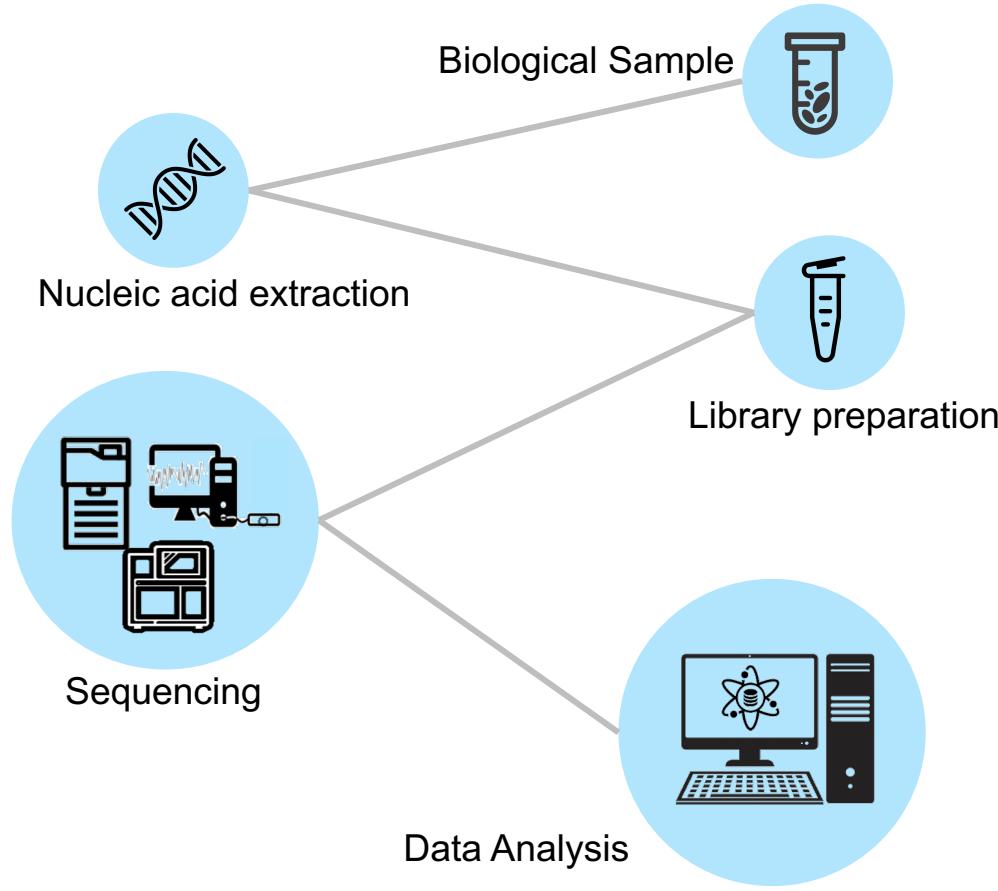


Basic long-read sequencing QC

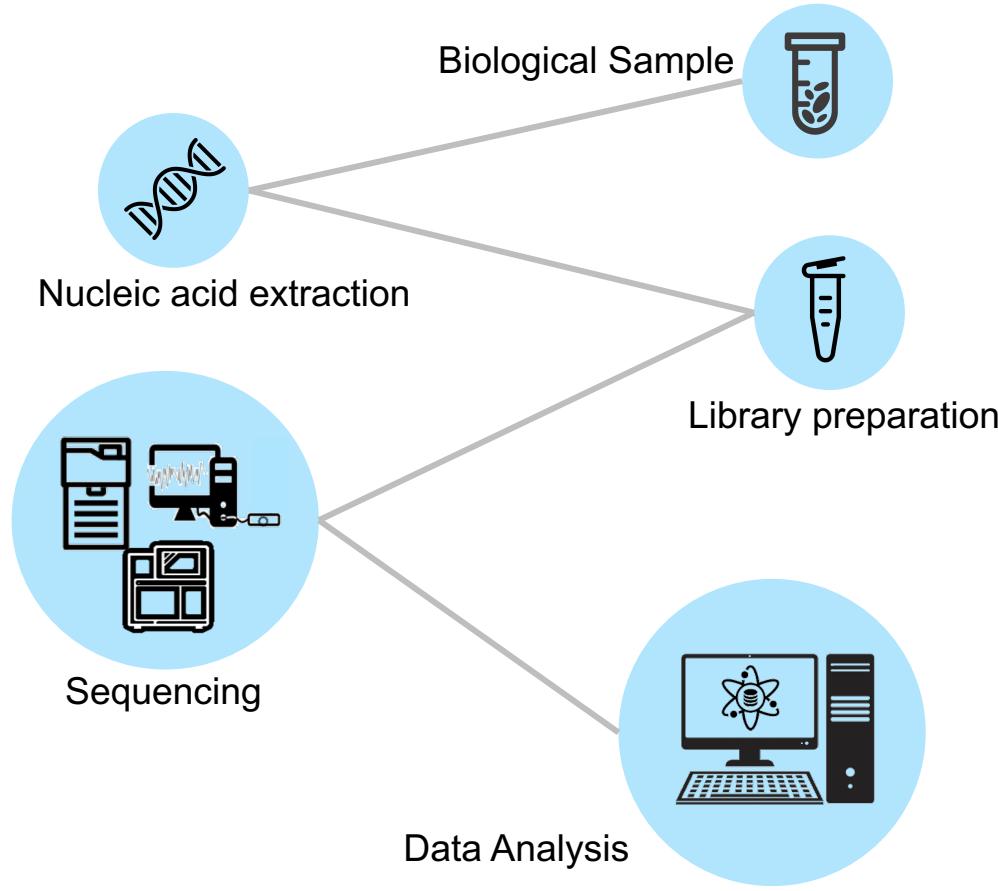
Where do I start? Fundamental analyses of Nanopore and PacBio sequencing data

Gabriel Rech, Ph.D.

Workflow in sequencing projects

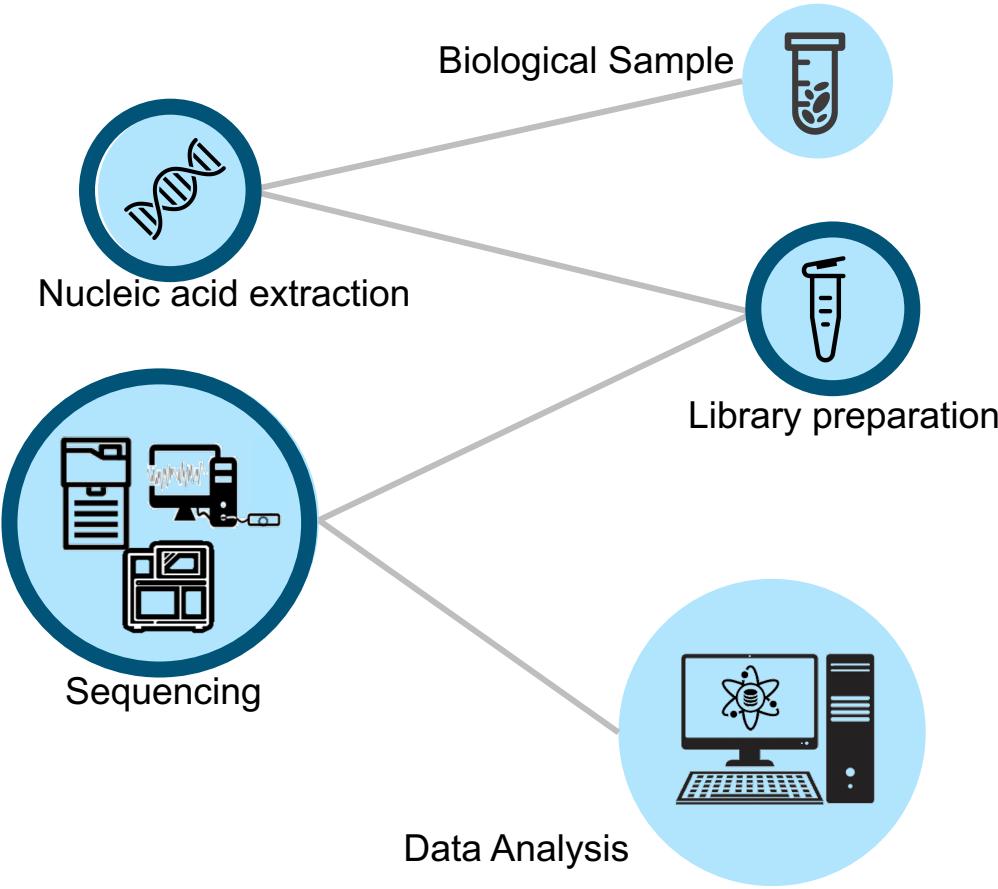


Workflow in sequencing projects



Quality Control (QC) at different steps!

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. $Q = -10 \log_{10} P \quad \rightarrow \quad P = 10^{\frac{-Q}{10}}$

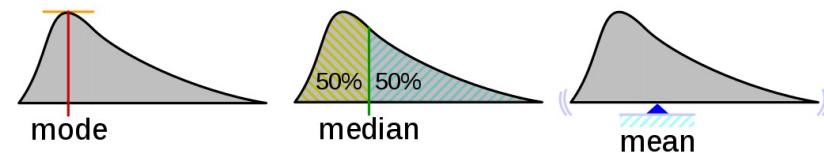
Phred quality scores are logarithmically linked to error probabilities

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

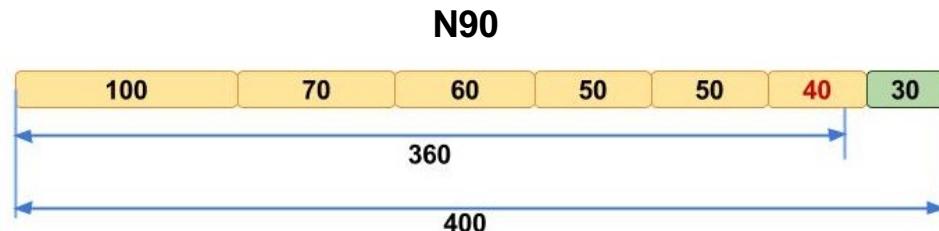
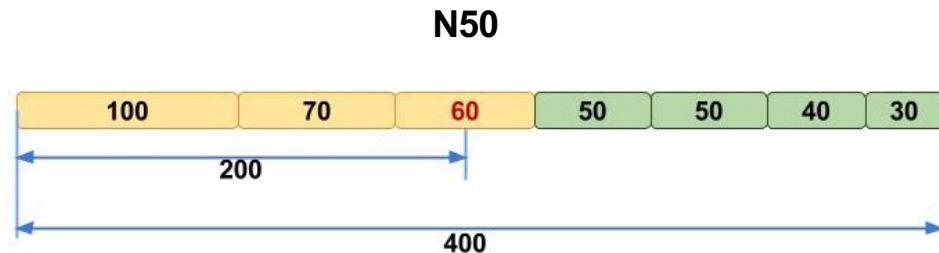
Table 1 ASCII Characters Encoding Q-scores 0-40

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



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*
- **Flow cell version**
 - SMRT Cell 1M vs. 8M
 - R9.4.1 vs R10.4

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



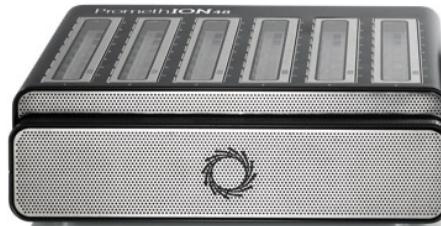
Sequel system

Sequel II system

Sequel Ile system

Oxford Nanopore Technologies (ONT)

PromethION



GridION



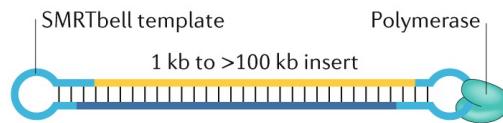
MinION



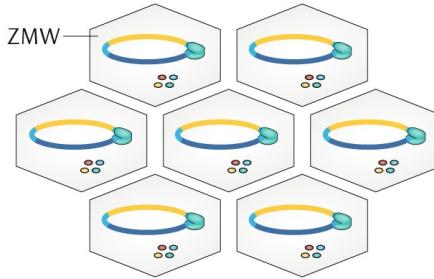
PacBio

a PacBio SMRT sequencing

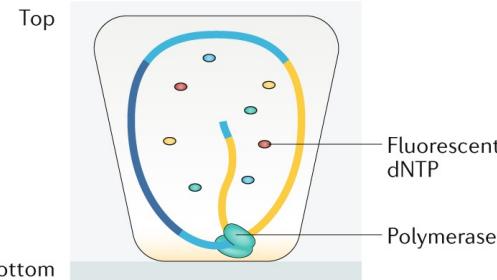
Template topology



Flow cell (top view)



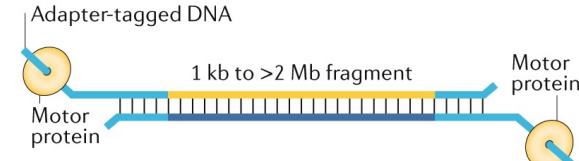
Single ZMW (cross section)



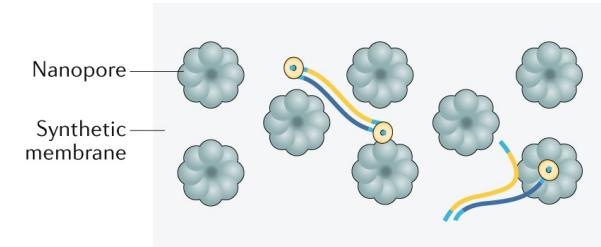
ONT

b ONT sequencing

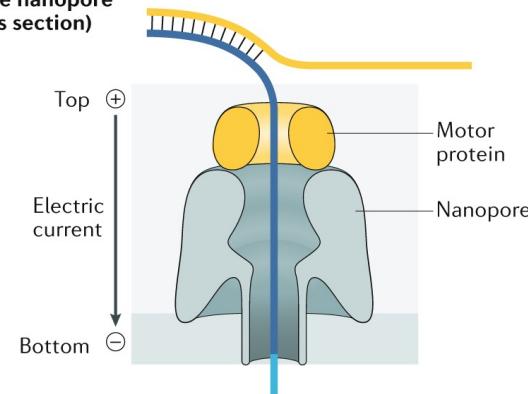
Template topology



Flow cell (top view)

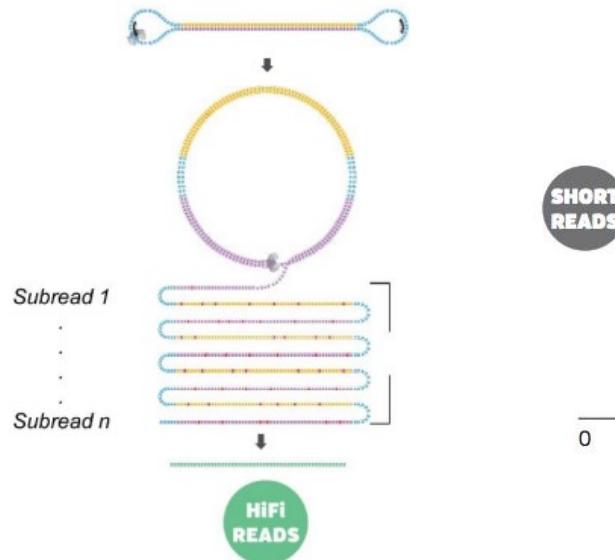


Single nanopore (cross section)

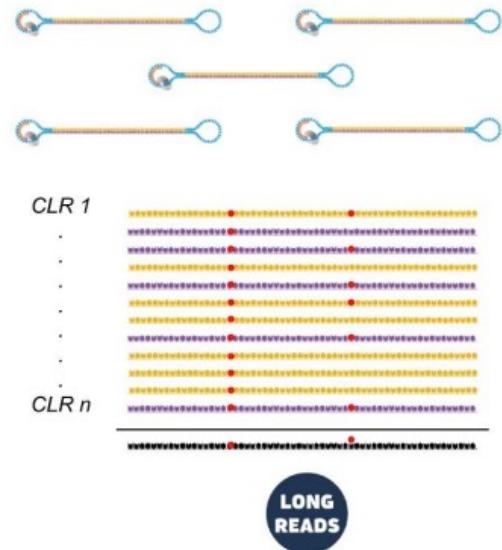
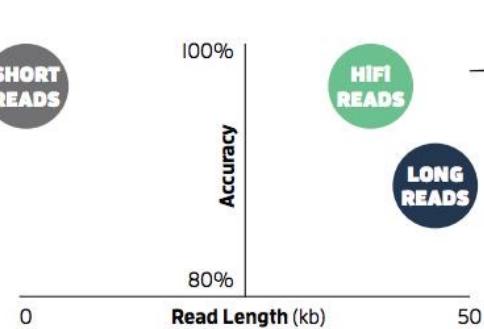


Two modes of PacBio (SMRT) sequencing

Circular Consensus Sequencing (CCS)

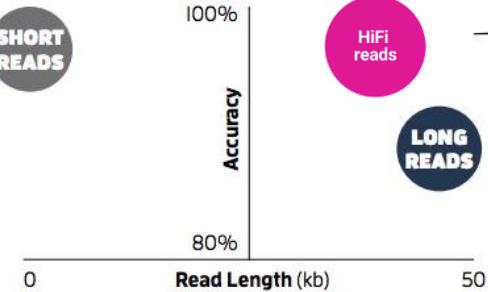
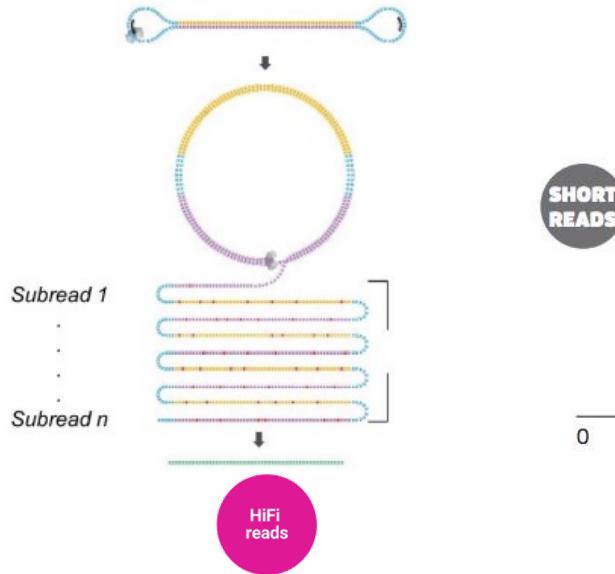


Continuous Long Read (CLR) Sequencing



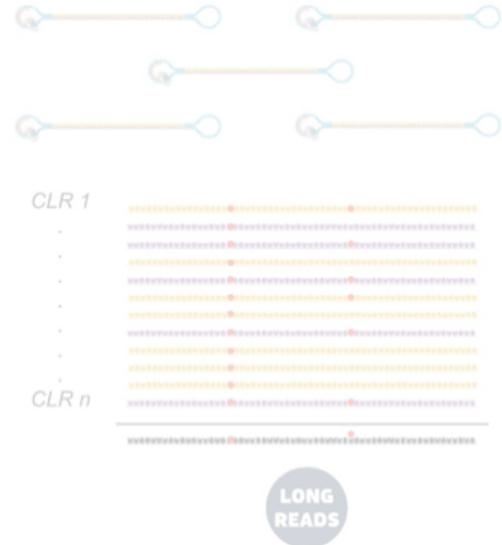
Two modes of PacBio (SMRT) sequencing

Circular Consensus Sequencing (CCS)



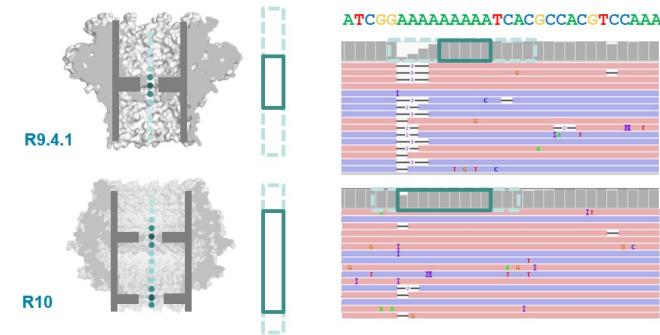
15-20 kb (up to 25 kb)
Accuracy >99.9%

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%

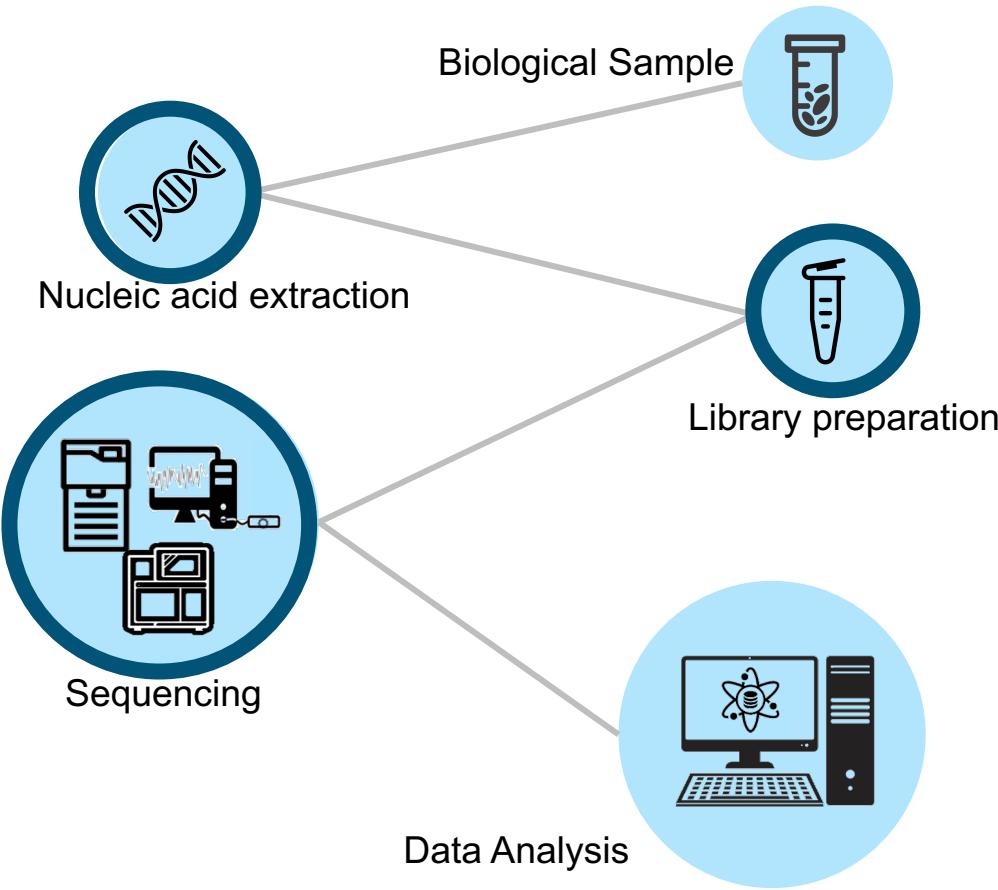
ONT flow cell and kits chemistry

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

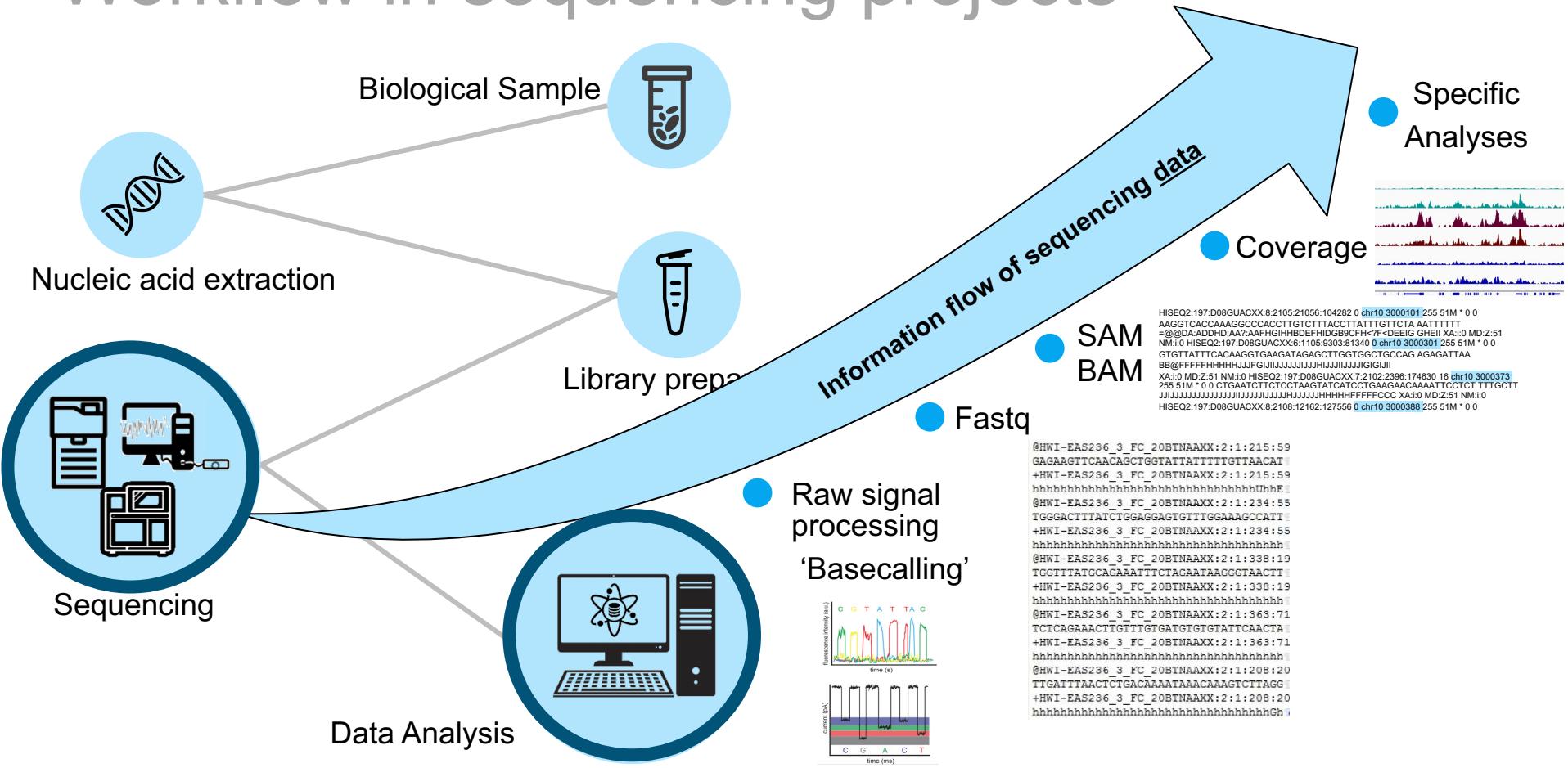


Quality Control (QC) at different steps!

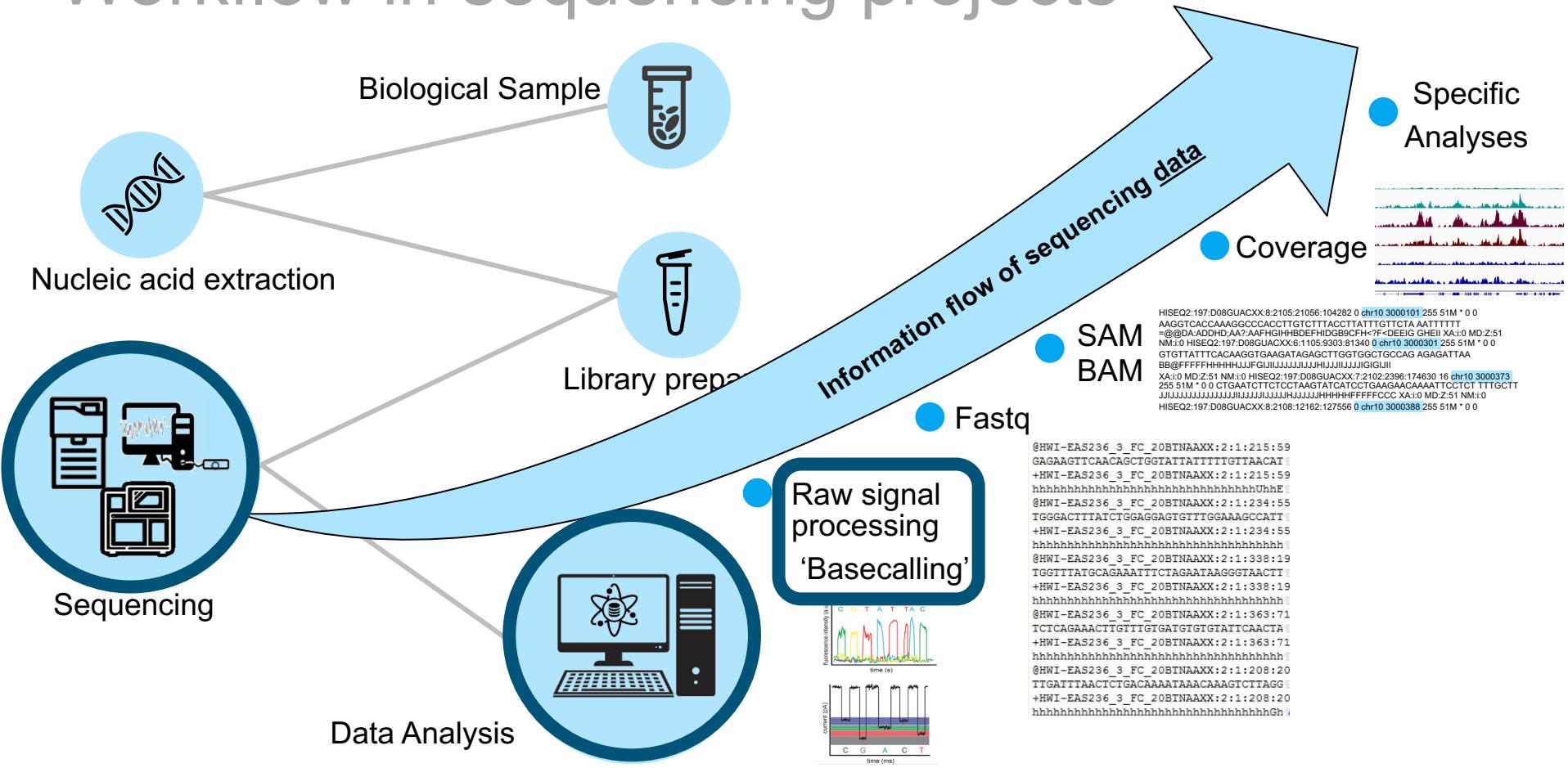
Determine:

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

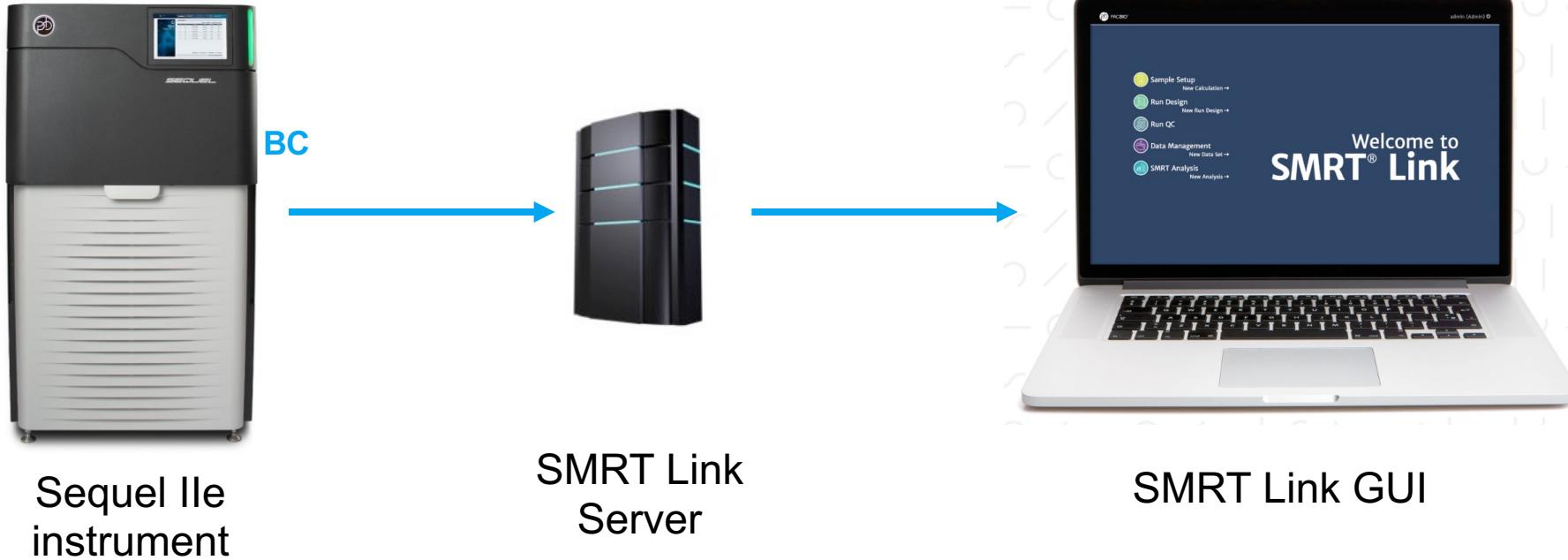
Workflow in sequencing projects



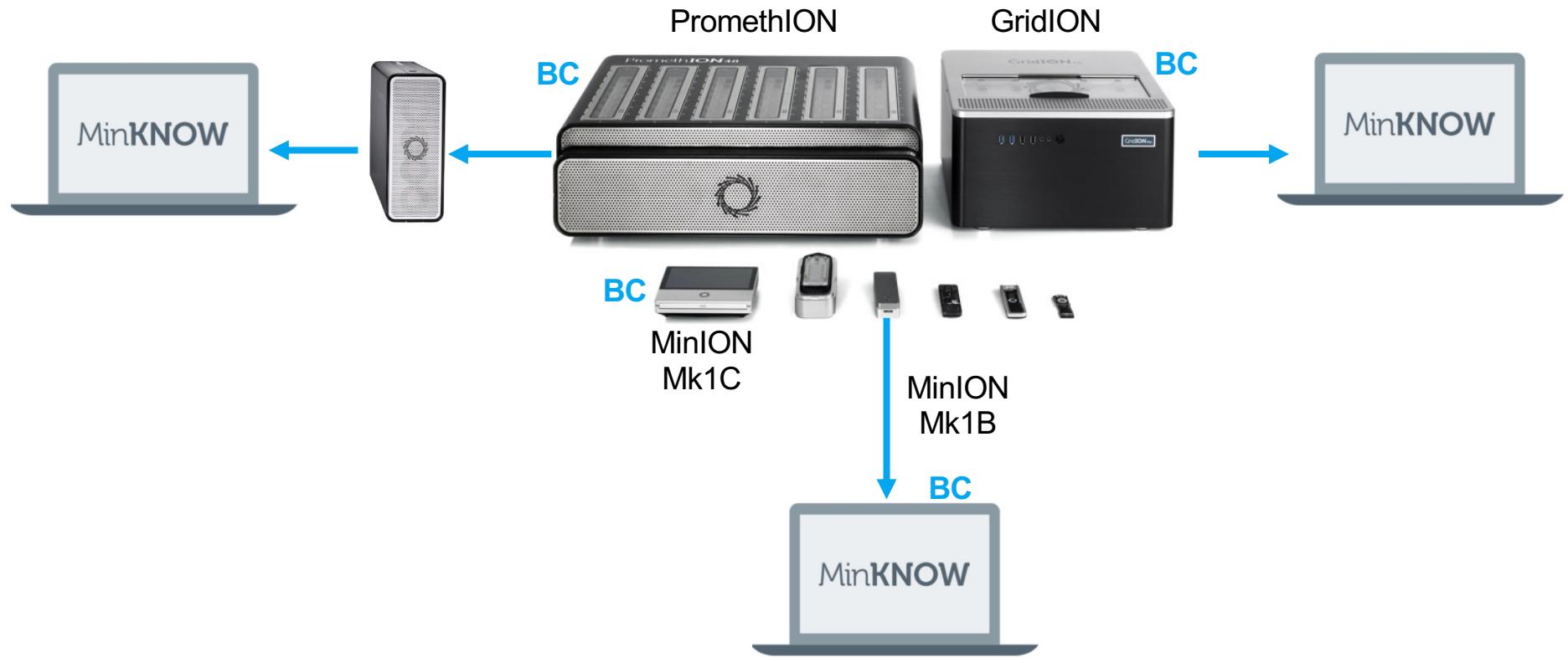
Workflow in sequencing projects



Data workflow in PacBio



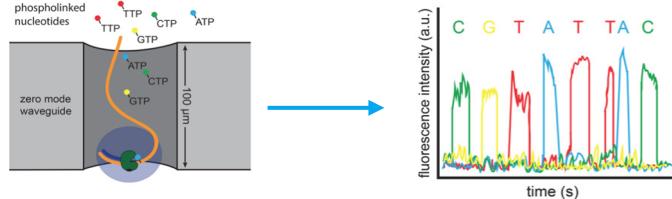
Data workflow in ONT



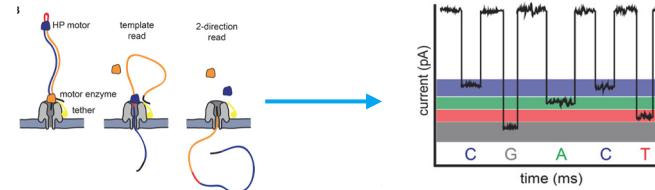
Basecalling

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.

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.

- Measure the ionic current fluctuations.
- The sequence of bases can be inferred from the specific patterns of current variation.

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.

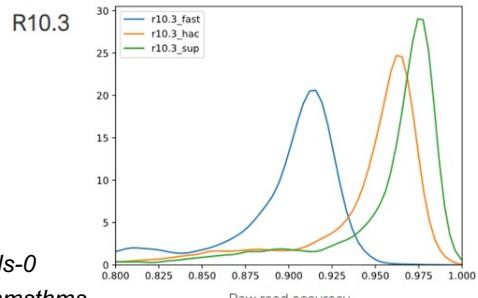
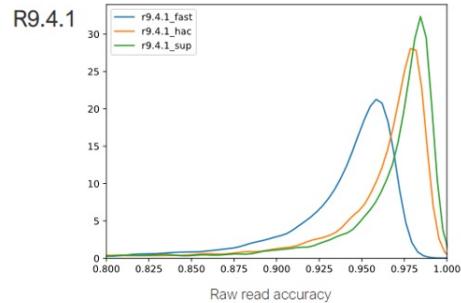
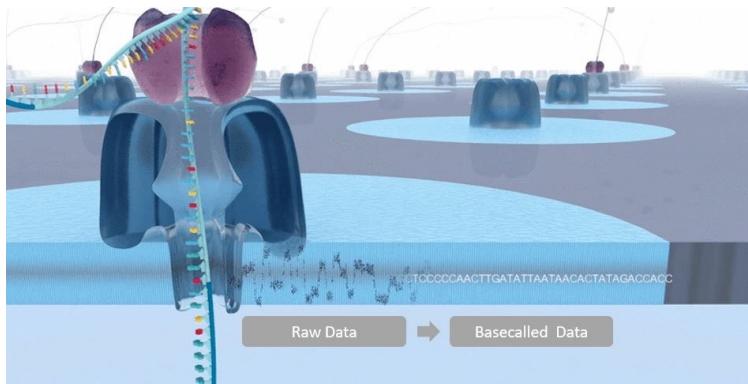
SMRT basecallers are mainly developed internally.

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



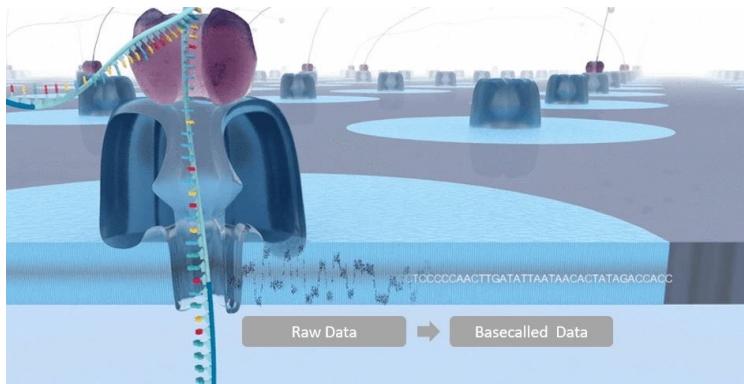
<https://nanoporetech.com/about-us/news/oxford-nanopore-tech-update-new-duplex-method-q30-nanopore-single-molecule-reads-0>

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

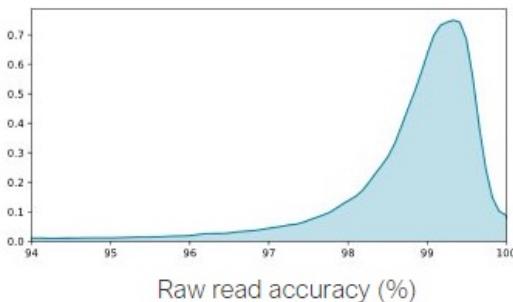
ONT Basecalling - Guppy

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



Raw read modal 99.3%, >Q20
Q20+ chemistry + Guppy sup



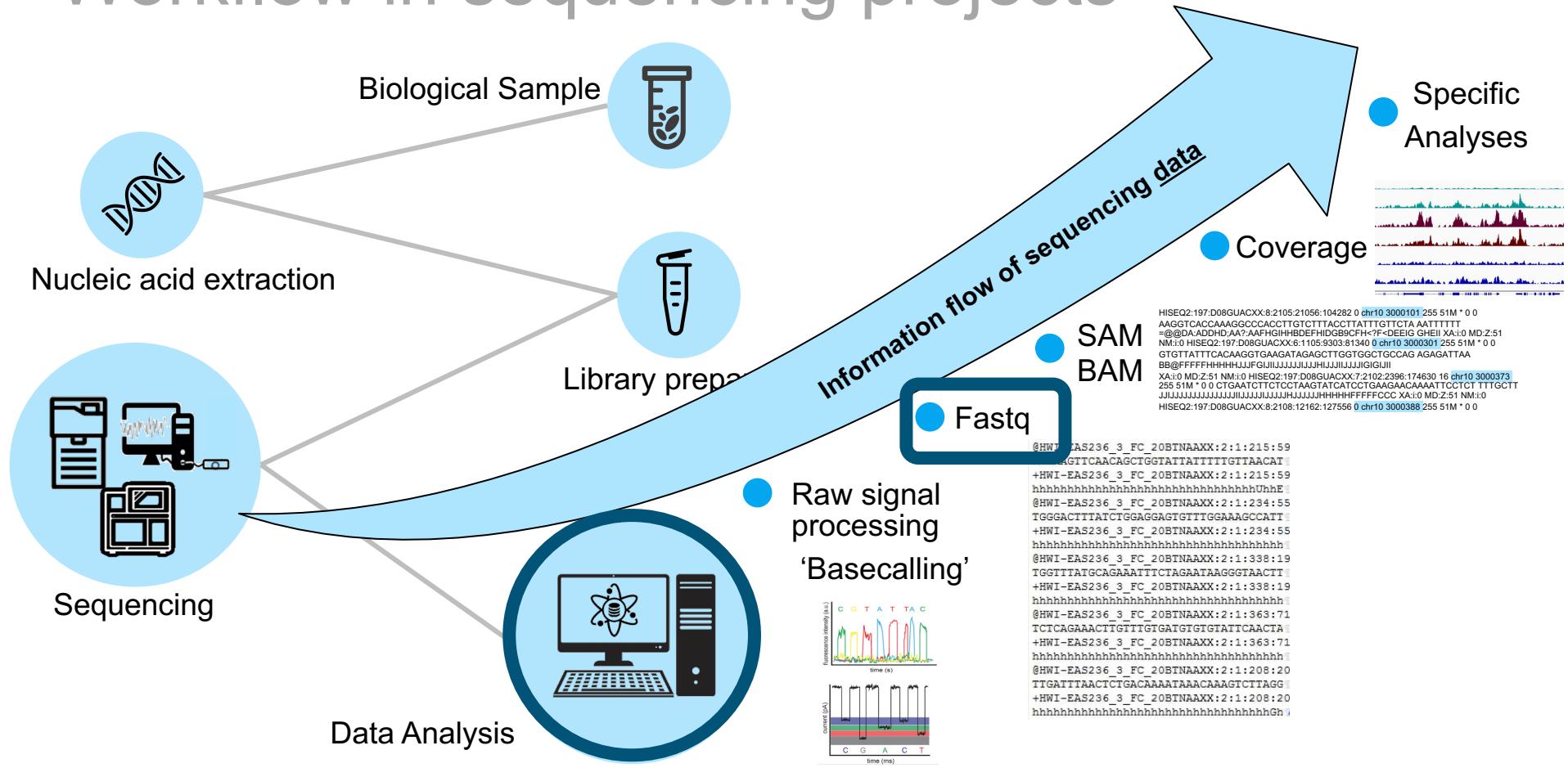
Raw read accuracy (%)

Oxford Nanopore Tech Update
Duplex method trending towards Q30
(May 2021)

<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.

Fastq

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.

sequencing_ summary.txt

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.

ONT output file types

Display the entire contents of the FAST5 file:

Fast5

```
$ 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" {
```

Fastq

List all groups inside a FAST5 file:

sequencing_
summary.txt

```
$ 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



James Ferguson @Psy_Fer_ · 1d
So @Hasindu2008 and I have been
converting all our archived @nanopore
FAST5s to SLOW5 (future data live
converted)

Here are final results

Data from 2019 - Jan 2022:

560 samples +
FAST5 405 TB *
BLOW5 163 TB ^

~60% reduction in storage use

More below

hasindu2008/
slow5tools



Slow5Tools is a toolkit for converting (FAST5 <->
SLOW5), compressing, viewing, indexing and
manipulating data in SLOW5 format.

5 Contributors 2 Issues 42 Stars 0 Forks

github.com
GitHub - hasindu2008/slow5tools
Slow5Tools is a toolkit for converting ...

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

nature
biotechnology

BRIEF COMMUNICATION

<https://doi.org/10.1038/s41587-021-0147-4>



OPEN

Fast nanopore sequencing data analysis with
SLOW5

Hasindu Gamaarachchi^{1,2}✉, Hiruna Samarakoon^{1,2}, Sasha P. Jenner¹, James M. Ferguson³✉,
Timothy G. Amos³, Jillian M. Hammond³, Hassaan Saadat², Martin A. Smith^{1,4}, Sri Parameswaran²
and Ira W. Deveson³^{✉,5}

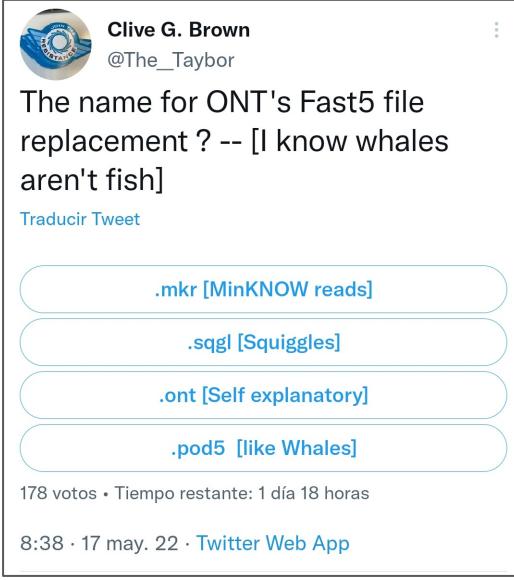
Slow5tools is a toolkit for converting
(FAST5 <-> SLOW5), compressing,
viewing, indexing and manipulating
data in SLOW5 format.

ONT output file types

Fast5

Fastq

sequencing_
summary.txt



A screenshot of a Twitter poll card. At the top is a profile picture of Clive G. Brown (@The_Taylor) with a blue whale logo. Below it is the poll question: "The name for ONT's Fast5 file replacement? -- [I know whales aren't fish]". Below the question are four options in blue text: ".mkr [MinKNOW reads]", ".sqgl [Squiggles]", ".ont [Self explanatory]", and ".pod5 [like Whales]". At the bottom of the card are the poll statistics: "178 votos • Tiempo restante: 1 día 18 horas" and the timestamp "8:38 · 17 may. 22 · Twitter Web App".



A screenshot of a Twitter poll card. At the top is a profile picture of Clive G. Brown (@The_Taylor) with a blue whale logo. Below it is the poll question: "We went for .pod5". Below the question is the poll result: "Clive G. Brown @The_Taylor · May 17 · The name for ONT's Fast5 file replacement? -- [I know whales aren't fish]" followed by a link "Show this poll". At the bottom of the card are the poll statistics: "3", "1", "18", and an upward arrow icon.

ONT output file types

Fast5

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

Fastq

A Fastq file uses four lines per sequence:

```
@SEQ_ID  
GATTGGGGTTCAAAGCAGTATCGATCAAA  
+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

sequencing_
summary.txt

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_duration	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

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.scrapes.bam
m64012_201204_044926.scrapes.bam.pbi
```

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.scrapes.bam  
m64012_201204_044926.scrapes.bam.pbi
```

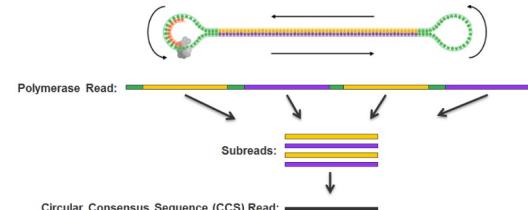
Note: If CCS Analysis is run on the Sequel IIe System, the subreads.bam, scraps.bam and scraps.bam.pbi files are no longer generated or available. If CCS Analysis is run in SMRT Link, Sequel IIe System instrument output includes the subreads.bam file, and optionally, the scraps.bam and scraps.bam.pbi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



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.scrapbs.bam  
m64012_201204_044926.scrapbs.bam.pbi
```

`reads.bam` files contains one read per productive ZMW and consist of **both HiFi Reads (\geq 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 (\geq 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)

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:coordinate @SQ SN:ref LN:45										Header section
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *										Optional fields in the format of TAG:TYPE:VALUE
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *										QUAL: read quality; * meaning such information is not available
r003 0 ref 9 30 5S6M * 0 0 GCCTAAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;										SEQ: read sequence
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *										TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;										PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1										RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.
										CIGAR: summary of alignment, e.g. insertion, deletion
										MAPQ: mapping quality
										POS: 1-based position
										RNAME: reference sequence name, e.g. chromosome/transcript id
										FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.
										QNAME: query template name, aka. read ID

<https://bioinformaticamente.com/2021/03/03/sam-bam/>

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:pmerge-1.7.0 VN:1.15 CL:/usr/local/bin/samtools vie
m64039_220208_221544/2/ccs 4 * 0 255 * * 0 0 ACACATCTCGT... 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 GTCCCTTCTATTCCCTCTGACCC euYE~muO~pdkl^~}A~K
m64039_220208_221544/32/ccs 4 * 0 255 * * 0 0 CCAGCCAAGGCCAGCCAAGCC ~~Xrhc~T~~oH~nU}V~n
m64039_220208_221544/35/ccs 4 * 0 255 * * 0 0 GTCCCTTCTATTCCCTCTGACCC i#j1{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 CCAGCCAAGGCCAGCCAAGCC ~~Xrhc~T~~oH~nU}V~n
m64039_220208_221544/52/ccs 4 * 0 255 * * 0 0 CCAGCCAAGGCCAGCCAAGCA M]ta?v^9|SmEx
m64039_220208_221544/54/ccs 4 * 0 255 * * 0 0 GTCCCTTCTATTCCCTCTGACCC euYE~muO~pdkl^~}A~K
m64039_220208_221544/56/ccs 4 * 0 255 * * 0 0 CCAGCCAAGGCCAGCCAAGCC c~X~~S~~~M~~~~~W~~~T~~
m64039_220208_221544/57/ccs 4 * 0 255 * * 0 0 GTCCCTTCTATTCCCTCTGACCC i#j1{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 CCAGCCAAGGCCAGCCAAGCA ~~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 or 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 (\geq 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  Output HiFi BAM.
```

Options:

<code>-h, --help</code>	Show this help and exit.
<code>--version</code>	Show application version and exit.

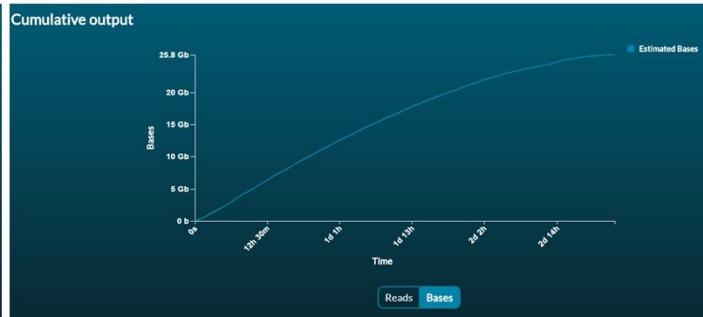
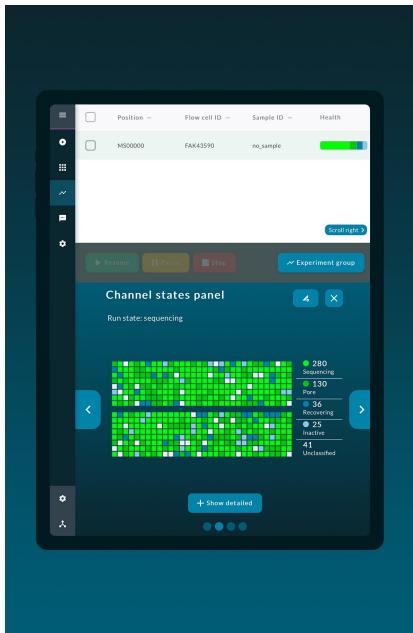
PacBio Yield and file sizes

	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 MinNOW QC



Pacbio SMRT Link: Run QC

Expand All

Well	Name	Movie Time (h..)	Status	Productivity (%)			Reads >		Control >			Template >			
				Total Bas...	Unique M...	P0	P1	P2	≥Q2..	Yield	Mean...	Medi...	Poly RI Me...	Local Base Rate	Missing Ad...
A01	Rhino_BM1_ABC_HG...	30	Complete	637.19	88.69	25.3	73.5	2.2	2817..	365.00	12967	Q37	98761	2.79	0
B01	Rhino_BM1_NSC_HG...	30	Complete	531.02	133.92	39.5	67.0	2.2	1044..	28.7..	15608	Q32	91774	2.55	0
C01	Rhino_BM1_ABC_HG...	30	Complete	673.07	119.39	15.7	61.8	2.5	2802..	45.4..	16218	Q33	94451	2.69	0
D01	Rhino_BM1_ABC_HG...	30	Complete	570.69	138.54	19.6	76.8	3.6	2178..	51.1..	14282	Q34	87702	2.58	0

Plots

▼ A01

Polymerase Read Length

Control Polymerase RL

Control Concordance

Base Yield Density

Normalized Read Length Versus Polymerase Read Length

Read Length Density

HIFI Read Length Distribution

Read Quality Distribution

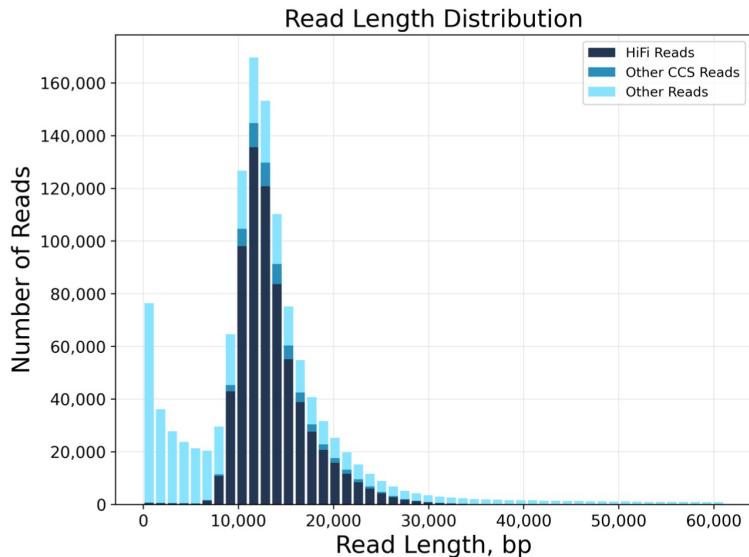
Predicted Accuracy vs. Read Length

CpG Methylation in Reads

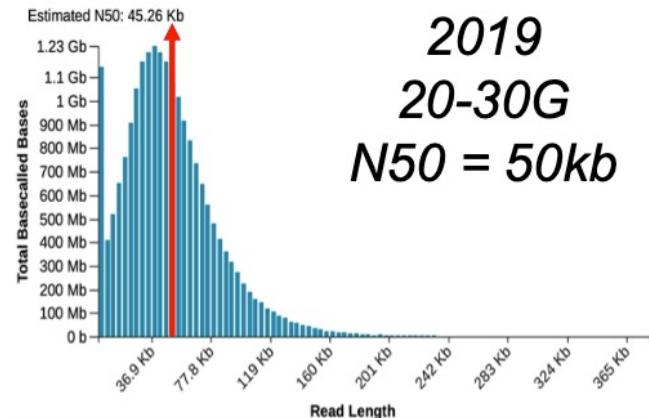
SmC Detections

Read Length

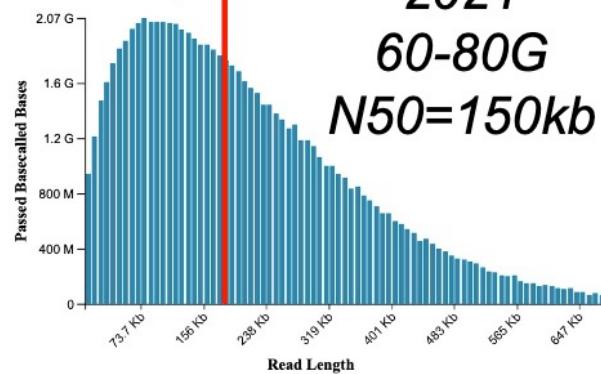
PacBio



ONT
2019
20-30G
N50 = 50kb

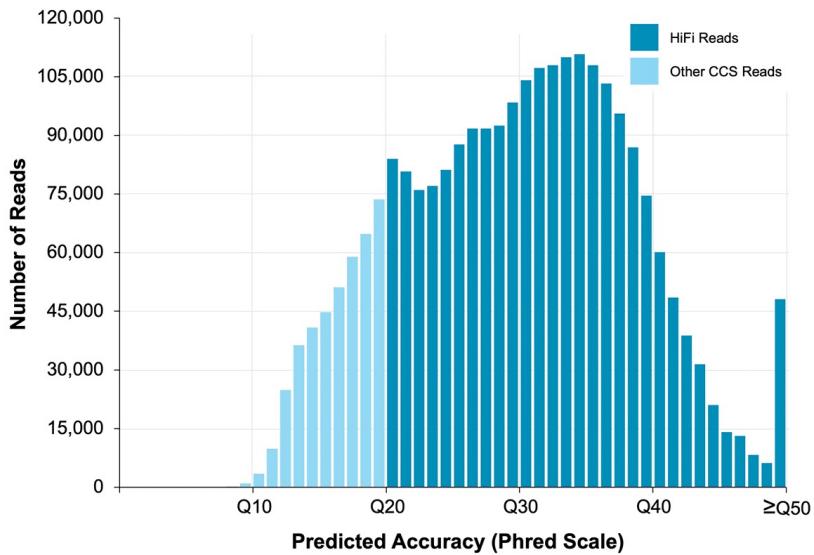


Read Length Histogram Basecalled Bases
Estimated N50: 181.02 K
2021
60-80G
N50=150kb



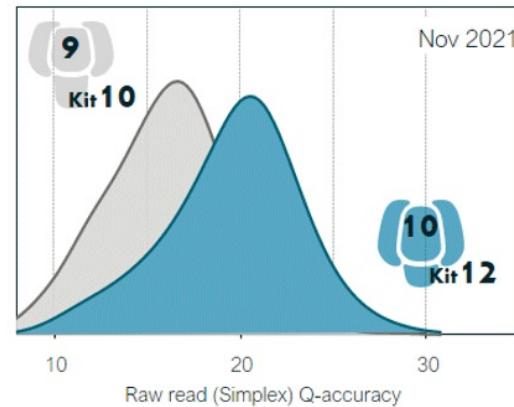
Read Accuracy

PacBio

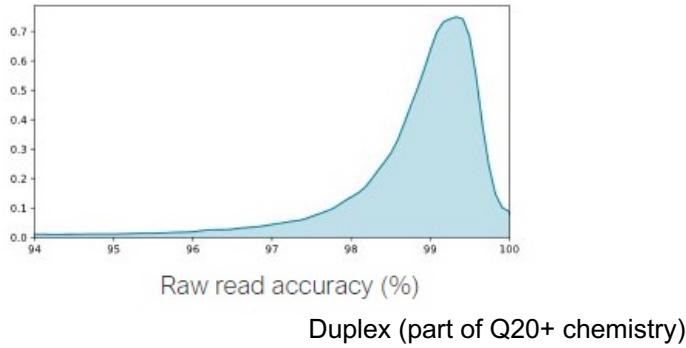


Accuracy mean = 99.96% (Q34)

ONT



Raw read modal 99.3%, >Q20



Duplex (part of Q20+ chemistry)

Tutorial:

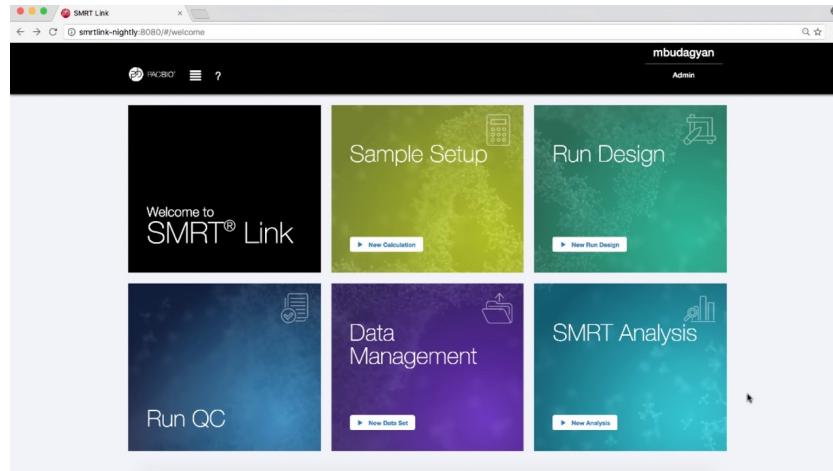
<https://gabyrech.github.io/LongReadsQC/>

Table of Contents

- [Tutorial Set Up](#)
 - [Unix](#)
 - [I don't have Unix, what do I do?](#)
 - [Tools](#)
 - [Data Set](#)
 - [Toy dataset](#)
- [Quick QC using SeqKit](#)
- [Quick QC using SeqFu](#)
- [Comprehensive QC using NanoPack](#)
 - [NanoPlot](#)
 - [NanoComp](#)
- [Comprehensive QC using pycoQC](#)

Secondary Analyses options offered by...

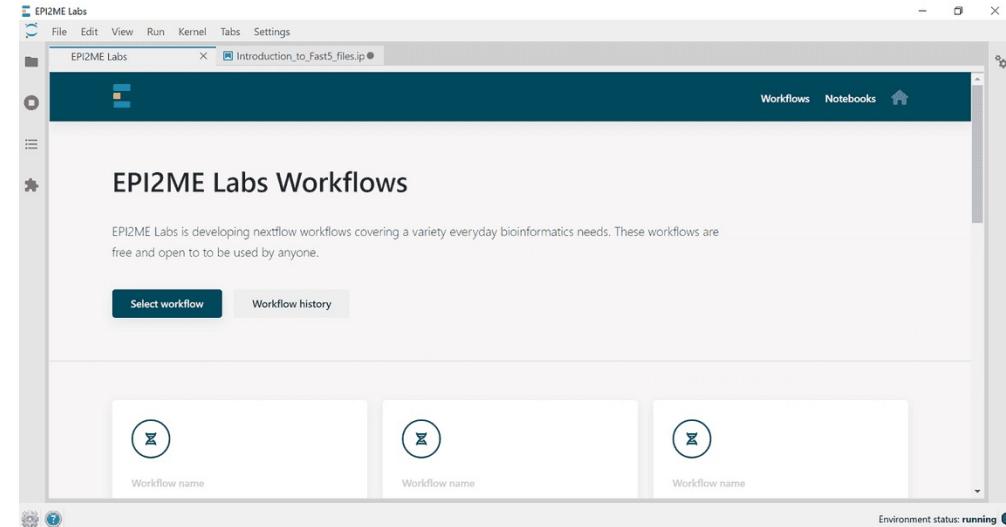
PacBio



SMRT® Link



ONT

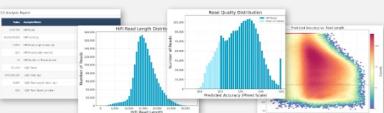
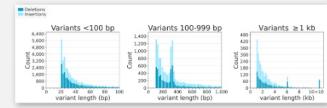


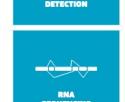
EPI2ME

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.

	<ul style="list-style-type: none"> - Generate sample preparation protocols using step-by-step guidelines - Save and retrieve sample calculations
	<ul style="list-style-type: none"> - Design runs for multiple instruments across any of the Sequel System instruments - Save favorite run parameters for fast and easy setup
	<ul style="list-style-type: none"> - Monitor run status in real time - Obtain key run metrics, including read length and quality, throughput, and loading efficiency
	<ul style="list-style-type: none"> - Organize data into data sets and projects, and generate reports with key QC metrics - Manage access permissions to projects for SMRT Link users <p></p> <p>Data visualizations for rapid interpretation</p>
	<ul style="list-style-type: none"> - Analyze sequencing data generated by one or more sequencing runs - Use the suite of analysis applications to obtain easy-to-interpret results <p></p> <p>SMRT® Analysis interface showing results of a Structural Variant Calling analysis</p>

Application	Analysis Applications	Features
	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
	Structural Variant Calling	Identifies indels and structural variants (default: ≥20 bp) in a sample or set of samples relative to a reference
	Iso-Seq® Analysis	Characterizes transcripts and splice variants (de novo or reference-based)
	Minor Variant Analysis	Detects, quantifies, and phases minor single nucleotide substitution variants in complex populations
	Amplicon Analysis	Identifies phased consensus sequences from a heterogeneous pool of amplicons
	Base Modification Analysis	Detects common bacterial epigenetic modifications (6mA, 4mC) and optionally analyzes the methyltransferase recognition motifs
	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 	<ul style="list-style-type: none">- Generate sample preparation protocols using step-by-step guidelines- Save and retrieve sample calculations
Run Design 	<ul style="list-style-type: none">- Design runs for multiple instruments across any of the Sequel System instruments- Save favorite run parameters for
Run QC 	<ul style="list-style-type: none">- Monitor run status in real time- Obtain key run metrics, including loading efficiency

Application	Analysis Applications	Features
	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
		Characterizes transcripts and splice variants (de novo or reference-based)

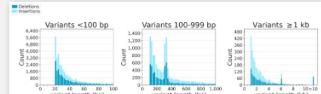
PacBio command-line tools

SEQUENCING

<https://github.com/PacificBiosciences/pbbioconda>

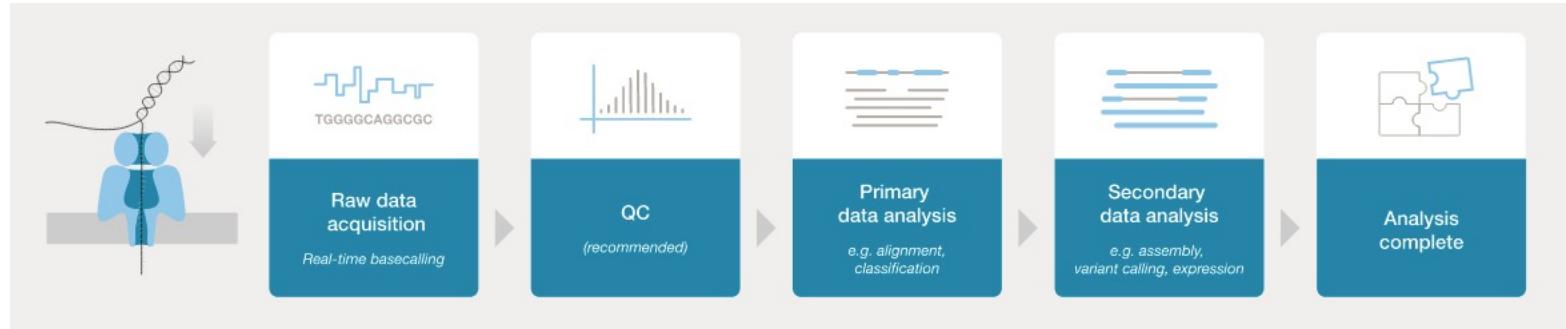
PacBio & BIOCONDA®

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

SMRT Analysis 	<ul style="list-style-type: none">- Use the suite of analysis applications to obtain easy-to-interpret results
 SMRT® Analysis interface showing results of a Structural Variant Calling analysis	

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 ONT



	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 ①	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
\(FASTQ WIMP\)^①](#)

[FASTQ WIMP –
human + viral^①](#)

[16S taxonomic
classification^①](#)

[ARMA - antimicrobial
resistance^①](#)

[FASTQ ARTIC +
Nextclade^①](#)

Human genome analysis

[Human Exome^①](#)

[FASTQ SV caller^①](#)

Alignment

[Human Alignment
GRCh38^①](#)

[FASTA reference
upload^①](#)

[FASTQ Custom
Alignment^①](#)

Quality control and raw data processing

[Barcoding^①](#)

[FASTQ DNA Control
Experiment^①](#)

[FASTQ RNA Control
Experiment^①](#)

[Connection Test^①](#)



EPI2ME

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!



ARTICLE

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

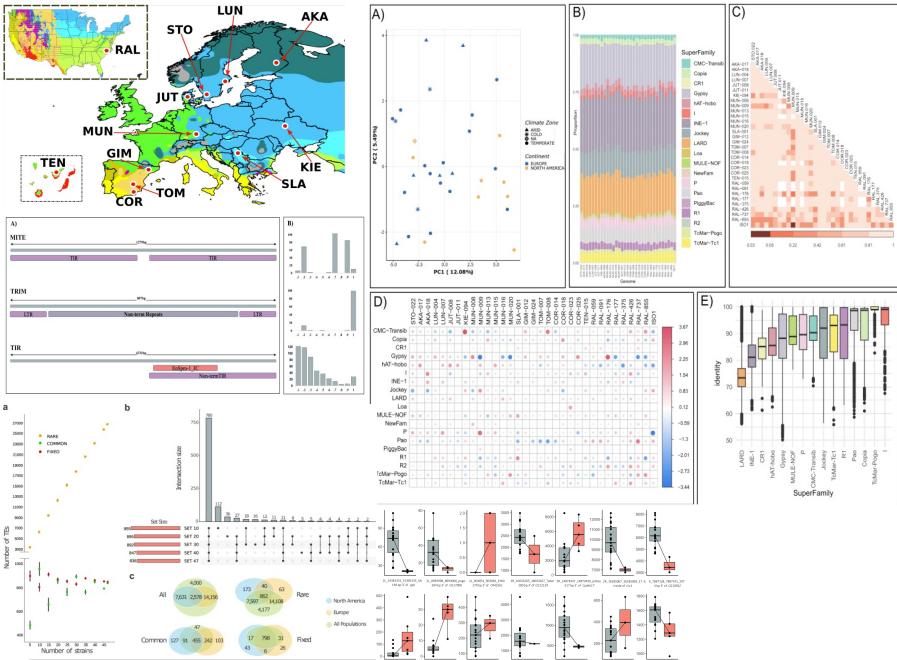
OPEN



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

Gabriel E. Rech¹, Santiago Radio¹, Sara Guirao-Rico¹, Laura Aguilera¹, Vivien Horvath¹, Llewellyn Green¹, Hannah Lindstadt¹, Véronique Jamilloux², Hadi Quesneville² & Josefa González¹

<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 methods were missed by short-reads methods.

Associations with gene expression