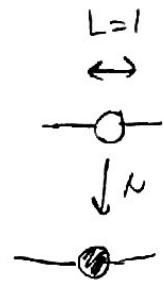


DNA sequencing and genomics (\pm)

(1)

So far, we have focused on dynamics of mutations at a single site in the genome (or technically, in genomes that only have one site, $L=1$), where the presence of a mutation produced a qualitative phenotype that could be screened by colony counting.

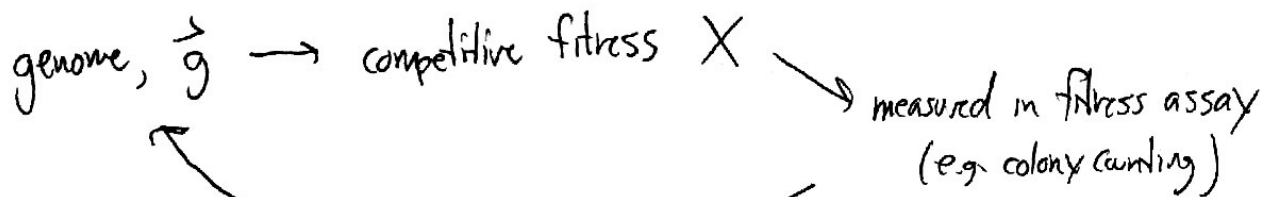


a single site genome

\Rightarrow In practice, genomes contain $L \gg 1$ sites & don't know what kind of phenotypes these mutations produce or how to measure them w/ colony counting assay.

($L \sim 10^{4-5}$ for viruses
 $L \sim 10^{6-7}$ for bacteria
 $L \sim 10^9$ for human cells)

Historically, experimental evolution relied on mapping from



statistics of X w/in & between populations tells us about evolutionary dynamics of \vec{g} ...

\Rightarrow Luria-Delbrück experiment is one classic example.

\Rightarrow downside: indirect. many ^{different} dynamics @ genetic level consistent w/ ^{same} dynamics at fitness level. $\vec{g} \rightarrow X$ poorly understood.

One of the biggest advantages that we have today is that we have the technology to read genomes

2


⇒ if you think about it, this is an incredible task: information is encoded in \bullet order of elements in a single chemical molecule. How do we get it out?

Today we'll present some background about how this is done.


* the details will not be so important for this class, but the basic constraints will be, so we'll try to focus on these.

⇒ much modern progress in biology comes from ability to recognize when your experiment can be shoehorned into constraints of DNA sequencing experiment. (or alternatively, when your theory predicts something that can be measured in genomic data.)

Recall:

genome =  complementary strands.

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

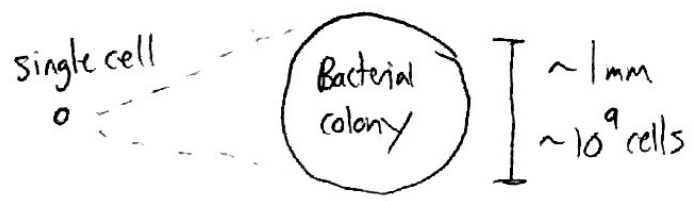
Step 1 for reading genomes = amplification 

(need macroscopic quantities of your DNA molecule to work w/)

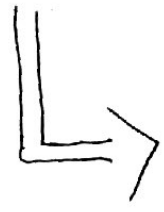
e.g. in physics, we use photomultiplier tube to turn single photon into macroscopic # of photons that we can observe

How do we do this for DNA molecules?

For bacteria, easy! use their built-in tools to exponentially proliferate



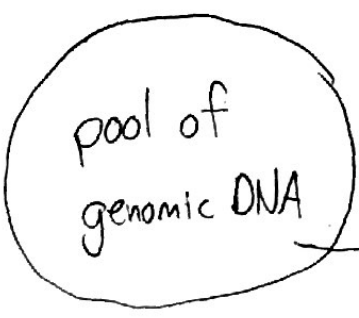
⇒ saw that colonies will be nearly clonal w/ wildtype mutation rates ($\sim 10^{-10}$)



techniques for breaking apart cells (lysis) and extracting just the DNA molecules.

"DNA Extraction"

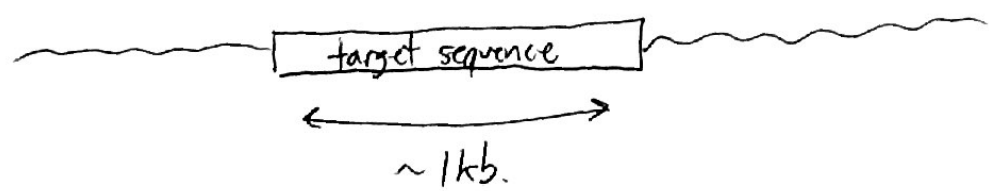
⇒ can do in 96-well plate for <\$5 per well.



from billions of individual cells

But $L = 10^6$ sized genomes are too difficult to do anything w/ directly.

⇒ most sequencing methods work with short sequences, $l < 10^3$ bp



How to get a macroscopic amount of just this region?

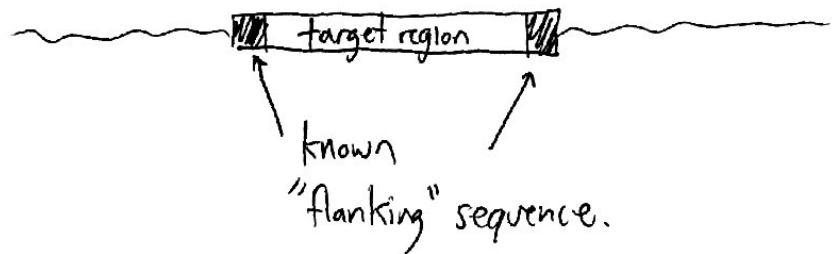
Answer: PCR (Polymerase Chain Reaction)

(4)

⇒ ~~cell-free~~ cell-free chemical reaction that's like exponential growth, but just for part of genome.

⇒ again, ~~PCR~~ takes advantage of machinery invented by bacteria to replicate DNA.

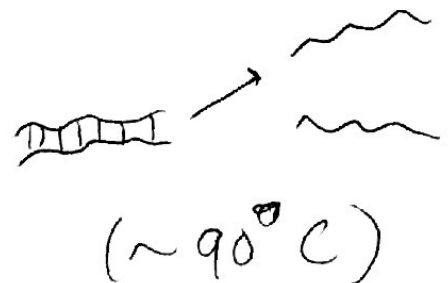
⇒ but requires us to know some of the sequence that surrounds our target region:



① can have company synthesize "primers" (short sequences of ~ 20 bp) corresponding to known sequence ($\sim \$0.30/\text{bp}$ for $\sim 1000^3$ reactions)

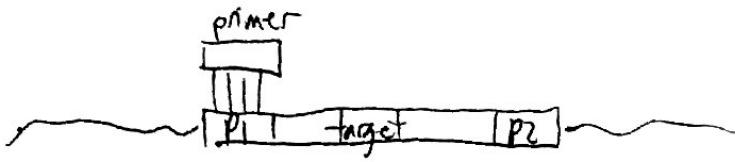
↪ mix w/ purified DNA polymerase, dNTP (free A's, C's, T's, G's) + your genomic DNA.

② heat sample so that strands "melt":



③ Now cool sample so that primers "anneal" to input DNA

⑤



⇒ melting & annealing is physics problem

$$\text{e.g. } \frac{p(\text{bound})}{p(\text{not})} = e^{-\Delta E/T}$$

∠ $\Delta E \propto \#$ ~~matched~~ matched bases
(very crudely)

⇒ want your primer to bind to known region, but not to anywhere else.

* saw in homework that most of E. coli genome identifiable w/
 $l=20\text{bp}$ sequence, so $\sim 20\text{bp}$ primers sound reasonable.

⇒ more generally, 4^l possible sequences of length l .

$L-l \approx L$ distinct sequences of length $l \ll L$ in
genome of total length L .

if E. coli genome is random, chance of overlap small if $L \left(\frac{1}{4^l}\right) \ll 1$

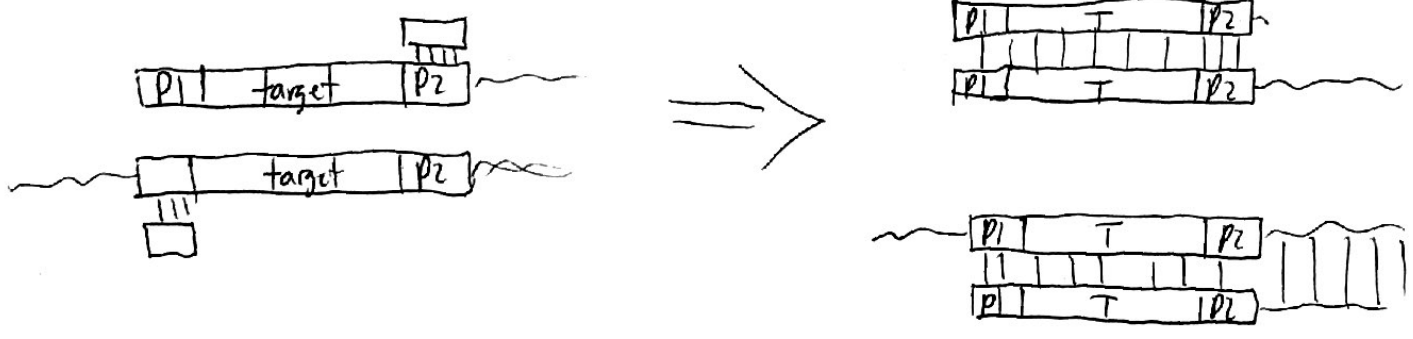
$$\Rightarrow l \gtrsim \log_4(L) \approx \begin{cases} \sim 10\text{bp} & (\text{Bacteria}) \\ \sim 16\text{bp} & (\text{humans}) \end{cases} \Rightarrow \text{lower bound}$$

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


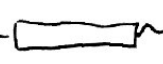
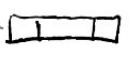


("extension phase")

5 Melt, anneal, and extend again

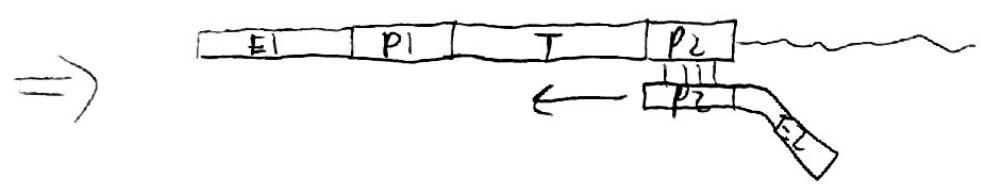
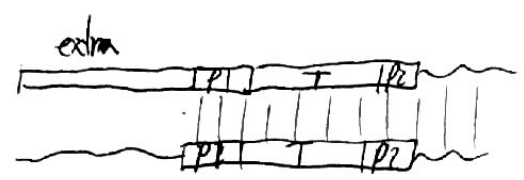
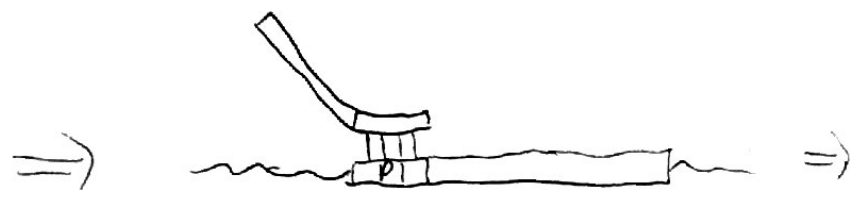
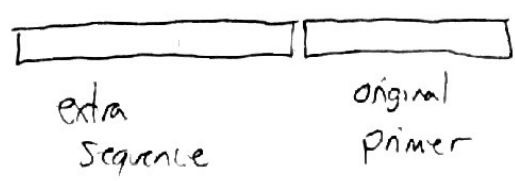


6 Repeat for K ~ 20-30 cycles

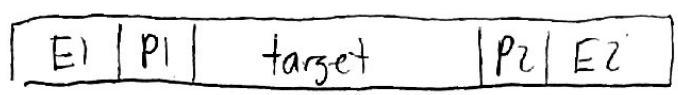
=> exponential amplification of  sequence
 (overwhelms  and  versions)

7 "clean up" step to remove extra primers, etc.

Note: can also use PCR to add extra bit of sequence to your target sequence:

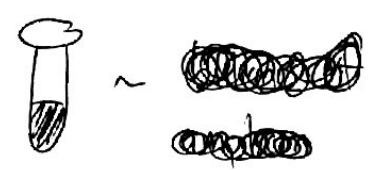


⇒ amplified sequence



as long as you're willing to pay for (\$0.30/bp) up to ~ 100

Now we have tube of PCR product



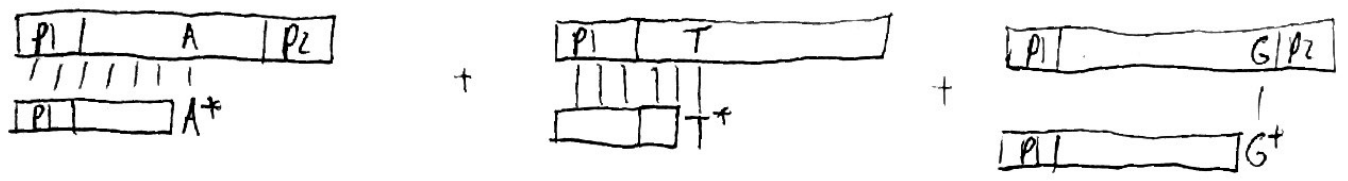
⇒ ~~how~~ how can we read out information?

Sanger sequencing: like an extra round of PCR.

⇒ idea: mix PCR product + primers + DNAp + dNTP + (P1)

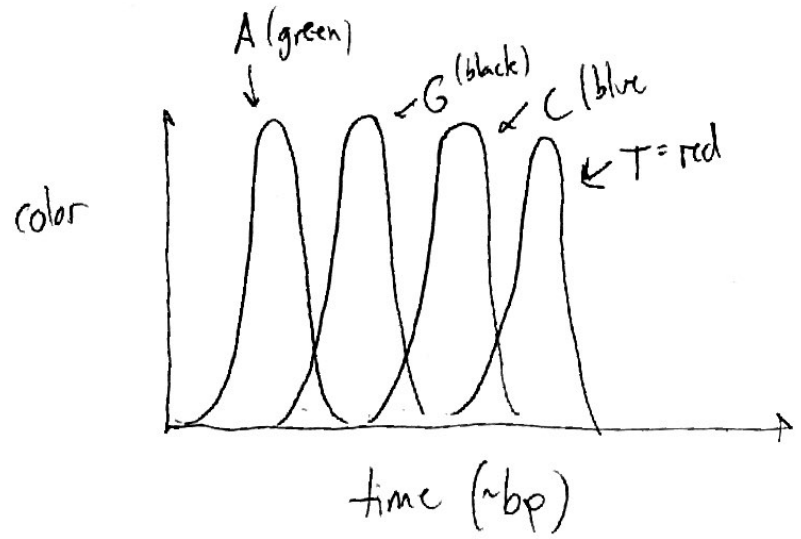
 fluorescent dNTPs that block DNAp

After 1 round of extension, will randomly have mixture of



flow in electric field
(shorter fragments move faster)


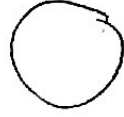
laser to read fluorescence.




"chromato-gram"

↳ can read off sequence.

⇒ costs ~\$5 (send away overnight).

⇒ so have seen how to go from  single cell →  colony

to sequence of  region for ~\$10 + \$5 for primers.
1kb

In problem 1 of last week's homework, Lang & Murray used Sanger sequencing to sequence the URA3 gene in their 720 yeast colonies \Rightarrow \sim \$7000

\Rightarrow but quickly gets expensive to sequence a whole genome.

e.g. E. coli, 1 clone = 10^6 bp = 10^3 sanger sequences \sim \$5k

Humans, 1 person = 10^9 bp = 10^6 sanger sequences \sim \$5M!

\Rightarrow this is roughly how first Human genome project was done.
(hence need for massive consortium)

\Rightarrow Now things are much cheaper w/ Next-generation sequencing
("Illumina" sequencing)



same basic idea but higher throughput.

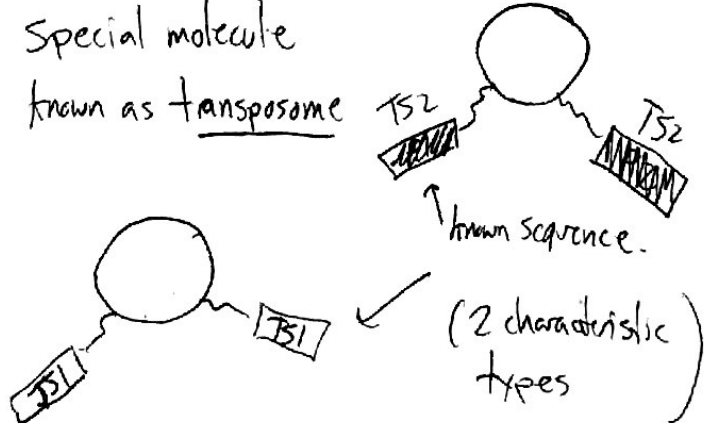
Step 1

genomic DNA

+

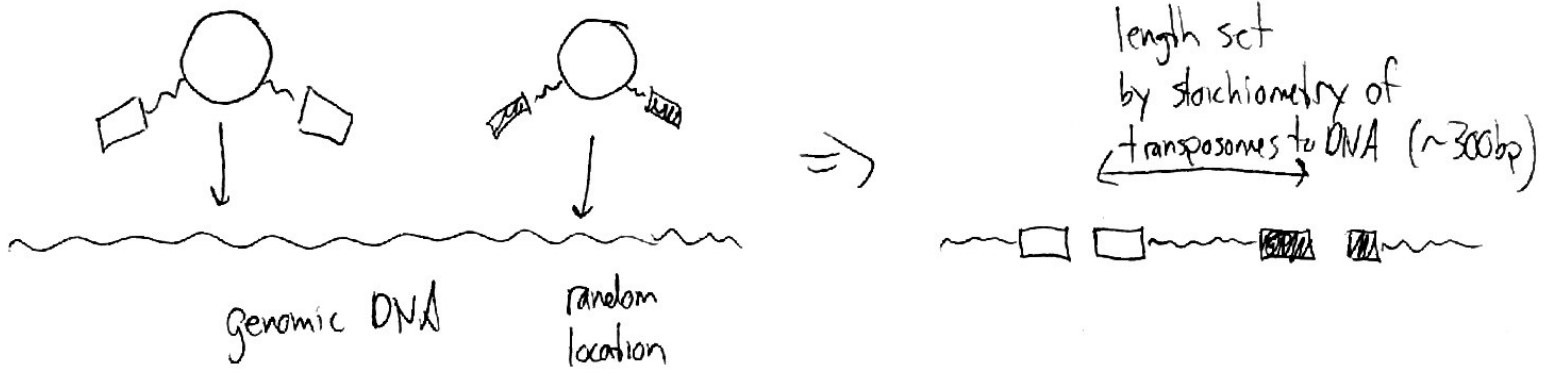
Special molecule known as transposome

"library prep"

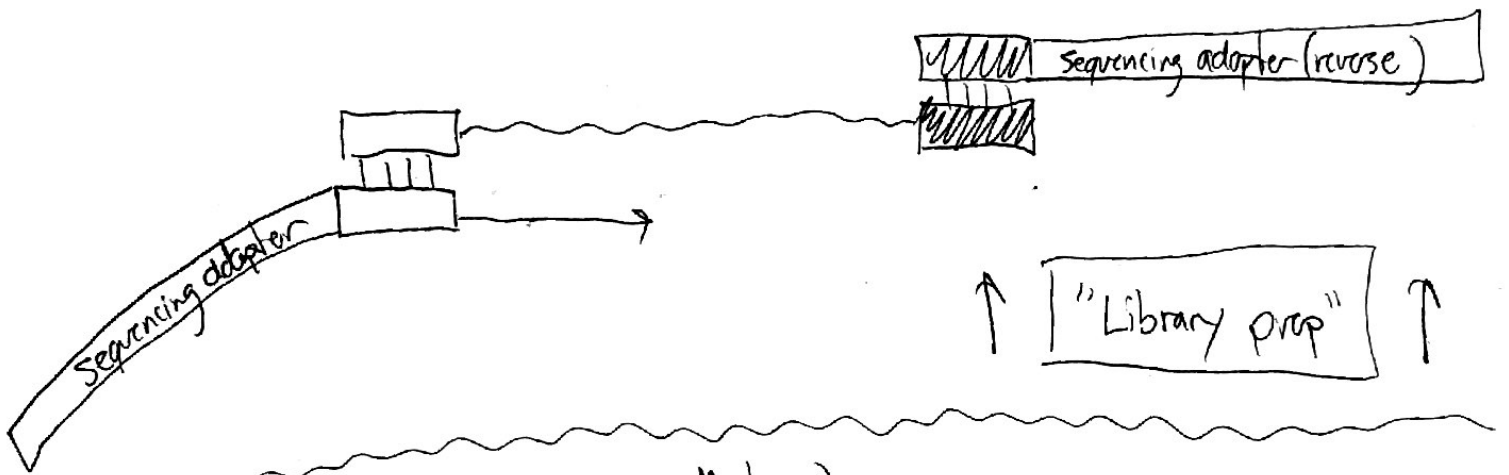


Transposomes simultaneously cut DNA & insert known sequence:

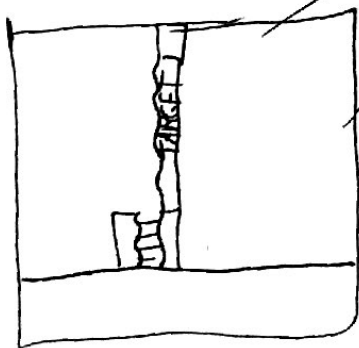
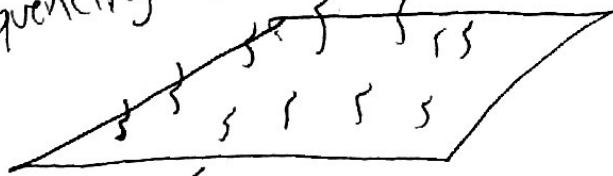
(10)



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

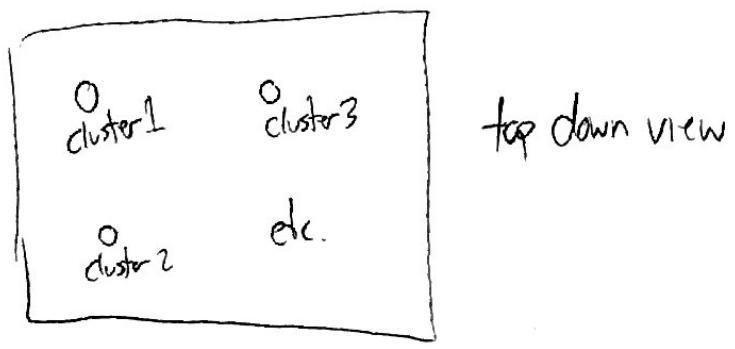
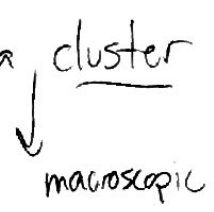


Step 3: "sequencing" (Illumina Machine)



flow amplified fragments onto chip that has lots of probes that bind sequencing adaptor ("flow cell")

Step 4: Now do some PCRs directly on banded DNA to turn each banded molecule into a cluster of identical molecules



Step 5: Now flow fluorescent dNTPs ~~one~~ that incorporate once and then stop.

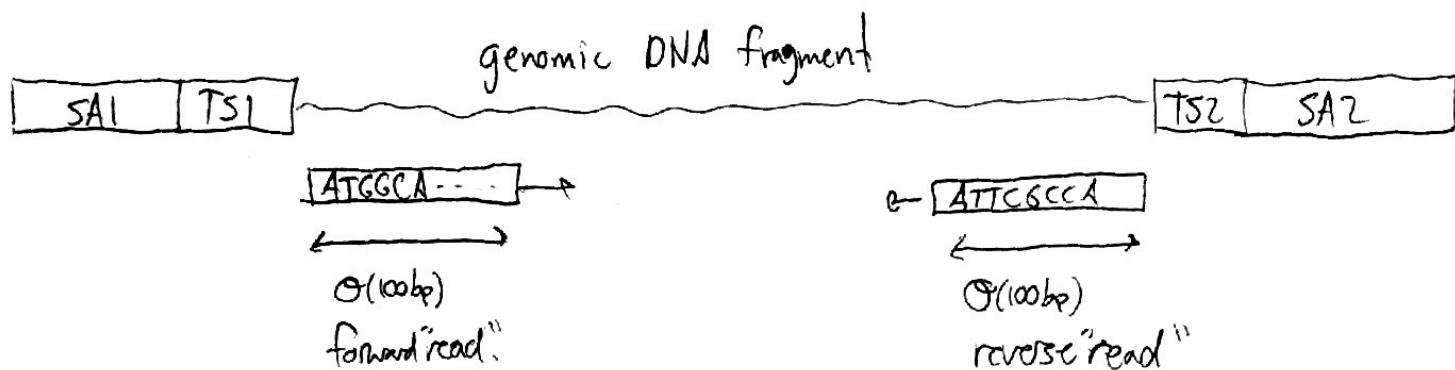
⇒ if take picture of flow cell, each cluster has different color ≈ nucleic acid of first position.

Step 6: can remove fluorescent part that blocks incorporation of new base & repeat w/ new round of fluorescent dNTPs. (2nd position)

Step 7: can repeat for ~ $O(100)$ cycles until pictures get desynchronized too much.

⇒ read ~ $O(100)$ bases from one end of each molecule.

Step 8: can repeat for reverse direction:



⇒ get $\sim O(100bp)$ from each end of single DNA molecule!
"read pair"

⇒ w/ modern Illumina machines, this process is very high throughput. can get $\sim 4 \times 10^8$ read pairs in a few days for $\sim \$2000$. (catch: you can't do smaller batches)

often limited by quality of camera.

⇒ 4×10^8 read pairs is $\sim 100Gbp$ of sequence

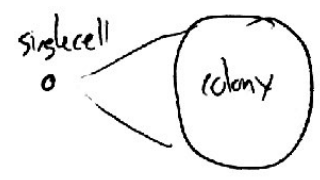
- Human genome $\times 30$
- E.coli genome $\times 3 \times 10^4$ (overkill!)

⇒ easy to "multiplex". Add barcode ^{sample specific} during library prep:

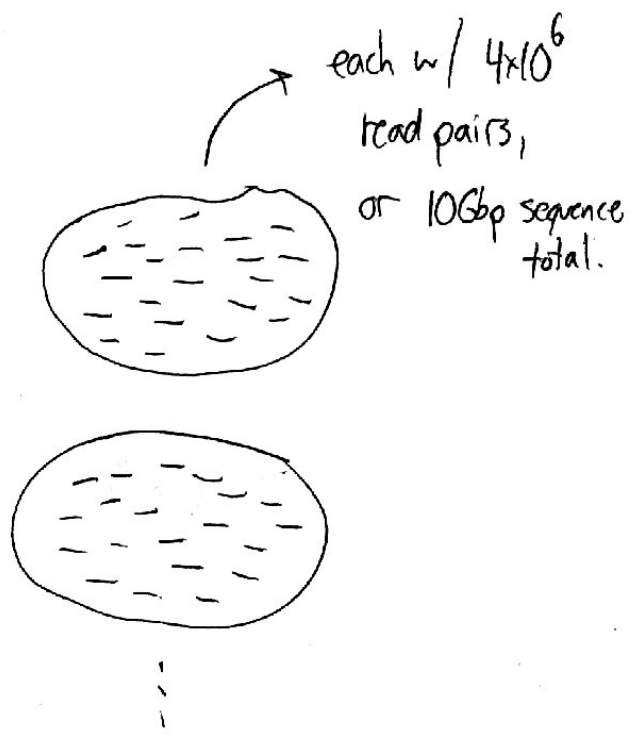


⇒ can sequence ~ 100 E.coli libraries on one flow cell ("lane") and get 300-fold coverage of E.coli genome. (still pretty good!)

Summary:



library prep + sequencing (~\$3000) →



x 100

What do we do with this kind of data?