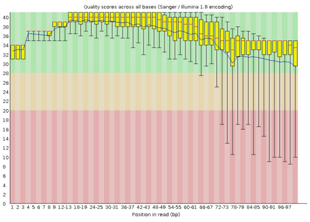


# of demultiplexing and QCs

Laure Lemee, Thomas Cokelaer, Etienne Kornobis

v1 , 17 Aug 2020
https://biomics.pasteur.fr/drylab

Disclaimer

This document is in constant evolution and progress. Please use it with care. Examples and up-to-date test cases are posted on the following wikis: https://github.com/sequana/sequana_demultiplex/wiki and https://github.com/sequana/sequana_fastqc/wiki.

Introduction

Sequencing platforms generate large amounts of sequencing data from short and long read technologies. Short-read sequencers are mostly originating from Illumina company that offers a wide range of sequencers (ISeq100, NextSeq, HiSeq, etc). In a given experiment, samples are indexed (tagged) with specific index to be able to perform multiplexing runs. Once the run is finished, some sequencers will provide the raw data that will later need demultiplexing to be transformed into FastQ files. We are speaking of hundreds of millions of sequences in a single run. Most of the sequencers will perform the demultiplexing automatically but not all of them (e.g., NextSeq). Although experimentalists have access to software to check the quality of the run during the run itself, most analysts will need to perform a QC of the FastQ files and possibly a demultiplexing of the raw data. Although the demultiplexing and QCs of the FastQs are routine tasks for sequencing platforms, it is not for students or scientists new to sequencing analysis or who need to perform these tasks occasionally. In this document, we share our experience on quality checks that are required before sequence analysis and biological conclusions can be drawn.

Sequana pipelines

The demultiplexing and QC of samples are made with tools such as *FastQC* and *bcl2fastq*. In practice, we have 2 dedicated pipelines called *sequana_fastqc* and *sequana_demultiplex*. Those pipelines have online documentation on:

https://github.com/sequana/sequana_fastqc
https://github.com/sequana/sequana_demultiplex

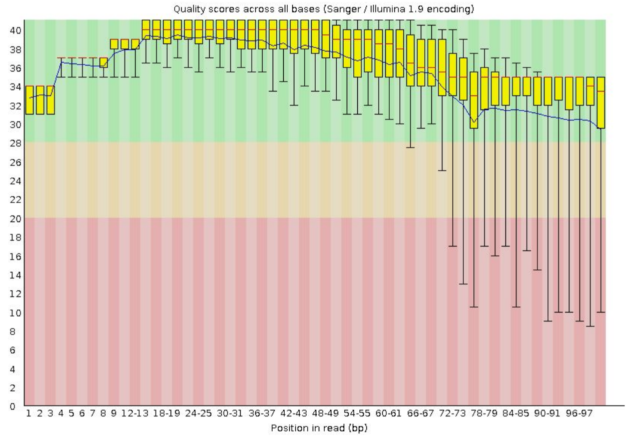
A summary is also is available on https://github.com/sequana/of_demultiplexing_and_fastqc and information available on https://biomics.pasteur.fr/drylab/of_demultiplexing_and_fastqc.html

Interpretation fastqc

A *fastqc* report created by *fastqc* [fastqc] tool can be created for each run. The tool *sequana_fastqc* generalises this task by performing a QC on each sample. At the end, it also creates a summary in the form of a *multiqc* report [multiqc]. There are lots of images created for you. We summarize here below those different types of plots, either for a single sample or for several samples.

Section quality histograms (per base sequence quality)

Each position has a mean quality score based on a Phred scale, ranging from 0 to 40 (although this can go higher with pacbio data). The Phred score is computed as $10^{-(score/10)}$. So, a score of 40 means an error of 1 in 10,000. A score of 10 means 10% of errors. The average quality is represented by the thin blue line. For each position, a boxplot is drawn. The central red line in each box indicates the median value, the yellow box represents the interquartile range (25-75%) and upper and lower whiskers represent the 10 and 90% points. The background color divides the scores into good (green), average (orange) and poor quality scores (red). In this example, the x-axis shows the position from 1 to 100. Note that for clarity, some bases positions are indicated as a range (e.g 12-13 on the x-axis). Provide a bad quality example.



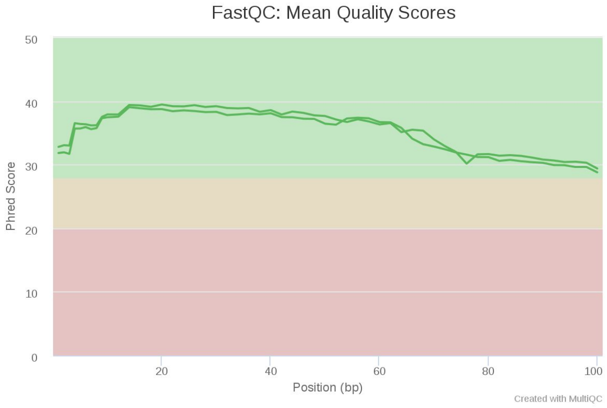
An orange warning (or red failure) will be issued if the lower quartile for any base is less than 10 (5), or if the median for any base is less than 25 (20).

Note that there is a general degradation of quality over the duration of long runs. Typically for reads longer than 100 bases, it is not uncommon to see the mean quality in the orange background. If so, trimming poor quality bases is recommended [cutadapt]. The decrease of the sequence quality at the 3' end of the reads can be observed in most runs.

Another issue (reported in fastqc web page) is the short loss of quality earlier in the run, which then recovers to produce later good quality sequence. This can happen if there is a transient problem with the run (bubbles passing through a flowcell for example). The bubble effect should also be seen in the per-tile quality plot (see later). Note that in such situation, trimming is not advisable as it may remove sequence that have a good overall quality.

In the multi-sample report, we obtain a summary of all samples into a plot called ‘FastQC: Mean quality Scores” as shown in the following figure. The interpretation of the plots background colour is the same as above. The main difference is that for clarity, the boxplot indicating the error bars are not indicated. The great advantage here is that all samples quality can be compared.

Note that with paired-end reads the average quality scores for read 1 will almost always be higher than for read 2.

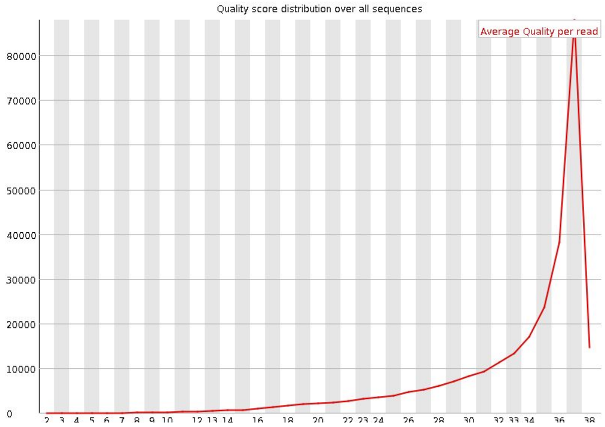


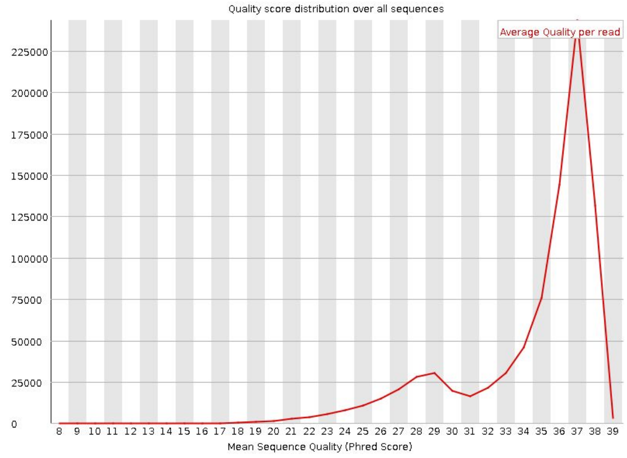
Section per sequence quality score

Another plot related to the quality is the per sequence quality score. The x-axis reports the average phred score along with a read, and the y-axis the number of reads with that score. This plot allows us to check whether a subset has a lower quality than the overall quality. If your average quality is e.g. 35 you should see a narrow peak around that number. Be aware that x-axis is not always linear.

A bimodal distribution may indicate a systematic problem. For example, the end of a run being of poor quality or a broken tile.

A warning is raised if the mean quality is below 27 (0.2% error rate) and an error is below 20 (1% error rate).

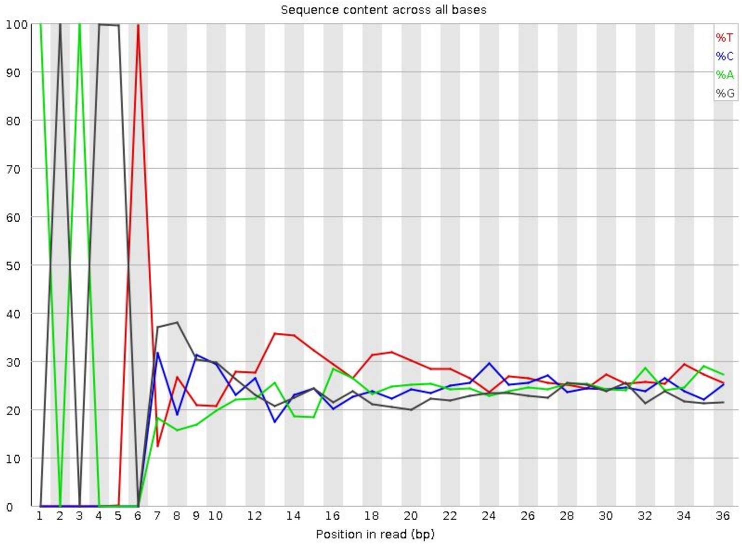




Section per base sequence content

The per base sequence content plots the percentage of each of the four normal DNA bases on the y-axis at each base position (x-axis). The 4 lines should run parallel with each other in a random library. This plot should reflect the content of the sequenced genome. The presence of biases (changes in different bases) usually indicates the presence of an overrepresented sequence. If consistent along the x-axis the library was probably already biased [fastqc]. A warning is raised if difference between ACGT >10% in any position and an error is raised if difference are greater than 20%.

For whole-genome shotgun DNA sequencing the 4 bases should remain constant with a similar percentage of A=T and G=C. For RNA-seq, when using TruSeq RNA library preparation, one should see a non-uniform distribution of bases for the first 10-15 nucleotides. This is normal even though warning or error may be reported by fastqc. Indeed it is not uncommon to have either AT-rich nucleotides introduced by random hexamer priming used in cDNA synthesis

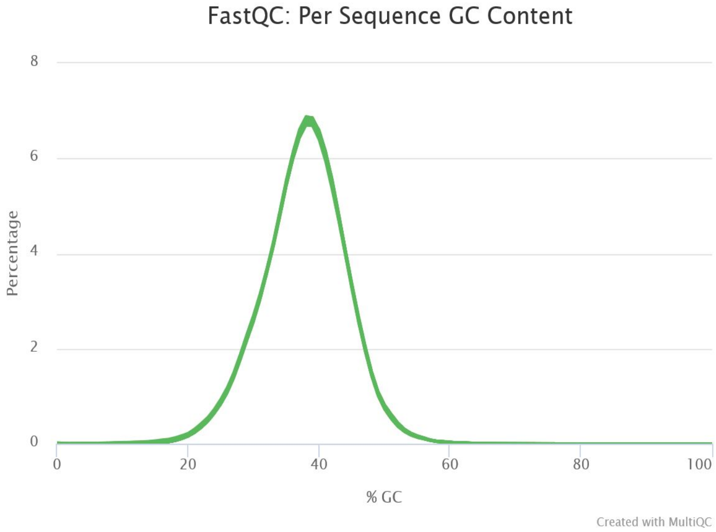


Random hexamer priming results in a bias in the nucleotide composition at the start of sequencing reads [3], GC bias at the beginning of sequences is found when using SMARTer Low/pico input RNA Kit for library preparation (typically for degraded or low quality samples).

At the start, one may have those 7-10 bases bias coming from other kit such as Nextera. You may trim these bases for subsequent analysis.

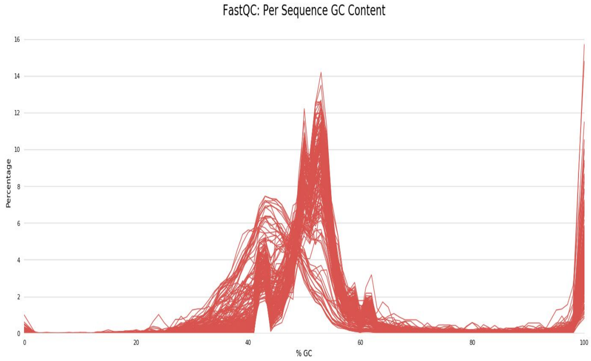
Section Per sequence GC content

The per-sequence GC plot shows the distribution of the GC content found in each sequence. The distribution is compared to a normal distribution that best fit the data (mean GC is taken since the underlying GC content is unknown from the tool). In practice, most organisms have a GC content



between 20 and 80% except for a few special organisms (e.g., plasmodium) and we usually see a distribution centered around 50% in such case. If a specific organism is sequenced, the GC curve should follow a gaussian distribution centered around the mean GC content of the organism.

A warning (failure) is raised if the sum of the deviations from the normal distribution represents more than 15% of the read (30%).

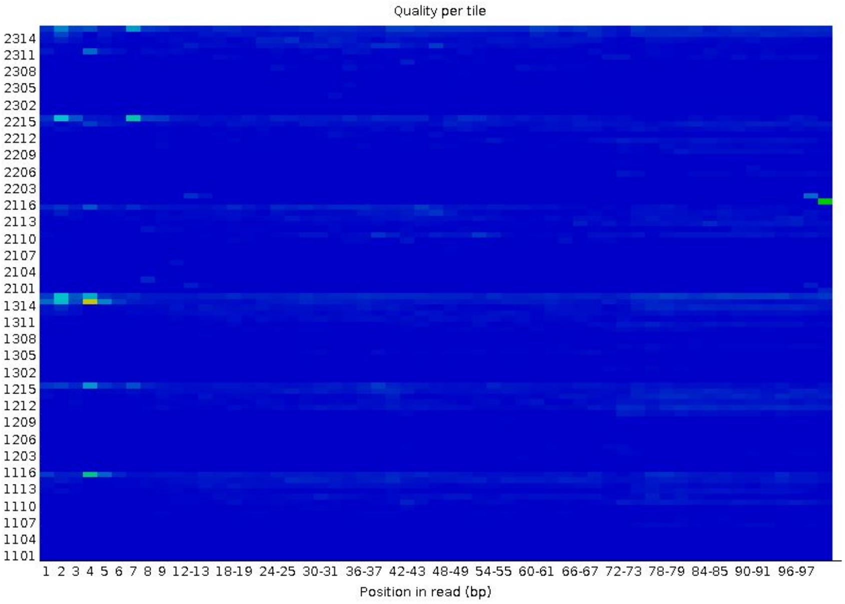


Such failure may indicate a contaminated library. A normal distribution which is shifted indicates some systematic bias which is independent of base position. This will not be flagged as an error since the genome's GC content is unknown.

Yet, it may happen that for a given organism, that additional peaks are found indicating the presence of ribosomal content.

Section per tile sequence quality

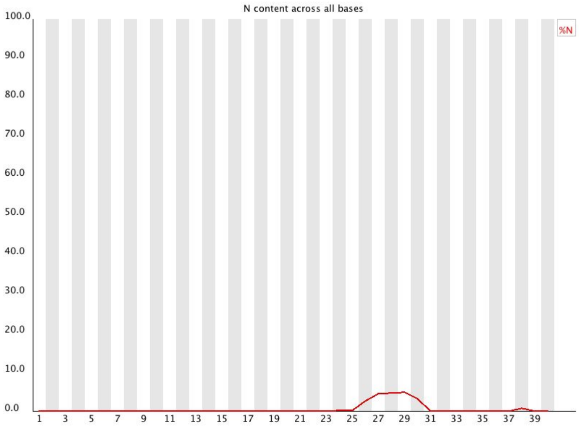
coming soon



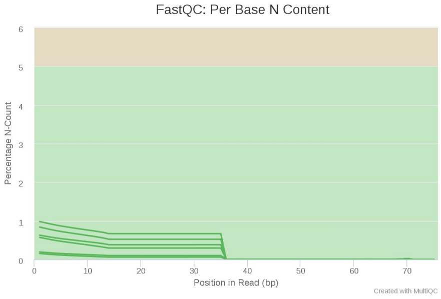
Section N base

If the sequencer is unable to make a based call, the base is replaced with a N rather than a conventional base. This plot shows the percentage of N calls at each position at each position.

According to fastqc manual, if a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a base call. With 2-color sequencers it is normally G's. Yet, it is not uncommon to get N's. If N content >5% fastqc reports a warning, if above 20%, an error. The following image from fastqc manual shows small bump of N's at a specific location.



In the multi-sample figure in the right hand side, we have stretches of N's present in proportion of 1%. This is a symptom of lots of dimers of adapters present in the library. A supplementary cleaning of the data should be done; although this is not always possible due to the library quality. Those reads are actually 35 bases long and should be removed [cutadapt].



Section over-represented sequences

In general, most sequences are unique. If a sequence is over-represented, it may be for a biological reason or indicates a contaminated library coming from adapters, primers ribosomal RNA, etc.

This section gives a table with over-represented sequences. Only sequences that make up more than 0.1% of the total are shown. Again only the first 200,000 sequences are considered.

Each over-represented sequence is then matched to a database of common contaminants. Hits must have 20bp with no more than 1 mismatch. any reads over 75bp is truncated to 50bp.

The over-represented sequences should represent less than 5% of the reads

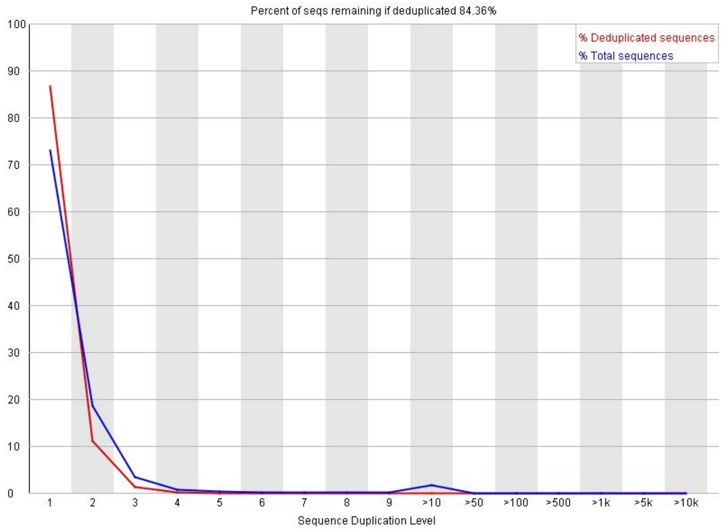
Section duplication levels

In a library made of various genomes, most sequences will occur only once. Therefore A low level of duplication is expected. A high level of duplication may indicate an enrichment bias (e.g. PCR amplification due to a poor complexity of the DNA library). The duplication level plot shows the percentage of reads that have no duplication (level 1), the percentage of reads that appear twice (level 2) and so on and so forth.

According to fastqc manual [fastqc] only the first 200,000 sequences are used to build the duplication levels even though all sequences are then read to estimate each level (not clear from fastqc doc). Any sequence with more than 10 duplicates is placed into the 10 to 50 duplicates category, then 50-100, 100-500, etc. not clear if above 10, above 50, above 500

At the top we have the percent of sequences remaining if deduplicated.

From the fastqc manual: “Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to underrepresent highly duplicated sequences.”

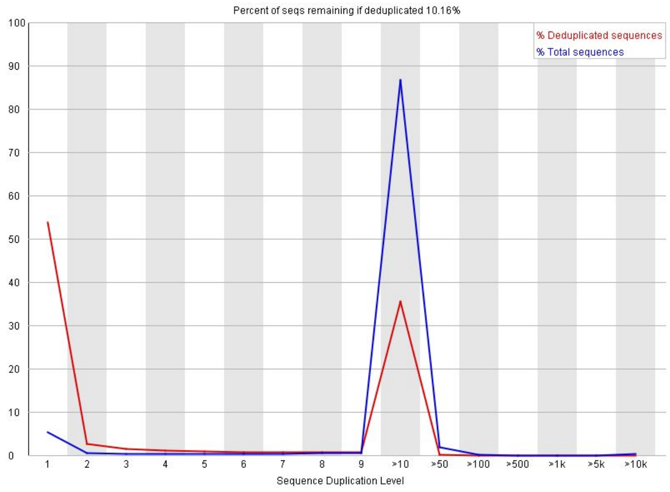


Warning (failure) if non-unique sequences make up more than 20% (50%) of the total.

in RNA-seq experiment, it is not possible to distinguish highly-expressed genes from possible PCR biases

One proxy for library complexity is to look at the sequence duplication levels on the FastQC report. here below are two examples with high and low complexity

Note that with mammalian genomes, the level of duplication is not expected to be large if the sequence library is smaller than the genome size (e.g. 50Mreads) while for bacterial genome size of typically 3-4Mb multiple matches are expected along the whole genome.



Section sequence length distribution

Most sequencers produce reads of equal length. Depending on the demultiplexing, reads may be trimmed or not for adapters. Long reads technologies will have different lengths. This plot shows the distribution of the fragment sizes. If sequences are not of the same length a warning is raised. If any sequence has zero-length a failure is raised.

Section over-represented kmers

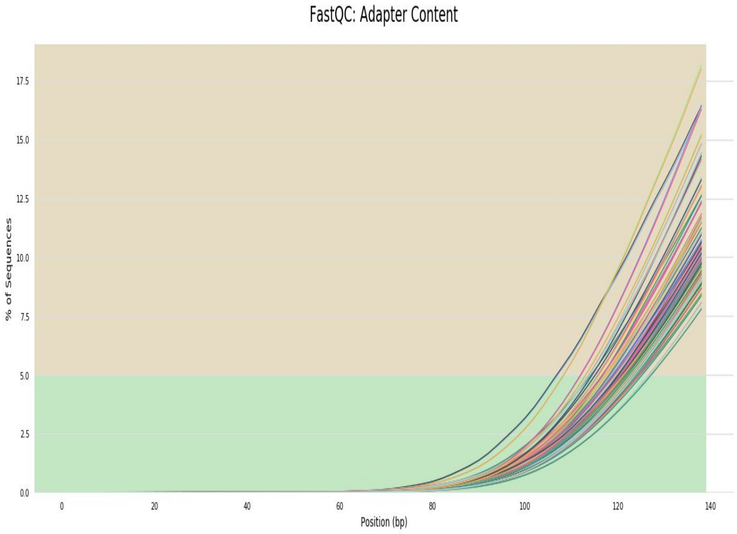
fastqc doc: “This module counts the enrichment of every 5-mer within the sequence library. It calculates an expected level at which this k-mer should have been seen based on the base content of the library as a whole and then uses the actual count to calculate an observed/expected ratio for that k-mer. In addition to reporting a list of hits it will draw a graph for the top 6 hits to show the pattern of enrichment of that Kmer across the length of your reads. This will show if you have a general enrichment, or if there is a pattern of bias at different points over your read length.”

Section adapters

https://support.illumina.com/bulletins/2016/04/adapter-trimming-why-are-adapter-sequences-trimmed-from-only-the--ends-of-reads.html

adapters are only on the 3’ side. Here is a example of presence of adapters at the end of the sequences in a multi-sample plot

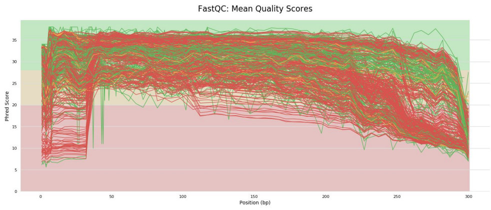
full list of adapters is available here:



https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-13.p df

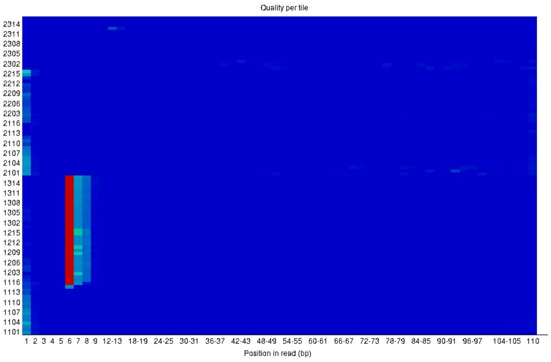
Conclusions

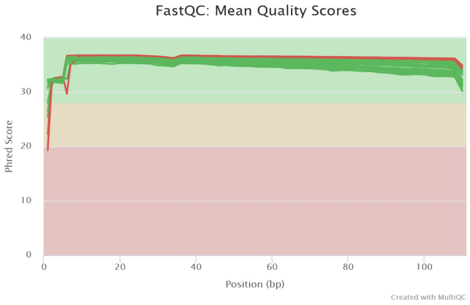
This document is not complete but should already give you a good idea of what fastqc and multiqc plots can goffer you. An important point is that the green/orange/red colors used to indicate that a QC is good/average/bad should be taken with care. To give just one example, the main plot showing the overall quality is essential in DNA and RNA analysis but is known to be red for metagenomics, which is fine. The following plot is a typical example



Test cases

An interesting example is the run HiSeq/190405_D00395_0549_BCDN37ANXX(B1270) for which quality exhibits low quality at the beginning. This show the sample as red but only the first few bases are bad. The rest of the 300 bases are usable. The per-tile image shows an red warning indeed.





References

[1] Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016)

[2] https://biof-edu.colorado.edu/videos/dowell-short-read-class/day-4/fastqc-manual

[3] Hansen, Kasper D., Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming. *Nucleic Acids Research* 38.12 (2010)

[4] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal.* (2011)

https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/