

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

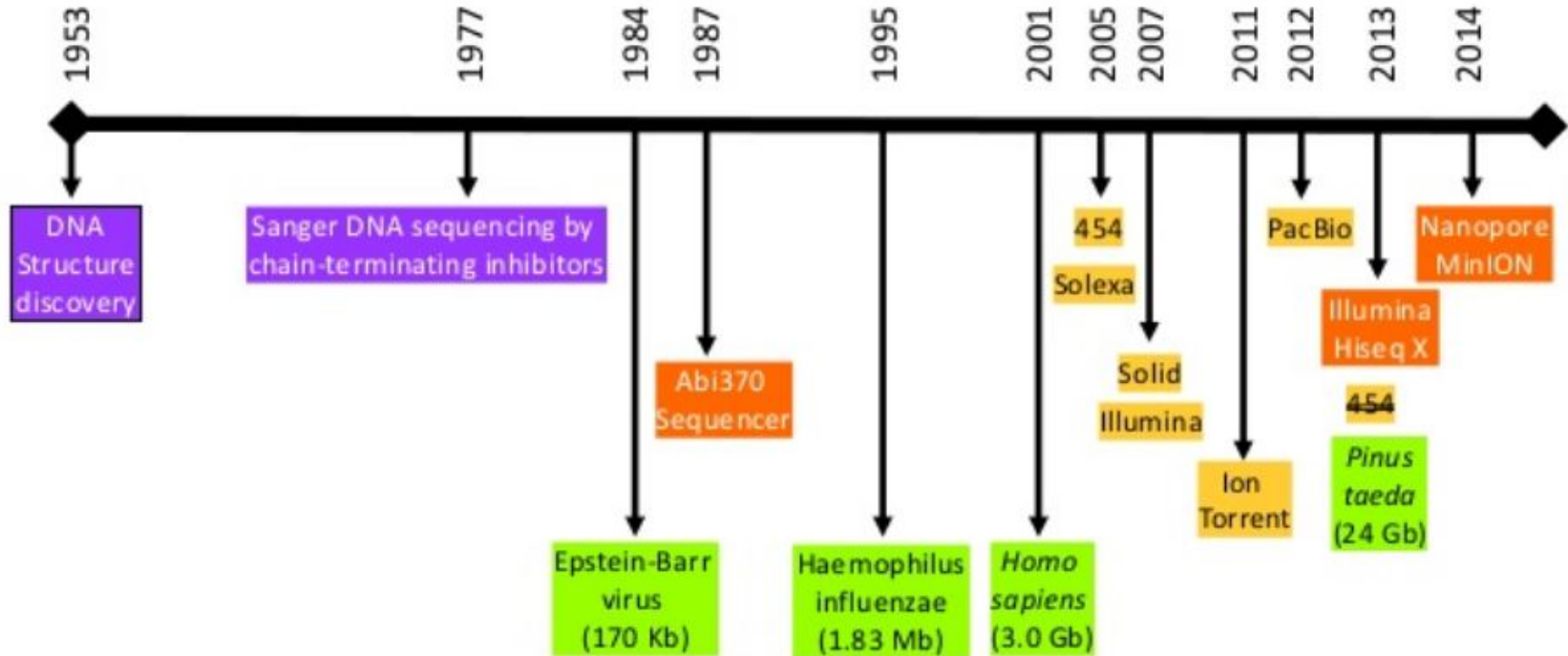
High-throughput Sequencing

- **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



Sequencing over the Ages



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



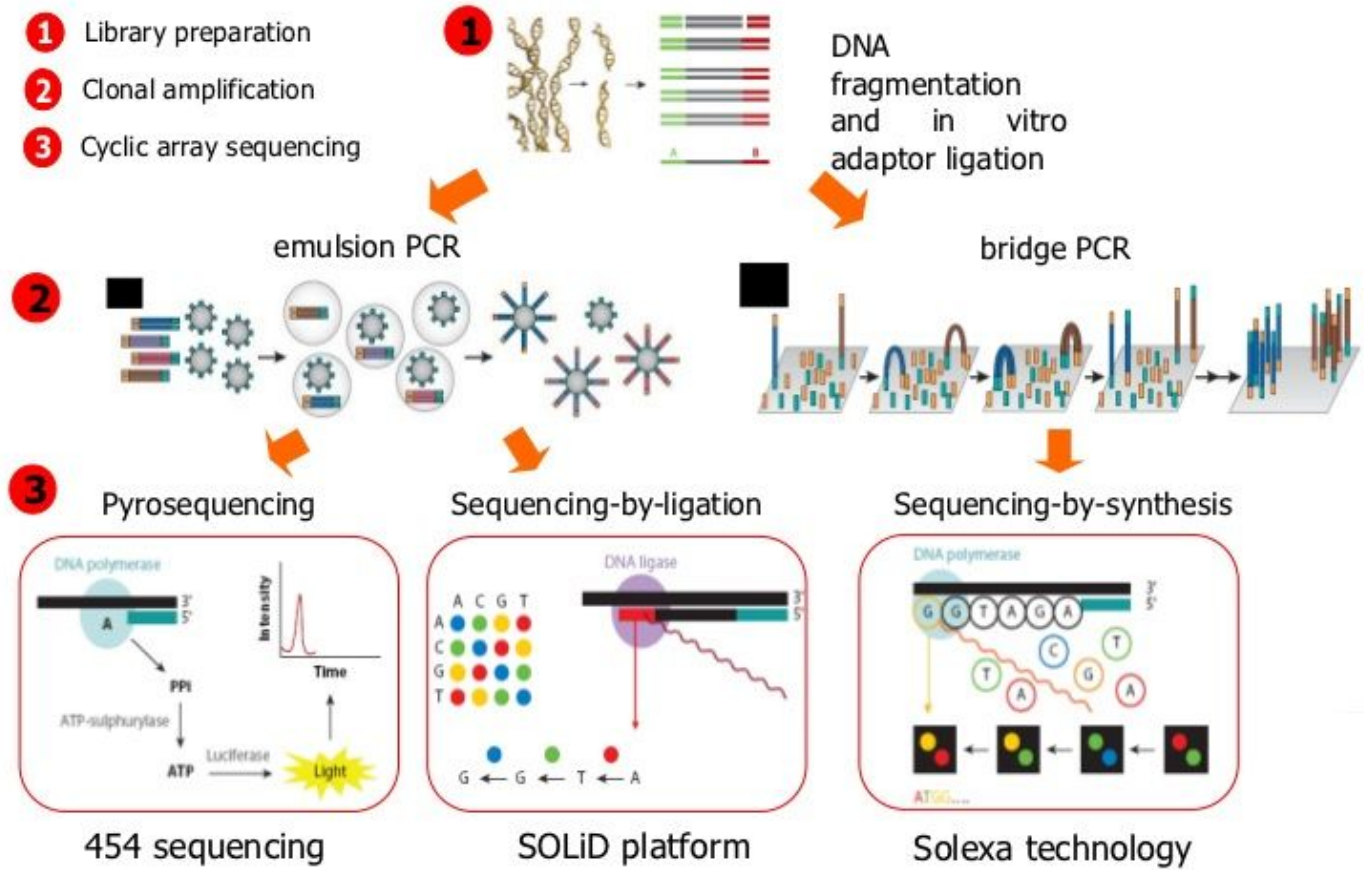
Illumina HiSeq 2500



Illumina NovaSeq 6000



Experiment



Comparison

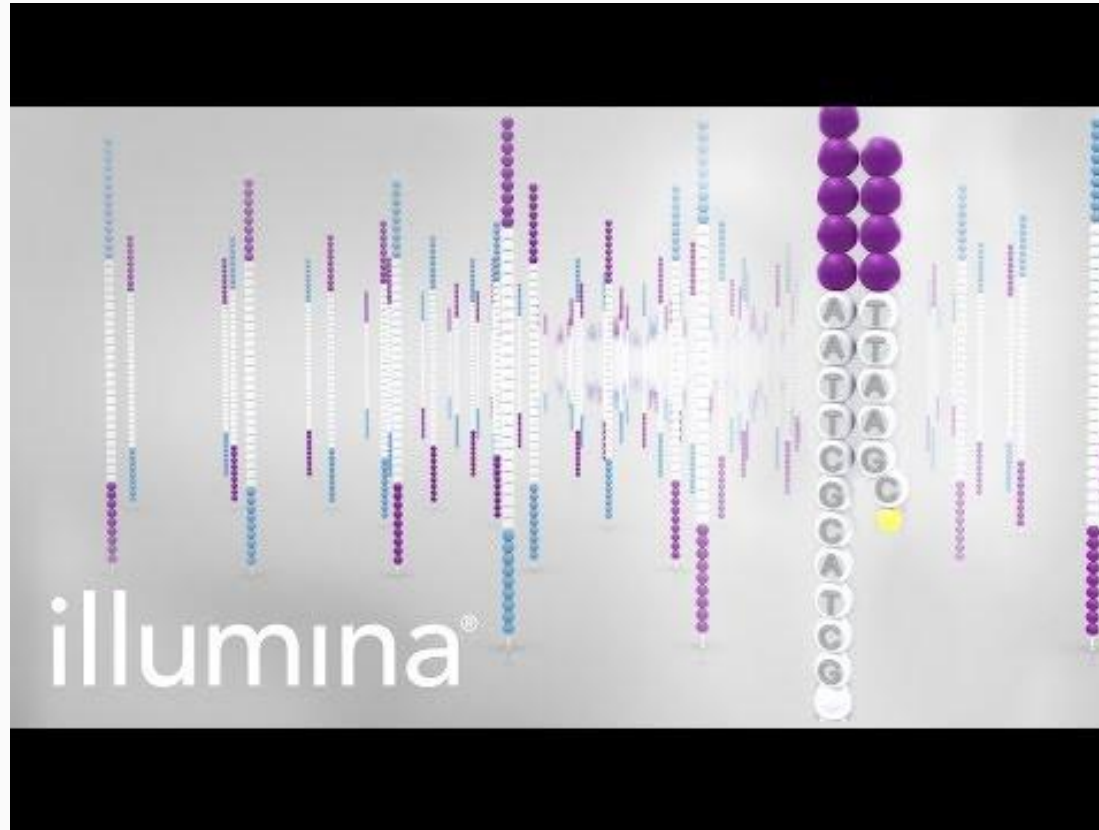
| | Run Time | Read Length | Quality | Total nucleotides sequenced | Cost /MB |
|---------------------|------------|--------------|-------------------------------|-----------------------------|-----------------|
| 454 Pyrosequencing | 24h | 700 bp | Q20-Q30 | 1 GB | \$10 |
| Illumina Miseq | 27h | 2x300bp | > 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

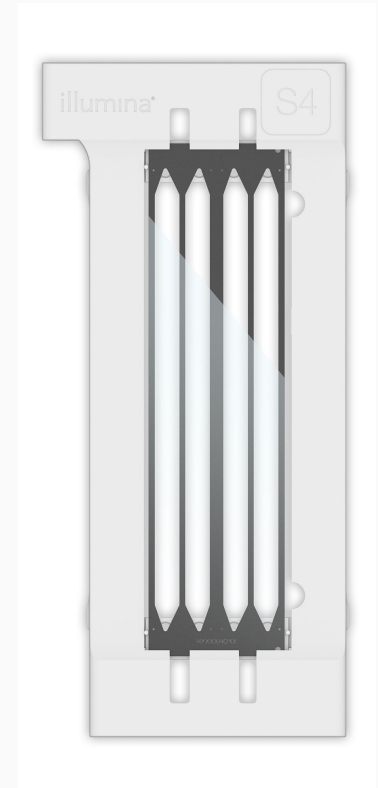
Illumina Sequencing Process



<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



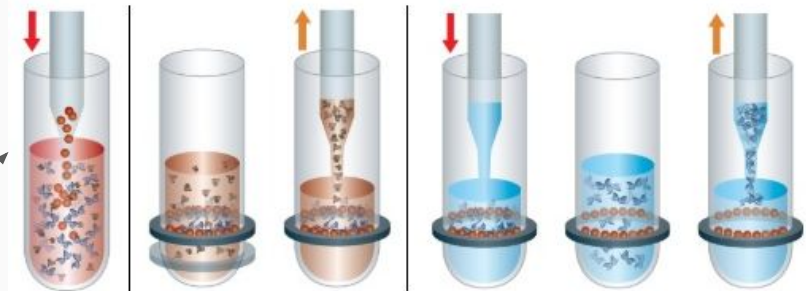
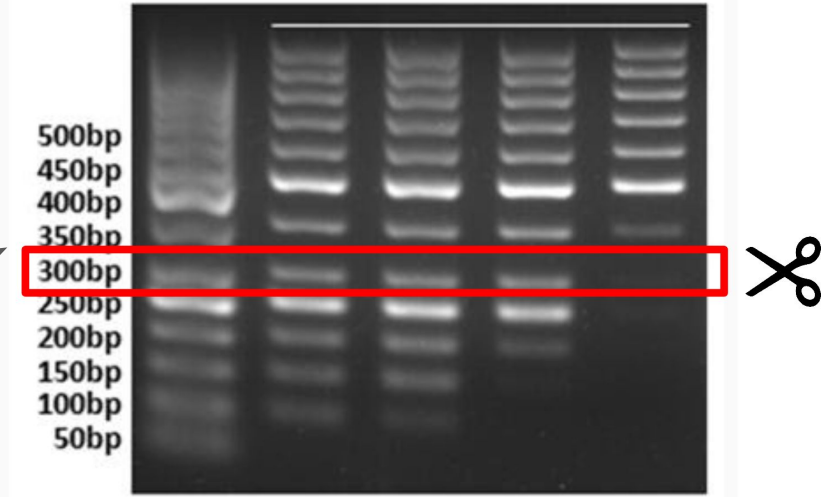
NovaSeq Flowcell
Courtesy of Illumina, Inc. 6

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

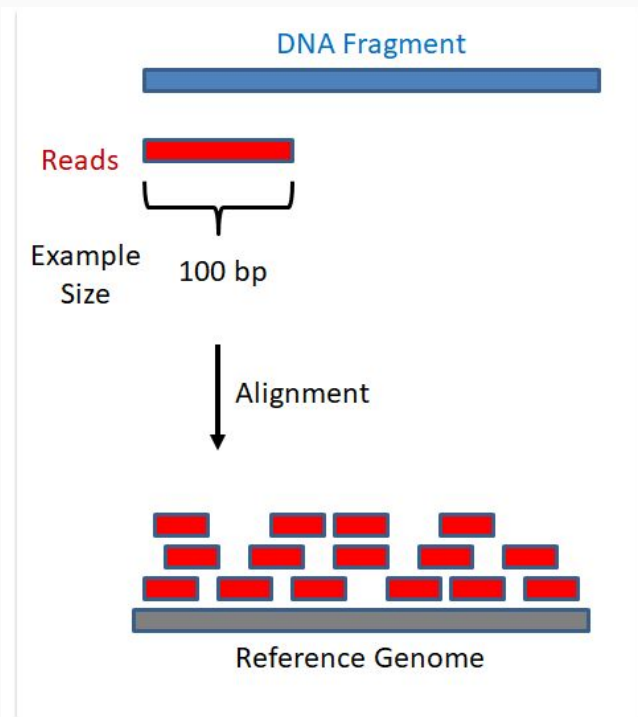
Design Choice: Fragment Length

- 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)

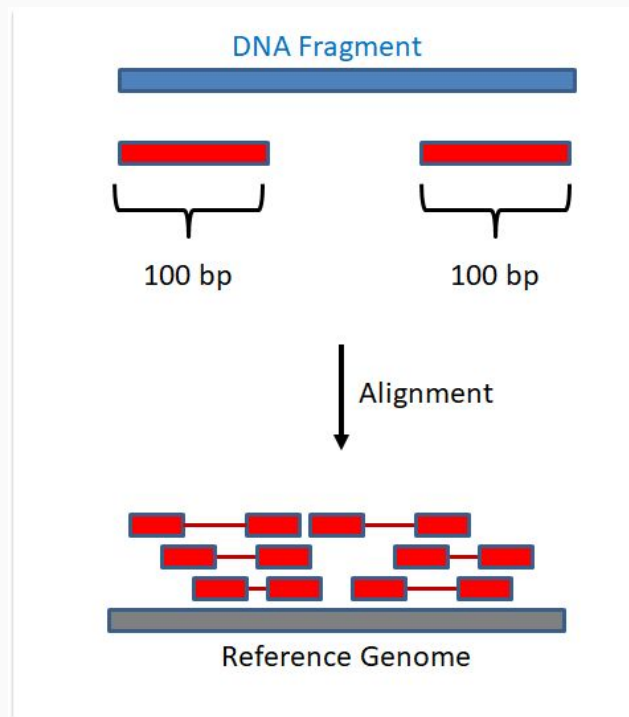


Design Choice: Single End vs Paired End

Single End



Paired End



Design Choice: Number of Reads

- Each sequencing run generates a # of total reads
- # of reads/sample \sim # 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

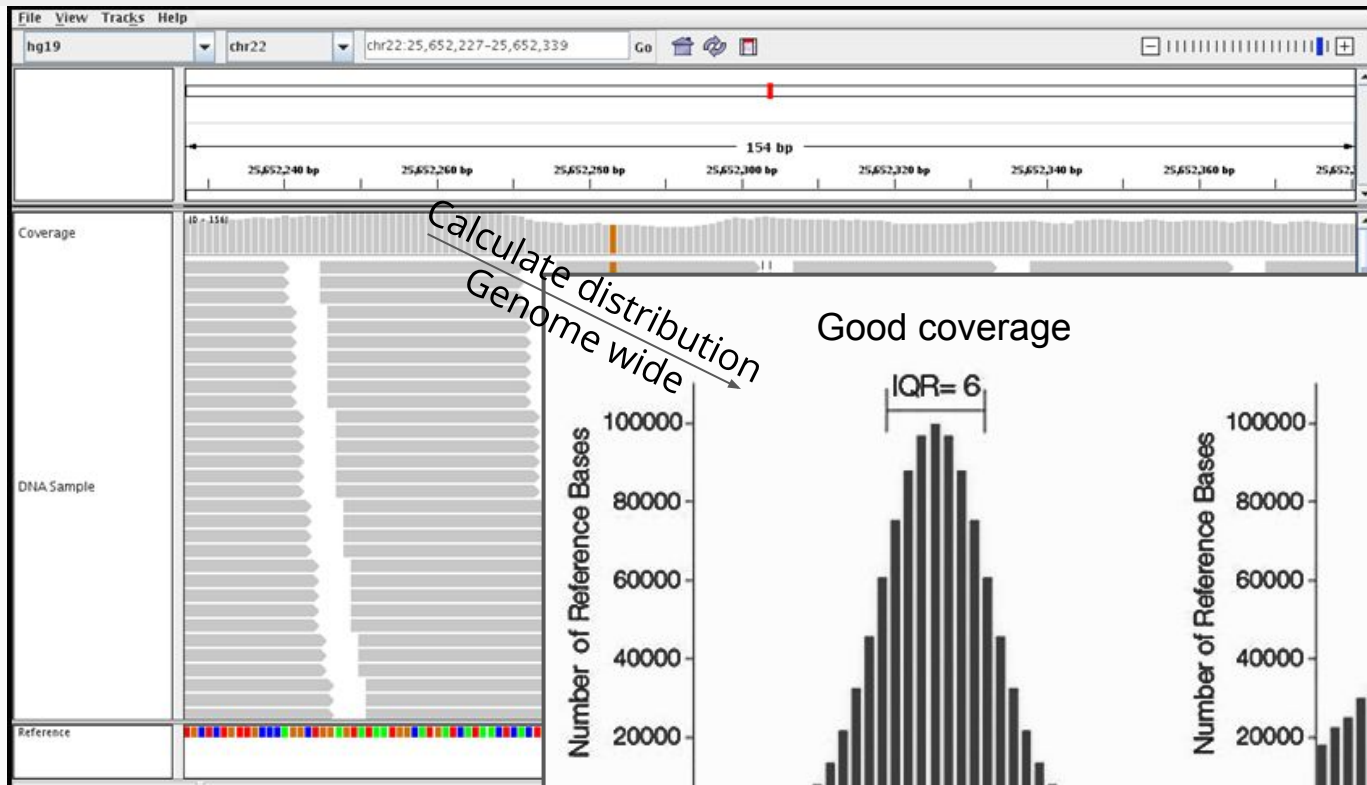
Mapped or Aligned reads



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

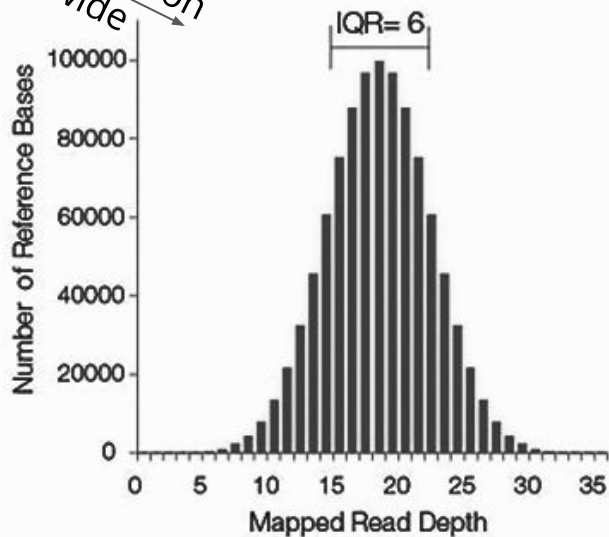
Coverage:
fraction of genomic locus covered by at least one read

Coverage - Whole Genome Sequencing

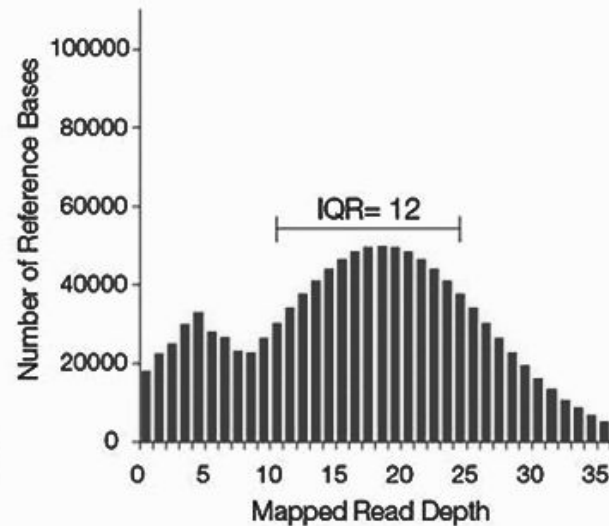


Calculate distribution
Genome wide

Good coverage



Bad coverage



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
NTTGTAGCTGAGGAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBFBFBFBFBFBF#####
```

fastq format

unique read header

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBFBFBFBFBFBF#####
```

fastq format

comments separated by space, could be anything

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBFBFBFBF#####
```


fastq format

Sequence of the read

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBFBFBFBFBFBF#####
```

fastq format

start quality line



```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTNNNNNN
+SRR1997412.2 2 length=125
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBFBFBFBFBFBF#####
```

fastq format

repeat read header and comment, not required, often blank

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBFBFBFBFBFBF#####
```

fastq format

Quality sequence of the read, in ASCII

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAGGTACCAGCACTCAGATGGAATGGTTTTGAACTCAGTCCA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997412.2 2 length=125
GTATTTAGTCATGTAAGACTCCTTAACCAGCTAACTTAAGAAAGACTTCTAGGACAGAATAGGTTACACTAGTTATAATTTTNNNNNN
+SRR1997412.2 2 length=125
BFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBFBFBFBF####
```


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/>)
- REPOSITORY (<https://discover.repositive.io/datasets/>)
- GDC (<https://portal.gdc.cancer.gov/>)
- Seven Bridges (<https://igor.sbgenomics.com/>)

NCBI SRA portal

Secure https://www.ncbi.nlm.nih.gov/sra/?term=NA12878

Apps cancer phylogeny METADE BU Other bookmarks

NCBI Resources How To Sign in to NCBI

SRA SRA NA12878 Search
Create alert Advanced Help

Access
Controlled (19)
Public (2,714)

Source
DNA (2,705)
RNA (24)

Type
exome (93)
genome (2,149)

Other
aligned data (358)

Summary 20 per page Send to: Filters: Manage Filters

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

Search results

Items: 1 to 20 of 2733 << First < Prev Page 1 of 137 Next > Last >>

- [WES of reference sample NA12878 with AmpliSeq Exome](#)
 1. 1 ION_TORRENT (Ion Torrent Proton) run: 38.3M spots, 7.3G bases, 3.8Gb downloads
Accession: SRX3570665
- [WES of reference sample NA12878 with AmpliSeq Exome](#)
 2. 1 ION_TORRENT (Ion Torrent Proton) run: 46M spots, 7.6G bases, 4.3Gb downloads
Accession: SRX3570664
- [Illumina paired-end 151 bp WGS of NA12878](#)
 3. 1 ILLUMINA (Illumina HiSeq 2500) run: 855.6M spots, 258.4G bases, 77.7Gb downloads
Accession: SRX3538696
- [Targeted Sequencing of Tandem repeats](#)
 4. 1 ILLUMINA (Illumina MiSeq) run: 5.8M spots, 3.5G bases, 1.9Gb downloads
Accession: SRX3472658
- [Targeted Sequencing of Tandem repeats](#)
 5. 1 ILLUMINA (Illumina MiSeq) run: 3.9M spots, 2G bases, 1.1Gb downloads
Accession: SRX3472657

Results by taxon

Top Organisms [\[Tree\]](#)
Homo sapiens (2731)
unidentified (1)
artificial sequences (1)

Top Bioprojects

Production ENCODE epigenomic... (6)
Production ENCODE functional... (5)
Production ENCODE transcript... (2)

Search in related databases

| Database | Access | | all |
|--------------|---------------------|--------------------|---------------------|
| | public | controlled | |
| BioSample | 265 | 18 | 283 |
| BioProject | 32 | 1 | 33 |
| dbGaP | | 20 | 20 |
| GEO Datasets | 36 | | 36 |

Find related data

Database: [Select](#)

Illumina BaseSpace

Public Data

[NextSeq 550: SureCell WTA3 DP \(PBMC samples\)](#)

rna-seq

[NextSeq 550: Transcriptome Panel \(UHR, Brain, Lung, and HL60\)](#)

targeted-sequencing rna-seq differential-expression ampliseq

[NextSeq 550: Focus DNA Panel \(Coriell and Horizon Samples\)](#)

targeted-sequencing variant-analysis ampliseq

[NextSeq 550: Exome Panel \(Coriell and Horizon Samples\)](#)

targeted-sequencing variant-analysis ampliseq

[NextSeq 550: Comprehensive v3 DNA Panel \(Coriell and Horizon Samples\)](#)

targeted-sequencing variant-analysis ampliseq

[NextSeq 550: Comprehensive Cancer Panel \(Horizon Samples\)](#)

targeted-sequencing variant-analysis ampliseq

Search Public Data

Categories

| | |
|---------------------------|--------------------------|
| Exome (12) | Resequencing (42) |
| Small RNA (1) | Targeted Sequencing (30) |
| De Novo Assembly (6) | RNA-Seq (20) |
| Gene Fusion Detection (1) | ChIP-Seq (0) |
| Methyl-Seq (6) | Metagenomics (3) |

contact us

3rd Generation Sequencing: PacBio

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



1. generate amplicon

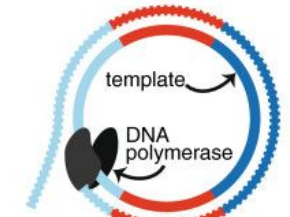
5' forward strand 3'

3' reverse strand 5'

2. ligate adaptors



3. sequence



4. data analysis

raw long read

processed long read

single-molecule fragments

circular consensus sequence (ccs)

1° analysis

3rd Gen Sequencing: Oxford Nanopore

- “Real time single molecule”

