From sequencer to polished reads for bacteria

September 2023

General information

- 2 sessions - 9-13 GMT+2
  - 12th of September
  - 14th of September
- Sessions will be recorded!
- Q&A bottom to ask questions
- Please take your time to evaluate on EVA
- Tech expert
  - Rodrigo

Overall objectives

- Inspect the raw fastq files from the sequencer
- Understand how sequences can differ based on the preparation methodology
- Explain the differences between sequencing reads from Illumina and Nanopore
- Assess the quality of sequence data and trim low quality data
- Learn when to decontaminate and when to re-sequence
- Generate an overview of an entire dataset using different tools

Course intro

Day 1
Sep 12th

Theory

- Course intro
- Illumina sequencing theory
- Nanopore sequencing theory
- Sequencing technique comparison

Practical

- Practical 1 intro
- First look at dataset
- Contamination control
- Raw read QC

Day 2
Sep 14th

Interactive

- Session 1 recap
- Interpret raw read QC results
- Interpret contamination control results

Practical

- Short-read trimming
- Long-read trimming
-
- QC parameters for a whole dataset
- Course summary

Presenters

Kasper Thystrup Karstensen

Education:

- Bachelor's in Medical Laboratory Technology
- Master's in Bioinformatics and Systems Biology

Experience:

- Background in RNA-related work
- Employed at SSI since January 2022, specializing in antimicrobial resistance surveillance

Astrid Rasmussen

Education:

- Bachelor's in Biology
- Master's in Bioinformatics and Systems Biology

Experience:

- Employed at SSI since January 2022, specializing in antimicrobial resistance surveillance and Nanopore sequencing

Before we start…

Previously…

A very brief recap

Illumina technology

Theoretical

- Differences between single-end and paired-end reads
- Insights into paired end read fragments as well as size selection
- Introduction to sequencing by synthesis

Illumina technology

Theoretical

- Differences between single-end and paired-end reads
- Insights into paired end read fragments as well as size selection
- Introduction to sequencing by synthesis

Practical

- Introduction to FastQC for generating QC reports
- Hands-on execution of FastQC

Nanopore technology

Theoretical

- A fundamental understanding on the Nanopore chemistry
- An overview of different machinery as well as library prep options
- Knowledge of possible modifications post novel insight as well as challenges

Nanopore technology

Theoretical

- A fundamental understanding on the Nanopore chemistry
- An overview of different machinery as well as library prep options
- Knowledge of possible modifications post novel insight as well as challenges

Practical

- Introduction to NanoPlot for visualising QC statistics
- Hands-on execution of NanoPlot

FastQC

Quality assessment of read data

Basic statistics

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

Basic statistics

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

✓ Same amount of sequences

Basic statistics

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

✓Same amount of sequences！ Somewhat same amount of bases

Basic statistics

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

✓ Same amount of sequences ！ Somewhat same amount of bases
✓ No poor quality flags

Basic statistics

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

✓ Same amount of sequences
✓ No poor quality flags

! Somewhat same amount of bases
! Somewhat identical sequence sizes

Basic statistics

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

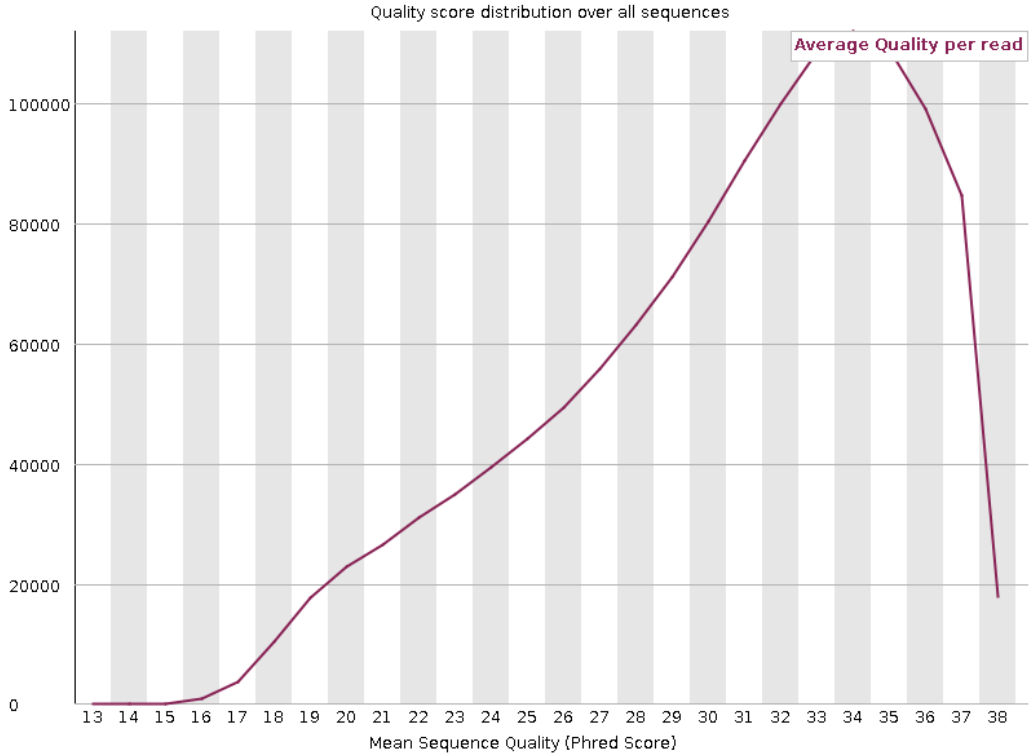
Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

✓ Same amount of sequences
✓ No poor quality flags
✓ Identical GC contents

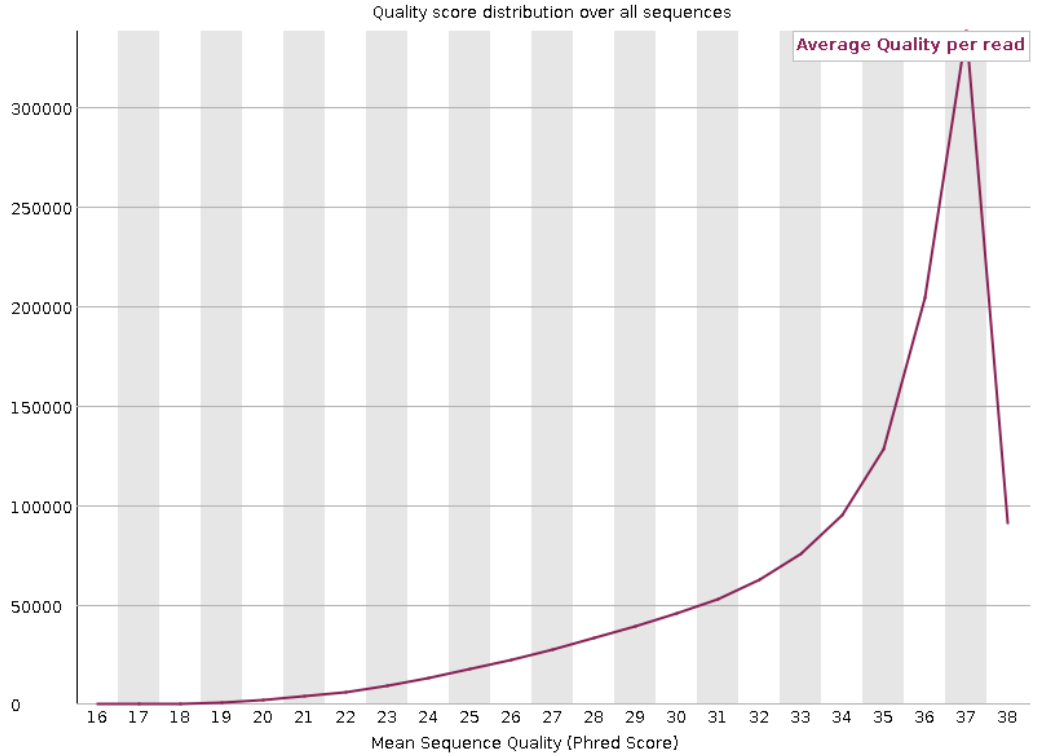
! Somewhat same amount of bases
! Somewhat identical sequence sizes

Per sequence quality

Ec005_R1

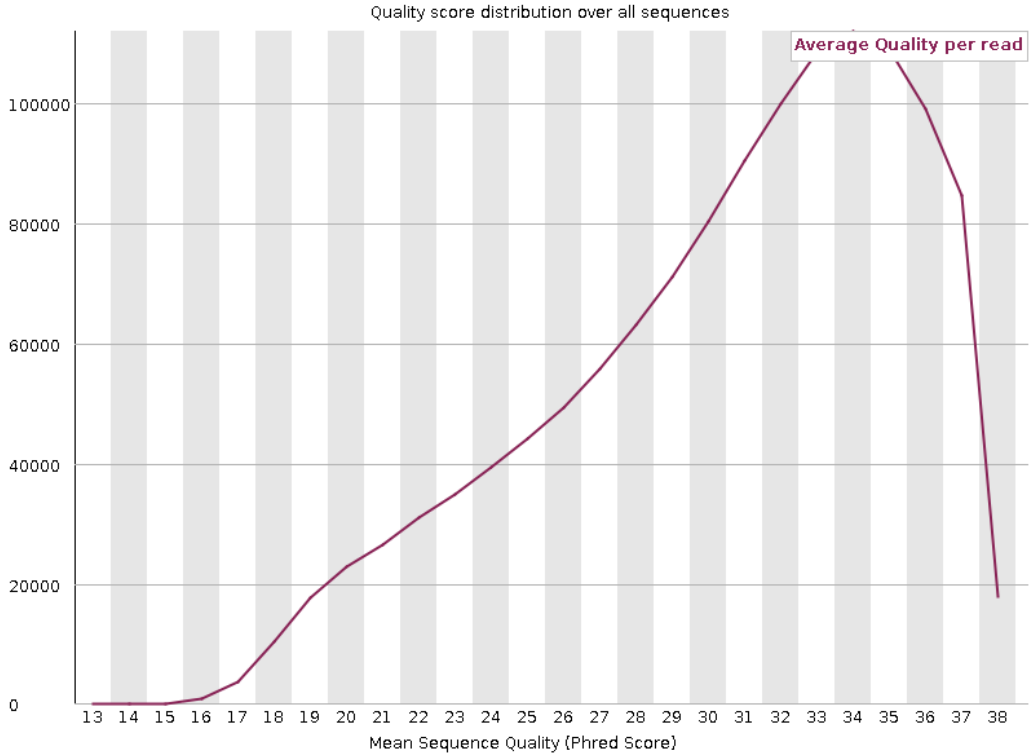


Ec005_R2

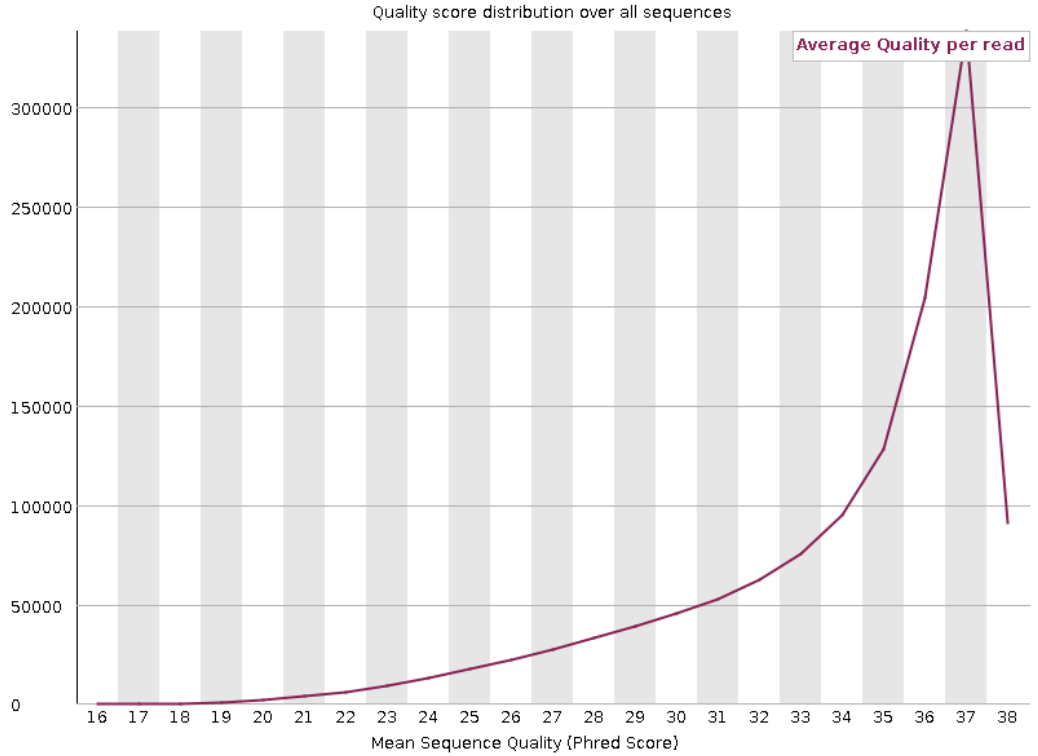


Per sequence quality

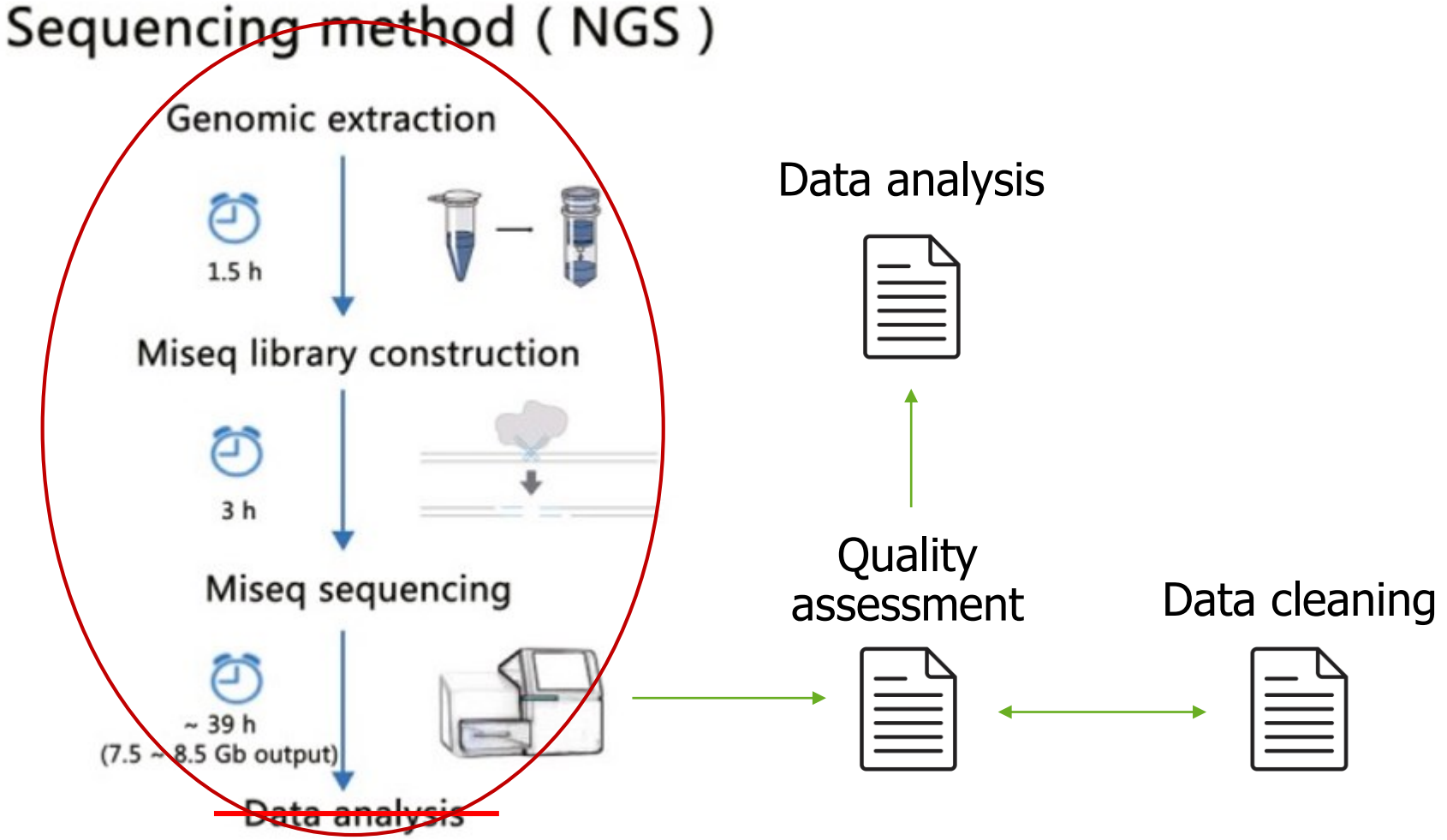
Ec005_R1



Ec005_R2



✓ Majoirty of average read quality > Q30



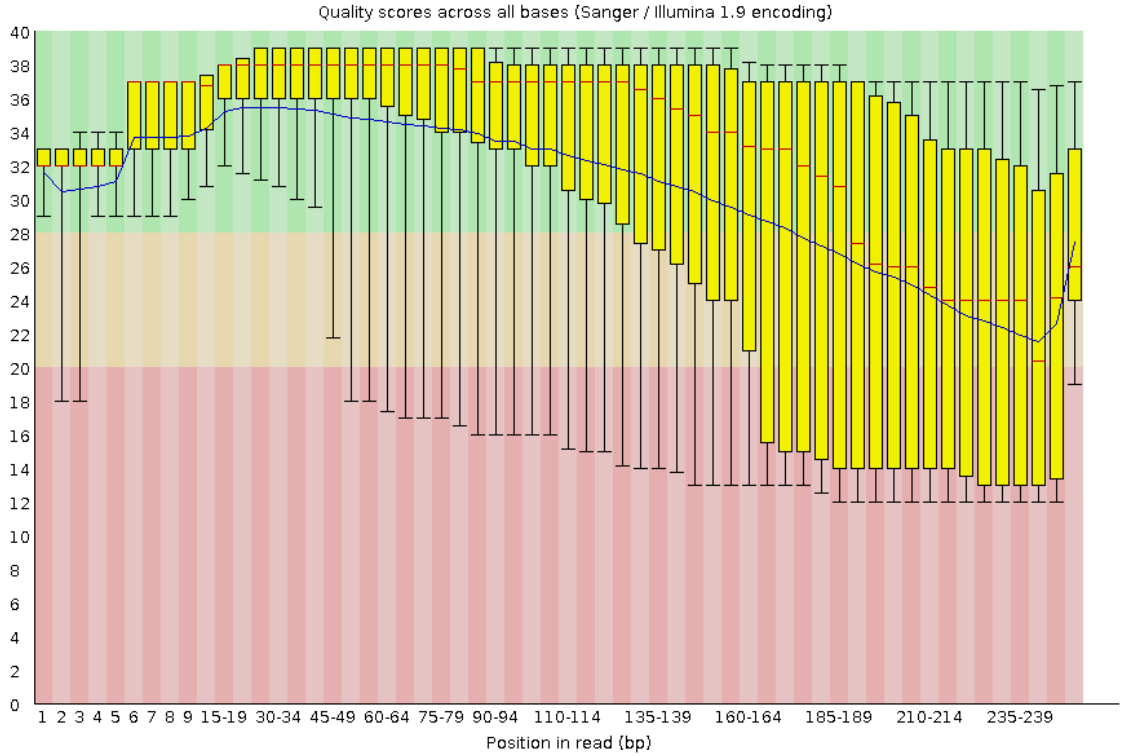
Data analysis

Quality assessment

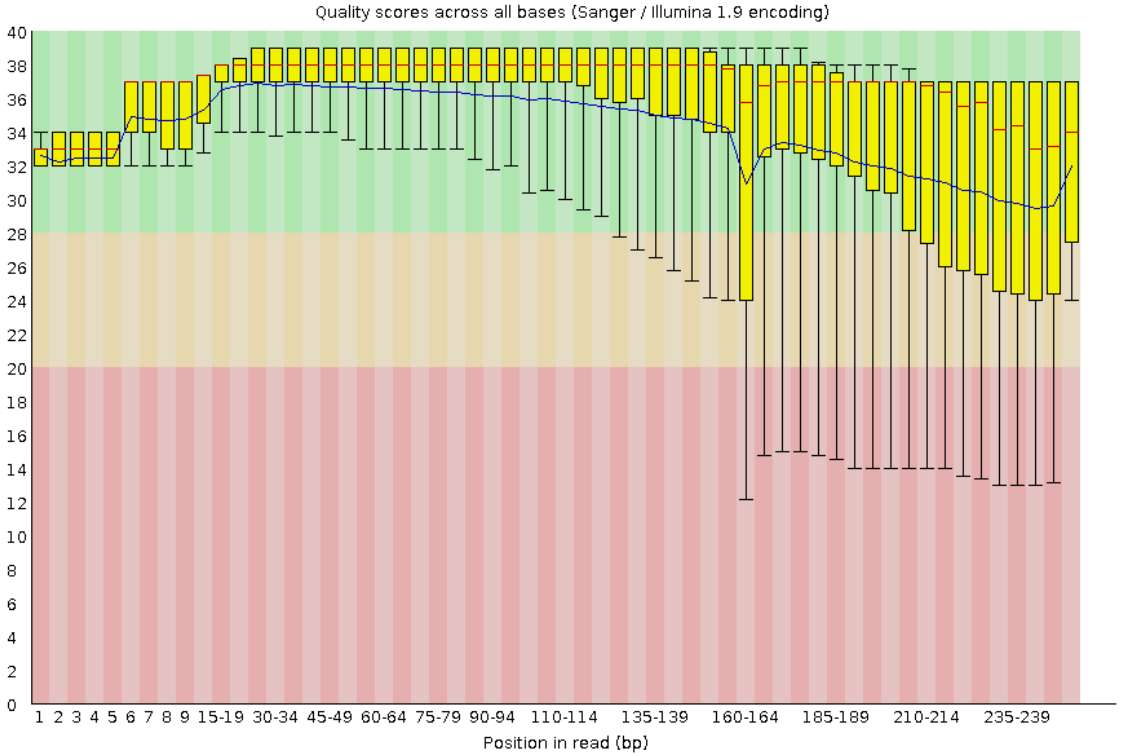
Data cleaning

Per base sequence quality

Ec005_R1

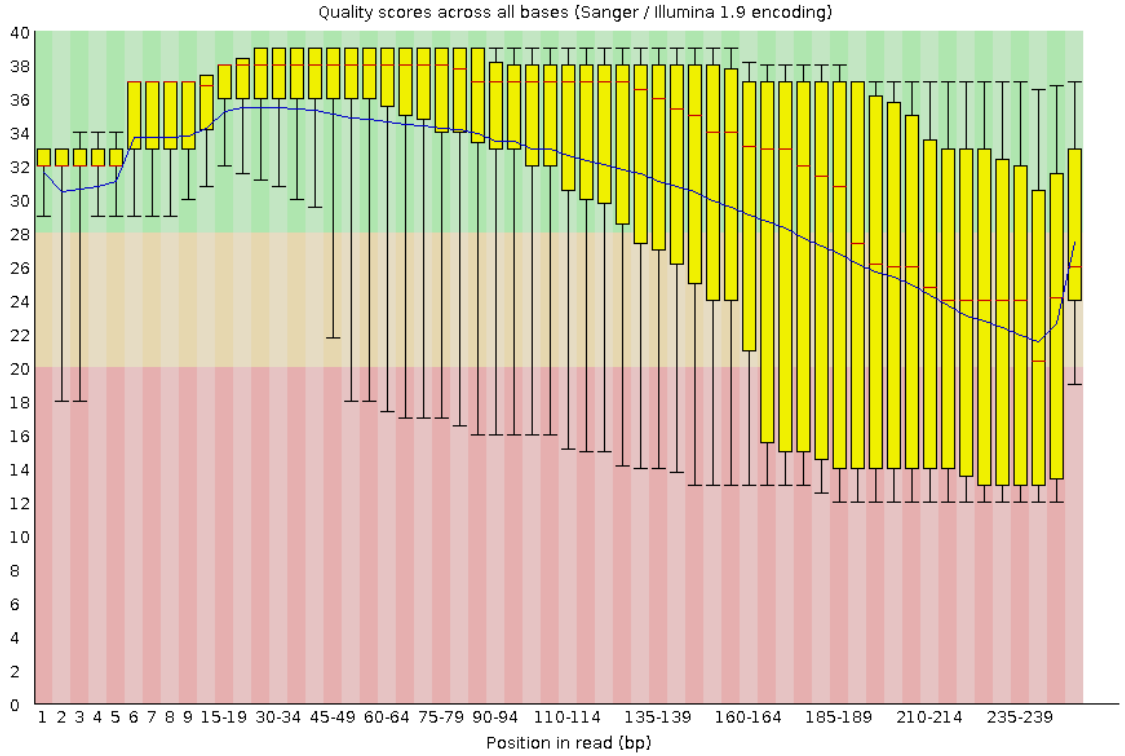


Ec005_R2

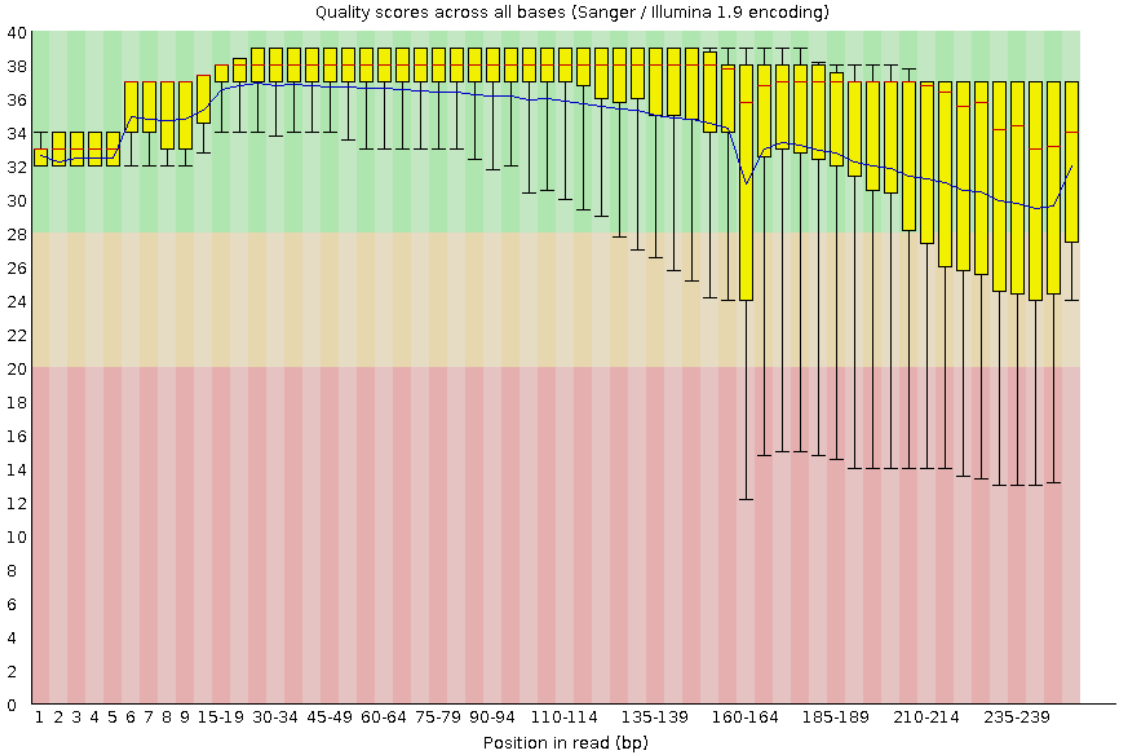


Per base sequence quality

Ec005_R1



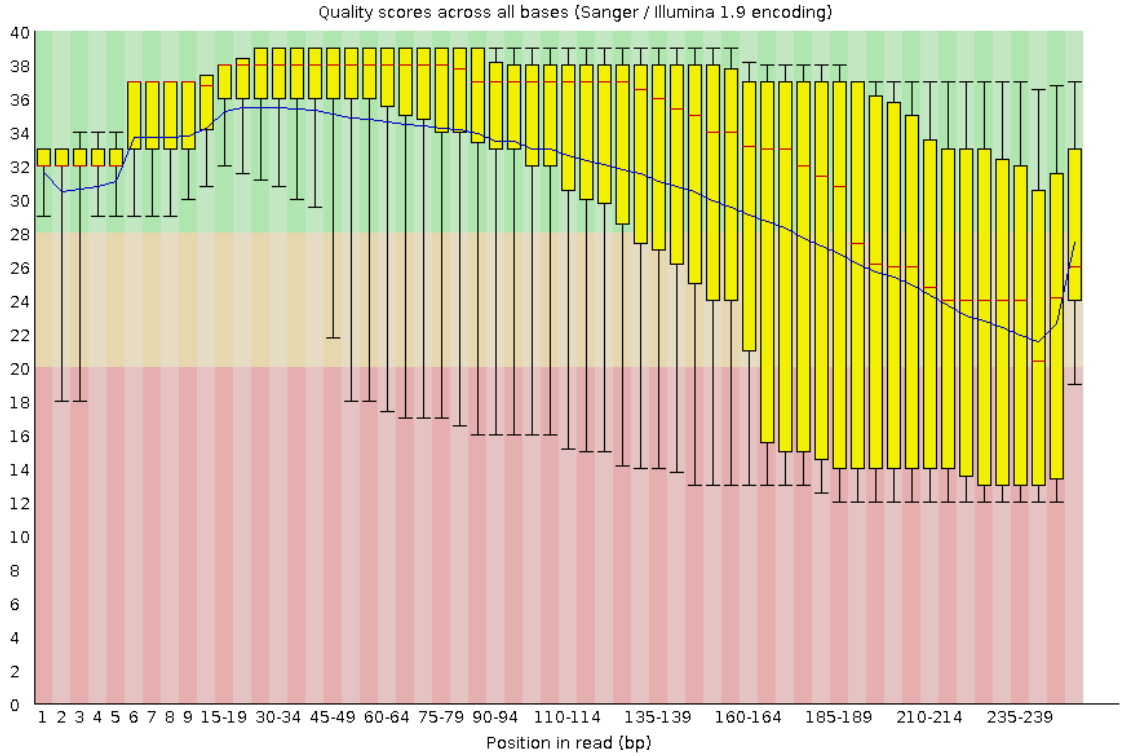
Ec005_R2



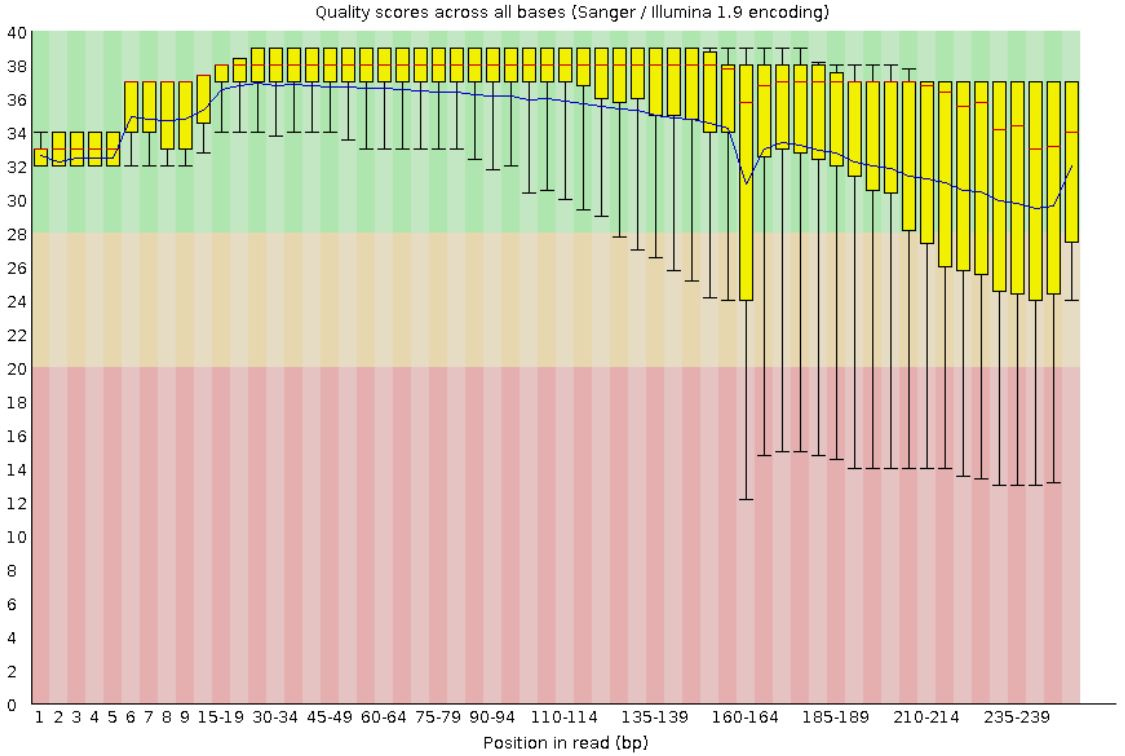
✓ All above Q20: Accuracy > 99%

Per base sequence quality

Ec005_R1



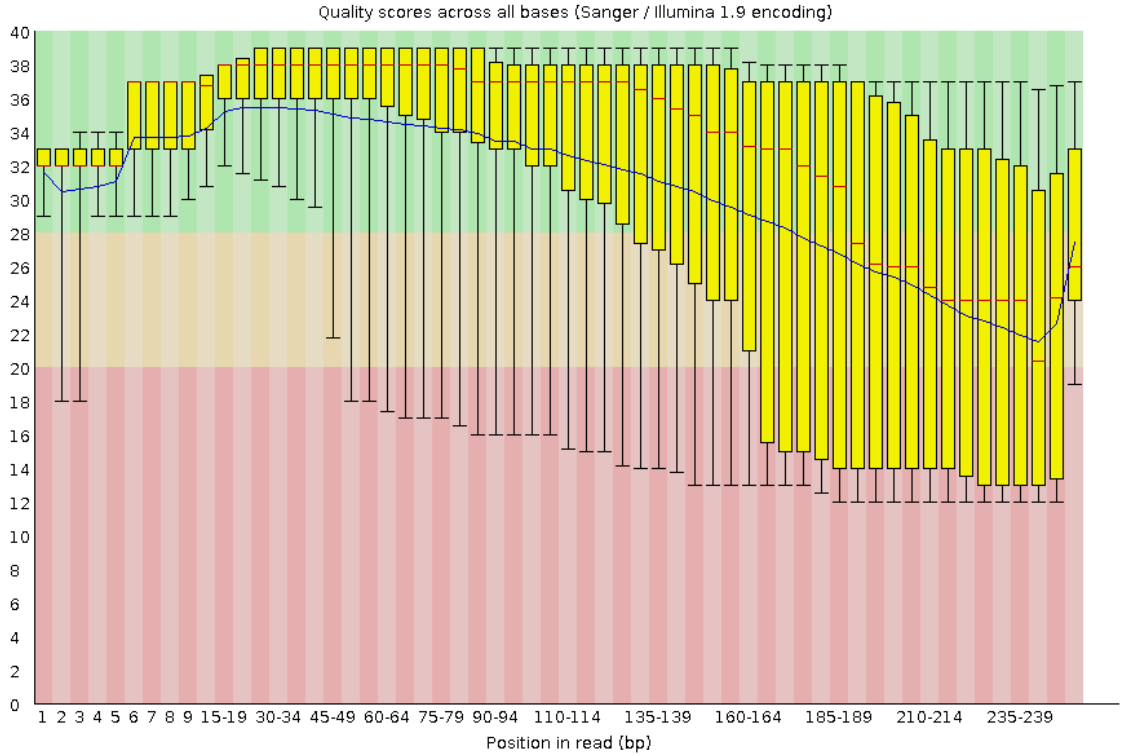
Ec005_R2



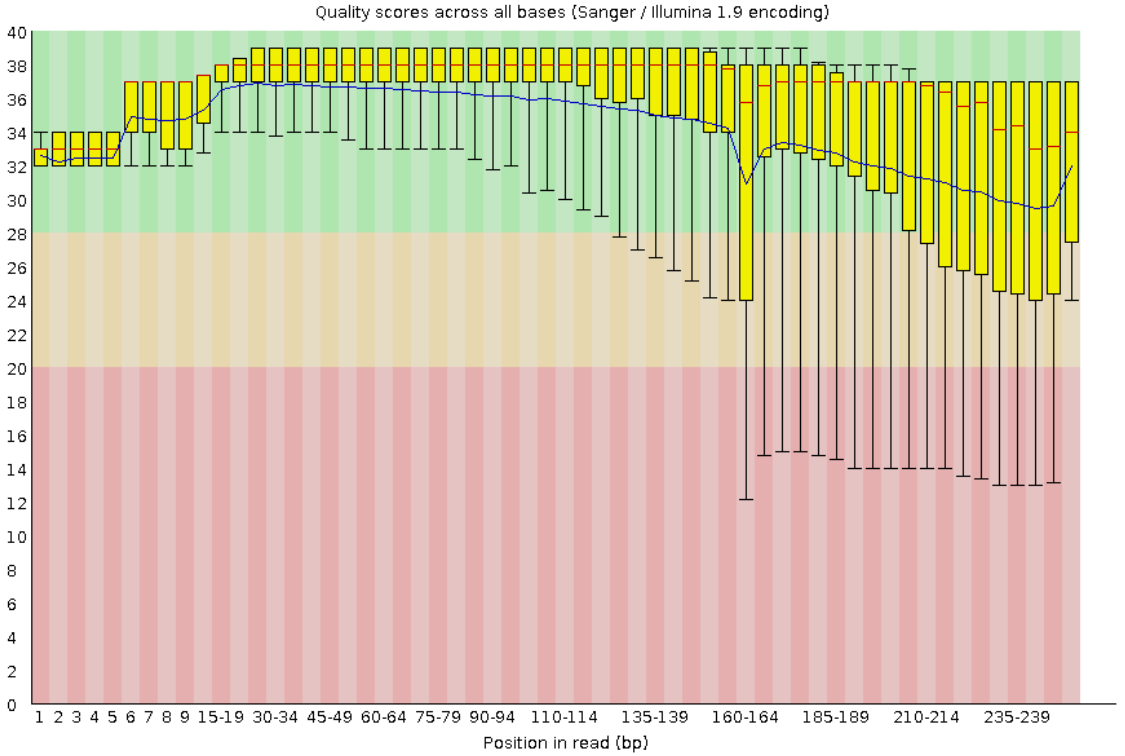
- ✓ All above Q20: Accuracy > 99%
- ✓ R2 above Q30: Accuracy > 99.9%

Per base sequence quality

Ec005_R1



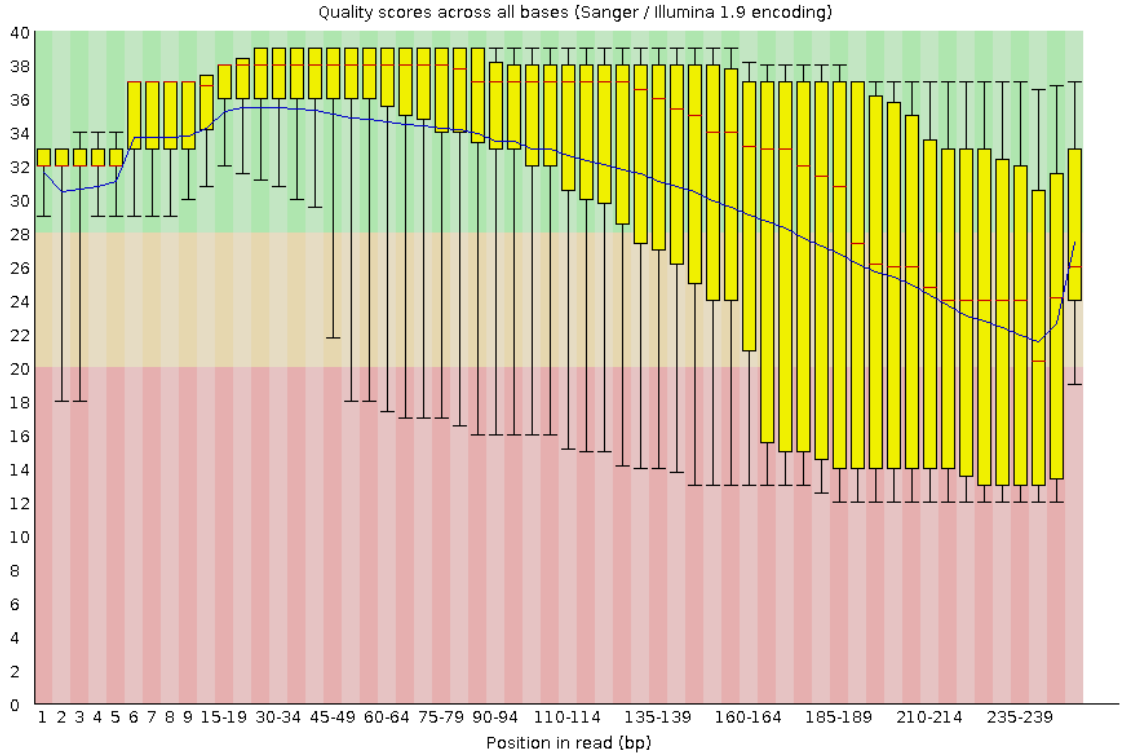
Ec005_R2



- ✓ All above Q20: Accuracy > 99%
- ✓ R2 above Q30: Accuracy > 99.9%
- ? R1: Very high quality distribution

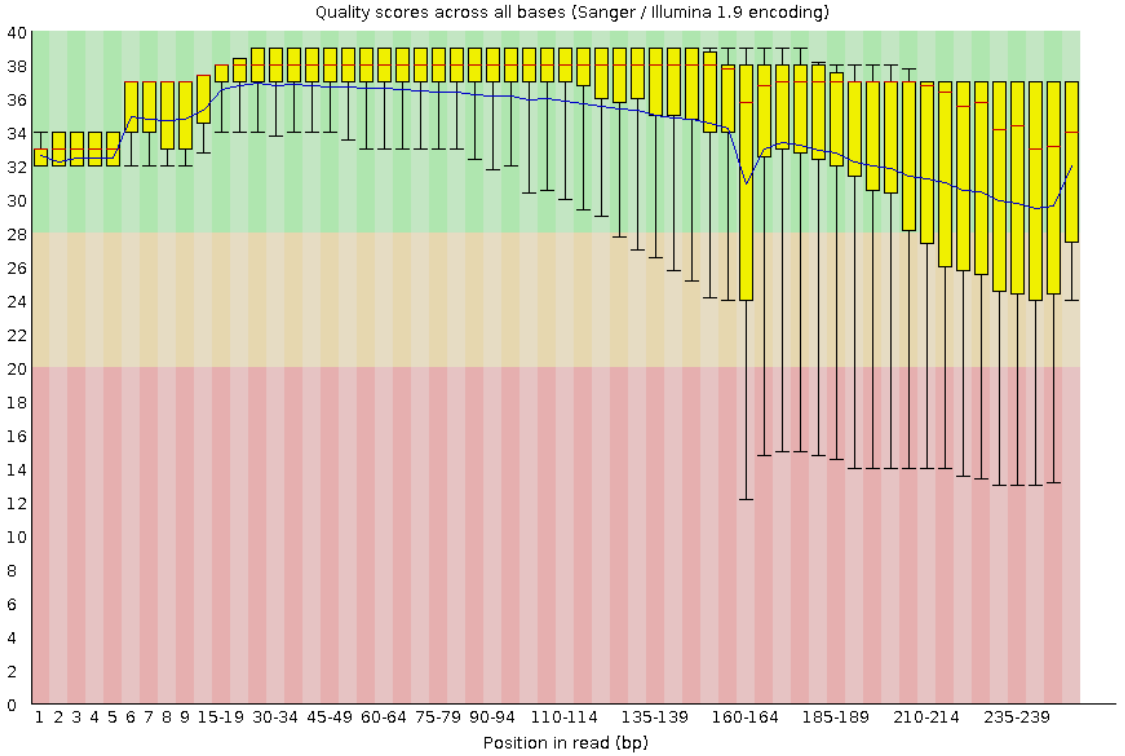
Per base sequence quality

Ec005_R1



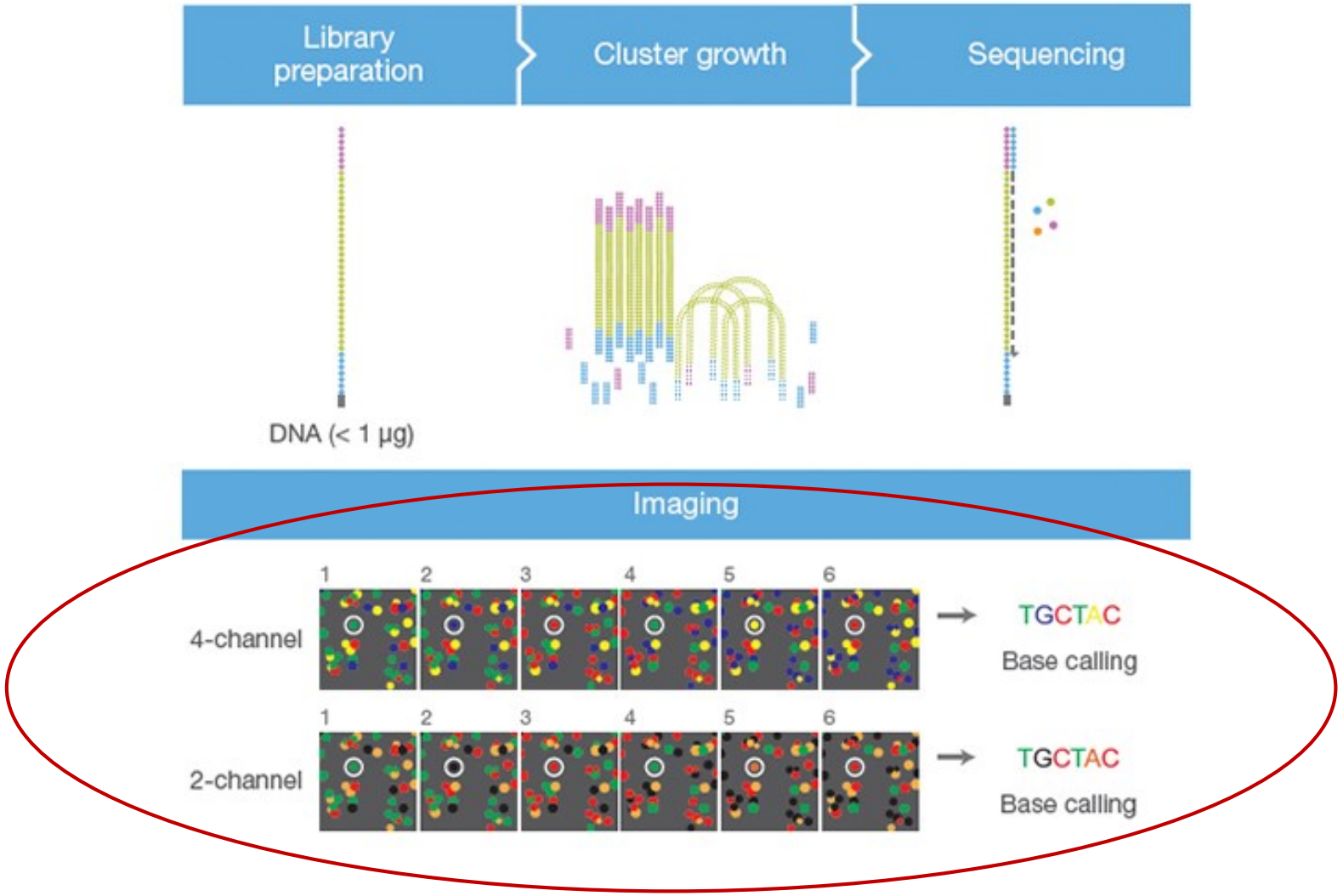
✓ All above Q20: Accuracy > 99%
✓ R2 above Q30: Accuracy > 99.9%
? R1: Very high quality distribution

Ec005_R2



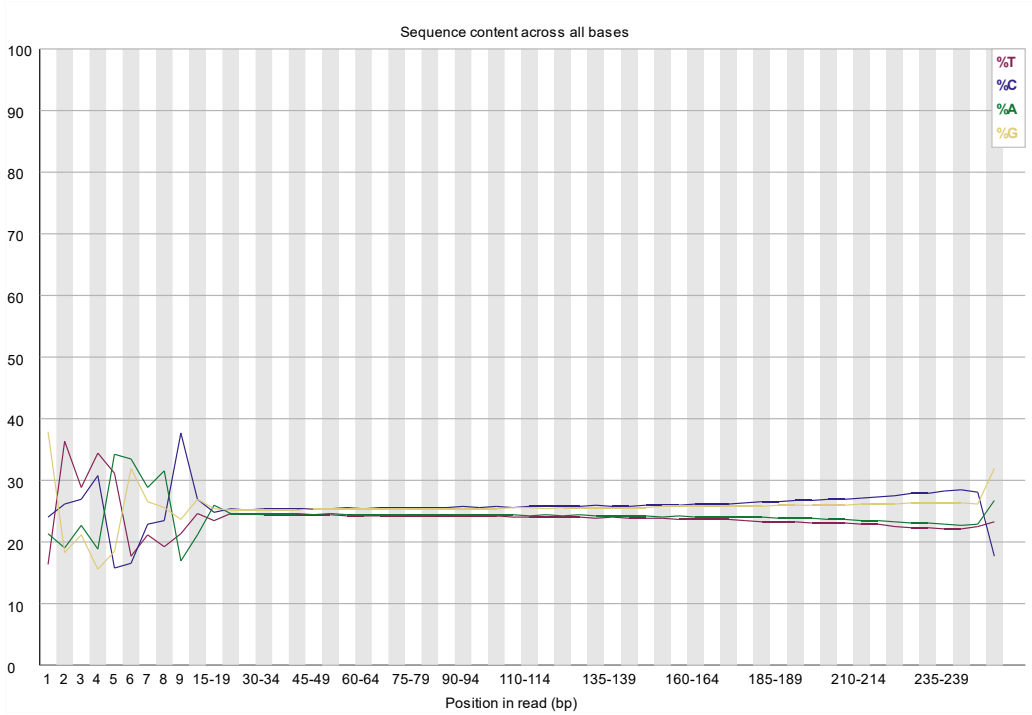
! R1: Not all base calls above Q30

Illumina cluster generation

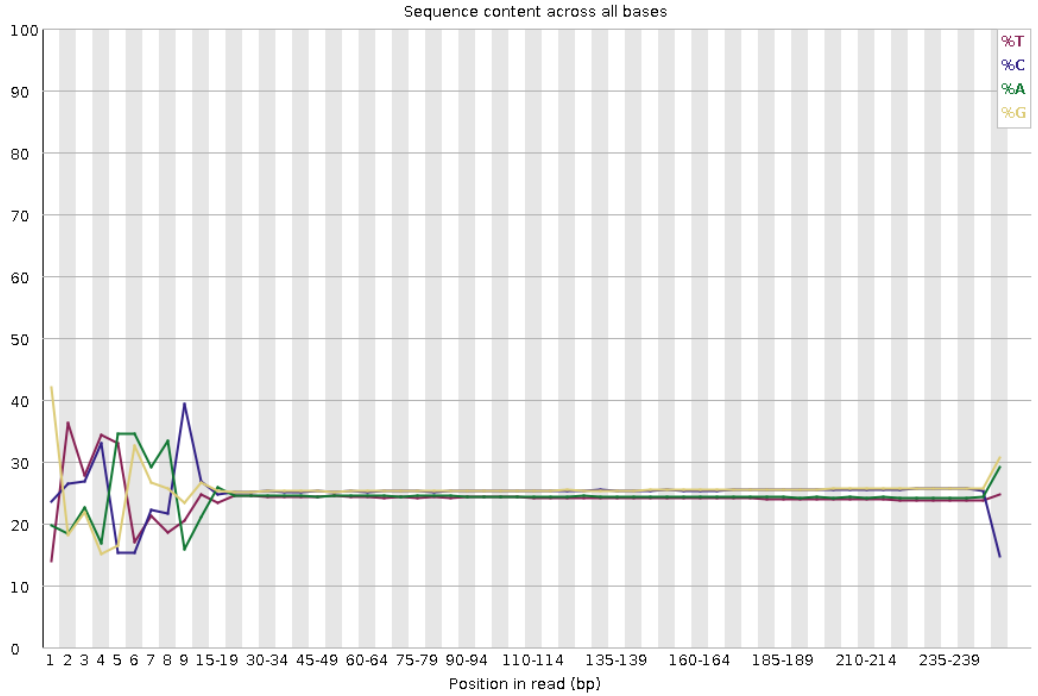


Per base sequence contents

Ec005_R1

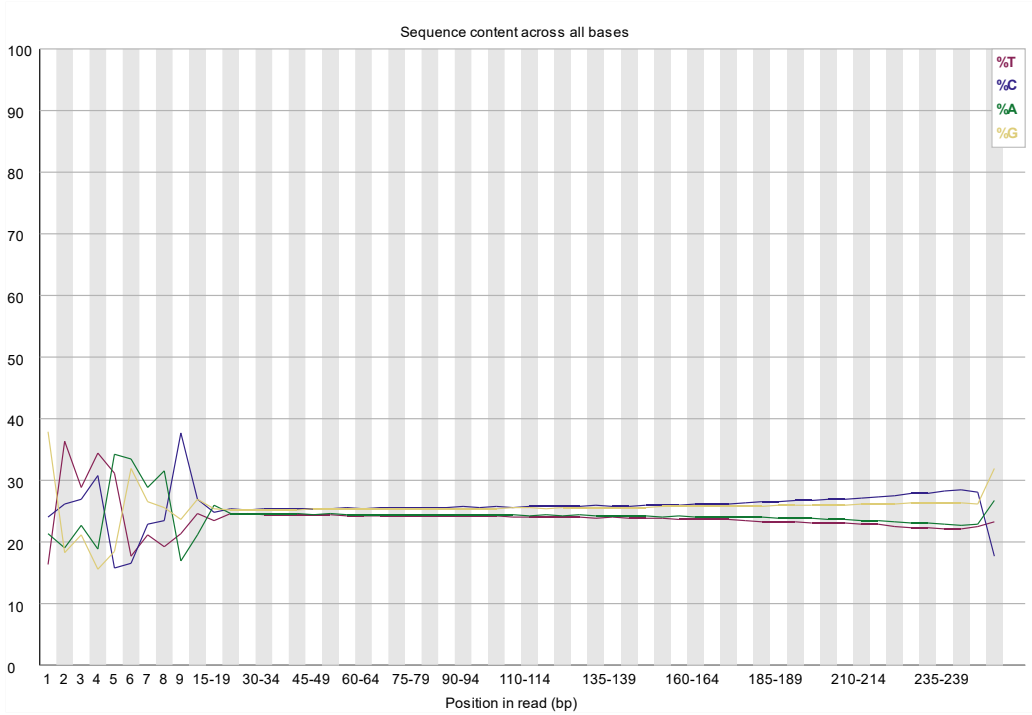


Ec005_R2

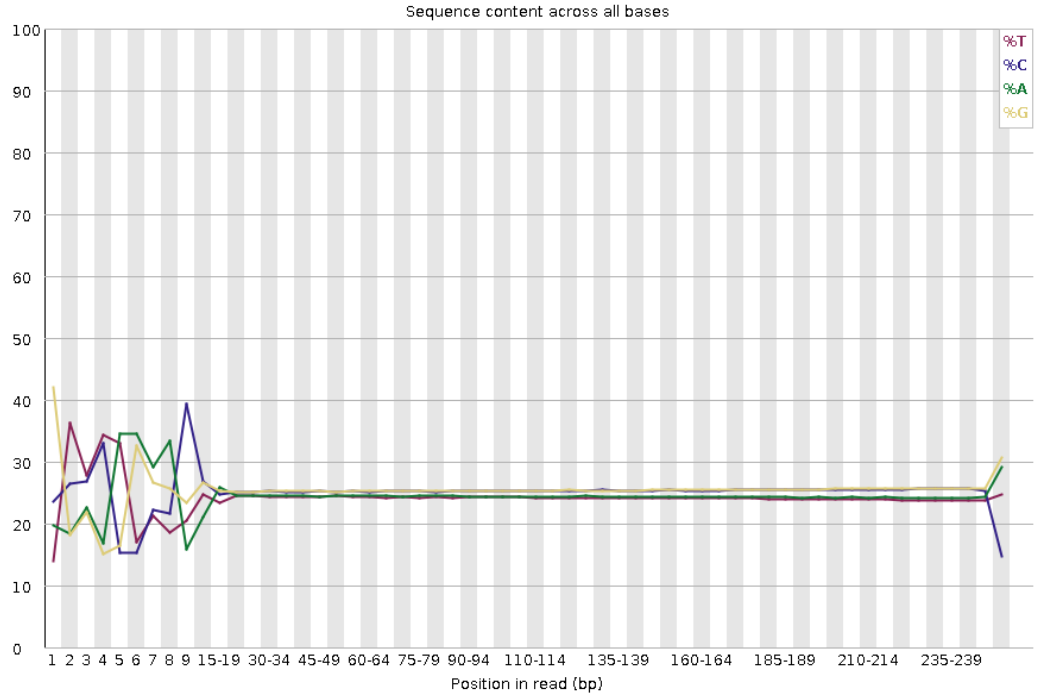


Per base sequence contents

Ec005_R1



Ec005_R2



✓ ACTG difference in beginning related to tagmantation

Illumina Nextera XT in summary

1) Sample prep

- gDNA extraction

- Pre-normalization

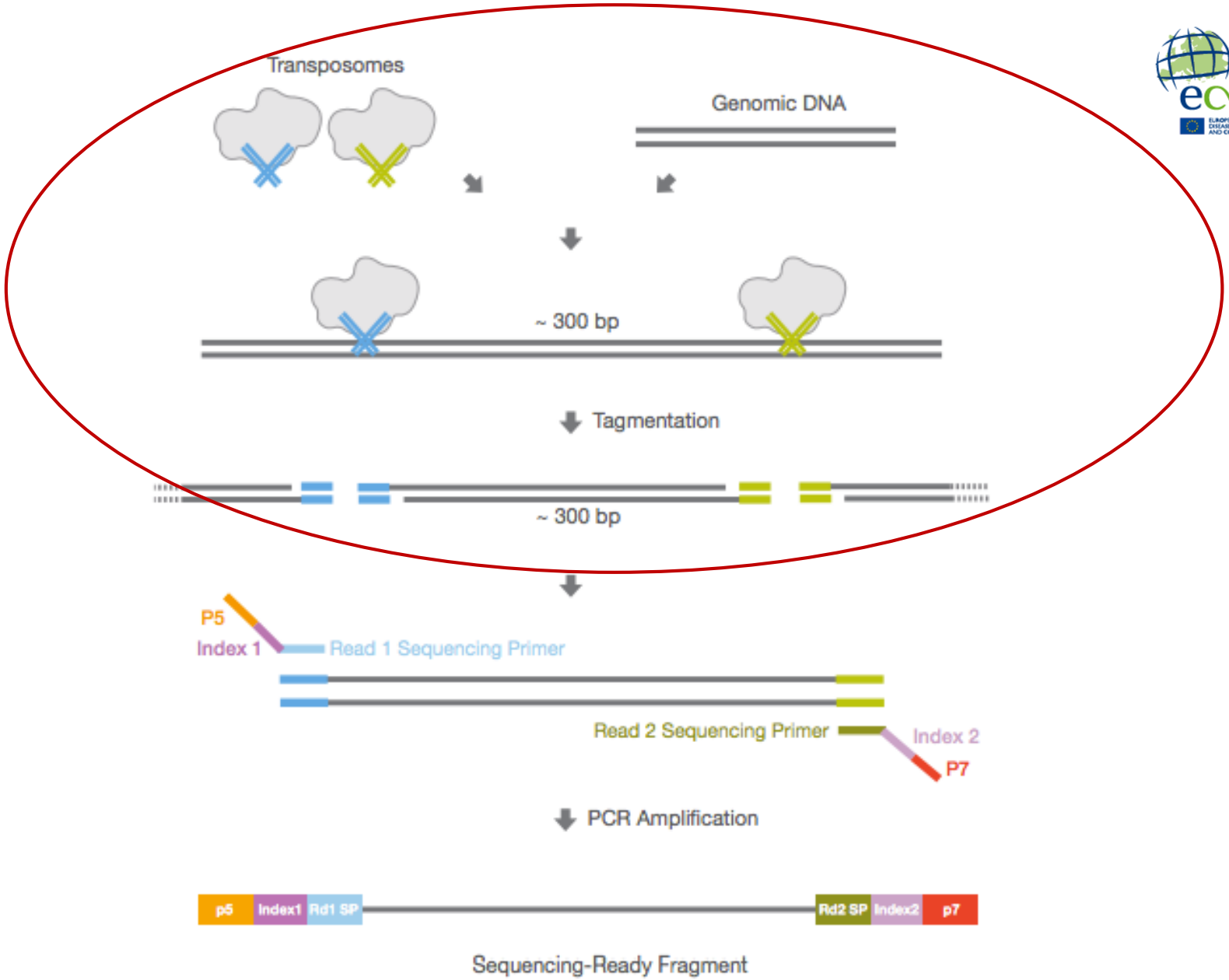
2) Library preparation

- Tagmentation

- Index PCR

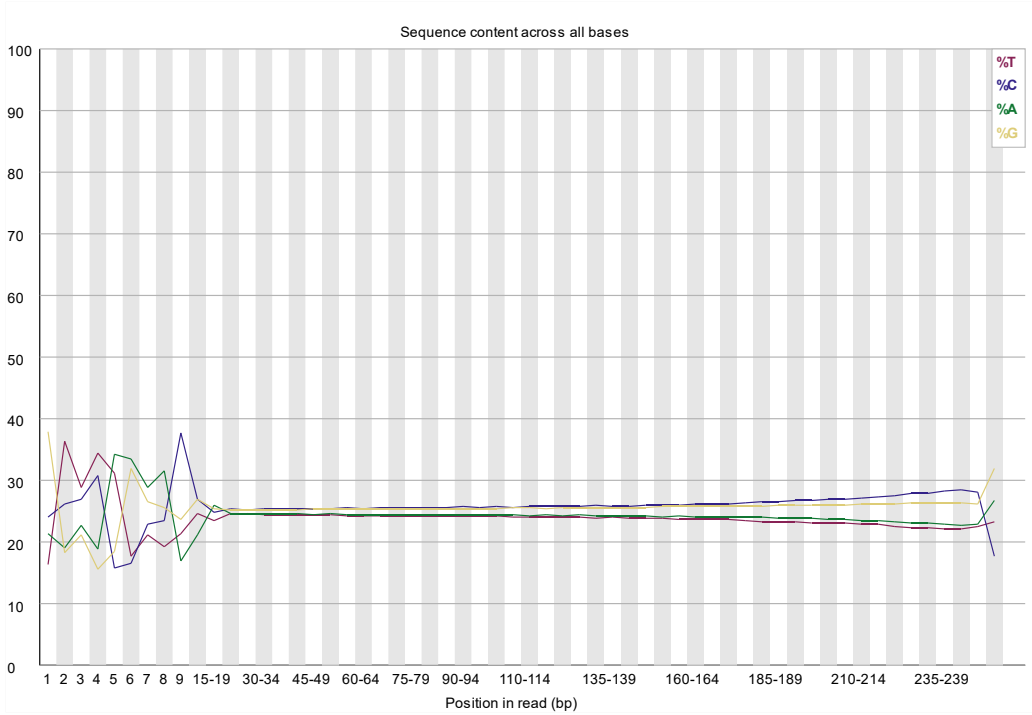
- Normalization and pool

3) Sequencing

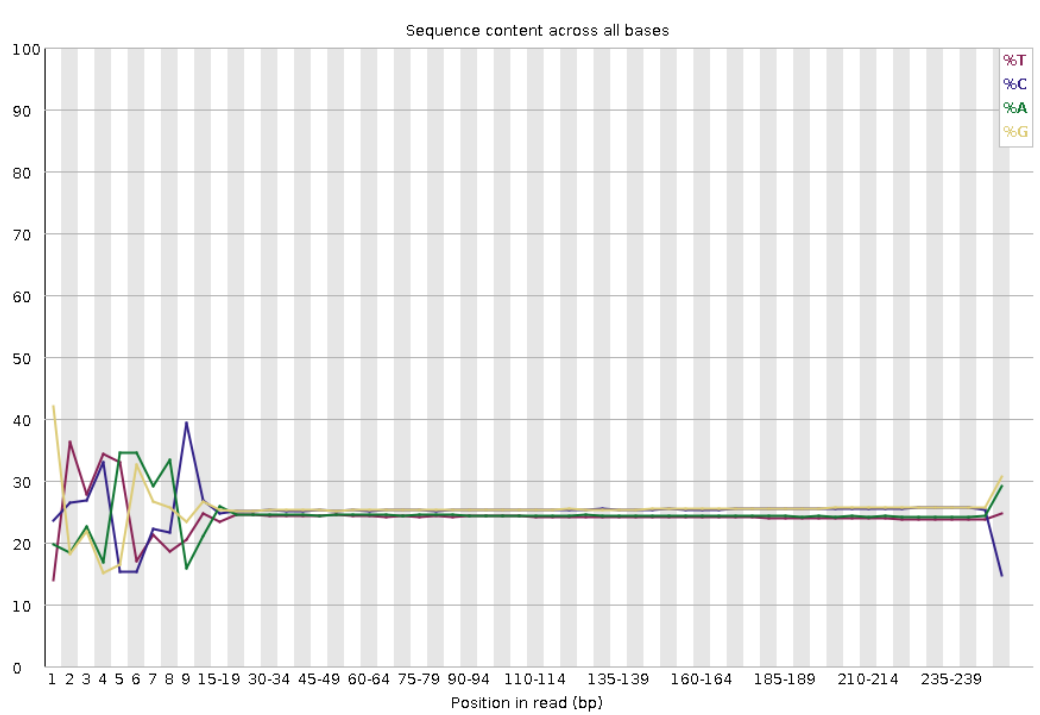


Per base sequence contents

Ec005_R1



Ec005_R2

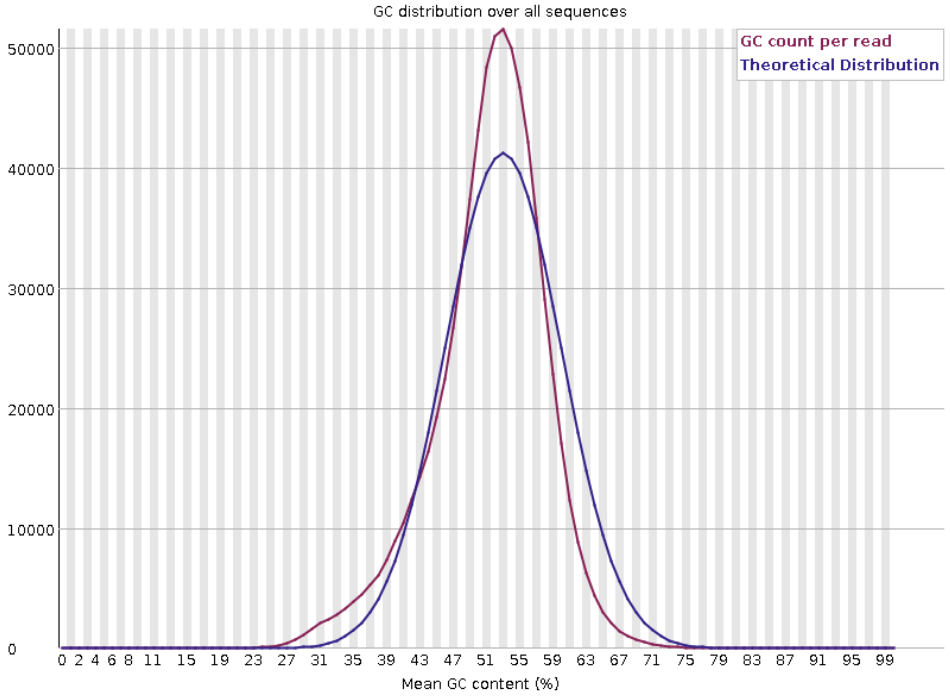


✓ ACTG difference in beginning related to tagmantation

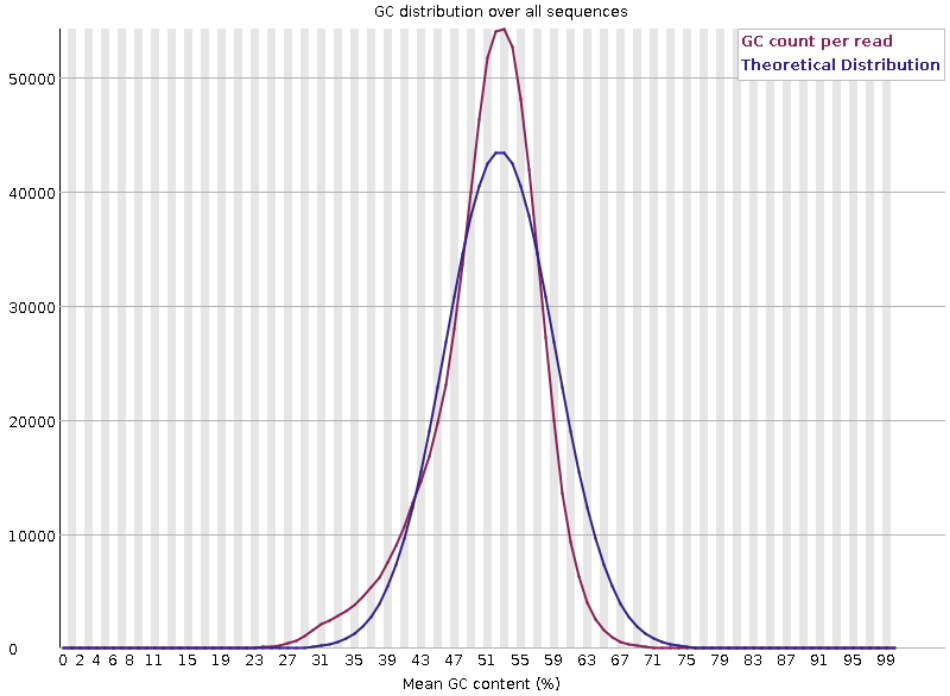
? Difference trails off in read ends

Per sequence GC contents

Ec005_R1

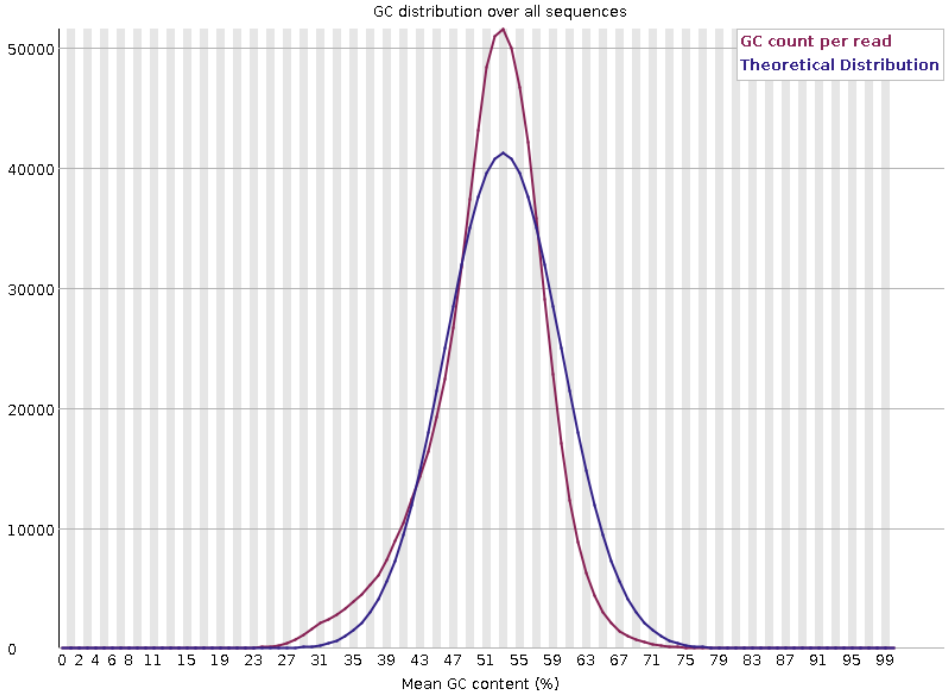


Ec005_R2

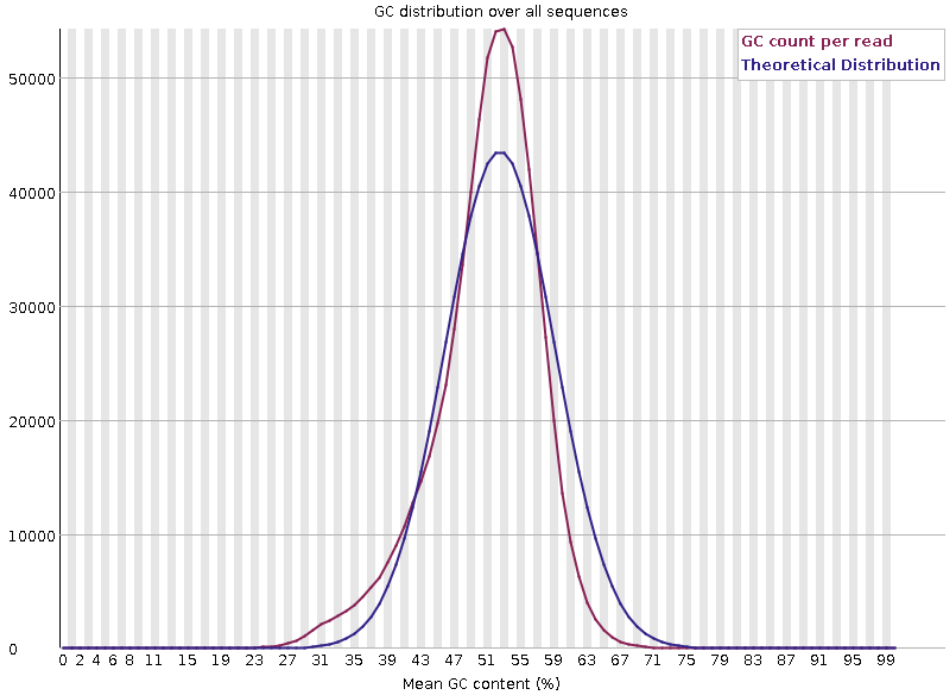


Per sequence GC contents

Ec005_R1



Ec005_R2

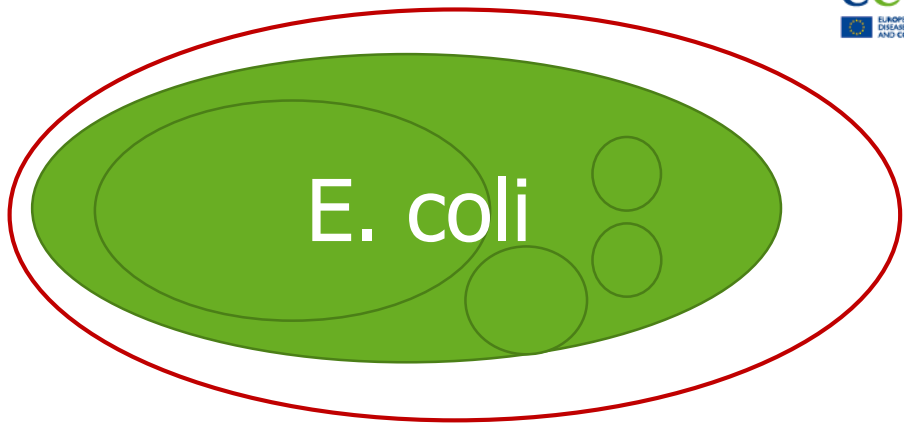


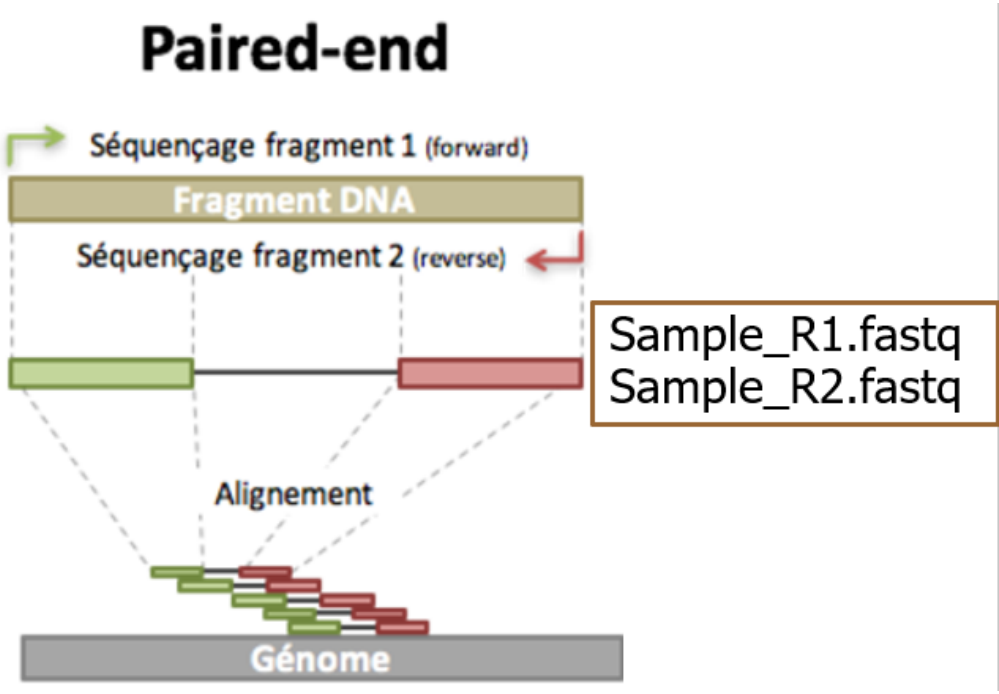
? Offset in GC contents relative to theoretical distribution

What to expect

Whole genome sequencing of E. coli

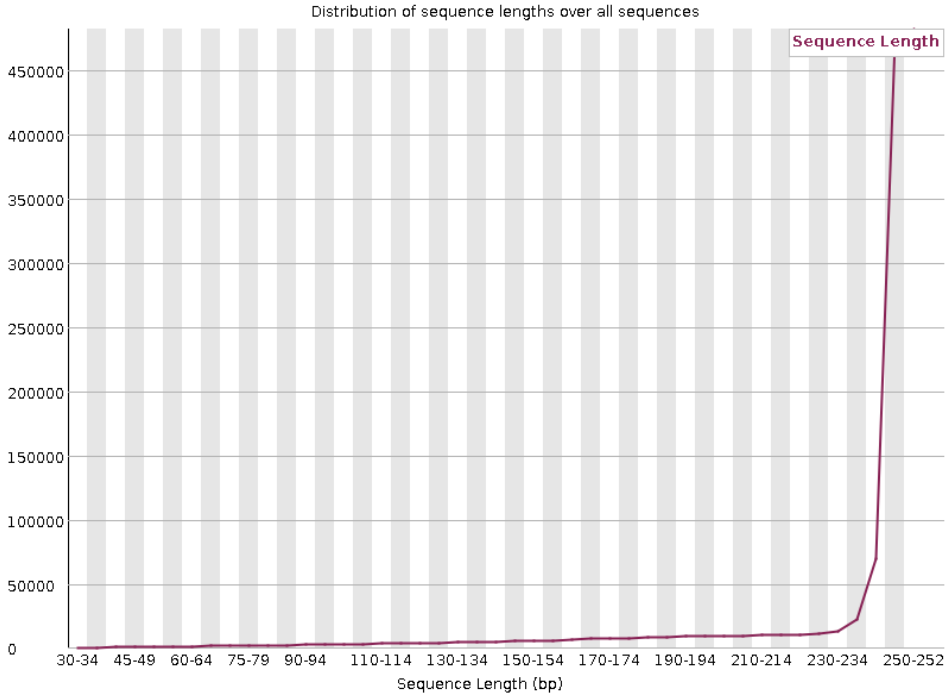
Illumina 2 x 250 bp paired end reads



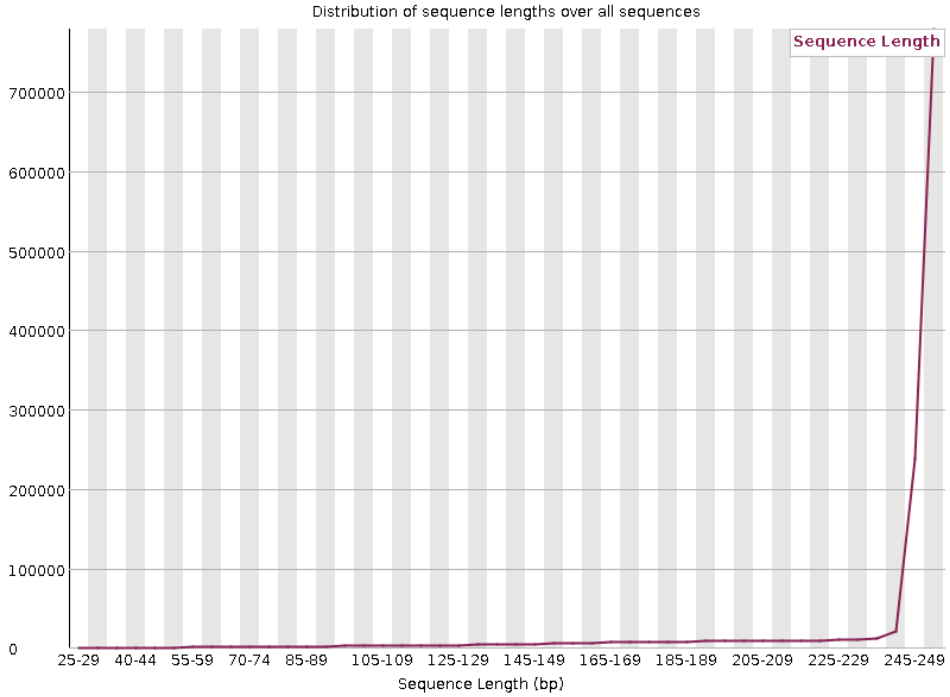


Sequence length distribution

Ec005_R1

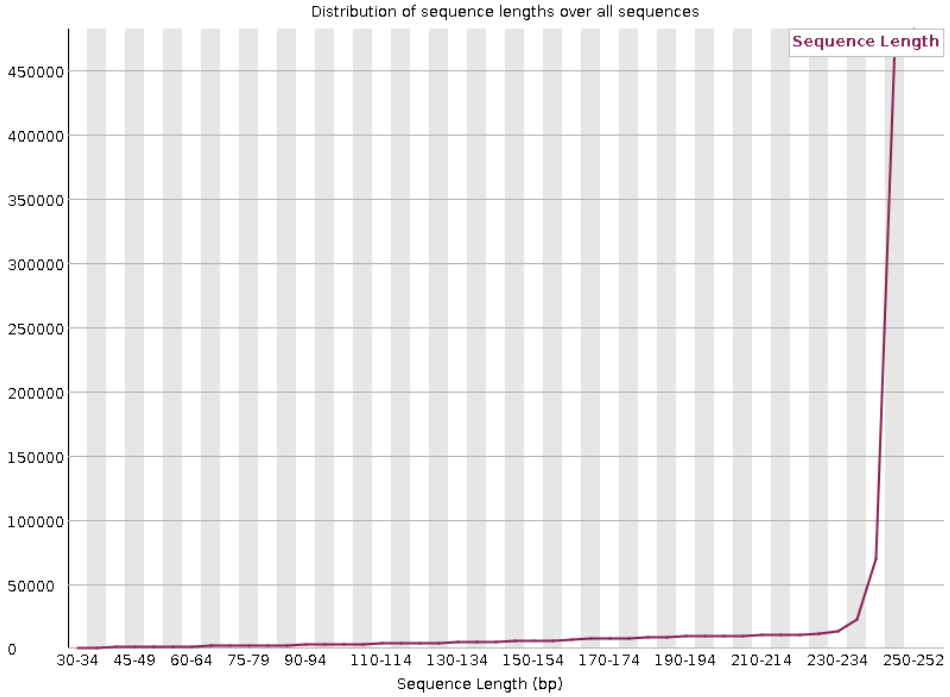


Ec005_R2

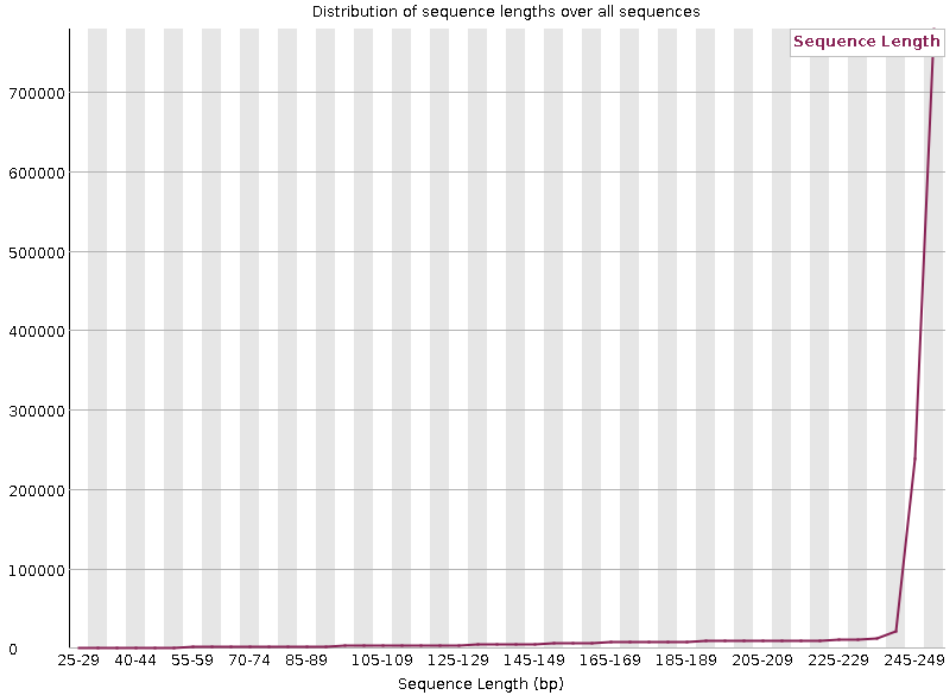


Sequence length distribution

Ec005_R1

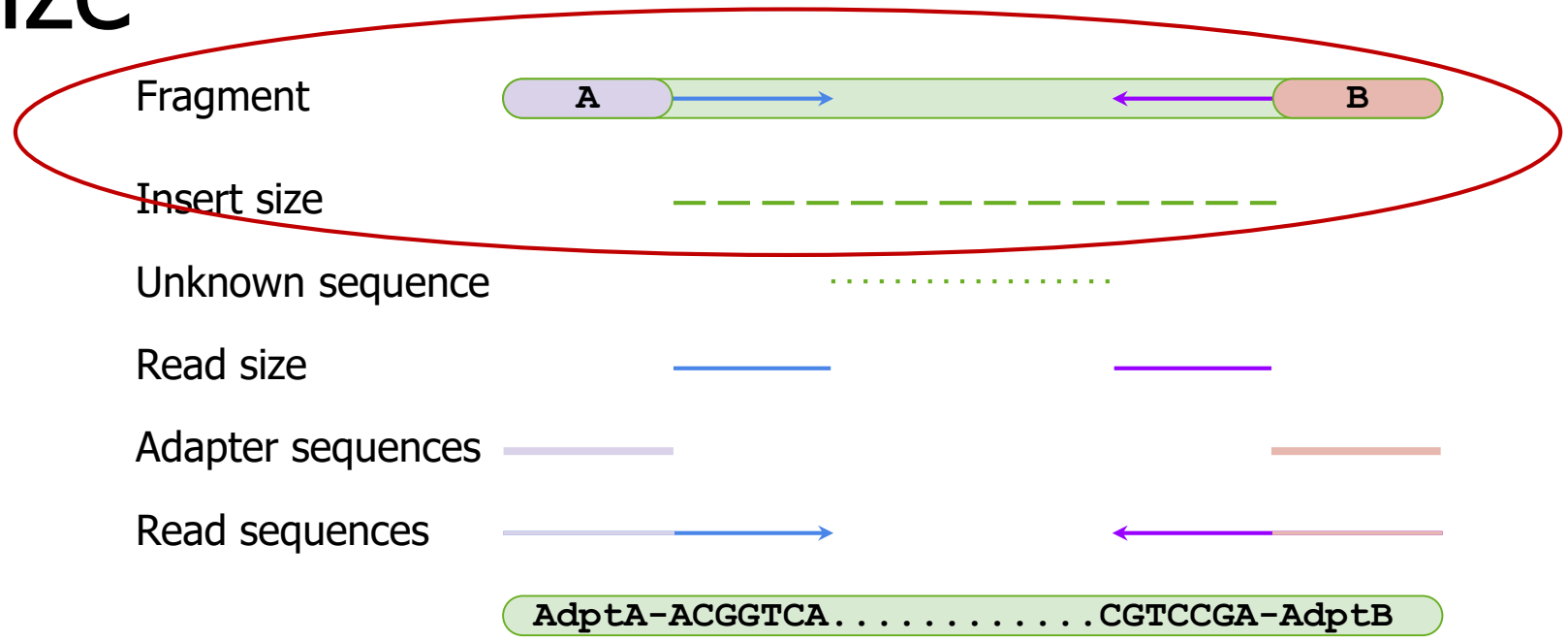


Ec005_R2



? Very short sequences does occur

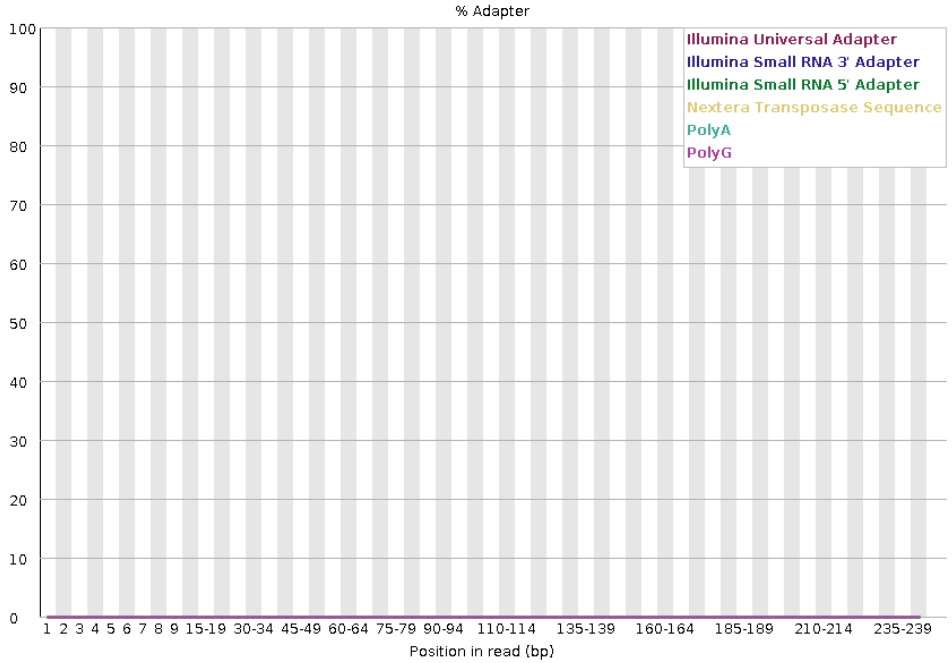
Read size



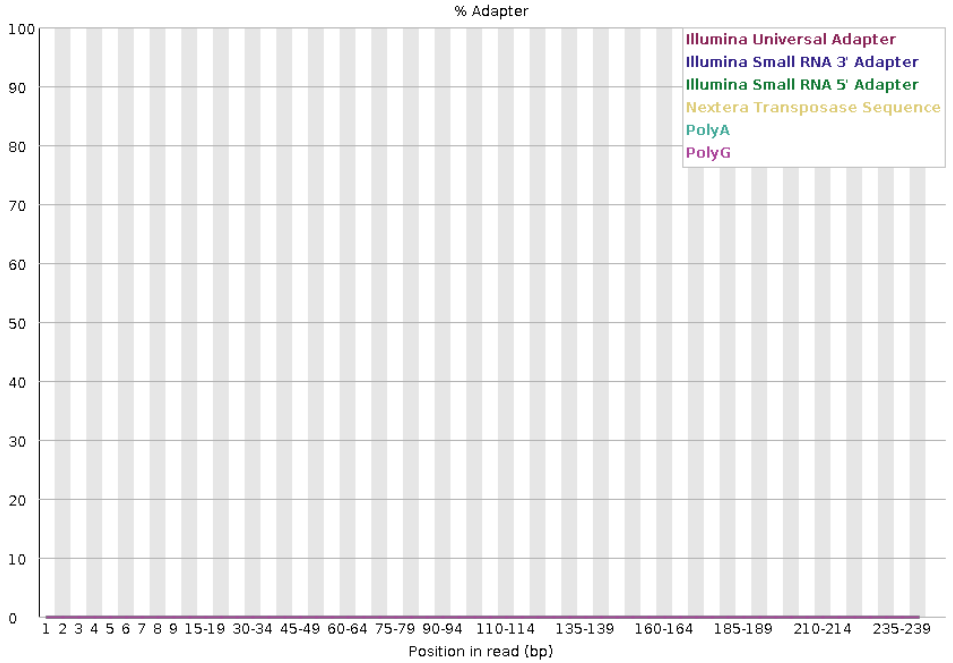

```
      @sample1_Mate1_readX
AdptA-ACGGTCA
      ...
      @sample1_Mate2_readX
AdptB-AGCCTGC
      …
      @sample1_Mate1_readY
```

Adapter contents

Ec005_R1

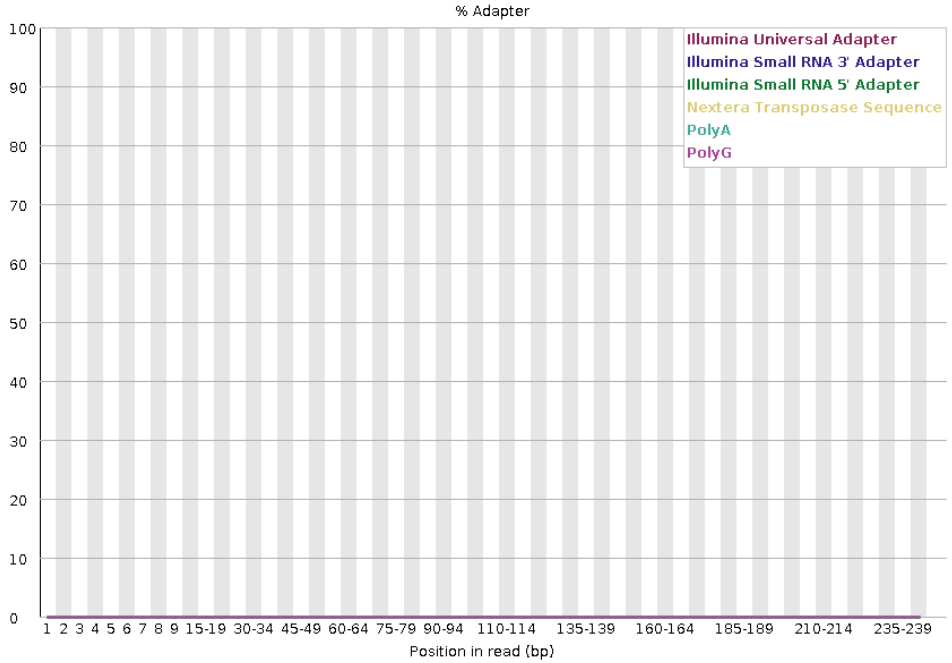


Ec005_R2

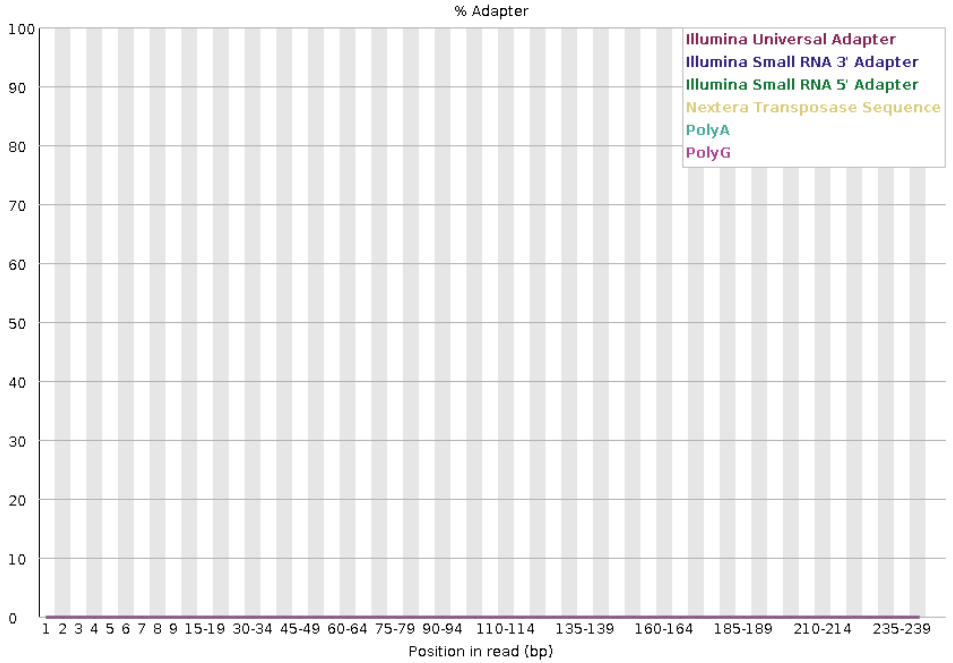


Adapter contents

Ec005_R1



Ec005_R2

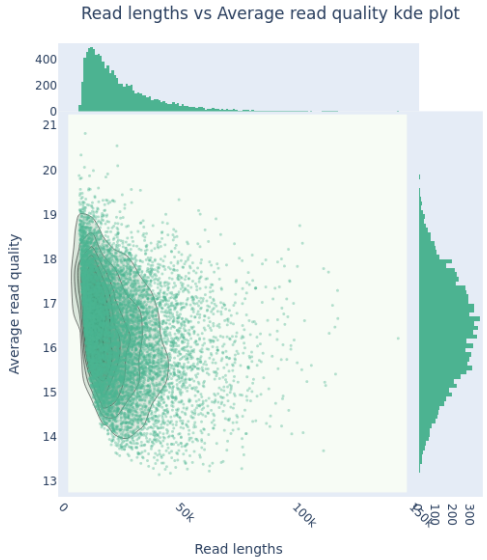


✓ No adapters detected

More information on Modules?

More information on Modules?

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/



Interpretation of results from Nanoplot on Nanopore data

NanoPlot recap

NanoPlot on Ec001

3. Run `NanoPlot`. You can run the simple command shown below but try to add more options.

(This should take a few minutes depending on the number of cores you have available.)

```
NanoPlot --fastq [Ec001_fastq_reads]
```

4. NanoPlot should now have generated a bunch of output files which we will examine further in Session 2.

How to run NanoPlot on multiple files

Bash For Loop

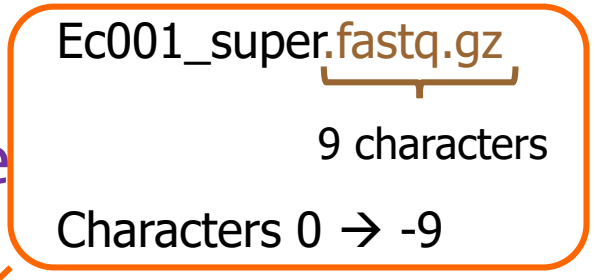
cd folder_with_nanopore_reads

Print all *.fastq.gz filenames in folder to screen:

```
for file in *.fastq.gz; do echo $file; done
```

Run NanoPlot on all *.fastq.gz files in folder

```
for file in *.fastq.gz; do NanoPlot -t 4 --fastq_rich $file --outdir "${file:0:-9}"_nanoplot; done
```

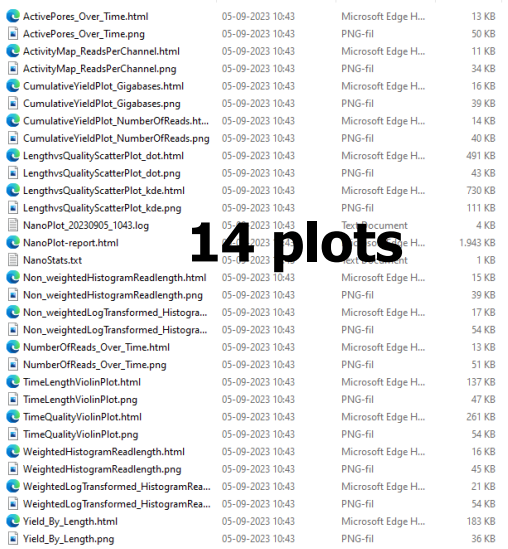


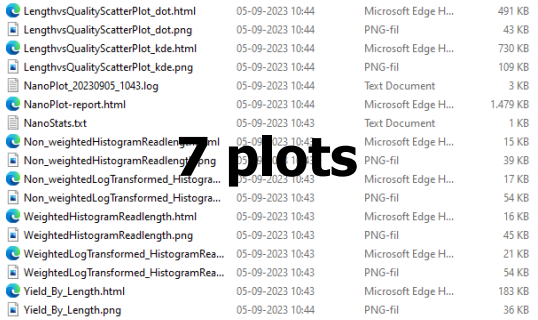
NanoPlot output

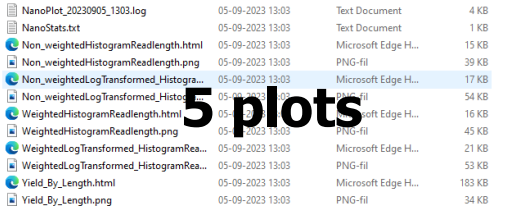
NanoPlot -t 4 --fastq_rich Ec001_super.fastq.gz --outdir Ec001_super_fastqrich_nanoplot

NanoPlot -t 4 --fastq Ec001_super.fastq.gz --outdir Ec001_super_fastq_nanoplot

NanoPlot -t 4 --fastq_minimal Ec001_super.fastq.gz --outdir Ec001_super_fastqminimal_nanoplot





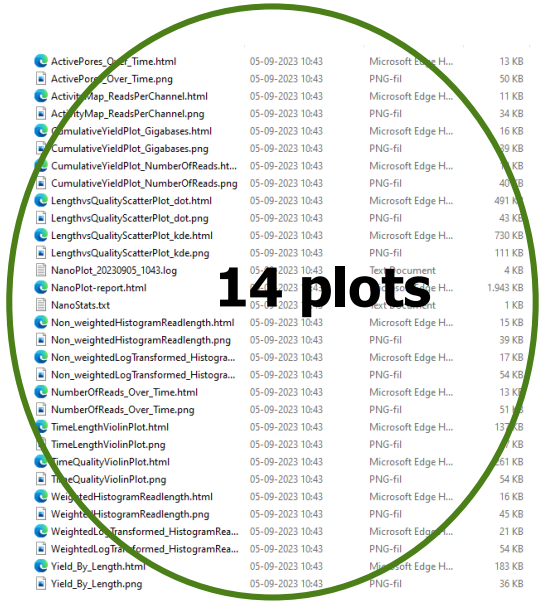


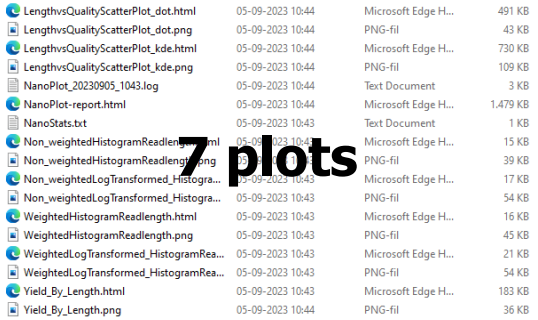
NanoPlot output

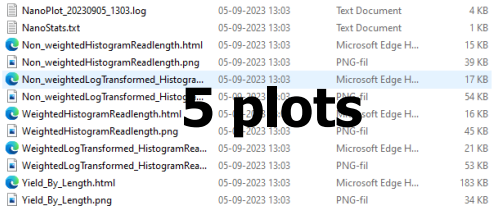
NanoPlot -t 4 --fastq_rich Ec001_super.fastq.gz --outdir Ec001_super_fastqrich_nanoplot

NanoPlot -t 4 --fastq Ec001_super.fastq.gz --outdir Ec001_super_fastq_nanoplot

NanoPlot -t 4 --fastq_minimal Ec001_super.fastq.gz --outdir Ec001_super_fastqminimal_nanoplot

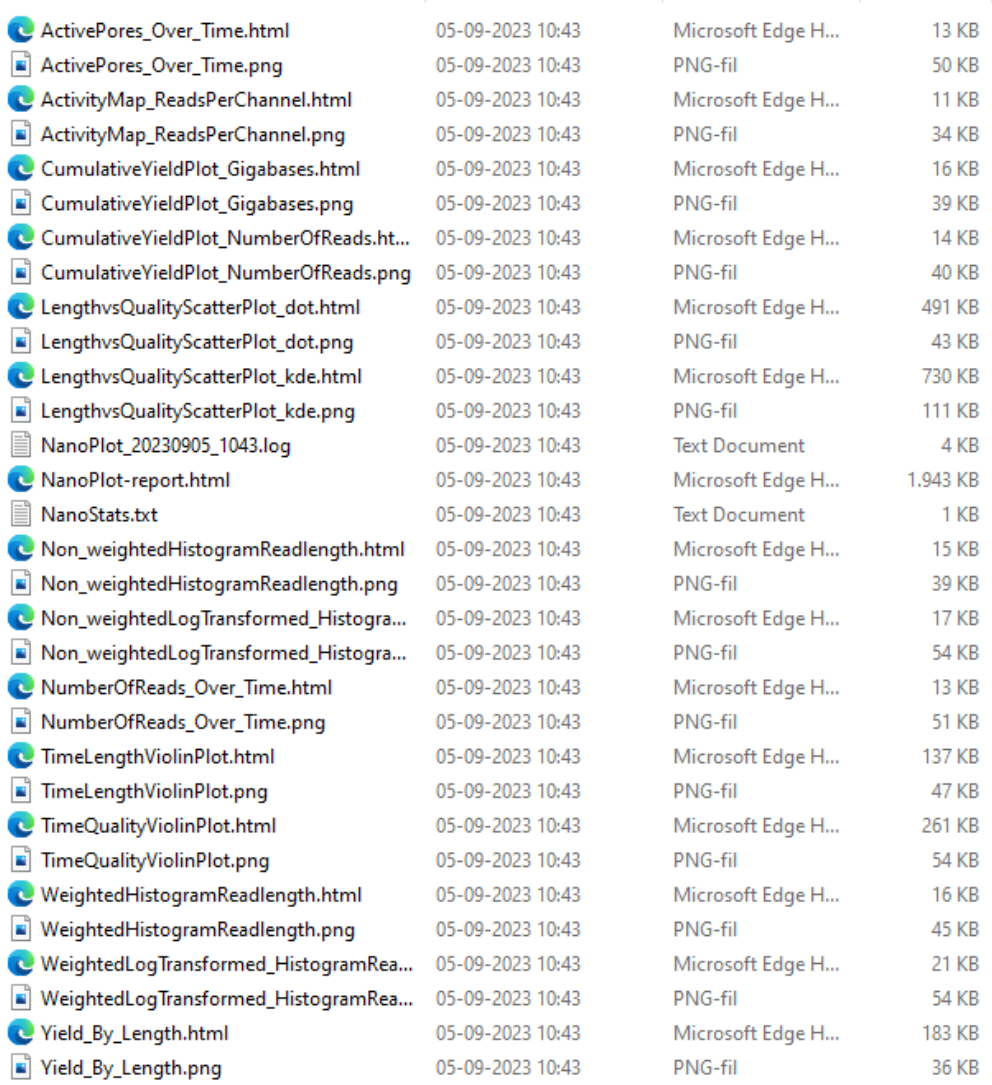






NanoPlot output

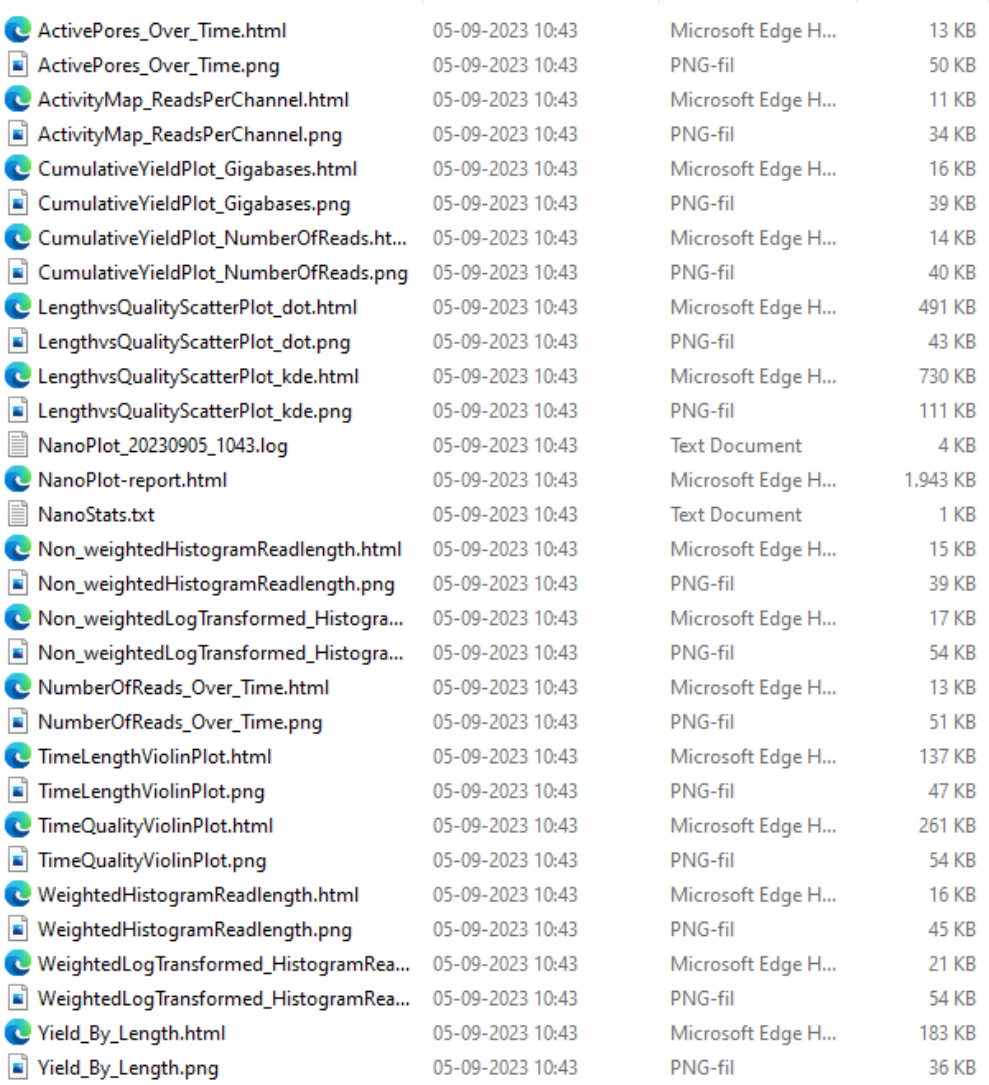
- 14 plots in html and png format
- Log file
- NanoStats.txt file
- NanoPlot-report.html file



ActivePores_Over_Time.html	05-09-2023 10:43	Microsoft Edge H...	13 KB
ActivePores_Over_Time.png	05-09-2023 10:43	PNG-fil	50 KB
ActivityMap_ReadsPerChannel.html	05-09-2023 10:43	Microsoft Edge H...	11 KB
ActivityMap_ReadsPerChannel.png	05-09-2023 10:43	PNG-fil	34 KB
CumulativeYieldPlot_Gigabases.html	05-09-2023 10:43	Microsoft Edge H...	16 KB
CumulativeYieldPlot_Gigabases.png	05-09-2023 10:43	PNG-fil	39 KB
CumulativeYieldPlot_NumberOfReads.ht...	05-09-2023 10:43	Microsoft Edge H...	14 KB
CumulativeYieldPlot_NumberOfReads.png	05-09-2023 10:43	PNG-fil	40 KB
LengthvsQualityScatterPlot_dot.html	05-09-2023 10:43	Microsoft Edge H...	491 KB
LengthvsQualityScatterPlot_dot.png	05-09-2023 10:43	PNG-fil	43 KB
LengthvsQualityScatterPlot_kde.html	05-09-2023 10:43	Microsoft Edge H...	730 KB
LengthvsQualityScatterPlot_kde.png	05-09-2023 10:43	PNG-fil	111 KB
NanoPlot_20230905_1043.log	05-09-2023 10:43	Text Document	4 KB
NanoPlot-report.html	05-09-2023 10:43	Microsoft Edge H...	1.943 KB
NanoStats.txt	05-09-2023 10:43	Text Document	1 KB
Non_weightedHistogramReadlength.html	05-09-2023 10:43	Microsoft Edge H...	15 KB
Non_weightedHistogramReadlength.png	05-09-2023 10:43	PNG-fil	39 KB
Non_weightedLogTransformed_Histogra...	05-09-2023 10:43	Microsoft Edge H...	17 KB
Non_weightedLogTransformed_Histogra...	05-09-2023 10:43	PNG-fil	54 KB
NumberOfReads_Over_Time.html	05-09-2023 10:43	Microsoft Edge H...	13 KB
NumberOfReads_Over_Time.png	05-09-2023 10:43	PNG-fil	51 KB
TimeLengthViolinPlot.html	05-09-2023 10:43	Microsoft Edge H...	137 KB
TimeLengthViolinPlot.png	05-09-2023 10:43	PNG-fil	47 KB
TimeQualityViolinPlot.html	05-09-2023 10:43	Microsoft Edge H...	261 KB
TimeQualityViolinPlot.png	05-09-2023 10:43	PNG-fil	54 KB
WeightedHistogramReadlength.html	05-09-2023 10:43	Microsoft Edge H...	16 KB
WeightedHistogramReadlength.png	05-09-2023 10:43	PNG-fil	45 KB
WeightedLogTransformed_HistogramRea...	05-09-2023 10:43	Microsoft Edge H...	21 KB
WeightedLogTransformed_HistogramRea...	05-09-2023 10:43	PNG-fil	54 KB
Yield_By_Length.html	05-09-2023 10:43	Microsoft Edge H...	183 KB
Yield_By_Length.png	05-09-2023 10:43	PNG-fil	36 KB

NanoPlot output

- 14 plots in html and png format
- Log file
- NanoStats.txt file
- NanoPlot-report.html file



ActivePores_Over_Time.html	05-09-2023 10:43	Microsoft Edge H...	13 KB
ActivePores_Over_Time.png	05-09-2023 10:43	PNG-fil	50 KB
ActivityMap_ReadsPerChannel.html	05-09-2023 10:43	Microsoft Edge H...	11 KB
ActivityMap_ReadsPerChannel.png	05-09-2023 10:43	PNG-fil	34 KB
CumulativeYieldPlot_Gigabases.html	05-09-2023 10:43	Microsoft Edge H...	16 KB
CumulativeYieldPlot_Gigabases.png	05-09-2023 10:43	PNG-fil	39 KB
CumulativeYieldPlot_NumberOfReads.ht...	05-09-2023 10:43	Microsoft Edge H...	14 KB
CumulativeYieldPlot_NumberOfReads.png	05-09-2023 10:43	PNG-fil	40 KB
LengthvsQualityScatterPlot_dot.html	05-09-2023 10:43	Microsoft Edge H...	491 KB
LengthvsQualityScatterPlot_dot.png	05-09-2023 10:43	PNG-fil	43 KB
LengthvsQualityScatterPlot_kde.html	05-09-2023 10:43	Microsoft Edge H...	730 KB
LengthvsQualityScatterPlot_kde.png	05-09-2023 10:43	PNG-fil	111 KB
NanoPlot_20230905_1043.log	05-09-2023 10:43	Text Document	4 KB
NanoPlot-report.html	05-09-2023 10:43	Microsoft Edge H...	1.943 KB
NanoStats.txt	05-09-2023 10:43	Text Document	1 KB
Non_weightedHistogramReadlength.html	05-09-2023 10:43	Microsoft Edge H...	15 KB
Non_weightedHistogramReadlength.png	05-09-2023 10:43	PNG-fil	39 KB
Non_weightedLogTransformed_Histogra...	05-09-2023 10:43	Microsoft Edge H...	17 KB
Non_weightedLogTransformed_Histogra...	05-09-2023 10:43	PNG-fil	54 KB
NumberOfReads_Over_Time.html	05-09-2023 10:43	Microsoft Edge H...	13 KB
NumberOfReads_Over_Time.png	05-09-2023 10:43	PNG-fil	51 KB
TimeLengthViolinPlot.html	05-09-2023 10:43	Microsoft Edge H...	137 KB
TimeLengthViolinPlot.png	05-09-2023 10:43	PNG-fil	47 KB
TimeQualityViolinPlot.html	05-09-2023 10:43	Microsoft Edge H...	261 KB
TimeQualityViolinPlot.png	05-09-2023 10:43	PNG-fil	54 KB
WeightedHistogramReadlength.html	05-09-2023 10:43	Microsoft Edge H...	16 KB
WeightedHistogramReadlength.png	05-09-2023 10:43	PNG-fil	45 KB
WeightedLogTransformed_HistogramRea...	05-09-2023 10:43	Microsoft Edge H...	21 KB
WeightedLogTransformed_HistogramRea...	05-09-2023 10:43	PNG-fil	54 KB
Yield_By_Length.html	05-09-2023 10:43	Microsoft Edge H...	183 KB
Yield_By_Length.png	05-09-2023 10:43	PNG-fil	36 KB

NanoPlot log file

- Filename will tell you when file was created

- Info inside file:
  - Options used
  - NanoPlot version
  - Input filename
  - Plots created

```
2023-09-05 10:43:01,834 NanoPlot 1.41.6 started with arguments Namespace(threads=2, verbose=False, store=False, raw=False, huge=False, outdir='Ec001_super_nanoplot', no_static=False, prefix='', tsv_stats=False,
info_in_report=False, maxlength=None, minlength=None, drop_outliers=False, downsample=None, loglength=False, percentqual=False, alength=False, minqual=None, runtime_until=None, readtype='1D', barcoded=False,
no_supplementary=False, color='#4CB391', colormap='Greens', format=['png'], plots=['kde', 'dot'], legacy=None, listcolors=False, listcolormaps=False, no_N50=False, N50=True, title=None, font_scale=1, dpi=100,
hide_stats=False, fastq=None, fasta=None, fastq_rich=['Ec001_super.fastq.gz'], fastq_minimal=None, summary=None, bam=None, ubam=None, cram=None, pickle=None, feather=None, path='Ec001_super_nanoplot/')
2023-09-05 10:43:01,834 Python version is: 3.10.12 | packaged by conda-forge | (main, Jun 23 2023, 22:40:32) [GCC 12.3.0]
2023-09-05 10:43:01,959 Nanoget: Starting to collect statistics from rich fastq file.
2023-09-05 10:43:01,968 Nanoget: Decompressing gzipped fastq Ec001_super.fastq.gz
2023-09-05 10:43:37,341 Reduced DataFrame memory usage from 2.4891433715820312Mb to 2.4891433715820312Mb
2023-09-05 10:43:37,362 Nanoget: Gathered all metrics of 20232 reads
2023-09-05 10:43:37,383 Calculated statistics
2023-09-05 10:43:37,384 Using sequenced read lengths for plotting.
2023-09-05 10:43:37,389 NanoPlot:  Valid color #4CB391.
2023-09-05 10:43:37,390 NanoPlot:  Valid colormap Greens.
2023-09-05 10:43:37,392 NanoPlot:  Creating length plots for Read length.
2023-09-05 10:43:37,393 NanoPlot: Using 20232 reads with read length N50 of 29257bp and maximum of 151104bp.
2023-09-05 10:43:39,351 Saved Ec001_super_nanoplot/WeightedHistogramReadlength  as png (or png for --legacy)
2023-09-05 10:43:40,354 Saved Ec001_super_nanoplot/WeightedLogTransformed_HistogramReadlength  as png (or png for --legacy)
2023-09-05 10:43:40,904 Saved Ec001_super_nanoplot/Non_weightedHistogramReadlength  as png (or png for --legacy)
2023-09-05 10:43:41,429 Saved Ec001_super_nanoplot/Non_weightedLogTransformed_HistogramReadlength  as png (or png for --legacy)
2023-09-05 10:43:45,964 Saved Ec001_super_nanoplot/Yield_By_Length  as png (or png for --legacy)
2023-09-05 10:43:45,966 Created length plots
2023-09-05 10:43:45,967 NanoPlot: Creating Read lengths vs Average read quality plots using 20232 reads.
2023-09-05 10:43:46,993 Saved Ec001_super_nanoplot/LengthvsQualityScatterPlot_dot  as png (or png for --legacy)
2023-09-05 10:43:48,520 Saved Ec001_super_nanoplot/LengthvsQualityScatterPlot_kde  as png (or png for --legacy)
2023-09-05 10:43:48,525 Created LengthvsQual plot
2023-09-05 10:43:48,525 Nanoplotter: Creating heatmap of reads per channel using 20232 reads.
2023-09-05 10:43:49,029 Saved Ec001_super_nanoplot/ActivityMap_ReadsPerChannel  as png (or png for --legacy)
2023-09-05 10:43:49,036 Created spatialheatmap for succesfull basecalls.
2023-09-05 10:43:49,036 Nanoplotter: Creating timeplots using 20232 (full) or 10000 (subsampled dataset) reads.
2023-09-05 10:43:49,615 Saved Ec001_super_nanoplot/CumulativeYieldPlot_Gigabases  as png (or png for --legacy)
2023-09-05 10:43:50,165 Saved Ec001_super_nanoplot/CumulativeYieldPlot_NumberOfReads  as png (or png for --legacy)
2023-09-05 10:43:50,989 Saved Ec001_super_nanoplot/NumberOfReads_Over_Time  as png (or png for --legacy)
2023-09-05 10:43:51,857 Saved Ec001_super_nanoplot/ActivePores_Over_Time  as png (or png for --legacy)
2023-09-05 10:43:52,561 Saved Ec001_super_nanoplot/TimeLengthViolinPlot  as png (or png for --legacy)
2023-09-05 10:43:53,159 Saved Ec001_super_nanoplot/TimeQualityViolinPlot  as png (or png for --legacy)
2023-09-05 10:43:53,163 Created timeplots.
2023-09-05 10:43:53,163 Writing html report.
2023-09-05 10:43:53,197 Finished!
```

NanoPlot NanoStats.txt file

```
General summary:
Active channels:                   466.0
Mean read length:               24,715.5
Mean read quality:                  15.7
Median read length:             19,995.0
Median read quality:                16.4
Number of reads:                20,232.0
Read length N50:                29,257.0
STDEV read length:              15,487.5
Total bases:                500,043,638.0
Number, percentage and megabases of reads above quality cutoffs
>Q5:    20232 (100.0%) 500.0Mb
>Q7:    20232 (100.0%) 500.0Mb
>Q10:   20232 (100.0%) 500.0Mb
>Q12:   20232 (100.0%) 500.0Mb
>Q15:   17638 (87.2%) 414.2Mb
Top 5 highest mean basecall quality scores and their read lengths
1:      20.8 (9807)
2:      20.5 (23414)
3:      20.5 (6995)
4:      20.3 (8002)
5:      20.2 (8532)
Top 5 longest reads and their mean basecall quality score
1:      151104 (16.5)
2:      146754 (15.4)
3:      143743 (16.2)
4:      139141 (15.6)
5:      121077 (17.8)
```

NanoPlot-report.html file

NanoPlot reports

Summary statistics

General summary	
Active channels	466.0
Mean read length	24,715.5
Mean read quality	15.7
Median read length	19,995.0
Median read quality	16.4
Number of reads	20,232.0
Read length N50	29,257.0
STDEV read length	15,487.5
Total bases	500,043,638.0
Number, percentage and megabases of reads above quality cutoffs	
>Q5	20232 (100.0%) 500.0Mb
>Q7	20232 (100.0%) 500.0Mb
>Q10	20232 (100.0%) 500.0Mb
>Q12	20232 (100.0%) 500.0Mb
>Q15	17638 (87.2%) 414.2Mb
Top 5 highest mean basecall quality scores and their read lengths	
1	20.8 (9807)
2	20.5 (23414)
3	20.5 (6995)
4	20.3 (8002)
5	20.2 (8532)
Top 5 longest reads and their mean basecall quality score	
1	151104 (16.5)
2	146754 (15.4)
3	143743 (16.2)
4	139141 (15.6)
5	121077 (17.8)

NanoPlot

If you run Nanoplot on the same dataset multiple times it will produce slightly different plots.

"The plotting function will randomly sample up to 10000 reads for the plot. This is mainly for the speed of plotting and disk size of the plots. It may lead to subtle differences in outliers, but should not affect the bulk of your data." –Wouter De Coster

Interpretation of Kraken output

Kraken recap

Kraken 1 | KrakenUniq | Kraken 2 | Kraken2Uniq | MiniKraken

Illumina

```
kraken2 --output [output/name] --db [db] --report
sample_report –paired --gzip-compressed
[input_R1.fastq.gz] [input_R2.fastq.gz]
```

Nanopore

```
kraken2 --output [output/name] --db [db] --report
sample_report --gzip-compressed [input.fastq.gz]
```

Kraken recap

- Kraken1 is more accurate than Kraken2.
- Kraken1 database size ~300Gb.
- Kraken2 database size ~30-50Gb.
- KrakenUniq will add 1 extra column to Kraken report with exact k-mer count.
- MiniKraken will use a smaller database.

Kraken2 output

```
kraken2 --output [output/name] --db [db] --report
sample_report
--paired --gzip-compressed [input_R1.fastq.gz]
[input_R2.fastq.gz]
```

Report
File with 1 line per taxon

Standard output
File with one line per read/read-pair

Kraken2 output

Each sequence (or sequence pair, in the case of paired reads) classified by Kraken 2 results in a single line of output.
Kraken 2's output lines contain five tab-delimited fields; from left to right, they are:

- "C"/"U": a one letter code indicating that the sequence was either classified or unclassified.
- The sequence ID, obtained from the FASTA/FASTQ header.
- The taxonomy ID Kraken 2 used to label the sequence; this is 0 if the sequence is unclassified.
- The length of the sequence in bp. In the case of paired read data, this will be a string containing the lengths of the two sequences in bp, separated by a pipe character, e.g. "98|94".

A space-delimited list indicating the LCA mapping of each k-mer in the sequence(s). For example, "562:13 561:4 A:31 0:1 562:3" would indicate that:

- the first 13 k-mers mapped to taxonomy ID #562
- the next 4 k-mers mapped to taxonomy ID #561
- the next 31 k-mers contained an ambiguous nucleotide
- the next k-mer was not in the database
- the last 3 k-mers mapped to taxonomy ID #562

Note that paired read data will contain a "|:|" token in this list to indicate the end of one read and the beginning of another.

Kraken 2 Manual – (oregonstate.edu)

Kraken2 output report

Like Kraken 1, Kraken 2 offers two formats of sample-wide results. Kraken 2's standard sample report format is tab-delimited with one line per taxon. The fields of the output, from left-to-right, are as follows:

- Percentage of fragments covered by the clade rooted at this taxon
- Number of fragments covered by the clade rooted at this taxon
- Number of fragments assigned directly to this taxon
- A rank code, indicating (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. Taxa that are not at any of these 10 ranks have a rank code that is formed by using the rank code of the closest ancestor rank with a number indicating the distance from that rank. E.g., "G2" is a rank code indicating a taxon is between genus and species and the grandparent taxon is at the genus rank.
- NCBI taxonomic ID number
- Indented scientific name

Kraken 2 Manual – (oregonstate.edu)

Kraken2 output report

Filter Kraken report to only contain hits above 1%

```
cat [sample_report] | awk -F "\t" '{if ($1>1) {print}}’ >
[sample_filtered_1p_report]
```

Example of contamination

```
99.18   4384411 23992    R    1    root
98.64   4360294 40   R1   131567     cellular organisms
98.58   4357718 9632     D    2        Bacteria
98.34   4347248 14825    P    1224           Proteobacteria
97.99   4331833 38624    C    1236             Gammaproteobacteria
97.09   4291848 120479   O    91347              Enterobacterales
94.31   4168957 1863629  F    543                Enterobacteriaceae
37.63   1663415 878068   G    570                  Klebsiella
16.98   750700  719553   S    573                    Klebsiella pneumoniae
13.90   614375  344228   G    547                  Enterobacter
 5.97   263939  189088   G1   354276                     Enterobacter cloacae complex
 1.07   47468   30788    S    158836                       Enterobacter hormaechei
```

KrakenTools

README.md

Kraken Tools

KrakenTools is a suite of scripts to be used alongside the Kraken, KrakenUniq, Kraken 2, or Bracken programs. These scripts are designed to help Kraken users with downstream analysis of Kraken results.

For news and updates, refer to the github page: https://github.com/jenniferlu717/KrakenTools/

Citation

KrakenTools has been published on September 28, 2022 as part of a protocol paper for using the Kraken software suite. Please cite the following when using any KrakenTools script:

[Lu J, Rincon N, Wood D E, Breitwieser F P, Pockrandt C, Langmead B, Salzberg S L, Steinegger M. Metagenome analysis using the Kraken software suite. Nature Protocols, doi: 10.1038/s41596-022-00738-y (2022)] (https://www.nature.com/articles/s41596-022-00738-y)

Extract unwanted reads

```
extract_kraken_reads.py -k contaminated_kraken2 -s
consample_R1.fastq.gz -s2 consample_R2.fastq.gz -t 547 -o
consample_R1_decon.fastq -o2 consample_R2_decon.fastq --exclude --
include-children --fastq-output -r consample_kraken_report
```

Will output unzipped fastq files (one for each input file) WITHOUT the reads assigned at belonging to taxon 547 or any reads assigned to more specific subgroups

```
99.61   1903841 9689     R    1     root
99.10   1894096 186 R1   131567      cellular organisms
99.05   1893210 2708     D    2         Bacteria
98.90   1890278 6470     P    1224            Proteobacteria
98.55   1883616 16559    C    1236              Gammaproteobacteria
97.64   1866267 49500    O    91347               Enterobacterales
94.99   1815480 727858   F    543               Enterobacteriaceae
55.17   1054543 507634   G    570                 Klebsiella
27.40   523625  500331   S    573                   Klebsiella pneumoniae
 1.09   20896   17497    S1   72407                       Klebsiella pneumoniae subsp. pneumoniae
```

Trimmatic

Trimming out the trash

Trimmomatic Architcture

trimmomatic [SE|PE]

SE = Single-end

PE = Paired-end

Trimmomatic Architcture

trimmomatic [SE|PE] sample1.fastq.gz

Trimmomatic Architcture

trimmomatic [SE|PE] sample1{_R1}.fastq.gz {sample1_R2.fastq.gz}

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz
sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz
sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz

NOTE: Specify trimmed and unpaired output for each read mates.

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz
sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz
sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz

NOTE: Specify trimmed and unpaired output for each read mates.

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:

ILLUMINACLIP = Mandatory flag

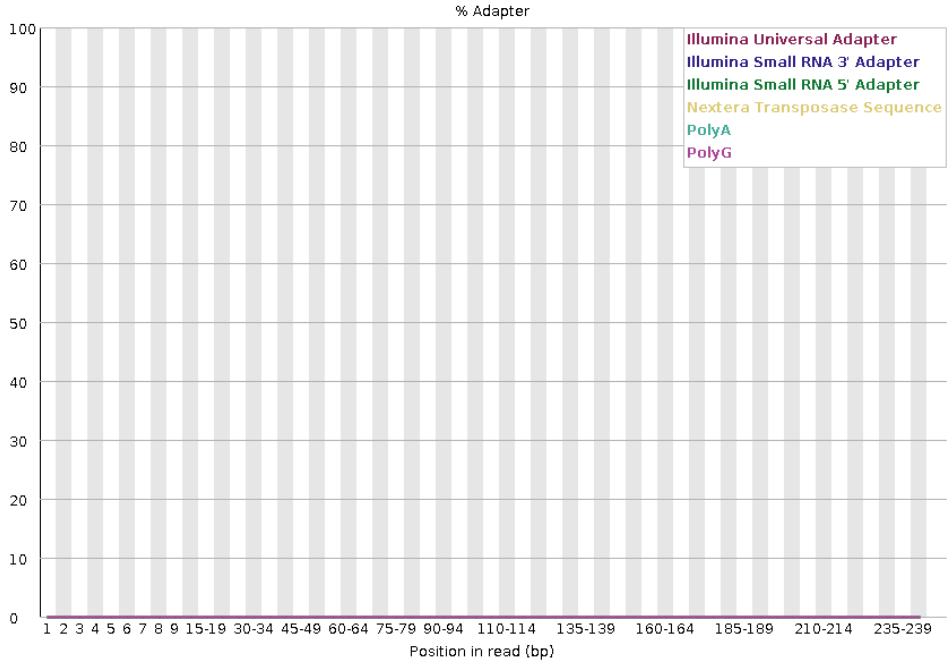
Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta

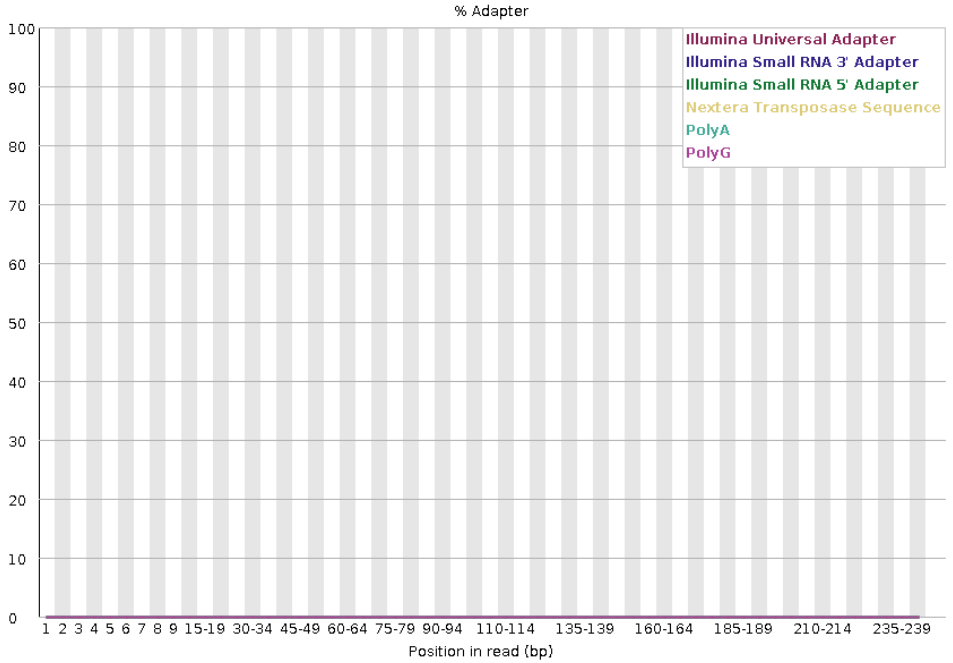
ILLUMINACLI:<path/2/adapters.fasta>

Adapter contents

Ec005_R1



Ec005_R2



Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta:2

ILLUMINACLI:<path/2/adapters.fasta>:<seed mismatches>

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta:2:30

ILLUMINACLI:<path/2/adapters.fasta>:<seed mismatches>:<palindrome clip threshold>

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta:2:30:10

ILLUMINACLI:<path/2/adapters.fasta>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>

Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta:2:30:10 LEADING:30

LEADING = Cut off X bases from start of read if below Q30

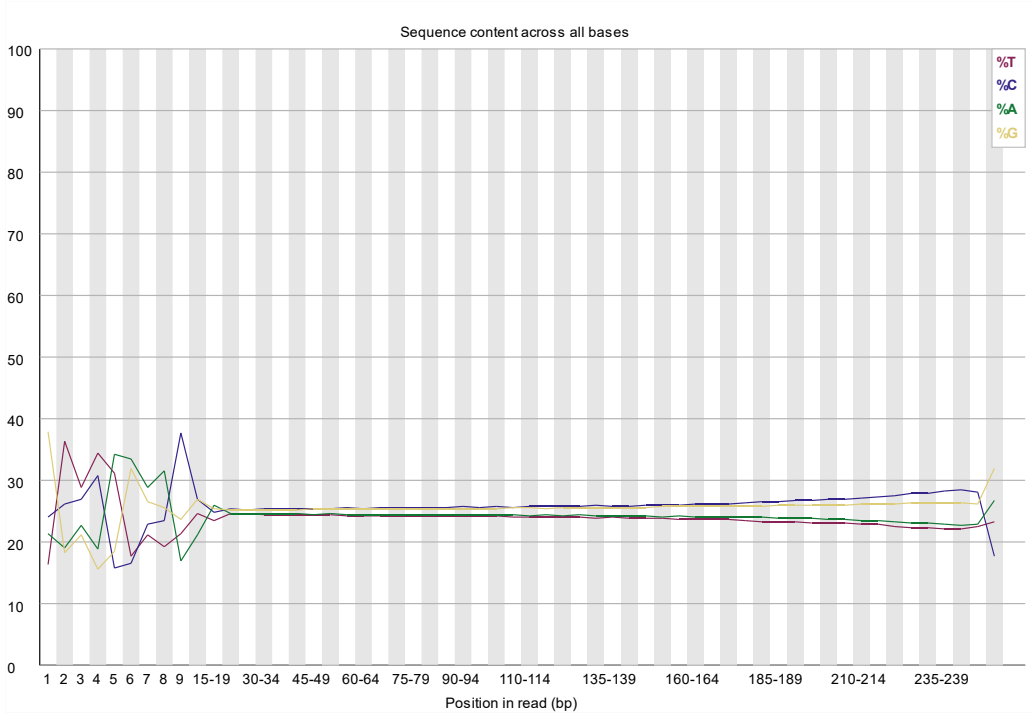
Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta:2:30:10 LEADING:30 TRAILING:30

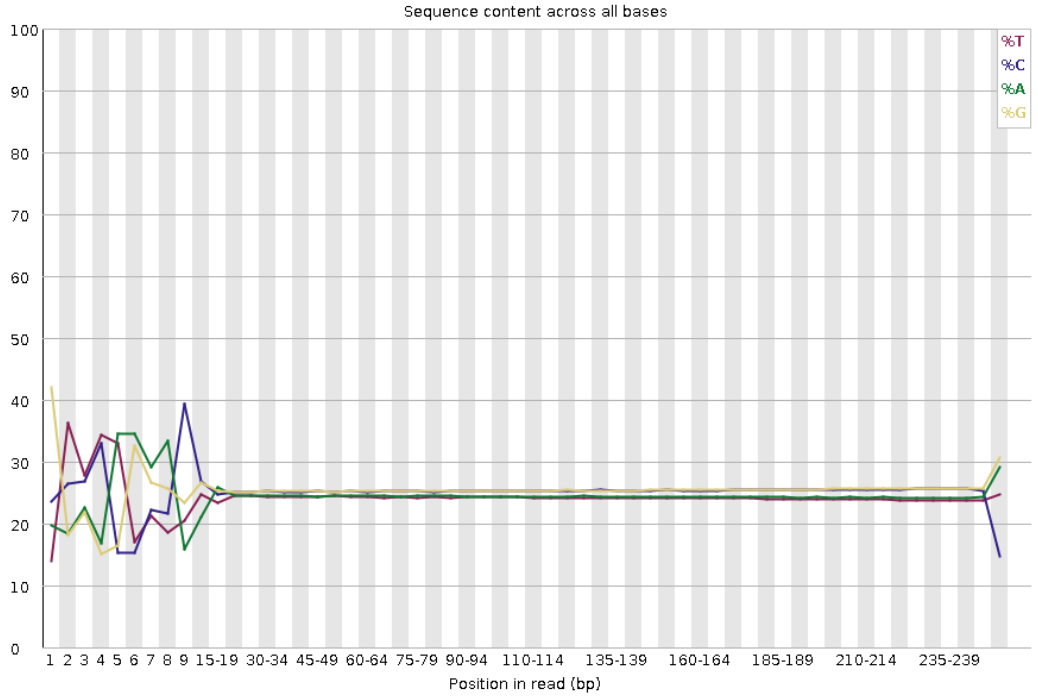
LEADING = Cut off X bases from start of read if below Q30
TRAILING = Cut off Y bases from end of read if below Q30

Per base sequence contents

Ec005_R1

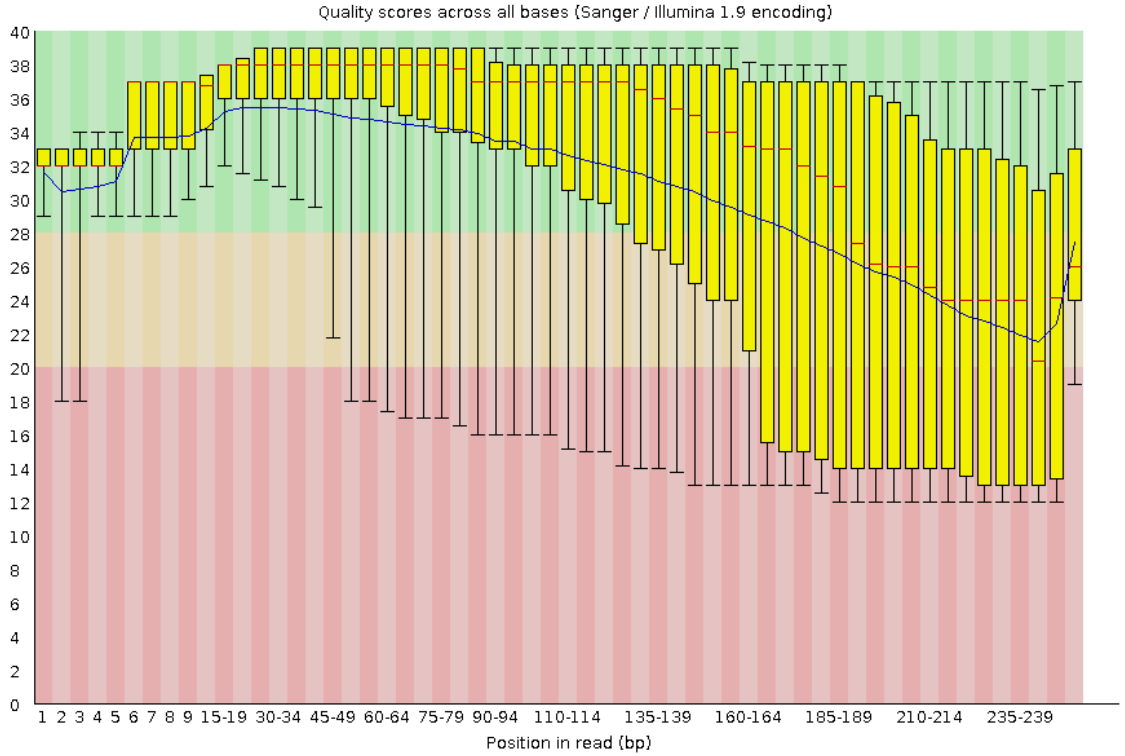


Ec005_R2

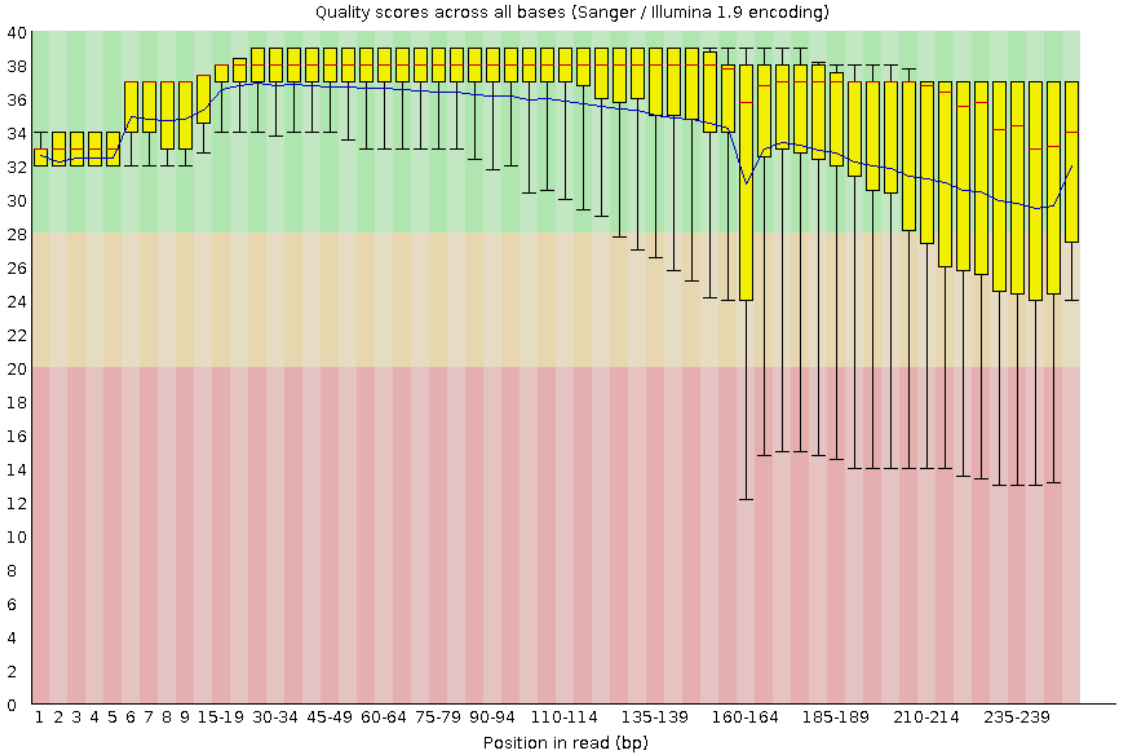


Per base sequence quality

Ec005_R1



Ec005_R2



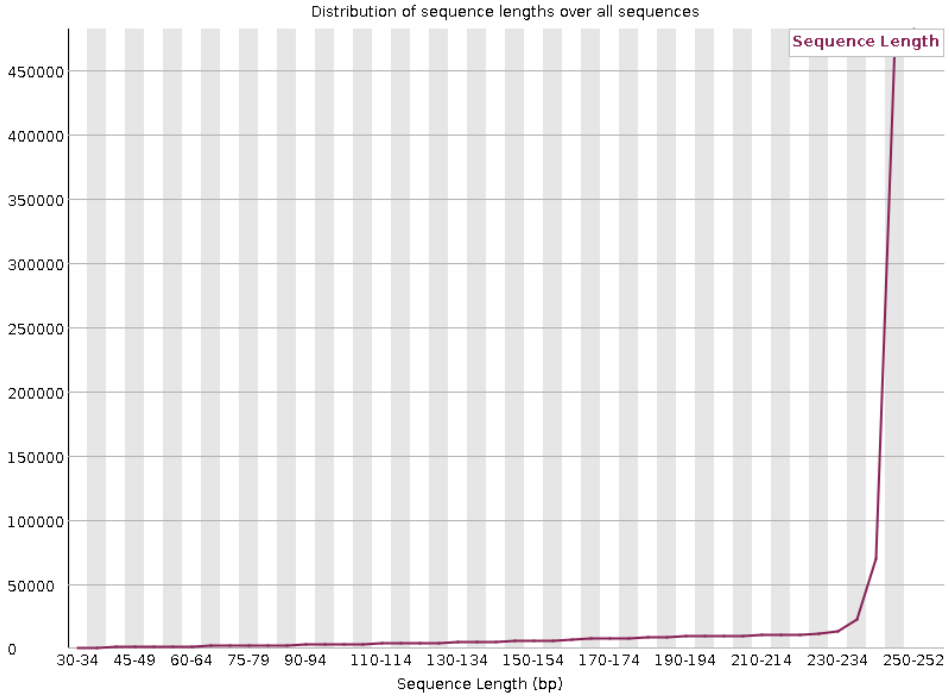
Trimmomatic Architcture

trimmomatic PE sample1_R1.fastq.gz sample1_R2.fastq.gz sample1_trimmed_R1.fastq.gz sample1_unpaired_R1.fastq.gz sample1_trimmed_R2.fastq.gz sample1_unpaired_R2.fastq.gz ILLUMINACLIP:relevant_adapters.fasta:2:30:10 LEADING:30 TRAILING:30 MINLEN:120

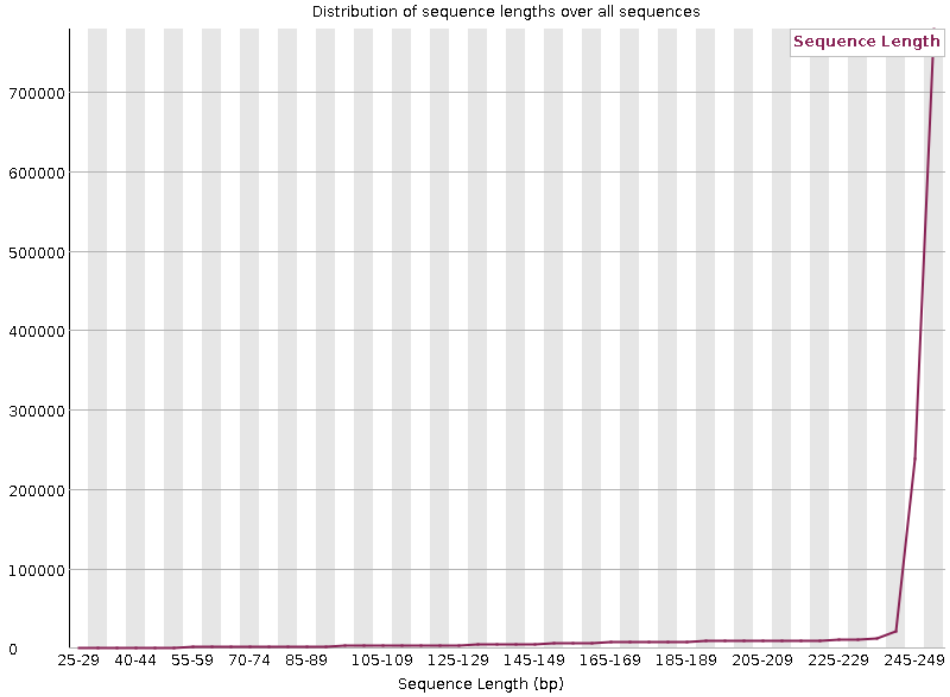
LEADING = Cut off X bases from start of read if below Q30
TRAILING = Cut off Y bases from end of read if below Q30
MINLEN = Exclude reads shorter than Z

Sequence length distribution

Ec005_R1



Ec005_R2



What are the effects of trimmomatic?

Results for Ec005

Stat	Value
Input Read Pairs	1273619
Both Surviving Reads	1233679
Both Surviving Read Percent	96,86
Forward Only Surviving Reads	313
Forward Only Surviving Read Percent	0,02
Reverse Only Surviving Reads	227
Reverse Only Surviving Read Percent	0,02
Dropped Reads	39400
Dropped Read Percent	3,09

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

Results for Ec005

Ec005_R1

Measure	Value
Filename	Ec005.illumina_R1.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	298.6 Mbp
Sequences flagged as poor quality	0
Sequence length	31-251
%GC	51

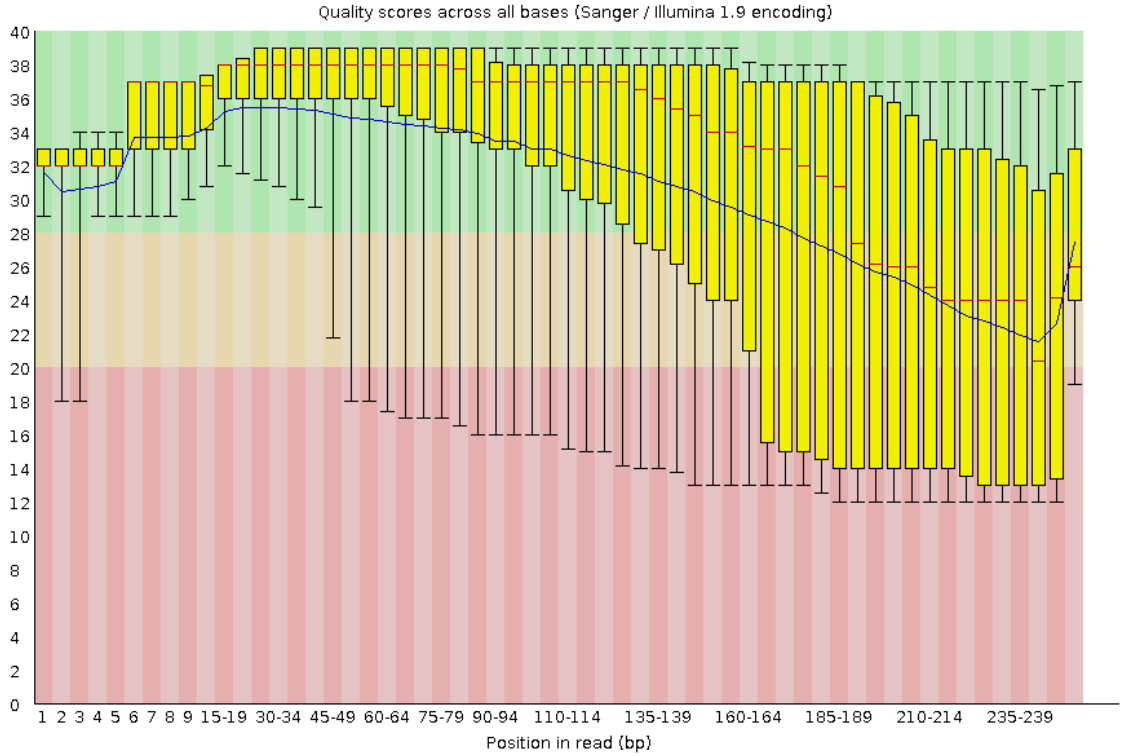
Measure	Value
Filename	Ec005.illumina_trimmed_R2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1233679
Total Bases	296.4 Mbp
Sequences flagged as poor quality	0
Sequence length	120-251
%GC	51

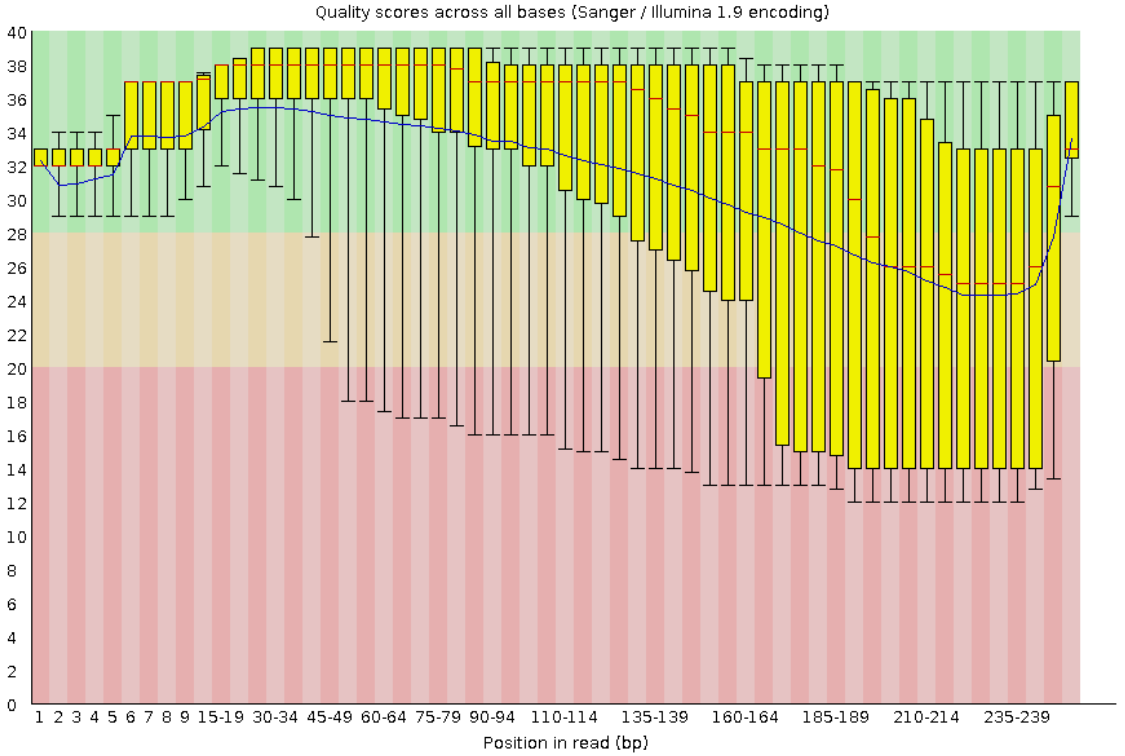
Ec005_R2

Measure	Value
Filename	Ec005.illumina_R2.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1273619
Total Bases	299.9 Mbp
Sequences flagged as poor quality	0
Sequence length	28-251
%GC	51

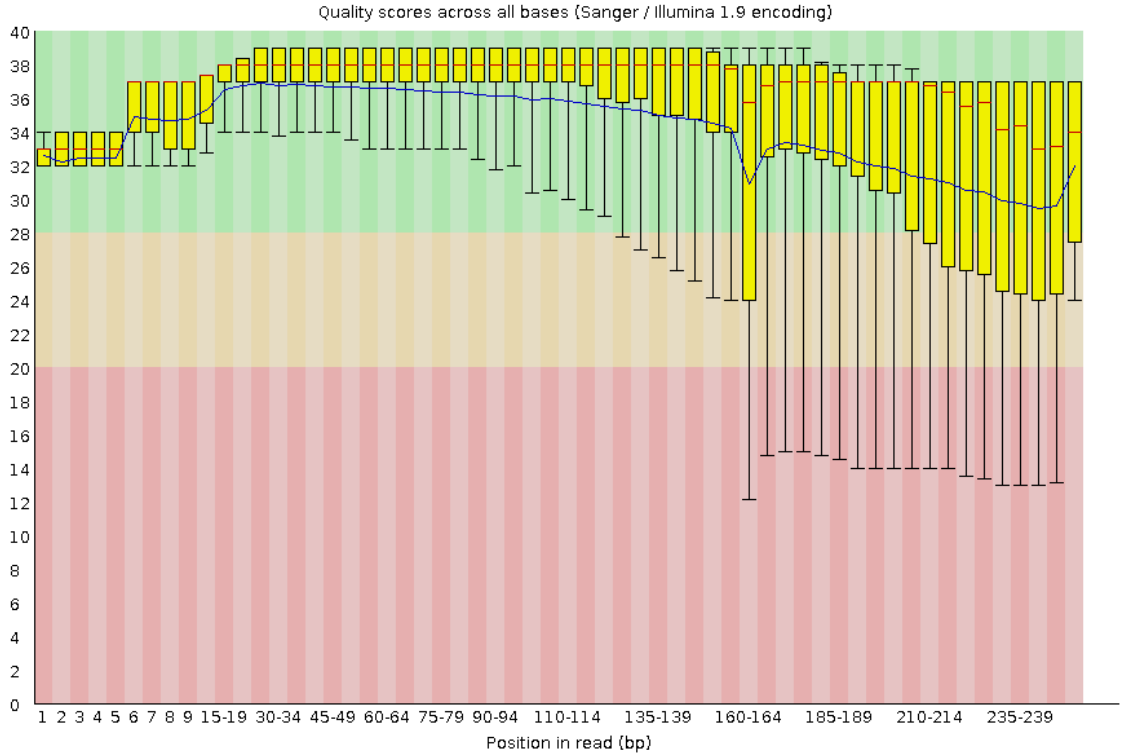
Measure	Value
Filename	Ec005.illumina_trimmed_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1233679
Total Bases	295 Mbp
Sequences flagged as poor quality	0
Sequence length	120-251
%GC	51

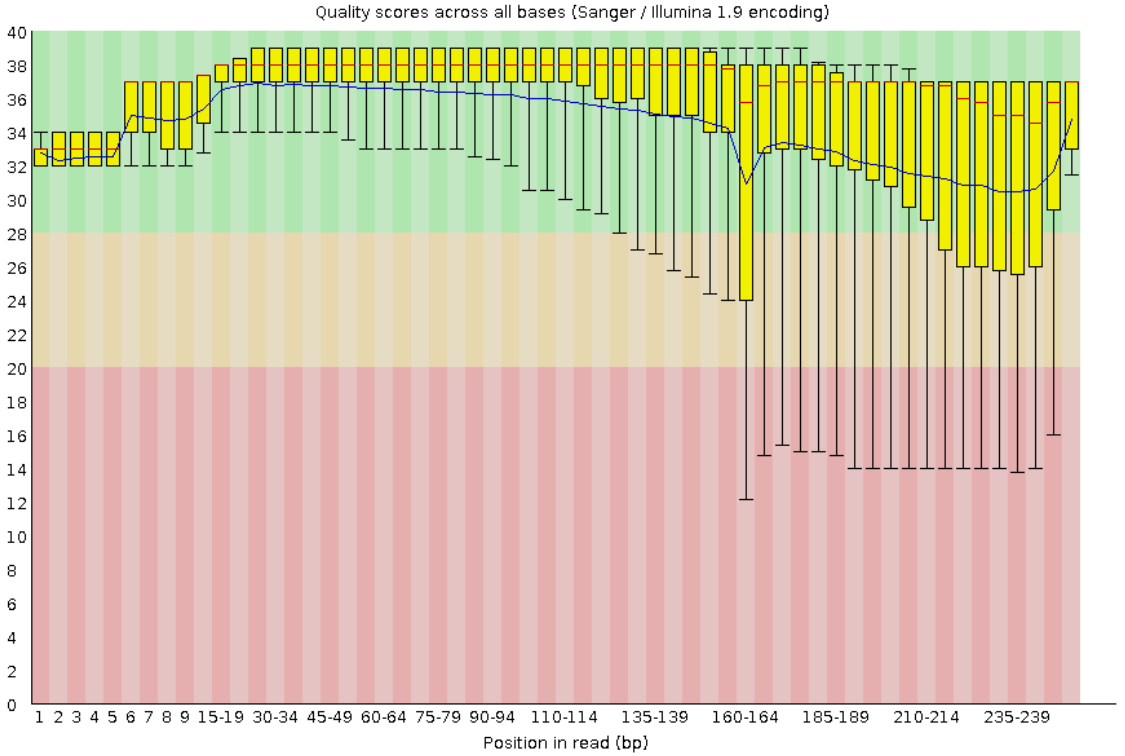
Per base sequence quality Ec005_R1





Per base sequence quality Ec005_R2





Now its your turn!

Now its your turn!

Go to the trimmomatic exercise on EVA ☺

Trimming and filtering of Nanopore reads

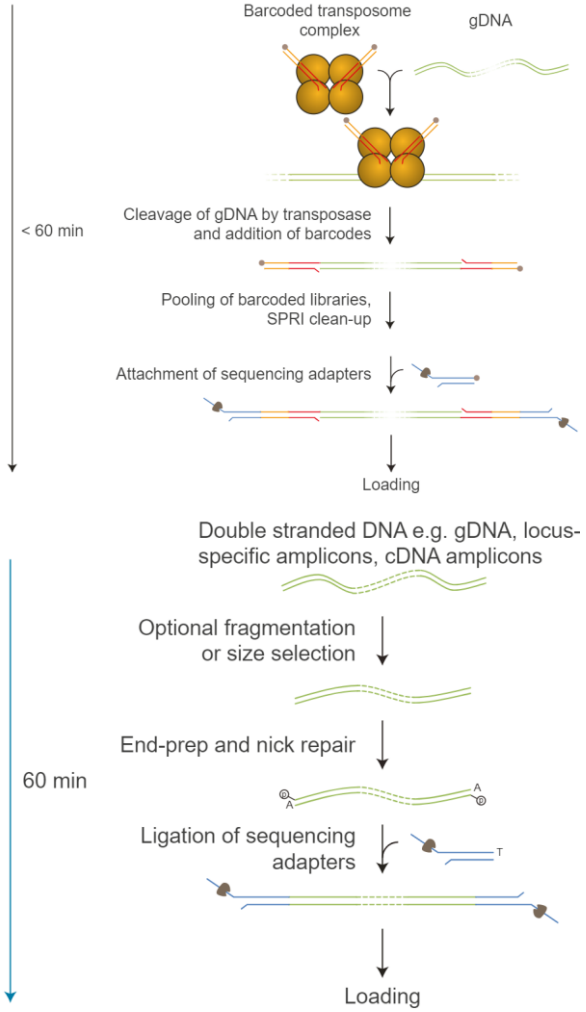
Adapters and barcodes

Adapters will always be added to the ends of DNA strands when preparing a Nanopore sequencing library.

If barcoding, you can enable barcode+adapter trimming in MinKNOW.

Guppy can detect and trim barcodes and adapters (Dorado can hopefully soon too)

Porechop_abi can look for overrepresented sequences and remove them



Porechop_abi

Porechop_abi to detect adapter sequences in your reads:

```
porechop_abi -go --format fastq -i [input.fastq.gz] -t [threads]
```

```
-go == --guess_adapter_only
```

Porechop_abi to removes adapter sequences in your reads:

```
porechop_abi -abi --format fastq -i [input.fastq.gz] -o
[output_trimmed.fastq.gz]
```

```
-abi == --ab_initio
```

Practical

Head to the course page on EVA

Navigate to session two and find: –Long-read_trimming_filtering

Enjoy!

Quality assessment from multiple Illumina sample reports

MultiQC

MultiQC

It might be a challenging to set up for all modules but...

MultiQC

It might be a challenging to set up for all modules but...

...once it works, it’s a blast!

Adventure time!

Go to the MultiQC exercise on EVA ☺

Quality assessment on multiple Nanopore samples

NanoPlot

NanoPlot on a multiplexed run

```
NanoPlot -t [threads] --summary  sequencing_summary.txt --barcoded --outdir [output directory]
```

No report

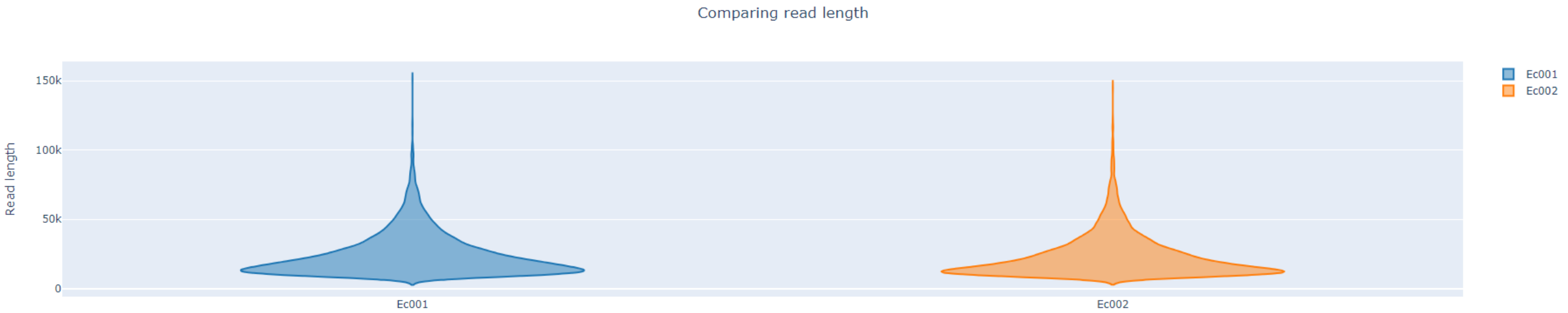
Plots for all barcodes individually

Large stat files with info on all barcodes

Nanocomp

Example

```
NanoComp --fastq reads1.fastq.gz reads2.fastq.gz
reads3.fastq.gz reads4.fastq.gz --names run1 run2 run3 run4
```



https://github.com/wdecoster/nanocomp

Final quiz

Acknowledgements

The creation of this training material was commissioned by ECDC to Statens Serum Institut (SSI) with the direct involvement of Sharmin Baig, Søren Hallstrøm, Kasper Thystrup Karstensen & Astrid Rasmussen