SISG 2016

Introduction to Genetics and Genomics

2. Molecular Biology of the Genome

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

Outline

- Information flow
- Molecular biology
- Connections
- Variation
- Technology

The Double Helix XX-XY
Sculpture by:
Franco Castellucio

To what extent does structure imply function?

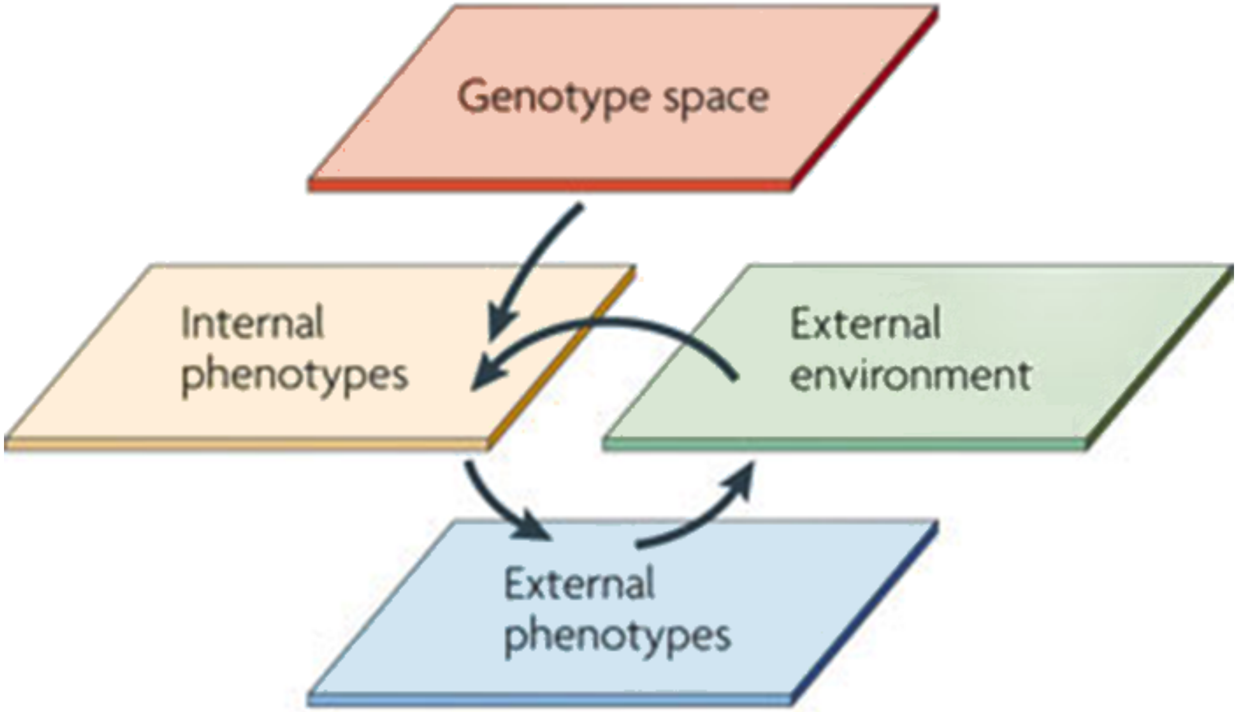
Terminology

- **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

Information flow

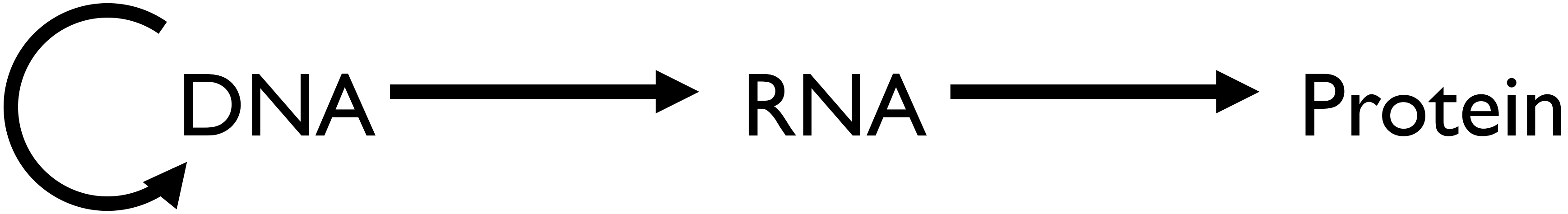
Image rights: Ramona Saldamando

Genotype-phenotype map



Nature Reviews | Genetics

Central Dogma of Molecular Biology*



DNA

RNA

Protein

*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

Image rights: Wikimedia Commons

Molecular biology

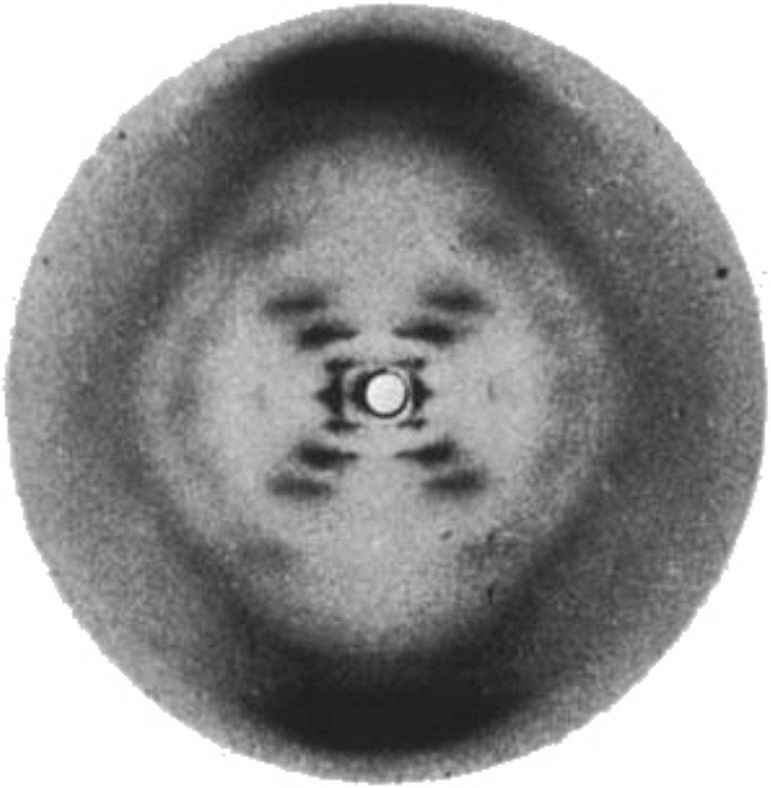


Photo 51: X-ray diffraction of DNA
(Gosling and Franklin)

DNA

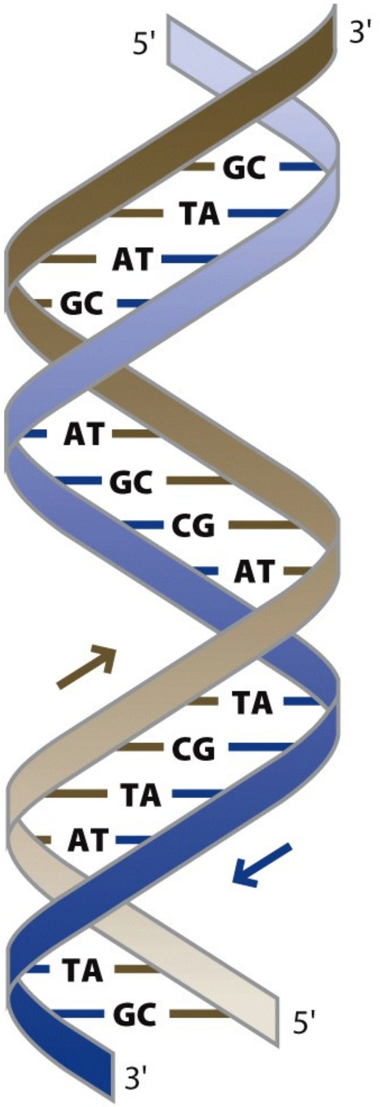


Figure 2.4b Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

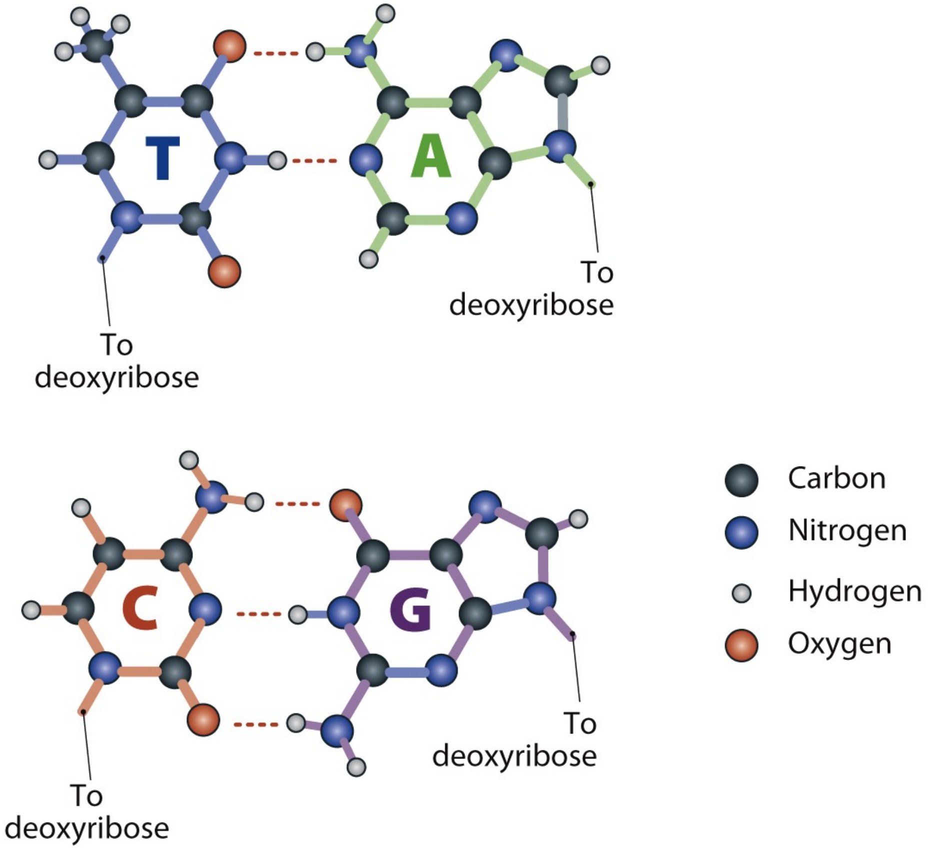


Figure 2.5 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

DNA packaging

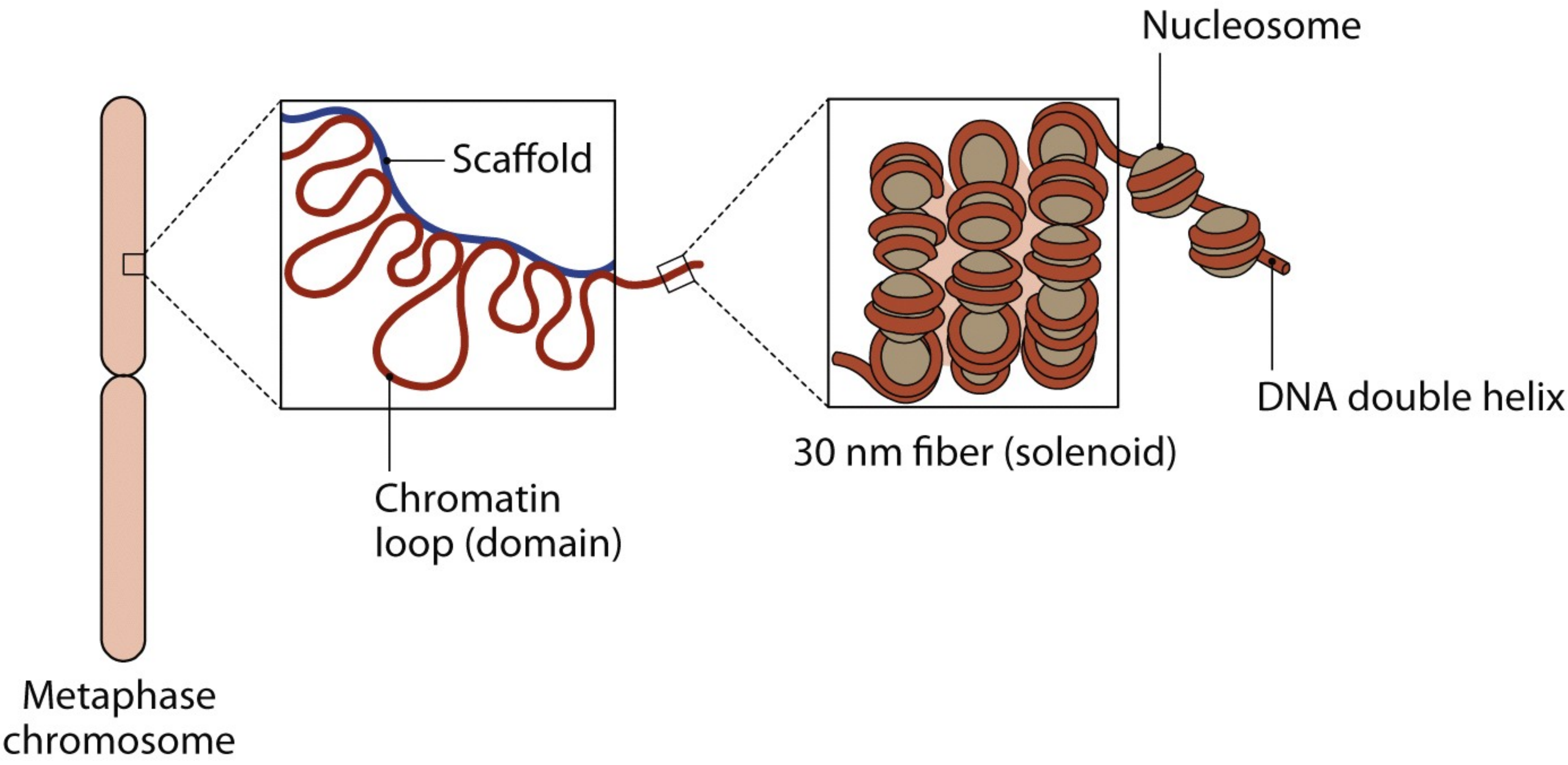
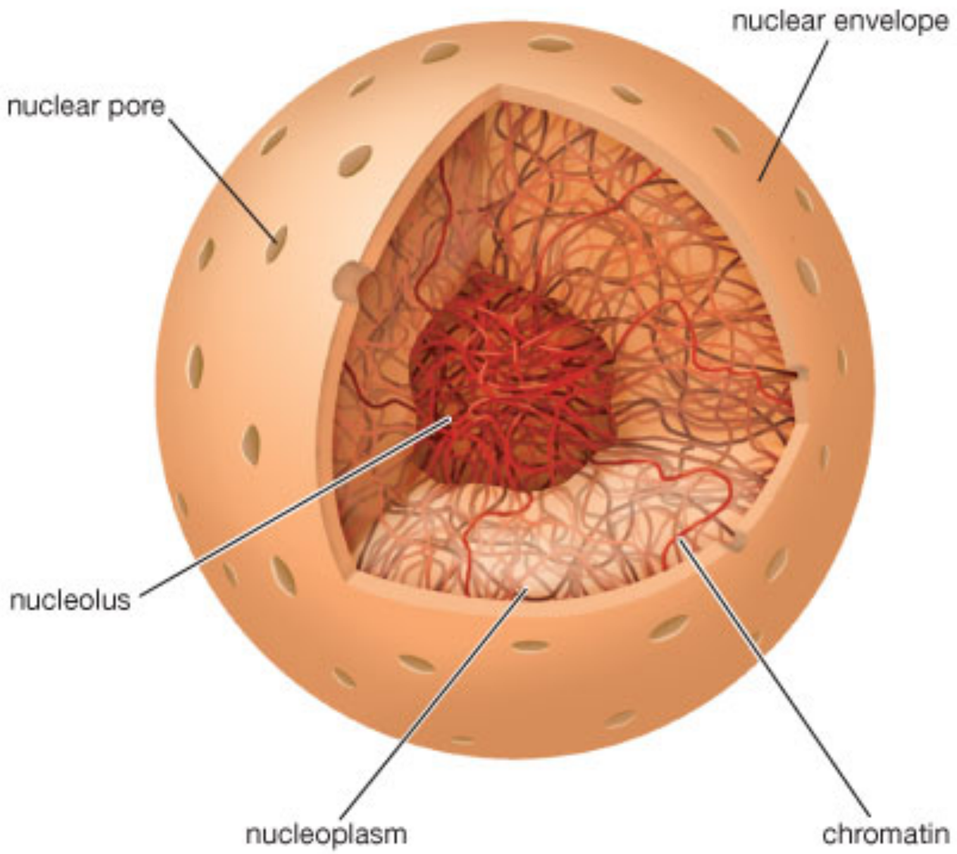


Figure 2.11 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

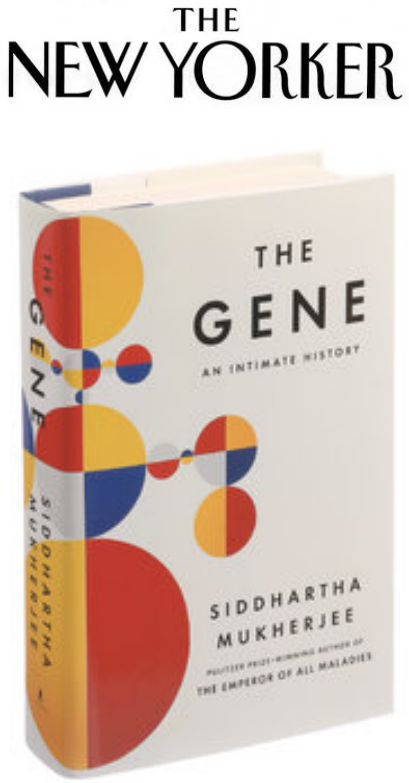
DNA packaging (movie clip)

Chromatin


nuclear envelope
nuclear pore
nucleolus
nucleoplasm
chromatin

DNA packaging: implications

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



Ploidy

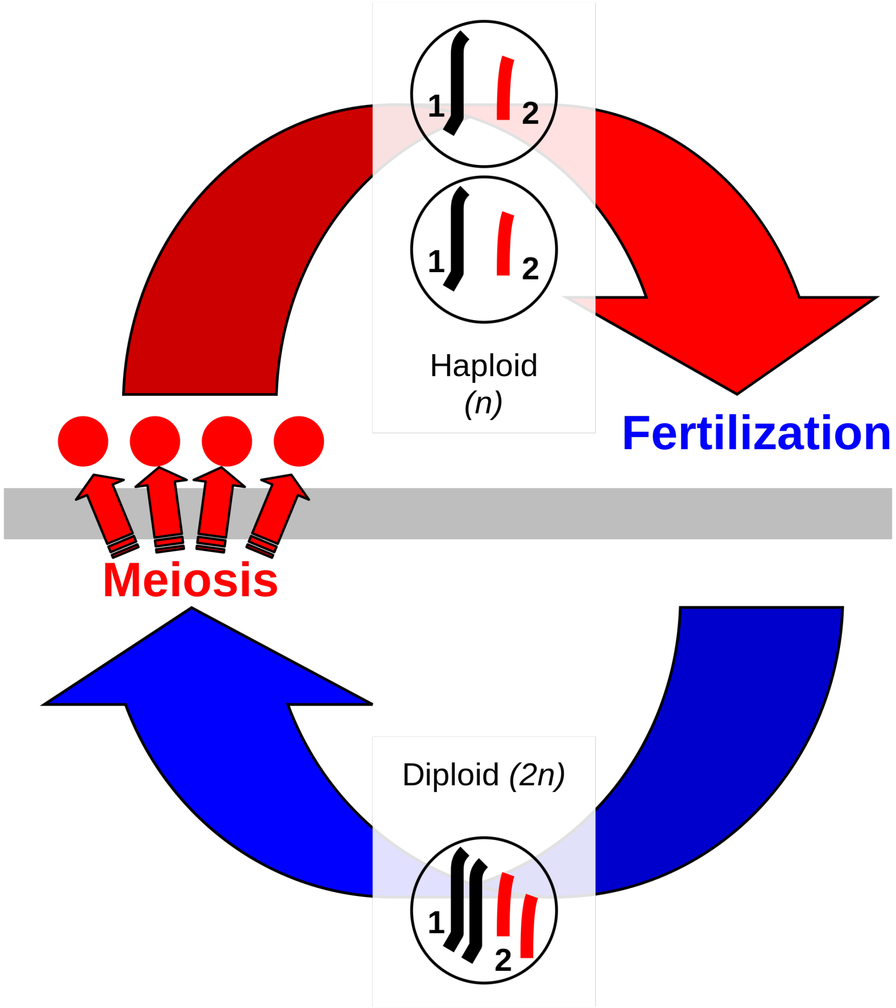
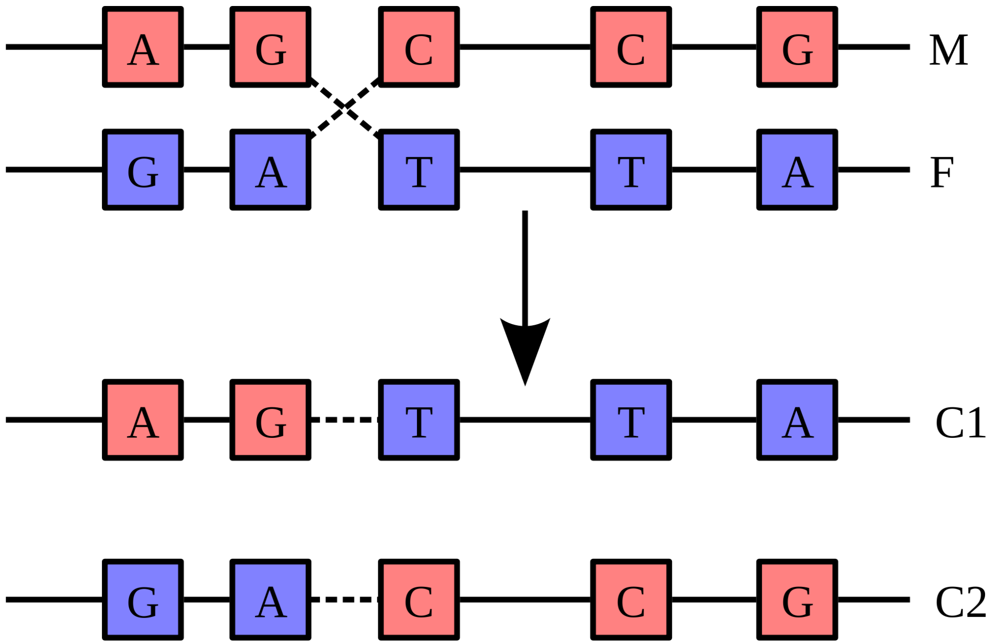


Image rights: Wikimedia Commons

Recombination

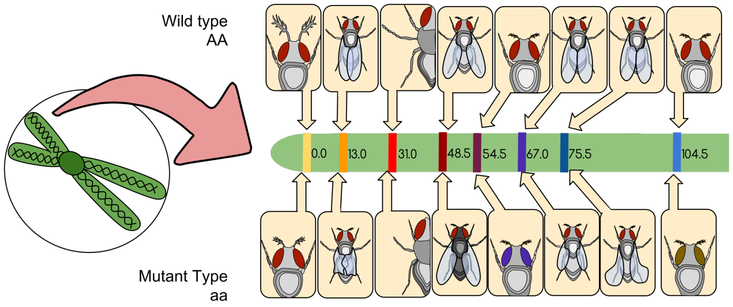


- Occurs in meiosis
- Byproduct of the need to pair homologous chromosomes

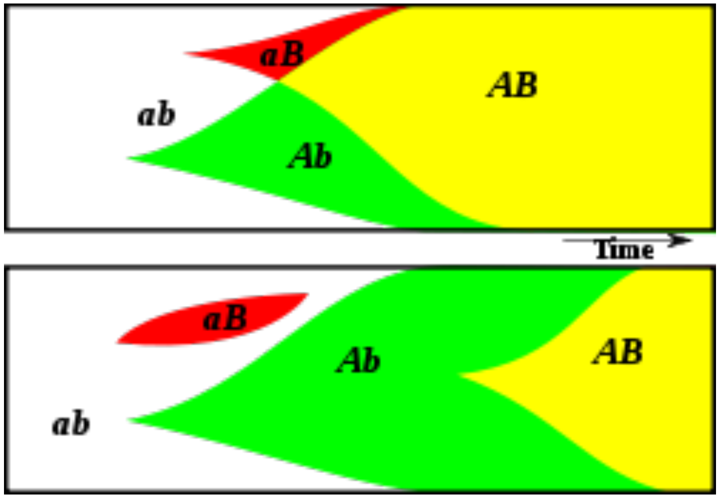
Image rights: Wikimedia Commons

Recombination: implications

- Genetic maps and linkage disequilibrium



- Benefits of sex



Sexual reproduction
(recombination)

Asexual reproduction

Image rights: Wikimedia Commons

DNA replication

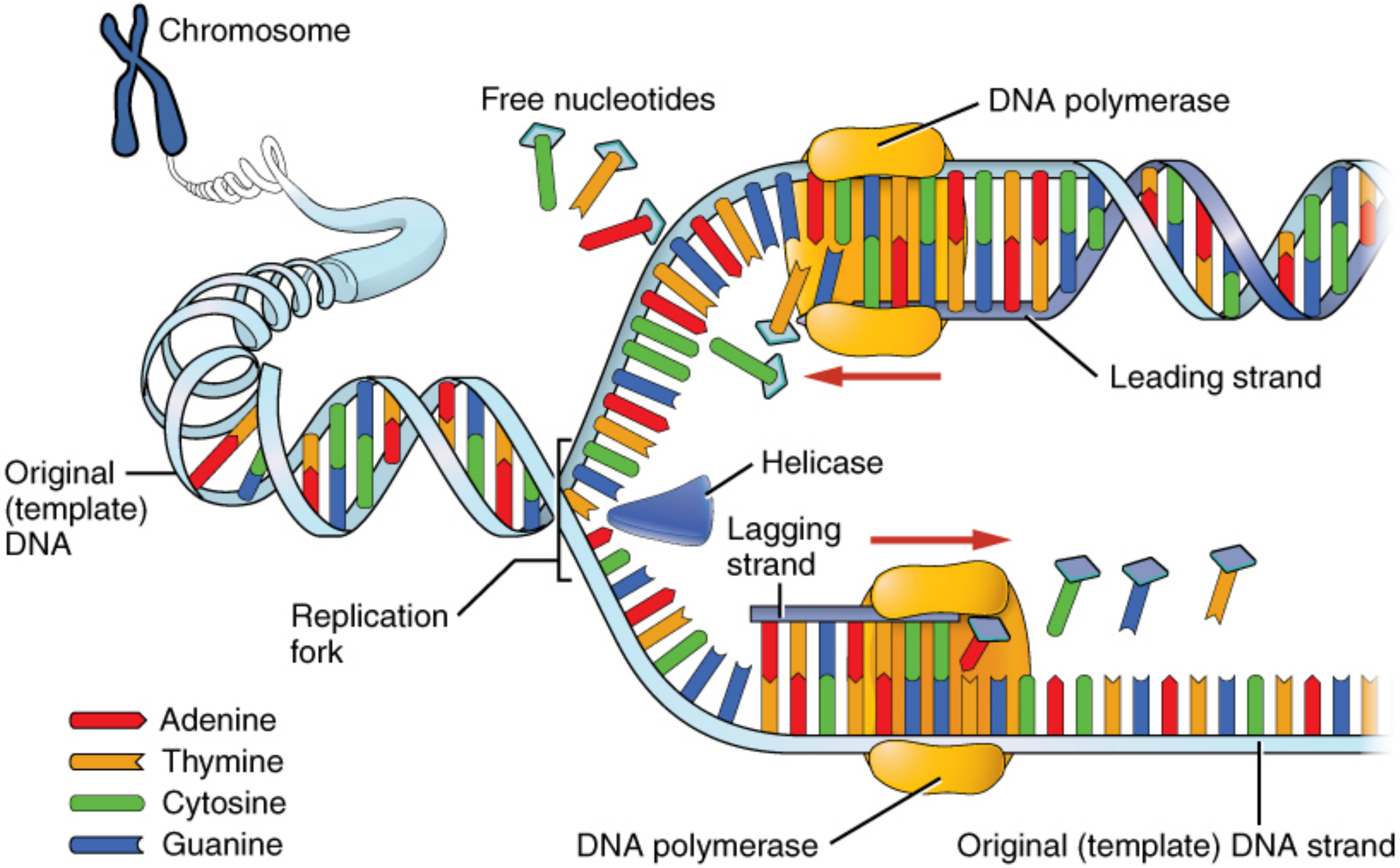
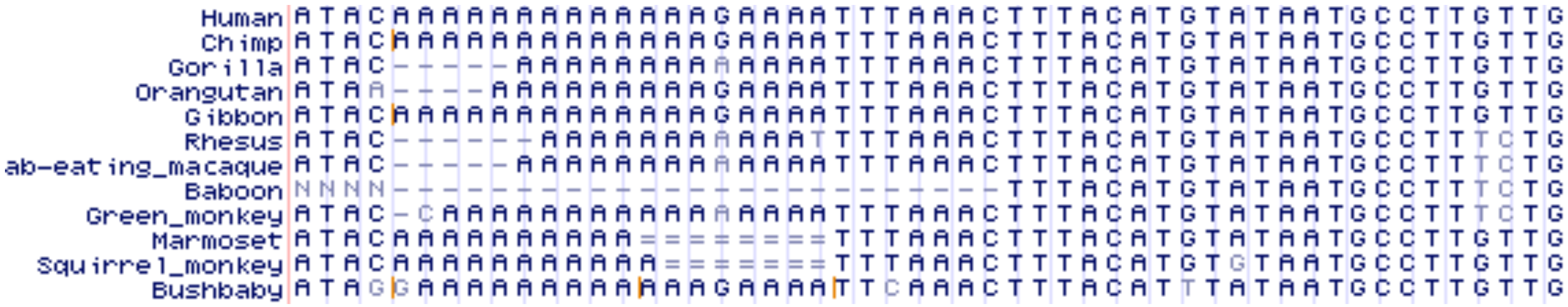


Image rights: Wikimedia Commons

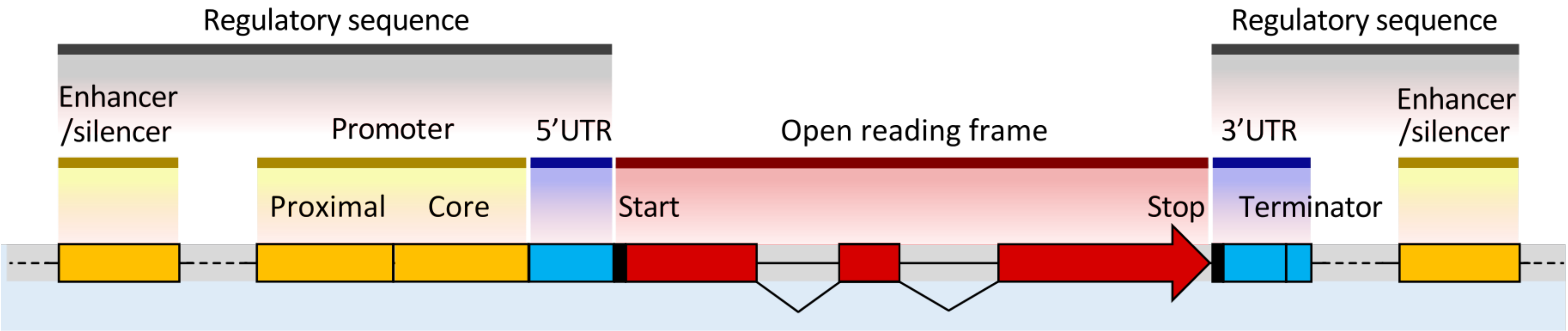
DNA replication: implications

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



```
            Human ATACAAAAAAAAAAAAAGAAAATTTAAACTTTACATGTATAATGCCTTGTTG
            Chimp ATACAAAAAAAAAAAAAGAAAATTTAAACTTTACATGTATAATGCCTTGTTG
          Gorilla ATAC-----AAAAAAAAAAAAATTTAAACTTTACATGTATAATGCCTTGTTG
        Orangutan ATAA----AAAAAAAAAGAAAATTTAAACTTTACATGTATAATGCCTTGTTG
           Gibbon ATACAAAAAAAAAAAAAGAAAATTTAAACTTTACATGTATAATGCCTTGTTG
           Rhesus ATAC------AAAAAAAAAAATTTTAAACTTTACATGTATAATGCCTTTCTG
ab-eating_macaque ATAC-----AAAAAAAAAAAAATTTAAACTTTACATGTATAATGCCTTTCTG
           Baboon NNNN-------------------------TTTACATGTATAATGCCTTTCTG
     Green_monkey ATAC-CAAAAAAAAAAAAAAAATTTAAACTTTACATGTATAATGCCTTTCTG
         Marmoset ATACAAAAAAAAAA========TTTAAACTTTACATGTATAATGCCTTGTTG
  Squirrel_monkey ATACAAAAAAAAAAA=======TTTAAACTTTACATGTGTAATGCCTTGTTG
         Bushbaby ATAGGAAAAAAAAAAAAGAAAATTCAAACTTTACATTTATAATGCCTTGTTG
```

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**?*

Image rights: Wikimedia Commons

Transcription factors and gene regulation

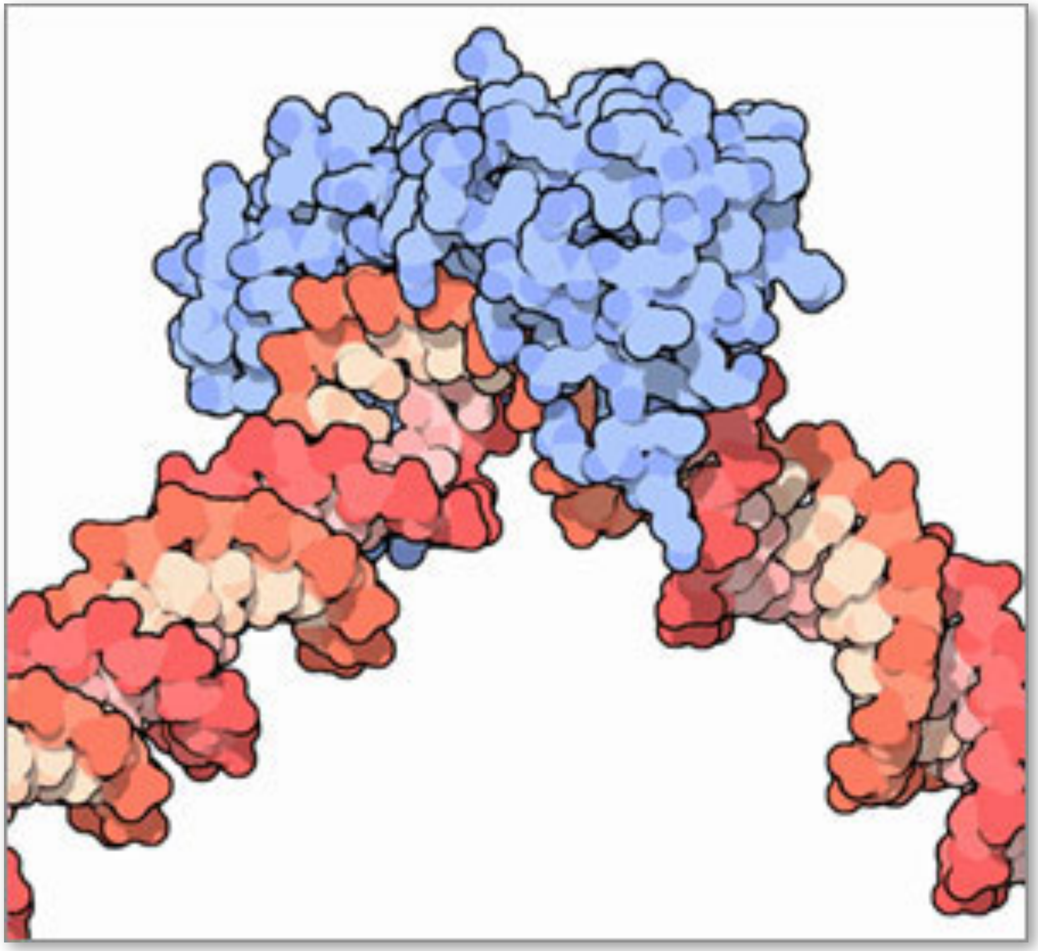
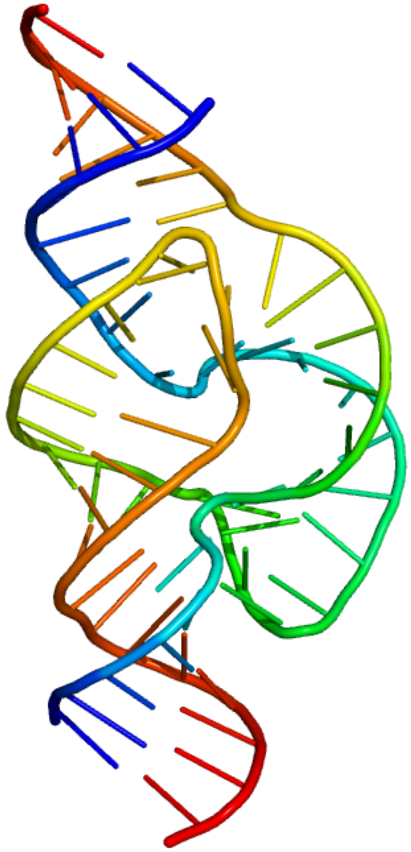


Image rights: Broad Institute abd RCSB Protein Data Bank

RNA comes in many different flavors

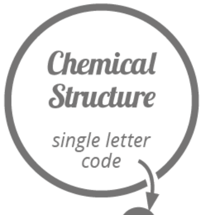
- mRNA: messenger RNA
- tRNA: transfer RNA
- rRNA: ribosomal RNA
- Regulatory RNAs (miRNA, siRNA, piRNA)



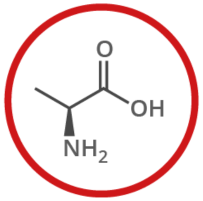
Ribozyme
Image rights: Wikimedia Commons

Proteins are made of amino acids

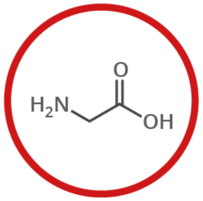
Chart Key: ● ALIPHATIC ● AROMATIC ● ACIDIC ● BASIC ● HYDROXYLIC ● SULFUR-CONTAINING ● AMIDIC ○ NON-ESSENTIAL ◌ ESSENTIAL



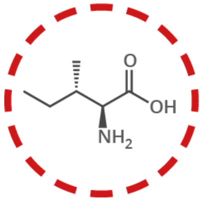
NAME A
three letter code
DNA codons



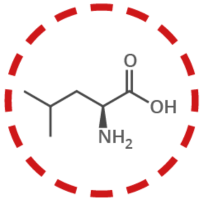
ALANINE A
Ala
GCT, GCC, GCA, GCG



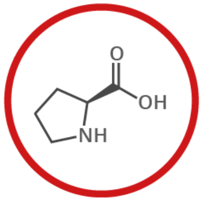
GLYCINE G
Gly
GGT, GGC, GGA, GGG



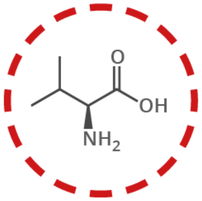
ISOLEUCINE I
Ile
ATT, ATC, ATA



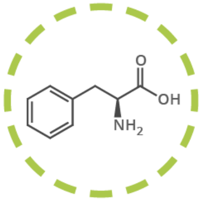
LEUCINE L
Leu
CTT, CTC, CTA, CTG, TTA, TTG



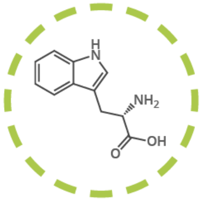
PROLINE P
Pro
CCT, CCC, CCA, CCG



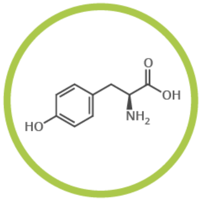
VALINE V
Val
GTT, GTC, GTA, GTG



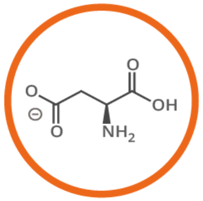
PHENYLALANINE F
Phe
TTT, TTC



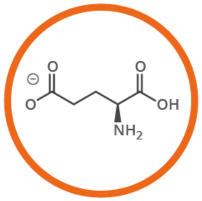
TRYPTOPHAN W
Trp
TGG



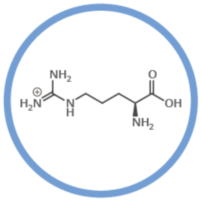
TYROSINE Y
Tyr
TAT, TAC



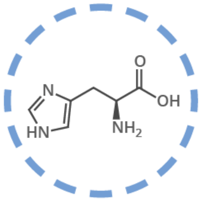
ASPARTIC ACID D
Asp
GAT, GAC



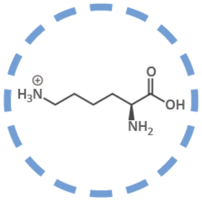
GLUTAMIC ACID E
Glu
GAA, GAG



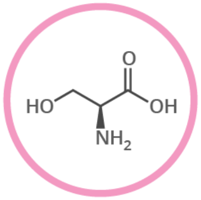
ARGININE R
Arg
CGT, CGC, CGA, CGG, AGA, AGG



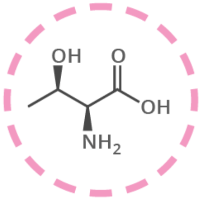
HISTIDINE H
His
CAT, CAC



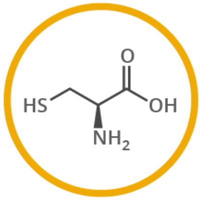
LYSINE K
Lys
AAA, AAG



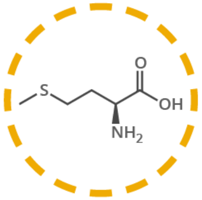
SERINE S
Ser
TCT, TCC, TCA, TCG, AGT, AGC



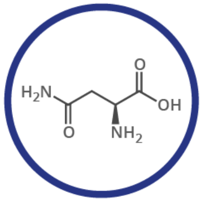
THREONINE T
Thr
ACT, ACC, ACA, ACG



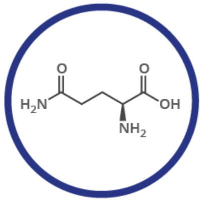
CYSTEINE C
Cys
TGT, TGC



METHIONINE M
Met
ATG



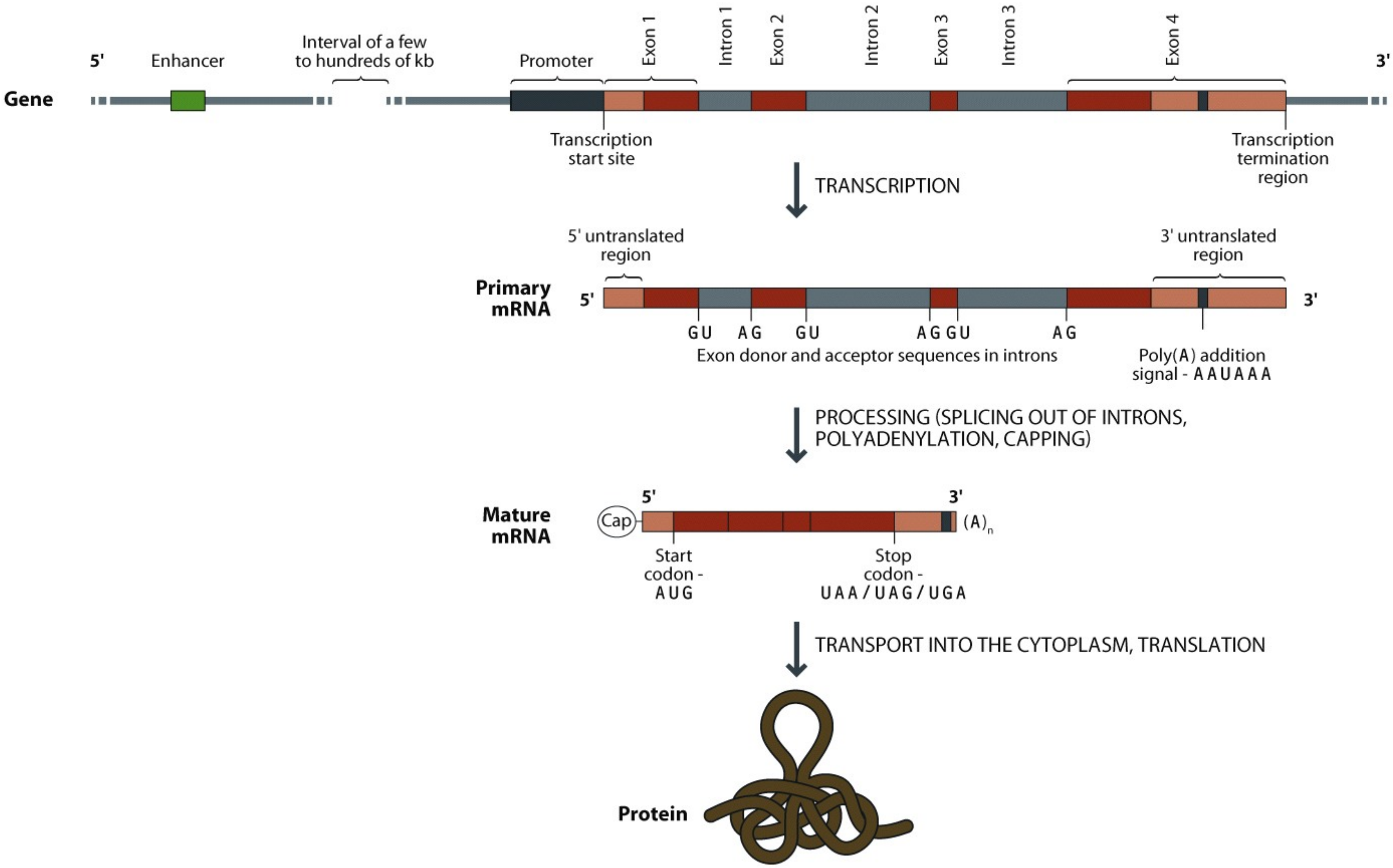
ASPARAGINE N
Asn
AAT, AAC



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



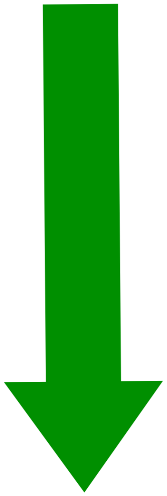
Gene
5'
Enhancer
Interval of a few to hundreds of kb
Promoter
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Intron 3
Exon 4
3'
Transcription start site
Transcription termination region
TRANSCRIPTION
5' untranslated region
3' untranslated region
Primary mRNA
5'
3'
GU
AG
GU
AG GU
AG
Exon donor and acceptor sequences in introns
Poly(A) addition signal - AAUAAA
PROCESSING (SPLICING OUT OF INTRONS, POLYADENYLATION, CAPPING)
Mature mRNA
Cap
5'
3'
$(A)_n$
Start codon - AUG
Stop codon - UAA / UAG / UGA
TRANSPORT INTO THE CYTOPLASM, TRANSLATION
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)

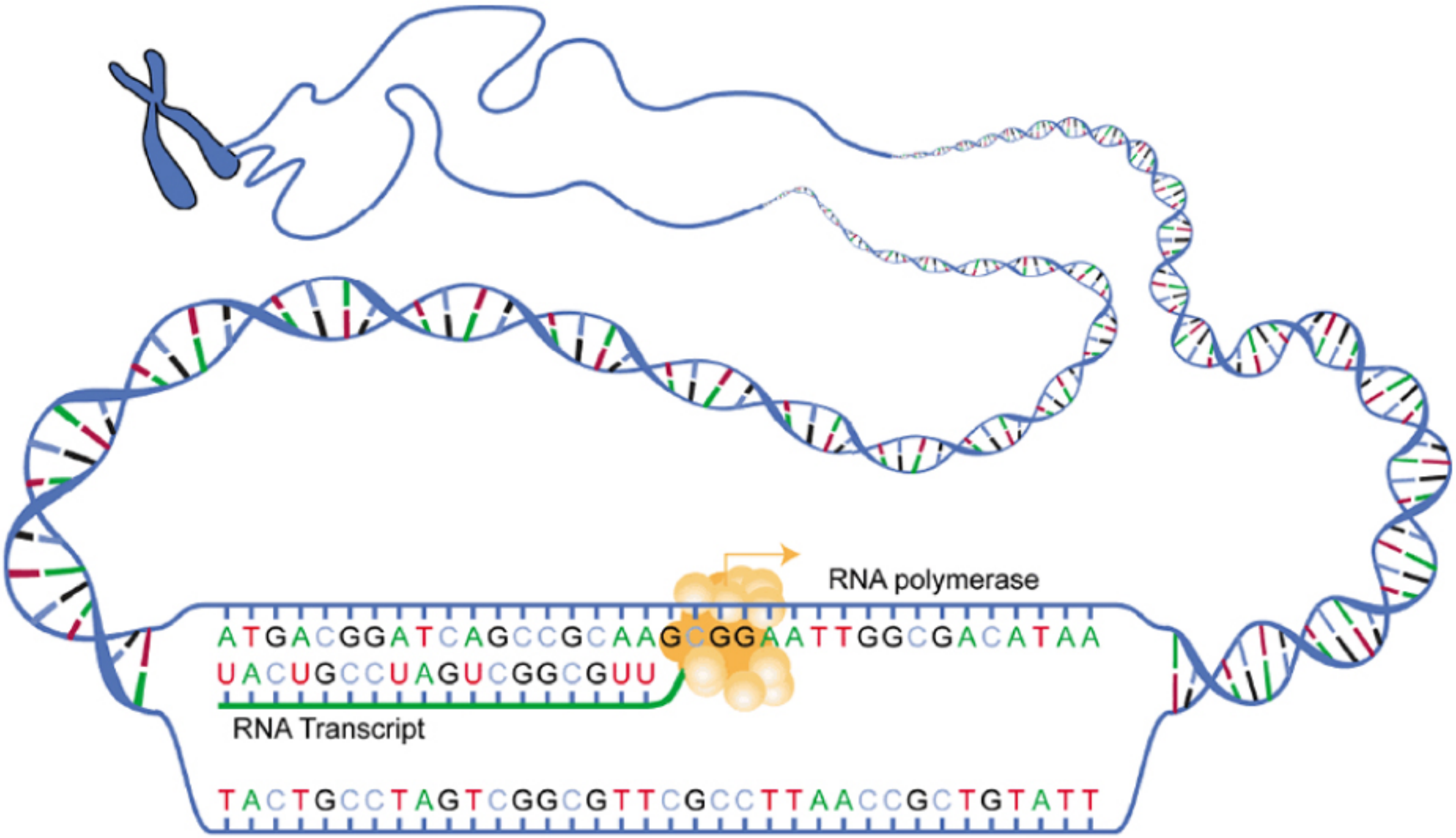

RNA polymerase
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
UACUGCCUAGUCGGCGUU
RNA Transcript
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Image rights: Wikimedia Commons

Transcription (movie clip)

Splicing

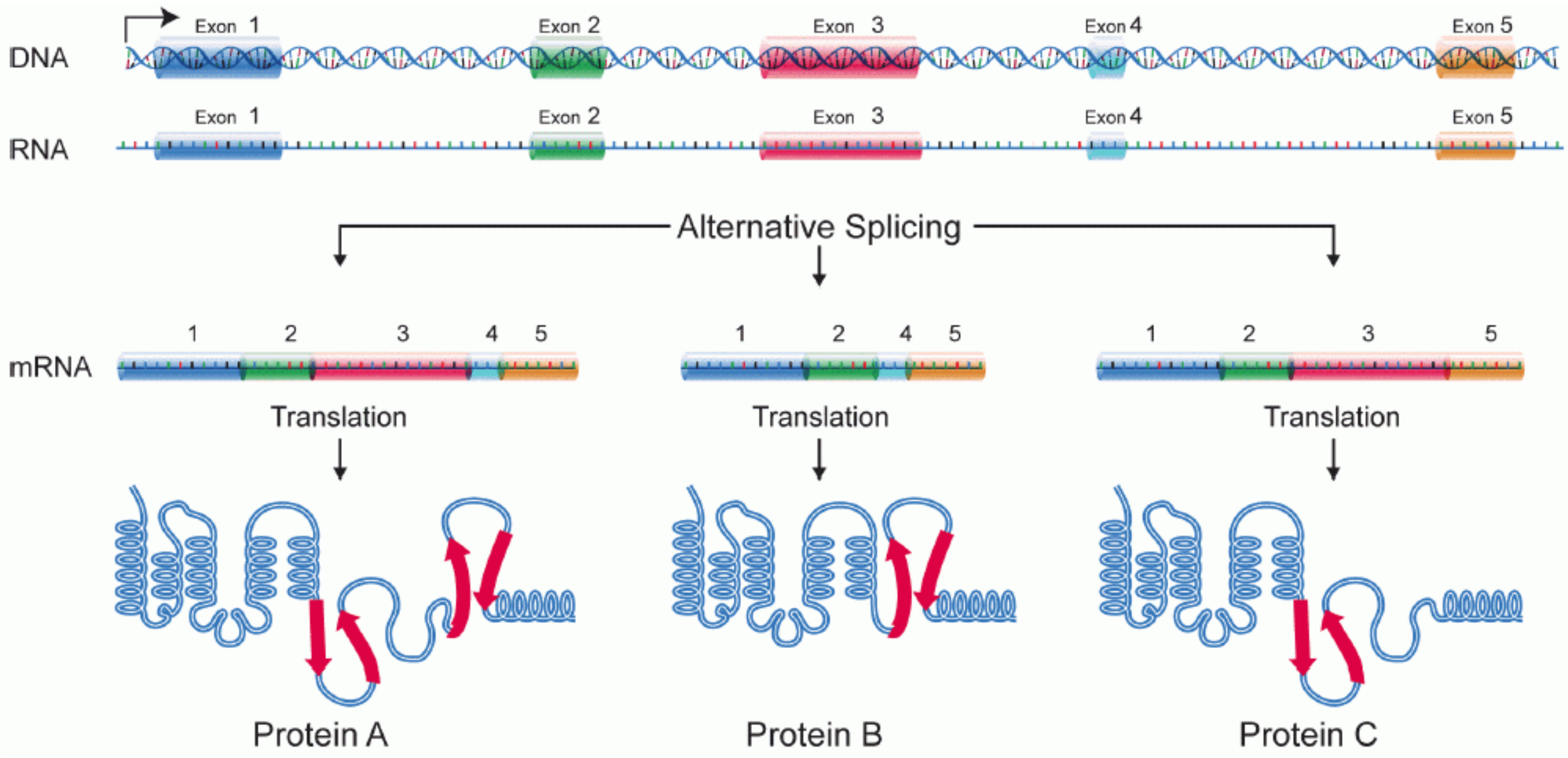


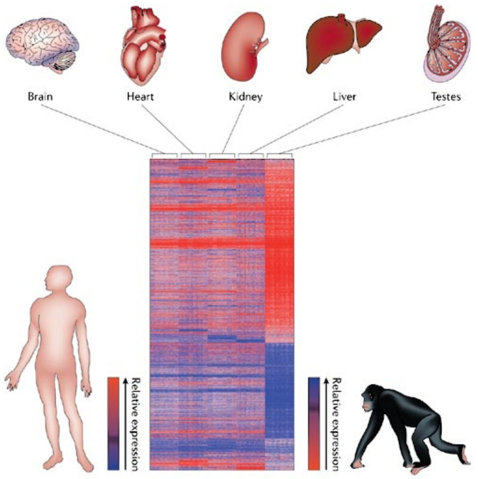
Image rights: Wikimedia Commons

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



Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

Translation (RNA to protein)

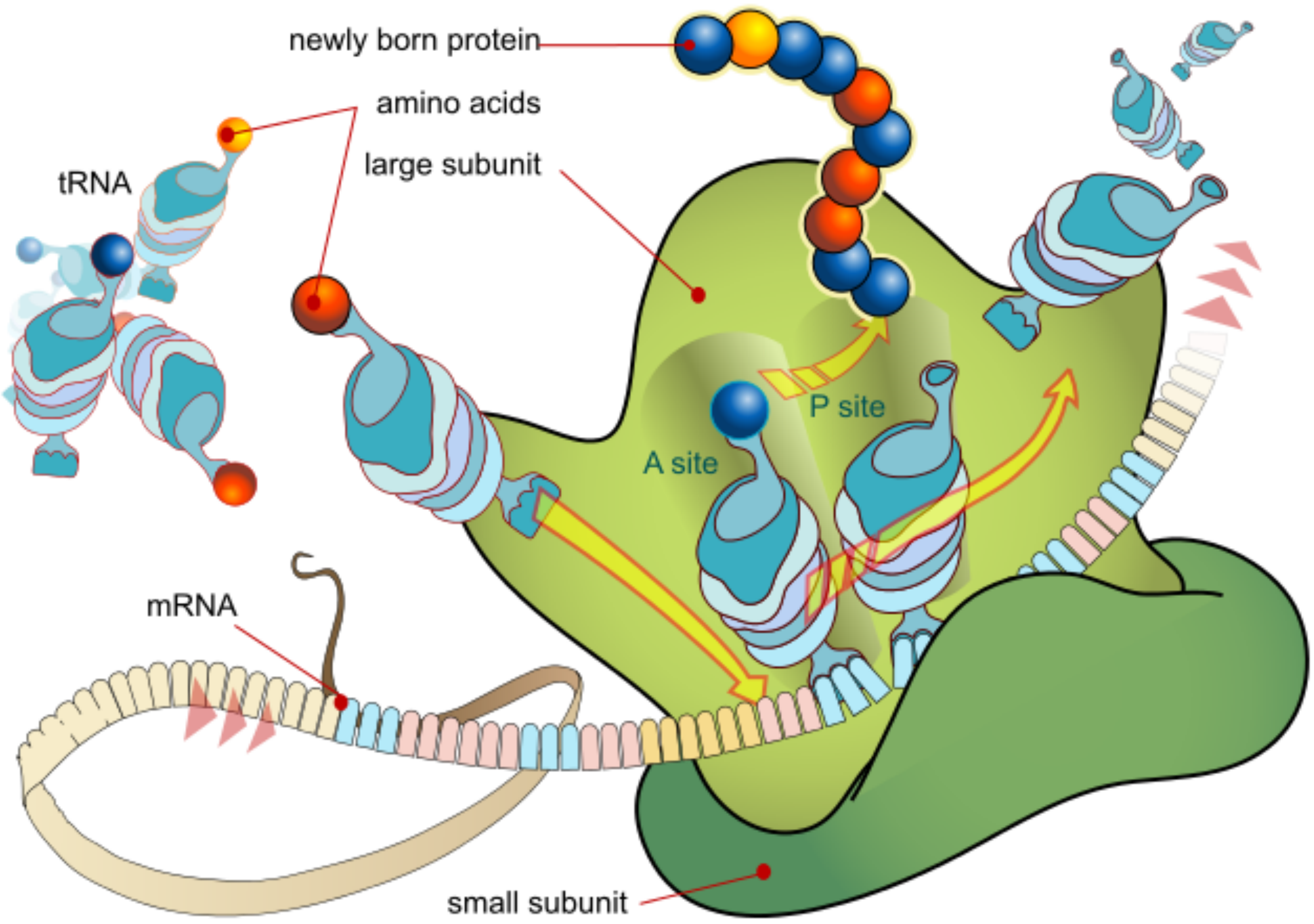
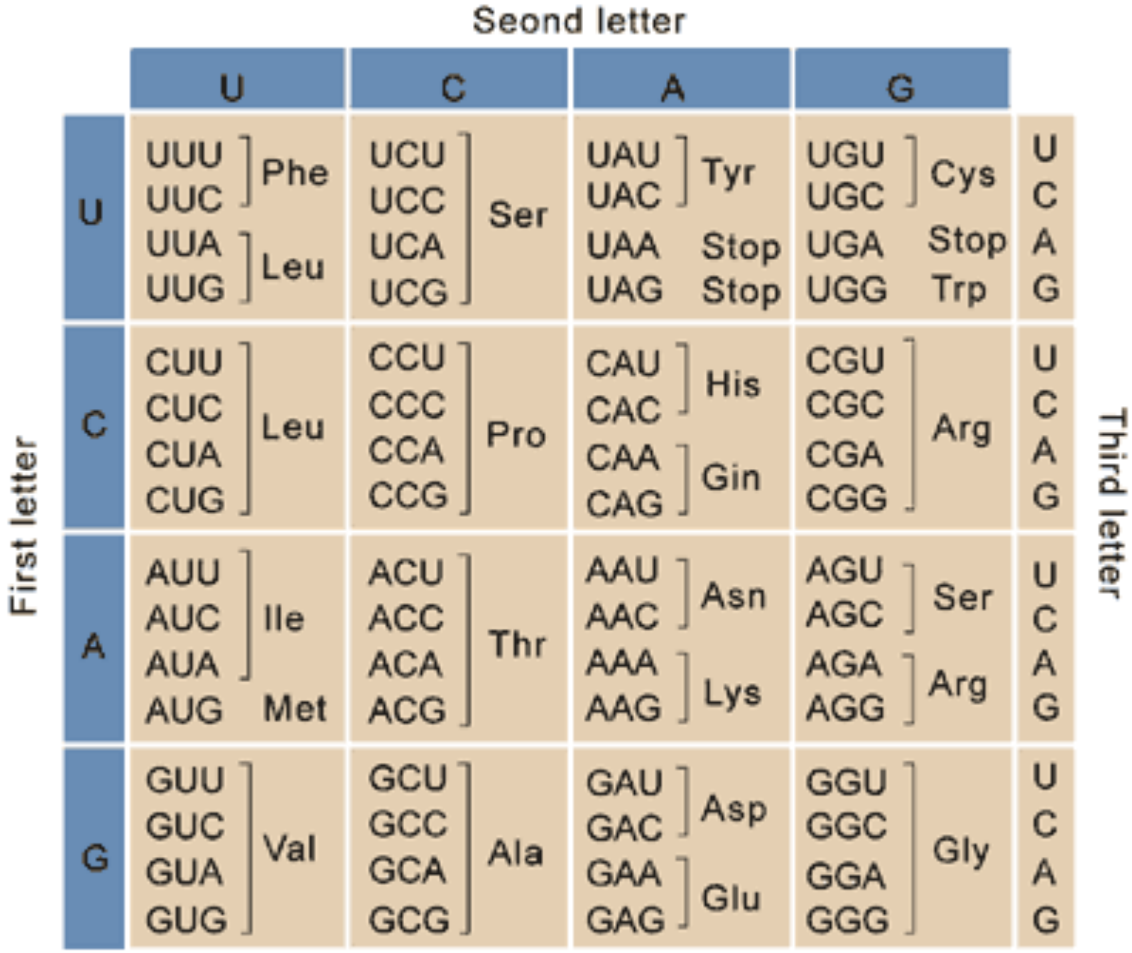


Image rights: Wikimedia Commons

Translation (movie clip)

The genetic code



	Seond letter								
First letter	U		C		A		G		Third 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	Gin	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

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

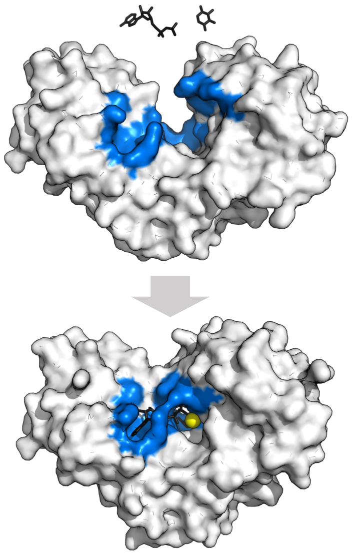
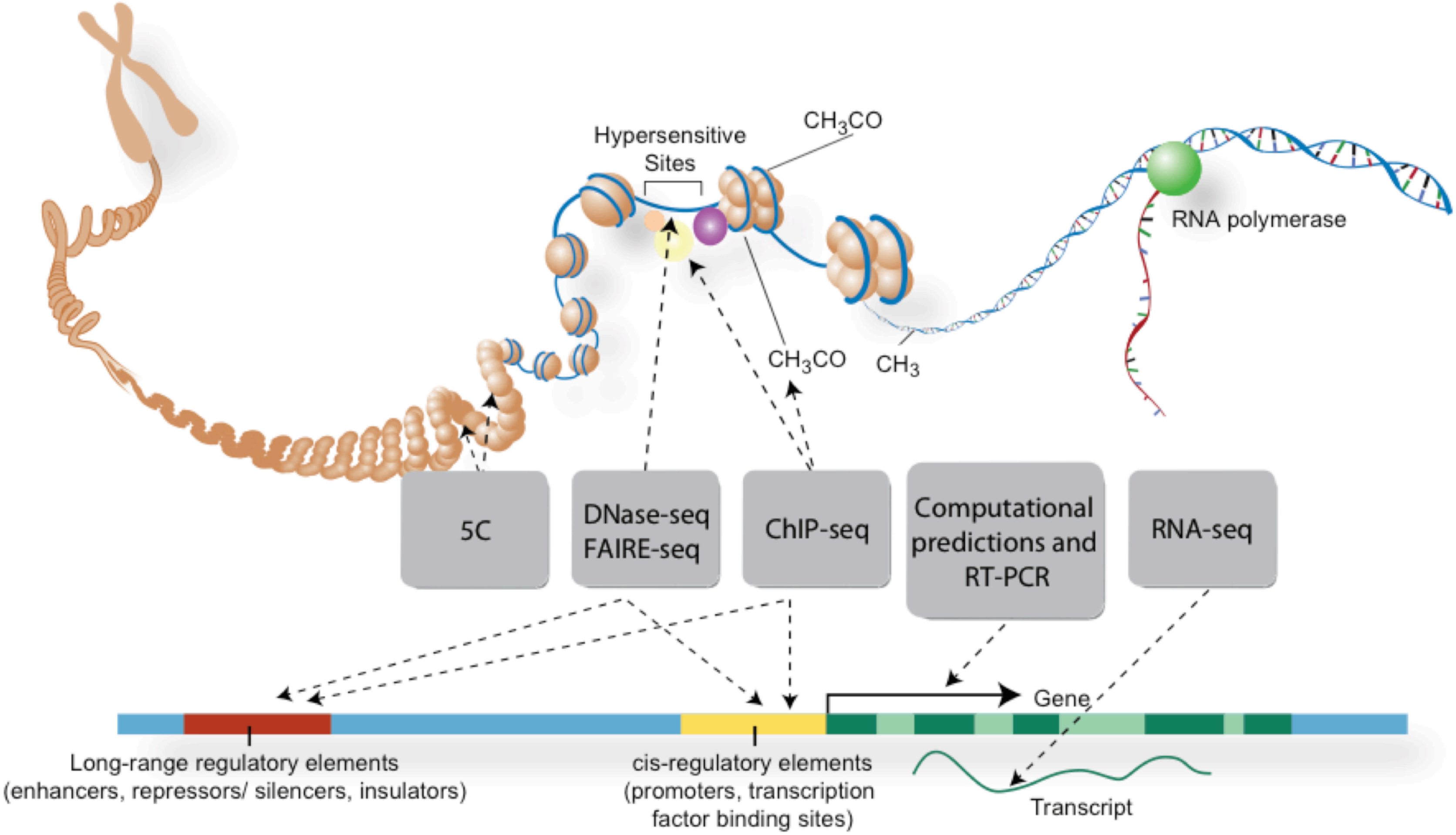


Image rights: Wikimedia Commons

ENCODE and the debate about "function"



How would you differeniate functional and nonfunctional DNA?

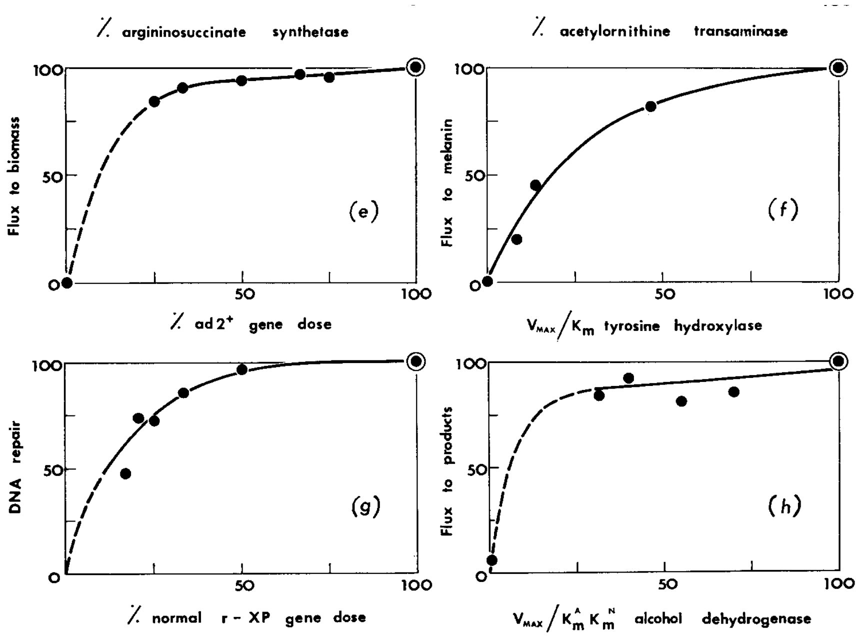
	Prokaryotes	Eukaryotes
Internal structures	No organelles	Organelles
DNA	No histones Circular No introns DNA in cytoplasm	Histones Linear Introns DNA in nucleus
Genome size	Most <5Mb	10Mb-100,000Mb
Chromatin	No histones	Histones
Ploidy	Haploid	Usually diploid
Reproduction	Asexual (binary fission)	Asexual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

Connections between molecular and classical genetics

Dominance and recessivity

- Kacser and Burns 1981 (*Genetics*)

- Dominance can arise as an emergent property of metabolic flux



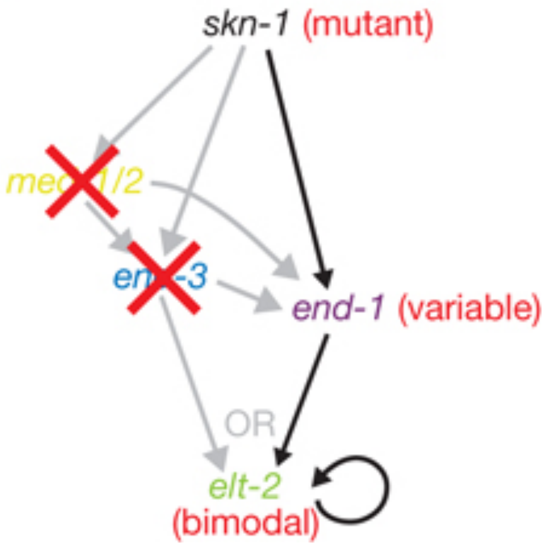
- Haldane’s Sieve: mutations that reach fixation tend to be dominant

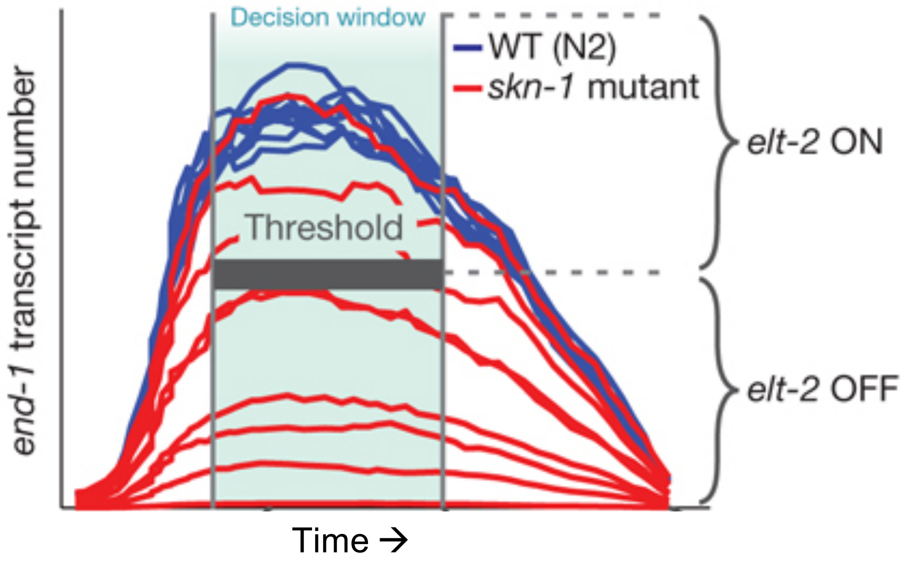
Pleiotropy

- 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

Image rights: telegraph.co.uk

Incomplete penetrance





- **Penetrance**: proportion of individuals with a given genotype that show the expected phenotype

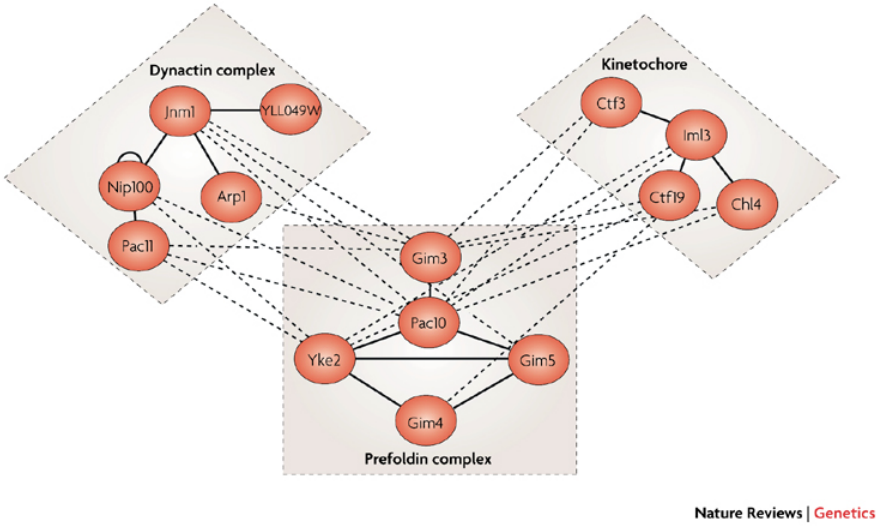
- Raj et al 2010 (*Nature*)

- Variability in gene expression + threshold → incomplete penetrance

Epistasis (genetic interactions)

- Epistasis can arise from physical interactions

- Think of transcription factors and cis-regulatory elements…

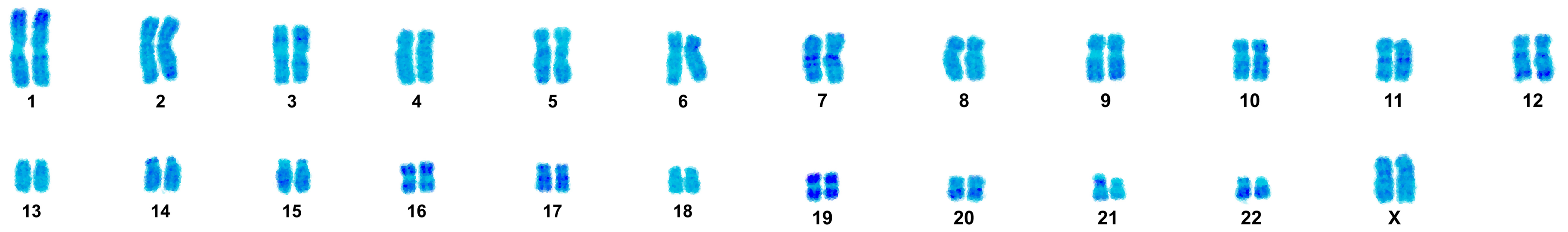


- Fitness interaction networks vs. physical interaction networks: not the same!
  - Beyer et al 2007 (*NRG*)

Variation

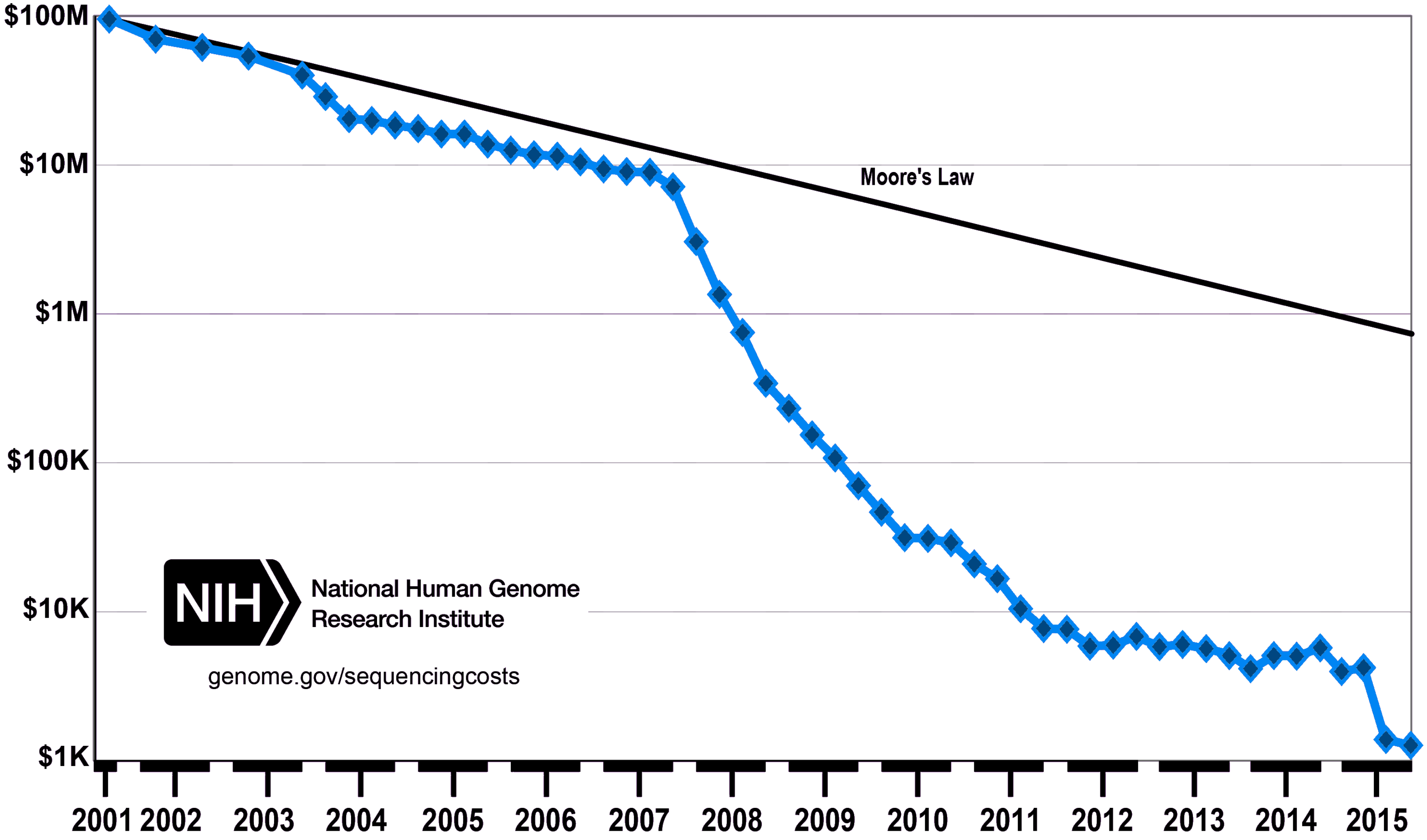
National Geographic

The human genome



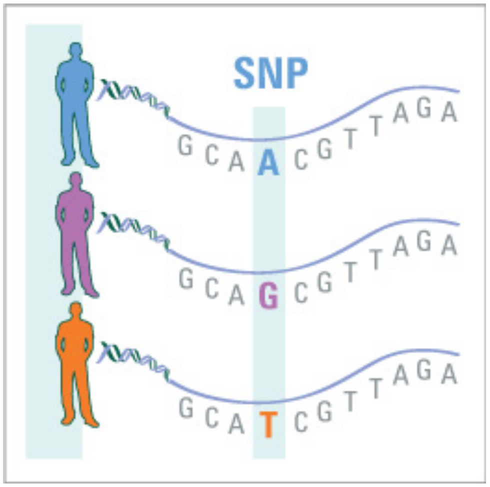
- Approximately 3.2 billion base pairs
- 23 pairs of chromosomes
  - 22 autosomes
  - One pair of sex chromosomes (XX or XY)
  - mtDNA (16.6kb)
- A typical genome
  - Heterozygous at 1 out of every 1000 sites
  - 44% transposable elements!!
  - 1.1% coding DNA

Declining sequencing costs

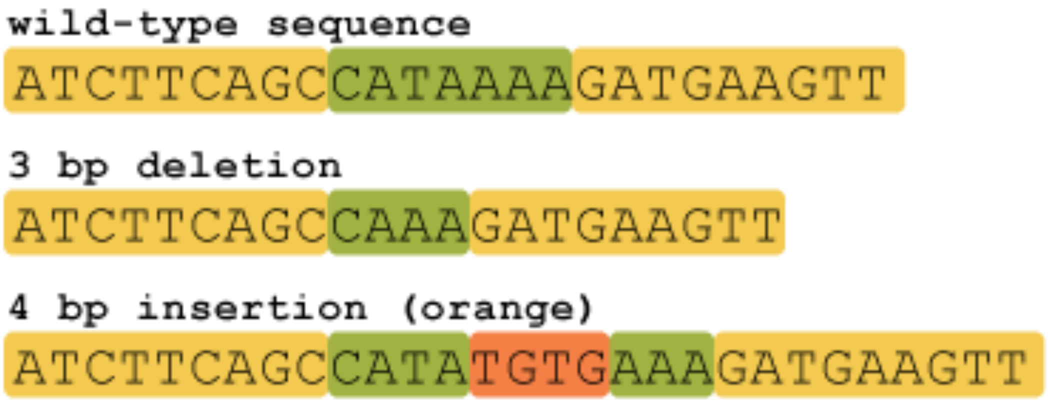


SNPs

- **<u>S</u>**ingle **<u>N</u>**ucleotide **<u>P</u>**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



wild-type sequence

ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (orange)

ATCTTCAGCCATATGTGAAAGATGAAGTT

- **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

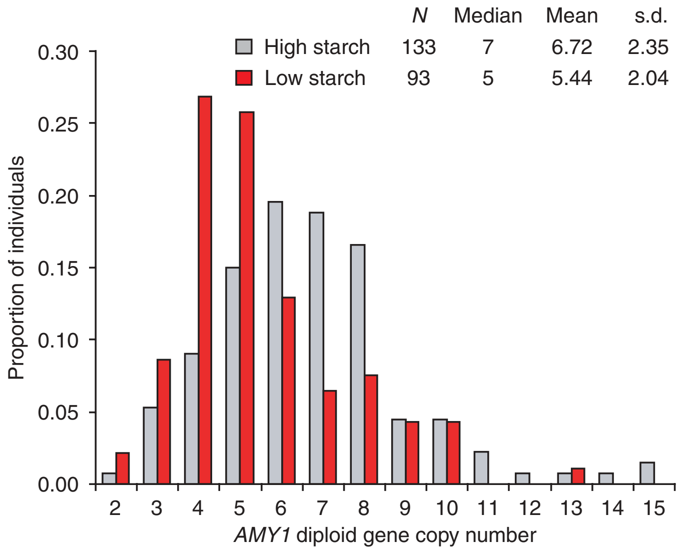
- **C**opy **N**umber **V**ariations (CNVs): when the number of copies of a gene differs from one person to the next

- 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

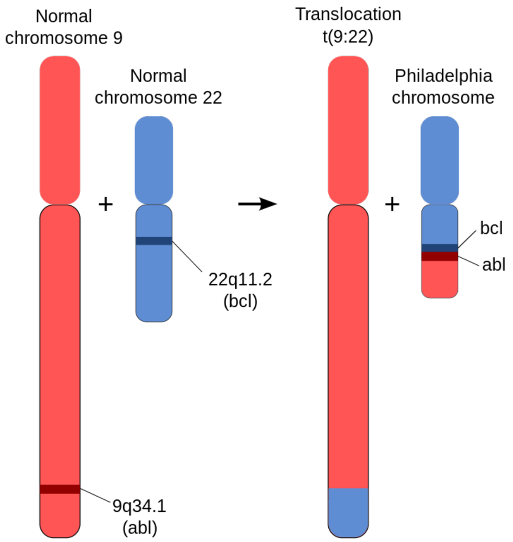
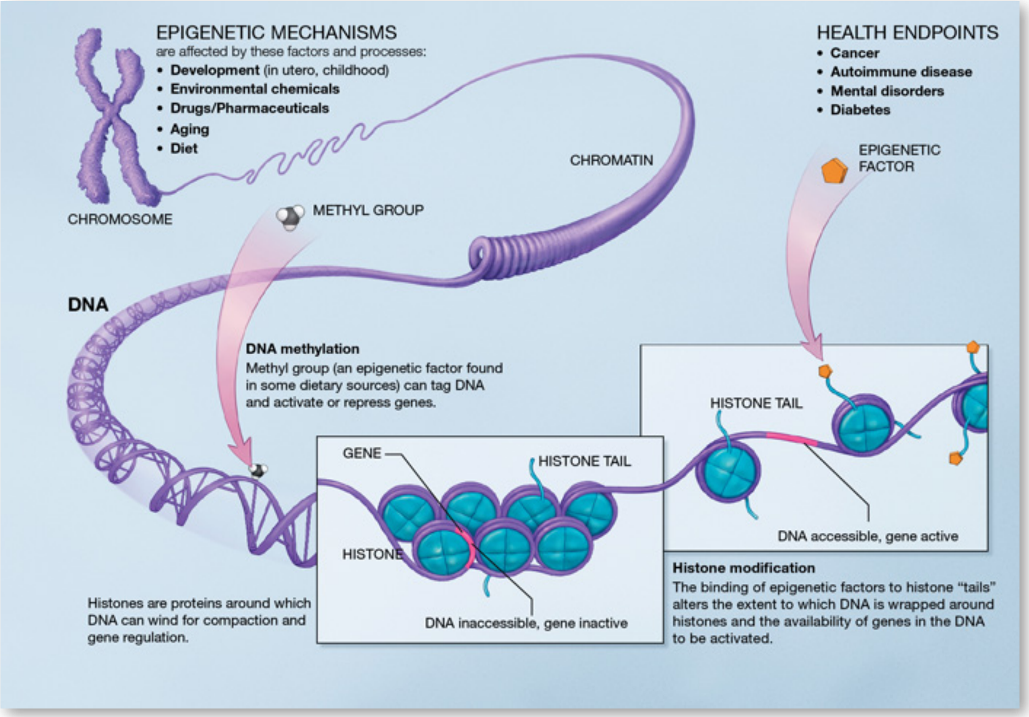


Image from Wikipedia

- 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



EPIGENETIC MECHANISMS
are affected by these factors and processes:

- **Development** (in utero, childhood)
- **Environmental chemicals**
- **Drugs/Pharmaceuticals**
- **Aging**
- **Diet**

HEALTH ENDPOINTS

- **Cancer**
- **Autoimmune disease**
- **Mental disorders**
- **Diabetes**

DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

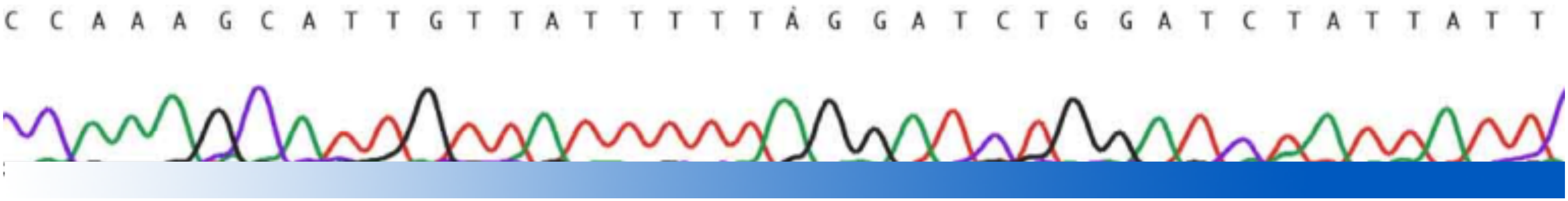
Histones are proteins around which DNA can wind for compaction and gene regulation.

Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

(Image from Wikipedia)

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

Genotyping technologies



Sanger sequencing

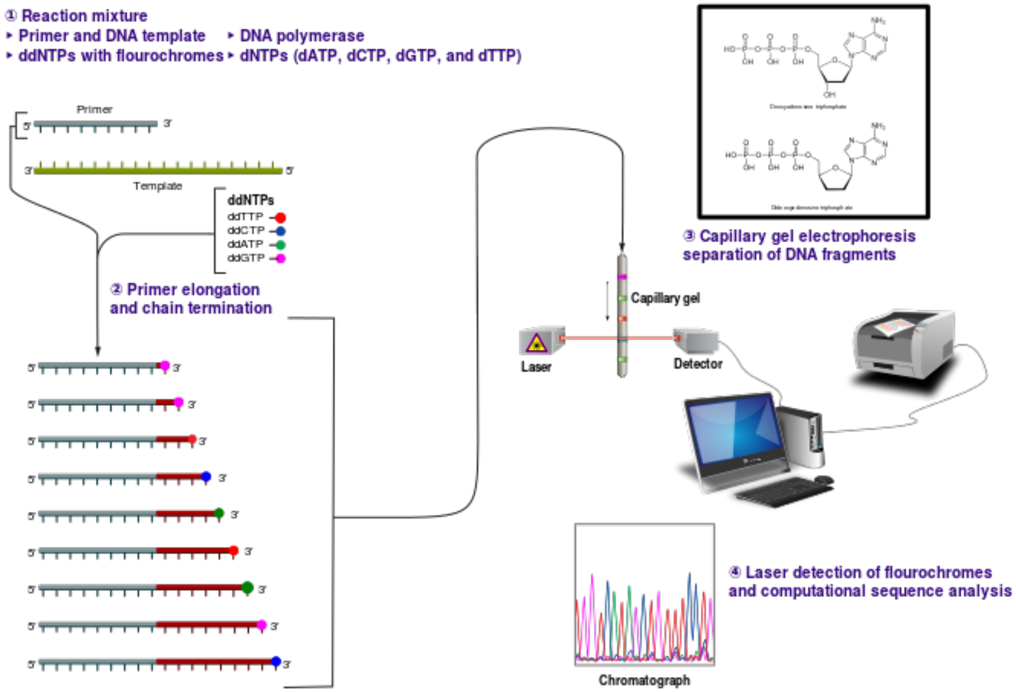

① Reaction mixture
‣ Primer and DNA template
‣ DNA polymerase
‣ ddNTPs with flourochromes
‣ dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
③ Capillary gel electrophoresis separation of DNA fragments
Capillary gel
Laser
Detector
④ Laser detection of flourochromes and computational sequence analysis
Chromatograph

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)

©2008, Illumina Inc. All rights reserved.

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

illumina® HiSeq 2500

Whole genome sequencing: how it works

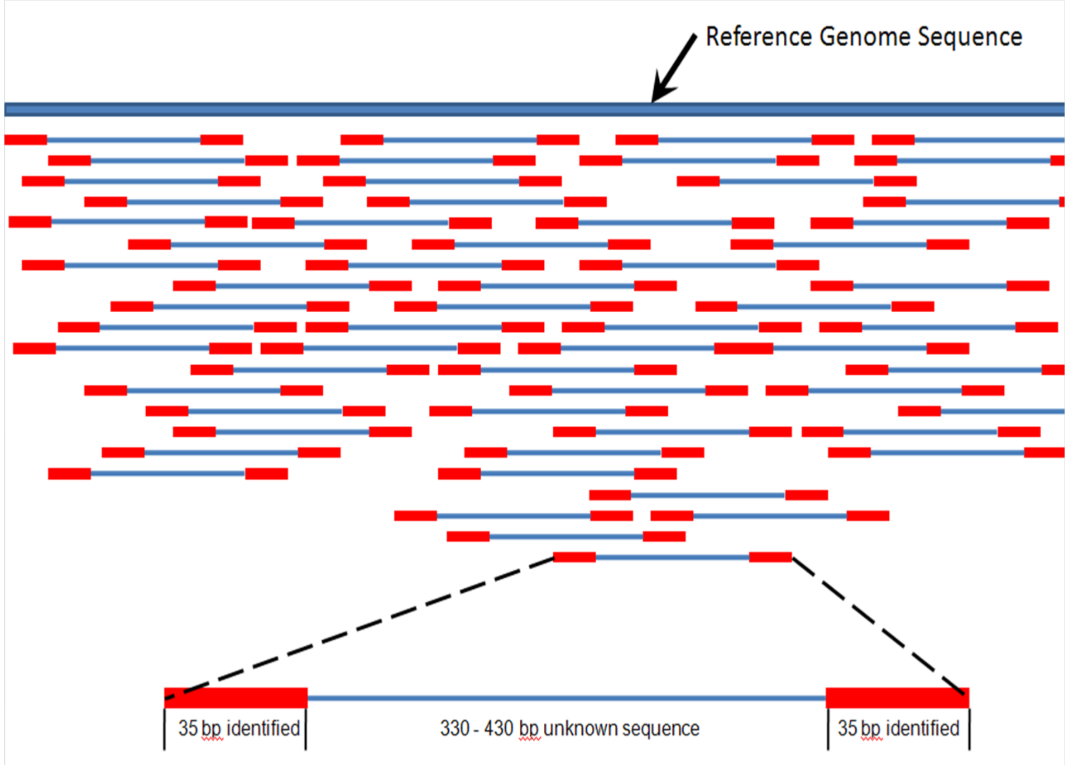


Image rights:
Wikimedia Commons

- 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

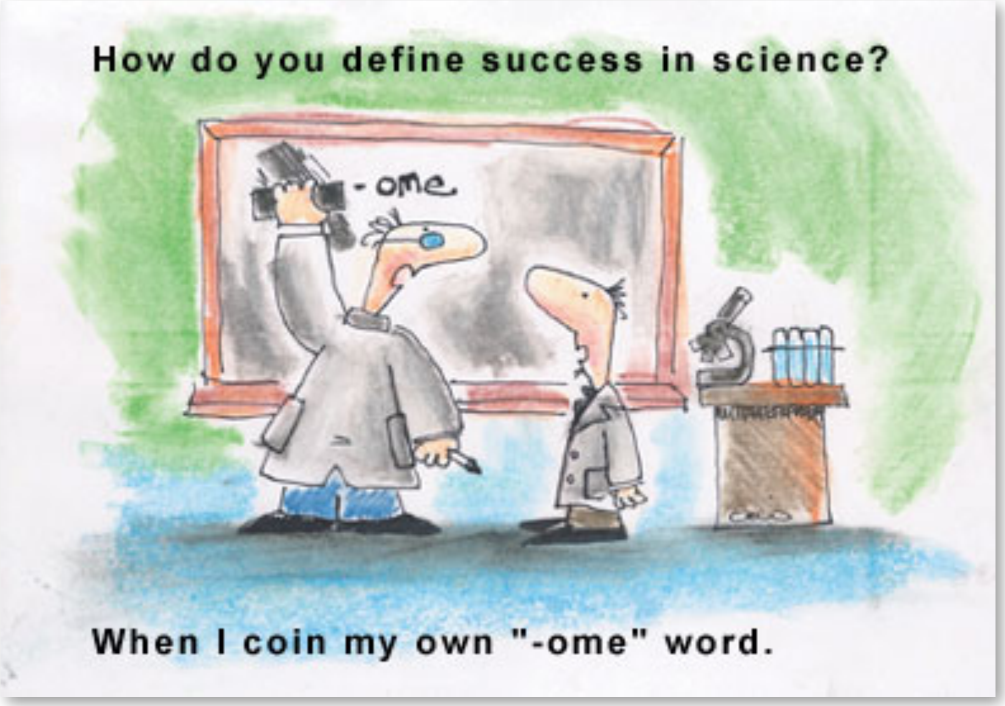

How do you define success in science?
- ome
When I coin my own "-ome" word.

Image rights: Anthony Canamucio (*TheScientist*, 2002)

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