EPFL

Genomic solutions to sustainable development

Week 10 — Real-time genomics

29 April 2025

Sebastian M. Waszak, Ph.D.
Assistant Professor, Life Sciences, EPFL
Associate Adjunct Professor, Neurology, UCSF

École
polytechnique
fédérale
de Lausanne

The Atomic bomb, the Alta Summit, and the Human Genome Project

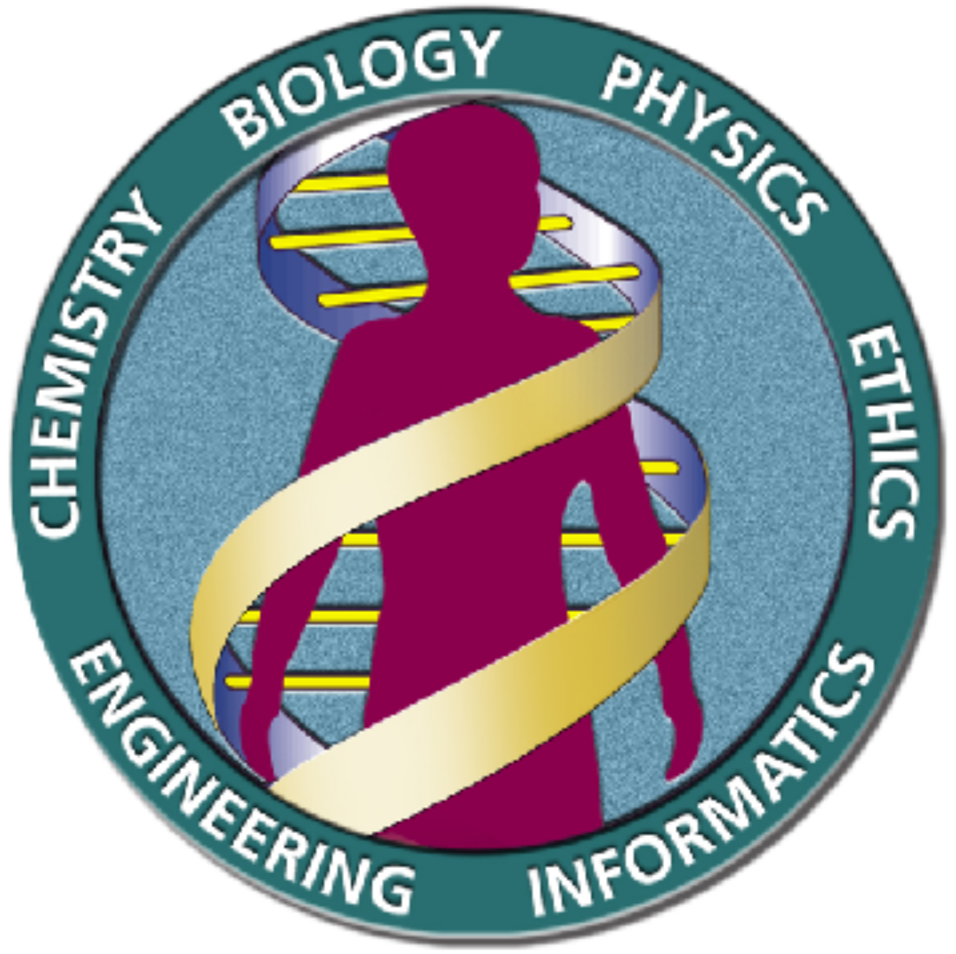
- The Alta Summit, December 1984, was sponsored by the US Department of Energy (DOE) and the International Commission for Protection Against Environmental Mutagens and Carcinogen. It gathered leading scientists, policymakers, and funding agency representatives to deliberate on the feasibility and implications of initiating a large-scale project to map and sequence the human genome.

- Key question asked asked to participants during the Alta Summit: **"Could new methods permit detection of mutations, and more specifically could any increase in the mutation rate among survivors of the Hiroshima and Nagasaki bombings be detected (in them or in their children)?"**

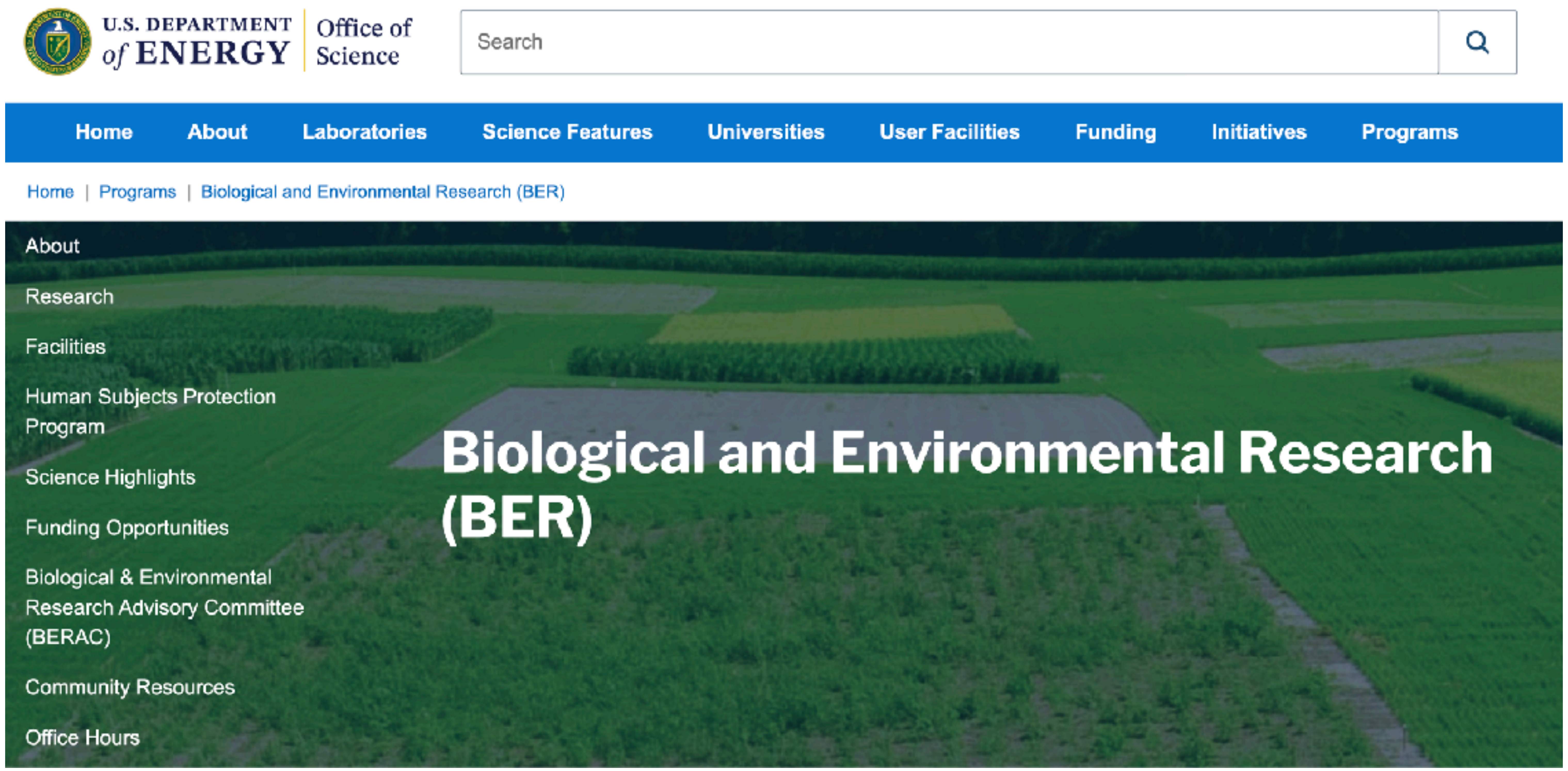
- Prior to this meeting, a meeting in Hiroshima, March 1984, concluded that new DNA analytical tools second highest priority for human mutation research. Why? Existing methods had failed to detect an increase in mutation rates in 12,000 children of Hiroshima-Nagasaki survivors due to low throughput of existing techniques. The meeting concluded that detecting the base rate of human mutations required sequencing the entire human genome.

- In 1986, Charles DeLisi, then the director of the **DOE's Health & Environmental Research Programs**, proposed to Congress what was to become the Human Genome Project (1990-2003)

- The Alta Summit is the bridge between the DOE's historical research program in mutation research (WWII, the Manhattan Project, and first atomic bomb) and push for a Human Genome Project



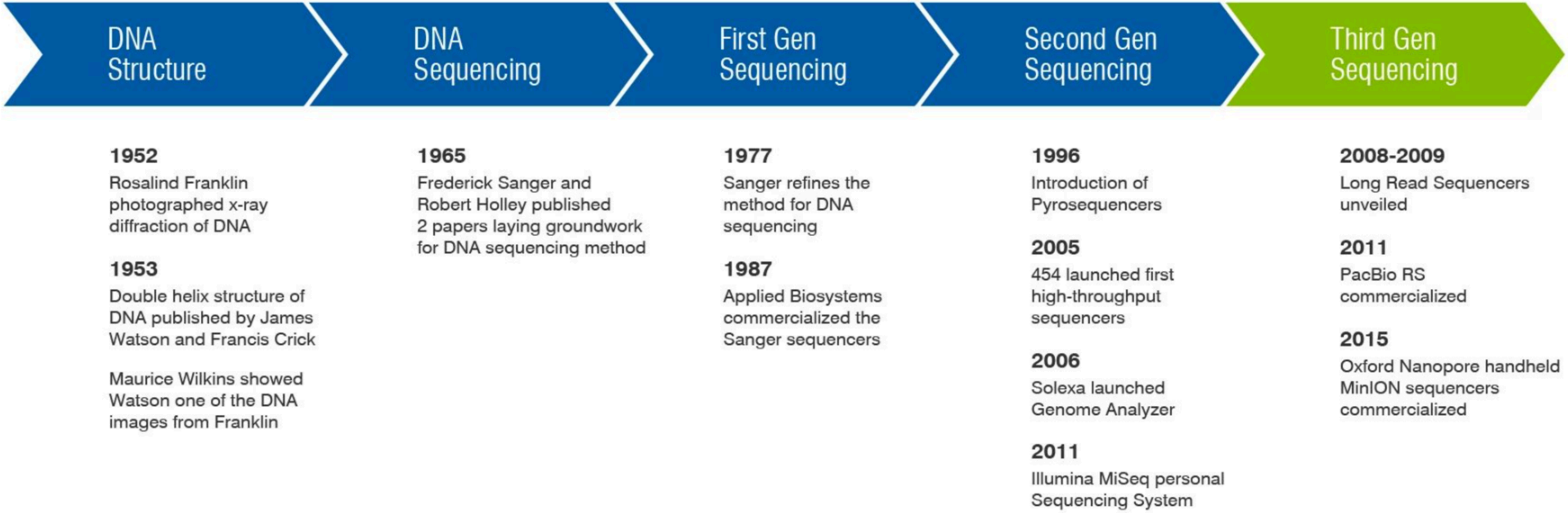
https://doe-humangenomeproject.ornl.gov/human-genome-project-timeline/



Biological and Environmental Research (BER)

The mission of the Biological and Environmental Research (BER) program is to support transformative science and scientific user facilities to achieve a predictive understanding of complex biological, earth, and environmental systems for energy and infrastructure security, independence, and prosperity. **The program seeks to understand the biological, biogeochemical, and physical processes that span from molecular and genomics-controlled scales to the regional and global scales that govern changes in watershed dynamics, climate, and the earth system.**

History of DNA sequencing



DNA Structure

1952
Rosalind Franklin photographed x-ray diffraction of DNA

1953
Double helix structure of DNA published by James Watson and Francis Crick

Maurice Wilkins showed Watson one of the DNA images from Franklin

DNA Sequencing

1965
Frederick Sanger and Robert Holley published 2 papers laying groundwork for DNA sequencing method

First Gen Sequencing

1977
Sanger refines the method for DNA sequencing

1987
Applied Biosystems commercialized the Sanger sequencers

Second Gen Sequencing

1996
Introduction of Pyrosequencers

2005
454 launched first high-throughput sequencers

2006
Solexa launched Genome Analyzer

2011
Illumina MiSeq personal Sequencing System

Third Gen Sequencing

2008-2009
Long Read Sequencers unveiled

2011
PacBio RS commercialized

2015
Oxford Nanopore handheld MinION sequencers commercialized

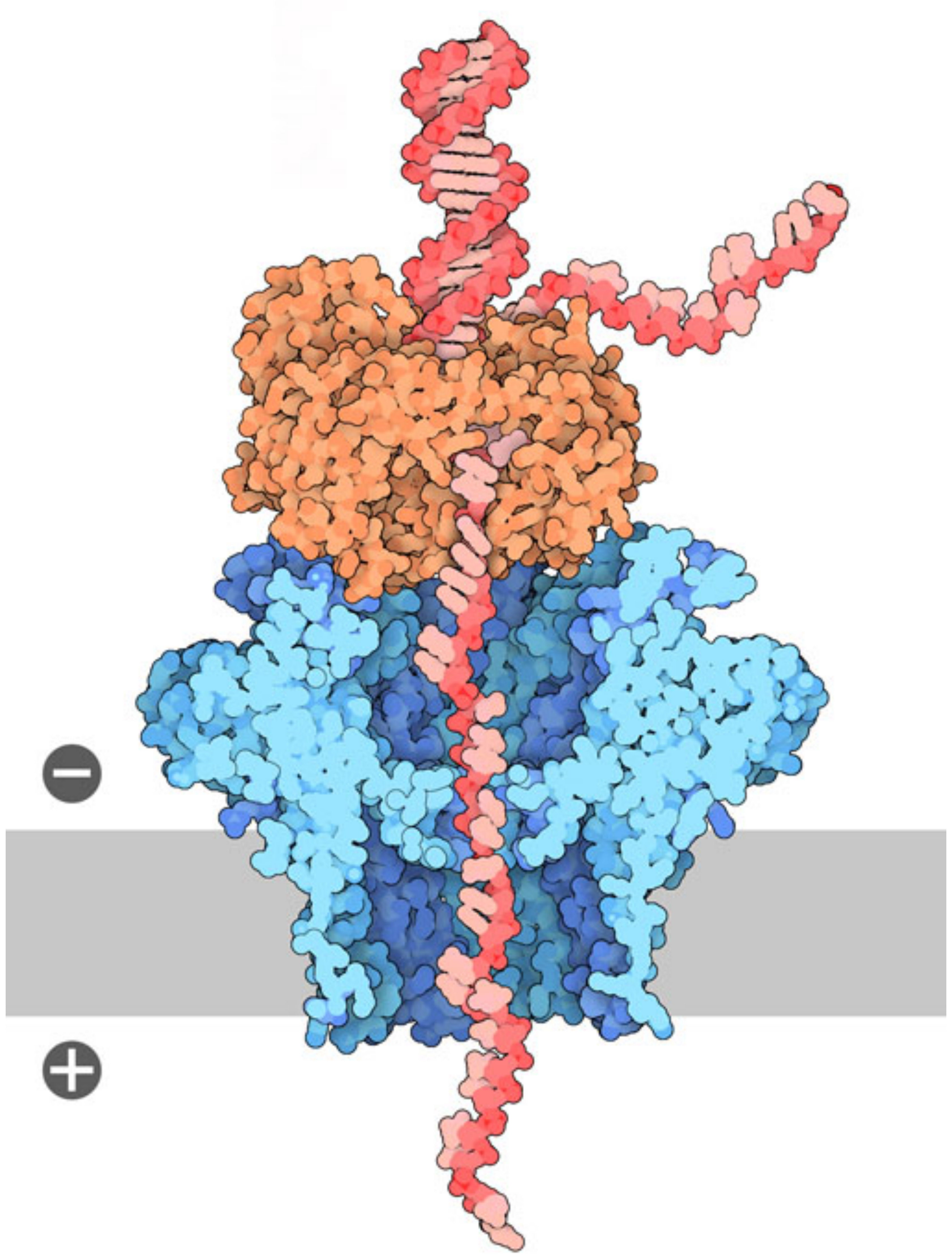
Charles river

Nanopore DNA sequencing

- Molecules pass through a pore in a membrane (eg, Curli production assembly/transport component CsgG, a membrane protein of E. coli)

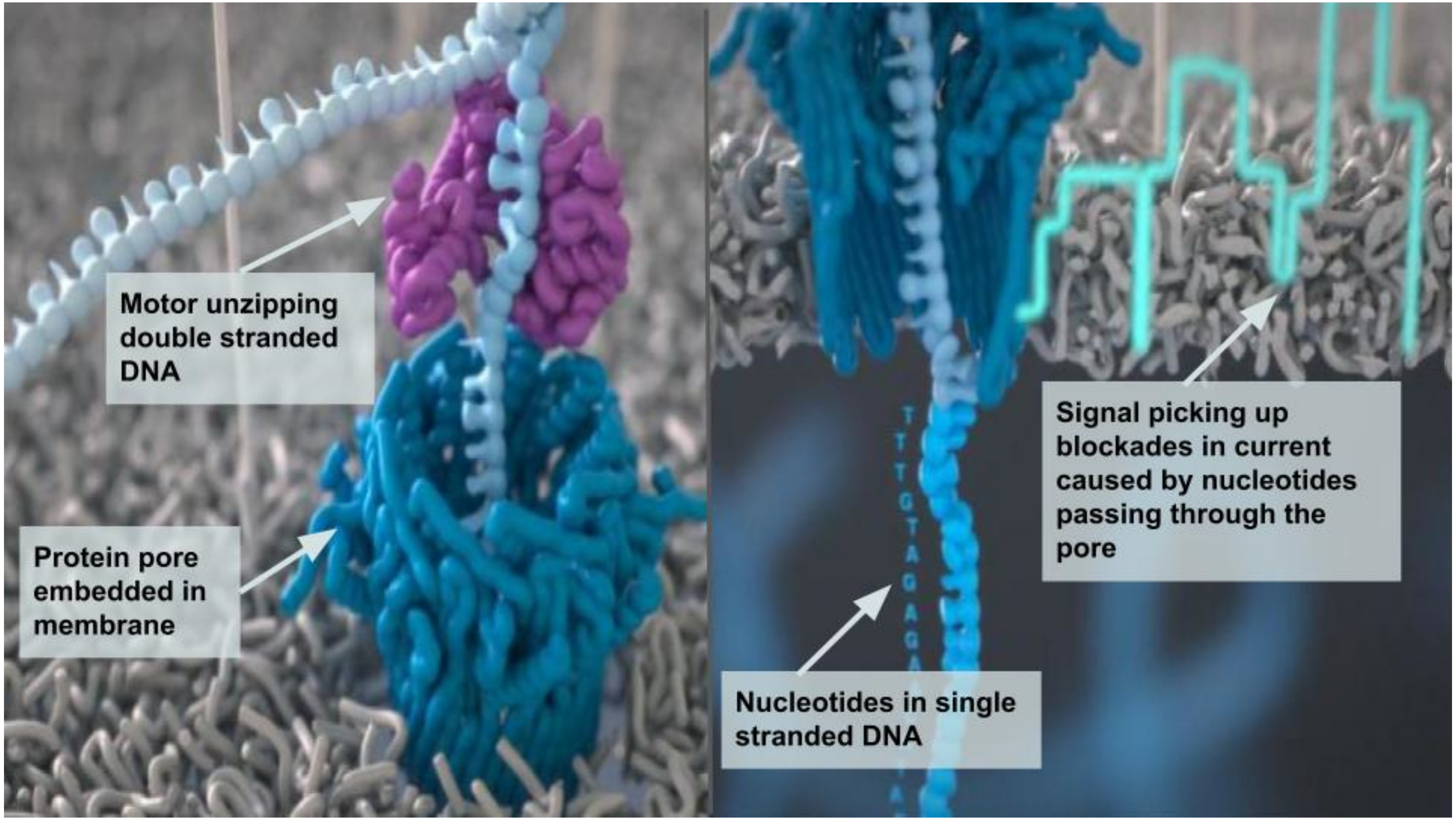
- Passing molecules block ion current and the length of of ion current blockage is sequence specific

- Ion current change profiles can be translated into unmodified and modified sequence (eg, A, C, T, G, 5mC)



BIO-312 Genomic solutions to sustainable development

Nanopore DNA sequencing



Motor unzipping double stranded DNA

Protein pore embedded in membrane

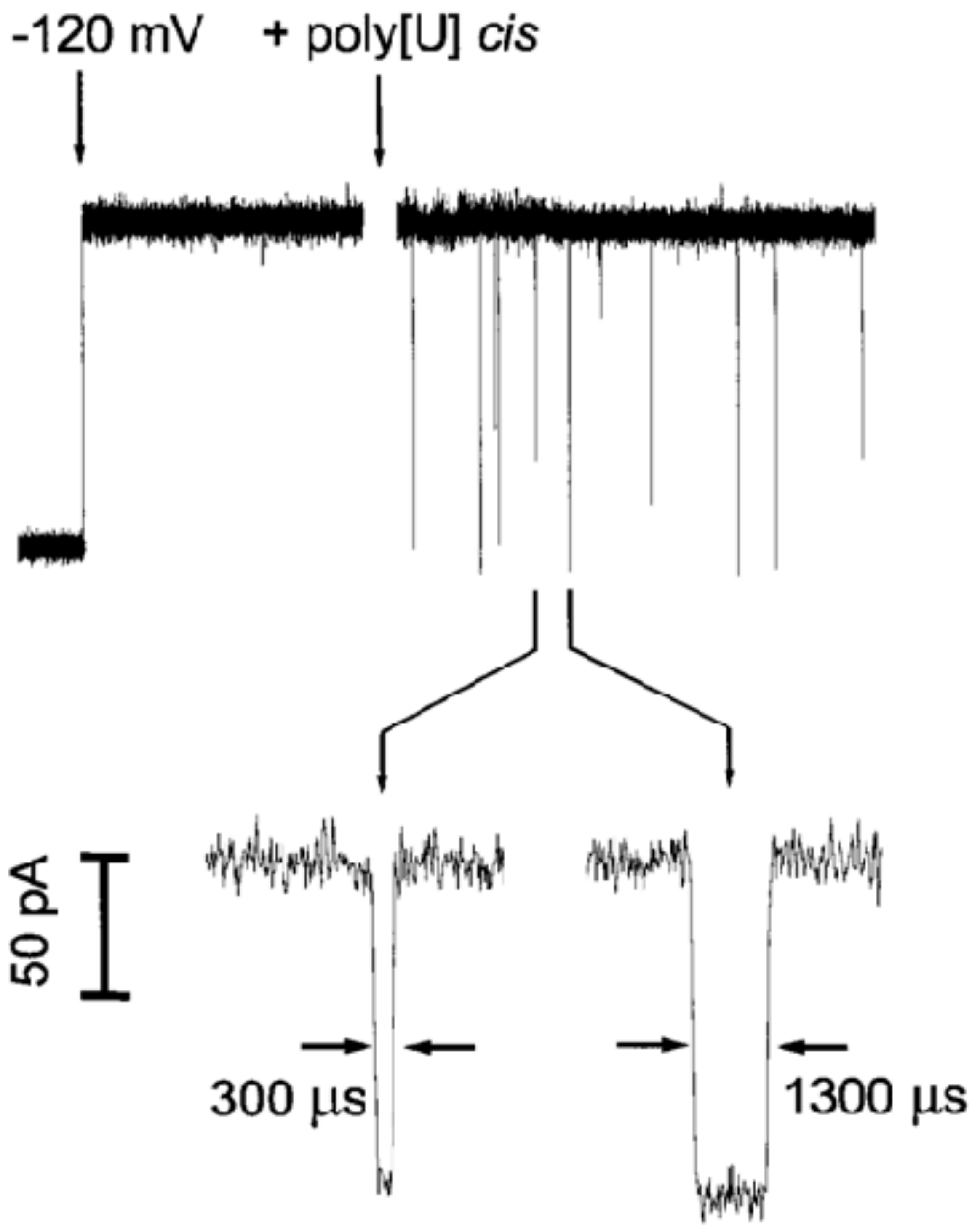
Signal picking up blockades in current caused by nucleotides passing through the pore

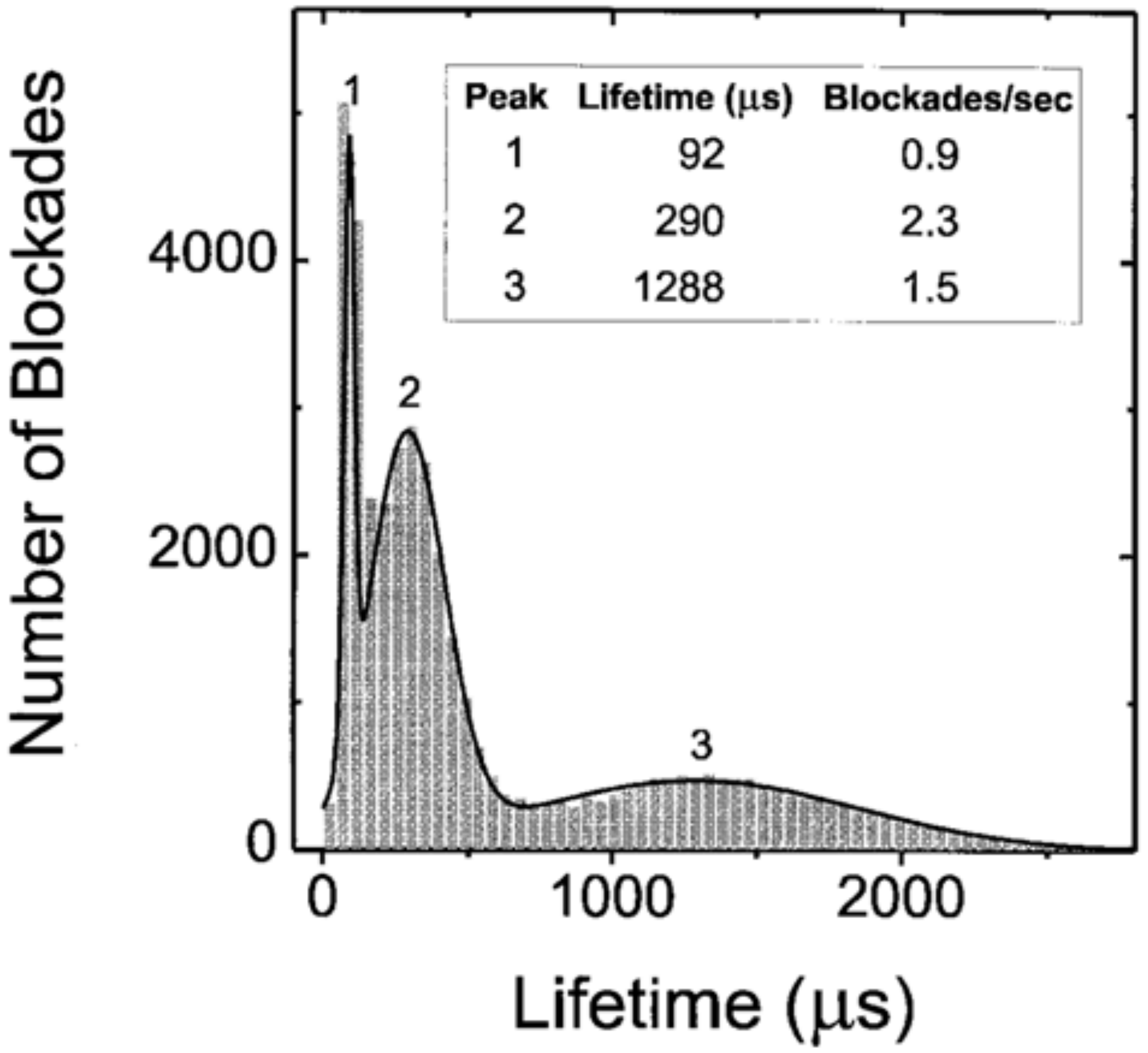
Nucleotides in single stranded DNA

Nanopore DNA sequencing

How it all started

Characterization of individual polynucleotide molecules using a membrane channel





Peak	Lifetime (µs)	Blockades/sec
1	92	0.9
2	290	2.3
3	1288	1.5

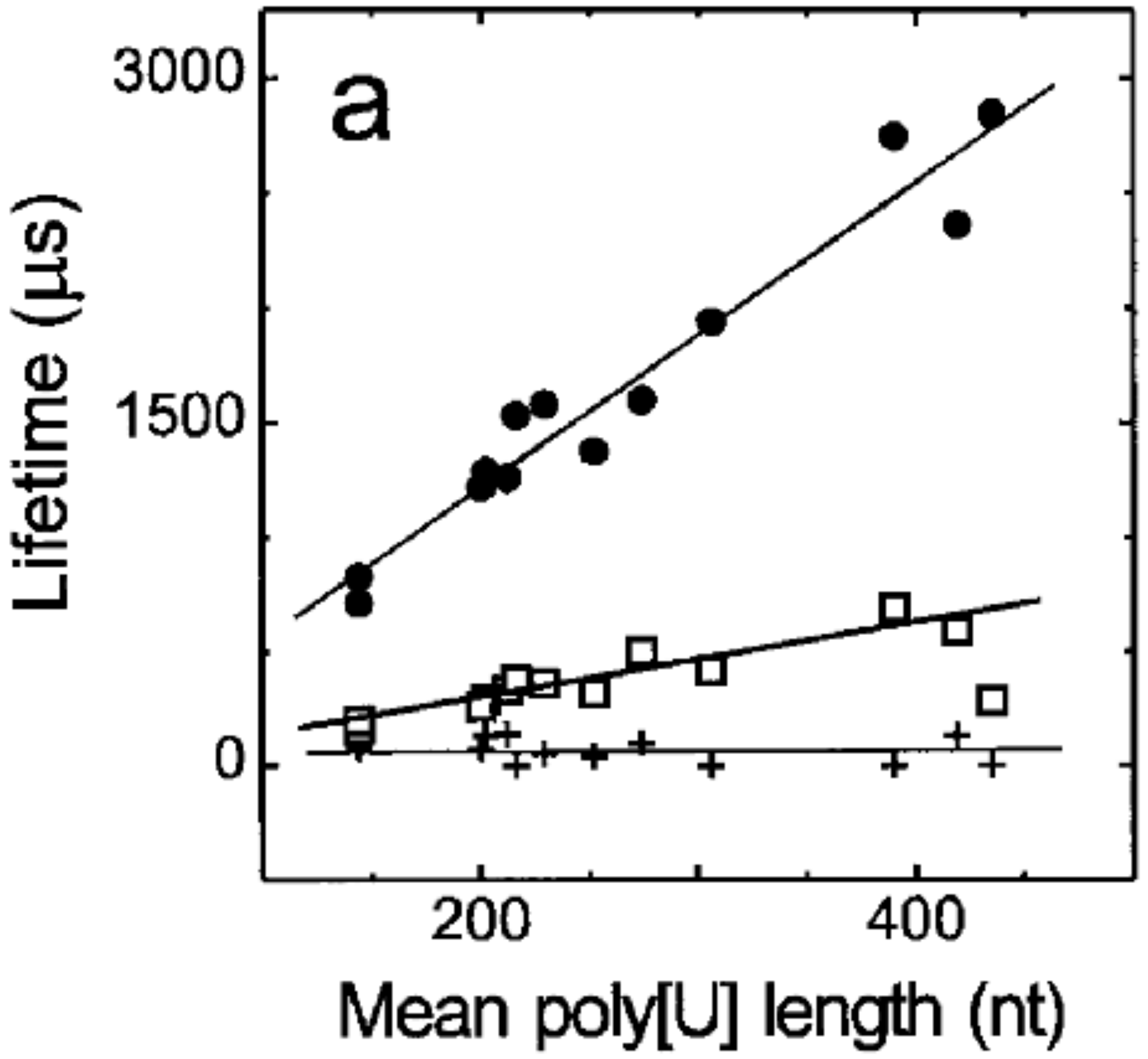


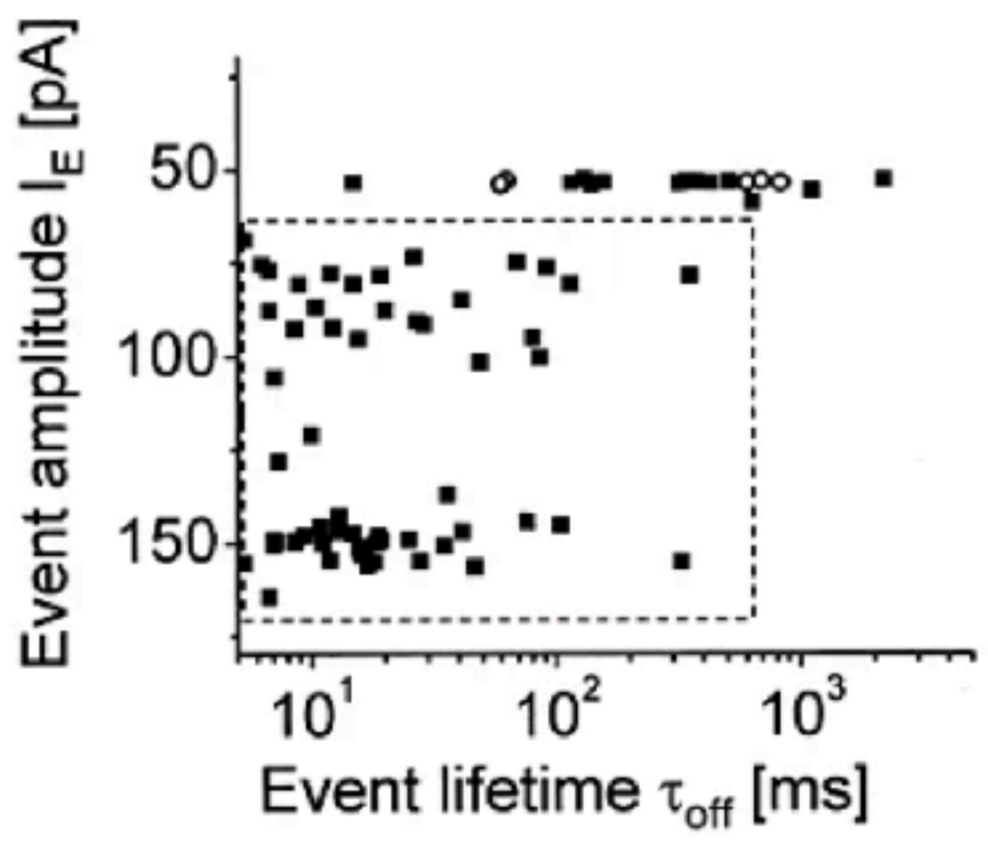
FIG. 3. Poly[U]-induced channel blockade lifetimes were proportional to (*a*) mean polymer length and (*b*) inversely proportional to applied voltage. The plots show lifetimes for (*a*) peaks 1 (+), 2 (□), and 3 (●) in experiments using $V = -120$ mV with 13 different size selected poly[U]s and (*b*) for peaks 2 (□), and 3 (●) with poly[U] of mean length 215 nt at the indicated voltages. Although the peak 1 lifetime appeared to be independent of the applied voltage (data not shown), because this lifetime (≈100 µs) was barely a factor of 2 greater than the time resolution of our system, a slight voltage dependence could not be ruled out.

Proc. Natl. Acad. Sci. USA Vol. 93, pp. 13770–13773, November 1996

Sequence-specific detection of individual DNA strands using engineered nanopores

Stefan Howorka[1*], Stephen Cheley[1], and Hagan Bayley[1,2]

We describe biosensor elements that are capable of identifying individual DNA strands with single-base resolution. Each biosensor element consists of an individual DNA oligonucleotide covalently attached within the lumen of the α-hemolysin (αHL) pore to form a "DNA–nanopore". The binding of single-stranded DNA (ssDNA) molecules to the tethered DNA strand causes changes in the ionic current flowing through a nanopore. On the basis of DNA duplex lifetimes, the DNA–nanopores are able to discriminate between individual DNA strands up to 30 nucleotides in length differing by a single base substitution. This was exemplified by the detection of a drug resistance–conferring mutation in the reverse transcriptase gene of HIV. In addition, the approach was used to sequence a complete codon in an individual DNA strand tethered to a nanopore.



○ oligo-181Y: 5'-ACAAAATCCAGACATAGTTATCT**A**TCAATA-3'

■ oligo-181C: 5'-ACAAAATCCAGACATAGTTATCTGTCAATA-3'

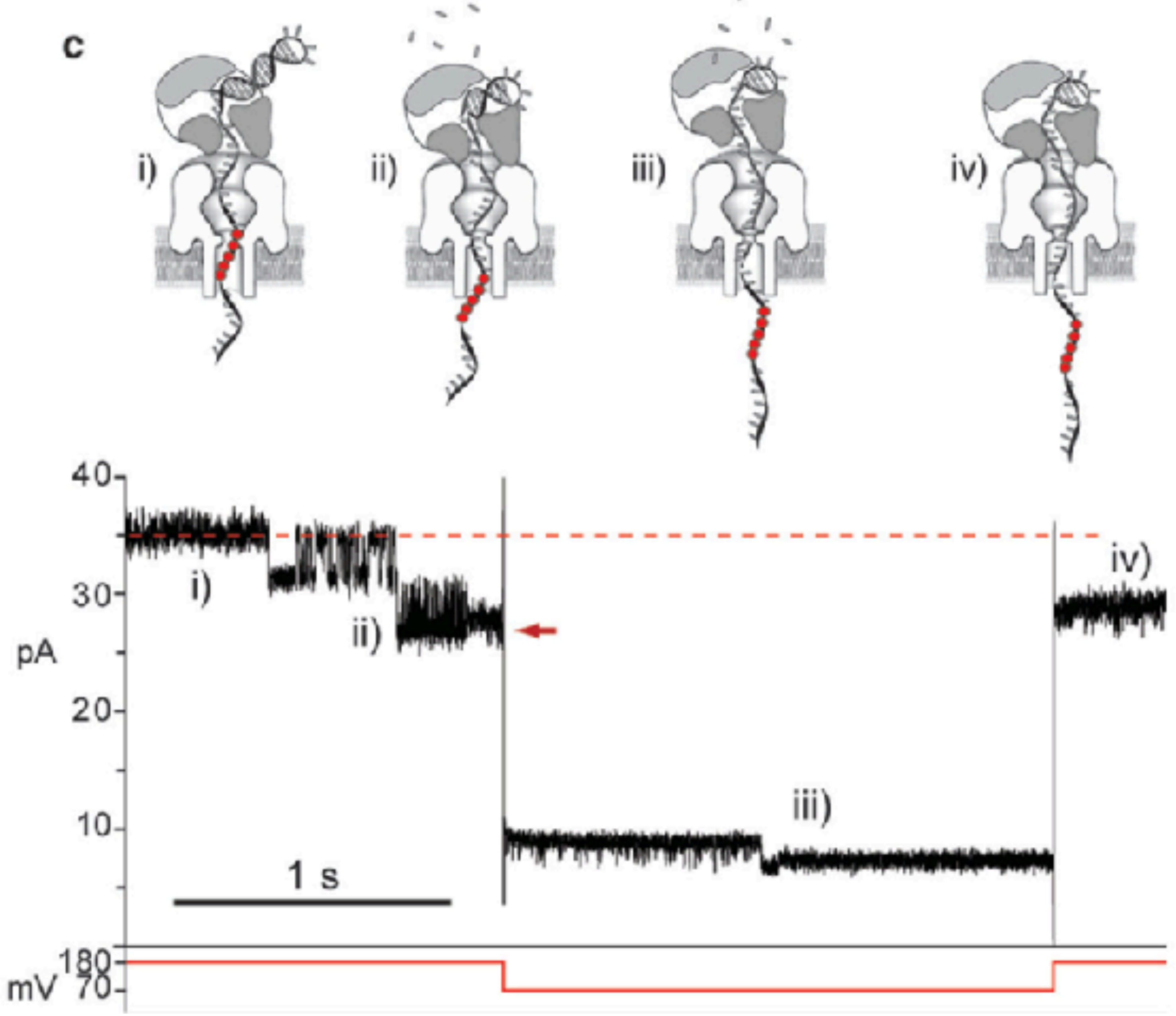
3'-TAGACAGT-5'-SS-αHL

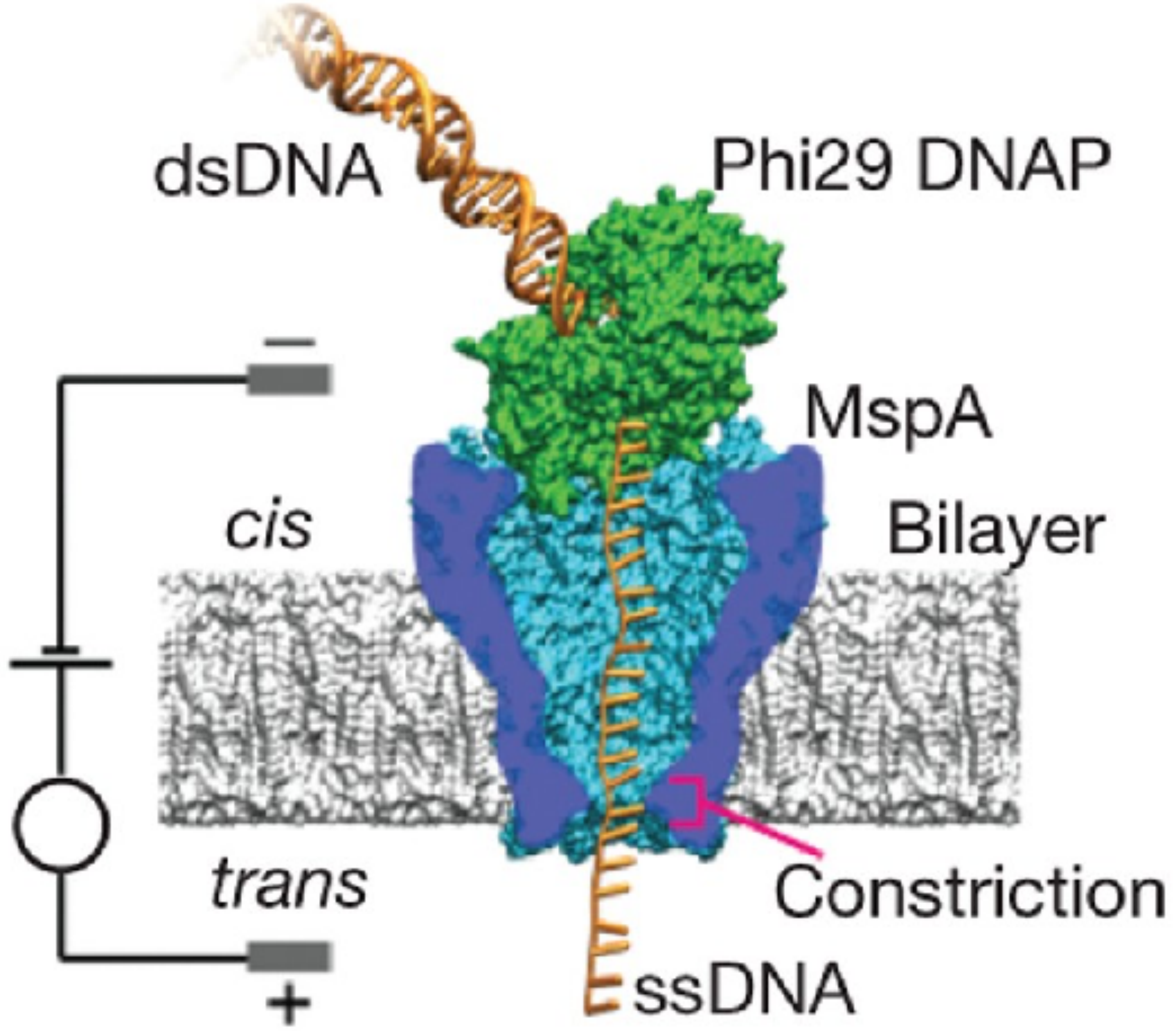
Howorka, S., Cheley, S. & Bayley, H. *Nat Biotechnol* **19**, 636–639 (2001). https://doi.org/10.1038/90236

EPFL

Nanopore DNA sequencing

Controlling the movement of DNA with the DNA polymerase phi29



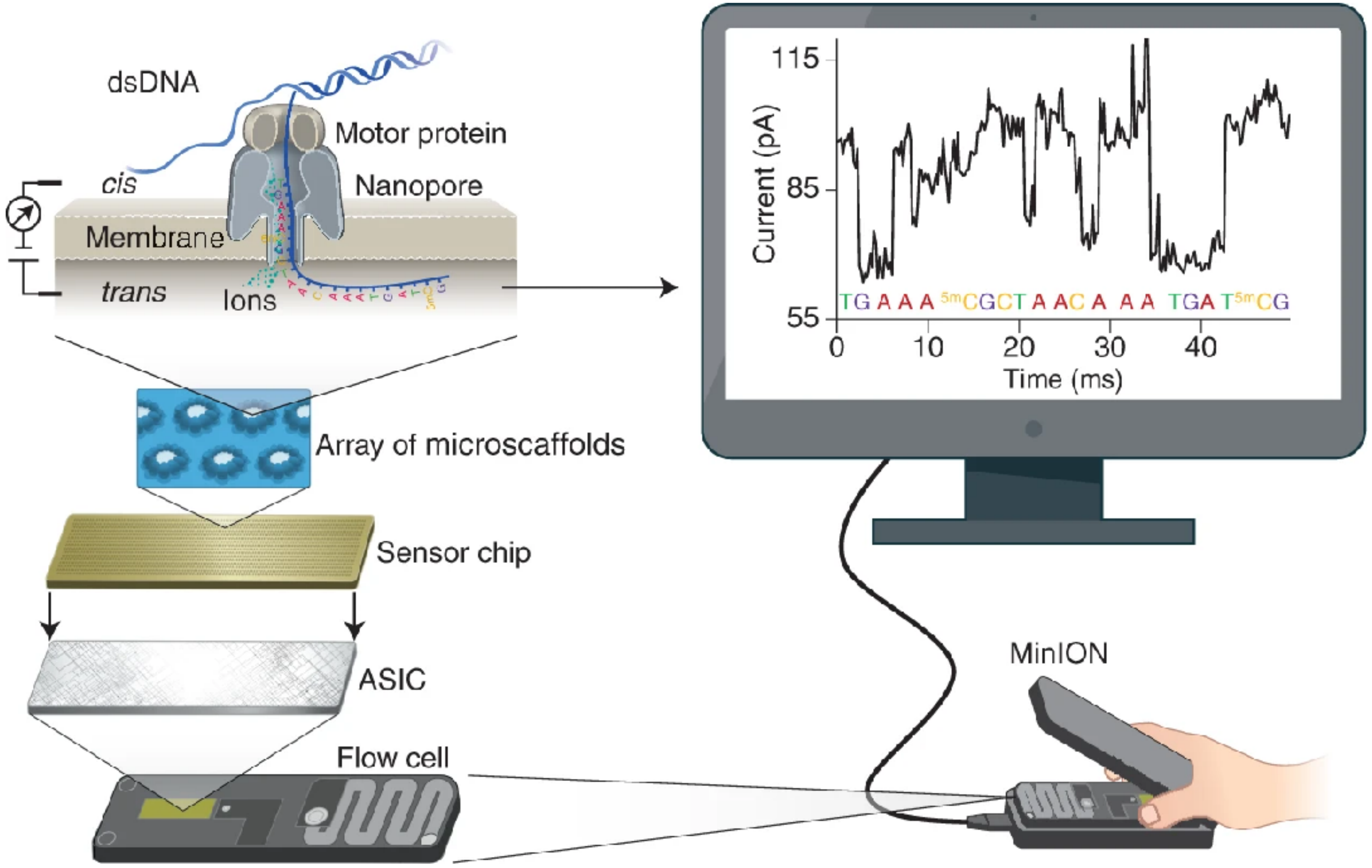


- Improved signal to noise due to slow DNA translocation with the help of phi29 through a nanopore
- Motor protein unwinds dsDNA and ssDNA (negatively charged) passes through the nanopore (driven by voltage)

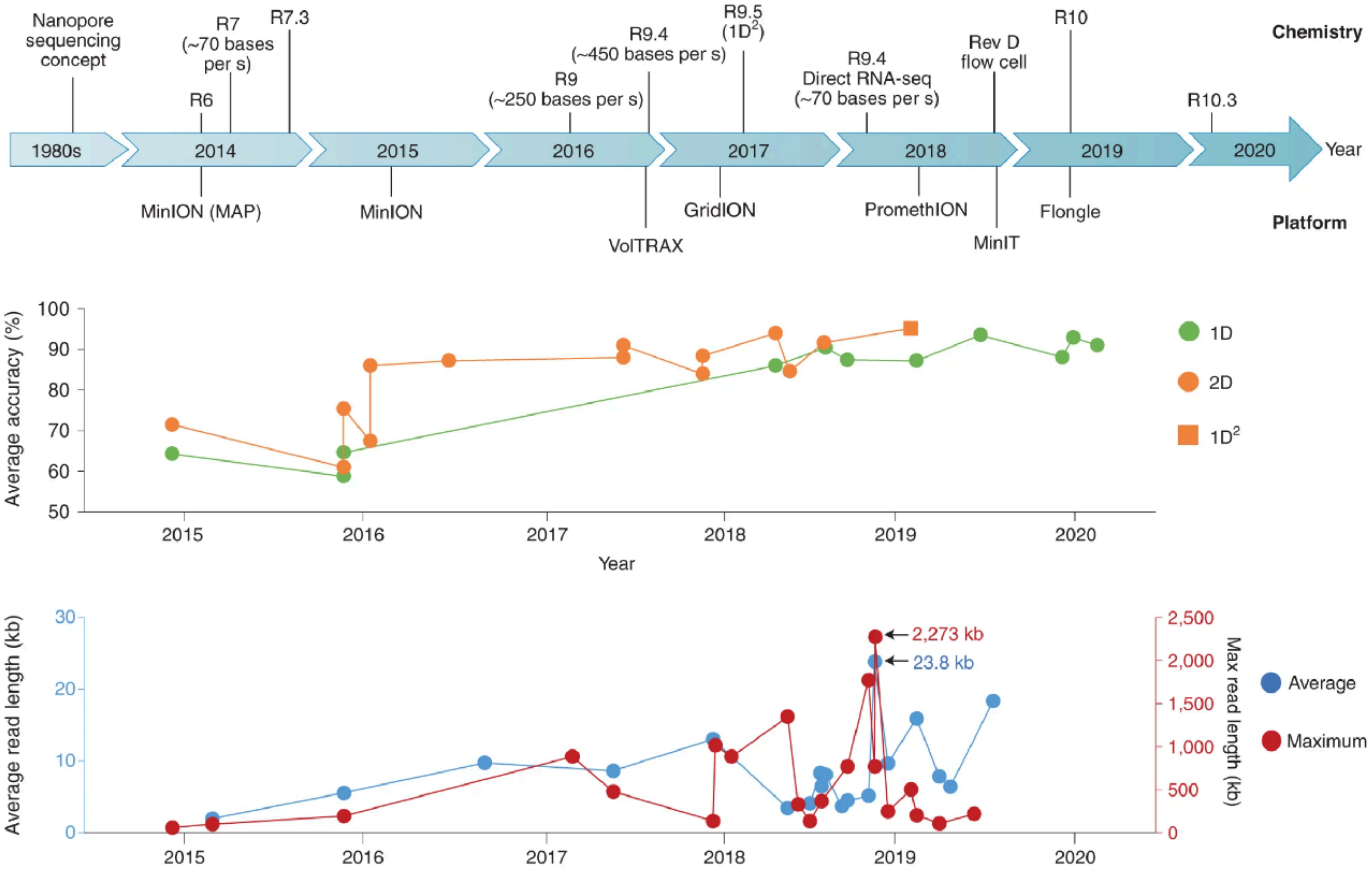
J. Am. Chem. Soc. 2010, 132, 50, 17961–17972

Nanopore flow cells

512 channels with 4 nanopores per channel (2,048 nanopores)



Major chemistry updates during the past 10 years



Oxford Nanopore Technologies

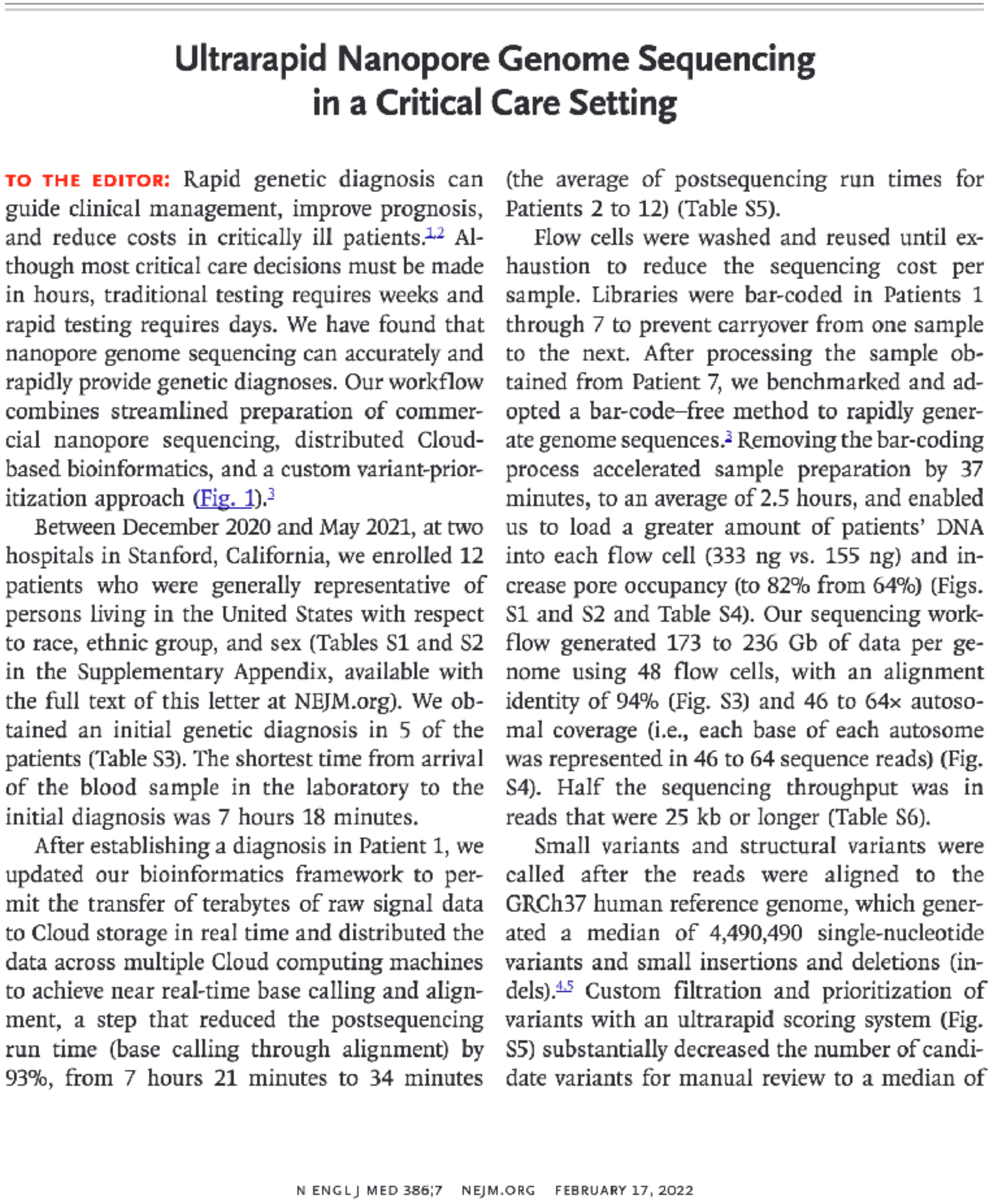
DNA sequencing devices

30x whole human genome in 3 days

20M reads with a typical length of 15-20 kb

BIO-312 Genomic solutions to sustainable development

7 hours from blood draw, to genome sequence, and initial diagnosis



Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting

TO THE EDITOR: Rapid genetic diagnosis can guide clinical management, improve prognosis, and reduce costs in critically ill patients.[1,2] Although most critical care decisions must be made in hours, traditional testing requires weeks and rapid testing requires days. We have found that nanopore genome sequencing can accurately and rapidly provide genetic diagnoses. Our workflow combines streamlined preparation of commercial nanopore sequencing, distributed Cloud-based bioinformatics, and a custom variant-prioritization approach (Fig. 1).[3]

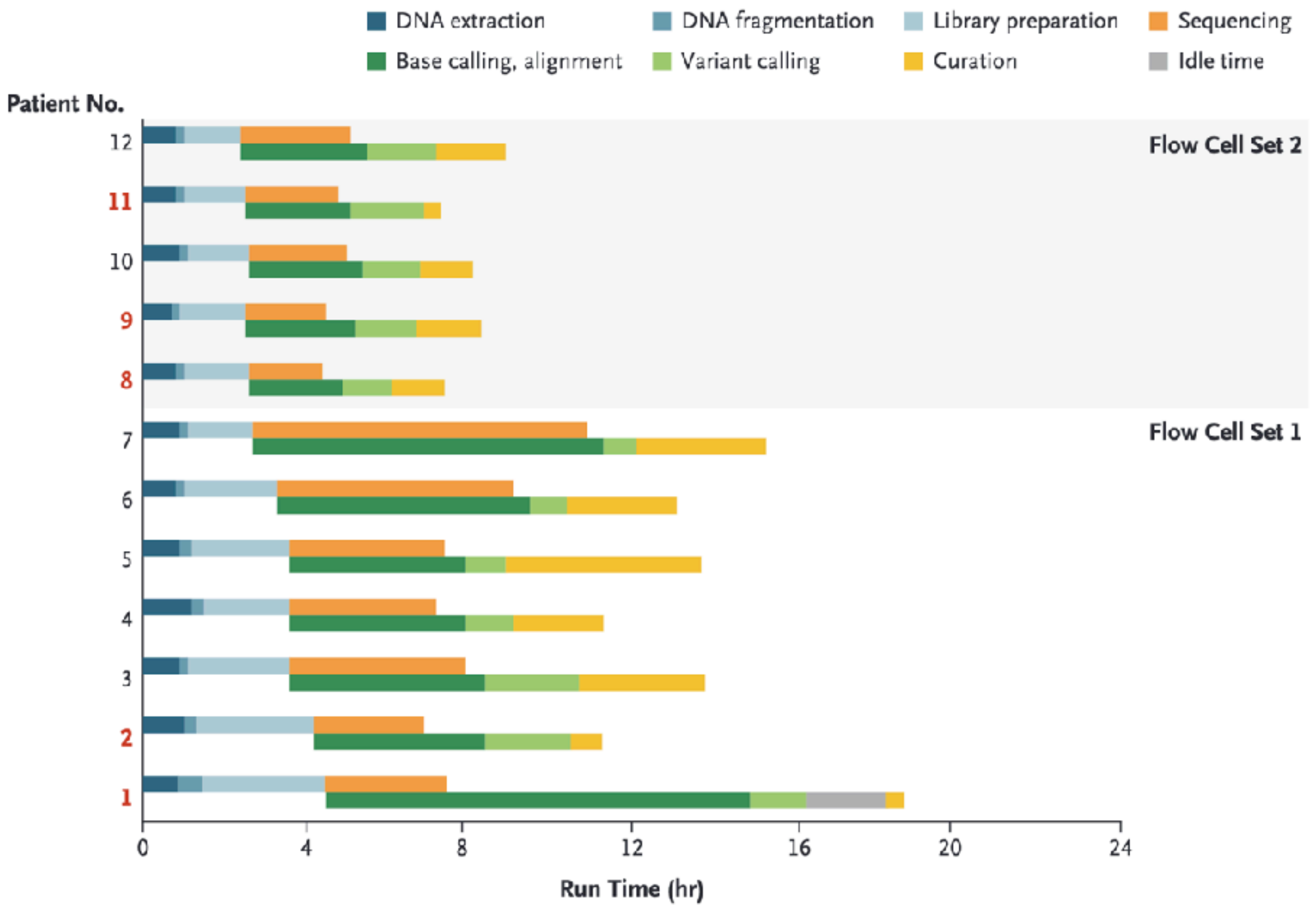
Between December 2020 and May 2021, at two hospitals in Stanford, California, we enrolled 12 patients who were generally representative of persons living in the United States with respect to race, ethnic group, and sex (Tables S1 and S2 in the Supplementary Appendix, available with the full text of this letter at NEJM.org). We obtained an initial genetic diagnosis in 5 of the patients (Table S3). The shortest time from arrival of the blood sample in the laboratory to the initial diagnosis was 7 hours 18 minutes.

After establishing a diagnosis in Patient 1, we updated our bioinformatics framework to permit the transfer of terabytes of raw signal data to Cloud storage in real time and distributed the data across multiple Cloud computing machines to achieve near real-time base calling and alignment, a step that reduced the postsequencing run time (base calling through alignment) by 93%, from 7 hours 21 minutes to 34 minutes (the average of postsequencing run times for Patients 2 to 12) (Table S5).

Flow cells were washed and reused until exhaustion to reduce the sequencing cost per sample. Libraries were bar-coded in Patients 1 through 7 to prevent carryover from one sample to the next. After processing the sample obtained from Patient 7, we benchmarked and adopted a bar-code–free method to rapidly generate genome sequences.[3] Removing the bar-coding process accelerated sample preparation by 37 minutes, to an average of 2.5 hours, and enabled us to load a greater amount of patients' DNA into each flow cell (333 ng vs. 155 ng) and increase pore occupancy (to 82% from 64%) (Figs. S1 and S2 and Table S4). Our sequencing workflow generated 173 to 236 Gb of data per genome using 48 flow cells, with an alignment identity of 94% (Fig. S3) and 46 to 64× autosomal coverage (i.e., each base of each autosome was represented in 46 to 64 sequence reads) (Fig. S4). Half the sequencing throughput was in reads that were 25 kb or longer (Table S6).

Small variants and structural variants were called after the reads were aligned to the GRCh37 human reference genome, which generated a median of 4,490,490 single-nucleotide variants and small insertions and deletions (indels).[4,5] Custom filtration and prioritization of variants with an ultrarapid scoring system (Fig. S5) substantially decreased the number of candidate variants for manual review to a median of

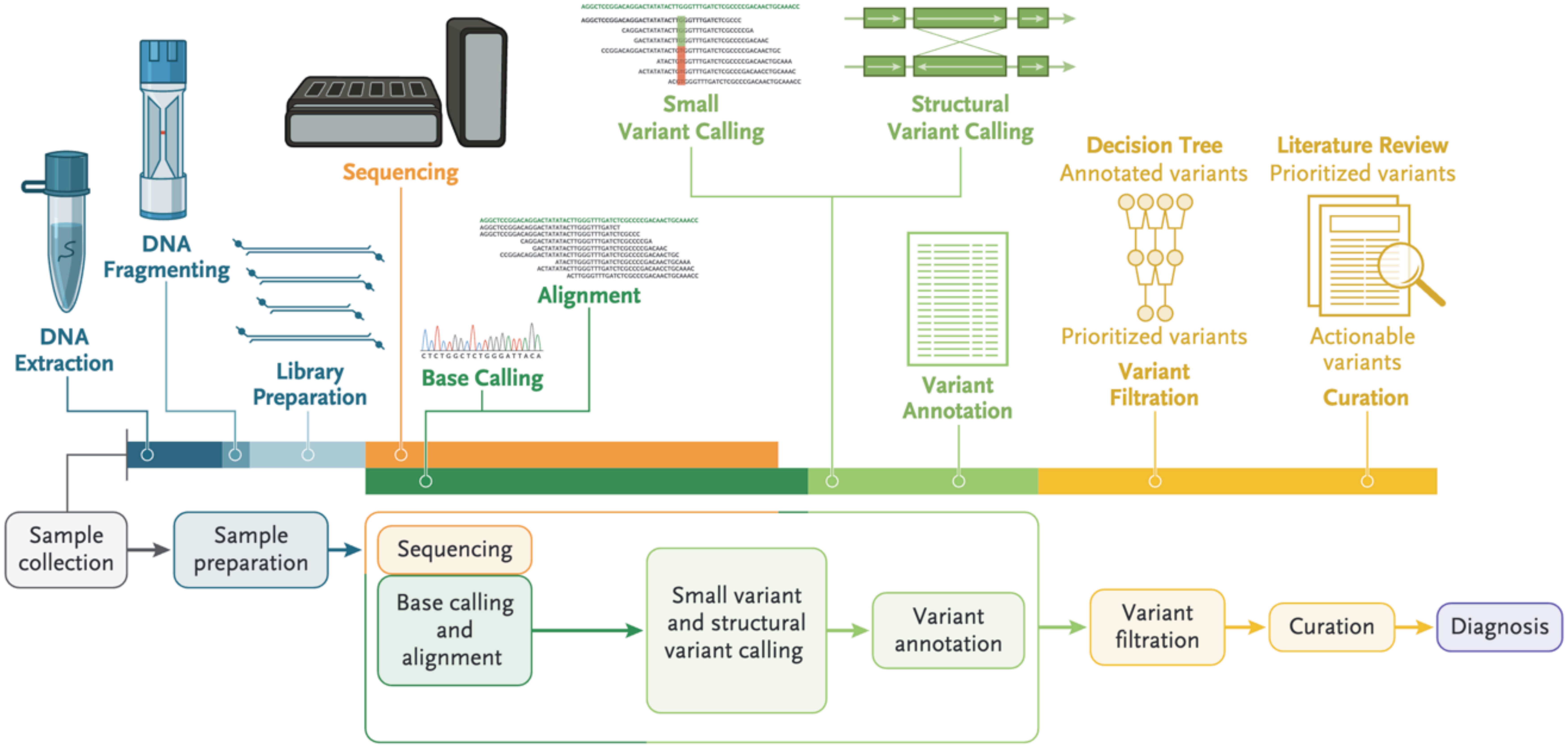
N ENGL J MED 386;7 NEJM.ORG FEBRUARY 17, 2022



École polytec fédéral de Lau

EPFL

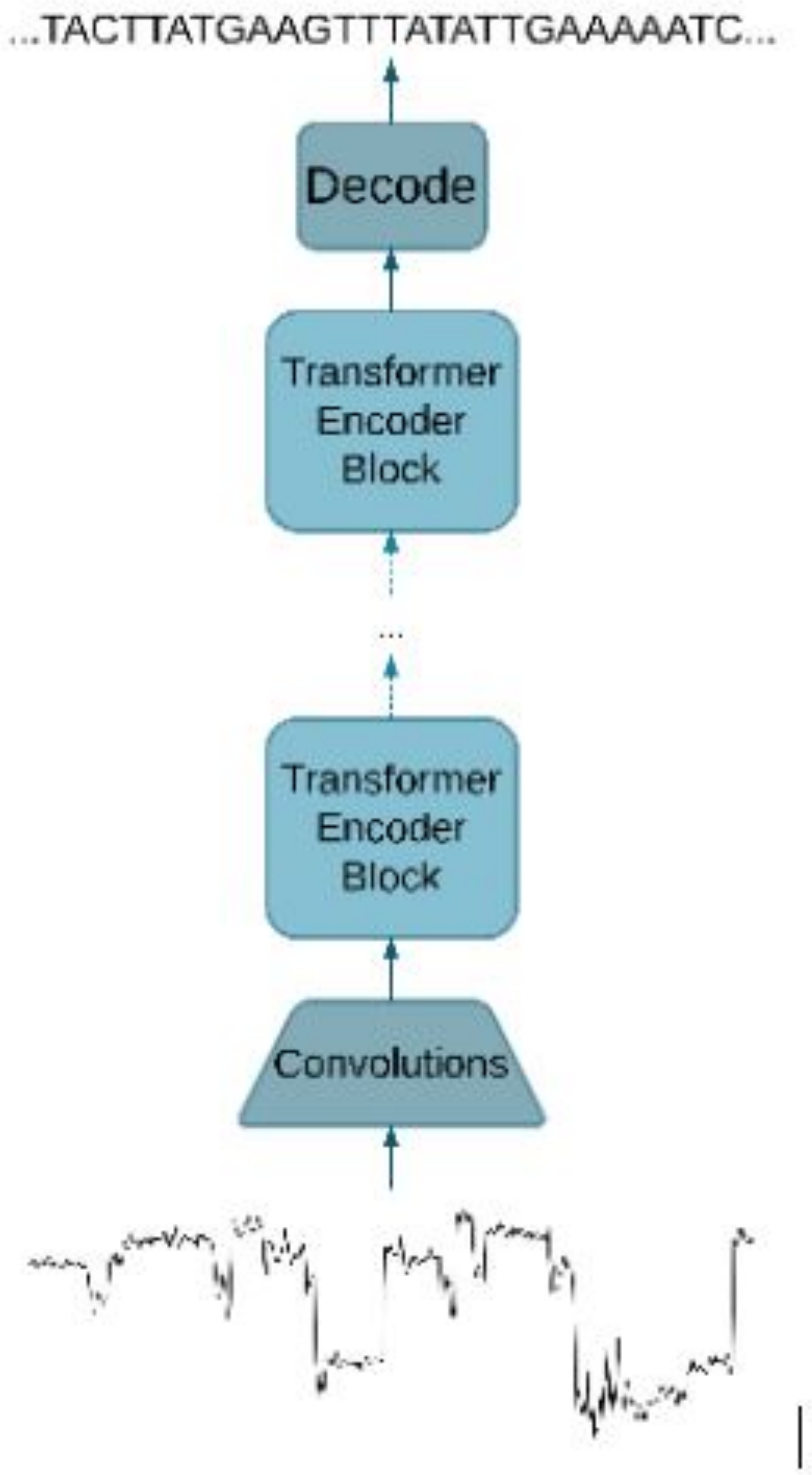
Nanopore genome sequencing workflow



Basecalling

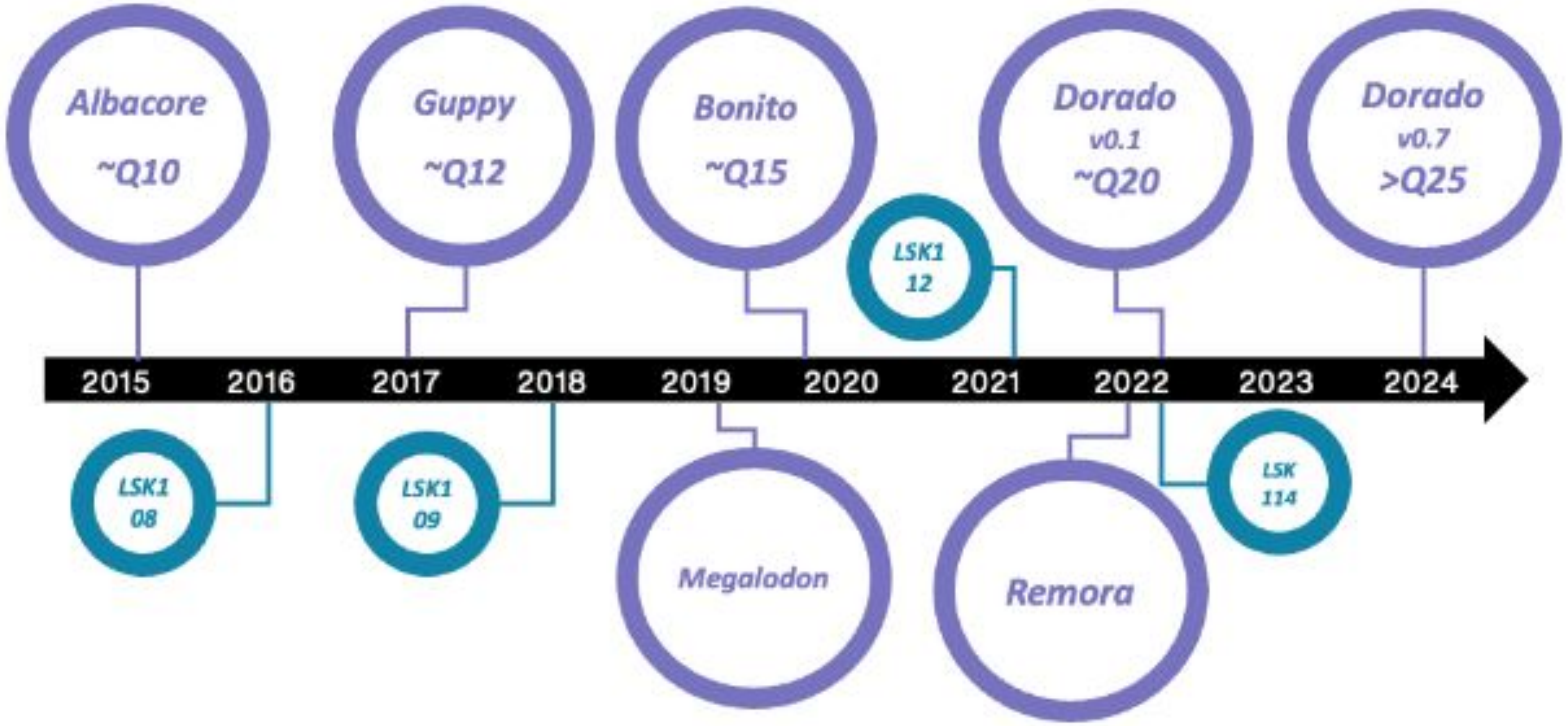
From electrical signals to nucleotide sequences

- Basecalling = converting raw electrical signals to nucleotide sequences
- Input: POD5 file
- Output: FASTQ file


...TACTTATGAAGTTTATATTGAAAAATC...
Decode
Transformer Encoder Block
...
Transformer Encoder Block
Convolutions

Basecalling algorithms

From hidden Markov models to transformer-based architectures



https://nanoporetech.com/blog/transforming-basecalling-in-genomic-sequencing

https://github.com/nanoporetech/dorado

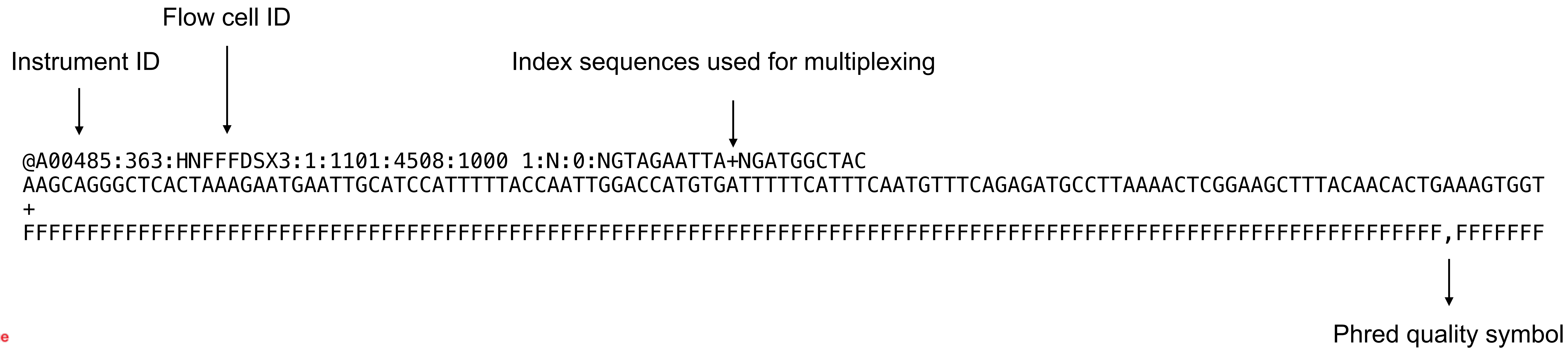
DNA sequencing data types

- FASTA: format to store nucleotide or amino acid sequences
- FASTQ: format to store nucleotide sequence, meta data, and quality scores
- POD5: format to store raw Nanopore and meta data
- SAM/BAM: format to store sequence alignments

FASTQ

Format to store raw sequencing data

- Text file and not compressed
- Blocks of 4 lines correspond (=1 read)
- Header, DNA sequence, header (+ symbol), per-base quality



Flow cell ID

Instrument ID

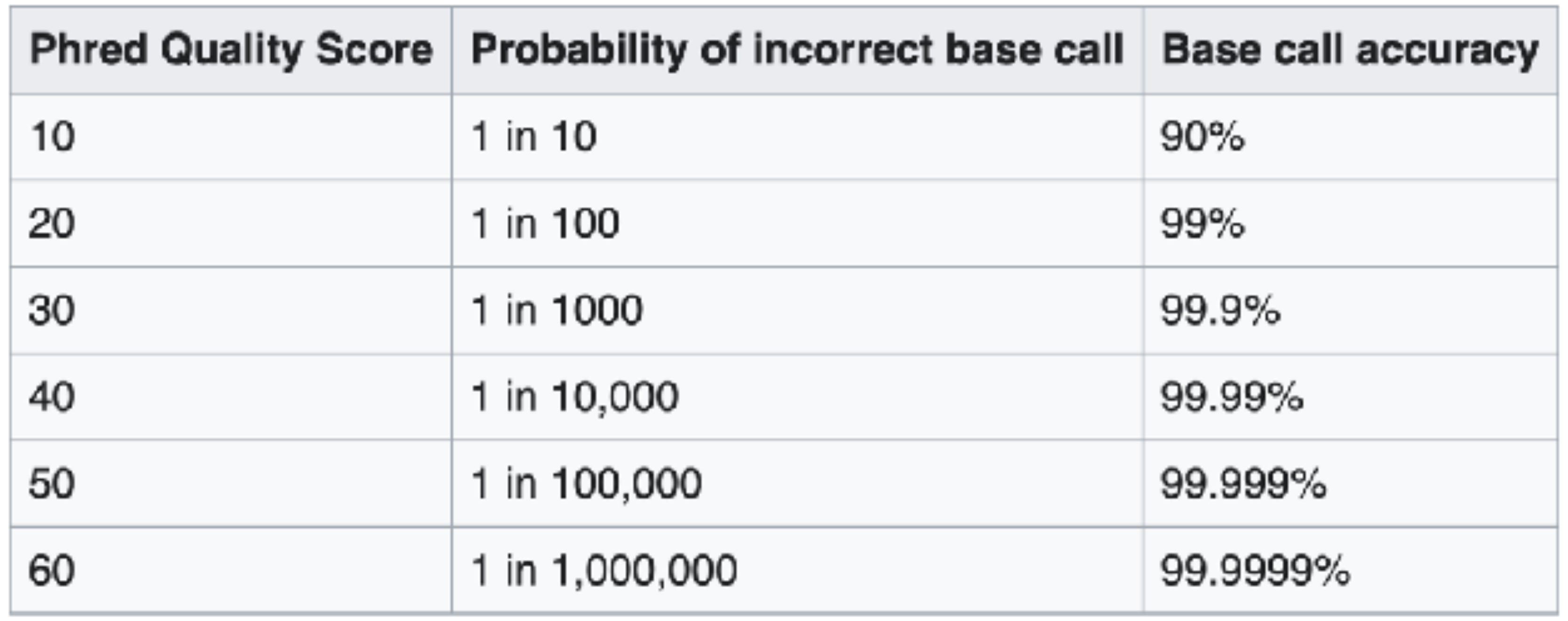
Index sequences used for multiplexing

```
@A00485:363:HNFFFDSX3:1:1101:4508:1000 1:N:0:NGTAGAATTA+NGATGGCTAC
AAGCAGGGCTCACTAAAGAATGAATTGCATCCATTTTTACCAATTGGACCATGTGATTTTTCATTTCAATGTTTCAGAGATGCCTTAAAACTCGGAAGCTTTACAACACTGAAAGTGGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFF
```

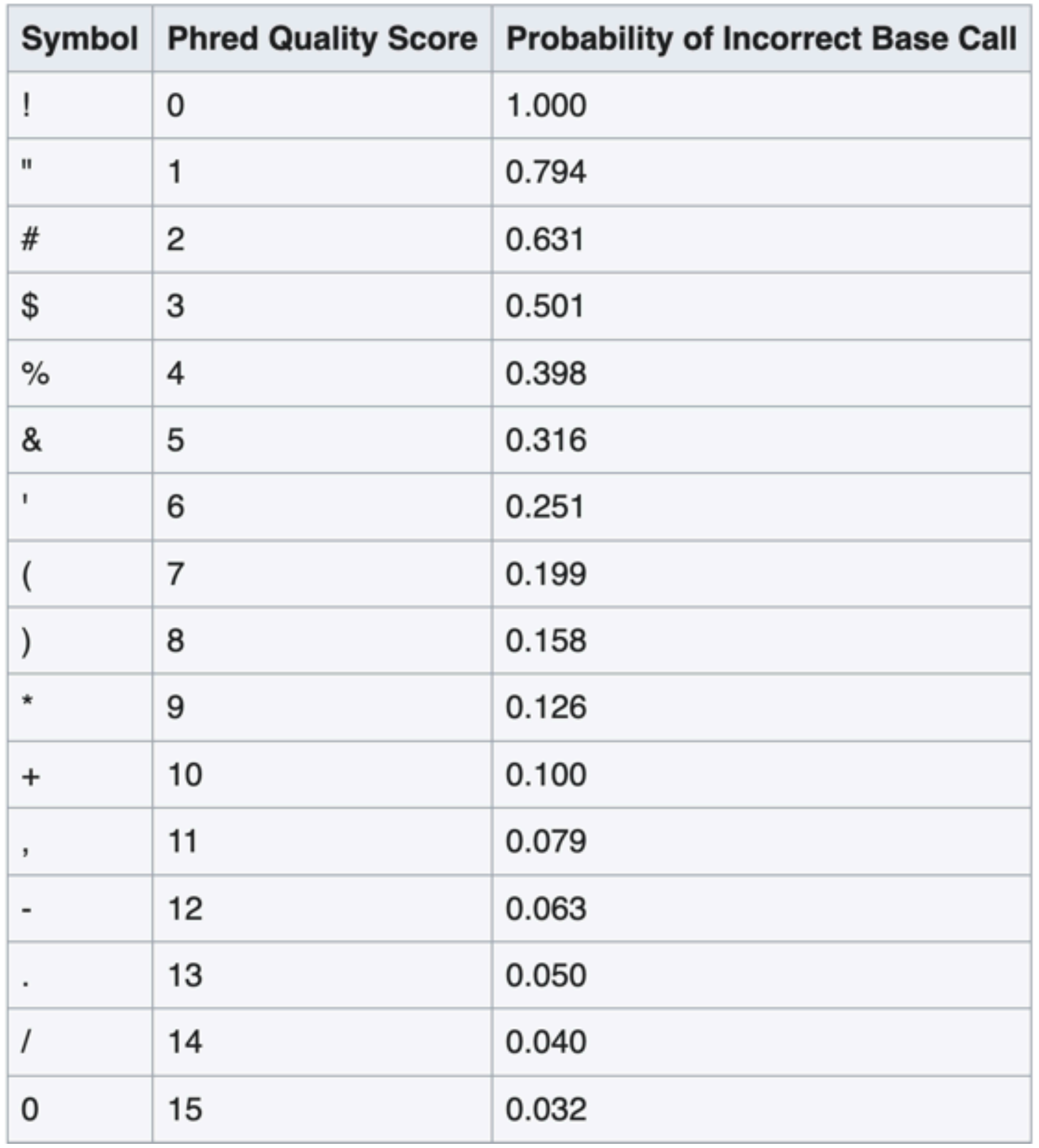
Phred quality symbol

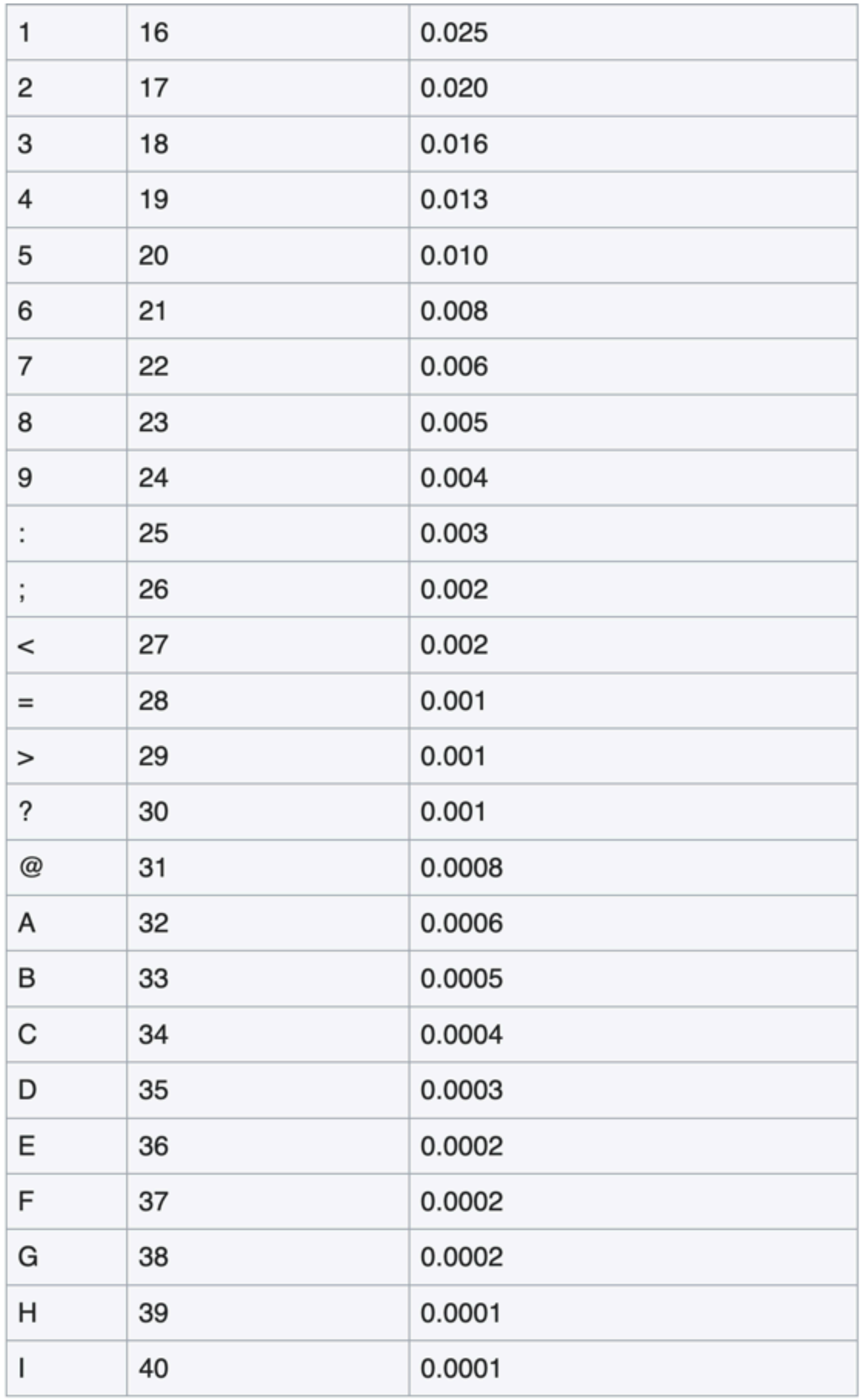
Phred base quality scores

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



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%





Symbol	Phred Quality Score	Probability of Incorrect Base Call
!	0	1.000
"	1	0.794
#	2	0.631
$	3	0.501
%	4	0.398
&	5	0.316
'	6	0.251
(	7	0.199
)	8	0.158
*	9	0.126
+	10	0.100
,	11	0.079
-	12	0.063
.	13	0.050
/	14	0.040
0	15	0.032
1	16	0.025
2	17	0.020
3	18	0.016
4	19	0.013
5	20	0.010
6	21	0.008
7	22	0.006
8	23	0.005
9	24	0.004
:	25	0.003
;	26	0.002
<	27	0.002
=	28	0.001
>	29	0.001
?	30	0.001
@	31	0.0008
A	32	0.0006
B	33	0.0005
C	34	0.0004
D	35	0.0003
E	36	0.0002
F	37	0.0002
G	38	0.0002
H	39	0.0001
I	40	0.0001

Data type conversions

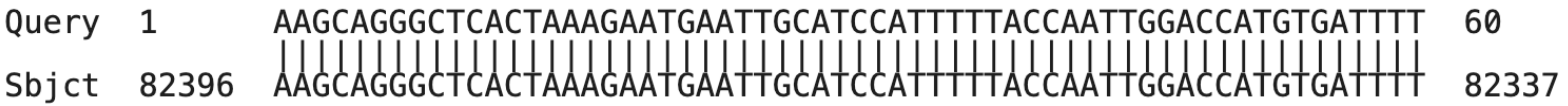
From raw sequence reads to annotated sequence reads

- Raw data (FASTQ or POD5)
- Reference genome (FASTA)
- Alignment (SAM/BAM/CRAM)
- Genome annotation files (GTF, BED)
- Variant files (VCF, BCF)

DNA sequence alignment

From sequences to regions in reference genomes with high similarity

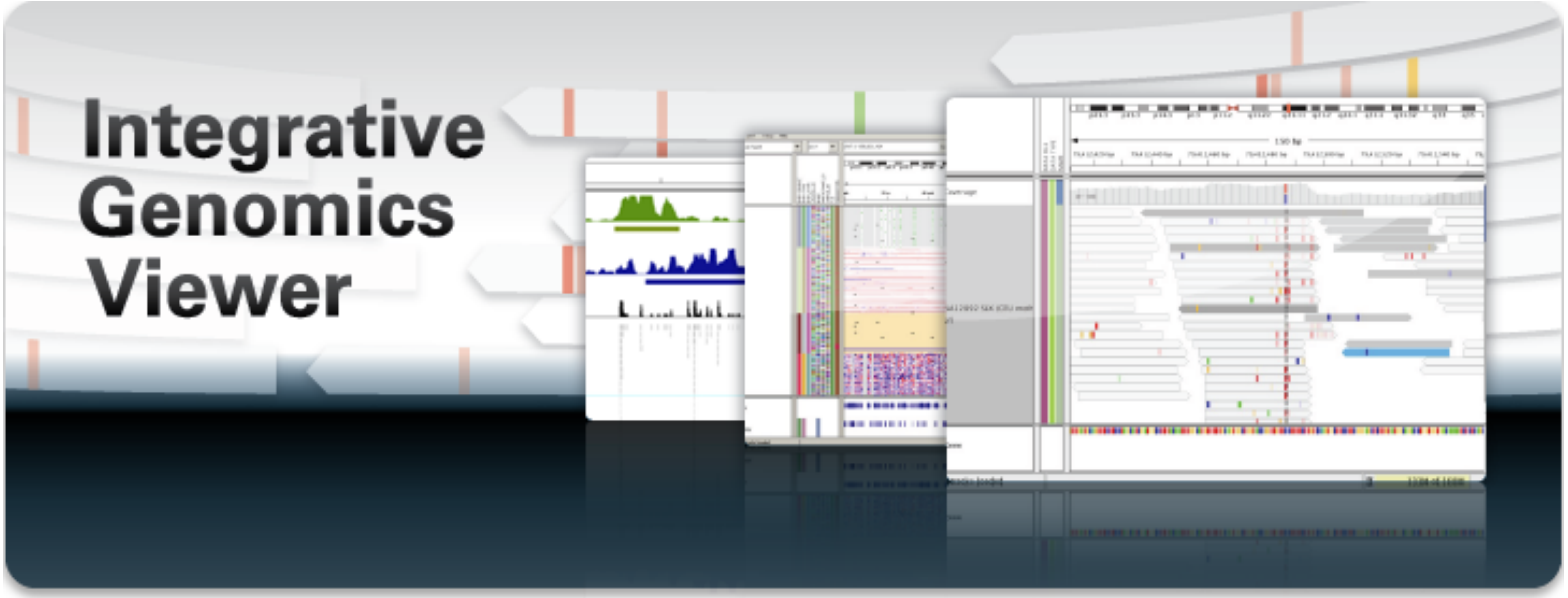
- Identify, or map, DNA sequences to a reference or multiple reference genomes
- Important to detect DNA sequence variation, to quantify species abundance, and many other applications
- Alignment methods/algorithms:
  - Many short reads (100-300 bp): bwa-mem2 or bowtie2 (fast and memory efficient)
  - Many long reads (up to Mb): minimap2 (handles high error rates, very long sequences, and structural variation)
  - Few reads: blastn (local alignment with similarity search)



```
Query  1      AAGCAGGGCTCACTAAAGAATGAATTGCATCCATTTTTACCAATTGGACCATGTGATTTT  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  82396  AAGCAGGGCTCACTAAAGAATGAATTGCATCCATTTTTACCAATTGGACCATGTGATTTT  82337
```

Integrative Genomics Viewer

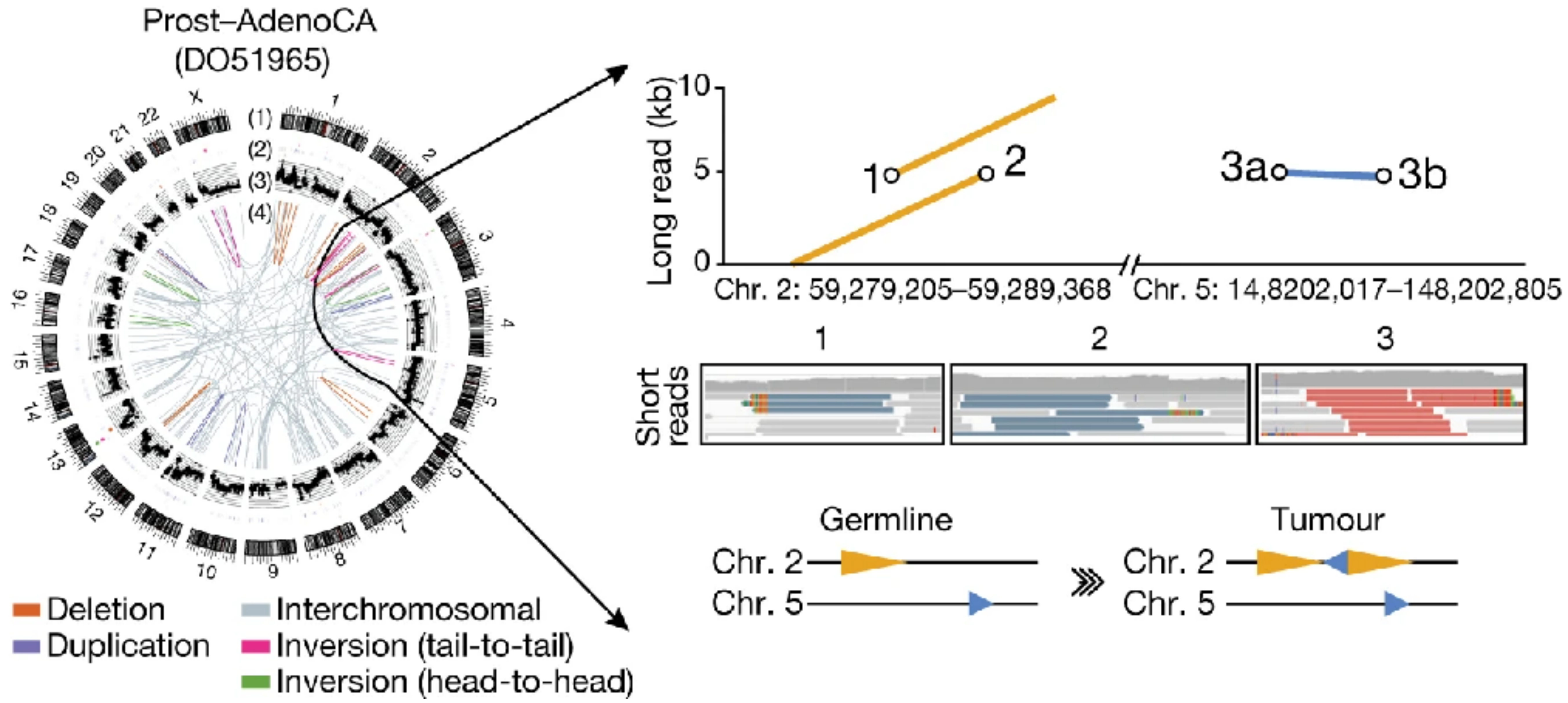
High-performance and interactive program for visual exploration of genomic data



https://igv.org/doc/desktop/#

Complete human genomes enabled by Nanopore DNA sequencing

- Single nucleotide variants (SNVs)
- Small insertions/deletions (5-50 bp)
- Structural variants (>50 to several kb)
- Copy number variants (CNVs)
- Long-range haplotypes (maternal/paternal)
- *De novo* genome assembly
- DNA methylation (5mC and 5hmC)



Waszak, Tiao, Zhu, Rausch *et al. bioRxiv* 2017 (PCAWG Consortium, *Nature* 2020)

Open source code: https://github.com/nanoporetech