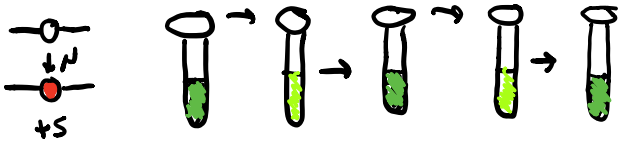
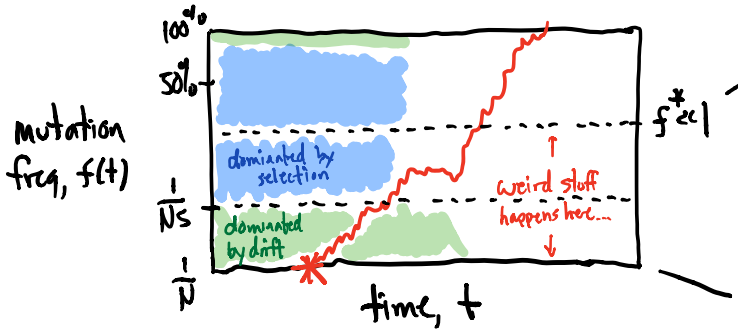
Last time: Quick review - how did we get here?



$$\Rightarrow \quad \frac{df}{dt} = \overbrace{sf(1-f)}^{\text{selection}} + \overbrace{\sqrt{\frac{f(1-f)}{N}}\eta(t)}^{\text{genetic drift}} + \underbrace{[\mu(1-f) - \nu f]}_{\text{mutation}}$$



"Linear Branching Process"

$$\frac{df}{dt} = \mu + sf + \sqrt{\frac{f}{N}}\eta(t)$$

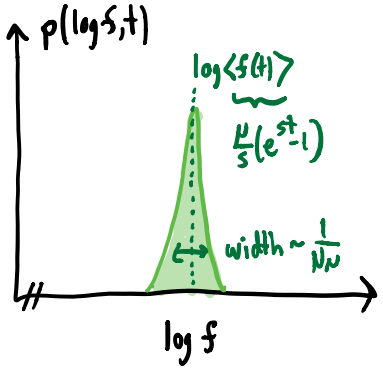
Dynamic mut-sel-drift balance:

$$p(f,t) \propto f^{2N\mu-1} e^{-f/f_{max}(t)} \qquad \text{w/ } f_{max}(t) = \langle f_1(t) \,|\, f_1 > 0 \rangle = \frac{e^{st}-1}{2Ns}$$

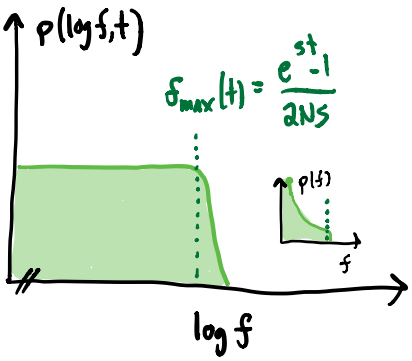
(size of single mut conditioned on survival)

$\Rightarrow$ 2 characteristic behaviors depending on $N\cdot\mu$:

$N\mu \gg 1$



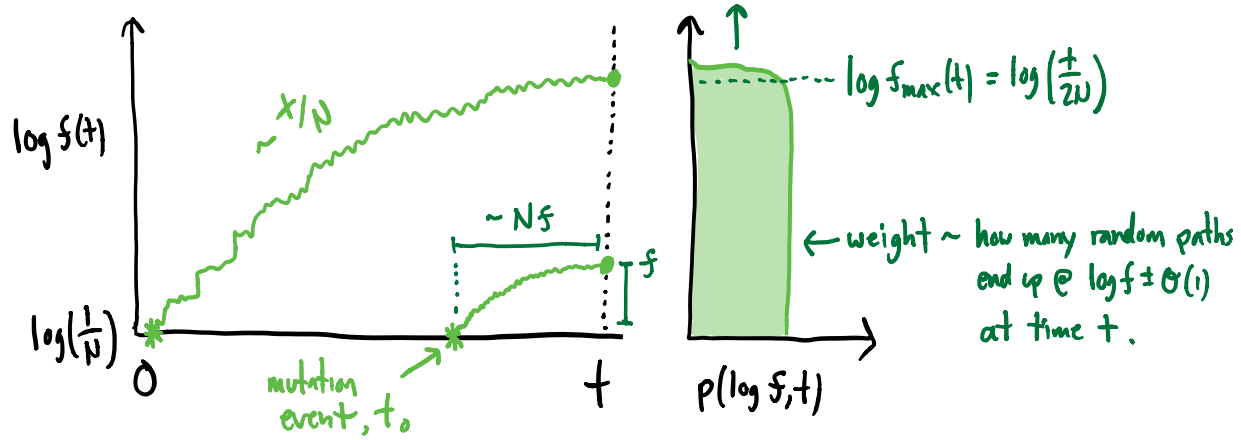
$N\mu \ll 1$



Today: (1) what's going on in $N\mu \approx 1$ case? (2) DNA sequencing

E.g. for neutral mutation, $\boxed{p(f,t) \approx 2N\mu f^{-1} e^{-2Nf/t} \quad (N\mu \ll 1)}$

* can understand dist'n as contribution from @ most 1 mutation event:



* What range of times contribute to $p(\log f)$?

$\Rightarrow$ since $f \sim t/N \Rightarrow \log f \pm O(1) \quad \Delta t_0 \sim Nf$

* Putting everything together $\Rightarrow$ heuristic formula for $p(\log f)$

$$p(\log f, t) \cdot \underbrace{\Delta \log f}_{O(1)} \sim \underbrace{N\mu}_{\substack{p(\text{mutation}) \\ \text{per gen.}}} \times \underbrace{\Delta t_0}_{\substack{\text{range of} \\ \text{origination} \\ \text{times}}} \times \underbrace{p\left(\tfrac{1}{N} \to f\right)}_{\substack{\text{probability of} \\ \text{drifting to } f}} \sim N\mu \times Nf \times \frac{1}{Nf} \to 1$$

$$\Rightarrow p(\log f, t) \sim N\mu \quad (f \ll f_{max}) \checkmark \Rightarrow p(f,t) = \frac{N\mu}{f} \checkmark \quad (f \ll f_{max})$$

What about selected mutations?

$\Rightarrow f_{max}(t) = \frac{e^{st}-1}{2Ns} \xrightarrow{t \ll 1/|s|} \frac{t}{2N}$ $\Rightarrow$ just like single trajectory, indistinguishable from neutral mutations when $t \ll 1/|s|$

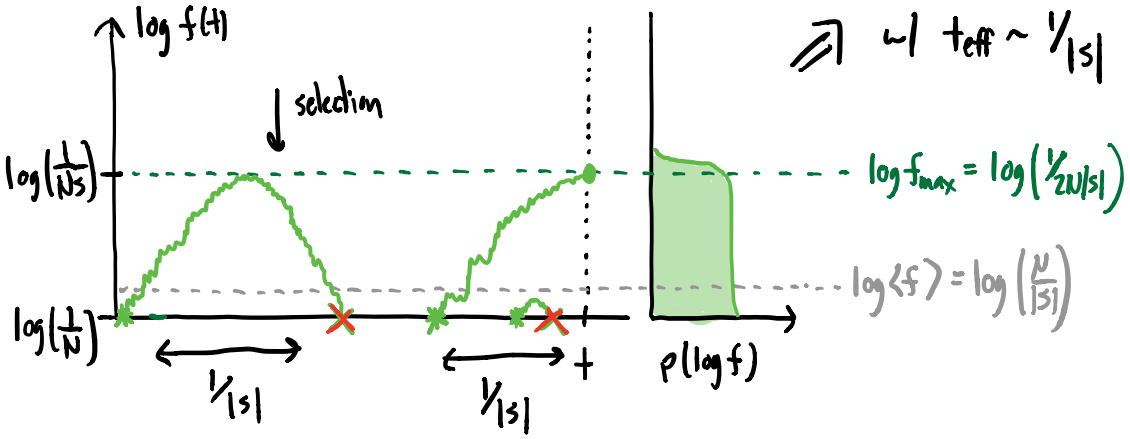
$\Rightarrow$ At longer times, strongly depends on sign of s...

(1) For deleterious mutations: $f_{max}(t) \xrightarrow{t \gg 1/|s|} 1/2N|s|$ ← independent of time!

$\Rightarrow$ $\boxed{p(f,t) \approx 2N\mu f^{-1} e^{-2N|s|f}}$

Intuition: "mostly $f=0$, but small probability ($\sim N\mu$) of growing as large as $f_{max} \sim 1/N|s|$"

$\Rightarrow$ can again visualize with paths:



$\Rightarrow$ Typical frequencies very different from avg!

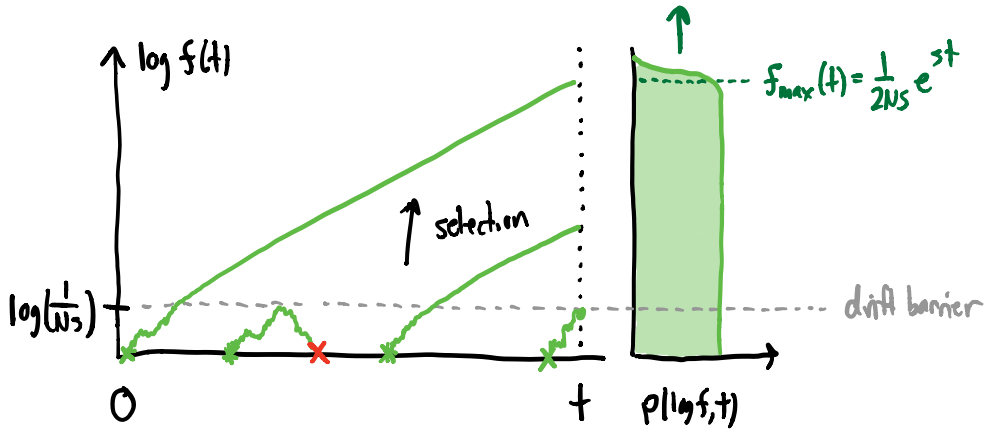
e.g. ABX resistance mut w/ cost $|s| \sim 10^{-2}$ in absence of drug.

w/ $\mu \sim 10^{-10}$, $N \sim 10^{6}$

$\Rightarrow N \cdot \frac{\mu}{|s|} \approx 10^{-2}$ cells w/ mutation ≈ 0 $\Rightarrow$ but $\frac{1}{2N|s|} \cdot N \approx 100$ cells!

(2) For beneficial mutations, $f_{max}(t) \xrightarrow{t \gg 1/s} \frac{1}{2Ns} e^{st}$ $(\gg 1/Ns)$

can again visualize w/ paths:



⇒ Distribution is still broad, but paths deterministic once $f \gg \frac{1}{Ns}$

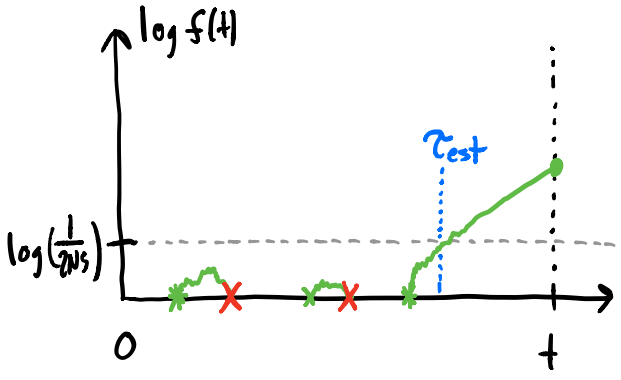
⇒ can again try to catch randomness in single random #:

$$f(t) \equiv \nu(t)\, e^{st}$$ (just like for $\mu = 0$ case)

⇒ find that $\nu(t) \xrightarrow{t \gg 1/s} \nu \sim \text{Gamma}\left(2N\mu, \underbrace{\frac{1}{2Ns}}_{\nu_{max}}\right)$

independent of time!

$\Rightarrow$ also useful to rewrite ν as a time, $f(t) \equiv \frac{1}{2Ns} e^{s(t-\tau_{est})}$



$\Rightarrow \tau_{est} \equiv \frac{1}{s}\log\left(\frac{1}{2Ns\nu}\right)$

"establishment time"

$\approx$ "when mutation arose & survived drift"

("time that f(t) would have reached $\frac{1}{2Ns}$ if it grew deterministically back in time.")

$\Rightarrow$ when $N\mu \ll 1$ $\Rightarrow \tau_{est} \sim \underbrace{\text{Exponential}\left(\frac{1}{2N\mu s}\right)}_{\text{randomness in when successful mutation arose}} \pm \underbrace{O\left(\frac{1}{s}\right)}_{\text{randomness in path from } \frac{1}{N} \to \frac{1}{Ns}}$

$\Rightarrow$ interpretation: (1) mutations arise @ rate $N\cdot\mu$ per gen

(2) survive drift ("establish") w/ prob $\sim s$

$\Rightarrow$ successful mutations occur as Poisson process w/ rate $N\cdot\mu\cdot s$

$\Rightarrow$ similarly, for $t_{1/2}$: $\quad f(t_{1/2}) = \dfrac{\frac{1}{2Ns} e^{s(t_{1/2} - \tau_{est})}}{\frac{1}{2Ns} e^{s(t_{1/2} - \tau_{est})} + 1} = \dfrac{1}{2}$

$\Rightarrow \quad t_{1/2} = \underbrace{\frac{1}{s} \log(Ns)}_{\text{from single mutation trajectory}} + \underbrace{\tau_{est}}_{\text{from when mutation occurred}} \xrightarrow{N\mu \to 0} \frac{c}{2N\mu s} \sim \text{Exp}(1)$

"limited by supply of new mutations"

↳ e.g. increase N or increase μ by const factor
$\Rightarrow$ decrease $t_{1/2}$ by the same amount.

$\Rightarrow$ compare to $N\mu \gg 1$ case: $\quad f(t) \approx \langle f(t) \rangle = \frac{\mu}{s} e^{st}$

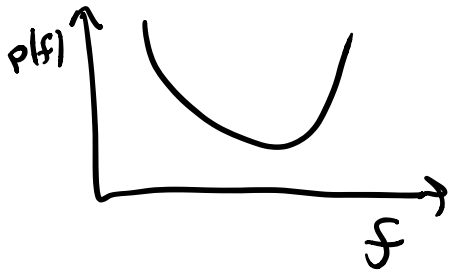
$\Rightarrow$ if try to set $\quad f(t) \equiv \frac{1}{2Ns} e^{s(t - \tau_{est})}$ ← set equal

$\Rightarrow \quad \tau_{est} = -\frac{1}{s} \log(N\mu)$ ← large & negative (+ deterministic)

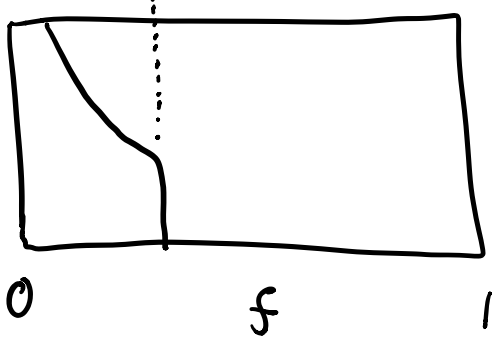
$\Rightarrow \quad t_{1/2} \approx \frac{1}{s} \log\left(\frac{s}{\mu}\right)$ ← independent of N
↖ weakly dependent on N

How long does it take for $p(f) \propto f^{2Nu-1}(1-f)^{2Nu-1}$ to equilibrate.

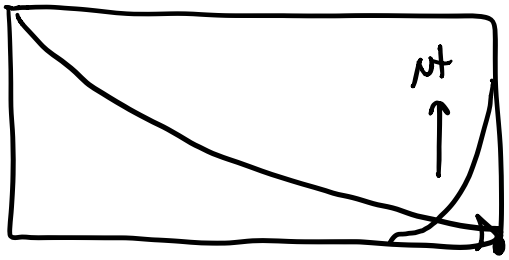
($s=0$)



① $f_{max} = t/N \longrightarrow$ as $t \to \infty$



* need $t \sim N$ for left "L" to form from mutations from $f=0$

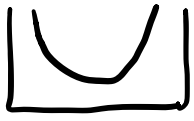
② 

New chance for right half of u-shape to form from back mutations from $f=1$ state.

$\Rightarrow$ height of right half is smaller than left half.

$\Rightarrow$ Rate that mutations reach $f=1$ is

$$Nu \times \left(\frac{1}{N}\right) = u$$

$\Rightarrow$ need $\sim 1/u$ generations to reach 

$\Rightarrow$ $t_{eq} \sim 1/\nu$ e.g. humans – $\nu \sim 10^{-8}$

gen ~ 20 yrs

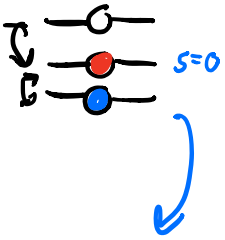
$\Rightarrow$ $t_{eq} \sim 2 \times 10^{9}$ years!

$\Rightarrow$ not enough time to equilibrate!

$\Rightarrow$ Instead, more useful is quasi-stationary dist'n

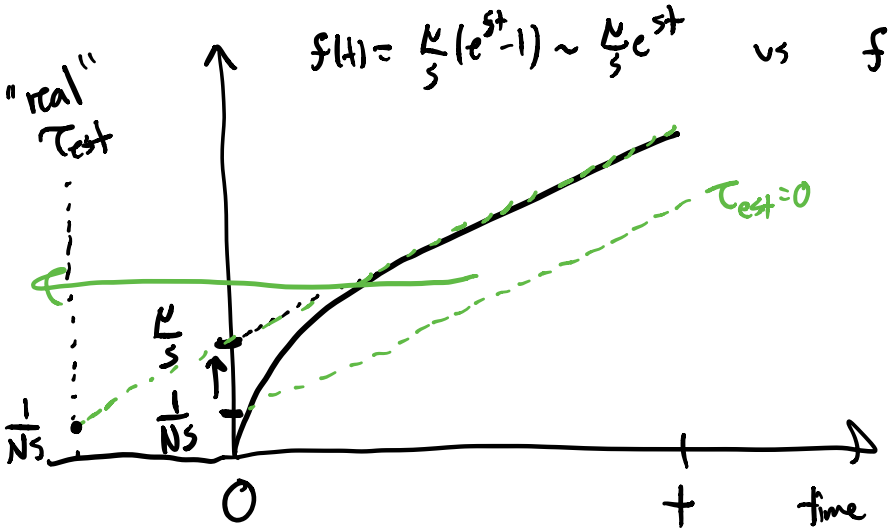
$$p(f) \sim \frac{2N_e\nu}{f}$$ (valid for $s=0$, $t \gg N$, but $t \ll 1/\nu$)

(will see more of this in the coming weeks...)



compare to strong selection: $p(f) = \frac{2N_e\nu}{f} e^{-2N|s|f}$

Question: why is τ_{est} negative when $N\mu \gg 1$?



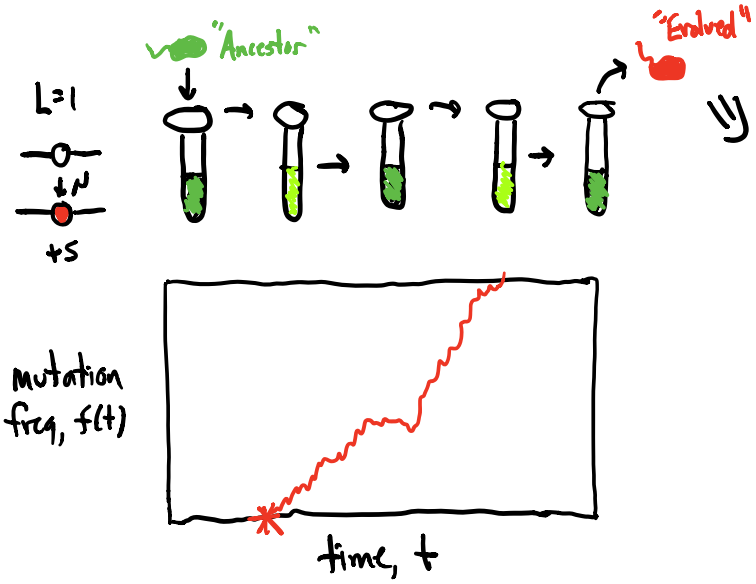
$f(t) = \frac{\mu}{s}(e^{st}-1) \sim \frac{\mu}{s}e^{st}$ vs $f(t) \equiv \frac{1}{Ns}e^{s(t-\tau)}$

How many mutations effectively contribute?

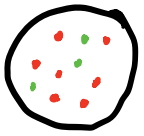
$$\frac{\left(\frac{\mu}{s}\right)}{\left(\frac{1}{Ns}\right)} = N\mu$$

DNA Sequencing & Genomics

So far....



Qualitative phenotype (e.g. red vs green) that could be detected by colony counting

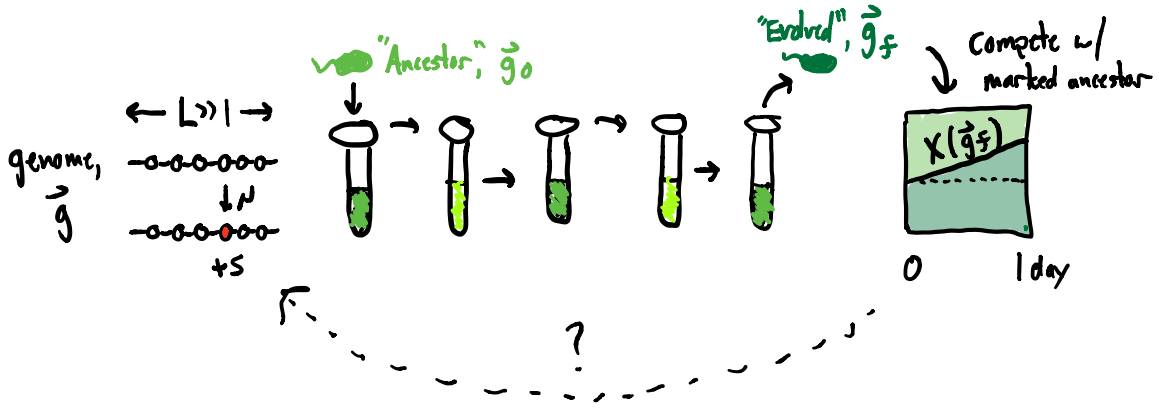


$\Rightarrow$ In practice, genomes contain many sites

$\Rightarrow$ don't know what phenotypes mutations @ these sites produce or how to measure them w/ colony counting assay...

$L \sim 10^4\text{-}10^5$ for viruses
$L \sim 10^6\text{-}10^7$ bacteria
$L \sim 10^9$ for humans

Historically, experimental evolution relied on <u>competitive fitness</u>

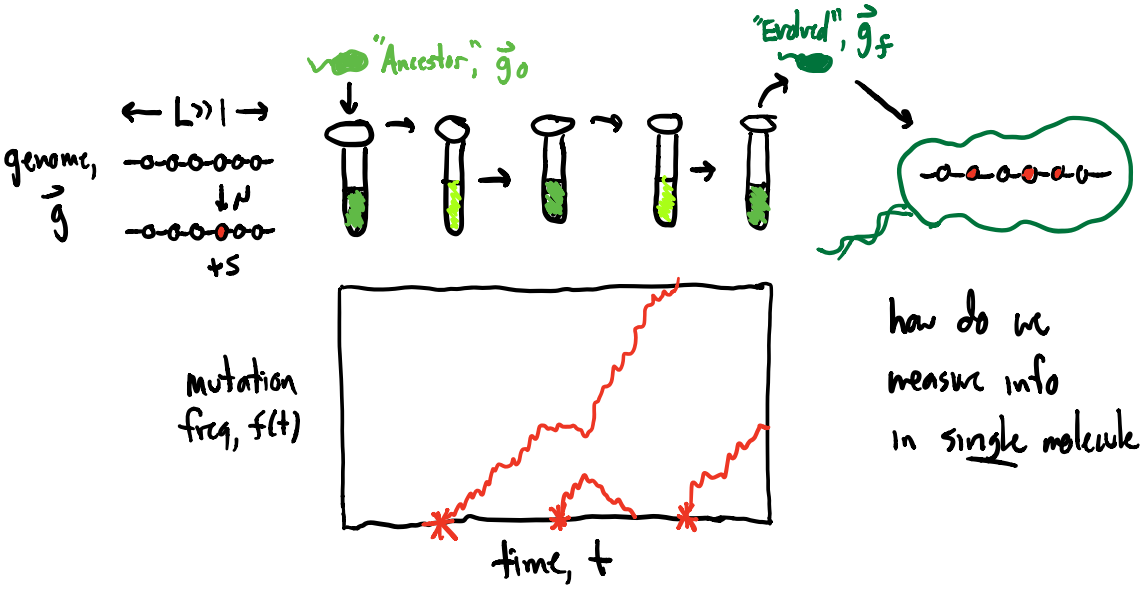


$\Rightarrow$ statistics of $X(\vec{g})$ w/in & between populations tell us something about evolutionary dynamics of $\vec{g}$

$\Rightarrow$ <u>downside</u>: indirect! many diff dynamics of $\vec{g}$ consistent w/ same dynamics of $X(\vec{g})$...

+ $\vec{g} \rightarrow X(\vec{g})$ poorly understood...

Now: DNA sequencing allows us to measure genomes directly*

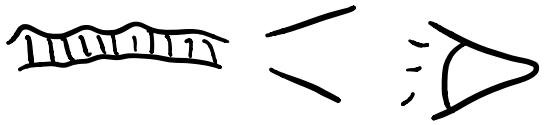


how do we measure info in single molecule?

Recall: genome = ATCGTA... / TA... (complementary strands)

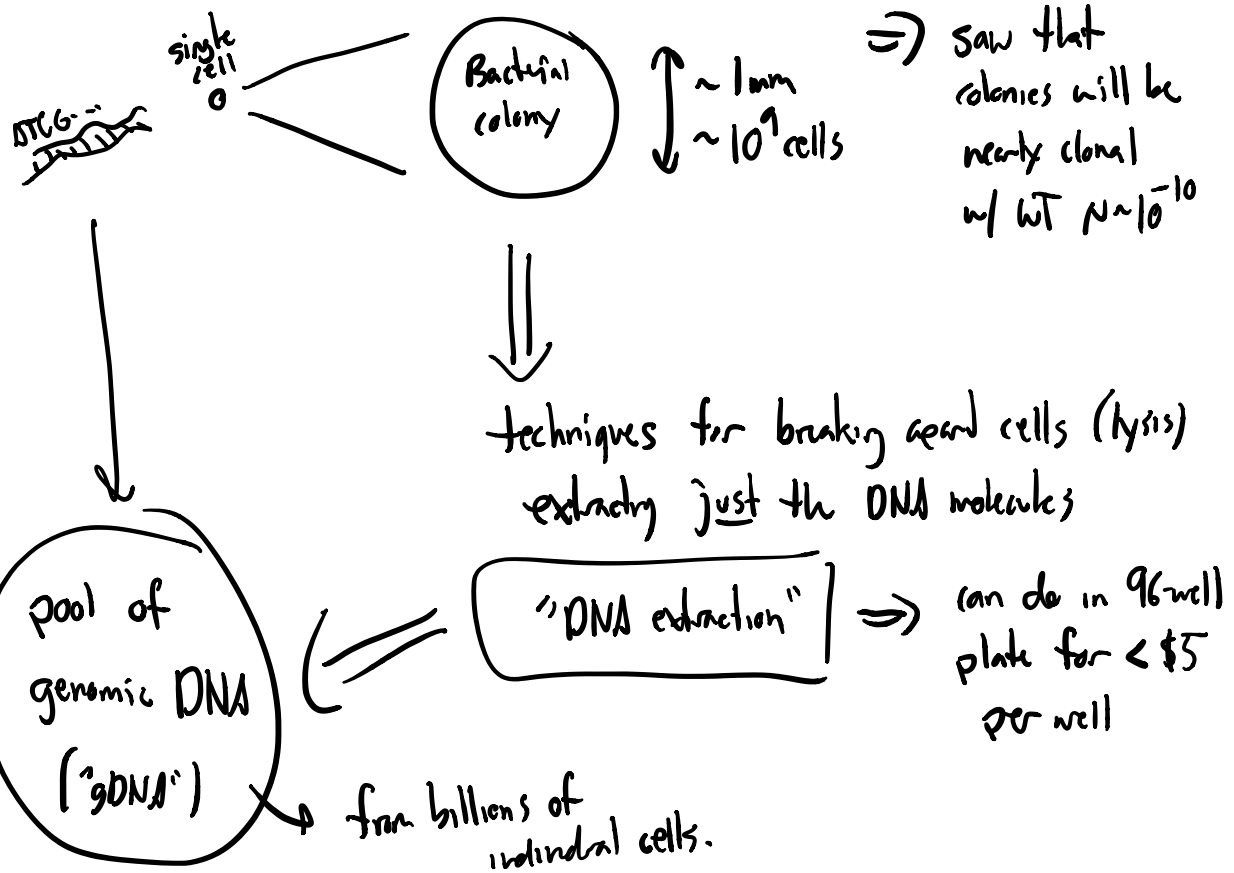
$L \sim 10^{4-5}$ virus
$L \sim 10^{6}$ bacteria
$L \sim 10^{9}$ humans.

Step 1 for reading genomes: amplification!



need macroscopic quantities of our DNA molecule to work with.

For bacteria: easy! use built in ability to grow exp.



⇒ <u>Next time</u>: how do we read these out?