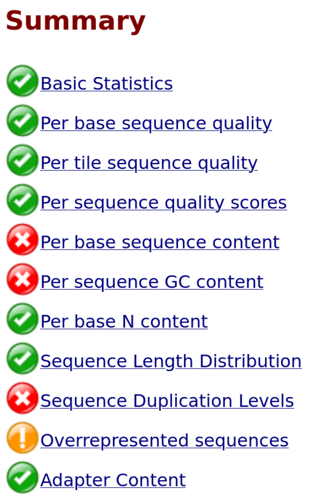
Home work 3

- **Organism** - Mus musculus
- **Sequencing platform** - Illumina
- **Reads (paired/unpaired)** - paired

SAMPLE: ERR9974118

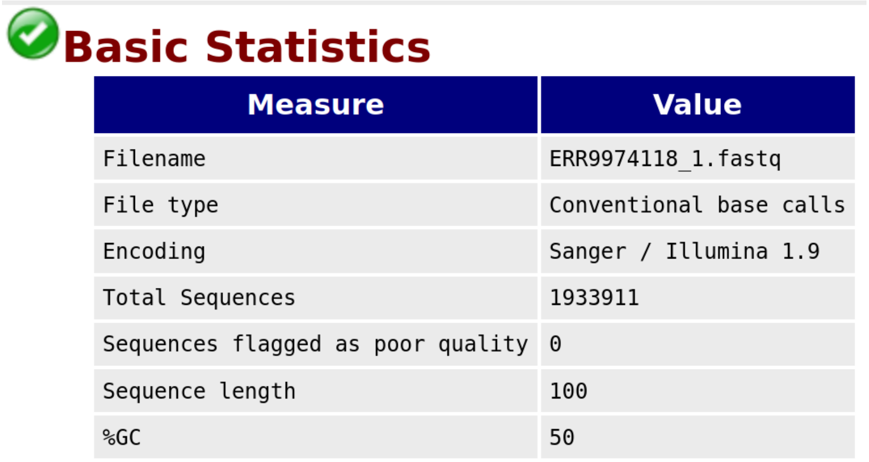


Summary

- ✓ 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

The fastq file has problems with Per base sequence content, Per sequence GC content, Sequence Duplication Levels and also a moderate quality of Overrepresented sequences.

- **General statistics: number of reads and their length (5 point)**

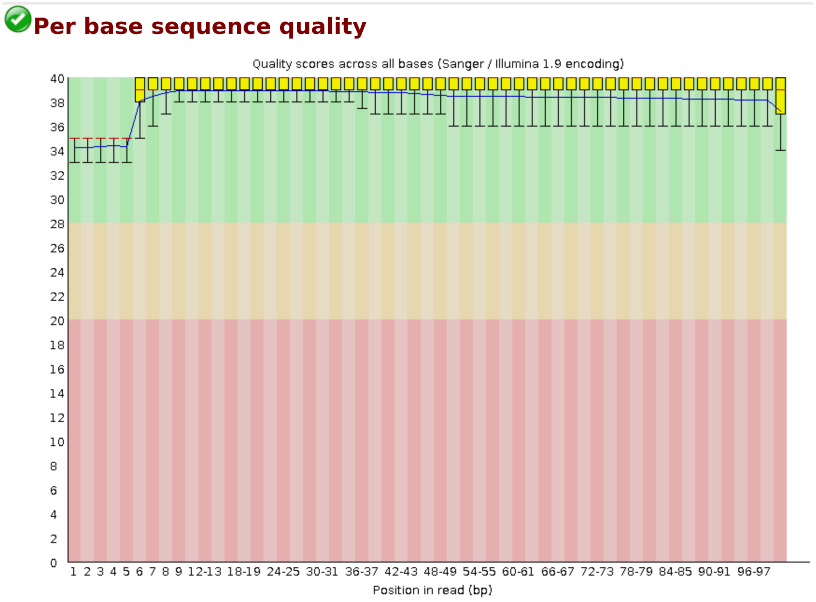


✓ Basic Statistics

Measure	Value
Filename	ERR9974118_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1933911
Sequences flagged as poor quality	0
Sequence length	100
%GC	50

Total number of reads is 1933911
Length of reads is 100

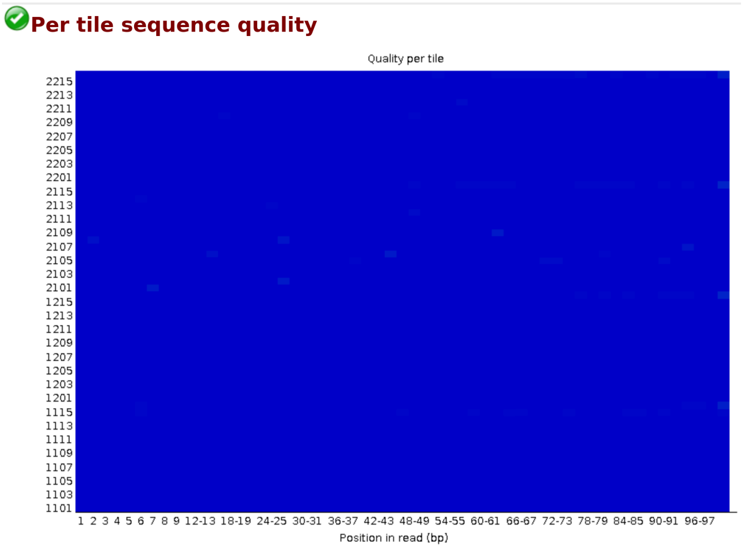
- **Quality of individual nucleotides and average quality of reads (5 point)**



✓ Per base sequence quality

An extremely good per base sequence quality and as so the average quality of reads is also good. Minor fall in quality at the beginning of reads (adapters?), but still very good.

- **Per Tile Sequence Quality (5 point)**

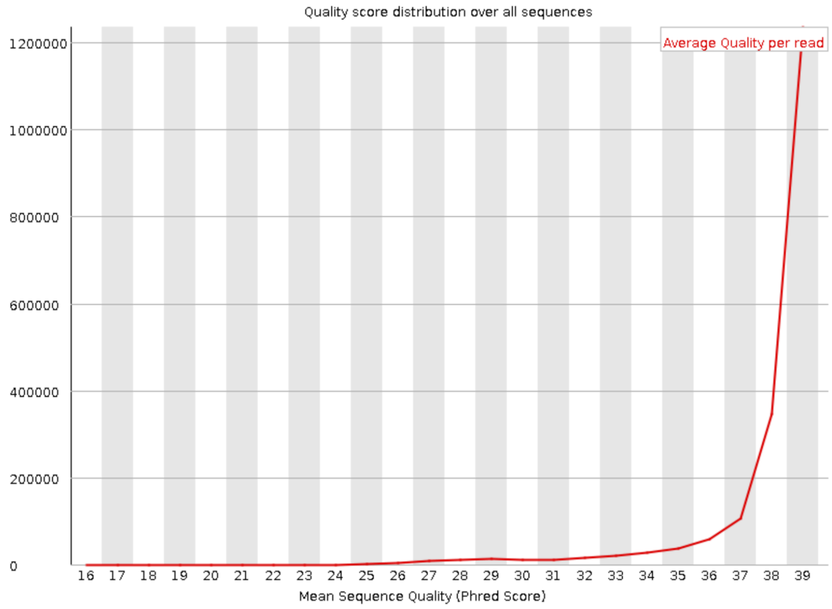


✓ Per tile sequence quality

As can be seen from picture per tile sequence quality distributed equally across all tiles. So no problem can be admitted with this point.

- **Per Sequence Quality Scores (5 point)**

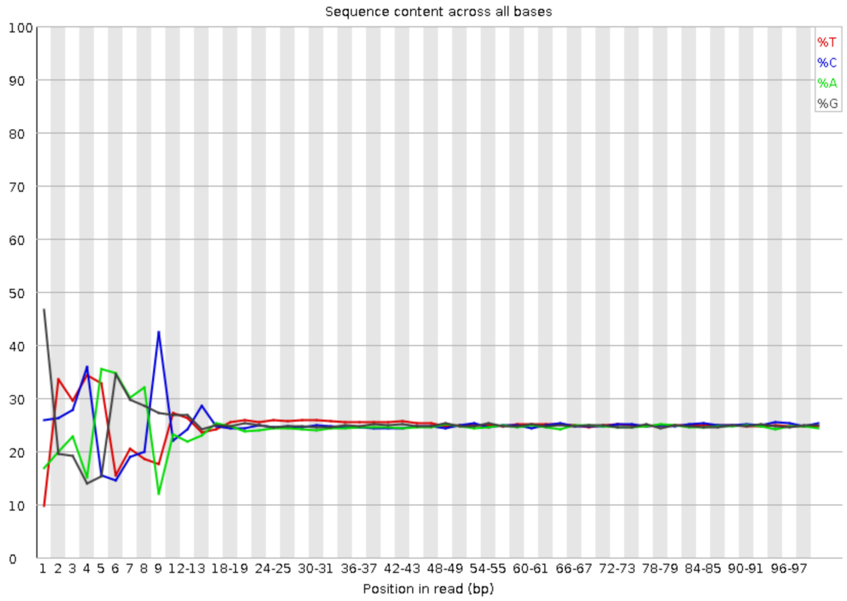
Per sequence quality scores



Average quality per read has one peak at 38 Phred Score (which is good). And no other peaks could be seen. So no problem.

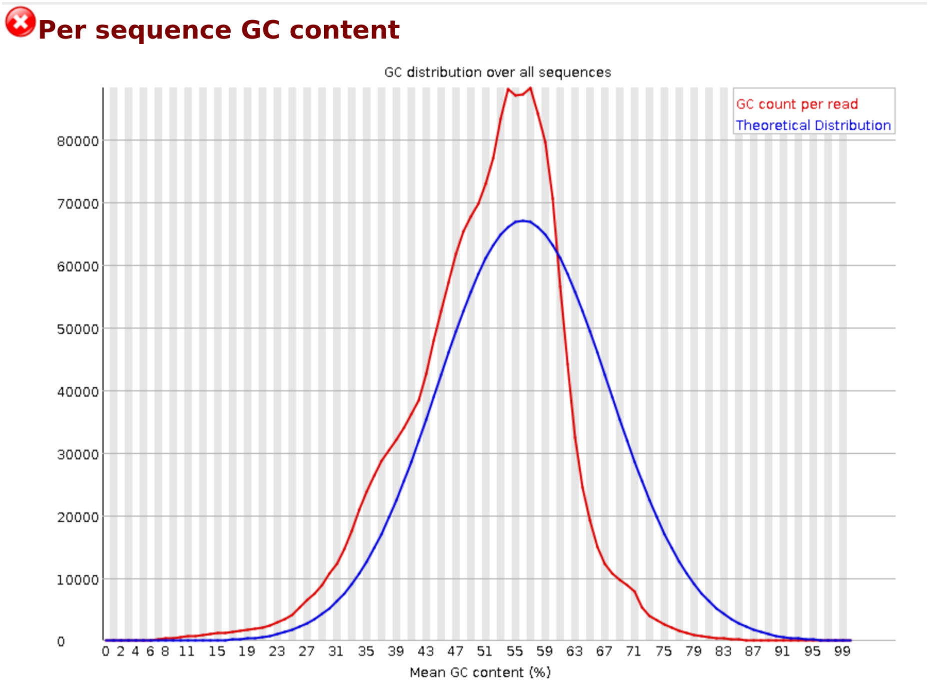
- **Per Base Sequence Content (5 point)**

Per base sequence content



There is some preference in relation to the location of nucleotides at the 5’ end of reads. (for example, at position 9, A is more common in reads). This can be explained by the presence of adapters. In order to get rid of adapters, reads can be trimmed, for example, using trimmomatic.

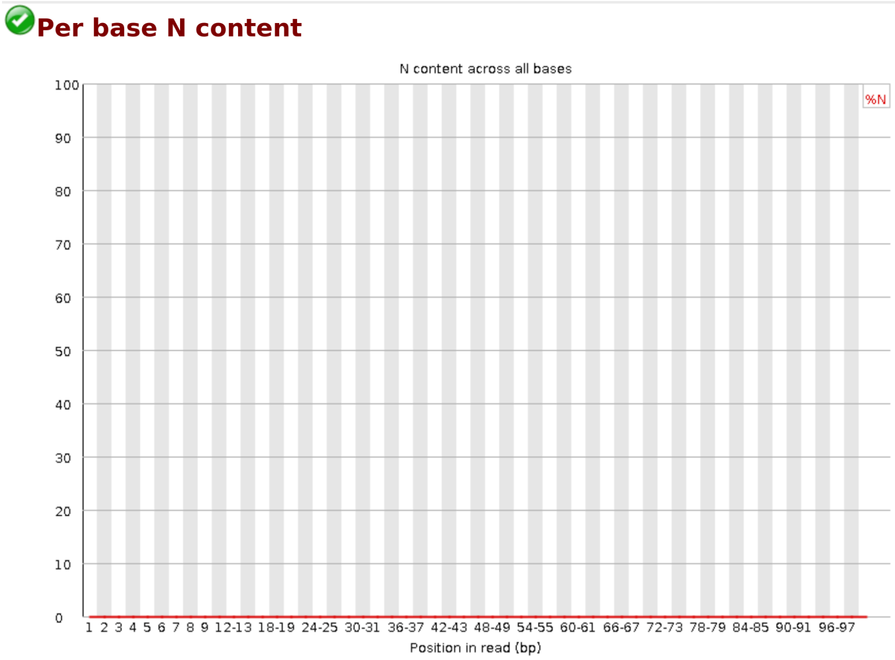
- **Per Sequence GC Content, what distribution do you see? (5 point)**



Per sequence GC content

GC distribution is close to the normal one. However it has two peaks, which might be an artifact. I don’t know if it’s a problem and if it is, how to fix it.

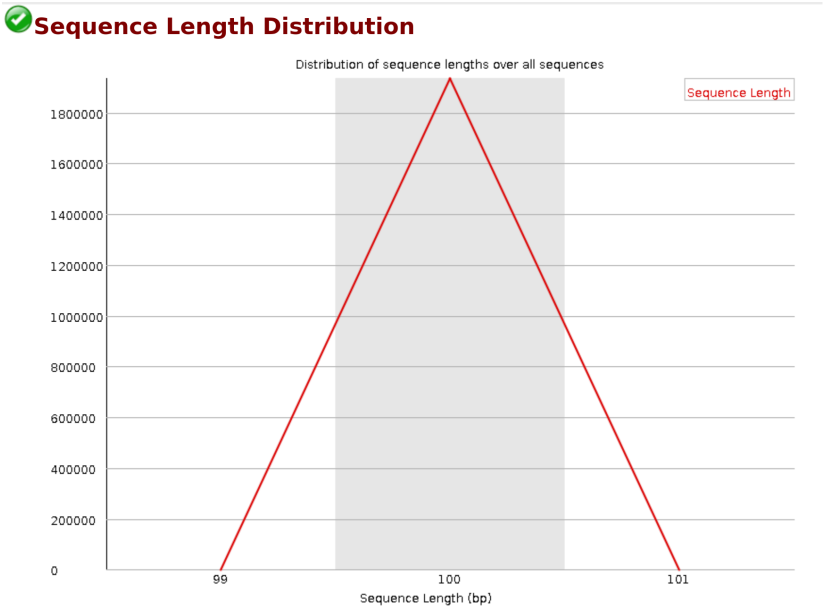
- **Per Base N Content (5 point)**



Per base N content

Zero Ns in every position, so no problem.

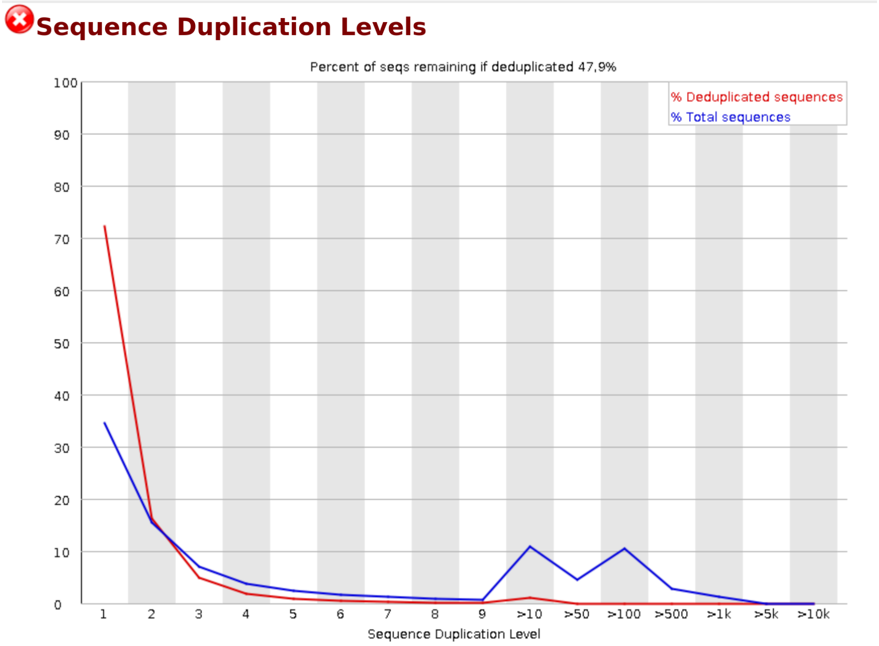
- **Sequence Length Distribution, are there sequences that differ in length? (5 point)**



✓ Sequence Length Distribution

All sequences have the same length.

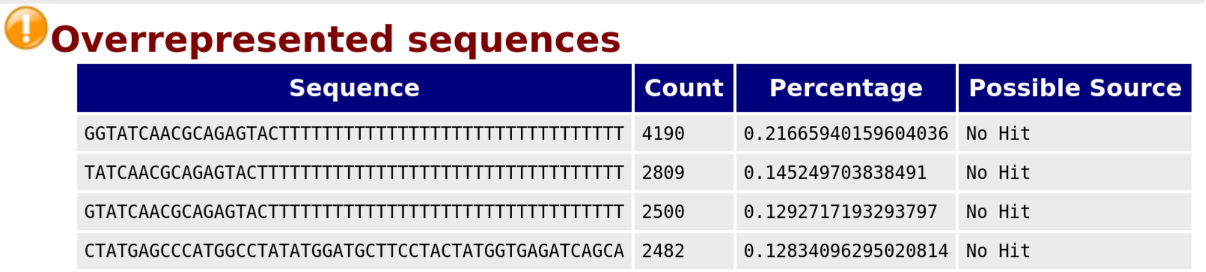
- **Duplicate Sequences,low or high duplication? (5 point)**



☒ Sequence Duplication Levels

Given that this is RNA sequencing, the presence of overrepresented sequences such as very abundant transcripts is expected. So no problem.

- **Overrepresented Sequences, do they exist? if yes, is anything known about them? (5 point)**

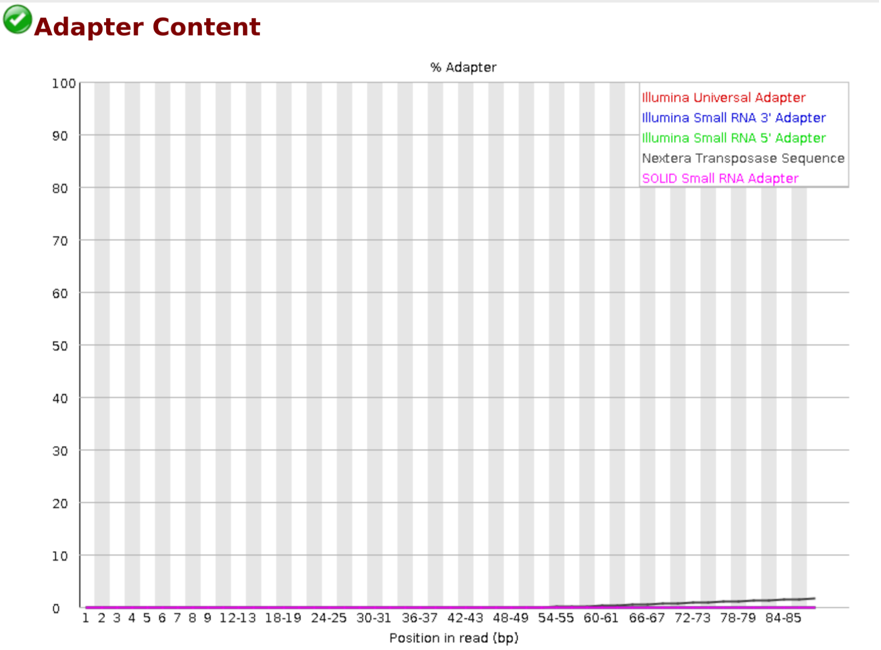


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	4190	0.21665940159604036	No Hit
TATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	2809	0.145249703838491	No Hit
GTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	2500	0.1292717193293797	No Hit
CTATGAGCCCATGGCCTATATGGATGCTTCCTACTATGGTGAGATCAGCA	2482	0.12834096295020814	No Hit

There are overrepresented sequences in this experiment.
The first three sequences don't align to anything specific, so they could just be poly A tails of the mRNA.
The latter sequence aligns perfectly using BLAST to mouse progastricsin (pepsinogen C) mRNA, which, given that the sequence was from a mouse, is probably normal.

- **Availability of Adapter Content, if it will (5 point)**



Adapter Content

No problems with adapters.