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

Learning the Quality Filter Method of Sequence by Quality Cut-off Value and Base Percent

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: Quality filter method; Sequence; Quality cut-off value; Base percent; Bioinformatics

Abstract. For the next bioinformatics analysis of sequence reads, the initial sequence reads with next-generation sequencing technology should be clean with the quality filter method. The basic quality filter method can filter the initial sequence reads by quality cut-off value and base percent. We select some initial sequence reads, and do a lot of quality filter experiments. For learning the quality filter method, we compare the results quality filter the experiment sequence reads by varies quality cut-off value and base percent.

Introduction

The emergence of next generation sequencing technology has made it possible for individual investigators to generate gigabases of sequencing data per week. Effective analysis and manipulation of these data is limited due to large file sizes, so even simple tasks such as data filtration and quality assessment have to be performed in several steps [1]. For the next bioinformatics analysis of sequence reads, the initial sequence reads with next-generation sequencing technology should be clean with the quality filter method. The quality filter method of bioinformatics means that filters sequences based on Phred quality score [2] [3], and the basic quality filter method can filter the initial sequence reads by quality cut-off value and base percent.

In this paper, we introduce Phred quality score. Then, for more clear observe the quality value of experiment sequence reads, we convert the quality format of the selecting experiment sequence reads from ASCII to numeric with a tool of quality format converter [4]. At last, we do a lot of experiments with quality filter by varies quality cut-off value and base percent, and compare and analysis the results.

Phred Quality Score

Phred quality scores have become widely accepted to characterize the quality of DNA sequences, 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 used for assessment of sequence quality, recognition and removal of low-quality sequence (end clipping), and determination of accurate consensus sequences

Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P [2].

$$P = 10^{\frac{-Q}{10}} \tag{1}$$

Phred quality scores are logarithmically linked to error probabilities as Table 1

Table 1 Phred quality scores, probability of incorrect base call and base call accuracy

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality Format Converter

The quality format of the initial sequence reads always are ASCII format. The following is the quality value characters in left-to-right increasing order of quality with ASCII format for Illumina 1.8+-assigned [5] identifier

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

The above quality value characters are 0 to 41 with ASCII format. The quality format converter is a tool for converting ASCII format to/from numeric format. The following are the selecting experiment sequence reads.

@HWI-ST1268:95:D1MY0ACXX:4:1101:3854:2076 1:N:0:AGTTCC
NTTCCGAAGGTCTAAAGGATCTGGANTNNTNNGGTGCCAAGGAACTCCAG
+
#4=DDFFFHHHGHIHHIIJIIFIII#1##1##1::BGHIGGGIJJJJJJJ
@HWI-ST1268:95:D1MY0ACXX:4:1101:2584:2217 1:N:0:AGTTCC
CTTATTCTATAAAAGGACCCCTTGGAATTCTCGGGTGCCAAGGAACTCCA
+
;@@ABDDDAFD<AAE;FABBFFHEEC4ACCGH==F1?DD**0899??FDB
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
NCGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
#1=DDF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGIF<BFHF
@HWI-ST1268:95:D1MY0ACXX:4:1101:2385:2186 1:N:0:AGTTCC
TCGTAGTTGAACCTTGGGCCTGGCTGGCCTGGAATTCTCGGGTGCCAAGG
+
@@CDDFFFFHHHGDGIIJIJJIIIIGIGIG;FCDBH<F9?BF8@;CFH@F

We convert the selecting experiment sequence reads from ASCII format to numeric format with quality format converter. The results are as the following.

@HWI-ST1268:95:D1MY0ACXX:4:1101:3854:2076 1:N:0:AGTTCC
NTTCCGAAGGTCTAAAGGATCTGGANTNNTNNGGTGCCAAGGAACTCCAG
+
2 19 28 35 35 37 37 37 39 39 39 38 39 40 39 39 40 40 41 40 40 37 40 40 40 2 16 2 2 16 2 2 16 25 25 33 38 39 40 38 38 38 40 41 41 41 41 41 41 41
@HWI-ST1268:95:D1MY0ACXX:4:1101:2584:2217 1:N:0:AGTTCC
CTTATTCTATAAAAGGACCCCTTGGAATTCTCGGGTGCCAAGGAACTCCA
+
26 31 31 32 33 35 35 35 32 37 35 27 32 32 36 26 37 32 33 33 37 37 39 36 36 34 19 32 34 34 38 39 28 28 37 16 30 35 35 9 9 15 23 24 24 30 30 37 35 33
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
NCGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+

2 16 28 35 35 37 30 36 39 39 38 39 37 38 40 38 40 32 37 39 40 39 36 38 40 40 36 40 39 38 40 23 31 30 33 37 38 26 33 37 39 38 38 40 37 27 33 37 39 37

@HWI-ST1268:95:D1MY0ACXX:4:1101:2385:2186 1:N:0:AGTTCC

TCGTAGTTGAACCTTGGGCCTGGCTGGCCTGGAATTCTCGGGTGCCAAGG

+

31 31 34 35 35 37 37 37 37 39 39 39 38 35 38 40 40 41 40 41 41 40 40 40 40 38 40 38 40 38 26 37 34 35 33 39 27 37 24 30 33 37 23 31 26 34 37 39 31 37

Experiment with Quality Filter

Experiment Setting on Galaxy. The quality filter method of bioinformatics can filter the initial sequence reads by quality cut-off value and base percent. Galaxy[6][7][8] is a scientific workflow, data integration[9][10], and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience. Although it was initially developed for genomics research, it is largely domain agnostic and is now used as a general bioinformatics workflow management system [11].

We do a lot of experiments with quality filter by varies quality cut-off value and base percent on Galaxy as in Fig. 1. Then, we set base percent as 90%, and respectively set quality cut-off value as 20, 25, 30, 35, 40.

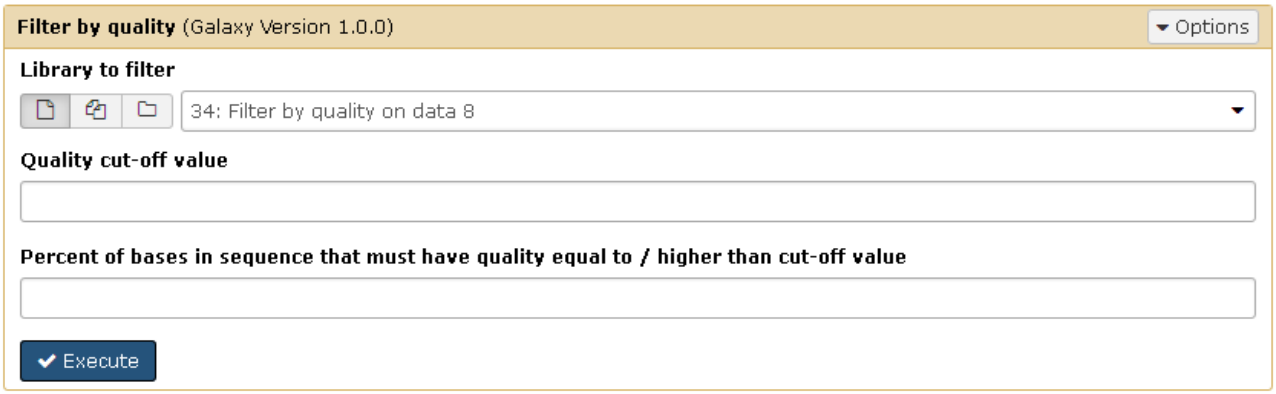

Figure 1. Quality filter by varies quality cut-off value and base percent on Galaxy

Experiment Results and Analysis. First, we set base percent as 90% and quality cut-off value as 20, and quality filter the above selecting experiment sequence reads. We can see that one sequence read is discarded. For the length of all sequence reads are 50, we set base percent as 90% and quality cut-off value as 20. It means that the bases number (quality value is less than 20) of the sequence read is 10 and it is more than 5, and this sequence read should be discarded. The following is the discarded sequence read, and we set the bases (quality value is less than 20) as bole italic type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:3854:2076 1:N:0:AGTTCC

NTTCCGAAGGTCTAAAGGATCTGGANTNNTNNGGTGCCAAGGAACTCCAG

+

2 19 28 35 35 37 37 37 39 39 39 38 39 40 39 39 40 40 41 40 40 37 40 40 40 ***2 16 2 2 16 2 2 16*** 25 25 33 38 39 40 38 38 38 40 41 41 41 41 41 41 41

In the same way, when setting base percent as 90% and quality cut-off value as 25, we can see the following sequence read is discarded, and the bases number (quality value is less than 25) of the following sequence read is 8 and it is more than 5. We can also see the bases number (quality value is less than 20) of the following sequence read is 5 and it is equal to5, and then, the following sequence read should not be discarded with base percent as 90% and quality cut-off value as 20. The following is the discarded sequence read, and we set the bases (quality value is less than 25) as bole

italic type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:2584:2217 1:N:0:AGTTCC
CTTATTCTATAAAAGGACCCCTTGGAATTCTCGGGTGCCAAGGAACTCCA
+
26 31 31 32 33 35 35 35 32 37 35 27 32 32 36 26 37 32 33 33 37 37 39 36 36 34 ***19*** 32 34 34 38 39 28 28 37 ***16*** 30 35 35 ***9 9 15 23 24 24*** 30 30 37 35 33

When setting base percent as 90% and quality cut-off value as 30, we can see the following sequence read is discarded. We set the bases (quality value is less than 30) as bole italic type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
NCGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
2 16 28 35 35 37 30 36 39 39 38 39 37 38 40 38 40 32 37 39 40 39 36 38 40 40 36 40 39 38 40 ***23*** 31 30 33 37 38 ***26*** 33 37 39 38 38 40 37 ***27*** 33 37 39 37

When setting base percent as 90% and quality cut-off value as 35, we can see the following sequence read is discarded. We set the bases (quality value is less than 35) as bole italic type.

@HWI-ST1268:95:D1MY0ACXX:4:1101:2385:2186 1:N:0:AGTTCC
TCGTAGTTGAACCTTGGGCCTGGCTGGCCTGGAATTCTCGGGTGCCAAGG
+
31 31 34 35 35 37 37 37 37 39 39 39 38 35 38 40 40 41 40 41 41 40 40 40 40 38 40 38 40 38 26 37 34 35 33 39 ***27*** 37 ***24*** 30 33 37 ***23 31 26 34*** 37 39 ***31*** 37

When setting base percent as 90% and quality cut-off value as 40, we can see all selecting sequence read are discarded for the bases number (quality value is less than 40) of all selecting sequence reads are more than 5.

Summary

With the quality filter method of bioinformatics, we can quality filter the sequence reads. There are many quality filter method, and the basic quality filter method can filter the initial sequence reads by quality cut-off value and base percent.

In this paper, we introduce Phred quality score and convert the quality format of the selecting experiment sequence reads, and do a lot of quality filter experiments with quality filter by varies quality cut-off value and base percent. The experiment results show that quality filter by varies quality cut-off value and base percent can filter the initial sequence reads quickly and exactly. For other filter method such as filter by quality score and length, we can also learning with the same way.

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] G. Georgiy, K. Kamil, R. Mark, M. A. Antonio, J. H. Jesse, B. Efren, G. Sharu, W. William, F. Yuriy: Slim-Filter: an interactive windows-based application for illumina genome analyzer data assessment and manipulation. BMC Bioinformatics, 2012. **13**: p. 166

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

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

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

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

[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. Blankenberg, N.Coraor, G. K. Von, J. Taylor, A. Nekrutenko: Integrating diverse databases into a unified analysis framework: A Galaxy approach. Database. 2011

[10] 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.

[11] Information on https://wiki.galaxyproject.org/PublicGalaxyServers.