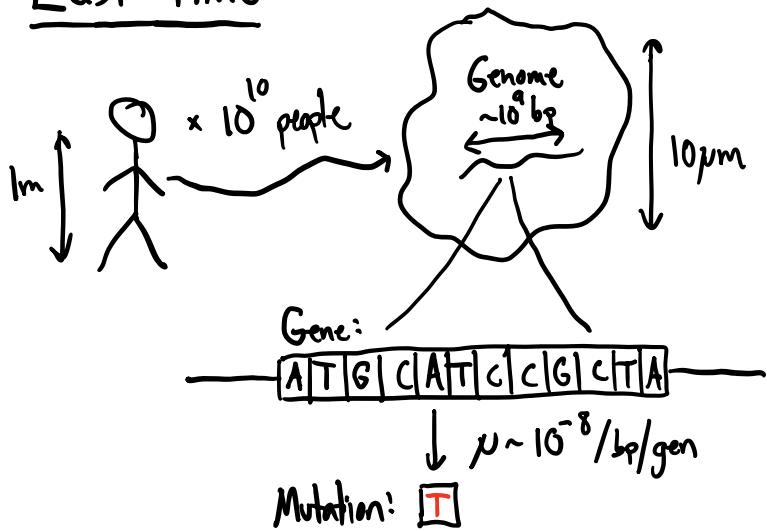


Last time:



"Fermi problem" (mutation supply)

$$\left( \begin{array}{l} \# \text{ individuals} \\ \text{in population} \end{array} \right) \times \left( \begin{array}{l} \text{Pr[mutation]} \\ \text{per site} \\ \text{per generation} \end{array} \right) = \left( \begin{array}{l} \# \text{ new mutations produced in pop'n} \\ \text{per site per generation} \end{array} \right)$$

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

Empirical observation:

Avg # differences between  
my genome and yours is

$\sim 10^{-3} / \text{bp}$

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)

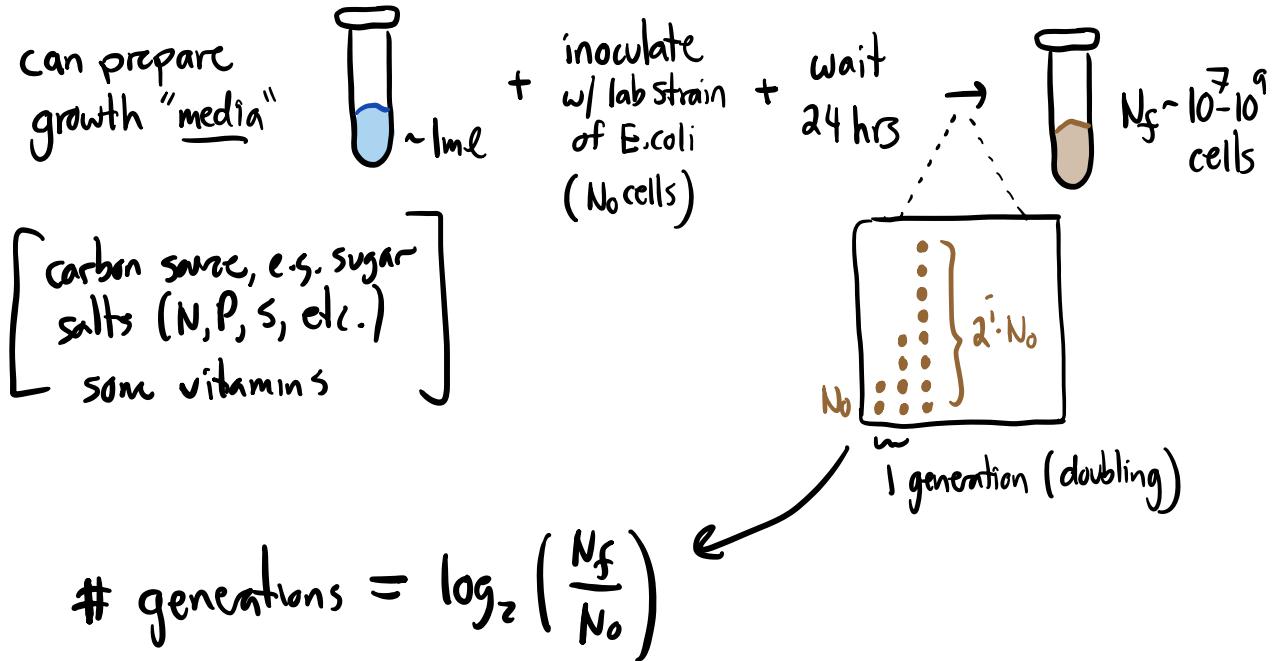
⇒ Here, we'll take a different approach:  
base our model on experiments we can do in lab

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...

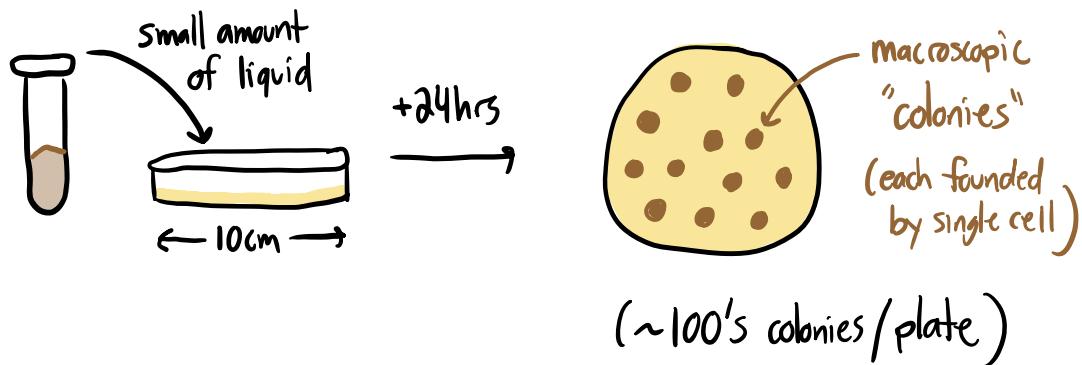
① Need a population of organisms (ideally, small  
& fast growing)

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



How can we measure  $N_0$  &  $N_f$ ?

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



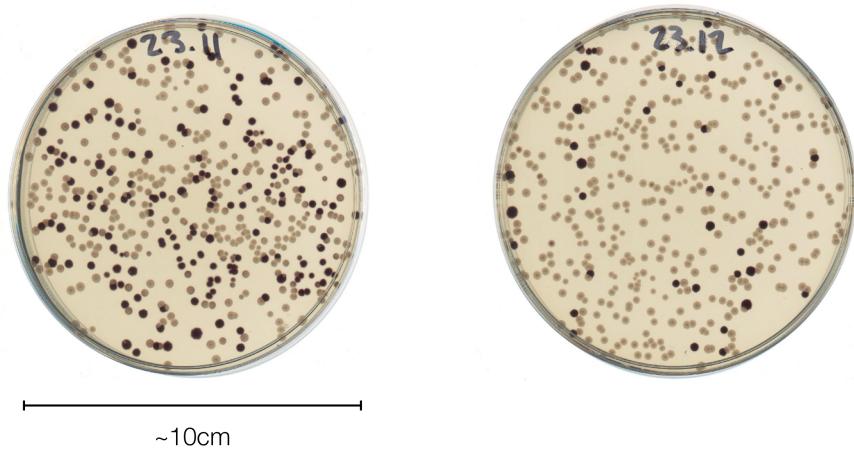
$$\Rightarrow \begin{matrix} \text{observed} \\ \# \text{ colonies} \\ \text{on plate} \end{matrix} \sim \text{Poisson} \left( N_f \times \frac{V_{\text{spread}}}{V_{\text{tot}}} \times \text{plating efficiency, } p \right)$$

(can measure) ↑ (can measure)

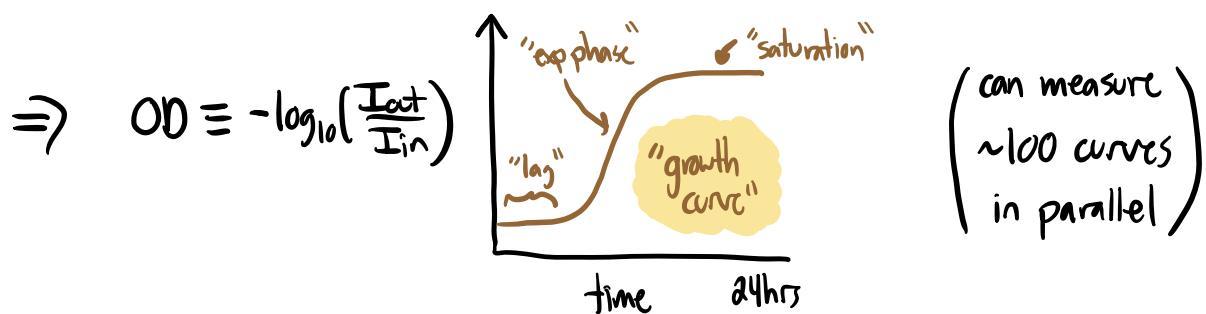
$\Rightarrow$  can infer  $N_f \cdot p$  (colony forming units / CFUs)

Example data:

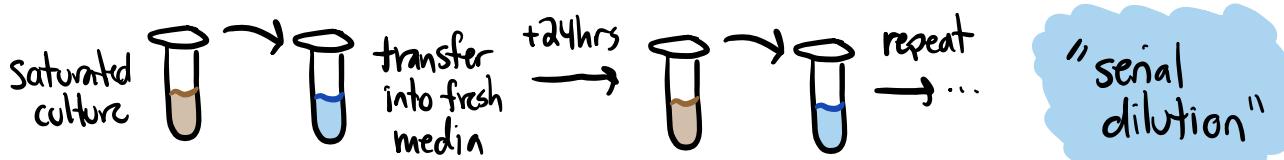
(2 different color colonies)



② More modern method: measure "optical density" / "OD"  
(e.g. w/ lasers)



② Basic idea of experimental evolution:



$\Rightarrow$  For simplicity, imagine following scenario:

- i Start w/  $N_0$  cells & grow for fixed time  $\Delta t$

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

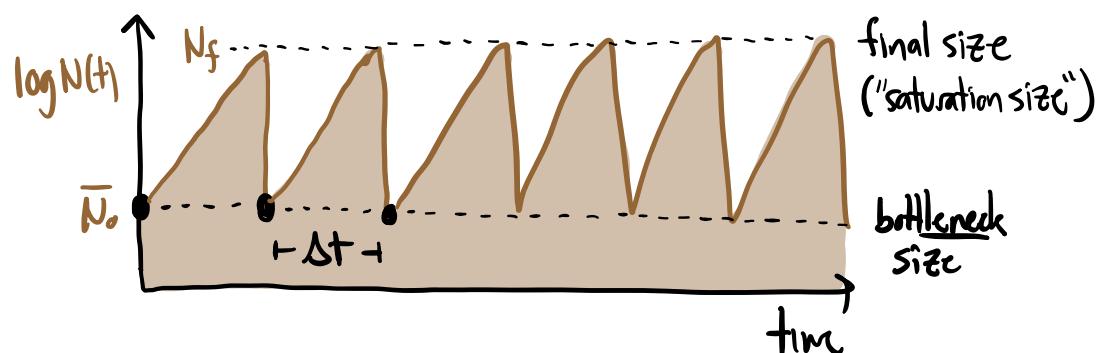
"growth rate" ( $\approx \log(2)$  if  $\Delta t$  measured in gens)

[ 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  $\Delta t$ , choose dilution factor such that expect  $\bar{N}_0$  cells in fresh tube

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

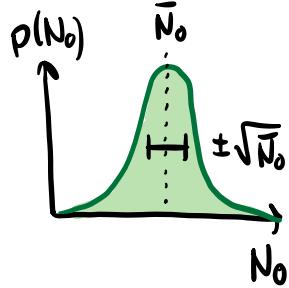
- iii Repeat steps i + ii over & over...



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

"dilution factor"

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



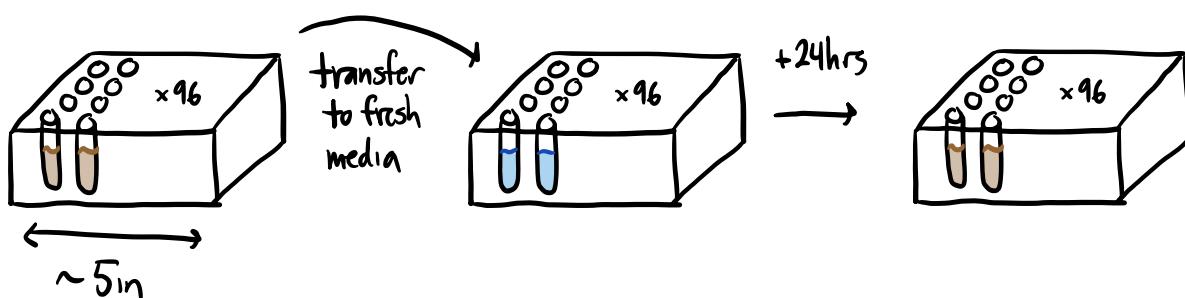
"Case 1" dist'n  
(fuzzy noise)

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

e.g. 100-fold dilution  $\Rightarrow$  6.6 gens/day  $\Rightarrow$  100 gens in  
 $\sim 2$  weeks  
 1000-fold "  $\Rightarrow$  10 gens/day

$\Rightarrow$  if  $N_0 \sim 10^6$  cells  $\Rightarrow N_f \sim 10^8 - 10^9$  cells ( $\sim 1\text{ml}$ )

$\Rightarrow$  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)

"Δsugar X"

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

(e.g. resistance to ABX Y)

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

or:

Total Pop'n Size

$$N_{\text{tot}}(t) \equiv N_1(t) + N_2(t)$$

Relative frequency

$$f(t) \equiv N_2(t) / N_{\text{tot}}(t)$$

How do they change over time?

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

⇒ strain 2 grows slightly faster in growth media:

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

some empirical param  $s > 0$

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

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

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

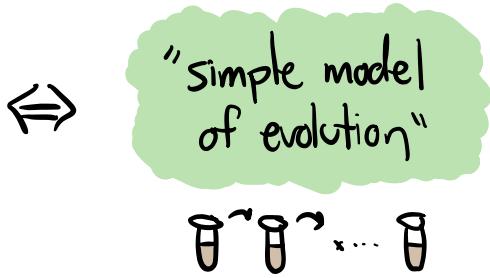
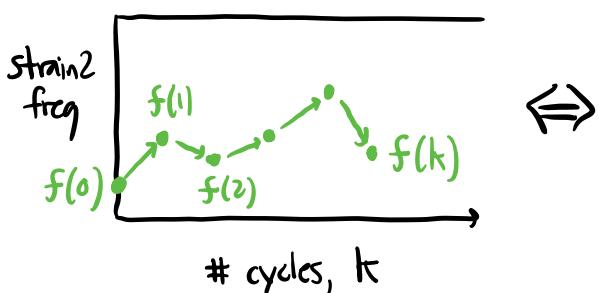
$$N_2(k+1) \sim \text{Poisson}\left(\bar{N}_0 \cdot \frac{f(k) e^{s\Delta t}}{1 - f(k) + f(k) e^{s\Delta t}}\right)$$

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

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

$\Rightarrow$  repeat to generate sequence of freqs,

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



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

$\Rightarrow$  model reduces to:  $N_2(k+1) \sim \text{Poisson}(N_0 f(k))$

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

$\Rightarrow$  can derive some basic properties:

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

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

due to symmetry  
(exchangeability)

$\Rightarrow$  unconditional mean:

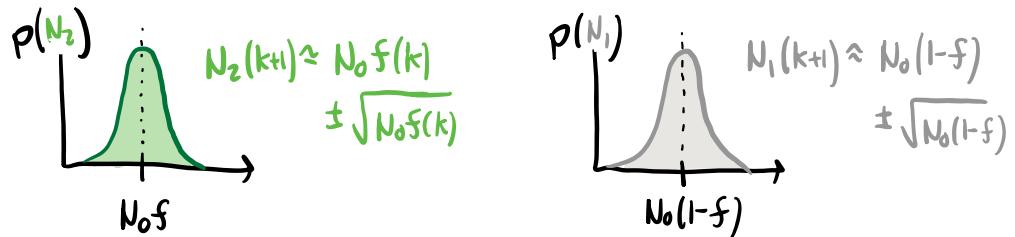
$$E[f(k+1)] \equiv \sum_{f(k)} \underbrace{E[f(k+1) | f(k)]}_{f(k)} p(f(k)) = E[f(k)]$$

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

i.e. average is constant in time!

$\Rightarrow$  in practice, fluctuations around avg value

$\Rightarrow$  if  $N_0 f(k) \gg 1$  &  $N_0(1-f(k)) \gg 1 \Rightarrow$  "case 1" noise:



$\Rightarrow$  New frequency:

$$f(k+1) = \frac{N_0f \pm \sqrt{N_0f}}{N_0f \pm \sqrt{N_0f} + N_0(1-f) \pm \sqrt{N_0(1-f)}} = \frac{f \pm \sqrt{\frac{f}{N_0}}}{1 \pm \sqrt{\frac{f}{N_0}} \pm \sqrt{\frac{1-f}{N_0}}}$$

↓

Taylor expand  
for large  $N_0$

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

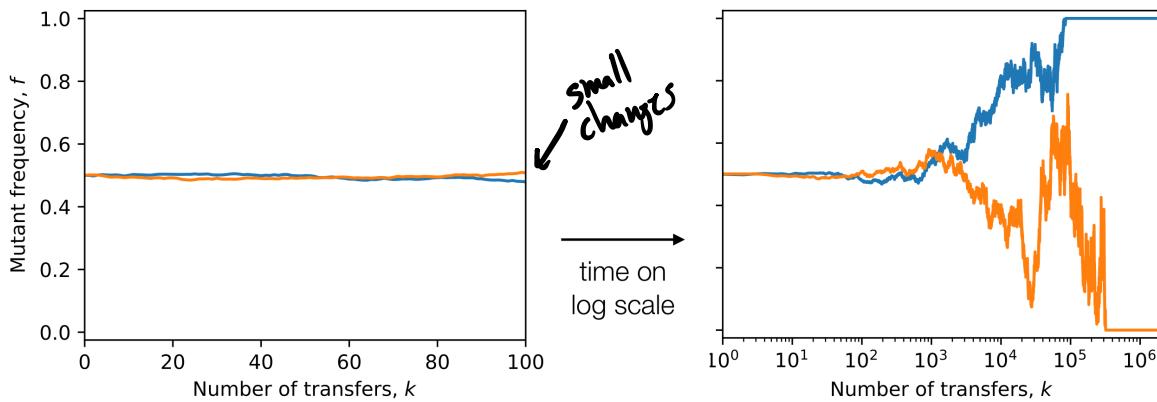
"genetic drift"

$\Rightarrow$  if  $N_0$  is large  $\Rightarrow$  genetic drift is pretty small!

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

$\Rightarrow$  but it is repetitive! (i.e. compounds over time)

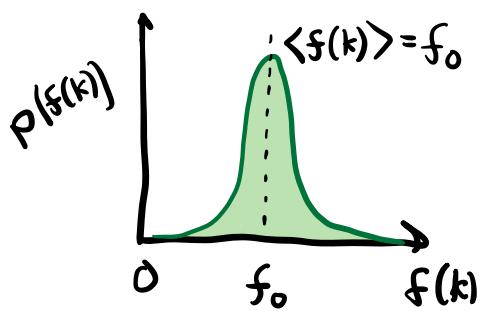
Computer simulations of model with  $s = 0$ ,  $N_0 = 10^5$ ,  $f(0) = 50\%$



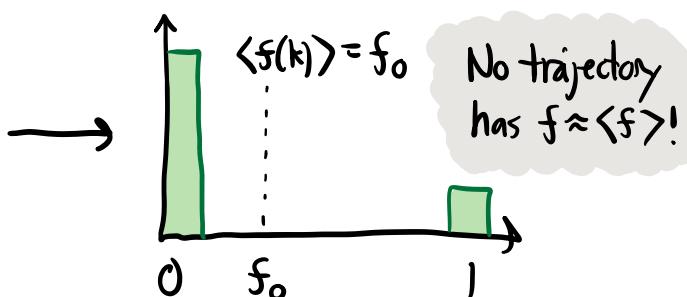
$\Rightarrow$  in 2nd case, something "singular" happens:

- ① if  $f=0$  @ one time  $\Rightarrow f=0$  @ all later times
  - ② if  $f=1$  @ "  $\Rightarrow f=1$  @ all "
- "fixation"  
"extinction"

$\Rightarrow$  Short times ("case 1")



Long times ("case 2")



$\Rightarrow$  Instead, avg is mixture of 2 outcomes:

$$\langle f(\infty) \rangle = 0 \times \Pr[f=0] + 1 \times \Pr[f=1] = f_0 \quad \text{from neutrality}$$

$$\Rightarrow \Pr[f=1] = f_0$$

"Fixation probability"  
of neutral mutation

$\Rightarrow$  but timescale required is quite long...

$\Rightarrow$  will show for short times:  $f(k) \approx f_0 \pm \Theta\left(\sqrt{\frac{k}{N_0}}\right)$   
"random walk"

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

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

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

$\Rightarrow$  selection will often be more important

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

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

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

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

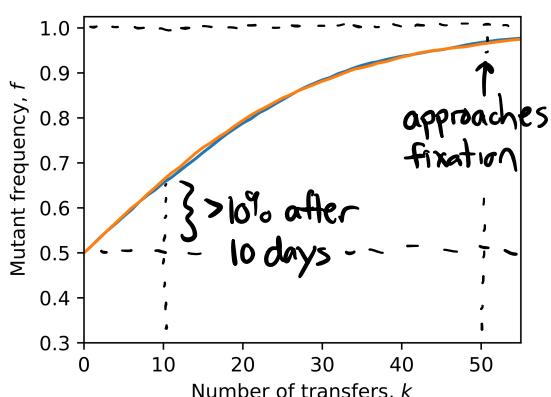
$$f(t) = \frac{f(0)e^{st}}{1-f(0)+f(0)e^{st}} \Leftrightarrow \text{"Logistic growth"} \quad \frac{df}{dt} = sf(1-f)$$

$\Rightarrow$  Now can get a big change:

e.g. if  $s=0.01$ ,

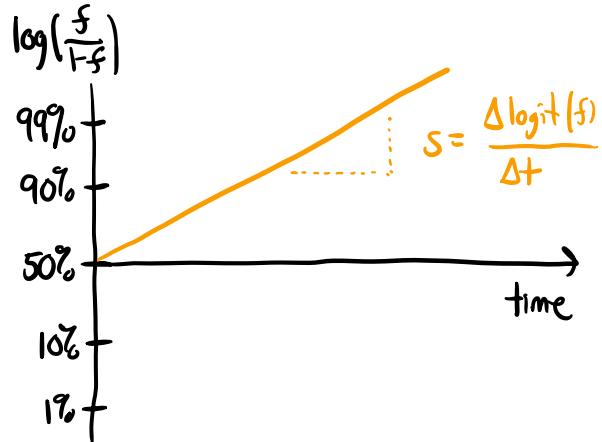
$$\Delta t = \log_2(100) \approx 7$$

$$\therefore N_0 \approx 10^5$$



$\Rightarrow$  Sometimes helpful to plot on "logit" scale:

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



$\Rightarrow$  upshot: can notice big change when  $s t \gtrsim 1$

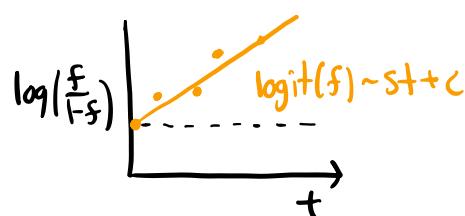
$\Rightarrow t \gtrsim 1/s$  "selection timescale"

$\Rightarrow$  So far, if know  $s$  (e.g. from previous expt's on underlying growth rate,  $r \rightarrow r+s$ )  
 can predict  $f(t)$  ...

$\Rightarrow$  Can also turn around & use as definition of  $s$ :

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

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

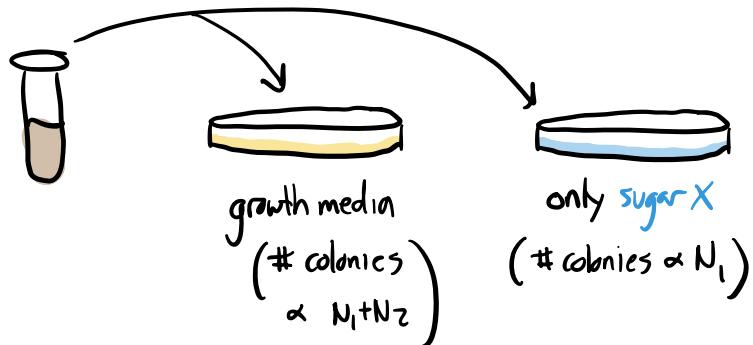


Question: How do we measure  $f(t)$ ?

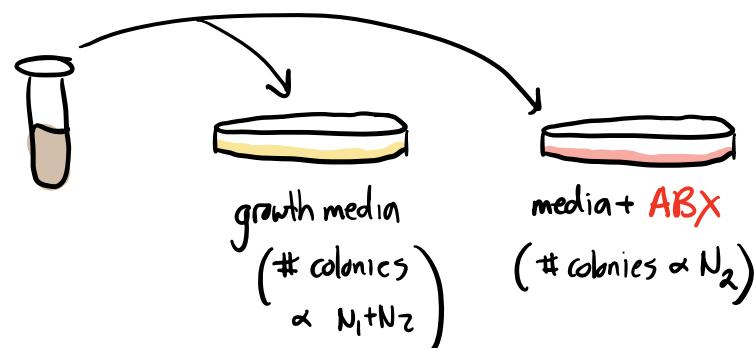
(in principle hard to distinguish similar-looking strains like WT,  $\Delta$ sugarX...)

① Old fashioned way: make them distinguishable & count colonies

e.g.  $\Delta$ sugarX :



e.g. + ABX resistance :

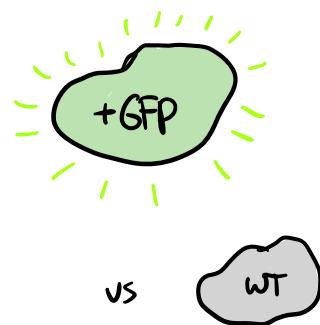


② Fluorescence + lasers ("flow cytometry")

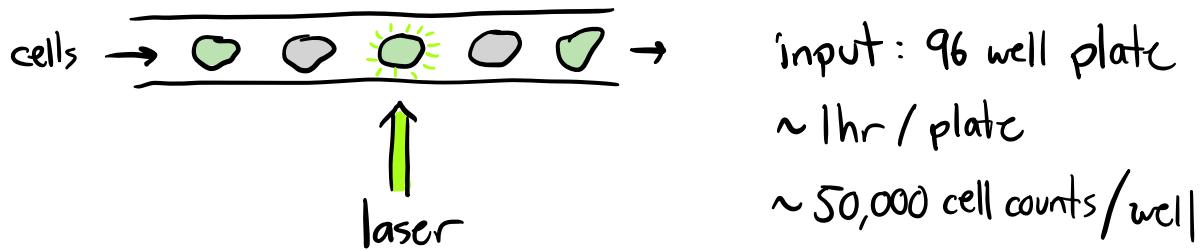


1. gene producing  
fluorescent protein  
(GFP, RFP, ...)

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



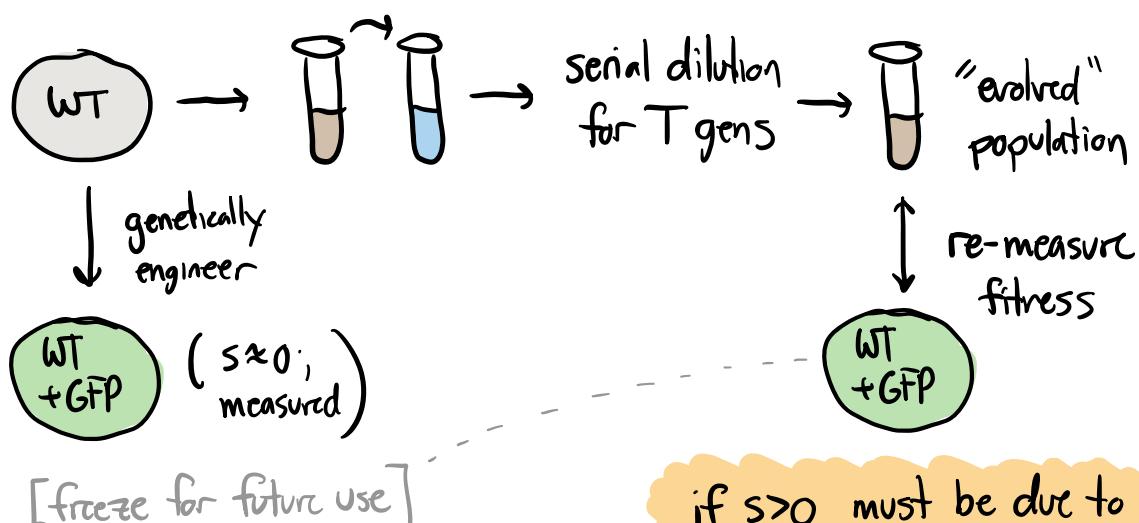
3. can count on "flow cytometer:



### 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:



if  $S > 0$ , must be due to mutations that arose in pop'n during expt.