

# Assignment report, Advanced sequencing and genome assembly

MSc Applied Bioinformatics,

Marie Schmit

# Assembly

## Quality control and genome size

Illumina reads

A control quality is perform on the raw reads using FastQC. Read1 has 2 475 000 reads and a sequence length of 101. The mean quality of it’s reads is very poor, with a value of 17. No sequence is flagged as poor quality.

Chart

Description automatically generated

Figure Quality score across bases of first Illumina read

Its per sequence quality score is also too low, with a mean sequence quality of 17 (more than 1% error rate). The per base content of the first read is good, with little to no differences between bases: no bias was introduced in the reads.

The GC contents presents a warning as its distribution is slightly shifted, which could indicate a bias that does not depend on base position.

Chart, histogram

Description automatically generated

Figure GC content of first Illumina read

All the other metrics (per base N content, sequence length distribution, overrepresented sequence and adapter content) are good.

The results are the same for Illumina second reads. The quality per base sequence are problematic. Indeed, they are all the same and very low. Thought, no sequence is flagged as poor quality. A problem might have occurred with the sequencing. I will neither trim nor filter those data to improve their general quality, since all the values are the same and are poor. No data would be left after trimming or filtering.

A k-mer analysis is then performed with Jellyfisch, from MaSURCA on both short reads. 2 threads, 1G of elements hash are used. Different values of k-mer are tested: 19; 25; 30. The number of kmer is calculation with “count” function and a histogram is computed using the function “histo”. The results are plot with R.

|  |  |  |
| --- | --- | --- |
| 19mer, R1 | 25mer, R1 | 30mer, R1 |
|  |  |  |
|  |  |  |

The results are the same for the second reads.

The first major peak of the histogram, at count 1, is the error component. Errors boost the number of kmers appearing once. Apart from this error peak, the distribution presents one major peak, which is characteristic of a rather good data quality.

A correction is applied to try to decrease the error peak, with KmerFreq\_AR and Corrector\_AR functions (parameters: 1 thread, 33 ASCII shift). The k parameter for the correction is 15 for lighter computation.

|  |  |  |
| --- | --- | --- |
| 19mer, R1 after correction | 25mer, R1 after correction | 30mer, R1 after correction |
|  |  |  |

The number of reads went from 2 475 000 to after 2 378 197 reads correction.

For 19 kmer, the error peak is still very high. Its density value has decreased, going from 5e7 before correction to 8e5 after correction, but the amelioration is small (only a difference of e2). The improvement is poorer with 25mer: the density went from 6e7 to 1e7. The same goes for 30mer, with a density going from 6e7 to 2e7 after correction.

The rest of the analysis will still be done with the correction, since the control quality with FastQC gave poor quality results.

PacBio reads

For this file, the number of sequences is 33 413. Their minimal length is 167 and the maximal is 246451. No sequences are flagged as poor quality and the percentage of GC content is 37.

The per base sequence quality is very poor, with a mean of 8.5. The quality is the same for almost all the reads, with a very light drop of 0.5 at the end. Since the quality is certainly bad but constant, no filtering nor trimming will be applied: all the values are bad.

Chart, bar chart

Description automatically generated

Figure Per base sequence quality, PacBio reads

The per sequence quality score is also flagged as failed. The mean quality value is around 8, and the maximal value is 10. It confirms the per base sequence quality: both are problematic.

Chart, line chart

Description automatically generated

Figure Average quality per read (Quality score distribution over all sequences), PacBio reads

The per base content is also flagged as a fail. The percentage of adenine is for instance 30% higher than the percentage of guanine for the last bases (between 23500 and 24000).

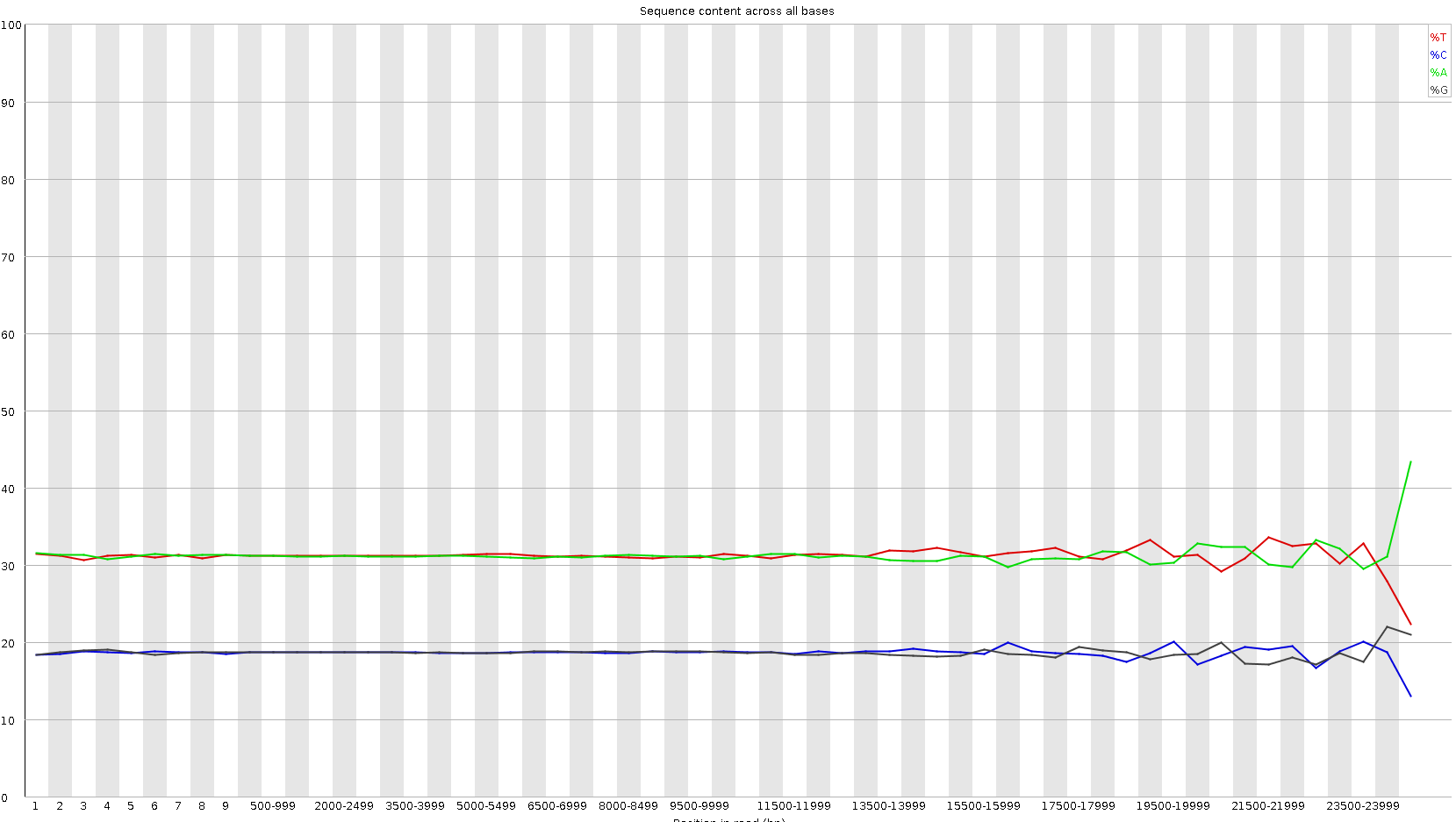


Figure Sequence content across all bases

As for the short reads, the per sequence GC content has a warning. The same goes for the sequence length distribution. The length of the sequences indeed vary, with a majority of sequences (5000) having a length between 1500 and 2500 bp.

Background pattern

Description automatically generated

Figure Sequence length distribution over all sequences for PacBio reads

The adapter and per base N contents, the sequence duplication level are good and there is no overrepresented sequences.