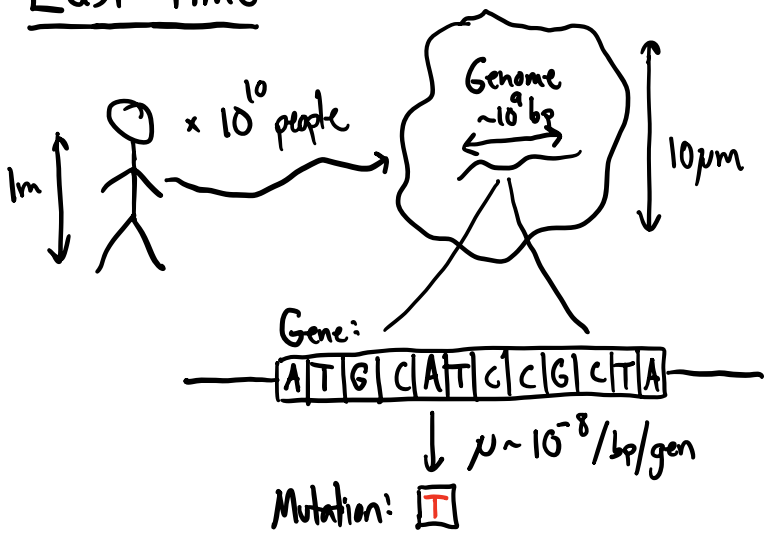
Chapter 4

A Simple Model of Evolution

Last time:



"Fermi problem" (mutation supply)

(# individuals in population) × (Pr[mutation] per site per generation) = (# new mutations produced in pop'n per site per generation)

E.g. Humans: $N \sim 10^{10}$ × $\mu \sim 10^{-8}$ = ~ 100 / bp / gen

Empirical observation:

Avg # differences between my genome and yours is $\sim 10^{-3}$ /bp

How do we connect these 2 observations?

↓

Evolutionary dynamics!

Today: A Simple Model of Evolution

⇒ Traditionally: start w/ abstract math model
(e.g. "balls & urns" in pop gen, 1920's)

⇒ Here, we'll take a different approach:
base our model on experiments we can do in lab

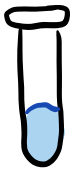
Payoff: will enable operational definitions for quantities that can be difficult to interpret...
(e.g. "fitness" / "genetic drift")

+ keep us grounded in some concrete data...

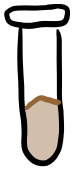
① Need a population of organisms (ideally, small & fast growing)

⇒ model microorganisms (e.g. E. coli) grown in lab

can prepare growth "media"

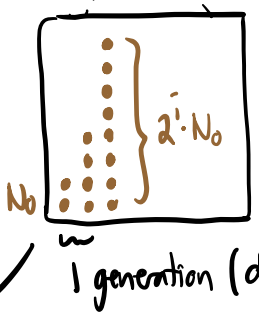


+ inoculate w/ lab strain of E.coli (N_0 cells) + wait 24 hrs →



$N_f \sim 10^7 - 10^9$ cells

[carbon source, e.g. sugar
salts (N, P, S, etc.)
some vitamins]

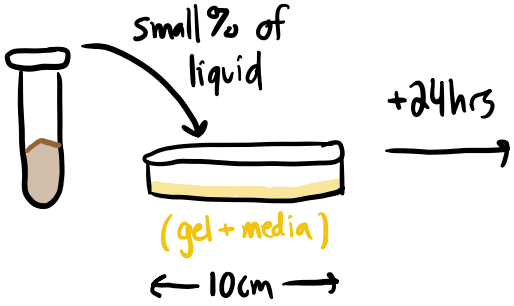


~ 1 generation (doubling)

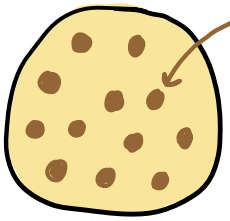
$$\#\ \text{generations} = \log_2\left(\frac{N_f}{N_0}\right)$$

How can we measure N_0 & N_f? [in principal hard b.c. must count large #'s of microscopic things...]

(i) Old fashioned way: dilute & grow on plates ("Petri dish")



+24hrs →



macroscopic "colonies" (each founded by single cell)

(~100's colonies/plate)

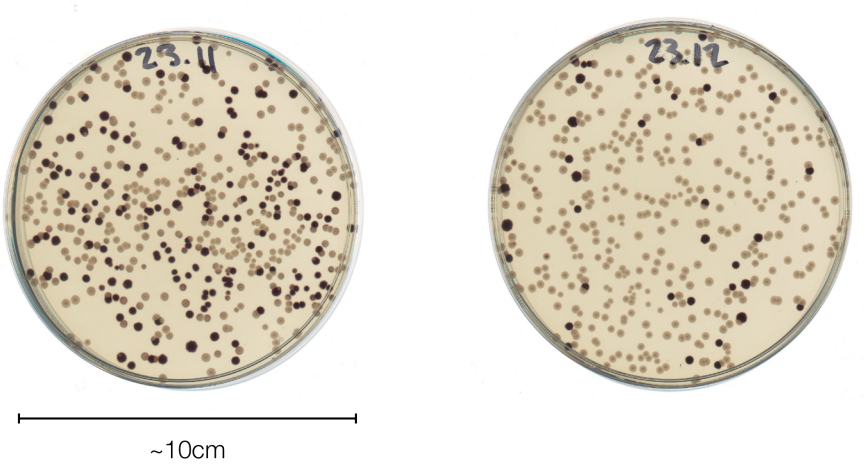
$$\Rightarrow \text{observed \# colonies on plate} \sim \text{Poisson}\left(N_f \times \frac{V_{spread}}{V_{tot}} \times \text{plating efficiency}, p\right)$$

(can measure)

($\frac{V_{spread}}{V_{tot}}$: dilution factor, can measure)

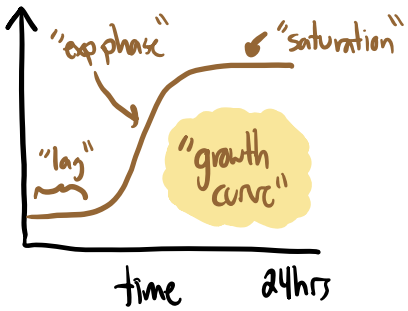
$\Rightarrow$ can infer $N_f \cdot p$ (colony forming units / CFUs)

Example data:
(2 different color colonies)



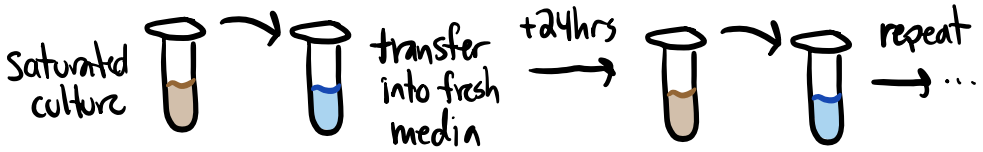
(ii) More modern method: measure "optical density" / "OD"
(e.g. w/ lasers)

$\Rightarrow$ $OD \equiv -\log_{10}\left(\frac{I_{out}}{I_{in}}\right)$



(can measure ~100 curves in parallel)

(2) **Basic idea of experimental evolution:**



"serial dilution"

⇒ For simplicity, imagine following scenario:

(i) Start w/ N_0 cells & grow for fixed time Δt

$$\Rightarrow N(t) = N_0 e^{rt} \Rightarrow N_f = N_0 e^{r\Delta t}$$

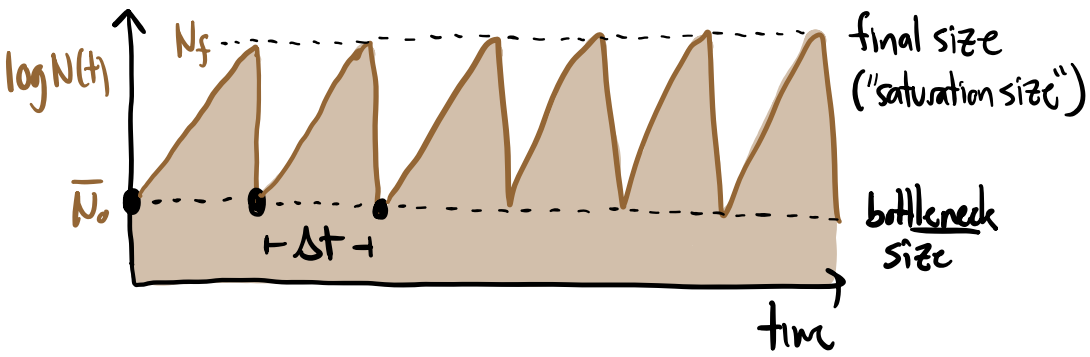
r: "growth rate" ($\approx \log(2)$ if Δt measured in gens)

[technically, assumes that $\Delta t \ll$ time where cells deplete media... can always do this in theory - though in practice we often don't]

(ii) Measure N_f @ time Δt, choose dilution factor such that expect $\bar{N}_0$ cells in fresh tube

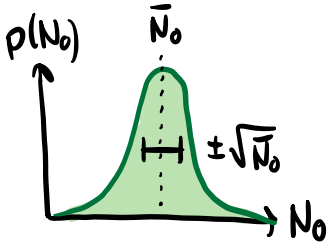
$$\Rightarrow N_0(k+1) \sim \text{Poisson}(\bar{N}_0) \equiv \#\text{ cells in fresh tube @ beginning of day } k+1$$

(iii) Repeat steps i & ii over & over...



$\Rightarrow$ # gens / cycle $= \log_2\left(\frac{N_f}{N_0}\right)$ "dilution factor"

$\Rightarrow$ # cells @ bottleneck (N_0) is stochastic



"Case 1" dist'n (fuzzy noise)

$N_0 \approx \bar{N}_0 \pm \sqrt{\bar{N}_0}$

$\Rightarrow$ avg is good guess: $\bar{N}_0 \approx 10^3 \Rightarrow N_0 \approx 10^3 \pm 30$ (3% error)

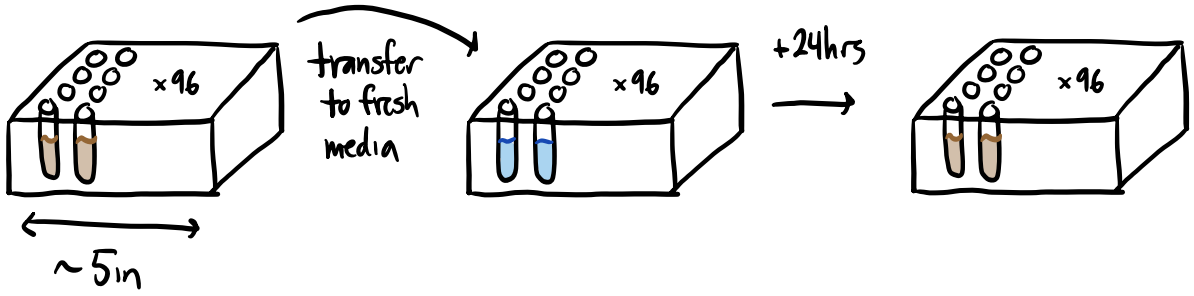
$\Rightarrow$ Example dilution factors:

100-fold dilution $\Rightarrow$ 6.6 gens/day $\Rightarrow$ 100 gens in ~2 weeks

1000-fold " $\Rightarrow$ 10 gens/day

$\Rightarrow$ if $N_0 \sim 10^6$ cells $\Rightarrow$ $N_f \sim 10^8 - 10^9$ cells (~1 mL)

$\Rightarrow$ not just test tubes... can also grow in "96-well plates"



How do we think about evolution in this scenario?

let's imagine mixing 2 E. coli strains together in 50-50 ratio

Strain 1: normal lab strain (WT)

Strain 2: some gene deleted (e.g. can't grow on fancy sugar X that's not in growth media... → "ΔsugarX")

(e.g. resistance to ABX Y)

⇒ Now 2 #'s to keep track of: $N_1(t)$, $N_2(t)$

or:

Total Pop'n Size: $N_{tot}(t) \equiv N_1(t) + N_2(t)$

&

Relative frequency: $f(t) \equiv N_2(t) / N_{tot}(t)$

How do they change over time?

⇒ suppose ΔsugarX frees up resources (e.g. for ribosomes)

⇒ strain 2 grows slightly faster in growth media:

⇒ $N_1(t) = N_1(0)e^{rt}$, $N_2(t) = N_2(0)e^{(r+s)t}$

s: some empirical param $s > 0$

$\Rightarrow$ if freq @ beginning of day is $f(0)$, freq @ end of day is:

$$f(\Delta t) \equiv \frac{N_2(\Delta t)}{N_1(\Delta t) + N_2(\Delta t)} = \frac{N_0 f(0) e^{(r+s)t}}{N_0(1-f)e^{rt} + N_0 f e^{(r+s)t}} = \frac{f(0)\, e^{s\Delta t}}{1 - f(0) + f(0)\, e^{s\Delta t}}$$

$\Rightarrow$ # cells of each type transferred to next day's flask:

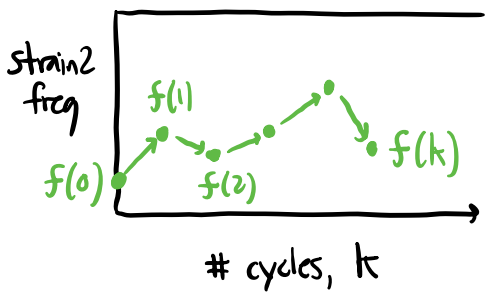
$$N_2(k+1) \sim \text{Poisson}\left(\bar{N}_0 \cdot \frac{f(k)\, e^{s\Delta t}}{1 - f(k) + f(k)\, e^{s\Delta t}}\right)$$

$$N_1(k+1) \sim \text{Poisson}\left(\bar{N}_0 \cdot \frac{1 - f(k)}{1 - f(k) + f(k)\, e^{s\Delta t}}\right)$$

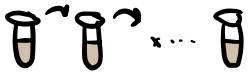
$\Rightarrow$ New freq. $f(k+1) \equiv \dfrac{N_2(k+1)}{N_1(k+1) + N_2(k+1)}$

$\Rightarrow$ repeat to generate sequence of freqs,

$f(0) \rightarrow f(1) \rightarrow f(2) \rightarrow \cdots f(k)$ ("Markov process")



$\Leftrightarrow$ "simple model of evolution"



Simplest Case: $s = 0$ (no growth rate diffs, "neutrality")

$\Rightarrow$ model reduces to:

$$N_2(k+1) \sim \text{Poisson}\left(N_0 f(k)\right)$$
$$N_1(k+1) \sim \text{Poisson}\left(N_0 \cdot (1 - f(k))\right)$$

$\Rightarrow$ can derive some basic properties:

e.g. conditional mean (i.e. known value of $f(k)$)

$$E\left[f(k+1) \,\middle|\, f(k)\right] \equiv \sum_{f(k+1)} f(k+1) \cdot p\left(f(k+1) \,\middle|\, f(k)\right) = f(k)$$

↓ due to symmetry (exchangeability)

$\Rightarrow$ unconditional mean:

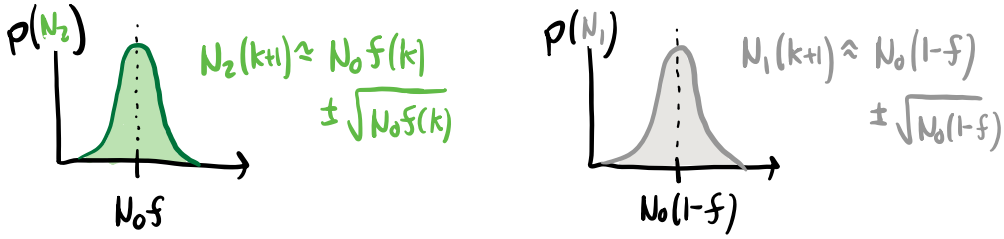
$$E\left[f(k+1)\right] \equiv \sum_{f(k)} \underbrace{E\left[f(k+1) \,\middle|\, f(k)\right]}_{f(k)} p\left(f(k)\right) = E\left[f(k)\right]$$

$$\Rightarrow E\left[f(k)\right] = E\left[f(k-1)\right] = \cdots = E\left(f(0)\right) \equiv f_0$$

i.e. average is constant in time!

⇒ in practice, fluctuations around avg value

⇒ if $N_0 f(k) \cdot N_0(1-f(k)) \gg 1$ ⇒ "case 1" noise:



⇒ New frequency:

$$f(k+1) = \frac{N_0 f \pm \sqrt{N_0 f}}{N_0 f \pm \sqrt{N_0 f} + N_0(1-f) \pm \sqrt{N_0(1-f)}} = \frac{f \pm \sqrt{\frac{f}{N_0}}}{1 \pm \sqrt{\frac{f}{N_0}} \pm \sqrt{\frac{1-f}{N_0}}}$$

Taylor expand for large N_0

$$\approx f(k) \pm \mathcal{O}\left(\frac{1}{\sqrt{N_0}}\right)$$

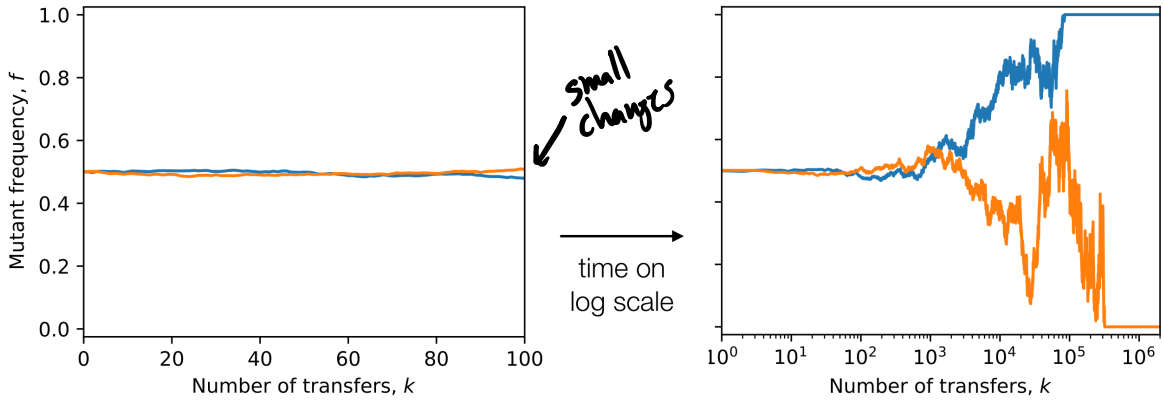
"genetic drift"

⇒ if N_0 is large ⇒ genetic drift is pretty small!

e.g. $N_0 \sim 10^5$ cells ⇒ $\frac{1}{\sqrt{N_0}} \sim 0.3\%$

⇒ but it is relentless! (i.e. compounds over time)

Computer simulations of model with $s = 0$, $N_0 = 10^5$, $f(0) = 50\%$



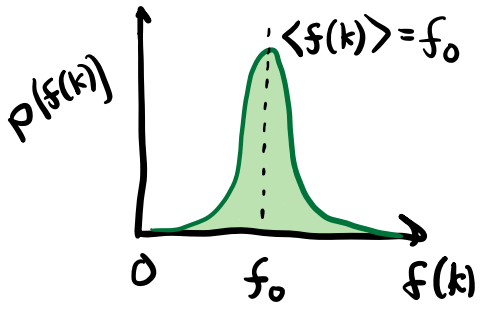
⇒ in 2nd case, something "singular" happens:

① if $f=0$ @ one time ⇒ $f=0$ @ all later times

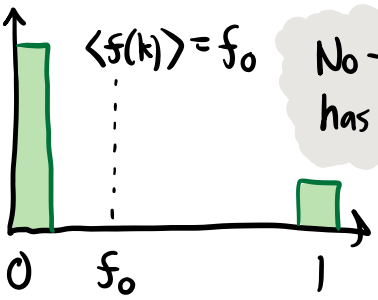
② if $f=1$ @ " ⇒ $f=1$ @ all "

- "fixation"
- "extinction"

⇒ Short times ("case1")



Long times ("case2")



No trajectory has $f \approx \langle f \rangle$!

⇒ Instead, avg is mixture of 2 outcomes:

$$\langle f(\infty) \rangle = 0 \times Pr[f=0] + 1 \times Pr[f=1] = f_0$$

from neutrality

⇒ $Pr(f=1) = f_0$

"Fixation probability" of neutral mutation

⇒ but timescale required is quite long...

⇒ will show for short times: $f(k) \approx f_0 \pm O\left(\sqrt{\frac{k}{N_0}}\right)$

"random walk"

⇒ need $k \sim N_0$ before we can start to think about fixation

⇒ e.g. $N_0 \sim 10^5$ cells ⇒ 10^5 days ⇒ 300 yrs!

⇒ Upshot: genetic drift is very weak on lab timescales* (*for mutations @ 50% frequency)

⇒ selection will often be more important

Natural selection and "fitness"

Now consider $s \neq 0$. (For simplicity, assume $N_0 = \infty$ i.e. no diff for now...)

$$\Rightarrow f(k) = \frac{f(k-1)e^{s\Delta t}}{1-f(k-1)+f(k-1)e^{s\Delta t}} = \frac{\frac{f(k-2)e^{s\Delta t}}{1-f(k-2)+f(k-2)e^{s\Delta t}} \cdot e^{s\Delta t}}{\frac{1-f(k-2)}{1-f(k-2)+f(k-2)e^{s\Delta t}} + \frac{f(k-2)e^{s\Delta t}}{1-f(k-2)+f(k-2)e^{s\Delta t}} \cdot e^{s\Delta t}}$$

$$= \frac{f(0)e^{ks\Delta t}}{1-f(0)+f(0)e^{ks\Delta t}} \leftarrow \frac{f(k-2)e^{2s\Delta t}}{1-f(k-2)+f(k-2)e^{2s\Delta t}}$$

denominators cancel...

$\Rightarrow$ if measure time in generations, $t \equiv k \cdot \Delta t$,

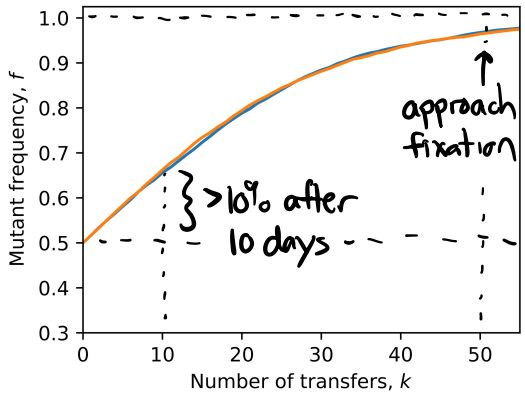
$$f(t) = \frac{f(0)e^{st}}{1-f(0)+f(0)e^{st}} \Leftrightarrow \text{"Logistic growth"} \quad \frac{df}{dt} = sf(1-f)$$

$\Rightarrow$ Now can get a big change:

e.g. if $s = 0.01$,

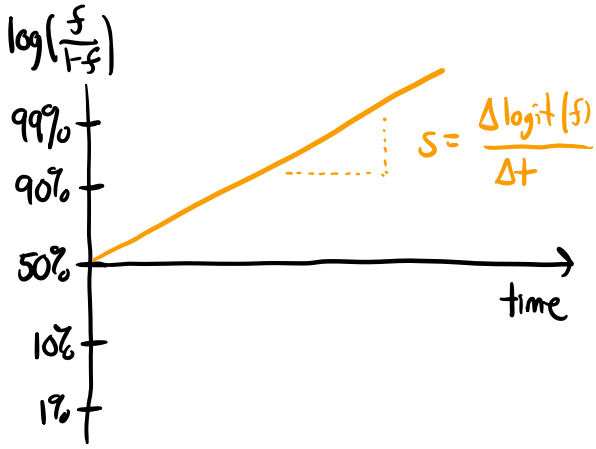
$\Delta t = \log_2(100) \approx 7$

$N_0 \approx 10^5$



⇒ sometimes helpful to plot on "logit" scale:

$$\text{logit}(f) \equiv \log\left(\frac{f}{1-f}\right)$$



⇒ upshot: can notice big change when $st \gtrsim 1$

⇒ $t \geq 1/s$ "selection timescale"

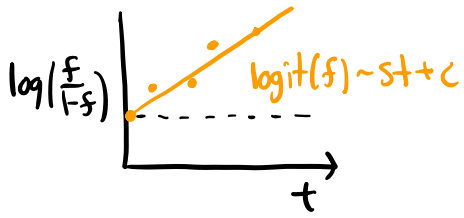
⇒ So far, if know s (e.g. from previous expt's on underlying growth rate, $r \to r+s$)

↳ can predict $f(t)$...

⇒ can also turn around & use as definition of s:

⇒ if measure $f(t)$ ⇒ $s = \frac{1}{t}\log\left(\frac{f(t)}{1-f(t)} \cdot \frac{1-f(0)}{f(0)}\right)$

$s \equiv$ "fitness difference" -or- "competitive fitness"

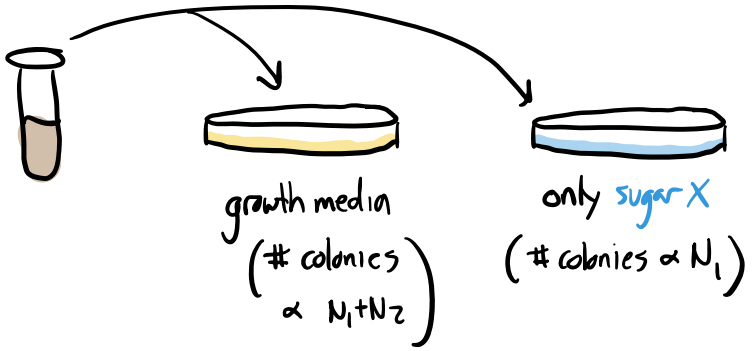


Question: How do we measure $f(t)$?

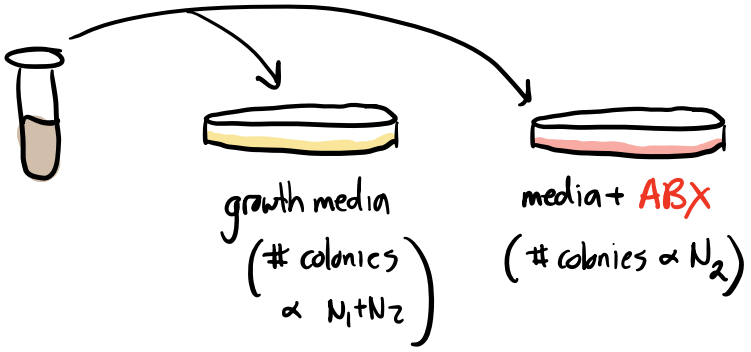
(in principle hard to distinguish similar-looking strains like WT, ΔsugarX...)

(1) Old fashioned way: make them distinguishable & count colonies

e.g. ΔsugarX:



e.g. + ABX resistance:

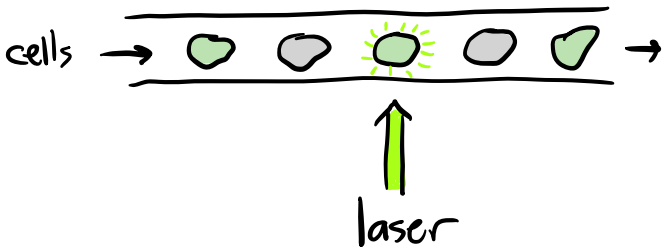


(2) Fluorescence + lasers ("flow cytometry")

1. gene producing fluorescent protein (GFP, RFP, ...)

2. insert into one strain (requires genetic engineering...)

3. can count on "flow cytometer":

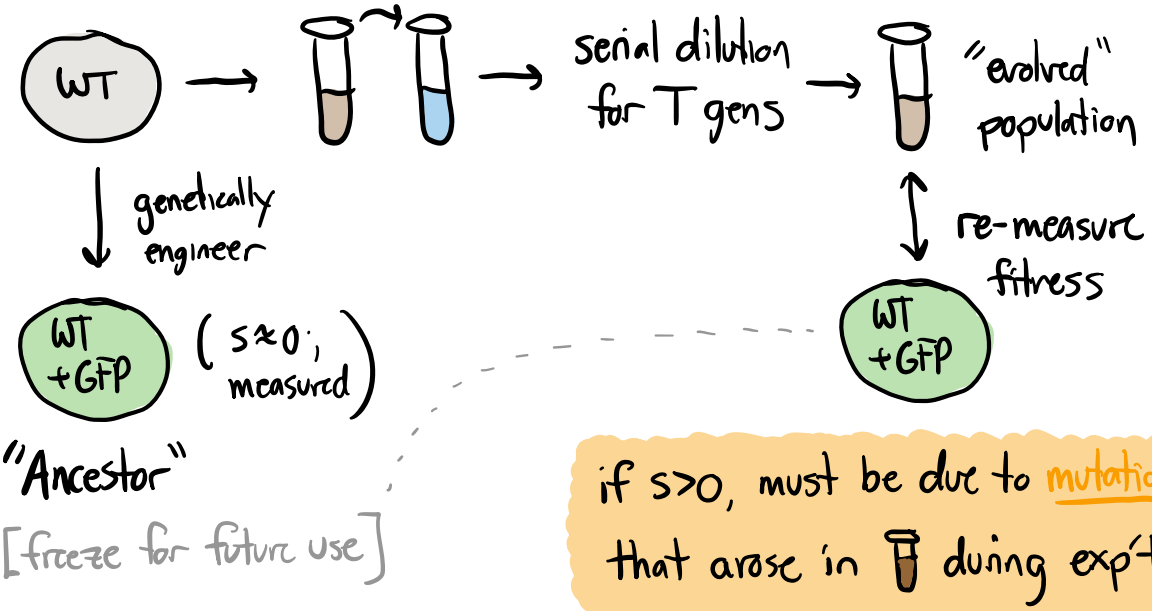


input: 96 well plate
~ 1hr / plate
~ 50,000 cell counts / well

(3) DNA sequencing (will introduce later)

Upshot: now have way of measuring fitness operationally
(mix @ 50-50 and measure short-term $f(t)$)

⇒ Consider following experiment:



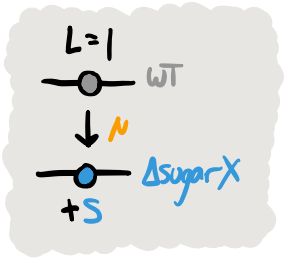
⇒ How do we model these?

Spontaneous mutations

Start w/ simplest case:

1. suppose there is a single target for mutations (e.g. WT → ΔsugarX)
2. mutations happen w/ probability μ per division ($\mu \ll 1$)

⇒ known as a "single locus" model

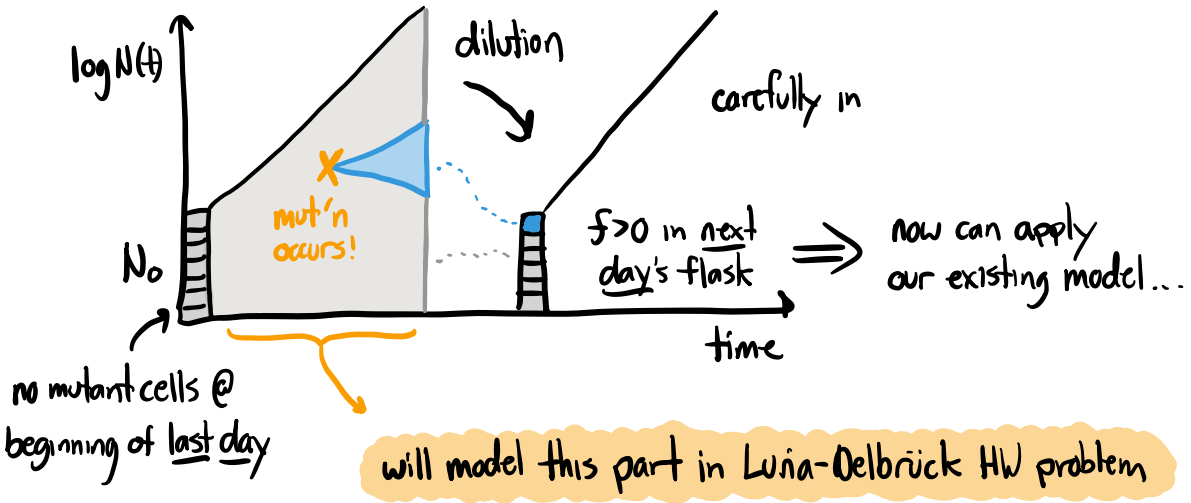


⇒ equivilant to genome w/ a single site

⇒ seems unrealistic... but can learn a lot about evolution by studying this simple case (~ single particle QM)

⇒ will learn how to generalize to bigger genomes later...

⇒ Let's zoom in on cycle where the mutation first occurs:

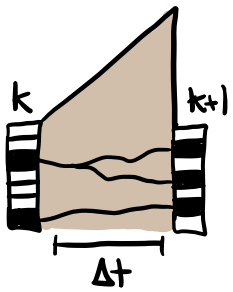


To model this process, assume for simplicity:

① mutation doesn't exert fitness benefit until next day's cycle

[not such a crazy assumption biologically...
e.g. ΔsugarX, need a few divisions to dilute out WT protein]

⇒ Note: every cell @ beginning of next day's flask traces back to ancestor cell alive @ beginning of previous day...



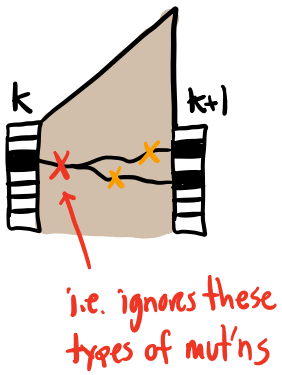
[this is our first example of genealogical thinking, which will be very useful throughout this course!]

⇒ by definition, $\Delta t = \log_2\left(\frac{N_f}{N_0}\right)$ divisions separate them

⇒ $Pr\left[\begin{array}{c}\text{present day cell}\\ \text{has acquired mutation}\end{array}\right] \approx \mu \times \Delta t$

② Approximation is that each cell acquires mutations ≈ independently...

will show in HW that this can be an OK approx when N_0 is large...



Implies that: $N_2(k+1) \sim \text{Poisson}(\bar{N}_0 \cdot \mu \cdot \Delta t)$ new mutants

$N_1(k+1) \sim \text{Poisson}(\bar{N}_0 \cdot (1-\mu \cdot \Delta t))$ everyone who didn't mutate

$$\Rightarrow f(k+1) = \frac{N_2(k+1)}{N_1(k+1)+N_2(k+1)}$$ (then previous dynamics apply...)

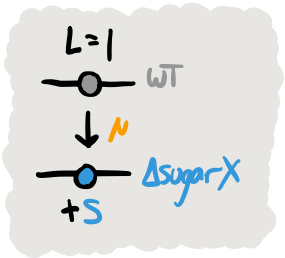
$\Rightarrow$ can combine w/ $f(k)>0$ case to get "full model":

"Microscopic model" of serial dilution:

$$N_2 \sim \text{Poisson}\left(\bar{N}_0 \cdot \frac{f(k)e^{s\Delta t}}{1-f(k)+f(k)e^{s\Delta t}}\right) + \text{Poisson}\left(\bar{N}_0 \mu \Delta t \cdot \frac{1-f(k)}{1-f(k)+f(k)e^{s\Delta t}}\right)$$

$$N_1 \sim \text{Poisson}\left(\bar{N}_0 \cdot (1-\mu\Delta t) \cdot \frac{1-f(k)}{1-f(k)+f(k)e^{s\Delta t}}\right)$$

$$\hookrightarrow f(k+1) = \frac{N_2}{N_2+N_1}$$



$\Rightarrow$ can implement w/ simple computer program (HW)

$\Rightarrow$ can also add "back mutations"  (exercise for reader)