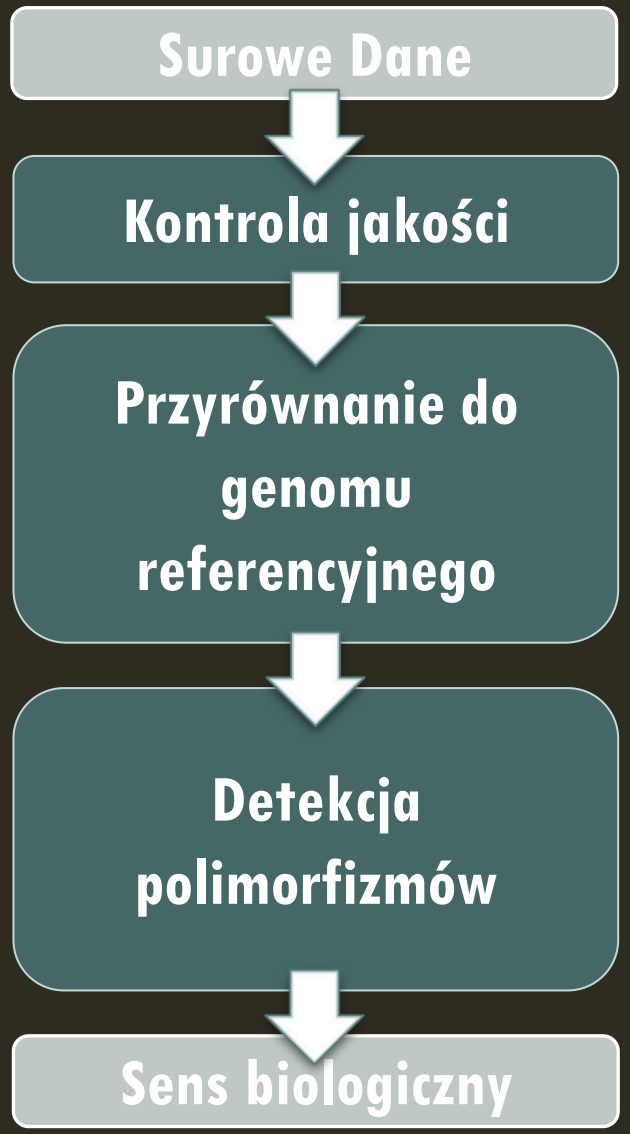
KONTROLA JAKOŚCI I EDYCJA DANYCH NGS

Analiza danych NGS
Wykład 4

PIPELINE

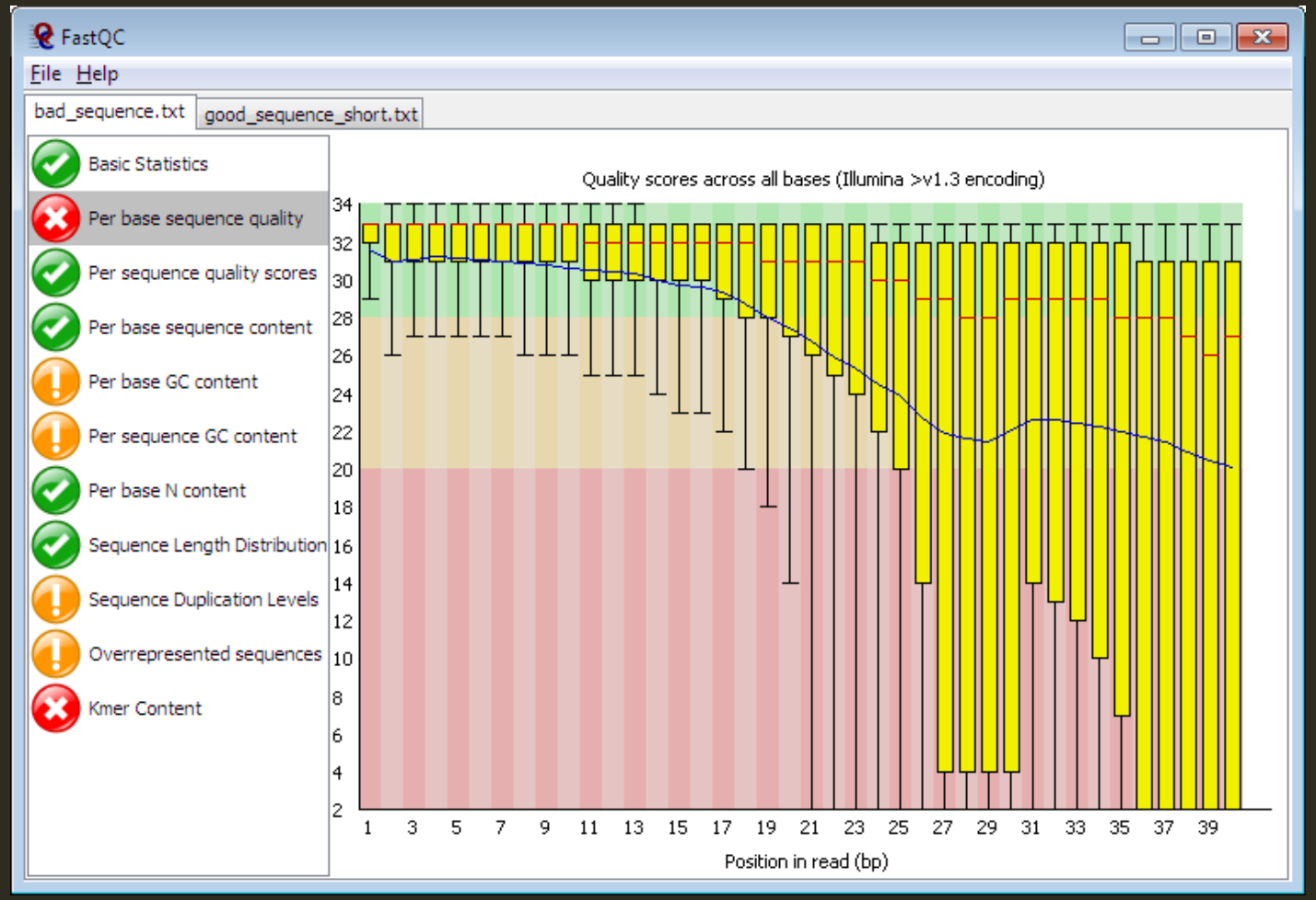
Standardowe kroki w analizie danych NGS



KONTROLA JAKOŚCI

FASTQC

KONTROLA JAKOŚCI



Surowe Dane

Kontrola jakości

Przyrównanie do genomu referencyjnego

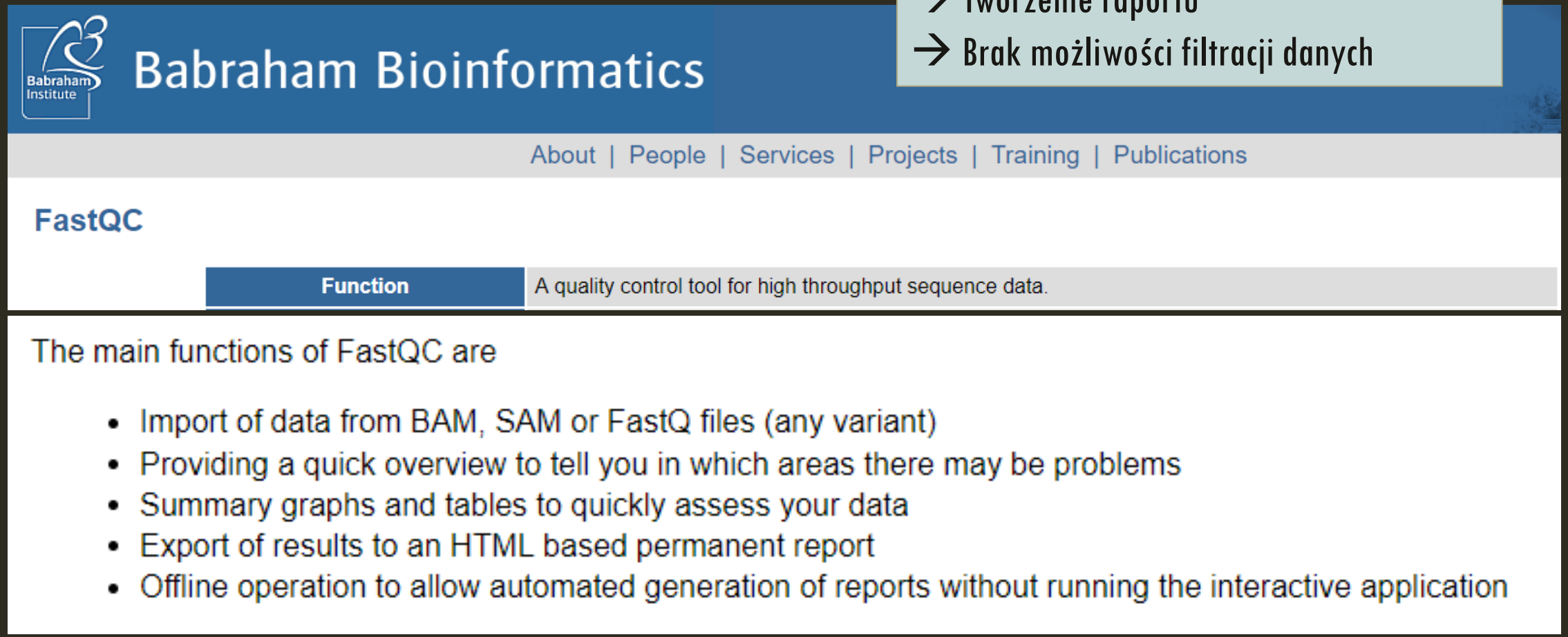
Detekcja polimorfizmów

Sens biologiczny

www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC

- → Kontrola jakości danych
- → Graficzne przedstawienie sekwencji
- → Tworzenie raportu
- → Brak możliwości filtracji danych



FASTQC - PRZYKŁADY

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

Documentation

A copy of the FastQC documentation is available for you to try before you buy (well download..).

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

RAPORT W FORMIE GRAFICZNEJ

FASTQC

FASTQC - PRZYKŁADY

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

Documentation

A copy of the FastQC documentation is available for you to try before you buy (well download..).

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

FASTQC - PRZYKŁADY

The main functions of FastQC are

- Import of data from BAM, SAM
- Providing a quick overview to t
- Summary graphs and tables to
- Export of results to an HTML b
- Offline operation to allow auto

Documentation

A copy of the FastQC documentation i

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated ru
- Small RNA with read-through a
- Reduced Representation BS-S
- PacBio
- 454

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

www.bioinformatics.babraham.ac.uk/projects/fastqc/

BASIC STATISTICS

Basic Statistics

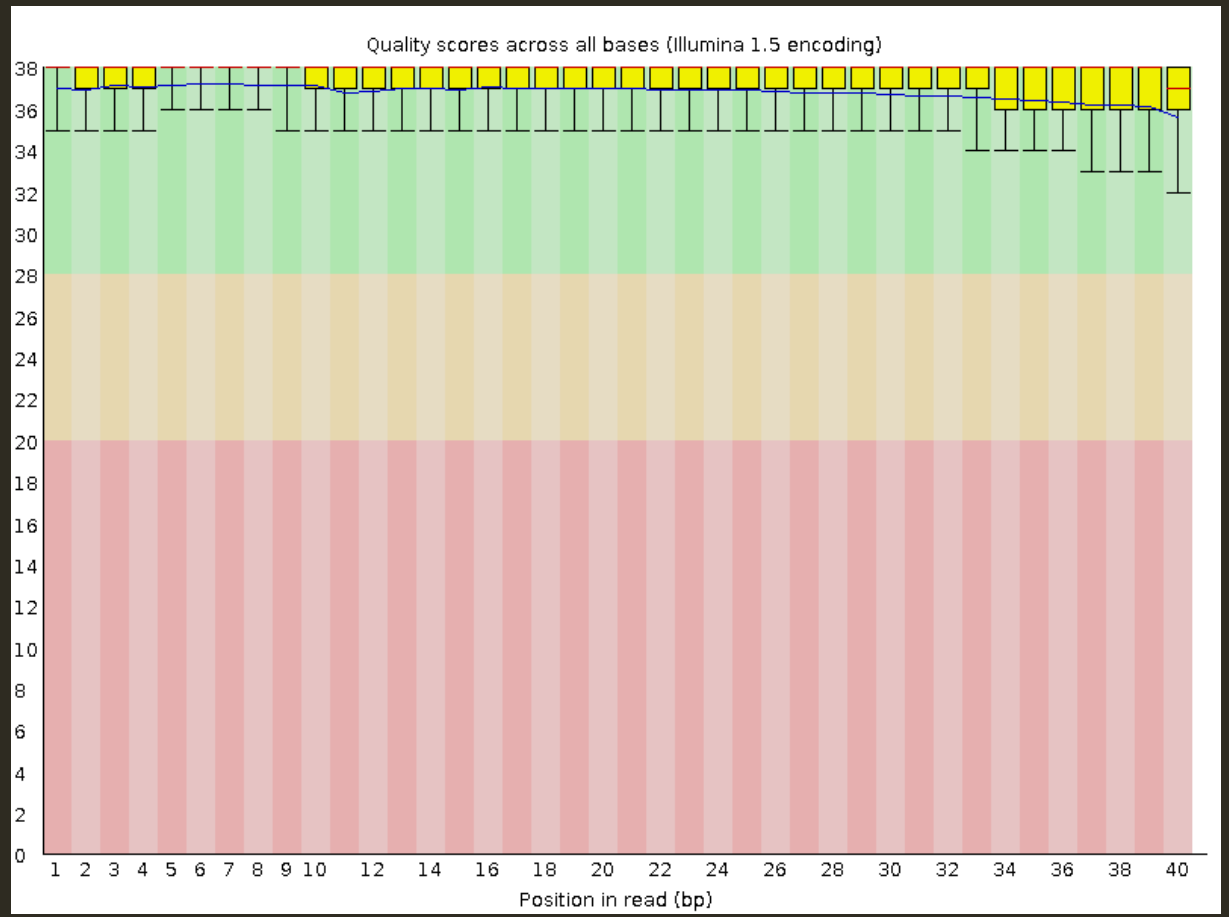
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

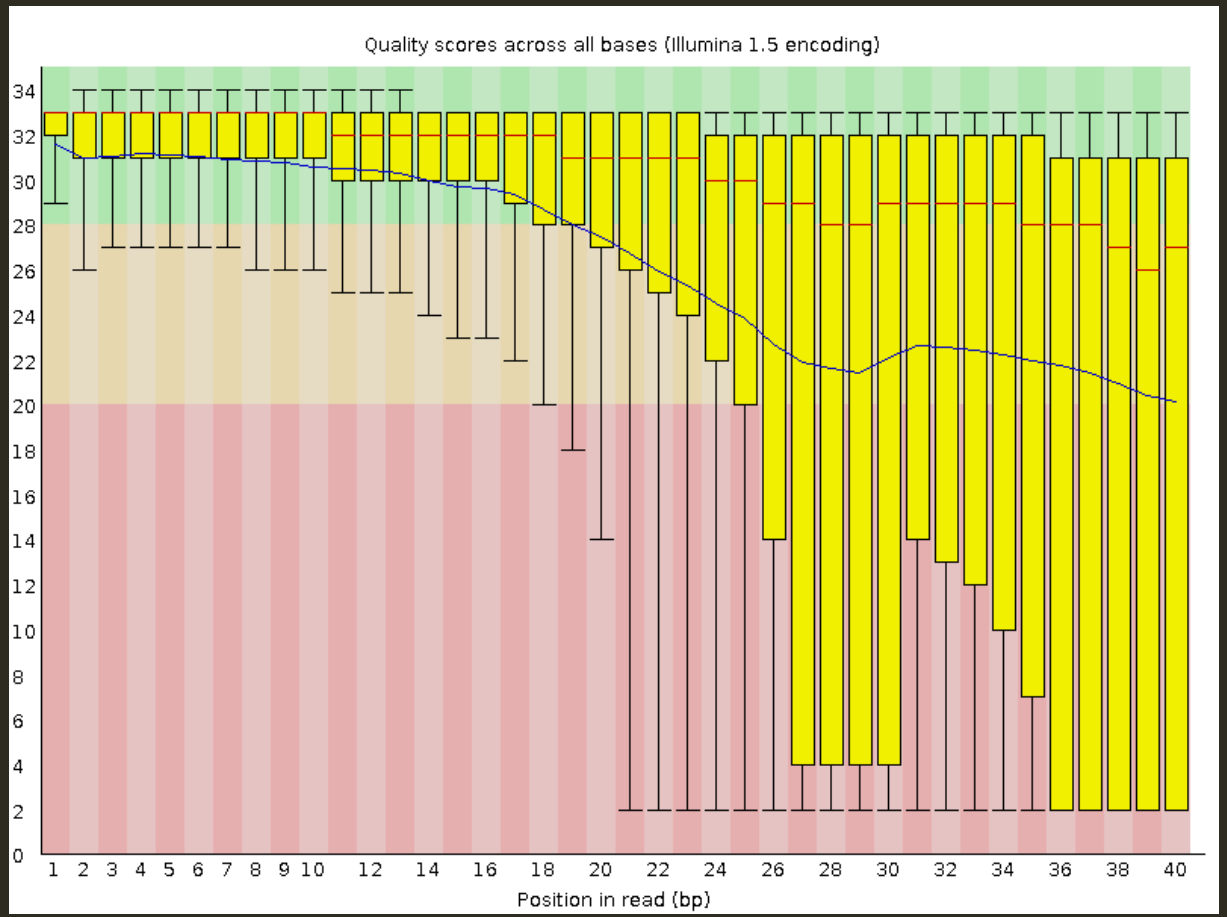
Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

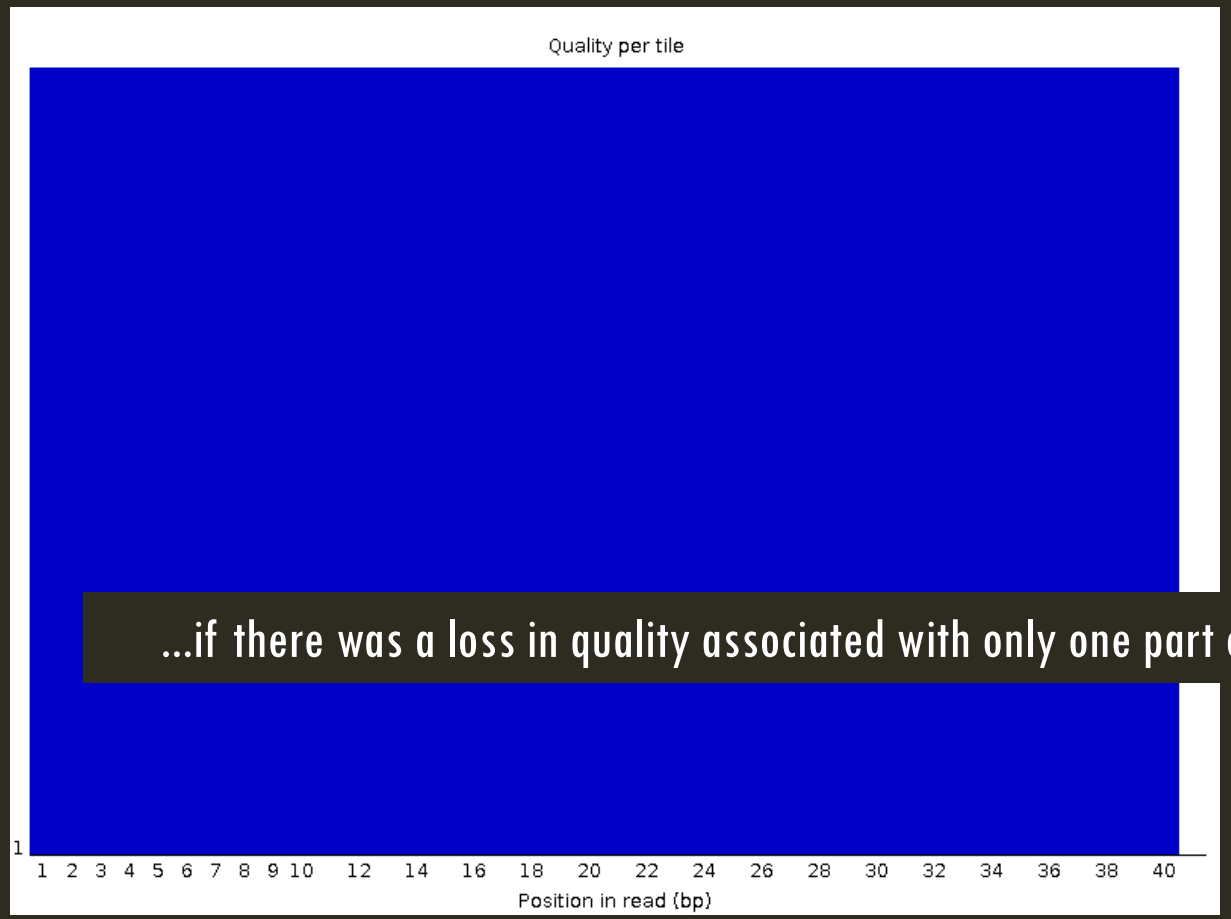
www.bioinformatics.babraham.ac.uk/projects/fastqc/

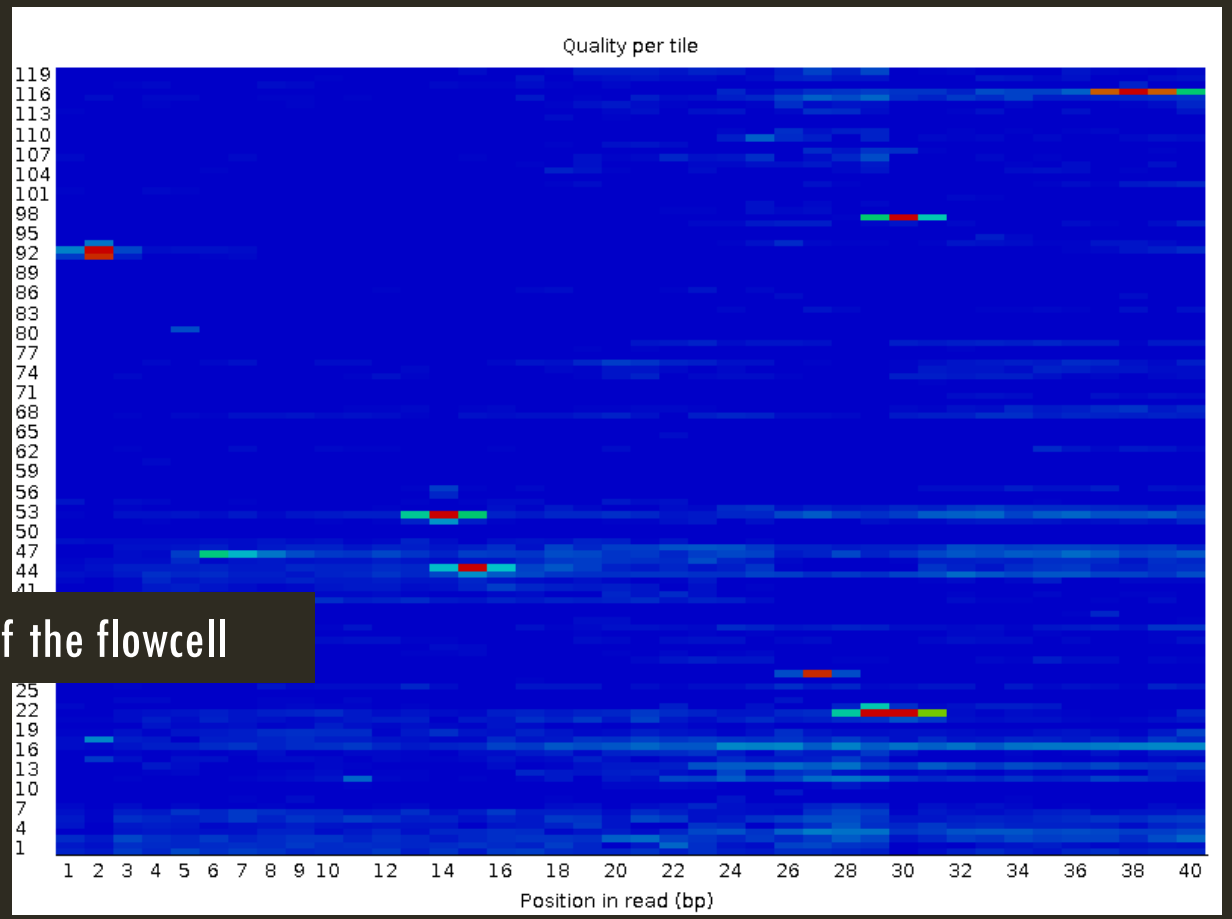
PER BASE SEQUENCE QUALITY





PER TILE SEQUENCE QUALITY

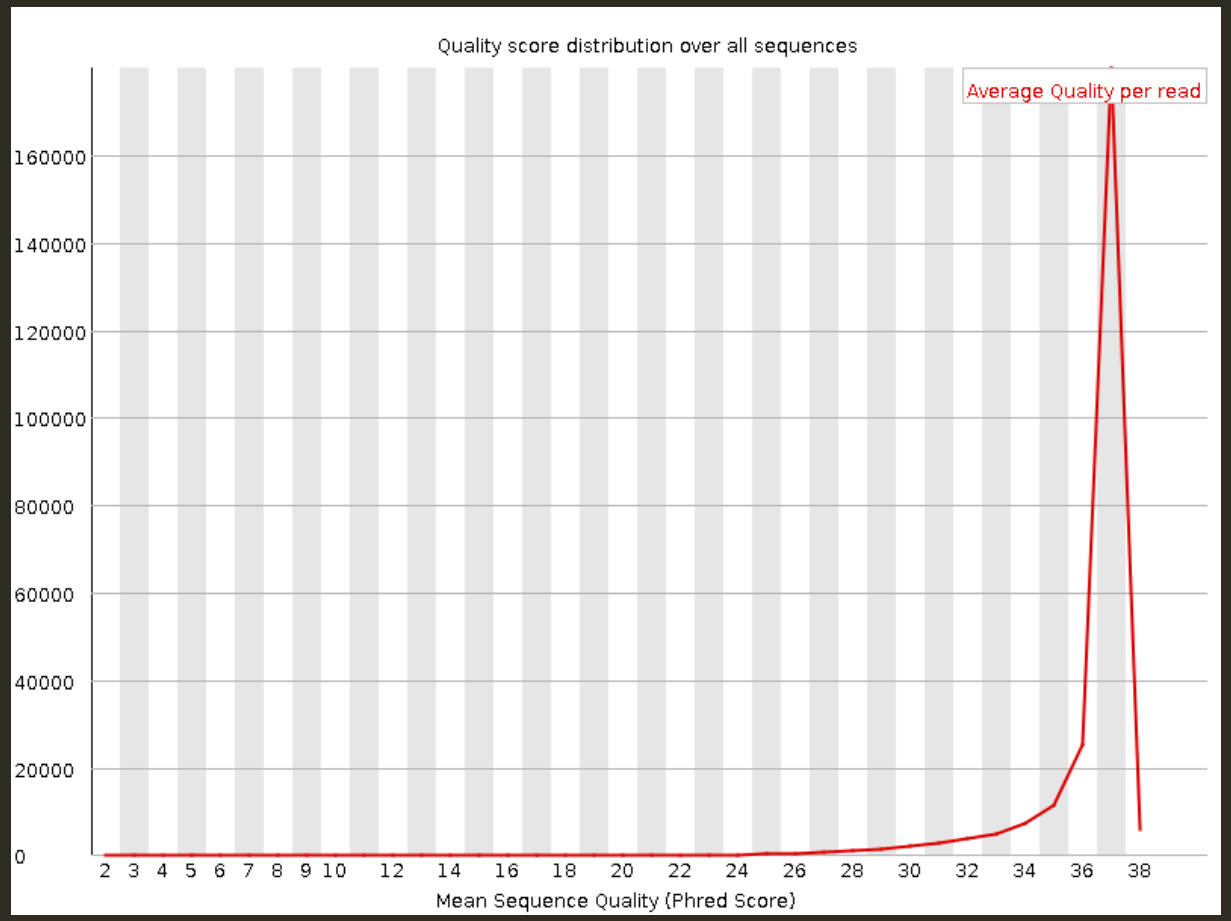


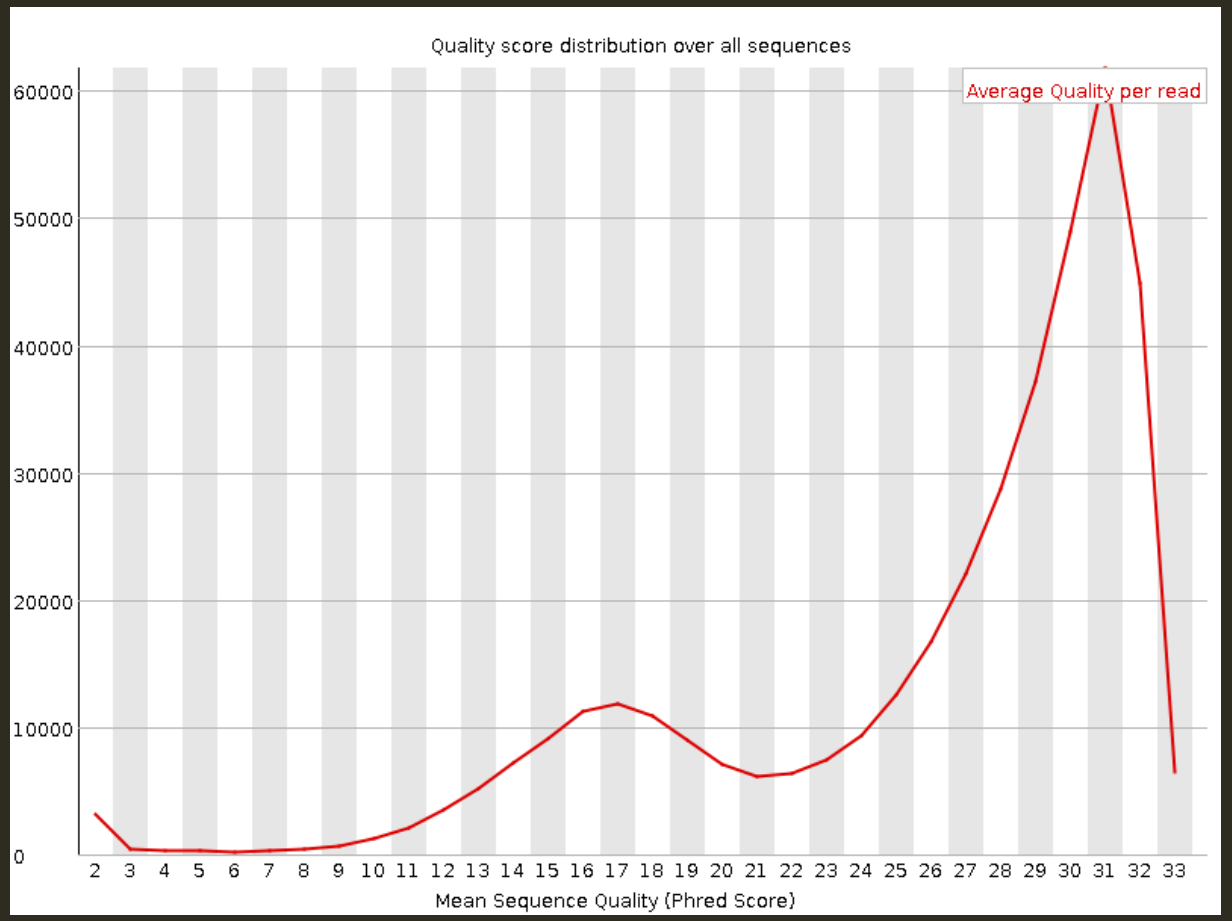


...if there was a loss in quality associated with only one part of the flowcell

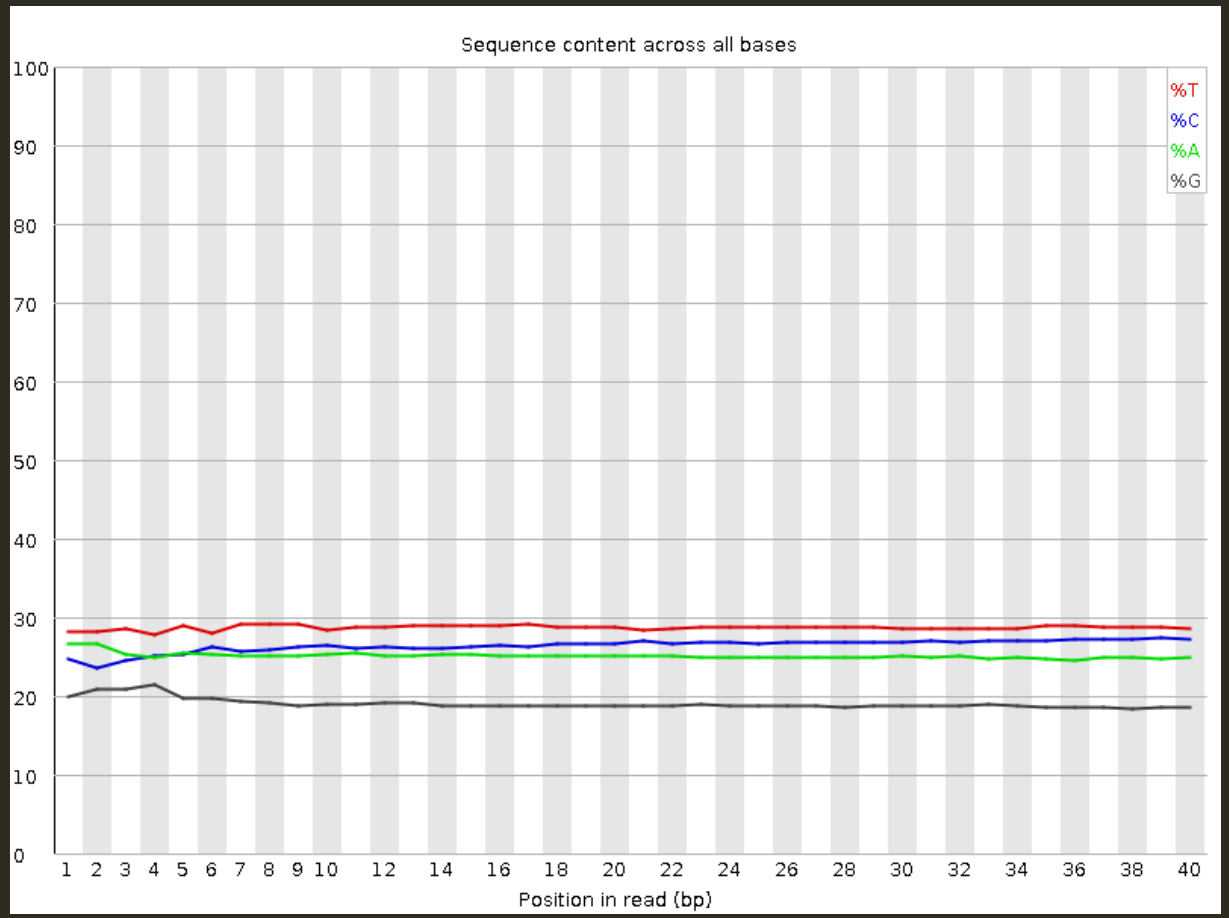
www.bioinformatics.babraham.ac.uk/projects/fastqc/

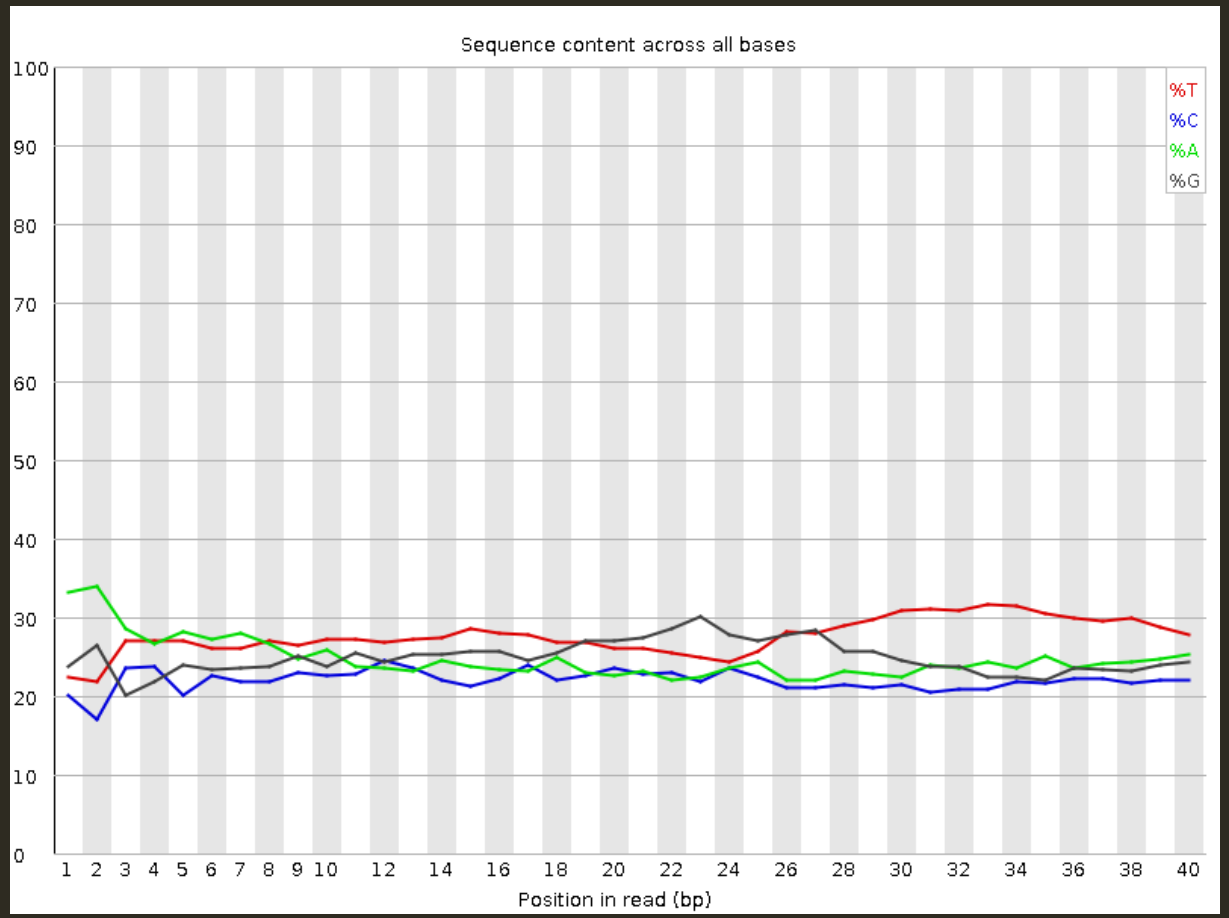
PER SEQUENCE QUALITY SCORES



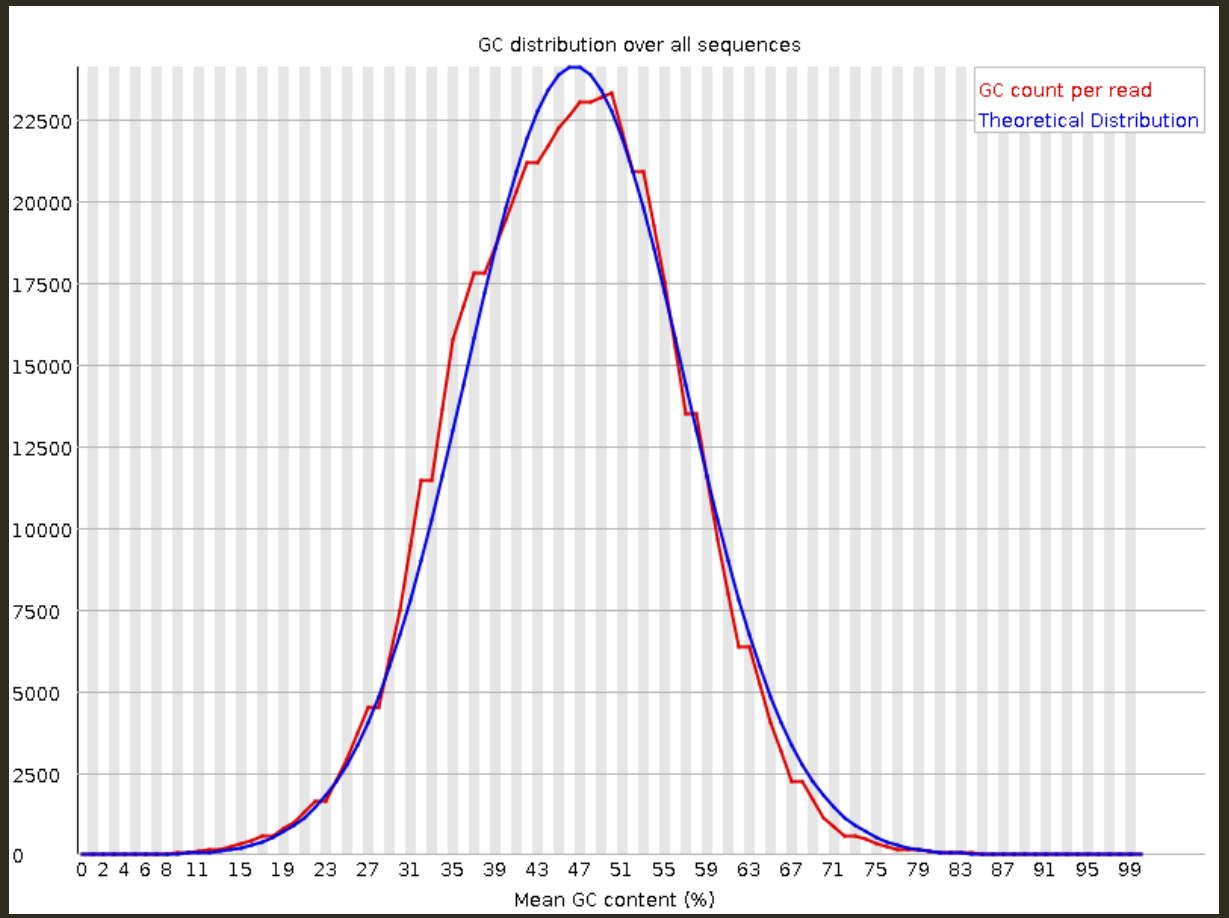


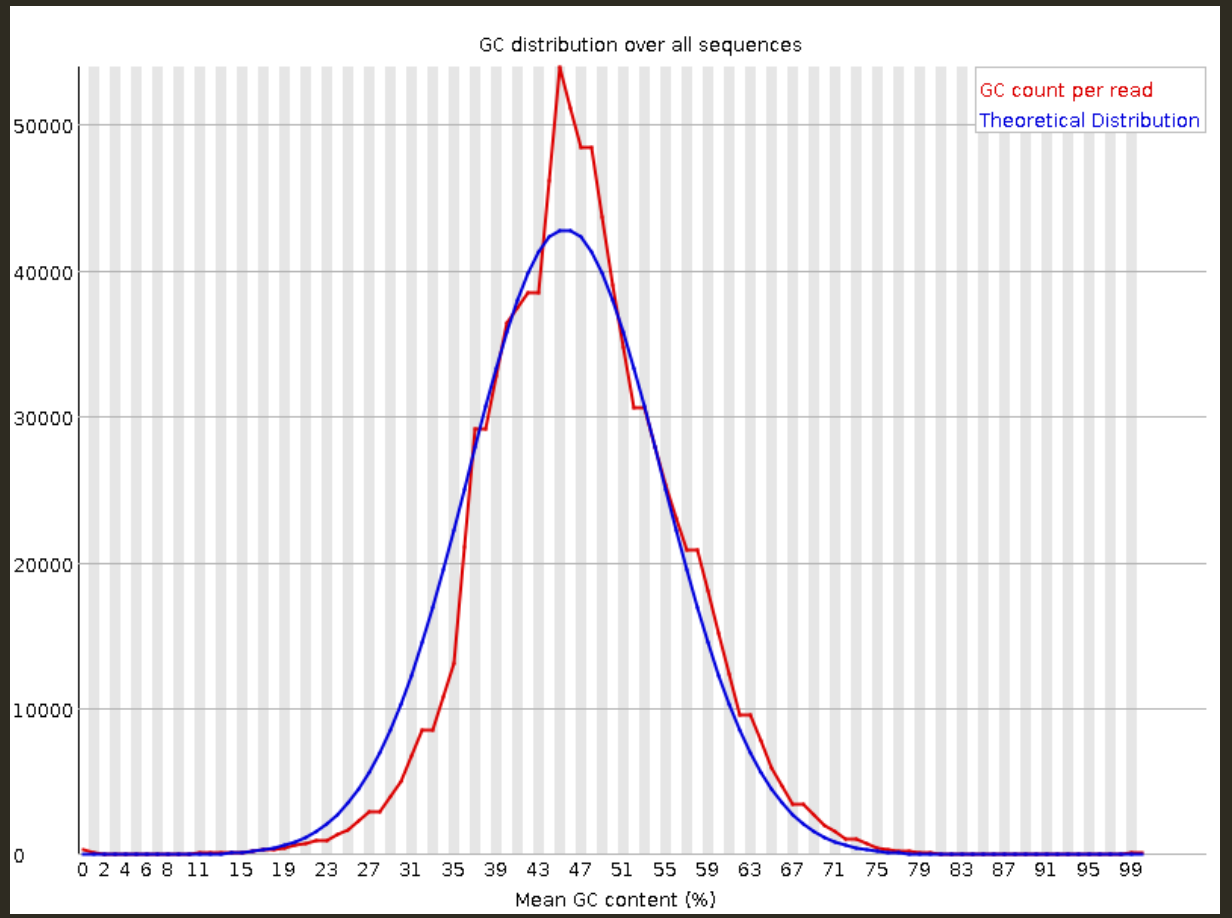
PER BASE SEQUENCE CONTENT





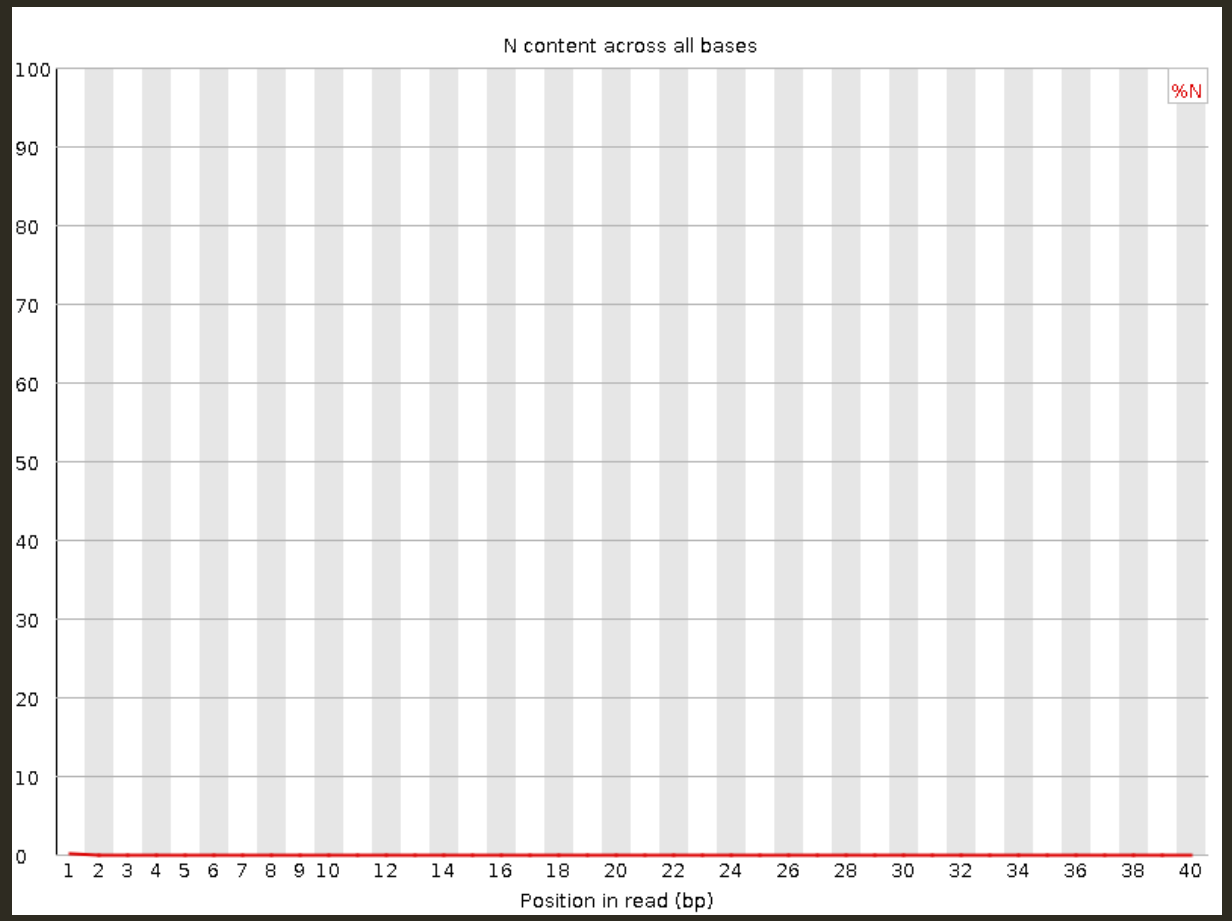
PER SEQUENCE GC CONTENT

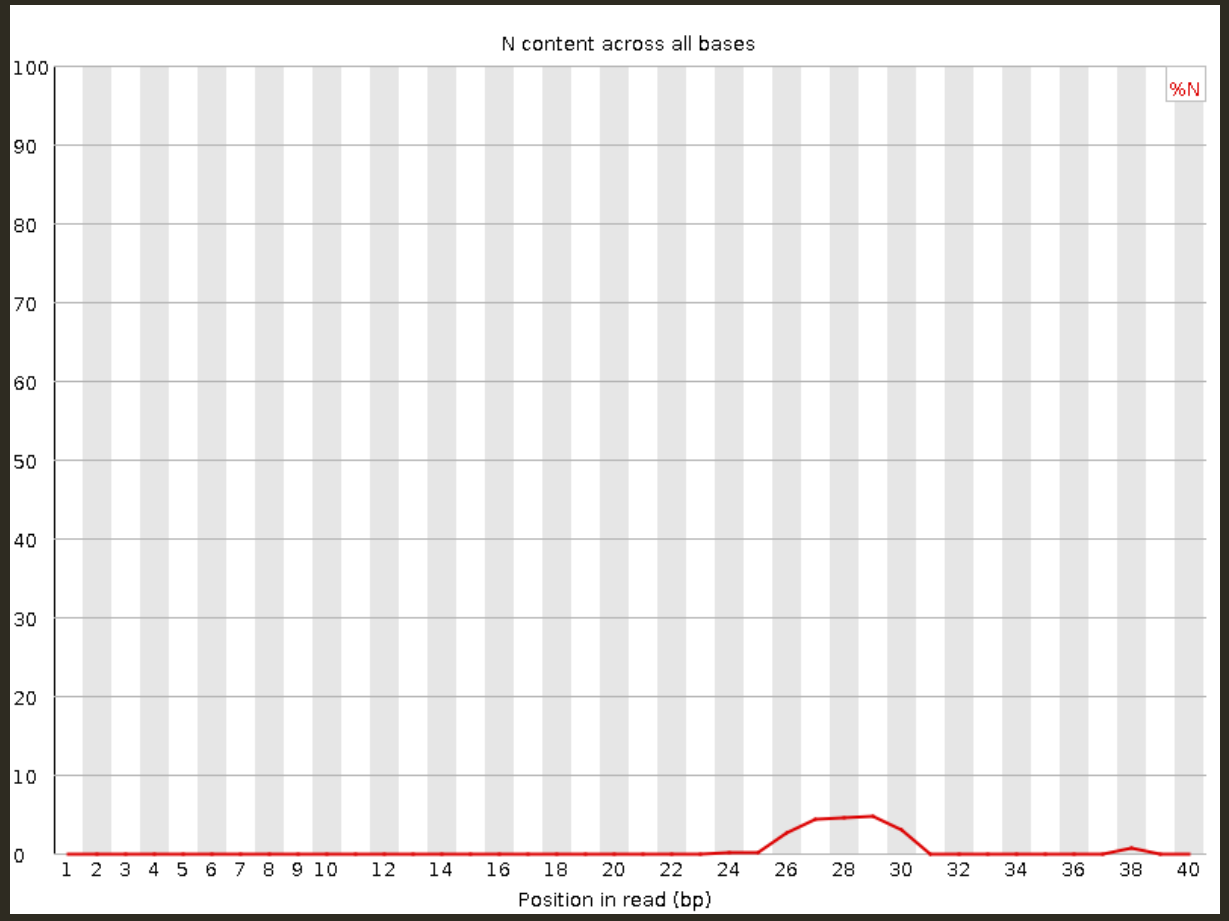




www.bioinformatics.babraham.ac.uk/projects/fastqc/

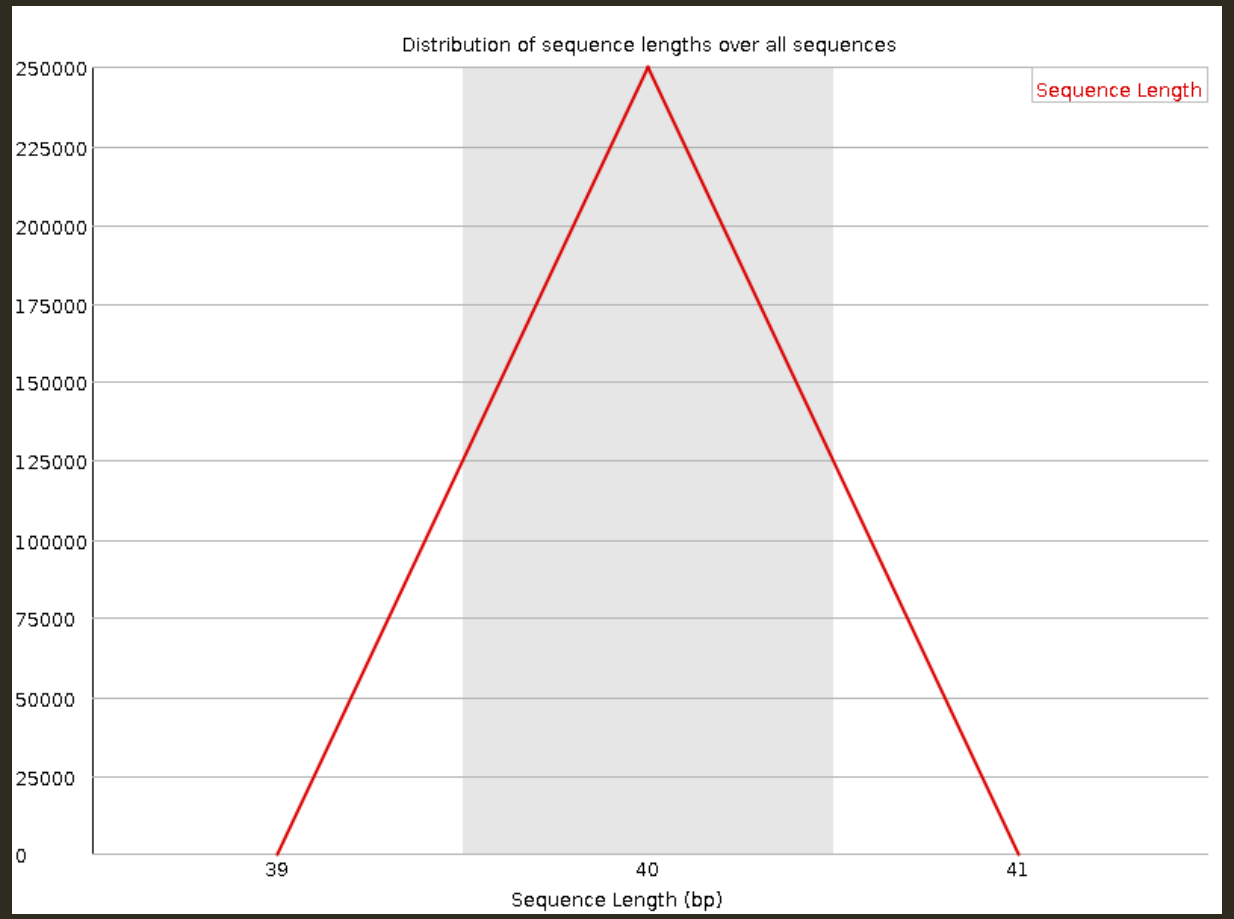
PER BASE N CONTENT

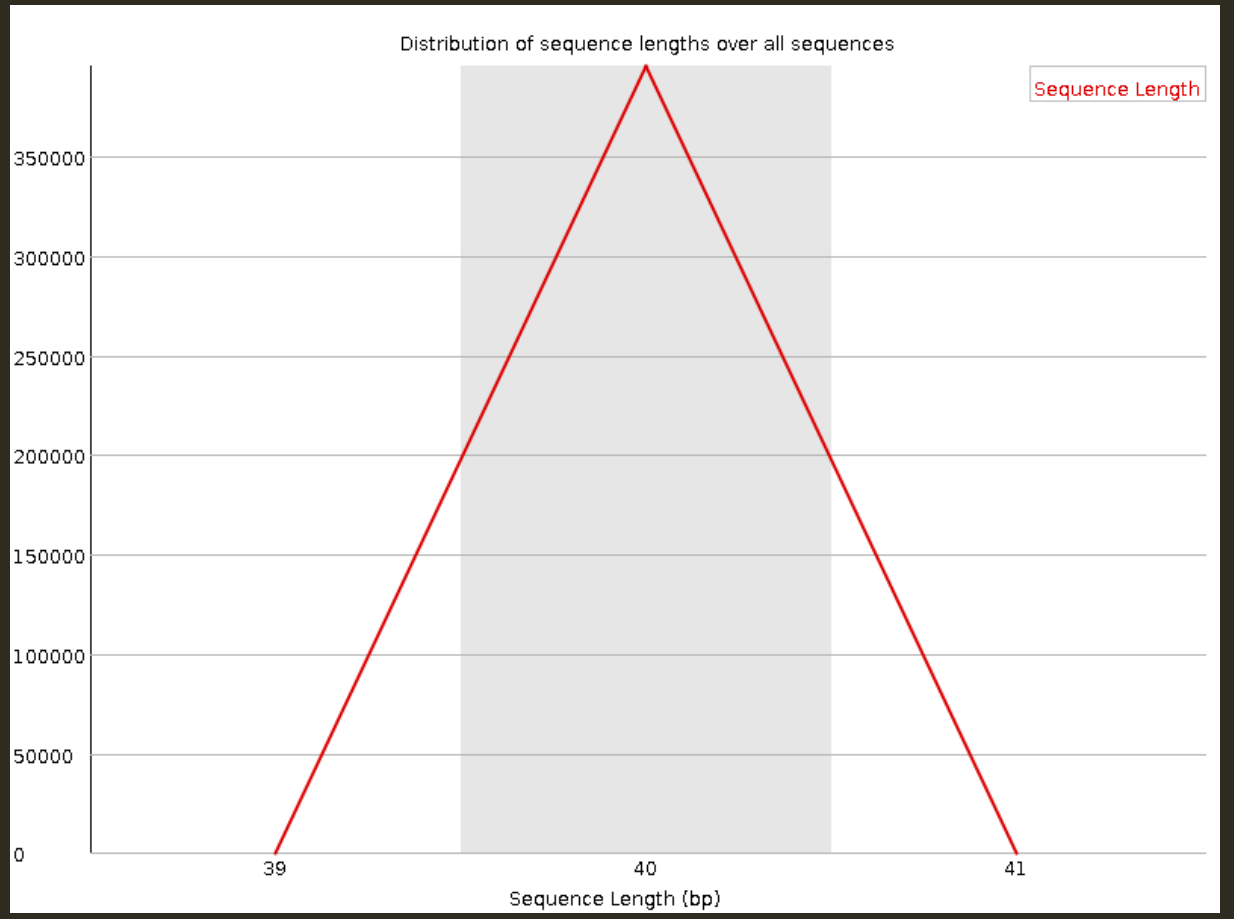




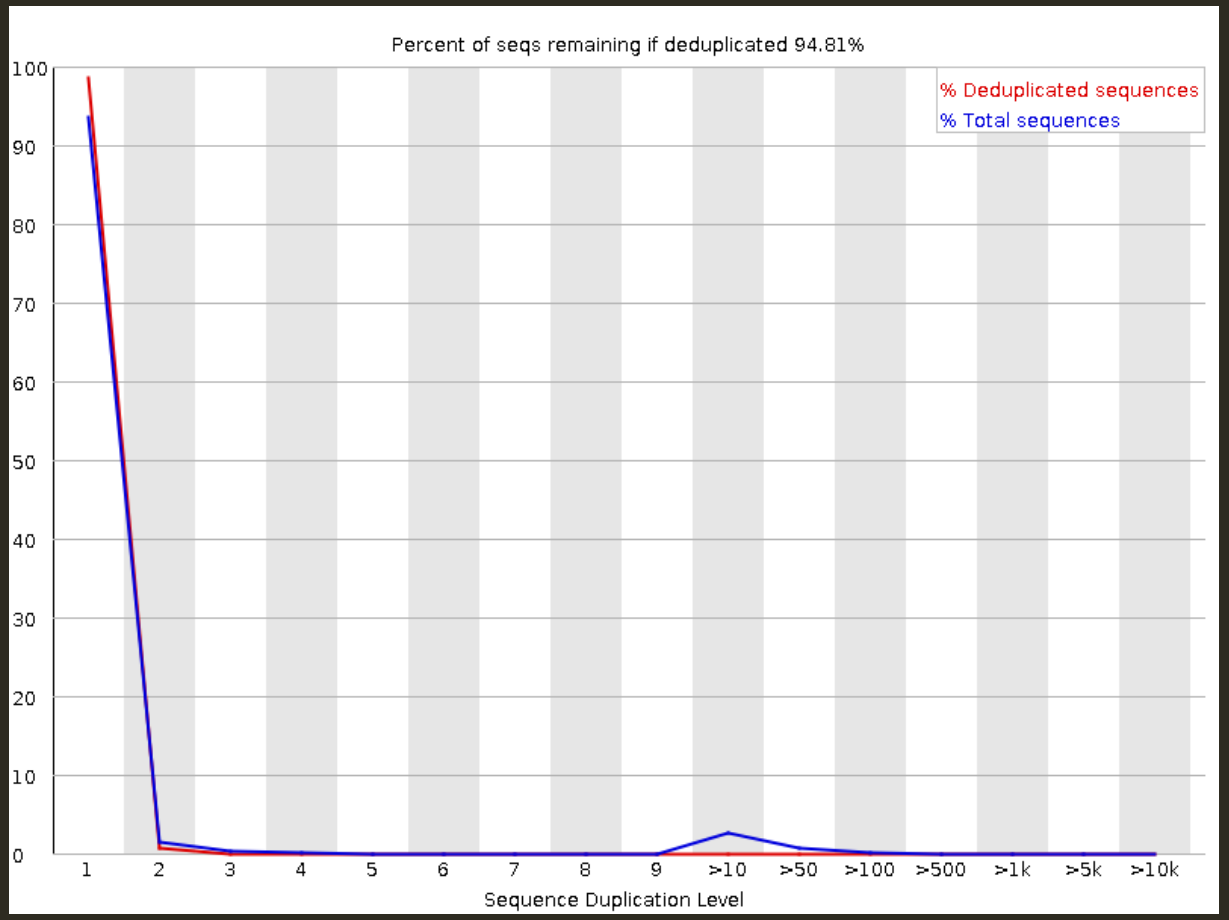
www.bioinformatics.babraham.ac.uk/projects/fastqc/

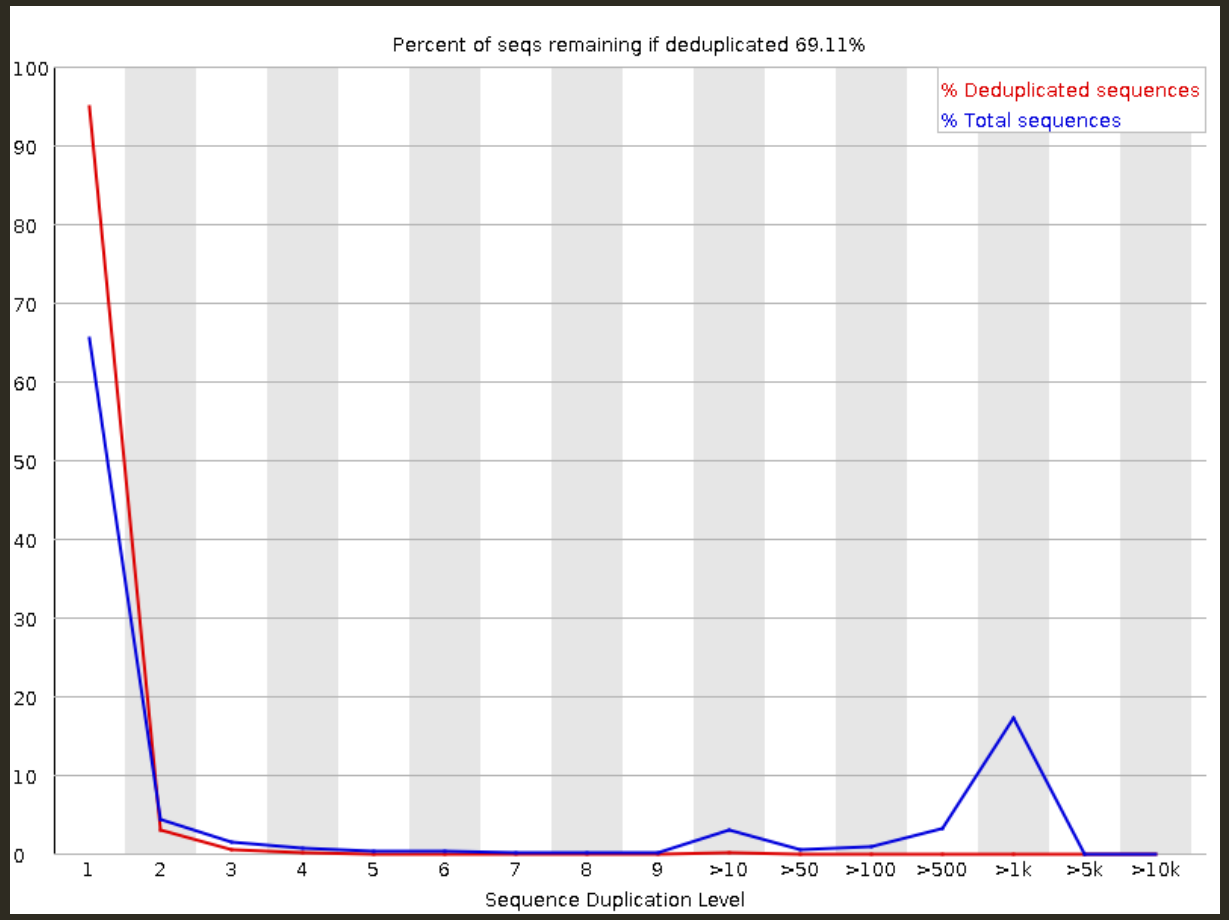
SEQUENCE LENGTH DISTRIBUTION





SEQUENCE DUPLICATION LEVELS

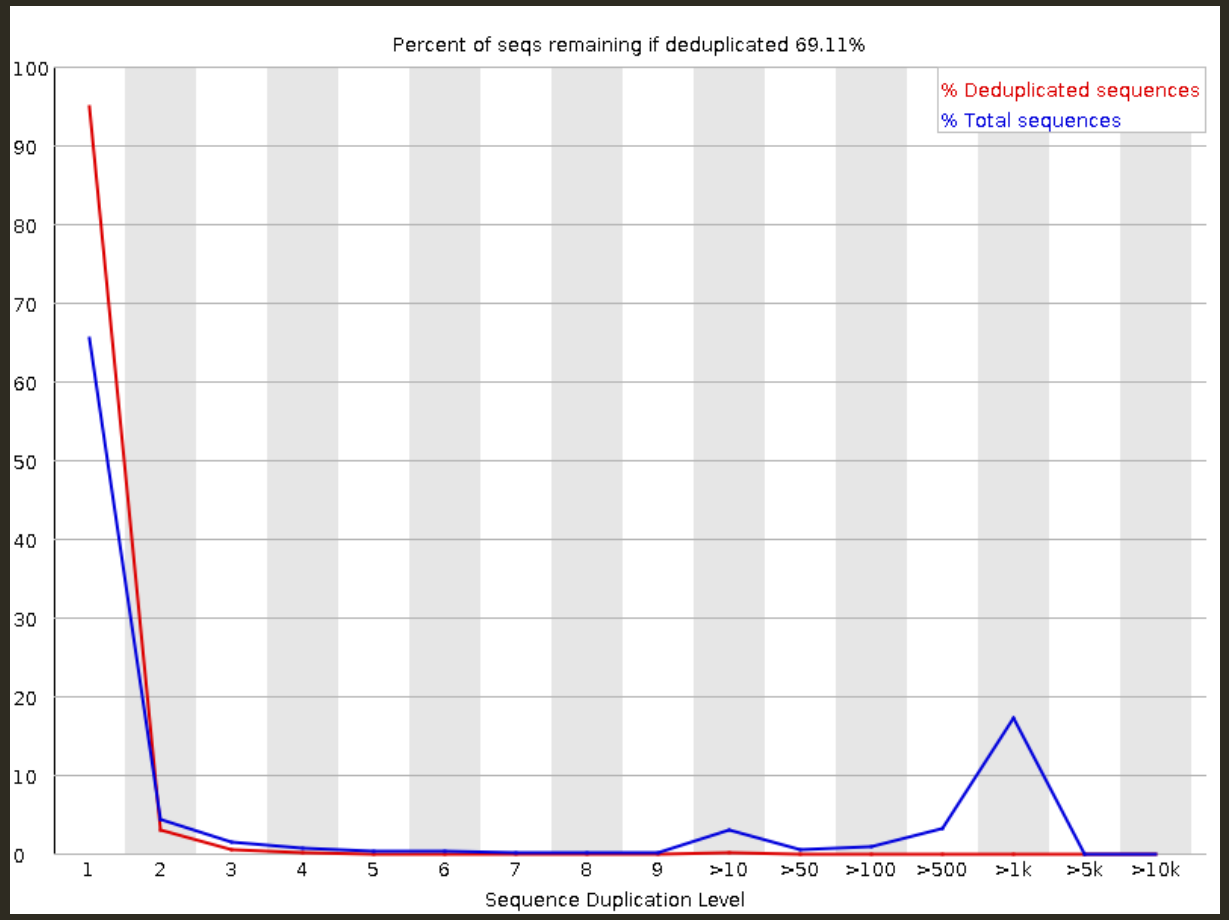




SEQUENCE DUPLICATION LEVELS

To cut down on the memory requirements for this module only sequences which first appear in the first 100,000 sequences in each file are analysed, but this should be enough to get a good impression for the duplication levels in the whole file.

Because the duplication detection requires an exact sequence match over the whole length of the sequence, any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to underrepresent highly duplicated sequences.



OVERREPRESENTED SEQUENCES

Ostrzeżenie – którakolwiek sekwencja występuje > 0,1%

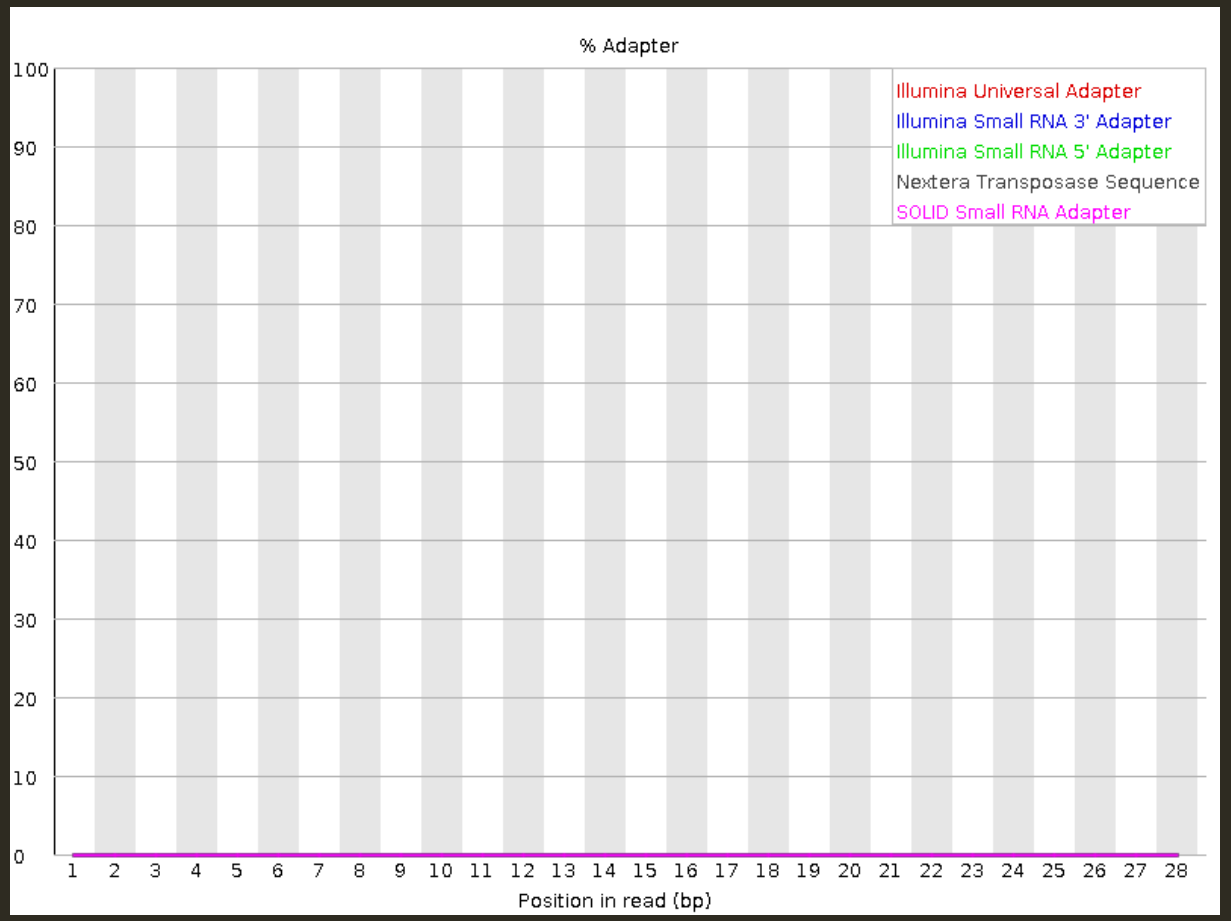
Awaria – którakolwiek sekwencja występuje > 1%

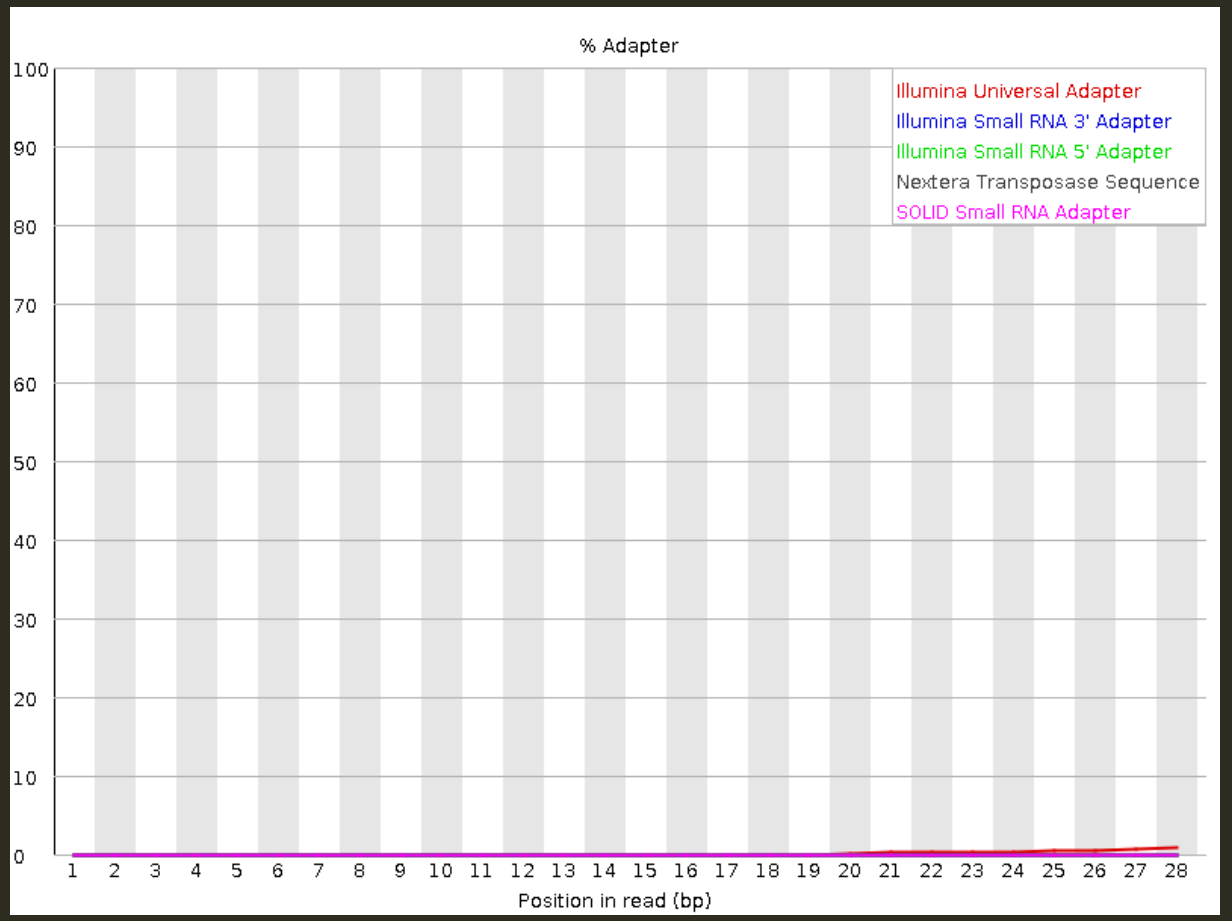
No overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGC	532	0.1345854162028698	No Hit
CTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCGC	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTTG	411	0.10397482341988626	No Hit

www.bioinformatics.babraham.ac.uk/projects/fastqc/

ADAPTER CONTENT





FASTQC - PRZYKŁADY

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive appli

Documentation

A copy of the FastQC documentation is available for you to try before you buy (well download..).

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

FASTQC - PRZYKŁADY

The main functions of FastQC are

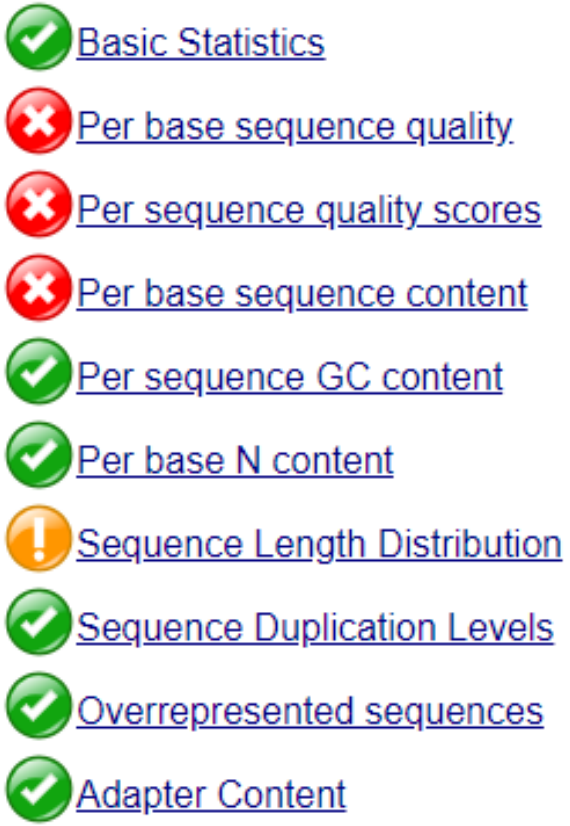
- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interact

Documentation

A copy of the FastQC documentation is available for you to try before you buy (well dov

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

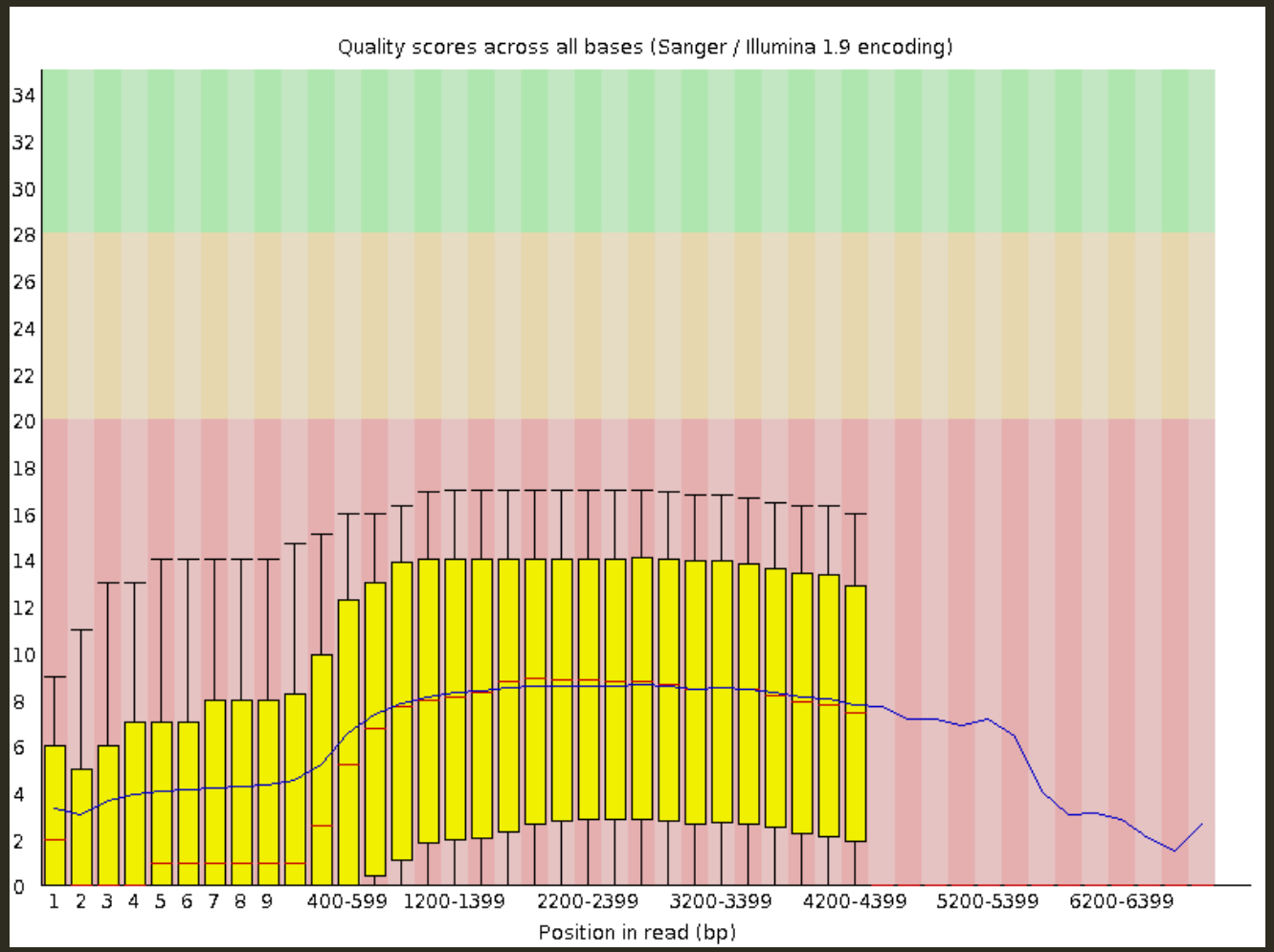


BASIC STATISTICS

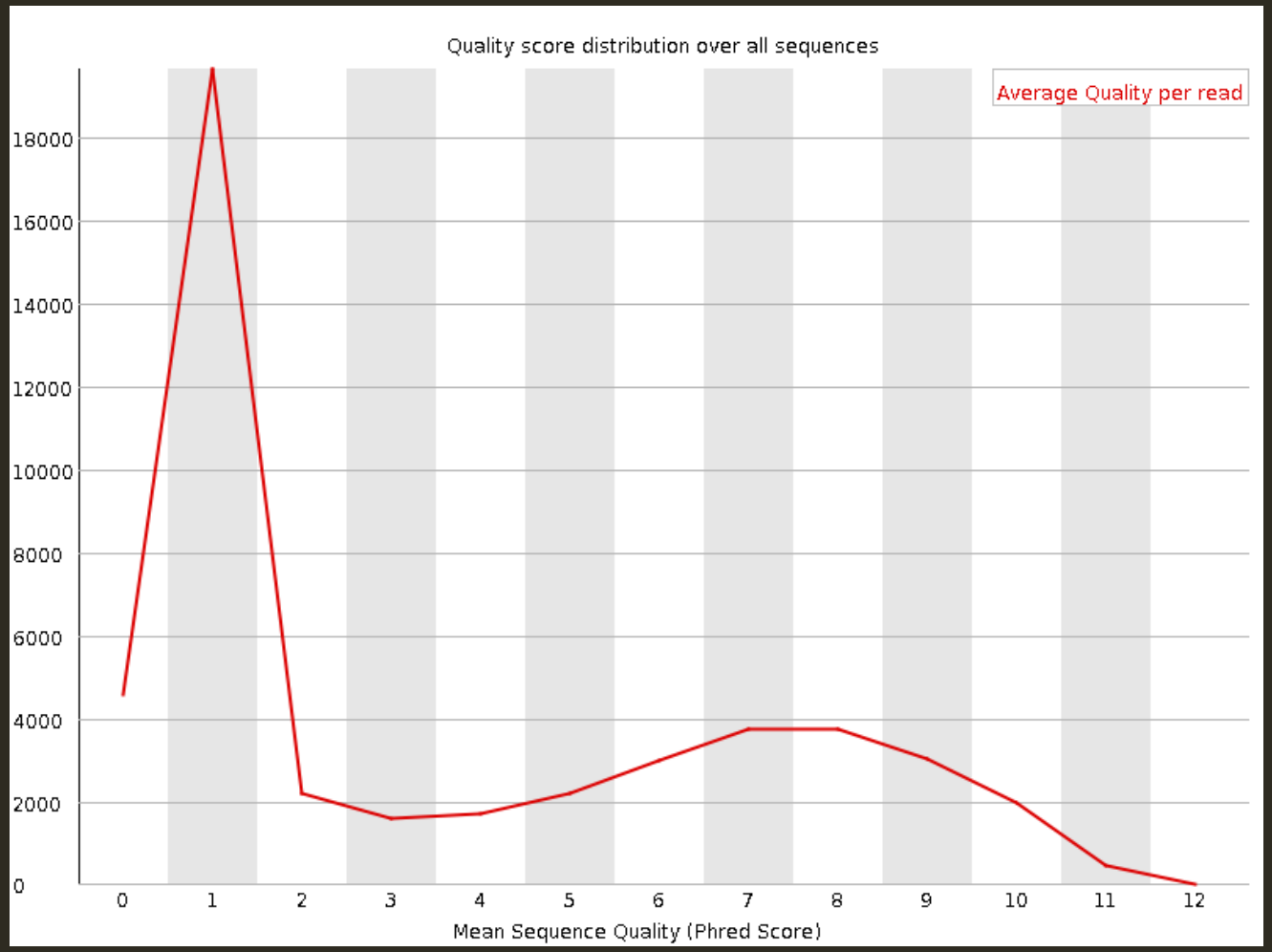
Basic Statistics

Measure	Value
Filename	pacbio_srr075104.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	48053
Sequences flagged as poor quality	0
Sequence length	82-6919
%GC	52

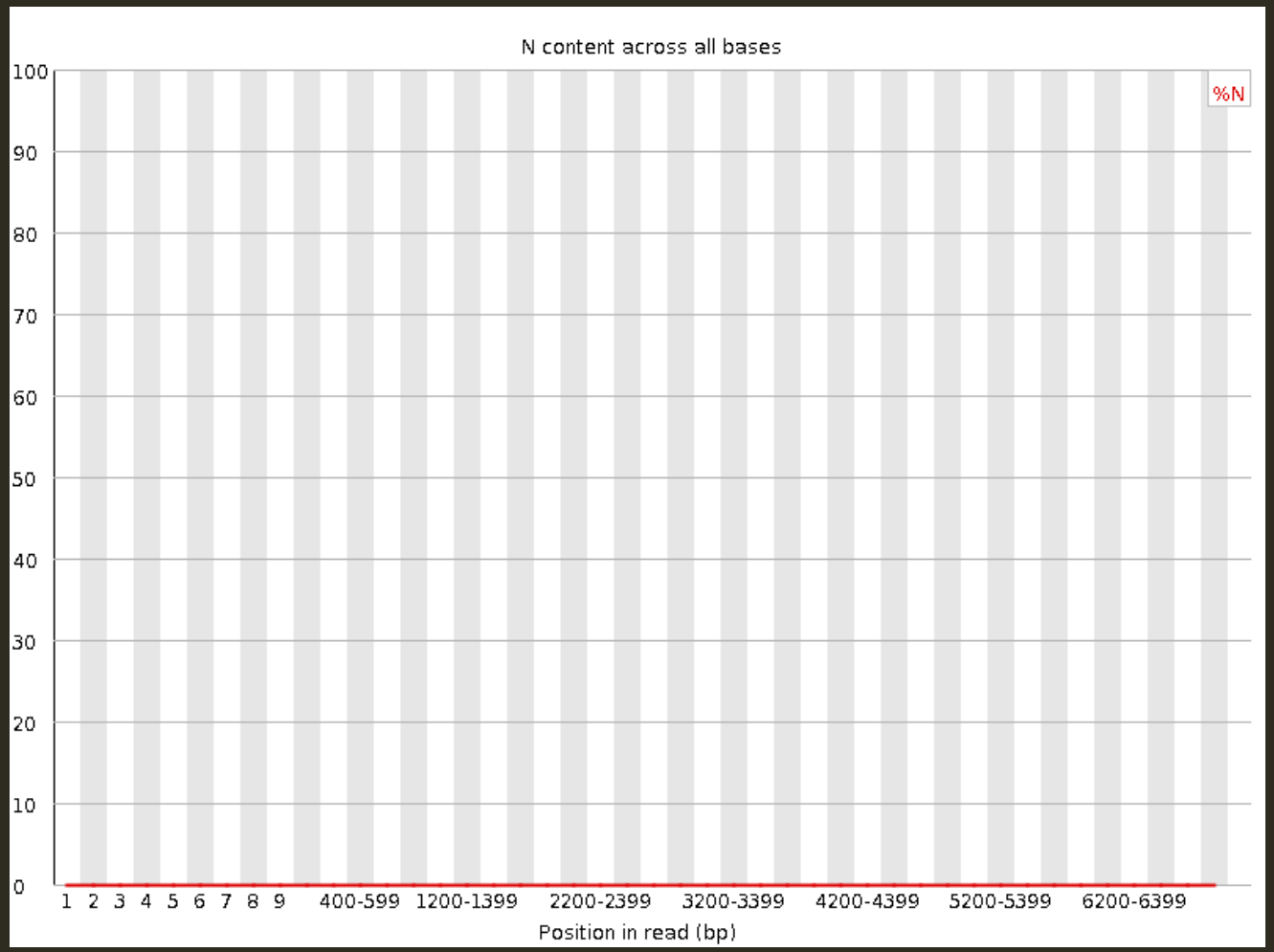
PER BASE SEQUENCE QUALITY



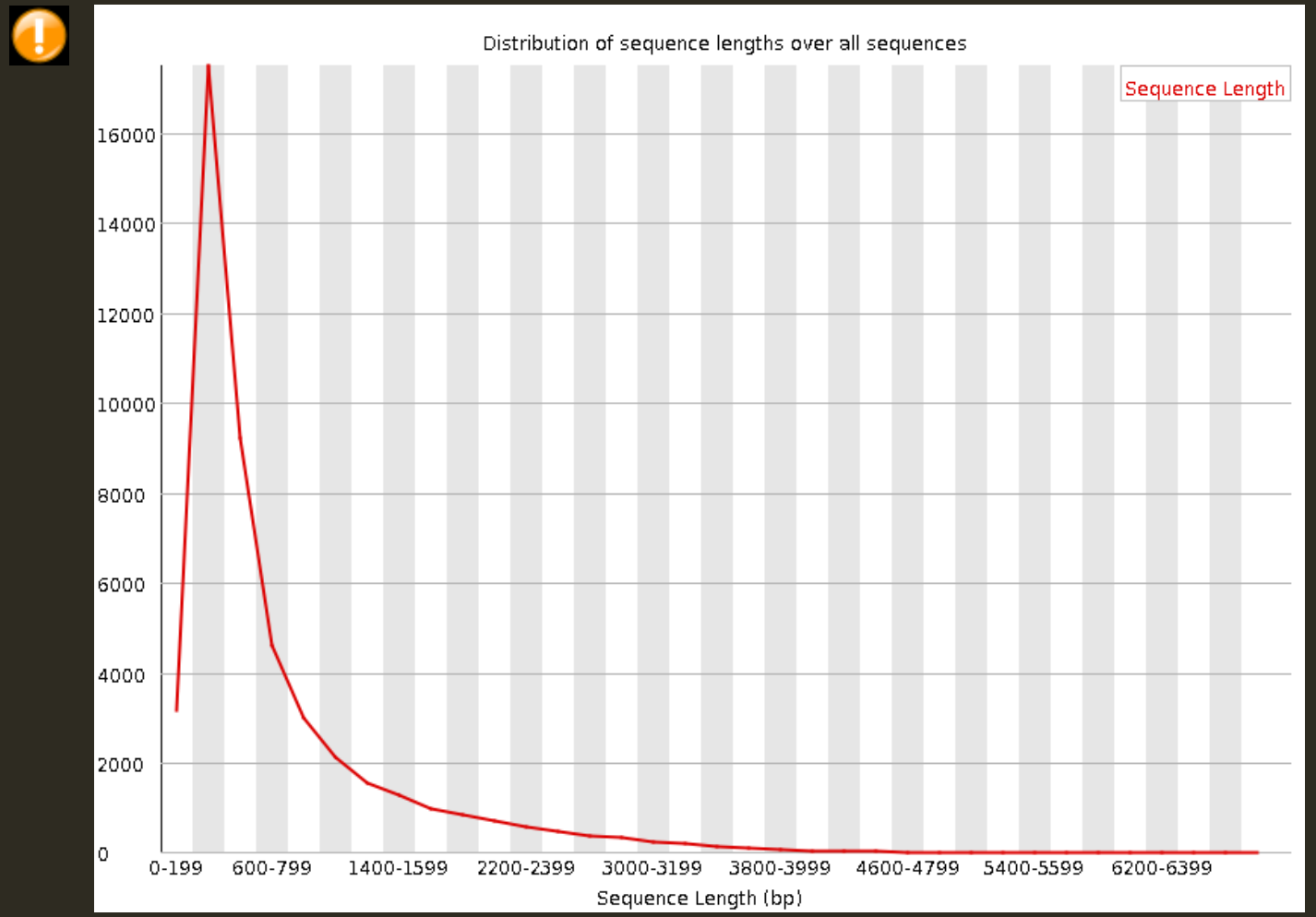
PER SEQUENCE QUALITY SCORES



PER BASE N CONTENT



SEQUENCE LENGTH DISTRIBUTION



FASTQC - PRZYKŁADY

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

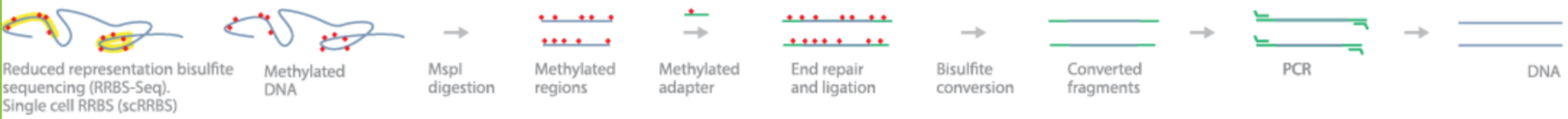
Documentation

A copy of the FastQC documentation is available for you to try before you buy (well download..).

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

RRBS-Seq/scRRBS



Reduced-representation bisulfite sequencing (RRBS-Seq) or single-cell reduced-representation bisulfite sequencing (scRRBS) is a protocol that uses one or multiple restriction enzymes on the genomic DNA to produce sequence-specific fragmentation. The fragmented genomic DNA is then treated with bisulfite and sequenced. This is the method of choice to study specific regions of interest. It is particularly effective where methylation is high, such as in promoters and repeat regions.

- Genome-wide coverage of CpGs in islands at single-base resolution
- Areas dense in CpG methylation are covered

Cons:

- Restriction enzymes cut at specific sites, providing biased sequence selection
- Method measures 10-15% of all CpGs in genome
- Cannot distinguish between 5mC and 5hmC
- Does not cover non-CpG areas, genome-wide CpGs, and CpGs in areas without the enzyme restriction site

CpG islands are short stretches of DNA with an unusually high GC content and a higher frequency of CpG dinucleotides

FASTQC - PRZYKŁADY

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

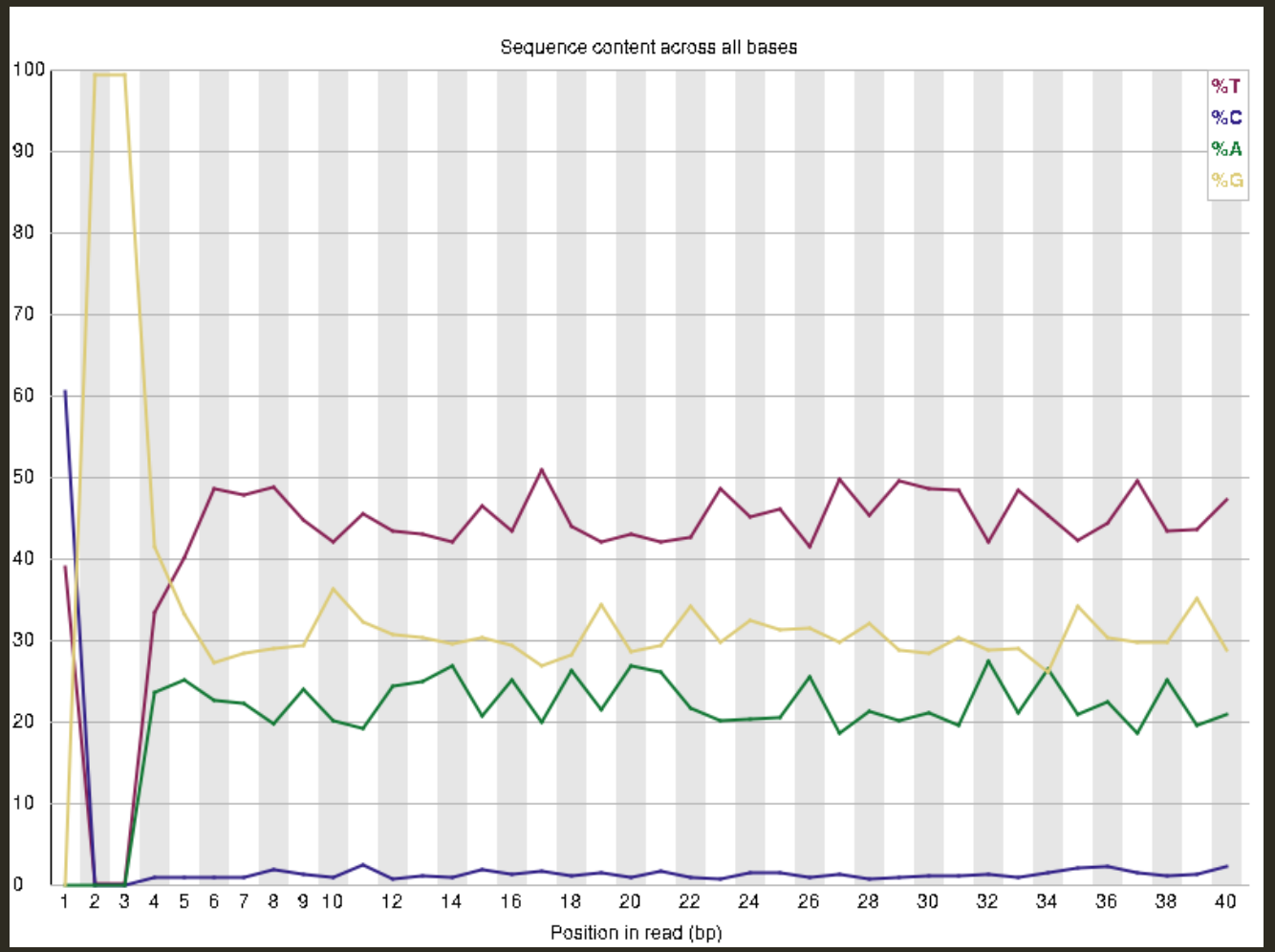
Documentation

A copy of the FastQC documentation is available for you to try before you buy (well download..).

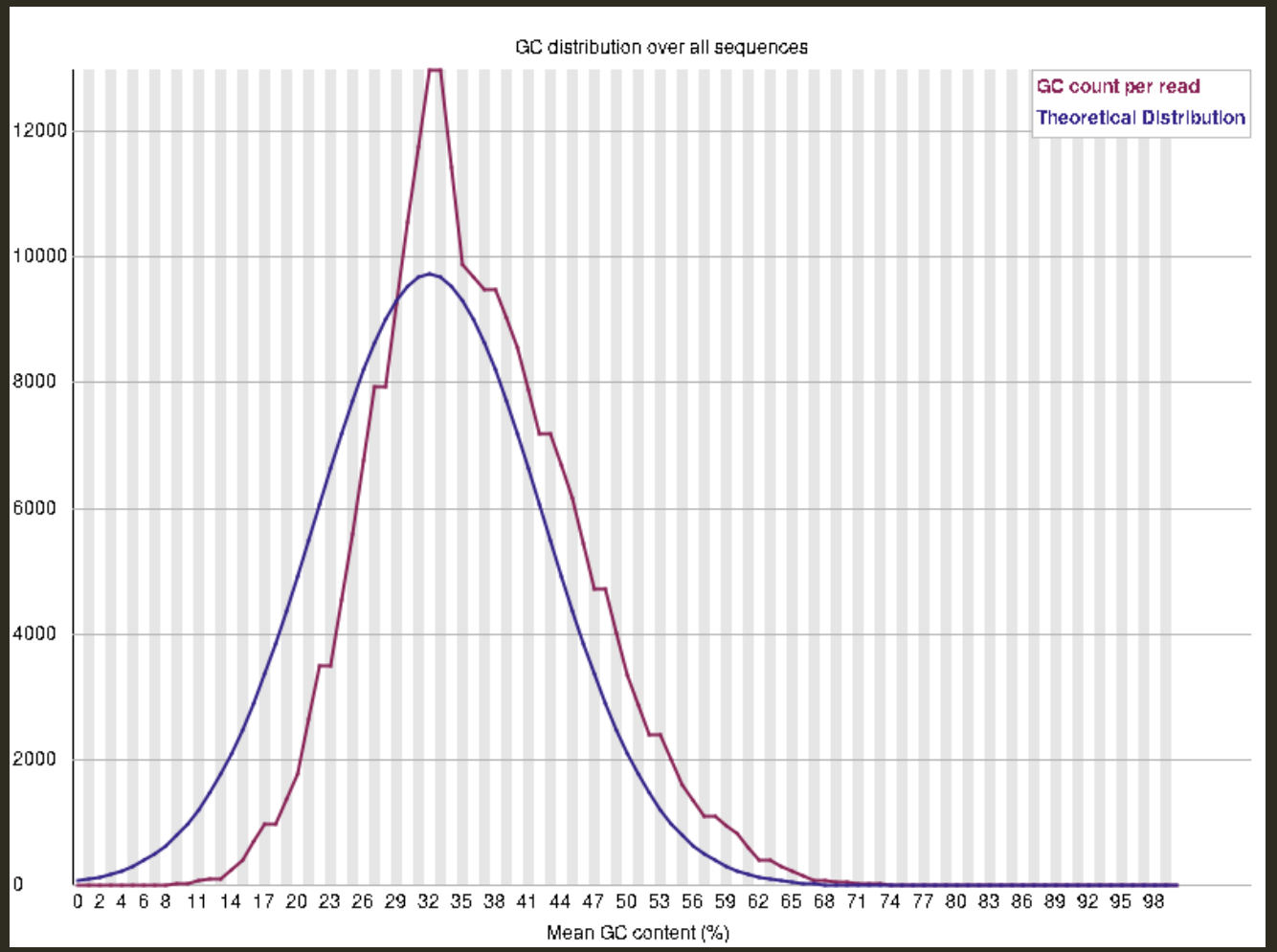
Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

PER BASE SEQUENCE CONTENT



PER SEQUENCE GC CONTENT



RAPORT W FORMIE TEKSTOWEJ

FASTQC

```
drwxr-xr-x 2 mielczarek users 4.0K Nov  8 09:23 Images
drwxr-xr-x 2 mielczarek users 4.0K Nov  8 09:23 Icons
-rw-r--r-- 1 mielczarek users  818 Nov  8 09:23 summary.txt
-rw-r--r-- 1 mielczarek users 343K Nov  8 09:23 fastqc_report.html
-rw-r--r-- 1 mielczarek users  11K Nov  8 09:23 fastqc.fo
-rw-r--r-- 1 mielczarek users 375K Nov  8 09:23 fastqc_data.txt
```

```
PASS    Basic Statistics        NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Per base sequence quality       NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Per tile sequence quality       NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Per sequence quality scores     NG-22253_E3_lib357991_6555_1_2.fastq.gz
FAIL    Per base sequence content       NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Per sequence GC content NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Per base N content      NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Sequence Length Distribution    NG-22253_E3_lib357991_6555_1_2.fastq.gz
FAIL    Sequence Duplication Levels     NG-22253_E3_lib357991_6555_1_2.fastq.gz
WARN    Overrepresented sequences       NG-22253_E3_lib357991_6555_1_2.fastq.gz
PASS    Adapter Content NG-22253_E3_lib357991_6555_1_2.fastq.gz
```

```
drwxr-xr-x 2 mielczarek users 4.0K Nov  8 09:23 Images
drwxr-xr-x 2 mielczarek users 4.0K Nov  8 09:23 Icons
-rw-r--r-- 1 mielczarek users  818 Nov  8 09:23 summary.txt
-rw-r--r-- 1 mielczarek users 343K Nov  8 09:23 fastqc_report.html
-rw-r--r-- 1 mielczarek users  11K Nov  8 09:23 fastqc.fo
-rw-r--r-- 1 mielczarek users 375K Nov  8 09:23 fastqc data.txt
```

```
##FastQC        0.11.4
>>Basic Statistics      pass
#Measure        Value
Filename        NG-22253_E3_lib357991_6555_1_2.fastq.gz
File type       Conventional base calls
Encoding        Sanger / Illumina 1.9
Total Sequences 19743654
Sequences flagged as poor quality       0
Sequence length 151
%GC     46
>>END_MODULE
>>Per base sequence quality      pass
#Base   Mean    Median  Lower Quartile  Upper Quartile  10th Percentile 90th Percentile
1       35.83468860424722       37.0    37.0    37.0    37.0    37.0
2       36.005622464818316      37.0    37.0    37.0    37.0    37.0
3       36.102991168706666      37.0    37.0    37.0    37.0    37.0
4       36.01433316244298       37.0    37.0    37.0    37.0    37.0
5       36.07038656572892       37.0    37.0    37.0    37.0    37.0
6       36.02822881721894       37.0    37.0    37.0    37.0    37.0
```

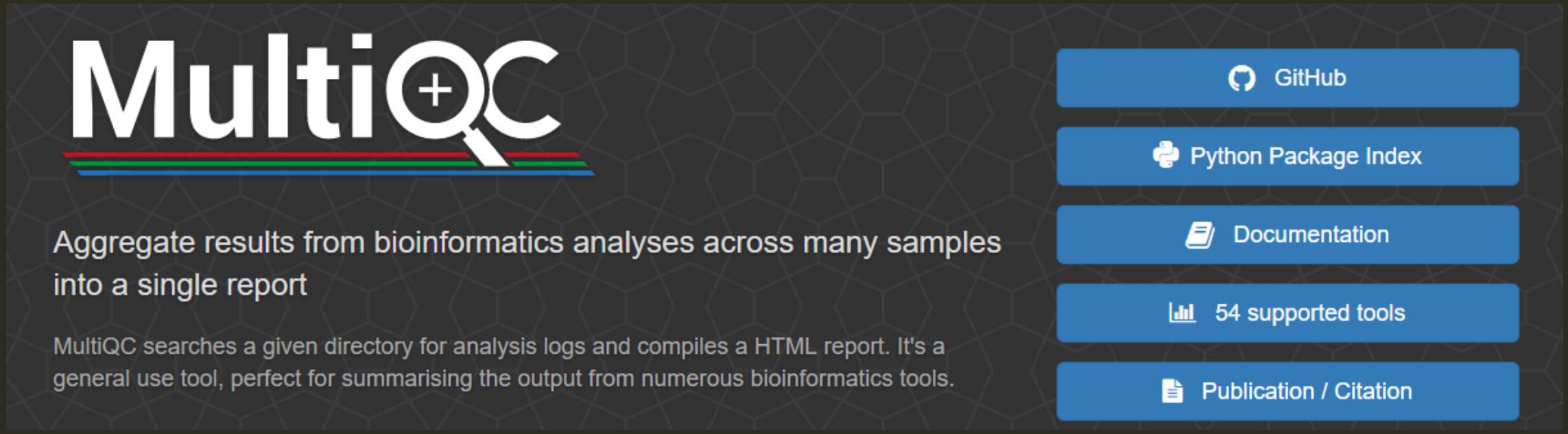
ŁĄCZENIE RAPORTÓW

MultiQC

http://multiqc.info/

MULTIQC

→ 1 RAPORT DLA WIELU PRÓB



YouTube PL

www.youtube.com/watch?time_continue=5&v=BbScv9TcaMg

www.youtube.com/watch?time_continue=1&v=qPbIIO_KWN0

http://multiqc.info/

MULTIQC

→ 1 RAPORT DLA WIELU PRÓB

General Statistics

Copy table | Configure Columns | Showing 16/16 rows and 3/5 columns.

Sample Name
151114_I191_FCH3Y35BCXX_L1_wHAIPI023992-37_1
151114_I191_FCH3Y35BCXX_L1_wHAIPI023992-37_2
151114_I191_FCH3Y35BCXX_L2_wHAMPI023991-66_1
151118_I137_FCH3KNJBBXX_L5_wHAXPI023905-96_2
160103_I137_FCH3V5YBBXX_L3_WHOSTibkDCABDLAAPEI-62_1
160103_I137_FCH3V5YBBXX_L3_WHOSTibkDCABDLAAPEI-62_2
160103_I137_FCH3V5YBBXX_L3_WHOSTibkDCACDTAAPEI-75_1
160103_I137_FCH3V5YBBXX_L3_WHOSTibkDCACDTAAPEI-75_2
160103_I137_FCH3V5YBBXX_L4_WHOSTibkDCABDLAAPEI-62_1

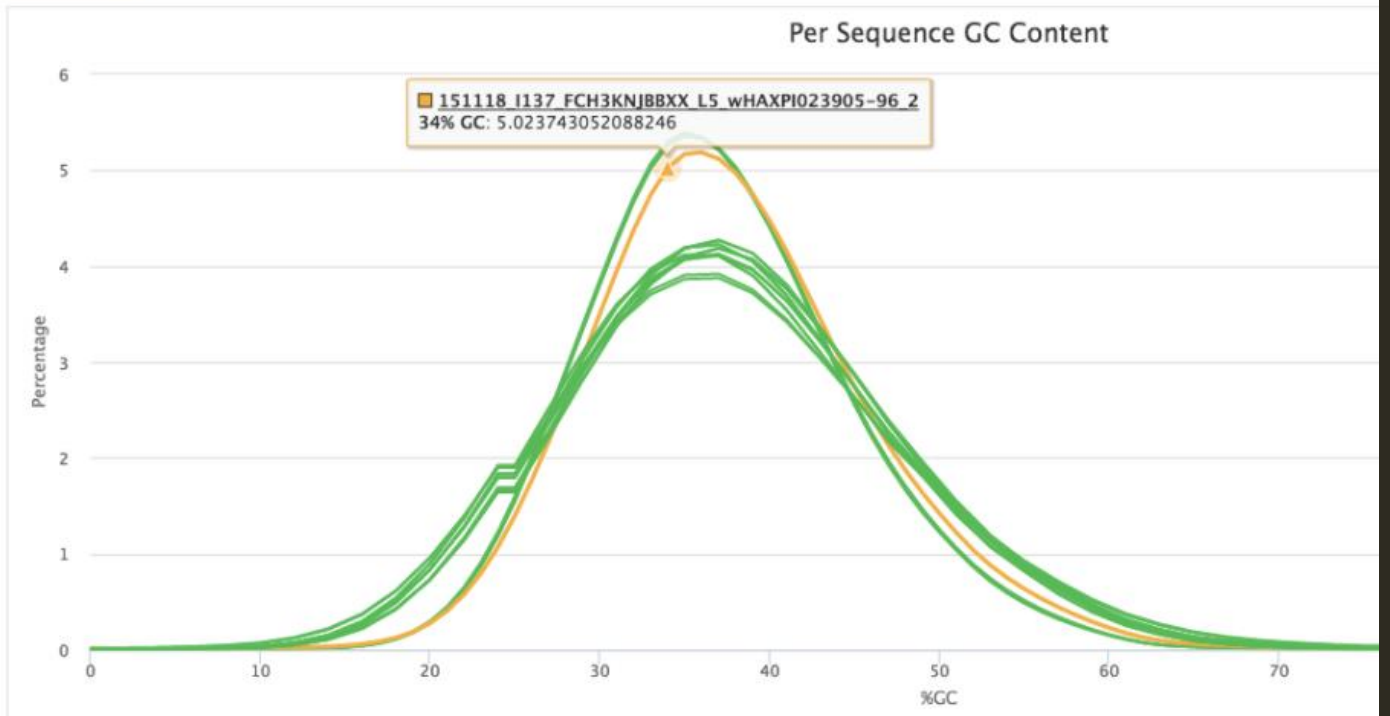
https://sr320.github.io/Oly-read-check/

Per Sequence GC Content

15 1

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the FastQC help.

Percentages | Counts



EDYCJA SEKWENCJI

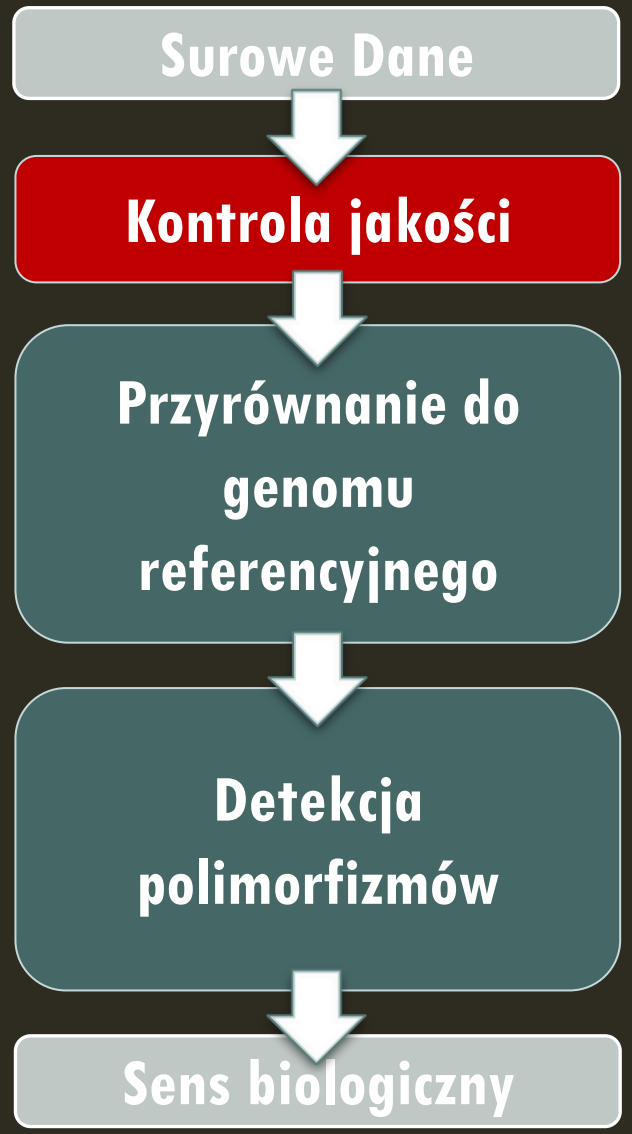
→ MOTYWACJA

Błędne dane mogą prowadzić do:

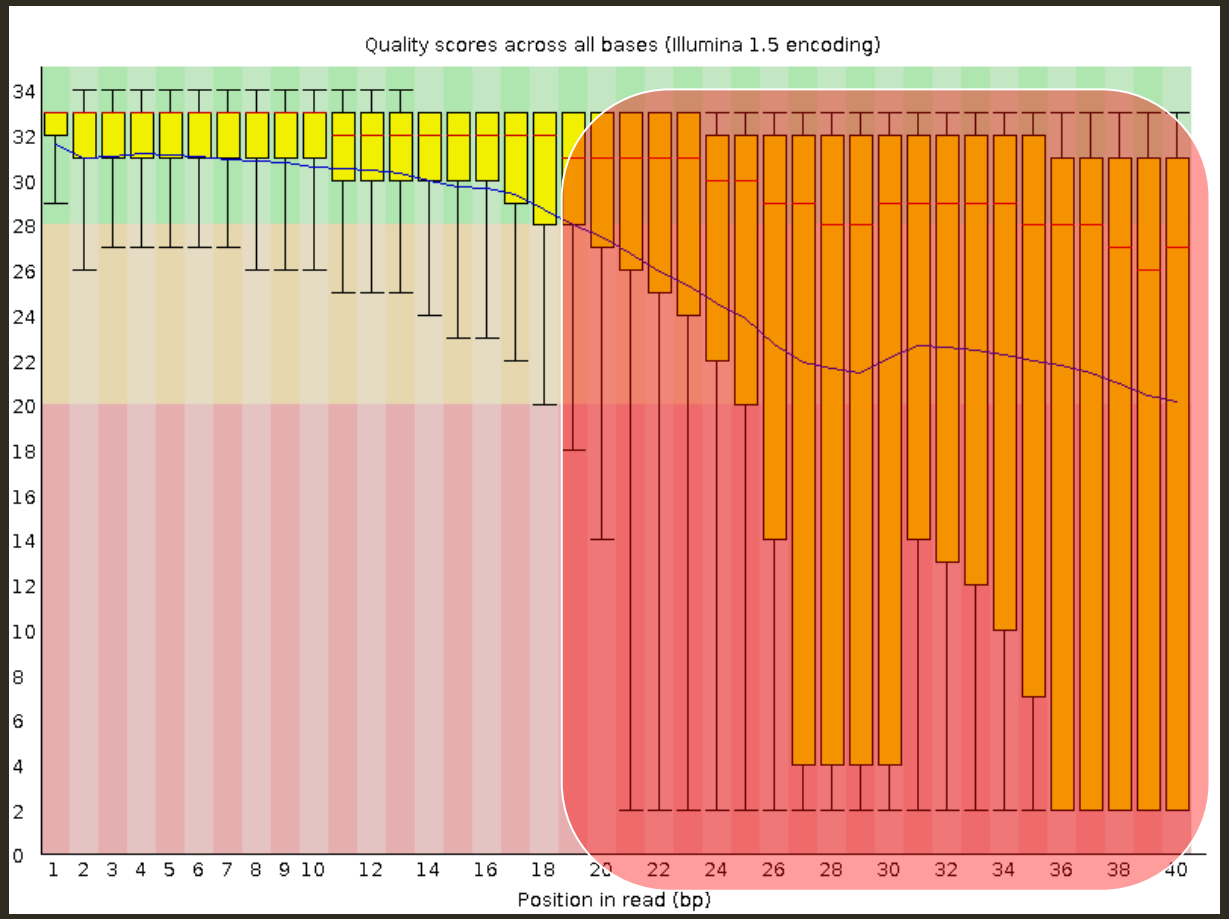
- wolniejszego działania oprogramowania
- generowania słabej jakości/niewłaściwych wyników

Czyszczenie danych:

- zwiększa średnią jakość sekwencji
- daje lepsze rezultaty przyrównania
- redukuje rozmiar danych



@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 1:N:0:ACAGTG
AGAAATGCCAGGCTAGATGAGTTACAATCNAGTATCAAGATAGGC
+
@@@FFDFFGHGHHFDDDGHHHDDDD**44#$%&,344+400/01234**
(…)



Surowe Dane

Kontrola jakości

Przyrównanie do genomu referencyjnego

Detekcja polimorfizmów

Sens biologiczny

CZYSZCZENIE DANYCH

Trimmomatic

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Paired End:

```
java -jar trimmomatic-0.35.jar PE -phred33 input_forward.fq.gz input_reverse.fq.gz
output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz
output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

Description

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.The selection of trimming steps and their associated parameters are supplied on the command line.

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

It works with FASTQ (using phred + 33 or phred + 64 quality scores, depending on the Illumina pipeline used), either uncompressed or gzipp'ed FASTQ. Use of gzip format is determined based on the .gz extension.

For single-ended data, one input and one output file are specified, plus the processing steps. For paired-end data, two input files are specified, and 4 output files, 2 for the 'paired' output where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not.

KONTROLA JAKOŚCI CZYSZCZENIE DANYCH

PRINSEQ

http://prinseq.sourceforge.net/

PRINSEQ

→ AMBIGUOUS BASES

Sequences can contain the ambiguous base N for positions that could not be identified as a particular base. A high number of Ns can be a sign for a low quality sequence or even dataset.

Ambiguous bases can cause problems during downstream analysis. The different programs deal with the problem in different ways. Some programs replace ambiguous bases with a random base (e.g. BWA) and others with a fixed base (e.g. SHAHA2 and Velvet replace Ns with As). This can result in misassemblies or false mapping of sequences to a reference sequence and therefore, sequences with a high number of Ns should be removed before downstream analysis.

Filtering reads containing more than 1% of ambiguous bases is advised.

PRINSEQ

→ MINIMUM AND MAXIMUM READ LENGTH

Short sequences are more likely to match at a random position by chance than longer sequences and may therefore result in false positive functional or taxonomical assignments.

In some cases, sequences can be much longer than several standard deviations above the mean length (e.g. 1,500+ bp for a 500 bp mean length with a sd100 bp). Those sequences should be used with caution as they likely contain long stretches of homopolymer runs. Homopolymers are a known issue of pyrosequencing technologies.

A rule of thumb for sequence length thresholds of longer-read datasets is to filter sequences shorter than 60 bp (20 amino acids) and longer than twice the mean length.