In this example, we are unzipping a compressed FASTQ file and printing the uncpmressed data to a new file called "output.txt." The STDERR stream is redirected to a file called "error.log" using the `2>` operator. This means that any error messages generated by the `gzip` command will be written to the "error.log" file instead of being printed to the terminal.

Note that the `2>` operator must be used after the command you want to redirect STDERR for. In this example, we used it after the `gzip -cd file.fastq.gz` command to redirect STDERR to "error.log."

This can be especially useful when dealing with large datasets, where error messages may be numerous and difficult to keep track of in the terminal. By redirecting STDERR to a file, you can keep track of any errors that occur during your analysis without cluttering up your terminal output.

Most bioinformatics tools offer STDERR streams as logs. One should always keep and use them

If you make Linux do something stupid, such as get stuck in a loop, you can always get out by typing C**trl+C**. We will not make you do that in the course but you should remember **Ctrl+C**.

Getting Data

Now let's get some files this will might take a while

As this takes a while we needed to start this on day 2 or before the exercises on day 3

mkdir course
cd course
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR176/006/SRR17655906/SRR17655906.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR176/003/SRR17655903/SRR17655903.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR176/098/SRR17655898/SRR17655898.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR176/007/SRR17655907/SRR17655907.fastq.gz

Day 3 Working with some real data

Questions and exercises

1. What might these files be? [use linux/unix tools]
2. Get a basic idea how much data you downloaded
3. Analyze how much bases (approximately) and reads (exactly) you have per file
4. Get the md5sum of the files
5. Write a short script that counts the length of each sequence line in a fastq file: use sort und uniq on the output of the script as a basic quality control tool.

Experimental Design

Before you run an experiment, consider a few principles. Some NGS experiments are still somewhat expensive and can sometimes fail. Hence consider good experimental design and practice.

Replication strategy – When planning experiments using RNA sequencing, ATAC etc, there are several factors to consider to ensure **robust and reliable results**. One important consideration is how many replicates to use. Although cost considerations may make it tempting to avoid deeper replication, it is important to keep in mind that biological variation is significant and we encounter technical noise in RNAseq as well. Technical replicates do not help!

When deciding on the number of replicates to include, it is important to consider the expected variability in your data and the statistical power needed to detect meaningful differences.

Sequencing Depth and preparation

In addition, you also have to decide on sequencing depth as the more reads you have – the better you can detect and quantify lowly expressed transcripts. This is another cost factor to consider. Another thing is stranded versus unstranded RNAseq experiments. Here the difference in cost is not high but you gain more biological inishgts.

RNA quality

Finally, it is of high importance that the RNA (same holds for anything really) is of high quality and not degraded.

Pitfalls– **<u>DNA contaminated RNA samples literally cause nightmares</u>**

In summary, when designing RNA-seq experiments, it is important to consider the trade-off between replication and sequencing depth, and to strike a balance that allows for reliable (and needed) statistical inferences. Including biological replicates and using high-quality RNA samples are also critical for ensuring the robustness and reliability of your data.

Famous last words
I am sure the samples were not switched.
Two reps is good enough EASY
Keeping backup sample makes no sense
Sending stuff by UPS/DHL/Post on dry ice works

Plant C Value database

The Plant C Value Database is a resource for researchers interested in the genome size, also known as C value, of plants. Genome size refers to the total amount of DNA in a cell, typically measured in picograms (pg) or base pairs (bp). The Plant C Value Database contains information on the genome size of several thousand plant species. (But take note there might be data not contained in the database and it is only as good as the underlying data).

The database is maintained by the Royal Botanic Gardens, Kew in the United Kingdom, and is freely available to the scientific community. The data in the database is compiled from published studies, including both original research and compilations of previously published data.

One can use the Plant C Value Database to explore the range of genome sizes among plant species, as well as to investigate patterns of genome size evolution.

In addition to providing information on plant genome size, the Plant C Value Database also includes data on the ploidy level of each species. Ploidy level refers to the number of sets of chromosomes in a cell, and can have important implications for plant development and evolution.

The Plant C Value Database offers several search options, including by plant family, common name, or scientific name. Users can also download the entire dataset for further analysis.

Student Exercises Plant C Value:

1. Use the Plant C Value Database to investigate the range of genome sizes for species of the genus *Solanum*.
2. Explore the ploidy level data available in the Plant C Value Database of *Solanum*.
3. Use the database to identify *Solanum* species with particularly large and small genome sizes.

NCBI and EBI

The NCBI SRA and the EBI ENA databases are two of the largest and most widely used public repositories of nucleotide sequence data. These databases serve as central repositories for **raw sequence data** generated by high-throughput sequencing technologies, including Illumina, PacBio, Oxford Nanopore but also discontinued technologies like 454 and microarray expression data.

The SRA database is maintained by the National Center for Biotechnology Information (NCBI), which is a part of the National Institutes of Health (NIH) in the United States. The ENA database is maintained by the European Bioinformatics Institute (EBI), which is a part of the European Molecular Biology Laboratory (EMBL). Both databases operate under the International Nucleotide Sequence Database Collaboration (INSDC), which is a global partnership that aims to provide comprehensive, high-quality nucleotide sequence data to the scientific community. Th

The SRA and ENA databases allow researchers to submit, access, and analyze raw sequence data in a standardized format. The data in these databases are annotated with metadata, such as sample information, experimental conditions, and sequencing platform details, which allows researchers to evaluate the quality and relevance of the data before using it for downstream analyses.

In addition to raw sequencing data, the SRA and ENA databases also contain processed data, such as assembled genomes, transcriptomes, and annotations. These databases provide researchers with a wealth of data that can be used to answer a wide range of biological questions, from understanding the genetic basis of disease to investigating the evolution of species.

One of the most significant benefits of the SRA and ENA databases is that they promote data sharing and reuse. By making raw sequence data and associated metadata freely available to the scientific community, these databases enable researchers to use the same data for different analyses and to compare results across studies. This helps to accelerate scientific discovery and advances our understanding of the natural world.

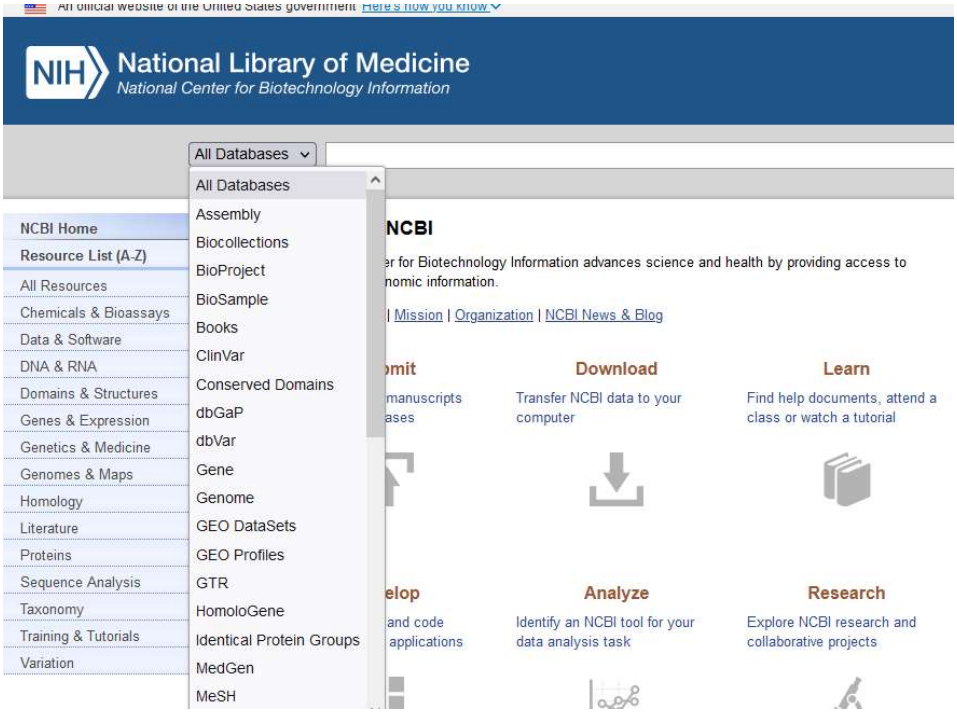


Figure 1 NCBI entry https://www.ncbi.nlm.nih.gov/

Organism Overview ; Genome Assembly and Annotation report [27] ; Organelle Annotation Report [8] ID: 7

Solanum lycopersicum (tomato)

Solanum lycopersicum Organism overview

Lineage: Eukaryota[11443]; Viridiplantae[1241]; Streptophyta[1145]; Embryophyta[1137]; Tracheophyta[1123]; Spermatophyta[1108]; Magnoliopsida[1084]; eudicotyledons[866]; Gunneridae[866]; Pentapetalae[866]; asterids[267]; lamiids[173]; Solanales[62]; Solanaceae[53]; Solanoideae[39]; Solaneae[34]; Solanum[33]; Solanum subgen. Lycopersicon[12]; Solanum lycopersicum[1]

Solanum lycopersicum has been used for a number of firsts in molecular genetics of plants. The use of RFLP (restriction fragment length polymorphism) to generate a linkage map of a complete plant genome was done with tomato (Bernatzky and Tanksley. Genetics 1986; 112:887-898). Tomato was the organism where quantitative traits were resolved into More...

Summary

Sequence data: genome assemblies: 27; sequence reads: 161 (See Genome Assembly and Annotation report)
Statistics: median total length (Mb): 793.815
median protein count: 37604
median GC%: 34.4976
NCBI Annotation Release: 103

Publications (limited to 20 most recent records)

1. De novo genome assembly of two tomato ancestors, Solanum pimpinellifolium and Solanum lycopersicum var. cerasiforme, by long-read sequencing. Takei H, et al. DNA Res 2021 Jan 19
2. A high-continuity and annotated tomato reference genome. Su X, et al. BMC Genomics 2021 Dec 15
3. Linkage between the I-3 gene for resistance to Fusarium wilt race 3 and increased sensitivity to bacterial spot in tomato. Li J, et al. Theor Appl Genet 2018 Jan

More...

Representative (genome information for reference and representative genomes)

Reference genome:

Solanum lycopersicum SL3.1

Submitter: Solanaceae Genomics Project

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	1	NC_015438.3	CM001064.3	98.46	35.6	4,790	12	89	1,172	3,898	197
	Chr	2	NC_015439.3	CM001065.3	55.98	35.8	3,997	-	65	1,276	3,017	80
	Chr	3	NC_015440.3	CM001066.3	72.29	36.4	3,820	2	70	743	3,019	87
	Chr	4	NC_015441.3	CM001067.3	66.56	35.2	3,124	-	66	654	2,586	102
	Chr	5	NC_015442.3	CM001068.3	66.72	35.2	2,437	2	54	559	2,142	111

Figure 2: The tomato genome in the NCBI genome section

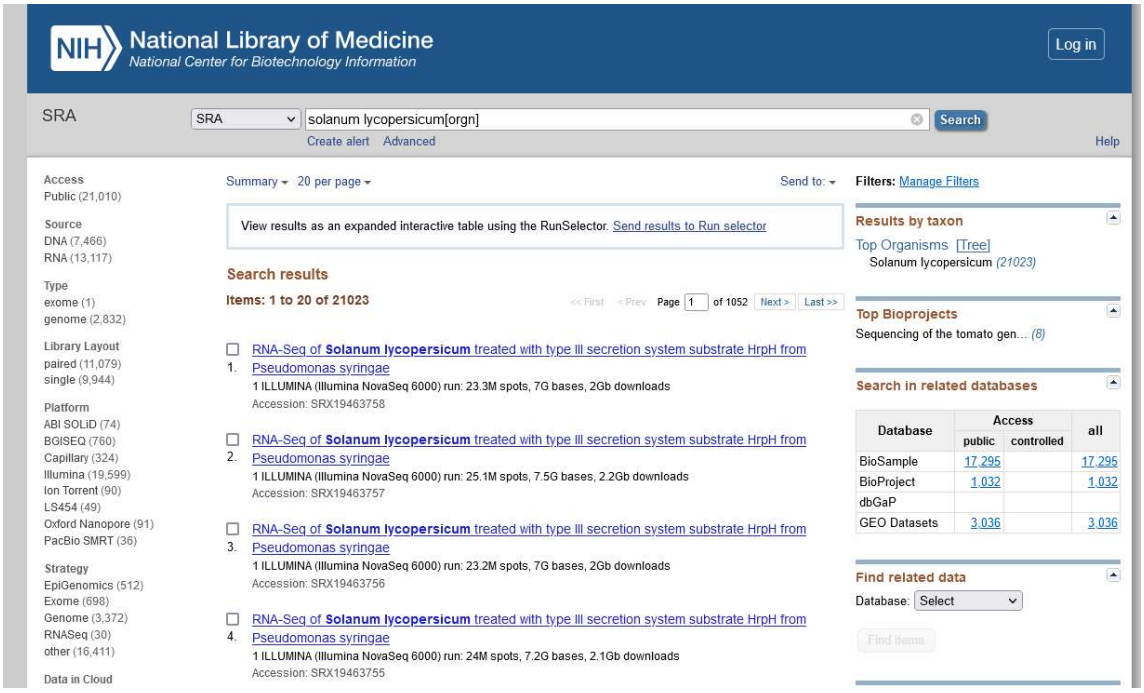


Figure 3: tomato short read archive (SRA) in the NCBI genome section

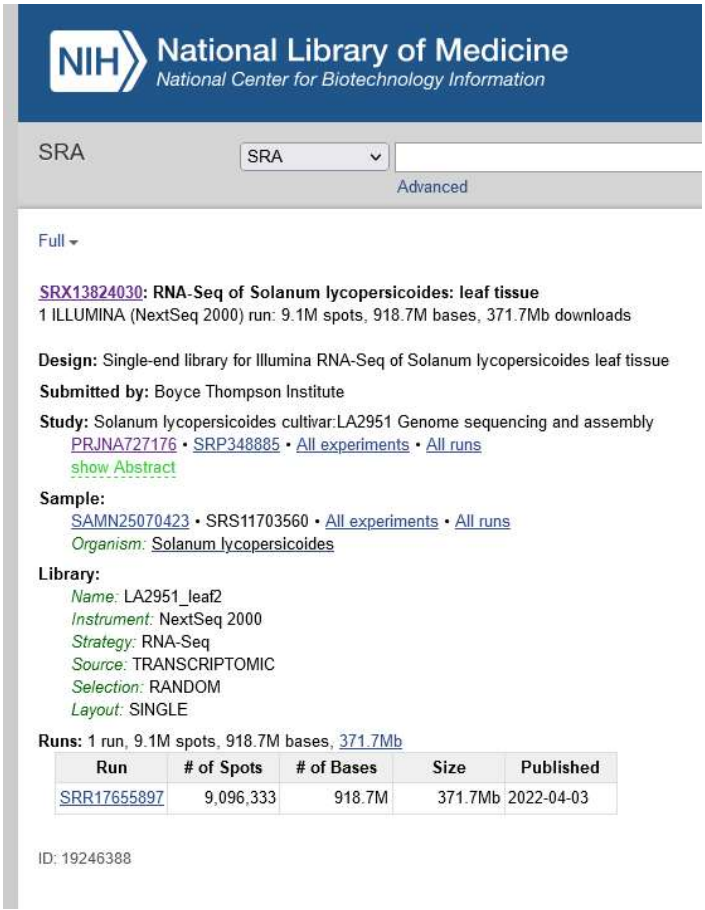


Figure 4: specific read set for *Solanum lycopersicoides*

Student Exercises Sequence Database:

1. Search for the genomes of *Solanum lycopersicum, Solanum pennellii* and *Solanum lycopersicoides*. Compare the differences.
2. Explore the different search options and filters available in the SRA or ENA database. Use these options to find all the sequencing data available for a specific gene or genomic region.
3. Find all the RNA-seq datasets available for *Solanum lycopersicoides* in the SRA or ENA database. Filter the results based on a specific tissue (e.g. leaf)
4. Locate the data for SRX13824030
5. Gather all metadata for the read sets that you downloaded

JASPAR

The JASPAR database contains storing transcription factor binding sites

Student Exercises

1. find all binding sites for *Solanum lycopersicum*
2. Rate how much data you have in JASPAR- will this be helpful for a plant researcher / a person working with mice?

Weblogos

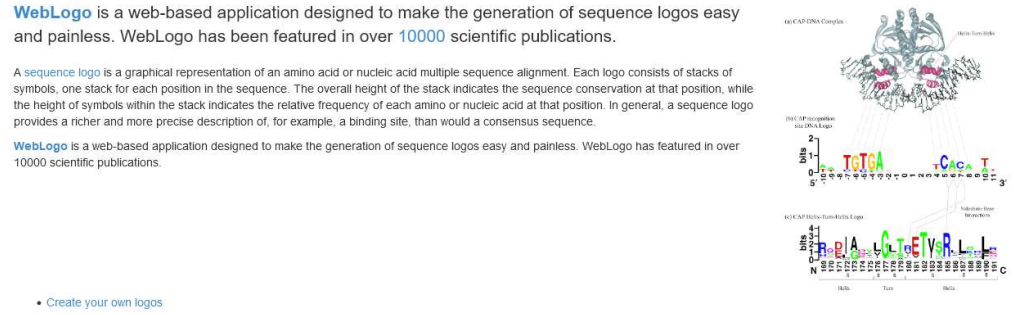
Sequence or binding site weblogos are widely used in biomedical research to analyze DNA or protein sequences for patterns or motifs that are indicative of functional or structural properties. These logos are graphical representations of the sequence conservation at each position in the sequence alignment.

In a sequence logo, each position in the alignment is represented as a stack of letters, with the height of each letter proportional to its frequency (or information content) at that position.

To create a sequence logo, the first step is to generate a multiple sequence alignment (MSA) of the sequences of interest. Once the MSA is generated, the next step is to calculate proportions and/or the information content (IC) of each position in the alignment. The IC is a measure of the degree of conservation at each position and is calculated based on the frequency of each residue at that position and the background frequency of that residue in the entire sequence set.

Sequence logos have many applications in biomedical research, including the identification of transcription factor binding sites, protein interaction domains, and post-translational

modification sites.



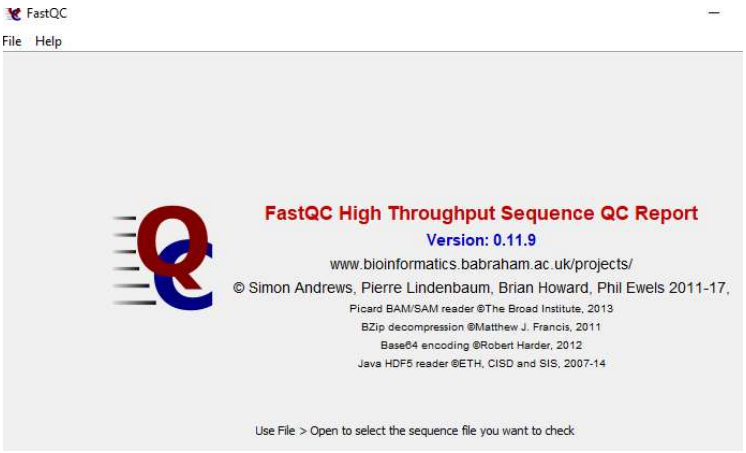
Student Exercises

Explore the weblogo website create some weblogos using e.g. the examples and try to modify them

Day 3/4 Quality Control

FastQC

FastQC is a widely used tool for assessing the quality of Next-Generation Sequencing (NGS) data, including RNA-seq data. It provides a comprehensive set of quality metrics that can help identify issues that may affect downstream analysis, such as low sequencing quality, adapter contamination, and overrepresented sequences. FastQC is a free, open-source software package that can be easily installed and run on most operating systems.



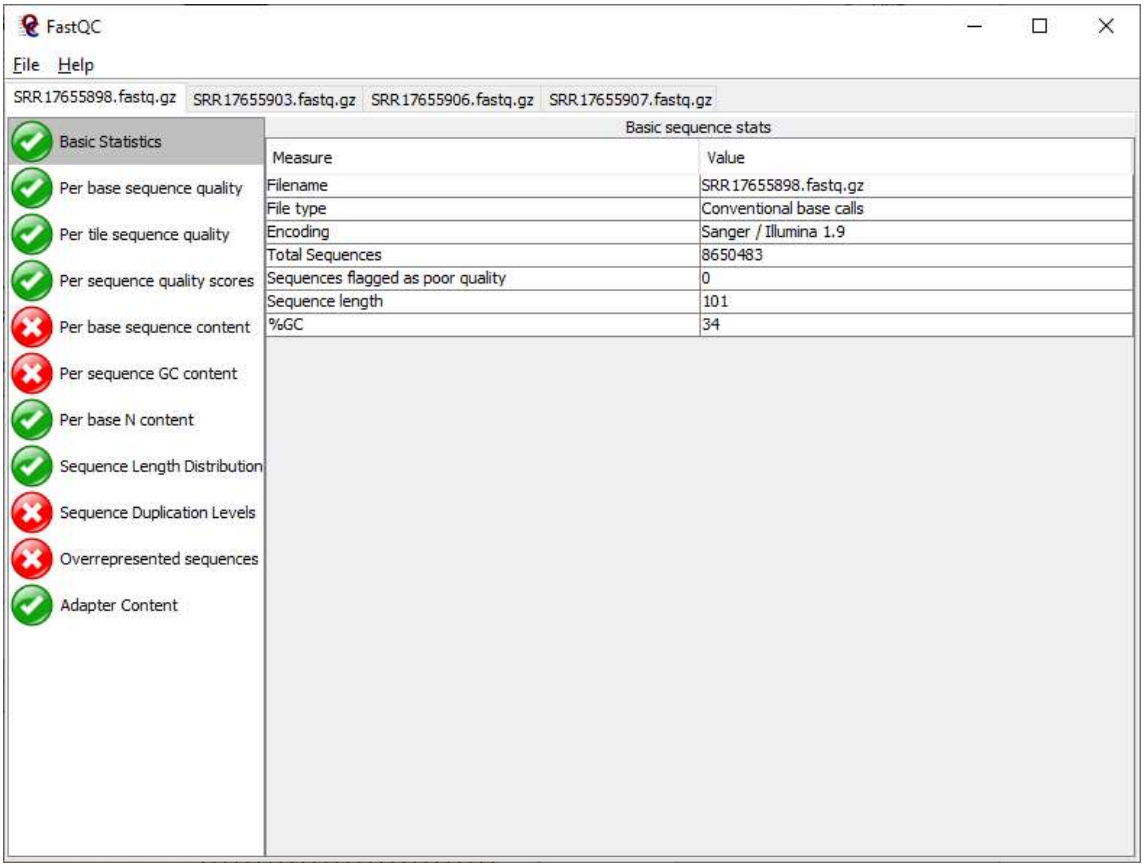
To use FastQC, the first step is to launch the program and load your dataset. Like most other tools working on sequencing files, FastQC allows for compressed (gz) files. After the data has been loaded, FastQC will perform a series of quality control checks and generate a detailed report with visualizations that allow you to quickly assess the quality of your data.

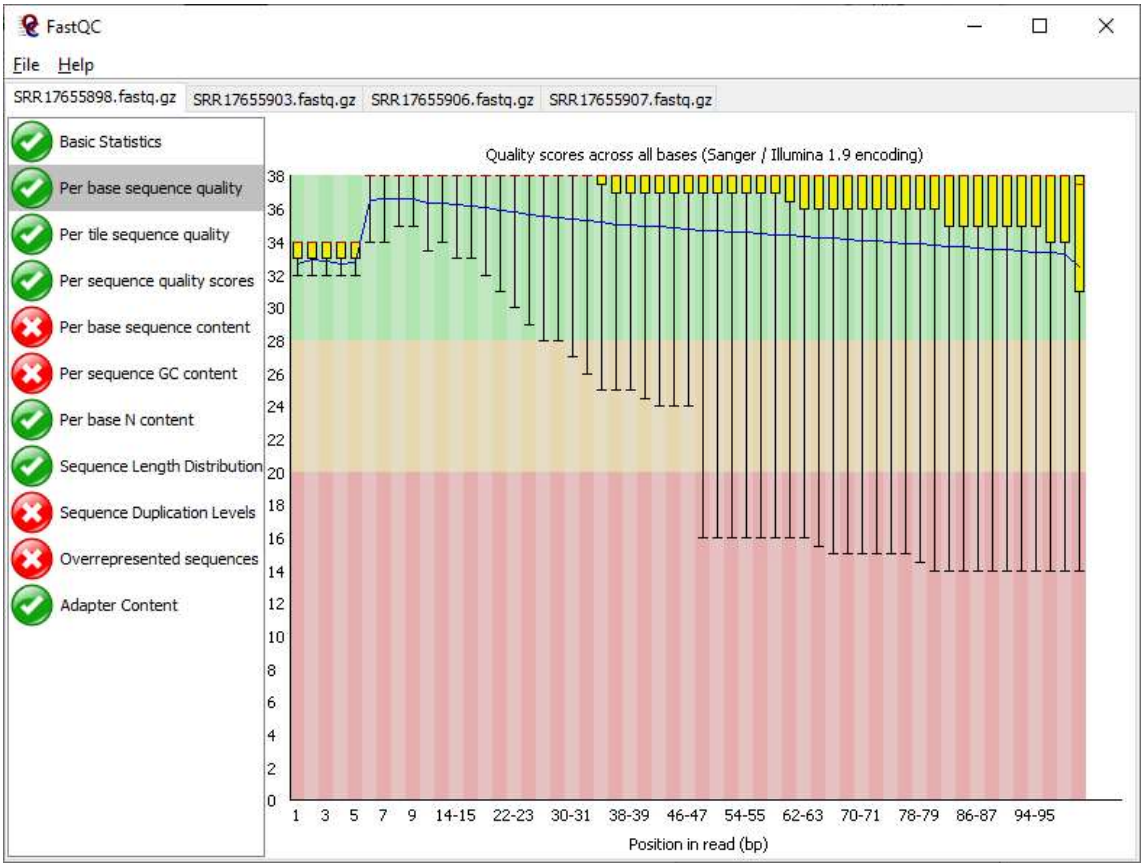
If FastQC doesn't see your WSL2 linux folder. Go there once in the File explore so it appears in your recent items. Then select the recent one in FASTQC

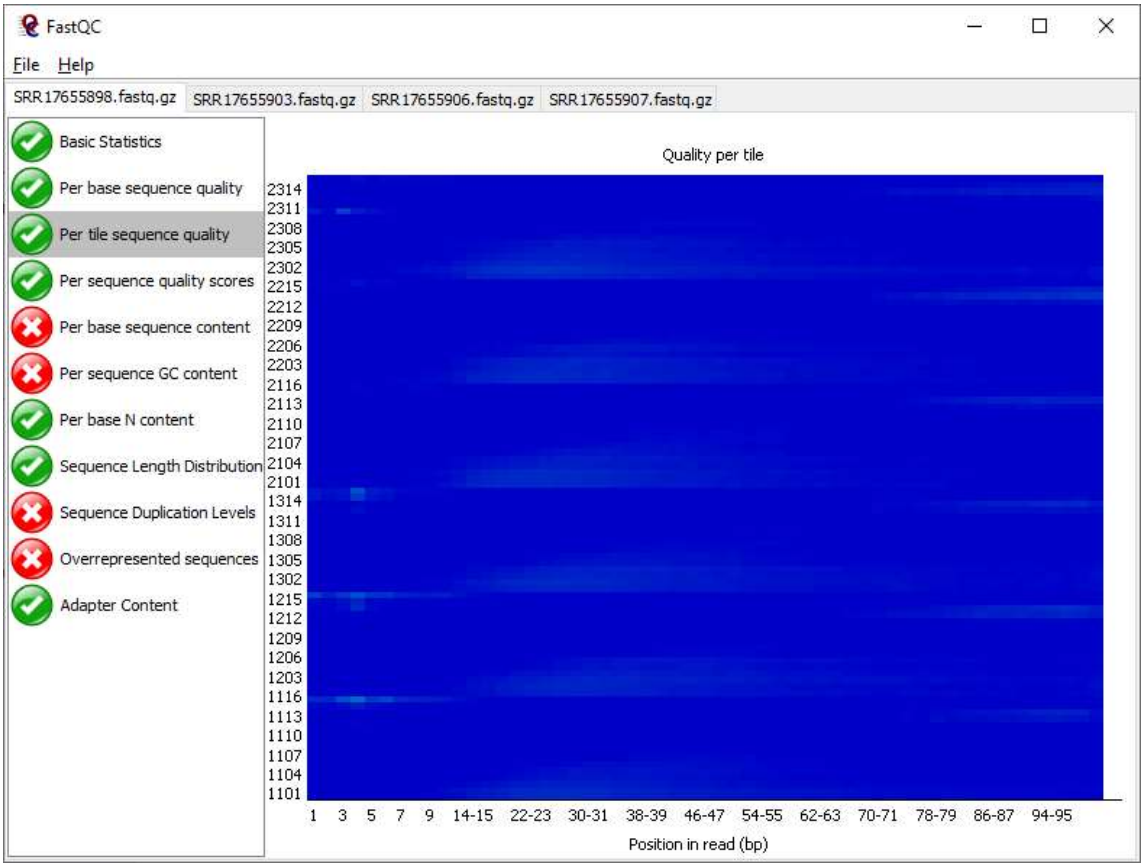
The FastQC report is organized into several tabs, each of which displays different quality metrics. These tabs include:

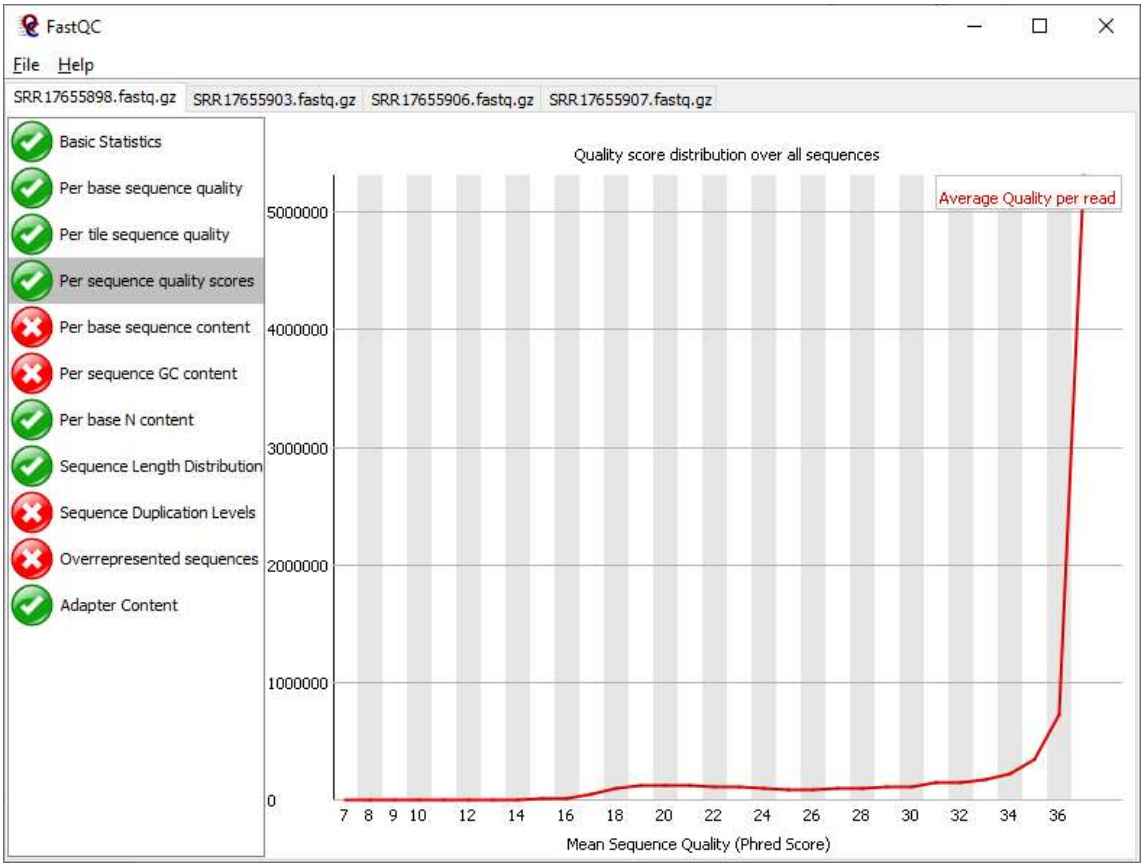
1. Basic Statistics: This tab provides basic information about the data, such as the total number of reads, the percentage of reads that pass the quality filters, and the average read length.
2. Per Base Sequence Quality: This tab shows the quality scores of each base in the sequence, which can help identify issues with sequencing quality that may affect downstream analysis.
3. Per Tile Sequence quality
4. Per Sequence Quality Scores: This tab provides a histogram of the quality scores for all sequences in the dataset, which can help identify any biases or unusual patterns in the data.
5. Per Base Sequence Content: This tab shows the percentage of each base at each position in the sequence, which can help identify issues with sequence bias or contamination.
6. Per Sequence GC Content: This tab shows the distribution of GC content in the dataset, which can help identify issues with GC bias or contamination.
7. Per Base N Content: This tab shows the percentage of Ns (undetermined bases) at each position in the sequence, which can help identify issues with incomplete sequencing or sample contamination.
8. Sequence Length Distribution: This tab shows the distribution of read lengths in the dataset, which can help identify any issues with sequencing library preparation.
9. Sequence Duplication Levels
10. Overrepresented Sequences: This tab identifies sequences that are overrepresented in the dataset, which can help identify potential contamination or adapter sequences that may need to be trimmed.
11. Adapter Content: This tab shows the percentage of reads that contain adapter sequences, which can help identify issues with adapter contamination in the sequencing library.

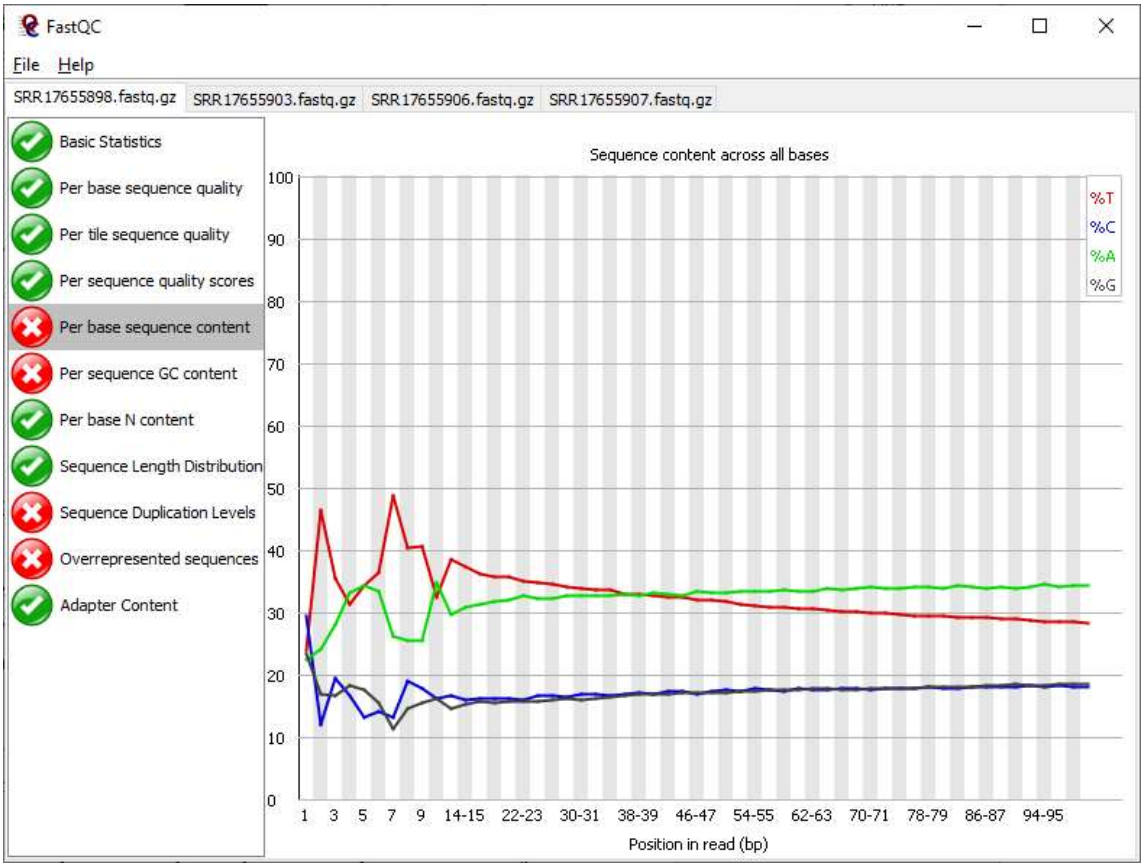
Each tab includes an evaluation for each module, which indicates whether the results are normal (green tick), slightly abnormal (orange triangle), or very unusual (red cross). By carefully reviewing these evaluations and visualizations, you can quickly identify any issues with your data and take appropriate corrective action, such as re-sequencing or trimming the data.

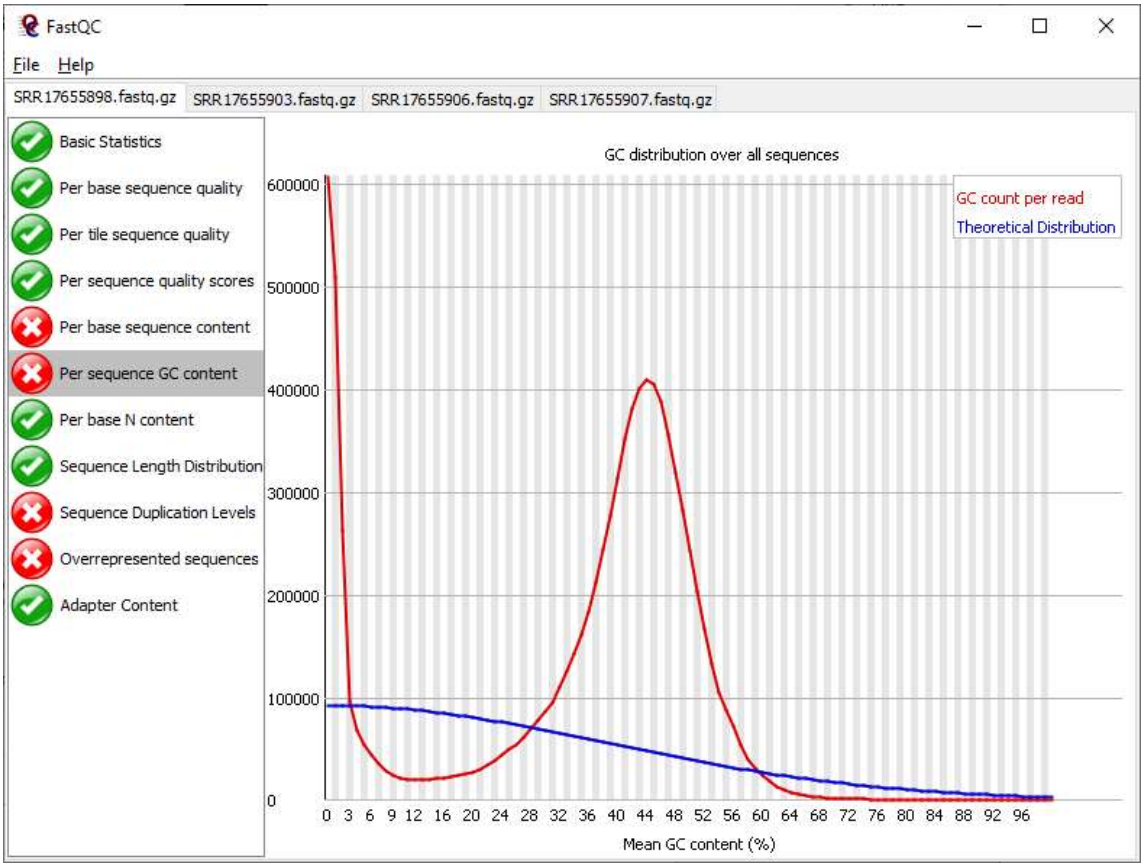

FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Basic sequence stats
Measure
Value
Filename
SRR17655898.fastq.gz
File type
Conventional base calls
Encoding
Sanger / Illumina 1.9
Total Sequences
8650483
Sequences flagged as poor quality
0
Sequence length
101
%GC
34

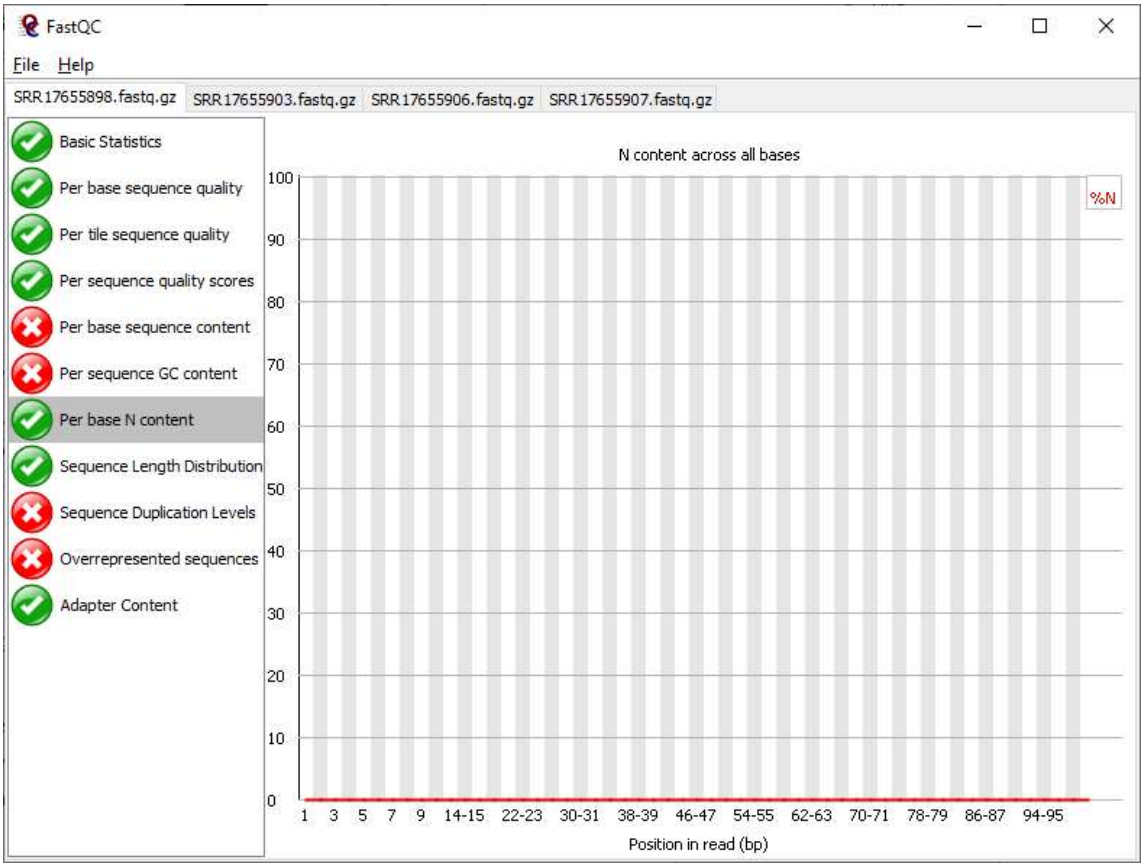

FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 3 5 7 9 14-15 22-23 30-31 38-39 46-47 54-55 62-63 70-71 78-79 86-87 94-95
Position in read (bp)

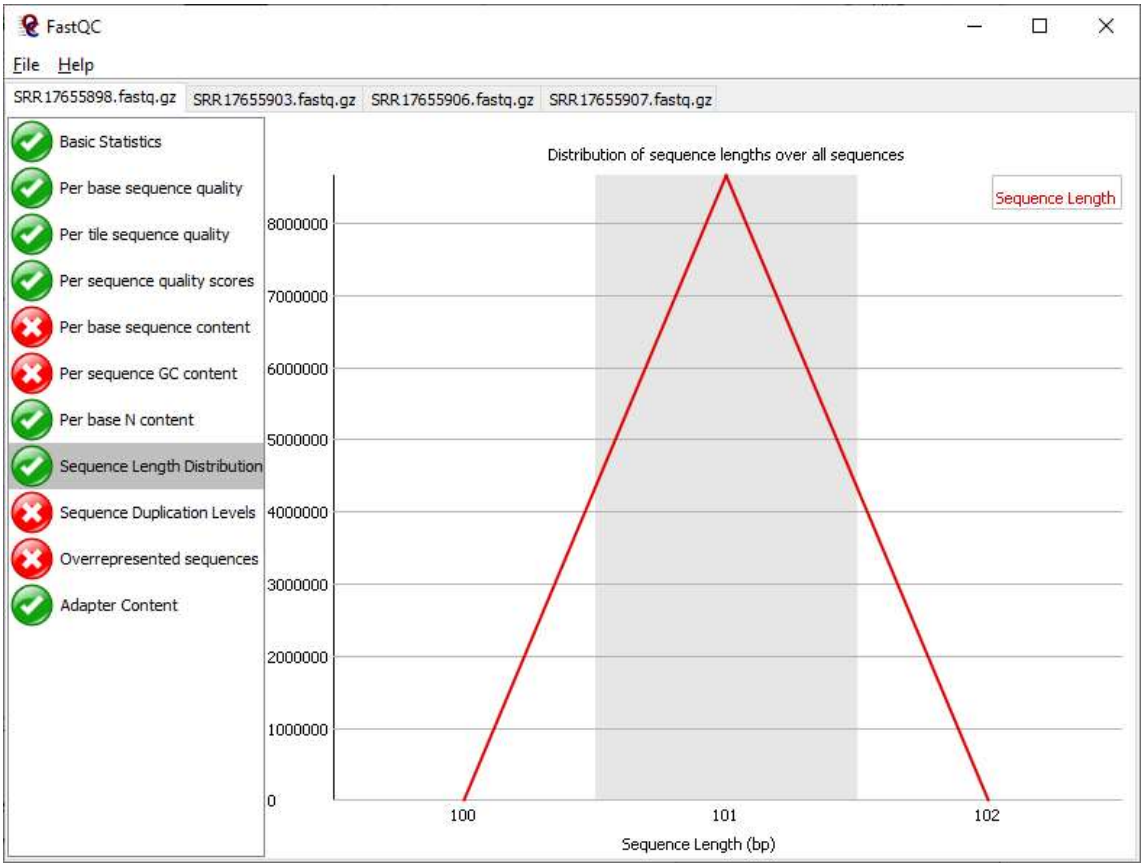

FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Quality per tile
2314
2311
2308
2305
2302
2215
2212
2209
2206
2203
2116
2113
2110
2107
2104
2101
1314
1311
1308
1305
1302
1215
1212
1209
1206
1203
1116
1113
1110
1107
1104
1101
1 3 5 7 9 14-15 22-23 30-31 38-39 46-47 54-55 62-63 70-71 78-79 86-87 94-95
Position in read (bp)

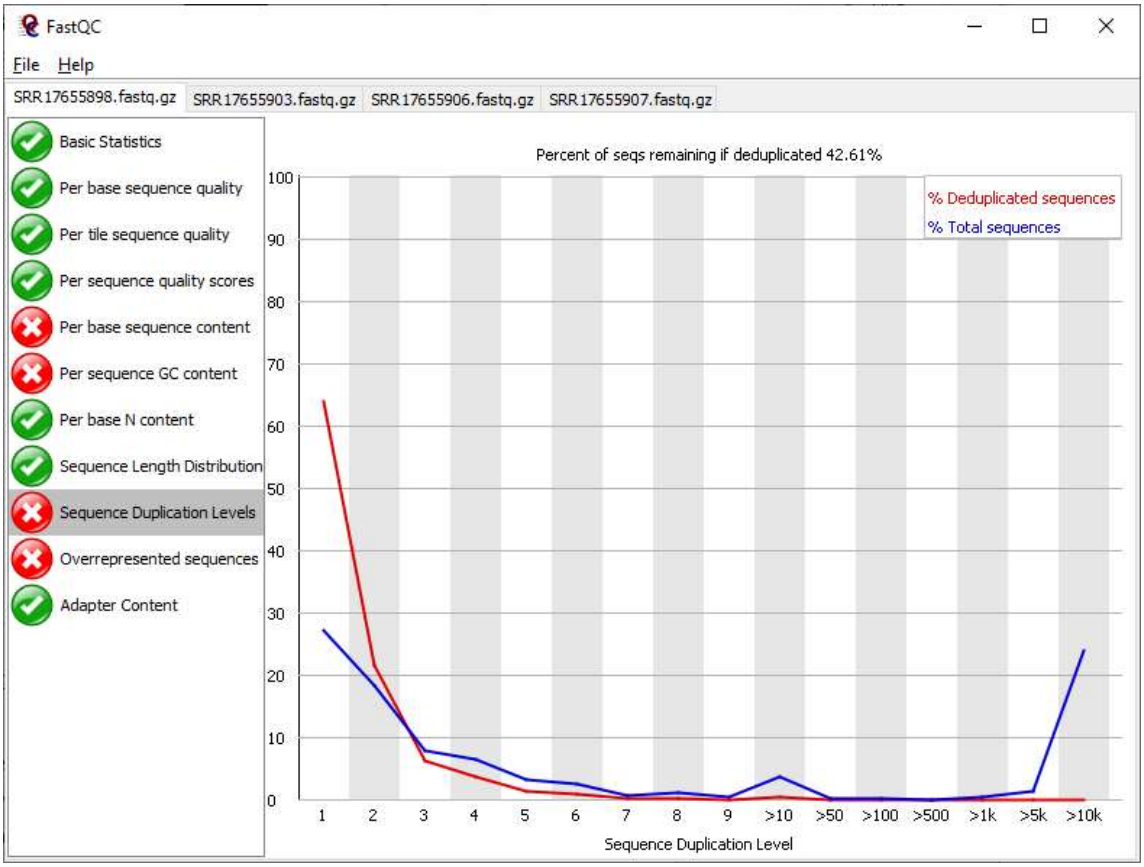

FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Quality score distribution over all sequences
Average Quality per read
5000000
4000000
3000000
2000000
1000000
0
7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Mean Sequence Quality (Phred Score)

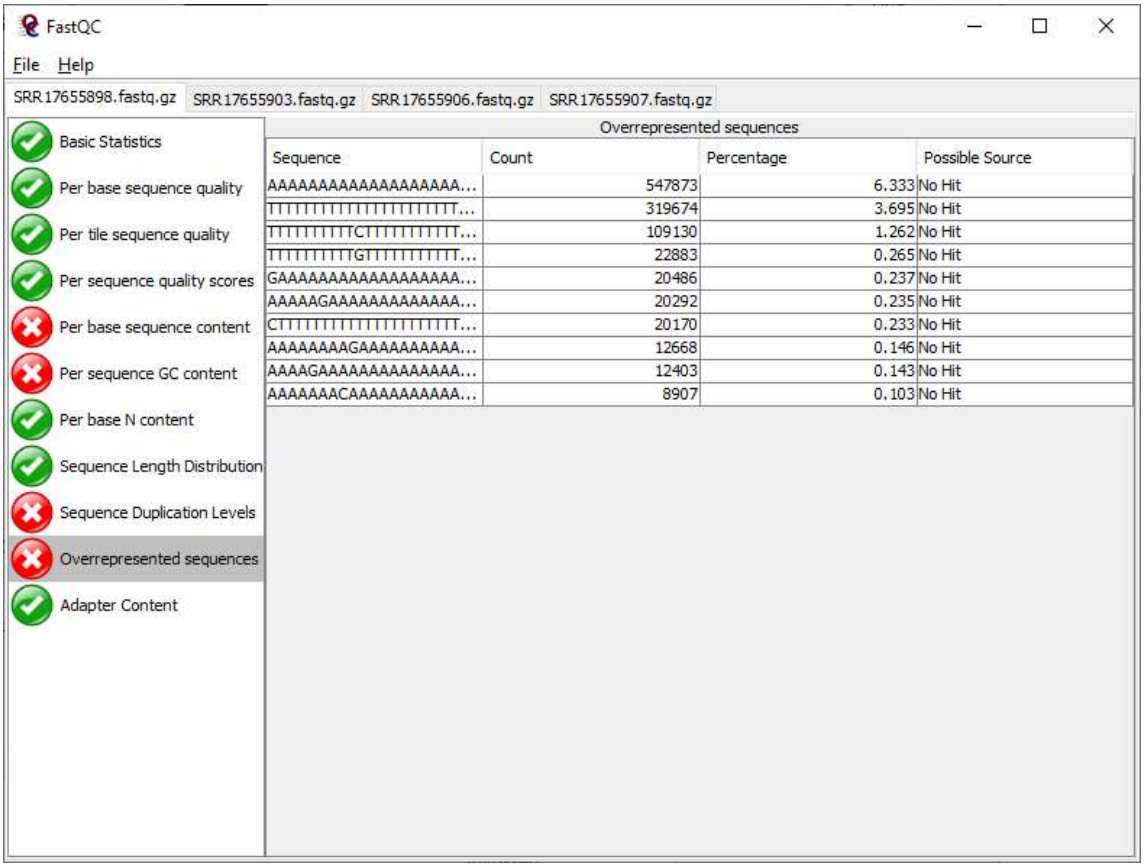

FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 3 5 7 9 14-15 22-23 30-31 38-39 46-47 54-55 62-63 70-71 78-79 86-87 94-95
Position in read (bp)

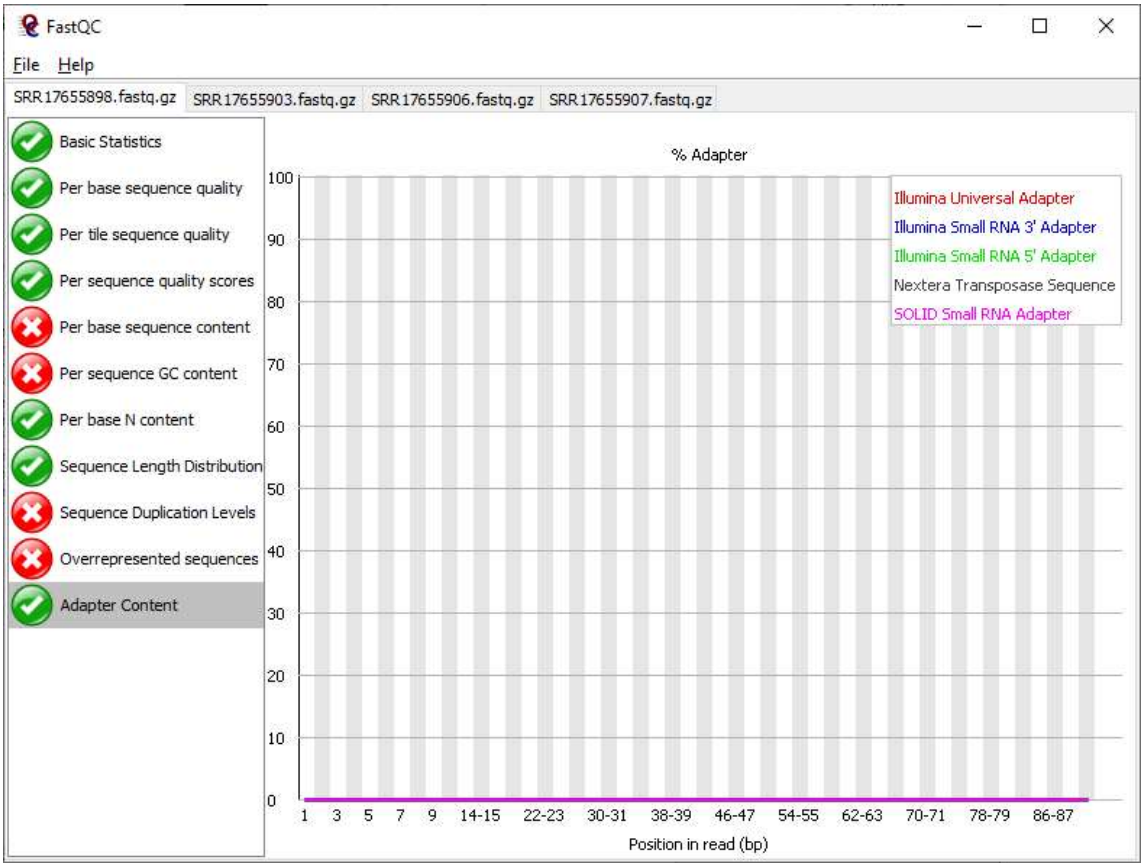

FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
GC distribution over all sequences
GC count per read
Theoretical Distribution
600000
500000
400000
300000
200000
100000
0
0 3 6 9 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96
Mean GC content (%)


FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
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
N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
1 3 5 7 9 14-15 22-23 30-31 38-39 46-47 54-55 62-63 70-71 78-79 86-87 94-95
Position in read (bp)


FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
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
Distribution of sequence lengths over all sequences
Sequence Length
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
100
101
102
Sequence Length (bp)


FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
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
Percent of seqs remaining if deduplicated 42.61%
% Deduplicated sequences
% Total sequences
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k
Sequence Duplication Level


FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
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
Overrepresented sequences
Sequence Count Percentage Possible Source
AAAAAAAAAAAAAAAAAAA... 547873 6.333 No Hit
TTTTTTTTTTTTTTTTTTTT... 319674 3.695 No Hit
TTTTTTTTTCTTTTTTTTTT... 109130 1.262 No Hit
TTTTTTTTTGTTTTTTTTTT... 22883 0.265 No Hit
GAAAAAAAAAAAAAAAAAA... 20486 0.237 No Hit
AAAAAGAAAAAAAAAAAAA... 20292 0.235 No Hit
CTTTTTTTTTTTTTTTTTTT... 20170 0.233 No Hit
AAAAAAAAGAAAAAAAAAA... 12668 0.146 No Hit
AAAAGAAAAAAAAAAAAAA... 12403 0.143 No Hit
AAAAAAACAAAAAAAAAAA... 8907 0.103 No Hit


FastQC
File Help
SRR17655898.fastq.gz
SRR17655903.fastq.gz
SRR17655906.fastq.gz
SRR17655907.fastq.gz
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
% Adapter
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
100
90
80
70
60
50
40
30
20
10
0
1 3 5 7 9 14-15 22-23 30-31 38-39 46-47 54-55 62-63 70-71 78-79 86-87
Position in read (bp)