Learning the Sequences Quality Control of Bioinformatics Analysis Method

Henghua Shi[1, a*], Xin Xu[2,b]

[1]School of Computer and Information Engineering, Beijing University of Agriculture, China

[2]Communication Technology Bureau, Xinhua News Agency, China

[a]henghuashi@163.com, [b]youges@163.com

Keywords: Bioinformatics; Quality Control; Sequences; Analysis Method; FastQC

Abstract. The quality control of the original segment obtained by sequencing has a direct impact on the subsequent data analysis. With the application of next-generation sequencing technology, bioinformatics analysis method for sequences have developed rapidly. The sequence quality control has become an important part of bioinformatics analysis method to learning and teaching. With FastQC as a tool for sequences quality control, we do a learning quality controls analysis experiment, and analyze all quality controls analysis steps including Per sequence quality scores, Per sequence GC content, Sequence Length Distribution, Sequence Duplication Levels, et al. Through learning the sequences quality control, we can study bioinformatics analysis method with the same way and can learn other bioinformatics analysis method easier.

1. Introduction

Learning the sequences is the main point of bioinformatics analysis. Biological sequences are including DNA-seq, RNA-seq, Protein sequences, and etc al. With the application of next-generation sequencing (NGS) technology, bioinformatics analysis method for sequences have developed rapidly. Then, the quality control of the original segment obtained by sequencing has a direct impact on the subsequent data analysis. The sequences quality control of bioinformatics analysis method has become the main step of the biological sequences learning.

For RNA-Seq is one of the applications of NGS, the NGS technology has become an important research on the transcriptomics [1]. In this paper, we study FastQC as a tool for sequences quality control, and does an RNA-seq quality controls analysis experiment. The experiment results show Per sequence quality scores, Per sequence GC content, Sequence length distribution, Sequence duplication levels, et al.

2. Sequences Quality Control

Most sequencers will generate a quality control report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself. There are many sequences quality controls tools such as FastQC[2], FastX[3], Sickle[4], and RNA-SeQC[5].

FastQC can visually view the quality of the segment. FastX can remove the low quality of the call or read segment. For double-ended sequences, Sickle can simultaneously remove the corresponding reverse read segment while filtering out the forward - segment of a lot of low-quality base, and vice versa. RNA-SeQC[32] calculation of RNA-seq data quality indicators used to guide experimental design, quality control and optimization analysis, such as sequences depth (depth of coverage), the alignment area (intron, exon, gene region), rDNA content and so on. RNA-SeQC can also be the length of the sequences alignment of the results of statistical analysis, to get a number of quality control indicators.

In this paper, we study FastQC as the example tool for sequences quality control, and does an RNA-seq quality controls analysis experiment to learning the sequences quality control of bioinformatics analysis method.

3. FastQC

FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material. The analysis in FastQC is performed by a series of analysis modules. The left hand side of the main interactive display or the top of the HTML report show a summary of the modules which were run, and a quick evaluation of whether the results of the module seem entirely normal (green tick), slightly abnormal (orange triangle) or very unusual (red cross).

It is important to stress that although the analysis results appear to give a pass/fail result, these evaluations must be taken in the context of what you expect from your library. A 'normal' sample as far as FastQC is concerned is random and diverse. Some experiments may be expected to produce libraries which are biased in particular ways. You should treat the summary evaluations therefore as pointers to where you should concentrate your attention and understand why your library may not look random and diverse.

4. Experiment Results

We do an RNA-seq quality controls analysis experiment with FastQC tool, and the experiment results show Per sequence quality scores, Per sequence GC content, Sequence length distribution, Sequence duplication levels as following.

4.1 Per sequence quality scores.

Quality scores distribution over all sequences shows as Fig. 1. The x-axis on the graph shows the quality score and the y-axis on the graph shows the reads numbers.

The per sequence quality score [7] report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences. If a significant proportion of the sequences in a run have overall low quality then this could indicate some kind of systematic problem - possibly with just part of the run (for example one end of a flow cell).

A warning is raised if the most frequently observed mean quality is below 27 (this equates to a 0.2% error rate). An error is raised if the most frequently observed mean quality is below 20 (this equates to a 1% error rate). In Fig. 1., the most frequently observed mean quality is over 27. The experiment result of this quality controls analysis step is entirely normal.

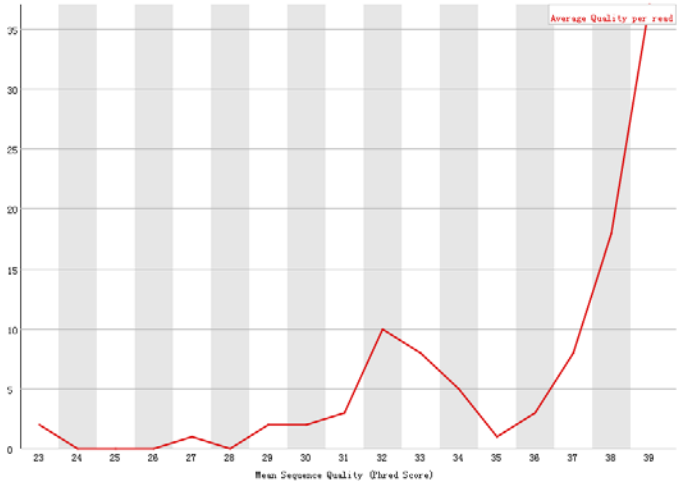

Fig. 1. Quality scores distributiong over all sequences

4.2 Per sequence GC content.

GC distribution over all sequences shows as Fig. 2. The x-axis on the graph shows the mean percentage of GC of per sequence, and the y-axis on the graph shows the reads numbers. It is means the summation mean percentage of the percentage of G and C of per sequence.

This module measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content. In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution.

An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what your genome's GC content should be.

A warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads. This module will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads. The red line represents GC count per read, and the blue line represents the theoretical distribution in Fig. 2., and the sum of the deviations from the normal distribution represents more than 30% of the reads. The experiment result of this quality controls analysis step is very unusual.

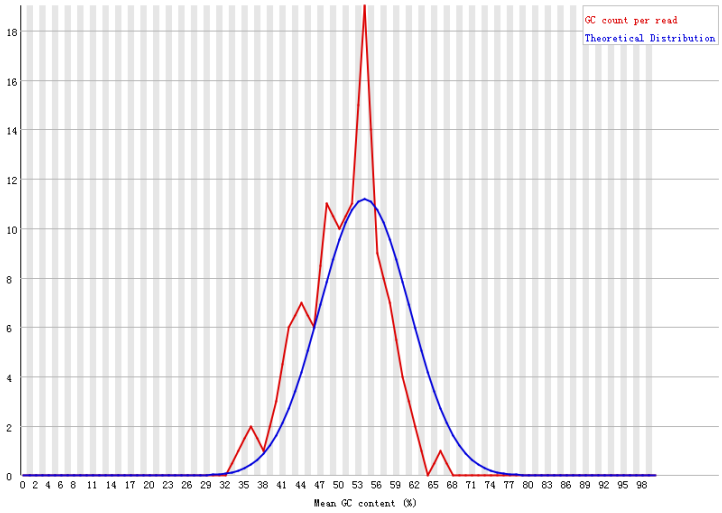

Fig. 2. GC distribution over all sequences

4.3 Sequence length distribution.

Distribution of sequence lengths over all sequences shows as Fig. 3. The x-axis on the graph shows the sequence lengths (bp) and the y-axis on the graph shows the mean percentage of all sequences.

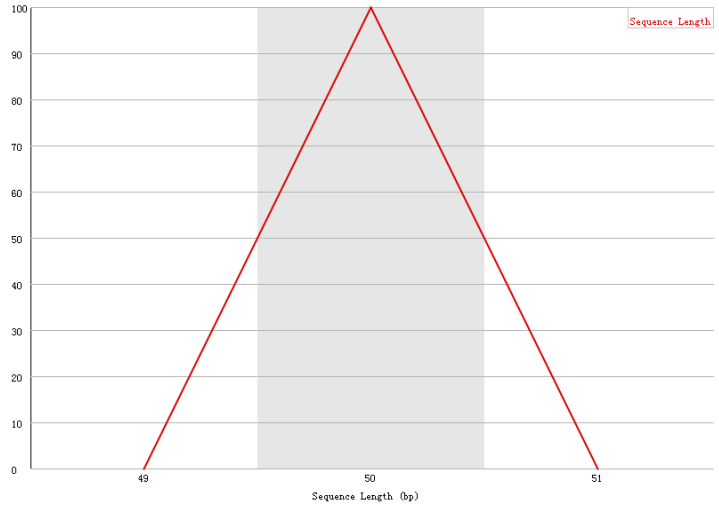

Fig. 3. Distribution of sequence lengths over all sequences

Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end. This module generates a graph showing the

distribution of fragment sizes in the file which was analyzed. In many cases this will produce a simple graph showing a peak only at one size, but for variable length FastQ files this will show the relative amounts of each different size of sequence fragment.

This module will raise a warning if all sequences are not the same length. This module will raise an error if any of the sequences have zero length. The length of all sequences is 50 in Fig. 3. The experiment result of this quality controls analysis step is entirely normal.

4.4 Sequence Duplication Levels.

Sequence duplication levels over all sequences shows as Fig. 4. The x-axis on the graph shows the sequence duplication levels and the y-axis on the graph shows the mean percentage of all sequences.

In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification). This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication.

This module will issue a warning if non-unique sequences make up more than 20% of the total. This module will issue an error if non-unique sequences make up more than 50% of the total. The sequence duplication levels of all sequences are not more than level 2 in Fig. 4. The experiment result of this quality controls analysis step is entirely normal.

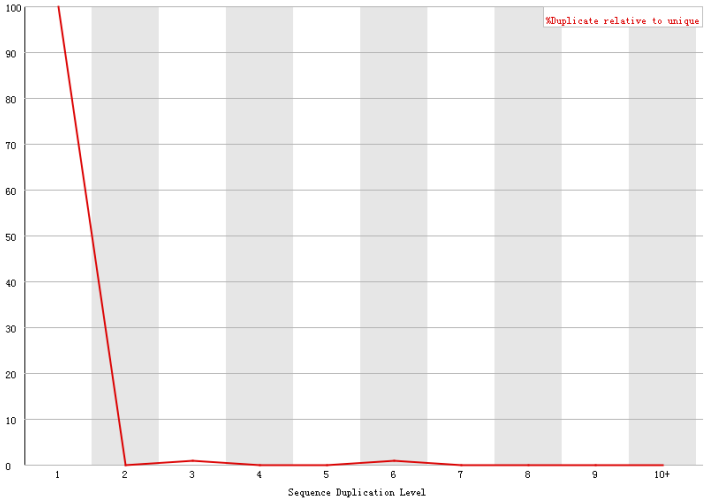

Fig. 4. Sequence duplication levels over all sequences

5. Conclusion

Learning the sequences is the main point of bioinformatics analysis. The sequences quality control of bioinformatics analysis method has become the main step of the biological sequences learning. With FastQC as a tool for sequences quality control, we do a learning quality controls analysis experiment, and analyze all quality controls analysis steps including Per sequence quality scores, Per sequence GC content, Sequence Length Distribution, Sequence Duplication Levels, et al. All the experiment results are entirely normal, but Per sequence GC content is very unusual.

Through learning the sequences quality control in this paper, we can study bioinformatics analysis with the same way and can learn other bioinformatics analysis method easier.

6. Acknowledgement

Corresponding author is Henghua Shi. The authors would like to acknowledge the supports provided by 2016 General Scientific Research Project of Beijing Municipal Education Commission (PXM2016_014207_000008).

7. References

[1]. Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009.10(1): p. 57-63.

[2]. Simon A., Felix K., Anne S. P., et al, http://www.bioinformatics.babraham.ac.uk/projects/festqc/

[3]. http://hannonlab.cshl.edu/fastx_toolkit/

[4]. https://github.com/ucdavis-bioinformatics/sickle

[5]. DeLuca, D.S., et al., RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics, 2012.28(11): p. 1530-1532.

[6]. Ewing B, Green P, Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8 (3): 186–194. doi:10.1101/gr.8.3.186. PMID 9521922.

[7]. Dear S, Staden R, A standard file format for data from DNA sequencing instruments. DNA Seq. 3 (2): 107–110. doi:10.3109/10425179209034003. PMID 1457811.