ATLANTIS PRESS
Advances in Intelligent Systems Research, volume 130
6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016)

Learning the Sequence Reads Information of Bioinformatics Analysis Method

Henghua Shi[1, a*] and 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: Sequence reads; Quality statistics; Nucleotide distribution; Quality score; Bioinformatics

Abstract. The main information of the sequence reads are quality statistics and nucleotide distribution. There are many bioinformatics analysis method to compute quality statistics and create the quality scores and the nucleotide distribution. For learning the sequence reads information of bioinformatics analysis method, we do a sequence reads analysis experiment, and compute quality statistics and create the quality scores and nucleotide distribution graph with the corresponding bioinformatics analysis tools.

Introduction

The quality statistics and nucleotide distribution are the main information of the sequence reads. The quality statistics show the quality scores of each sequence reads such as the min value, the max value, the median value, 1st quartile quality score, 3rd quartile quality score, IQR (Inter-Quartile range), et al. The nucleotide distribution shows the distribution of four nitrogenous bases. 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].

In this paper, we do a sequence reads analysis experiment with two steps. The first step is to compute the quality statistics and nucleotide distribution with compute quality statistics software [3], and we analysis the result. The second step is to create the quality scores and nucleotide distribution graph with draw quality score boxplot software and draw nucleotides distribution chart software [3].

Compute Quality Statistics

Bioinformatics analysis method for sequences has developed rapidly with the application of next-generation sequencing (NGS) technology [4]. For the initial sequence reads, we should analysis the main information of the sequence reads such as quality statistics and nucleotide distribution. Compute quality statistics is the software to compute the quality statistics and nucleotide distribution of the sequence reads, and is integrated into the Galaxy scientific workflow [5][6][7]. The computing results contain the following fields:

- column = column number.
- min = Lowest quality score value found in this column.
- max = Highest quality score value found in this column.
- Q1 = 1st quartile quality score.
- med = Median quality score.
- Q3 = 3rd quartile quality score.
- IQR = Inter-Quartile range (Q3-Q1).
- 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.

We compute the quality statistics and nucleotide distribution for the select sequence reads with compute quality statistics software, and the output file of compute quality statistics is as Table 1.

Table 1 The quality statistics and nucleotide distribution

column	min	max	Q1	med	Q3	IQR	A_Count	C_Count	G_Count	T_Count	N_Count
1	2	34	31	33	34	3	34	13	8	27	18
2	16	34	31	34	34	3	21	38	15	26	0
3	25	34	31	34	34	3	20	21	39	20	0
4	32	37	35	37	37	2	31	17	33	19	0
5	33	37	35	37	37	2	37	16	29	18	0
6	32	37	37	37	37	0	28	25	23	24	0
7	30	37	37	37	37	0	24	29	26	21	0
8	33	37	37	37	37	0	42	19	20	19	0
9	32	39	39	39	39	0	27	12	37	24	0
10	27	39	39	39	39	0	26	18	27	29	0
11	23	39	39	39	39	0	20	30	23	27	0
12	19	39	39	39	39	0	26	24	21	29	0
13	23	39	39	39	39	0	29	21	24	26	0
14	10	41	39	41	41	2	27	32	17	24	0
15	27	41	40	41	41	1	44	16	18	22	0
16	26	41	40	41	41	1	21	14	32	33	0
17	29	41	40	41	41	1	30	17	19	34	0
18	32	41	39	40	41	2	31	29	19	21	0
19	33	41	40	41	41	1	28	26	19	27	0
20	33	41	40	41	41	1	32	29	20	19	0

Create the Quality Scores and the Nucleotide Distribution Graph

For learning and analysis the sequence reads information, we should create the quality scores and nucleotide distribution graph after compute the quality statistics and nucleotide distribution. Draw quality score boxplot is the software to create quality scores graph, and draw nucleotides distribution chart is the software to create nucleotide distribution graph. This software is based on the output file of compute quality statistics, and is also integrated into the Galaxy scientific workflow.

Create the Quality Scores Boxplot Graph. We create quality scores graph base on the above output file of compute quality statistics with draw quality score boxplot software. Quality score boxplot graph 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 score. The quality score is Fred [8] [9].

In Fig. 1, the black horizontal lines are medians, the rectangular red boxes show the Inter-quartile Range (IQR) (top value is Q3, bottom value is Q1), and the whiskers show outliers at max. 1.5*IQR.

Create the Nucleotide Distribution Chart Graph. After draw the quality scores boxplot graph, we create a stacked-histogram graph for the nucleotide distribution base on the above output file of compute quality statistics with draw nucleotides distribution chart software. Nucleotide distribution chart graph shows as Fig. 2.

In Fig. 2, 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 [10]. 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. 2. The nucleotides distribution is from Table 1. For example, A count is 34, C count is 13, G count is 8, T count is 27, and N count is 18 in column 1. Then, the summation count is 100. We can compute the percent of each nitrogenous base. The percent of A is 34%, the percent of C is 13%, the percent of G is 8%, the percent of T is 27%, and the percent of N is 18%. All the above results are corresponding to the length of the charts in Fig. 2.

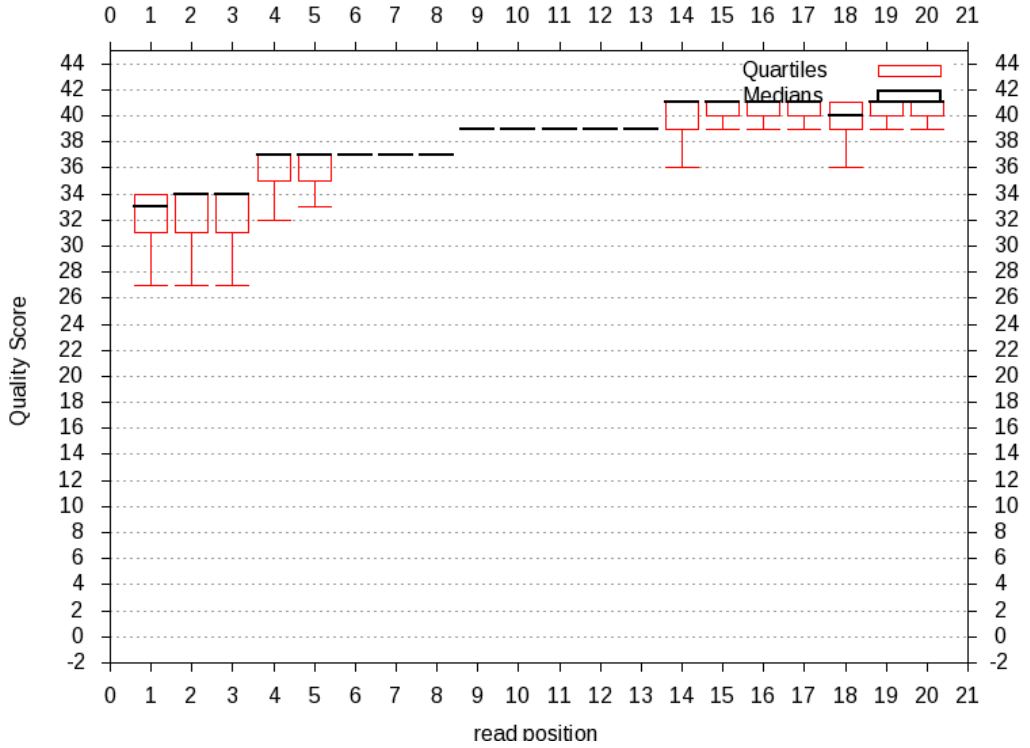

Figure 1. Quality scores boxplot graph

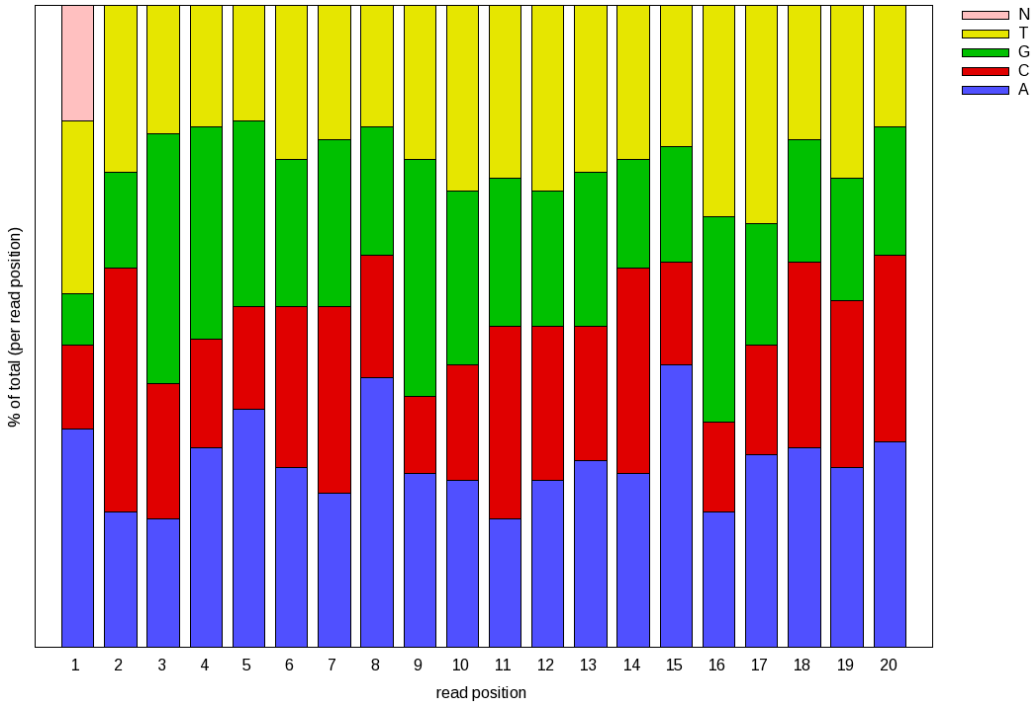

Figure 2. Nucleotide distribution chart graph

Summary

The quality statistics and nucleotide distribution are the two elements of the sequence reads information. In the bioinformatics analysis method, compute quality statistics is the software for compute the quality statistics and nucleotide distribution, and draw quality score boxplot and draw nucleotides distribution chart are the software for create the quality scores and nucleotide distribution graph.

For learning the sequence reads information of bioinformatics analysis method, we do a sequence reads analysis experiment to compute quality statistics and create the quality scores and nucleotide distribution graph. The experiment results show that compute quality statistics software, draw quality score boxplot software, and draw nucleotides distribution chart software can compute quickly and truly, and draw corresponding graph very well.

Acknowledgements

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

[5] J. Goecks, A. Nekrutenko, 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.

[6] 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.

[7] J. Taylor, I Schenck, D. Blankenberg, A. Nekrutenko: Using Galaxy to Perform Large-Scale Interactive Data Analyses". In Andreas D. Baxevanis. Current Protocols in Bioinformatics.

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

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

[10] 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.