Learning the Base Sequence Quality and Content of Bioinformatics Analysis Method

Henghua Shi[1, a], Weiyu Li[2, b] and Xin Xu[3, c]

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

[2]College of Plant Science and Technology, Beijing University of Agriculture, China

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

[a]henghuashi@163.com, [b]youges@163.com, [c]lwy@bac.edu.cn

Keywords: Bioinformatics; Base sequence; Quality; Content; Next-generation sequencing

Abstract. With the application of next-generation sequencing technology, bioinformatics analysis method for sequences have developed rapidly. There are many variable quantities and sub-methods to evaluate the quality of the base sequence. With comparing and analysis the tools of learning quality and content of the base sequence, we do a learning quality and content analysis experiment, and analyze the base sequence quality and content analysis steps including Per base sequence quality, Per base sequence content, Per base GC content, Per base N content, et al.

Introduction

The base sequence quality and content learning is one of the most important of bioinformatics analysis. The base sequence is including DNA-seq, RNA-seq, and is percentage of the four normal DNA bases such as A, G, C, T. DNA has four nitrogenous bases: (A) adenine, (T) thymine, (C) cytosine, and (G) guanine. RNA contains three of these bases - (A), (C), and (G) but not (T). Uracil (U) is found in its place and complements adenine (A) instead in transcription. Transcription is the system that produces a complementary RNA sequence from a strand of DNA [1] [2]. Transcription is initiated when RNA polymerase connects to the DNA template strand in an area called the promoter region and starts stringing together a new complementary RNA strand in the 3' to 5' direction. The resulting new strand of mRNA has complementary base pairs to the original DNA template. Therefore if the original DNA template strand read ACGT, the RNA strand will attach uracil to adenine so the complementary RNA strand will read UGCA.

For the quality evaluation of the base sequence, we do a learning quality and content analysis experiment. The experiment results show the base sequence quality and content with Per base sequence quality, Per base sequence content, Per base GC content, Per base N content, et al.

The Quality Evaluation of the Base Sequence

There are many sequences quality control tools such as FastQC [2], FastX [3], Sickle [4], and RNA-SeQC [5], and some of these tools are including many variable quantities and sub-methods to evaluate the quality of the base sequence. We analysis all these tools and summarize four items to evaluate the quality of the base sequence such as Per base sequence quality, Per base sequence content, Per base GC content, Per base N content. After comparing the FastQC, FastX, Sickle, and RNA-SeQC, we select FastQC as a tool for learning quality and content of the base sequence with Per base sequence quality, Per base sequence content, Per base GC content, Per base N content, et al.

Experiment Results

Per Base Sequence Quality. Quality scores across all bases shows as Fig. 1. The x-axis on the graph shows the position in read, and the y-axis on the graph shows the quality scores. The quality scores is is Fred [6] [7].

In Fig. 1, the background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read. The blue line represents the mean quality, and its quality is all over 28 and is very good quality calls. The experiment result of this quality and content analysis step is entirely normal.

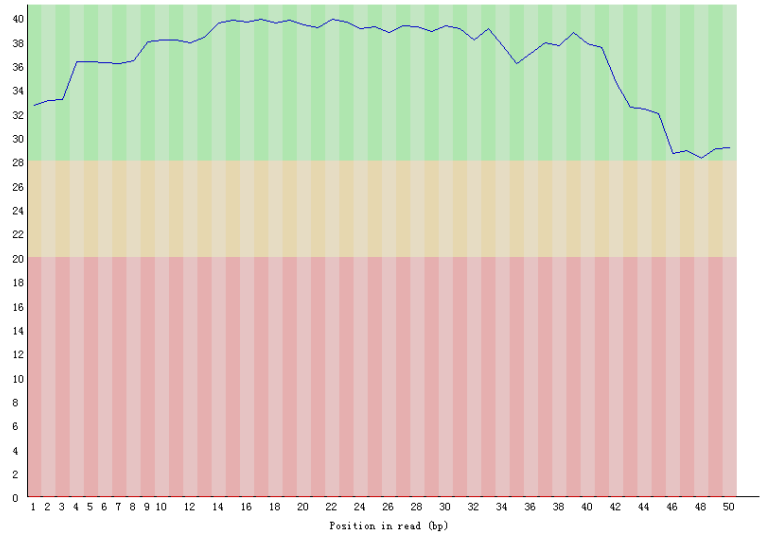

Figure 1. Quality scores across all bases

Per Base Sequence Content. Base sequence content across all bases shows as Fig. 2. The x-axis on the graph shows the position in read, and the y-axis on the graph shows the percentage of the four normal DNA bases such as A, G, C, T [8].

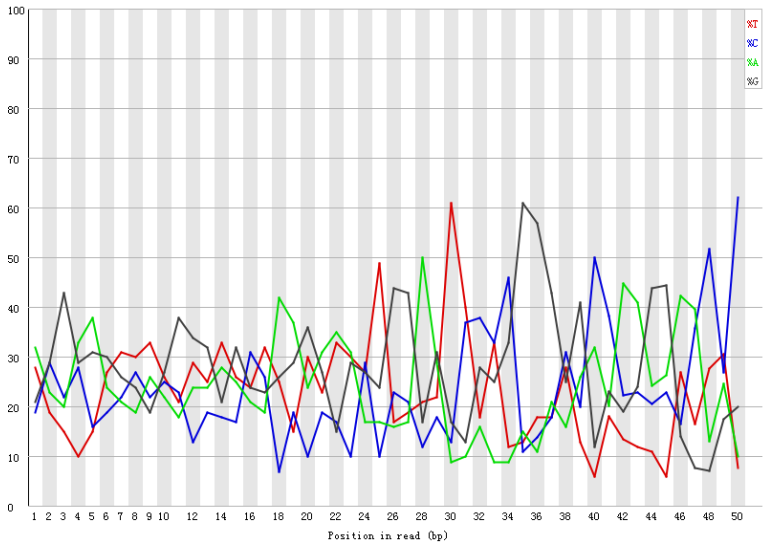

Figure 2. Content across all bases

Per base sequence content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called. In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall amount of these bases in your genome, but in any case they should not be hugely imbalanced from each other.

This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position. This module will fail if the difference between A and T, or G and C is greater than 20% in any position. The red line, the blue line, the green line, and the black line respectively represent the

percentage of T, C, A, G in Fig. 2. The difference between A and T is greater than 50% in 30 position in read, and the difference between G and C is also greater than 50% in 35 position in read. The experiment result of this quality and content analysis step is very unusual.

Per Base GC Content. GC content [9] [10] across all bases shows as Fig. 3. The x-axis on the graph shows the position in read, and the y-axis on the graph shows the percentage of G and C.

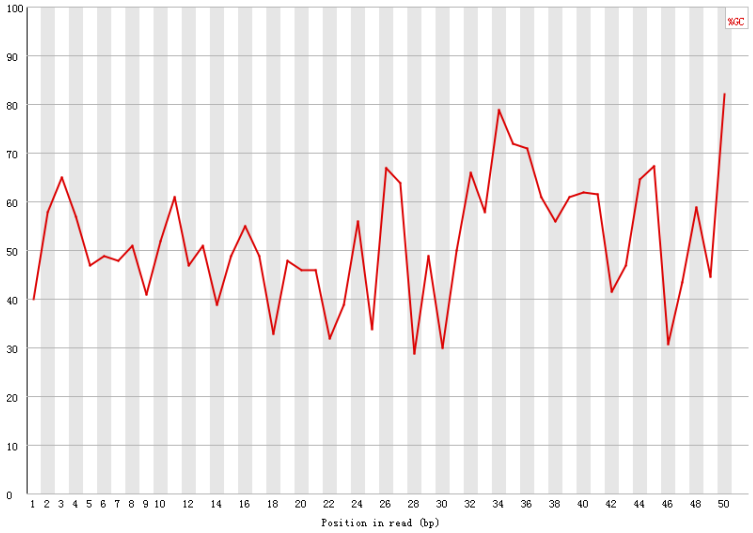

Figure 3. GC content across all bases

In a random library, there would be little to no difference between the different bases of a sequence run, so the line in this plot should run horizontally across the graph. The overall GC content should reflect the GC content of the underlying genome. If a GC bias which changes in different bases then this could indicate an overrepresented sequence which is contaminating library. A bias which is consistent across all bases either indicates that the original library was sequence biased, or that there was a systematic problem during the sequencing of the library.

This module issues a warning it the GC content of any base strays more than 5% from the mean GC content. This module will fail if the GC content of any base strays more than 10% from the mean GC content. The red line represents the percentage of G and C in Fig. 3. There is greater than 50% in all position in read. The experiment result of this quality and content analysis step is very unusual.

Per Base N Content. N content across all bases shows as Fig. 4. The x-axis on the graph shows the position in read, and the y-axis on the graph shows N content.

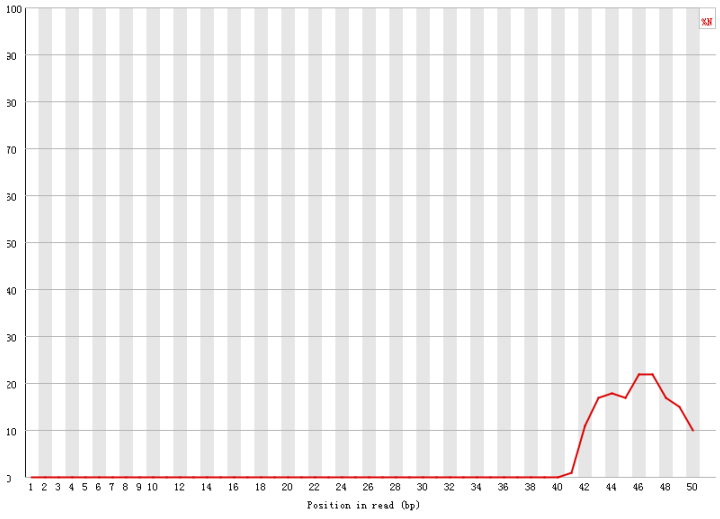

Figure 4. N content across all bases

If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called. It's not unusual to see a very low proportion of Ns appearing in a sequence, especially nearer the end of a sequence. However, if this proportion rises above a few percent it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls.

This module raises a warning if any position shows an N content more than 5%. This module will raise an error if any position shows an N content more than 20%. The red line represents N content in Fig. 5, and N content is more than 20% in 46 and 47 position in read. The experiment result of this quality controls analysis step is very unusual.

Conclusion

There are many variable quantities and sub-methods to evaluate the quality of the base sequence. For the quality evaluation of the base sequence, we study FastQC as a selected tool for learning quality and content of the base sequence, and do a learning quality and content analysis experiment. The experiment results show the base sequence quality and content with Per base sequence quality, Per base sequence content, Per base GC content, Per base N content, et al. Some of the experiment result of this quality controls analysis step is entirely normal or very unusual.

With this paper for learning the base sequence quality and content, we can study and analysis base sequence better, and learning other bioinformatics analysis method more easy.

Acknowledgement

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

References

[1] D. L. Nelson, M C. Michael: *Lehninger Principles of Biochemistry, ed. 5*, W.H. Freeman and Company 2008.

[2] F. A. Carey: *Organic Chemistry, ed. 6*, Mc Graw Hill 2008.

[3] Information on http://www.bioinformatics.babraham.ac.uk/projects/festqc/

[4] Information on http://hannonlab.cshl.edu/fastx_toolkit/

[5] Information on https://github.com/ucdavis-bioinformatics/sickle

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

[7] B. Ewing, P. Green: Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8(3): p.186-194.

[8] L. D. Hurst, A. R. Merchant: High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. Biol. Sci. 268 (1466): p. 493-497.

[9] M. T. Madigan, J. M. Martinko: *Brock biology of microorganisms (10th ed.)*. Pearson-Prentice Hall. ISBN 84-205-3679-2.

[10] N. Galtier, J. R. Lobry: Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in Prokaryotes. Journal of Molecular Evolution. 44 (6): p. 632-636.

