From sequencer to polished reads for bacteria

September 2023

General information

- 2 sessions - 9-13 GMT+2
  - 12th of September
  - 14th of September
- Sessions will be recorded!
- Q&A bottom to ask questions
- Please take your time to evaluate on EVA
- Two tech experts
  - Rodrigo

Overall objectives

- Inspect the raw fastq files from the sequencer
- Understand how sequences can differ based on the preparation methodology
- Explain the differences between sequencing reads from Illumina and Nanopore
- Assess the quality of sequence data and trim low quality data
- Learn when to decontaminate and when to re-sequence
- Generate an overview of an entire dataset using different tools

Course intro

Day 1
Sep 12th

Theory
- Course intro
- Illumina sequencing theory
- Nanopore sequencing theory
- Sequencing technique comparison

Practical
- Practical 1 intro
- First look at dataset
- Contamination control
- Raw read QC

Day 2
Sep 14th

Interactive
- Session 1 recap
- Interpret raw read QC results
- Interpret contamination control results

Practical
- Short-read trimming
- Long-read trimming
- Decontamination
- QC parameters for a whole dataset
- Course summary

Presenters

Kasper Thystrup Karstensen

Education:

- Bachelor's in Medical Laboratory Technology
- Master's in Bioinformatics and Systems Biology

Experience:

- Background in RNA-related work
- Employed at SSI since January 2022, specializing in antimicrobial resistance surveillance

Astrid Rasmussen

Education:

- Bachelor's in Biology
- Master's in Bioinformatics and Systems Biology

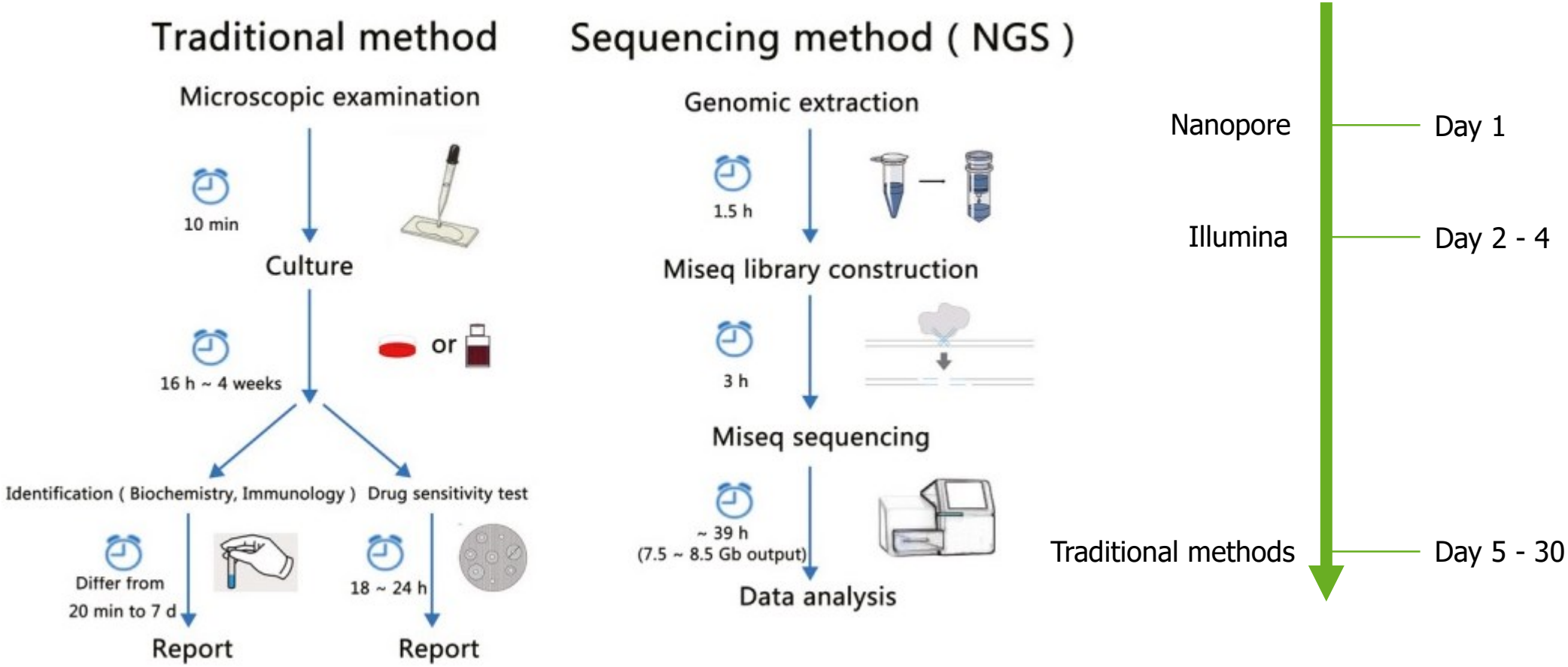
Experience:

- Employed at SSI since January 2022, specializing in antimicrobial resistance surveillance and Nanopore sequencing

How about you?

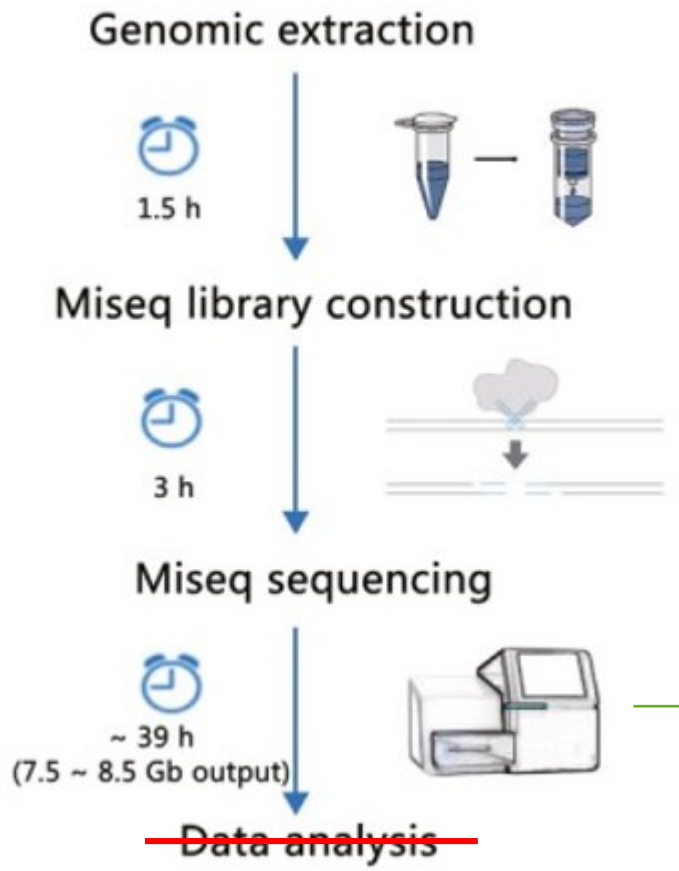
Please have a phone or another browser window at hand…

Sequencing for bacterial typing – A perspective from a clinical setting



Yu *et al.*, 2019, J Cell Mol Med

Sequencing method (NGS)

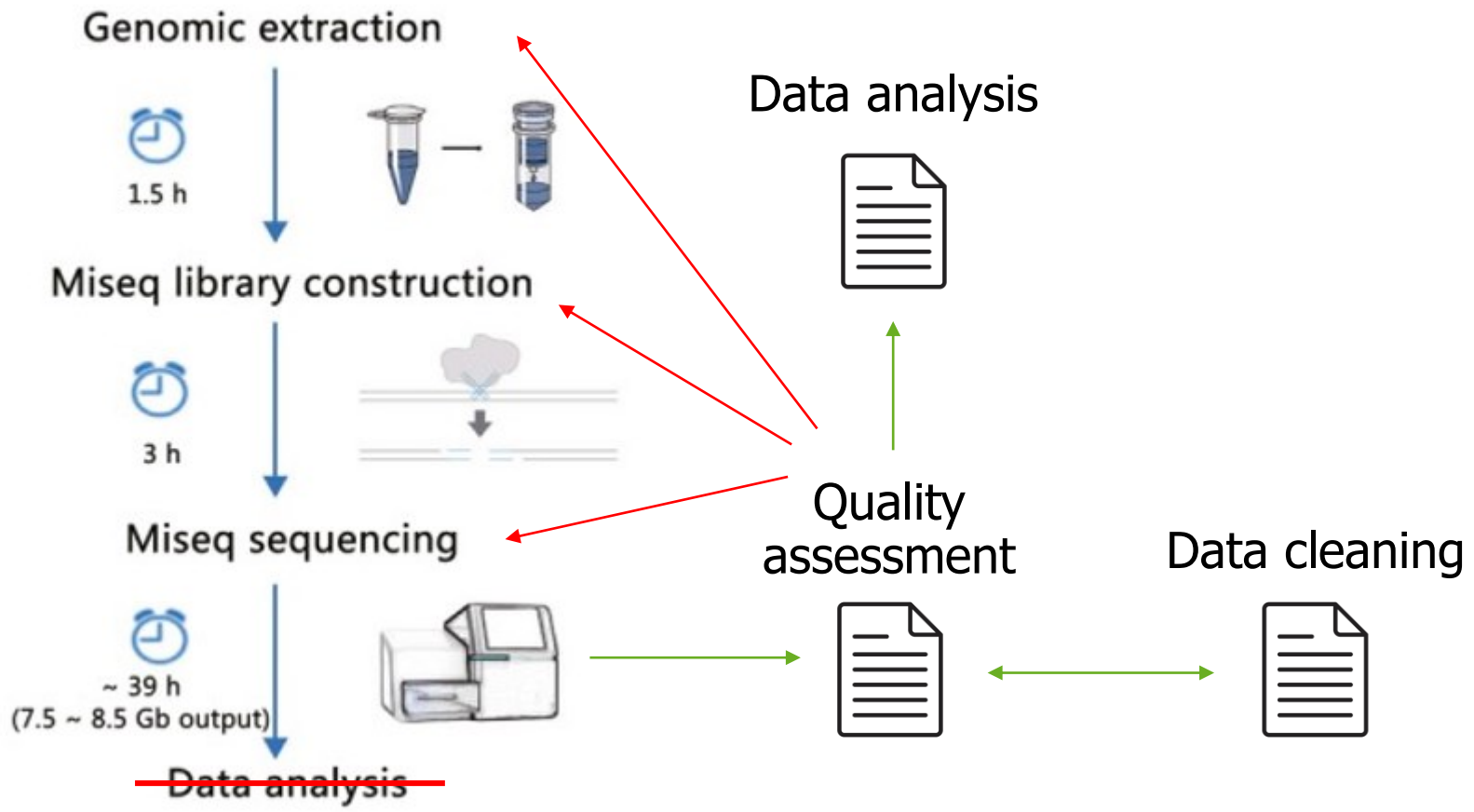


Data analysis

Quality assessment

Data cleaning

Sequencing method (NGS)



Course page

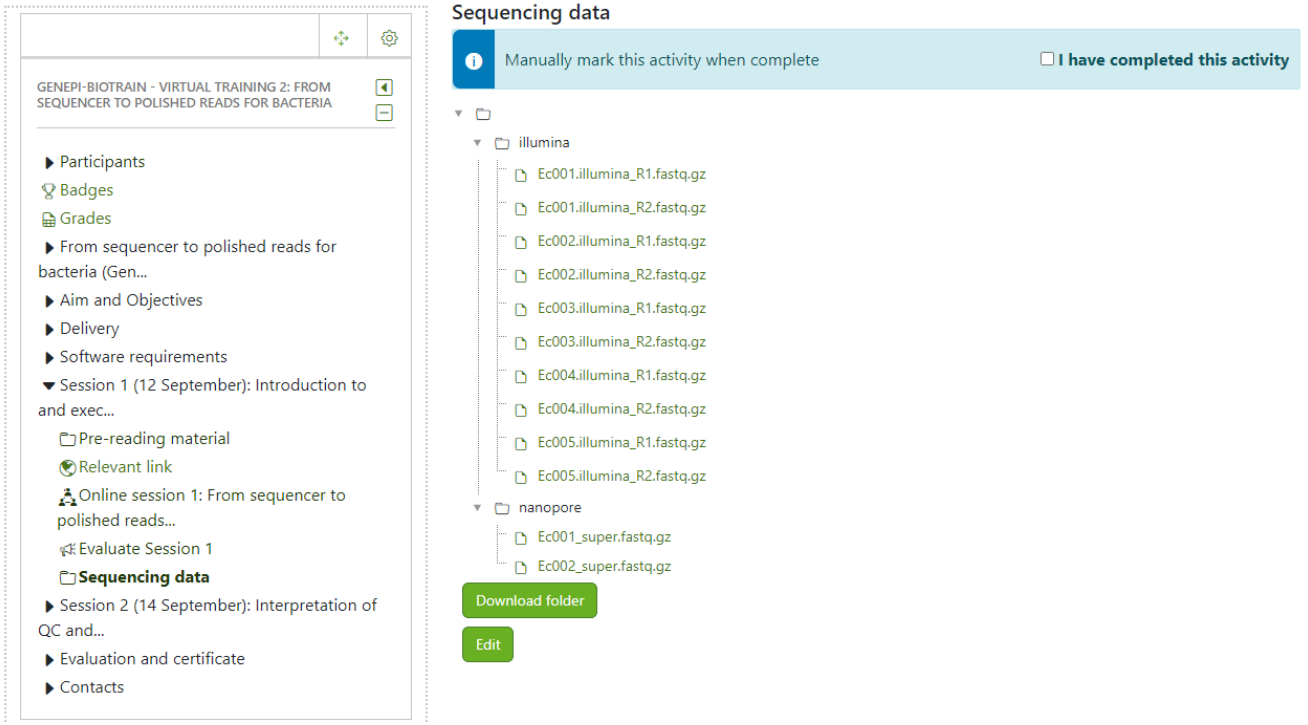
Life demo

Dataset

Find datasets on EVA

Session1 > Sequencing data

We assume that you have pre-installed FastQC and NanoPlot



From sequencer to polished reads for bacteria

Illumina sequencing theory

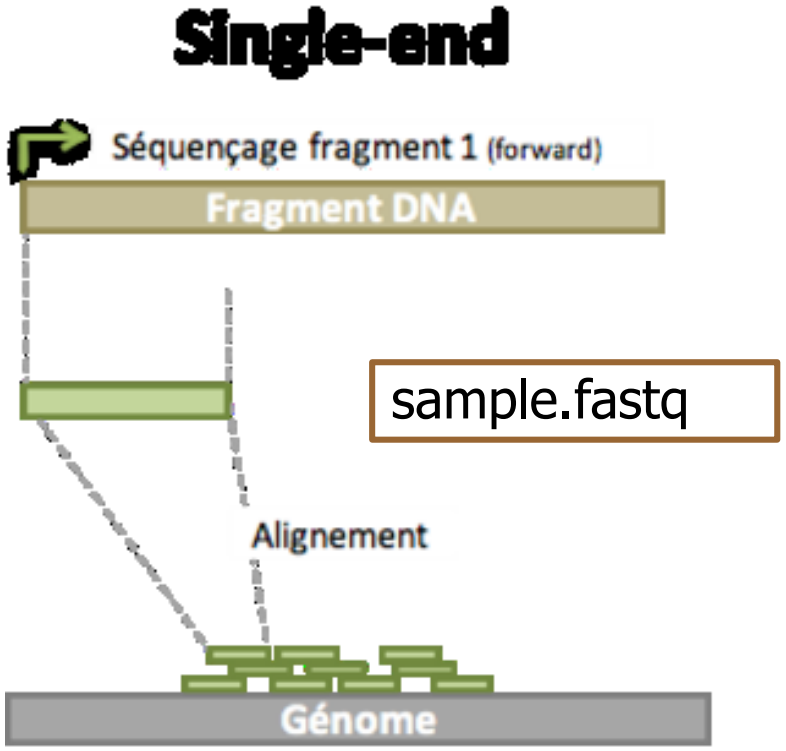
September 2023, Statens Serum Institut

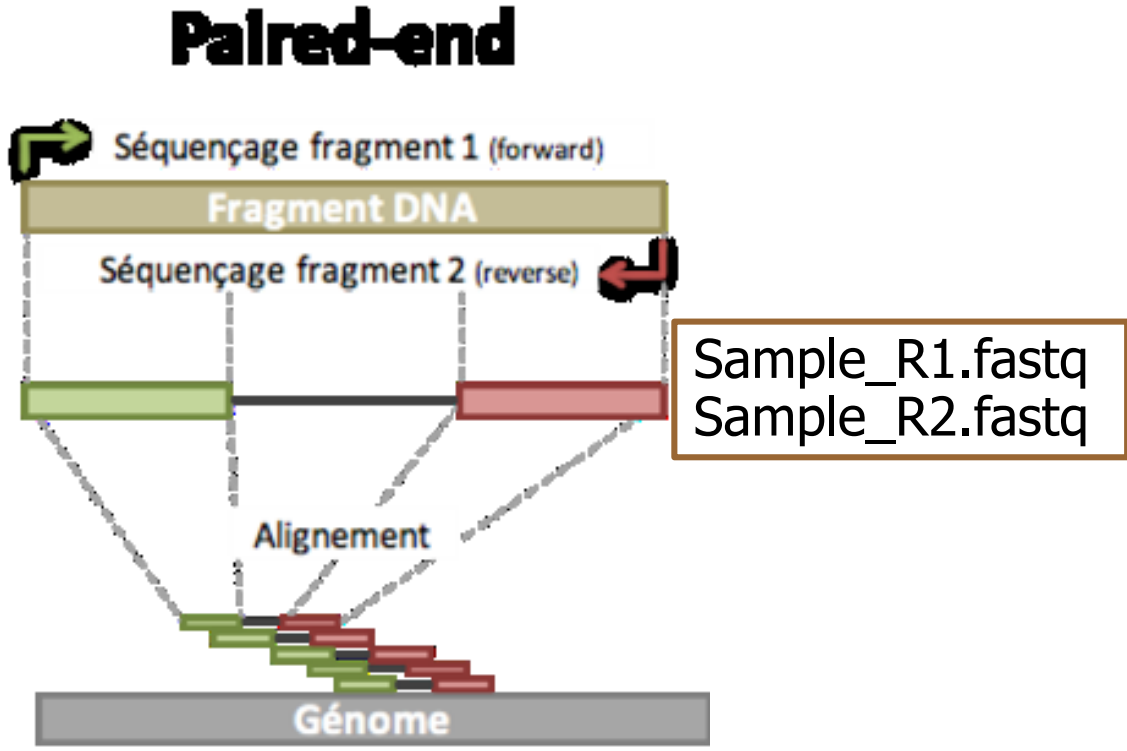
ILO's

- A fundamental understanding of read type and read fragment sizes
- A crude understanding of tagmantation, size selection of fragments, as well as sequencing by synthesis
- Finally, knowledge of the output from the sequencer.

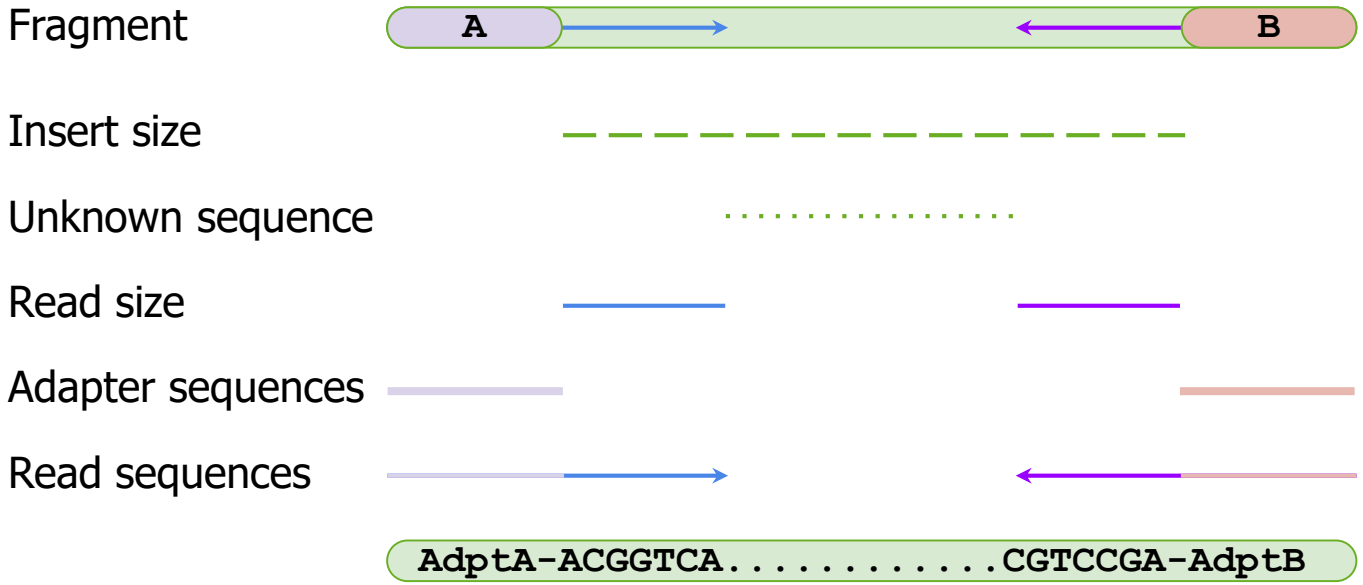
Illumina read types

Single-end vs Paired-end reads



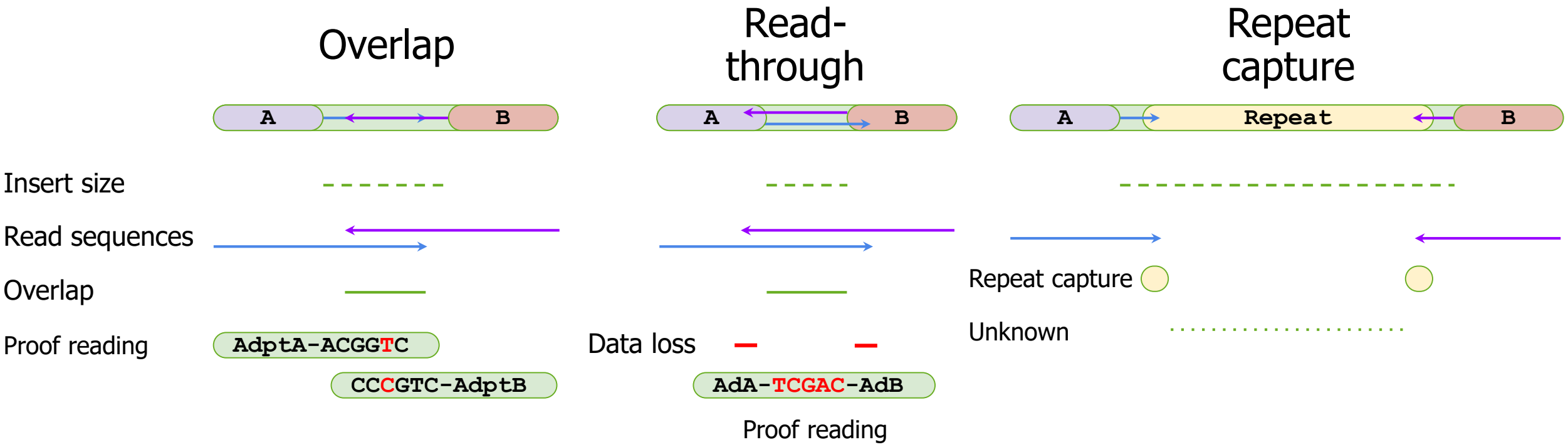


Read size




```
      @sample1_Mate1_readX
AdptA-ACGGTCA
      ...
      @sample1_Mate2_readX
AdptB-AGCCTGC
      …
      @sample1_Mate1_readY
```

Scenarios


Overlap
Read-
through
Repeat
capture
A
B
A
B
A
Repeat
B
Insert size
Read sequences
Overlap
Repeat capture
Unknown
Proof reading
AdptA-ACGGTC
CCCGTC-AdptB
Data loss
AdA-TCGAC-AdB
Proof reading

Illumina Nextera XT in summary

1) Sample prep

- gDNA extraction

- Pre-normalization

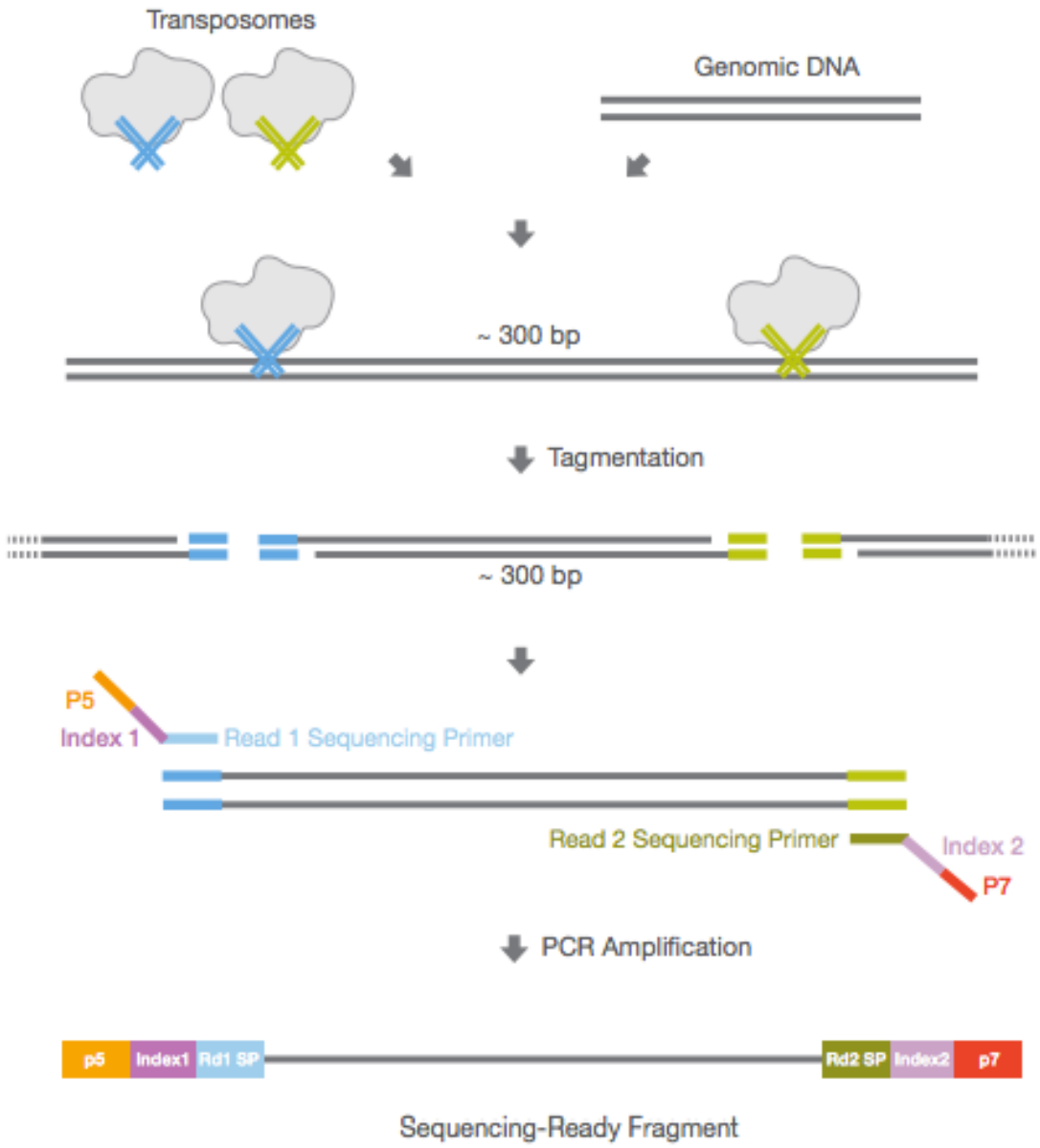
2) Library preparation

- Tagmentation

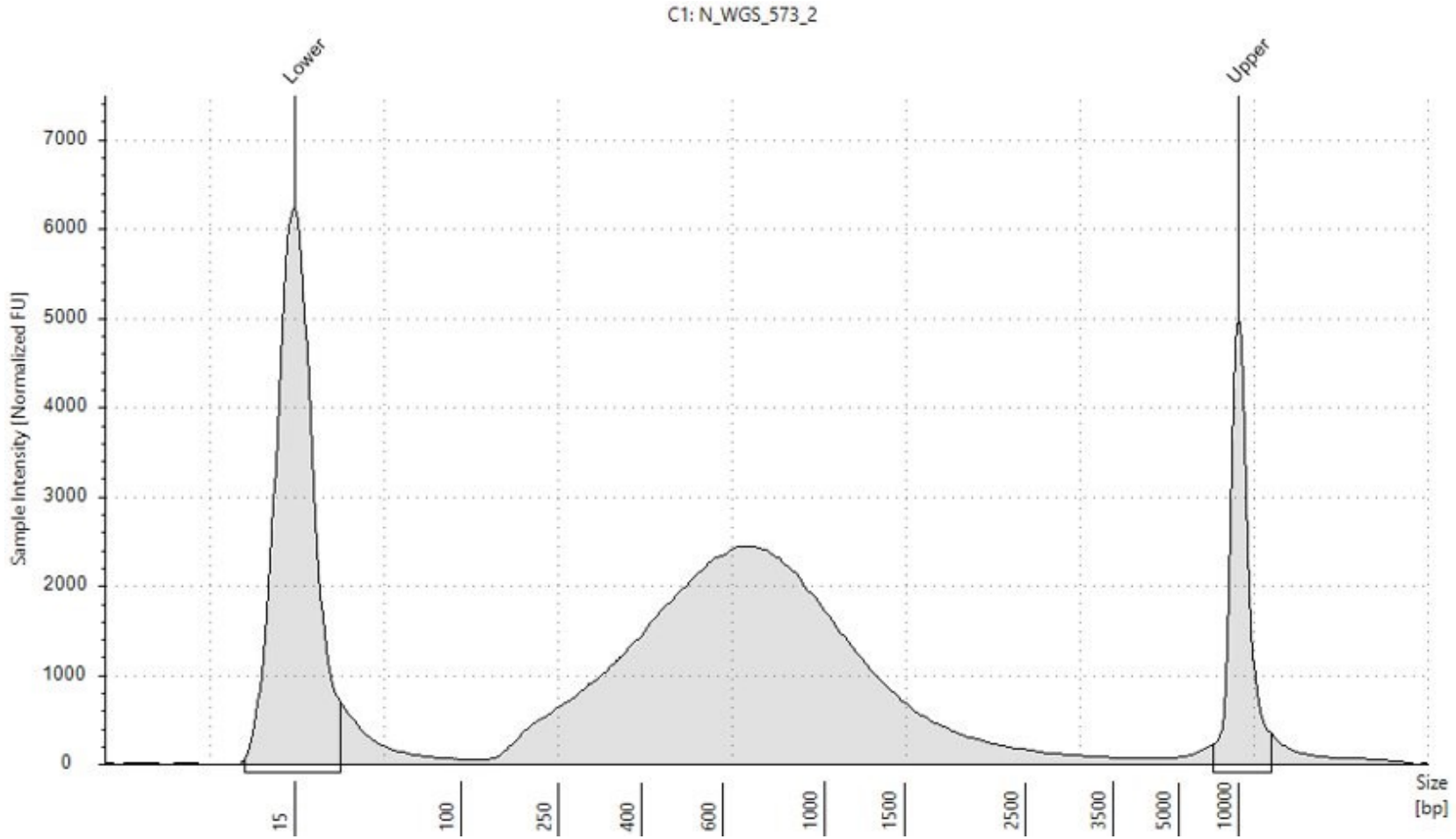
- Index PCR

- Normalization and pool

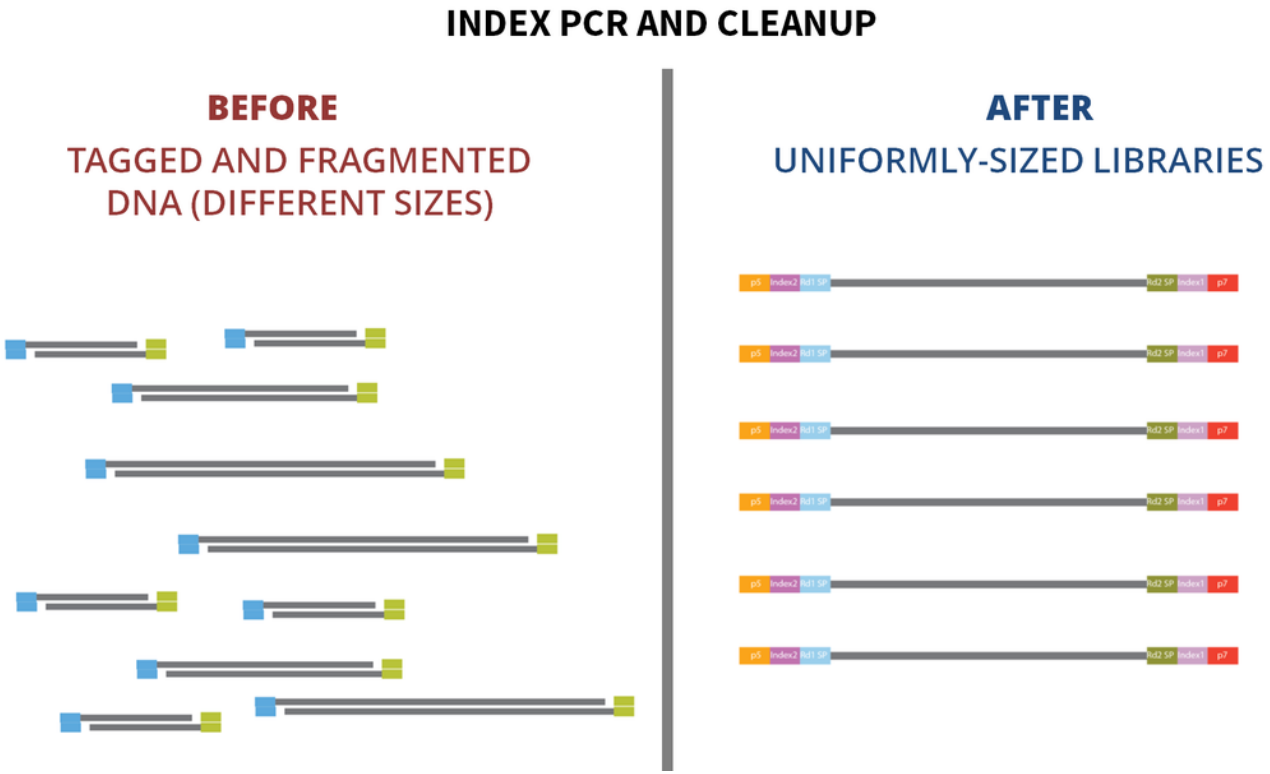
3) Sequencing

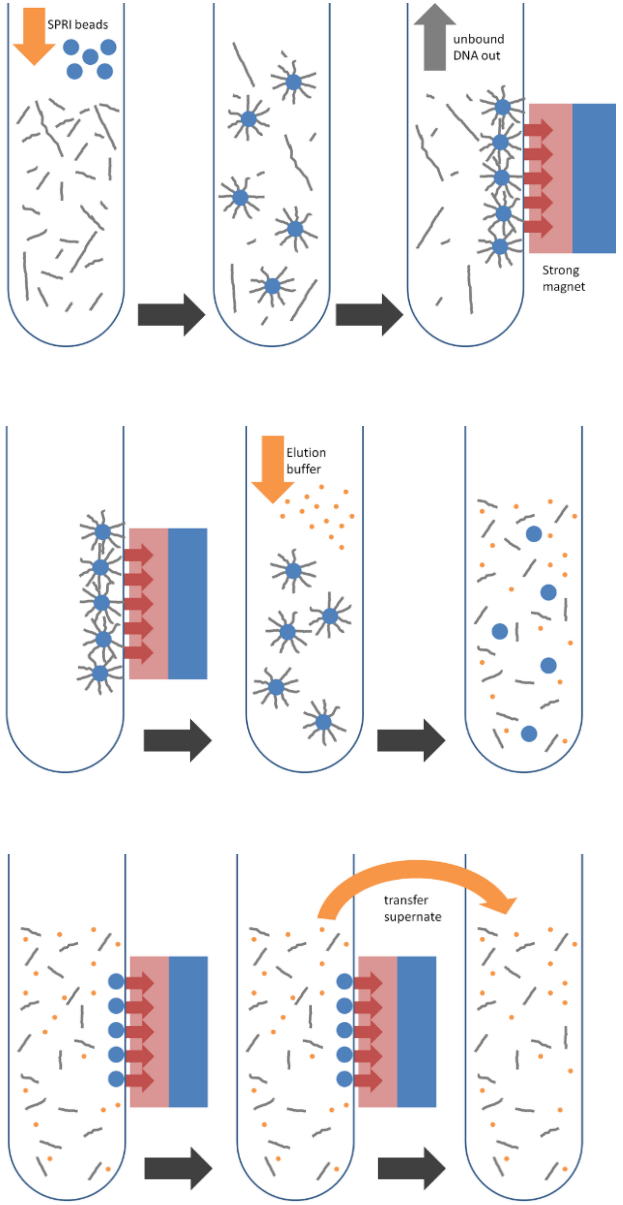


Library size selection
Bead-based size selection

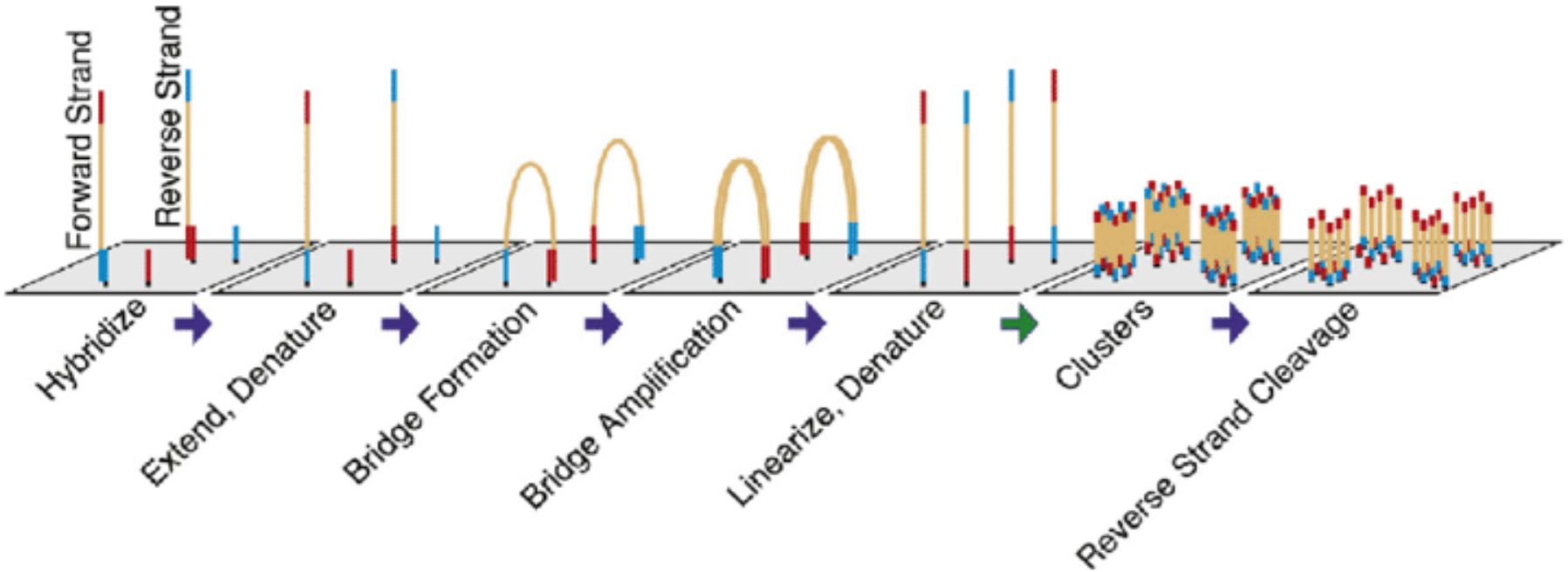


Library size selection
Bead-based size selection

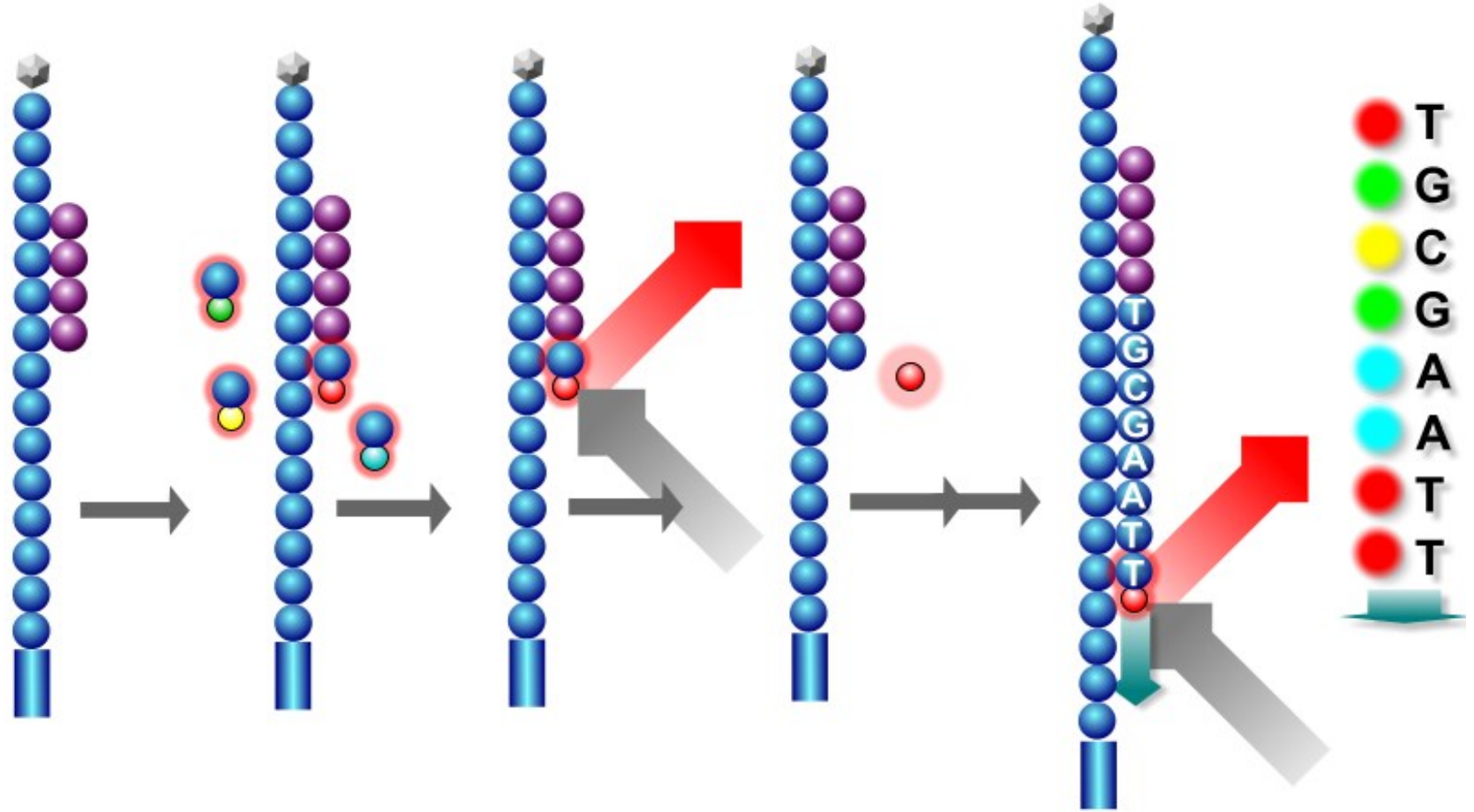




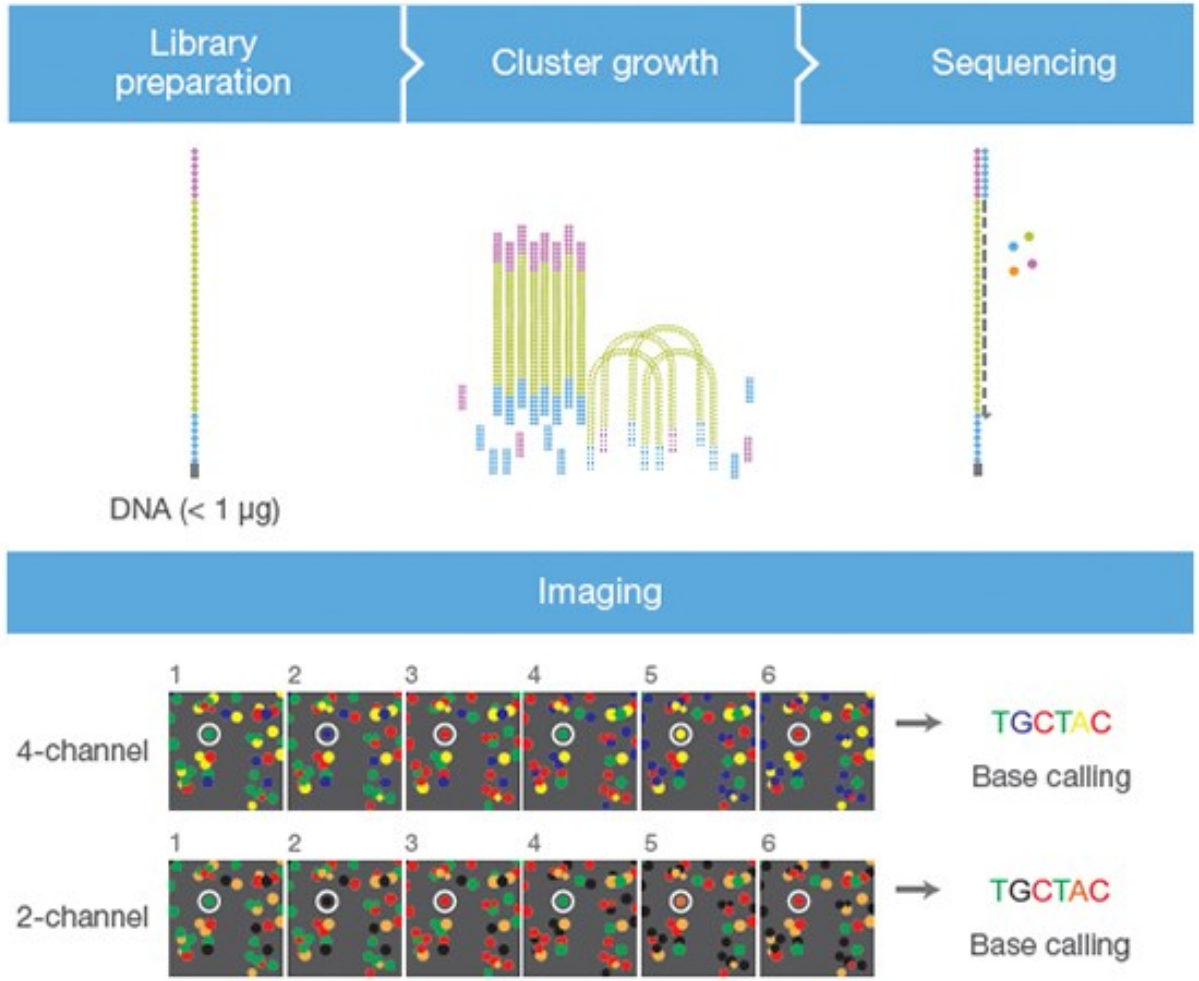
Illumina cluster generation



Sequencing by Synthesis



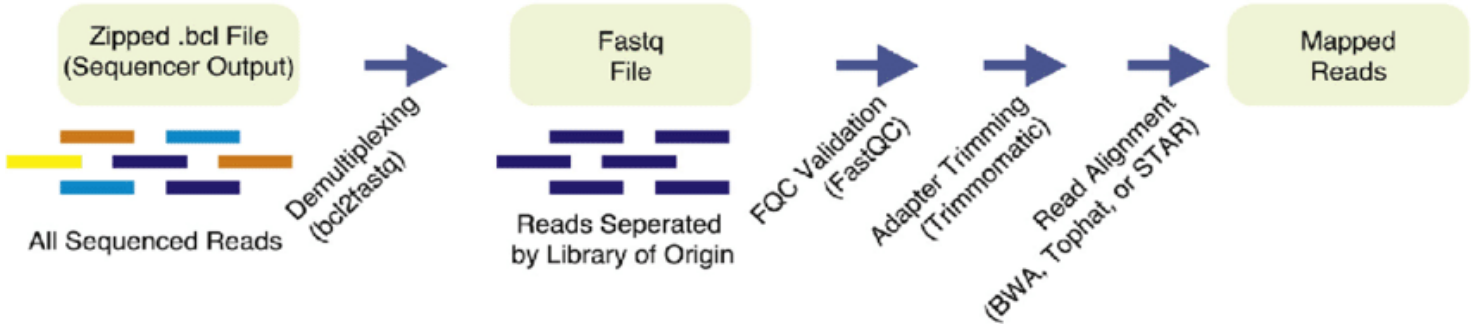
Illumina cluster generation



Illumina sequence data files

Illumina sequencer generates .bcl

Translated to fastq file format on the machine using bcl2fastq




QUESTIONS

Break!

From sequencer to polished reads for bacteria

Nanopore sequencing theory

September 2023, Statens Serum Institut

ILOs

- Understand the fundamentals of Nanopore sequencing.
- Explain how library preparation affects Nanopore sequencing data.
- Distinguish between Nice-to-Know and Need-to-Know information.

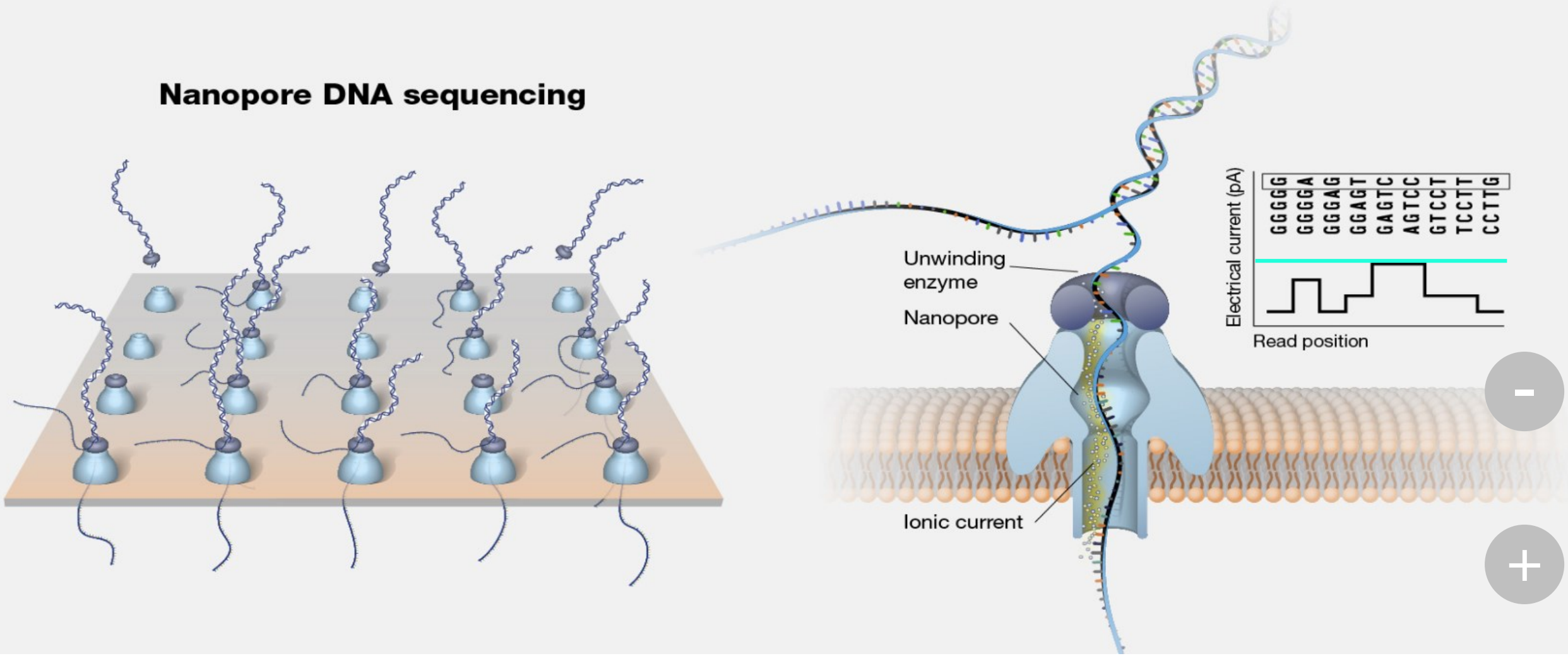
Sequencing fundamentals

Extraction

DNA

Library preparation

Nanopore sequencing



Nanopore DNA Sequencing (genome.gov)

Nanopore sequencing platform overview

MinION and Flongle Flow Cell compatible

PromethION Flow Cell compatible

Configuration	Platform			Techniques	Tech specifications			
Number of flow cells per device	1	1	1	5	2	2	24	48
Maximum number of channels per flow cell	512	512	512	512	2,675	2,675	2,675	2,675
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours
Device TMO†	50 Gb	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 Tb	~14 Tb
Maximum number of flow cells per year*	104	104	104	520	208	208	2,596	4,992
Offer sequencing as a service	No	No	No	Yes	Yes	Yes	Yes	Yes

Product specifications (nanoporetech.com)

Nanopore sequencing platform overview

MinION and Flongle F

PromethION Flow Cell compatible

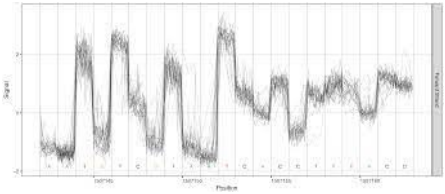
Configuration	Platform				Tech specifications			
Number of flow cells per device	1	1			2	2	24	48
Maximum number of channels per flow cell	512	512			675	2,675	2,675	2,675
Run time	72 Hours	72 Hours			Hours	72 Hours	72 Hours	72 Hours
Device TMO†	50 Gb	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 Tb	~14 Tb
Maximum number of flow cells per year*	104	104	104	520	208	208	2,596	4,992
Offer sequencing as a service	No	No	No	Yes	Yes	Yes	Yes	Yes

MinKNOW

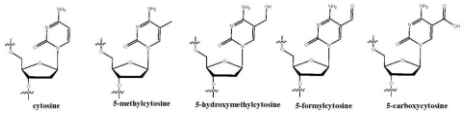
Product specifications (nanoporetech.com)

Nanopore sequencing data

- The raw electrical signal is saved in **pod5 files** and is translated, or base called, into nucleotide bases which are stored in fastq files.



- Each DNA strand will be represented as a read and the **length is determined by input DNA length**.
- With enough computer power, squiggles are base called in **real time** and you can start to analyse while still sequencing.
- **Base modifications** (methylation, glycosylation etc.) will appear as abnormal squiggles. Very interesting but difficult to basecall correctly.



- Relatively **high error-rate** specifically in homopolymeric regions and because of base modifications.

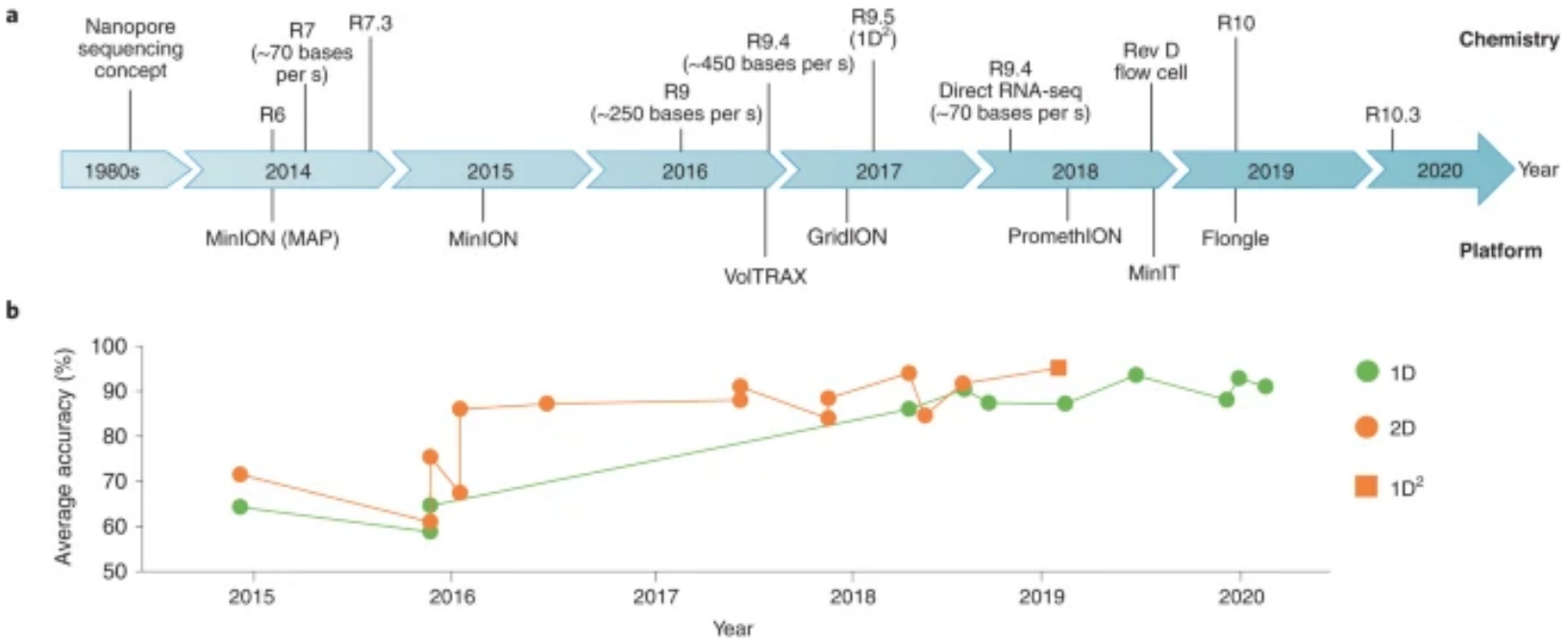
https://nanoporetech.com/applications/dna-nanopore-sequencing

A bit of history…

Nanopore sequencing was made available to the public in 2014.

- Since then the chemistry has changed many times..

Fig. 2: ONT sequencing data improvement over time.

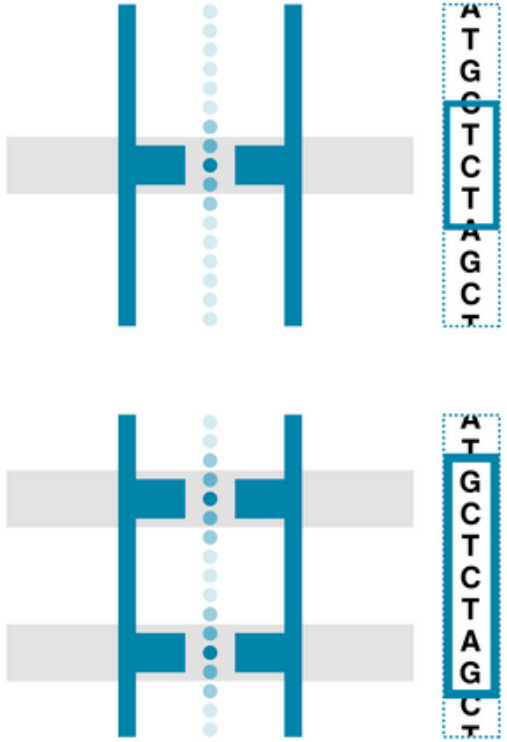


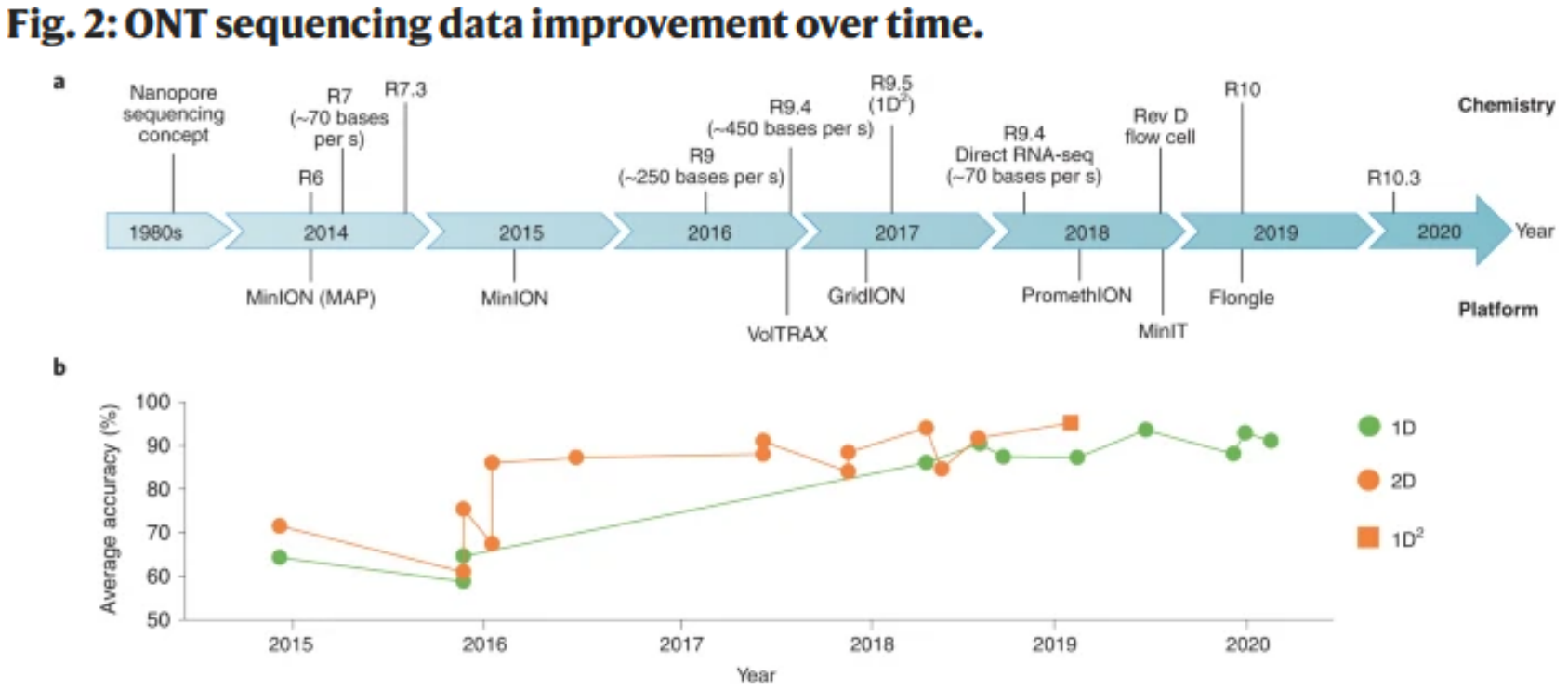
Wang, Y., Zhao, Y., Bollas, A. *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021). https://doi.org/10.1038/s41587-021-01108-x

A bit of history…

Nanopore sequencing was made available to the public in 2014.

- Since then the chemistry has changed many times..





Wang, Y., Zhao, Y., Bollas, A. *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021). https://doi.org/10.1038/s41587-021-01108-x

A bit of history…

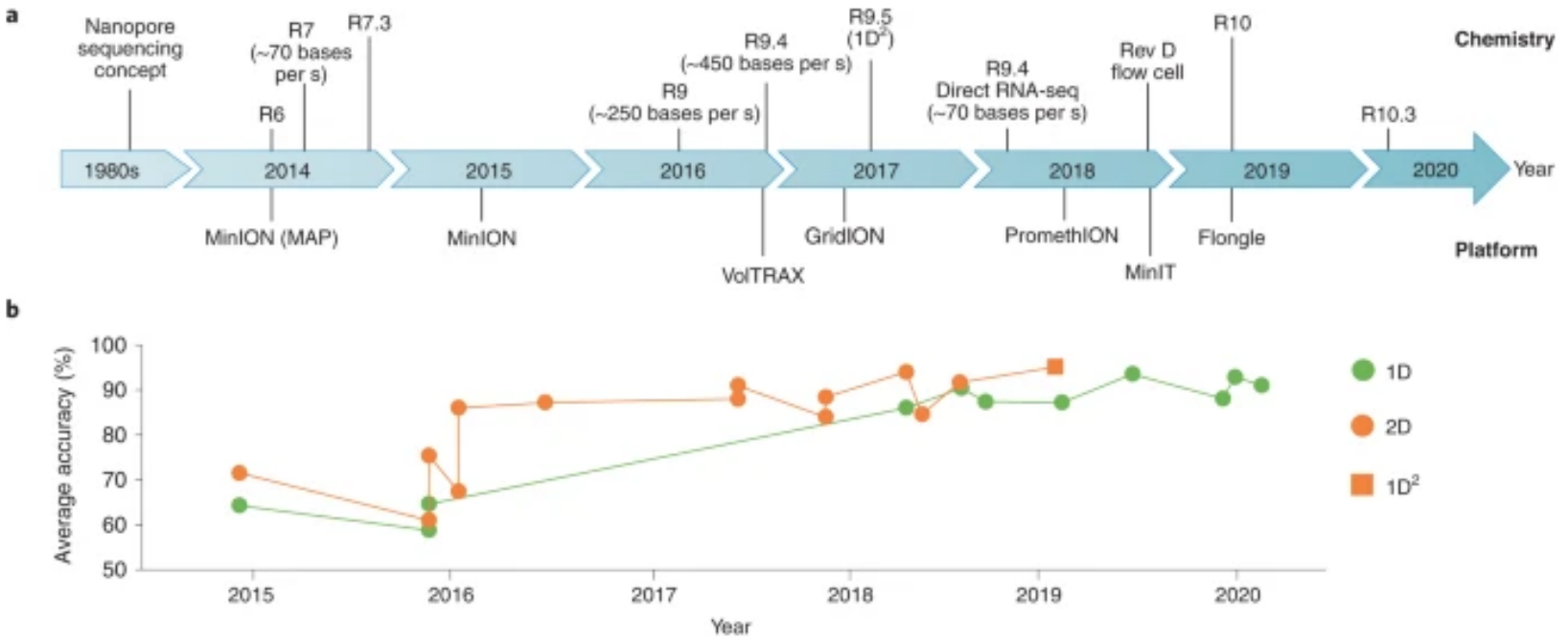
Nanopore sequencing was made available to the public in 2014.

- Since then the chemistry has changed many times..

..and so has the software

- 2017 == Albacore
- 2019 == Guppy
- 2023 == Dorado

Fig. 2: ONT sequencing data improvement over time.



Wang, Y., Zhao, Y., Bollas, A. *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021). https://doi.org/10.1038/s41587-021-01108-x

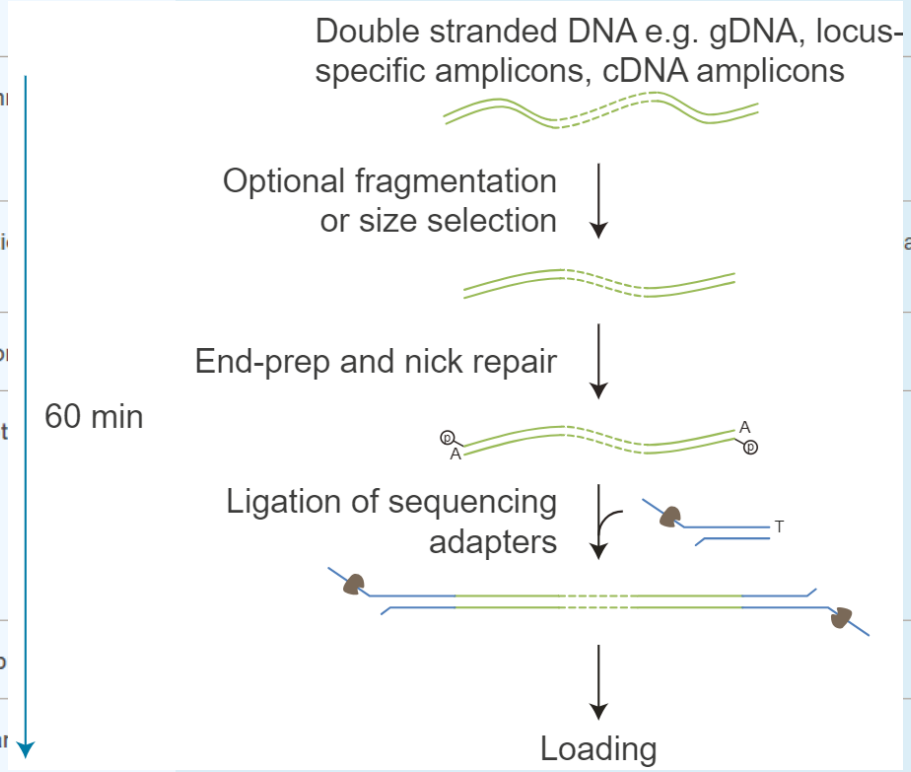
Nanopore library prep

	Output optimised	Speed optimised	Ultra-long reads optimised	Low input optimised	Targeted sequencing	
	Ligation Sequencing Kit	Rapid Sequencing Kit	Ultra-Long DNA Sequencing Kit	Rapid PCR Barcoding Kit	16S Barcoding Kit	Cas9 Sequencing Kit
Preparation time	60 minutes	10 minutes	200* mins +1xO/N incubation	15 mins + PCR	10 mins + PCR	110 minutes
Input recommendation	1000 ng gDNA 100-200 fmol for amplicons	50 - 100 ng	6M Cells	1 - 5 ng	10 ng	1 - 10 µg
Fragmentation	Optional	Transposase-based	Transposase-based	Transposase-based	-	Cas9-dependent cleavage
Amplification	No	No	No	Yes	Yes	No
Barcode options	Native Barcoding Kit 24 Native Barcoding Kit 96	Rapid Barcoding Kit 24 Rapid Barcoding Kit 96	-	12 plex	24 plex	*In development*
Typical output	●●●	●●○	●●○	●●○	●●○	●○○
Adaptive sampling	✓	✓	✓	✓	-	*In development*
Methylation included	✓	✓	✓	-	-	✓

Resource centre (nanoporetech.com)

Nanopore library prep

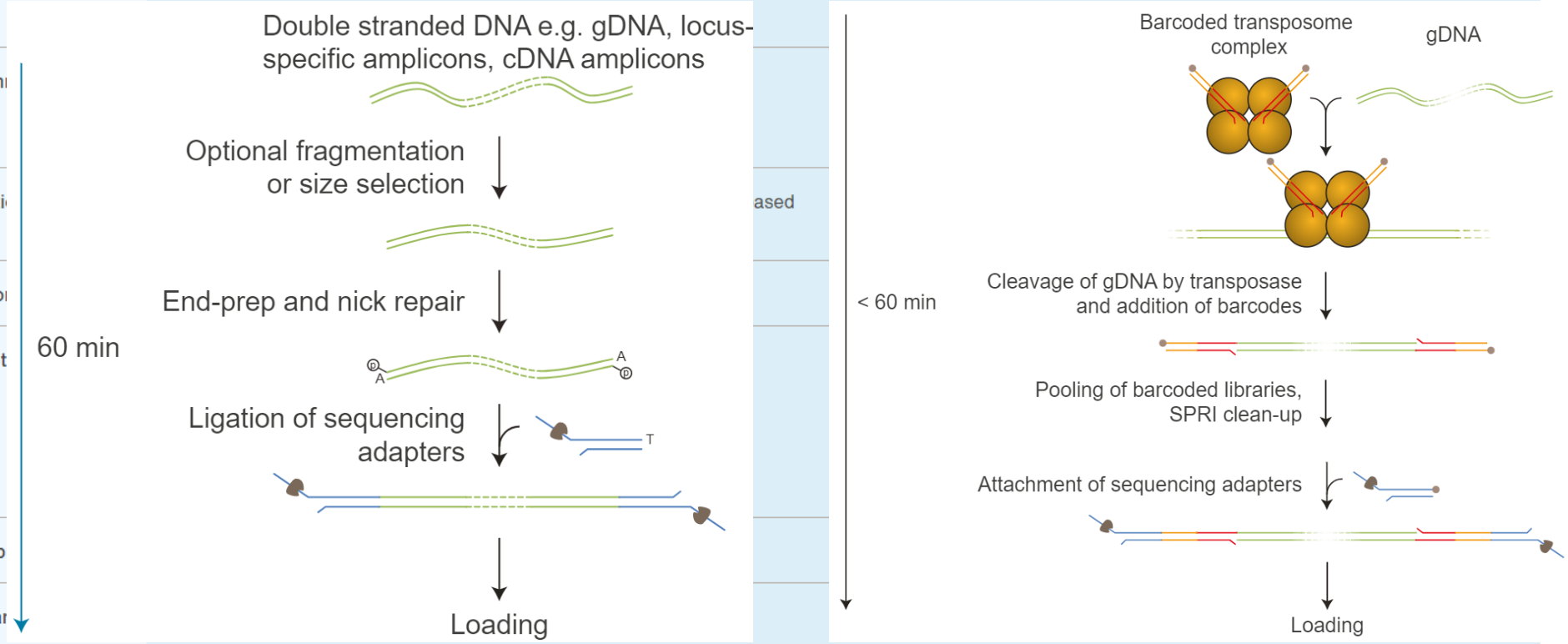
	Output optimised	Speed optimised	Ultra-long reads optimised	Low input optimised	Targeted sequencing	
	Ligation Sequencing Kit	Rapid Sequencing Kit	Ultra-Long DNA Sequencing Kit	Rapid PCR Barcoding Kit	16S Barcoding Kit	Cas9 Sequencing Kit
Preparation time	60 minutes	10 minutes	200* mins +1xO/N	15 mins + PCR	10 mins + PCR	110 minutes
Input recom				1 - 5 ng	10 ng	1 - 10 µg
Fragmentati			ased	Transposase-based	-	Cas9-dependent cleavage
Amplificatio				Yes	Yes	No
Barcode opt				12 plex	24 plex	*In development*
Typical outp				●●○	●●○	●○○
Adaptive sa				✓	-	*In development*
Methylation included	✓	✓	✓	-	-	✓



Resource centre (nanoporetech.com)

Nanopore library prep

	Output optimised	Speed optimised	Ultra-long reads optimised	Low input optimised	Targeted sequencing	
	Ligation Sequencing Kit	Rapid Sequencing Kit	Ultra-Long DNA Sequencing Kit	Rapid PCR Barcoding Kit	16S Barcoding Kit	Cas9 Sequencing Kit
Preparation time	60 minutes	10 minutes	200* mins +1xO/N	15 mins + PCR	10 mins + PCR	110 minutes
Input recom						
Fragmentati			ased			
Amplificatio						
Barcode opt						
Typical outp						
Adaptive sa						
Methylation included	✓	✓	✓	-	-	✓



Resource centre (nanoporetech.com)

Nanopore library prep

Choice of library kit and sequencing platform will affect:

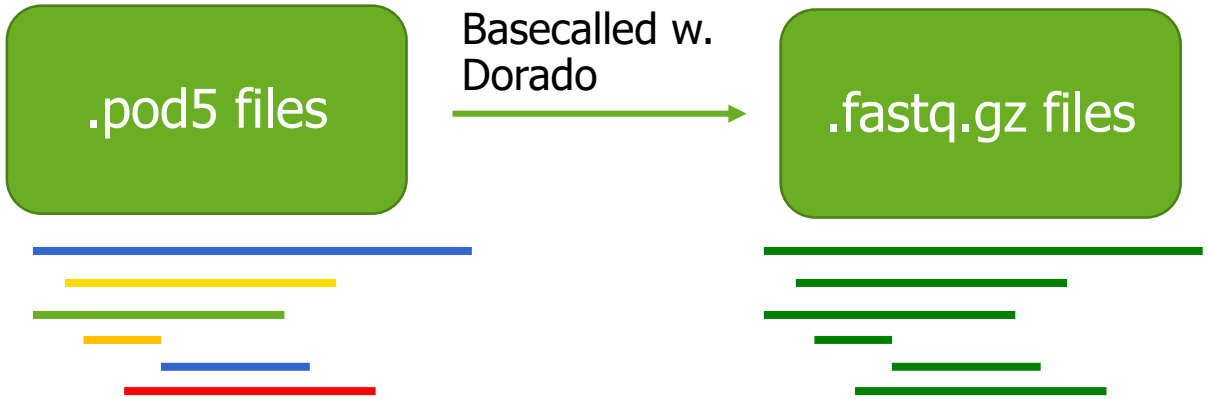
- Amount of data produced
- Length of reads
- Coverage
- Overall quality

Nanopore sequence data files

Nanopore sequencer generates demultiplexed .pod5 files

Translated to fastq file format with dorado

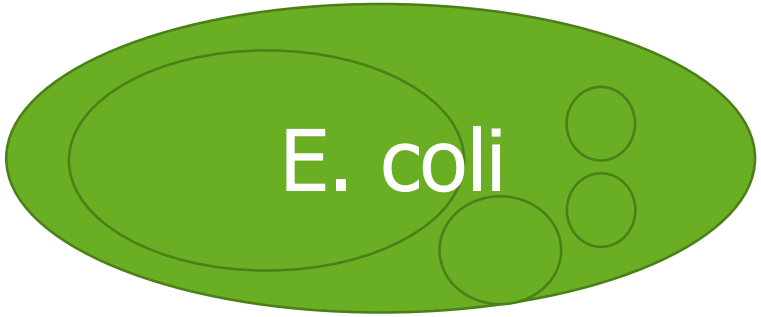
- Enable barcode trimming in MinKNOW

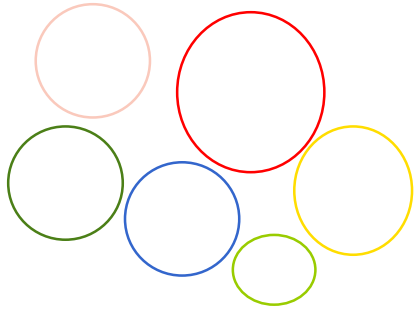


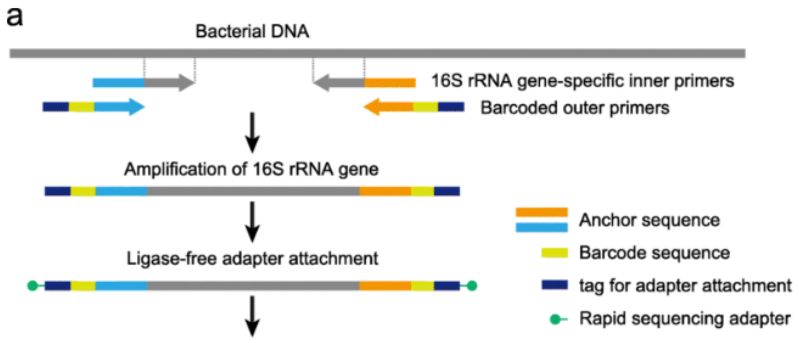
Need-to-Know

What was sequenced?

- Single bacterial isolate
- 16S
- Metagenomic sample







Matsuo, Yoshiyuki & Komiya, Shinnosuke & Yasumizu, Yoshiaki & Yasuoka, Yuki & Mizushima, Katsura & Takagi, Tomohisa & Kryukov, Kirill & Fukuda, Aisaku & Morimoto, Yoshiharu & Naito, Yuji & Okada, Hidetaka & Bono, Hidemasa & Nakagawa, So & Hirota, Kiichi. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. BMC Microbiology. 21. 35. 10.1186/s12866-021-02094-5.

Nice-to-Know

- Which flow cell version was used?
- Which library was used?
- Which base caller was used?
- Has it been quality filtered, if so on which q-score?
- Have adapters/barcodes been trimmed?


QUESTIONS

Summary

Summary of the two methods

Illumina

- Library
  - \+ 1 ηg/μl input required
  - − Amplification bias introduced
- Reads
  - \+ High accuracy > 99%
  - − Very short sequences
- Seqeuncing
  - \+ High yield: Gigabases
  - − 4h to 2.5 days

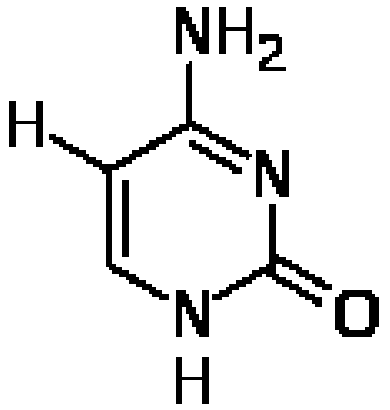
Nanopore

- Library
  - \+ Bias-free
  - − High DNA purity required
- Reads
  - \+ Very long reads (> 2Mb)
  - − Lower accuracy 92 – 99%
- Seqeuncing
  - \+ Real-time
  - − Fixed max. Yield / flowcell

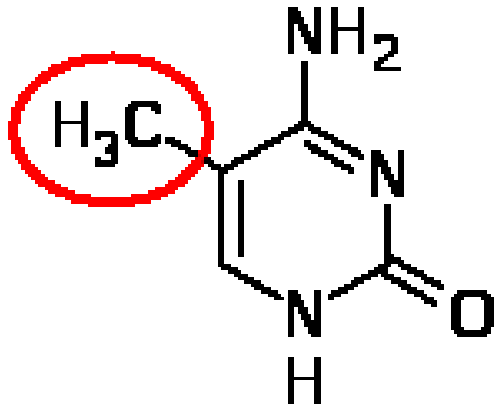
Nanopore sequencing

Important differences from NGS

- Read lengths are determined by input DNA length
- Relatively high error-rate
- Base modifications



cytosine



methylated cytosine

Which one to chose

It depends…

We use both Illumina and Nanopore data..

We even combine their data.

Break!

And now to something completely different…

Intro to Session 1 Practical

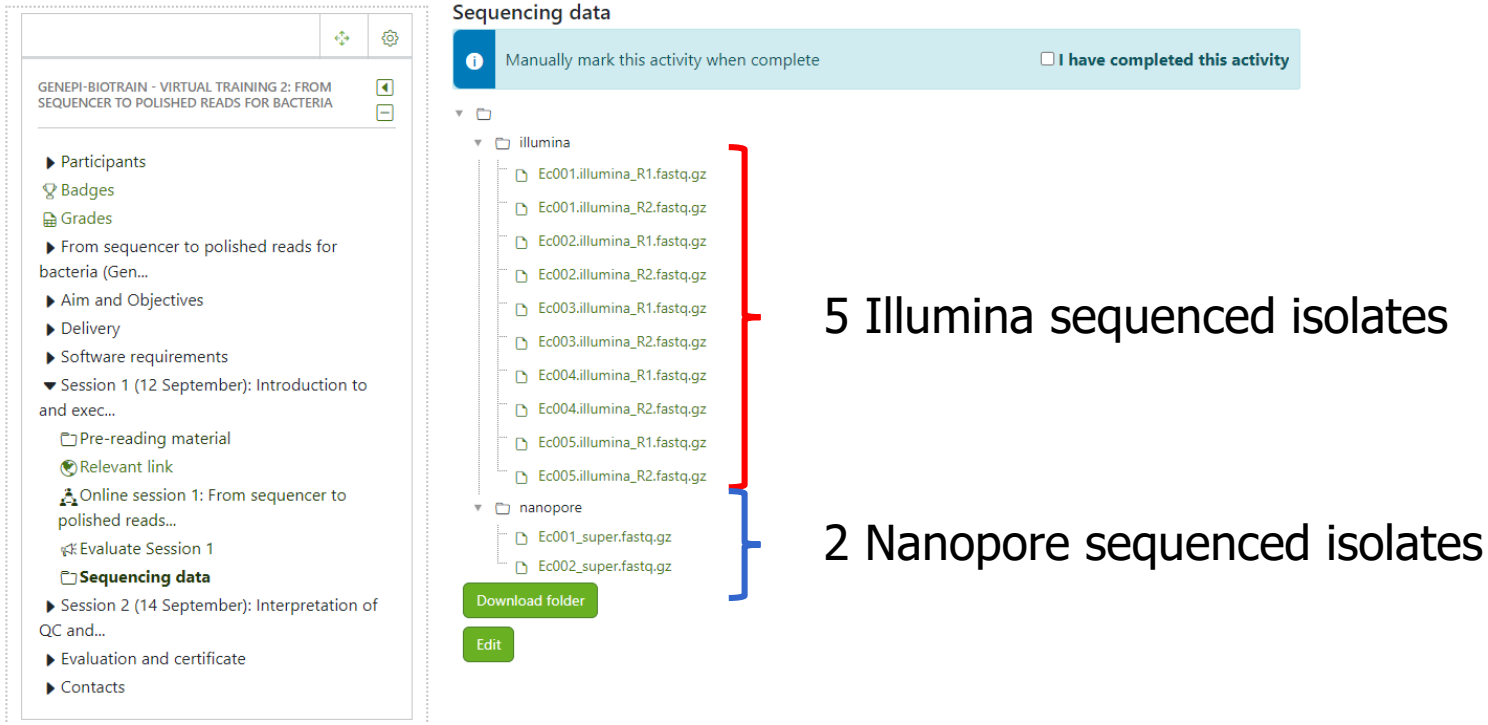
- Get to know your dataset
- Intro to contamination control
- Perform quality assessment on Illumina data
- Perform quality assessment on Nanopore data

From sequencer to polished reads for bacteria

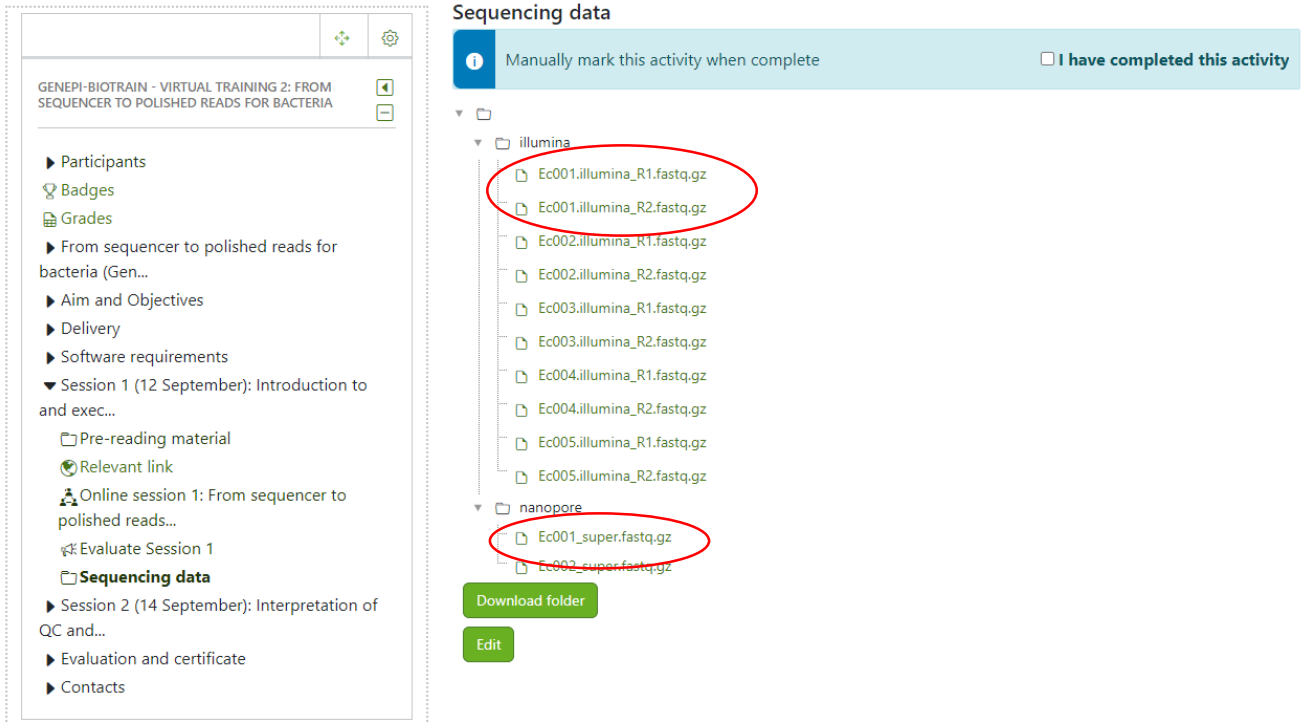
Get to know your dataset

September 2023, Statens Serum Institut

Whole genome sequencing data



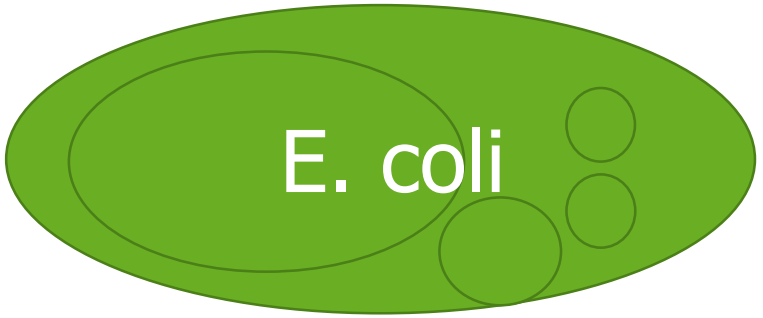
Whole genome sequencing data

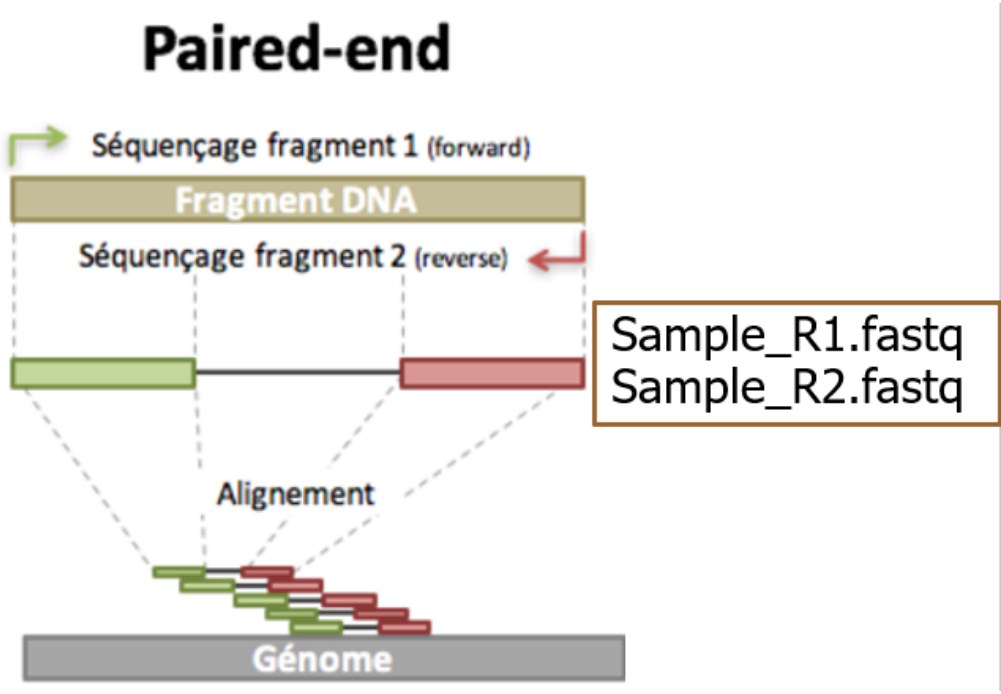


What to expect

Whole genome sequencing of E. coli

Illumina 2 x 250 bp paired end reads





First look at our sequencing data

The ‘less’ command will allow you to look at gzipped files.

```
less Ec001_super.fastq.gz
```

Press ‘q’ to exit

The FASTQ format

Read name:	@Sequence_ID							
Read Seq:	A	T	C	G	A	T	C	G
Separator:	+							
Phred-ASCII:	F	D	G	F	J	J	S	K
Phred:	36	34	38	36	40	40	41	39

fastq

Phred quality scores

Phred quality scores are logarithmically linked to error probabilities

Phred Quality 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%

$$P = 10^{-Q/10}$$

$$Q = -10 \log_{10}(P)$$

P = the probability that a base is incorrect

Q = Phred quality score

Number of reads and bases in fastq files

A pipe

Read count

```
echo $(zcat sample.fastq.gz|wc -l)/4|bc
```

Character count

```
zcat sample.fastq.gz | paste - - - - | cut -f 2 | tr -d '\n' | wc –c
```

Number of reads and bases in fastq files

A pipe

Read count

```
echo $(zcat sample.fastq.gz|wc -l)/4|bc
```

Character count

```
zcat sample.fastq.gz | paste - - - - | cut -f 2 | tr -d '\n' | wc -c
```

```
@seqID1     ATGCTGAAATCA     +     HVIKILIFDPFBI
@seqID2     GCTAGGTACCAT     +     2222;;;<=?CCH
@seqID3     ATGCTGAAATCA     +     (+.9>;;76667<
@seqID4     GCTAGGTACCAT     +      @(&&&&&&%'
@seqID5     ATGCTGAAATCA     +      DEGHKHJJKKEC
@seqID6     GCTAGGTACCAT     +      LEDCCEIB@@@
```

From sequencer to polished reads for bacteria

Contamination

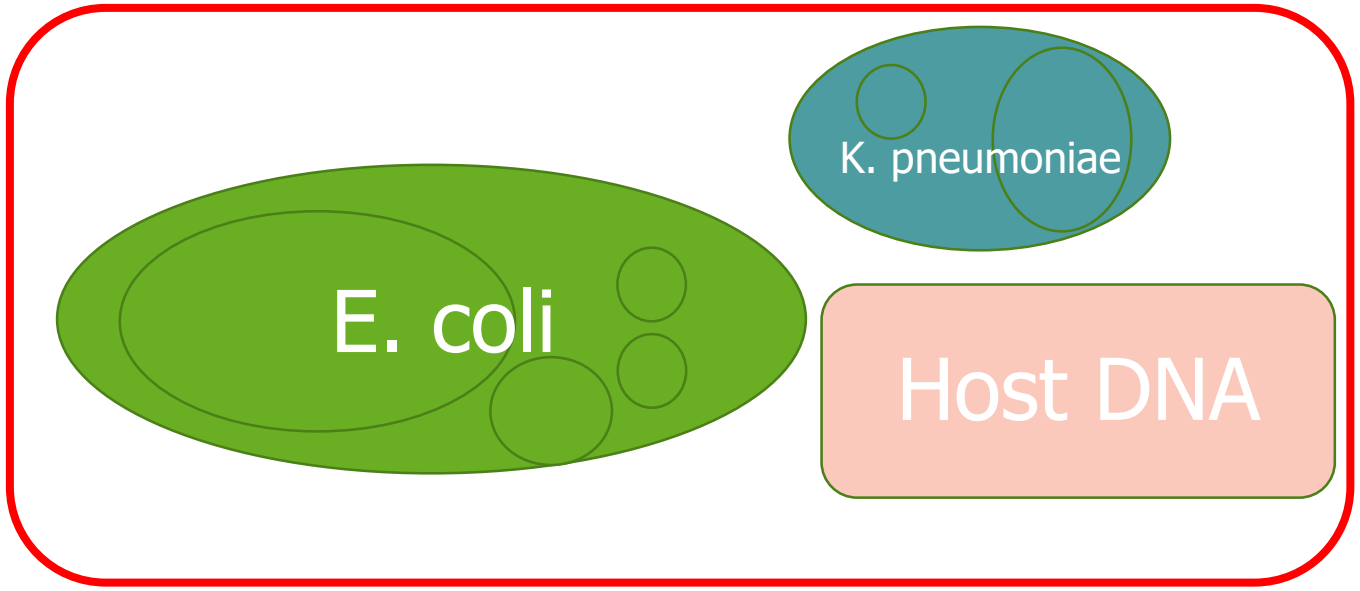
September 2023, Statens Serum Institut

Contamination in the context of sequencing data

- Single isolate sequencing: any sequence that does not belong to the organism sequenced.
- Can be interspecies contamination: sequences from another species or intraspecies contamination: sequences from the same species.

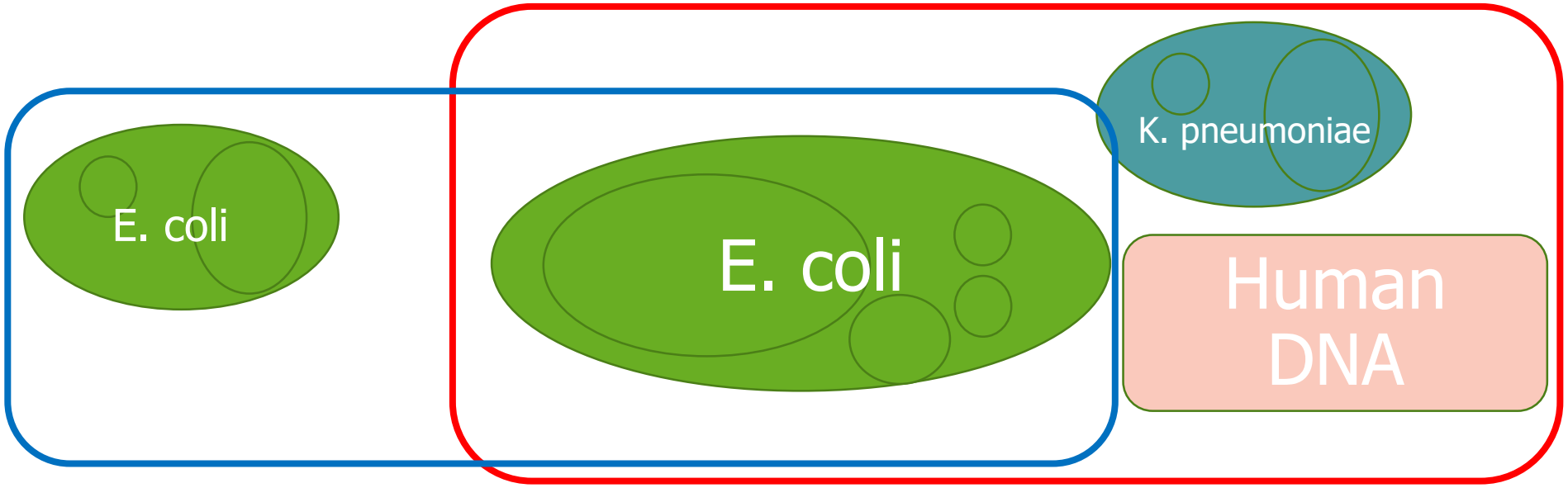
Contamination in the context of sequencing data

- Single isolate sequencing: any sequence that does not belong to the organism sequenced.
- Can be interspecies contamination: sequences from another species or intraspecies contamination: sequences from the same species.



Contamination in the context of sequencing data

- Single isolate sequencing: any sequence that does not belong to the organism sequenced.
- Can be interspecies contamination: sequences from another species or intraspecies contamination: sequences from the same species.



Possible sources of contamination in sequencing data

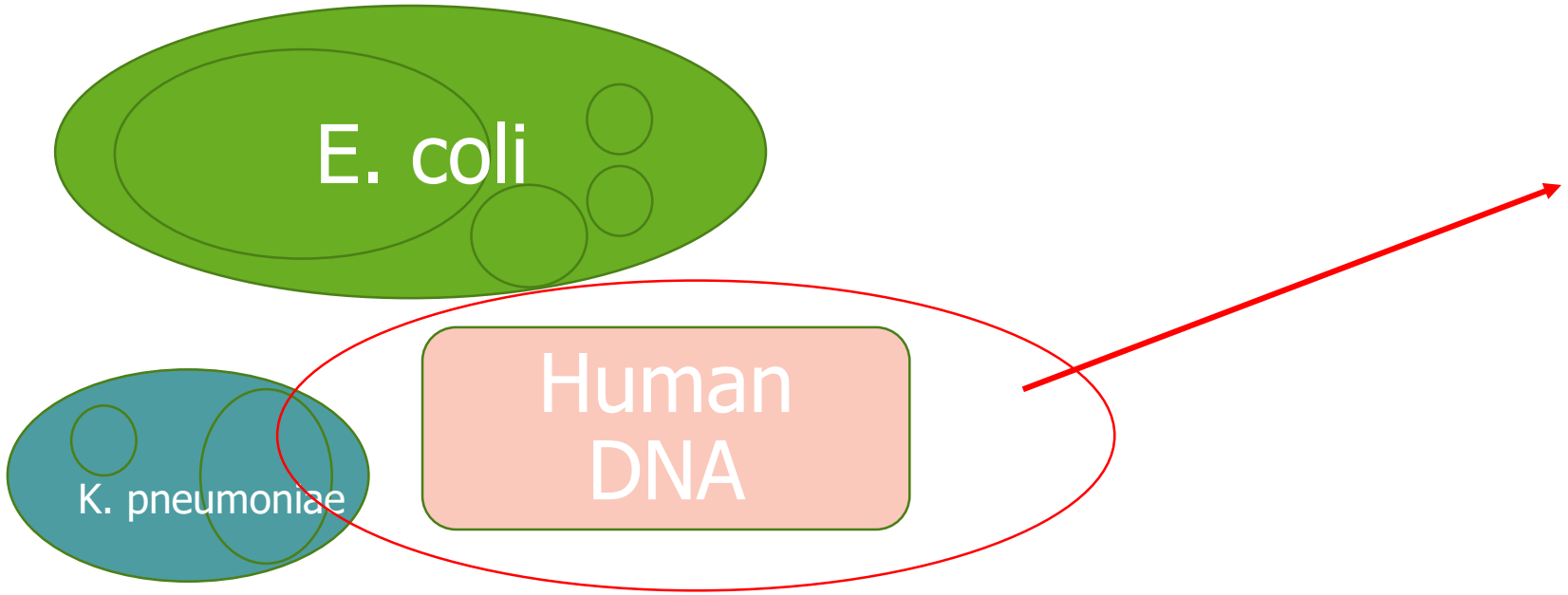
- Bacterial contamination in laboratory reagents and sequencing kits
- Cross-contamination across samples
- Host contamination
- Carryover contamination from previous sequencing runs

When can you do something about contaminated sequencing data

- You can always remove unrelated sequences if you know what your target sequences are.
    - You rarely know exactly what you are looking for.

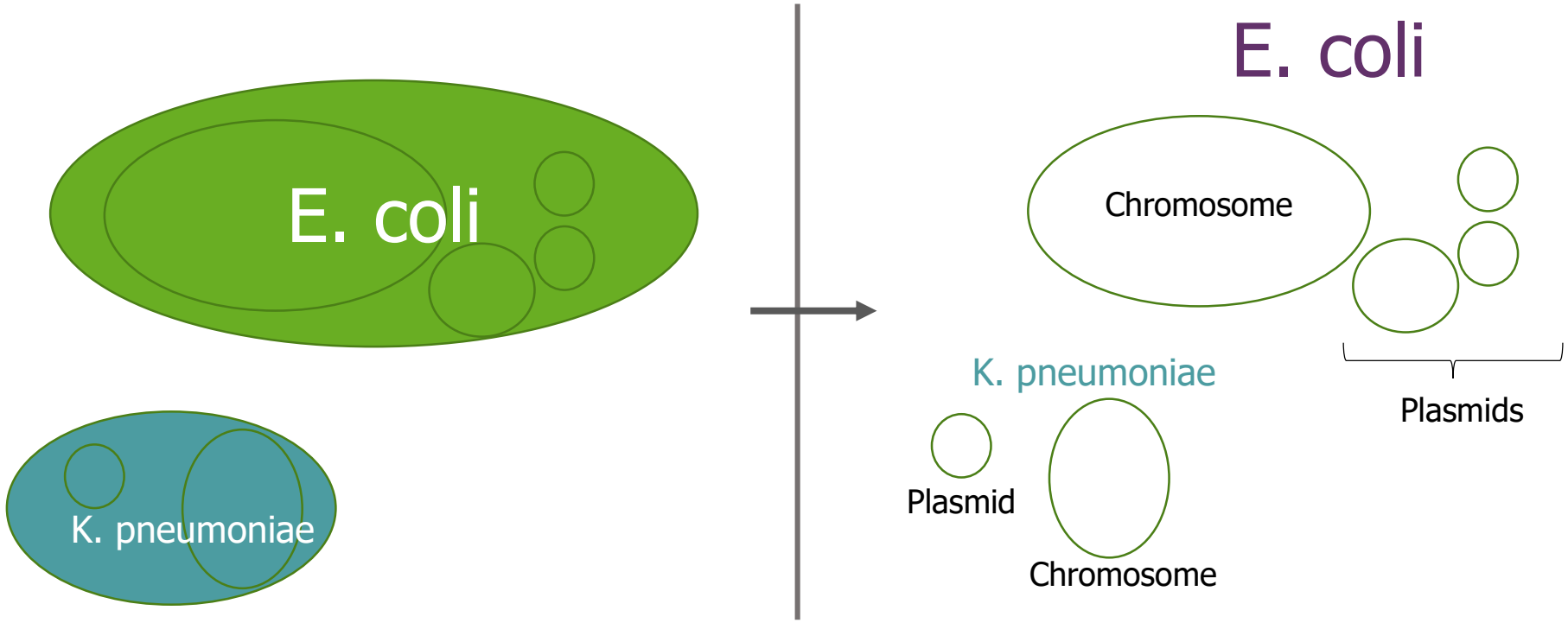
When can you do something about contaminated sequencing data?

- You can always remove unrelated sequences if you know what your target sequences are.
    - You rarely know exactly what you are looking for but you might know exactly what you are <u>not</u> looking for.



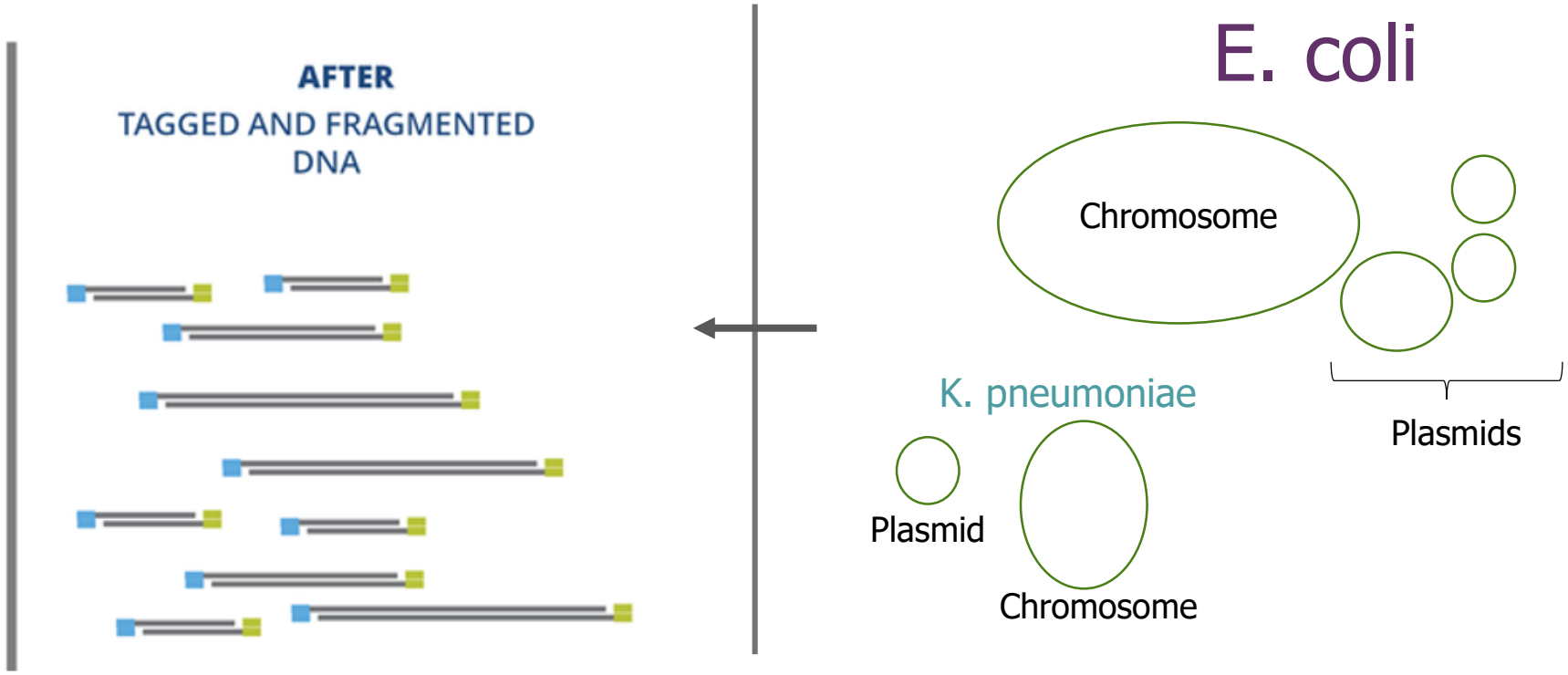
When can you do something about contaminated sequencing data?

- You can always remove **unrelated** sequences if you know what your target sequences are.



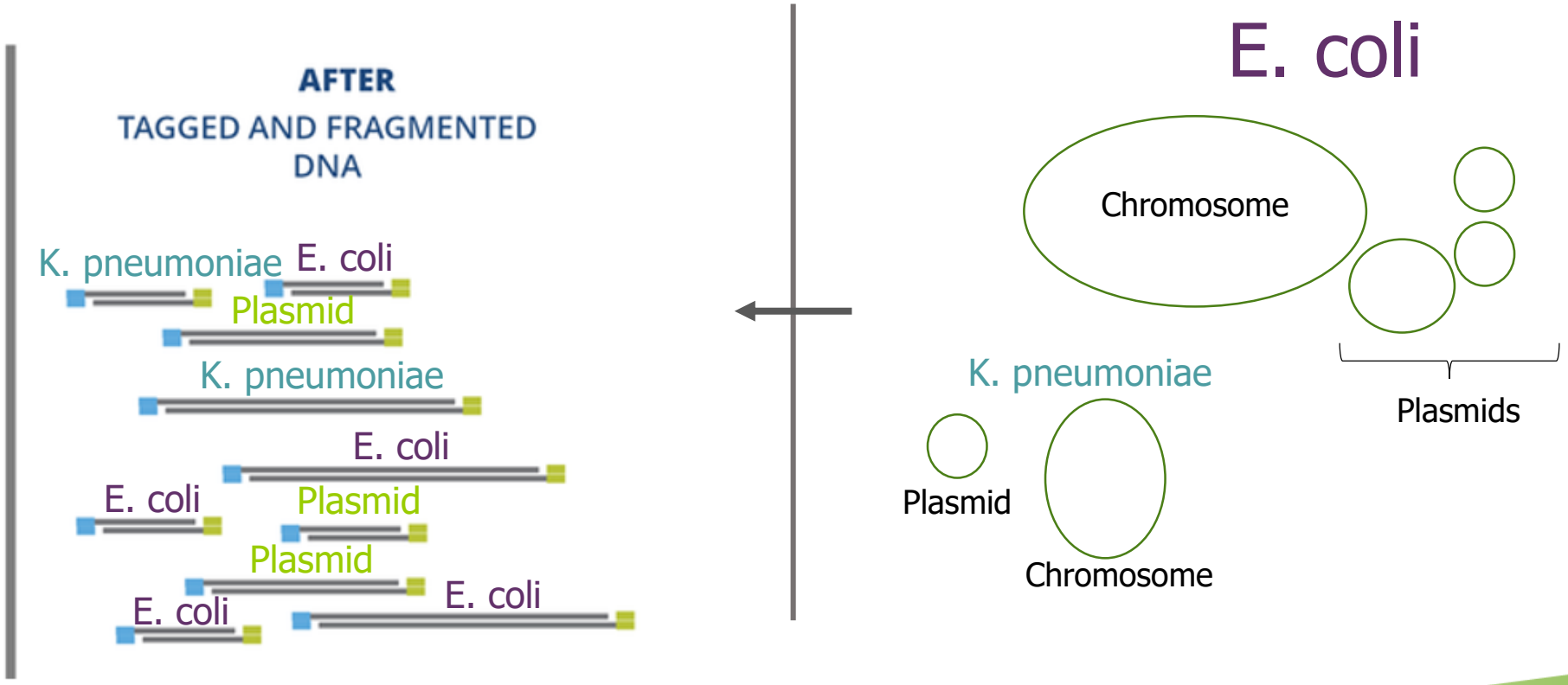
When can you do something about contaminated sequencing data?

- You can always remove **unrelated** sequences if you know what your target sequences are.



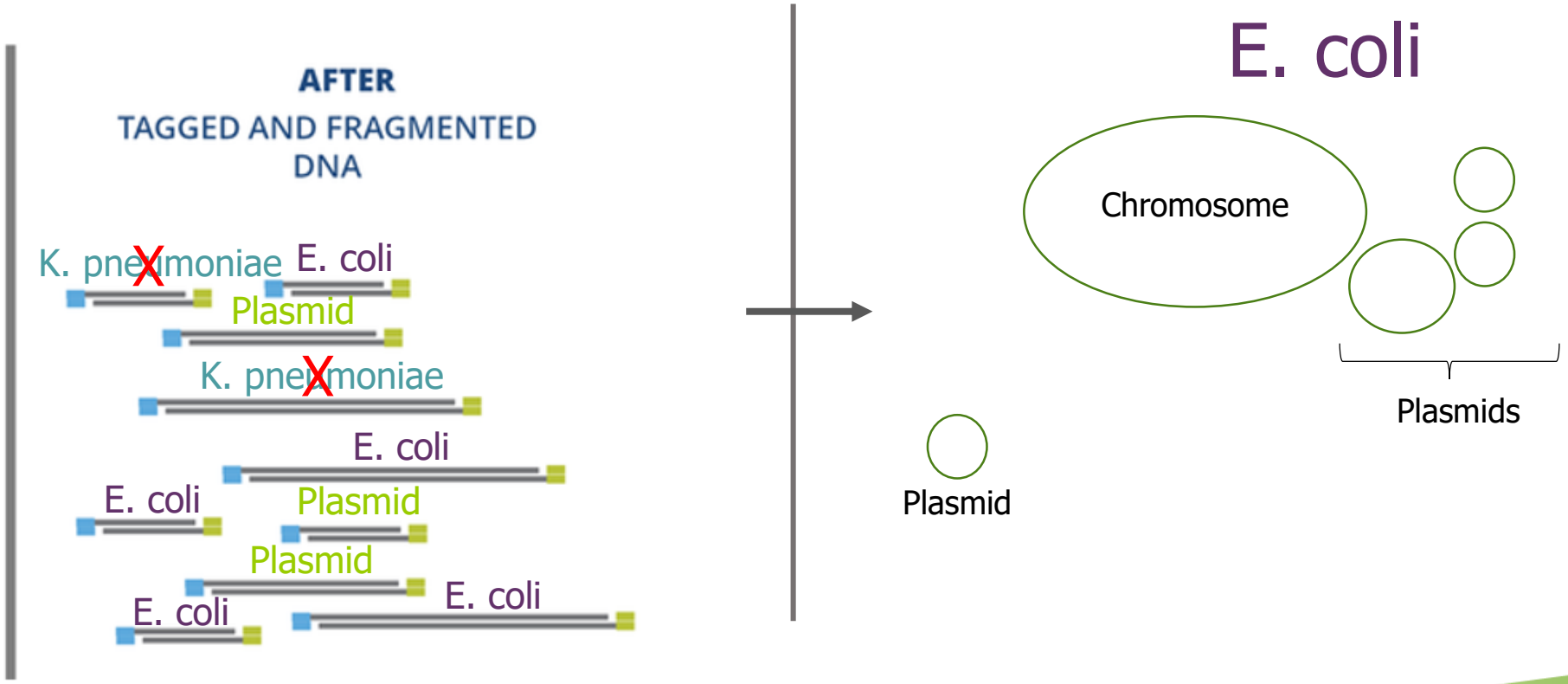
When can you do something about contaminated sequencing data?

- You can always remove **unrelated** sequences if you know what your target sequences are.



When can you do something about contaminated sequencing data?

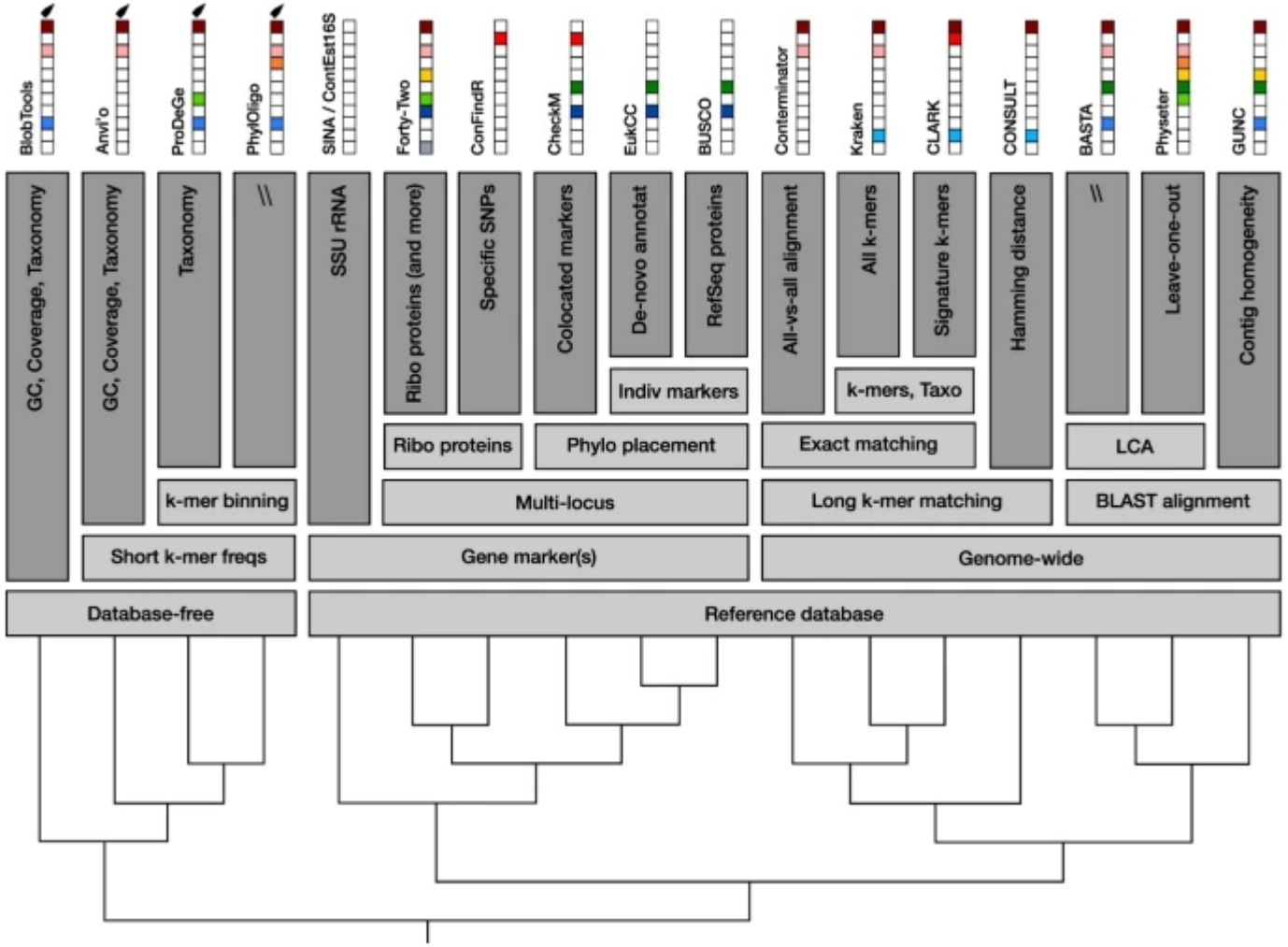
- You can always remove **unrelated** sequences if you know what your target sequences are.



When can you do something about contaminated sequencing data?

- You can always remove **unrelated** sequences if you know what your target sequences are.
- Consider which questions you want to answer with your sequencing data.
- When in doubt, re-sequence.

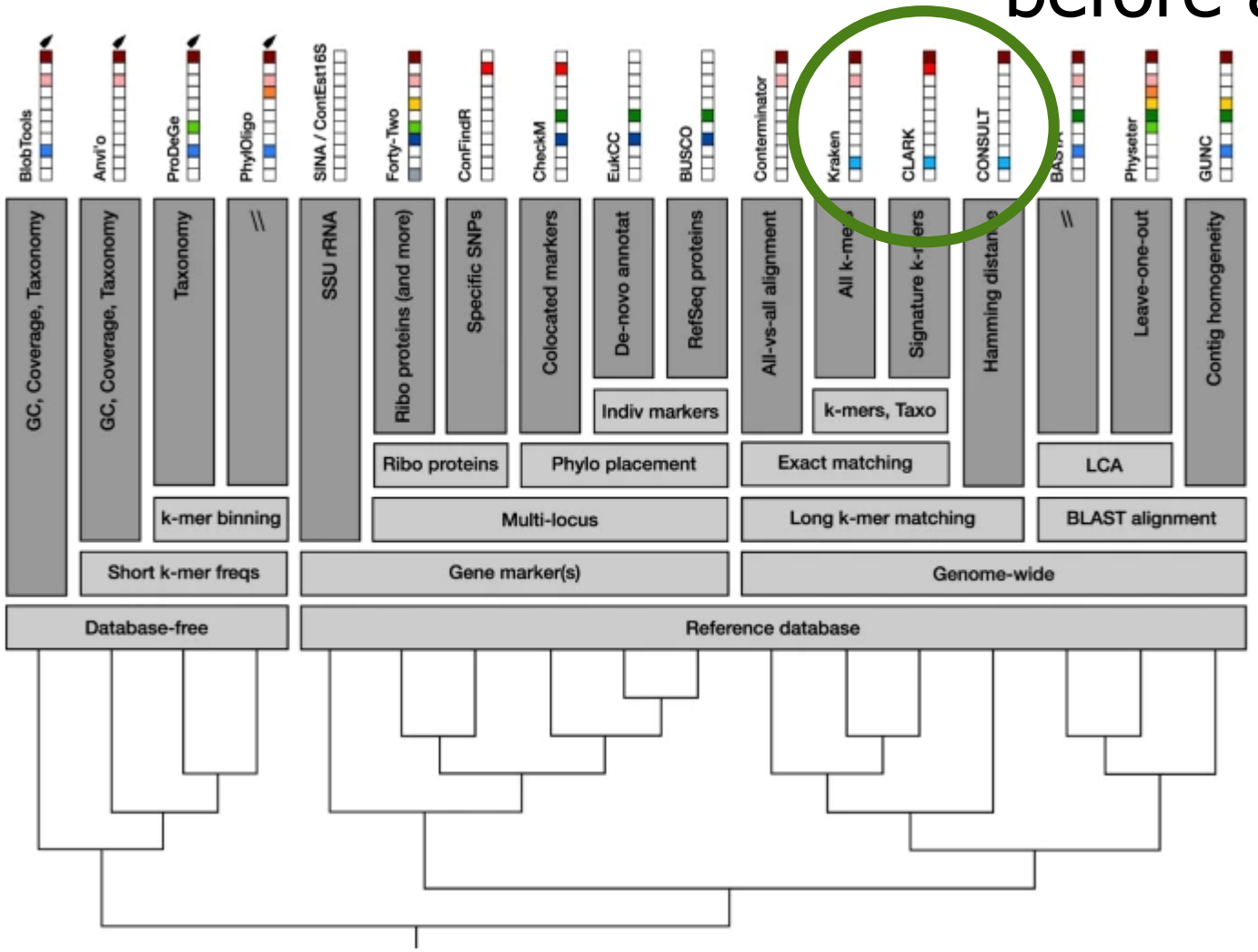
Contamination control tools



Cornet, L., Baurain, D. Contamination detection in genomic data: more is not enough. *Genome Biol* **23**, 60 (2022). https://doi.org/10.1186/s13059-022-02619-9

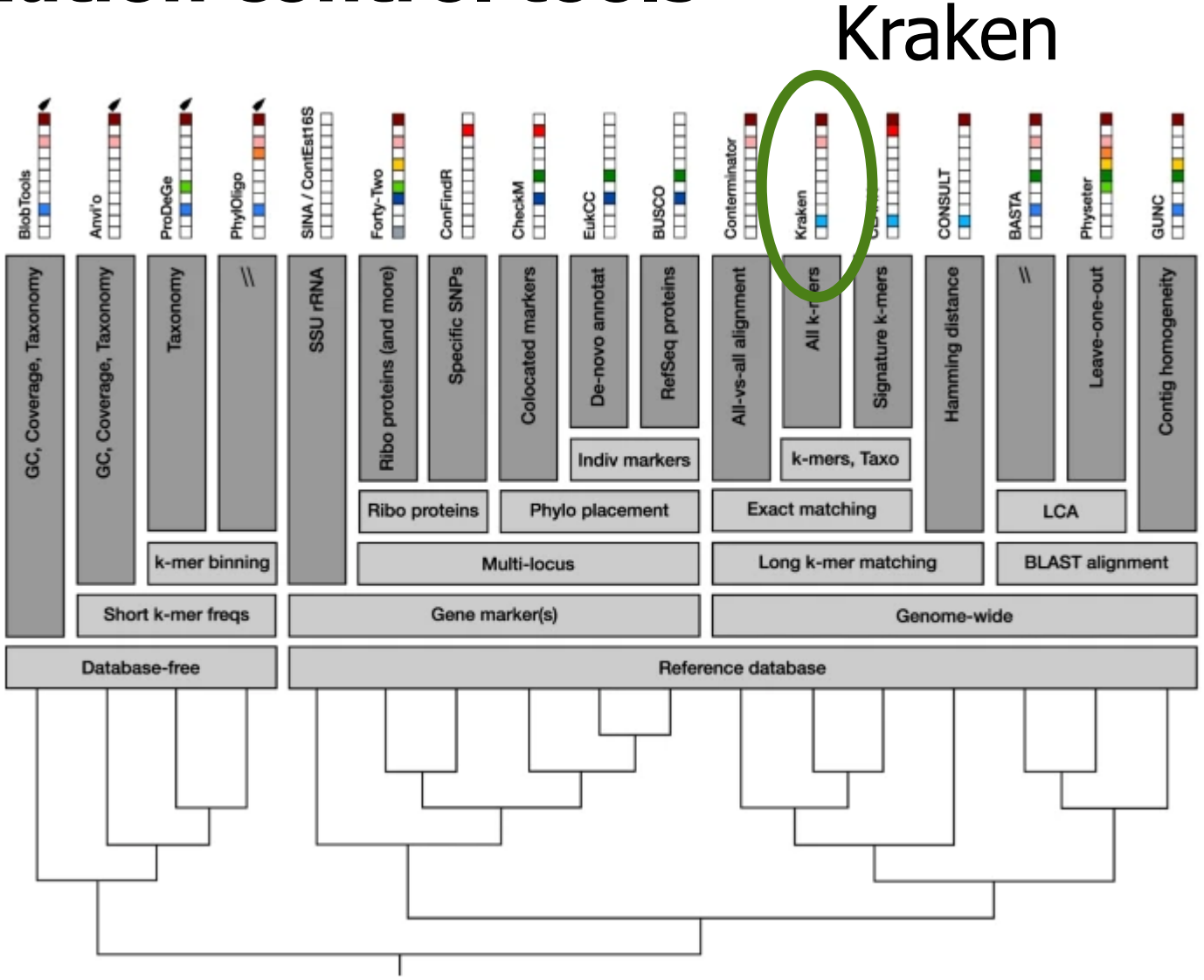
Contamination control tools

Can estimate contamination before assembly



Cornet, L., Baurain, D. Contamination detection in genomic data: more is not enough. *Genome Biol* **23**, 60 (2022). https://doi.org/10.1186/s13059-022-02619-9

Contamination control tools



Cornet, L., Baurain, D. Contamination detection in genomic data: more is not enough. *Genome Biol* **23**, 60 (2022). https://doi.org/10.1186/s13059-022-02619-9

Contamination control tools

Kraken, a taxonomic sequence identifier.

- Exact alignments of k-mers and a novel classification algorithm.
- (https://github.com/DerrickWood/kraken)
- made for short reads.

Kraken 1 | KrakenUniq | Kraken 2 | Kraken2Uniq | MiniKraken

kraken2 --output [output/name] --db [db] --report sample_report --gzip-compressed [input.fastq.gz]

Contamination control tools

KMA

- A mapping method for both short and long reads
- https://bitbucket.org/genomicepidemiology/kma/src/master/

#download fasta files for database
#index database
kma index -i [templates.fsa.gz] -o [database/name]
#run kma
kma -i [someLongReads.fq.gz] -o [output/name] -t_db [database/name] -bcNano -bc 0.7

Part 1

Illumina quality assessment

Practicals

Head to the course page on EVA

Navigate to session one and click into the Part 1 – Illumina Quality Assessment page

GL HF!

Part 2

Nanopore quality assessment

Practicals

Head to the course page on EVA

Navigate to session one and click into the Part 2 –Nanopore quality control page

Enjoy!

Acknowledgements

The creation of this training material was commissioned by ECDC to Statens Serum Institut (SSI) with the direct involvement of Sharmin Baig, Søren Hallstrøm, Kasper Thystrup Karstensen & Astrid Rasmussen