NGS data formats and Quality Control

Jacqui Keane

@drjkeane
drjkeane@gmail.com

Adapted from slides provided by Petr Danecek
petr.danecek@sanger.ac.uk

Quality Control

Quality Control

The commands I run:

```
samtools stats file.bam > file.bam.stats
plot-bamstats -p plots/ file.bam.stats
```

The questions I want to answer:

- Do I have enough read coverage with my reads?
- Was the library creation process efficient and problem-free?
- Did the sequencing process create artefacts?

Read coverage

Read coverage / depth

- is every genomic position “covered” to a sufficient depth?
- maximum average depth: (number-of-reads x read-length) / target-size
- average depth: (number-of-reads-mapped x read-length) / target-size
  - the whole bacteria genome target-size = reference sequence length = 4Mbp

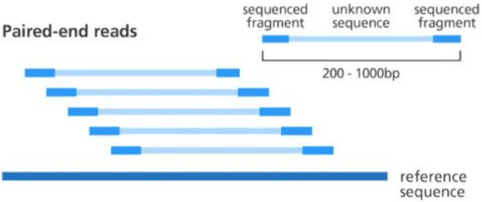


Image credit: Genome Research Limited

Useful coverage

- 10x ok for variants calling
- 30x ok for most things (variant calling, assembly)
- 100x more than enough, pipelines subsample down to this

Library prep biases: PCR duplicates

Experiments start with small amounts of DNA

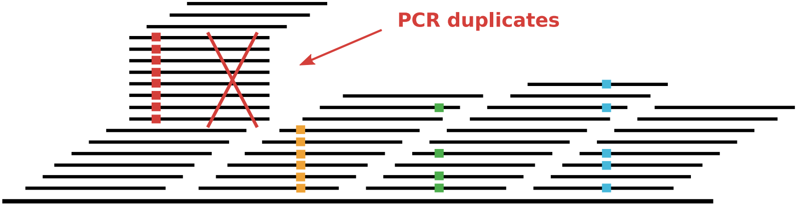
- a PCR amplification step is necessary for Illumina sequencing: one molecule => many identical molecules

Problem:

- additional PCR-copy molecules are not informative

Solution:

- infer and mark PCR-dupliates, discount in later analysis
  - mark if reads and their mates start at the same position
- use `picard MarkDuplicates` or `samtools markdup`
- typical dup rates: Exomes ~ 15-20%, Genomes < 5%

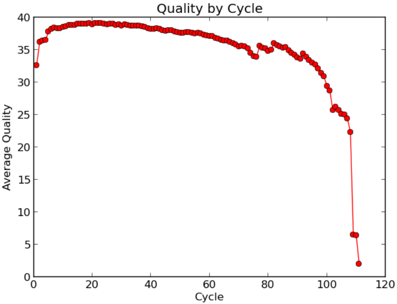


Base quality

Sequencing by synthesis: dephasing

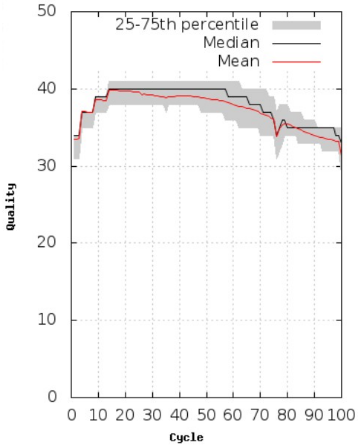
- growing sequences in a cluster gradually desynchronize
- error rate increases with read length

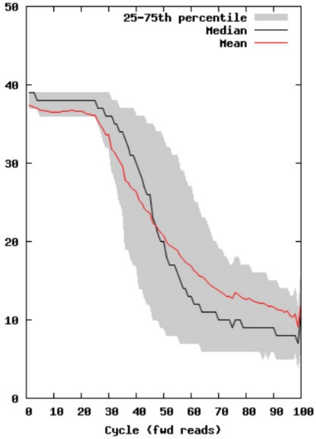
Calculate the average quality at each position across all reads



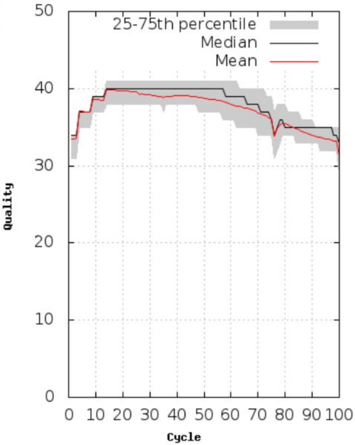
Quality	Probability of error	Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

Base quality

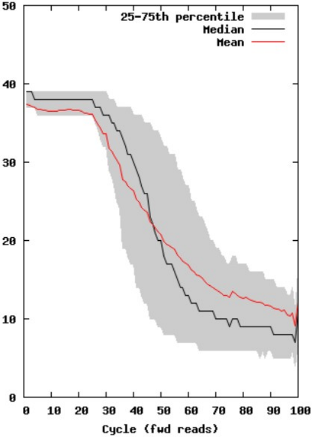




Base quality



✓

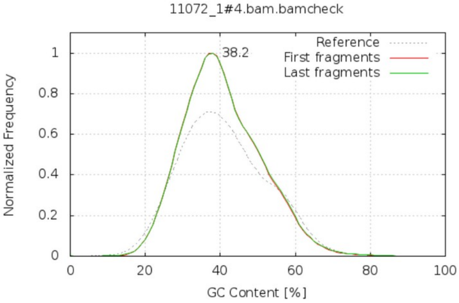


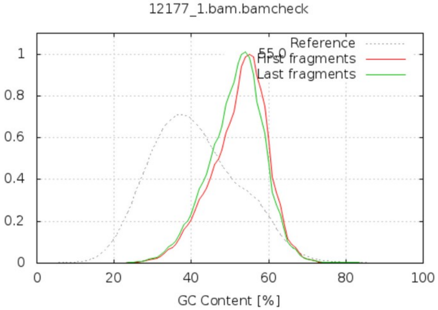
✗

GC bias

GC- and AT-rich regions are more difficult to amplify

- compare the GC content against the expected distribution (reference sequence)

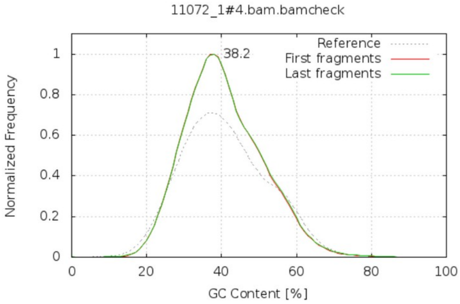


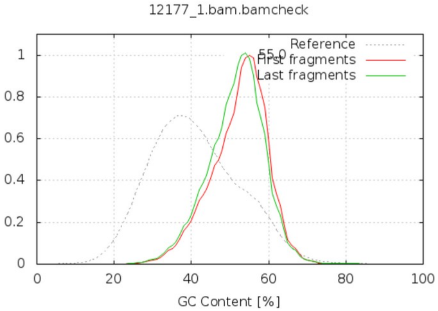


GC bias

GC- and AT-rich regions are more difficult to amplify

- compare the GC content against the expected distribution (reference sequence)





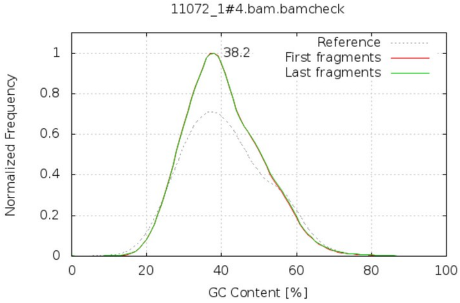
✓

✗

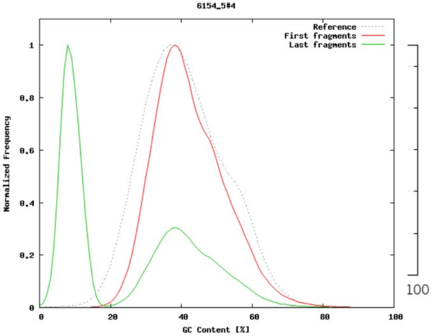
GC bias

GC- and AT-rich regions are more difficult to amplify

- compare the GC content against the expected distribution (reference sequence)

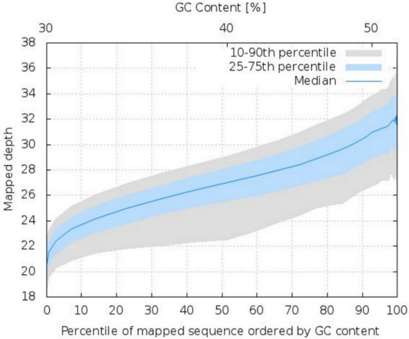


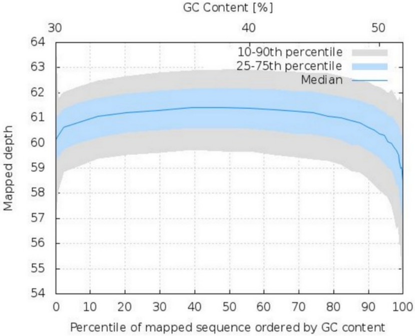
✓



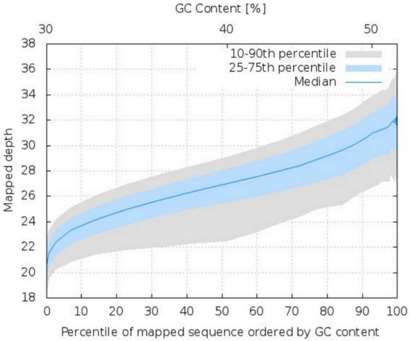
✗

GC content vs depth

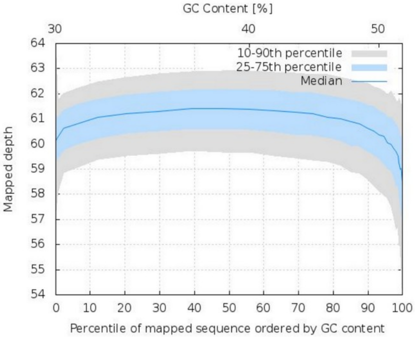




GC content vs depth



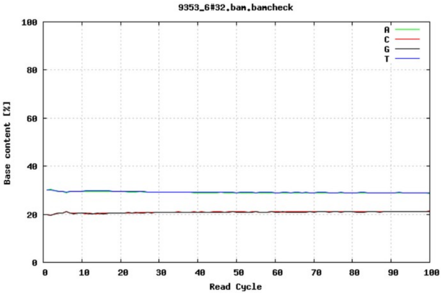
✗

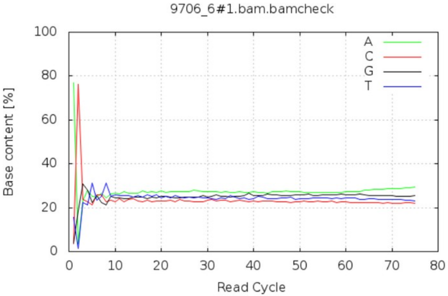


✓

GC content by cycle

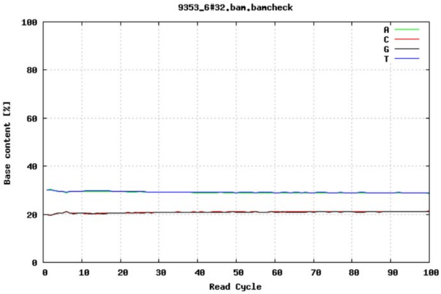
Was the adapter sequence trimmed?

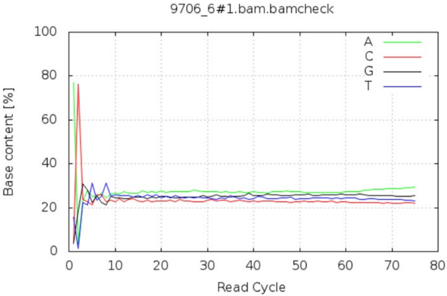




GC content by cycle

Was the adapter sequence trimmed?



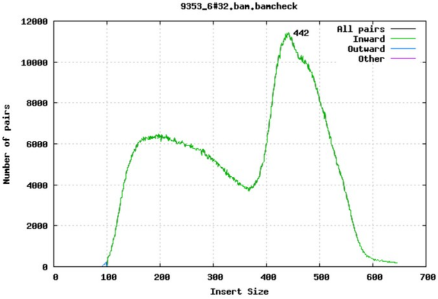


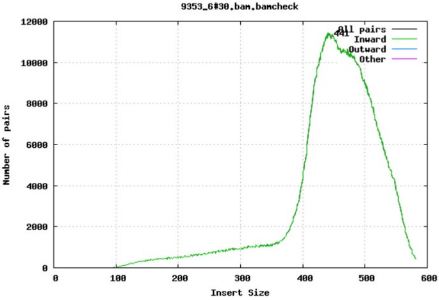
✓

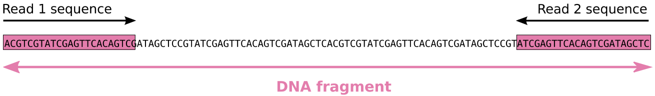
✗

Fragment size

Paired-end sequencing: the size of DNA fragments matters






Read 1 sequence

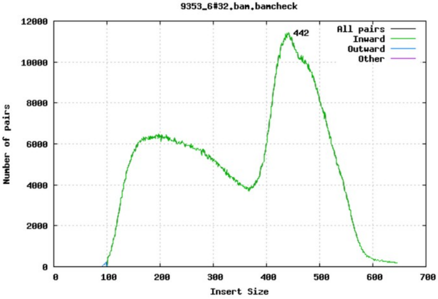
Read 2 sequence

ACGTCGTATCGAGTTCACAGTCGATAGCTCCGTATCGAGTTCACAGTCGATAGCTCACGTCGTATCGAGTTCACAGTCGATAGCTCCGTATCGAGTTCACAGTCGATAGCTC

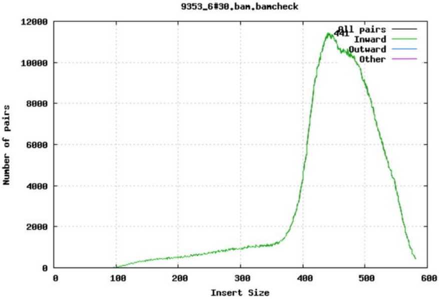
DNA fragment

Fragment size

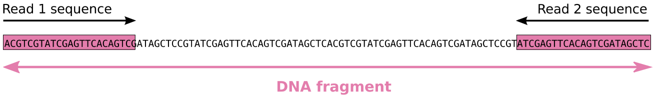
Paired-end sequencing: the size of DNA fragments matters



✗



✓



Read 1 sequence

Read 2 sequence

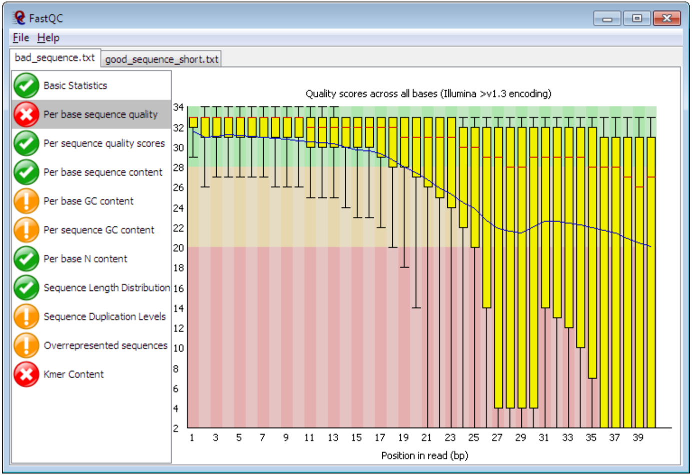
ACGTCGTATCGAGTTCACAGTCGATAGCTCCGTATCGAGTTCACAGTCGATAGCTCACGTCGTATCGAGTTCACAGTCGATAGCTCCGTATCGAGTTCACAGTCGATAGCTC

DNA fragment

Quality Control

FastQC/MultiQC are alternative tools for QC

```
fastqc *.fastq.gz
multiqc .
```


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)

Quality Control

Other tools I use:

```
kraken - a taxonomic classification tool for sequence data
bactinspector - determines the most probable species based on sequence data
confindr - detection of intraspecies and cross-species
           contamination in bacterial whole-genome sequence data
```

Other important questions I ask

- Is my sequence data the species I think it is?
- Is there any contamination in my samples?
  - Intraspecies contamination e.g. heterozygous SNPs
  - Cross-species contamination e.g. GC content, bactinspector/confindr