STAT 992: Science of Large Language Models

Genomics foundation models

Spring 2026

Zhexuan Liu

DNA Foundation Models

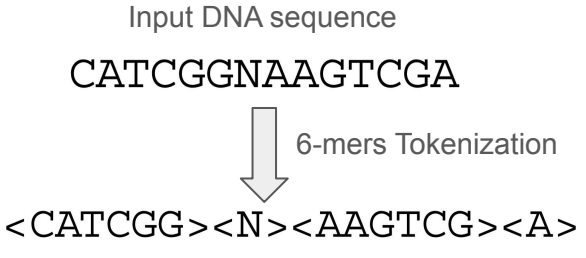
LLM concepts and genomics equivalents

Concepts	Large Language Models (LLMs)	DNA Foundation Models
Corpus	Massive text datasets (web scrapes, books, Wikipedia, code)	Genomes (reference + many individuals / species)
Document	An article, a book chapter, a webpage, or a chat thread	A genomic window (e.g., 200 bp - 100 kb)
Tokens	Words, subwords (BPE, WordPiece), or characters	Nucleotides, k-mers, or BPE subsequences
Pretrain Objective	Masked language modeling (MLM) or next-token prediction	Masked modeling or next-token prediction
Downstream Tasks	Translation, summarization, sentiment analysis, Q&A, …	Promoters/enhancers, TF binding, splicing, variant effect

Table initially generated by Gemini-3

Tokenization

- Nucleotides
  - Vocabulary: A, C, G, T, N (unknown/ambiguous nucleotide)
  - Models: HyenaDNA, …
- K-mers
  - Vocabulary: Combinations of nucleotides.
  - Example: 3-mers (<AAA>, <AAC>, <ACG>, …)
  - Models: Nucleotide Transformer, DNABERT, …
- BPE subsequences
  - Vocabulary: Learned variable-length DNA “subwords” (frequent substrings)
  - Example: Start with {A, C, G, T, N} -> learn tokens like <CG>, <ACG>, …
  - Models: DNABERT-2, GROVER, …



Dalla-Torre, Nucleotide Transformer: building and evaluating robust foundation models for human genomics, 2024

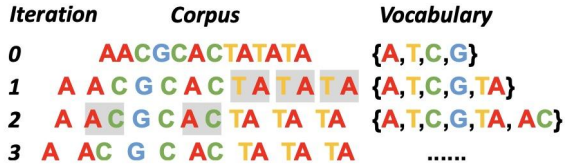


Figure 2: Illustration of the BPE vocabulary constructions.

Zhou, DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes, 2024

Downstream applications

- Zero-shot
  - **Variant effect**: Zero-shot scoring of reference vs. alternate allele via changes in model likelihood
  - …
- Fine-tuning / Supervised
  - **Regulatory Element Classification**: Identifying enhancers, promoters, etc.
  - **Epigenetic / Chromatin Marks**: Profiling histone marks, multi-label chromatin profiles.
  - **TF Binding & Prediction**: TF-related chromatin feature profiling.
  - …
- Sequence generation
  - Generate de novo DNA or autocomplete / infill missing segments from the pretrained model distribution.
  - Virtual cell?

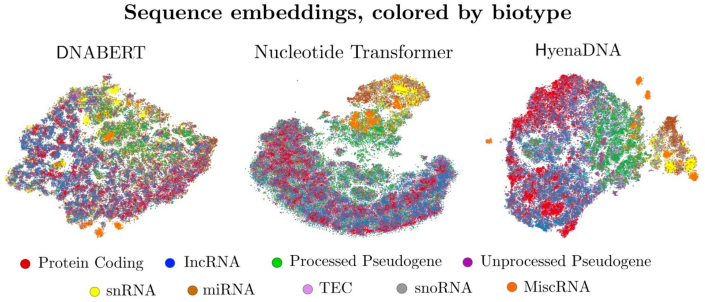
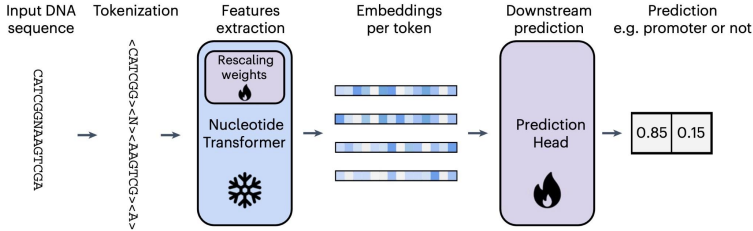


Figure 4.3: **Embedding visualisation.** t-SNE of the embeddings generated by DNABERT, Nucleotide Transformer and HyenaDNA coloured by Ensembl biotype annotations.

Nguyen, HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution, 2023

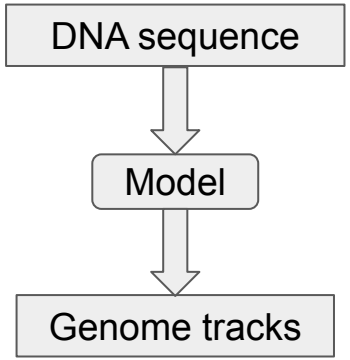


Dalla-Torre, Nucleotide Transformer: building and evaluating robust foundation models for human genomics, 2024

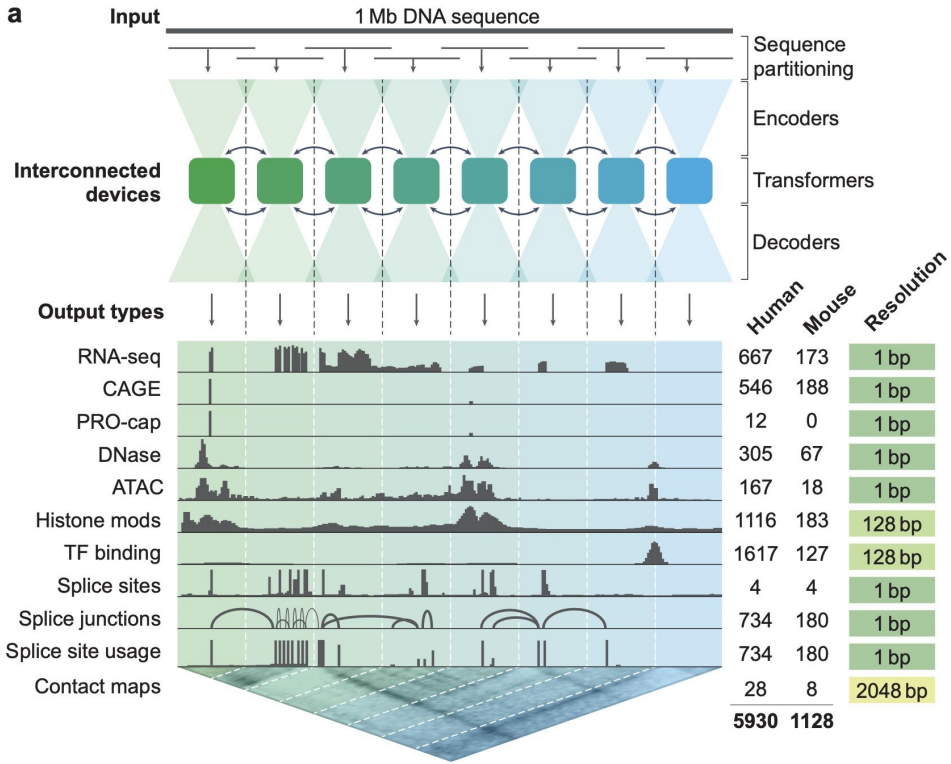
DNA Sequence-to-function Models

DNA Sequence-to-function models

Model	Release year	Context length	Architecture
Basenji	2018	~131 kb	CNN
Enformer	2021	~196 kb	CNN + Transformer
AlphaGenome	2025	1 Mb	CNN + Transformer



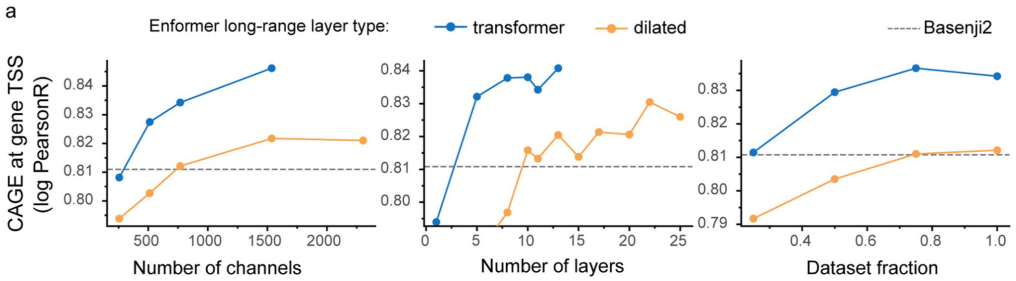
* One genome track: one experimental measurement of genomic regulation across the genome

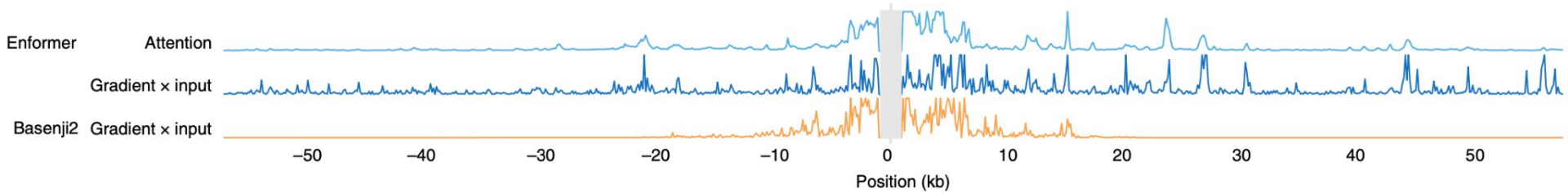


Avsec, Advancing regulatory variant effect prediction with a unified DNA sequence model, 2026

Ablation study by Enformer

- Superior long-range interaction by transformer.
  - If the attention layers are replaced with dilated convolutions, we notice a significant performance drop.





Avsec, Effective gene expression prediction from sequence by integrating long-range interactions, 2026

Single-cell Foundation Models

LLM concepts and genomics equivalents

Feature / Dimension	Large Language Models (LLMs)	Single-Cell Foundation Models
Corpus	Massive text datasets (web scrapes, books, Wikipedia, code)	Large-scale single-cell transcriptomic atlases (millions of cells)
Document	An article, a book chapter, a webpage, or a chat thread	A single cell (its complete gene expression profile)
Tokens	Words, subwords (BPE, WordPiece), or characters	Genes (often ranked by expression level) or gene-expression bins
Pretrain Objective	Masked language modeling (MLM) or next-token prediction	Masked gene modeling (predicting masked genes or expression values)
Downstream Tasks	Translation, summarization, sentiment analysis, Q&A, …	Cell type annotation, perturbation prediction, batch integration, gene network inference

Table initially generated by Gemini-3

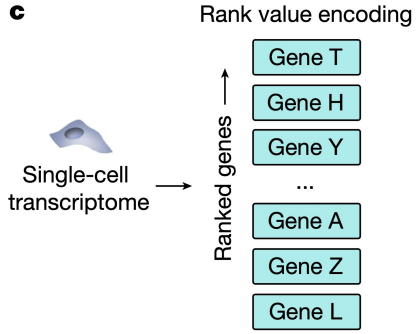
Training data

Matrix X	**Gene 1**	**Gene 2**	**Gene 3**	**Gene 4**	**...**	**Gene 20,000**
Cell 1	65	1	0	0	...	0
Cell 2	50	0	18	4	...	0
Cell 3	48	0	15	0	...	1
Cell 4	72	0	0	0	...	0
...	...	...	...	...	...	...
Cell N	55	2	19	0	...	0

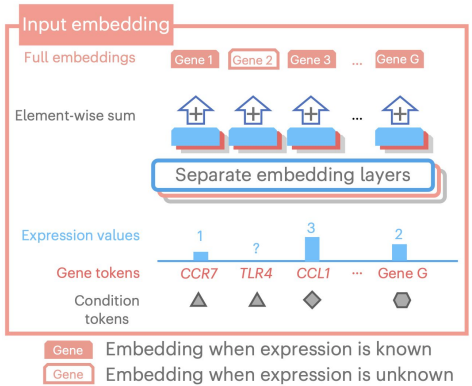
The entry in the matrix (e.g., the number 65) represents the number of RNA molecules (transcripts) from a specific gene that were physically detected inside that single cell.

Tokenization

- Rank-ordered Genes
  - Vocabulary: Gene name / symbols (unique identifier of each gene, e.g., ~20,000 to 30,000 human genes using Ensembl IDs or symbols)
  - Tokenization: Genes act as standard sequence tokens, ordered from highest to lowest expression within the cell.
  - Example: Genes ordered sequentially from highest to lowest expression in a specific cell (<MALAT1>, <B2M>, <ACTB>, …)
  - Models: Geneformer
- Genes + Binned Expression
  - Embedding strategy: $E_{input} = E_{gene} + E_{bin} + E_{condition}$
  - Models: scGPT, scBERT
- Continuous Gene Expression
  - Embedding strategy: $E_{input} = E_{gene} + E_{value}$
  - Models: scFoundation, Universal Cell Embeddings (UCE)



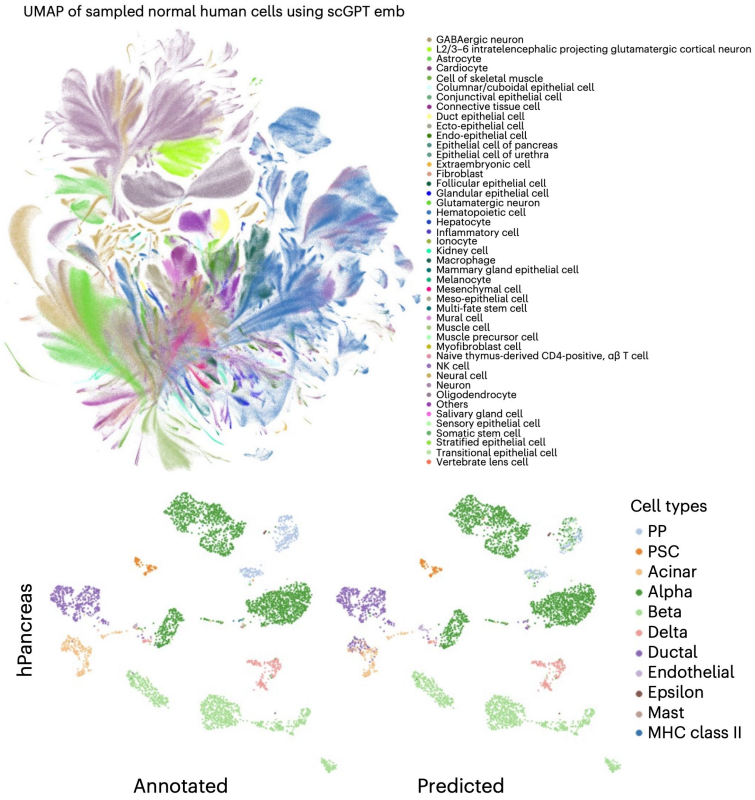
Theodoris, Transfer learning enables predictions in network biology, 2023



Cui, scGPT: toward building a foundation model for single-cell multi-omics using generative AI, 2024

Downstream applications

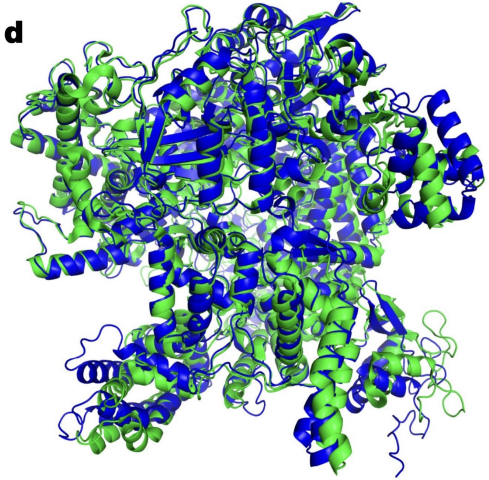
- Zero-shot
  - Cell embeddings: embed new cells and do reference mapping.
  - In silico perturbation: predicting how a cell's entire expression profile shifts when a specific gene is knocked out or overexpressed.
    - Gene deleterious effect, measured by the cosine similarity of embeddings before and after gene removal.
    - …
  - …
- Fine-tuning / Supervised
  - Cell Type Annotation.
  - Disease & Drug Response Prediction
  - …



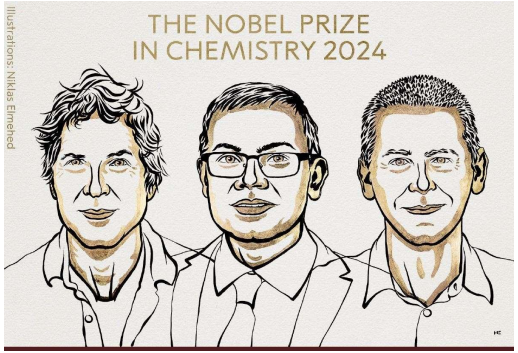
Cui, [scGPT: toward building a foundation model for single-cell multi-omics using generative AI](), 2024

Protein Models

- Protein Language Models.
  - Vocabulary: 20 amino acids.
  - Applications: Mutation effect prediction, protein family detection, …
- Structure Prediction Models.
  - Models like AlphaFold.
  - Predicts a protein's 3D structure from its amino acid sequence.
  - Applications: structure-based drug discovery, protein engineering, …



Jumper, Highly accurate protein structure prediction with AlphaFold, 2021



The Controversy: Are Foundation Models Truly Outperforming Baselines?

DNA Foundation Models

- **Expectation**: Scaling up sequence language models will automatically decode the complex, long-range "grammar" of gene regulation.
- **Reality**: Recent benchmarks reveal that DNA foundation models frequently struggle to outperform simple, task-specific CNNs trained entirely from scratch.

Single-cell Foundation Models

- **Expectation**: Massive models internalize profound biological rules to predict perturbation effects and identify novel cell types.
- **Reality**: Recent paper shows that deep-learning perturbation models failed to beat simple linear baselines on some tasks.

Patel, DART-Eval: A Comprehensive DNA Language Model Evaluation Benchmark on Regulatory DNA, 2024

Ahlmann-Eltze, Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines, 2025
Souza, Parameter-free representations outperform single-cell foundation models on downstream benchmarks, 2026

Course Survey

Link: https://forms.gle/FTeqgn3cvL8iLUFZ8