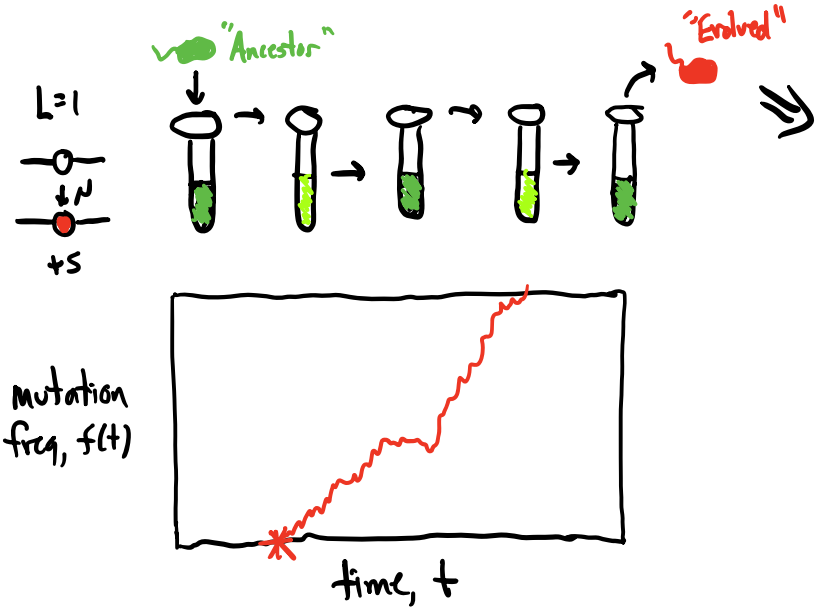
Chapter 8

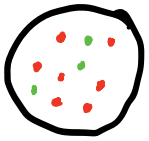
DNA sequencing & genomics

DNA Sequencing & Genomics

So far....



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



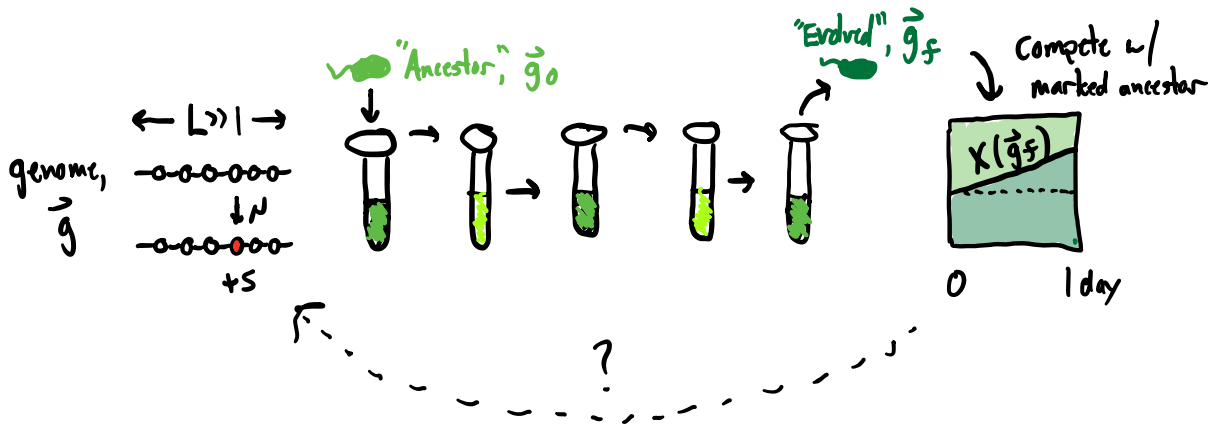
e.g. ΔsugarX

⇒ In practice, genomes contain many sites

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

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

Historically, experimental evolution relied on competitive fitness

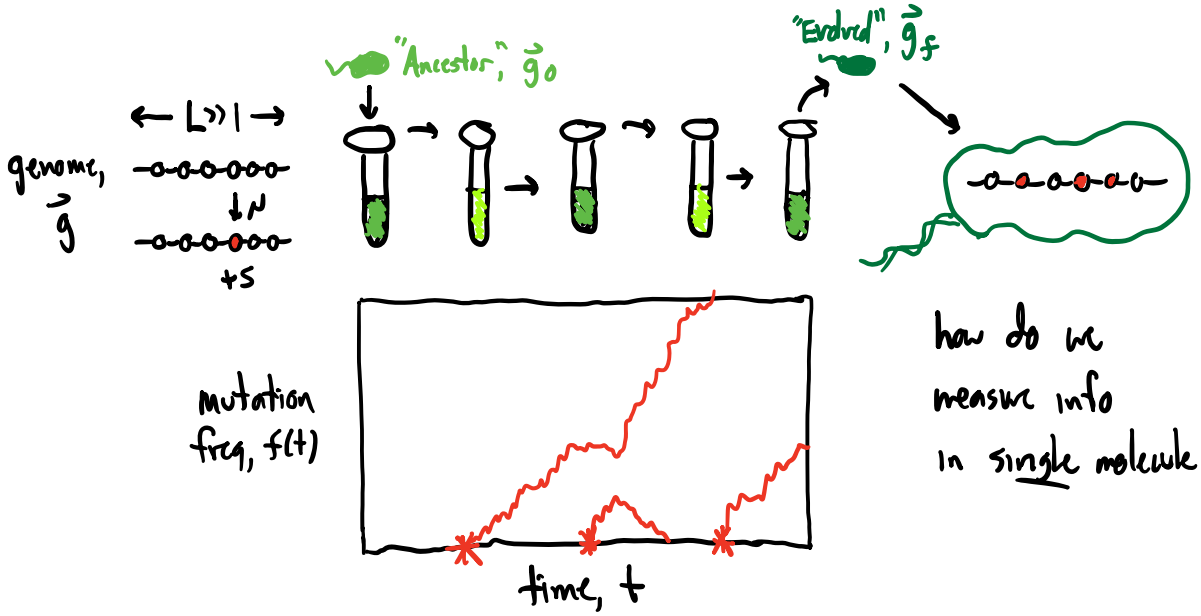


$\Rightarrow$ statistics of $X(\vec{g}_f)$ w/in & between populations tell us something about evolutionary dynamics of $\vec{g}$

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

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

Now: DNA sequencing allows us to measure genomes directly*

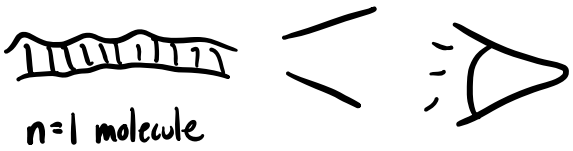


how do we measure info in single molecule?

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

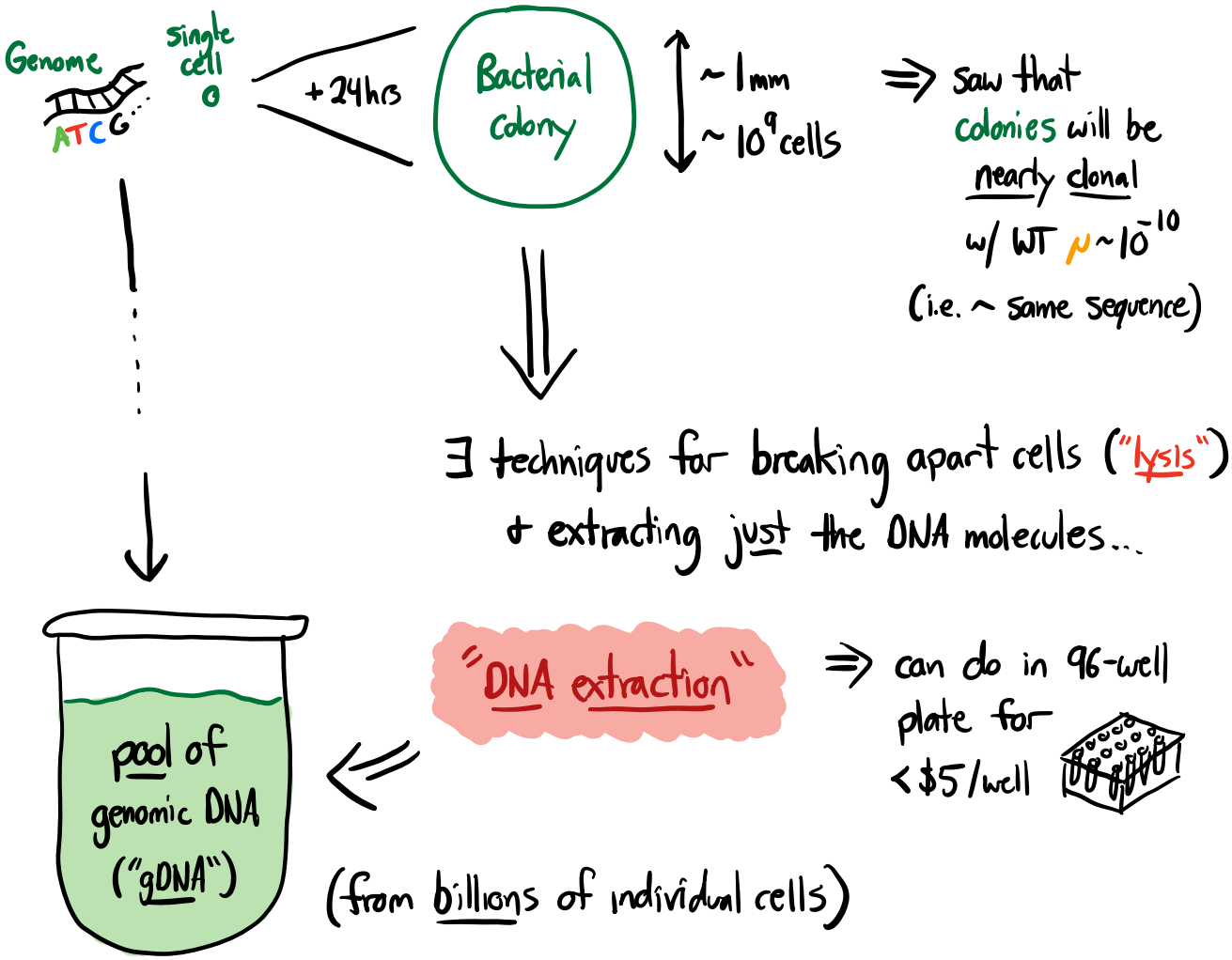
$L \sim 10^{4\text{-}5}$ viruses
$L \sim 10^{6}$ bacteria
$L \sim 10^{9}$ humans

$\Rightarrow$ Step 1 for reading genomes: amplification!

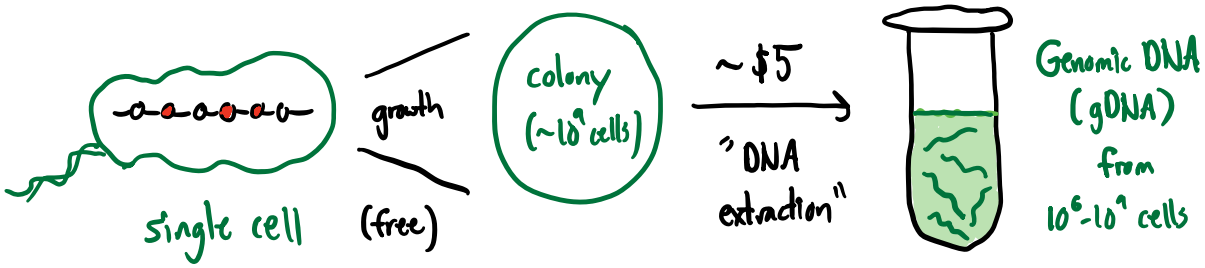


need macroscopic quantities of our DNA molecule (w/ same DNA sequence) to work with...

⇒ easy for lab bacteria! use built-in ability to grow exponentially:

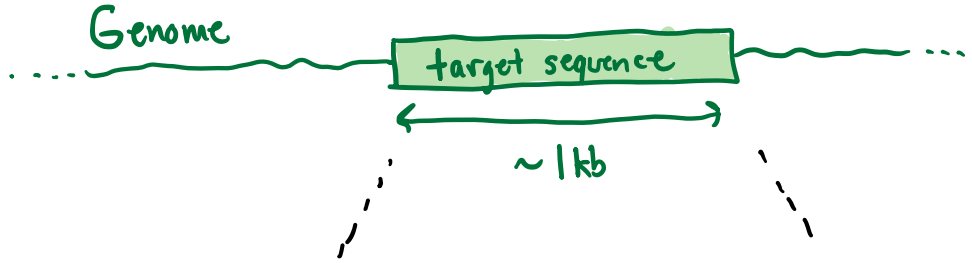


Recap: Step 1: Amplification ("get a macroscopic amount of them")



Problem: $L = 10^6$ sized genomes too difficult do measure directly

⇒ most sequencing methods work w/ short sequences ($\lesssim 10^3$ bp)



How to get a macroscopic amount of just this region?

Answer: PCR ("polymerase chain reaction")

⇒ cell-free chemical reaction that's like 

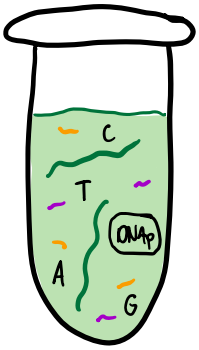
but just for part of genome.

⇒ takes advantage of DNA replication machinery invented by bacteria

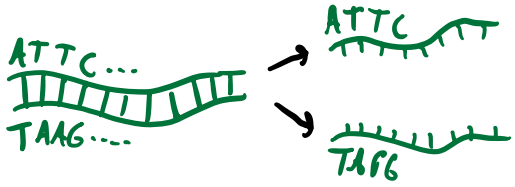
* But requires us to know some of the sequence near target

(1) can have company synthesize "primers" (short sequences of ~20bp) that correspond to P1 & P2. (~ $0.30/bp for ~ 10^3 reactions)

(2) mix w/ DNA polymerase,
dNTP (free A's, C's, T's, G's)
+ your genomic DNA.



(3) Heat sample so that DNA strands "melt" (~90°C)



(4) Now cool sample (~60°C) so that primers "anneal" to input DNA

$\Rightarrow$ melting & annealing is physics problem: $\frac{p(\text{bound})}{p(\text{unbound})} \sim e^{-\frac{\Delta E}{kT}}$

w/ $\Delta E \approx$ # matched bp (very roughly...)

⇒ want primers to bind to known region, but not anywhere else!

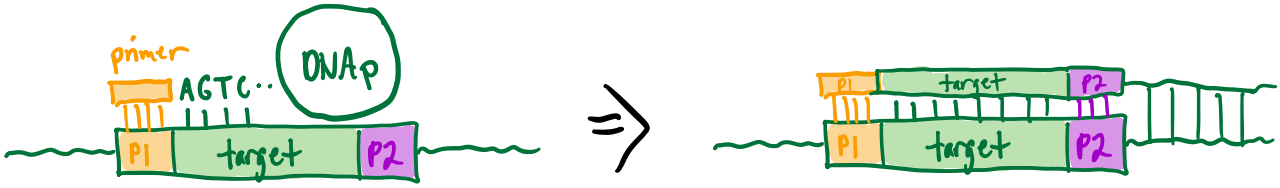
* we saw in PSET 1 that most of E. coli genome is uniquely identifiable w/ $\ell \approx 20$bp sequence ⇒ ~20bp primers sounds ok *

(more generally ⇒ "information theory argument" :

$$\#\text{unique locations } (L) \lesssim \#\text{unique sub-sequences } (4^{\ell})$$

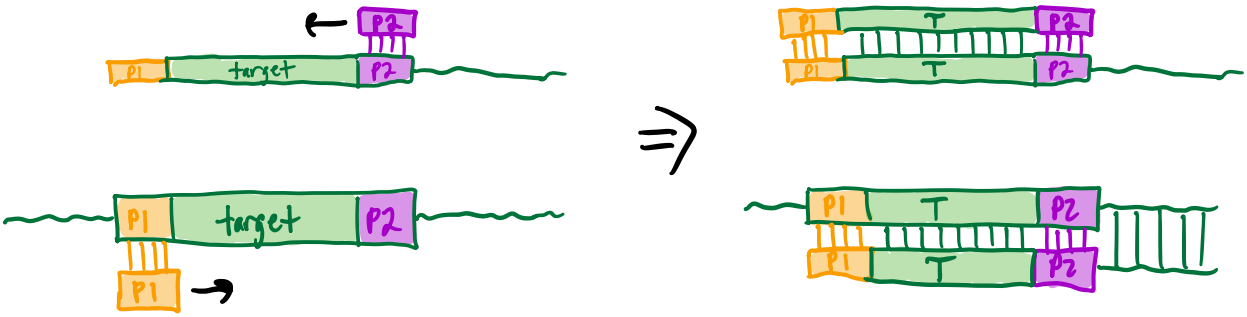
$$\Rightarrow \ell^* \gtrsim \log_4(L) \begin{cases} \sim 10 \text{ bp (E.coli)} \\ \sim 15 \text{ bp (Humans)} \end{cases})$$

(5) After primers are bound, DNA polymerase will start incorporating dNTPs onto primer to create complementary strand...



("extension phase")

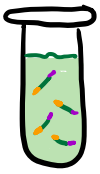
(6) Melt, anneal, & extend again:



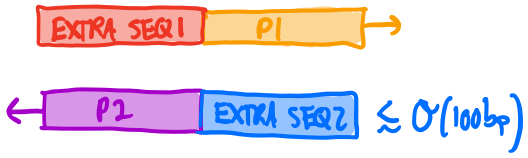
(7) Repeat for $k \sim 20\text{-}30$ cycles...

$\Rightarrow$ exponential amplification of  sequence!

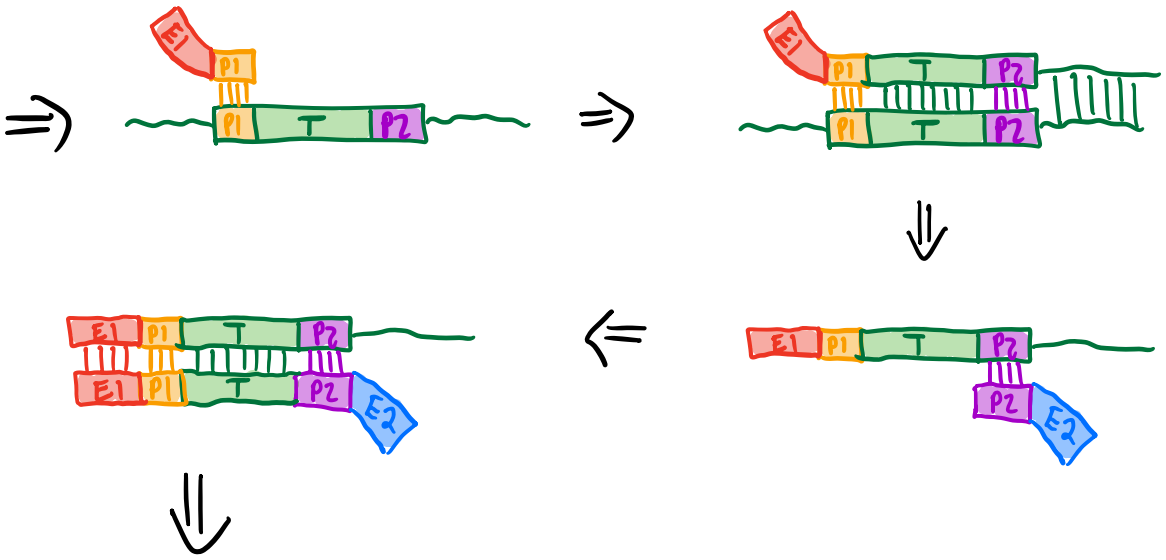
("amplicon")

(8) "clean up" to remove leftover primers, etc. $\Rightarrow$  macroscopic # of amplicons!

$\Rightarrow$ Note: can also use PCR to add extra bit of DNA sequence to your target region...

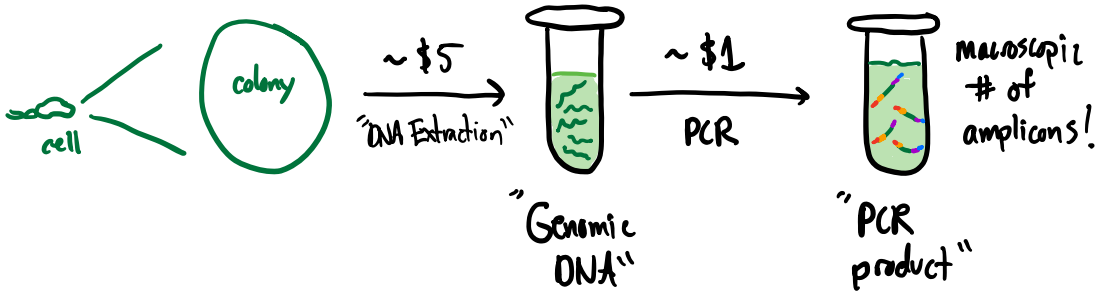
$\Rightarrow$ order modified primers: 

$\lesssim O(100\text{bp})$



Amplified sequence:

Recap:

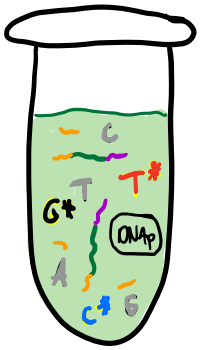


⇒ How do we read out information from PCR products?

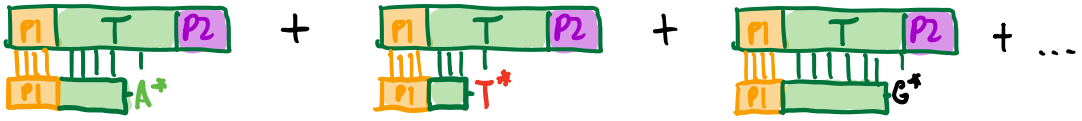
Traditional approach: Sanger sequencing

(like 1 extra round of PCR...)

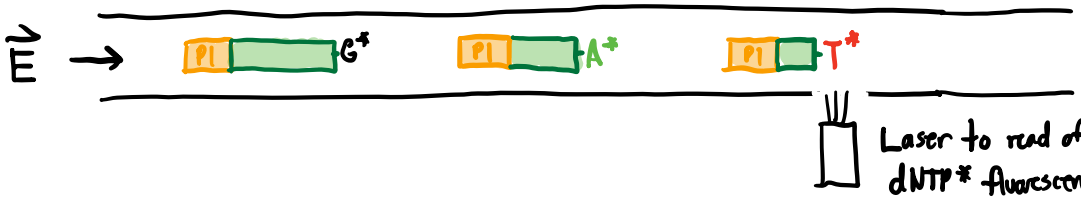
Idea: mix PCR product w/ P1, DNAp, dNTPs,
+ special fluorescent dNTPs that block DNAp
(A^*, G^*, C^*, T^*)



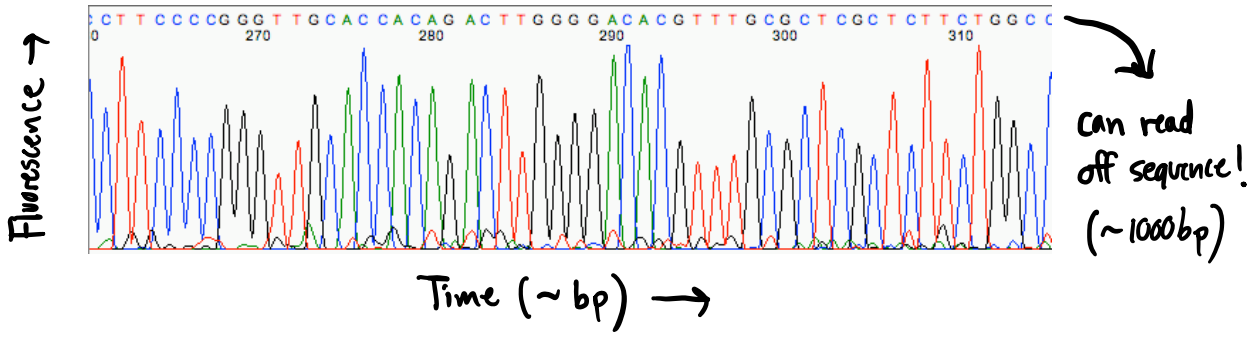
⇒ After 1 round of extension, random mixture of



⇒ Flow in electric field (shorter fragments move faster)



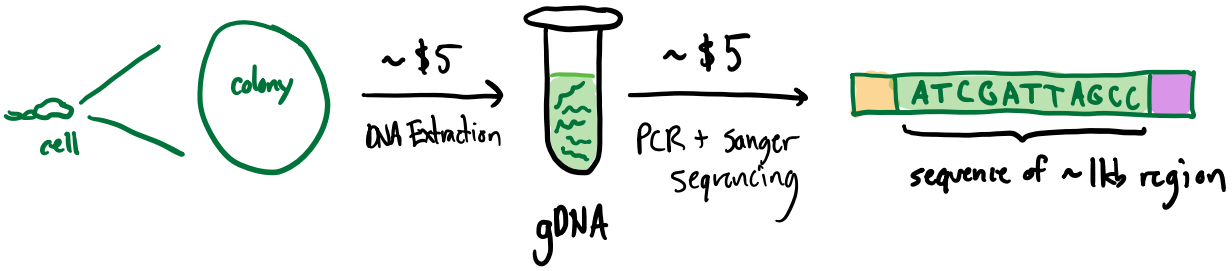
⇓ "chromato-gram"



⇒ costs ~$5 (send away overnight: 

)

⇒ so have seen how to go from:



⇒ in Problem 1 of PSET 2, Layd Murray used sanger seq to sequence URA3 genes in ~300 yeast colonies

⇒ $1500 total

⇒ But expensive to sequence whole genomes!

e.g. E. coli: 1 clone = 10^6 bp ⇒ 10^3 Sanger seq runs → $5k

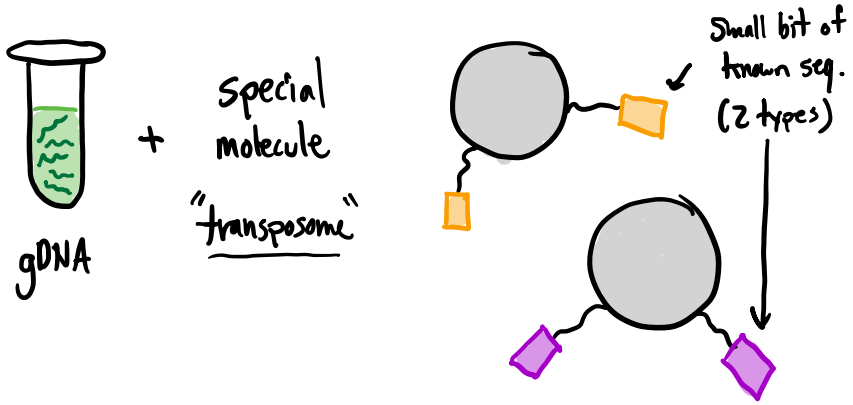
e.g. Humans: 1 person = 10^9 bp ⇒ 10^6 Sanger seq runs → $5M

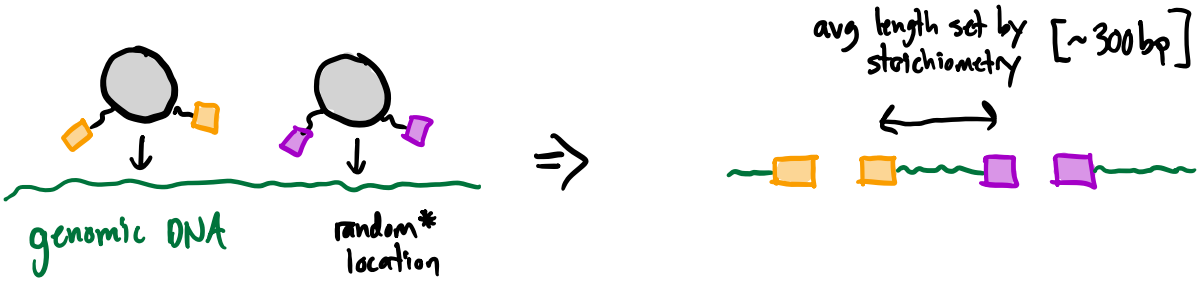
⇒ Now things are much cheaper w/ Next-gen sequencing ("Illumina sequencing")

⇒ same idea, but higher throughput!

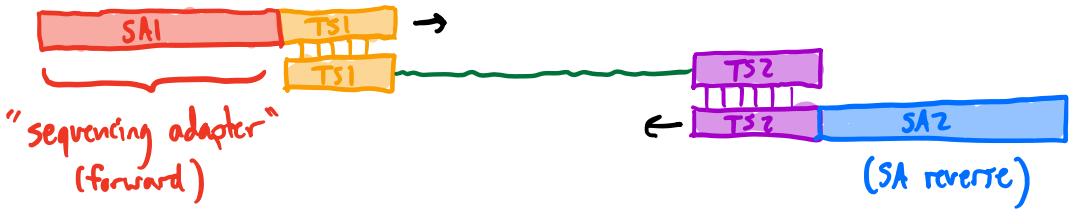
Step 1:

"Library prep"

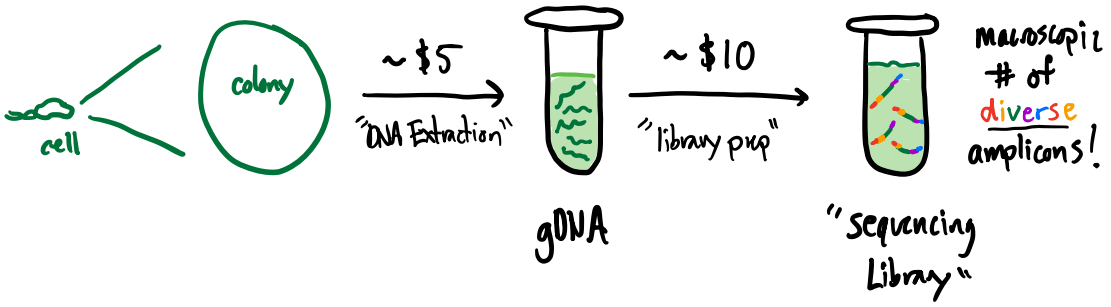




Step 2: Now in position to do PCR + add extra known sequence:

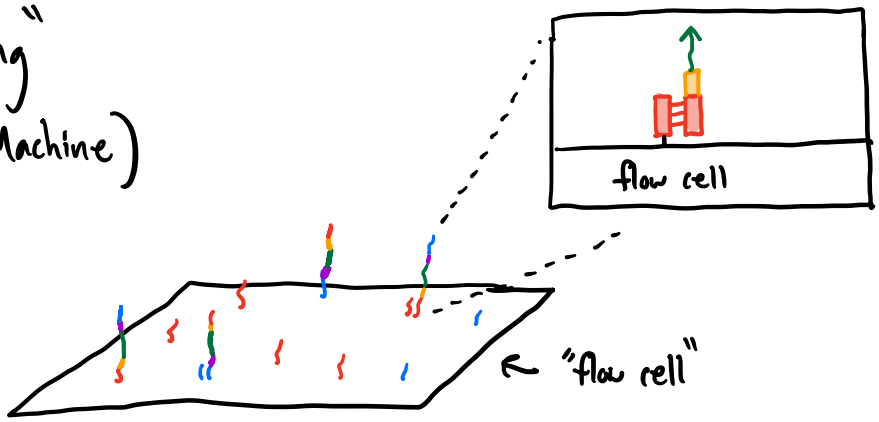


⇒ This gives us:



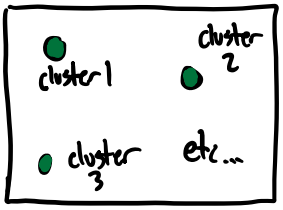
Step 3: "Sequencing" (Illumina Machine)

Flow library onto chip w/ probes that bind sequencing adapter



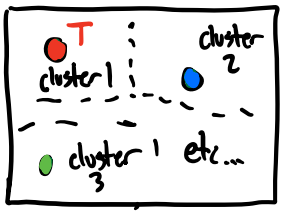
Step 4: do more PCR on bound fragments to turn each molecule into macroscopic cluster of identical* molecules

top down view:



Step 5: flow primer (SA1) + fluorescent dNTPs that incorporate once (then stop)

⇒ if take picture of flow cell:



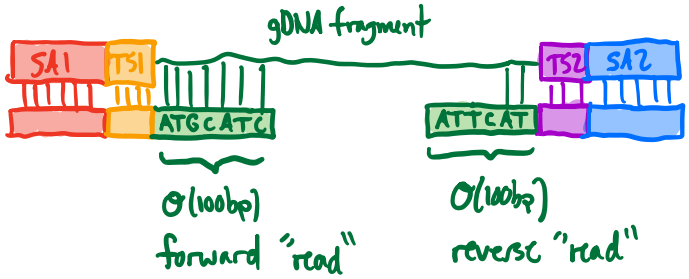
⇒ different colors = nucleic acid @ first position

Step 6: remove fluorescent part (+ block), + repeat w/ new round of dNTPs ⇒ 2nd position.

Step 7: can repeat for O(100) cycles [until pictures ≈ desynchronized]

Step 8: can also repeat for reverse direction (SA2)

Net result:



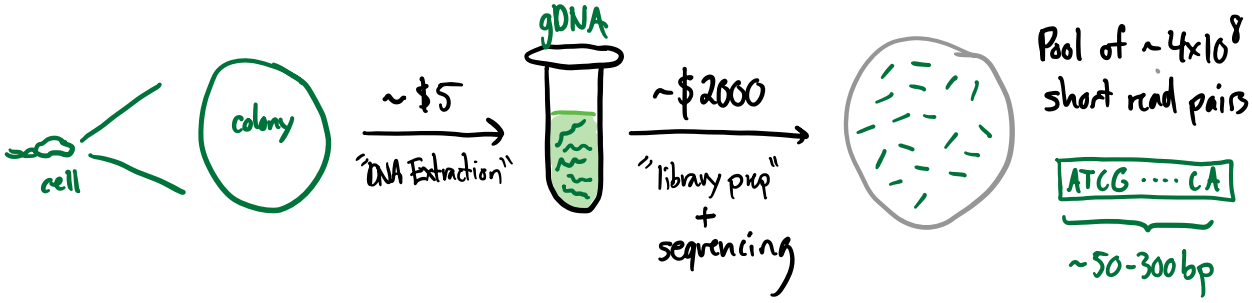
⇒ get O(100 bp) read out from each end of single DNA fragment

⇒ w/ modern Illumina machines, process is very high throughput!

⇒ ~ 4×10^8 read pairs in ~ few days for ~ $2000

(catch: can't do smaller batches)

Recap: Now have method for going from:

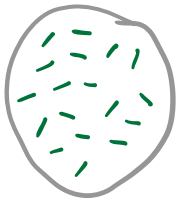


Next: what can we do with this kind of data?

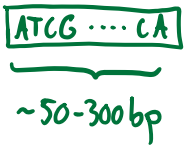
⇒ need to put puzzle back together... 2 main methods

① "De novo Genome Assembly"

(common programs: spades, Velvet, ...)



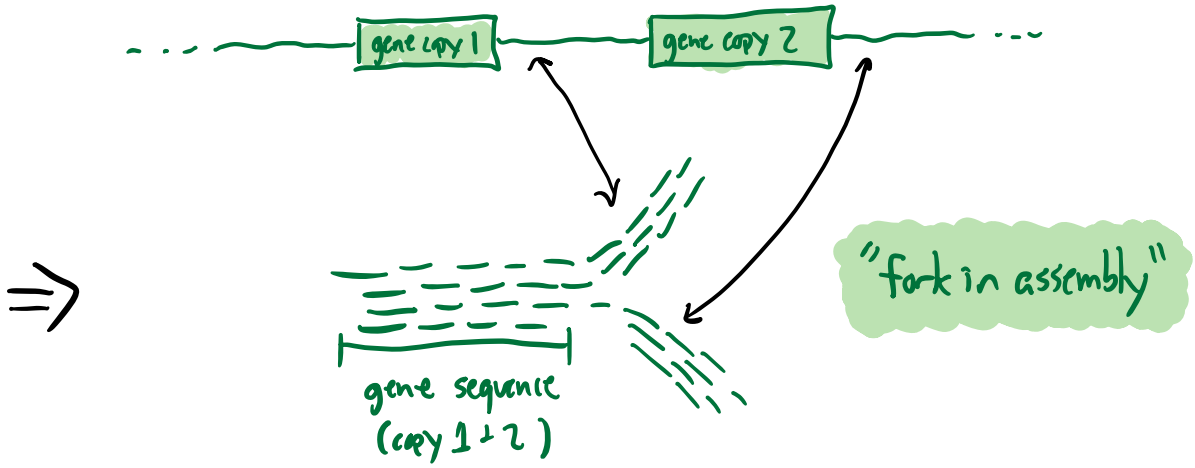
Pool of ~4×10^8 short read pairs



Idea: look for reads that partially overlap (e.g. ≥ 20bp)

⇒ simple in principle, but lots of corner cases...

⇒ e.g. what if 2 regions of genome are identical for ≥ 100bp (or length of read?)

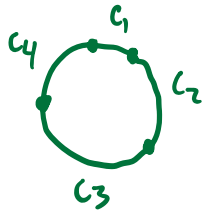


⇒

⇒ for these & other reasons, assembly typically yields collection of discrete contigs,

each ~10^3-10^5 bp long: 

⇒ much harder (+ manual effort)

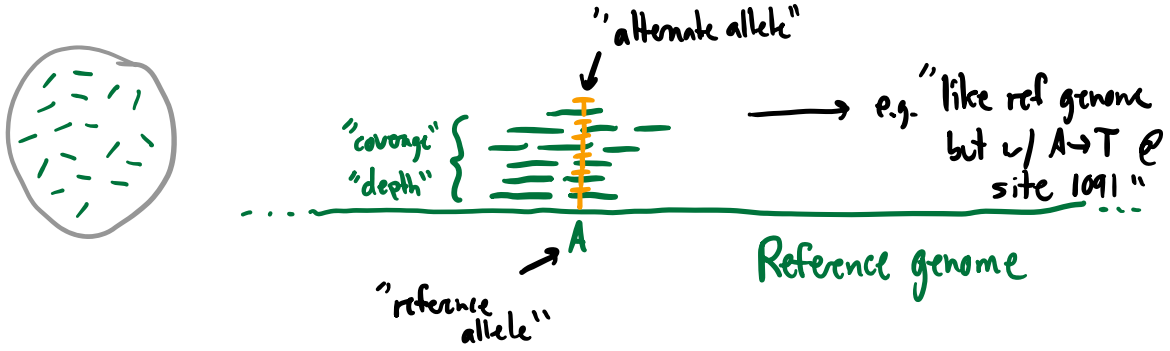
to "finish" into complete genome 

+ HUGE memory requirements (~32Gb-1Tb depending on L)

(since need to compare all* pairs of reads to each other...)

② Alignment of reads to Reference Genome

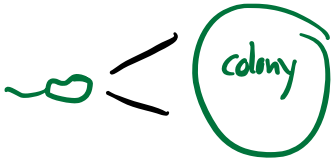
⇒ if already have assembled genome from related strain...
can align reads to best matching place & look for changes
(common programs: BLAST, Bowtie2, BWA-MEM, + mpileup)



Major wrinkle: sequencing errors (A→T by chance, "shot noise")

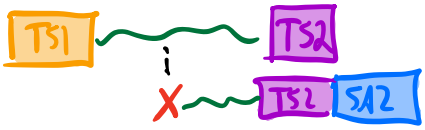
where could these errors come from?

① initial grow up:



⇒ but happen w/ low rate ($\mu \sim 10^{-10}$/bp/gen)

② library prep & PCR



$\Rightarrow$ higher rates $\sim 10^{-6}$/bp/cycle $\rightsquigarrow$ 10^{-4}/bp total.

(3) Errors on sequencing machine $\Rightarrow$ estimated $\sim 10^{-3}$/bp

(cluster generation PCR errors)
wrong fluorescent base...

* but can vary from site-to-site
(up to 10^{-2} sometimes...)

$\Rightarrow$ dominant source of noise $\Rightarrow$ $P_{err} \sim 10^{-3}$ (up to $\sim 10^{-2}$)
in special cases...

$\Rightarrow$ low rate per site, but # sites $L \gg 1$!

$\Rightarrow$ # errors per genome = $L \times P_{err}$ $\sim 10^{3}$ errors / E.coli
$\sim 10^{6}$ errors / human

$\Rightarrow$ fortunately, can correct errors w/ consensus sequences!

$\Rightarrow$ higher coverage is helpful... how much do we need?

$$Pr(\text{consensus error}) = \sum_{A=\bar{D}/2}^{\infty} \frac{(P_{err}\times\bar{D})^A}{A!} e^{-P_{err}\bar{D}} \simeq \frac{(P_{err}\bar{D})^{\bar{D}/2} e^{-P_{err}\bar{D}}}{(\bar{D}/2)!}$$

$$\#\text{ errors in genome} = L \times Pr(\text{consensus error})$$

$$\simeq \exp\left[\log L - \frac{\bar{D}}{2}\left[\log\left(\frac{1}{2P_{err}}\right) - 1\right]\right] \lesssim O(1)$$

$$\Rightarrow \text{solve for } \bar{D} \Rightarrow \quad \bar{D} \simeq \frac{2\log L}{\log\left(\frac{1}{2P_{err}}\right) - 1} \approx \begin{cases} 5 \text{ if } L\sim 10^{+6},\ P_{err}\sim 10^{-3} \\ 8 \text{ if } L\sim 10^{9} \end{cases}$$

$\Rightarrow$ so need coverage of $\geq 10\times$ coverage to detect single mutations.

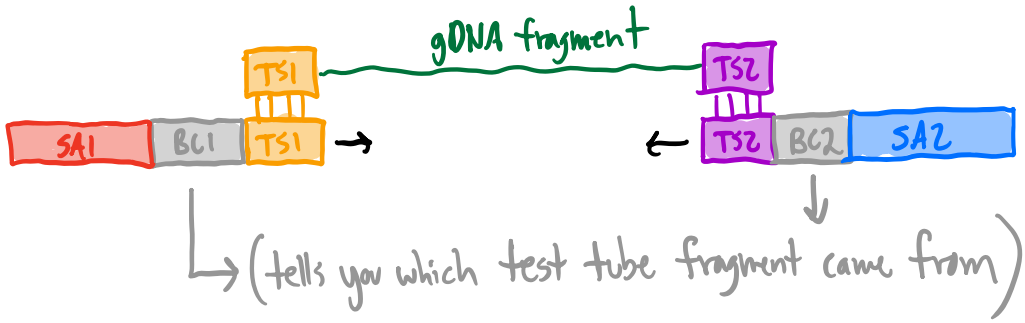
$\Rightarrow$ How many reads is this?

E.coli: $10 \frac{\text{reads}}{\text{site}} \times 10^6 \text{ bp} \Rightarrow 10^7 \text{ bp} \Rightarrow 10^5 \text{ reads}$

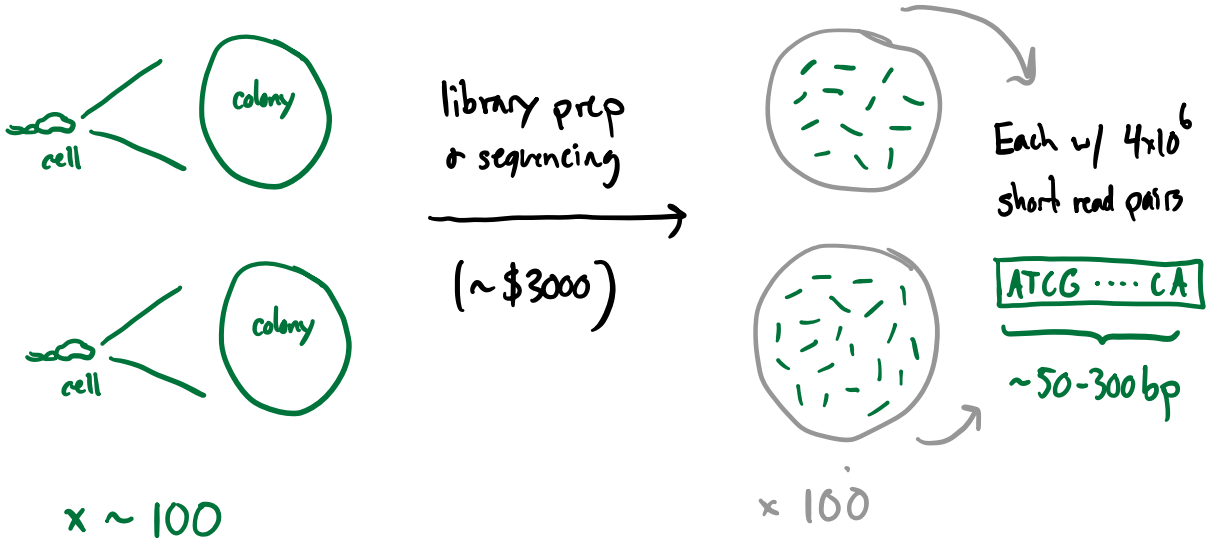
$\Rightarrow$ but a single run of Illumina seq produces 4×10^8 reads...

$\Rightarrow$ overkill! (wastes money...)

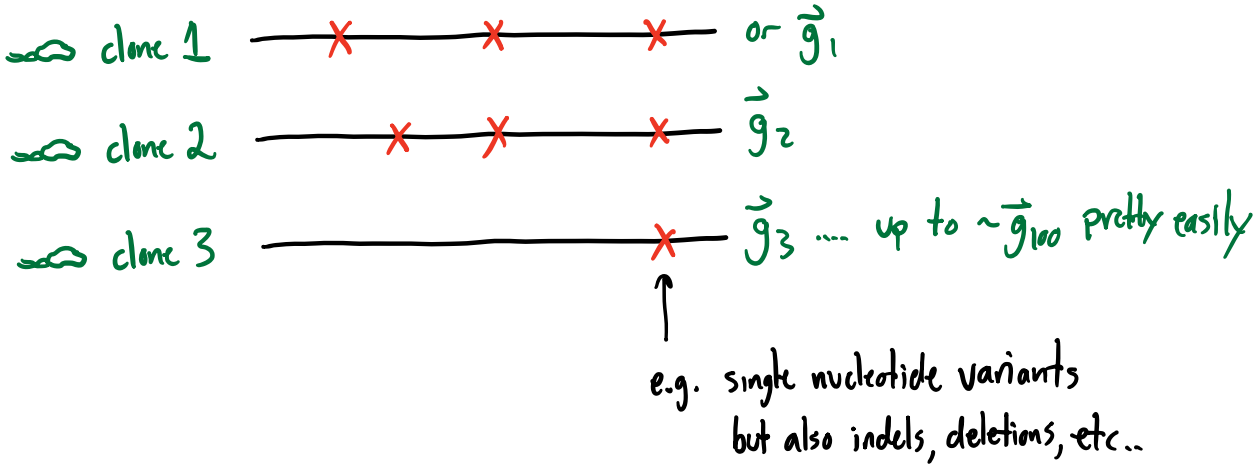
Solution: "multiplex": Add sample specific "barcode" sequence during library prep step



Upshot: can sequence ~100 E. coli libraries on one flow cell & get 300-fold coverage of E. coli genome



⇒ After aligning reads & detecting "true" mutations, get sequences of genomes:



How are sampled genomes related to dist'n of genomes in pop'n?

⇒ let $n_{\vec{g}}$ = # sampled clones w/ genome $\vec{g}$ (random from sampling)

$f(\vec{g})$ = frequency of genome $\vec{g}$ in population (random from evolution)

⇒ then $\Pr\left[\{n_{\vec{g}}\} \middle| n, \{f(\vec{g})\}\right]$ = multinomial distribution

(n: total # clones sampled)

$$\propto \prod_{\vec{g}} \frac{f(\vec{g})^{n_{\vec{g}}}}{n_{\vec{g}}!}$$

⇒ Genotype space is huge!

⇒ often coarse-grain to summary statistics.

e.g. n_ℓ = # individuals w/ mutation @ site ℓ = $\sum_{\vec{g}} g_\ell n_{\vec{g}}$

(g_ℓ: 1 if mutⁿ @ site ℓ, 0 else)

⇒ can show $\Pr[n_\ell | n, \{f(\vec{g})\}] = \text{Binomial}(n, f_\ell)$

$f_\ell \equiv \sum_{\vec{g}} g_\ell f(\vec{g})$

e.g. total # of mutations separating 2 genomes

⇒ since depends on genome length, often normalized by L:

$$\frac{\#\text{ mutations between 2 random clones}}{L} = \begin{cases} \text{"heterozygosity" } (\pi) & \text{if from same pop'n} \\ \text{"divergence" } (d) & \text{if from diff "species" (or isolated sub-pops)} \end{cases}$$

e.g. heterozygosity (π) in humans is $\sim 10^{-3}$

divergence (d) between humans & chimps is $\sim 10^{-2}$

heterozygosity (π) between E.coli in different humans is $\sim 10^{-2}$

Can we relate π to genotype distribution, $f(\vec{g})$?

Note that: ($g_{1\ell}, g_{2\ell} \in \{0,1\}$)

$$\pi = \frac{1}{L}\sum_{\ell=1}^{L}\left[\, g_{1\ell}(1-g_{2\ell}) + (1-g_{1\ell})g_{2\ell} \,\right]$$

$$\langle \pi \,|\, \{f(\vec{g})\} \rangle_{\text{sampling}} = \frac{1}{L}\sum_{\ell=1}^{L}\left[\underbrace{\langle g_{1\ell}(1-g_{2\ell})\rangle}_{f_\ell(1-f_\ell)} + \underbrace{\langle (1-g_{1\ell})g_{2\ell}\rangle}_{(1-f_\ell)f_\ell} \right]$$

$$= \frac{1}{L}\sum_{\ell=1}^{L} 2f_\ell(1-f_\ell)$$

Remember: $f_\ell \equiv \sum_{\vec{g}} g_\ell f(\vec{g})$

$\Rightarrow$ averaging over f_ℓ:

$$\langle \pi \rangle = \frac{1}{L}\sum_{\ell=1}^{L}\langle 2f_\ell(1-f_\ell)\rangle = \frac{1}{L}\sum_{\ell=1}^{L}\int 2f_\ell(1-f_\ell)\,P(f_\ell)\,df_\ell$$

e.g. if genome is collection of neutral sites

$$\Rightarrow p(f_\ell) \approx \frac{2N_e\nu}{f_\ell} \quad \text{(quasi-stationary dist'n)}$$

$$\langle \pi \rangle = \int 2f(1-f)\,\frac{2N_e\nu}{f}\,df = \boxed{2N_e\nu}\ ^*$$

$\rightarrow$ population can only spread out so far in sequence space!

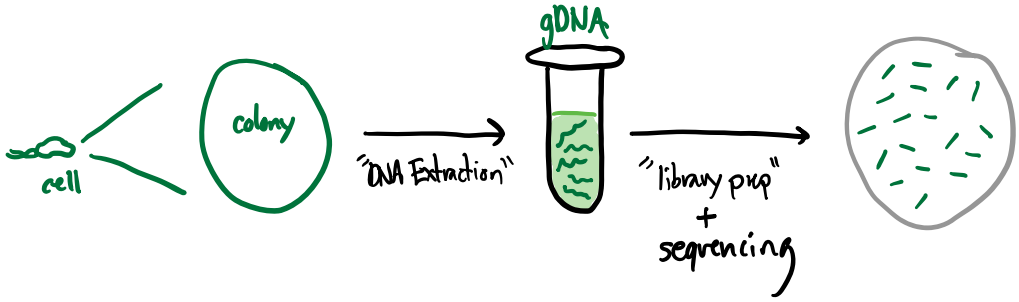
$$\Rightarrow \text{fit } N_e \equiv \frac{\langle \pi \rangle}{2\nu} \approx \frac{10^{-3}}{2\times 10^{-8}}$$

$\Rightarrow$ Variance of π $\Rightarrow$ much more complicated!

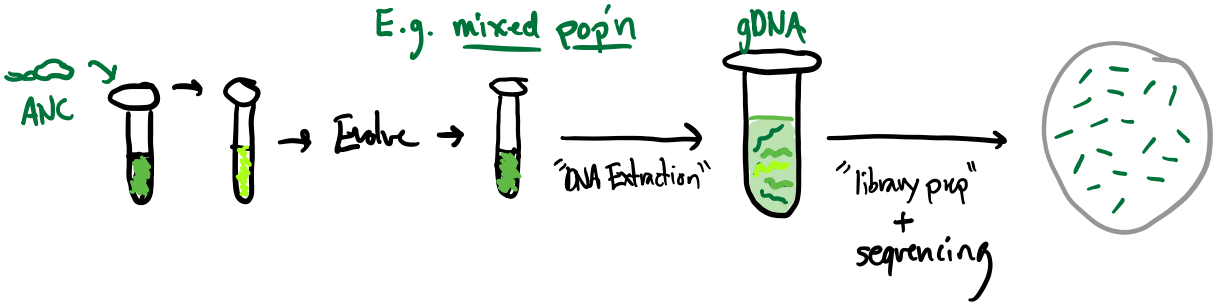
$\Rightarrow$ correlations between g_ℓ & $g_{\ell'}$

$\Rightarrow$ will see more later!

So far, have focused on clones



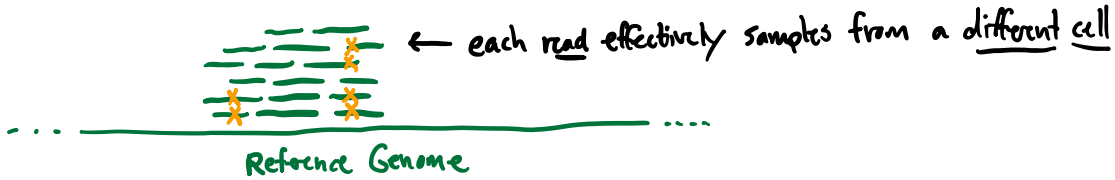
⇒ lots of other things we could put in our library prep...



⇒ known as "pooled sequencing" / "metagenomic sequencing"

⇒ in this case, assembly very hard!

⇒ Payoff comes from reference mapping:



e.g. if $A_\ell \equiv$ # reads w/ mutations @ site ℓ

$$\Rightarrow \Pr\left[A_\ell \,\middle|\, D_\ell, \{f(\vec{g})\}\right] = \text{Binomial}\left(D_\ell, f_\ell\right)$$

(D_ℓ: total coverage @ site ℓ)

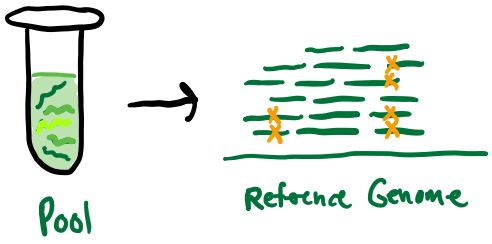
+ sequencing/PCR errors

$\Rightarrow$ since we can sequence ~100 E.coli genomes @ >100x coverage in 1 run of Illumina sequencing

$\Rightarrow$ can effectively sample ~100 clones ~100x more cheaply by sequencing pools!

(much cheaper way to track freqs of individual mut'ns)

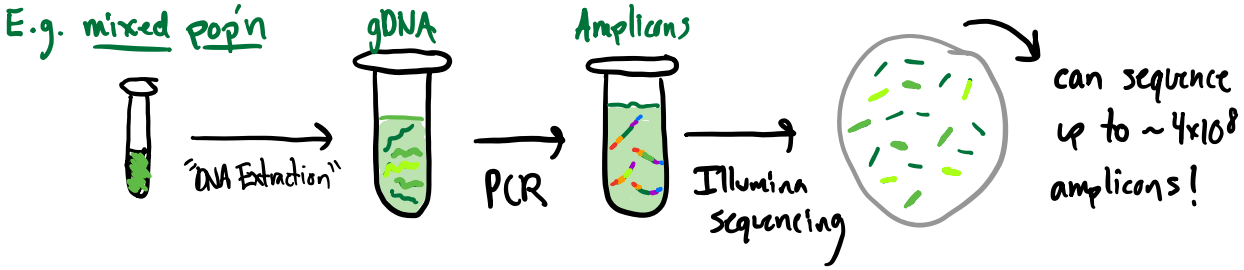
Downsides:



1. sequencing errors!
2. which mutations are in same cells? ("linkage information")

e.g.  vs

Can also sequence pools of amplicons:



Two common targets:

all bacteria have it

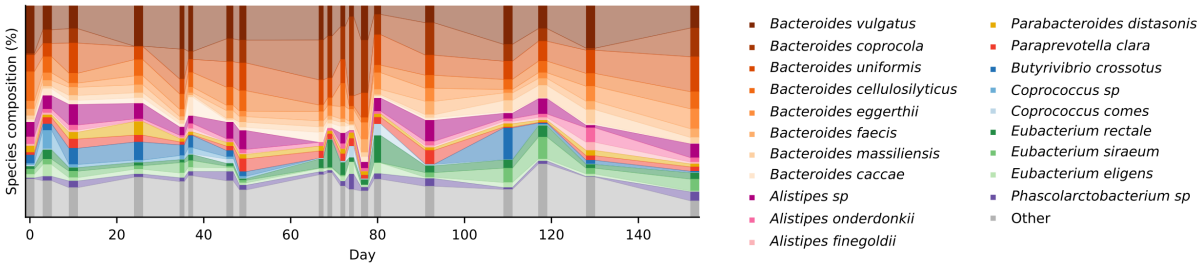
↳ track abundances of different species

can be added to lab strains

↳ track closely related strains

⇓

E.g. 1 person's microbiome over time



⇓

E.g. Homework Problem