Illumina data QC & basic NGS tools

From the very beginning

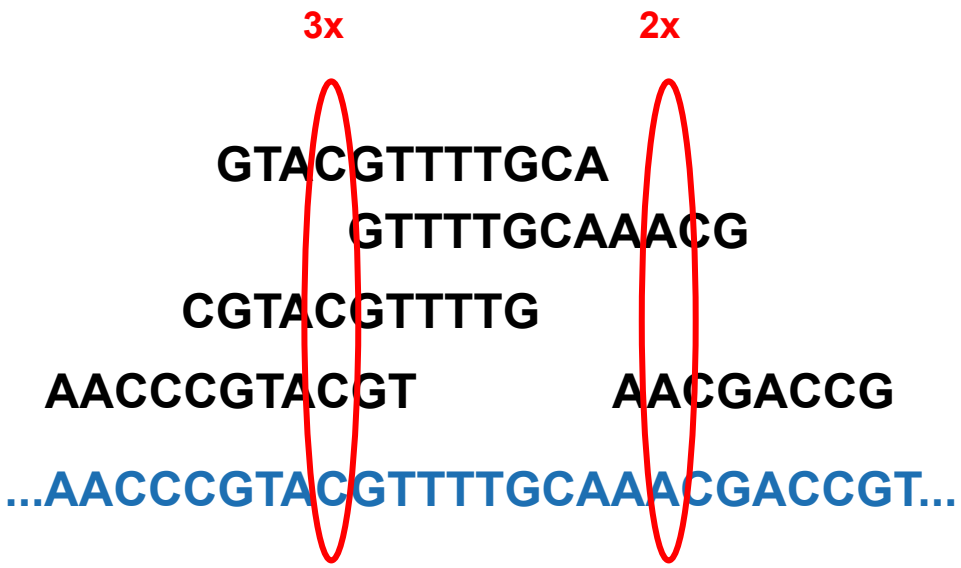
...AACCCGTACGTTTTGCAAACGACCGT...

From the very beginning

- Sequencing

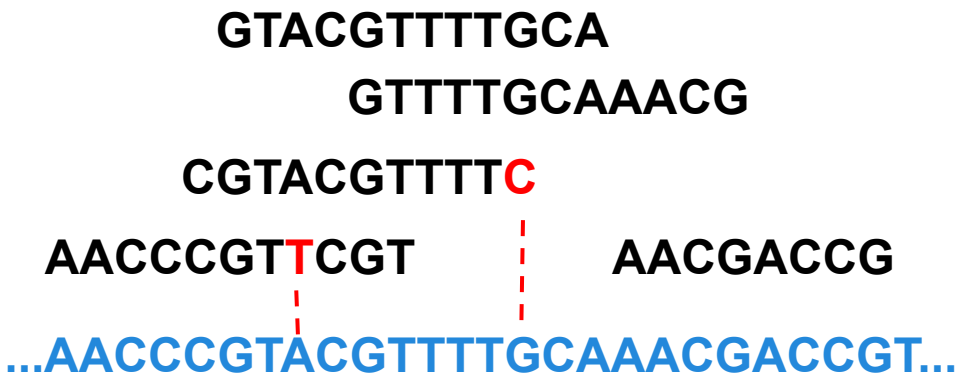
From the very beginning

- Sequencing
- Coverage



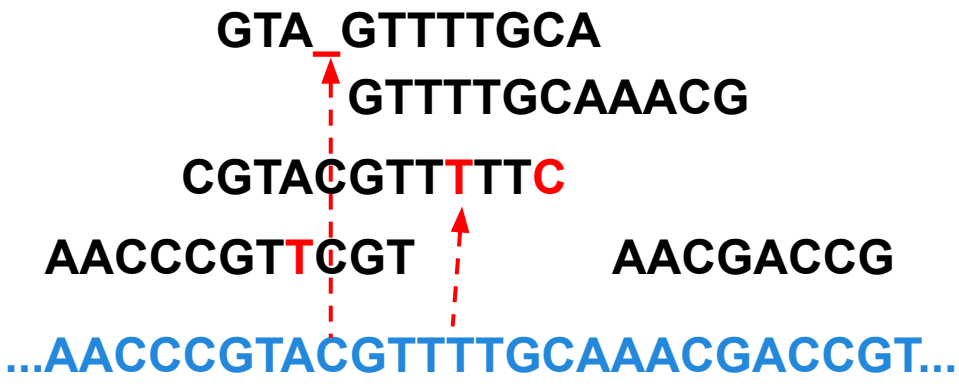
From the very beginning

- Sequencing
- Coverage
- Errors
  - Mismatches



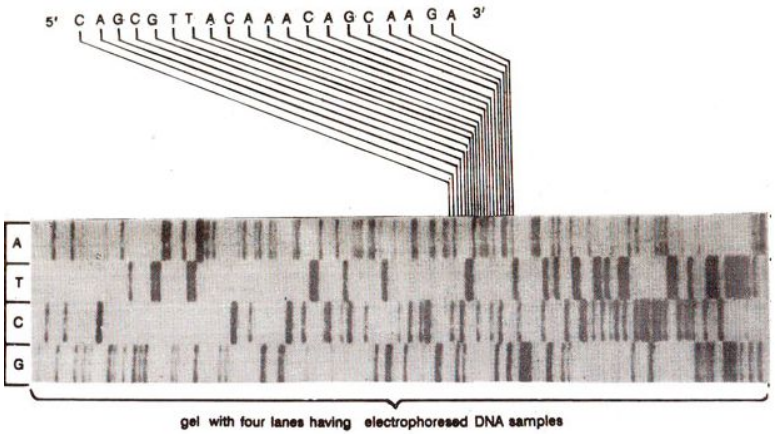
From the very beginning

- Sequencing
- Coverage
- Errors
  - Mismatches
  - Indels

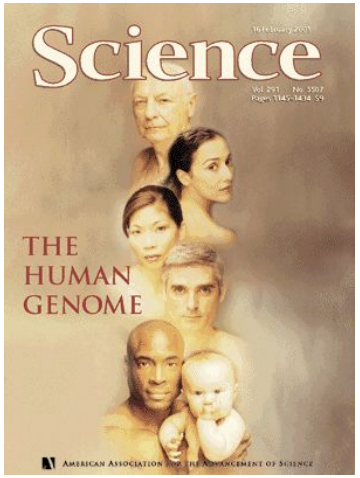


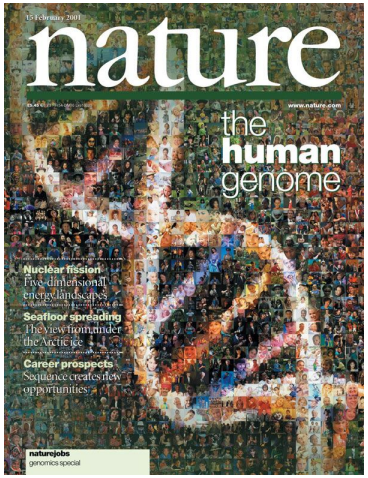
Early days

- Sanger sequencing
  - Long reads (~900 bp)
  - Low coverage (< 10x)
  - Extreme cost



- Human genome project
  - 3 Gbp
  - 3 billion USD
  - 10 years





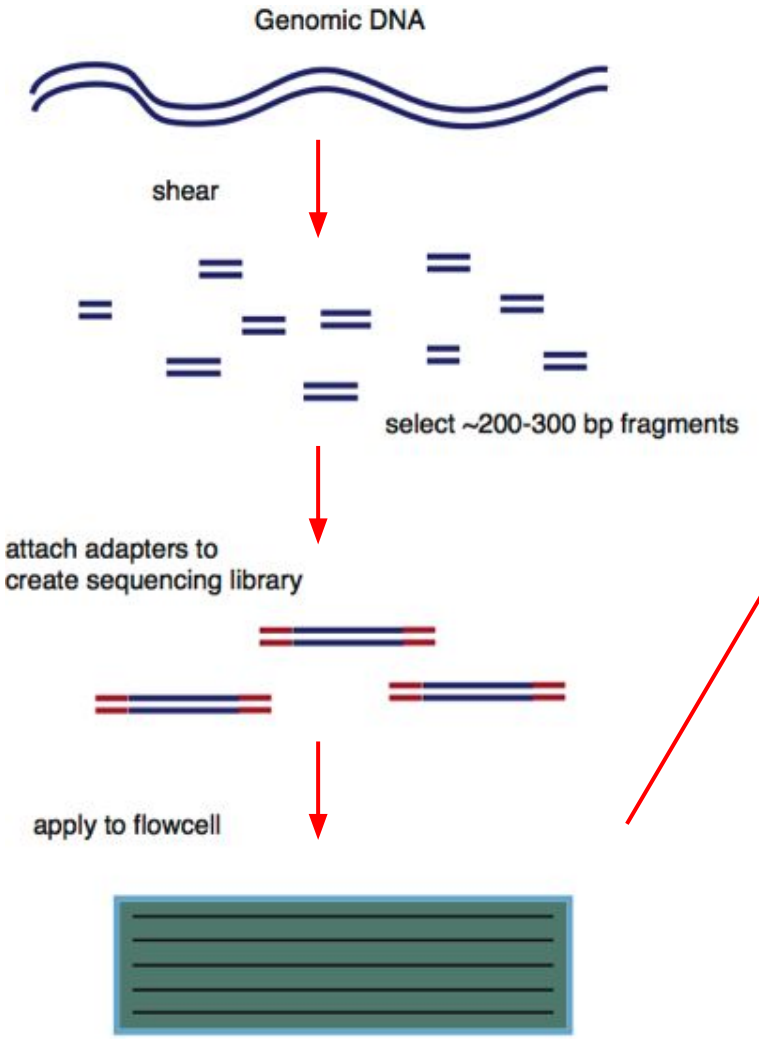
NGS

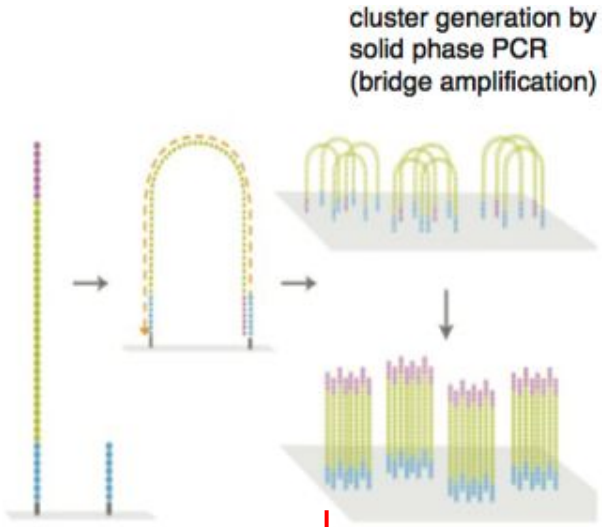
- Shorter reads (25-400bp)
- High coverage (50-1000x)
- Huge amount of data
- Low cost
- More applications
- **Required completely new algorithms**

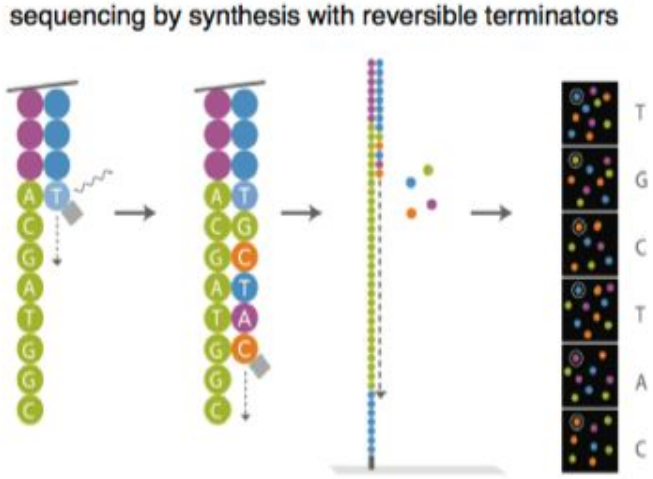
NGS technologies

	illumina	Roche 454 SEQUENCING	ion torrent by life technologies	PACIFIC BIOSCIENCES	Oxford NANOPORE Technologies
Read length, bp	25-300	400-1100	200-400	1000-70000	5000-900000
Error rate	0.1-1%	1%	1-2%	10-20%	10-30%
Error type	Mismatches only	Indels & Mismatches	Indels & Mismatches	Indels & Mismatches	Indels & Mismatches
Comments	Error rate grows at the end of read	Problems with homopolymers	Problems with homopolymers	Errors distributed randomly	Typically several deletions in a row
$ per 1 Mbp	0.05 - 0.5	30	0.5 - 20	2+	0.01
Sequencer cost	100-500 K	100 K	80K	700 K	1 K

Illumina sequencing

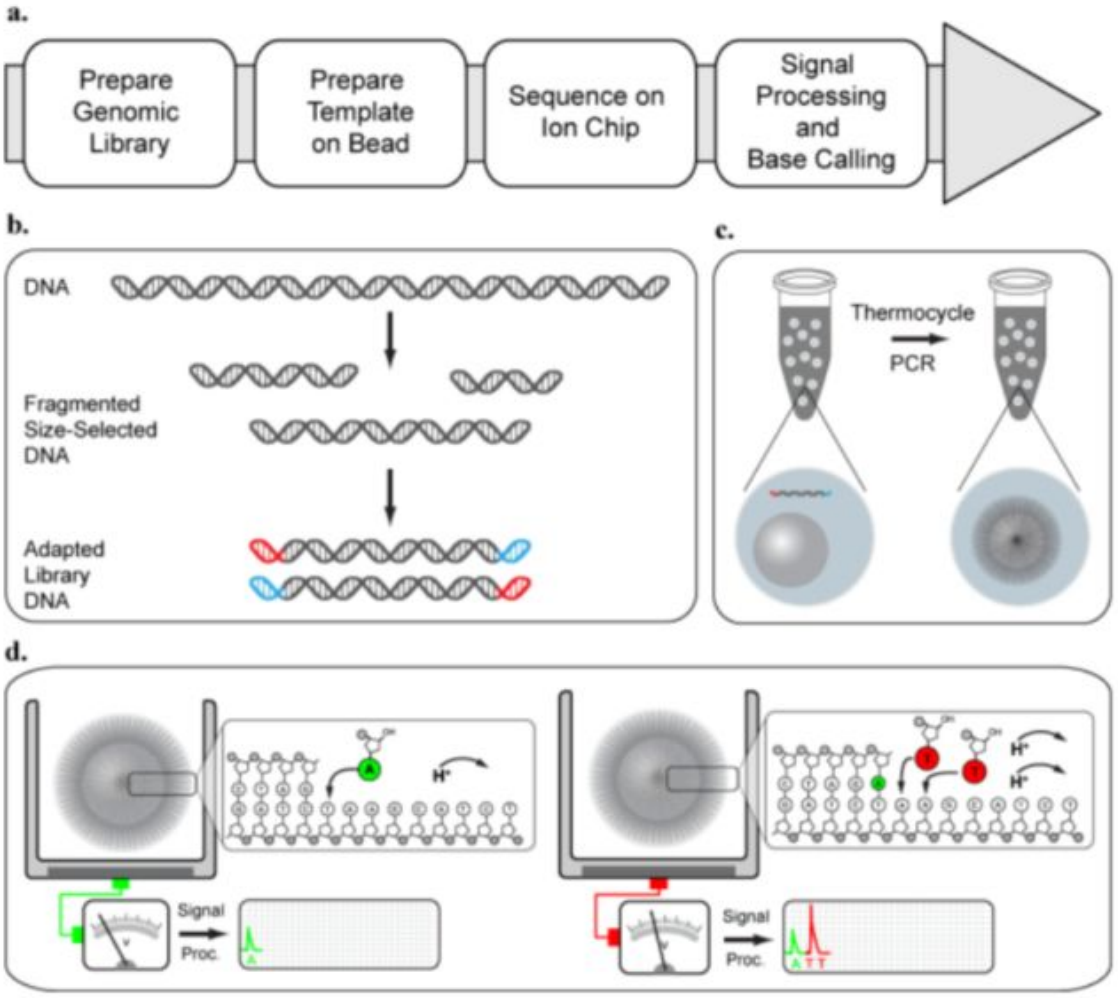






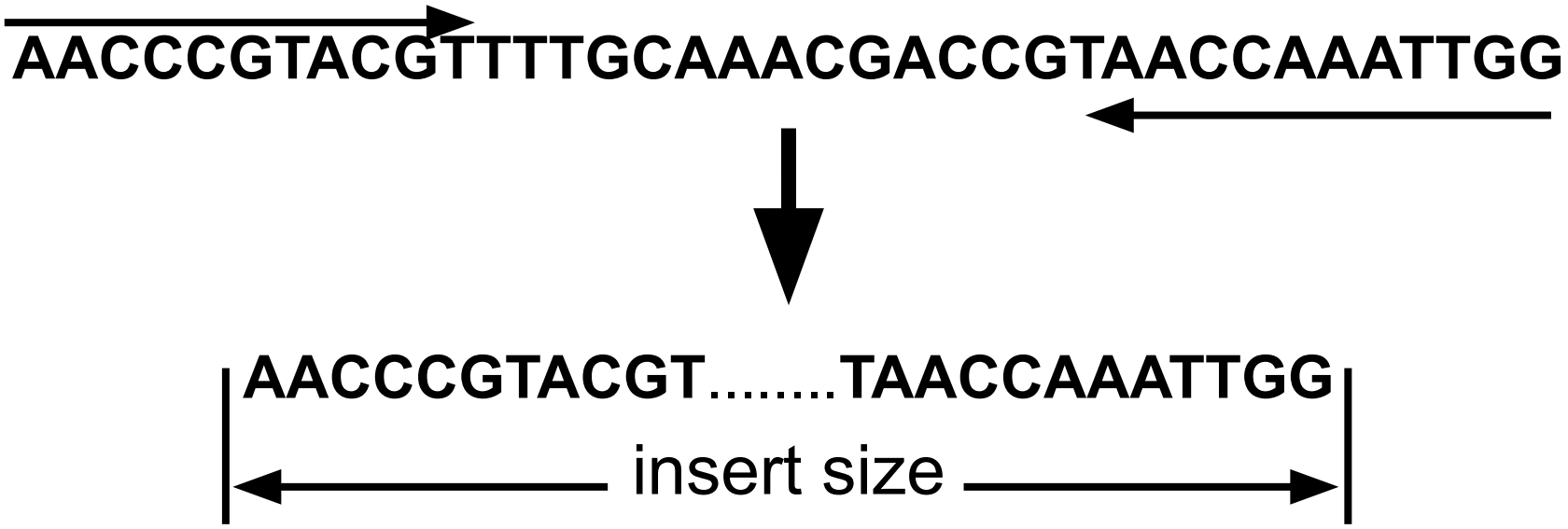
http://www.youtube.com/watch?v=77r5p8IBwJk

IonTorrent sequencing



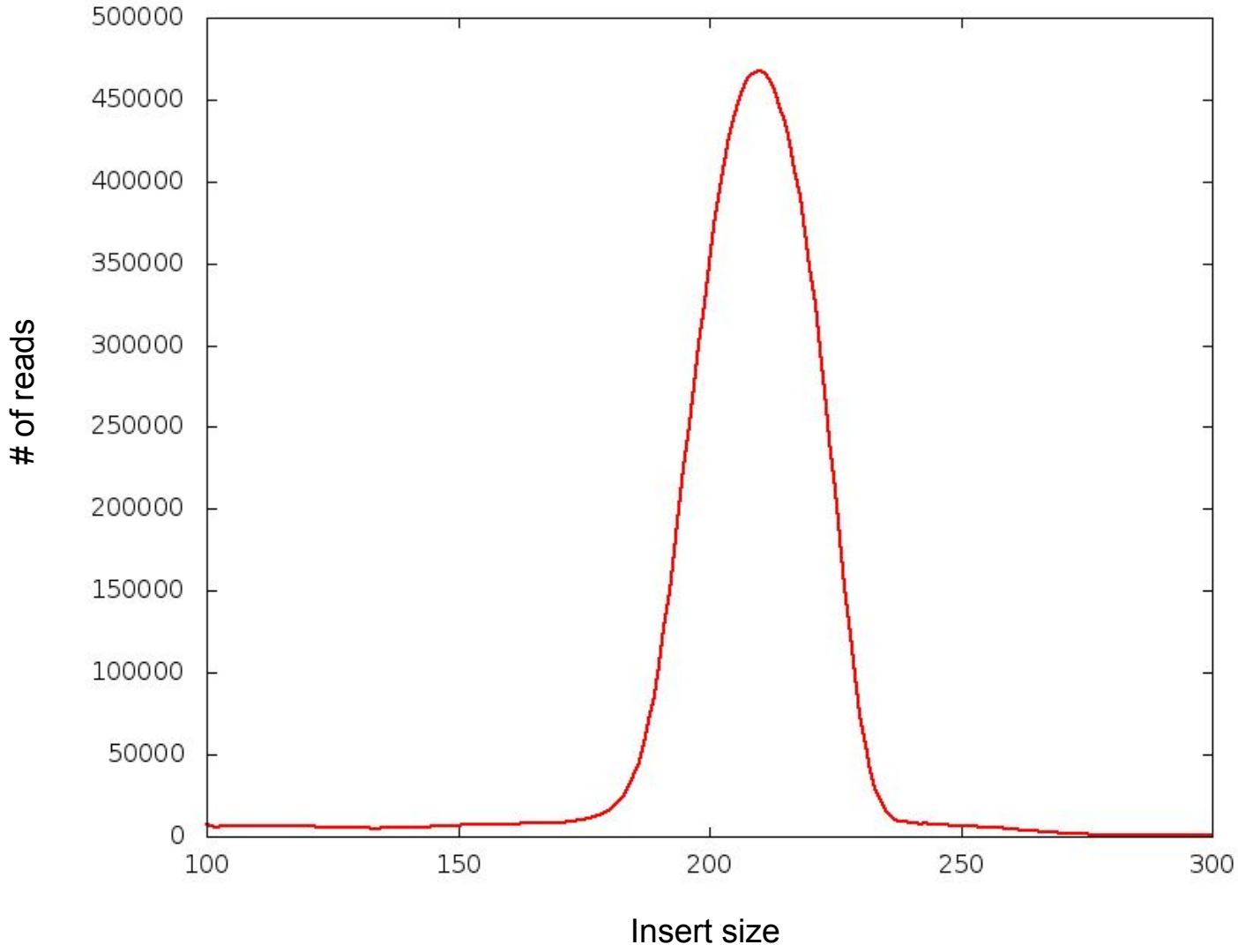
https://www.youtube.com/watch?v=WYBzbxIfuKs

Paired reads



- Paired-end (< 1 kbp)
- Mate-pairs (1 - 20 kbp)

Insert size distribution



FASTA/FASTQ

- FASTA

```
>EAS20_8_6_1_9_1972/1
ACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGC
>EAS20_8_6_1_163_1521/1
GCAGAAAACGTTCTGCATTTGCCACTGATGTACCGCCGAACTTCAACACTCGCA
```

- FASTQ

```
@EAS20_8_6_1_1477_92/1
ACCGTTACCTGTGGTAATGGTGATGGTGGTGGTAATGGTGGTGCTAATGCGTTT
+EAS20_8_6_1_1477_92/1
HHGHFHHHHHHHHHGFFHHHBG?GGC8DD9GF??=FFBCGBAF>FGCFHGHGGG
```

- Phred quality

$$Q = [- 10 \log_{10} p / (1 - p)]$$

seqtk utility

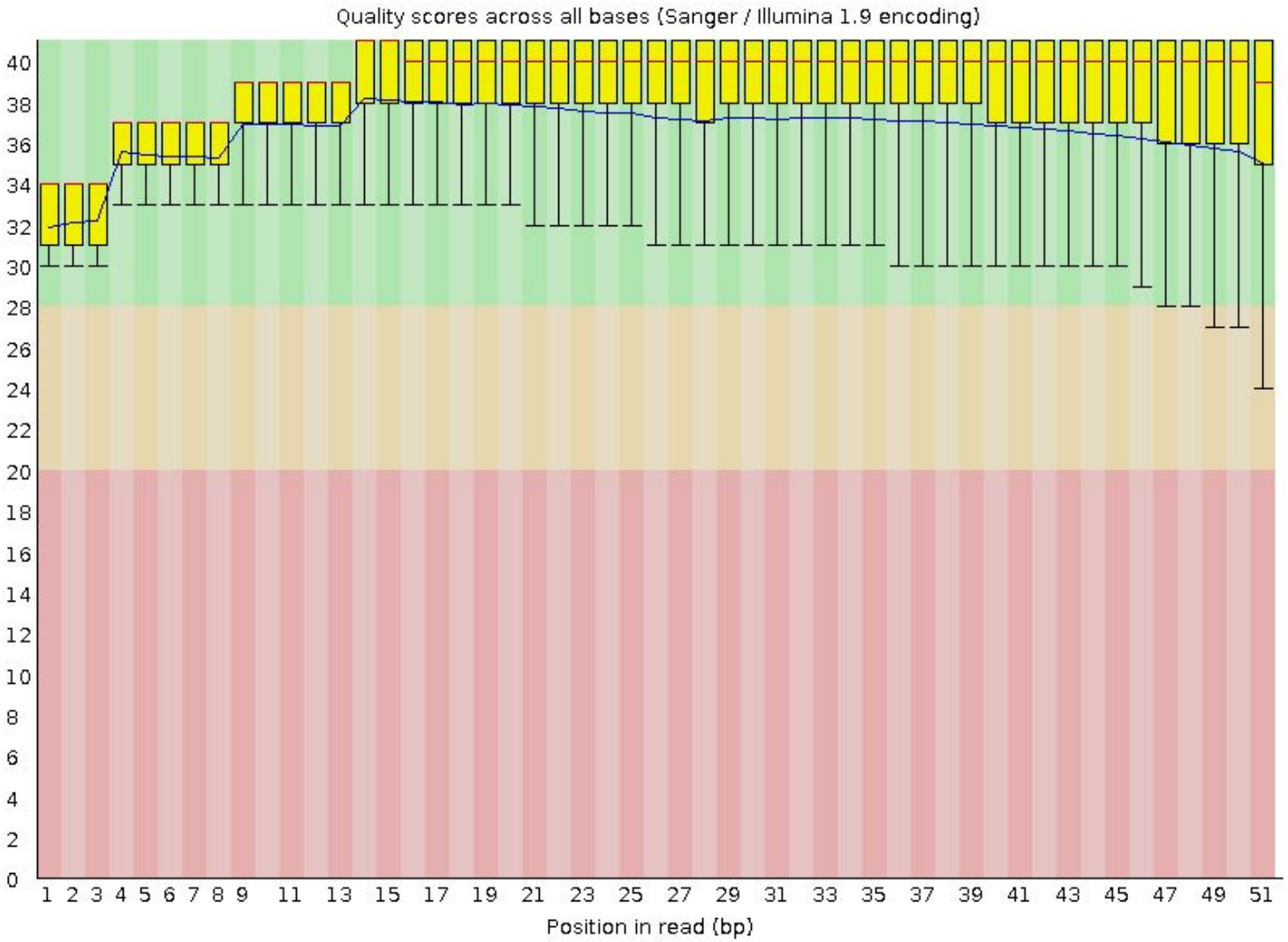
- Subsampling
  sample
- Converting between interleaved/paired files
  mergepe, seq -1/-2
- fastq->fasta
  seq -A
- Quality trimming
- Shifting the quality
- Modifying names
- etc...

Quality Control

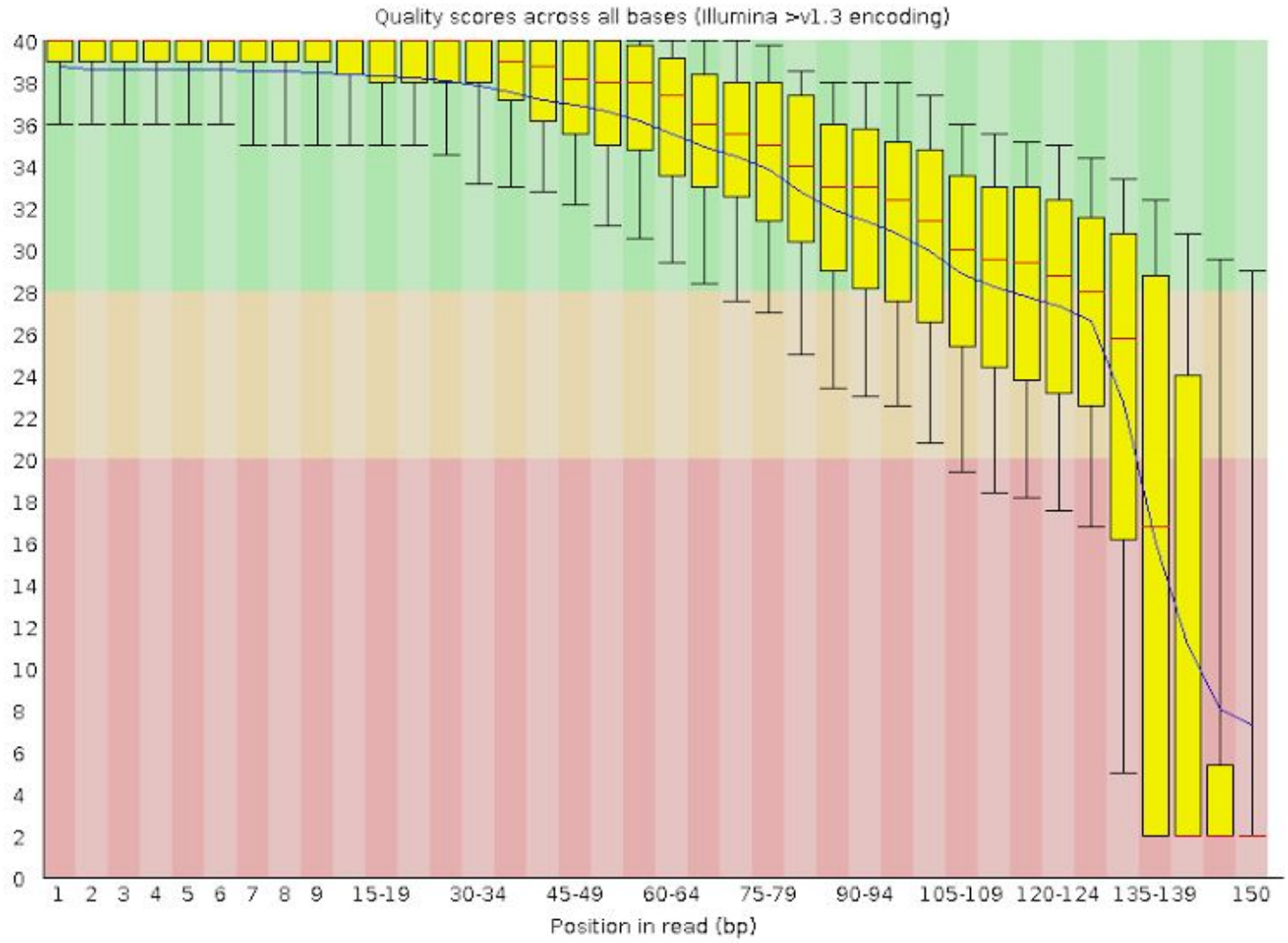
FastQC

- Easy and lightweight quality control for sequencing data
- Does not require reference genome

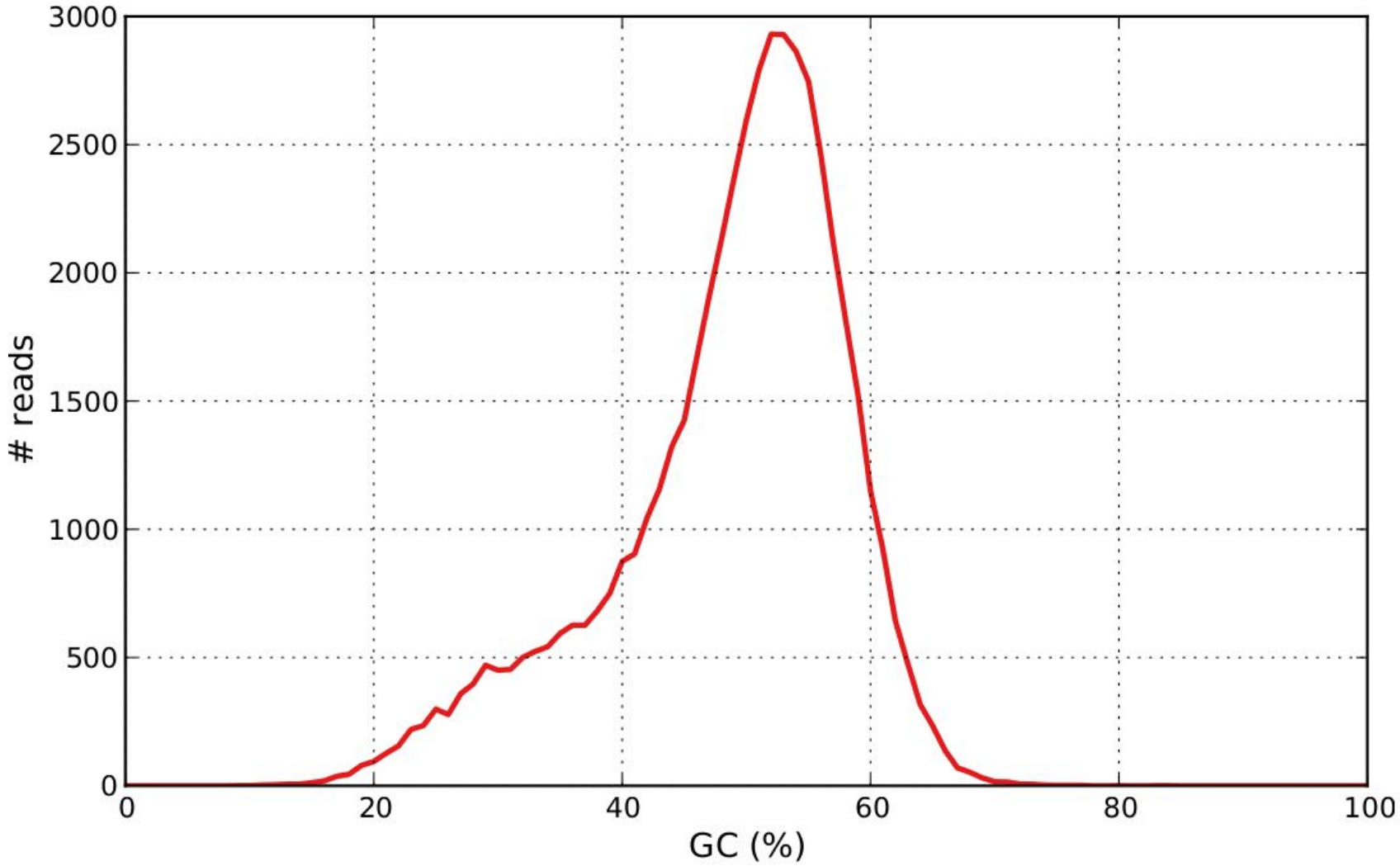
Per base sequence quality



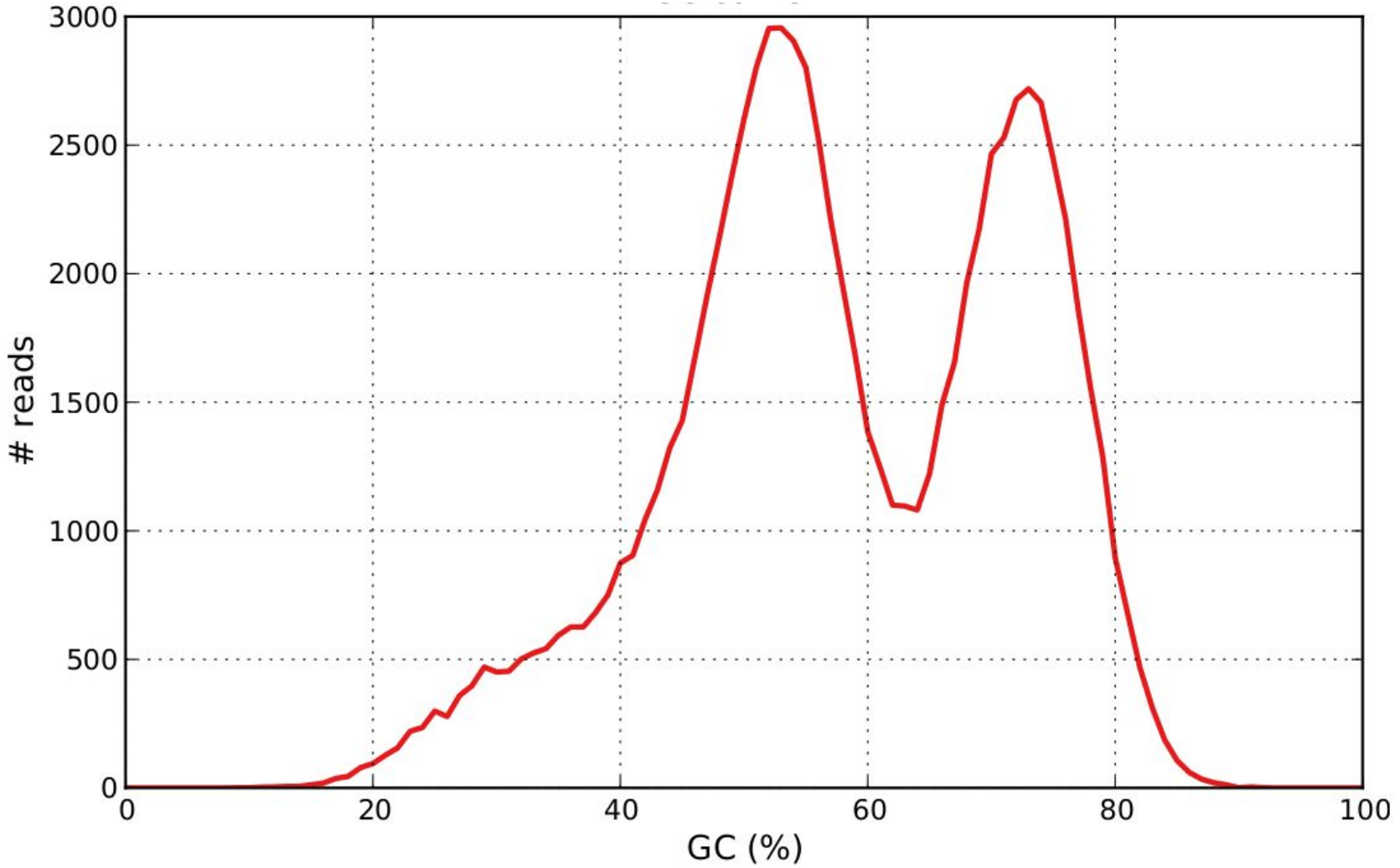
Per base sequence quality



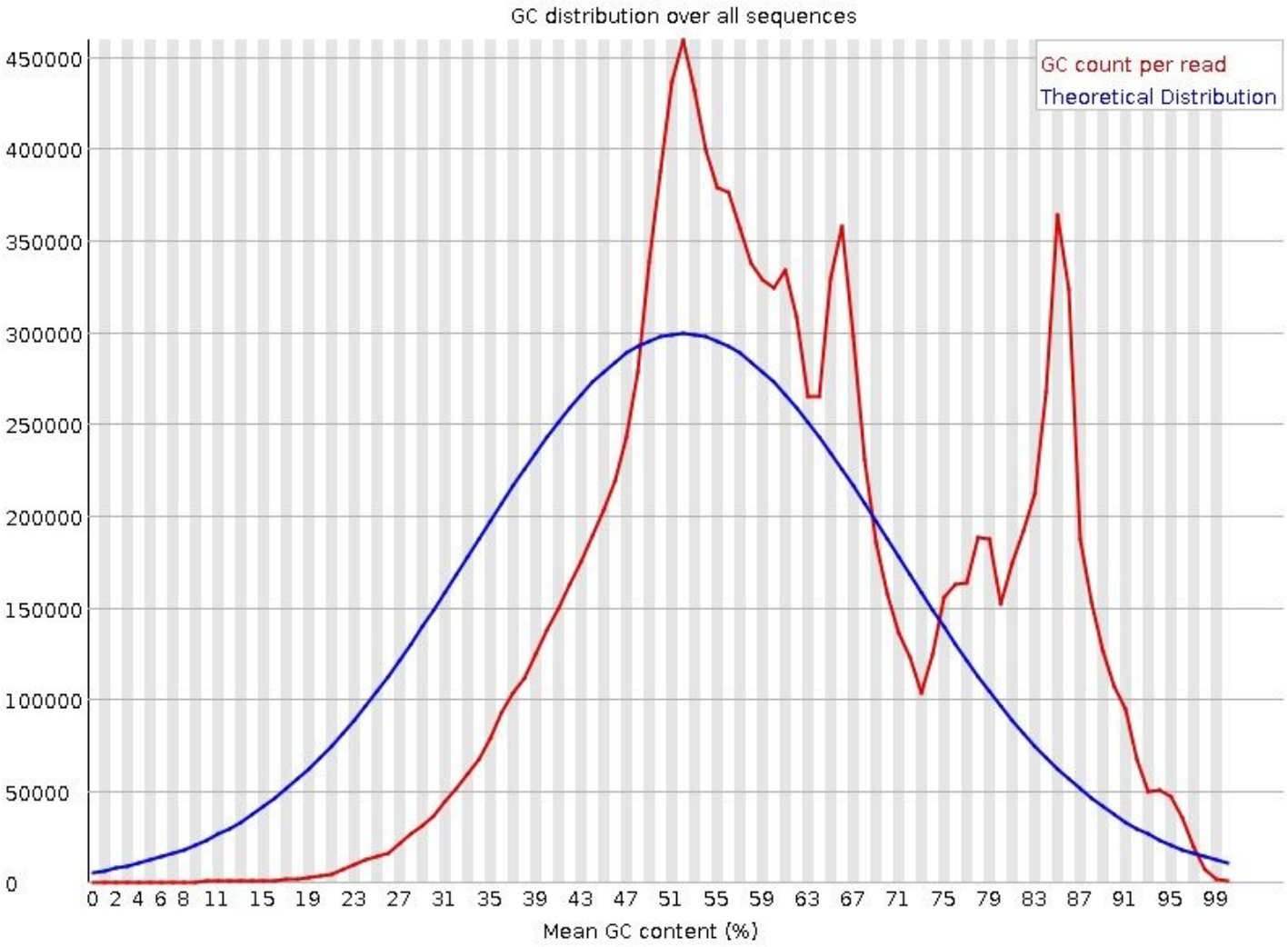
Per sequence GC content



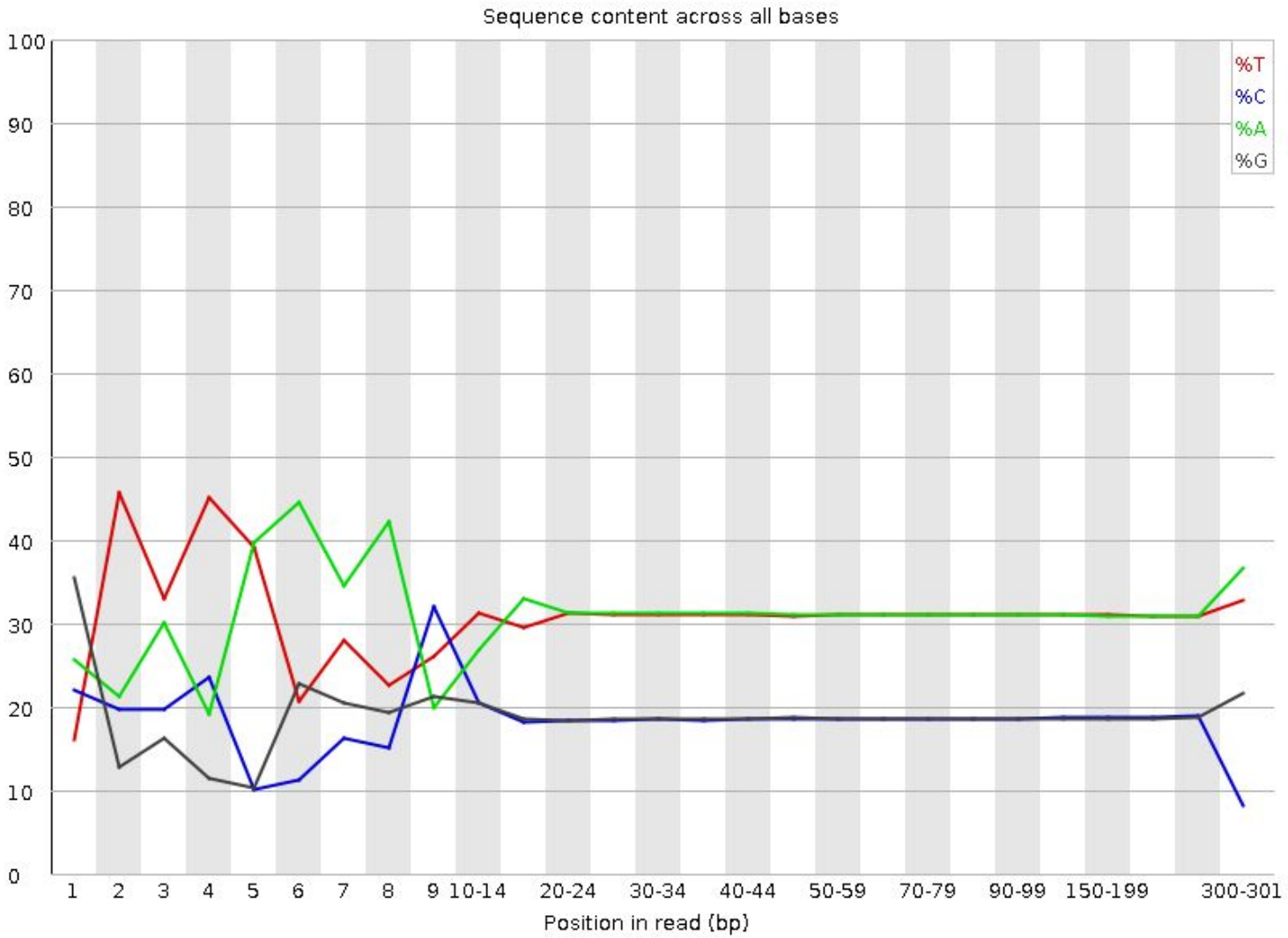
Per sequence GC content



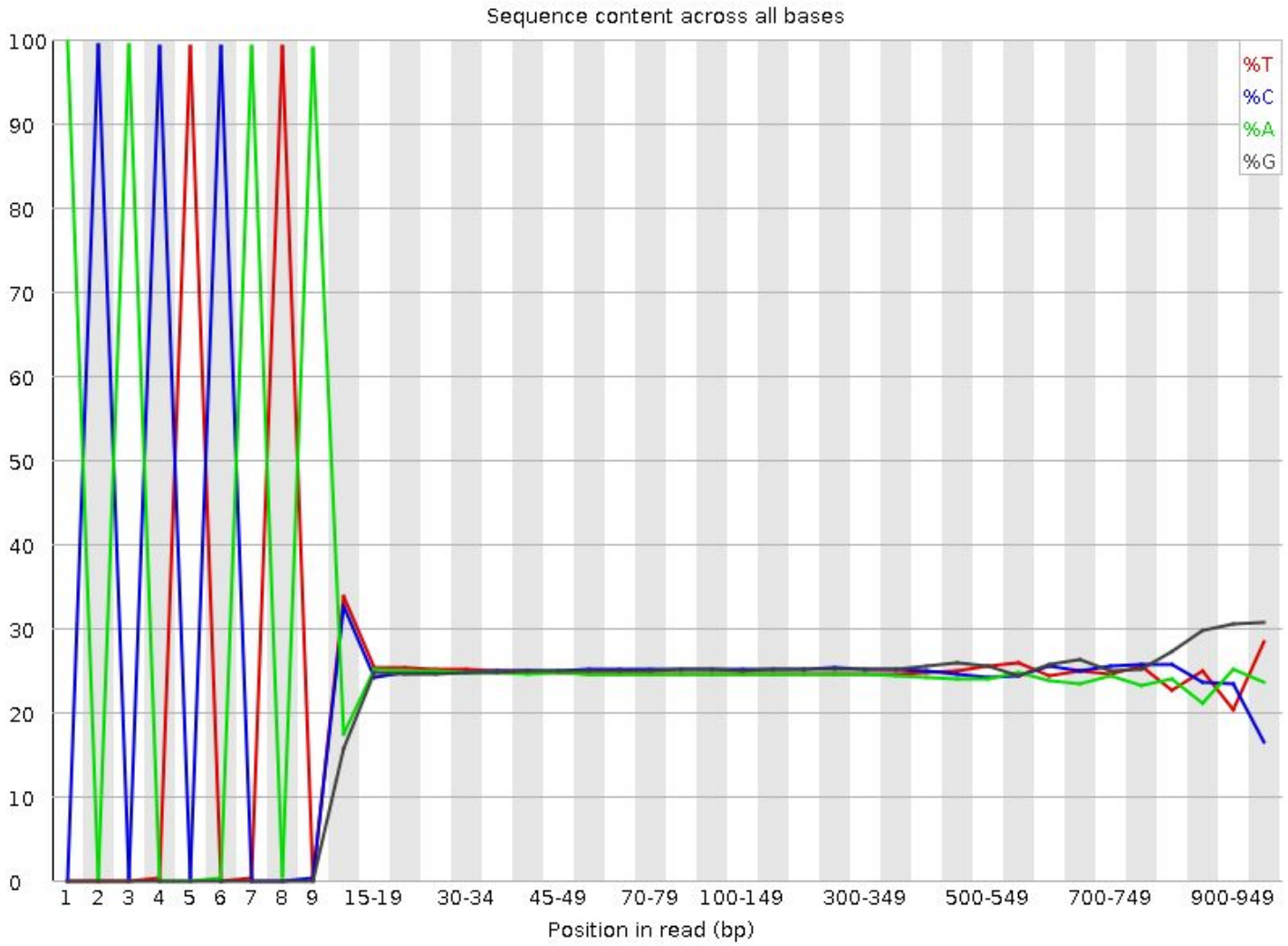
Per sequence GC content



Per base sequence content



Per base sequence content



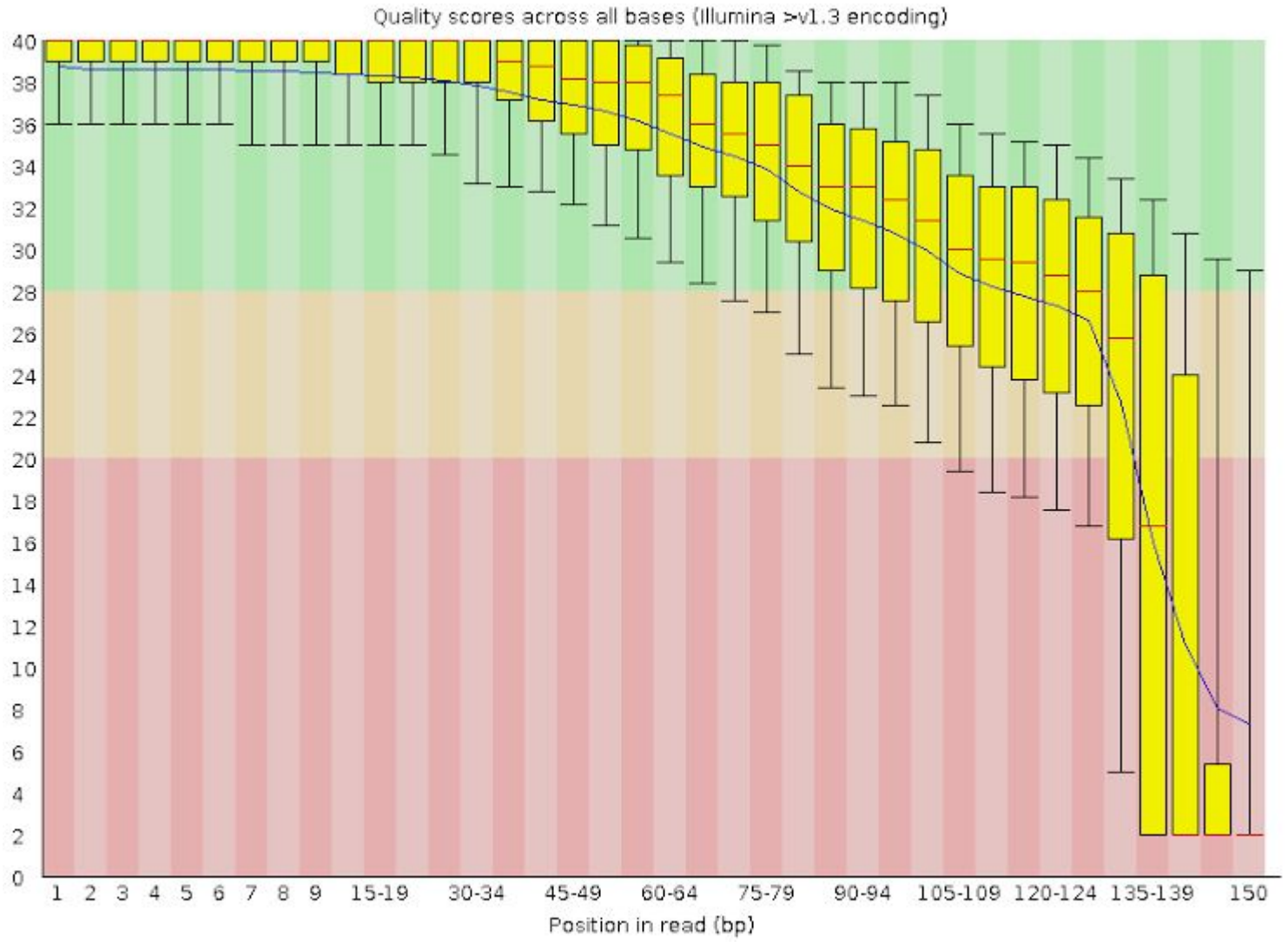
FastQC

- **fastqc** -h
- **mkdir** <output>
- **fastqc** <file1.fastq> <file2.fastq> …

  -o <output>

Error correction

Per base sequence quality



Trimmomatic

- **SE** <input reads> <output reads> LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

- Remove leading low quality or N bases (below quality 3) (LEADING:3)
- Remove trailing low quality or N bases (below quality 3) (TRAILING:3)

Trimmomatic

- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

Trimmomatic

- **PE** <left reads> <right reads> <left paired> <left unpaired> <right paired> <right unpaired> OPTIONS
- ILLUMINACLIP:<path to adapters>
  - ILLUMINACLIP:TruSeq3-PE.fa

Adapter trimming

ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>

ILLUMINACLIP:NexteraPE-PE.fa:2:10:30

Thank you!