When AI meets life sciences
当人工智能遇上生命科学

2025 清华大学 AI创新实践课程

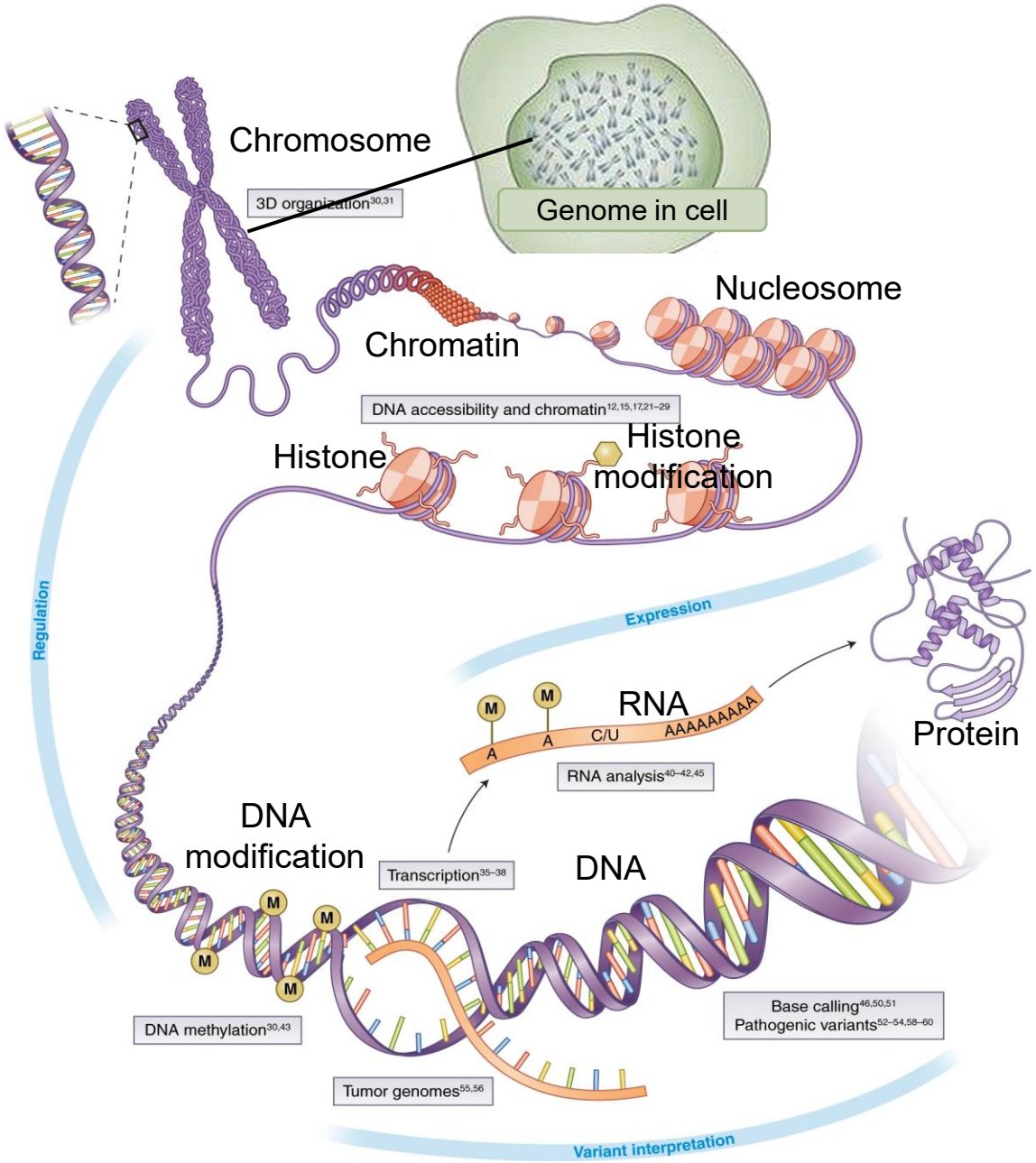
张强锋，清华大学

Nov 17, 2025

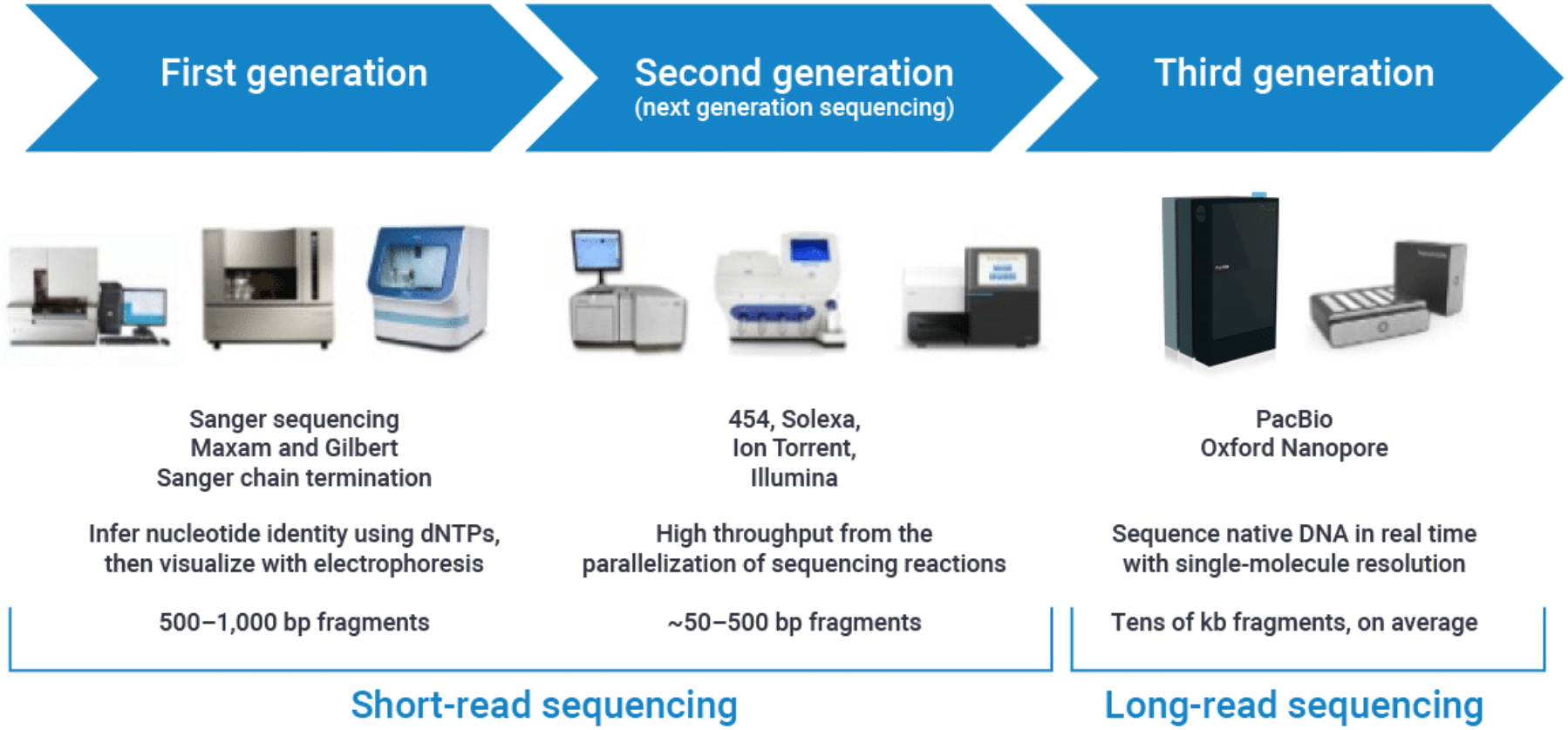
Outline

- Genome and genomics sequencing
- Genome assembly and annotation
- Comparative genomics and evolution
    - Variants calling and mutation detection

Genome

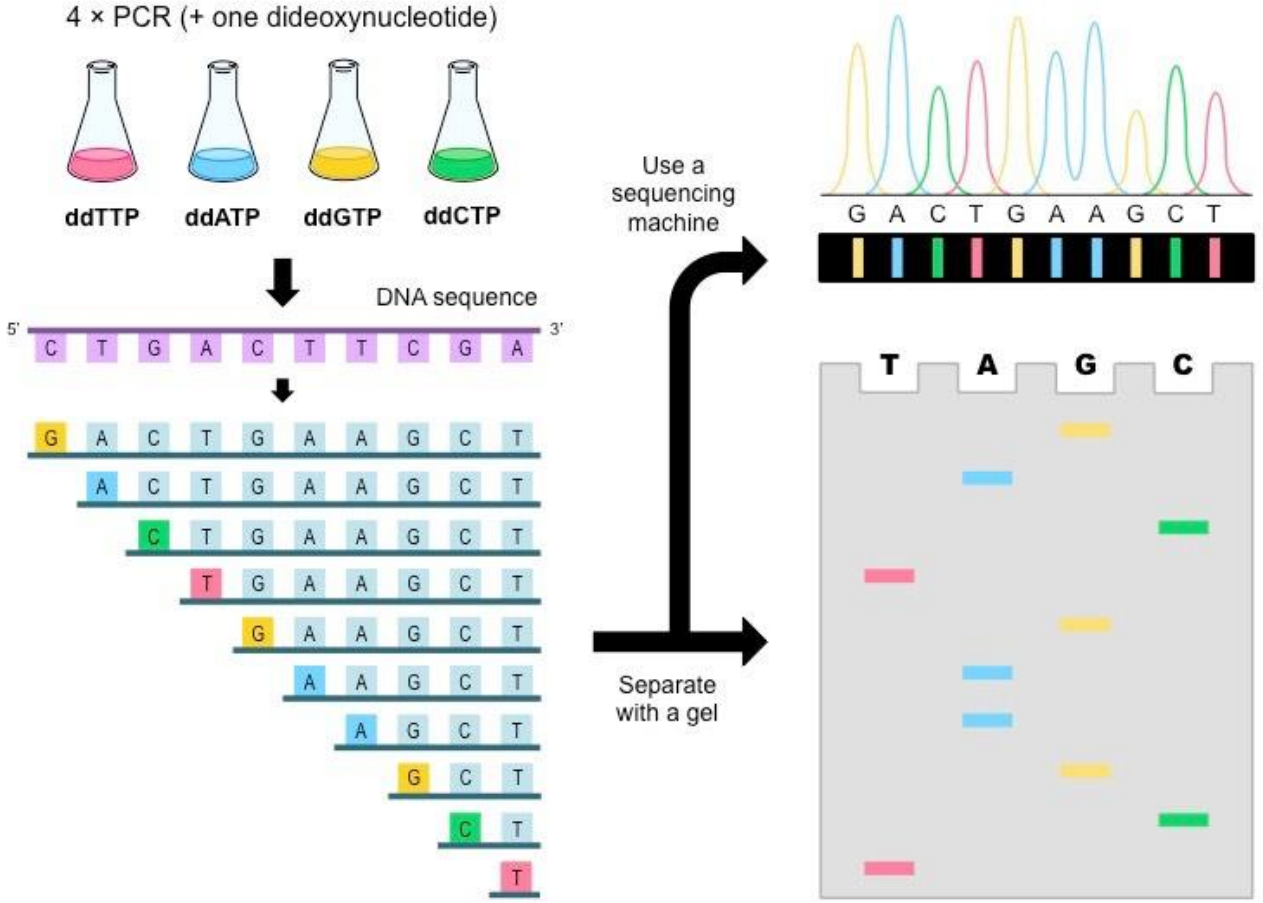

Chromosome
3D organization[30,31]
Genome in cell
Nucleosome
Chromatin
DNA accessibility and chromatin[12,15,17,21–29]
Histone
Histone modification
Regulation
Expression
RNA
M
M
A
A
C/U
AAAAAAAAA
Protein
RNA analysis[40–42,45]
DNA modification
Transcription[35–38]
DNA
M
M
M
M
DNA methylation[30,43]
Base calling[46,50,51]
Pathogenic variants[52–54,58–60]
Tumor genomes[55,56]
Variant interpretation

Recap: evolution of sequencing technologies



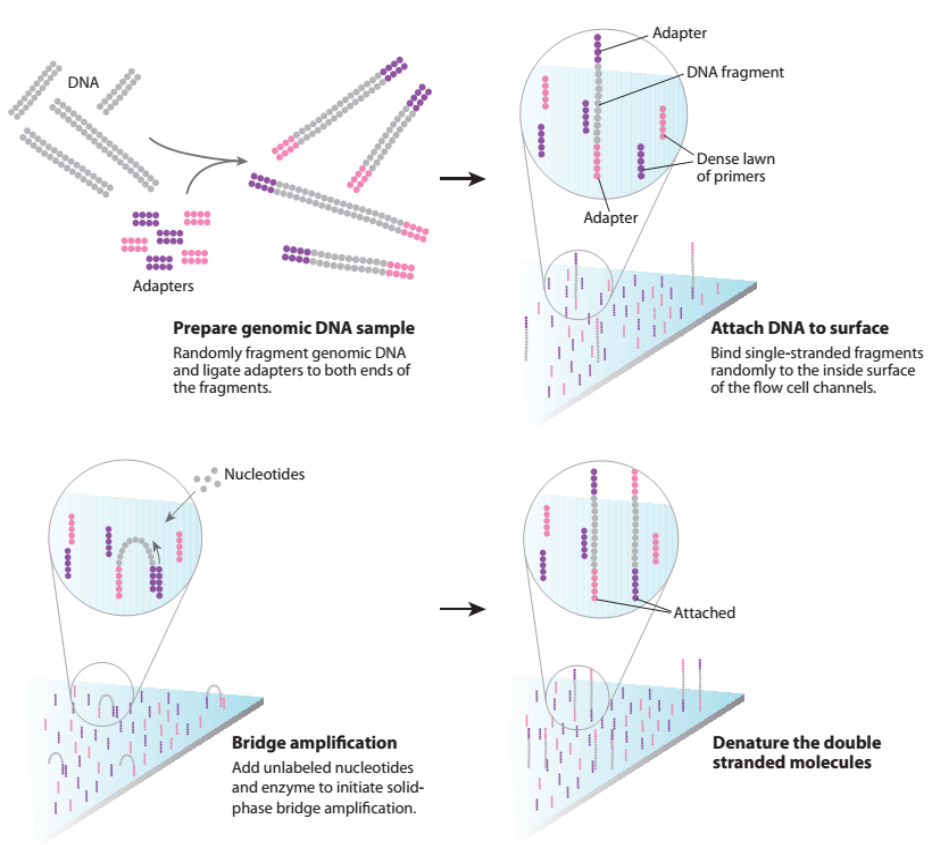
Recap: first generation sequencing — Sanger sequencing

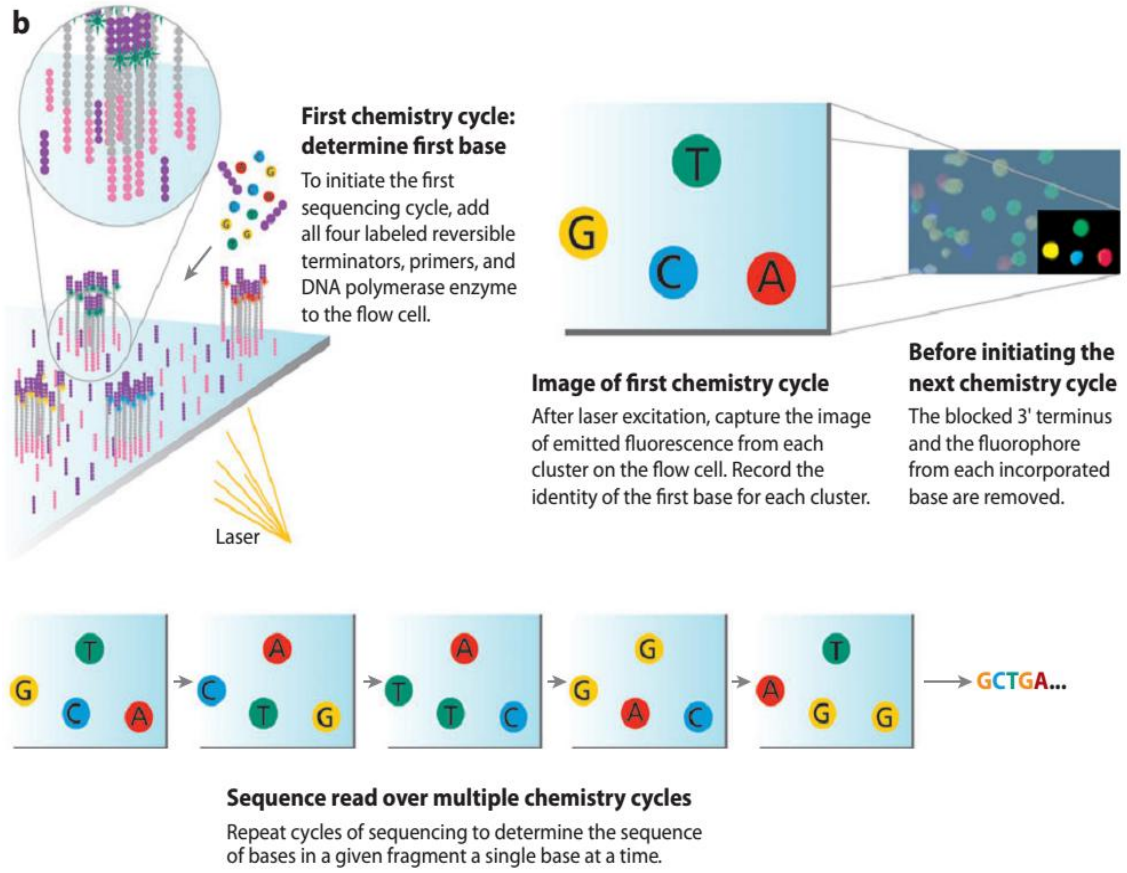
- The principle of Sanger sequencing is based on the termination of DNA strand elongation by ddNTPs
- 500-1,000 bp fragments
- Served as the bedrock of the Human Genome Project



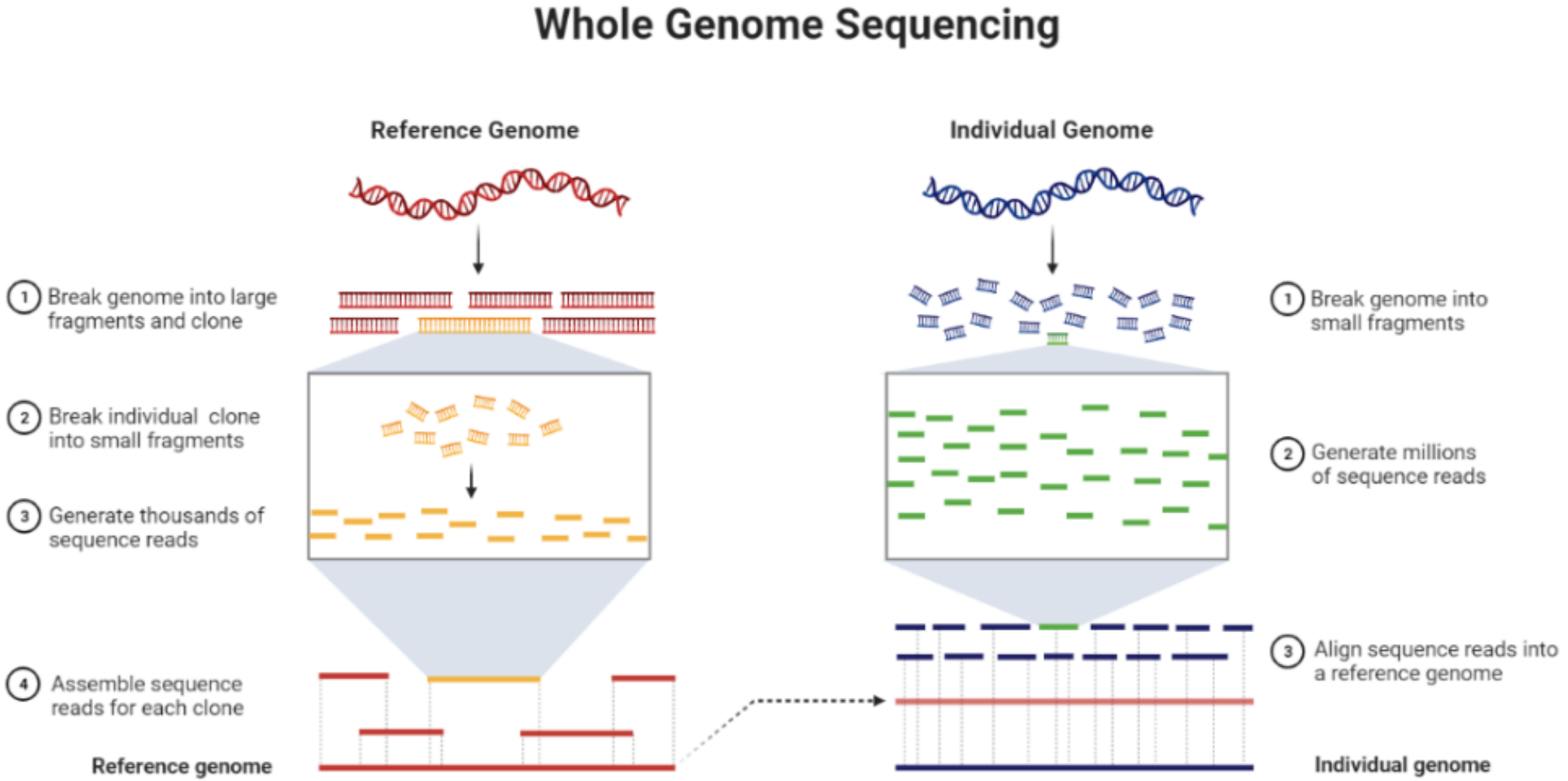
Recap: second generation sequencing — Illumina

- Allows for sequencing at very low cost
- Typically produce reads of ~50–500 bp in length

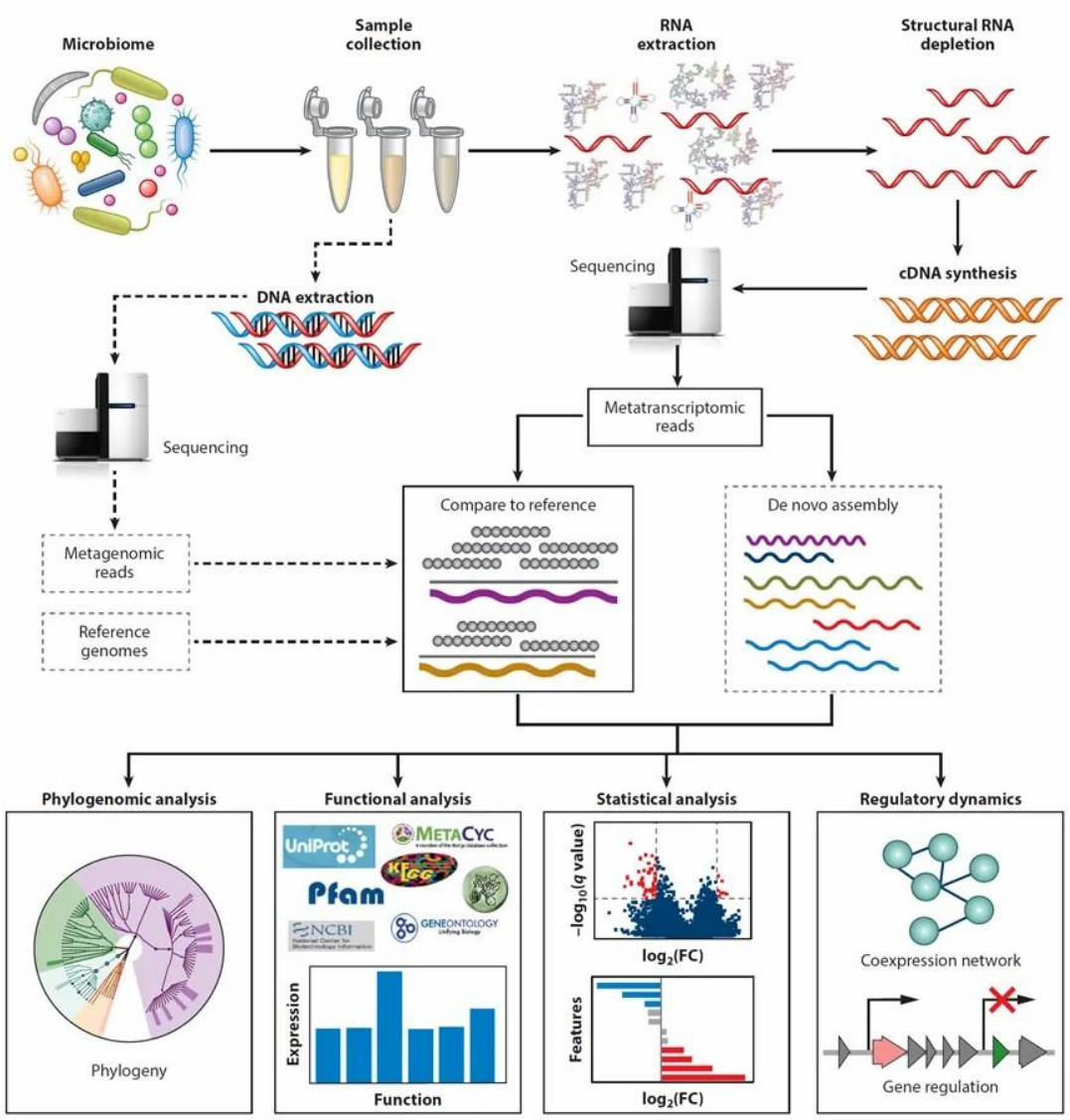


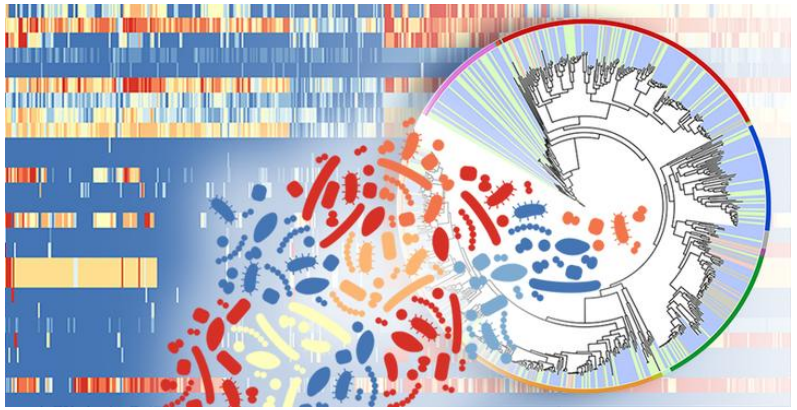


Genomic level: WGS



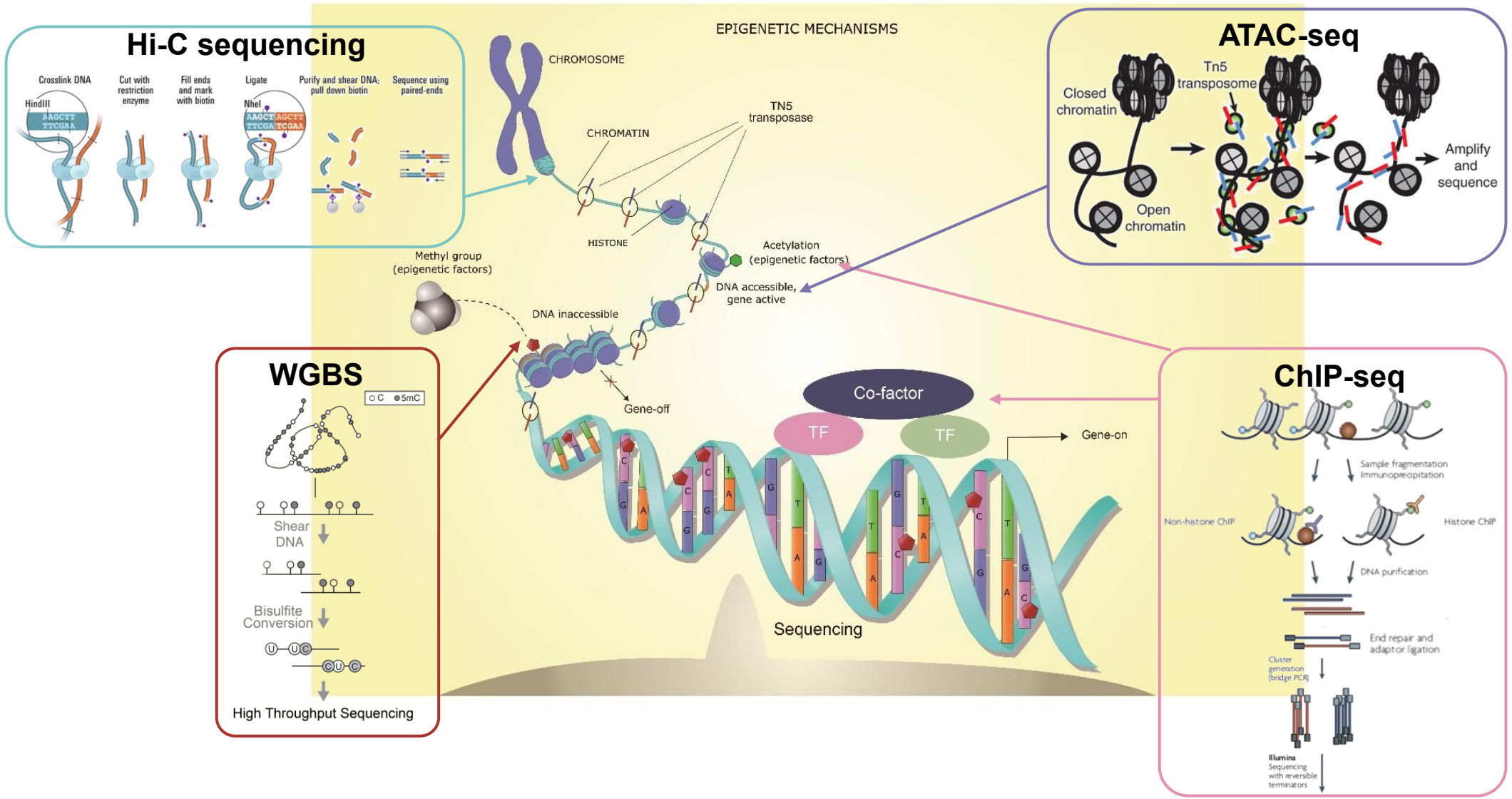
Metagenomic sequencing for microbiome





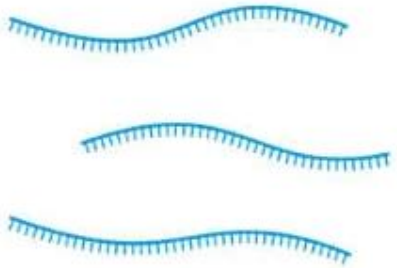
- No cultivation required: directly sequences all DNA from a complex sample
- Provides a census of the entire microbial community
- Mixed and complex data output
- Used for exploration and discovery of novel genes, pathways, and even entirely new branches of life

Epigenetic level: from chromatin to modifications



Transcriptomic level: RNA-seq

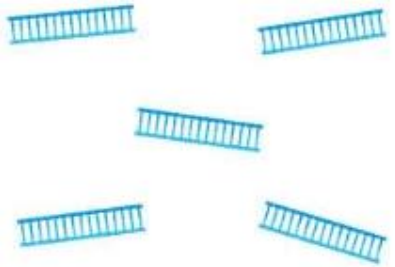
1 Isolate RNA from samples



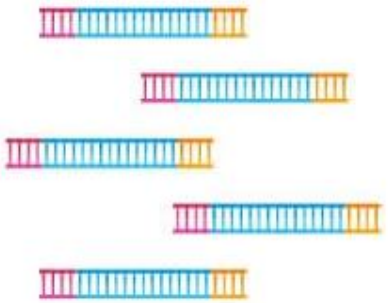
2 Fragment RNA into short segments



3 Convert RNA fragments into cDNA



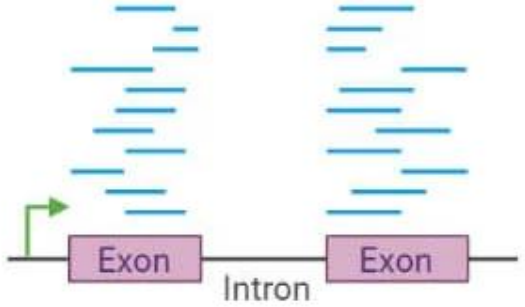
4 Ligate sequencing adapters and amplify



5 Perform NGS sequencing



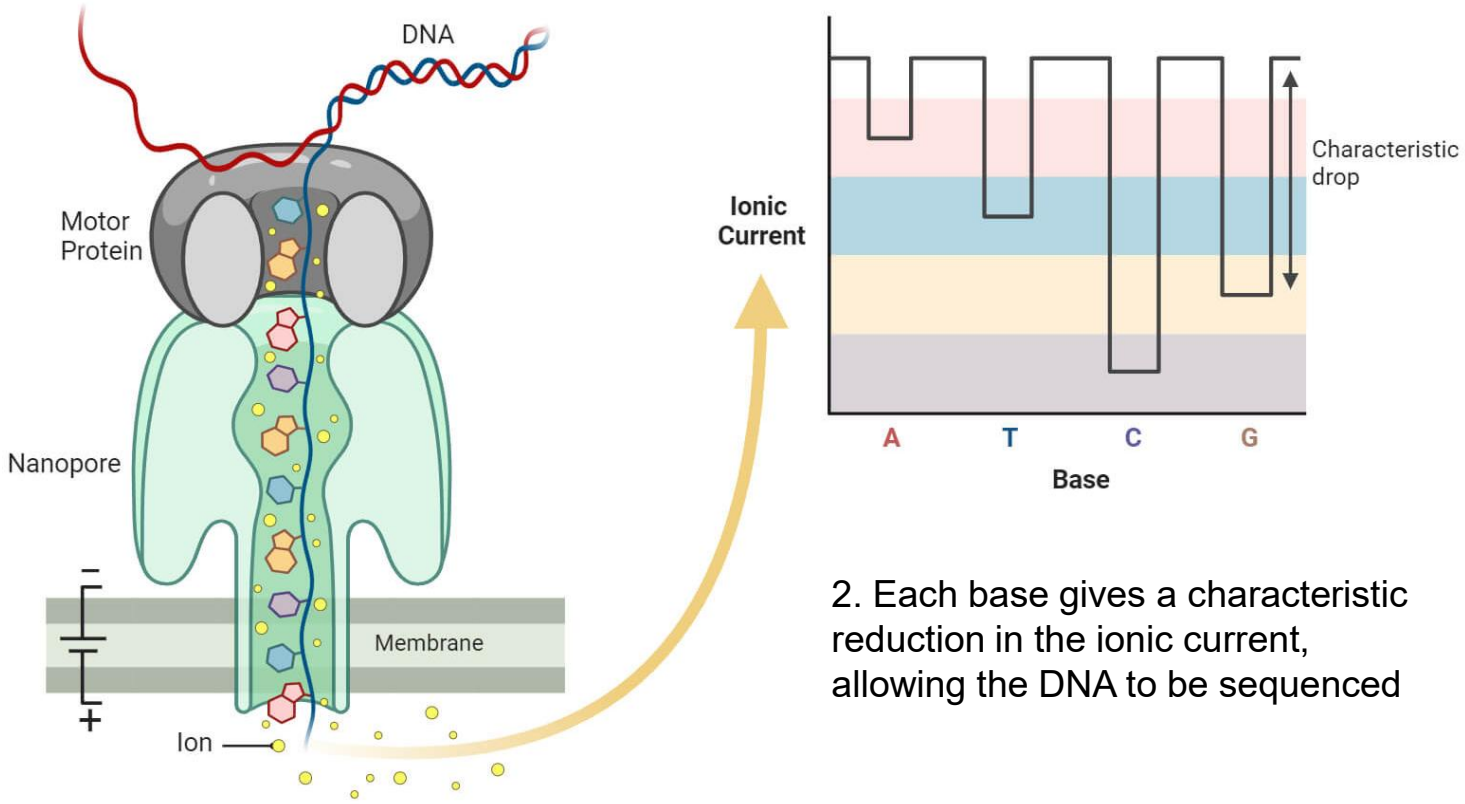
6 Map sequencing reads to the transcriptome/genome



Third generation sequencing — Nanopore

- Long-read sequencing using pore-forming proteins and electrical detection
- Tens of kb fragments

1. DNA is unwound by the motor protein and one strand is translocated through the pore to the +ve side of the membrane

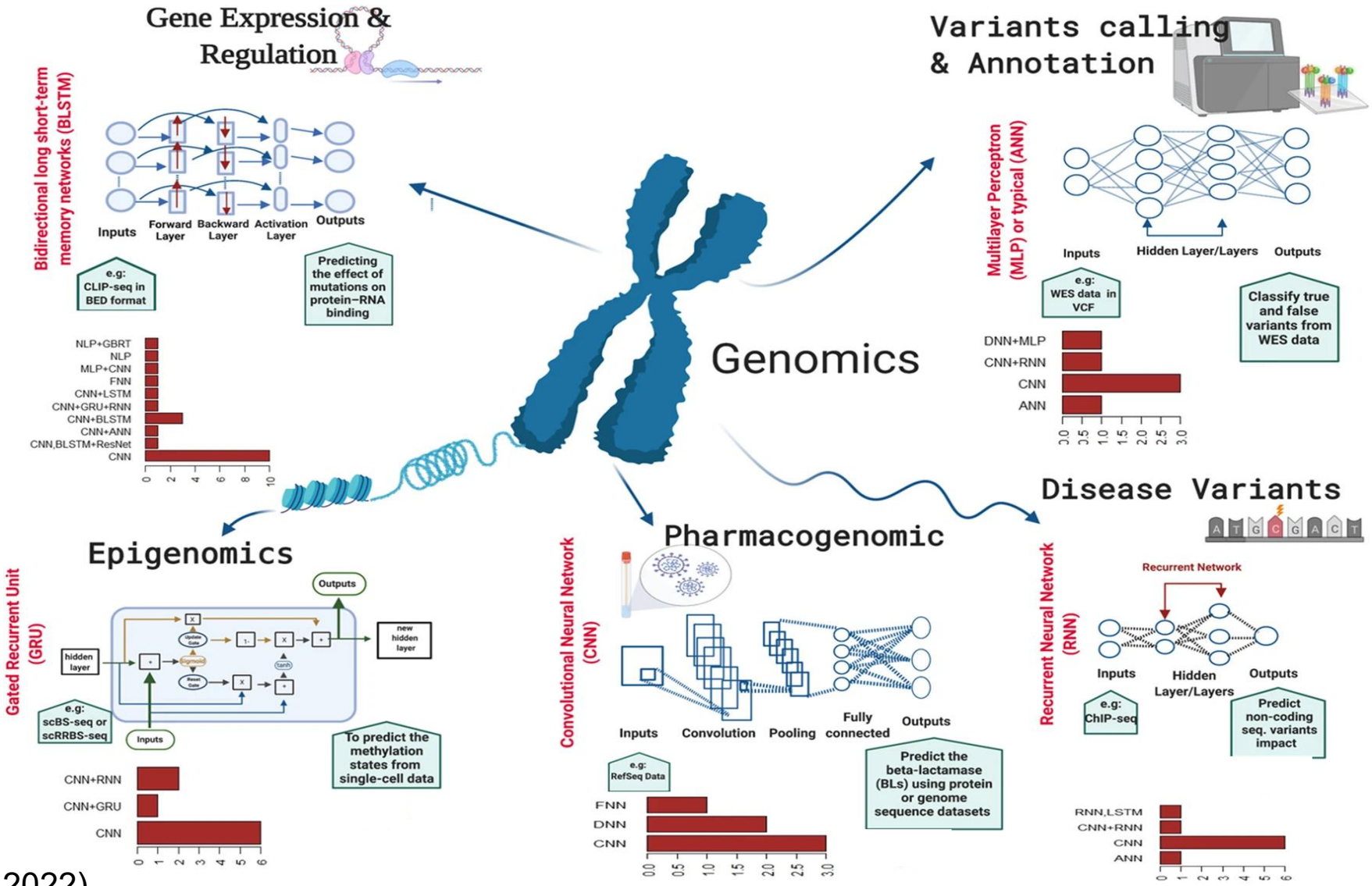


2. Each base gives a characteristic reduction in the ionic current, allowing the DNA to be sequenced

AI in genomics

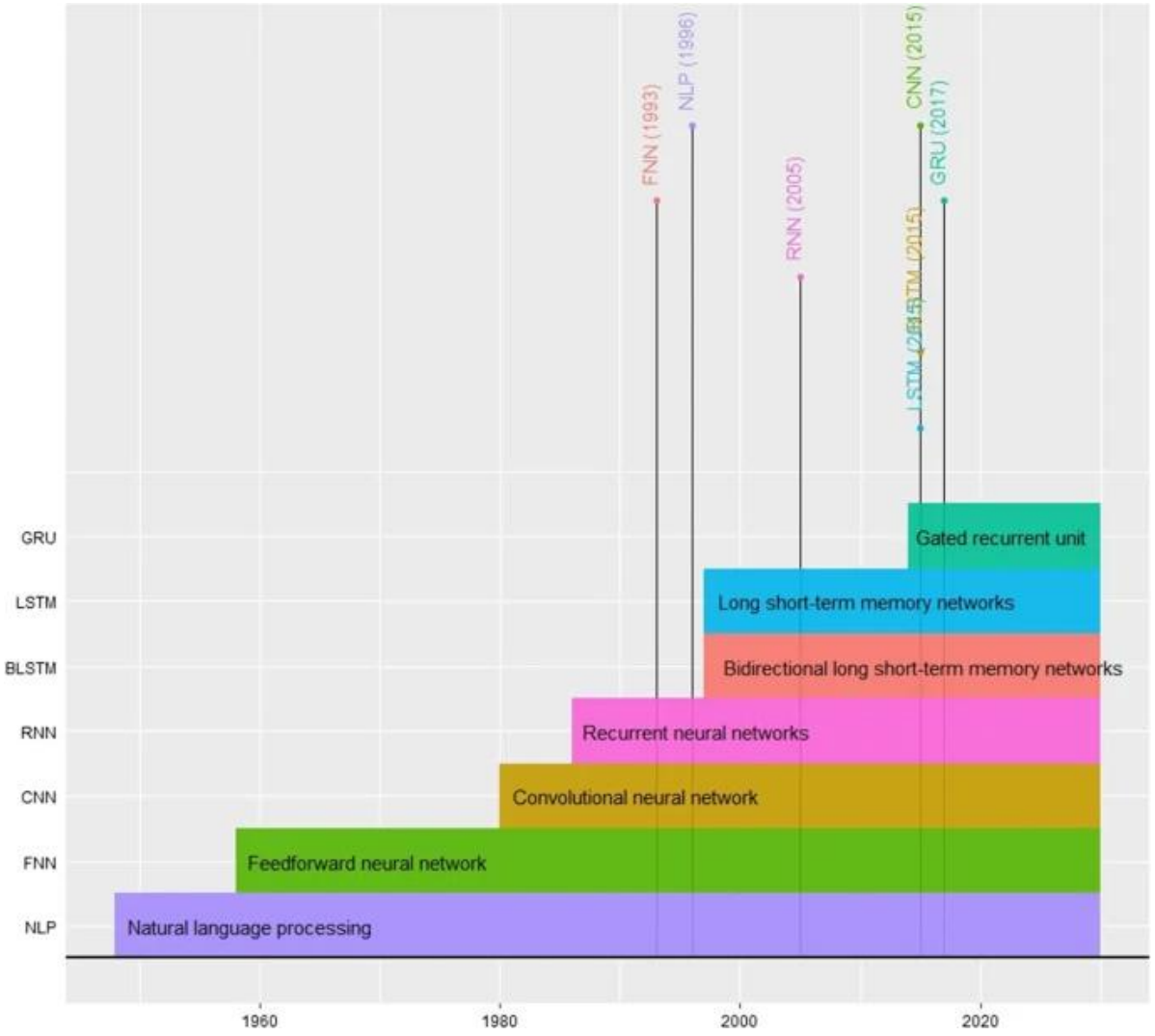
- Genome assembly and annotation
  - AI accelerates genome assembly
  - AI simplifies genomic annotation
  - Foundation models provide an end-to-end solution for all genome annotation tasks.
- Comparative genomics and evolution
  - Traditional methods and databases
  - AI is capable of understanding evolution
- Variants calling and mutation detection

AI in genome and genomics



(*Human Genomics*, 2022)

The debut of AI methods in the genomics



(*Human Genomics*, 2022)

- CNN: Alipanahi et al,. DeepBind (2015). Prediction of variant impacts on expression and disease risk, predicting drug response of tumours from genomic profiles, and pharmacogenomics.
- RNN: Maraziotis, Dragomir and Bezerianos. Recurrent neuro-fuzzy technique (2005). Predicting transcription factor binding sites, for Alignment and SNV identification.
- BLSTMs/LSTMs: Quang and Xie. DanQ model (2015). DNA function predictions and prediction of protein localisation, predict miRNA precursor, Enhancer–promoter interaction (EPI) prediction.
- GRU: Angermueller et al,. DeepCpG (2017). Enhancers and methylation states predictions.

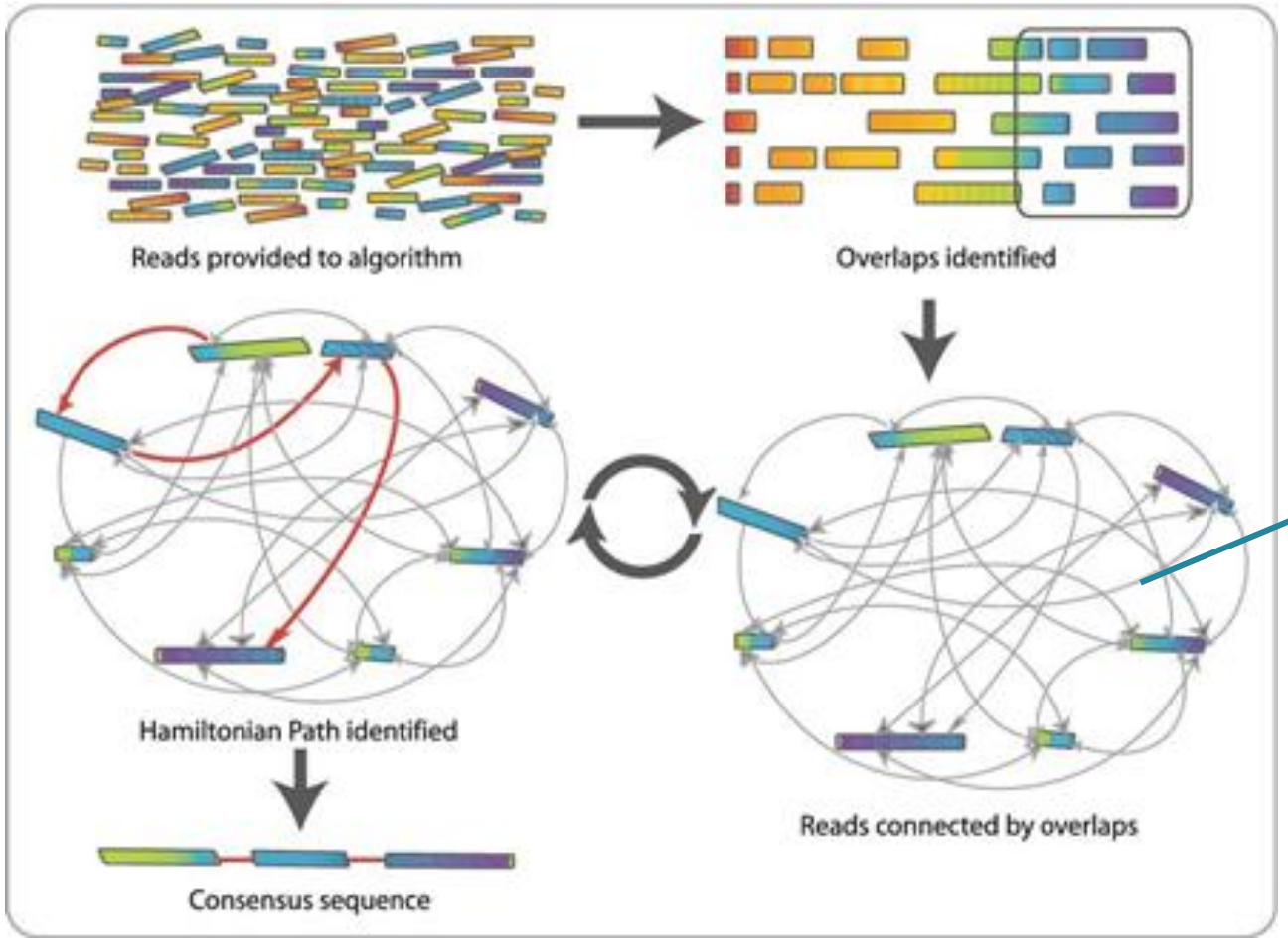
AI in genome assembly and annotation

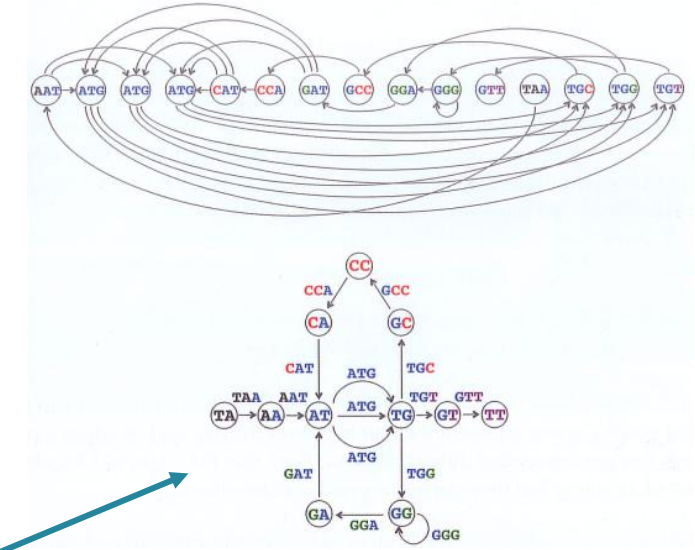
- **Problem definition**: given a vast of raw sequencing reads as inputs, output a complete, annotated genome sequence with structural features and functional elements identified.

- **Application**:
  - Genome construction and reads mapping
  - “Structural” genome annotation is the process of identifying genes and their intron–exon structures.
  - “Functional” genome annotation is the process of attaching meta-data such as gene ontology terms to structural annotations.

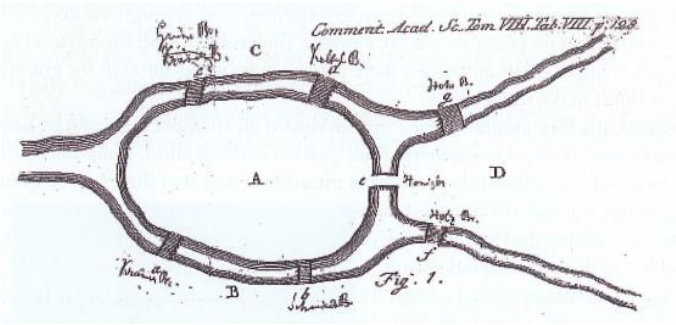


Genome assembly



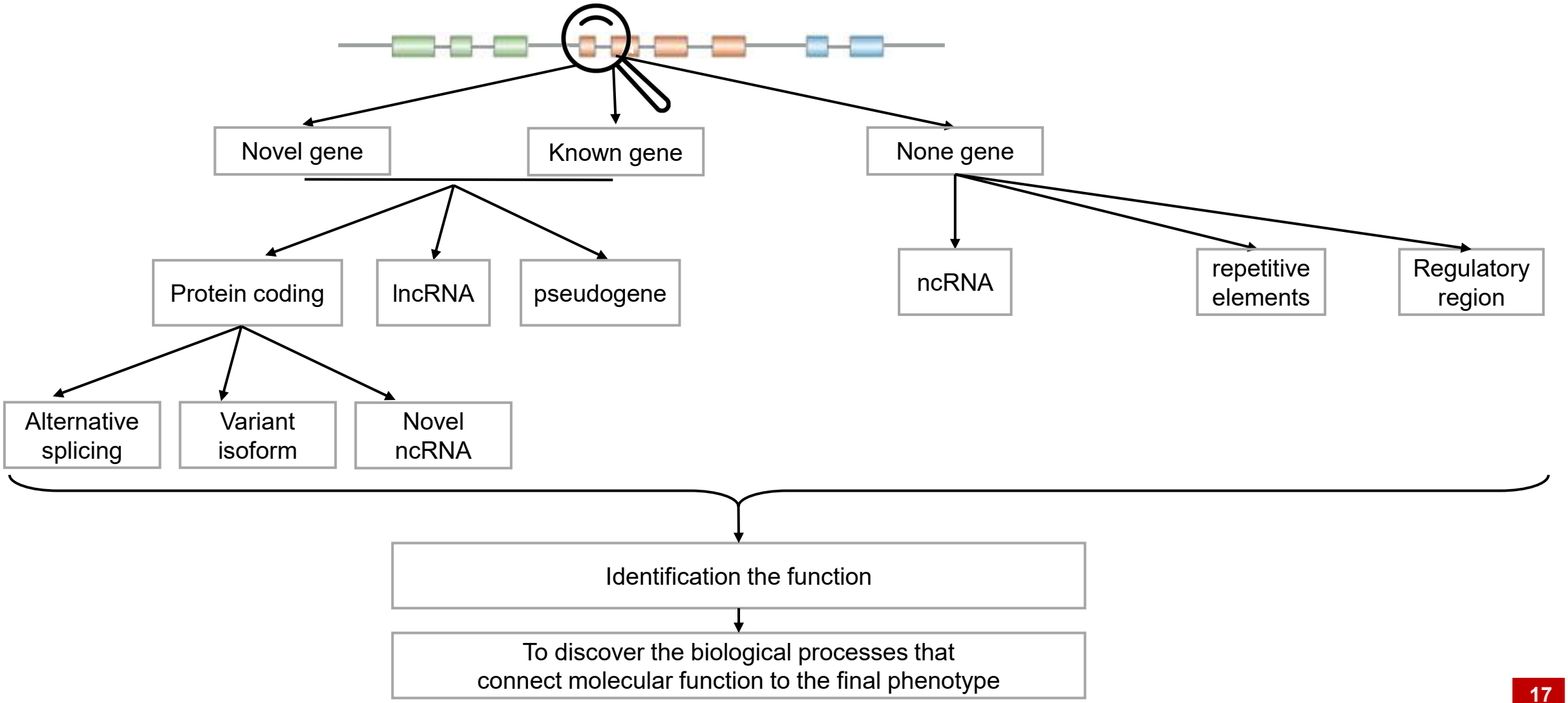


The overlap graph and de Bruijn graph of 3-mer DNA assembly

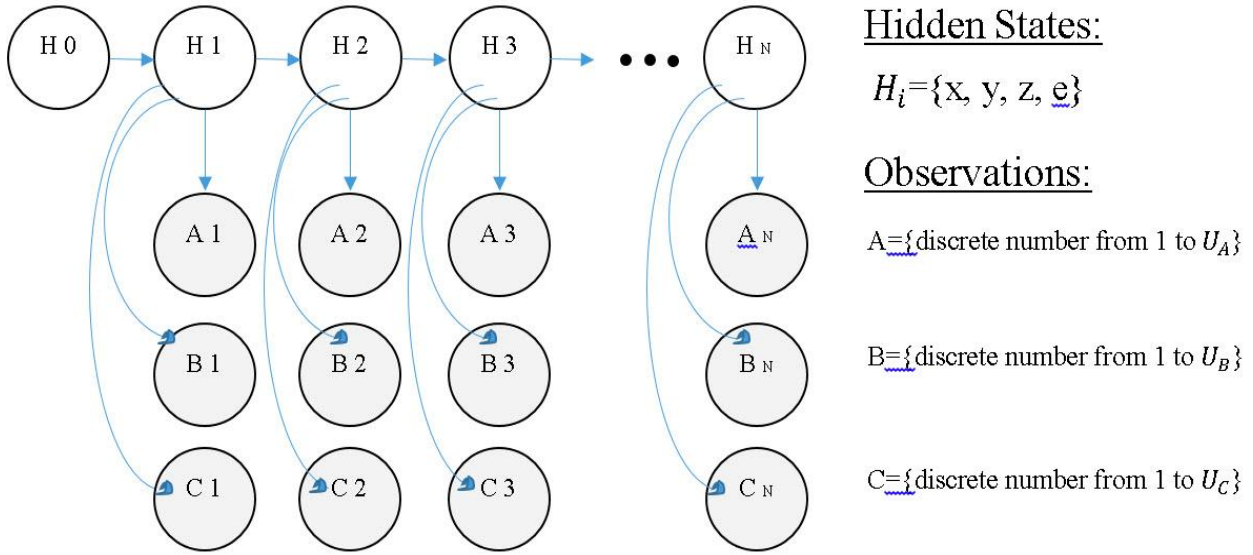


Euler's 1735 illustration of Königsberg

Genome annotation

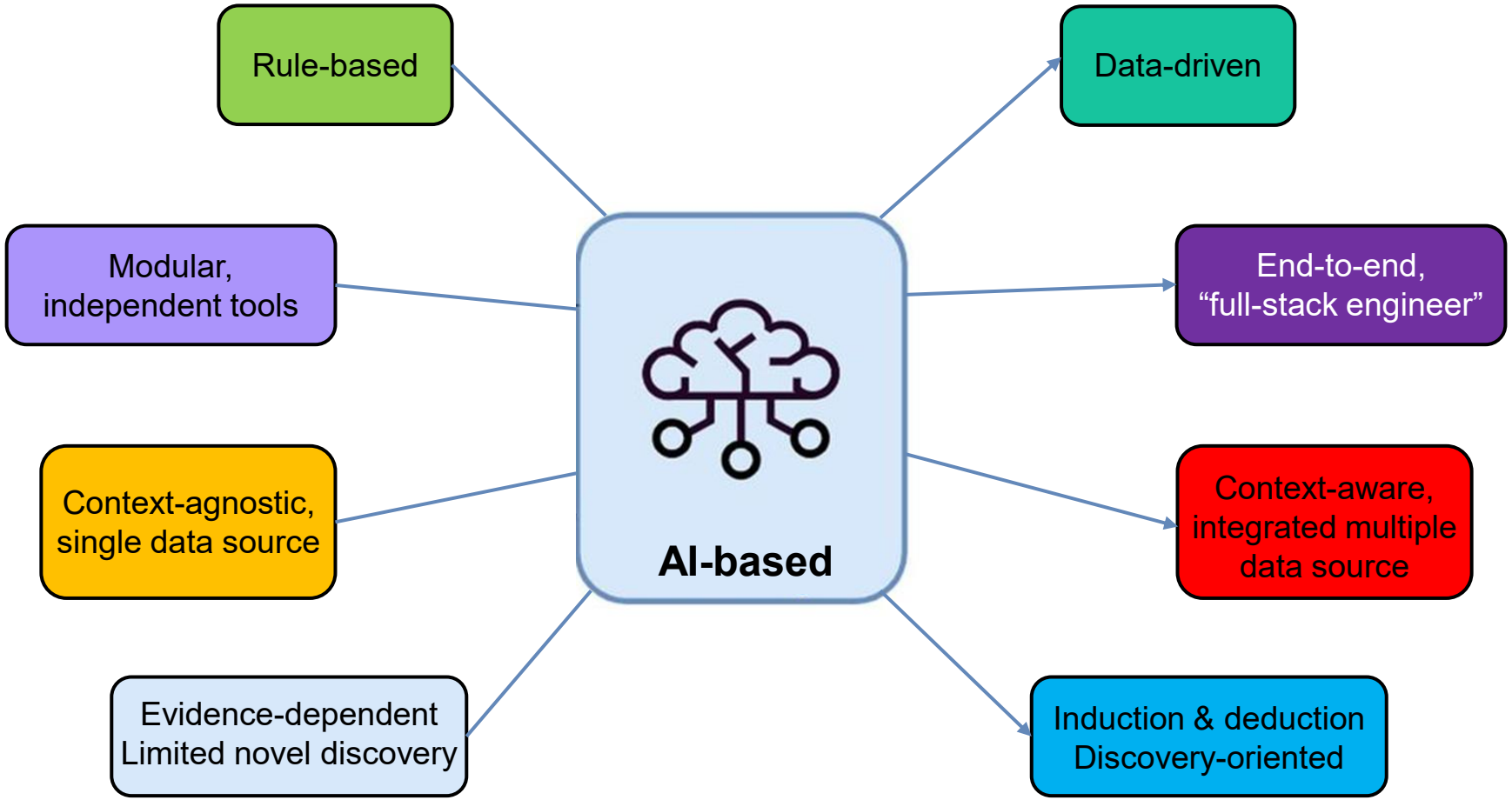


Genome annotation

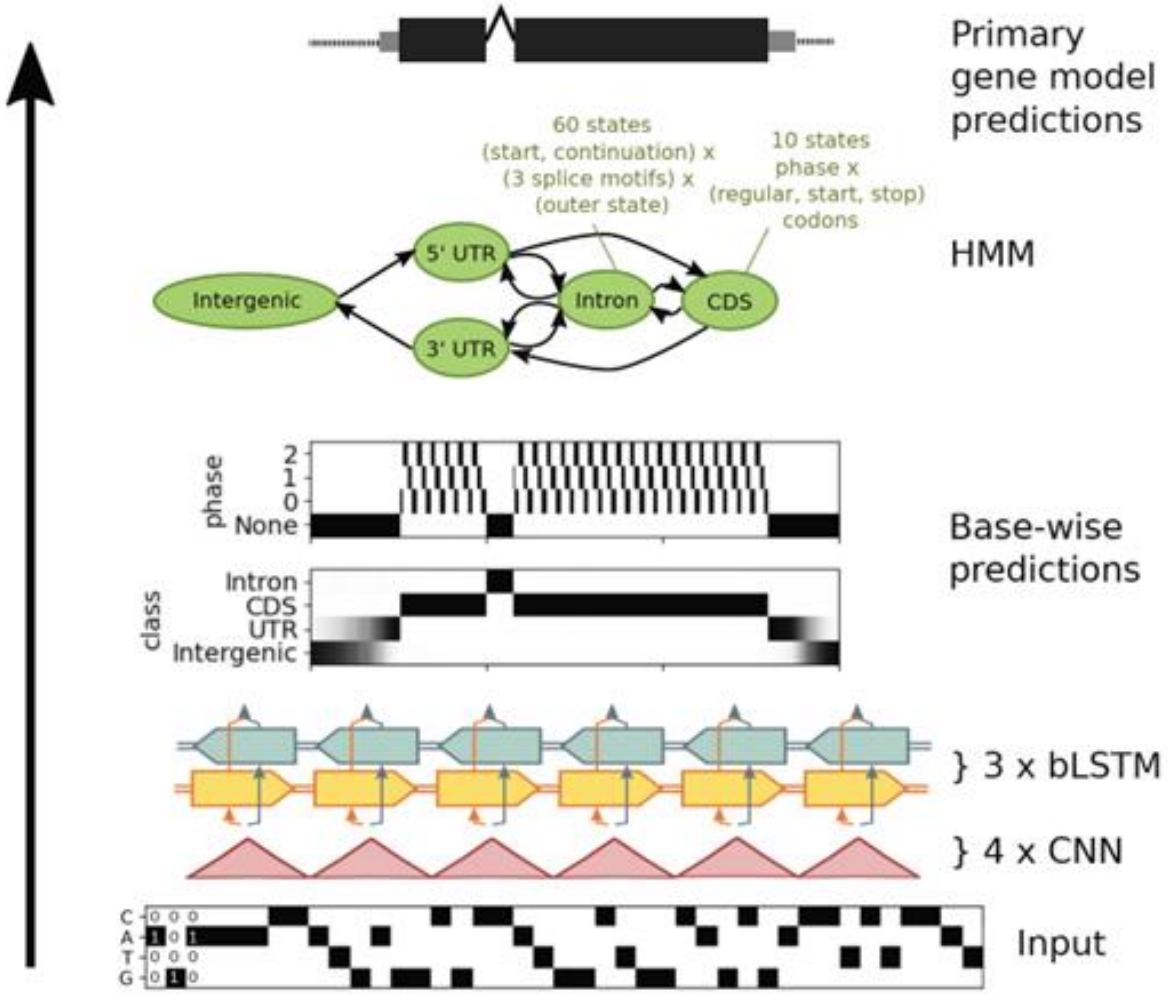


- GenScan (http://genes.mit.edu/GENSCAN.html) 5th-order HMM
  - Combined hexamer frequencies with coding signals, Initiation codons, TATA boxes, CAP site, Poly-A
  - Trained on Arabidopsis and maize data
  - Extensively used in human genome project
- HMMgene (http://www.cbs.dtu.dk/services/HMMgene)
  - Identified sub regions of exons from cDNA or proteins
  - Locks such regions and used HMM extension into neighboring regions

The genome annotation paradigm is undergoing a fundamental shift, driven by AI methods



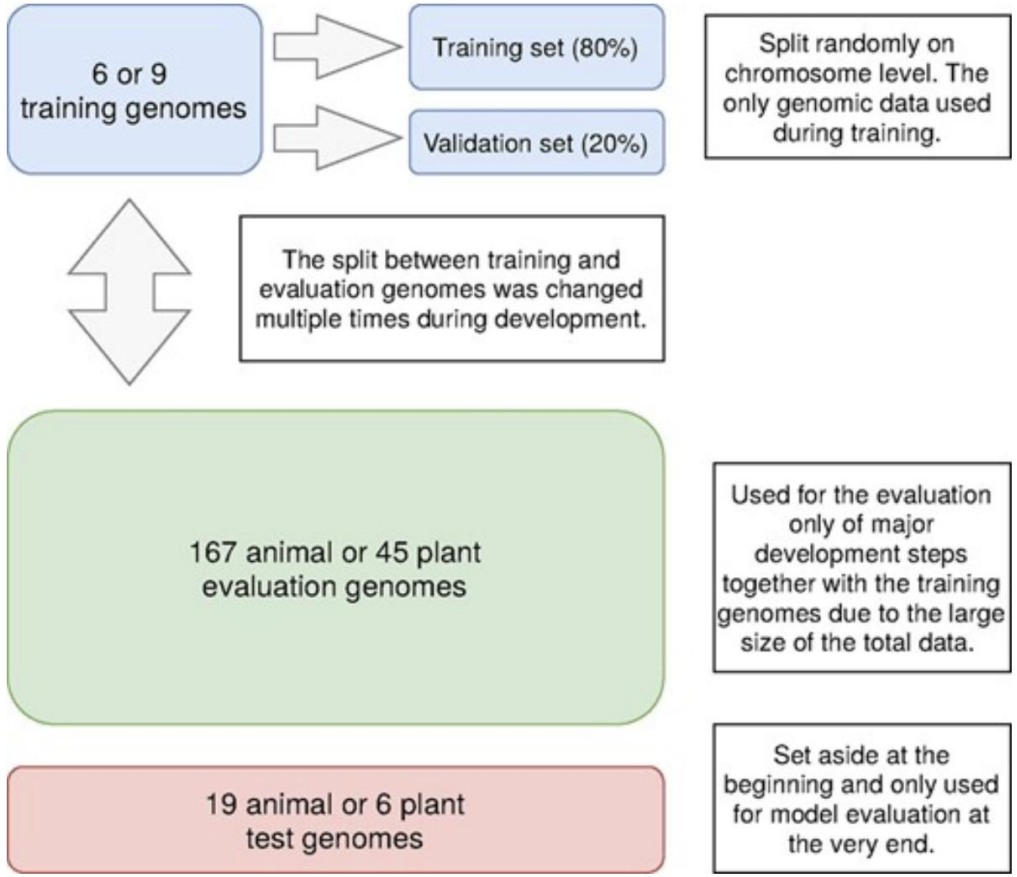
Helixer: A deep learning-based gene prediction tool



- Utilizes a deep learning model trained on diverse animal and plant genomes, reducing the need for retraining on closely related species.
- Provides probabilities for every base, enabling detailed predictions for intergenic, untranslated, coding, and intronic regions.
- Outperforms traditional tools like AUGUSTUS in both base-wise metrics and RNA-Seq consistency.
- Handles large and complex eukaryotic genomes effectively, even for species with longer genes.
- Requires no additional data except genome.
- Cannot predict alternative splicing events or spliced start codons.

(*Bioinformatics, 2020*)

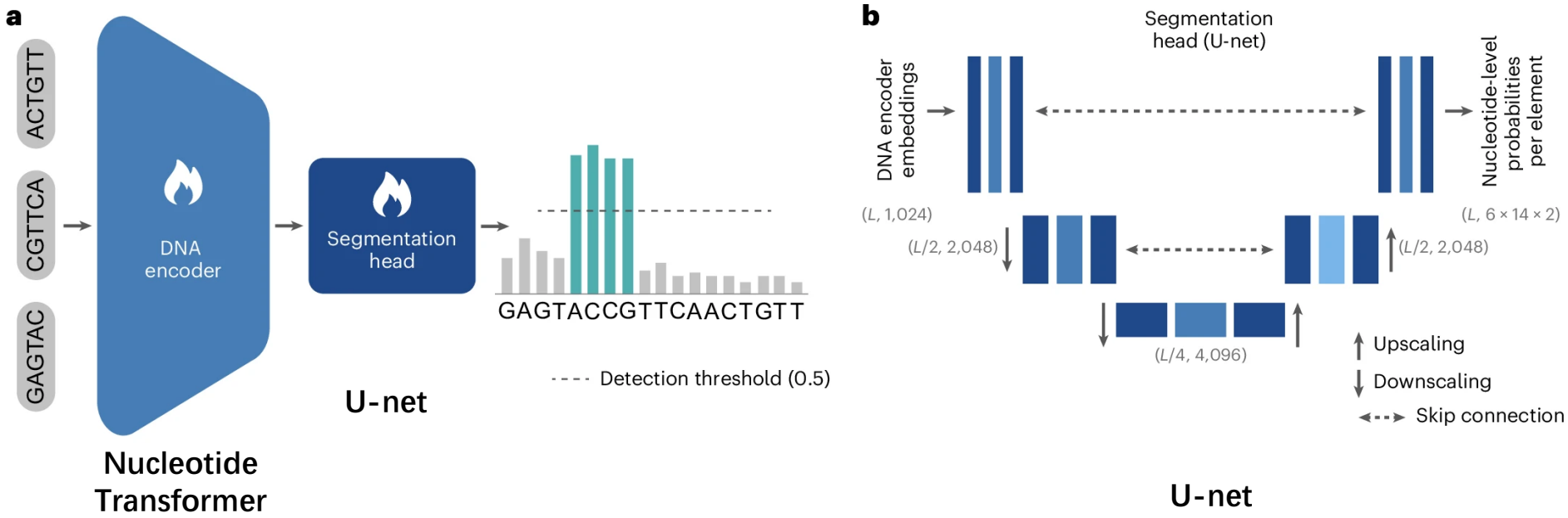
Data and representation



Name	Information
Input	Genomic sequence in the 4-dimensional IUPAC encoding
	(one hot encoding for non-ambiguous bases)
Output	Labels in a 4-dimensional one hot encoding representing
	the classes *Intergenic*, *UTR*, *CDS* and *Intron*
Sample weights	One of {0, 1}; whether there is an error at a base

(*Bioinformatics, 2020*)

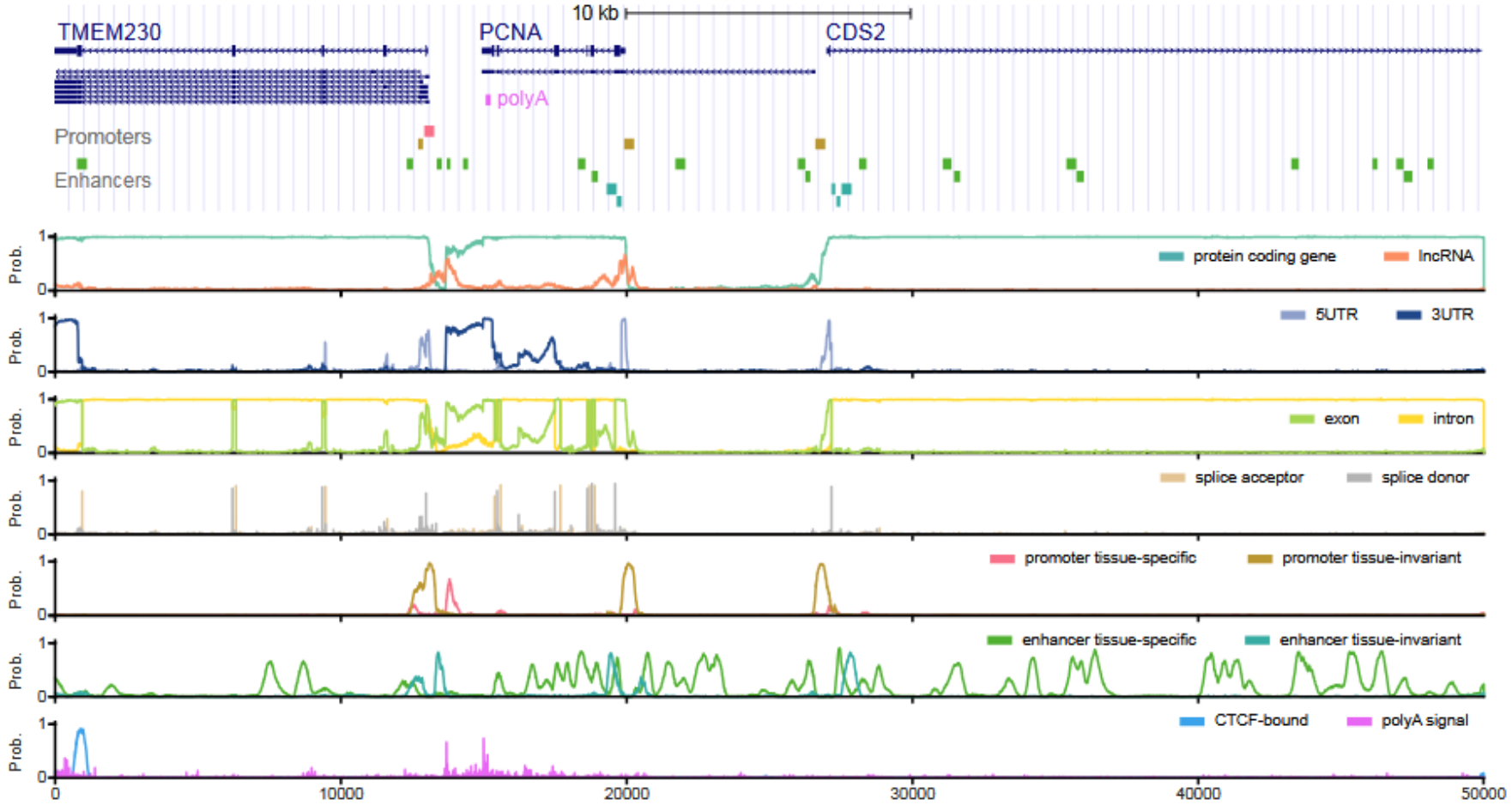
SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models



- **Generalized** annotation tool for 14 types of genic elements on unseen species
- **Genome annotation as semantic segmentation**: nucleotide-level precision
- The SegmentNT framework consists of two main components:
  - The DNA Encoder: Nucleotide Transformer (NT)
  - The Segmentation Head: 1D U-Net Architecture

(*Almeida, et al*, 2024)

SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models



(*Almeida, et al*, 2024)

AI in comparative genomics and evolution

- **Problem definition**: Given multiple genome sequences across species or strains, the goal is to jointly reconstruct their evolutionary information.

- **Application**:
  - Reconstructing the “tree of life” and tracing the evolutionary relationships between species
  - Revealing how genomic elements function and identifying evolutionarily conserved non-coding regions
  - helping identify genetic variations linked to disease susceptibility, drug response, and the emergence of antibiotic resistance or viral escape mutants
  - Guiding protein engineering and synthetic biology

Sequence alignment and BLAST

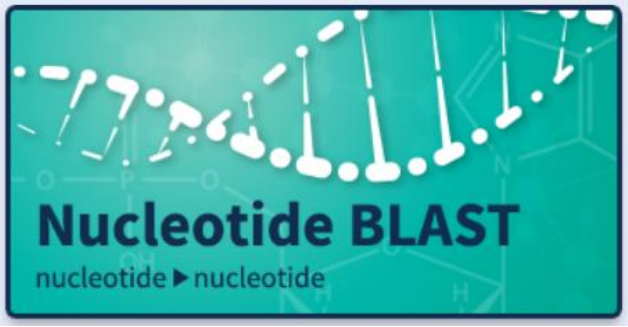
```
seq1  ACGTTTGCT

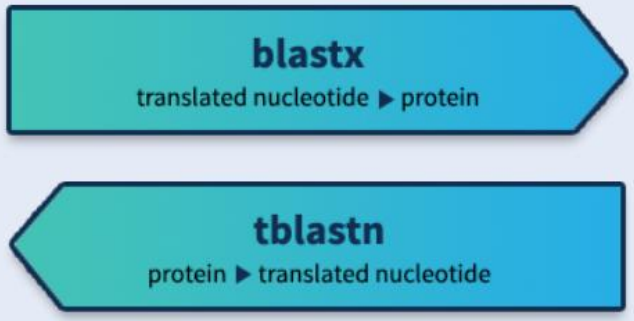
seq2  ATGTTGCAT
```

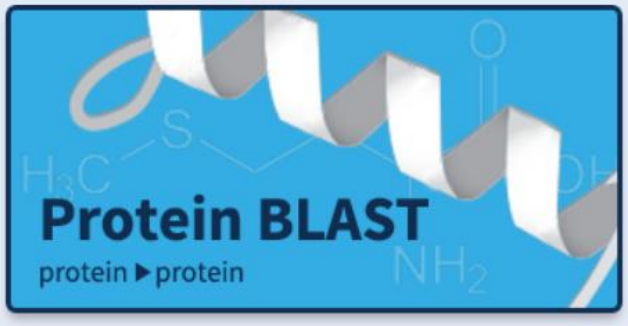


```
seq1  ACGTTTGCC-T
      |  |||  |||  |
seq2  ATGTT-GCCAT
```

Substitution Deletion Insertion

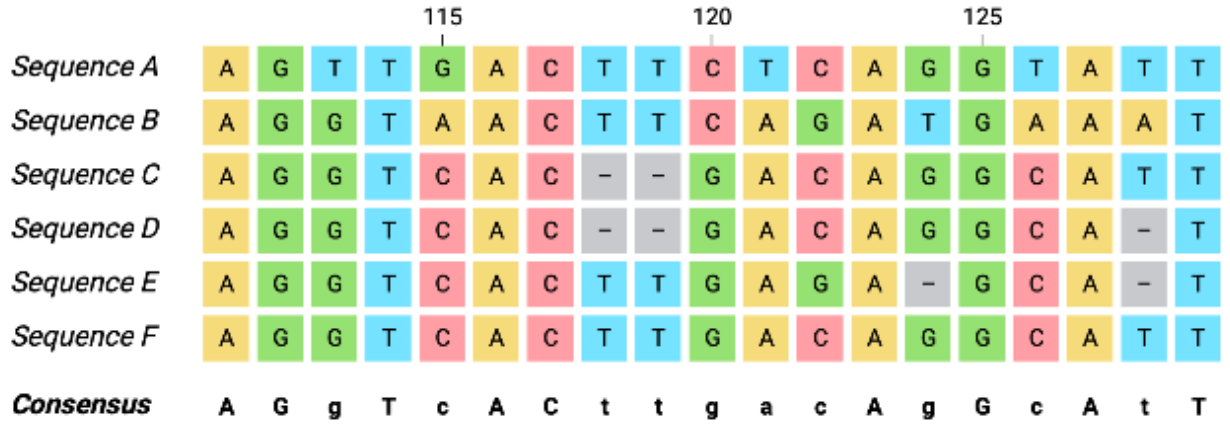


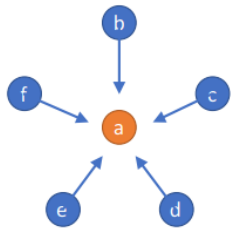




Homology ⟷ Similarity

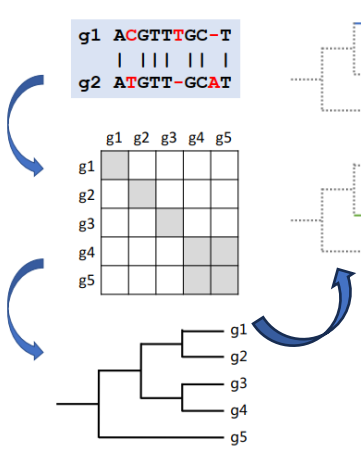
Multiple sequence alignment





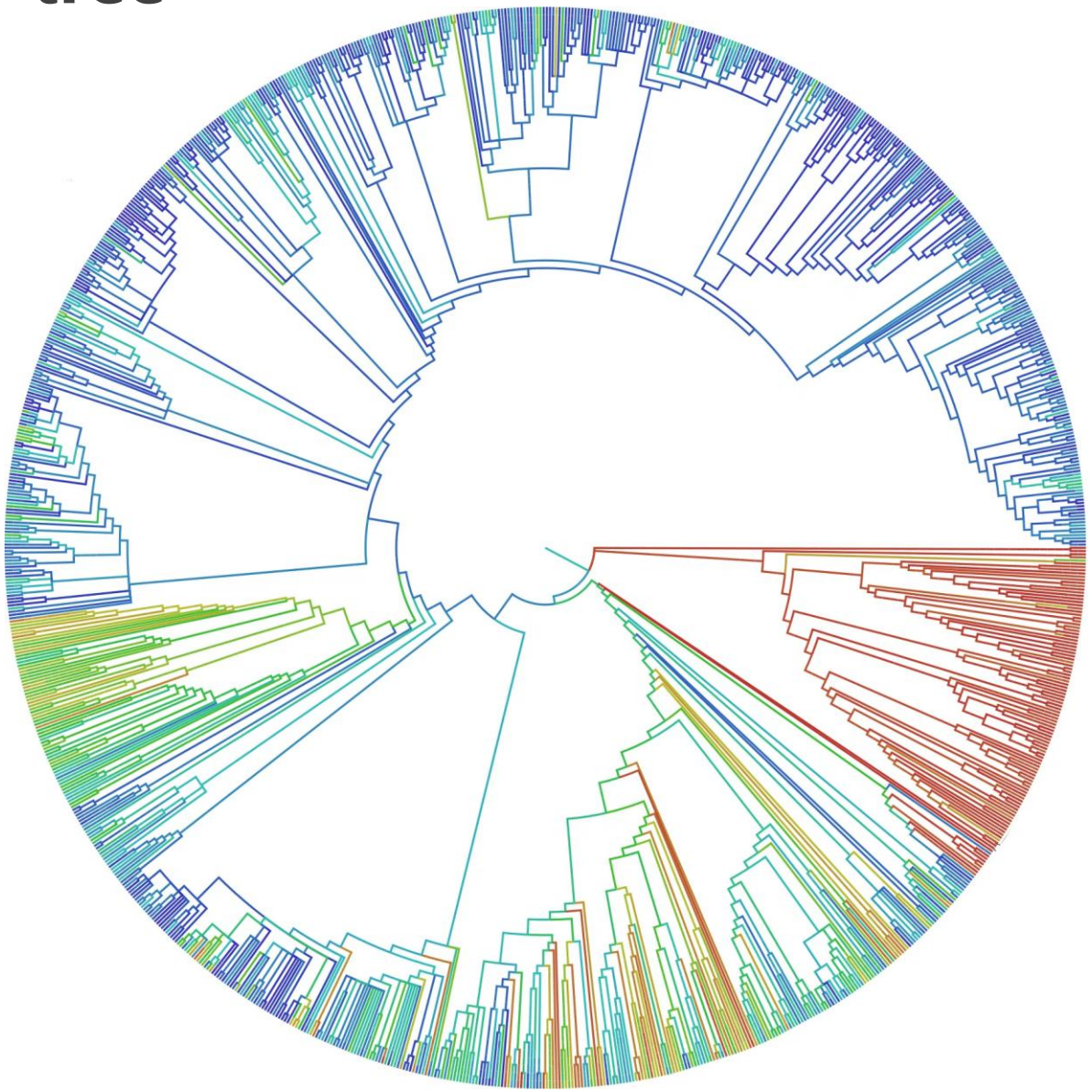
Star Alignment:
Select one sequence as the center (reference sequence) and map all other sequences onto it

```
g1 ACGTTTGC-T      g1 ACGTTTGC-T      g1 ACGTTTGC-T      g1 ACGTTTGC-T
   |||||  ||  |  +     |  |||  ||  |  +     |  |  |  ||  |  =  g2 ACGTTAGCAT
g2 ACGTTAGCAT      g3 ATGTT-GCAT      g4 ATGAT-GCAT      g3 ATGTT-GCAT
                                                         g4 ATGAT-GCAT
                                                         g5 ... ...
```

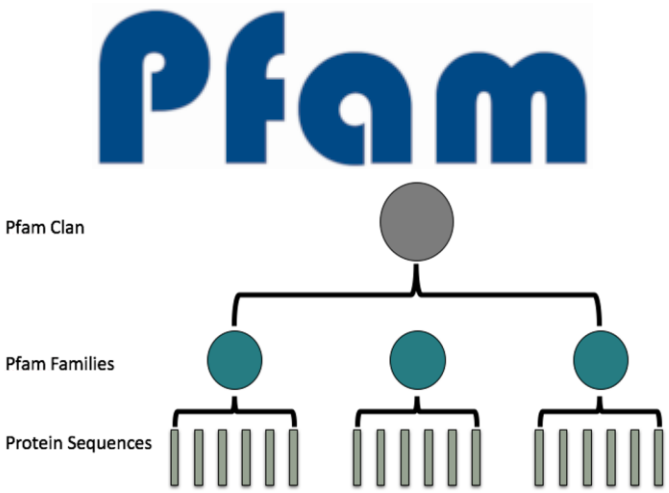


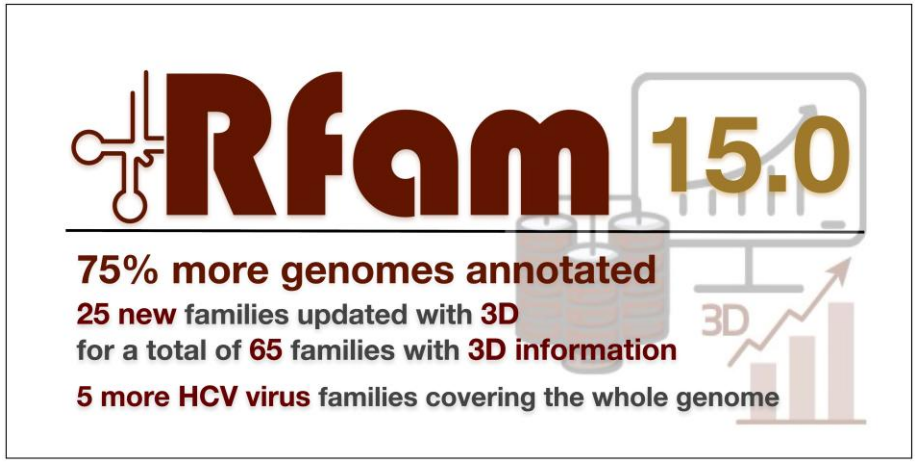
Tree Alignment:
Similarity is calculated through pairwise sequence alignments to build a hierarchical cluster, based on which progressive pairwise alignment is conducted from the bottom up

Evolutionary tree



Pfam & Rfam

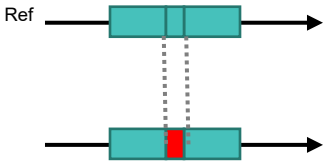




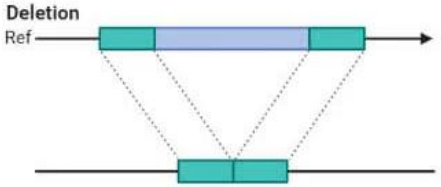
- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments (MSA) and hidden Markov models (HMMs).
- The Rfam database is a collection of RNA sequence families of structural RNAs including non-coding RNA genes as well as cis-regulatory elements. Each family is represented by a multiple sequence alignment (MSA) and a covariance model (CM).

Genome variations

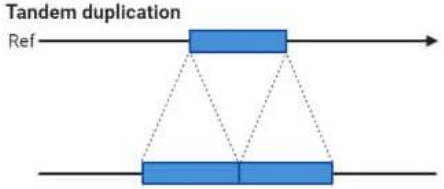
Mutation



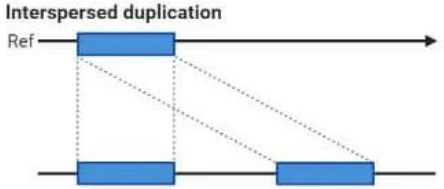
Deletion



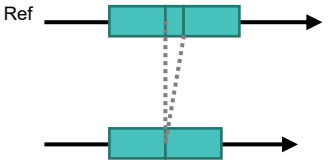
Tandem duplication



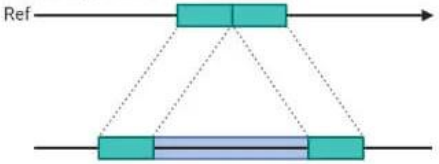
Interspersed duplication



Deletion (1nt)



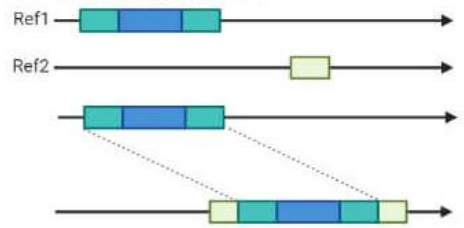
Novel sequence insertion



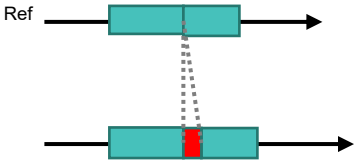
Translocation



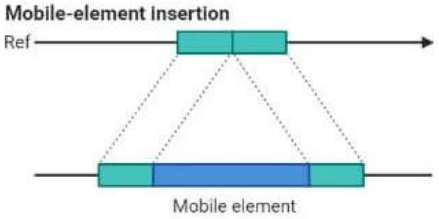
Inter-chromosome insertion



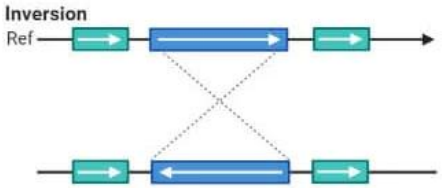
Insertion (1nt)



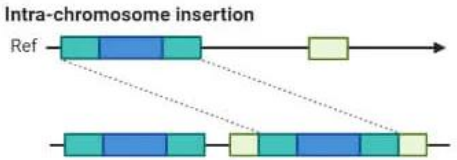
Mobile-element insertion



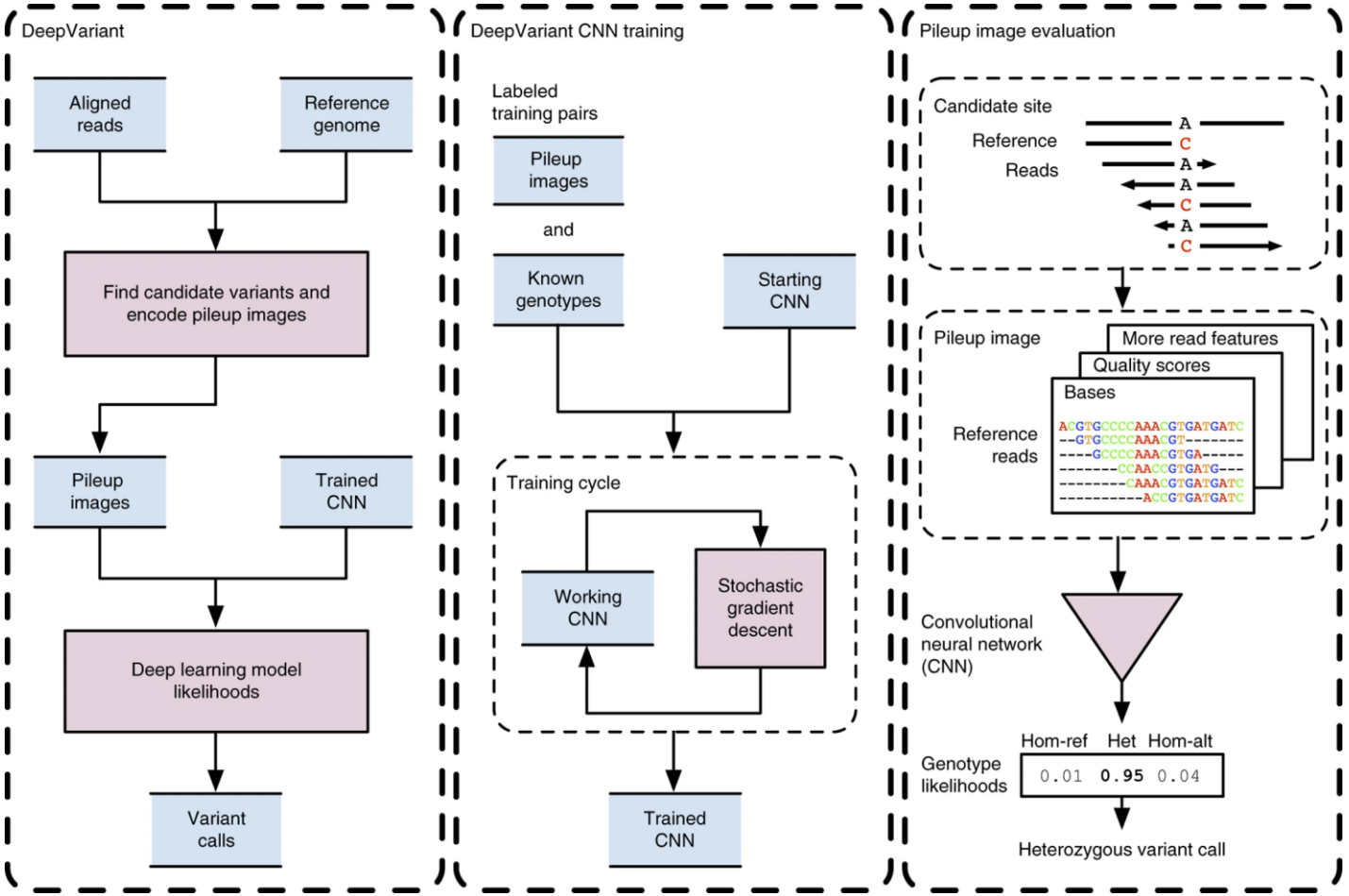
Inversion



Intra-chromosome insertion



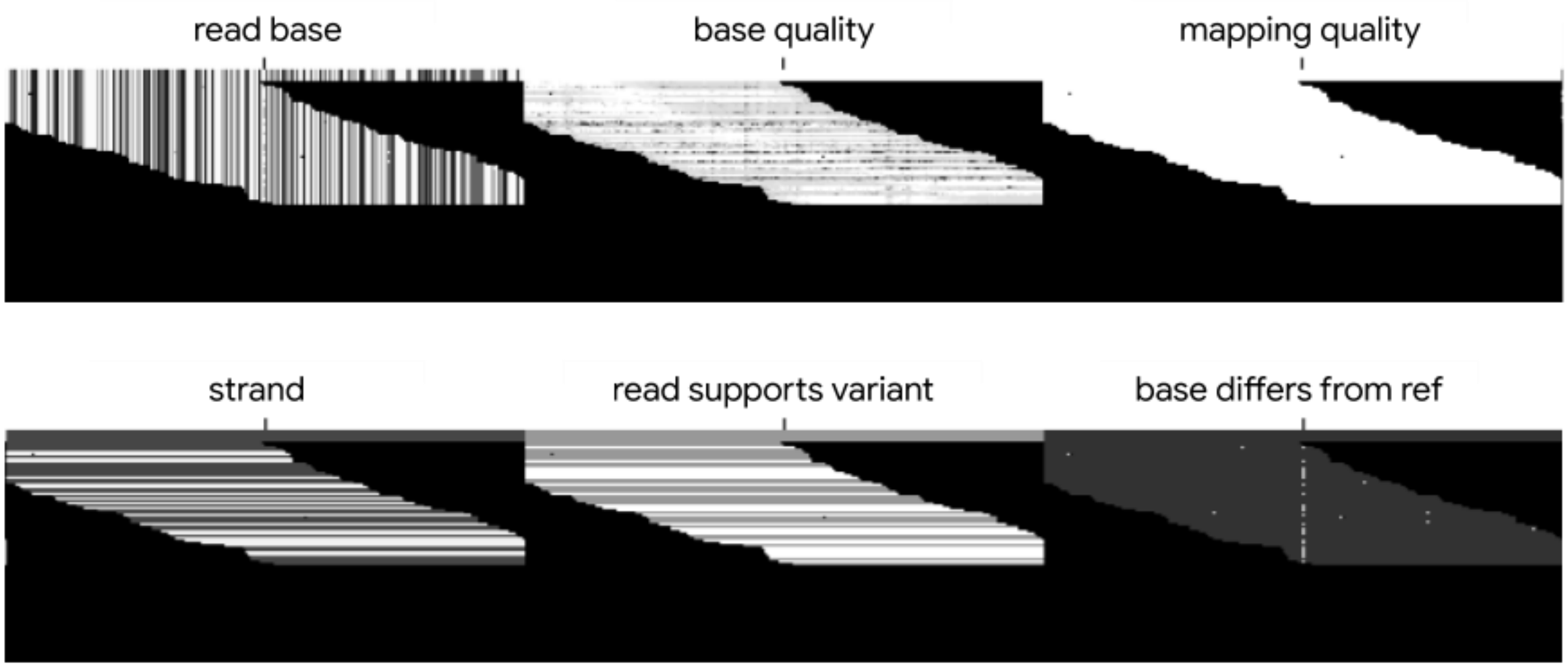
DeepVariant



- DeepVariant is a deep learning-based variant caller
- It takes aligned reads (in BAM or CRAM format), produces pileup image tensors from them, classifies each tensor using a convolutional neural network, and finally reports the results in a standard VCF or gVCF file.
- Application in variation detection of tumor tissues.

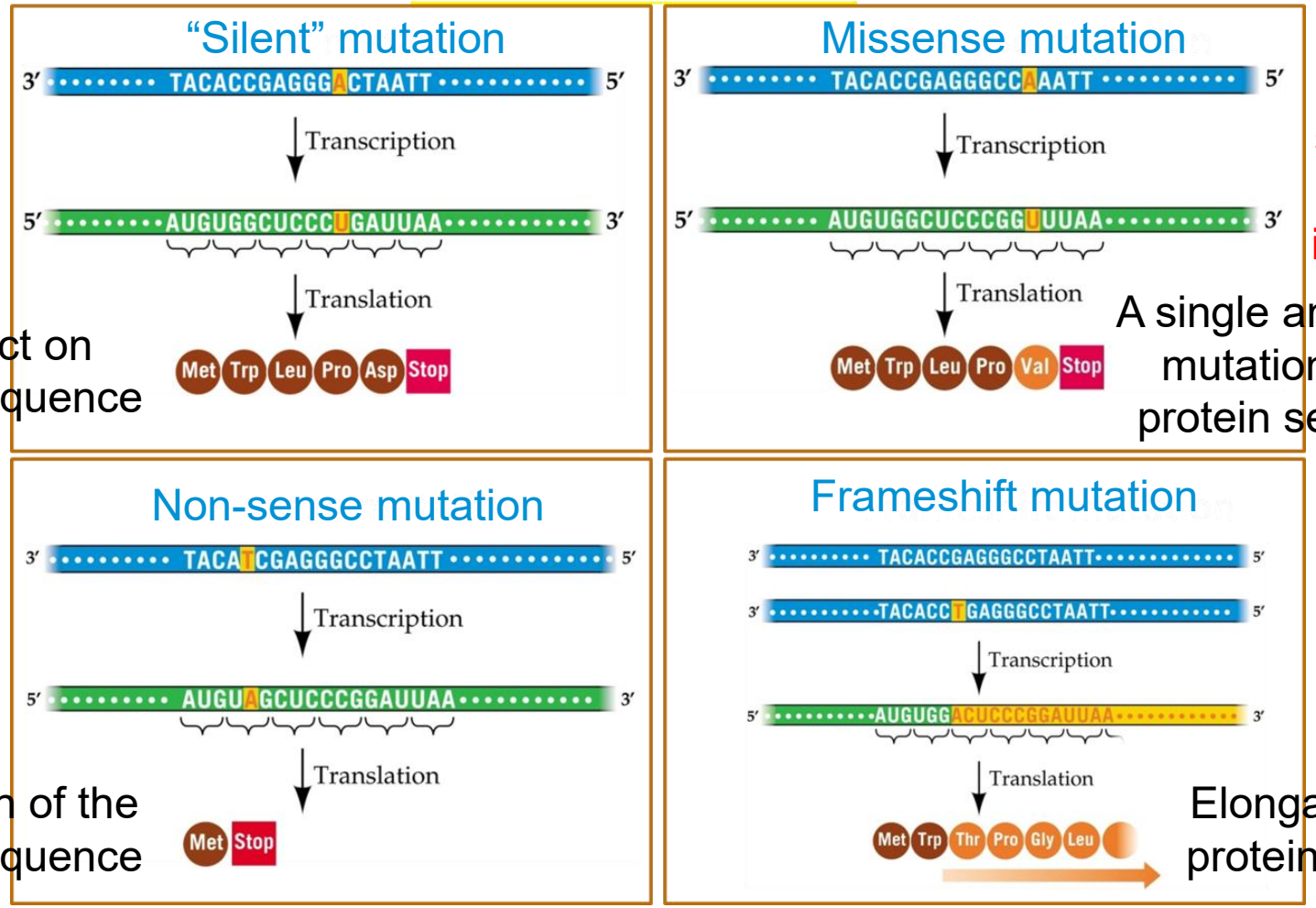
(*Nature Biotechnology*, 2018)

DeepVariant



(*Nature Biotechnology*, 2018)

The types of single-nucleotide variant (SNV)



No effect on protein sequence

A single amino acid mutation in the protein sequence

Can minimal perturbations to a protein sequence lead to decisive effects on its structure, interactions, and functions?

? Function?

Truncation of the protein sequence

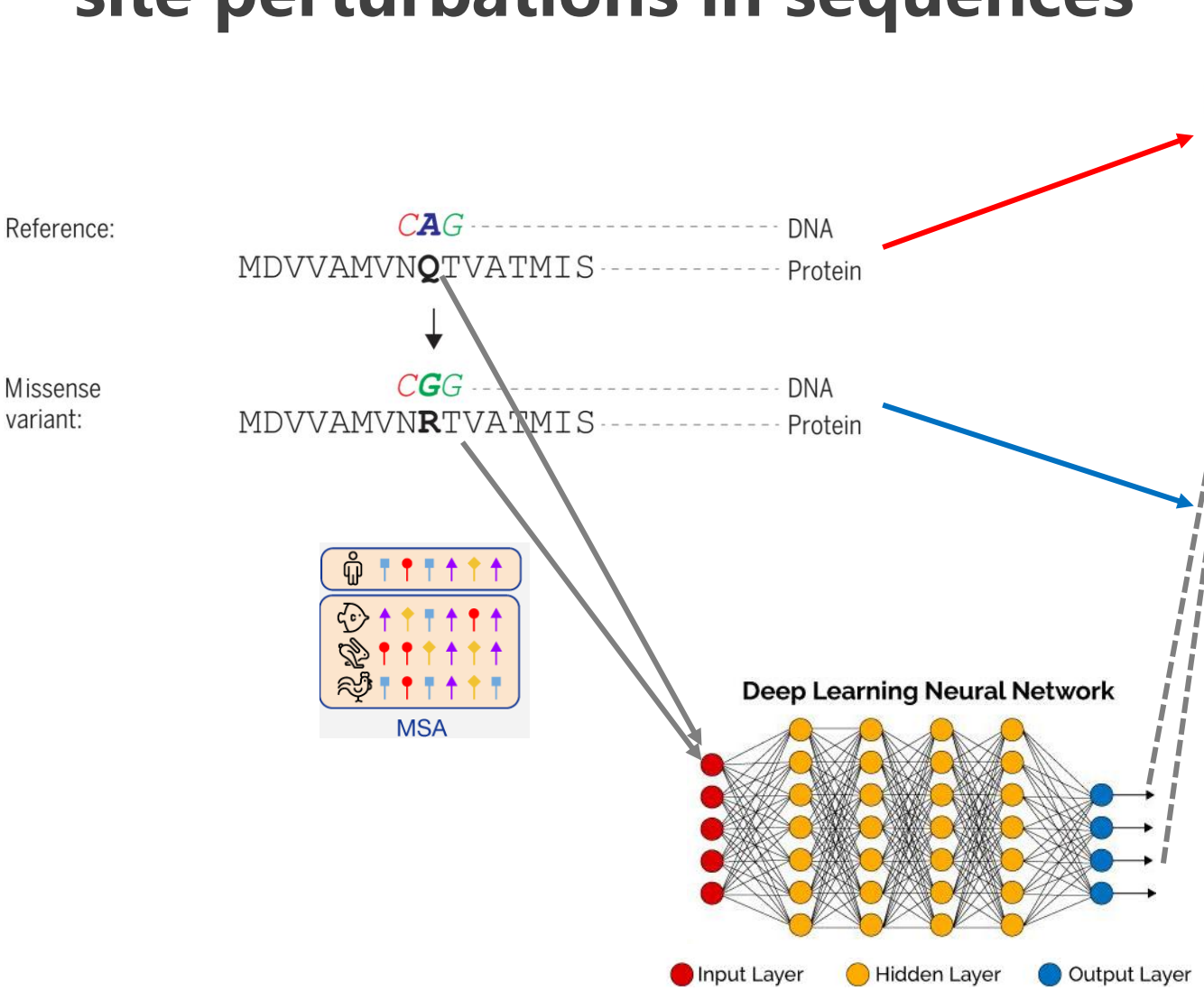
Elongation of the protein sequence

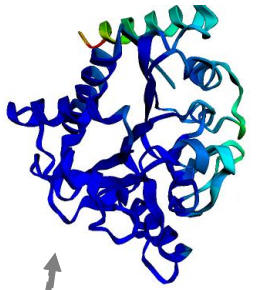
Genetic variation and disease risk



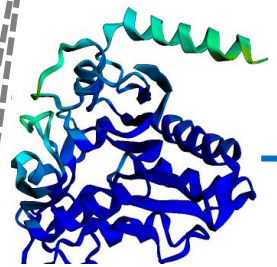
- **VUS Dilemma:** Variants of Uncertain Significance (VUS) cause diagnostic uncertainty and anxiety for patients.
- **Scale:** The vast number of potential genetic variants makes it impossible to test them all experimentally.
- **Tool Limitations:** Current predictive tools are not accurate enough for clinical use.

Most of deep learning models are insensitive to single-site perturbations in sequences





Normal function

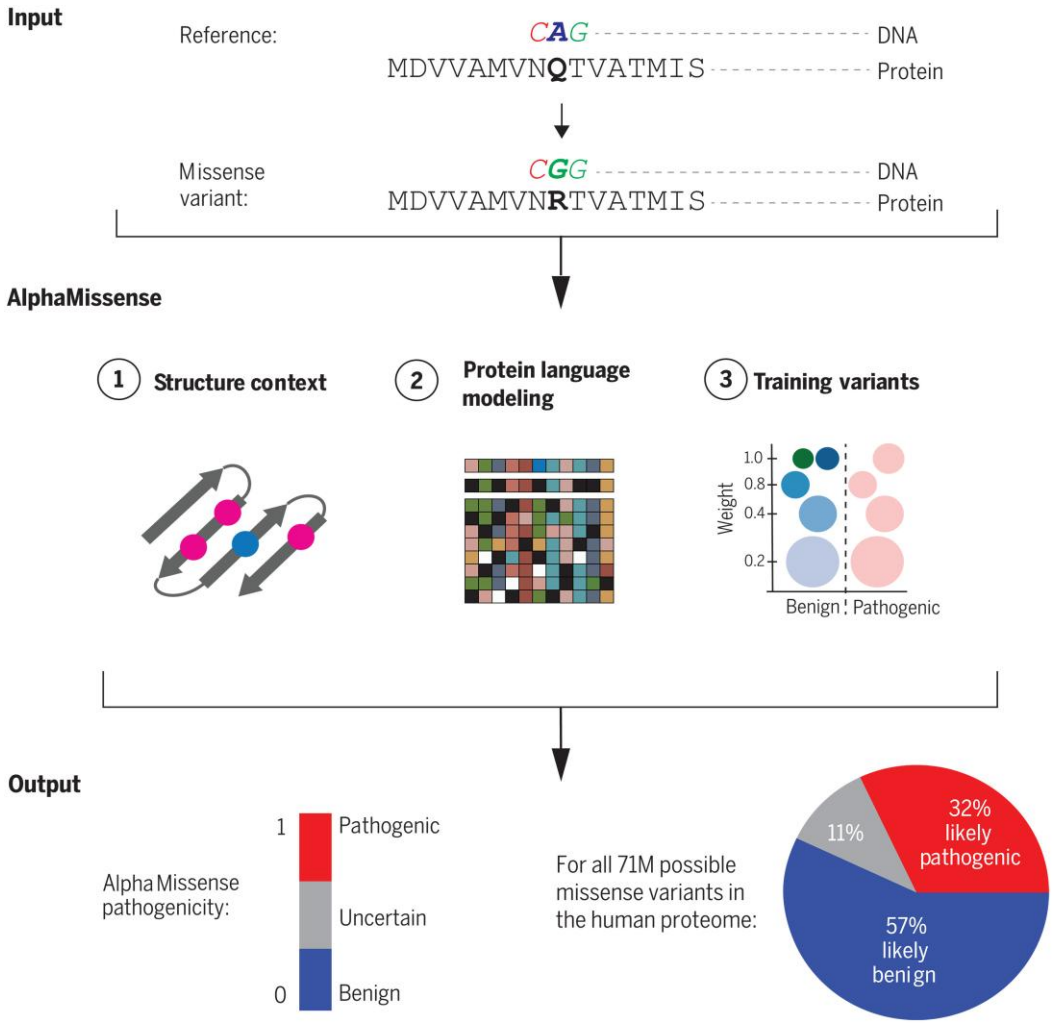


Unknown function

When similar sequences are used as input, conventional deep learning models tend to overlook minor perturbations and predict highly similar structures.

For example, AlphaFold2 calculates the MSA of sequences. For two sequences with a single-residue mutation, their MSAs are nearly identical, resulting in almost identical predicted structures.

AlphaMissense is proteome-wide missense variant effect classification model



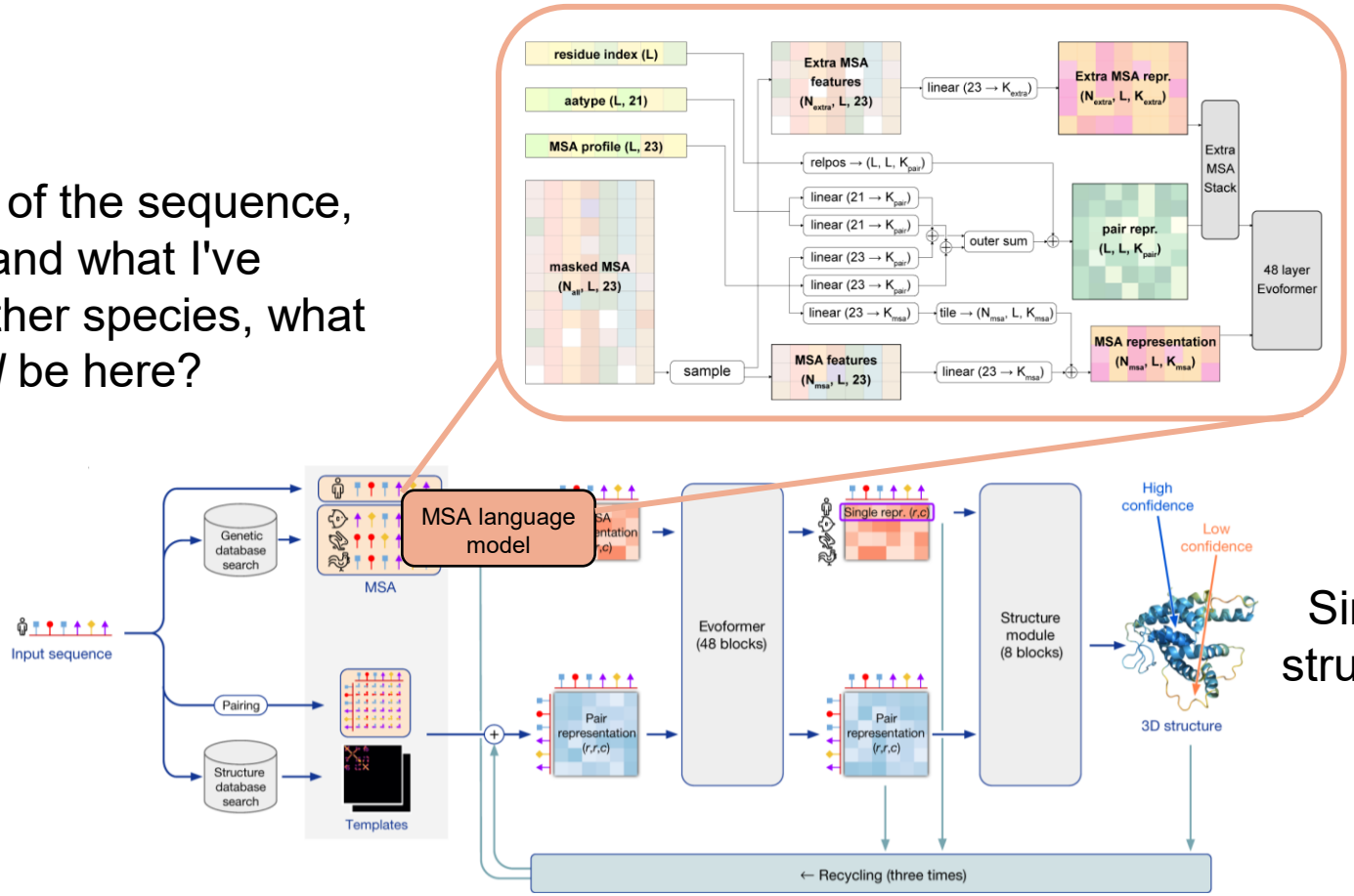
- AlphaMissense model to predict missense variant effect
- AlphaMissense is built upon the technology of **AlphaFold**. This is crucial because a protein's function is determined by its 3D shape.
- The core idea is, however, it does not predict/use the structures, but directly learns these patterns buried in amino sequences (similar to ChatGPT)

(*Science*, 2023)

AlphaMissense have two training stages

Stage1: the network is trained to perform single-chain structure prediction along with protein language modeling by predicting the identity of the amino acids masked at random positions in the MSA

Based on the rest of the sequence, the 3D structure, and what I've learned from all other species, what amino acid *should* be here?



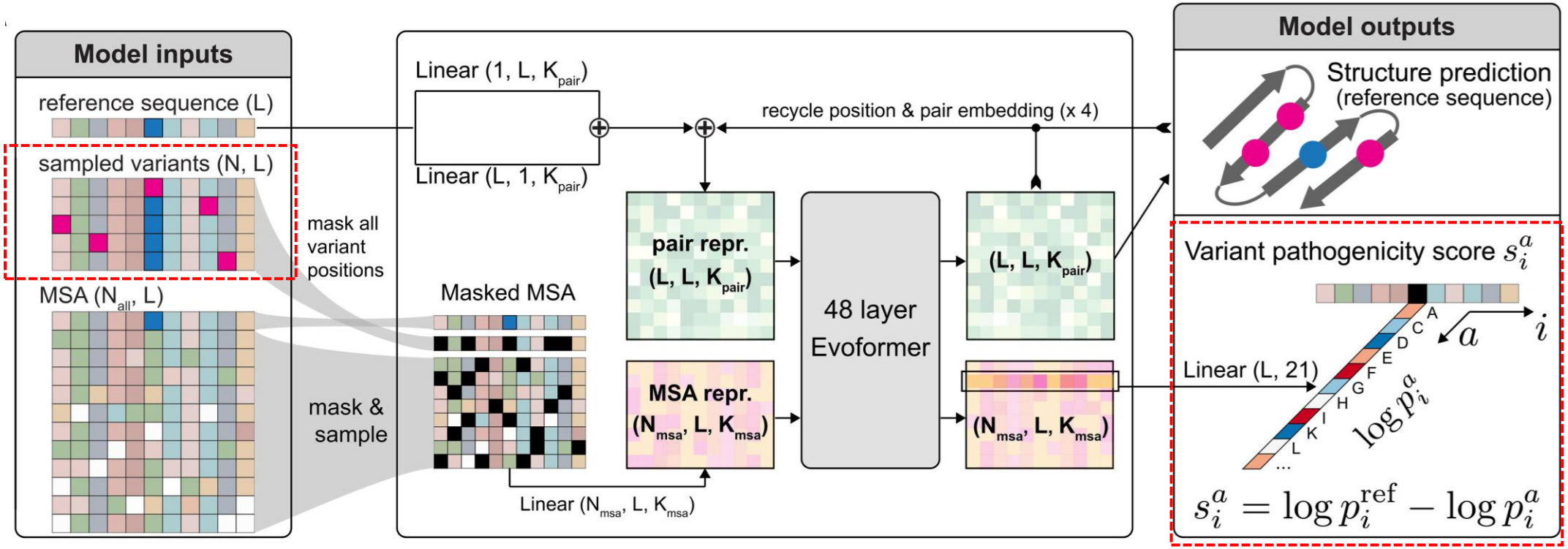
MSA mask and reconstruction

Single sequence structure prediction

AlphaMissense have two training stages

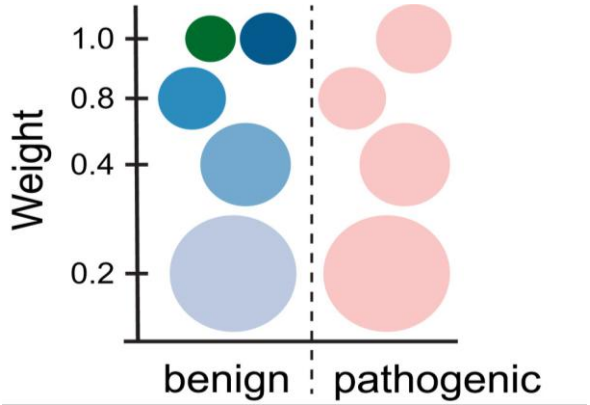
- Stage2: It compares its prediction to the *actual* mutation.

  - If the mutation results in an amino acid that the model predicted was **very unlikely** to be there, it gets a high score (closer to 1.0), suggesting it's likely **pathogenic**.
  - If the mutation results in an amino acid that the model considered a **plausible alternative**, it gets a low score (closer to 0.0), suggesting it's likely **benign**.



AlphaMissense outputs a three-dimensional protein structure with a pathogenicity score for each residue.

Label processing



Primates:
- MAF > 1e-3, total 9.7e4

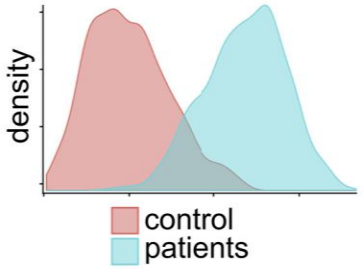
Humans:
- MAF > 2e-4, total 1.2e5
- 1e-5 < MAF < 2e-4, total 1.1e6
- sample from unobserved, match counts of benign

- The pathogenicity score → binary classification:
  - **benign** (observed or frequent missense variants in human or primate populations)
  - **pathogenic** (unobserved human missense variants)
- Split the benign variants into clusters by their minor allele frequency (MAF), introduce weights in the loss function that reduce the contribution of rare variants.
- Positive and negative sample: For each observed variant in the benign set, sampled a missense variant (1:1) from the pathogenic set and assign it the same loss weight as for the benign variant

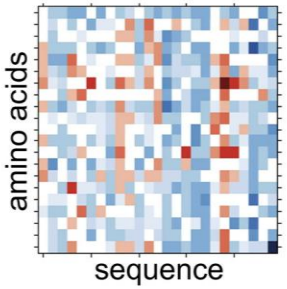
AlphaMissense performs the best performance on three different benchmark datasets



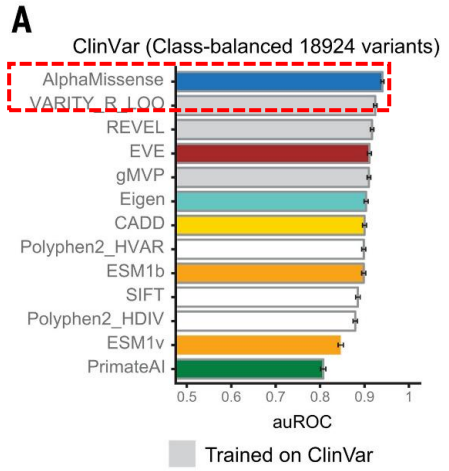
Known benign and pathogenic variants (ClinVar)

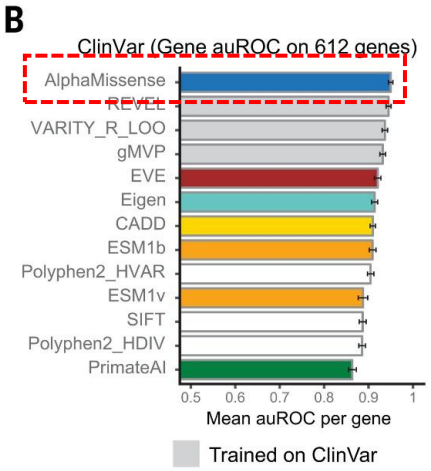


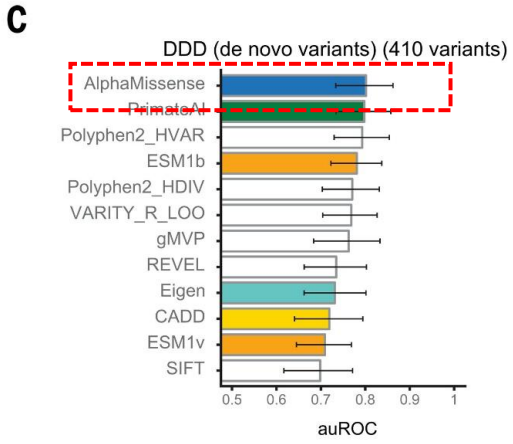
De novo variants of rare disease

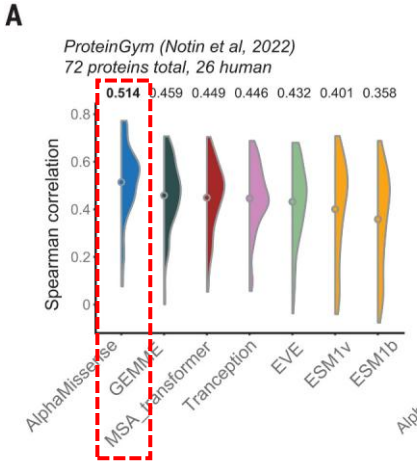


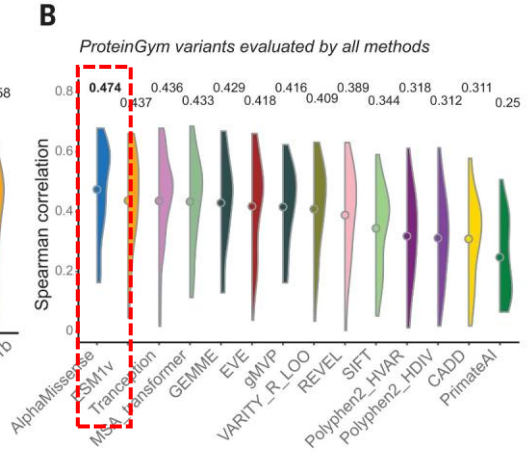
Multiplexed assay of variant effect (ProteinGym)



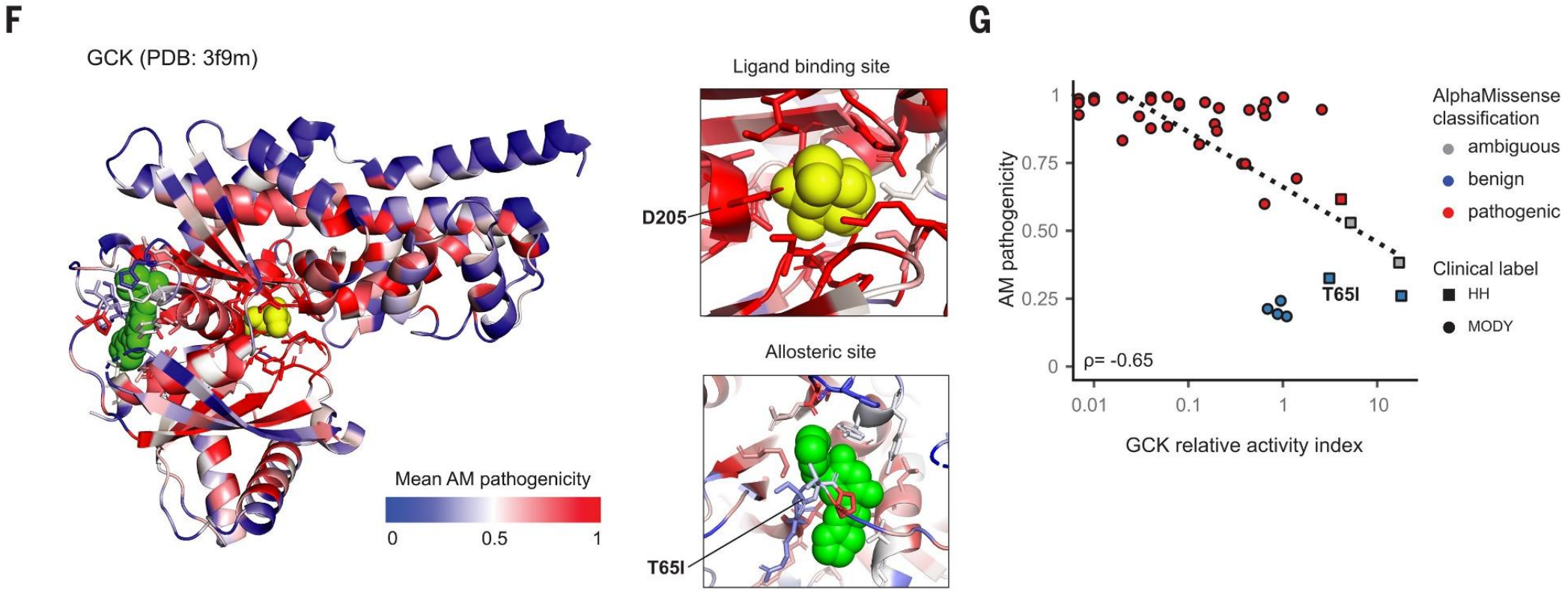








AlphaMissense predictions show a strong correlation with both the active site of GCK and its relative activity index



(*Science*, 2023)