Bioinformatic challenges

Whole molecule sequencing methods and their applications

PROF. DR. WOJCIECH MAKAŁOWSKI, INSTITUTE OF BIOINFORMATICS, WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

> The double helix is indeed a remarkable molecule. Modern man is perhaps 50,000 years old, civilization has existed for scarcely 10,000 years and the United States for only just over 200 years; but DNA and RNA have been around for at least several billion years. All that time the double helix has been there, and active, and yet we are the first creatures on Earth to become aware of its existence.

Francis Crick (1916–2004)

DNA story

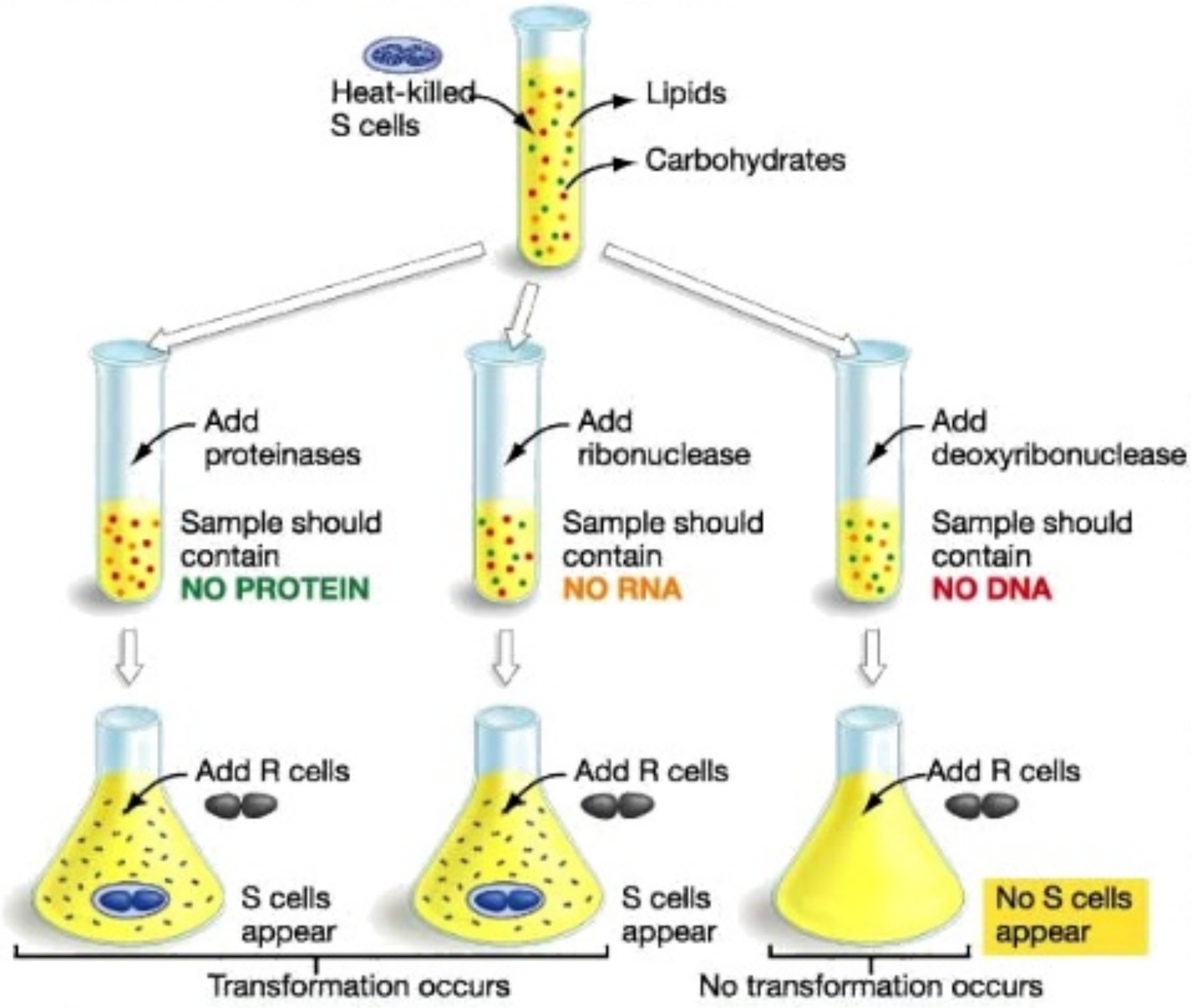
1870

Friedrich Miescher discovers DNA

1944

Oswald Avery proves that DNA is a genetic material

DETERMINING THAT DNA IS THE HEREDITARY MATERIAL



1. Remove the lipids and carbohydrates from a solution of heat-killed S cells. Proteins, RNA, and DNA remain.

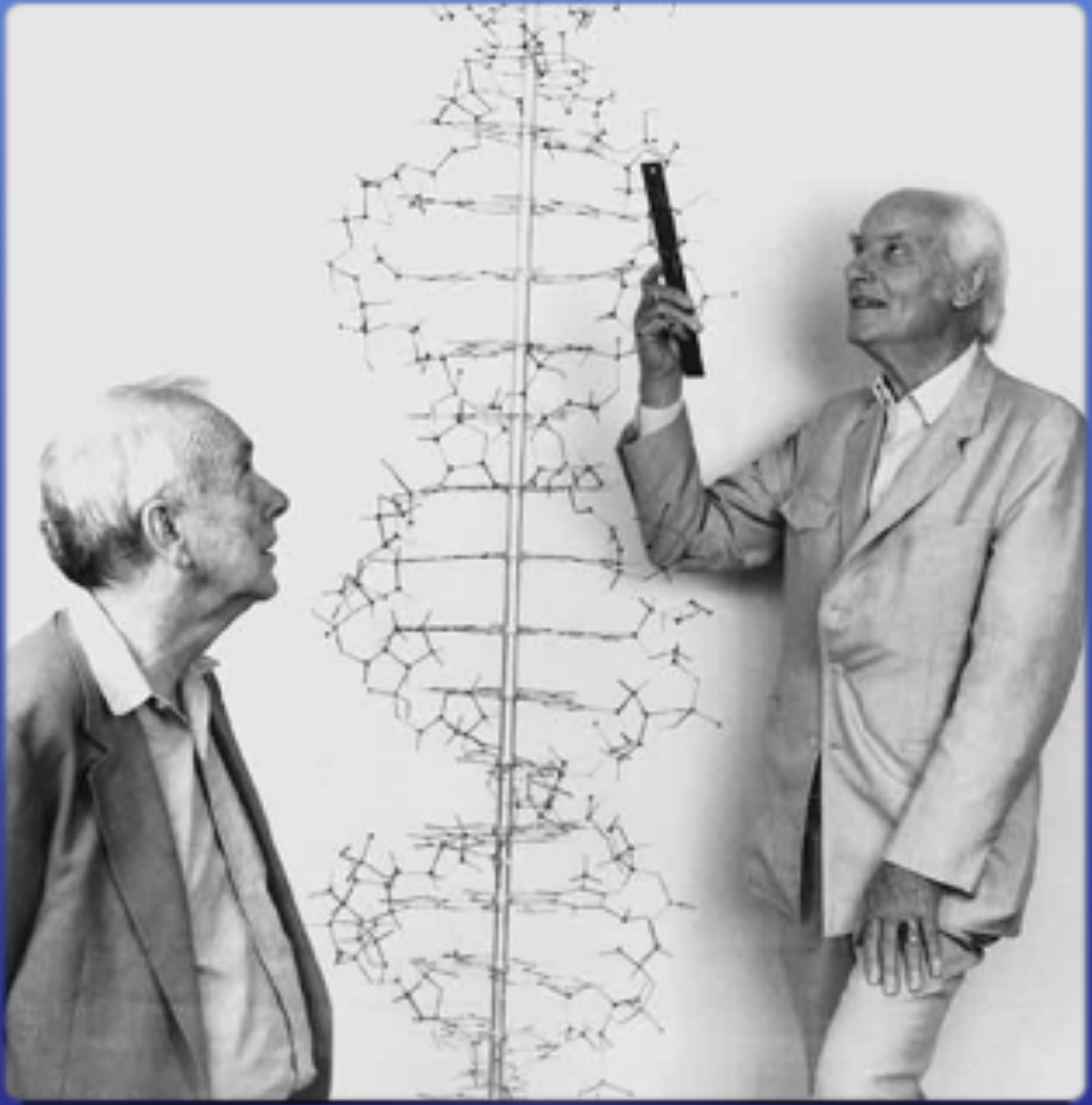
2. Subject the solution to treatments of enzymes to destroy either the proteins, RNA, or DNA.

3. Add a small portion of each sample to a culture containing R cells. Observe whether transformation has occurred by testing for the presence of virulent S cells.

DNA story

1953

James Watson and Francis Crick discover DNA structure



Sequencing: beginnings

1964

Robert W. Holley determines nucleotide sequences (77 nt) of the yeast Alanine tRNA

J. Biol. Chem. 240: 2122-2128

1968

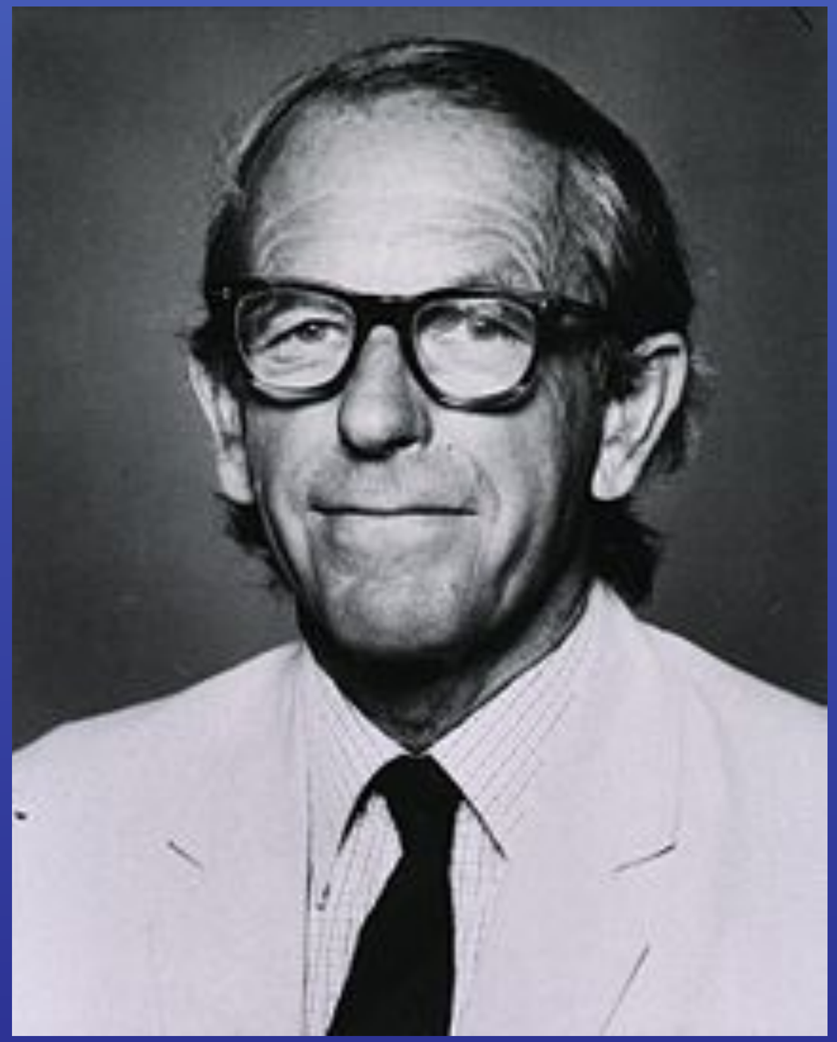
Ray Wu and A. Dale Kaiser sequenced 12 bases (!) of λ phage's 5' cohesive ends of its DNA, using radioactively labeled nucleotides and polyacrylamide gel electrophoresis

J. Mol. Biol. 35: 523-537

Sequencing: 1st generation sequencing

1977
Allan Maxam and Walter Gilbert develop DNA sequencing method by chemical degradation
J. Biol. Chem. 240: 2122-2128

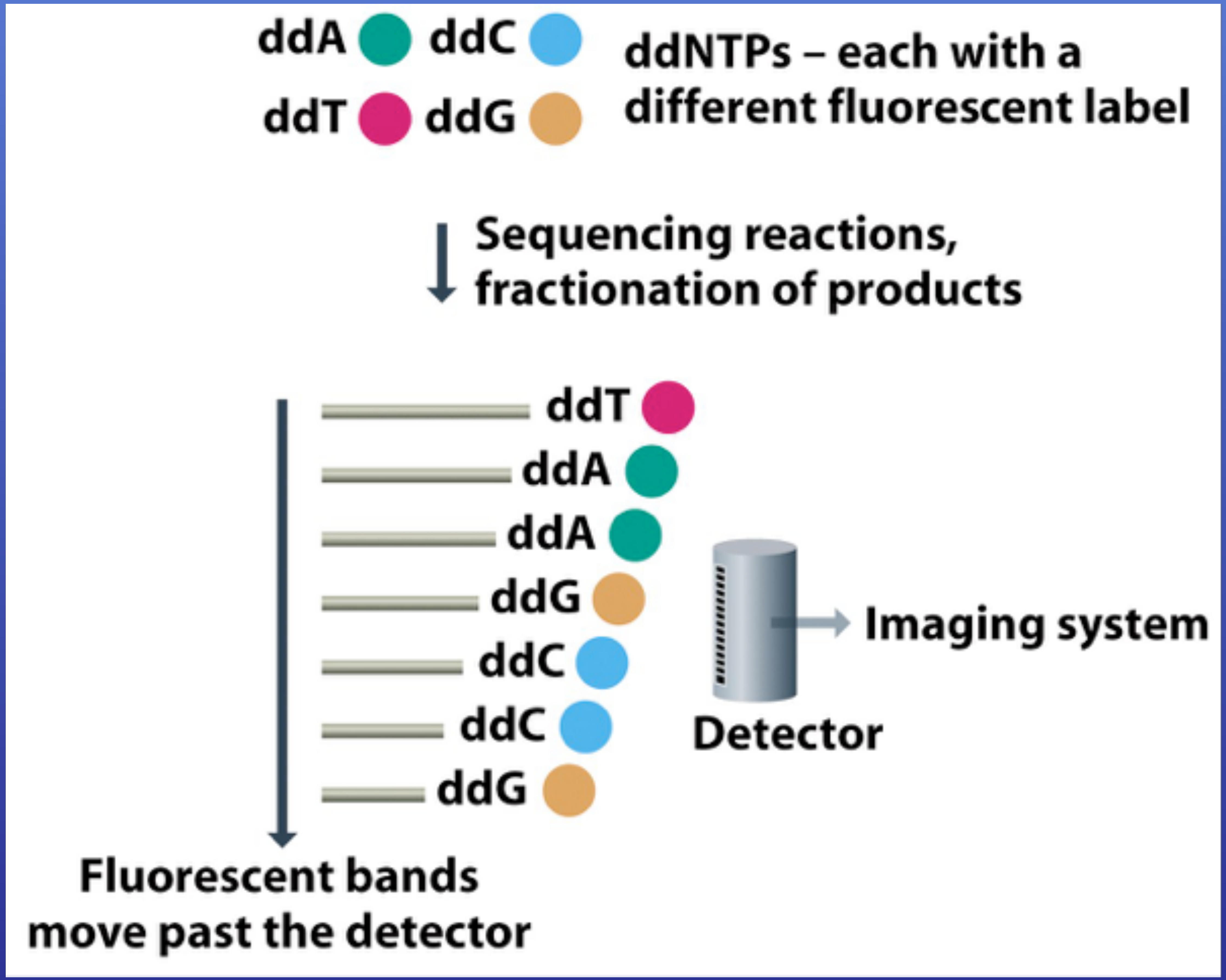
1977
Fred Sanger develops 2’,3’-dideoxy chain termination method
J. Mol. Biol. 35: 523-537



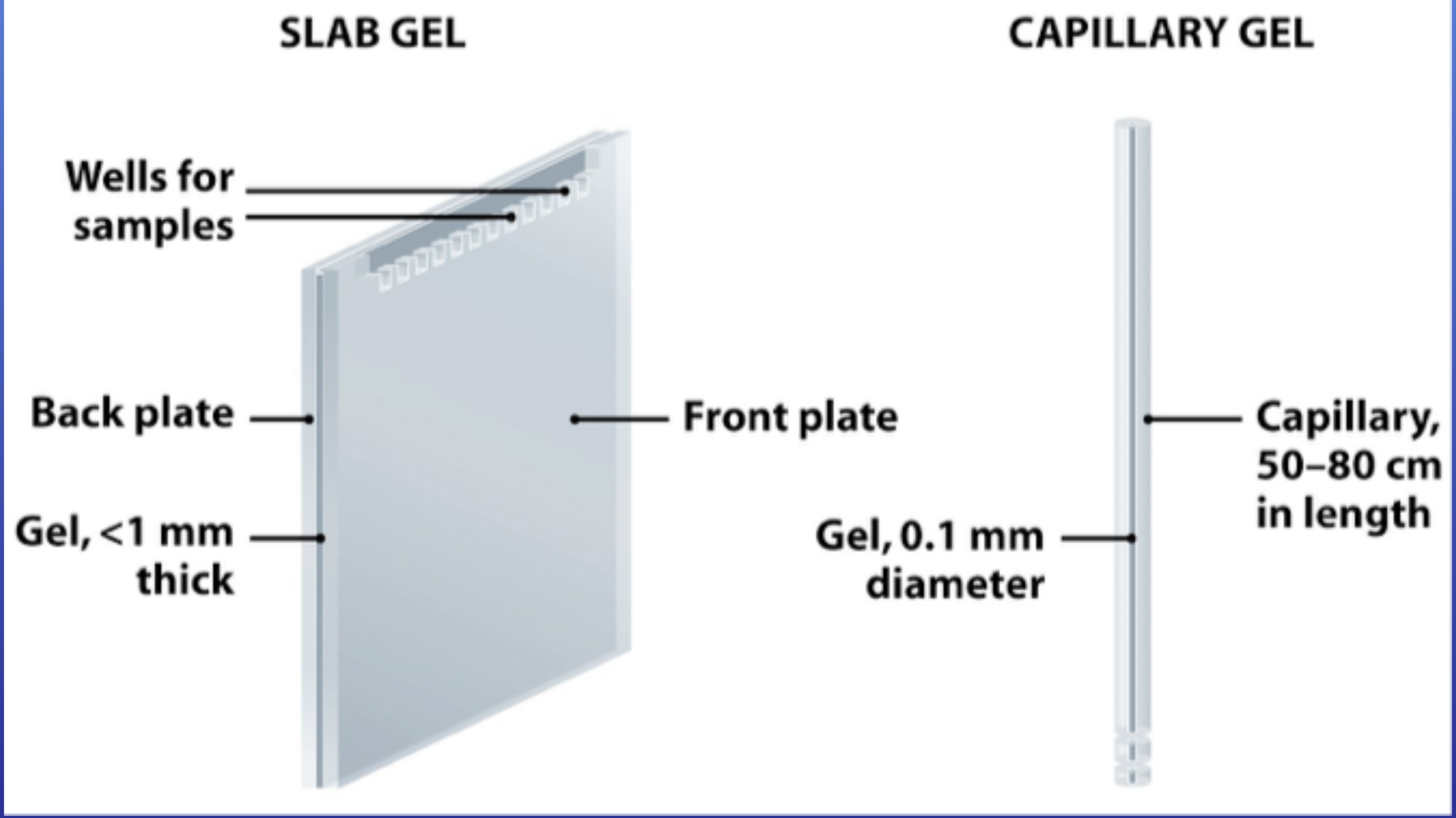
Sequencing: maturation

- 1983 - Marvin Caruthers developed a method to construct fragments of DNA of predetermined sequence from five to about 75 base pairs long. He and Leroy Hood invented instruments that could make such fragments automatically.
- 1983 - Kary Mullis invented the polymerase chain reaction (PCR) technique
- 1987 - ABI 370; first fully automated sequencing machine
- 1995 - Craig Venter uses whole-genome shotgun sequencing technique to determine complete genome of bacterium *Haemophilus influenzae*
- 2005 - introduction of GS20 sequencing machine (454 Life Sciences); first in the line of “Next Generation Sequencing”
- 2010 - PacBio introduced first single molecule, long reads instrument marking Third Generation Sequencing.

Sequencing: maturation


ddA ● ddC ●
ddT ● ddG ●
ddNTPs – each with a different fluorescent label
Sequencing reactions, fractionation of products
ddT ●
ddA ●
ddA ●
ddG ●
ddC ●
ddC ●
ddG ●
Detector
Imaging system
Fluorescent bands move past the detector

Sequencing: maturation



Next Generation Sequencing

- Massive parallelization of the sequencing process
- Relatively short reads
- Different approaches from improving Sanger’s technique to direct “observation” of DNA through a microscope

Sequencing: 3rd generation sequencing

2010

PacBio - SMRT technology

Eid at al. (2009) Science 323: 133-138

2014

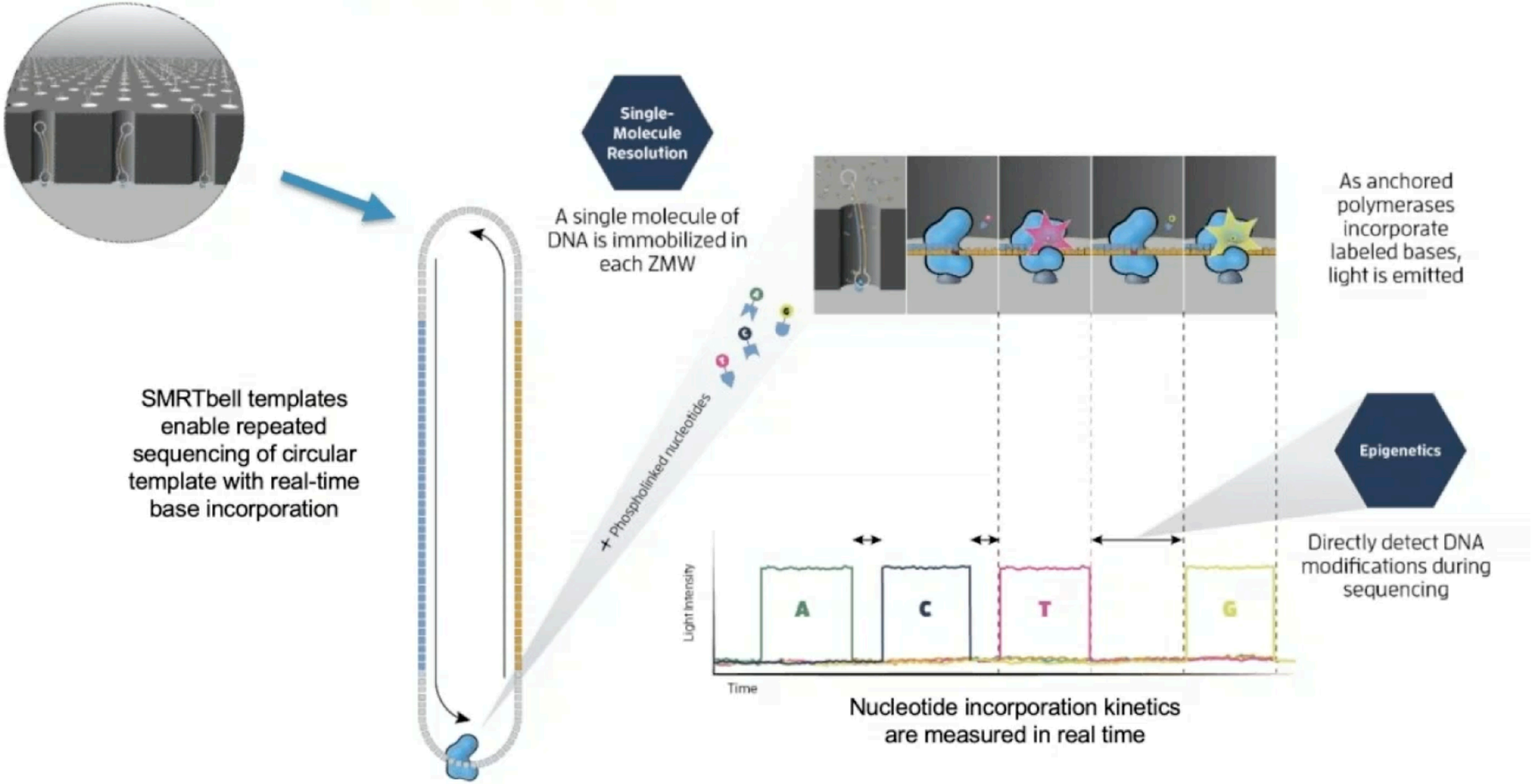
Oxford Nanopore Technologies

MinION

Kasianowicz et al. (1996) PNAS 93: 3770-13773

Single Molecule, Real-Time (SMRT) Sequencing https://www.youtube.com/watch?v=_ID8JyAbwEo

PacBio - key technology



SMRTbell templates enable repeated sequencing of circular template with real-time base incorporation

Single-Molecule Resolution

A single molecule of DNA is immobilized in each ZMW

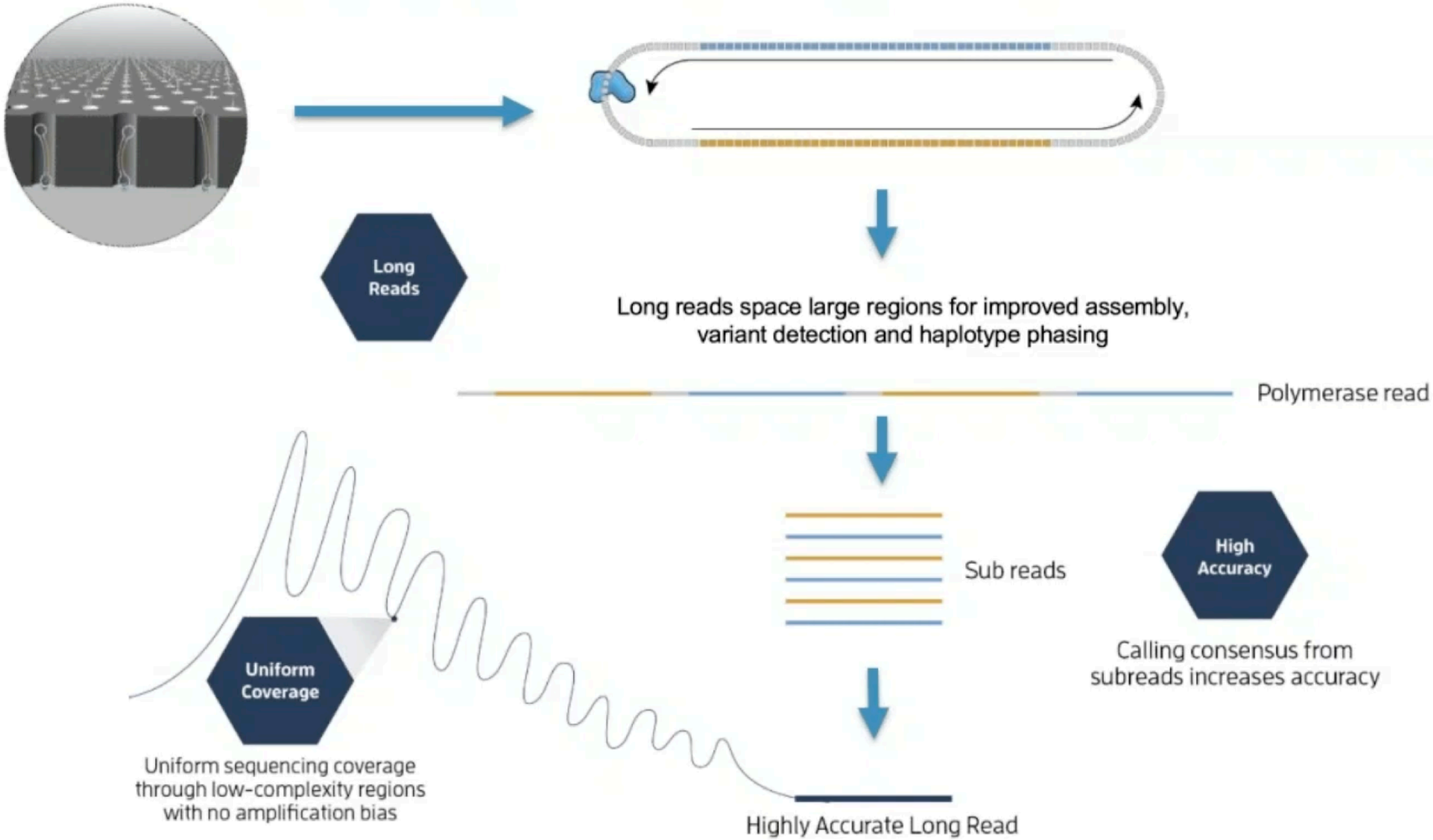
As anchored polymerases incorporate labeled bases, light is emitted

Epigenetics

Directly detect DNA modifications during sequencing

Nucleotide incorporation kinetics are measured in real time

PacBio - key technology



Long Reads

Long reads space large regions for improved assembly, variant detection and haplotype phasing

Polymerase read

Sub reads

High Accuracy

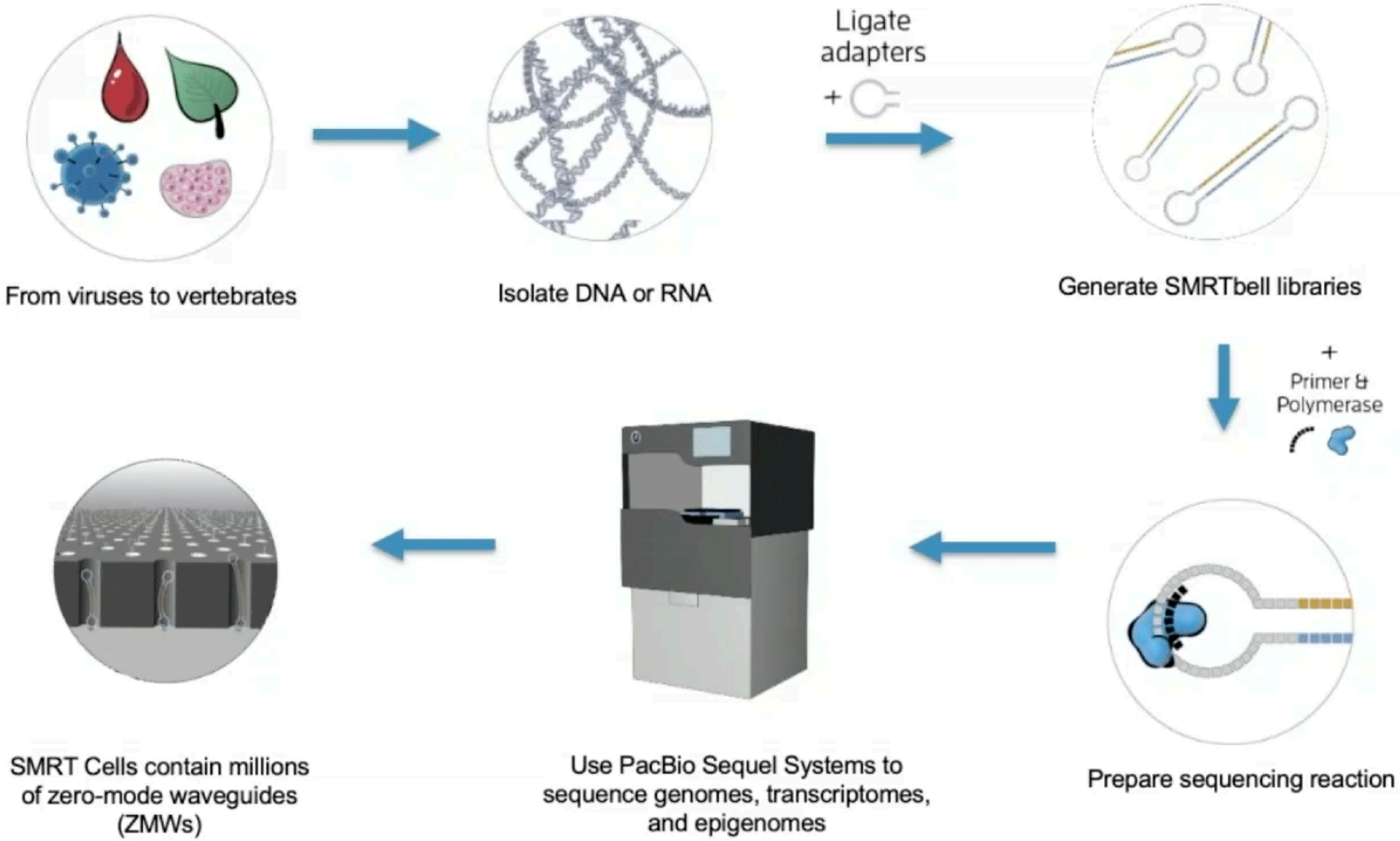
Calling consensus from subreads increases accuracy

Uniform Coverage

Uniform sequencing coverage through low-complexity regions with no amplification bias

Highly Accurate Long Read

PacBio - from sample to sequence



From viruses to vertebrates

Isolate DNA or RNA

Ligate adapters

+

Generate SMRTbell libraries

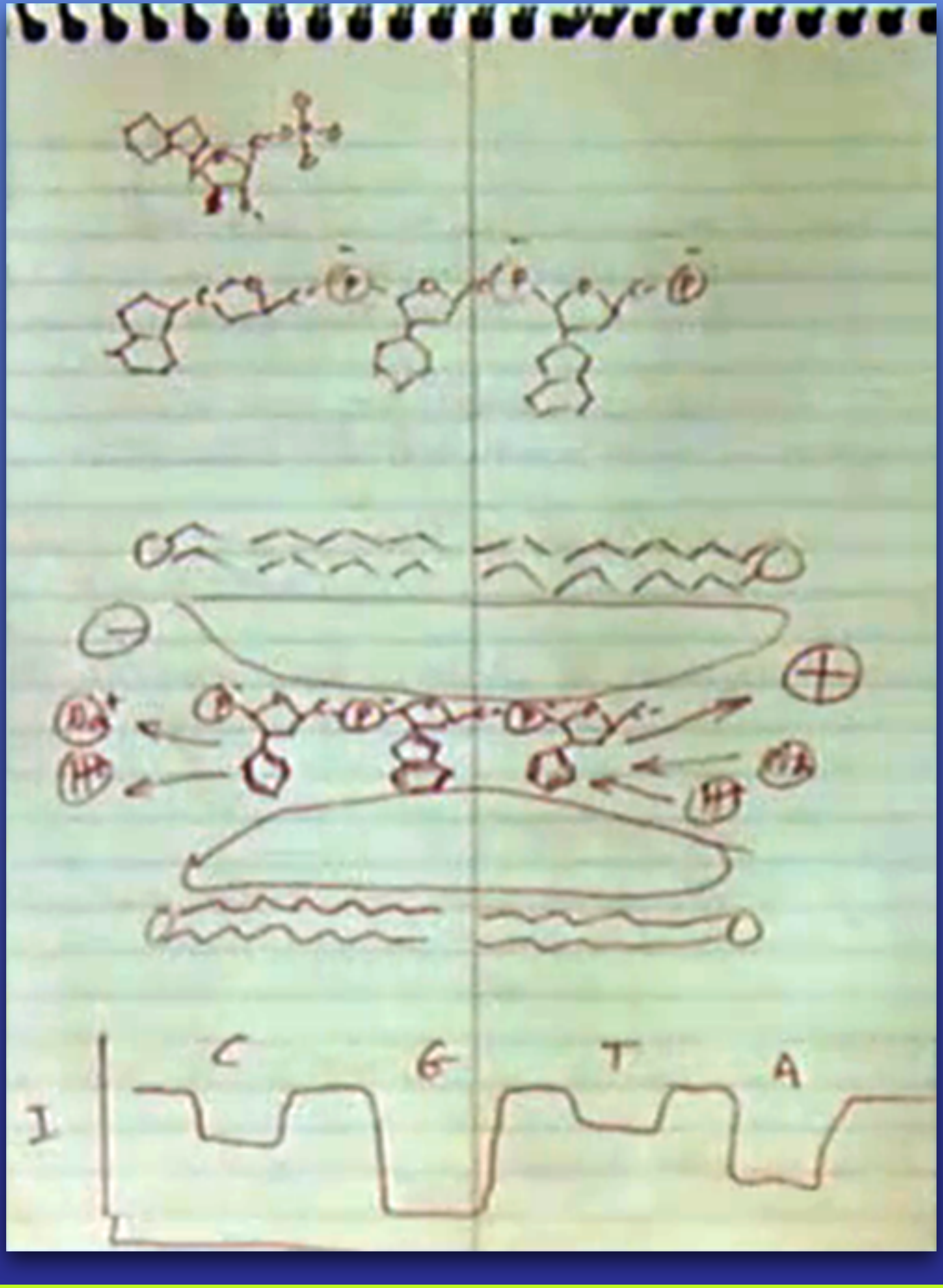
+

Primer & Polymerase

Prepare sequencing reaction

Use PacBio Sequel Systems to sequence genomes, transcriptomes, and epigenomes

SMRT Cells contain millions of zero-mode waveguides (ZMWs)



Sequencing using nanopores

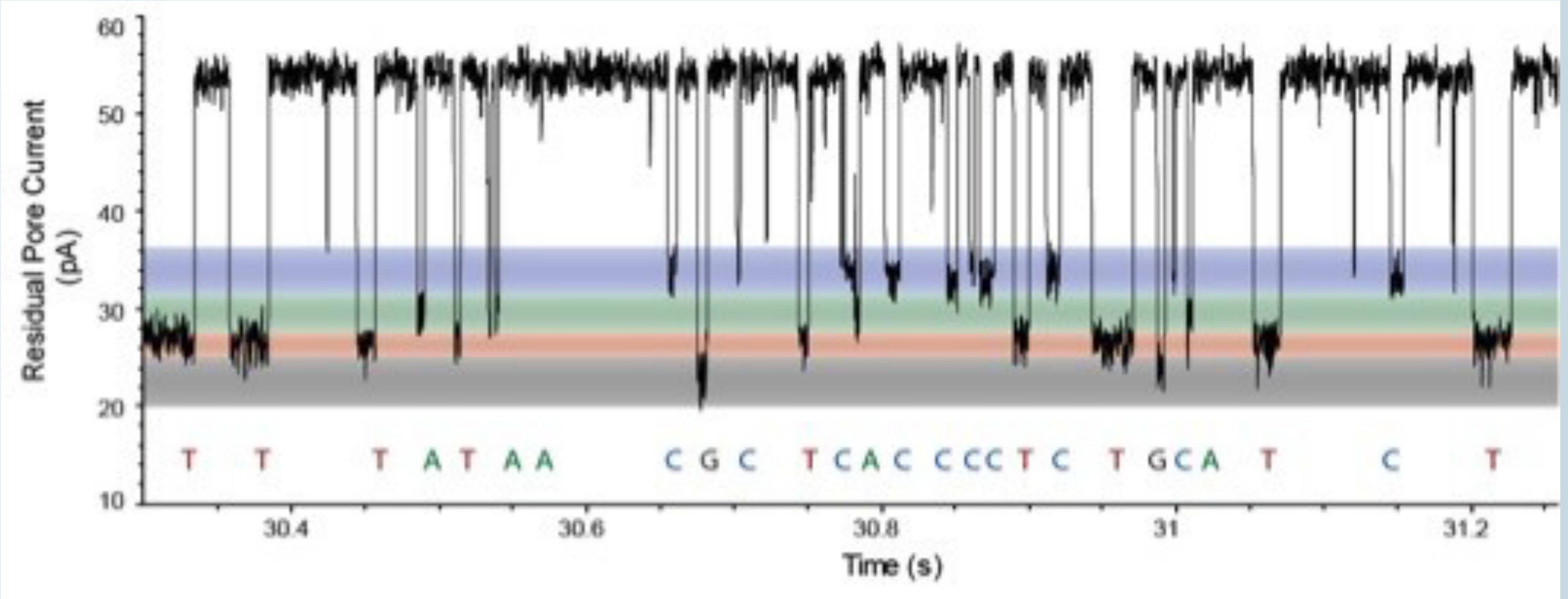
Nanopores as polymer sensors.

The idea emerged in early 1990s.

Fundamental work done by David Deamer and Daniel Branton in collaboration with John Kasianowicz. (PNAS 1996 146:13770-13773)

Biologicaly relevant experiments – since 2010.

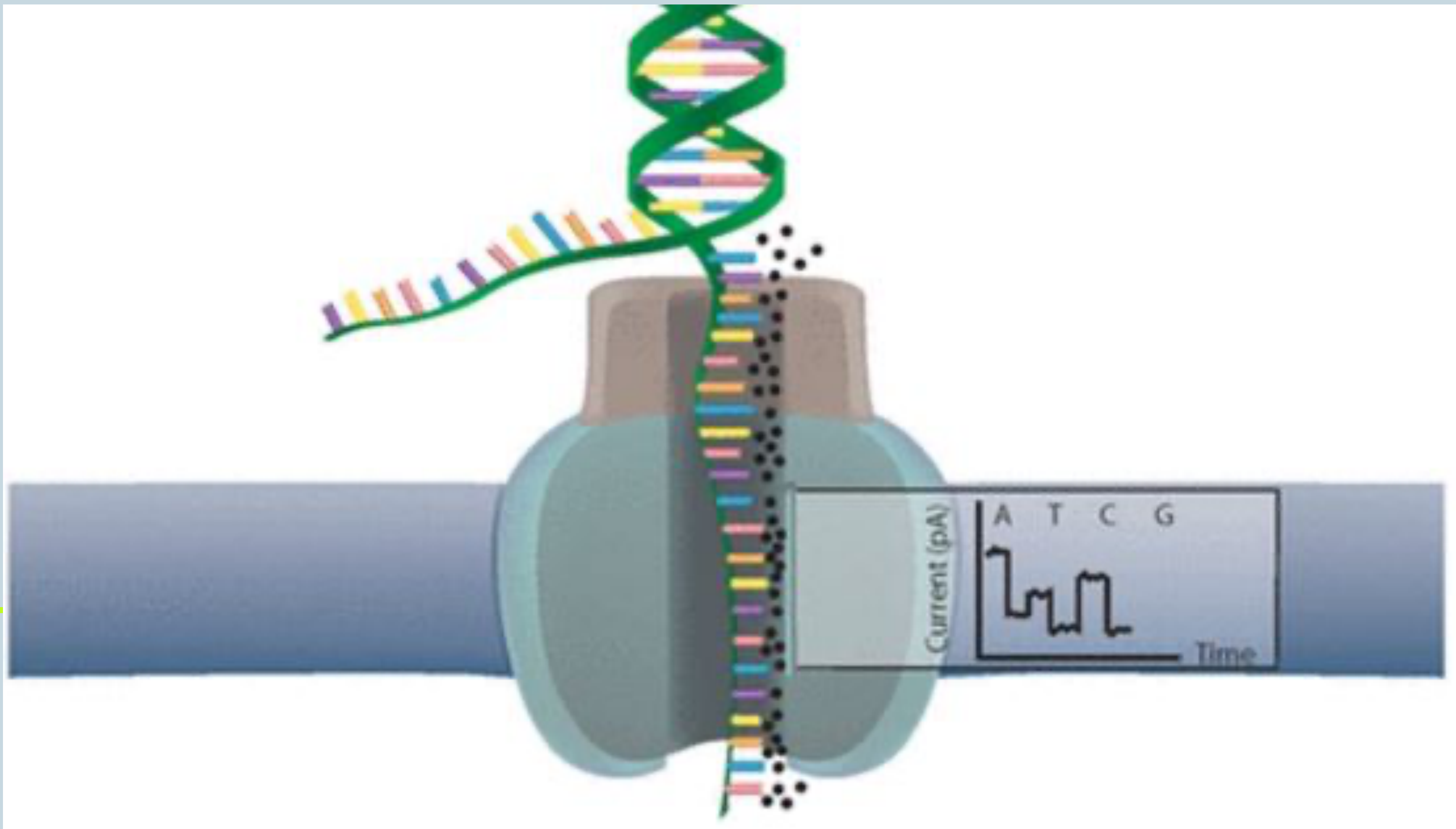
Current flow
current



https://nanoporetech.com/how-it-works

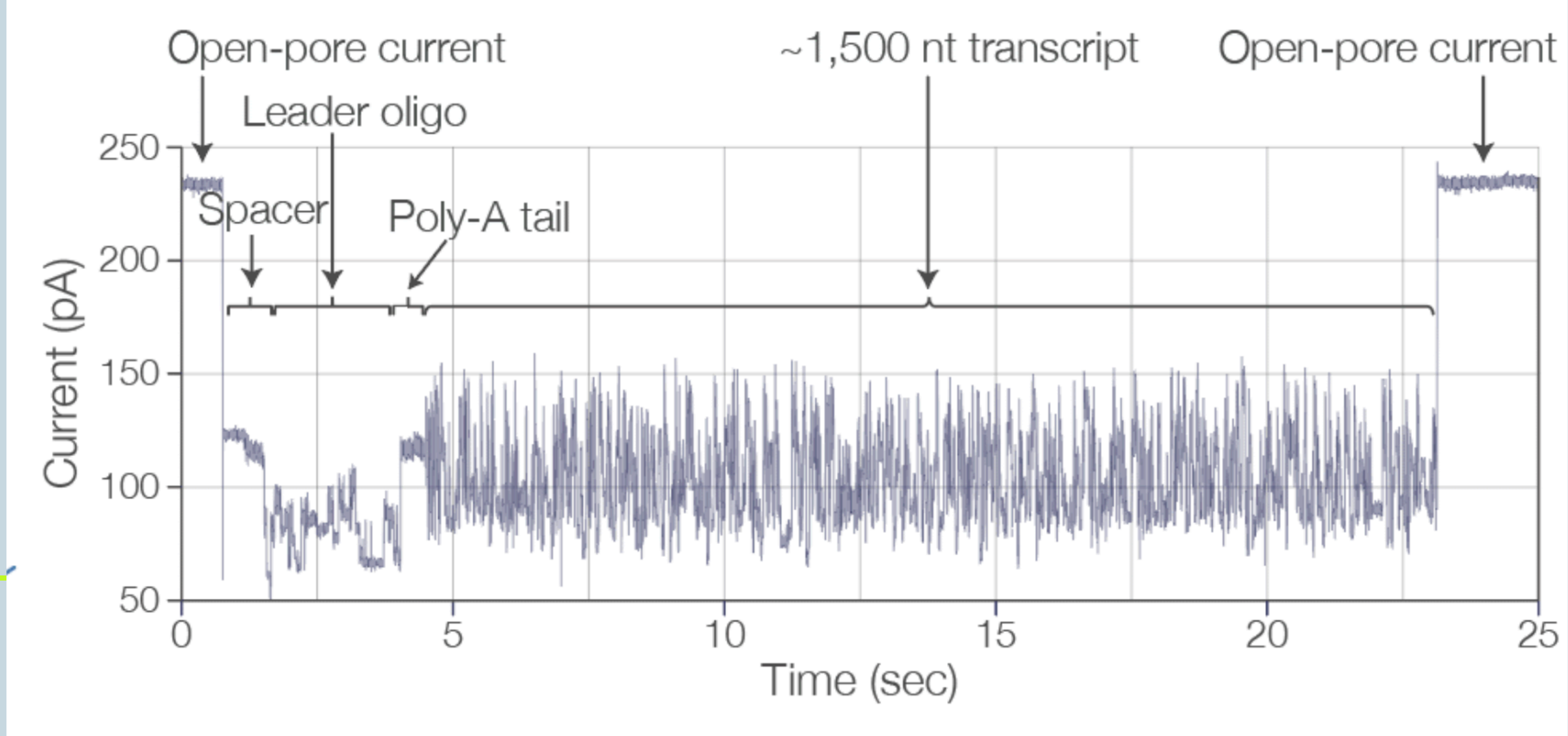
Nanopore basics

- Oxford Nanopore's first generation of technology uses bespoke, proprietary pore-forming proteins to create pores in membranes. Pore-forming proteins are common in nature.
- For example, the protein α-hemolysin and similar protein pores are found naturally in cell membranes, where they act as channels for ions or molecules to be transported in and out of cells.
- α-hemolysin is a heptameric protein pore with an inner diameter of 1 nm, about 100,000 times smaller than that of a human hair. This diameter is the same scale as many single molecules, including DNA. The pore is highly stable.
- Membrane is synthetic
- Non-destructive motor protein
- Read speed: about 400 bases per second



Nanopore basics - basecalling

- Raw electrical signal has to be translated to nucleotide sequence
- Originally, Hidden Markov Model based algorithms were used but performance was not so good; about 60-70% accuracy
- All recent basecallers are based on neutral networks.
- Electric signal produced by four nucleotides occupying a pore is processed at a time.
- Current accuracy of a single read is about 95% with consensus sequence produced at the accuracy level of 99.9%



ONT devices

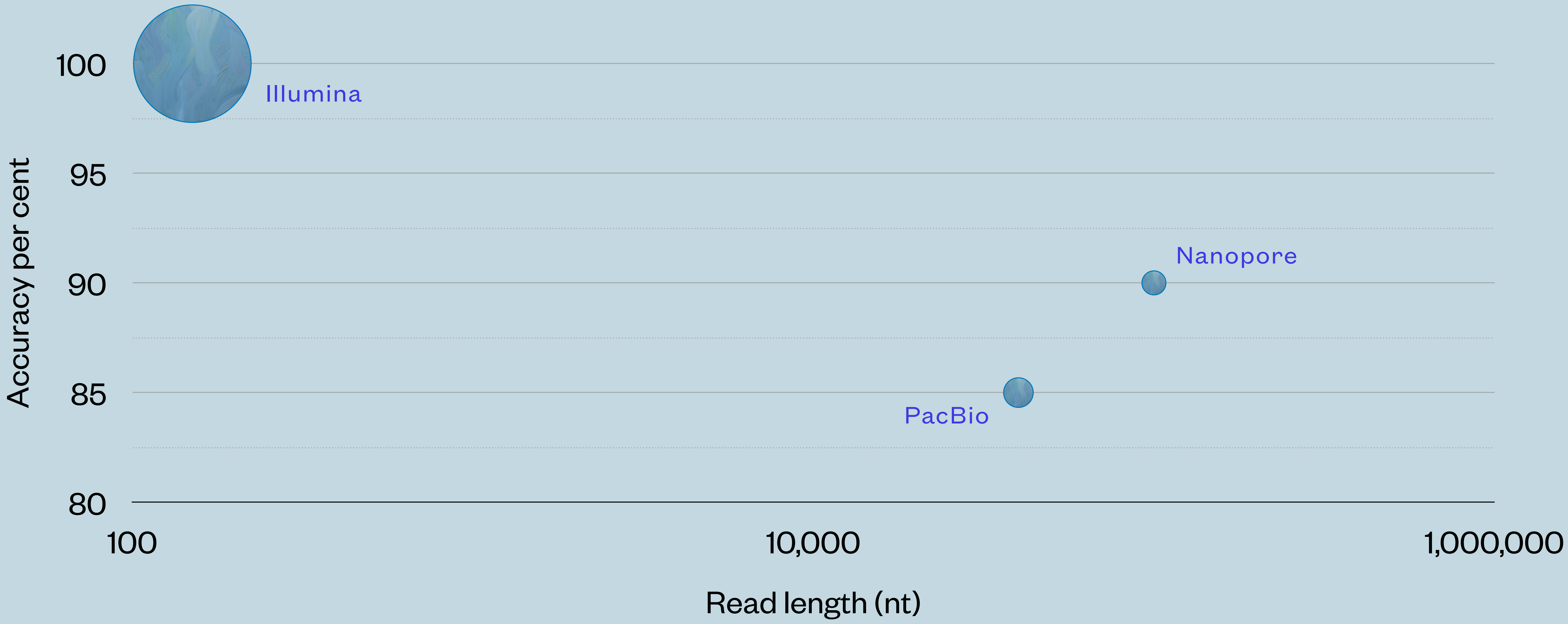
	Number of channels per flow cell	Yield per flow cell	Yield per device	Maximum run time	Application
Flongle	126	2 Gb	2 Gb	16 hr	Amplicons, panels/targeted sequencing, quality testing, small sequencing tests
MinION	512	50 Gb	50 Gb	48 hr	Whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome (cDNA), smaller transcriptomes (direct RNA), multiplexing for smaller samples
GridION	512	50 Gb	250 Gb	48 hr	Larger genomes or projects, whole transcriptomes (direct RNA or cDNA), large numbers of samples
PromethION 24	3000	220 Gb	5.2 Tb	72 hr	Very large genomes or projects, population-scale human, whole transcriptomes, very large numbers of samples
PromethION 48	3000	220 Gb	10.5 Tb	72 hr	

Pricing

	PacBio	Nanopore
Initial investment	$495,000 (Sequel II)	$1,000 - $327k
Single run	$1,300	$900
De novo small genome	$1,300	$900
De novo large genome	$2,600	$900 (?)
Whole transcriptome	$1,300	$900
Metagenomics (full-length 16S)	$15 (multiplexing up to 96 samples)	?

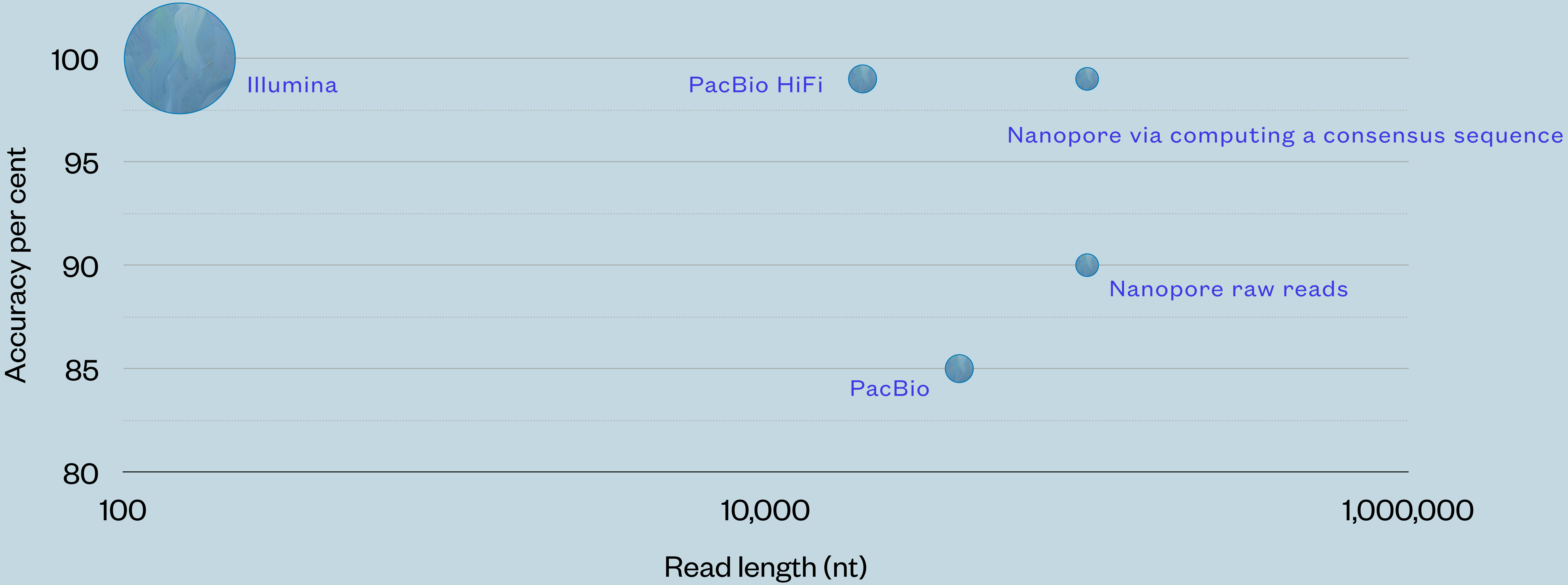
The Old Sequencing Paradigm

Sequence reads are long OR accurate



The New Sequencing Paradigm

Sequence reads can be long AND accurate



NGS vs 3rd Generation Sequencing

Technology	Read Length	Read Accuracy	Genome Characterization
Short Reads	300 bp	99.9%	single nucleotide variants, indels
Long Reads	>20 kb	89.0%	structural variants, assembly

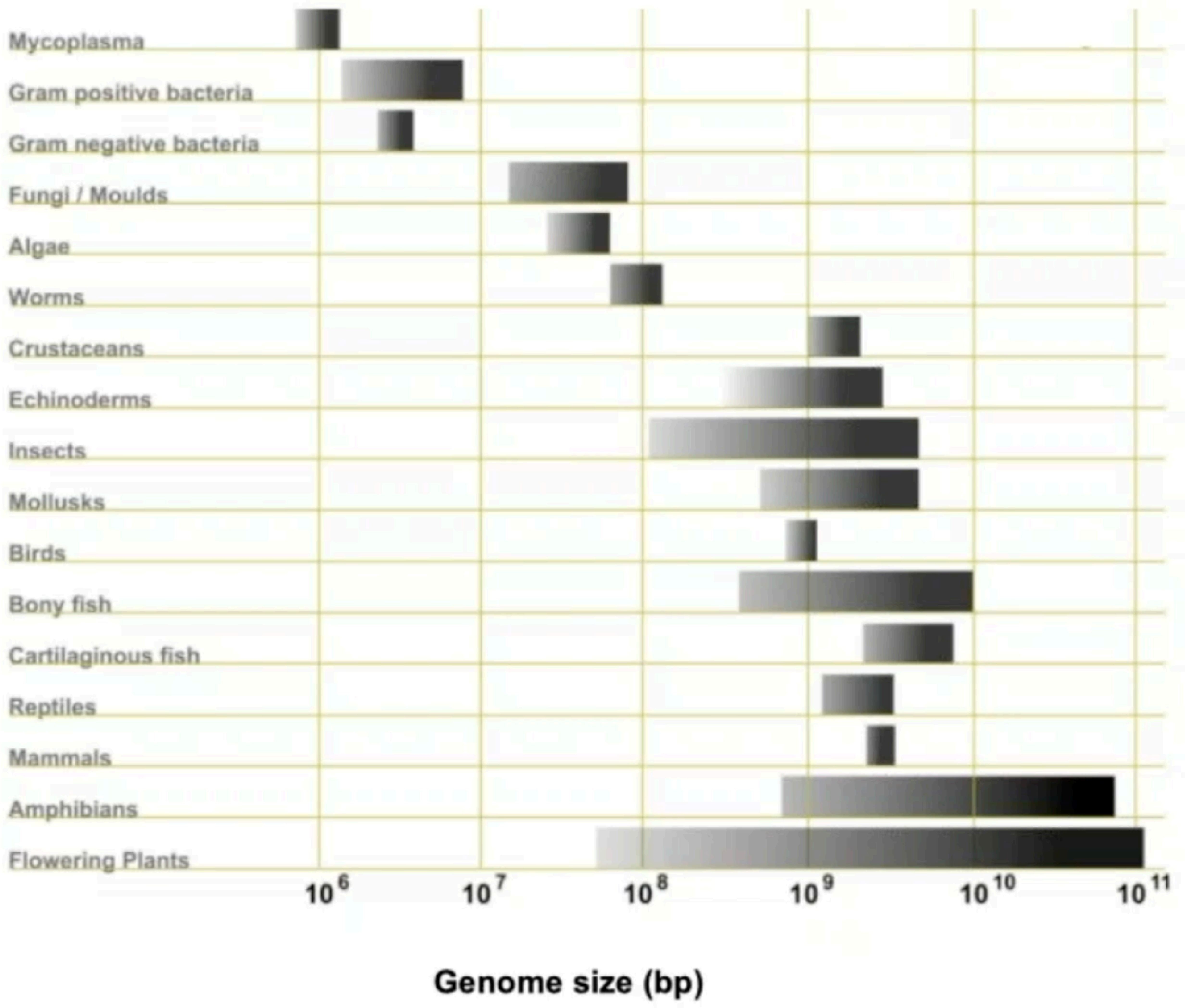
NGS vs 3rd Generation Sequencing

Technology	Read Length	Read Accuracy	Genome Characterization
Short Reads	300 bp	99.9%	single nucleotide variants, indels
Long Reads	>20 kb	89.0%	structural variants, assembly
PacBio CCS	10-20 kb	99.8%	comprehensive

Benefits of Long Reads

- Highly accurate *de novo* genome assembly
- Phase variants into haplotypes
- More accurate variant detection
- Sequencing full-length transcripts
- Exploring metagenomes in high resolution
- Epigenetics

Genome assembly



Challenges of Genome Assembly

- Size and complexity
  - human genome over 3 billion base pair
  - plants often have larger genomes
- Extreme repeat content
  - maize over 60%
  - wheat over 80%
- Each project is unique
  - ranges in size, ploidy, heterozygosity
  - custom strategy is required

Draft Versus Complete Genome

Short reads

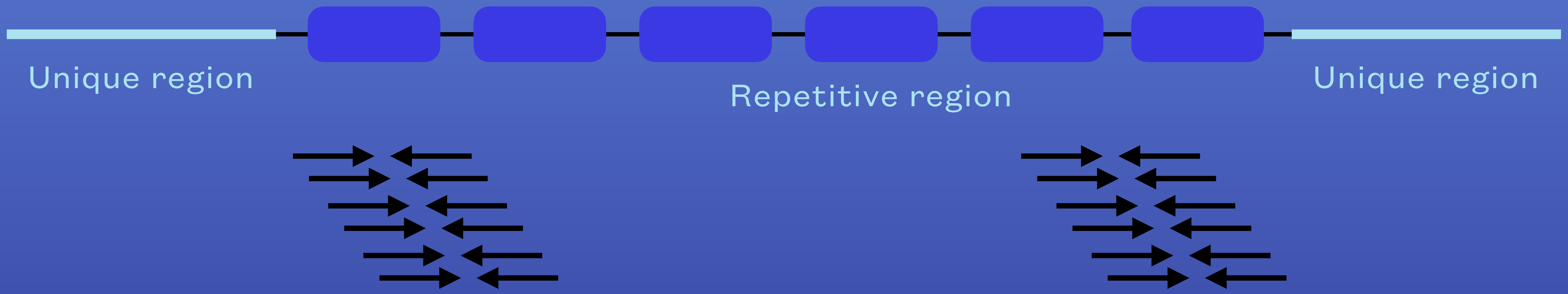
Long reads

Genome
Sequence Gaps
NNNNNN
NNNNN
NNNNNN
Draft Genome
Genome
Complete Genome

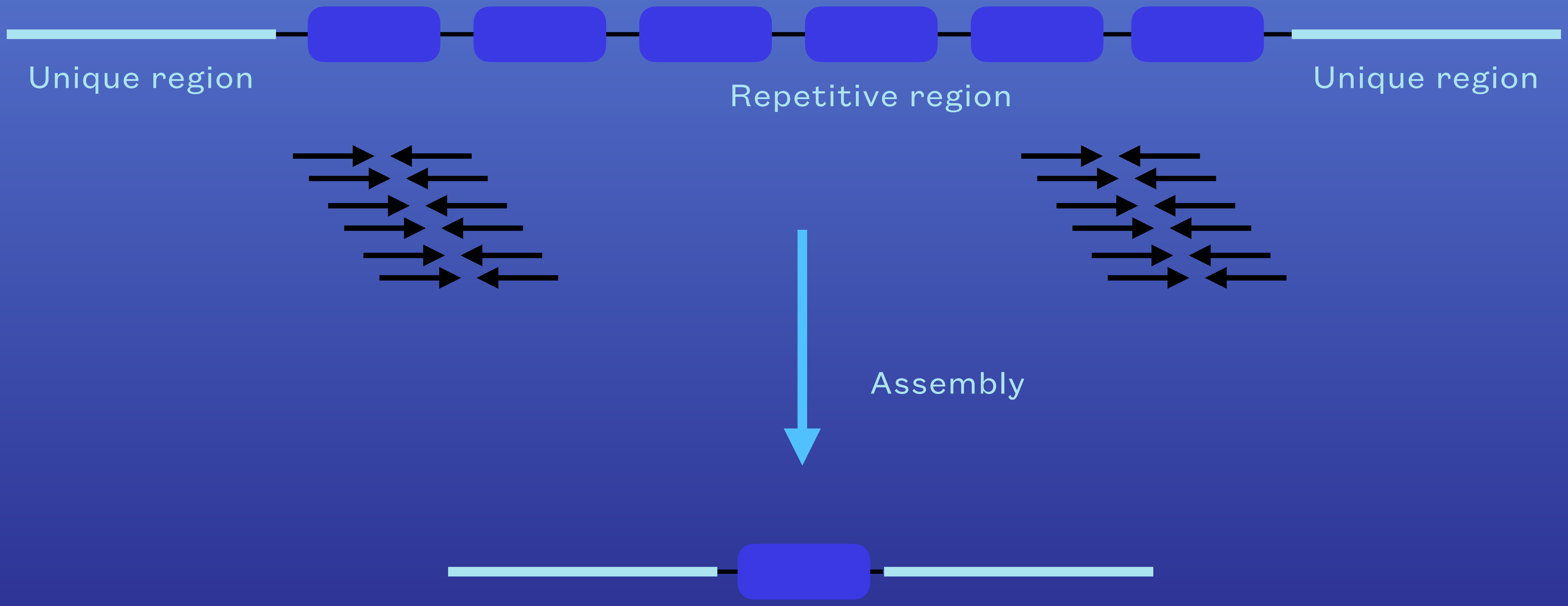
Missing sequencing leads to missed genes and limits biological interpretation

A comprehensive structural, functional and organizational picture of the genome

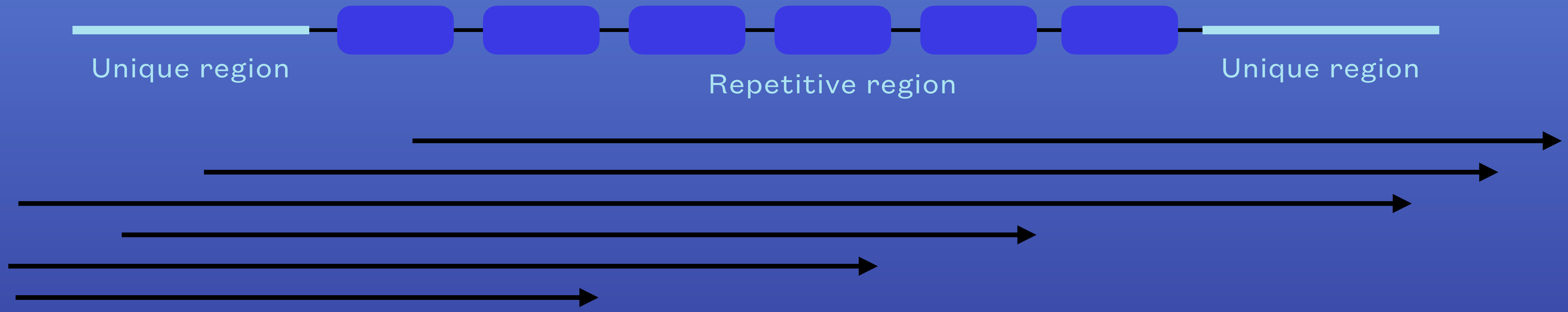
Sequencing through repetitive regions



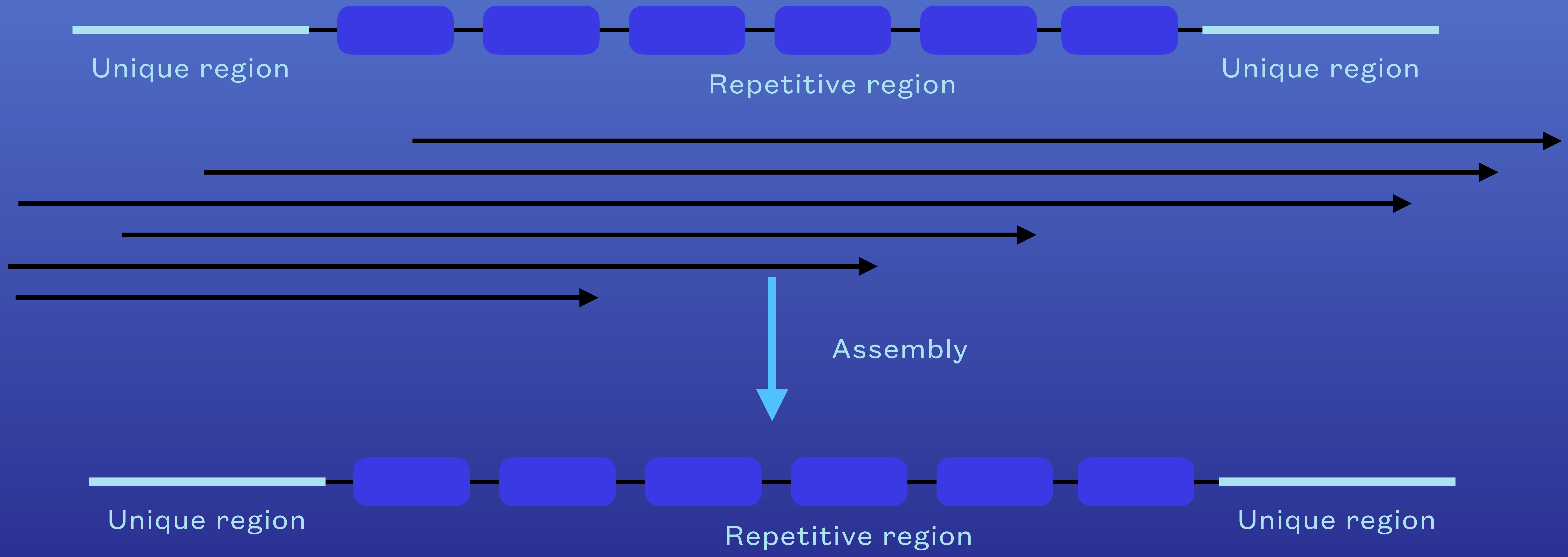
Sequencing through repetitive regions



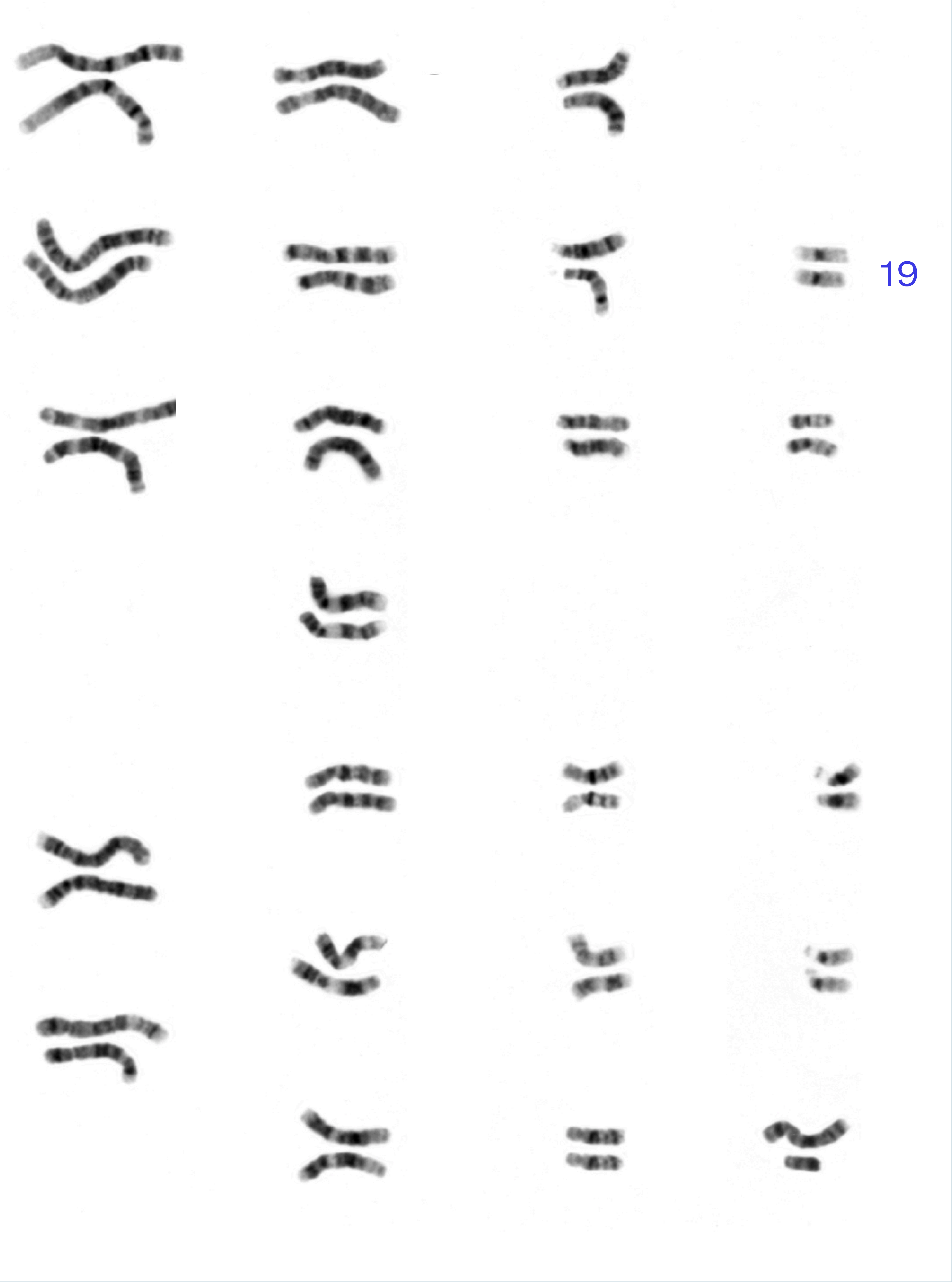
Sequencing through repetitive regions

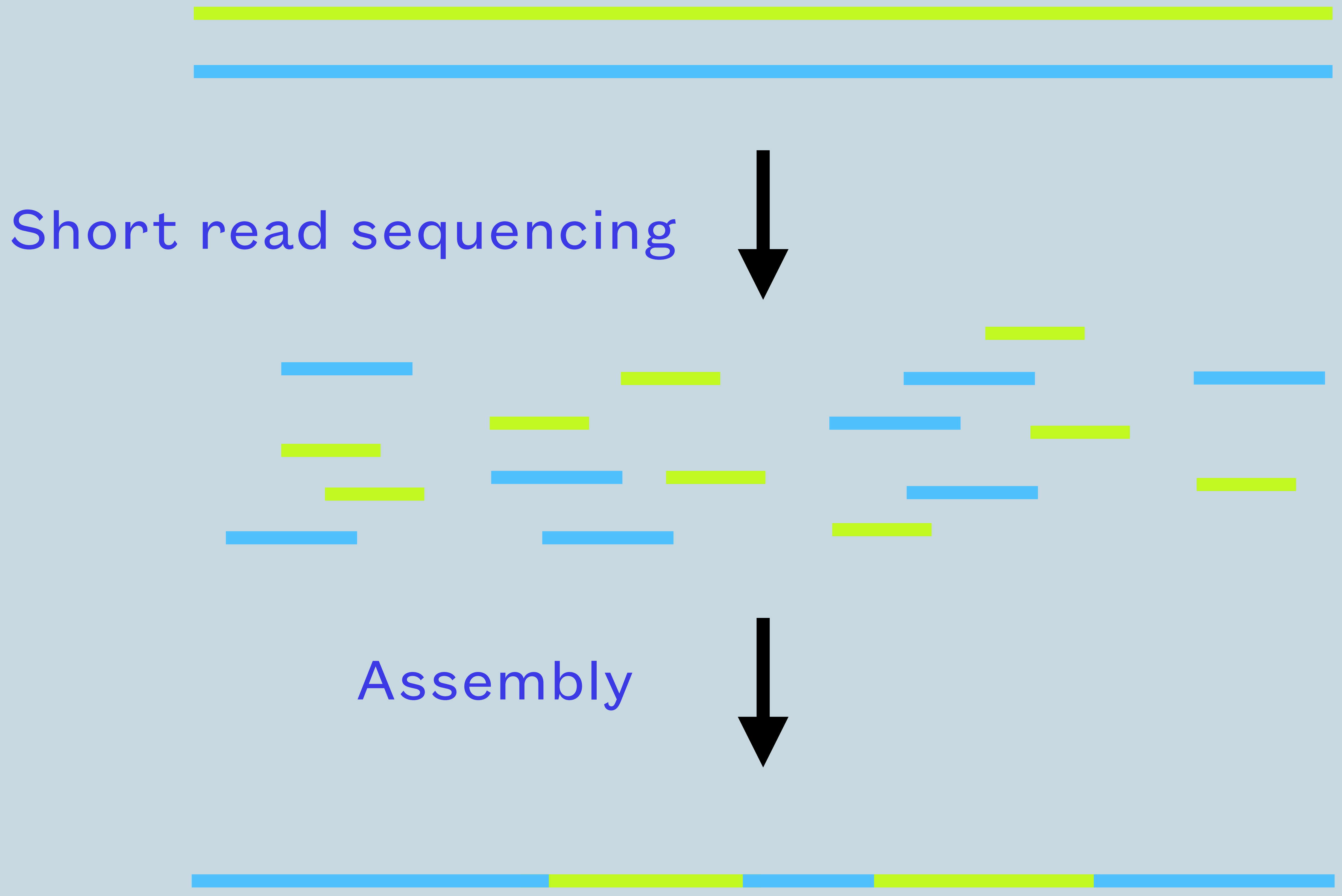


Sequencing through repetitive regions

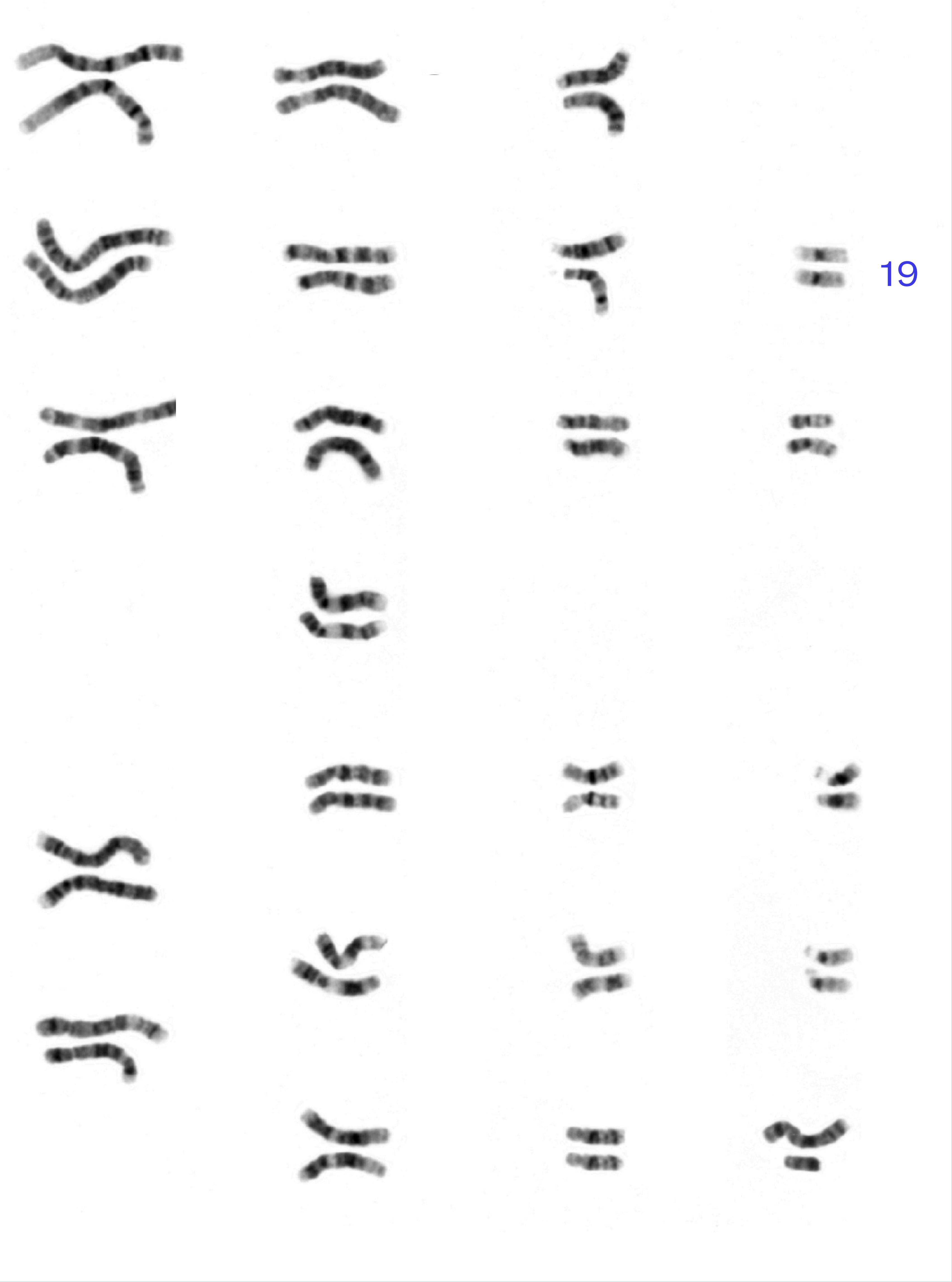


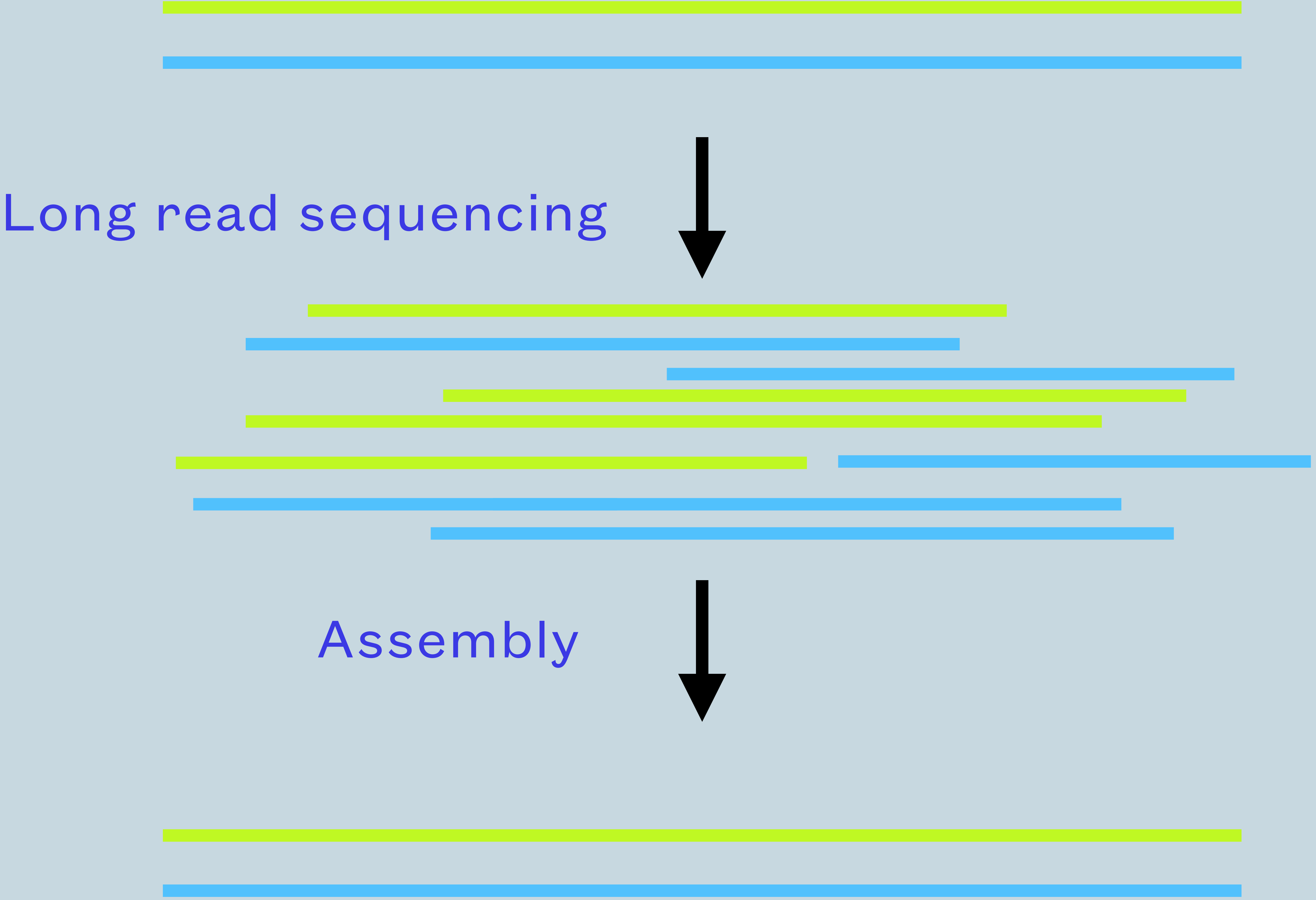
Haplotyping (phasing)





Haplotyping (phasing)



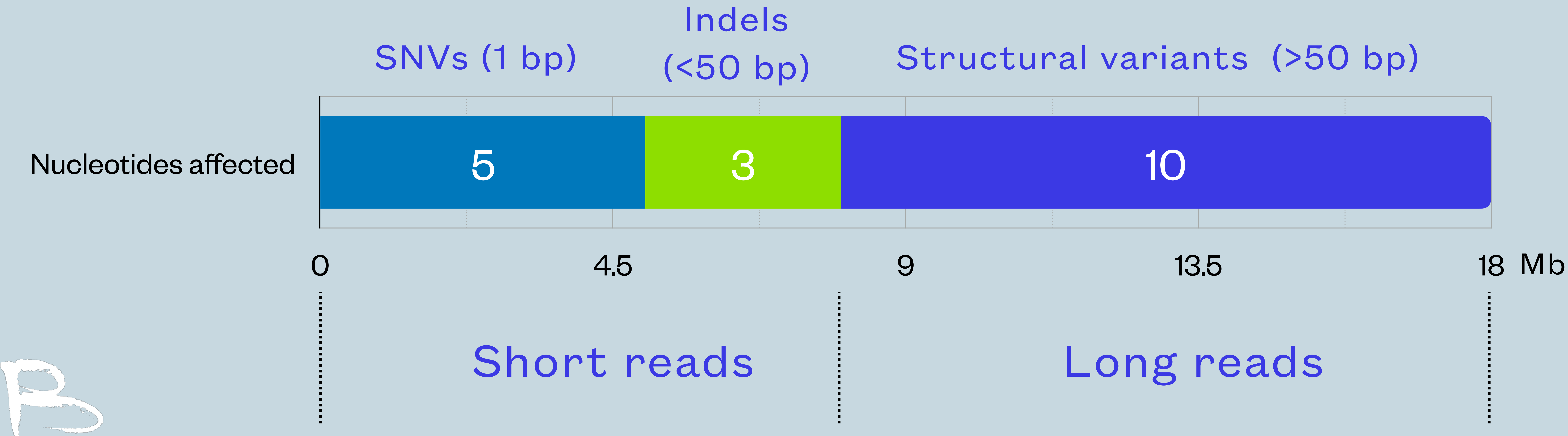


Variant detection

Type of variants

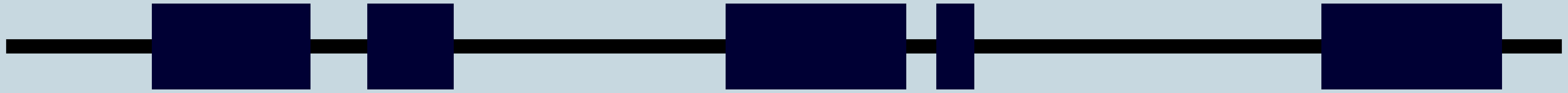
Single nucleotide variant	AAGTGGCATTACGTAG	Individual 1
	AAGTGTCATTACGTAG	Individual 2
Deletion	AAGTGGCATTACGTAG	Individual 1
	AAGTGCATTACGTAG	Individual 2
Insertion	AAGTGCATTACGTAG	Individual 1
	AAGTGGCATTACGTAG	Individual 2
Tandem duplication	AAGTGGCATTACGTAG	Individual 1
	AAGTGGCTGCATTACGTAG	Individual 2
Translocation	AAGTGCATTACGTAG	Individual 1
	AAGTTTACGTGCAAG	Individual 2
Inversion	AAGTGGCATTACGTAG	Individual 1
	AAGTTGCCTTACGTAG	Individual 2
Copy number variant	1 2 AAGTGCAGCATTACGTAG	Individual 1
	AAGTGCAGCAGCAGCATTACGTAG 1 2 3 4	Individual 2

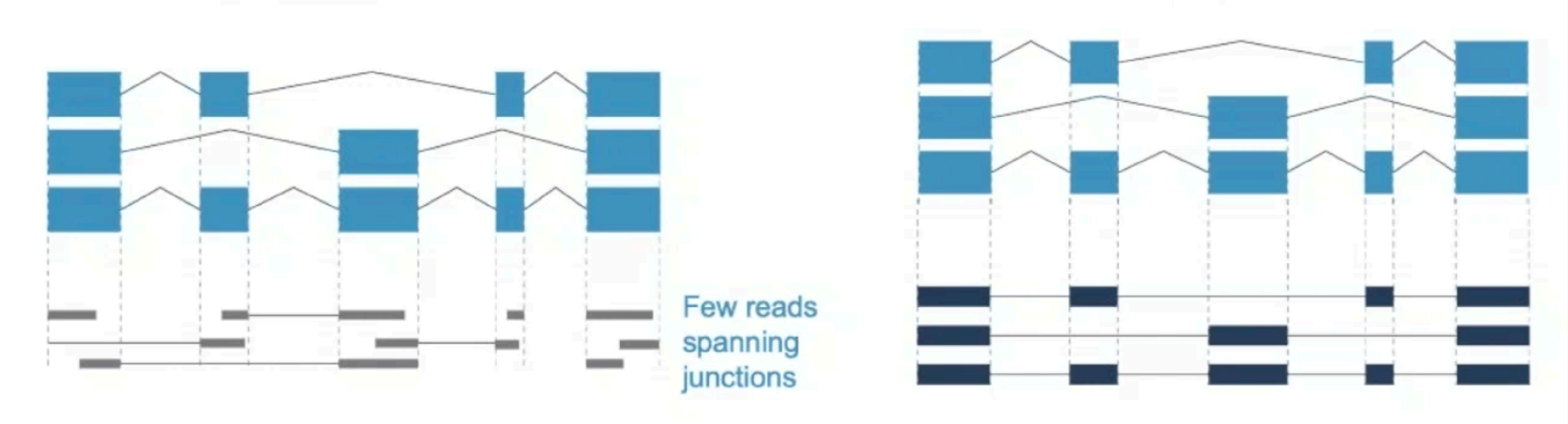
Genetic variation occurs at small and large scale



Isoforms (alternative splicing)

A gene consisting of five exons





Additional analysis required to recover all isoforms

Full length transcripts recover all isoforms

After sequencing is done

- Basecalling
- Mapping
- Sequence Assembly
- Variant detection
- and more

Software For Long Reads

Base callers for nanopore sequencing

Tool	Read qscore#	Consensus qscore#	Availability
Albacore	9.2	21.9	Only to ONT customers
BasecRAWller	N/A	N/A	https://basecrawller.lbl.gov/ (seems to be down)
Chiron	7.7	21.4	https://github.com/haotianteng/Chiron
DeepNano	N/A	N/A	https://bitbucket.org/vboza/deepnano/src/master/
Flappie	9.6	22.0	https://github.com/nanoporetech/flappie
Guppy	9.7	23.0	Only to ONT customers
Metrichor	N/A	N/A	Only to ONT customers
Nanocall	N/A	N/A	https://github.com/mateidavid/nanocall
Scrappie	9.3	22.4	https://github.com/nanoporetech/scrappie

Makałowski and Shabardina (2020) Bioinformatics of nanopore sequencing. J. Hum. Genet. 65:61-67.

Aligners

Tool	Algorithm	Availability
BWA	Burrows-Wheeler Aligner's Smith-Waterman Alignment	http://bio-bwa.sourceforge.net
GraphMap	Gapped spaced seeds	https://github.com/isovic/graphmap
Kart	Divide and conquer	https://github.com/hsinnan75/Kart
LAMSA	Sparse dynamic programming (SDP)-based split alignment	https://github.com/hitbc/LAMSA
LAST	Adaptive seeds approach	http://last.cbrc.jp/
Minimap2	Hash table approach	https://github.com/lh3/minimap
NGMLR	k-mer search followed by a banded Smith-Waterman alignment algorithm	https://github.com/philres/ngmlr
winnowmap	weighted-minimizer sampling algorithm	https://github.com/marbl/winnowmap

Makałowski and Shabardina (2020) Bioinformatics of nanopore sequencing. J. Hum. Genet. 65:61-67.

Assemblers (selected)

Tool	Description	Availability
Canu	A hierarchical assembly pipeline based on Celara Assembler	https://github.com/marbl/canu
Flye	De novo assembler for single molecule sequencing reads	https://github.com/fenderglass/Flye
MECAT	An ultra-fast mapping, error correction and de novo assembly tool for long reads	https://github.com/xiaochuanle/MECAT
Medaka	A tool to create a consensus sequence of nanopore sequencing data using neural networks	https://nanoporetech.github.io/medaka/index.html
NanoPipe	A pipeline that includes a consensus sequence calculation based on LAST alignment to a reference sequence	http://bioinformatics.uni-muenster.de/tools/nanopipe2/index.hbi
Nanopolish	Software package for signal-level analysis of Oxford Nanopore sequencing data, including consensus sequence calculation	https://github.com/jts/nanopolish
Shasta	Using a run-length representation of the read sequence and a representation of the read sequence based on *markers*, a fixed subset of short k-mers (k ≈ 10).	https://github.com/chanzuckerberg/shasta

Makałowski and Shabardina (2020) Bioinformatics of nanopore sequencing. J. Hum. Genet. 65:61-67.

Variant calling

Tool	Description	Availability
Clair	Deep neural network based variant caller	https://github.com/HKU-BAL/Clair
HapCUT2	It is a maximum-likelihood-based tool for assembling haplotypes.	https://github.com/vibansal/HapCUT2
IDP-ASE	Haplotyping and quantification of allele-specific expression	http://augroup.org/IDP-ASE/IDP-ASE
Medaka	An experimental pipeline to call SNPs	https://nanoporetech.github.io/medaka/index.html
NanoPipe	A pipeline that includes a consensus sequence calculation based on LAST alignment to a reference sequence	http://bioinformatics.uni-muenster.de/tools/nanopipe2/index.hbi https://github.com/IOB-Muenster/nanopipe2
Nanopolish	Software package for signal-level analysis of Oxford Nanopore sequencing data, including SNP and indel calling	https://github.com/jts/nanopolish
PBHoney	An implementation of variant-identification designed for long reads	https://sourceforge.net/projects/pb-jelly/
Sniffles	Sniffles is a structural variation (over 10 bp) caller using third generation sequencing	https://github.com/fritzsedlazeck/Sniffles
WhatsHap	It is a software for phasing genomic variants	https://whatshap.readthedocs.io/en/latest/

NanoPipe

Institute of Bioinformatics Münster
HOME
TEACHING
PUBLICATIONS
THESES
TOOLS
INTERN
MYWWU
UNIVERSITY CLINICS MUENSTER
BIOLOGY
DE
NanoPipe
About
Usage
Run the Pipeline
Contact
NanoPipe

Welcome to

The NanoPipe pipeline analyzes reads generated by the Oxford Nanopore sequencing devices. As a result, it provides alignments to any target of interest, alignment statistics and information about polymorphisms.

The NanoPipe development team is open to suggestions and requests for implementing new targets.

Presented by Victoria Shabardina, Norbert Grundmann, Felix Manske, Tabea Kischka and Wojtek Makalowski

2019-10-21 14:12

http://bioinformatics.uni-muenster.de/tools/nanopipe2/index.hbi?lang=en

NanoPipe
About
Usage
Run the Pipeline
View All Requests
Contact
Previous Runs / Views
Aug-20: 156628333047847
EGFR_1D.fa
Run
View Human Test Case
View Plasmodium Test Case
Set Default Parameters
Reset

NanoPipe New Request 2020/11/04

Previous Request

ID ?	

New Request

Target ?	Upload File
Target File ?	
Query File ?	
Minimum Sequence Length ?	100
Email ?	
Title ?	

Last Parameters

Substitution Matrix Load Init ?

Use Matrix or Match Score / Mismatch Cost

	A	C	G	T
A	5	-3	-2	-14
C	-7	6	-6	-9
G	-4	-6	6	-14
T	-14	-9	-8	5

Gap Existence Cost (-a) ?	10
Gap Extension Cost (-b) ?	4
Insertion Existence Cost (-A) ?	17
Insertion Extension Cost (-B) ?	3
Score Matrix applies to Forward Strand (-S) ?	1
Initial Matches Position (-k) ?	
Maximum Score Drop (-x) ?	
Polymorphism threshold ?	0.2
Target threshold ?	0.8
Coverage threshold ?	0.3

MetaG

Institute of Bioinformatics Münster
HOME TEACHING PUBLICATIONS THESES TOOLS INTERN
MYWWU
UNIVERSITY CLINICS MUENSTER
BIOLOGY
DE
MetaG
About
Usage
Tutorial ...
Run the Pipeline
Download
Contact
News
NEW
virus database
MetaG

2020-07-29
ICTV updated to MSL35 version August 1 changes

2020-08-13
Stay updated via RSS »

2020-07-29
ICTV updated to MSL35 version May 1 changes

2020-04-23
Improved pathogen graph for viruses

2020-03-15
Preprint is available »

2020-03-13
Desktop version release »

2020-03-13
Updated testcases

Welcome to

This metagenomics pipeline analyzes reads from both targeted and whole genome sequencing. It is suitable for long and short-read sequencing technologies. Each read will be reported with its respective taxonomy. Additionally, you will receive a summary covering all organisms in your sample. Our databases cover organisms throughout the tree of life: Bacteria, archaea, fungi and further selected eukaryotes. As an additional feature, pathogenic bacteria or archaea are predicted in your sample. All results are visualized in intuitive interactive graphs.

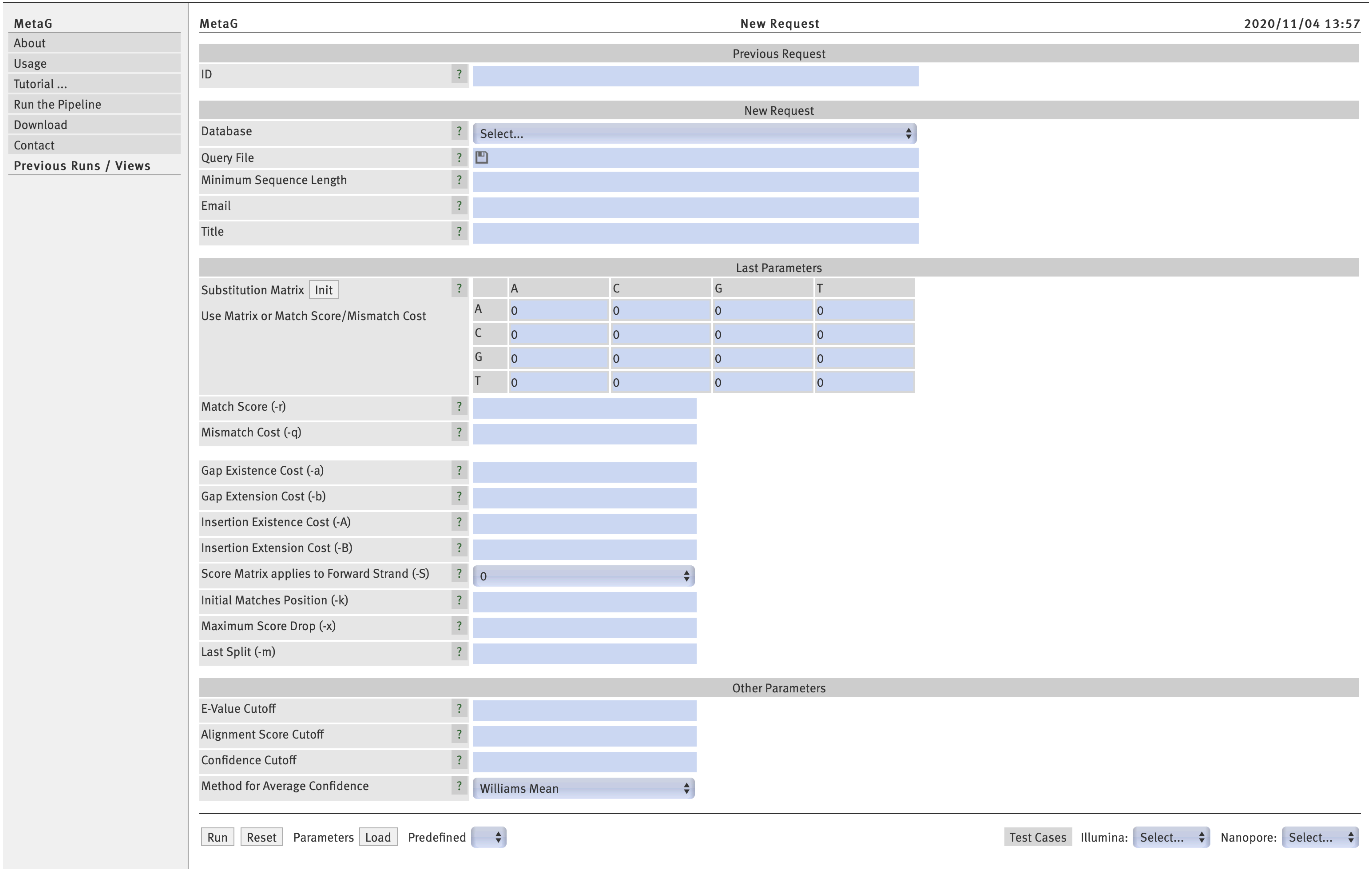
The Metagenomics development team is open to suggestions and requests for implementing new targets.

Presented by Felix Manske, Norbert Grundmann and Wojtek Makalowski

Logo created by Maciej Makalowski.

2019-11-18 10:35

http://bioinformatics.uni-muenster.de/tools/metag/index.hbi?lang=en



MetaG New Request 2020/11/04 13:57

Previous Request

ID ?

New Request

Database ? Select...

Query File ?

Minimum Sequence Length ?

Email ?

Title ?

Last Parameters

Substitution Matrix Init ?

Use Matrix or Match Score/Mismatch Cost

	A	C	G	T
A	0	0	0	0
C	0	0	0	0
G	0	0	0	0
T	0	0	0	0

Match Score (-r) ?

Mismatch Cost (-q) ?

Gap Existence Cost (-a) ?

Gap Extension Cost (-b) ?

Insertion Existence Cost (-A) ?

Insertion Extension Cost (-B) ?

Score Matrix applies to Forward Strand (-S) ? 0

Initial Matches Position (-k) ?

Maximum Score Drop (-x) ?

Last Split (-m) ?

Other Parameters

E-Value Cutoff ?

Alignment Score Cutoff ?

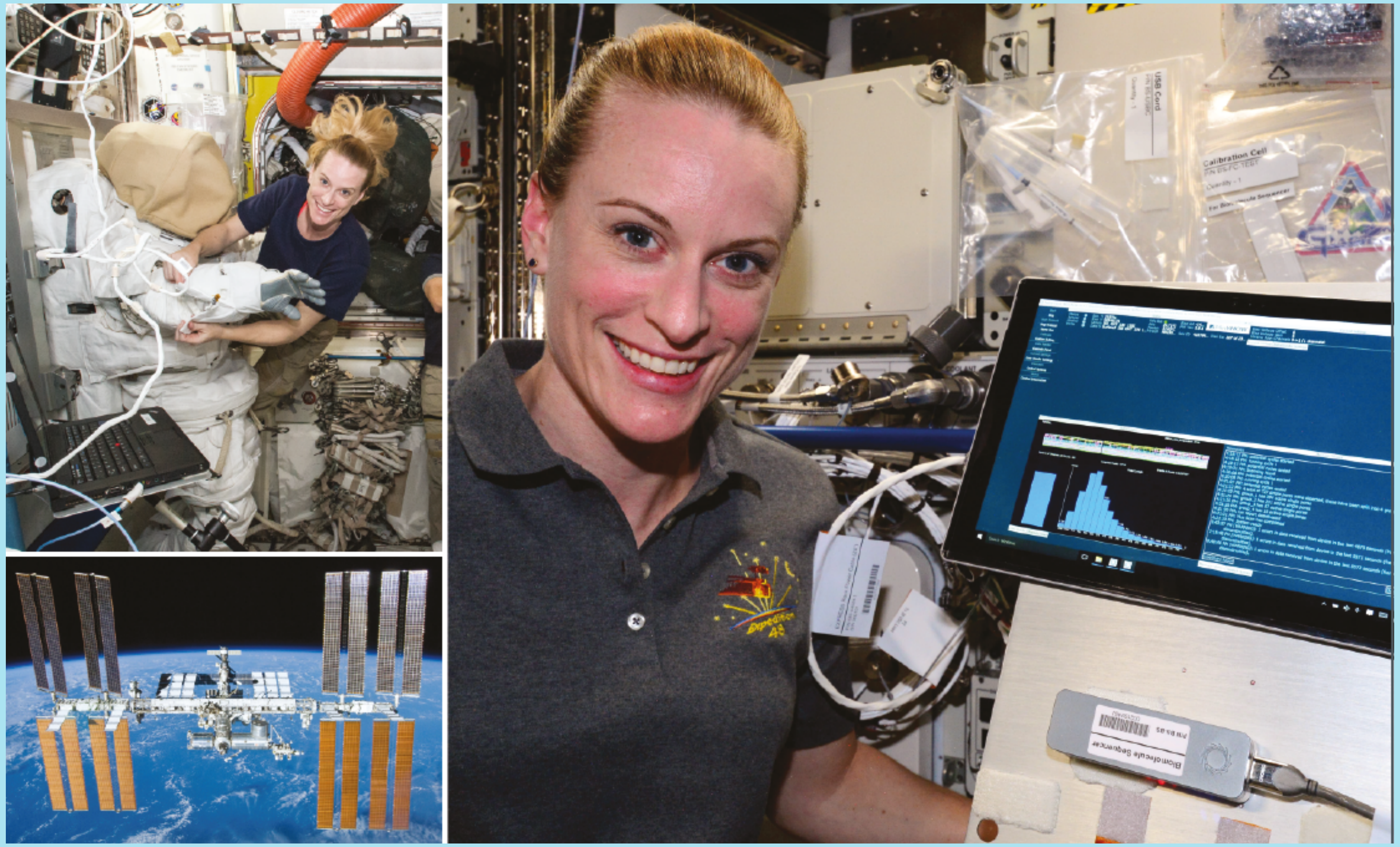
Confidence Cutoff ?

Method for Average Confidence ? Williams Mean

Cool Projects

Bringing Sequencing to the Masses

- Sequencing literally anywhere



Astronaut Dr. Kate Rubins on the ISS

Dr. Jacqueline Goordial, University of Guelph, Canada

Bringing Sequencing to the Masses

- Sequencing in rural areas of underdeveloped countries, helping to fight infectious diseases.

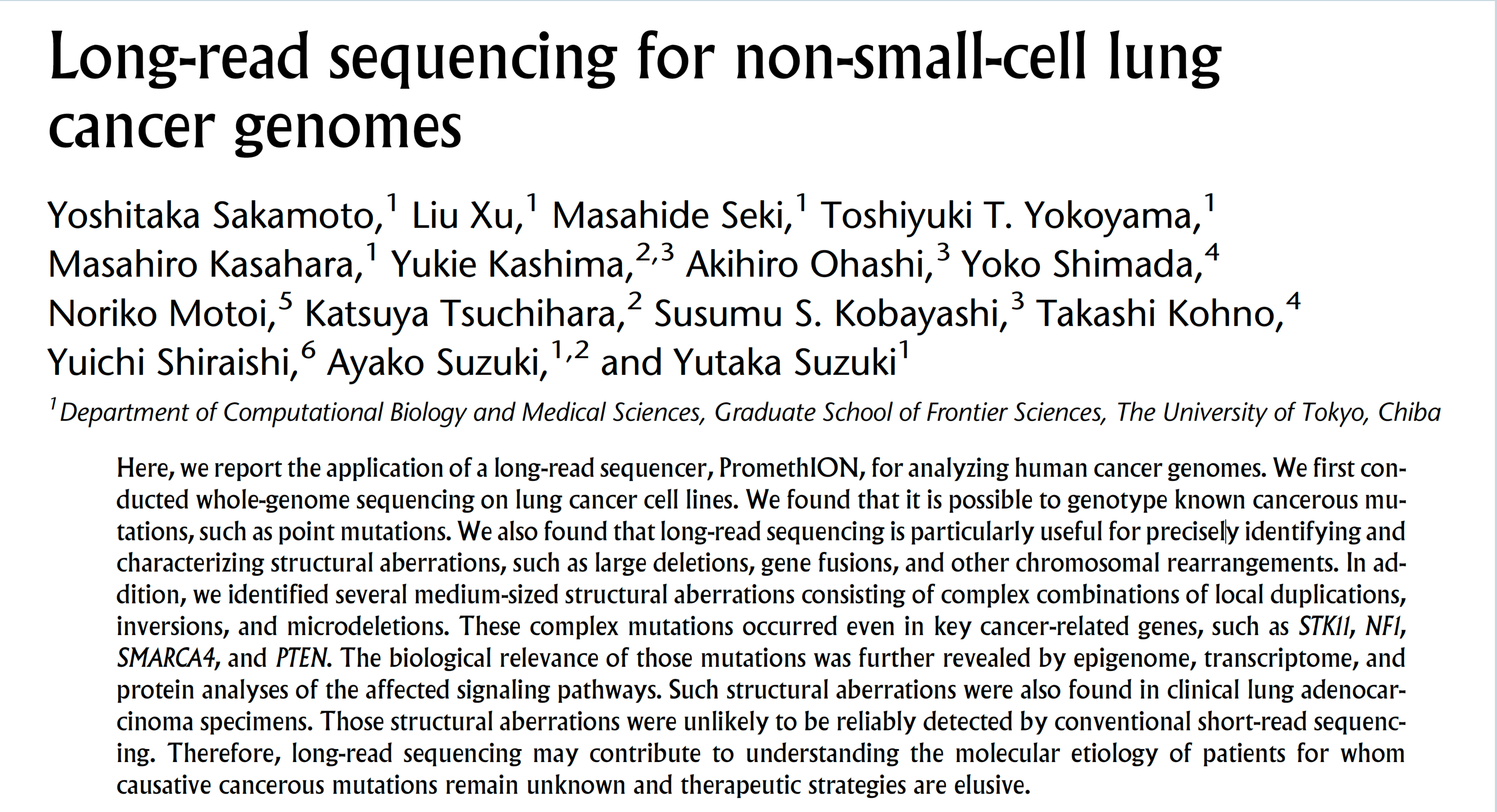
MinION workshop in Manado, Indonesia

and Bangkok

Yamagishi J, Runtuwene LR, Hayashida K, Mongan AE, Thi LAN, Thuy LN, Nhat CN, Limkittikul K, Sirivichayakul C, Sathirapongsasuti N, Frith M, Makalowski W, Suzuki Y (2017) Serotyping dengue virus with isothermal amplification and a portable sequencer. Scientific Reports 7: 3510

Runtuwene LR, Tuda JSB, Mongan AE, Makalowski W, et al. Y. (2018) Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum. Sci Rep. 8:8286.

Cancer Genomics



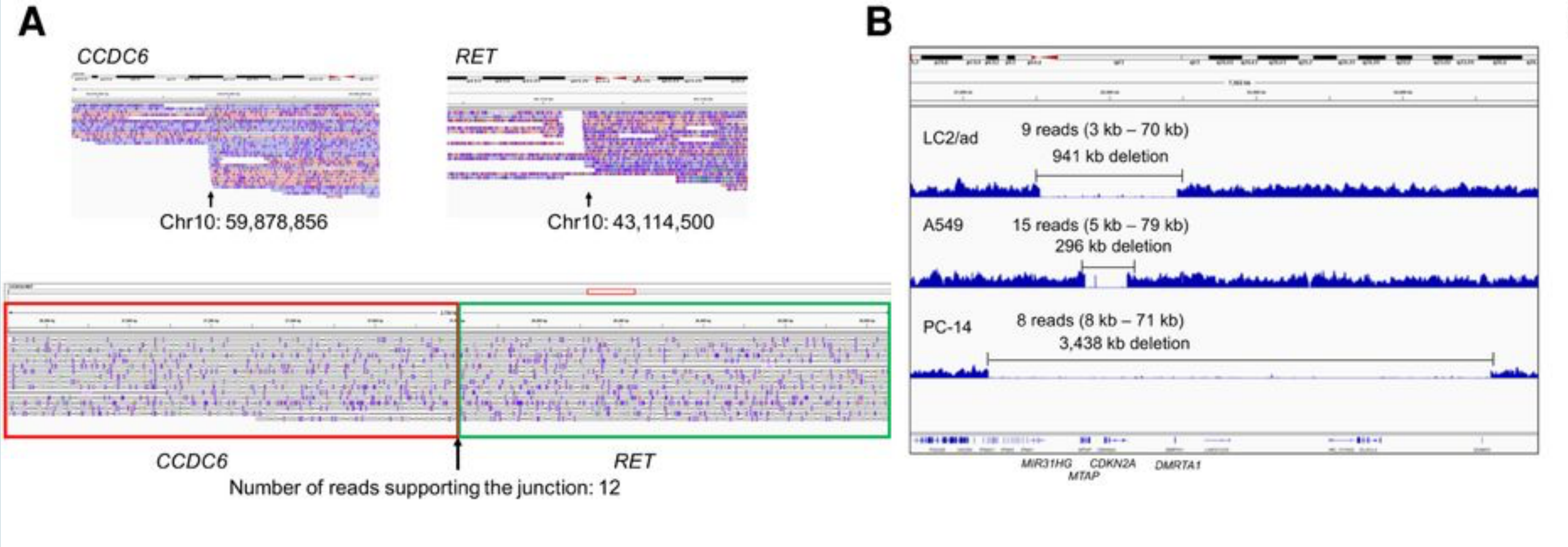
Long-read sequencing for non-small-cell lung cancer genomes

Yoshitaka Sakamoto,[1] Liu Xu,[1] Masahide Seki,[1] Toshiyuki T. Yokoyama,[1] Masahiro Kasahara,[1] Yukie Kashima,[2,3] Akihiro Ohashi,[3] Yoko Shimada,[4] Noriko Motoi,[5] Katsuya Tsuchihara,[2] Susumu S. Kobayashi,[3] Takashi Kohno,[4] Yuichi Shiraishi,[6] Ayako Suzuki,[1,2] and Yutaka Suzuki[1]

[1] Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba

Here, we report the application of a long-read sequencer, PromethION, for analyzing human cancer genomes. We first conducted whole-genome sequencing on lung cancer cell lines. We found that it is possible to genotype known cancerous mutations, such as point mutations. We also found that long-read sequencing is particularly useful for precisely identifying and characterizing structural aberrations, such as large deletions, gene fusions, and other chromosomal rearrangements. In addition, we identified several medium-sized structural aberrations consisting of complex combinations of local duplications, inversions, and microdeletions. These complex mutations occurred even in key cancer-related genes, such as *STK11*, *NF1*, *SMARCA4*, and *PTEN*. The biological relevance of those mutations was further revealed by epigenome, transcriptome, and protein analyses of the affected signaling pathways. Such structural aberrations were also found in clinical lung adenocarcinoma specimens. Those structural aberrations were unlikely to be reliably detected by conventional short-read sequencing. Therefore, long-read sequencing may contribute to understanding the molecular etiology of patients for whom causative cancerous mutations remain unknown and therapeutic strategies are elusive.

Cancer Genomics



Sakamoto et al. (2020) Long-read sequencing for non-small-cell lung cancer genomes. Genome Res. 30:1243-1257

The Telomere-toTelomere (T2T) consortium

- Community-based effort to generate the first complete assembly of a human genome.
- The consortium aims to finish remaining unresolved regions and generate the first truly complete assembly of a human genome. These regions include segmental duplications, ribosomal rRNA gene arrays, and satellite arrays that harbor unexplored variation of unknown consequence.
- Data: 50X coverage of ultra-long Oxford Nanopore sequencing for the CHM13hTERT cell line, including 44 Gb of sequence in reads 100 kb+ and a maximum read length exceeding 1 Mb.
- This coverage of ultra-long reads enabled the resolution of most repeats in the genome, including large fractions of the centromeric satellite arrays and short arms of the acrocentrics. A de novo assembly combining this nanopore data with 70X of existing PacBio data achieved an NG50 contig size of 75 Mb (compared to 56 Mb for GRCh38), with some chromosomes broken only at the centromere.
- Using this assembly as a basis, they manually finished the X chromosome. The few unresolved segmental duplications were assembled using ultra-long reads spanning the individual copies, and the ~2.8 Mbp X centromere was assembled by identifying unique variants within the array and using these to anchor overlapping ultra-long reads.

Miga KH, Koren S, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature, 2020.

Logsdon GA, et al. The structure, function, and evolution of a complete human chromosome 8. bioRxiv, 2020.

Conclusions

- Long reads are not necessarily "noisy" any more
- Sequencing is getting not only affordable but also easy to use almost at any place
- Computational analyses lag behind sequencing technology development
- "Sequencing for the masses" is the present not the future!

Acknowledgments