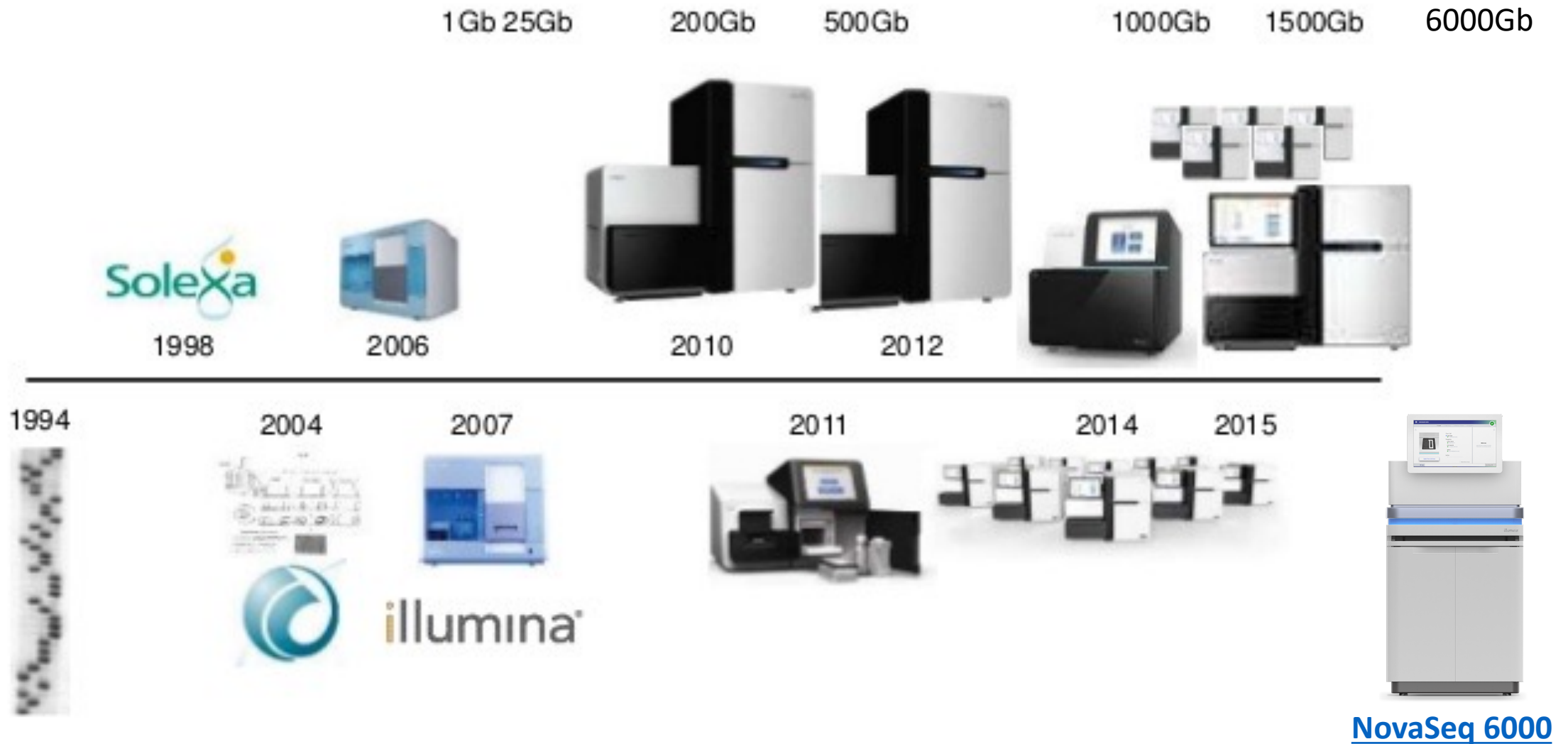
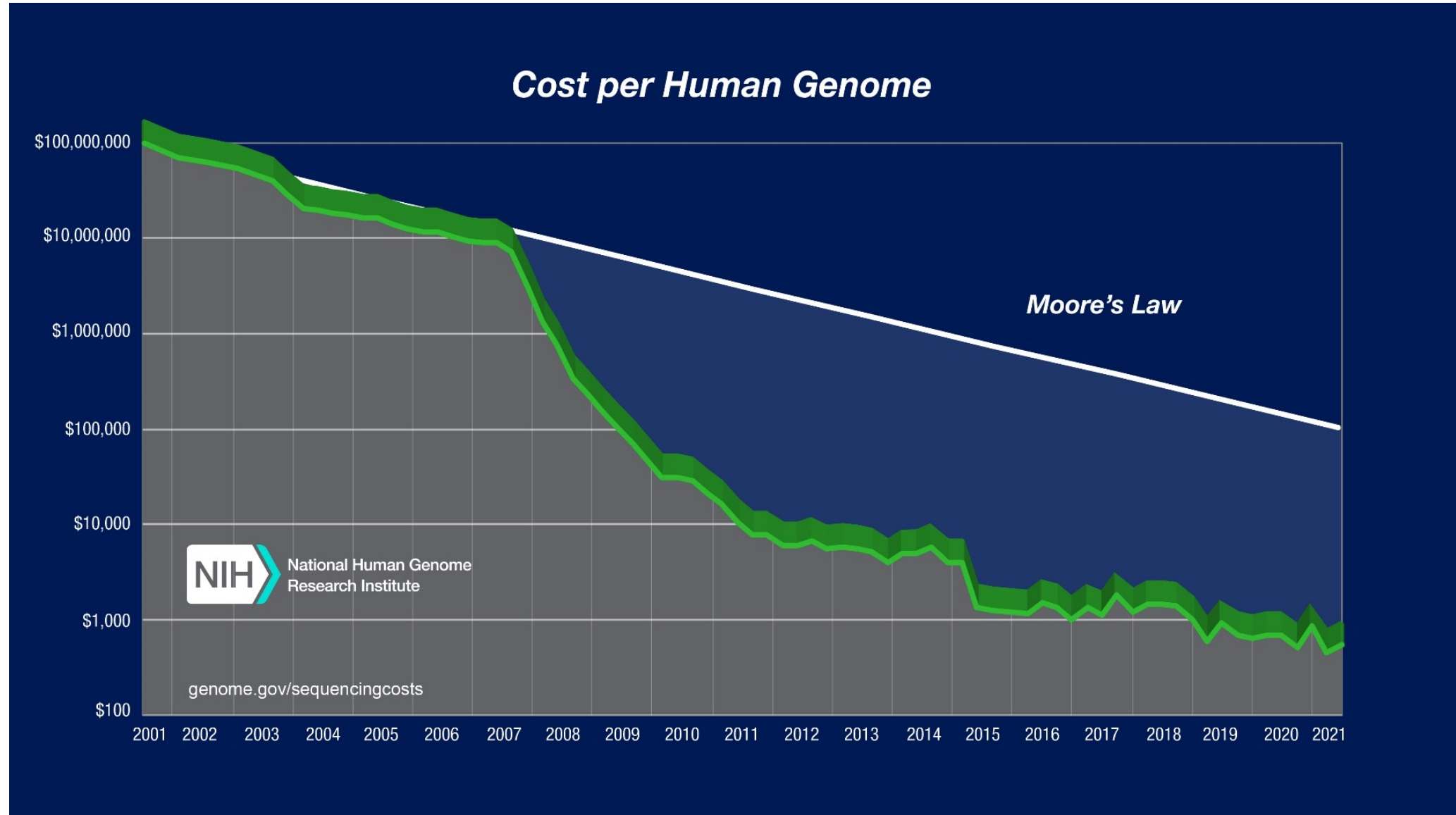


High Throughput sequencing or Next Generation Sequencing (NGS)

History of Illumina sequencing



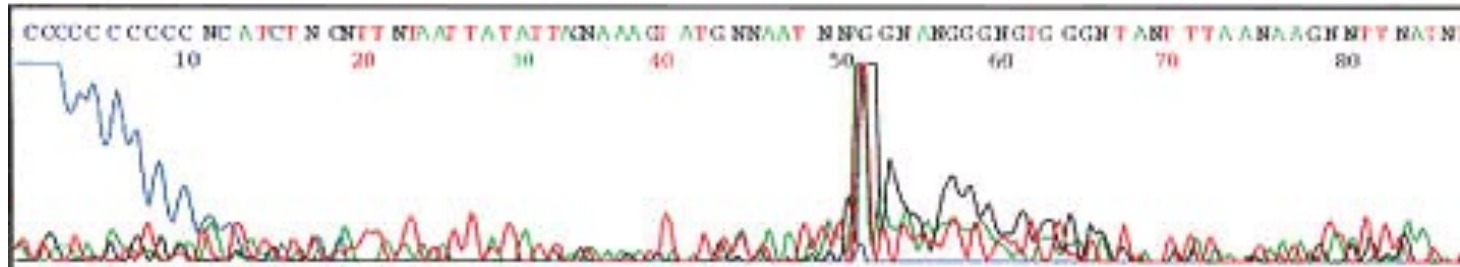
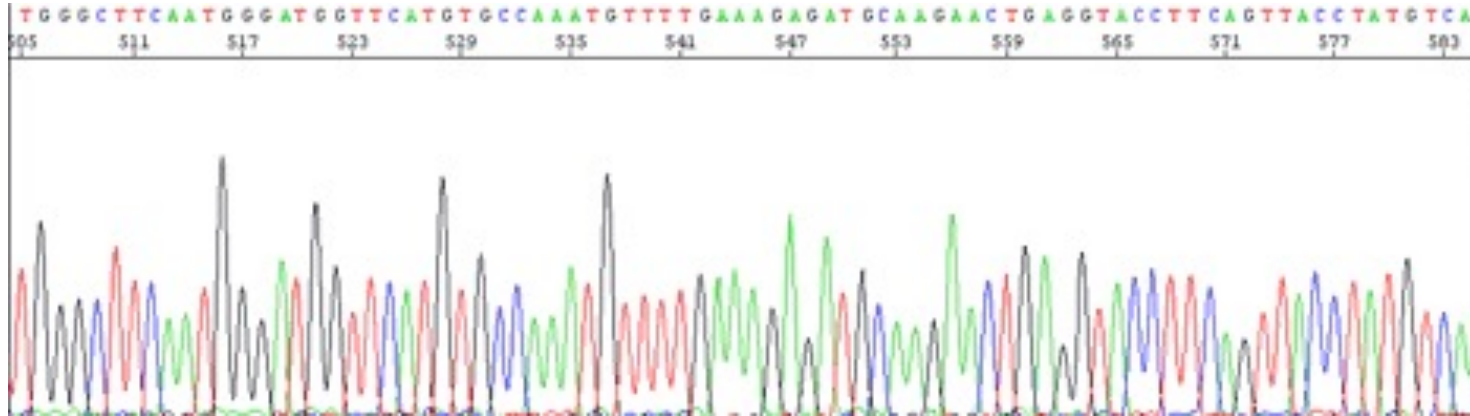
Sequencing costs have decreased massively over time



High Throughput Sequencing

- **Short-read sequencing technologies** (2nd generation):
 - Sequence millions of clonally amplified molecules at the same time
 - Reads are 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 or Nanopore

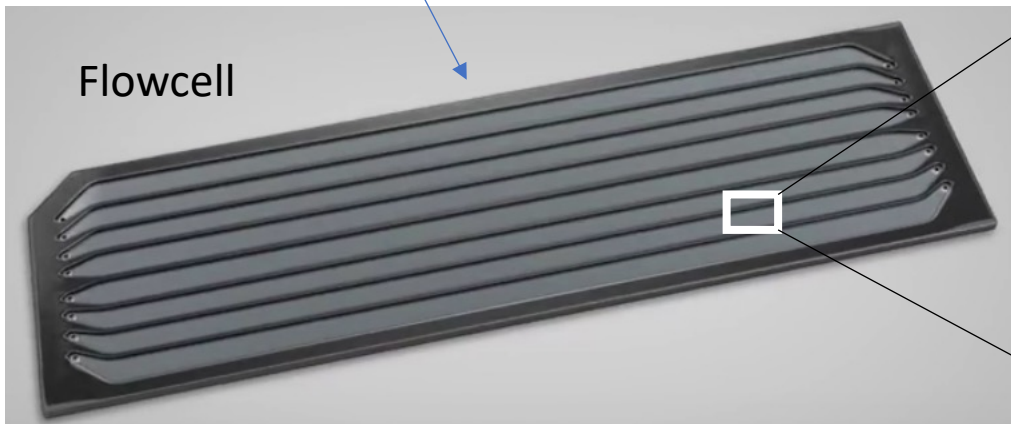
Sanger Sequencing



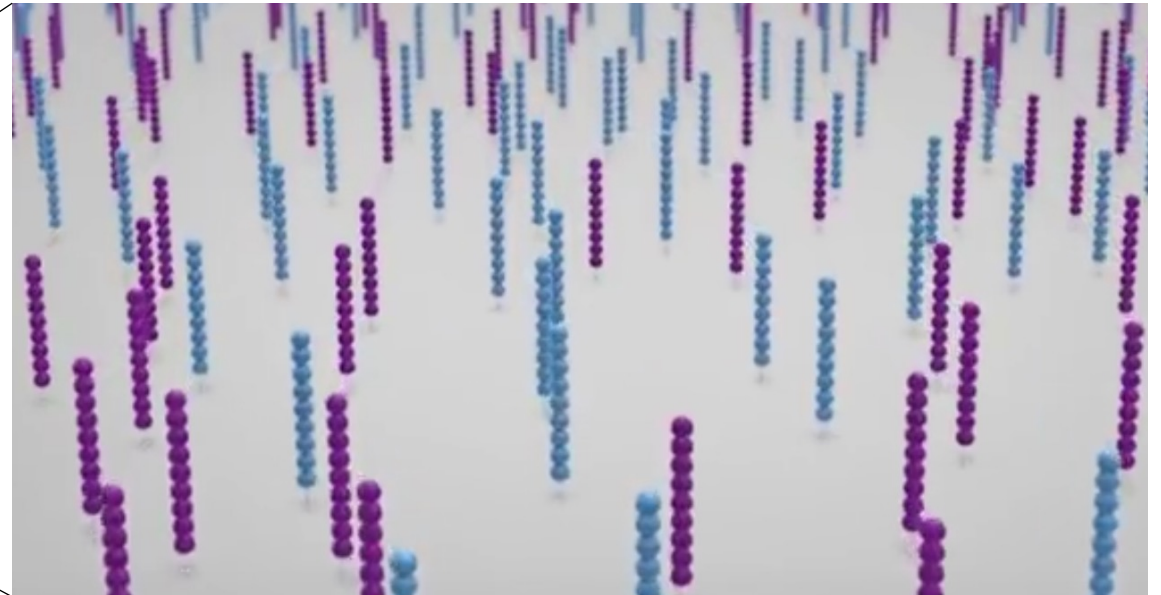
- Manually check each sequence
- Resequenced failed sequences

Illumina flowcell: millions of DNA sequences

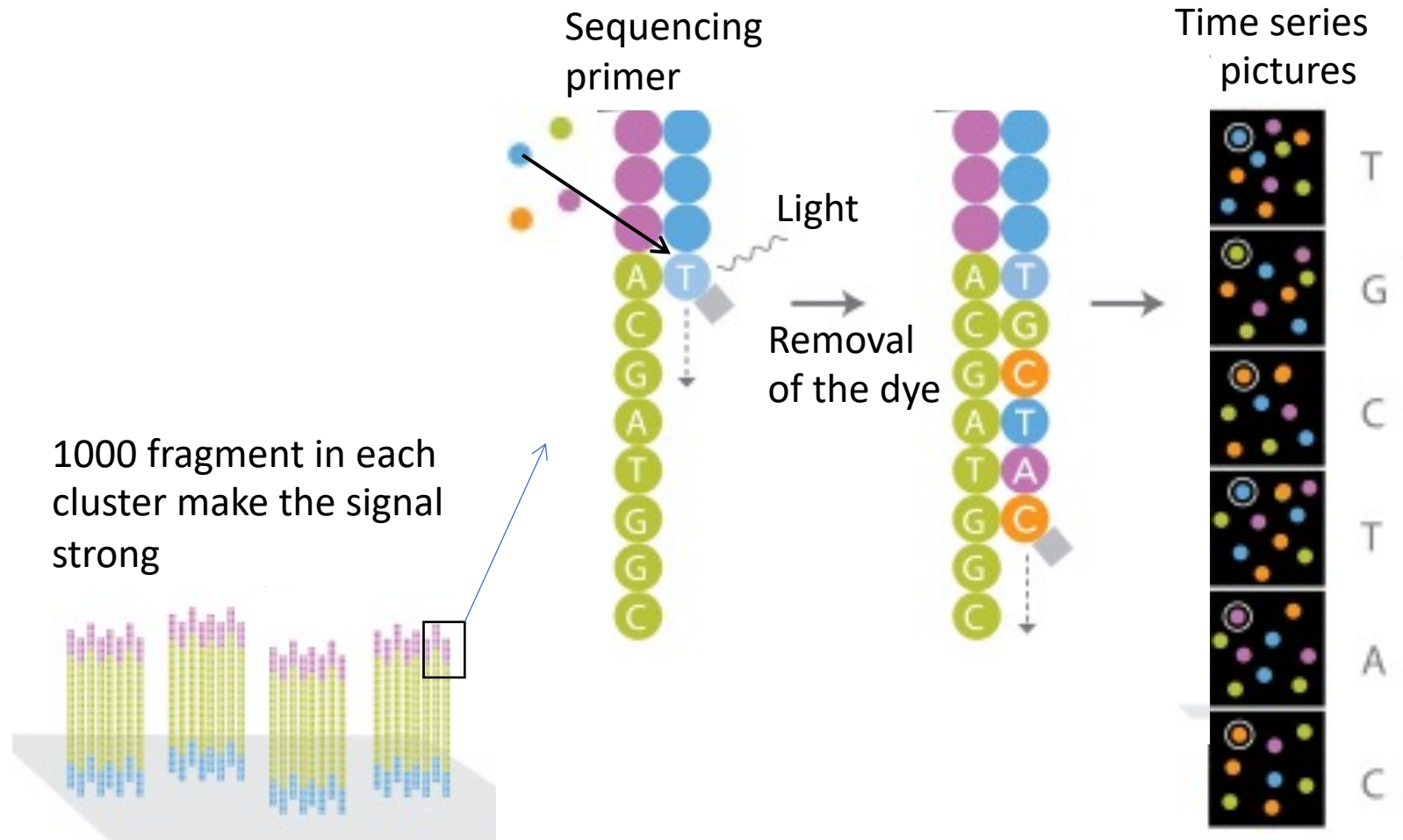
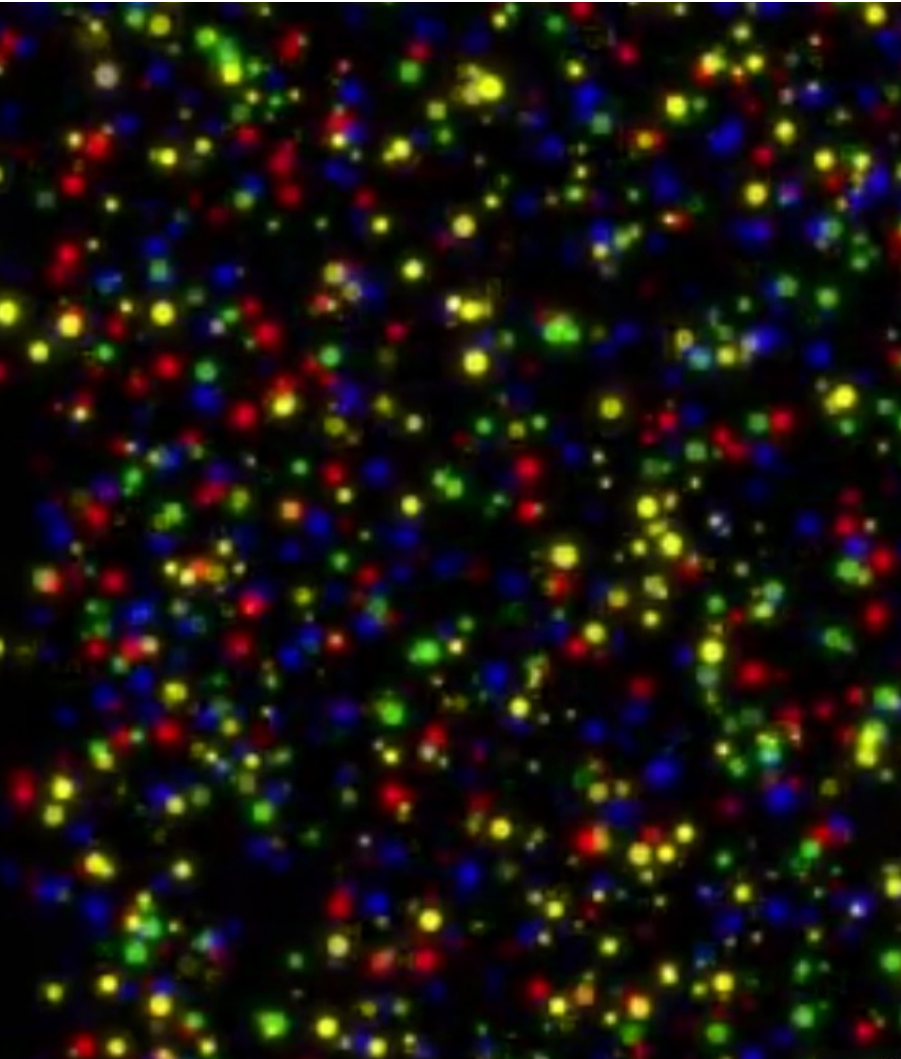
DNA fragments with Illumina adapters



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)

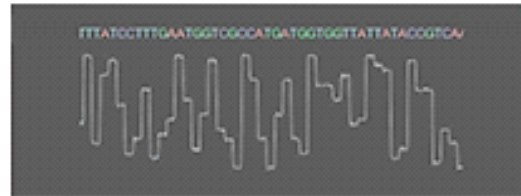
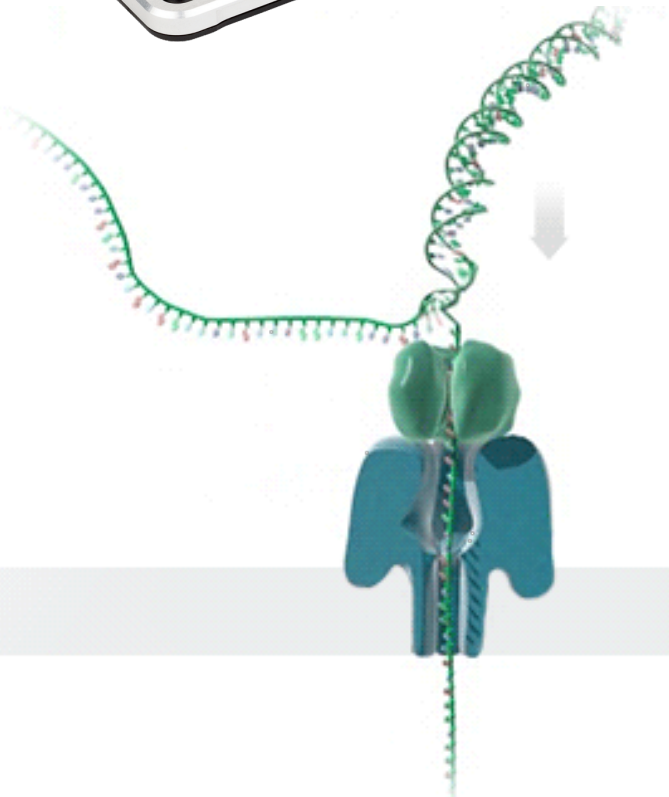
4-Channel Chemistry				
	 A	 G	 T	 C
Image 1				
Image 2				
Image 3				
Image 4				
Result	A	G	T	C

2-Channel Chemistry				
	 A	G	 T	 C
Image 1				
Image 2				
Result	A	G	T	C

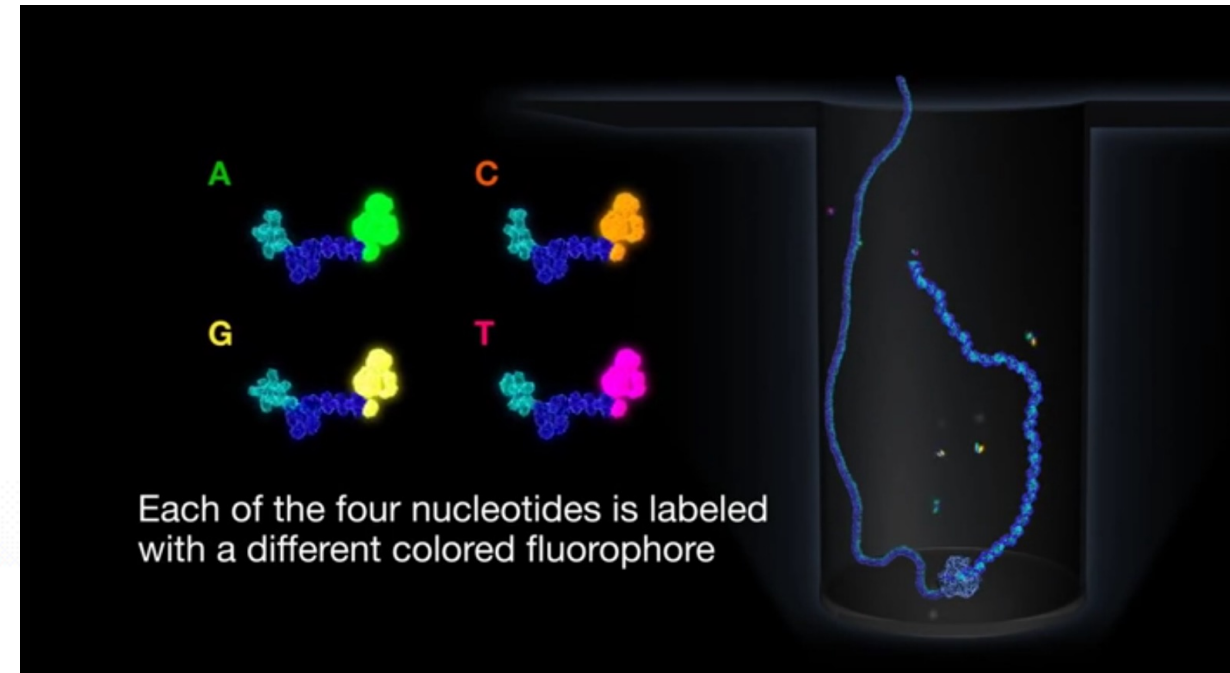
Long read sequencing technologies



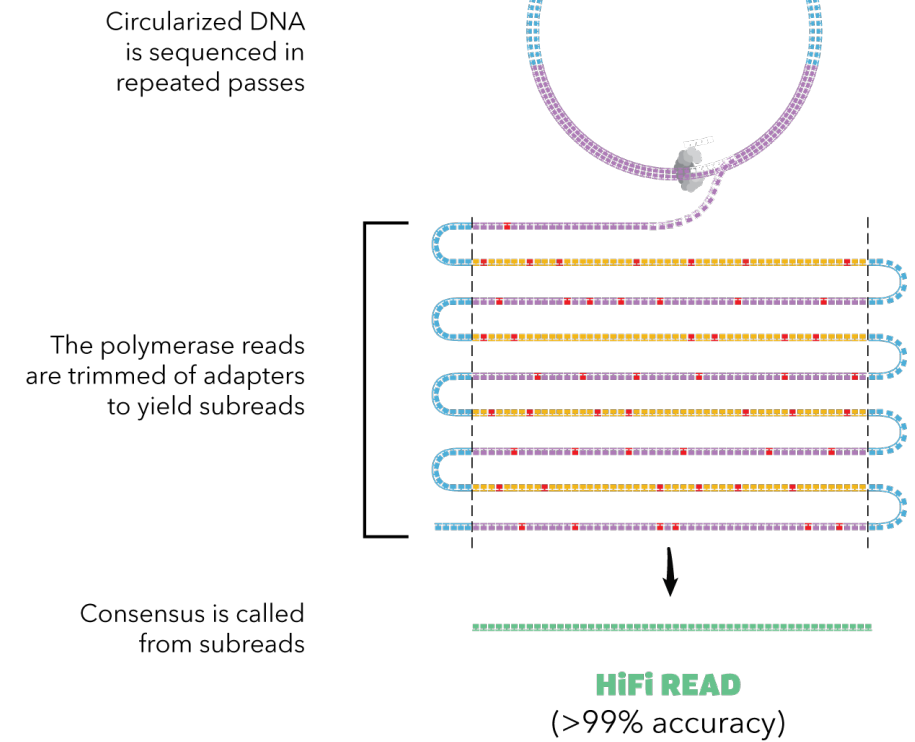
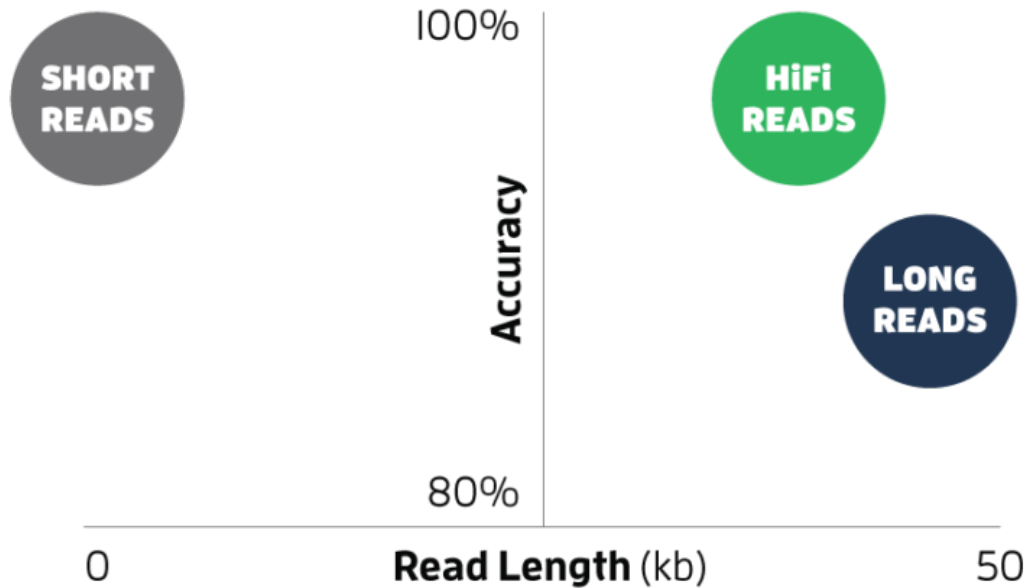
Nanopore



PacBio



PacBio HiFi reads



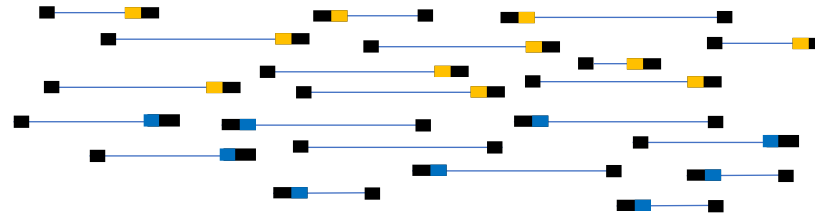
Whole-genome sequencing (shotgun sequencing)

DNA

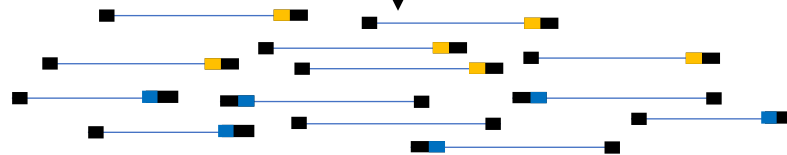
Random shearing



Illumina adapter ligation
incl. individual index



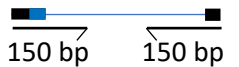
Size selection



paired-end sequencing

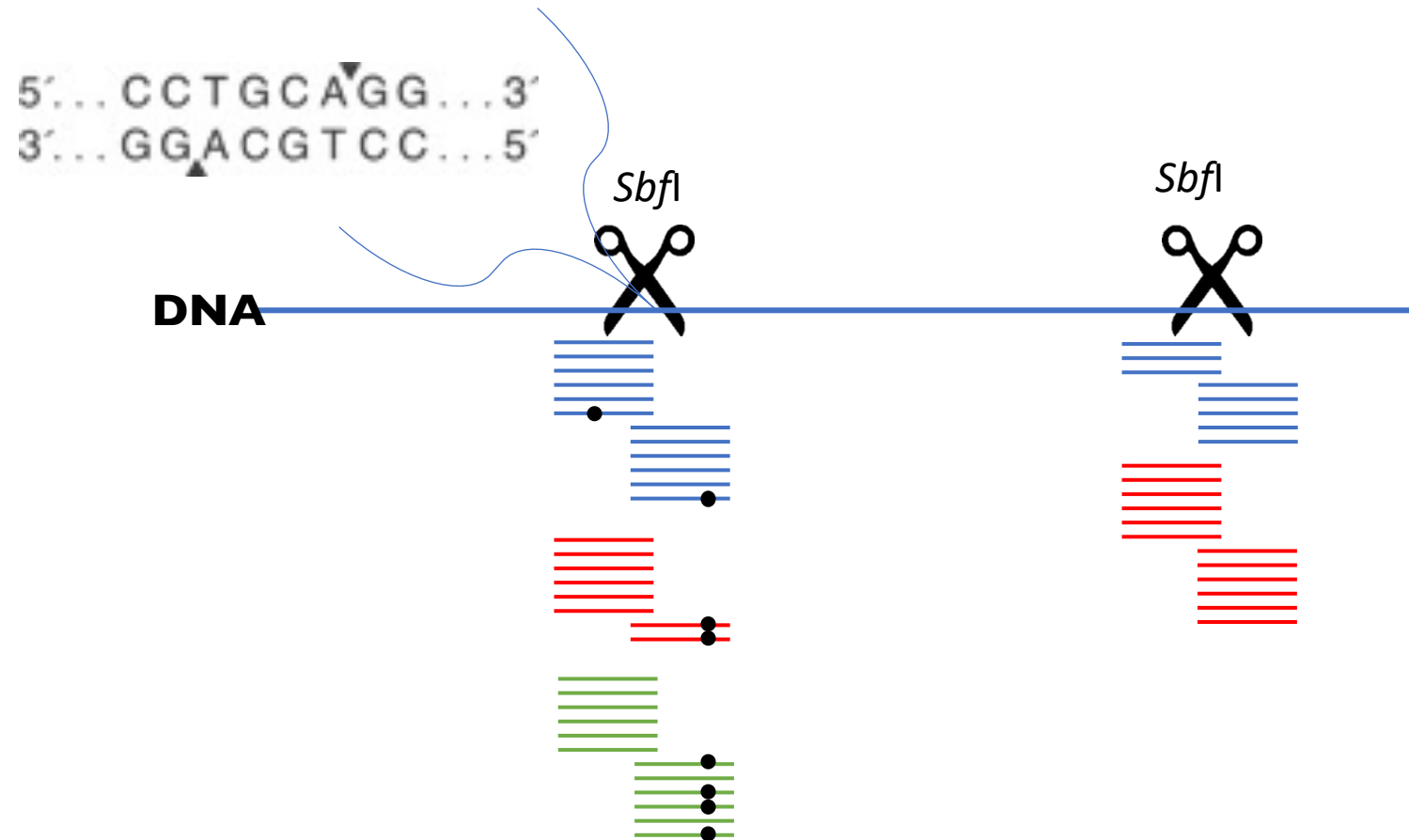


Up to 20 billion
read pairs



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 RAD sequencing at a sequencing depth of 25

Fastq format

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

Header (must start with @)

Base calls (sequence)

Quality scores

Quality scores

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

ASCII encoding

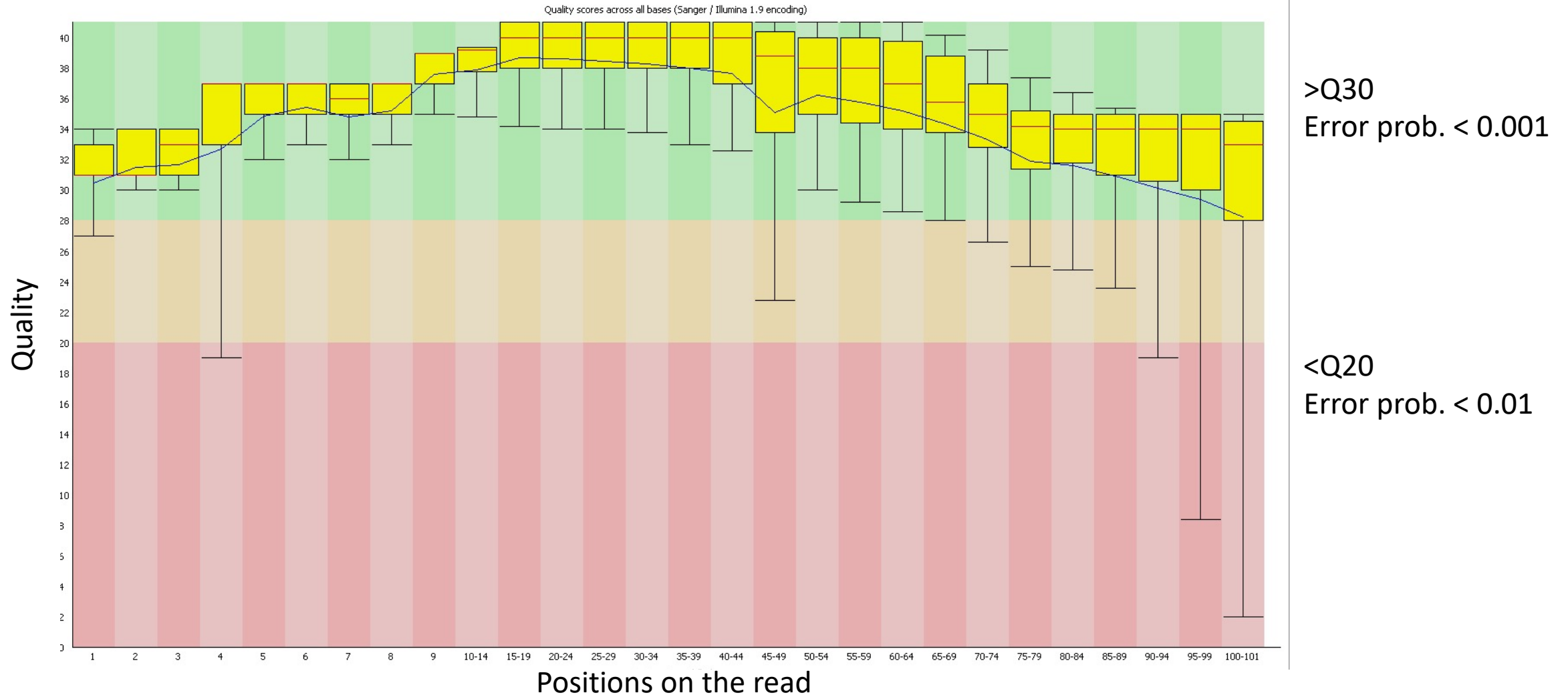
40:@	90:Z	141:a
41:A	91:[142:b
42:B	92:\	143:c
43:C	93:]	144:d
44:D	94:^	145:e
45:E	95:_	146:f
... :...	... :...	... :...

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

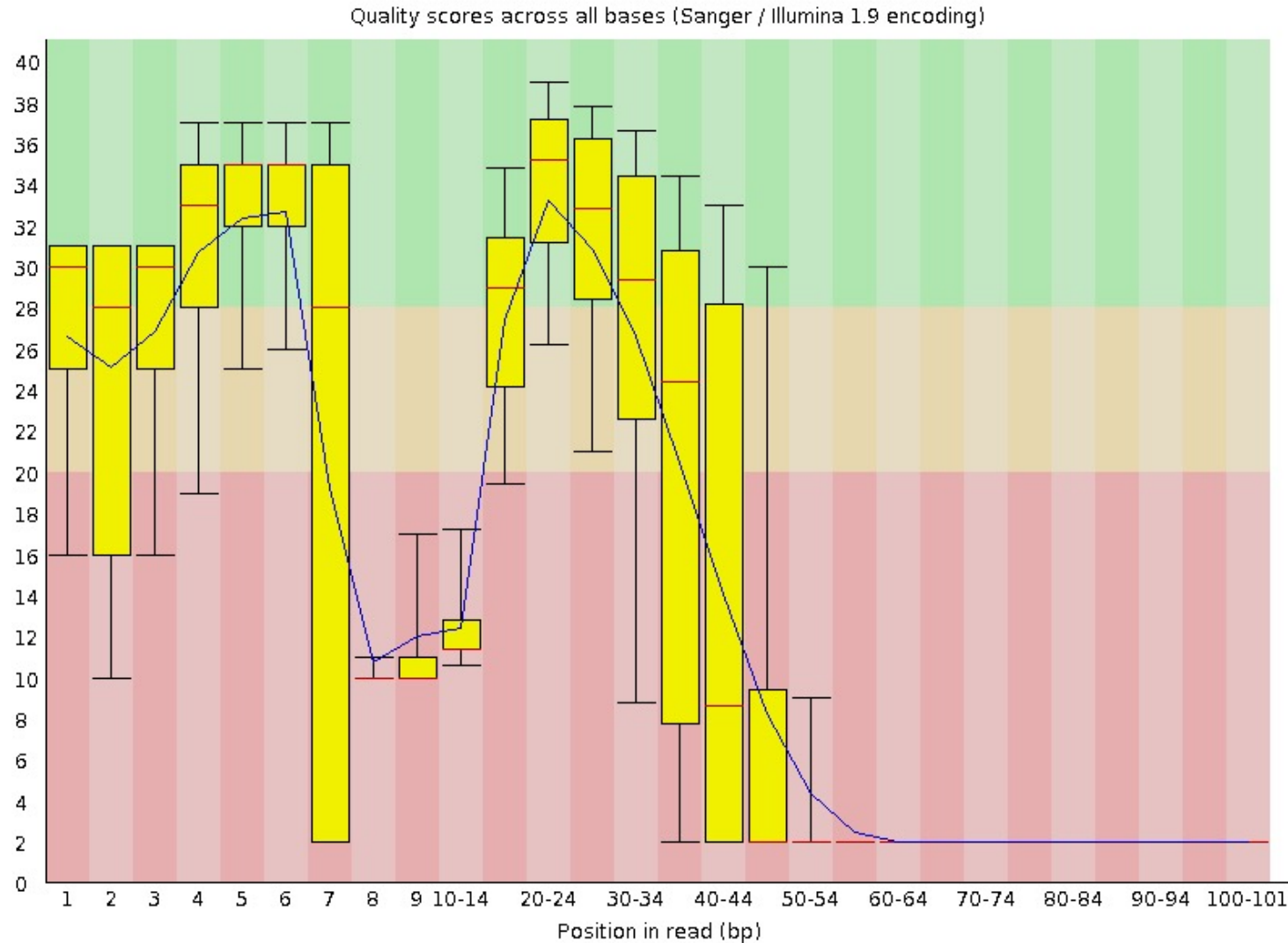
p = Probability call is incorrect

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)



FastQC: Quality across bases (bad example)



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