Assignment RNA sequence work flow for Week 5

Paula Carrio Cordo - 24/10/2016

Assignment RNA sequence work flow for Week 5

Option 1. Download the fastq file(s) for one publicly available sample. Briefly describe the sample, run FastQC on the file(s) and comment on the results.

This analysis is based on a fastq file downloaded from the European Nucleotide Archive (ENA). The Study selected is PRJNA28911 (sample accession SAMN00001622). It is part of 1000 Genomes Project Pilot 1 (low coverage sequencing of 180 Hapmap individuals from multiple populations), a study in *Homo Sapiens*.

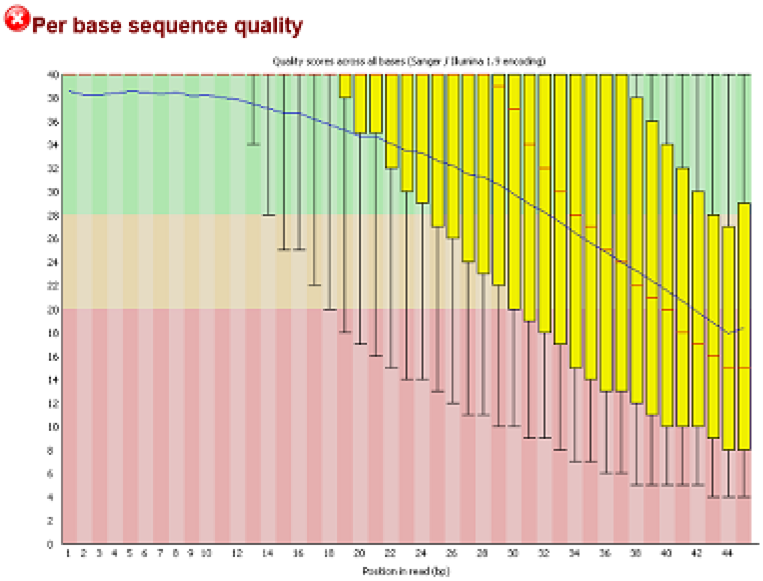
More details were obtained after running a FastQC on the file. FastQC as a quality control tool for high sequencing data allows to check for low-quality data, remaining adapters, contamination,. . . Overall, this analysis shows that the sample has a good quality since the FastQC Report does not show many errors or warnings.

Basic Statistics

The sample ERR000051_1.fastq.gz, has a file type of Conventional base calls. It was encoding by Sanger/Illumina 1.9. The total sequences are 8145207. Sequences flagged as poor quality has a null value. Interestingly the sequence length is 45 which is an acceptable value. The content of GC is 44%.

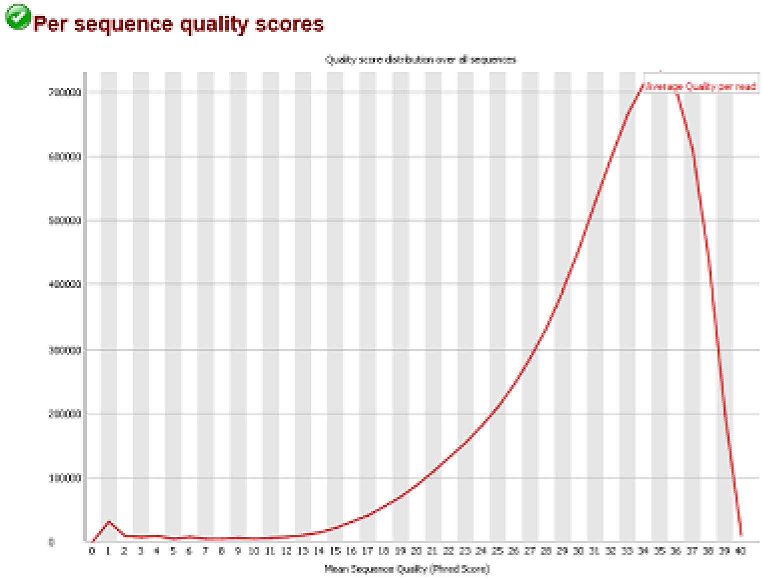
Per Base Sequence Quality

The overview of the range of quality values across all bases at each position in the FastQ file shows a failure. For more than half of the bases the median is less than 20. In addition, for some bases the lower quartile is less than 5.
The quality scores across all bases shows an unusual pattern that may come from a general degradation of quality over the duration of long runs.



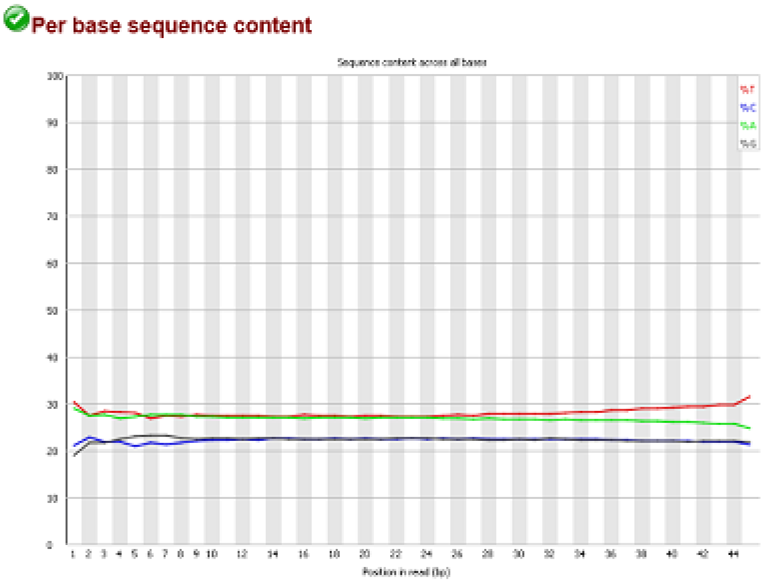
Per Sequence Quality Scores

The graphic shows the quality score distribution over all sequences. In this case there is not a subset of sequences with universally low quality values. The Average Quality per read was 35.



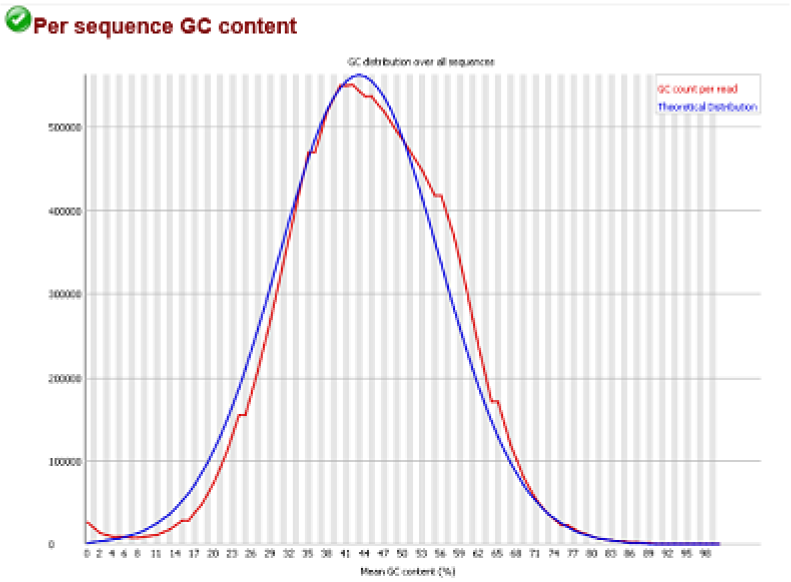
Per Base Sequence Content

There is a normal proportion of each base position for our sample file for which each of the four normal DNA bases has been called. For the bases A and T the values are more or less constant around the 30%, whereas for the bases G and C, the values goes around 20%.



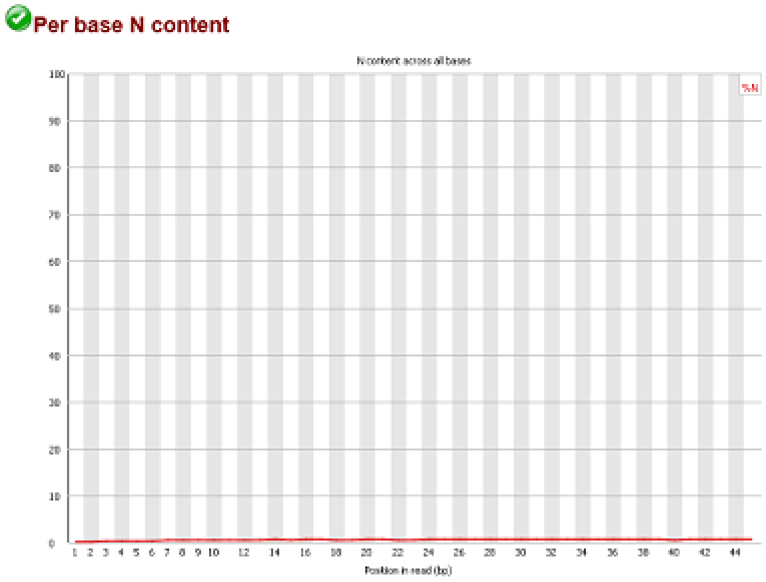
Per sequence GC content

The obtained distribution of GC count per read is similar to the theoretical distribution of GC content, showing a normal distribution. The central peak corresponds to the overall GC content of the genome of the sample.



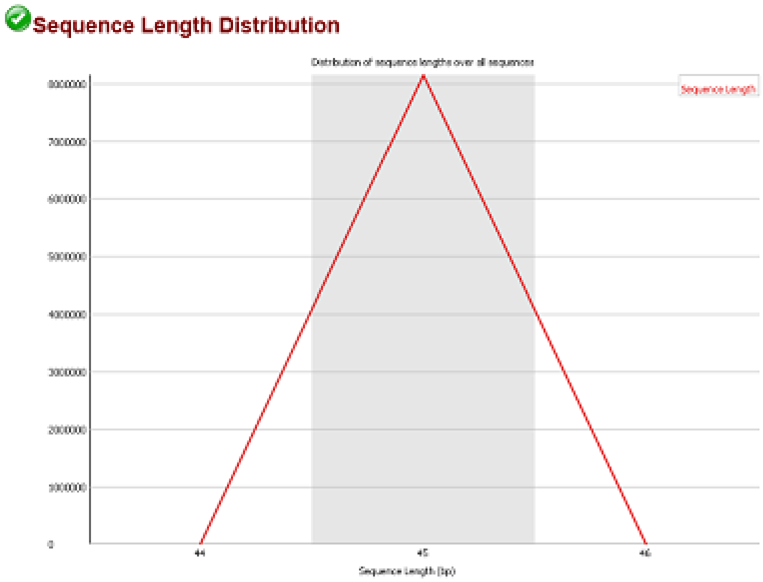
Per Base N Content

A very low proportion of Ns appear for all the positions. This is a good indicative that the sequencer was able to make a base call with sufficient confidence during the whole process of sequenciation.



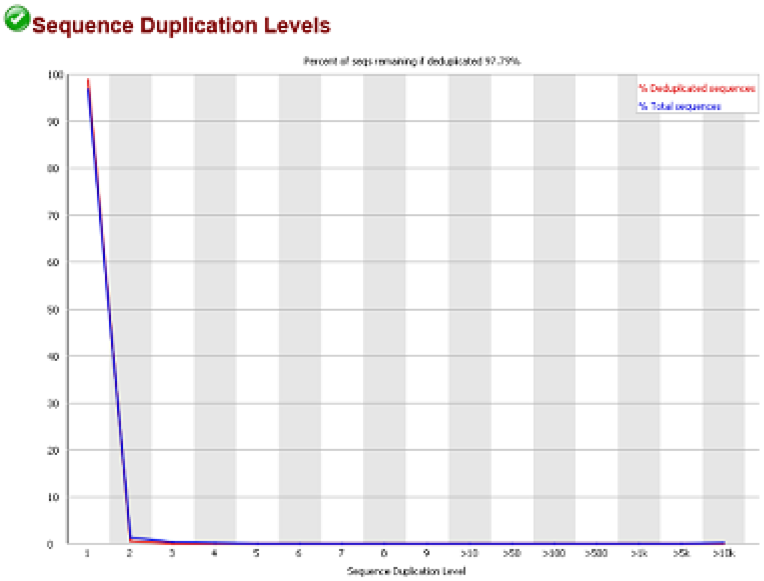
Sequence Length Distribution

This high throughput sequencer genereted sequence fragments of uniform length. The graph shows that most of the fragments had a sequence length of 45 bp.



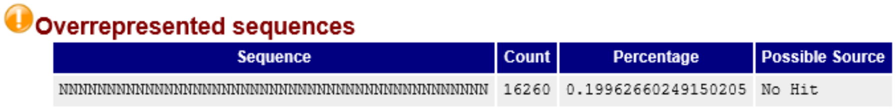
Sequence Duplication Levels

The relative number of sequences with different degrees of duplication has good levels.The sequences fall into the far left of the plot in both the red and blue lines. This indicates that the sample had a properly diverse library.



Overrepresented Sequences

There is a reported warning in this section. The analysis FastQC has found a sequence “NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN…” is representing more than 1% of the total. This is an indicator that the library may be contaminated, or not as diverse as expected.

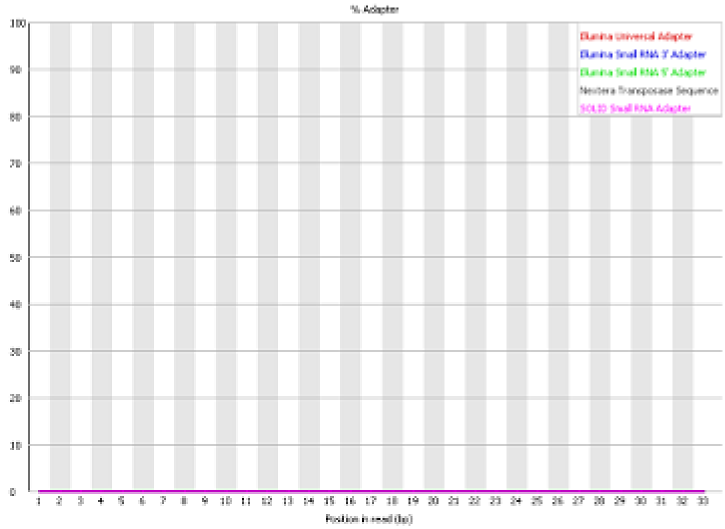

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	16260	0.19962660249150205	No Hit

Adapter Content

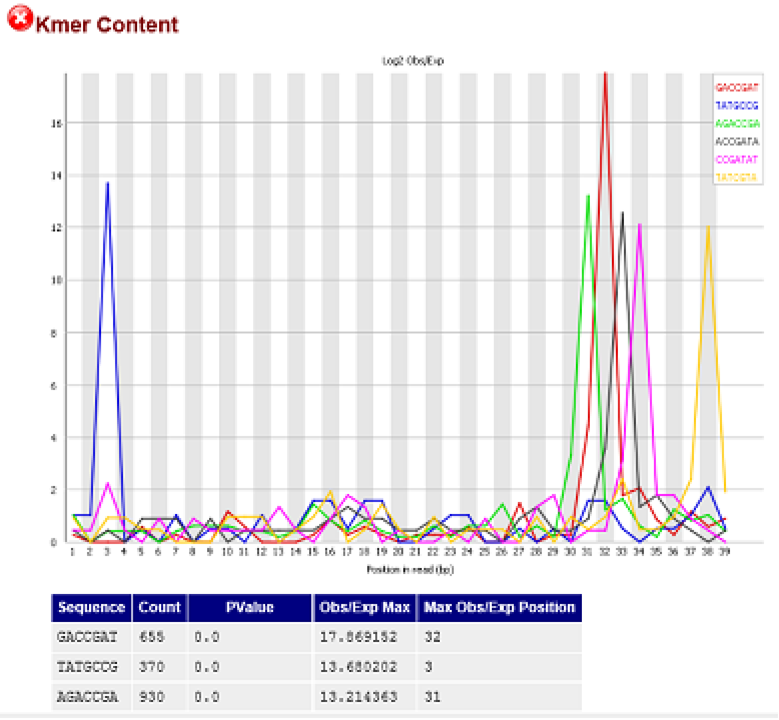
The Kmer Content module after doing a generic analysis of all of the Kmers in the library did not find uneven coverage through the length of the reads. A positive result is reported since there were not any sequence presented in more than 5% of all reads.

✓ Adapter Content



Kmer Content

This analysis of overrepresented sequences shows a failure. This may be indicative of long sequences with poor sequence quality. It is possible that random sequencing errors had dramatically reduce the counts for exactly duplicated sequences.



Kmer Content

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GACCGAT	655	0.0	17.869152	32
TATGCCG	370	0.0	13.680202	3
AGACCGA	930	0.0	13.214363	31