CMSC 829C: Algorithms, AI, and Hardware for Biological Data

Lecture 3: Introduction to Genomics and Sequencing

Fall 2026

September 8, 2026

Can Firtina | canfirtina@gmail.com | cs.umd.edu/~firtina

Assistant Professor of Computer Science at UMD

UNIVERSITY OF MARYLAND

Agenda

- A quick introduction to genomes and genome analysis
- Sequencing technologies
  - Particular focus: Nanopore sequencing and real-time analysis

Recall: Analyzing Genomes Reveals Key Insights

Personalized Medicine

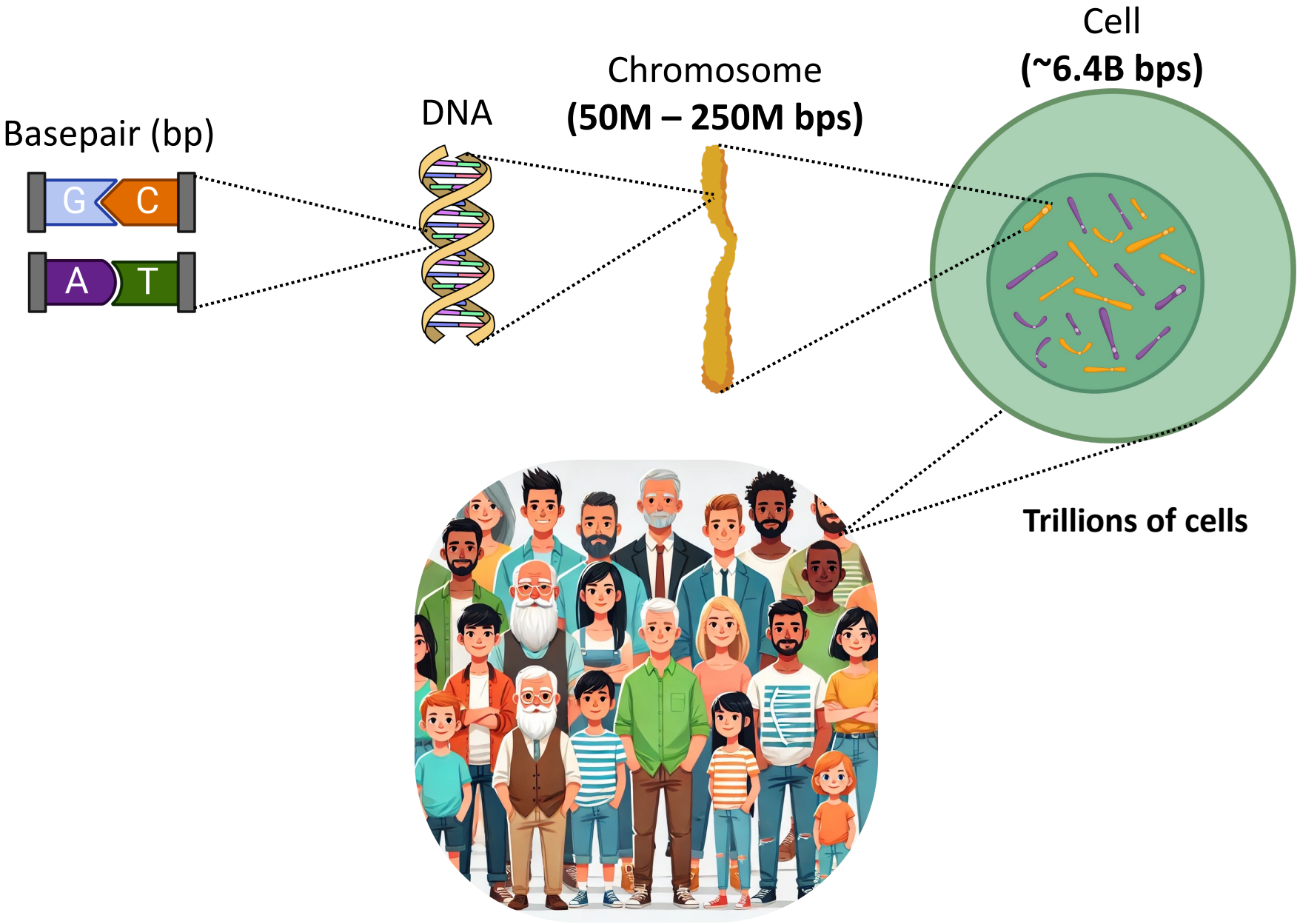
Genome Editing

Detecting **pathogens** in the environment

Rapid surveillance of **disease outbreaks**

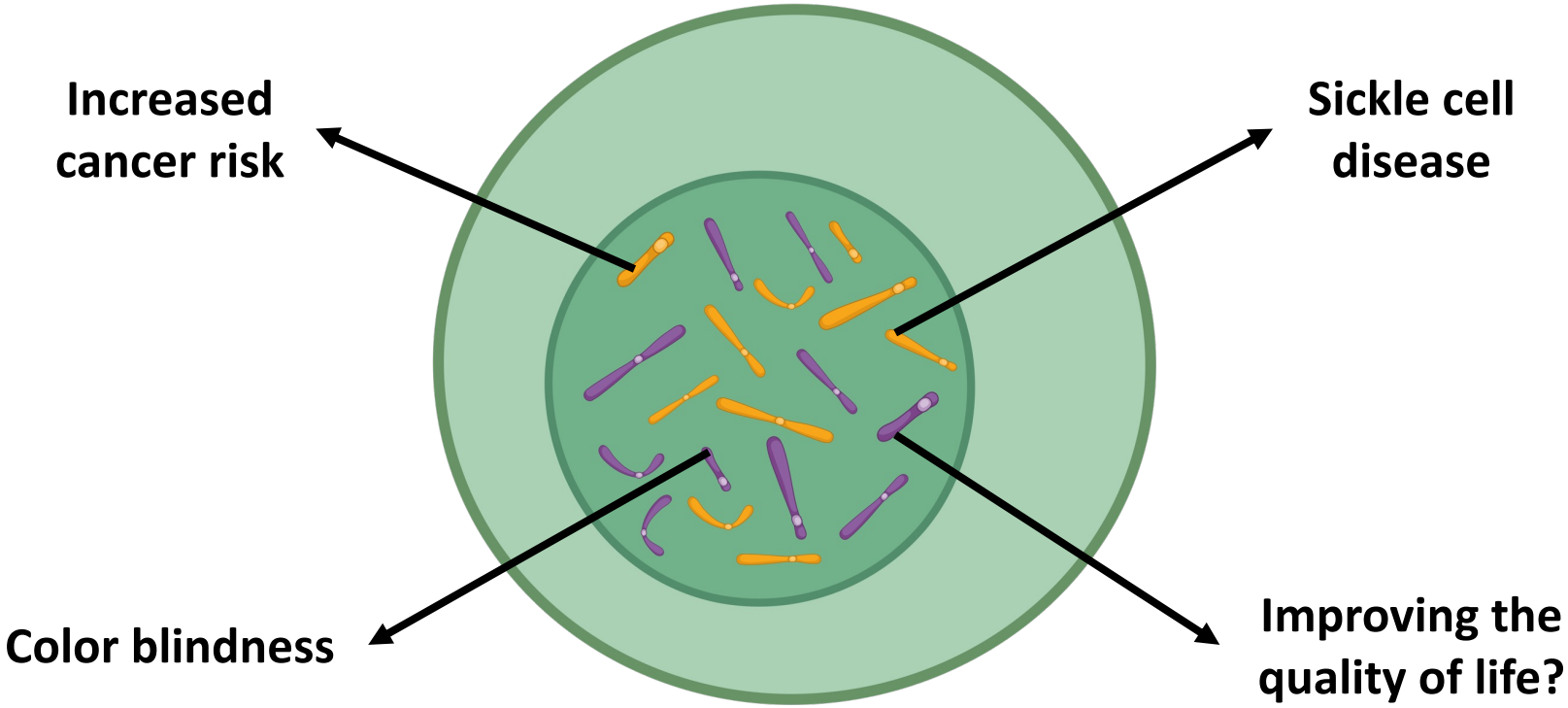
What is a Genome?

DNA – The Code of Life

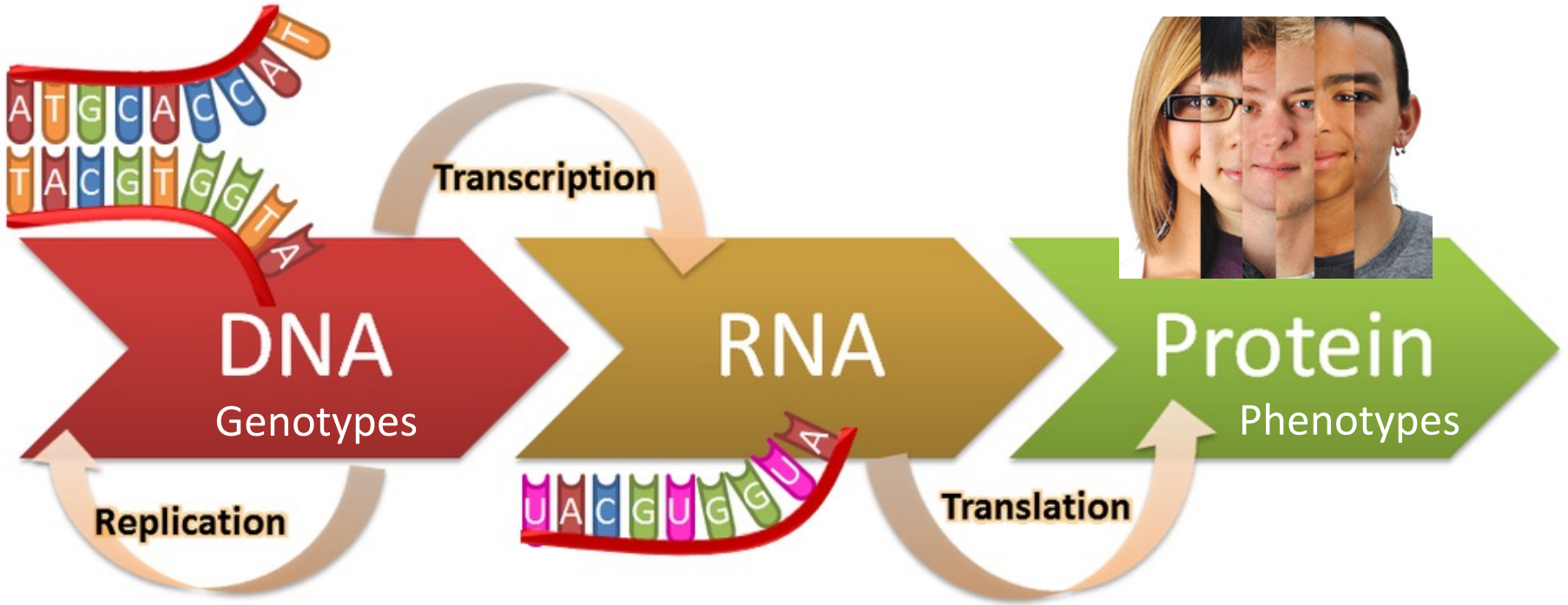


Whole Genome Analysis is Critical

Even a **single character** can have huge impact on health



The Central Dogma of Molecular Biology



How Large Is a Human Genome?

~170 meters

Washington Monument, DC

~200 meters

~6.4 billion bases*

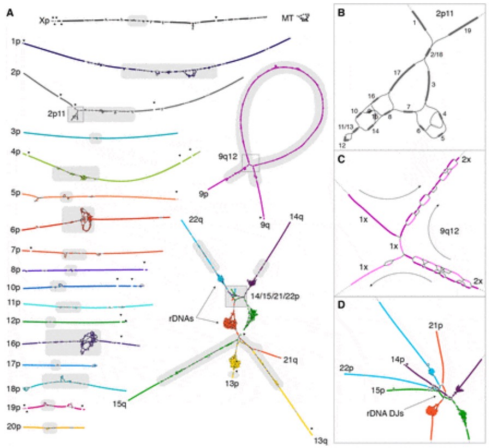
https://simple.wikipedia.org/wiki/Washington_Monument

*~200 meters = 80 characters per line, 40 lines per A4 page, each page is 0.1mm thick

Constructing the Human Reference Genome

2022: A gapless human reference genome

The complete sequence of a human genome



Abstract

In 2001, Celera Genomics and the International Human Genome Sequencing Consortium published their initial drafts of the human genome, which revolutionized the field of genomics. While these drafts and the updates that followed effectively covered the euchromatic fraction of the genome, the heterochromatin and many other complex regions were left unfinished or erroneous. Addressing this remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium has finished the first truly complete 3.055 billion base pair (bp) sequence of a human genome, representing the largest improvement to the human reference genome since its initial release. The new T2T-CHM13 reference includes gapless assemblies for all 22 autosomes plus Chromosome X, corrects numerous errors, and introduces nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding. The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies for the first time.

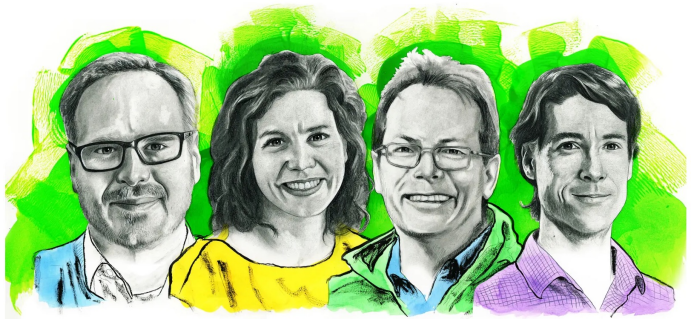
TIME

← THE 100 MOST INFLUENTIAL PEOPLE OF 2022

PRESENTED BY

Booking.com

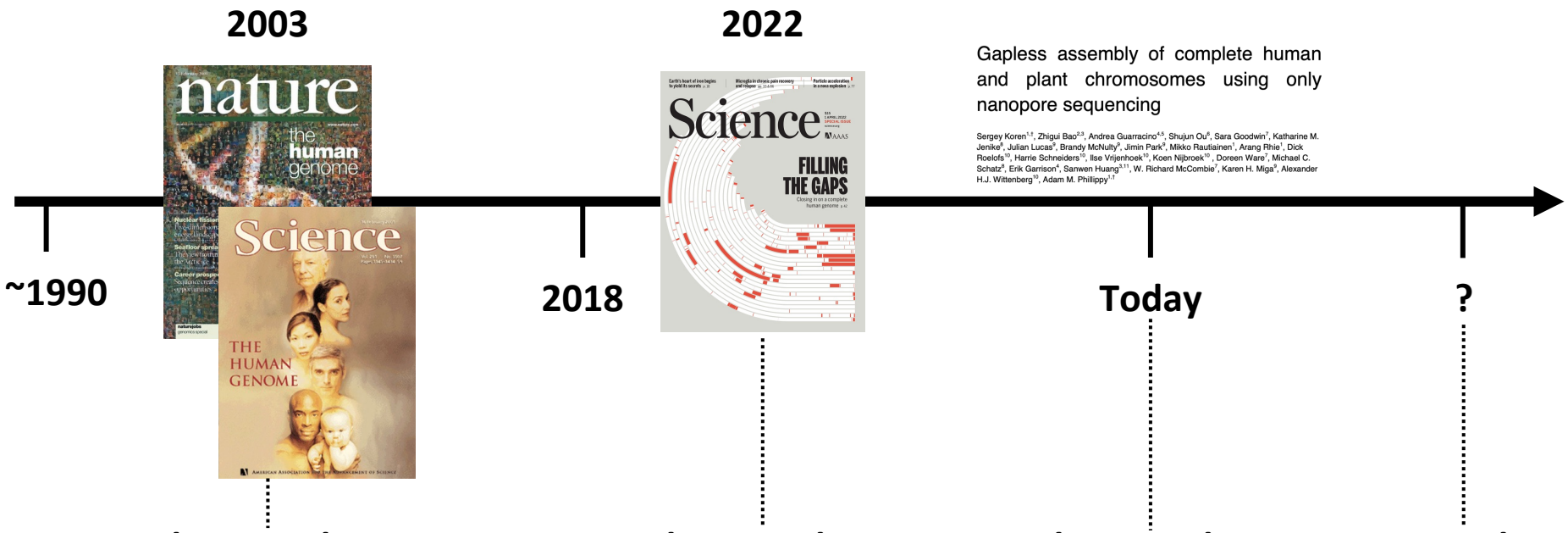
Michael Schatz, Karen Miga, Evan Eichler, and Adam Phillippy



https://genomics.ucsc.edu/2021/06/09/the-complete-sequence-of-a-human-genome/

https://time.com/collection/100-most-influential-people-2022/...

Humanity's Longest Puzzle Solving Challenge



The Complete Human Reference Genome

Significant resources, expertise, and time (yet incomplete)

The Complete Human Reference Genome (Really)

Significant resources and expertise

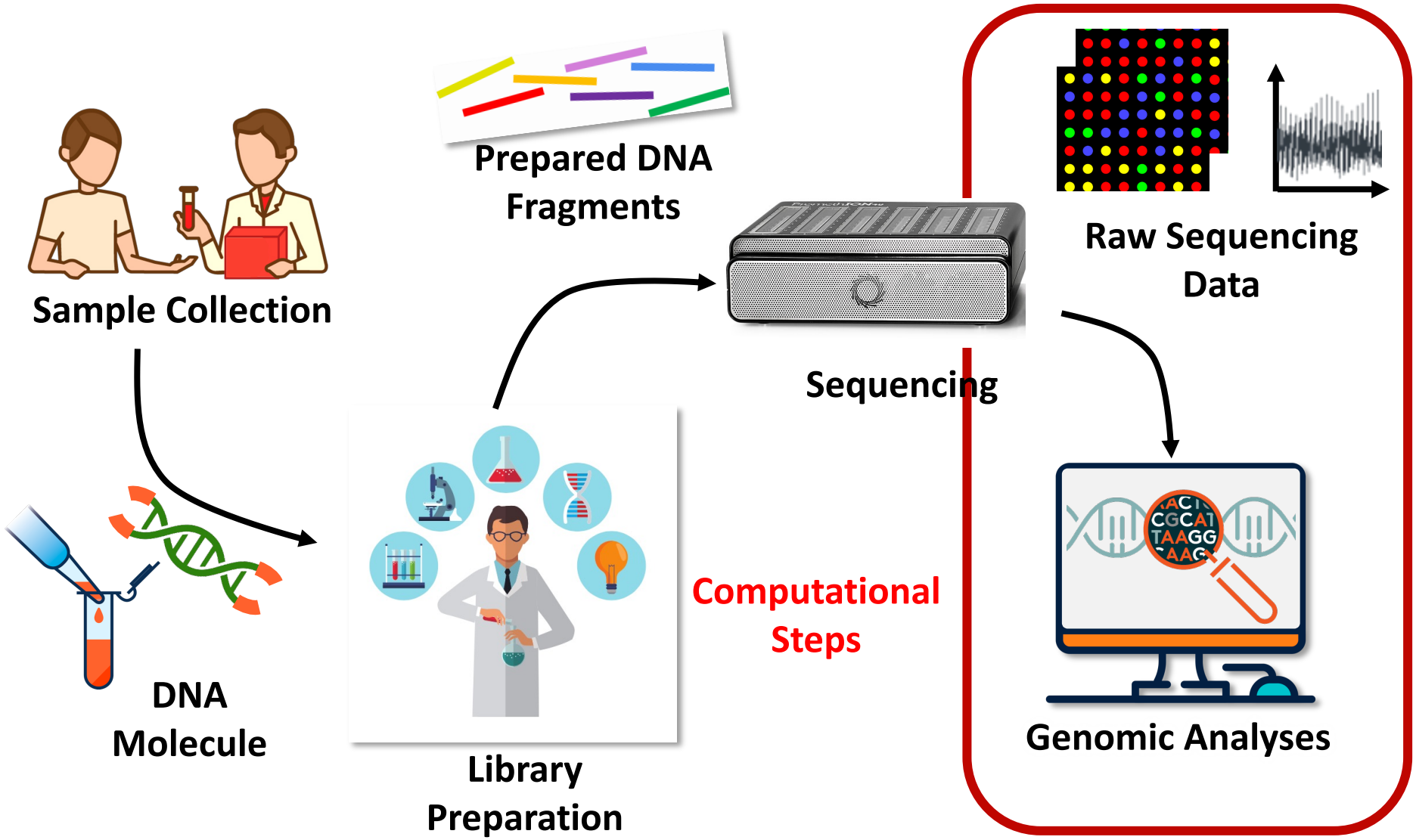
The Complete Human & Plant Chromosomes

Significant expertise

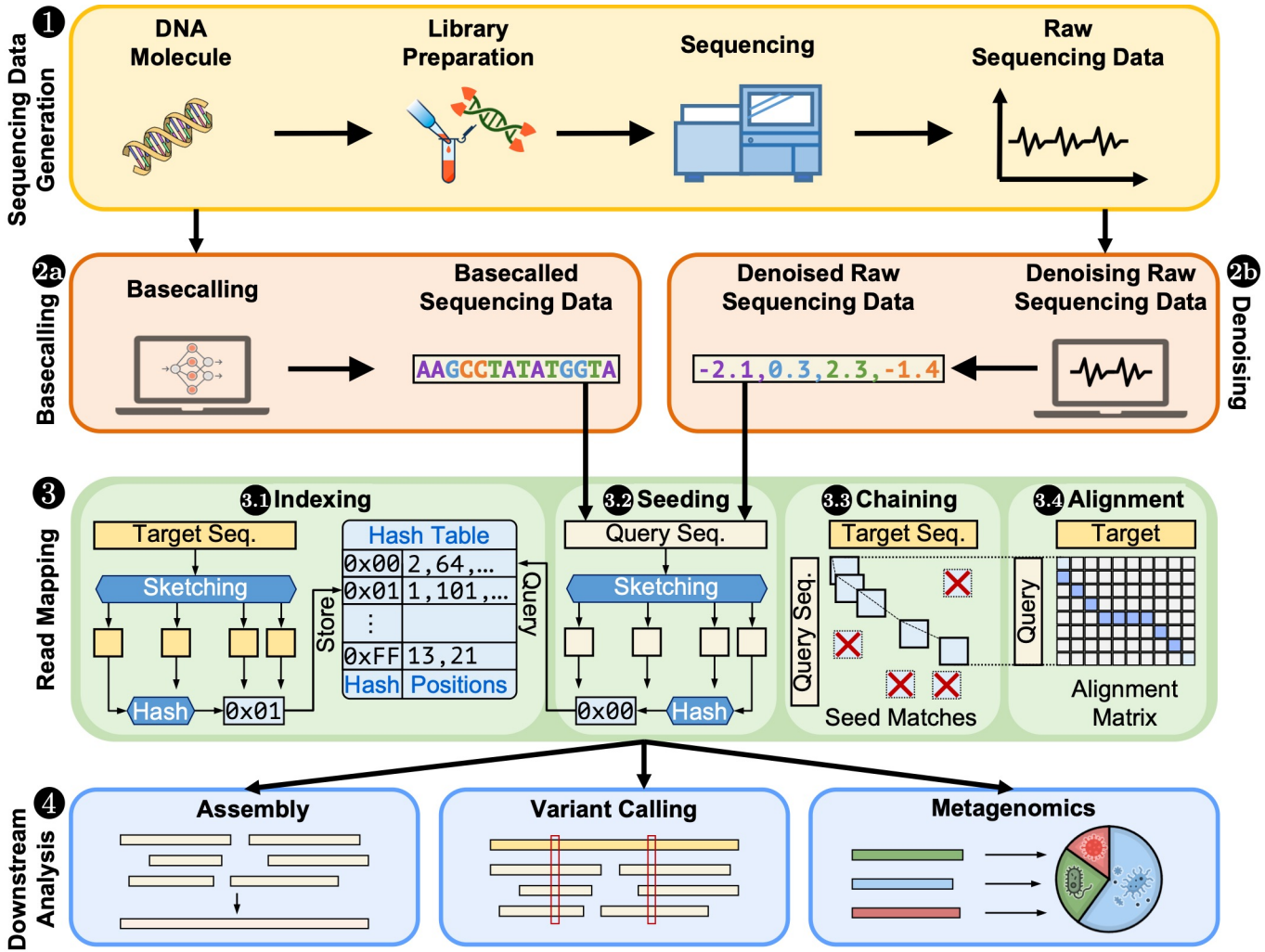
A Complete Human Genome

No expertise, Limited resources, quick, and accurate

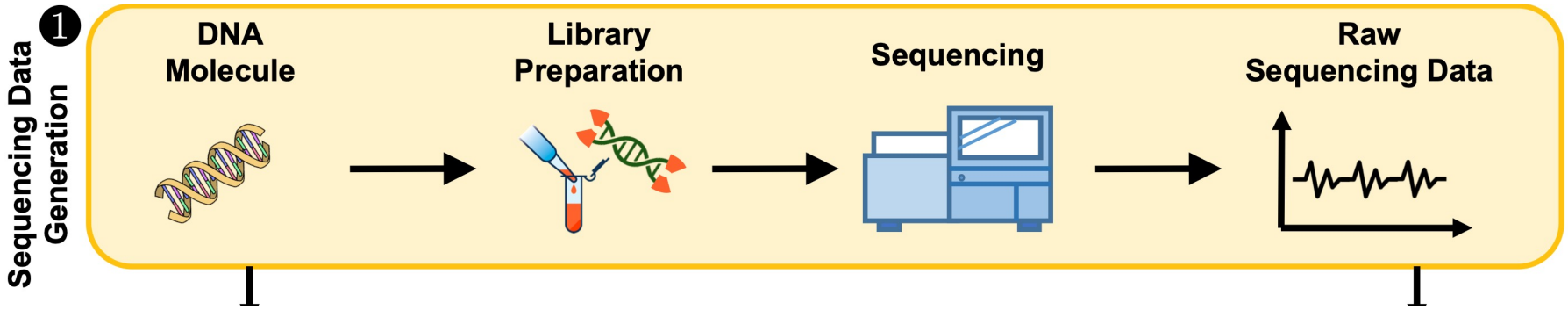
Genome Analysis in Real Life



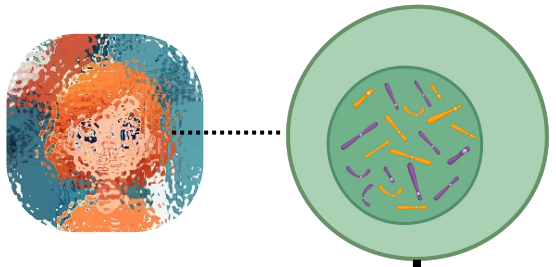
Genome Analysis Pipeline



Sequencing Data Generation

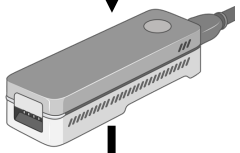


Recall: Generating the Entire Genomic Sequence



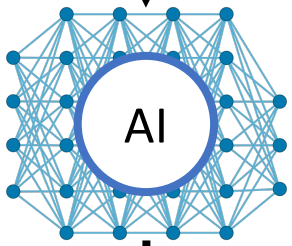
Whole Genome: 6.4B bps
(Non-human-readable)

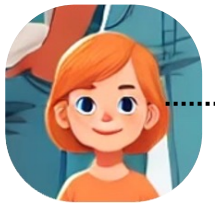
Sequencing Device



Raw Genomic Data
(Human-readable?)

Translator
(Basecalling)

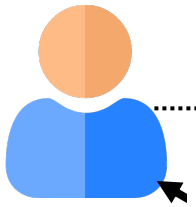




CGAAATGCC...AGATTAAACGCT...TGCCCTAA...GAATGGCGT

Whole Human Genome: 6.4B bps (Human-readable)

Genomic Data: Giant Jigsaw Puzzle to Solve



Human (Reference) Genome: 3.2 Billion bps (haploid)

CGAAATGCC...AGATTAAACGCT...TGCCCTAA...GAATGGCGT

?

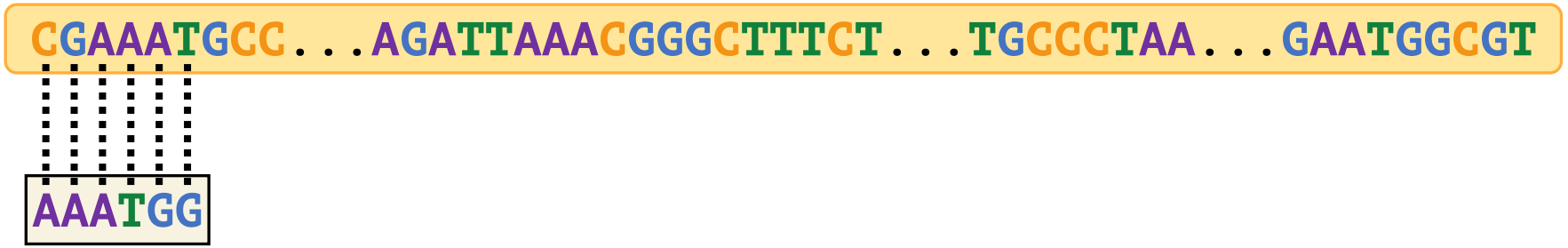
CGAAATGCC...AGATTAAATGGT...TGCCCTAA...GAATGGCGT

Entire Genomic Sequence of an Organism

- **Small pieces (reads)**
- **Unknown origins**
- **Non-identical reference genome**
- **Large volume of data to analyze**

Solving the Puzzle: The Naïve Way

Human (Reference) Genome: 3.2 Billion bps (haploid)



Pairwise Approximate String Matching (Alignment)
$O(n^2)$

→

Entire Genome
$O(n^2m)$

→

For All Reads
$O(n^2mk)$
$\cong O(10^{20})$

≅Years

We will explain the practical solutions

Sequencing Technologies

Oxford Nanopore Technologies (ONT)

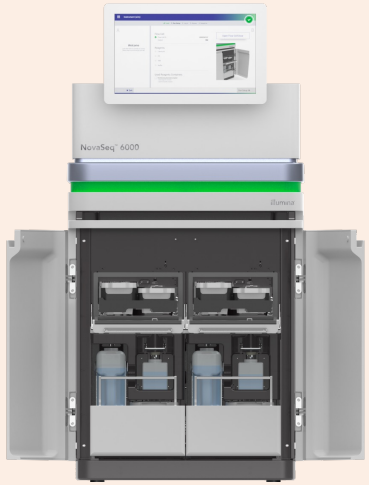
PacBio (HiFi, CLR)

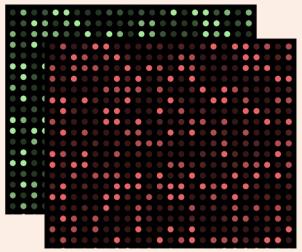
Illumina

… and more! All produce data with different properties.

Different Raw Sequencing Data

Illumina (SBS)

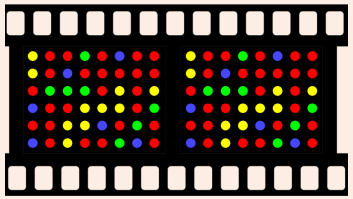




Multiple images

.BCL/.CBCL

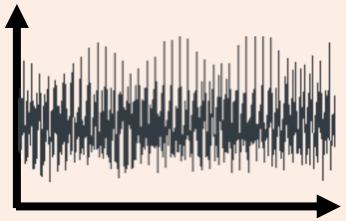
PacBio (SMRT)



Fluorescence pulses

.BAM

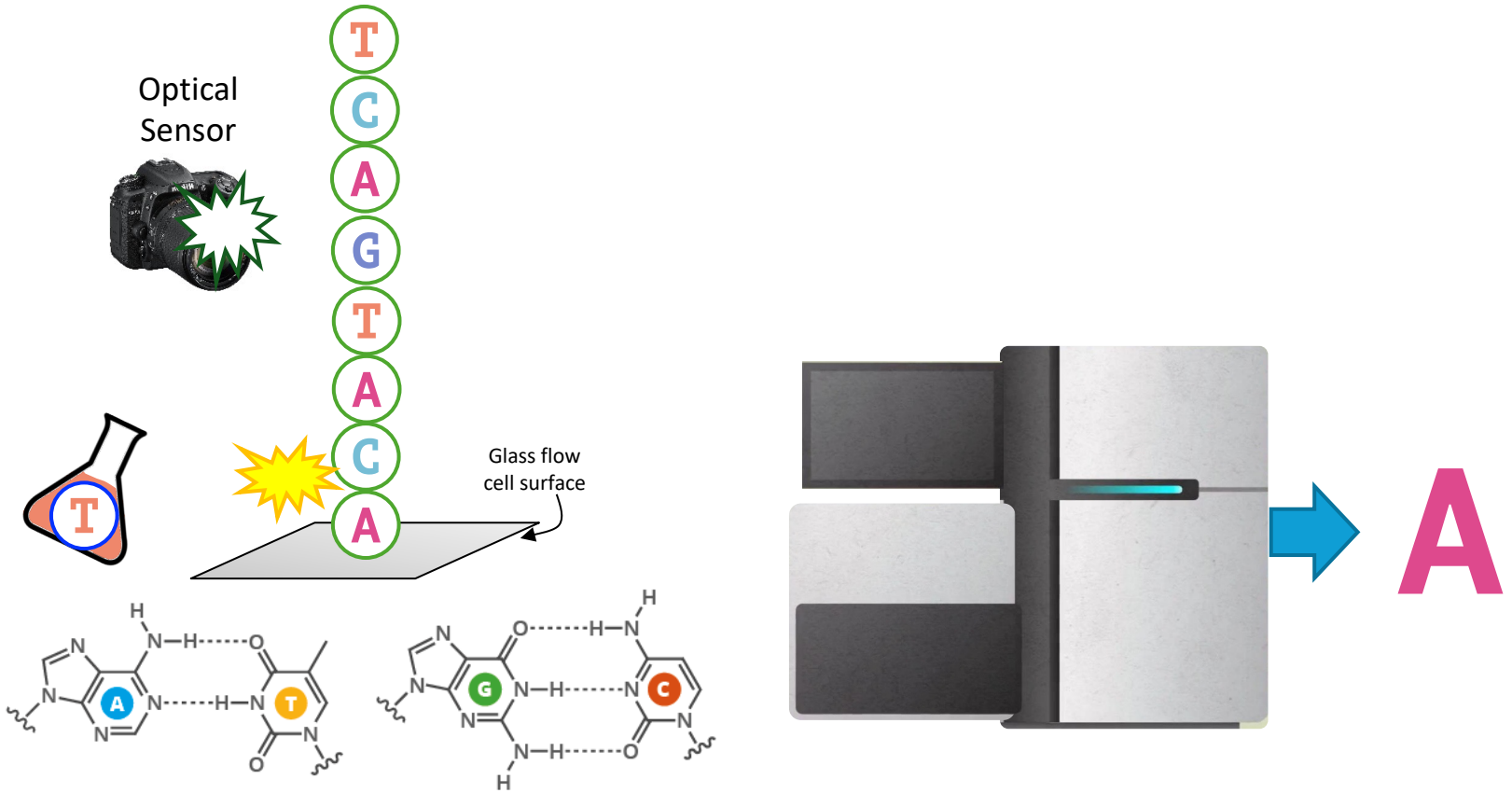
Nanopore



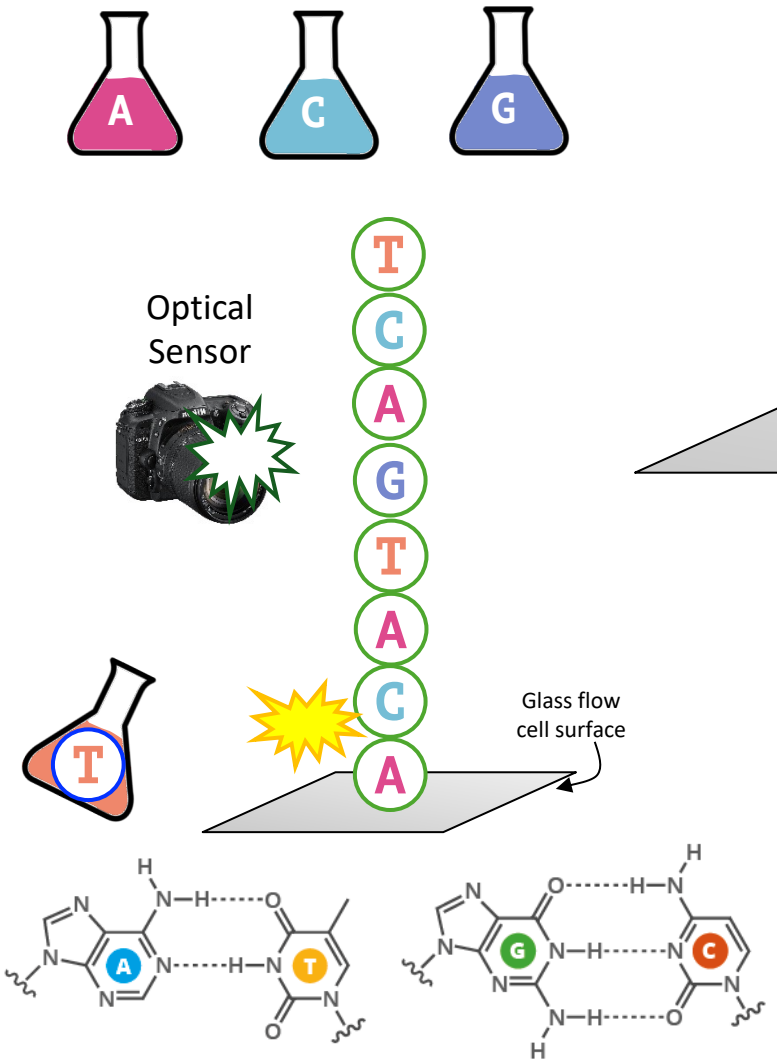
Electrical Signals

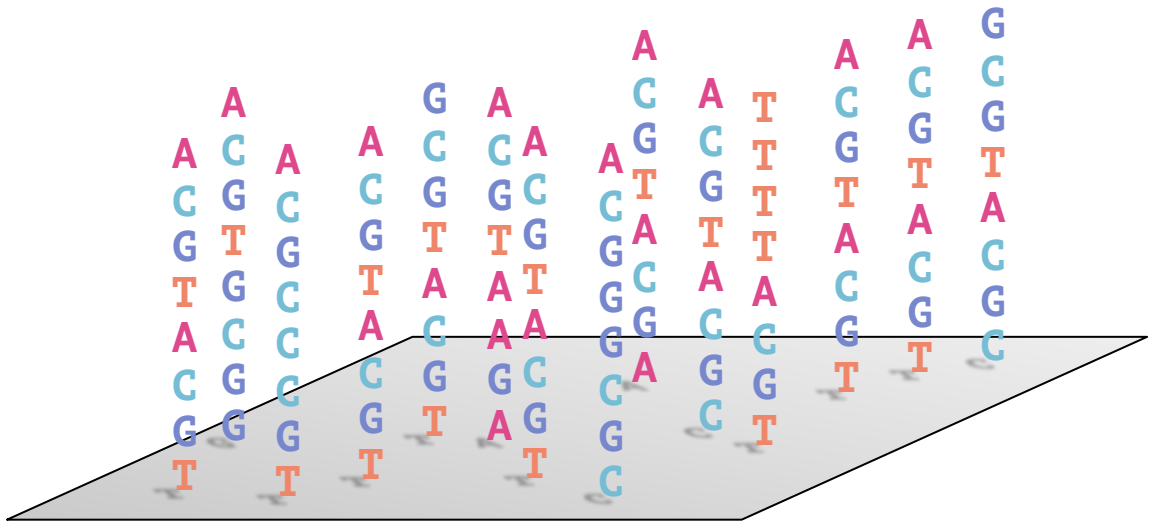
.POD5

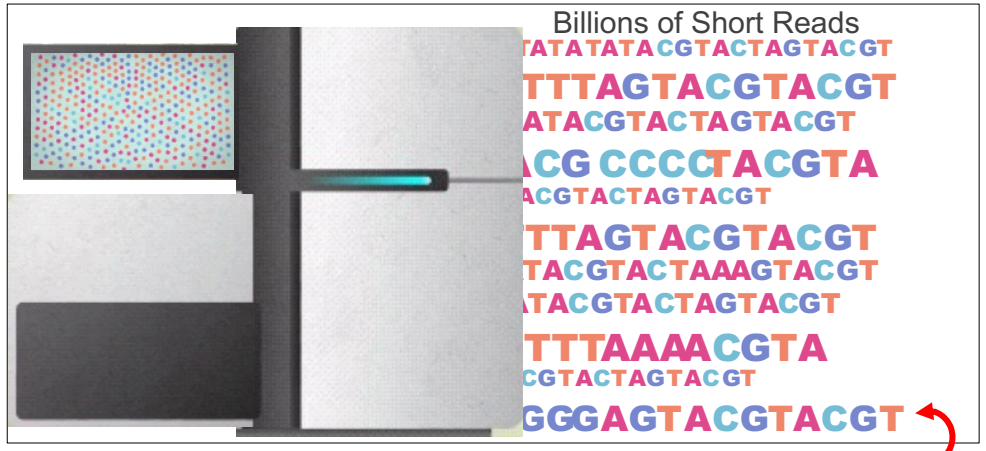
Sequencing using the Illumina Machines



Sequencing using the Illumina Machines



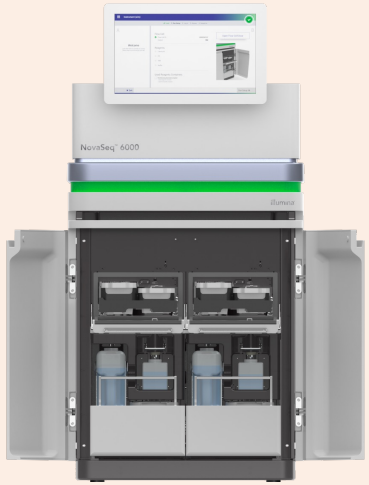


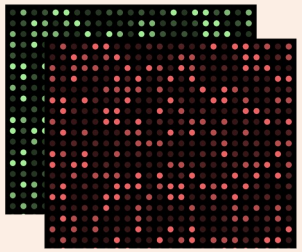


Sequenced bases form a read

Different Raw Sequencing Data

Illumina (SBS)

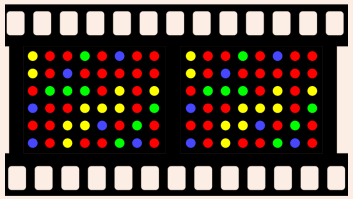




Multiple images

.BCL/.CBCL

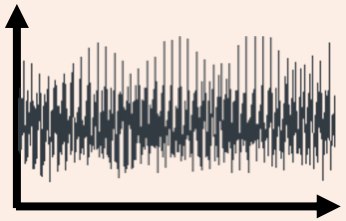
PacBio (SMRT)



Fluorescence pulses

.BAM

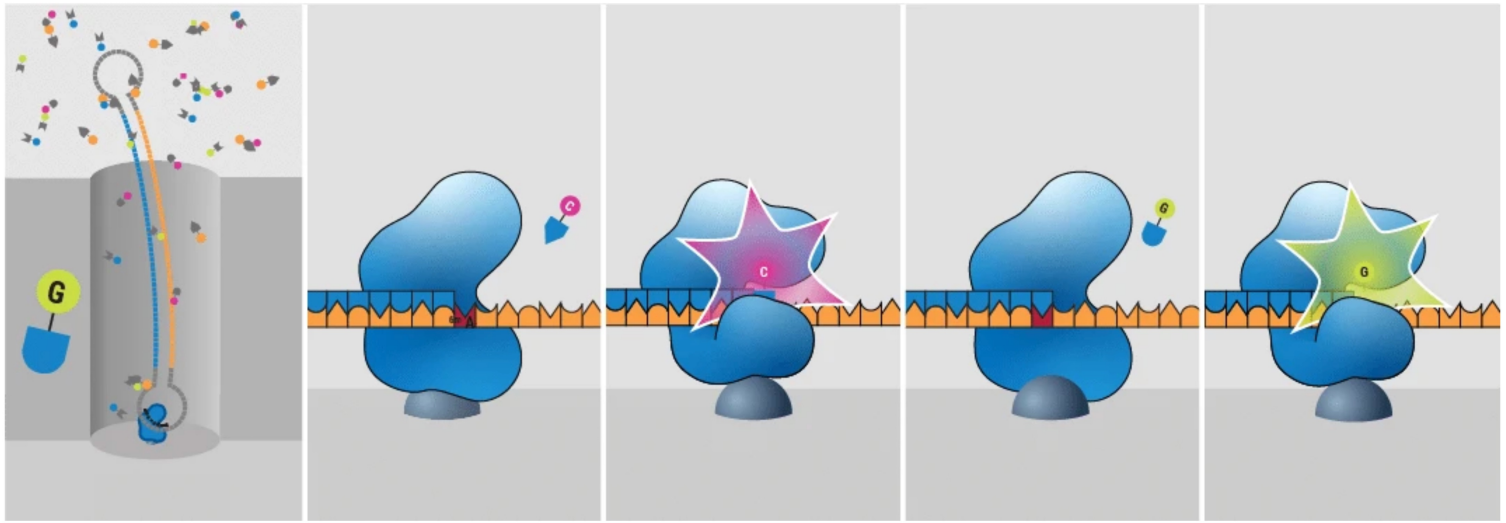
Nanopore

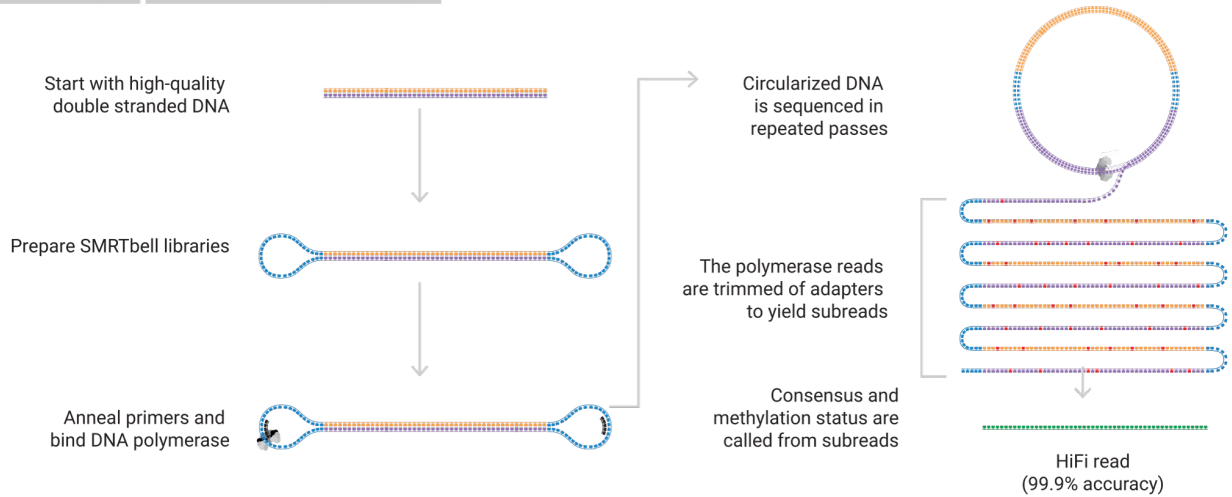


Electrical Signals

.POD5

PacBio HiFi Sequencing

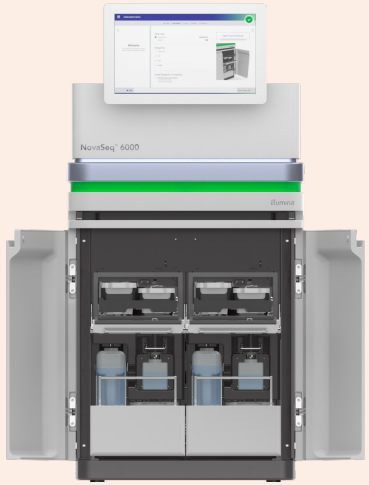


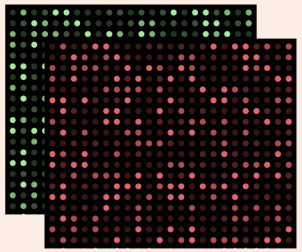


Repeated passes over the same DNA insert produce one HiFi read

Different Raw Sequencing Data

Illumina (SBS)

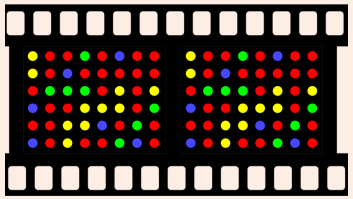




Multiple images

.BCL/.CBCL

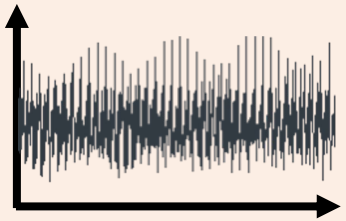
PacBio (SMRT)



Fluorescence pulses

.BAM

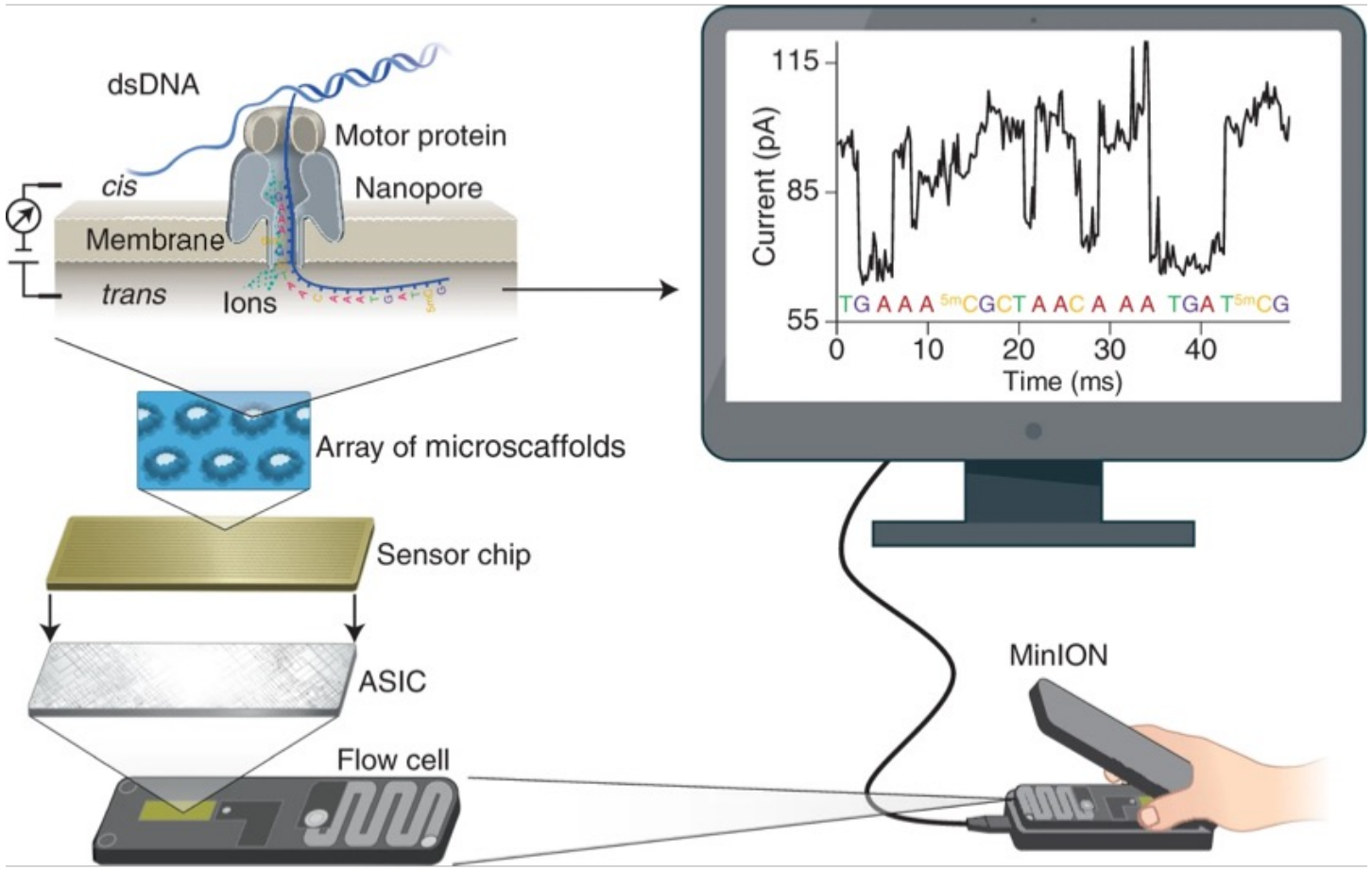
Nanopore



Electrical Signals

.POD5

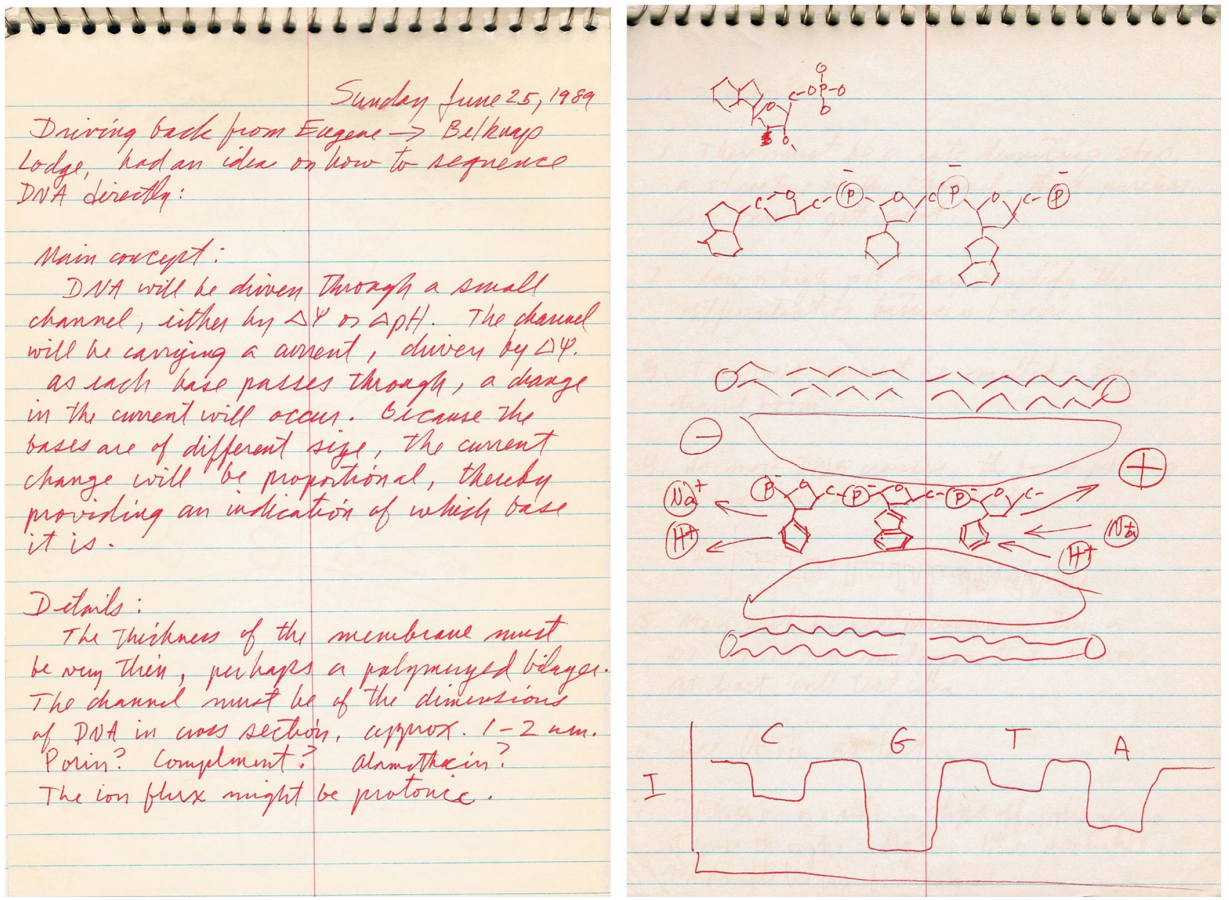
Nanopore Sequencing



Wang+, "Nanopore sequencing technology, bioinformatics and applications," *Nature Biotechnology*, 2021.

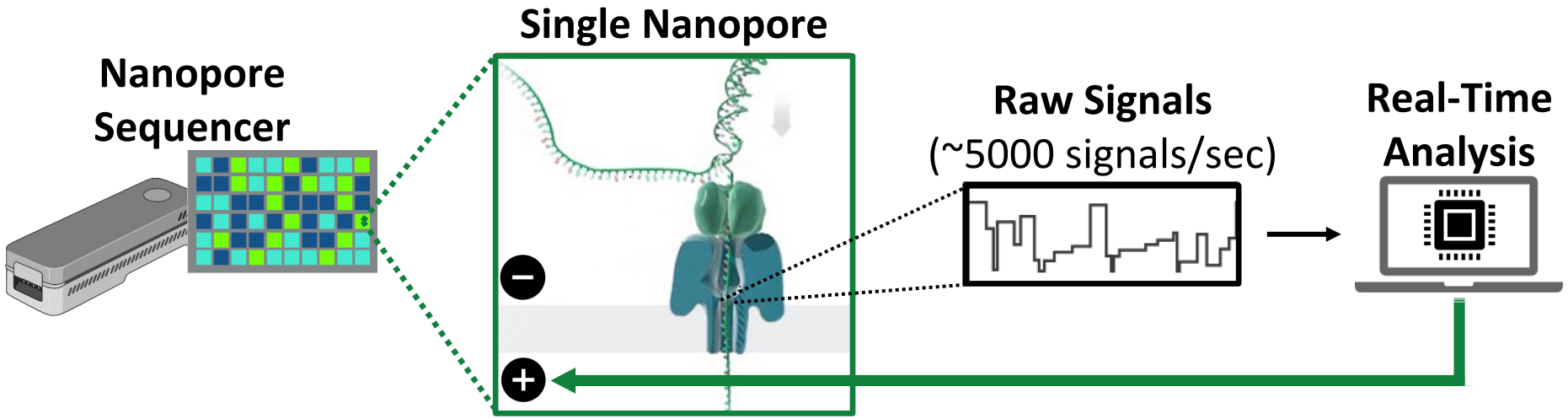
Nanopore Sequencing: The Original Idea

David Deamer's handwriting on nanopore sequencing (June 25, 1989)



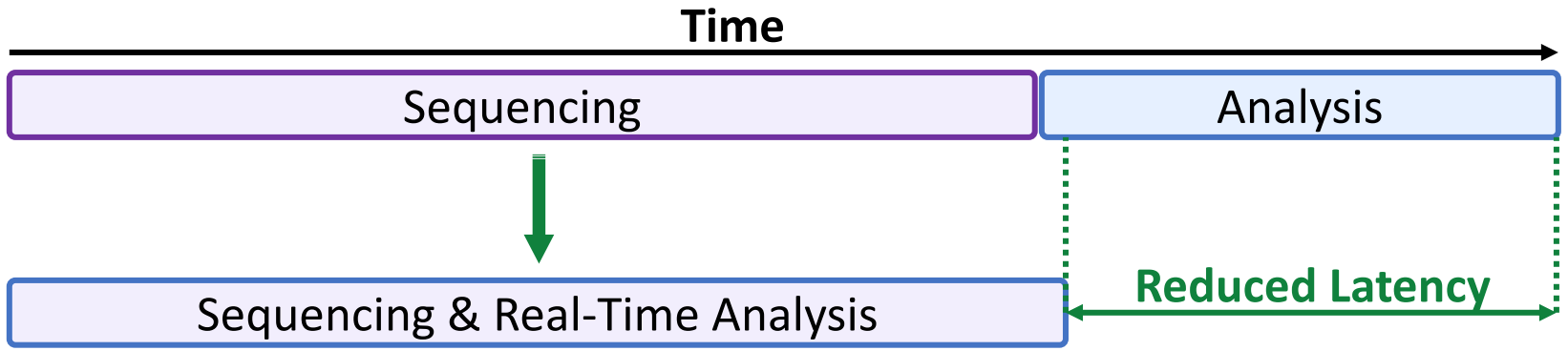
Deamer+, "Three decades of nanopore sequencing," *Nature Biotechnology*, 2016.

Recall: Generating Electrical Signals from DNA

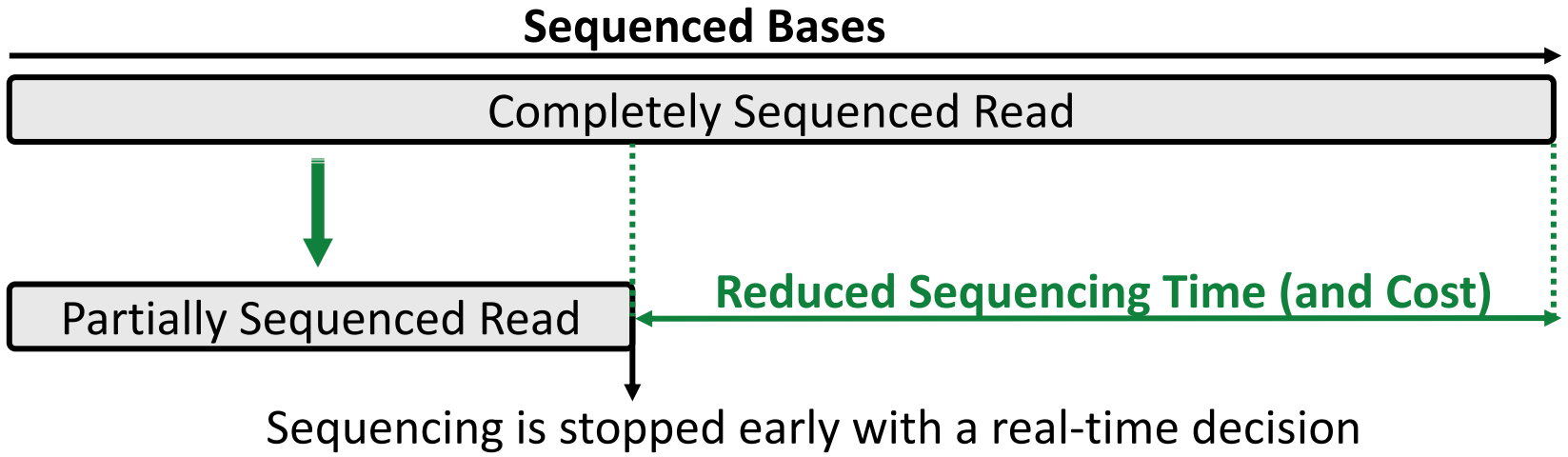


Recall: Benefits of Real-Time Analysis

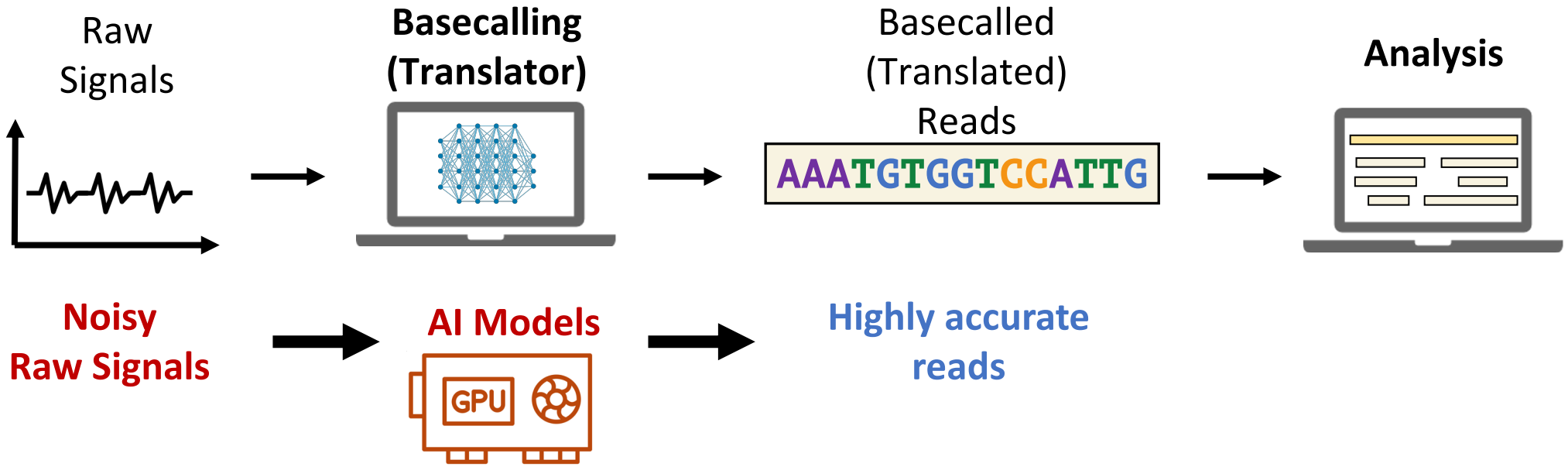
Reducing latency by overlapping analysis with sequencing



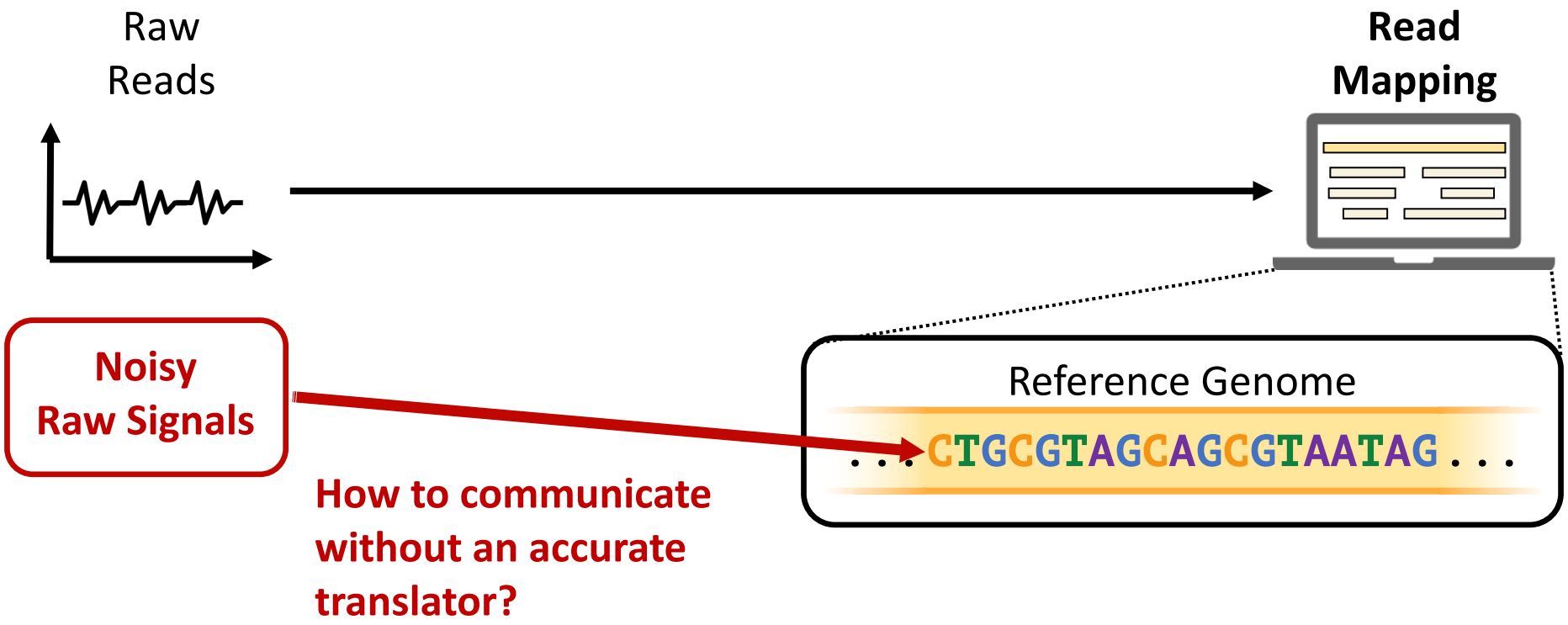
Reducing sequencing time and cost by stopping sequencing early



Recall: A Common Genome Analysis Pipeline

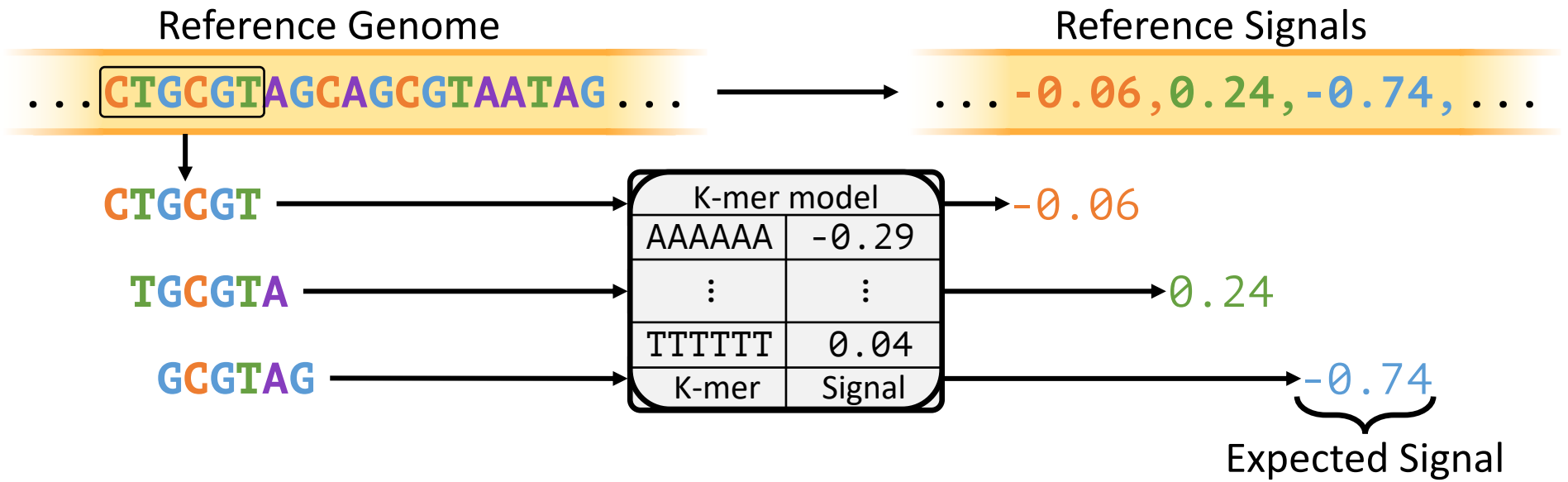


Can We Survive Without Translation?



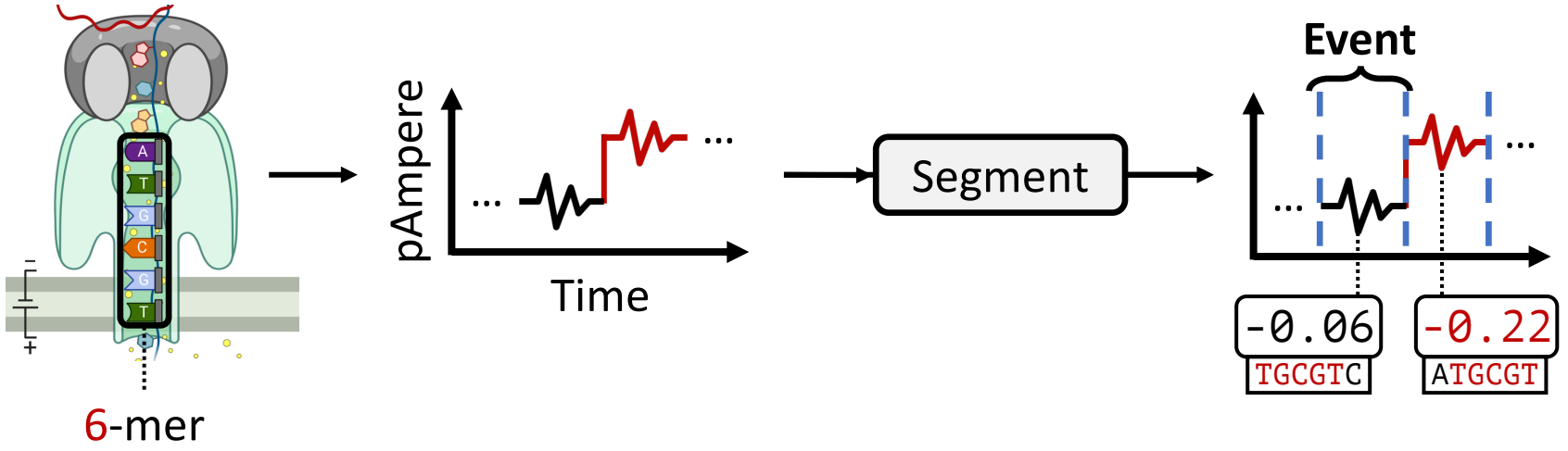
Reference-to-Signal Conversion

- **Goal:** Convert the reference sequence into expected current levels.
- **K-mer model:** A lookup table of expected signals for all k-mers.
- Slide the k-mer window along the reference to synthesize a signal.

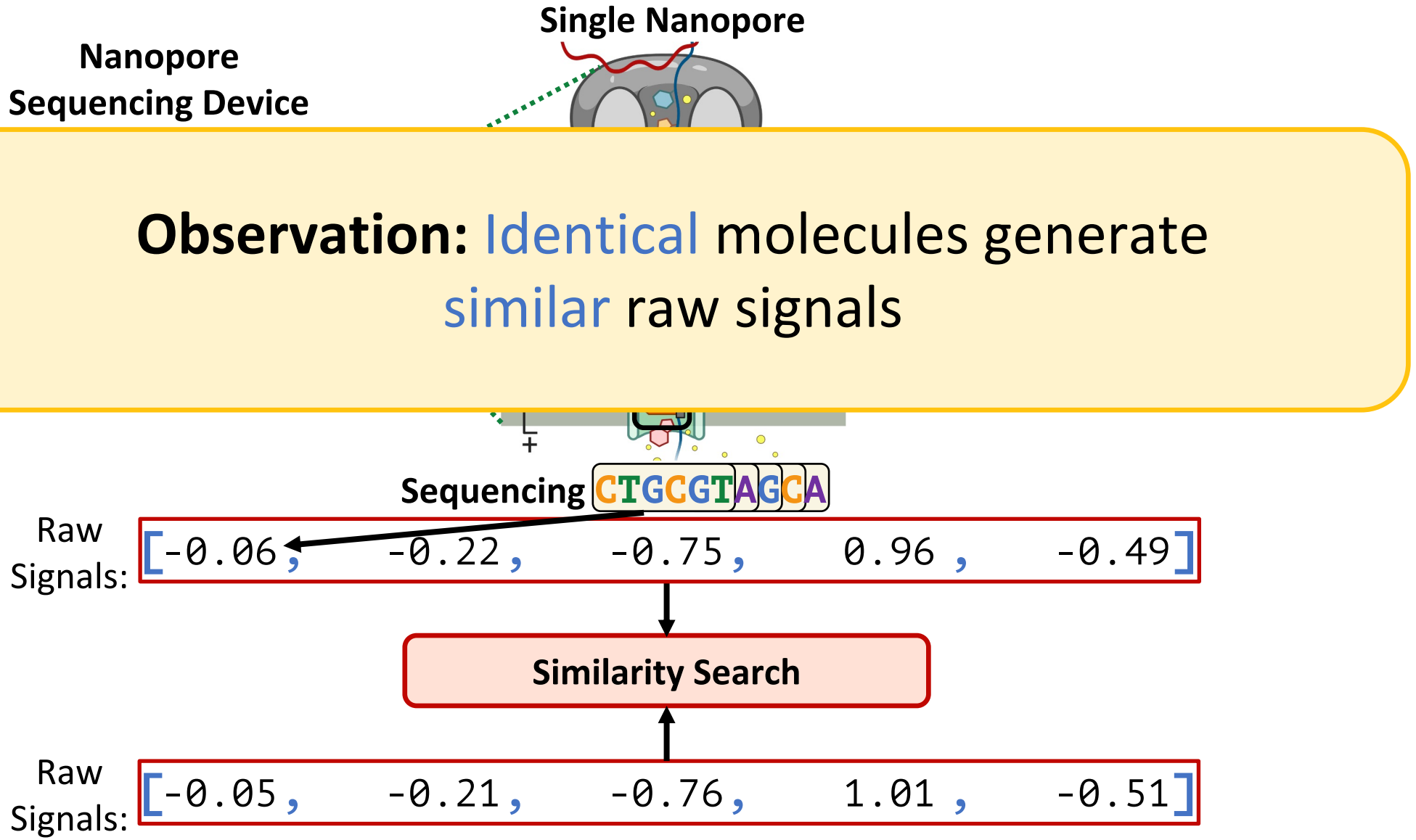


Processing Raw Nanopore Signals

- K bases influence the signal at a time.
- **Observation:** Signal levels change as DNA moves through the pore.
- **Goal:** Segment the trace into events that reflect local sequence.



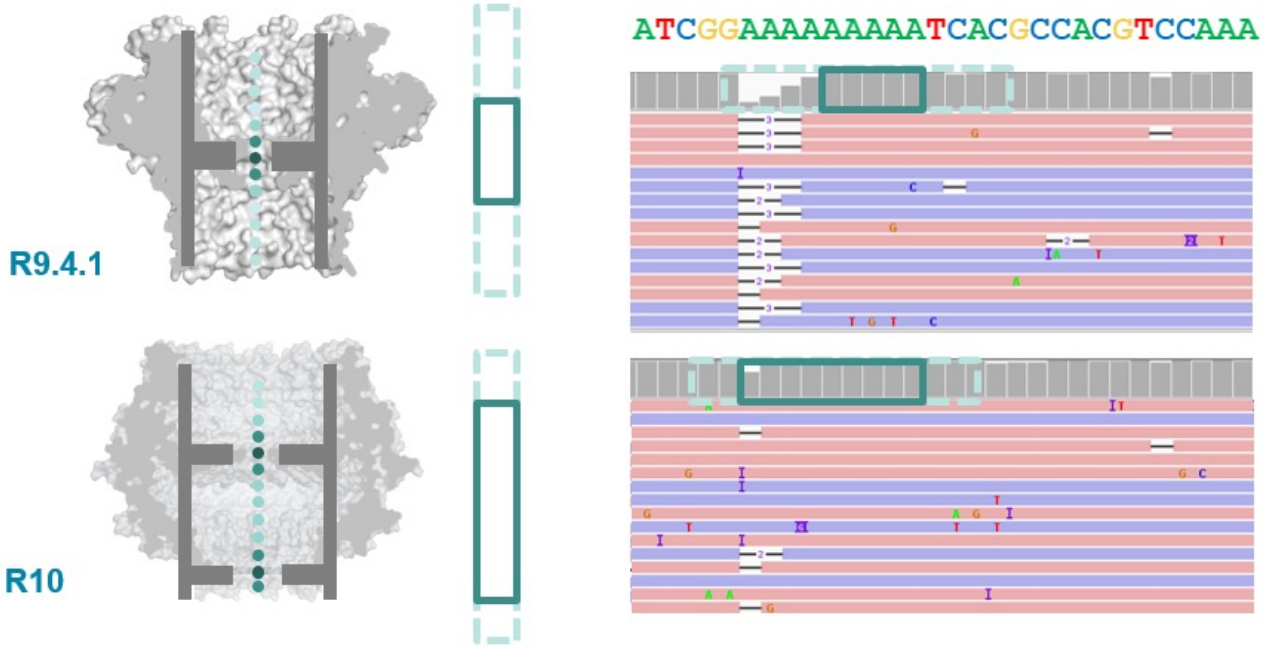
Recall: Noise in Raw Electrical Signals



Source of Noise in Nanopore Sequencing

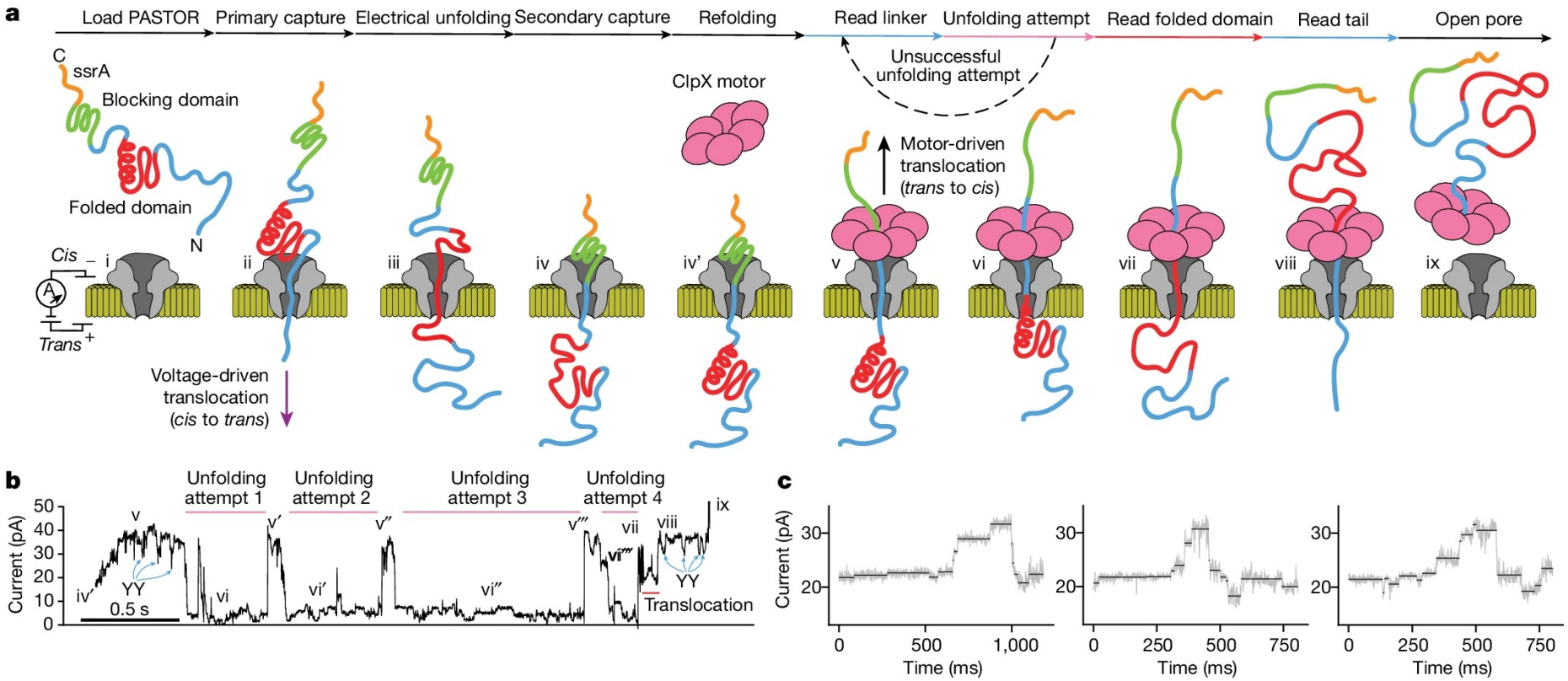
- **Stochastic thermal fluctuations in the ionic current**
  - Random ionic movement due to inherent thermal energy (Brownian motion)
- **Variations in the translocation speed**
  - Mainly due to the motor protein
- **Environmental factors**
  - **Temperature:** Affecting enzymes including the motor protein
  - **pH levels:** Affecting charge and the shape of molecules
- **Maybe:** Aging & material-related noise between nanopores

Various Nanopore Chemistries



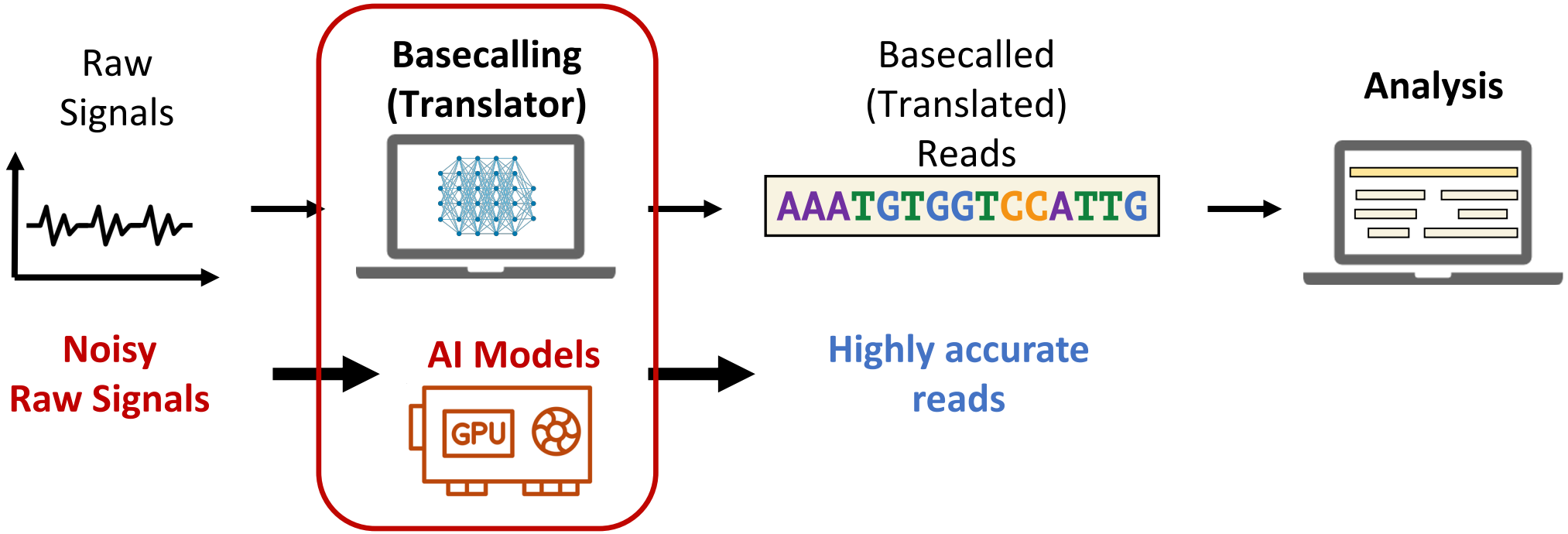
- Dual reader head in R10
- Motor protein with more consistent translocation speed in R10

Proteomics with Nanopores

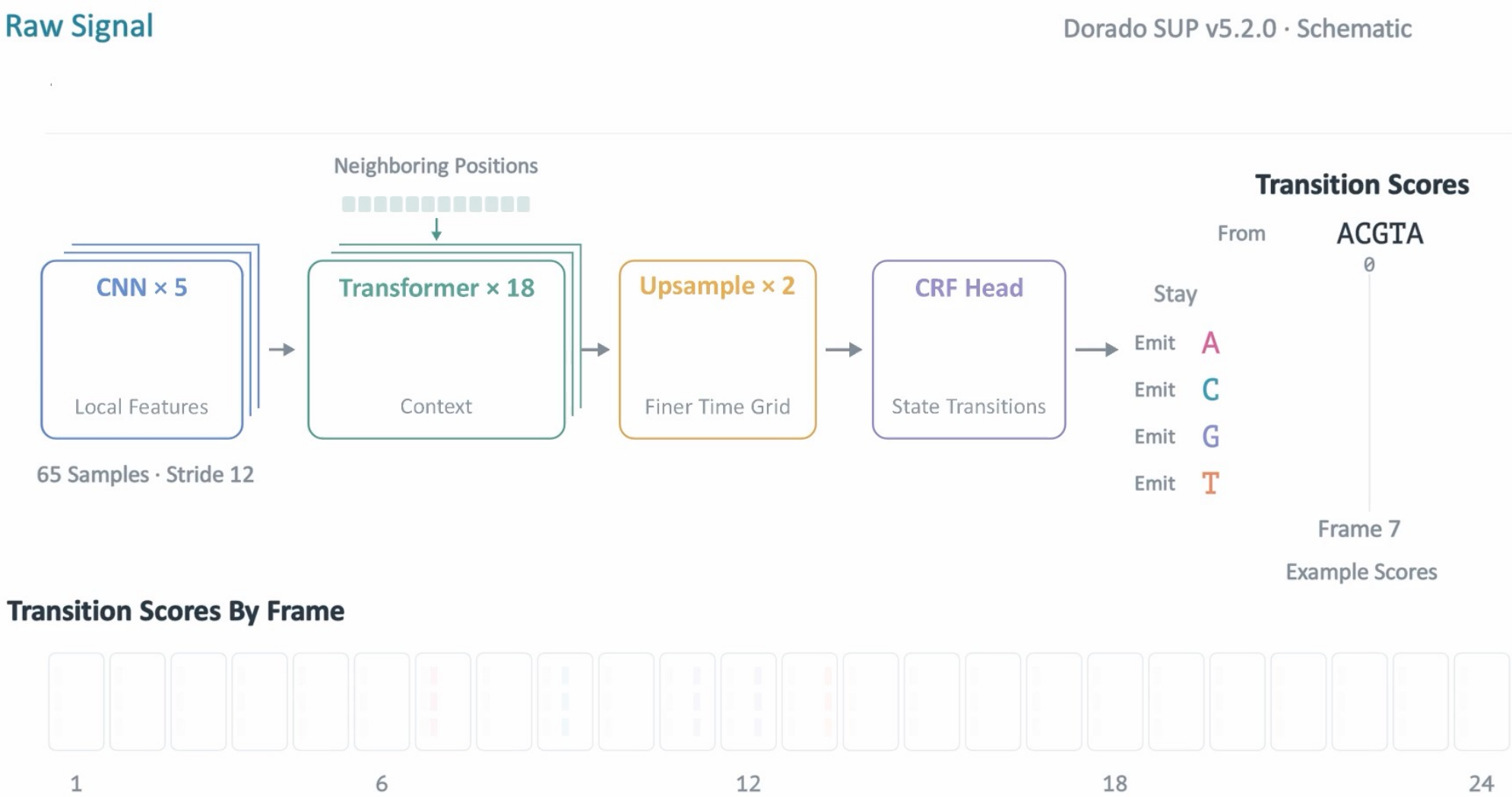


Motone+, "Multi-pass, single-molecule nanopore reading of long protein strands", Nature 2024.

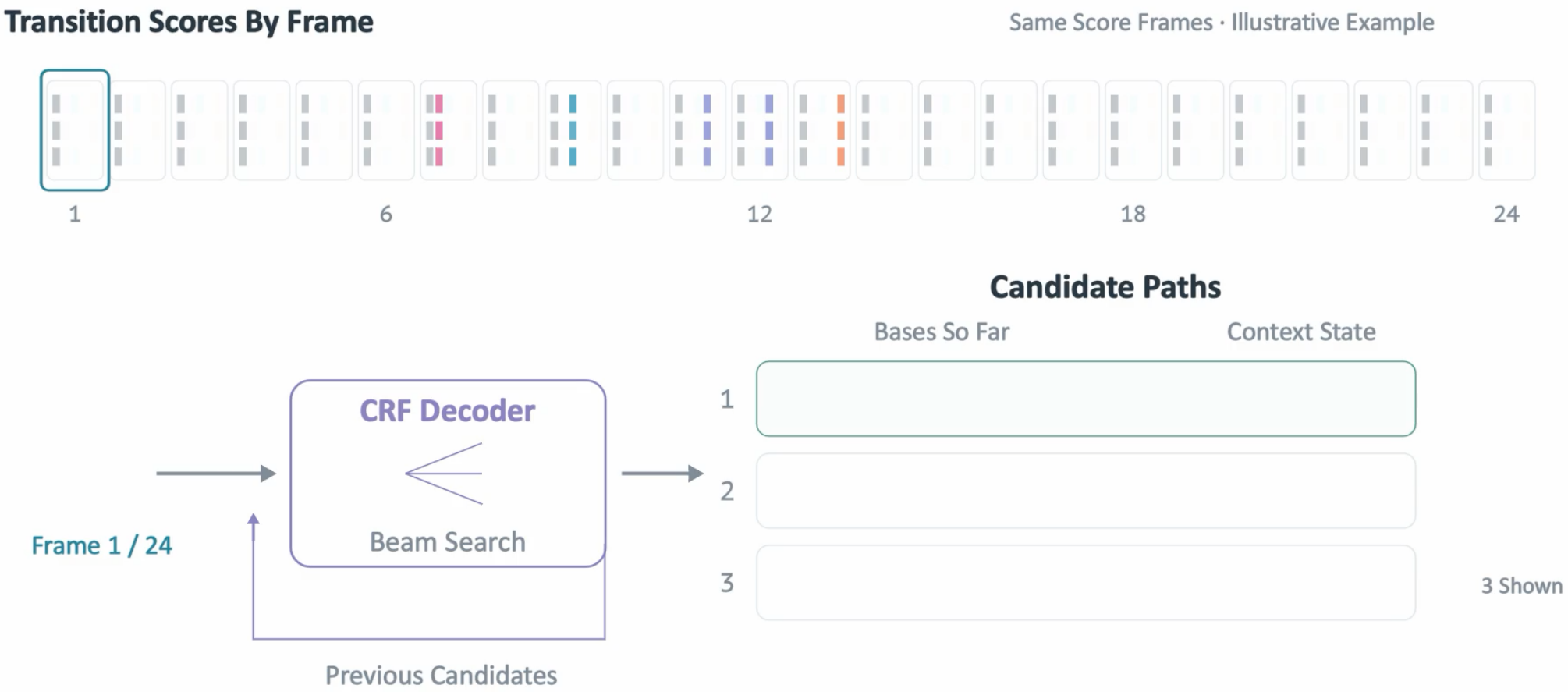
Recall: A Common Genome Analysis Pipeline



Basecalling Overview

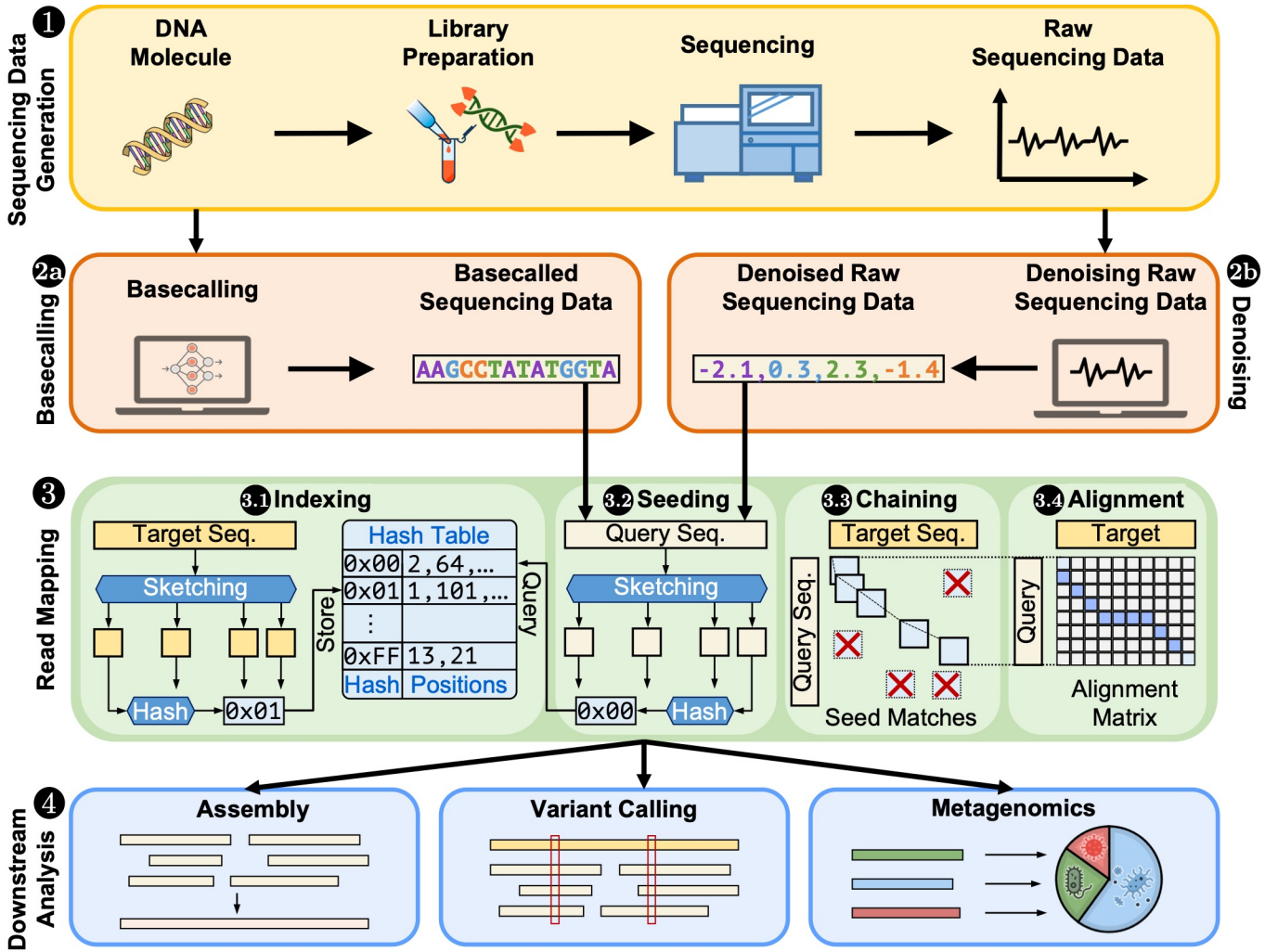


CTC-CRF Decoding



Questions?

Genome Analysis Pipeline



Next Lecture

- **Reminder:** Finalize your paper selection by today 11:59pm!

CMSC 829C: Algorithms, AI, and Hardware for Biological Data

Lecture 3: Introduction to Genomics and Sequencing

Fall 2026

September 8, 2026

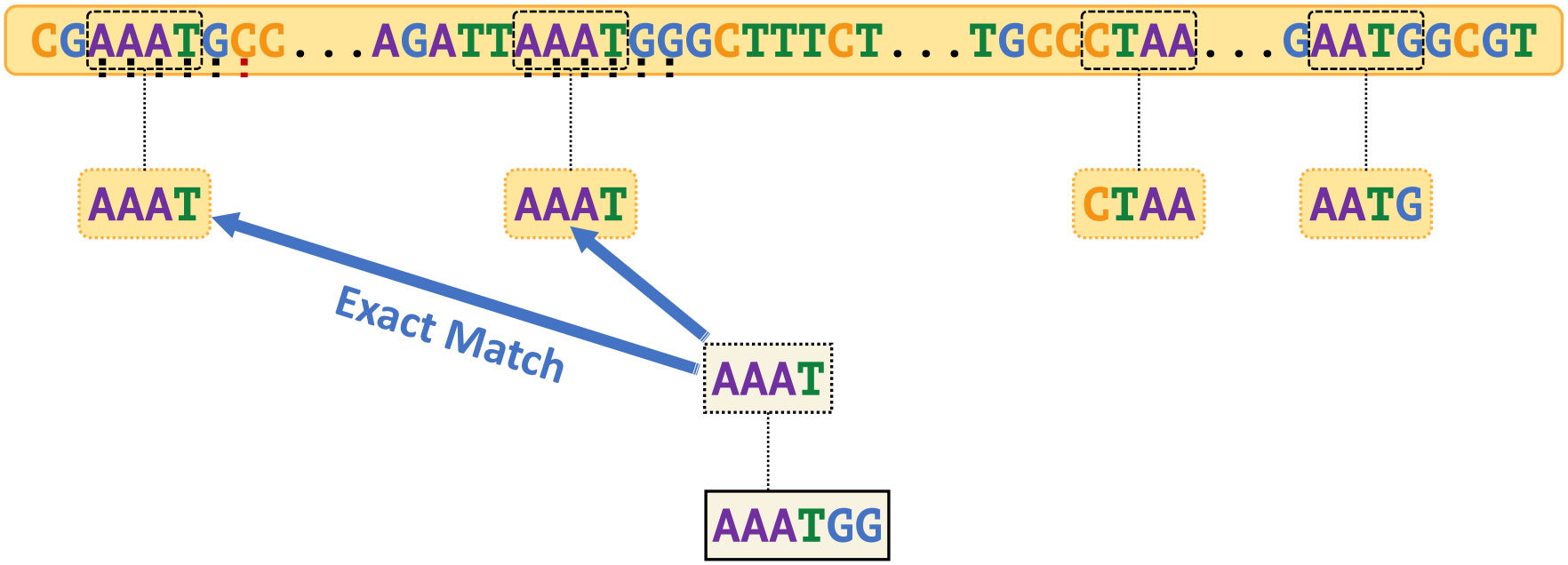
Can Firtina | canfirtina@gmail.com | cs.umd.edu/~firtina

Assistant Professor of Computer Science at UMD

UNIVERSITY OF MARYLAND

Solving the Puzzle: The Practical Way

Goal: Quickly and accurately reduce the search space



Read Mapping