Lecture 1: Introduction to Genomic Data

What Genomic Data Looks Like

Yen-Yi Ho

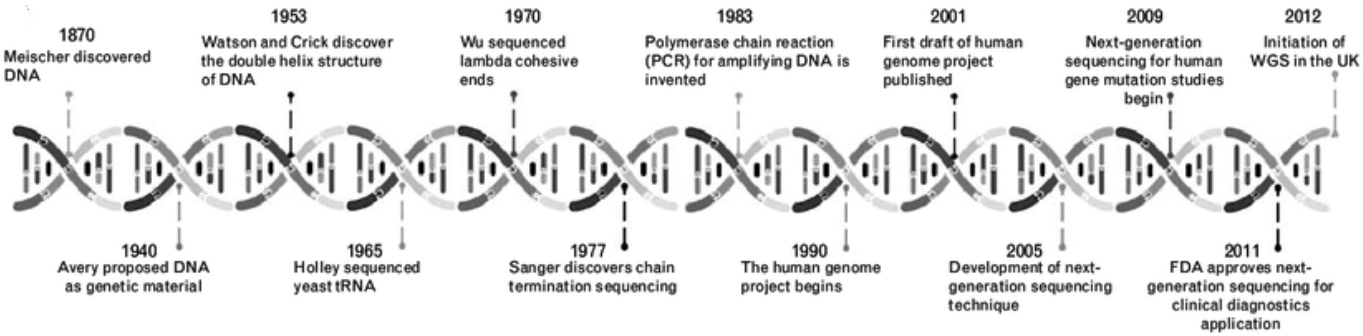
Department of Statistics

Roadmap

- What is genomic data?
- Sequencing technologies and data generation
- Downstream analyses: RNA-seq, variant interpretation, population genomics, epigenomics

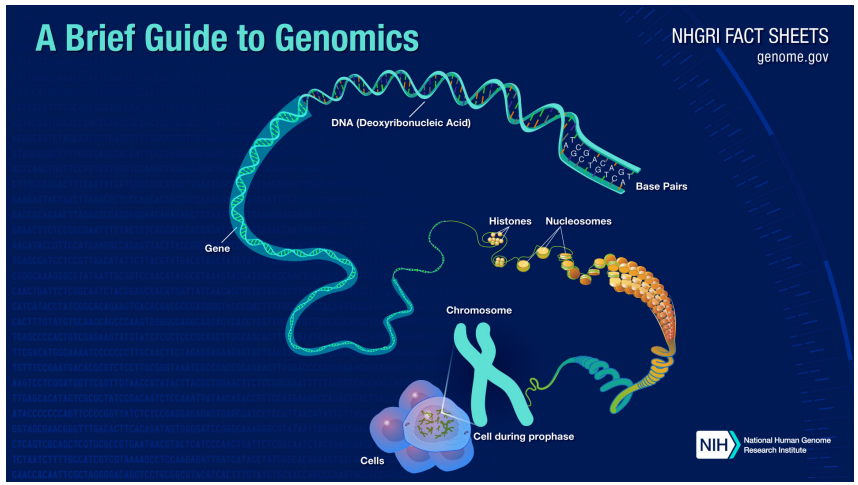

DECEMBER 24, 2012
Egypt Divided / Pot's Big Moment / Best of 2012 Movies, Music, Books & More
TIME
Want to Know My Future?
Alzheimer's
Dementia
Diabetes
Asthma
Cancer
Hemochromatosis
Cystic fibrosis
Colon cancer
Breast cancer
Tay-Sachs disease
Huntington's disease
Epilepsy
Burkitt's lymphoma
Parkinson's
Malignant melanoma
Glaucoma
Phenylketonuria
Prostate cancer
Obesity
New genetic tests can point to risks—but not always a cure
BY BONNIE ROCHMAN
www.time.com

A long time ago in a galaxy far, far away....



What is Genomic Data? — Big picture

- **Genomic data** = information derived from an organism's entire DNA sequence (the *genome*).
- Contains nucleotide sequences (A, C, G, T) and observed variation between individuals.
- Used to study *variation*, *disease*, *evolution*, and *biology*.

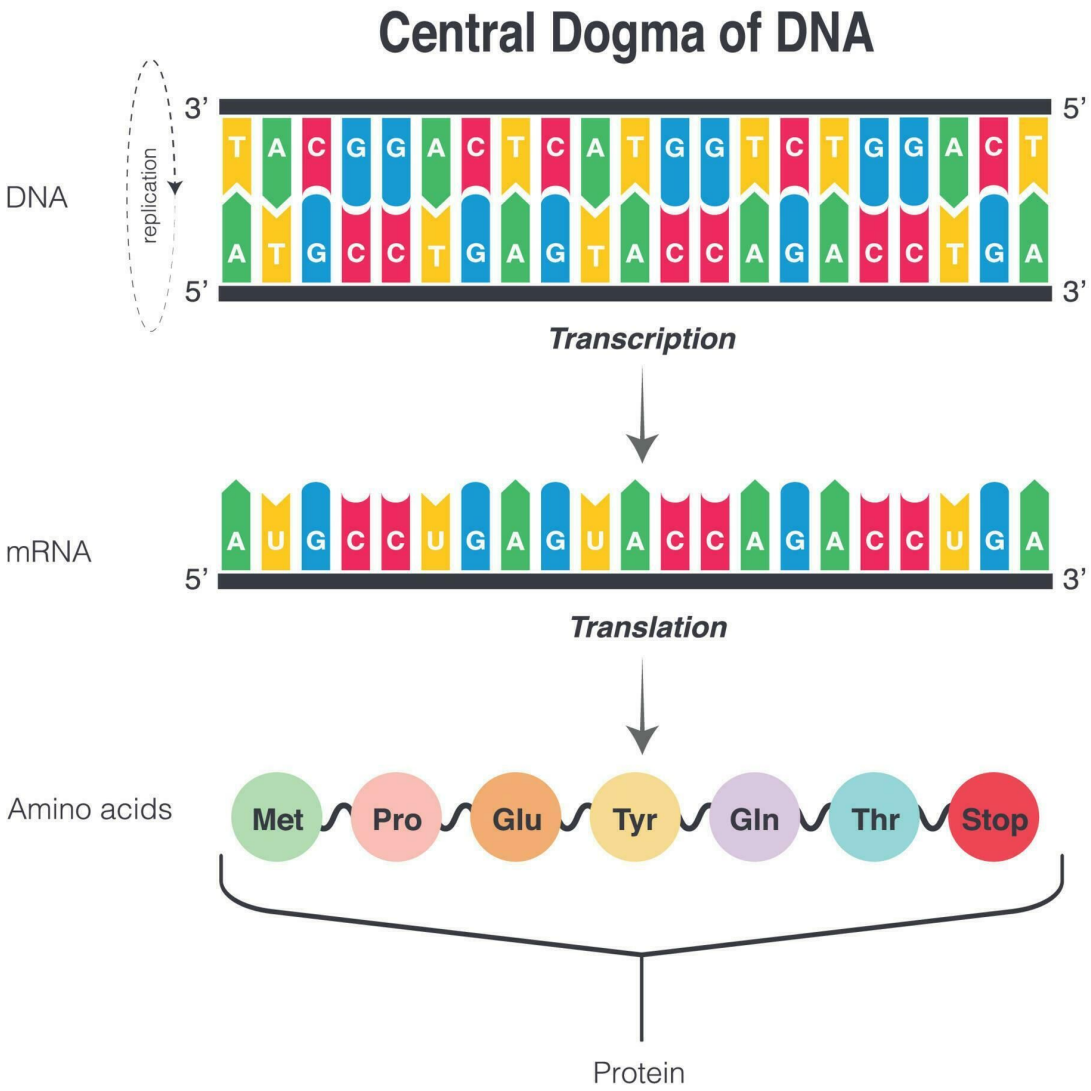


In human genome, every single cell in the body contains a complete copy of the approximately 3 billion (3×10^9) DNA base pairs (letters).

Structure of genomic data: DNA

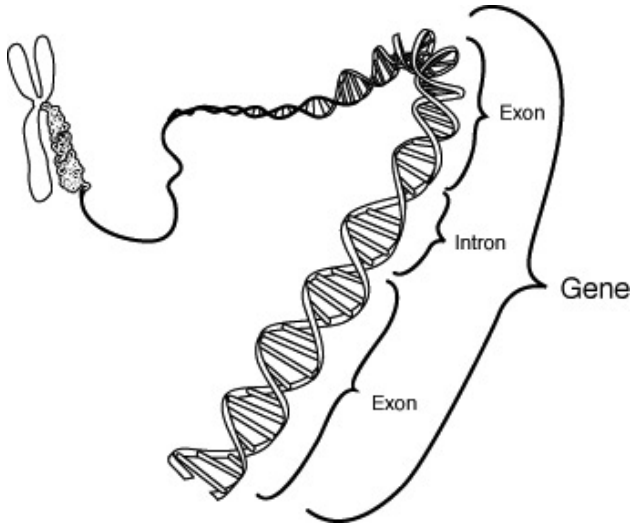
- Data is composed of nucleotides: `A`, `C`, `G`, `T`.
- Organized into **reads**, **contigs**, **chromosomes**, and **annotations**.
- Common file formats: `FASTQ`, `FASTA`, `VCF`, `GTF/GFF`.

Central Dogma of Molecular Biology



Gene

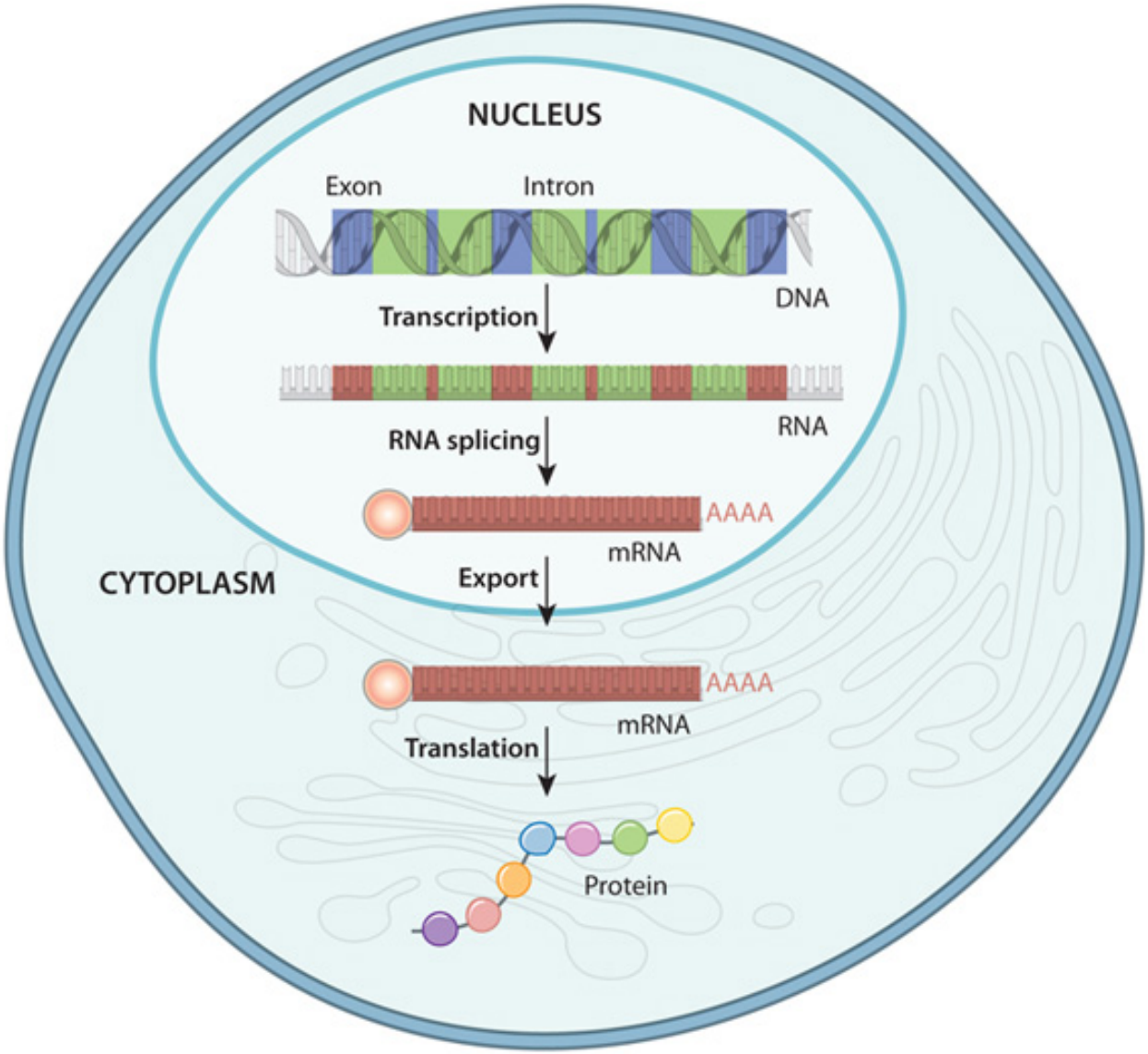
- A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions.
- Or simply, a piece of "useful" DNA sequence.

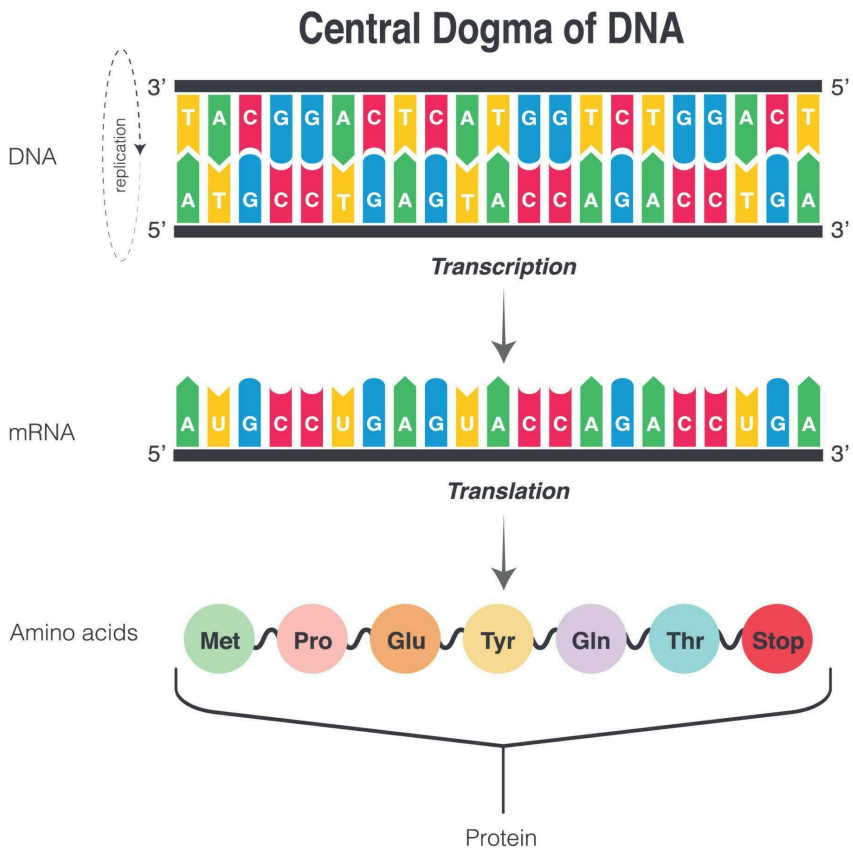


Functional Genomics

- **enhancer**: a region for enhancing gene expression. Not necessarily closes to the gene.
- **promoter**: at the beginning of the gene, helps transcription.
- **exons**: the "useful" part of the gene, will appear in the mRNA product.
- **introns**: the "spacer" between exons, will NOT be in the mRNA product.
- **splicing**: the process to remove introns and join exons.
- **alternative splicing**: different splicing patterns for the same pre-mRNA. For example, mRNA could be from exons 1 and 2 or exons 1 and 3. Those are different **transcripts** of the same gene.

Gene structure and splicing

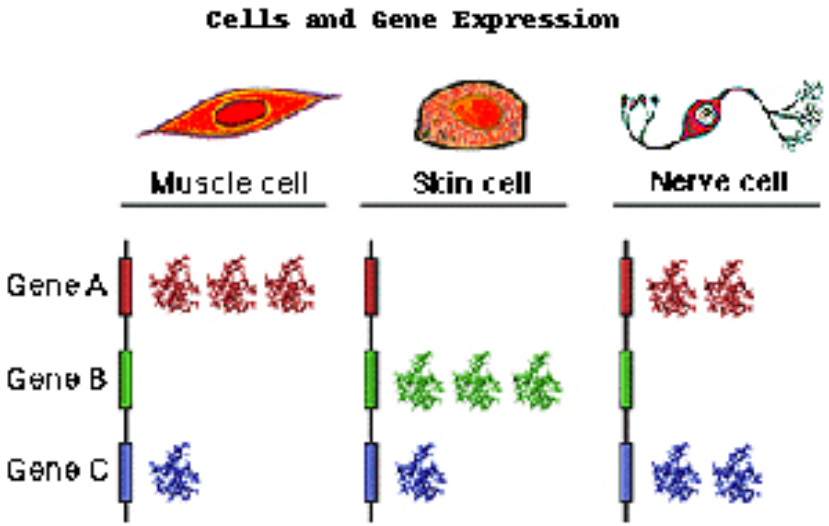



Central Dogma of DNA
DNA
replication
3'
5'
T A C G G A C T C A T G G T C T G G A C T
A T G C C T G A G T A C C A G A C C T G A
5'
3'
Transcription
mRNA
5'
A U G C C U G A G U A C C A G A C C U G A
3'
Translation
Amino acids
Met
Pro
Glu
Tyr
Gln
Thr
Stop
Protein

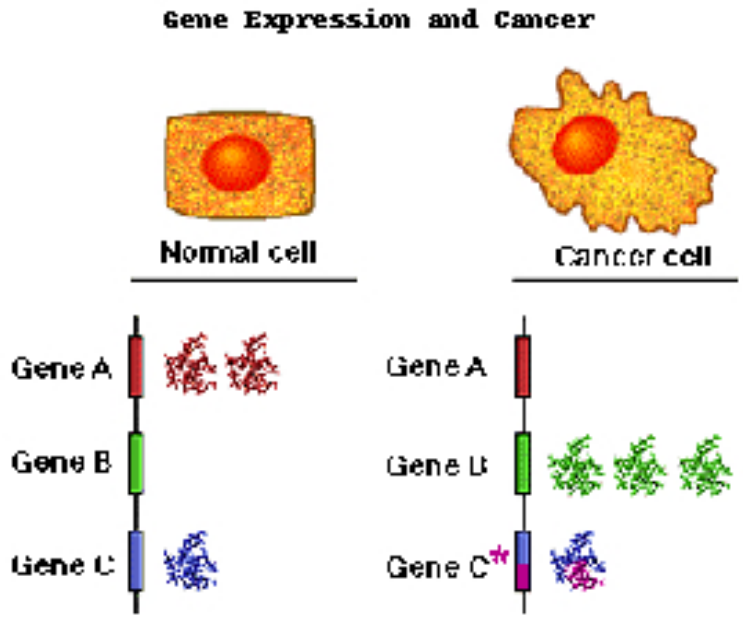
Gene Expression

Gene expression is a term that is used to describe the entire process of translation and transcription of a gene. Gene expression is a highly specific process. Only a small fraction of the genes are expressed, or turned "on," in any particular type of cell.

gene expression in different tissues

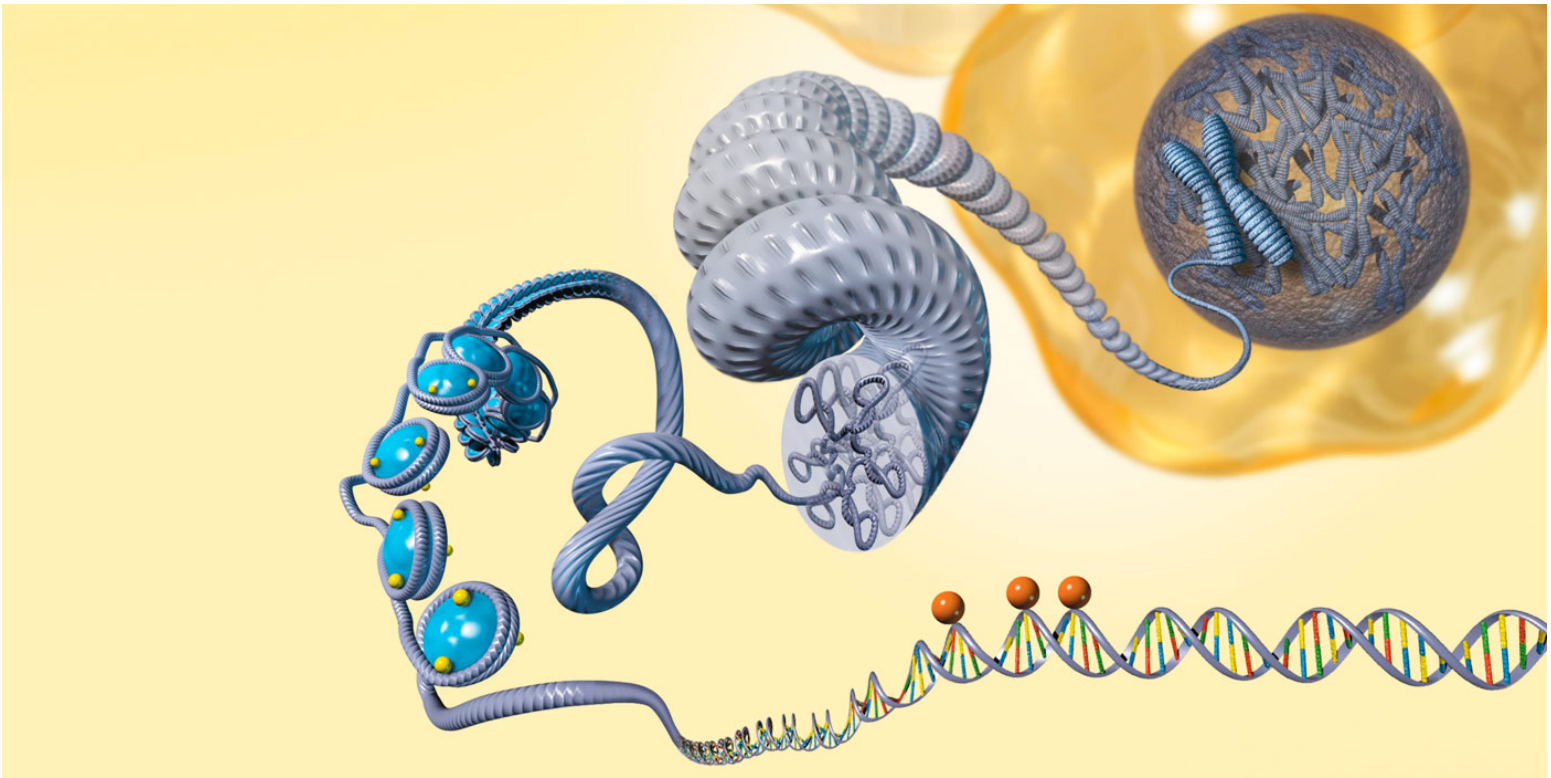


gene expression in the same tissue, but different points in time



Epigenetics

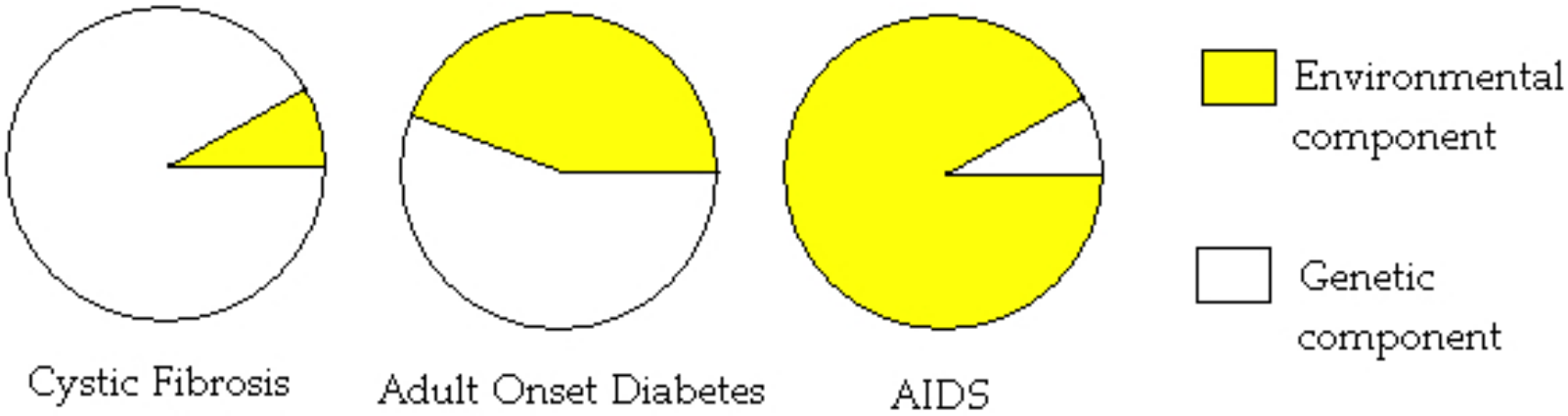
Non-DNA sequence related, heritable mechanisms to control gene expressions. Example: DNA methylation, histone modifications.



Source of Variations

Environment Vs. Gene

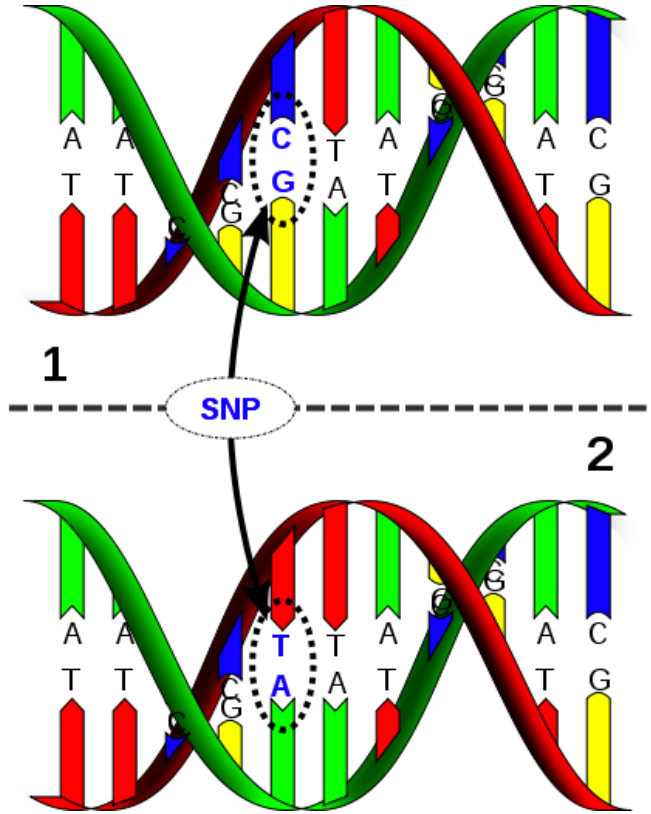
Any two individuals are 99.9% identical in their DNA



Genetic Variations (Polymorphisms)

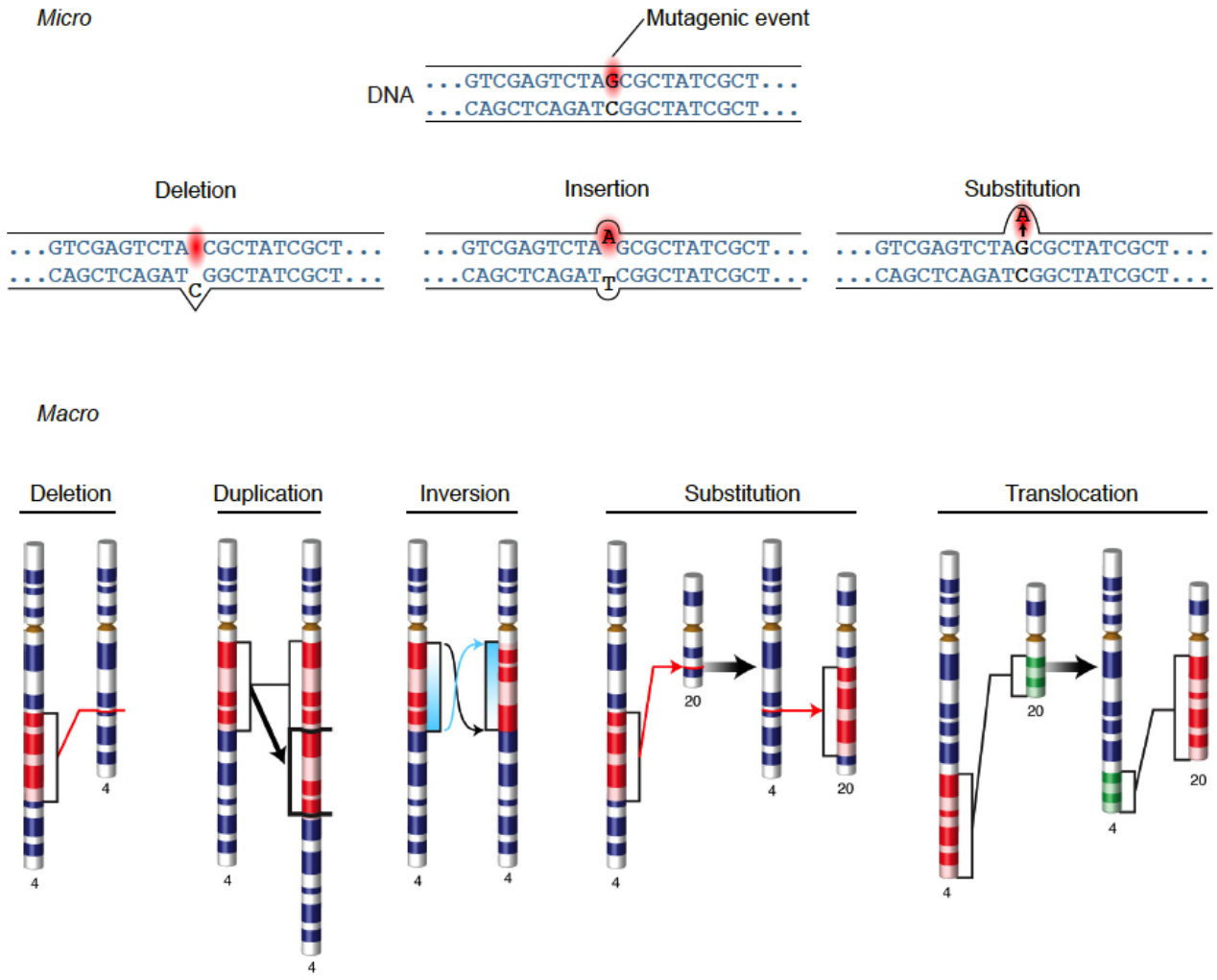
That 0.1 % is very important in defining our differences

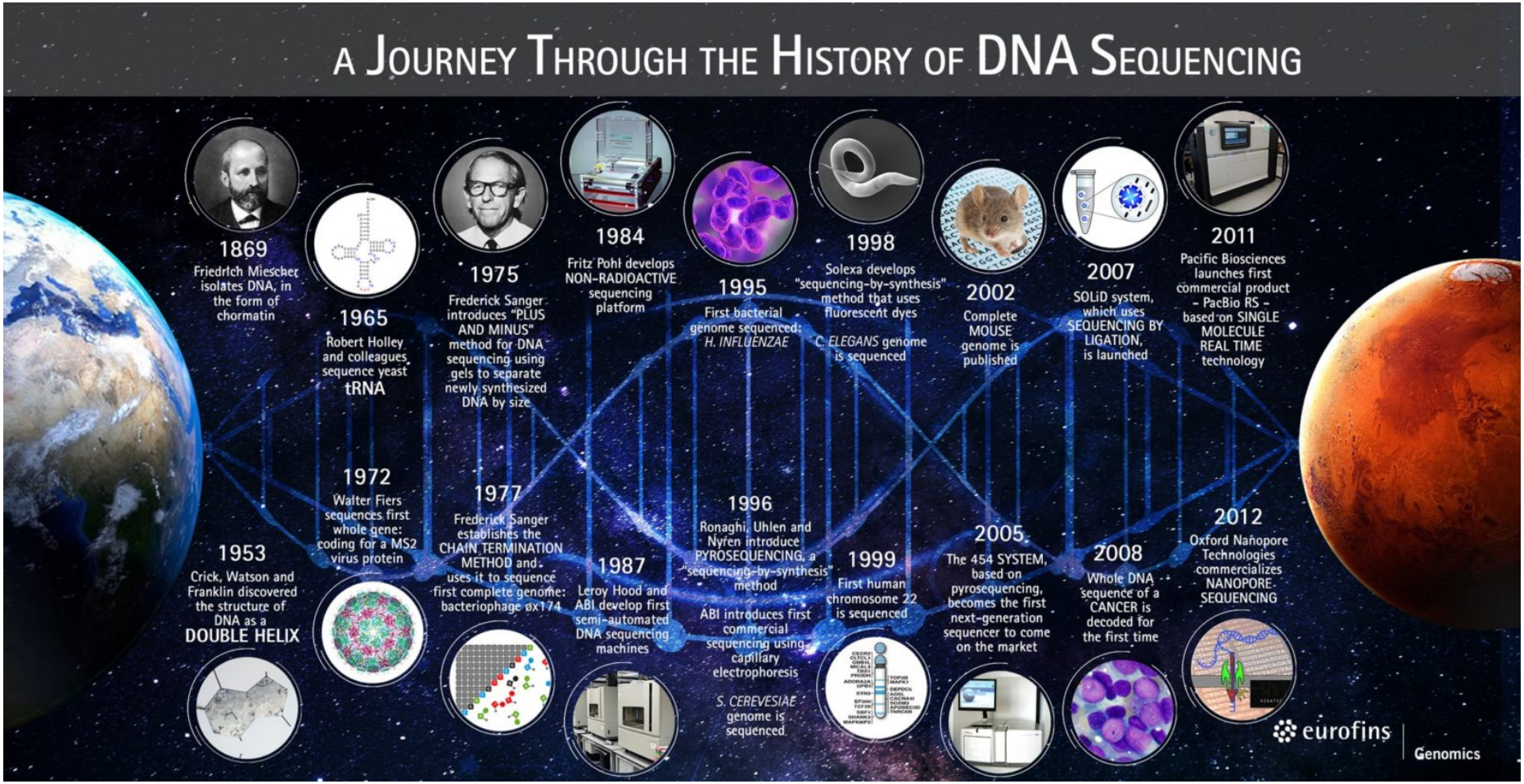
- single nucleotide polymorphisms (SNPs, every 300 nucleotide on average)
- small-scale mutation, insertions, deletions
- copy number variations (AAGAAGAAGAAG)



source: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp

Mutations


Micro
Mutagenic event
DNA
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Deletion
Insertion
Substitution
Macro
Deletion
Duplication
Inversion
Substitution
Translocation


A JOURNEY THROUGH THE HISTORY OF DNA SEQUENCING
1869
Friedrich Miescher isolates DNA, in the form of chormatin
1953
Crick, Watson and Franklin discovered the structure of DNA as a DOUBLE HELIX
1965
Robert Holley and colleagues sequence yeast tRNA
1972
Walter Fiers sequences first whole gene: coding for a MS2 virus protein
1975
Frederick Sanger introduces "PLUS AND MINUS" method for DNA sequencing using gels to separate newly synthesized DNA by size
1977
Frederick Sanger establishes the CHAIN TERMINATION METHOD and uses it to sequence first complete genome: bacteriophage ϕx174
1984
Fritz Pohl develops NON-RADIOACTIVE sequencing platform
1987
Leroy Hood and ABI develop first semi-automated DNA sequencing machines
1995
First bacterial genome sequenced: H. INFLUENZAE
1996
Ronaghi, Uhlen and Nyren introduce PYROSEQUENCING, a "sequencing-by-synthesis" method
ABI introduces first commercial sequencing using capillary electrophoresis
S. CEREVESIAE genome is sequenced
1998
Solexa develops "sequencing-by-synthesis" method that uses fluorescent dyes
C. ELEGANS genome is sequenced
1999
First human chromosome 22 is sequenced
2002
Complete MOUSE genome is published
2005
The 454 SYSTEM, based on pyrosequencing, becomes the first next-generation sequencer to come on the market
2007
SOLiD system, which uses SEQUENCING BY LIGATION, is launched
2008
Whole DNA sequence of a CANCER is decoded for the first time
2011
Pacific Biosciences launches first commercial product - PacBio RS - based on SINGLE MOLECULE REAL TIME technology
2012
Oxford Nanopore Technologies commercializes NANOPORE SEQUENCING
eurofins | Genomics

How Sequencing Works?

- Determining the order of nucleotides (A, T, C, G)
- Technologies differ in chemistry, accuracy, read length, throughput
- Two major eras:
  - Next-Generation Sequencing (NGS)
  - Long-read (3rd generation) sequencing

NGS Overview

- Massively parallel sequencing
- Short reads (50–300 bp)
- High throughput, low cost
- Dominant platforms: Illumina, Ion Torrent

Illumina Sequencing

Sequencing-by-Synthesis (SBS)

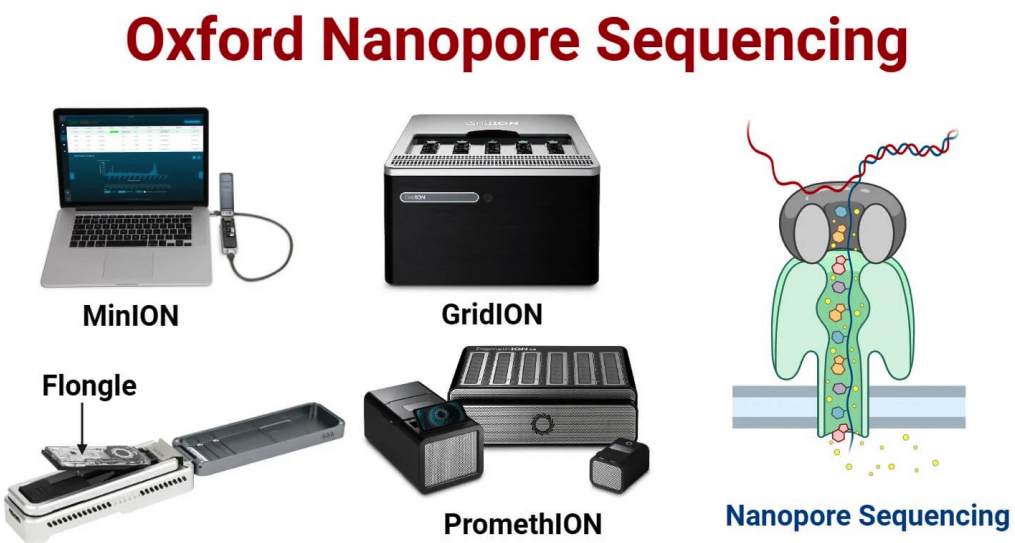
- Bridge amplification on flowcell
- Reversible fluorescent terminators
- Very high accuracy
- Applications: WGS, RNA-seq, ChIP-seq

Ion Torrent Sequencing

- Semiconductor detection of pH change
- No optical system needed
- Read length: 200–400 bp
- Limitations: homopolymer errors

Oxford Nanopore Technologies (ONT)

- Nanopores embedded in membranes measure ionic current
- Very long reads (10 kb–1 Mb)
- Portable devices (MinION)
- Applications: assembly, metagenomics, transcriptomics



PacBio SMRT Sequencing (Long Reads)

- Single molecule real-time sequencing
- HiFi reads: 10–25 kb with very high accuracy (Q20+)
- Ideal for structural variants and de novo assembly

How Sequencing Data Is Generated

- **Step 1: Sample Preparation**
  - Extract DNA or RNA
  - Assess purity and concentration
- **Step 2: Library Preparation**
  DNA is fragmented and adapters are added.
- **Step 3: Sequencing Run**
  Depends on platform:
  - Illumina: sequencing-by-synthesis with fluorescent nucleotides
  - Nanopore: strands pass through nanopores
  - PacBio: circular consensus sequencing
- **Step 4: Base Calling**
  Raw signals $\rightarrow$ nucleotide sequences
  - Illumina: fluorescence intensities
  - ONT: electrical current patterns
  - PacBio: light pulses

Data Output Formats

- FASTQ — sequences with quality scores
- BAM/CRAM — aligned reads (compressed)
- VCF — variants
- FASTA — reference sequences

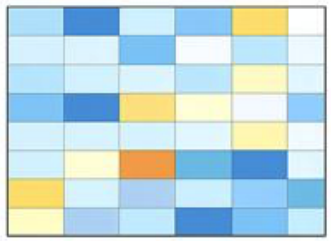
```
@SEQ_ID
GATTTGGGGTC AAGAAAAGCA
+
!!"((((((**....**)%%}}
```

Raw Reads

Assembled Genome

CHROM	POS	REF	ALT
17	1234567	C	T
20	1110696	G	D
3	16512	T	AG

Variant Calls



Functional Genomics

Applications in Research & Clinical Genomics

- Whole-genome/exome sequencing
- RNA-seq for transcriptomics
- Metagenomics for microbial community profiling
- Epigenomics (whole genome bisulfite sequencing, ATAC-seq)
- Structural variants and long-read phasing

Spatial Biology is the Next Life Sciences Revolution

Bulk

2006

Genome Analyzer

Next Generation Sequencing
2009 Method of the Year

Single Cell

2012

Chromium

Single Cell Genomics
2013 Method of the Year

Spatial

2019

GeoMx® DSP

Spatially-resolved Transcriptomics
2020 Method of the Year

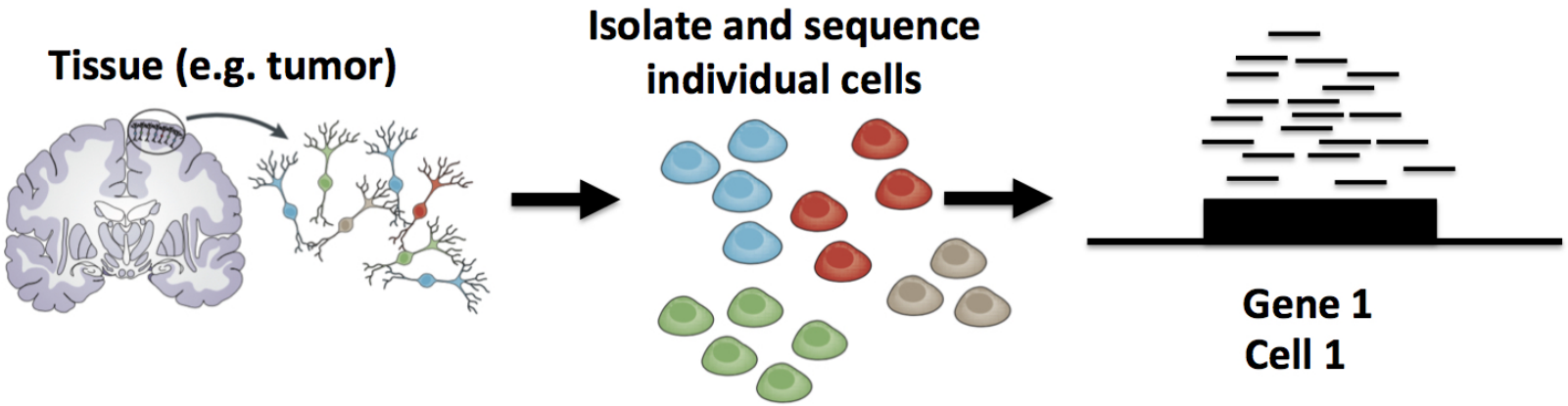
nanoString

Adapted from concept and images by Dr. Aviv Regev of The Broad Institute; *Nature Methods*

Single-Cell Sequencing: Overview

- Measures transcriptomic or epigenomic profiles at **single-cell resolution**.
- Enables discovery of:
  - Cell types and subtypes
  - Developmental trajectories
  - Cellular heterogeneity in disease
- Common platforms:
  - Droplet-based (10x Genomics)
  - Plate-based (SMART-seq2)
- Typical output: **cell** $\times$ **gene** count matrix (usually sparse).

Single-cell RNA-Seq (scRNA-Seq)



Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

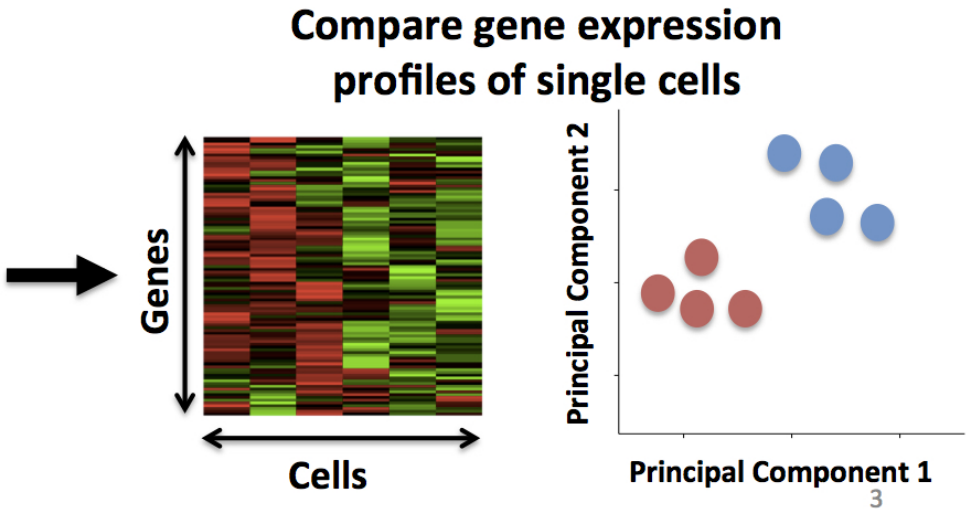


Figure adapted from https://speakerdeck.com/stephaniehicks/welcome-to-the-world-of-single-cell-rna-sequencing

Spatial Transcriptomics

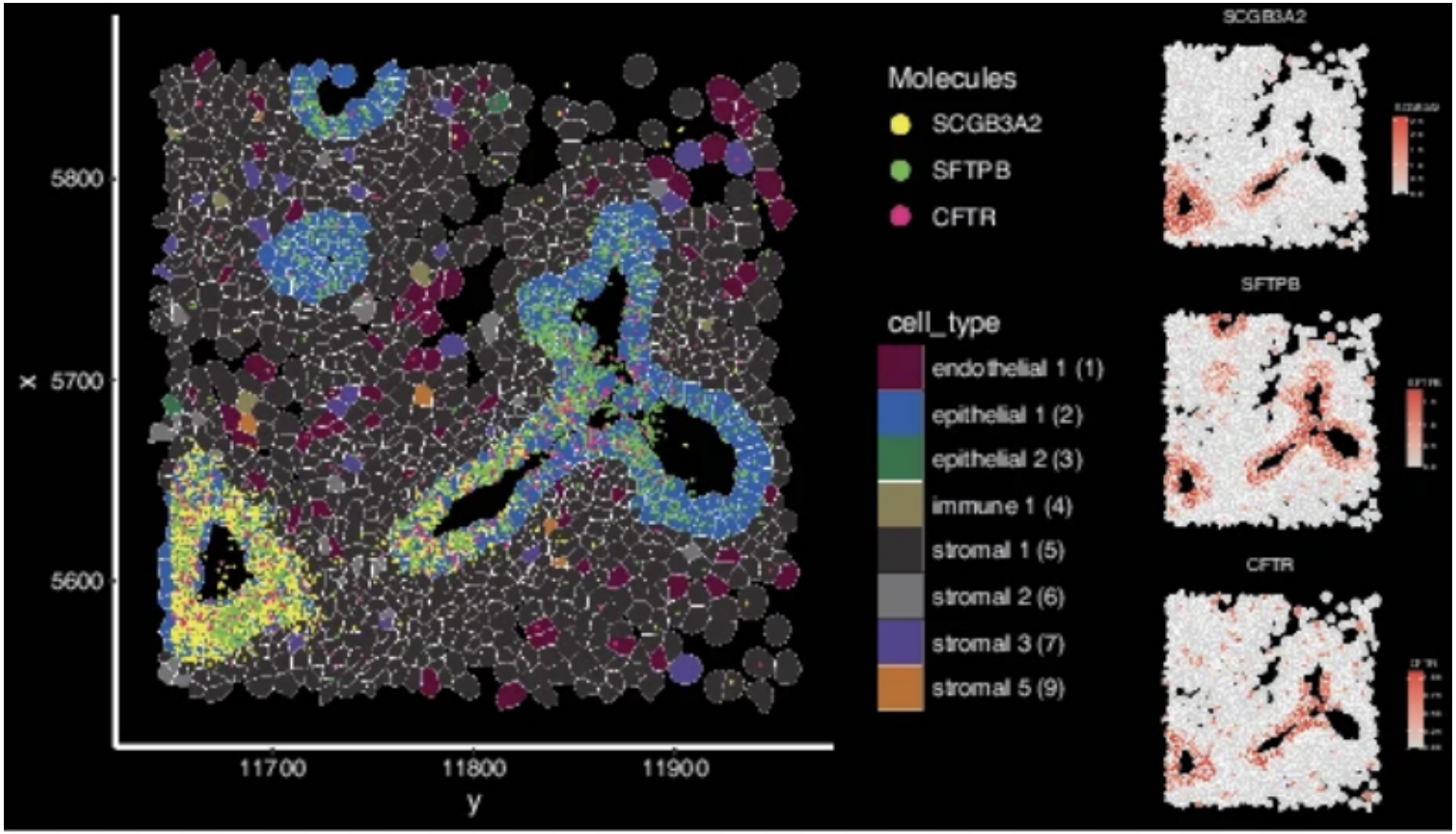


Figure adapted from Quach H, et al. Early human fetal lung atlas reveals the temporal dynamics of epithelial cell plasticity. Nat Commun 15: 5898 (2024). doi: 10.1038/s41467-024-50281-5

Spatial Transcriptomics: Overview

- Measures gene expression while preserving **spatial context** in tissues.
- Key question: *Where are gene programs active within the tissue microenvironment?*
- Major platform categories:
  - **Spot-based sequencing platforms**
    - 10x Visium
    - Slide-seq / Slide-seqV2
  - **High-resolution imaging-based platforms**
    - **10x Xenium** (RNA in situ imaging, subcellular resolution)
    - **NanoString CosMx SMI** (Spatial Molecular Imager; single-cell/subcellular resolution)
    - MERFISH, seqFISH
- Output varies by platform:
  - Spot-based: **spot × gene** counts (multi-cell)
  - Imaging-based: **cell × gene** or **subcellular transcripts with coordinates**

So ... what about analytical approachces?

Much to learn, you still have . . . .

Popular neural network architectures

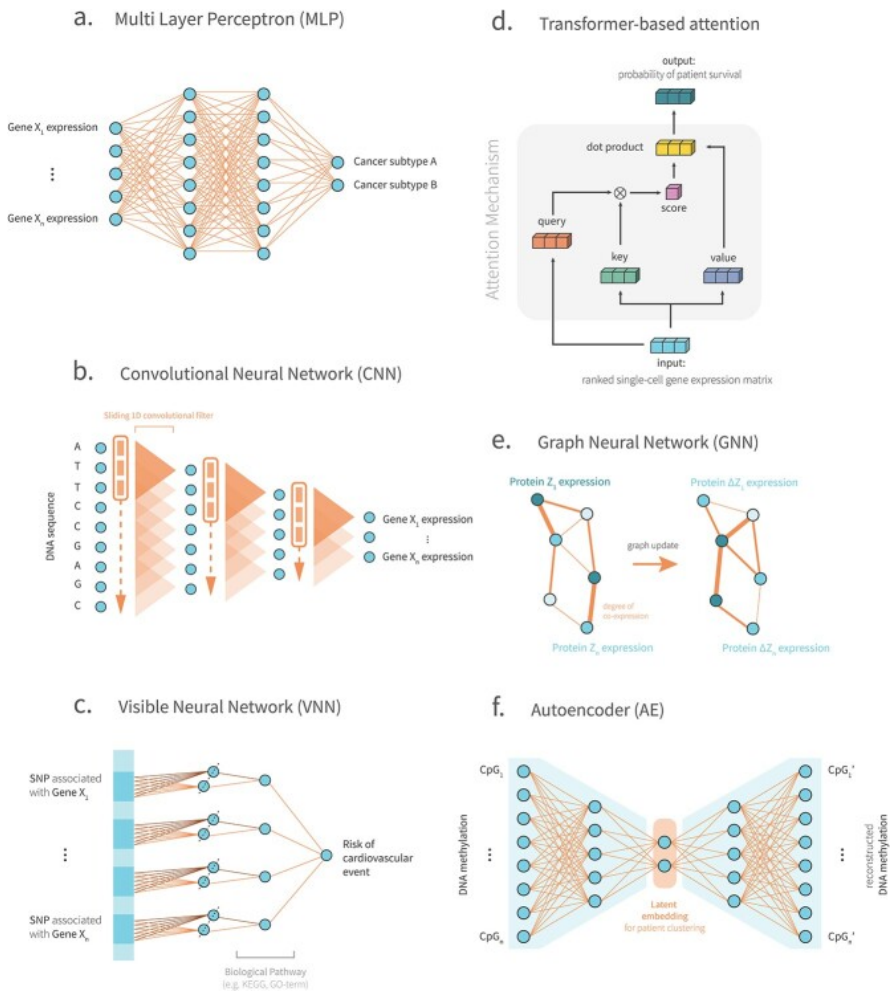


Figure adapted from Van Hilten, A., Katz, S., Saccenti, E., Niessen, W. J., & Roshchupkin, G. V. (2024). Designing interpretable deep learning applications for functional genomics: a quantitative analysis. Briefings in Bioinformatics, 25(5).

Data Types and Biological Fields

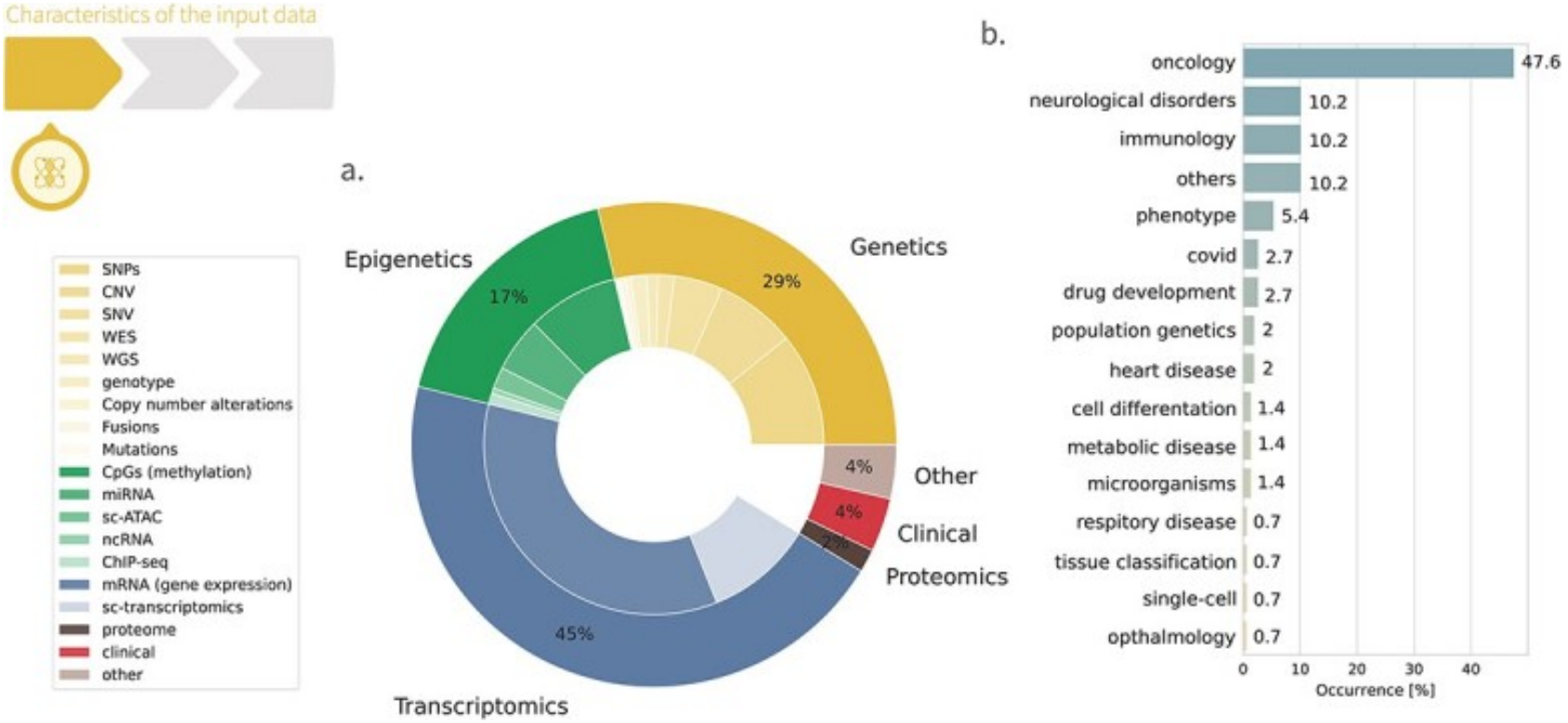


Figure adapted from Van Hilten, A., Katz, S., Saccenti, E., Niessen, W. J., & Roshchupkin, G. V. (2024). Designing interpretable deep learning applications for functional genomics: a quantitative analysis. Briefings in Bioinformatics, 25(5).

What makes genomic data unique?

- **Very large:** single-cell & individual genomes are gigabytes; cohorts are terabytes.
- **Hierarchical:** bases $\rightarrow$ genes $\rightarrow$ chromosomes $\rightarrow$ genomes.
- **Complex formats:** multiple specialized file formats and metadata.
- **Sensitive:** potential to identify individuals — ethical and legal considerations.