

Learning the Computing Quality Statistics Method of the Sequence Reads

Henghua Shi^{1, a*} and Xin Xu^{2, b}

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

²Communication Technology Bureau, Xinhua News Agency, China

^ahenghuashi@163.com, ^byouges@163.com

Keywords: Bioinformatics; Quality statistics; Sequence reads; Quality score; FastQC

Abstract. There are many bioinformatics analysis method to compute the quality statistics of sequence reads. With the application of next-generation sequencing technology, bioinformatics analysis method for sequences have developed rapidly. The sequence quality statistics has become an important part of bioinformatics analysis method to learning and teaching. For learning the computing quality statistics method of the sequence reads, we do a compute the sequence reads quality statistics experiment. We show and analysis the experiment results for per base sequence quality and per sequence GC content with FastQC software and for nucleotide distribution with draw nucleotides distribution chart software.

Introduction

Learning the sequence quality statistics is the main point of bioinformatics analysis. For the sequence quality analysis, the most important is the study per base sequence quality, per sequence GC content and nucleotide distribution. 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]. Per base sequence quality shows the mean quality score of each sequence reads. Per sequence GC content shows the GC content of each sequence reads. Nucleotide distribution shows the distribution of four nitrogenous bases.

In this paper, we introduce compute quality statistics software [3] to compute the quality statistics of experiment sequence reads. Then, we study FastQC as a tool for per base sequence quality and per sequence GC content, and introduce draw nucleotides distribution chart software for nucleotide distribution.

Compute Quality Statistics

Compute quality statistics is the software to compute the quality statistics of the sequence reads, and is integrated into the Galaxy scientific workflow[4][5][6]. Some computing results contain the following fields.

- column = column number (position on the read)
- mean = Mean quality score value for this column.
- A_Count = Count of 'A' nucleotides found in this column.
- C_Count = Count of 'C' nucleotides found in this column.
- G_Count = Count of 'G' nucleotides found in this column.
- T_Count = Count of 'T' nucleotides found in this column.
- N_Count = Count of 'N' nucleotides found in this column.

The computing the quality statistics of the sequence reads is as in Table 1.

Table 1 The mean quality score and A, T, C, G count

column	mean	A_Count	C_Count	G_Count	T_Count	N_Count
1	28.44	36	18	14	18	14
2	30.69	28	18	19	35	0
3	32.35	36	19	21	24	0
4	36.21	23	17	41	19	0
5	36.46	36	19	25	20	0
6	36.36	38	16	24	22	0
7	36.17	32	20	30	18	0
8	36.47	20	15	27	38	0
9	38.37	29	28	19	24	0
10	38.07	26	23	34	17	0
11	38.31	24	21	25	30	0
12	38.07	23	27	22	28	0
13	37.87	23	21	24	32	0
14	39.49	27	20	32	21	0
15	39.2	32	20	23	25	0
16	39.86	29	21	17	33	0
17	39.82	23	18	24	35	0
18	39.49	23	23	27	27	0
19	39.55	33	20	23	24	0
20	39.6	29	18	27	26	0
21	39.66	26	21	22	31	0
22	39.34	28	21	24	27	0
23	39.12	22	24	28	26	0
24	39.59	17	30	15	38	0
25	39.5	15	8	20	57	0
26	39.25	8	7	73	11	1
27	39.35	16	7	57	20	0
28	38.98	65	12	11	11	1
29	38.89	54	12	12	20	2
30	38.84	2	12	11	75	0
31	37.87	4	27	10	55	4
32	36.29	3	56	16	18	7
33	37.88	6	19	20	55	0
34	37.96	5	59	30	6	0
35	38.19	9	4	77	10	0
36	39.05	13	8	76	3	0
37	38.97	11	14	59	16	0
38	38.41	9	10	23	58	0
39	39.12	16	22	54	8	0
40	38.83	14	69	9	8	0
41	39.29	18	62	16	4	0
42	39.15	72	9	14	5	0
43	38.53	66	7	22	5	0
44	38.38	12	10	72	6	0
45	38.12	19	14	55	12	0
46	37.46	67	11	10	12	0
47	38.21	55	31	6	8	0
48	37.81	10	70	3	17	0
49	38.27	16	22	9	53	0
50	38.51	9	68	15	8	0

Per Base Sequence Quality and Per Sequence GC Content with FastQC

FastQC. There are many sequences quality control tools such as FastQC [7], FastX [8], Sickle [9], and RNA-SeQC [10]. 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. 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.

Per Base Sequence Quality. We show the experiment results as in Table 1 with FastQC, and per base sequence quality of the experiment sequence reads 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.

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 the mean quality of most is all over 28 and is very good quality calls. The experiment result of this quality and content analysis step is entirely normal.

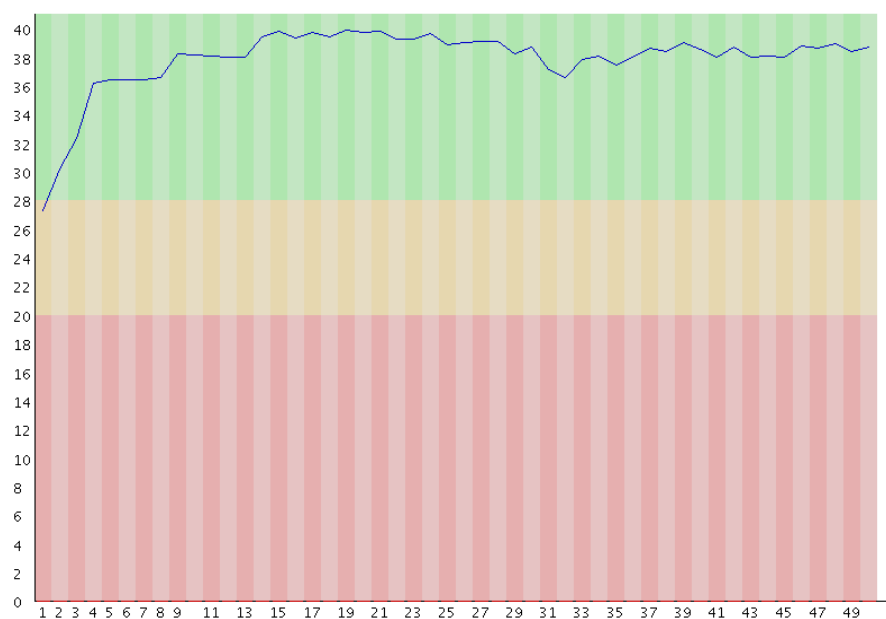


Figure 1. Per base sequence quality

Per Sequence GC Content. We show the experiment results as in Table 1 with FastQC, and per sequence GC content of the experiment sequence reads 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.

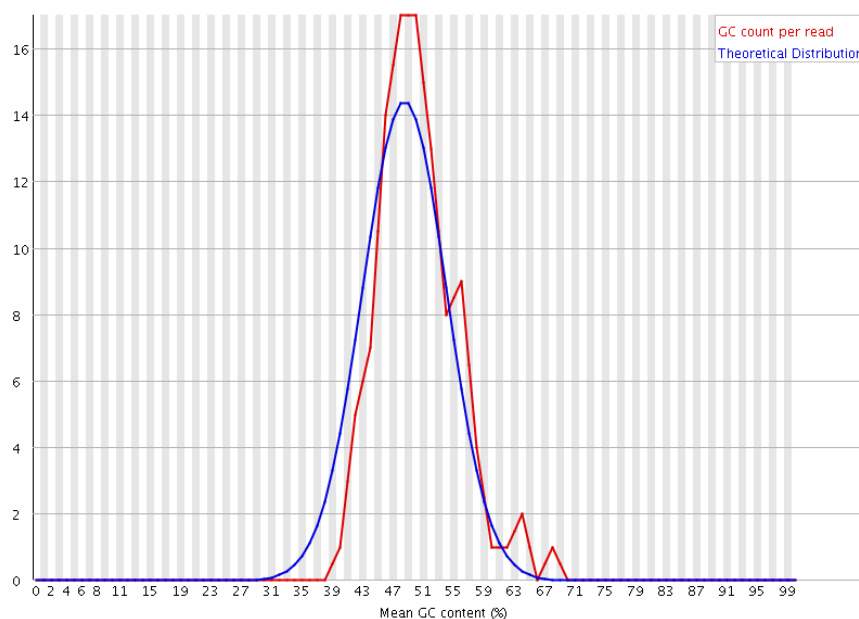


Figure 2. Per sequence GC content

Nucleotide Distribution

For the nucleotide distribution with draw nucleotides distribution chart software, we create the experiment results as in Table 1 with a stacked-histogram graph. Nucleotide distribution chart graph of the experiment sequence reads shows as Fig. 3.

In Fig. 3, The x-axis on the graph shows the position in read, and the y-axis on the graph shows the percent of each nitrogenous bases such as A, G, C, T [11]. 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. In the initial sequence reads, there is many N.

The blue chart is A, the red chart is C, the green chart is G, the orange chart is T, and the pink chart is N in Fig. 3. The nucleotides distribution is from Table 1. For example, A count is 36, C count is 18, G count is 14, T count is 18, and N count is 14 in column 1. Then, the summation count is 100. We can compute the percent of each nitrogenous base. The percent of A is 36%, the percent of C is 18%, the percent of G is 14%, the percent of T is 18%, and the percent of N is 14%. All the above results are corresponding to the length of the charts in Fig. 3.

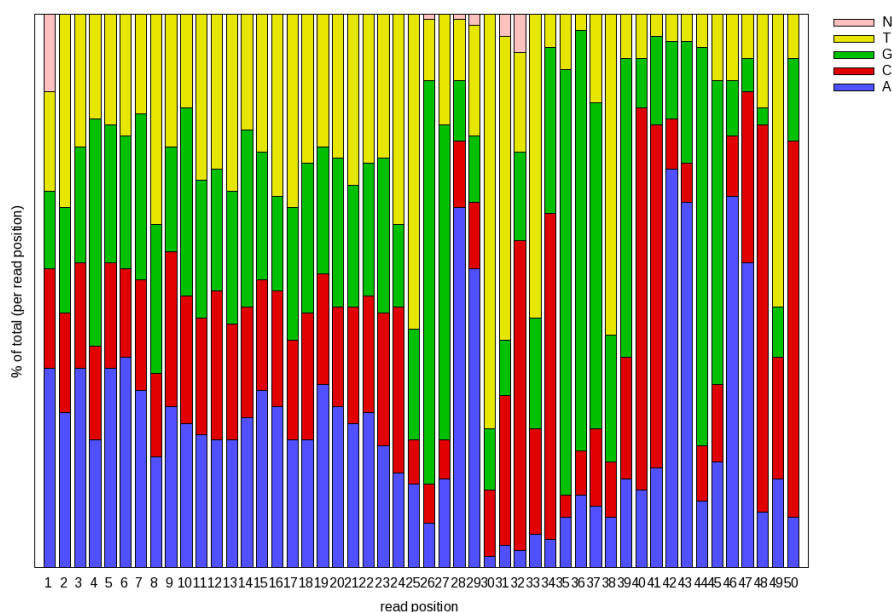


Figure 3. Nucleotide distribution chart graph

Summary

Learning the sequences quality statistics is the main point of bioinformatics analysis. The computing quality statistics method of the sequence reads has become the main step of the biological sequences learning. With FastQC as a tool for sequences quality control and per sequence GC content, we do a learning computing quality statistics analysis experiment, show the experiment results for per base sequence quality and per sequence GC content with FastQC software and for nucleotide distribution with draw nucleotides distribution chart software. With the same way, we can learn and study 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://hannonlab.cshl.edu/fastx_toolkit/
- [4] J. Goecks, A. Nekrutenko, and A. J. Taylor: Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*. 11 (8): p. 86.
- [5] D. Blankenberg, G. V. Kuster N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, J. Taylor: Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. In Frederick M. Ausubel. *Current Protocols in Molecular Biology*.
- [6] J. Taylor, I Schenck, D. Blankenberg, and A. Nekrutenko: Using Galaxy to Perform Large-Scale Interactive Data Analyses". In Andreas D. Baxevanis. *Current Protocols in Bioinformatics*.
- [7] Information on <http://www.bioinformatics.babraham.ac.uk/projects/festqc/>
- [8] Information on http://hannonlab.cshl.edu/fastx_toolkit/
- [9] Information on <https://github.com/ucdavis-bioinformatics/sickle>
- [10] D. S. DeLuca: RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 2012.28(11): p. 1530-1532.
- [11] V. Boeva, A. Zinovyev, K. Bleakley, J. Vert, I. Janoueix-Lerosey, O. Delattre, E. Barillot: Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 2011, 27(2): p. 268.