Today’s Agenda

- ✓ Look at two *M. tuberculosis* datasets
- ✓ Run fastqc to look at fastq data quality
- ✓ Trim poor quality reads with Trim Galore
- ✓ Mapping module

ILLUMINA DATA QC

Illumina

fastq

Nanopore

fast5

Ion Torrent

bam

Fastq format

```
1   @SEQ_ID
2   GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGT
3   TT +
4   !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

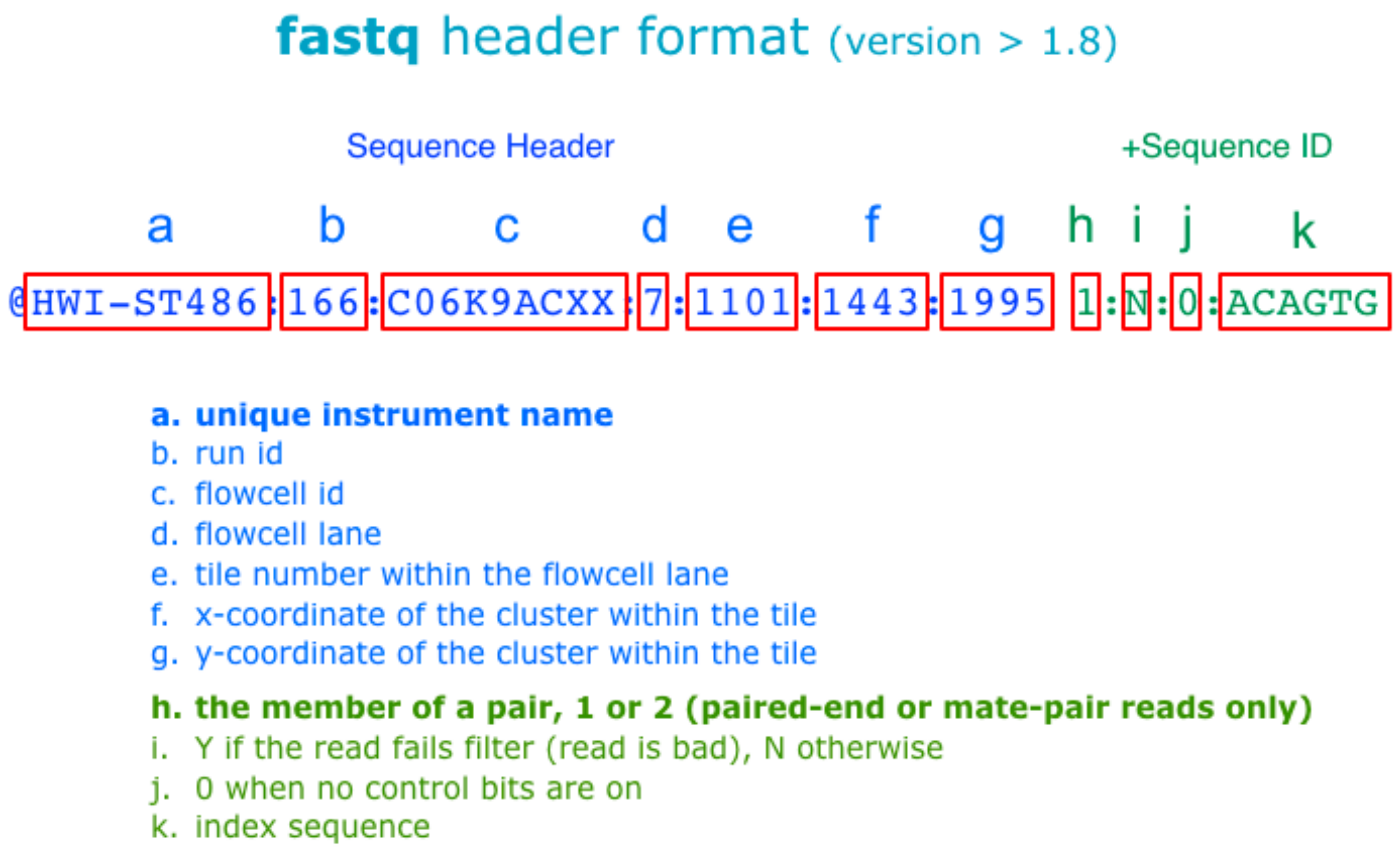
Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Fastq format



fastq header format (version > 1.8)

Sequence Header

+Sequence ID

```
@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG
```

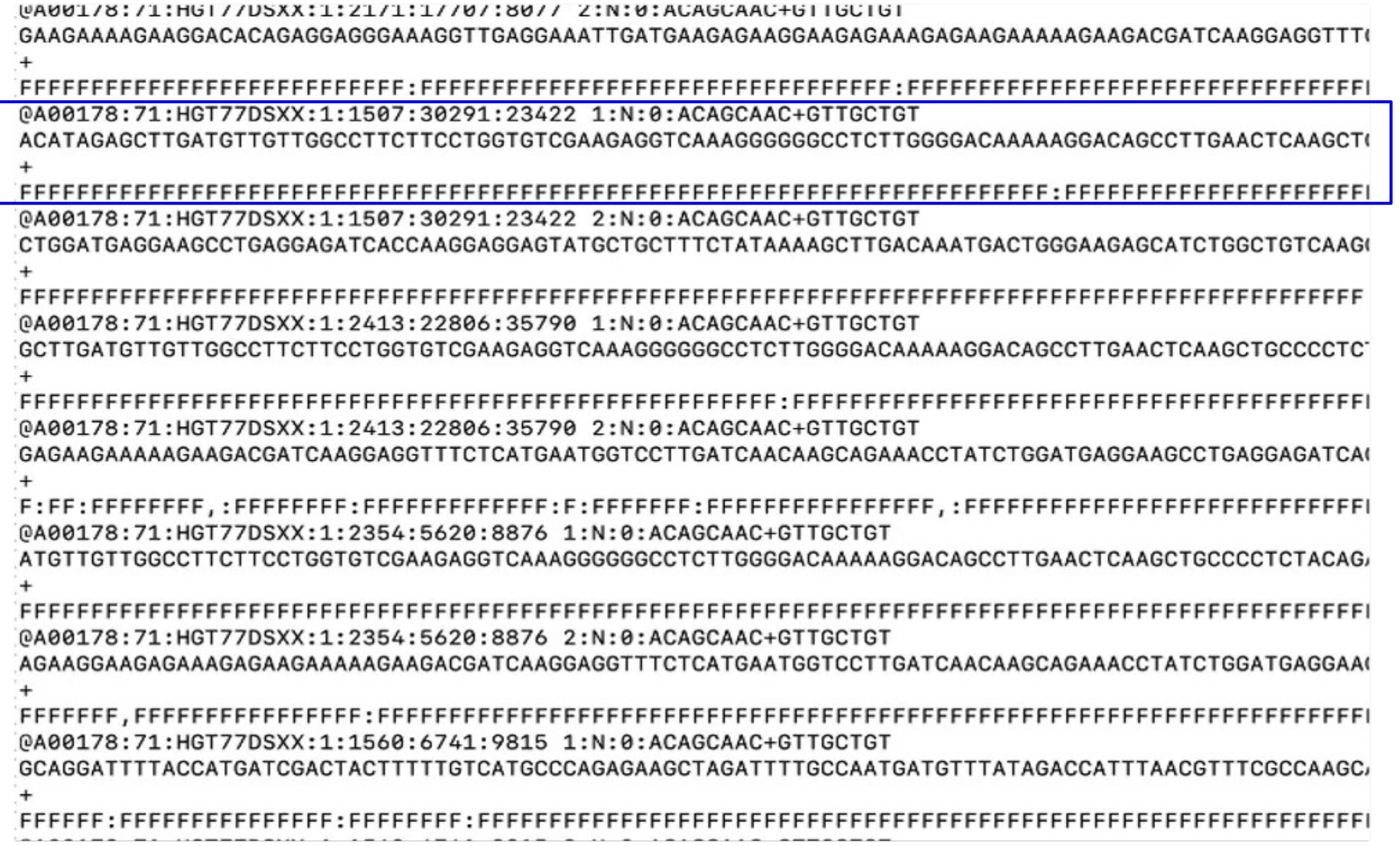
a. unique instrument name
b. run id
c. flowcell id
d. flowcell lane
e. tile number within the flowcell lane
f. x-coordinate of the cluster within the tile
g. y-coordinate of the cluster within the tile

h. the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
i. Y if the read fails filter (read is bad), N otherwise
j. 0 when no control bits are on
k. index sequence

https://www.gdc-docs.ethz.ch/MDA/site/getdata/

Fastq format

Each sequence read is represented by 4 lines

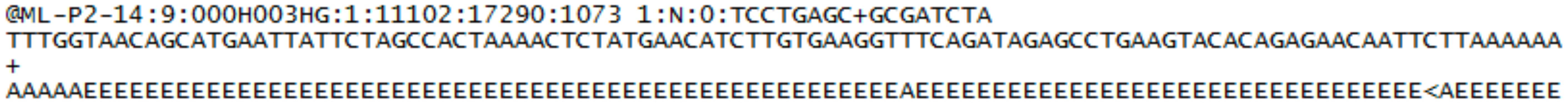


```
@A00178:71:HGT77DSXX:1:2171:17707:8077 2:N:0:ACAGCAAC+GTTGCTGT
GAAGAAAAGAAGGACACAGAGGAGGGAAAGGTTGAGGAAATTGATGAAGAGAAGGAAGAGAAAGAGAAGAAAAAGAAGACGATCAAGGAGGTTT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:1507:30291:23422 1:N:0:ACAGCAAC+GTTGCTGT
ACATAGAGCTTGATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAAGGACAGCCTTGAACTCAAGCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:1507:30291:23422 2:N:0:ACAGCAAC+GTTGCTGT
CTGGATGAGGAAGCCTGAGGAGATCACCAAGGAGGAGTATGCTGCTTTCTATAAAAGCTTGACAAATGACTGGGAAGAGCATCTGGCTGTCAAG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2413:22806:35790 1:N:0:ACAGCAAC+GTTGCTGT
GCTTGATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAAGGACAGCCTTGAACTCAAGCTGCCCCTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2413:22806:35790 2:N:0:ACAGCAAC+GTTGCTGT
GAGAAGAAAAAGAAGACGATCAAGGAGGTTTCTCATGAATGGTCCTTGATCAACAAGCAGAAACCTATCTGGATGAGGAAGCCTGAGGAGATCA
+
F:FF:FFFFFFFF,:FFFFFFFF:FFFFFFFFFFFFF:F:FFFFFFF:FFFFFFFFFFFFFFF,:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2354:5620:8876 1:N:0:ACAGCAAC+GTTGCTGT
ATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAAGGACAGCCTTGAACTCAAGCTGCCCCTCTACAG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2354:5620:8876 2:N:0:ACAGCAAC+GTTGCTGT
AGAAGGAAGAGAAAGAGAAGAAAAAGAAGACGATCAAGGAGGTTTCTCATGAATGGTCCTTGATCAACAAGCAGAAACCTATCTGGATGAGGAA
+
FFFFFFF,FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:1560:6741:9815 1:N:0:ACAGCAAC+GTTGCTGT
GCAGGATTTTACCATGATCGACTACTTTTTGTCATGCCCAGAGAAGCTAGATTTTGCCAATGATGTTTATAGACCATTTAACGTTTCGCCAAGC
+
FFFFFF:FFFFFFFFFFFFFFFF:FFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

1. Fastq header

2. The read sequence

3. Sequence/quality line separator

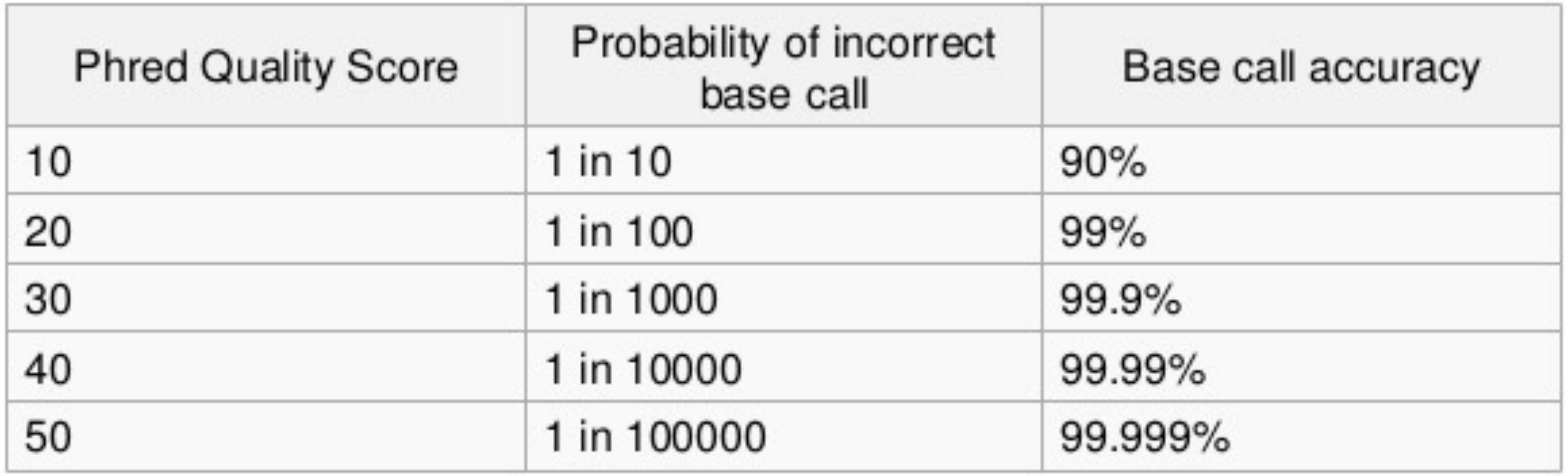


```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

4. Sequence quality. There is one character for each nucleotide. The characters relate to a sequence quality score e.g. how likely is the nucleotide correct? Known as **Phred score**

Quality score interpretation

$$Q = -10\ \log_{10} P \quad \longrightarrow \quad P = 10^{\frac{-Q}{10}}$$



Phred 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%
50	1 in 100000	99.999%

The quality (Q), also called Phred score, is the probability (P) that the corresponding basecall is incorrect.

fast5 format

Binary file (not human readable)

Contains:

- Sequence of a read
- Raw signal data from pore
- Additional log files

Typically convert fast5 to fastq for downstream analyses

BAM format for read data

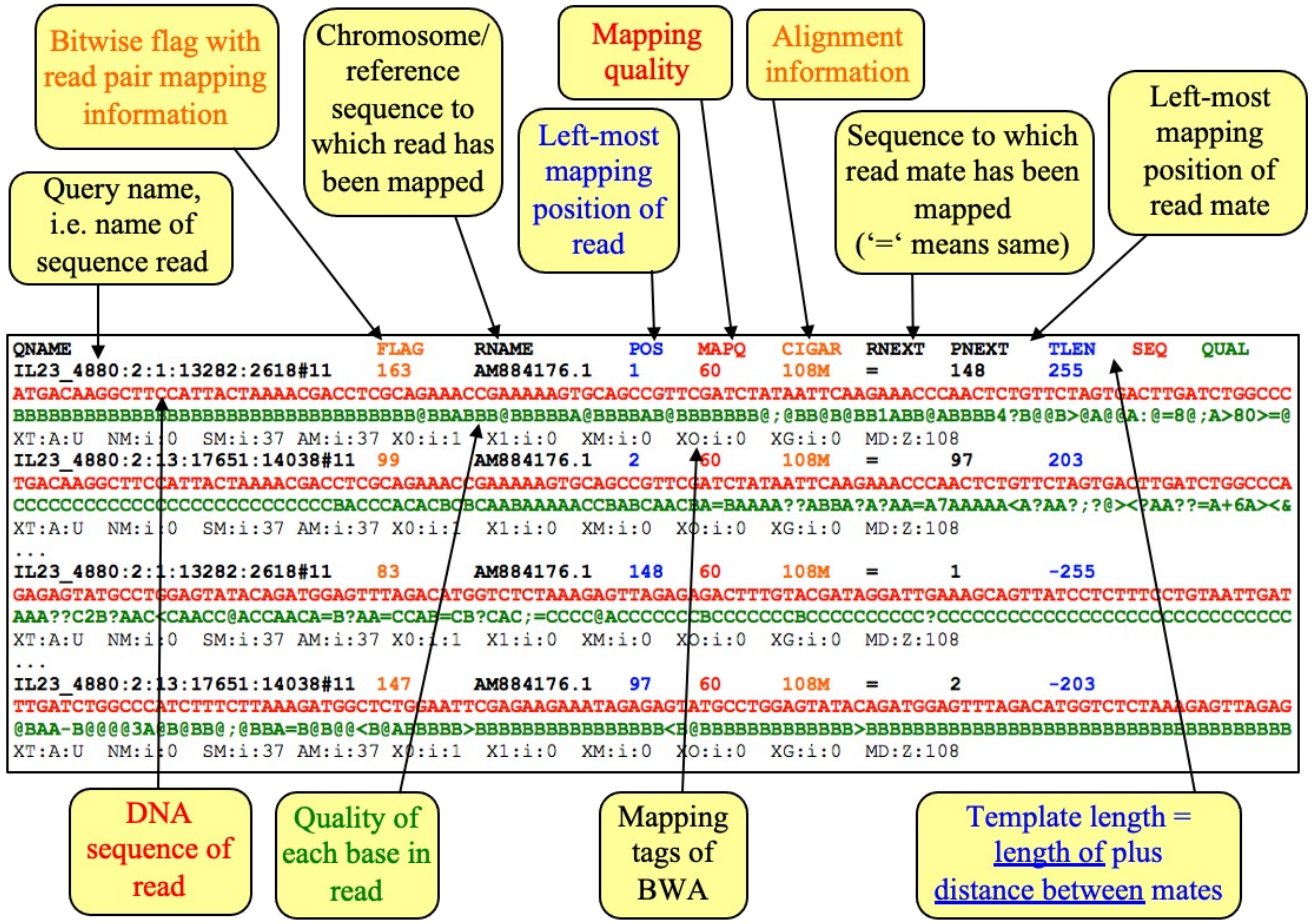
Binary Alignment Map format

Binary conversion of the Sequence Alignment Map (SAM) file

Typically convert bam to fastq for downstream analyses

File format: SAM / BAM (each line: one aligned sequence read)

The SAM/BAM file format is very powerful. It is unlikely that you will need to work with the contents of a SAM/BAM file directly, but it is very informative to visualize it in a viewer and it is a great format to do further analysis with. The format specifications are at http://samtools.sourceforge.net/SAM1.pdf. Below is a brief overview of the information contained in such files.



Query name, i.e. name of sequence read

Bitwise flag with read pair mapping information

Chromosome/ reference sequence to which read has been mapped

Left-most mapping position of read

Mapping quality

Alignment information

Sequence to which read mate has been mapped ('=' means same)

Left-most mapping position of read mate

```
QNAME                        FLAG     RNAME        POS    MAPQ    CIGAR   RNEXT   PNEXT     TLEN     SEQ     QUAL
IL23_4880:2:1:13282:2618#11  163      AM884176.1   1      60      108M    =       148       255
ATGACAAGGCTTCCATTACTAAAACGACCTCGCAGAAACCGAAAAAGTGCAGCCGTTCGATCTATAATTCAAGAAACCCAACTCTGTTCTAGTGACTTGATCTGGCCC
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB@BBABBB@BBBBBA@BBBBAB@BBBBBBB@;@BB@B@BB1ABB@ABBBB4?B@@B>@A@@A:@=8@;A>80>=@
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:108
IL23_4880:2:13:17651:14038#11  99     AM884176.1   2      60      108M    =       97        203
TGACAAGGCTTCCATTACTAAAACGACCTCGCAGAAACCGAAAAAGTGCAGCCGTTCGATCTATAATTCAAGAAACCCAACTCTGTTCTAGTGACTTGATCTGGCCCA
CCCCCCCCCCCCCCCCCCCCCCCCCCCBACCCACACBCBCAABAAAAACCBABCAACBA=BAAAA??ABBA?A?AA=A7AAAAA<A?AA?;?@><?AA??=A+6A><&
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:108
...
IL23_4880:2:1:13282:2618#11  83       AM884176.1   148    60      108M    =       1         -255
GAGAGTATGCCTGGAGTATACAGATGGAGTTTAGACATGGTCTCTAAAGAGTTAGAGAGACTTTGTACGATAGGATTGAAAGCAGTTATCCTCTTTCCTGTAATTGAT
AAA??C2B?AAC<CAACC@ACCAACA=B?AA=CCAB=CB?CAC;=CCCC@ACCCCCCCBCCCCCCCBCCCCCCCCCC?CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:108
...
IL23_4880:2:13:17651:14038#11  147    AM884176.1   97     60      108M    =       2         -203
TTGATCTGGCCCATCTTTCTTAAAGATGGCTCTGGAATTCGAGAAGAAATAGAGAGTATGCCTGGAGTATACAGATGGAGTTTAGACATGGTCTCTAAAGAGTTAGAG
@BAA-B@@@@3A@B@BB@;@BBA=B@B@@<B@ABBBBB>BBBBBBBBBBBBBBBB<B@BBBBBBBBBBBBB>BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:108
```

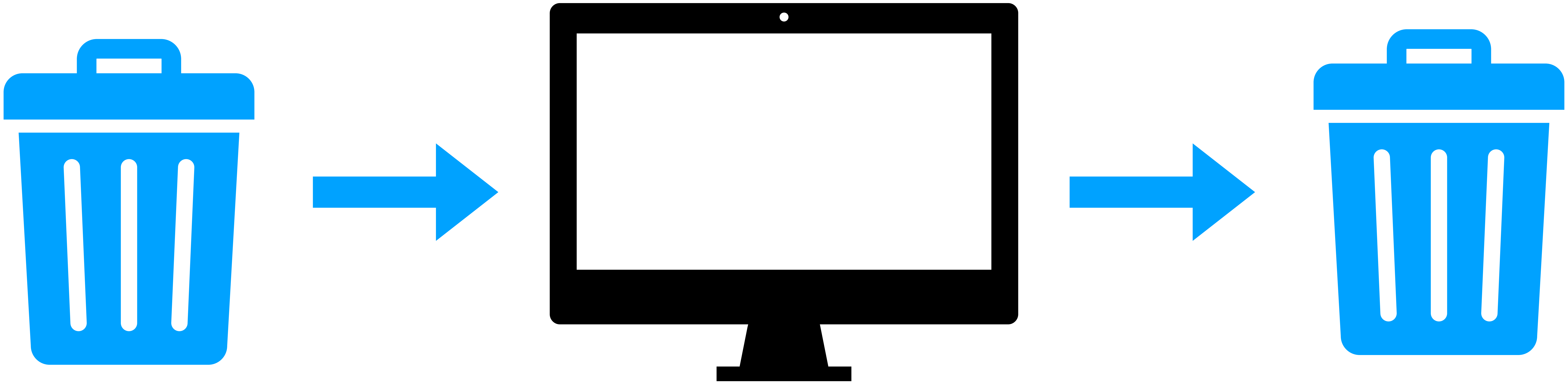
DNA sequence of read

Quality of each base in read

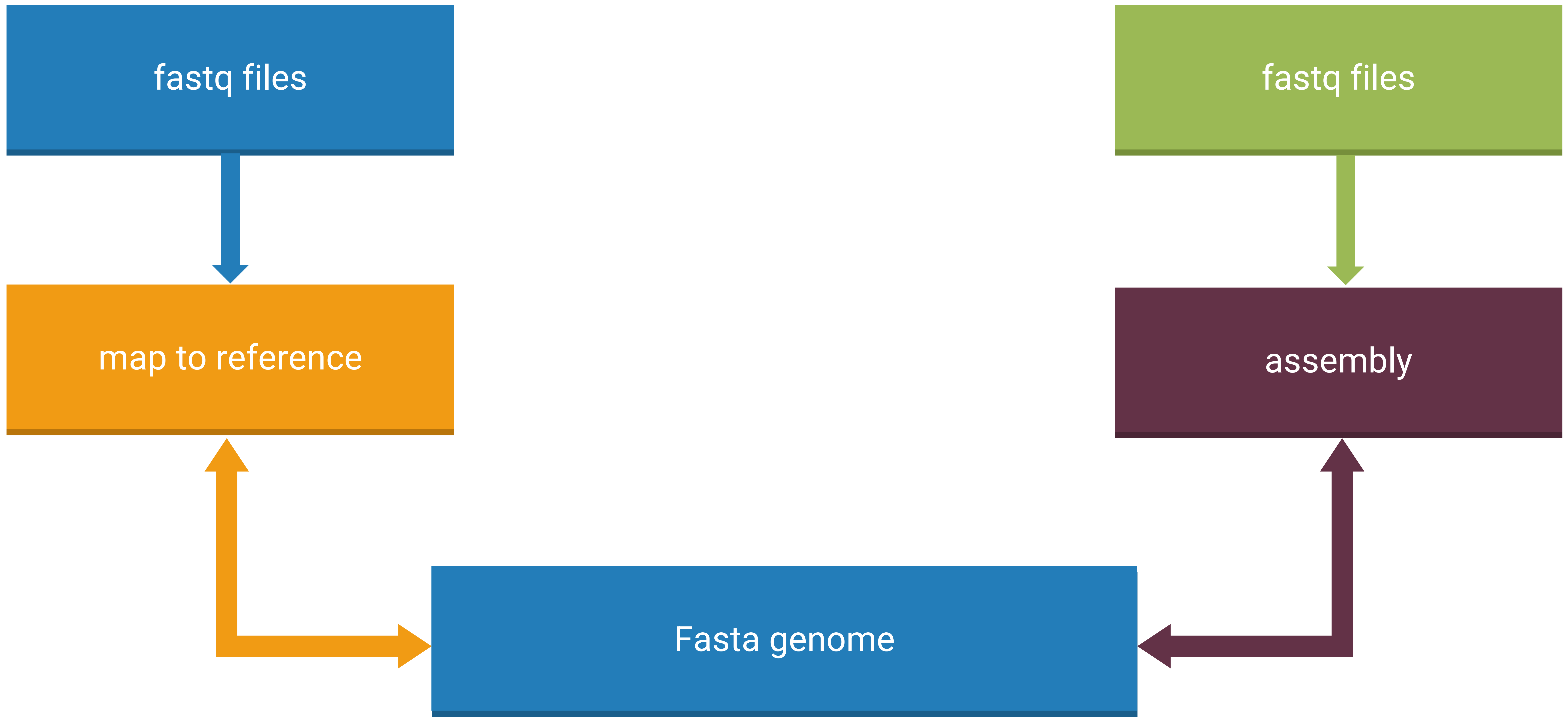
Mapping tags of BWA

Template length = length of plus distance between mates

The “Golden” Rule



Having good quality fastq data is important!!



fastq files

map to reference

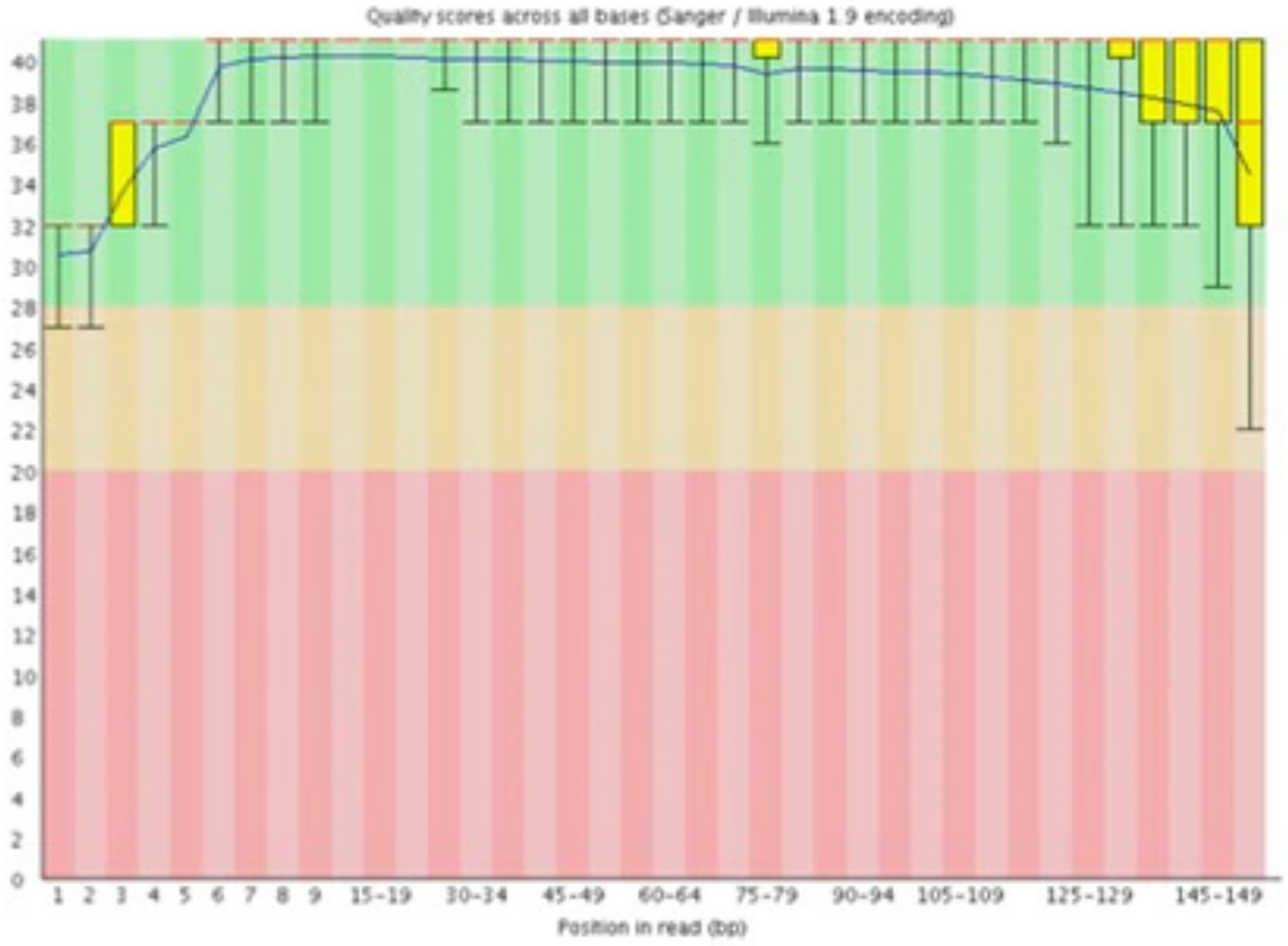
fastq files

assembly

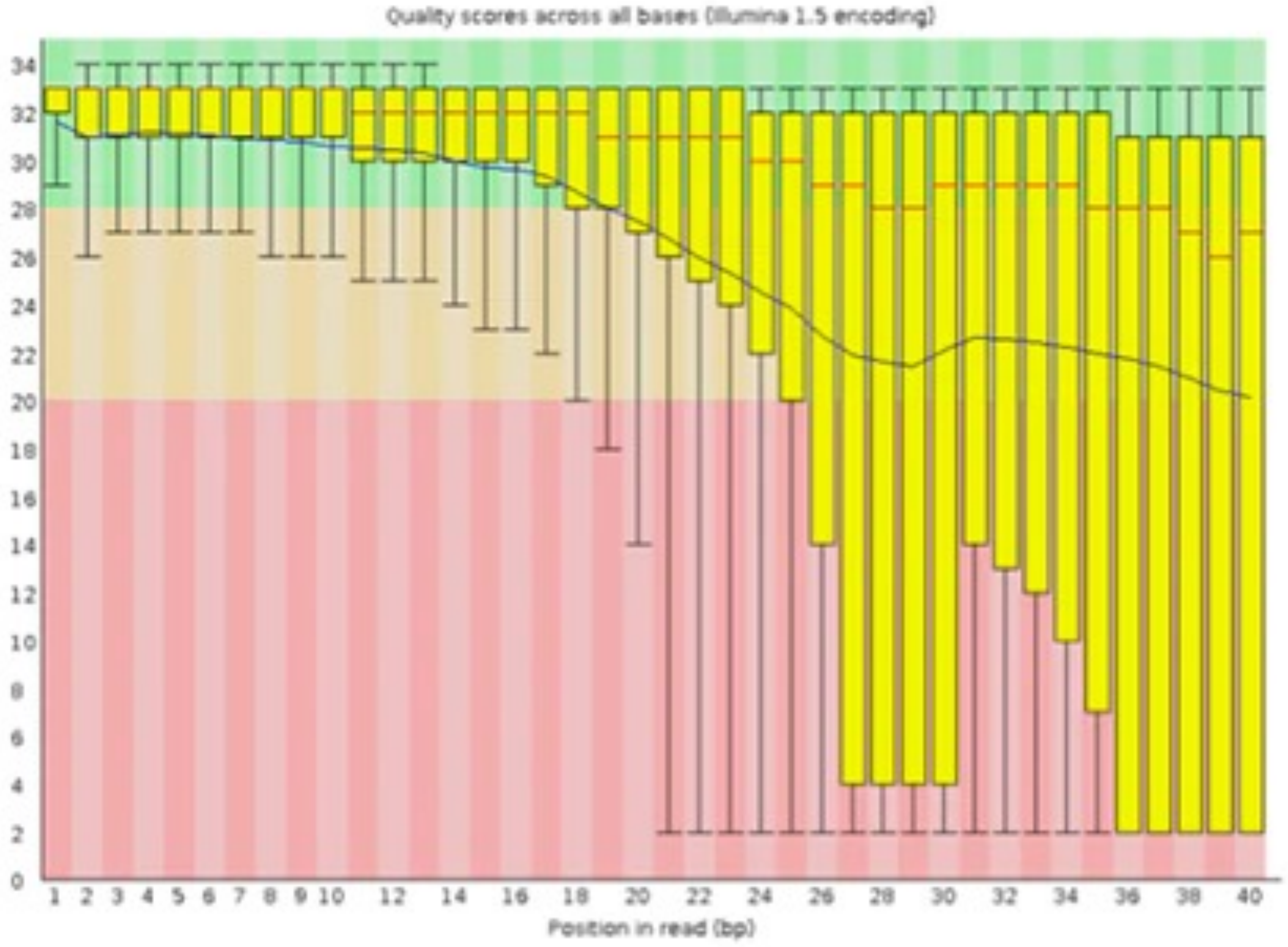
Fasta genome

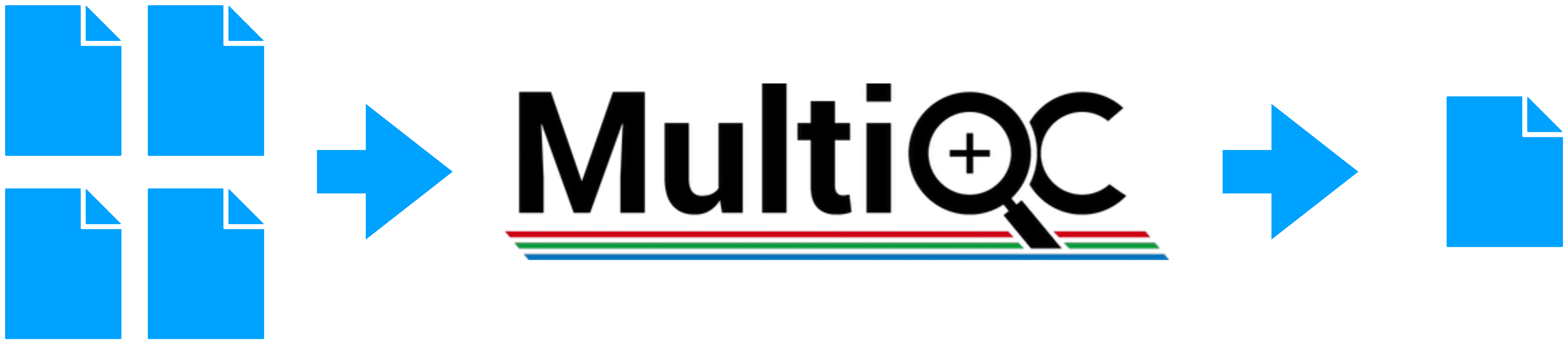
FastQC: Per base sequence quality

Good data



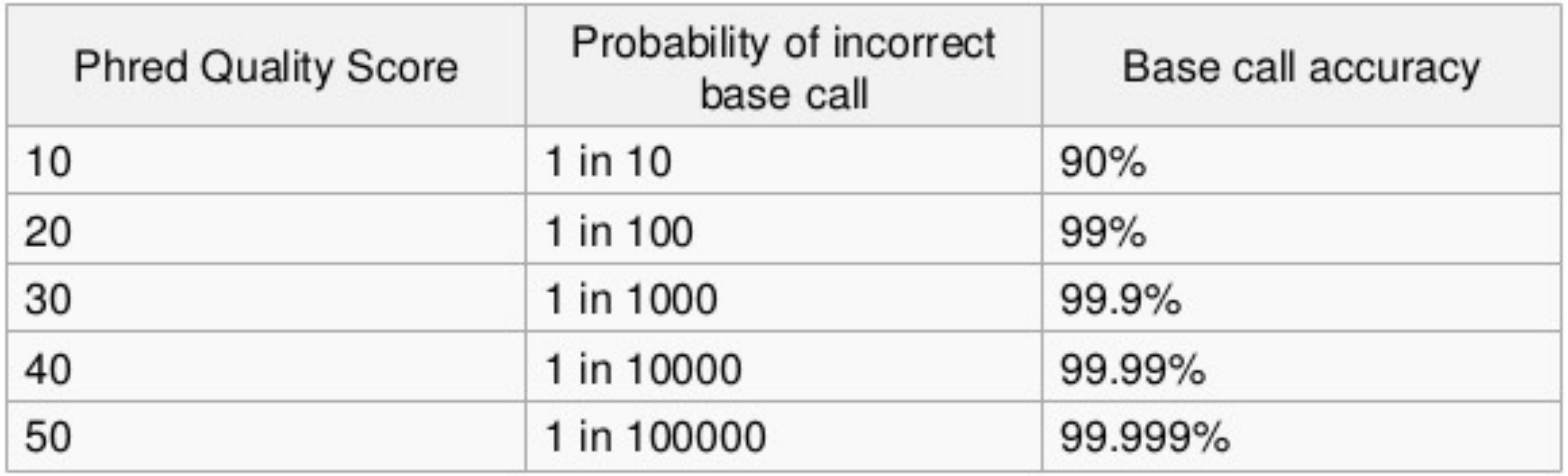
Bad data





Quality score interpretation

$$Q = -10\ \log_{10} P \quad \Rightarrow \quad P = 10^{\frac{-Q}{10}}$$



Phred 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%
50	1 in 100000	99.999%

The quality (Q), also called Phred score, is the probability (P) that the corresponding basecall is incorrect.

Many tools for trimming

- Trimmomatic
- sickle
- fastP
- bbduk
- cutadapt
- Trim Galore

Questions?