Learning the Masking and Reverse Complement Method of the Sequence Reads

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: Reverse complement; Masking; Bioinformatics; Sequence reads; Quality scores

Abstract. There are some base characters with specified quality score value in the initial sequence reads. The base characters need the bioinformatics analysis method to mask such as mask by quality score. In the other hand, there are many sequence reads need the bioinformatics analysis method to reverse complement such as reverse-complement. For learning the masking and reverse complement method of the sequence reads, we select some initial sequence reads for the masking and reverse complement experiments, and compare the experiment results of mask by quality score and reverse-complement.

Introduction

With the application of next-generation sequencing (NGS) technology, bioinformatics analysis method for sequences have developed rapidly. Biological sequences are including DNA-seq, RNA-seq, Protein sequences, and etc al. For RNA-Seq is one of the applications of NGS, the NGS technology has become an important research on the transcriptomics [1]. FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity. It was originally developed at the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data, but has recently become the de facto standard for storing the output of high-throughput sequencing instruments such as the Illumina Genome Analyzer [2].

In the initial sequence reads with FASTQ format, we can mask the base characters with specified quality score value. The corresponding software of bioinformatics analysis method is mask by quality score [3] [4]. However, we can reverse complement the initial sequence reads including the quality scores, and the software of bioinformatics analysis method is reverse-complement [5]. For the above two software have been integrated into Galaxy [6] [7] [8]. In this paper, we select some initial sequence reads, and do the experiments on Galaxy for learning the masking and reverse complement method of the sequence reads. Then, we compare the experiment results of mask by quality score and reverse-complement.

Phred Quality Score in FASTQ Format

Phred quality scores [9] [10] have become widely accepted to characterize the quality of the sequence reads, and can be used to compare the efficacy of different sequencing methods. Perhaps the most important use of Phred quality scores is the automatic determination of accurate, quality-based consensus sequences. Phred quality scores are defined as a property which is logarithmically related to the base-calling error probabilities in FASTQ format file.

The following is the quality value characters in left-to-right increasing order of quality with ASCII format for Illumina 1.8+-assigned [11] identifier and the quality value characters are 0 to 41 with ASCII format

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
```

The following is an initial sequence reads for the experiment.

@HWI-ST1268:95:D1MY0ACXX:4:1101:5643:2140 1:N:0:AGTTCC
TCTGATCAGATGGTGTAAAATTTGGAATTCTCGGGTGCCAAGGAACTCCA
+
?@?BBDDDD<:<D+<<?EIBAEHDEHAEHFI>B?6:C@?90:00?BDD

Mask by Quality Score Experiment

In this paper, we respectively set quality score as 20, 25, 30, and respectively mask the above initial sequence reads with mask by quality score software on Galaxy as in Fig. 1.

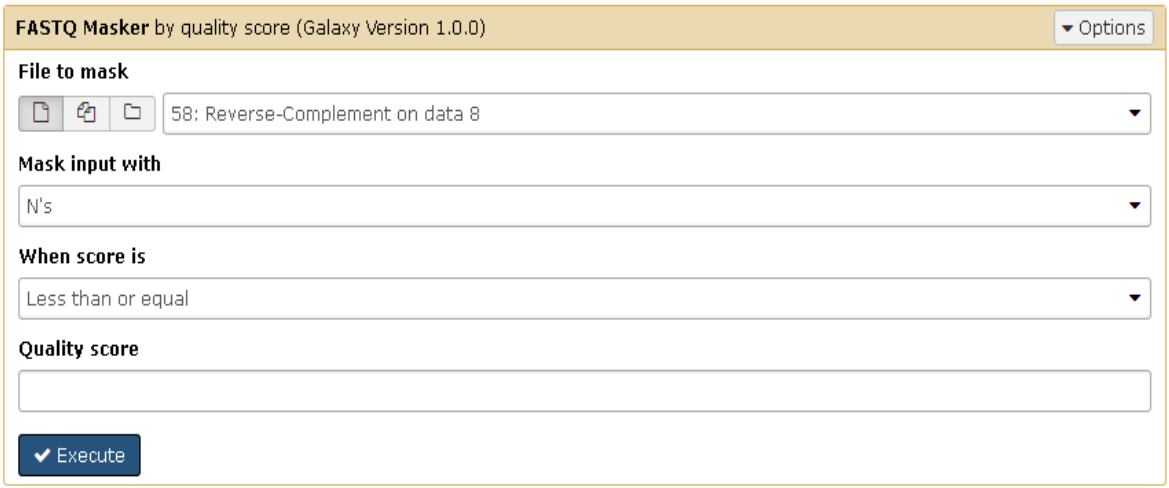

Figure 1. Mask by quality score on Galaxy

In the first experiment result for setting quality score as 20, we can see that the base characters (the quality value characters is at left of "5" ASCII format) are masked by "N", and we set "N" as red bole type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:5643:2140 1:N:0:AGTTCC
TCTGATCAGATGGNGTAAAATTTGGAATTCTCGGGTGCCANGNNACTCCA
+
?@?BBDDDD<:<D+<<?EIBAEHDEHAEHFI>B?6:C@?90:00?BDD>?

In the second experiment result for setting quality score as 25, we can see that the base characters (the quality value characters is at left of ":" ASCII format) are masked by "N", and we set "N" as red bole type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:5643:2140 1:N:0:AGTTCC
TCTGATCAGANGGNGTAAAATTTGGAATTCTCGGNNGCCNNNNNACTCCA
+
?@?BBDDDD<:<D+<<?EIBAEHDEHAEHFI>B?6:C@?90:00?BDD>?

In the third experiment result for setting quality score as 30, we can see that the base characters (the quality value characters is at left of "?" ASCII format) are masked by "N", and we set "N" as red bole type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:5643:2140 1:N:0:AGTTCC
NCNGATCAGNNNGNNNNAAATTTGGAATTCTNGNNNGCNNNNNNNCTCNN
+

?@?BBDDDD<:<D+<<?EIBAEHDEHAEHFI>B?6:C@?90:00?BDD>?

In the third experiment result for setting quality score as 35, we can see that the base characters (the quality value characters is at left of "D" ASCII format) are masked by "N", and we set "N" as red bole type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:5643:2140 1:N:0:AGTTCC
NNNNNNNNNNNNNNNNNAANNTTNGANTTCTNNNNNNNNNNNNNNNNNNNN
+
?@?BBDDDD<:<D+<<?EIBAEHDEHAEHFI>B?6:C@?90:00?BDD>?

Reverse-complement Experiment

In this paper, we reverse complement the above initial sequence reads with reverse-complement software on Galaxy as in Fig. 2.

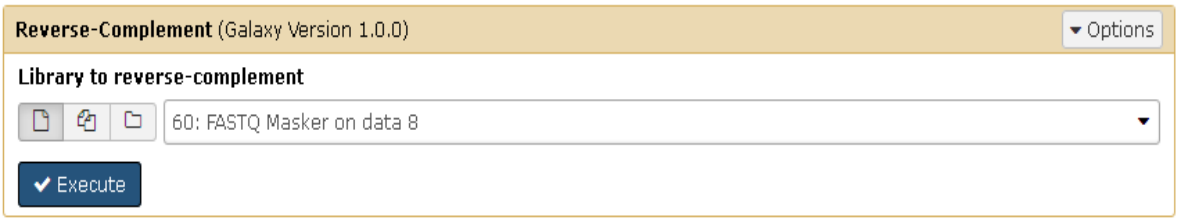

Figure 2. Reverse-complement software on Galaxy

Reverse is means that reverse the base characters order of the sequence reads, and complement is means that complement between the four nitrogenous bases. The four nitrogenous bases is (A) adenine, (T) thymine, (C) cytosine, and (G) guanine. A complement with T, and C complement with G. If the base characters include the quality scores, the quality scores should be reversed too. The experiment result is as following.

@HWI-ST1268:95:D1MY0ACXX:4:1101:5643:2140 1:N:0:AGTTCC
TGGAGTTCCTTGGCACCCGAGAATTCCAAATTTTACACCATCTGATCAGA
+
?>DDB?00:09?@C:6?B>IFHEAHEDHEABIE?<<+D<:<DDDDBB?@?

From the experiment result, we can see that the base characters complement each by each, and also see that all the complement characters and its quality scores are reversed.

Summary

In the initial sequence reads, there are some bases characters with specified quality score value need mask with mask by quality score software. Then, there are many sequence reads need the reverse complement with reverse-complement software. For mask by quality score software, the base characters are always masked by "N". For reverse-complement software, reverse is means that reverse the base characters order of the sequence reads, and complement is means that complement between the four nitrogenous bases. The above two software have been integrated into Galaxy.

For learning the masking and reverse complement method of the sequence reads, we do the experiments on Galaxy with the selecting initial sequence reads. In the masking experiment, we respectively set quality score as 20, 25, 30 with mask by quality score software, and can see the experiment results are correct with each quality score setting. In the reverse complement experiment, we can see the experiment result is very exact including the quality scores of the base characters.

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

[2] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, P. M. Rice: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 38 (6): p. 1767-1771.

[3] D.Blankenberg, A. Gordon, G. K.Von. N. Coraor, J. Taylor, A. Nekrutenko: Manipulation of FASTQ data with Galaxy. Bioinformatics. 26 (14): p. 1783–1785.

[4] Information on https://usegalaxy.org/

[5] Information on http://hannonlab.cshl.edu/fastx_toolkit/index.html

[6] 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.

[7] 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.

[8] 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.

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

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

[11] Information on http://www.illumina.com/