DNA - the code of life

- **DNA:** Deoxyribonucleic acid - double helix structure
- Composed of **four nucleotides:** Adenine (**A**), Cytosine (**C**), Guanine (**G**), Thymine (**T**)
- Base pairing: **A-T** and **C-G**
- Human genome: **~3 billion base pairs**, in 23 chromosomes
- **Computational representation:** Sequences of A, C, G, T (string data)

Introduction to genomics

Oxford dictionaries

"The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping** of **genomes**."

where are the genes and other interesting bits?

how do sequences *change* over evolutionary time?

what does all the DNA *do*?

what are the physical shapes of the genome and its products?

Collins English Dictionary

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

Genetics		Genomics
Targeted studies of one or a few genes	scope ⟷	Studies considering all genes in a genome
Targeted, low-throughput experiments	technology ⟷	Global, high-throughput experiments
Clever experimental design, painstaking experimentation	hard part ⟷	Tons of data, uncertainty, computation

Genomics in medical research

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

Collins English Dictionary

How is genotype related to health phenotypes?

What's the difference between DNA in a tumor vs DNA in healthy tissue?

Can genomic data help predict what drugs might be appropriate for:

- a particular cancer patient?
- a particular genetic disorder?

Can genomic data reveal weaknesses in the defenses of pathogens?

Can genomic data help us predict what flu strains will prevail next year?

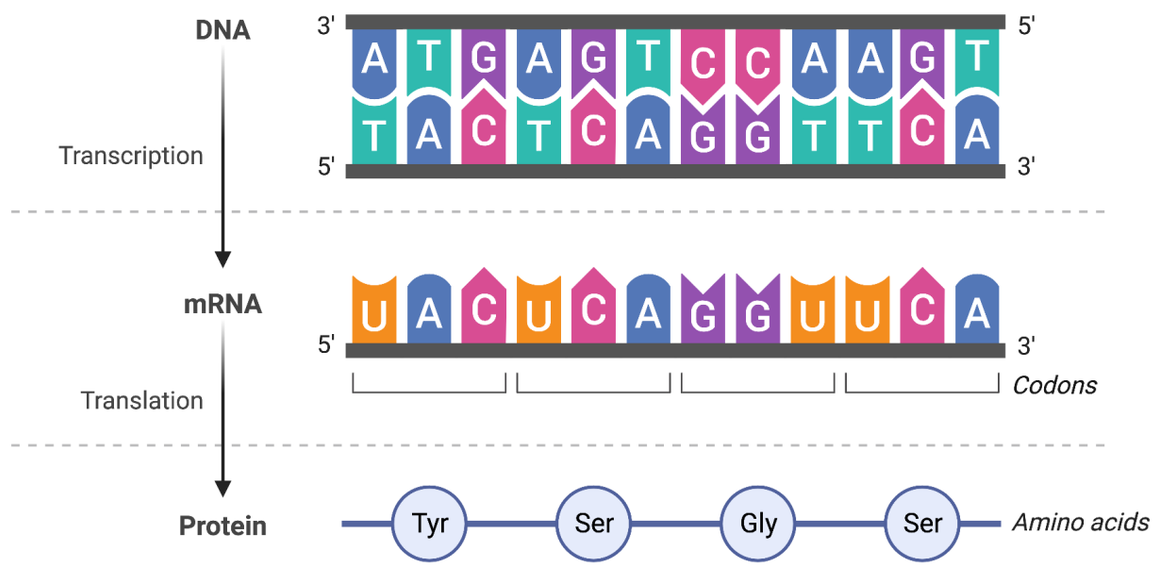
Deep Learning in Genomics: our course

- **Can we detect regulatory motifs in DNA?** → CNNs (Weeks 1–3)
- **Can we learn the "language" of DNA?** → GPT / Transformers (Weeks 4–5)
- **Can we predict gene regulation from sequence?** → Enformer (Weeks 6–7)
- **Can we model microbial communities?** → microbiome-based prediction (Weeks 8–9)

In each unit: understand the biological problem → learn the architecture → train a model → apply to real data → present results

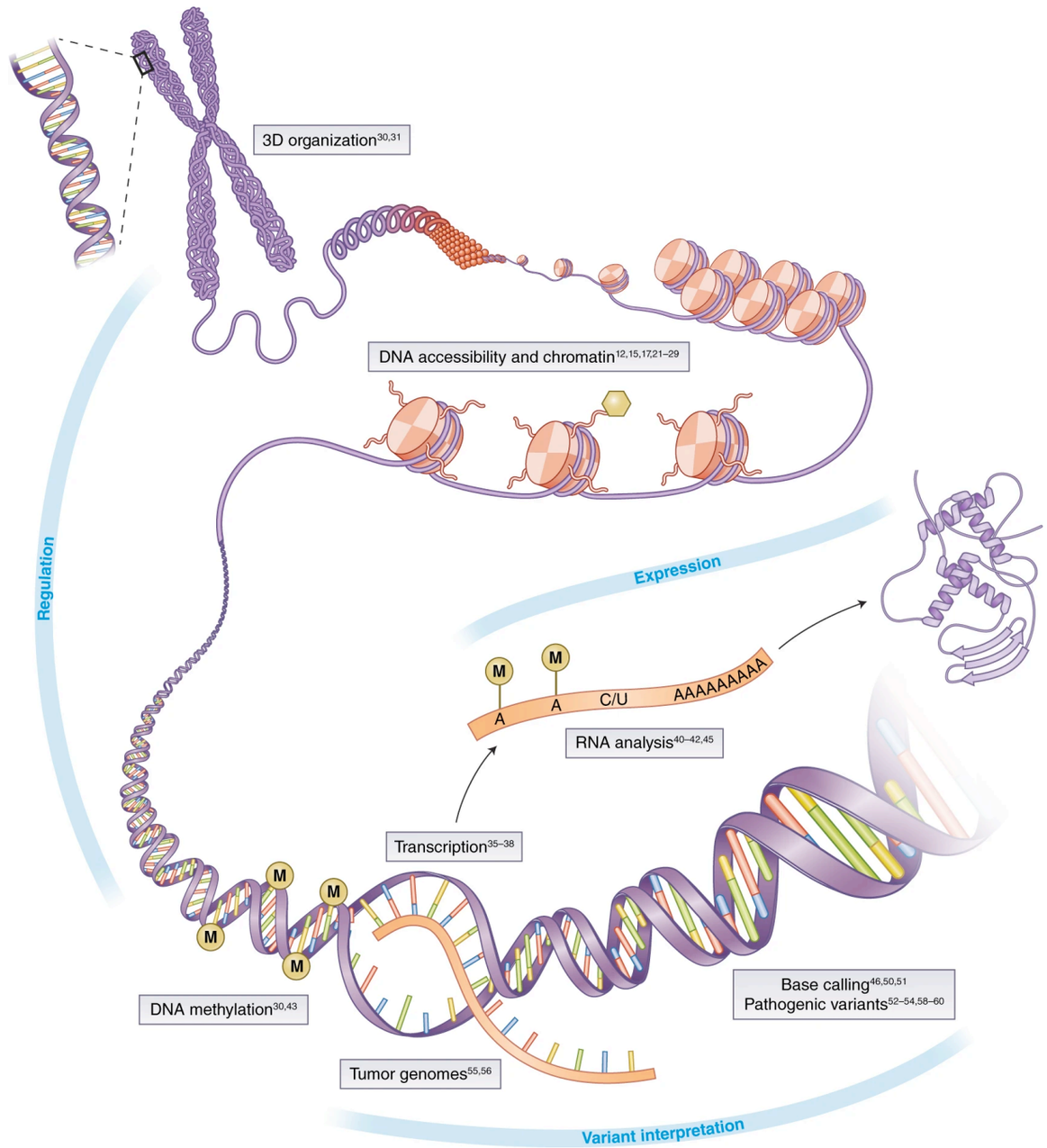
Genes and gene expression

- **Gene:** Basic unit of heredity; DNA segment that codes for protein or RNA
- **Human genome:** ~20,000-25,000 genes (~1-2% of genome)
- **Central Dogma:** DNA → RNA → Protein
- **Gene expression:** Process by which information from a gene is used to synthesize a functional product
- **Machine learning opportunity:** Predicting expression levels, regulatory effects



Genome organization and structure

- **Chromosomes:** Organized structures of DNA (humans have 23 pairs)
- **Regulatory elements:** Promoters, enhancers, silencers
- **Non-coding regions:** Once called "junk DNA," now known to be important for regulation
- **3D genome structure:** DNA folds in complex ways that affect function
- **Computational challenge:** Understanding the significance of sequence and structure



Genetic variation

- **Single Nucleotide Polymorphisms (SNPs):** Single base differences between individuals
- **Structural variations:** Insertions, deletions, duplications, inversions
- **Copy Number Variations (CNVs):** Repeated sections of the genome
- **Importance:** Associated with disease risk, drug response, phenotypic traits
- **AI application:** Variant interpretation and disease risk prediction

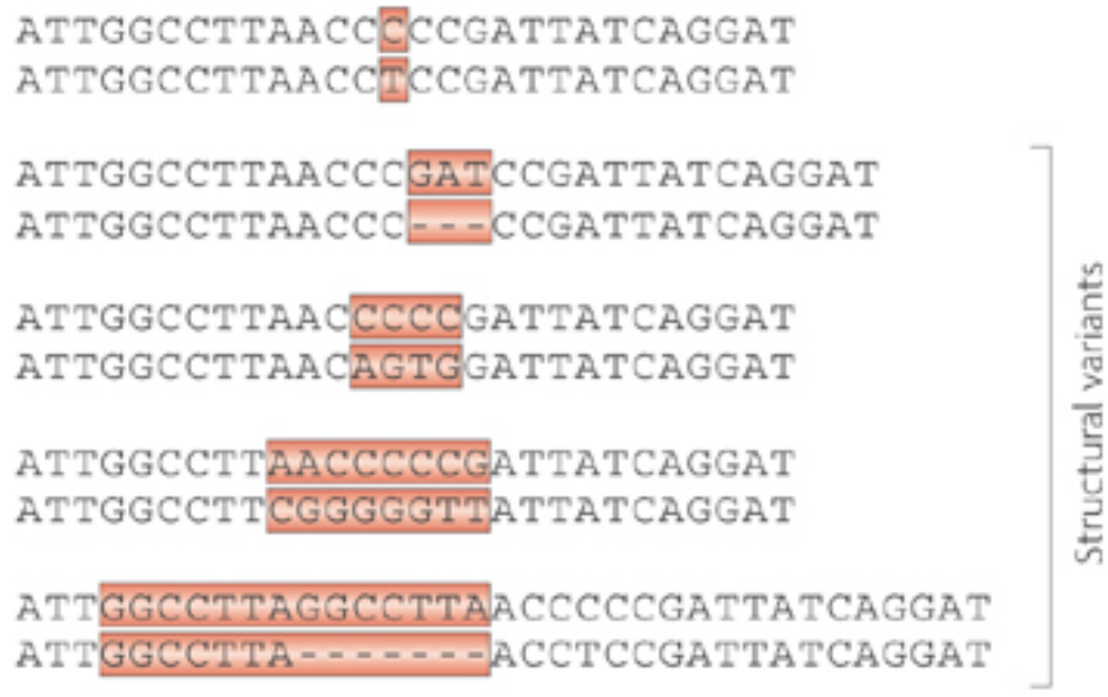
Types of genetic variation

SNP/SNV - Single nucleotide variant

Indel - Insertion–deletion variant

Structural variant (SV)

- Block substitution
- Inversion variant
- Copy number variant



Nature Reviews | Genetics

Frazer et al, Nat Rev Genet 2009
Coriell

Types of genetic variation

Nature Reviews | Genetics

SNP: Single Nucleotide Polymorphism

Frazer et al, Nat Rev Genet 2009
Coriell

Genetic variation

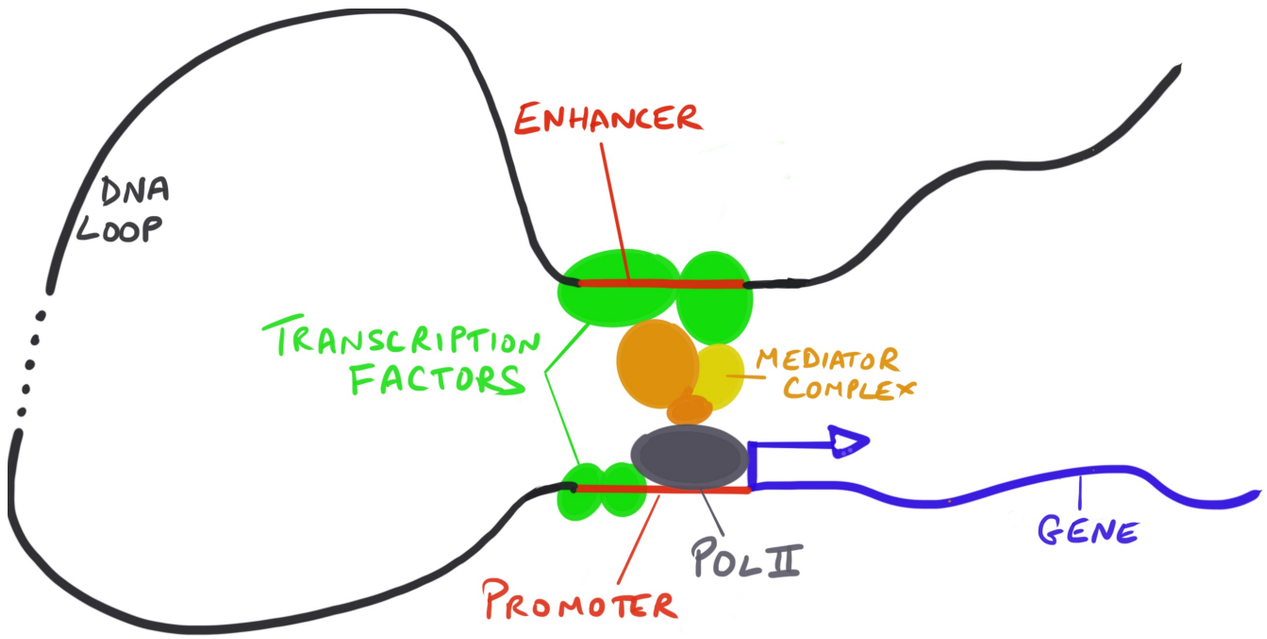
"We find that a typical [human] genome differs from the reference human genome at **4.1 million to 5.0 million sites**. Although **›99.9% of variants consist of SNPs and short indels**, structural variants affect more bases: the typical genome contains an estimated **2,100 to 2,500 structural variants** (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), **affecting ~20 million bases of sequence.**

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

- **Importance:** Associated with disease risk, drug response, phenotypic traits
- **AI applications:** Variant calling, interpretation, and disease risk prediction

Gene regulation and transcription factors



- **Promoters:** Regions near genes where transcription begins. RNA polymerase binds here to start reading the gene
- **Enhancers:** Distant DNA elements that boost gene expression. Can act from thousands of base pairs away
- **Transcription factors:** Proteins that bind to specific short DNA sequences (motifs) to activate or repress genes
- **Motifs:** Short DNA patterns (6–12 bp) recognized by TFs. The same motif can appear in many genes' regulatory regions

The microbiome

What is the microbiome?

- Trillions of microorganisms (bacteria, viruses, fungi) living in and on the human body
- The gut microbiome alone contains ~1,000 species
- Linked to digestion, immunity, mental health, and disease
- Varies enormously between individuals

Metagenomic data

- Shotgun sequencing: sequence all DNA in a sample (not just one organism)
- 16S rRNA: targeted sequencing for species identification
- Produces species abundance tables + functional annotations
- Challenge: high dimensionality, compositionality, sparsity

Genomic technologies

	Bases per day
1977: Fred Sanger	700 bp/day
↓	
1985: ABI 370 (first automated sequencer)	5000 bp/day
↓	
1995: ABI 377 (Bigger gels, better chemistry & optics, more sensitive dyes, faster computers)	19,000 bp/day
↓	
1999: ABI 3700 (96 capillaries, 96 well plates, fluid handling robots)	400,000 bp/day

Genomic technologies

	Bases per day	Years to complete a human genome
1977: Fred Sanger	700 bp/day	118,000 years
1985: ABI 370 (first automated sequencer)	5000 bp/day	16,000 years
1995: ABI 377 (Bigger gels, better chemistry & optics, more sensitive dyes, faster computers)	19,000 bp/day	4,400 years
1999: ABI 3700 (96 capillaries, 96 well plates, fluid handling robots)	400,000 bp/day	205 years

Next generation sequencing

Illumina NovaSeq X

- 16 Tb per run (17 – 48 hrs)
- 128 human genomes (at 30× coverage)

Modern genomic technologies

- **DNA Sequencing:** Reading the order of nucleotides
- **Next-Generation Sequencing (NGS):** Massively parallel sequencing
- **Single-cell technologies:** Analysis at individual cell level
- **Multi-omics:** Integrating genomics with proteomics, transcriptomics, etc.
- **Big data challenge:** genomics sequencing approaches can generate a huge amount of raw data

Types of genomic data

- **Whole Genome Sequencing (WGS):** Complete DNA sequence
- **RNA-Seq:** Gene expression measurement via RNA
- **ChIP-Seq:** Protein-DNA interactions
- **ATAC-Seq:** Chromatin accessibility
- **Hi-C:** 3D chromosome conformation
- **Metagenomics:** microbiome characterization
- **Deep learning applications:** Each data type enables specific predictive tasks - from variant calling (WGS) to regulatory networks (RNA-Seq) to 3D structure modeling (Hi-C), as well as integration of different data types

Genomic data challenges for deep learning

- **Dimensionality:** Millions to billions of features
- **Sparsity:** Important signals can be rare
- **Noise:** Biological and technical variability
- **Interpretability:** Connecting predictions to biological mechanisms
- **Data integration:** Combining heterogeneous data types

Common genomic problems addressed by deep learning

- **Variant calling:** Identifying genetic variations
- **Gene expression prediction:** From DNA sequence to expression levels
- **Functional element identification:** Finding regions with specific roles
- **Protein structure prediction:** From sequence to 3D structure
- **Disease risk assessment:** Connecting genomic patterns to health outcomes