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

Learning the Trimming Method of Sequence for Removing the Barcodes or Noise

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: Trimming method; Barcodes or noise; Sequence; Bioinformatics; Offsets

Abstract. The initial sequence reads with next-generation sequencing technology have some barcodes or noise, and need to be removed by trimming method of bioinformatics. The trimming method includes trimmer by column with absolute for fixed length reads or percentage for variable length reads and quality trimmer by sliding window. For learning the trimming method of sequence for removing the barcodes or Noise, we do some trimming sequence reads experiments, study the trimmer by column with absolute for fixed length reads or percentage for variable length reads, and compare the results for the different setting of quality score values for the quality trimmer by sliding window.

Introduction

With the application of next-generation sequencing (NGS) technology [1], bioinformatics analysis method for sequences has developed rapidly. The initial sequence reads with next-generation sequencing technology have some barcodes or noise, and need to be removed by trimming method of bioinformatics.

In this paper, we study the trimmer by column with absolute for fixed length reads or percentage for variable length reads with a sample experiment. Then, we introduce Phred quality score [2][3] and the range of scores of Illumina 1.8+-assigned [4] identifier, and do comparing experiments for the different setting of quality score values for the quality trimmer by sliding window.

Experiment of the Trimmer by Column

In the trimmer by column, the base offsets can be defined as absolute for fixed length reads or percentage for variable length reads. The following is an Illumina 1.8+-assigned identifier example with foure lines per sequence.

```
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
NCGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
#1=DDF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGIF<BFHF
```

First, we choose the base offsets defined as absolute for fixed length reads, and set the offset from 5' end as 4 and the offset from 3' end as 6. The first sample experiment result as following.

```
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
TTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTC
+
DF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGI
```

In the first sample experiment result, we can see the sequence read be removed 4 in the beginning and 6 in the end.

Then, we choose the base offsets defined as percentage for variable length reads, and set the offset from 5' end as 10% and the offset from 3' end as 15%. The second sample experiment result as following.

@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
TGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAAC
+
F?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHG

In the second sample experiment result, we can see the sequence read be removed 5 in the beginning. For the read length is 50, 10% corresponds to absolute offset as 5. In the same way, 15% corresponds to absolute offset as 8 for offset is rounded to the nearest integer. Then, we can see the sequence read be removed 8 in the end

Experiment of the Quality Trimmer by Sliding Window

For the comparing experiments for the different setting of quality score values, we introduce Phred quality score and the range of scores of Illumina 1.8+-assigned identifier at first. Then, we do comparing experiments for the different setting of quality score values.

Phred Quality Score and the Range of Scores of Illumina 1.8+-Assigned Identifier. A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing [5]. It was originally developed for Phred base calling to help in the automation of DNA sequencing in the Human Genome Project. Phred quality scores are assigned to each nucleotide base call in automated sequencer traces [6].

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. For sequence raw reads with various FASTQ formats, the range of scores will depend on the technology and the base caller used. Quality scores are normally stored together with the nucleotide sequence in the widely accepted FASTQ format.

A FASTQ file normally uses four lines per sequence [7] [8].

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence. Here are the quality value characters in left-to-right increasing order of quality (ASCII):

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

The range of scores of various FASTQ formats will depend on the technology and the base caller used. Fig. 1 is the range of scores of Illumina 1.8+-assigned identifier.

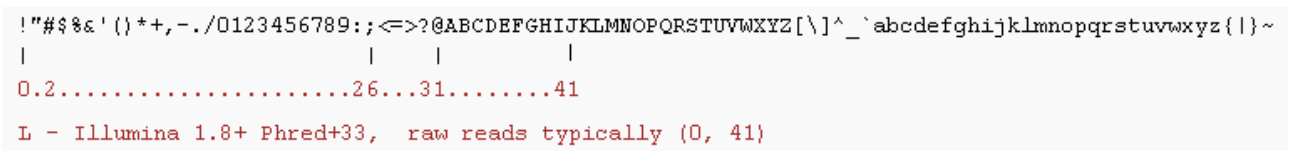

Figure 1. The range of scores of Illumina 1.8+-assigned identifier

The Comparing Experiment Result. In the quality trimmer by sliding window, we can trim the ends of reads based upon the aggregate value of quality scores found within a sliding window. When the sliding window of size is 1, it is equivalent to 'simple' trimming of the ends. We can specify the aggregating action (min, max, and sum, mean) to perform on the quality score values found within the sliding window to be used with the defined comparison operation and comparison value. We can provide a maximum count of bases that can be excluded from the aggregation within the window. When set, this tool will first check the aggregation of the entire window, then after removing 1 value, then after removing 2 values, up to the number declared. Setting this value to be equal to or greater than the window size will cause no trimming to occur [9].

In this paper, we do comparing experiments for the different setting of quality score values with the quality trimmer by sliding window, and set all the experiment elements expect quality score as the default values [10] as in Fig. 2.

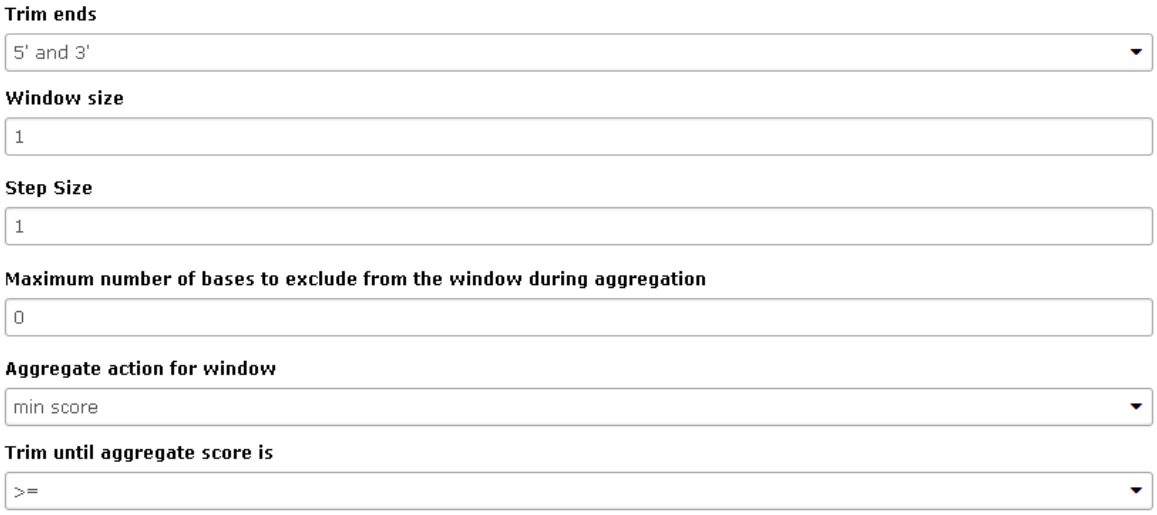

Figure 2. The setting of all the experiment elements expect quality score

Then, we do comparing experiments for the different setting of quality score values such as 10, 20, and 30. The following is an Illumina 1.8+-assigned identifier example with foure lines per sequence.

```
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
NCGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
#1=DDF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGIF<BFHF
```

The comparing experiment result as following. The first result is the setting of quality score values as 10, the second result is the setting of quality score values as 20, and the third result is the setting of quality score values as 30.

```
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
CGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
1=DDF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGIF<BFHF

@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
GATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
=DDF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGIF<BFHF

@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
ATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
```

+
DDF?EHHGHFGIGIAFHIHEGIIEIHGI8@?BFG;BFHGGIF<BFHF

Comparing with Illumina 1.8+-assigned identifier example, we can see that the "N" is removed as the barcodes or noise in the head of the sequence reads in the first result. It is because the corresponding quality score is "#" and quality score values is 2 (less than 10) in Fig. 1. In the same way, we can see that the "NC" is removed as the barcodes or noise in the head of the sequence reads in the second result. However, we can that the "NCG" is removed as the barcodes or noise in the head of the sequence reads in the third result.

Conclusions

By trimming method of bioinformatics, we can remove the barcodes or noise of the initial sequence reads with next-generation sequencing technology. There are many trimming method such as trimmer by column and quality trimmer by sliding window. Then, the trimmer by column includes with absolute for fixed length reads and with percentage for variable length reads.

In this paper, we do some experiments of the trimmer by column, and respectively study the experiment results with absolute for fixed length reads and with percentage for variable length reads. The experiment results show that the trimmer by column is a sample and convenient trimming method of sequence for removing the barcodes or noise. Then, we do a comparing experiment of the quality trimmer by sliding window. The experiment results show that the quality trimmer by sliding window is a complex trimming method than the trimmer by column, but it is a more effective trimming method of sequence for removing the barcodes or noise. In the future, we should choose the suitable trimming method with the requirements according to the specific quality analysis.

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

[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://www.illumina.com/

[5] B. Ewing, L. Hillier, M. C. Wendl, P. Green: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8(3): p.175-185.

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

[7] H. H. Shi, Xin Xu: Learning the comparing and converting method of sequence phred quality score. *Proc. 2016 4th International Conference on Management, Education, Information and Contro* (Shen yang, China, September 24-26, 2016). In press.

[8] H. H. Shi, Xin Xu: Learning the sequences quality control of bioinformatics analysis method *Proc. 2016 4th International Conference on Management, Education, Information and Contro* (Shen yang, China, September 24-26, 2016). In press

[9] D. Blankenberg, A. Gordon, K. G. Von, N. Coraor, J. Taylor, A. Nekrutenko: Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15; 26(14):p.1783-1785.

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