Introduction to Genetics and Genomics

2. Molecular Genetics

lachance.joseph@gmail.com
https://popgen.gatech.edu/

2a. Historical perspective

2b. Molecular Biology

Break

2c. Genetic Variation

2d. Technology and Bias

The Double Helix XX-XY
Sculpture by:
Franco Castellucio

Terminology

- **Allele**: One of two or more alternative forms of a gene (e.g. A or G)
- **Gene**: DNA sequence that encodes a functional protein or RNA molecule
- **Genome**: the complete set of genetic material in a cell or organism
- **Chromosome**: threadlike structure of nucleic acids and proteins found in the nucleus
- **Haplotype**: A set of linked alleles that are inherited together
- **kb** (**kilobase**): one thousand base pairs, **Mb** (**megabase**): 1 million bp

Mendel's laws of inheritance

- **Law of segregation** (1st law)
  - Parental pairs of alleles separate during gamete formation
- **Law of independent assortment** (2nd law)
  - Pairs of alleles for different traits segregate independently
- **Law of dominance** (3rd law)
  - Heterozygotes manifest the trait associated with the dominant allele
- *These rules are often broken!*

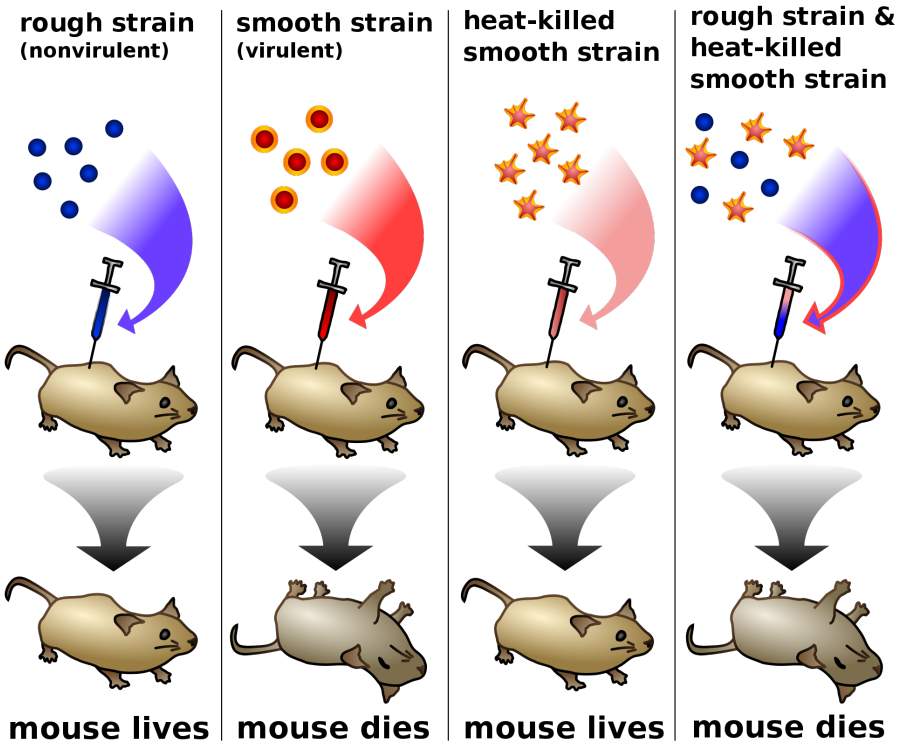
Morgan

- Sex linkage
- Chromosomal theory of inheritance
- Genetic linkage and crossing over

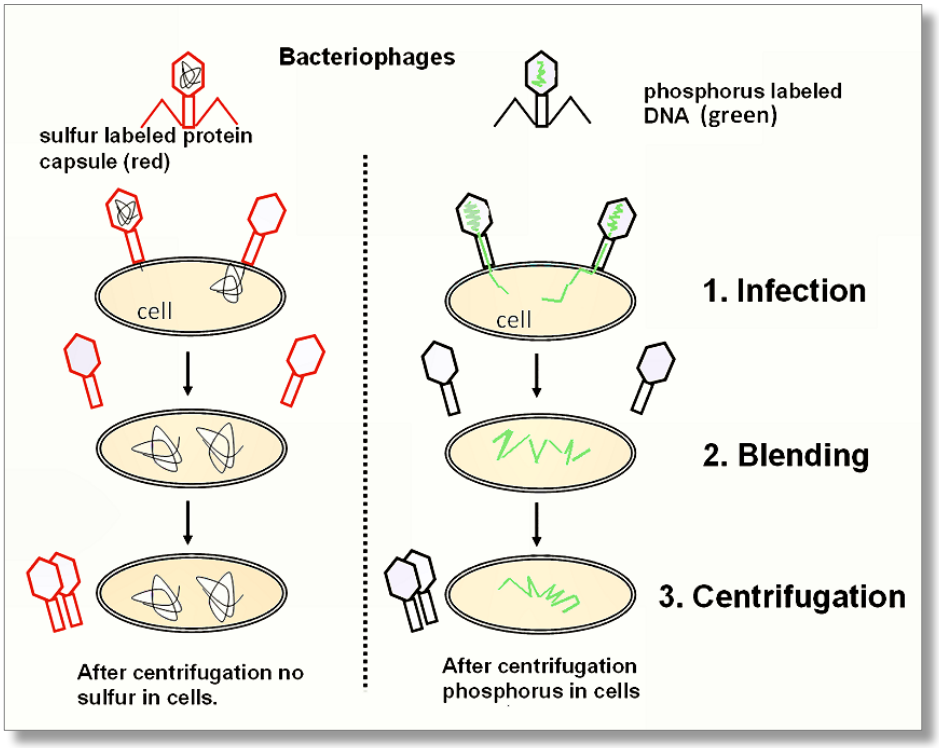
The search for the transforming factor

- Griffith

- Avery, MacLeod, and McCarty



Hershey-Chase: DNA is the hereditary material



Watson and Crick: double helix structure of DNA

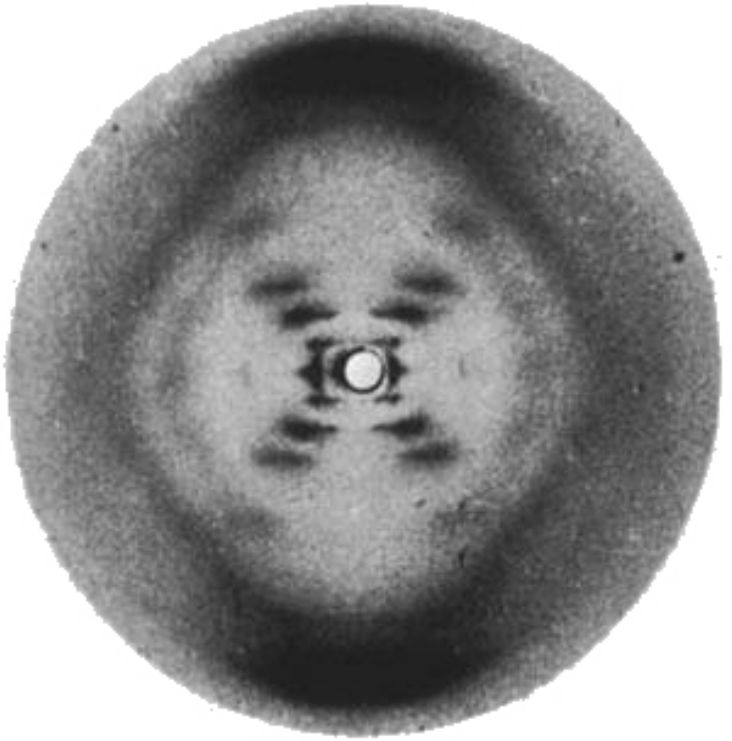
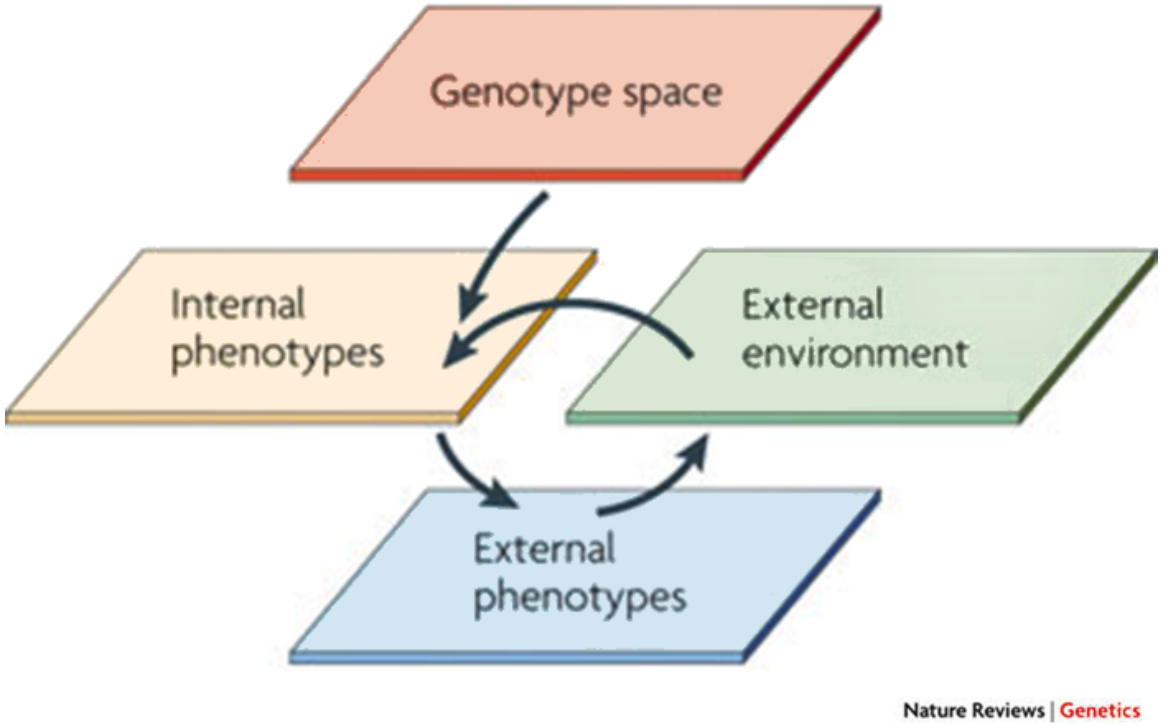


Photo 51: X-ray diffraction of DNA
(Gosling and Franklin)

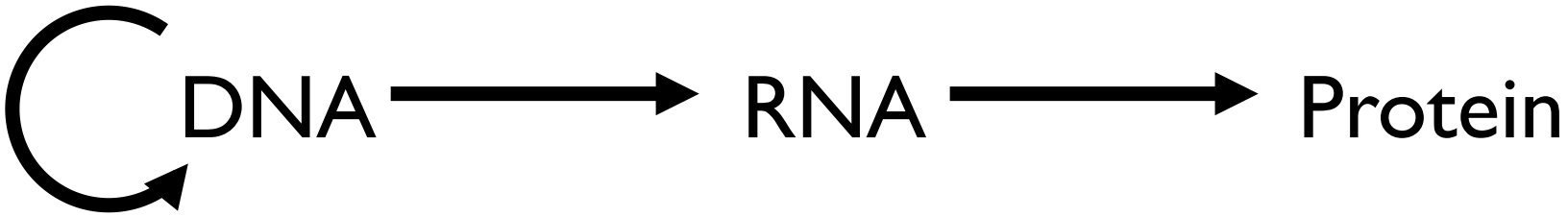
Information flow in genetics

Image rights: Ramona Saldamando

Genotype-phenotype map



Central Dogma of Molecular Biology*



*Things are not quite this simple!

What are some exceptions to the Central Dogma?

Central Dogma: implications

- Mendelism vs. Lamarckism (acquired characteristics)
- Germline vs. Soma (Weismann)
- Genes as information - decoupling of structure and function
- Biological “laws” are full of exceptions

DNA

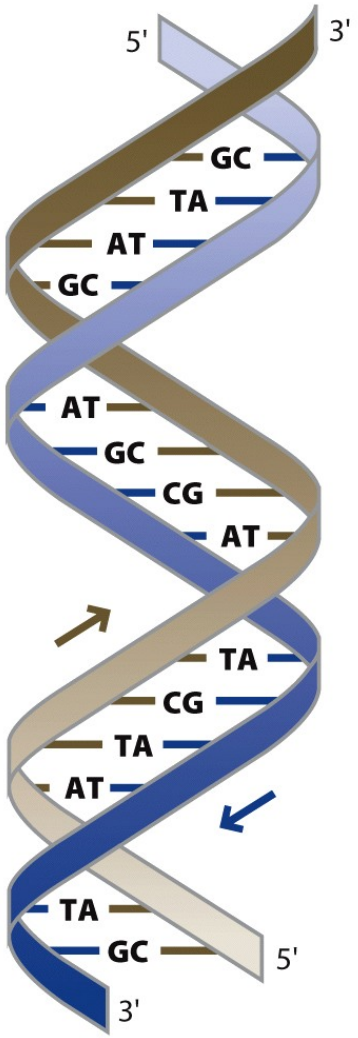


Figure 2.4b Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

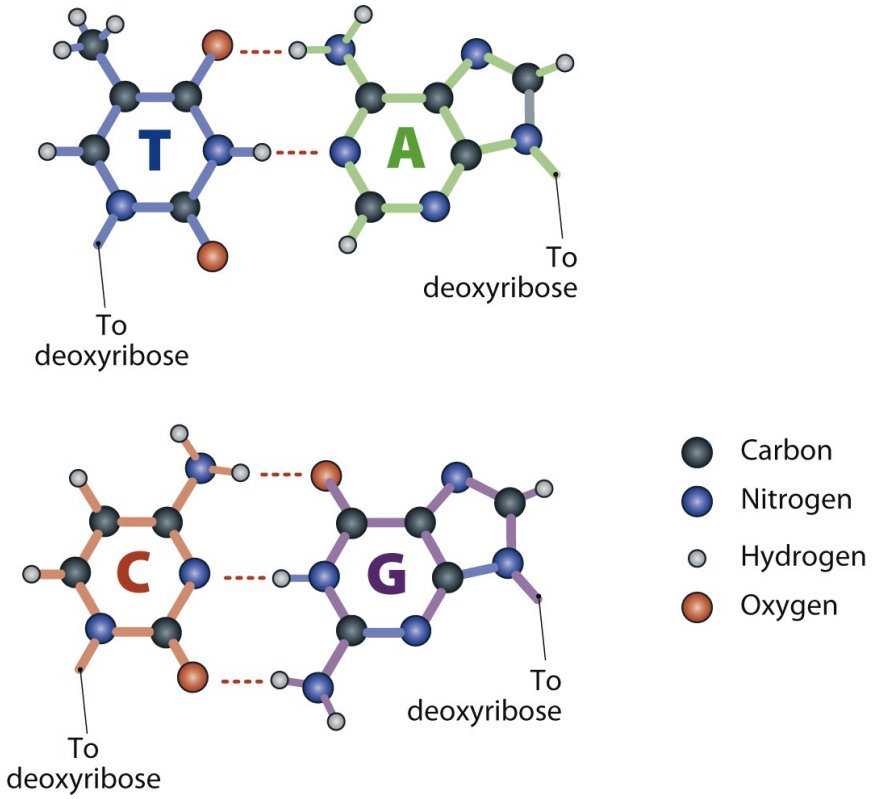


Figure 2.5 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

DNA packaging

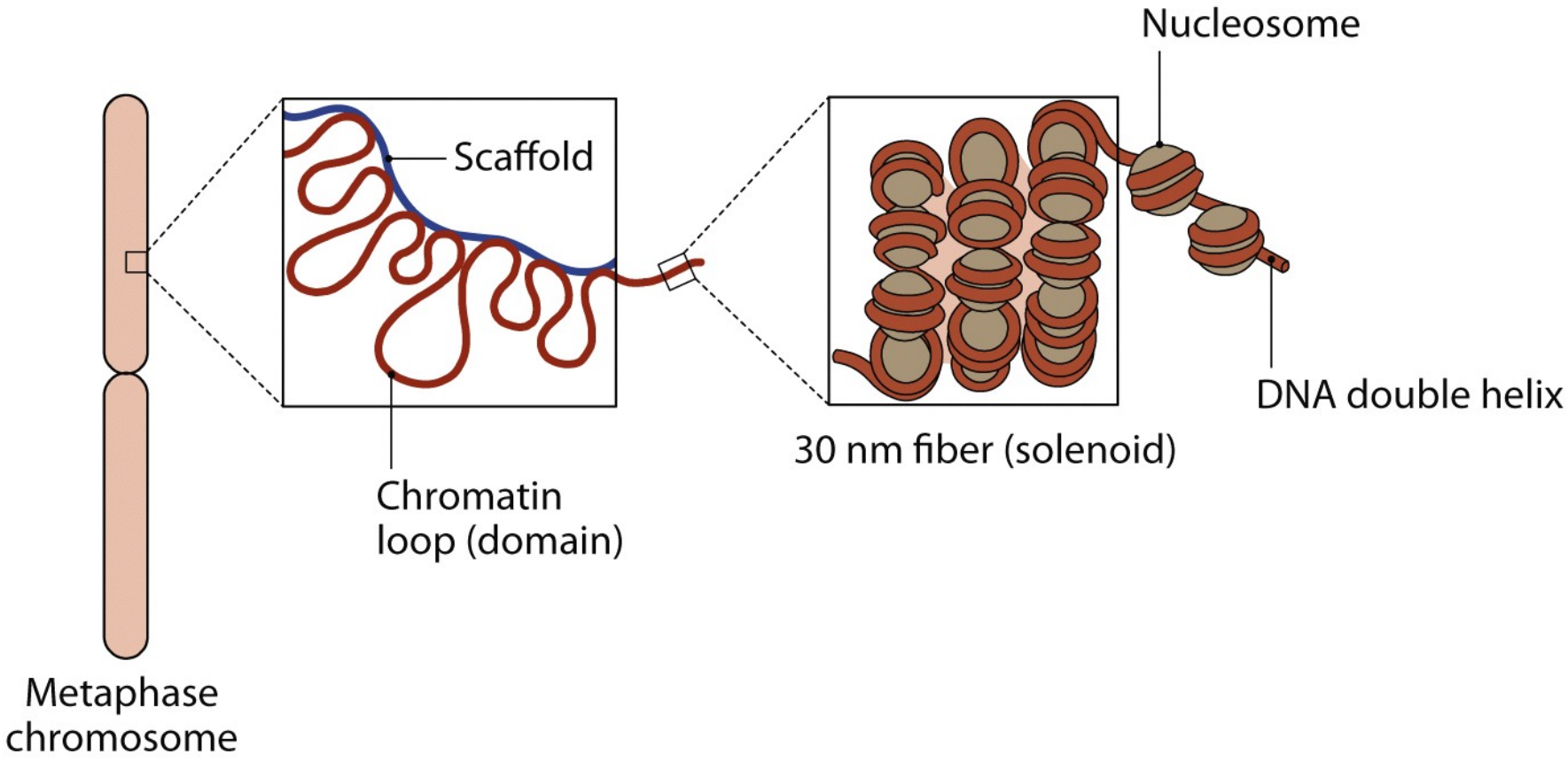
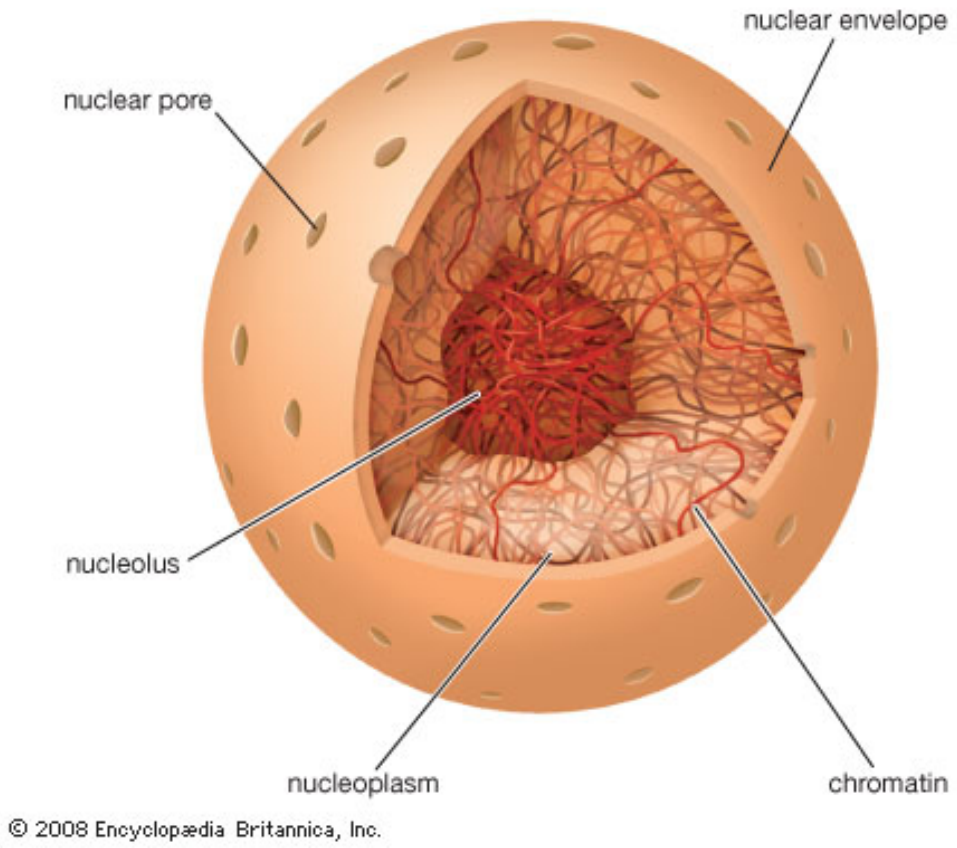


Figure 2.11 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

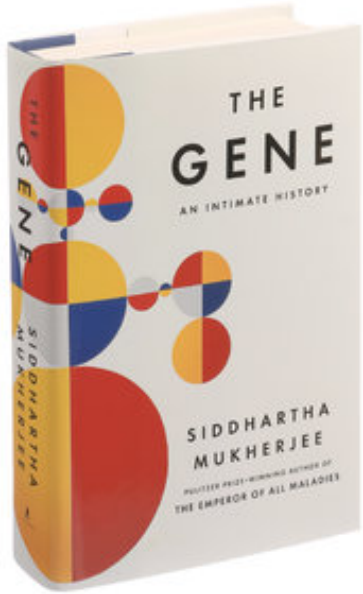
DNA packaging (movie clip)

Chromatin

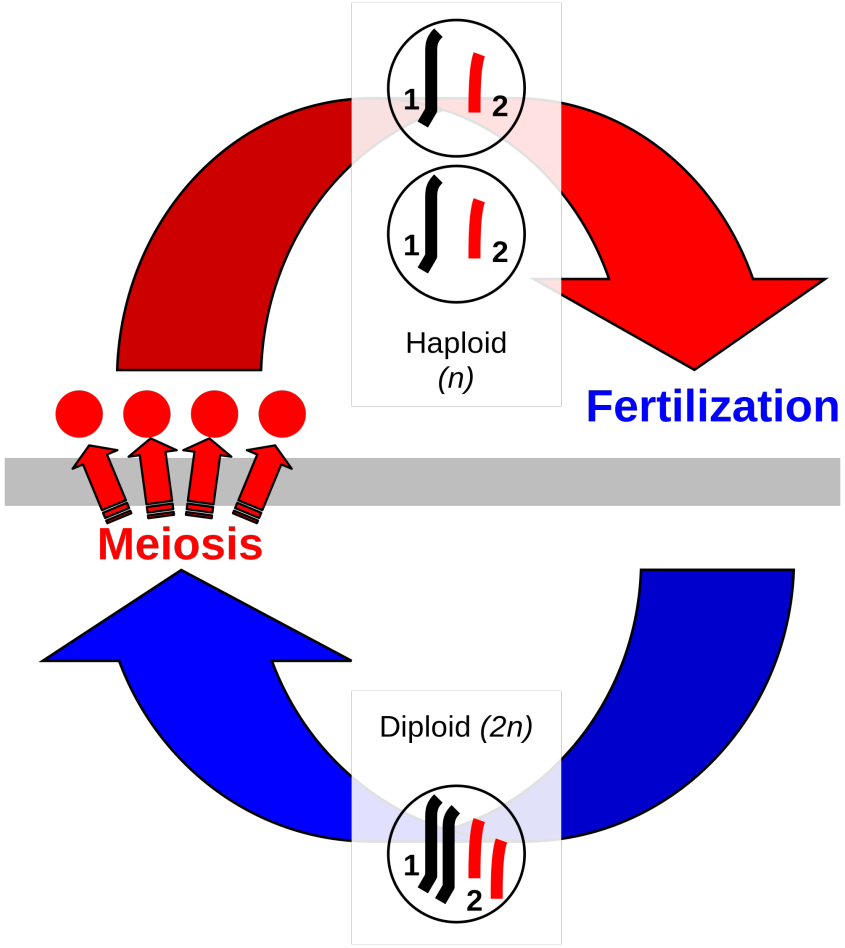


DNA packaging: implications

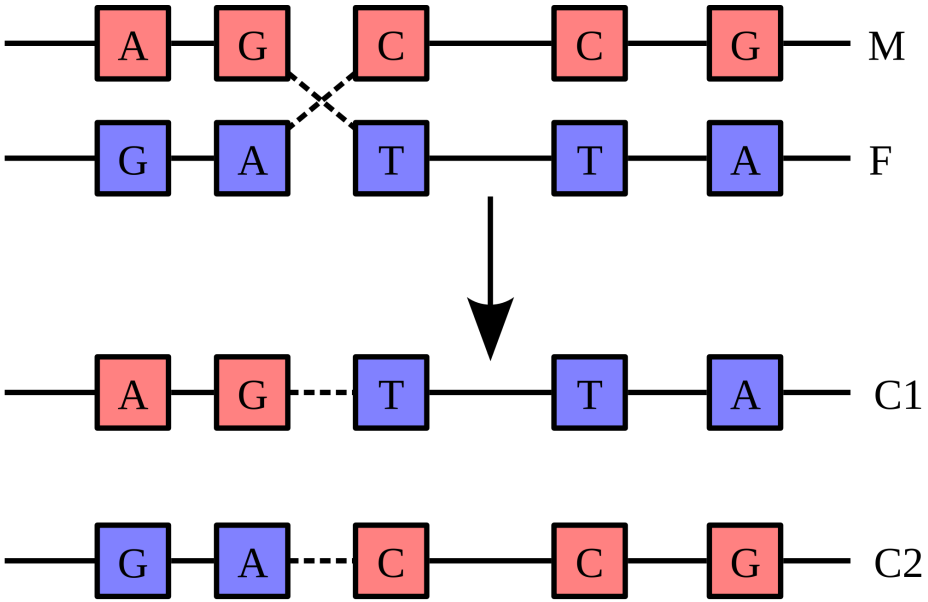
- Exposed DNA is more likely to be functional
- Proximity in 3D space matters
- Histone code
- Mukherjee: overstating the importance of epigenetics?



Ploidy



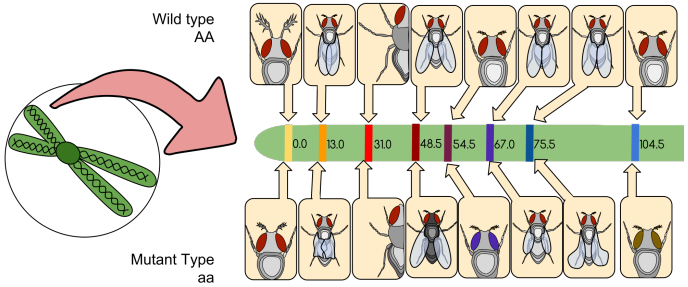
Recombination



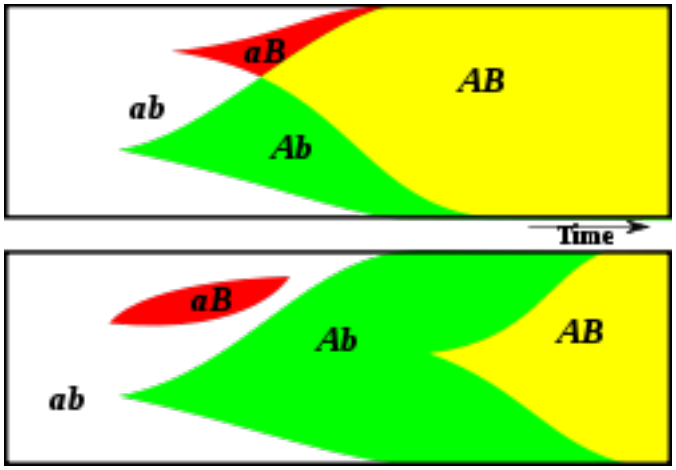
- Occurs in meiosis
- Byproduct of the need to pair homologous chromosomes

Recombination: implications

- Genetic maps and linkage disequilibrium



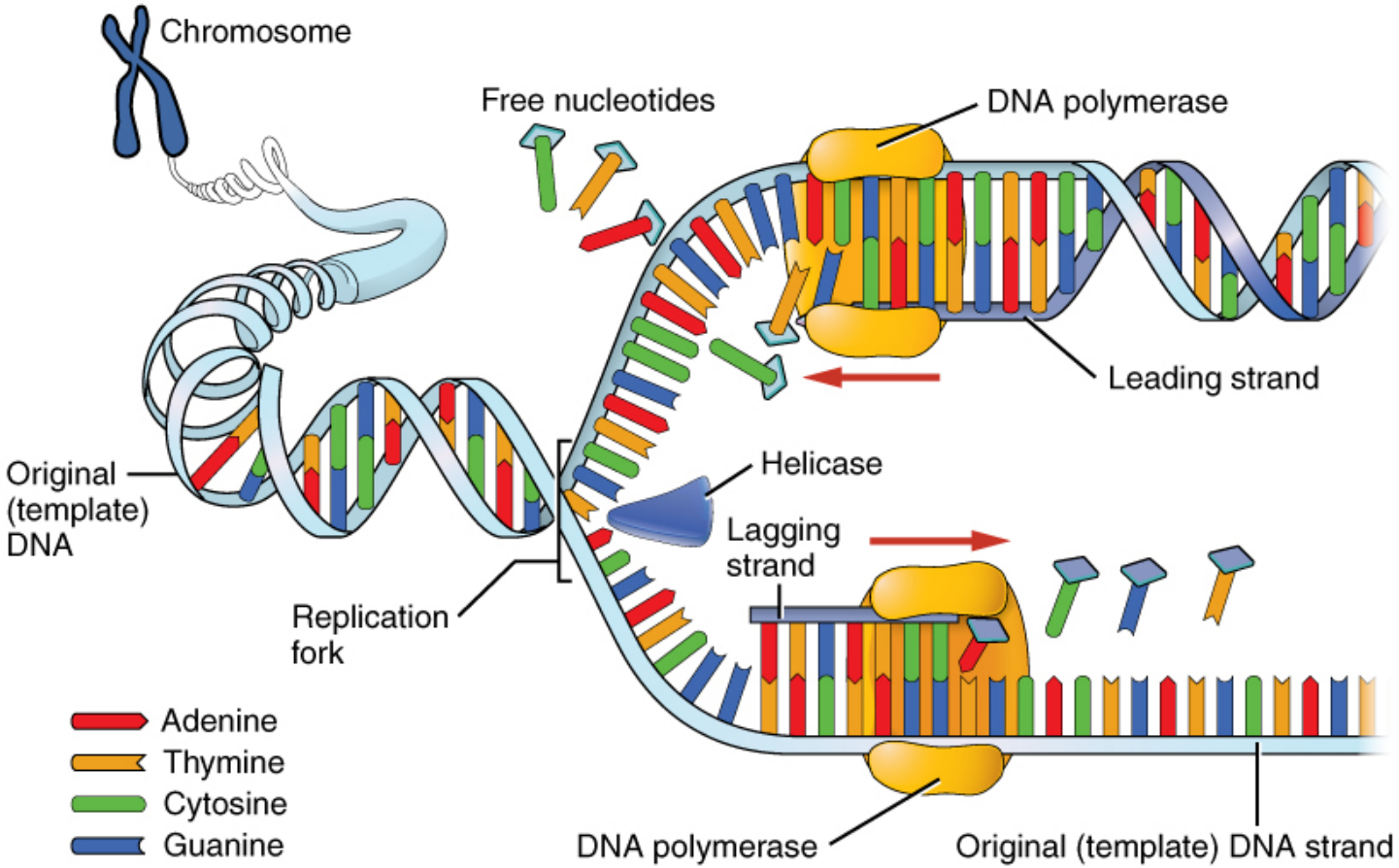
- Benefits of sex



Sexual reproduction (recombination)

Asexual reproduction

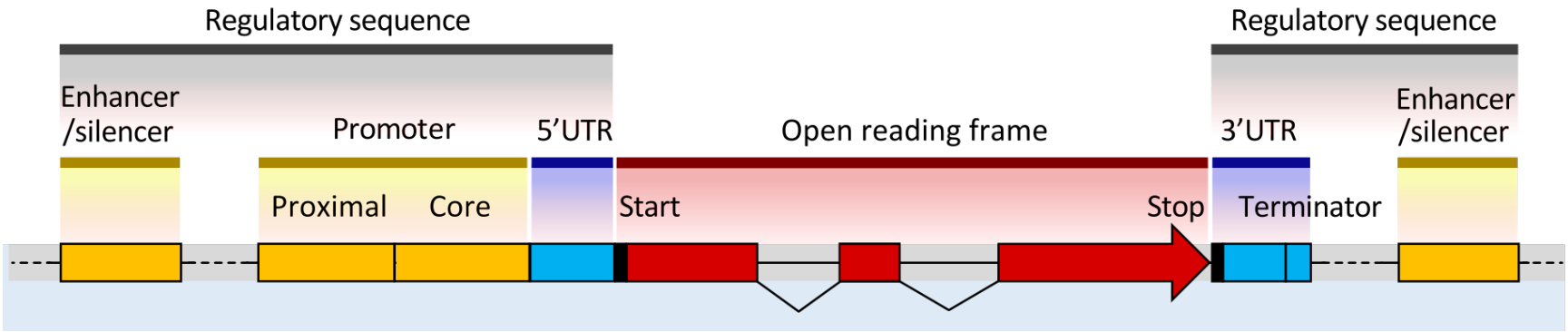
DNA replication



DNA replication: implications

- Semi-conservative replication
- 5′ → 3′ directionality causes problems (solved by evolution)
- Potential for miscopying → **mutations**
- Digital information enables comparative genomics

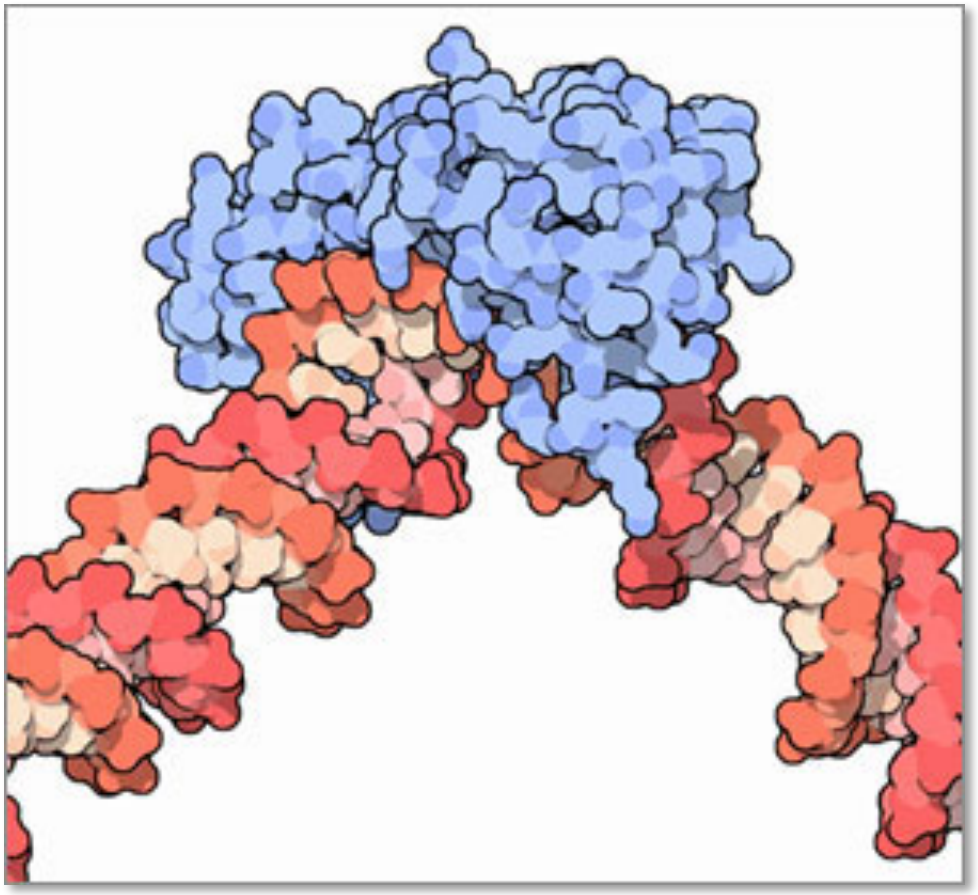
The structure of (protein coding) genes



- Cis-regulatory elements
  - Enhancers: increase the likelihood of transcription when bound to activators
  - Silencers: decrease likelihood of transcription when bound to repressors
  - Promoters: region of DNA where transcription is initiated
- UTRs: untranslated regions
- Exons: nucleotide sequence not removed by splicing (coding DNA)
- Introns: nucleotide sequence removed by splicing (noncoding DNA)

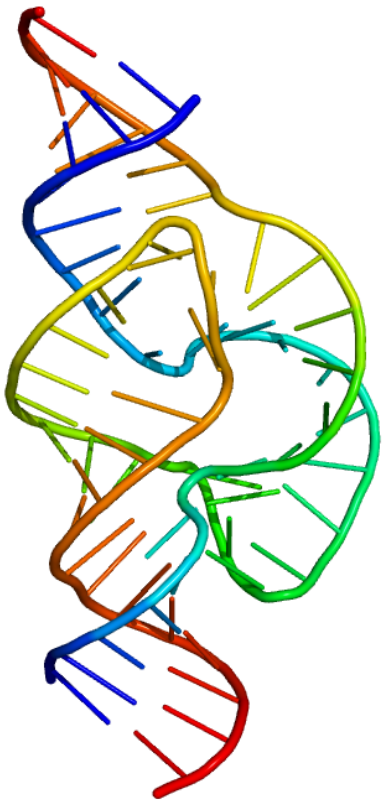
- *How would you define a **gene**?*

Transcription factors and gene regulation



RNA comes in many different flavors

- mRNA: messenger RNA
- tRNA: transfer RNA
- rRNA: ribosomal RNA
- Regulatory RNAs (miRNA, siRNA, piRNA)



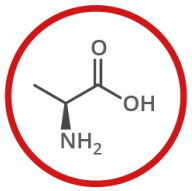
Ribozyme

Proteins are made of amino acids

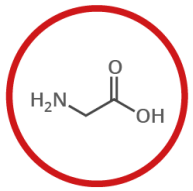
Chart Key: ALIPHATIC · AROMATIC · ACIDIC · BASIC · HYDROXYLIC · SULFUR-CONTAINING · AMIDIC · NON-ESSENTIAL · ESSENTIAL

Chemical Structure — single letter code

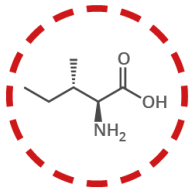
NAME (A) — three letter code — DNA codons



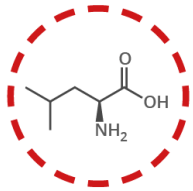
ALANINE (A)
Ala
GCT, GCC, GCA, GCG



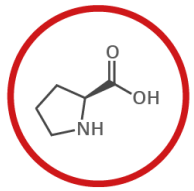
GLYCINE (G)
Gly
GGT, GGC, GGA, GGG



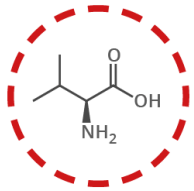
ISOLEUCINE (I)
Ile
ATT, ATC, ATA



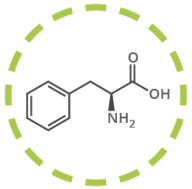
LEUCINE (L)
Leu
CTT, CTC, CTA, CTG, TTA, TTG



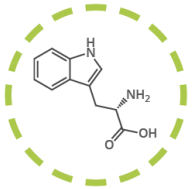
PROLINE (P)
Pro
CCT, CCC, CCA, CCG



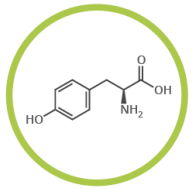
VALINE (V)
Val
GTT, GTC, GTA, GTG



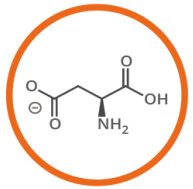
PHENYLALANINE (F)
Phe
TTT, TTC



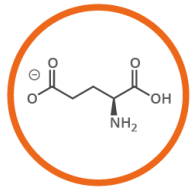
TRYPTOPHAN (W)
Trp
TGG



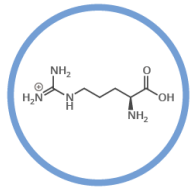
TYROSINE (Y)
Tyr
TAT, TAC



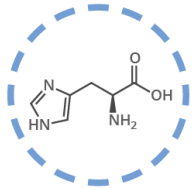
ASPARTIC ACID (D)
Asp
GAT, GAC



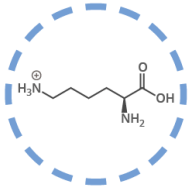
GLUTAMIC ACID (E)
Glu
GAA, GAG



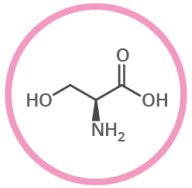
ARGININE (R)
Arg
CGT, CGC, CGA, CGG, AGA, AGG



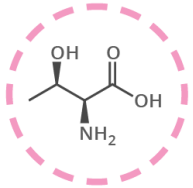
HISTIDINE (H)
His
CAT, CAC



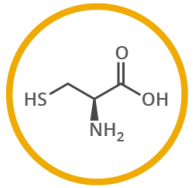
LYSINE (K)
Lys
AAA, AAG



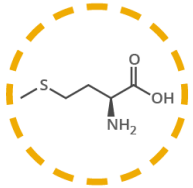
SERINE (S)
Ser
TCT, TCC, TCA, TCG, AGT, AGC



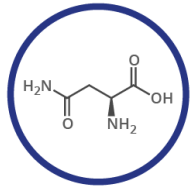
THREONINE (T)
Thr
ACT, ACC, ACA, ACG



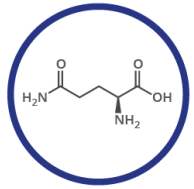
CYSTEINE (C)
Cys
TGT, TGC



METHIONINE (M)
Met
ATG



ASPARAGINE (N)
Asn
AAT, AAC



GLUTAMINE (Q)
Gln
CAA, CAG

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

From DNA to RNA to protein

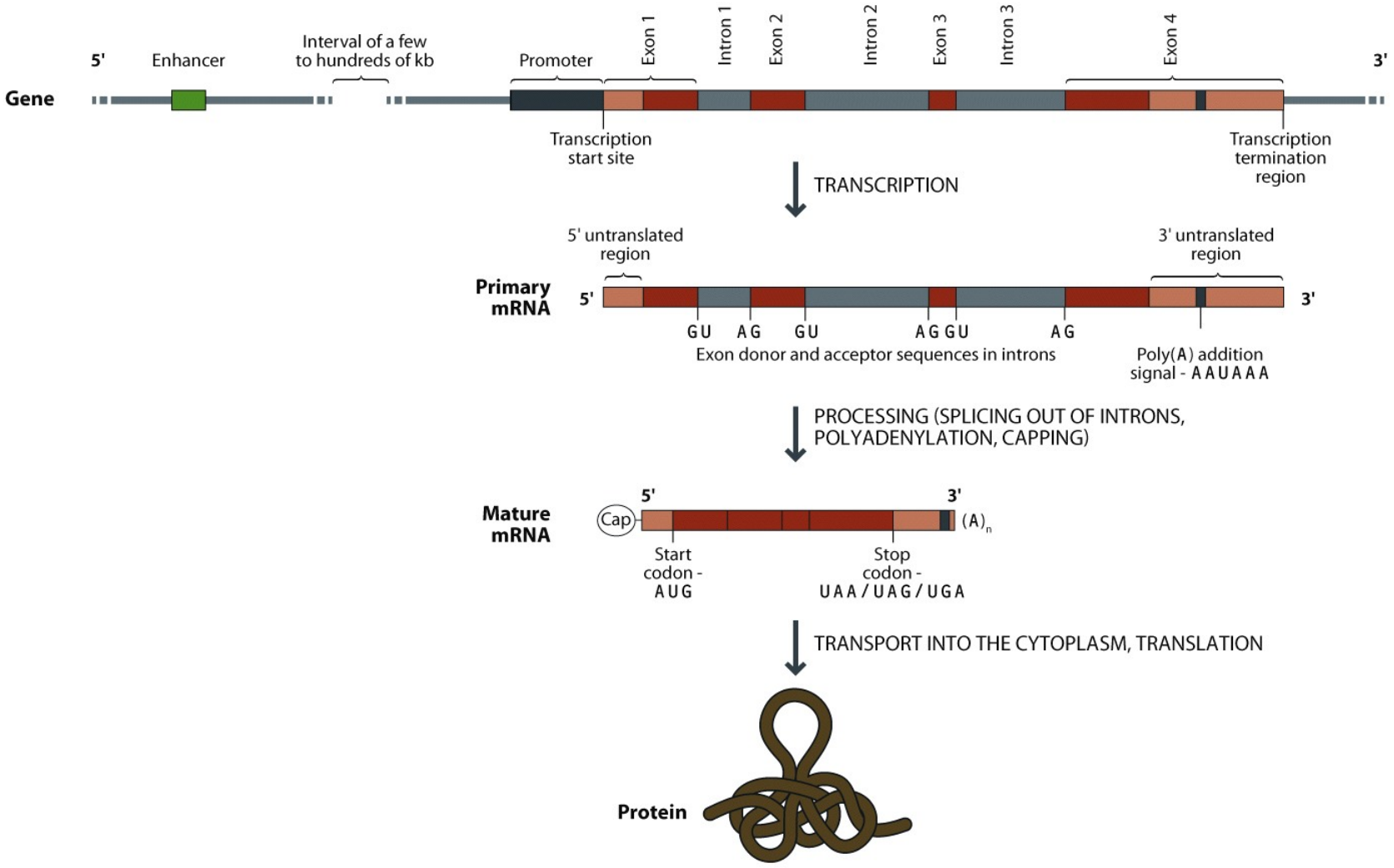


Figure 2.6 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

Transcription: DNA serves as a template

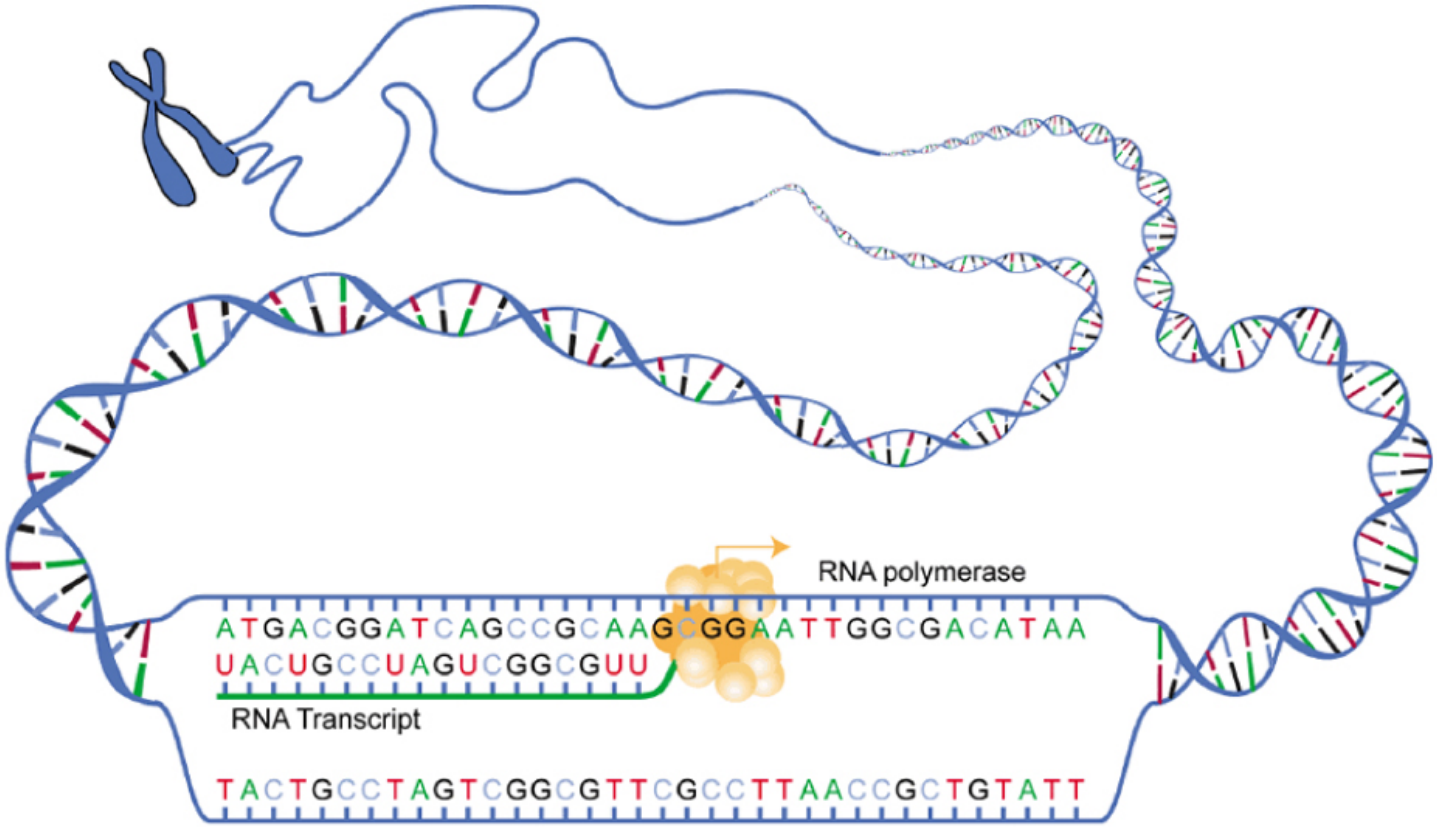
5′ … CGATCGGACTACGGACTAGCGACTACGA … 3′ **Sense strand of DNA**

3′ … GCTAGCCTGATGCCTGATCGCTGATGCT … 5′ **Antisense strand of DNA**

Transcription of antisense strand

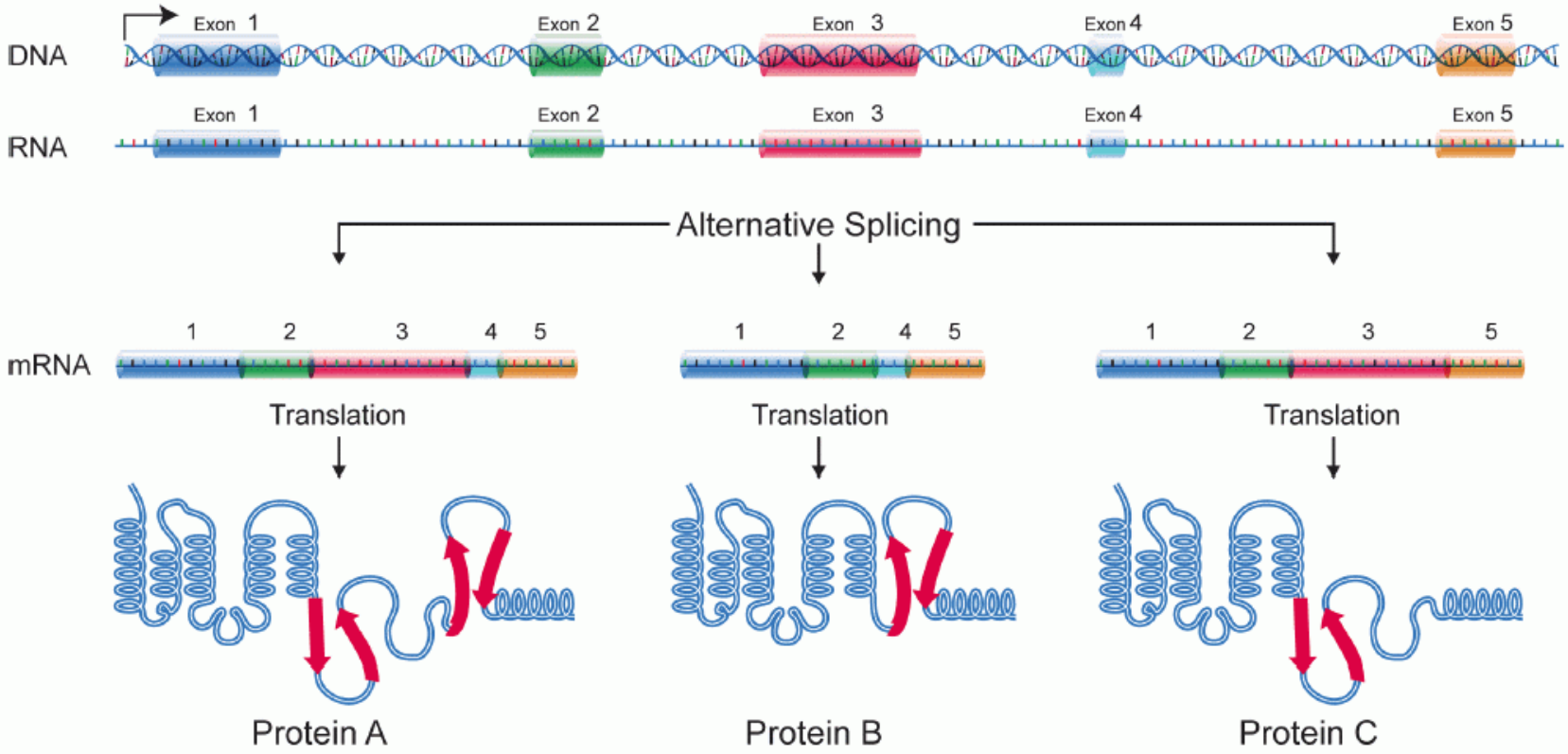
5′ … CGAUCGGACUACGGACUAGCGACUACGA … 3′ **RNA**

Transcription (DNA to RNA)



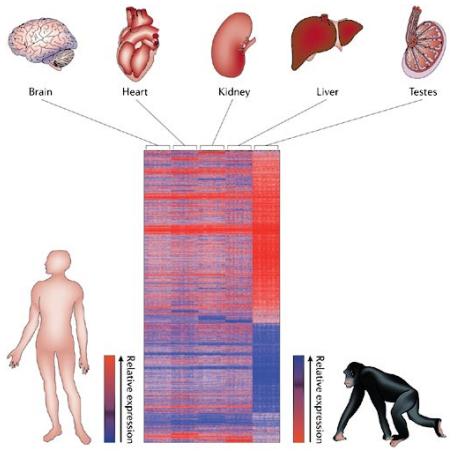
Transcription (movie clip)

Splicing

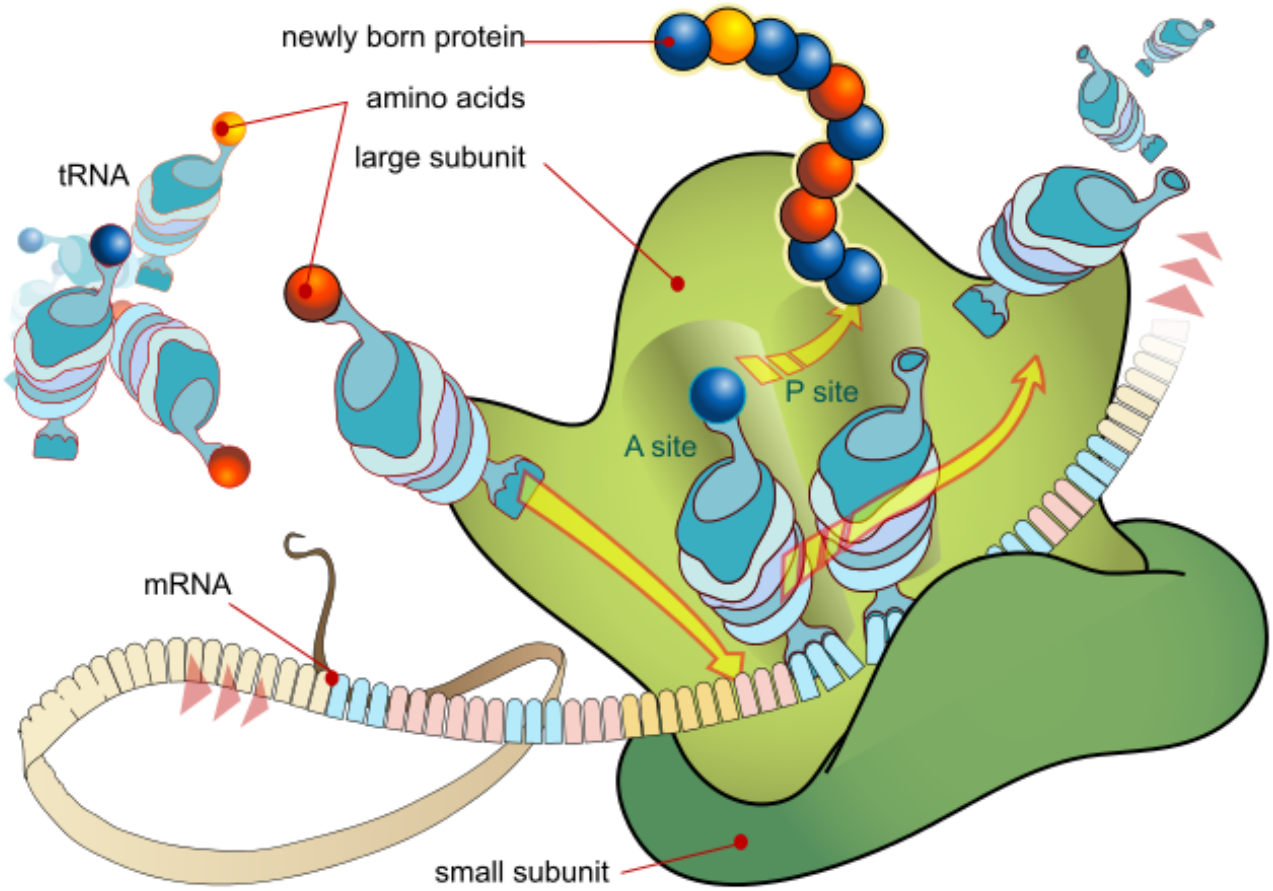


Transcription: implications

- **Gene expression**: transcriptional activity of a gene that results in RNA
- Inducible system that allows organisms to respond to environments
- Helps explain how different cell types can share same DNA

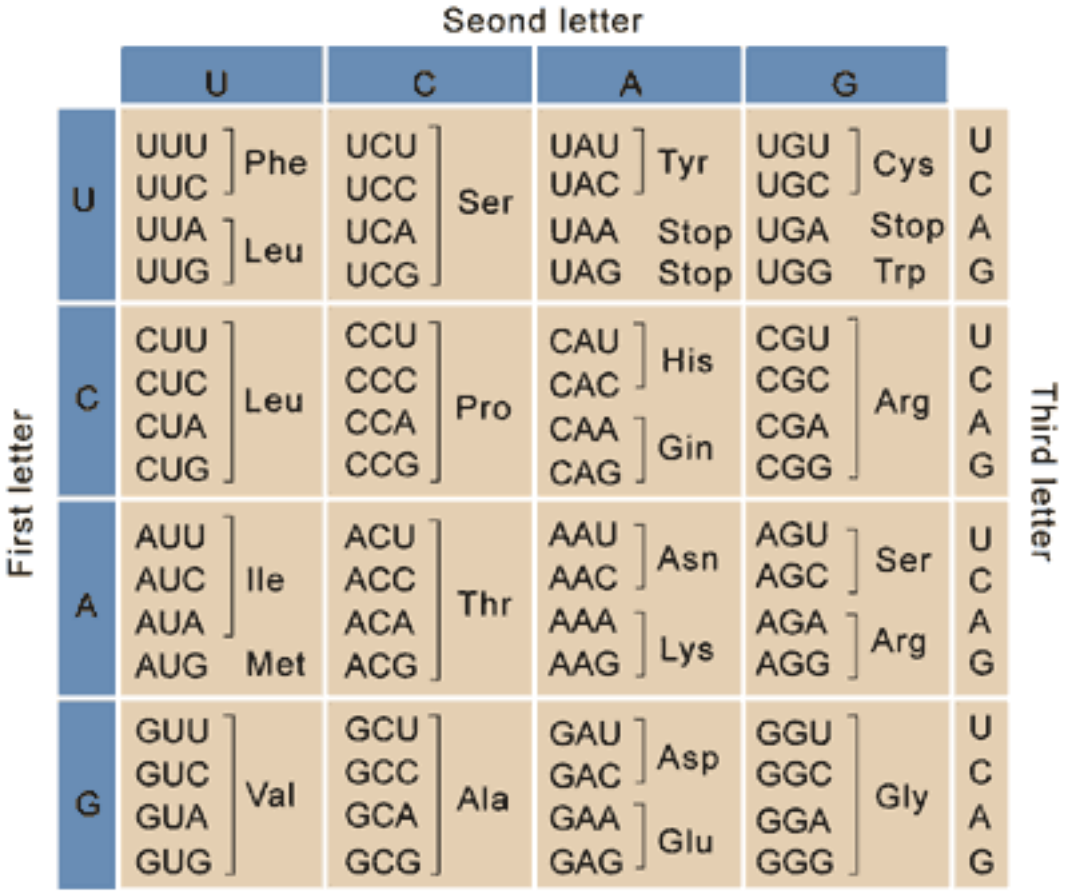


Translation (RNA to protein)



Translation (movie clip)

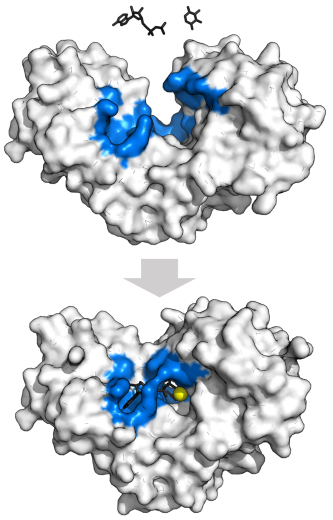
The genetic code



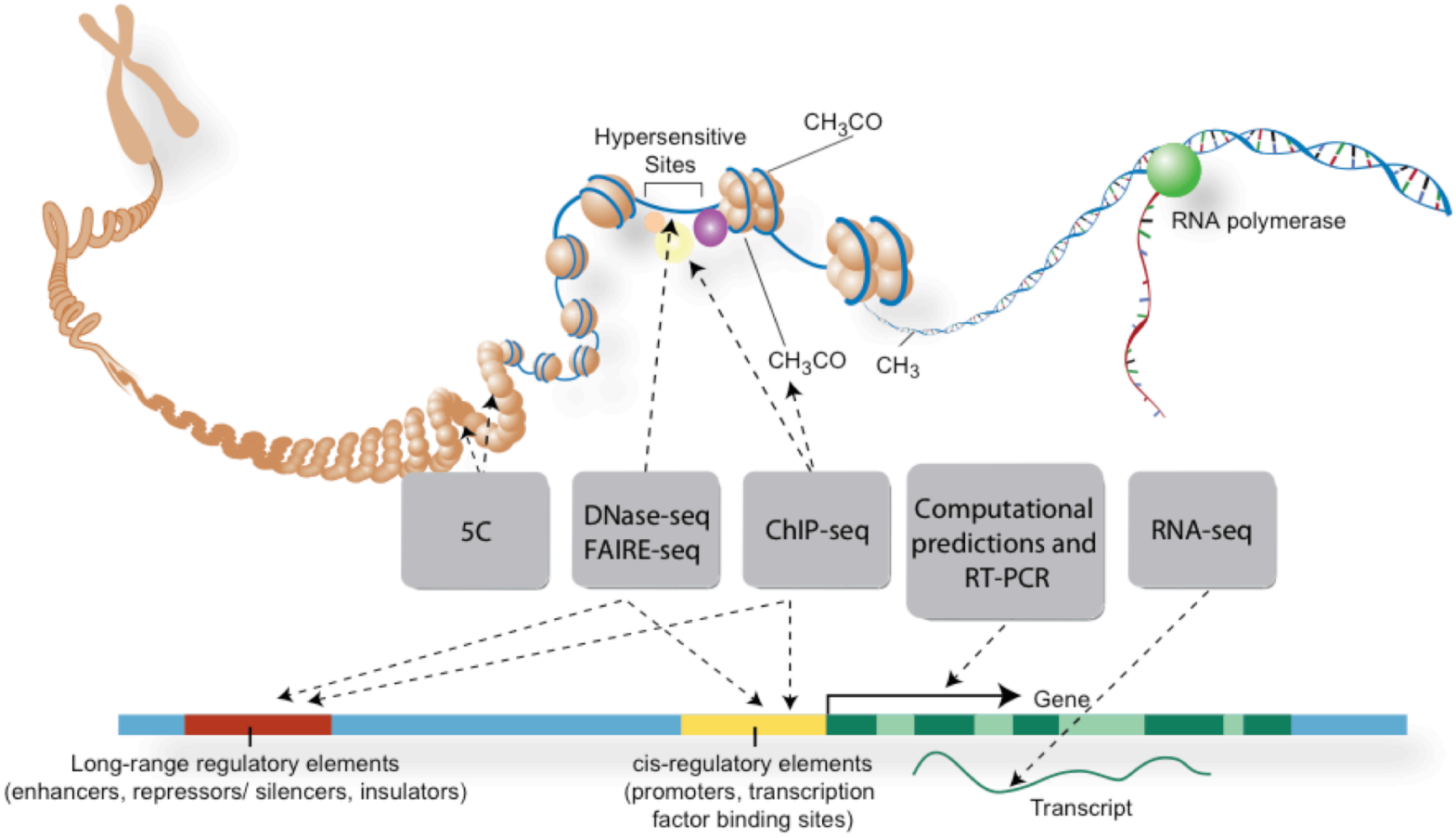
What about alternative codes?

Translation: implications

- The genetic code is (relatively) arbitrary... frozen accident?
- Phase
- Post-translational modifications (e.g. glycosylation)
- **Enzymes**: a substance produced by a living organism that catalyzes a specific biochemical reaction. Enzymes are made of proteins



ENCODE and the debate about “function”

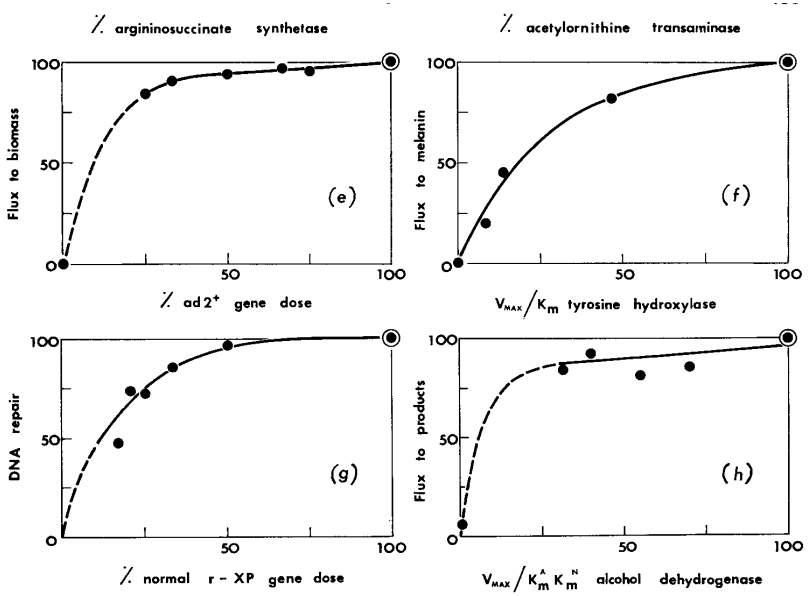


How would you differeniate functional and nonfunctional DNA?

	Prokaryotes	Eukaryotes
Internal structures	No organelles	Organelles
DNA	No histones Circular No introns DNA in cytoplasm	Histones Linear Introns DNA in nucleus
Genome size	Most <5Mb	10Mb-100,000Mb
Chromatin	No histones	Histones
Ploidy	Haploid	Usually diploid
Reproduction	Asexual (binary fission)	Asexual (mitosis) and sexual (meiosis)

Dominance and recessivity

- Kacser and Burns 1981 (*Genetics*)
- Dominance can arise as an emergent property of metabolic flux

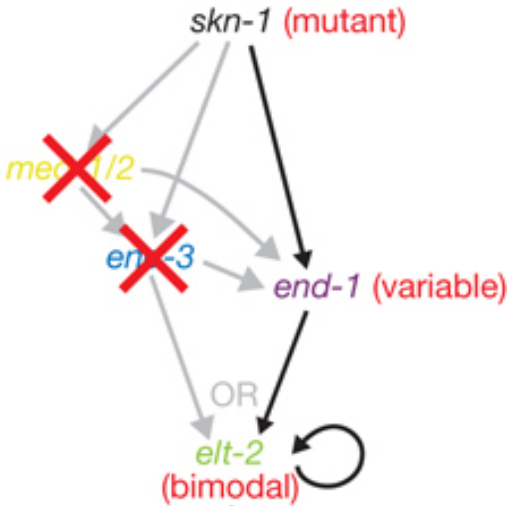


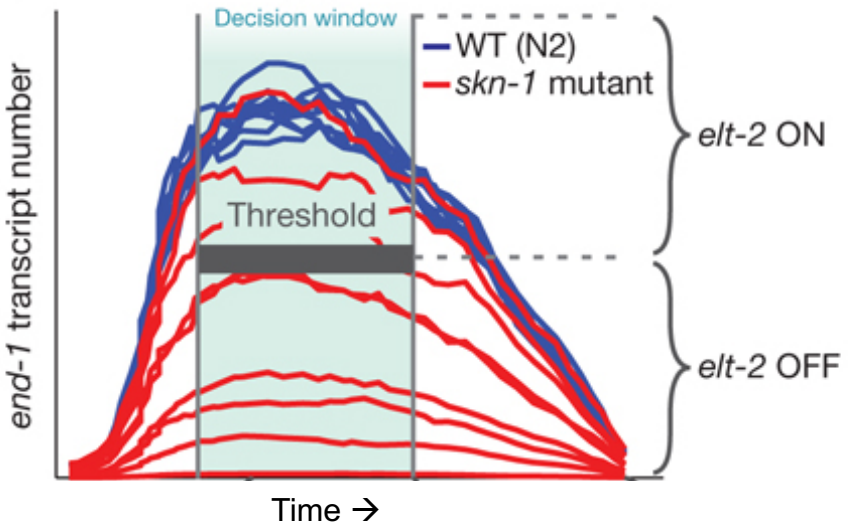
- Haldane's Sieve: mutations that reach fixation tend to be dominant

Pleiotropy

- It is incorrect talk say that something is “a blank gene” (e.g. a cancer gene)
- **Pleiotropy**: when a gene produces multiple phenotypic effects
- Indirect result of the Central Dogma of Molecular Biology
- *Frizzle* mutation results in feathers that curve outward, fewer eggs laid, and high temperatures

Incomplete penetrance

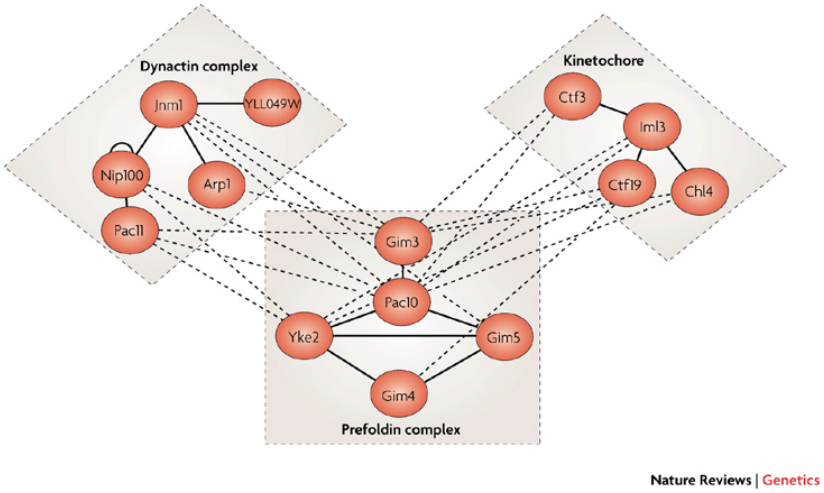




- **Penetrance**: proportion of individuals with a given genotype that show the expected phenotype
- Raj et al 2010 (*Nature*)
- Variability in gene expression + threshold → incomplete penetrance

Epistasis (genetic interactions)

- Epistasis can arise from physical interactions
- Think of transcription factors and cis-regulatory elements…



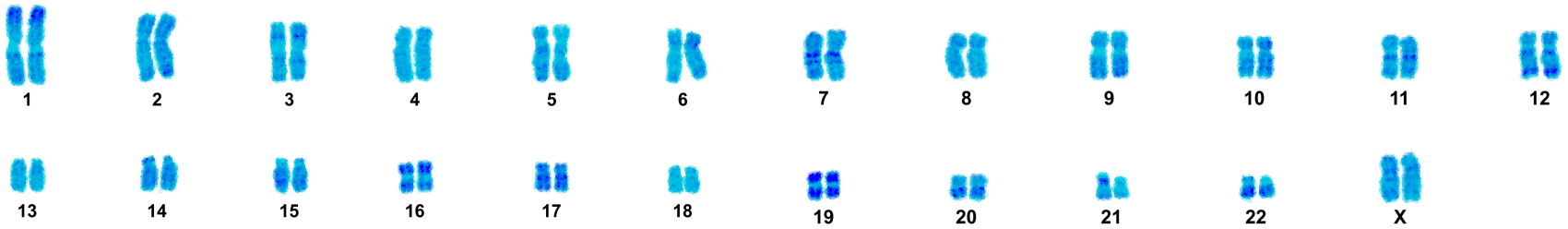
- Fitness interaction networks vs. physical interaction networks: not the same!
  - Beyer et al 2007 (*NRG*)

Break

Variation

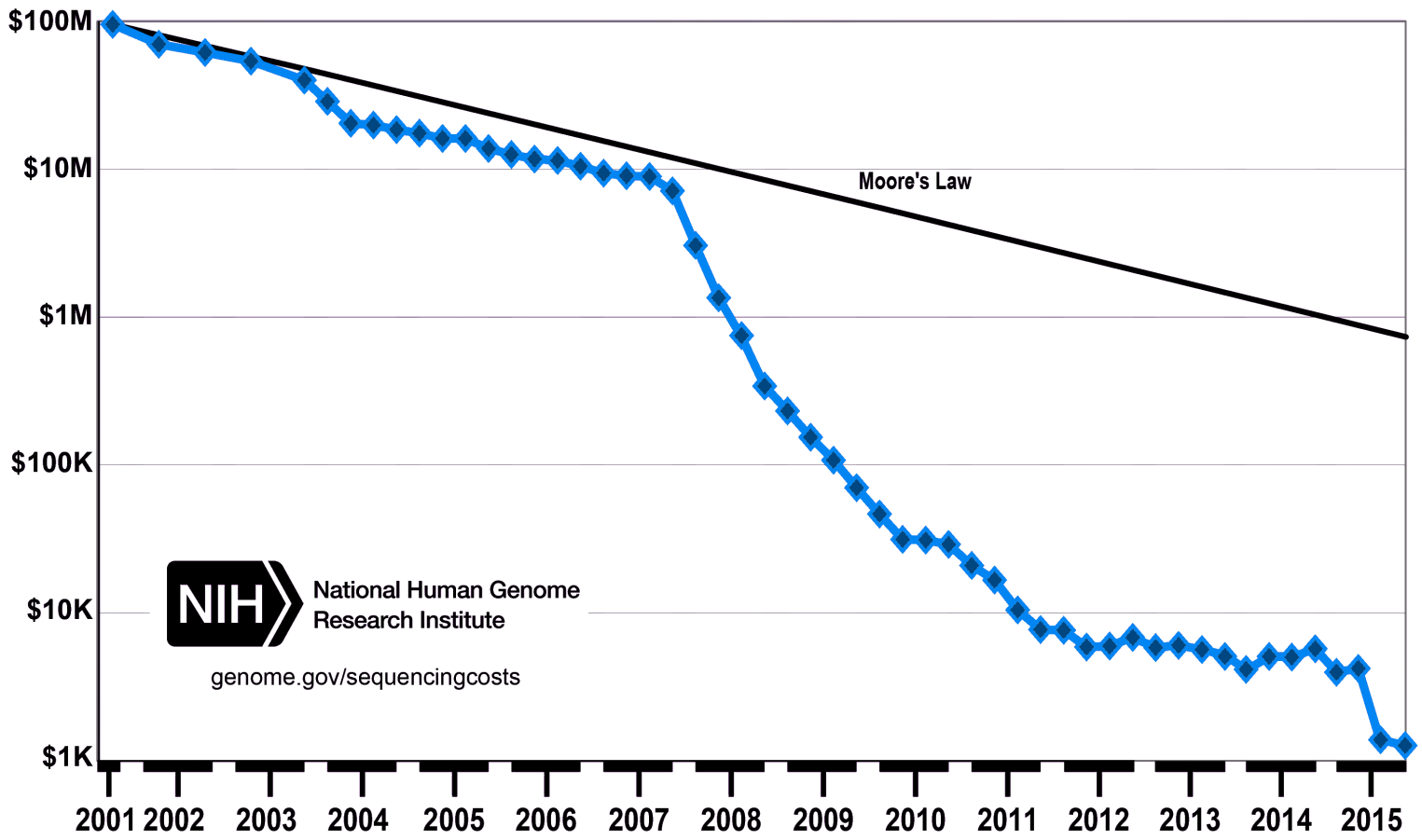
National Geographic

The human genome



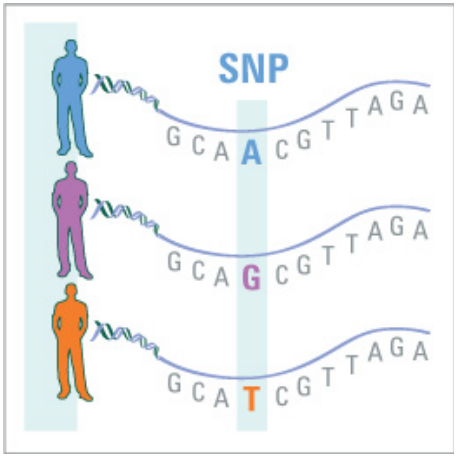
- Approximately 3.2 billion base pairs
- 23 pairs of chromosomes
  - 22 autosomes
  - One pair of sex chromosomes (XX or XY)
  - mtDNA (16.6kb)
- A typical genome
  - Heterozygous at 1 out of every 1000 sites
  - 44% transposable elements!!
  - 1.1% coding DNA

Declining sequencing costs

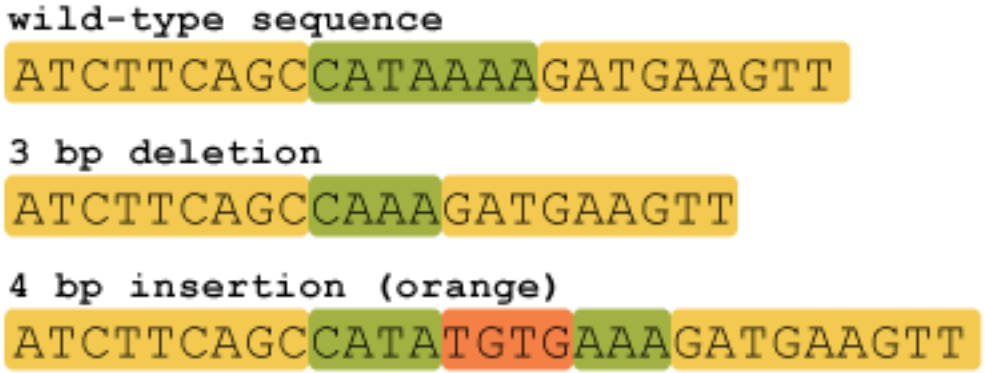


SNPs

- **S**ingle **N**ucleotide **P**olymorphisms (SNPs): single letter changes in DNA
- Human genomes have between 3.5 million to 4.3 million SNPs (African genomes have more SNPs)
- dbSNP: 153 million SNPs and counting...
- Most SNPs are biallelic
- Most SNPs have a rare a rare derived allele and a common ancestral allele



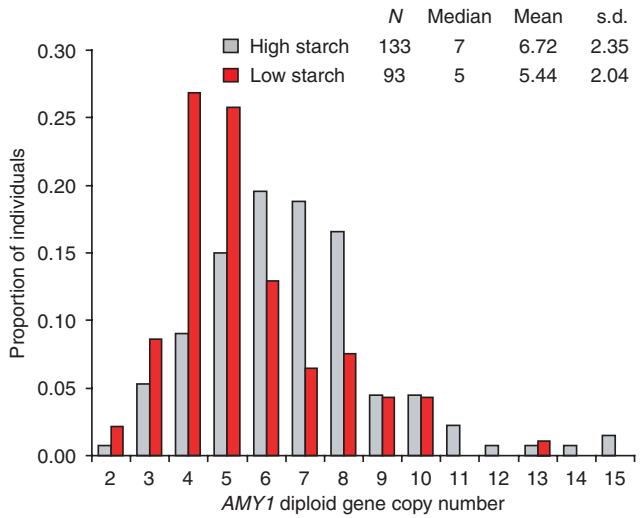
Indels



- **In**sertions or **del**etions (indels)
- Each human genome has between 540k and 625k indels
- Most indels are small
- *Indels in coding regions tend to be multiples of 3bp. Why?*

CNVs

- **C**opy **N**umber **V**ariations (CNVs): when the number of copies of a gene differs from one person to the next
- Can be identified by CGH or depth of coverage (tricky!)
- Amylase copy number and diet
- Perry et al 2007 (Nature Genetics)
- refSeq genes:
  - *AMY1A, AMY1B, AMY1C, AMY2A, AMY2B*



Microsatellites

- Microsatellites are DNA sequences that contain a number of repeated 2-6bp sequences (also called short tandem repeats, STRs)

- Example:
  - AGAGAGAGAGAGAGAG
  - $(AG)_8$

- Different alleles have different numbers of repeats

- Huntingon's disease: $(CAG)_{40}$ is pathogenic

- Microsatellites have high mutation rates

- Microsatellites tend to be polymorphic (useful for DNA fingerprinting)

Folk singer Woody Guthrie

(Image from Wikipedia)

Structural variation

- Structural variation includes inversions, translocations
- Also includes large (>1kb) insertions or deletions

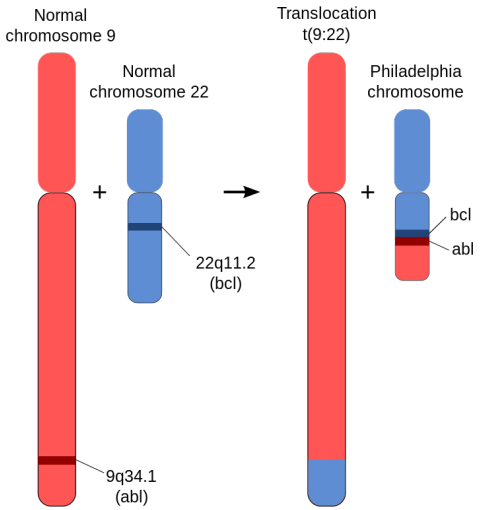
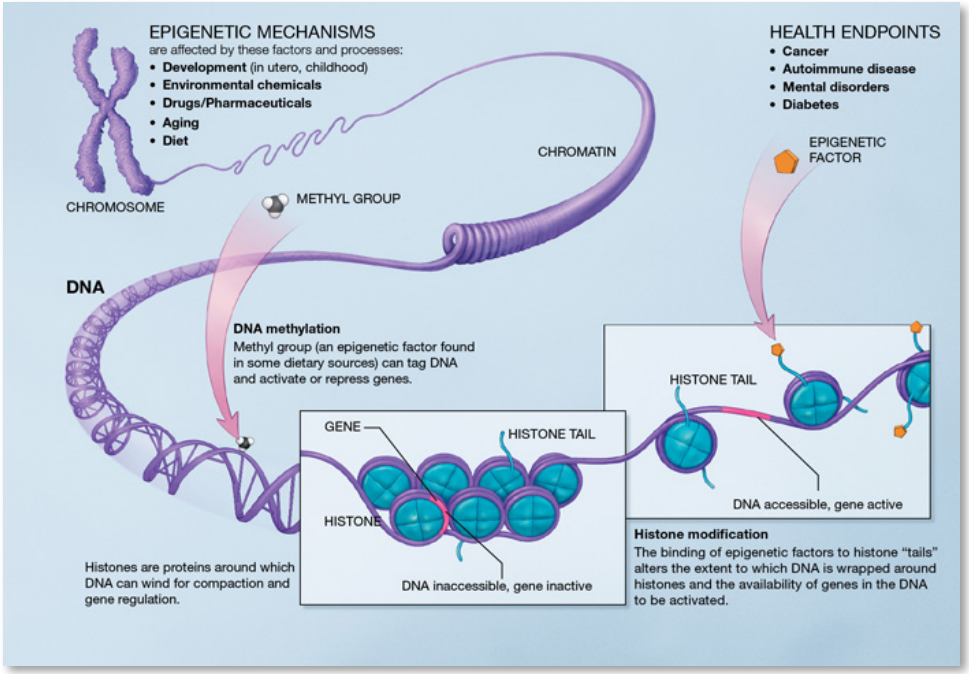


Image from Wikipedia

- Philadelphia chromosome
  - Reciprocal translocation between chromosome 9 and 22
  - Causes chronic myelogenous leukemia (CML)

Epigenetic variation

- DNA methylation (methylated CpGs)
- Histone modification
- X-inactivation
- Genomic imprinting



(Image from Wikipedia)

- Different people have different epigenetic marks
- Almost all of these epigenetic marks are erased each generation

Genotyping technologies

Sanger sequencing

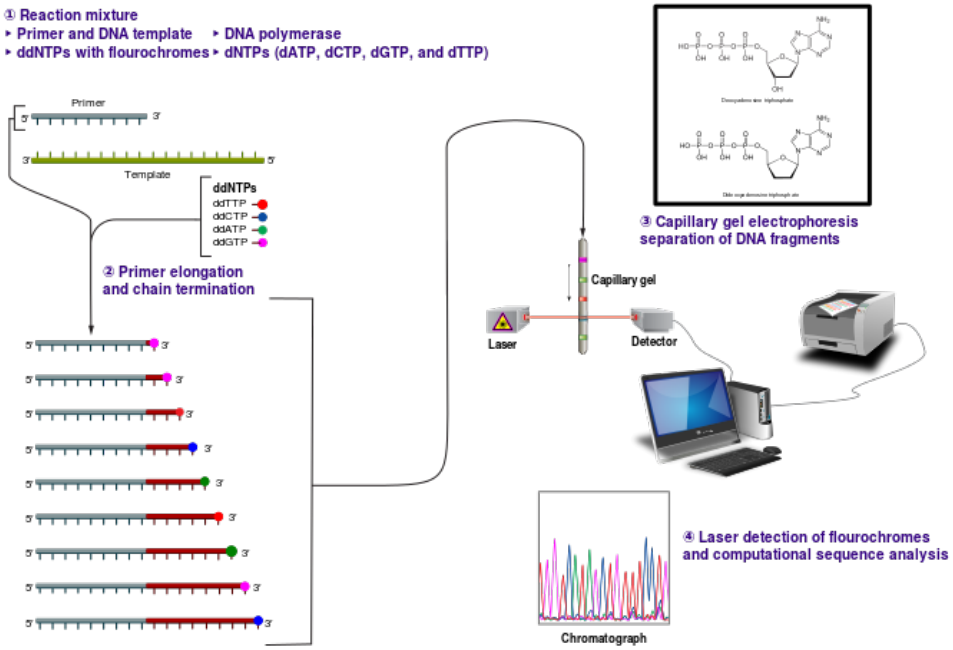


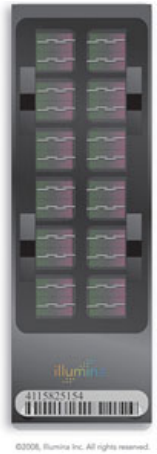
Image rights: Wikimedia Commons

The Engineered Biosystems Building has a window motif that resembles a radioactively labeled sequencing gel

- Developed in 1977. Despite being a gold-standard, it is **not** high-throughput!
- Yields ~700bp reads (targeted sequencing)
- Uses a single-stranded DNA template, DNA primer, DNA polymerase, normal dNTPs, and labeled ddNTPs which terminate DNA strand elongation

SNP genotyping arrays: overview

- Microarrays contain collections of DNA spots attached to a surface\
- Can contain probes for over 1M different SNPs
- Limitation: unable to detect novel variants
- Previously ascertained SNPs can lead to biased results
- Relatively inexpensive
- One error per 10,000 SNPs
- Useful for GWAS (SNPs on arrays tag genomic regions)



Whole genome sequencing (WGS): overview

- WGS is sometimes called next-generation sequencing
- Depth of coverage: average number of reads per base pair in a genome (low coverage = 5-10X, high coverage: >30X)
- One error per 100,000 base pairs (high coverage)
- Relatively expensive
- Allows you to discover new variants
- Neutral intergenic variants can be used to infer demographic history

Whole genome sequencing: how it works

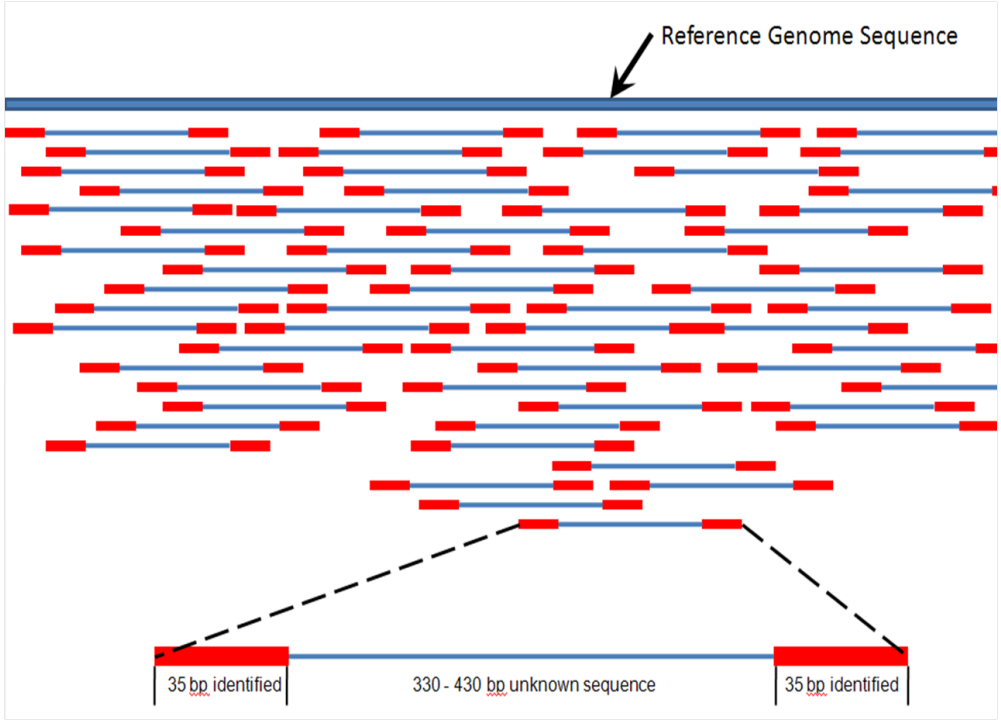


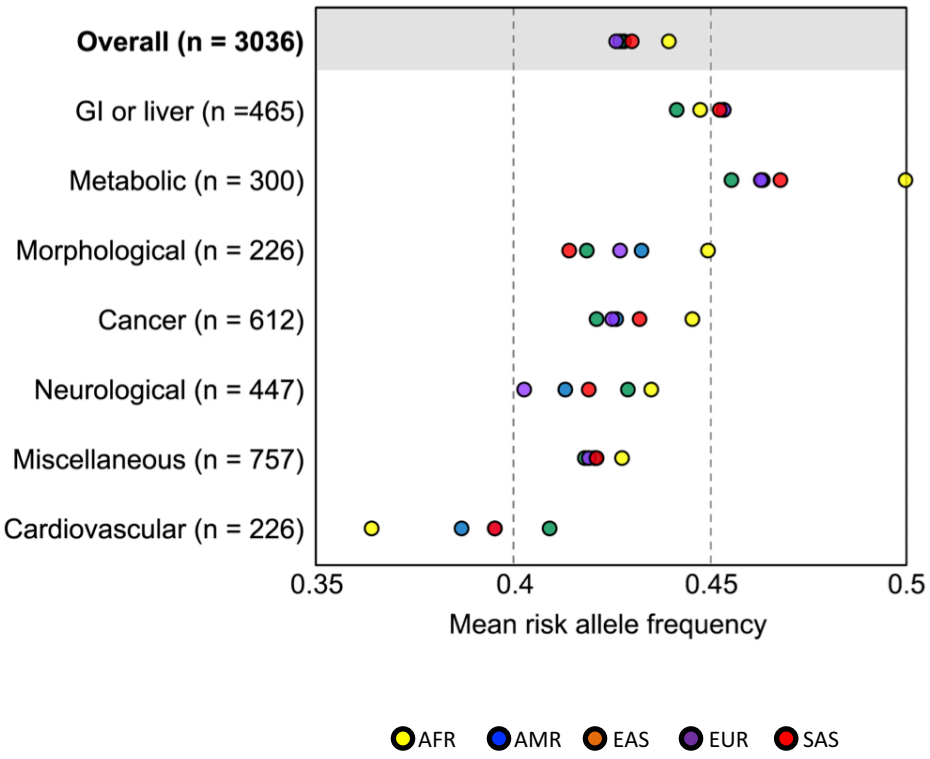
Image rights:
Wikimedia Commons

- DNA broken up into small fragments
- Paired-end reads generated (~35bp fragments with spacers)
- Reads mapped to the human reference genome and SNPs are called
- Approximately 5% of the human genome is unmappable repetitive DNA

'omics: an overused suffix?

- Genomics: the study of all the entire set of genes in a cell
- Transcriptomics: the study of all mRNA molecules in a cell
- Proteomics: the study of all protein molecules in a cell
- Metabolomics: the study of all metabolites in a cell
- Epigenomics: the study of the entire set of epigenetic modifications
- Microbiomics: the study of the microorganisms that share our body space
- Connectomics: the study of connections in an organism's nervous system

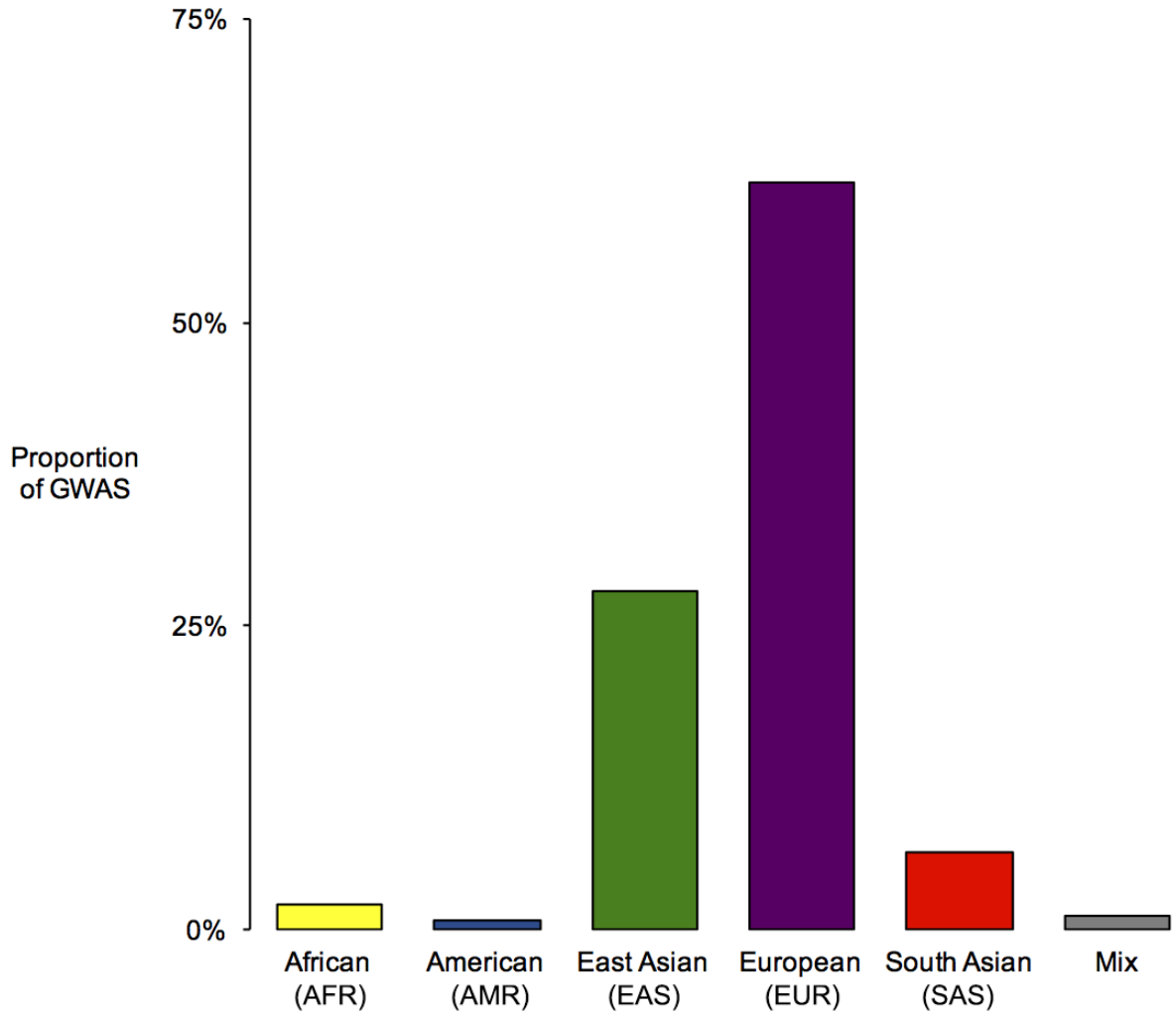
SNP ascertainment bias



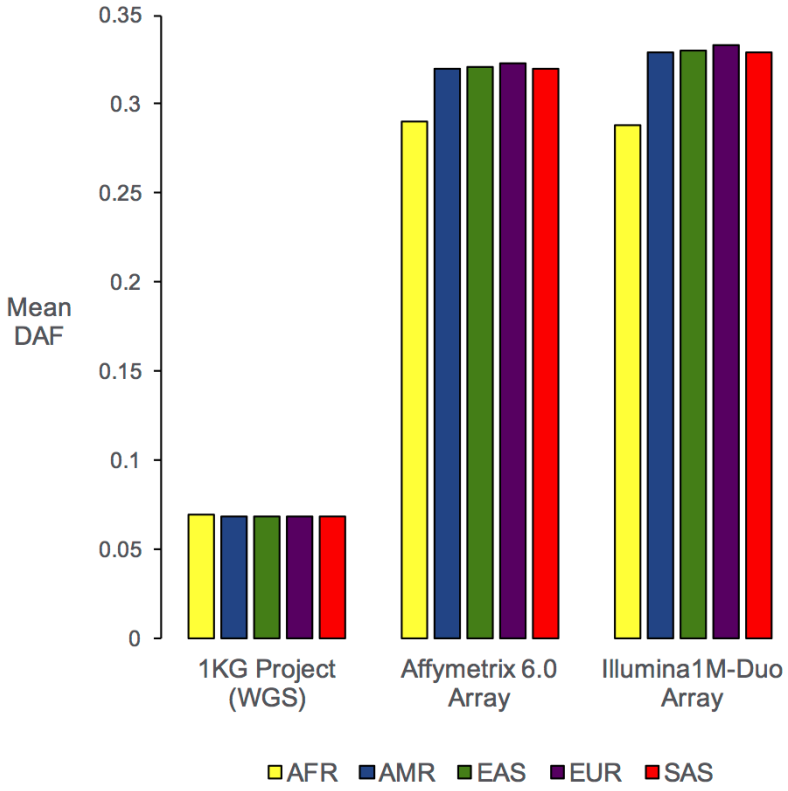
Michelle Kim
Georgia Institute of Technology

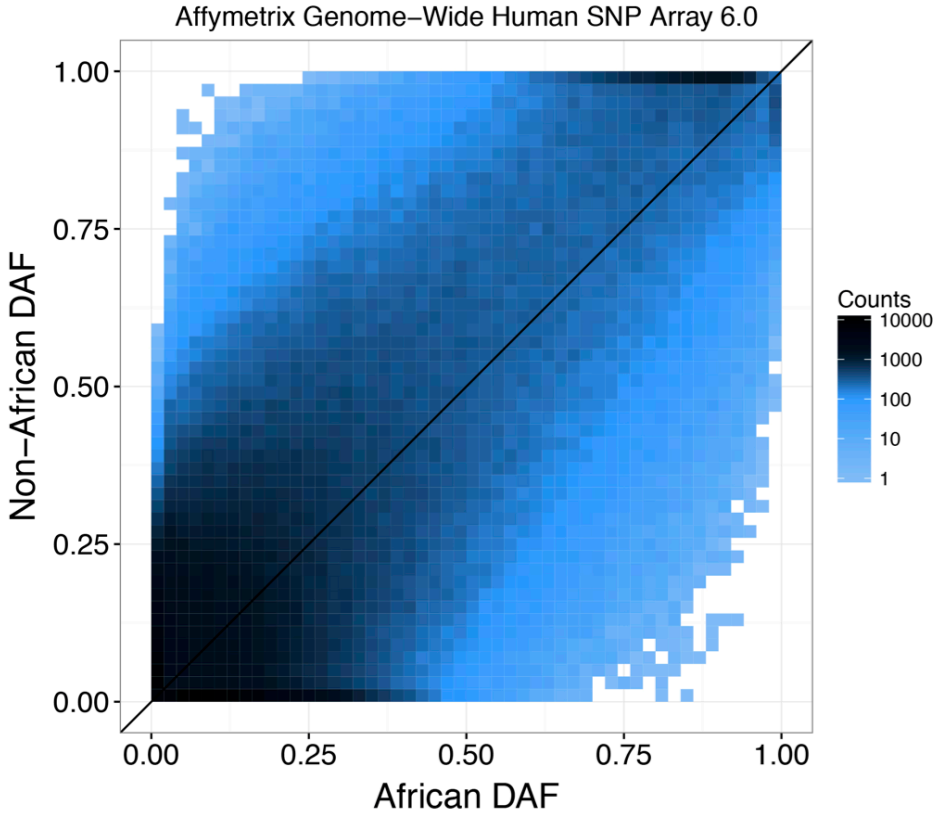
Kane Patel
Georgia Institute of Technology

Most GWAS use European or Asian samples



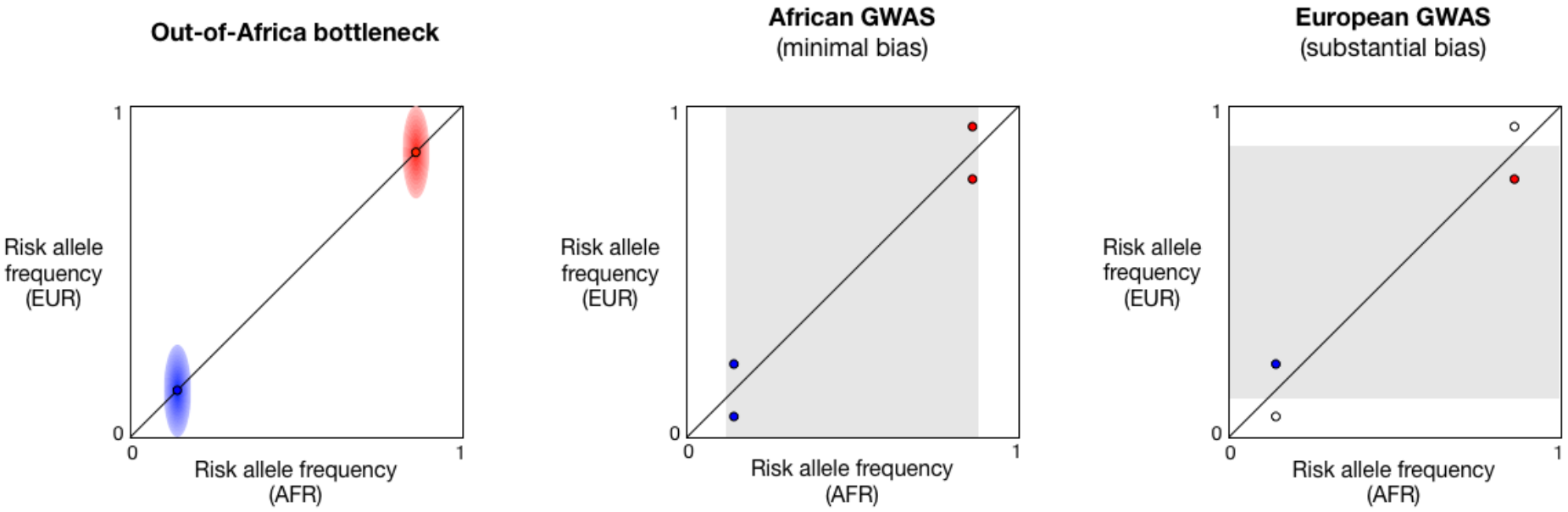
Arrays have biased allele frequencies





DAF: derived allele frequency

Bottlenecks cause some disease SNPs to be missed in European GWAS



Rumsfeldian science?

"There are **known knowns**. These are things we know that we know. There are **known unknowns**. That is to say, there are things that we know we don't know. But there are also **unknown unknowns**. There are things we don't know we don't know."

- Donald Rumsfeld, 2002