# SISG 2023 - Module 2

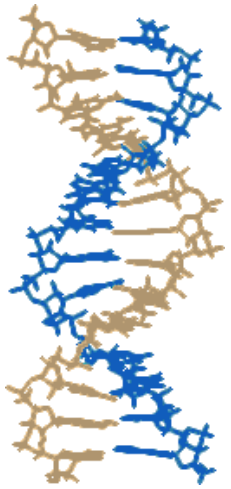
## Introduction to Genetics and Genomics

# Molecular Evolution

Block #6 – Tuesday, July 11

Joe Lachance and Greg Gibson

joseph.lachance@biology.gatech.edu



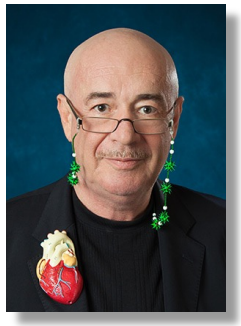
# Functional DNA

- What does it mean to say that a part of the genome is functional?

- What fraction of the human genome is functional?

# ENCODE and the debate about functionality

- ENCODE refers to the **enc**yclopedia **o**f **D**NA **e**lements
- ENCODE: functional elements encode a defined product or display a reproducible biochemical signature
- Evolutionary conservation is another way to infer functional DNA
- Ewan Birney: "80% of the human genome has a biochemical function"
- Dan Graur: "An example of function that fits the ENCODE definition: shoes binding to chewing gum"
- 10-15% might be a better estimate



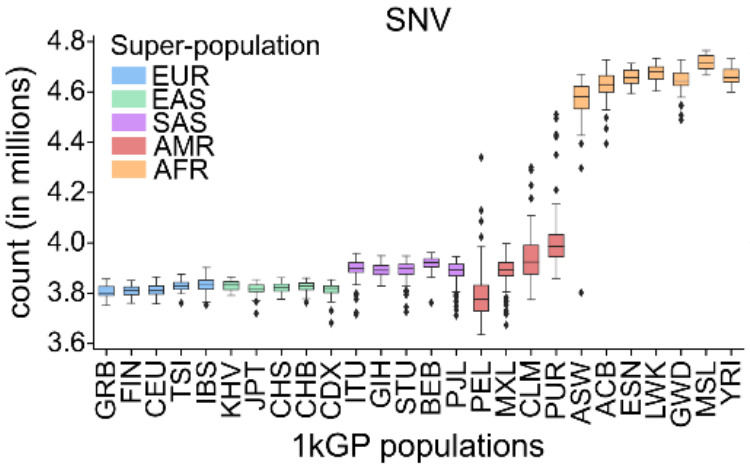
# Junk DNA and the evolution of genomes

- **Junk DNA** refers to sequences that have no known function
- Species with small population sizes tend to have more junk DNA
- Genomes are not static – they change over evolutionary timescales
- Junk DNA can be repurposed

Image: Disney · PIXAR

# SNPs

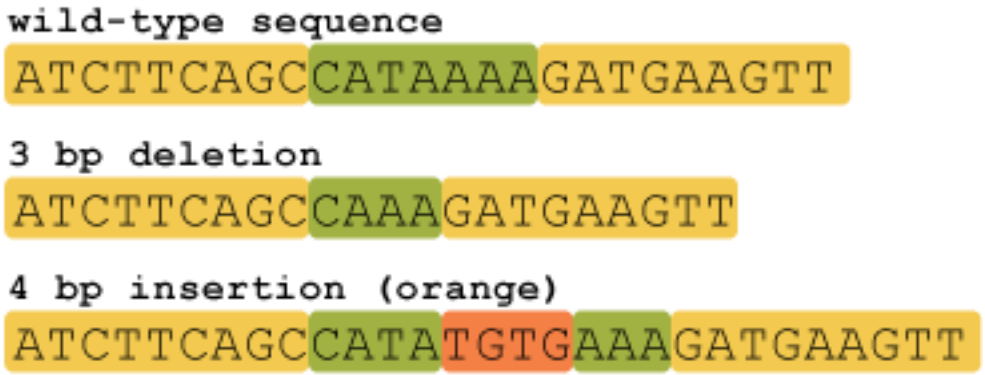
- **SNVs** refer to Single Nucleotide Variants (e.g., **A** or **G**), and when minor allele frequencies are above 1% these variants are called **SNPs** (Single Nucleotide Polymorphisms)




- Each human genome has between 3.8 million to 4.7 million SNVs
  - Genotyping error might overestimate counts of heterozygous sites
  - African genomes contain more genetic diversity than non-African genomes
- Most SNPs are biallelic (they have two alleles)
- More than 660 million polymorphisms are known at present (dbSNP)

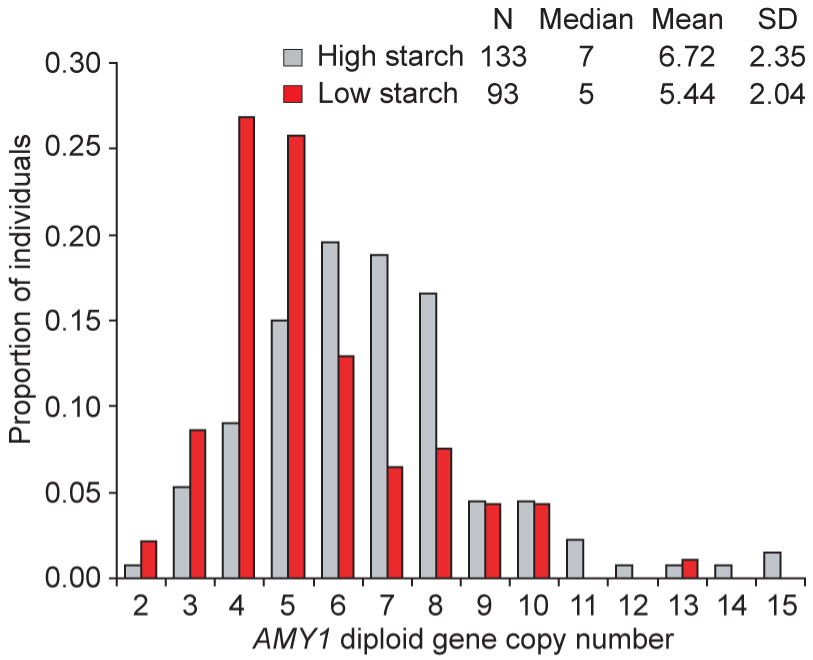
Figure from Byrska-Bishop et al. (bioRxiv)

# Indels



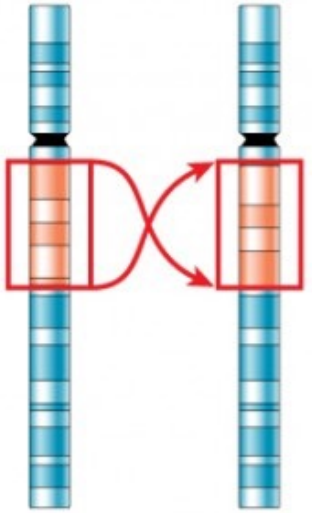

- **In**sertions or **del**etions (indels)
- Human genomes have between 540k and 625k indels
- Most indels are small
- Indels in coding regions tend to be multiples of 3bp.  Why?

# CNVs




- **CNV**: copy number variation
- Data from Perry et al. (*Nature Genetics*, 2007)
- Humans with high starch diets have more copies of *amylase genes*

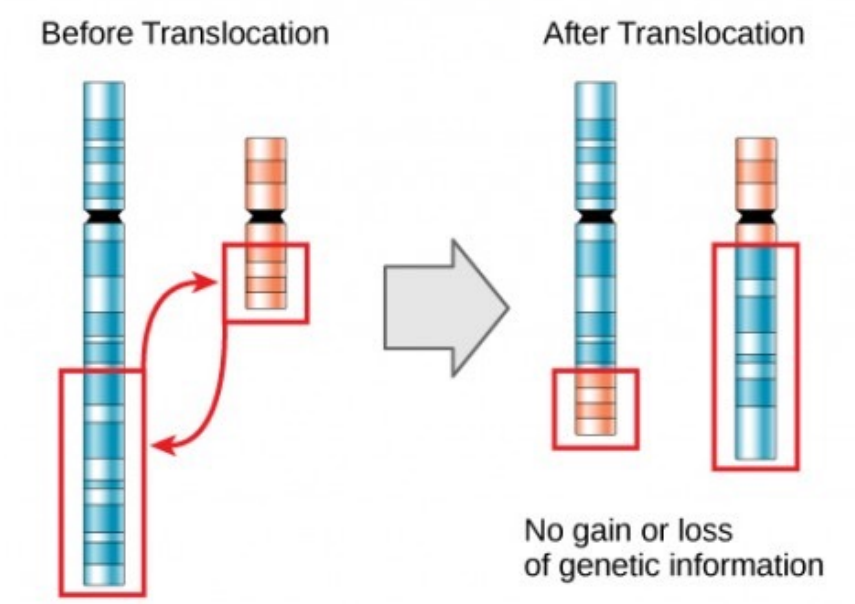
# Inversions



- **Inversions** are chromosomal rearrangements in which a segment of a chromosome is reversed from end to end

- Inversions inhibit recombination (crossover products are not recovered)

# Translocations




- **Translocations** are chromosomal rearrangements in which genetic material is exchanged between chromosomes

- Can cause genes to be mis-regulated and problems during meiosis

# Translocation example

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

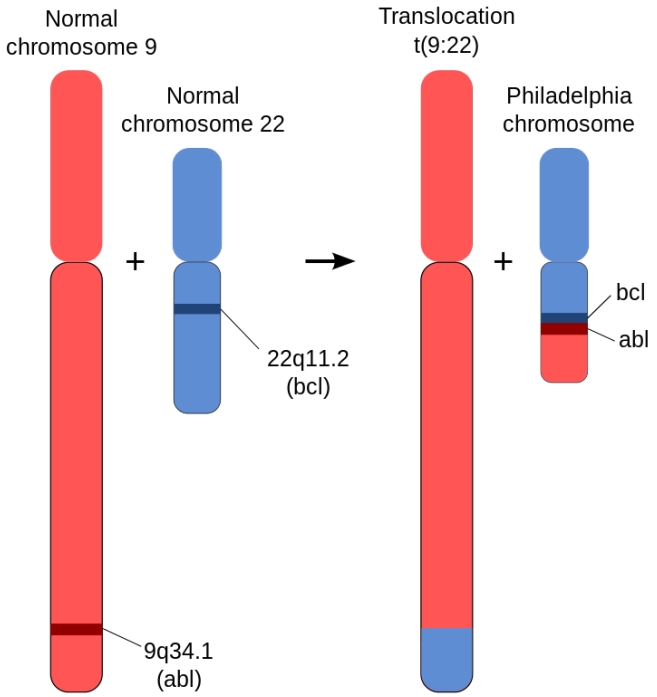
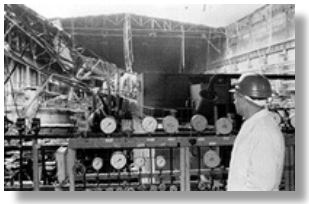



Image from Wikipedia

# Causes of mutations

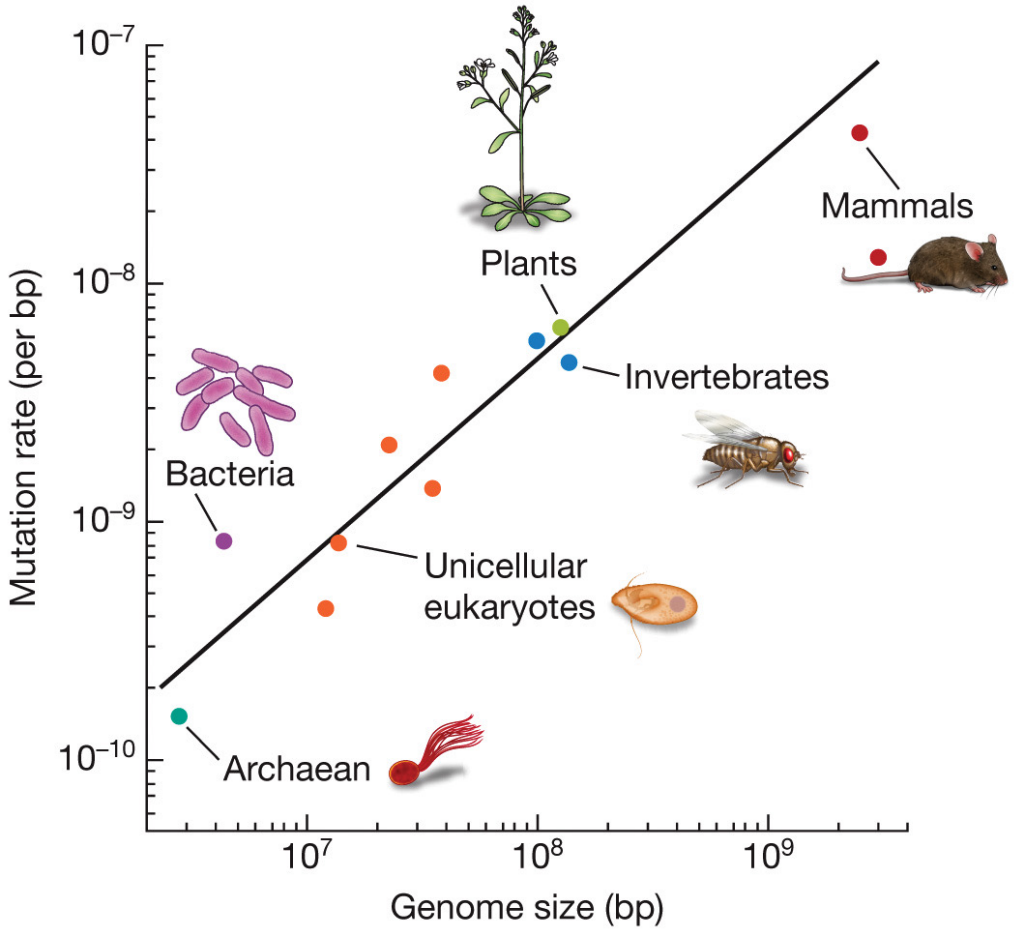
- DNA replication errors
- Chemical mutagens
  (think of the Ames test)
- Radiation (X-rays and UV)



- What are the evolutionary impacts of the Three Mile Island, Chernobyl, and Fukushima Daiichi disasters?

Image sources: Rick and Morty (Adult Swim) and NPR

# Mutation rates vary widely across species




*EVOLUTION 4e*, Figure 4.15

# Different estimates of mutation rates in humans

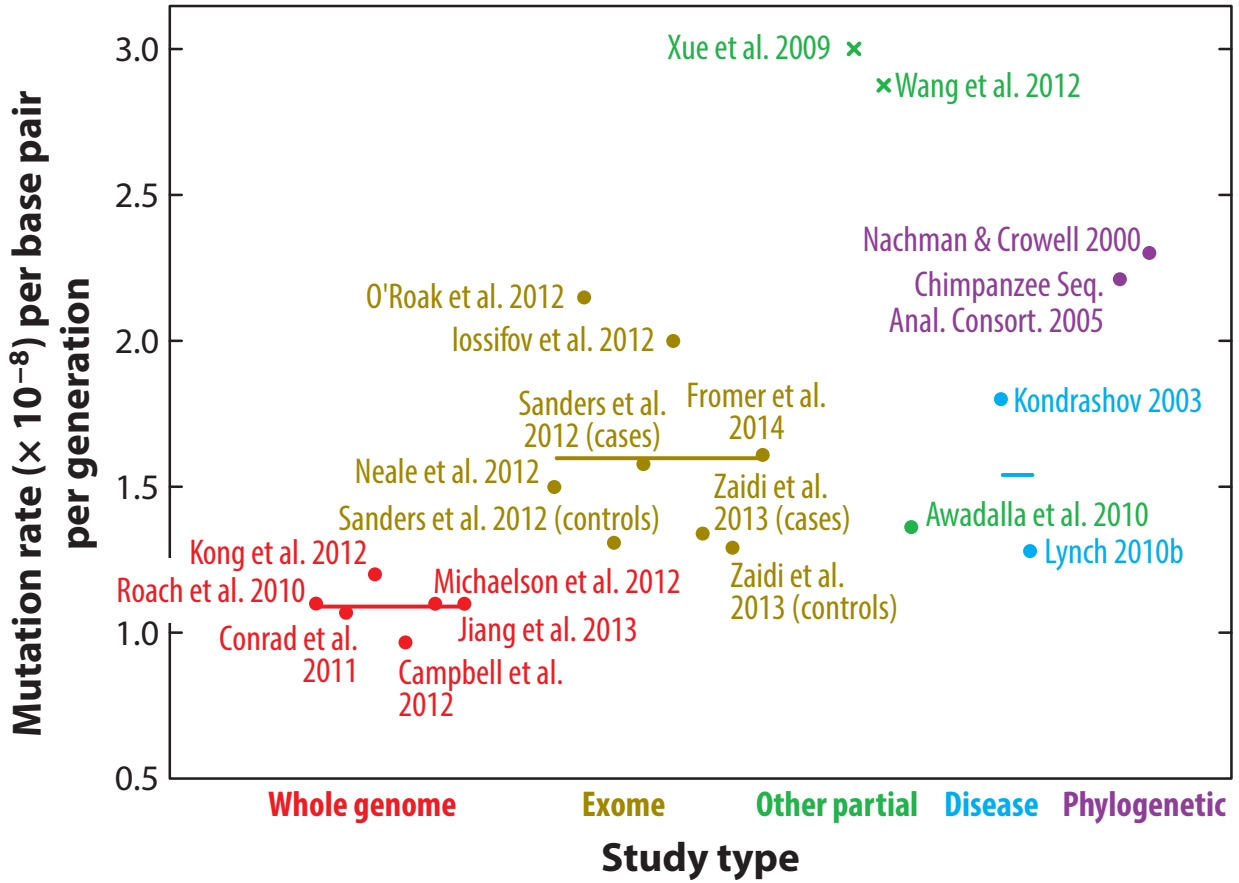


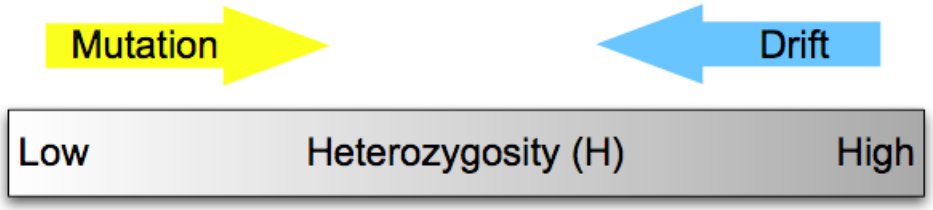

Figure from Ségurel et al. (*Annual Review of Genomics and Human Genetics*, 2015)

# Neutral theory of molecular evolution

- Motoo Kimura (1968)
- Most polymorphisms are neutral (neither good nor bad)
- Examples of neutral variation:
  - Synonymous changes (codon change, but same amino acid)
  - Pseudogenes: “dead genes” that are no longer expressed
  - Intergenic DNA

# Neutral theory of molecular evolution

- A balance exists between a decrease in variation due to random chance (genetic drift) and an increase in variation due to mutation




$$\hat{H} = \frac{4N_e\mu}{1 + 4N_e\mu}$$

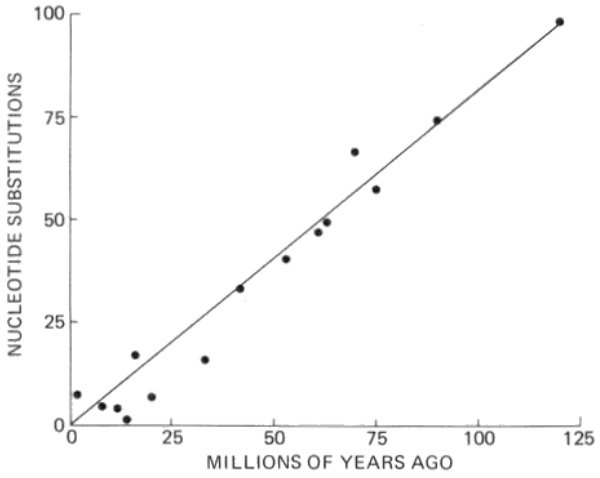
- Large populations have more genetic variation than small populations
- Highly mutable parts of genomes contain more genetic variation
- The neutral theory provides a null hypothesis for studies of molecular evolution

# Neutral-selectionist debate

- What is more important: neutral evolution or natural selection?
- Historical: Motoo Kimura (neutral) vs. John Gillespie (selection)
- Modern day: Jeff Jensen (neutral) vs. Matt Hahn (selection)

Image of the Legislative Yuan in Taiwan: BBC/AFP

# Molecular clock




$$d = 2\mu t$$

$d$ = divergence (proportion of sites)

$\mu$ = mutation rate

$t$ = time (generations)

- Mutations at neutral sites accumulate in a clocklike fashion (but not like a metronome!)
- Genetic data can be used to infer divergence times between species
- First proposed by Zuckerlandl and Pauling in 1962

# $D_n/D_s$ ratios and MK tests

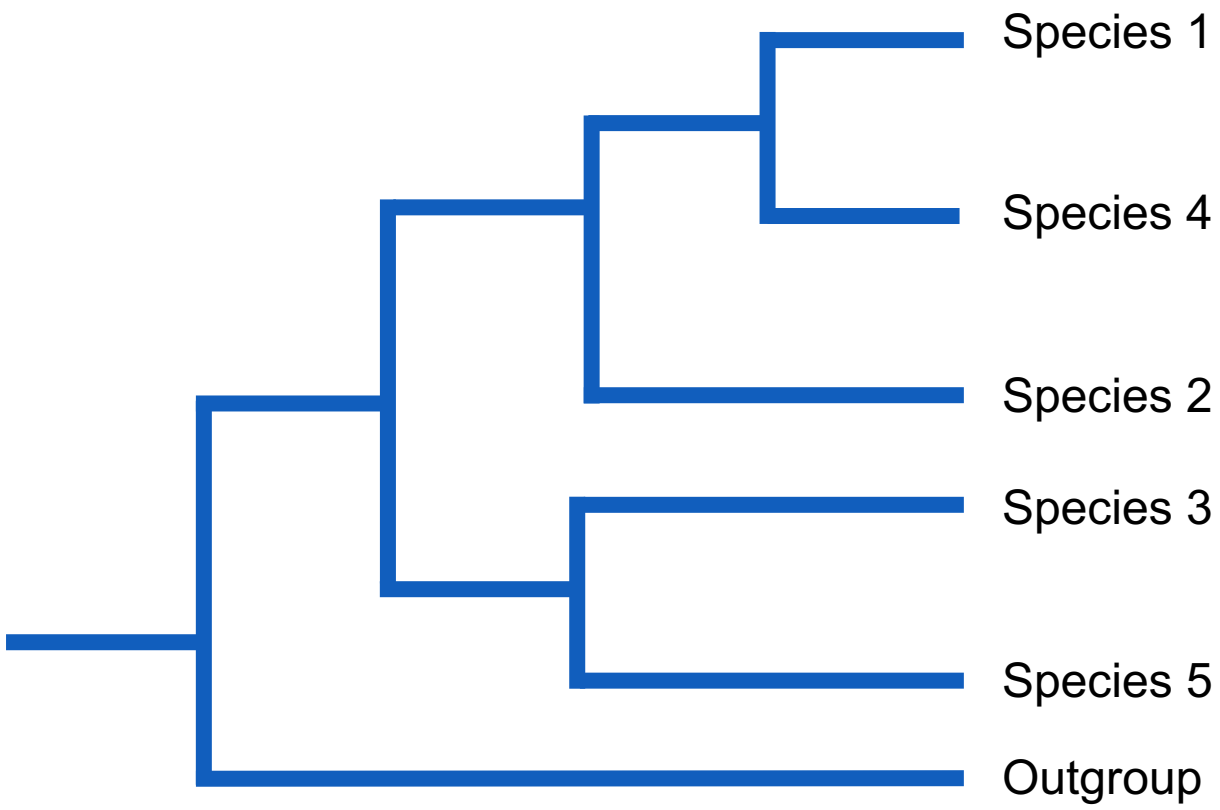
| | Fixed differences between species | Polymorphic within species |
|---|---|---|
| **Nonsynonymous (a.a. change)** | $D_n$ | $P_n$ |
| **Synonymous (no a.a. change)** | $D_s$ | $P_s$ |

$$\text{Neutrality Index } (NI) = \frac{P_n/P_s}{D_n/D_s}$$

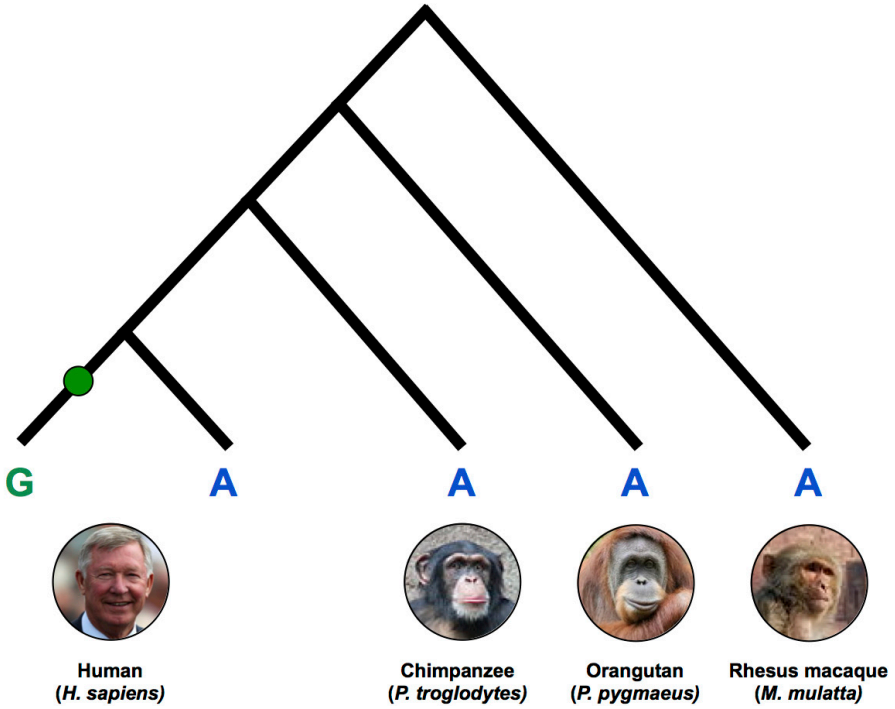
$$NI > 1 \implies \text{negative selection}$$
$$NI < 1 \implies \text{positive selection}$$

- Comparative genomics can reveal which genes have been under selection
- Positively selected genes have an excess of nonsynonymous substitutions
- McDonald-Kreitman (MK) test compares fixed differences and polymorphisms

# Phylogenies describe evolutionary relationships



# Ancestral vs. derived traits (or alleles)

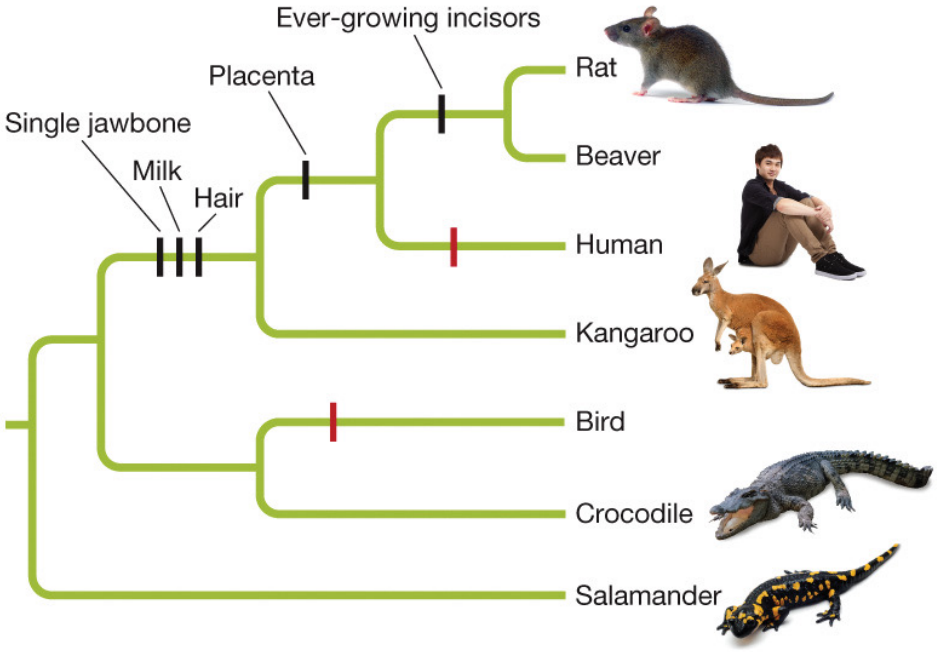



Ancestral allele: **A**

Derived allele: **G**

- **Ancestral** traits are shared with related species
- **Derived** traits are due to recent mutations
- How is this information relevant to hereditary disease risks?

# Phylogenetically informative characters




- **Synapomorpy:** shared derived character

- Synapomorphies are phylogenetically informative characters

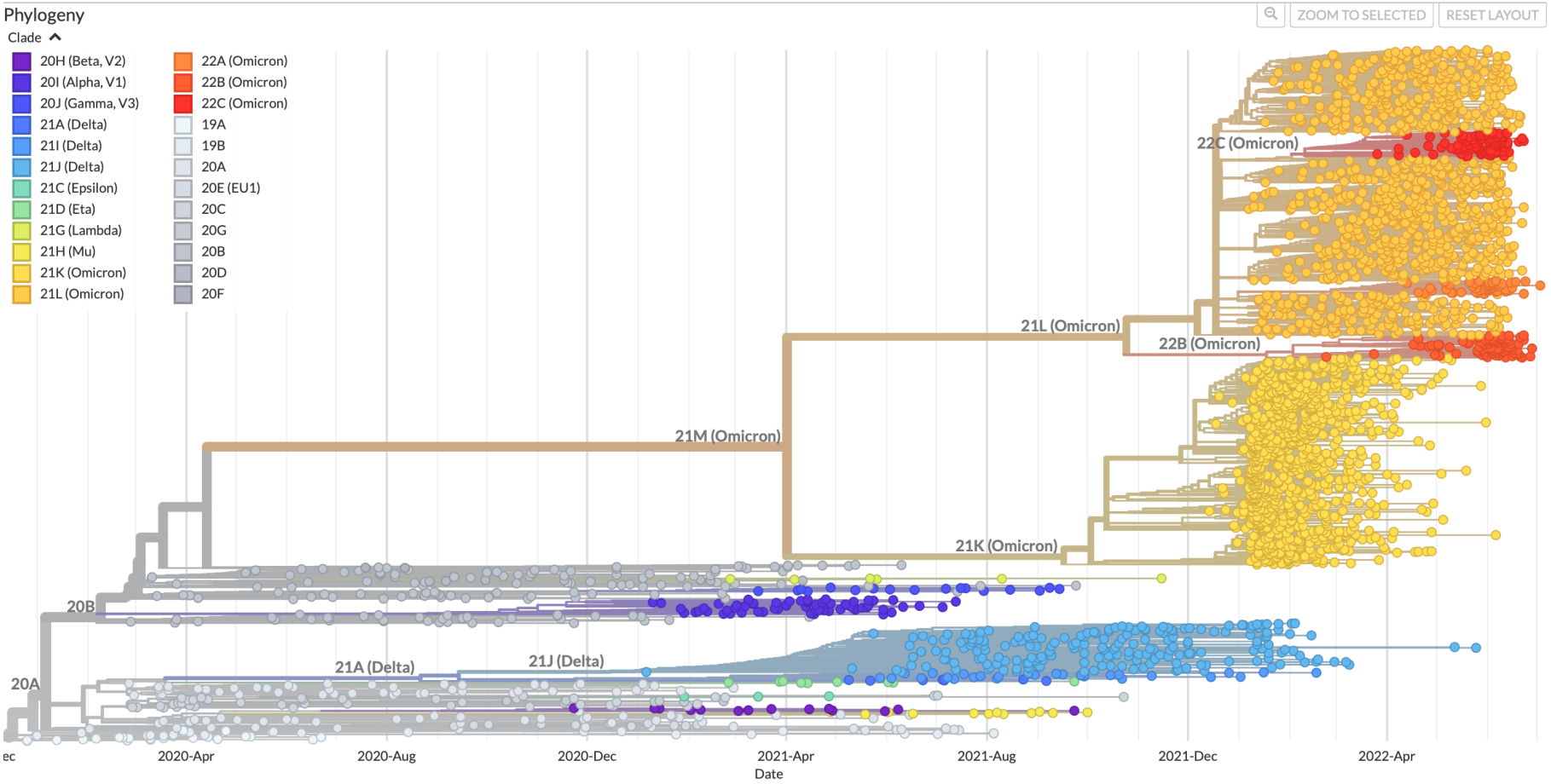
Figure from *Evolution 4e* (Sinauer Associates)

# Genetic data can be used to build phylogenies

| | Species **1** | Species **2** | Species **3** | Species **4** | Species **5** |
|---|---|---|---|---|---|
| Species **1** | 0 | 3 | 7 | 1 | 7 |
| Species **2** | 3 | 0 | 7 | 3 | 7 |
| Species **3** | 7 | 7 | 0 | 6 | 2 |
| Species **4** | 1 | 3 | 7 | 0 | 7 |
| Species **5** | 7 | 7 | 2 | 7 | 0 |

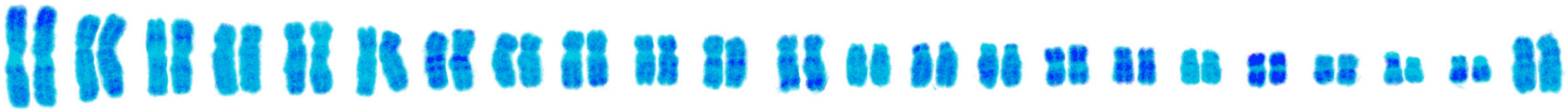
- Pairwise distance matrix calculated by counting the number of sites that differ between each pair of species

# SARS-CoV-2 phylogenetics




Figures nextstrain.org

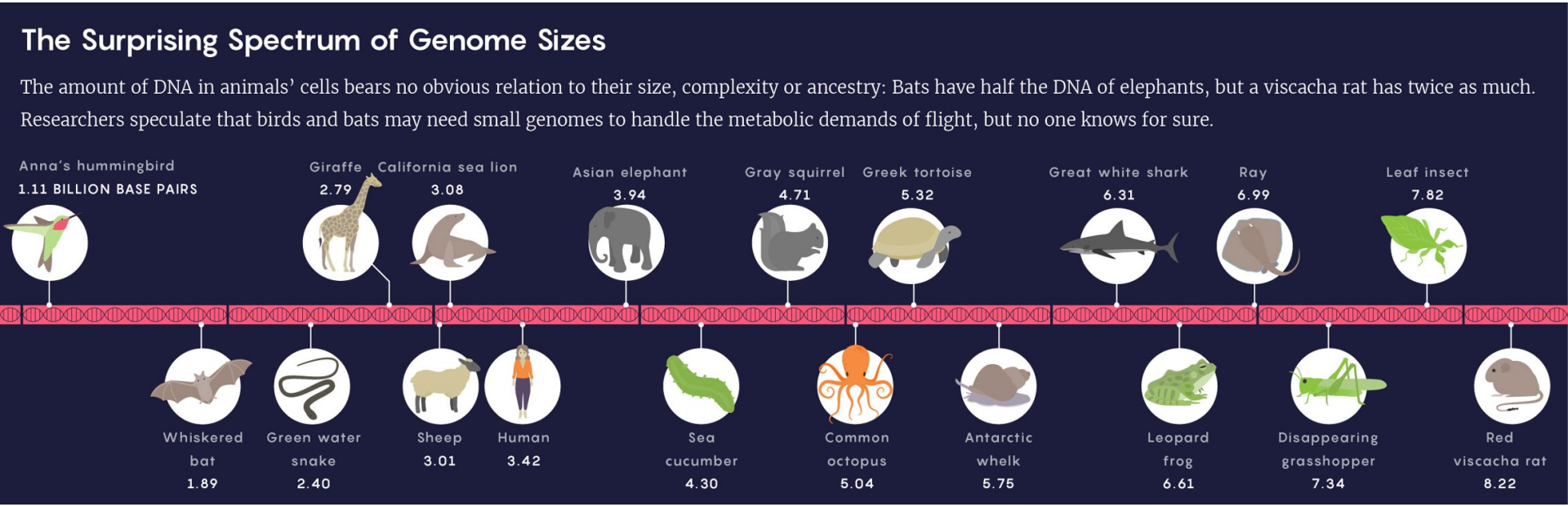
# Variation in the number of chromosomes



- **Karyotype**: the number of chromosomes in the nucleus of a species

- Diploid (2N) chromosome numbers for different species
  - Human *(Homo sapiens*): 46
  - Chimpanzee (*Pan troglodytes*): 48
  - Jack jumper ant (*Myrmecia pilosula*): 2
  - Fern (*Ophioglossum reticulatum*): 1260
  - Ciliate (*Oxytricha trifallax*): 32000

# Genome sizes vary greatly across species

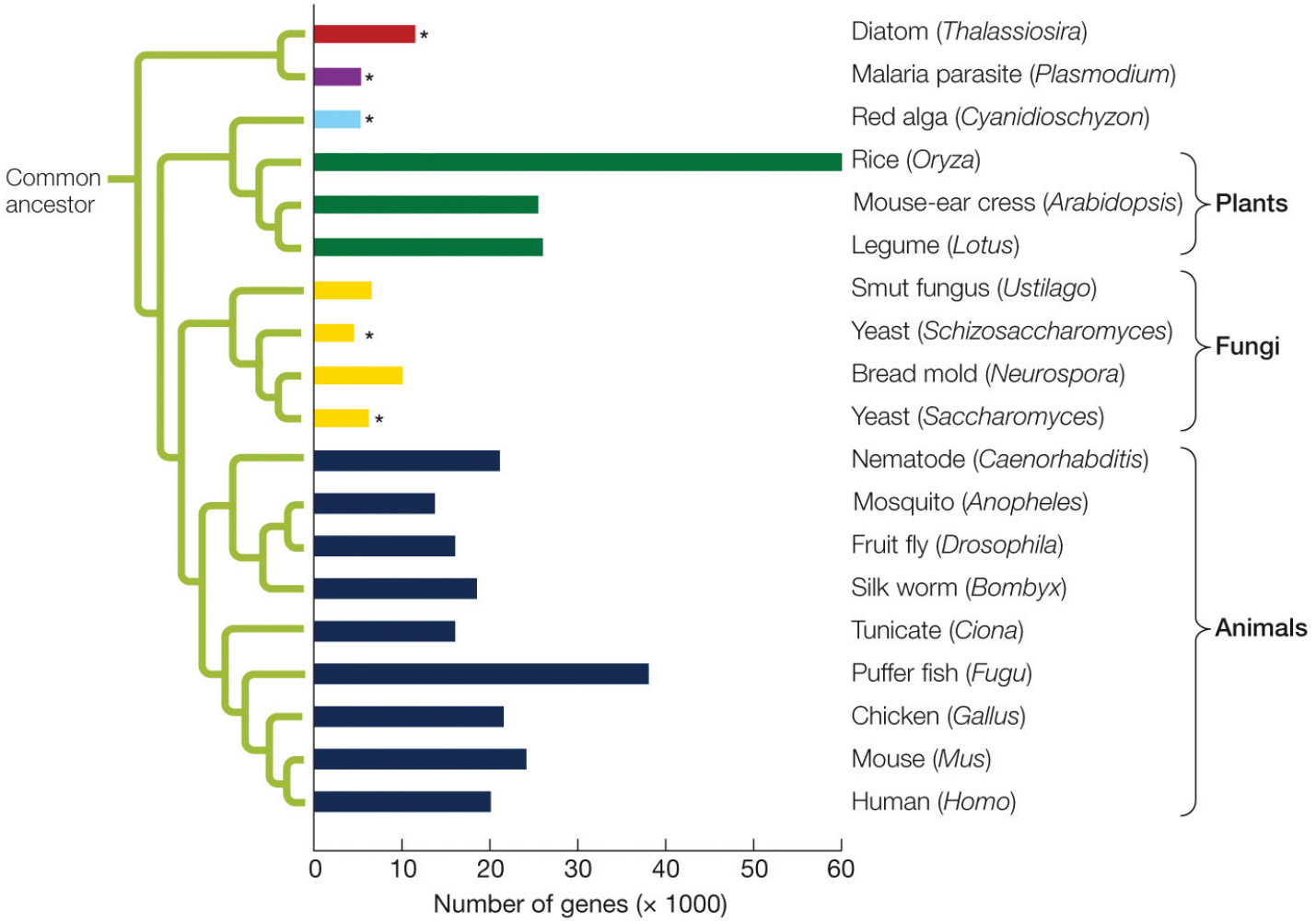



- **C-value paradox**: complex organisms don't always have big genomes (C-value refers to the total amount of DNA in each genome)

- Why might this be the case?

Image source: Quanta Magazine (Lucy Reading-Ikkanda)

# The number of protein coding genes varies by species




*EVOLUTION 4e*, Figure 14.2

# Fates of duplicated genes

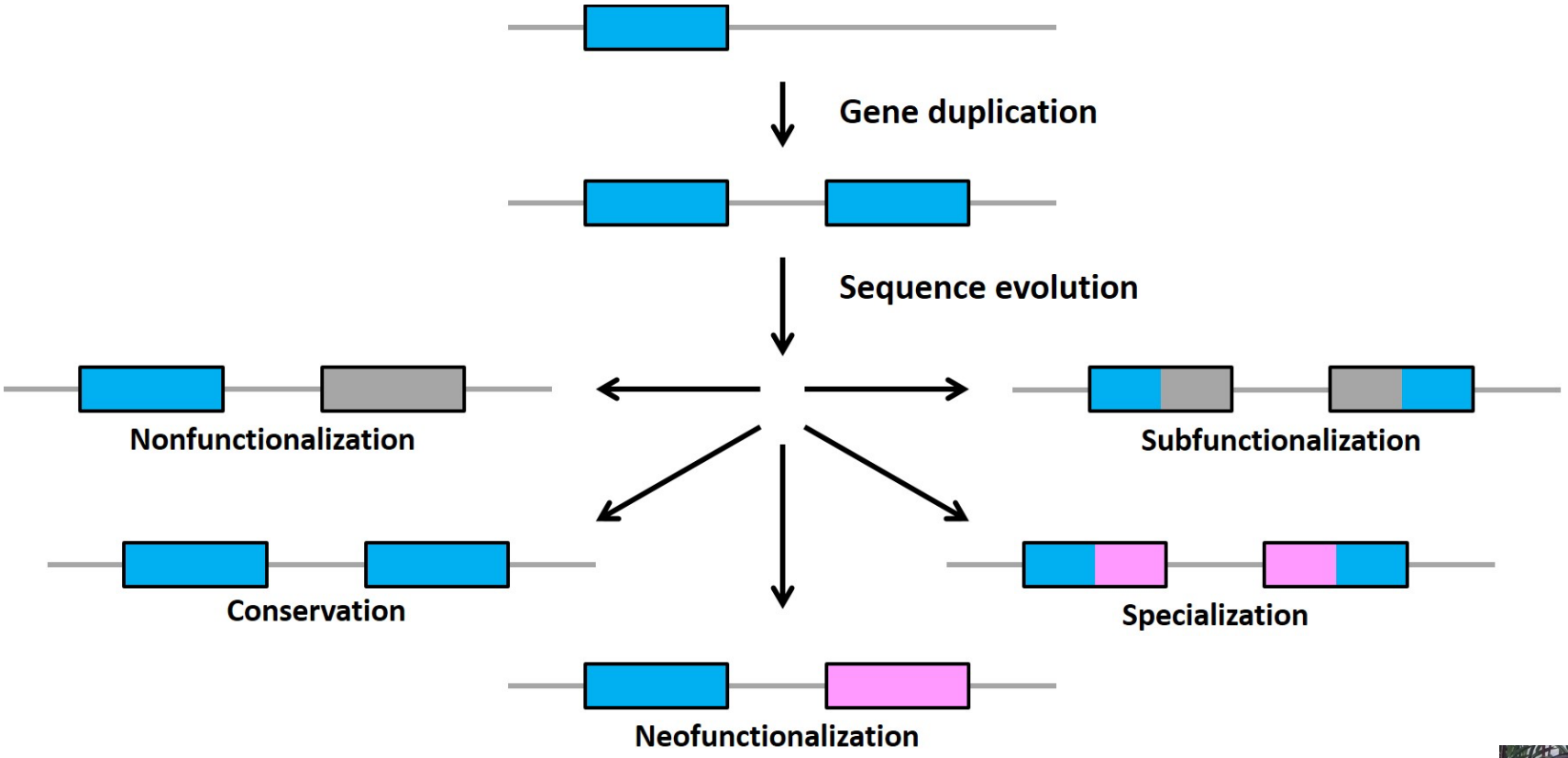
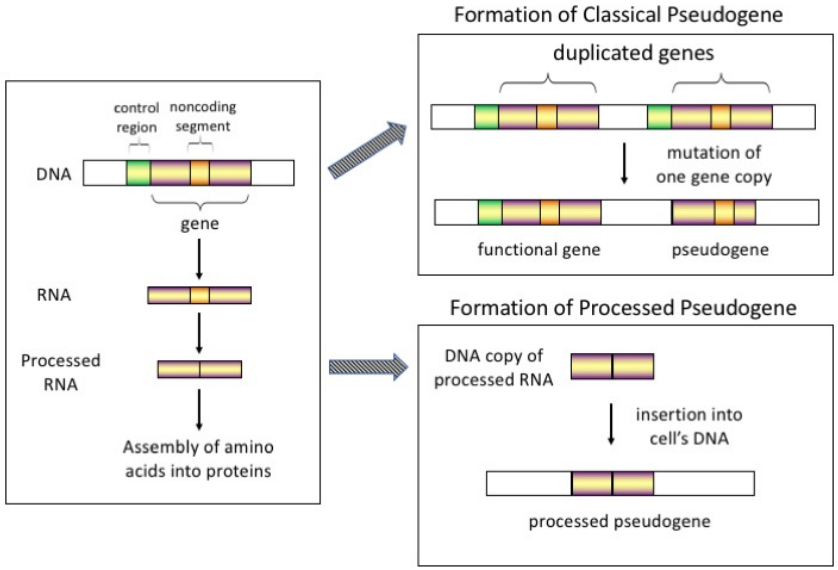



Figure from the Assis Lab at Florida Atlantic University

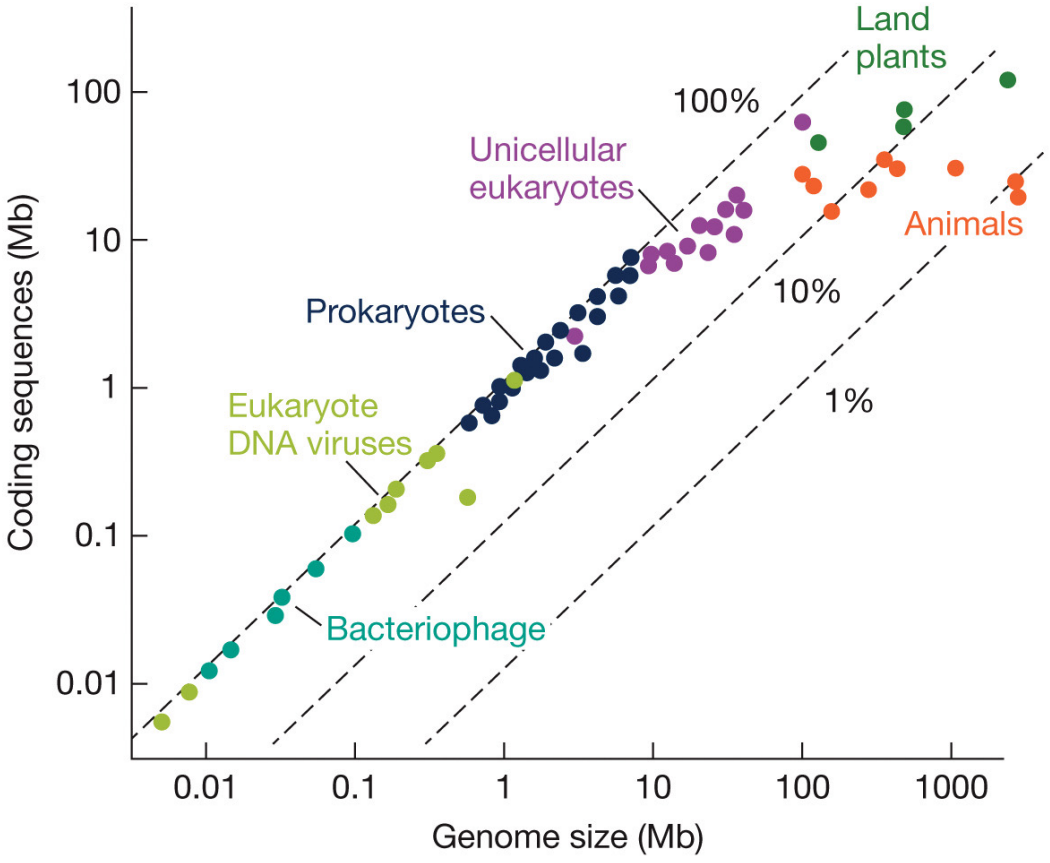
# Pseudogenes




- **Pseudogenes** are nonfunctional versions of normal genes
  - Causes include mutations of premature stop codons
- Classical pseudogenes contain introns
- Processed pseudogenes do not contain introns (they are due to reverse transcription of mRNA into chromosomal DNA)

Image from Wikipedia

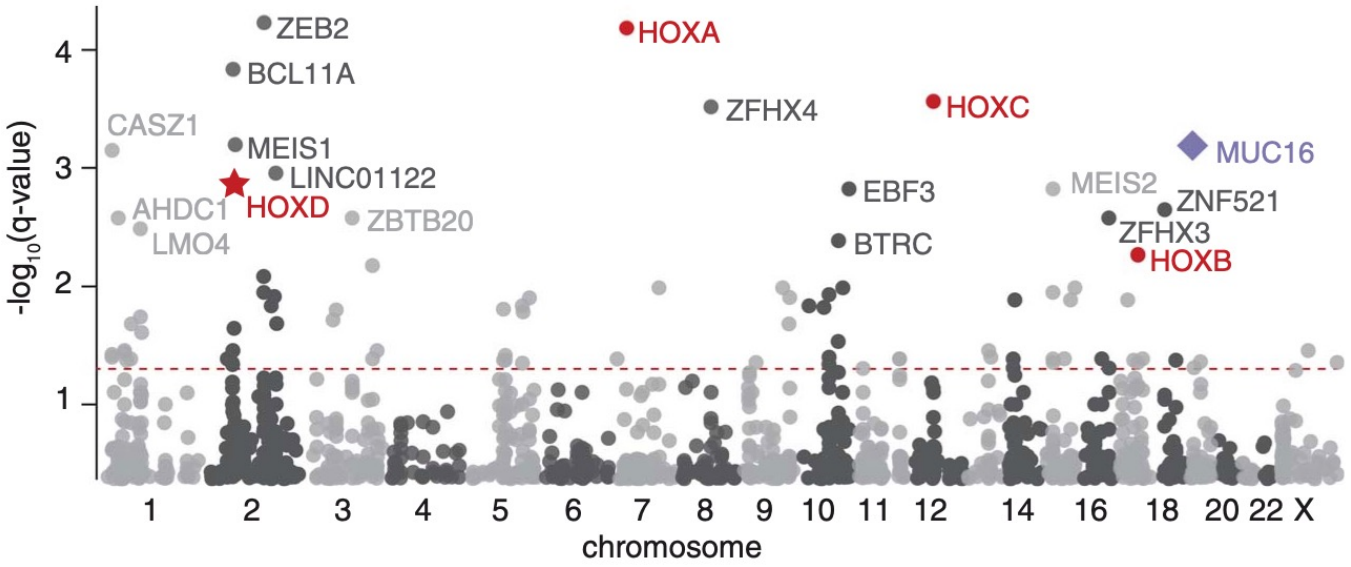
# The coding fraction of genomes varies by taxa




*EVOLUTION 4e*, Figure 14.18


How to explain this pattern?

# Evolutionary constraint in mammalian genomes




- 240 species compared (a major challenge was sequence alignment)
- Evolutionary constraint was quantified on a 100kb-scale

Figure from Christmas et al. (*Science*, 2023)