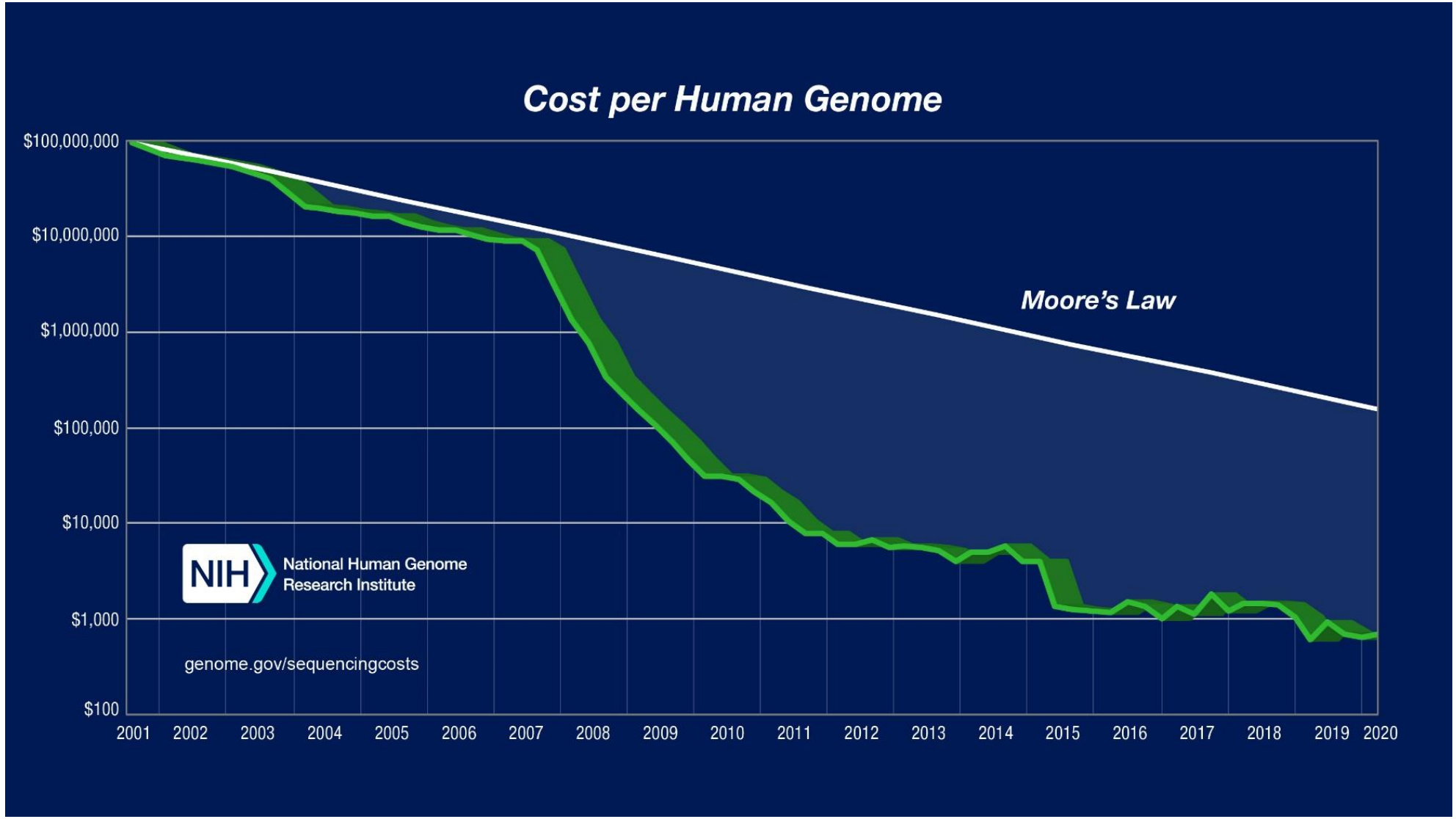
Next generation sequencing (NGS) introduction

High Throughput sequencing

History of Illumina sequencing

NovaSeq 6000

Sequencing costs have decreased massively over time

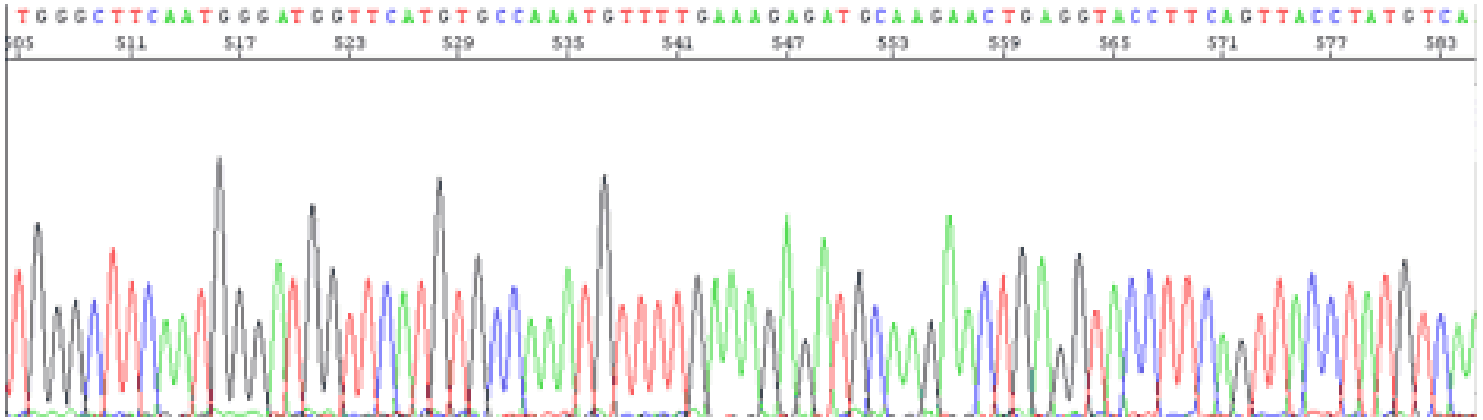


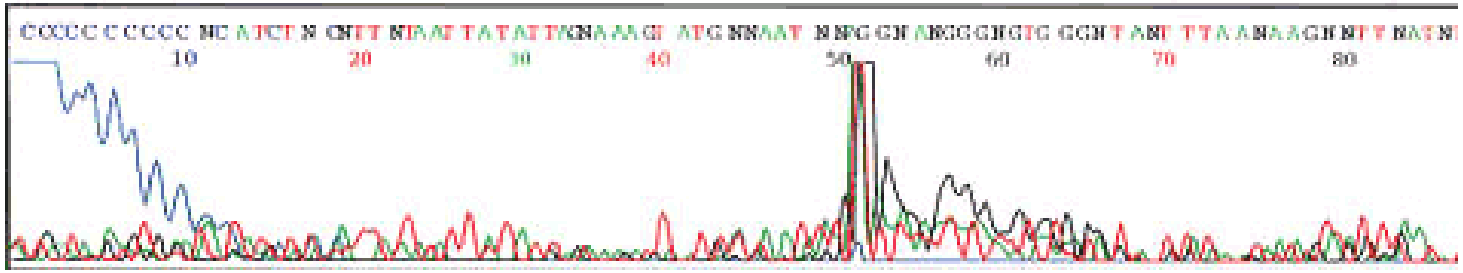
High Throughput Sequencing (=Next Generation Sequencing)

- **Short-read sequencing technologies** (2nd generation):
  - Sequence millions of clonally amplified molecules
  - Reads typically 150 bp long
  - Illumina

- **Long-read technologies** (3rd generation):
  - Single molecules are sequenced in real-time, fast but expensive and high error rates
  - Reads typically kb long
  - PacBio
  - Nanopore

Sanger Sequencing

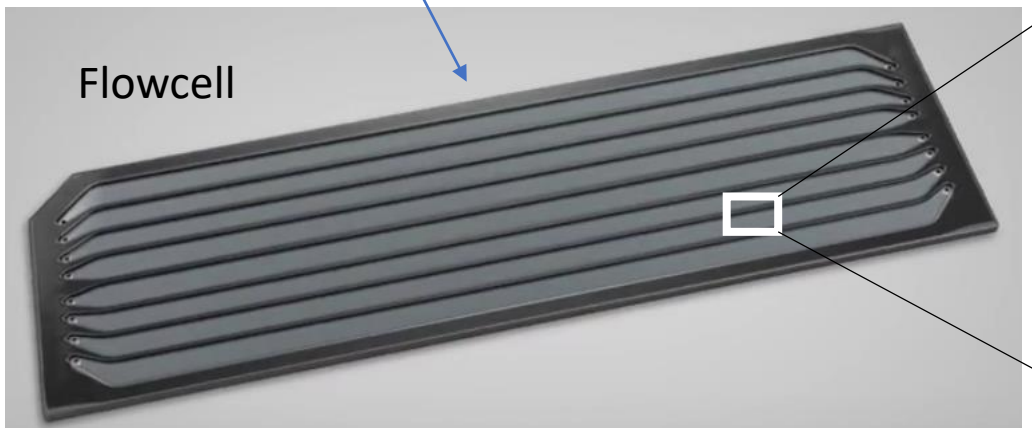




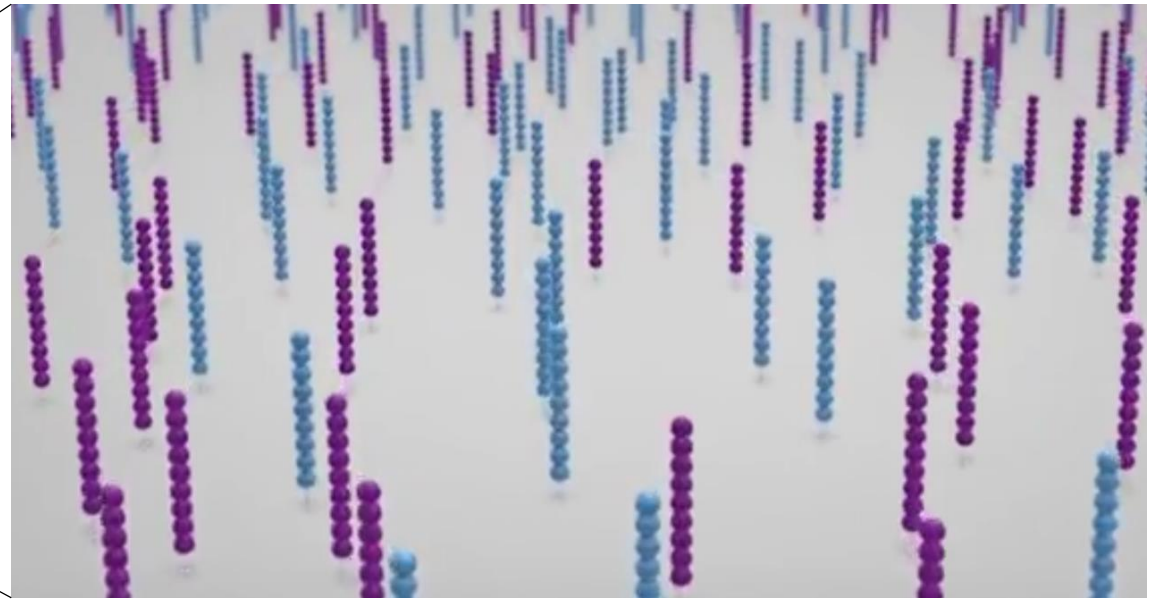
- Manually check each sequence
- Resequence failed sequences

Illumina flowcell: millions of DNA sequences

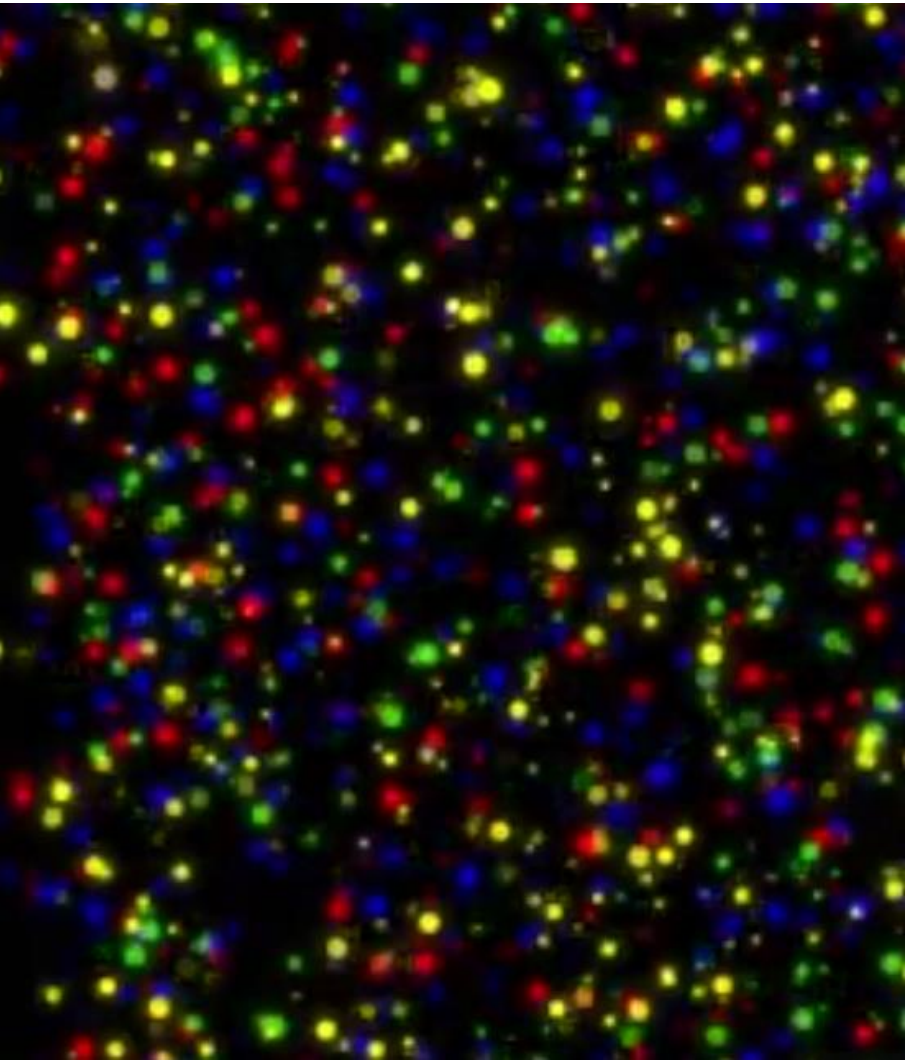
DNA fragments with Illumina adapters

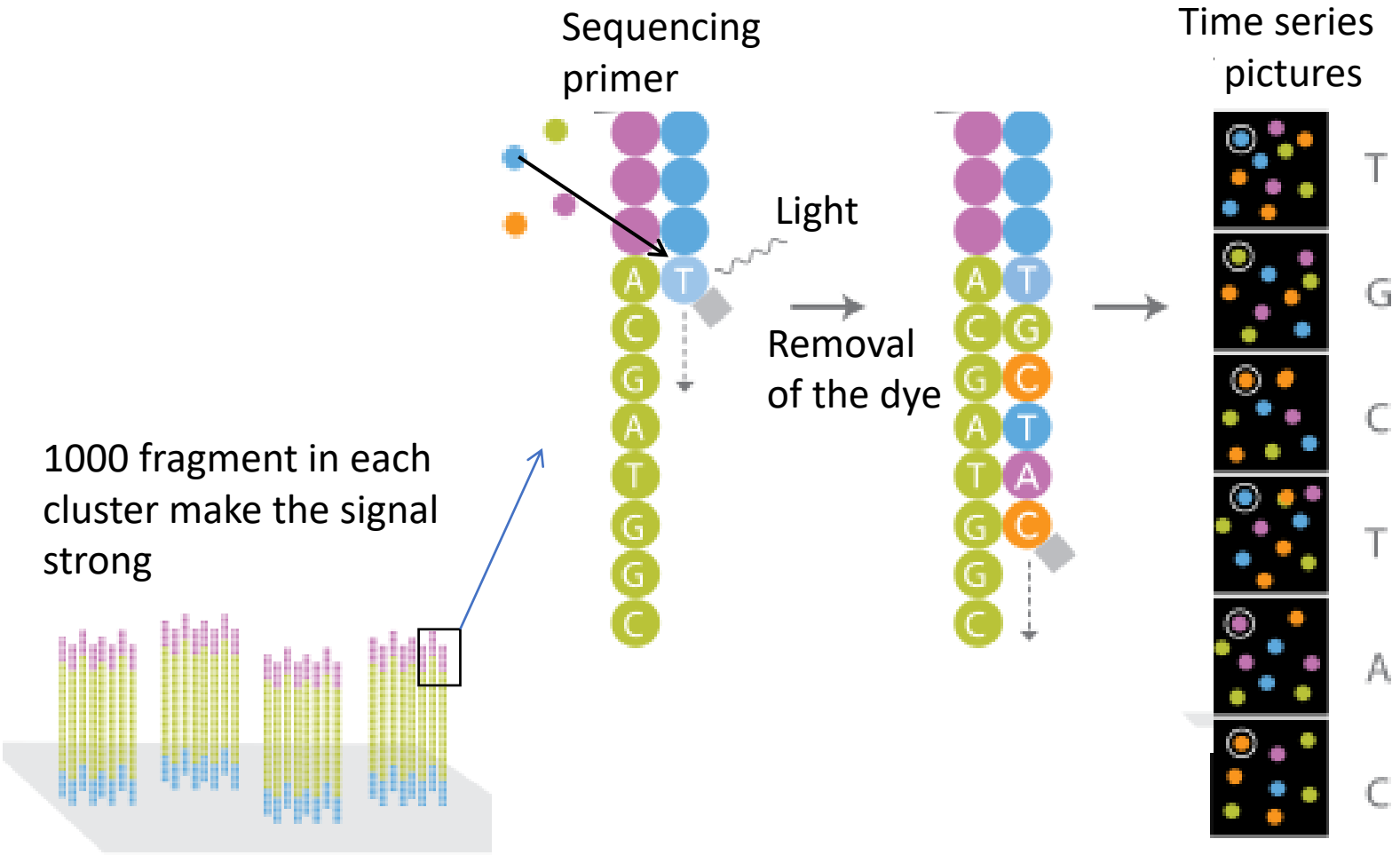


Each lane contains a dense lawn of Illumina primers



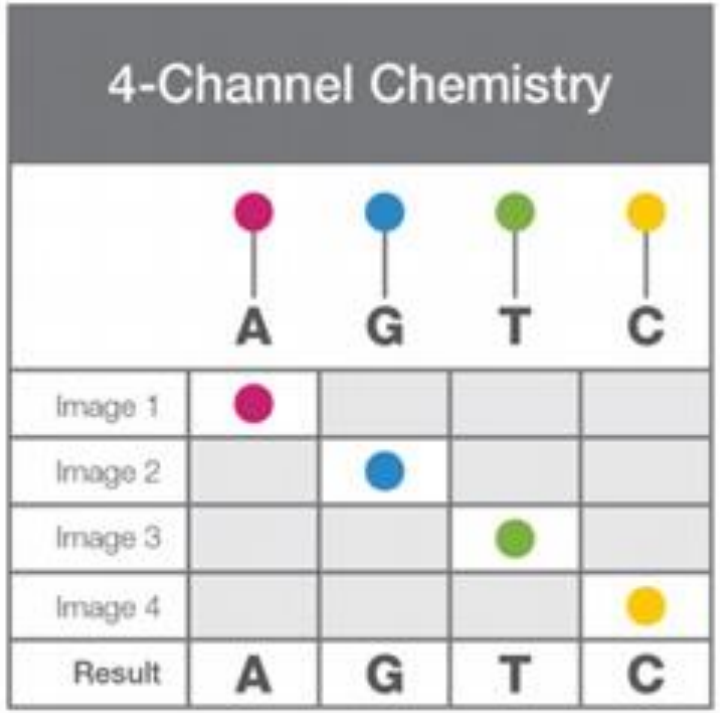
Sequencing by synthesis by Illumina

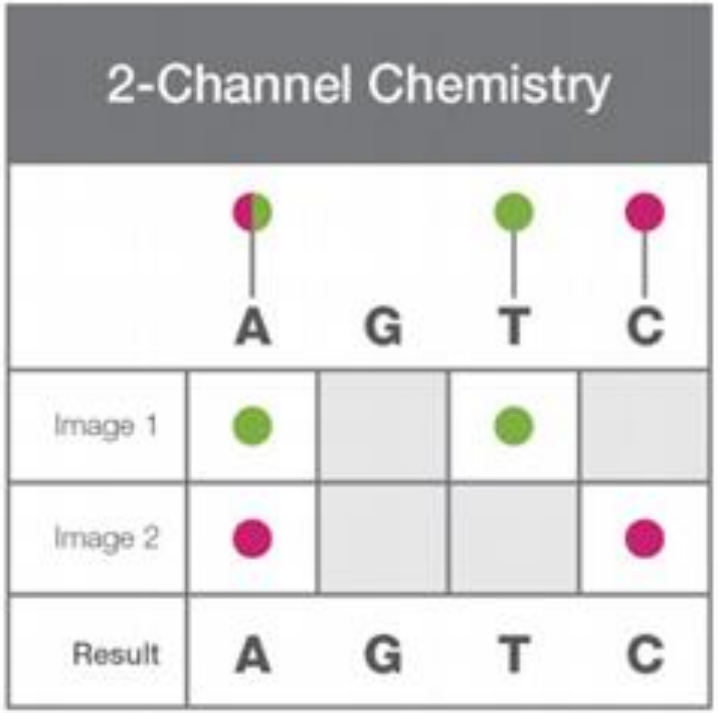




2-channel sequencing by synthesis

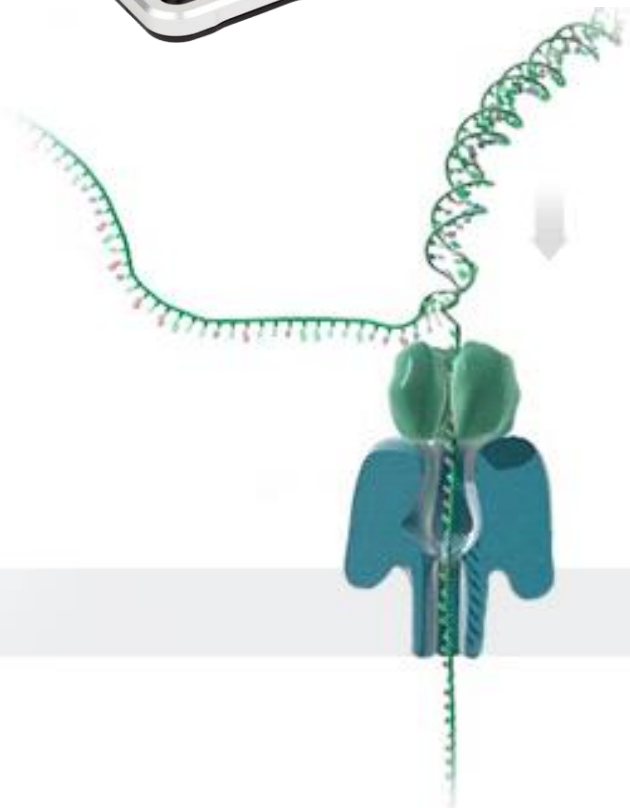
(used by these Illumina machines: Novaseq, Nextseq, MiniSeq)

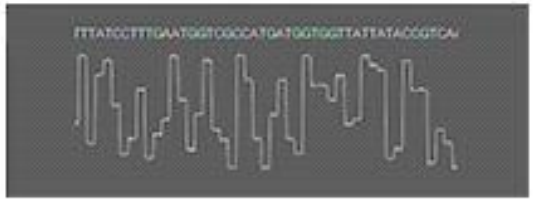




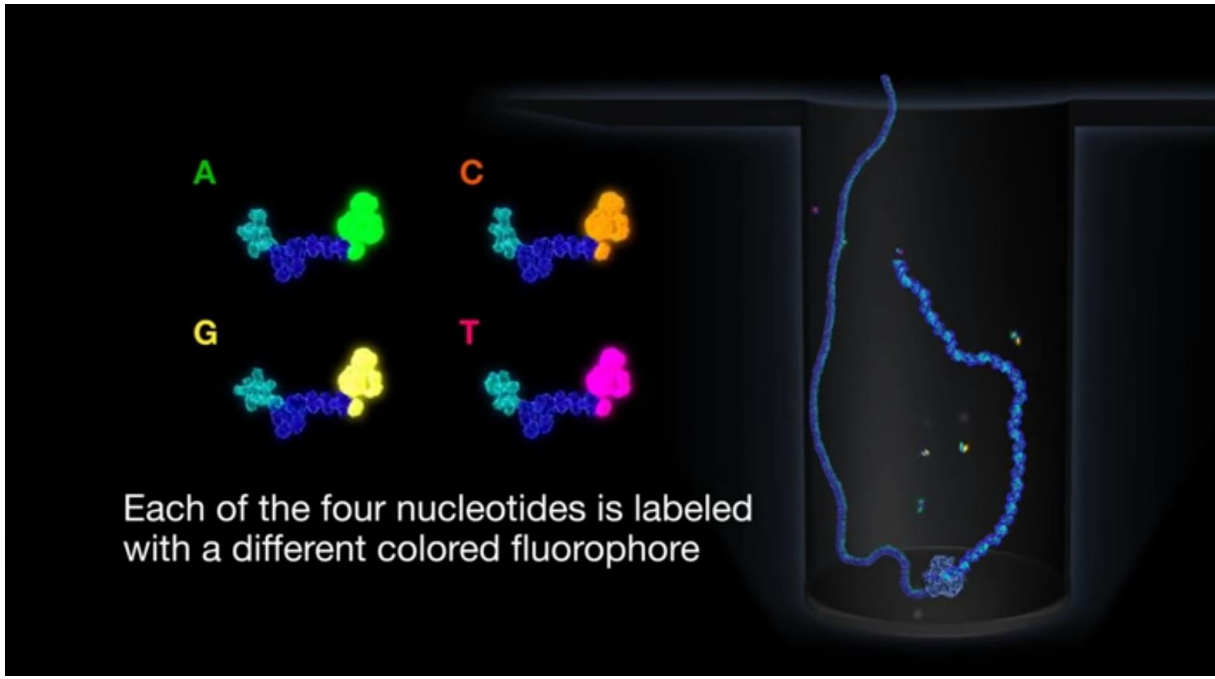
Long read sequencing technologies

Nanopore

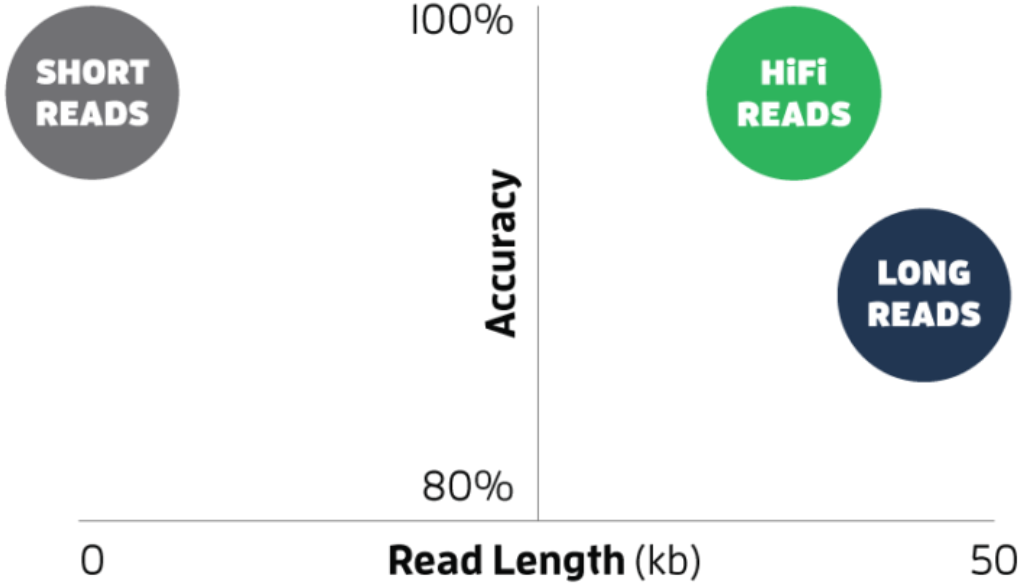


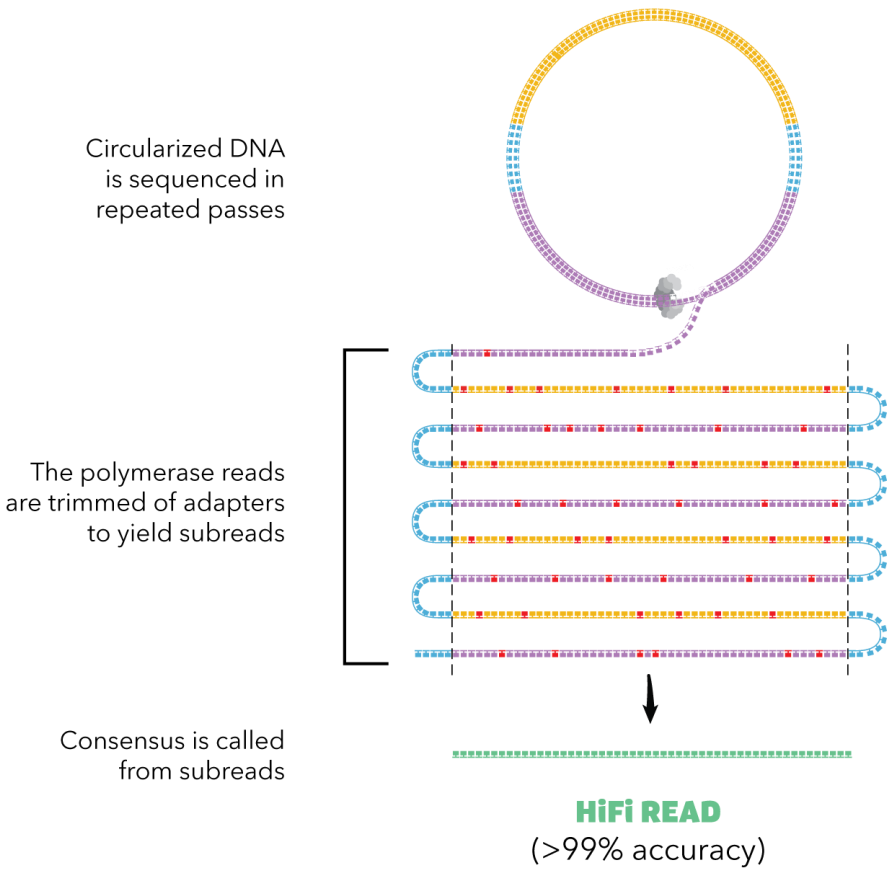


PacBio

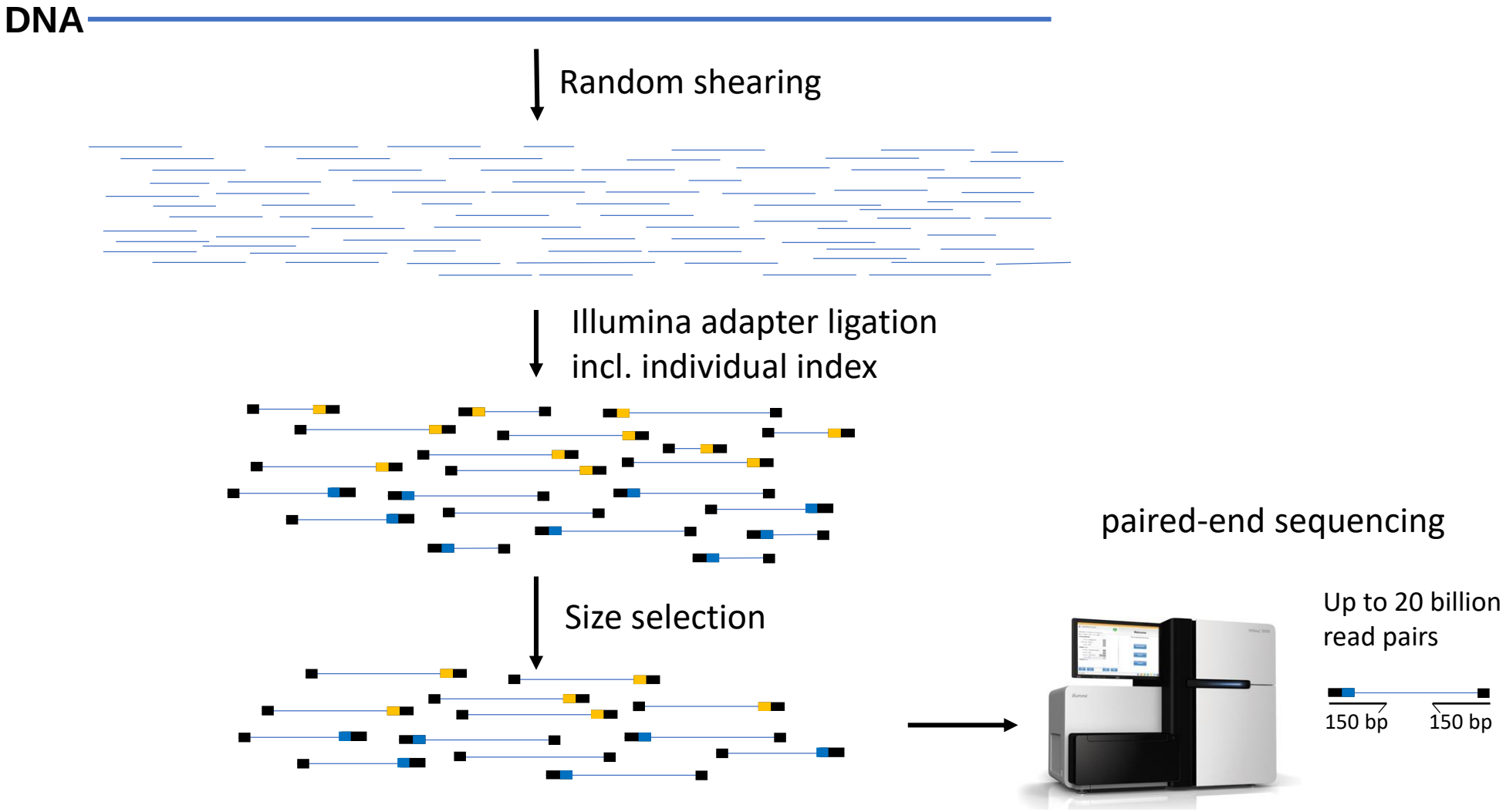


PacBio HiFi reads

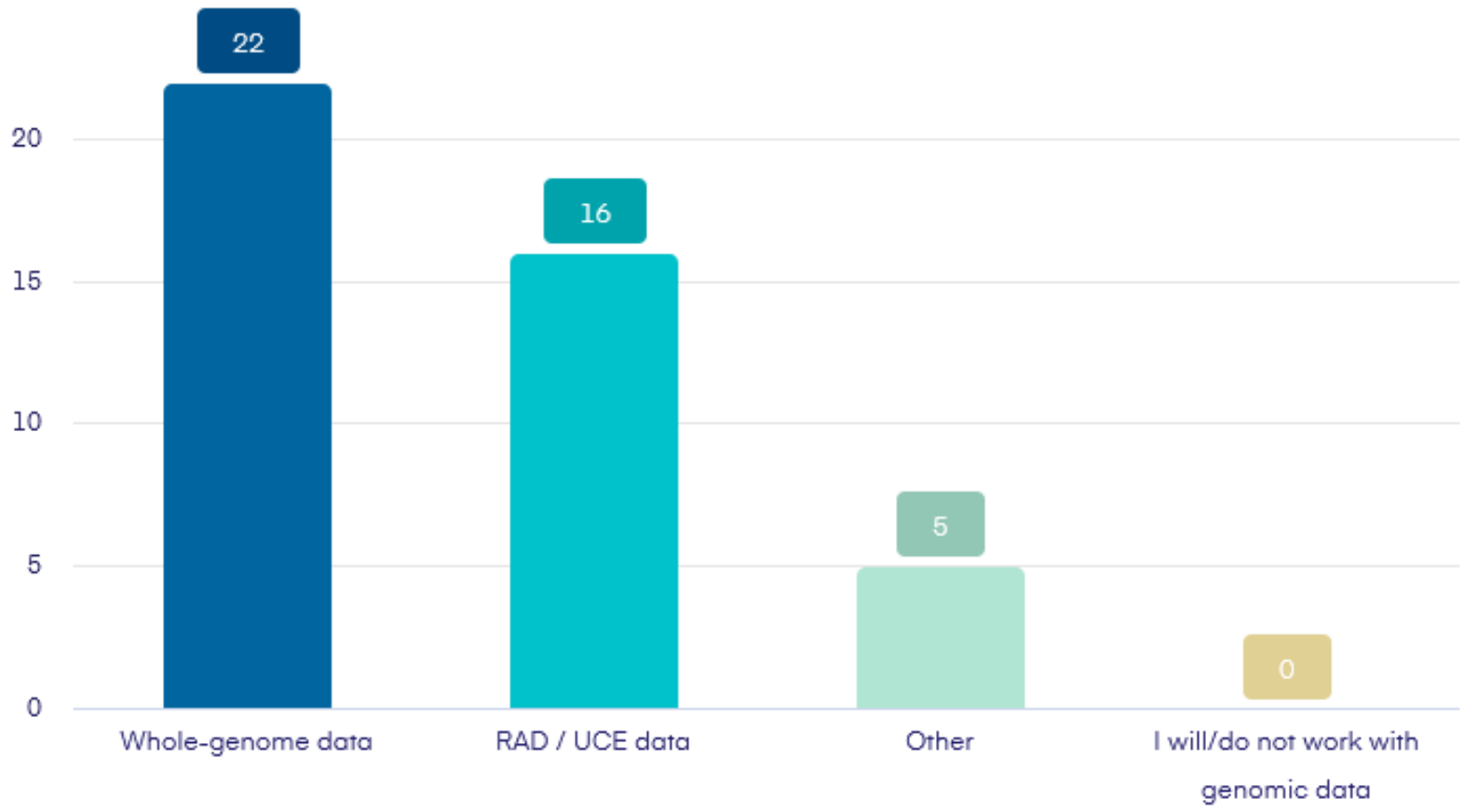




Whole-genome sequencing (shotgun sequencing)

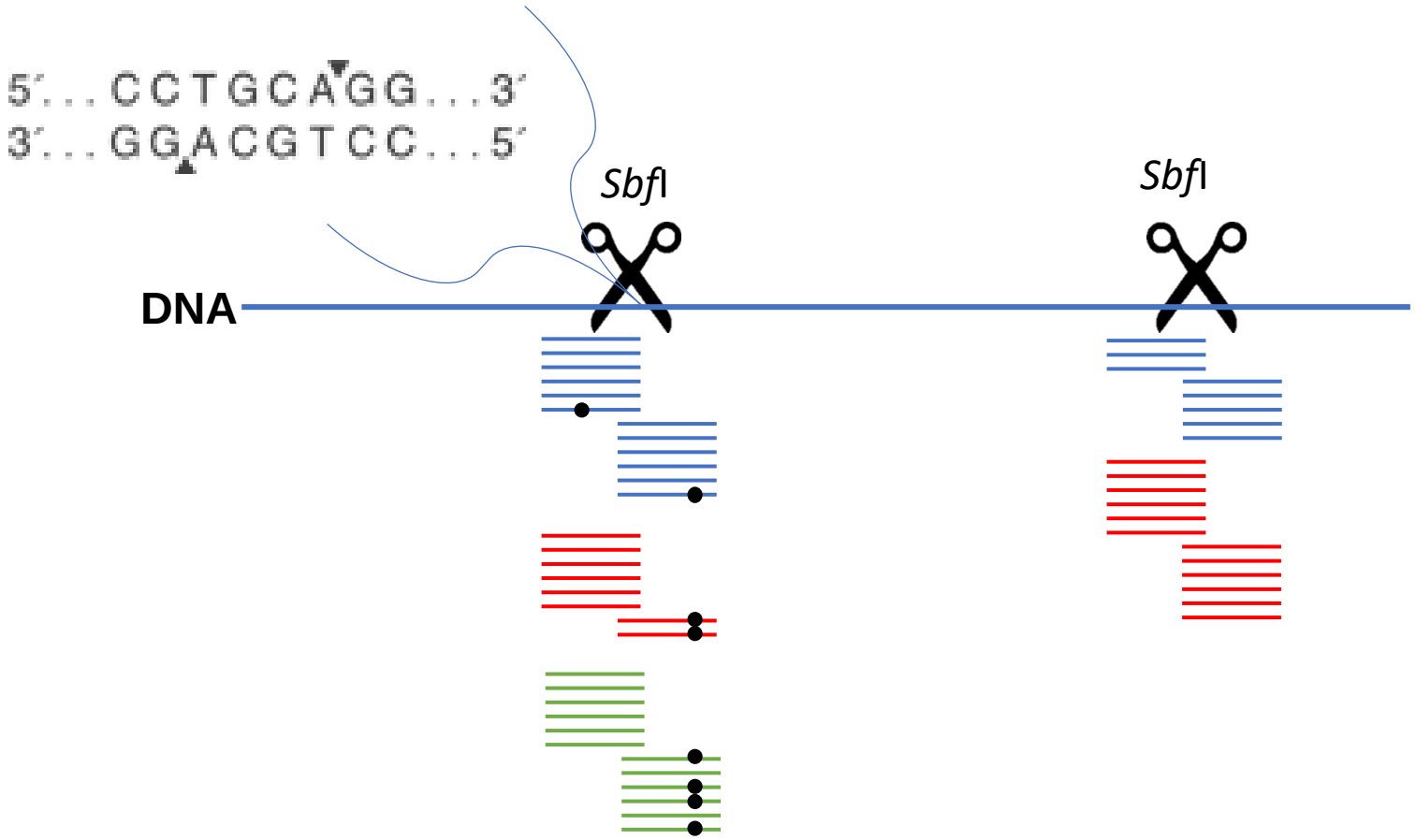


DNA preparation methods



RAD sequencing

Restriction **A**ssociated **D**NA **sequencing**



Trade-offs: Splitting reads (i.e. costs) among:

- Number of sites to sequence
- Number of samples
- Sequencing depth

- Example: 1 Hiseq2500 flow cell
  ~250 mio read pairs of 125 bp each -> 75 Gb data
  - 5 whole-genomes of a species with 1 Gb genome size at 15x coverage
  - 50 whole-genomes of a species with 500 Mb genome size at 3x coverage
  - 30 Mbp sequenced for 100 samples with a reduced-representation technique at a sequencing depth of 25

Considerations in choosing the library preparation and sequencing techniques

- Research question and planned analyses
- Genome size
- Availability & quality of reference genome (no ref genome -> not wgs)
- Available budget
- Number of samples to sequence (tradeoff with sequencing depth)
- Amounts of DNA available
- Sequencing depth aimed at
- SNP density required
- Divergence between samples
- Heterozygosity of samples
- Phase required
- Accuracy of each single position (if high needed, avoid PCR-based methods)
- Importance of annotations
- Neutral dataset or specific regions wanted

Fastq format

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
@HWUSI-EAS611:34:6669YAAXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
DF=DBD<BBFGGGGGGGBD@GGGD4@CA3CGG>DDD:D,B
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

Header (must start with @)

Base calls (sequence)

Quality scores

Quality scores

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
CCGTCAATTCATTAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA:9@::::??@@::FFAAAAACCAA:::BB@@?A?
```

ASCII encoding

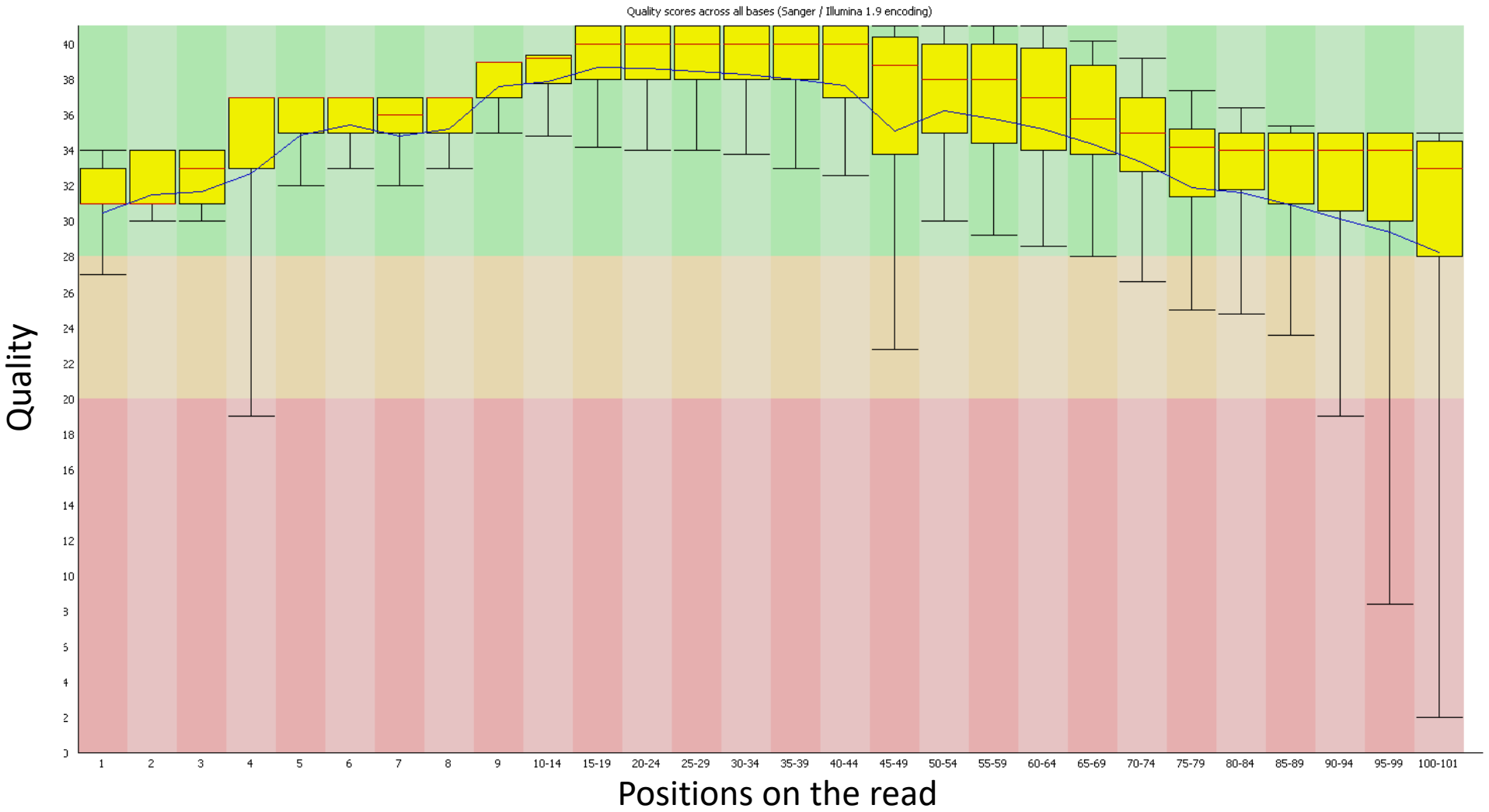
40:@	90:Z	141:a
41:A	91:[	142:b
42:B	92:\	143:c
43:C	93:]	144:d
44:D	94:^	145:e
45:E	95:_	146:f
… :…	… :…	… :…

$$\text{Phred} = -10 \log_{10} p$$

p = Probability call is incorrect

Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

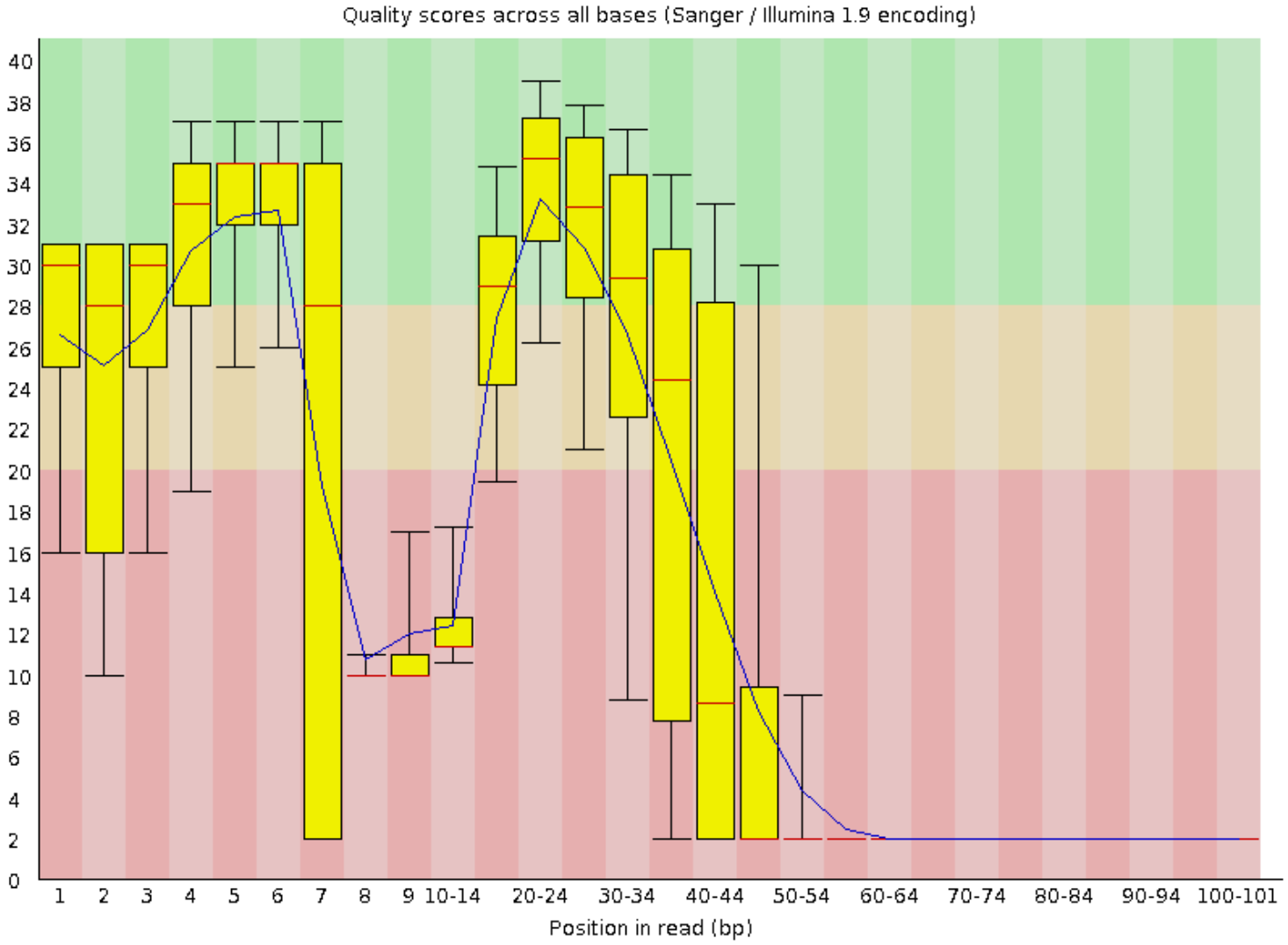
FastQC: Quality across bases (good example)



>Q30
Error prob. < 0.001

<Q20
Error prob. < 0.01

FastQC: Quality across bases (bad example)



Let’s have a look at the first few sequences and check the sequencing quality with fastqc