Learning the Comparing and Converting Method of Sequence Phred Quality Score

Henghua Shi[1, a], Weiyu Li[2, b] and Xin Xu[3, c]

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

[2]College of Plant Science and Technology, Beijing University of Agriculture, China

[3]Communication Technology Bureau, Xinhua News Agency, China

[a]henghuashi@163.com, [b]youges@163.com, [c]lwy@bac.edu.cn

Keywords: Phred quality scores; Sequence; Bioinformatics; FASTQ format

Abstract. The Phred quality score can measure the sequence quality, and quality scores are normally stored together with the nucleotide sequence in the widely accepted FASTQ format. For sequence raw reads with various FASTQ formats, the range of scores will depend on the technology and the base caller used. With the different technology, the range of scores is different. For learning the comparing and converting method of sequence Phred quality score, we do an experiment, and analyze the different between various FASTQ quality formats.

Introduction

A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing [1]. 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 [2].

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.

In this paper, we study Phred quality score and FASTQ format. Then, we make a RNA sequence as the experiment resource and do an experiment to analyze the different between various FASTQ quality formats for learning the comparing and converting method of sequence Phred quality score.

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.

Originally, Phred quality scores were primarily used by the sequence assembly program Phrap. Phrap was routinely used in some of the largest sequencing projects in the Human Genome Sequencing Project and is currently one of the most widely used DNA sequence assembly programs in the biotech industry. Phrap uses Phred quality scores to determine highly accurate consensus sequences and to estimate the quality of the consensus sequences. Phrap also uses Phred quality scores to estimate whether discrepancies between two overlapping sequences are more likely to arise from random errors, or from different copies of a repeated sequence.

Within the Human Genome Project, the most important use of Phred quality scores was for automatic determination of consensus sequences. Before Phred and Phrap, scientists had to carefully look at discrepancies between overlapping DNA fragments; often, this involved manual determination of the highest-quality sequence, and manual editing of any errors. Phrap's use of Phred quality scores effectively automated finding the highest-quality consensus sequence; in most cases, this completely circumvents the need for any manual editing. As a result, the estimated error rate in assemblies that were created automatically with Phred and Phrap is typically substantially lower than the error rate of manually edited sequence.

In 2009, many commonly used software packages make use of Phred quality scores, albeit to a different extent. Programs like Sequencher use quality scores for display, end clipping, and consensus determination; other programs like CodonCode Aligner also implement quality-based consensus methods.

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

$$P = 10^{\frac{-Q}{10}} \qquad (1)$$

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 scores are normally stored together with the nucleotide sequence in the widely accepted FASTQ format. They account for about half of the required disk space in the FASTQ format (before compression), and therefore the compression of the quality values can significantly reduce storage requirements and speed up analysis and transmission of sequencing data. Both lossless and lossy compression are recently being considered in the literature. For example, the algorithm QualComp [3] performs lossy compression with a rate (number of bits per quality value) specified by the user. Based on rate-distortion theory results, it allocates the number of bits so as to minimize the MSE (mean squared error) between the original (uncompressed) and the reconstructed (after compression) quality values. Other algorithms for compression of quality values include SCALCE [4] and Fastqz.[5] Both are lossless compression algorithms that provide an optional controlled lossy transformation approach. For example, SCALCE reduces the alphabet size based on the observation that “neighboring” quality values are similar in general.

FASTQ Format

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 [6].

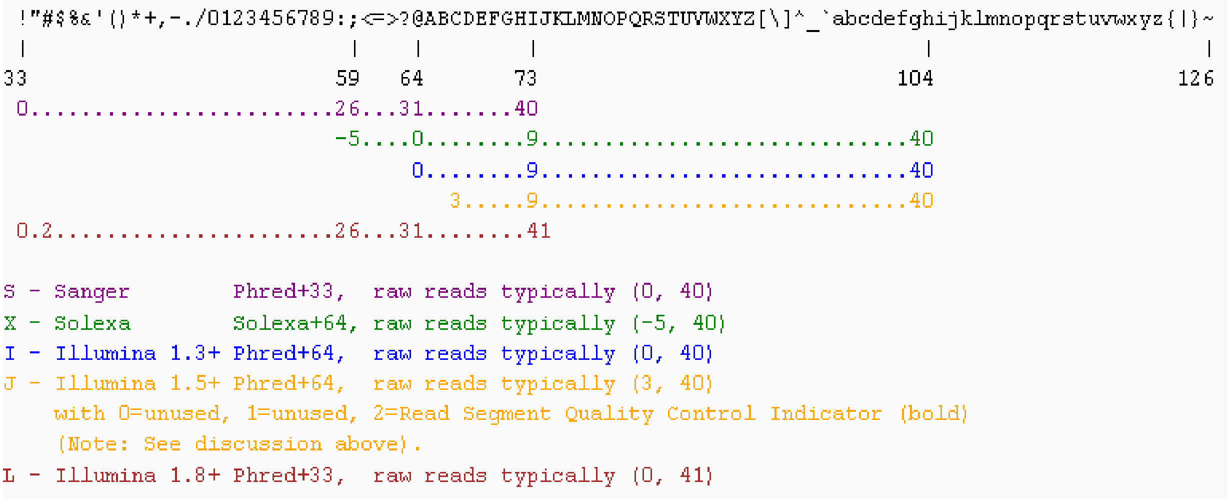
A FASTQ file normally uses four lines per sequence.

- 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 comparing of different range of scores for Sanger, Illumina and Solexa.



```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0.........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Figure 1. The different range of scores for Sanger, Illumina and Solexa

Experiment Results

FASTQ read simulation has been approached by several tools [7] [8]. A comparison of those tools can be seen in [9]. In this paper, we do an experiment with the above tools to analyze the different between various FASTQ quality formats for learning the comparing and converting method of sequence Phred quality score. The following is an Illumina 1.8+-assigned [10] 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
```

We respectively convert the above example to a Sanger-assigned and a Solexa-assigned. The experiment results of converting from Illumina 1.8+-assigned to Sanger-assigned is as following, and we can see that Sanger-assigned is the same with Illumina 1.8+-assigned. It is because the range of scores is similar as in Fig. 1.

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

The experiment results of converting from Illumina 1.8+-assigned to Solexa-assigned is as following, and we can see that Solexa-assigned is different with Illumina 1.8+-assigned. It is because the range of scores is different as in Fig. 1.

```
@HWI-ST1268:95:D1MY0ACXX:4:1101:2314:2086 1:N:0:AGTTCC
NCGATTGAATGGTCCGGTTGGAATTCTCGGGTGCCAAGGAACTCCAGTCA
+
;;;BBE;CGGFGEFHFH;EGHGCFHHCHGFH;;;>EF;>EGFFHE;>EGE
```

Summary

Phred quality scores are assigned to each nucleotide base call in automated sequencer traces. The range of scores of sequence raw reads with various FASTQ formats is different, and will depend on the technology and the base caller used. For learning the comparing and converting method of sequence Phred quality score, we do the converting from Illumina 1.8+-assigned to Sanger-assigned and Solexa-assigned experiments in this paper, and compare and analyze the different between various FASTQ quality formats.

With this paper for the comparing and converting method of sequence Phred quality score, we can covert and compare sequence Phred quality score between FASTQ quality formats better, and make a basic for learning 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] 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.

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

[3] I. Ochoa, et al. QualComp: a new lossy compressor for quality scores based on rate distortion theory. BMC Bioinformatics 14.1 (2013): p.187.

[4] F. Hach, I. Numanagić, C. Alkan, S. C. Sahinalp: SCALCE: boosting sequence compression algorithms using locally consistent encoding. Bioinformatics2012, 28(23): p. 3051-3057.

[5] Information on http://mattmahoney.net/dc/fastqz

[6] 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.

[7] W. Huang, L. Li, J. R. Myers, G. T. Marth: ART: a next-generation sequencing read simulator. Bioinformatics 28, p.593-594.

[8] D. Pratas, A. J. Pinho, O. S. Rodrigues, J. M. XS: a FASTQ read simulator. BMC Res. Notes 7, p. 40.

[9] M. Escalona, S. Roch, D. Posada: A comparison of tools for the simulation of genomic next-generation sequencing data. Nature Reviews Genetics, 17, p.459-469.

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