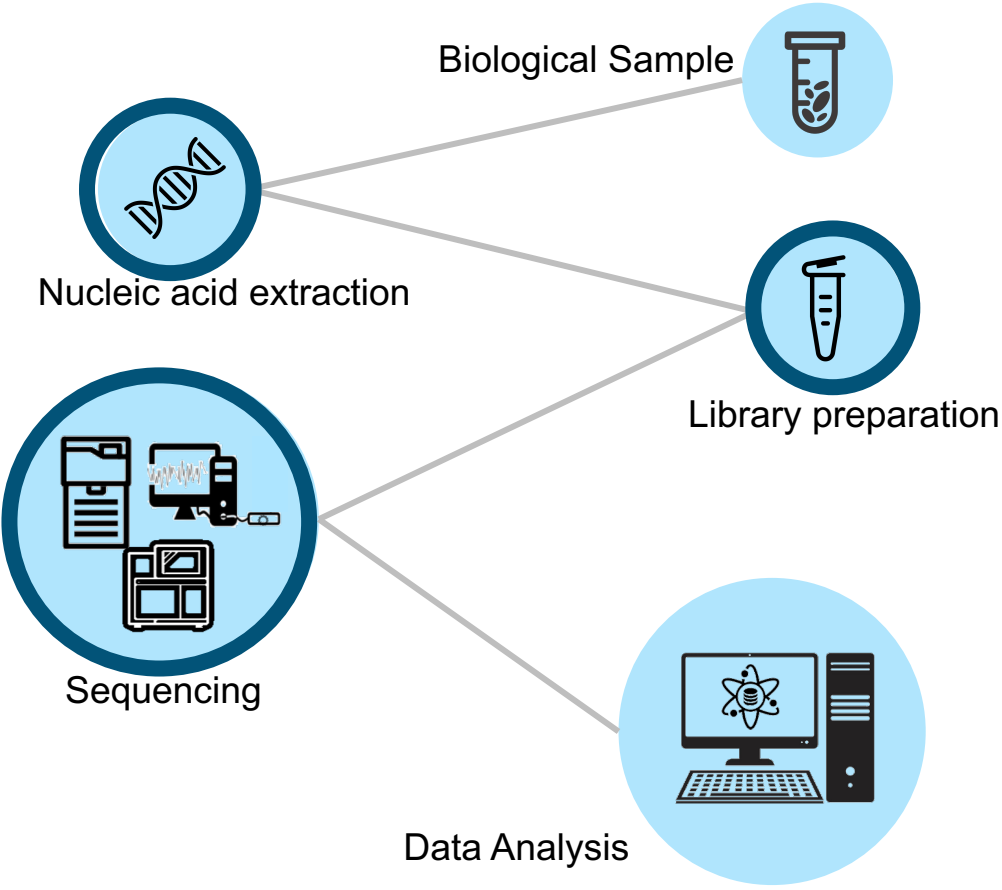
Basic long-read sequencing QC

Where do I start? Fundamental analyses of Nanopore and PacBio sequencing data

Gabriel Rech, Ph.D.

@rechgab

Workflow in sequencing projects



Quality Control (QC) at different steps!

Determine:

- **Quality (accuracy)**
- **Read Length**
- **Yield**

How do we measure **Quality (accuracy)?**

Phred Quality Score

Quality of the identification of the nucleobases generated by sequencing

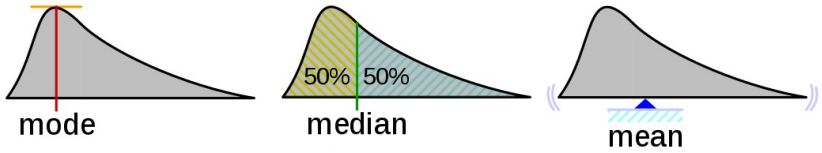
A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect. $Q = -10 \log_{10} P \longrightarrow P = 10^{\frac{-Q}{10}}$

Phred quality scores are logarithmically linked to error probabilities		
Phred Quality Score (Q-Score)	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

The FASTQ format encodes Phred scores as ASCII characters

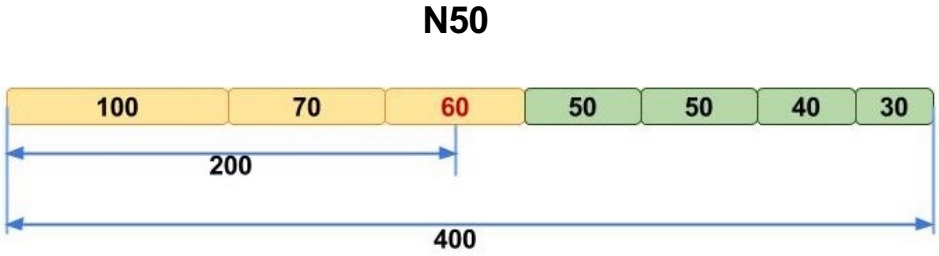
Table 1 ASCII Characters Encoding Q-scores 0–40

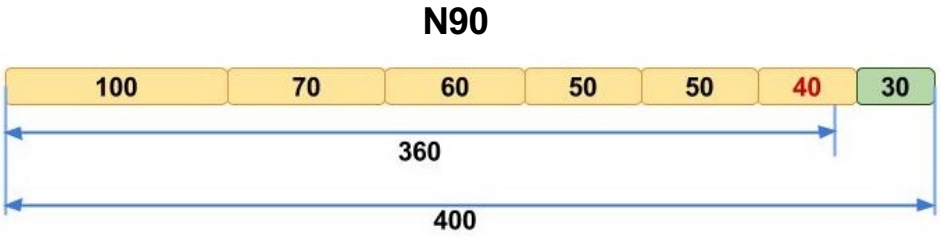
Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			



How do we measure **Read Length?**

Min
Max
Mean
N50
N90





https://www.molecularecologist.com/2017/03/29/whats-n50/

How do we measure **yield?**

- **Total amount of bases sequenced per run/flowcell**

Determinants of **quality**, **read length** and **yield**

- **DNA quality:**

  High-molecular weight DNA

  Purity (free of proteins, phenol and other unwanted organic compounds)

- **Platform (ONT/PacBio) / Instrument**

- **Library Preparation / Sequencing mode:**

  ONT: *Sequencing kit (Rapid, Ligation, Q20+, etc.…)*
  PacBio: ~~CCS vs. CLR~~ **HiFi**

- **Flow cell version**

  SMRT Cell 1M vs. 8M → 64M?(R&D)
  R9.4.1 vs R10.4

Sequencing instruments

Pacific Biosciences (PacBio)

RS II

Discontinued &
Unsupported since 2021

Sequel system
Sequel II system
Sequel IIe system

Oxford Nanopore Technologies (ONT)

PromethION

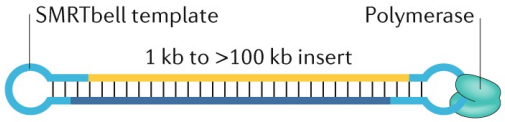
GridION

MinION

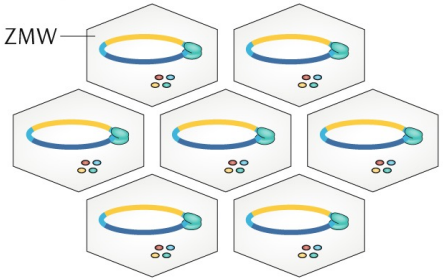
PacBio

a PacBio SMRT sequencing

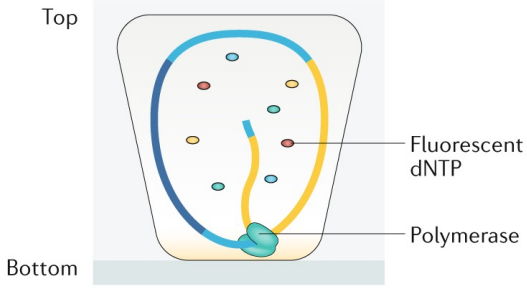
Template topology



Flow cell (top view)



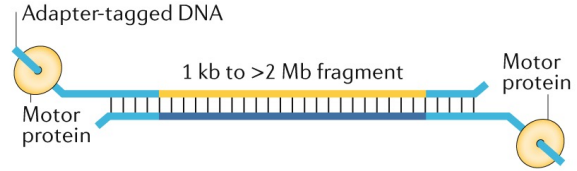
Single ZMW (cross section)



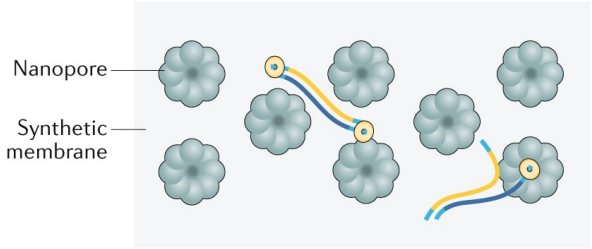
ONT

b ONT sequencing

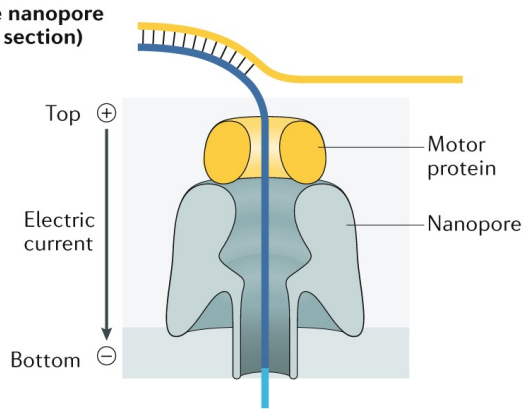
Template topology



Flow cell (top view)



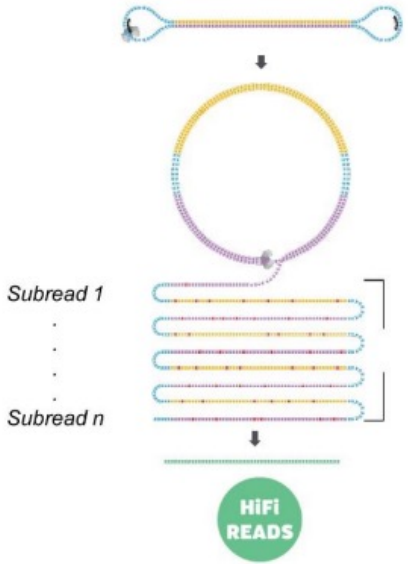
Single nanopore (cross section)



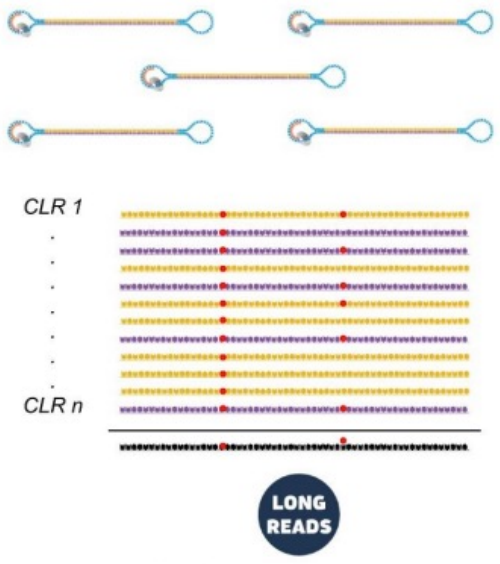
Logsdon et al (2020)

Two modes of PacBio (SMRT) sequencing

Circular Consensus Sequencing (CCS)

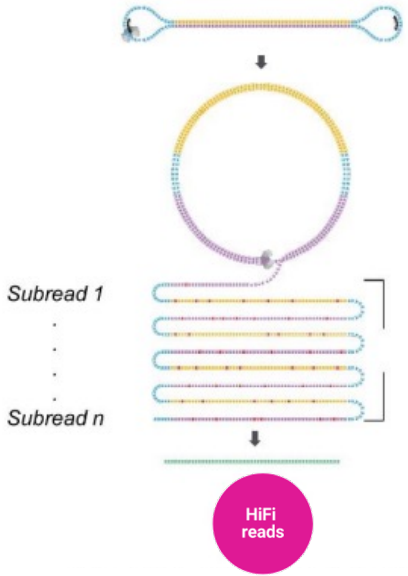


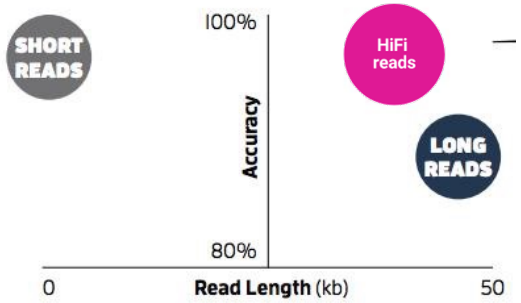
Continuous Long Read (CLR) Sequencing



Two modes of PacBio (SMRT) sequencing

Circular Consensus Sequencing (CCS)



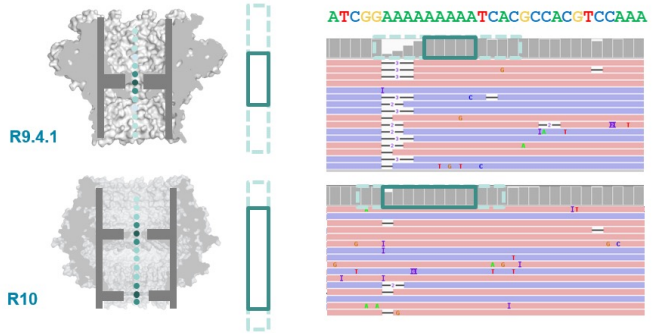


15-20 kb (up to 25 kb)
Accuracy >99.9%

Continuous Long Read (CLR) Sequencing

ONT flow cell and kits chemistry

- Ultra-long sequencing kit
- R9.4.1 vs R10.4
- Chemistry for Q20+ sequencing



Sequencing mode	Pore	N50	Yield / flow cell	Median read accuracy
Long	R9.4	30-50 kb	~120-150 Gb	96%
Ultra-Long	R9.4	60-100 kb	~ 100 Gb	96%
Q20 EA	R10.4	40-50 kb	~45 Gb	99%

Miten Jain, NanoporConf 2021

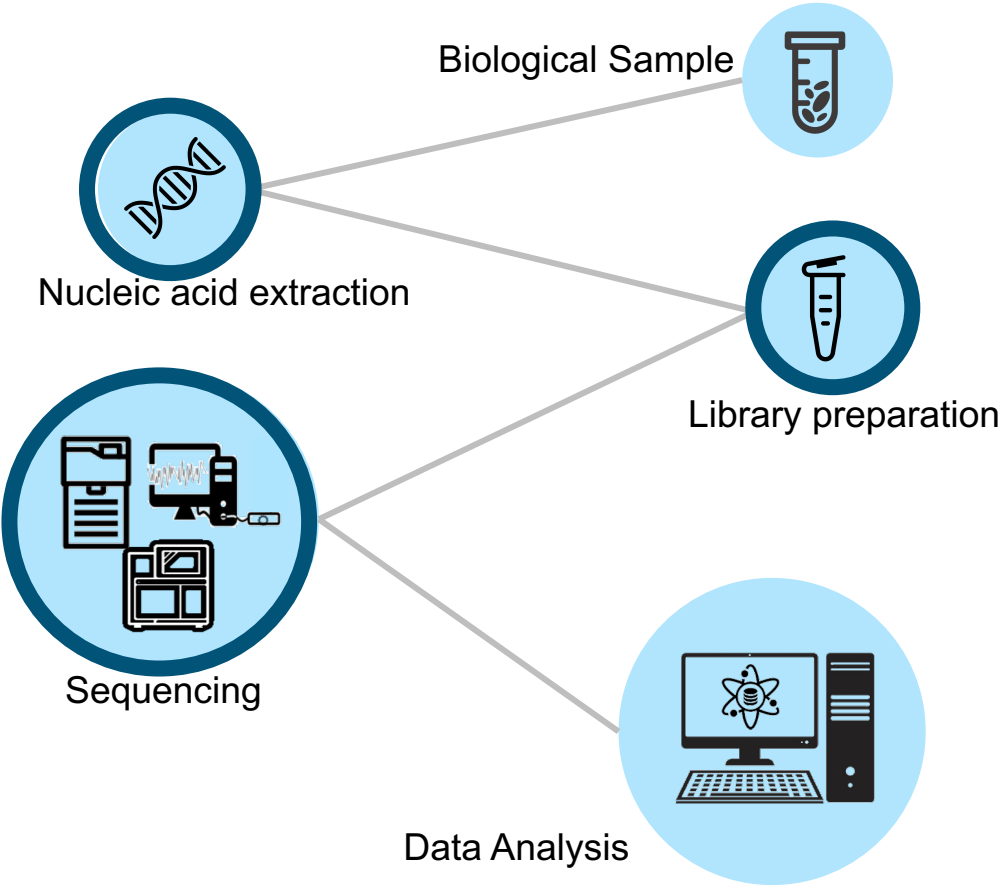
ONT flow cell and kits chemistry

Overview

	Flongle	MinION Mk1B	GridION Mk1	PromethION 24	PromethION 48
Read length	Nanopores read the length of DNA presented to them. Longest read so far: > 4 Mb.				
Number of flow cells per device	1	1	5	24	48
Theoretical maximum output per device, Kit 10/11 chemistry	2.8 Gb	50 Gb	250 Gb	Up to 7 Tb	Up to 14 Tb
Theoretical maximum output per flow cell, Kit 10/11 chemistry	2.8 Gb	50 Gb	50 Gb	Up to 290 Gb	Up to 290 Gb
Theoretical maximum output per device, Kit 12 chemistry	Up to 1.6 Gb	Up to 30 Gb	Up to 150 Gb	Up to 4 Tb	Up to 8 Tb
Theoretical maximum output per flow cell, Kit 12 chemistry	Up to 1.6 Gb	Up to 30 Gb	Up to 30 Gb	Up to 170 Gb	Up to 170 Gb

https://nanoporetech.com/products/comparison

Workflow in sequencing projects



Quality Control (QC) at different steps!

Determine:

- **Quality (accuracy)**
- **Read Length**
- **Yield**

Workflow in sequencing projects

Biological Sample

Nucleic acid extraction

Library prepa

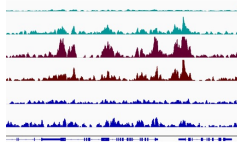
Sequencing

Data Analysis

Information flow of sequencing data

Specific Analyses

Coverage

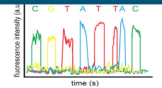


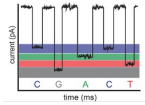
SAM BAM

```
HISEQ2:197:D08GUACXX:8:2105:21056:104282 0 chr10 3000101 255 51M * 0 0
AAGGTCACCAAAGGCCCACCTTGTCTTTACCTTATTTGTTCTA AATTTTTT
=@@DA:ADDHD;AA?:AAFHGIHHBDEFHIDGB9CFH<?F<DEEIG GHEII XA:i:0 MD:Z:51
NM:i:0 HISEQ2:197:D08GUACXX:6:1105:9303:81340 0 chr10 3000301 255 51M * 0 0
GTGTTATTTCACAAGGTGAAGATAGAGCTTGGTGGCTGCCAG AGAGATTAA
BB@FFFFFHHHHHJJJFGJIIJJJJJJJJHIJJJIIJJJJIGIGJJI
XA:i:0 MD:Z:51 NM:i:0 HISEQ2:197:D08GUACXX:7:2102:2396:174630 16 chr10 3000373
255 51M * 0 0 CTGAATCTTCTCCTAAGTATCATCCTGAAGAACAAAATTCCTCT TTTGCTT
JJIJJJJJJJJJJJJJJIJJJJIJJJJJHJJJJJJHHHHHFFFFFCCC XA:i:0 MD:Z:51 NM:i:0
HISEQ2:197:D08GUACXX:8:2108:12162:127556 0 chr10 3000388 255 51M * 0 0
```

Fastq

Raw signal processing ‘Basecalling’






```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:59
GAGAAGTTCAACAGCTGGTATTATTTTTGTTAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:59
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:55
TGGGACTTTATCTGGAGGAGTGTTTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:55
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:19
TGGTTTATGCAGAAATTTCTAGAATAAGGGTAACTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:19
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:71
TCTCAGAAACTTGTTTGTGATGTGTGTATTCAACTA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:71
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:20
TTGATTTAACTCTGACAAAATAAACAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:20
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

Data workflow in PacBio

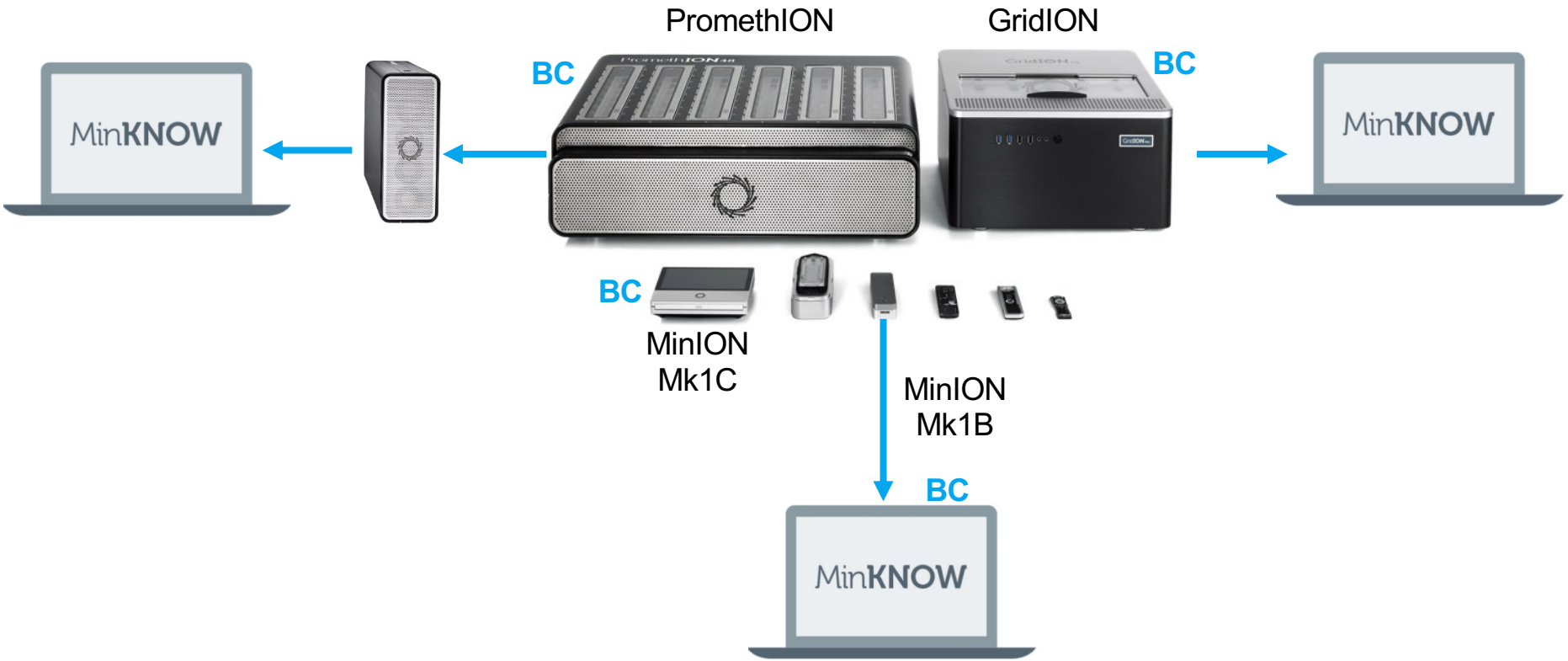
BC

Sequel IIe instrument

SMRT Link Server

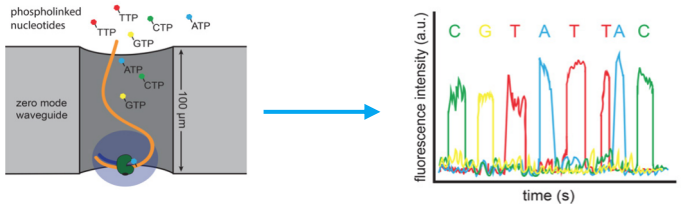
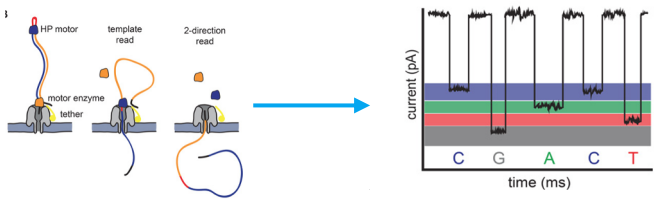
SMRT Link GUI

Data workflow in ONT



Basecalling

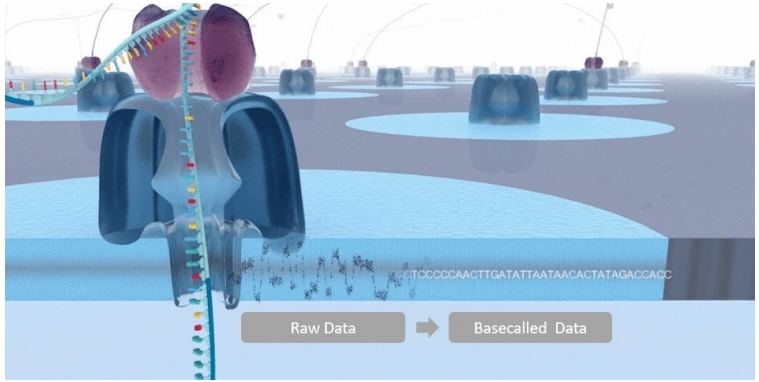
Reuter et al. 2015
Amarasinghe et al. 2020

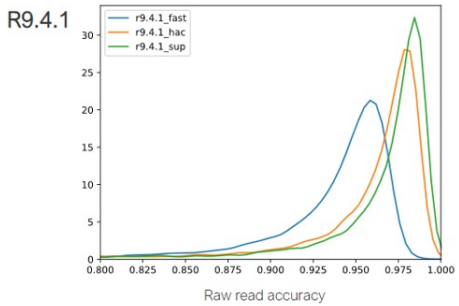
PacBio	***ONT***
	
• Fluorescence flashes are recorded as a movie • Segmenting the fluorescence trace into pulses and converting the pulses into bases.	• Measure the ionic current fluctuations. • The sequence of bases can be inferred from the specific patterns of current variation.
Production basecaller: ***ccs*** (current version: **6.4.0**) *ccs* combines multiple subreads of the same SMRTbell molecule using a statistical model to produce one highly accurate consensus sequence (HiFi reads) along with base quality values.	Production basecaller: **Guppy** (current version: **6.1.2**) Guppy uses neural networks that have been trained on a range of example DNA sequences. The network learns how to translate the series of measurements into the sequence.
SMRT basecallers are mainly developed internally.	**Area of active research**, algorithms are quickly evolving.

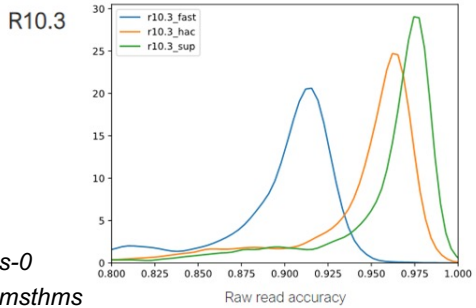
ONT Basecalling - Guppy

Fast, High Accuracy and Super Accurate models

Guppy model	Fast	High Accuracy (HAC)	Super Accurate (sup)
Observations	Designed to keep up with data generation on Oxford Nanopore devices (MinION Mk1C, GridION, PromethION)	Provides a higher raw read accuracy than the Fast model and is currently 5-8 times more computationally-intensive	Has an even higher raw read accuracy, and is ~3 times more intensive than the HAC model.
R9.4.1 modal accuracy	95.8	97.8	98.3
R10.3 modal accuracy	91.4	95.7	97.5
Speed Gbp/hr (GridION)	34	5	2



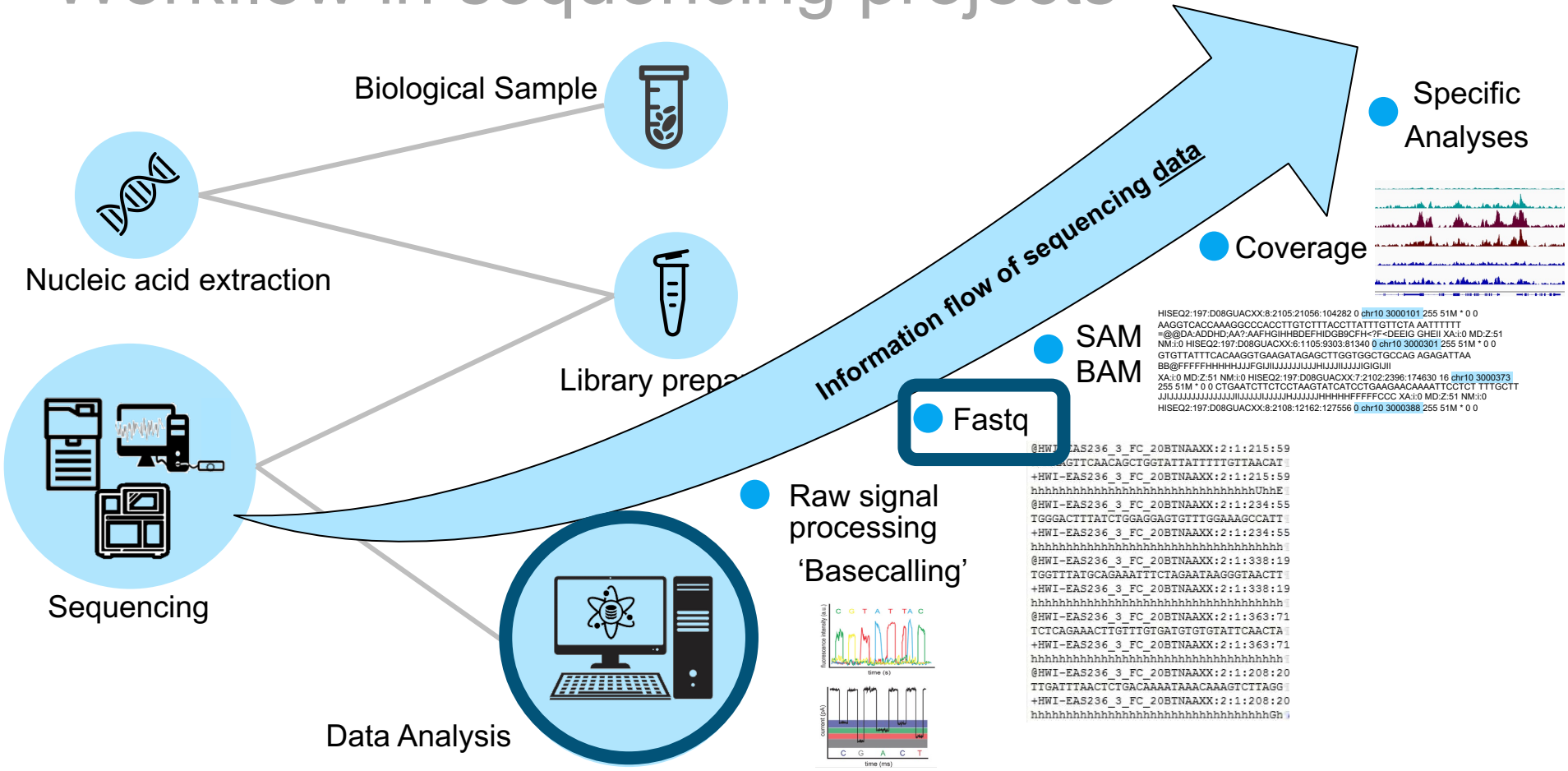




https://nanoporetech.com/about-us/news/oxford-nanopore-tech-update-new-duplex-method-q30-nanopore-single-molecule-reads-0

https://community.nanoporetech.com/technical_documents/data-analysis/v/datd_5000_v1_revo_22aug2016/basecalling-algorithmsthms

Workflow in sequencing projects



ONT output file types

Fast5

Fastq

sequencing_summary.txt

ONT output file types

Fast5

Contain all information needed for **analyzing** nanopore sequencing data and **tracking** it back to its source.

Are the **input of the basecalling software**.

As default, each Fast5 file contain **4000 reads** although this can be configured when starting a run.

The sequencing data is stored in an **HDF5** file, a “container” for storing data of a variety of types in a single file. HDF5 files have a **hierarchical structure** containing “**groups**” which are a bit like directories, and “datasets” which are multidimensional arrays of data elements.

ONT offers two sets of **tools for working with .fast5** files that users may find helpful:

- ont_fast5_api: Provides a simple interface to the .fast5 format, including tools for converting between single- and multi-read formats.
- ont_h5_validator: Provides a tool for validating .fast5 file structures against official Oxford Nanopore Technologies file schemas.

Fastq

sequencing_

summary.txt

ONT output file types

Fast5

Fastq

sequencing_

summary.txt

Display the entire contents of the FAST5 file:

```
$ h5dump FAST5FILE | less
GROUP "/" {
          ATTRIBUTE "file_version" {
                    DATATYPE H5T_IEEE_F64LE
                    DATASPACE SCALAR
                    DATA {
                    (0): 1
                    }
          }
          GROUP "Analyses" {
          GROUP "Basecall_1D_000" {
          ATTRIBUTE "chimaera version" {
```

List all groups inside a FAST5 file:

```
$ h5ls r FAST5FILE | less

/ Group
/Analyses Group
/Analyses/Basecall_1D_000 Group
/Analyses/Basecall_1D_000/BaseCalled_complement Group
/Analyses/Basecall_1D_000/BaseCalled_complement/Events Dataset {678}
/Analyses/Basecall_1D_000/BaseCalled_complement/Fastq Dataset {SCALAR}
/Analyses/Basecall_1D_000/BaseCalled_complement/Model Dataset {4096}
/Analyses/Basecall_1D_000/BaseCalled_template Group
/Analyses/Basecall_1D_000/BaseCalled_template/Events Dataset {657}
```

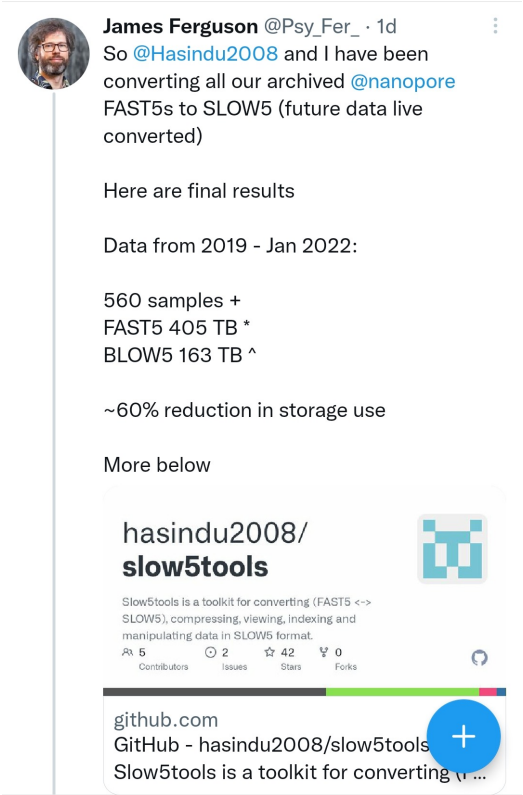
ONT output file types

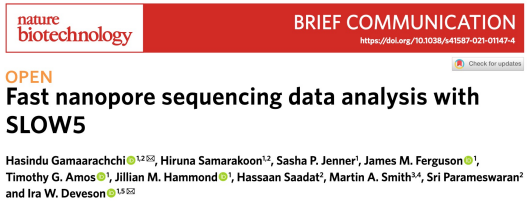
Fast5

Fastq

sequencing_

summary.txt





Slow5tools is a toolkit for converting **(FAST5 <-> SLOW5), compressing, viewing, indexing and manipulating data in SLOW5 format.**

https://hasindu2008.github.io/slow5tools/archive.html

ONT output file types

Fast5

Fastq

sequencing_

summary.txt

ONT output file types

Fast5

Fastq

sequencing_

summary.txt

Text-based format for storing both biological sequences and corresponding quality scores.

A Fastq file uses four lines per sequence:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAA
+SEQ_ID(Optional)
!''*((((***+))%%%++)(%%%%).1**
```

line 1: Sequence ID and Sequence description
line 2: Sequence line e.g. ATCGs
line 3: plus symbol (can additionally have description)
line 4: Sequence line qualities

A Fastq file may contain multiple records. Default is **4000** in a ONT run.

ONT output file types

Fast5

Produced **during base-calling with Guppy** software.

This summary file contains metadata for each sequence read produced during a run.

Fastq

filename_fastq	filename_fast5	read_id	run_id
FAO60506_pass_e2c64338_18.fastq	FAO60506_pass_e2c64338_18.fa	7713567a-585b-4394-a626-c2768c95ea09	e2c64338e3ec82c24e907e4ac77dba8a
FAO60506_pass_e2c64338_32.fastq	FAO60506_pass_e2c64338_32.fa	55ce692a-b84e-493a-b94a-85de4f398ef2	e2c64338e3ec82c24e907e4ac77dba8a
FAO60506_pass_e2c64338_30.fastq	FAO60506_pass_e2c64338_30.fa	f6a02e13-09b3-4f2e-87eb-fd7f98e7e720	e2c64338e3ec82c24e907e4ac77dba8a

channel	mux	start_time	duration	num_events	passes_filter	template_start	num_events	template_du	sequence_length_template
125	2	13536.8655	1143.12325	2286246	TRUE	13536.904	2286169	1143.08475	454884
196	3	26395.8528	1905.0715	3810143	TRUE	26395.88025	3810088	1905.044	738684
181	4	24914.2323	1732.63125	3465262	TRUE	24914.24575	3465235	1732.61775	643839

mean_qscore_template	strand_score_template	median_template	mad_template	pore_type	experiment_id	sample
11.96183	0	69.716446	9.018047	not_set	201119_20-gtct-072_GXB02	GT20-1
12.28224	0	77.000252	9.53832	not_set	201119_20-gtct-072_GXB02	GT20-1
11.243103	0	73.705193	9.191471	not_set	201119_20-gtct-072_GXB02	GT20-1

sequencing_summary.txt

ONT output file sizes

1 Gbase of sequence data ~ 11 Gbytes of storage

- 90% .fast5
- 9% .fastq
- 1% sequence summary file (.txt)

Output (Gbases)	.fast5 storage (Gbytes)	FASTQ storage (Gbytes)	.fast5 + FASTQ storage (Gbytes)
50	500	50	550
100	1000	100	1100
200	2000	200	2200

https://community.nanoporetech.com/requirements_documents/promethion-it-reqs.pdf

PacBio output file types

```
m64039_220208_221544.hifi_reads.bam
m64039_220208_221544.hifi_reads.fasta.gz
m64039_220208_221544.hifi_reads.fastq.gz
m64012_201204_044926.baz2bam_1.log
m64012_201204_044926.ccs.log
m64012_201204_044926.ccs_reports.json
m64012_201204_044926.ccs_reports.txt
m64012_201204_044926.consensusreadset.xml
m64012_201204_044926.reads.bam
m64012_201204_044926.reads.bam.pbi
m64012_201204_044926.sts.xml
m64012_201204_044926.transferdone
m64012_201204_044926.zmw_metrics.json.gz
m64012_201204_044926.subreads.bam
m64012_201204_044926.subreads.bam.pbi
m64012_201204_044926.subreadset.xml
m64012_201204_044926.scraps.bam
m64012_201204_044926.scraps.bam.pbi
```

https://www.pacb.com/wp-content/uploads/Sequel_II_and_IIe_Data_Files.pdf

PacBio output file types

```
m64039_220208_221544.hifi_reads.bam
m64039_220208_221544.hifi_reads.fasta.gz
m64039_220208_221544.hifi_reads.fastq.gz
m64012_201204_044926.baz2bam_1.log
m64012_201204_044926.ccs.log
m64012_201204_044926.ccs_reports.json
m64012_201204_044926.ccs_reports.txt
m64012_201204_044926.consensusreadset.xml
m64012_201204_044926.reads.bam
m64012_201204_044926.reads.bam.pbi
m64012_201204_044926.sts.xml
m64012_201204_044926.transferdone
m64012_201204_044926.zmw_metrics.json.gz
m64012_201204_044926.subreads.bam
m64012_201204_044926.subreads.bam.pbi
m64012_201204_044926.subreadset.xml
m64012_201204_044926.scraps.bam
m64012_201204_044926.scraps.bam.pbi
```

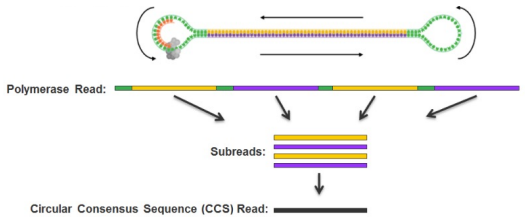
Note: If CCS Analysis is run on the Sequel IIe System, the `subreads.bam`, `scraps.bam` and `scraps.bam.pbi` files are **no longer** generated or available. If CCS Analysis is run in SMRT Link, Sequel IIe System instrument output includes the `subreads.bam` file, and optionally, the `scraps.bam` and `scraps.bam.pbi` files.

`scraps.bam`:

Sequence data **outside of the high-quality region**, rejected subreads, excised adapter and possible barcode sequences, as well as spike-in control sequences.

`subreads.bam`

Are the **native output** data file of the Sequel System and the Sequel II Systems. These are also produced by the Sequel IIe System if users choose to skip on-instrument CCS analysis. Contain the **individual sequencing passes** (subreads) from every productive ZMW. Subreads and HiFi Reads have different error models, and subreads should not be used in HiFi Read applications or vice versa.



https://www.pacb.com/wp-content/uploads/Sequel_II_and_IIe_Data_Files.pdf

PacBio output file types

```
m64039_220208_221544.hifi_reads.bam
m64039_220208_221544.hifi_reads.fasta.gz
m64039_220208_221544.hifi_reads.fastq.gz
m64012_201204_044926.baz2bam_1.log
m64012_201204_044926.ccs.log
m64012_201204_044926.ccs_reports.json
m64012_201204_044926.ccs_reports.txt
m64012_201204_044926.consensusreadset.xml
m64012_201204_044926.reads.bam
m64012_201204_044926.reads.bam.pbi
m64012_201204_044926.sts.xml
m64012_201204_044926.transferdone
m64012_201204_044926.zmw_metrics.json.gz
m64012_201204_044926.subreads.bam
m64012_201204_044926.subreads.bam.pbi
m64012_201204_044926.subreadset.xml
m64012_201204_044926.scraps.bam
m64012_201204_044926.scraps.bam.pbi
```

`reads.bam` files contains one read per productive ZMW and consist of **both HiFi Reads (≥QV 20) and non-HiFi reads (<QV 20)**. It is the native output file of the Sequel IIe System when running on-instrument CCS analysis. A `reads.bam` file is also generated when running CCS analysis in SMRT Link v10.0 or v10.1
~50GB, 5x if kinetic info x5)

`hifi_reads.bam` files contains **PacBio HiFi Reads (≥QV 20**) and can be used directly as input for PacBio and third-party analysis tools designed to work with HiFi Reads.
(Usually <50GB, x 5 if kinetic info)

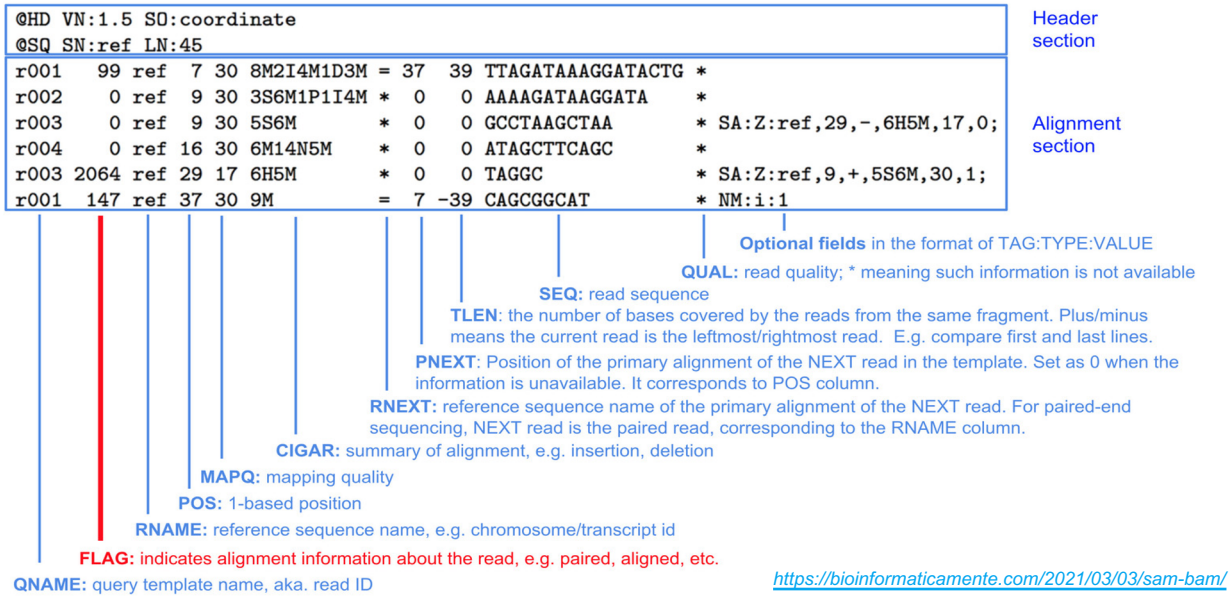
`hifi_reads.fastq.gz` files includes the same reads as `hifi_reads.bam` files, but contain less information about individual reads.
(Usually <50GB, x 5 if kinetic info)

https://www.pacb.com/wp-content/uploads/Sequel_II_and_IIe_Data_Files.pdf

PacBio output file types

BAM

The BAM format is a binary, compressed, record-oriented container format for **raw or aligned** sequence reads. SAM format is a text representation of the same data.



PacBio uses **unaligned BAM** files as the native format to store read information.

PacBio output file types

BAM

The BAM format is a binary, compressed, record-oriented container format for **raw or aligned** sequence reads. SAM format is a text representation of the same data.

```
@HD     VN:1.5  SO:unknown      pb:3.0.1
@RG     ID:3634952b     PL:PACBIO       DS:READTYPE=CCS;BINDINGKIT=101-820-500;SEQUENCINGKIT=101-826-1
00;BASECALLERVERSION=5.0.0;FRAMERATEHZ=100.000000 LB:2-7567597-Oreoseq_pool       PU:m64039_220208_221 544 SM:A
PM:SEQUELII     CM:S/P4.1-C2/5.0-8M
@PG     ID:ccs-6.2.0    PN:ccs  VN:6.2.0        DS:Generate circular consensus sequences (ccs) from su
@PG     ID:samtools     PN:samtools     PP:pbmerge-1.7.0        VN:1.15 CL:/usr/local/bin/samtools vie
m64039_220208_221544/2/ccs 4 * 0 255 * * 0 0 ACACATCTCGTG… QEAGHawxQ~TrmI…
m64039_220208_221544/25/ccs 4 * 0 255 * * 0 0 CCAGCCAACCAGCCAAG S~~S~rOI~~~qkJ~~~Ky
m64039_220208_221544/30/ccs 4 * 0 255 * * 0 0 GTCCCTTTCTATTCCTCTGACCC euYE~muO~pdkl^~}A~K
m64039_220208_221544/32/ccs 4 * 0 255 * * 0 0 CCAGCCAAGCCAGCCAAGCC ~~Xrhc~T~~oH~nU}V~n
m64039_220208_221544/35/ccs 4 * 0 255 * * 0 0 GTCCCTTTCTATTCCTCTGACCC i#jl{BwW9XV~oqH~Y0&e
m64039_220208_221544/43/ccs 4 * 0 255 * * 0 0 CCAGCCAACCAGCCAAG c~X~~S~~~M~~~~~W~~~T~~
m64039_220208_221544/46/ccs 4 * 0 255 * * 0 0 CCAGCCAAGCCAGCCAAGCC ~~Xrhc~T~~oH~nU}V~n
m64039_220208_221544/52/ccs 4 * 0 255 * * 0 0 CCAGCCAAGCCAGCCAGCA M]ta?v^9|SmEx
m64039_220208_221544/54/ccs 4 * 0 255 * * 0 0 GTCCCTTTCTATTCCTCTGACCC euYE~muO~pdkl^~}A~KN
m64039_220208_221544/56/ccs 4 * 0 255 * * 0 0 CCAGCCAAGCCAGCCAAGCC c~X~~S~~~M~~~~~W~~~T~~
m64039_220208_221544/57/ccs 4 * 0 255 * * 0 0 GTCCCTTTCTATTCCTCTGACCC i#jl{BwW9XV~oqH~Y0&e
m64039_220208_221544/66/ccs 4 * 0 255 * * 0 0 CCAGCCAACCAGCCAAG c~X~~S~~~M~~~~~W~~~T~~
m64039_220208_221544/68/ccs 4 * 0 255 * * 0 0 CCAGCCAAGCCAGCCAGCA ~~Xrhc~T~~oH~nU}V~n
m64039_220208_221544/71/ccs 4 * 0 255 * * 0 0 CCAGCCAACCAGCCAAG M]ta?v^9|SmEx
```

PacBio uses **unaligned BAM** files as the native format to store read information.

https://pacbiofileformats.readthedocs.io/en/11.0/BAM.html

Generating HiFi reads from reads.bam

Manually Generating HiFi Reads Files from the Sequel IIe System reads.bam File

If the **Export Reads** analysis application did **not** run automatically, you can run this application manually in SMRT Link. First go to **Data Management > Dataset Details** and click **Analyses > Completed Analyses** to determine if any Export Reads analysis application job has already been completed for your Data Set.

If **no** completed **Export Reads** analysis results are listed, follow the steps below to run the **Export Reads** analysis application for the Data Set of interest:

1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
3. Click **+ Create New Analysis**.
4. Enter a name for the analysis.
5. Select the type of data to use for the analysis: **HiFi Reads**.
6. In the Data Sets table select the Data Set to export to HiFi.
7. Click **Next**.
8. Select the **Export Reads** analysis application from the dropdown list.
9. Fill in the required parameters: **Output FASTA File** (ON or OFF), **Output BAM file** (ON on OFF), **Min CCS Predicted Accuracy** (Default QV 20.)
10. Click **Start**.

Extracting HiFi Reads from the Sequel IIe System reads.bam File Using the Command Line

The **Export Reads** analysis application in the SMRT Link GUI has its command line-version counterpart in our developmental repository at `pbbioconda/GitHub: extracthifi`.

The `extracthifi` tool extracts HiFi Reads (≥ Q20) from the full CCS `reads.bam` output. For more information, see https://github.com/PacificBiosciences/extracthifi/.

To use `extracthifi`, follow the installation instructions in the `pbbioconda` GitHub home page: https://github.com/PacificBiosciences/pbbioconda.

Usage:

```
extracthifi [options] <input.bam> <output.bam>
input.bam   STR  Input CCS BAM.
output.bam  STR  Ouput HiFi BAM.
```

Options:

```
-h,--help         Show this help and exit.
--version         Show application version and exit.
```

https://www.pacb.com/wp-content/uploads/Sequel_II_and_IIe_Data_Files.pdf

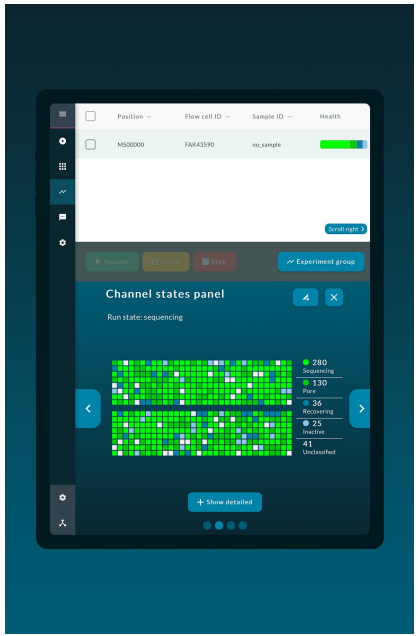
PacBio Yield and file sizes

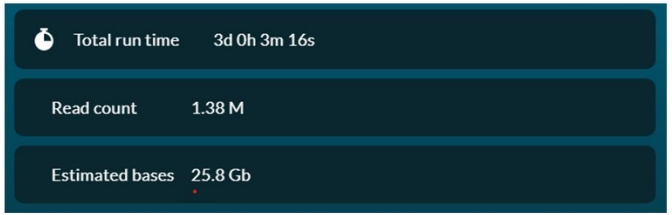
	Sequel IIe system	Sequel II system	Sequel system
Supported SMRT Cell	SMRT Cell 8M	SMRT Cell 8M	SMRT Cell 1M
Number of HiFi reads >99%* accuracy	Up to 4,000,000 **(~ 50Gb)**	Up to 4,000,000 **(~ 50Gb)**	Up to 500,000 **(~ 5Gb)**
Sequencing runtime per SMRT Cell	Up to 30 hrs	Up to 30 hrs	Up to 20 hrs
Instrument control software	v10.1	v10.1	v8.0
SMRT Link	v11.0	v11.0	v10.2
Performance data	Sequel IIe system release	Sequel II system release	Sequel system release

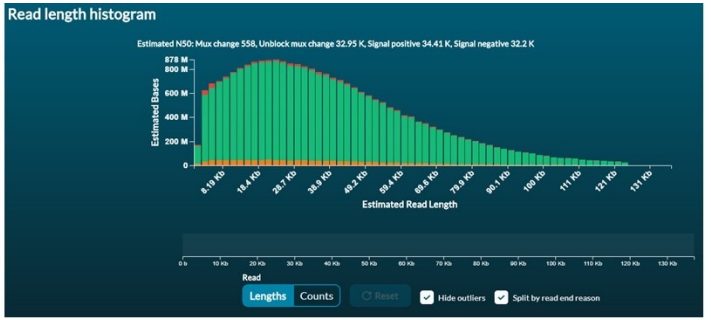
```
$ seqkit stats m64039_220208_221544.hifi_reads.fastq.gz
file                                  format  type   num_seqs        sum_len
m64039_220208_221544.hifi_reads.fastq.gz  FASTQ   DNA   1,928,838  2,839,684,933  (2.8 Gbp)

1.1G m64039_220208_221544.hifi_reads.bam
495M m64039_220208_221544.hifi_reads.fasta.gz
1.1G m64039_220208_221544.hifi_reads.fastq.gz
243G m64039_220208_221544.subreads.bam
```

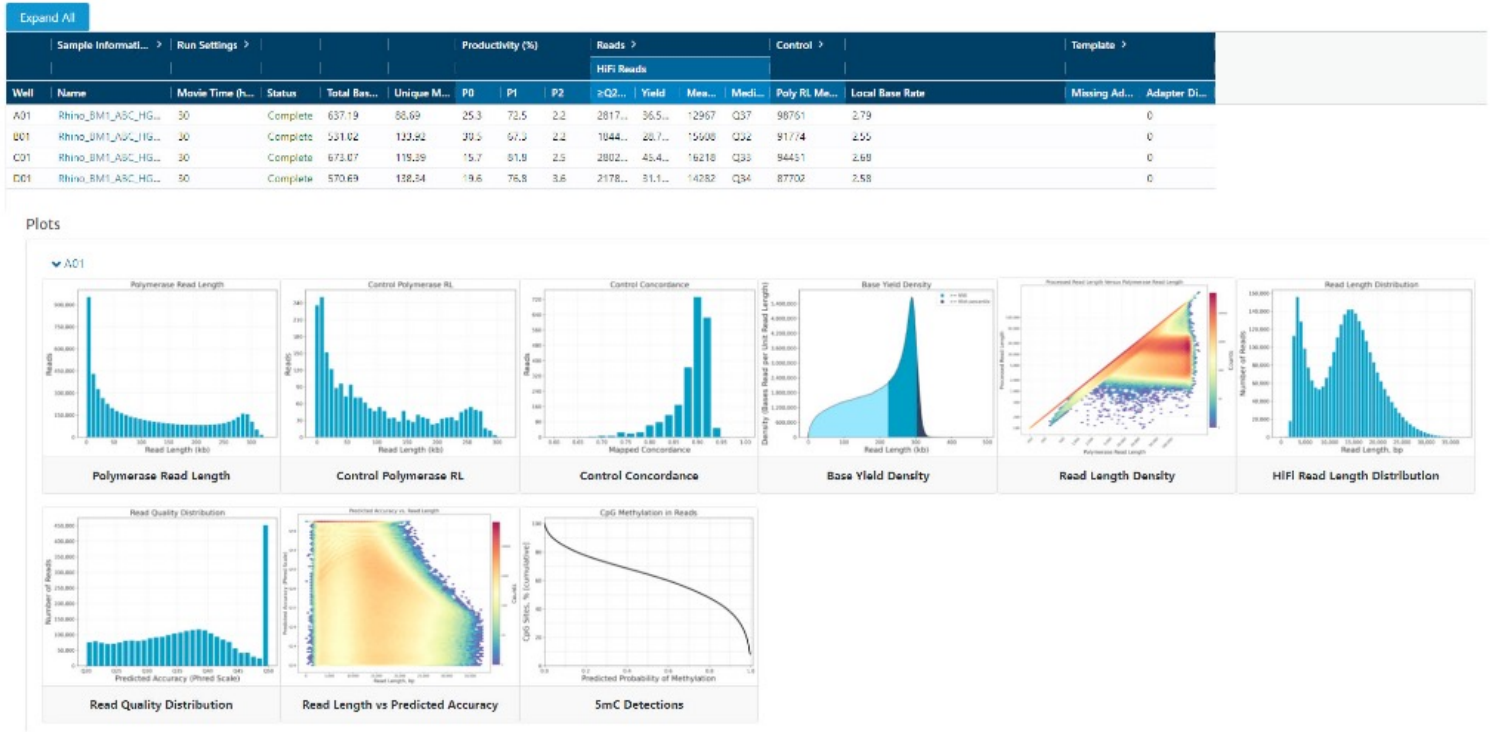
ONT MinKNOW QC







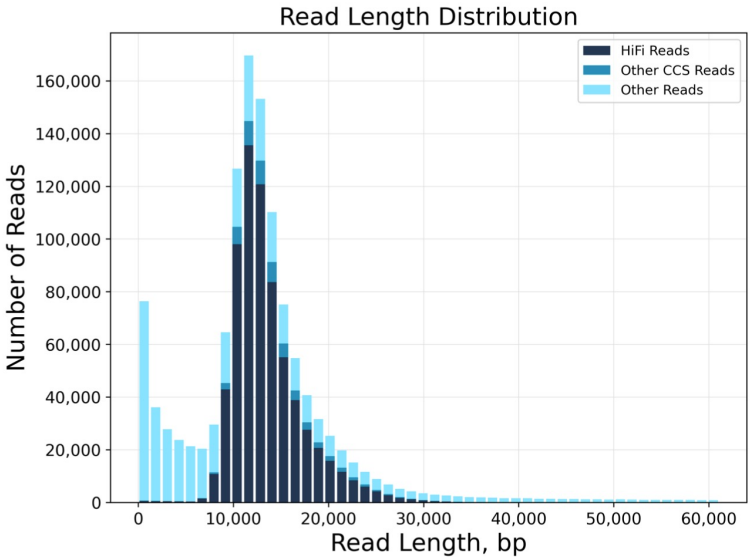
Pacbio SMRT Link: Run QC



https://www.pacb.com/wp-content/uploads/SMRT_Link_User_Guide_v11.0.pdf

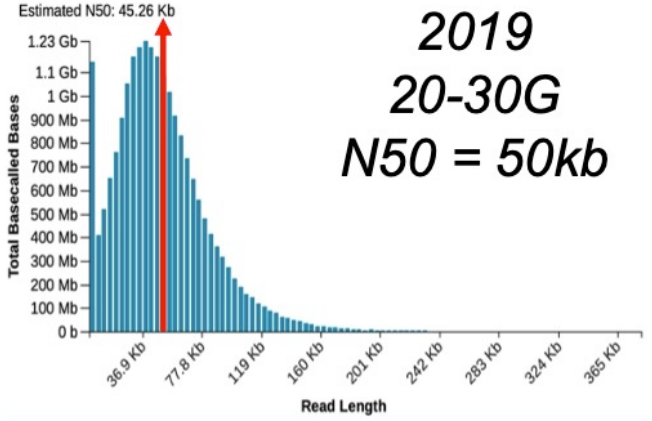
Read Length

PacBio

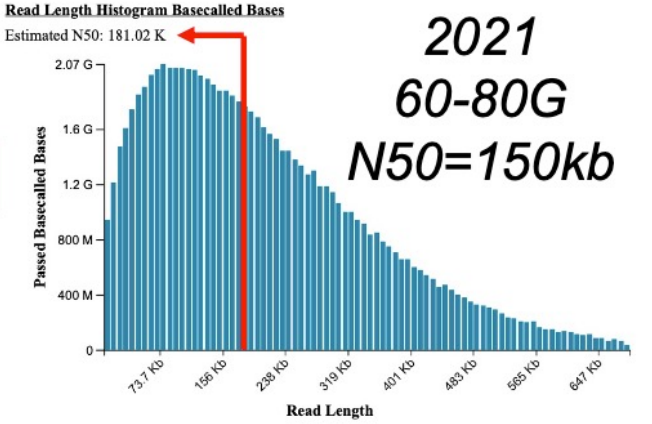


HiFi N50 = 18kb

ONT



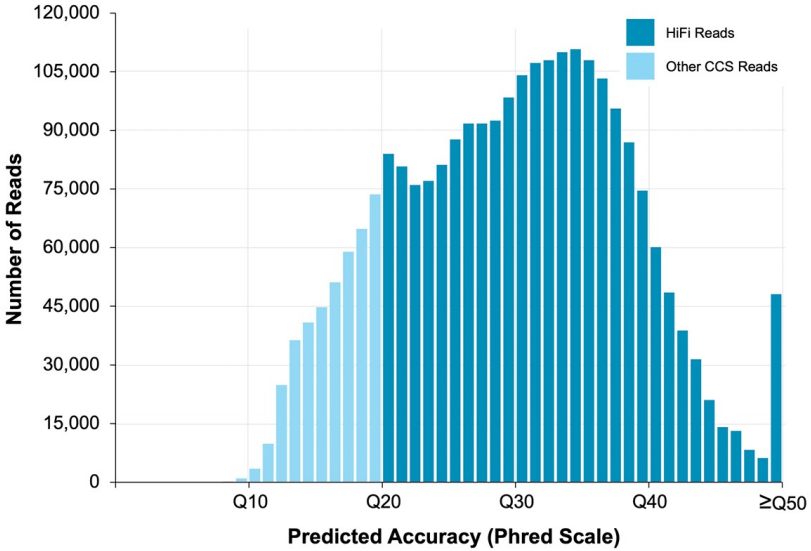
2019
20-30G
N50 = 50kb



2021
60-80G
N50=150kb

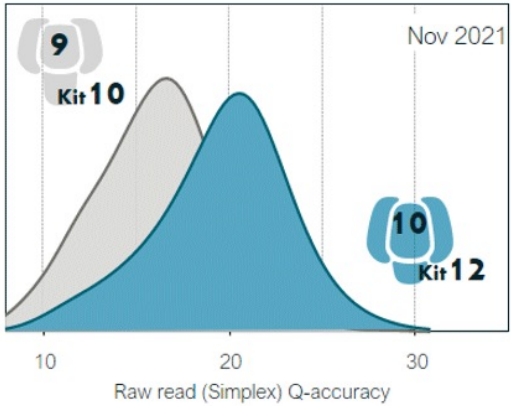
Read Accuracy

PacBio

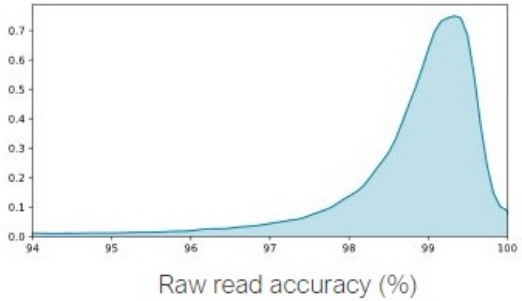


Accuracy mean = 99.96% (Q34)

ONT



Raw read modal 99.3%, >Q20



Duplex (part of Q20+ chemistry)

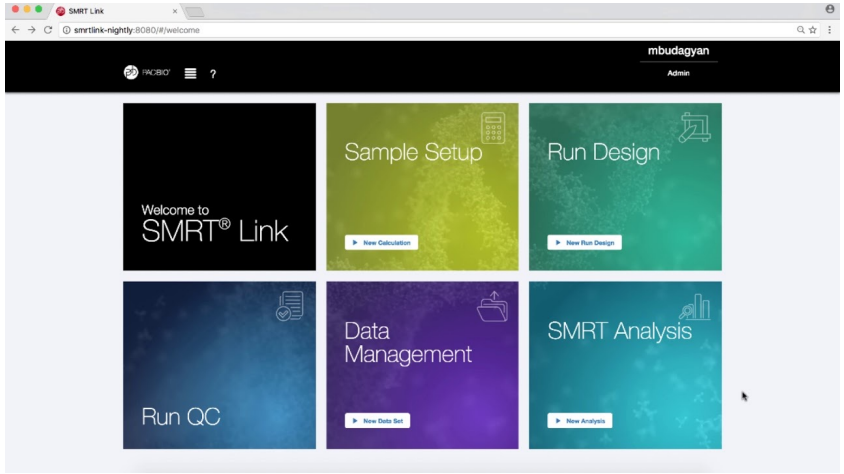
Tutorial:

https://gabyrech.github.io/LongReadsQC/

Table of Contents

Secondary Analyses options offered by…

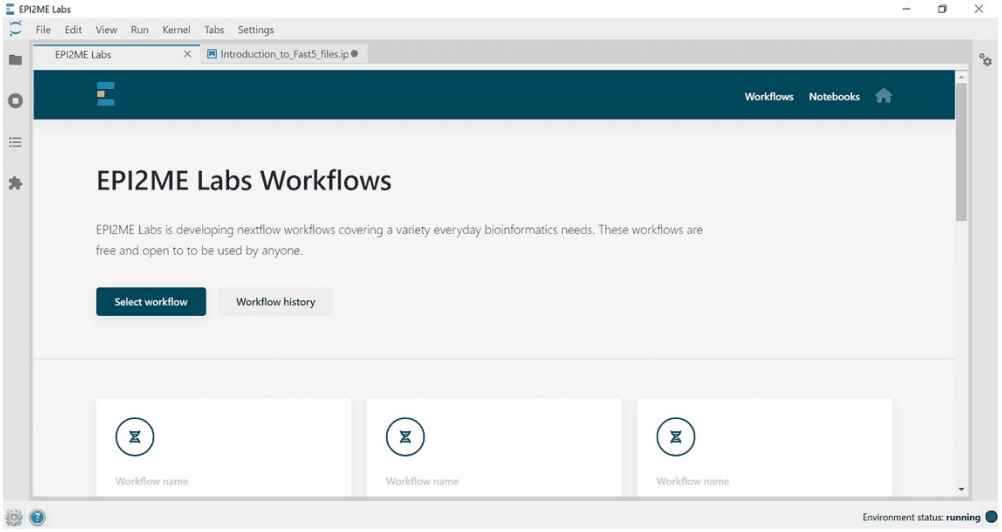
PacBio



SMRT® Link

SMRT Analysis

ONT



EPI2ME

Secondary Analyses options offered by **PacBio**

SMRT Link Modules

SMRT Link includes five easy-to-use modules to guide you from setting up samples for sequencing through data analysis.

Sample Setup

- Generate sample preparation protocols using step-by-step guidelines
- Save and retrieve sample calculations

Run Design

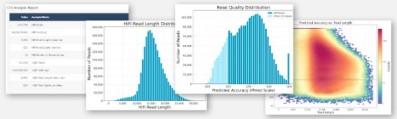
- Design runs for multiple instruments across any of the Sequel System instruments
- Save favorite run parameters for fast and easy setup

Run QC

- Monitor run status in real time
- Obtain key run metrics, including read length and quality, throughput, and loading efficiency

Data Management

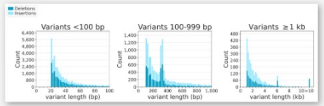
- Organize data into data sets and projects, and generate reports with key QC metrics
- Manage access permissions to projects for SMRT Link users



Data visualizations for rapid interpretation

SMRT Analysis

- Analyze sequencing data generated by one or more sequencing runs
- Use the suite of analysis applications to obtain easy-to-interpret results



SMRT® Analysis interface showing results of a Structural Variant Calling analysis

Application	Analysis Applications	Features
WHOLE GENOME SEQUENCING	**Genome Assembly**	Generates high-quality *de novo* assemblies using HiFi reads
	Microbial Assembly	Generates *de novo* assemblies of small prokaryotic genomes 1.9 Mb – 10 Mb and companion plasmids 2 kb – 220 kb
	Mapping	Aligns sequencing reads to a user-provided reference sequence
VARIANT DETECTION	**Structural Variant Calling**	Identifies indels and structural variants (default: ≥20 bp) in a sample or set of samples relative to a reference
RNA SEQUENCING	**Iso-Seq® Analysis**	Characterizes transcripts and splice variants (*de novo* or reference-based)
COMPLEX POPULATIONS	**Minor Variant Analysis**	Detects, quantifies, and phases minor single nucleotide substitution variants in complex populations
TARGETED SEQUENCING	**Amplicon Analysis**	Identifies phased consensus sequences from a heterogeneous pool of amplicons
EPIGENETICS	**Base Modification Analysis**	Detects common bacterial epigenetic modifications (6mA, 4mC) and optionally analyzes the methyltransferase recognition motifs
Additional Tools	**Circular Consensus Sequencing**	Generates HiFi reads (>Q20) by identifying consensus sequences for single DNA molecules
	Trim gDNA Amplification Adapters and Mark PCR Duplicates	Trims PCR Adapters from a HiFi Reads data set created using an ultra-low DNA sequencing library and removes duplicate reads
	Various utility applications	Demultiplex barcodes, convert BAM to FASTX, export reads

Secondary Analyses options offered by **PacBio**

SMRT Link Modules

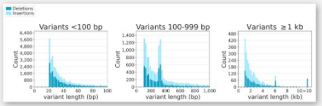
SMRT Link includes five easy-to-use modules to guide you from setting up samples for sequencing through data analysis.

Sample Setup
- Generate sample preparation protocols using step-by-step guidelines
- Save and retrieve sample calculations

Run Design
- Design runs for multiple instruments across any of the Sequel System instruments
- Save favorite run parameters fo

Run QC
- Monitor run status in real time
- Obtain key run metrics, includin loading efficiency

SMRT Analysis
- Use the suite of analysis applications to obtain easy-to-interpret results



SMRT® Analysis interface showing results of a Structural Variant Calling analysis

Application	Analysis Applications	Features
WHOLE GENOME SEQUENCING	Genome Assembly	Generates high-quality *de novo* assemblies using HiFi reads
	Microbial Assembly	Generates *de novo* assemblies of small prokaryotic genomes 1.9 Mb – 10 Mb and companion plasmids 2 kb – 220 kb
	Mapping	Aligns sequencing reads to a user-provided reference sequence
		Identifies indels and structural variants (default: ≥20 bp) in a sample or set of samples relative to a reference
SEQUENCING		Characterizes transcripts and splice variants (*de novo* or reference-based)
Additional Tools	Circular Consensus Sequencing	Generates HiFi reads (>Q20) by identifying consensus sequences for single DNA molecules
	Trim gDNA Amplification Adapters and Mark PCR Duplicates	Trims PCR Adapters from a HiFi Reads data set created using an ultra-low DNA sequencing library and removes duplicate reads
	Various utility applications	Demultiplex barcodes, convert BAM to FASTX, export reads

PacBio command-line tools

https://github.com/PacificBiosciences/pbbioconda

PacBio & BIOCONDA®

https://www.pacb.com/wp-content/uploads/SMRT_Tools_Reference_Guide_v11.0.pdf

Secondary Analyses options offered by **ONT**

MinKNOW

EPI2ME

	EPI2ME	EPI2ME Labs Tutorials	EPI2ME Labs Workflows	Community-developed tools	Research software and custom analysis
Bioinformatic capability needed	● ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ● ○	● ● ● ●
How	Use the cloud-based EPI2ME platform for real-time analysis workflows.	Use EPI2ME Labs for local, post-run analysis and data exploration.	Access EPI2ME Labs Tutorials via simplified, automated Nextflow workflows.	Run open-source tools written and developed by the Nanopore Community.	Access the latest research algorithms from Oxford Nanopore
Where	EPI2ME dashboard ⓘ	EPI2ME Labs Tutorials	EPI2ME Labs Workflows	Community-developed tools	Oxford Nanopore GitHub

Secondary Analyses options offered by **ONT**

Microbial classification

Analysis workflow

- What's In My Pot (FASTQ WIMP)
- FASTQ WIMP – human + viral
- 16S taxonomic classification
- ARMA - antimicrobial resistance
- FASTQ ARTIC + Nextclade

Human genome analysis

- Human Exome
- FASTQ SV caller

Alignment

- Human Alignment GRCh38
- FASTA reference upload
- FASTQ Custom Alignment

Quality control and raw data processing

- Barcoding
- FASTQ DNA Control Experiment
- FASTQ RNA Control Experiment
- Connection Test

EPI2ME

https://nanoporetech.com/nanopore-sequencing-data-analysis

Other resources…

- NanoPlot Online: http://nanoplot.bioinf.be/
- Nexflow: https://github.com/nf-core/nanoseq
- NanoGalaxy: https://nanopore.usegalaxy.eu/
- Long-read tools: https://long-read-tools.org/

Thanks!!!

Check my last publication using long-read sequencing!

nature COMMUNICATIONS

ARTICLE

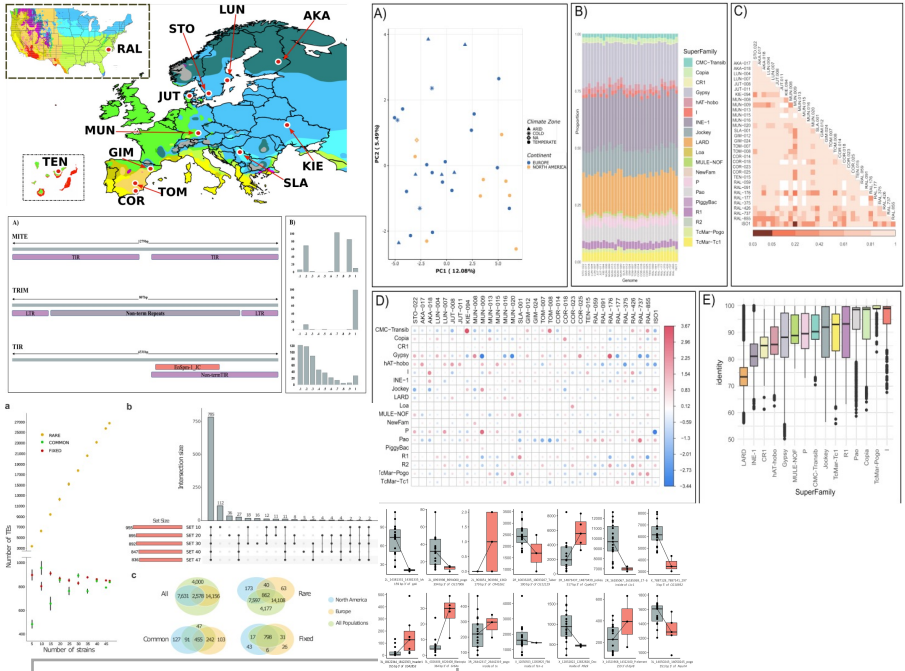
https://doi.org/10.1038/s41467-022-29518-8 OPEN

Check for updates

Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*

Gabriel E. Rech[1], Santiago Radío[1], Sara Guirao-Rico[1], Laura Aguilera[1], Vivien Horvath[1], Llewellyn Green[1], Hannah Lindstadt[1], Véronique Jamilloux[2], Hadi Quesneville[2] & Josefa González[1]✉

https://www.nature.com/articles/s41467-022-29518-8



Reconstruction of new TE families

Evidences of positive selection

Up to **57**% of the insertions identified using long-reads were missed by short-reads methods.

Associations with gene expression

@rechgab