

## 9.3b Overview of Basic Molecular Biology

**10x10 – Introduction to Biomedical Informatics & Artificial Intelligence**  
**William Hersh, MD, FACMI, FAMIA, FIAHSI**  
**Copyright 2026, All Rights Reserved**



# Overview of Basic Molecular Biology

---

The cell, DNA, and RNA

Central Dogma

Protein structure and function

Mendelian heredity

Human genome

Genetic variation

Epigenetics

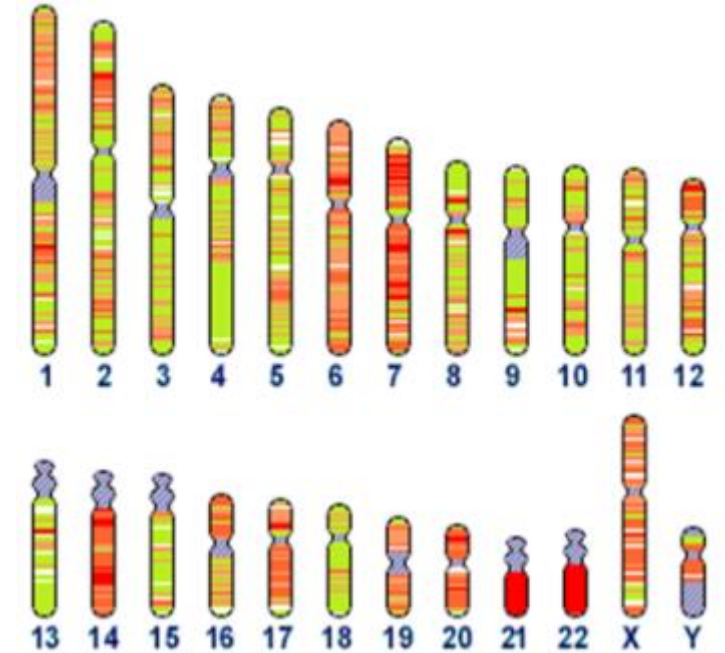
Technologies advancing bioinformatics

# Human genome

Project to sequence human genome [began in 1988](#)

In Feb. 2001, the NIH-based project published “first draft” (Lander, 2001) simultaneously with private effort from Craig Venter of Celera Genomics (Venter, 2001)

Nucleotide sequencing technology continues to evolve substantially with rising capacity and falling costs



# Human Genome Project

---

Completed in 2003 (Collins, 2003)

## Major revelations

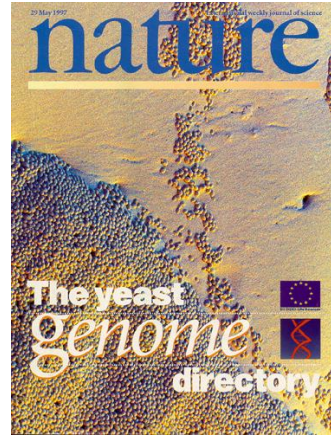
- Estimated number of human genes, thought originally to be as high as 100,000, was lowered to <30,000 (Pennisi, 2003)
- Human DNA contained more “junk” sequences than most lower animals, but with genes crowded together

Last 8% of genome not fully sequenced until 2022 (Nurk, 2022), mainly due to difficulty of sequencing Y chromosome (Rhie, 2023)

# Other genomes sequenced too



1995  
Bacterium  
1.6 Mb  
1600 genes  
0.5 mm



1997  
Yeast  
13 Mb  
6000 genes  
4.4 mm



1998  
Worm  
100 Mb  
20,000 genes  
34 mm



2001  
Human  
3,000 Mb  
35,000 genes  
>1000 mm

- Others – mouse, rat, rice, viruses, including SARS-CoV-2 (Wu, 2020) and variants (Maxmen, 2021)
- Ancient humans – Neandertal (Green, 2010), Old Kingdom Egyptian (Morez Jacobs, 2025)

# Genetic variation

---

Genome of one individual (Craig Venter) showed 4.1 million variants versus “standard” genome build (Levy, 2007)

But should really think of “pangenome” for any species, based on areas of genome variation (Sherman, 2020)

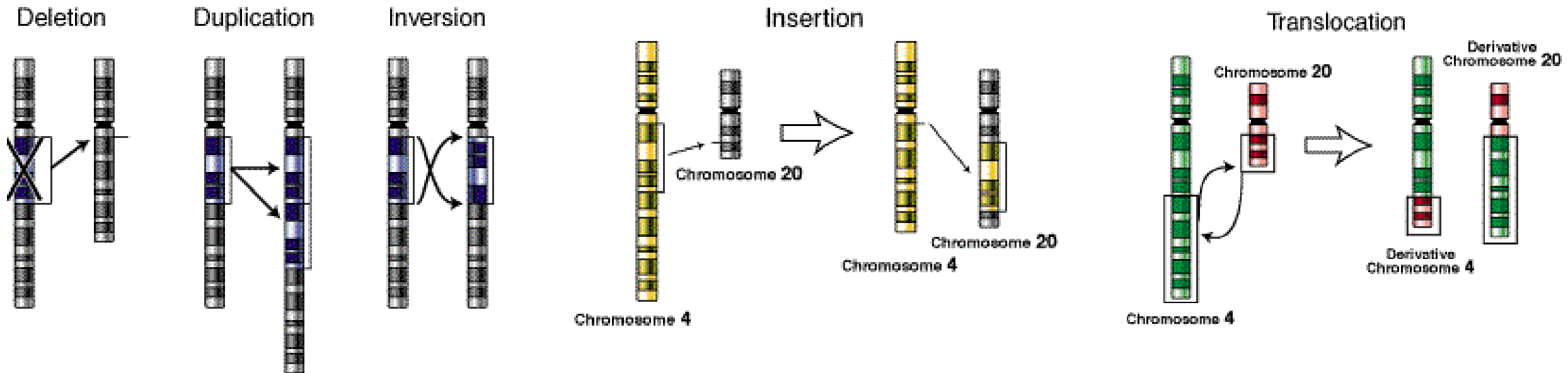
- Recent pangenome for humans published based on genetic variants of 47 individuals (Liao, 2023)
- Also must account for non-coding RNA genes (Amaral, 2023)

Caused by mutations: inversions, duplications, insertion-deletions

Important manifestations include:

- Single-nucleotide polymorphisms (SNPs)
- Copy number variations (CNVs)
- Gene regulation
- Epigenetics

# Much genetic variation arises from mutations



# Single-nucleotide polymorphisms (SNPs)

---

Number of identified human SNPs is 400-500 million and growing (Hakenberg, 2016)

Variation not evenly distributed; some SNPs would be incompatible with life, e.g., mutation in insulin protein

99.9% of genome is identical across humans

- Remainder allows for much variation
- Still leaves 300,000 base pairs to differ!

# ENCODE project “rewrote” some major thinking on genetics

---

Detailed study of 1% of genome in Encyclopedia of DNA Elements (ENCODE) project (Nature, 2007)

Major “rethinking” included (Pennisi, 2007)

- Many genes overlap with each other on chromosome
- “Junk” DNA contains variety of messengers and is even transcribed, although not translated
- Regulatory regions not necessarily adjacent to genes
- Inhibit translation and protein production
- Promote mRNA degradation

ENCODE expanded to entire genome and made new assertions

- 80% of genome, well outside protein-coding regions, plays role in regulating genome and other functions (Nature, 2012)

# Copy number variations (CNVs)

---

CNVs are repeats of DNA sequences of 50+ nucleotides

Venter genome showed 22% of variants not SNPs (Levy, 2007)

About 4.8-9.5% of genome is CNVs (Zarrei, 2015)

- Some can be deleterious, others not
- Unbalanced rearrangements are genomic variants that involve loss (deletion) or gain (duplication) of segments of genome

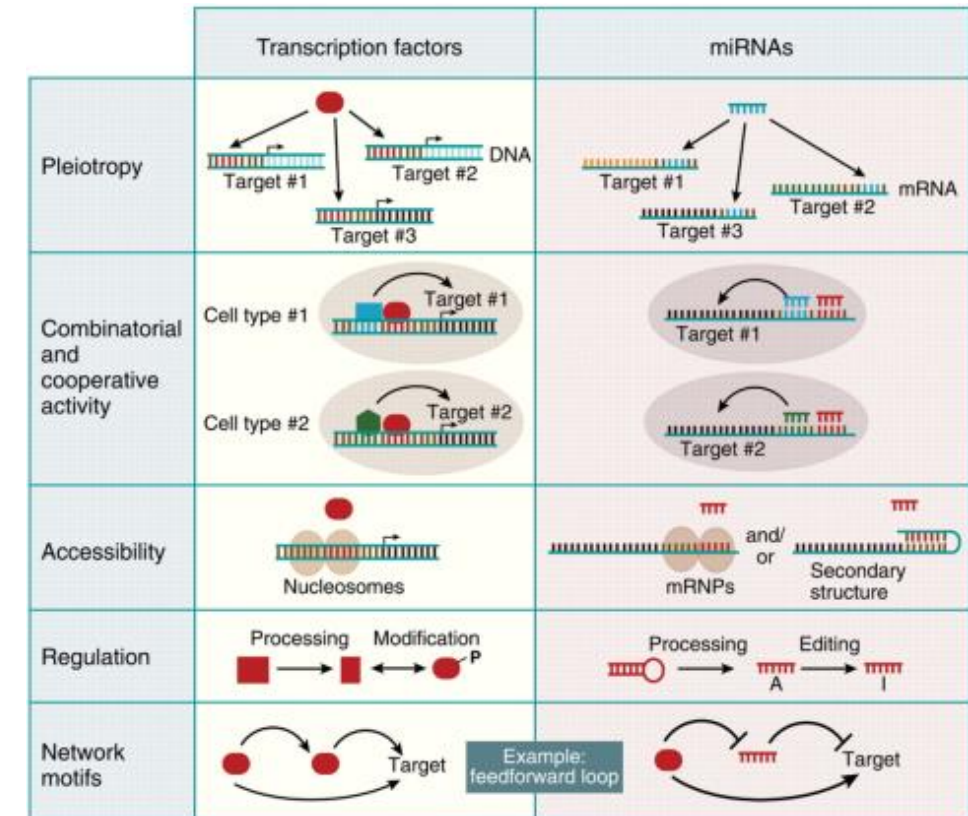
# Gene regulation

As noted already, much similarity in genomes across organisms

Part of difference in phenotype due to gene regulation

Two major approaches to gene regulation (Hobert, 2008)

- Transcription factors
- MicroRNAs – growing understanding of role in disease and as possible diagnostic or therapeutic tools



# Epigenetics

---

Even further “violation” of the central dogma: variation in phenotype or gene expression caused by mechanisms other than DNA sequence differences (Heard, 2014)

Molecular mechanisms beginning to be unraveled, e.g., DNA methylation (Kim, 2017)

Influence of environment shows genes “aren’t destiny” (Cavalli, 2019)

Maybe Lamarck was not completely wrong after all (Handel, 2010)?

# Technologies advancing bioinformatics

---

Polymerase chain reaction (PCR)

Molecular cloning

Mass spectrometry

Nucleotide sequencing

Next-generation sequencing (NGS)

Gene expression

Protein structure prediction

Clustered regularly interspaced short palindromic repeats (CRISPR)

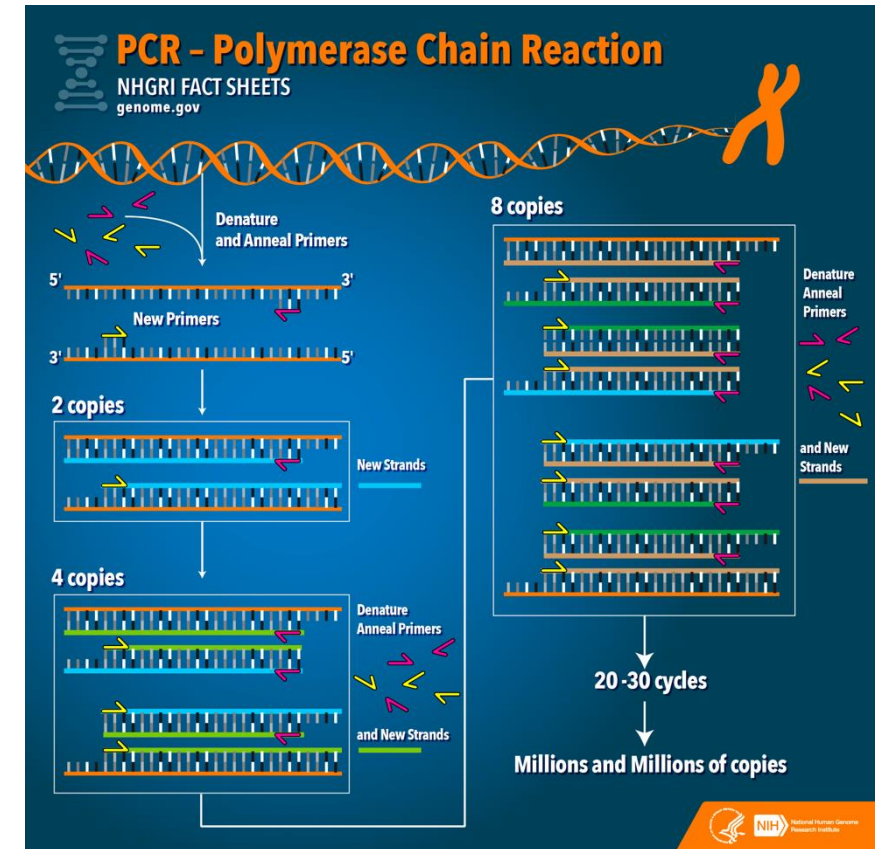


# Polymerase chain reaction (PCR)

A method of creating copies of specific fragments of DNA

Rapidly amplifies a single DNA molecule into many billions of molecules

Many uses, including viral diagnostics, e.g. SARS-CoV-2 virus of COVID-19 (Mohanty, 2021)

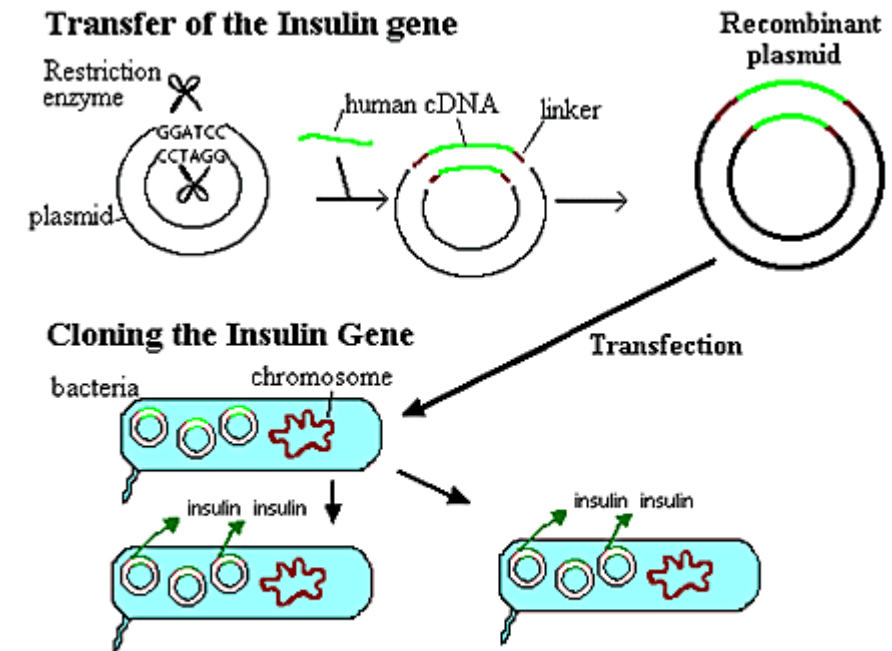


# Molecular cloning

Amplifying a gene or DNA sequence by insertion into the genome of a self-replicating element

Plasmids are strands of DNA inserted into viral or bacterial genomes

Example is production of human insulin



# Mass spectrometry (Garg, 2025)

---

Enables characterization and quantification of proteins

- Complements and extends older x-ray crystallography techniques

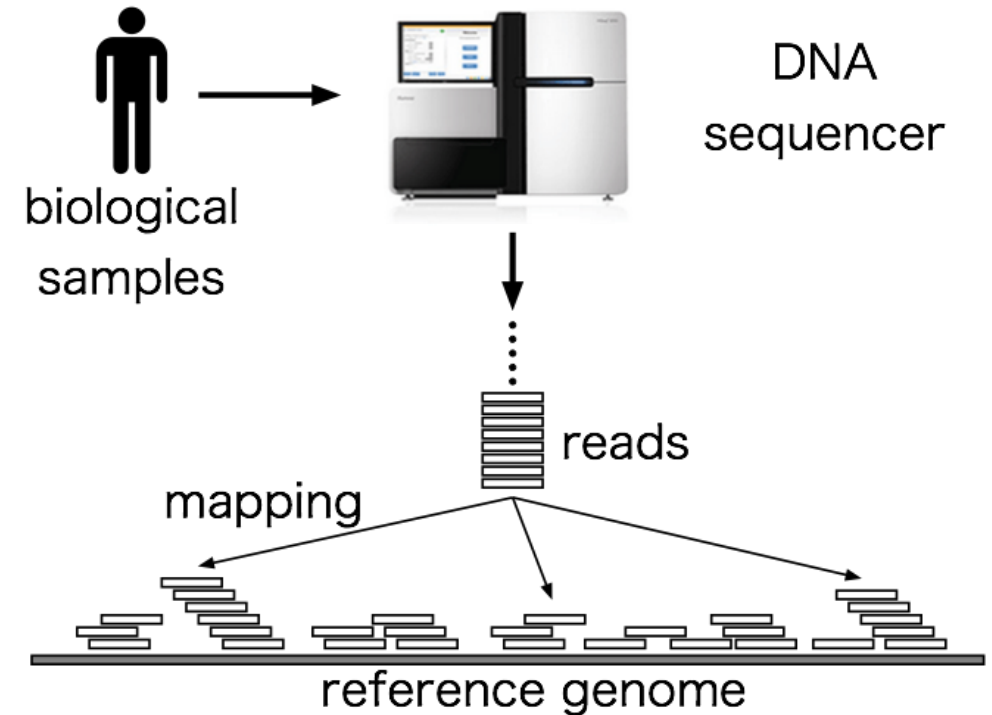
Use “shotgun algorithms” to break into peptides and match to known proteins

- Challenges arise when new proteins are discovered (i.e., not in reference databases)

# Nucleotide sequencing

Identifying the nucleotide sequence of the entire genome, i.e., gene sequencing

Done by fragmenting chromosomes into many pieces and elucidating sequence by overlapping regions



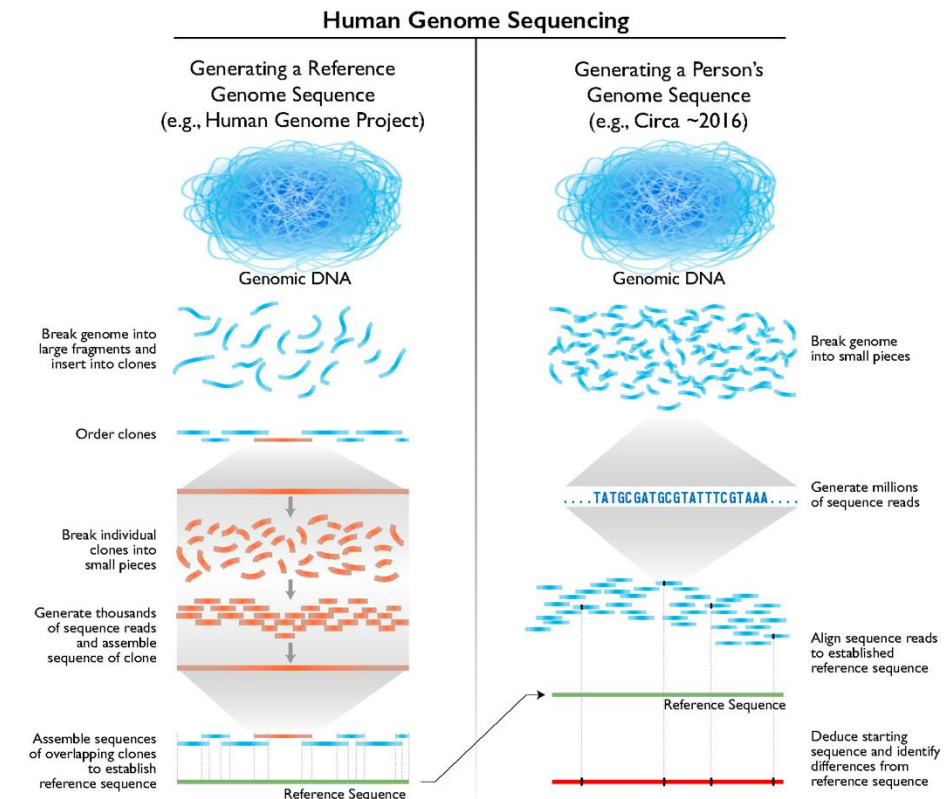
# Next-generation sequencing (NGS)

Sequencing enhanced in last decade by NGS (Goodwin, 2016; Kappelmann-Fenzl, 2021)

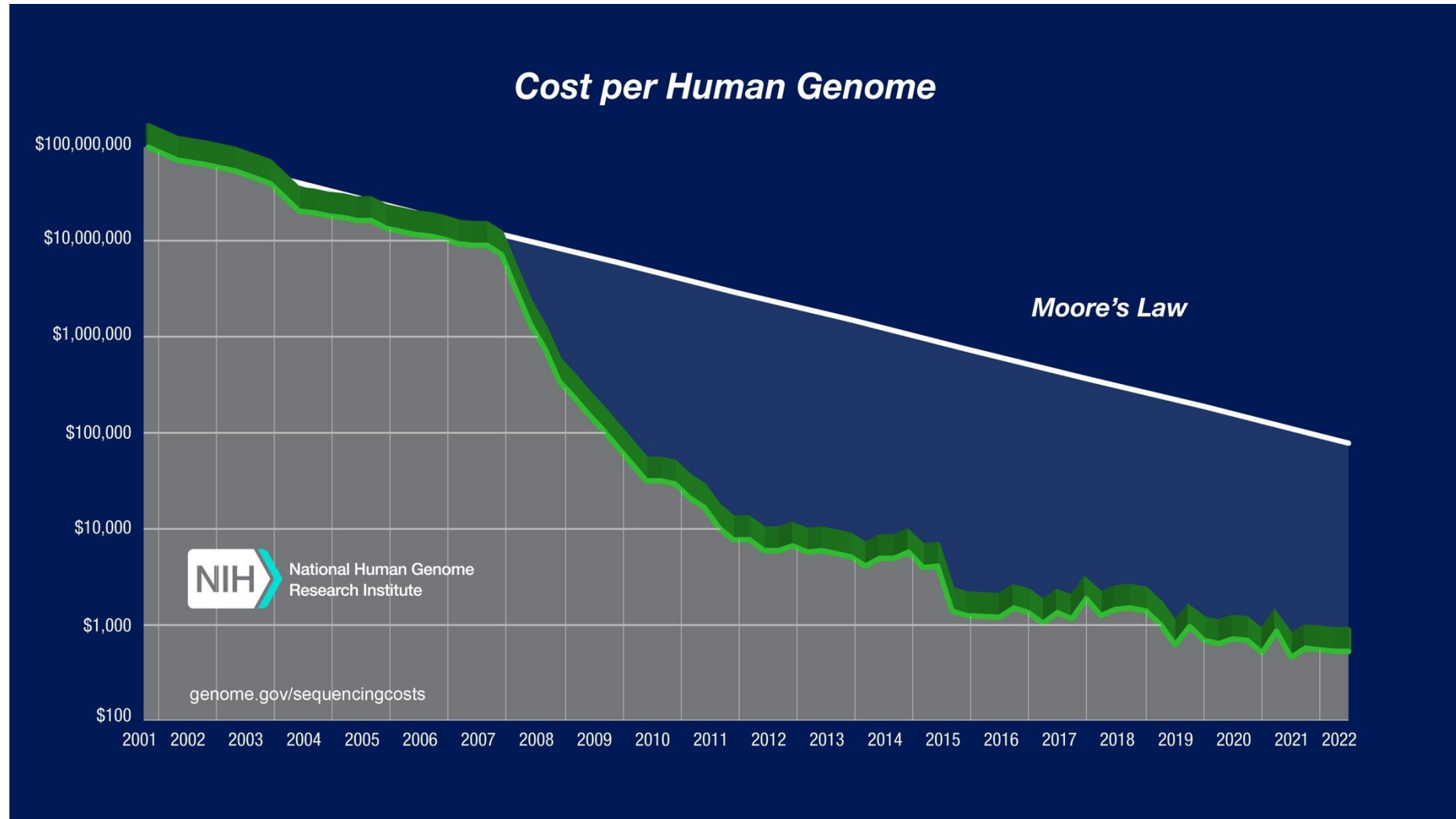
Allows detection of SNPs and CNVs of whole genome, leading to

- Whole genome sequencing – entire genome (Johansen-Taber, 2014; Venner, 2022)
- Whole exome sequencing – focus only on exons

Recent advance is long-read sequencing – allows faster and more efficient reading of longer sequences of nucleotides (Kavaka, 2023; Nature, 2023)



# Cost per genome has fallen rapidly



# Further advances from NGS

---

Extracellular RNA in single droplet of (blood) serum to detect RNA expression in tissues (Zhou, 2019)

“Liquid biopsies” of blood to detect products related to cancer, e.g., DNA, RNA, proteins, etc. (Cohen, 2023; Budnik, 2024; Alix-Panabières, 2025)

- Additional clinical validation needed (Nature Medicine, 2025)

Identification of DNA and RNA from infectious pathogens (Gwinn, 2019)

“Same-day” sequencing (Wojcik, 2025)

# Gene expression (Goodwin, 2016)

---

Sequencing of gene is just a “parts list;” what genes are actually expressed

Early gene expression measurement used microarrays, allowing measurement of many different biological substances, such as DNA, RNA, and protein on a chip

- Initial application was to measure expression of genes (mRNA)

Microarrays largely supplanted by RNA-seq, which more reliably (though expensively) measure mRNA expression (Stark, 2019)

# Protein structure prediction

---

Long-unsolved research problem was prediction of protein folding, i.e., how linear DNA and their amino acids fold into 3-D structures

Solved by Google researchers with AlphaFold (Jumper, 2021)

- Awarded [2024 Nobel Prize for Chemistry](#)

Led to development of [AlphaFold Protein Structure Database](#) (Bertoni, 2025)

- Subsequent ability to predict interactions of proteins and other biomolecules, [accessible via server](#)
- Has revolutionized many aspects of biological science (Callaway, 2025)

Most recent advance is AlphaGenome, with ability to predict impact of genetic variants (Avsec, 2026)

# Clustered regularly interspaced short palindromic repeats (CRISPR)

---

Allows “editing” of genes (Doudna, 2020; Zhang, 2022; Ahmad, 2022)

- Awarded Nobel Prize in Chemistry in 2020

## Uses

- [Somatic Cell Genome Editing \(SCGE\)](#) program (Saha, 2021)
- Applications in treatment of
  - Heterozygous familial hypercholesterolemia (Horie, 2024)
  - Blood disorders sickle cell anemia and beta thalassemia (Joseph, 2023)
  - Treatment of infant with carbamoylphosphate synthetase 1 (CPS1) deficiency at 7 months (Musunuru, 2025; Gropman, 2025)

Growing use of AI for genome target selection and prediction of functional outcomes (Thomson, 2025)

## Concerns

- Rogue scientist editing of gene in embryo to disable gene playing role in HIV infection (Kolata, 2018)
- Unexpected mutations to other parts of genome (Aussel, 2025)
- “Heritable polygenic editing” possible (Visscher, 2025) but may be unsafe and unproven (Carmi, 2025)

# More information about genetics and genomics

---

National Human Genome Research Institute (NHGRI) – [About Genomics](#)

New England Journal of Medicine – [Illustrated Glossary](#)

MedlinePlus – [Help Me Understand Genetics](#)