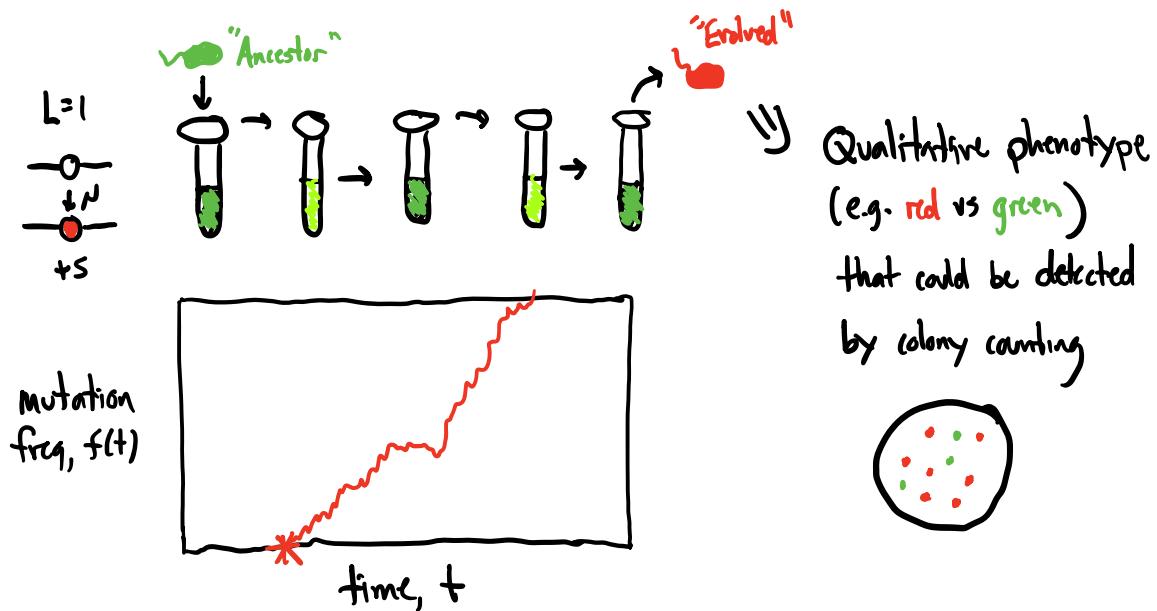


Chapter 8

DNA sequencing & genomics

DNA Sequencing + Genomics

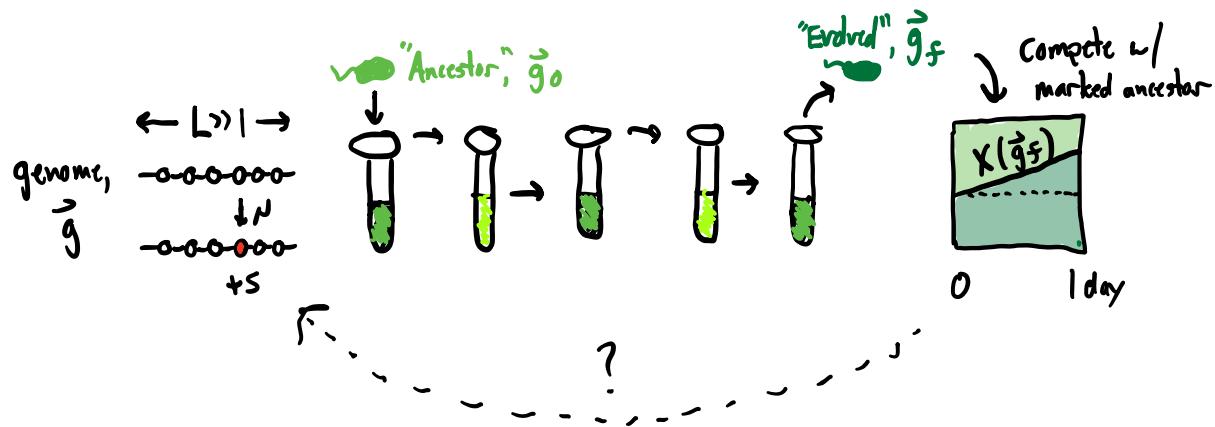
So far....



⇒ In practice, genomes contain many sites
⇒ don't know what phenotypes mutations
@ these sites produce or how to
measure them w/ colony counting assay...

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

Historically, experimental evolution relied on competitive fitness

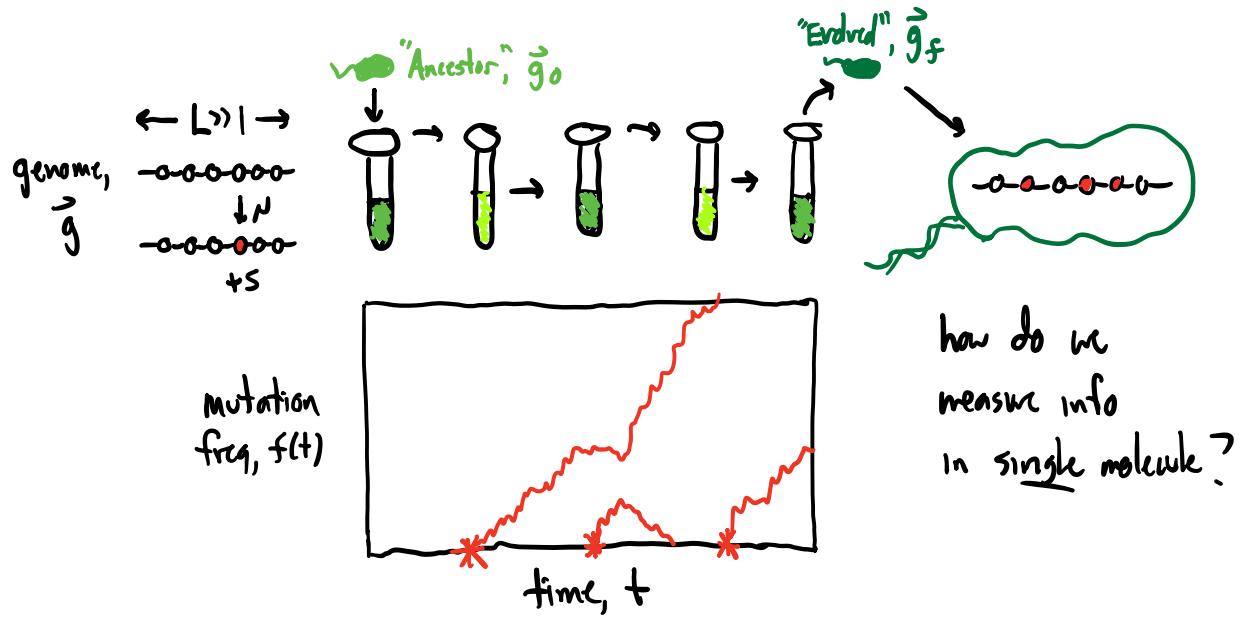


\Rightarrow statistics of $X(\vec{g}_f)$ w/in & between populations
tell us something about evolutionary dynamics of \vec{g}

\Rightarrow downside: 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*

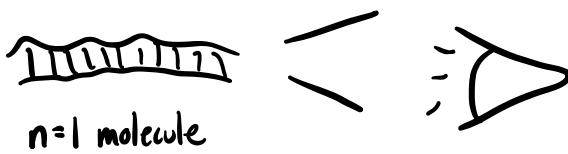


Recall: genome = $\begin{matrix} \text{A} & \text{T} & \text{C} & \text{G} & \text{T} & \text{A} \\ \text{T} & \text{A} & \cdots & \cdots & \cdots & \end{matrix}$

complementary strands

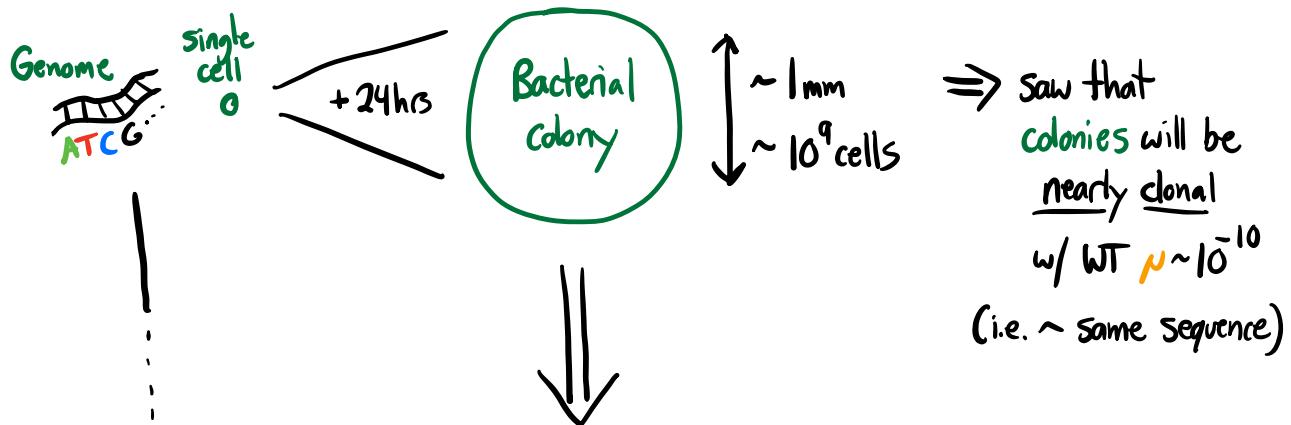
$L \sim 10^{4-5}$ viruses
 $L \sim 10^6$ bacteria
 $L \sim 10^9$ humans

\Rightarrow Step 1 for reading genomes: amplification!

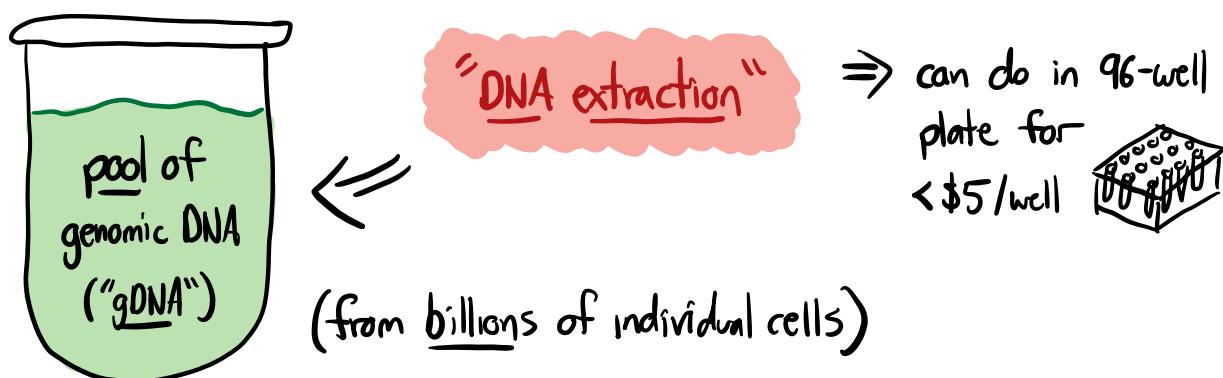


need macroscopic quantities
 of our DNA molecule
 (w/ same DNA sequence)
 to work with...

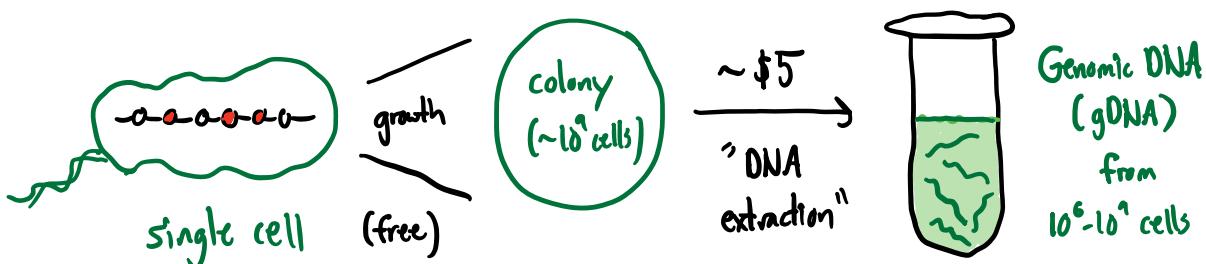
⇒ easy for lab bacteria! use built-in ability to grow exponentially:



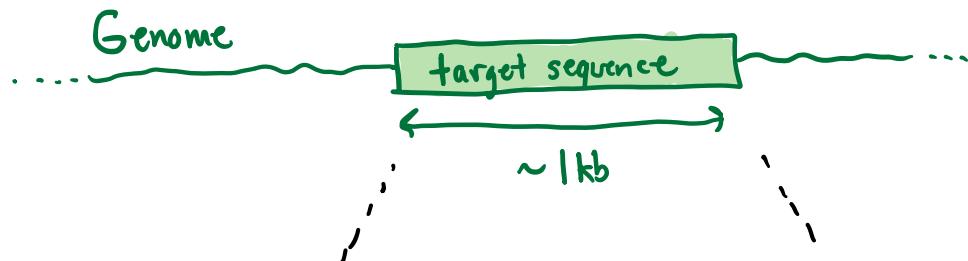
∃ techniques for breaking apart cells ("lysis")
+ extracting just the DNA molecules...



Step 1: Amplification (i.e. "get a macroscopic amount of them")



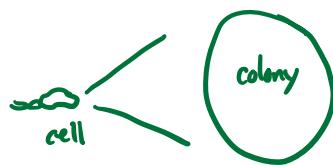
Problem: $L = 10^6$ sized genomes too difficult to measure directly
⇒ most sequencing methods work w/ short sequences ($\leq 10^3$ bp)



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.



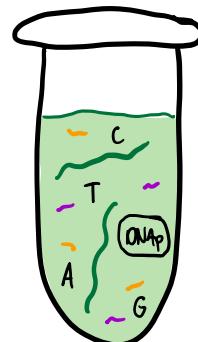
⇒ takes advantage of DNA replication machinery invented by bacteria

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

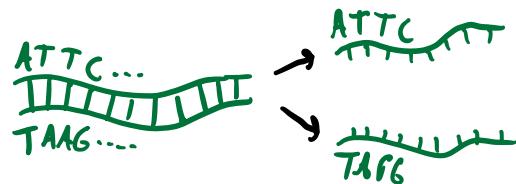


① can have company synthesize "primers" (short sequences of ~20bp) that correspond to P1 + P2. (~\$0.30/bp for ~10³ reactions)

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)



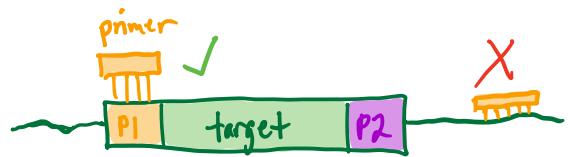
④ Now cool ^ 60°C sample so that primers "anneal" to input DNA



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

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

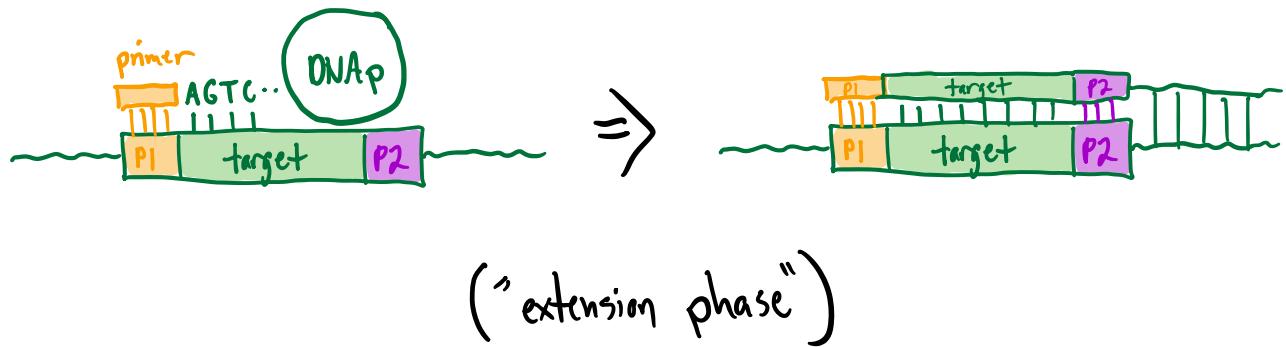
\Rightarrow want primers to bind to known region, but not anywhere else!



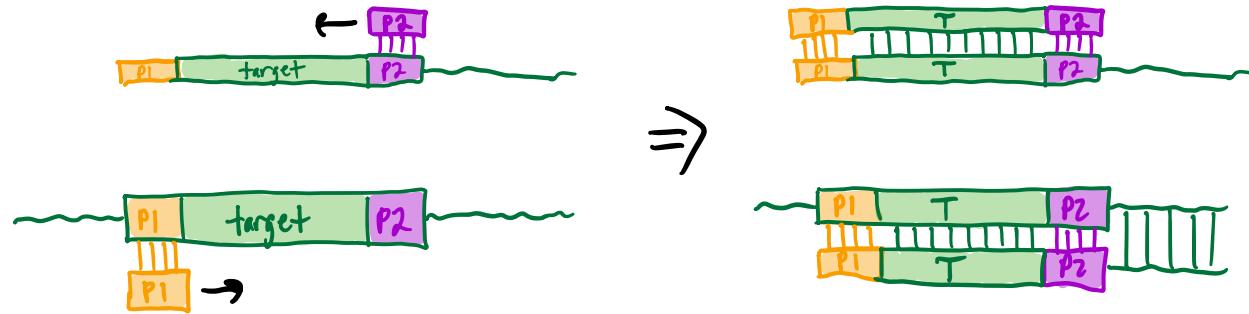
* we saw in PSET 1 that most of E.coli genome is uniquely identifiable w/ $l \approx 20\text{bp}$ sequence \Rightarrow ~20bp primers sounds ok *

(more generally \Rightarrow information theory argument, see p.5 of notes...)

④ After primers are bound, DNA polymerase will start incorporating dNTPs onto **primer** to create complementary strand...



⑤ Melt, anneal, & extend again :



⑥ Repeat for $k \sim 20-30$ cycles ...

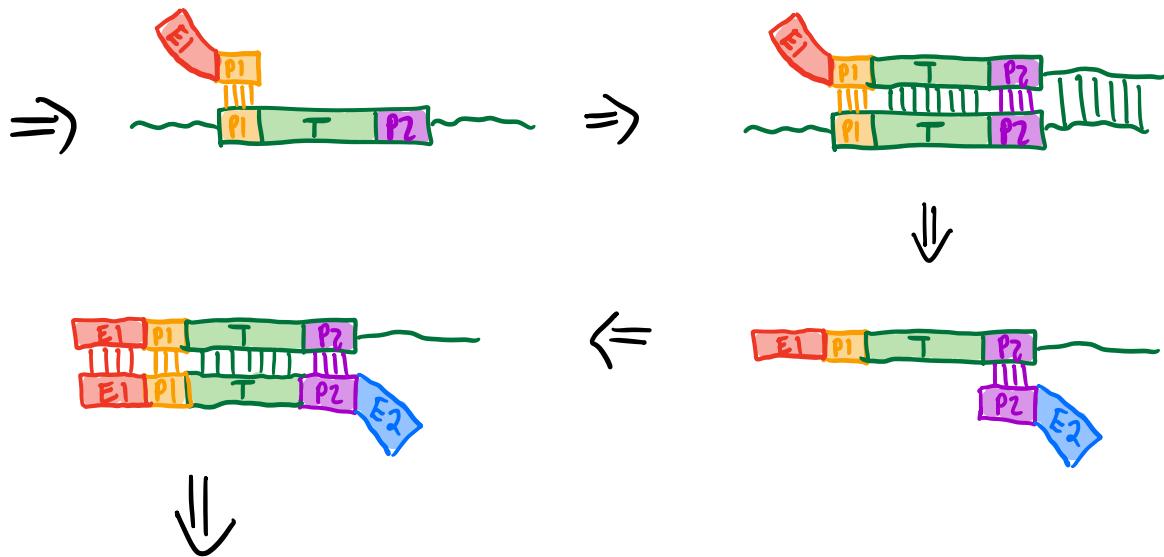
\Rightarrow exponential amplification of sequence!
("amplicon")

⑦ "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...

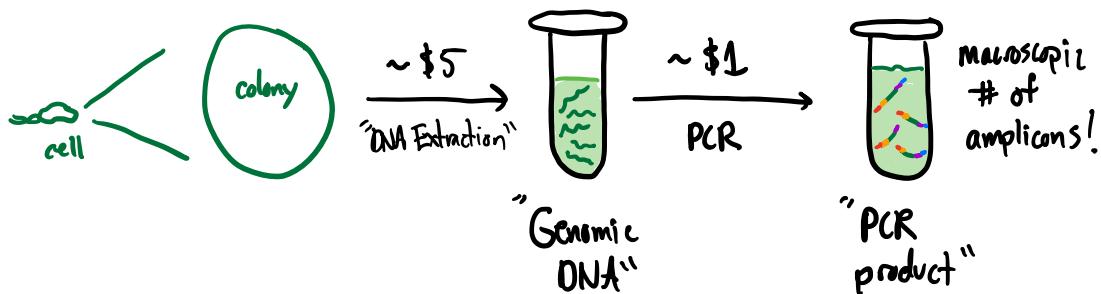
\Rightarrow order modified primers:

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



Amplified sequence: 

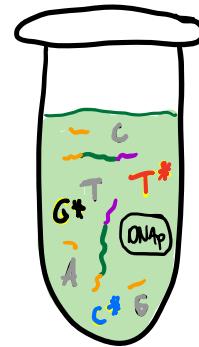
Recap:



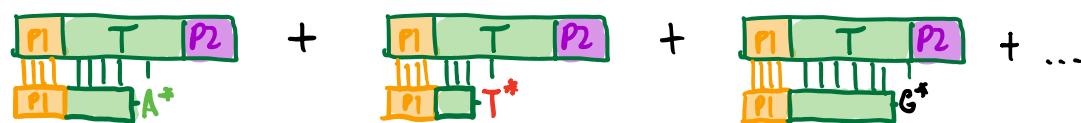
⇒ How do we read out information from PCR products?

Traditional approach: Sanger sequencing
(like 1 extra round of PCR...)

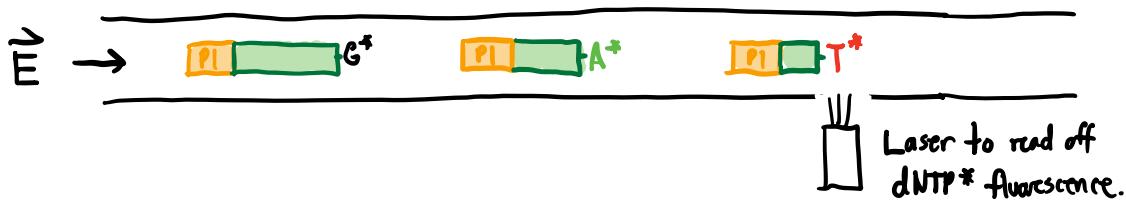
Idea: mix PCR product w/ **P1**, DNAP, dNTPs,
+ special fluorescent dNTPs that block DNAP
(**A***, **G***, **C***, **T***)



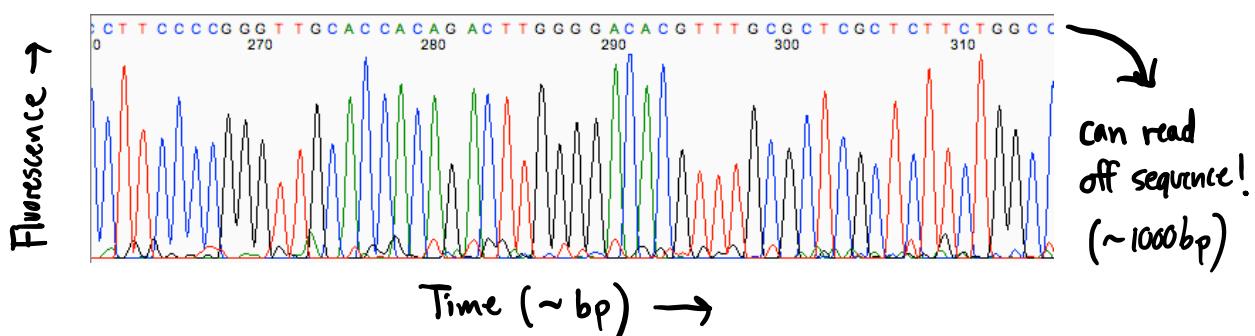
⇒ After 1 round of extension, random mixture of



⇒ Flow in electric field (shorter fragments move faster)



⇓ "chromato-gram"

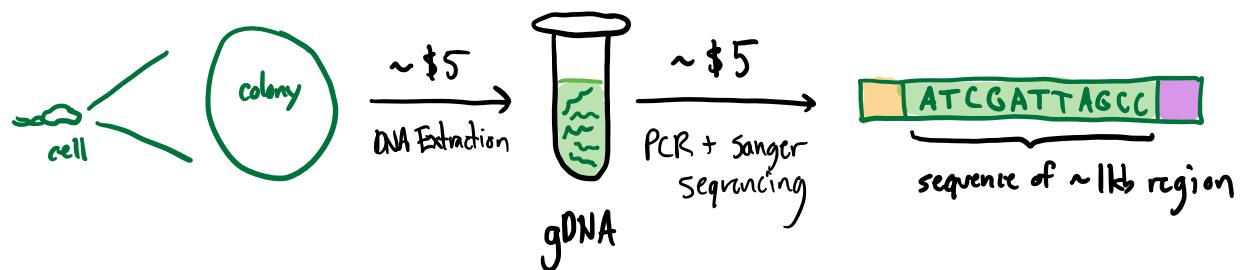


\Rightarrow costs $\sim \$5$ (send away overnight:



)

\Rightarrow so have seen how to go from:



\Rightarrow in Problem 1 of PSET 2, Lang & Murray used sanger seq to sequence URA3 genes in ~ 300 yeast colonies

\Rightarrow \$1500 total

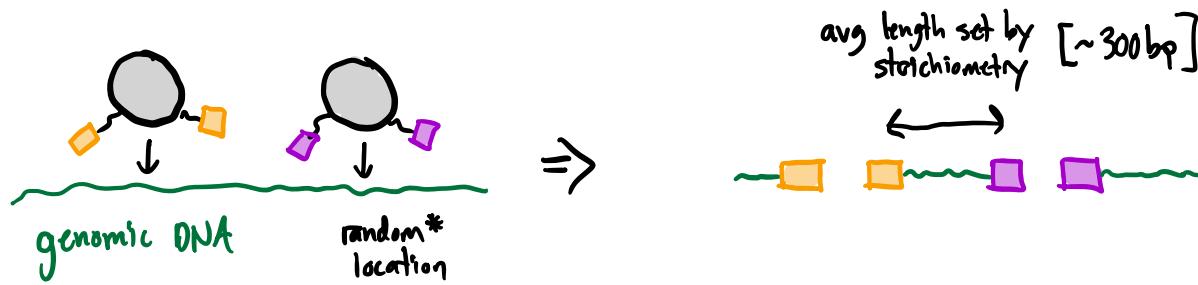
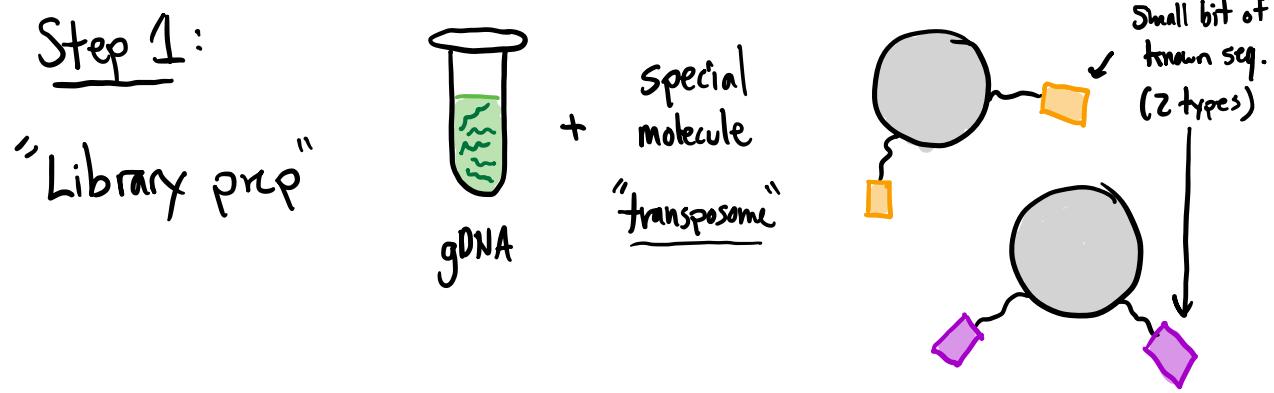
\Rightarrow But expensive to sequence whole genomes!

e.g. E. coli: 1 clone = $10^6 \sim 10^3$ sanger seq runs $\rightarrow \$5k$

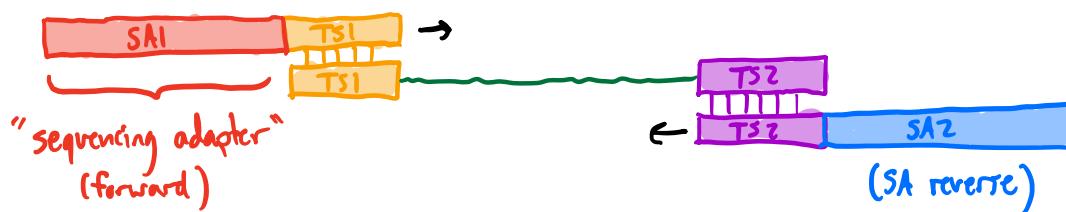
e.g. Humans: 1 person = 10^9 bp $\Rightarrow 10^6$ sanger seq runs $\sim \$5M$

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

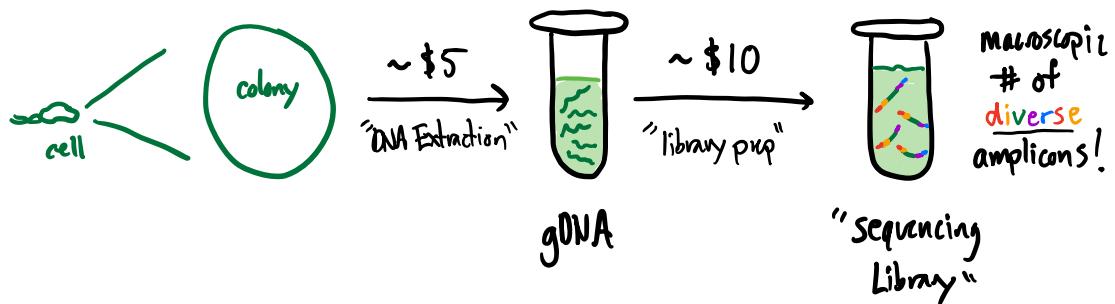
⇒ same idea, but higher throughput!



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

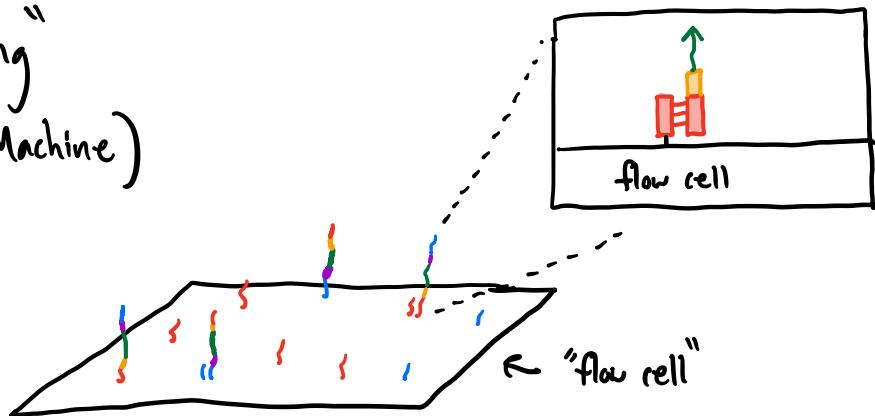


⇒ This gives us:



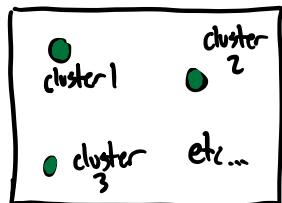
Step 3: "Sequencing"
(Illumina Machine)

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

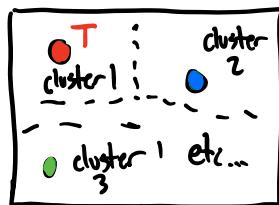
top down view:



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

⇒ if take picture
of flow cell:

:
G | G
T | C
A | T



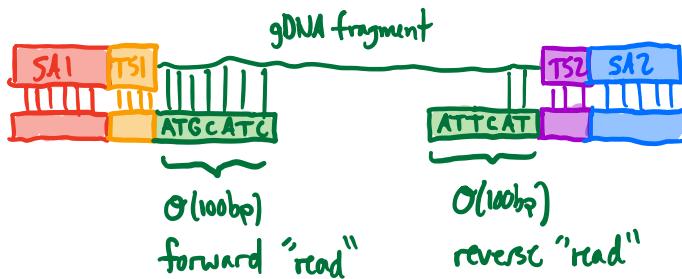
⇒ different colors
= nucleic acid
@ first position

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

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

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

Net result:

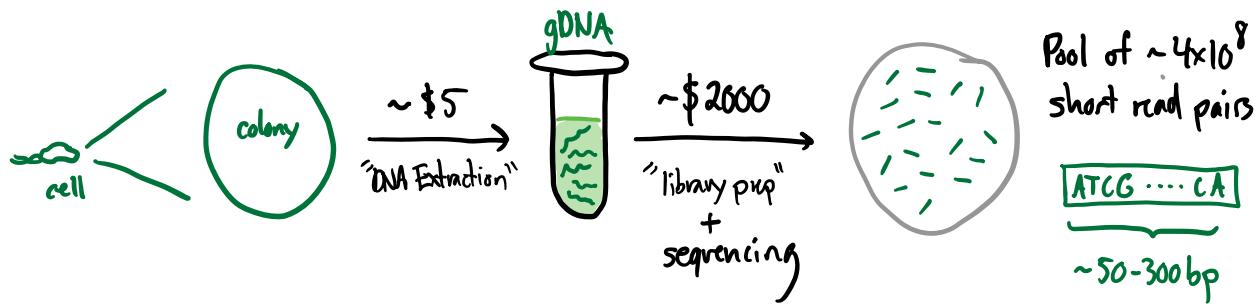


⇒ get $\Theta(100\text{ bp})$ read out from
each end of single DNA fragment

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

\Rightarrow $\sim 4 \times 10^8$ read pairs in ~few days for $\sim \$2000$
(catch: can't do smaller batches)

Recap: Now have method for going from:

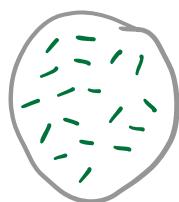


Next: What can we do with this kind of data?

\Rightarrow need to put puzzle back together... 2 main methods

① "De novo Genome Assembly"

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



Pool of $\sim 4 \times 10^8$
short read pairs

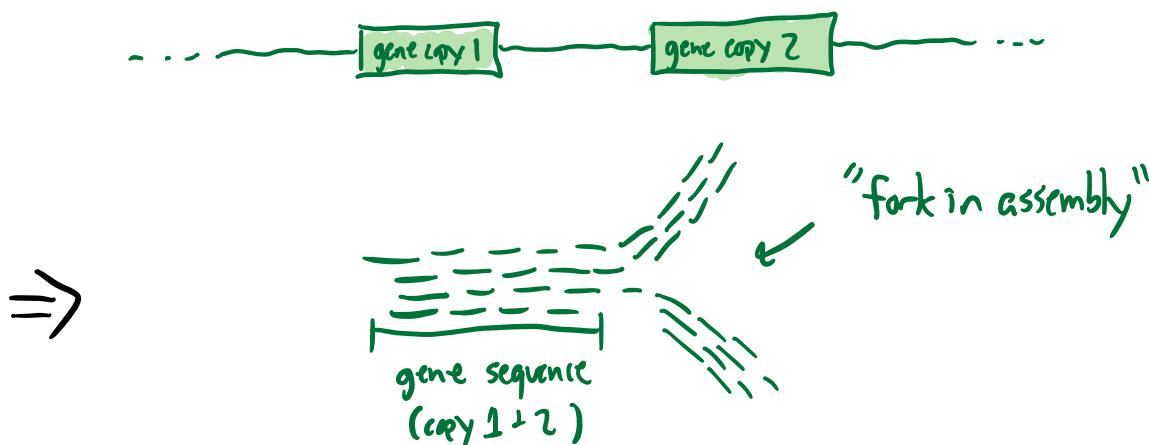
 $\sim 50\text{-}300\text{ bp}$

Idea: look for reads that partially overlap

 $\sim 20\text{ bp}$

\Rightarrow simple in principle, but lots of corner cases ...

\Rightarrow e.g. what if 2 regions of genome are identical
for $\geq 100\text{ bp}$ (or length of read?)

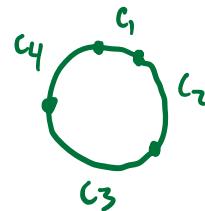


\Rightarrow for these + other reasons, assembly typically yields collection of discrete contigs,

each $\sim 10^3 - 10^5$ bp long : 

\Rightarrow much harder (+ manual effort)

to "finish" into complete genome



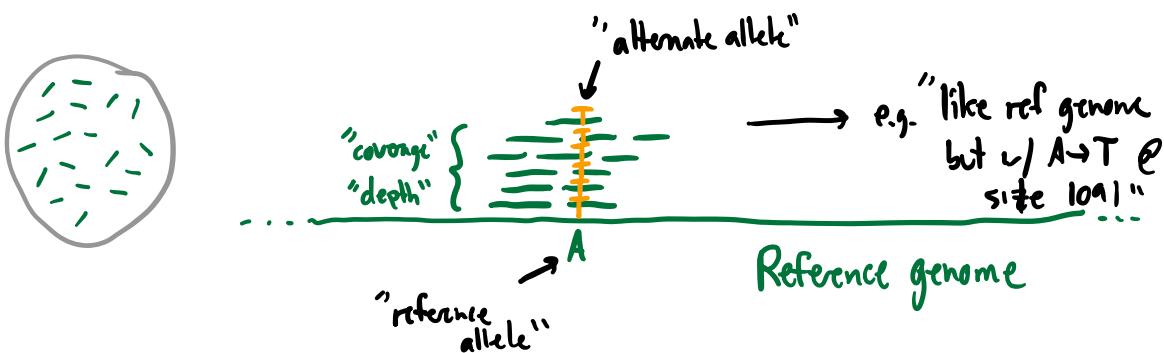
+ HUGE memory requirements ($\sim 32\text{Gb} - 1\text{Tb}$ depending on L)
(since need to compare all* pairs of reads to each other...)

② Alignment of reads to Reference Genome

⇒ if already have assembled genome from related strain...

can align reads to best matching place or look for changes

(common programs: Bowtie2, BWA-MEM, + samtools)



Major wrinkle: sequencing errors ($A \rightarrow T$ by chance, "shot noise")

where could these errors come from?

① initial grow up: ⇒ but happen w/ low rate ($\mu = 10^{-10} / \text{bp/gen}$)

② library prep & PCR



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

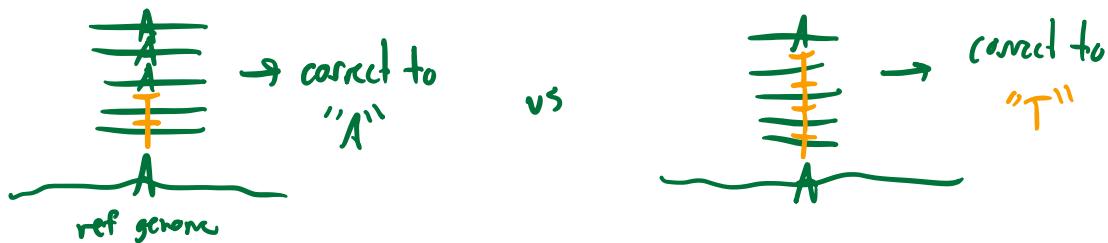
③ Errors on sequencing machine \Rightarrow estimated $\sim 10^{-3}$ / bp ↑
larger
(clustering PCR errors)
wrong fluorescent base... * but can vary from site to site
(up to 10^{-2} sometimes)

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

\Rightarrow low rate, but $L \gg 1$!

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

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



\Rightarrow higher coverage is helpful... how much?

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

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

$$\approx \exp\left(\log L - \frac{\bar{D}}{2} \left[\log\left(\frac{1}{2\text{Perr}}\right) - 1\right]\right] \lesssim \Theta(1)$$

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

\Rightarrow so need coverage of $\gtrsim 10x$ coverage to detect single muts.

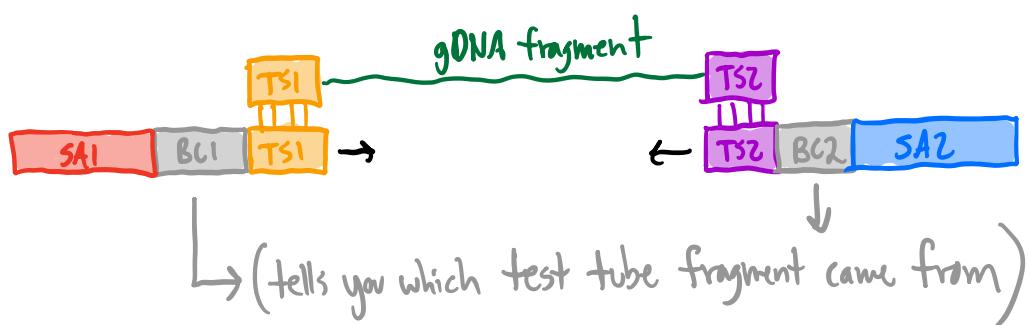
\Rightarrow How many reads is this?

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

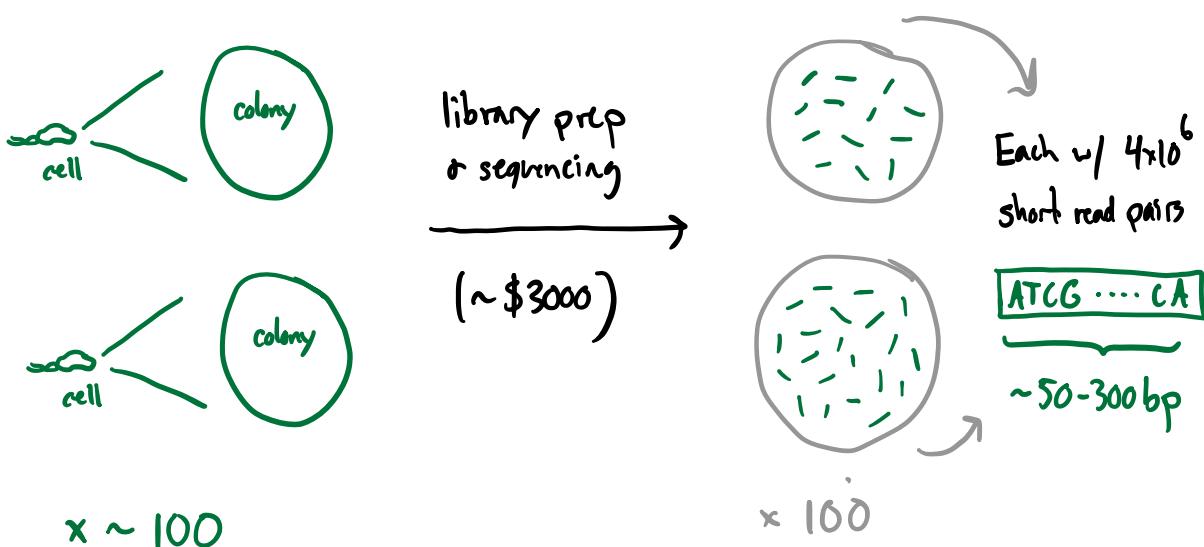
\Rightarrow but a single run of Illumina seq produces 4×10^8 reads

\Rightarrow overkill! (wastes money)

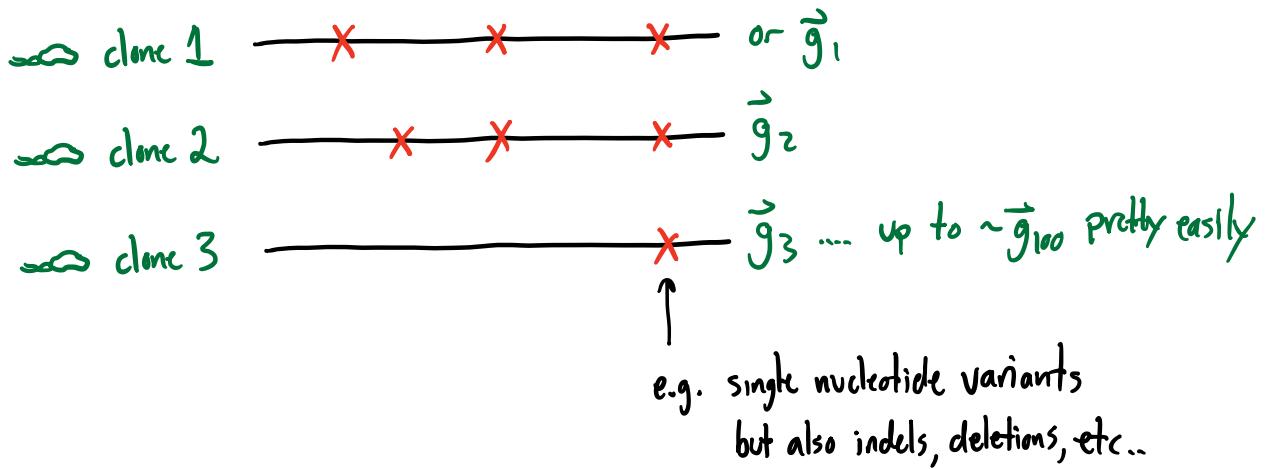
Solution: "multiplex": Add sample specific "barcode" sequence during library prep step



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



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



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

\Rightarrow 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)

$$\Rightarrow \text{then } \Pr[\{n_{\vec{g}}\} \mid n, \{f(\vec{g})\}] = \begin{matrix} \text{multinomial} \\ \text{distribution} \end{matrix}$$

\uparrow
 total # clones sampled

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

\Rightarrow Genotype space is huge!

\Rightarrow often coarse-grain to summary statistics.

$\left\{ \begin{array}{l} 1 \text{ if mut@ site} \\ 0 \text{ else} \end{array} \right.$

e.g. $n_e = \# \text{ individuals w/ mutation @ site } l = \sum_g g_e n_g$

$$\Rightarrow \text{can show } \Pr[n_e | n, \{f(\vec{g})\}] = \text{Binomial}(n, f_e)$$
$$f_e \equiv \sum_g g_e f(\vec{g})$$

e.g. total # of mutations separating 2 genomes

\Rightarrow Since depends on genome length, often normalized by L:

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

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:

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

$\in \{0,1\}$

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

$$= \frac{1}{L} \sum_{e=1}^L 2f_e (1-f_e)$$

Remember:
 $f_e \equiv \sum_{\vec{g}} g_e f(\vec{g})$

\Rightarrow averaging over f_e :

$$\langle \pi \rangle = \frac{1}{L} \sum_{e=1}^L \langle 2f_e (1-f_e) \rangle = \frac{1}{L} \sum_{e=1}^L \int 2f_e (1-f_e) p(f_e) df_e$$

e.g. if genome is collection of neutral sites

$$\Rightarrow \rho(f_e) \approx \frac{2Ne^N}{f_e} \quad (\text{quasi-stationary dist'n})$$

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

↳ fit $N_c \equiv \frac{\langle \pi \rangle}{2N} = \frac{10^{-3}}{2 \times 10^{-8}}$

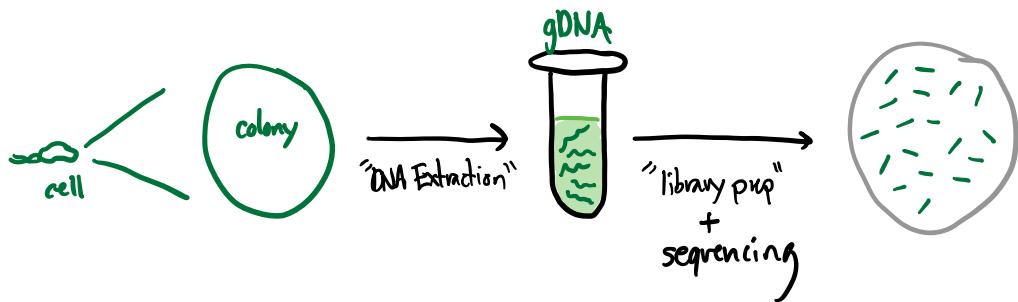
→ population can only
spread out so far
in sequence space!

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

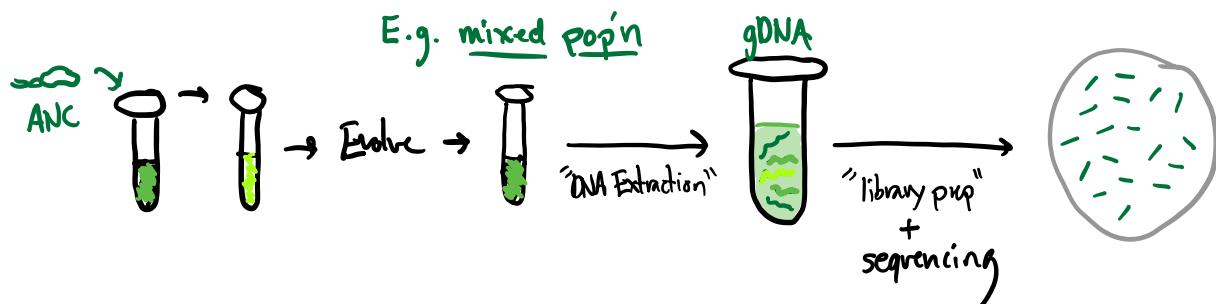
\Rightarrow correlations between g_e " $g_{e'}$

\Rightarrow will see more later!

So far, have focused on clones



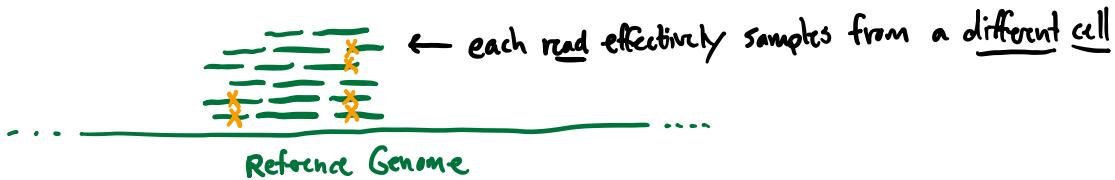
⇒ lots of other things we could put in our library prep...



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

⇒ in this case, assembly very hard!

⇒ Payoff comes from reference mapping:



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

$$\Rightarrow \Pr[A_e | D_e, \{f(g)\}] = \text{Binomial}(D_e, f_e)$$

↑
total coverage
@ site ℓ

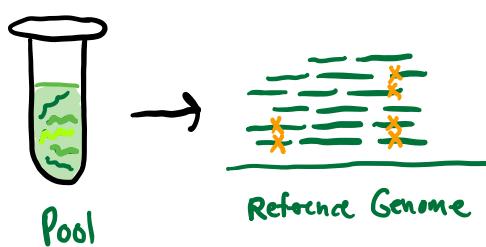
+ sequencing/PCR errors

\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's)

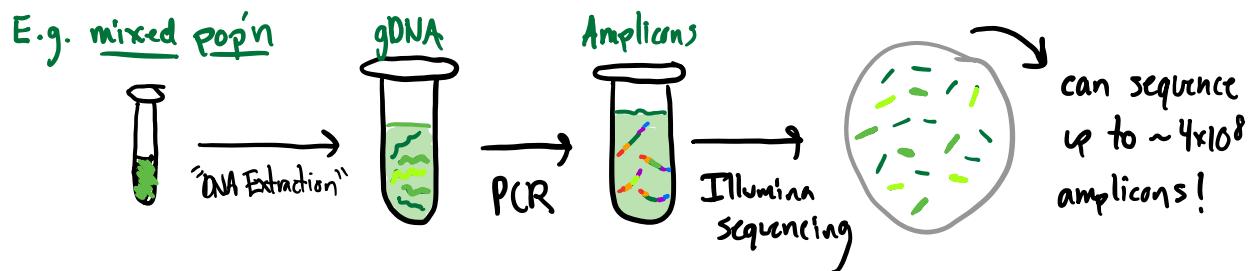
Downsides:



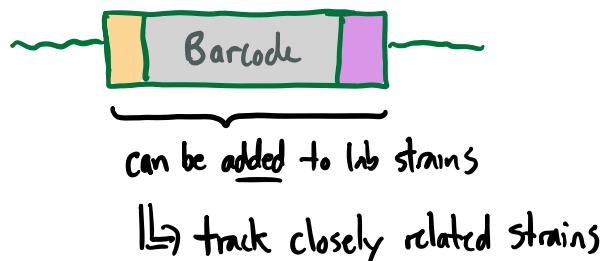
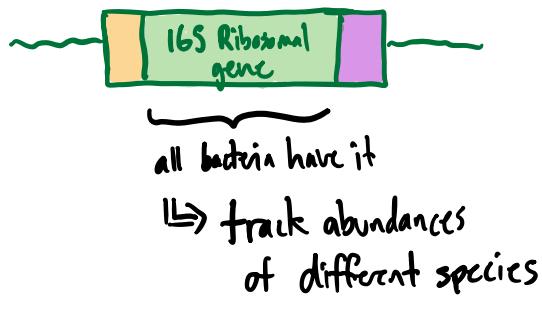
- ① sequencing errors!
- ② which mutations are in same cells? ("linkage information")

e.g. vs

Can also sequence pools of amplicons:

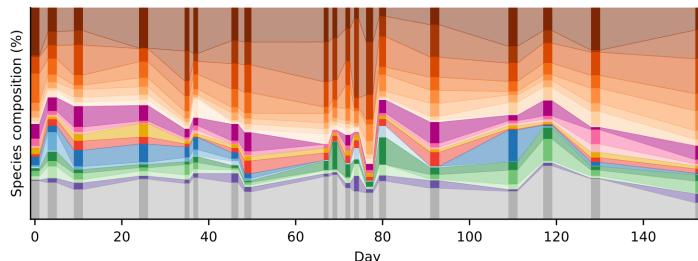


Two common targets:



↓

E.g. 1 person's microbiome over-time



- *Bacteroides vulgatus*
- *Bacteroides coprocota*
- *Bacteroides uniformis*
- *Bacteroides cellulosilyticus*
- *Bacteroides eggerthii*
- *Bacteroides faecis*
- *Bacteroides massiliensis*
- *Bacteroides caccae*
- *Alistipes sp*
- *Alistipes onderdonkii*
- *Alistipes finegoldii*
- *Parabacteroides distasonis*
- *Paraprevotella clara*
- *Butyrivibrio crosstotus*
- *Coprococcus sp*
- *Coprococcus comes*
- *Eubacterium rectale*
- *Eubacterium siraeum*
- *Eubacterium eligens*
- *Phascolarctobacterium sp*
- Other

↓

E.g. Homework Problem