Nanopore sequencing

– for dummies and smarties

oslohack:22

Why / What / How

Why this talk?

- Biology is rapidly becoming an informatics problem
- In many ways, biotech of today is where the computer industry was in the 70s
- By reverse engineering the “programs” embedded in our DNA, we can improve our health
- Almost all of the software for the Oxford Nanopore (ONT) sequencers are open source

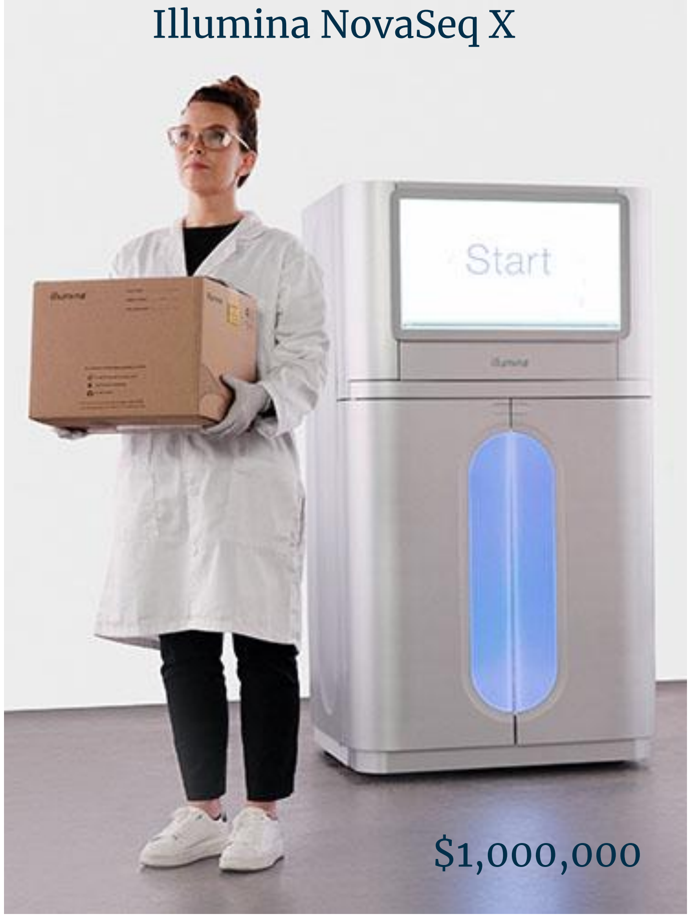
What will I talk about?

- How to accurately digitize the information stored in and around our DNA

How will I proceed?

- Walk-through of a low-cost approach to DNA sequencing using nanopore sequencing

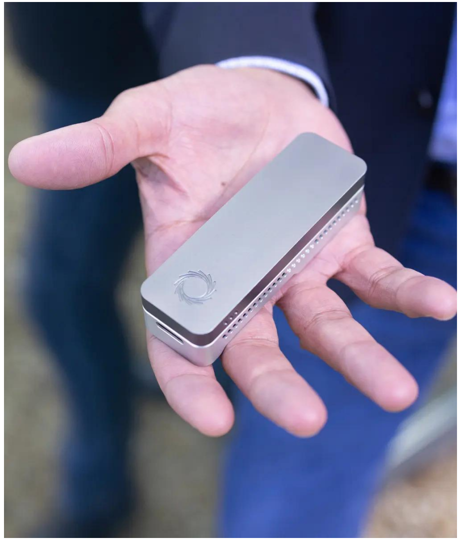
Illumina NovaSeq X



Source: Illumina

$1,000,000

Oxford Nanopore MinION



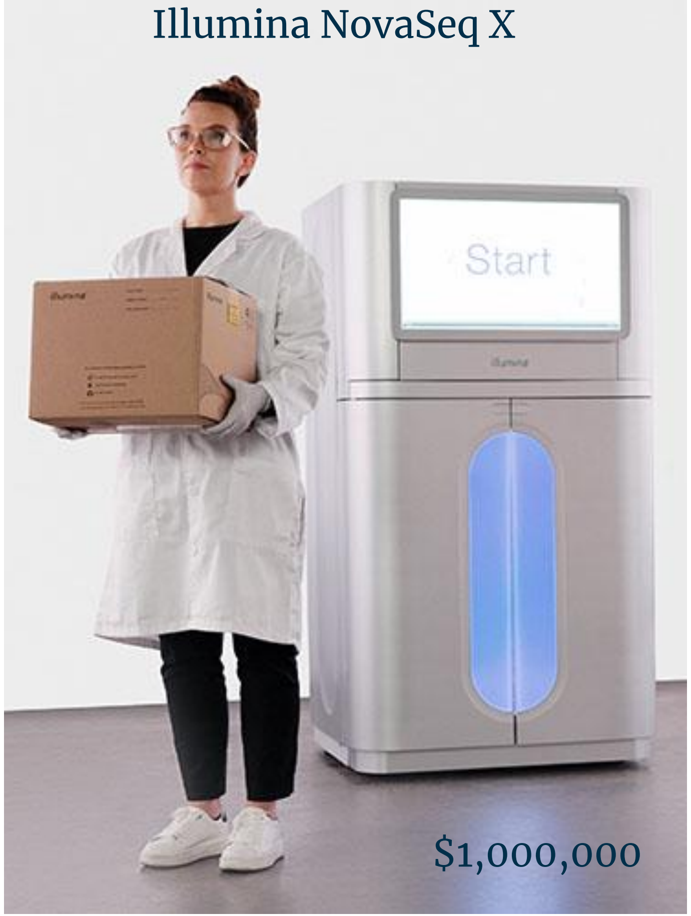
Source: Oxford Nanopore

$1,000

Next generation sequencing

Next next generation sequencing

Illumina NovaSeq X

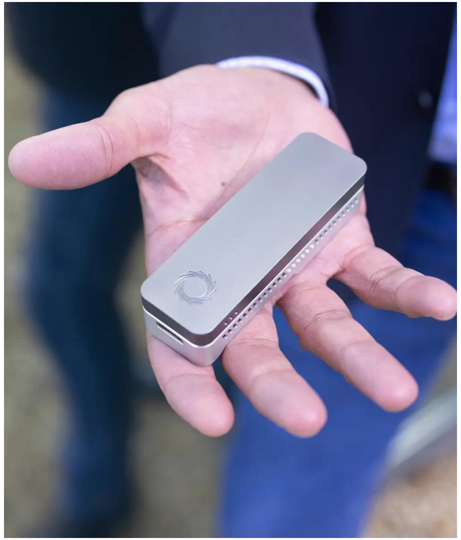


Source: Illumina

$1,000,000

Very closed source

Oxford Nanopore MinION



Source: Oxford Nanopore

$1,000

Mostly open source

Outline

- How the biology works
- How the nanopore hardware and wetware works
- How to do the data analysis
- Examples of applications and application areas
- How to get started yourself
- Wrap-up

How the biology works

DNA is the “source code” for us

DNA is our primary information carrier

We are but a shell for our genes

Identical DNA = identical twins

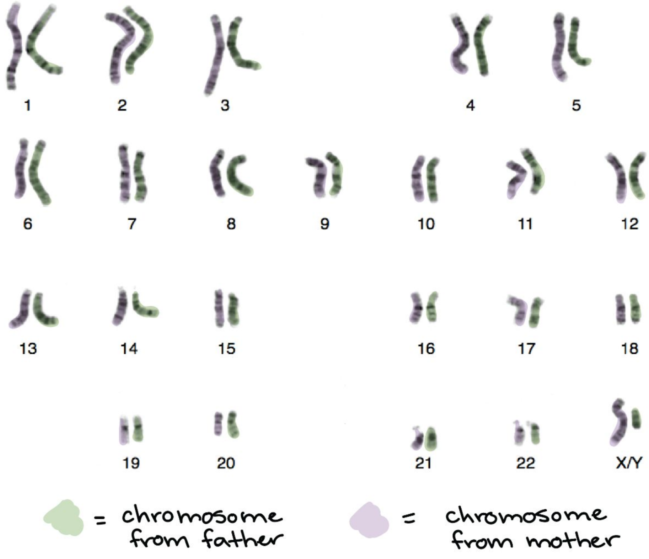
Source: New Scientist

DNA is split into chromosomes*

Split into 23 pairs of chromosomes*

- 22 autosome pairs
- 1 sex chromosome pair

Not unlike splitting your program into 23 x 2 binary files



Source: Khan Academy

* in humans, at least

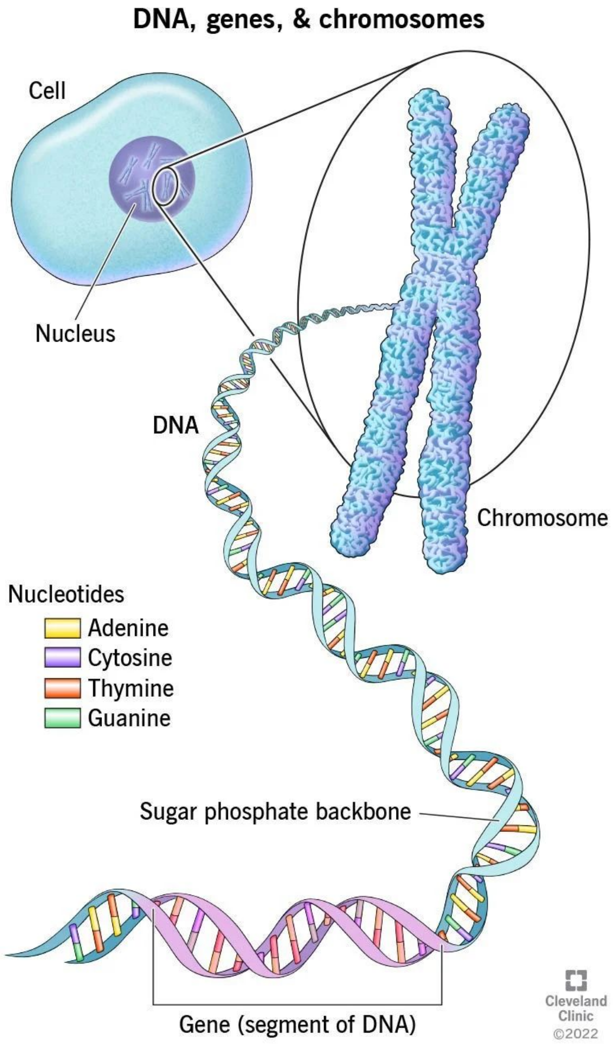
Deoxyribo
Nucleic
Acid

Is found in the core (nucleus) of the cell

Is formed as a double helix

Is a long string of nucleotides

- Bases: A, C, T, G



Cleveland Clinic ©2022

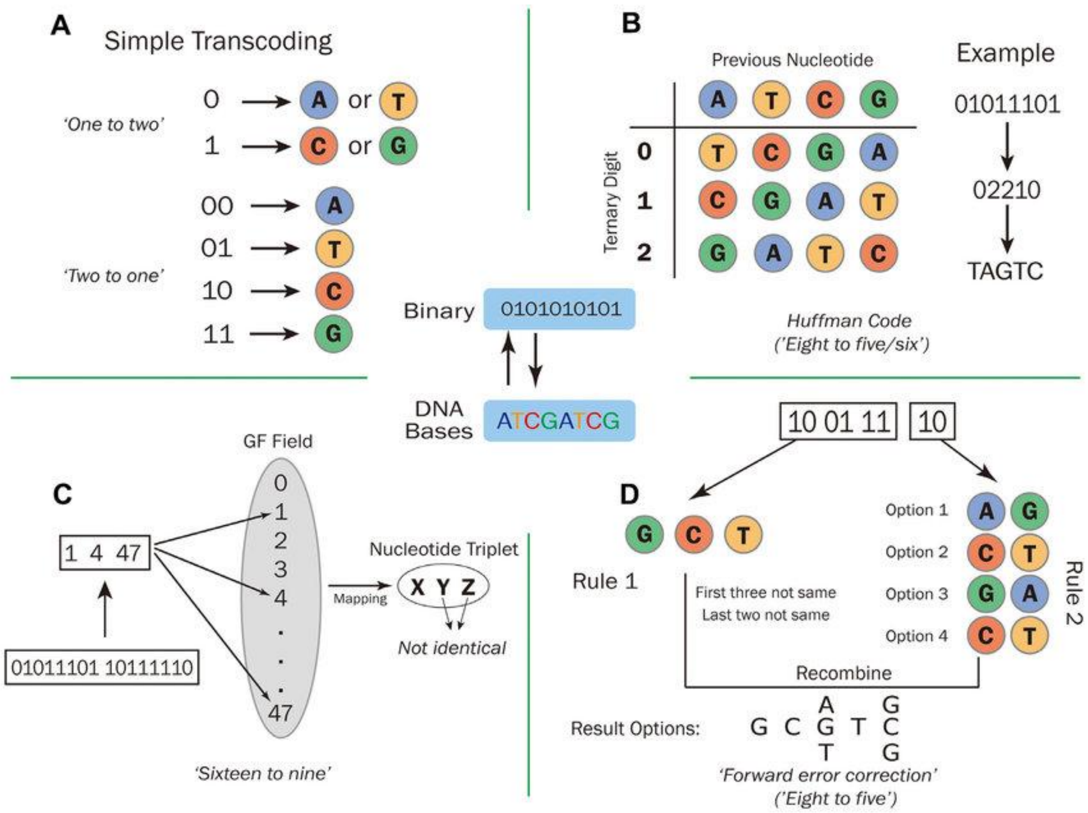
Deoxyribo Nucleic Acid

Is extremely space-efficient

- ~4 gigabases in every cell
- White blood cell is 130 µm
- Sperm cell is 30 µm

Is being explored as a storage medium

Comes with its own configuration

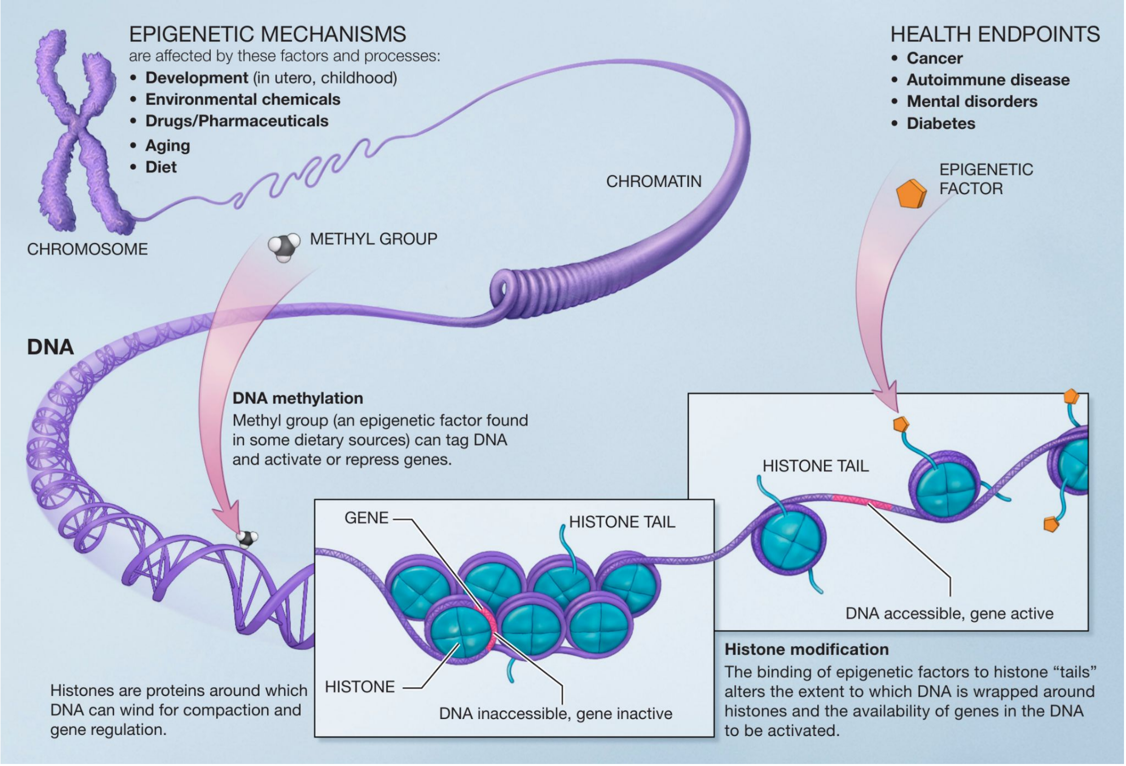


Ping, Zhi & Ma, Dongzhao & Huang, Xiaoluo & Chen, Shihong & Liu, Longying & Guo, Fei & Zhu, Sha & Shen, Yue. (2019). *Carbon-based archiving: current progress and future prospects of DNA-based data storage*. GigaScience. 8. 10.1093/gigascience/giz075.

Epigenetics is the “config setting” for every cell

Q: All cells have the same DNA → how can they be different then?

A: Per-cell configuration = epigenetics


EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
EPIGENETIC FACTOR
METHYL GROUP
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
HISTONE TAIL
DNA accessible, gene active
GENE
HISTONE TAIL
HISTONE
DNA inaccessible, gene inactive
Histones are proteins around which DNA can wind for compaction and gene regulation.
Histone modification
The binding of epigenetic factors to histone “tails” alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

Source: Wikipedia

How the nanopore hardware and wetware works

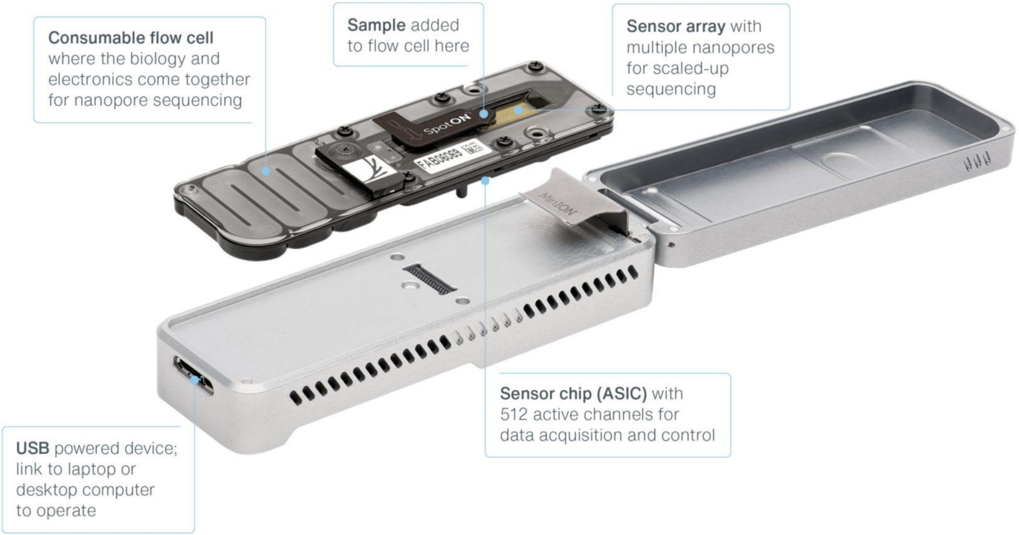
ONT MinION

Follows the inkjet business model

- Sequencer is cheap ($1000)
- Consumable flow cells are expensive ($90 - $900)

Can produce <= 50 Gb per flow-cell

Requires a PC to run



Consumable flow cell where the biology and electronics come together for nanopore sequencing

Sample added to flow cell here

Sensor array with multiple nanopores for scaled-up sequencing

Sensor chip (ASIC) with 512 active channels for data acquisition and control

USB powered device; link to laptop or desktop computer to operate

Source: Oxford Nanopore Technologies

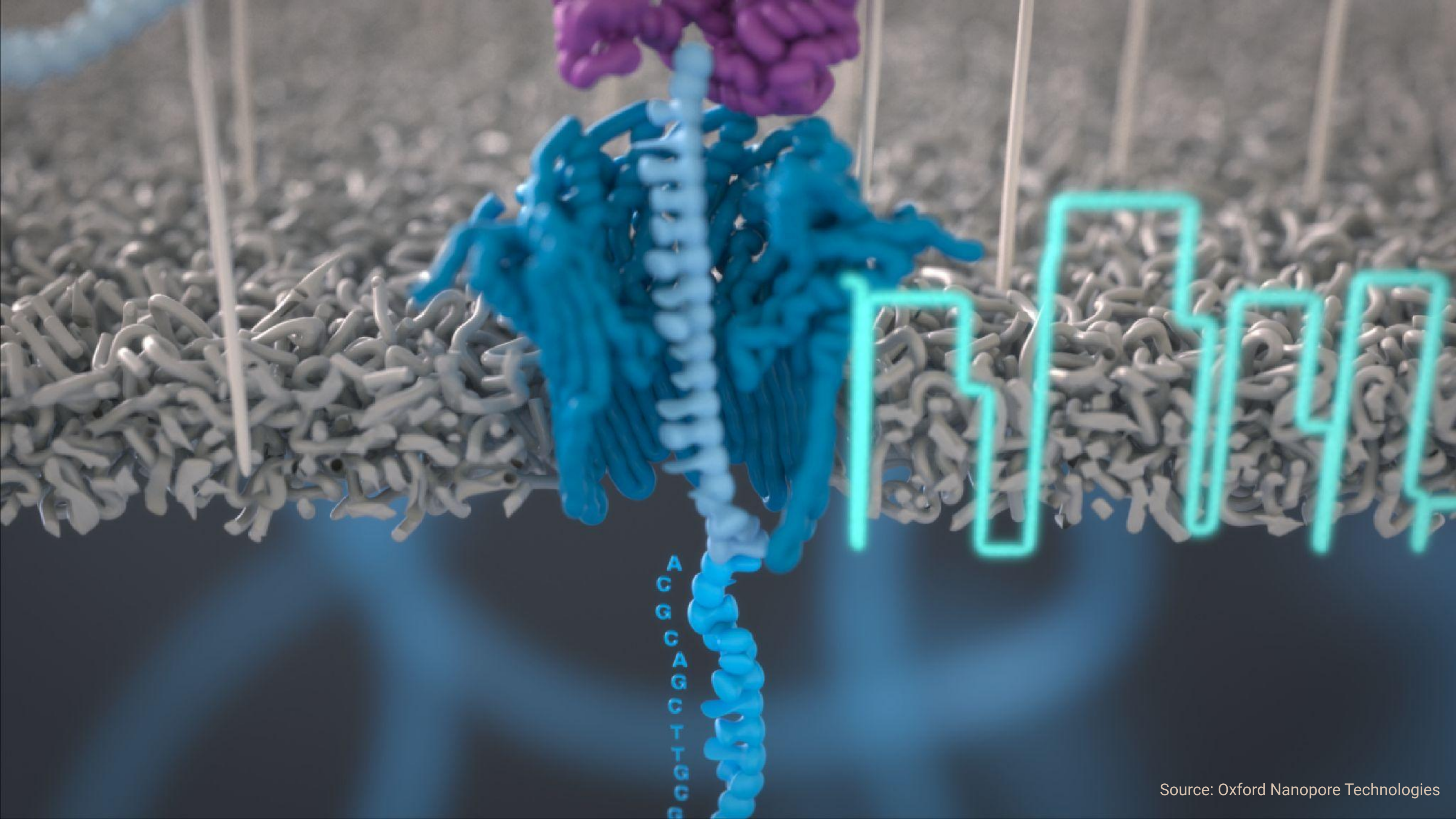
The first DNA sequencer that works in the field

Used on all continents

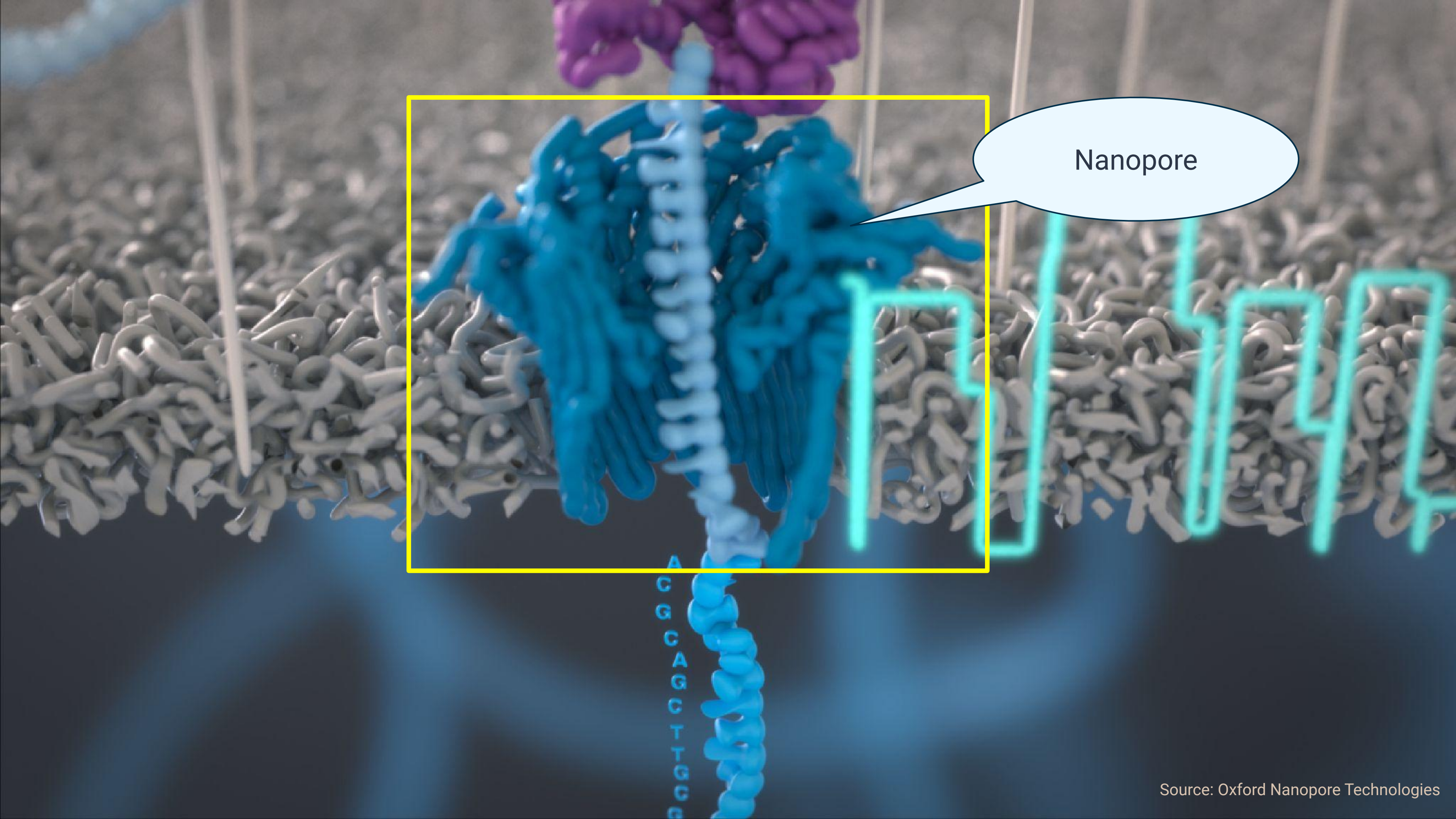
Sequencing has become easy

(Sample preparation remains a hassle)

Credit: Sarah S. Johnson *et al.*

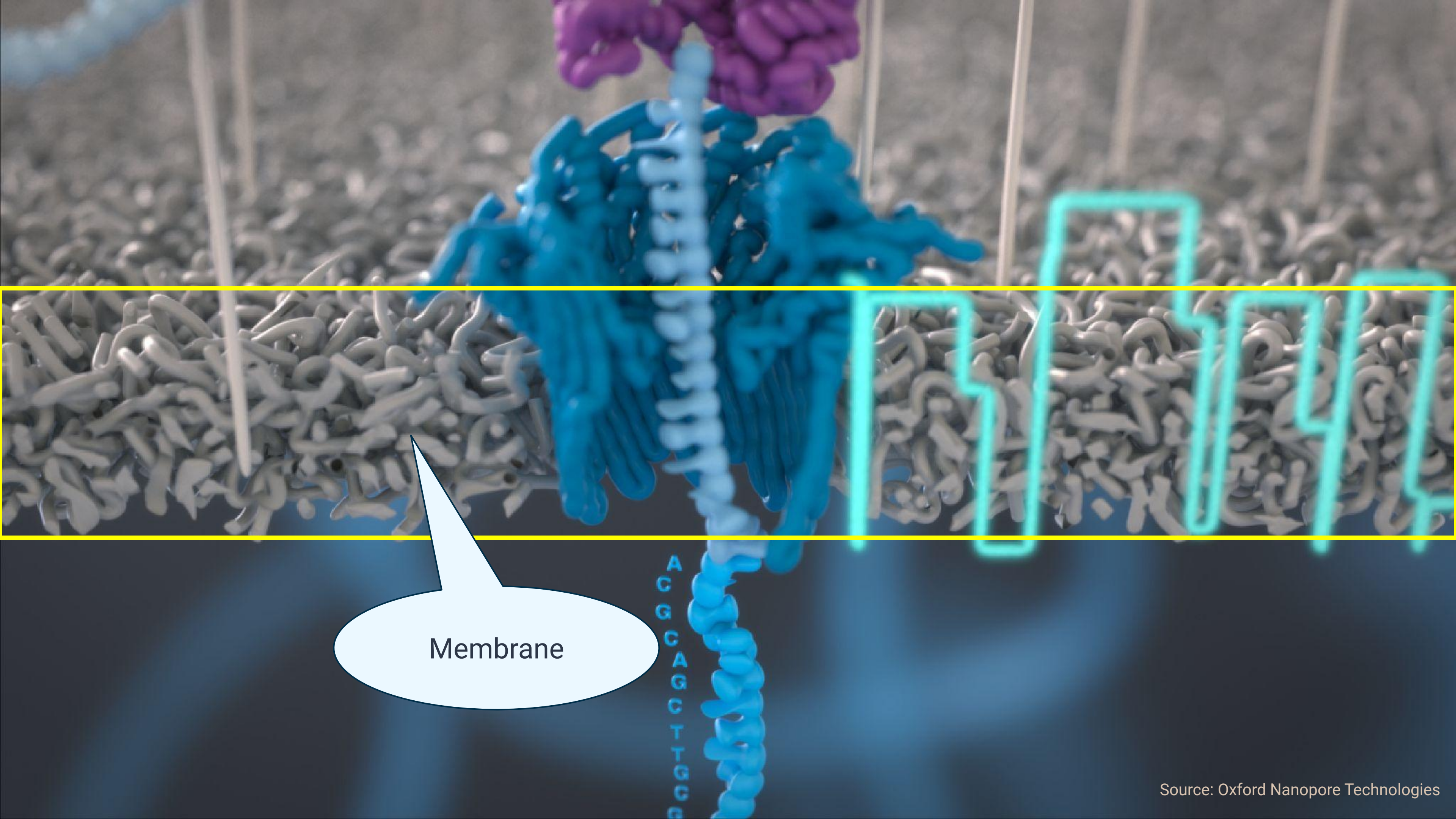


Source: Oxford Nanopore Technologies



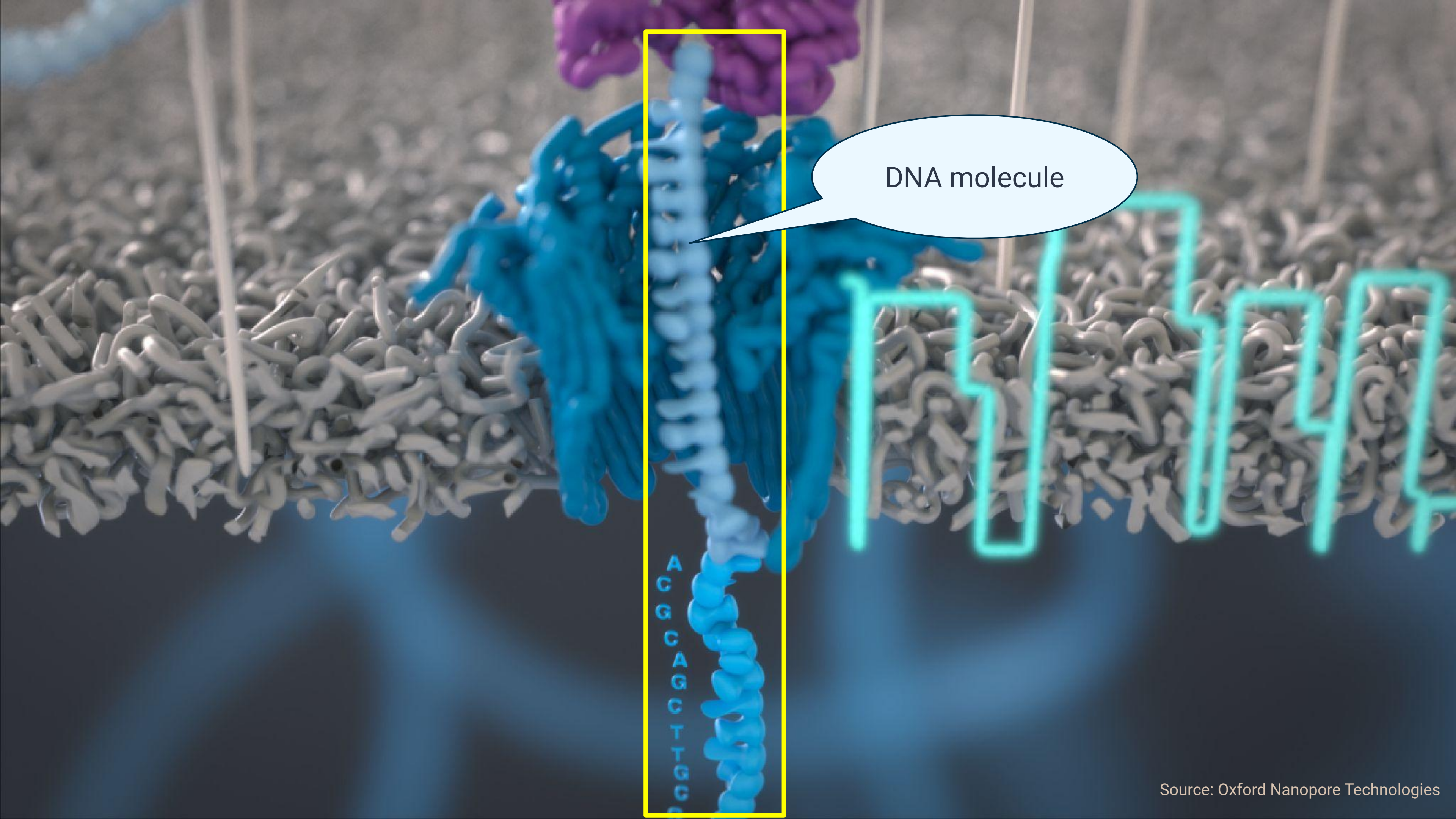
Nanopore

Source: Oxford Nanopore Technologies



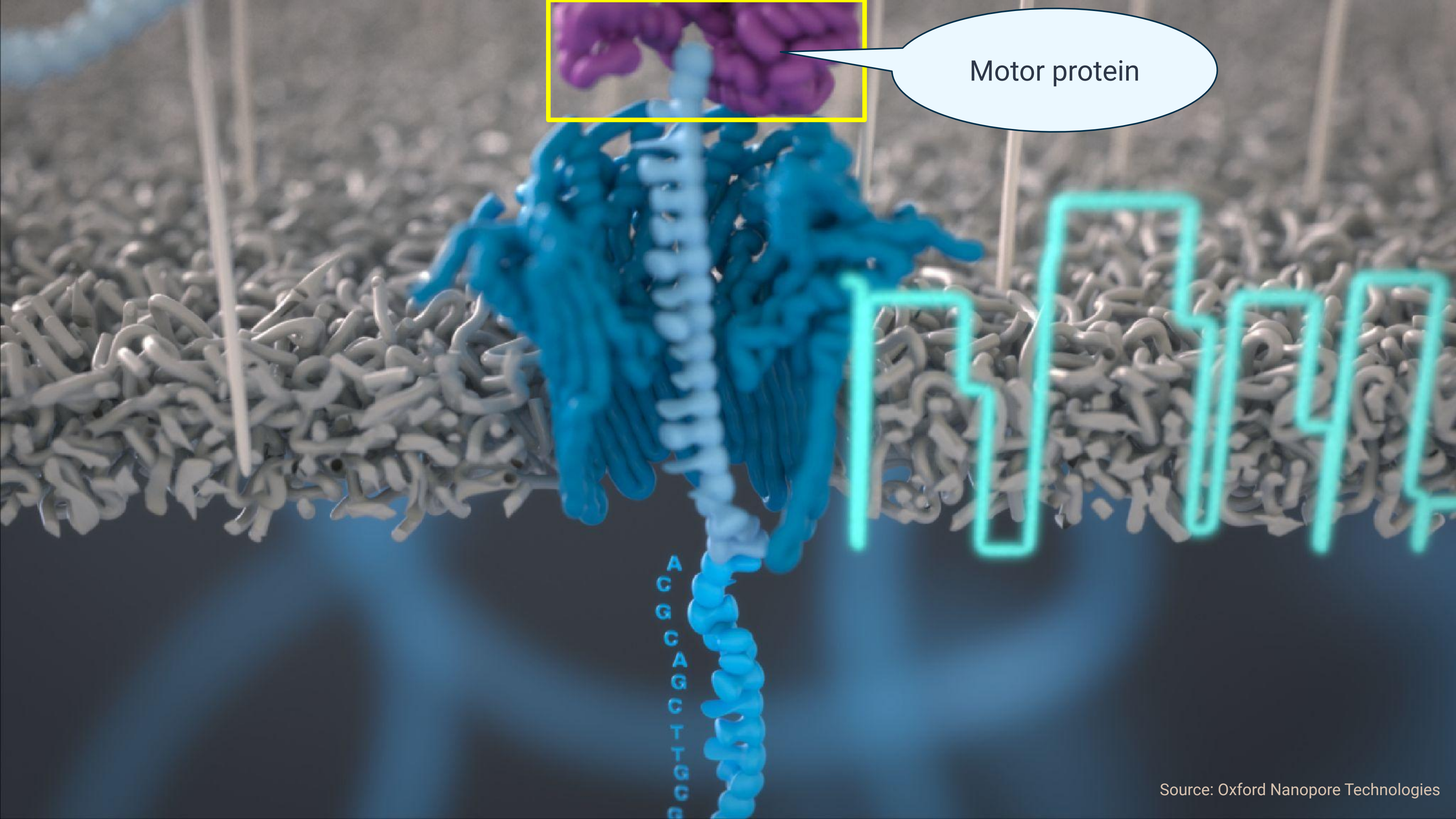
Membrane

Source: Oxford Nanopore Technologies

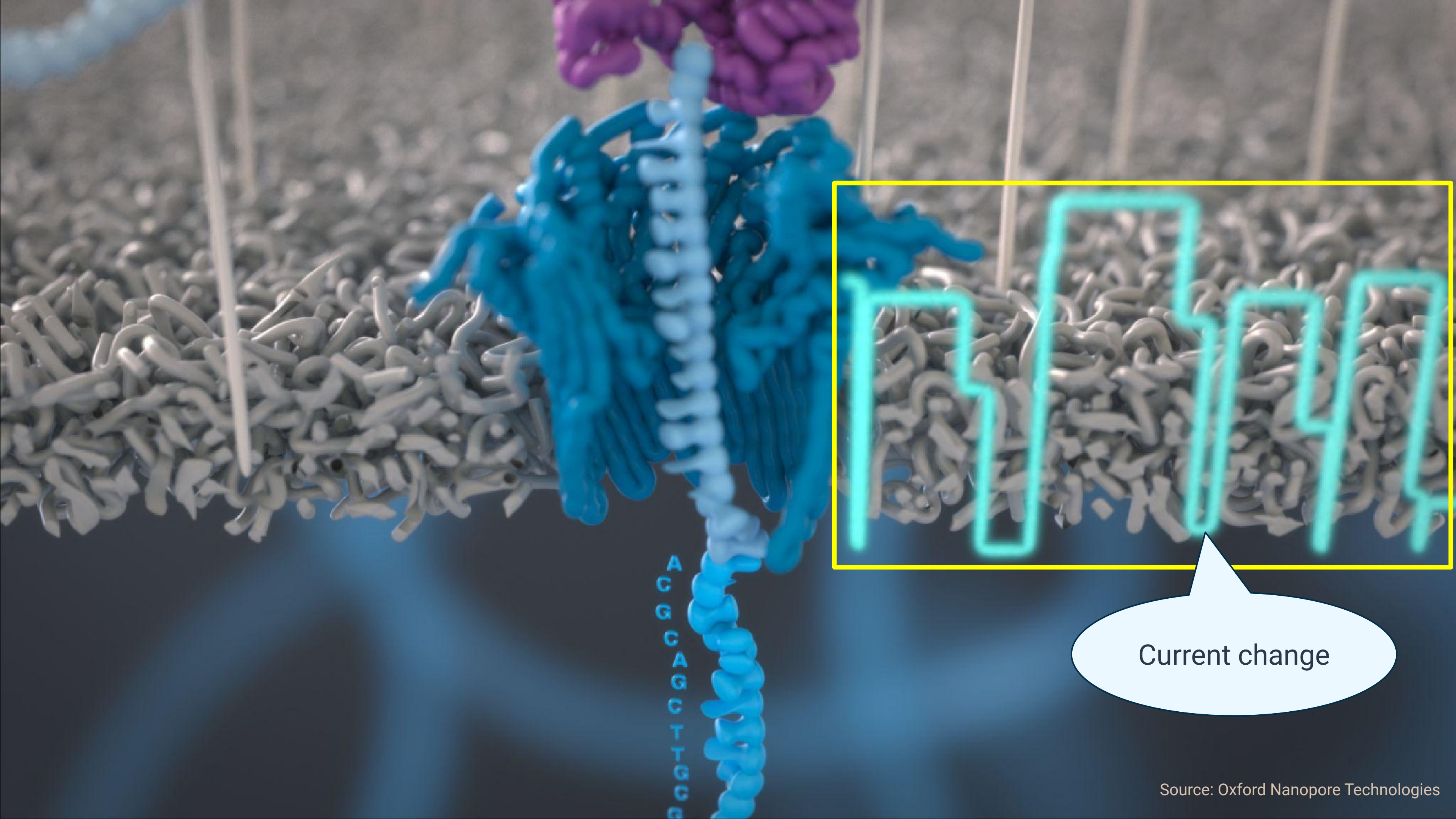


DNA molecule

Source: Oxford Nanopore Technologies

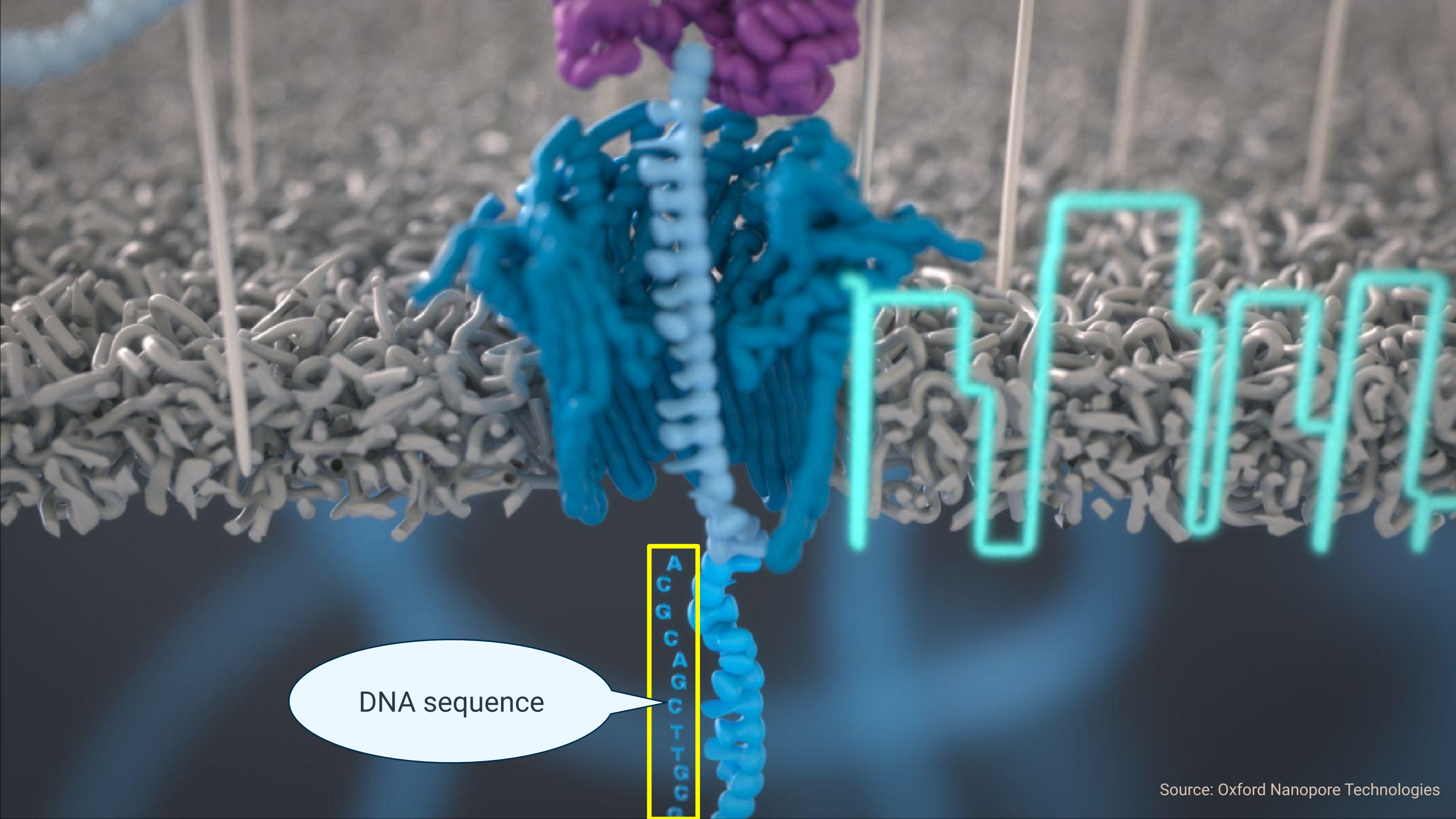


Source: Oxford Nanopore Technologies



Current change

Source: Oxford Nanopore Technologies



DNA sequence

Source: Oxford Nanopore Technologies

The basics of how a nanopore works

1 - Two chambers, filled with fluid with ions, separated by a membrane with a tiny channel, the nanopore

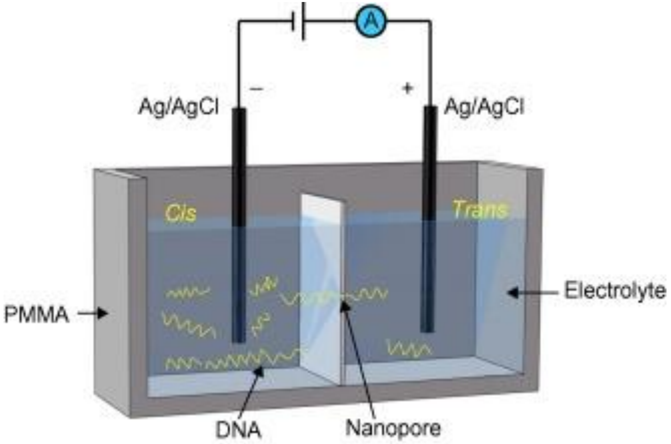
2 - Voltage difference between two chambers

3 - Electrolytes moving through the pores → base current *I*

5 - Whenever a DNA molecule slips through a pore, it transiently blocks the pore

6 - This leads to transient changes in *I*.

→ Just plug in a digital ammeter, a neural network, and we have a DNA sequencer!



Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). *Nanopore-based Fourth-generation DNA Sequencing Technology*. In Genomics, Proteomics & Bioinformatics (Vol. 13, Issue 1, pp. 4–16). Elsevier BV. https://doi.org/10.1016/j.gpb.2015.01.009

PMMA = polymethyl methacrylate = Plexiglas®

DNA moves through pore

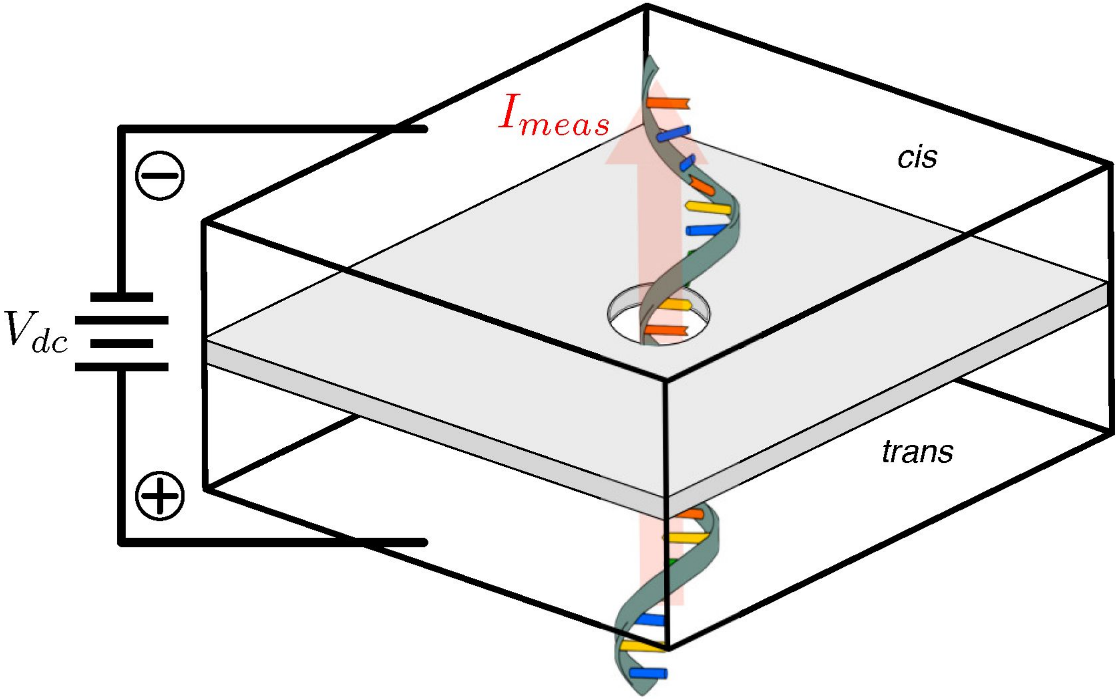
↓

Change in I

Oblique view of simplified nanopore structure.

A thin, pore-infused, membrane separates the *cis* and *trans* chambers biased with DC voltage V_{dc}.

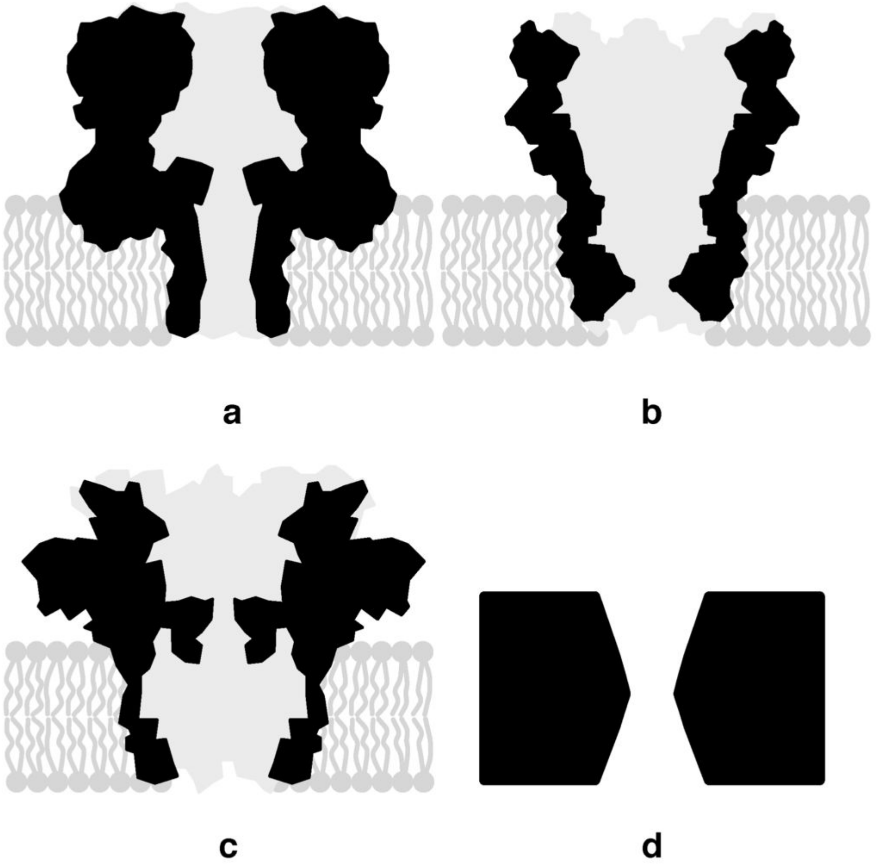
As the DNA blocks the pore, we can read corresponding changes in I_{meas}.



Magierowski, S., Huang, Y., Wang, C., & Ghafar-Zadeh, E. (2016). *Nanopore-CMOS Interfaces for DNA Sequencing*. In Biosensors (Vol. 6, Issue 3, p. 42). MDPI AG. https://doi.org/10.3390/bios6030042

Pore proteins taken from various bacteria

Cross-section of (a) the α−HL nanopore infused in a lipid bilayer membrane support structure, (b) MspA nanopore, (c) CsgG nanopore and (d) solid-state nanopore.



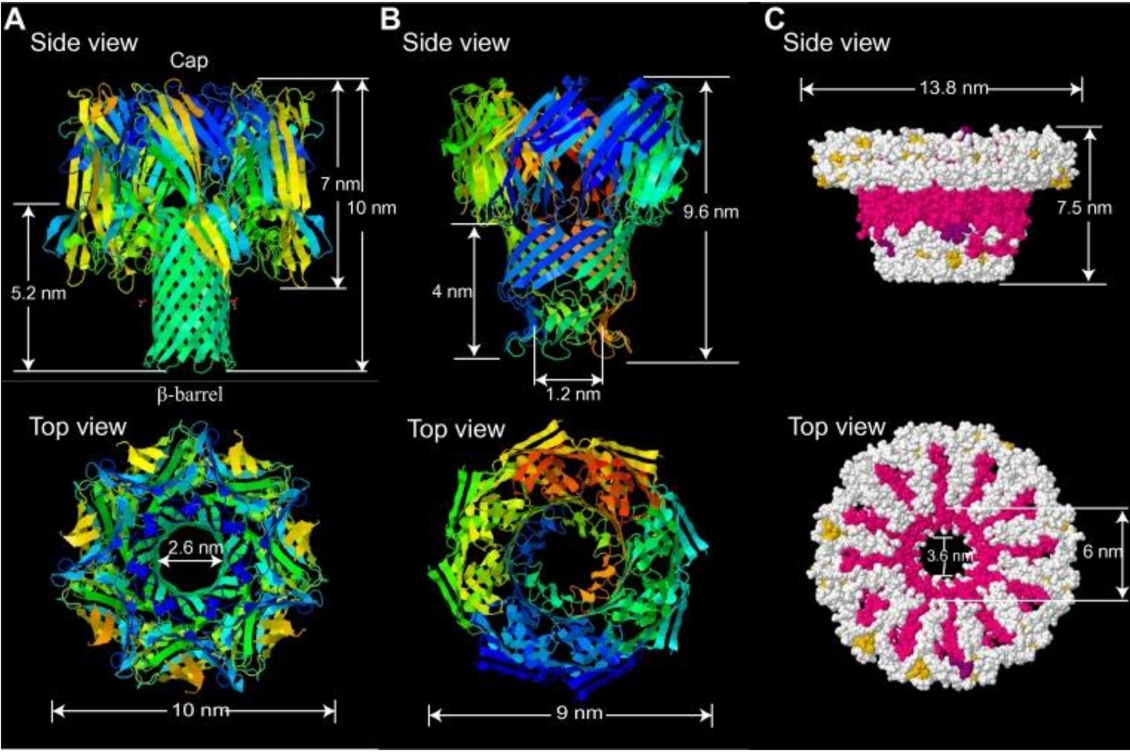
Magierowski, S., Huang, Y., Wang, C., & Ghafar-Zadeh, E. (2016). *Nanopore-CMOS Interfaces for DNA Sequencing*. In Biosensors (Vol. 6, Issue 3, p. 42). MDPI AG. https://doi.org/10.3390/bios6030042

Structural view of a nanopore

α-Hemolysin (α-HL, also called α-toxin) is the first and most commonly used biological nanopore

α-HL is an exotoxin secreted by the bacterium *Staphylococcus aureus*

(S. aureus is one of the most important bacteria that cause disease in humans)



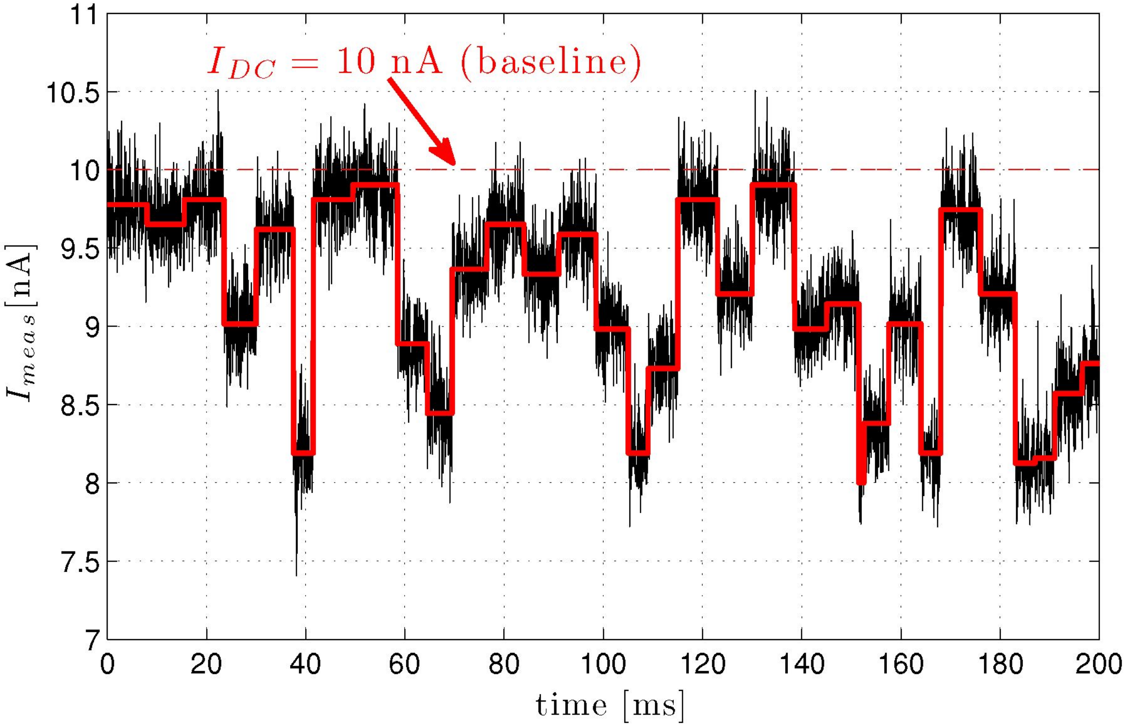
Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). *Nanopore-based Fourth-generation DNA Sequencing Technology*. In Genomics, Proteomics & Bioinformatics (Vol. 13, Issue 1, pp. 4–16). Elsevier BV. https://doi.org/10.1016/j.gpb.2015.01.009

As DNA slips through the pore, current changes

Example illustration of modulated current through a nanopore.

The modulation of the current curve corresponds to the DNA sequence of the ssDNA going through the pore

The relationship between DNA sequence and current levels is complex

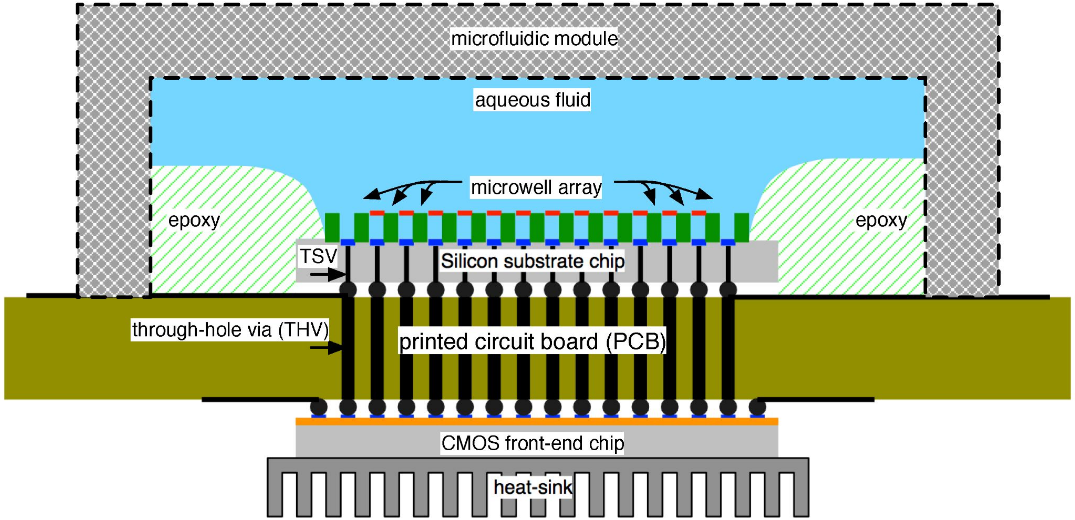


Magierowski, S., Huang, Y., Wang, C., & Ghafar-Zadeh, E. (2016). *Nanopore-CMOS Interfaces for DNA Sequencing*. In Biosensors (Vol. 6, Issue 3, p. 42). MDPI AG. https://doi.org/10.3390/bios6030042

Nanopores on a chip

Cross-section of an example construct for connecting the nanopore sensor system to external system components.

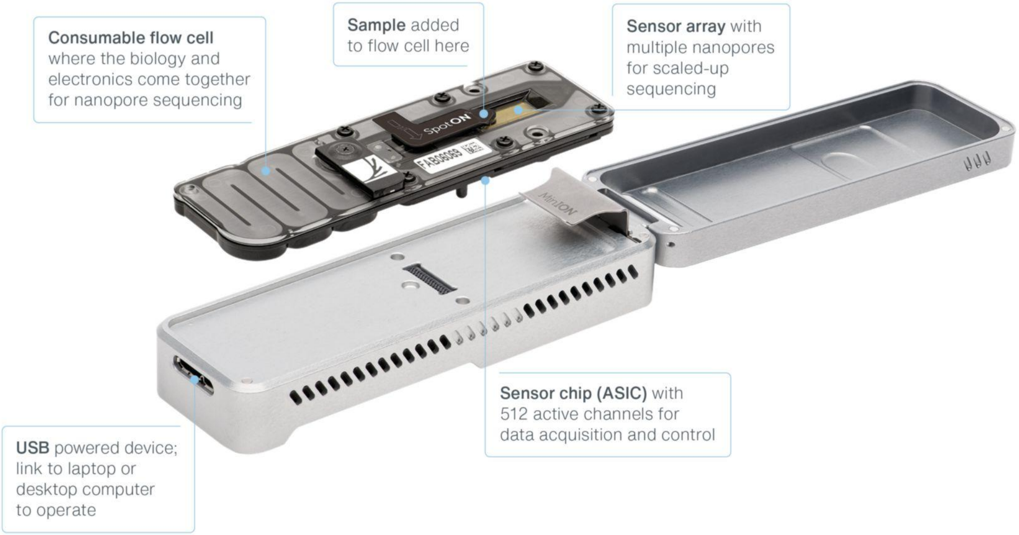
The construction of sensing structures and electronic structures on separate silicon substrates offers substantial room for the optimization of each component.



Magierowski, S., Huang, Y., Wang, C., & Ghafar-Zadeh, E. (2016). *Nanopore-CMOS Interfaces for DNA Sequencing*. In Biosensors (Vol. 6, Issue 3, p. 42). MDPI AG. https://doi.org/10.3390/bios6030042

Specs for the ONT MinION nanopore

- 512 channels per flow cell
- 4 kHz sampling rate
- ~400 bps (bases per second)
- ~48 hours per flow-cell
- ~50 Gb per flow-cell



Consumable flow cell where the biology and electronics come together for nanopore sequencing

Sample added to flow cell here

Sensor array with multiple nanopores for scaled-up sequencing

USB powered device; link to laptop or desktop computer to operate

Sensor chip (ASIC) with 512 active channels for data acquisition and control

Source: Oxford Nanopore Technologies

DNA fragmentation

The DNA is fragmented in the process

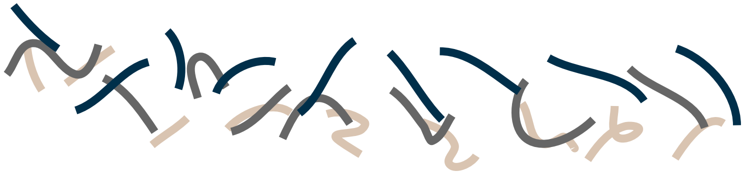
Each fragment is read separately

Leads to a puzzle game at the other end

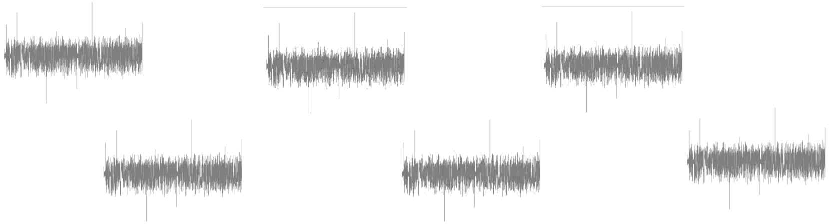
AGGAACTGCCGATCTTAATGGATGGCCGGAGG

True DNA sequence

Multiple copies of the DNA (each from its own cell)

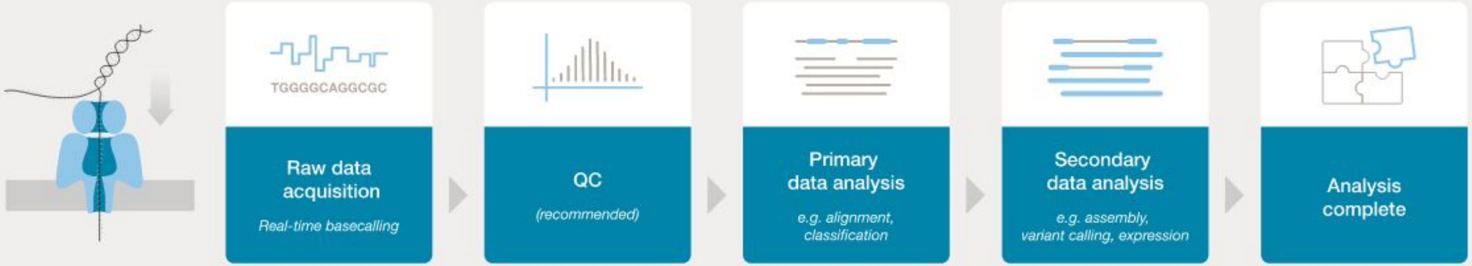


Fragmented and sequenced together



One squiggle per fragment

How to do the data analysis



Raw data acquisition

Real-time basecalling

QC

(recommended)

Primary data analysis

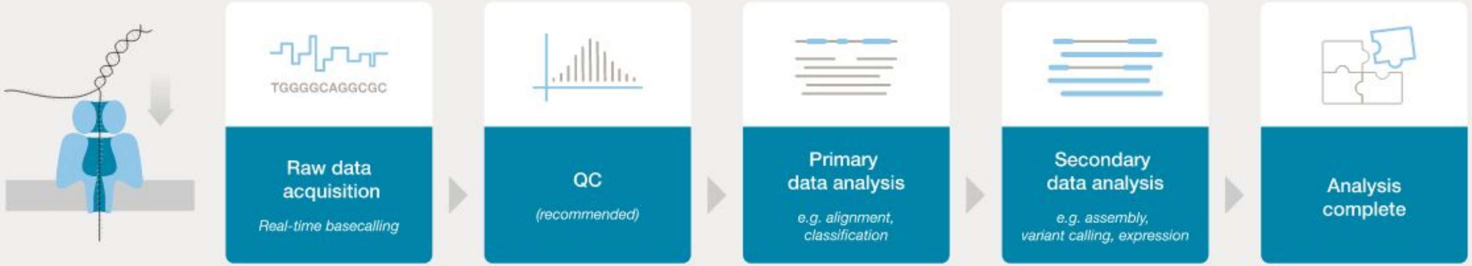
e.g. alignment, classification

Secondary data analysis

e.g. assembly, variant calling, expression

Analysis complete

Data analysis pipeline at a glance



Practically all of the tooling for the ONT sequencers are open source.

Data analysis pipeline at a glance

The MinION raw data

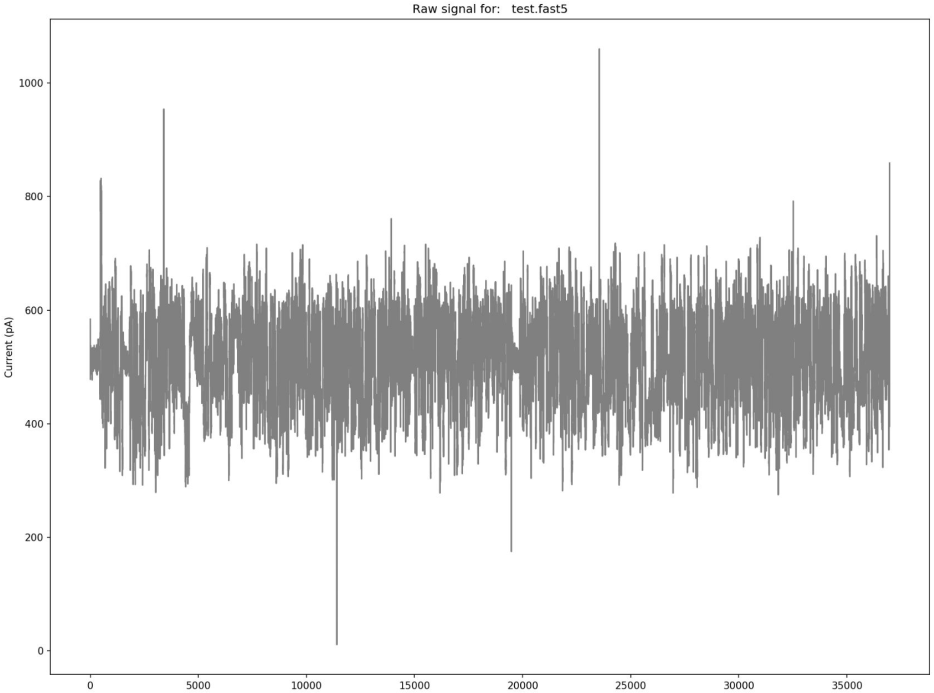
Old format: FAST5 (pre 2022)

New format: POD5 (post 2022)

Formats are functionally equivalent

- Raw sensor signal
- Metadata
- DNA sequence itself (optional)

```
> python SquigglePlot.py -i ~/data/test.fast5
```



Source: SquiggleKit

Terminology

In biology/informatics/bioinformatics, everything tends to have at least two names

I will try to be consistent in this talk

Base calling – assigning nucleobases to electrical current changes resulting from nucleotides passing through a nanopore

Alignment – the process of comparing and detecting similarities between DNA sequences

Assembly – aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence

Fragment - a piece of DNA (“substring”)

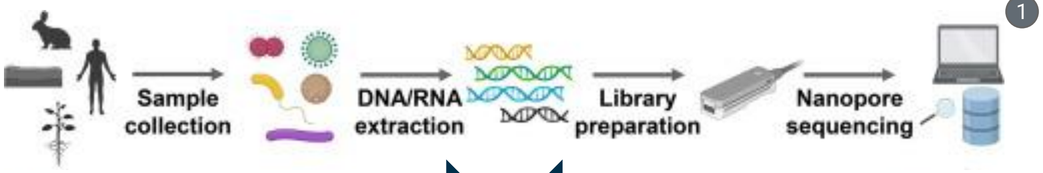
Read - sequence of bases corresponding to a fragment

DNA sequencing at 20,000 ft

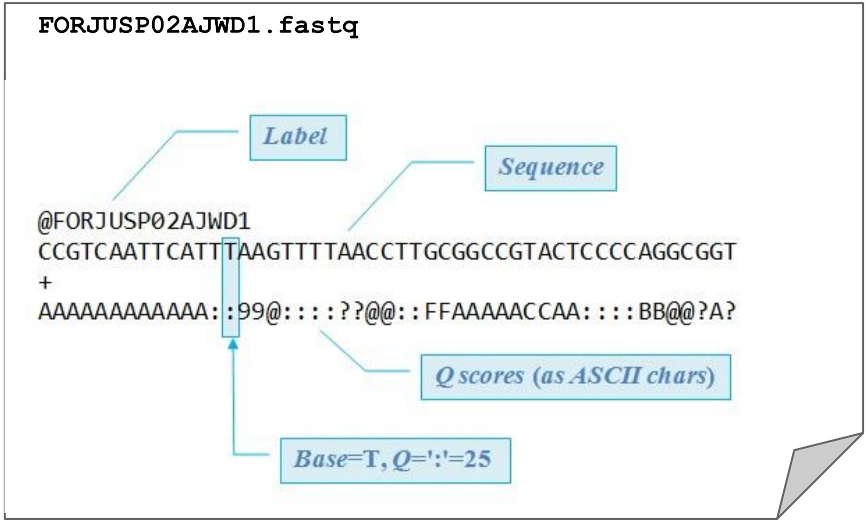
DNA molecule read using chemistry and electronics

↓ *(base calling)*

File with nucleotide sequence (typically FASTQ files)



[1]



FORJUSP02AJWD1.fastq

```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Label
Sequence
Q scores (as ASCII chars)
Base=T, Q=':'=25

[2]

[1] Laura Ciuffreda, Héctor Rodríguez-Pérez, Carlos Flores, *Nanopore sequencing and its application to the study of microbial communities*, Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 1497-1511.

[2] @RobertEdgarPHD

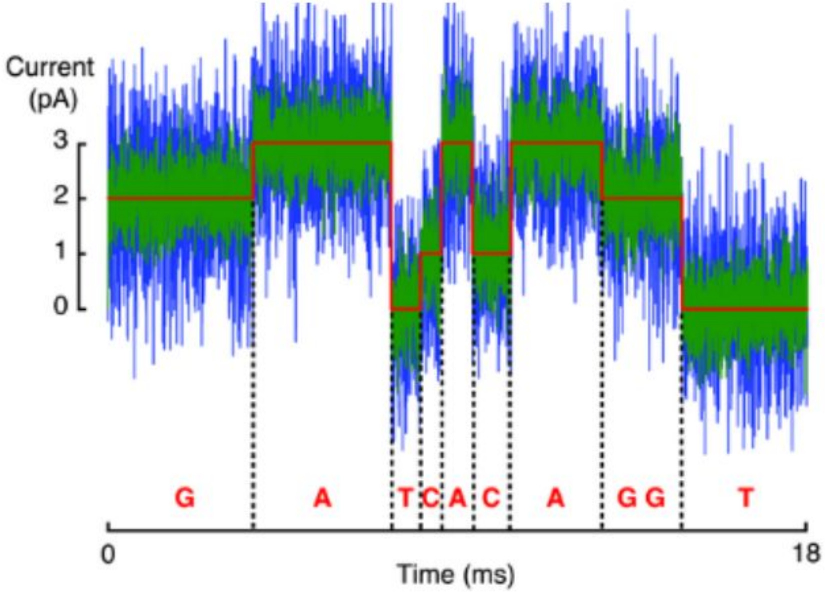
Base calling

Convert the raw signal (current over time) to a sequence of nucleotides (ACGT)

Area of very active research

Neural networks work well for this

Effect: FAST5 file → FASTQ file



Source: Ivan Gesteira Costa, IZKF Research Group Bioinformatics

The dorado basecaller is the latest and greatest from ONT. Open-source, on GitHub.

Quality control

Q scores

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

Q is the phred* quality score

P is probability

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%

Source: Wikipedia

** Phred is a classical computer program for base calling*

FASTQ files

Each “read” in a FASTQ file

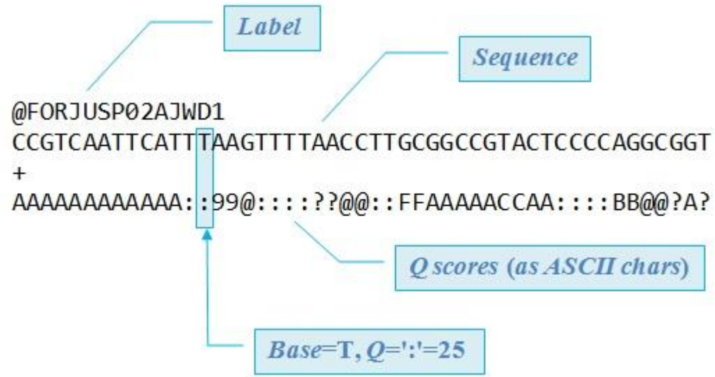
- Takes up four lines
  - Label
  - Sequence
  - Separator (always +)
  - Base call quality scores

A FASTQ file may contain many reads

Label - A sequence identifier, no globally accepted standard for this

Sequence - [ACGT]+

Quality - runs from
    0x21 (lowest quality; '!' in ASCII) to
    0x7e (highest quality; '~' in ASCII).



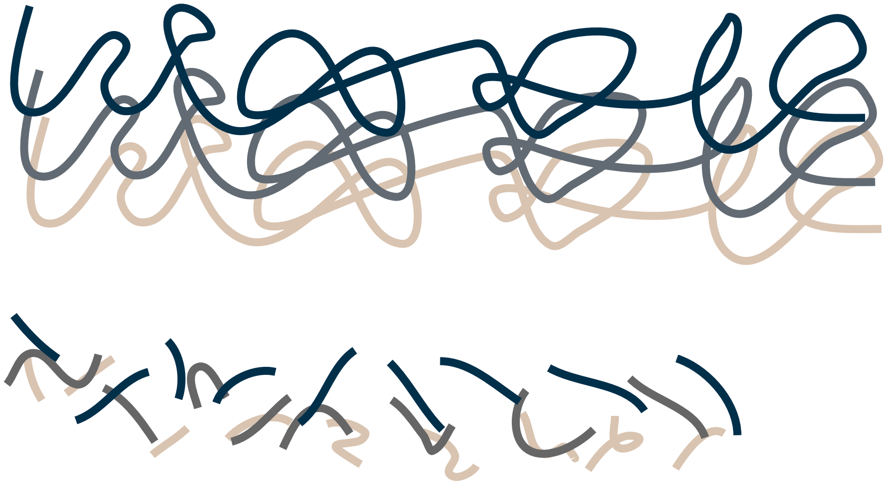
```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Alignment

After base calling, we have a large collection of DNA fragments that we have to order in some way

```
AGGAACTGCCGATCTTAATGGATGGCCGGAGG
```

True DNA sequence



Multiple copies of the DNA (each from its own cell)

Fragmented and sequenced together

```
GAT AACTGCC CTGCC AGG TCTT GATC
TTAATG AGGAA CTTAATG AACTG AGG
AGG GATGG GATG CCGA GCCGG TGGC
AATGGA CGGAGG CCGGAGG
```

Puzzle game galore!

Alignment

If we know the species we're sequencing for, we can use a ***reference genome***

AGGAACTGCCGATCTTAATGGATGGCCGGAGG

True DNA sequence

AGGAACT**C**CCGATCTTA**T**TGGATG**T**CCGGAGG

Reference DNA sequence

GAT AACTGCC CTGCC AGG TCTT GATC
TTAATG AGGAA CTTAATG AACTG AGG
AGG GATGG GATG CCGA GCCGG TGGC
AATGGA CGGAGG CCGGAGG

Puzzle game galore!

Alignment

If we know the species we’re sequencing for, we can use a ***reference genome***

Match each fragments to the most similar part of the reference

AGGAACTGCCGATCTTAATGGATGGCCGGAGG

True DNA sequence

AGGAACT**C**CCGATCTTA**T**TGGATG**T**CCGGAGG

Reference DNA sequence

GAT AACTGCC CTGCC **AGG** TCTT GATC
TTAATG **AGGAA** CTTAATG AACTG AGG
AGG GATGG GATG CCGA GCCGG TGGC
AATGGA CGGAGG CCGGAGG

Puzzle game galore!

Alignment

If we know the species we're sequencing for, we can use a ***reference genome***

Match each read to the most similar part of the reference

AGGAACTGCCGATCTTAATGGATGGCCGGAGG

True DNA sequence

AGGAA**CTCCC**GATCTTA**T**TGGATG**T**CCGGAGG

Reference DNA sequence

GAT AACTGCC **CTGCC** ~~AGG~~ TCTT GATC

TTAATG ~~AGGAA~~ CTTAATG AACTG AGG

~~AGG~~ GATGG GATG CCGA GCCGG TGGC

AATGGA CGGAGG CCGGAGG

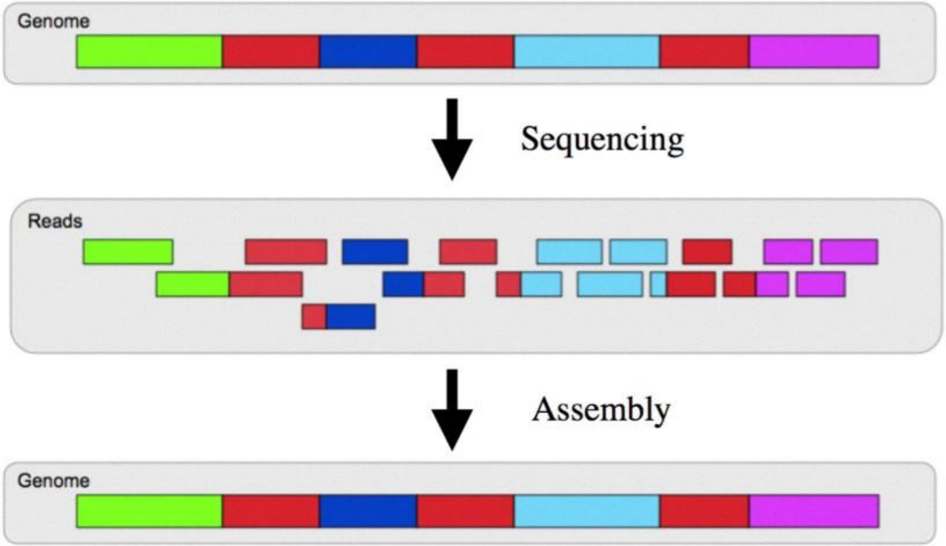
Puzzle game galore!

Assembly

De novo assembly — when the DNA is entirely unknown

Two algorithmic approaches

- Greedy: local optimum
- Graph-based: global optimum



Source: Ontario Institute of Cancer Research

Examples of applications and application areas

DNA sequencing – in space!

Problem: … in space!

Kate brought DNA from

- A mouse
- The E.coli bacteria
- The lambda phage virus

Next mission: sequence true aliens?

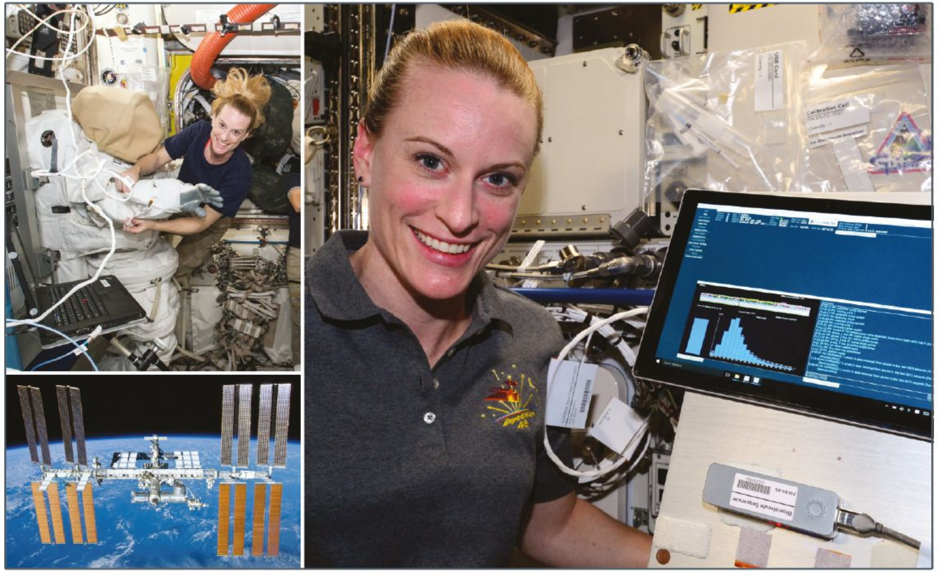


Fig. 1 Astronaut Kate Rubins on the ISS

Source: NASA

DNA sequencing – in the field

Problem: Which species live in this area?

“The remarkable accuracy and low computational demand of [our] pipeline, together with the inexpensive equipment and simple protocols, make the proposed workflow particularly suitable for tracking species under field conditions.”

Sequence all the poop!

Source: Ph. Francesco Ciccotti / Getty Images

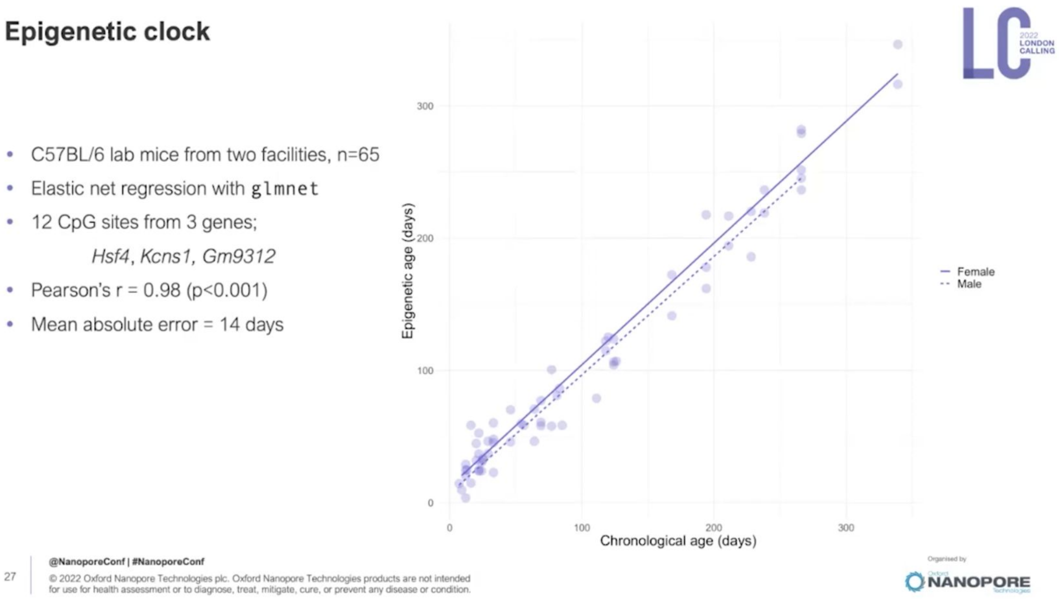
Maestri, Cosentino, Paterno, Freitag, Garces, Marcolungo, Alfano, Njunjić, Schilthuizen, Slik, Menegon, Rossato, & Delledonne. (2019). *A Rapid and Accurate MinION-Based Workflow for Tracking Species Biodiversity in the Field.* In Genes (Vol. 10, Issue 6, p. 468). MDPI AG. https://doi.org/10.3390/genes10060468

Going beyond DNA sequencing

More poop sequencing!

Problem: *How old was the mouse when it took laid down this particular crap?*

(Yes, scientists actually get paid to answers these important questions!)



Epigenetic clock

- C57BL/6 lab mice from two facilities, n=65
- Elastic net regression with `glmnet`
- 12 CpG sites from 3 genes;

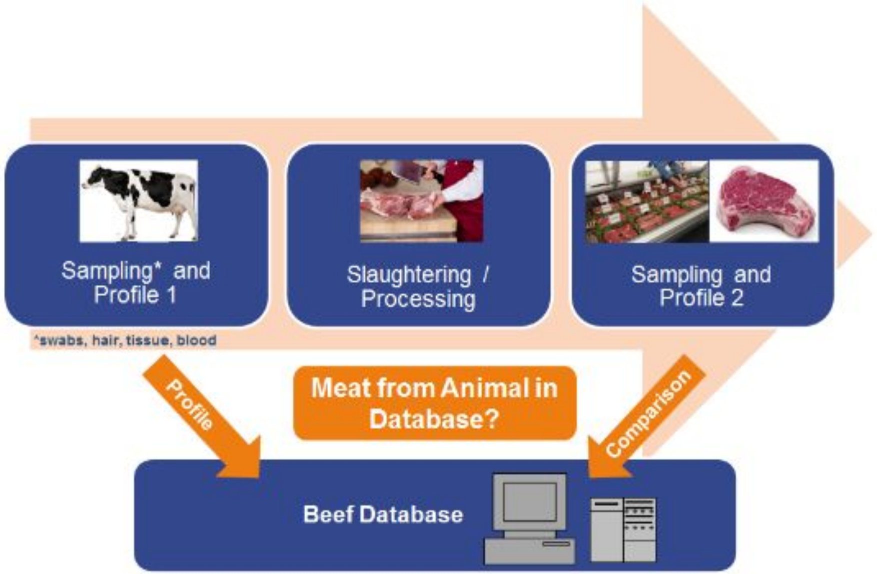
  Hsf4, *Kcns1*, *Gm9312*
- Pearson's r = 0.98 (p<0.001)
- Mean absolute error = 14 days

@NanoporeConf | #NanoporeConf

[Eveliina Hanski: A non-invasive, MinION-based method for determining the epigenetic age of mice]()

Genomic meat sourcing

Problem: Where did this beef really come from?


Sampling* and Profile 1
Slaughtering / Processing
Sampling and Profile 2
*swabs, hair, tissue, blood
Profile
Meat from Animal in Database?
Comparison
Beef Database

Source: Eurofins Food & Feed Testing Germany

How to get started yourself

Things to do if you’re curious

Without buying a sequencer

- Look at the EPI2ME Labs Tutorials
- Download the EPI2ME Labs
- Download the CliveOME from S3

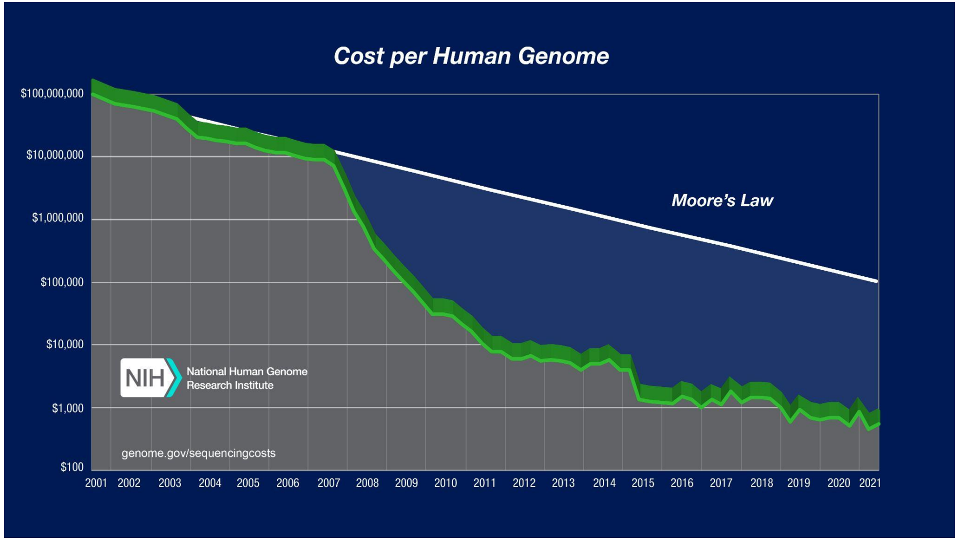
Buying a sequencer

- Complete a (wet) lab course
- Get a MinION
- Sequence the lambda phage

Wrap-up

Conclusion

- The democratizing* craze has reached DNA sequencing
- You can buy a brand new DNA sequencer for $1000
- You can do DNA sequencing in the field
- All software necessary to interpret the raw output from the sequencer is open source

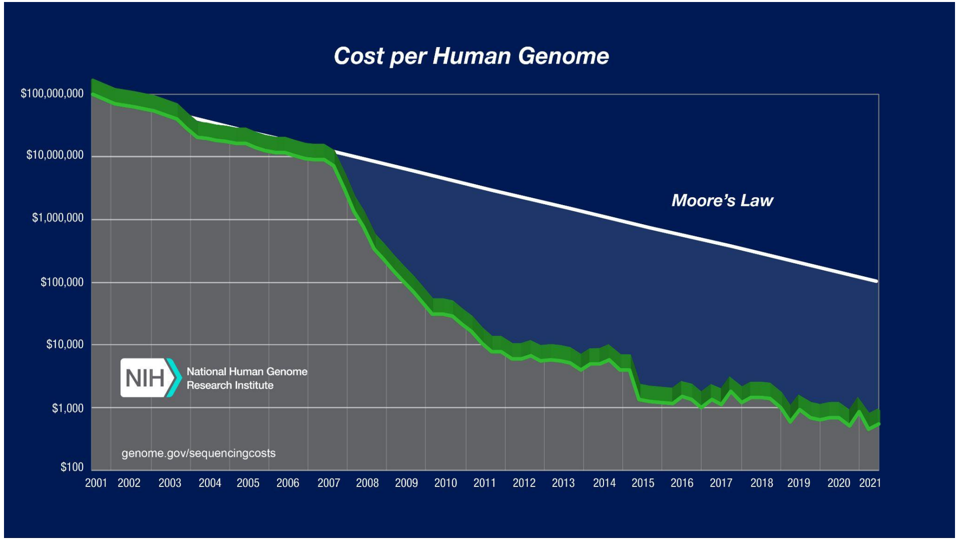


Source: National Human Genome Research Institute

* Democratization of technology refers to the process by which access to technology rapidly continues to become more accessible to more people.

Conclusion

- The democratizing* craze has reached DNA sequencing
- You can buy a brand new DNA sequencer for $1000
- You can do DNA sequencing in the field
- All software necessary to interpret the raw output from the sequencer is open source
- ***From a piece of poop, you can determine someone's age, gender and eye color***



Source: National Human Genome Research Institute

* Democratization of technology refers to the process by which access to technology rapidly continues to become more accessible to more people.