Quality control and artefact removal

FastQC, Cutadapt, Trimmomatic, Fastx toolkit

Dóra Bihary

MRC Cancer Unit, University of Cambridge

CRUK CI Bioinformatics Summer School

July 2018

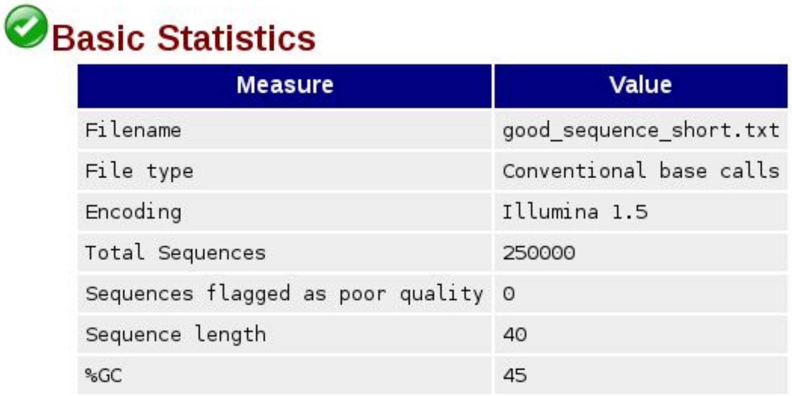
Overview

- Quality control
    - FastQC
- Artefact removal
    - Cutadapt/TrimGalore, Trimmomatic

Why do we need quality control?

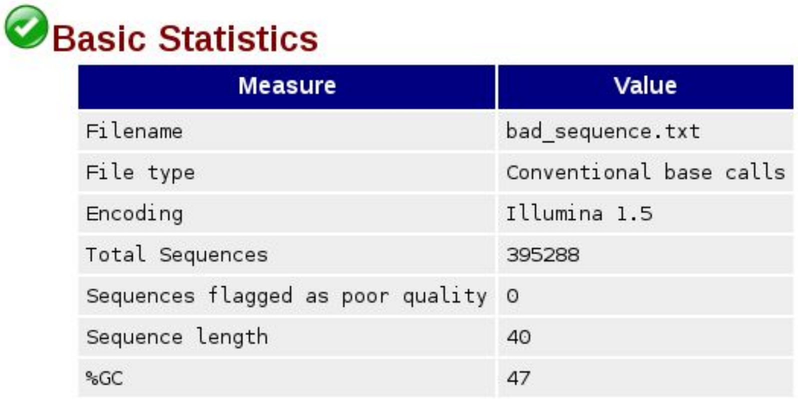
- NGS sequencing generates highly accurate data, but it can have certain types of errors:
    - Contamination with adapters
    - Technical duplication in the library
    - Failure at specific parts of the flowcell
    - PCR duplicates
    - Etc.
- This is why it is important to check the data quality before alignment
- FastQC:
    - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
    - Reads in fastq files and generates reports based on the quality information that the sequencer provided
    - Command line and interactive mode
    - Outputs an html report and a .zip file with the raw quality data
- MultiQC:
    - http://multiqc.info/
    - Aggregates FastQC results of multiple analyses into a single report

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



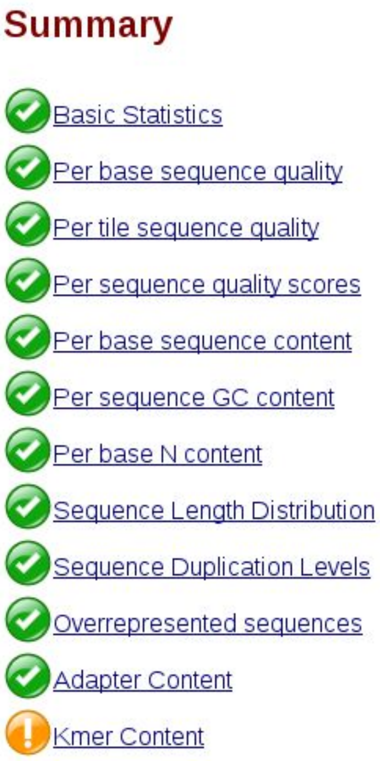
✓ Basic Statistics

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

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

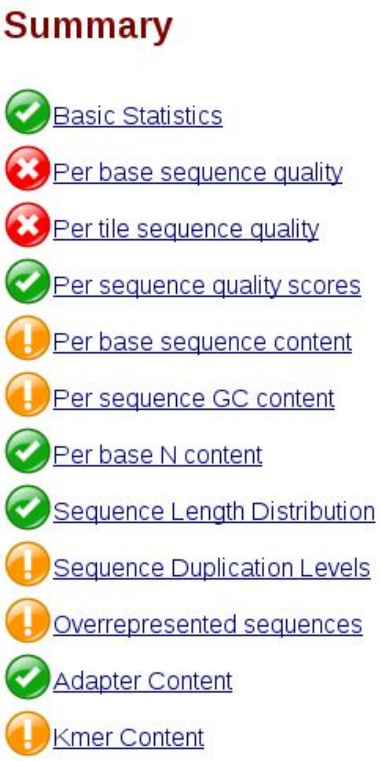
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC - summary



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
- ! Kmer Content

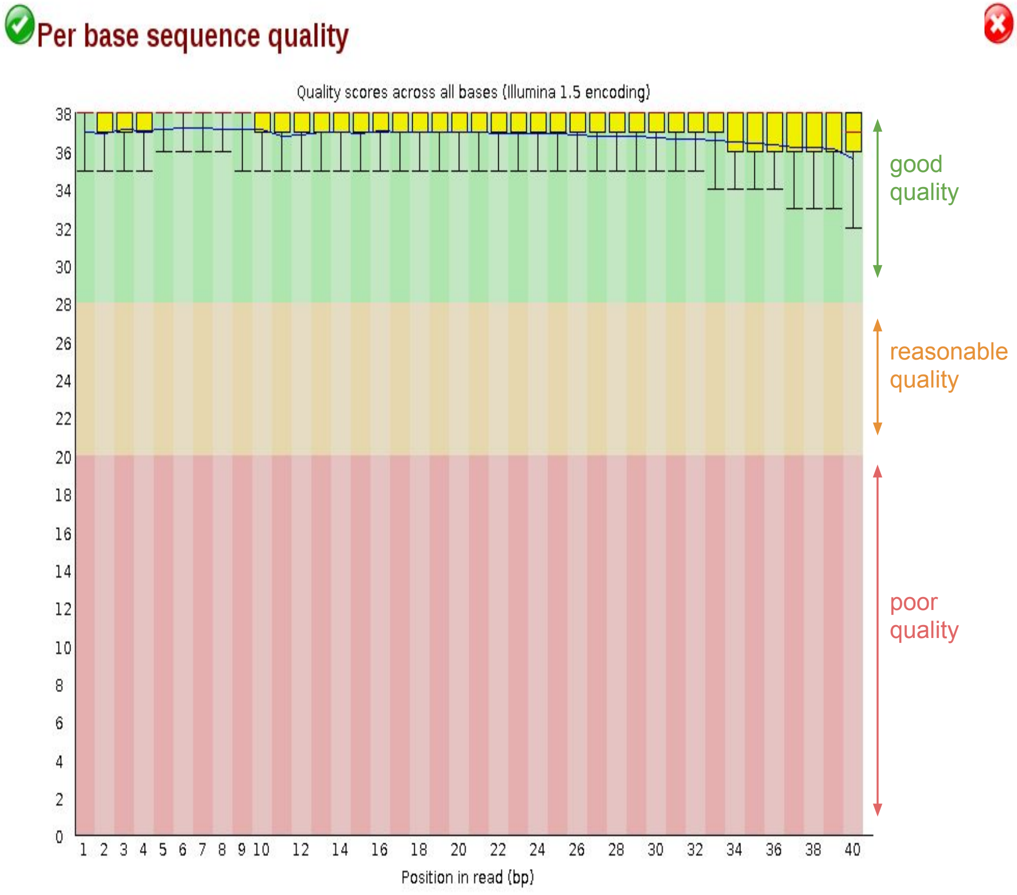


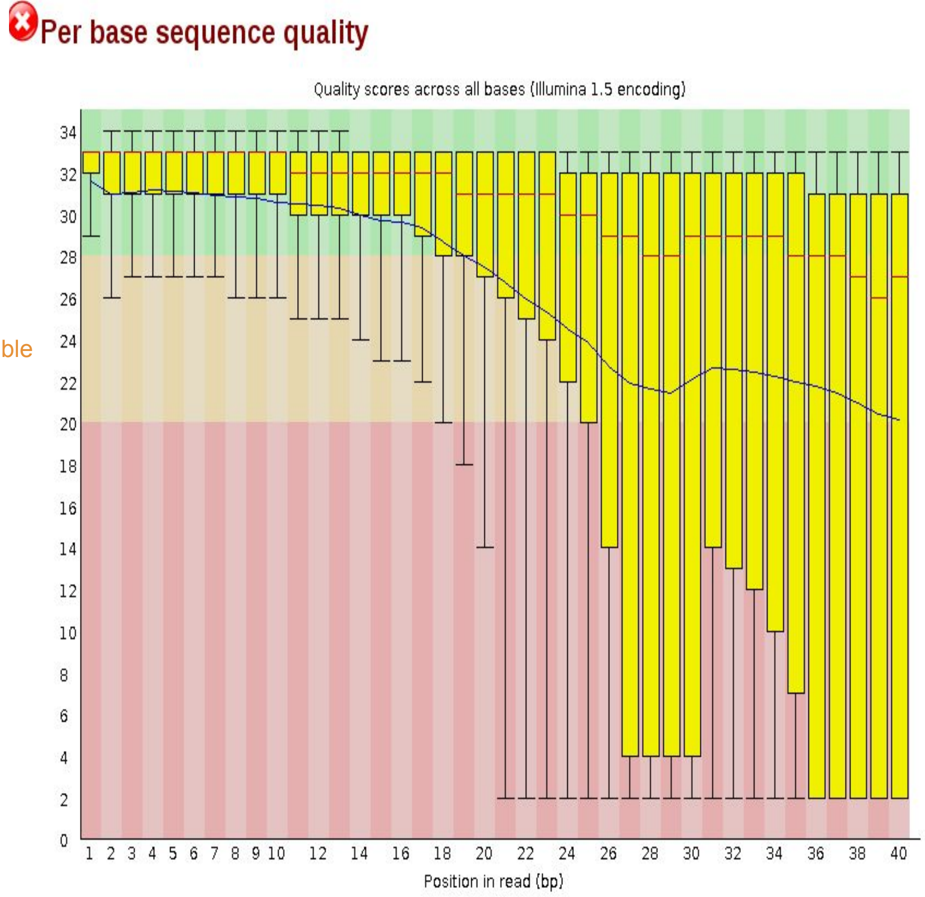
Summary

- ✓ 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
- ! Kmer Content

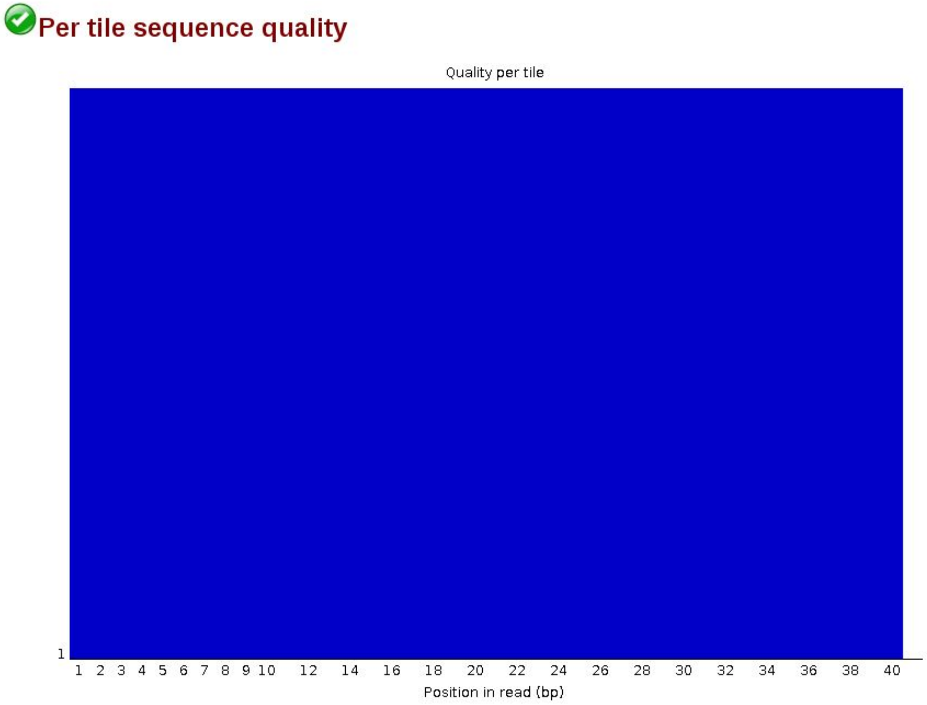
FastQC - per base sequence quality

🟥	median value
🟦	mean value
🟨	inter-quartile range (25-75%)
whiskers	10% and 90% of points

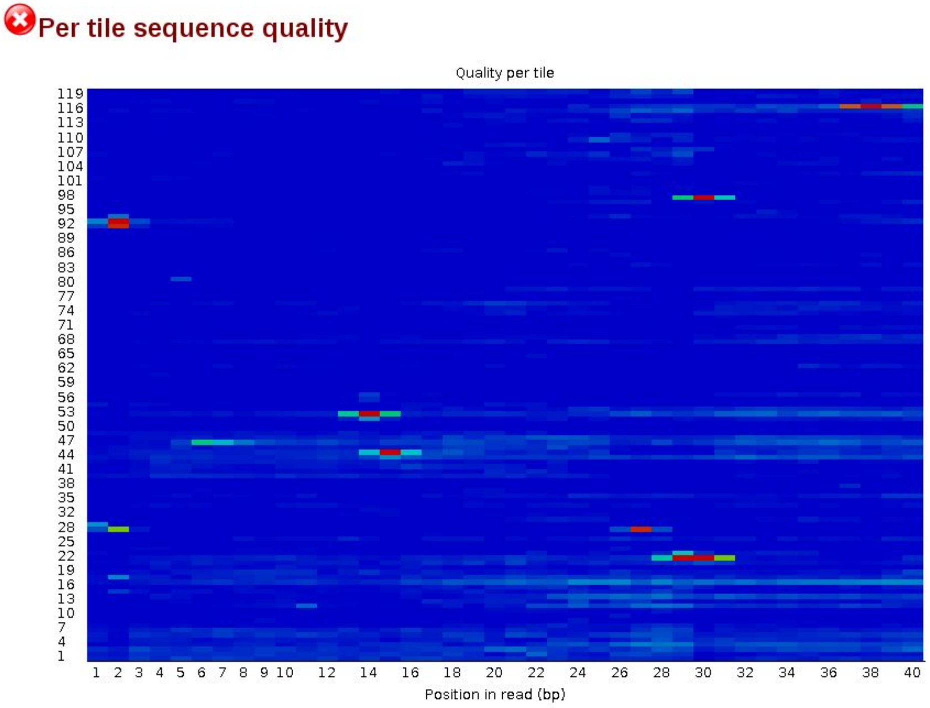




FastQC - per tile sequence quality

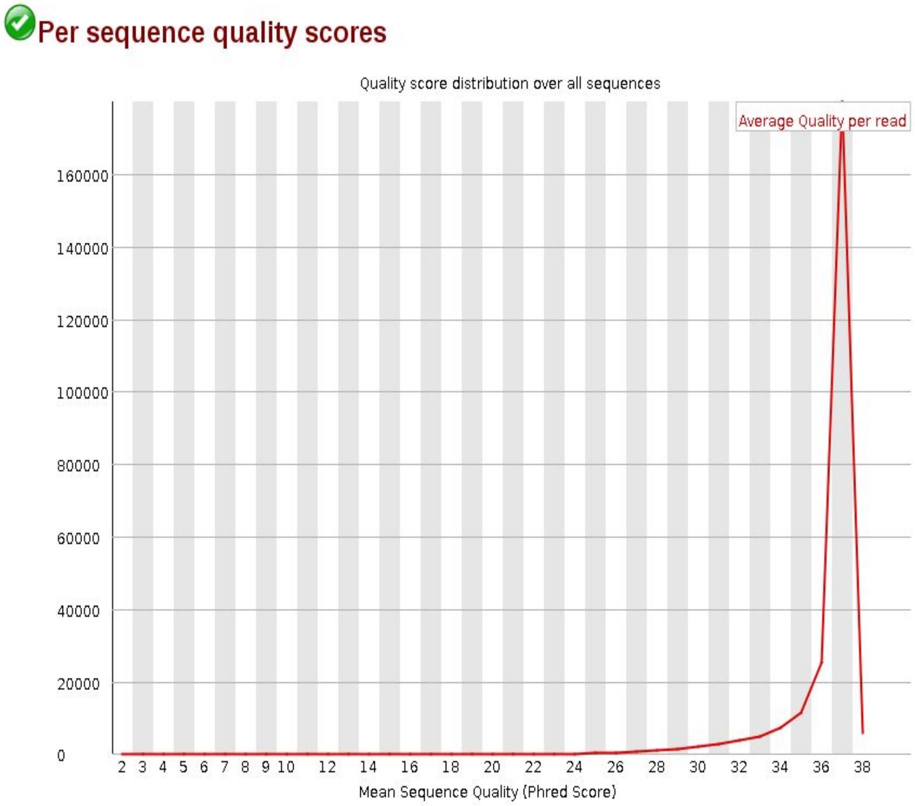


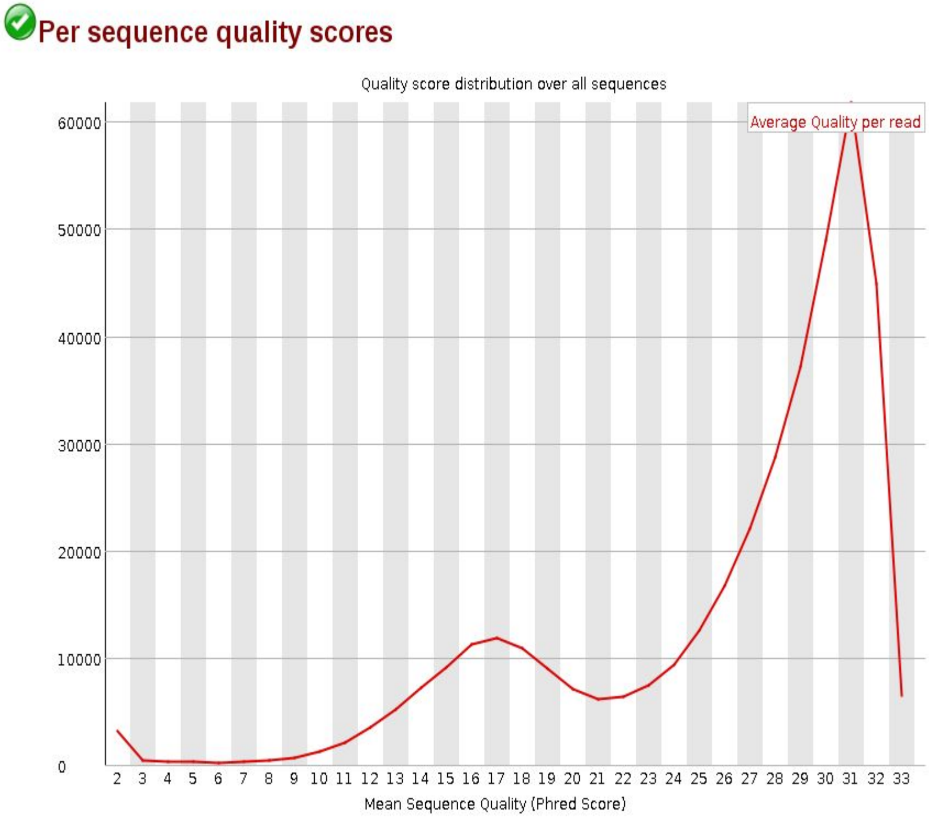
✓ Per tile sequence quality



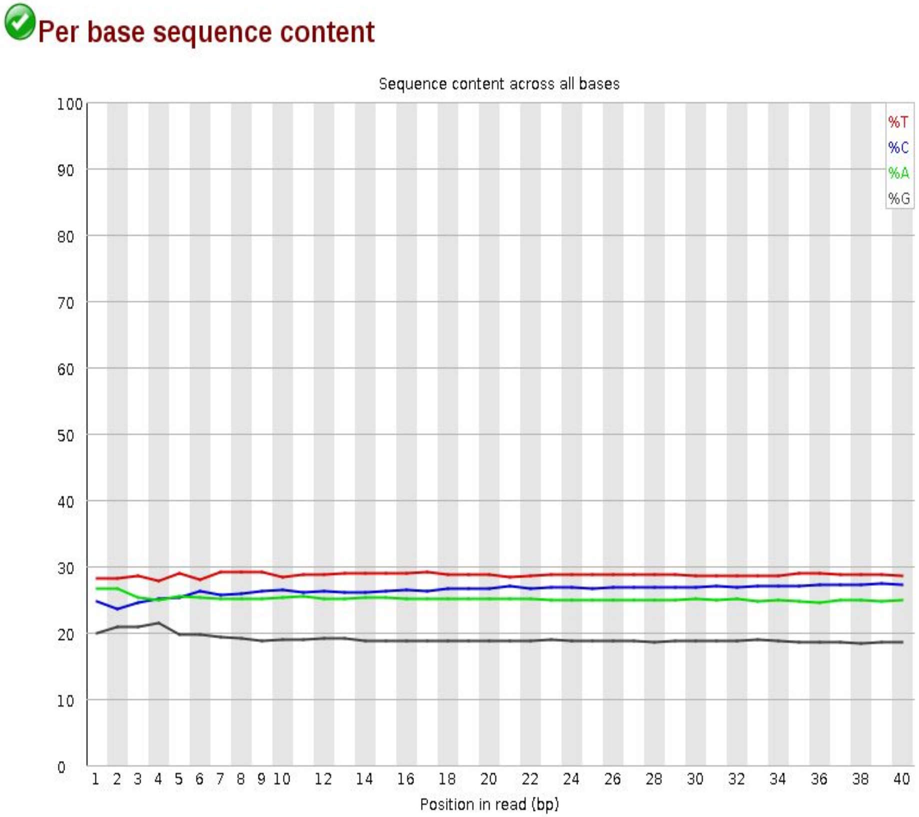
✗ Per tile sequence quality

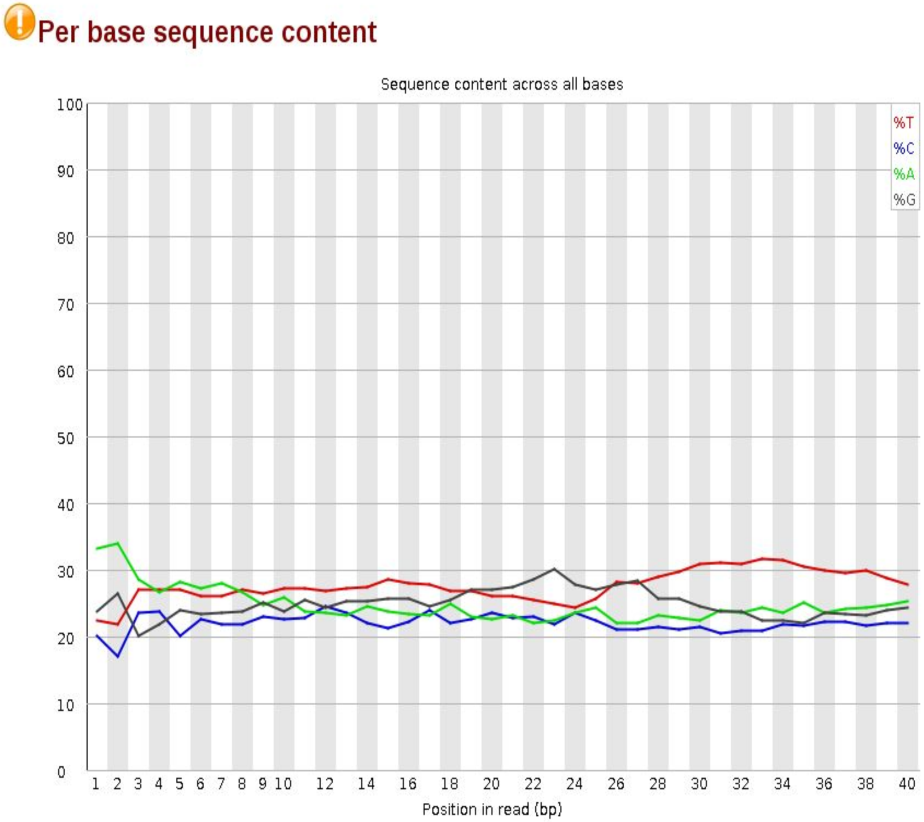
FastQC - per sequence quality scores


✓ Per sequence quality scores

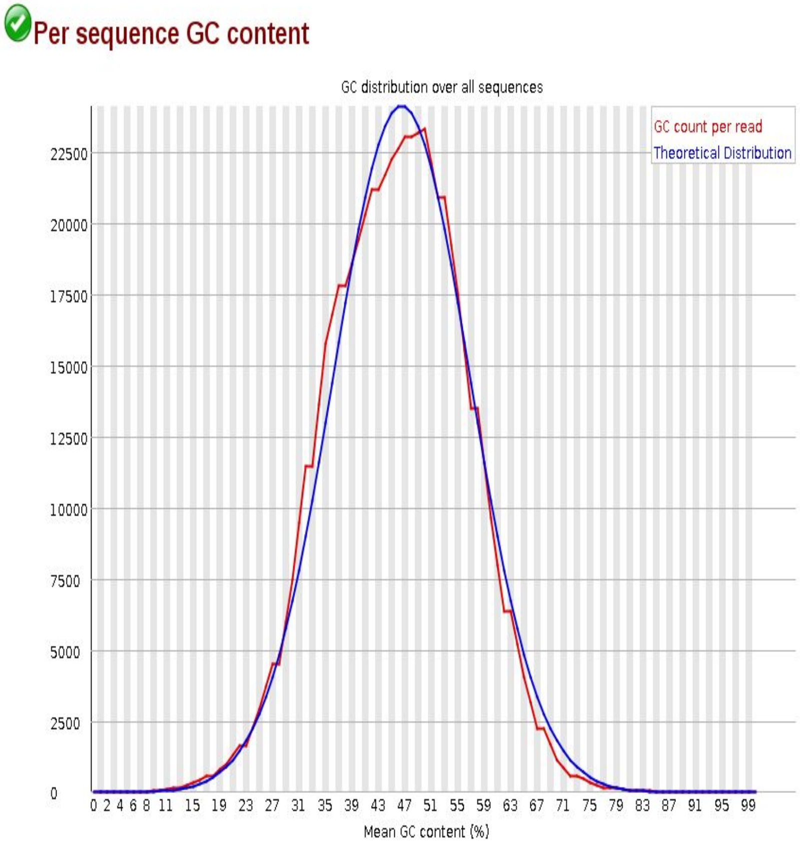

✓ Per sequence quality scores

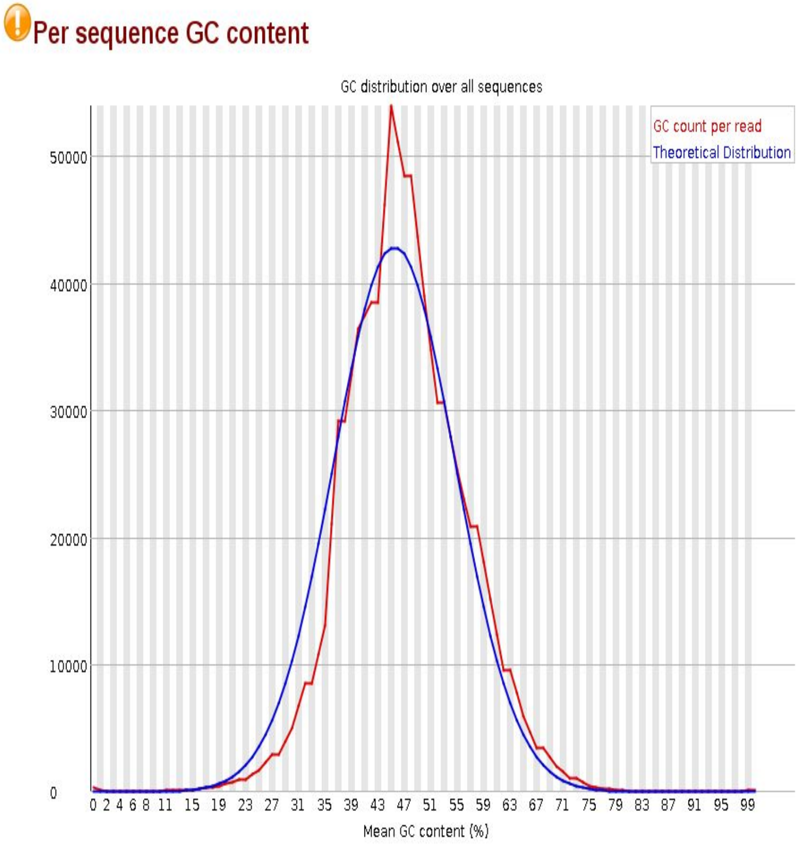
FastQC - per base sequence content





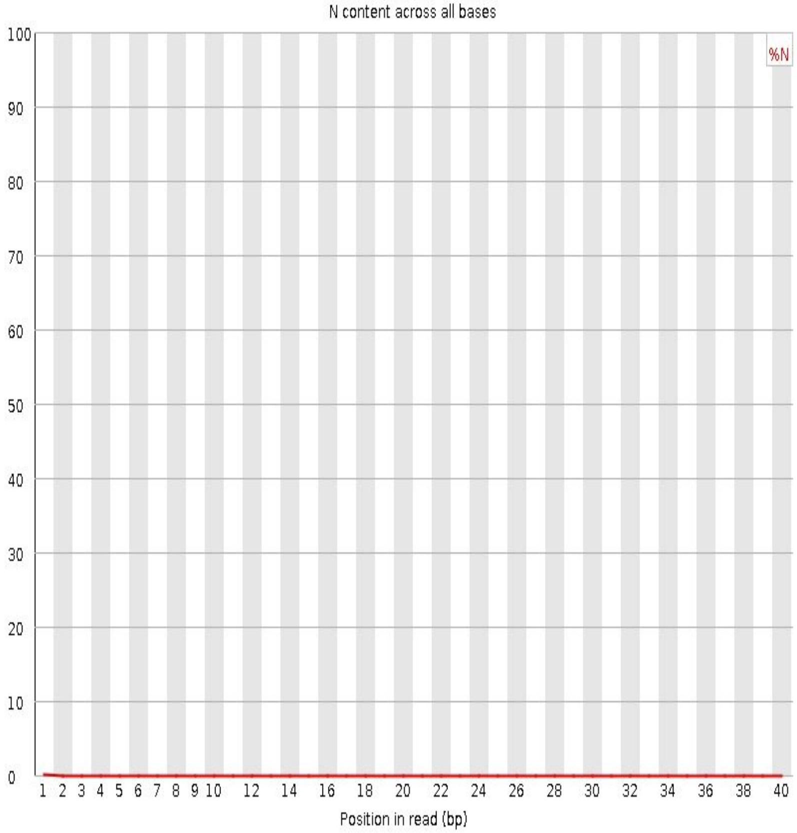
FastQC - per sequence GC content



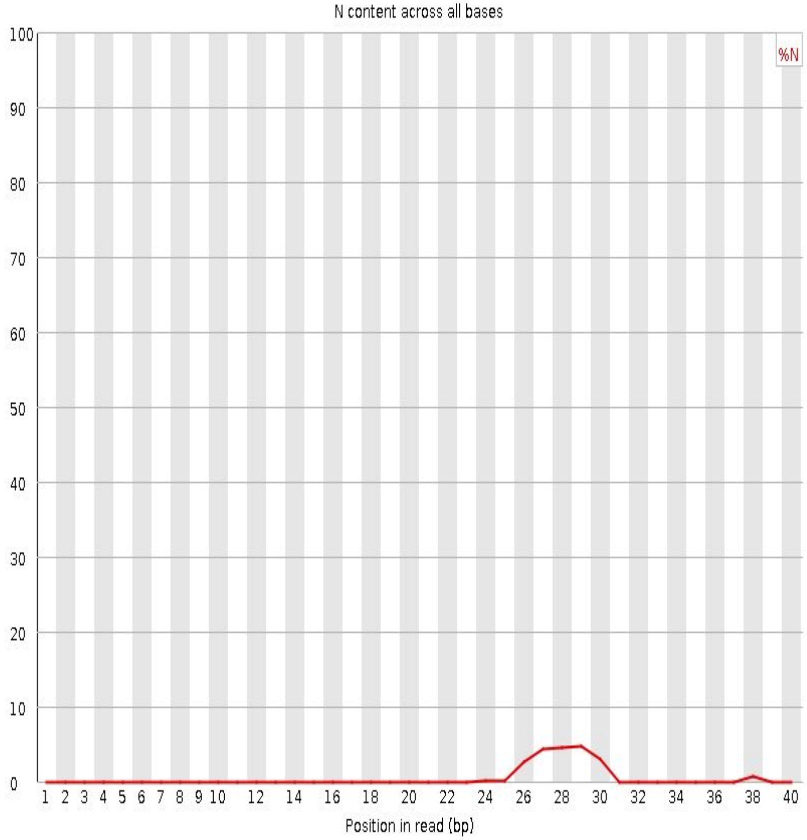


FastQC - per base N content

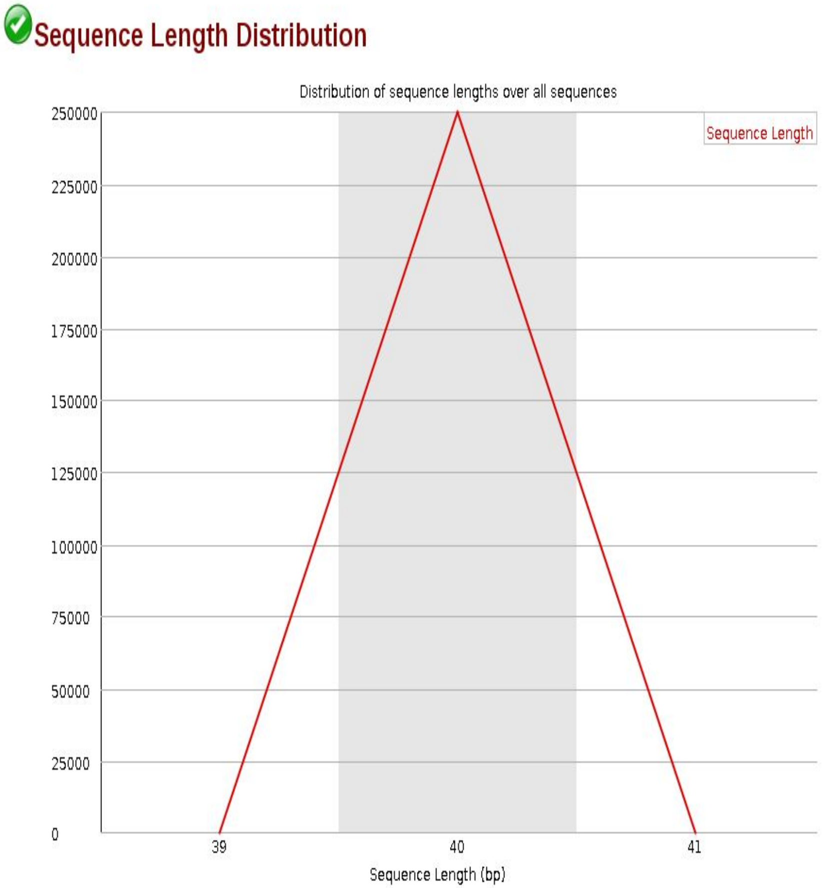
✓ Per base N content

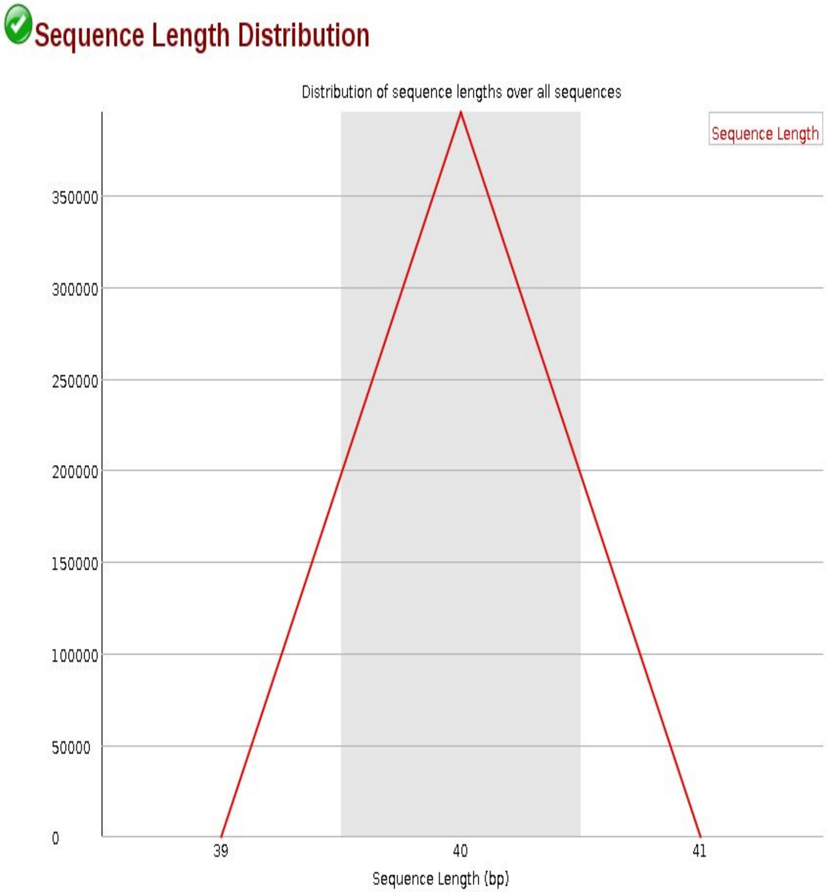


✓ Per base N content

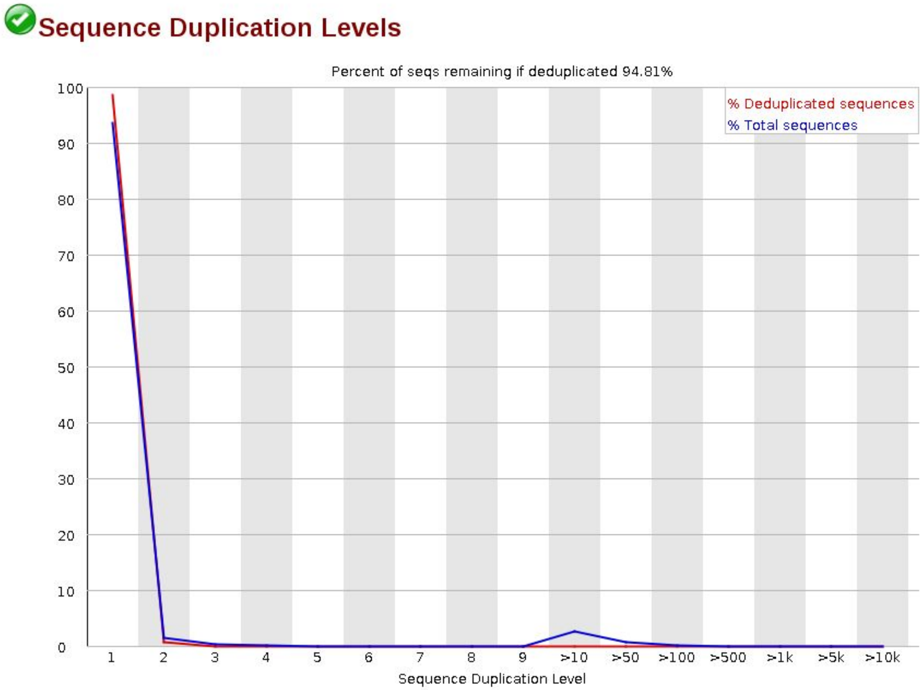


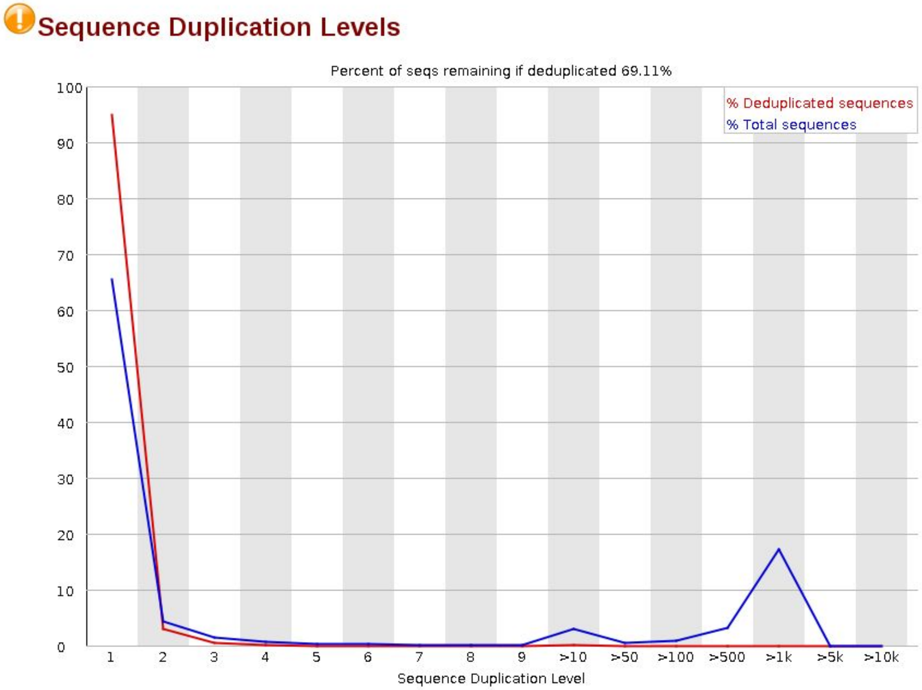
FastQC - sequence length distribution



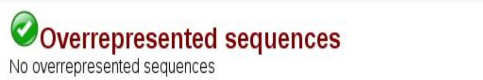


FastQC - sequence duplication level



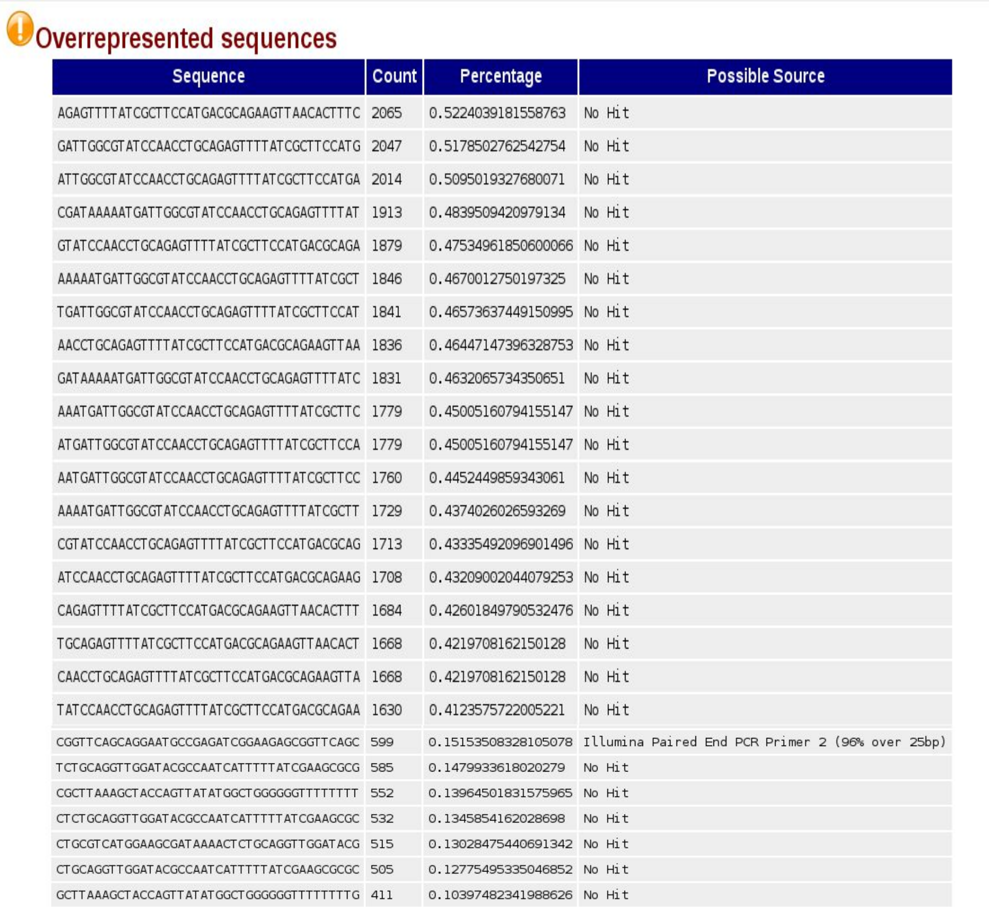


FastQC - overrepresented sequences



Overrepresented sequences

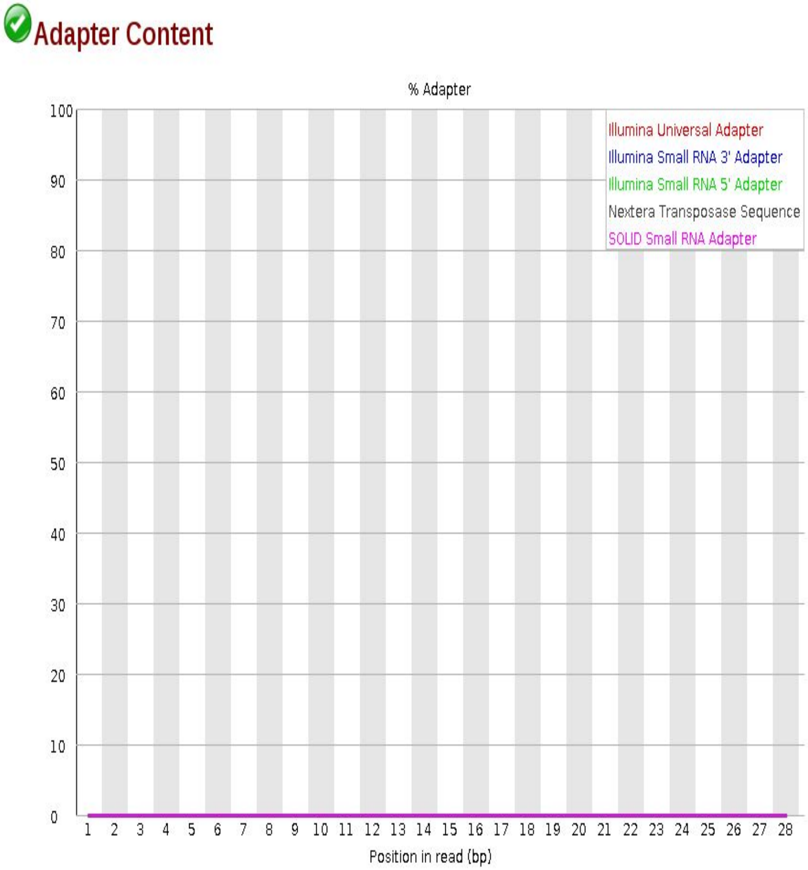
No overrepresented sequences

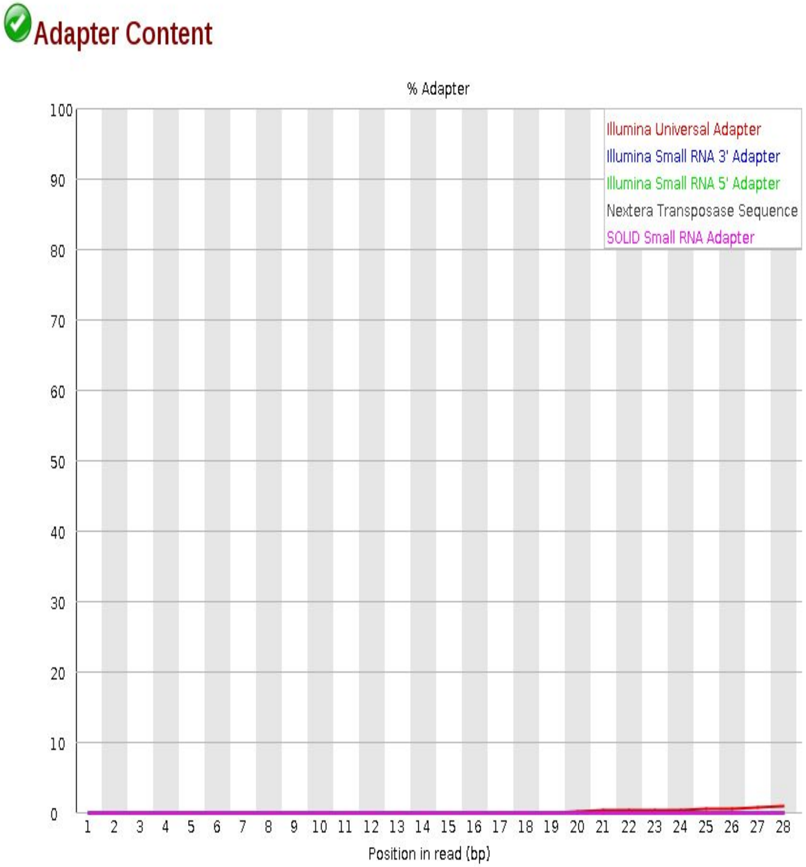


Overrepresented sequences

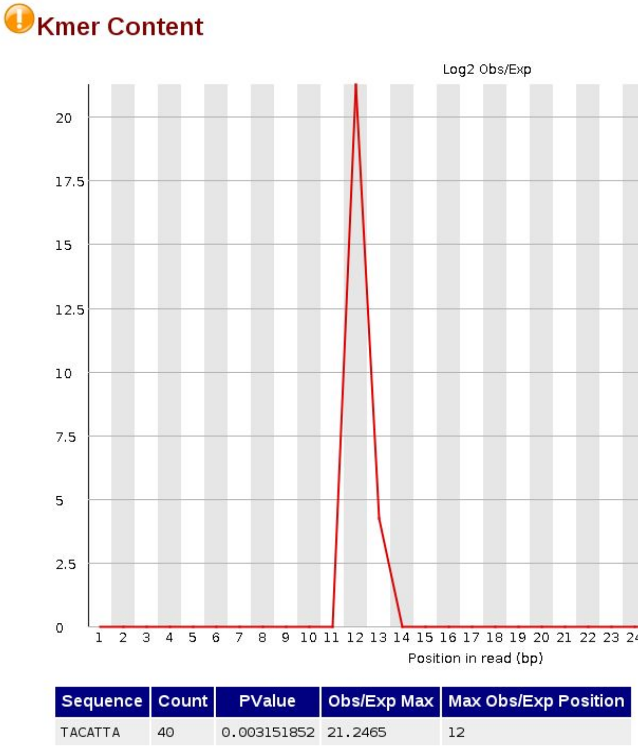
Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	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

FastQC - adapter content



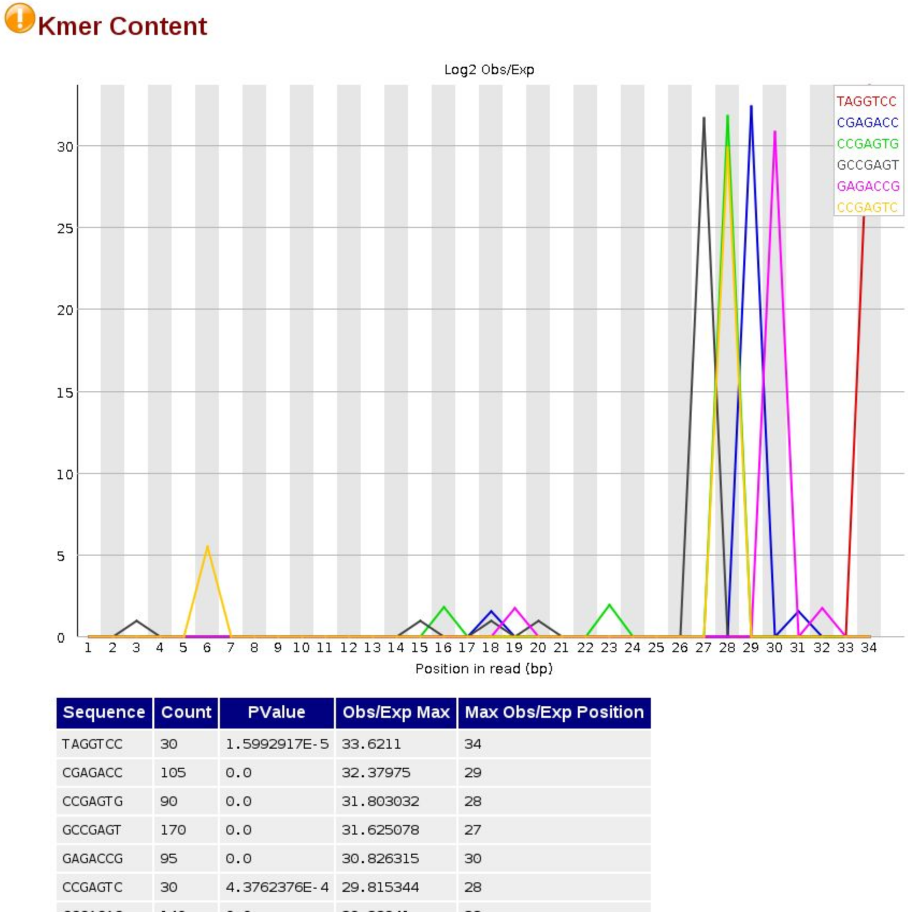


FastQC - kmer content



Kmer Content

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TACATTA	40	0.003151852	21.2465	12



Kmer Content

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TAGGTCC	30	1.5992917E-5	33.6211	34
CGAGACC	105	0.0	32.37975	29
CCGAGTG	90	0.0	31.803032	28
GCCGAGT	170	0.0	31.625078	27
GAGACCG	95	0.0	30.826315	30
CCGAGTC	30	4.3762376E-4	29.815344	28

Overview

- Quality control
  - FastQC
- Artefact removal
  - Cutadapt/TrimGalore, Trimmomatic

Artefact removal

- Important when the quality needs to be increased
- Adapter trimming
  - Based on “Overrepresented Sequences”, “Adapter Content” and/or “Kmer Content” you might identify certain adapter contaminations that needs to be trimmed
  - Spikes in “Per sequence GC content” usually indicate adapter contamination
- Quality-based trimming
  - When the quality drops eg. towards the end of reads (“Per base sequence quality”)
  - When the “Per base sequence content” shows bias in sequence composition towards beginning/end
  - You can trim regions below a certain quality threshold (eg. 20)
  - You can trim *n* bases from beginning/end of all your reads

Artefact removal - paired-end data

- We want to preserve the pairs so that aligners will know which reads belong together
- We have to keep track of the pairs of those reads that are removed from one of the paired files
- Four output files will be produced, two with the trimmed paired reads and two with the unpaired ones

Artefact removal - tools

- Cutadapt/TrimGalore
    - http://cutadapt.readthedocs.io/en/stable/index.html
    - TrimGalore: wrapper around Cutadapt
- Trimmomatic
    - http://www.usadellab.org/cms/?page=trimmomatic
- Fastx toolkit
    - http://hannonlab.cshl.edu/fastx_toolkit/
    - short read pre-processing tool
    - fastx_trimmer: fixed length trimmer
    - fastq_quality_filter: quality based trimmer
    - fastx_artifacts_filter: artefact remover