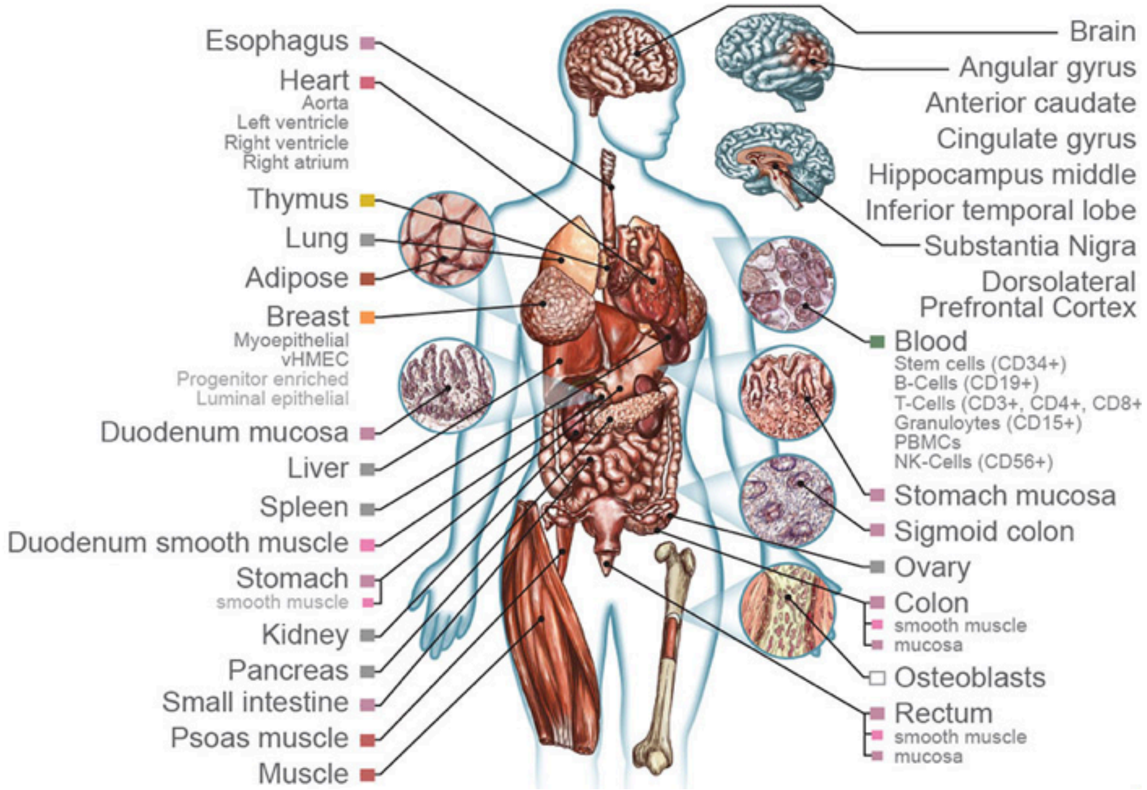
Introduction to Molecular and Cell Biology

Jian Ma

Ray and Stephanie Lane Professor of Computational Biology
Ray and Stephanie Lane Computational Biology Department
Director, Center for AI-Driven Biomedical Research
School of Computer Science
Carnegie Mellon University

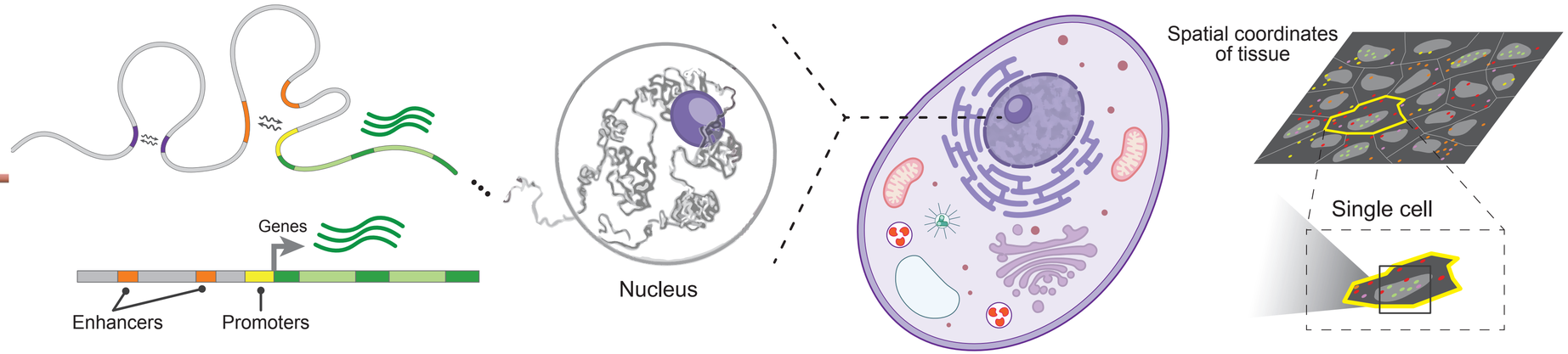
August 26, 2025 | 02541 / 02741 / 11481 / 11781

We don’t quite understand ourselves



Yao et al. *Nature* 2023

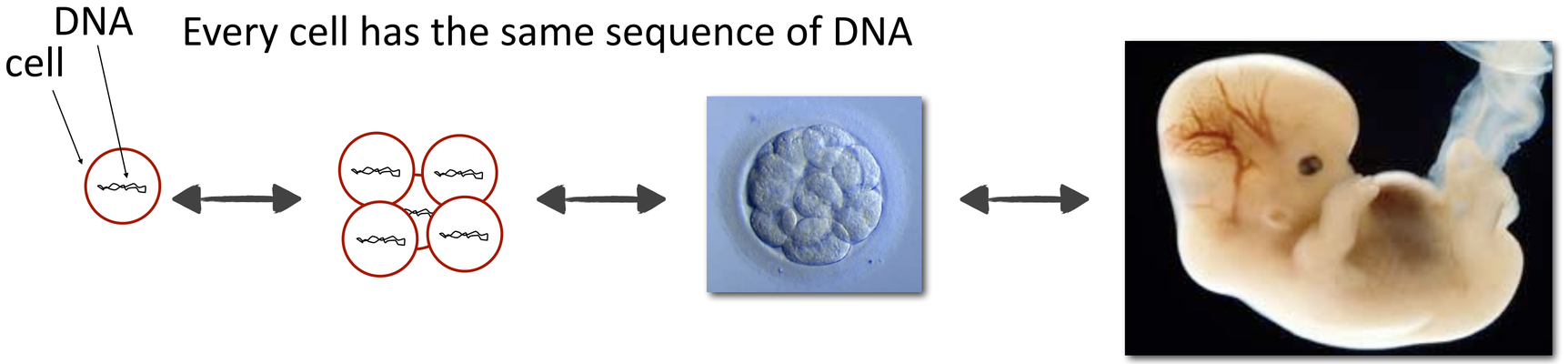
Decoding the “language” of genomes, cells, and tissues



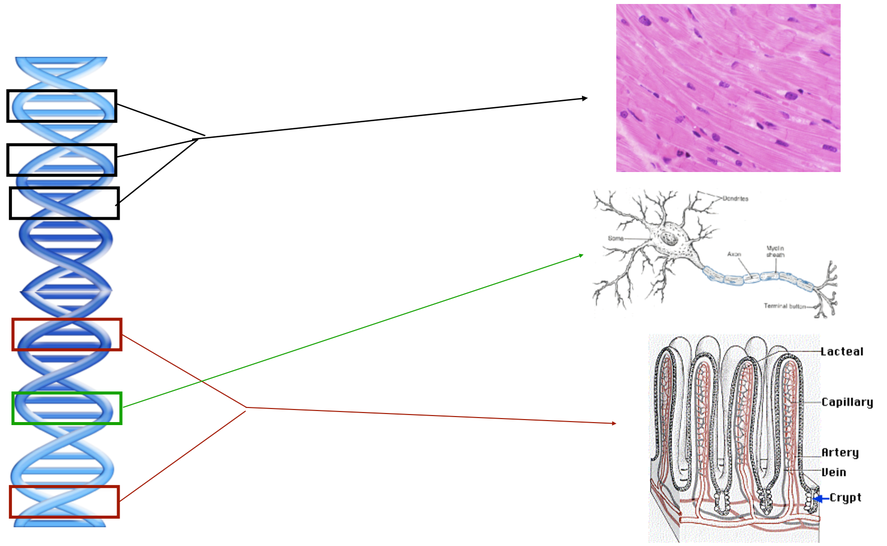
sequence — structure — function

Biology is multiscale. So must our models be.

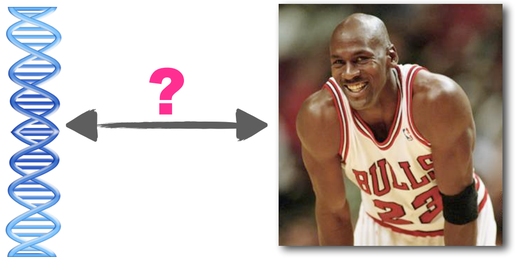
Building an organism



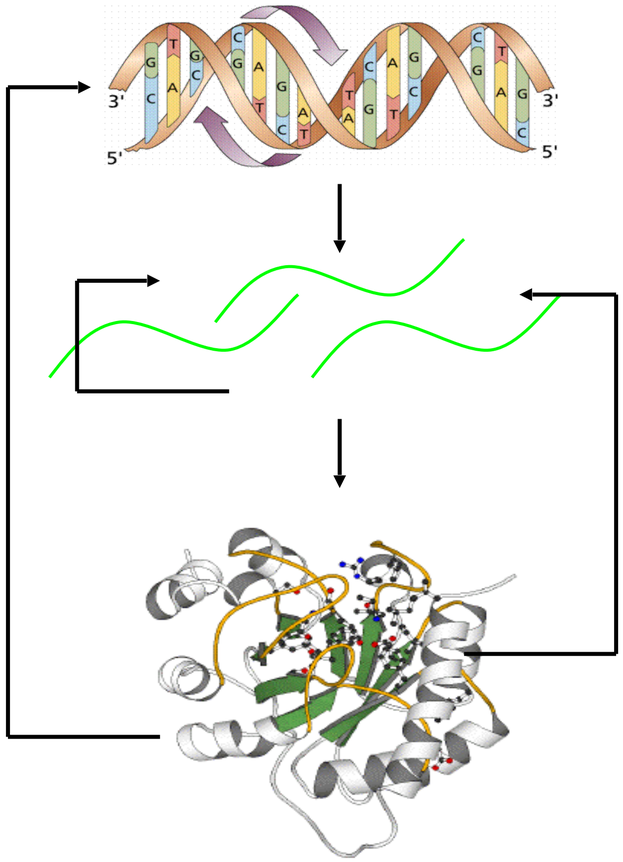
Subsets of the DNA sequence determine the identity and function of different cells



Central dogma



Proteins do most of the work in biology, and are encoded by subsequences of DNA, known as genes.



DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

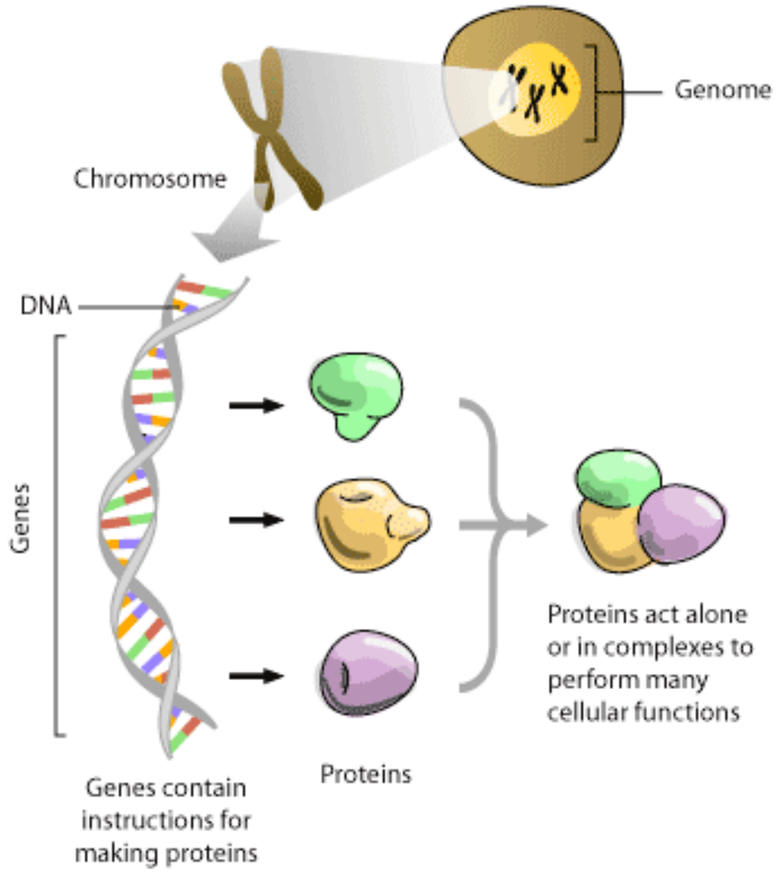
The human genome: the “blueprint” of our body

10^{13} different cells in an adult human

The cell is the basic unit of life

DNA = linear molecule inside the cell that carries instructions needed throughout the cell’s life ~ long string(s) over a small alphabet

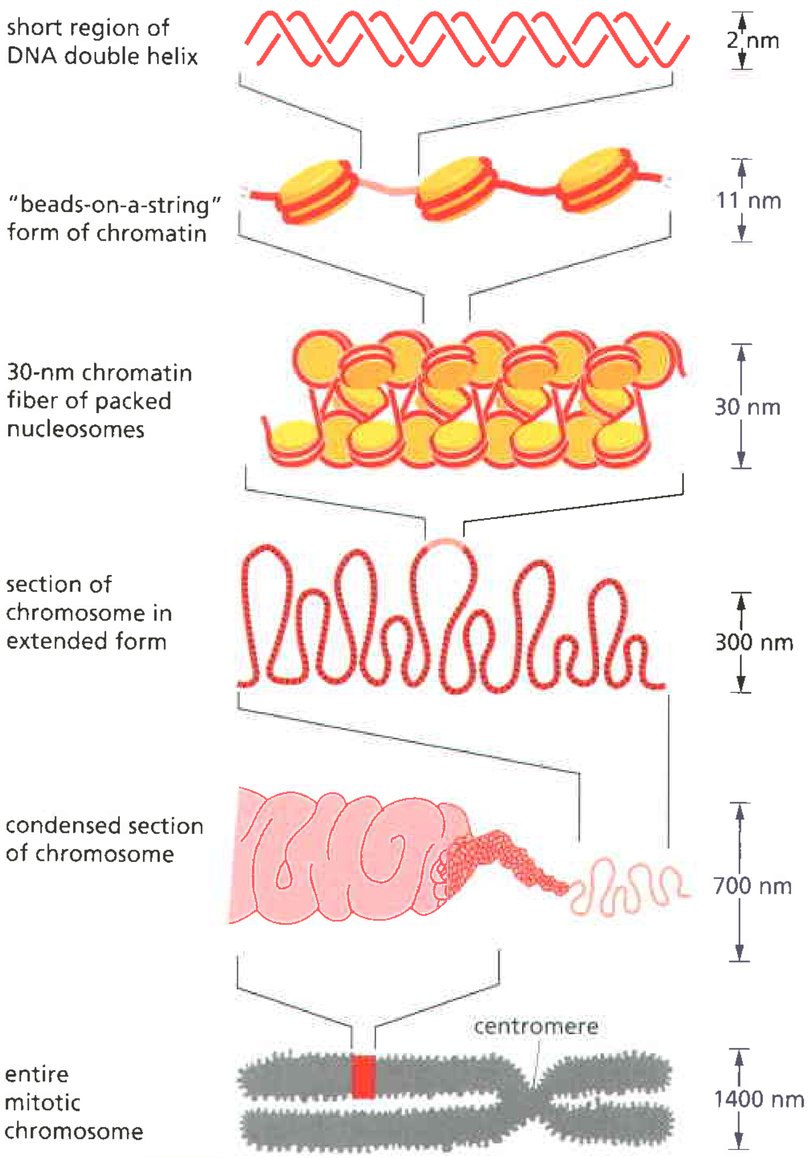
Alphabet of four (nucleotides/bases) {A,C,G,T}

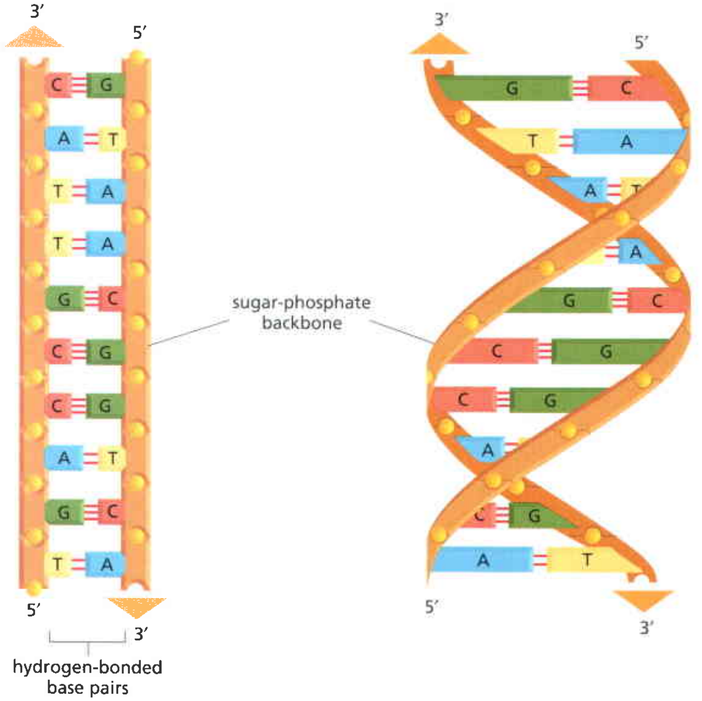


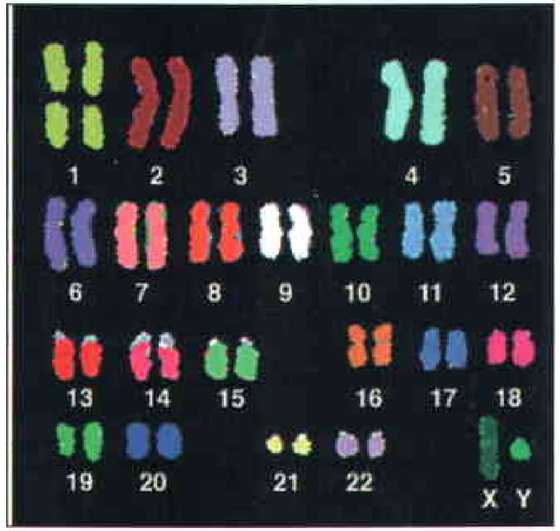
GTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGA
TTCTACAATACTAGCTTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCC
CCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATGATAATGCGATT
AGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCT
ATTAACAGATATATAAATGGAAAAGCTGCATAACCACTTTAACTAATACTTTCAAC
ATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAATT

February 15, 2001

DNA, Chromosome, and Genome



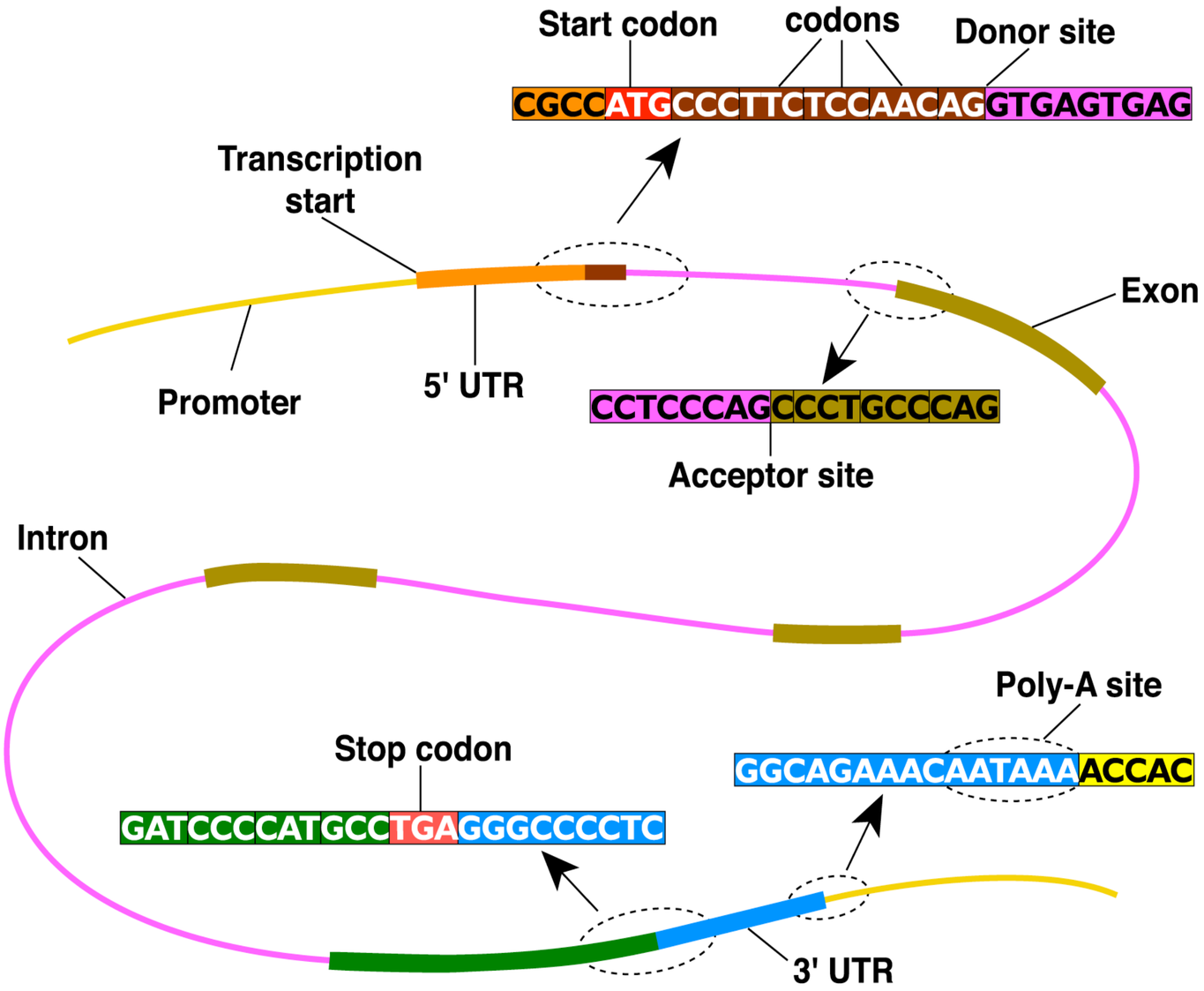




Genome

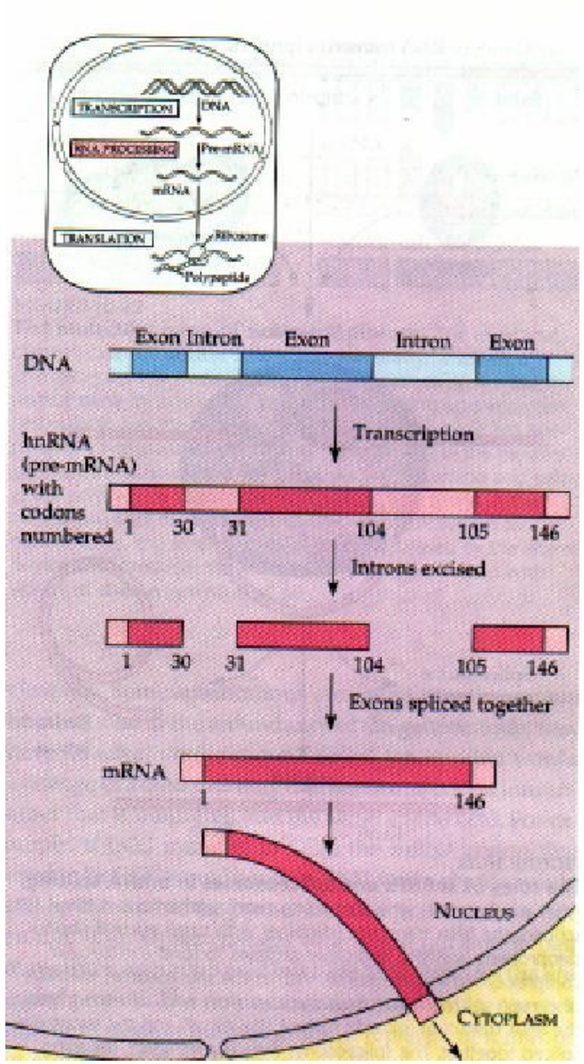
- A genome is an organism's complete set of DNA (including its genes)
- In humans, less than 2% of the genome actually encodes for genes.
- However, a much larger percentage of the genome is transcribed (miRNAs, lncRNAs, and other non-coding RNAs…)
- … and a large fractions of the rest of the genome serves as a control regions, i.e., these regions are involved in determining when genes are turned on and off depending on the context

What is a gene?



Structure of genes in mammalian cells

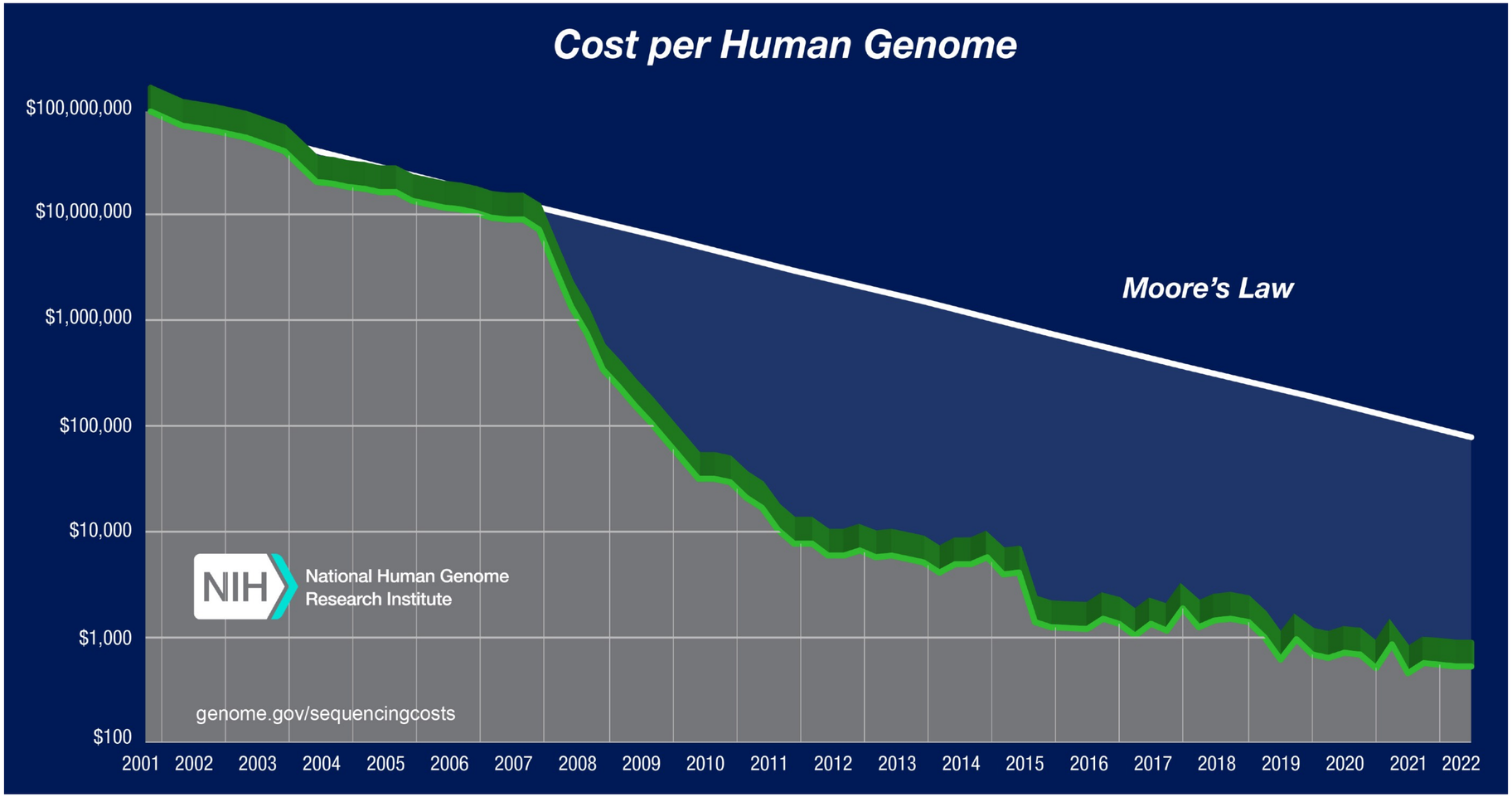
- Within coding DNA genes there can be un-translated regions (**Introns**)
- The translated parts are termed **Exons**. These are the segments of DNA that contain the protein coding information
- Following transcription Introns are removed and Exons are **spliced** to make the **protein**
- **Alternative splicing** increases the potential number of different proteins, allowing the generation of millions of proteins from a small number of genes



Genes encode for proteins

1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

High-throughput sequencing

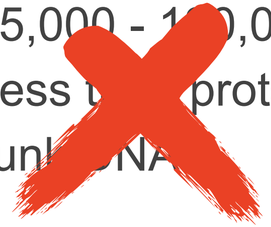


- High-throughput DNA sequencing — One of the most disruptive technologies in the past decade

Interpret the genetic code
i.e. How the Human Genome works?

Understanding the genome

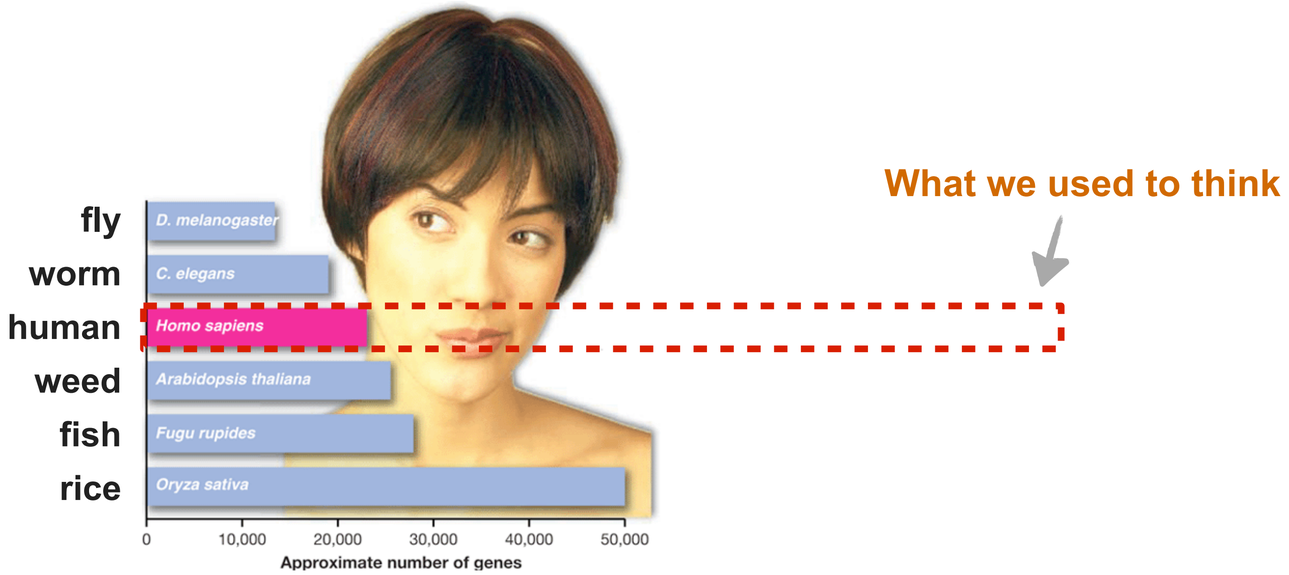
- View from 2000
  - Protein-coding genes 35,000 - 1[illegible]0,000
  - Regulatory sequence Less t[illegible] protein-coding information
  - Transposons Junk DNA



- We now know ...

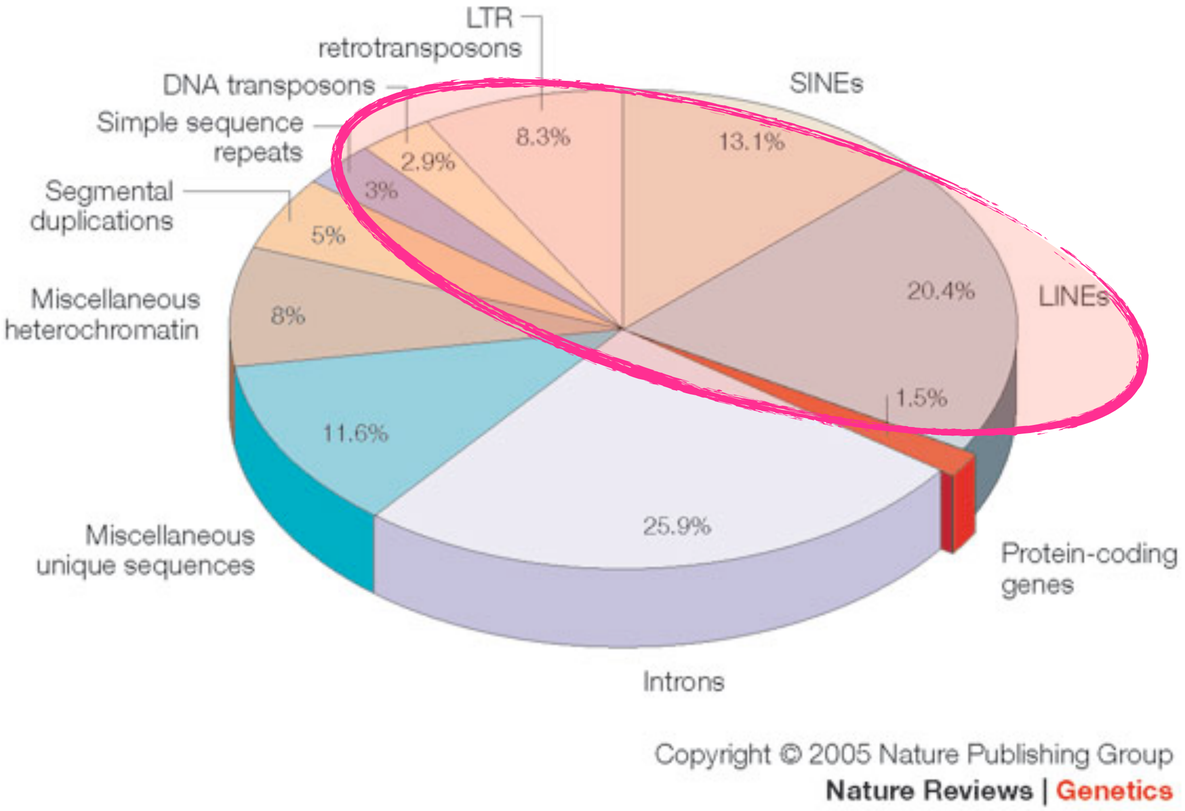
All these are WRONG!

How many genes do we have?



Gene numbers do not correlate with organism complexity.
Many gene families are surprisingly old.

Main components in the human genome



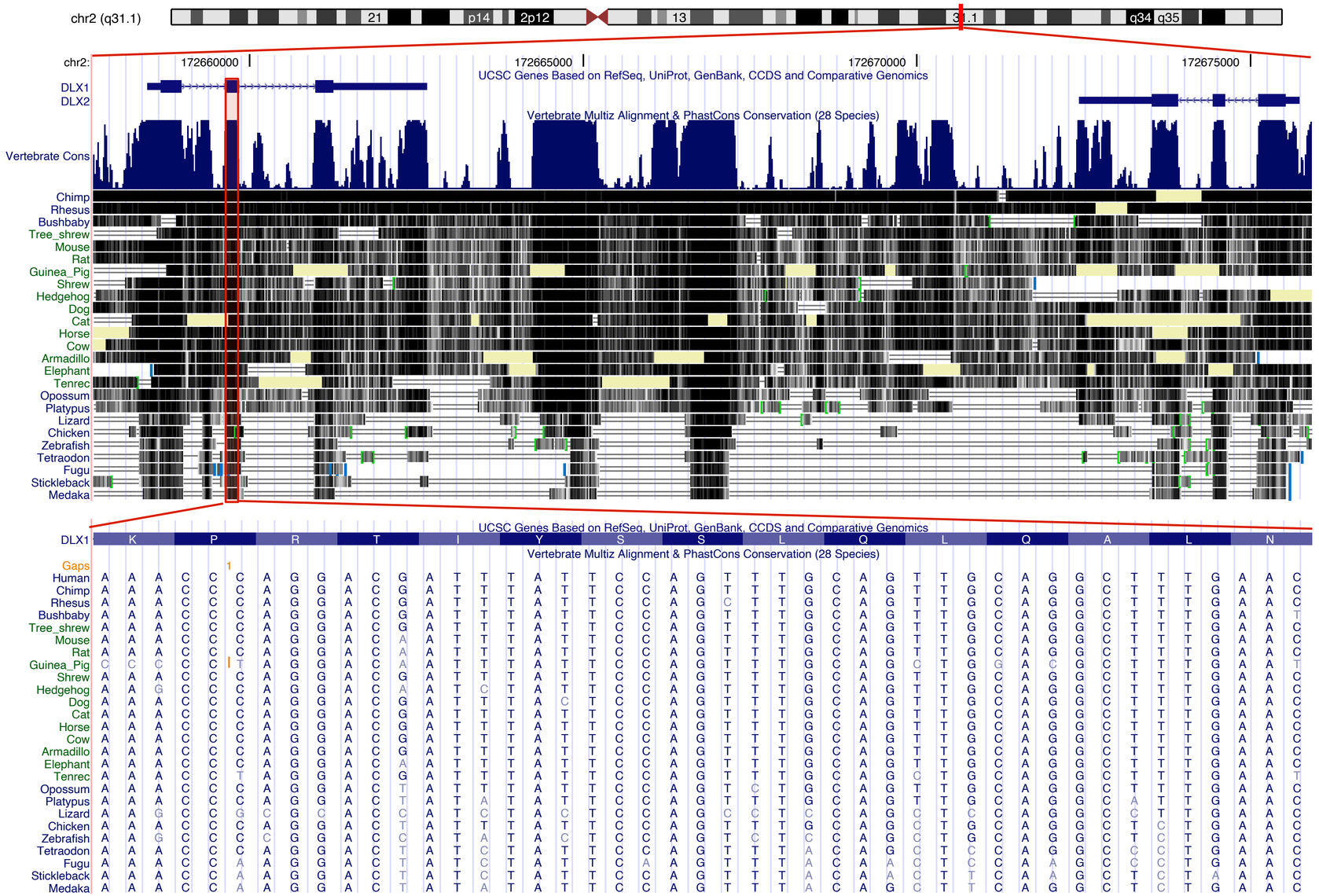
Barbara McClintock

Only 1.5% of the human genome are protein-coding regions

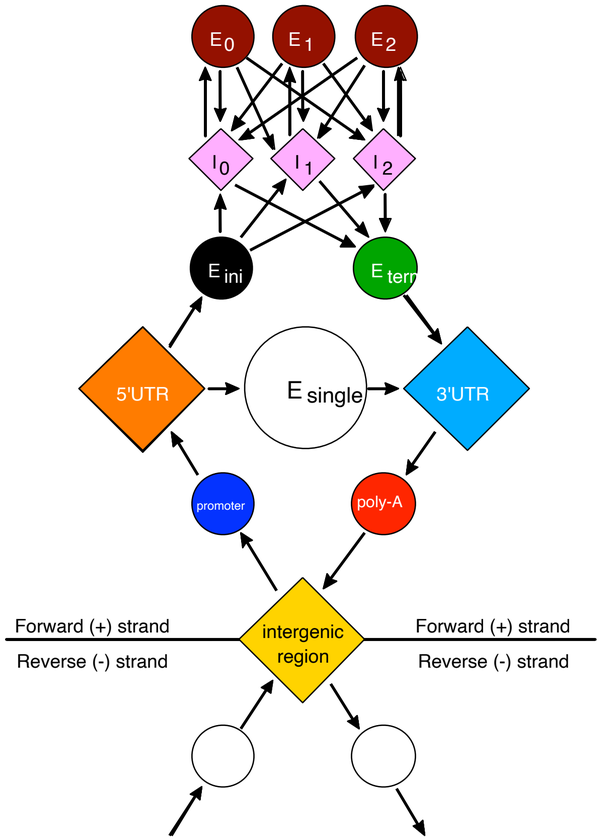
Transposable elements make up almost half of the human genome

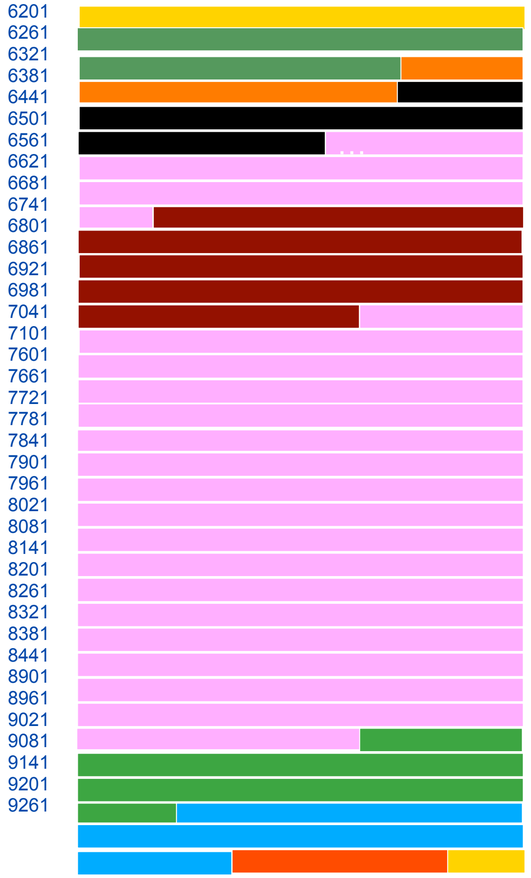
Most functional information is non-coding

- 5% highly conserved, but only 1.5% encodes proteins

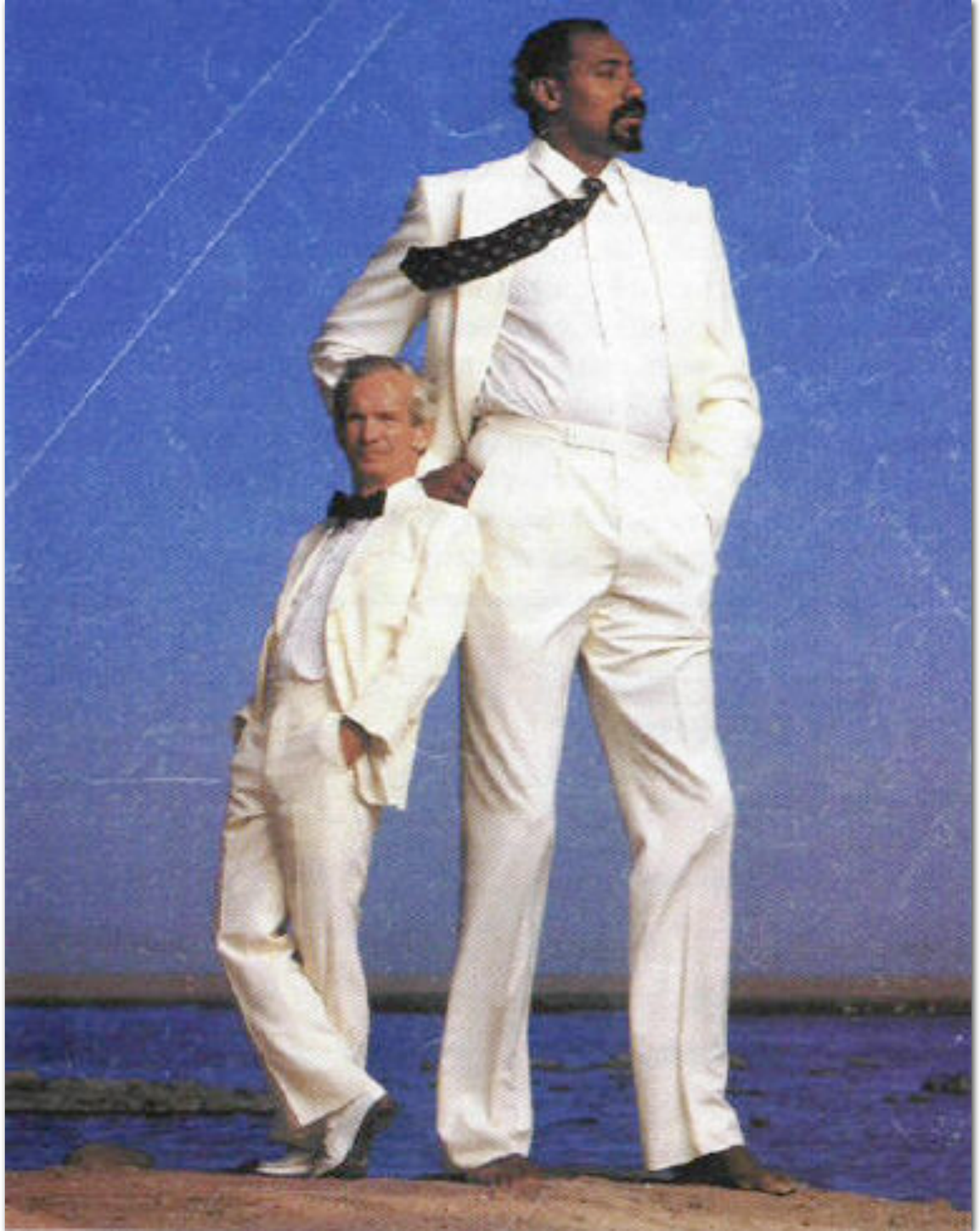


Genscan (Burge and Karlin, 1998)



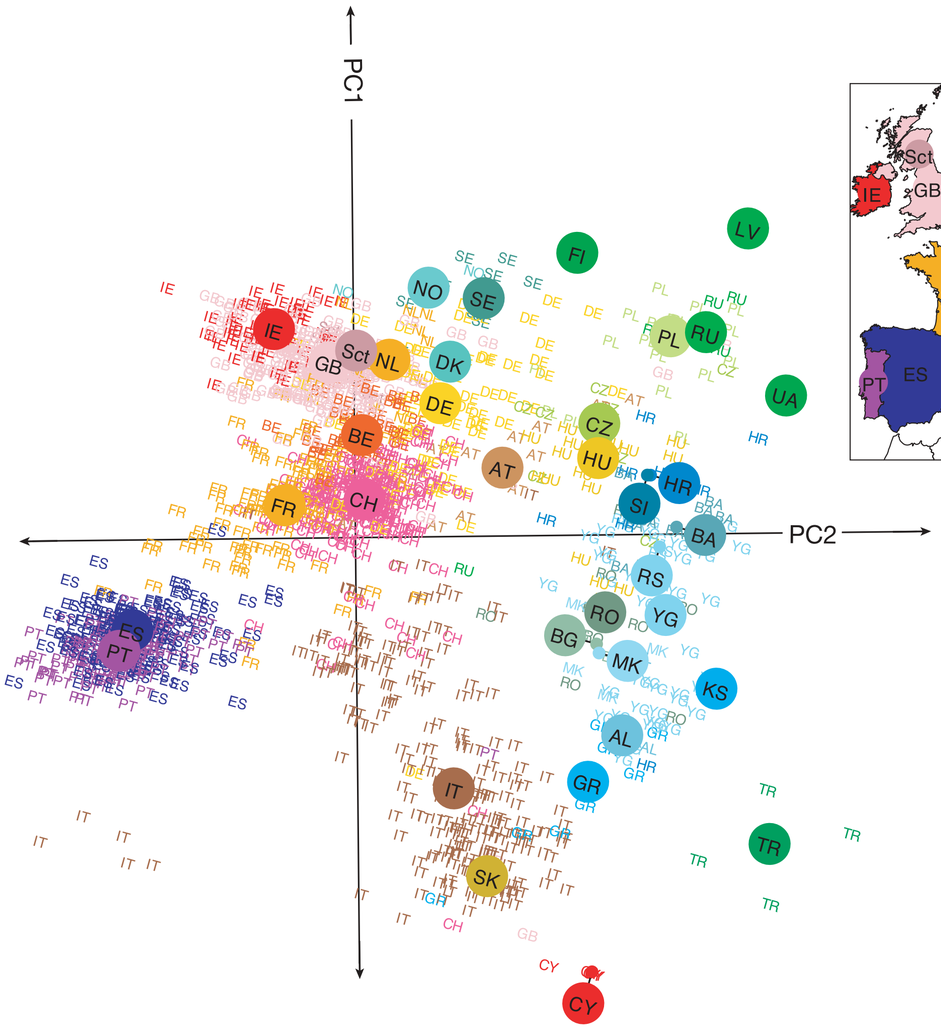


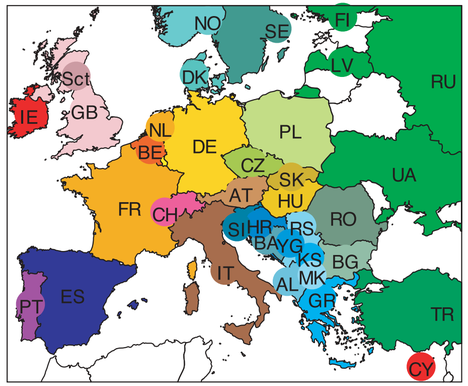

GTACT
CAAGT
CTA
TAC
CCGA
TAGG
ATCC
TGG
ATCC



The spectrum of genetic variation

Variation	Rearrangement type	Size range[a]
Single base-pair changes	Single nucleotide polymorphisms, point mutations	1 bp
Small insertions/deletions	Binary insertion/deletion events of short sequences (majority <10 bp in size)	1–50 bp
Short tandem repeats	Microsatellites and other simple repeats	1–500 bp
Fine-scale structural variation	Deletions, duplications, tandem repeats, inversions	50 bp to 5 kb
Retroelement insertions	SINEs, LINEs, LTRs, ERVs[b]	300 bp to 10 kb
Intermediate-scale structural variation	Deletions, duplications, tandem repeats, inversions	5 kb to 50 kb
Large-scale structural variation	Deletions, duplications, large tandem repeats, inversions	50 kb to 5 Mb
Chromosomal variation	Euchromatic variants, large cytogenetically visible deletions, duplications, translocations, inversions, and aneuploidy	~5 Mb to entire chromosomes



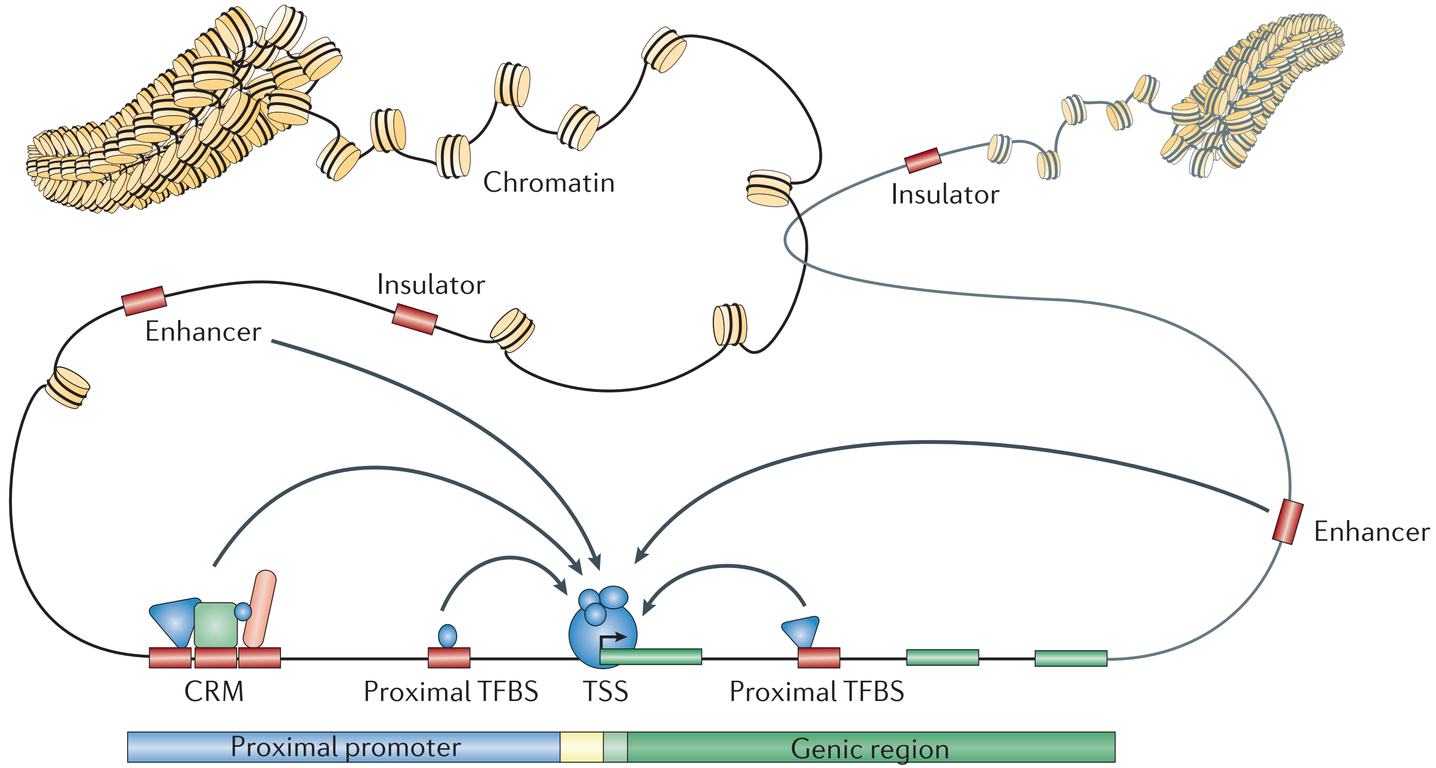


Principal component analysis

Data:
1,387 Europeans
197,146 SNPs

Novembre *et al.* Nature 2008

Transcriptional regulation

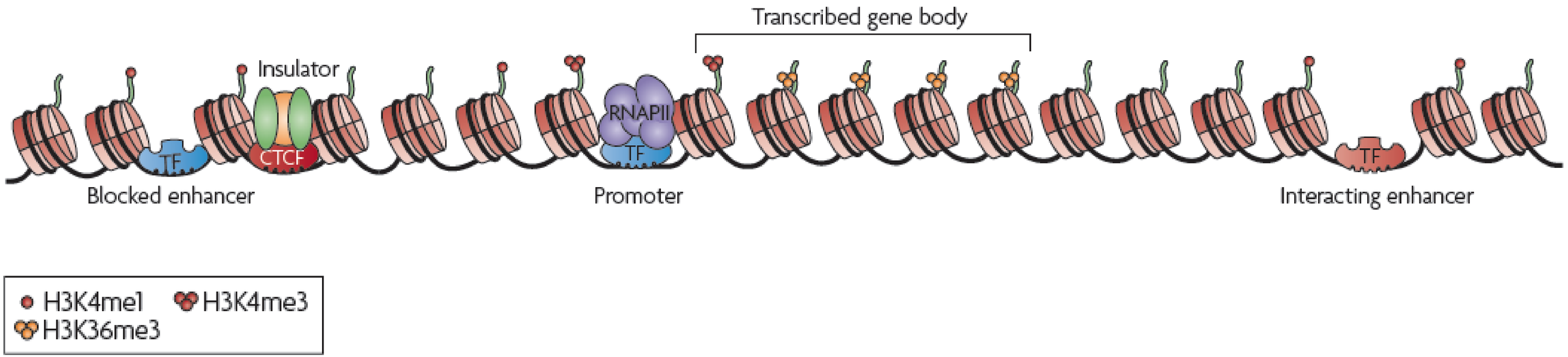


- Transcription start site (TSS)
- Transcription factor binding sites (TFBS)
- Cis-regulatory module (CRM)
- Proximal promoter and distal enhancer

Transcription factors (TFs) in the human genome

- 300 TFs bind to core promoter regions
  - General transcription machinery (e.g., RNA polymerase)
  - Required for transcription of most protein-coding genes

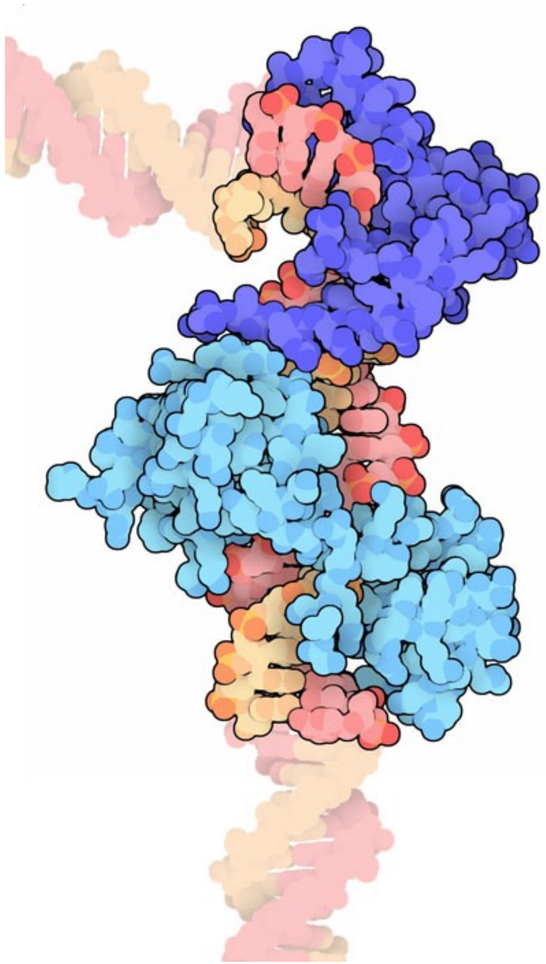
- 1500 TFs bind to other regions in the genome
  - Proximal promoter, enhancer, silencer
  - Regulate a subset of genes



How is specificity of binding achieved?

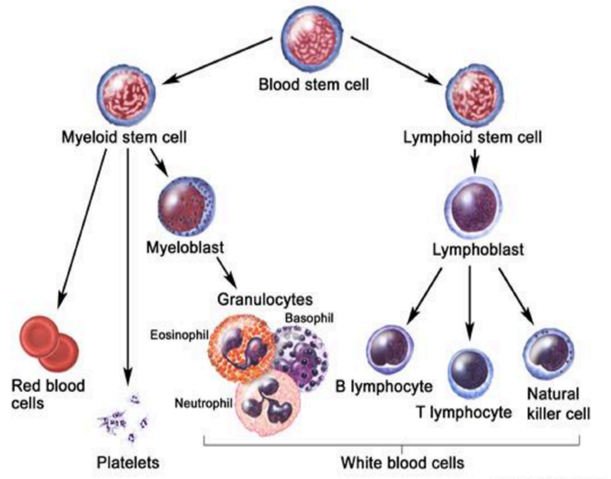
- Consensus motif
  - Commonly found sequence
  - Shows which nucleotide is most abundant at each position, represented as a Position Weight Matrix (PWM)
  - E.g., GATA1 binding motif: [AT] G A T A [AG]

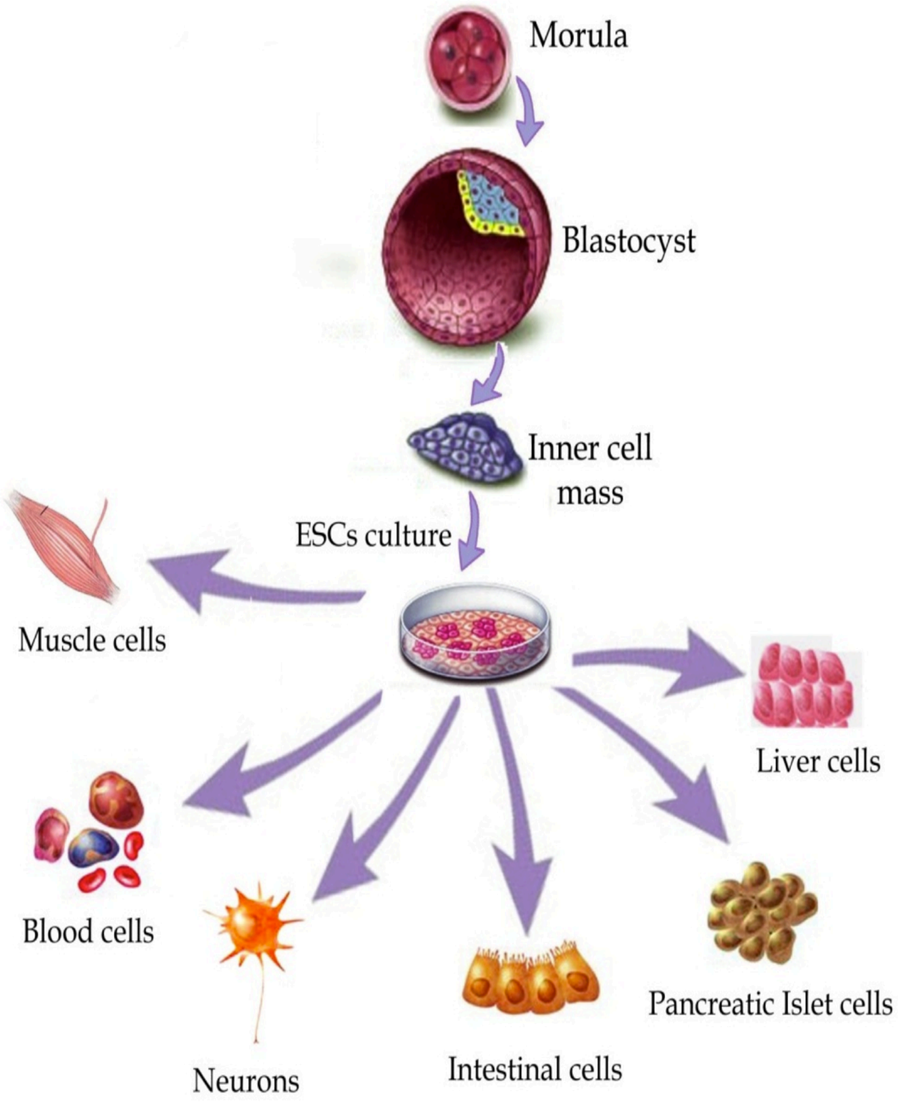
- Motif — A subsequence (substring) that occurs in multiple sequences with a biological importance.
- Motifs can be totally constant or have variable elements.
- Motifs for regulatory elements
  - Binding sites for proteins
  - Short sequences (5-25)
  - Up to certain range, e.g., 1000 bp (or farther), from gene
  - Inexactly repeating patterns (challenge!)



From one cell to trillions

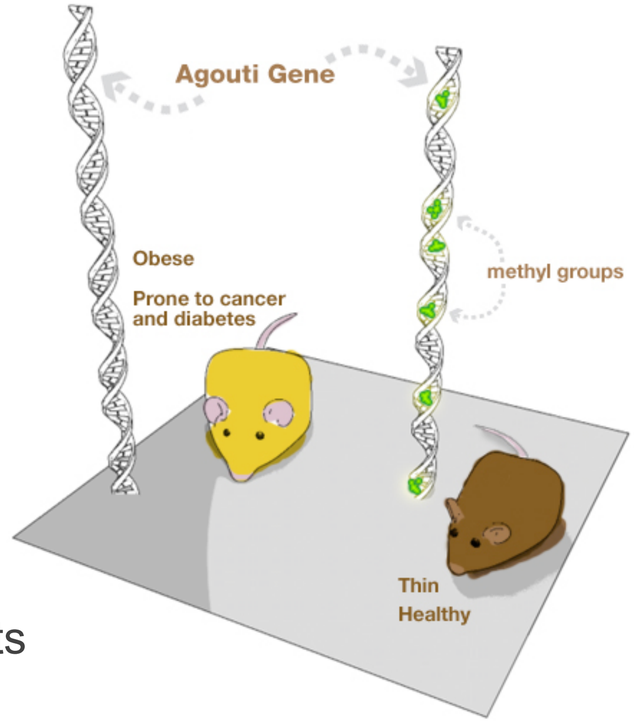
- Humans have multiple types of cells
- Individual cell types also differentiate into sub-cell types
- Lots more cell types left to discover!

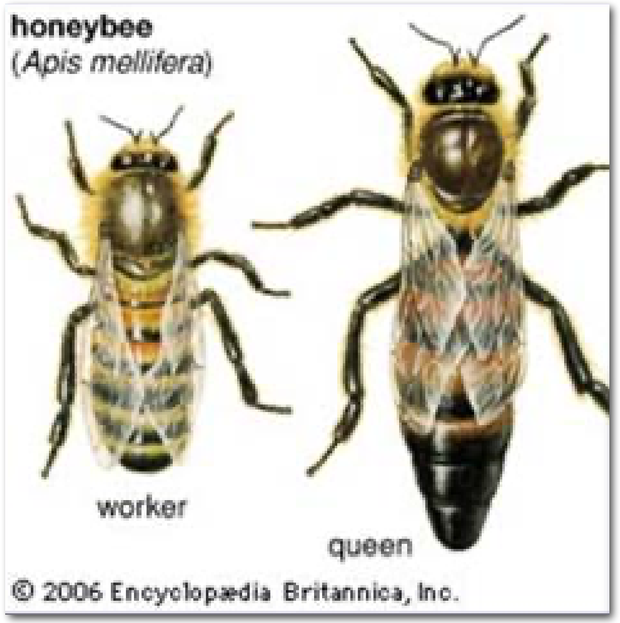




DNA is only half the story

- Variations in DNA alone may not entirely account for variations in phenotypic traits
- Organisms with identical DNA often exhibit distinct phenotypes
  - Plants, Insects, Mammals





Yellow mouse
- High risk of cancer, diabetes, obesity;
- Reduced lifespan

Agouti mouse
- Agouti gene gets methylated
- Low risk of cancer, diabetes, obesity;
- Prolonged lifespan

What is epigenetics/epigenomics?

- A mitotically or meiotically heritable state of different gene activity and expression (**phenotype**) that is independent of differences in DNA sequence (**genotype**)
  - – based on Conrad Waddington, 1942

- The sum of the alterations to the chromatin template that collectively establish and propagate different patterns of gene expression (transcription) and silencing from the same genome.

- Epigenetic changes influence the phenotype without altering the genotype.

- While **epigenetics** often refers to the study of single genes or sets of genes, **epigenomics** refers to more global analyses of epigenetic changes across the entire genome.

Epigenetic mechanisms

- DNA methylation
  - Normal cells — role in gene expression and chromosome stability
  - Cancer cells — consequences of aberrant hypo- and hyper-methylation

- Histone modification
  - Normal cells — the histone code
  - Cancer cells — consequences of altered histone modifying enzymes

- Interaction between DNA methylation, histone modifications, and other players such as non-coding RNAs

- Cell and tissue type specificity

- Gene-environment interaction, disease susceptibility

- If DNA is like the alphabet, epigenetic marks are like the accents and punctuation

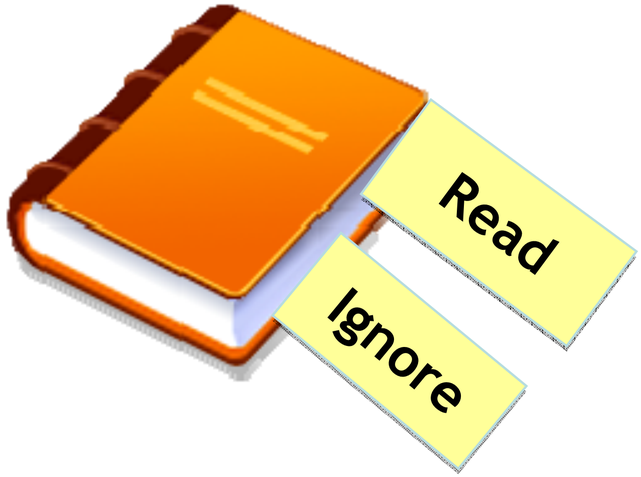
DNA sequence

TAG CAT ACT

TAG! CAT? ~~ACT~~

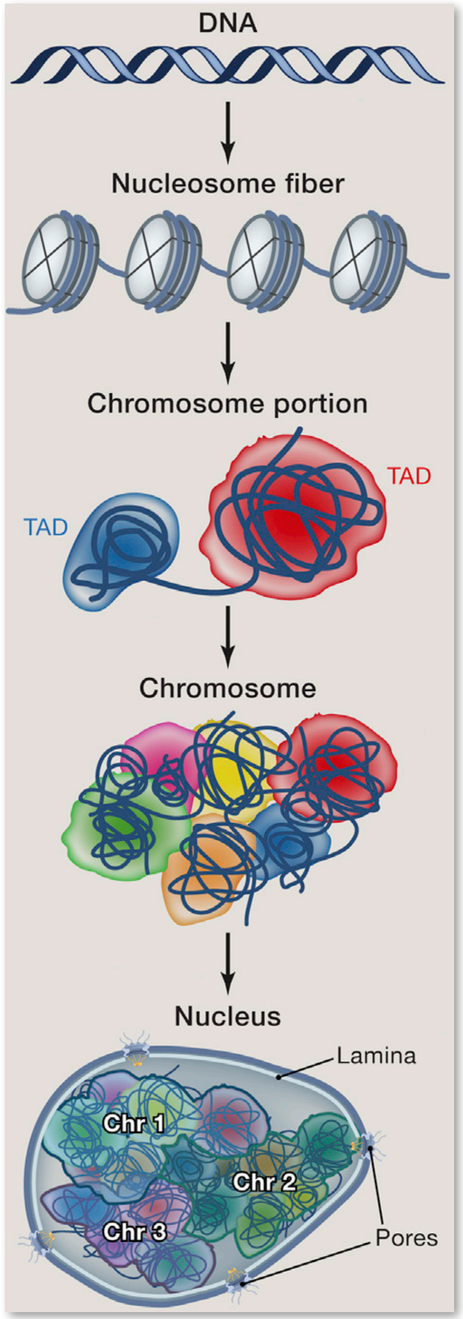
Epigenetic marks

- If DNA is like a book, epigenetic marks are like sticky notes



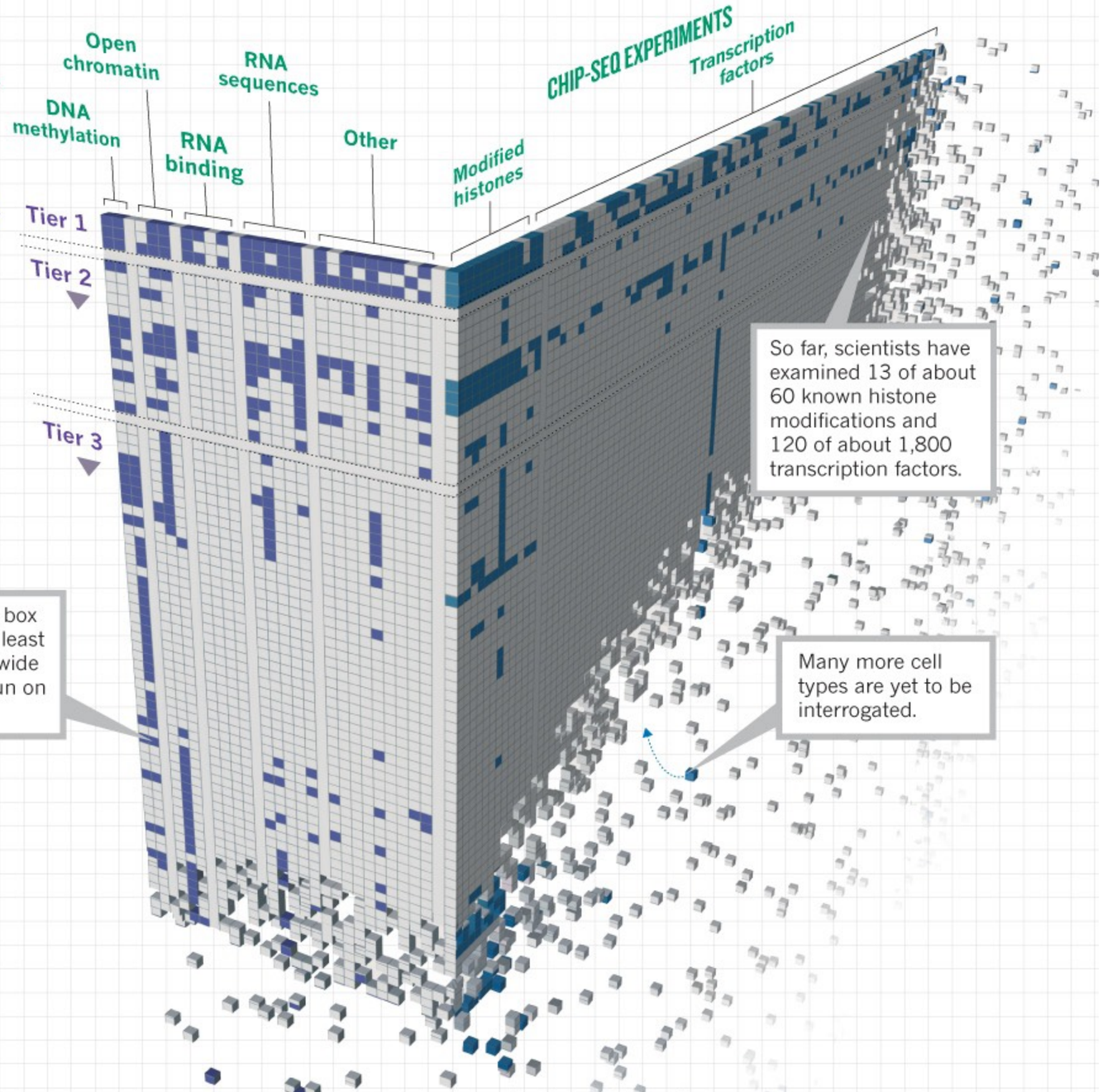
Chromatin structures

- Eukaryotic chromatin structure can be viewed as a series of superimposed organizational layers.
- At the root are the DNA sequence and its direct chemical modification by **cytosine methylation.**
- The DNA is folded into **nucleosomes** — the fundamental units of chromatin — that comprise approximately 147 bp of DNA wrapped around a histone octamer.
- The nucleosomal **histones** H2A, H2B, H3 and H4 can be chemically modified and exchanged with variants. The nucleosome positions along with histone variants and modifications make up the primary structure of chromatin.
- Finally, three-dimensional models of chromatin in nuclei are now being developed with increasing precision and propose that there are additional sophisticated layers of genome regulation through **higher-order organization** and **nuclear compartmentalization.**



MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.



EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

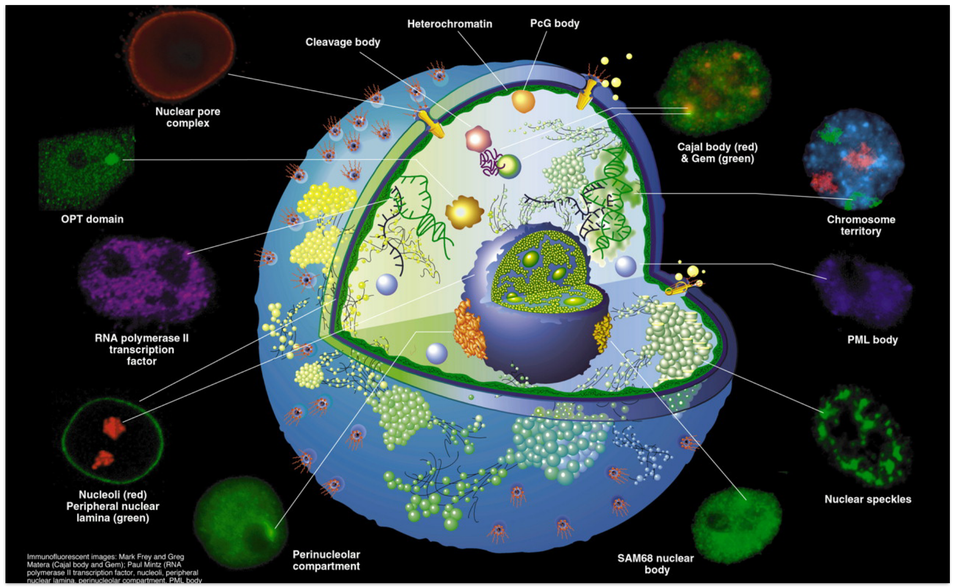
Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.

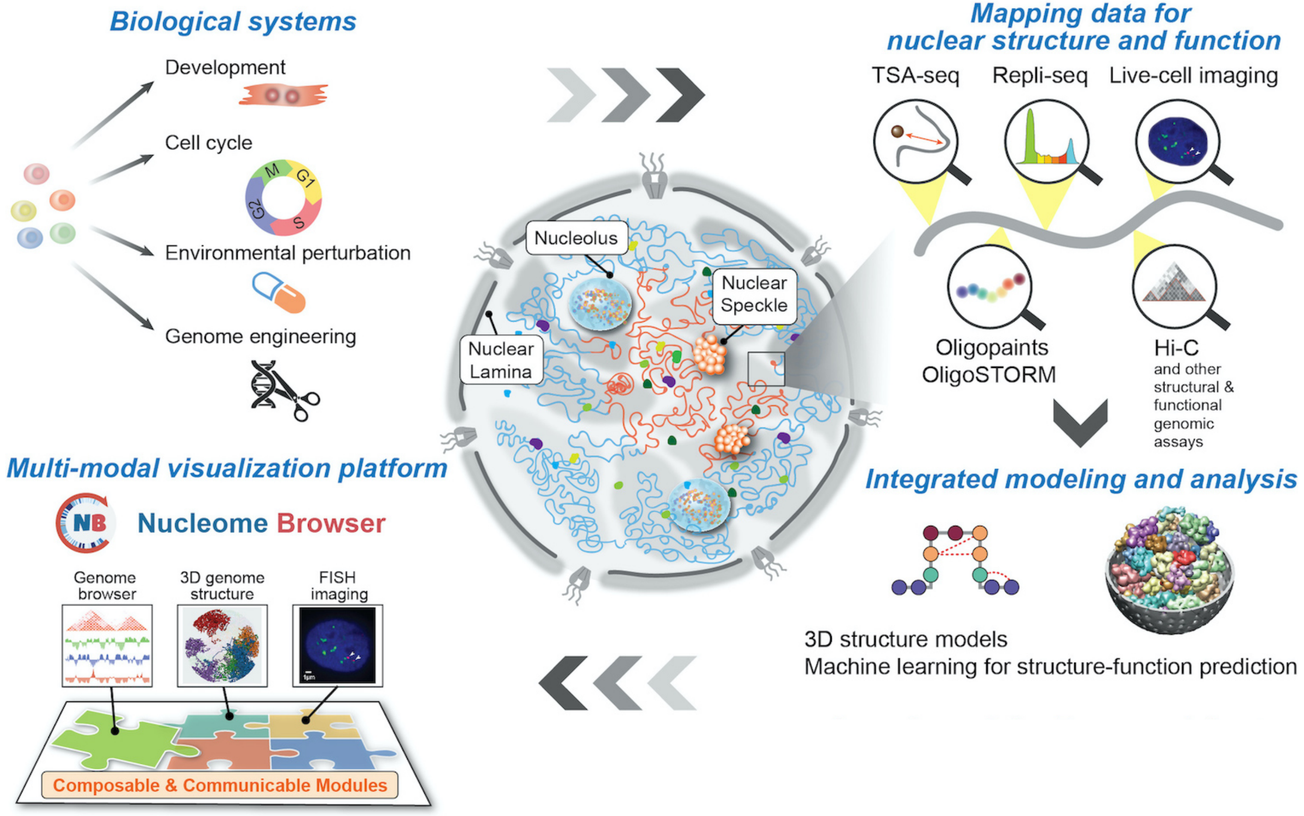
Higher-order genome organization

- We can now determine a genome's sequence and annotate linear chromatin composition, but our genome is not linear
- Our knowledge of the 3D organization of the genome remains limited
- Need to achieve high spatial and temporal resolution

Large-scale genome organization in the nucleus



Spector, *J of Cell Science* 2001

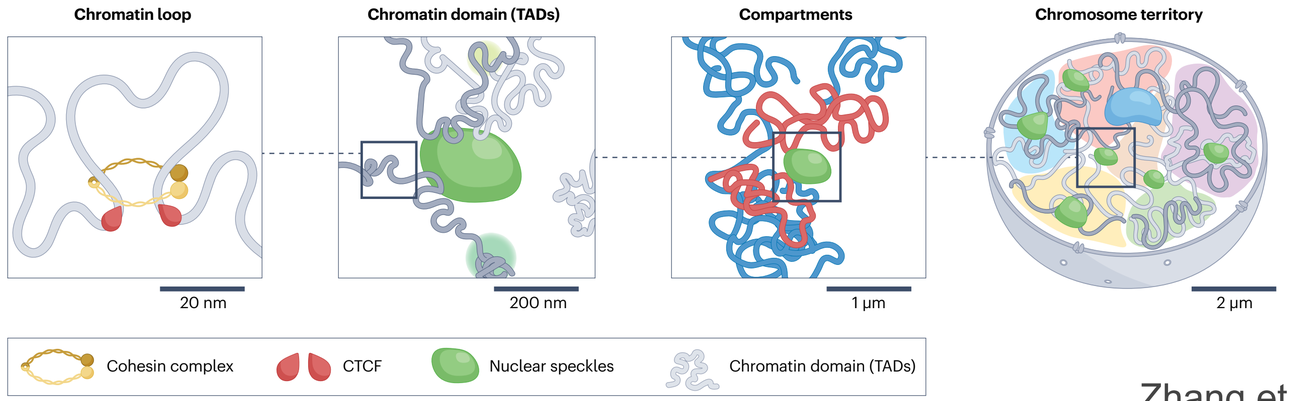


Multiscale Analyses of 4D Nucleome Structure and Function by Comprehensive Multimodal Data Integration

4D Nucleome

Carnegie Mellon University · ILLINOIS University of Illinois at Urbana-Champaign · Allen Institute for Cell Science · Harvard Medical School · UCLA · NIH National Cancer Institute · University of Dundee · San Diego Biomedical Research Institute · Brown

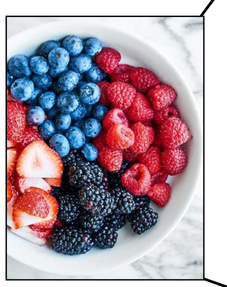
Multiscale 3D genome organization in single cells



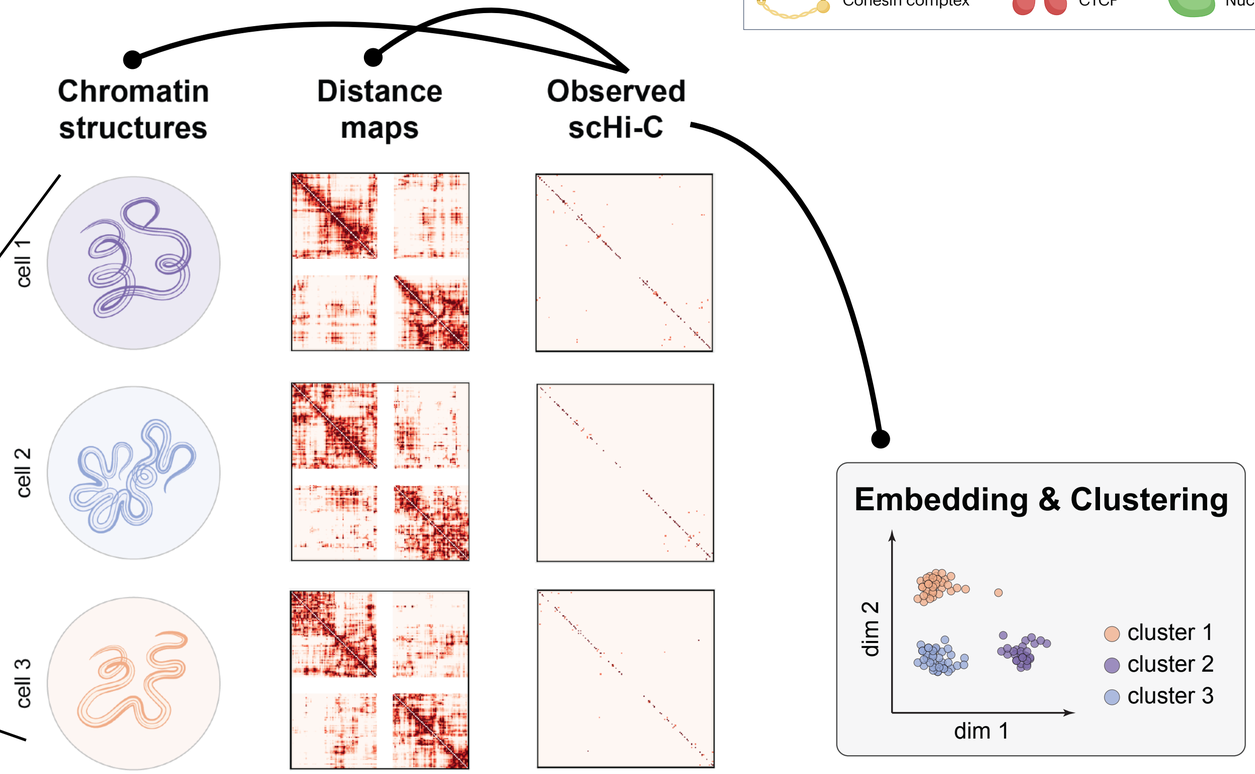
Zhang et al.
Nat Rev Genetics 2024

Bulk

Single cell

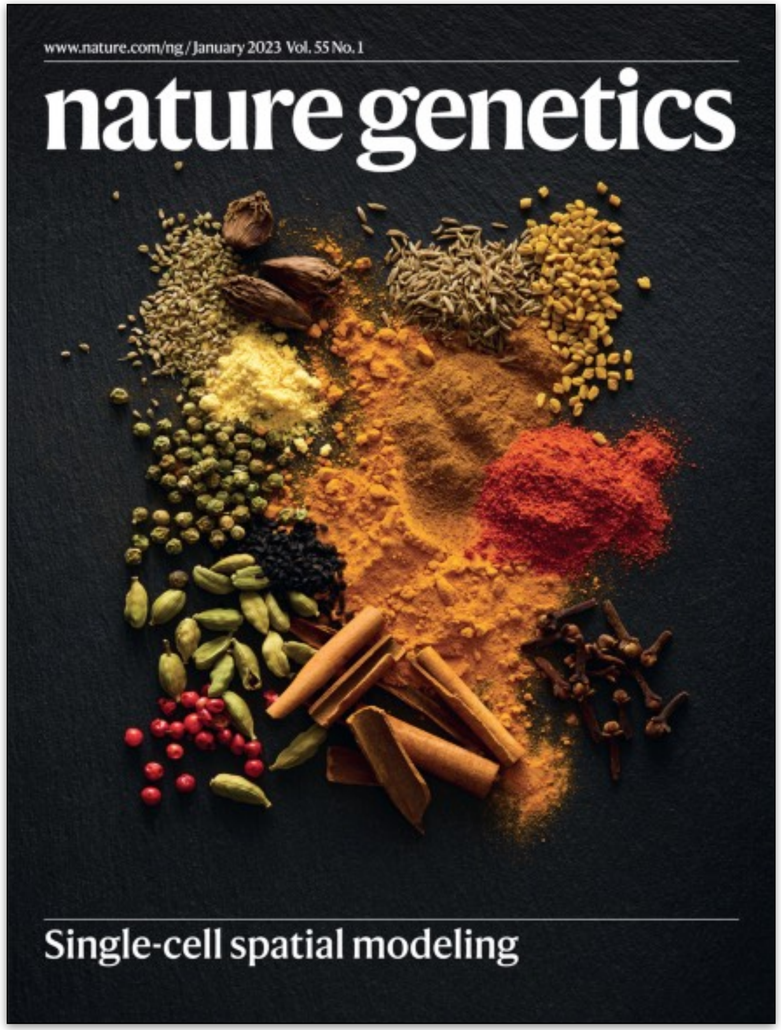


Recover missing information

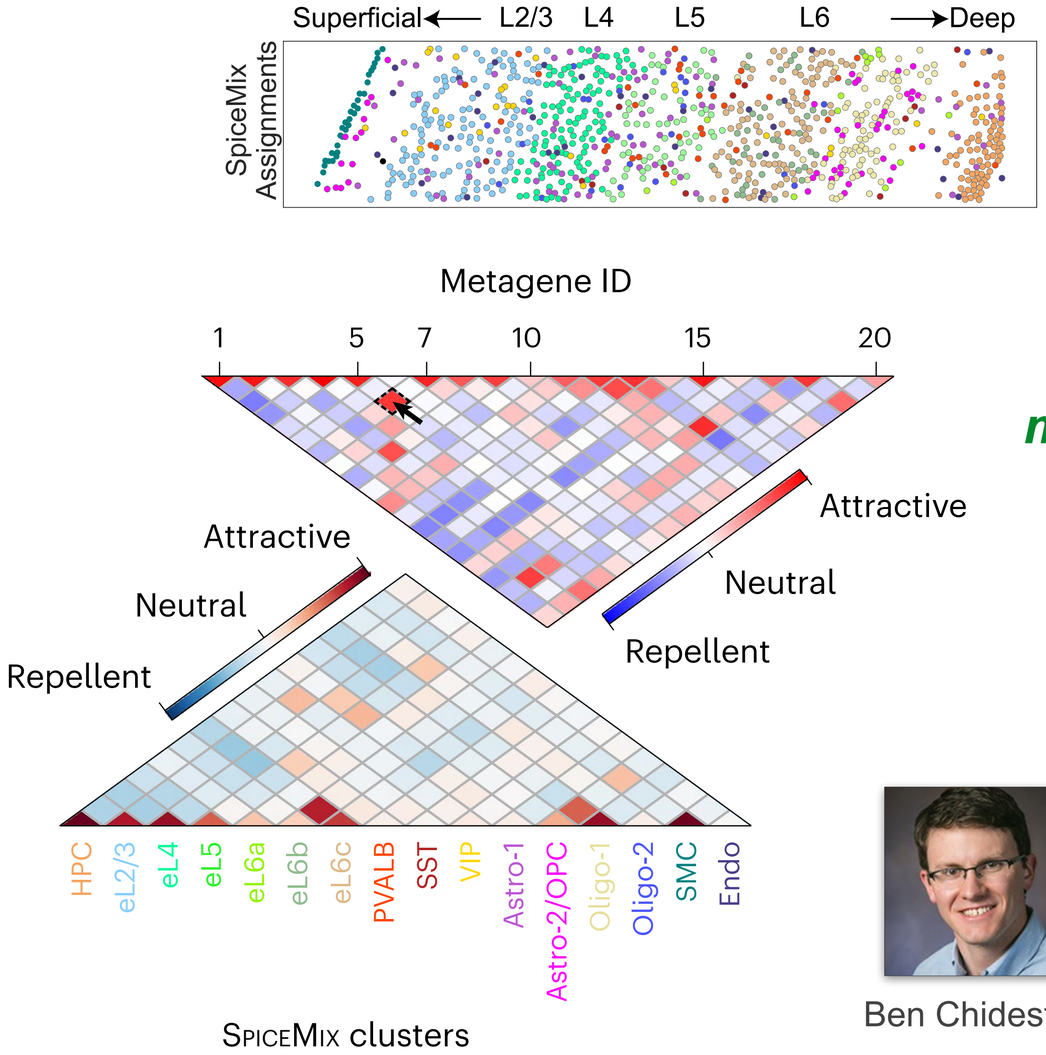


- What does 3D genome look like in single cells?
- How do multiscale 3D features vary in individual cells and different cell types?
- Structure-function relationships

SPICEMIX enables integrative single-cell spatial modeling

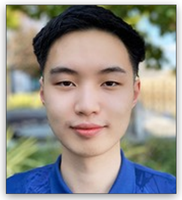


January 2023 issue



metagenes

Ben Chidester



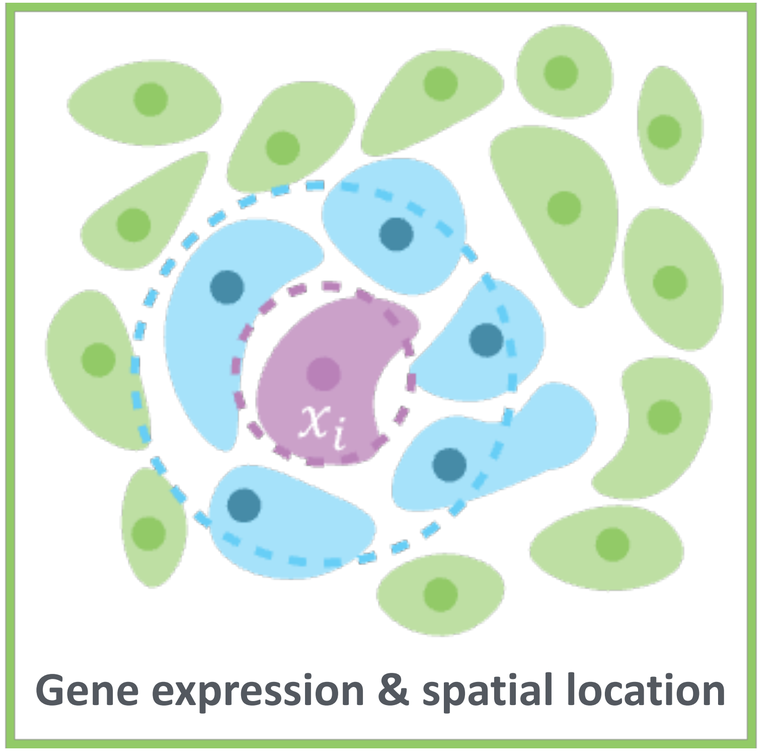
Tianming Zhou

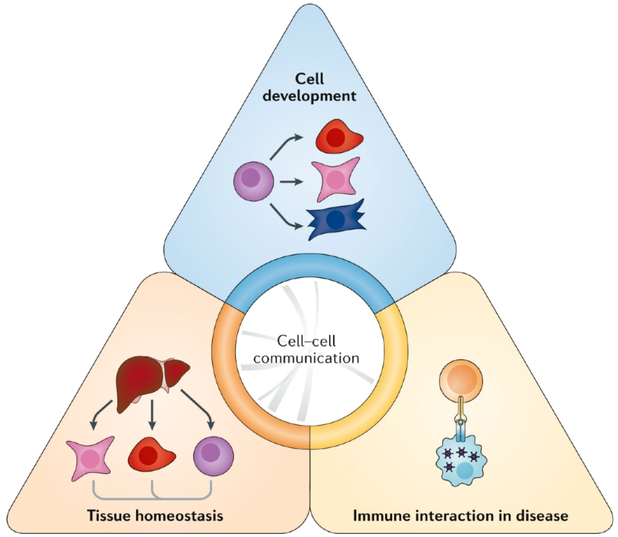
Chidester #, Zhou #, Alam, and Ma. *Nature Genetics,* 2023

STEAMBOAT: modeling cell-cell interactions

- The molecular profile of cells is a result of superimposing:
  - **intrinsic factors**
  - **interactions at multiple scales**

- **How do we decompose them and model such multiscale interactions?**





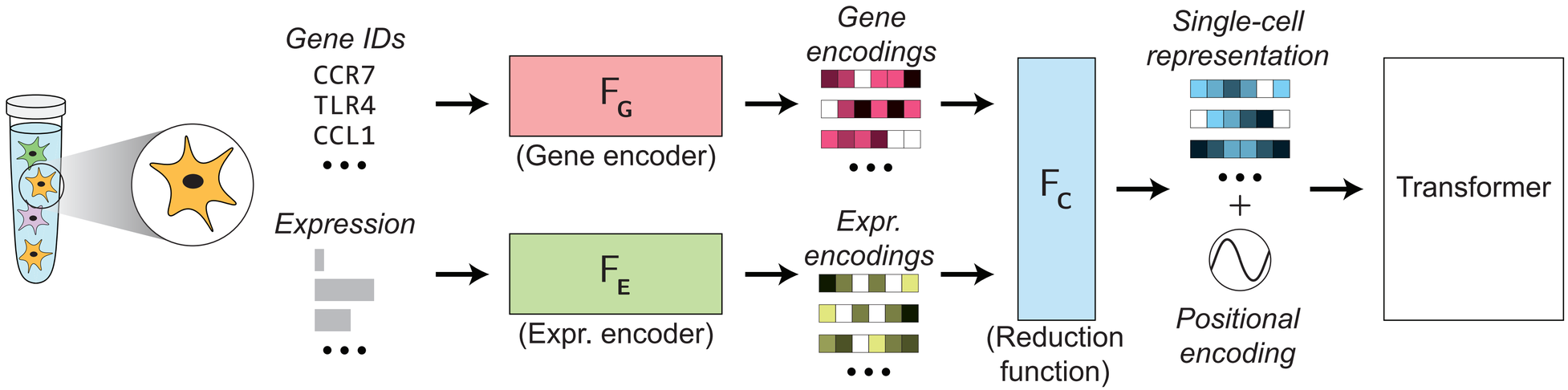
Armingol et al. *Nat Rev Genet* 2021

Attention-based multiscale delineation of cellular interactions in tissues

Shaoheng Liang

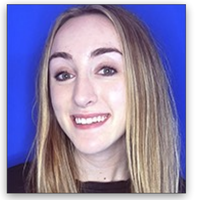
Liang et al. *bioRxiv* 2025

Heimdall explores design choices in single-cell FMs



F_G examples		F_E examples		F_C examples	
Random	*ID* CCR7	Sorting	CCR7, TLR4, CCL1 → TLR4, CCL1, APOE	Filter	F_G → Filter by F_E ← F_E
ESM2	*Amino acid sequence* MDL...	Binning	# Genes / Count → CCR7, TLR4, CCL1	Sum	F_G → F_G token + F_E token ← F_E
GenePT	*ID* CCR7 → *NCBI* → *ChatGPT*	Autobin	CCR7, TLR4, CCL1 → CCR7, TLR4, CCL1		
Gene2vec	*Co-expression* CCR7, TLR4, CCL1				

Make design choices composable and interpretable

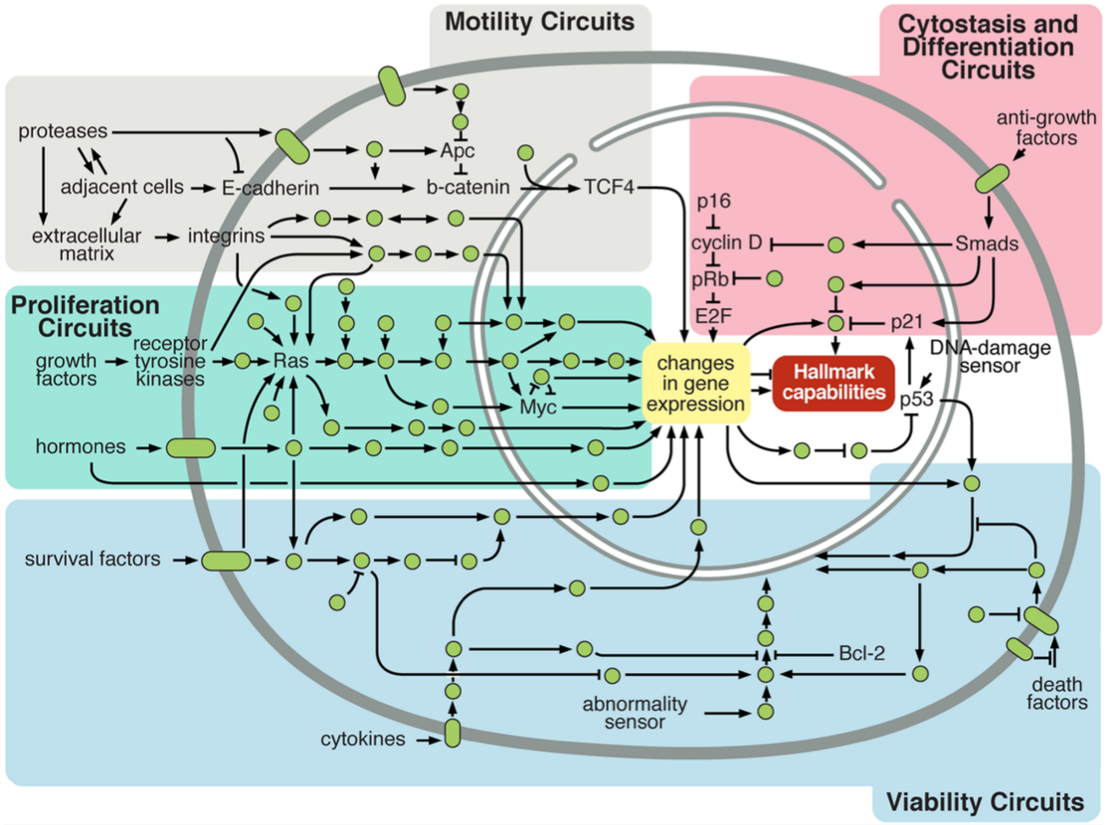

Ellie Haber

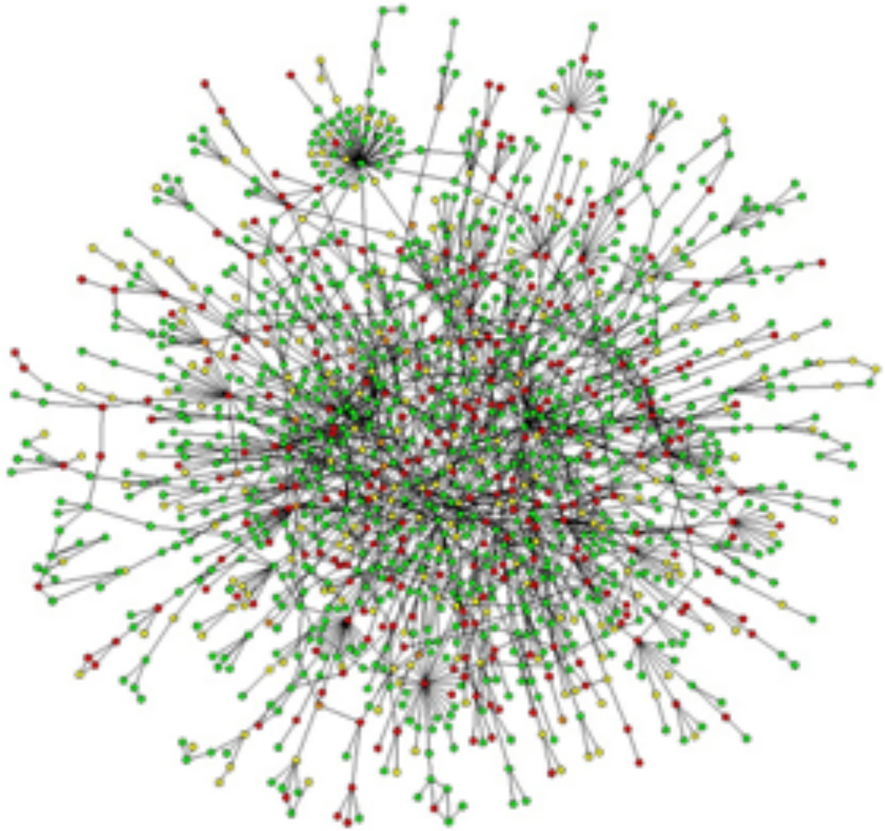

Spencer Krieger


Shahul Alam

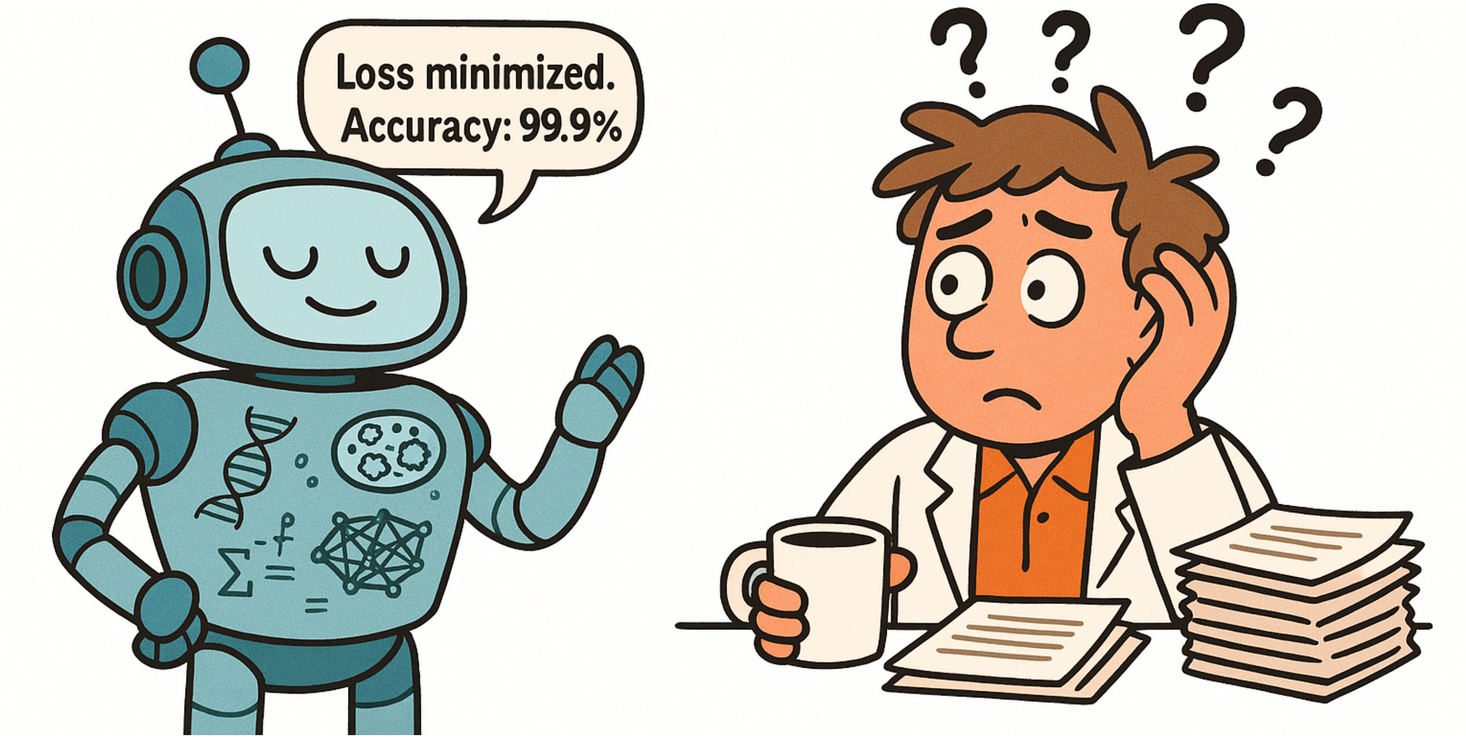

Nick Ho

Systems level perturbations lead to disease





- Different constituents in the cell do not function as single entities
- Disease causing mutations exhibit high degree of heterogeneity among individuals; but the impact of network level perturbations may be more important to model
- Different components in the tissue are also interconnected



The machine is learning something — but are we?


A FILM FROM DANIELS
EVERYTHING EVERYWHERE ALL AT ONCE
天 馬 行 空
COMING SOON


SPECIAL EDITION
CLINT EASTWOOD
THE GOOD THE BAD AND THE UGLY
2-DISC DVD COLLECTOR'S SET