PacBio

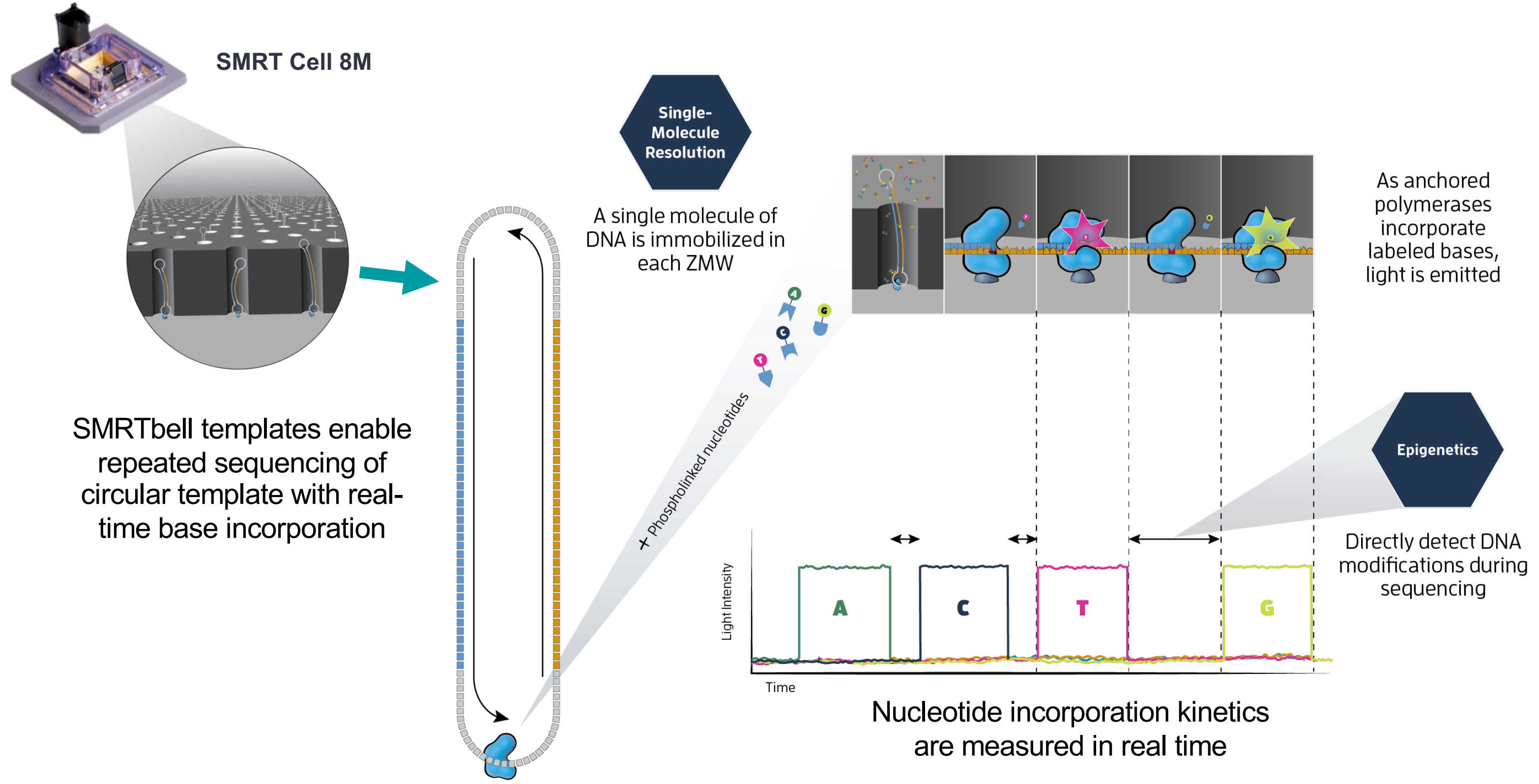
Unlock the promise of genomics through PacBio sequencing

Single Molecule Real-time Sequencing Analysis Overview

04 July 2023

彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

Single Molecule, Real-Time (SMRT) Sequencing



SMRT Cell 8M

Single-Molecule Resolution

A single molecule of DNA is immobilized in each ZMW

As anchored polymerases incorporate labeled bases, light is emitted

SMRTbell templates enable repeated sequencing of circular template with real-time base incorporation

Epigenetics

Directly detect DNA modifications during sequencing

Nucleotide incorporation kinetics are measured in real time

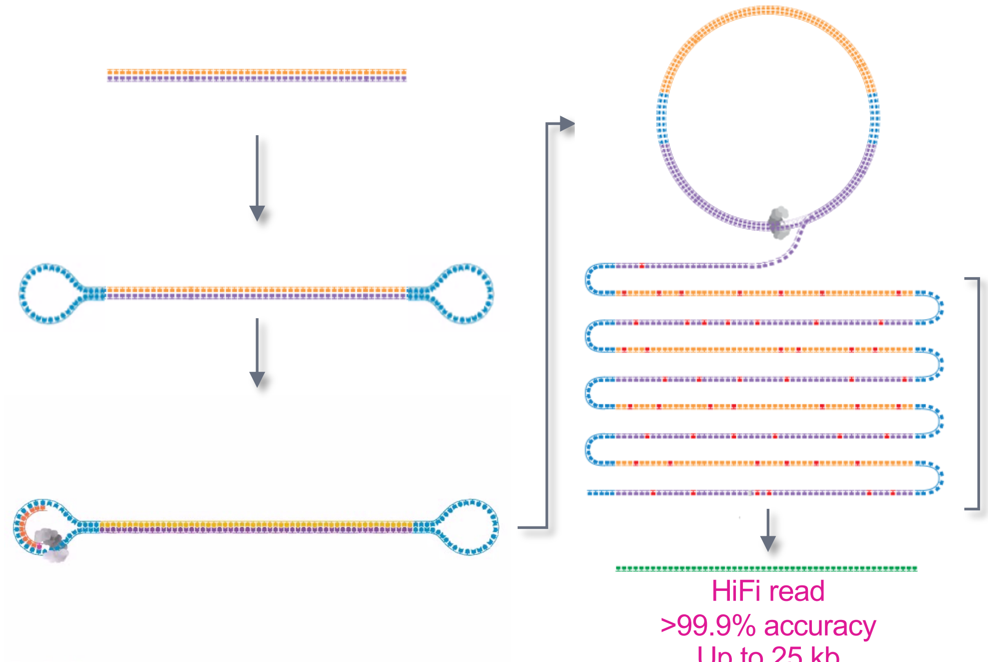
PacBio

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



Start with high-quality dsDNA

Prepare SMRTbell libraries

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

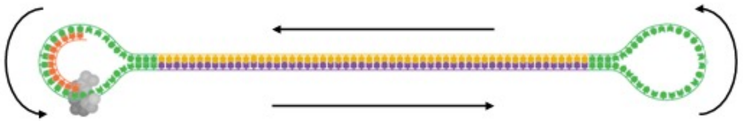
Polymerase reads are trimmed of adapters to yield subreads

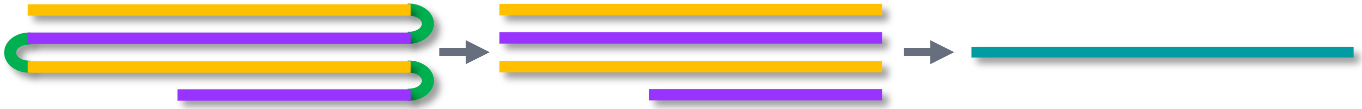
Consensus and methylation status are called from subreads

HiFi read
>99.9% accuracy
Up to 25 kb

Summary of Read Metrics Definitions and their utility

SMRTbell Template





Polymerase Read

Definition:

- Sequence of nucleotides incorporated by polymerase while reading a template
- High Quality region only
- Includes adapters
- 1 molecule → 1 Polymerase Read

Purpose:

- Sequencing run performance QC

Subreads

Definition:

- Adapters removed
- 1 molecule → 1 or more Subreads

Purpose:

- Library insert size QC
- Used for applications requiring multi-molecule consensus sequence generation

CCS Reads

Definition:

- Represents highest-quality single-sequence for an insert
- 1 molecule → 1 CCS read

HiFi read >99% (Q20) accuracy

Purpose:

- Library insert size QC
- Used for applications requiring intra-molecular consensus sequence generation

PacBio

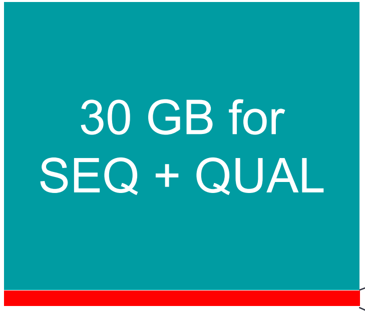
PacBio Glossary of Terms (http://www.pacb.com/wp-content/uploads/2015/09/Pacific-Biosciences-Glossary-of-Terms.pdf)

Representation of 5mC CpG data uses BAM format standard

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

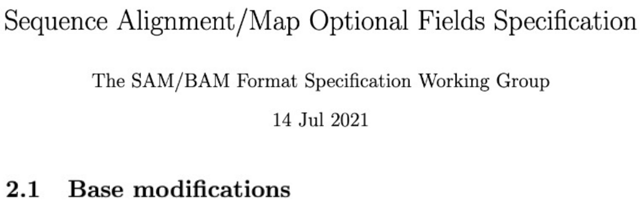
Sequel IIe system

hifi_reads.bam



30 GB for SEQ + QUAL

+5% for methylation



Sequence Alignment/Map Optional Fields Specification

The SAM/BAM Format Specification Working Group

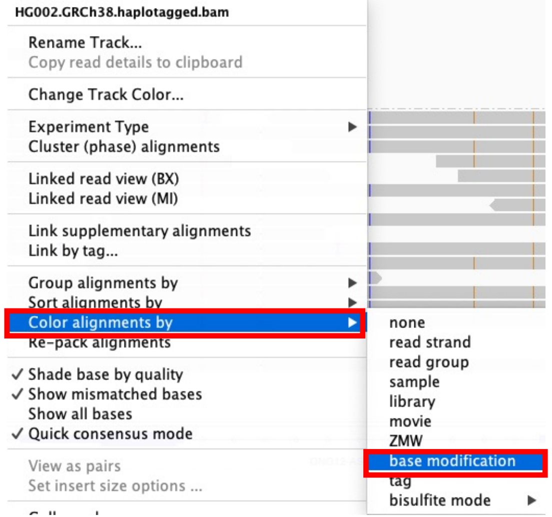
14 Jul 2021

2.1 Base modifications

```
MM:Z:C+m,5,12,0
ML:B:C,204,89,26
```

Supported in IGV 2.10

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...

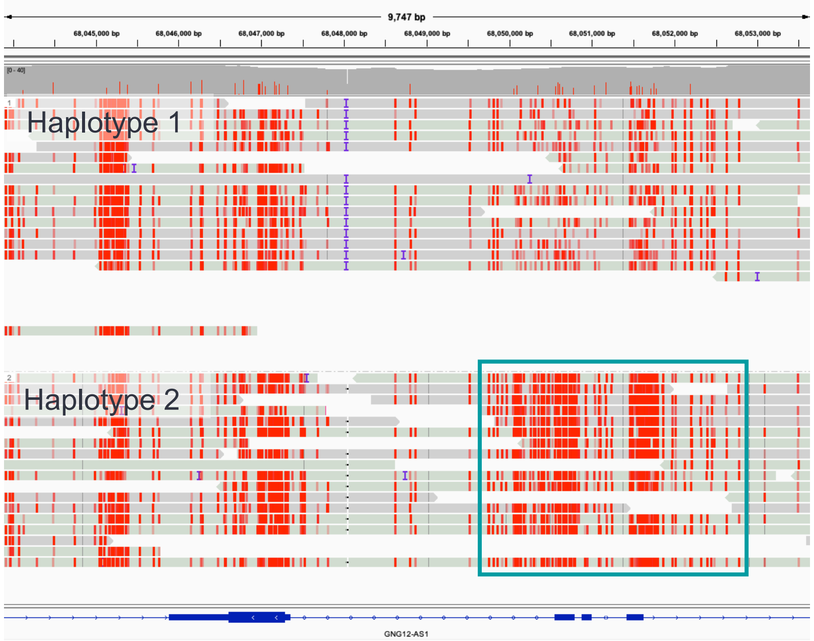
Group alignments by ▶
Sort alignments by ▶
Color alignments by ▶
Re-pack alignments

✓ Shade base by quality
✓ Show mismatched bases
Show all bases
✓ Quick consensus mode

View as pairs
Set insert size options ...

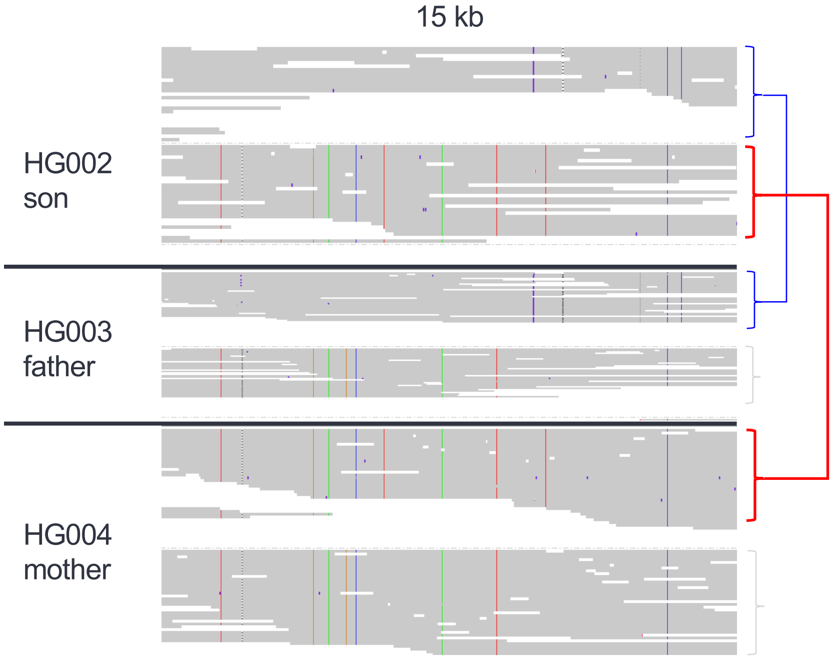
none
read strand
read group
sample
library
movie
ZMW
base modification
tag
bisulfite mode ▶

Supported in IGV 2.10



Allele-specific methylation
(imprinting)

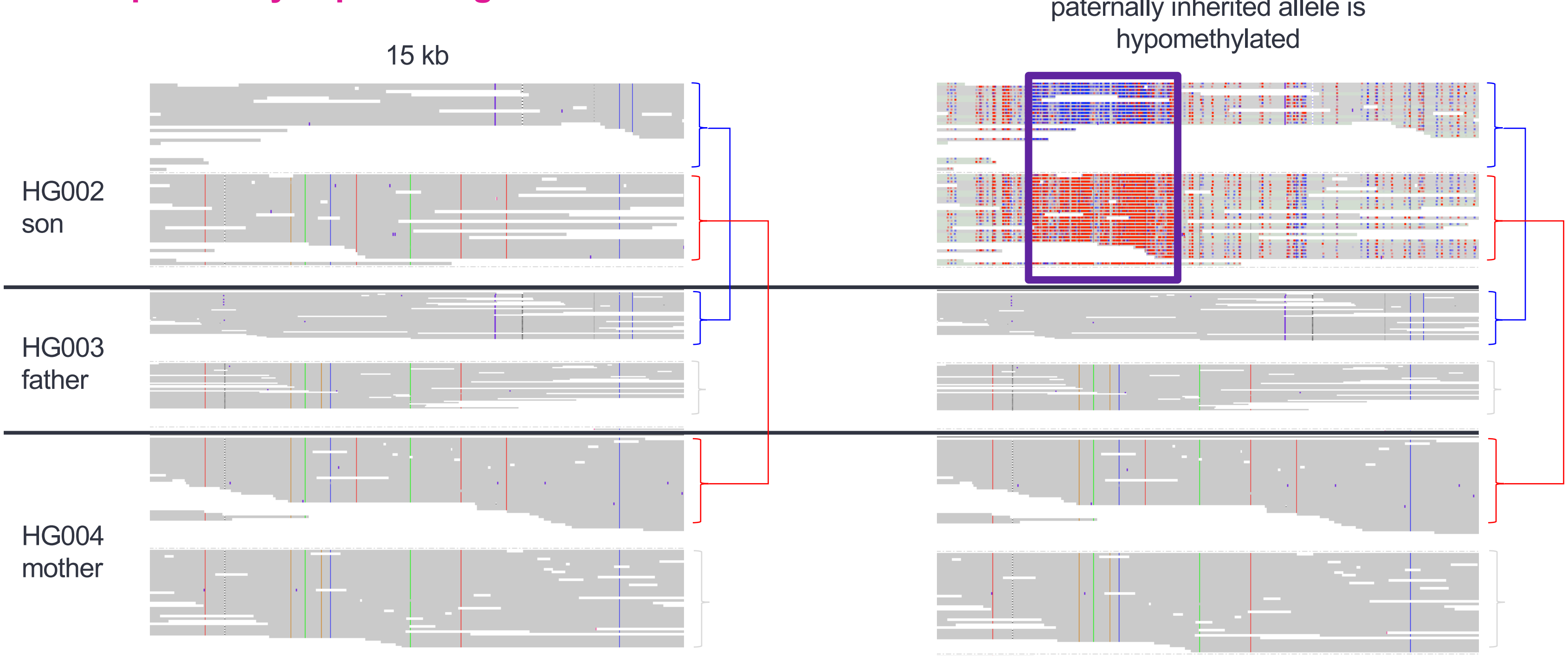
Haplotype phasing reveals parental imprinting in human



https://downloads.pacbcloud.com/public/dataset/HG002-CpG-methylation-202202/

Haplotype phasing reveals parental imprinting in human

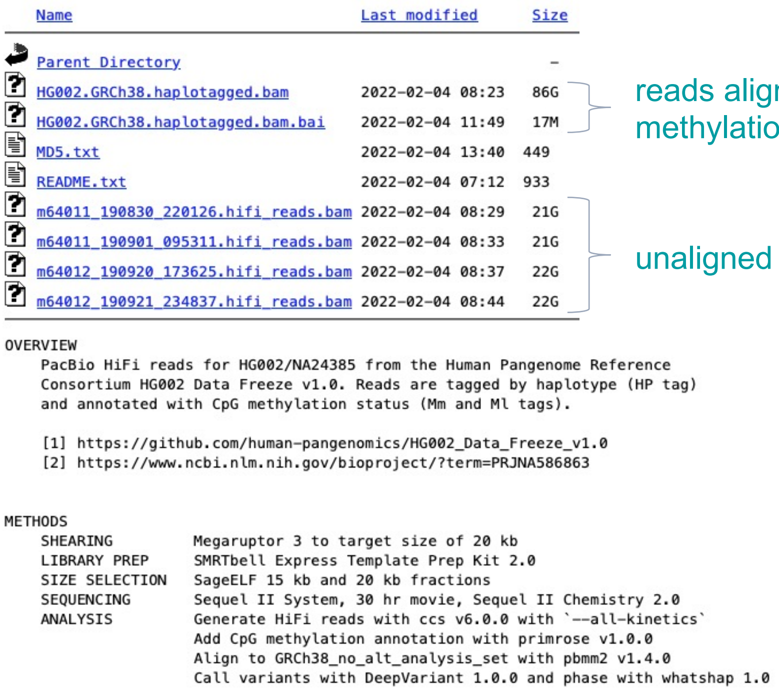
PEG3 = paternally expressed gene 3



https://downloads.pacbcloud.com/public/dataset/HG002-CpG-methylation-202202/

HG002 Sample Dataset – 34× coverage

https://downloads.pacbcloud.com/public/dataset/HG002-CpG-methylation-202202/



Name	Last modified	Size
Parent Directory		-
HG002.GRCh38.haplotagged.bam	2022-02-04 08:23	86G
HG002.GRCh38.haplotagged.bam.bai	2022-02-04 11:49	17M
MD5.txt	2022-02-04 13:40	449
README.txt	2022-02-04 07:12	933
m64011_190830_220126.hifi_reads.bam	2022-02-04 08:29	21G
m64011_190901_095311.hifi_reads.bam	2022-02-04 08:33	21G
m64012_190920_173625.hifi_reads.bam	2022-02-04 08:37	22G
m64012_190921_234837.hifi_reads.bam	2022-02-04 08:44	22G

```
OVERVIEW
    PacBio HiFi reads for HG002/NA24385 from the Human Pangenome Reference
    Consortium HG002 Data Freeze v1.0. Reads are tagged by haplotype (HP tag)
    and annotated with CpG methylation status (Mm and Ml tags).

    [1] https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0
    [2] https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA586863


METHODS
    SHEARING         Megaruptor 3 to target size of 20 kb
    LIBRARY PREP     SMRTbell Express Template Prep Kit 2.0
    SIZE SELECTION   SageELF 15 kb and 20 kb fractions
    SEQUENCING       Sequel II System, 30 hr movie, Sequel II Chemistry 2.0
    ANALYSIS         Generate HiFi reads with ccs v6.0.0 with `--all-kinetics`
                     Add CpG methylation annotation with primrose v1.0.0
                     Align to GRCh38_no_alt_analysis_set with pbmm2 v1.4.0
                     Call variants with DeepVariant 1.0.0 and phase with whatshap 1.0
```

reads aligned to GRCh38
methylation tags Mm/Ml & haplotype tags PS/HP

unaligned reads with methylation tags Mm/Ml

Understanding Sequencing Coverage and Depth

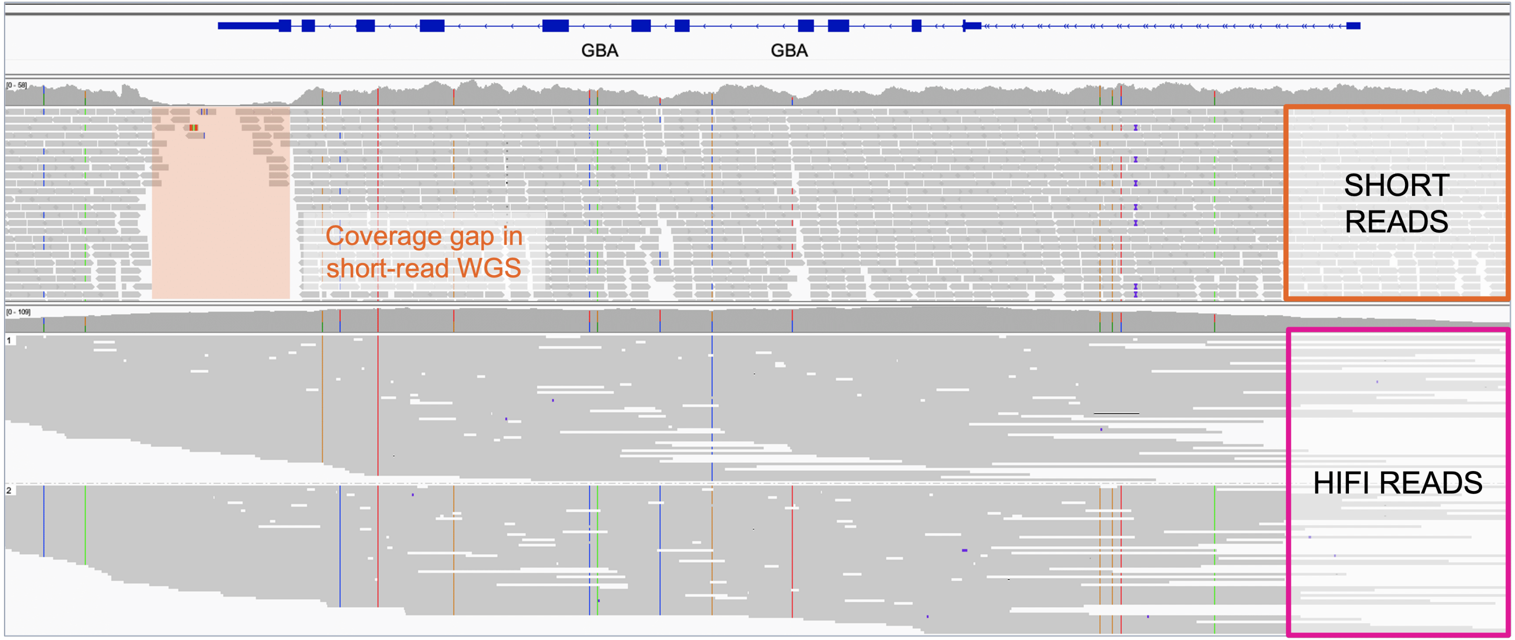
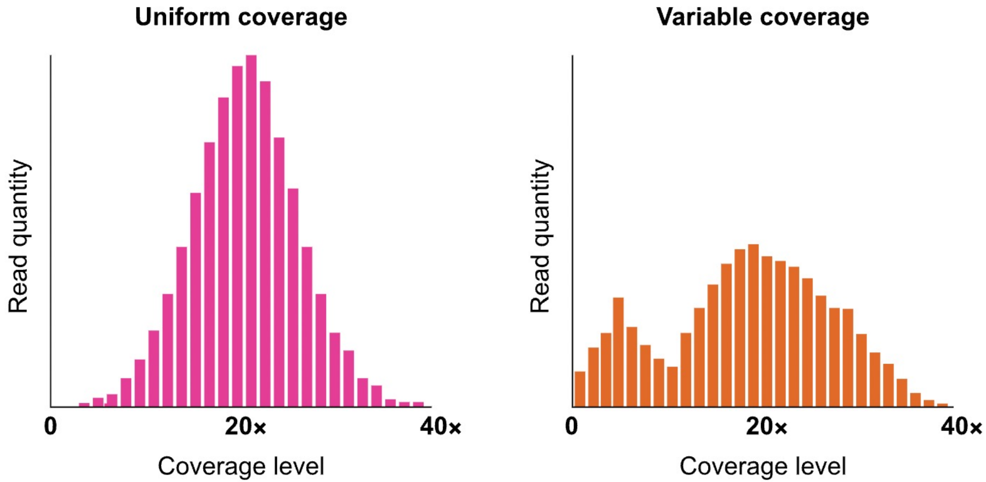
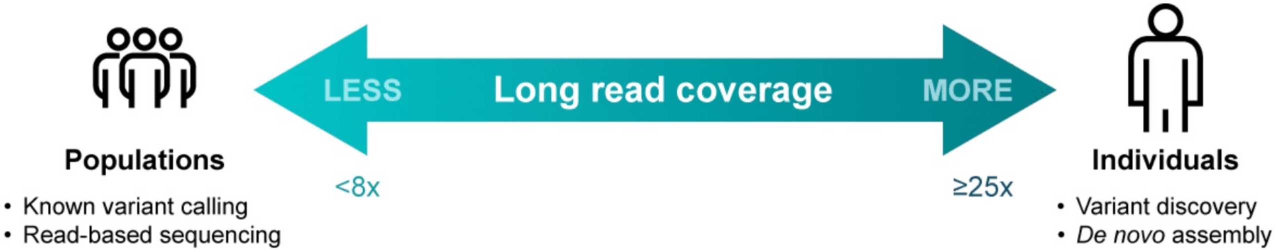


Figure 1. IGV generated image of PacBio long reads (purple section) and short-read alternative (orange section) covering a genomic reference region (blue line and bars at top). Note the area not covered by any reads (grey strips) in the short-read sequence alignment.

Understanding Sequencing Coverage and Depth



Coverage uniformity tells us how evenly distributed individual reads are across the genome or region of interest.



Populations

- Known variant calling
- Read-based sequencing

Individuals

- Variant discovery
- *De novo* assembly

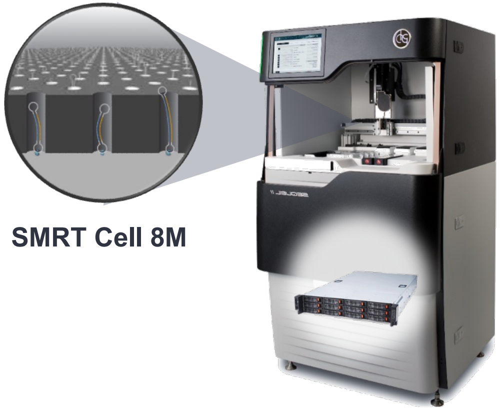
HiFi 5-base sequencing: a complete genome & epigenome

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

A
C
G
T
+ 5mC

SMRT Link software overview

Sequel IIe System and Software v12


SMRT Cell 8M

Sequel IIe System - the only sequencer with highly accurate long reads

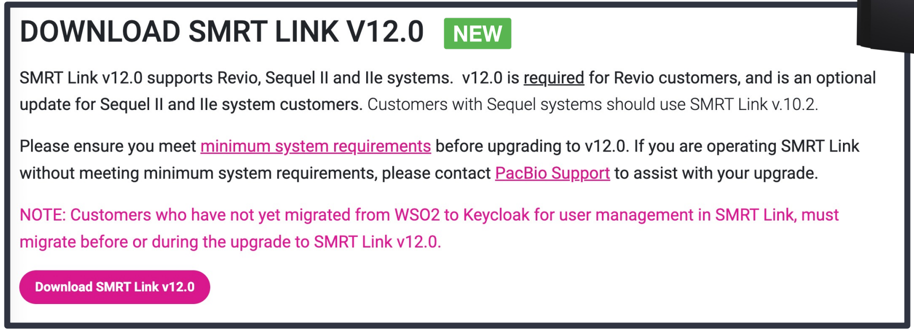
- off the box
    - Fast time to results, significantly less compute needs, greatly reduced storage
    - Lower overall solution cost resulting in more accessible system

SMRT Link – PacBio's open source SMRT Analysis software suite.

- Support intuitive GUI or command-line interface

SMRT Sequencing Data on a Network Server

Software Download


DOWNLOAD SMRT LINK V12.0 NEW

SMRT Link v12.0 supports Revio, Sequel II and IIe systems. v12.0 is required for Revio customers, and is an optional update for Sequel II and IIe 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.

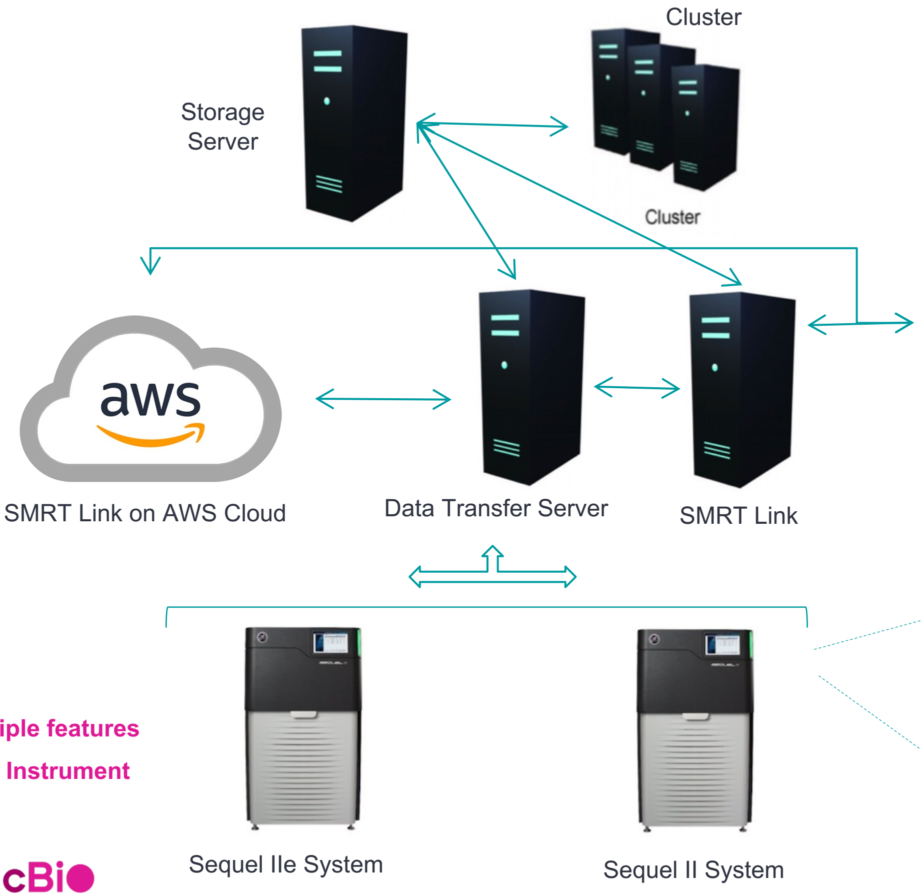
NOTE: Customers who have not yet migrated from WSO2 to Keycloak for user management in SMRT Link, must migrate before or during the upgrade to SMRT Link v12.0.

Download SMRT Link v12.0

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

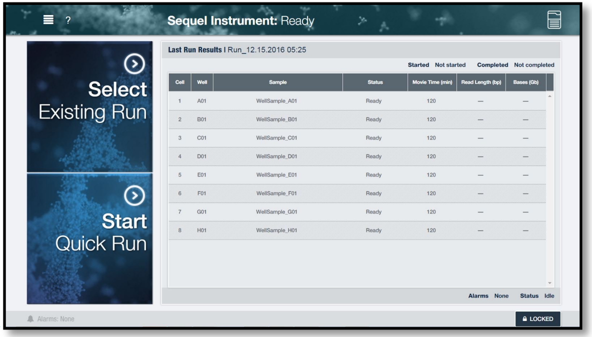
SMRT Link

SMRT Link system



Multiple features

On Instrument



Sequel Instrument: Ready

Last Run Results I Run_12.15.2016 05:25

Started Not started **Completed** Not completed

Cell	Well	Sample	Status	Movie Time (min)	Read Length (bp)	Bases (Gb)
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

Alarms: None

Compute requirements Sequel IIe 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
Shared Data Storage	
Sequencing Data	20 TB[a]
Analysis Data	40 TB[a]
Network	
10 GbE strongly recommended, 1GbE required[b]	

64 x 4 = 256 GB RAM

[a] Storage is calculated for one Sequel IIe System, assuming 100 human genomes per year at 30-fold coverage, *de novo* assembly

[b] Connection between the Head Node and Sequel IIe System

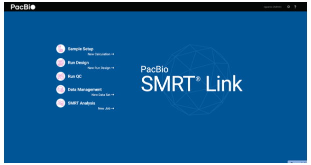
Server OS: CentOS 7.x and 8.x, and Ubuntu 18.04 and 20.04 64-bit Linux® distributions
(This also applies to SMRT Link compute nodes.)

PacBio

PacBio Software suite and analysis pipeline for SMRT data

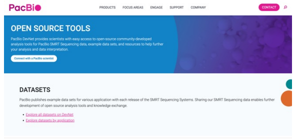
Denovo assembly	Improved Phased Assembly (IPA)
Variant Calling	DeepVariant + whatshap + pbsv
Structure variant	pbsv
Isoform detection	Iso-Seq
Single cell isoform	MAS-Seq
Metagenome	HiFi + Third party tools
16S Full-length	HiFi + Third party tools

- 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



OPEN SOURCE TOOLS

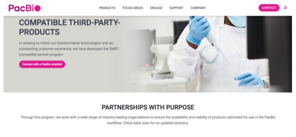
PacBio DevNet provides scientists with easy access to open-source community-developed analysis tools for PacBio SMRT Sequencing data, example data sets, and resources to help further your analysis and data interpretation.

DATASETS

PacBio publishes example data sets for various application with each release of the SMRT Sequencing Systems. Sharing our SMRT Sequencing data enables further development of open source analysis tools and knowledge exchange.

- Explore all datasets on DevNet
- Explore datasets by application

Datasets including example datasets



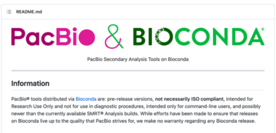
COMPATIBLE THIRD-PARTY-PRODUCTS

In working to match our transformative technologies with an outstanding customer experience, we have developed the SMRT compatible partner program.

PARTNERSHIPS WITH PURPOSE

Through this program, we work with a wide range of industry-leading organizations to ensure the availability and viability of products optimized for use in the PacBio workflow. Check back soon for an updated directory.

SMRT Compatible Analysis Products (partner solutions)



PacBio Secondary Analysis Tools on Bioconda

Information

PacBio® tools distributed via Bioconda are: pre-release versions, **not necessarily ISO compliant**, intended for Research Use Only and not for use in diagnostic procedures, intended only for command-line users, and possibly newer than the currently available SMRT® Analysis builds. While efforts have been made to ensure that releases on Bioconda live up to the quality that PacBio strives for, we make no warranty regarding any Bioconda release.

pbbioconda

(developmental tools)

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

Access SMRT Link using Google Chrome web browser (v83+)

Instruments – Displays instrument status, SMRT Cell status and sequencing ZMW plots for in-progress runs

↓

Sample Setup – Calculate annealing / binding / cleanup (ABC) RXs for preparing samples for sequencing

↓

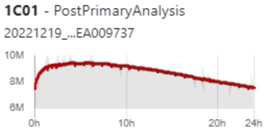
Runs – Create sequencing run designs and review primary sequencing quality metrics

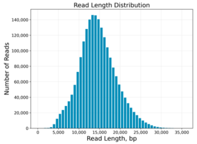
↓

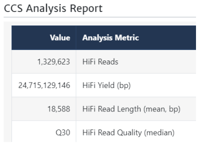
Data Management – Create projects and group data sets; import / delete reference and barcode files

↓

SMRT Analysis – Perform secondary analysis (e.g., *de novo* assembly, full-length RNA transcript analysis)

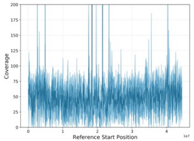






CCS Analysis Report

Value	Analysis Metric
1,329,623	HiFi Reads
24,715,129,146	HiFi Yield (bp)
18,588	HiFi Read Length (mean, bp)
Q30	HiFi Read Quality (median)



PacBio

For more detailed information on SMRT Link modules, see the SMRT Link user guide (102-877-300).

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

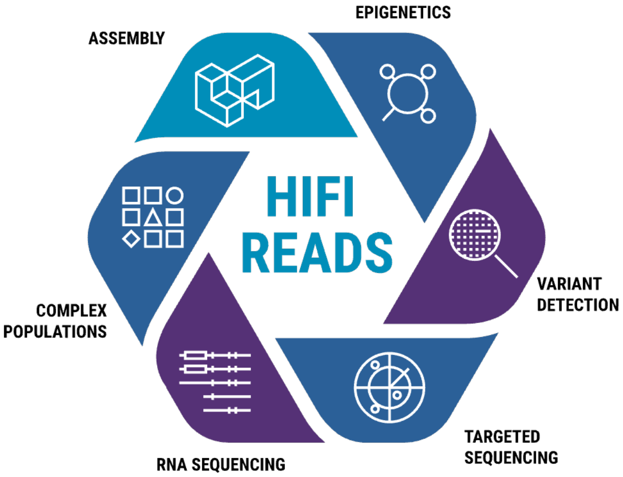
- Application note – MAS-Seq for single cell isoform sequencing (102-326-549)

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

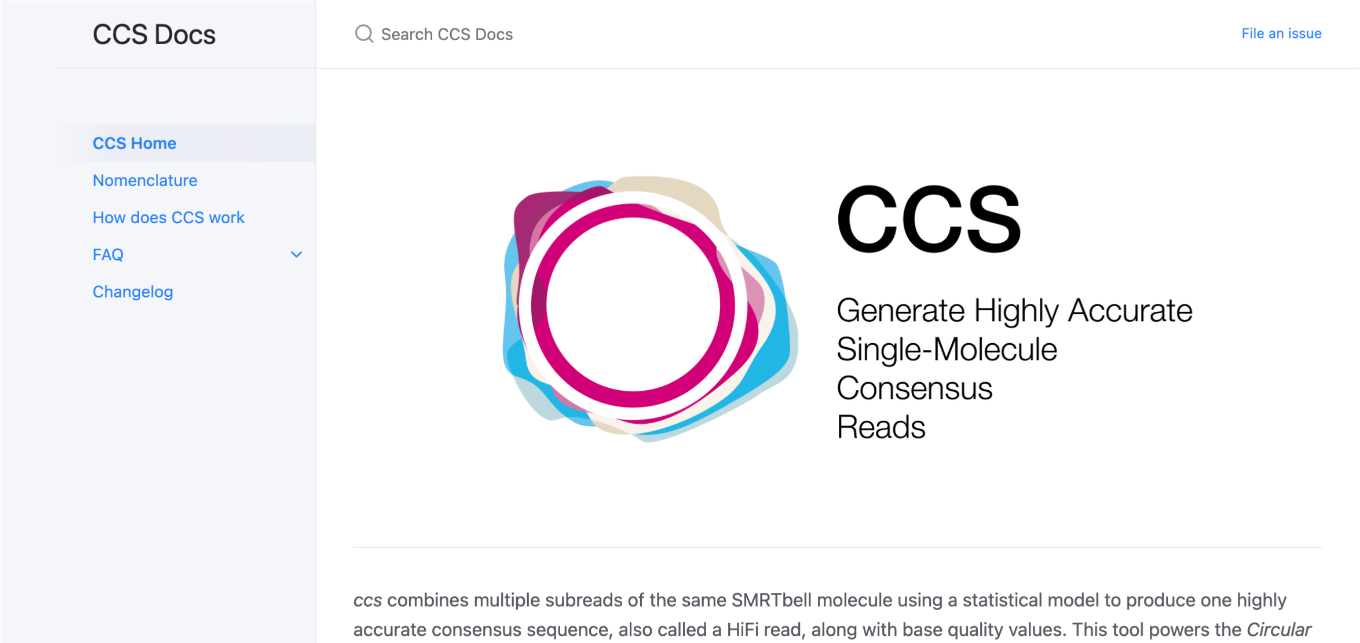
- Technical overview – MAS-Seq library preparation using the MAS-Seq for 10x Single Cell 3’ kit (102-829-300)
- Technical overview – Multiplexed amplicon library preparation using SMRTbell prep kit 3.0 (102-395-900)
- Technical overview – Nanobind HT kits for automated HMW DNA extraction (Coming soon)
- 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)

how.to ccs repository



CCS Docs

Search CCS Docs

File an issue

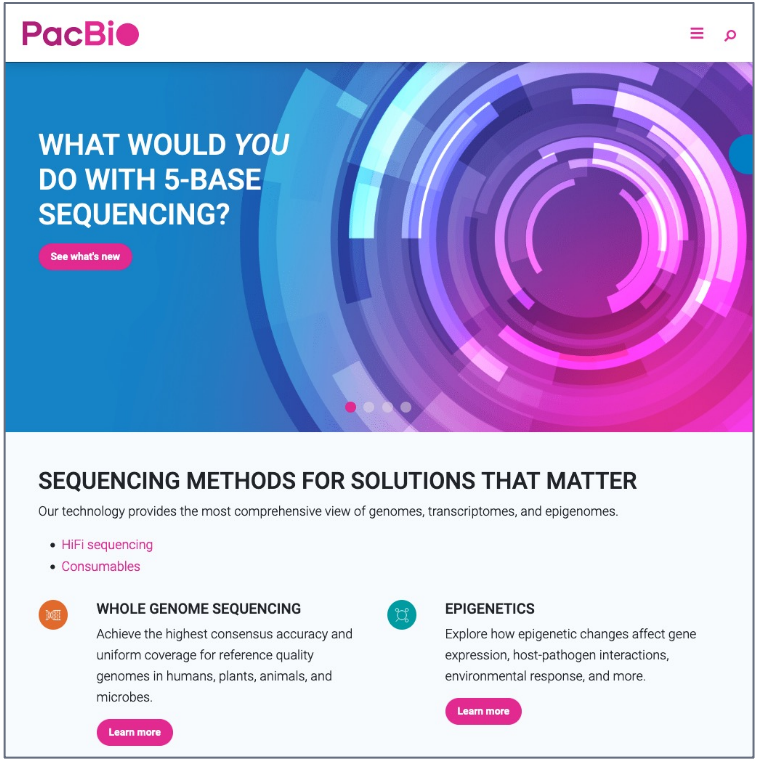
CCS

Generate Highly Accurate Single-Molecule Consensus Reads

ccs combines multiple subreads of the same SMRTbell molecule using a statistical model to produce one highly accurate consensus sequence, also called a HiFi read, along with base quality values. This tool powers the *Circular*

https://ccs.how/

5-base sequencing available now!



PacBio

WHAT WOULD *YOU* DO WITH 5-BASE SEQUENCING?

See what's new

SEQUENCING METHODS FOR SOLUTIONS THAT MATTER

Our technology provides the most comprehensive view of genomes, transcriptomes, and epigenomes.

- HiFi sequencing
- Consumables

WHOLE GENOME SEQUENCING

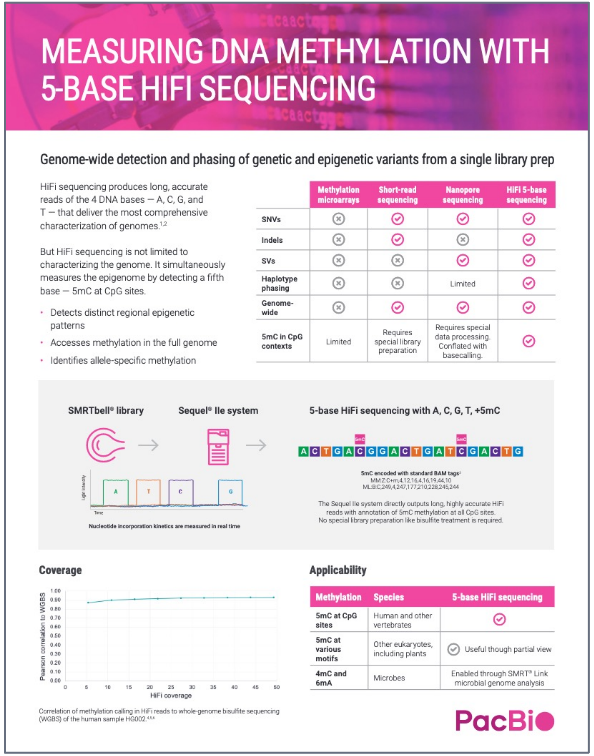
Achieve the highest consensus accuracy and uniform coverage for reference quality genomes in humans, plants, animals, and microbes.

Learn more

EPIGENETICS

Explore how epigenetic changes affect gene expression, host-pathogen interactions, environmental response, and more.

Learn more



MEASURING DNA METHYLATION WITH 5-BASE HIFI SEQUENCING

Genome-wide detection and phasing of genetic and epigenetic variants from a single library prep

HiFi sequencing produces long, accurate reads of the 4 DNA bases — A, C, G, and T — that deliver the most comprehensive characterization of genomes.[1,2]

But HiFi sequencing is not limited to characterizing the genome. It simultaneously measures the epigenome by detecting a fifth base — 5mC at CpG sites.

- Detects distinct regional epigenetic patterns
- Accesses methylation in the full genome
- Identifies allele-specific methylation

	Methylation microarrays	Short-read sequencing	Nanopore sequencing	HiFi 5-base sequencing
SNVs	⊗	✓	✓	✓
Indels	⊗	✓	⊗	✓
SVs	⊗	⊗	✓	✓
Haplotype phasing	⊗	⊗	Limited	✓
Genome-wide	⊗	✓	✓	✓
5mC in CpG contexts	Limited	Requires special library preparation	Requires special data processing. Conflated with basecalling.	✓

SMRTbell® library
Sequel® IIe system
5-base HiFi sequencing with A, C, G, T, +5mC
Light intensity
Time
A
T
C
G
5mC
5mC
A C T G A C G G A C T G A T C G A C T G

Nucleotide incorporation kinetics are measured in real time

5mC encoded with standard BAM tags[3]
MM:Z:C+m,4,12,16,4,16,19,44,10
ML:B:C,249,4,247,177,210,228,245,244

The Sequel IIe system directly outputs long, highly accurate HiFi reads with annotation of 5mC methylation at all CpG sites. No special library preparation like bisulfite treatment is required.

Coverage

Correlation of methylation calling in HiFi reads to whole-genome bisulfite sequencing (WGBS) of the human sample HG002.[4,5,6]

Applicability

Methylation	Species	5-base HiFi sequencing
5mC at CpG sites	Human and other vertebrates	✓
5mC at various motifs	Other eukaryotes, including plants	✓ Useful though partial view
4mC and 6mA	Microbes	Enabled through SMRT® Link microbial genome analysis

PacBio

https://www.pacb.com/literature/application-brief-measuring-dna-methylation-with-5-base-hifi-sequencing/

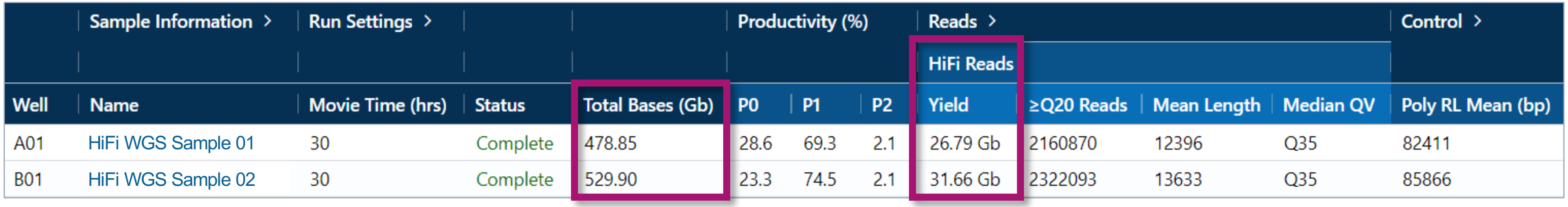
PacBio

PacBio sequencing Run QC interpretation

Field Application Scientist | 應用科學家
Steiner Chen | 陳冠安

Primary analysis table overview

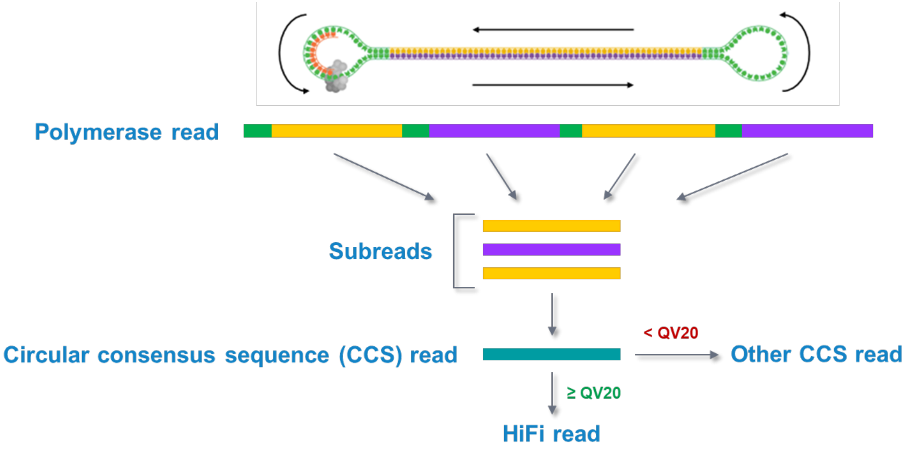
Primary analysis metrics summary table: Data yield



	Sample Information >	Run Settings >			Productivity (%)			Reads >				Control >
								HiFi Reads				
Well	Name	Movie Time (hrs)	Status	Total Bases (Gb)	P0	P1	P2	Yield	≥Q20 Reads	Mean Length	Median QV	Poly RL Mean (bp)
A01	HiFi WGS Sample 01	30	Complete	478.85	28.6	69.3	2.1	26.79 Gb	2160870	12396	Q35	82411
B01	HiFi WGS Sample 02	30	Complete	529.90	23.3	74.5	2.1	31.66 Gb	2322093	13633	Q35	85866

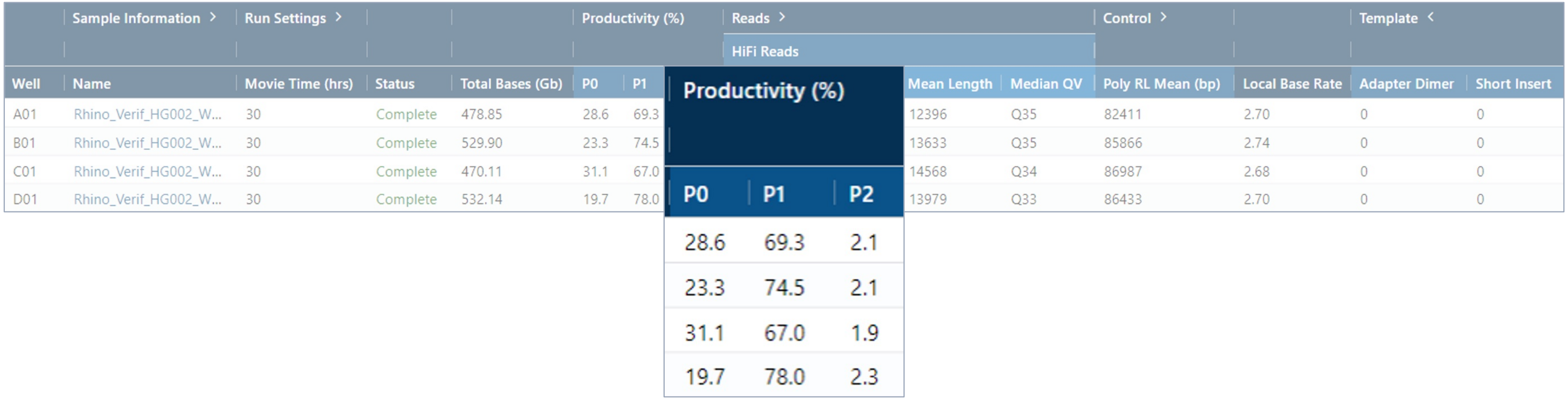
1. **Total bases (GB):** Calculated by multiplying the number of Productive (*P1*) ZMWs by the mean Polymerase read length; displayed in Gigabases.

2. **HiFi base yield:** The total yield of the CCS reads whose quality value (QV) is equal to or greater than 20; displayed in Gigabases.



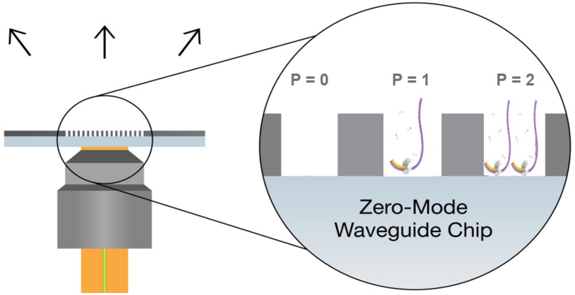
PacBio

Primary analysis metrics summary table: Productivity



	Sample Information ›	Run Settings ›			Productivity (%)		Reads ›		Control ›		Template ‹	
							HiFi Reads					
Well	Name	Movie Time (hrs)	Status	Total Bases (Gb)	P0	P1	Mean Length	Median QV	Poly RL Mean (bp)	Local Base Rate	Adapter Dimer	Short Insert
A01	Rhino_Verif_HG002_W...	30	Complete	478.85	28.6	69.3	12396	Q35	82411	2.70	0	0
B01	Rhino_Verif_HG002_W...	30	Complete	529.90	23.3	74.5	13633	Q35	85866	2.74	0	0
C01	Rhino_Verif_HG002_W...	30	Complete	470.11	31.1	67.0	14568	Q34	86987	2.68	0	0
D01	Rhino_Verif_HG002_W...	30	Complete	532.14	19.7	78.0	13979	Q33	86433	2.70	0	0

Productivity (%)		
P0	P1	P2
28.6	69.3	2.1
23.3	74.5	2.1
31.1	67.0	1.9
19.7	78.0	2.3



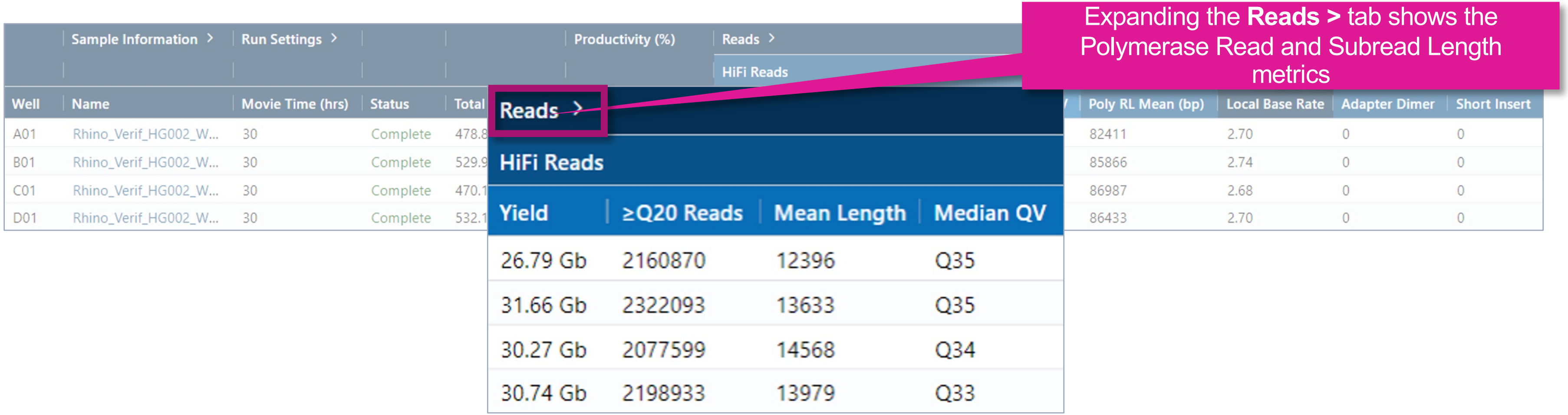
Productivity

- ***P0***: Empty ZMW; no signal detected.
- ***P1***: ZMW with a high quality (HQ) read generated.
- ***P2***: Other – signal detected but no HQ read generated

Recommended t**arget *P1* is ~50% to 85%** for optimal HiFi data yield per SMRT Cell. ***If P0 values are <10% then the SMRT Cell is overloaded.***

Primary analysis metrics summary table: HiFi read metrics

Expanding the **Reads >** tab shows the Polymerase Read and Subread Length metrics

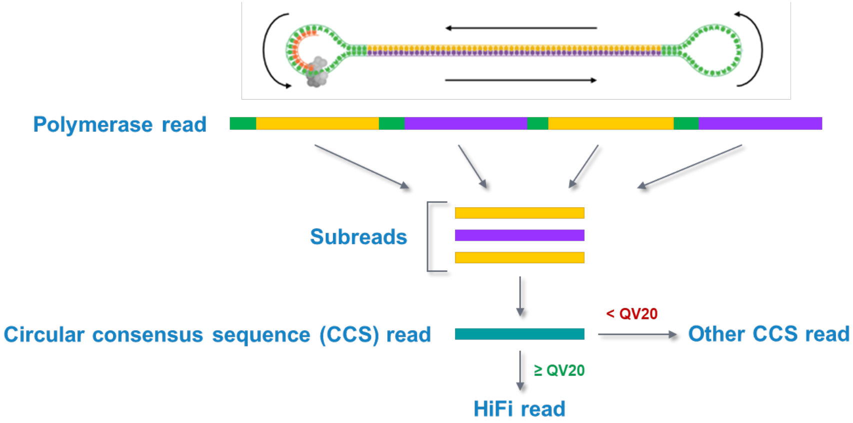


	Sample Information >	Run Settings >			Productivity (%)	Reads >				
						HiFi Reads				
Well	Name	Movie Time (hrs)	Status	Total			Poly RL Mean (bp)	Local Base Rate	Adapter Dimer	Short Insert
A01	Rhino_Verif_HG002_W...	30	Complete	478.8			82411	2.70	0	0
B01	Rhino_Verif_HG002_W...	30	Complete	529.9			85866	2.74	0	0
C01	Rhino_Verif_HG002_W...	30	Complete	470.1			86987	2.68	0	0
D01	Rhino_Verif_HG002_W...	30	Complete	532.1			86433	2.70	0	0

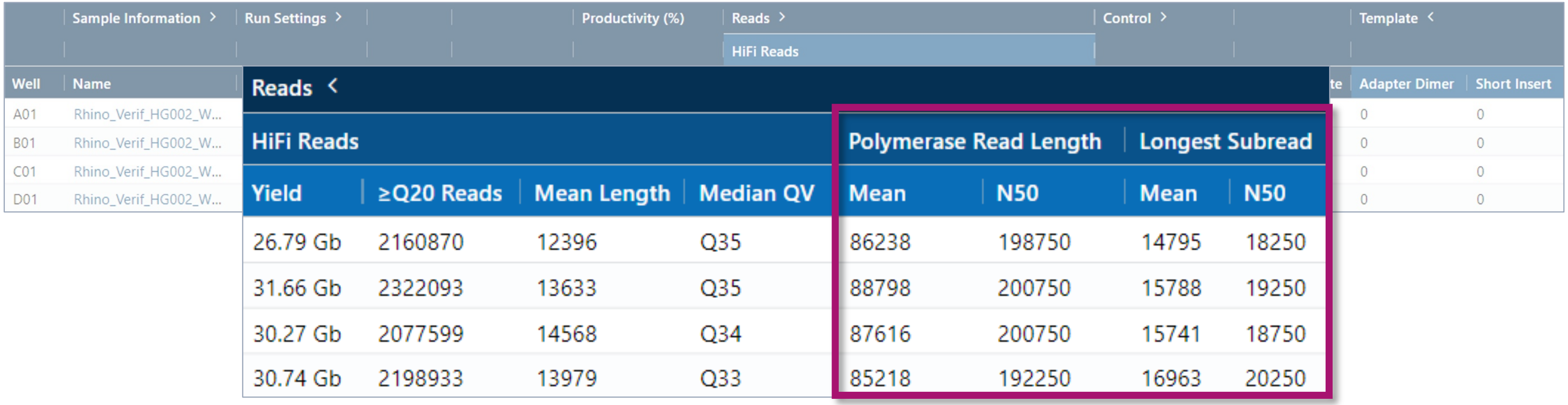
Reads >			
HiFi Reads			
Yield	≥Q20 Reads	Mean Length	Median QV
26.79 Gb	2160870	12396	Q35
31.66 Gb	2322093	13633	Q35
30.27 Gb	2077599	14568	Q34
30.74 Gb	2198933	13979	Q33

HiFi Read Metrics

- **HiFi Reads ≥Q20 Reads:** The total number of CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Reads Yield:** The total yield (in base pairs) of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Reads Mean Length:** The mean read length of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Reads Mean QV:** The mean quality value of CCS Reads whose QV is equal to or greater than 20.



Primary analysis metrics summary table: Polymerase read & subread metrics



Sample Information › | Run Settings › | Productivity (%) | Reads › | HiFi Reads | Control › | Template ‹

Well	Name	te	Adapter Dimer	Short Insert
A01	Rhino_Verif_HG002_W...		0	0
B01	Rhino_Verif_HG002_W...		0	0
C01	Rhino_Verif_HG002_W...		0	0
D01	Rhino_Verif_HG002_W...		0	0

Reads ‹

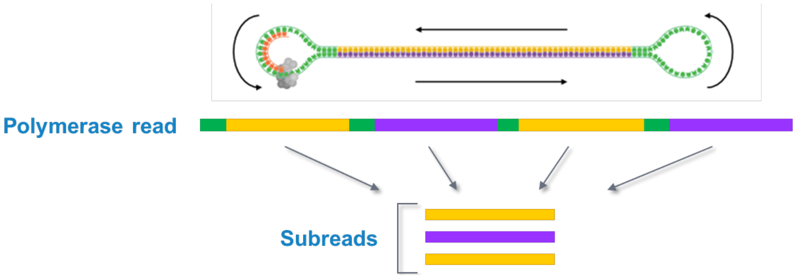
HiFi Reads				Polymerase Read Length		Longest Subread	
Yield	≥Q20 Reads	Mean Length	Median QV	Mean	N50	Mean	N50
26.79 Gb	2160870	12396	Q35	86238	198750	14795	18250
31.66 Gb	2322093	13633	Q35	88798	200750	15788	19250
30.27 Gb	2077599	14568	Q34	87616	200750	15741	18750
30.74 Gb	2198933	13979	Q33	85218	192250	16963	20250

Polymerase read length metrics

- **Polymerase Mean:** The mean high-quality read length of all polymerase reads. The value includes bases from adapters as well as multiple passes around a circular template.
- **Polymerase N50:** 50% of all read bases came from polymerase reads longer than this value.

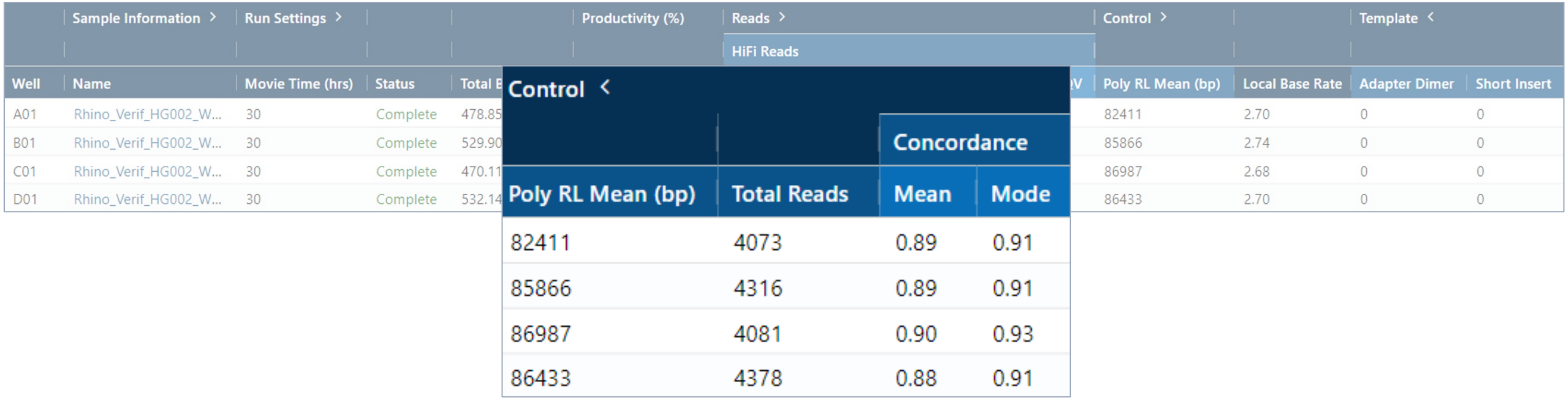
Subread length metrics

- **Longest Subread Mean:** The mean subread length, considering only the longest subread from each ZMW.
- **Longest Subread N50:** 50% of all read bases came from subreads longer than this value when considering only the longest subread from each ZMW.



PacBio

Primary analysis metrics summary table: Control read metrics

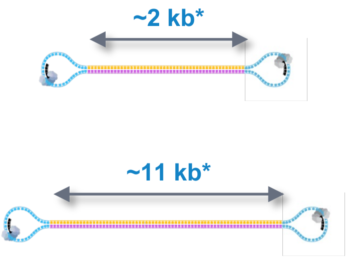


	Sample Information >	Run Settings >			Productivity (%)	Reads >	Control >		Template <	
						HiFi Reads				
Well	Name	Movie Time (hrs)	Status	Total B		V	Poly RL Mean (bp)	Local Base Rate	Adapter Dimer	Short Insert
A01	Rhino_Verif_HG002_W...	30	Complete	478.85			82411	2.70	0	0
B01	Rhino_Verif_HG002_W...	30	Complete	529.90			85866	2.74	0	0
C01	Rhino_Verif_HG002_W...	30	Complete	470.11			86987	2.68	0	0
D01	Rhino_Verif_HG002_W...	30	Complete	532.14			86433	2.70	0	0

Control <			
		Concordance	
Poly RL Mean (bp)	Total Reads	Mean	Mode
82411	4073	0.89	0.91
85866	4316	0.89	0.91
86987	4081	0.90	0.93
86433	4378	0.88	0.91

Control read metrics

- **Total Reads:** The number of control reads obtained.
- **Poly RL Mean:** The mean polymerase read length of the control reads.
- **Concordance Mean:** The average concordance (agreement) between the control raw reads and the control reference sequence.
- **Concordance Mode:** The modal concordance (agreement) between the control raw reads and the control reference sequence.



* Not to scale

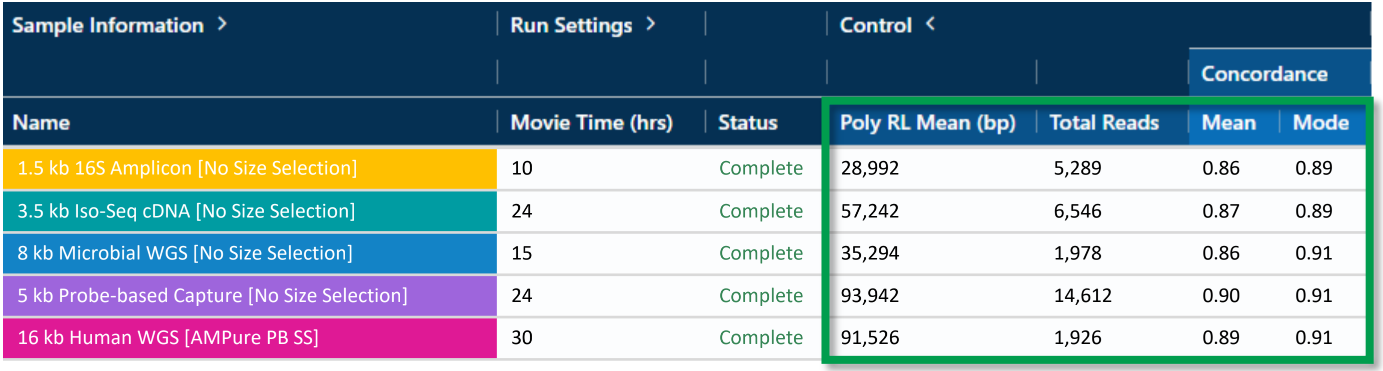
Sequel II DNA internal control 3.1 is aligned to the known 2 kb control reference sequence.

- Control 3.1 polymerase read length is typically ≥15 kb for a 15-hr movie time and ≥30 kb for a 30-hr movie time

Sequel II DNA internal control 3.2 is aligned to the known 11 kb control reference sequence.

- Control 3.2 polymerase read length is typically ≥40 kb for a 15-hr movie time and ≥80 kb for a 30-hr movie time

Control reads: Example expected performance for DNA internal control



	Sample Information ›	Run Settings ›		Control ‹			
						Concordance	
	Name	Movie Time (hrs)	Status	Poly RL Mean (bp)	Total Reads	Mean	Mode
DNA Control 3.1	1.5 kb 16S Amplicon [No Size Selection]	10	Complete	28,992	5,289	0.86	0.89
DNA Control 3.1	3.5 kb Iso-Seq cDNA [No Size Selection]	24	Complete	57,242	6,546	0.87	0.89
DNA Control 3.1	8 kb Microbial WGS [No Size Selection]	15	Complete	35,294	1,978	0.86	0.91
DNA Control 3.2	5 kb Probe-based Capture [No Size Selection]	24	Complete	93,942	14,612	0.90	0.91
DNA Control 3.2	16 kb Human WGS [AMPure PB SS]	30	Complete	91,526	1,926	0.89	0.91

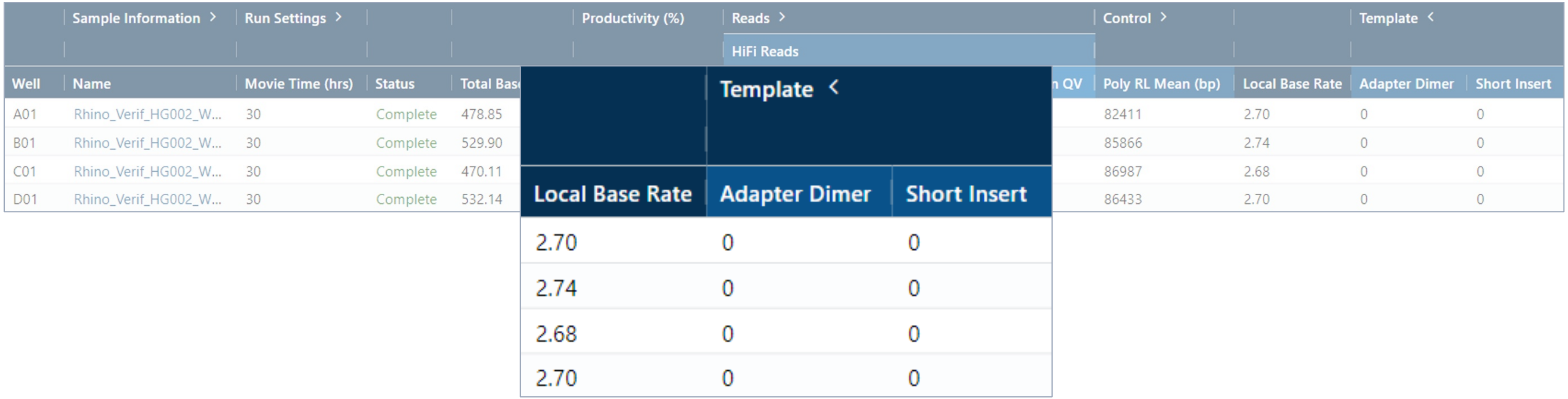
	Expected performance range			
	Sequel II DNA internal control 3.1		Sequel II DNA internal control 3.2	
Metric	15 hr movie	30 hr movie	15 hr movie	30 hr movie
Control read count	≥500	≥500	≥1000	≥1000
Control polymerase read length (Mean)	≥15 kb	≥30 kb	≥40 kb	≥80 kb
Control concordance (Mean)	≥0.85	≥0.85	≥0.87	≥0.87

Note: DNA internal control 3.2 is in part derived from sequences with high homology to lambda phage

- As a result, in sequencing runs with microbial genomes containing integrated phage sequence, a small fraction of reads may be misidentified as internal control reads.
- Such reads will display low concordance to the control sequence.

PacBio

Primary analysis metrics summary table: Local base rate & other metrics



	Sample Information >	Run Settings >			Productivity (%)	Reads >	Control >		Template <	
						HiFi Reads				
Well	Name	Movie Time (hrs)	Status	Total Bas		n QV	Poly RL Mean (bp)	Local Base Rate	Adapter Dimer	Short Insert
A01	Rhino_Verif_HG002_W...	30	Complete	478.85			82411	2.70	0	0
B01	Rhino_Verif_HG002_W...	30	Complete	529.90			85866	2.74	0	0
C01	Rhino_Verif_HG002_W...	30	Complete	470.11			86987	2.68	0	0
D01	Rhino_Verif_HG002_W...	30	Complete	532.14			86433	2.70	0	0

	Template <	
Local Base Rate	Adapter Dimer	Short Insert
2.70	0	0
2.74	0	0
2.68	0	0
2.70	0	0

Local base rate

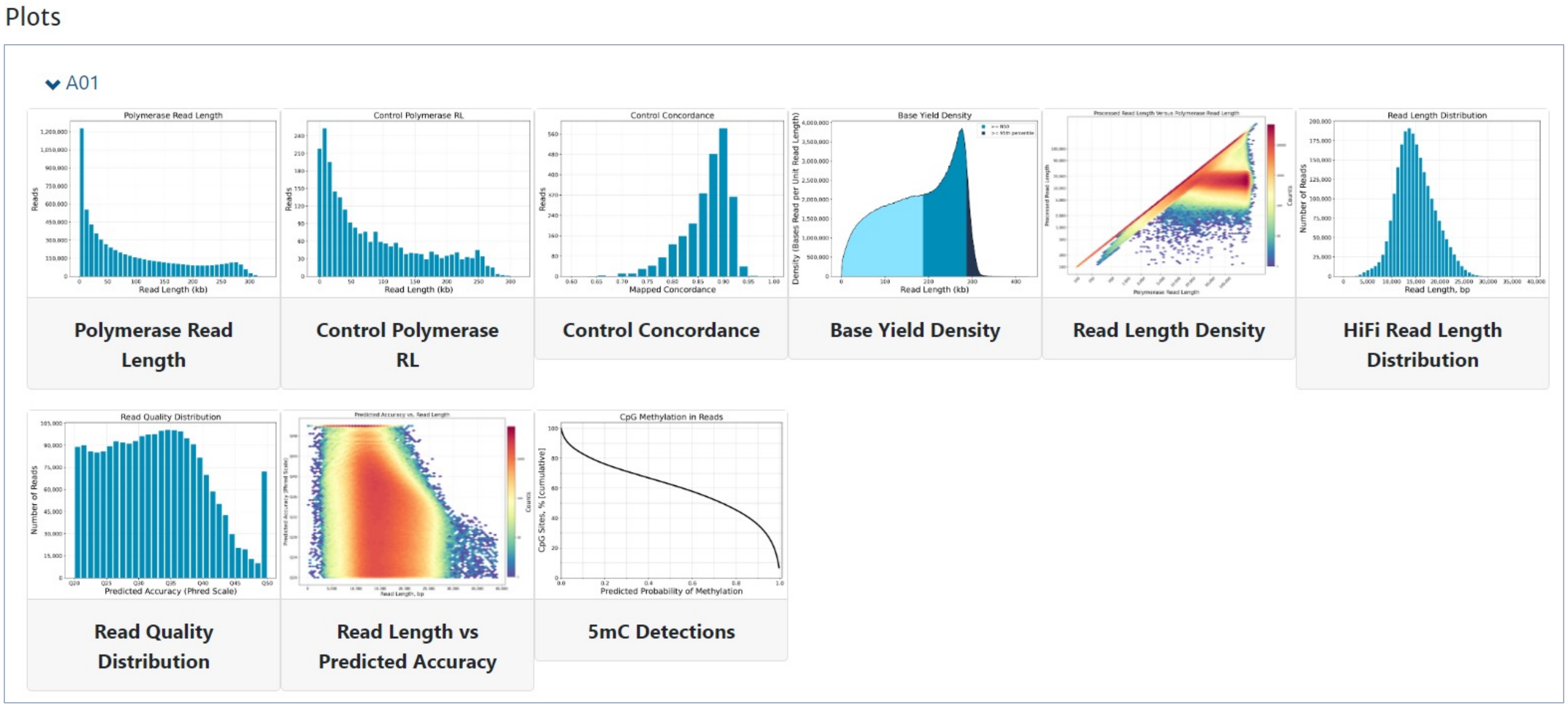
- The average base incorporation rate, excluding polymerase pausing events.
  - For Sequel II Binding kit 3.1, local base rate is typically ~1.7 – 2.2 bases per second
  - For Sequel II Binding kit 3.2, local base rate is typically ~2.2 – 3.0 bases per second

Template

- **Adapter Dimer:** The % of pre-filter ZMWs which have observed inserts of 0-10 bp. These are likely adapter dimers.
  - Purified SMRTbell libraries should typically show <2% adapter dimer levels
- **Short Insert:** The % of pre-filter ZMWs which have observed inserts of 11-100 bp. These are likely short fragment contamination.
  - Purified SMRTbell libraries should typically show <2% short insert levels

Run QC report plots

Run QC report plots: Overview

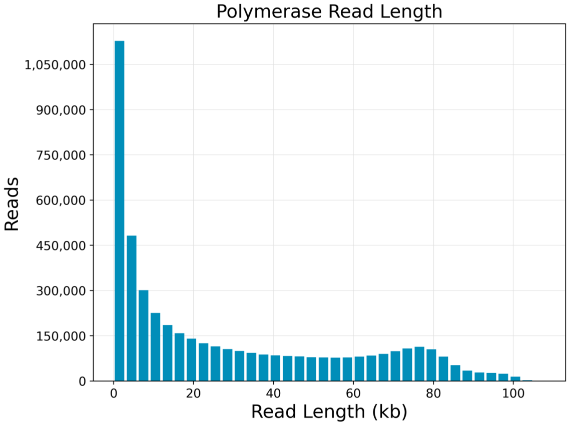


- Click the **>** arrow to expand rows to view Run QC Report plots for each SMRT Cell
- Clicking on an individual plot displays an expanded view.

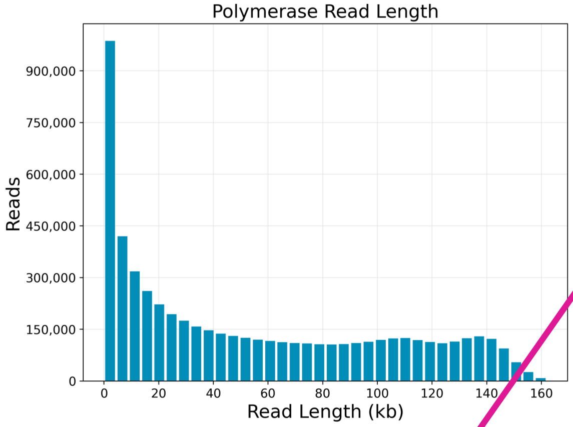
Run QC report plots: Polymerase read length

Example polymerase read length plots for different sample library types

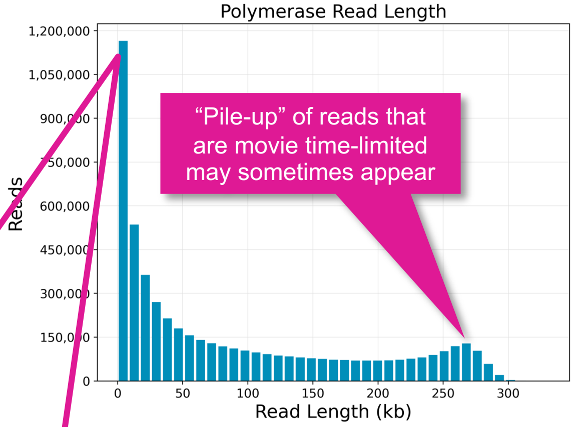
1.5 kb 16S Amplicon library
[10 h Movie Time]



8 kb Microbial WGS library
[15 h Movie Time]



16 kb Human WGS library
[30 h Movie Time]



"Pile-up" of reads that are movie time-limited may sometimes appear

Polymerase read length

- Plots the number of reads against the polymerase read length
- Read count typically decreases as the polymerase length increases

Early-terminating reads typically appear as a major left-hand peak and can be caused by:

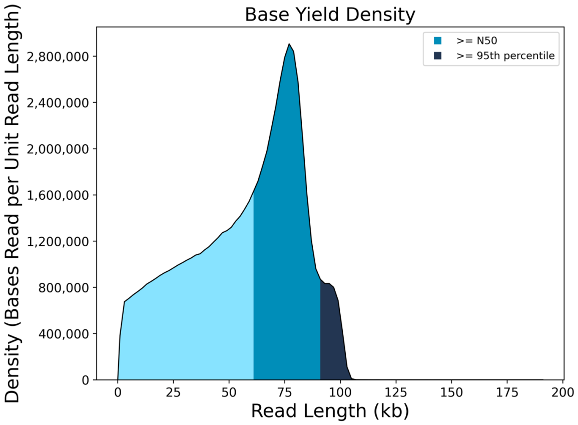
- Adapter hairpin oligo quality issues or incomplete adapter ligation
- Presence of nicks or other DNA damage
- Disassociation of the polymerase from the template
- Laser-induced photodamage that stops polymerase activity

Run QC report plots: Base yield density

Example base yield density plots for different sample library types

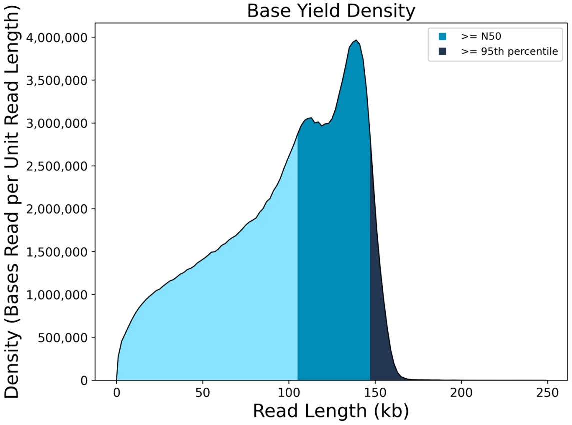
1.5 kb 16S Amplicon library

[10 h Movie Time]



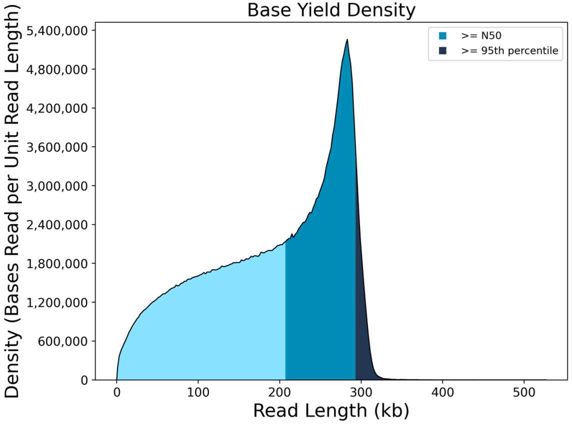
8 kb Microbial WGS library

[15 h Movie Time]



16 kb Human WGS library

[30 h Movie Time]



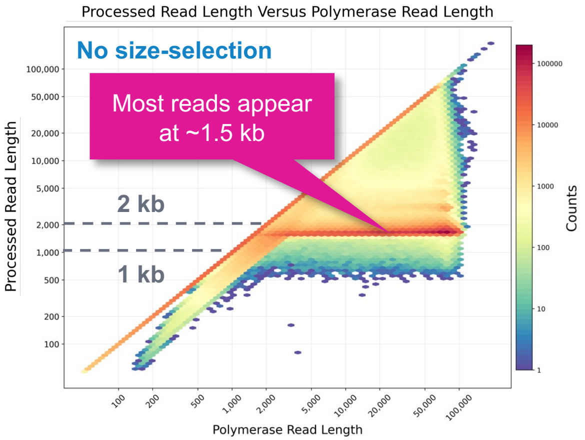
Base yield density: Displays the number of bases sequenced in the collection according to the length of the read in which they were observed.

- Values displayed are per unit of read length (i.e., the base yield density) and are averaged over 2000 bp windows to gently smooth the data.
- Regions of the graph corresponding to bases found in reads longer than the **N50** and **N95** values are shaded in **medium** and **dark blue**, respectively

Run QC report plots: Read length density

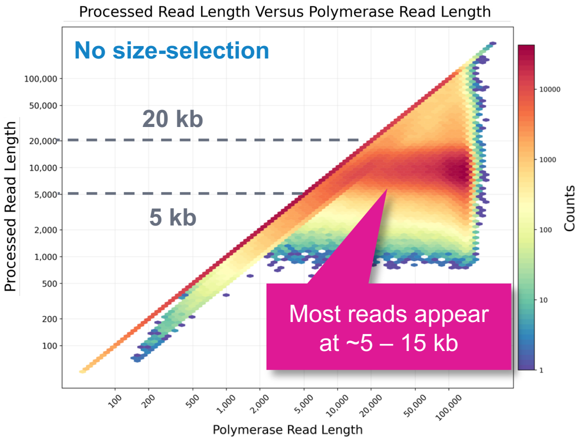
Example read length density plots for different sample library types

1.5 kb 16S Amplicon library
[10 h Movie Time]



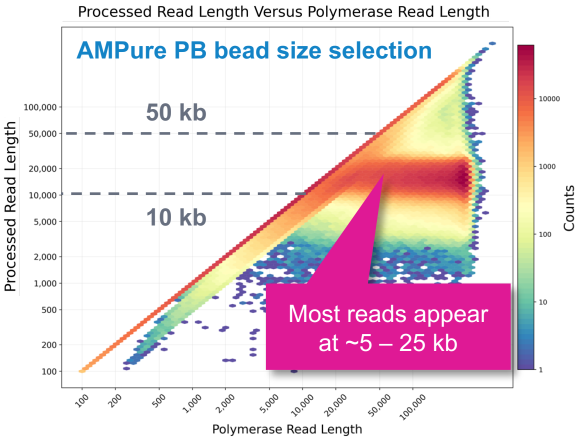
Most reads appear at ~1.5 kb

8 kb Microbial WGS library
[15 h Movie Time]



Most reads appear at ~5 – 15 kb

16 kb Human WGS library
[30 h Movie Time]



Most reads appear at ~5 – 25 kb

Read length density: Displays a (log scale) density plot of reads, binned according to their estimated insert read length* and polymerase read length

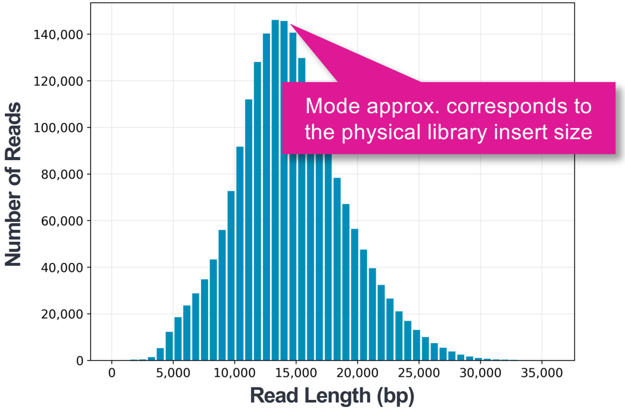
- This plot is useful for quickly visualizing aspects of library quality (e.g., insert size distributions and reads terminating at adapters)
- Reads that are concordant with the expected physical library insert size should ideally appear as strong horizontal features with a high density of counts (i.e., appear as a "dark red" color)

* In RLD plots, the **'Processed Read Length'** value corresponds to the estimated insert size of the library and is taken to be equal to the median subread length for a ZMW

Run QC report plots: HiFi data-specific plots

The following HiFi data-specific plots below are generated for any run where CCS processing is performed on-instrument (Sequel IIe system) or in SMRT Link (Sequel II systems)

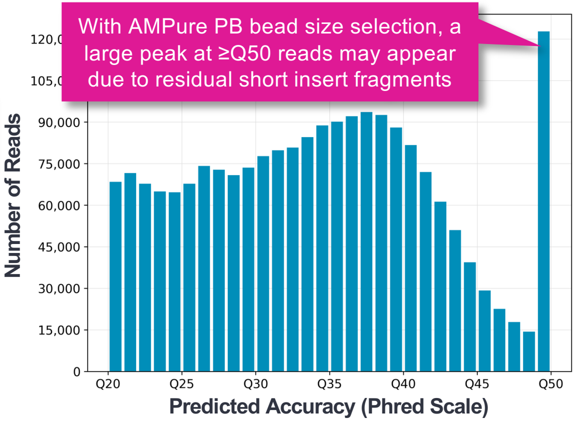
Read Length Distribution



Mode approx. corresponds to the physical library insert size

Displays a histogram distribution of HiFi Reads (QV ≥20)

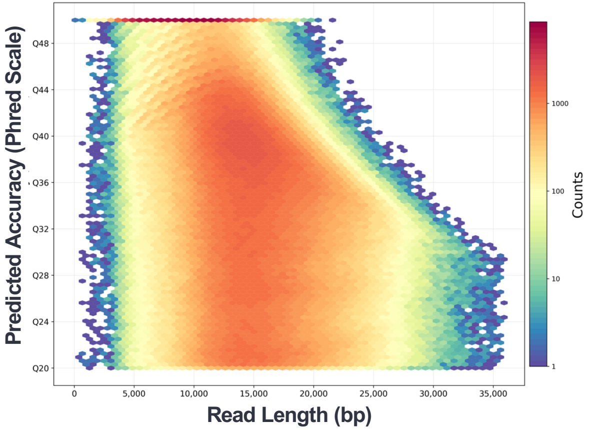
Read Quality Distribution



With AMPure PB bead size selection, a large peak at ≥Q50 reads may appear due to residual short insert fragments

Displays a histogram distribution of HiFi Reads (QV ≥20)

Predicted Accuracy vs. Read Length



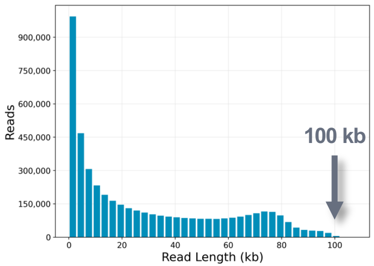
Displays a heat map of HiFi read lengths and predicted accuracies.

Run QC plots interpretation & example case studies

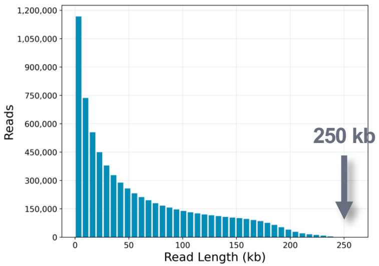
Polymerase read length plots: Example ideal *vs.* suboptimal performance

Example ideal performance with high-quality DNA samples

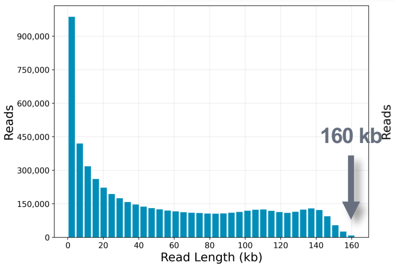
Plot should ideally show many reads **exceeding** the expected ideal mean pol RL range.*



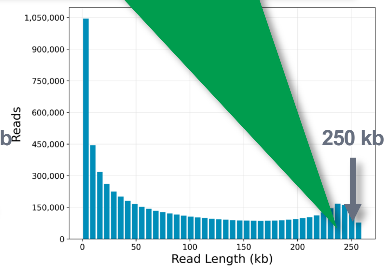
P1 = 59.4%
Mean Pol RL = 35,567 bp



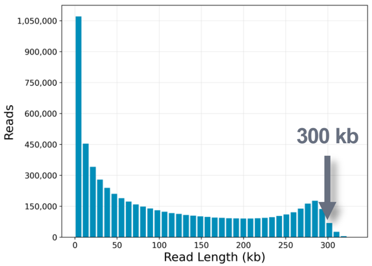
P1 = 84.5%
Mean Pol RL = 59,523 bp



P1 = 72.0%
Mean Pol RL = 54,428 bp



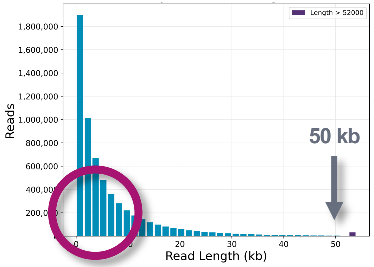
P1 = 79.7%
Mean Pol RL = 94,081 bp



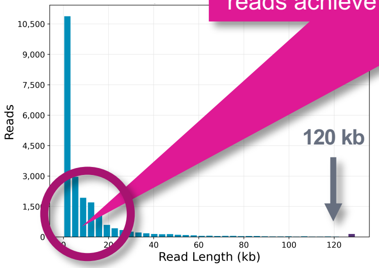
P1 = 74.9%
Mean Pol RL = 106,616 bp

Example suboptimal performance

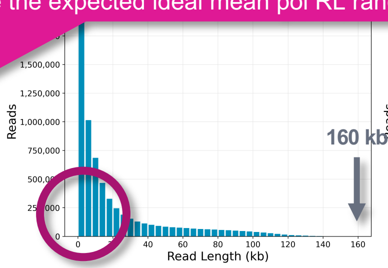
Most reads show **short** pol RL and relatively few reads achieve the expected ideal mean pol RL range.



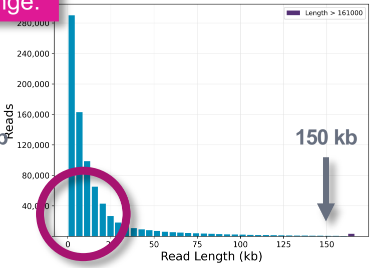
P1 = 75.6%
Mean Pol RL = 6,401 bp



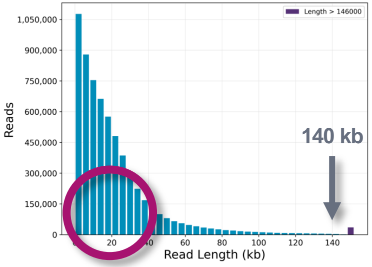
P1 = 0.47%
Mean Pol RL = 10,903 bp



P1 = 80.6%
Mean Pol RL = 21,396 bp



P1 = 80.2%
Mean Pol RL = 16,518 bp



P1 = 78.8%
Mean Pol RL = 21,327 bp

≤3 kb Amplicon
10 h Movie Time

2 – 5 kb Iso-Seq cDNA
24 h Movie Time

7 – 12 kb Microbial WGS
15 h Movie Time

5 – 10 kb Probe-based Capture
24 h Movie Time

15 – 18 kb Large Genome WGS
30 h Movie Time

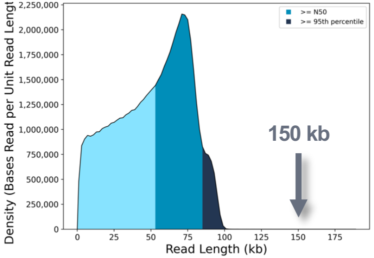
PacBio

* Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size.

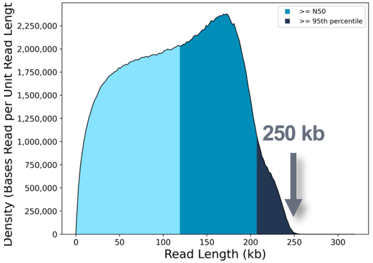
Base yield density: Example ideal *vs.* suboptimal performance

Example ideal performance with high-quality DNA samples

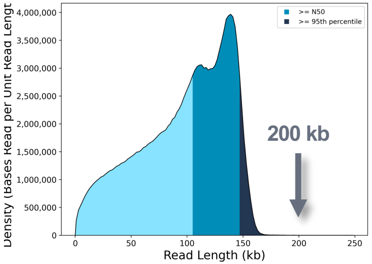
Plot should ideally show a **rightward bias**, indicating long polymerase RLs.*



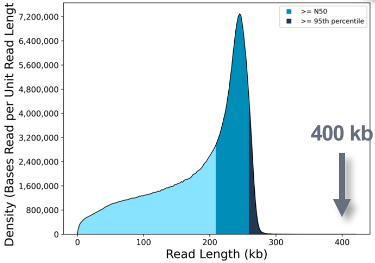
P1 = 59.4%
Mean Pol RL = 35,567 bp



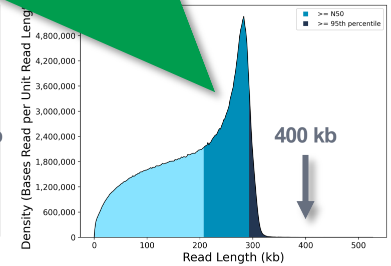
P1 = 84.5%
Mean Pol RL = 59,523 bp



P1 = 72.0%
Mean Pol RL = 54,428 bp



P1 = 79.7%
Mean Pol RL = 94,081 bp

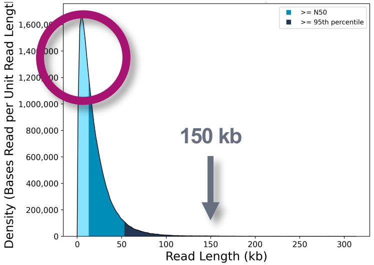


P1 = 74.9%
Mean Pol RL = 106,616 bp

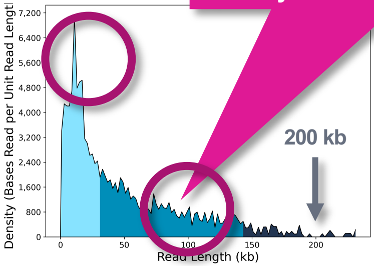
Example suboptimal performance

Samples with low *P1* yield show **low density counts** (Y-axis) and a **noisier** plot

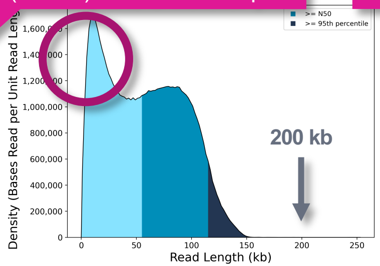
Samples with short polymerase RLs tend to show a **leftward bias** in the plot.



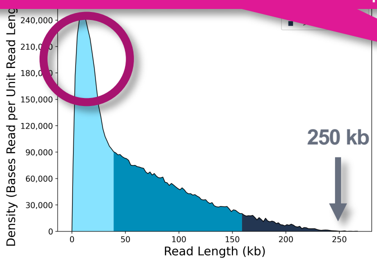
P1 = 75.6%
Mean Pol RL = 6,401 bp



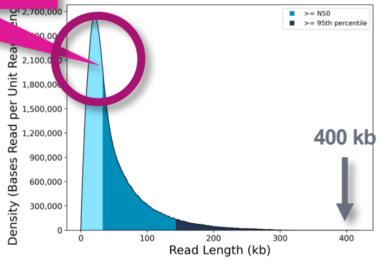
P1 = 0.47%
Mean Pol RL = 10,903 bp



P1 = 80.6%
Mean Pol RL = 21,396 bp



P1 = 80.2%
Mean Pol RL = 16,518 bp



P1 = 78.8%
Mean Pol RL = 21,327 bp

≤3 kb Amplicon
10 h Movie Time

2 – 5 kb Iso-Seq cDNA
24 h Movie Time

7 – 12 kb Microbial WGS
15 h Movie Time

5 – 10 kb Probe-based Capture
24 h Movie Time

15 – 18 kb Large Genome WGS
30 h Movie Time

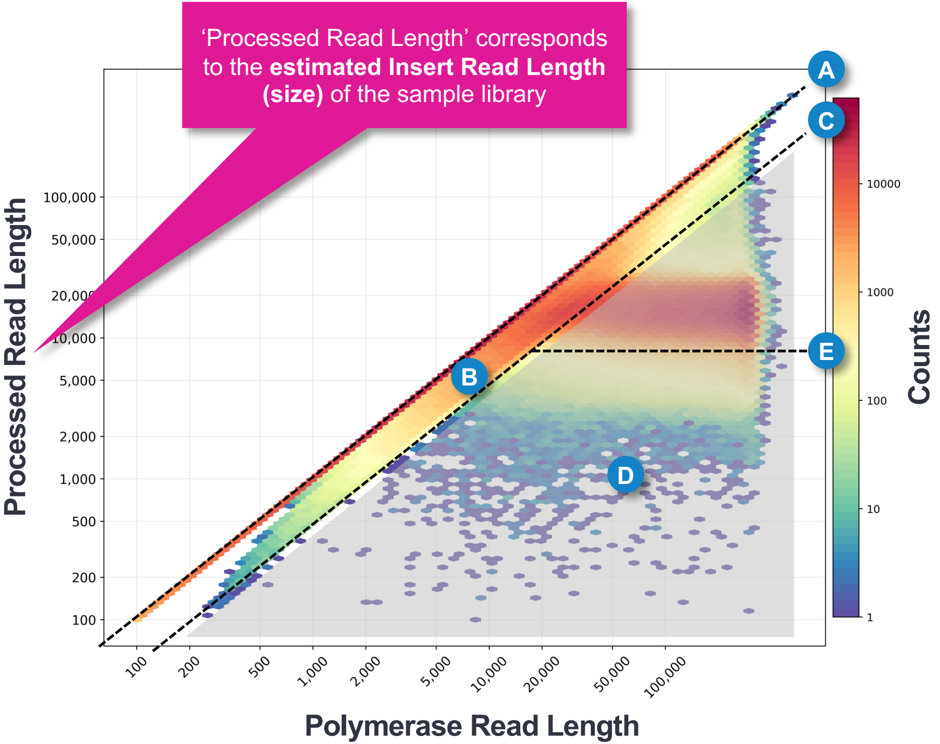
* Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size.

Read length density plot interpretation

Displays a (log scale) density plot of reads, binned according to their estimated Insert Read Length* and Polymerase Read Length

This plot is useful for **quickly visualizing aspects of library quality**, including insert size distributions, reads terminating at adapters, relative abundance of CCS reads, etc.

- **A** **(Primary diagonal line)** Reads terminating in the first observed pass along the SMRTbell template
- **B** **(Area between line A and line C)** Reads terminating in the second observed pass along the SMRTbell template
- **C** **(Secondary diagonal line)** Reads terminating at the second SMRTbell adapter
- **D** **(Grayed area below line C)** CCS (HiFi) reads
- **E** **(Horizontal line)** Size selection cutoff boundary



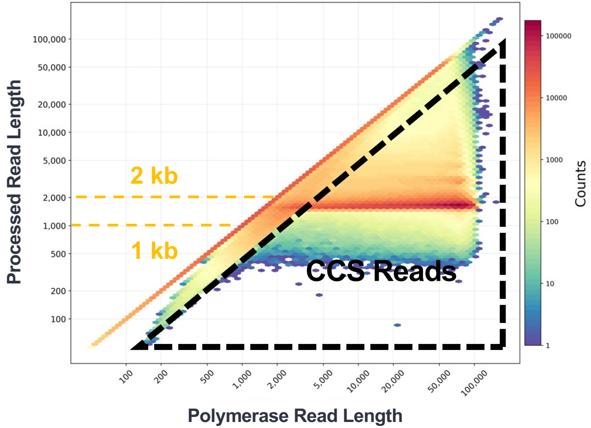
‘Processed Read Length’ corresponds to the **estimated Insert Read Length (size)** of the sample library

* **Note:** Estimated Insert Read Length (IRL) is labeled as ‘Processed Read Length’ on the Y-axis in Read Length Density plots. Estimated insert read length values are based on the **Median Subread Length** value for each ZMW.

Read length density plot: Example ideal performance

1.5 kb 16S Amplicon library
10 h Movie time

No size-selection

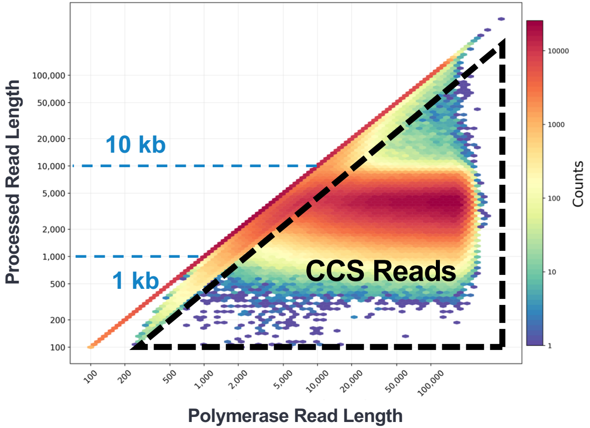


Most Insert Read Lengths are centered around ~1.5 kb, consistent with the expected size of this (non-size selected) 16S amplicon library.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
177.39	77.7	28,530	2.16

3.5 kb Iso-Seq library
24 h Movie time

No size-selection

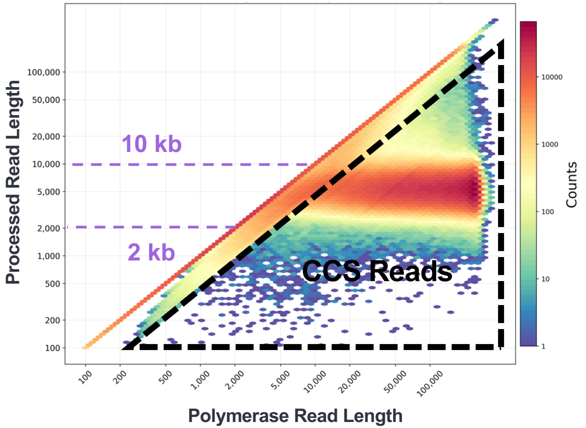


Most Insert Read Lengths are centered around ~1 kb – 10 kb, consistent with the expected size of this (non-size selected) Iso-Seq method cDNA library.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
402.67	84.1	59,523	2.35

5 kb Probe-based capture library
24 h Movie time

No size-selection



Most Insert Read Lengths are centered around ~3 kb – 10 kb, consistent with the expected size of this AMPure PB bead-size selected, HiFi target enrichment human gDNA library.

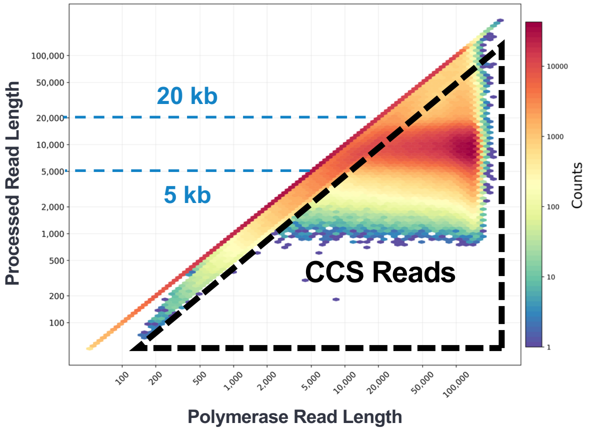
Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
552.72	63.1	109,311	3.00

Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size.

Read length density plot: Example ideal performance (cont.)

8 kb Microbial WGS library
15 h Movie time

No size-selection

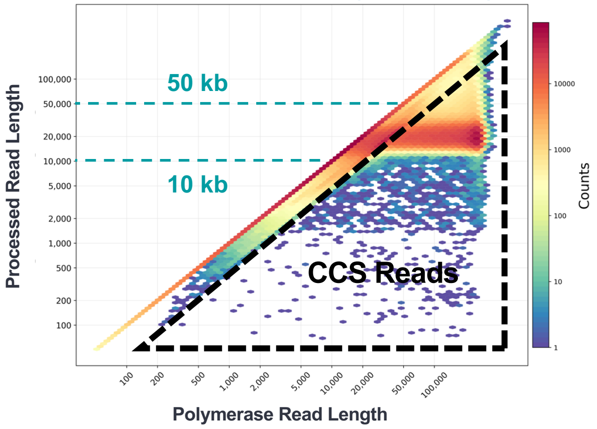


Most Insert Read Lengths are centered around ~5 kb – 15 kb, consistent with the expected size of this (non-size selected) multiplexed microbial gDNA library.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
313.97	72.0	54,428	2.5

18 kb Human WGS library
30 h Movie time

>10 kb size-selection (BluePippin)

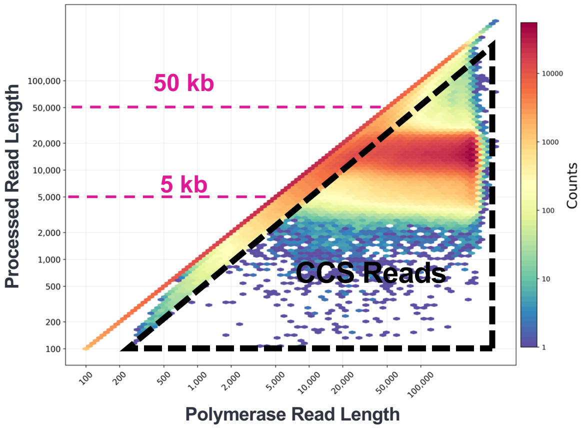


Most Insert Read Lengths are centered around ~10 kb – 25 kb, consistent with the expected size of this >10 kb BluePippin size-selected human gDNA library.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
418.85	70.5	74,108	2.70

16 kb Human WGS library
30 h Movie time

AMPure PB bead size-selection



Most Insert Read Lengths are centered around ~5 kb – 25 kb, consistent with the expected size of this AMPure PB bead size-selected human gDNA library.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
655.94	76.3	107,306	2.88

PacBio

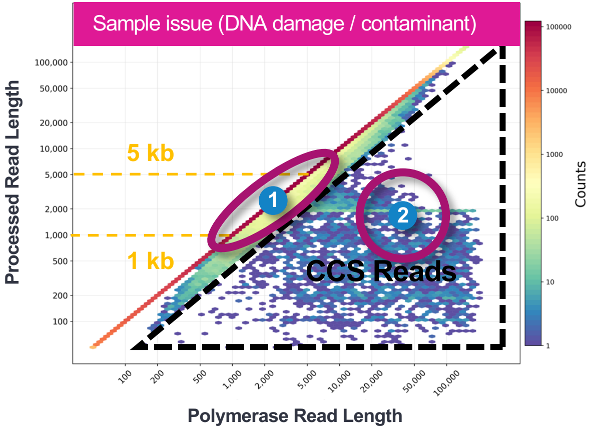
Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size.

Read length density plot: Example suboptimal performance

3 kb Amplicon library
20 h Movie time

No size-selection

Sample issue (DNA damage / contaminant)



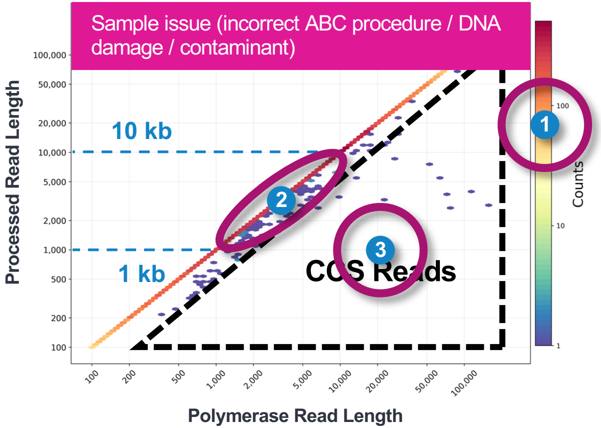
- A high density of reads (red) appears on the primary diagonal line **(1)**, indicating a large proportion of reads terminating in the first pass along the SMRTbell template – thus leading to short pol RL and lower than ideal CCS read density at the target insert size **(2)**.
- Sample also shows a low base rate for Polymerase 2.1 and low total raw bases.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
38.81	75.6	6,401	0.69

1.5 kb Iso-Seq library
30 h Movie time

No size-selection

Sample issue (incorrect ABC procedure / DNA damage / contaminant)



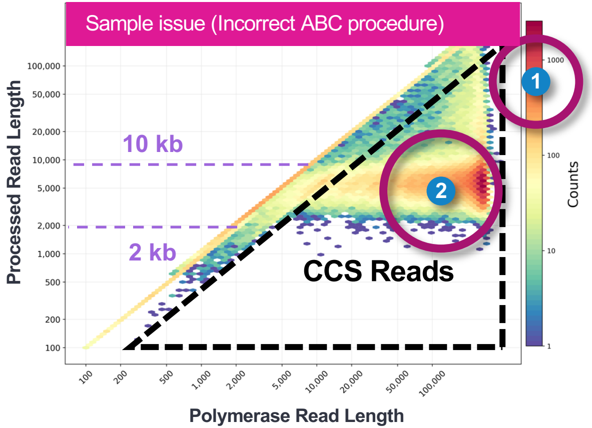
- Sample shows a very low overall density of *P1* counts **(1)**, short pol RL (due to a high relative number of reads terminating in the first pass **(2)**), and almost no CCS reads generated **(3)**.
- Sample also shows a low base rate for Polymerase 2.1 and low total raw bases.

Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
0.24	0.5	10,903	1.37

5 kb Probe-based Capture library
24 h Movie time

No size-selection

Sample issue (Incorrect ABC procedure)



- Sample shows a very low overall density of *P1* counts **(1)** – thus leading toa lower than ideal CCS read density at the target insert size **(2)**.
- Sample also shows low total raw bases due to the low overall *P1* count

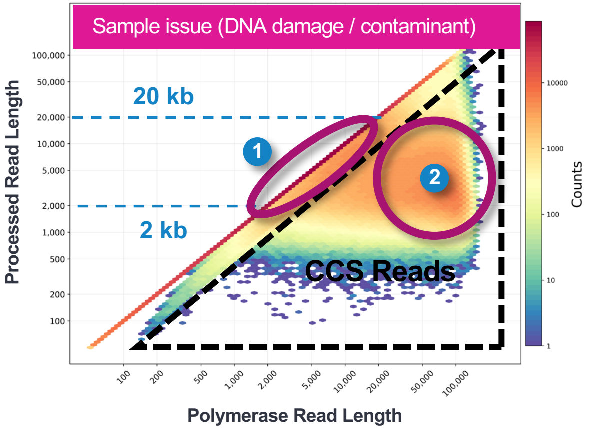
Total raw bases (Gb)	*%P1*	Pol RL (bp)	Base rate
16.19	1.5	147,837	2.96

Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size.

Read length density plot: Example suboptimal performance (cont.)

6 kb Microbial WGS library
15 h Movie time

No size-selection

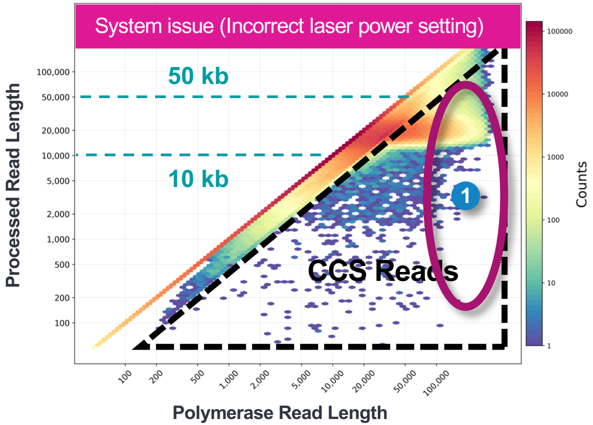

Sample issue (DNA damage / contaminant)

- Sample shows a short mean pol RL (due to a high relative number of reads terminating in the first pass (1)) and lower than ideal CCS read density at the target insert size (2) with a substantial presence of insert read lengths ≤2 kb, suggesting a low-quality gDNA sample.
- Sample also shows a low base rate for Polymerase 2.2 and a low total raw base yield.

Total raw bases (Gb)	%P1	Pol RL (bp)	Base rate
137.80	80.68	21,396	1.21

18 kb Human WGS library
30 h Movie time

>10 kb size-selection (BluePippin)

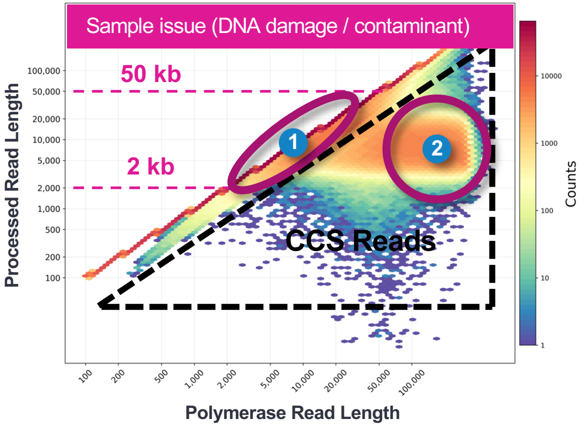

System issue (Incorrect laser power setting)

- Density of CCS reads noticeably decreases as pol RL increases (1) due to a short mean pol RL value.
- Sample also shows a low total raw base yield.

Total raw bases (Gb)	%P1	Pol RL (bp)	Base rate
134.78	78.8	21,327	2.53

18 kb Plant WGS library
30 h Movie time

AMPure PB bead-size selection


Sample issue (DNA damage / contaminant)

- A high proportion of reads are terminating in the first pass (1), thus leading to a shorter than ideal mean pol RL and lower than ideal CCS read density at the target insert size (2).
- Sample also shows a lower than ideal base rate for Polymerase 2.2 and a low total raw base yield.

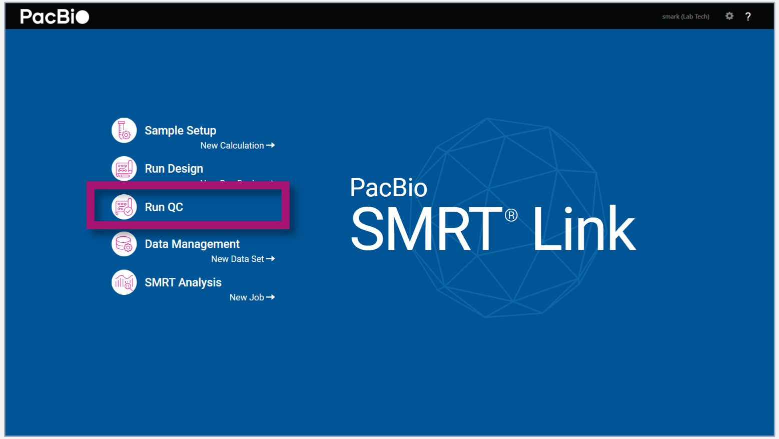
Total raw bases (Gb)	%P1	Pol RL (bp)	Base rate
215.48	63.8	42,093	2.08

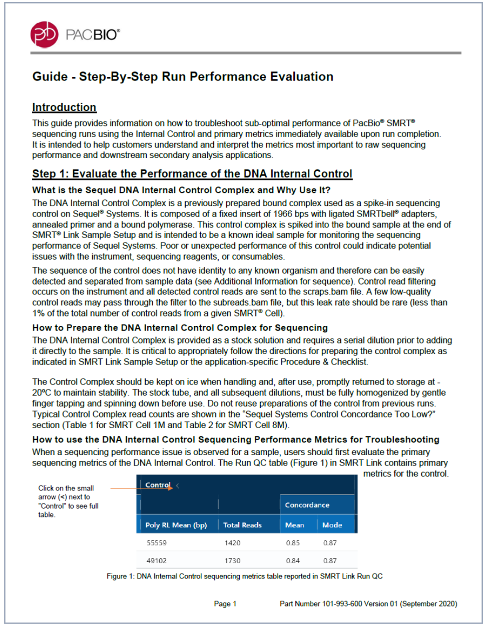
Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size.

Run QC report interpretation & example case studies

The data yields achievable through SMRT sequencing and the diverse number of applications available highlight how important the quality control of a sequencing run is before starting any bioinformatic analysis.

- This section describes how to use PacBio's SMRT Link Run QC and Data Management reports to evaluate primary analysis metrics and overall sequencing performance trends for your sample library
- For more detailed sequencing performance troubleshooting guidance, refer to troubleshooting resources available on PacBio's Documentation website.





PACBIO®

Guide - Step-By-Step Run Performance Evaluation

Introduction

This guide provides information on how to troubleshoot sub-optimal performance of PacBio® SMRT® sequencing runs using the Internal Control and primary metrics immediately available upon run completion. It is intended to help customers understand and interpret the metrics most important to raw sequencing performance and downstream secondary analysis applications.

Step 1: Evaluate the Performance of the DNA Internal Control

What is the Sequel DNA Internal Control Complex and Why Use It?

The DNA Internal Control Complex is a previously prepared bound complex used as a spike-in sequencing control on Sequel® Systems. It is composed of a fixed insert of 1966 bps with ligated SMRTbell® adapters, annealed primer and a bound polymerase. This control complex is spiked into the bound sample at the end of SMRT® Link Sample Setup and is intended to be a known ideal sample for monitoring the sequencing performance of Sequel Systems. Poor or unexpected performance of this control could indicate potential issues with the instrument, sequencing reagents, or consumables.

The sequence of the control does not have identity to any known organism and therefore can be easily detected and separated from sample data (see Additional Information for sequence). Control read filtering occurs on the instrument and all detected control reads are sent to the scraps.bam file. A few low-quality control reads may pass through the filter to the subreads.bam file, but this leak rate should be rare (less than 1% of the total number of control reads from a given SMRT® Cell).

How to Prepare the DNA Internal Control Complex for Sequencing

The DNA Internal Control Complex is provided as a stock solution and requires a serial dilution prior to adding it directly to the sample. It is critical to appropriately follow the directions for preparing the control complex as indicated in SMRT Link Sample Setup or the application-specific Procedure & Checklist.

The Control Complex should be kept on ice when handling and, after use, promptly returned to storage at -20ºC to maintain stability. The stock tube, and all subsequent dilutions, must be fully homogenized by gentle finger tapping and spinning down before use. Do not reuse preparations of the control from previous runs. Typical Control Complex read counts are shown in the "Sequel Systems Control Concordance Too Low?" section (Table 1 for SMRT Cell 1M and Table 2 for SMRT Cell 8M).

How to use the DNA Internal Control Sequencing Performance Metrics for Troubleshooting

When a sequencing performance issue is observed for a sample, users should first evaluate the primary sequencing metrics of the DNA Internal Control. The Run QC table (Figure 1) in SMRT Link contains primary metrics for the control.

Click on the small arrow (<) next to "Control" to see full table.

Control <			
		Concordance	
Poly RL Mean (bp)	Total Reads	Mean	Mode
55559	1420	0.85	0.87
49102	1730	0.84	0.87

Figure 1: DNA Internal Control sequencing metrics table reported in SMRT Link Run QC

Page 1 Part Number 101-993-600 Version 01 (September 2020)

Step-By-Step Run Performance Evaluation Guide (101-993-600) provides information on how to troubleshoot sub-optimal sequencing performance using the DNA Internal control and primary metrics available through SMRT Link Run QC.

SMRT sequencing troubleshooting

Troubleshooting guidance summary table for samples showing poor sequencing performance

Symptom	Potential causes	Possible actions / solutions
Sample shows short polymerase read length (DNA internal control sequencing performance is normal)	Excessively high sample OPLC	Reduce sample OPLC
	Sample quality issue (e.g., highly fragmented DNA, high amount of DNA damage, presence of a contaminant)	Confirm sample QC; re-purify sample (consider changing methodology)
	Incorrect or insufficient pre-extension time specified in Run Design	Use SMRT Link recommended or longer pre-extension time setting
	Inefficient adapter ligation reaction	Verify adapter ligation conditions used during SMRTbell library construction; reperform ligation reaction step if needed
Sample shows low *P1* productivity metric (DNA internal control sequencing performance is normal)	Manual pipetting error during primer annealing, polymerase binding or complex cleanup (ABC) steps	Redo sample ABC steps
	Sample quality issue (e.g., highly fragmented DNA, high amount of DNA damage, presence of a contaminant)	Confirm sample QC (DNA molecular weight, purity and concentration); re-purify sample (consider changing methodology)
Sample shows high *P2* productivity metric (DNA internal control sequencing performance is normal)	Excessively high sample OPLC	Reduce OPLC
	Sample quality issue (e.g., contaminant)	Confirm sample QC (DNA purity); re-purify sample (consider changing methodology)
	Excessively high unbound polymerase carryover	Redo polymerase binding step
	Use of non-recommended plastic consumables (leaching of fluorescent contaminants)	Use recommended plastic consumables for all sample extraction/purification, library construction and sequencing preparation steps
Insert read length density plot shows larger than expected library insert size (DNA internal control sequencing performance is normal)	Concatemers present due to inefficient adapter ligation reaction	Verify adapter ligation conditions used during SMRTbell library construction; redo ligation reaction step if needed
	Sample quality issue (e.g., incorrect library insert size)	Confirm sample QC and molarity (DNA molecular weight); redo sample size selection step

Refer to **Step-By-Step Run Performance Evaluation Guide** (101-993-600) for more detailed information about troubleshooting SMRT sequencing performance issues.

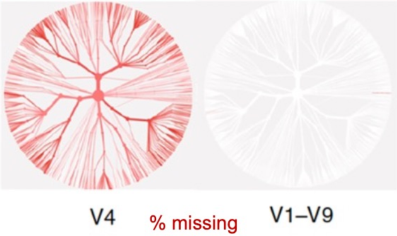
PacBio HiFi Sequencing for High-Resolution Microbiome Research

4 July 2023

彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

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

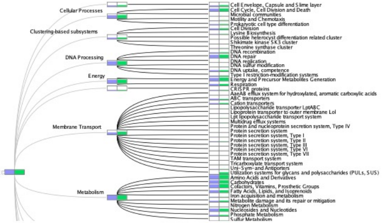
Full-length 16S sequencing



Species/strain-level resolution

Reveals true sample diversity

Shotgun metagenome profiling

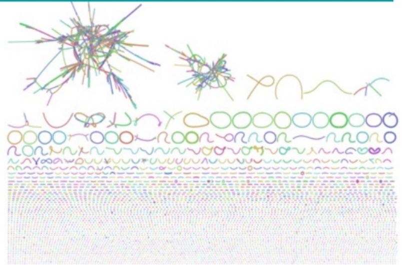

Cellular Processes
Clustering-based subsystems
DNA Processing
Energy
Membrane Transport
Metabolism
Cell Envelope, Capsule and Slime layer
Cell Cycle, Cell Division and Death
Microbial communities
Motility and Chemotaxis
Prokaryotic cell type differentiation
Cell Division
Lysine Biosynthesis
Possible heterocyst differentiation related cluster
Shikimate kinase SK3 cluster
Threonine synthase cluster
DNA recombination
DNA repair
DNA replication
DNA sulfur modification
DNA uptake, competence
Type I restriction-modification systems
Energy and Precursor Metabolites Generation
Respiration
CRISPR proteins
AaeAB efflux system for hydroxylated, aromatic carboxylic acids
ABC transporters
Cation transporters
Lipopolysaccharide transporter LptABC
Lipoprotein transporter to outer membrane Lol
Lpt lipopolysaccharide transport system
Multidrug efflux systems
Protein and nucleoprotein secretion system, Type IV
Protein secretion system
Protein secretion system, Type I
Protein secretion system, Type II
Protein secretion system, Type III
Protein secretion system, Type VI
Protein secretion system, Type VII
TAM transport system
Tricarboxylate transport system
Uni-Sym-and Antiporters
Utilization systems for glycans and polysaccharides (PULs, SUS)
Amino Acids and Derivatives
Carbohydrates
Cofactors, Vitamins, Prosthetic Groups
Fatty Acids, Lipids, and Isoprenoids
Iron acquisition and metabolism
Metabolite damage and its repair or mitigation
Nitrogen Metabolism
Nucleosides and Nucleotides
Phosphate Metabolism
Sulfur Metabolism

~8 genes per sequence read

~90% of reads with at least 1 gene

Profile taxonomy with high precision and recall

Shotgun metagenome assembly



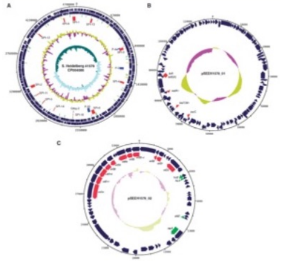
Obtain many high-quality MAGs

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%

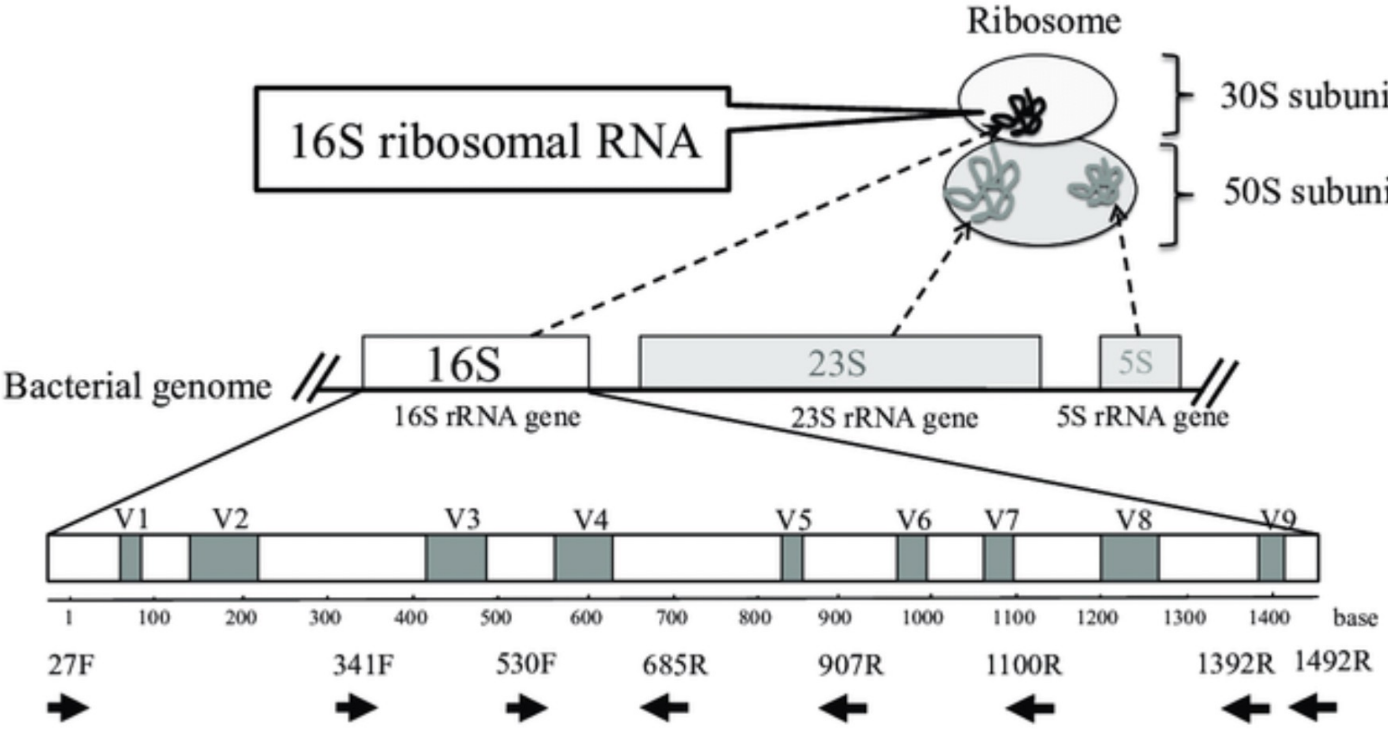
Detection of R-M system motifs

Short-read polishing (hybrid assembly) not needed

PacBio

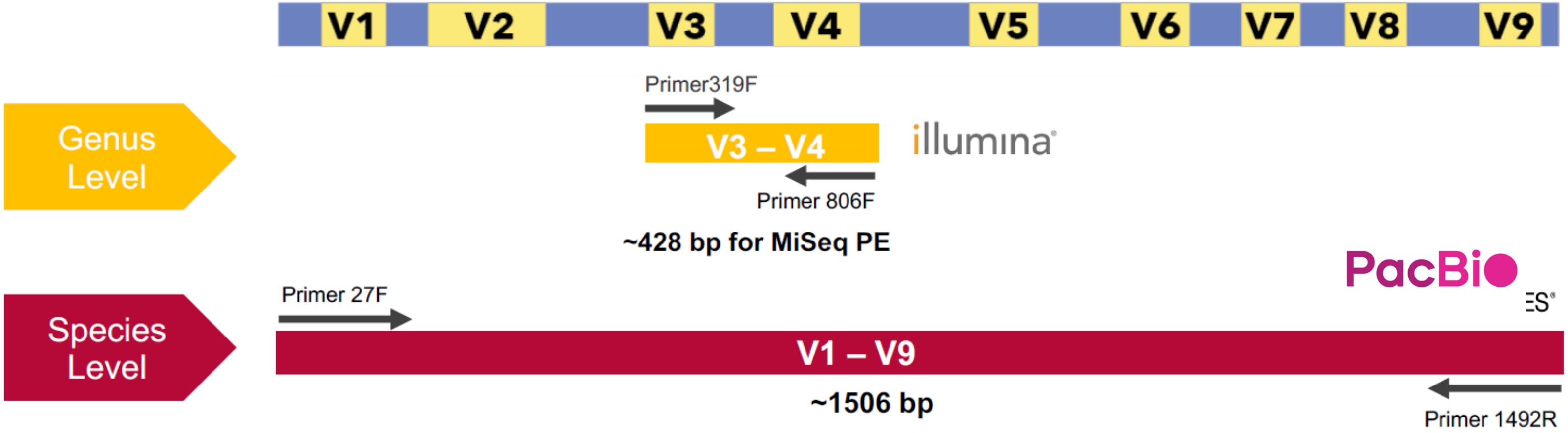
Full-Length 16S Pipeline Overview

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



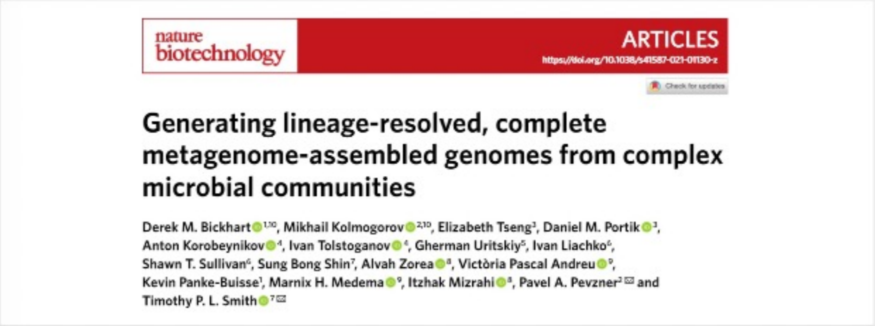
Fukuda K, Ogawa M, Taniguchi H, Saito M. Molecular Approaches to Studying Microbial Communities: Targeting the 16S Ribosomal RNA Gene. J UOEH. 2016 Sep;38(3):223-32. doi: 10.7888/juoeh.38.223. PMID: 27627970.

Amplicons can Target 16s rRNA and Beyond



Longer amplicons enable higher resolution taxonomic identification

PacBio HiFi sequencing is setting a new gold standard in 16S/metagenomics



nature biotechnology | ARTICLES

https://doi.org/10.1038/s41587-021-01130-z

Check for updates

Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities

Derek M. Bickhart[1,10], Mikhail Kolmogorov[2,10], Elizabeth Tseng[3], Daniel M. Portik[3], Anton Korobeynikov[4], Ivan Tolstoganov[4], Gherman Uritskiy[5], Ivan Liachko[6], Shawn T. Sullivan[6], Sung Bong Shin[7], Alvah Zorea[8], Victòria Pascal Andreu[9], Kevin Panke-Buisse[1], Marnix H. Medema[9], Itzhak Mizrahi[8], Pavel A. Pevzner[2] ✉ and Timothy P. L. Smith[7] ✉

genomeweb

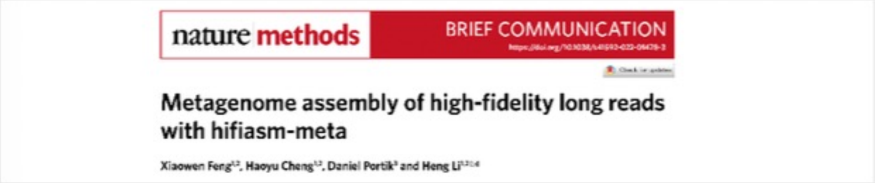
Tech-Boosted Genome Assembly Helps Resolve Closely Related Microbes in Metagenomic Sample

Jan 03, 2022 | Andrew P. Han

MICROBIAL GENOMICS

MAGs achieve lineage resolution

Taylor E. Reiter & C. Titus Brown ✉



nature methods | BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01478-3

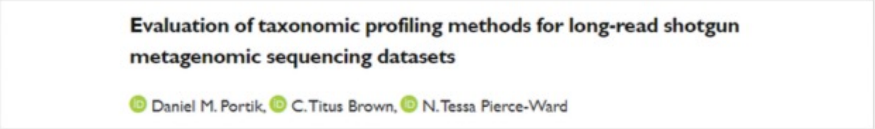
Check for updates

Metagenome assembly of high-fidelity long reads with hifiasm-meta

Xiaowen Feng[1,2], Haoyu Cheng[1,2], Daniel Portik[3] and Heng Li[1,2] ✉[4]

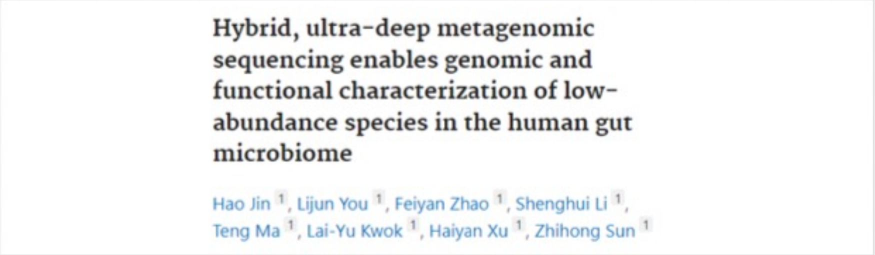
HiFi Metagenomic Sequencing Enables Assembly of Accurate and Complete Genomes from Human Gut Microbiota

Chan Yeong Kim, Junyeong Ma, Insuk Lee



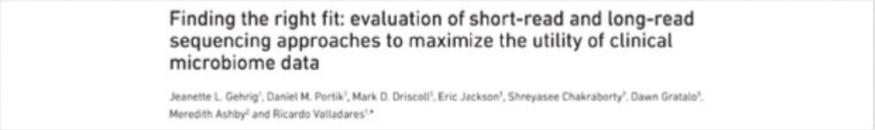
Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets

Daniel M. Portik, C. Titus Brown, N. Tessa Pierce-Ward



Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome

Hao Jin[1], Lijun You[1], Feiyan Zhao[1], Shenghui Li[1], Teng Ma[1], Lai-Yu Kwok[1], Haiyan Xu[1], Zhihong Sun[1]

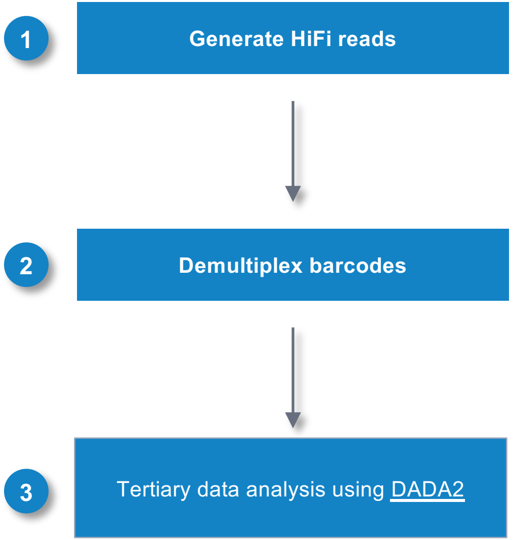


Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data

Jeanette L. Gehrig[1], Daniel M. Portik[2], Mark D. Driscoll[1], Eric Jackson[1], Shreyasee Chakraborty[2], Dawn Gratalo[3], Meredith Ashby[2] and Ricardo Valladares[1,*]

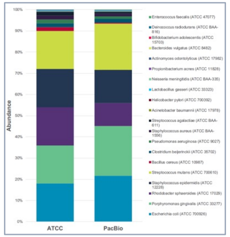
https://www.nature.com/articles/s41587-021-01130-z; https://www.nature.com/articles/s41587-021-01130-z; https://www.nature.com/articles/s41564-021-01027-2; https://www.genomeweb.com/sequencing/tech-boosted-genome-assembly-helps-resolve- closely-related-microbes-metagenomic-sample; https://www.tandfonline.com/doi/full/10.1080/19490976.2021.2021790; https://www.biorxiv.org/content/10.1101/2022.01.31.478527v1?ct=; https://www.biorxiv.org/content/10.1101/2022.02.09.479829v1; https://www.nature.com/articles/s41592-022-01478-3; https://pubmed.ncbi.nlm.nih.gov/35302439/

16S Data Analysis Workflow Recommendations


1 Generate HiFi reads

2 Demultiplex barcodes

3 Tertiary data analysis using DADA2



1. **Perform CCS analysis** on-instrument (Sequel IIe system only) or in SMRT Link to generate highly accurate (≥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

DADA2 Amplicon Sequencing. Exactly. Version 1.8

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

qiime2

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

PacBio

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.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

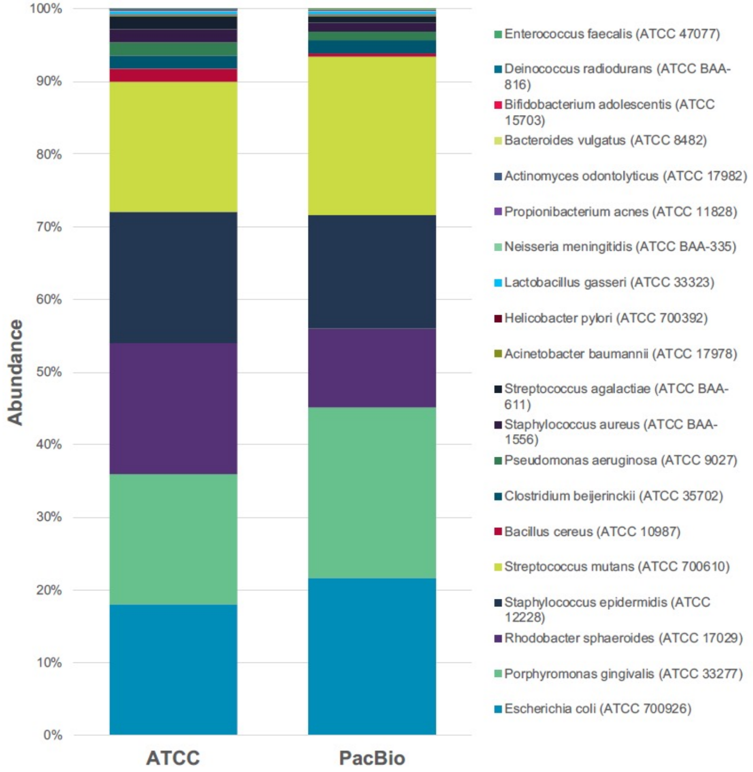
ATCC®

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

https://www.atcc.org/products/all/MSA-1003.aspx

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

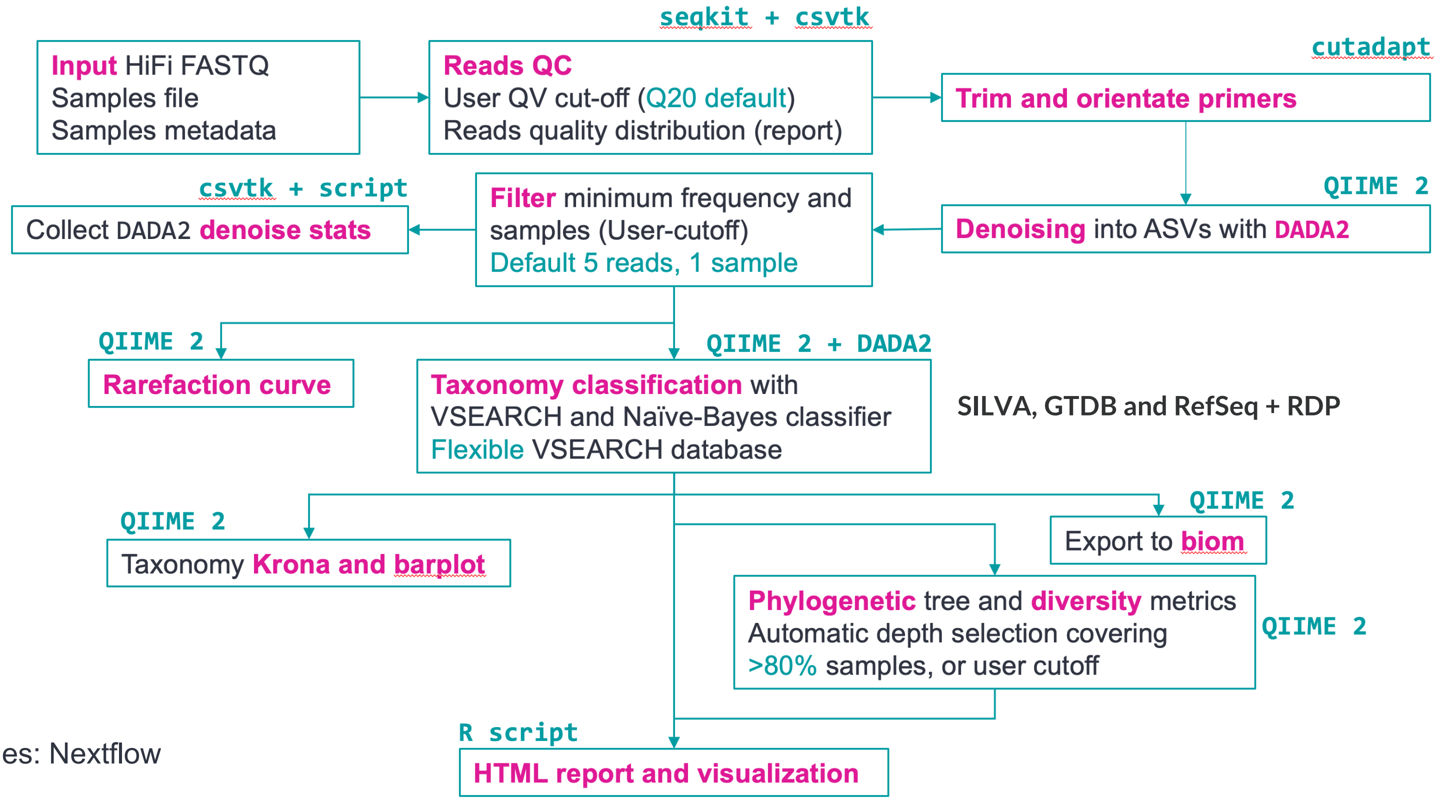
Yield of >99% accurate 16S reads matches the expected composition of the MSA-1003 mock community sample

GC content ranging from 30 ~ 69% can be identified

Download and explore this 16S HiFi dataset further

Full-length (V1-V9) 16S amplicon samples were pooled at 96-Plex and sequenced on a single SMRT Cell 8M (Sequel II System Chemistry 2.0). PacBio results shown in bar graph reflect the average abundance values derived from the pooled MSA-1003 replicate samples.

pb-16S-nf overview



seqkit + csvtk

cutadapt

Input HiFi FASTQ
Samples file
Samples metadata

Reads QC
User QV cut-off (Q20 default)
Reads quality distribution (report)

Trim and orientate primers

QIIME 2

Denoising into ASVs with **DADA2**

Filter minimum frequency and samples (User-cutoff)
Default 5 reads, 1 sample

csvtk + script

Collect DADA2 **denoise stats**

QIIME 2

Rarefaction curve

QIIME 2 + DADA2

Taxonomy classification with VSEARCH and Naïve-Bayes classifier
Flexible VSEARCH database

SILVA, GTDB and RefSeq + RDP

QIIME 2

Taxonomy **Krona and barplot**

QIIME 2

Export to **biom**

Phylogenetic tree and **diversity** metrics
Automatic depth selection covering >80% samples, or user cutoff

QIIME 2

R script

HTML report and visualization

Languages: Nextflow

Compatible with : Conda, Singularity, Docker

Step-by-step guideline

1. Clone repository:

```
git clone https://github.com/PacificBiosciences/pb-16S-nf.git
```

2. Install Anaconda/Miniconda and Nextflow:

```
conda install mamba -n base -c conda-forge
conda install -c bioconda nextflow
```

3. Download Databases:

```
nextflow run main.nf -download_db
```

4. Run pipeline:

```
nextflow run main.nf -input sample.tsv \
--metadata metadata.tsv \
-profile conda \
--outdir results
```

If using Docker, just add "`-profile docker`".

Modify "nextflow.config" to utilizes HPC job scheduler if desirable

Input & Metadata

A file giving a sample name for each of the FASTQ file that we are going to analyze.

```
# PB_sample.tsv
sample-id absolute-filepath
A-1 /home/smrtuser/16Sdata/BC2079_5p--BC2038_3p.hifi_reads.fastq.gz
Z-1 /home/smrtuser/16Sdata/BC2080_5p--BC2038_3p.hifi_reads.fastq.gz
A-4 /home/smrtuser/16Sdata/BC2079_5p--BC2076_3p.hifi_reads.fastq.gz
Z-4 /home/smrtuser/16Sdata/BC2080_5p--BC2076_3p.hifi_reads.fastq.gz
```

And a file giving the status/info/condition of the sample

```
#PB_metadata.tsv
sample_name condition
A-1 Control_A
Z-1 Control_Z
A-4 Control_A
Z-4 Control_Z
```

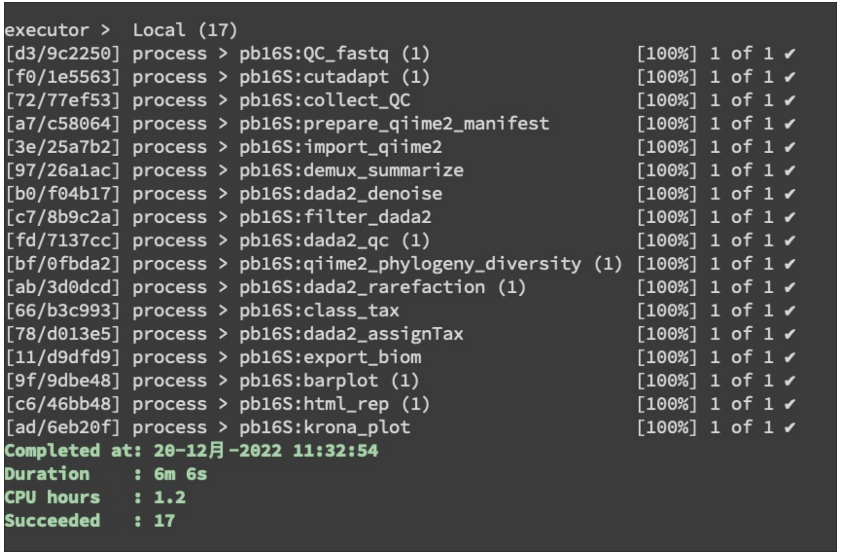
```
$nextflow run main.nf \
--input PB_sample.tsv  --metadata PB_metadata.tsv -profile conda \
--dada2_cpu 80 --vsearch_cpu 80 \
--outdir PB_16S_2023-03
```

Parameters tuning

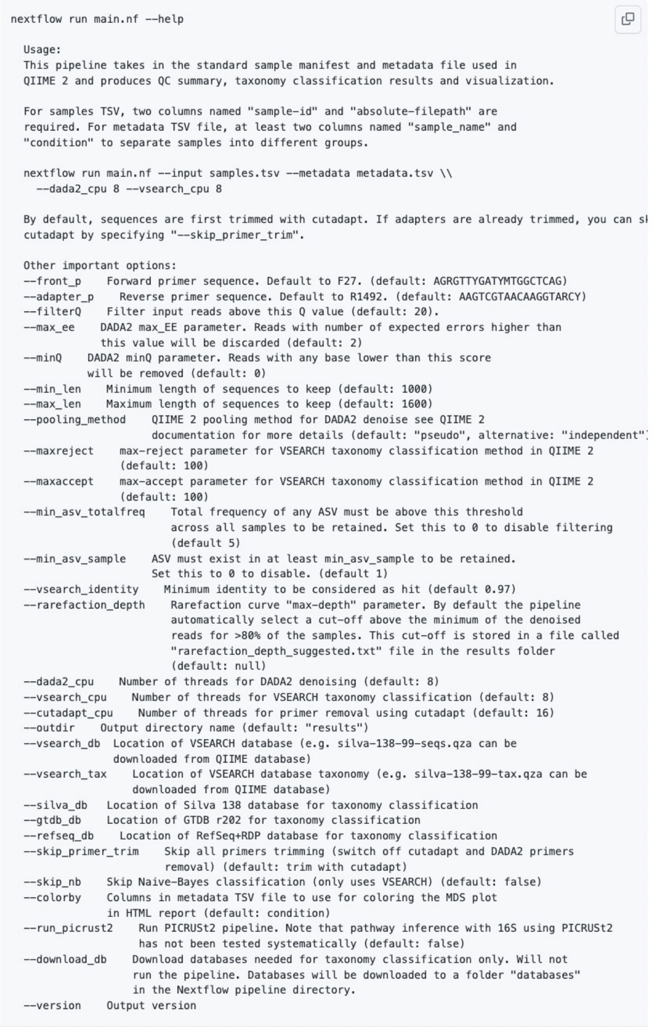
Parameters can be changed when running the pipeline, e.g. to change the default quality filter threshold to Q30:

```
nextflow run main.nf -input sample.tsv \
--filterQ 30
```

Pipeline will report progress:



```
executor >  Local (17)
[d3/9c2250] process > pb16S:QC_fastq (1)                   [100%] 1 of 1 ✔
[f0/1e5563] process > pb16S:cutadapt (1)                   [100%] 1 of 1 ✔
[72/77ef53] process > pb16S:collect_QC                     [100%] 1 of 1 ✔
[a7/c58064] process > pb16S:prepare_qiime2_manifest        [100%] 1 of 1 ✔
[3e/25a7b2] process > pb16S:import_qiime2                  [100%] 1 of 1 ✔
[97/26a1ac] process > pb16S:demux_summarize                [100%] 1 of 1 ✔
[b0/f04b17] process > pb16S:dada2_denoise                  [100%] 1 of 1 ✔
[c7/8b9c2a] process > pb16S:filter_dada2                   [100%] 1 of 1 ✔
[fd/7137cc] process > pb16S:dada2_qc (1)                   [100%] 1 of 1 ✔
[bf/0fbda2] process > pb16S:qiime2_phylogeny_diversity (1) [100%] 1 of 1 ✔
[ab/3d0dcd] process > pb16S:dada2_rarefaction (1)          [100%] 1 of 1 ✔
[66/b3c993] process > pb16S:class_tax                      [100%] 1 of 1 ✔
[78/d013e5] process > pb16S:dada2_assignTax                [100%] 1 of 1 ✔
[11/d9dfd9] process > pb16S:export_biom                    [100%] 1 of 1 ✔
[9f/9dbe48] process > pb16S:barplot (1)                    [100%] 1 of 1 ✔
[c6/46bb48] process > pb16S:html_rep (1)                   [100%] 1 of 1 ✔
[ad/6eb20f] process > pb16S:krona_plot                     [100%] 1 of 1 ✔
Completed at: 20-12月-2022 11:32:54
Duration    : 6m 6s
CPU hours   : 1.2
Succeeded   : 17
```



```
nextflow run main.nf --help

  Usage:
  This pipeline takes in the standard sample manifest and metadata file used in
  QIIME 2 and produces QC summary, taxonomy classification results and visualization.

  For samples TSV, two columns named "sample-id" and "absolute-filepath" are
  required. For metadata TSV file, at least two columns named "sample_name" and
  "condition" to separate samples into different groups.

  nextflow run main.nf --input samples.tsv --metadata metadata.tsv \\
    --dada2_cpu 8 --vsearch_cpu 8

  By default, sequences are first trimmed with cutadapt. If adapters are already trimmed, you can sk
  cutadapt by specifying "--skip_primer_trim".

  Other important options:
  --front_p    Forward primer sequence. Default to F27. (default: AGRGTTYGATYMTGGCTCAG)
  --adapter_p    Reverse primer sequence. Default to R1492. (default: AAGTCGTAACAAGGTARCY)
  --filterQ    Filter input reads above this Q value (default: 20).
  --max_ee    DADA2 max_EE parameter. Reads with number of expected errors higher than
              this value will be discarded (default: 2)
  --minQ    DADA2 minQ parameter. Reads with any base lower than this score
            will be removed (default: 0)
  --min_len    Minimum length of sequences to keep (default: 1000)
  --max_len    Maximum length of sequences to keep (default: 1600)
  --pooling_method    QIIME 2 pooling method for DADA2 denoise see QIIME 2
                      documentation for more details (default: "pseudo", alternative: "independent")
  --maxreject    max-reject parameter for VSEARCH taxonomy classification method in QIIME 2
                 (default: 100)
  --maxaccept    max-accept parameter for VSEARCH taxonomy classification method in QIIME 2
                 (default: 100)
  --min_asv_totalfreq    Total frequency of any ASV must be above this threshold
                         across all samples to be retained. Set this to 0 to disable filtering
                         (default 5)
  --min_asv_sample    ASV must exist in at least min_asv_sample to be retained.
                      Set this to 0 to disable. (default 1)
  --vsearch_identity    Minimum identity to be considered as hit (default 0.97)
  --rarefaction_depth    Rarefaction curve "max-depth" parameter. By default the pipeline
                         automatically select a cut-off above the minimum of the denoised
                         reads for >80% of the samples. This cut-off is stored in a file called
                         "rarefaction_depth_suggested.txt" file in the results folder
                         (default: null)
  --dada2_cpu    Number of threads for DADA2 denoising (default: 8)
  --vsearch_cpu    Number of threads for VSEARCH taxonomy classification (default: 8)
  --cutadapt_cpu    Number of threads for primer removal using cutadapt (default: 16)
  --outdir    Output directory name (default: "results")
  --vsearch_db  Location of VSEARCH database (e.g. silva-138-99-seqs.qza can be
                downloaded from QIIME database)
  --vsearch_tax    Location of VSEARCH database taxonomy (e.g. silva-138-99-tax.qza can be
                   downloaded from QIIME database)
  --silva_db   Location of Silva 138 database for taxonomy classification
  --gtdb_db    Location of GTDB r202 for taxonomy classification
  --refseq_db    Location of RefSeq+RDP database for taxonomy classification
  --skip_primer_trim    Skip all primers trimming (switch off cutadapt and DADA2 primers
                        removal) (default: trim with cutadapt)
  --skip_nb    Skip Naive-Bayes classification (only uses VSEARCH) (default: false)
  --colorby    Columns in metadata TSV file to use for coloring the MDS plot
               in HTML report (default: condition)
  --run_picrust2    Run PICRUSt2 pipeline. Note that pathway inference with 16S using PICRUSt2
                    has not been tested systematically (default: false)
  --download_db    Download databases needed for taxonomy classification only. Will not
                   run the pipeline. Databases will be downloaded to a folder "databases"
                   in the Nextflow pipeline directory.
  --version    Output version
```

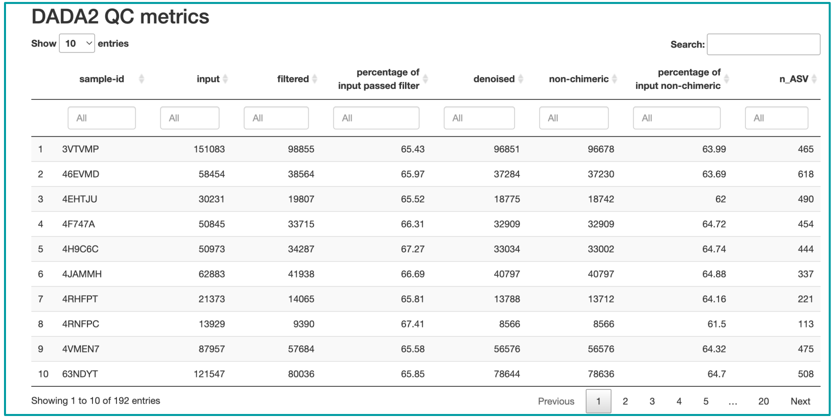
PacBio

Results from pb-16S-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



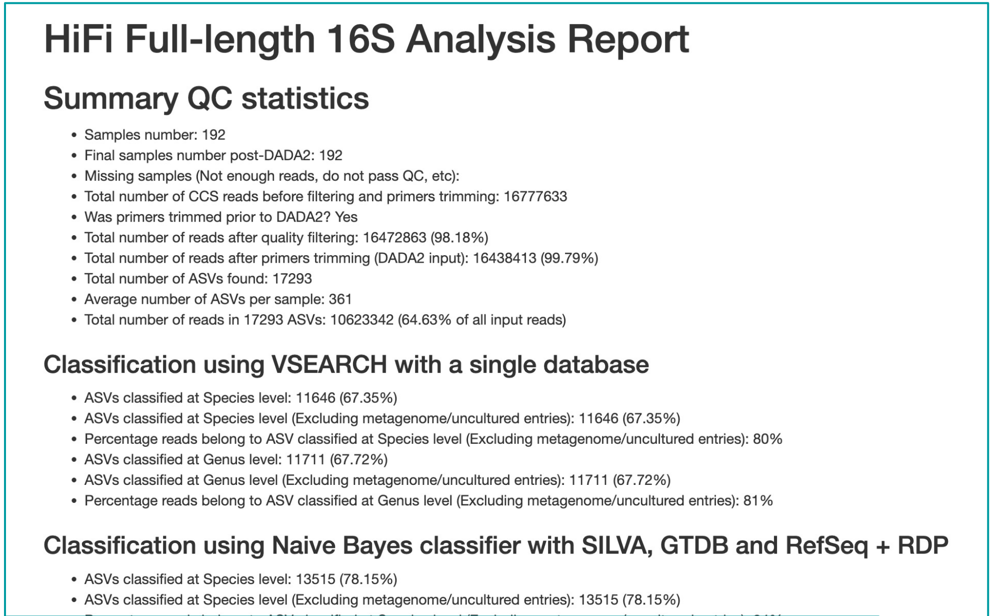
DADA2 QC metrics

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	3VTVMP	151083	98855	65.43	96851	96678	63.99	465
2	46EVMD	58454	38564	65.97	37284	37230	63.69	618
3	4EHTJU	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	66.69	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



HiFi Full-length 16S Analysis Report

Summary QC statistics

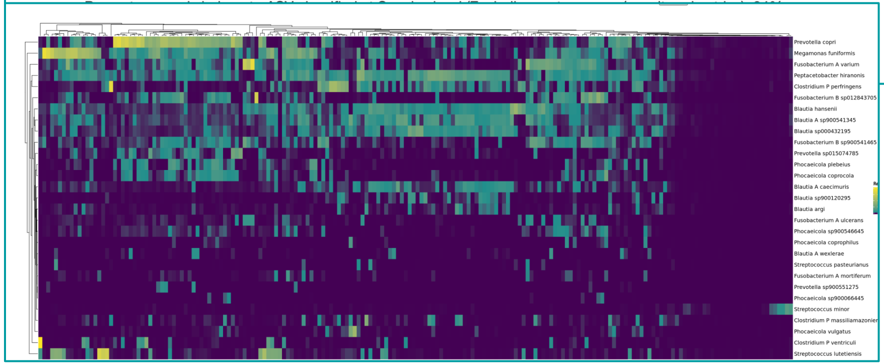
- Samples number: 192
- Final samples number post-DADA2: 192
- Missing samples (Not enough reads, do not pass QC, etc):
- Total number of CCS reads before filtering and primers trimming: 16777633
- Was primers trimmed prior to DADA2? Yes
- Total number of reads after quality filtering: 16472863 (98.18%)
- 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

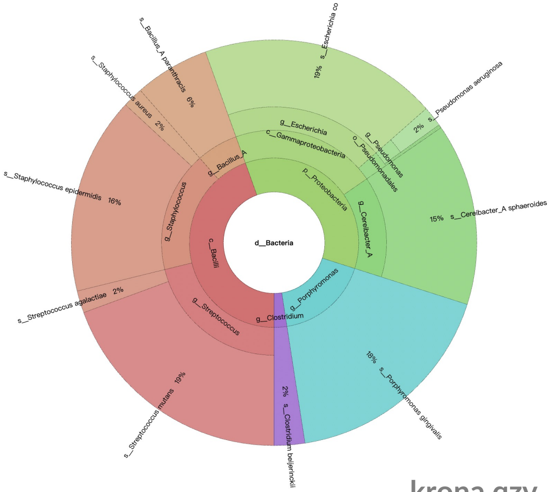
- ASVs classified at Species level: 13515 (78.15%)
- ASVs classified at Species level (Excluding metagenome/uncultured entries): 13515 (78.15%)



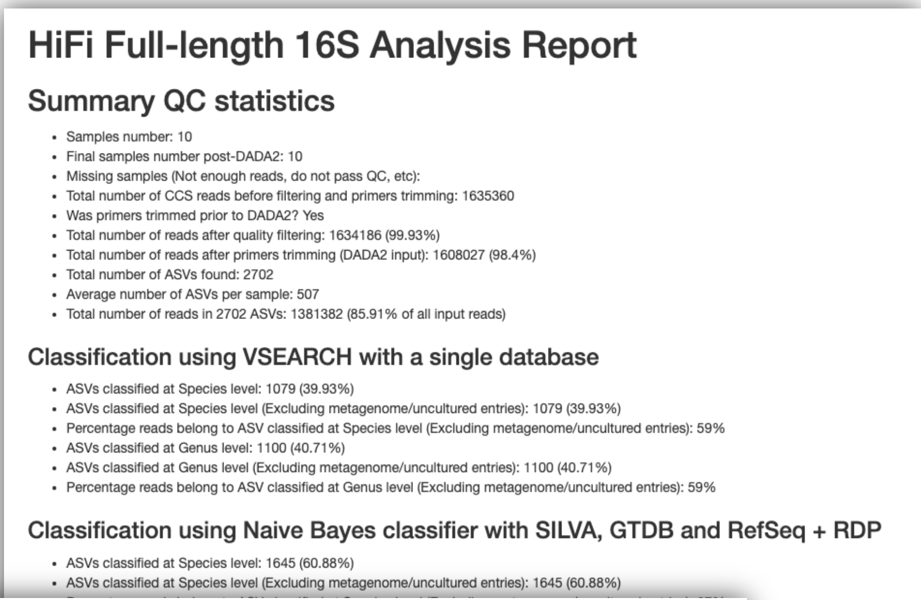
https://github.com/PacificBiosciences/pb-16S-nf

Results from pb-16S-nf pipeline

- HTML report provides useful metrics and visualizations
- Important outputs are in QIIME2-compatible format and TSV format for easy importing
- Outputs documentation: https://github.com/PacificBiosciences/pb-16S-nf



krona.qzv



HiFi Full-length 16S Analysis Report

Summary QC statistics

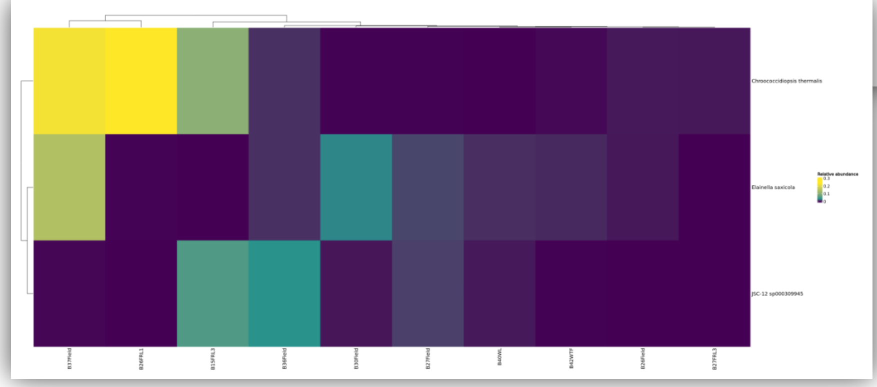
- Samples number: 10
- Final samples number post-DADA2: 10
- Missing samples (Not enough reads, do not pass QC, etc):
- Total number of CCS reads before filtering and primers trimming: 1635360
- Was primers trimmed prior to DADA2? Yes
- Total number of reads after quality filtering: 1634186 (99.93%)
- Total number of reads after primers trimming (DADA2 input): 1608027 (98.4%)
- Total number of ASVs found: 2702
- Average number of ASVs per sample: 507
- Total number of reads in 2702 ASVs: 1381382 (85.91% of all input reads)

Classification using VSEARCH with a single database

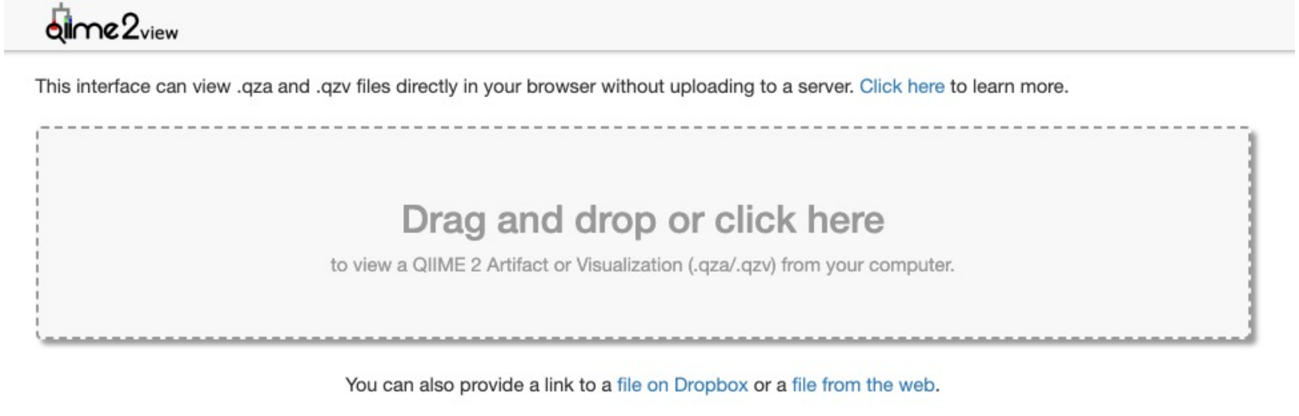
- ASVs classified at Species level: 1079 (39.93%)
- ASVs classified at Species level (Excluding metagenome/uncultured entries): 1079 (39.93%)
- Percentage reads belong to ASV classified at Species level (Excluding metagenome/uncultured entries): 59%
- ASVs classified at Genus level: 1100 (40.71%)
- ASVs classified at Genus level (Excluding metagenome/uncultured entries): 1100 (40.71%)
- Percentage reads belong to ASV classified at Genus level (Excluding metagenome/uncultured entries): 59%

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

- ASVs classified at Species level: 1645 (60.88%)
- ASVs classified at Species level (Excluding metagenome/uncultured entries): 1645 (60.88%)



Results from pb-16S-nf pipeline

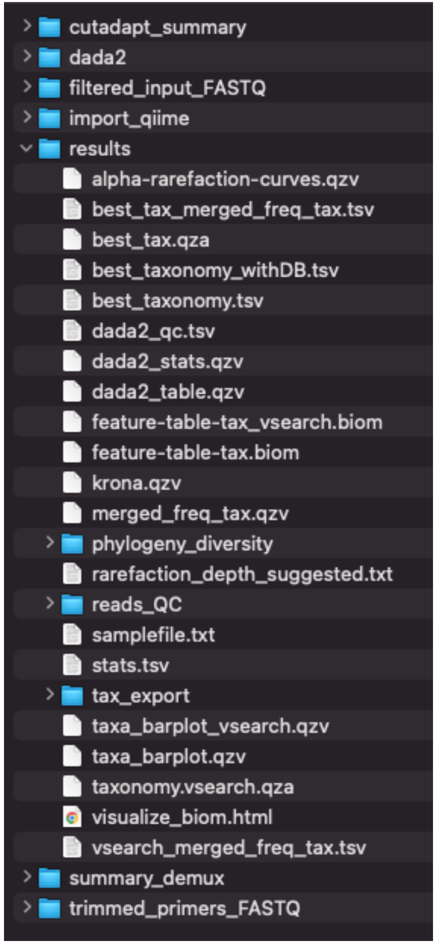


This interface can view .qza and .qzv files directly in your browser without uploading to a server. Click here to learn more.

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
v 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
```

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[1]	891	8.3m	5417	5104663 (62%)	**87%**	**91%**	2.5h	34 GB
Gut[2]	192	2.2m	1593	996965 (45%)	**96%**	**99%**	2h	30 GB
Animal gut[3]	192	16.7m	17293	10623342 (65%)	**67%***	**81%**	13h	87 GB
Animal gut[3]	192	2.2m (99.3%)	10917	1789875 (83%)	**70%**	**79%**	5.5h	30 GB
Wastewater full[4]	33	2.14m	11462	1969683 (92%)	**39%***	**63%**	12h	47 GB
Wastewater 10k/sample[5]	33	326k	3974	265137 (82%)	**44%***	**65%**	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.

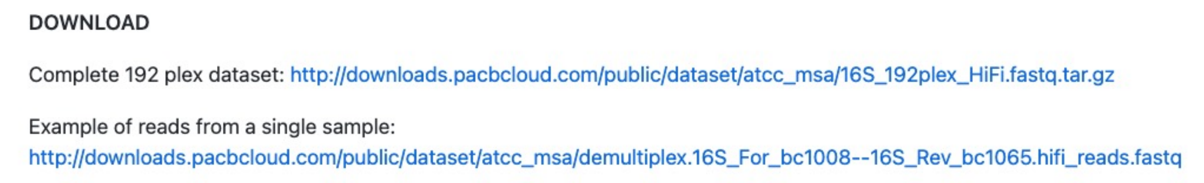
pb-16S-nf analysis

ATCC MSA-1003-16S

Analysis PacBio HiFi Mock Community 16S Data

DEMO SAMPLE

20 Strain Staggered Mix Genomic Material (ATCC® MSA-1003™)



DOWNLOAD

Complete 192 plex dataset: http://downloads.pacbcloud.com/public/dataset/atcc_msa/16S_192plex_HiFi.fastq.tar.gz

Example of reads from a single sample:
http://downloads.pacbcloud.com/public/dataset/atcc_msa/demultiplex.16S_For_bc1008--16S_Rev_bc1065.hifi_reads.fastq

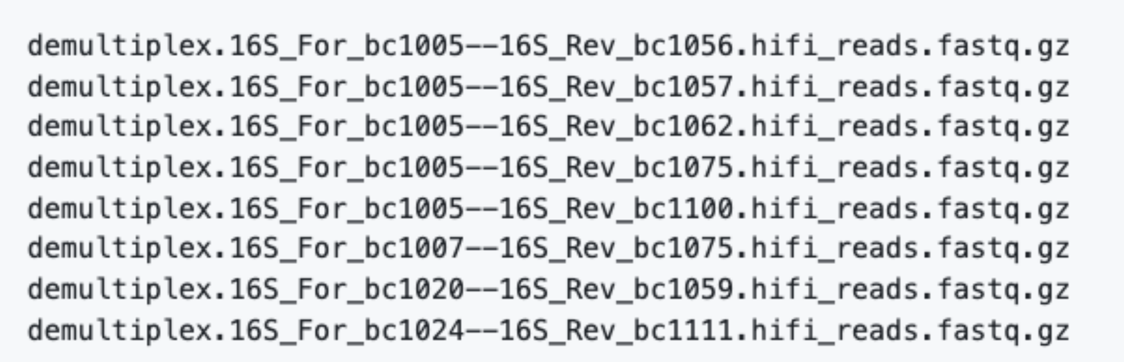
Download from Sequel II System

16S HiFi dataset

METHODS

- 16S protocol with Barcoded Primers
- Library prep: SMRTbell Express Template Prep Kit 2.0
- Sequencing: Sequel II System binding kit
- Run time: 0.5 hour pre-extension; 10 hour movie
- CCS Analysis: SMRT Link v10.0 Circular Consensus Sequencing Application (ccs 5.0.0)

ATCC MSA-1003 Mock Community



```
demultiplex.16S_For_bc1005--16S_Rev_bc1056.hifi_reads.fastq.gz
demultiplex.16S_For_bc1005--16S_Rev_bc1057.hifi_reads.fastq.gz
demultiplex.16S_For_bc1005--16S_Rev_bc1062.hifi_reads.fastq.gz
demultiplex.16S_For_bc1005--16S_Rev_bc1075.hifi_reads.fastq.gz
demultiplex.16S_For_bc1005--16S_Rev_bc1100.hifi_reads.fastq.gz
demultiplex.16S_For_bc1007--16S_Rev_bc1075.hifi_reads.fastq.gz
demultiplex.16S_For_bc1020--16S_Rev_bc1059.hifi_reads.fastq.gz
demultiplex.16S_For_bc1024--16S_Rev_bc1111.hifi_reads.fastq.gz
```

https://github.com/PacificBiosciences/pb-16S-analysis/wiki/Analyzing-PacBio-HiFi-Mock-Community-16S-Data-with-QIIME-2

Input: Sample & Metadata tsv

A file giving a sample name for each of the FASTQ file that we are going to analyze.

```
# pb_sample.tsv
sample-id absolute-filepath
A-1 <path_to_dataset>/demultiplex.16S_For_bc1005--16S_Rev_bc1056.hifi_reads.fastq
A-2 <path_to_dataset>/demultiplex.16S_For_bc1005--16S_Rev_bc1057.hifi_reads.fastq
A-3 <path_to_dataset>/demultiplex.16S_For_bc1005--16S_Rev_bc1062.hifi_reads.fastq
A-4 <path_to_dataset>/demultiplex.16S_For_bc1005--16S_Rev_bc1075.hifi_reads.fastq
A-5 <path_to_dataset>/demultiplex.16S_For_bc1005--16S_Rev_bc1100.hifi_reads.fastq
A-6 <path_to_dataset>/demultiplex.16S_For_bc1007--16S_Rev_bc1075.hifi_reads.fastq
A-7 <path_to_dataset>/demultiplex.16S_For_bc1020--16S_Rev_bc1059.hifi_reads.fastq
A-8 <path_to_dataset>/demultiplex.16S_For_bc1024--16S_Rev_bc1111.hifi_reads.fastq
```

And a file giving the status/info/condition of the sample

```
# pb_metadata.tsv
sample_name condition
A-1 RepA
A-2 RepA
A-3 RepA
A-4 RepA
A-5 RepB
A-6 RepB
A-7 RepB
A-8 RepB
```

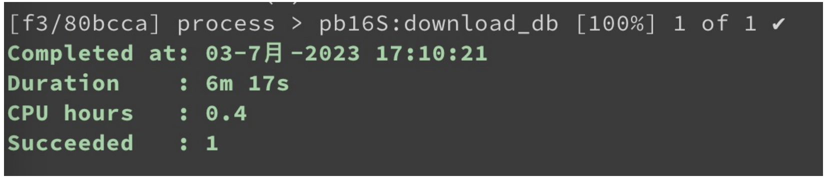
Download Database and run pipeline

1. Download Databases:

```
nextflow run main.nf -download_db
```

# With docker (If you use docker, add -profile docker to all Nextflow-related command)

```
nextflow run main.nf -download_db -profile docker
```



```
[f3/80bcca] process > pb16S:download_db [100%] 1 of 1 ✔
Completed at: 03-7月-2023 17:10:21
Duration     : 6m 17s
CPU hours    : 0.4
Succeeded    : 1
```

2. Run pipeline:

```
nextflow run main.nf -input sample.tsv \
--metadata metadata.tsv \
-profile conda \
--outdir results
```

if using Docker, just add "`-profile docker`".

Modify "nextflow.config" to utilizes HPC job scheduler if desirable

Run analysis

Usage

```
$nextflow run main.nf \
--input pb_sample.tsv \
--metadata pb_metadata.tsv \
-profile conda \
--dada2_cpu 80 --vsearch_cpu 80 \
--outdir PB_16S_2023-03
```

By default, sequences are first trimmed with cutadapt. If adapters are already trimmed, you can skip cutadapt by specifying "`--skip_primer_trim`".

Other important options:

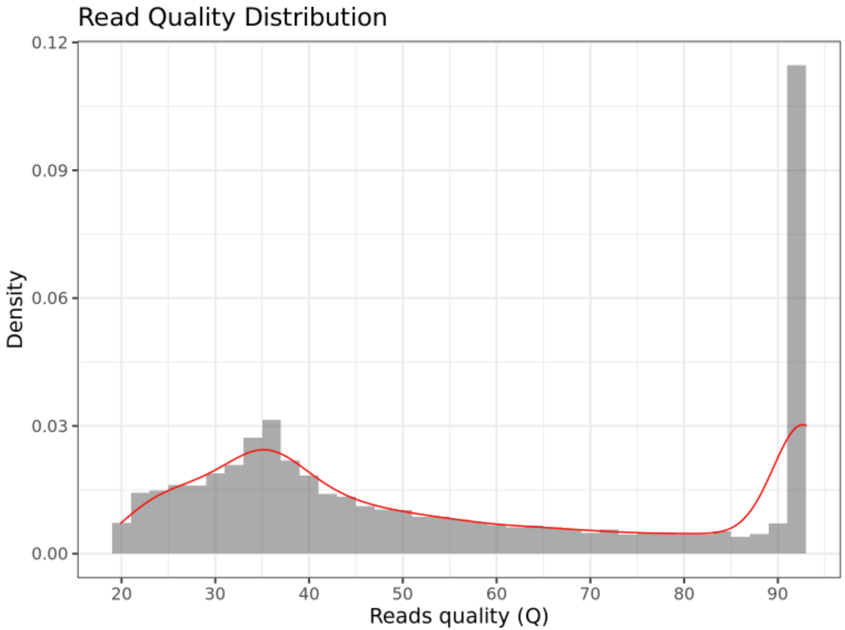
`--dada2_cpu` Number of threads for DADA2 denoising (default: 8)

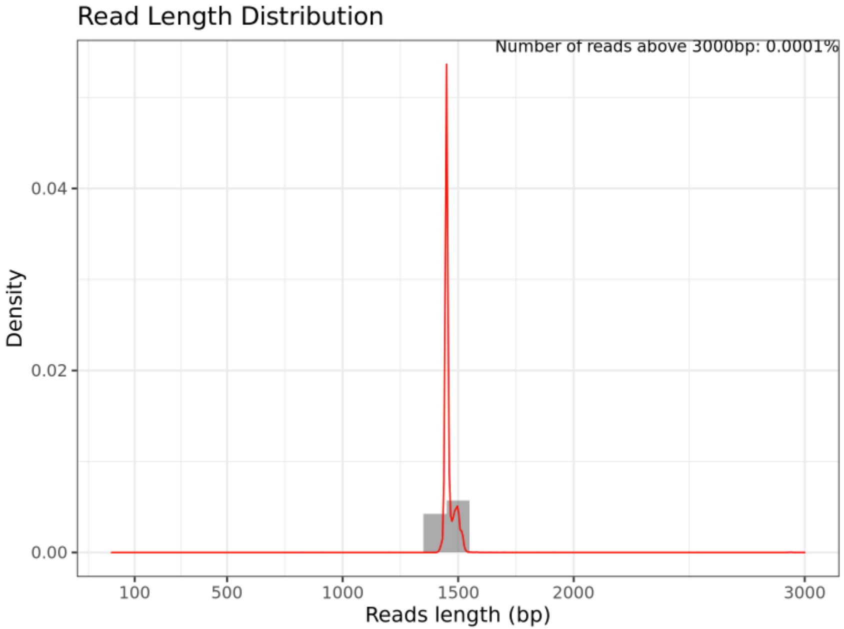
`--vsearch_cpu` Number of threads for VSEARCH taxonomy classification (default: 8)

`--cutadapt_cpu` Number of threads for primer removal using cutadapt (default: 16)

Results from pb-16S-nf pipeline

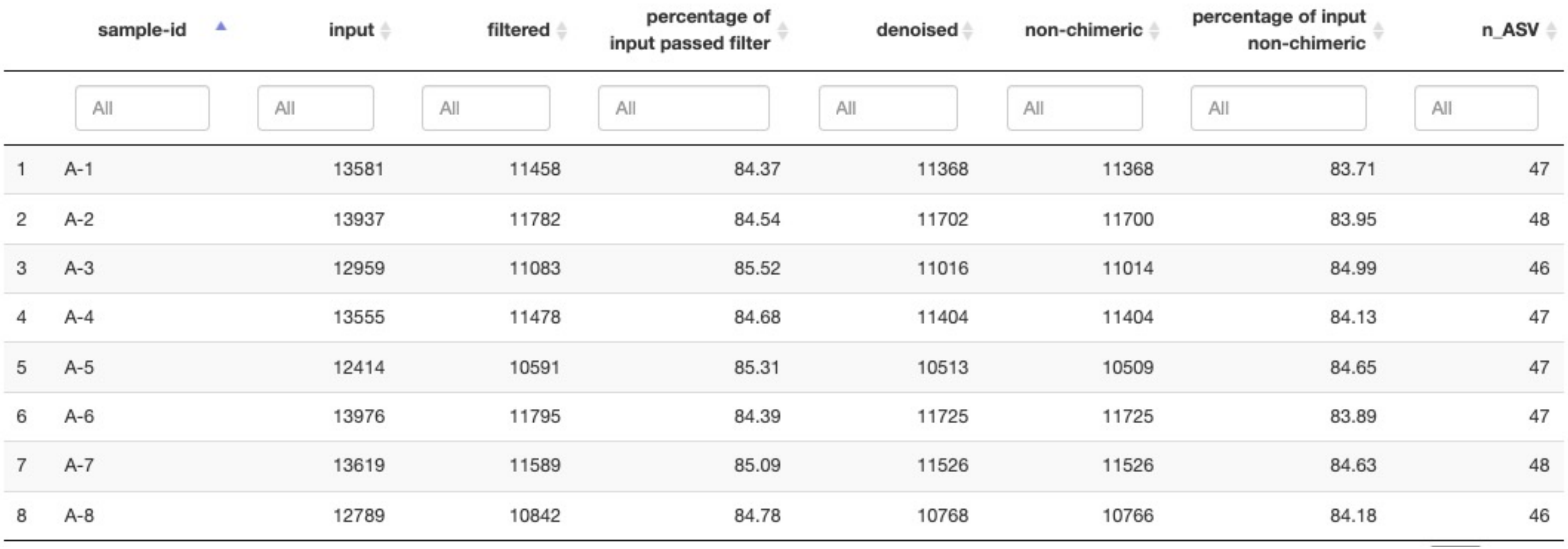
Input reads QC (Before filtering and primers removal)





DADA2 QC metrics

Summarizing Denoised Statistics



	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	A-1	13581	11458	84.37	11368	11368	83.71	47
2	A-2	13937	11782	84.54	11702	11700	83.95	48
3	A-3	12959	11083	85.52	11016	11014	84.99	46
4	A-4	13555	11478	84.68	11404	11404	84.13	47
5	A-5	12414	10591	85.31	10513	10509	84.65	47
6	A-6	13976	11795	84.39	11725	11725	83.89	47
7	A-7	13619	11589	85.09	11526	11526	84.63	48
8	A-8	12789	10842	84.78	10768	10766	84.18	46

https://github.com/PacificBiosciences/pb-16S-analysis/wiki/Analyzing-PacBio-HiFi-Mock-Community-16S-Data-with-QIIME-2

Default pipeline parameters with all data

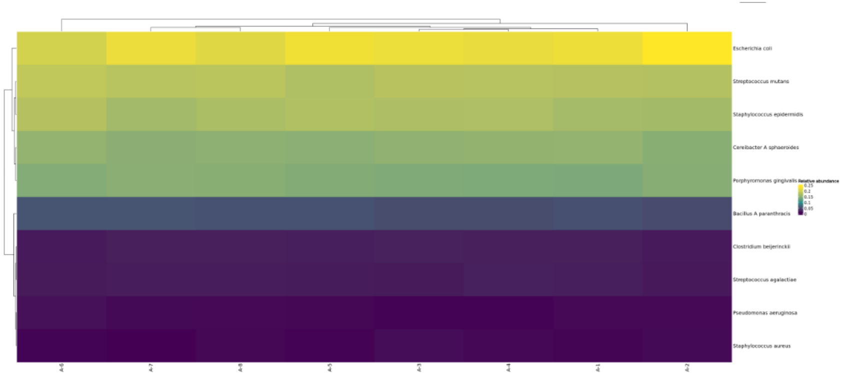
Summary QC statistics

Samples number: 8

Total number of ASVs found: 50

Average number of ASVs per sample: 47

Total number of reads in 50 ASVs: 90002 (84.25% of all input reads)



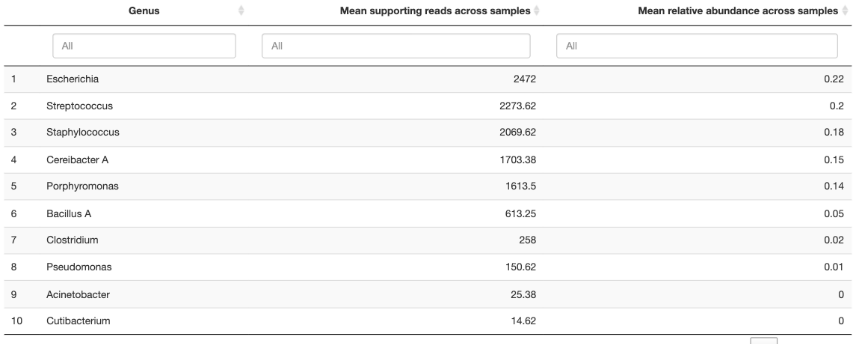
Classification using VSEARCH (GTDB r207)

ASVs classified at Species level: 50 (100%)

Percentage reads belong to ASV classified at Species level: 100%

ASVs classified at Genus level: 50 (100%)

Percentage reads belong to ASV classified at Genus level 100%

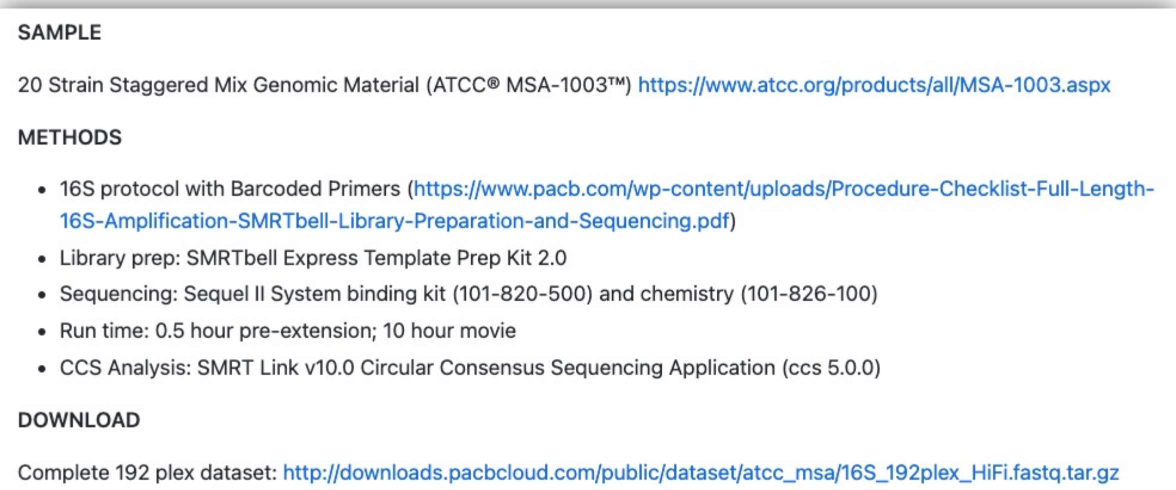


	Genus	Mean supporting reads across samples	Mean relative abundance across samples
	All	All	All
1	Escherichia	2472	0.22
2	Streptococcus	2273.62	0.2
3	Staphylococcus	2069.62	0.18
4	Cereibacter A	1703.38	0.15
5	Porphyromonas	1613.5	0.14
6	Bacillus A	613.25	0.05
7	Clostridium	258	0.02
8	Pseudomonas	150.62	0.01
9	Acinetobacter	25.38	0
10	Cutibacterium	14.62	0

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.pacbcloud.com/public/dataset/atcc_msa/16S_192plex_HiFi.fastq.tar.gz

- pb-16S-nf

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

PacBio

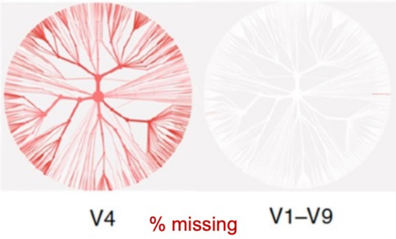
Microbial Assembly Analysis Application

04 July 2023

彭彥菱 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

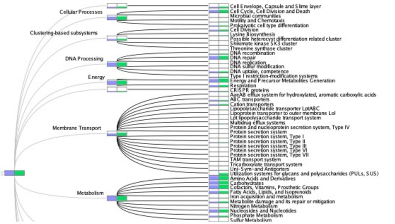
HiFi sequencing delivers the most comprehensive and highest quality data for microbial genomics

Full-length 16S sequencing



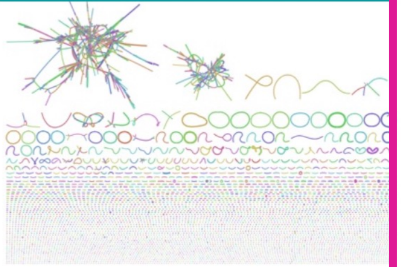
- Species/strain-level resolution
- Reveals true sample diversity

Shotgun metagenome profiling


Cellular Processes
Clustering-based subsystems
DNA Processing
Energy
Membrane Transport
Metabolism
Cell Envelope, Capsule and Slime layer
Cell Cycle, Cell Division and Death
Microbial communities
Motility and Chemotaxis
Prokaryotic cell type differentiation
Cell Division
Lysine Biosynthesis
Possible heterocyst differentiation related cluster
Shikimate kinase SK3 cluster
Threonine synthase cluster
DNA recombination
DNA repair
DNA replication
DNA sulfur modification
DNA uptake, competence
Type I restriction-modification systems
Energy and Precursor Metabolites Generation
Respiration
CRISPR proteins
AaeAB efflux system for hydroxylated, aromatic carboxylic acids
ABC transporters
Cation transporters
Lipopolysaccharide transporter LptABC
Lipoprotein transporter to outer membrane Lol
Lpt lipopolysaccharide transport system
Multidrug efflux systems
Protein and nucleoprotein secretion system, Type IV
Protein secretion system
Protein secretion system, Type I
Protein secretion system, Type II
Protein secretion system, Type III
Protein secretion system, Type VI
Protein secretion system, Type VII
TAM transport system
Tricarboxylate transport system
Uni- Sym- and Antiporters
Utilization systems for glycans and polysaccharides (PULs, SUS)
Amino Acids and Derivatives
Carbohydrates
Cofactors, Vitamins, Prosthetic Groups
Fatty Acids, Lipids, and Isoprenoids
Iron acquisition and metabolism
Metabolite damage and its repair or mitigation
Nitrogen Metabolism
Nucleosides and Nucleotides
Phosphate Metabolism
Sulfur Metabolism

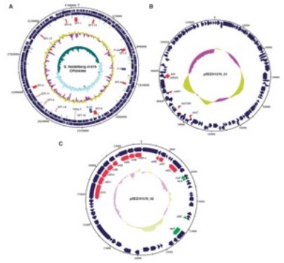
- ~8 genes per sequence read
- ~90% of reads with at least 1 gene
- Profile taxonomy with high precision and recall

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

Microbial whole genome sequencing and assembly with HiFi data

Complete microbial genomes

including chromosomes and plasmids

High contiguity

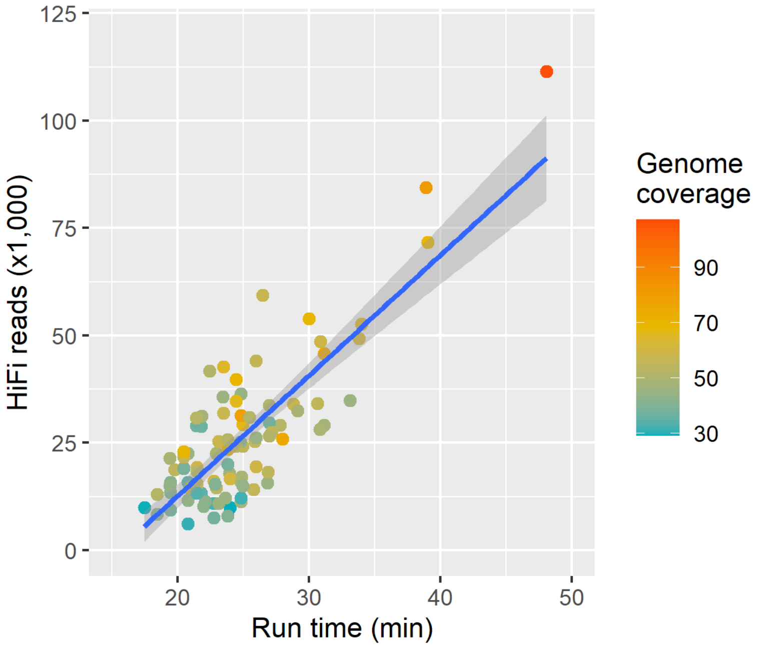
high per-base quality of final microbial assemblies

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

$$\longrightarrow\ Target\ HiFi\ Base\ Yield = [Microbe\ Genome\ Size\ (Mb)]x\ [Target\ HiFi\ Coverage\ per\ Microbe]$$

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)

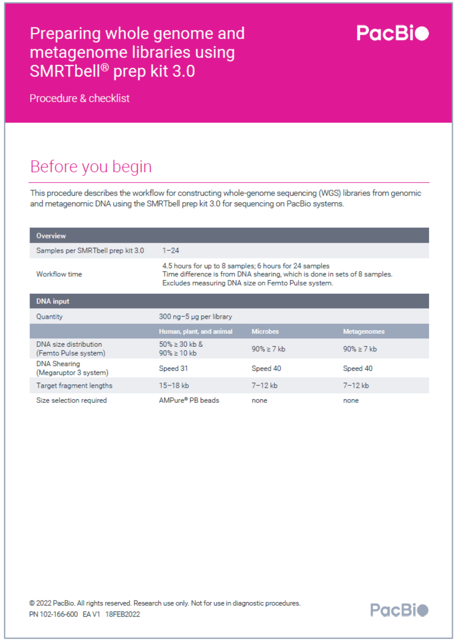
** Minimum Compute Requirements: Head Node - Cores: 32, RAM: 64 GB, 1 TB local tmp, 256 GB local db_datadir; Compute Nodes – Cores 64, RAM: 4GB per core, 1 TB local tmp, 256 GB local db_datadir*

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

PacBio

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

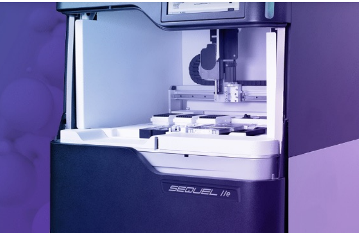
PacBio

PacBio Documentation (102-166-600)

APPLICATIONS

WHOLE GENOME SEQUENCING

De Novo assembly & variant detection
Microbial assembly
Shotgun metagenomics



Example performance

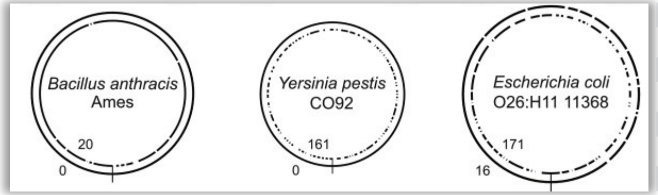
https://downloads.pacbcloud.com/public/dataset/2021-11-Microbial-96plex/

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



Class I **Class II** **Class III**

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.

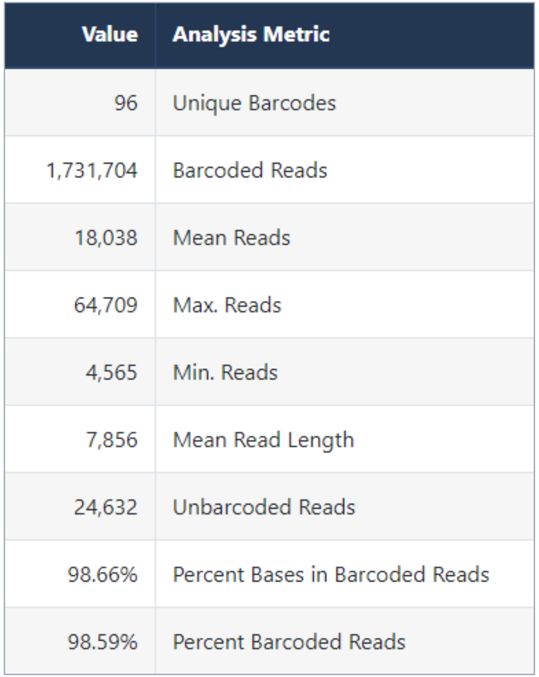
Class III - Contain large, often phage-related, repeats >7 kb.

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

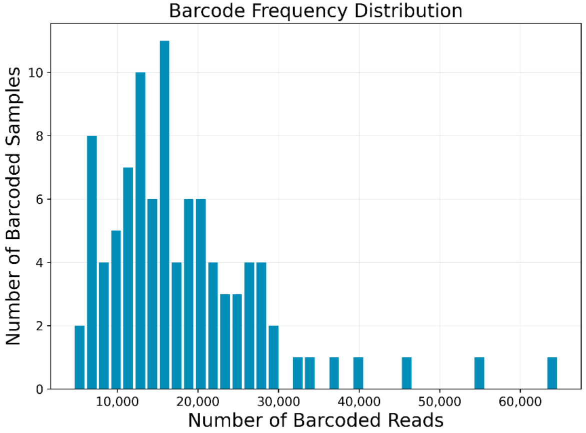
** Koren, S. et al. (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14, R101*

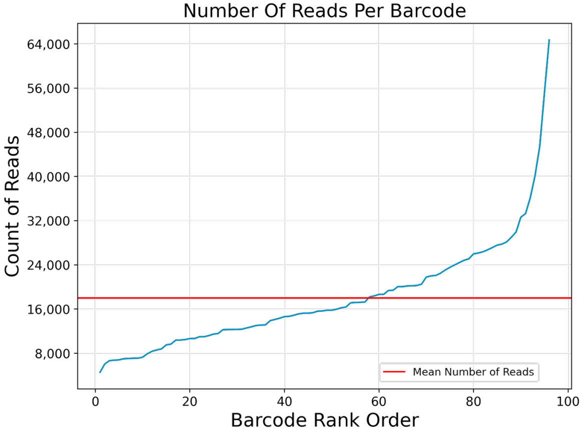
Example sequencing performance for a 96-plex microbial WGS library prepared with SMRTbell prep kit 3.0 (cont.)

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



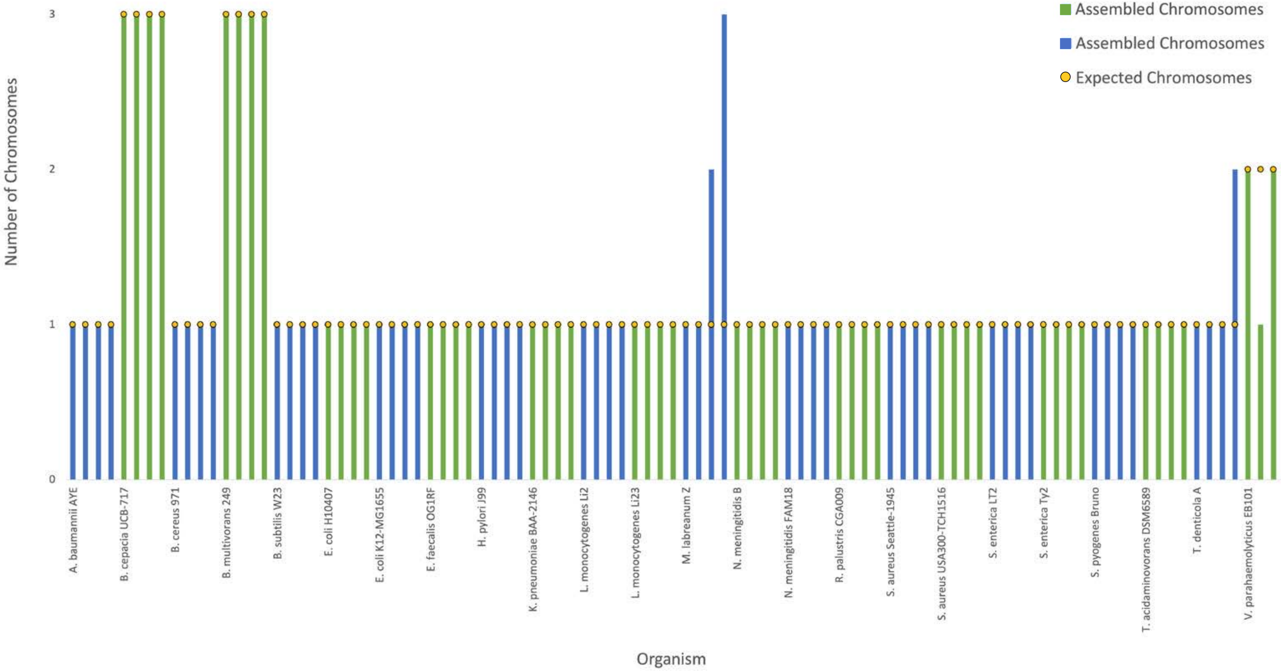


- **All 96 barcodes detected**
- Mean # of barcoded HiFi reads per microbe is ~18,000
- Mean HiFi base coverage per microbe is 36-fold (Range is 19- to 63-fold)

PacBio

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

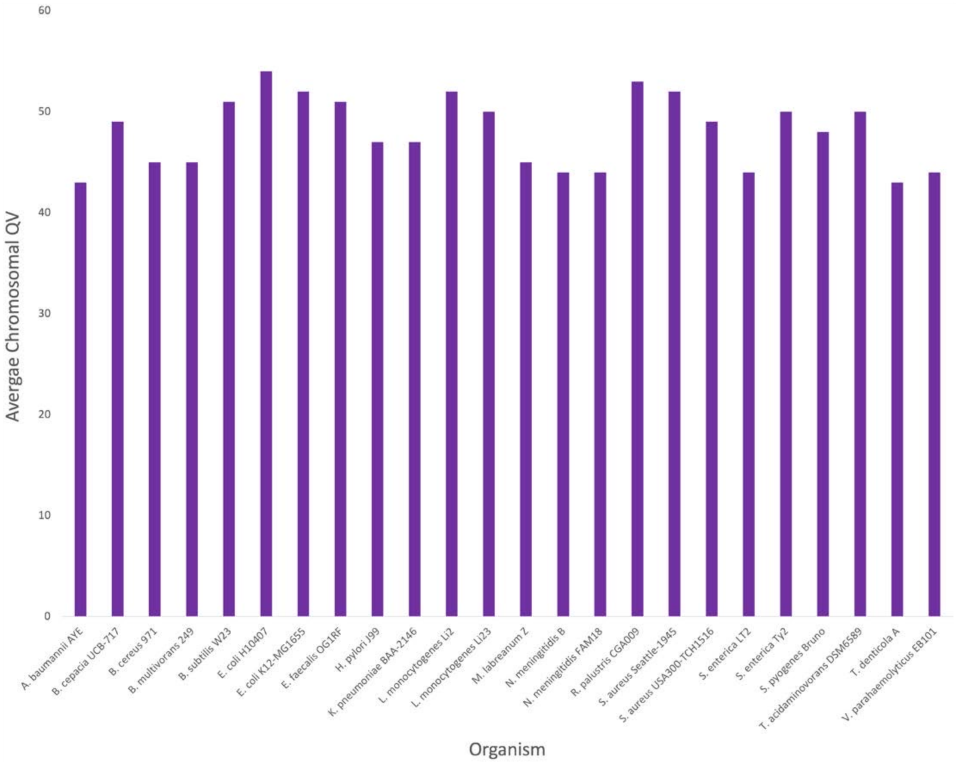


- Achieved **1 Contig / Chromosome** for 92 out of 96 assemblies
- For all 96 microbes, chromosomal assemblies were **complete** and **of the expected sizes**

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.

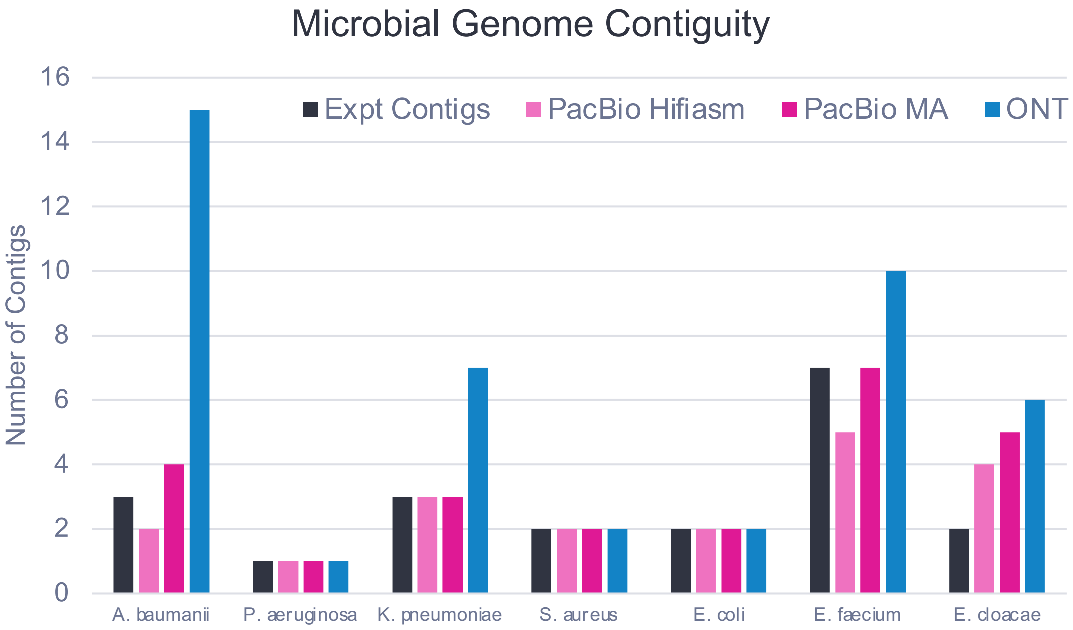
Example sequencing performance for a 96-plex microbial WGS library prepared with SMRTbell prep kit 3.0 (cont.)

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**

PacBio superior data quality has real-world consequences for antimicrobial surveillance and susceptibility testing

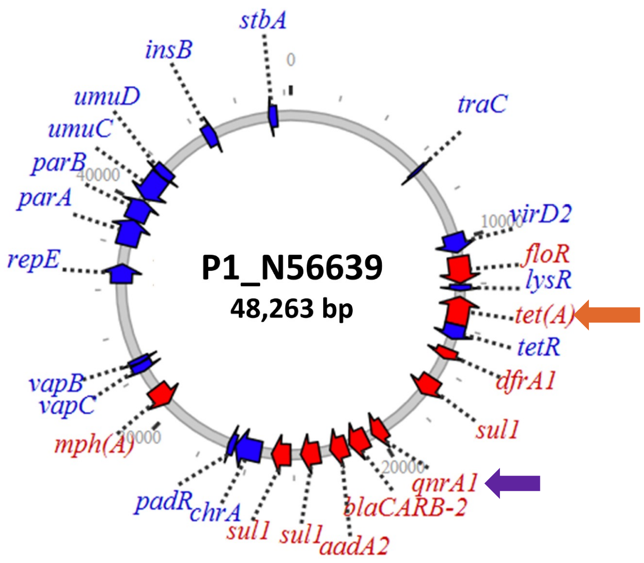


- The customer wanted to evaluate the use of different sequencing technologies in their genomics-based AMR and antibiotic susceptibility pipeline
- PacBio produces more contiguous assemblies than either Illumina or ONT

How does dangerous antibiotic resistance develop and spread?

- Scientists at National Antimicrobial Resistance Monitoring System (NARMS) sequenced *E. coli* found on retail meats
- Identified plasmids mediated quinolone resistance (PMQR) genes on novel plasmid backbones
- Saw evidence of co-selection of resistance to quinolone and antimicrobials used in animal feed and to treat infections in humans
- Read the blog

"These details are important in assessing the nature of resistant microbial hazards in food and other sources."[1]



48-kb IncR plasmid

[1] Tyson, G. H. et al. (2019) Diverse fluoroquinolone resistance plasmids from retail meat *E. coli* in the United States. *Frontiers in Microbiology.*

Analysis workflow overview

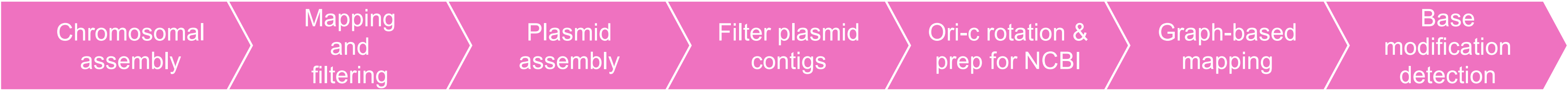
HiFi microbial assembly workflow

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



Chromosomal assembly

Mapping and filtering

Plasmid assembly

Filter plasmid contigs

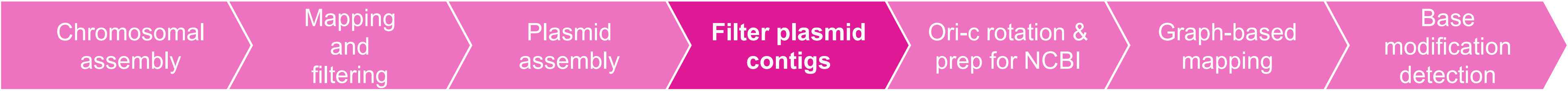
Ori-c rotation & prep for NCBI

Graph-based mapping

Base modification detection

Filter plasmid contigs

HiFi microbial assembly workflow stages



Chromosomal assembly

Mapping and filtering

Plasmid assembly

Filter plasmid contigs

Ori-c rotation & prep for NCBI

Graph-based mapping

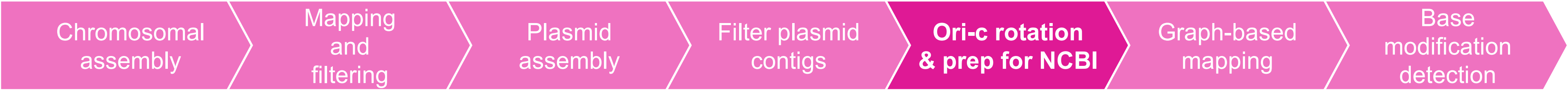
Base modification detection

Task: filtering of plasmid contigs

Method: map (pbmm2) plasmid contigs to chromosomal contig(s) and filter out contigs with more than 90% gap compressed identity and longer than 300 kb (default)

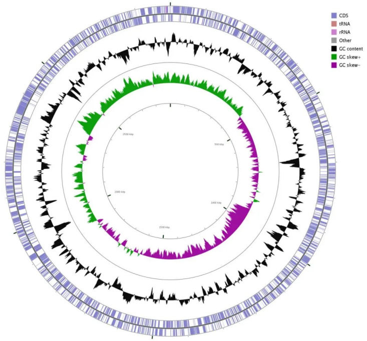
Ori-c rotation

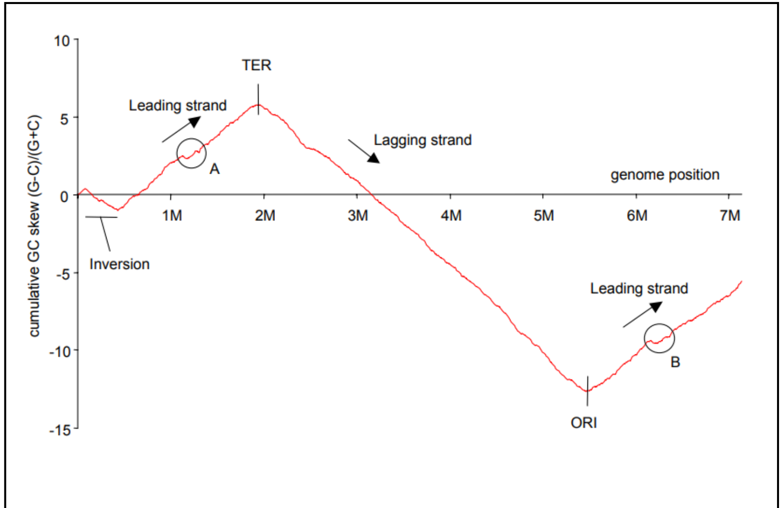
HiFi microbial assembly workflow stages



Task: find origin of replication, header and file formatting

Method: GC-skew for origin of replication detection

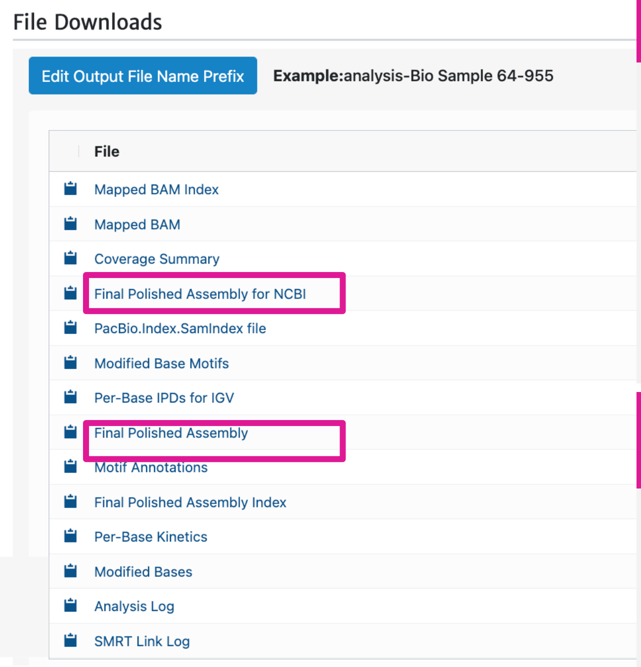




Plot of the cumulative GC-skew ((G-C)/(G+C)) in *Pirellula* sp. strain1.
The line shows the inversion. A and B indicate minor irregularities in the genome.

SMRT Analysis report

Data



File Downloads

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]
TCAATTGTGAATAACTTTTTGCACATCCTGTGGATAAATTATCACATAAACTTATCCACAATCCATAAAGACAATAAAAACAGAGTTA
TCAACAGTTCAAATATATGTTTTTTAAATTTAAAACTGTGAAATCCACAAGAAAAGTCCACACTAATAAGAATAAATTTAAATTTTAA
AATTTGAATTTATTTAATAGGGCTGATCCAAATTGTGGATAACTAAAAAATATGAATTTAAATTCAAATATACCAAATCAAAACCAAC
TTCACATCAAGGTTTGTTGGTAAGTATGTAAATAAGAAGTGTATATCTTAAAAGTCTTAATAAAAATAAACAATTACTTTGTGCATAA
CTTTTAAATAAGAAAAATAGGCTAAATATAAAGAGAAGATAAAAAGTTAAAAATTTGACTTAAATACAAAACTTTCACGGTTTTTCAT
TGACAGCGTAAACATTGCACAATAAAATCGCGGACCTTTATAGAAAGATCATTTTTGGGAGTTTCGATATGAAACGTACTTTCCAACC
ATCTGAATTAAA
```

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]

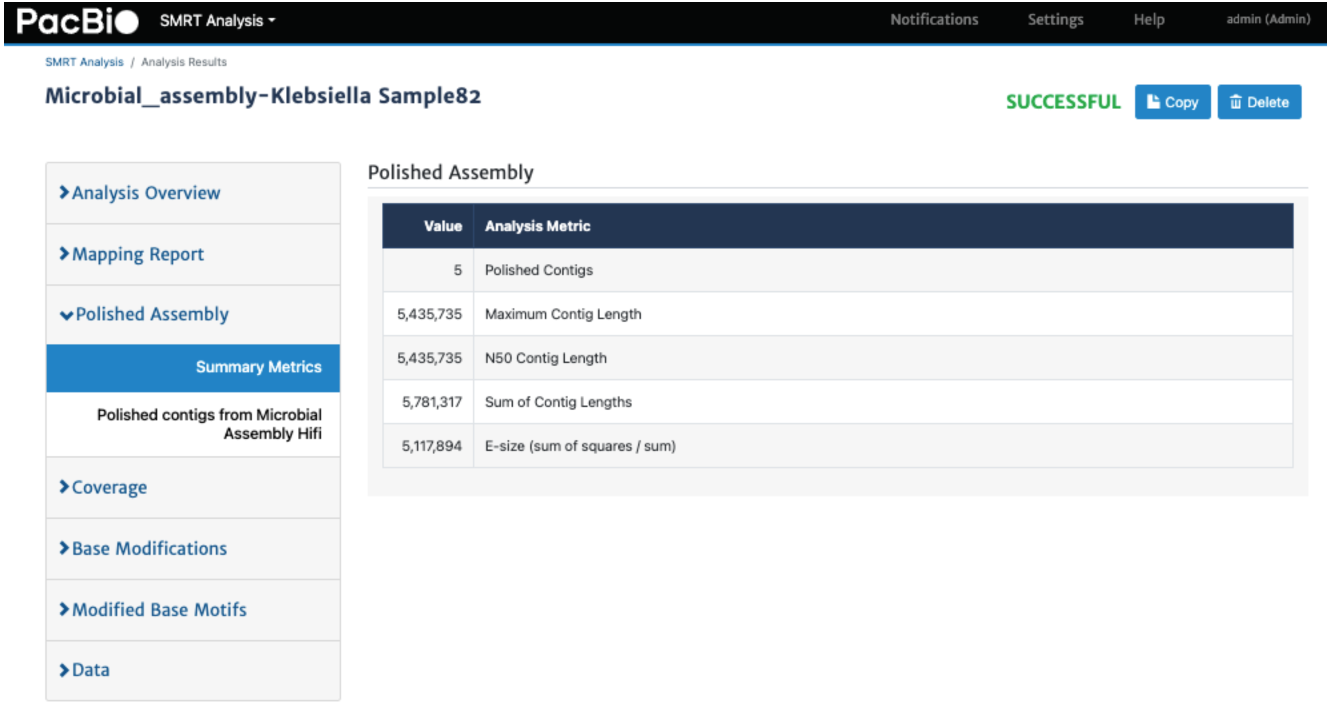
```
>ctg.s1.000000F shifted_by_bp:-1218400/3943308
TCAATTGTGAATAACTTTTTGCACATCCTGTGGATAAATTATCACATAAACTTATCCACAATCCATAAAGACAATAAAAACAGAGTTA
TCAACAGTTCAAATATATGTTTTTTAAATTTAAAACTGTGAAATCCACAAGAAAAGTCCACACTAATAAGAATAAATTTAAATTTTAA
AATTTGAATTTATTTAATAGGGCTGATCCAAATTGTGGATAACTAAAAAATATGAATTTAAATTCAAATATACCAAATCAAAACCAAC
TTCACATCAAGGTTTGTTGGTAAGTATGTAAATAAGAAGTGTATATCTTAAAAGTCTTAATAAAAATAAACAATTACTTTGTGCATAA
CTTTTAAATAAGAAAAATAGGCTAAATATAAAGAGAAGATAAAAAGTTAAAAATTTGACTTAAATACAAAACTTTCACGGTTTTTCAT
TGACAGCGTAAACATTGCACAATAAAATCGCGGACCTTTATAGAAAGATCATTTTTGGGAGTTTCGATATGAAACGTACTTTCCAACC
ATCTGAATTAAA
```

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

Analysis results guide

SMRT Analysis report

Polished Assembly



Microbial_assembly-Klebsiella Sample82

SUCCESSFUL

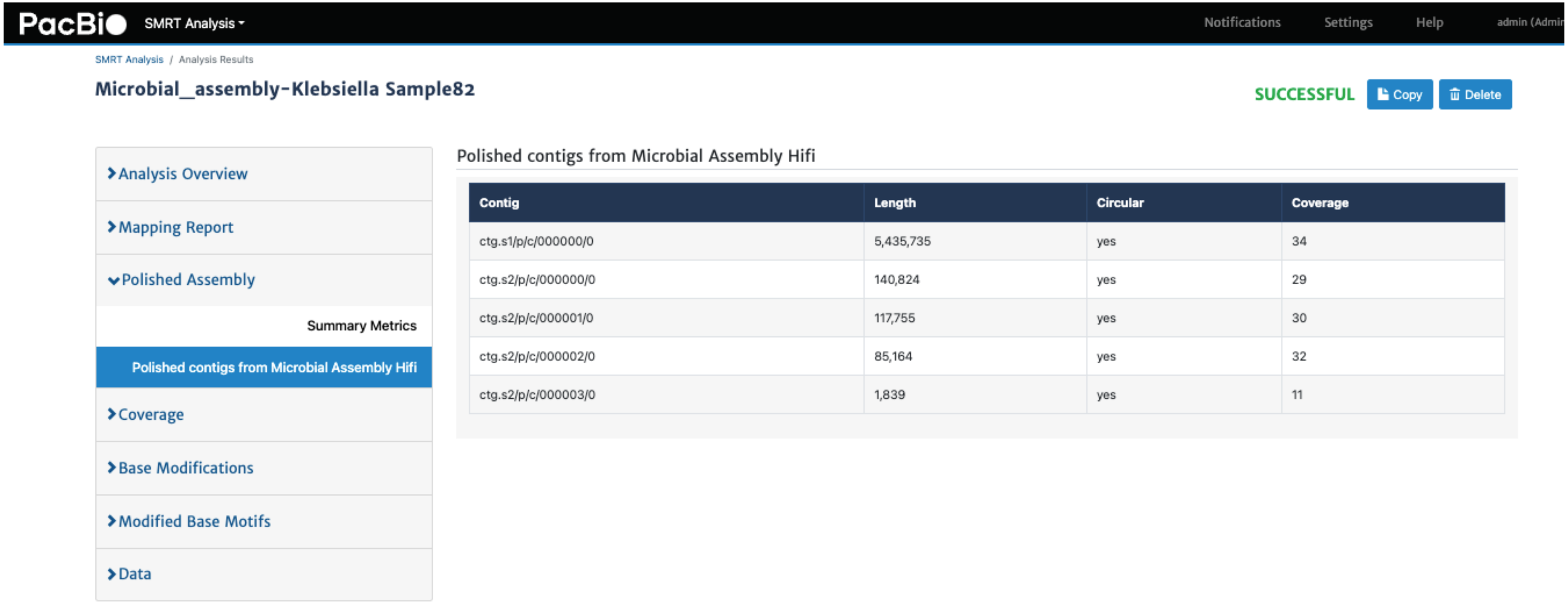
Polished Assembly

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)

Klebsiella pneumoniae BAA-2146

SMRT Analysis report

Polished Assembly



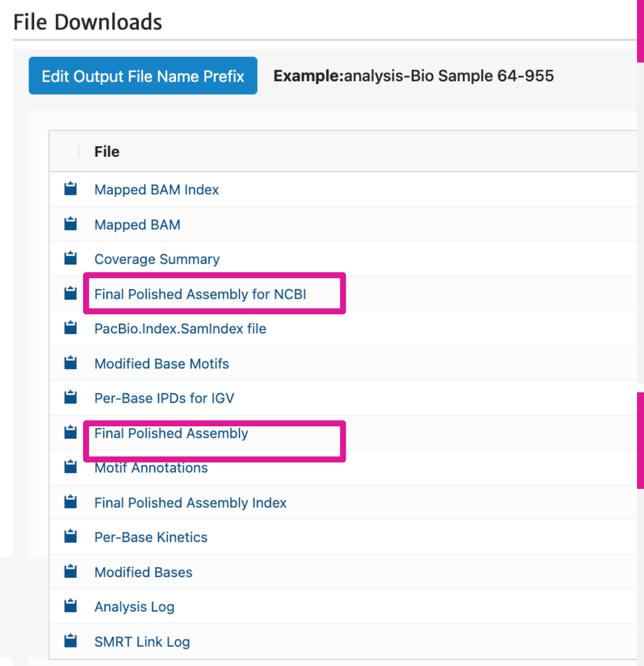
Microbial_assembly-Klebsiella Sample82

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

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]
TCAATTGTGAATAACTTTTTGCACATCCTGTGGATAAATTATCACATAAACTTATCCACAATCCATAAAGACAATAAAAACAGAGTTA
TCAACAGTTCAAATATATGTTTTTTAAATTTAAAACTGTGAAATCCACAAGAAAAGTCCACACTAATAAGAATAAATTTAAATTTTAA
AATTTGAATTTATTTAATAGGGCTGATCCAAATTGTGGATAACTAAAAAATATGAATTTAAATTCAAATATACCAAATCAAAACCAAC
TTCACATCAAGGTTTGTTGGTAAGTATGTAAATAAGAAGTGTATATCTTAAAAGTCTTAATAAAAATAAACAATTACTTTGTGCATAA
CTTTTAAATAAGAAAAATAGGCTAAATATAAAGAGAAGATAAAAAGTTAAAAATTTGACTTAAATACAAAACTTTCACGGTTTTTCAT
TGACAGCGTAAACATTGCACAATAAAATCGCGGACCTTTATAGAAAGATCATTTTTGGGAGTTTCGATATGAAACGTACTTTCCAACC
ATCTGAATTAAA
```

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
TCAATTGTGAATAACTTTTTGCACATCCTGTGGATAAATTATCACATAAACTTATCCACAATCCATAAAGACAATAAAAACAGAGTTA
TCAACAGTTCAAATATATGTTTTTTAAATTTAAAACTGTGAAATCCACAAGAAAAGTCCACACTAATAAGAATAAATTTAAATTTTAA
AATTTGAATTTATTTAATAGGGCTGATCCAAATTGTGGATAACTAAAAAATATGAATTTAAATTCAAATATACCAAATCAAAACCAAC
TTCACATCAAGGTTTGTTGGTAAGTATGTAAATAAGAAGTGTATATCTTAAAAGTCTTAATAAAAATAAACAATTACTTTGTGCATAA
CTTTTAAATAAGAAAAATAGGCTAAATATAAAGAGAAGATAAAAAGTTAAAAATTTGACTTAAATACAAAACTTTCACGGTTTTTCAT
TGACAGCGTAAACATTGCACAATAAAATCGCGGACCTTTATAGAAAGATCATTTTTGGGAGTTTCGATATGAAACGTACTTTCCAACC
ATCTGAATTAAA
```

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

Cromwell workflow key output files

basemods.csv

basemods.gff

modifications.report.json

motifs.csv

motifs.gff

motifs.report.json

ipds.bw

Base modification and motifs

collected_circ.txt

coverage.gff

Coverage Summary

coverage.report.json

mapped.bam

mapped.bam.bai

Mapped BAM and Index

mapped.consensusalignmentset.xml

mapping_stats.report.json

polished_assembly.fasta

polished_assembly.fasta.fai

polished_assembly.report.json

assembly.rotated.polished.renamed.fsa

Final Polished Assembly for NCBI

p_ctg_oric.fasta

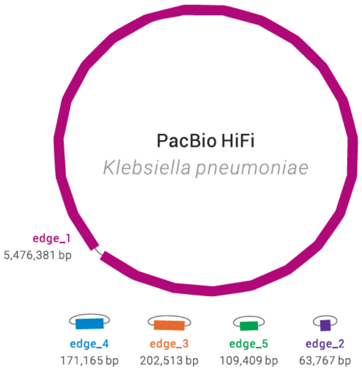
final_assembly.fasta.fai

Final Polished Assembly and Index

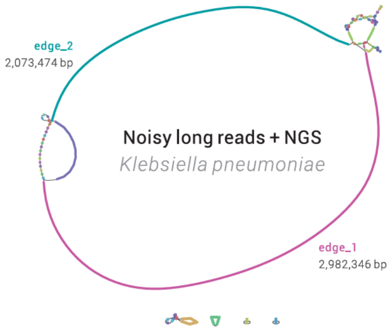
Case Study Sharing

Visualization and comparison of WGS assemblies for *K. pneumoniae*

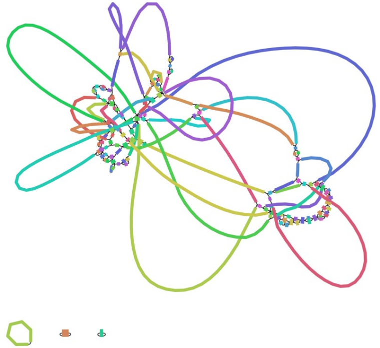
PacBio HiFi



Hybrid assembly

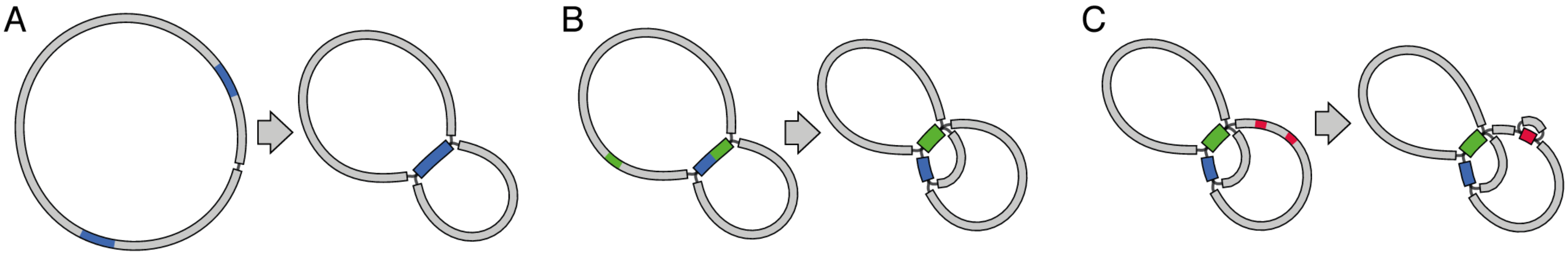


NGS ILMN



	PacBio HiFi	ONT + ILMN	ILMN
Coverage	40X	69X (ONT), 34X (ILMN)	34X
Contig N50	5.47 Mb	2.1 Mb	0.3 Mb
Number of contigs	5	47	220
Assembler	Flye	Unicycler	Unicycler

Limitations of short reads



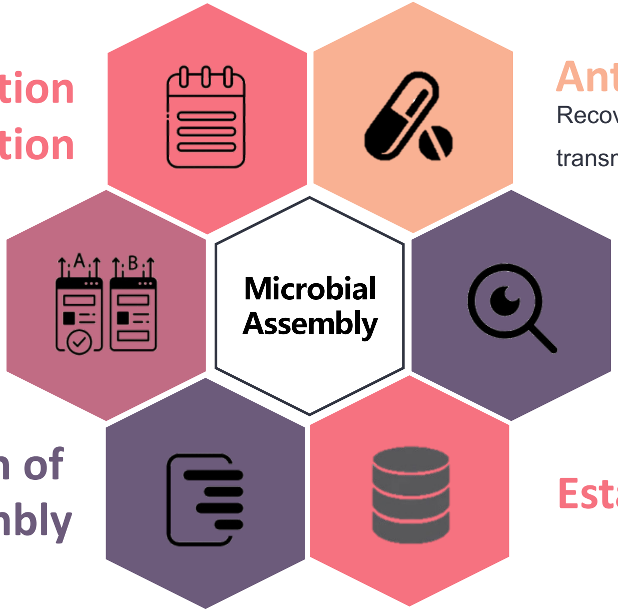
The main reason we can ‘t get a complete assembly from short reads is that DNA usually contains ***repeats*** – the same sequence occurring two or more times in the genome.

To complete a bacterial genome assembly (i.e. find the one correct sequence for each chromosome/plasmid), we need to resolve the repeats. This means finding which way into a repeat matches up with which way out. **Short reads don't have enough information for this but *long reads* do**.

Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595.

Downstream Applications

Downstream Application



Genome predication and Annotation

Antibiotic resistance gene

Recover plasmids to track drug resistance and transmission paths

Comparative genome analyses

Detect variant of virulence

Clarify the role of transposons, phage insertions, and other structural variants in the evolution of virulence

Validation of the assembly

Establish Database

Useful tools for further analysis

01

Genome Annotation

http://rast.theseed.org/FIG/rast.cgi

RAST Rapid Annotation using Subsystem Technology version 2.0

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

KBase PREDICTIVE BIOLOGY

http://kbase.us/

PROKKA

https://github.com/tseemann/prokka

02

Comparative Analysis

http://quast.sourceforge.net/quast

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

Assemblytics

Analyze your assembly by comparing it to a reference genome

http://assemblytics.com/

MUMMER 3+

http://mummer.sourceforge.net/

03

Visualization

Ribbon

http://genomeribbon.com/

BUSCO

https://busco.ezlab.org/

igv Integrative Genomics Viewer

https://igv.org/

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
- 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

Flye assembler

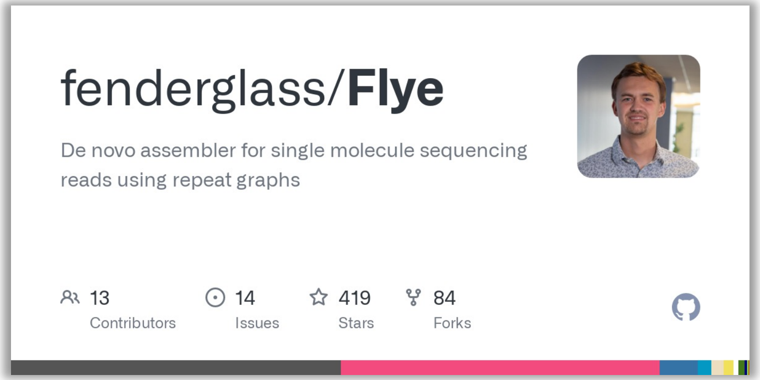
De novo assembler for single molecule sequencing reads.

It is designed for a wide range of datasets, from small bacterial projects to large mammalian-scale assemblies.

The package represents a complete pipeline.

Supported Input Data:

- Oxford Nanopore (ONT reads)
- PacBio (raw, corrected and HiFi reads)

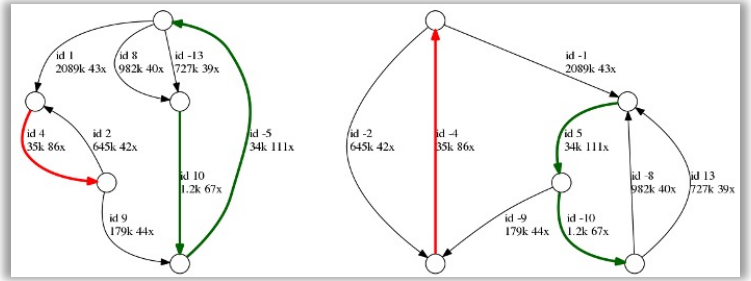


fenderglass/**Flye**

De novo assembler for single molecule sequencing reads using repeat graphs

13	14	419	84
Contributors	Issues	Stars	Forks

Repeat graph

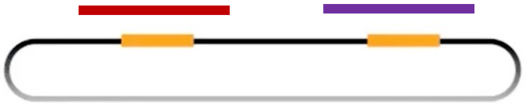


Repetitive edges are colored / **Unique** edges are black

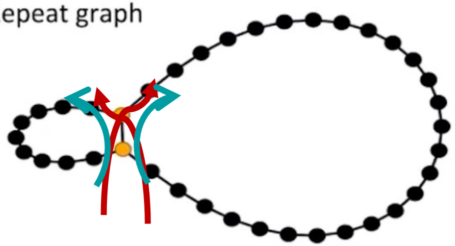
Mikhail Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith and Pavel A. Pevzner "metaFlye: scalable long-read metagenome assembly using repeat graphs", Nature Methods, 2020. doi:s41592-020-00971-x

Untangling Repeat Graph

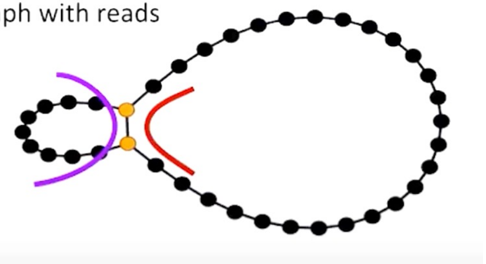
Genome with one repeat



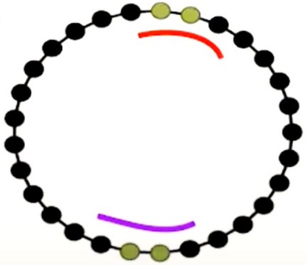
Repeat graph



Repeat graph with reads



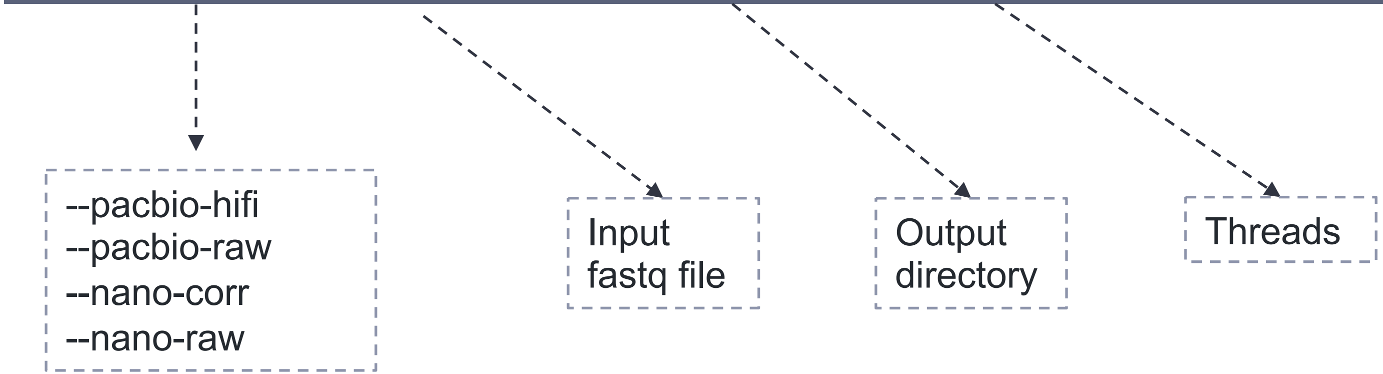
Simplified graph



Quick usage for Flye assembler

E. coli K12 PacBio data

```
flye --pacbio-hifi <fastqfile> --out-dir out_pacbio --threads 4
```



--pacbio-hifi
--pacbio-raw
--nano-corr
--nano-raw

Input fastq file

Output directory

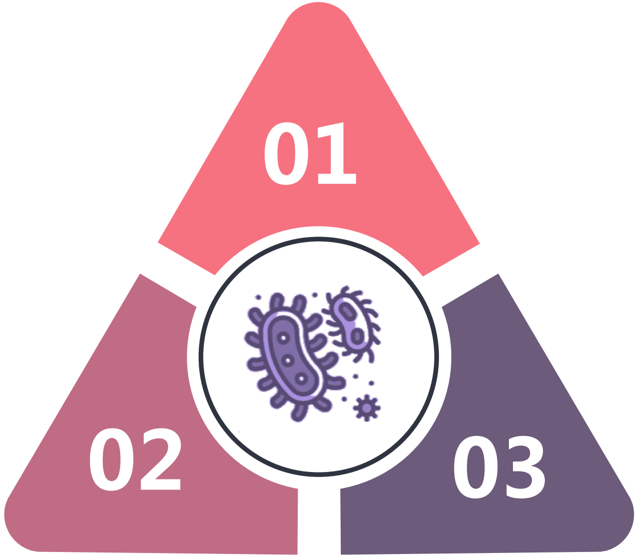
Threads

- For **PacBio HiFi** use the `--pacbio-hifi` mode. The default error-rate is 0.001 (in HPC space), and works well for the default CCS algorithm settings (e.g. 3+ polymerase passes).

The original dataset is available at the 2021-11-Microbial-96-plex

PacBio https://github.com/fenderglass/Flye/blob/flye/docs/USAGE.md

Analysis Interpretation



Assemble Quality

Contig length & number, contigN50, circular…etc.

Sequence consistency

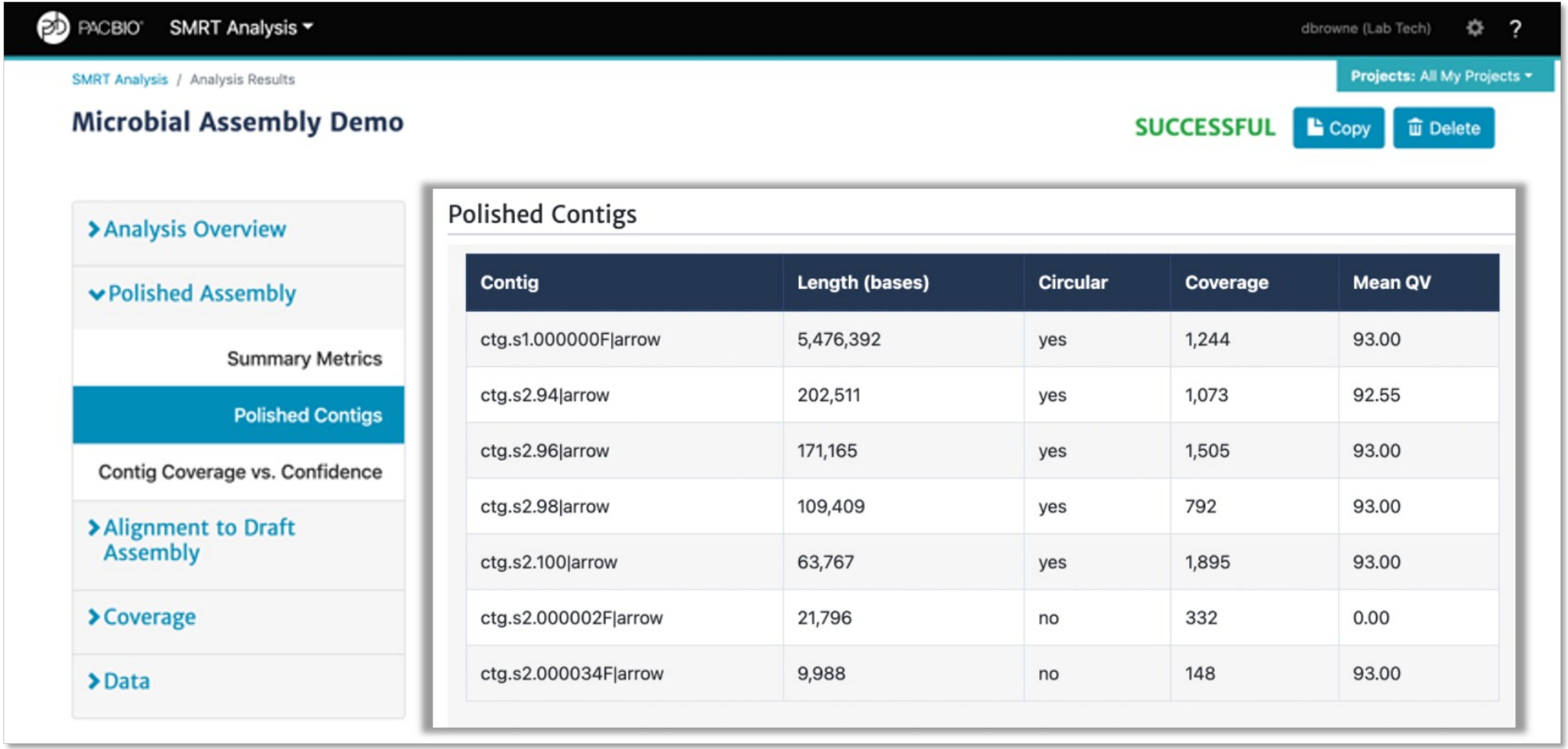
Mapping rate and coverage, Mean Concordance (mapped)

Assembly Complete

BUSCO (Benchmarking Universal Single-Copy Orthologs)

Evaluate the results of genome assembly

The analysis results of SMRT Analysis



Microbial Assembly Demo

SUCCESSFUL

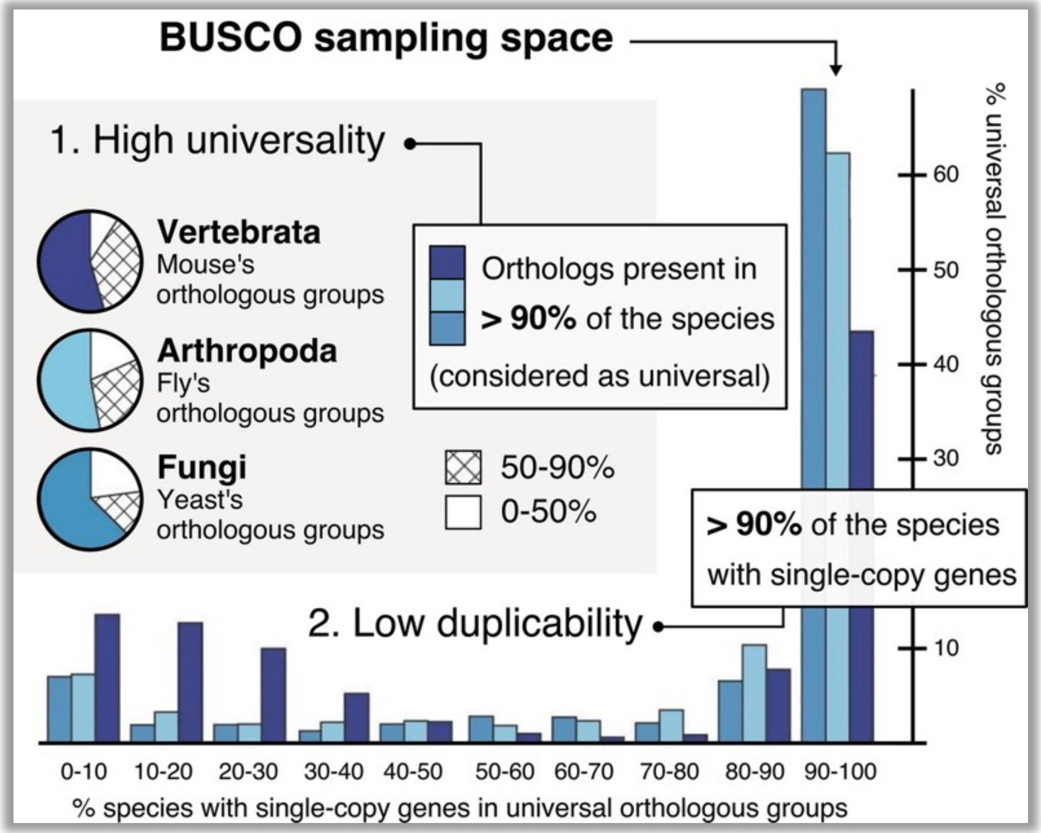
Polished Contigs

Contig	Length (bases)	Circular	Coverage	Mean QV
ctg.s1.000000F|arrow	5,476,392	yes	1,244	93.00
ctg.s2.94|arrow	202,511	yes	1,073	92.55
ctg.s2.96|arrow	171,165	yes	1,505	93.00
ctg.s2.98|arrow	109,409	yes	792	93.00
ctg.s2.100|arrow	63,767	yes	1,895	93.00
ctg.s2.000002F|arrow	21,796	no	332	0.00
ctg.s2.000034F|arrow	9,988	no	148	93.00

Assembly Complete - BUSCO

BUSCO attempts to provide a quantitative assessment of the completeness in terms of expected gene content of a genome assembly, transcriptome, or annotated gene set.

The latest BUSCO versions introduce new functionalities for assessments of **eukaryotic**, **prokaryotic**, and **viral data**.



Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution [Internet]. Available from: https://doi.org/10.1093/molbev/msab199

Assembly Complete - BUSCO

- Report results in simple BUSCO notation:

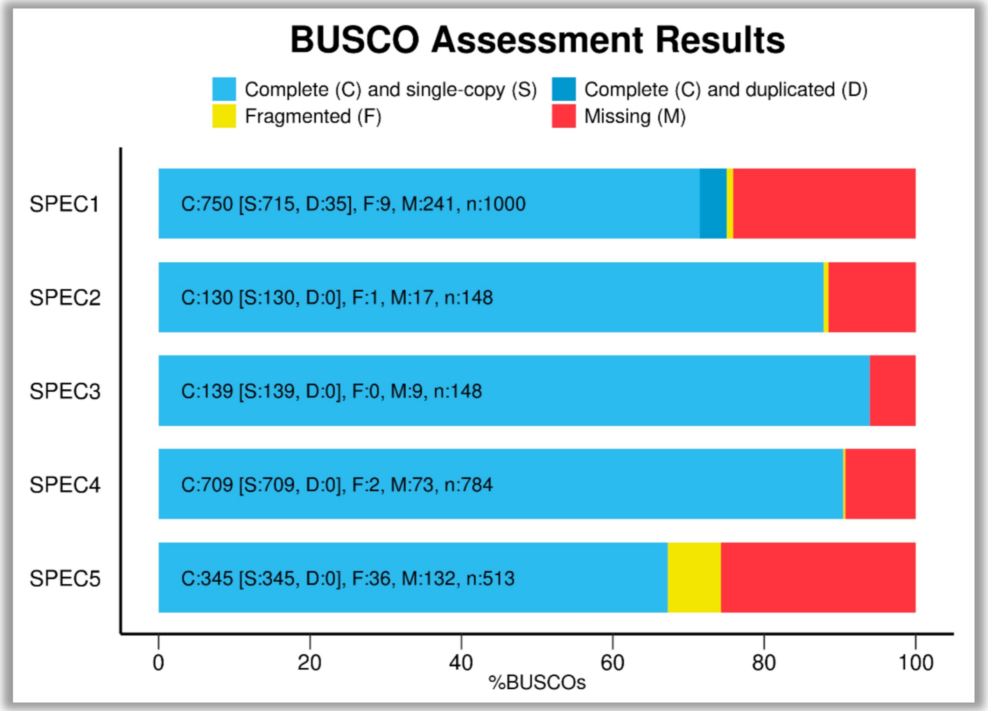
```
 ***** Results: *****

C:75%[S:71.5%,D:3.5%],F:0.9%,M:24.1%,n:1000
        750     Complete BUSCOs (C)
        715     Complete and single-copy BUSCOs (S)
        35      Complete and duplicated BUSCOs (D)
        9       Fragmented BUSCOs (F)
        241     Missing BUSCOs (M)
        1000    Total BUSCO groups searched
```

`short_summary_*.txt`

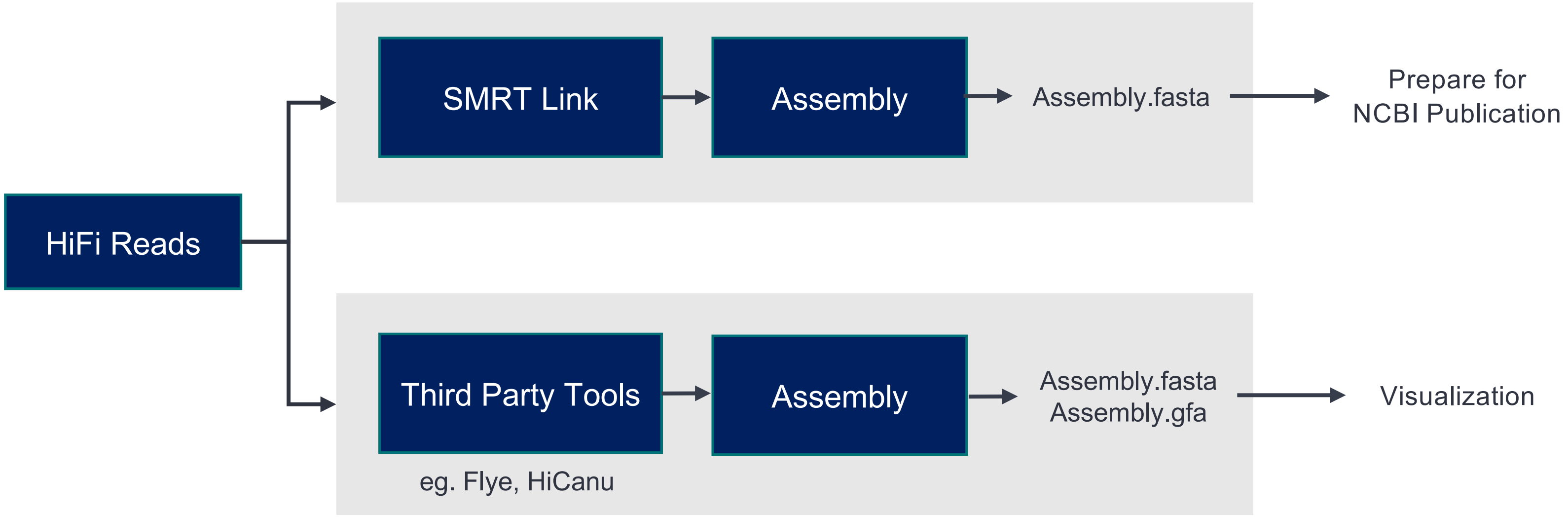
- Use the generate_plot.py script to produce simple graphical summaries for your publication’ s supporting online information.

- Highly recommend using the BUSCO container, whose version is sufficient to safely reproduce a run.



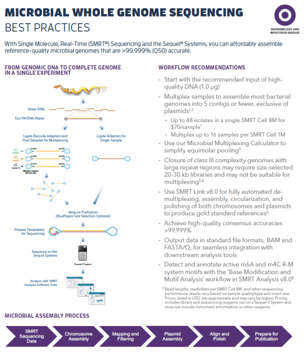
Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution [Internet]. Available from: https://doi.org/10.1093/molbev/msab199

Bioinformatics workflow for microbial assembly



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 >99.999% (Q50) accurate.

FROM GENOMIC DNA TO COMPLETE GENOME IN A SINGLE EXPERIMENT

MICROBIOLOGY AND INFECTIOUS DISEASE
Shear DNA
Exo VII/DNA Repair
Ligate Barcode Adapters and Pool Samples for Multiplexing
Ligate Adapters for Single Sample
Ampure Purification (BluePippin Size Selection Optional)
Prepare Template(s) for Sequencing
Sequence on the Sequel Systems
Sequel II System
Analyze with SMRT Analysis Software Suite

WORKFLOW RECOMMENDATIONS

- Start with the recommended input of high-quality DNA (1.0 µg)
- Multiplex samples to assemble most bacterial genomes into 5 contigs or fewer, exclusive of plasmids[1,2]
  - Up to 48 isolates in a single SMRT Cell 8M for $70/sample*
  - Multiplex up to 16 samples per SMRT Cell 1M
- Use our Microbial Multiplexing Calculator to simplify equimolar pooling[3]
- Closure of class III complexity genomes with large repeat regions may require size-selected 20-30 kb libraries and may not be suitable for multiplexing[3,4]
- Use SMRT Link v8.0 for fully automated de-multiplexing, assembly, circularization, and polishing of both chromosomes and plasmids to produce gold standard references[5]
- Achieve high-quality consensus accuracies >99.999%
- Output data in standard file formats, BAM and FASTA/Q, for seamless integration with downstream analysis tools
- Detect and annotate active m6A and m4C R-M system motifs with the 'Base Modification and Motif Analysis' workflow in SMRT Analysis v8.0[6]

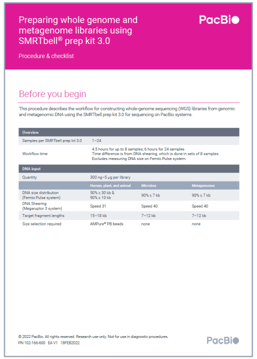
* Read lengths, reads/data per SMRT Cell 8M, and other sequencing performance results vary based on sample quality/type and insert size. Prices, listed in USD, are approximate and may vary by region. Pricing includes library and sequencing reagents run on a Sequel II System and does not include instrument amortization or other reagents.

MICROBIAL ASSEMBLY PROCESS

SMRT Sequencing Data → Chromosome Assembly → Mapping and Filtering → Plasmid Assembly → Align and Polish → Prepare for Publication

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

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



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

Procedure & checklist

PacBio

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

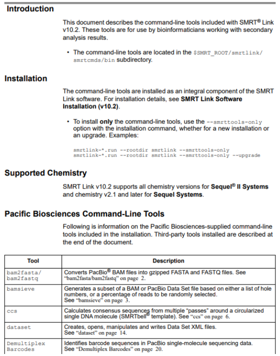
PacBio

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.



Introduction

This document describes the command-line tools included with SMRT® Link v10.2. These tools are for use by bioinformaticians working with secondary analysis results.

- The command-line tools are located in the `$SMRT_ROOT/smrtlink/smrtcmds/bin` subdirectory.

Installation

The command-line tools are installed as an integral component of the SMRT Link software. For installation details, see **SMRT Link Software Installation (v10.2)**.

- To install **only** the command-line tools, use the `--smrttools-only` option with the installation command, whether for a new installation or an upgrade. Examples:

```
smrtlink-*.run --rootdir smrtlink --smrttools-only
smrtlink-*.run --rootdir smrtlink --smrttools-only --upgrade
```

Supported Chemistry

SMRT Link v10.2 supports all chemistry versions for **Sequel® II Systems** and chemistry v2.1 and later for **Sequel Systems**.

Pacific Biosciences Command-Line Tools

Following is information on the Pacific Biosciences-supplied command-line tools included in the installation. Third-party tools installed are described at the end of the document.

Tool	Description
bam2fasta/ bam2fastq	Converts PacBio® BAM files into gzipped FASTA and FASTQ files. See "bam2fasta/bam2fastq" on page 2.
bamsieve	Generates a subset of a BAM or PacBio Data Set file based on either a list of hole numbers, or a percentage of reads to be randomly selected. See "bamsieve" on page 3.
ccs	Calculates consensus sequences from multiple "passes" around a circularized single DNA molecule (SMRTbell® template). See "ccs" on page 6.
dataset	Creates, opens, manipulates and writes Data Set XML files. See "dataset" on page 14.
Demultiplex Barcodes	Identifies barcode sequences in PacBio single-molecule sequencing data. See "Demultiplex Barcodes" on page 20.

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.

Human WGS Variant Calling

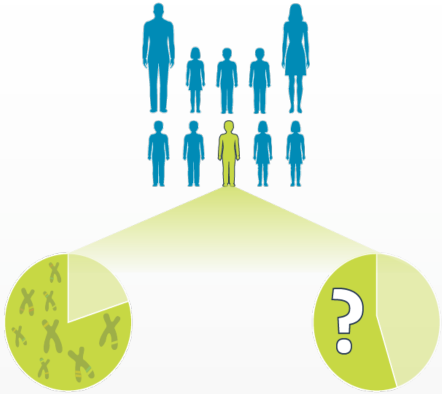
4 July 2023

彭彥蓤 Lynn Peng | Bioinformatics Engineer, Blossombio Taiwan

Rare & inherited diseases

RARE DISEASES

affect **1** in **10** individuals



80%

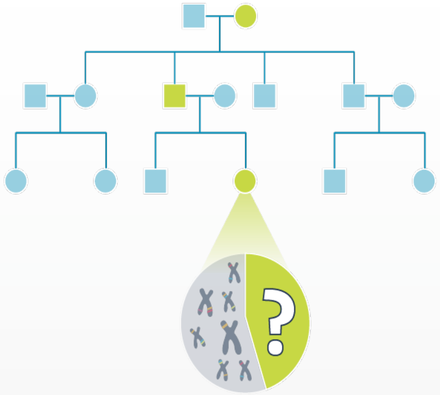
are genetic in origin

>50%

of cases remain unsolved after short-read exome or WGS

MENDELIAN DISEASES

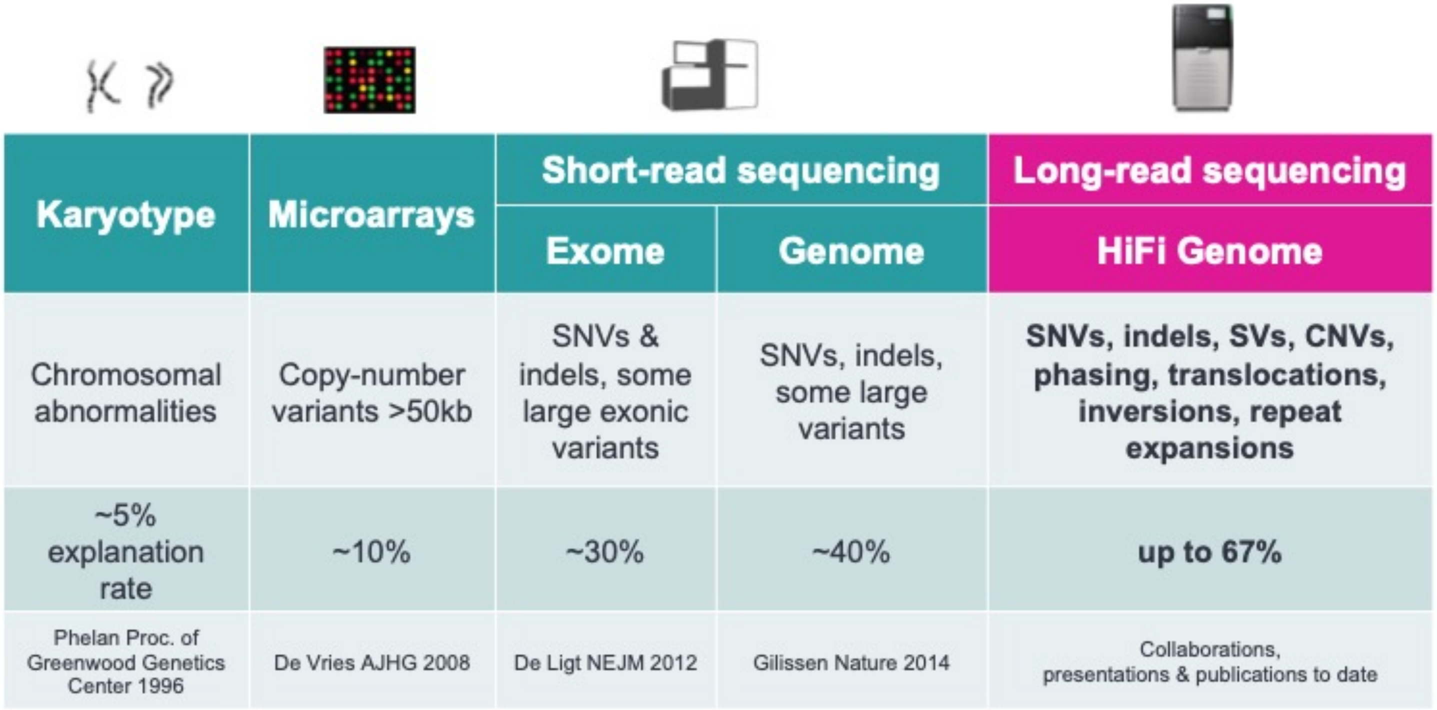
include over **8,500** known disorders



40%

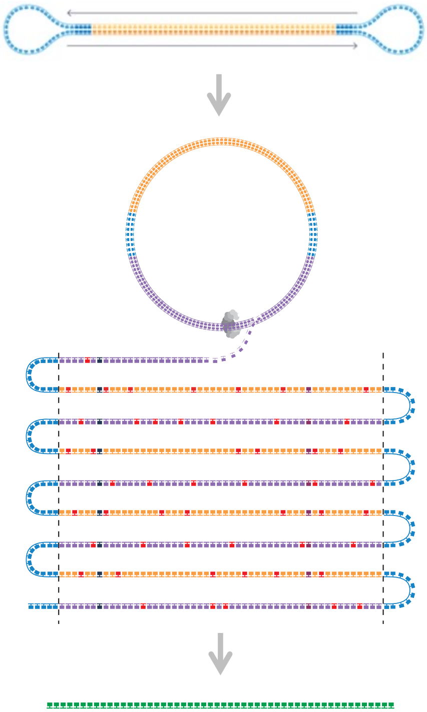
have unknown genetic cause

More complete detection yields more insights



Karyotype	Microarrays	Short-read sequencing		Long-read sequencing
		Exome	Genome	HiFi Genome
Chromosomal abnormalities	Copy-number variants >50kb	SNVs & indels, some large exonic variants	SNVs, indels, some large variants	**SNVs, indels, SVs, CNVs, phasing, translocations, inversions, repeat expansions**
~5% explanation rate	~10%	~30%	~40%	**up to 67%**
Phelan Proc. of Greenwood Genetics Center 1996	De Vries AJHG 2008	De Ligt NEJM 2012	Gilissen Nature 2014	Collaborations, presentations & publications to date

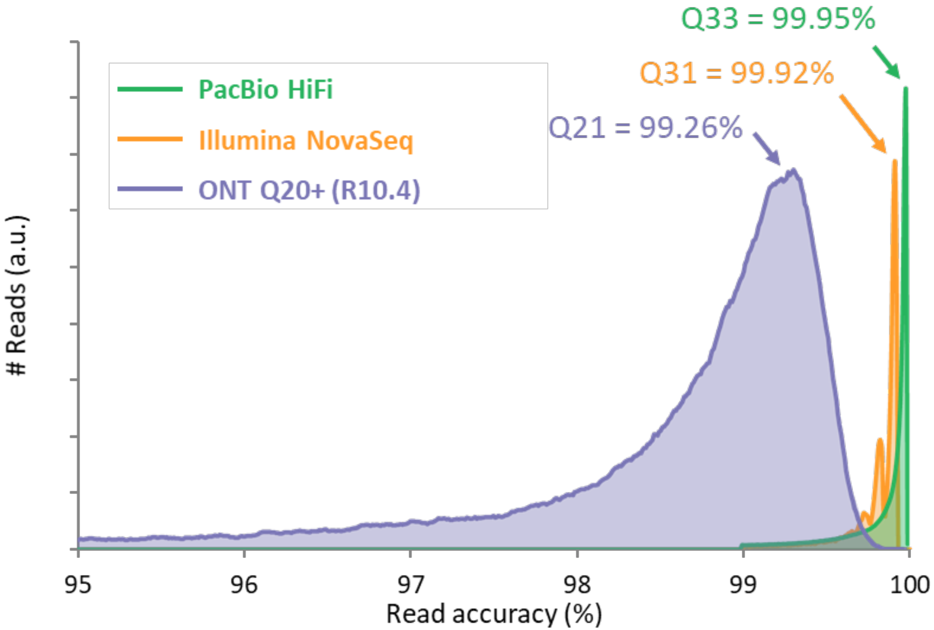
HiFi reads are long and accurate sequence reads



HiFi read

>99.9% accuracy

Up to 25 kb



PacBio HiFi: HG003 18 kb library, Sequel II System Chemistry 2.0, precisionFDA Truth Challenge V2

Illumina: HG002 2×150 bp NovaSeq library, precisionFDA Truth Challenge V2

ONT: Q20+ chemistry (R10.4, Kit 12), Oct 2021 GM24385 Dataset Release

HiFi reads underlie first telomere-to-telomere assembly of a human genome

“High accuracy long-read sequencing has finally removed this technological barrier, enabling comprehensive studies of genomic variation across the entire human genome, which we expect to drive future discovery in human genomic health and disease.”

T2T-CHM13 v2.0 assembly with sequences soft-masked using the repeat models discovered by the T2T team

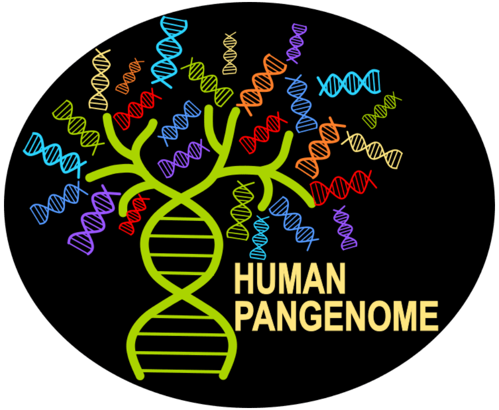
Earth's heart of iron begins to yield its secrets p. 18
Microglia in chronic pain recovery and relapse pp. 33 & 86
Particle acceleration in a nova explosion p. 77
Science
$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS
FILLING THE GAPS
Closing in on a complete human genome p. 42

PacBio

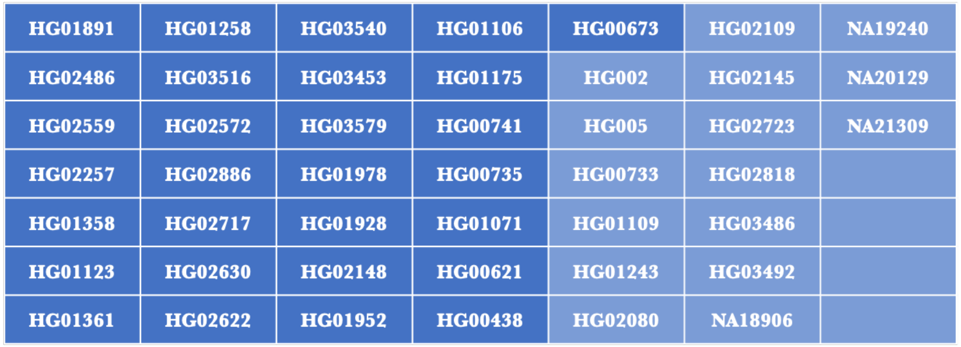
Nurk S. et al., 2021. The complete sequence of a human genomes. *bioRxiv* doi:10.1101/2021.05.26.445798

Fast, high-quality human *de novo* assemblies with HiFi reads

Human Pangenome Reference Consortium



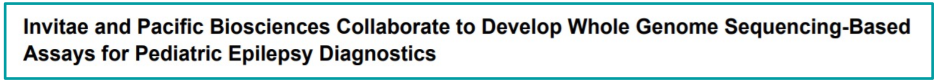
New references from
350 human genomes



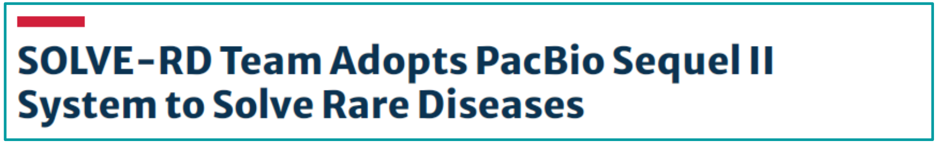
HG01891	HG01258	HG03540	HG01106	HG00673	HG02109	NA19240
HG02486	HG03516	HG03453	HG01175	HG002	HG02145	NA20129
HG02559	HG02572	HG03579	HG00741	HG005	HG02723	NA21309
HG02257	HG02886	HG01978	HG00735	HG00733	HG02818	
HG01358	HG02717	HG01928	HG01071	HG01109	HG03486	
HG01123	HG02630	HG02148	HG00621	HG01243	HG03492	
HG01361	HG02622	HG01952	HG00438	HG02080	NA18906	

https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0

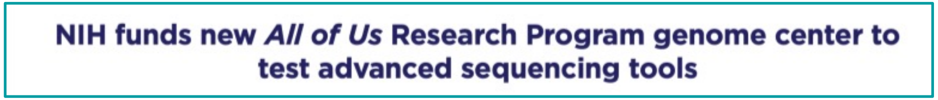
Adoption by leading medical institutes + consortia



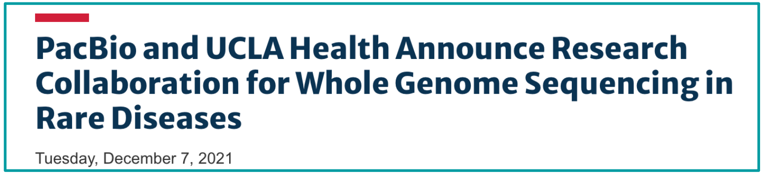
Invitae and Pacific Biosciences Collaborate to Develop Whole Genome Sequencing-Based Assays for Pediatric Epilepsy Diagnostics



SOLVE-RD Team Adopts PacBio Sequel II System to Solve Rare Diseases



NIH funds new *All of Us* Research Program genome center to test advanced sequencing tools



PacBio and UCLA Health Announce Research Collaboration for Whole Genome Sequencing in Rare Diseases

Tuesday, December 7, 2021

Children's Mercy KANSAS CITY

INVITAE

HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY

Solve RD Solving the Unsolved Rare Diseases

Stanford University

Genomic ANSWERS for KiDS

All of Us RESEARCH PROGRAM

IN·DEI·NOMINE·FELICITER

Radboud Universiteit

YCU 横浜市立大学 YOKOHAMA CITY UNIVERSITY

DISCOVERY LIFE SCIENCES

KK Women's and Children's Hospital SingHealth

UCLA Health

Genomics England

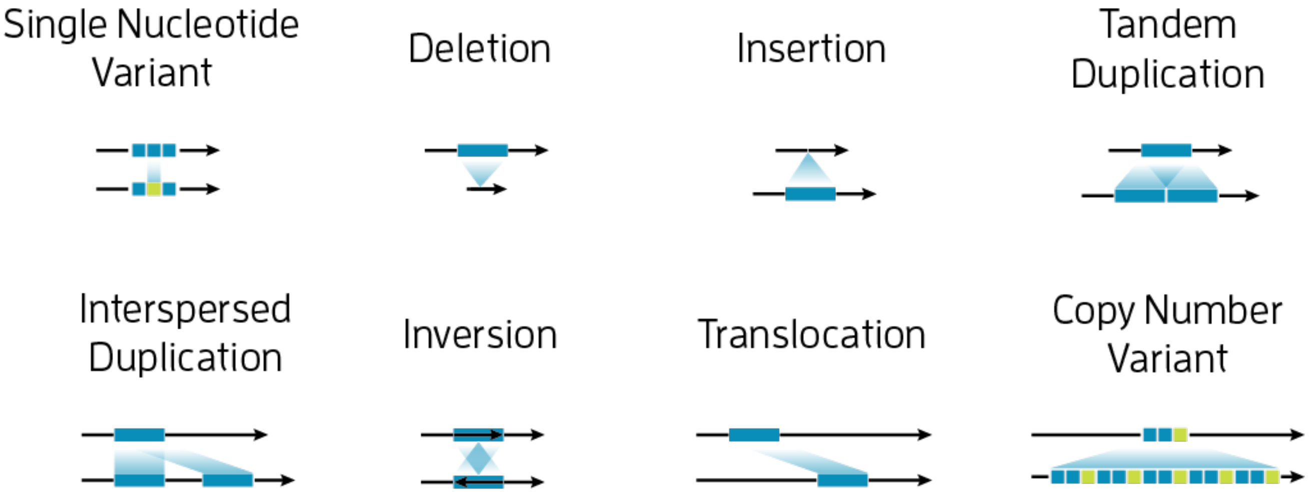
Cincinnati Children's® Hospital Medical Center

W UNIVERSITY of WASHINGTON

https://www.pacb.com/blog/solve-rd-team-adopts-pacbio-sequel-ii-system-to-solve-rare-diseases/
https://investor.pacificbiosciences.com/news-releases/news-release-details/childrens-mercy-kansas-city-teams-pacific-biosciences-fight-rare
https://allofus.nih.gov/news-events-and-media/announcements/nih-funds-new-all-us-research-program-genome-center-test-advanced-sequencing-tools
https://investor.pacificbiosciences.com/node/11431/pdf

Defining and detecting structural variants

Types of variants in a genome



Single Nucleotide Variant

Deletion

Insertion

Tandem Duplication

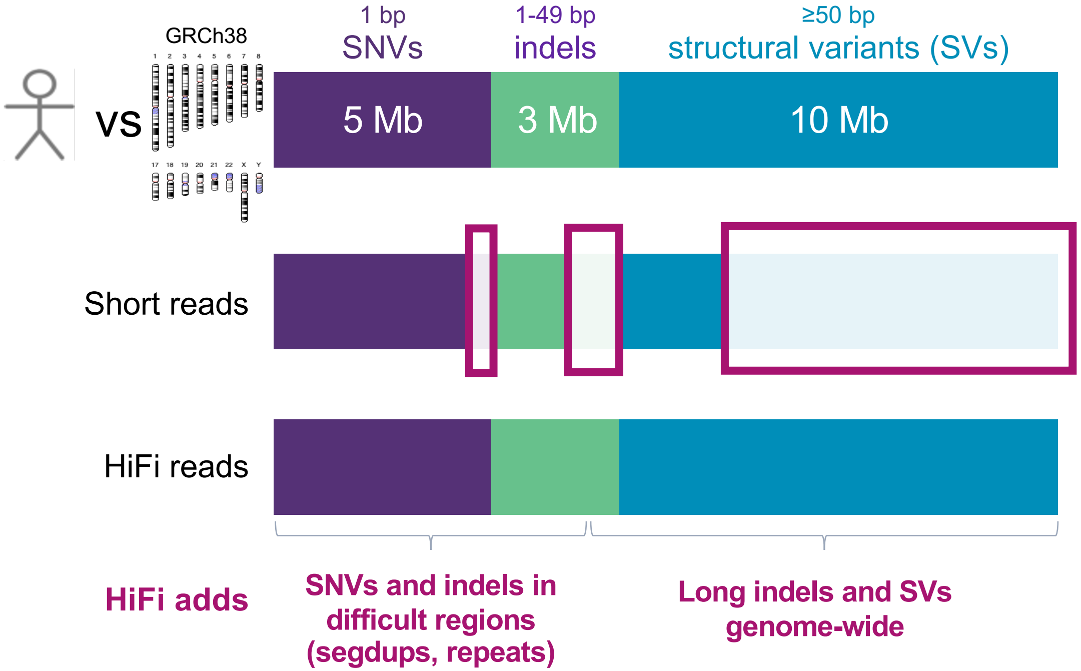
Interspersed Duplication

Inversion

Translocation

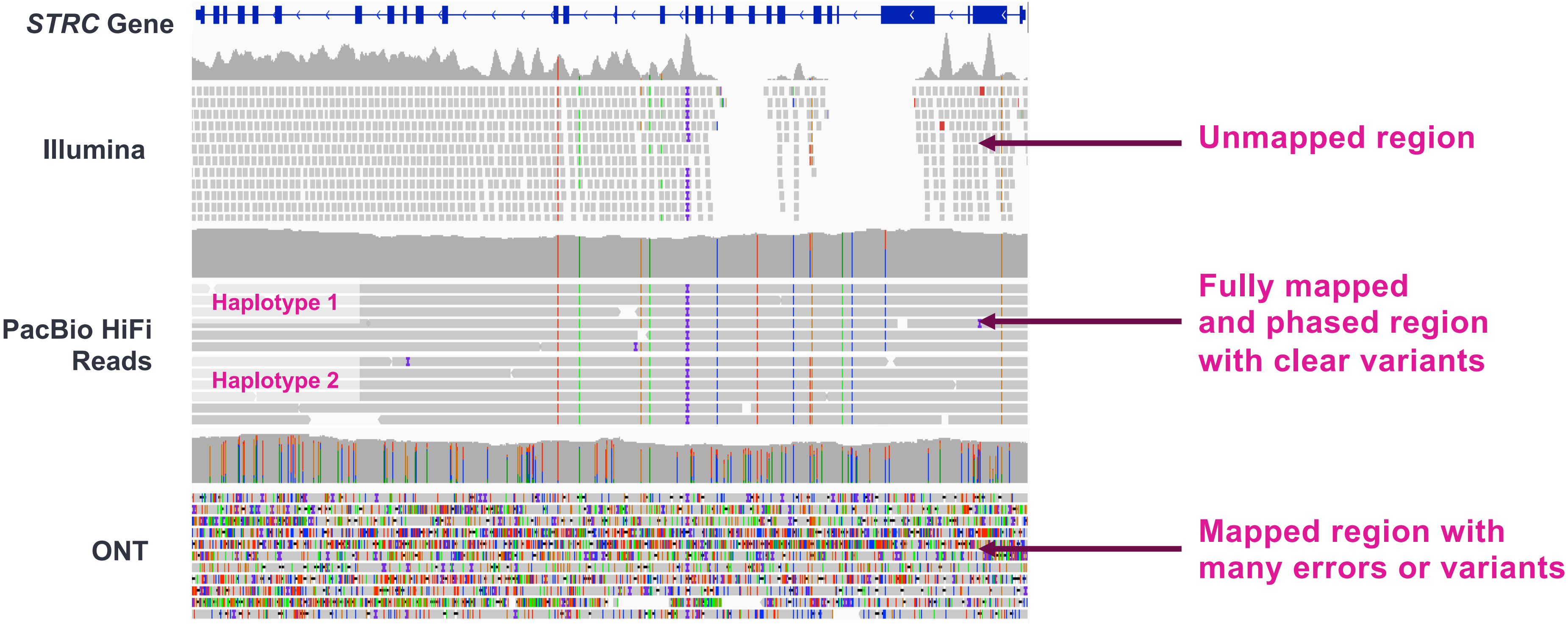
Copy Number Variant

HiFi reads provide a comprehensive view of variation in the genome


GRCh38
vs
1 bp
SNVs
1-49 bp
indels
≥50 bp
structural variants (SVs)
5 Mb
3 Mb
10 Mb
Short reads
HiFi reads
HiFi adds
SNVs and indels in difficult regions (segdups, repeats)
Long indels and SVs genome-wide

Huddleston et al. (2017) Genome Research 27(5):677-85.

Detect more variants in medically relevant genes



STRC Gene

Illumina

PacBio HiFi Reads

ONT

Unmapped region

Fully mapped and phased region with clear variants

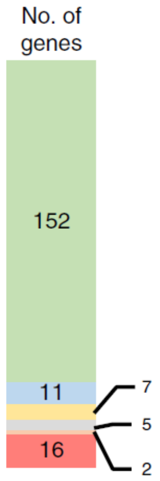
Mapped region with many errors or variants

HG002 GRCh38 chr15:43,599,422-43,619,001 (19 kb)

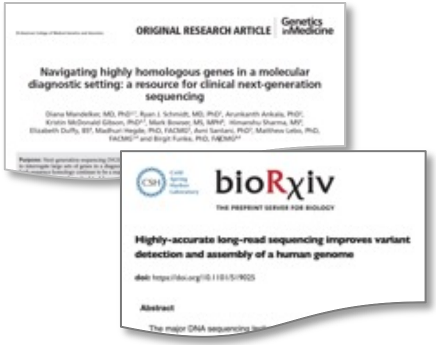
Illumina: HiSeq 2x250 (NHGRI_Illumina300X_AJtrio_novoalign_bams)
ONT: ultralong (guppy-V3.2.4_2020-01-22) / PacBio HiFi: Sequel II System Chemistry 2.0 (PacBio_CCS_15kb_20kb_chemistry2)
STRC gene alignments from Genome in a Bottle (GIAB), HG002_NA24385_son. (IGV settings)

More variants + higher accuracy in “challenging” medically-relevant genes

Percentage of problem exons mappable	Genes
100	*ABCC6, ABCD1, ACAN, ACSM2B, AKR1C2, ALG1, ANKRD11, BCR, CATSPER2, CD177, CEL, CES1, CFH, CFHR1, CFHR3, CFHR4, CGB, CHEK2, CISD2, CLCNKA, CLCNKB, CORO1A, COX10, CRYBB2, CSH1, CYP11B1, CYP11B2, CYP21A2, CYP2A6, CYP2D6, CYP2F1, CYP4A22, DDX11, DHRS4L1, DIS3L2, DND1, DPY19L2, DUOX2, ESRRA, F8, FAM120A, FAM205A, FANCD2, FCGR1A, FCGR2A, FCGR3A, FCGR3B, FLG, FLNC, FOXD4, FOXO3, FUT3, GBA, GFRA2, GON4L, GRM5, GSTM1, GYPA, GYPB, GYPE, HBA1, HBA2, HBG1, HBG2, HP, HS6ST1, IDS, IFT122, IKBKG, IL9R, KIR2DL1, KIR2DL3, KMT2C, KRT17, KRT6A, KRT6B, KRT6C, KRT81, KRT86, LEFTY2, LPA, MST1, MUC5B, MYH6, MYH7, NEB, NLGN4X, NLGN4Y, NOS2, NOTCH2, NXF5, OPN1LW, OR2T5, OR51A2, PCDH11X, PCDHB4, PGAM1, PHC1, PIK3CA, PKD1, PLA2G10, PLEKHM1, PLG, PMS2, PRB1, PRDM9, PROS1, RAB40AL, RALGAPA1, RANBP2, RHCE, RHD, RHPN2, ROCK1, SAA1, SDHA, SDHC, SFTPA1, SFTPA2, SIGLEC14, SLC6A8, SMG1, SPATA31C1, SPTLC1, SRGAP2, SSX7, STAT5B, STK19, STRC, SULT1A1, SUZ12, TBX20, TCEB3C, TLR1, TLR6, TMEM231, TNXB, TRIOBP, TRPA1, TTN, TUBA1A, TUBB2B, UGT1A5, UGT2B15, UGT2B17, UNC93B1, VCY, VWF, WDR72, ZNF419, ZNF592, ZNF674*
[75, 100)	*ANAPC1, C4A, C4B, CHRNA7, CR1, DUX4, FCGR2B, HYDIN, OTOA, PDPK1, TMLHE*
[50, 75)	*ADAMTSL2, CDY2A, DAZ1, GTF2I, NAIP, OCLN, RPS17*
[25, 50)	*DAZ2, DAZ3, KIR3DL1, OPN1MW, PPIP5K1*
(0, 25)	*NCF1, RBMY1A1*
0	*BPY2, CCL3L1, CCL4L1, CDY1, CFC1, CFC1B, GTF2IRD2, HSFY1, MRC1, OR4F5, PRY, PRY2, SMN1, SMN2, TSPY1, XKRY*



PacBio resolves most (152/193) of these genes completely with 13.5 kb reads


ORIGINAL RESEARCH ARTICLE

Genetics inMedicine

Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing

Diana Mandelker, MD, PhD, Ryan J. Schmidt, MD, PhD, Arunkanth Ankala, PhD, Kristin McDonald Gibson, PhD, Mark Bowser, MS, MPH, Himanshu Sharma, MS, Elizabeth Duffy, BS, Madhuri Hegde, PhD, FACMG, Avni Santani, PhD, Matthew Lebo, PhD, FACMG and Birgit Funke, PhD, FACMG

Purpose: [illegible]

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

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

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

Abstract

The major DNA sequencing

PacBio https://www.nature.com/articles/s41587-019-0217-9

HiFi sequencing in a rare disease cohort

Genomic ANSWERS for KiDS

Children's Mercy KANSAS CITY

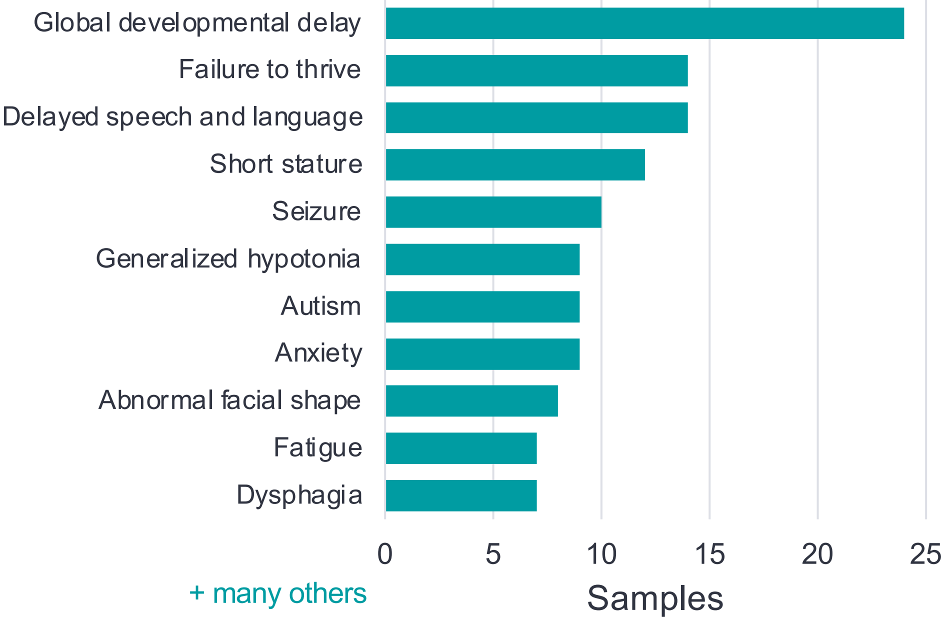
Emily Farrow

Neil Miller

Tomi Pastinen

80 singletons

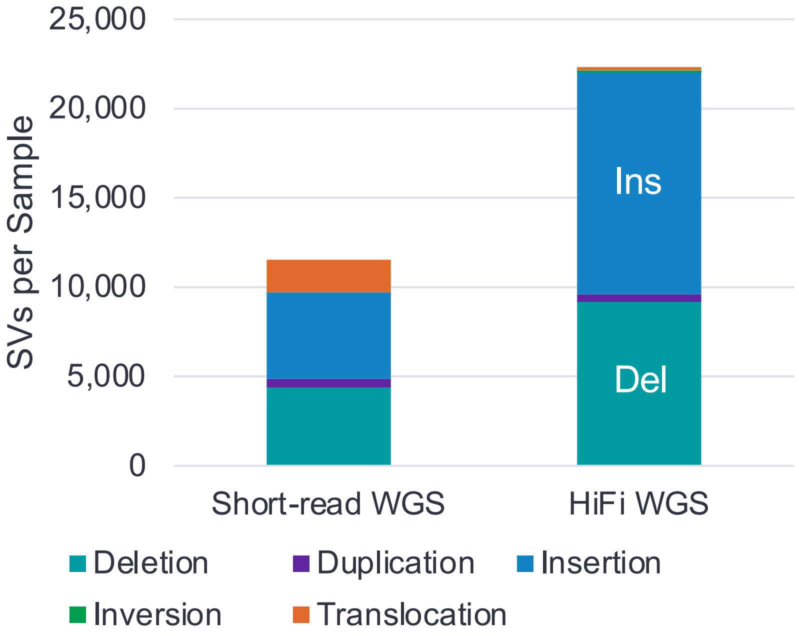
with prior short-read WGS



+ many others

PacBio

Structural variants

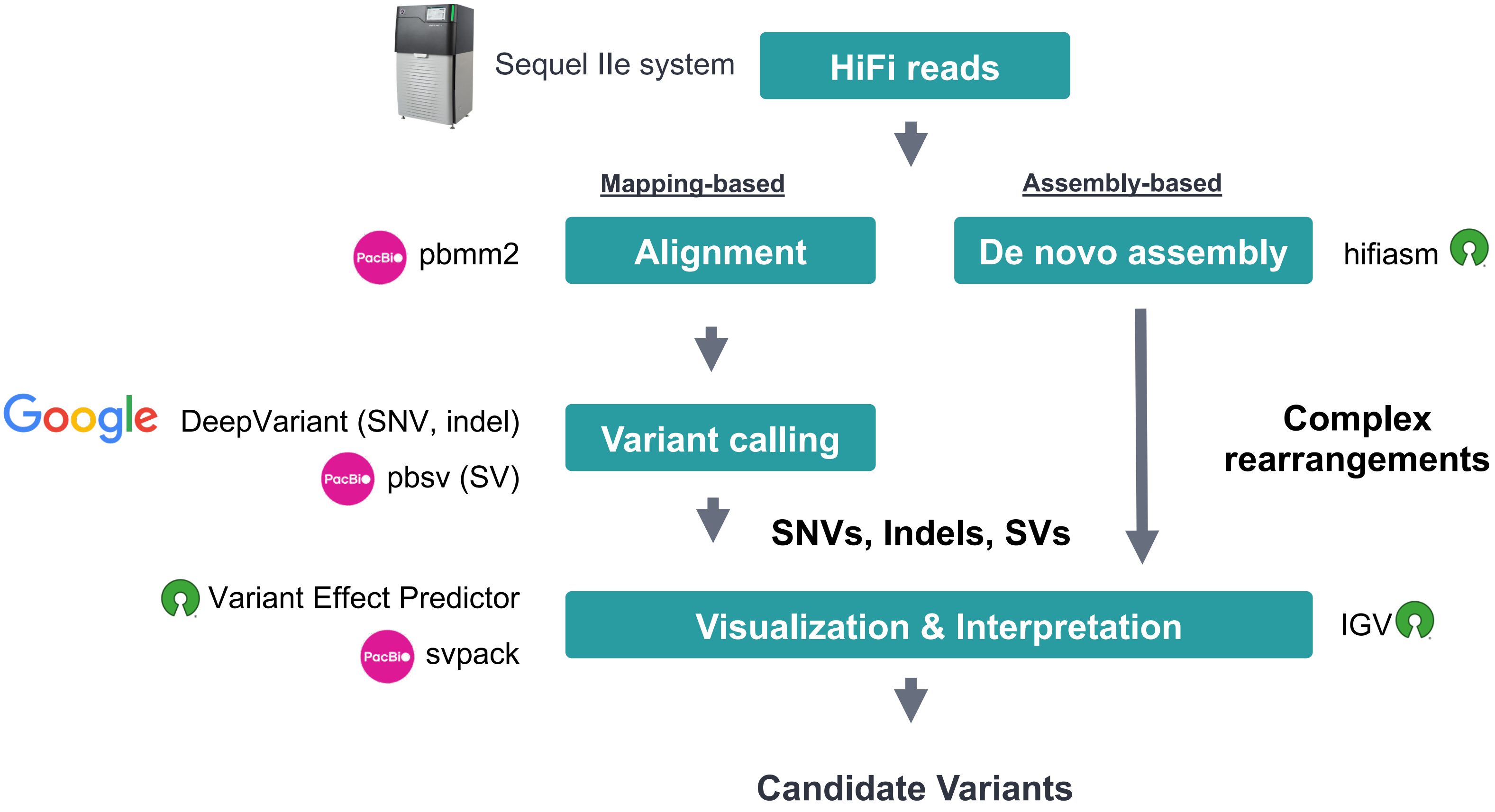


	Short-read WGS	HiFi WGS	Expected[1,2]
Deletion	4,374	9,174	9,219
Duplication	488	442	408
Insertion	4,844	12,437	14,456
Inversion	-	94	117
Translocation	1,823	162	113
Total	**11,529**	**22,309**	**24,313**

Deletion, Insertion, Inversion (average of 32 genomes): Ebert, P. et al. (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. doi:10.1126/science.abf7117; Duplication, Translocation (HG003): Sedlazeck, F.J. et al. (2018) Accurate detection of complex structural variations using single molecule sequencing. Nat Methods. doi:10.1038/s41592-018-0001-7

Workflow for WGS data analysis

Roughly 500-1000 CPU hours per sample. Can be accelerated with GPU



Sequel IIe system

HiFi reads

Mapping-based

Assembly-based

pbmm2

Alignment

De novo assembly

hifiasm

DeepVariant (SNV, indel)

pbsv (SV)

Variant calling

Complex rearrangements

SNVs, Indels, SVs

Variant Effect Predictor

svpack

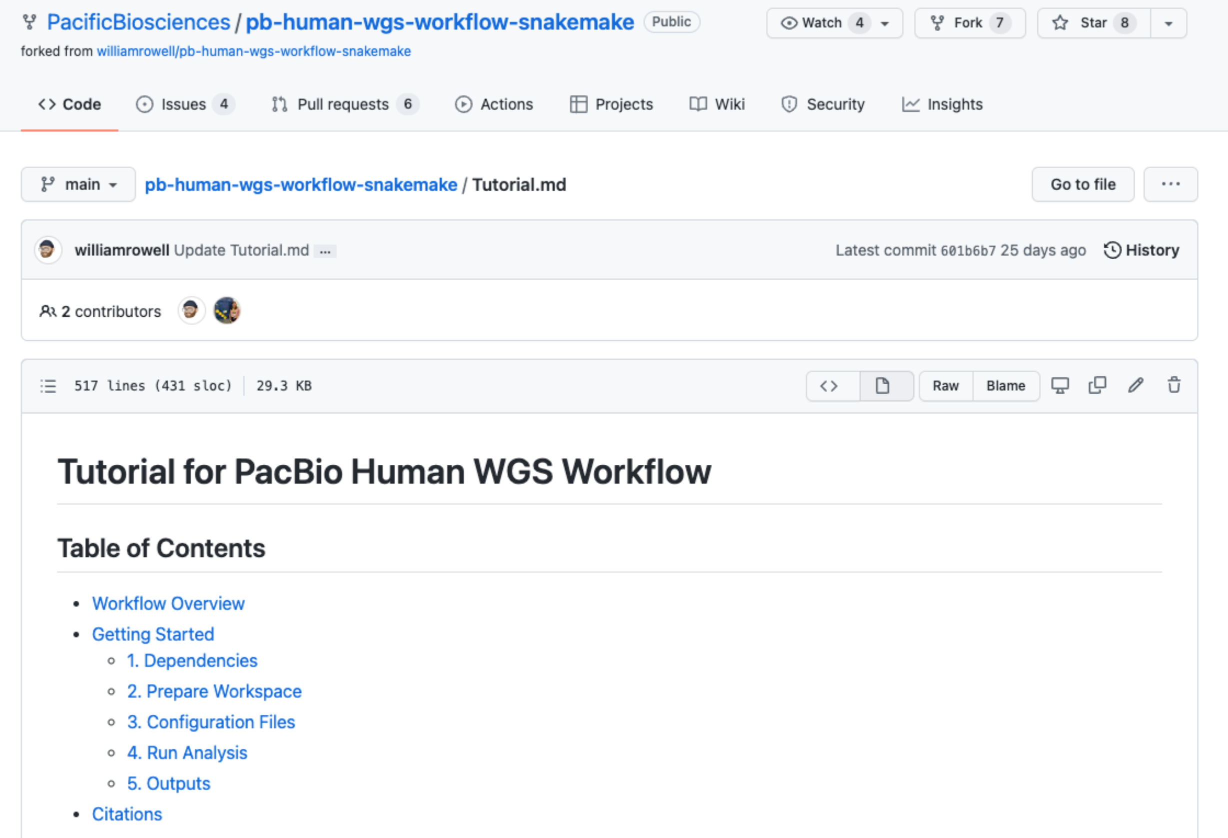
Visualization & Interpretation

IGV

Candidate Variants

PacificBiosciences/pb-human-wgs-workflow- snakemake/

PB human WGS workflow snakemake



PacificBiosciences / pb-human-wgs-workflow-snakemake Public

forked from williamrowell/pb-human-wgs-workflow-snakemake

pb-human-wgs-workflow-snakemake / Tutorial.md

williamrowell Update Tutorial.md ... Latest commit 601b6b7 25 days ago History

2 contributors

517 lines (431 sloc) | 29.3 KB

Tutorial for PacBio Human WGS Workflow

Table of Contents

- Workflow Overview
- Getting Started
  - 1. Dependencies
  - 2. Prepare Workspace
  - 3. Configuration Files
  - 4. Run Analysis
  - 5. Outputs
- Citations

https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake/blob/main/Tutorial.md

PB human WGS workflow snakemake

Process smrtcells

Aligns HiFi reads reference genome also for QC to confirm.

pbmm2

Align HiFi reads to reference genome (GRCh38)

mosdepth

- Calculate aligned coverage depth
- Generate read length and QC
- Calculate depth ratio (chrX:chrY)

jellyfish

Count kmers in HiFi reads to dump and Export modimers for sample swap detection.

Process sample

Variant discovery, variant calling, and assembly for each sample.

pbsv

Call structural variants

DeepVariant

Call small variants

Whatshap

Phased small variants and generate merged, haplotagged BAM

Hifiasm

Assemble reads

TRGT

Genotype tandem repeat

pb-cpg-tools

Generate list of CpG/5mC sites and modification probabilities

Process cohort

Variants are prioritized, annotated, and filtered find candidate rare variants with functional consequence.

pbsv

Joint call structural variants

GLnexus

Joint call small variants

slivar

Annotate and filter small variant with population AF from gnomAD and HRTC

svpack

Annotate and filter structural variant

calN50

Calculate assembly status

https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake/blob/main/Tutorial.md

PB human WGS workflow snakemake

1. Dependencies

- singularity >= 3.5.3 installed by root
- conda
- other
  - lockfile==0.12.2
  - python3
  - snakemake>=5.19
  - mamba (optional, but recommended)

Recommend at least 80 cores and 1TB RAM for local execution. Local execution will use all available cores.

The following command creates a conda environment named `pacbio-human-wgs` with the final requirements.

```
# create conda environment
conda install mamba -n base -c
conda-forge conda activate base
mamba create -c conda-forge -c bioconda -n pb-human-wgs snakemake=6.15.3 tabulate=0.8.10 pysam=0.16.0.1 python=3
conda activate pacbio-human-wgs
```

PB human WGS workflow snakemake

2.1 Prepare Workspace

- These snakemake workflows require a very **specific directory structure** in order to function properly.
- Empty directories that will store input and output files from the analysis were not built into the repo.

1

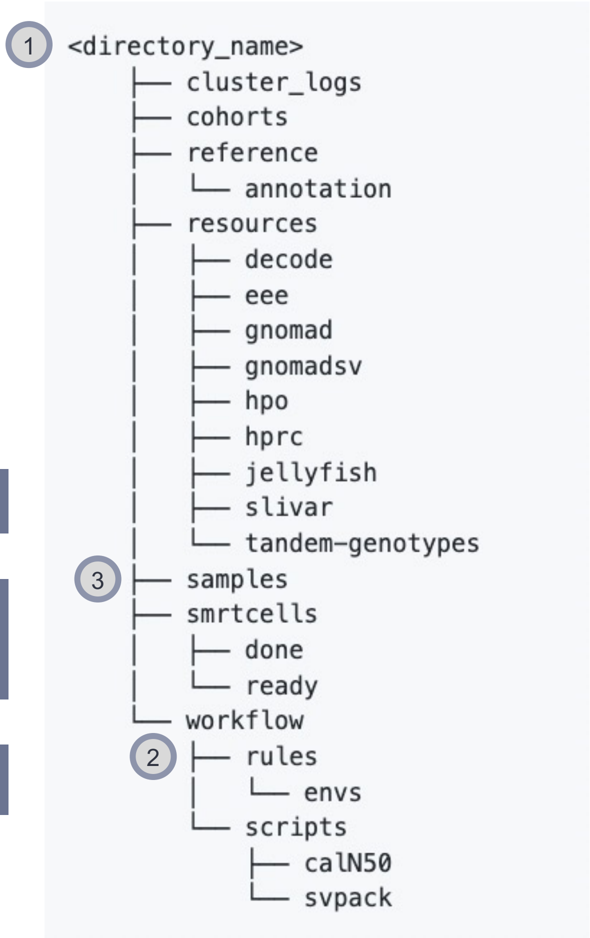
```
mkdir <directory_name> cd <directory_name>
```

2

```
git clone --recursive https://github.com/PacificBiosciences/pb-human-
wgs-workflow-snakemake.git workflow
```

3

```
mkdir -p cluster_logs smrtcells/ready smrtcells/done samples cohorts
```



```
<directory_name>
    ├── cluster_logs
    ├── cohorts
    ├── reference
    │   └── annotation
    ├── resources
    │   ├── decode
    │   ├── eee
    │   ├── gnomad
    │   ├── gnomadsv
    │   ├── hpo
    │   ├── hprc
    │   ├── jellyfish
    │   ├── slivar
    │   └── tandem-genotypes
    ├── samples
    ├── smrtcells
    │   ├── done
    │   └── ready
    └── workflow
        ├── rules
        │   └── envs
        └── scripts
            ├── calN50
            └── svpack
```

PB human WGS workflow snakemake

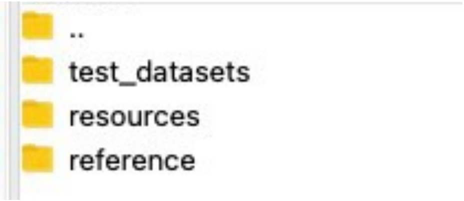
2.2 Additional folders

There are two additional folders (`reference/` and `resources/`) which contain content necessary for these workflows to run.

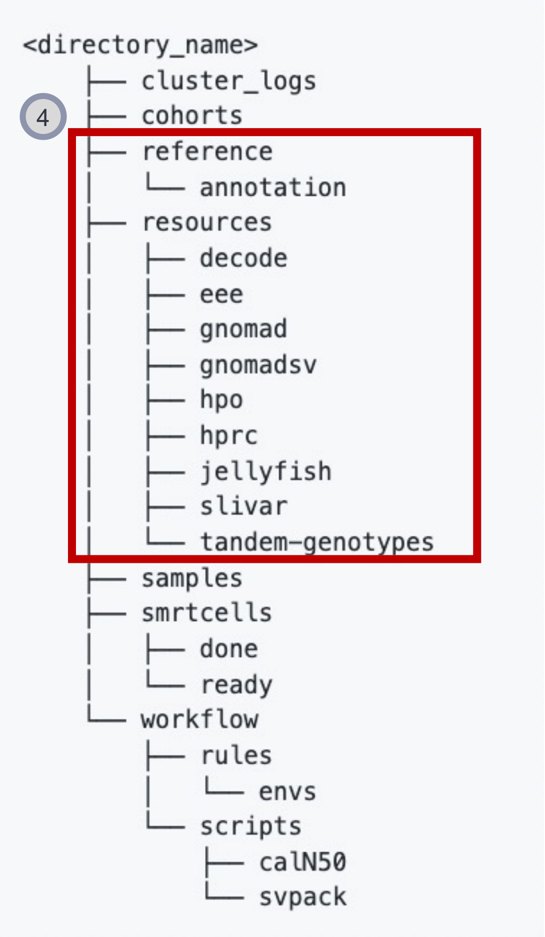
These folders can be downloaded from the ftp account below:

There are also some test datasets available from this ftp account, for making sure that the workflow runs as expected. These include only chromosomes chr2, chrX, chrY, and chrM for samples HG002, HG003, and HG004.

FileZilla



```
..
test_datasets
resources
reference
```



```
<directory_name>
        ├── cluster_logs
        ├── cohorts
        ├── reference
        │   └── annotation
        ├── resources
        │   ├── decode
        │   ├── eee
        │   ├── gnomad
        │   ├── gnomadsv
        │   ├── hpo
        │   ├── hprc
        │   ├── jellyfish
        │   ├── slivar
        │   └── tandem-genotypes
        ├── samples
        ├── smrtcells
        │   ├── done
        │   └── ready
        └── workflow
            ├── rules
            │   └── envs
            └── scripts
                ├── calN50
                └── svpack
```

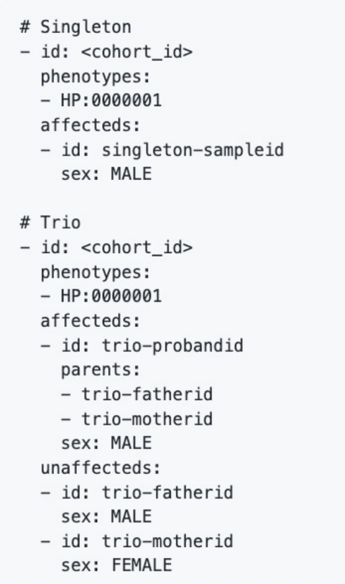
4

PB human WGS workflow snakemake

3.1 Analysis Configuration Files

Configuration files are written with `yaml` syntax. The following configuration files require your attention before running the workflows.

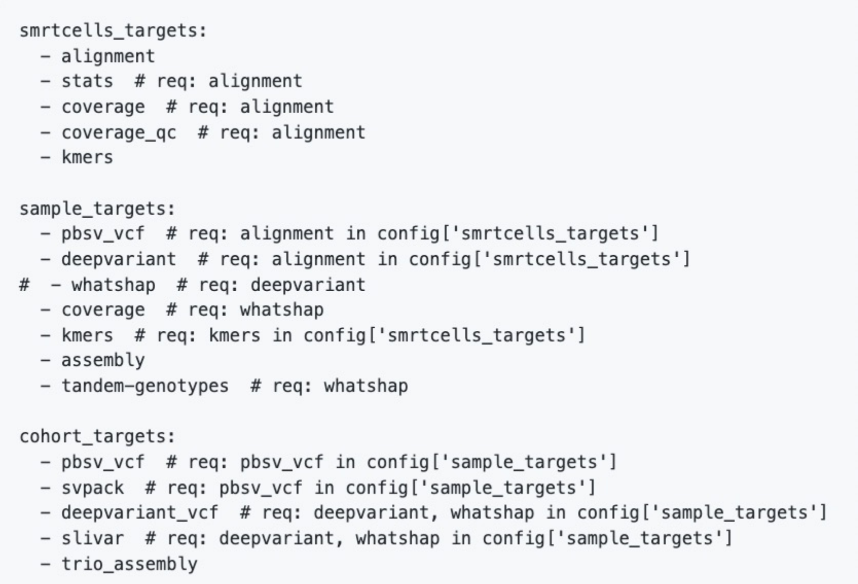
`cohort.yaml`



```
# Singleton
- id: <cohort_id>
  phenotypes:
  - HP:0000001
  affecteds:
  - id: singleton-sampleid
    sex: MALE

# Trio
- id: <cohort_id>
  phenotypes:
  - HP:0000001
  affecteds:
  - id: trio-probandid
    parents:
    - trio-fatherid
    - trio-motherid
    sex: MALE
  unaffecteds:
  - id: trio-fatherid
    sex: MALE
  - id: trio-motherid
    sex: FEMALE
```

`config.yaml`



```
smrtcells_targets:
  - alignment
  - stats  # req: alignment
  - coverage  # req: alignment
  - coverage_qc  # req: alignment
  - kmers

sample_targets:
  - pbsv_vcf  # req: alignment in config['smrtcells_targets']
  - deepvariant  # req: alignment in config['smrtcells_targets']
#  - whatshap  # req: deepvariant
  - coverage  # req: whatshap
  - kmers  # req: kmers in config['smrtcells_targets']
  - assembly
  - tandem-genotypes  # req: whatshap

cohort_targets:
  - pbsv_vcf  # req: pbsv_vcf in config['sample_targets']
  - svpack  # req: pbsv_vcf in config['sample_targets']
  - deepvariant_vcf  # req: deepvariant, whatshap in config['sample_targets']
  - slivar  # req: deepvariant, whatshap in config['sample_targets']
  - trio_assembly
```

PB human WGS workflow snakemake

3.2 Human Phenotype Ontology

The Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality, such as Deafness. The HPO is currently being developed using the medical literature, Orphanet, DECIPHER, and OMIM. HPO currently contains over 13,000 terms and over 156,000 annotations to hereditary diseases.

human phenotype ontology
All
Search for phenotypes, diseases or genes...
e.g. Arachnodactyly | Marfan syndrome | FBN1

https://hpo.jax.org/app/

PB human WGS workflow snakemake

4. Run Analysis

- Input data

Create a directory for each sample in smrtcells/ready. The names of these directories must match the sample IDs specified in cohort.yaml.

```
mkdir smrtcells/ready/<sample_id>
```

Put PacBio HiFi reads into their respective directories. The easiest way to do this is with a symlink. **Note: unaligned BAM and FASTQ filenames must be identifiable as HiFi reads, i.e. have the following format.**

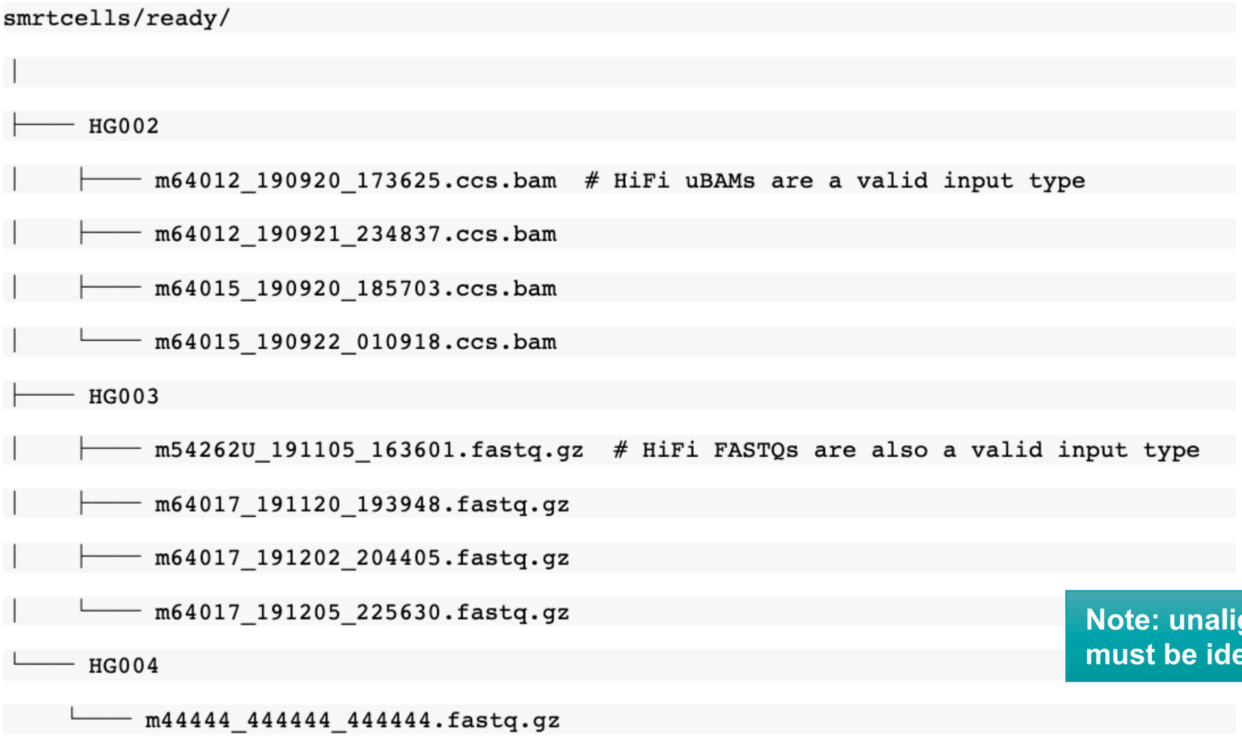
4. regex for BAM: /m\d{5}[Ue]?_\d{6}_\d{6}.(ccs|hifi_reads).bam
    4. example: m54119U_210108_012126.ccs.bam
    5. example: m64013e_210917_004210.hifi_reads.bam
5. regex for FASTQ: /m\d{5}[Ue]?_\d{6}_\d{6}.fastq.gz
    4. example: m54119U_210108_012126.fastq.gz
    5. example: m64013e_210917_004210.fastq.gz

```
ln -s /path/to/HiFi/BAM/or/FASTQ/<hifi_reads_filename> smrtcells/ready/<sample_id>/
```

PB human WGS workflow snakemake

4. Run Analysis

- Example Trio sample



```
smrtcells/ready/
│
├── HG002
│   ├── m64012_190920_173625.ccs.bam  # HiFi uBAMs are a valid input type
│   ├── m64012_190921_234837.ccs.bam
│   ├── m64015_190920_185703.ccs.bam
│   └── m64015_190922_010918.ccs.bam
├── HG003
│   ├── m54262U_191105_163601.fastq.gz  # HiFi FASTQs are also a valid input type
│   ├── m64017_191120_193948.fastq.gz
│   ├── m64017_191202_204405.fastq.gz
│   └── m64017_191205_225630.fastq.gz
└── HG004
    └── m44444_444444_444444.fastq.gz
```

Note: unaligned BAM and FASTQ filenames must be identifiable as HiFi reads

PB human WGS workflow snakemake

4. Run Analysis

- Run process workflow

This will process all samples located in `smrtcells/ready`. If you have samples in this folder that you don't want to process, move them to `smrtcells/done,` and make sure to re-activate the conda environment before submitting the job

```
sbatch workflow/process_smrtcells.sh
```

```
sbatch workflow/process_sample.sh <sample_id>
```

```
sbatch workflow/process_cohort.sh <cohort_id>
```

The following instructions are specific to a slurm cluster (i.e. sbatch). If not, just use bash command (i.e. sh `workflow/process_smrtcells.sh`).

PB human WGS workflow snakemake

Outputs

Process smartcells

```
$ tree -dL 1 samples/<sample_id>
samples/<sample_id>
├── aligned
├── benchmarks
├── deepvariant
├── hifiasm
├── jellyfish
├── logs
├── mosdepth
├── pbsv
├── smrtcell_stats
├── tandem-genotypes
├── trgt
└── whatshap

12 directories
```

Process sample

BAM

VCF

FASTA

Process cohort

```
$ tree -dL 1 cohorts/<cohort_id>
cohorts/<cohort_id>
├── benchmarks
├── glnexus
├── hifiasm
├── logs
├── pbsv
├── slivar
├── svpack
└── whatshap

8 directories
```

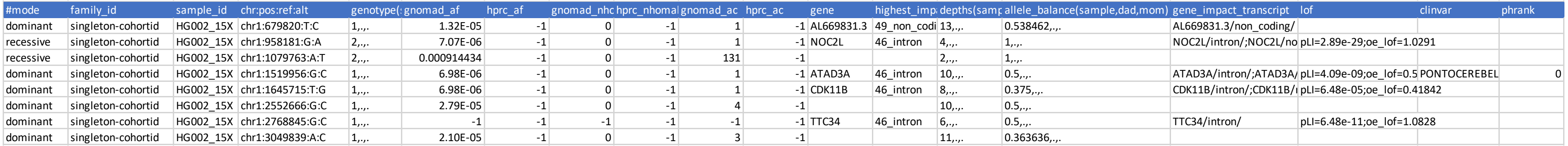
VCF/TSV

The following are some of the key output files from these workflows. The haplotype-resolved assembly is only produced when a cohort includes one or more trios (child and both parents).

Annotated Small and Structural Variants

Small variant calls

Small variants and compound heterozygotes that are filtered based on **population frequency** and annotated with **cohort information**, **population AF**, **gene, functional impact**, etc by slivar.

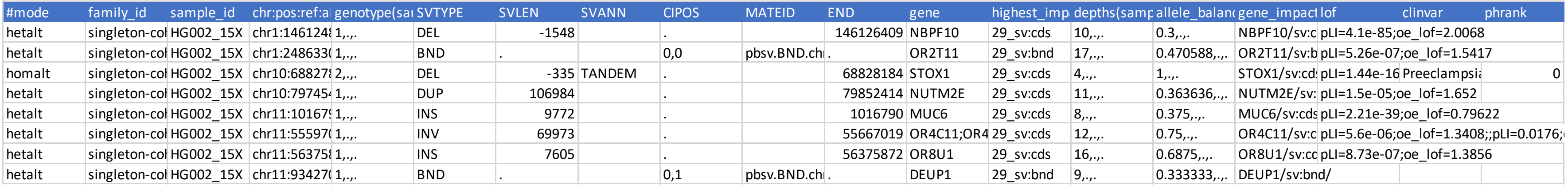


#mode	family_id	sample_id	chr:pos:ref:alt	genotype(	gnomad_af	hprc_af	gnomad_nhc	hprc_nhomal	gnomad_ac	hprc_ac	gene	highest_imp	depths(samp	allele_balance(sample,dad,mom)	gene_impact_transcript	lof	clinvar	phrank
dominant	singleton-cohortid	HG002_15X	chr1:679820:T:C	1,.,.	1.32E-05	-1	0	-1	1	-1	AL669831.3	49_non_codi	13,.,.	0.538462,.,.	AL669831.3/non_coding/			
recessive	singleton-cohortid	HG002_15X	chr1:958181:G:A	2,.,.	7.07E-06	-1	0	-1	1	-1	NOC2L	46_intron	4,.,.	1,.,.	NOC2L/intron/;NOC2L/no	pLI=2.89e-29;oe_lof=1.0291		
recessive	singleton-cohortid	HG002_15X	chr1:1079763:A:T	2,.,.	0.000914434	-1	0	-1	131	-1			2,.,.	1,.,.				
dominant	singleton-cohortid	HG002_15X	chr1:1519956:G:C	1,.,.	6.98E-06	-1	0	-1	1	-1	ATAD3A	46_intron	10,.,.	0.5,.,.	ATAD3A/intron/;ATAD3A/	pLI=4.09e-09;oe_lof=0.5	PONTOCEREBEL	0
dominant	singleton-cohortid	HG002_15X	chr1:1645715:T:G	1,.,.	6.98E-06	-1	0	-1	1	-1	CDK11B	46_intron	8,.,.	0.375,.,.	CDK11B/intron/;CDK11B/	pLI=6.48e-05;oe_lof=0.41842		
dominant	singleton-cohortid	HG002_15X	chr1:2552666:G:C	1,.,.	2.79E-05	-1	0	-1	4	-1			10,.,.	0.5,.,.				
dominant	singleton-cohortid	HG002_15X	chr1:2768845:G:C	1,.,.	-1	-1	-1	-1	-1	-1	TTC34	46_intron	6,.,.	0.5,.,.	TTC34/intron/	pLI=6.48e-11;oe_lof=1.0828		
dominant	singleton-cohortid	HG002_15X	chr1:3049839:A:C	1,.,.	2.10E-05	-1	0	-1	3	-1			11,.,.	0.363636,.,.				

`singleton-cohortid.GRCh38.deepvariant.phased.slivar.tsv`

Structural Variant calls

Structural variants that are filtered based on **population frequency** and annotated with **cohort information**, **population AF**, **gene, functional impact**, etc by svpack.



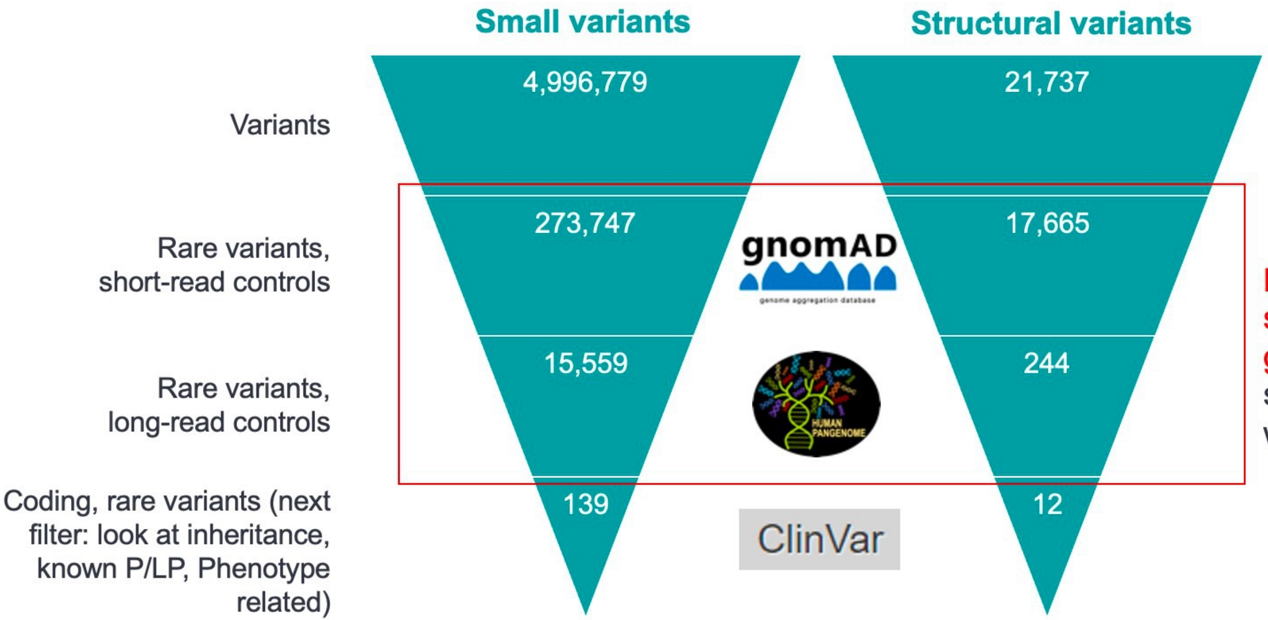
#mode	family_id	sample_id	chr:pos:ref:al	genotype(sa	SVTYPE	SVLEN	SVANN	CIPOS	MATEID	END	gene	highest_imp	depths(samp	allele_balan	gene_impact	lof	clinvar	phrank
hetalt	singleton-col	HG002_15X	chr1:1461248	1,.,.	DEL	-1548		.		146126409	NBPF10	29_sv:cds	10,.,.	0.3,.,.	NBPF10/sv:c	pLI=4.1e-85;oe_lof=2.0068		
hetalt	singleton-col	HG002_15X	chr1:248633	1,.,.	BND	.		0,0	pbsv.BND.ch	.	OR2T11	29_sv:bnd	17,.,.	0.470588,.,.	OR2T11/sv:b	pLI=5.26e-07;oe_lof=1.5417		
homalt	singleton-col	HG002_15X	chr10:688278	2,.,.	DEL	-335	TANDEM	.		68828184	STOX1	29_sv:cds	4,.,.	1,.,.	STOX1/sv:cd	pLI=1.44e-16	Preeclampsi	0
hetalt	singleton-col	HG002_15X	chr10:797454	1,.,.	DUP	106984		.		79852414	NUTM2E	29_sv:cds	11,.,.	0.363636,.,.	NUTM2E/sv:	pLI=1.5e-05;oe_lof=1.652		
hetalt	singleton-col	HG002_15X	chr11:101679	1,.,.	INS	9772		.		1016790	MUC6	29_sv:cds	8,.,.	0.375,.,.	MUC6/sv:cds	pLI=2.21e-39;oe_lof=0.79622		
hetalt	singleton-col	HG002_15X	chr11:555970	1,.,.	INV	69973		.		55667019	OR4C11;OR4	29_sv:cds	12,.,.	0.75,.,.	OR4C11/sv:c	pLI=5.6e-06;oe_lof=1.3408;;pLI=0.0176;		
hetalt	singleton-col	HG002_15X	chr11:563758	1,.,.	INS	7605		.		56375872	OR8U1	29_sv:cds	16,.,.	0.6875,.,.	OR8U1/sv:cc	pLI=8.73e-07;oe_lof=1.3856		
hetalt	singleton-col	HG002_15X	chr11:934270	1,.,.	BND	.		0,1	pbsv.BND.ch	.	DEUP1	29_sv:bnd	9,.,.	0.333333,.,.	DEUP1/sv:bnd/			

`singleton-cohortid.GRCh38.pbsv.svpack.tsv`

PacBio

Population Frequency Filtering Is Necessary for NGS Genetic Disease analysis/interpretation

Frequency database (gnomAD) and database like Clinvar, HGMD etc are the real power behind 3rd analysis. Without this data, interpretation would not fully extract benefit of increased SV detection



	Small variants	Structural variants
Variants	4,996,779	21,737
Rare variants, short-read controls	273,747	17,665
Rare variants, long-read controls	15,559	244
Coding, rare variants (next filter: look at inheritance, known P/LP, Phenotype related)	139	12

PacBio Current State: Using summary data from 40 long read genomes for freq. filtering – Building something with more power is what we propose

Filter for rare SNVs that impact a gene
- max_gnmad_af: 0.01
- max_hprc_af:0.01

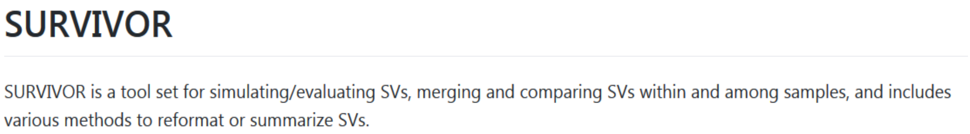
Filter for rare SVs that impact a gene
- confident SV calls (PASS calls)
- SV calls not seen in population controls (rare variants)
- SV calls that impact a coding gene

PacBio

The Genome Aggregation Datatbase (gnomAD)
https://gnomad.broadinstitute.org/

Human Pangenome Reference Consortium (HPRC)
https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0

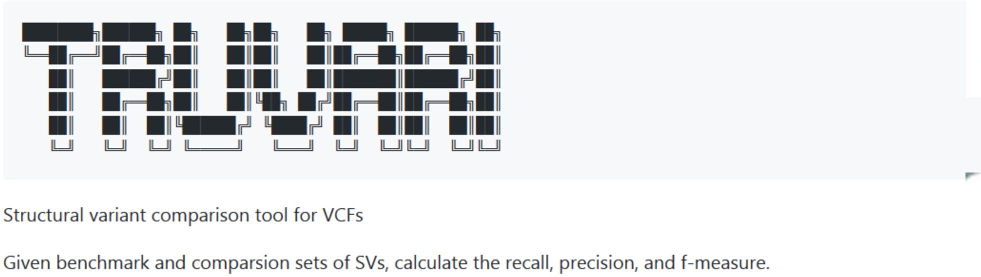
Downstream tools



SURVIVOR

SURVIVOR is a tool set for simulating/evaluating SVs, merging and comparing SVs within and among samples, and includes various methods to reformat or summarize SVs.

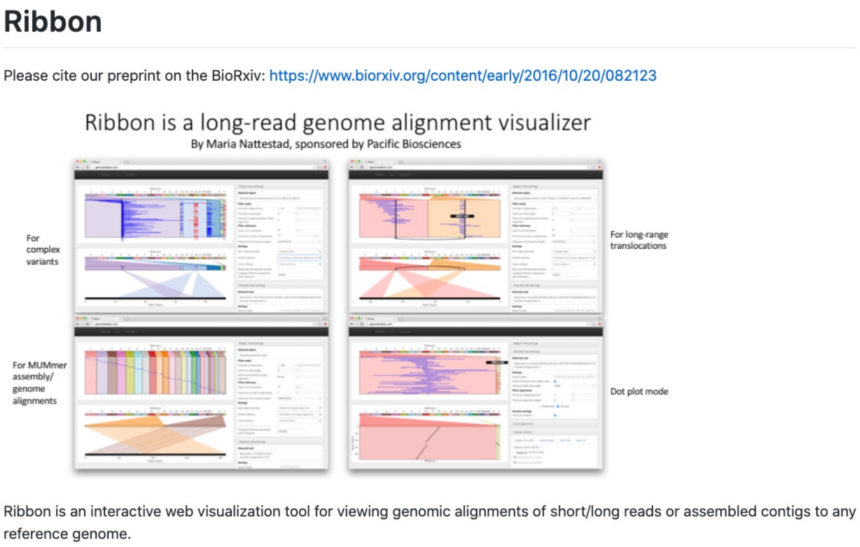
https://github.com/fritzsedlazeck/SURVIVOR



Structural variant comparison tool for VCFs

Given benchmark and comparsion sets of SVs, calculate the recall, precision, and f-measure.

https://github.com/spiralgenetics/truvari



Ribbon

Please cite our preprint on the BioRxiv: https://www.biorxiv.org/content/early/2016/10/20/082123

Ribbon is a long-read genome alignment visualizer

By Maria Nattestad, sponsored by Pacific Biosciences

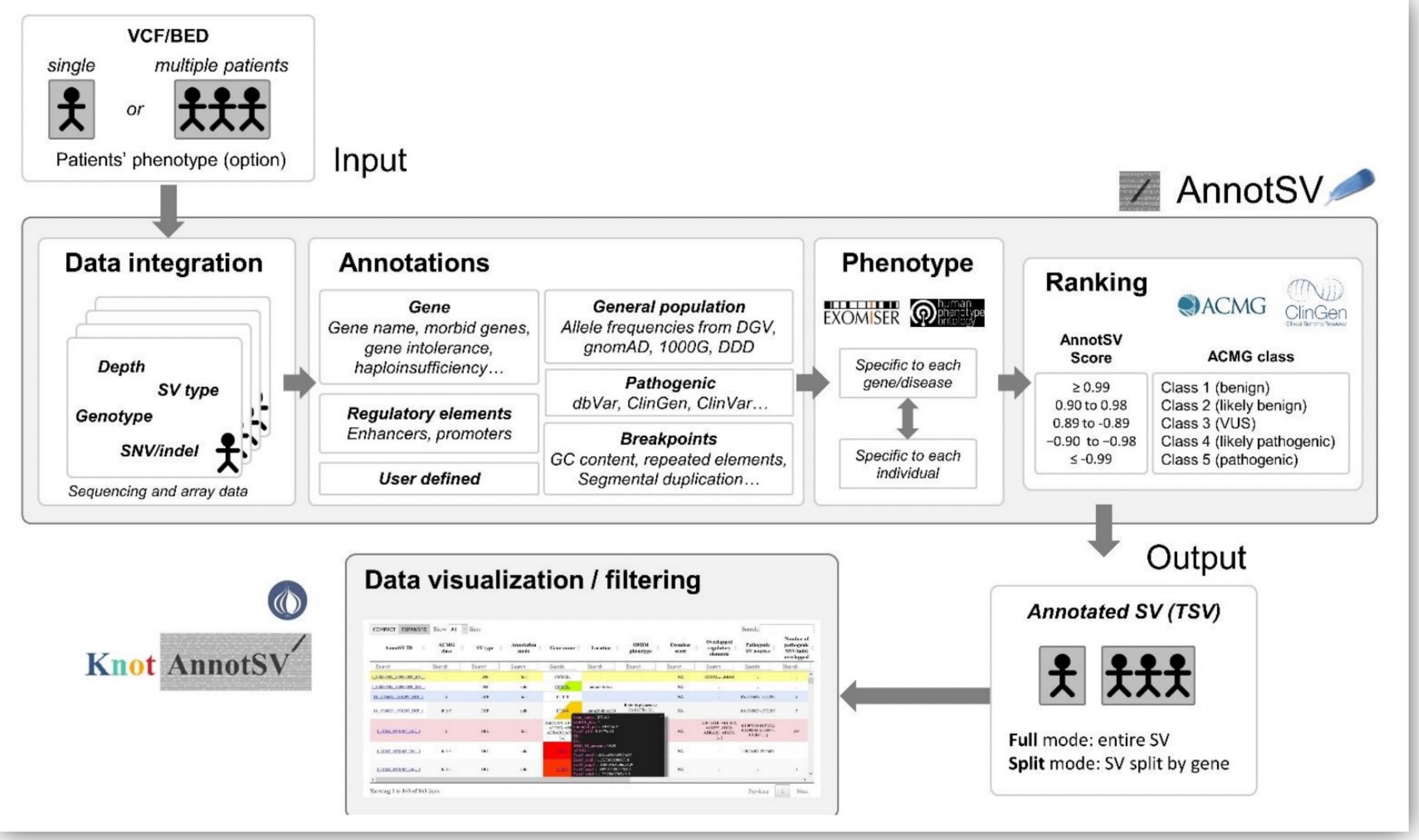
For complex variants
For long-range translocations
For MUMmer assembly/ genome alignments
Dot plot mode

Ribbon is an interactive web visualization tool for viewing genomic alignments of short/long reads or assembled contigs to any reference genome.

https://github.com/MariaNattestad/ribbon

http://software.broadinstitute.org/software/igv/

AnnotSV



VCF/BED

single *or* *multiple patients*

Patients' phenotype (option)

Input

AnnotSV

Data integration

Depth

SV type

Genotype

SNV/indel

Sequencing and array data

Annotations

Gene
Gene name, morbid genes, gene intolerance, haploinsufficiency…

Regulatory elements
Enhancers, promoters

User defined

General population
Allele frequencies from DGV, gnomAD, 1000G, DDD

Pathogenic
dbVar, ClinGen, ClinVar…

Breakpoints
GC content, repeated elements, Segmental duplication…

Phenotype

EXOMISER human phenotype ontology

Specific to each gene/disease

Specific to each individual

Ranking

ACMG ClinGen

AnnotSV Score	ACMG class
≥ 0.99	Class 1 (benign)
0.90 to 0.98	Class 2 (likely benign)
0.89 to -0.89	Class 3 (VUS)
-0.90 to -0.98	Class 4 (likely pathogenic)
≤ -0.99	Class 5 (pathogenic)

Output

Annotated SV (TSV)

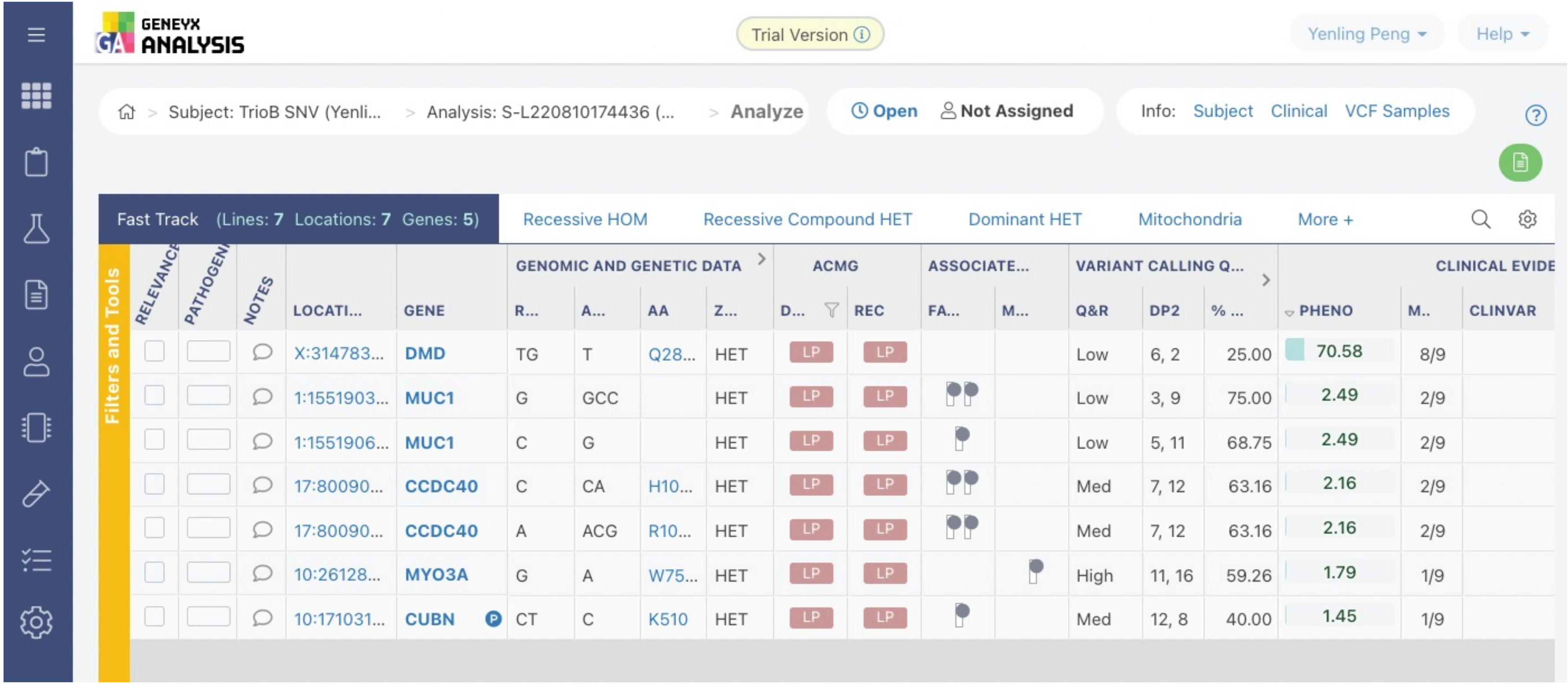
Full mode: entire SV

Split mode: SV split by gene

Data visualization / filtering

Knot AnnotSV

GENEYX



⌂ > Subject: TrioB SNV (Yenli... > Analysis: S-L220810174436 (... > Analyze

Open · Not Assigned

Info: Subject Clinical VCF Samples

Fast Track (Lines: 7 Locations: 7 Genes: 5) | Recessive HOM | Recessive Compound HET | Dominant HET | Mitochondria | More +

					GENOMIC AND GENETIC DATA				ACMG		ASSOCIATE...		VARIANT CALLING Q...			CLINICAL EVIDE		
RELEVANCE	PATHOGEN	NOTES	LOCATI...	GENE	R...	A...	AA	Z...	D...	REC	FA...	M...	Q&R	DP2	% ...	PHENO	M..	CLINVAR
□	□	💬	X:314783...	DMD	TG	T	Q28...	HET	LP	LP			Low	6, 2	25.00	70.58	8/9	
□	□	💬	1:1551903...	MUC1	G	GCC		HET	LP	LP	●●		Low	3, 9	75.00	2.49	2/9	
□	□	💬	1:1551906...	MUC1	C	G		HET	LP	LP	●		Low	5, 11	68.75	2.49	2/9	
□	□	💬	17:80090...	CCDC40	C	CA	H10...	HET	LP	LP	●●		Med	7, 12	63.16	2.16	2/9	
□	□	💬	17:80090...	CCDC40	A	ACG	R10...	HET	LP	LP	●●		Med	7, 12	63.16	2.16	2/9	
□	□	💬	10:26128...	MYO3A	G	A	W75...	HET	LP	LP		●	High	11, 16	59.26	1.79	1/9	
□	□	💬	10:171031...	CUBN Ⓟ	CT	C	K510	HET	LP	LP	●		Med	12, 8	40.00	1.45	1/9	

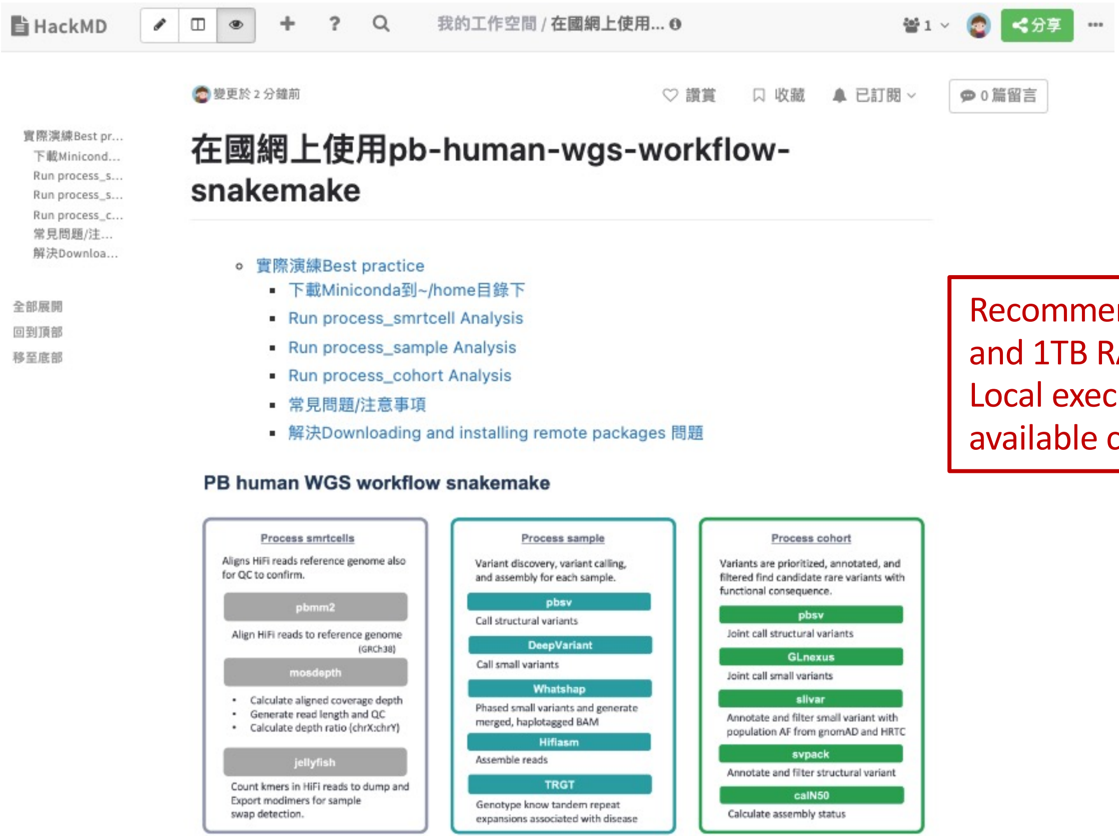
PacBio https://geneyx.com/

References

https://www.pacb.com/applications/whole-genome-sequencing/structural-variation/

- Application Brief: Structural Variant Detection Using Whole Genome Sequencing Best Practices
- Structural Variation Project Calculator
- Whitepaper
- Video (Tutorials and Conference Proceedings)
- Publications
- Example datasets: https://github.com/PacificBiosciences/DevNet/wiki/Datasets
- SMRT Link User Guide PDF (GUI)
- SMRT Tools Reference Guide PDF (CLI)
- *pbsv* online documentation
- minimap2 repository

Using pb-human-wgs-workflow-snakemake on NCHC



在國網上使用pb-human-wgs-workflow-snakemake

PB human WGS workflow snakemake

Process smrtcells

Aligns HiFi reads reference genome also for QC to confirm.

pbmm2

Align HiFi reads to reference genome (GRCh38)

mosdepth

- Calculate aligned coverage depth
- Generate read length and QC
- Calculate depth ratio (chrX:chrY)

jellyfish

Count kmers in HiFi reads to dump and Export modimers for sample swap detection.

Process sample

Variant discovery, variant calling, and assembly for each sample.

pbsv

Call structural variants

DeepVariant

Call small variants

Whatshap

Phased small variants and generate merged, haplotagged BAM

Hifiasm

Assemble reads

TRGT

Genotype know tandem repeat expansions associated with disease

Process cohort

Variants are prioritized, annotated, and filtered find candidate rare variants with functional consequence.

pbsv

Joint call structural variants

GLnexus

Joint call small variants

slivar

Annotate and filter small variant with population AF from gnomAD and HRTC

svpack

Annotate and filter structural variant

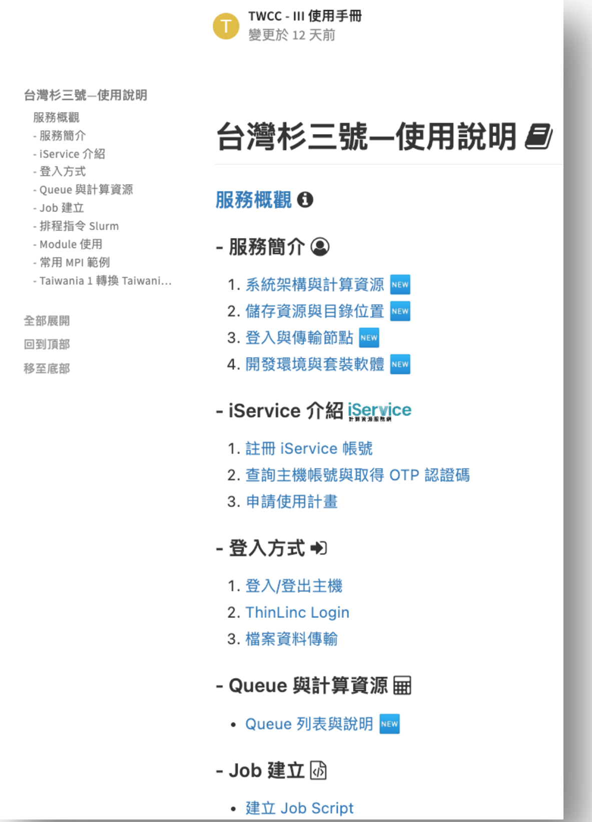
calN50

Calculate assembly status

Recommend at least 80 cores and 1TB RAM for local execution. Local execution will use all available cores.

https://hackmd.io/@YenlingPeng/SJSG29KH2

Using pb-human-wgs-workflow-snakemake on NCHC



TWCC - III 使用手冊
變更於 12 天前

台灣杉三號—使用說明
服務概觀
- 服務簡介
- iService 介紹
- 登入方式
- Queue 與計算資源
- Job 建立
- 排程指令 Slurm
- Module 使用
- 常用 MPI 範例
- Taiwania 1 轉換 Taiwani...

全部展開
回到頂部
移至底部

台灣杉三號—使用說明

服務概觀

- 服務簡介

1. 系統架構與計算資源 NEW
2. 儲存資源與目錄位置 NEW
3. 登入與傳輸節點 NEW
4. 開發環境與套裝軟體 NEW

- iService 介紹 iService

1. 註冊 iService 帳號
2. 查詢主機帳號與取得 OTP 認證碼
3. 申請使用計畫

- 登入方式

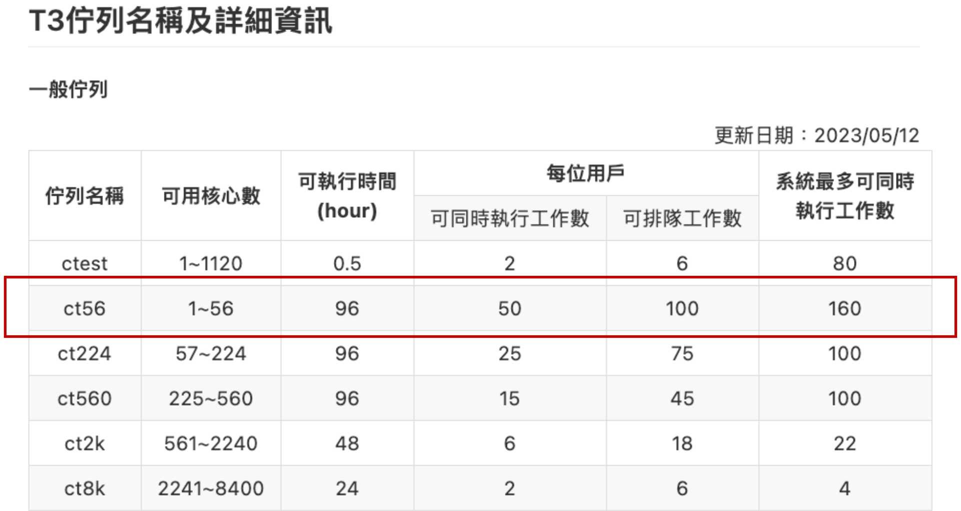
1. 登入/登出主機
2. ThinLinc Login
3. 檔案資料傳輸

- Queue 與計算資源

- Queue 列表與說明 NEW

- Job 建立

- 建立 Job Script



T3佇列名稱及詳細資訊

一般佇列

更新日期：2023/05/12

佇列名稱	可用核心數	可執行時間 (hour)	每位用戶		系統最多可同時執行工作數
			可同時執行工作數	可排隊工作數	
ctest	1~1120	0.5	2	6	80
ct56	1~56	96	50	100	160
ct224	57~224	96	25	75	100
ct560	225~560	96	15	45	100
ct2k	561~2240	48	6	18	22
ct8k	2241~8400	24	2	6	4

https://man.twcc.ai/@TWCC-III-manual/H1bEXeGcu

Enabling full-featured genomes with HiFi sequencing

Comprehensive bioinformatics solutions

July 04, 2023 | PacBio BFX

Wilson Cheng | Senior Bioinformatics Scientist, Field Applications, PacBio APAC

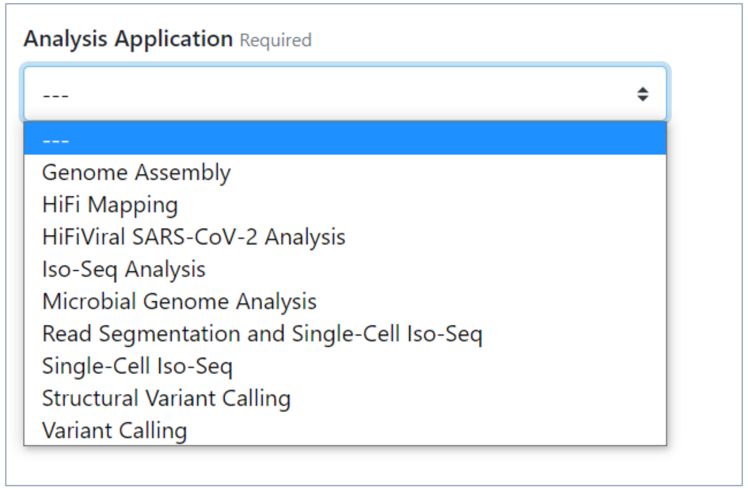
SMRT Link v12.0 GUI application enables user-friendly data management and analysis

+

- Instruments
- Sample Setup
- Runs
- Data Management
- SMRT Analysis

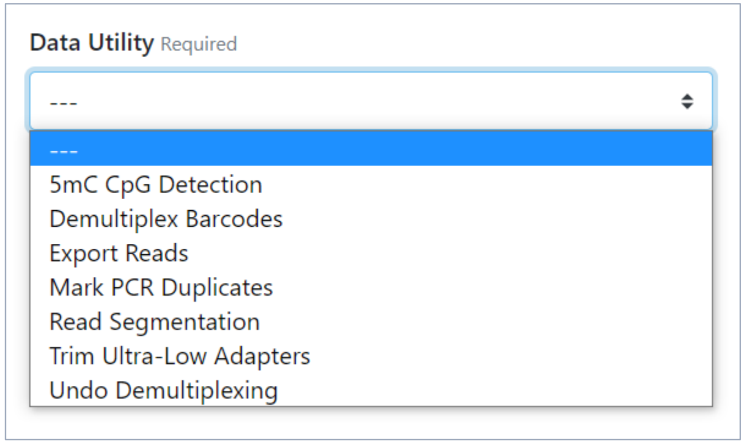
- Revio systems
- Sequel IIe systems
- Sequel II systems

SMRT Link v12.0 Analysis Applications



Analysis Application Required

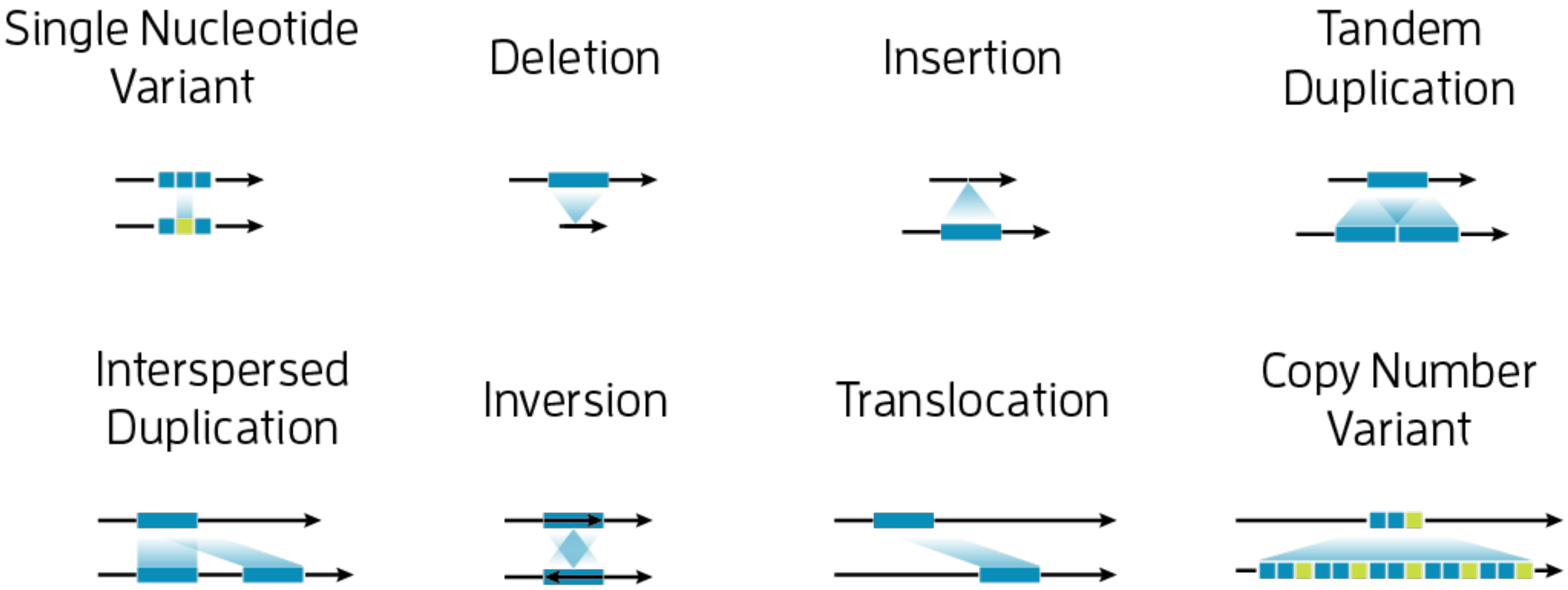
Genome Assembly
HiFi Mapping
HiFiViral SARS-CoV-2 Analysis
Iso-Seq Analysis
Microbial Genome Analysis
Read Segmentation and Single-Cell Iso-Seq
Single-Cell Iso-Seq
Structural Variant Calling
Variant Calling



Data Utility Required

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

Types of variants in a genome



Single Nucleotide Variant

Deletion

Insertion

Tandem Duplication

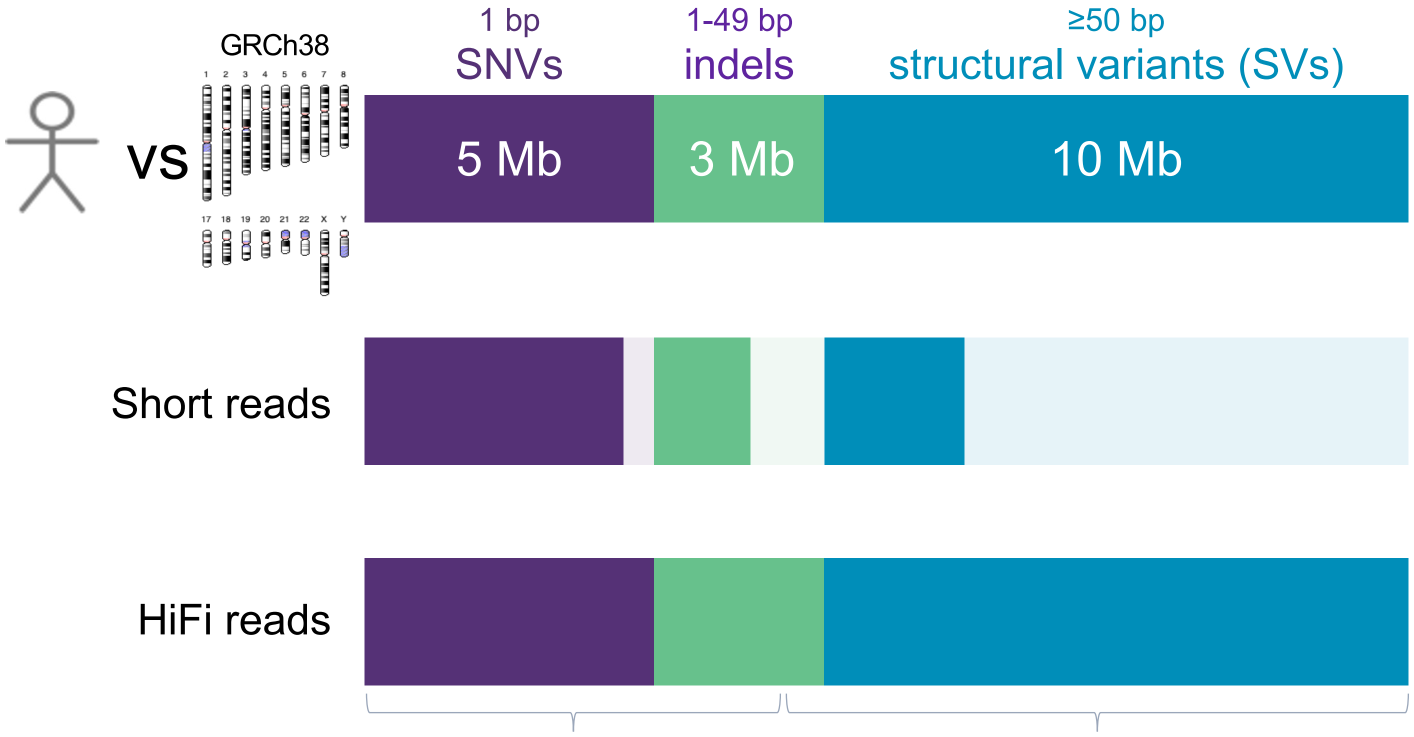
Interspersed Duplication

Inversion

Translocation

Copy Number Variant

HiFi reads provide a comprehensive view of the genome



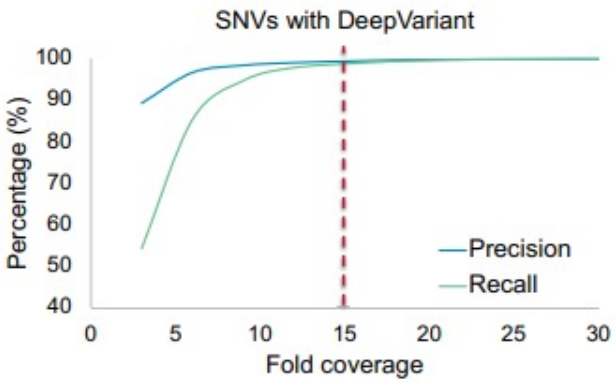
HiFi adds

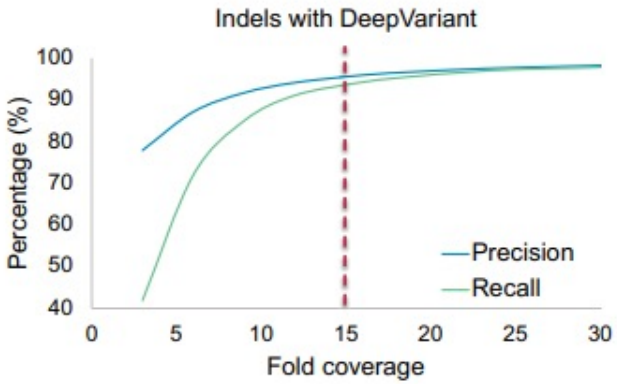
SNVs and indels in difficult regions (segdups, repeats)

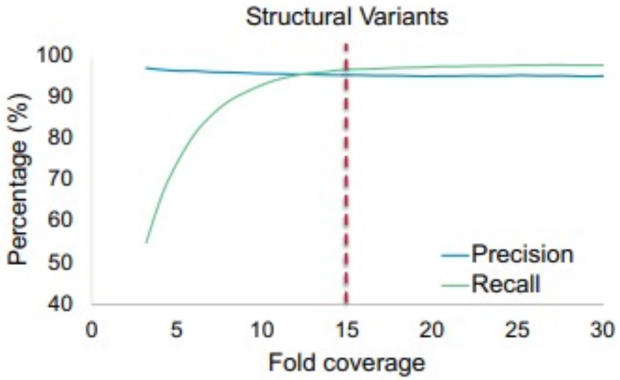
Long indels and SVs genome-wide

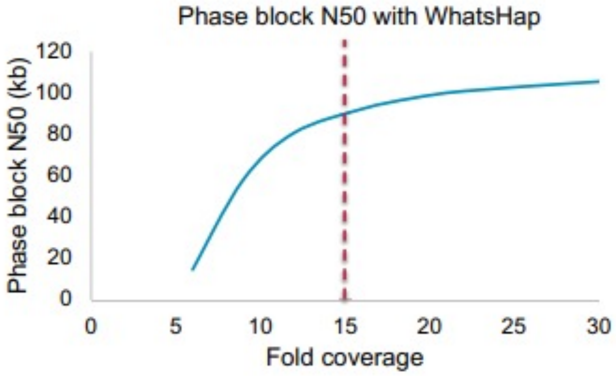
PacBio

15-fold HiFi read coverage recommendation for comprehensive variant detection applications









15-fold HiFi (≥Q20) Coverage
[2 SMRT Cells 8M for a 3 Gb genome]
provides a good trade-off between cost and results

Wenger, A. et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology. 37:1155–1162

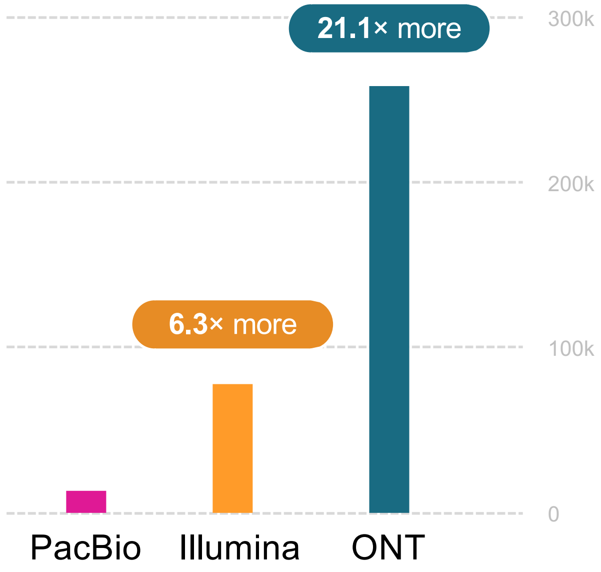
HiFi reads outperform other approaches for variant detection

precisionFDA Truth Challenge V2 & Genome in a Bottle SV Benchmark v0.6

HiFi reads improve detection of

- ✓ Structural variants
- ✓ Repeat expansions
- ✓ SNVs and indels in difficult-to-map regions

precisionFDA

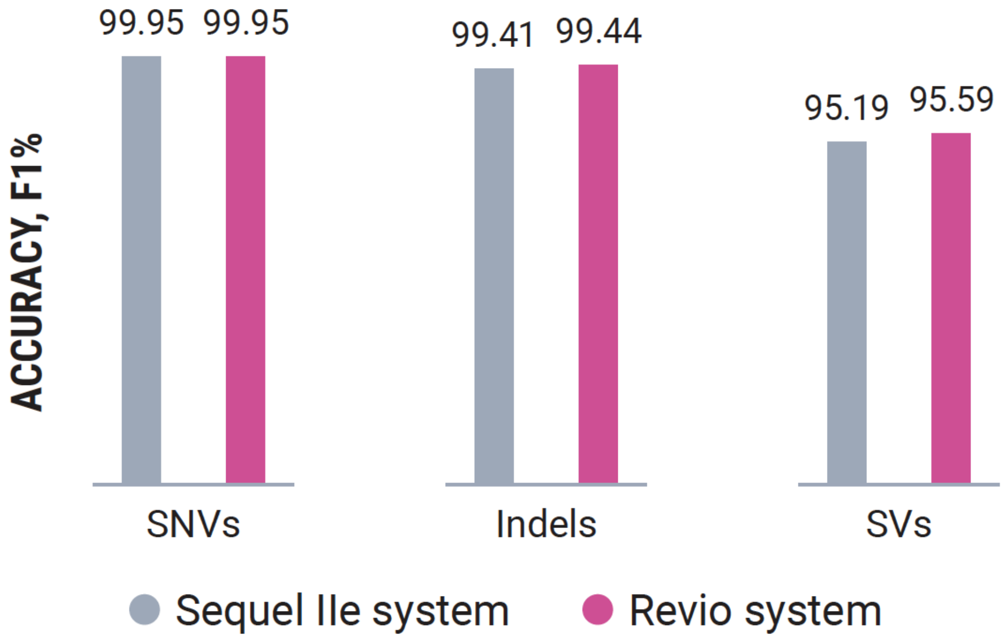


Total errors (SNV + indel + SV)

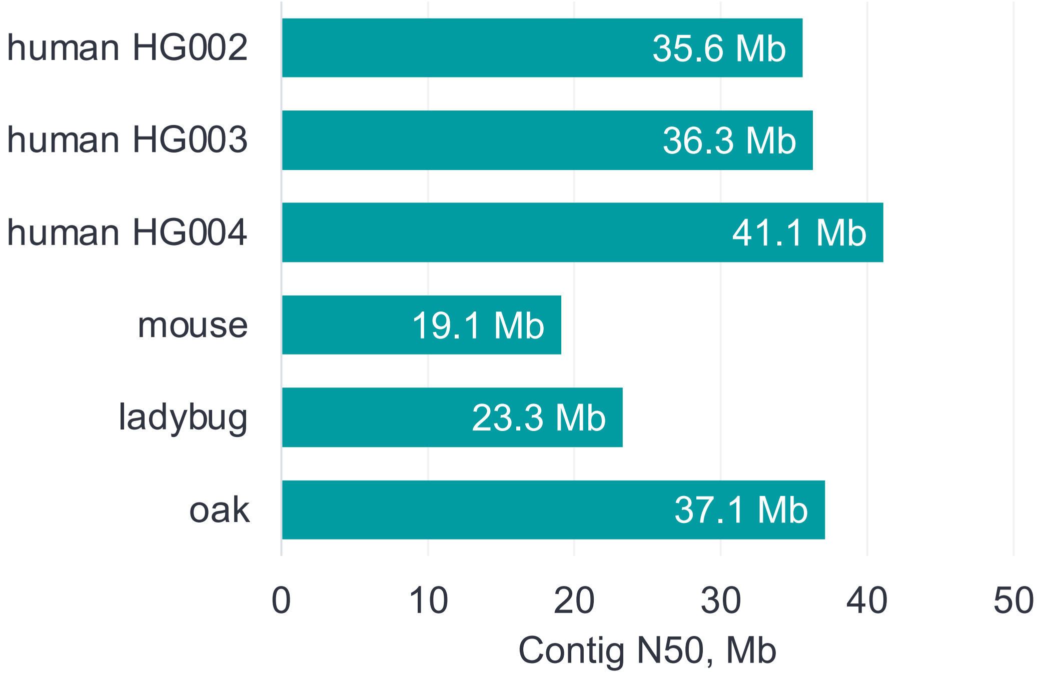
PacBio: 35× HiFi (Sequel II, Chemistry 2.0); Illumina: 35× NovaSeq ; ONT: 60× PromethION R9.4.1
Small variant calling: https://precision.fda.gov/challenges/10; Structural variant calling: HG002 against Genome in a Bottle v0.6 benchmark

Revio system has exceptional application performance

Revio system **matches *precisionFDA*-winning variant calling performance** of Sequel IIe



Revio system has **excellent genome assembly** performance

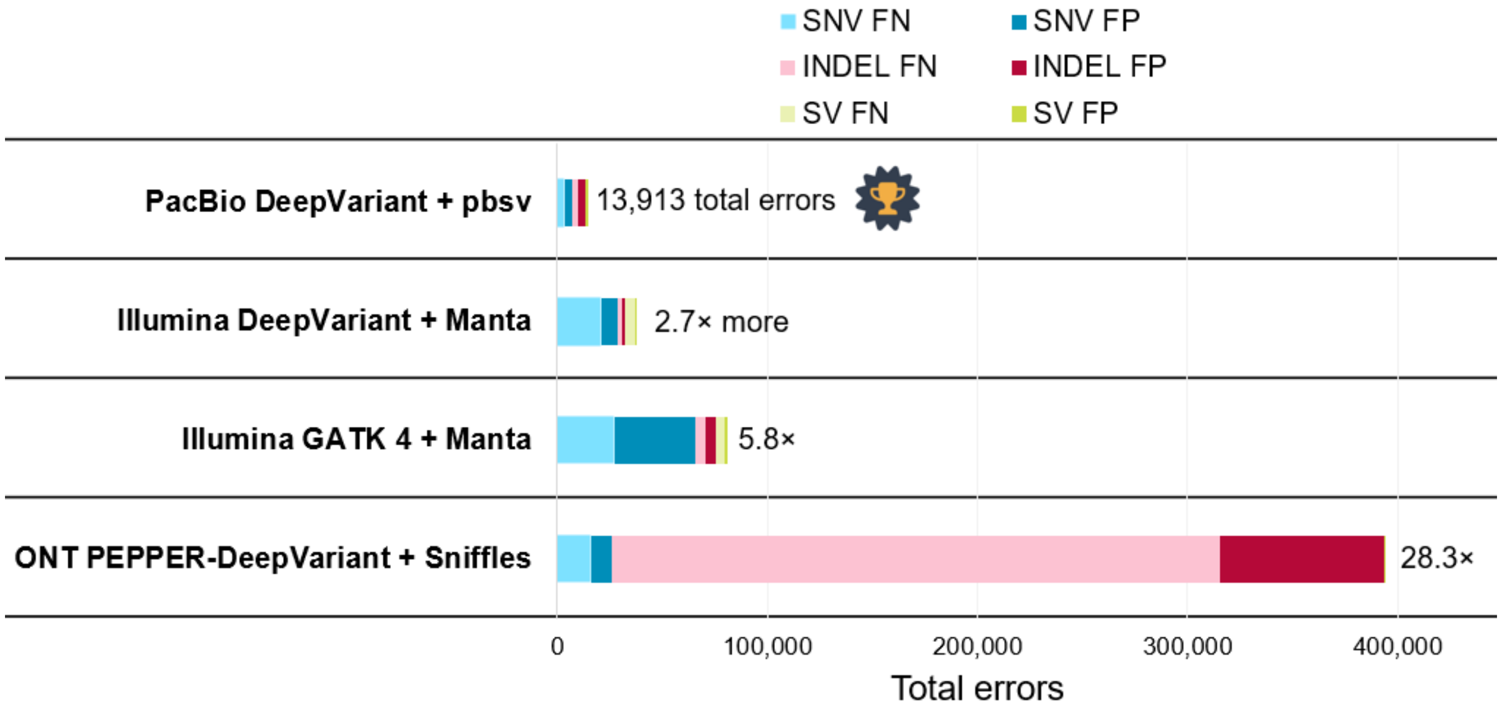


PacBio

GETTING DOWN TO THE BASIC OF SEQUENCING ACCURACY

Precision medicine also needs to be accurate medicine

Platform Comparison
PrecisionFDA Truth Challenge results (HG0003)

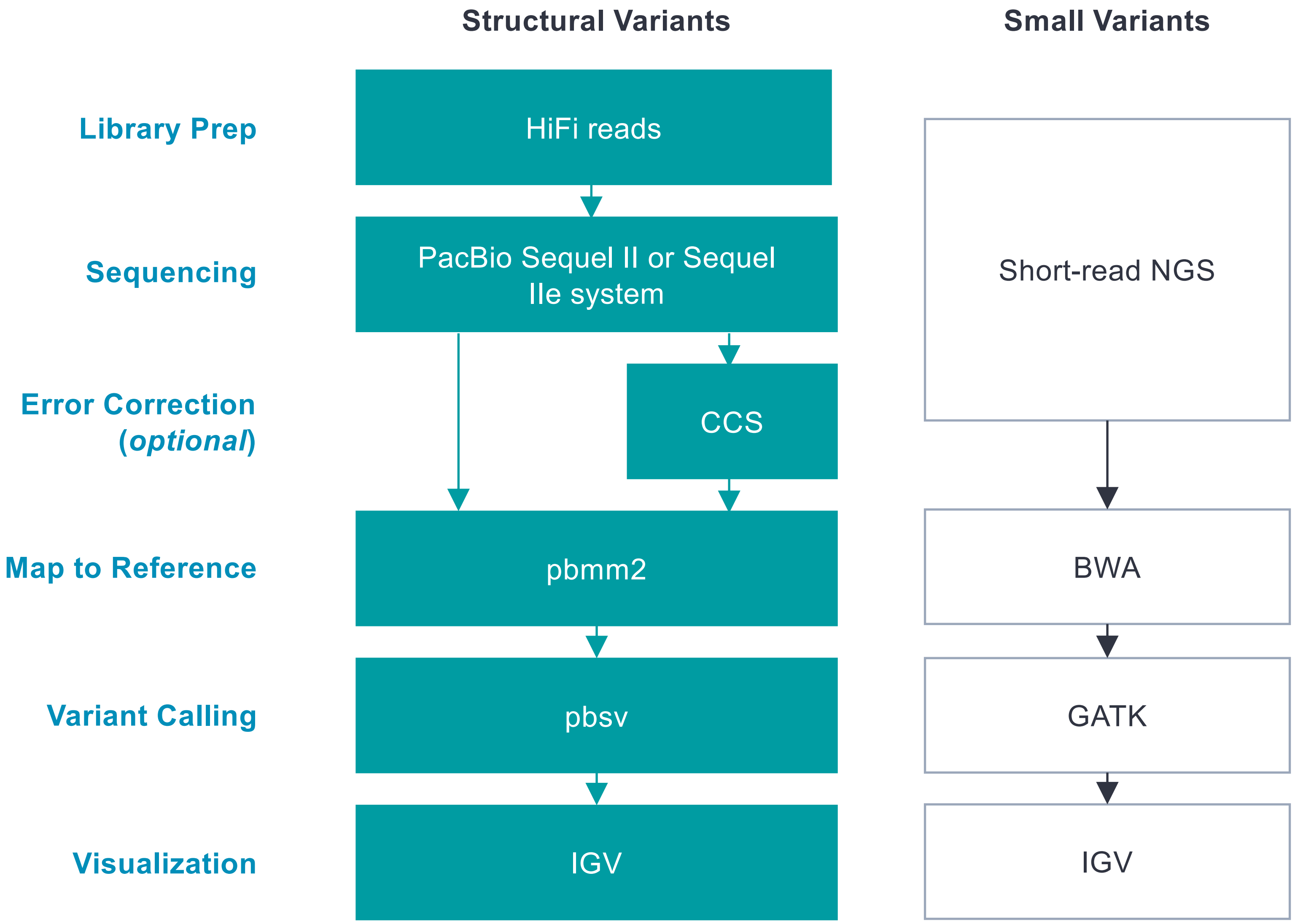


https://www.youtube.com/watch?v=dddFT7e2vP4&t=3436s

Algorithm deep dive

pbsv

Workflow to detect variants

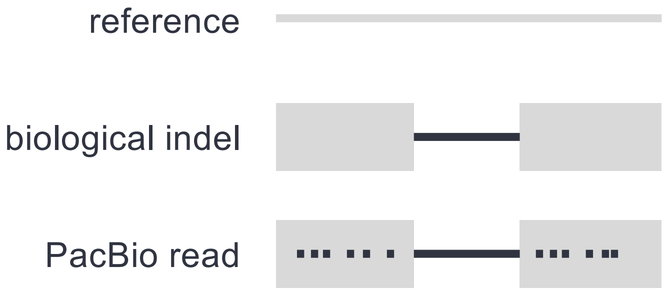


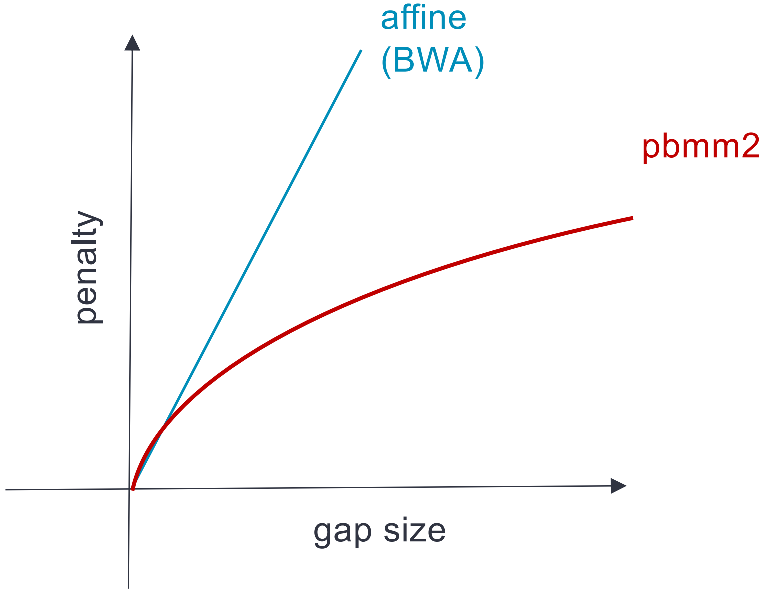
	Structural Variants	Small Variants
Library Prep	HiFi reads	Short-read NGS
Sequencing	PacBio Sequel II or Sequel IIe system	
Error Correction (*optional*)	CCS	
Map to Reference	pbmm2	BWA
Variant Calling	pbsv	GATK
Visualization	IGV	IGV

https://github.com/PacificBiosciences/pbsv

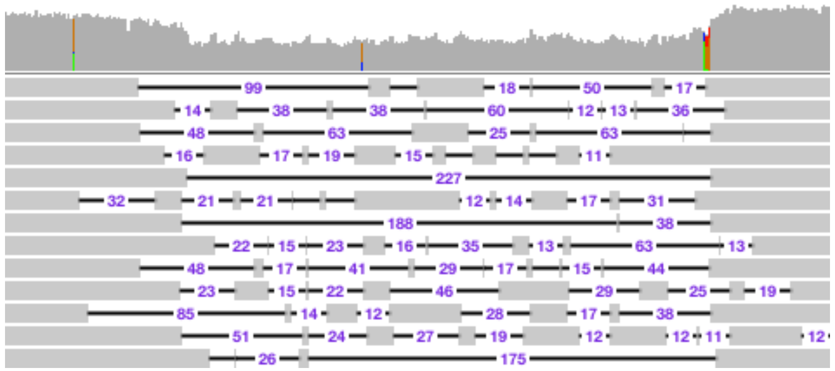
Map to reference

pbmm2

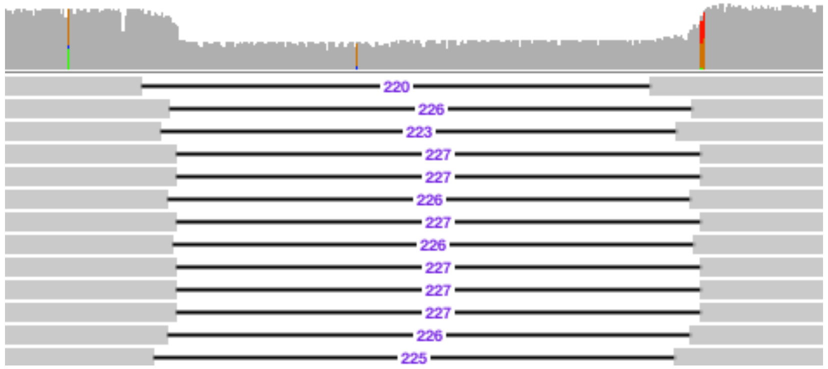




BWA



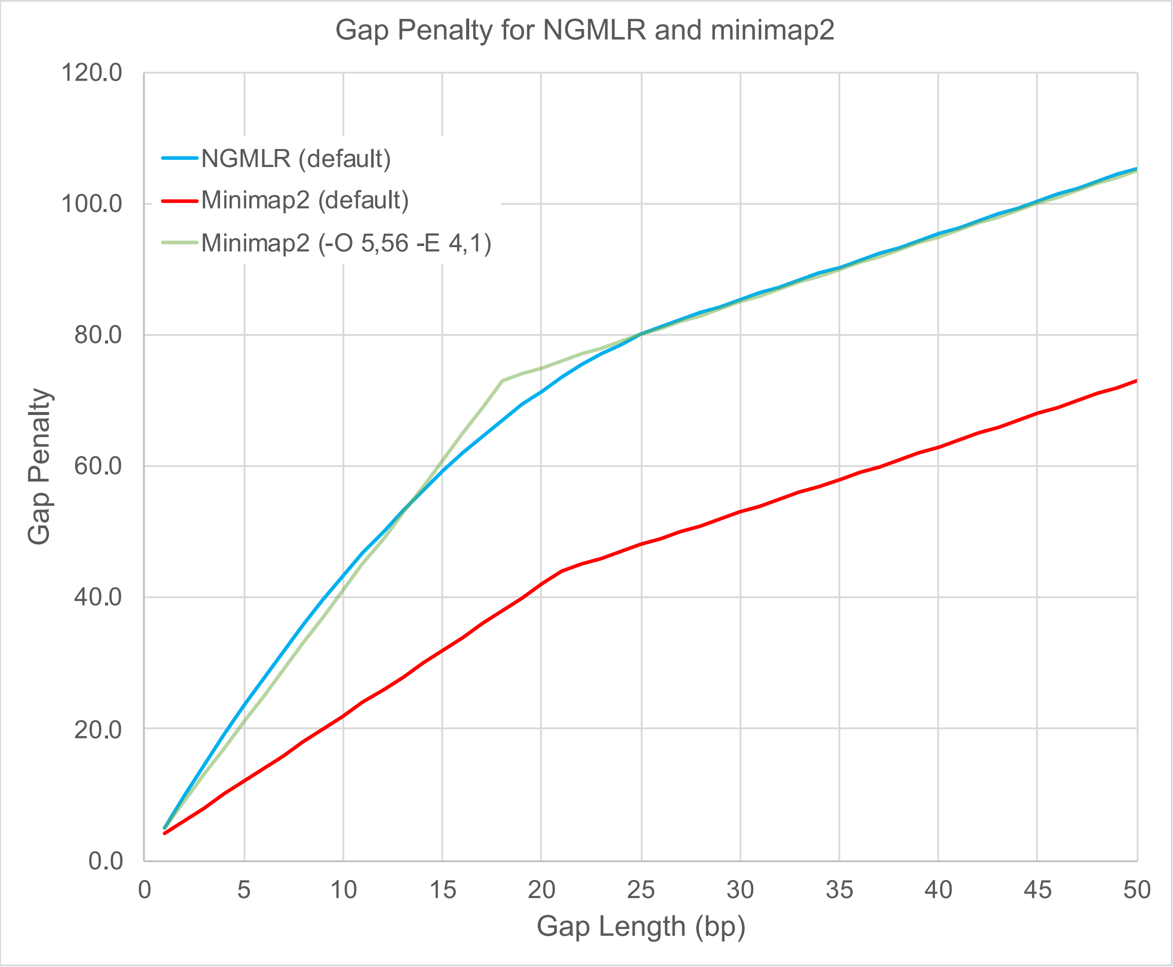
pbmm2



Sedlazeck et al. (2017) bioRxiv. doi:10.1101/169557.

Map to reference: Why pbmm2?

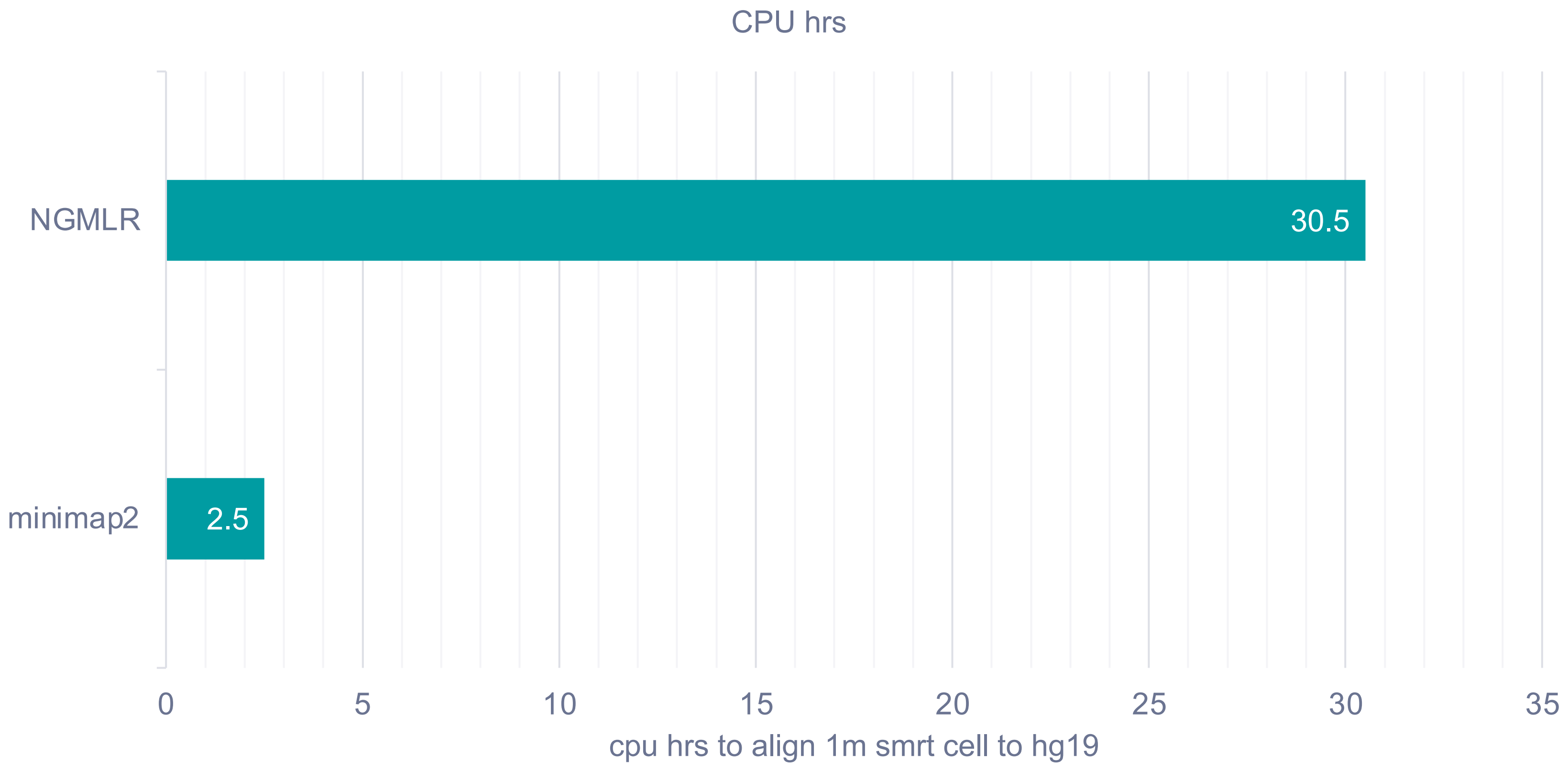
NGMLR (convex) vs pbmm2 / minimap2 (piecewise linear)



PacBio

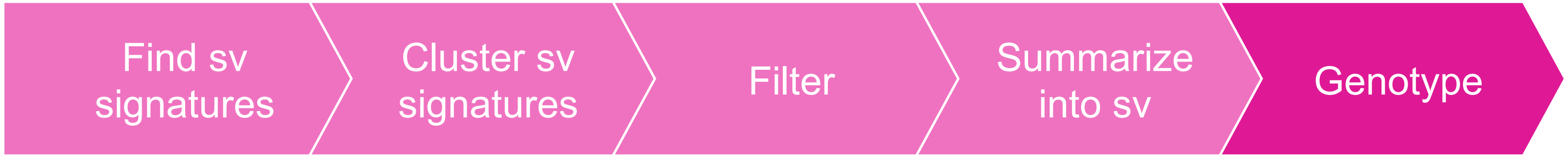
Map to reference: Why pbmm2?

Improved run time



Variant calling

Workflow



Find sv signatures

Cluster sv signatures

Filter

Summarize into sv

Genotype

SVSIG file

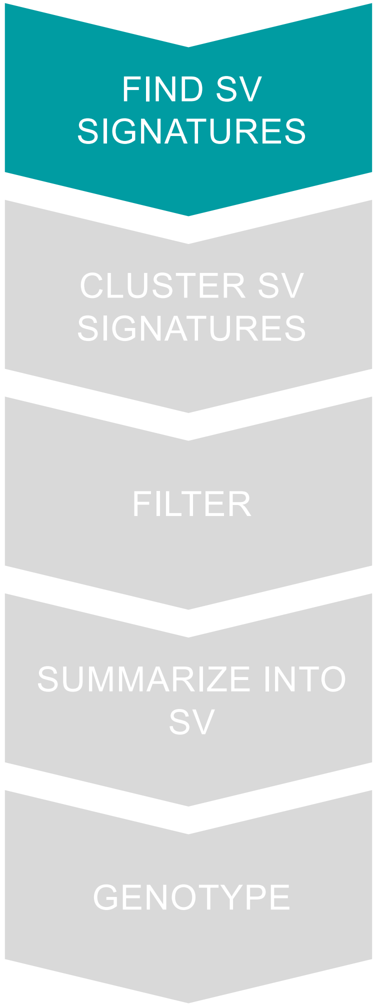
nearby with similar sequence

≥1 and ≥10% reads HiFi support

POA consensus of supporting reads + realign with AGE

supporting reads / covering reads

Variant calling



FIND SV SIGNATURES

SVSIG file

CLUSTER SV SIGNATURES

nearby with similar sequence

FILTER

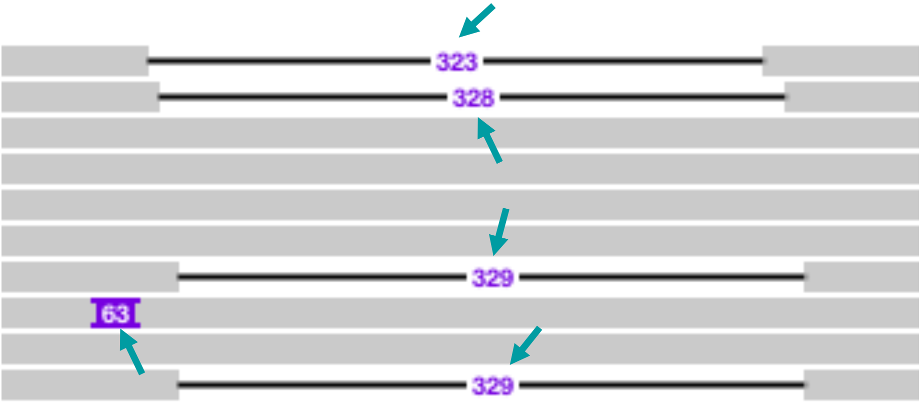
≥2 and ≥20% reads support

SUMMARIZE INTO SV

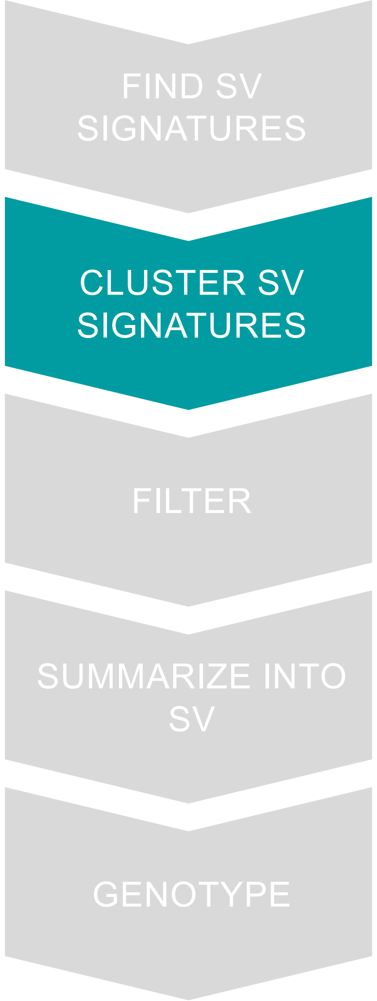
POA consensus of supporting reads + realign with AGE

GENOTYPE

supporting reads / covering reads



VARIANT CALLING



FIND SV SIGNATURES

SVSIG file

CLUSTER SV SIGNATURES

nearby with similar sequence

FILTER

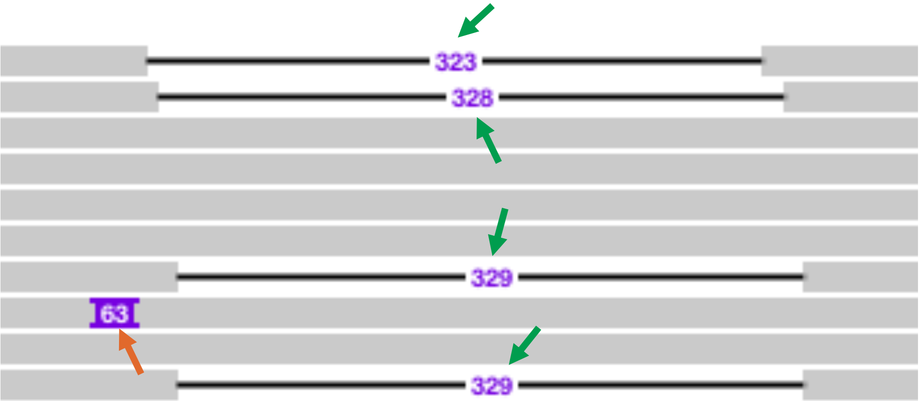
≥2 and ≥20% reads support

SUMMARIZE INTO SV

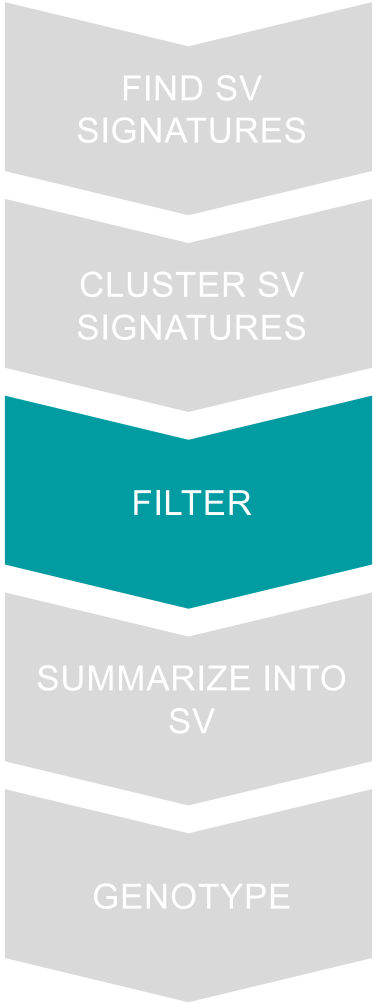
POA consensus of supporting reads + realign with AGE

GENOTYPE

supporting reads / covering reads



VARIANT CALLING



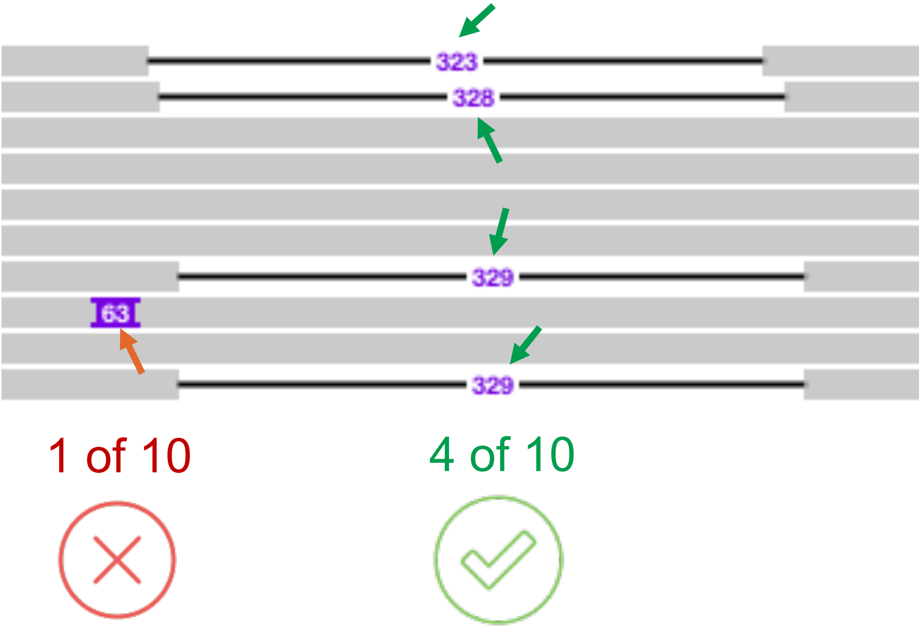
SVSIG file

nearby with similar sequence

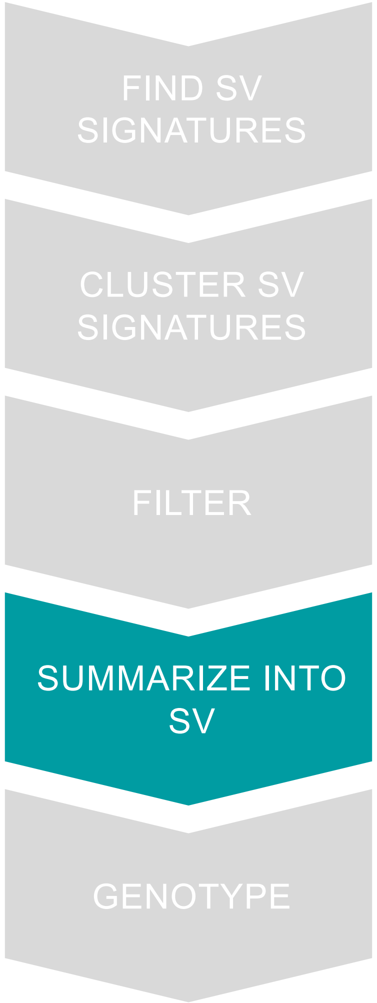
≥2 and ≥20% reads support

POA consensus of supporting reads + realign with AGE

supporting reads / covering reads



VARIANT CALLING



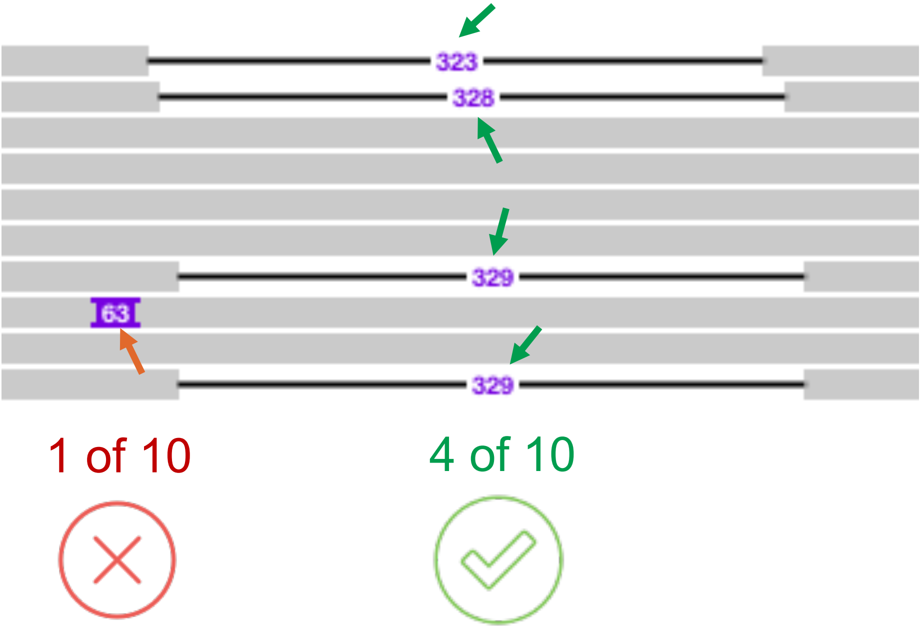
SVSIG file

nearby with similar sequence

≥2 and ≥20% reads support

POA consensus of supporting reads + realign with AGE

supporting reads / covering reads



329 bp deletion

VARIANT CALLING

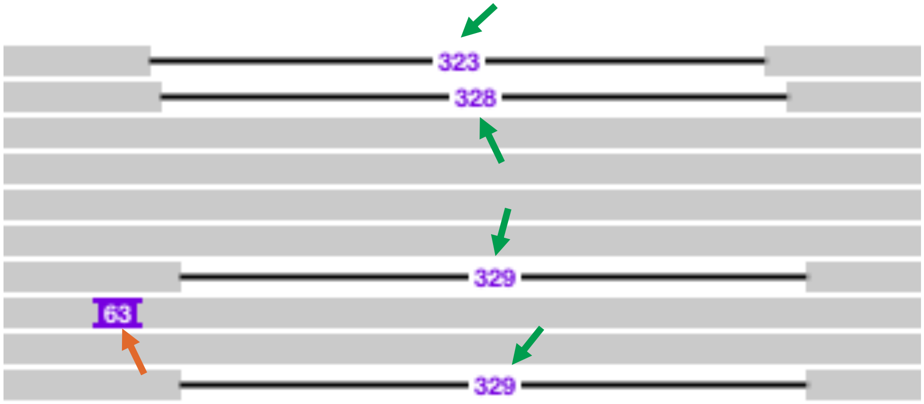
FIND SV SIGNATURES — SVSIG file

CLUSTER SV SIGNATURES — nearby with similar sequence

FILTER — ≥2 and ≥20% reads support

SUMMARIZE INTO SV — POA consensus of supporting reads + realign with AGE

GENOTYPE — supporting reads / covering reads



1 of 10 ✗

4 of 10 ✓

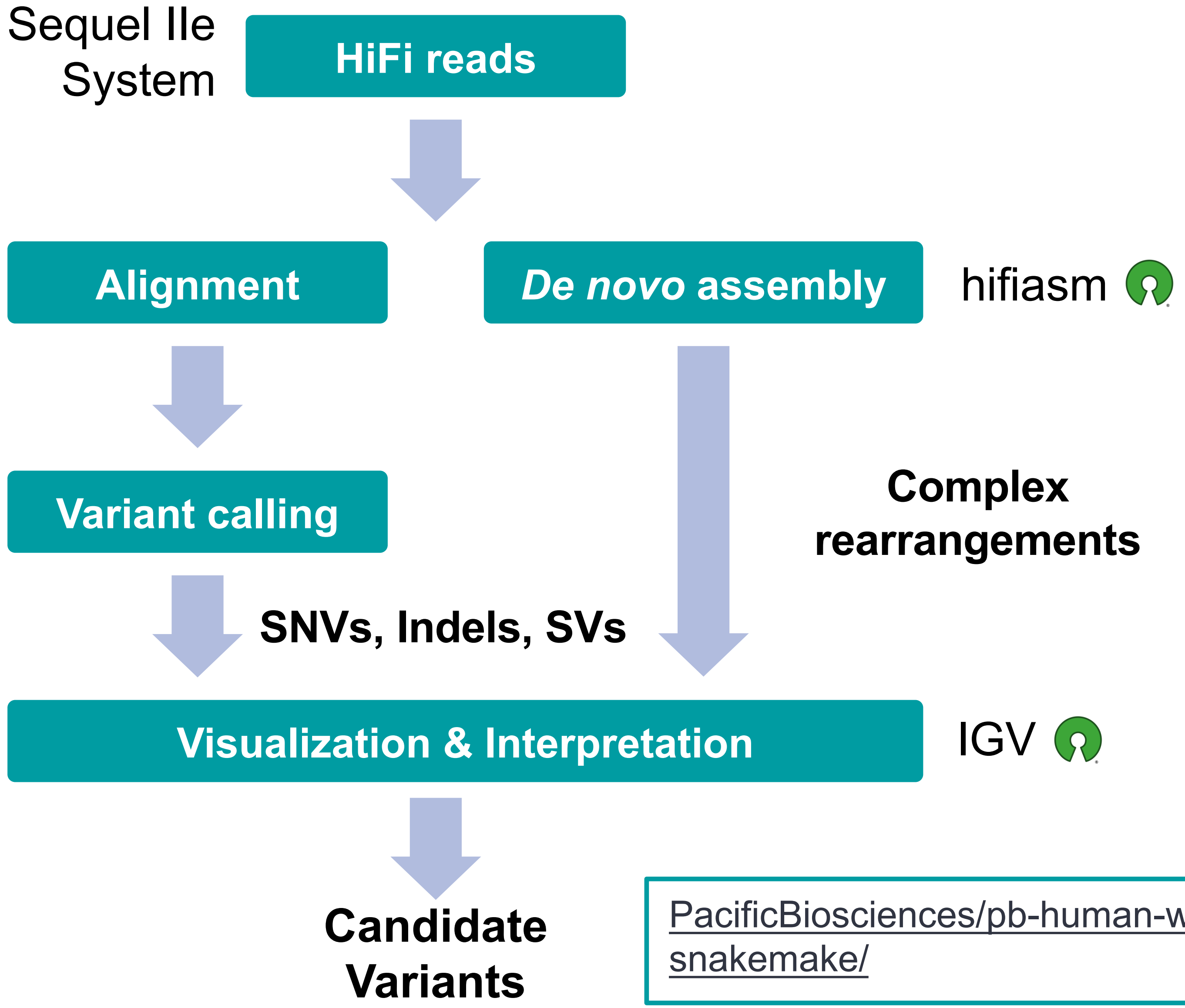
329 bp deletion

heterozygous (4 of 10)

New workflows

Tools in pbbioconda

Workflow for WGS data analysis



Sequel IIe System

HiFi reads

PACBIO pbmm2

Alignment

De novo assembly

hifiasm

Google DeepVariant (SNV, indel)

PACBIO pbsv (SV)

Variant calling

Complex rearrangements

SNVs, Indels, SVs

Variant Effect Predictor

PACBIO svpack

Visualization & Interpretation

IGV

Candidate Variants

PacificBiosciences/pb-human-wgs-workflow-snakemake/

PacBio

PB human WGS workflow snakemake

Process smrtcells

Aligns HiFi reads reference genome also for QC to confirm.

pbmm2

Align HiFi reads to reference genome (GRCh38)

mosdepth

- Calculate aligned coverage depth
- Generate read length and QC
- Calculate depth ratio (chrX:chrY)

jellyfish

Count kmers in HiFi reads to dump and Export modimers for sample swap detection.

Process sample

Variant discovery, variant calling, and assembly for each sample.

pbsv

Call structural variants

DeepVariant

Call small variants

Whatshap

Phased small variants and generate merged, haplotagged BAM

Hifiasm

Assemble reads

Trgt

Genotype tandem repeat

pb-cpg-tools

Gerneate list of CpG/5mC sites and modification probabilities

Process cohort

Variants are prioritized, annotated, and filtered find candidate rare variants with functional consequence.

pbsv

Joint call structural variants

GLnexus

Joint call small variants

slivar

Annotate and filter small variant with population AF from gnomAD and HRTC

svpack

Annotate and filter structural variant

calN50

Calculate assembly status

PacBio

https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake/blob/main/Tutorial.md

The Consortium for long-read sequencing variant frequency database

Eichler Lab
Department of Genome Sciences,
University of Washington

UW Genome Sciences

HudsonAlpha
Institute for Biotechnology

Johns Hopkins
University

Amsterdam UMC
Universitair Medische Centra

Amsterdam UMC Genome Diagnostics

In·Dei·Nomine·Feliciter

Radboud Universiteit

Children's Mercy
Kansas City

The Jackson Laboratory

UC Santa Cruz Genomics Institute

School of Science
The University of Tokyo

Yale School of Medicine

Yale Center for Genomic Health

Mission: Establish database of long read specific variants to fully realize the utility of long-read sequencing in human health applications.

Membership: Founded by leading institutions and experts with significant interest and experience in population scale variant frequency databases

Goals:

- Establish a globally accessible long-read variant database (>2,000 genomes by end 2023) to be hosted in **NHGRI AnVIL** (Analysis Visualization and Informatics Lab-space)
- Incorporate standardized data and pipeline required to normalize heterogeneous data sets from contributors
- Write a manuscript describing analysis of variant data and database can be used to screen potentially pathogenic variants in clinical samples

The International Children Hospitals' Consortium to Increase Diagnostic Yield in Rare and Inherited Diseases (RID)

Children's Mercy KANSAS CITY

CHEO

SickKids

国立研究開発法人 国立成育医療研究センター National Center for Child Health and Development

GENOMICS CORE LEUVEN KU LEUVEN UZ LEUVEN

BROAD INSTITUTE

Children's Hospital of Philadelphia

KK Women's and Children's Hospital SingHealth

Faculty of Medicine CHULALONGKORN UNIVERSITY

Radboud Universiteit

UNIVERSITÄTS KLINIKUM TÜBINGEN

Mission: Generate evidence and establish clinically-informed best practices on the utility of HiFi sequencing technology to potentially increase diagnostic yield of unsolved rare and inherited diseases.

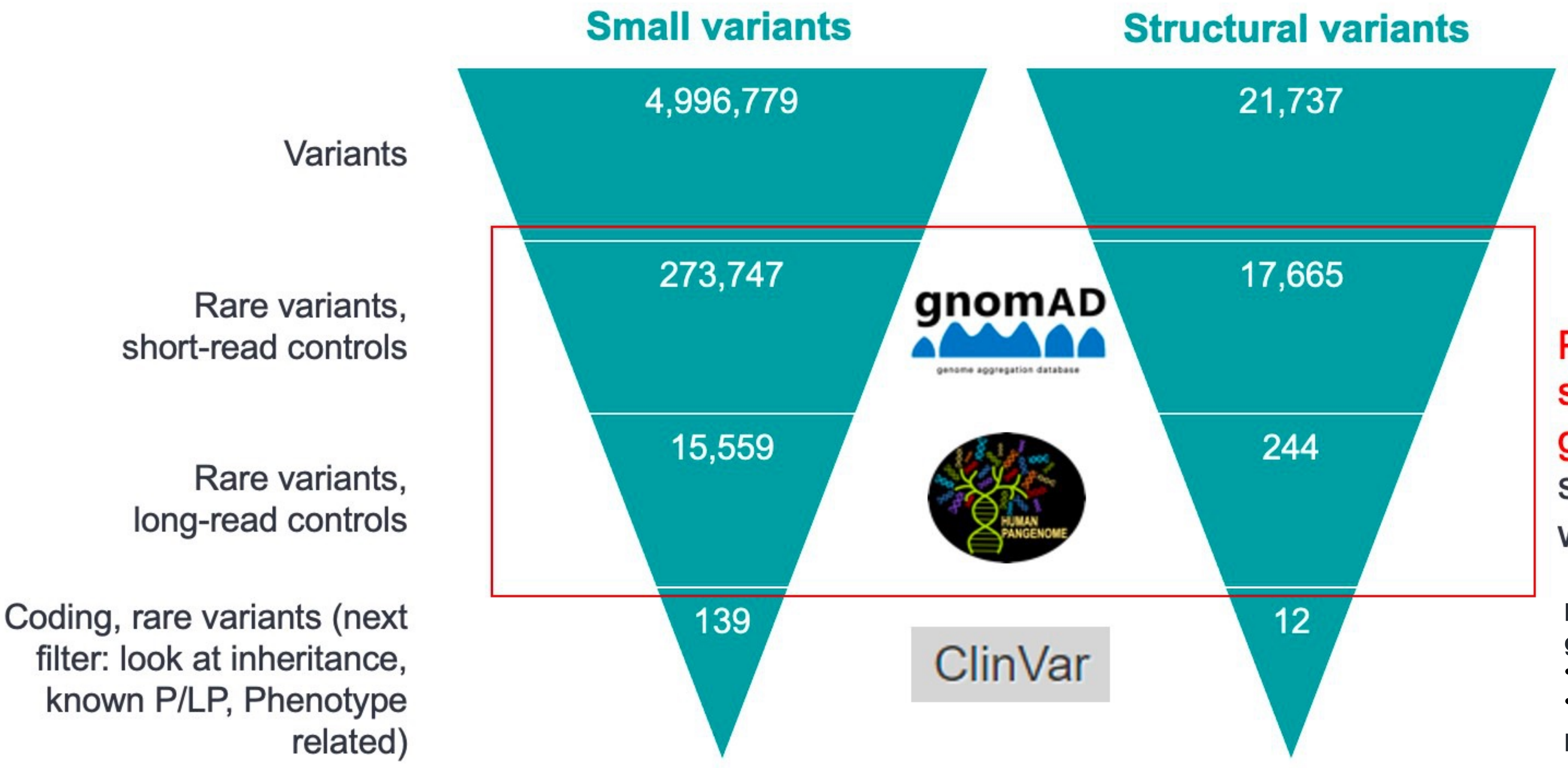
Membership: Founded by leading institutions and children's hospitals with significant interest and experience in [evaluation / diagnosis] of rare and inherited diseases.

Goals:

- Establish a globally accessible variant database of HiFi WGS genomes in RID (at least 2,500 genomes by end 2023)
- Publish a series of clinically-informed best practices in adoption of HiFi WGS in RID
- Provide access to testing facilities offering HiFi long-read WGS

Population Frequency Filtering Is Necessary for NGS Genetic Disease analysis/interpretation

Frequency database (gnomAD) and database like Clinvar, HGMD etc are the real power behind 3rd analysis. Without this data, interpretation would not fully extract benefit of increased SV detection


Small variants
Structural variants
Variants
4,996,779
21,737
Rare variants, short-read controls
273,747
17,665
gnomAD
genome aggregation database
Rare variants, long-read controls
15,559
244
HUMAN PANGENOME
Coding, rare variants (next filter: look at inheritance, known P/LP, Phenotype related)
139
12
ClinVar

PacBio Current State: Using summary data from 40 long read genomes for freq. filtering – Building something with more power is what we propose

Filter for rare SNVs that impact a gene

- max_gnmad_af: 0.01
- max_hprc_af:0.01

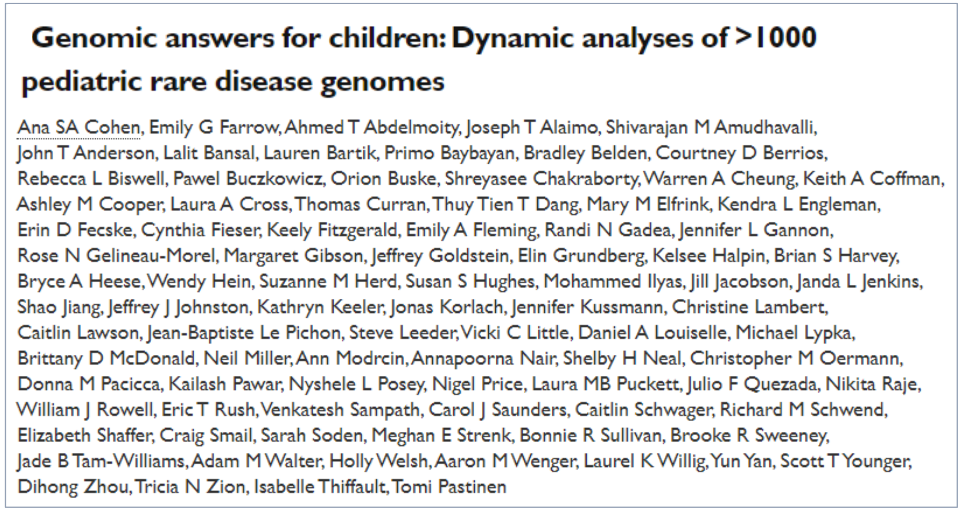
Filter for rare SVs that impact a gene

- confident SV calls (PASS calls)
- SV calls not seen in population controls (rare variants)
- SV calls that impact a coding gene

The Genome Aggregation Datatbase (gnomAD)
https://gnomad.broadinstitute.org/

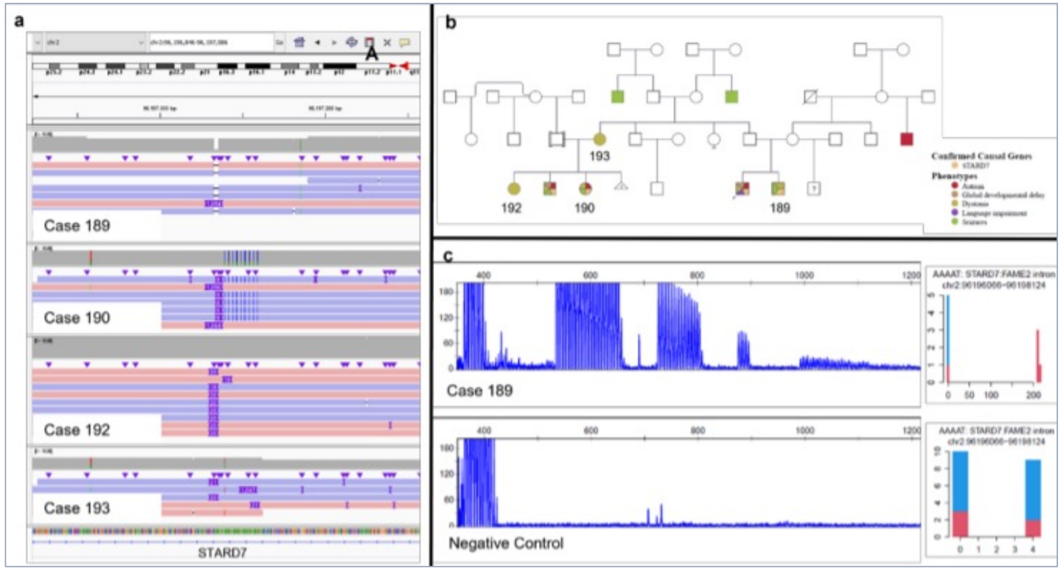
Human Pangenome Reference Consortium (HPRC)
https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0

Increased explanation rate through PacBio HiFi WGS

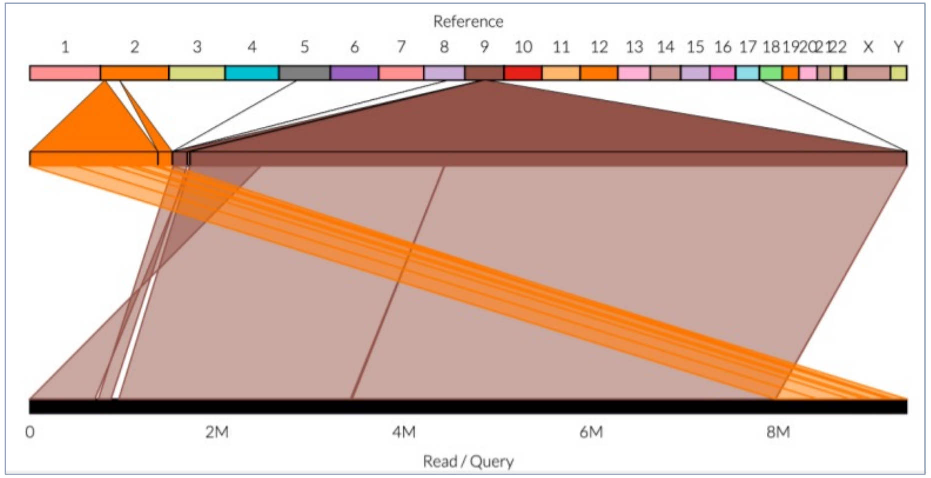


Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes

Ana SA Cohen, Emily G Farrow, Ahmed T Abdelmoity, Joseph T Alaimo, Shivarajan M Amudhavalli, John T Anderson, Lalit Bansal, Lauren Bartik, Primo Baybayan, Bradley Belden, Courtney D Berrios, Rebecca L Biswell, Pawel Buczkowicz, Orion Buske, Shreyasee Chakraborty, Warren A Cheung, Keith A Coffman, Ashley M Cooper, Laura A Cross, Thomas Curran, Thuy Tien T Dang, Mary M Elfrink, Kendra L Engleman, Erin D Fecske, Cynthia Fieser, Keely Fitzgerald, Emily A Fleming, Randi N Gadea, Jennifer L Gannon, Rose N Gelineau-Morel, Margaret Gibson, Jeffrey Goldstein, Elin Grundberg, Kelsee Halpin, Brian S Harvey, Bryce A Heese, Wendy Hein, Suzanne M Herd, Susan S Hughes, Mohammed Ilyas, Jill Jacobson, Janda L Jenkins, Shao Jiang, Jeffrey J Johnston, Kathryn Keeler, Jonas Korlach, Jennifer Kussmann, Christine Lambert, Caitlin Lawson, Jean-Baptiste Le Pichon, Steve Leeder, Vicki C Little, Daniel A Louiselle, Michael Lypka, Brittany D McDonald, Neil Miller, Ann Modrcin, Annapoorna Nair, Shelby H Neal, Christopher M Oermann, Donna M Pacicca, Kailash Pawar, Nyshele L Posey, Nigel Price, Laura MB Puckett, Julio F Quezada, Nikita Raje, William J Rowell, Eric T Rush, Venkatesh Sampath, Carol J Saunders, Caitlin Schwager, Richard M Schwend, Elizabeth Shaffer, Craig Smail, Sarah Soden, Meghan E Strenk, Bonnie R Sullivan, Brooke R Sweeney, Jade B Tam-Williams, Adam M Walter, Holly Welsh, Aaron M Wenger, Laurel K Willig, Yun Yan, Scott T Younger, Dihong Zhou, Tricia N Zion, Isabelle Thiffault, Tomi Pastinen



- **13% of new explanations in previously unsolved cases by incorporating SVs**
- HiFi WGS over srWGS:
  - 2x more SVs
  - 4x more rare transmitted SVs
  - 5% (~200,000 additional) SNVs
  - Long-range phasing (~400 kb)

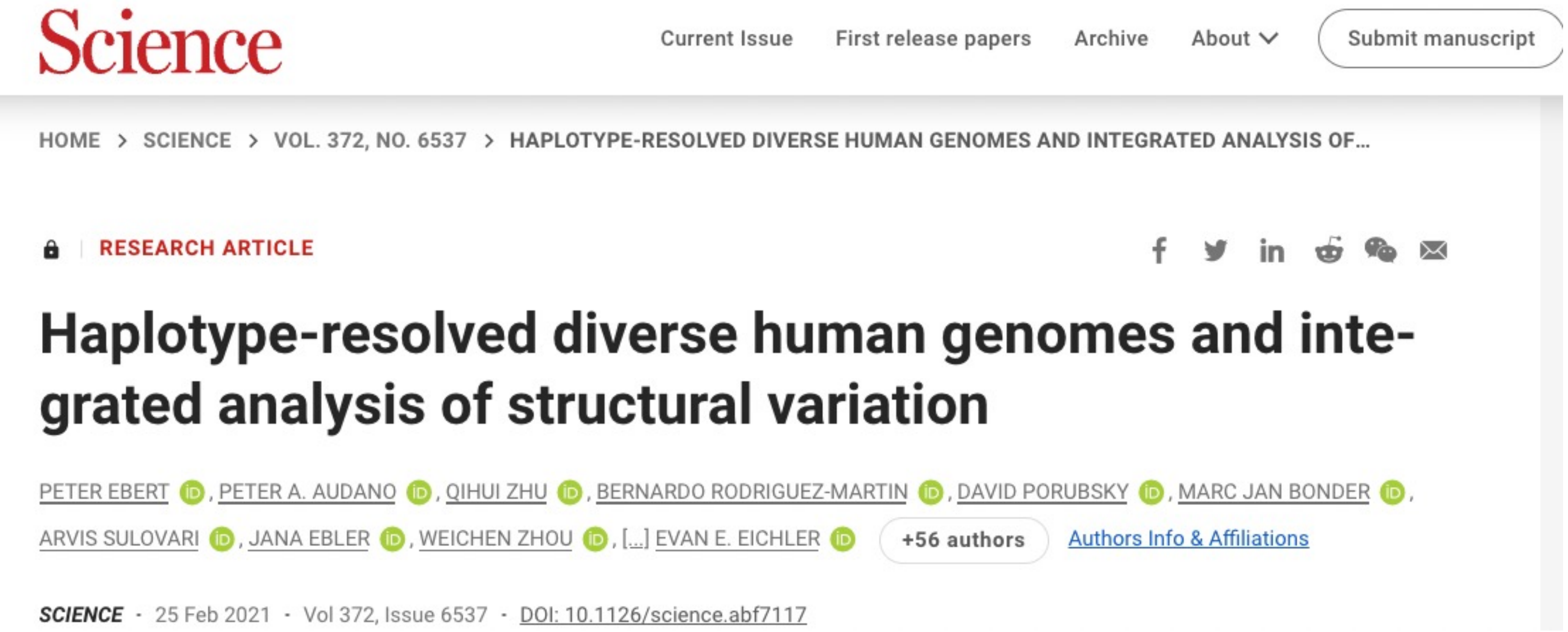


https://www.sciencedirect.com/science/article/pii/S1098360022006530; 1https://github.com/ChildrensMercyResearchInstitute/GA4K

Phased Assembly Variant Caller (pav)

PAV is a tool for discovering variation using assembled genomes aligned to a reference.

- It is designed explicitly for phased assemblies, however, it can be used for squashed assemblies by providing an empty FASTA for the second haplotype.
- PAV was developed for the Human Genome Structural Variation Consortium (HGSVC)



HOME > SCIENCE > VOL. 372, NO. 6537 > HAPLOTYPE-RESOLVED DIVERSE HUMAN GENOMES AND INTEGRATED ANALYSIS OF…

RESEARCH ARTICLE

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

PETER EBERT, PETER A. AUDANO, QIHUI ZHU, BERNARDO RODRIGUEZ-MARTIN, DAVID PORUBSKY, MARC JAN BONDER, ARVIS SULOVARI, JANA EBLER, WEICHEN ZHOU, [...] EVAN E. EICHLER +56 authors Authors Info & Affiliations

SCIENCE · 25 Feb 2021 · Vol 372, Issue 6537 · DOI: 10.1126/science.abf7117

https://github.com/EichlerLab/pav
Ebert et al., "Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation," Science, February 25, 2021, eabf7117, https://doi.org/10.1126/science.abf7117.

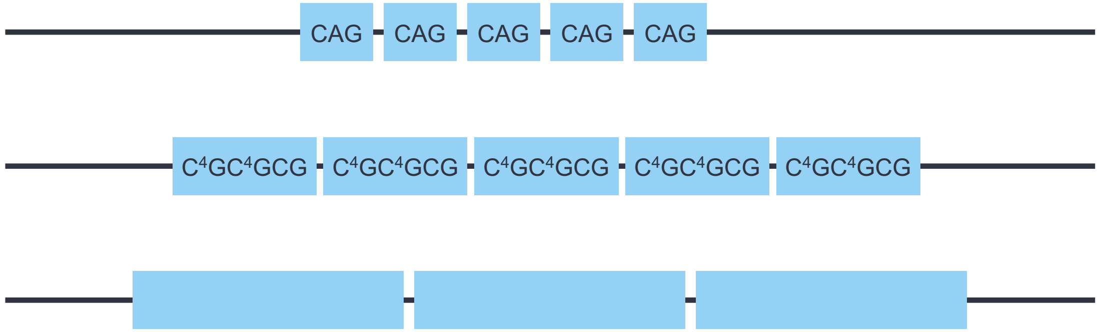
Computational tools drive discovery

SNPs / small indels	✓ DeepVariant	
SVs	✓ pbsv	>22,000 SVs
Tandem Repeats	✓ TRGT/TRVZ	Sequence, repeat length, methylation
Segmental duplications	✓ Paraphase	Pseudogenes, paralogs
Phasing	✓ HiPhase	N50=493kb
CNVs	✓ HiFiCNV	Advanced CNV calling
Methylation	✓ 5mC included	>28M CpG sites

Compute power

Google DeepConsensus and more on board

Tandem repeats play a key role in human health and disease



Microsatellite
(unit size 1–9 bp)
Short tandem repeats (STRs)

Minisatellite
(unit size 10–100 bp)
Variable length tandem repeats (VNTRs)

Macrosatellite
(unit size >100 bp)
Macrosatellite

Most abundant class of variation in the human genome[1]

- > 1M tandem repeats in the human genome
- > 10 higher mutation rate than any other variant class

Known to cause disease

- >50 repeat expansion disorders caused by STRs[1]
- Several VNTRs linked to diseases like Alzheimer's, Autism, Epilepsy, ALS and others[2,3]

Accurate characterization is essential to diagnose disease[1]

- Accurate repeat count
- Identification of medically relevant interruption sequences
- Methylation status

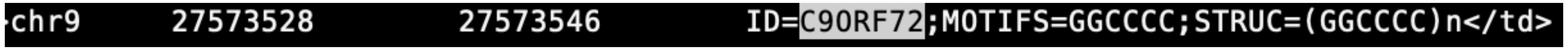
Resolving the unsolved: Comprehensive assessment of tandem repeats at scale

Egor Dolzhenko, Adam English, Harriet Dashnow, Guilherme De Sena Brandine, Tom Mokveld, William J. Rowell, Caitlin Karniski, Zev Kronenberg, Matt C. Danzi, Warren Cheung, Chengpeng Bi, Emily Farrow, Aaron Wenger, Verónica Martínez-Cerdeño, Trevor D Bartley, Peng Jin, David Nelson, Stephan Zuchner, Tomi Pastinen, Aaron R. Quinlan, Fritz J. Sedlazeck, Michael A Eberle

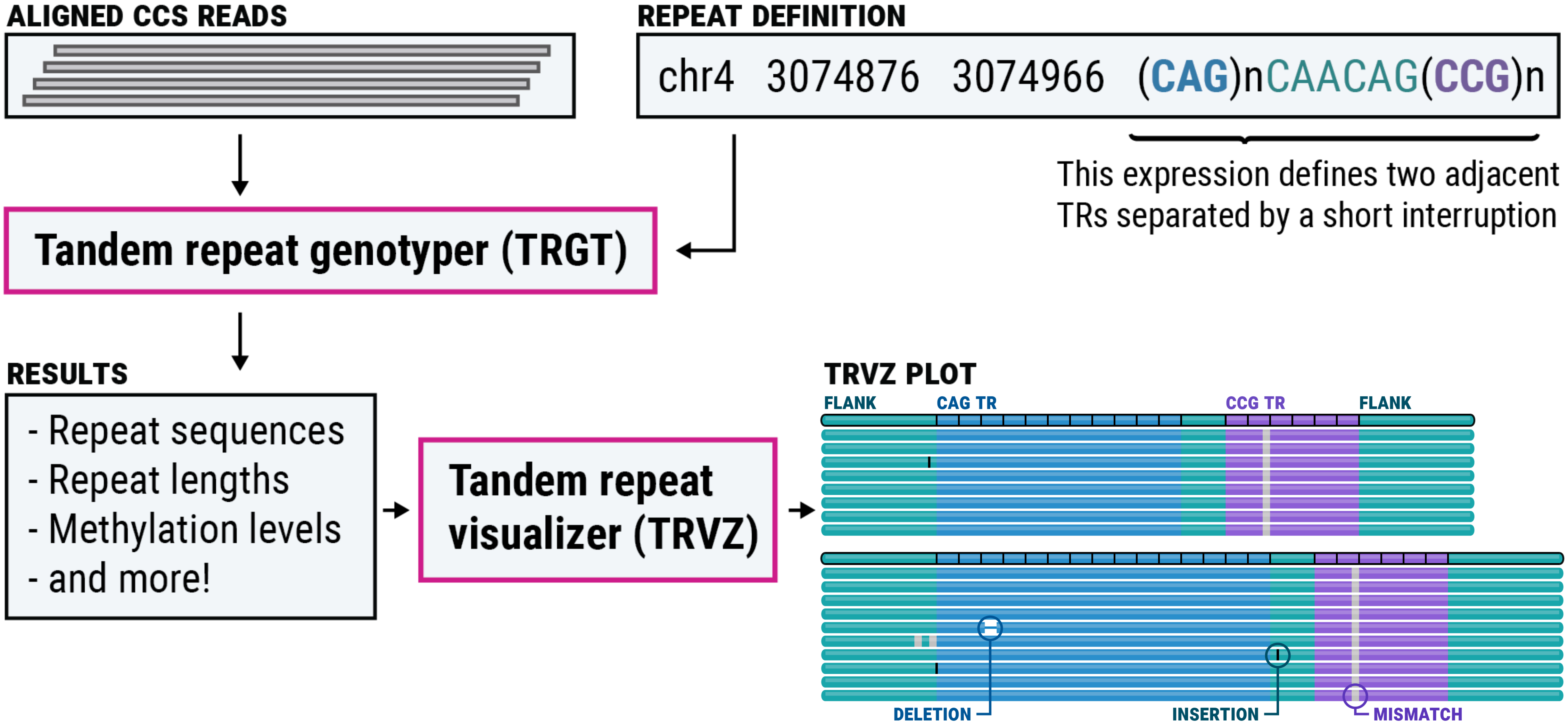
doi: https://doi.org/10.1101/2023.05.12.540470

1. Depienne et. al (2021) 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? AJHG,108(5): 764-785; 2) Ebbert et. Al (2020)
2. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 20(1):97
3. Raybould (2021) Searching the dark genome for Alzheimer's disease risk variants. Brain Sci. 11(3):332

Mapping of *HTT* repeat expansion with TRGT and visualizing with TRVZ



```
chr9	27573528	27573546	ID=C9ORF72;MOTIFS=GGCCCC;STRUC=(GGCCCC)n</td>
```



ALIGNED CCS READS

REPEAT DEFINITION

chr4 3074876 3074966 (CAG)nCAACAG(CCG)n

This expression defines two adjacent TRs separated by a short interruption

Tandem repeat genotyper (TRGT)

RESULTS

- Repeat sequences
- Repeat lengths
- Methylation levels
- and more!

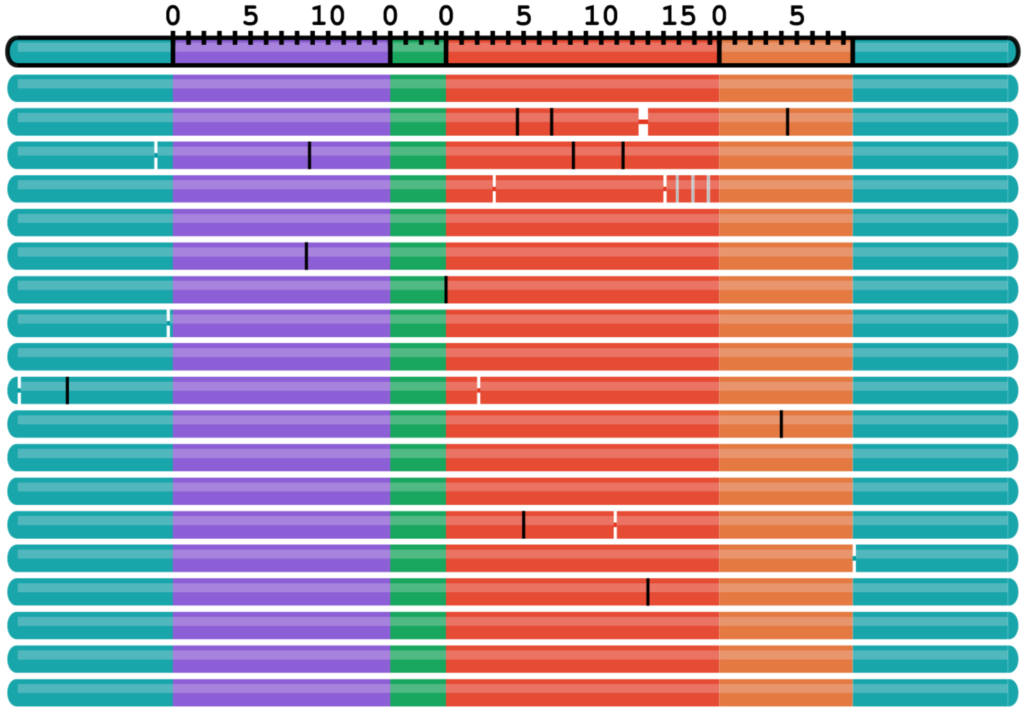
Tandem repeat visualizer (TRVZ)

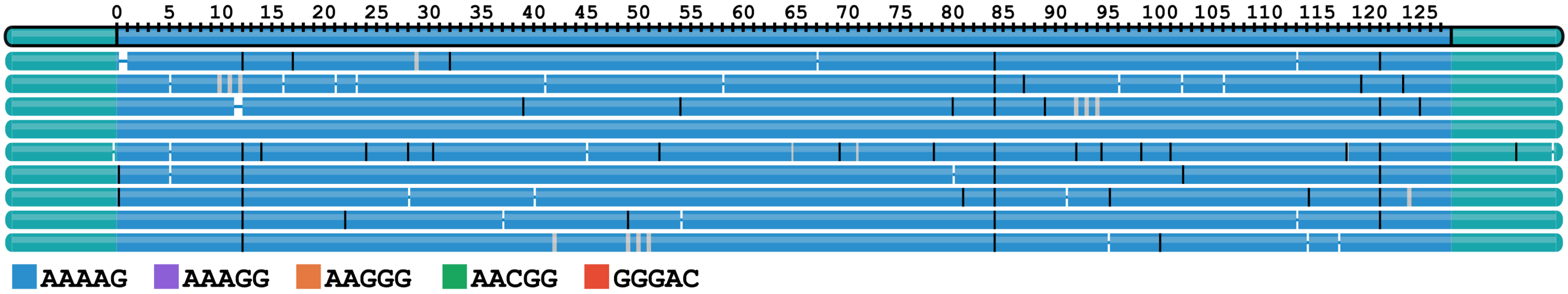
TRVZ PLOT

Egor et. al. (2023) Resolving the unsolved: Comprehensive assessment of tandem repeats at scale. bioRxive
doi: https://doi.org/10.1101/2023.05.12.540470

A polymorphic repeat whose expansions cause CANVAS

- *RFC1* repeat is polymorphic in length and sequence composition
- Biallelic *RFC1* expansions consisting of AAGGG motif cause cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS)
- Here are *RFC1* repeat alleles in HG00733:





PacBio

Egor et. al. (2023) Resolving the unsolved: Comprehensive assessment of tandem repeats at scale. bioRxive
doi: https://doi.org/10.1101/2023.05.12.540470

HiFiCNV – calling copy number variants from HiFi datasets

Many variant types of clinical research interest (*i.e.*, human genomics) are covered by existing tools for HiFi data.

- Small variants – DeepVariant; SNV and indel
- Structural Variants (SVs) – pbsv; deletion, insertion, and inversion
- Short Tandem Repeats (STRs) – TRGT
- Targeted tools – Paraphase, Pangu

HiFiCNV aims to call copy number variants (CNVs)

- Large scale copy number gains and losses (typically >100 kb)
- Frequently caused by segmental duplications and/or sequence homology
- Main goal: create a tool that can leverage read-depth signature from HiFi data to detect CNVs

Repo: https://github.com/PacificBiosciences/HiFiCNV

HiFiCNV outputs – variant calls and IGV visuals

VCF v4.2

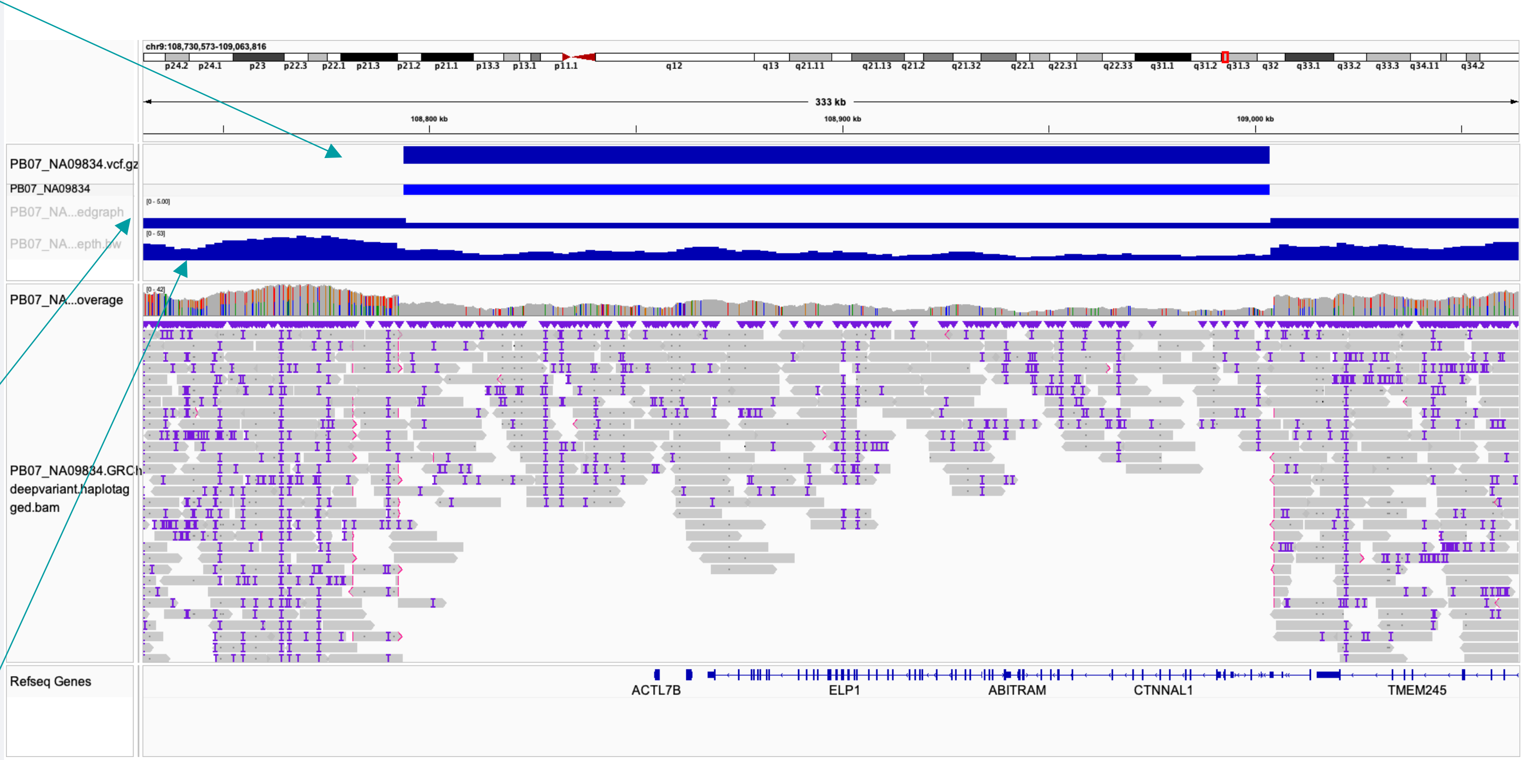
- Contains the variant calls (deviations from expected CN), usually < 50 PASS calls
- `TARGET_SIZE` filter – if event < 100 kb
- `QUAL` – based on next-most-likely CN state

Copy number track

- Reports CN from HMM
- Deviations from expected are in VCF

Depth track

- One entry per bin



Repo: https://github.com/PacificBiosciences/HiFiCNV

PacBio

HiFiCNV performance

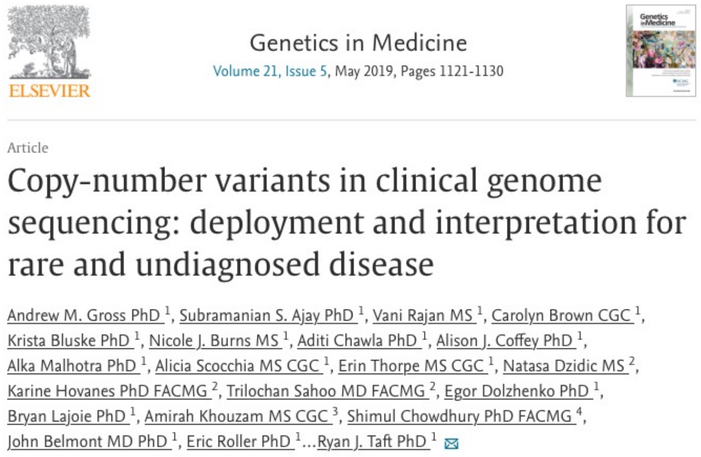
Evaluated on clinically-relevant CNVs from Gross *et al.*, 2019

- 17 samples
- 25 clinical events
  - Some are small (<100 kb)
  - Large gains and losses
  - Whole chromosome triplication

HiFiCNV accurate calls large CNVs

- 100% recall of all large (>100 kb) CNVs
- 80% recall for whole test set
- Complements pbsv for 100% recall

Metric	Value
HiFiCNV recall	80% (20 / 25)
HiFiCNV + pbsv recall	**100% (25 / 25)**
HiFiCNV base recall	97.43%
HiFiCNV base precision	58.52%



ELSEVIER

Genetics in Medicine

Volume 21, Issue 5, May 2019, Pages 1121-1130

Article

Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease

Andrew M. Gross PhD [1], Subramanian S. Ajay PhD [1], Vani Rajan MS [1], Carolyn Brown CGC [1], Krista Bluske PhD [1], Nicole J. Burns MS [1], Aditi Chawla PhD [1], Alison J. Coffey PhD [1], Alka Malhotra PhD [1], Alicia Scocchia MS CGC [1], Erin Thorpe MS CGC [1], Natasa Dzidic MS [2], Karine Hovanes PhD FACMG [2], Trilochan Sahoo MD FACMG [2], Egor Dolzhenko PhD [1], Bryan Lajoie PhD [1], Amirah Khouzam MS CGC [3], Shimul Chowdhury PhD FACMG [4], John Belmont MD PhD [1], Eric Roller PhD [1]...Ryan J. Taft PhD [1] ✉

Repo: https://github.com/PacificBiosciences/HiFiCNV

HiPhase – phasing SNVs, indels, and structural variants from HiFi datasets

Current phasing approaches are limited

- WhatsHap is most prominent tool for HiFi read-backed phasing
- Only phases small variants (SNVs and indels)
- Leaves ~10% of genes with multiple phase blocks
- Downsamples the data to 15x by default
- Does not span homozygous deletions or reference gaps
- Practical issues: hard to parallelize

HiPhase aims to phase all variant types called from HiFi reads

- *Jointly* phases SNVs, indels, and structural variants (insertions and deletions)
- No downsampling
- Includes logic to span coverage gaps with supplemental mappings
- Quality of life additions: innate multithreading, simultaneous statistics gathering and haplotagging

Repo: https://github.com/PacificBiosciences/HiPhase

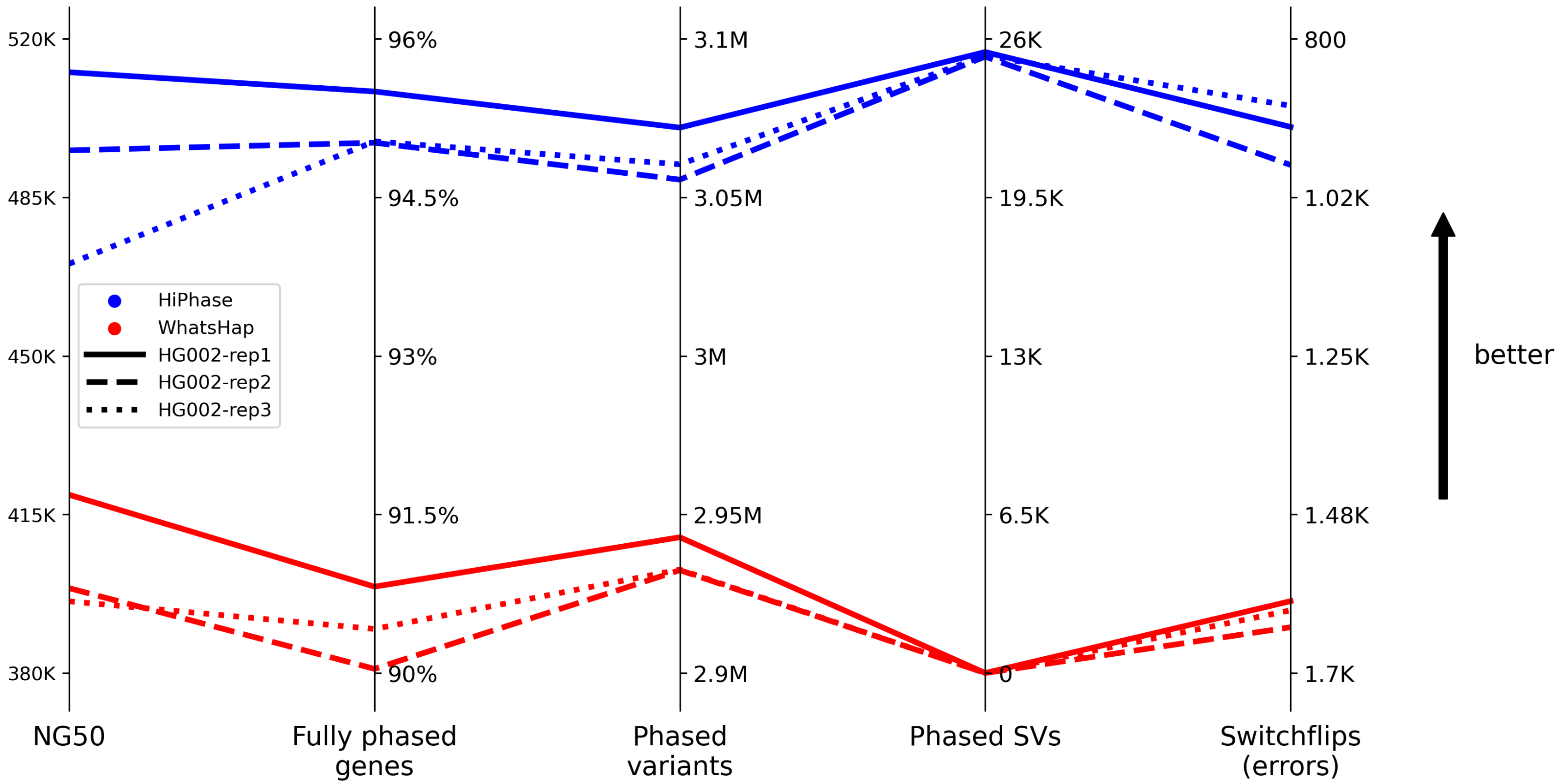
HiPhase performance

Datasets

- Three HG002 replicates
- Revio system

HiPhase improves over existing approach

- Per-replicate averages
- Block NG50: 493 kb
- Phased variants: 3.1M
- Phased SVs: 25K
- Fully phased genes: 95.2%
- Errors (switchflips): 933

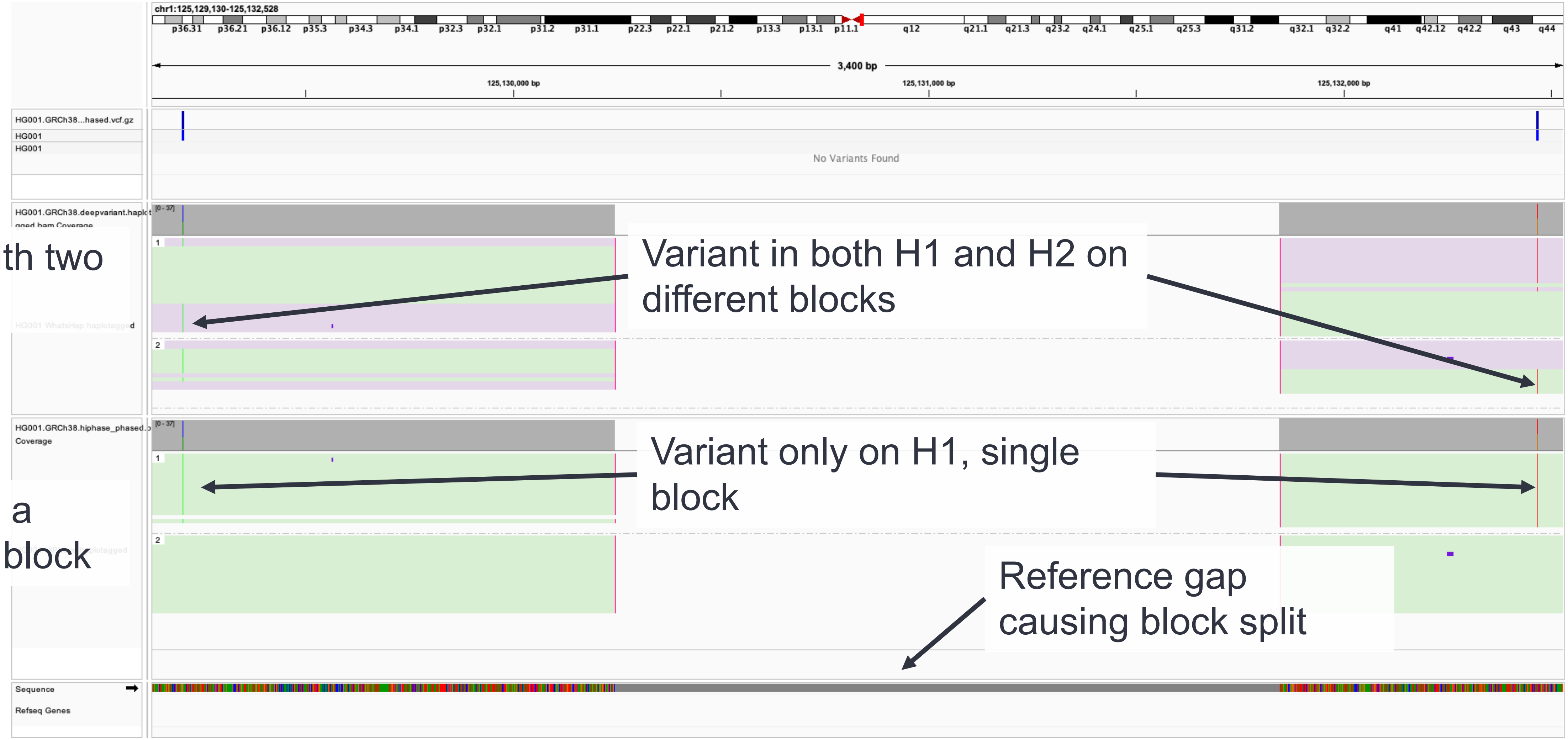


Datasets: https://downloads.pacbcloud.com/public/revio/2022Q4/

PacBio

Repo: https://github.com/PacificBiosciences/HiPhase

HiPhase – example gap spanning


chr1:125,129,130-125,132,528
p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 p11.1 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.2 q41 q42.12 q42.2 q43 q44
3,400 bp
125,130,000 bp
125,131,000 bp
125,132,000 bp
HG001.GRCh38...hased.vcf.gz
HG001
HG001
No Variants Found
HG001.GRCh38.deepvariant.haplotagged.bam Coverage
[0 - 37]
1
2
HG001.WhatsHap haplotagged
HG001.GRCh38.hiphase_phased.bam Coverage
[0 - 37]
1
2
Sequence
Refseq Genes

WhatsHap with two phase blocks

Variant in both H1 and H2 on different blocks

HiPhase with a single phase block

Variant only on H1, single block

Reference gap causing block split

Grouped by haplotype ID
Colored by phase block id

Repo: https://github.com/PacificBiosciences/HiPhase

HiFi target enrichment

Twist + PacBio partner to deliver off-the-shelf long-read panels

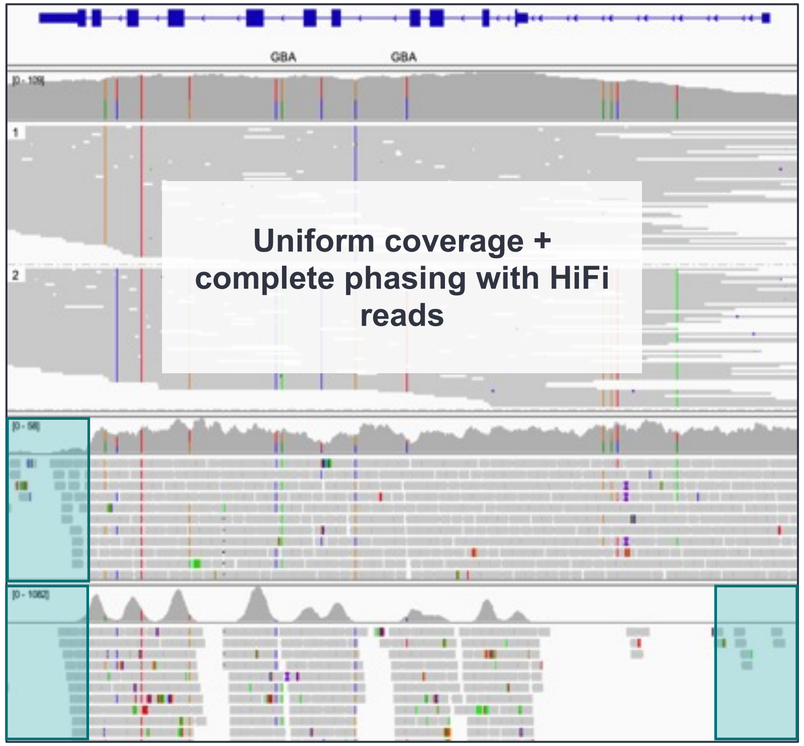
Targeted HiFi sequencing at scale

TWIST BIOSCIENCE

Initial product portfolio focuses on challenging genes

Twist *Alliance* panel	*Long Read PGx*	*Dark Genes*
Number of genes	49 + mtDNA[4]	389
Panel size	2 Mb	22 Mb
Samples/Sequel IIe SMRT Cell 8M	24	4
Sample/Revio SMRT Cell	72	12

Full gene coverage of medically relevant genes



GBA gene

HiFi target enrichment[1]

ILMN WGS[2]

ILMN WES[3]

Uniform coverage + complete phasing with HiFi reads

Haplotype 1

Haplotype 2

Variable coverage gaps in UTRs

No phasing with short reads

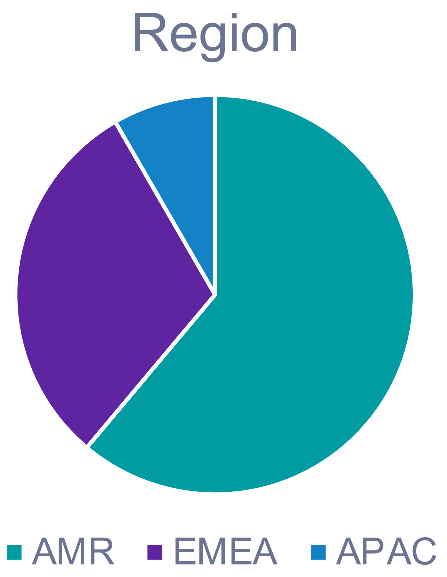
Available for sale now

https://www.twistbioscience.com/products/ngs/Long-Read-Sequencing-Panel

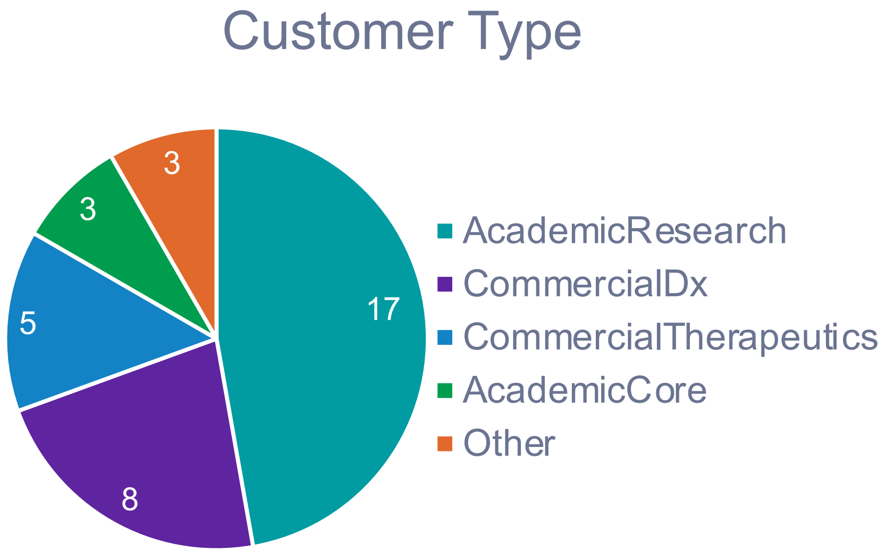
1. BCM-HGSC Twist *Alliance* panel, HG001 Sequel IIe system
2. https://www.biorxiv.org/content/10.1101/2020.12.11.422022v1.full (HG001 30x PCR free NovaSeq)
3. https://www.biorxiv.org/content/10.1101/2020.12.11.422022v1.full (HG001 75x TruSeq NovaSeq)
4. mtDNA spike-in probes available from Twist

Enthusiasm for Twist Panels since launch of custom panels last year

36 project in all regions



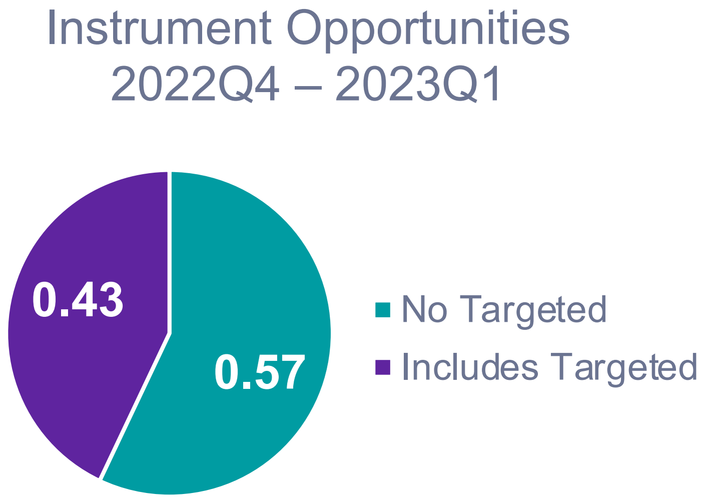
New customers + markets



31% Off-the-shelf PGx + Dark Genes

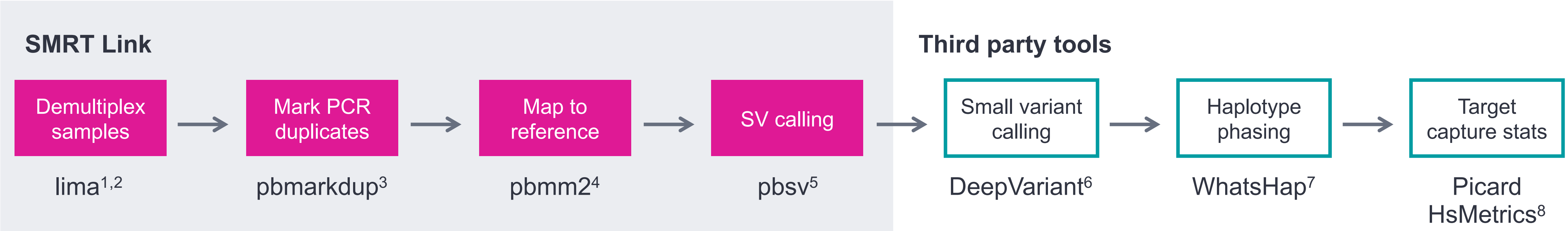
69% Custom panels HLA, Repeat Expansion, BRCA, P&A, cDNA

Expanding menu, driving instrument opportunities



Bioinformatics analysis recommendations

1. SMRT Link delivers mapped BAM compatible with third-party tools



SMRT Link

Demultiplex samples → Mark PCR duplicates → Map to reference → SV calling →

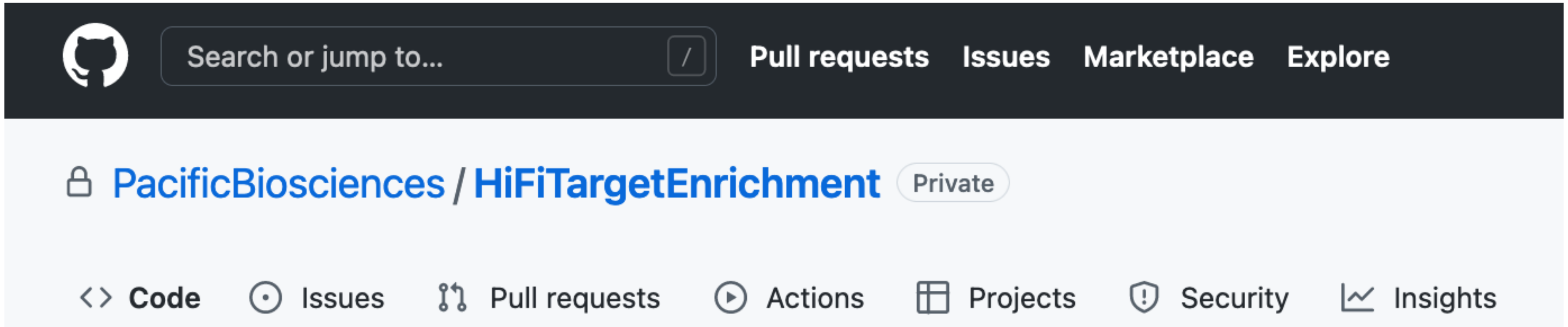
lima[1,2] pbmarkdup[3] pbmm2[4] pbsv[5]

Third party tools

Small variant calling → Haplotype phasing → Target capture stats

DeepVariant[6] WhatsHap[7] Picard HsMetrics[8]

2. GitHub command line pipeline delivers phased VCF, QC stats + plots for **advanced users**



PacificBiosciences / **HiFiTargetEnrichment** Private

1. https://github.com/PacificBiosciences/barcoding
2. Twist Barcode download: https://www.pacb.com/wp-content/uploads/Twist_Universal_Adapter_System_384.FASTA_.zip
3. https://github.com/PacificBiosciences/pbmarkdup
4. https://github.com/PacificBiosciences/pbmm2
5. https://github.com/PacificBiosciences/pbsv
6. https://github.com/google/deepvariant
7. https://github.com/WhatsHap/WhatsHap
8. https://snakemake-wrappers.readthedocs.io/en/stable/wrappers/picard/collecthsmetrics.html

Twist Alliance *Dark Genes* panel

- Comprehensive 22 Mb panel: full gene coverage for 389 **challenging medically-relevant genes**[2]
- Uncover genes in "**NGS dead zone**" that are difficult to sequence or map with short reads[2,3]
- Genes with pseudogenes, paralogs, repetitive sequence, or contained within segmental duplications.

Panel content[4.5]

A4GALT, ABCG8, ABO, ABR, ADAMTS10, ADAMTSL2, AFP, AGL, AGRN, ALOXE3, ANKRD11, ANO7, APOBEC1, APOBEC3H, APOC1, APOC2, APOC4, ARHGEF10, ASIP, ATPAF2, AXIN1, B3GAT3, BAX, BFSP2, BLOC1S3, BRAF, BSG, BTRC, C1R, C3, CABIN1, CALR3, CANT1, CASP10, CBR3, CBS, CCL3L1, CD247, CD320, CD4, CD55, CDH15, CDH17, CEL, CFC1, CFC1B, CFD, CFHR1, CFHR3, CHL1, CHMP1A, CHRNA4, CLCN7, CLIP2, CNR2, COL18A1, COL6A1, COL6A2, COX14, COX6B1, CR1, CREB3L3, CRYAA, CTDP1, CYB5R3, CYP2D6, CYP2G1P, CYP4F12, CYP4F3, D2HGDH, DAXX, DAZL, DCLRE1C, DEAF1, DGCR6, DIP2C, DLGAP2, DMPK, DNMT3L, DOK7, DPP6, DPY19L2, DRD4, DSPP, DUX4, DUX4L1, ECHS1, EEF1A2, EHMT1, EIF2B5, EIF4E, ELANE, ENO3, ESPN, ESRRA, ETFB, ETHE1, EXTL2, F7, FAM20C, FAT1, FCGR1A, FCGR2B, FCGR3A, FGF3, FGFRL1, FKBP8, FLAD1, FLG, FLT4, FOXN1, FSCN2, FTCD, FUT1, FUT3, FXN, G6PC3, GAK, GALNT9, GALR1, GALT, GBA, GCGR, GCSH, GDF3, GIP, GIPC3, GNPTG, GOLGA3, GP1BA, GP6, GPI, GPIHBP1, GRIN1, GRK1, GSTM1, GTF2I, GTF2IRD2, GUSB, GYPA, GYPB, GYPE, H19, HBG1, HBM, HCN2, HCN3, HES7, HLA-B, HLA-DQB1, HLA-DRB1, HMGCL, HMX1, HNF1A, HOMER2, HOXB8, HPD, HSD11B2, HYAL1, HYDIN, IFITM3, IFNL3, IGHA1, IGHG1, IGHG2, IGHM, IGHV3-21, IGKC, IGKV1-5, IKBKB, IKZF1, IMPA1, INPP5D, INPP5E, INSL3, INSR, JAG2, KANSL1, KATNAL2, KCNE1, KCNJ18, KCNV2, KDM2B, KIR2DL1, KIR2DL3, KIR3DL1, KISS1, KISS1R, KLF11, KLF14, KLK4, KMT2C, KNG1, KRTAP1-1, LAMB1, LBR, LCE3B, LHFPL5, LIPN, LIX1, LMF1, LMNB2, LPA, LRIG2, LRPAP1, LZTFL1, MAFA, MAN1B1, MAP2K3, MARVELD2, MASP2, MBOAT7, MC1R, MDK, MEST, MLC1, MLPH, MOGS, MPG, MRC1, MST1R, MUC1, MUC16, MUC3A, MUC4, MUC5B, MUSK, MYO9B, MYOT, MYT1, NACA, NAIP, NAPRT, NBEAP1, NCF1, NCF1C, NCR3, NDUFA6, NDUFAF1, NDUFB1, NDUFV3, NFKBIL1, NLRP12, NLRP2, NLRP7, NOD1, NOTCH2, NPM1, NPPA, NSMF, NUTM2B, NUTM2D, OCLN, OPRL1, OR12D2, OR4F5, OR51A2, ORC6, P2RX2, P2RX5, PADI4, PAPSS2, PCBP1, PCCB, PCDHA10, PCMT1, PDE4DIP, PDE6B, PDLIM3, PDPK1, PDSS1, PEX5, PGAM5, PHKG2, PIGV, PKD1, PKN3, PLA2G10, PLTP, PMS2, PNKP, POLG2, PPIA, PPIP5K1, PRG4, PRKCG, PRODH, PROZ, PRSS2, PSPH, PTEN, PTK6, PTPRC, PTPRN2, PTPRQ, PXDN, RFX2, RGPD3, RHCE, RHOA, RNF212, RNF213, RPIA, RPL22, RPN1, RPS17, SAR1B, SBDS, SBK3, SDHA, SEC63, SEMG1, SERPINF2, SH2B1, SHANK2, SHANK3, SIGLEC16, SIRT3, SLC17A5, SLC22A1, SLC22A12, SLC26A9, SLC27A4, SLC27A5, SLC29A4, SLC5A11, SLC6A18, SLC6A3, SMG1, SMN1, SMN2, SMOC2, SNORD64, SNTG2, SOHLH1, SPATA31C1, SPI1, SPRN, SRGAP2, SRR, SSTR5, STK11, STXBP2, SULT1A1, SUZ12, TAPBP, TAS2R45, TAS2R46, TBXA2R, TCF3, TERT, TFPT, THBS2, TJP2, TM4SF19, TMC6, TMEM114, TNNI3, TNNT1, TNNT3, TPCN2, TPO, TRAPPC10, TRBV9, TRMT1, TRPM4, TTC37, TTLL1, TUBGCP6, TWIST2, TYK2, TYMS, U2AF1, UGT2A1, UGT2A2, UGT2B17, UGT2B28, UNKL, USP8, UVSSA, VANGL1, VKORC1, VPS53, ZAN, ZNF141, ZNF407, ZNF419, ZNF469, ZNF479

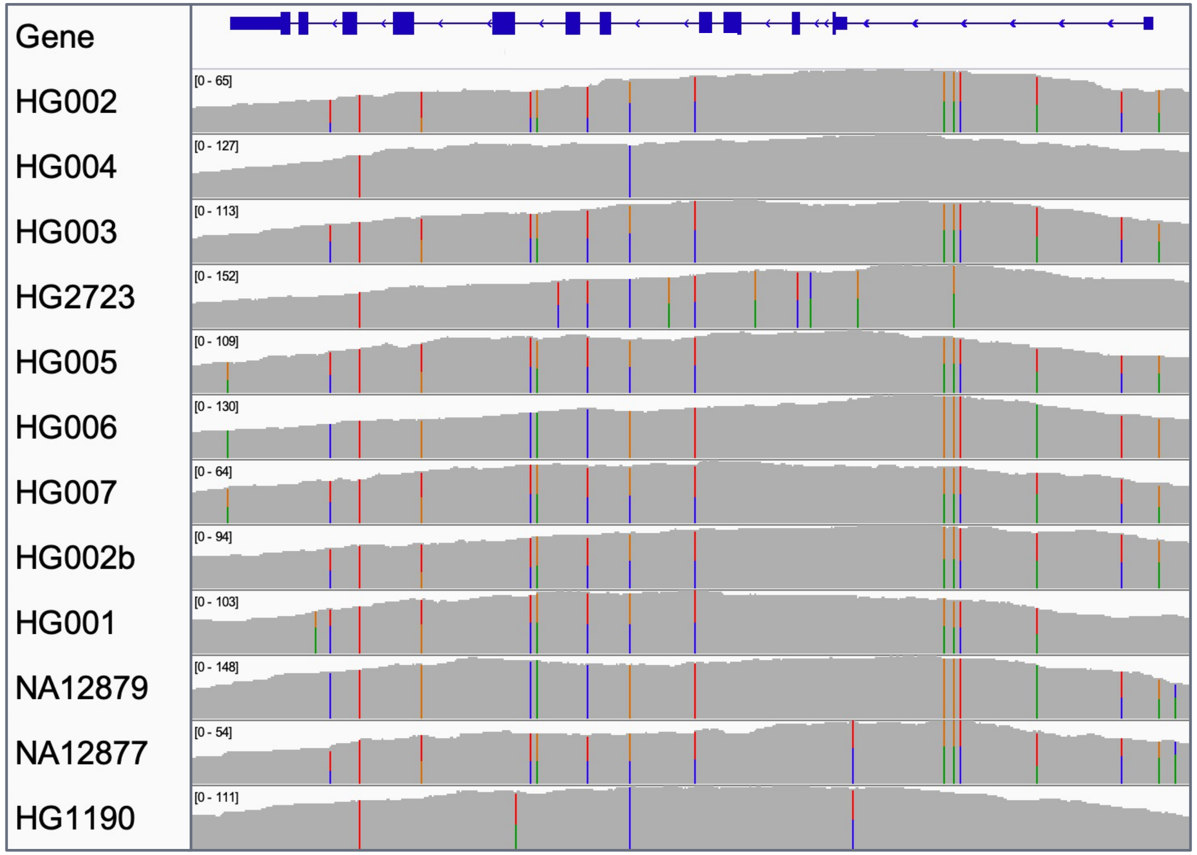
1. Ji *et al.* Characterizing the genetic polymorphisms in 370 challenging medically relevant genes using long-read sequencing data from 41 human individuals among 19 global populations. bioRXiv https://doi.org/10.1101/2022.08.03.502734
2. Mandelker *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genetics in Medicine 2016.
3. Wenger *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotech (2019)
4. https://downloads.pacbcloud.com/public/dataset/HiFiTE_Revio/Nov_2022/TwistAllianceDarkGene/TwistAllianceDarkGenes_GeneList.txt
5. BED file: https://www.twistbioscience.com/resources/data-files/twist-alliance-dark-genes-panel-bed-file

Twist — PacBio workflow compatible with Sequel IIe and Revio systems

Summary performance for *Dark Genes* panel

System	Sequel IIe	Revio
Samples / SMRT Cell	**4**	**12**
HiFi yield / SMRT Cell	19.53 Gb	51.43 Gb
Mean read length	5.2 kb	5.5 kb
Median read quality	Q40	Q38
Mean reads / sample	893,459	724,795
Mean target coverage	75-fold	75-fold
Target bases ≥10-fold	93%	93%
Fold enrichment	54-fold	65-fold

Uniform coverage at *GBA* across 12-plex on Revio system



https://www.pacb.com/connect/datasets/#targeted-datasets

Twist Alliance *Long Read PGx* panel

Why PGx?

- ~99% of adults have an actionable PGx variant (US, UK Biobank studies)
- ~50% of US adults are prescribed a drug for which there are CPIC guidelines
- > 100K adverse drug reactions per year in the US costing >$30B

Panel design

2 Mb target region
49 genes
full-length mtDNA spike-in available
39 full-length genes enable phasing
Includes all 20 genes with CPIC guidelines

CYP genes	HLA	Others	
CYP1A2*	***HLA-A***	***ABCB1***	*HTR2C*
CYP2B6$^+$	***HLA-B***	***ABCG2***	***IFNL3***
CYP2C19	***HLA-DQA1***	*ADD1*	***MT-RNR1***
CYP2C8	***HLA-DRB1***	***ADRA2A***	***MTHFR***
CYP2C9		***ANKK1***	***NAGS***
CYP2D6		***APOL1***	***NAT2***
CYP3A4		***BCHE***	***NUDT15***
CYP3A5		*CACNA1S*	***OPRD1***
CYP4F2		*CFTR*	***OPRK1***
		COMT	***OPRM1***
		CTBP2P2	*POLG*
		DPYD	***RYR1***
		DRD2	***SLC6A4***
		F2	***SLCO1B1***
		F5	***TPMT***
		G6PD	***UGT1A1***
		GBA	***UGT2B15***
		GRIK4	***VKORC1***
			YEATS4

***Bold** denotes full-gene coverage
$^+$Underline denotes inclusion in a CPIC guideline

- Ji Y et al. Preemptive pharmacogenetic testing: a comprehensive analysis of five actionable pharmacogenomic genes using next-generation DNA sequencing and a customized CYP2D6 genotyping cascade. *J Mol Diagn* (2015).
- Chanfreau-Coffinier C, et al. Projected prevalence of actionable pharmacogenetic variants and level A drugs prescribed among US veterans health administration pharmacy users. *JAMA Netw Open* (2019)
- BED file: https://www.twistbioscience.com/resources/data-files/twist-alliance-long-read-pgx-panel-bed-file

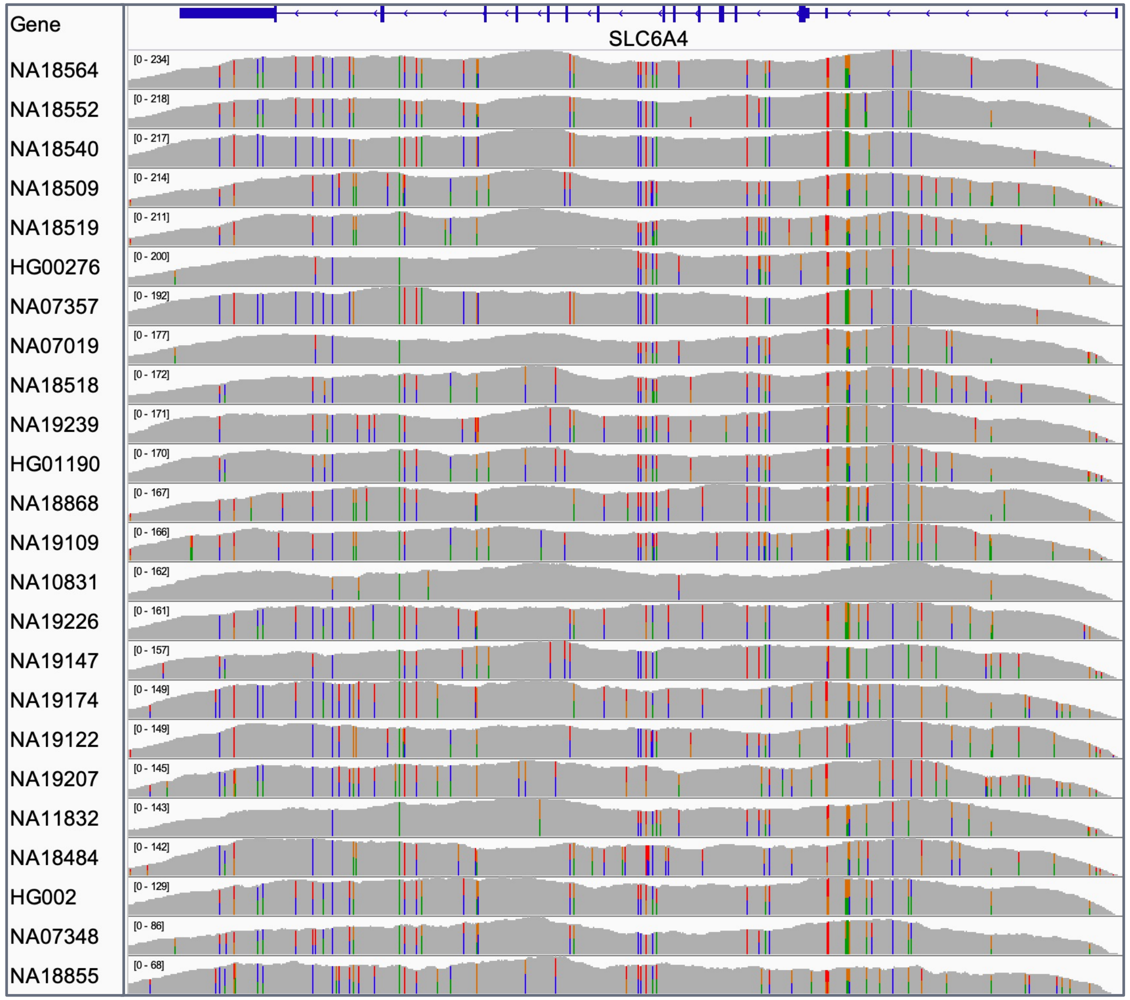
Twist Alliance *Long Read PGx* panel

Sequel IIe system 24-plex of reference samples

Panel size	2 Mb
HiFi yield per SMRT Cell	20.11 Gb
Mean read length	5.3 kb
Mean reads per sample	149,749
Mean target coverage	190-fold
Target bases ≥20-fold	96%
Fold enrichment	784-fold
PCR duplicate rate	2%
Demultiplex yield	96%

https://www.pacb.com/connect/datasets

Uniform coverage at *SLC6A4* across 24 samples on Sequel IIe system

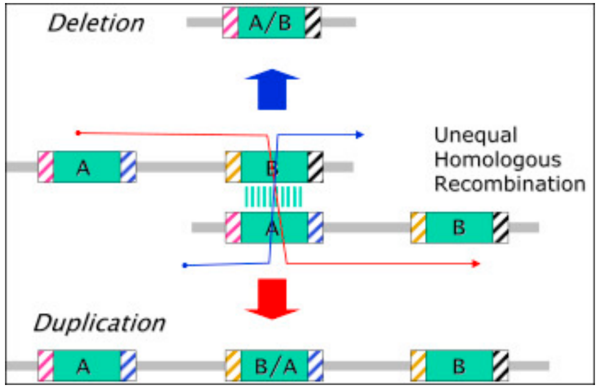


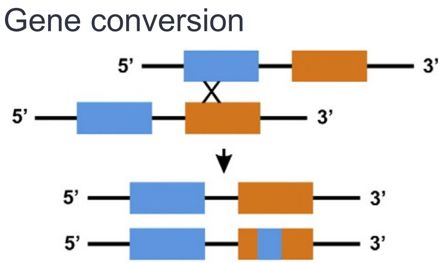
chr17: 30,192,000 - 30,236,000 (42 kb)

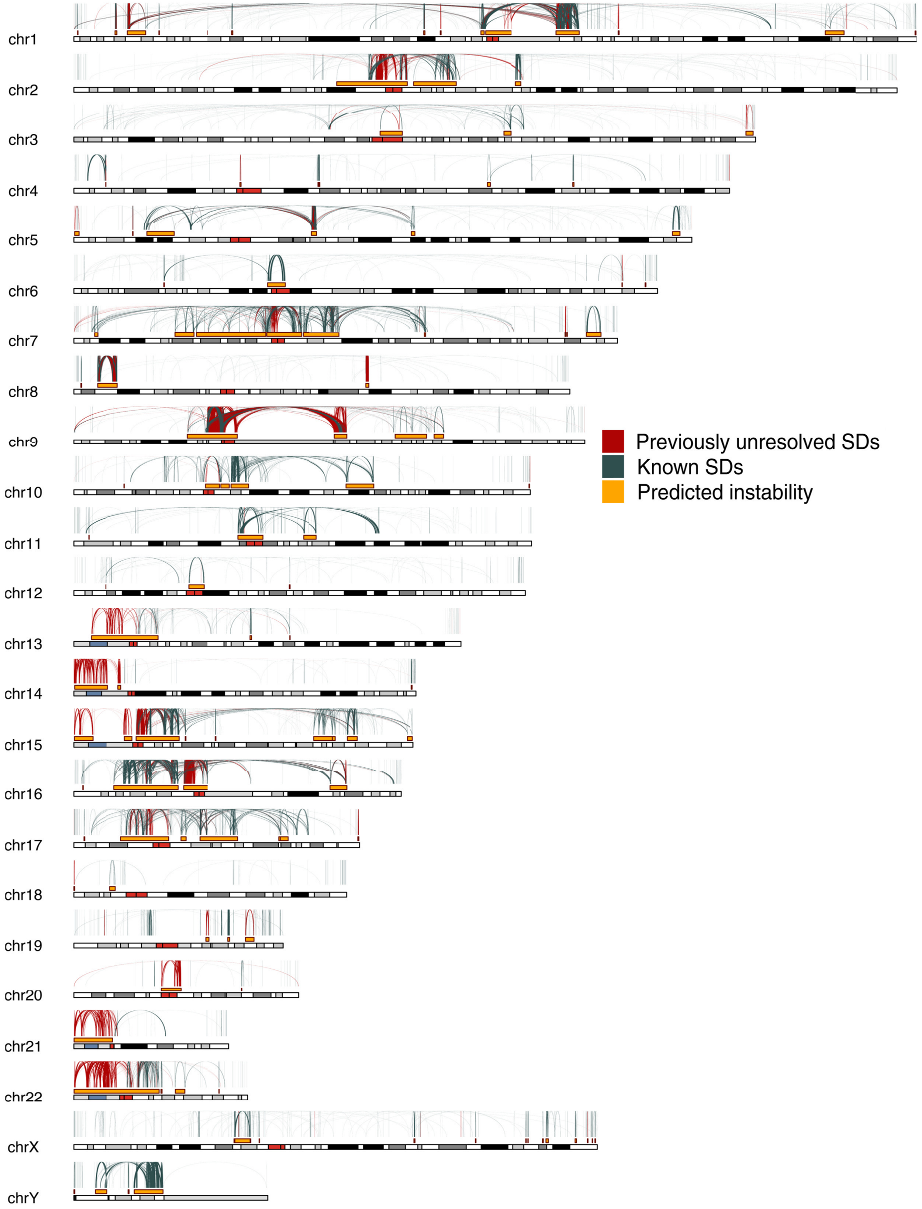
Full gene coverage + sample uniformity

Segmental duplications are informatically challenging

- Segmental duplications comprise 7% of the human genome
- Many clinically relevant genes fall into segmental duplications
- Segmental duplications are hotspots for structural variations, including deletions, duplications and gene conversions.
- High sequence similarity poses challenges to read alignment and variant calling

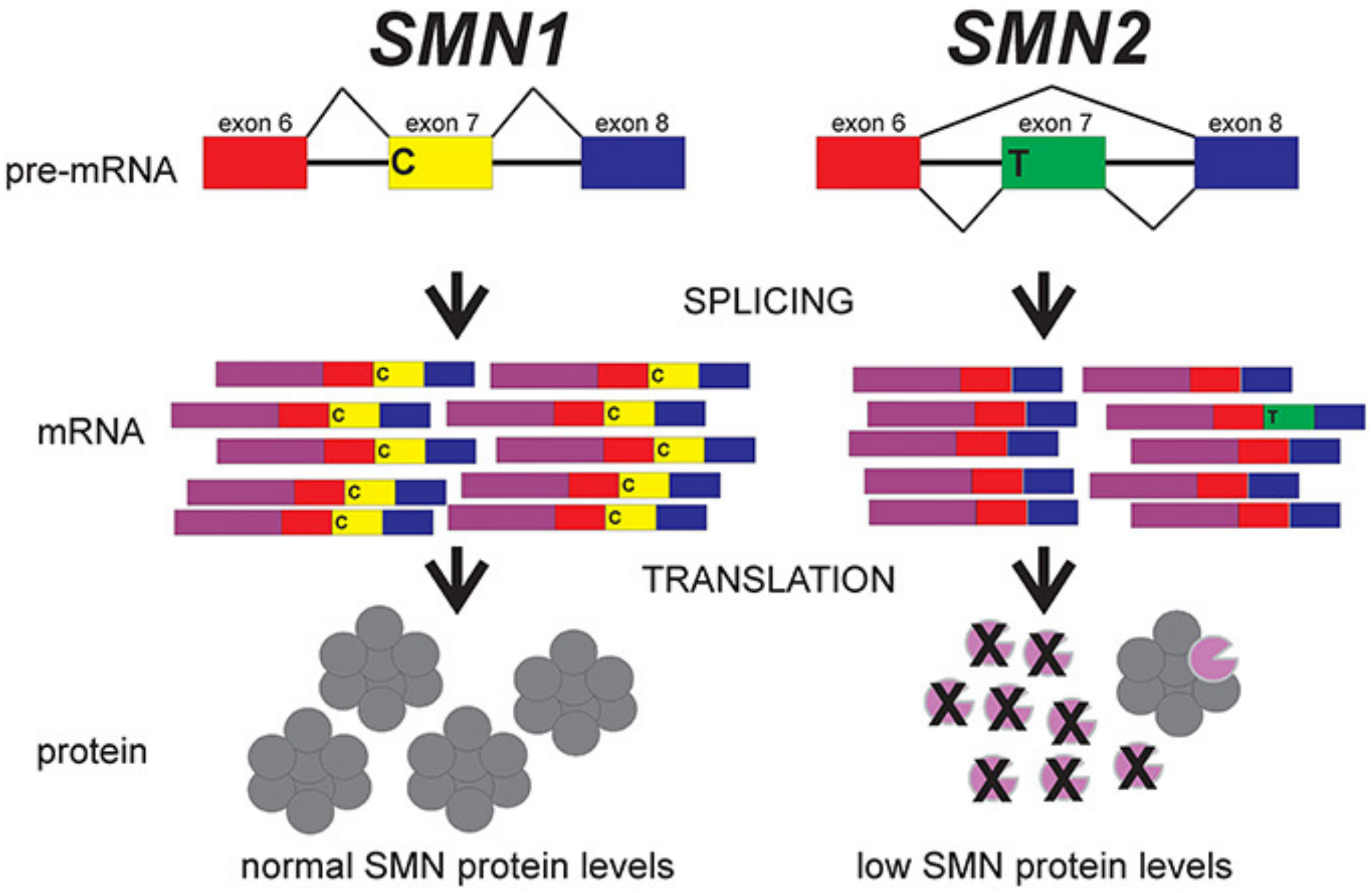



Gene conversion



Vollger et al. *Science*, 2022
Antonarakis, *Medical and Health Genomics*, 2016
Borg et al. *Clinical Biochemistry*, 2009

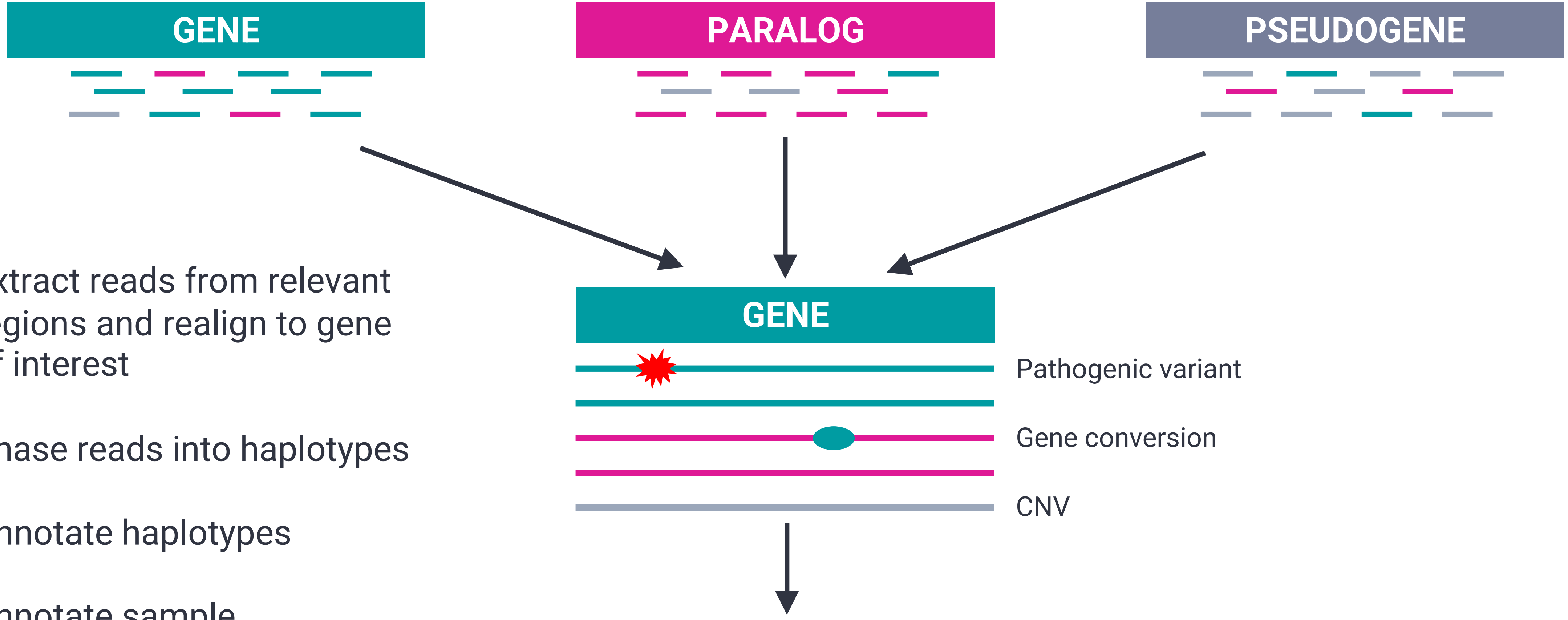
Spinal muscular atrophy (SMA) and *SMN1/SMN2*



- Spinal muscular atrophy (SMA) is an autosomal recessive neurodegenerative disorder - a leading cause of infant death
- Lack of *SMN1* leads to SMA
  - ~96% of SMA is due to biallelic absence of c.840C
  - ~4% involves other pathogenic variants within *SMN1*
- *SMN1* testing is usually done by dosage (copy number) testing with MLPA or qPCR (targeting c.840C)
- Number of copies of *SMN2* modifies disease severity
- *SMN1/SMN2* are ~30kb long and >99.9% homologous
  - 41% of HiFi WGS reads have a MAPQ lower than 5, and 85% of reads have a MAPQ lower than 20.

Butchbach et al, Front. Mol. Biosci, 2016
Keinath et al. *Applications of Clinical Genetics*, 2021

Paraphase: our approach for highly homologous genes



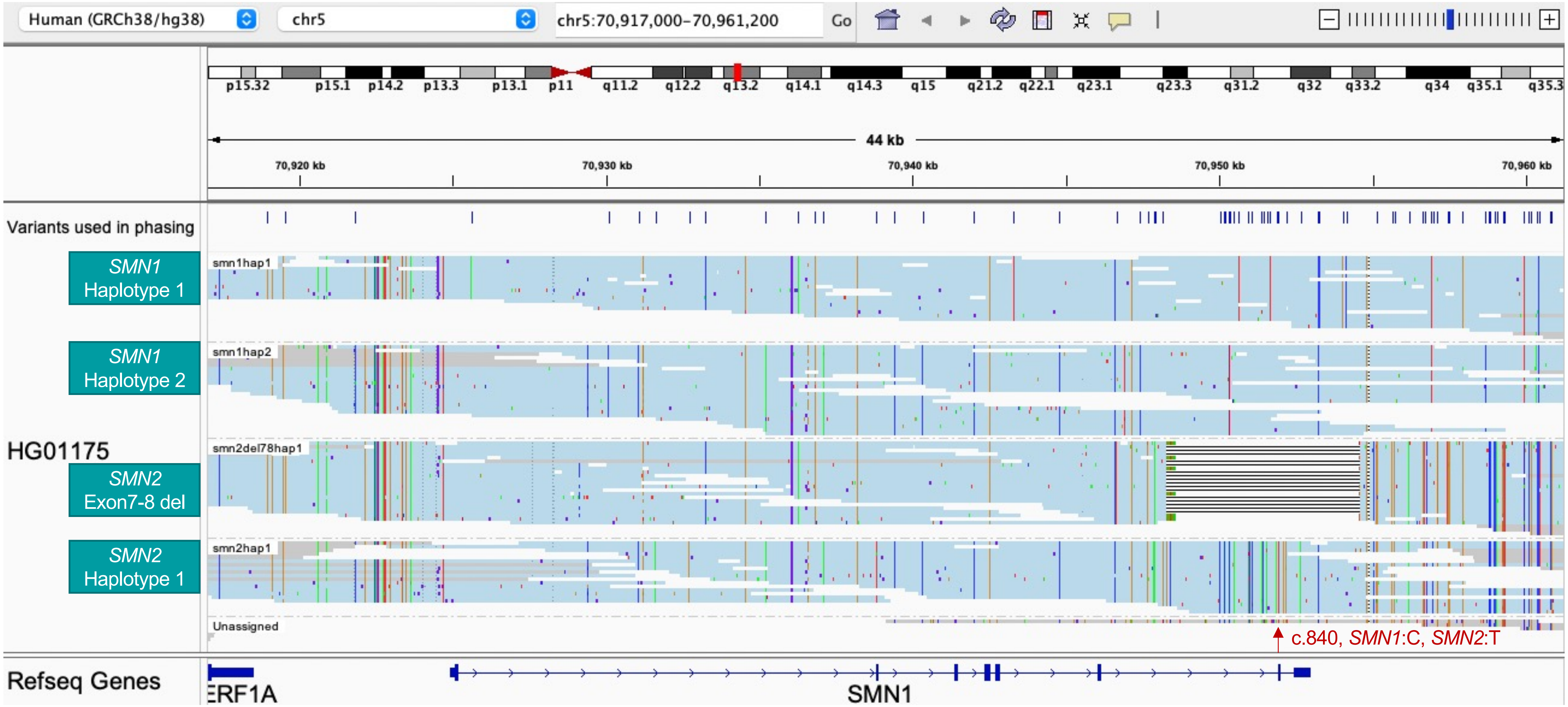
1. Extract reads from relevant regions and realign to gene of interest
2. Phase reads into haplotypes
3. Annotate haplotypes
4. Annotate sample

Report

- Copy number of the functional gene
- Disease/carrier status

PacBio Chen et al, AJHG, 2023

Phasing *SMN1/SMN2* haplotypes determines copy numbers



Chen et al. Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *The American Journal of Human Genetics.* 2023;0(0). doi:10.1016/j.ajhg.2023.01.001 | https://github.com/PacificBiosciences/paraphase

Now applying Paraphase to more clinically relevant segmental duplications

<u>Genes being assessed and associated diseases</u>

- *PMS2* (Lynch Syndrome)
- *STRC* (hereditary hearing loss and deafness)
- *IKBKG* (Incontinentia Pigmenti)
- *NCF1* (chronic granulomatous disease; Williams syndrome)
- *NEB* (Nemaline myopathy)
- *F8* (intron 22 inversion, Hemophilia A)
- *CFC1* (heterotaxy syndrome)
- and more…

Pharmacogenomics

Content of *Long Read PGx* panel

CYP genes	HLA	Others	
CYP1A2	*HLA-A*	*ABCB1*	*HTR2C*
CYP2B6	*HLA-B*	*ABCG2*	*IFNL3*
CYP2C19	*HLA-DQA1*	*ADD1*	*MT-RNR1*
CYP2C8	*HLA-DRB1*	*ADRA2A*	*MTHFR*
CYP2C9		*ANKK1*	*NAGS*
CYP2D6		*APOL1*	*NAT2*
CYP3A4		*BCHE*	*NUDT15*
CYP3A5		*CACNA1S*	*OPRD1*
CYP4F2		*CFTR*	*OPRK1*
		COMT	*OPRM1*
		CTBP2P2	*POLG*
		DPYD	*RYR1*
		DRD2	*SLC6A4*
		F2	*SLCO1B1*
		F5	*TPMT*
		G6PD	*UGT1A1*
		GBA	*UGT2B15*
		GRIK4	*VKORC1*
			YEATS4

Data release

https://www.pacb.com/connect/datasets/#targeted-datasets

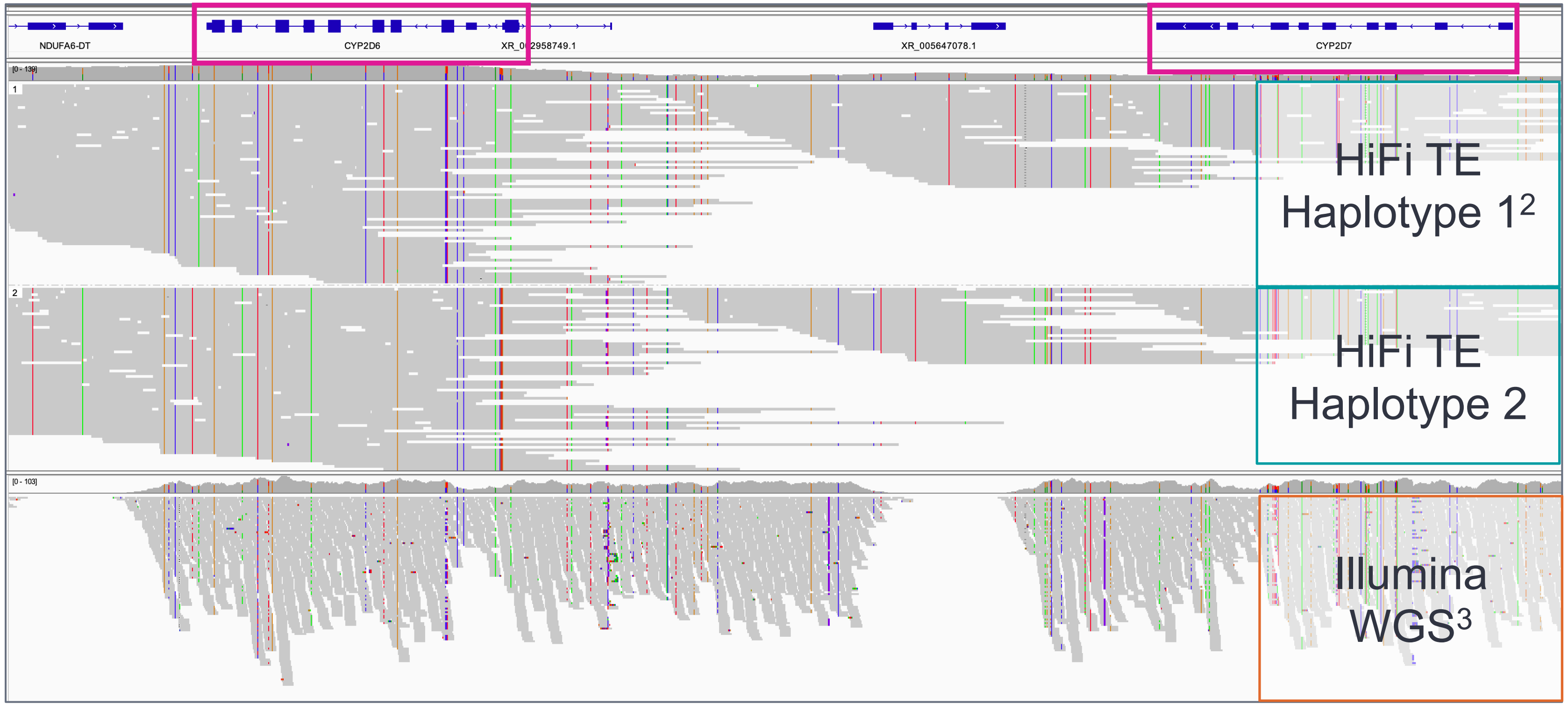
Sequel IIe system

- Samples
  - HG002
  - HG00276
  - HG01190
  - NA07348
  - NA11832
  - NA18518
  - NA19109
  - NA19174
  - NA19207
  - NA19226
- Files available
  - Aligned BAM (hg38 reference)
  - HiFi reads Fastq
  - Gene list

Accurate star allele calling at *CYP2D6*

Benchmarking PacBio *CYP2D6* genotyper, pangu[1]

Haplotype phasing spans *CYP2D6* and paralog *CYP2D7*



HG002, chr22:42,123,767-42,145,283

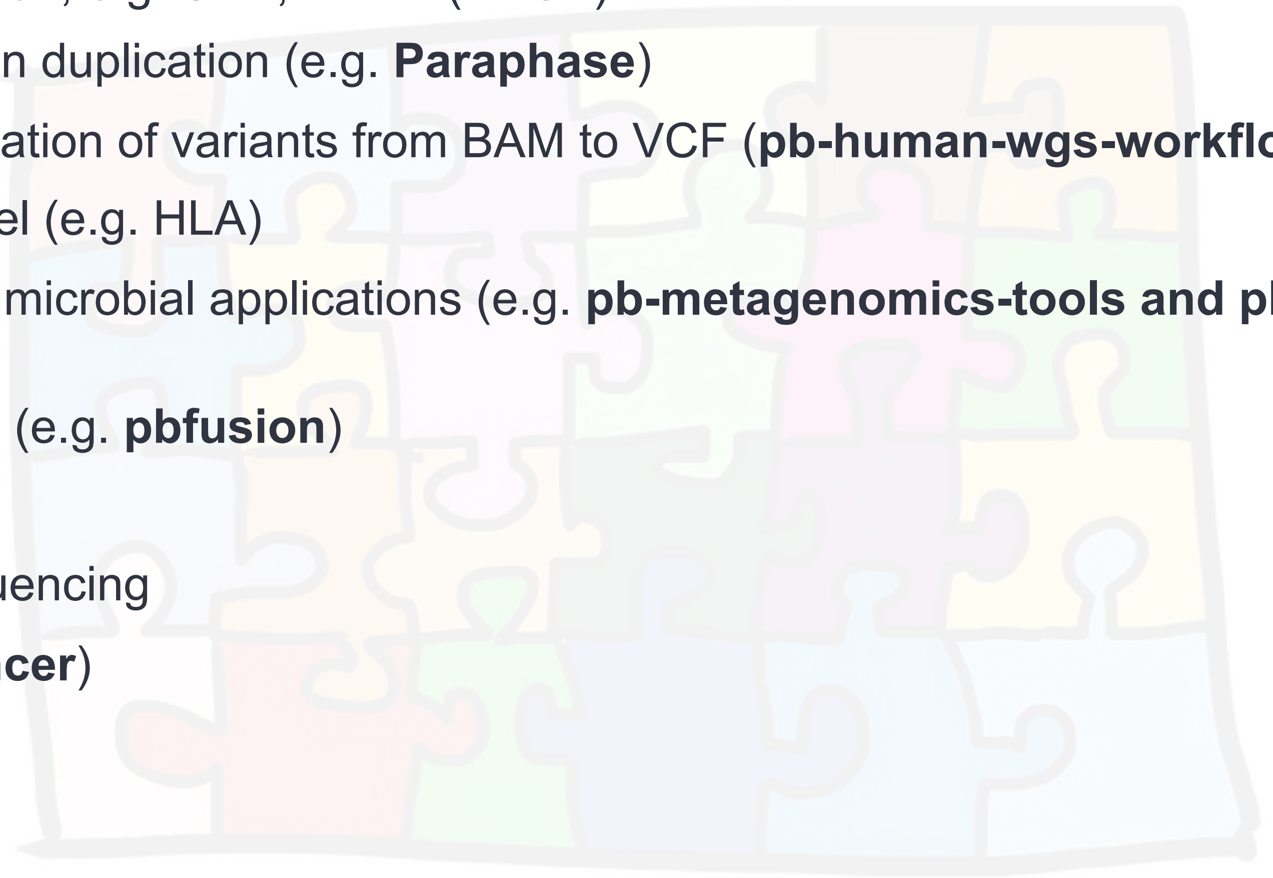
21 / 23 concordant genotypes[2]
1 corrected call

Sample	GeT-RM	PacBio	Concordance
HG002	*2/*4	*2/*4	✓
HG00276	*4/*5	*4/*5	✓
HG01190	*68+*4/*5	*68+*4/*5	✓
NA07019	*1/*4	*1/*4	✓
NA07348	*1/*6	*1/*6	✓
NA10831	*4/*5	*4/*5	✓
NA11832	*1/*68+*4	*1/*68+*4	✓
NA18484	*1/*17	*1/*17	✓
NA18509	*2/*17	*2/*17	✓
NA18518	*17/*29	*17/*29	✓
NA18519	*29/*1	*29/*106	✓ – corrected
NA18540	*36x2+*10/*41	*36+*10/*41	✗
NA18552	*1/*14	*1/*14	✓
NA18564	*2/*36+*10	*2/*36+*10	✓
NA18855	*1/*5	*1/(*5)	✓
NA18868	*2/*5	*2/*5	✓
NA19109	*2x2/*29	*2x2/*29	✓
NA19122	*2/*17	*2/*17	✓
NA19147	*17/*29	*17/*29	✓
NA19174	*4/*40	*4/*40	✓
NA19207	*2x2/*10	*2x2/*10	✓
NA19226	*2/*2x2	*2/*2x2	✓
NA19239	*15/*17	*15/*17	✓

[1] https://github.com/PacificBiosciences/Pangu
[2] https://downloads.pacbcloud.com/public/dataset/HiFiTE_SqIIe/Oct_2022/TwistAllianceLongReadPGx/
[3] https://storage.googleapis.com/brain-genomics-public/research/sequencing/grch38/bam/novaseq/wgs_pcr_free/50x/HG002.novaseq.pcr-free.50x.dedup.grch38.bam

Completing the puzzle and enabling full featured genomes

A laundry list of bioinformatics solutions required for different problems



Examples of problems the computational biology group in PacBio works on:

- Caller for complex repeats characterization, e.g. STR, VNTR (**TRGT**)
- Caller for genes involved in segmentation duplication (e.g. **Paraphase**)
- Workflow for comprehensive characterization of variants from BAM to VCF (**pb-human-wgs-workflow**)
- Characterizing genes from targeted panel (e.g. HLA)
- Maximizing utility of HiFi sequencing for microbial applications (e.g. **pb-metagenomics-tools and pb-16S-nf**)
- Extracting more from the transcriptomes (e.g. **pbfusion**)
- 5-base sequencing (5mC) and beyond
- Benchmarking the accuracy of HiFi sequencing
- Emerging applications (e.g. WGS in **cancer**)

PacBio

伯森生技

www.blossombio.com

PacBio

www.pacb.com