

Step 1. Learn about the experiment and get a file for analysis.

ENA

European Nucleotide Archive

Home

Submit

Search

Rulespace

About

Support

Examples: histone, BN000065

ERR9974118

View

Examples: Taxon:9606, BN000065, PRJEB402

The ENA Advanced Search API is changing on 2023-05-02! Details [here](#).

Run: ERR9974118

Illumina HiSeq 2500 paired end sequencing; Single cell RNA-seq of Pgc+ cells in homeostasis and upon injury

Organism: Mus musculus (house mouse)

Sample Accession: SAMEA11859817

Instrument Platform: ILLUMINA

Instrument Model: Illumina HiSeq 2500

Read Count: 3867822

Base Count: 386782200

Center Name: Department of New Biology, DGIST

Library Layout: PAIRED

Library Strategy: RNA-Seq

Library Source: TRANSCRIPTOMIC SINGLE CELL

View: XML

Download: XML

Navigation: Show

Read Files: Hide

Description of the experiment:

- Organism – Mus musculus (house mouse)
- Sequencing platform – Illumina HiSeq 2500
- Reads (paired/unpaired) – paired

Read Files

Show Column Selection

Download report: JSON TSV

Download Files as ZIP

Download selected files

Download All

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Generated FASTQ files: FTP	Size
PRJEB54870	SAMEA11859817	ERX9515112	ERR9974118	10090	Mus musculus	<input checked="" type="checkbox"/> ERR9974118_1.fastq.gz <input type="checkbox"/> ERR9974118_2.fastq.gz	2.1 MB

Items per page: 10 1 - 1 of 1

This PC > Data (D:) > ena_files > ERR9974118 > ERR9974118_1.fastq

Name	Date modified	Type	Size
ERR9974118_1	4/20/2023 4:09 PM	FASTQ Sequence T...	481,422 KB

Step 2. Install the FASTQC program.

```
Select magenta@magenta: ~/fastqc
inflating: FastQC/org/apache/commons/math3/exception/MathInternalError.class
inflating: FastQC/org/apache/commons/math3/exception/NotStrictlyPositiveException.class
creating: FastQC/org/apache/commons/math3/exception/util/
inflating: FastQC/org/apache/commons/math3/exception/util/ExceptionContextProvider.class
inflating: FastQC/org/apache/commons/math3/exception/util/ArgUtils.class
inflating: FastQC/org/apache/commons/math3/exception/util/ExceptionContext.class
inflating: FastQC/org/apache/commons/math3/exception/util/Localizable.class
inflating: FastQC/org/apache/commons/math3/exception/util/LocalizedFormats.class
inflating: FastQC/org/apache/commons/math3/exception/TooManyEvaluationsException.class
inflating: FastQC/org/apache/commons/math3/exception/MathIllegalStateException.class
inflating: FastQC/org/apache/commons/math3/exception/MathArithmeticException.class
inflating: FastQC/org/apache/commons/math3/exception/MathIllegalNumberException.class
inflating: FastQC/org/apache/commons/math3/exception/MathIllegalArgumentException.class
inflating: FastQC/org/apache/commons/math3/exception/OutOfRangeException.class
inflating: FastQC/org/apache/commons/math3/exception/NumberIsTooSmallException.class
inflating: FastQC/org/apache/commons/math3/exception/MaxCountExceededException.class
inflating: FastQC/org/apache/commons/math3/exception/NotPositiveException.class
inflating: FastQC/org/apache/commons/math3/exception/NoBracketingException.class
inflating: FastQC/org/apache/commons/math3/exception/NotFiniteNumberException.class
inflating: FastQC/org/apache/commons/math3/exception/NullArgumentException.class
inflating: FastQC/org/apache/commons/math3/exception/DimensionMismatchException.class
inflating: FastQC/org/apache/commons/math3/exception/ConvergenceException.class
inflating: FastQC/org/apache/commons/math3/exception/NumberIsTooLargeException.class
creating: FastQC/org/apache/commons/math3/random/
inflating: FastQC/org/apache/commons/math3/random/Well19937c.class
inflating: FastQC/org/apache/commons/math3/random/AbstractWell.class
inflating: FastQC/org/apache/commons/math3/random/RandomGenerator.class
inflating: FastQC/org/apache/commons/math3/random/RandomDataImpl.class
inflating: FastQC/org/apache/commons/math3/random/RandomData.class
inflating: FastQC/org/apache/commons/math3/random/BitStreamGenerator.class
creating: FastQC/org/apache/commons/math3/analysis/
inflating: FastQC/org/apache/commons/math3/analysis/UnivariateFunction.class
creating: FastQC/org/apache/commons/math3/analysis/solvers/
inflating: FastQC/org/apache/commons/math3/analysis/solvers/BaseAbstractUnivariateSolver.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/AllowedSolution.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/UnivariateSolverUtils.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/BrentSolver.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/UnivariateSolver.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/BaseUnivariateSolver.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/AbstractUnivariateSolver.class
inflating: FastQC/org/apache/commons/math3/analysis/solvers/BracketedUnivariateSolver.class
magenta@magenta:~$ cd FastQC
bash: cd: FastQC: No such file or directory
magenta@magenta:~$ cd FastQC
magenta@magenta:~/FastQC$ ls
class-jhdf5.jar  fastqc_icon.ico  INSTALL.txt  LICENSE_JHDF5.txt  org  RELEASE_NOTES.txt  uk
configuration  Help  jbrp2-0.9.jar  LICENSE.txt  README.md  run_fastqc.bat
fastqc  htseq.jar  LICENSE  net  README.txt  Templates
magenta@magenta:~/FastQC$ ./fastqc
```

Step 3. Analyze your file.

1. General statistics:

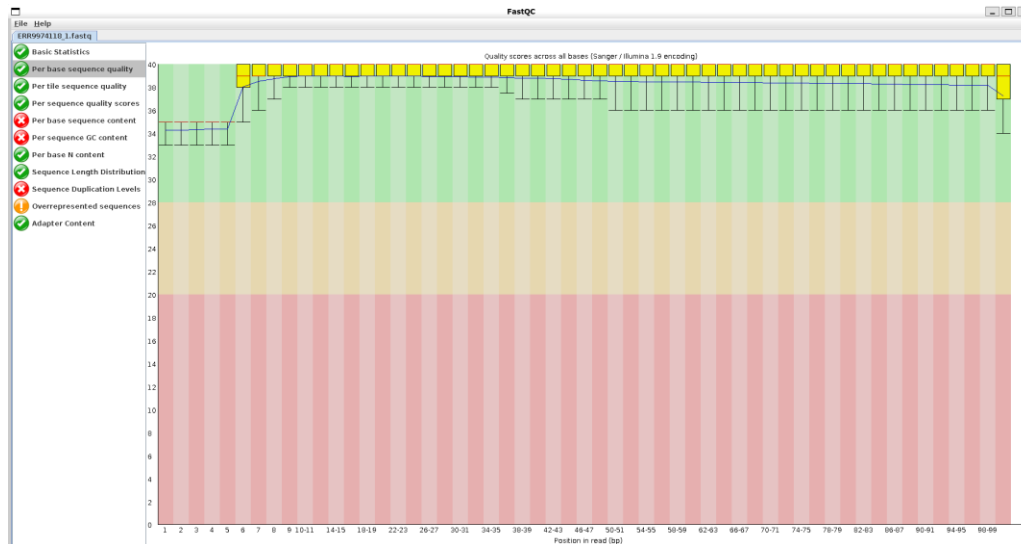
Sequence length – 100

Number of sequences – 1933911

FastQC			
File Help			
ERR9974118.1.fastq			
Basic Statistics		Basic sequence stats	
		Measure	Value
✓ Per base sequence quality	Filename	ERR9974118.1.fastq	
✓ Per tile sequence quality	File type	Conventional base calls	
✓ Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
✗ Per base sequence content	Total Sequences	1933911	
✗ Per sequence GC content	Total Sizes	193.3 Mbp	
✓ Per base N content	Sequences flagged as poor quality	0	
✓ Sequence Length Distribution	Sequence length	100	
✗ Sequence Duplication Levels	%GC	50	
⚠ Overrepresented sequences			
✓ Adapter Content			

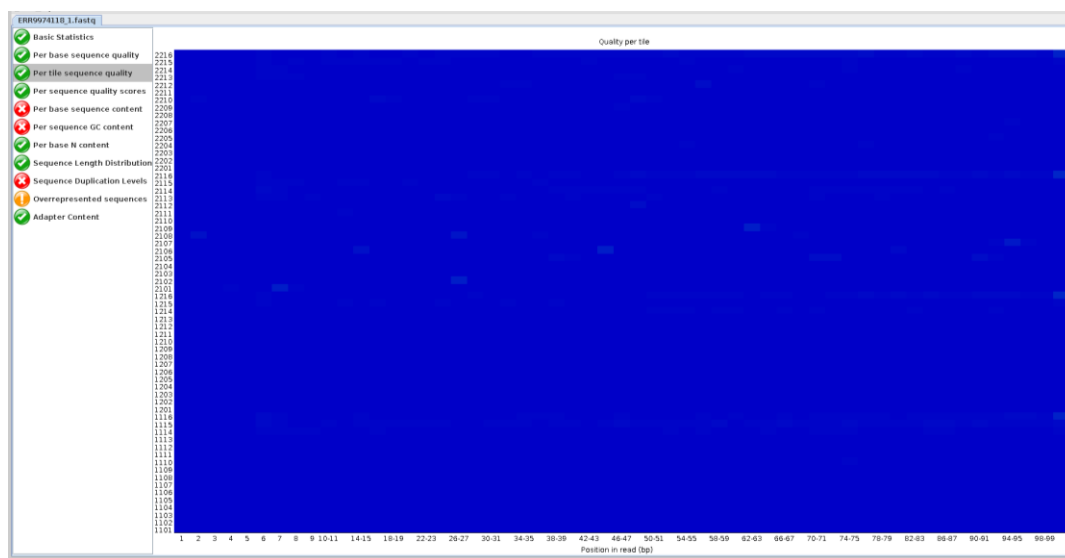
2. Quality of individual nucleotides and average quality of reads.

A box-and-whisker plot showing aggregated quality score statistics at each position along all reads in the file. It is normal with all Illumina sequencers for the median quality score to start out lower over the first 5-7 bases and to then rise. The average quality score will steadily drop over the length of the read. The blue line represents the mean quality. Average quality of reads is high (the whiskers are in the green zone). Individual nucleotides have high quality. The first 4 nucleotides have the quality score = 34, and the following ones have the quality score = 40.



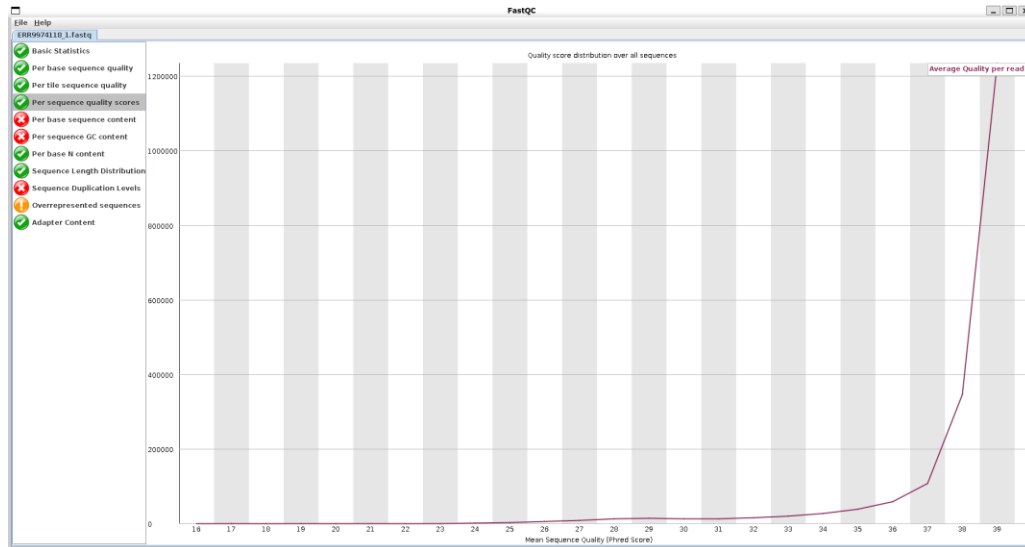
3. Per Tile Sequence Quality

A good plot should be blue all over. My picture shows an ideal case, where no quality loss is reported in any tile of the flow cell.



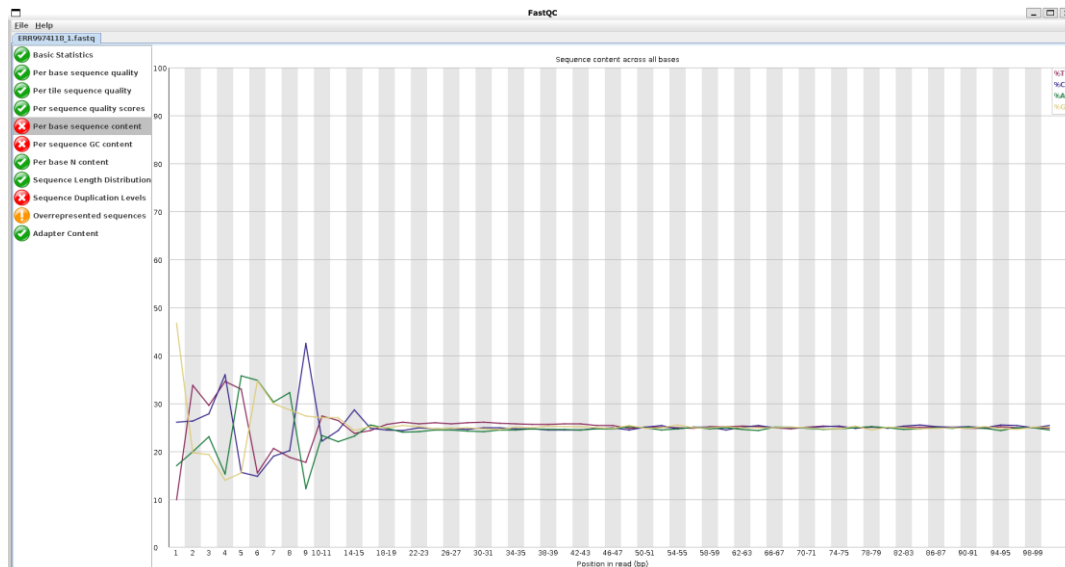
4. Per Sequence Quality Scores

A plot of the total number of reads vs the average quality score (Phred score) over full length of that read. The distribution of average read quality should be fairly tight in the upper range of the plot. My quality is high.



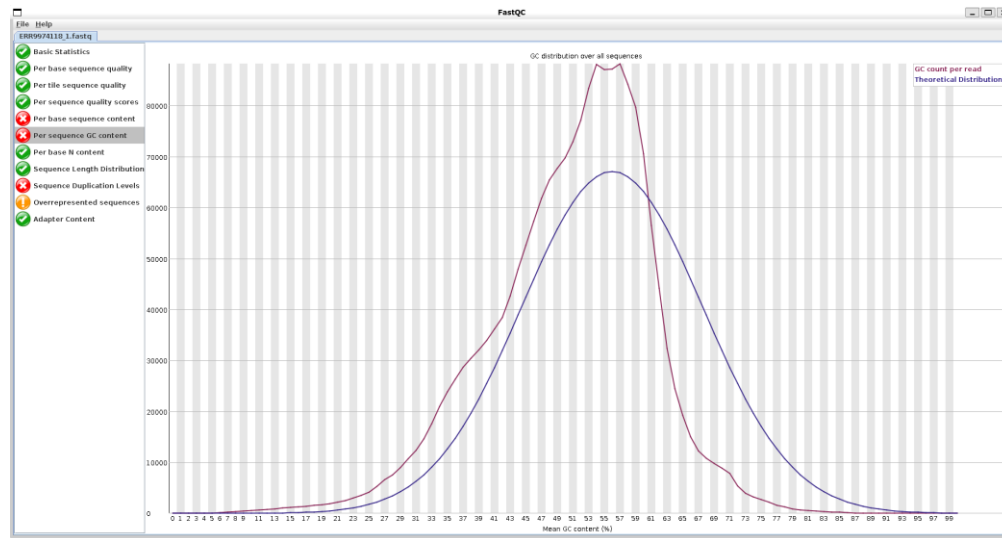
5. Per Base Sequence Content

This plot reports the percentage of each of the four nucleotides (T, C, A, G) at each position across all reads in the input sequence file. I work with RNA sequence. With most RNA-Seq library preparation protocols there is clear non-uniform distribution of bases for the first 10-15 nucleotides; this is normal and expected. RNA-Seq data showing this non-uniform base composition will always be classified as Failed by FastQC for this module even though the sequence is perfectly good.



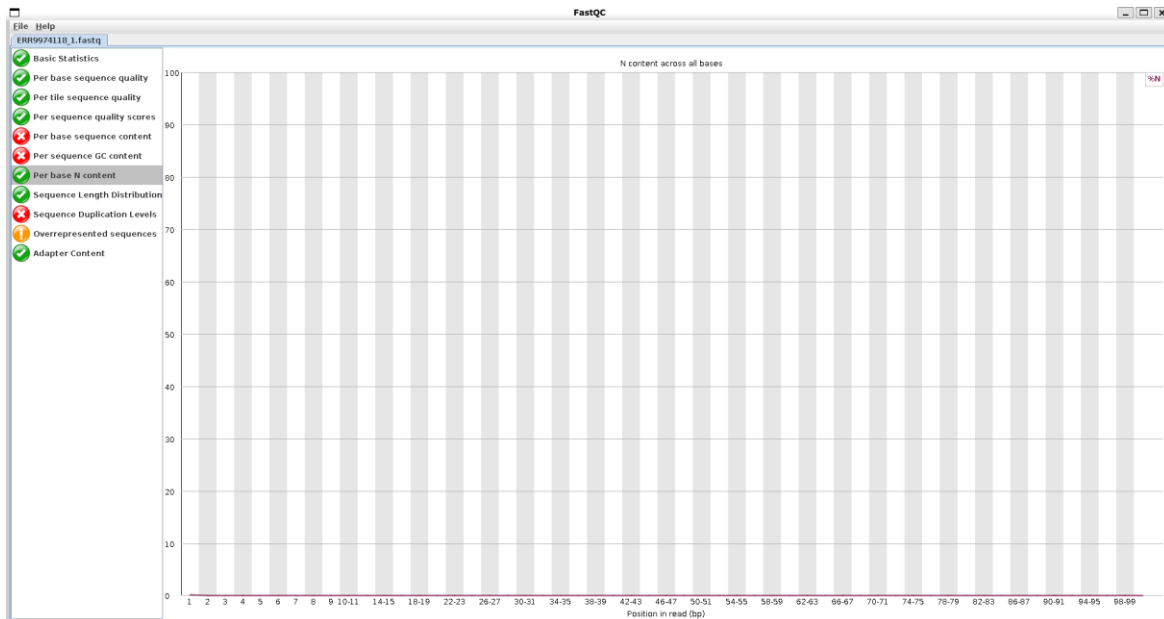
6. Per Sequence GC Content, what distribution do you see?

This module measures the GC content across the whole length of each sequence and compares it to a normal distribution of GC content. In RNA sequencing there may be a greater or lesser distribution of mean GC content among transcripts causing the observed plot to be wider or narrower than an ideal normal. My plot is from high quality RNA-Seq data, but FastQC classified it as Failed because the observed distribution is narrower than the theoretical and is shifted a bit.



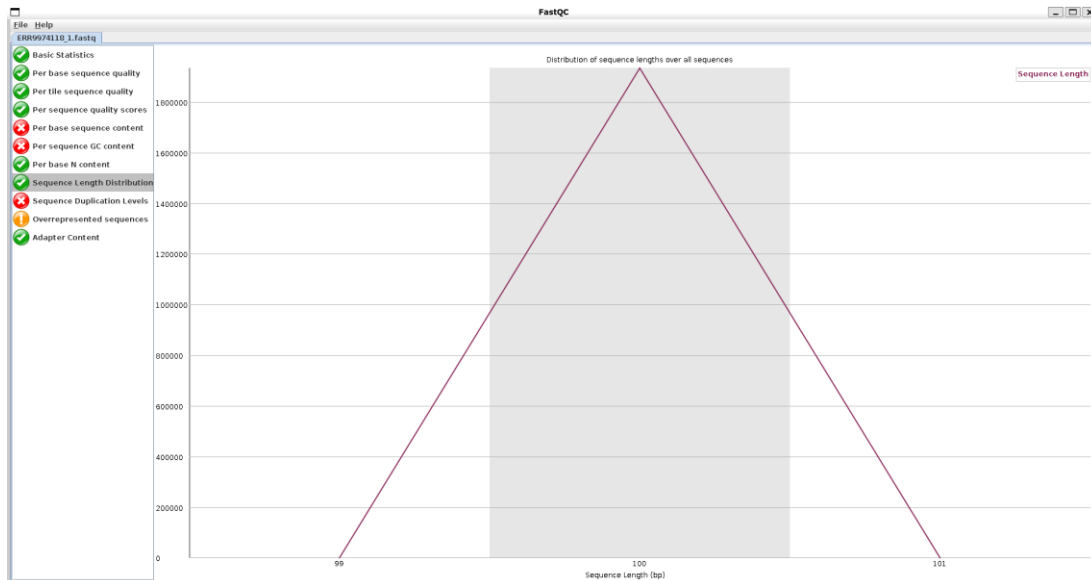
7. Per Base N Content

Percent of bases at each position or bin with no base call, i.e. 'N'. You should never see any point where this curve rises noticeably above zero. My plot shows that everything is good, there is no point where the curve rises above the zero.



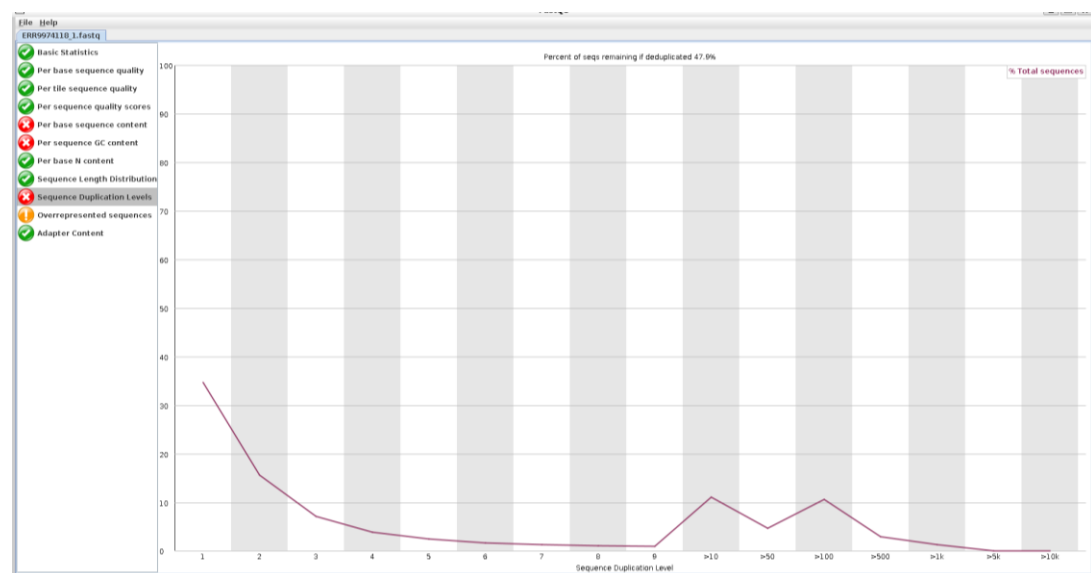
8. Sequence Length Distribution, are there sequences that differ in length?

It produced a simple graph showing a peak only at one size. No sequences that differ in length.



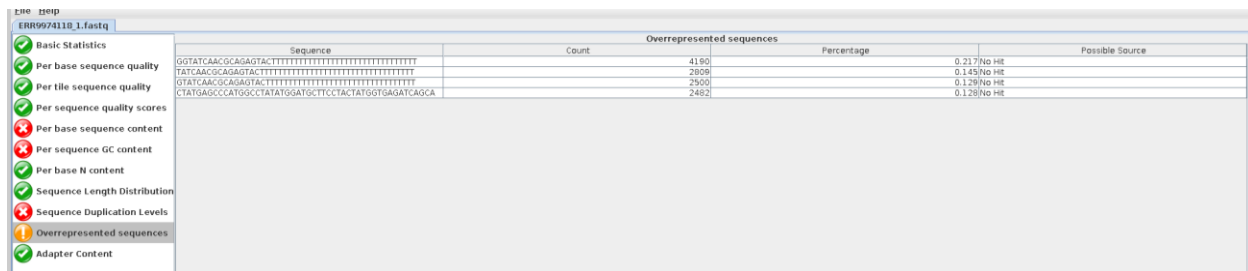
9. Duplicate Sequences, low or high duplication?

Percentage of reads of a given sequence in the file which are present a given number of times in the file. There are generally two sources of duplicate reads: PCR duplication in which library fragments have been over represented due to biased PCR enrichment or truly over represented sequences such as very abundant transcripts in an RNA-Seq library. When sequencing RNA there will be some very highly abundant transcripts and some lowly abundant. It is expected that duplicate reads will be observed for high abundance transcripts. My RNA-Seq data was flagged as Failed by FastQC. High duplication levels in my RNA-seq are normal and expected.



10. Overrepresented Sequences, do they exist? if yes, is anything known about them?

List of sequences which appear more than expected in the file. Only the first 50bp are considered. A sequence is considered overrepresented if it accounts for $\geq 0.1\%$ of the total reads. Each overrepresented sequence is compared to a list of common contaminants to try to identify it. For RNA-Seq data it is possible that there may be some transcripts that are so abundant that they register as overrepresented sequence. I have such overrepresented sequences.



Overrepresented sequences			
Sequence	Count	Percentage	Possible Source
GGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTT	4190	0.217%	No hit
TATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTT	2909	0.145%	No hit
GTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTT	2500	0.129%	No hit
CTATGAGCCCATGCTATATGGATGCTTCTACTATGOTGAGATCA	2482	0.128%	No hit

11. Availability of Adapter Content, if it will

Cumulative plot of the fraction of reads where the sequence library adapter sequence is identified at the indicated base position. Ideally Illumina sequence data should not have any adapter sequence present, however when using long read lengths it is possible that some of the library inserts are shorter than the read length resulting in read-through to the adapter at the 3' end of the read. This is more likely to occur with RNA-Seq libraries where the distribution of library insert sizes is more varied and likely to include some short inserts. My plot shows no adapter content.

