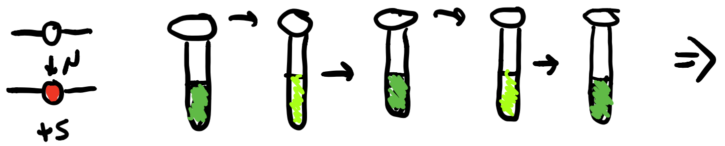
Announcements: Half way done!

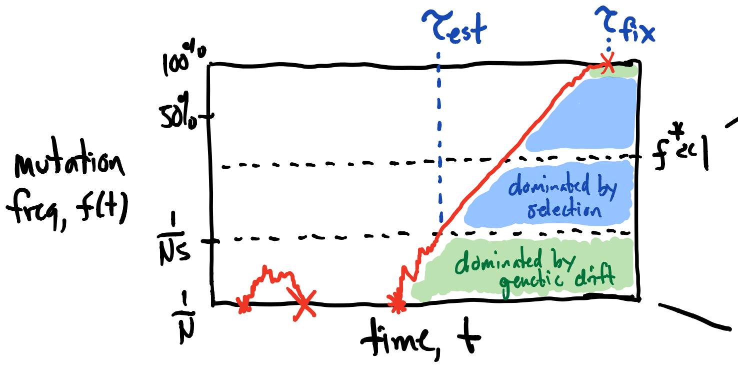
Recap: Evolution at a single genetic locus



$$\Rightarrow \frac{\partial f}{\partial t} = \overbrace{sf(1-f)}^{\text{selection}} + \overbrace{\sqrt{\frac{f(1-f)}{N}}\,\eta(t)}^{\text{genetic drift}}$$

$$+\left[\mu(1-f) - \nu f\right]\Big\}\text{ mutation}$$

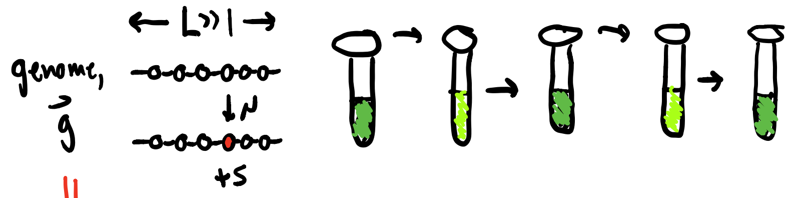
$$\Downarrow$$



"Linear Branching Process"

$$\frac{\partial f}{\partial t} = \mu + sf + \sqrt{\frac{f}{N}}\,\eta(t)$$

Where we're headed: Evolution of longer genomes!

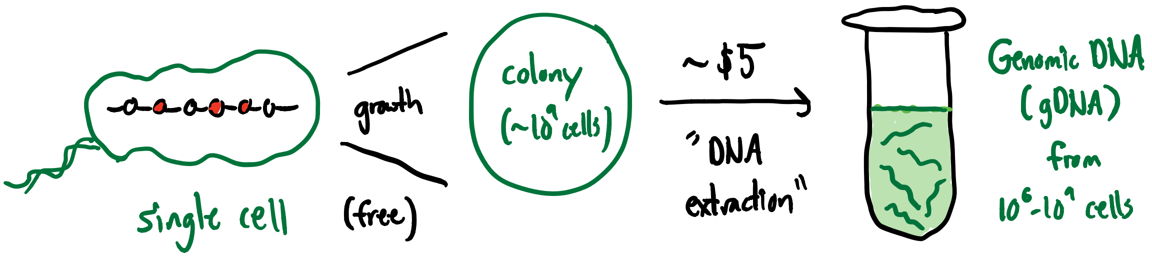


$$\frac{\partial f(\vec{g})}{\partial t} = ???$$

$\Downarrow$

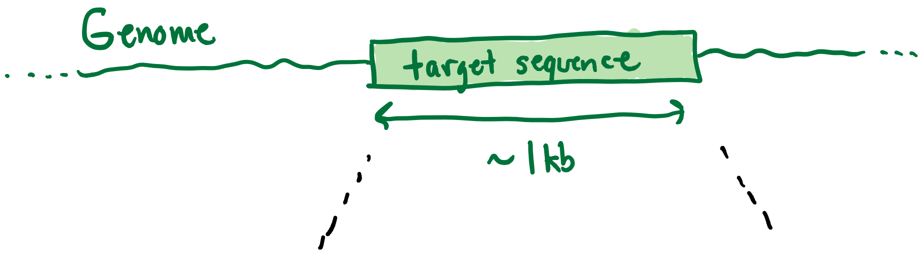
Last time: How do we measure information in single DNA molecules?

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



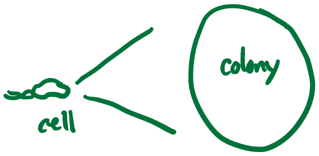
__Problem__: $L = 10^6$ sized genomes too difficult do measure directly

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



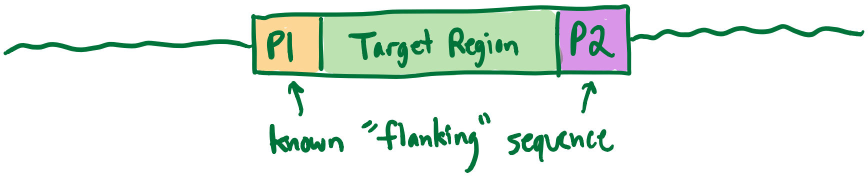
How to get a __macroscopic__ amount of just this region?

__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

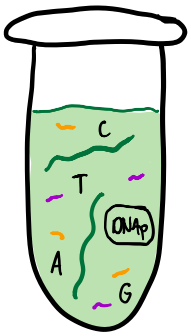


① can have company synthesize "primers" (short sequences of ~20bp) that correspond to P1 & P2. (~ \$0.30/bp for ~ 10^3 reactions)

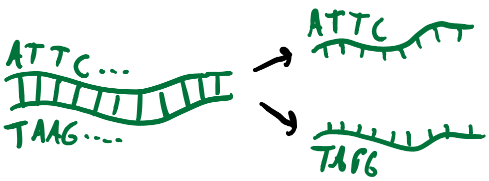
② mix w/ DNA polymerase,

dNTP (free A's, C's, T's, G's)

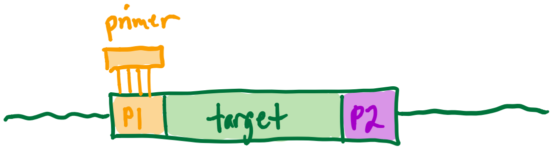
+ your genomic DNA.



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



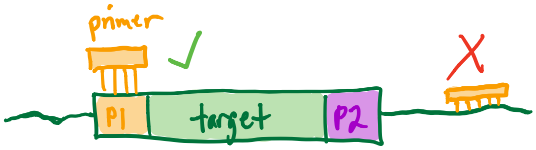
④ Now cool ~60°C sample so that primers "anneal" to input DNA



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

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

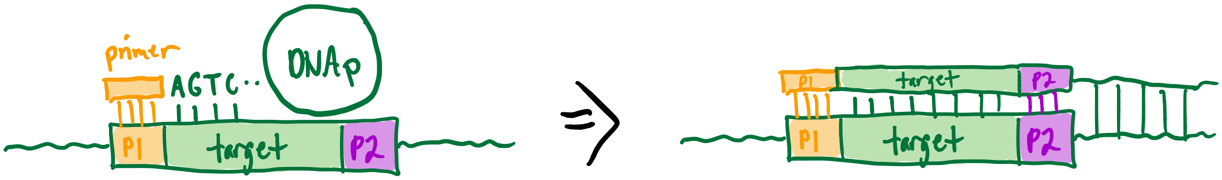
⇒ want primers to bind to known region, but not anywhere else!



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

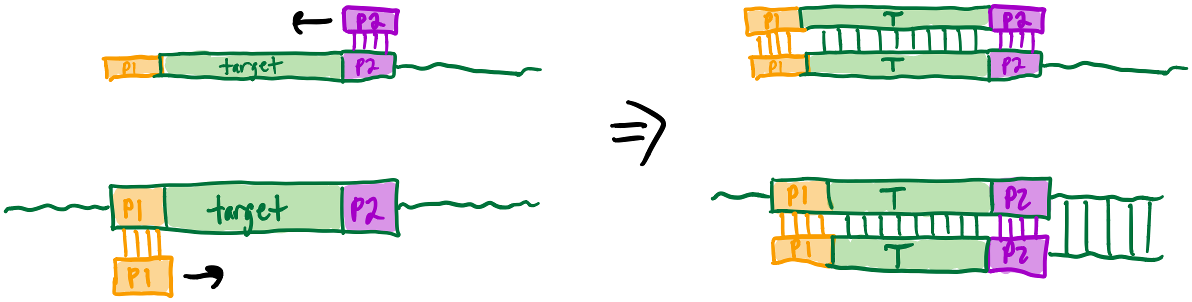
(more generally ⇒ information theory argument, see p.5 of notes...)

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

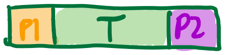


("extension phase")

⑤ Melt, anneal, & extend again:

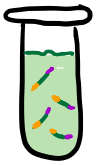


⑥ Repeat for K~20-30 cycles...

⇒ exponential amplification of  sequence!

("amplicon")

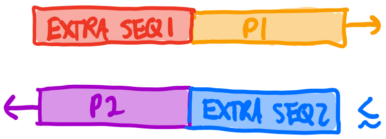
⑦ "clean up" to remove leftover primers, etc. ⇒



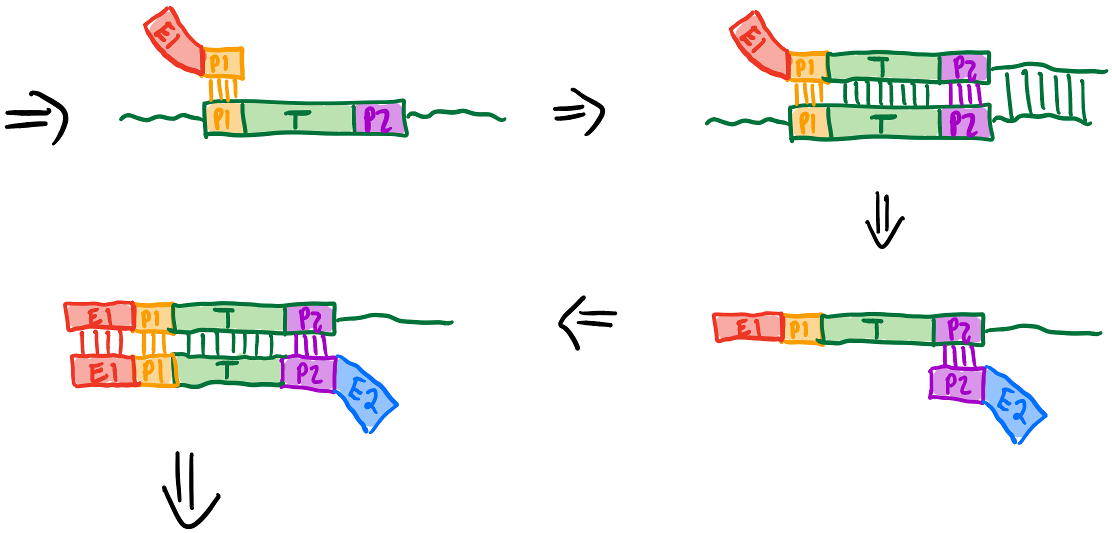
macroscopic # of amplicons!

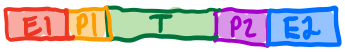
⇒ Note: can also use PCR to add extra bit of DNA sequence to your target region...

⇒ order modified primers:

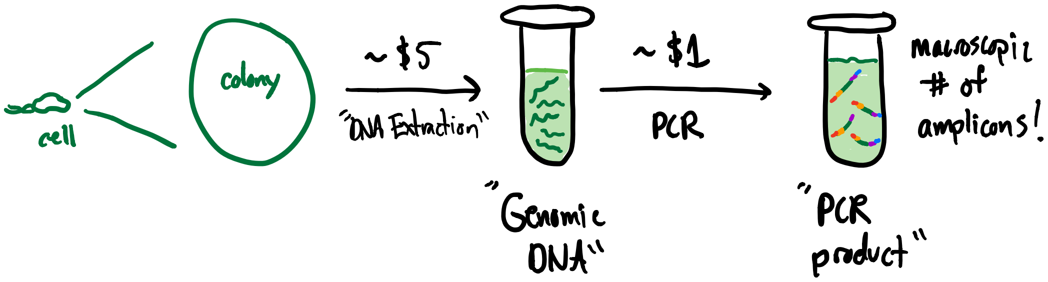


≲ $O(100bp)$



Amplified sequence: 

Recap:

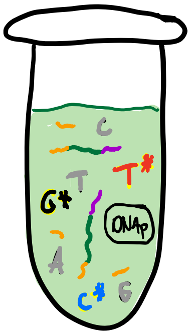


⇒ How do we read out information from PCR products?

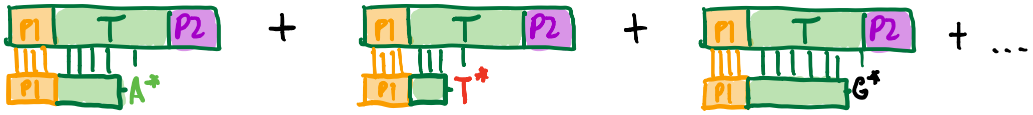
Traditional approach: Sanger sequencing

(like 1 extra round of PCR...)

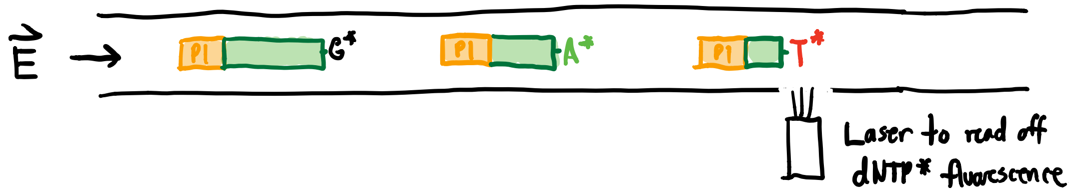
Idea: mix PCR product w/ P1, DNAp, dNTPs,
+ special fluorescent dNTPs that block DNAp
(A*, G*, C*, T*)



⇒ After 1 round of extension, random mixture of

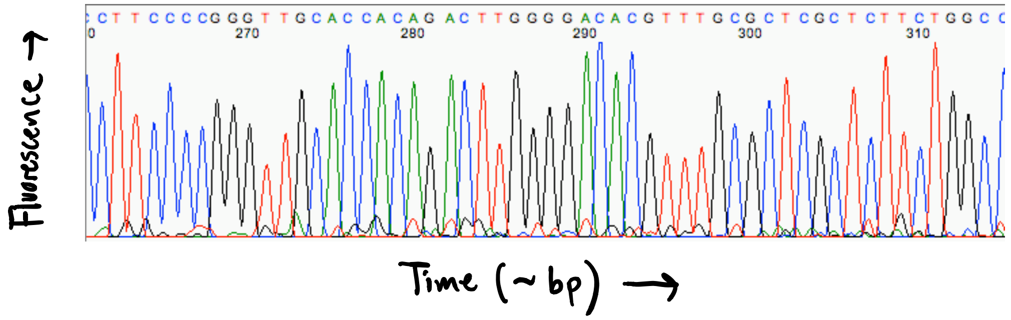


⇒ Flow in electic field (shorter fragments move faster)



Laser to read off dNTP* fluorescence.

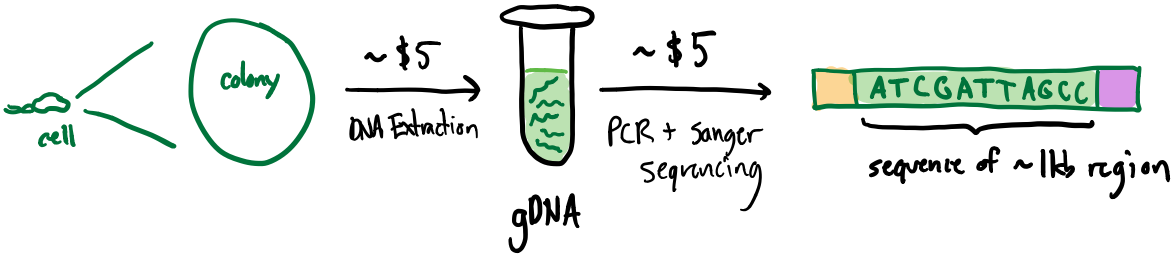
⇓ "chromato-gram"



can read off sequence! (~1000bp)

⇒ costs ~\$5 (send away overnight:
PLEASE PLACE SAMPLES HERE
GENEWIZ
Solid science. Superior service.
Albert Einstein College Of Medicine
Golding Building
DNA Sample Drop-Off Station
SAMPLE DROP CUT OFF TIME 5:00 PM
www.genewiz.com | Toll Free: 1-877-GENEWIZ
NOTICE:
DROP BOX SHIPPING IS ONLY FOR ROOM TEMPERATURE SAMPLES
-Samples packed with dry ice or ice packs should be sent by the researcher directly to GENEWIZ via overnight delivery service
)

⇒ so have seen how to go from:



⇒ in Problem 1 of PSET 2, Lloyd 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 ~ 10^3 sanger seq runs → \$5k

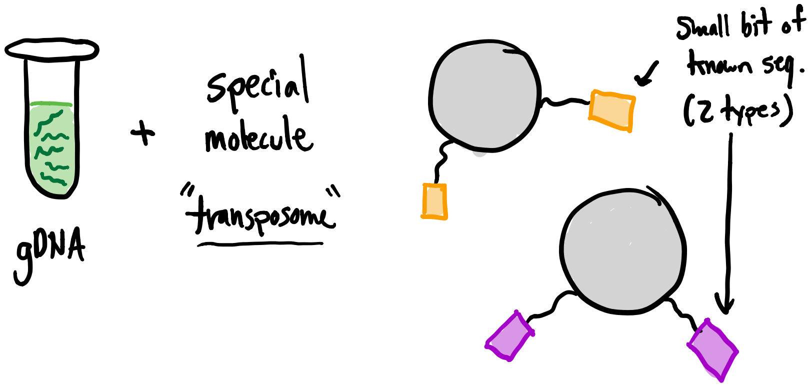
e.g. Humans: 1 person = 10^9 bp ⇒ 10^6 sanger seq runs ~ \$5M

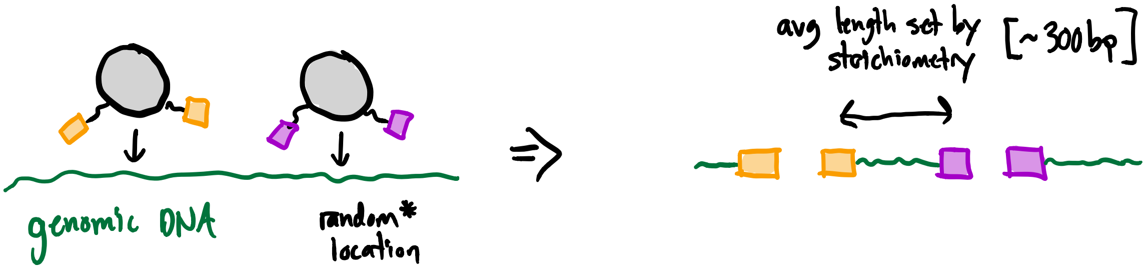
⇒ Now things are much cheaper w/ Next-gen sequencing
("Illumina sequencing")

⇒ same idea, but higher throughput!

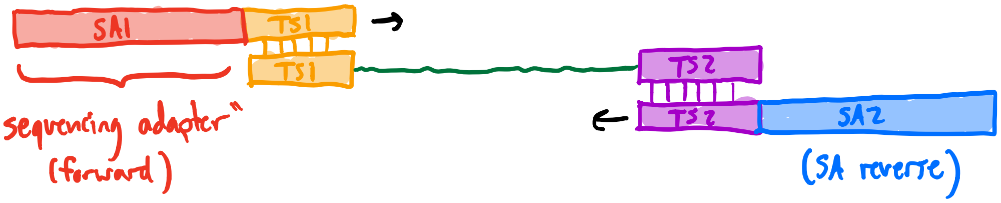
Step 1:

"Library prep"

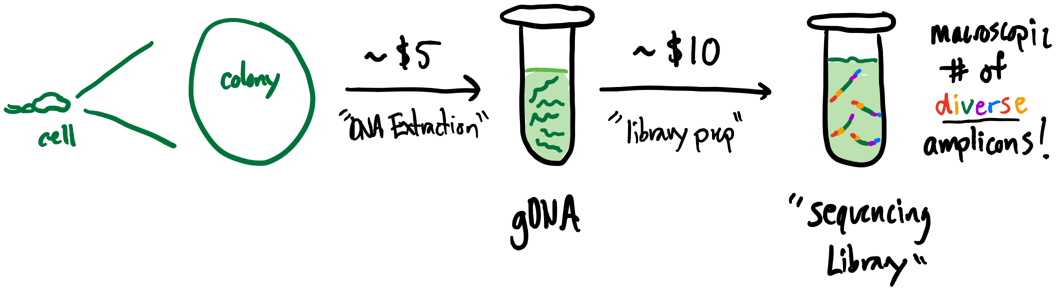




Step 2: Now in position to do PCR + add extra known sequence:

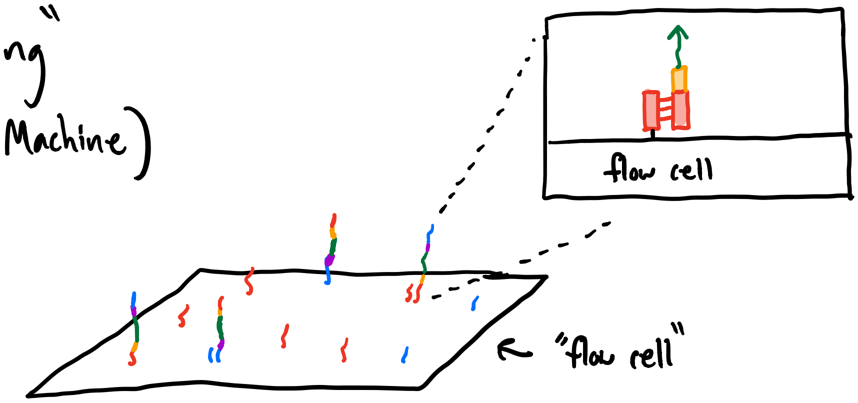


⇒ This gives us:



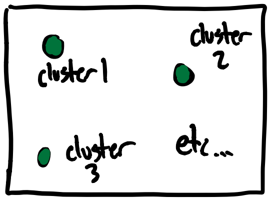
Step 3: "Sequencing" (Illumina Machine)

Flow library onto chip w/ probes that bind sequencing adapter



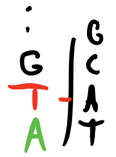
Step 4: do more PCR on bound fragments to turn each molecule into macroscopic cluster of identical* molecules

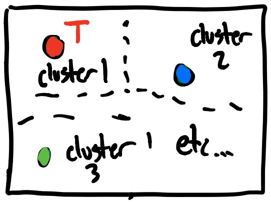
top down view:



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

⇒ if take picture of flow cell:





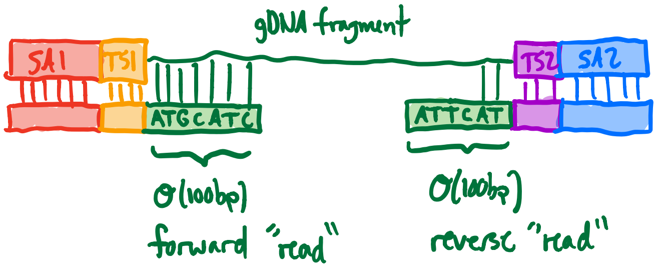
⇒ different colors = nucleic acid @ first position

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:



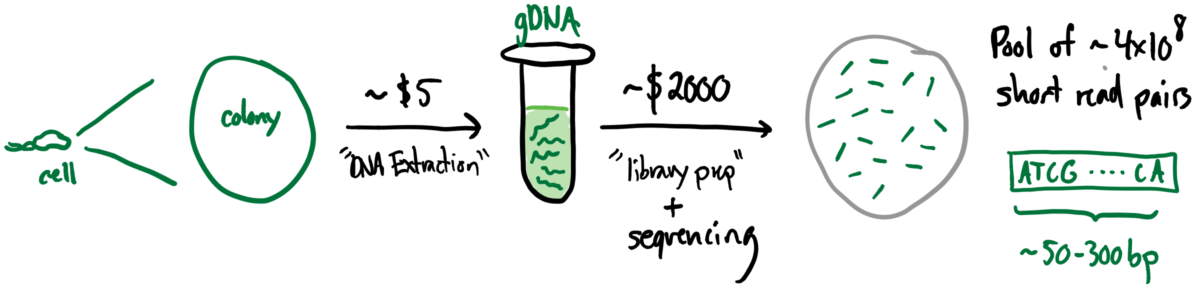
⇒ get $O(100\text{ bp})$ read out from each end of single DNA fragment

⇒ 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:



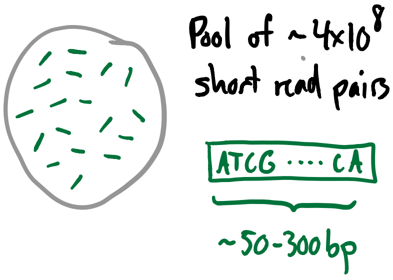
Next: what can we do with this kind of data?

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

① "De novo Genome Assembly"

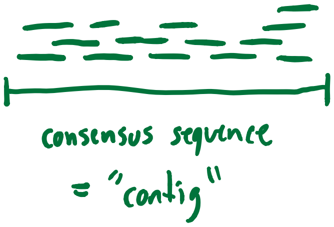
(common programs: spades, Velvet, ...)

Pool of ~4x10^8 short read pairs



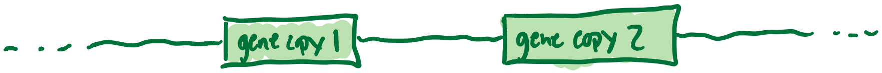
~20bp

Idea: look for reads that partially overlap

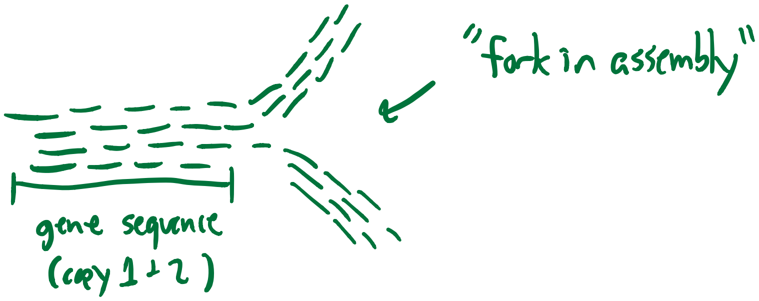


⇒ simple in principle, but lots of corner cases...

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

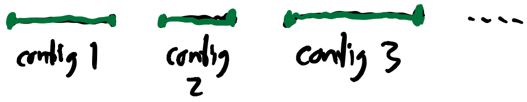


⇒



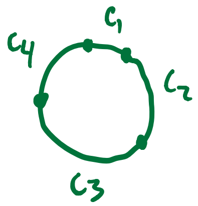
⇒ for these & other reasons, assembly typically yields collection of discrete contigs,

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



⇒ much harder (+ manual effort)

to "finish" into complete genome



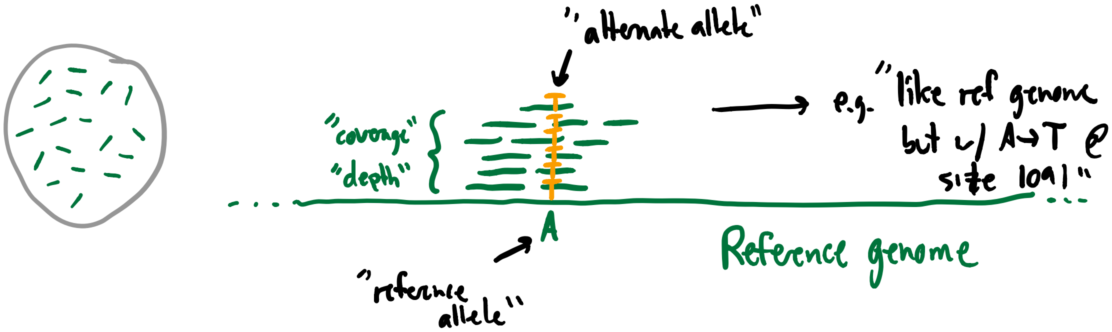
+ HUGE memory requirements (~32Gb - 1Tb depending on L)
(since need to compare all* pairs of reads to each other...)

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: Bowtie2, BWA-MEM, + mpileup)

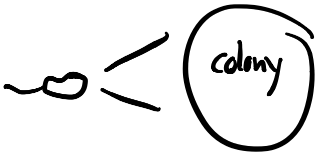


→ e.g. "like ref genome but w/ A→T @ site 1091"

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

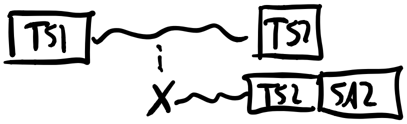
where could these errors come from?

① initial grow up:



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

② library prep & PCR



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

③ Errors on sequencing machine ⇒ estimated $\sim 10^{-3}$/bp ⤴ larger

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

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

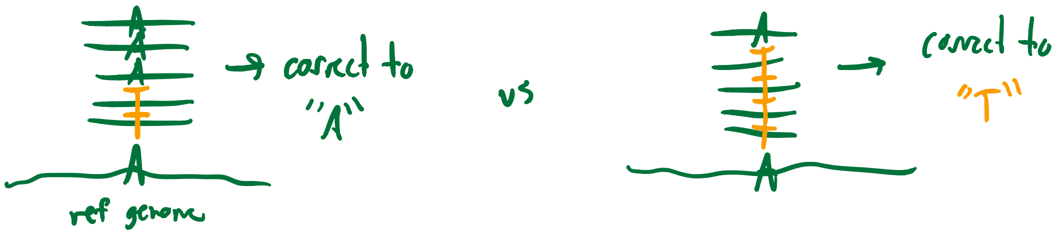
⇒ dominant source of noise: $P_{err} \sim 10^{-3}$ (up to $\sim 10^{-2}$
in special cases.)

⇒ low rate, but $L \gg 1$!

10^{3} errors per E.coli

⇒ # errors per genome = $L \times P_{err}$ ~ 10^{6} errors for Human

⇒ fortunately, can correct errors w/ consensus sequences:



⇒ higher coverage is helpful... how much?

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

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

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

$\Rightarrow$ solve for $\bar{D}$ $\Rightarrow$

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

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

$\Rightarrow$ How many reads is this?

E.coli: 10×10^{6} bp $\Rightarrow 10^{7}$ bp $\Rightarrow 10^{5}$ reads
per site

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

$\Rightarrow$ overkill! (wastes money)