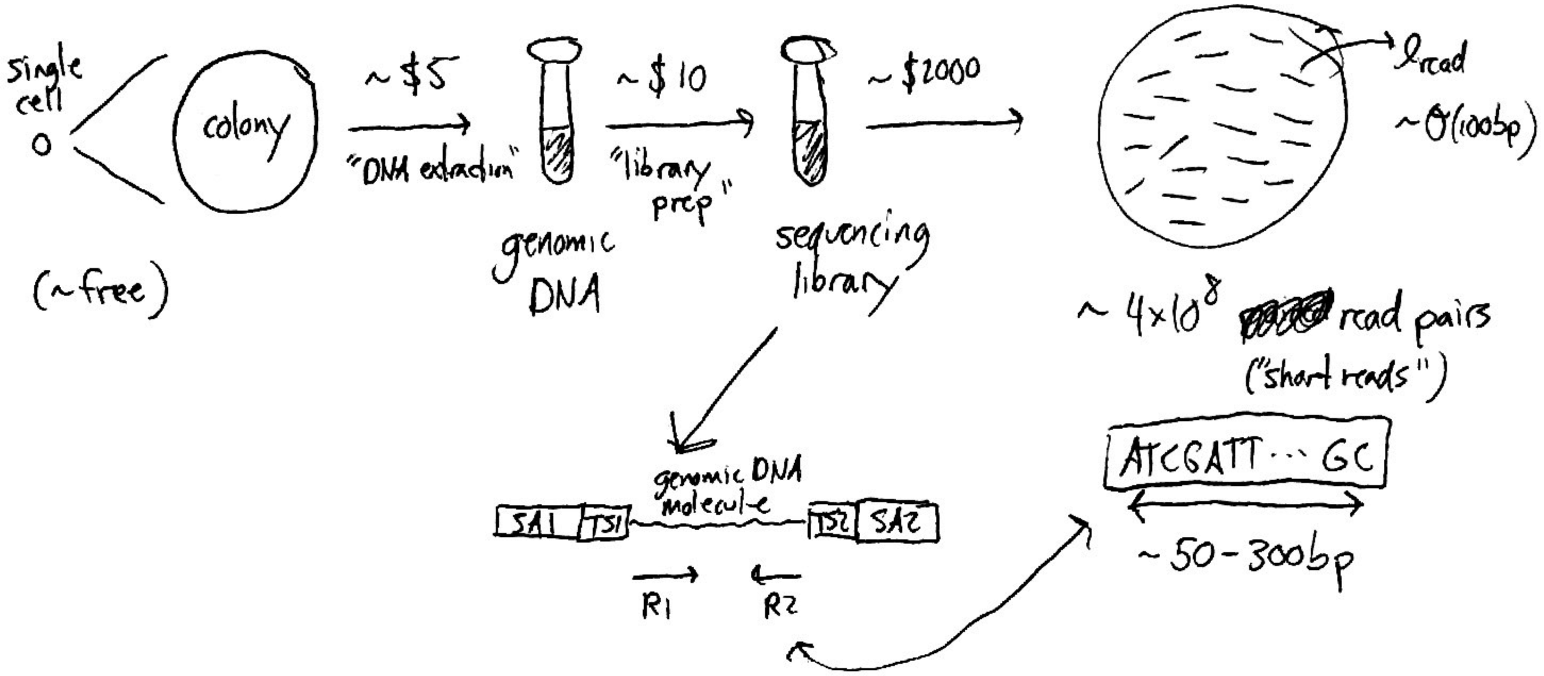
Sequencing & Genomics II

Last time: Next-gen / Illumina sequencing of bacterial isolates / "clones"

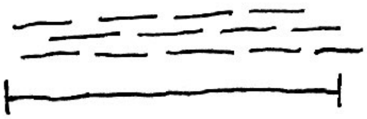


What can we do w/ this kind of data?

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

① De Novo Genome Assembly (common programs: Spades/Velvet/...)

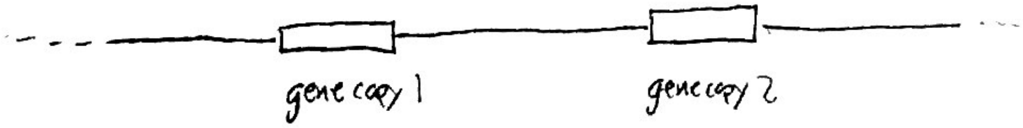
⇒ look for overlapping reads (≳ 20bp)

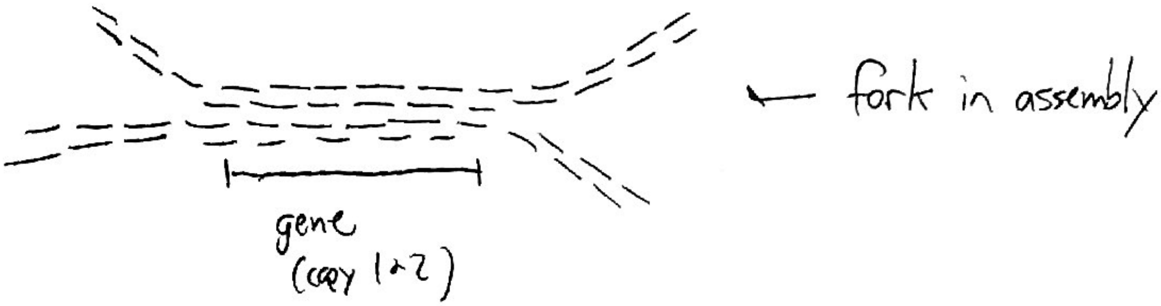


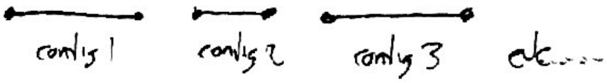
consensus sequence
= "contig"

Simple in principle, but lots of corner cases...

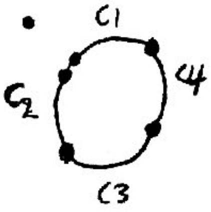
e.g. what if 2 regions of genome are <u>identical</u> for more than 100 bp (or 1 read)?





⟹ for these + other reasons, assembly often results in collections of different <u>contigs</u> 

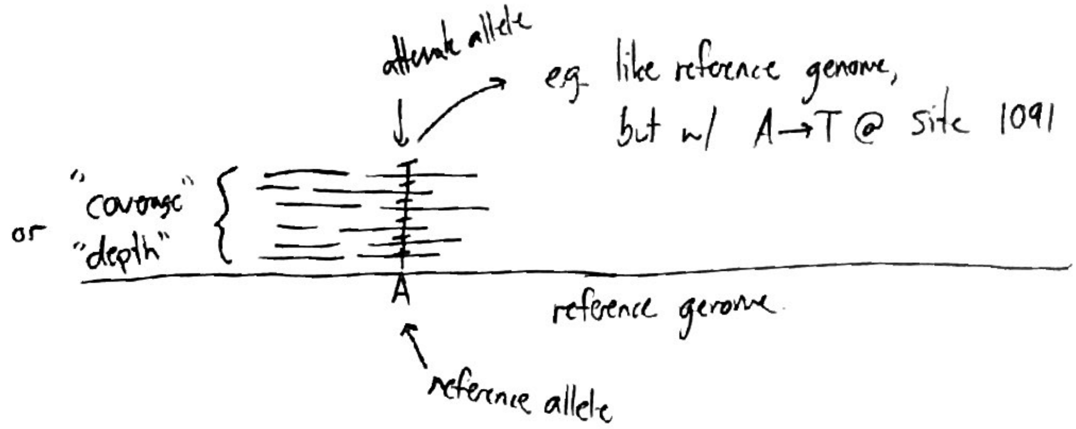
each ~ 1000-10^5 bp long.

↳ much harder (+ manual effort) to "finish" into complete genome: 

+ <u>HUGE</u> memory requirements (~32Gb-1Tb, depending on genome)
since have to compare all* reads to each other ...

2 Alignment of Reads to Reference Genome

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



⇒ this is O(#reads) & much lower memory footprint. (laptop)

⇒ saw that ~20bp sufficient to locate most sites in E. coli, so ~100bp reads are mostly ok.

(like difference between putting puzzle together from scratch, & putting together a puzzle when the completed one is right next to you.)

⇒ still some corner cases (but small fraction of genome)

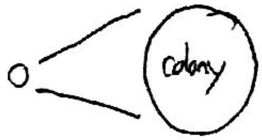
& works best when ref genome is "close" to sample (~ at most 1-2 diffs per read)

One winkle: sequencing errors

⇒ ability to sequence single molecules has its drawbacks:
random molecular errors ("shot noise") can lead to e.g. A→T by chance.

where could these errors come from?

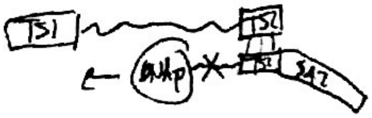
1) initial grow up



e.g. mutation during first division, 2nd, etc.
⇒ Luria-Delbrück process

but happen @ low rate, $\mu \sim 10^{-10}$/bp/gen

2) Library prep PCR



error during early rounds of PCR ("PCR errors")
also Luria-Delbrück-like process.

higher rates, $\sim 10^{-6}$/bp/round $\sim 10^{-4}$ total. ← presumably this too

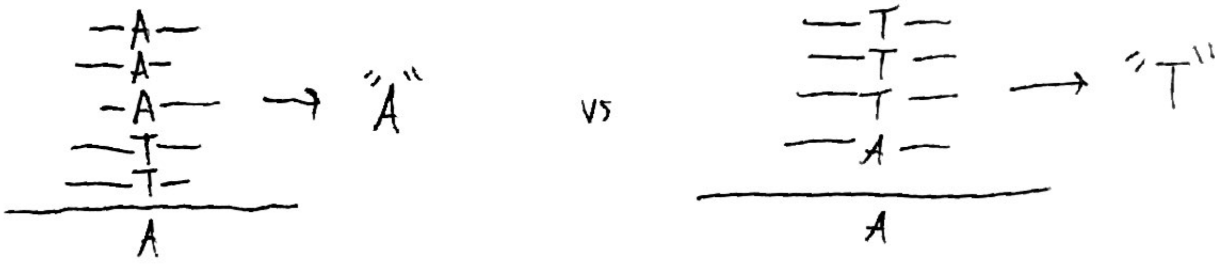
3) Error on sequencing machine
(e.g. cluster generation PCR error
wrong fluorescent base, etc.)

⇒ estimated to be $\sim 10^{-3}$/bp, but
can vary a lot from site-to-site.
⇒ dominant source of error

⇒ still low rate per site, but lots of sites in genome, so expect
@ least $L \times p_{err} \sim 10^{3}$ in Mb-sized genome.

(big problem for detecting single mutations...)

Fortunately, can correct many of these errors by taking <u>consensus</u> across independent reads covering same site:



⟹ higher coverage is helpful.

⟹ how much coverage necessary before we expect $\lesssim 1$ consensus errors in whole genome?

error in consensus require $\geq 1/2$ of all reads to have an error.

⟹ if coverage is Poisson process w/ mean $\bar{D}$

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

$$\#\text{ errors} = L \times \Pr(\text{consensus error}) = \exp\left[\log L + \frac{\bar{D}}{2}\log(P_{err}\bar{D}) - P_{err}\bar{D} - \frac{\bar{D}}{2}\log\left(\frac{\bar{D}}{2}\right) + \bar{D}/2\right]$$

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

⟹ crosses 1 when $\bar{D} \approx \frac{2\log L}{\log(\frac{1}{2P_{err}})-1} \approx \begin{cases} 5 \text{ if } L\approx 10^6, P_{err}\sim 10^{-3} \rightarrow 10 \text{ if } P_{err}\approx 10^{-2} \\ 8 \text{ if } L\approx 10^9 \rightarrow \end{cases}$

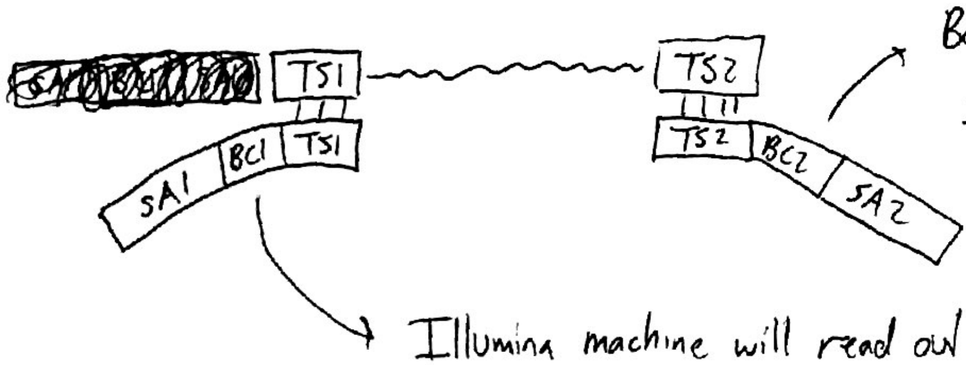
so need coverage $\gtrsim 10$ to eliminate errors (though still weird locs)

⇒ how many reads is this? 10×10^6 bp sequenced. $= 10^7$ bp.

→ 10^5 reads per genome.

⇒ but single run generates ~ 10^8 reads! overkill. (waste money)

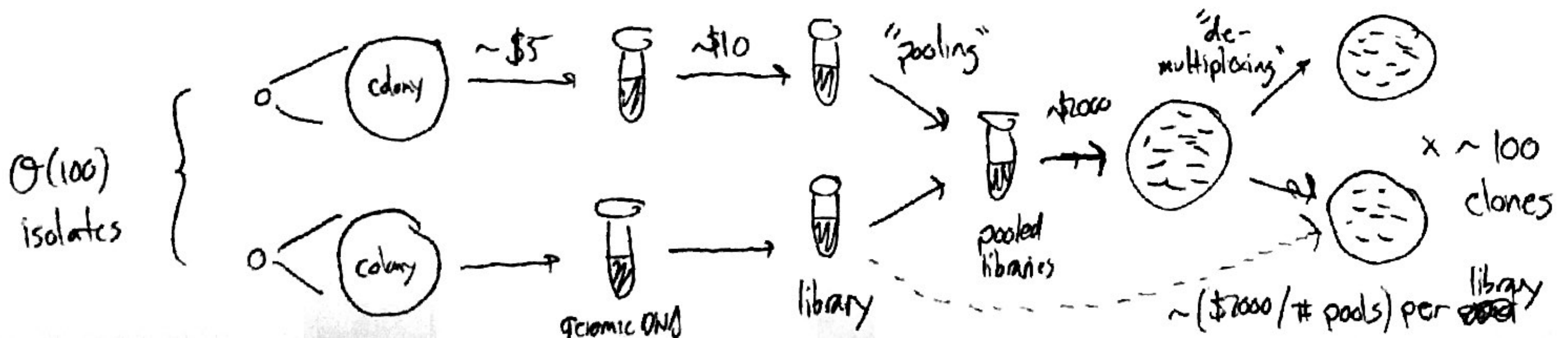
Solution: "multiplexing" w/ DNA barcodes. can design sequencing adapter w/ special sequence of letters (barcode) used to distinguish different samples.



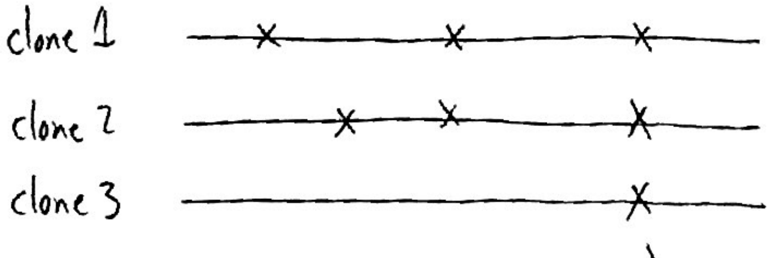
Barcode separate for each library prep. sample.

Illumina machine will read out barcode as well as reads (for "free")

⇒ can then tell which library each read pair came from.



⟹ after aligning reads, detecting "true" mutations, get sequences



or $\vec{g}_1$, $\vec{g}_2$, $\vec{g}_3$ … up to $\vec{g}_{100}$ pretty easily.

e.g. single nucleotide variants (but also indels, deletions, etc.)

How are these related to the distribution of genotypes in our population?

⟹ if clones are sampled randomly, then

$$\Pr\left[\{n_{\vec{g}}\} \,\middle|\, n, \{f(\vec{g})\}\right] = \text{multinomial dist'n} \propto \prod_{\vec{g}} f(\vec{g})^{n_{\vec{g}}} \frac{1}{n_{\vec{g}}!}$$

where $f(\vec{g})$ = fraction of population w/ genotype $\vec{g}$ (random from <u>evolution</u>)

$n_{\vec{g}}$ = # of clones in sample w/ genotype $\vec{g}$ (random from <u>sampling</u>)

⟹ genotype space is huge, so often reduce to <u>summary statistics</u>.

e.g. # of mutations separating 2 genomes

Since depends on length of genome, often normalize by L:

$$\frac{\#\text{ mutations between 2 randomly sampled clones}}{L} = \begin{cases} \text{"heterozygosity" } (\pi) \text{ if from same pop'n} \\ \text{"divergence" } (d) \text{ if from diff. "species"} \\ \qquad (\text{isolated pop'ns}) \end{cases}$$

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

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

heterozygosity among E.coli from different humans is $\sim 10^{-2}$

to relate π to genotype freqs, $f(\vec{g})$, note that

$$\pi = \frac{1}{L}\sum_{\ell=1}^{L}\left[\left[\underbrace{g_{1\ell}}_{\in\{0,1\}}(1-g_{2\ell})\right]+\left[(1-g_{1\ell})g_{2\ell}\right]\right] \longrightarrow$$

two ways of sampling genomes that differ at site ℓ.

so on average,

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

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

fraction of population w/ mutation @ site ℓ.

the f_ℓ's are themselves random (from evolution)

so technically, have only calculated $\langle \pi | \{f_\ell\} \rangle$

averaging over these, we have

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

If genome is <u>collection of neutral sites</u>, then $p(f_\ell) = p(f) = \frac{2N_e\mu}{f}$, and

$$\langle \pi \rangle = \langle 2f(1-f) \rangle = \int 2f(1-f) \cdot \frac{2N\mu}{f} \, df = 2N_e\mu$$

$\Rightarrow$ thus, connection between $\langle \pi \rangle$ and N_e

(sometimes people even call $\langle \pi \rangle$ N_e, as if N_e were an empirical property of the data. this is really bad, and we should stop doing it)

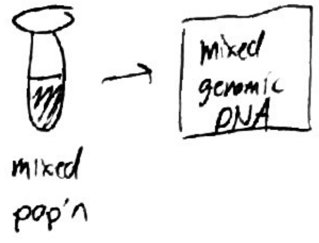
Note that Variance of π is much more complicated, since it depends on correlations between $g_{1\ell}$ and $g_{1\ell'}$ @ different sites.

However, related summaries that are <u>linear</u> in sites, but involve bigger samples can still be calculated.

(will see examples later on)

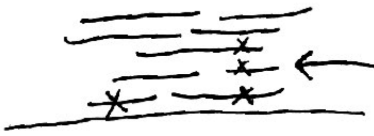
So far, have focused on sequencing clones, but lots of other things you could put in your library prep.

⇒ common one is population of bacteria ("pooled sequencing" or "metagenomic sequencing")



⇒ in this case, assembly can be very hard (since dealing w/ mixture of different genomes)

⇒ payoff comes from reference mapping:



for typical coverages & input pop'ns, each read samples from a different cell in pop'n.

e.g. # of reads w/ mutation @ site ℓ

is $Pr(A \mid D, f_\ell) \approx \text{Binomial}(D, f_\ell) + \text{sequencing error}$

e.g. if want to calculate $\langle \pi \mid f_\ell \rangle$,

$$\langle \pi \mid f_\ell \rangle = \left\langle \frac{A(D-A)}{\binom{D}{2}} \right\rangle + \text{sequencing error} = 2f_\ell(1-f_\ell) + \text{seq error}$$

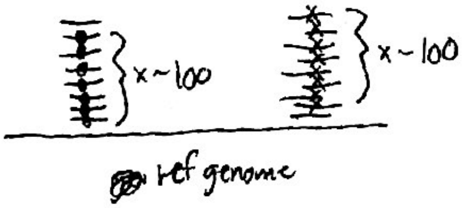
⇒ Since can get ~100x coverage for ~100 E. coli genomes in 1 run of Illumina seq, can effectively sample ~100 clones ~100x more cheaply by sequencing pools. (*much* cheaper way to measure $\langle\pi\rangle$ or to track frequencies of Individual mutations.)

Downsides: also *much* harder to distinguish low freq mutations from sequencing errors. $\langle\pi\rangle \gtrsim p_{err} \sim 10^{-3}$

If $p_{err} \sim 10^{-3}$, not even possible theoretically to measure freqs below this (even w/ infinite coverage) unless you do fancier things.

⇒ also lose information about which mutations are in same cells ("linkage information") unless you catch them on same sequencing read.

⇒ Sometimes can make progress w/ "pigeonhole principle", e.g.:



⇒ probably many cells w/ —•×—, since genotypes must add to 1:

$$1 = f(-) + f(-\bullet-) + f(-\times-) + f(-\bullet\times-) = f(-) - \underbrace{f(-\bullet\times-)}_{\text{Must be} > 0!} + \underbrace{f_\bullet + f_x}_{\gg 1}$$

⇒ Show example data from Lenski's LTEE

Don't have to sequence mixed population of same species

⇒ nothing keeping you from extracting DNA from mixed community of bacteria in native community (e.g. fecal samples in gut microbiome)

("shotgun metagenomic sequencing")

⇒ in this case, since don't have to grow the bacteria, can work even when bacteria hard to grow in lab / frozen / dead / etc.

"culture independent sequencing" → if genomes in sample sufficiently close (or sufficiently far), can use for de novo assembly to discover new bacteria/genes

often contrasted w/ amplicon sequencing (i.e. add Illumina sequencing adapters to PCR product)

└→ commonly used target in microbiology is 16S Ribosomal RNA gene. all bacteria have it, & there are a few regions that are highly conserved across tree of bactrial life → good targets for primers.

⇒ amplicon metagenomic sequencing ≈ distribution of species* abundances in sample.