Biochemistry/ selection

① Shotgun Sequencing

WGS
RNAseq
shotgun metagenom
ChIP seq

Sample (DNA/RNA) → All DNA → Fragmented

Digitized

1. Random sampling
2. Quantitative

Computer

Not Shotgun:
16s / barcode tag end
exome
RAD seq

## Illumina

DNA sample (or cDNA) → fragment → ~300-800 bp

"single molecule"

ligate primers

Paired end sequencing

300 read 1 → ??? ← read 2 300

200 bp

PCR
BAD

bias against GC/AT rich

flow cell

bridge amplify

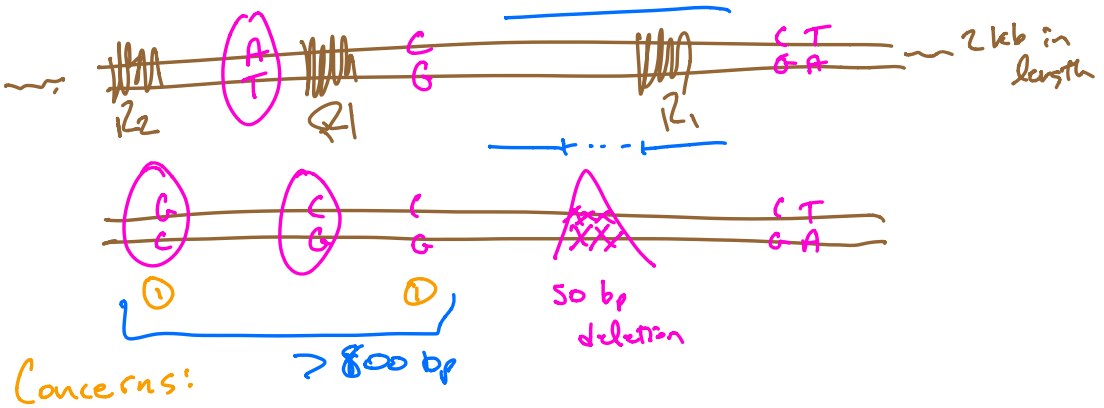
x 20m - 2 billion

sequence in situ

G C C A

2am-2bn sequences → 100-300 bp each

## What do we want? (Interlude)




Concerns:

1. we want haplotype resolution - which variants are on the same physical chr.
2. we want high coverage - we want to see all DNA at least 5-6 times

latency
throughput (bp per time)
cost
accessibility / realtime investment cost

### Long read sequencing

$$ but a lot longer than Illumina (~2kb - 100 kb)

- single molecule (no PCR)
- more input DNA
- lower thruput

Nanopore + Pacific Biosciences

Two reasons:

1. single molecule → PCR biases, don't use!! no amplify → minimal bias
2. long read
   - ↳ haplotype resolution
   - ↳ repeat resolution
   - ↳ good for structural variants: long indels, inversions, etc.

De novo sequencing

## Microbial genomes

Small
No/few repeats
Haploid
"Easy" to get lots of DNA

↓

Short read is OK

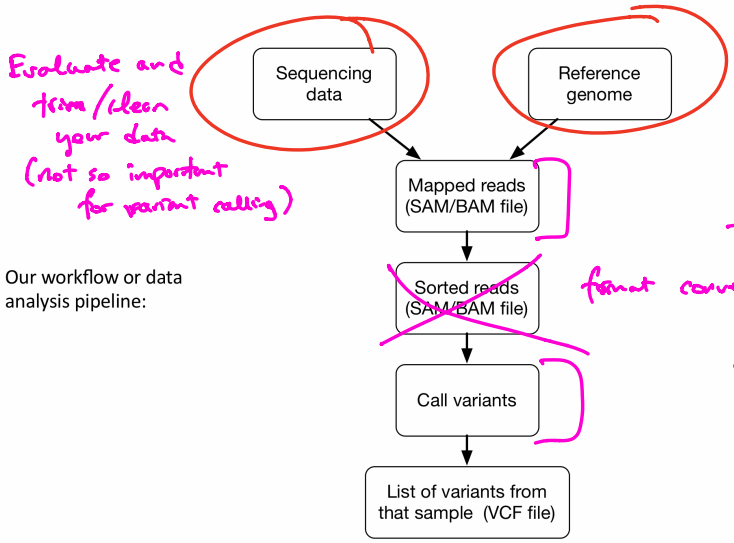
## Animal/Plant

Large
Repeat rich
diploid or worse
hard to culture

↓

long read better

Other:
- RNAseq
- Metagenomics




Our workflow or data analysis pipeline:

Evaluate and trim/clean your data (not so important for variant calling)

Choose the right reference!

99% similar

format conversion

Friday Lab 4:
- mapping parameters
- coverage
- pileup parameters