Chapter 3

Biological Background (#'s)

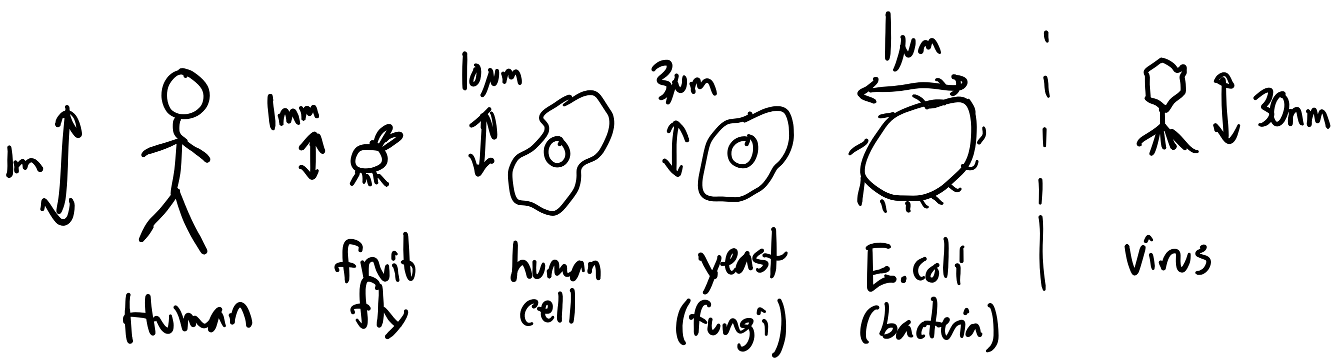
We promised on the syllabus that there are no biology prerequisites for this course, so I'll quickly review some of the background knowledge we'll need to get started. Many of these concepts will be familiar to students who have taken a previous introductory course in biology, but our emphasis on some the key numbers and order-of-magnitude estimates may be new. These numbers may initially seem like extraneous details, but they will turn out to be extremely useful when thinking about evolutionary problems. As we'll see several times throughout the course, they'll allow us to quickly make predictions about which kind of processes may be relevant in a given scenario [1].

[1] For broader introduction to order-of-magnitude estimation in biology, see Milo and Philipps, *Cell Biology by the Numbers*, `http://book.bionumbers.org/`

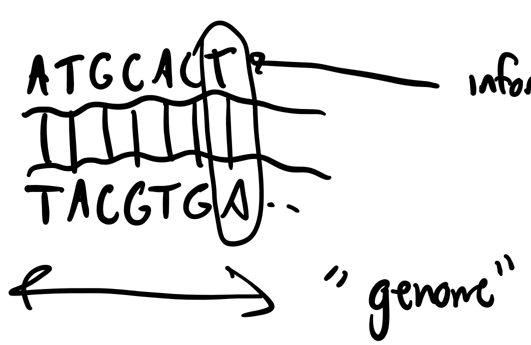
Biological background (key #'s and scales)

① Organisms come in huge range of shapes & sizes:

"Model organisms" we will encounter in this course:



② Despite diffs, these organisms are similar in that instructions to create them are encoded in a single* long molecule of DNA:



We take this for granted now, but pretty crazy from physics perspective! (important info in 1 molecule vs many)

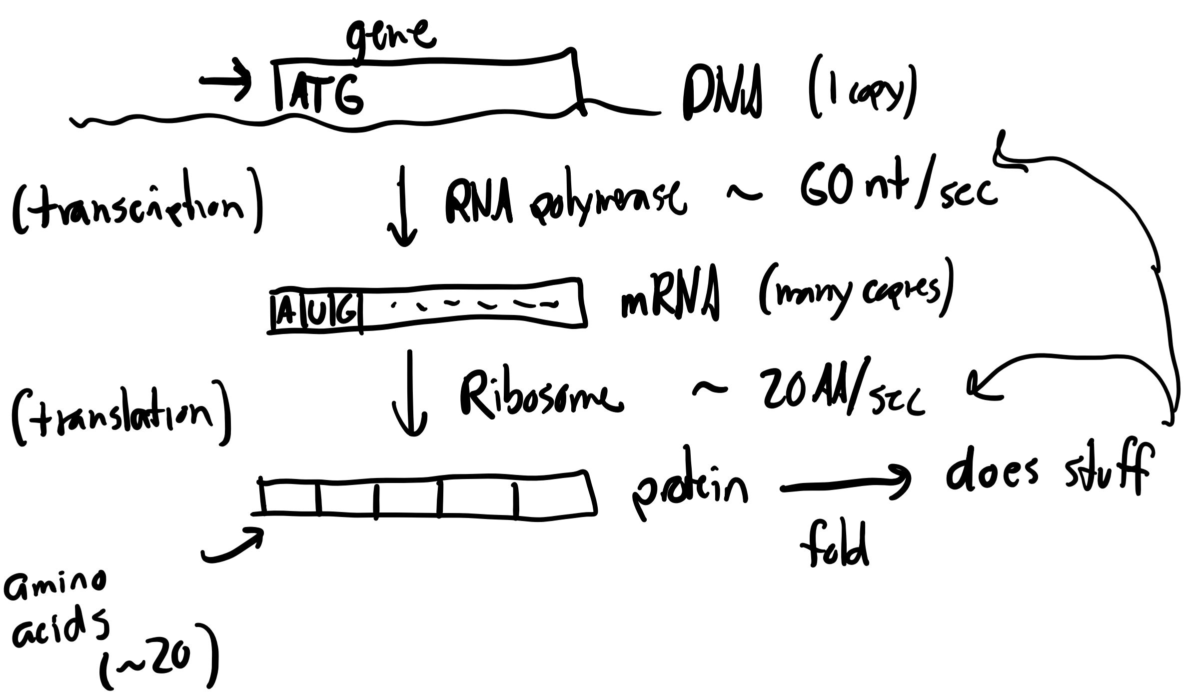
Lengths of genomes vary widely across species:

human: ~ 10^9 bp   yeast: 10^7 bp   virus: 10^4–10^5 bp

fruit fly: ~ 10^8 bp   bacteria: 10^6 bp

(1Gbp = 10^9 bp, 1Mb = 10^6 bp, 1kb = 1000bp.)

information often encoded in genes (make proteins)



How does ribosome do it?

A T T = "codon"

⇒ 1 amino acid (isoleucine)

$4^3 = 64$ different codons → 20 amino acids
+ "start codon"
+ "stop" codon

"genetic code"

⇒ has degeneracy

⇒ typical protein ~300 AA (1000bp of DNA)

⇒ # of genes varies widely across organisms:

humans: 20,000 genes   yeast: 6,000 genes

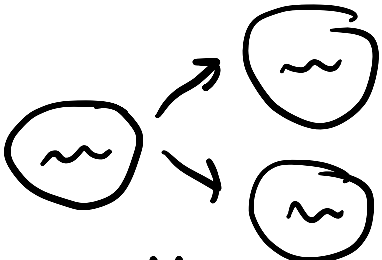
E. coli ~ 4,000 genes   viruses ~ 10 genes.

→ 1000x bigger genome ⇒ but 5x as many genes.

⇒ rest of genome is "noncoding" → regulation
→ "junk"

("coding" = genes)

⇒ net effect of doing all these things is that the organism makes a copy of itself:



Δt

$\longleftrightarrow$

1 doubling time
1 "generation"

① new cell wall, all other proteins (including ribosomes!)

② needs to copy its DNA (DNA polymerase) (not usually limiting factor in growth)

Some characteristic generation times:

humans: ~20 yrs

human cells: ~1 day (HeLa)

Virus: HIV ~ 13 hrs
SARS-CoV-2 ~ 10 hrs

E. coli ~ 20 mins – 1 hr (lab)
1 hr – 1 day? (in gut)

Prochlorococcus ~ 1 day

(ocean bacterium, one of the most abundant photosynthetic organisms on earth, $N \sim 10^{27}$)

⇒ Since n=1 genome, can make errors during copying

.... ATGCCA .... parent
.... ATGTCA .... offspring

"mutations"

⇒ simplest mutations are "point mutations" (A→T, T→C, ...)
aka "single nucleotide mutations" / "substitutions" / "SNPs"

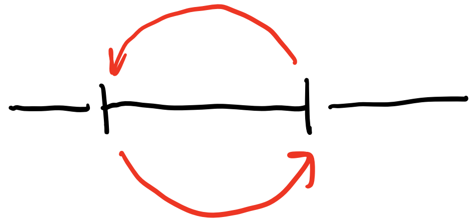
⇒ can also have "insertions": ...ATGTTTCA...
↓
...ATGTTTTTTCA...
(+3T)

⇒ or "deletions" ...ATGTTTCA...
↓
...ATGTCA ...
(-2)

e.g. slippage of DNA pol

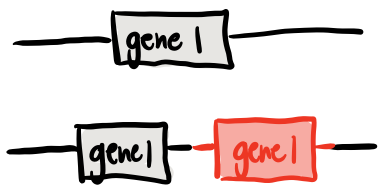
⇒ can also have larger "structural rearrangements":
(e.g. >1kb)

e.g. "inversions"

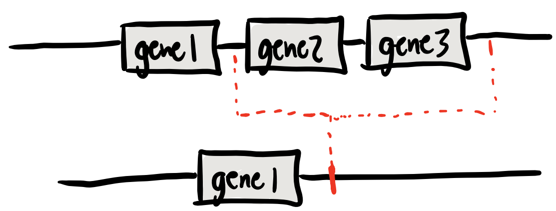


often mediated by special genes known as "transposons"

e.g. "duplications"



e.g. "deletions"



⇒ upshot: can get pretty complicated

⇒ cells have sophisticated machinery for fixing errors that occur in genome...

⇒ net mutation rates (μ) vary across organisms!

e.g. Humans: $\mu \sim 10^{-8}$ single nucleotide muts / bp / gen

Human cells: $\mu \sim 10^{-10}$/bp/division E. coli: $\mu \sim 10^{-10}$/bp/gen

viruses: up to $\mu \sim 10^{-5}$/bp/gen (SARS-CoV2 10^{-6}/bp/gen)

$\Rightarrow$ Using these #'s, can already make some interesting predictions...

Evolutionary "Fermi Problems"

e.g. in Humans genome is $L = 3 \times 10^9$ bp

(actually ×2, since there are 2 copies of each chromosome "diploid")

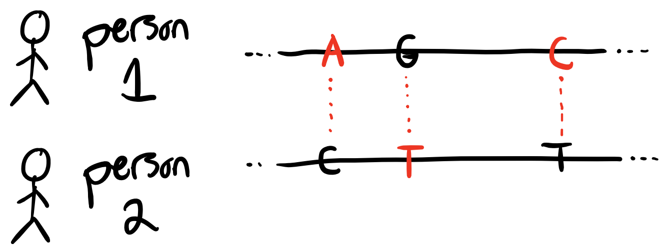
+ mutation rate $\mu \sim 10^{-8}$/bp/gen

$$\Rightarrow L \cdot \mu = 3 \times 10^9 \frac{\text{bp}}{\text{genome}} \times 10^{-8} \frac{\text{mutations}}{\text{bp} \cdot \text{gen}} \approx 30 \text{ mutations per genome per gen.}$$

$\Rightarrow$ there are $N \sim 10^{10}$ humans on earth, so

$$\Rightarrow N \times \mu \sim 10^{10} \times 10^{-8} \sim 100$$ mutations produced @ every site in human genome per generation (in some individual)

⟹ **but**, if we pick 2 random people + compare genomes:



Empirically: differ @ ~0.1% of genome

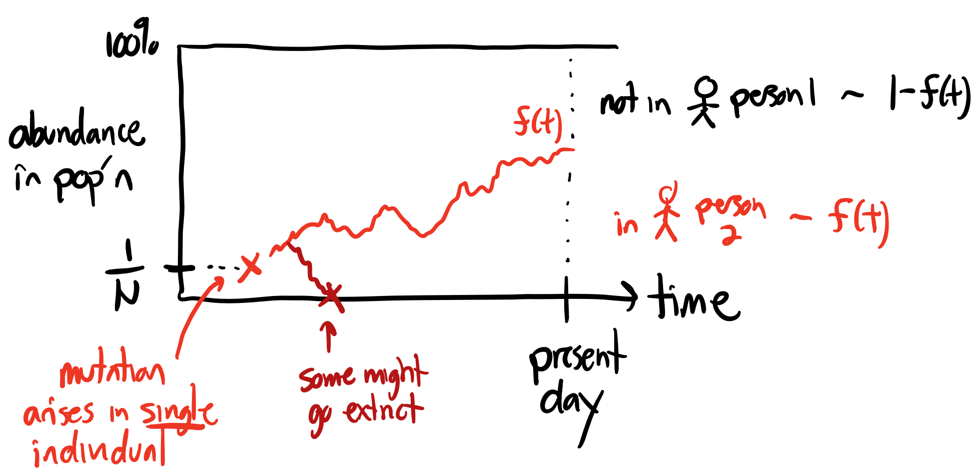
Question: what sets this scale? why not 10^{-4} or 10^{-2}?

⟹ one factor: **dynamics over time**

⟹ for mutations produced in previous generation,

$$\Pr\left[\begin{array}{l}\text{mut in person 1}\\ \text{or person 2}\end{array}\right] \approx 100\ \text{muts/bp} \times \frac{2}{10^{10}} \sim 10^{-8}/\text{bp}$$

⟹ many differences **observed** today **occurred** in past



$\Rightarrow$ must understand mut'n trajectories <u>over time</u> ("dynamics")

(goal of next several lectures...)

<u>Another Fermi calculation</u>:

* all <u>single mutations</u> produced every gen in humans but all <u>pairs of mutations</u> are not:

$$\Pr\left[\begin{array}{c}\text{site 1 \& site 2}\\ \text{mutated in some}\\ \text{newborn}\end{array}\right] \sim N \times \mu \times \mu \sim 10^{10} \times 10^{-8} \times 10^{-8} \sim 10^{-6}$$

$\Rightarrow$ must wait $\sim 10^{6}$ gens (20 million yrs!) for a given pair of sites to mutate @ same time

$\Rightarrow$ <u>Upshot</u>: past dynamics even more important for combinations of mut'ns.

$\Rightarrow$ can also repeat same calculations for E. coli...

$\Rightarrow$ genome is $L = 4 \times 10^{6}$ bp + $\mu \sim 10^{-10}$/bp/gen

$\Rightarrow L \times \mu \sim 4 \times 10^{-4}$ mutations/genome/gen

$\Rightarrow$ >1000 replications before a single error!

$\Rightarrow N_g \sim 10^{9}-10^{10}$ E. coli cells in single person's gut

$\Rightarrow N_g \times \mu \sim 0.1-1$ (almost every bp mutated w/in us each day)

$\Rightarrow N_h \sim 10^{10}$ guts in human pop'n,

$\Rightarrow N_h \times N_g \times \mu \times \mu \sim 0.1-1 \Rightarrow$ almost all double mutations produced in worldwide E. coli pop'n each day

$\Rightarrow$ but <u>not</u> triple mutants $\left(10^{10} \times 10^{10} \times (10^{-10})^{3} \ll 1\right)$

$\Rightarrow$ more generally, for single gene of $L \sim 1000$ bp

$\Rightarrow 4^{L} \simeq 10^{600}$ possible DNA sequences!

compare to $\sim 10^{82}$ atoms in universe

$\Rightarrow$ sequence space is <u>very</u> big (& sparsely populated)

What do mutations do? "genotype ⇒ phenotype map"

⇒ in general, we don't know a priori (even for model organisms like E.coli!)

⇒ but in special cases, can make some guesses based on structure of the genetic code...

e.g. if mutation occurs in a gene:

⇒ changes a codon (e.g. ATC → ATT)

① due to degeneracy, codon could code for same AA

⇒ doesn't change protein "synonymous mutation"

② could change to something else "nonsynonymous mut'n"

↳ e.g. other AA (small change?) "missense mut'n"

↳ e.g. stop codon ⇒ truncates gene (big change)

"loss-of-function" / "nonsense" mut'n