

---

# Gene & Genome Evolution

---

Kevin Duh

Molecular Biology Reading Group

---

# Introduction

- Natural selection works by applying selective pressure to individuals with different genotype/phenotype
    - But how did genome differences arise in the first place?
  - Outline:
    - Generating genome variation
    - Reconstructing phylogeny
    - Human genome
-

---

# Somatic vs. Germ Cells

- For sexually-reproducing organisms:
    - Germ cell: specialized reproductive cells
    - Somatic cells
  - Mutations in the germ line will be passed on
  - Mutations in somatic cells only affect individual
-

---

# Main Mechanisms for Genetic Change (1)

- Mutation within a gene:
    - Substitution or Insert/Deletion of nucleotide(s)
    - Result of mistake in DNA replication or repair
  - Gene duplication
    - Whole gene duplicated, then each undergo different genetic change
    - Common in Eukaryotes
  - Gene deletion
    - Result from chromosome breakage or repair failure
-

---

# Main Mechanisms for Genetic Change (1)

- Exon shuffling
    - Hybrid gene formed by combination of different genes
    - Joins occurs at intron, so exons stay intact
  - Horizontal (intercellular) transfer
    - Transfer of gene not to progeny, but another existing individual
    - Common in Prokaryotes
-

---

# Mutation rate (point mutation)

- Eukaryotes:
    - $10^{-9}$  per base, per year
    - $10^{-6} \sim 10^{-5}$  per gene, per year
  - HIV, Influenza A Virus:
    - $10^{-3} \sim 10^{-2}$  per base, per year
    - Mutation occurs every round of replication
  - Mutation is the only way new genetic material arise, but are often selectively neutral
    - Genetic drift can be used as evolutionary clock for reconstructing history
-

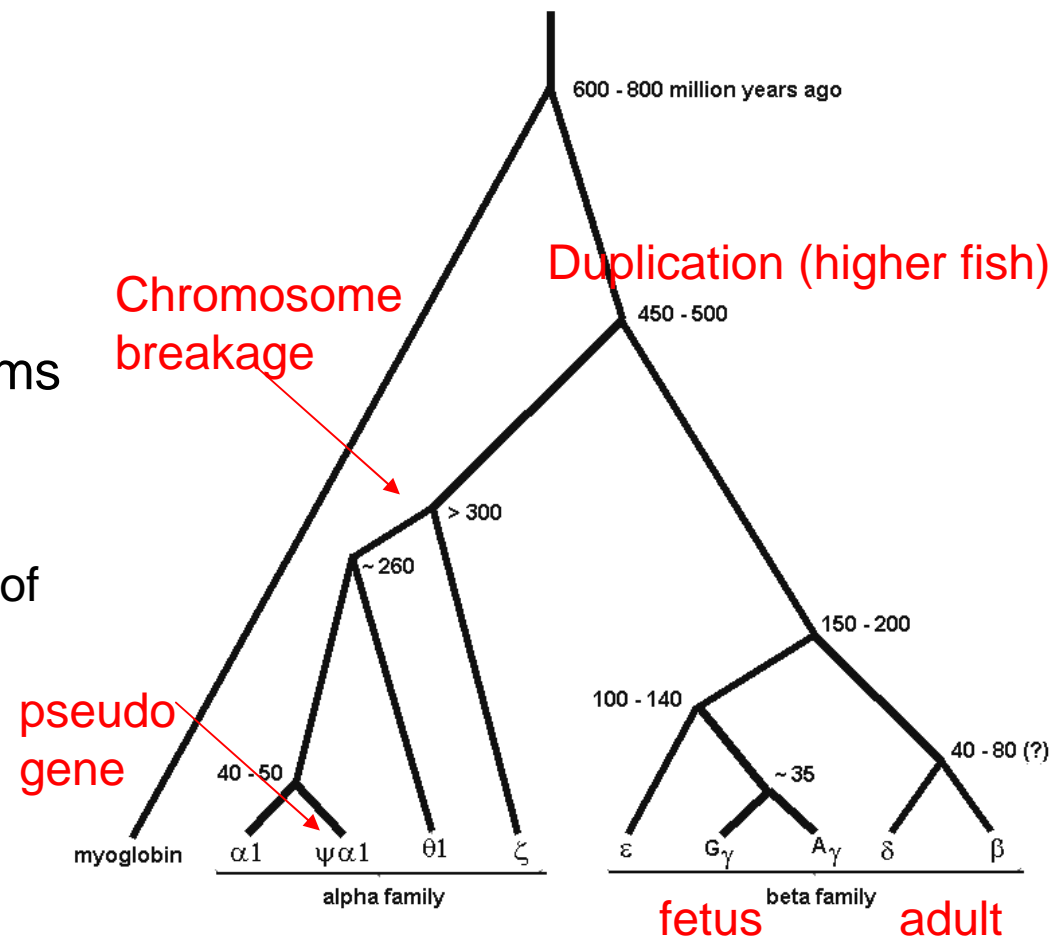
---

# Gene Duplication

- One gene can create a family of genes
    - Bacillus subtilis: >50% genes are in families
    - Almost all genes in vertebrates have multiple versions
      - Hypothesis: entire genome duplicated twice early on
  - Examples:
    - Different opsins (proteins that detect light at different wavelength) are expressed in different retinal cells
    - Hemoglobin
  - After duplication, different copies are free to mutate (without affecting core functionality, probably)
-

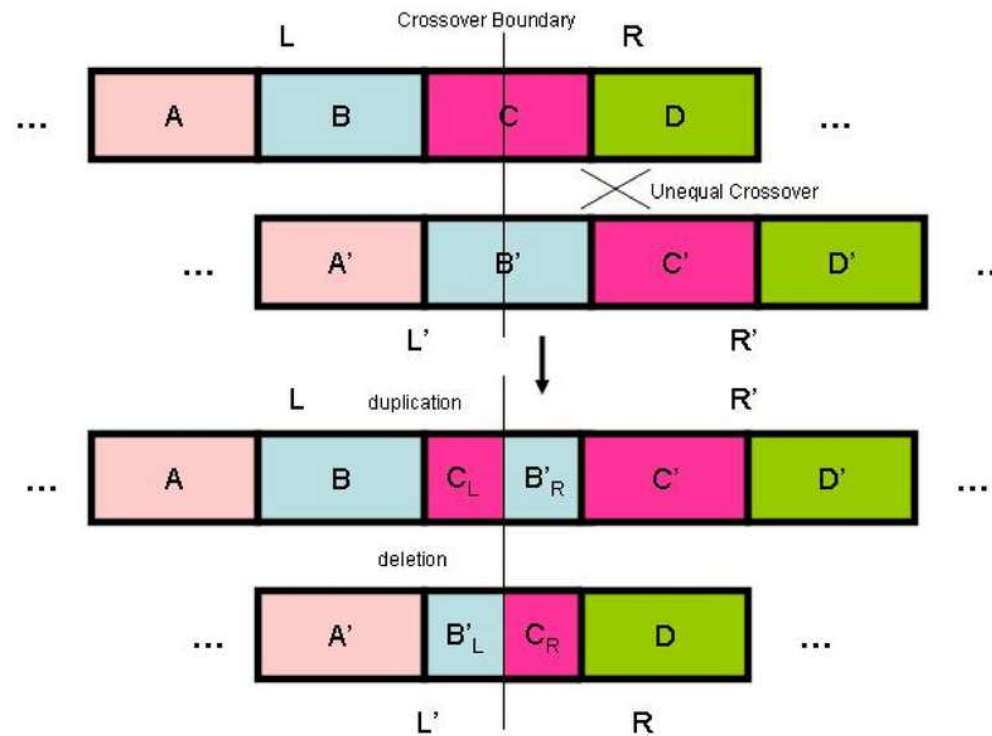
# Globin family

- Most primitive oxygen-carrying molecule in animals:
  - Polypeptide of 150 amino acids
  - Found in marine worms, insects, primitive fish
- Higher vertebrates:
  - hemoglobin composed of alpha & beta chains
  - more efficient
  - carries 4 oxygen



# What creates gene duplication?

## 1. Unequal crossover:



2. Whole genome duplication: e.g. in *Xenopus* frogs

3. Transposons

- [http://hc.ims.u-tokyo.ac.jp/JSBi/journal/GIW02/GIW02F010/GIW02F010\\_fig0002l.png](http://hc.ims.u-tokyo.ac.jp/JSBi/journal/GIW02/GIW02F010/GIW02F010_fig0002l.png)

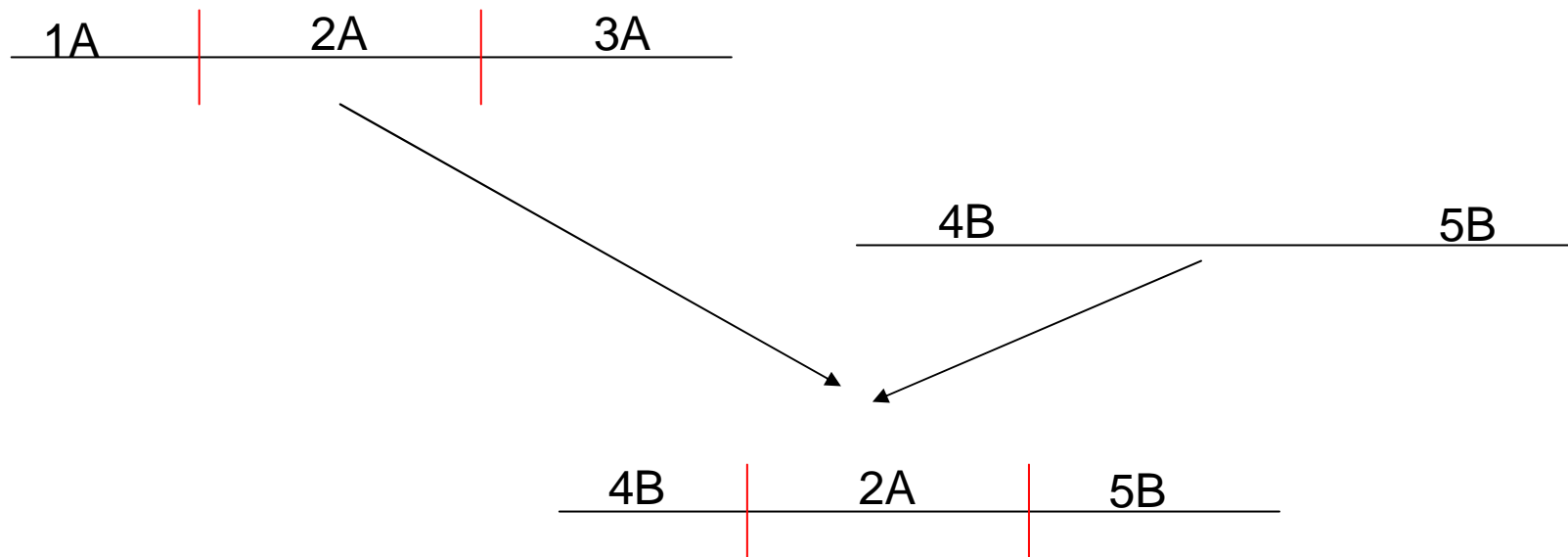
---

# Exon shuffling

- Duplication (unequal crossover) can also happen at intra-gene level:
    - exons are usually short, introns are long, making this process robust
  - ~30k genes in human probably arose from combination of a few thousand exons
-

# Genome evolution has been accelerated by transposable elements

- Transposons can carry exons from Gene A to Gene B



---

# Horizontal gene transfer

- E Coli:
    - 18% of genome is acquired from another species
  - Common way in which bacteria gain antibiotic resistance
  - Primordial cells may have been genetically promiscuous
    - Eukaryotes seem more similar to Archae in genes for replication, transcription, translation, but more similar to Eubacteria in genes for metabolic processes
-

---

# Reconstructing Life's Family Tree

- Homologous genes:
    - Genes that are similar in nucleotide sequence due to common ancestry
    - 50% of human genes are homologous to *C. elegans* or *Drosophila*
  - Highly conserved genes:
    - E.g. ribosomal RNA → useful for studying distant relationships
    - On the other hand, neutral mutations are useful for studying close organisms.
-

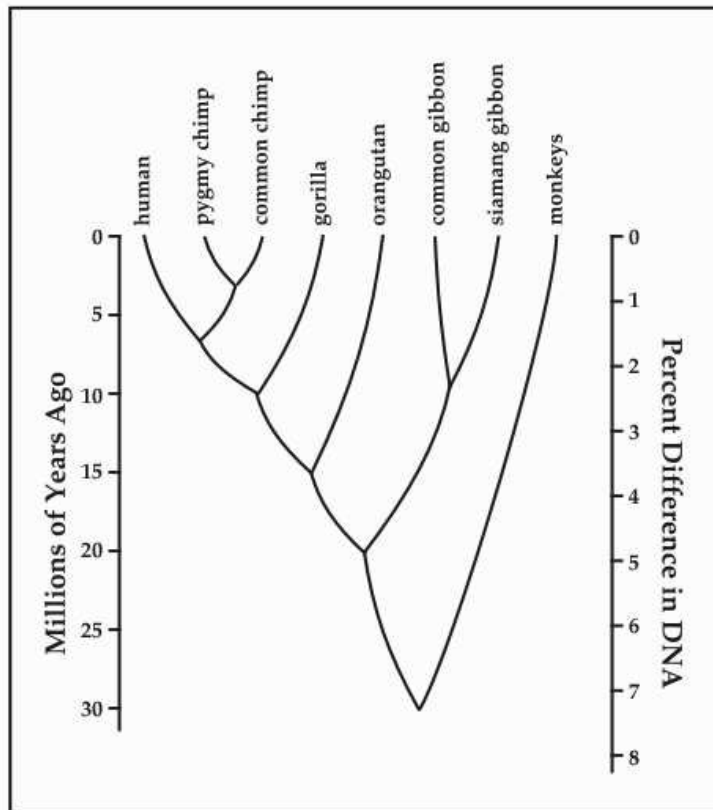
---

# Amount of gene divergence due to mutation can estimate time of speciation

- Assume:
    - 1 of  $10^{10}$  nucleotide mutates in each cell cycle
    - 5% of human genome code for protein and gene regulation; 95% selectively neutral (30k genes)
    - 200 cell division per germ line (from conception to production of egg/sperm)
    - Total DNA in a cell: 6 billion nucleotides
  - Then:
    - By mutation, 100 new differences between child and parent.
    - Between two families, the difference is 200 per generation
    - After 150 generations (early civilization): 30k nucleotide differences
    - After 5-10 million years: 1% genome difference (human vs. chimpanzee)
-

## Primate Family Tree

Humans did not descend from monkeys;  
rather, humans and monkeys  
share a common ancestor



- Both chromosome organization and DNA sequence are similar for human vs. chimpanzee
- 99%+ of the million copies of transposable elements (Alu) are similarly located
- Human chromosome 2 is fusion of 2 chromosomes in chimpanzee, gorilla, orangutan

---

## Human vs. Mouse

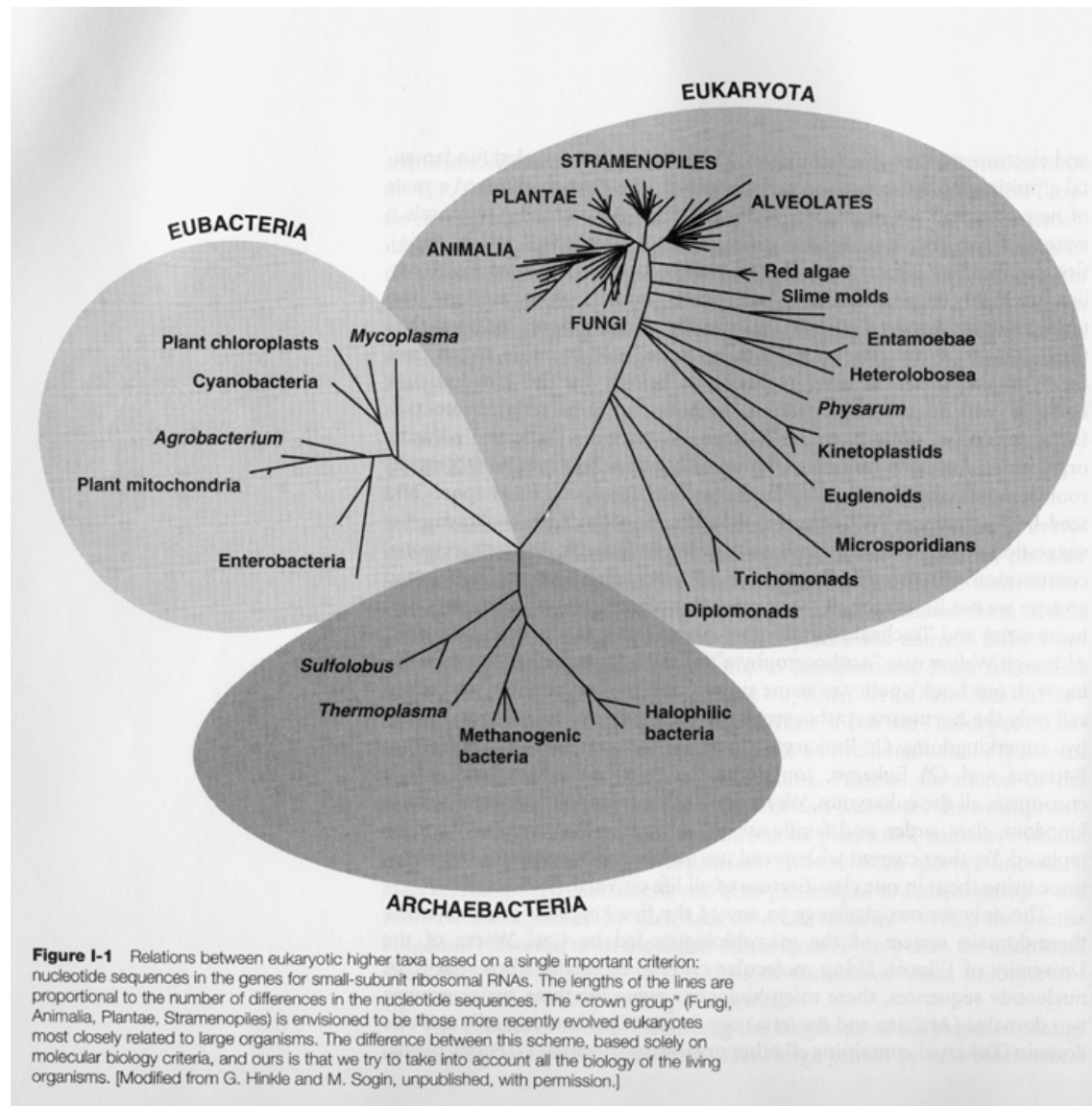
- Diverged 75 million years ago
  - Same number of genes, but transposon distribution differs
  - Centromeres in humans lie at chromosome center; in mouse, at the ends
  - 90% of gene can be partitioned and compared directly
-

---

# Human vs Fish

- Diverged 400 million years ago
  - Different size genomes, amount of gene duplication differs
  - Most sequence differ (except very high conserved ones)
-

- Focus on a highly conserved ribosomal RNA gene, which are present in all living species



- <http://www.geocities.com/herapeuter2002/treeoflife1.jpg>

---

# Human Genome

- $3.2 \times 10^9$  nucleotide pairs
    - on 22 autosomes & 2 sex chromosomes
    - Suppose each nucleotide is 1mm, then whole genome is 3200km (on average a gene every 300m, for 30m long, but only 1 meter of actual code)
  - Individual humans differ by 1 in 1000 nucleotide
    - Human Genome Project includes a variety of individuals
  - Characteristics:
    - Little (2%) protein-coding genes
    - Large average gene size of 27k nucleotides (long introns). Only ~1300 nucleotides needed to code average protein
    - Regulatory genes are spread all over
-

---

# Human genome

“It may resemble your garage/bedroom/fridge/life: highly individualistic, but unkempt; little evidence of organization; much accumulated clutter; and the few patently valuable items indiscriminately, apparently carelessly, scattered throughout.”

---

---

# Single-nucleotide polymorphism (SNP)

- 3 million+ SNP located so far
  - 90%+ of all genes contain at least one SNP
  - May be linked to specific traits/diseases → active area of medical research
  - Variation most likely present at the beginning of human ancestry
-

---

## CA repeats

- CA are replicated poorly due to slippage during replication
  - Individuals may vary by the number of CA repeats
  - Used in forensics, paternity identification
-

---

# Some research frontiers

- Identifying genes in the sea of “junk DNA”
    - E.g. Using comparative genomics
  - Understanding regulatory networks
    - Difference between humans and chimps are amplified by the “developmental program”, not just genetic sequence differences
    - Alternative splicing
      - Allows one gene to produce many different proteins
      - 60% human genes undergo this
-