_211206_183753.unbarcoded.hifi_reads.bam
|-- m64012e_211206_183753.unbarcoded.hifi_reads.bam.pbi
```

**Note:** The undemultiplexed `hifi_reads.bam` file is not transferred, it is partitioned into the file structure shown here

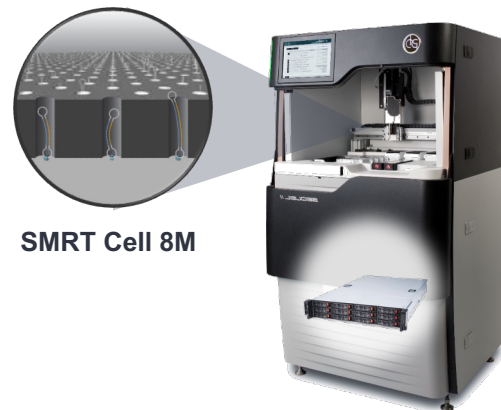


# SMRT Link software overview

# Sequel Ile System and Software v12

## Sequel Ile System - the only sequencer with highly accurate long reads

- off the box
  - Fast time to results, significantly less compute needs, greatly reduced storage
  - Lower overall solution cost resulting in more accessible system



SMRT Cell 8M

## SMRT Link – PacBio’s open source SMRT Analysis software suite.

- Support intuitive GUI or command-line interface

### Software Download

#### DOWNLOAD SMRT LINK V12.0 NEW

SMRT Link v12.0 supports Revio, Sequel II and Ile systems. v12.0 is **required** for Revio customers, and is an optional update for Sequel II and Ile system customers. Customers with Sequel systems should use SMRT Link v.10.2.

Please ensure you meet [minimum system requirements](#) before upgrading to v12.0. If you are operating SMRT Link without meeting minimum system requirements, please contact [PacBio Support](#) to assist with your upgrade.

**NOTE:** Customers who have not yet migrated from WSO2 to Keycloak for user management in SMRT Link, must migrate before or during the upgrade to SMRT Link v12.0.

[Download SMRT Link v12.0](#)

<https://www.pacb.com/support/software-downloads/>



SMRT Sequencing Data  
on a Network Server



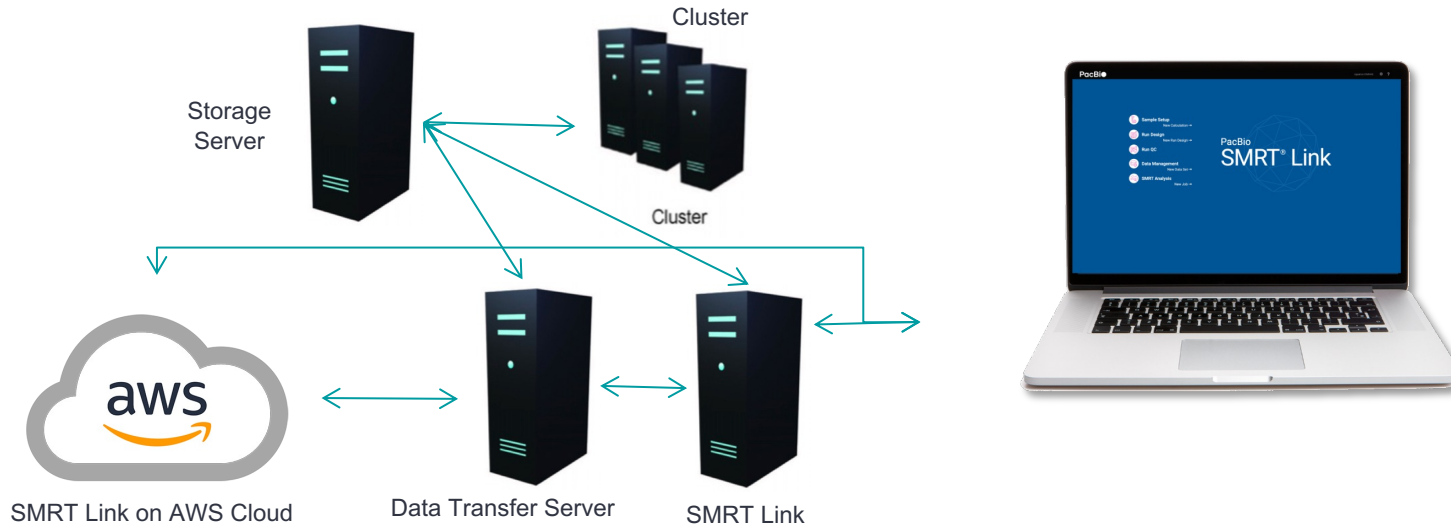
SMRT Link



# SMRT Link software overview



# SMRT Link system



Multiple features  
On Instrument



Sequel IIe System



Sequel II System

Sequel Instrument: Ready

Last Run Results | Run\_12\_15.2019 05:25

Col	Well	Sample	Status	Waste Time (sec)	Read Length (bp)	Items (kbp)
1	A01	WellSample_A01	Ready	120	---	---
2	B01	WellSample_B01	Ready	120	---	---
3	C01	WellSample_C01	Ready	120	---	---
4	D01	WellSample_D01	Ready	120	---	---
5	E01	WellSample_E01	Ready	120	---	---
6	F01	WellSample_F01	Ready	120	---	---
7	G01	WellSample_G01	Ready	120	---	---
8	H01	WellSample_H01	Ready	120	---	---

Alarms None Status Idle

LOCKED

# PacBio Software suite and analysis pipeline for SMRT data

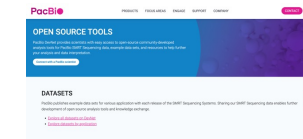
Denovo assembly	Improved Phased Assembly (IPA)
Variant Calling	DeepVariant + whatshap + pbsv
Structure variant	pbsv
Isoform detection	Iso-Seq
Single cell isoform	MAS-Seq
Metagenome	HiFi + Third party tools
16S Full-length	HiFi + Third party tools

- Fully automated analysis
- Efficient integration with LIMS and third-party analysis tools
- User-friendly UI design
- Industry-standard output formats: FASTA, FASTQ, SAM/BAM, VCF



[SMRT Link](#) with SMRT Analysis

[SMRT Link on AWS Cloud](#)



[Datasets](#) including example datasets



[SMRT Compatible Analysis Products](#) (partner solutions)



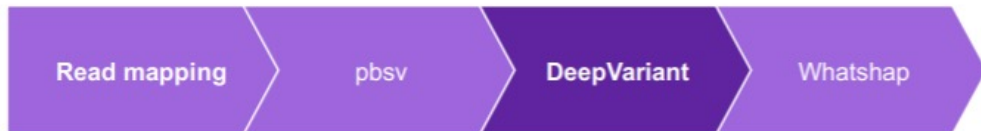
[pbbioconda](#)

(developmental tools)

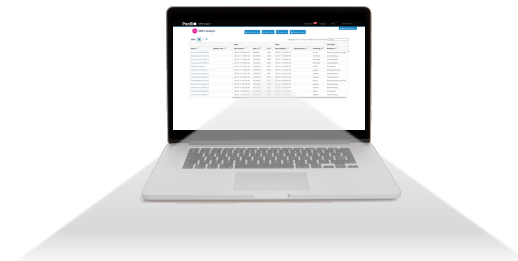
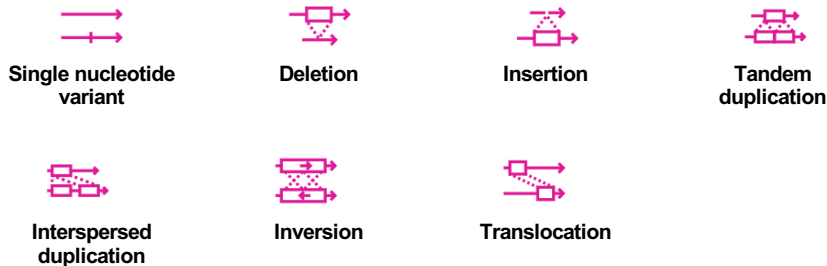
# SMRT Link variant calling analysis application overview

## New Variant Calling pipeline featuring DeepVariant

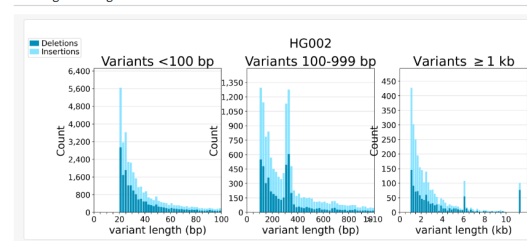
- New **Variant Calling** pipeline accepts HiFi reads (BAM format) as input and performs read mapping, structural variant-calling using pbsv, small variant calling using DeepVariant, and phasing using whatshap
- Use this application to identify **single-nucleotide variants, short insertions and deletions, and structural variants** for a WGS sample against a specific reference genome.
- Variants are **automatically phased and haplotagged**



## Types of variants



SV Length Histogram



Haplotagged Mapped BAM Index	20 MB	bam_bai
Haplotagged Mapped BAM	47 GB	bam



# Computer requirements

# Compute requirements Sequel Ile system

Head Node	
Cores	32
RAM	64 GB
Local Storage	1 TB SSD/Flash storage
Compute Nodes	
Cores (Total)	64
Minimum RAM per slot (1 slot = 1 core)	>4 GB
Local Storage	100 GB
64 x 4 = 256 GB RAM	
Shared Data Storage	
Sequencing Data	20 TB <sup>a</sup>
Analysis Data	40 TB <sup>a</sup>
Network	
10 GbE strongly recommended, 1GbE required <sup>b</sup>	

<sup>a</sup>Storage is calculated for one Sequel Ile System, assuming 100 human genomes per year at 30-fold coverage, *de novo* assembly

<sup>b</sup>Connection between the Head Node and Sequel Ile System

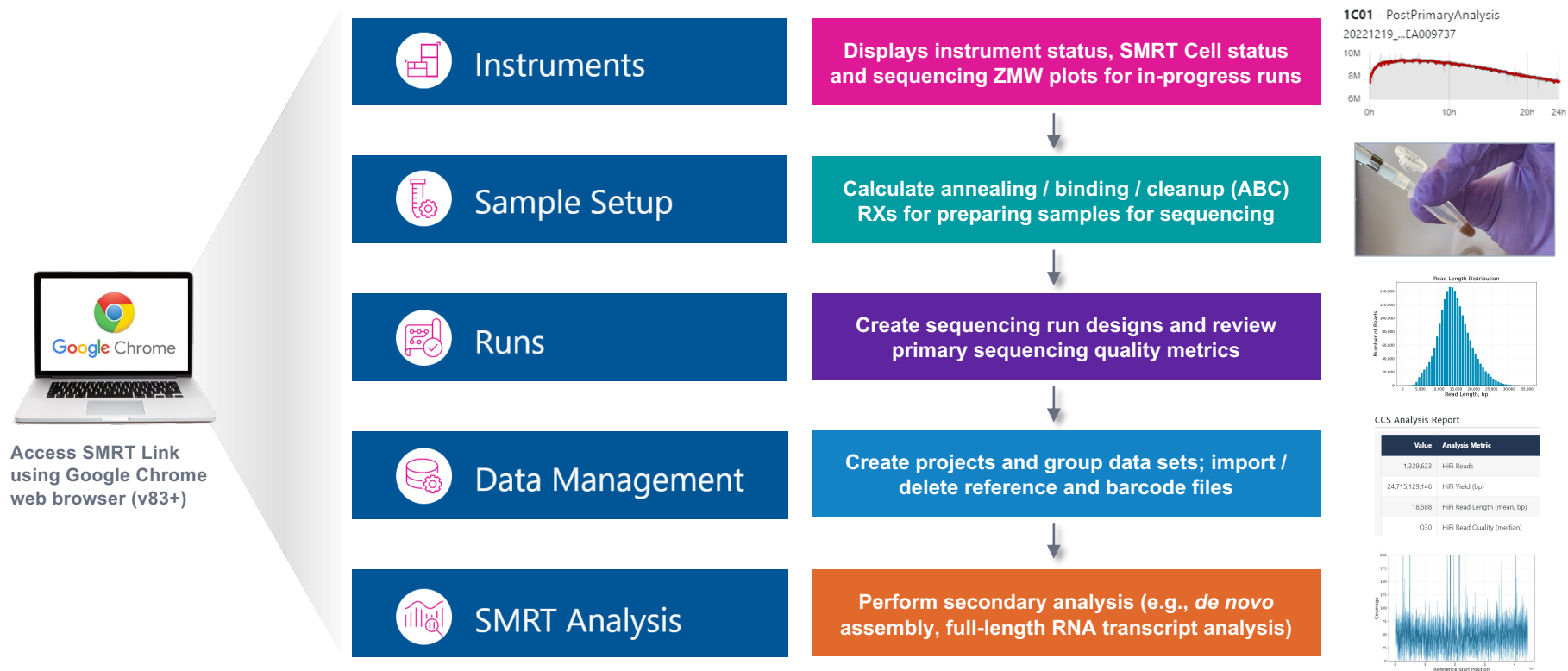
**Server OS:** CentOS 7.x and 8.x, and Ubuntu 18.04 and 20.04 64-bit Linux<sup>®</sup> distributions  
(This also applies to SMRT Link compute nodes.)



# SMRT Link GUI overview

# SMRT Link v12.0 core functions and organization

SMRT Link v12.0 enhances many core functions, features a new 'Instruments' module and combines Run Design & Run QC into a new 'Runs' module



# SMRT Link v12.0 Runs module

## Creating a new Run Design

+ Create New Run



To create a new Run Design, click on **Create New Run** on Runs module home page to go to **Create New Run Design** page

**PacBio** Runs

Alarms Request Error: Network Error Settings Help smark (Lab Tech)

Runs / Create New

### New Run Design

Cancel Delete Add Sample View Summary Save

#### Run Information

Instrument Type

Sequel II  Sequel IIe  Revio

Run Name

Run 12.23.2022 16:53

Plate 1 Required

Lot Serial Expiry

Plate 2

Lot Serial Expiry

Run Comments

Experiment Name

Experiment ID

#### Sample Information

Plate 1, Well A01: Copy Delete

Plate Well Required Plate 1, Well A01

Well Name Required

Well Comment

Library Type Required Standard

Polymerase Kit Required Revio polymerase kit

Application HiFi Reads

#### Samples

Adapters / Barcodes Required Revio SMRTbell adapters + barcodes

Sample Names Required Interactively From a File

> Run Options

> Data Options

> Analysis Options



# SMRT Analysis v12.0 for specifying workflow type

Data processing workflows are now separated into 'Analyses' and 'Utilities'

SMRT Analysis / Create New Analysis

1. Select Data 2. Select Analysis

Job Name Required  
SMRT Analysis Demo - Creating a New Analysis

Workflow Type  
 ANALYSIS  DATA UTILITY

Analysis of Multiple Data Sets  
One Analysis for All Data Sets

Choose an option when multiples Data Sets are selected.

only HiFi reads as input

Specify **One Analysis for All Data Sets** when performing variant calling analysis of multiple data sets

Datasets

Displaying rows 107 to 119 out of 2000 (scroll to load more)

	Data Set Details >			Sample Det...	Sample Det...	Run Data >	Metadata >				
<input type="checkbox"/>	Name <b>IT</b> ▾	Date Created ▾	Well Sample Name	Run Name <b>IT</b> ▾	Created By ▾	Bio Sample Name	Demultiplexed <b>S</b>	Barcode Name <b>↓</b>	Total Length <b>o</b>	Instrument Name	Version <b>IT</b> ▾
<input checked="" type="checkbox"/>	20221228_840...	2022-12-30,...	20221228_840...	20221228_8...	rdalal	Hg002-G7		bc2055--bc2055	88,725,681,2...	84023	3.0.1
<input checked="" type="checkbox"/>	20221228_840...	2022-12-30,...	20221228_840...	20221228_8...	rdalal	Hg002-G7_16...		bc2096--bc2096	740,226	84023	3.0.1

- **Analysis:** An analysis uses applications designed to produce biologically-meaningful results. These analysis applications **only** accept HiFi reads
- **Data Utilities:** Data processing utilities are used as intermediate steps to producing biologically-meaningful results. Some data utilities accept **only** HiFi reads whereas other data utilities accept **only** subreads (formerly known as “Continuous Long Reads” in previous SMRT Link versions)

# Analysis applications produce biologically-meaningful results

Analysis applications accept **only** HiFi reads as input

- **Genome Assembly**

- Generate de novo assemblies of genomes.

- **HiFi Mapping**

- Align (or map) reads to a user-provided reference sequence..

- **Iso-Seq Analysis**

- Characterize full-length transcript isoforms.

- **Microbial Genome Analysis**

- This combines and replaces the Microbial Assembly and Base Modification Analysis applications in the previous release.
- Generate de novo assemblies of small prokaryotic genomes, optionally identify 6mA and 4mC with associated nucleotide motifs.

- **Variants Analysis**

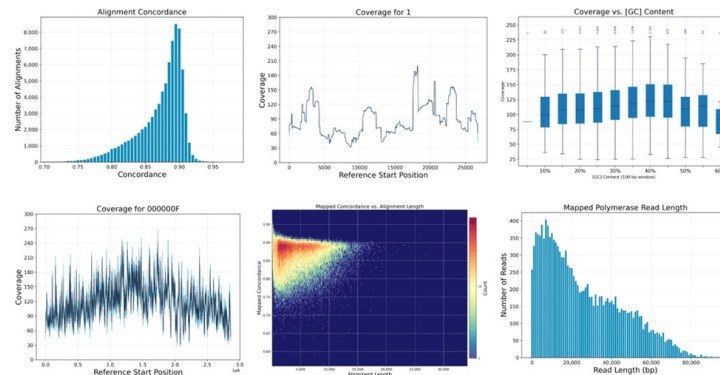
- Identify single-nucleotide variants, short insertions and deletions, and structural variants for a single sample against a specific reference genome.

- **Structural Variant Calling**

- Identify structural variants (Default:  $\geq 20$  bp) in a sample or set of samples relative to a reference. (support trio analysis)

- **Single-cell Iso-Seq**

- Enables analysis and functional characterization of full-length transcript isoforms with additional single cell information



# Data Utilities

PacBio Data Utilities are used as intermediate steps to producing biologically-meaningful results

The following data utilities accept only HiFi reads as input:

- **5mC CpG Detection**
  - Analyze the kinetic signatures of cytosine bases in CpG motifs to identify the presence of 5mC. (Sequel II only.)
- **Demultiplex Barcodes**
  - Separate reads by barcode.
- **Export Reads**
  - Export HiFi reads that pass filtering criteria as FASTA, FASTQ and BAM files.
  - For barcoded runs, you must first run the Demultiplex Barcodes application to create BAM files before using this application.
- **Mark PCR Duplicates**
  - Remove duplicate reads from a HiFi reads Data Set created using an ultra-low DNA sequencing protocol.
- **Trim Ultra-Low Adapters**
  - Trim PCR Adapters from a HiFi reads Data Set created using an ultra- low DNA sequencing library.

The following data utilities accept only subreads as input:

- **Circular Consensus Sequencing (CCS)**
  - Identify consensus sequences for single molecules.



# SMRT Analysis is HiFi centric

## Analysis Applications

PacBio SMRT Analysis

SMRT Analysis / Create New Analysis

1. Select Data 2. Select Analysis

**Analysis Application** Required

- ✓ ---
- Genome Assembly
- HiFi Mapping
- HiFIViral SARS-CoV-2 Analysis
- Iso-Seq Analysis
- Microbial Genome Analysis
- Read Segmentation and Single-Cell Iso-Seq
- Single-Cell Iso-Seq
- Structural Variant Calling
- Variant Calling

## Data Utilities

PacBio SMRT Analysis

SMRT Analysis / Create New Analysis

1. Select Data 2. Select Analysis

**Data Utility** Required

- ✓ ---
- 5mC CpG Detection
- Demultiplex Barcodes
- Export Reads
- Mark PCR Duplicates
- Read Segmentation
- Trim Ultra-Low Adapters
- Undo Demultiplexing



# Data Management report

## DataSet Overview

**PacBio** Data Management cguarco (Lab Tech) ?

Data Management / Dataset Details

### Rhino\_Verif\_HG002\_3ug5\_64441e\_C01\_3280358\_418 064\_2\_30hr\_85pM-Cell3 (CCS)

Copy Analyze... Export Delete

- Dataset Overview
- Status**
- Thumbnails
- Display All
- Adapter Report
- Control Report
- CCS Analysis Report
- Raw Data Report
- 5mC CpG Report
- Loading Report
- Analyses
- Data

#### Status

<b>Data Set</b>	Rhino_Verif_HG002_3ug5_64441e_C01_3280358_418064_2_30hr_85pM-Cell3 (CCS)
<b>Data Set ID</b>	189486
<b>Data Set UUID</b>	50d3d6cc-0a9c-4304-9939-044837af674d
<b>Well Sample Name</b>	Rhino_Verif_HG002_3ug5_64441e_C01_3280358_418064_2_30hr_85pM
<b>Biological Sample Name</b>	HG002
<b>Description</b>	ccs dataset converted
<b>Number of Records</b>	2,707,732
<b>Total Length</b>	39,236,168,651
<b>Status</b>	SUCCESSFUL
<b>Date Created</b>	2022-03-16, 10:48:33 AM
<b>Date Imported</b>	2022-03-16, 11:08:59 AM
<b>Date Updated</b>	2022-03-16, 11:13:20 AM
<b>Job ID</b>	51795
<b>Data Path</b>	/pb/collections/337/328/3280358/r64441e_20220311_211521/3_C01/m64441e_220314_191600.conser
<b>Run name</b>	RhinoVerif_64441
<b>Cell Index</b>	2
<b>Cell Id</b>	DA169999
<b>Instrument Name</b>	Sequel II Instrument
<b>Well Name</b>	C01
<b>Metadata Context Id</b>	m64441e_220314_191600

Analysis Parameters

# Data Management report

## CCS Analysis Report – Summary Metrics

Data Management / Dataset Details

Rhino\_Verif\_HG002\_3ug5\_64441e\_C01\_3280358\_418  
064\_2\_30hr\_85pM-Cell3 (CCS)

[Copy](#) [Analyze...](#) [Export](#) [Delete](#)

- Dataset Overview
- Adapter Report
- Control Report
- CCS Analysis Report
  - Summary Metrics
  - HiFi Read Length Summary
  - HiFi Read Quality Summary
  - Read Length Distribution

### CCS Analysis Report

Value	Analysis Metric
2,707,732	HiFi Reads
39,236,168,651	HiFi Yield (bp)
14,490	HiFi Read Length (mean, bp)
Q34	HiFi Read Quality (median)
12	HiFi Number of Passes (mean)

# Data Management report

## CCS Analysis Report – HiFi Read Length Summary

Data Management / Dataset Details

Rhino\_Verif\_HG002\_3ug5\_64441e\_C01\_3280358\_418  
o64\_2\_30hr\_85pM-Cell3 (CCS)

Copy Analyze... Export Delete

Dataset Overview

Adapter Report

Control Report

CCS Analysis Report

Summary Metrics

HiFi Read Length Summary

HiFi Read Quality Summary

Read Length Distribution

Number of Passes

Read Quality Distribution

### HiFi Read Length Summary

Read Length (bp)	Reads	Reads (%)	Yield (bp)	Yield (%)
≥ 0	2,707,732	100	39,236,168,651	100
≥ 5,000	2,664,322	98	39,051,919,399	100
≥ 10,000	2,353,137	87	36,541,368,326	93
≥ 15,000	1,164,272	43	21,435,305,025	55
≥ 20,000	294,460	11	6,522,779,501	17
≥ 25,000	21,062	1	559,040,421	1
≥ 30,000	1,012	0	35,294,569	0
≥ 35,000	388	0	15,240,023	0
≥ 40,000	129	0	5,578,841	0



# Data Management report

## CCS Analysis Report – HiFi Read Quality Summary

**PacBio** Data Management ▾ cguarco (Lab Tech) ⚙️ ?

Data Management / Dataset Details

**Rhino\_Verif\_HG002\_3ug5\_64441e\_C01\_3280358\_418  
064\_2\_30hr\_85pM-Cell3 (CCS)** Copy Analyze... Export Delete

- Dataset Overview
- Adapter Report
- Control Report
- ▼ CCS Analysis Report
  - Summary Metrics
  - HiFi Read Length Summary
  - HiFi Read Quality Summary**
  - Read Length Distribution

### HiFi Read Quality Summary

Read Quality (Phred)	Reads	Reads (%)	Yield (bp)	Yield (%)
≥ Q20	2,707,732	100	39,236,168,651	100
≥ Q30	1,811,377	67	25,413,473,886	65
≥ Q40	679,582	25	8,150,599,400	21
≥ Q50	146,257	5	1,355,549,531	3

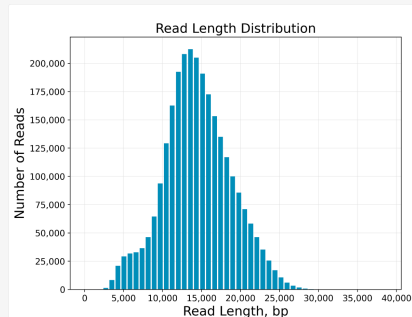
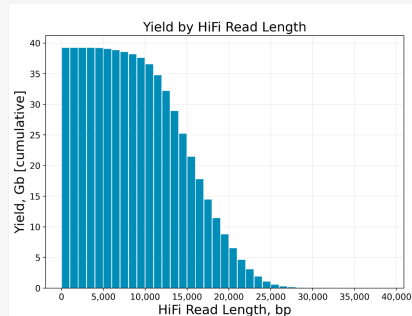
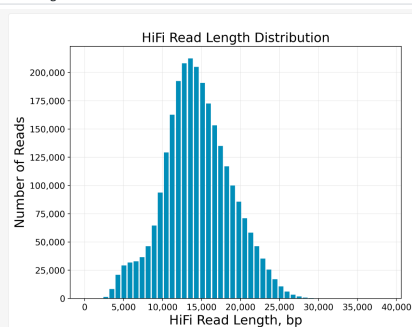
# Data Management report

## CCS Analysis Report

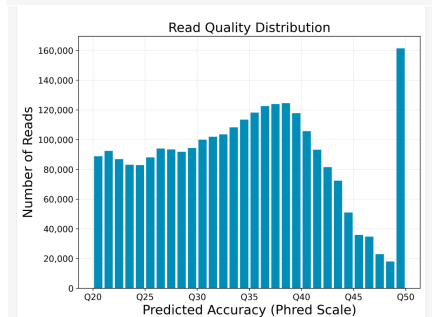
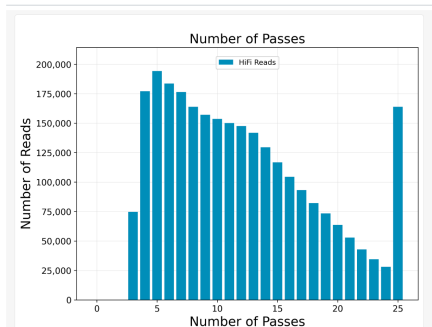
Distributions explain summary metrics

- ▼ CCS Analysis Report
  - Summary Metrics
    - HiFi Read Length Summary
    - HiFi Read Quality Summary
    - Read Length Distribution**
    - Number of Passes
    - Read Quality Distribution
    - Predicted Accuracy vs. Read Length

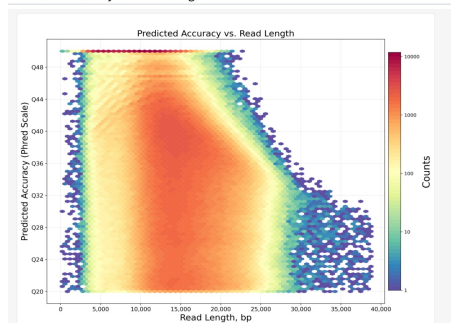
Read Length Distribution



Number of Passes

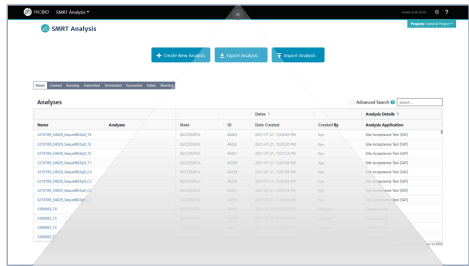
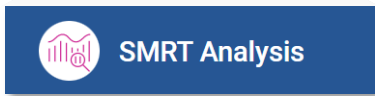


Predicted Accuracy vs. Read Length



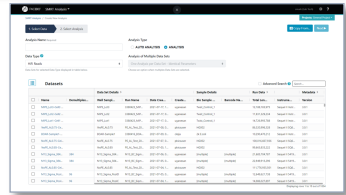
# SMRT Link SMRT Analysis workflow overview

Use **SMRT Link SMRT Analysis** to perform secondary analysis of SMRT sequencing data

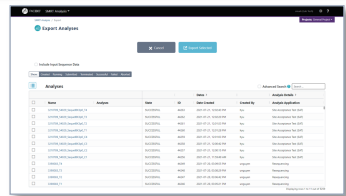


- + Create New Analysis
- Export Analysis
- Import Analysis

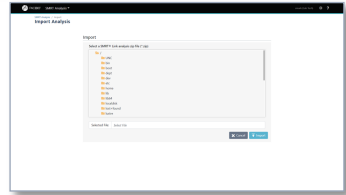
## Create a New Analysis



## Export an Analysis



## Import an Analysis

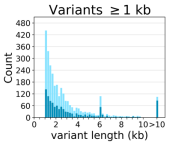


## Example Secondary Analysis Applications



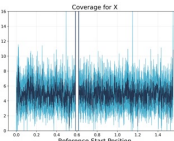
Whole Genome Sequencing

Genome Assembly  
Microbial Genome Analysis  
Structural Variant Calling



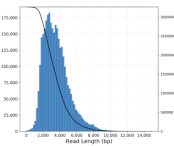
Targeted Sequencing

HiFi Mapping  
SARS-CoV-2 Analysis



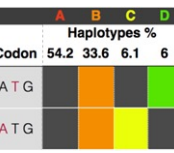
RNA Sequencing

Iso-Seq Analysis



Complex Populations

Minor Variants Analysis



# Applications support documentation

## Application notes & best practices guides

### Whole genome sequencing applications

- Application brief – Whole genome sequencing for de novo assembly – Best practices ([102-193-627](#))
- Application brief – Microbial whole genome sequencing – Best practices ([102-193-601](#))

### RNA sequencing applications

- Application note – MAS-Seq for single cell isoform sequencing ([102-326-549](#))

### Metagenomics applications

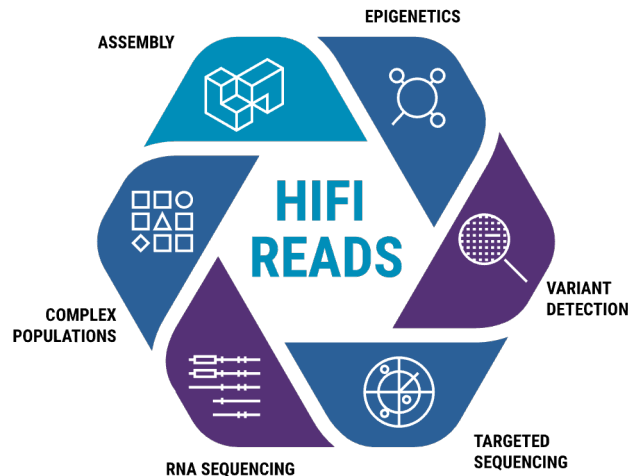
- Application brief – Metagenomic sequencing with HiFi reads – Best practices ([102-193-684](#))

### Targeted sequencing applications

- Application brief – HiFi target enrichment – Best practices ([102-193-603](#))
- Application brief – Targeted sequencing for amplicons – Best practices ([102-193-603](#))

### Application technical overviews

- Technical overview – MAS-Seq library preparation using the MAS-Seq for 10x Single Cell 3' kit ([102-829-300](#))
- Technical overview – Multiplexed amplicon library preparation using SMRTbell prep kit 3.0 ([102-395-900](#))
- Technical overview – Nanobind HT kits for automated HMW DNA extraction (Coming soon)
- Technical overview – Whole genome and metagenome library preparation using SMRTbell prep kit 3.0 ([102-390-900](#))



# Technical documentation & training resources

## SMRT Link & other data analysis documentation

- Brief primer and lexicon for PacBio SMRT sequencing webpage ([v12.0](#))
- PacBio bioinformatics file formats documentation webpage ([v12.0](#))
- SMRT Link v12.0 cloud reference guide ([102-978-000](#))
- SMRT Link v12.0 release notes ([102-877-200](#))
- SMRT Link v12.0 software installation guide ([102-878-100](#))
- SMRT Link v12.0 user guide ([102-877-300](#))
- SMRT Link v12.0 web services API use cases ([102-982-400](#))
- SMRT Tools v12.0 reference guide ([102-978-000](#))



# Assembly Analysis Application

27 June 2023

彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

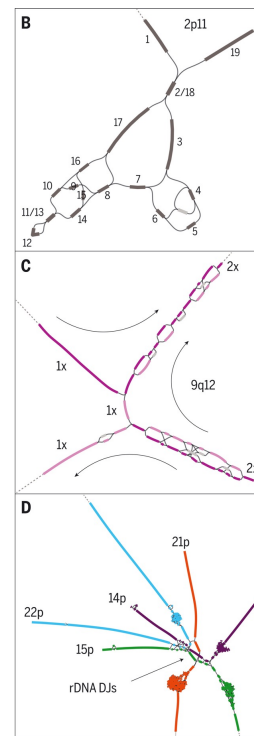
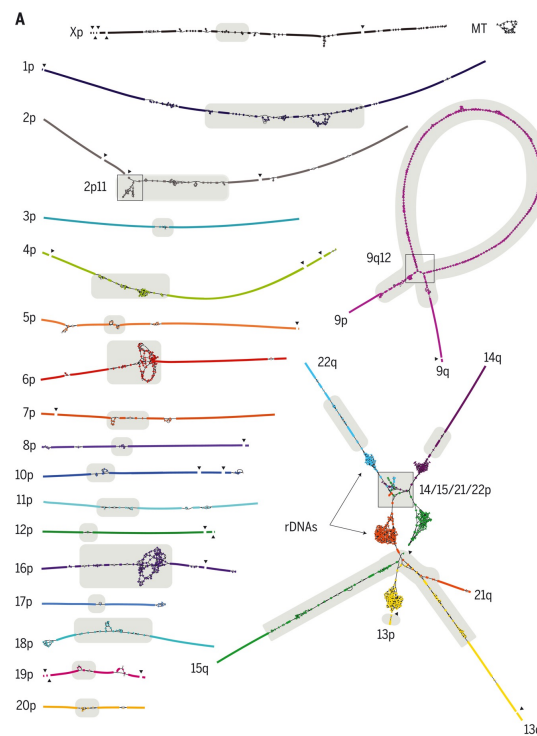
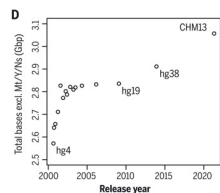
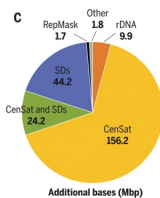
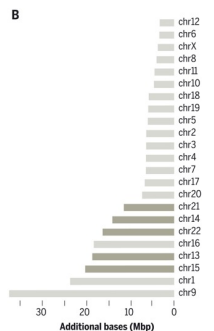
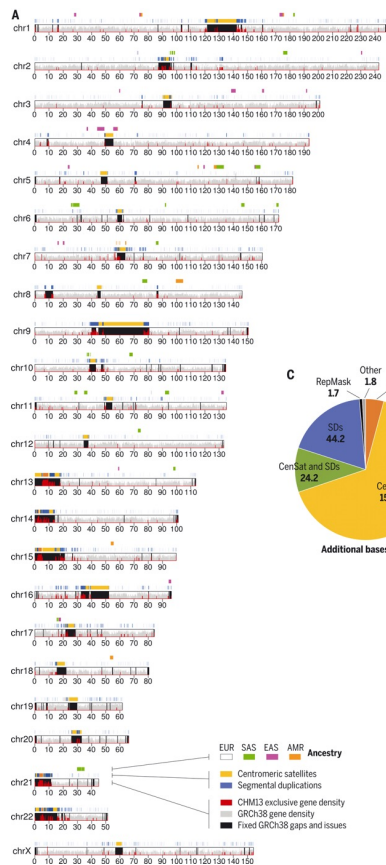
# Assembly of the First COMPLETE Human Genome With HiFi Reads



The basis of the first gapless human reference genome T2T-CHM13 assembly is a high-resolution assembly string graph built directly from HiFi reads

“In contrast to the first T2T assembly of chromosome X—which relied on ONT sequencing to create a backbone that was then polished with other technologies—we shifted to a new strategy that leverages the combined accuracy and length of HiFi reads to enable assembly of highly repetitive centromeric satellite arrays and closely related segmental duplications”

# High-resolution assembly string graph of the CHM13 genome



**“The all-star of this assembly has been PacBio HiFi.”**  
Adam Phillippy, NIH



# T2T-CHM13 Improves Understanding of the Genome

Table 1. Comparison of GRCh38 and T2T-CHM13 human genome assemblies.

Summary	GRCh38p13	CHM13v1.1	±%
Assembled bases (Gbp)	2.92	3.05	+4.5%
Unplaced bases (Mbp)	11.42	0	-100.0%
Gap bases (Mbp)	120.31	0	-100.0%
# Contigs	949	24	-97.5%
Ctg NG50 (Mbp)	56.41	154.26	+173.5%
# Issues	230	46	-80.0%
Issues (Mbp)	230.43	8.18	-96.5%
<b>Gene Annotation</b>			
# Genes	60,090	63,494	+5.7%
protein coding	19,890	19,969	+0.4%
# Exclusive genes	263	3,604	
protein coding	63	140	
# Transcripts	228,597	233,615	+2.2%
protein coding	84,277	86,245	+2.3%
# Exclusive transcripts	1,708	6,693	
protein coding	829	2,780	

200 million bp of novel sequence

Gapless assemblies of 22 autosomes, X chromosome, and mitochondrial genome

Adds 2,226 paralogous gene copies, including 115 predicted to be protein coding

# Build the best reference genomes

## Capturing Biodiversity



APR 28 2021 Research

### Project to Read Genomes of All 70,000 Vertebrate Species Reports First Discoveries

#### Summary

A bold project to read the complete genetic sequences of every known vertebrate species reaches its first milestone by publishing new methods and the first 25 high-quality genomes.



#### Scientist Profiles

**Erich D. Jarvis**  
The Rockefeller University  
Neuroscience, Molecular Biology

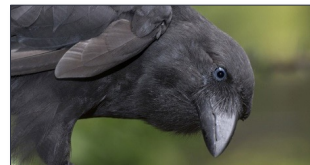
**David Haussler**  
University of California, Santa Cruz  
Computational Biology, Molecular Biology

**Beth Shapiro**  
University of California, Santa Cruz  
Evolutionary Biology, Genetics

FOR MORE INFORMATION  
Meghan Rosen  
301-215-8659  
rosenm@nih.gov

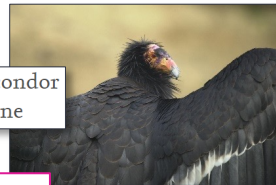
## Species Monitoring

A High-Quality, Long-Read *De Novo* Genome Assembly to Aid Conservation of Hawaii's Last Remaining Crow Species



Vaquita Genome Offers Hope for Species' Survival

Genome-wide diversity in the California condor tracks its prehistoric abundance and decline



Asian giant hornet (*Vespa mandarinia*)

Asian Giant Hornet Complete Genome Released by the Agricultural Research Service

## Plant & Animal Biology

Leveraging the secrets of hibernation to treat diabetes



The New York Times

A Question Hidden in the Platypus Genome: Are We the Weird Ones?

Researchers have produced the most comprehensive platypus genome yet, as well as that of another monotreme, an echidna.



Platypuses diverged from other mammals about 167 million years ago, making "very important for understanding mammalian evolution," says Takao Saito

# Mistletoe genome

From the Darwin *Tree of Life* project:

2022: The year we built the biggest genome in Britain and Ireland

Luke Lythgoe | 16 Dec 2022

**90 Gb genome size**  
(30 × human genome)

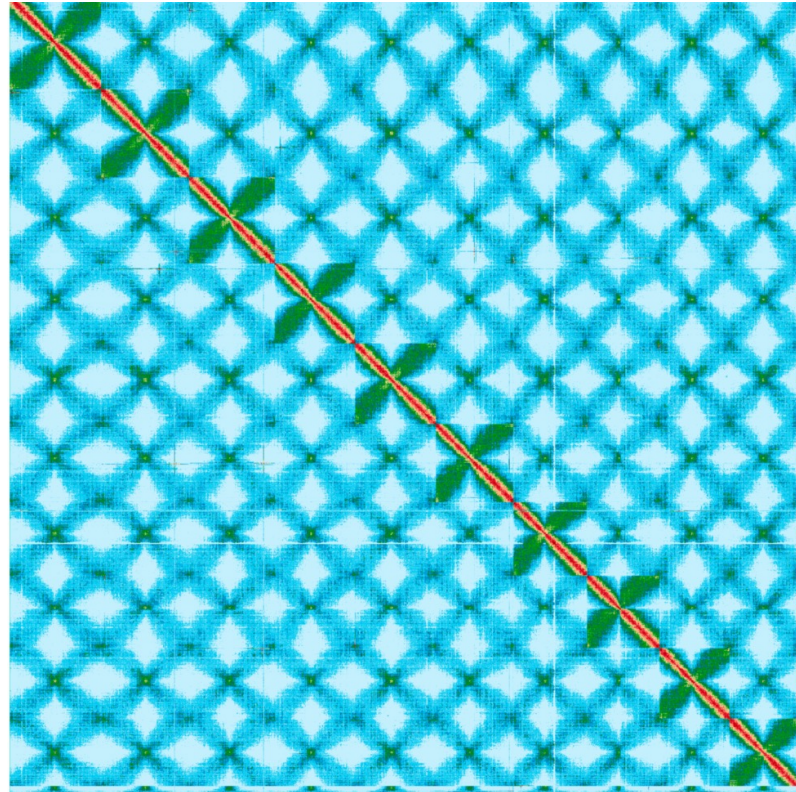
DECEMBER 22, 2022 | PLANT + ANIMAL BIOLOGY

**The HiFi difference Christmas edition  
– Big genomes**



**PacBio** 

[pacb.com/blog](https://pacb.com/blog)

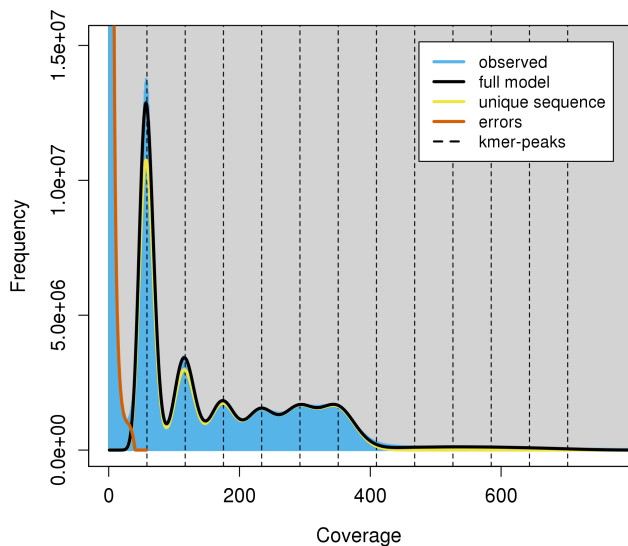


human genome

# Hexaploid persimmons ( $6 \times 800\text{Mb}$ )



Collaboration with Jeremy Schmutz (HudsonAlpha Institute for Biotechnology)  
& Scott Brainard (Savanna Institute)



HudsonAlpha Booth # 433



data from 1 SMRT Cell mapped to a primary hifiasm contig



PACBIO®

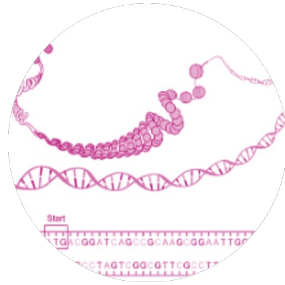
**Why switch to HiFi for *de novo* assemblies?**

# HiFi reads for improved assembly



## Contiguity

Resolve repetitive regions  
High contig n50



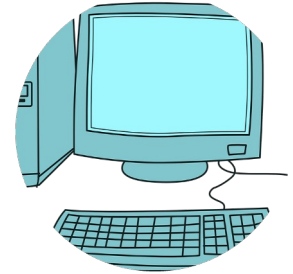
## Correctness

Base qv  
Phasing accuracy



## Completeness

Gene space  
Repetitive regions



## Compute

Cpu/wall time  
Ram  
Disk storage

# Hifiasm Propels Hifi to longer contiguity than any other long read technology

## Haplotype-resolved *de novo* assembly with phased assembly graphs

Haoyu Cheng<sup>1,2</sup>, Gregory T Concepcion<sup>3</sup>, Xiaowen Feng<sup>1,2</sup>, Haowen Zhang<sup>4</sup>, and Heng Li<sup>1,2,\*</sup>

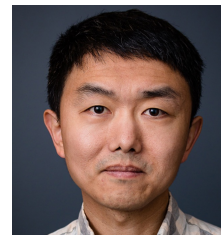
<sup>1</sup>Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

<sup>2</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>3</sup>Pacific Biosciences, Menlo Park, CA, USA

<sup>4</sup>College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

\*To whom correspondence should be addressed: hli@jimmy.harvard.edu



Dr. Heng Li  
Dana-Farber Cancer Institute, Harvard

Table 2. Statistics of human primary assemblies

Dataset	Assembly	Size (Gb)	NG50 (Mb)	NGA50 (Mb)	QV	Multi-copy genes retained (%)	Resolved BACs (%)	Gene completeness (asmgene)	
								Complete (%)	Duplicated (%)
CHM13 (HiFi 32× ONT 120×)	hifiasm	3.043	88.1	65.4	54.3	76.9	95.3	99.14	0.28
	HiCanu	3.047	76.3	59.4	53.9	76.7	96.5	99.13	0.33
	Peregrine	2.990	36.5	33.2	43.8	41.4	38.4	98.84	0.26
	Falcon	2.862	26.3	23.8	50.1	24.6	33.1	98.62	0.11
	Canu (ONT)	2.992	74.1	60.5	26.6	61.6	92.1	97.79	0.27
HG00733 (HiFi 33× ONT 50×)	hifiasm (purge)	3.039	70.0	56.8	49.8	67.3	83.2	99.09	0.31
	HiCanu (purge)	2.932	35.2	31.6	50.7	62.4	73.7	97.76	0.33
	Peregrine	3.035	30.1	30.1	40.5	37.2	38.5	98.70	0.31
	Falcon	2.861	24.4	23.2	46.3	33.6	38.0	96.51	0.15
	Canu (ONT)	2.834	40.5	35.1	22.7	22.5	69.3	91.26	0.14
HG002 (HiFi 36× ONT 80×)	hifiasm (purge)	3.063	98.7	65.4	51.4	74.8		99.31	0.35
	HiCanu (purge)	3.000	44.7	35.9	52.1	67.1		98.97	0.23
	Peregrine	3.081	33.4	32.5	41.3	42.5		99.14	0.36
	Falcon	2.955	30.4	29.0	46.7	36.6		99.00	0.20
	Canu (ONT)	2.831	32.3	30.5	21.9	19.6		88.94	0.21

<https://arxiv.org/abs/2008.01237>

# HiFi vs. CLR in human

- HiFi is as contiguous as CLR
- HiFi is more accurate than CLR

	CHM13 (Vollger <i>et al.</i> 2019)	
Data Type	CLR	HiFi
Coverage	77-fold	24-fold
Contig N50	<b>29.2</b>	<b>29.5</b>
Median Base QV	<b>40.7</b>	<b>45.0</b>
Method	FALCON, Arrow	Canu, Racon

bioRxiv THE PREPRINT SERVER FOR BIOLOGY

Confirmatory Results

**Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads**

Mitchell R. Vollger, Glennis A. Logsdon, Peter A. Audano, Arvis Sulovari, David Porubsky, Paul Peluso, Aaron M. Wenger, Gregory T. Concepcion, Zev N. Kronenberg, Katherine M. Munson, Carl Baker, Ashley D. Sanders, Diana C.J. Spierings, Peter M. Lansdorf, Urvashi Surti, Michael W. Hunkapiller, Evan E. Eichler

doi: <https://doi.org/10.1101/635037>

bioRxiv THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

New Results 2 comments

**Highly-accurate long-read sequencing improves variant detection and assembly of a human genome**

Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Toepfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, Michael W Hunkapiller

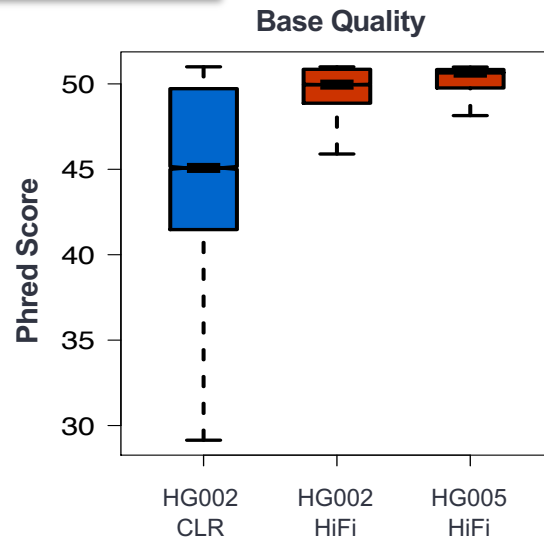
doi: <https://doi.org/10.1101/519025>

HG002 Data:

<https://bit.ly/2RW1b3I>

HG005 Data:

[PRJNA540706](https://bit.ly/2RW1b3I)





# COMPUTE: California redwood project



*Sequoia sempervirens*

## 17 days for entire project:

- sample collection
- library
- sequencing
- assembly

## hifiasm

Haoyu Cheng

Heng Li Lab

- 6 days wall time

- 64 cores with 512 GB RAM

- ~7,200 CPU hrs asm

- ~46,000 CPU hrs CCS

## Versus ONT + ILM Assembly

**Assembly took a while...**

- Maximum memory usage: 2 Tb
- Error correction: 330,000 CPU hours
- Assembly post-error-correction: 700,000 CPU hours
- Wall clock time: 5-6 months

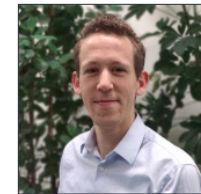
<https://medium.com/pacbio/a-genome-fit-for-a-giant-sequencing-the-california-redwood-ed722be9e49c>

# Polyploid genomes

Depends on type of polyploidy

Separate subgenomes accurately  
vs. one primary contig and multiple  
haplotigs

Future work to explore, HiFi data  
ideal to separate alleles



Stream A

Recent tetraploid: domesticated cotton

Year	Build	Size	Contig #	Contig N50
2015	Illumina Pooled Clone <sup>1</sup>	2.1 GB	44,816	0.68 MB
2015	WGS Illumina plus BES <sup>2</sup>	2.1 GB	265,279	0.63 MB
2019	Illumina plus 10x, DeNovoMap <sup>3</sup>	2.2 GB	53,849	0.11 MB
2018	RGB Long Read, MECAT + Clones V2 <sup>4</sup>	2.3 GB	6,743	0.78 MB
2020	SequelII Long Read, MECAT V3 <sup>5</sup>	2.3 GB	314	40.0 MB
2020	CCS, CANU2 build, unfinished <sup>6</sup>	2.3 GB	<100	70.8 MB

BIODIVERSITY GENOMICS 2020

Jeremy Schmutz

USDA National Science Foundation Cotton Recaptured ISS

Genetwister

## HQ long reads improves haplotyping

DH 'Old Blush' ref

Illumina

PacBio HiFi

<https://www.pacb.com/blog/rose/>

## Generate **contiguous** assemblies with HiFi

Dataset	Drosophila		Human	
Mode	Long Reads	HiFi Reads	Long Reads	HiFi Reads
Size Selection	BP >15 kb	19 kb ELF	BP >15 kb	15 kb ELF
Coverage	71-fold	20-fold	50-fold	22-fold
Contig N50 (Mb)	3.5	6.5	12.6	30.5

Contiguity is **equivalent, if not better** using lower coverage HiFi data. This is true even for highly repetitive genomes as you can now see **minute differences** between repeats.

# Generate **complete** and **correct** assemblies with HiFi

Dataset	Drosophila		Human	
Mode	Long Reads	HiFi Reads	Long Reads	HiFi Reads
Assembly Size	0.148	<b>0.150</b>	2.85	<b>2.92</b>
Base pair accuracy (Phred/Percentage)	Q44 / 99.996%	<b>Q50 / 99.999%</b>	Q41 / 99.992%	<b>Q49 / 99.9987%</b>
BUSCO complete	N=2,799 98.8%	N=2,799 <b>98.9%</b>	N=4,104 94.8%	N=4,104 <b>94.9%</b>
Species-specific genes in frame	N=13,947 98.8%	N=13,947 <b>99.5%</b>	N=19,313 96.4%	N=19,313 <b>99.5%</b>

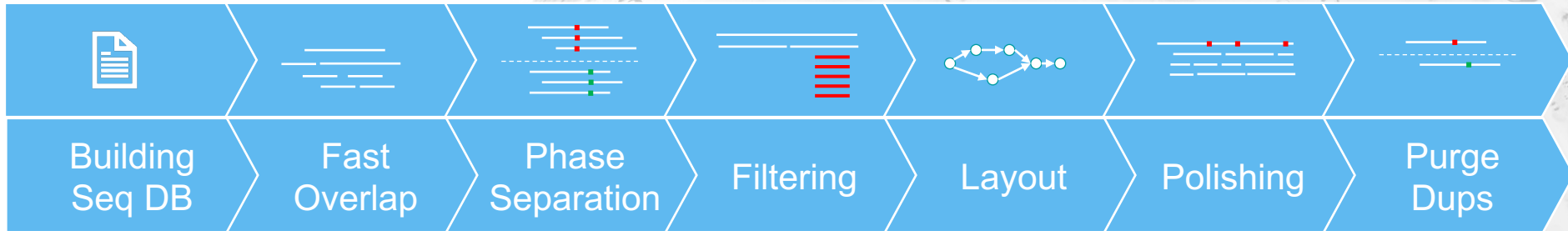
Accuracies approaching **Q50 (99.999%)**  
**>99%** of genes in frame  
**More** of the genome assembled

# Improved and Phased Assembly (IPA)



# IPA WORKFLOW

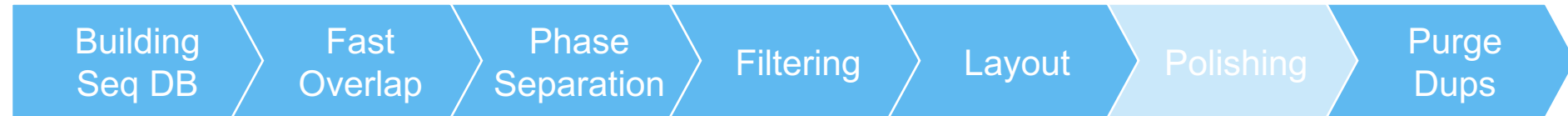
Phased workflow



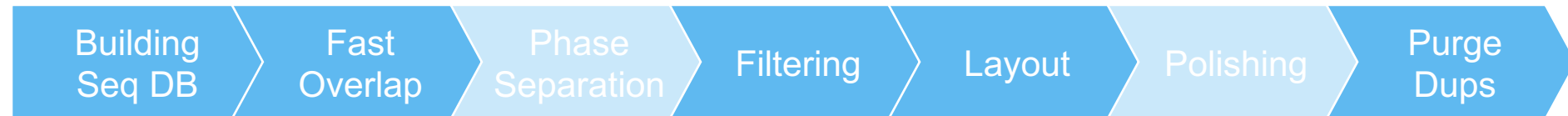
# IPA WORKFLOW Is Modular

Polishing can optionally be switched on/off: Fast draft assembly

Phased workflow



Haploid workflow



# HiFi WGS data analysis recommendations for large genomes

## Using HiFi reads for *de novo* assembly analysis of large genomes

- Perform CCS analysis on-instrument using the Sequel IIe system or in [SMRT Link](#) to generate highly accurate and long single-molecule reads (HiFi reads)
- **10- to 15-fold HiFi read coverage per haplotype** is recommended for most *de novo* assembly projects

→  $Target\ HiFi\ Base\ Yield = [Haploid\ Genome\ Size\ (Gb)] \times [Ploidy\ Level] \times [Target\ HiFi\ Coverage\ per\ Haplotype]$

E.g., for *de novo* assembly analysis of a 3 Gb diploid genome:

Recommended Minimum Target HiFi Base Yield = 3 Gb x 2 x 10 = 60 Gb

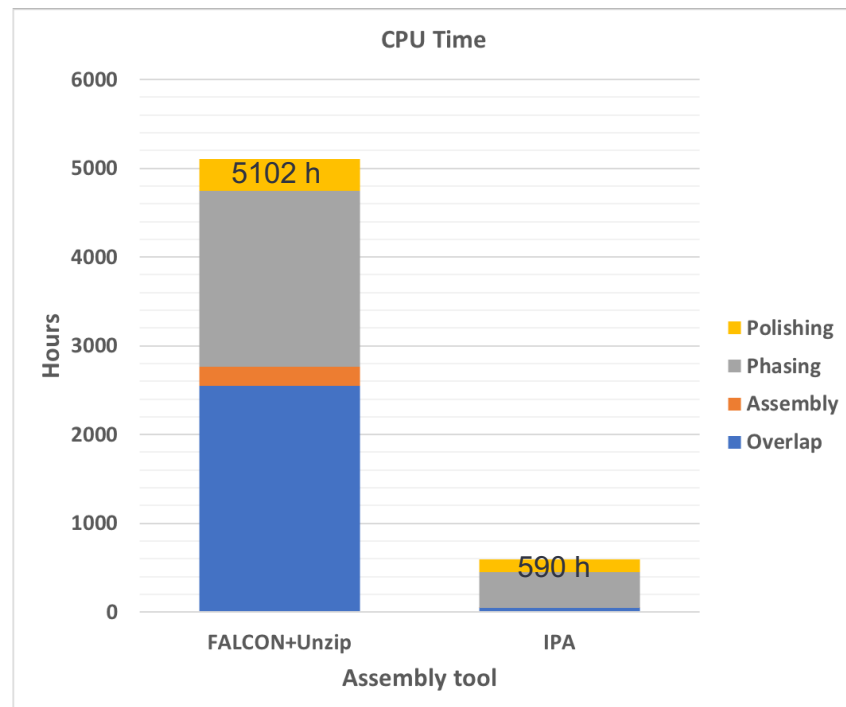
- Output data in standard file formats, (BAM and FASTA/Q) for seamless integration with downstream analysis tools
- Can use [SMRT Link](#) Genome Assembly analysis application (powered by [IPA](#)) or other third-party software for *de novo* assembly analysis using HiFi reads:
  - [Hifiasm](#)
  - [HiCanu](#)
- Contact PacBio Technical Support ([support@pacb.com](mailto:support@pacb.com)) or your local Bioinformatics Field Applications Scientist for additional information about data analysis recommendations



# Results: long phase blocks in human, high base QV

HPRC HG002 34x Dataset – Phased workflow with polishing – no “purge\_dups”

	FALCON-Unzip		IPA (Phased)	
	primary	haplotigs	primary	haplotigs
N50 [Mbp]	31.40	0.191	33.75	0.352
Max length [Mbp]	110.12	1.62	110.94	2.30
Total length [Gbp]	2.95	1.99	3.02	1.85
CPU time [h]	5102		590	
	<b>8.64x Faster!</b>			



# Results: great haplotig separation

## Atlantic Bluefin Tuna (0.85%)

	IPA + purge_dups	
	primary	haplotigs
Contig N50	20.12 Mb	4.21 Mb
Assembly size	<b>790 Mb</b>	<b>730 Mb</b>
BUSCO [C]	98.6%	92.8%



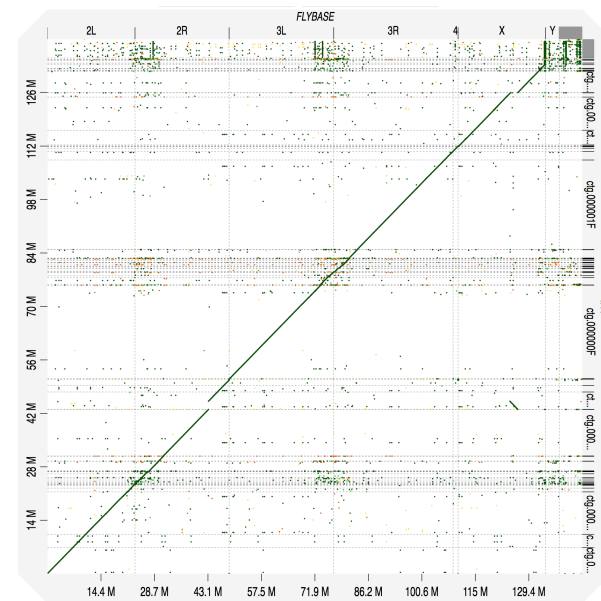
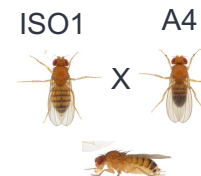
## Red Admiral Butterfly (1.1%)

	IPA + purge_dups	
	primary	haplotigs
Contig N50	12.12 Mb	4.80 Mb
Assembly size	<b>368 Mb</b>	<b>369 Mb</b>
BUSCO [C]	99.3%	97.7%



# Results: high PHASE accuracy

## *Drosophila melanogaster* F1 – Phased and polished



Hifiasm  
+ purge\_dups

IPA  
+ purge\_dups

primary

haplotigs

primary

haplotigs

N50 [Mbp] 22.55 1.28 13.49 **2.42**

Max length [Mbp] 28.13 6.81 23.47 12.48

Total length [Mbp] 160.19 149.87 **134.19** 115.26

Base QV 48.1 47.4 47.97 46.87

Phase accuracy 0.788 0.998 **0.826** **0.999**

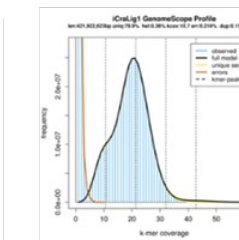
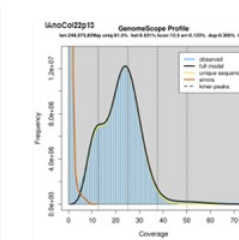
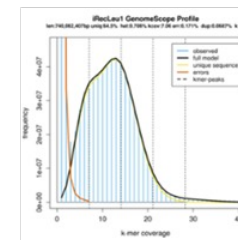
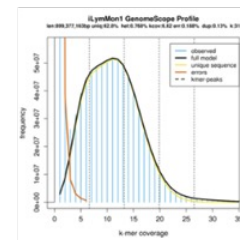
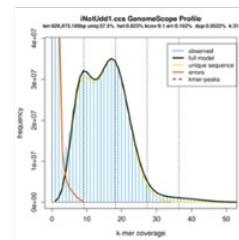
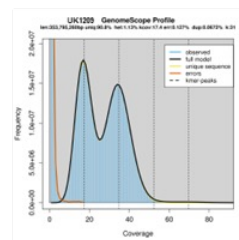
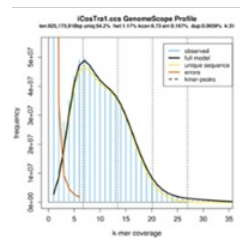
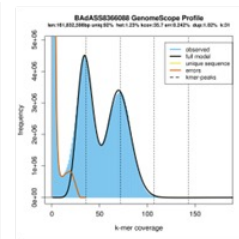
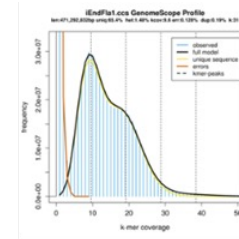
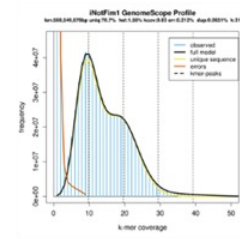
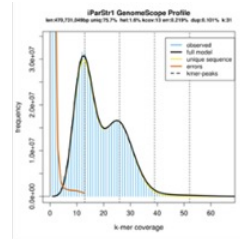
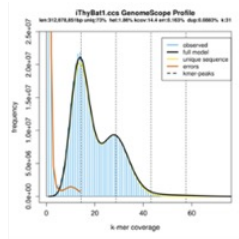
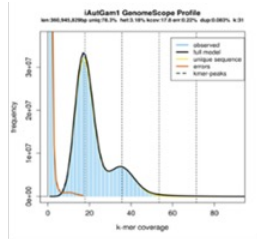
BUSCO of primary C:98.5% S:98.0%,D:0.5% C:98.7% S:98.2%,D:0.5%

# Results: Bug Genomes



Darwin Tree of Life, SANGER

Testing on real-world samples - butterflies, moths & mosquitoes



The background of the slide is a blurred laboratory setting. In the foreground, a multi-well plate is visible, with several wells containing a bright pink liquid. A pipette tip is positioned above one of the wells, with a single drop of the pink liquid about to fall. The word "PacBio" is overlaid in a bold, pink font in the upper right quadrant of the image.

**PacBio**

# **Microbial assembly analysis application**

# Microbial whole genome sequencing and assembly with HiFi data



## Complete microbial genomes

including chromosomes and plasmids



## High contiguity

high per-base quality of final microbial assemblies

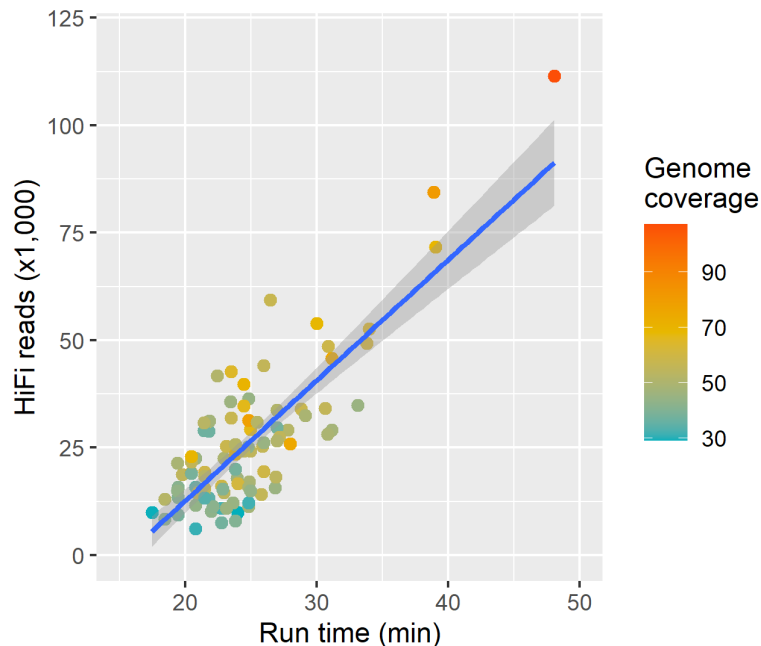


## Fast assembly,

easy to use, no need for parameter input/optimization

# Short turn-around times

Typical time to results for Microbial Assembly analysis is ~20 to 60 minutes\*



## Minimum compute requirements:


Head Node:

Cores: 32, RAM: 64 GB,  
1 TB local tmp, 256 GB local db\_datadir

Compute Nodes:

Cores: 64, RAM: 4GB per core,  
1 TB local tmp, 256 GB local db\_datadir





# Experimental design and input data requirements



# HiFi WGS data analysis recommendations small genomes (microbial multiplexing applications)

## Using HiFi reads for *de novo* assembly and base modification detection analysis of microbial genomes

- Perform CCS analysis on-instrument using the Sequel IIe System or in [SMRT Link](#) to generate highly accurate and long single-molecule reads (HiFi reads)
- **15-fold HiFi read coverage per microbe** is recommended for most *de novo* assembly projects
  - $Target\ HiFi\ Base\ Yield = [Microbe\ Genome\ Size\ (Mb)] \times [Target\ HiFi\ Coverage\ per\ Microbe]$ 

E.g., for *de novo* assembly analysis of a 5 Mb microbial genome:  
Recommended Minimum Target HiFi Base Yield = 5 Mb x 15 = 75 Mb
- Output data in standard file formats, (BAM and FASTA/Q) for seamless integration with downstream analysis tools
- Can use [SMRT Link](#) Microbial Genome analysis application for *de novo* assembly and base modification detection analysis using HiFi reads:
  - **Easy to use** (no requirement for laborious parameter input/optimization)
  - **Enables fast and efficient** microbial assembly results using HiFi reads (typical time to result is ~20-60 minutes\* for analysis of a 96-plex microbial data set (up to 375 total sum of genome sizes))
  - **Outputs complete, high-quality** microbial genome assemblies (including chromosomes and plasmids)

# WGS sample preparation procedure description

Procedure & Checklist – Preparing whole genome and metagenome libraries using SMRTbell prep kit 3.0 ([102-166-600](#)) describes a method for constructing SMRTbell libraries that are suitable for generating HiFi reads on the Sequel II and IIe systems for **WGS and metagenomic shotgun sequencing applications**.

## Procedure Highlights

- Uses **SMRTbell Prep Kit 3.0** (102-182-70) and supports high-throughput processing using **500 ng – 5 µg** of input genomic DNA amounts
  - We recommend starting with **≥1 µg of input DNA per SMRT Cell 8M** (or ~3 µg for up to a 3 Gb WGS sample to enable running 3 SMRT Cells 8M)
- Multiplexing of samples can be performed using **SMRTbell barcoded adapter plate 3.0** (102-009-200)
- Recommend shearing high-quality gDNA using a **Megaruptor 3 System** (Diagenode)
  - **15 kb – 18 kb** target insert size for large (plant / animal / human) genomes
  - **7 kb – 12 kb** target insert size for small (microbial) genomes
  - **7 kb – 12 kb** target insert size for shotgun metagenomic samples
- **4.5-hour workflow time** to process up to 8 samples from shearing to size selection (6 hours for 24 samples)
  - Time difference is from DNA shearing, which can be performed in sets of 8 samples.
  - Excludes time needed for DNA sizing QC analysis using a Femto Pulse system.
- WGS SMRTbell libraries can be **size-selected using AMPure PB Beads** without the need for third-party equipment

Preparing whole genome and metagenome libraries using SMRTbell® prep kit 3.0

Procedure & checklist

### Before you begin

This procedure describes the workflow for constructing whole-genome sequencing (WGS) libraries from genomic and metagenomic DNA using the SMRTbell prep kit 3.0 for sequencing on PacBio systems.

Overview			
Samples per SMRTbell prep kit 3.0	1–24		
Workflow time	4.5 hours for up to 8 samples, 6 hours for 24 samples Time difference is from DNA shearing, which is done in sets of 8 samples. Excludes measuring DNA size on Femto Pulse system.		
DNA input			
Quantity	300 ng–5 µg per library		
	Human, plant, and animal	Microbes	Metagenomes
DNA size distribution (Femto Pulse system)	50% ≥ 30 kb & 90% ≥ 10 kb	90% ≥ 7 kb	90% ≥ 7 kb
DNA Shearing (Megaruptor 3 system)	Speed 31	Speed 40	Speed 40
Target fragment lengths	15–18 kb	7–12 kb	7–12 kb
Size selection required	AMPure® PB beads	none	none

© 2022 PacBio. All rights reserved. Research use only. Not for use in diagnostic procedures.  
PN:102-166-600 EA V1 18FEB2022

PacBio

PacBio Documentation (102-166-600)

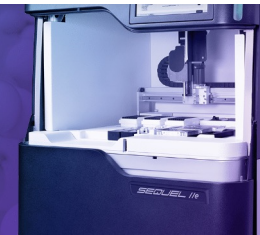
## APPLICATIONS

## WHOLE GENOME SEQUENCING

*De Novo assembly & variant detection*

*Microbial assembly*

*Shotgun metagenomics*





# Example performance

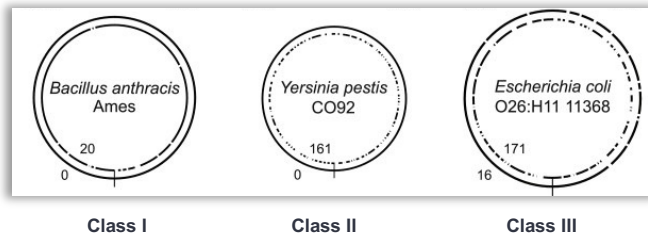
<https://downloads.pacbcloud.com/public/dataset/2021-11-Microbial-96plex/>

# Example sequencing performance for a 96-plex microbial WGS library prepared with SMRTbell prep kit 3.0

## Sample preparation workflow

### Experiment design

- 24 different microbes; each ligated independently to 4 different barcodes for 96-plex



### Microbial genome assembly complexity

**Class I** – Have few repeats except for the rDNA operon sized 5 to 7 kb

**Class II** - Class II genomes have many repeats, such as insertion sequence elements, but none greater than 7 kb.

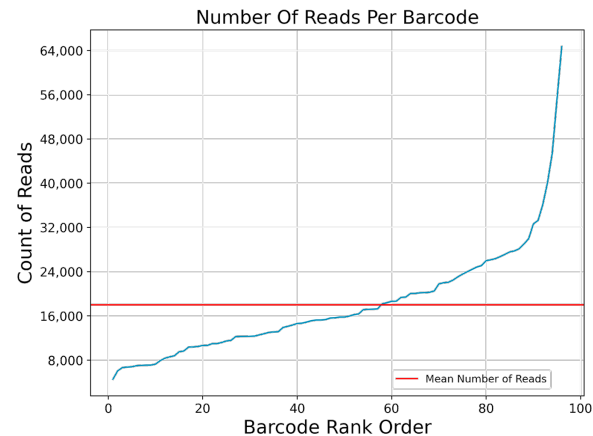
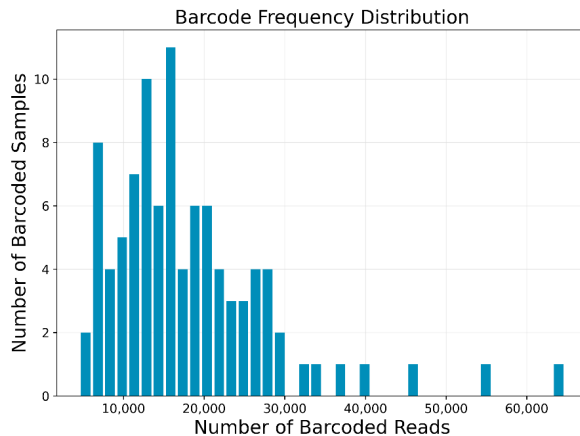
**Class III** - Contain large, often phage-related, repeats >7 kb.

Microbial species	Genome size (bp)	GC content (%)	Microbial genome complexity	Barcode names
<i>Acinetobacter baumannii</i> AYE	3,960,239	39.35	Class 3	bc2001 / bc2025 / bc2049 / bc2073
<i>Bacillus cereus</i> 971	5,430,163	35.29	Class 1	bc2002 / bc2026 / bc2050 / bc2074
<i>Bacillus subtilis</i> W23	4,045,592	43.5	Class 1	bc2003 / bc2027 / bc2051 / bc2075
<i>Burkholderia cepacia</i> UCB 717	8,569,621	66.6	Class 3	bc2004 / bc2028 / bc2052 / bc2076
<i>Burkholderia multivorans</i> 249	7,008,277	66.68	Class 3	bc2005 / bc2029 / bc2053 / bc2077
<i>Enterococcus faecalis</i> OG1RF	2,739,503	37.75	Class 1	bc2006 / bc2030 / bc2054 / bc2078
<i>Escherichia coli</i> H10407	5,393,109	50.71	Class 1	bc2007 / bc2031 / bc2055 / bc2079
<i>Escherichia coli</i> K12 MG1655	4,642,522	50.79	Class 1	bc2008 / bc2032 / bc2056 / bc2080
<i>Helicobacter pylori</i> J99	1,645,141	39.19	Class 1	bc2009 / bc2033 / bc2057 / bc2081
<i>Klebsiella pneumoniae</i> BAA-2146	5,780,684	56.97	Class 2	bc2010 / bc2034 / bc2058 / bc2082
<i>Listeria monocytogenes</i> Li2	2,950,984	37.99	Class 1	bc2011 / bc2035 / bc2059 / bc2083
<i>Listeria monocytogenes</i> Li23	2,979,685	38.19	Class 1	bc2012 / bc2036 / bc2060 / bc2084
<i>Methanococcus labreanum</i> Z	1,804,962	50.5	Class 1	bc2013 / bc2037 / bc2061 / bc2085
<i>Neisseria meningitidis</i> FAM18	2,194,814	51.62	Class 3	bc2014 / bc2038 / bc2062 / bc2086
<i>Neisseria meningitidis</i> Serogroup B	2,304,579	51.44	Class 1	bc2015 / bc2039 / bc2063 / bc2087
<i>Rhodospseudomonas palustris</i> CGA009	5,459,213	64.9	Class 3	bc2016 / bc2040 / bc2064 / bc2088
<i>Salmonella enterica</i> LT2	4,950,860	52.24	Class 1	bc2017 / bc2041 / bc2065 / bc2089
<i>Salmonella enterica</i> Ty2	4,791,947	52.05	Class 1	bc2018 / bc2042 / bc2066 / bc2090
<i>Staphylococcus aureus</i> Seattle 1945	2,806,348	32.86	—	bc2019 / bc2043 / bc2067 / bc2091
<i>Staphylococcus aureus</i> USA300_TCH1516	2,872,915	32.7	Class 1	bc2020 / bc2044 / bc2068 / bc2092
<i>Streptococcus pyogenes</i> Bruno	1,844,942	38.48	—	bc2021 / bc2045 / bc2069 / bc2093
<i>Thermanaerovibrio acidaminovorans</i> DSM6589	1,852,980	63.78	Class 1	bc2022 / bc2046 / bc2070 / bc2094
<i>Treponema denticola</i> A	2,842,721	37.87	—	bc2023 / bc2047 / bc2071 / bc2095
<i>Vibrio parahaemolyticus</i> EB101	5,146,979	45.33	Class 1	bc2024 / bc2048 / bc2072 / bc2096

# Example sequencing performance for a 96-plex microbial WGS library prepared with SMRTbell prep kit 3.0 (cont.)

## Barcode demultiplexing results

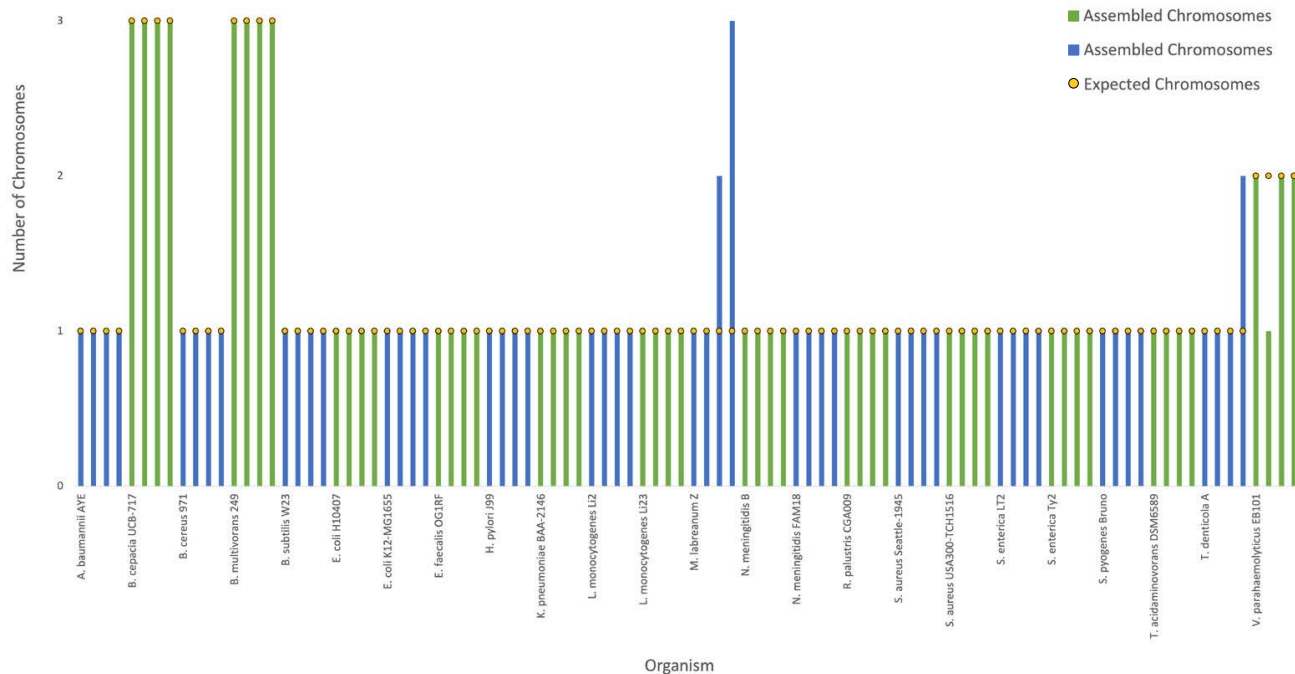
Value	Analysis Metric
96	Unique Barcodes
1,731,704	Barcoded Reads
18,038	Mean Reads
64,709	Max. Reads
4,565	Min. Reads
7,856	Mean Read Length
24,632	Unbarcoded Reads
98.66%	Percent Bases in Barcoded Reads
98.59%	Percent Barcoded Reads



- All 96 barcodes detected
- Mean # of barcoded HiFi reads per microbe is ~18,000
- Mean HiFi base coverage per microbe is 36-fold (Range is 19- to 63-fold)

# Example sequencing performance for a 96-plex microbial WGS library prepared with SMRTbell prep kit 3.0 (cont.)

## HiFi de novo assembly results – assembled chromosomes

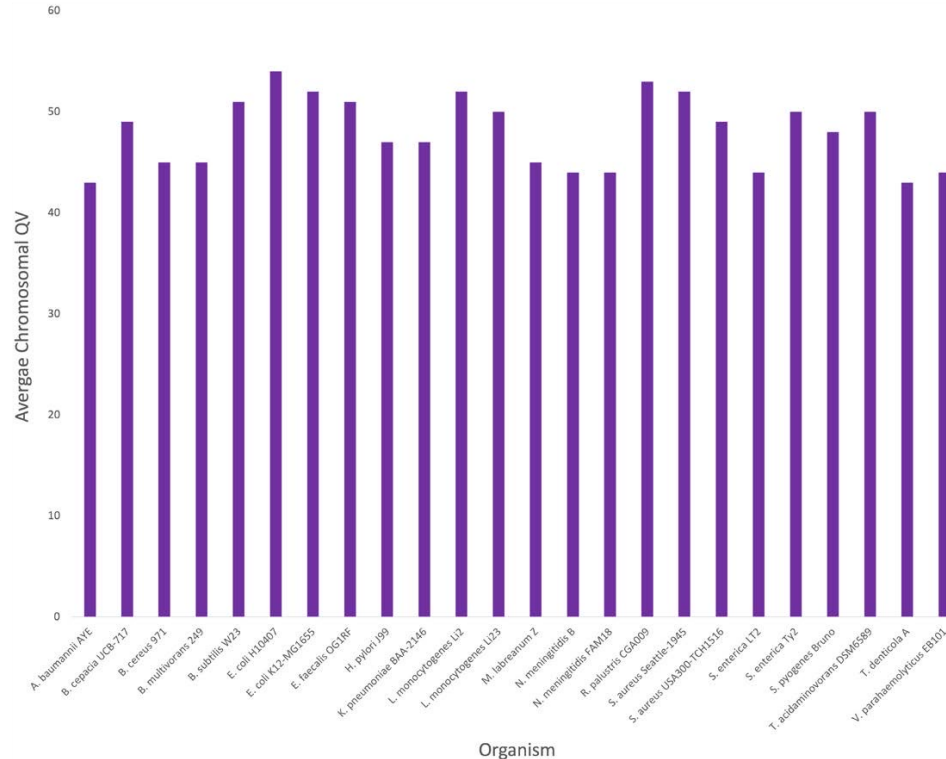


- Achieved 1 Contig / Chromosome for 92 out of 96 assemblies
- For all 96 microbes, chromosomal assemblies were complete and of the expected sizes

Microbial assembly statistics from a 96-plex pool of bacteria relevant to food safety and human health. These data were generated on the Sequel II system and assembled with the fully automated HiFi-based Microbial Assembly application in SMRT Link using the default parameters, without any manual curation. [Download](#) and explore the data yourself.

# Example sequencing performance for a 96-plex microbial WGS library prepared with SMRTbell prep kit 3.0 (cont.)

## HiFi de novo assembly results – representative assembly accuracies



With HiFi data and the Microbial Assembly application in SMRT Link, genome assemblies are **consistently >99.99% accurate**



# Analysis workflow overview



# HiFi microbial assembly workflow

## HiFi microbial assembly workflow stages

Assemble high-quality microbial chromosomes and plasmids

High contiguity, high per-base quality of final microbial assemblies

Fast assembly, easy to use, no need for parameter input/optimization



# Ori-c rotation

## HiFi microbial assembly workflow stages

Chromosomal  
assembly

Mapping  
and  
filtering

Plasmid  
assembly

Filter plasmid  
contigs

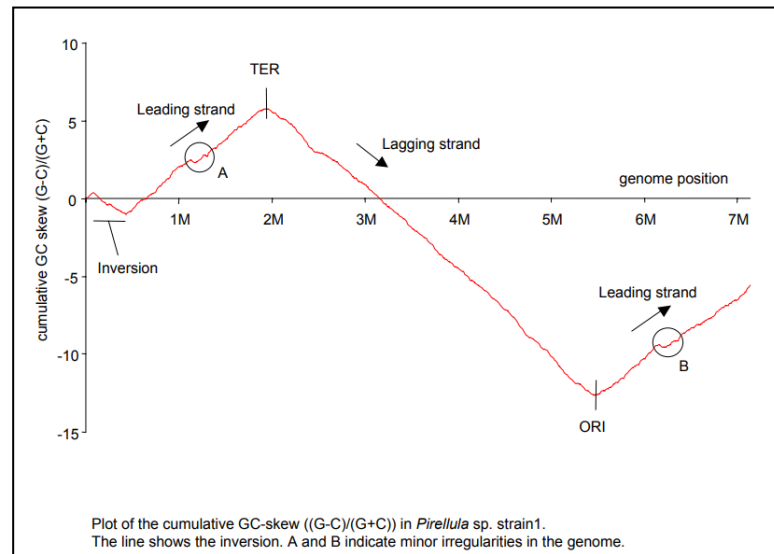
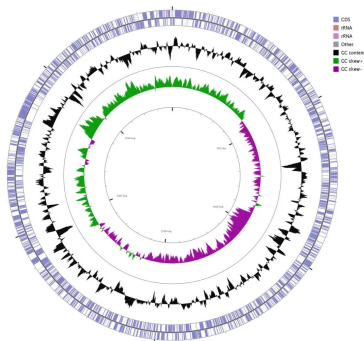
**Ori-c rotation  
& prep for NCBI**

Graph-based  
mapping

Base  
modification  
detection

Task: find origin of replication, header and  
file formatting

Method: GC-skew for origin of replication  
detection





# Analysis results guide

# SMRT Analysis report

## Polished Assembly

**PacBio** SMRT Analysis ▾      Notifications      Settings      Help      admin (Admin)

SMRT Analysis / Analysis Results

**Microbial\_assembly-Klebsiella Sample82**      **SUCCESSFUL**      Copy      Delete

- ▶ Analysis Overview
- ▶ Mapping Report
- ▼ Polished Assembly
  - Summary Metrics**
  - Polished contigs from Microbial Assembly Hifi
- ▶ Coverage
- ▶ Base Modifications
- ▶ Modified Base Motifs
- ▶ Data

### Polished Assembly

Value	Analysis Metric
5	Polished Contigs
5,435,735	Maximum Contig Length
5,435,735	N50 Contig Length
5,781,317	Sum of Contig Lengths
5,117,894	E-size (sum of squares / sum)

# SMRT Analysis report

## Polished Assembly

SMRT Analysis / Analysis Results

### Microbial\_assembly-Klebsiella Sample82

SUCCESSFUL

Copy

Delete

- ▶ Analysis Overview
- ▶ Mapping Report
- ▼ Polished Assembly
  - Summary Metrics
  - Polished contigs from Microbial Assembly Hifi
  - ▶ Coverage
  - ▶ Base Modifications
  - ▶ Modified Base Motifs
  - ▶ Data

#### Polished contigs from Microbial Assembly Hifi

Contig	Length	Circular	Coverage
ctg.s1/p/c/000000/0	5,435,735	yes	34
ctg.s2/p/c/000000/0	140,824	yes	29
ctg.s2/p/c/000001/0	117,755	yes	30
ctg.s2/p/c/000002/0	85,164	yes	32
ctg.s2/p/c/000003/0	1,839	yes	11

# SMRT Analysis report

## Data

### File Downloads

Edit Output File Name Prefix

Example:analysis-Bio Sample 64-955

File
Mapped BAM Index
Mapped BAM
Coverage Summary
Final Polished Assembly for NCBI
PacBio.Index.SamIndex file
Modified Base Motifs
Per-Base IPDs for IGV
Final Polished Assembly
Motif Annotations
Final Polished Assembly Index
Per-Base Kinetics
Modified Bases
Analysis Log
SMRT Link Log

### Final Polished Assembly for NCBI

[ analysis-A\_baumannii\_AYE\_bc2001 -45009-assembly.rotated.polished.renamed.fsa ]

```
>ctg.s1.000000F [topology=circular][completeness=complete]
TCAATTGTGAATAACTTTTTGCACATCCTGTGGATAAAATATCACATAAACTTATCCACAATCCATAAAGACAATAAAAAACAGAGTTA
TCAACAGTTCAAATATATGTTTTTAAATTTAAAAGTGTGAAATCCACAAGAAAAGTCCACACTAATAAGAATAAATTTAAATTTTAA
AATTTGAATTTATTTAATAGGGCTGATCCAAATTTGTGGATAACTAAAAAATATGAATTTAAATTCAAATATACCAAATCAAAACCAAC
TTCACATCAAGGTTTGTGGTAAGTATGTAATAAGAAGTGTATATCTTAAAAGTCTTAATAAAAAATAACAATTACTTTGTGCATAA
CTTTTAAATAAGAAAAATAGGCTAAATATAAAGAGAAGATAAAAAGTTAAAAATTTGACTTAAATACAAAACTTTCACGGTTTTTCAT
TGACAGCGTAAACATTGCACAATAAAATCGCGGACCTTTATAGAAAGATCATTTTTGGGAGTTTCGATATGAAACGTACTTTCCAACC
ATCTGAATTTAA
```

**Final Polished Assembly:** The final polished assembly with applied *oriC* rotation and header adjustment for NCBI submission, in FASTA format.

### Final Polished Assembly

[ analysis-A\_baumannii\_AYE\_bc2001 -45009-p\_ctg\_oric.fasta ]

```
>ctg.s1.000000F shifted_by_bp:-1218400/3943308
TCAATTGTGAATAACTTTTTGCACATCCTGTGGATAAAATATCACATAAACTTATCCACAATCCATAAAGACAATAAAAAACAGAGTTA
TCAACAGTTCAAATATATGTTTTTAAATTTAAAAGTGTGAAATCCACAAGAAAAGTCCACACTAATAAGAATAAATTTAAATTTTAA
AATTTGAATTTATTTAATAGGGCTGATCCAAATTTGTGGATAACTAAAAAATATGAATTTAAATTCAAATATACCAAATCAAAACCAAC
TTCACATCAAGGTTTGTGGTAAGTATGTAATAAGAAGTGTATATCTTAAAAGTCTTAATAAAAAATAACAATTACTTTGTGCATAA
CTTTTAAATAAGAAAAATAGGCTAAATATAAAGAGAAGATAAAAAGTTAAAAATTTGACTTAAATACAAAACTTTCACGGTTTTTCAT
TGACAGCGTAAACATTGCACAATAAAATCGCGGACCTTTATAGAAAGATCATTTTTGGGAGTTTCGATATGAAACGTACTTTCCAACC
ATCTGAATTTAA
```

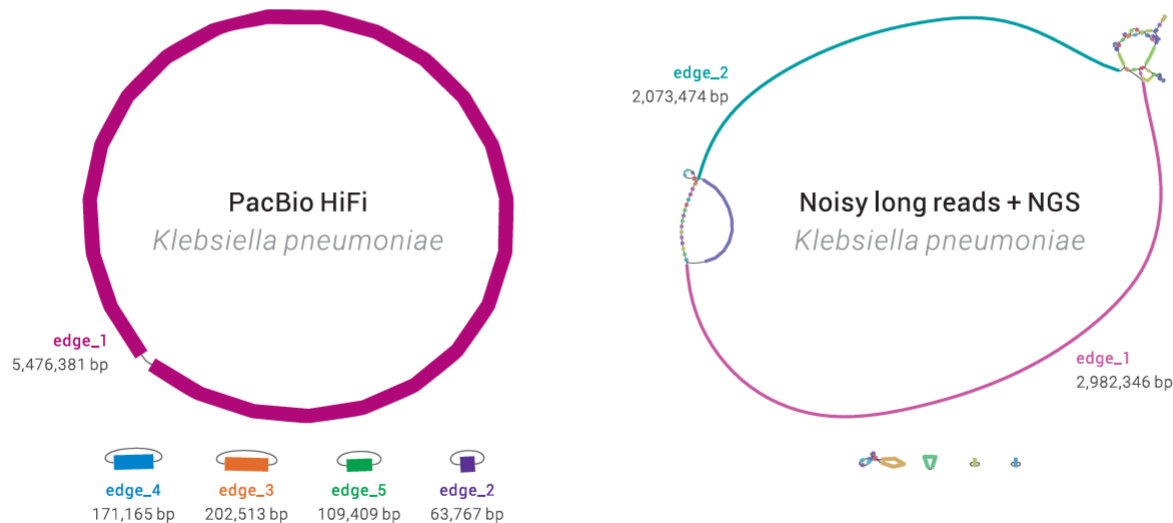
**Final Polished Assembly:** The final polished assembly with applied *oriC* rotation, in FASTA format.



# Case Study Sharing

# Case sharing: Microbial WGS via single technology

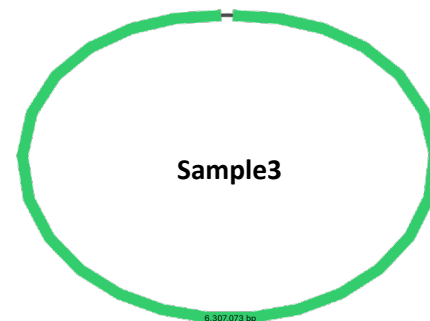
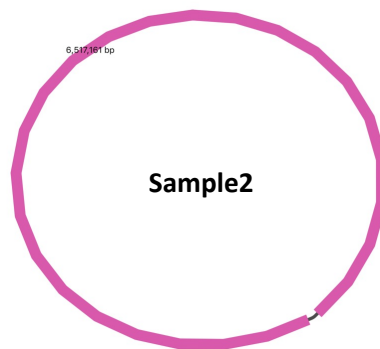
	PacBio HiFi	Noisy long reads + NGS
Coverage	40X	69X (ONT), 34X (ILMN)
Contig N50	5.47 Mb	2.1 Mb
Number of contigs	5	47
Run time	8 min	22 min
Assembler	Flye	Unicycler





# Complete sequence the closed genomes

Sample ID	Number of reads	CCS Reads Length Mean (mapped)	Contig number	Sum of Contig Lengths (bp)	Circular
Sample1	228,521	7,752	1	7,164,774	Y
Sample2	134,703	9,914	1	6,517,161	Y
Sample3	117,743	8,936	1	6,307,073	Y





# Downstream Applications

# Useful tools for further analysis

## Genome Annotation

- Kbase: <http://kbase.us/>
- Prokka: <https://github.com/tseemann/prokka>
- RAST: <http://rast.theseed.org/FIG/rast.cgi>

## Comparative Analysis

- QUAST: <http://quast.sourceforge.net/quast>
- MUMMER: <http://mummer.sourceforge.net/>
- Assemblytics: <http://assemblytics.com/>

## Visualization

- Ribbon: <http://genomeribbon.com/>
- IGV: <https://igv.org/>
- BUSCO: <https://busco.ezlab.org/>

## Other Genome Assembly tools

- FLYE: <https://github.com/fenderglass/Flye>
- Unicycler: <https://github.com/rswick/Unicycler>
- Ciclator: <https://sanger-pathogens.github.io/ciclator/>
- Canu(including Trio Binning Assembly):
  - <https://github.com/marbl/canu>
  - <https://canu.readthedocs.io/en/latest/quick-start.html>
- hifiasm: <https://hifiasm.readthedocs.io/en/latest/index.html>

# Proksee

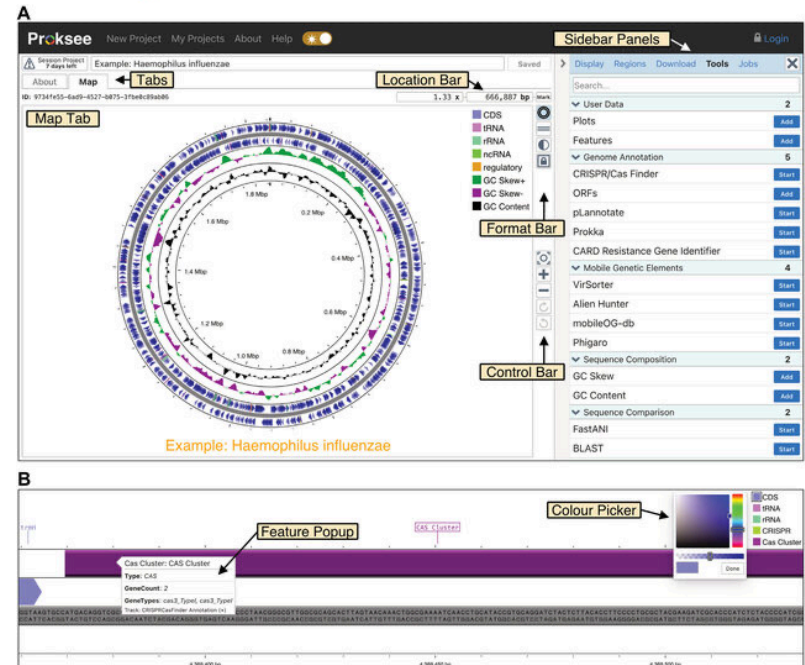
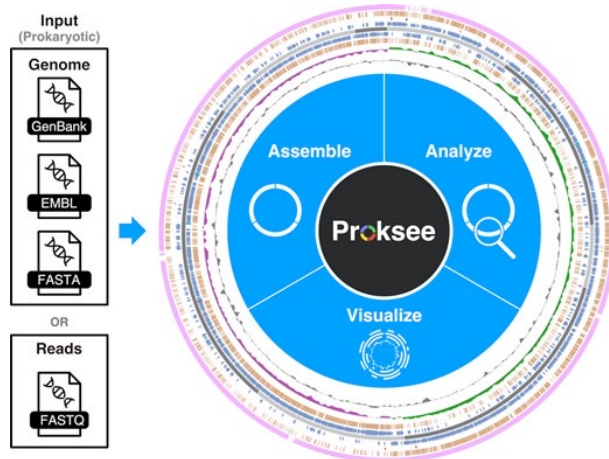
JOURNAL ARTICLE

## Proksee: in-depth characterization and visualization of bacterial genomes

Jason R Grant, Eric Enns, Eric Marinier, Arnab Mandal, Emily K Herman, Chih-yu Chen, Morag Graham, Gary Van Domselaar ✉, Paul Stothard ✉ Author Notes

*Nucleic Acids Research*, gkad326, <https://doi.org/10.1093/nar/gkad326>

Published: 04 May 2023 Article history ▼





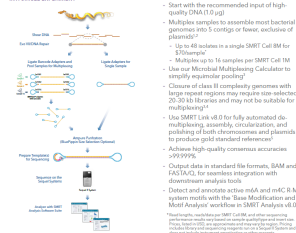
# Documentation

# Documentation

## MICROBIAL WHOLE GENOME SEQUENCING BEST PRACTICES

With Single Molecule, Real-Time (SMRT™) Sequencing and the Sequel™ Systems, you can affordably assemble reference-quality microbial genomes that are >99.99% (Q30) accurate.

### FROM GEOMIC DNA TO COMPLETE GENOME IN A SINGLE EXPERIMENT!



### WORKFLOW RECOMMENDATIONS

- Start with the recommended input of high-quality DNA (1.5 µg)
- Multiplex samples to assemble most bacterial genomes (10x coverage) or fewer numbers of plasmids?
- Up to 96 samples in a single SMRT Cell 1M for \$30/sample
- Use our Microbial Multiplexing Calculator to simplify experiment design!
- Closure of class II complexity genomes with large repeat regions may require size selection 20-30 kb libraries and may not be suitable for multiplexing!
- Use SMRT Link v4.0 for fully automated de-multiplexing, assembly, circularization, and polishing of both chromosomes and plasmids to produce high-accuracy references!
- Achieve high-quality consensus accuracies >99.99%
- Output data in standard file formats, BAM and FASTQ, for seamless integration with downstream analysis tools
- Diagnose and analyze error modes and SMRT cell system metrics with the Base Modification and Mist Analyzer software in SMRT Analysis v4.0!
- Highly accurate, reference-free SMRT Cell and sequencing data analysis tools are available in the SMRT Analysis v4.0 software.

### MICROBIAL ASSEMBLY PROCESS

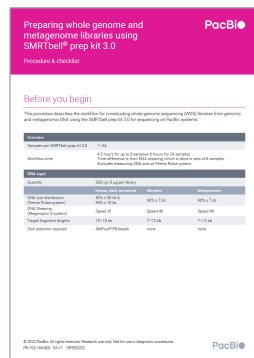


## Application Brief: Microbial whole genome sequencing – Best Practices (BP101-013020)

Summary overview of application-specific sample preparation and data analysis workflow recommendations

## Procedure & Checklist – Preparing whole genome and metagenome libraries using SMRTbell™ prep kit 3.0 (102-166-600)

Technical documentation containing sample library construction and sequencing preparation protocol details



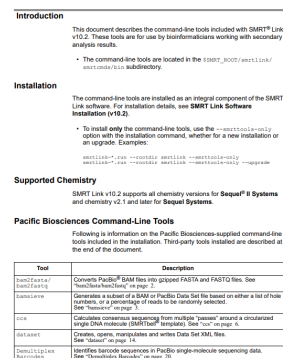
## SMRT Link User Guide – Sequel Systems (102-278-200)

Technical documentation describing how to use SMRT Link software. SMRT Link is the web-based end-to-end workflow manager for Sequel Systems.



## SMRT Tools Reference Guide (102-278-500)

Technical documentation describing command line tools included with SMRT Link. These tools are for use by bioinformaticians working with secondary analysis results.



Tool	Description
fastq2bam	Converts PacBio® BAM files into grouped FASTQ and FASTQ files. See <a href="#">fastq2bam</a> .
fastq2bam	Converts PacBio® BAM files into grouped FASTQ and FASTQ files. See <a href="#">fastq2bam</a> .
fastq2bam	Converts a subset of a BAM or PacBio Data Set file based on either a list of read numbers, or a percentage of reads to be randomly selected. See <a href="#">fastq2bam</a> .
fastq2bam	Calculates consensus sequences from multiple 'passes' around a circularized single DNA molecule (SMRTbell™). See <a href="#">fastq2bam</a> .
fastq2bam	Checks, opens, manipulates and writes Data Set XML files. See <a href="#">fastq2bam</a> .
fastq2bam	Identifies contigs in PacBio single-molecule sequencing data. See <a href="#">fastq2bam</a> .



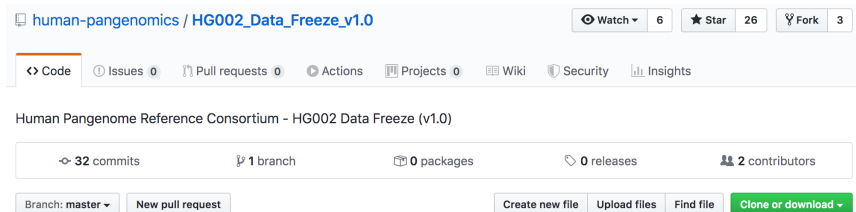
# Public data

# Public HiFi data

## HG002 Human Pan-Genome Reference Consortium

4 cells: 2 cells 20kb and 2 cells 15kbp

- ~34x coverage
- [https://github.com/human-pangenomics/HG002\\_Data\\_Freeze\\_v1.0](https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0)



human-pangenomics / HG002\_Data\_Freeze\_v1.0

Watch 6 Star 26 Fork 3

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights

Human Pangenome Reference Consortium - HG002 Data Freeze (v1.0)

32 commits 1 branch 0 packages 0 releases 2 contributors

Branch: master New pull request

Create new file Upload files Find file Clone or download

### Sequencing Data

The annotated table of sequence data can be downloaded [here](#).

#### HG002 Data Freeze (v1.0) Recommended downsampled data mix

We encourage assembly groups to use as much of the data from the HG002 freeze as possible to get the best assembly they can. However, as no two groups are likely to use exactly the same subset of data, making comparison more difficult, and the size and variety of the HG002 freeze is not representative of what is likely to be available in future freezes, we recommend that assembly groups also run their pipeline on the following set of 4 downsampled datasets from the HG002 (NA24385) human cell line:

##### PacBio HiFi:

~34X coverage of Sequel II System with Chemistry 2.0

15kb:

- [https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp\\_HG002\\_NA24385\\_son\\_v1/PacBio\\_HiFi/15kb/m64012\\_190920\\_173625.Q20.fastq](https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/15kb/m64012_190920_173625.Q20.fastq)
- [https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp\\_HG002\\_NA24385\\_son\\_v1/PacBio\\_HiFi/15kb/m64012\\_190921\\_234837.Q20.fastq](https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/15kb/m64012_190921_234837.Q20.fastq)

20kb:

- [https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp\\_HG002\\_NA24385\\_son\\_v1/PacBio\\_HiFi/20kb/m64011\\_190830\\_220126.Q20.fastq](https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/20kb/m64011_190830_220126.Q20.fastq)
- [https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp\\_HG002\\_NA24385\\_son\\_v1/PacBio\\_HiFi/20kb/m64011\\_190901\\_095311.Q20.fastq](https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/20kb/m64011_190901_095311.Q20.fastq)



# Public HiFi data

## CHM13 data from the HiCanu preprint

- 5 HiFi datasets
- <https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA530776>

- [WGS of CHM13 with PacBio CCS](#)
  - 1 PACBIO\_SMRT (Sequel II) run: 1M spots, 21G bases, 15.7Gb downloads  
Accession: SRX7897688
- [WGS of CHM13 with PacBio CCS](#)
  - 1 PACBIO\_SMRT (Sequel II) run: 1.4M spots, 28.7G bases, 21.7Gb downloads  
Accession: SRX7897687
- [WGS of CHM13 with PacBio CCS](#)
  - 1 PACBIO\_SMRT (Sequel II) run: 1.6M spots, 25.6G bases, 16.3Gb downloads  
Accession: SRX7897686
- [WGS of CHM13 with PacBio CCS](#)
  - 1 PACBIO\_SMRT (Sequel II) run: 1.6M spots, 25.1G bases, 16Gb downloads  
Accession: SRX7897685
- [WGS of CHM13 with PacBio CCS](#)
  - 4 PACBIO\_SMRT (Sequel II) runs: 6.9M spots, 75.6G bases, 47.3Gb downloads  
Accession: SRX5633451



bioRxiv  
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS | CHANNELS

Search  [Advanced Search](#)

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

[Comment on this paper](#)

[Previous](#)

[Next](#)

**HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads**

Sergey Nurk, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, Sergey Koren

doi: <https://doi.org/10.1101/2020.03.14.992248>

This article is a preprint and has not been certified by peer review [what does this mean?]

[Abstract](#) [Full Text](#) [Info/History](#) [Metrics](#) [Preview PDF](#)

Posted March 19, 2020.

**Download PDF**

Supplementary Material

Data/Code

XML

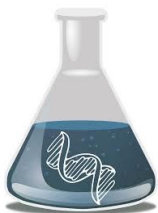
Revision Summary

Email

Share

Citation Tools

## Public HiFi data



### HG002

15 kb + 20 kb library

6 SMRT Cell 8M

[Data: PRJNA586863](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA586863)

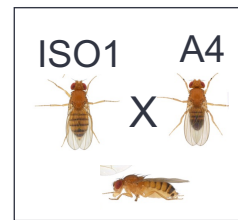


### *Oryza sativa indica* MH63

17 kb + 24 kb library

2 SMRT Cell 8M

[Data: PRJNA573706](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA573706)



### *Drosophila melanogaster* F1

19 kb + 24 kb library

2 SMRT Cell 8M

[Data: PRJNA573706](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA573706)

# Public HiFi data












**bioRxiv**

THE PREPRINT SERVER FOR BIOLOGY

New Results

## Highly accurate long-read HiFi sequencing data for five complex genomes

Ting Hon, Kristin Mars, Greg Young,  Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin,  Nicholas Maurer,  David Kudrna, Michael A. Hardigan,  Cynthia C. Steiner,  Steven J. Knapp,  Doreen Ware,  Beth Shapiro,  Paul Peluso,  David R Rank

doi: <https://doi.org/10.1101/2020.05.04.077180>



The background of the slide is a blurred laboratory setting. In the foreground, a multi-well plate is visible, with several wells containing a bright pink liquid. A pipette tip is positioned above one of the wells, with a single drop of the pink liquid about to fall. The word "PacBio" is overlaid in a bold, pink, sans-serif font in the upper right quadrant of the image.

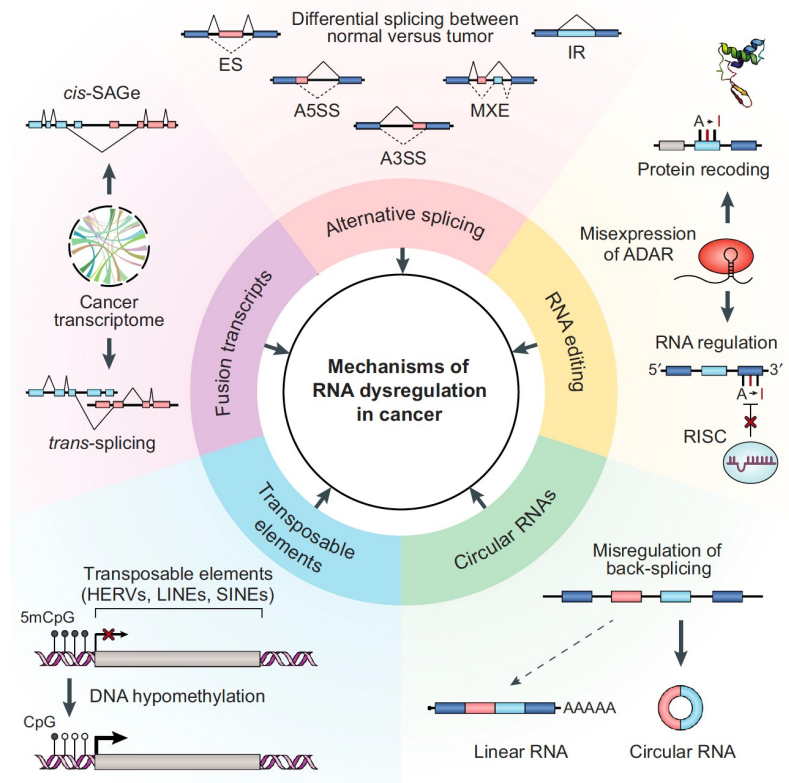
**PacBio**

# Iso-Seq analysis application overview

27 June 2023

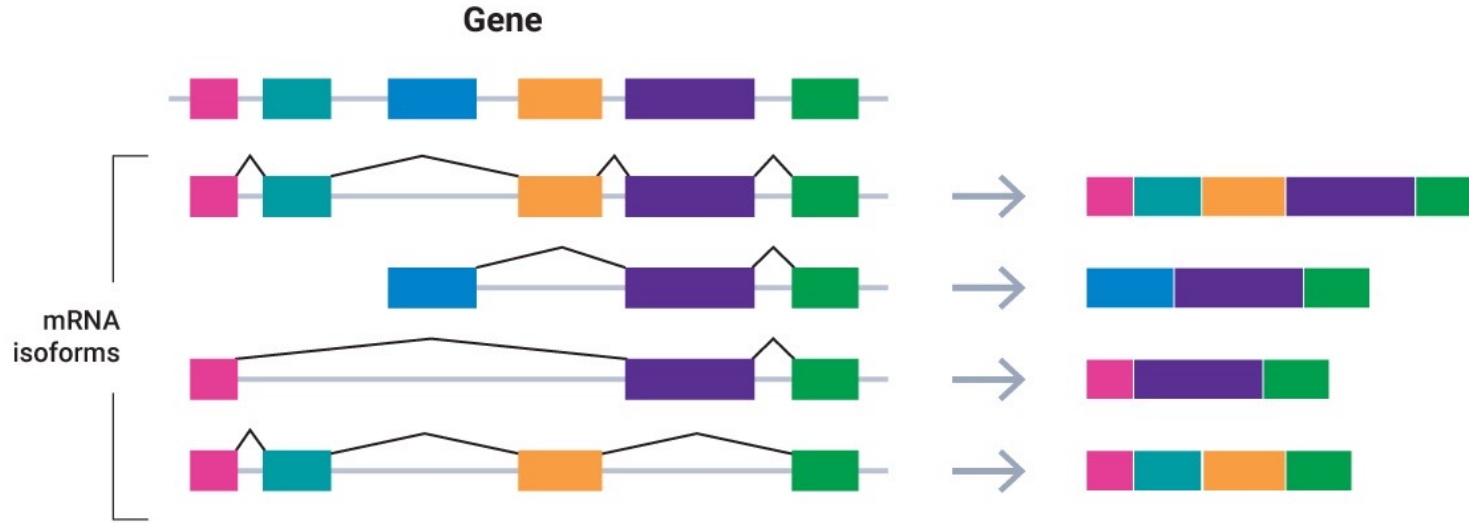
彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

# RNA dysregulation in cancer



Trends in Pharmacological Sciences

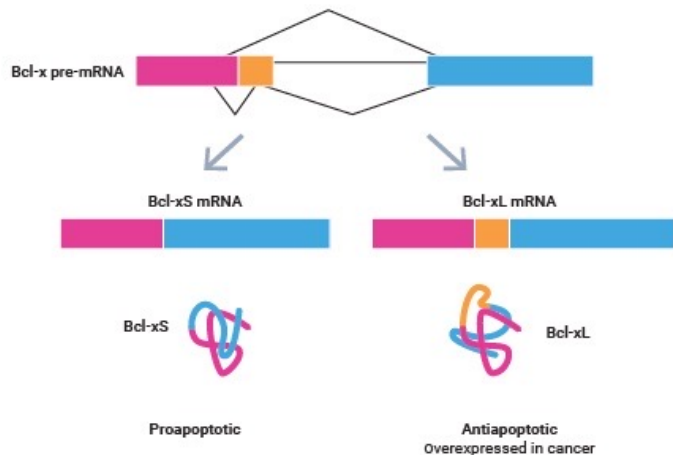
# Alternative splicing is fundamental in producing diversity of proteins in the human body



- >95% of genes undergo alternative splicing
- Cells express ~4 isoforms per gene

# Cancers display aberrant isoform regulation

## Splicing events that affect protein function



**Example:** The Bcl-x protein has two isoforms

- Bcl-xS (short isoform) is expressed in normal cells
- Bcl-xL (long isoform) is overexpressed in cancer cells, promotes cancer progression and resistance to chemotherapy.

## Cancer-specific switches in isoform expression

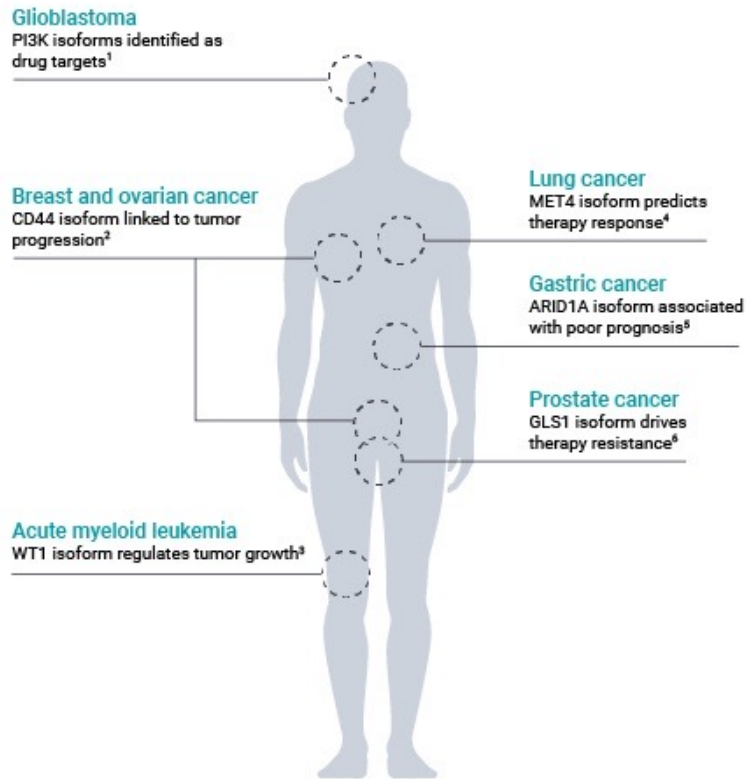


Alternative transcripts are more dominantly expressed in cancer tissues than in normal tissues.

# Isoforms are important source of novel cancer **biomarkers** and **drug targets**

Cancer-specific isoforms have been shown to be actionable **biomarkers** and are an untapped source of **drug targets** for oncology.

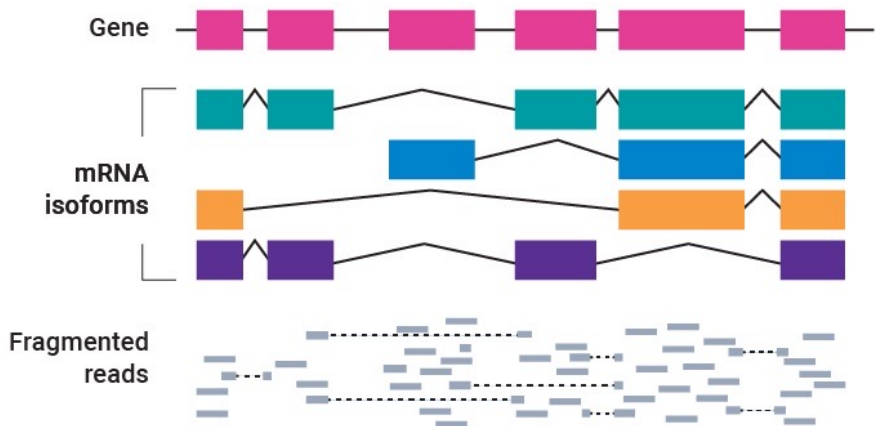
The vast majority of cancer-specific isoforms **remain unknown**.





# Long-read sequencing provides complete view of cancer transcriptome

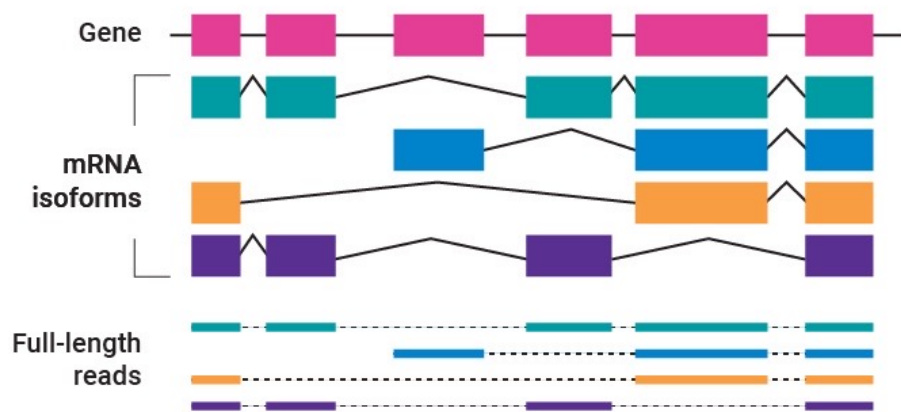
## Short read sequencing



Short-read sequencing can only assemble ~20 to 40% of human transcriptomes

**PARTIAL** view of cancer transcriptomes

## Long read sequencing



PacBio's long-read sequencing offers superior **isoform discovery power**

**COMPLETE** view of cancer transcriptomes

# Iso-seq identifies thousands of **novel isoforms** in breast cancer samples

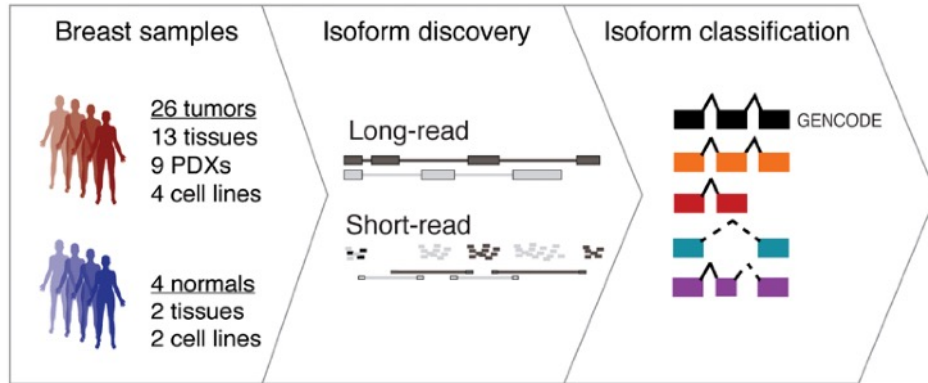
SCIENCE ADVANCES | RESEARCH ARTICLE

CANCER

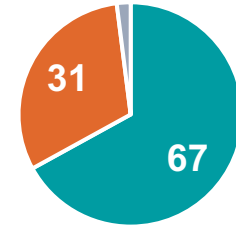
## A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer

Diogo F. T. Veiga<sup>1†</sup>, Alex Nesta<sup>1,2†</sup>, Yuqi Zhao<sup>1</sup>, Anne Deslattes Mays<sup>1</sup>, Richie Huynh<sup>1</sup>, Robert Rossi<sup>1</sup>, Te-Chia Wu<sup>1</sup>, Karolina Palucka<sup>1</sup>, Olga Anczukow<sup>1,2,3\*</sup>, Christine R. Beck<sup>1,2,3\*</sup>, Jacques Banchemareau<sup>1\*</sup>

Veiga et al., *Sci. Adv.* 8, eabg6711 (2022) 19 January 2022



142,514 splice isoforms detected



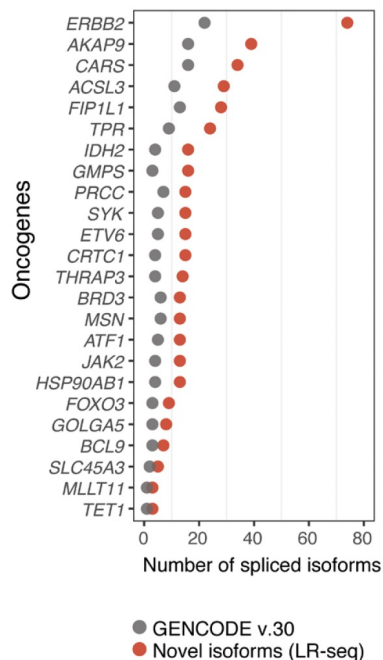
- Novel isoforms
- Known isoforms
- Other (antisense, intronic)

Two-third of identified isoforms are **novel** (NNC+NIC)

Proportion of novel isoforms is ~2-fold higher in tumor vs normal samples.

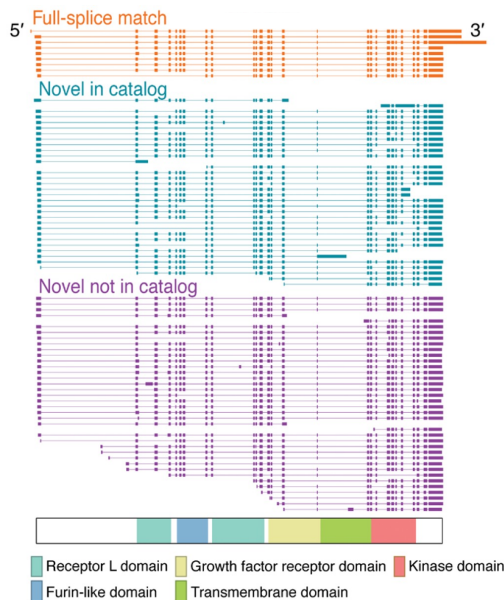
# Alternative splicing affects important functional domains in **oncogenes**

## Novel isoform increase in oncogenes



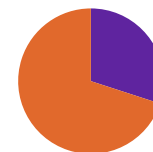
## Complex splicing regulation in

### *ERBB2/HER2* oncogene

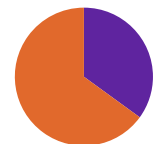


## Effects on important functional domains

Percent of novel isoforms with **affected conserved domains**



Percent of novel isoforms with predicted **protein localization effects**



# Example – gastric cancer

Huang et al. *Genome Biology* (2021) 22:44  
<https://doi.org/10.1186/s13059-021-02261-x>

Genome Biology

RESEARCH

Open Access

## Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer



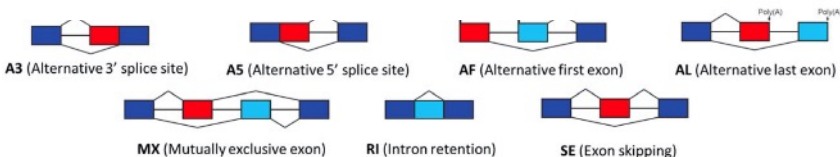
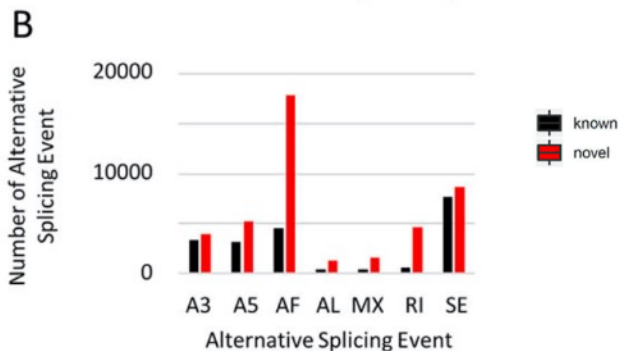
Kie Kyon Huang<sup>1</sup>, Jawen Huang<sup>1</sup>, Jeanie Kar Leng Wu<sup>1</sup>, Minghui Lee<sup>1</sup>, Su Ting Tay<sup>1</sup>, Vikrant Kumar<sup>1</sup>, Kapana Ramnarayanan<sup>1</sup>, Nisha Padmanabhan<sup>1</sup>, Chang Xu<sup>1</sup>, Angie Lay Keng Tan<sup>1</sup>, Charlene Chan<sup>2</sup>, Dennis Kappel<sup>1,3</sup>, Jonathan Gölke<sup>4</sup> and Patrick Tan<sup>1,2,4,5\*</sup>

- Gastric cancer is the 3<sup>rd</sup> leading cause of cancer death
- Tumor morphology gives limited guidance
- Molecular methods (sequencing) can better help subtype GC

# Example – gastric cancer

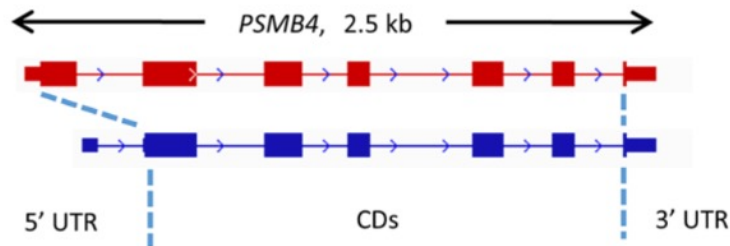
>60% Iso-Seq transcripts are novel

Majority of novelty comes from use of an alternative first exon (AF)

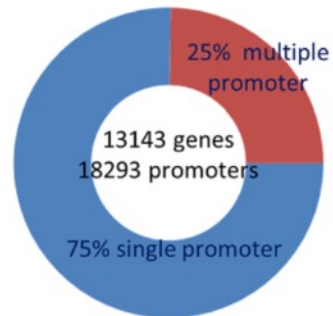


Alternative promoters change CDS

AFs can change the encoded protein



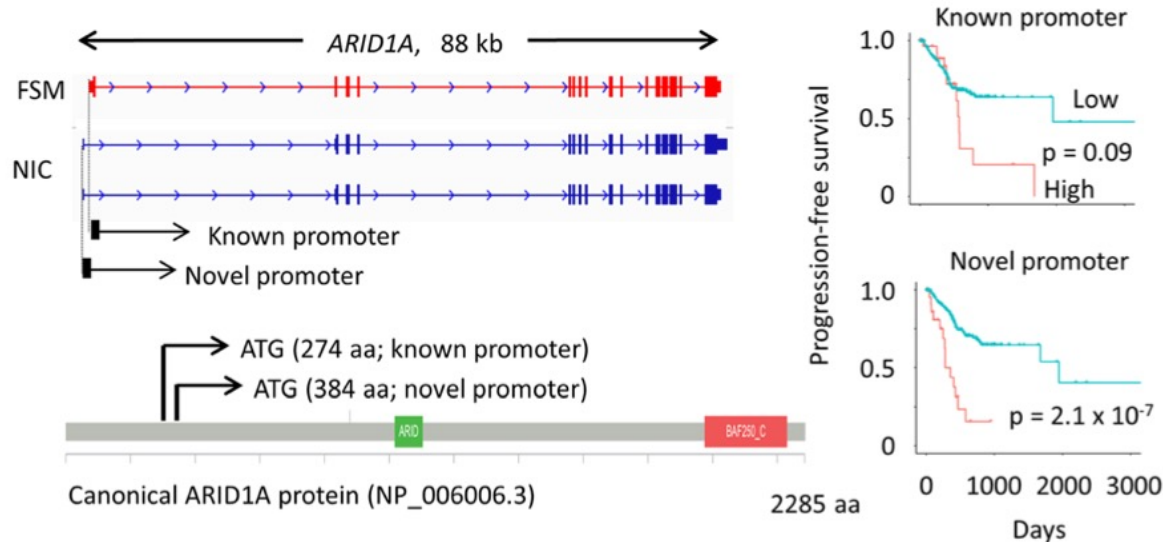
~25% genes have multiple promoters



# Linking novel promoters to potential clinical outcomes

## *Iso-Seq data identifies novel promoter in ARID1A*

Two novel (NIC) transcripts use a novel promoter that truncates the first 384 aa; it is associated with poor survival outcome. In contrast, the known (FSM) transcript uses a known promoter and is not significantly associated with poor survival.



# Advantages of the Iso-Seq method

The Iso-Seq method generates highly accurate full-length reads for bulk and single-cell transcriptome

- ✓ No transcript assembly required

---

- ✓ Full-length isoforms from 5' to 3' end

---

- ✓ Can be used for
  - genome annotation
  - novel isoform discovery
  - fusion gene finding
  - differential isoform expression analysis

# Iso-Seq workflow is an end-to-end solution



Library prep  
**1 DAY**



SMRT sequencing  
**1 DAY**



Data analysis  
**1 DAY**

**LONG-READ RNA SEQUENCING BEST PRACTICES**

**FROM RNA TO FULL-LENGTH TRANSCRIPTS**

- Reverse Transcription
- Template Switching
- Smear Switch (Optional)
- PCR Amplification
- DNA Damage Repair / End Repair / A-tailing
- Adapter Ligation
- Sequence on the Sequel System

**WORKFLOW RECOMMENDATIONS**

- Prepare full-length cDNA from 300 ng of total RNA using the NEBNext<sup>®</sup> Single Cell/Low Input cDNA Synthesis & Amplification Modules kit
- Use the SMART-Seq2 Express Template Prep Kit 2.0 to prepare libraries in one day
- Multiplex up to 12 samples
- Scale throughput on Sequel Systems
- Use the Sequel II System to generate up to 4 million full-length, non-chimeric FLNC reads per SMRT Cell 8M
- Or use the Sequel System to generate up to 500,000 FLNC reads per SMRT Cell 1M

**DETERMINATION OF TRANSCRIPT ISOFORMS**

Full length cDNA Sequence Reads  
Splice Isoform Curation – No Assembly Required

The Iso-Seq method allows you to produce evidence-based genome annotations, discover novel genes and isoforms, and improve 3'UTR and alternative splicing isoform annotations.

**WITH THE PACBIO ANALYTICAL PORTFOLIO**

FLNC reads are analyzed using the Iso-Seq software pipeline to output high-quality, full-length transcript FASTA sequences, with no assembly and a reference genome, and annotate the genome using community tools such as

**PLACING EVENTS IN SPECIFIC GENES**

The Iso-Seq method enables detection of complex alternative splicing of the syntenic alpha-CDNA gene using IsoSeq analysis

**TRANSCRIPTOME**

Full length cDNA sequence reads and detected 30,445 novel genes

With a single SMRT Cell 8M you can:

- Characterize a whole transcriptome
- Multiplex multiple tissues for genome annotation

[www.pacb.com/iso-seq](http://www.pacb.com/iso-seq)



# What can you get with 1 SMRT cell worth of Iso-Seq data?

SMRT sequencing

**1 DAY**



	FL reads	Unique genes	Unique transcripts
UHRR '19	4,734,362	16,328	183,689
Alzheimer Brain	4,277,293	17,670	162,290

- ✓ 1 SMRT Cells 8M yields up to 4 million full-length reads
- ✓ Each full-length read is a full-length transcript – no assembly required
- ✓ In human whole transcriptome, >100k unique isoforms can be observed

# Isoform characterization: A meal of its own



## PacBio genome + Iso-Seq



1 PacBio genome  
1 RNA tissue



1 PacBio genome  
1 individual



## PacBio Iso-Seq only



Existing ref genome  
3 tissues



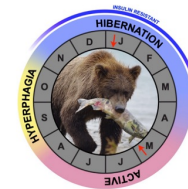
Existing ref genome  
19 samples total



Existing ref genome  
3 tissue x 3 conditions



Existing ref genome  
2 yeast strains



Existing ref genome  
3 bears x 3 tissues  
x 2 conditions

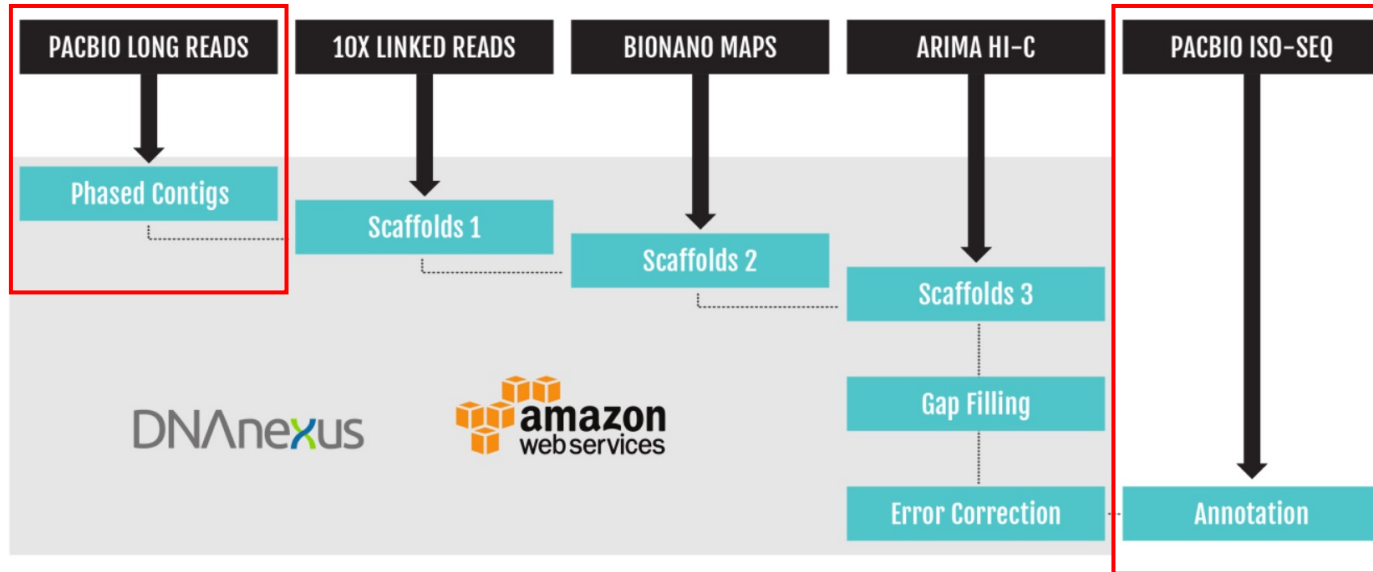


Existing ref genome  
10 rice cultivar RNA

# PacBio is the main technique of the VGP project

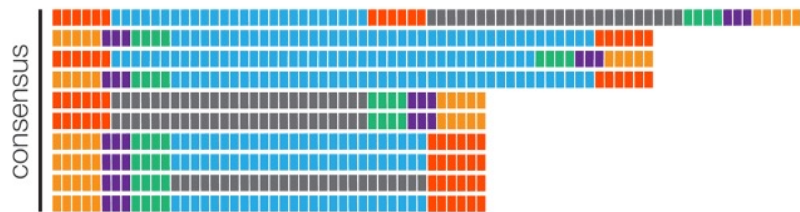


The analysis pipeline of the Vertebrate genomes projects (VGP)

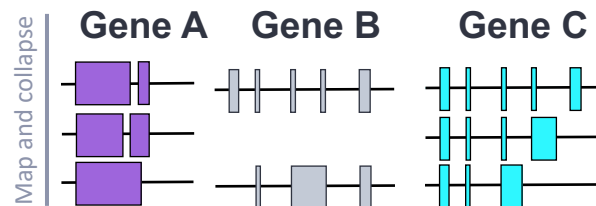


# Iso-Seq Analysis

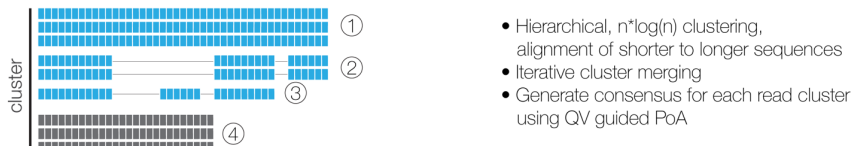
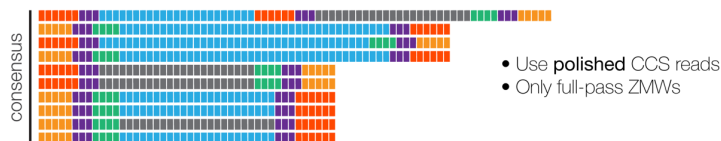
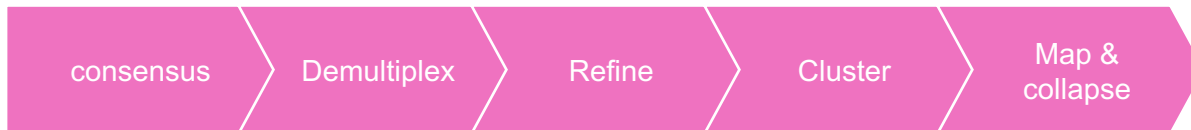
Transforms HiFi reads



Into high-quality, full-length isoforms



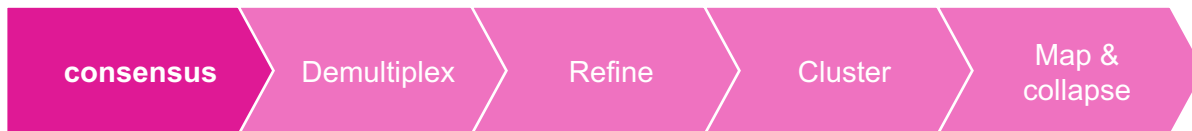
# Iso-Seq Analysis mid-level workflow



Transforms HiFi reads  
Into high-quality, full-length isoforms

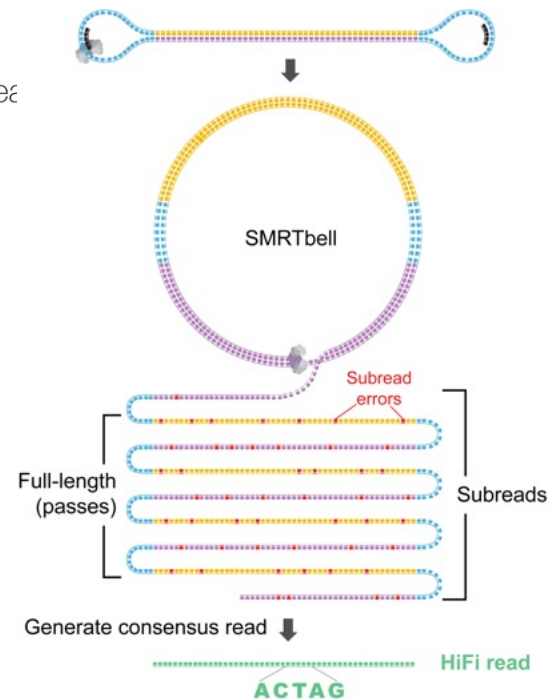


# Iso-Seq Analysis workflow

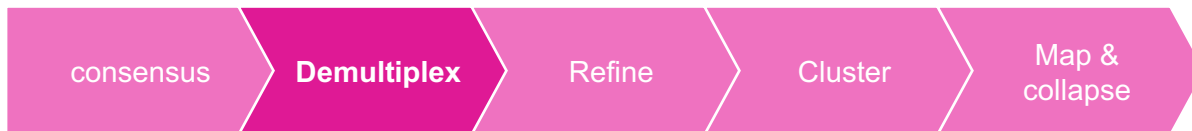


- Use **polished** CCS reads
- Only full-pass ZMWs

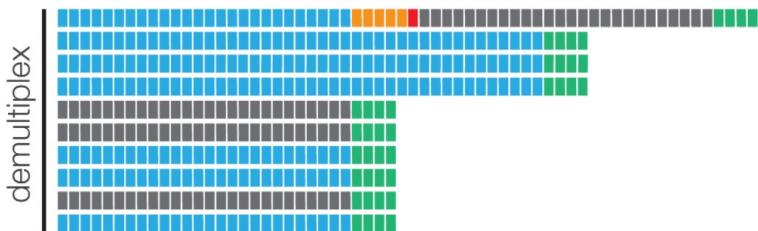
Iso-Seq analyzes **QV20** reads from hifi\_reads.bam.



# Iso-Seq Analysis workflow



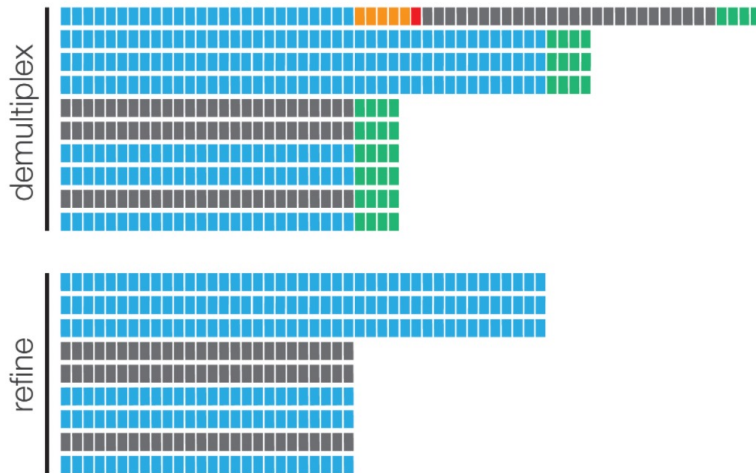
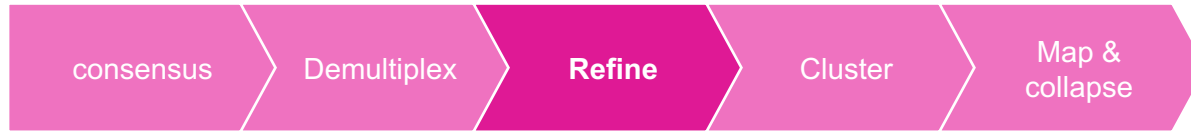
- Use **polished** CCS reads
- Only full-pass ZMWs



- Barcoded and unbarcoded cDNA primer removal
- Orientation
- Unwanted primer combination removal

Utilizes [demultiplex barcoding algorithm \(LIMA\)](#) with special `--isoseq` mode  
Full length reads have 5' and 3' cDNA primers, which are removed by LIMA  
DON'T USE DEMULTIPLEX BARCODES APPLICATION

# Iso-Seq Analysis workflow



- Barcoded and unbarcoded cDNA primer removal
- Orientation
- Unwanted primer combination removal

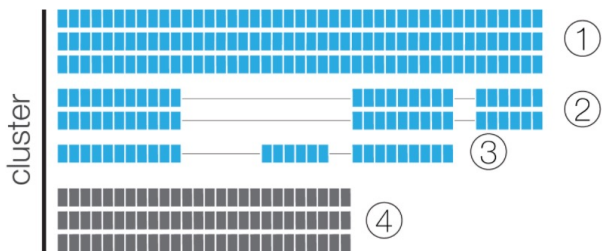
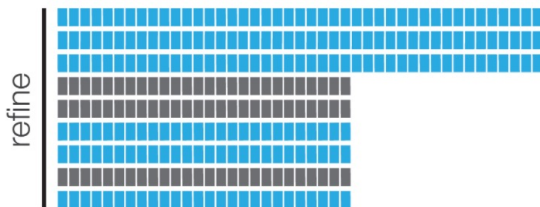
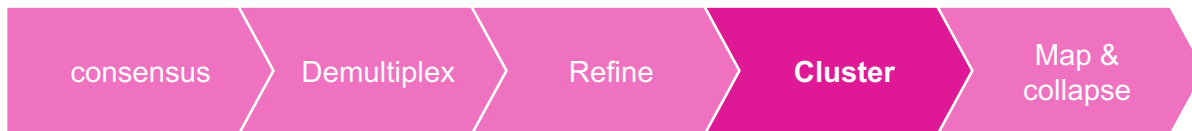
- PolyA tail trimming
- Concatemer removal

FLNC reads: CCS reads with 5' and 3' cDNA primers, polyA tail, and concatemers removed

If your sample has poly(A) tails, use `--require-polya` to filter for FL reads that have a poly(A) tail with at least 20 base pairs and remove the identified tail (GUI - turn off on advanced parameters )



# Iso-Seq Analysis workflow



- High Quality (HQ):

QV $\geq$ 20 and  $\geq$ 2 FLNC read support

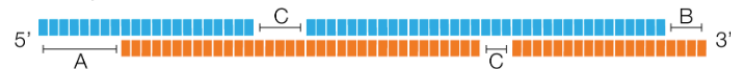
- Low Quality (LQ):

QV <99% and < 2 FLNC read support

- PolyA tail trimming
- Concatemer removal

- Hierarchical,  $n \cdot \log(n)$  clustering, alignment of shorter to longer sequences
- Iterative cluster merging
- Generate consensus for each read cluster using QV guided PoA

Similar transcripts:



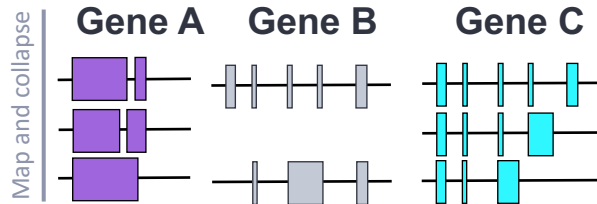
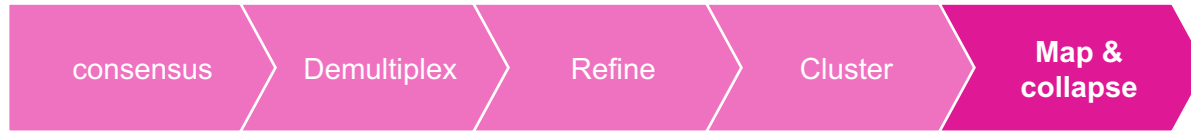
Two FLNC reads are considered the same isoform if:

A) <100 bp difference in 5' start

B) <30 bp difference in 3' end

C) <10 bp in internal gap, no limit on number of gaps

# Iso-Seq Analysis workflow



- Align to reference genome
- Remove redundancy
- pbmm2

# Iso-Seq terminology

NAME	ABBR	EXPLANATION
Full-Length Reads	FL Reads	CCS reads with 5' and 3' cDNA primers removed
Full-Length, Non-Concatemer Reads	FLNC Reads	CCS reads with 5' and 3' cDNA primers, polyA tail, and concatemers removed
High-Quality Isoforms	HQ Isoforms	Polished transcript sequences with predicted accuracy $\geq 99\%$ & $\geq 2$ FLNC
Low-Quality Isoforms	LQ Isoforms	Polished transcript sequences with predicted accuracy $< 99\%$ & $\geq 2$ FLNC

## Primer/barcode set required – demultiplexing step

### Primer Set :

Specify a primer sequence file in FASTA format to identify cDNA primers for removal. The primer sequence includes the 5' and 3' cDNA primers and (if applicable) barcodes.

Primer IDs must be specified using the suffix\_5p to indicate 5' cDNA primers and the suffix\_3p to indicate 3' cDNA primers. The 3' cDNA primer should not include the Ts and is written in reverse complement

Each primer sequence must be unique.

### Multiplex Samples:

For multiplexed datasets, **Iso-Seq Analysis reports and graphs now include isoforms per barcode in addition to the total number of isoforms across all barcodes.**

If barcodes were used, they should be included.

# Primer set required – demultiplex step- multiplex samples

## Barcoded Adapters

bc1001\_5p

CACATATCAGAGTGCGGCAATGAAGTCGCAGGGTTGGGG

>bc1002\_5p

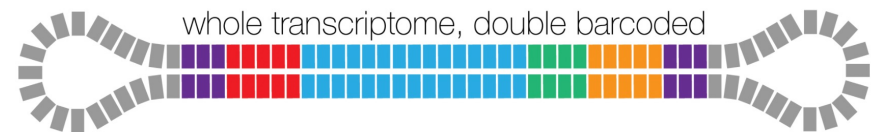
ACACACAGACTGTGAGGCAATGAAGTCGCAGGGTTGGGG

>bc1001\_3p

GTACTCTGCGTTGATACCACTGCTTCGCACTCTGATATGTG

>bc1002\_3p

GTACTCTGCGTTGATACCACTGCTTCTCACAGTCTGTGTGT





# Iso-Seq analysis artifacts

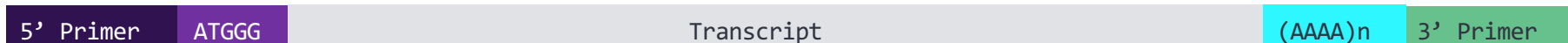
# Types of library artifacts

Type	Cause & phenotype	Detected By iso-seq analysis?
<b>TSO Artifacts</b>	C: TSO acting as a primer P: TSO on both ends	YES
<b>RT Artifacts</b>	C: Did not add TSO P: Missing TSO on 5' end	YES
<b>Artificial Concatemers</b>	C: Insufficient SMRT adapter P: cDNA primers in middle of read	YES
<b>PCR chimera</b>	C: PCR amplification P: Fusion of two transcripts	NO
<b>RT switching</b>	C: Secondary structure P: new intron with non-canonical junctions	NO (SQANTI can with mapping)
<b>Intrapriming</b>	C: dT priming off A-stretch P: genomic (A)s downstream of 3'	NO (SQANTI can with mapping)

# Types of library artifacts - representation

Full-Length (not artifact):

The "FL" in FLNC



RT Artifact:

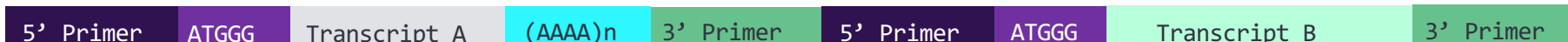


TSO Artifact:

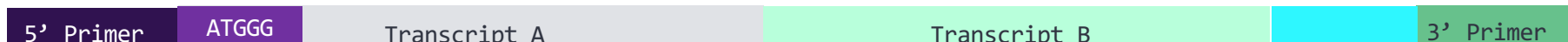


Artificial Concatemers:

The "NC" (non-artificial-concatemer) in FLNC



PCR Chimeras:





# Troubleshooting artifacts

Iso-Seq v3 will detect TSO & RT library artifacts and result in few FL reads

To determine which kind of artifacts it is:

1. Manually inspect the CCS fasta sequences
2. Inspect the LIMA report for hints

FILE: <job>/tasks/barcoding.tasks.lima-0/lima\_output.lima.summary

```
ZMWs input (A) : 486997
ZMWs above all thresholds (B) : 369009 (76%) # of FL
ZMWs below any threshold (C) : 117988 (24%) # of NFL

ZMW marginals for (C):
Below min length : 950 (1%)
Below min score : 0 (0%)
Below min end score : 37824 (32%)
Below min passes : 301 (0%)
Below min score lead : 0 (0%)
Below min ref span : 26830 (23%)
Without adapter : 301 (0%)
Undesired 5p--5p pairs : 16991 (14%)
Undesired 3p--3p pairs : 84015 (71%)
Undesired no hit : 301 (0%)
```

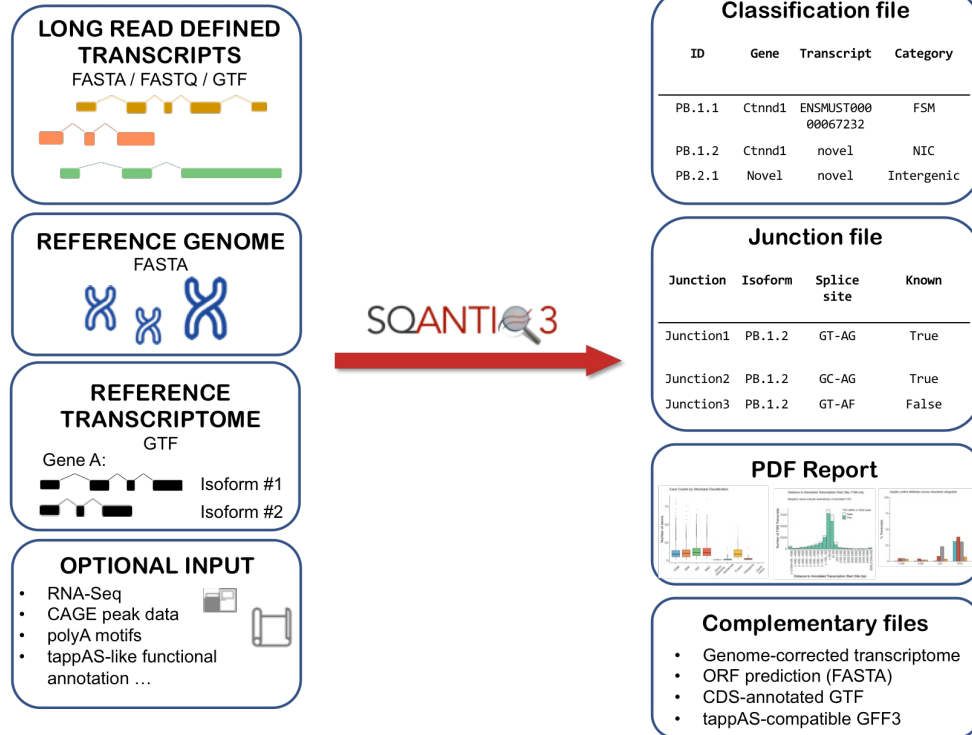
Reasons for nFL (Same read can have multiple reasons; sum can be over 100%)



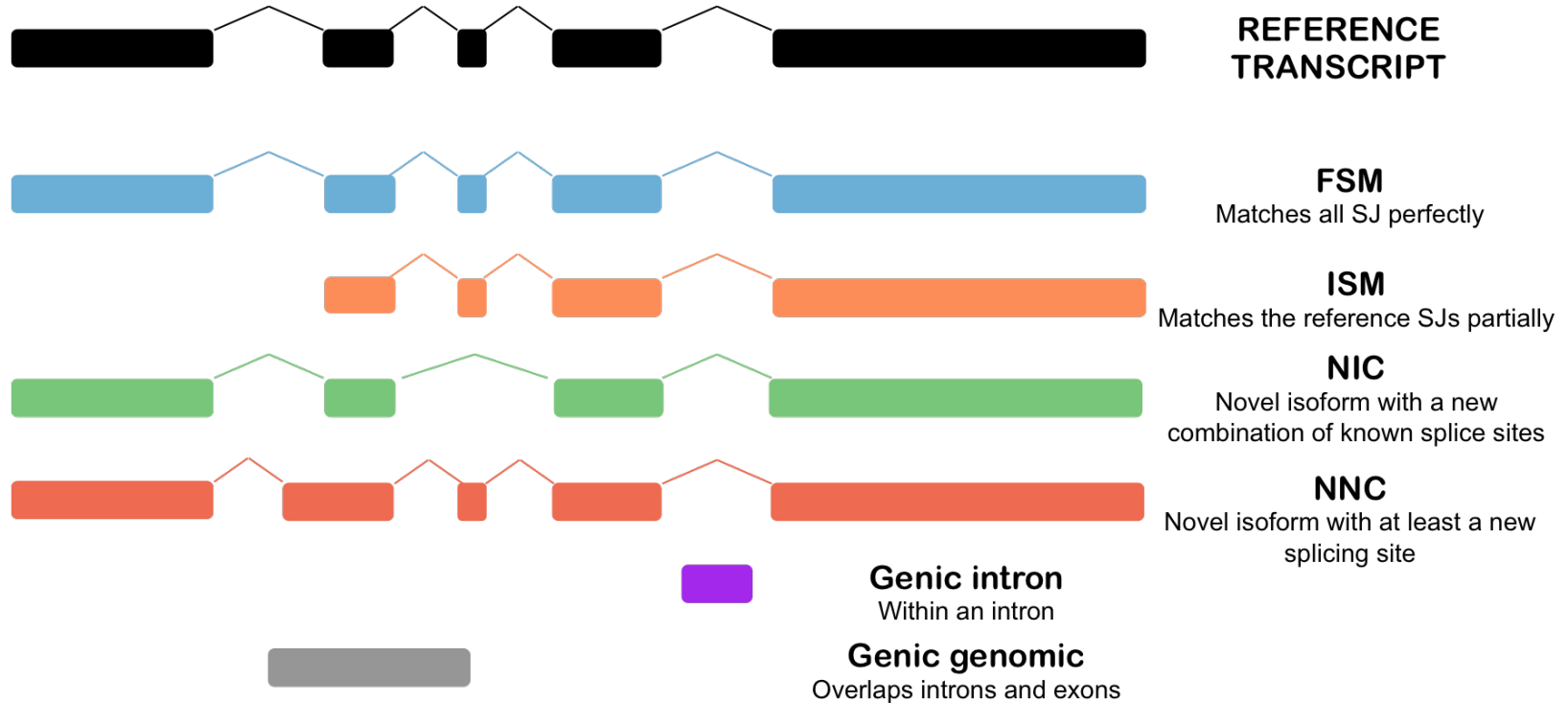
# Downstream analysis

Third party tools

# SQANTI3: Quality control transcriptomes

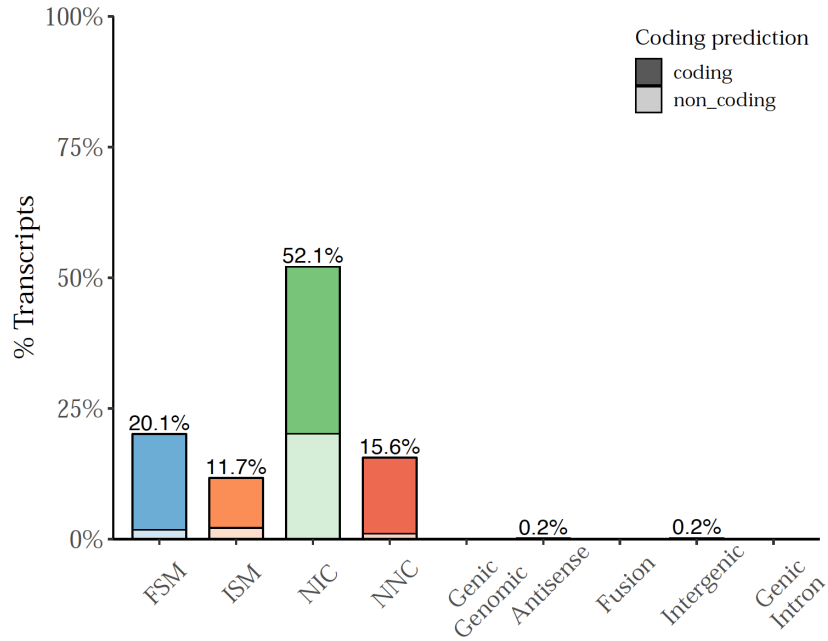


# SQANTI 3: isoform categorization

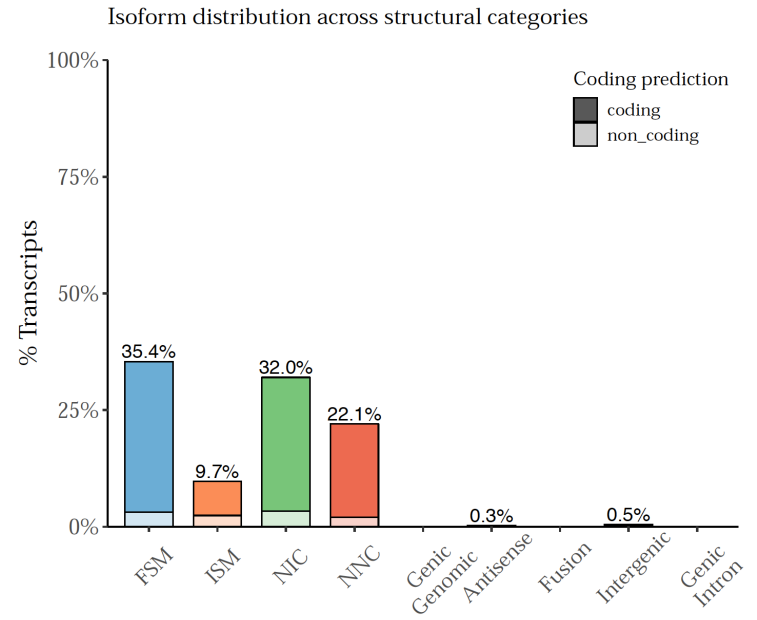


# Transcript distribution is highly sample-dependent

## Bulk Iso-Seq, Alzheimer brain



## Bulk Iso-Seq, UHRR



The PacBio logo is displayed in a bold, pink, sans-serif font. To the right of the text, a pink pipette tip is shown dripping a single drop of pink liquid. The background of the slide is a blurred laboratory setting with a rack of test tubes containing pink liquid.

PacBio

# PacBio HiFi Sequencing for High-Resolution Microbiome Research

27 June 2023

彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

# HiFi sequencing delivers the most comprehensive and highest quality data for microbial genomics

## Full-length 16S sequencing

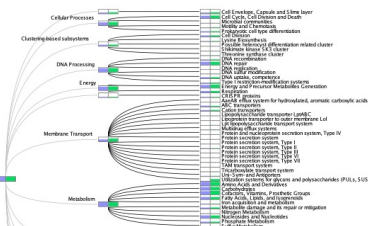


V4 % missing V1-V9

Species/strain-level resolution

Reveals true sample diversity

## Shotgun metagenome profiling

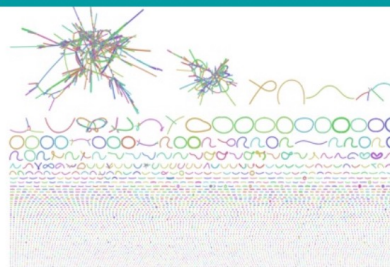


~8 genes per sequence read

~90% of reads with at least 1 gene

Profile taxonomy with high precision and recall

## Shotgun metagenome assembly



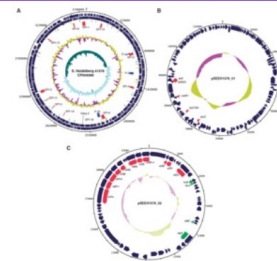
Obtain many high-quality MAGs

Complete, closed genomes

Resolves closely related strains

Short-read polishing (hybrid assembly) not needed

## Microbial whole genome sequencing



Obtain single contig chromosomes for most bacteria

Consensus accuracies >99.99%

Detection of R-M system motifs

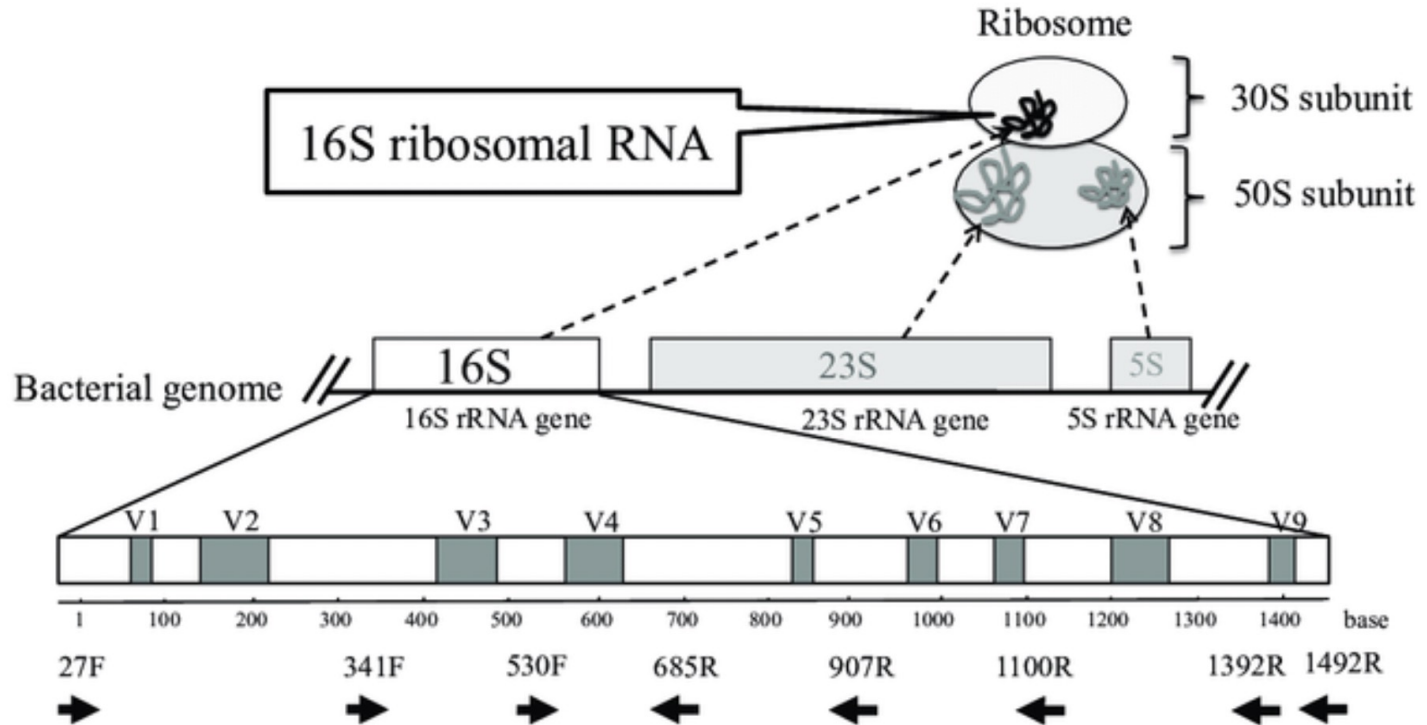
Short-read polishing (hybrid assembly) not needed



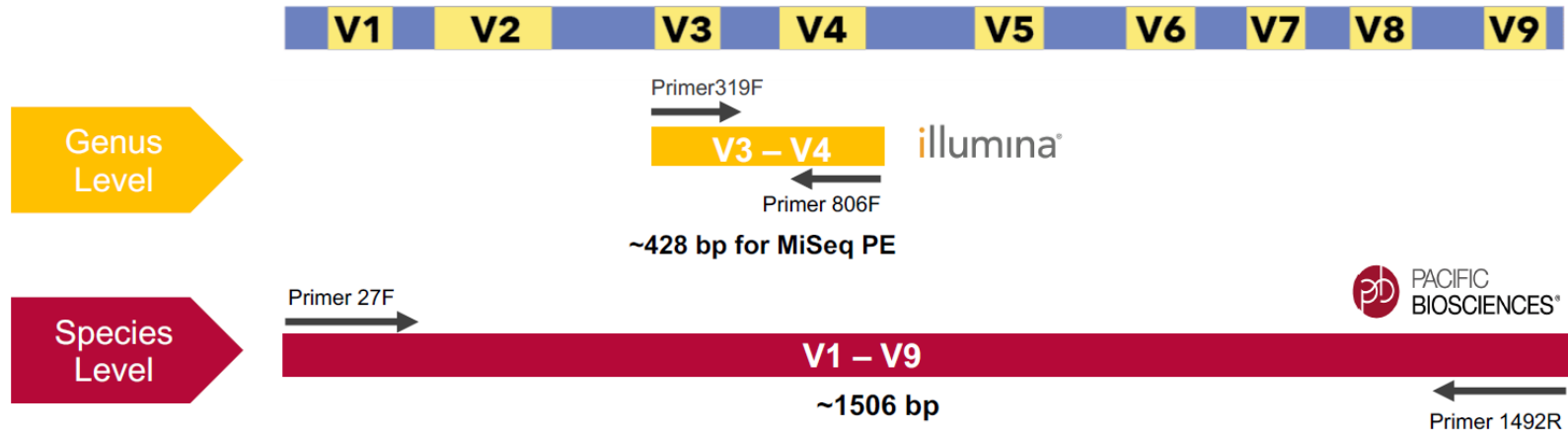
# Full-Length 16S Pipeline Overview



# 16s rRNA sequencing is a culture-free method to identify and compare bacterial diversity from complex microbiomes or environments



# Amplicons can Target 16s rRNA and Beyond



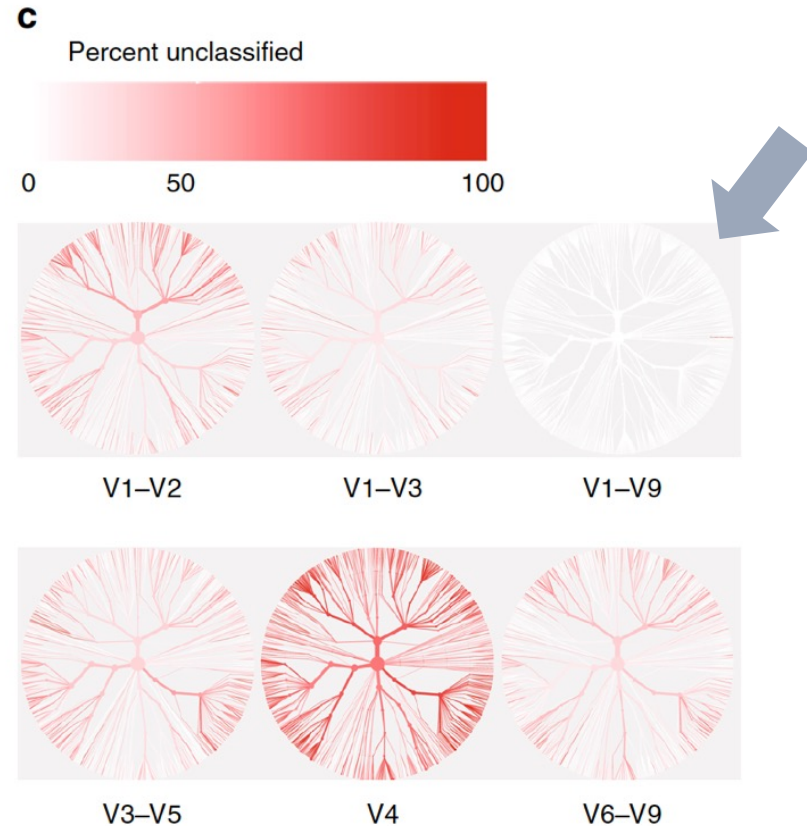
Longer amplicons enable higher resolution taxonomic identification

# Full Length 16S is crucial for complete taxonomy characterization in the human gut, without bias

**Full V1–V9 region: the only way to resolve ALL the clades that may be present in the human gut**

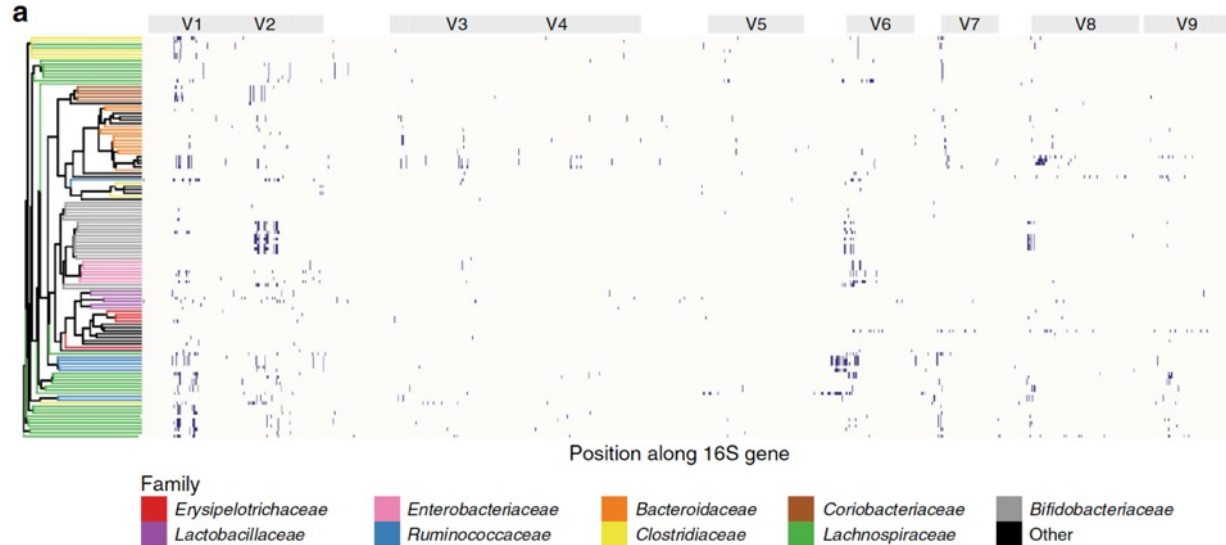
- V4: Consistently poor performance
- V1–V2: poor for Proteobacteria
- V3–V5: poor for Actinobacteria
- V1–V3: good results for *Escherichia* / *Shigella*
- V3–V5: good results for *Klebsiella*,
- V6–V9: good results for *Clostridium* and *Staphylococcus*

Proteobacteria 變形菌門 Actinobacteria 放線菌門 *Escherichia* 埃希氏菌屬 *Shigella* 志賀氏菌屬  
*Klebsiella* 克雷伯氏菌屬 *Clostridium* 梭菌屬 *Staphylococcus* 葡萄球菌屬



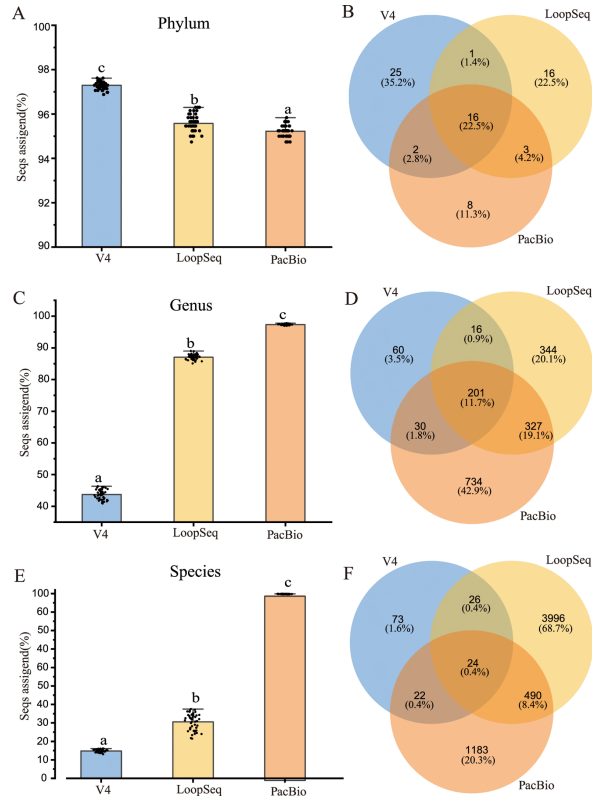
# Intragenomic 16S polymorphisms are highly prevalent in the human gut

“Multiple polymorphic 16S copies are not an inconvenience to be overlooked, rather they will enable the 16S gene to be used in strain-level microbiome analysis”



349 of 381 cultured isolates from the gut microbiome of the healthy individuals have intragenomic SNPs

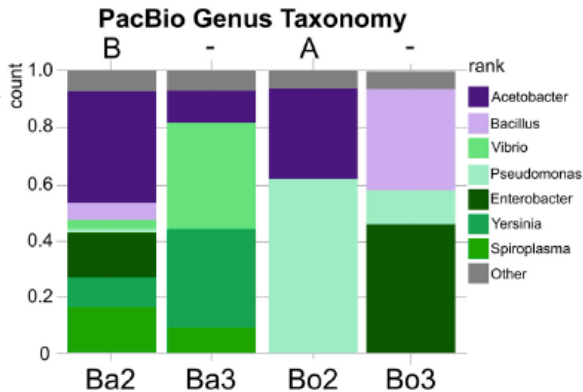
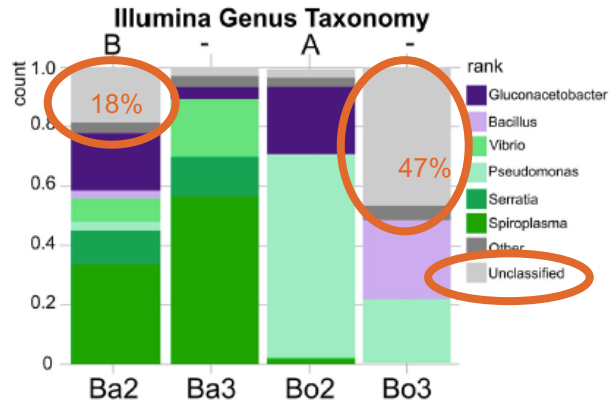
# More reads are classified to species and genus level with PacBio full-length 16S sequencing compared to V4 and synthetic long-read full-length 16S



**97%** sequences assigned to **genus level** using PacBio vs. LoopSeq at 87% and V4 at 44%

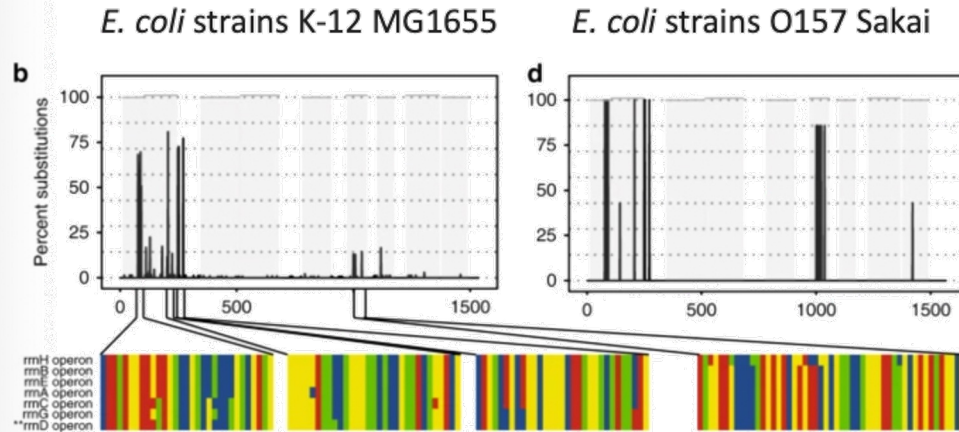
**99.7%** sequences assigned to **species level** using PacBio vs. LoopSeq at ~31% and V4 at ~15%

# Full-length 16S sequencing yields more in-depth taxonomic classification than short read V4



- “Our data show that Illumina data analysis provides a less effective taxonomic profiling than PacBio data analysis, with a high percentage of unclassified reads at the order and genus levels.”
- “Ultimately, while Illumina short-read sequencing is effective in microbiome analysis at a higher taxonomic level, long-read analyses show much greater power for in-depth classification.”

# HiFi reads can identify 16S polymorphisms



Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581 (2016).

[CAS](#) [PubMed Central](#) [Article](#) [PubMed](#)  
[Google Scholar](#)

Edgar R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Preprint at *bioRxiv* <https://doi.org/10.1101/081257> (2016).

Eren, A. M. et al. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**, 968–979 (2015).

[CAS](#) [Article](#) [Google Scholar](#)

# 16S Data Analysis Workflow Recommendations

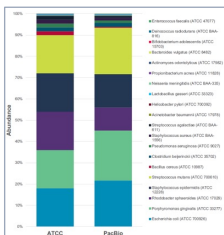
1 Generate HiFi reads



2 Demultiplex barcodes



3 Tertiary data analysis using either [DADA2](#) or [Qiime2](#)



1. **Perform CCS analysis** on-instrument (Sequel IIe system only) or in [SMRT Link](#) to generate highly accurate ( $\geq Q20$ ) single-molecule long reads (**HiFi reads**)

2. **Demultiplex barcodes** on-instrument (Sequel IIe system only) or in SMRT Link to separate HiFi reads by sample barcode

- Barcode FASTA files for demultiplexing can be downloaded from PacBio's [Multiplexing](#) website

3. Analyze 16S data using [DADA2](#) or [Qiime2](#)



- Open-source
- Well documented
- R package
- Easy and fast



An example HiFi read data set for a MSA-1003 mock community sample is available for download from PacBio ([Link](#))



<https://benjjneb.github.io/dada2/>

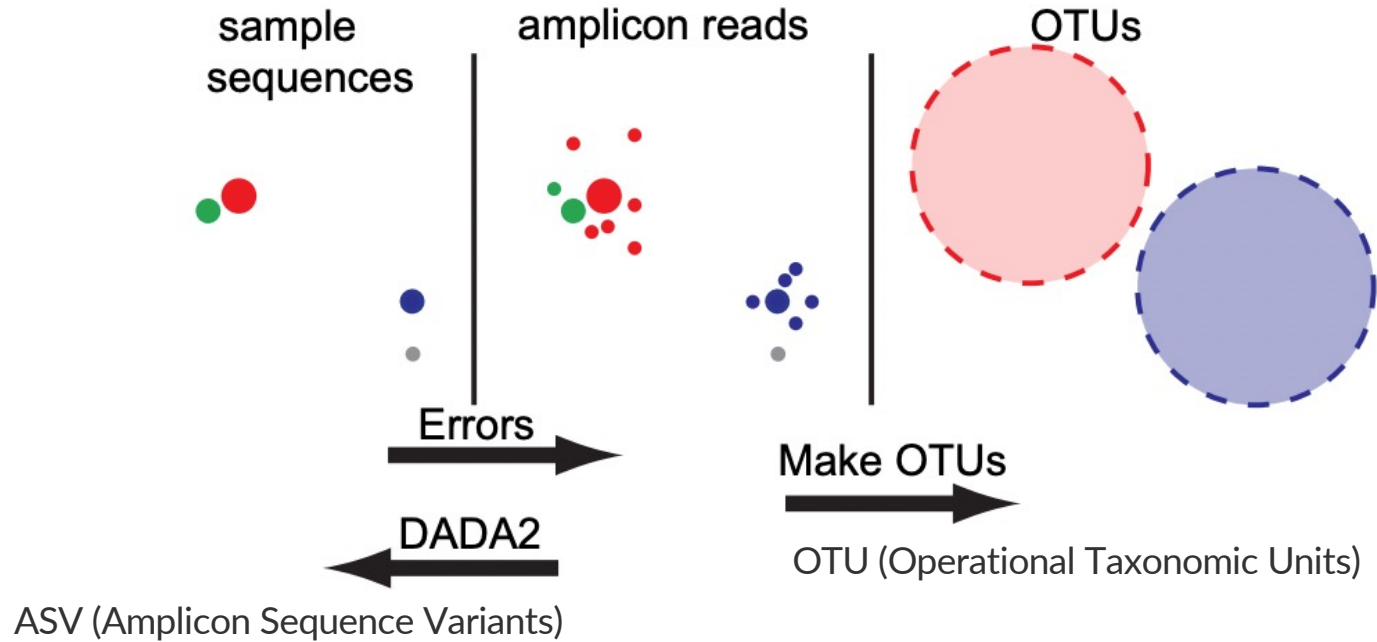
## High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution

Benjamin J Callahan , Joan Wong, Cheryl Heiner, Steve Oh, Casey M Theriot, Ajay S Gulati, Sarah K McGill, Michael K Dougherty

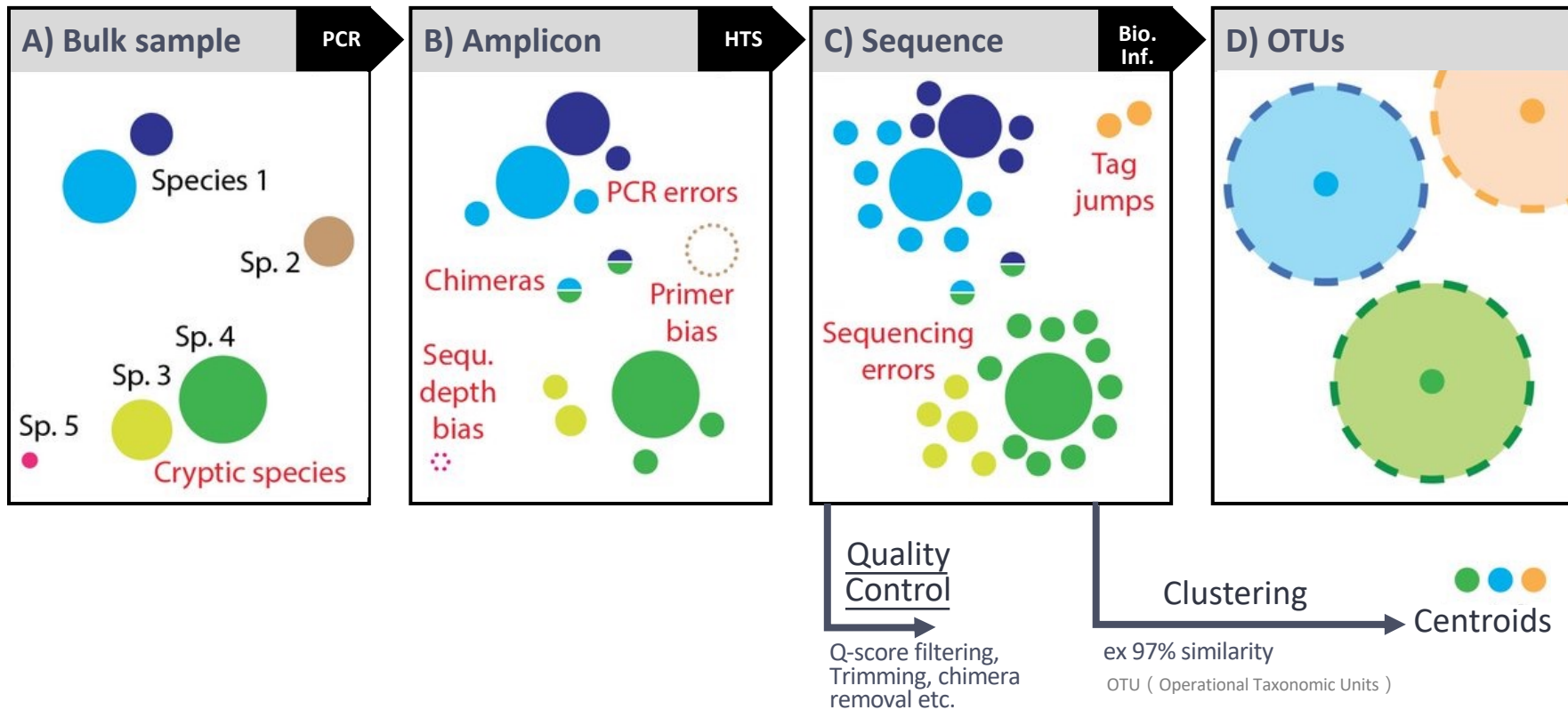
*Nucleic Acids Research*, Volume 47, Issue 18, 10 October 2019, Page e103,  
<https://doi.org/10.1093/nar/gkz569>

- PacBio CCS produced multiple distinct 16S sequences per bacterial genome
- 16S sequences appear in integer ratios that reflect their copy number

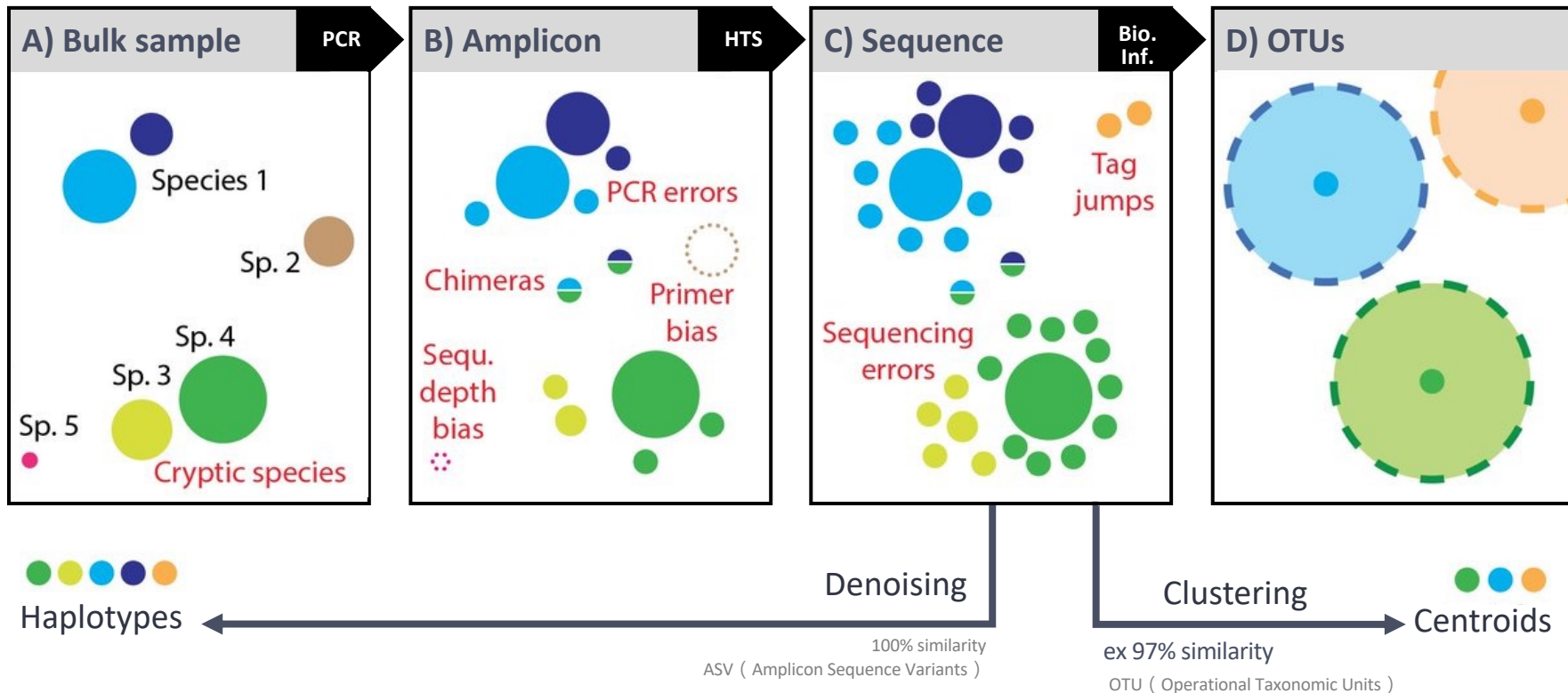
# Denoising and clustering



# Denoising and clustering



# Denoising and clustering



# Example workflow: 192-plex 16S amplicon library preparation using barcoded gene-specific primers

## MSA-1003 Mock Community Sample Description

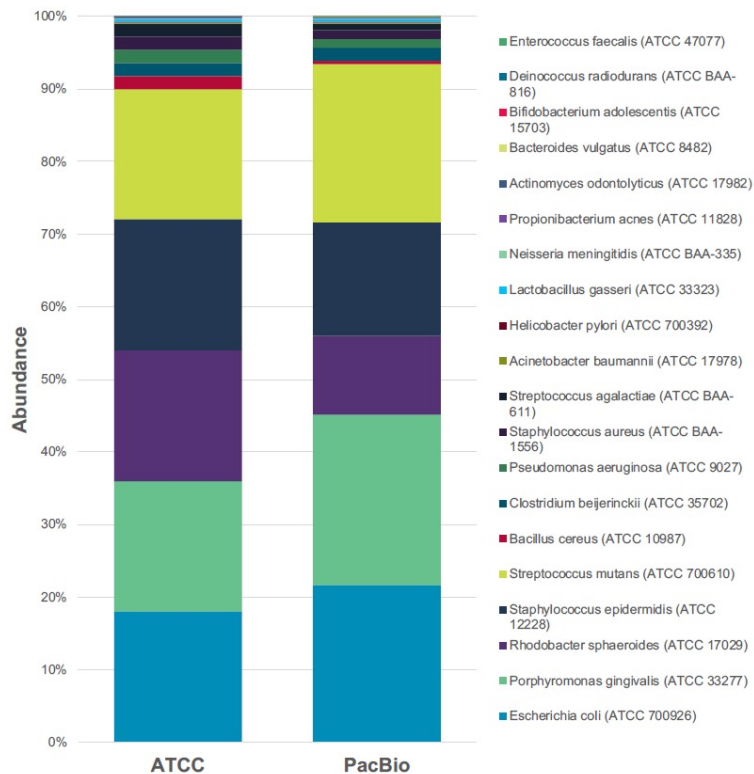
- MSA-1003 is a controlled, pre-defined, standardized reference material that can help with metagenomic analysis protocol development optimization, verification, and quality control
- 20 Strain Staggered Mix Genomic Material ([ATCC MSA-1003](https://www.atcc.org/products/all/MSA-1003))  
<https://www.atcc.org/products/all/MSA-1003.aspx>
- MSA-1003 sample is a mock microbial community that mimics mixed metagenomic samples
- MSA-1003 sample comprises genomic DNA prepared from fully sequenced, characterized, and authenticated ATCC Genuine Cultures that were selected by ATCC based on relevant phenotypic and genotypic attributes, such as Gram stain, GC content, genome size, and spore formation
- For the example data shown in this presentation, replicate MSA-1003 samples were processed in parallel to generate a 192-plex pooled 16S SMRTbell library using barcoded gene-specific primers and SMRTbell express template prep kit 2.0



%	MSA-1003 component
0.18	<i>Acinetobacter baumannii</i> (ATCC <a href="#">17978</a> )
1.80	<i>Bacillus cereus</i> (ATCC <a href="#">10987</a> )
0.02	<i>Bacteroides vulgatus</i> (ATCC <a href="#">8482</a> )
0.02	<i>Bifidobacterium adolescentis</i> (ATCC <a href="#">15703</a> )
1.80	<i>Clostridium beijerinckii</i> (ATCC <a href="#">35702</a> )
0.18	<i>Cutibacterium acnes</i> (ATCC <a href="#">11828</a> )
0.02	<i>Deinococcus radiodurans</i> (ATCC <a href="#">BAA-816</a> )
0.02	<i>Enterococcus faecalis</i> (ATCC <a href="#">47077</a> )
18.0	<i>Escherichia coli</i> (ATCC <a href="#">700926</a> )
0.18	<i>Helicobacter pylori</i> (ATCC <a href="#">700392</a> )
0.18	<i>Lactobacillus gasseri</i> (ATCC <a href="#">33323</a> )
0.18	<i>Neisseria meningitidis</i> (ATCC <a href="#">BAA-335</a> )
18.0	<i>Porphyromonas gingivalis</i> (ATCC <a href="#">33277</a> )
1.80	<i>Pseudomonas aeruginosa</i> (ATCC <a href="#">9027</a> )
18.0	<i>Rhodobacter sphaeroides</i> (ATCC <a href="#">17029</a> )
0.02	<i>Schaalia odontolytica</i> (ATCC <a href="#">17982</a> )
1.80	<i>Staphylococcus aureus</i> (ATCC <a href="#">BAA-1556</a> )
18.0	<i>Staphylococcus epidermidis</i> (ATCC <a href="#">12228</a> )
1.80	<i>Streptococcus agalactiae</i> (ATCC <a href="#">BAA-611</a> )
18.0	<i>Streptococcus mutans</i> (ATCC <a href="#">700610</a> )

# PacBio 16S Sequencing Faithfully Represents a Known Mock Community Sample

## 16S ANALYSIS OF THE MSA-1003 MOCK COMMUNITY



## MSA-1003 SAMPLE DESCRIPTION

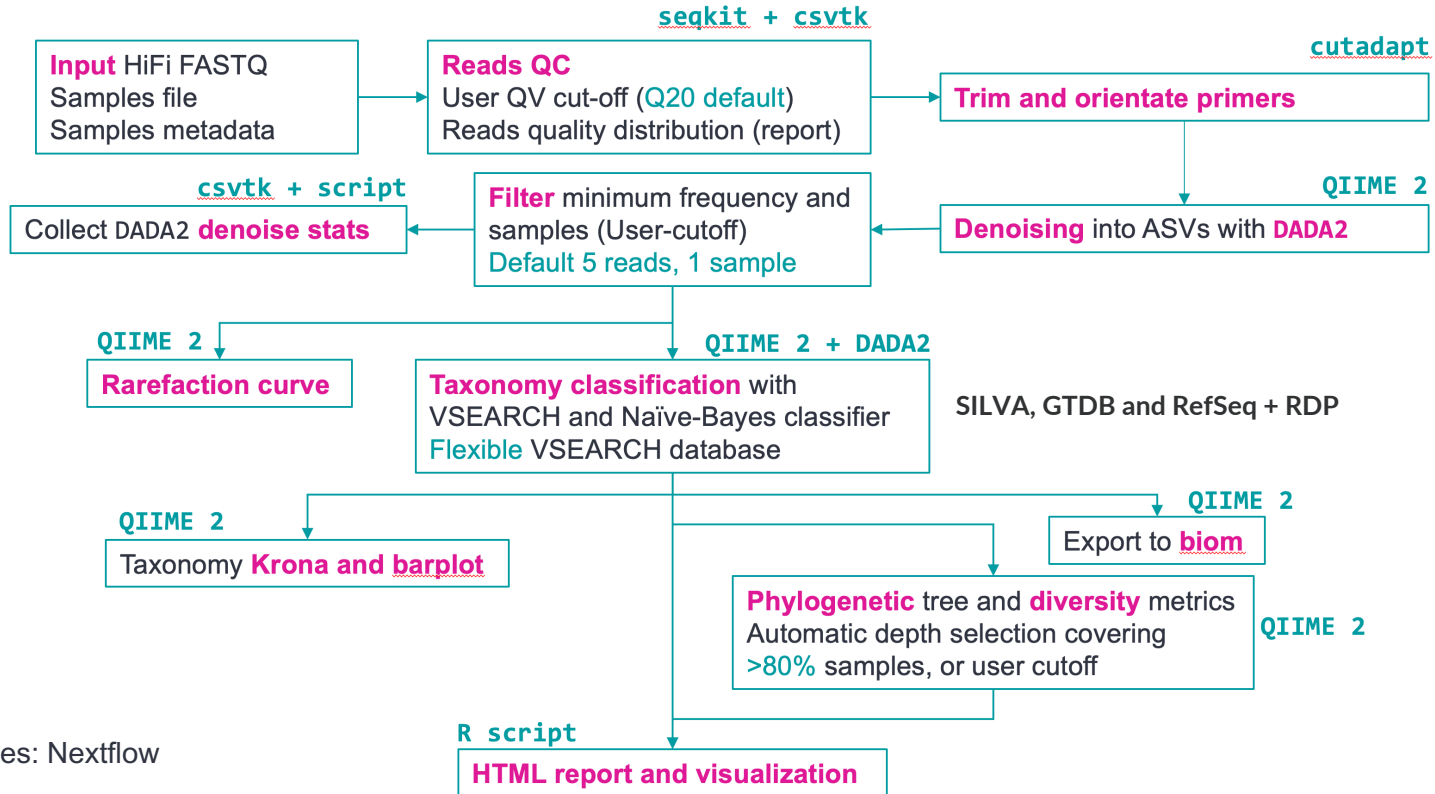
20 Strain Staggered Mix Genomic Material (ATCC® MSA-1003™)  
<https://www.atcc.org/products/all/MSA-1003.aspx>

**Yield of >99% accurate 16S reads matches the expected composition of the MSA-1003 mock community sample**

**GC content ranging from 30 ~ 69% can be identified**

[Download](#) and explore this 16S HiFi dataset further

# pb-16-nf overview



Languages: Nextflow

# Results from pb-16-nf pipeline

HTML report provides useful metrics and visualizations

Important outputs are in QIIME-compatible format and TSV format for easy importing

Outputs documentation:

[https://github.com/PacificBiosciences/pb-16S-nf/blob/main/pipeline\\_overview.md](https://github.com/PacificBiosciences/pb-16S-nf/blob/main/pipeline_overview.md)

DADA2 QC metrics								
Show	10	entries	Search:					
sample-id	input	filtered	percentage of input passed filter	denoised	non-chimeric	percentage of input non-chimeric	n_ASV	
All	All	All	All	All	All	All	All	
1 3VTVMF	151083	98855	65.43	96851	96678	63.99	465	
2 46EYMD	58454	38564	65.97	37230	37230	63.89	618	
3 4EH7JU	30231	19807	65.52	18775	18742	62	490	
4 4F747A	50845	33715	66.31	32909	32909	64.72	454	
5 4H9C6C	50973	34287	67.27	33034	33002	64.74	444	
6 4JAMMH	62883	41938	66.89	40797	40797	64.88	337	
7 4RHFPT	21373	14065	65.81	13788	13712	64.16	221	
8 4RNFPC	13929	9390	67.41	8566	8566	61.5	113	
9 4VMEN7	87957	57684	65.58	56576	56576	64.32	475	
10 63NDYT	121547	80036	65.85	78644	78636	64.7	508	

Showing 1 to 10 of 192 entries

Previous 1 2 3 4 5 ... 20 Next

## HiFi Full-length 16S Analysis Report

### Summary QC statistics

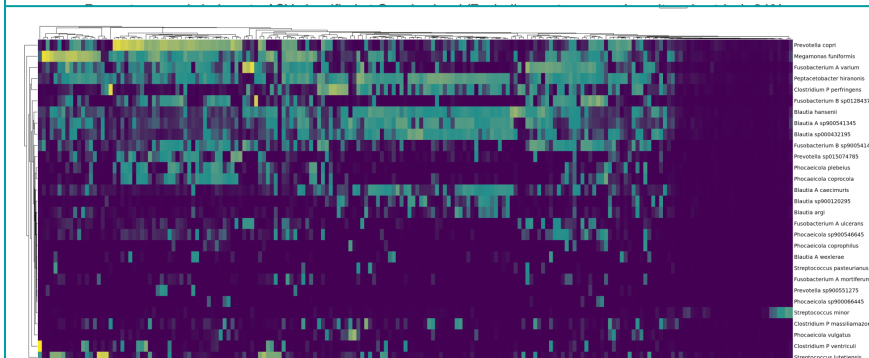
- Samples number: 192
- Final samples number post-DADA2: 192
- Missing samples (Not enough reads, do not pass QC, etc):
- Total number of CCS reads before filtering and primers trimming: 16777633
- Was primers trimmed prior to DADA2? Yes
- Total number of reads after quality filtering: 16472863 (98.18%)
- Total number of reads after primers trimming (DADA2 input): 16438413 (99.79%)
- Total number of ASVs found: 17293
- Average number of ASVs per sample: 361
- Total number of reads in 17293 ASVs: 10623342 (64.63% of all input reads)

### Classification using VSEARCH with a single database

- ASVs classified at Species level: 11646 (67.35%)
- ASVs classified at Species level (Excluding metagenome/uncultured entries): 11646 (67.35%)
- Percentage reads belong to ASV classified at Species level (Excluding metagenome/uncultured entries): 80%
- ASVs classified at Genus level: 11711 (67.72%)
- ASVs classified at Genus level (Excluding metagenome/uncultured entries): 11711 (67.72%)
- Percentage reads belong to ASV classified at Genus level (Excluding metagenome/uncultured entries): 81%

### Classification using Naive Bayes classifier with SILVA, GTDB and RefSeq + RDP

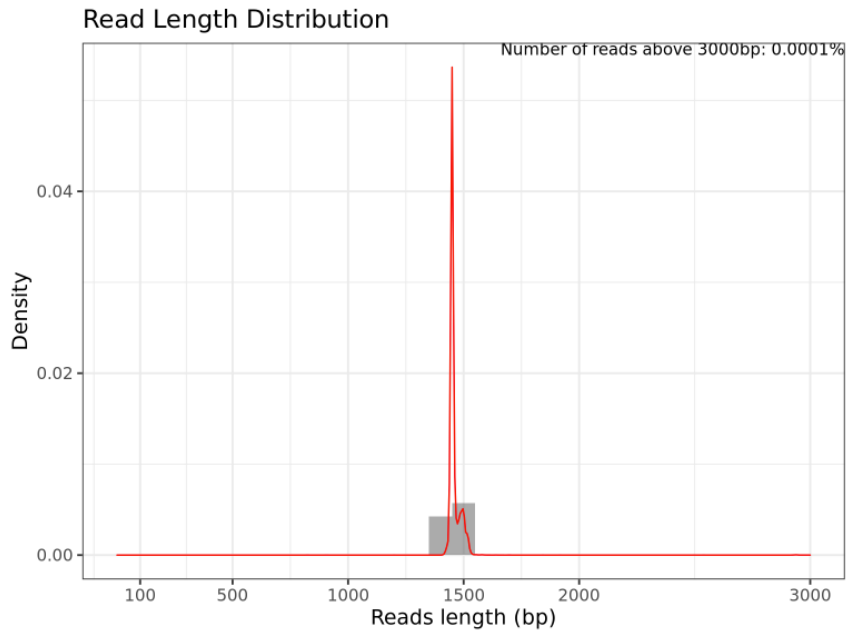
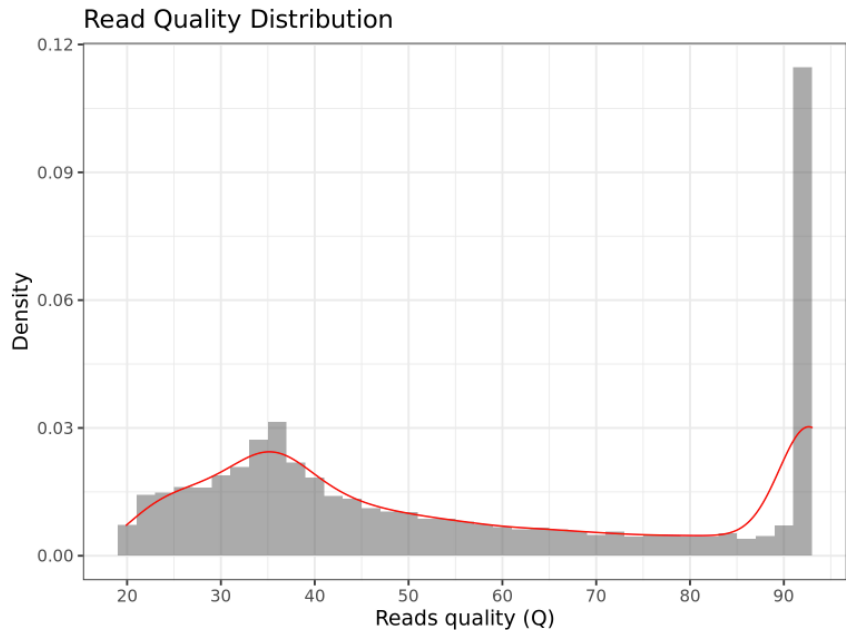
- ASVs classified at Species level: 13515 (78.15%)
- ASVs classified at Species level (Excluding metagenome/uncultured entries): 13515 (78.15%)





# Results from pb-16-nf pipeline

## Input reads QC (Before filtering and primers removal)





# Results from pb-16-nf pipeline



This interface can view .qza and .qzv files directly in your browser without uploading to a server. [Click here](#) to learn more.

Drag and drop or click here

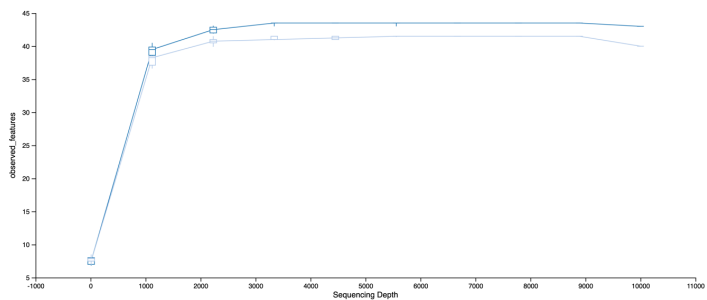
to view a QIIME 2 Artifact or Visualization (.qza/.qzv) from your computer.

You can also provide a link to a [file on Dropbox](#) or a [file from the web](#).

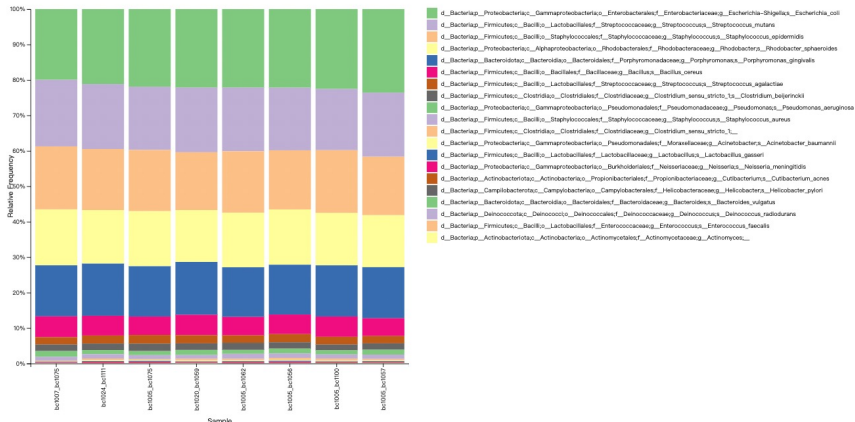
- > cutadapt\_summary
- > dada2
- > filtered\_input\_FASTQ
- > import\_qiime
- ▼ results
  - alpha-rarefaction-curves.qzv
  - best\_tax\_merged\_freq\_tax.tsv
  - best\_tax.qza
  - best\_taxonomy\_withDB.tsv
  - best\_taxonomy.tsv
  - dada2\_qc.tsv
  - dada2\_stats.qzv
  - dada2\_table.qzv
  - feature-table-tax\_vsearch.biom
  - feature-table-tax.biom
  - krona.qzv
  - merged\_freq\_tax.qzv
  - > phylogeny\_diversity
  - rarefaction\_depth\_suggested.txt
  - > reads\_QC
    - samplefile.txt
    - stats.tsv
  - > tax\_export
    - taxa\_barplot\_vsearch.qzv
    - taxa\_barplot.qzv
    - taxonomy.vsearch.qza
    - visualize\_biom.html
    - vsearch\_merged\_freq\_tax.tsv
- > summary\_demux
- > trimmed\_primers\_FASTQ

# Analysis PacBio HiFi Mock Community 16S Data with QIIME 2 view

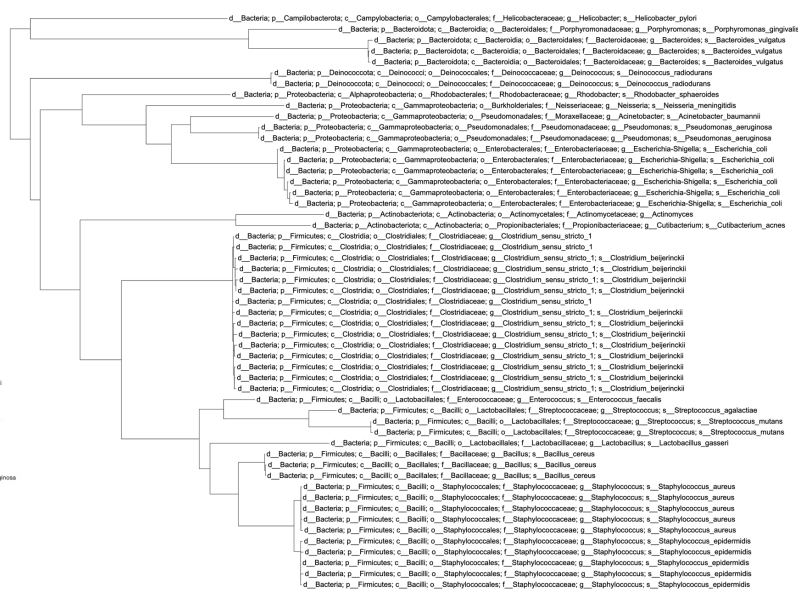
Rarefaction plot



Taxonomy bar plot



Phylogeny analysis



## How does it perform? (32 CPUs)

Sample types	Number of samples	Number of FL Q20 reads (FL%)	Total ASVs	Reads in ASVs	Classified species ASVs	Classified species reads	Pipeline run time	Pipeline max memory
Oral <sup>1</sup>	891	8.3m	5417	5104663 (62%)	<b>87%</b>	<b>91%</b>	2.5h	34 GB
Gut <sup>2</sup>	192	2.2m	1593	996965 (45%)	<b>96%</b>	<b>99%</b>	2h	30 GB
Animal gut <sup>3</sup>	192	16.7m	17293	10623342 (65%)	<b>67%*</b>	<b>81%</b>	13h	87 GB
Animal gut <sup>3</sup>	192	2.2m (99.3%)	10917	1789875 (83%)	<b>70%</b>	<b>79%</b>	5.5h	30 GB
Wastewater full <sup>4</sup>	33	2.14m	11462	1969683 (92%)	<b>39%*</b>	<b>63%</b>	12h	47 GB
Wastewater 10k/sample <sup>5</sup>	33	326k	3974	265137 (82%)	<b>44%*</b>	<b>65%</b>	4.6h	23 GB

\* Using MiDAS wastewater database increases classified species and reads to 85% for full dataset and 91% for down-sampled dataset

1. Data downloaded from SRA PRJDB12588, primers already trimmed.
2. Data downloaded from SRA PRJNA774819, primers already trimmed.
3. Customer collaboration dataset
4. Downloaded from SRA PRJNA846349, reads are Q30 filtered by author.
5. Downloaded from SRA PRJNA846349, reads are Q30 filtered by author. Down-sampled to 10k reads per sample.

# Mock Community HiFi Data available for download

- Full-length 16S Data Set

<https://github.com/PacificBiosciences/DevNet/wiki/16S-Data-Set-Sequel-II-System-2.0-Release>

## SAMPLE

20 Strain Staggered Mix Genomic Material (ATCC® MSA-1003™) <https://www.atcc.org/products/all/MSA-1003.aspx>

## METHODS

- 16S protocol with Barcoded Primers (<https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Full-Length-16S-Amplification-SMRTbell-Library-Preparation-and-Sequencing.pdf>)
- Library prep: SMRTbell Express Template Prep Kit 2.0
- Sequencing: Sequel II System binding kit (101-820-500) and chemistry (101-826-100)
- Run time: 0.5 hour pre-extension; 10 hour movie
- CCS Analysis: SMRT Link v10.0 Circular Consensus Sequencing Application (ccs 5.0.0)

## DOWNLOAD

Complete 192 plex dataset: [http://downloads.pacbccloud.com/public/dataset/atcc\\_msa/16S\\_192plex\\_HiFi.fastq.tar.gz](http://downloads.pacbccloud.com/public/dataset/atcc_msa/16S_192plex_HiFi.fastq.tar.gz)

- pb-16S-nf

<https://github.com/PacificBiosciences/pb-16S-nf>

# PacBio

 伯森生技

[www.pacb.com](http://www.pacb.com)

Research use only. Not for use in diagnostic procedures. © 2022 PacBio. All rights reserved.  
PacBio, the PacBio logo, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of PacBio.  
All other trademarks are the sole property of their respective owners.