Bridging the gaps in bioinformatics/Raw data QC

Overview of sequencing technologies

February 2025, Søren Hallstrøm, Statens Serum Institut, Denmark

Outline

This session include the following elements

1. The basics of Illumina and Nanopore sequencing
2. Library preparation (similarities and differences)
3. Comparison and overview of the two technologies

Objectives

Specific objectives of this session:

1. Explain the differences between short-read and long-read sequencing technologies
2. Describe Illumina and Nanopore sequencing, and the differences between them
3. Explain shortly about single-end reads vs paired-end reads
4. Outline common reasons for failed sequencing
5. Summarize pros and cons of each sequencing technology

Sequencing technologies used for Surveillance of Infectious Desease

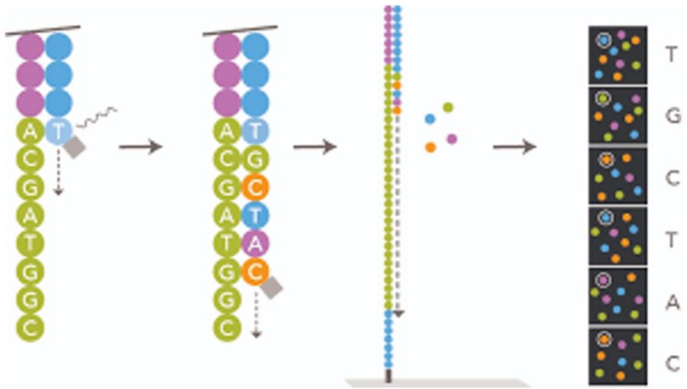
Illumina and Nanopore sequencing platforms

- The most widely used platforms for
    - Outbreak detection
    - Surveillance of infectious desease
    - Genomic epidemiology

Basic differences

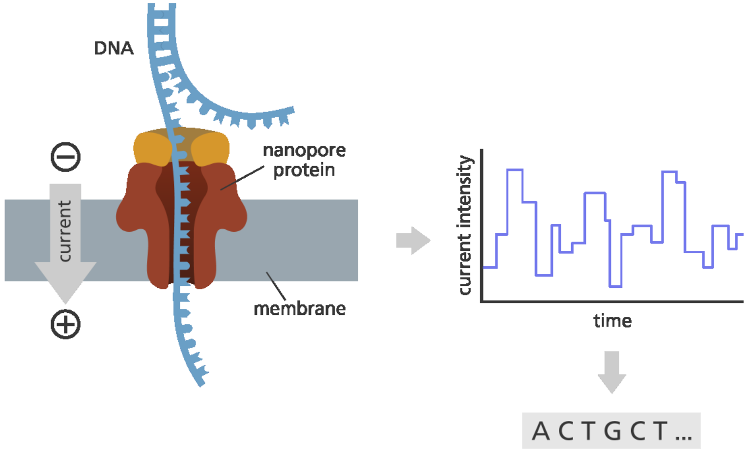
Illumina

- Sequencing by synthesis
- Read length restricted
  - 25-600 bp
- Output: 4 – 40 (100-1000) Gb
- 4 - 56 hours



Nanopore

- Sequencing by nanopores
- Virtually no read length restriction
  - Up to > 2 Mbp
- Output: 2 – 30 (50?) Gb
- Real time (1-2 days)



Main platforms

Illumina

MiSeq

NextSeq 500/550

NovaSeq

NextSeq 1000/2000

Nanopore

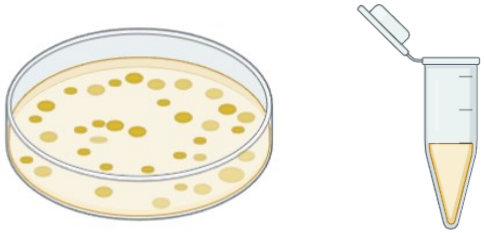
MinION

GridION

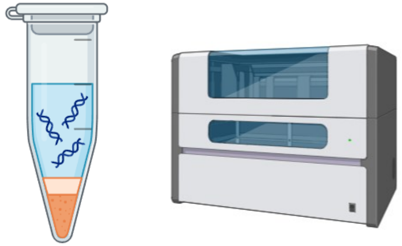
P2 solo

Steps invovled in Next-gen Sequencing

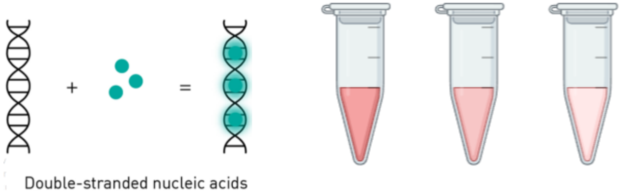
Culture Plating and cell lysis



Nucleotide Extraction

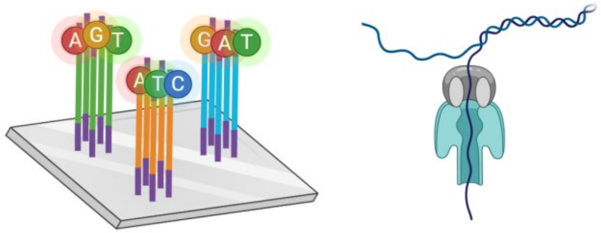


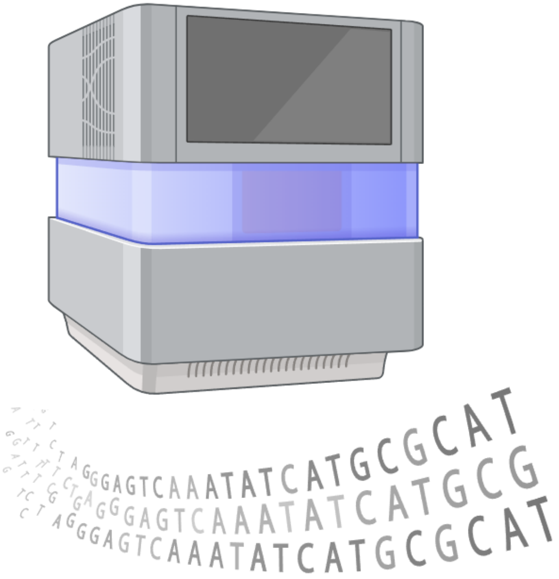
Pre-normalization
Quantify and dilute



Library preparation

Sequencing





Illumina Library preparation and sequencing

Illumina WGS library preparation Basics

1) Sample prepáration
    - gDNA extraction
    - Pre-normalization
2) Library preparation
    - Tagmentation
    - Index PCR
    - Size selection (clean-up)
    - Normalization and pool
3) Sequencing

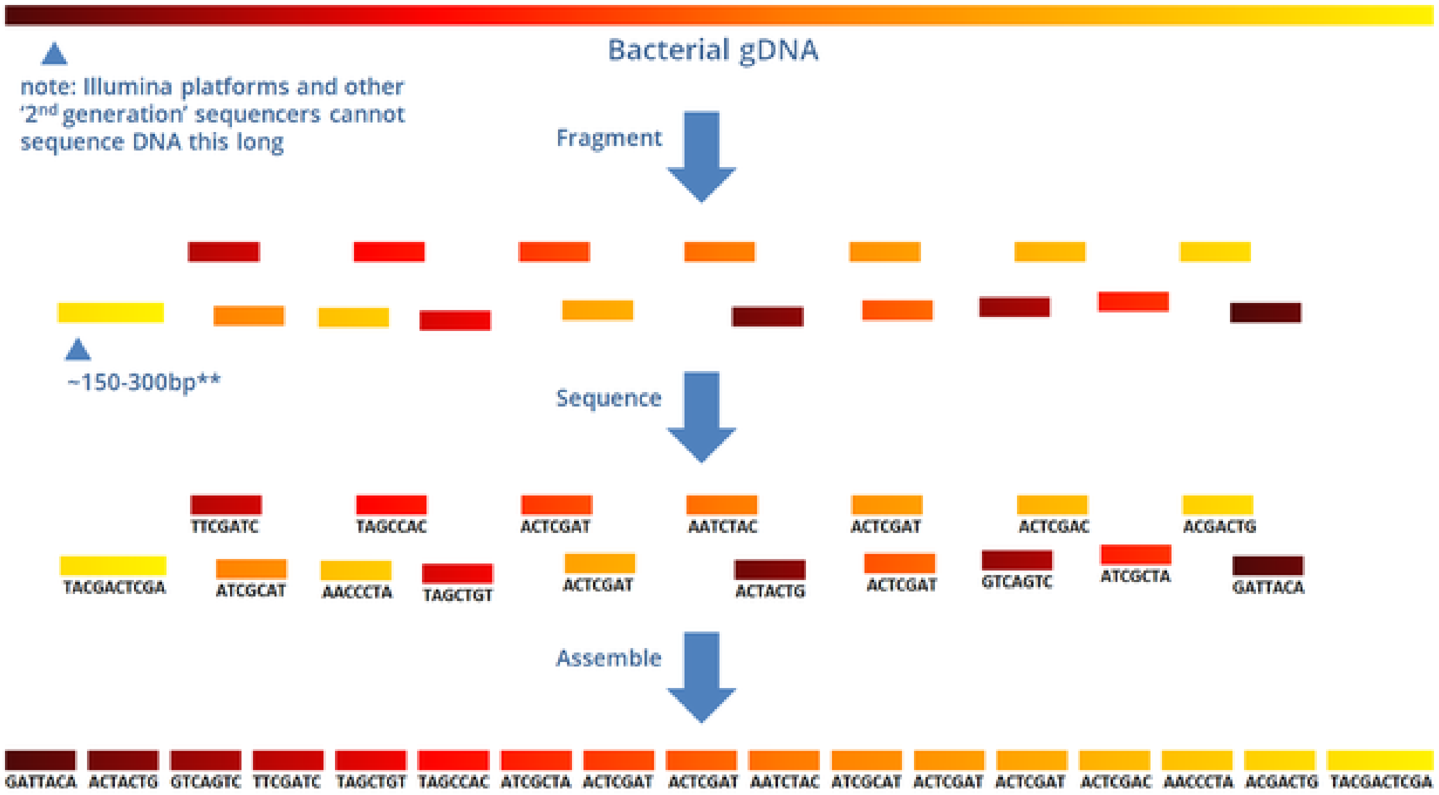


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

Library size selection
Large range of fragment sizes after tagmentation

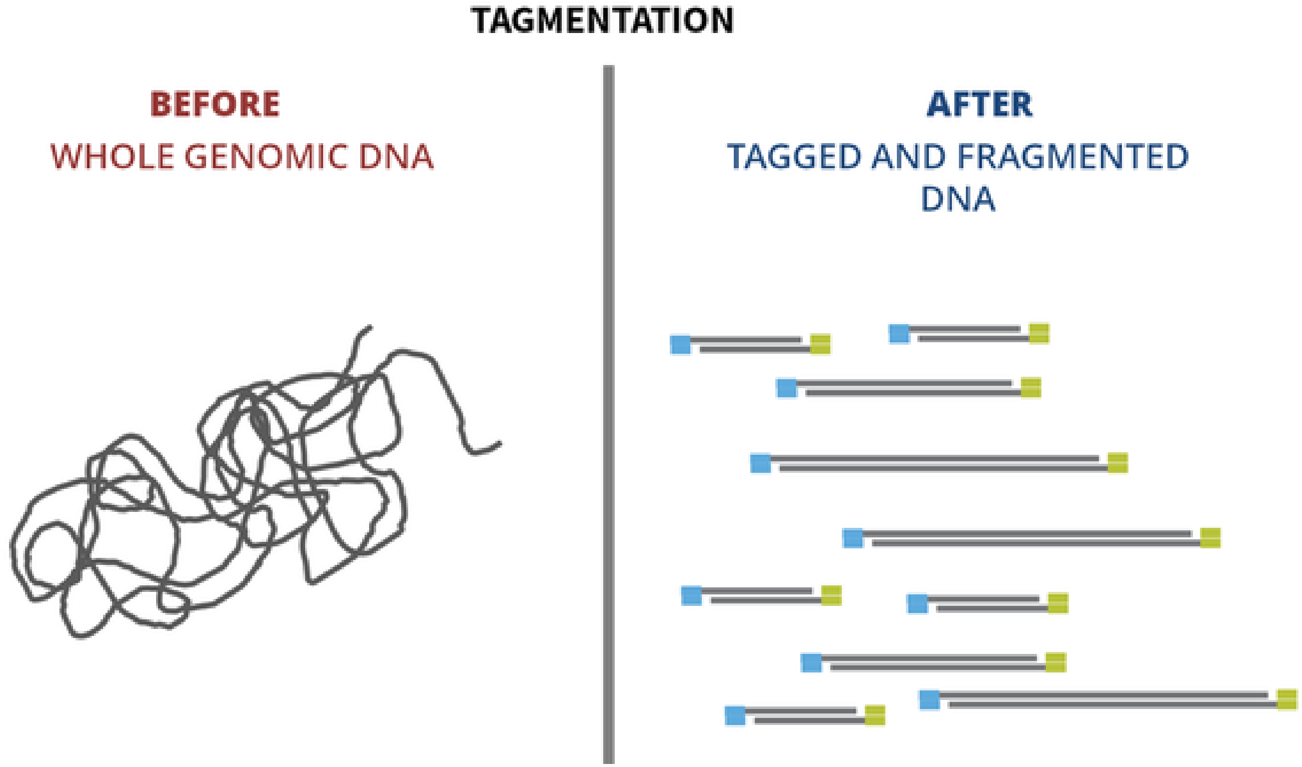


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

Illumina WGS Sequencing

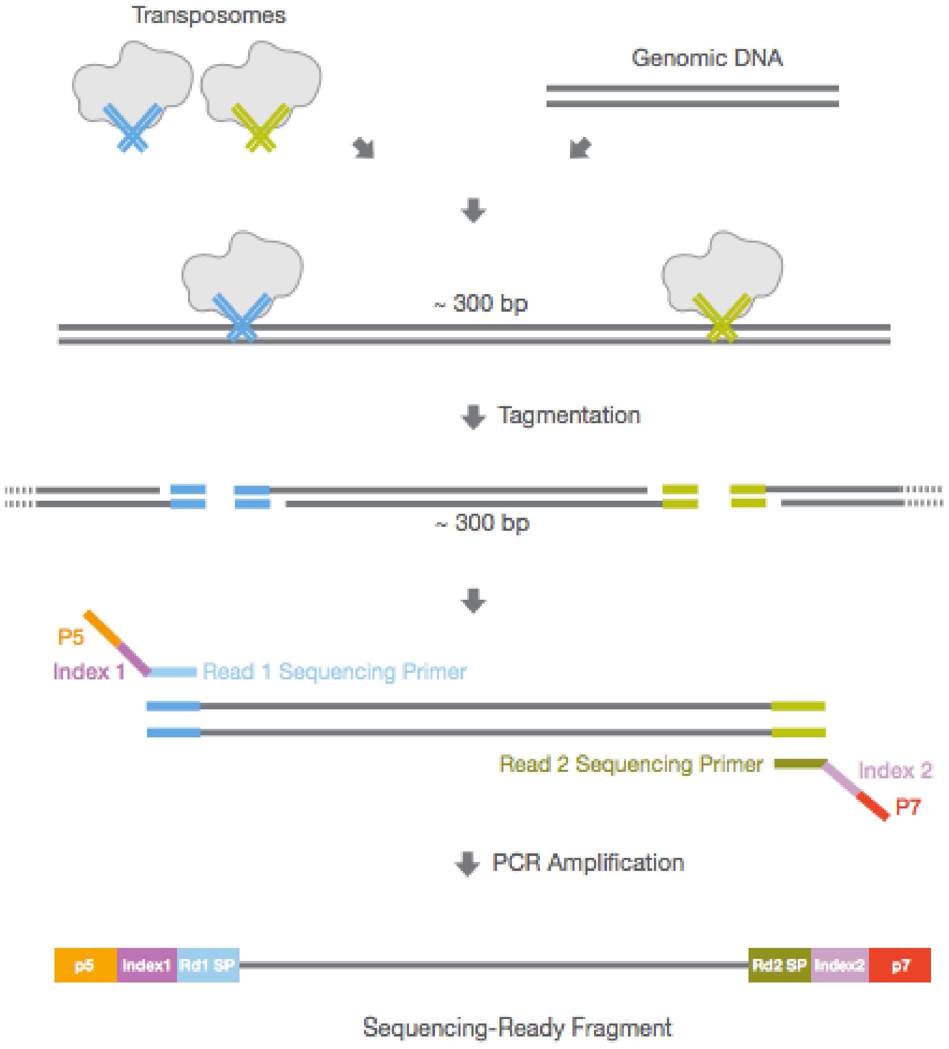
1) Sample prep

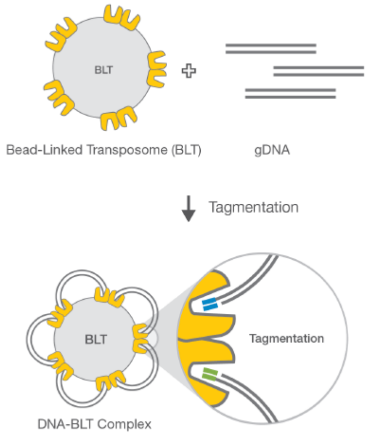
- gDNA extraction
- Pre-normalization

2) Library preparation

- Tagmentation
- Index PCR
- Size selection (clean-up)
- Normalization and pool

3) Sequencing





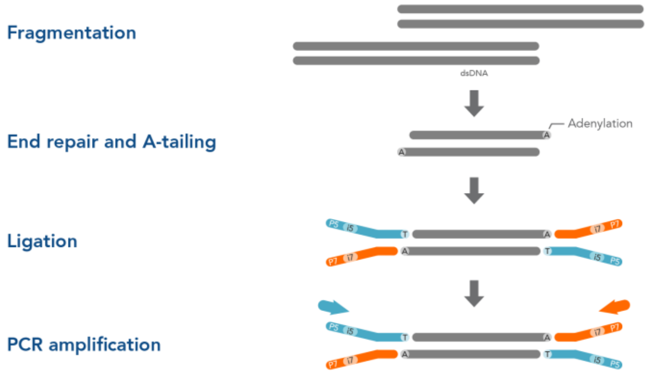
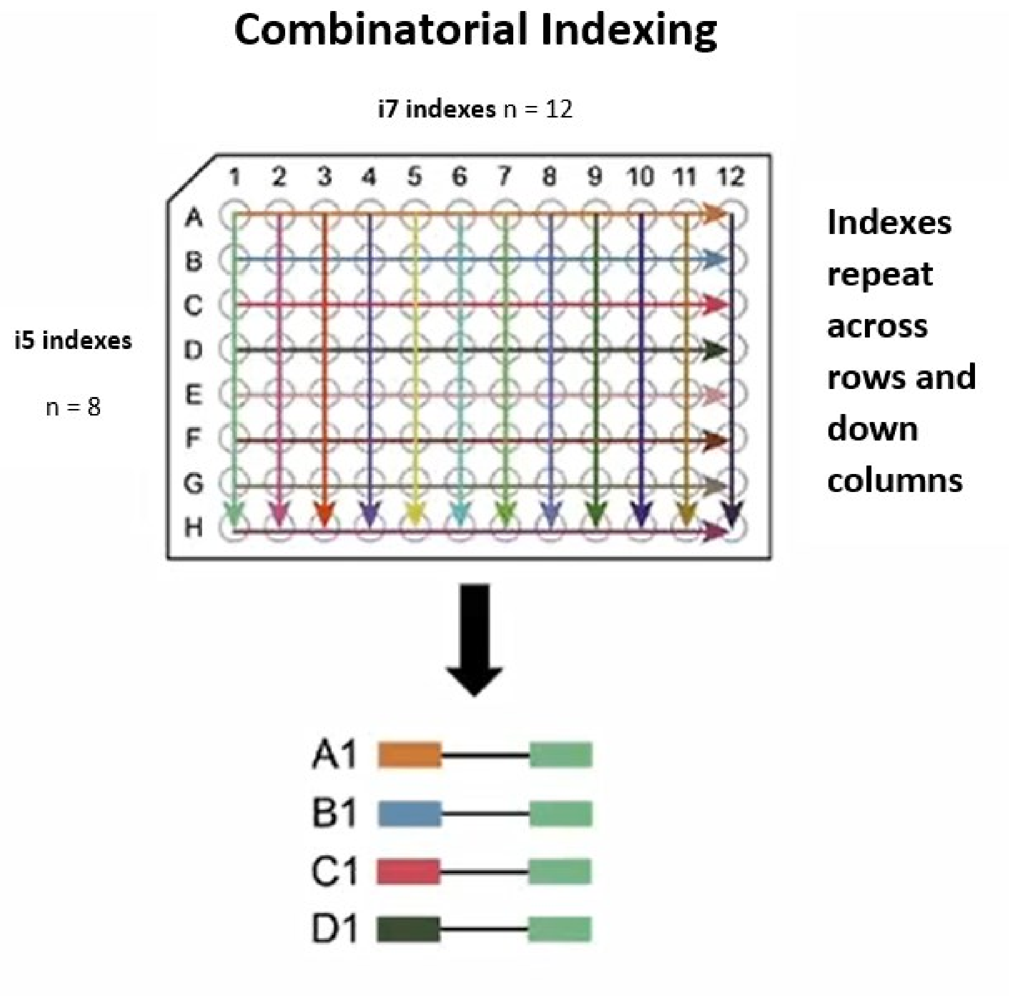


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)
https://eu.idtdna.com/pages/technology/next-generation-sequencing/library-preparation/ligation-based-library-prep

Multiplexing and Indexing



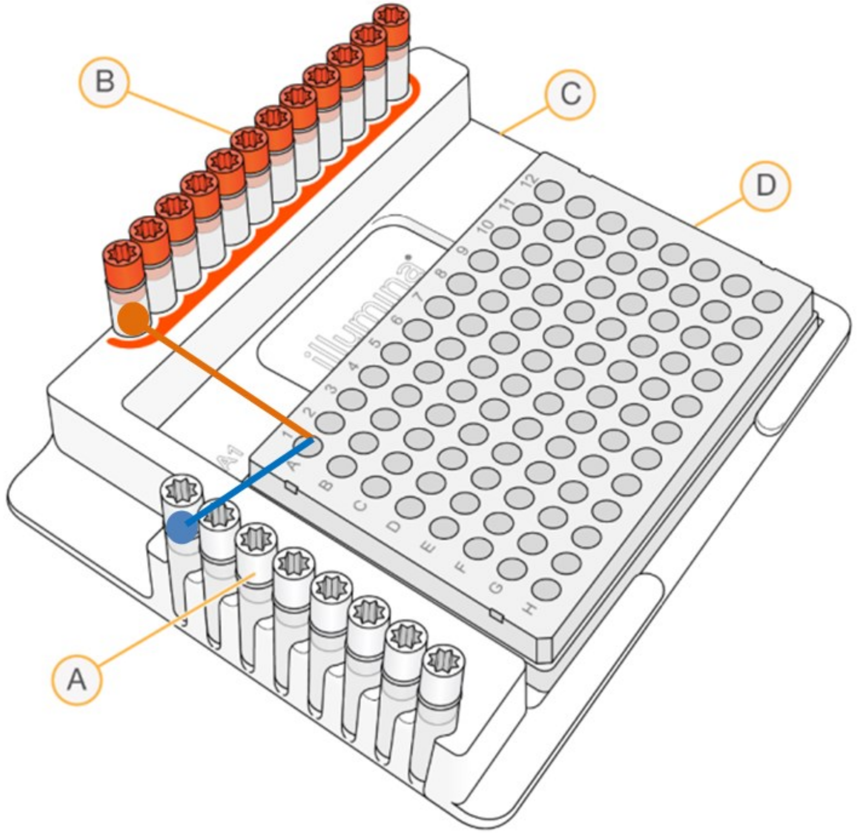
Combinatorial Indexing

i7 indexes n = 12

i5 indexes

n = 8

Indexes repeat across rows and down columns



A Rows A–H: Index 2 (i5) adapters (white caps)

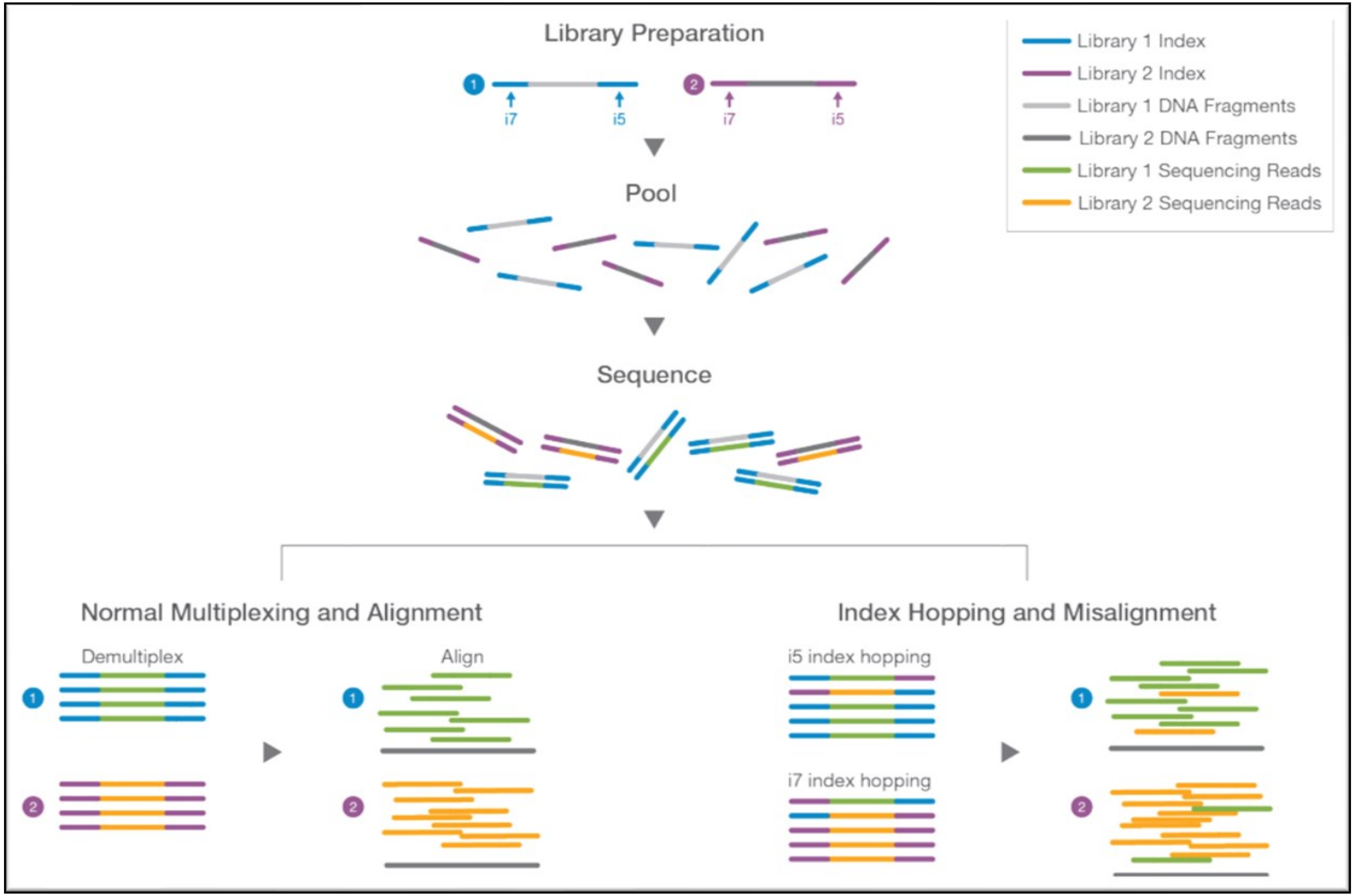
B Columns 1–12: Index 1 (i7) adapters (orange caps)

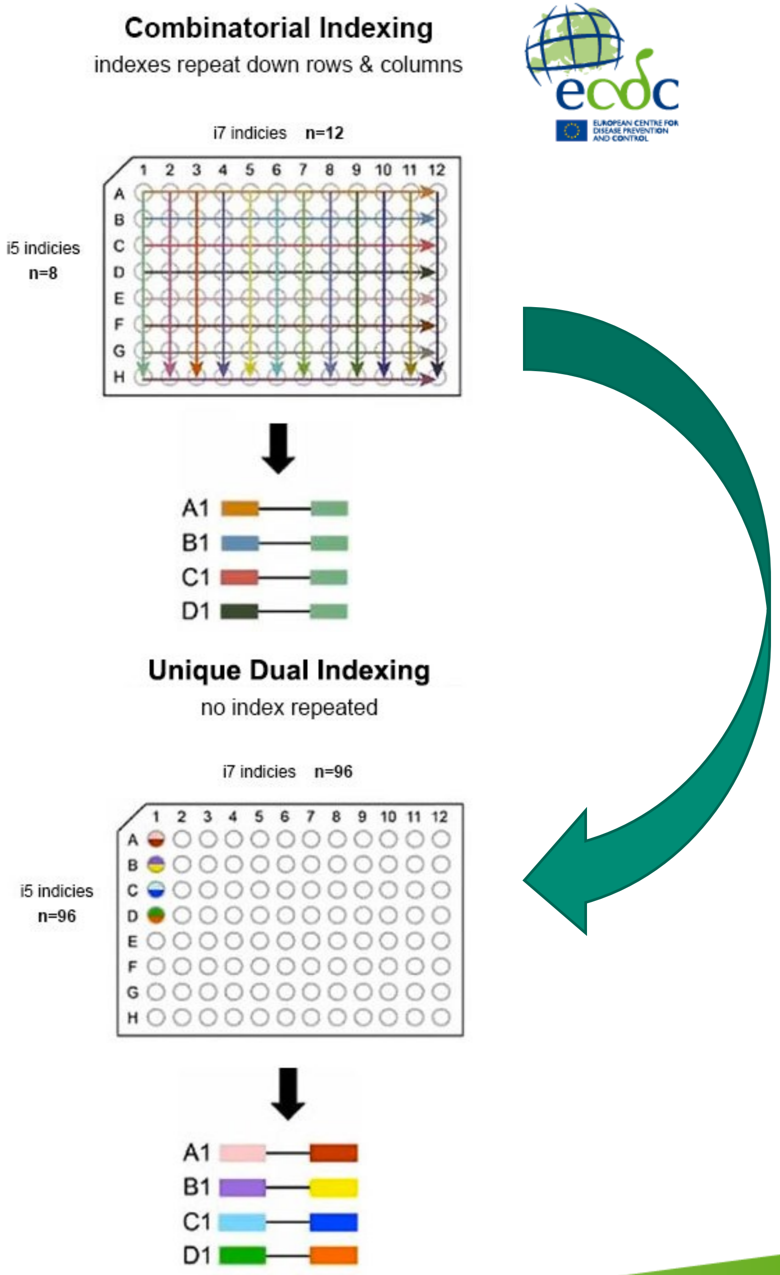
C TruSeq Index Plate Fixture

D Hard-Shell PCR plate

Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

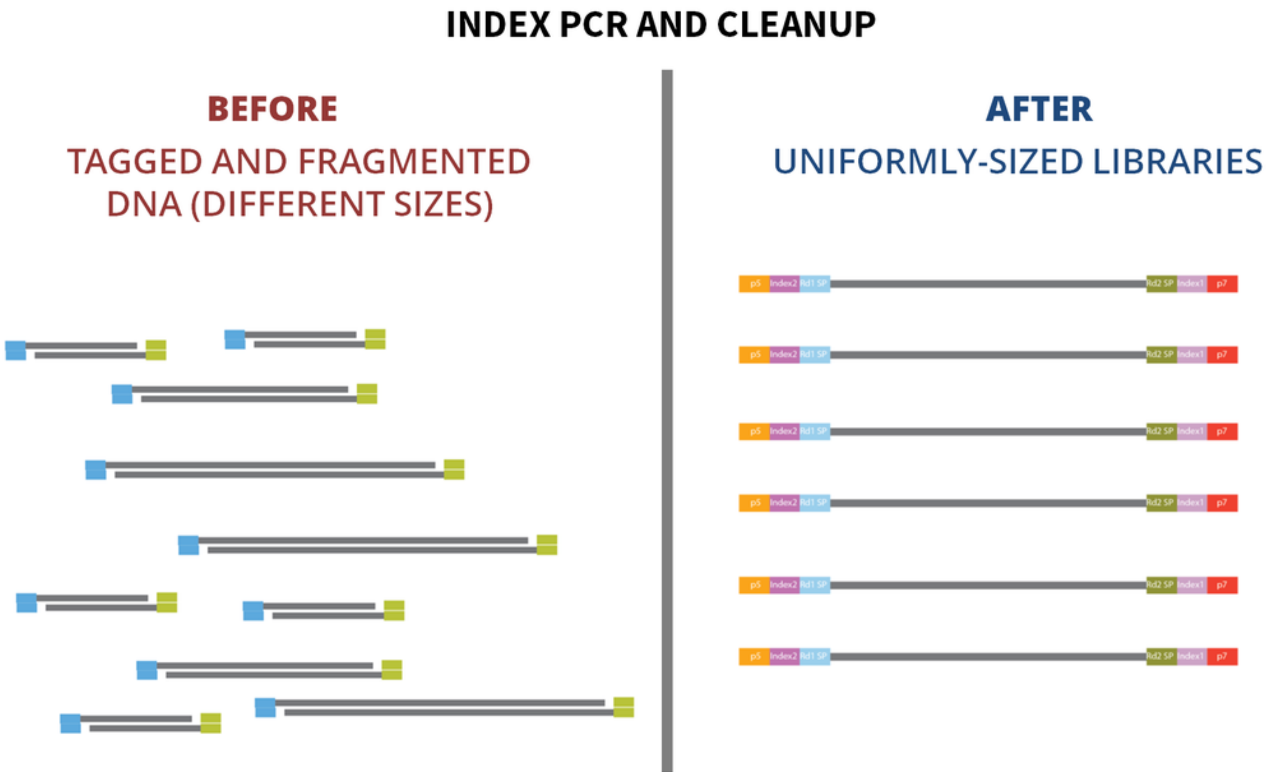
Unique Dual Indexing (UDI)
The issue of index hopping





https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/unique-dual-indexes.html
https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf?linkId=36607862

Library size selection
Bead-based size selection



INDEX PCR AND CLEANUP

BEFORE

TAGGED AND FRAGMENTED DNA (DIFFERENT SIZES)

AFTER

UNIFORMLY-SIZED LIBRARIES

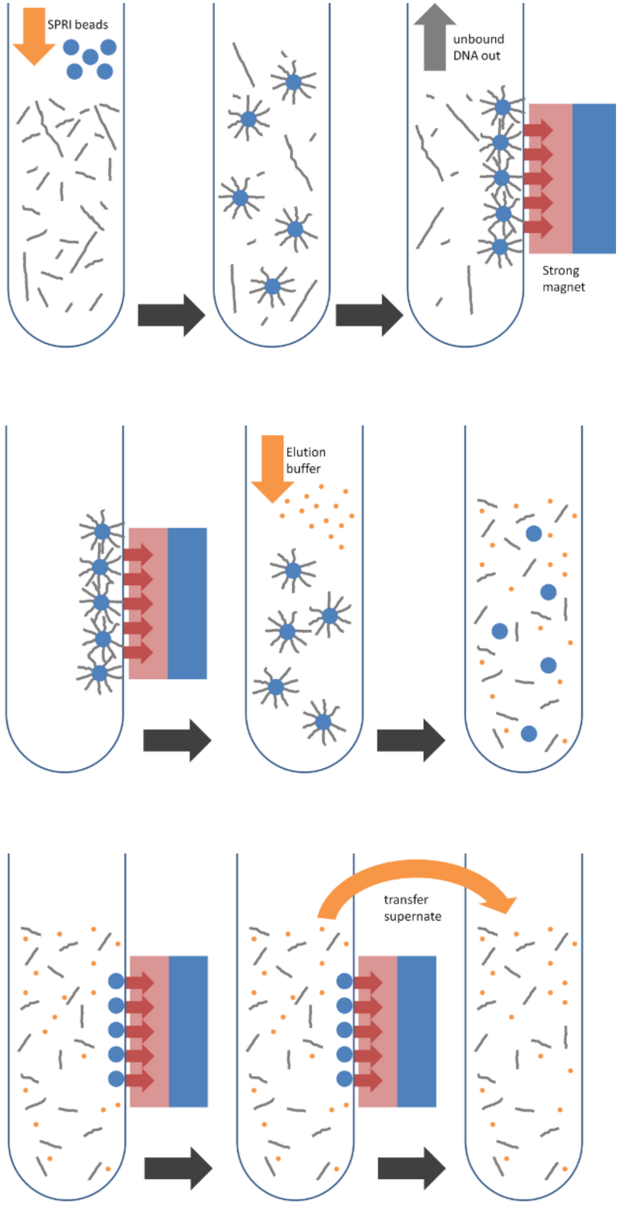
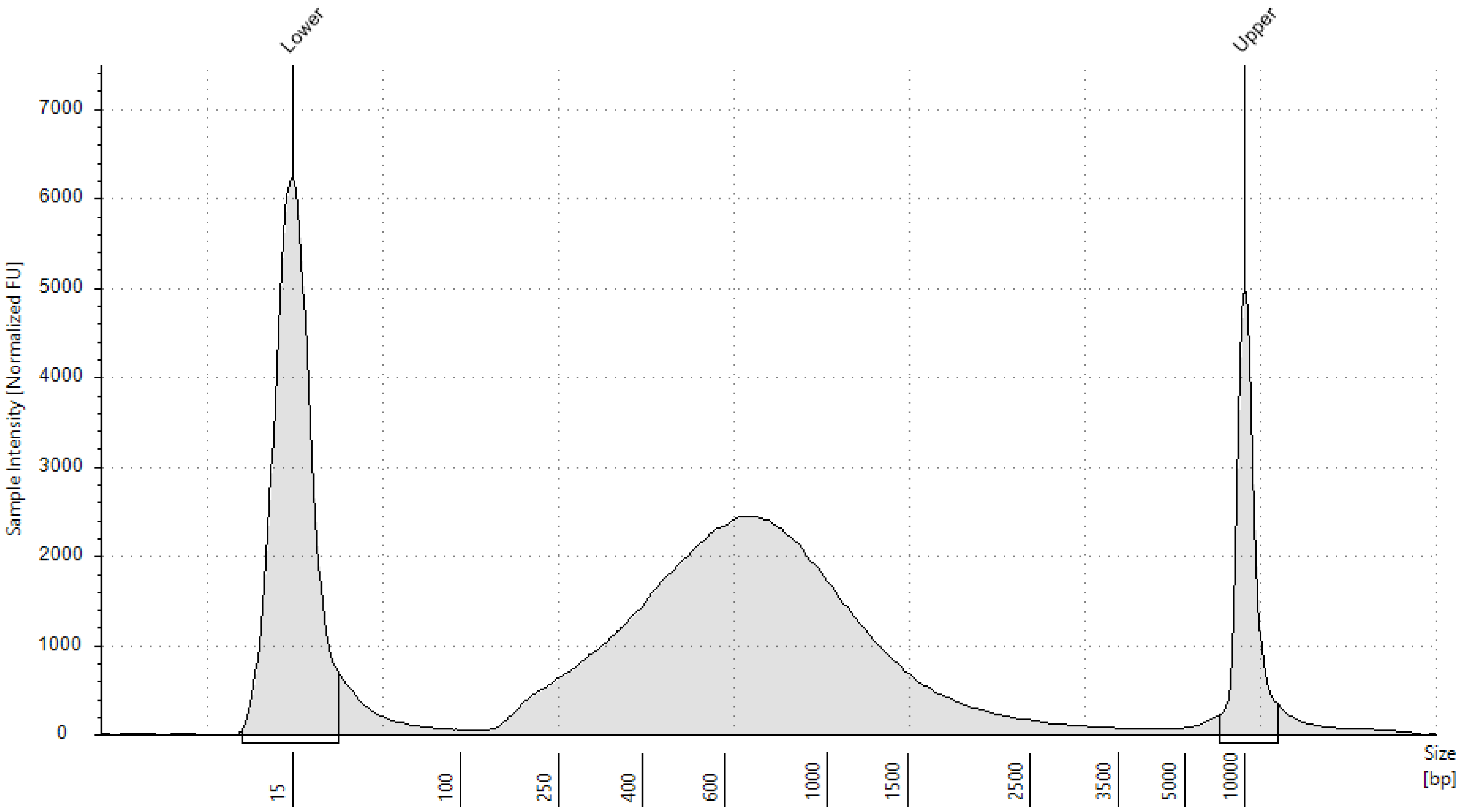


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

Library size selection
Resulting fragment distribution



Illumina cluster generation

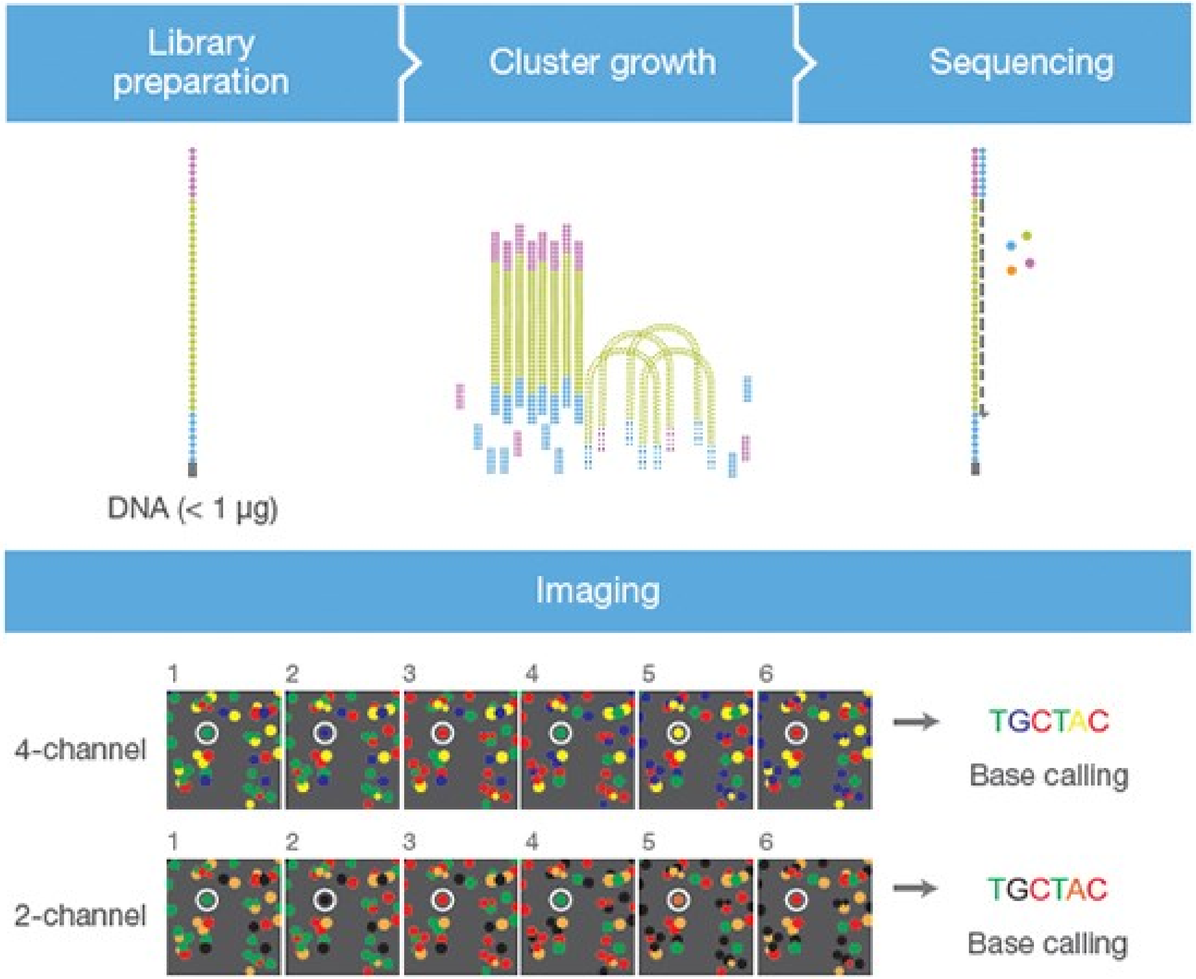
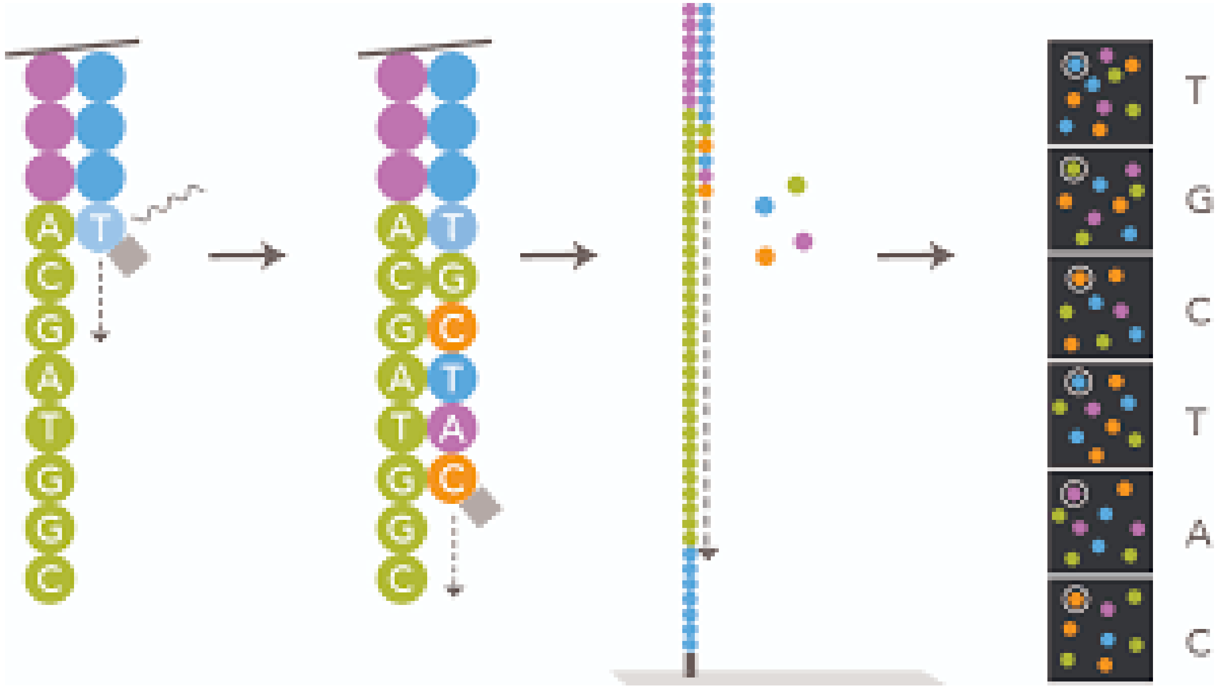
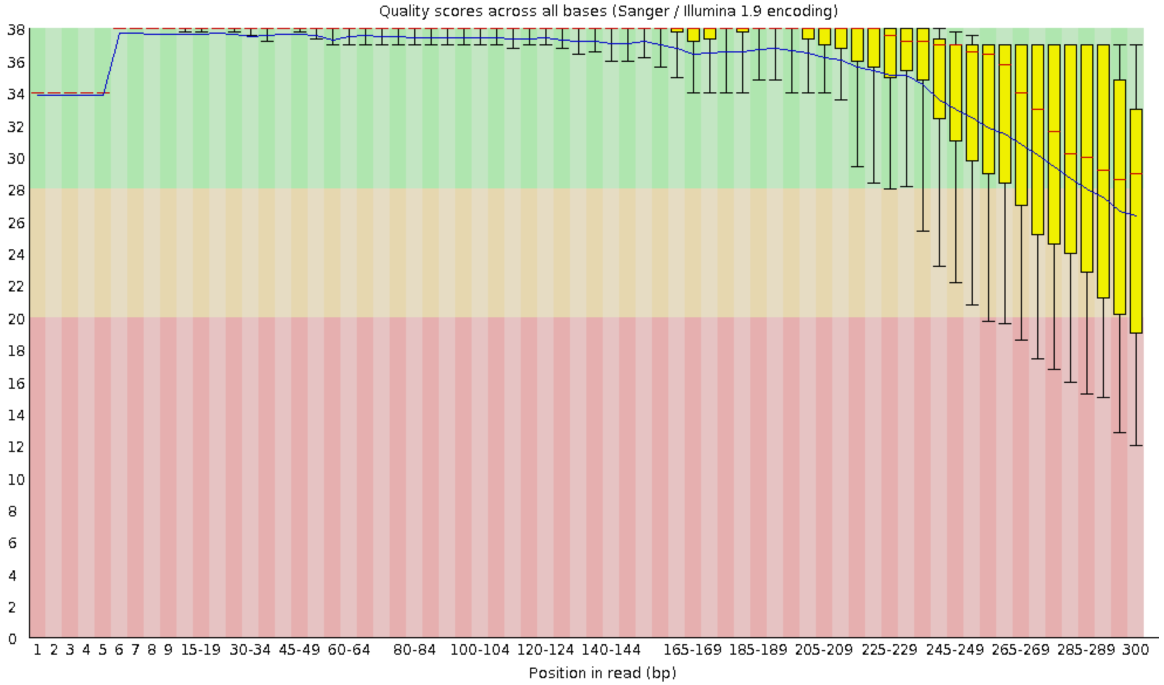


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

Sequencing by synthesis

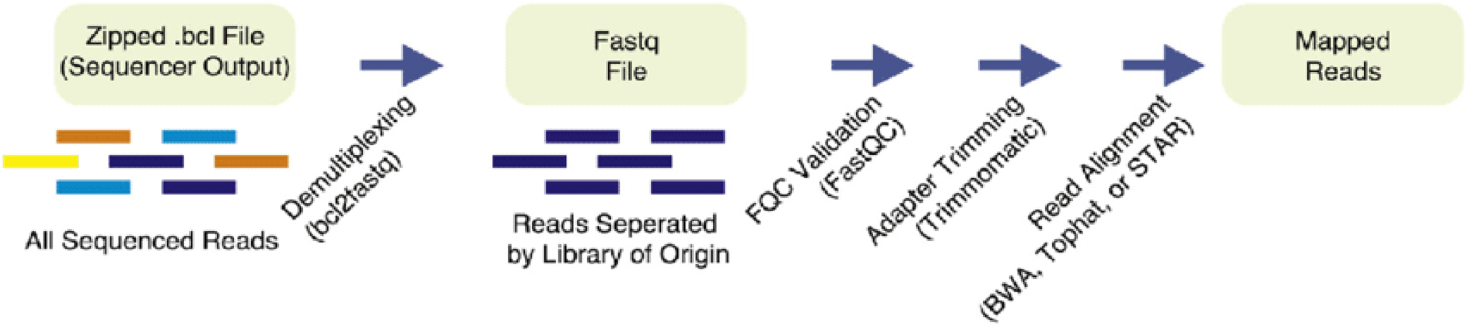




Illumina sequence data files

Illumina sequencer generates .bcl

Translated to fastq file format on the machine using bcl2fastq



Zipped .bcl File (Sequencer Output)

All Sequenced Reads

Demultiplexing (bcl2fastq)

Fastq File

Reads Seperated by Library of Origin

FQC Validation (FastQC)

Adapter Trimming (Trimmomatic)

Read Alignment (BWA, Tophat, or STAR)

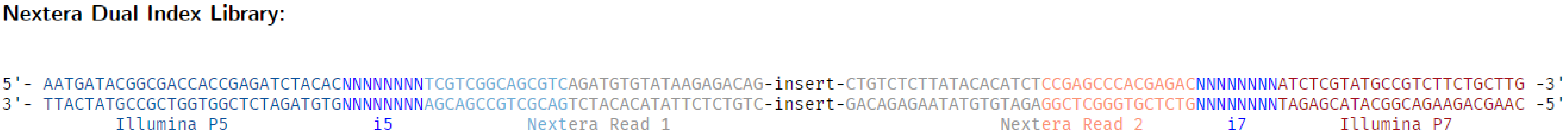
Mapped Reads

Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

Illumina sequencing constructs Nomenclature

Illumina adapters: P5 and P7

Illumina indices: Index1 (i7) and Index2 (i5)



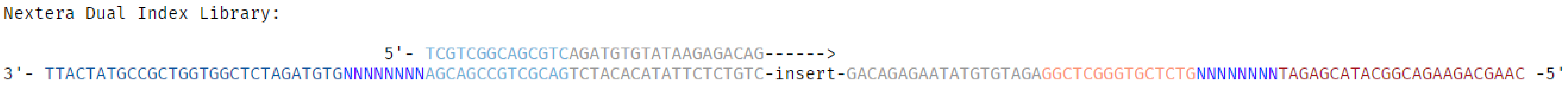
Nextera Dual Index Library:

```
5'- AATGATACGGCGACCACCGAGATCTACACNNNNNNNNTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-insert-CTGTCTCTTATACACATCTCCGAGCCCACGAGACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG -3'
3'- TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNNAGCAGCCGTCGCAGTCTACACATATTCTCTGTC-insert-GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC -5'
           Illumina P5               i5             Nextera Read 1                                 Nextera Read 2        i7          Illumina P7
```

https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html

Illumina sequencing
Four reads

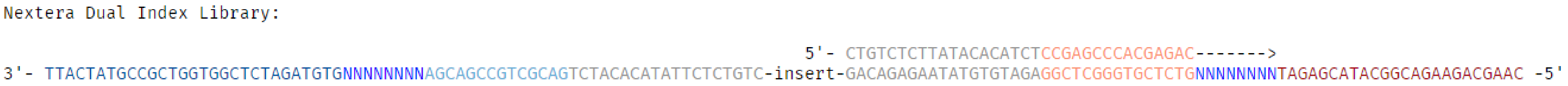
1) Read1



```
Nextera Dual Index Library:

                                     5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG------>
3'- TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNNAGCAGCCGTCGCAGTCTACACATATTCTCTGTC-insert-GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC -5'
```

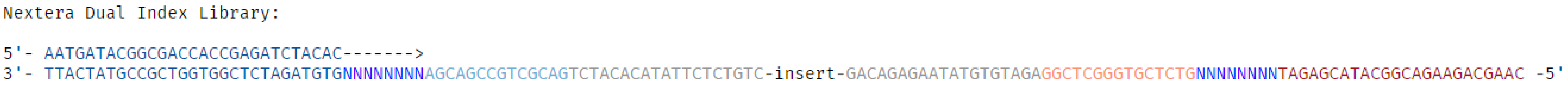
2) Index1 – i7



```
Nextera Dual Index Library:

                                                                              5'- CTGTCTCTTATACACATCTCCGAGCCCACGAGAC------->
3'- TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNNAGCAGCCGTCGCAGTCTACACATATTCTCTGTC-insert-GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC -5'
```

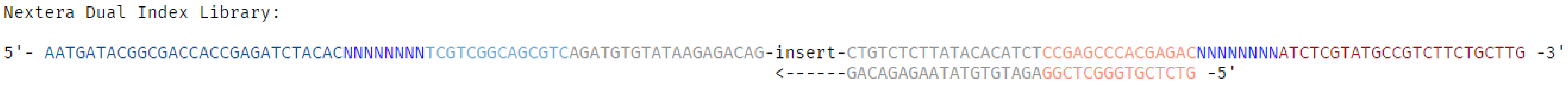
3) Index2 – i5



```
Nextera Dual Index Library:

5'- AATGATACGGCGACCACCGAGATCTACAC------->
3'- TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNNAGCAGCCGTCGCAGTCTACACATATTCTCTGTC-insert-GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC -5'
```

4)Read2

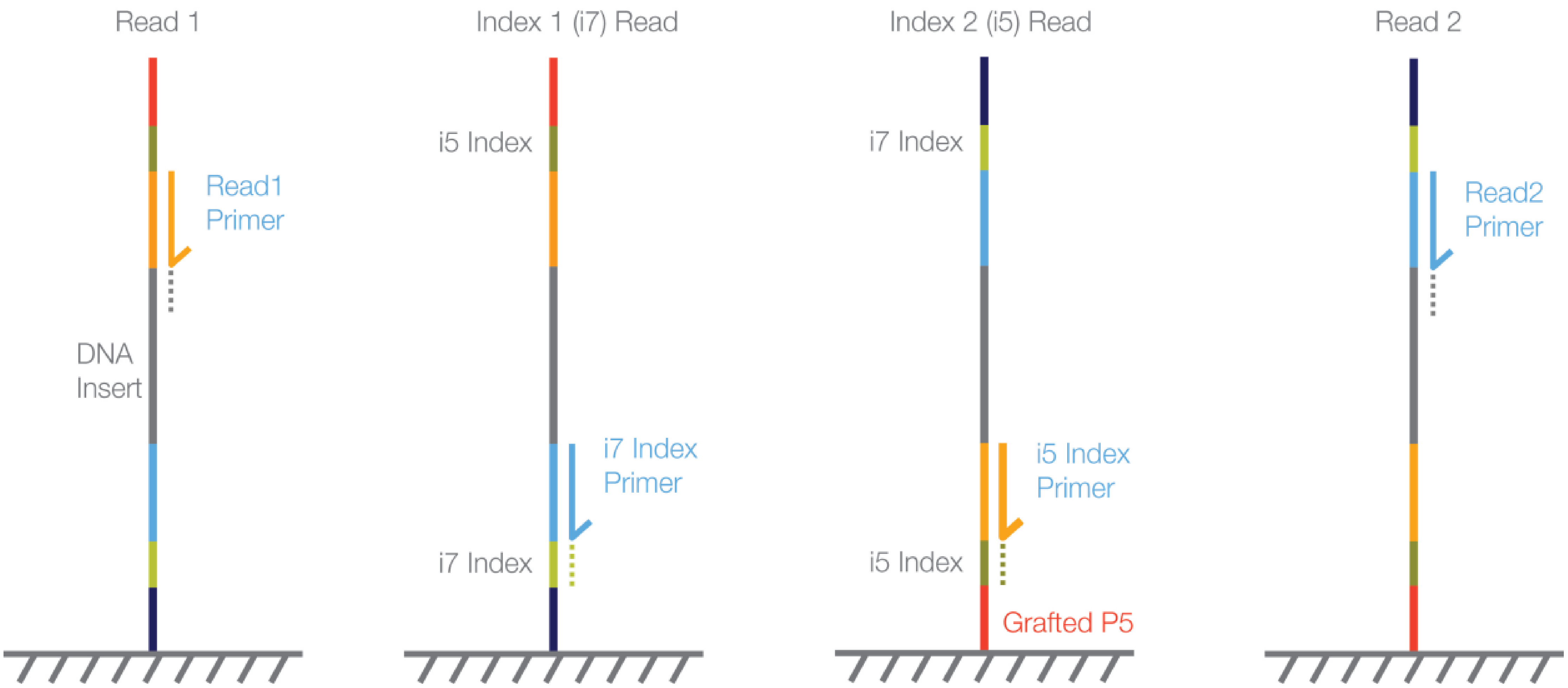


```
Nextera Dual Index Library:

5'- AATGATACGGCGACCACCGAGATCTACACNNNNNNNNTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-insert-CTGTCTCTTATACACATCTCCGAGCCCACGAGACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG -3'
                                                                           <------GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG -5'
```

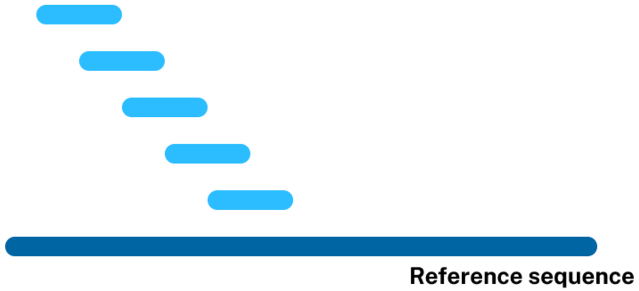
https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html

Illumina sequencing
Read orientation on the flowcell



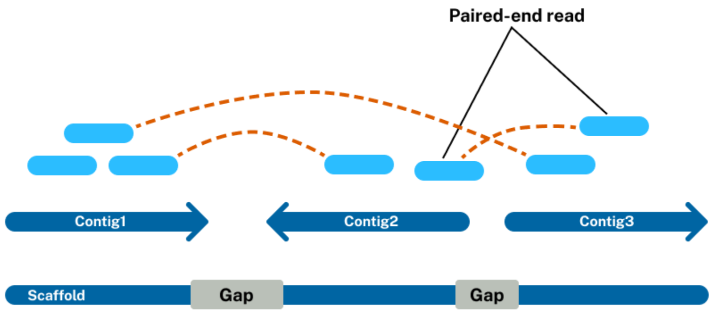
Illumina data types
Single-end vs Paired-end reads

Single-end reads

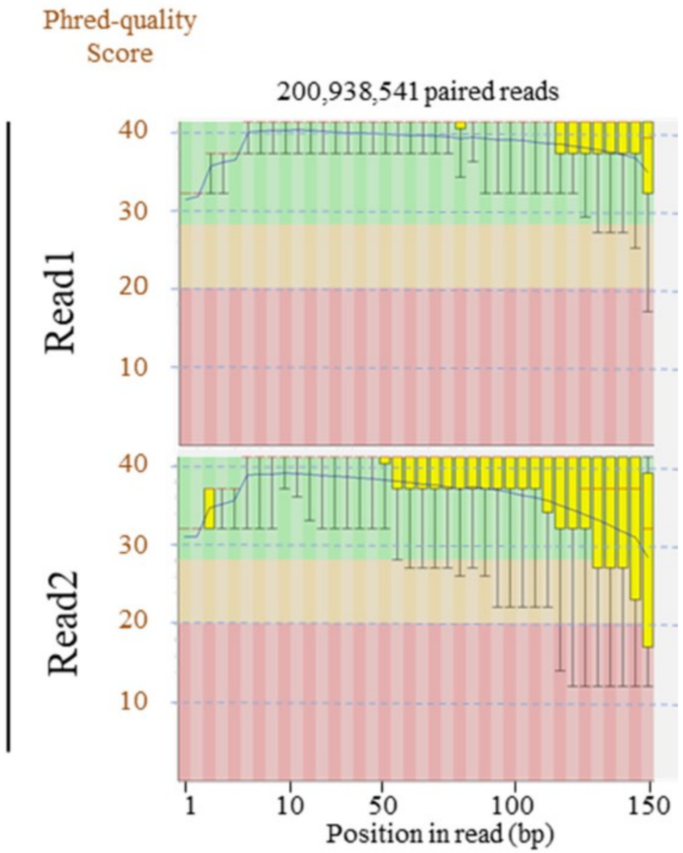


sample.fastq

Paired-end reads



Sample_R1.fastq
Sample_R2.fastq



https://du-bii.github.io/module-5-Methodes-Outils/seance1_NGS/slides.html#14
https://www.geneious.com/guides/assembling-dna-sequences

Sequencing by Synthesis

Pros ✓

Massive Parallel Sequencing

High data yield

High multiplex capacity

Possibility for paired end reads

Cons !

Read length restricted by the chemistry

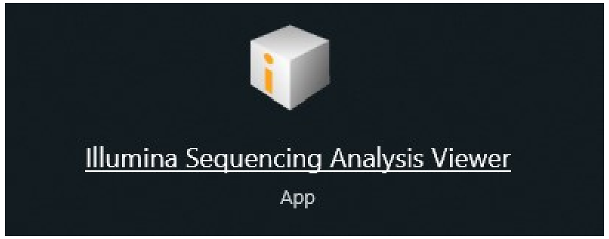
Quality drops during the strand synthesis – More pronounced in Read2

Data only available after run completion

Troubleshooting an Illumina sequencing run

Illumina Sequence Analysis Viewer (SAV)

The Illumina Sequence Analysis Viewer (SAV)

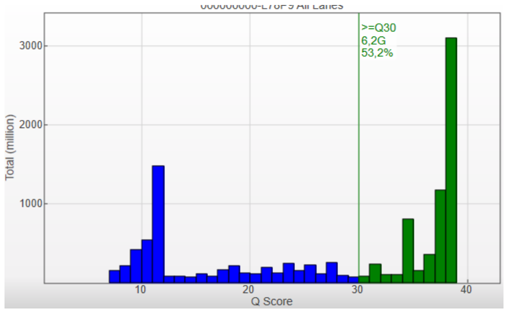


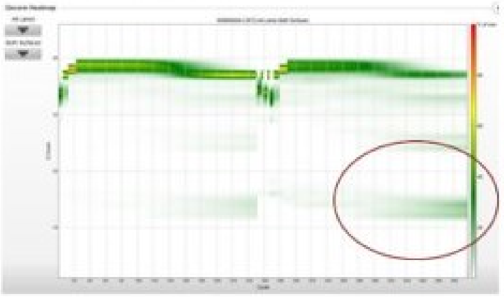
Illumina Sequencing Analysis Viewer

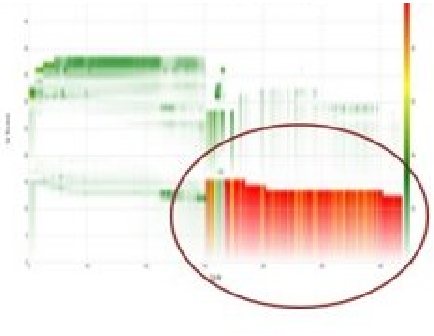
App

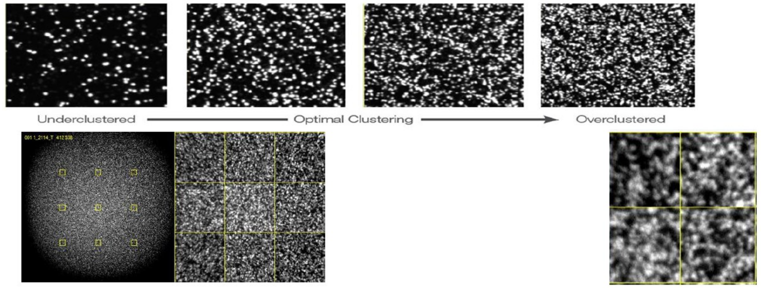
Evaluate key parameters

- Q30 data (Gb and %)
- Cluster density
- Reads passing filter (%)









Troubleshooting an Illumina sequencing run
Phred score – Quality of base call (Q-score)

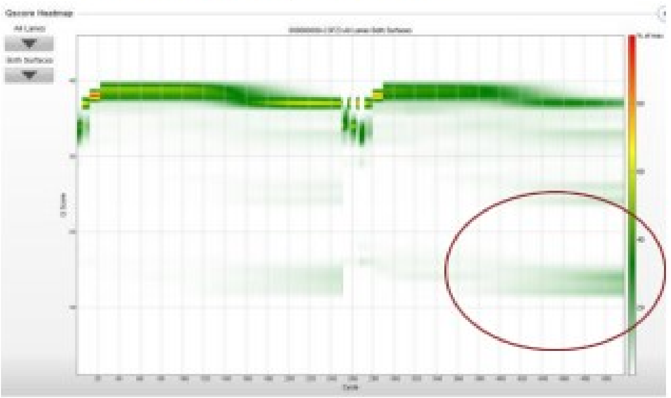
Q score is a quality indicator for individual reads

Log-scale -> Q score of 30 = 1 in 1000 may be incorrect

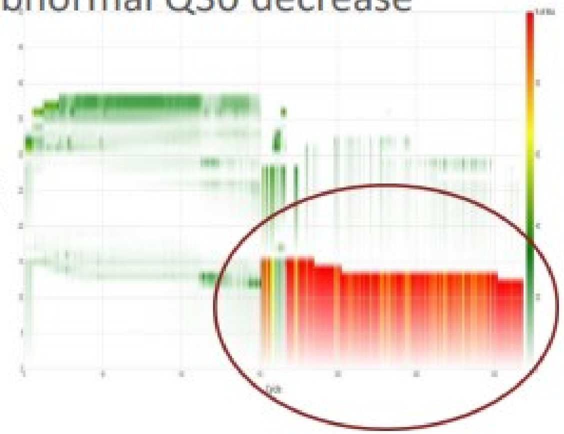
The longer the read length the lower the Q30 percentage

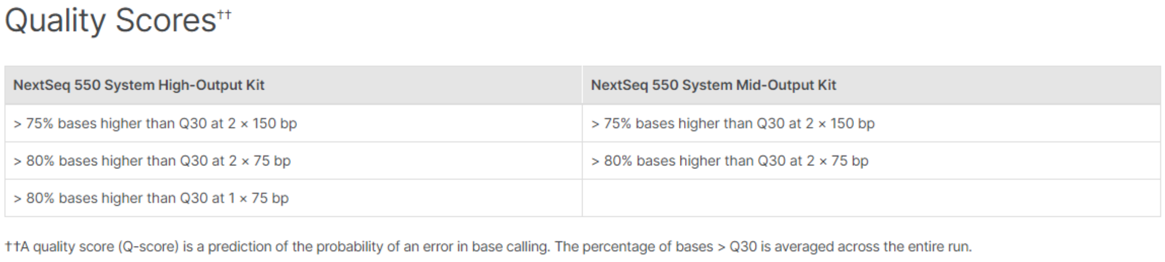
(Due to sequencing chemistry)

Normal drop in Q30



Abnormal Q30 decrease





Quality Scores[††]

NextSeq 550 System High-Output Kit	NextSeq 550 System Mid-Output Kit
> 75% bases higher than Q30 at 2 × 150 bp	> 75% bases higher than Q30 at 2 × 150 bp
> 80% bases higher than Q30 at 2 × 75 bp	> 80% bases higher than Q30 at 2 × 75 bp
> 80% bases higher than Q30 at 1 × 75 bp	

[††]A quality score (Q-score) is a prediction of the probability of an error in base calling. The percentage of bases > Q30 is averaged across the entire run.

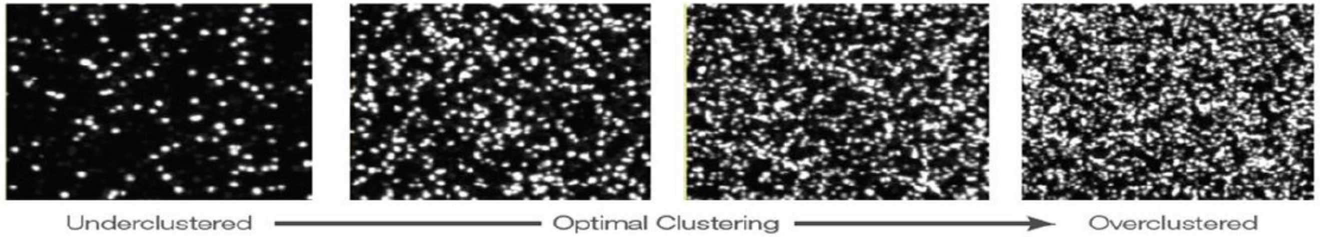
Troubleshooting an Illumina sequencing run
Cluster density and Passing filter %

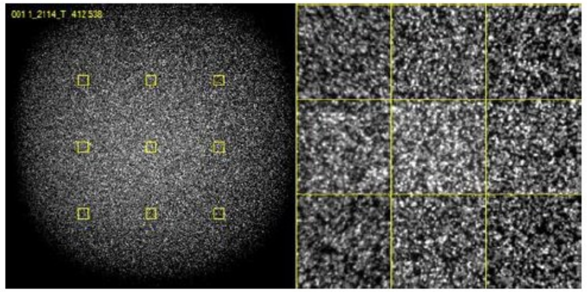
Cluster density is a measurement of how tight the clustering is on the flow cell
For each Illumina platform and kit chemistry a recommended cluster density is provided
Cluster density is linked to Clusters Passing filter (PF %)
Over clustering -> low data quality (low PF % but maybe higher data yield)
Under clustering -> high data quality (High PF % but lower data yield)

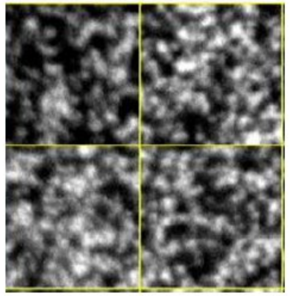
NB:
Imagine is a physical process
There is a trade off and optimization can often be required

Actual images from the sequencer

- can be accessed through SAV





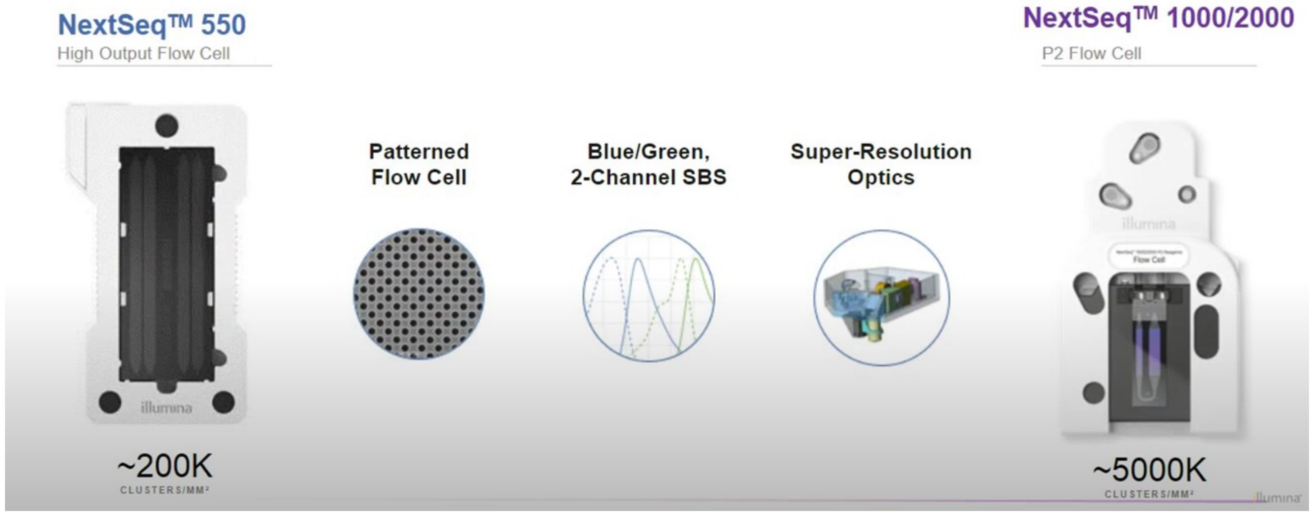


Troubleshooting an Illumina sequencing run NextSeq 1000/2000 (Illuminas new line of sequencers)

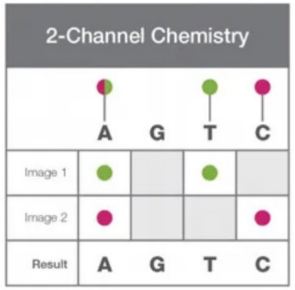
Cluster density and PF% **are not key parameters**

The reason

Patterned flow cell technolgy

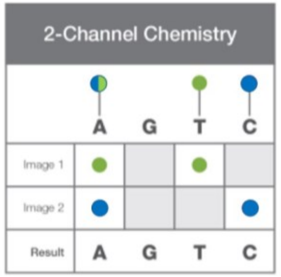


NextSeq 500/550



2-Channel Chemistry	A	G	T	C
Image 1	●		●	
Image 2	●			●
Result	A	G	T	C

NextSeq 1000/2000



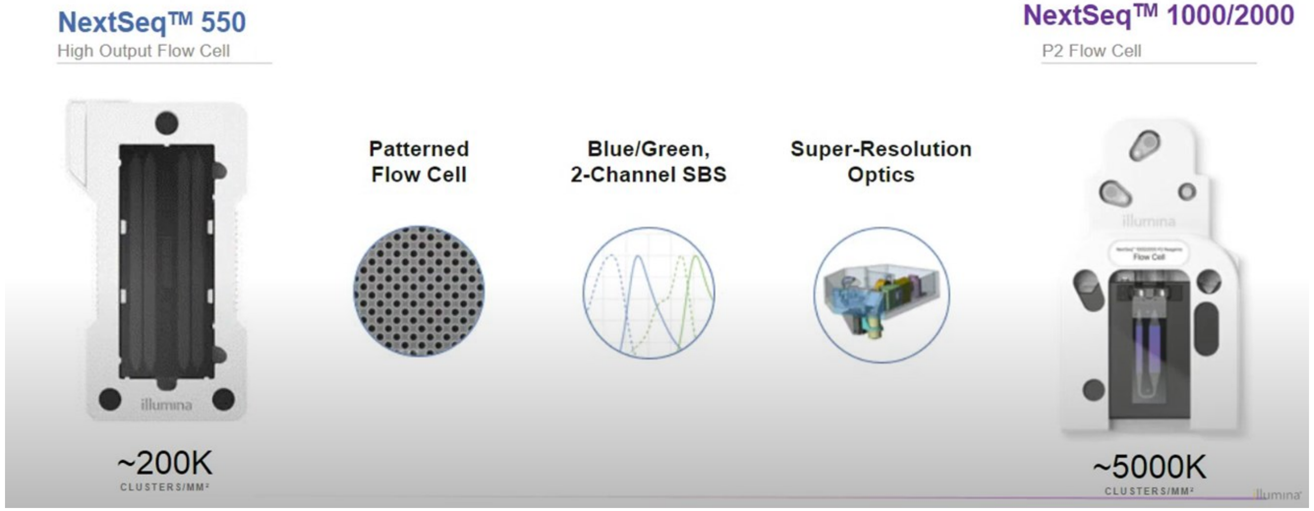
2-Channel Chemistry	A	G	T	C
Image 1	●		●	
Image 2	●			●
Result	A	G	T	C

Troubleshooting an Illumina sequencing run NextSeq 1000/2000 (Illuminas new line of sequencers)

Cluster density and PF% **are not key parameters**

The reason

Patterned flow cell technolgy



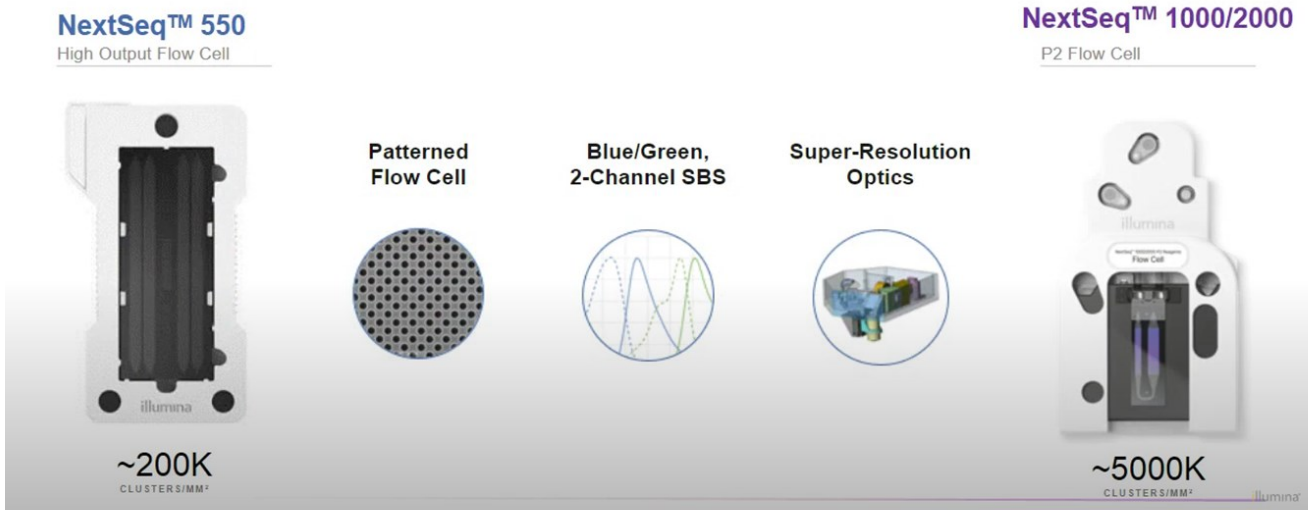
Troubleshooting an Illumina sequencing run
NextSeq 1000/2000 (Illuminas new line of sequencers)

Cluster density and PF% **are not key parameters**

The reason

Patterned flow cell technolgy

Instead look at PhiX% and 5 Loading concentration



Troubleshooting an Illumina sequencing run NextSeq 1000/2000 (Illuminas new line of sequencers)

Cluster density and PF% **are not key parameters**

The reason

Patterned flow cell technolgy

Instead look at PhiX% and 5 Loading concentration

In the PrimaryAnalysisMetrics.csv

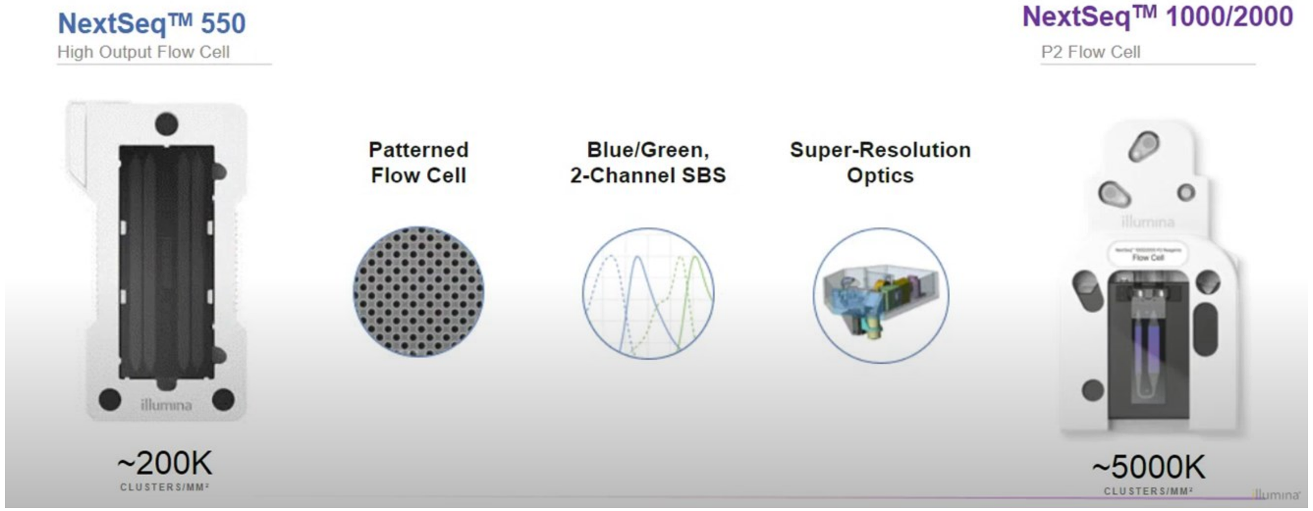
Metric, Unit, Value

≥ Q30, %, 92.61

Total Yield, Gbp, 39.07

Total Reads PF, M, 122.96

% Loading Concentration, %, 99.88



Concluding remarks on Illumina

Pros ✓

Massive Parallel Sequencing

Potentially high data yield

High multiplex capacity

High quality reads

Cons !

Read length

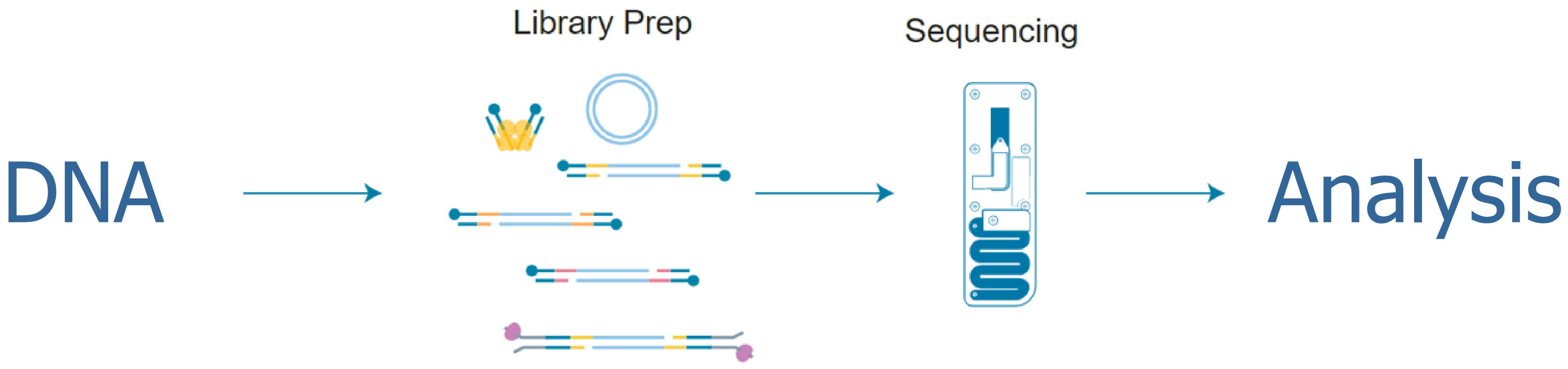
Quality drops during synthesis of strands

Data only available after run completion

Questions for
Illumina Library preparation and sequencing

Nanopore Library preparation and sequencing

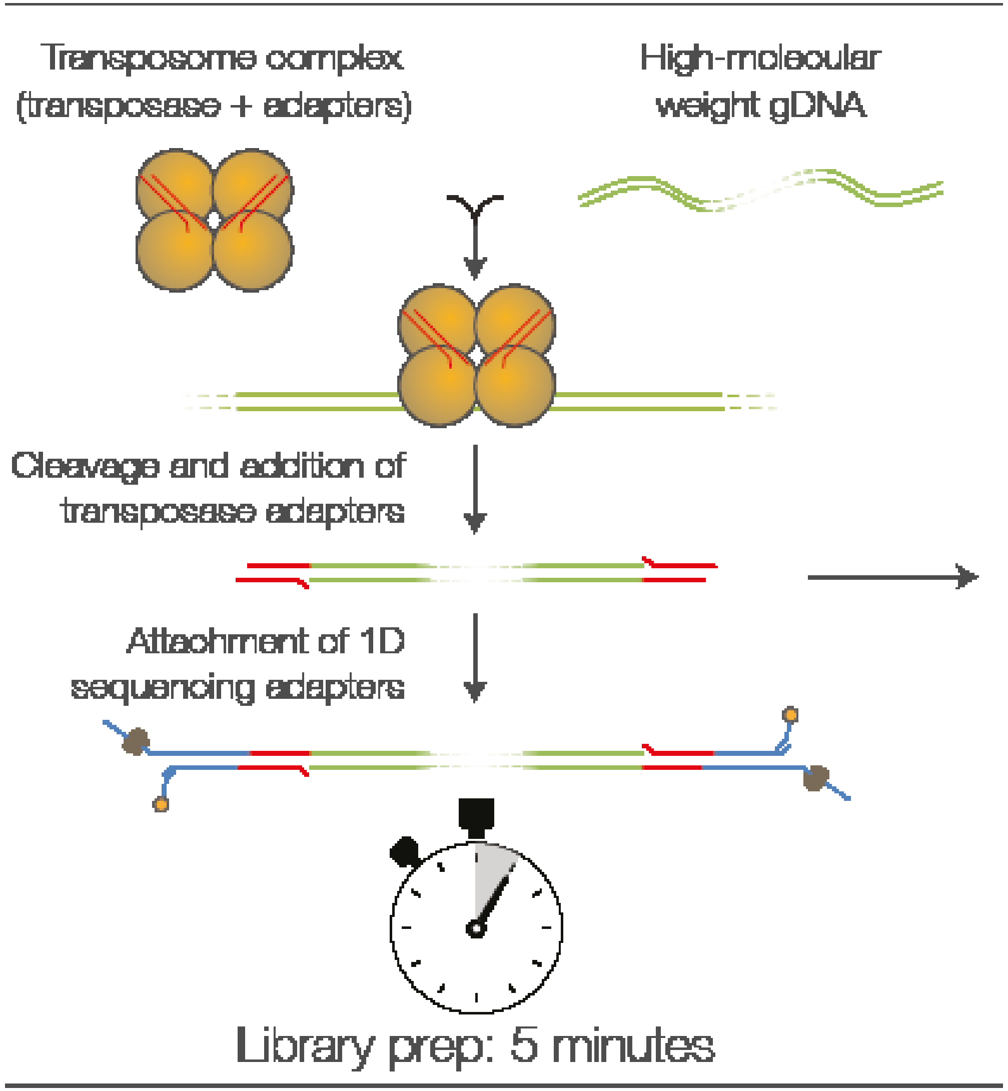
Nanopore sequencing



DNA → Library Prep → Sequencing → Analysis

Figure adapted from [Community - Protocol (nanoporetech.com)]

Nanopore Ligation library prep



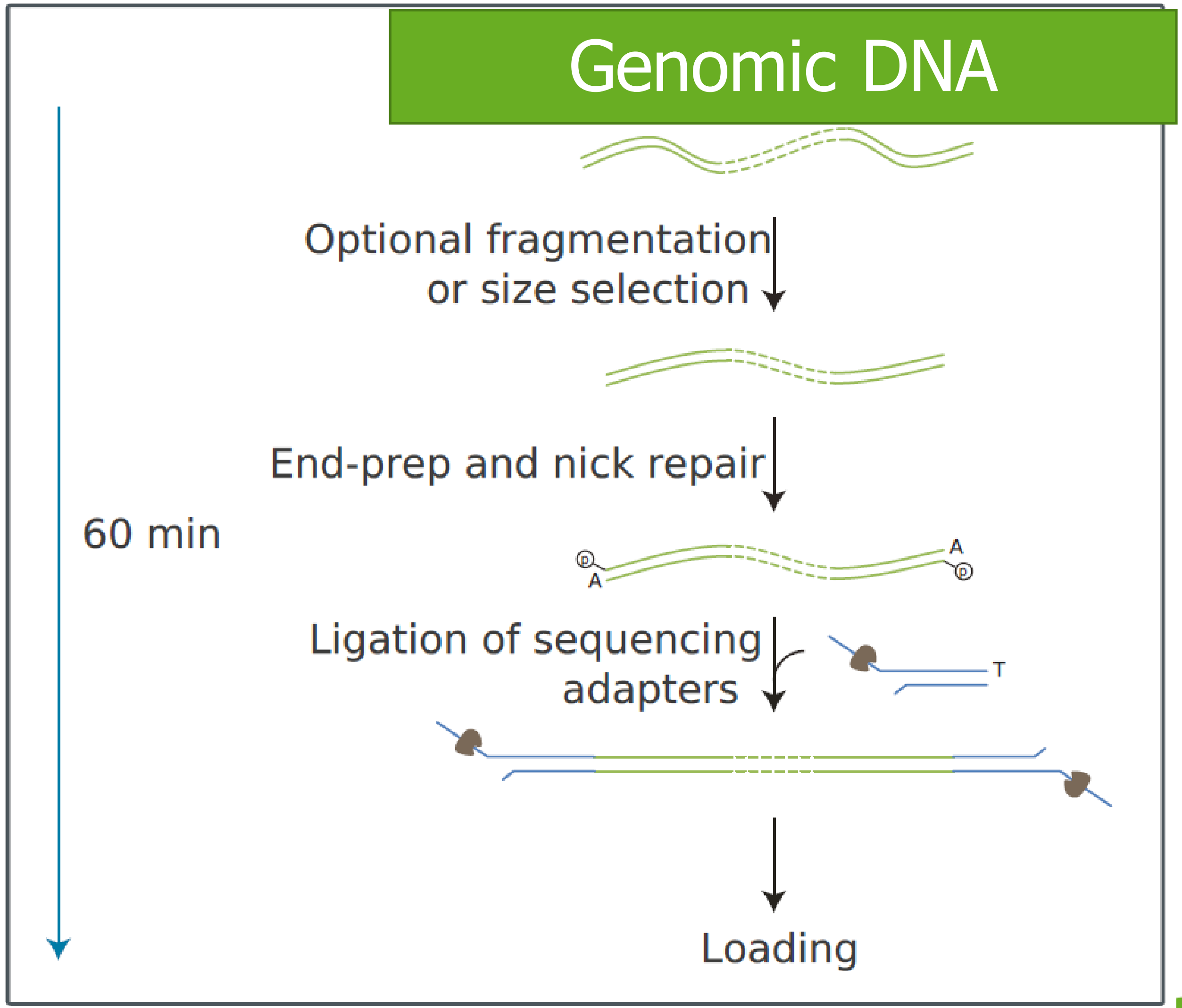
Transposome complex (transposase + adapters)

High-molecular weight gDNA

Cleavage and addition of transposase adapters

Attachment of 1D sequencing adapters

Library prep: 5 minutes



Genomic DNA

Optional fragmentation or size selection

End-prep and nick repair

Ligation of sequencing adapters

Loading

60 min

Image adapted from the SQK-LSK-114 Library Prep Kit Reference Guide (© 2023, Oxford Nanopore Technology)

Nanopore sequencing

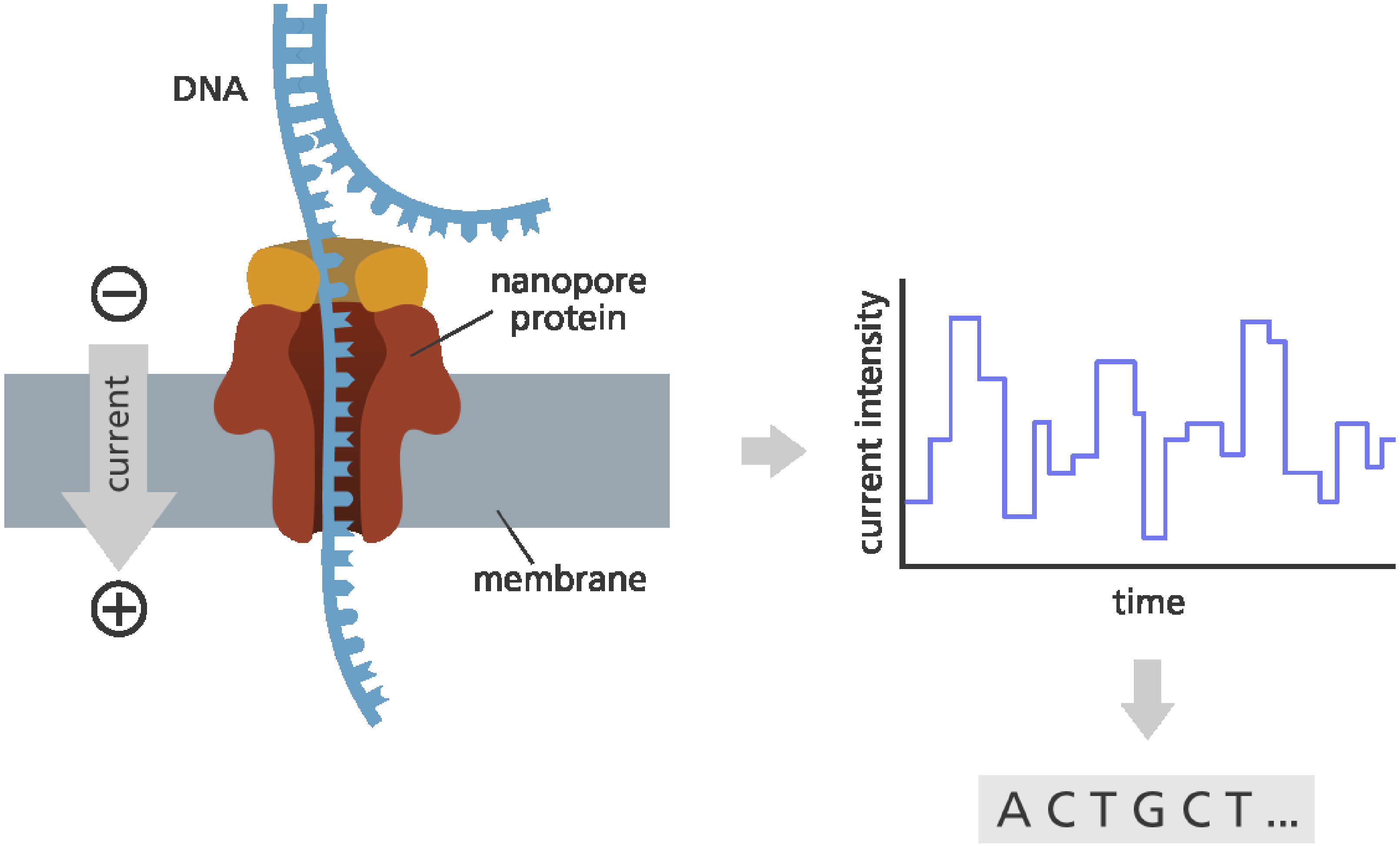


Image credit: Genome Research Limited.

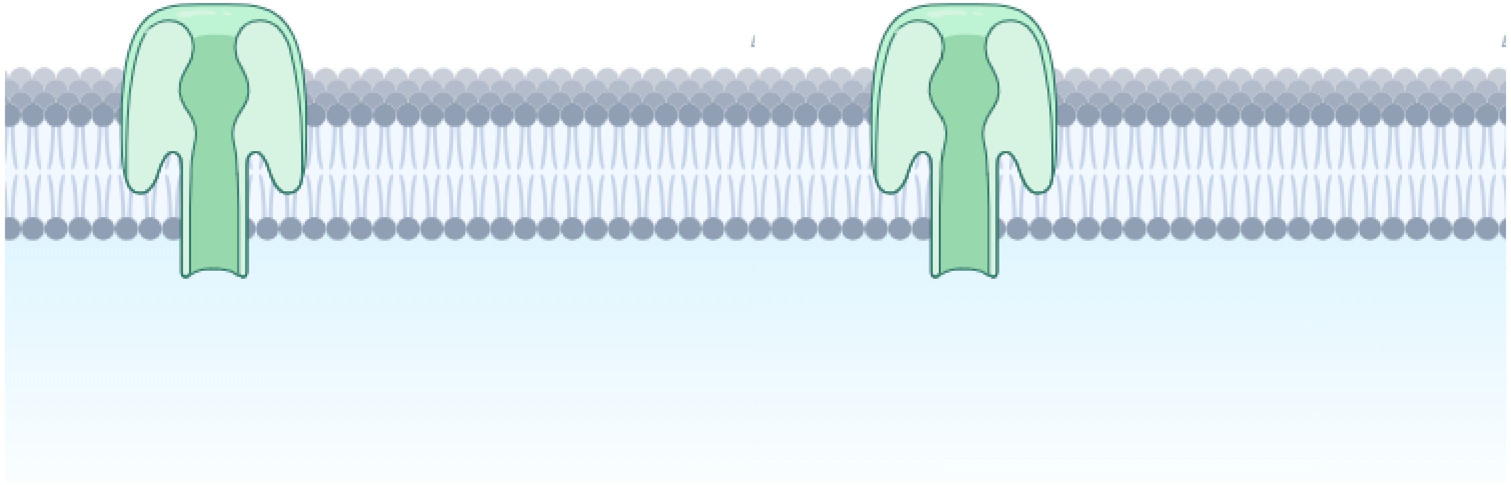


Image credit: Astrid Rasmussen

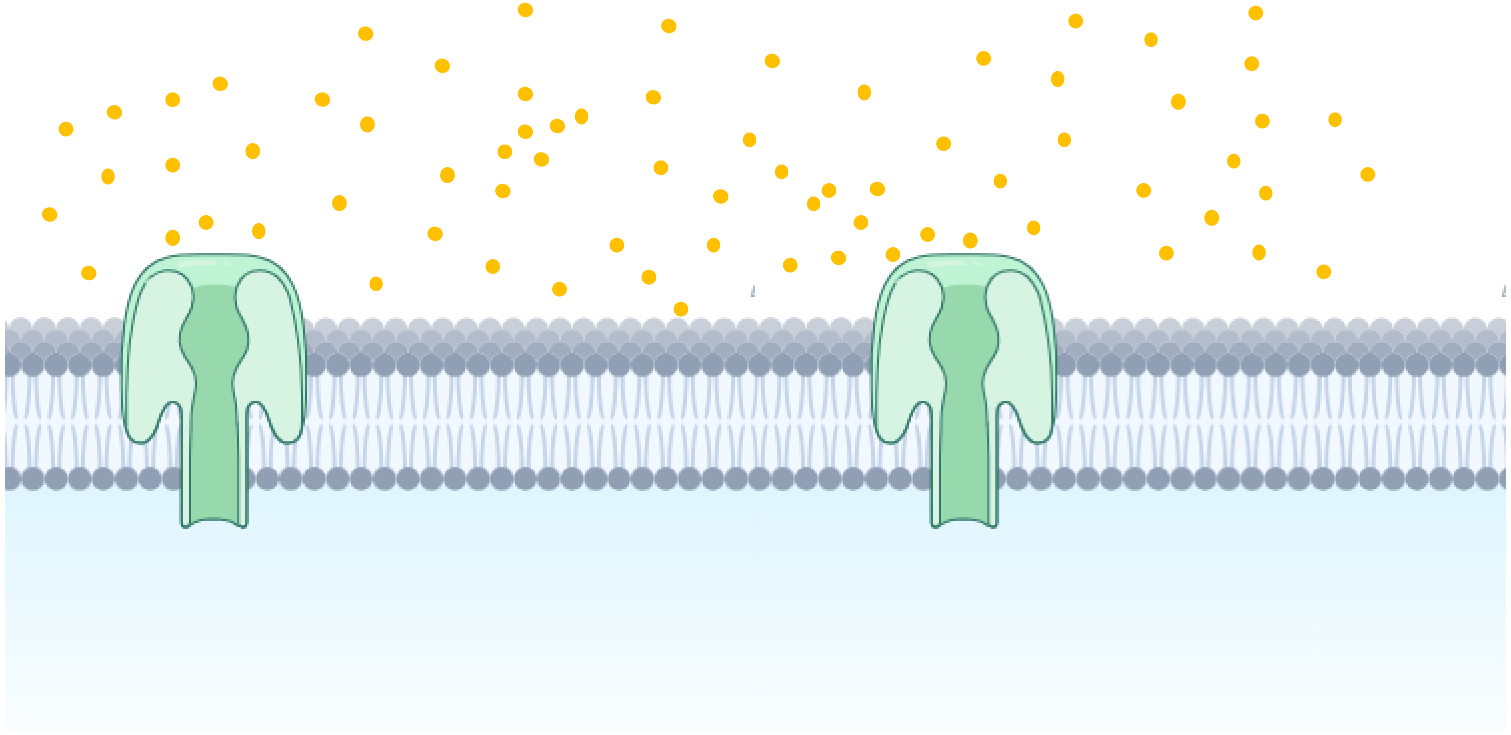


Image credit: Astrid Rasmussen

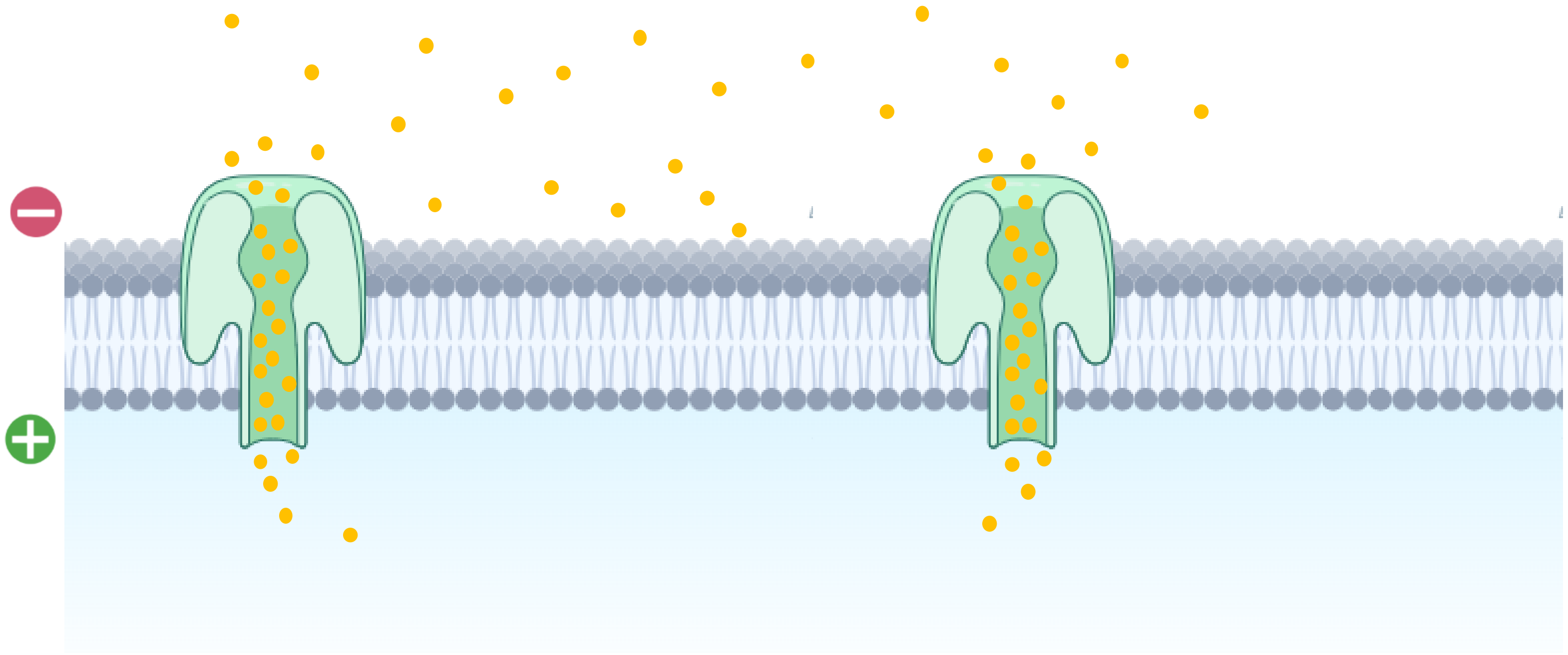


Image credit: Astrid Rasmussen

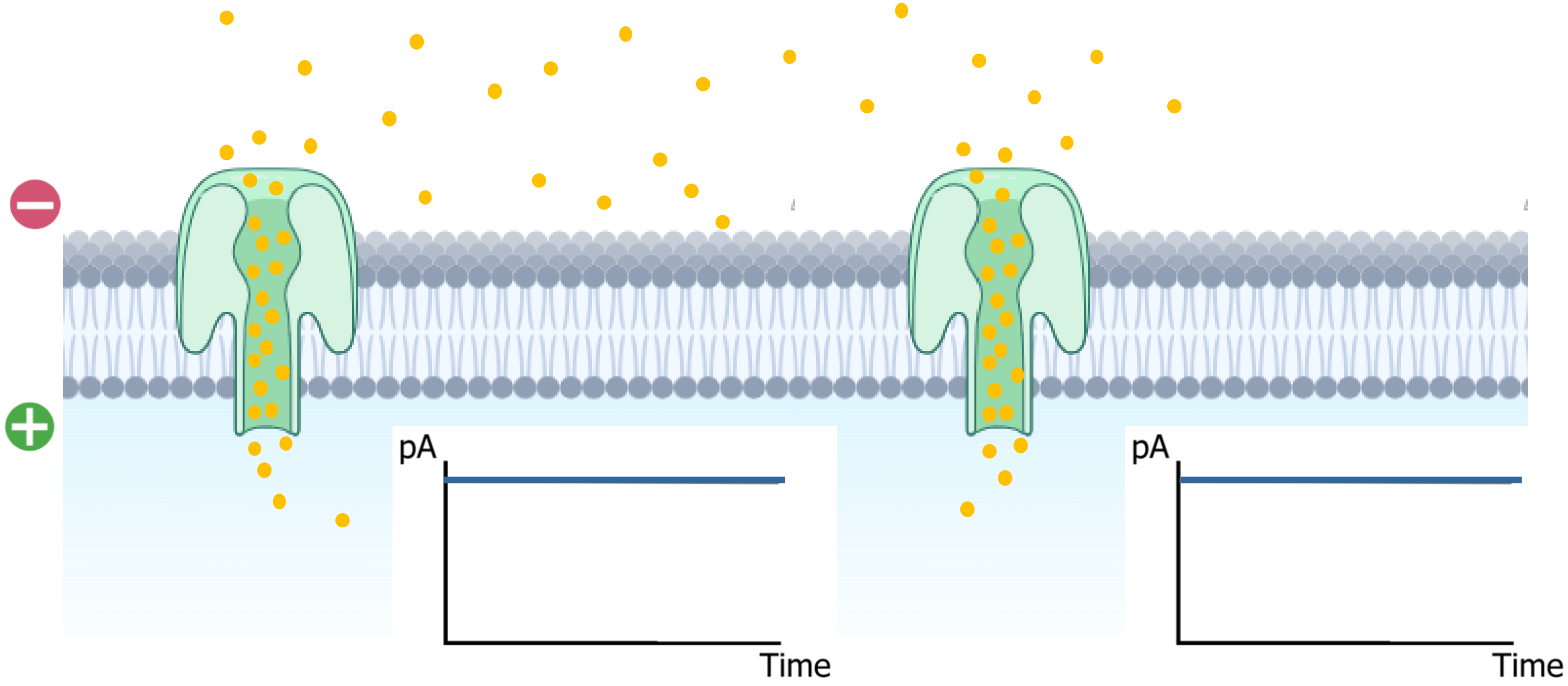


Image credit: Astrid Rasmussen

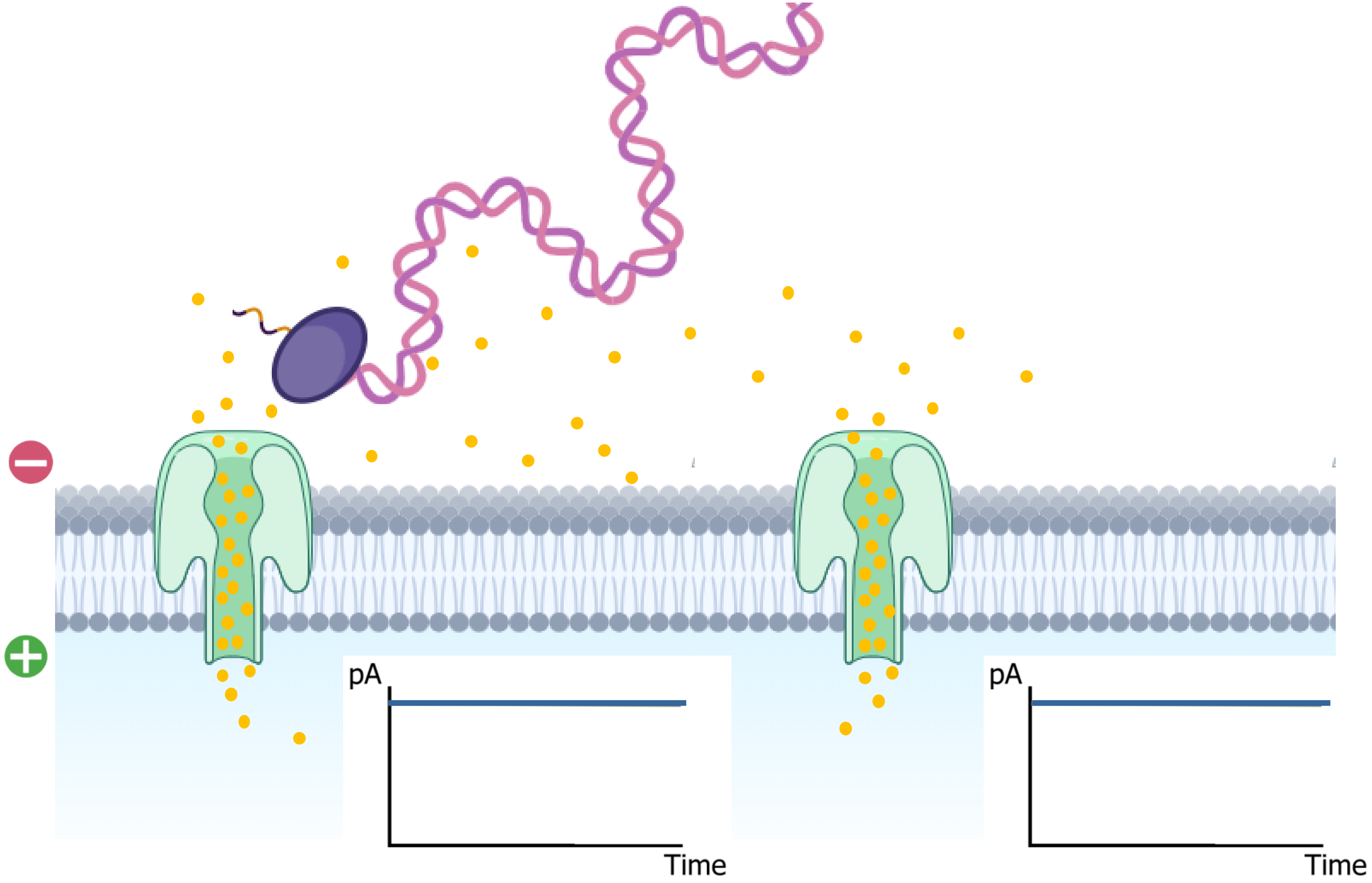


Image credit: Astrid Rasmussen

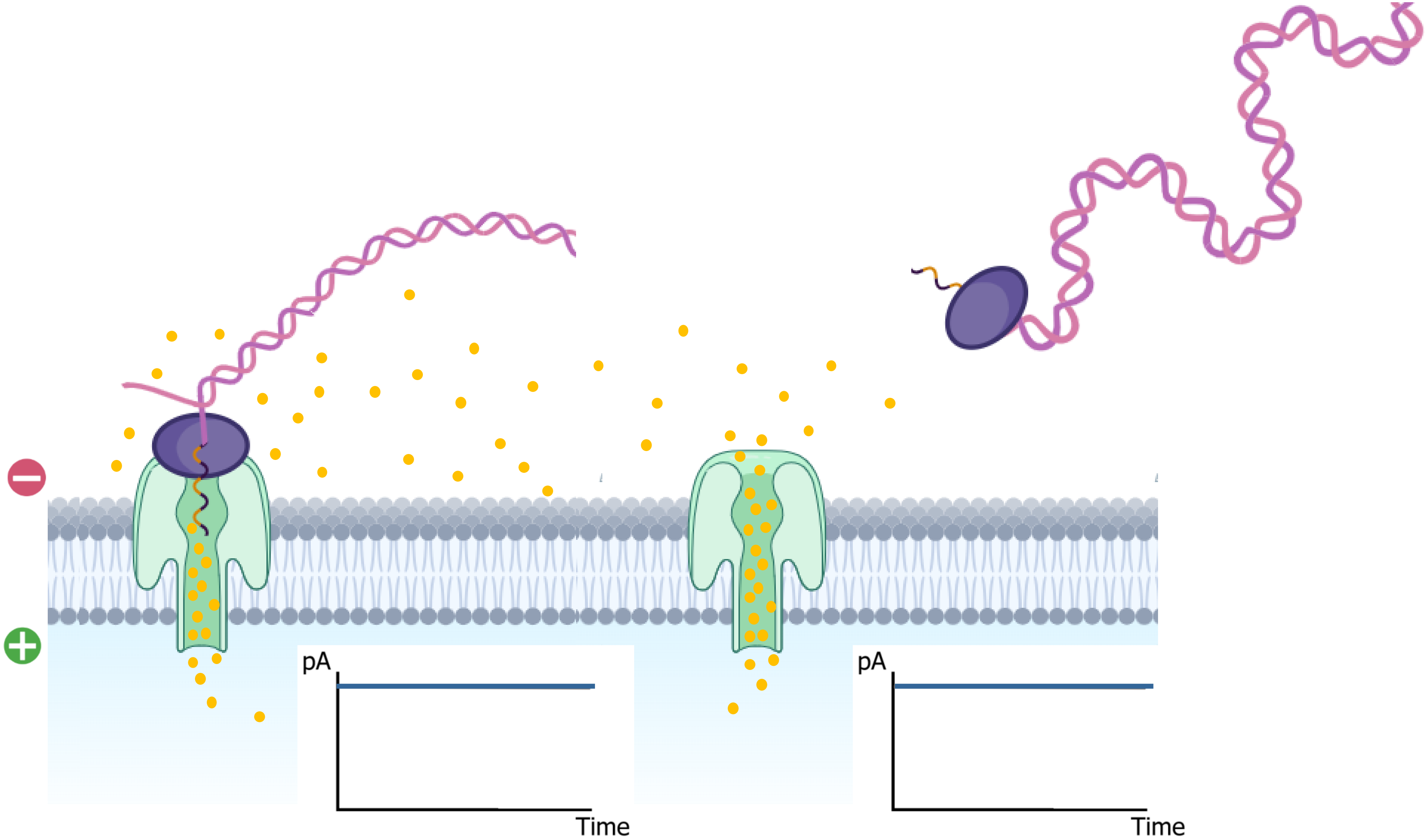


Image credit: Astrid Rasmussen

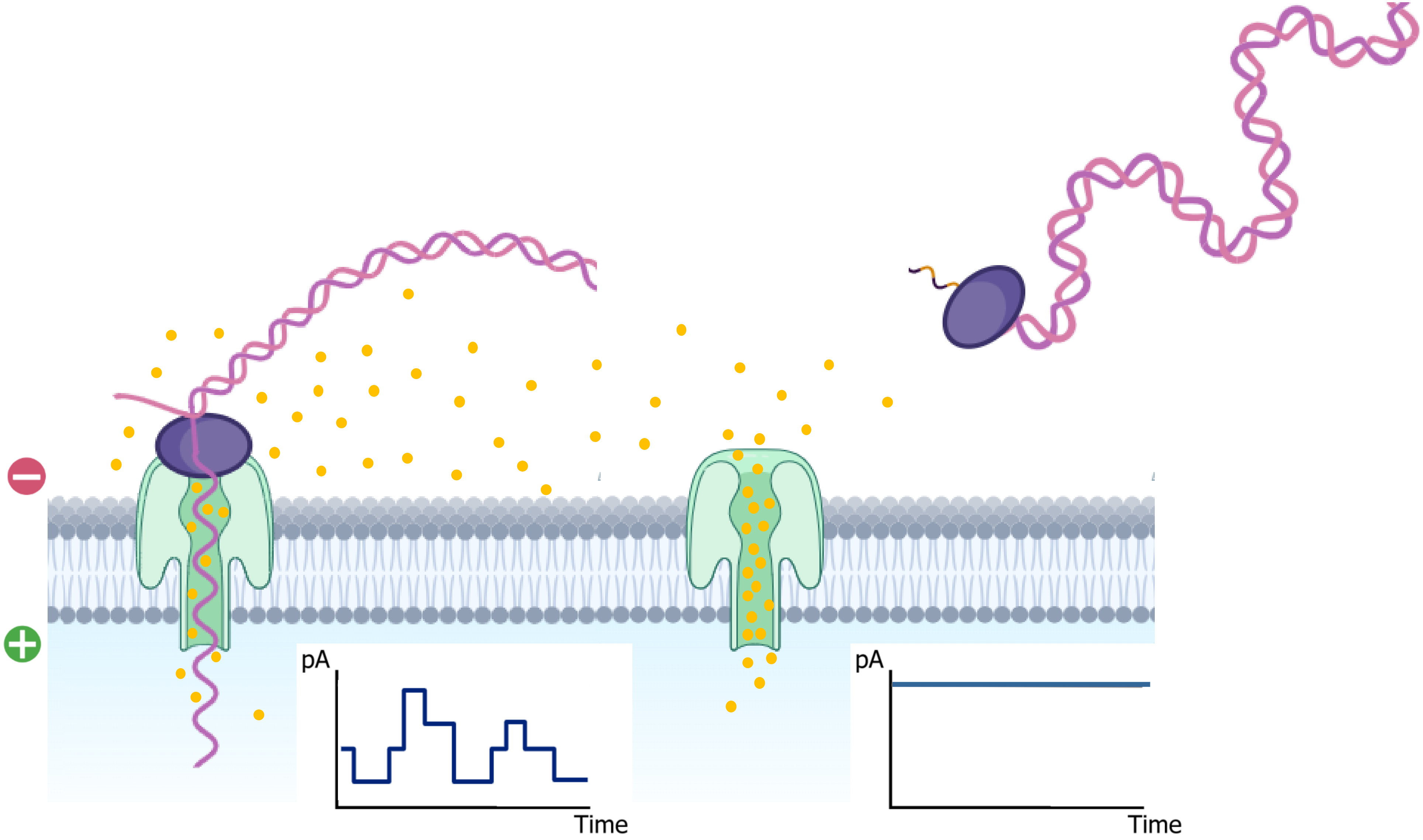
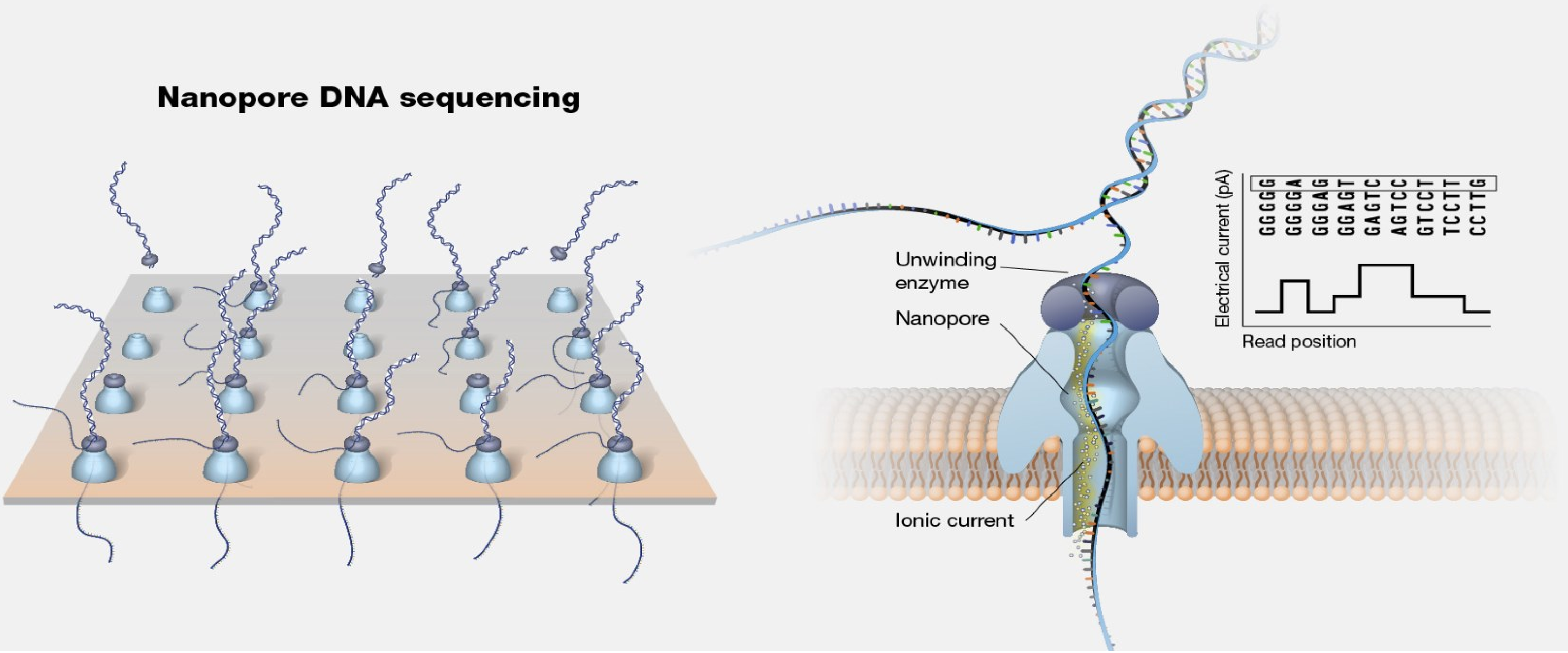


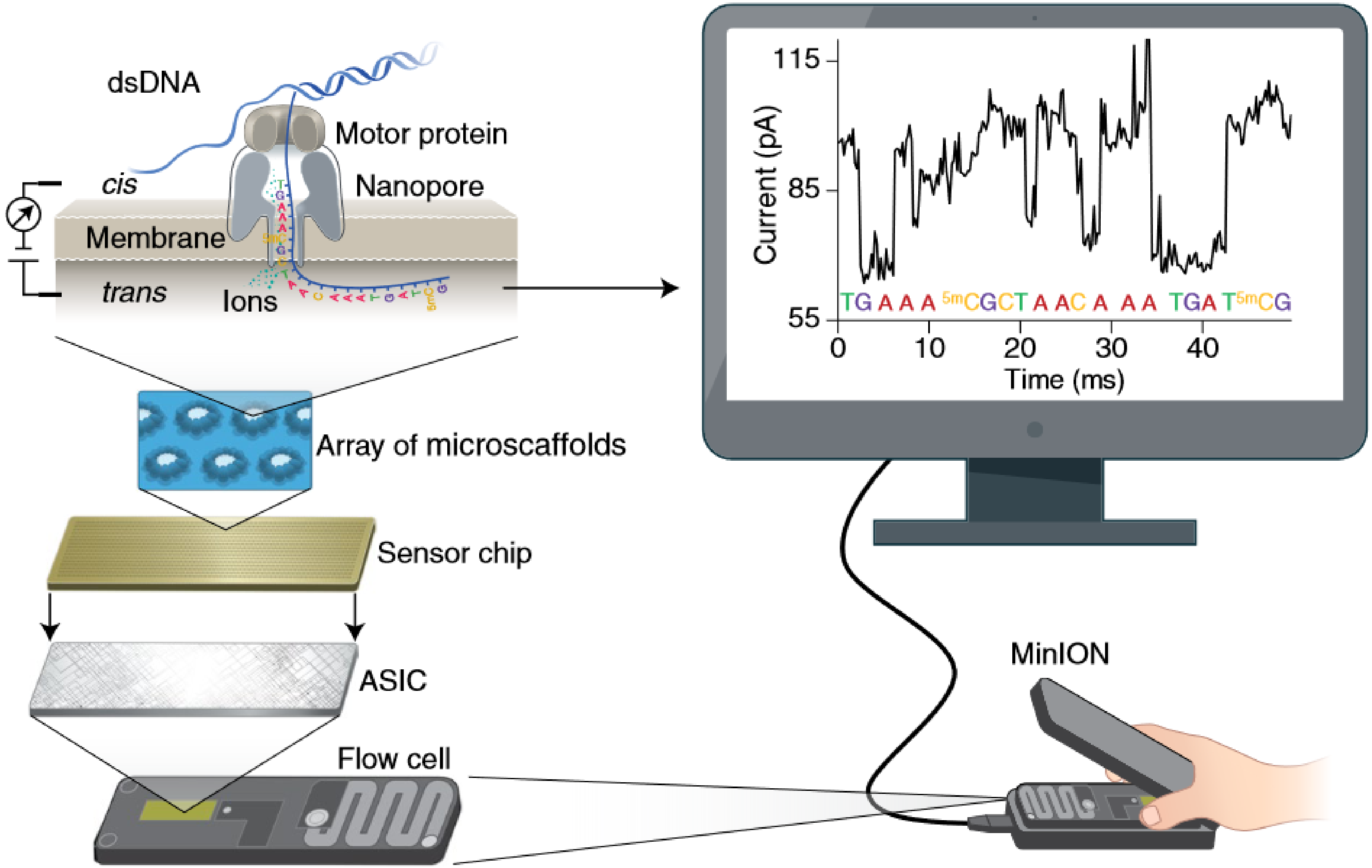
Image credit: Astrid Rasmussen

Nanopore sequencing



Nanopore DNA Sequencing (genome.gov)

Nanopore sequencing



Wang et al., **2019**, *Nature biotechnology*

Nanopore Rapid analysis pipeline

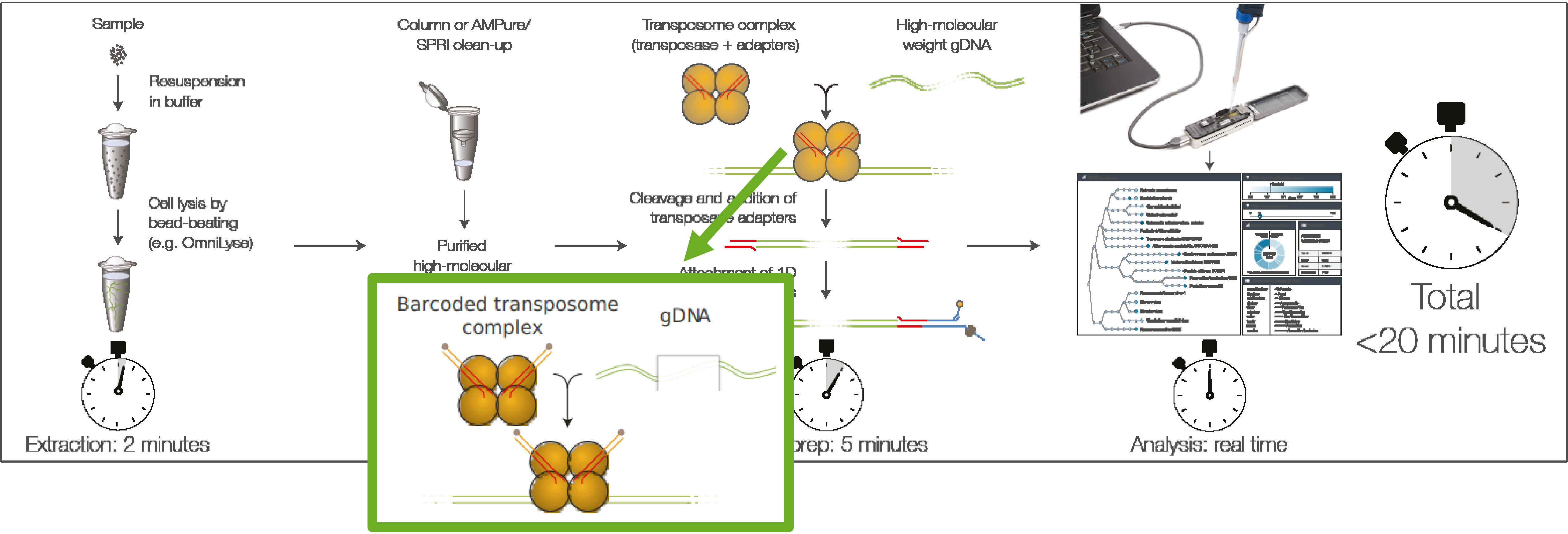

Sample
Resuspension in buffer
Cell lysis by bead-beating (e.g. OmniLyse)
Extraction: 2 minutes
Column or AMPure/ SPRI clean-up
Purified high-molecular
Transposome complex (transposase + adapters)
High-molecular weight gDNA
Cleavage and a dition of transpos e adapters
Attachment of 1D
prep: 5 minutes
Analysis: real time
Total <20 minutes
Barcoded transposome complex
gDNA

Image adapted from https://nanoporetech.com/sites/default/files/s3/posters/pdf/p17007-quick-extract-web.pdf (© 2023, Oxford Nanopore Technology)

Nanopore sequencing

Pros ✓

Parallel real time sequencing

Native DNA Sequencing (includes DNA modifications in signal)

PCR-free library preparation

Cons !

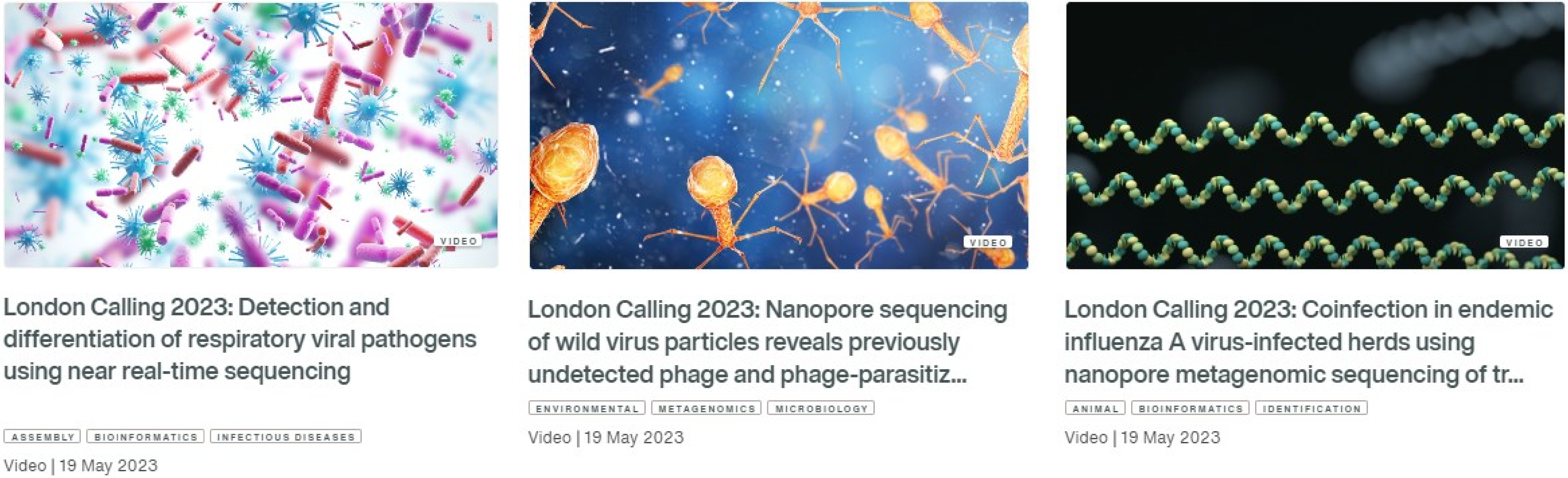
Base calling requires a advanced algorithm

- Can cause low quality
- Difficulty regions

Flow cell activity drops over time

- Less data is generated over time

Further 'reading'



London Calling 2023: Detection and differentiation of respiratory viral pathogens using near real-time sequencing

ASSEMBLY | BIOINFORMATICS | INFECTIOUS DISEASES

Video | 19 May 2023

London Calling 2023: Nanopore sequencing of wild virus particles reveals previously undetected phage and phage-parasitiz...

ENVIRONMENTAL | METAGENOMICS | MICROBIOLOGY

Video | 19 May 2023

London Calling 2023: Coinfection in endemic influenza A virus-infected herds using nanopore metagenomic sequencing of tr...

ANIMAL | BIOINFORMATICS | IDENTIFICATION

Video | 19 May 2023

Resource centre (nanoporetech.com)

Input DNA

Illumina

- Fragment Length
  - DNA is tagmented
- Dilution to >1 ng/µl
  - Purity not a hugh issue

Nanopore

- Long Fragments
  - Very important
- Little or No dilution
  - High purity required

In both cases magnetic bead based extraction is prefered

EVALUATING QUALITY OF DNA FOR NEXT GENERATION SEQUENCING

High molecular weight DNA

- Bioananalyzer/Tapestation or Agarose gel

Purity (measure absorbance ratios using e.g. Nanodrop)

- 260/280 ratio ~1.8 (No RNA contamination)
- 260/230 ratio >2.0 (No contaminants such as EDTA and salts)

Yield

- High concentration (>5ng/µl) = succesful lysis and extration

Part round-off discussion

1. What extraction and purification platforms do you use or have available?

2. Would you ever think of validating this part of the sample flow?

How many samples to load
Bacterial isolates

Total output (Gb) / Genome Size (Mb) / Coverage (50) = isolates

NextSeq 550 Mid output (300 cycles) ~30 Gb (Up to 35 Gb)

E. coli = 5.5 Mb

30 Gb / 5.5*10^-3 Gb / 50 =109

A genome load limit of 400 Mb:

400 Mb * 50x coverage = 20 Gb

When sequencing fails 1

! Low quality input material

- ✓Grow bacteria on non-selective plates (e.g. blood agar)
- ✓Prior validation of your extraction procedure
- ✓Measure concentrations and dilute accurately

When sequencing fails 2

! Library preparation issues

- ✓Careful index addition when multiplexing
- ✓Test for PCR/transposase inhibitors

! Size selection

- ✓Correct bead ratios – Bead resuspension (beads sediment fast)
- ✓Complete ethanol removal following bead wash

When sequencing fails 3

On the sequencer (Illumina)

- ! Over/under clustering (Bad cluster recognition OR Low data output)
    - ✓ Measure lib conc and dilute carefully
- ! Low diversity libraries (mainly amplicon)
    - ✓ add more phiX, heterogeneity

On the sequencer (Nanopore)

- ! Over/Under saturation of nanopores (Flowcell clutting OR Loss activity)
    - ✓ Measure concentration and dilute library if necesary (70-90 ng/µl)
- ! Flowcell is temperature sensitiv (34 – 37°C)
    - ✓ Keep sequencers under temperature controlled rooms when possible

Illumina	Nanopore
2nd generation Uses reversible dye terminators to detect sequence of DNA molecules	3rd generation Uses nanopores to detect sequence of DNA molecules
Accuracy: >99% (Q20-Q40)	Accuracy: 92 – 99% (Q12 – Q20)
Short read sequencing technology 25-300bp (up to 2x300 bp)	Long read sequencing technology 1.000 – 100.000 (> 2Mb)
4 – 56 hours	Real time
Potentially very high yield 4-40 (100-1000's Gb)	Fixed yield per flowcell (2-50?) Multiple flowcells in parallel

Microbiologist <-> Bioinformatician <-> Epidemiologist

What is the organization?

What level of understanding is required for successful collaboration?

What technology to choose?

General surveillance

Outbreak detection

Emerging pathogens

- Metagenomics
- Waste water surveillance

Plasmid-borne resistance

Further reading

Head SR, Komori HK, LaMere SA et al. (2014) Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56(2):61-64, 66, 68.

Yu, Xiaoling, Wenqian Jiang, Yang Shi, Hanhui Ye, og Jun Lin. "Applications of sequencing technology in clinical microbial infection". *Journal of Cellular and Molecular Medicine* 23, nr. 11 (november 2019): 7143–50. https://doi.org/10.1111/jcmm.14624.

Buytaers, Florence E., Assia Saltykova, Sarah Denayer, Bavo Verhaegen, Kevin Vanneste, Nancy H. C. Roosens, Denis Piérard, Kathleen Marchal, og Sigrid C. J. De Keersmaecker. "Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies". *Frontiers in Microbiology* 12 (2021). https://www.frontiersin.org/articles/10.3389/fmicb.2021.738284.

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

Acknowledgements

The creation of this training material was commissioned by ECDC to Statens Serum Institute (SSI) with the direct involvement of Søren Hallstrøm and Astrid Rasmussen