BF528 - 2nd (Next) Generation Sequencing

Introduction

Methods for obtaining nucleotide sequence information:

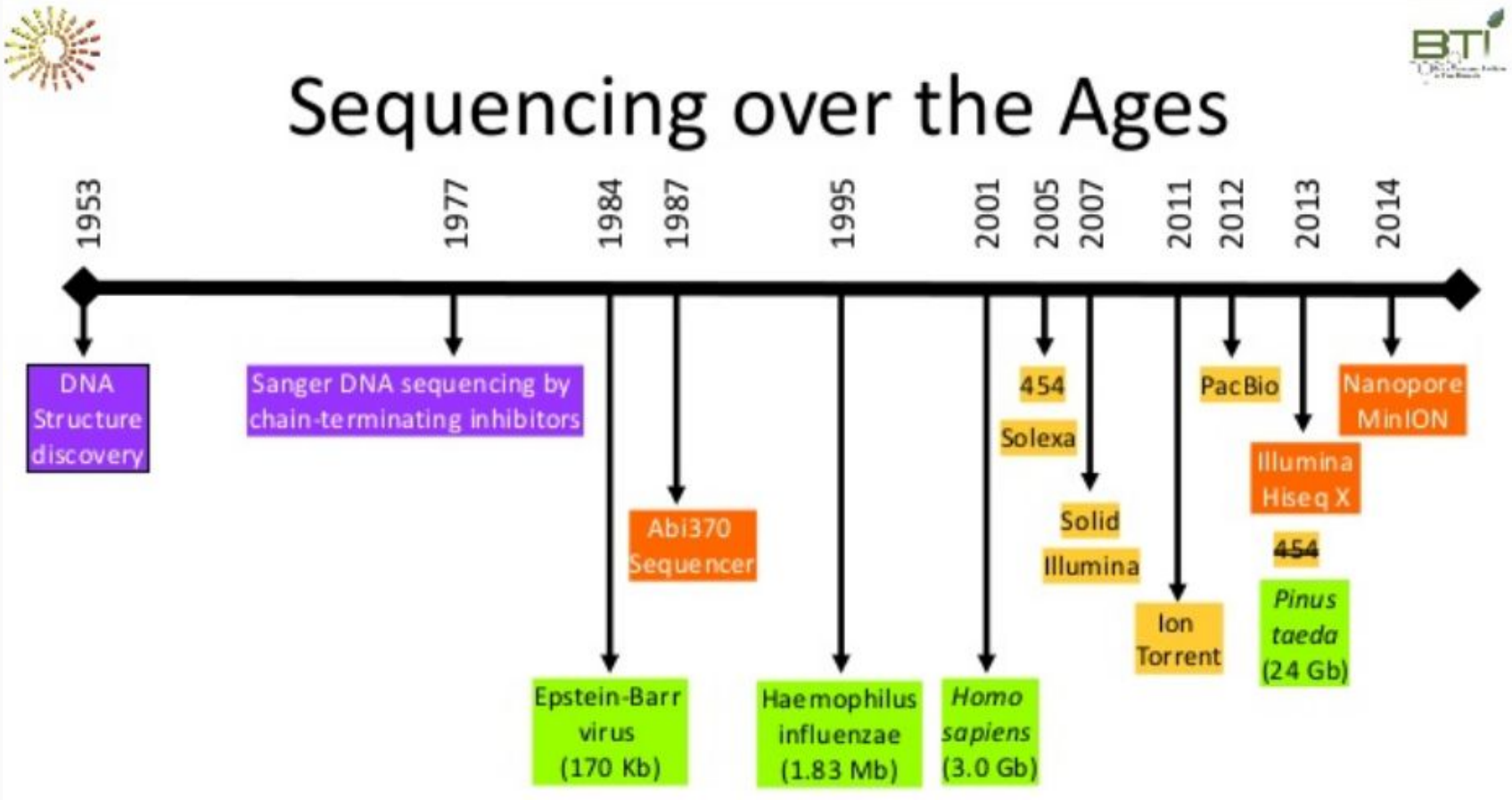
- Sanger sequencing (i.e. 1st generation)
  - sensitive, slow, low-throughput, expensive
- Microarray (not sequencing technology)
  - cheap, high-throughput, fast, known sequences only

We want to obtain information on **many sequences quickly**, **cheaply**, with **high confidence**, and **without knowing those sequences** beforehand

High-throughput Sequencing

- **High-throughput:** many nucleotides very quickly
- **Sequencing:** determine the linear sequence of nucleotides
- Use biochemical or biophysical approaches
- Current technologies vary by:
  - Number of individual sequences generated
  - Length of sequences
  - Confidence in sequences

Evolution of NGS



Surya Saha, Boyce Thompson Institute, Ithaca, NY (BTI plant bioinformatics course)

Sequencing machines

- Expensive to purchase (hundreds of thousands $USD)
- Expensive to operate (e.g. reagents, flow cells)
- You can sequence your genome at 30X depth for <$1000 USD.

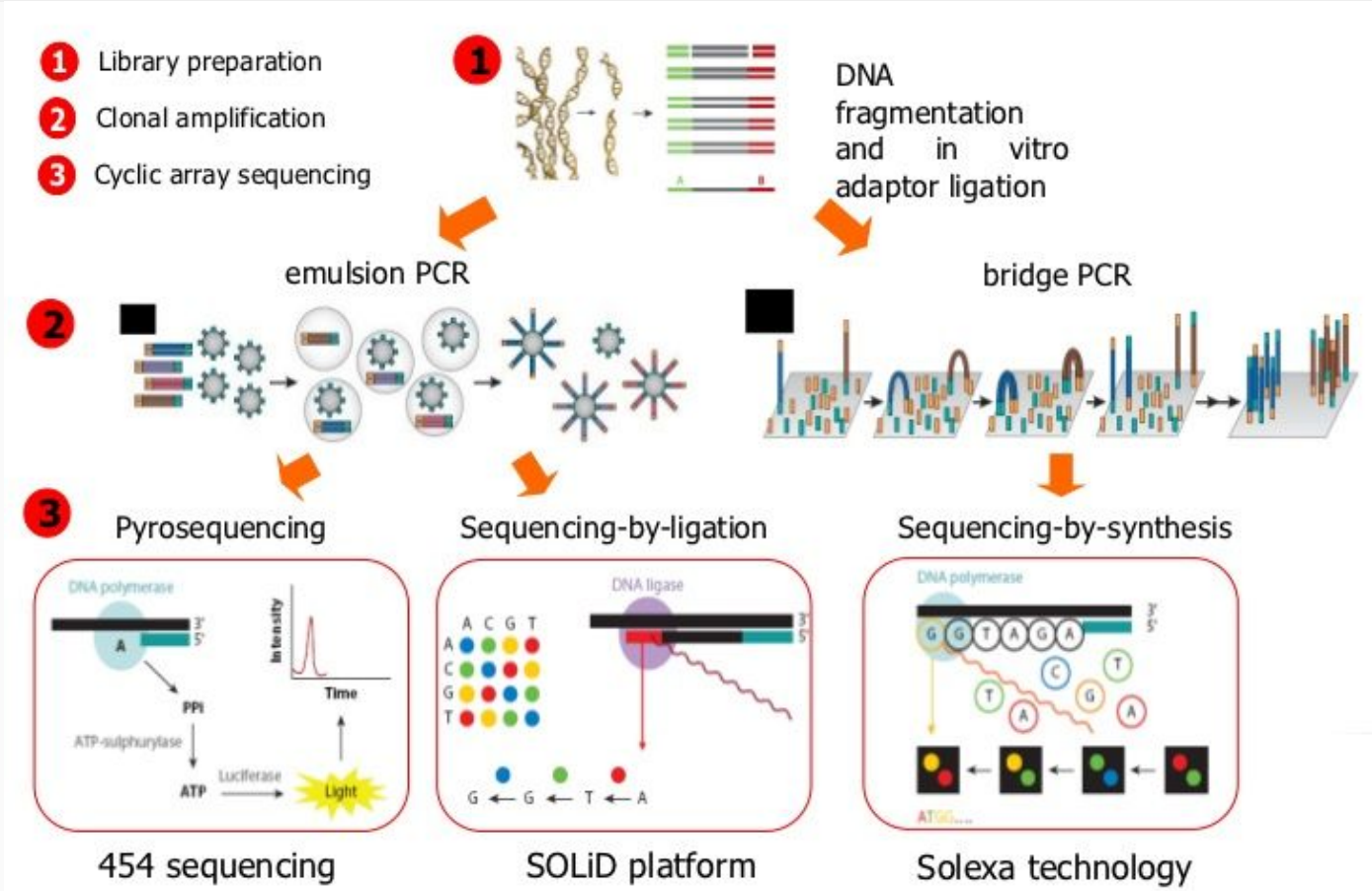
Roche 454

Ion Torrent

Illumina HiSeq 2500

Illumina NovaSeq 6000

Experiment


1 Library preparation
2 Clonal amplification
3 Cyclic array sequencing
1
DNA fragmentation and in vitro adaptor ligation
emulsion PCR
bridge PCR
2
3
Pyrosequencing
Sequencing-by-ligation
Sequencing-by-synthesis
DNA polymerase
Intensity
Time
PPi
ATP-sulphurylase
ATP
Luciferase
Light
DNA ligase
A C G T
DNA polymerase
454 sequencing
SOLiD platform
Solexa technology

Comparison

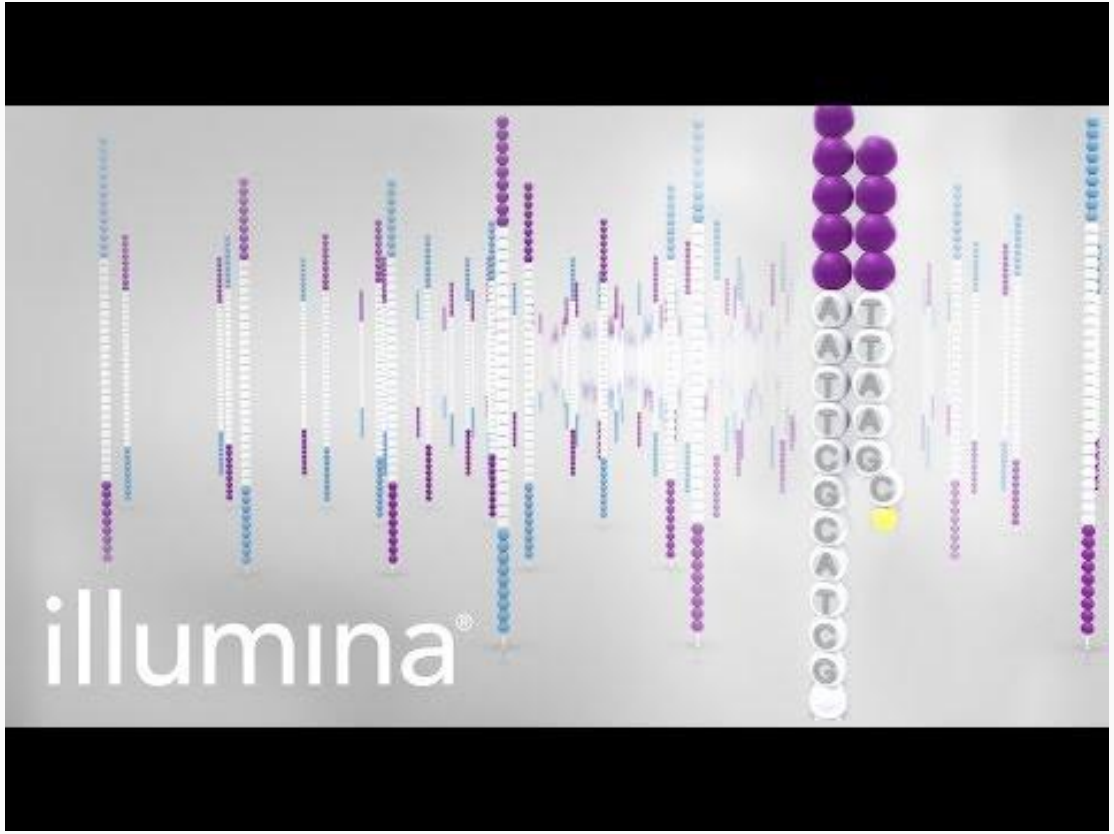
	Run Time	Read Length	Quality	Total nucleotides sequenced	Cost /MB
454 Pyrosequencing	24h	700 bp	Q20-Q30	1 GB	$10
Illumina Miseq	27h	2x300bp	> Q30	15 GB	$0.15
Illumina Hiseq 2500	1 - 10days	2x250bp	>Q30	3000 GB	$0.05
Ion torrent	2h	400bp	>Q20	50MB-1GB	$1
Pacific Biosciences	30m - 4h	10kb - >40kb	>Q50 consensus >Q10 single	500 - 1000MB /SMRT cell	$0.13 - $0.60

Quail, Michael A., et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC genomics*13.1 (2012): 341.

Illumina Sequencing

- Most common sequencing technology today
- Sequences any DNA
- Sequencing by synthesis method
- Sequences (**reads**) are short (<300bp)
- 2 gigabases - 6 terabases per run
- Hours to days to complete one run

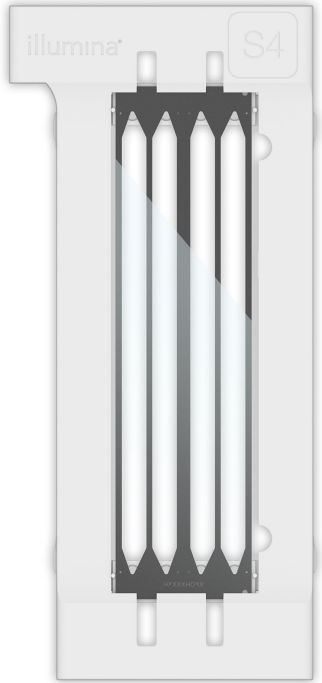
Illumina Sequencing Process



https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina Sequencing Properties

- Sequencing occurs on a **flow cell**
- Each flow cell has 1 to 8 **lanes**
- Number of reads for overall flow cell varies
- Length of reads is fixed (e.g. 250 bp)
- Read format:
  - **Single end** - one read per molecule
  - **Paired end** - two reads per molecule
- **Multiplexing**: sequence many samples at once using molecular barcodes



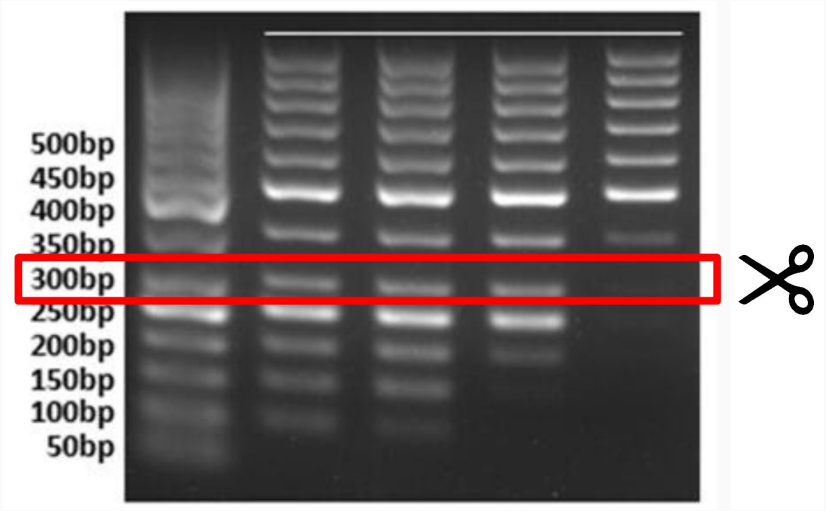
NovaSeq Flowcell
Courtesy of Illumina, Inc.

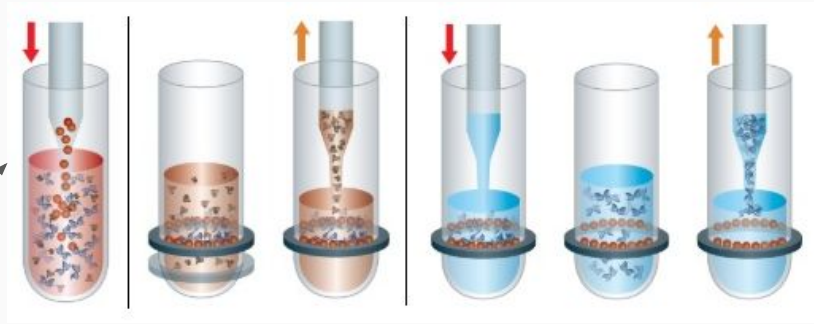
Sequencing Library Generation Workflow

- **Sequencing Library**: DNA prepared for sequencing
- Workflow:
  1. Extract RNA/DNA from sample
     - If RNA, reverse transcribe to cDNA
  2. Size select using gel cut or random shearing
  3. PCR amplify DNA if concentration is low
  4. Add sequencing adapters
     - If multiplexing, use barcoded adapters
  5. Pool samples, load across flow lanes for sequencing
- Typically only perform 1, sequencing cores do the rest

Design Choice: Fragment Length

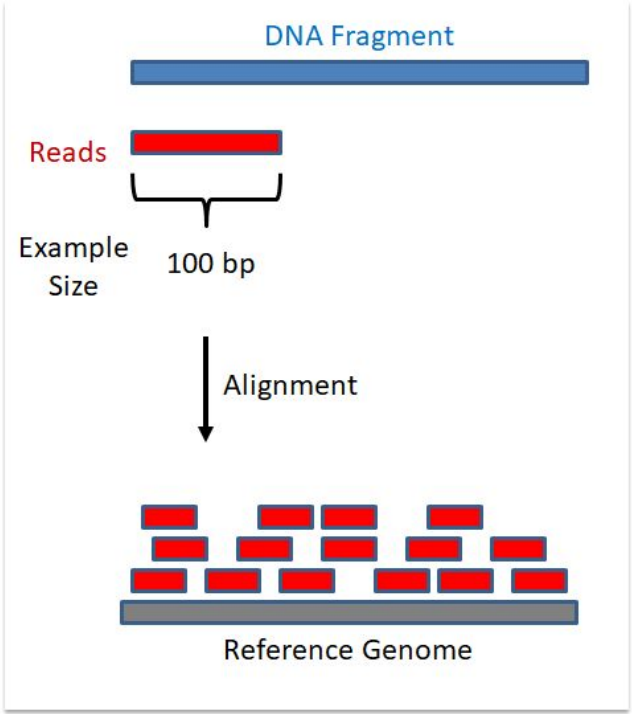
- Illumina sequencers can only sequence DNA fragments up to ~300nt long
- DNA must be *size-selected*, by one of:
- Gel cut (old method):
  - ~200-300nt band cut, purified, prepared for sequencing
  - Fragment length follows a normal distribution around target cut size
- SPRI Beads (current method)



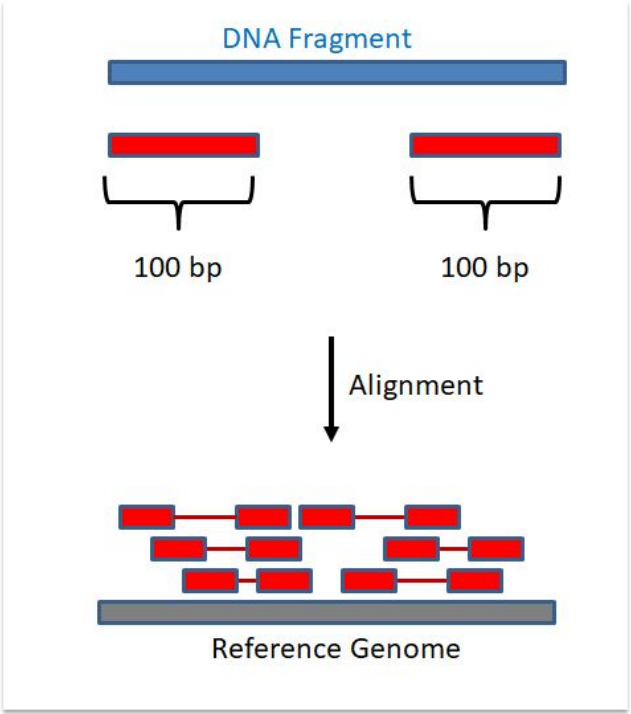


Design Choice: Single End vs Paired End

Single End



Paired End



Design Choice: Number of Reads

- Each sequencing run generates a # of total reads
- # of reads/sample ~= # total reads/number of samples
- # of reads for one sample: **library size**
- Choose target library size based on:
  - Desired **depth**
  - Desired **coverage**

For more see https://genohub.com/recommended-sequencing-coverage-by-application/

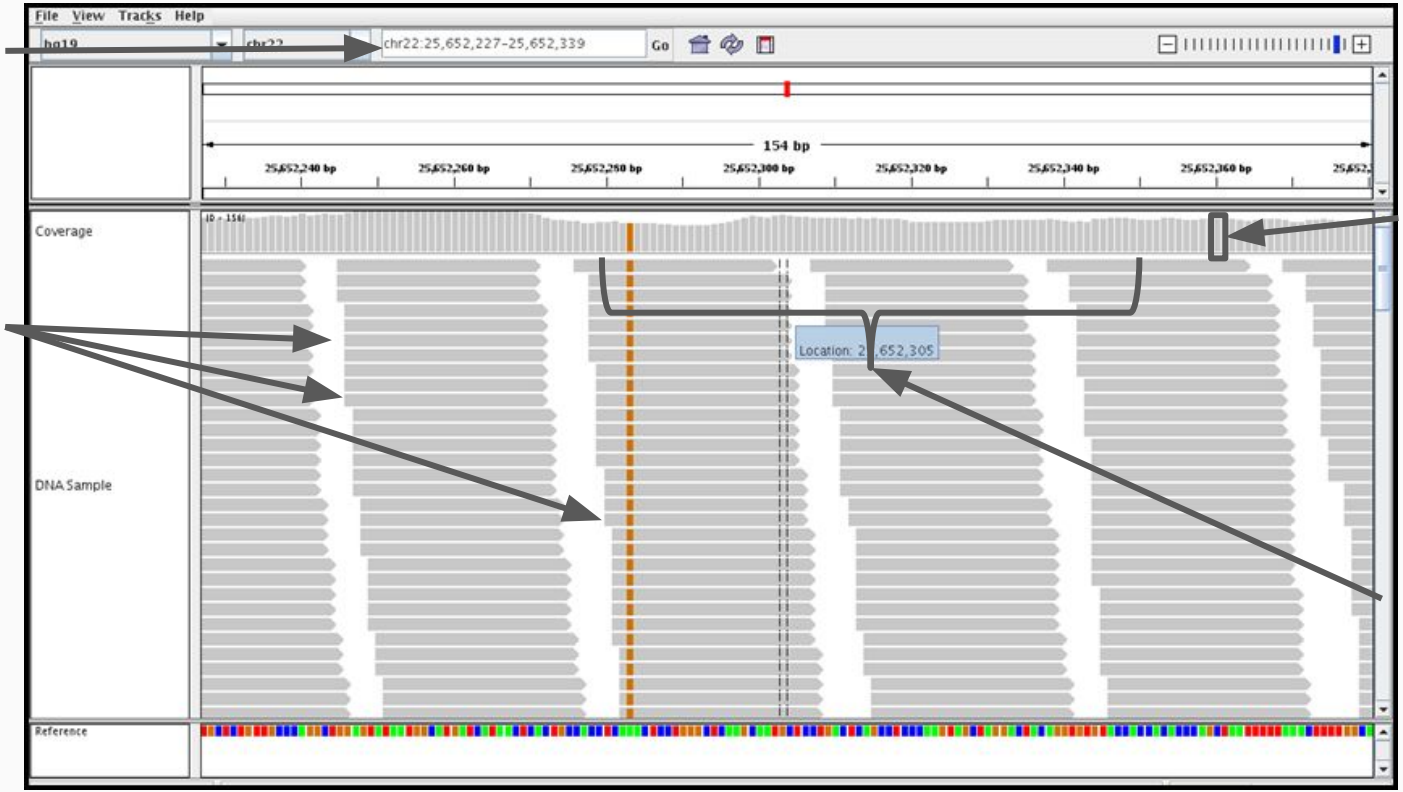
Critical Concept: Read Mapping

- **Question:**
  "Given a read and a reference sequence, where, if anywhere, in the reference does the read sequence occur?"
- E.g. chr3:2,358,092-2,358,193
- More on this next lecture

Mapped Read Terminology

Genome Locus

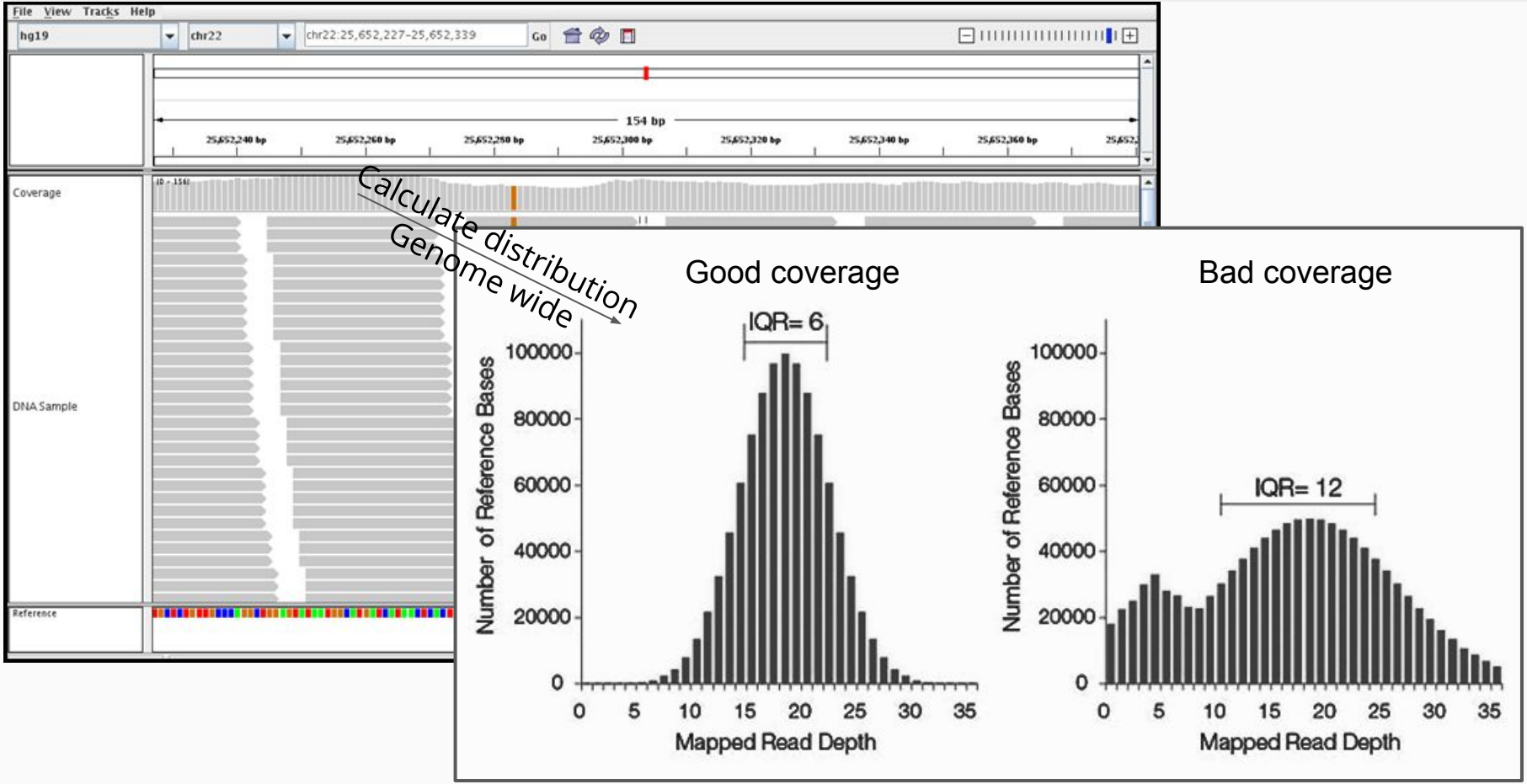
Mapped or Aligned reads



Depth: number of sequenced bases that map to a given location

Coverage: fraction of genomic locus covered by at least one read

Coverage - Whole Genome Sequenceing



Sequence Data Format: fastq

- Sequence reads provided in fastq format
- For each read there are 4 lines:

```
@read_header comment
read_sequence
+[quality_header]
phred_quality_scores
```

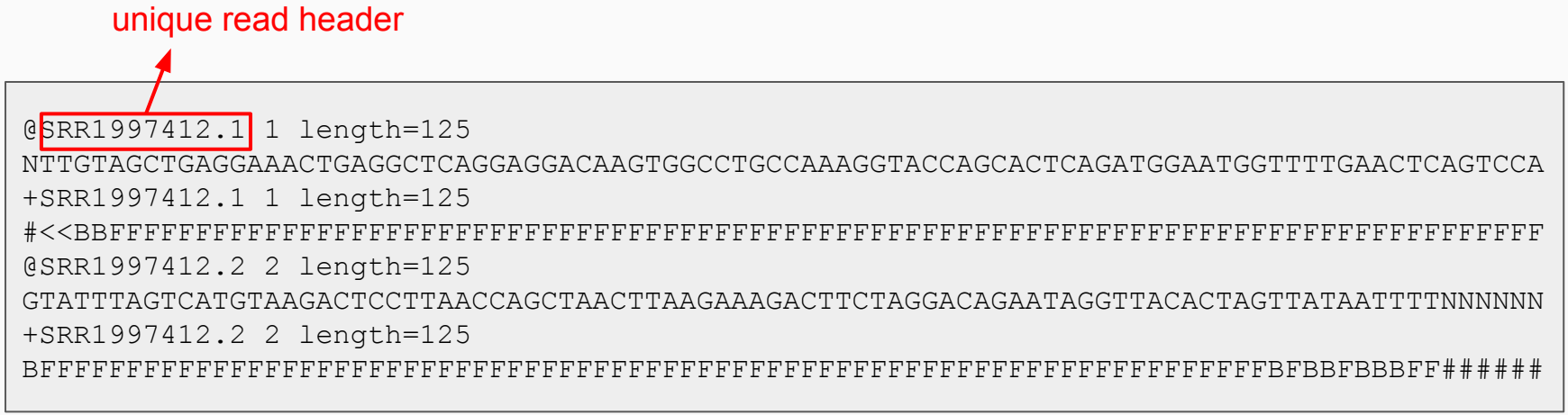
- Phred scores estimate the probability that a base is called incorrectly

fastq format

start new read

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBBFBBBFF######
```

fastq format

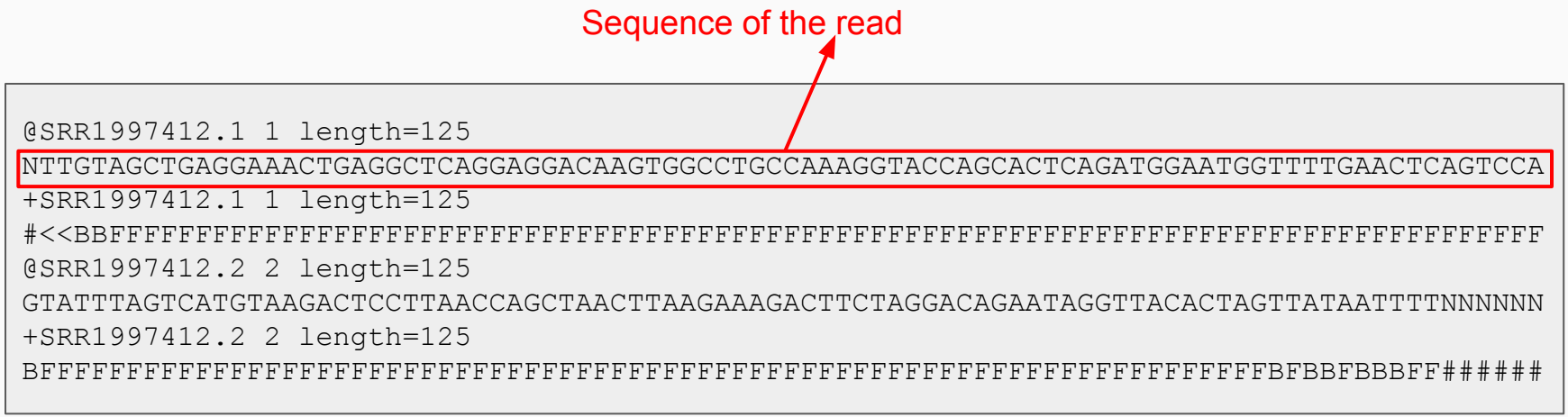


fastq format

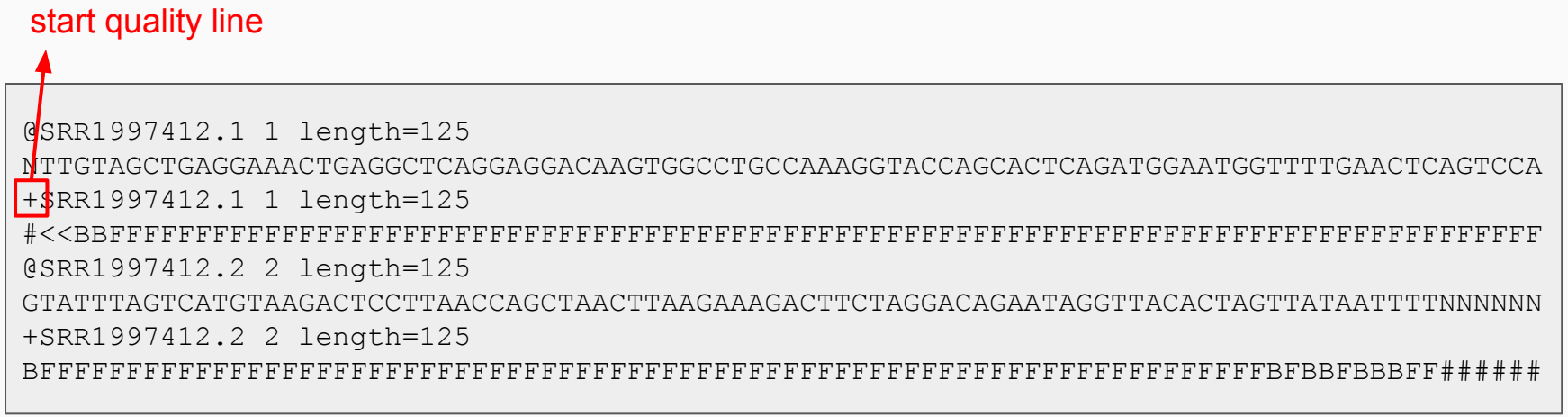
comments separated by space, could be anything

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBBFBBBFF######
```

fastq format



fastq format



fastq format

repeat read header and comment, not required, often blank

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBBFBBBFF######
```

fastq format

fastq format

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBBFBBBFF######
```

Next read

Sequence Quality Score - Phred

- Each read base has a corresponding quality score
- Score indicates probability of base being wrong

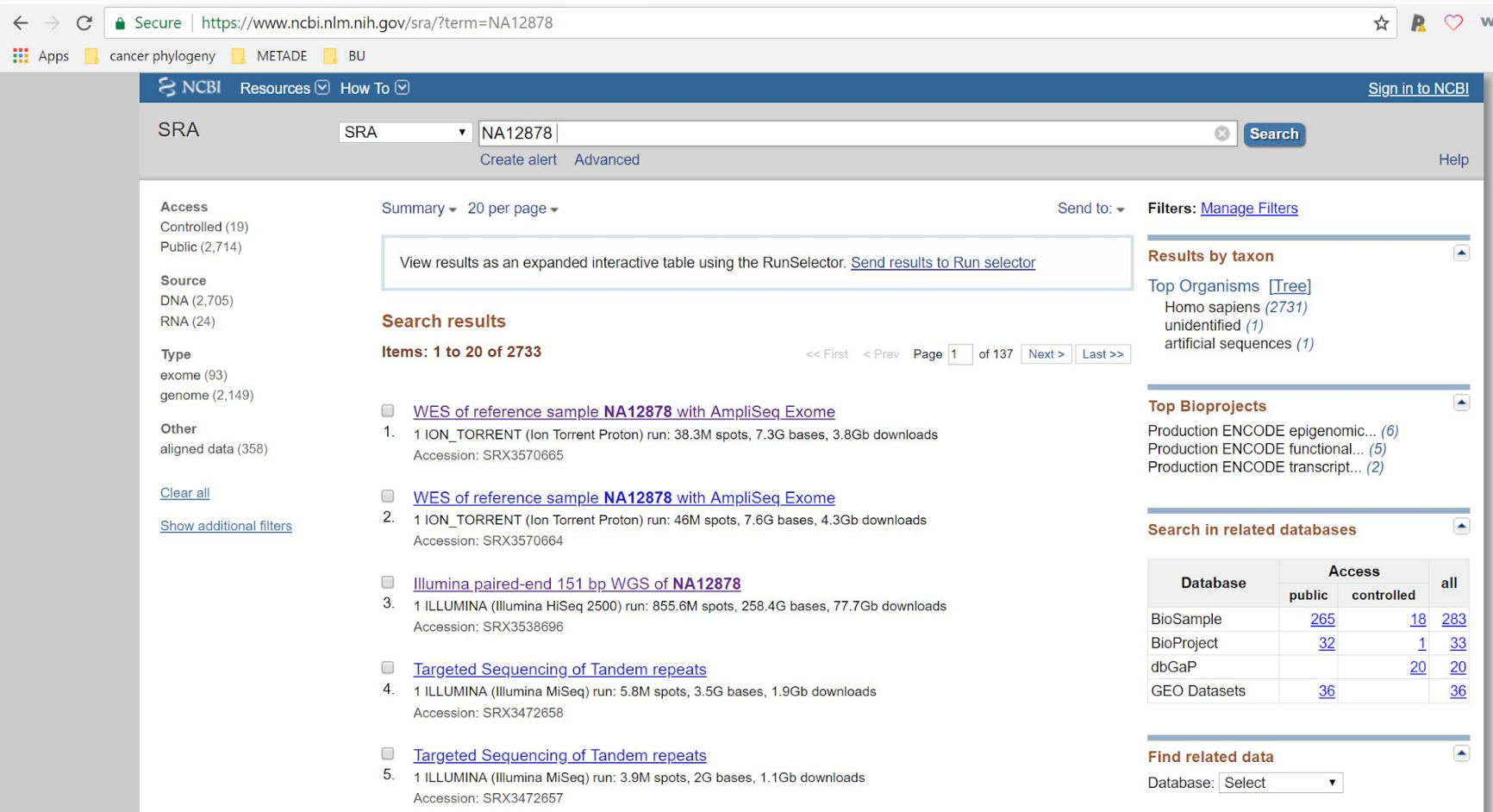
$$Q_{\text{phred}} = -10 \log_{10} e$$

- Scores quantized to integers, e.g. [-3,41]
  - e.g. e = 0.0123, Q = 19.1 = 19
- Q then mapped to ASCII space by adding offset
  - e.g. 19+64 = 83, 83 == ‘S’ in ASCII
- For more info: https://en.wikipedia.org/wiki/FASTQ_format

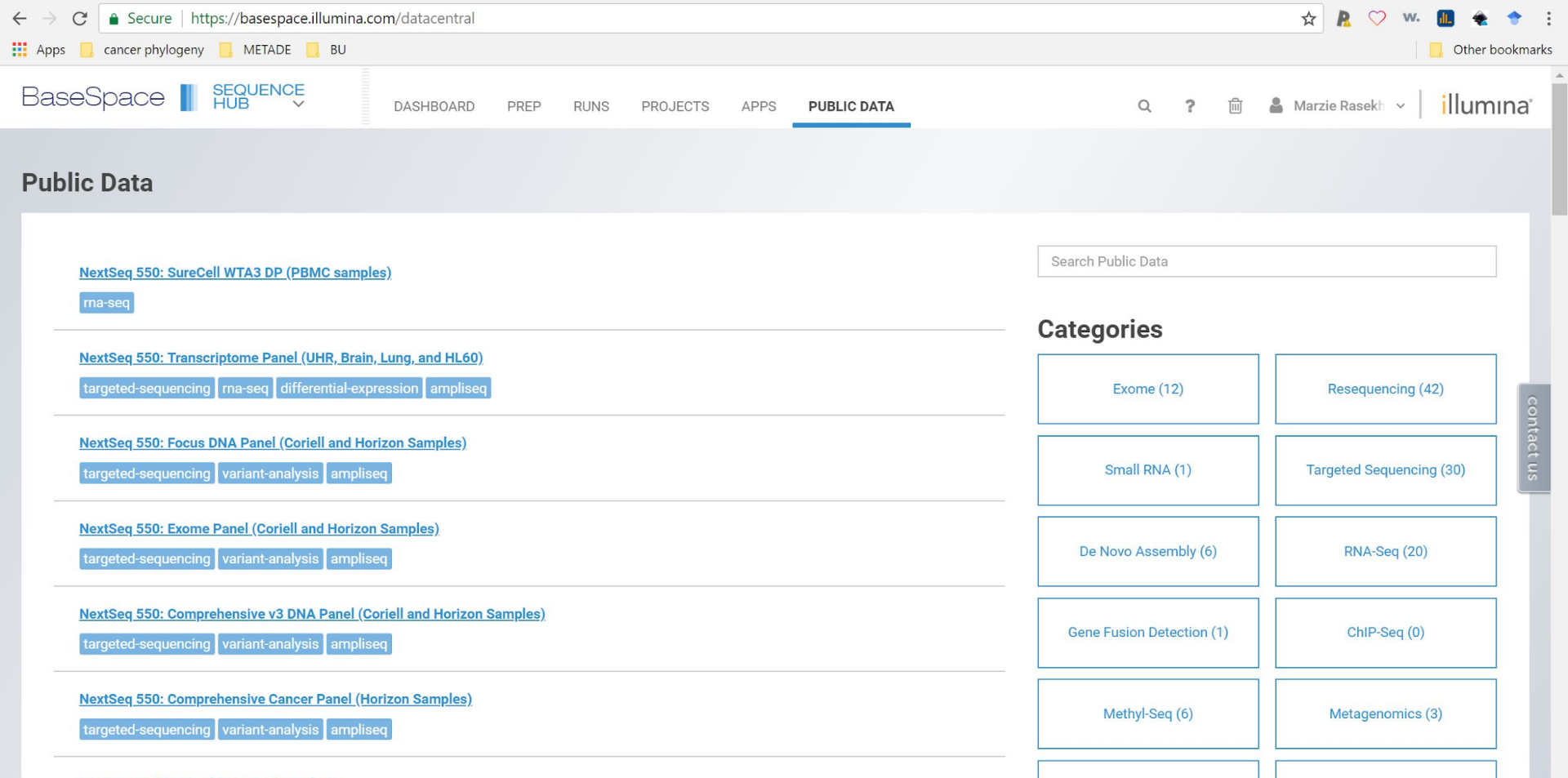
Public data and platforms

- NCBI (https://www.ncbi.nlm.nih.gov/sra)
- Illumina basespace (https://basespace.illumina.com/home/index)
- Google genomics cloud (https://console.cloud.google.com/genomics/)
- Genome In A Bottle (GIAB) (http://jimb.stanford.edu/giab/)
- REPOSITIVE (https://discover.repositive.io/datasets/)
- GDC (https://portal.gdc.cancer.gov/)
- Seven Bridges (https://igor.sbgenomics.com/)

NCBI SRA portal

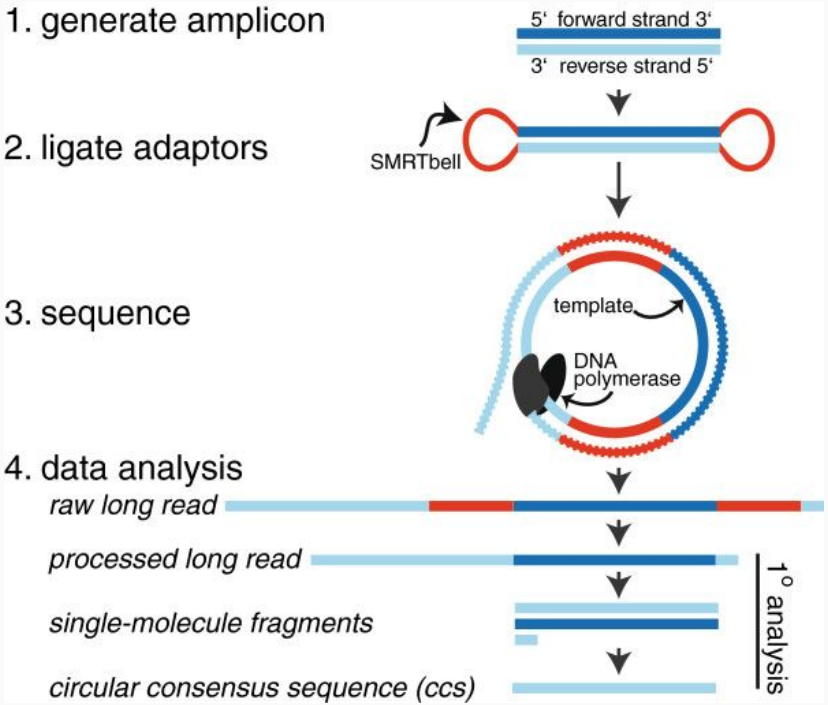


Illumina BaseSpace



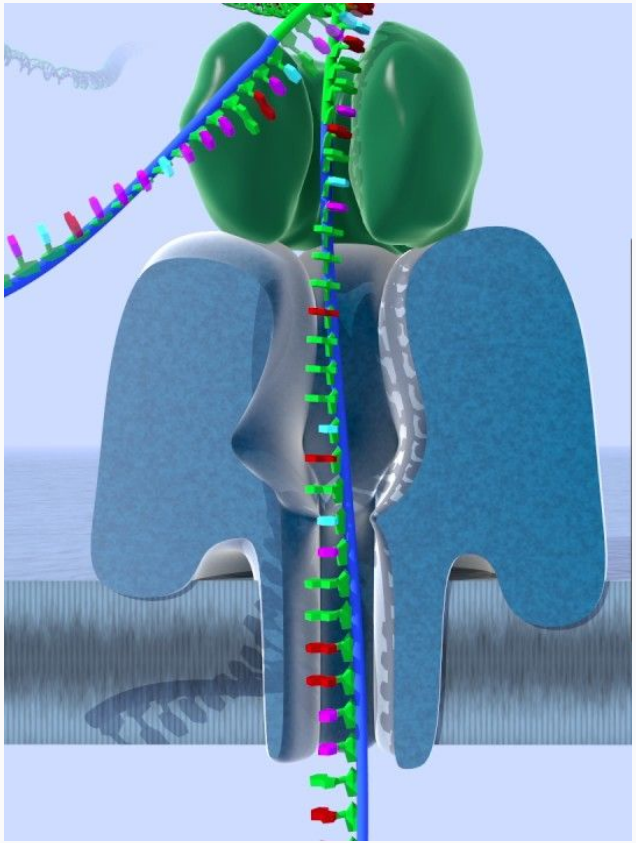
3rd Generation Sequencing: PacBio

- Pacific Biosciences (PacBio)
- “Long read”
- “Single molecule”
- SMRT



3rd Gen Sequencing: Oxford Nanopore

- “Real time single molecule”



https://www.youtube.com/watch?v=GUb1TZvMWsw