FASTQ and BAM formats and assessing quality

MSc in Genomic Medicine

Lucy Crooks

24/1/2017

Output from sequencing is list of bases for each read

Output from sequencing is list of bases for each read

• Ion Torrent PGM	5 million reads	1 GB
• MiSeq	25 million reads	6 GB
• HiSeq rapid run	600 million	150 GB
• HiSeq high-output	4 billion	1 TB

Equivalent to 40 HD movies

File sizes are for 100 bp reads, unzipped
Number of reads from thermofisher.com and Illumina.com

FASTQ format

- Text file
- Can be compressed as .gz
- Four lines per read

```
@M00969:31:000000000-A5GV2:1:1106:21539:11519 2:N:0:2
ATTAAATTCTCAAATTTAATTTTGAGAAGGTTGGTAGAATACTCC
+
CCCCCFFFFFFFGGGGGGGGGGHHGGCHGHHHHGFHHHHHHHHHH
```

Illumina read

FASTQ format

First line gives the read id

@M00969:31:000000000-A5GV2:1:1106:21539:11519 2:N:0:2
ATTAAATTCTCAAATTTAATTTTGAGAAGGTTGGTAGAATACTCC
+
CCCCCFFFFFFFGGGGGGGGGGHHGGCHGHHHHGFHHHHHHHHHH

FASTQ format

Second line gives the sequence of bases

```
@M00969:31:000000000-A5GV2:1:1106:21539:11519 2:N:0:2
ATTAAATTCTCAAATTTAATTTTGAGAAGGTTGGTAGAATACTCC
+
CCCCCFFFFFFFGGGGGGGGGGHHGGCHGHHHHGFHHHHHHHHHH
```

Written in order it is generated by the machine

Could be on the forward or reverse strand

FASTQ format

Fourth line gives a quality score for each base

```
@M00969:31:000000000-A5GV2:1:1106:21539:11519 2:N:0:2
ATTAAATTCTCAAATTTAATTTTGAGAAGGTTGGTAGAATACTCC
+
CCCCCFFFFFFFGGGGGGGGGGHHGGCHGHHHHGFHHHHHHHHHH
```

What do the quality characters mean?

- ASCII symbols to reduce file size
- ASCII are symbols like ! " # $ % & ' and characters
- You can look the values up in tables
- Subtract 33 and that gives you the quality score

What do the quality numbers mean?

- Quality tells you about the error probability
  e.g. chance that the base call is wrong
- Higher quality means less chance of error
- Values are Phred scaled

Phred scale

$$p = 10^{-\frac{Q}{10}}$$

p is the error probability
Q is the quality score

If Q = 30 $\quad -\frac{Q}{10} = -3 \quad p = 10^{-3}$

ie 0.001 or 1 in 1000 chance of error

Exercise: Looking at a FASTQ file

- Go to Galaxy and open the FASTQ file from yesterday

  (click on the eye icon)

- Chose a read and write down its id
- Write down the series of bases and the base qualities
- Work out the base quality scores in Phred and probabilities
- You can use the table on this website

  http://www.somewhereville.com/?p=1508

Paired-end reads

- Often we sequence both ends of the DNA fragments
- Two FASTQ files — read 1 and read 2
- Use the read id to match them back up

- Illumina read pairs have specific orientation

On opposite strands
Point towards each other on reference

Exercise: Looking at paired-end FASTQs

- Open one of the FASTQ files for NA12878
- Chose a read and write down
  - its id
  - First 5 and last 5 bases
  - ASCII characters for first 5 and last 5 bases
  - Whether it is read 1 or 2
- Open the other FASTQ file find the read with the same id
  write down the same details

Exercise: Using FastQC to assess read quality

- Click on NGS: QC and manipulation in the left hand column
- Click on FastQC Read Quality reports
- Choose the file from the drop down mean
- Click Execute
- Look at the Webpage output

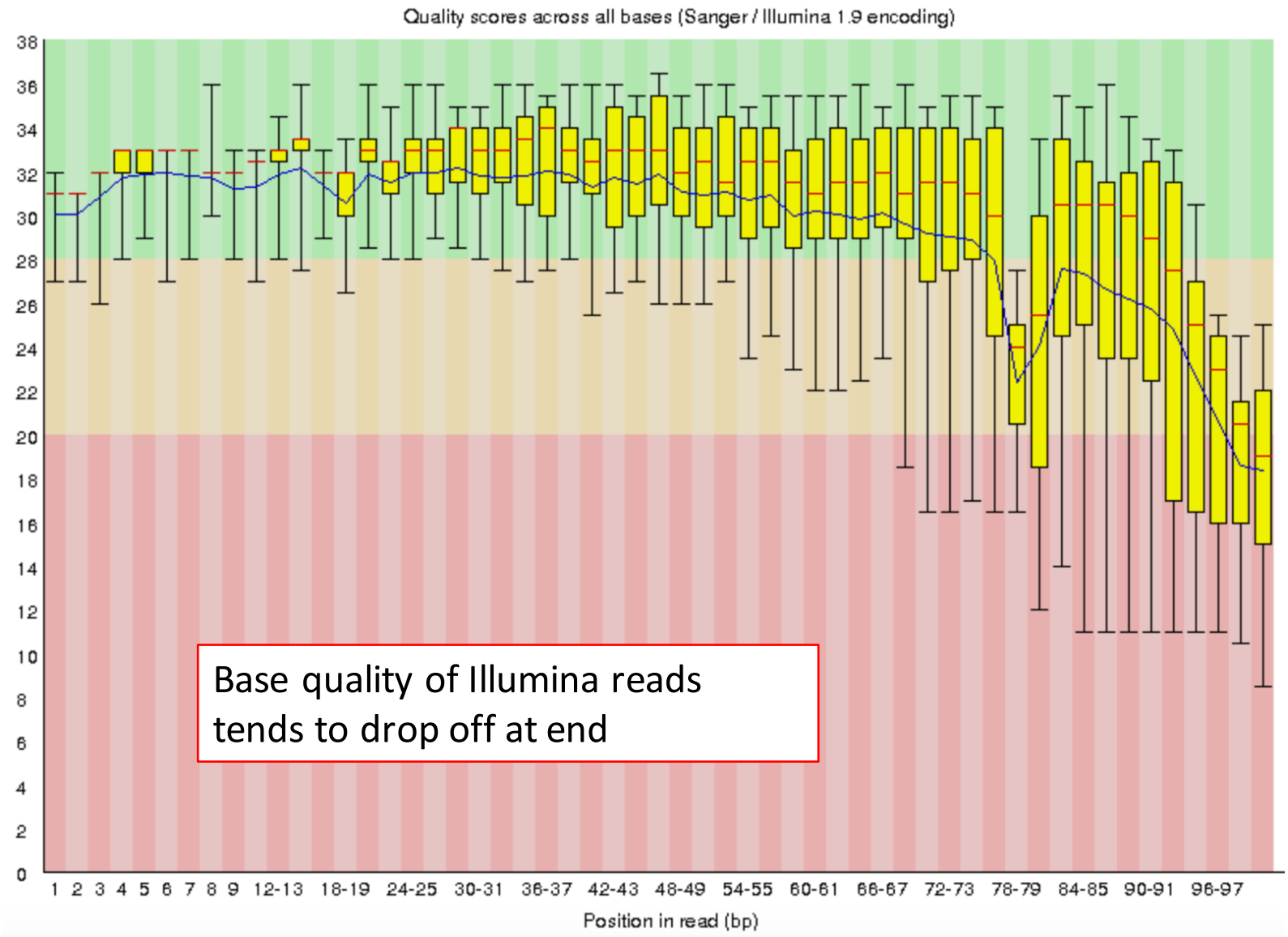
FastQC output

Basic Statistics

Measure	Value
Filename	NA12878_chrom_10_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	792910
Sequences flagged as poor quality	0
Sequence length	101
%GC	47

Per base sequence quality

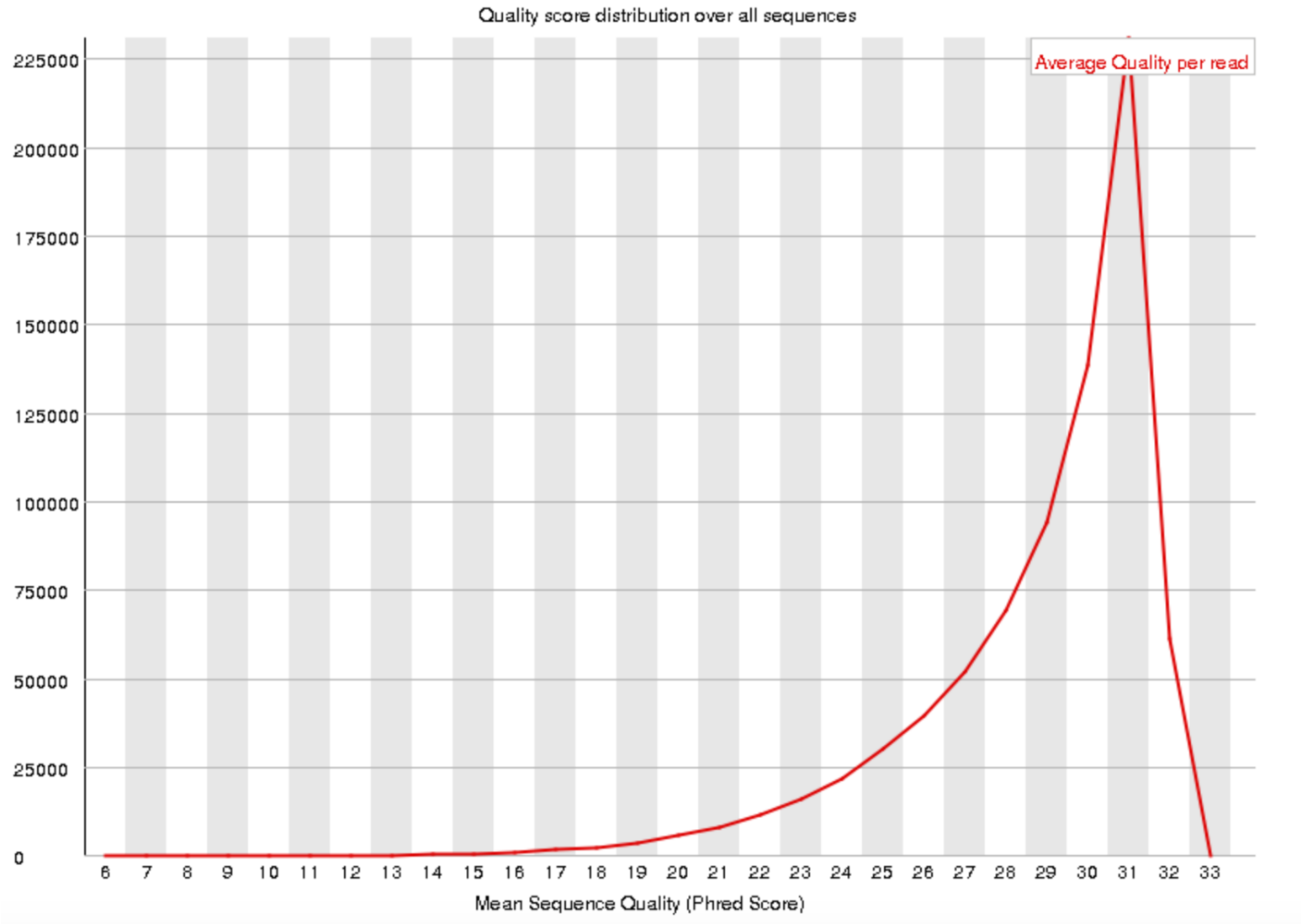
FastQC output



Base quality of Illumina reads tends to drop off at end

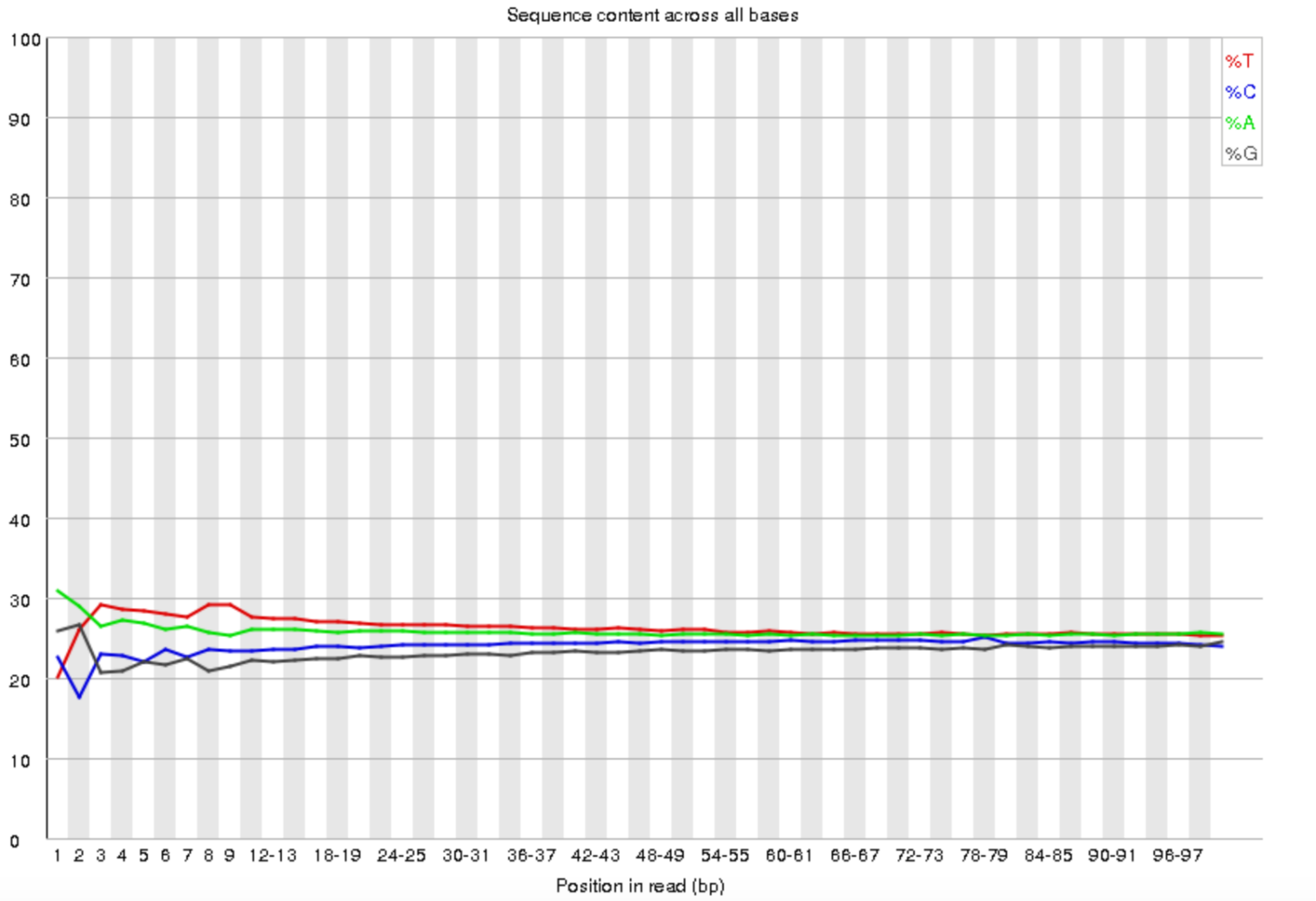
Per sequence quality scores

FastQC output



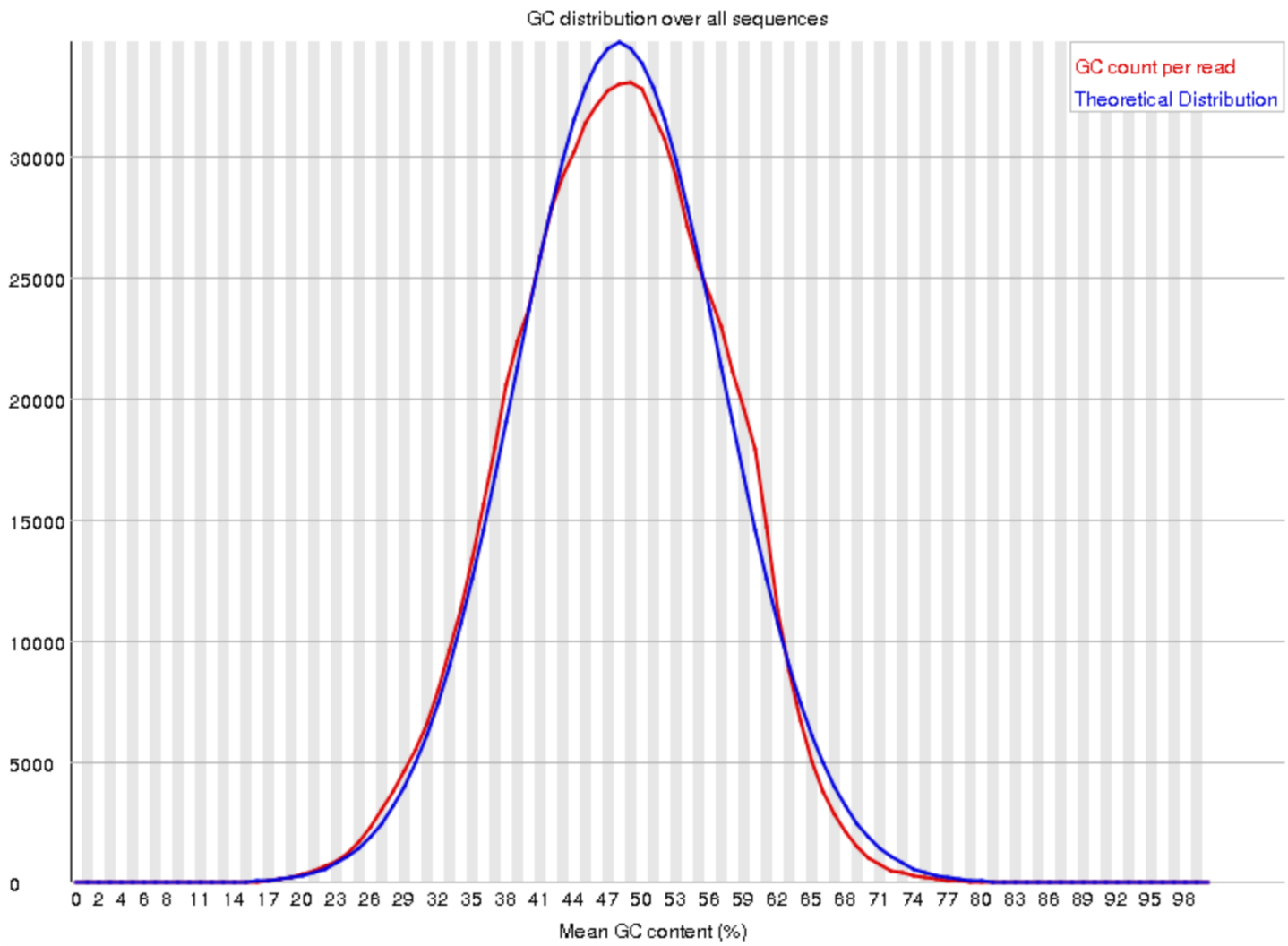
Per base sequence content

astQC output



Per sequence GC content

FastQC output



BAM is the format for storing aligned reads

- Is a binary format
- Human readable version is SAM
- Information on where read mapped, mapping quality and for paired-end reads where partner mapped

BAM format

One line per read

Same sequence identifier

SRR953254.23083 16 chr10 1890119 40 17M * 0 0
GAACGTCAATATCGCTA ,*,,///2444444444 AS:i:-3 XN:i:0
XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:14T2
YT:Z:UU

Many pieces of data (elements) separated by tabs

BAM format

Elements 2 and 3 give aligned position of 5' base

SRR953254.23083 16 chr10 1890119 40 17M * 0 0
GAACGTCAATATCGCTA ,*,,///2444444444 AS:i:-3 XN:i:0
XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:14T2
YT:Z:UU

BAM format

Element 4 is the mapping quality

```
SRR953254.23083 16      chr10   1890119 40      17M     *       0       0
GAACGTCAATATCGCTA       ,*,,///2444444444       AS:i:-3 XN:i:0
XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:14T2
      YT:Z:UU
```

- Is Phred scaled — gives probability that the alignment is wrong
- Typically aligners score multiple alignments and quality relates to score difference between best and second-best alignment
- Alignment scores can include penalties for gaps

BAM format

Element 5 is called a CIGAR string

SRR953254.23083 16 chr10 1890119 40 17M * 0 0
GAACGTCAATATCGCTA ,*,,///2444444444 AS:i:-3 XN:i:0
XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:14T2
YT:Z:UU

- Shows if there are gaps in the alignment

BAM CIGAR string

- Value before M is number of consecutive mapping bases (can be mismatches)
- Value before I is number of bases inserted relative to reference
- Value before D is number of bases deleted relative to reference

Example

142M**2I**7M	2 bp insertion after 142 bases
	then 7 aligned bases

BAM format

- The sequence of bases is element 9
- It is now written on the same strand as the reference

SRR953254.23083 16 chr10 189011 9 40 17M * 0 0
GAACGTCAATATCGCTA ,*,,///2444444444 AS:i:-3 XN:i:0
XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:14T2
YT:Z:UU

- And the base qualities are element 10

Exercise: Looking at a BAM file

You are going to look for the alignment of the Ion Torrent read you explored earlier

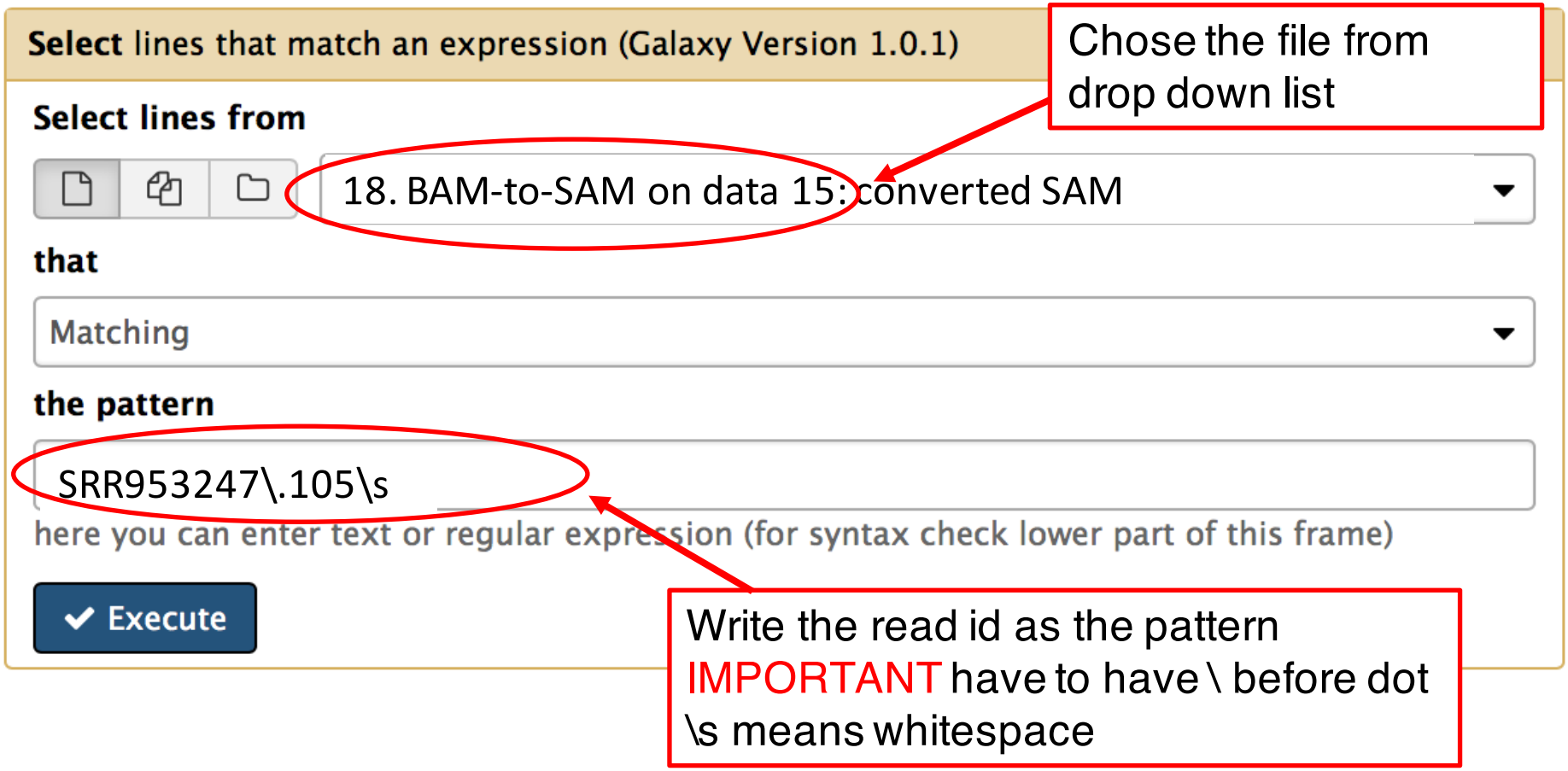
Exercise: Looking at a BAM file

Converting SAM to BAM

Use NGS: SAMtools then BAM-to-SAM

Finding a specific read in a SAM file

Click on Filter and Sort then Select lines that match an expression



Exercise: Looking at a BAM file

Write down all the information you can about the alignment of your read

Exercise: Looking at a BAM file

Write down all the information you can about the alignment of your read

- What is the probability that the alignment is wrong?
- Are there any gaps in the alignment?
- Is the sequence of bases the same as you have written down?
- Are the quality scores the same as you have written down?

BAM format

The first element is a number (flag) that summarises the alignment

SRR953254.23083 16 chr10 1890119 40 17M * 0 0
GAACGTCAATATCGCTA ,*,,///2444444444 AS:i:-3 XN:i:0
XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:14T2
YT:Z:UU

- For single reads, 16 means read is on reverse strand to reference
  0 means read is on forward strand
  4 means read did not mapped

- For paired-end reads the flags are more complicated

BAM format — paired-end reads

For paired-end reads the BAM file additionally gives the position of the mate

M00969:31:000000000-A5GV2:1:1106:21539:11519 163 chr15
86115312 60 151M = 86115317 156
ATTAAATTCTCAAATTTAATTTTGAGAAGGTTGGTAGAATACTCC
CCCCCFFFFFFFGGGGGGGGGGHHGGCHGHHHHGFHHHHHHHH
HH X0:i:1 X1:i:0 MD:Z:151 PG:Z:MarkDuplicates RG:Z:NDD
XG:i:0 AM:i:37 NM:i:0 SM:i:37 XM:i:0 XO:i:0 MQ:i:60 XT:A:U

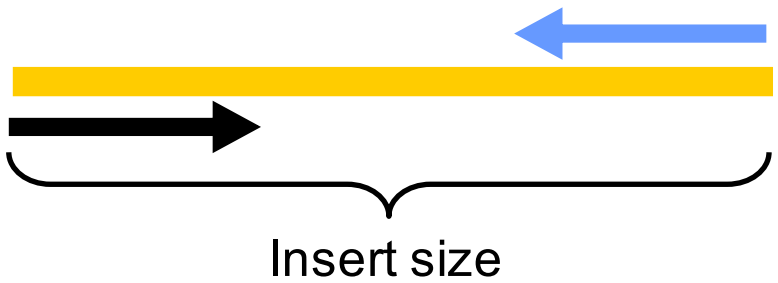
= means mate maps to same chromosome

- Get two lines for pair, one for each read

Exercise: Looking at paired-end SAM file

- Open the SAM file for NA12878
- Find the reads that you recorded earlier
- Comparing to the data from the FASTQs
  try to work out the relative alignment of the two reads
- Look up what the alignment flags mean using the tool at

  https://broadinstitute.github.io/picard/explain-flags.html



Element 8 of SAM gives insert size, but can be wrong!

Reads can overlap

Steps that can improve quality

- Detecting duplicates
  - Duplicates arise during PCR in library prep
  - Important because fragments are assumed to be independent samples from genome
  - NOT for Ion Torrent amplicons
- Realignment around InDels
- Set minimum base quality and mapping quality for reads to be considered in variant calling