35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

In-Depth Analysis of Interrelation between Quality Scores and Real Errors in Illumina Reads

Sunyoung Kwon, Seunghyun Park, Byunghan Lee, and Sungroh Yoon

Abstract— **In sequencing results, the quality score is reported for each base, representing the probability that the base is called incorrectly. The notion of quality scores was initially developed for conventional Sanger sequencing, but is widely used for next-generation sequencing techniques, including Illumina. In this paper, we carry out in-depth analysis of quality scores reported for Illumina reads and present how they are related to real errors in the reads. We confirmed strong interrelation between quality scores and real errors in Illumina reads, and observed that reverse reads tend to have lower quality scores than forward reads in paired-end reads do. In addition, we discovered other interesting patterns from quality score analysis. Our hope is that the findings in this paper will be helpful for designing error-correction and/or filtering methods for next-generation sequencing.**

I. Introduction

Since the emergence of next-generation sequencing (NGS) technologies, bioinformatics approaches have been very active. Among existing NGS methods, Illumina sequencing may be the most popular at the moment [1]. It can generate enormous reads per run and is highly cost effective compared to the other NGS methodologies although it is less accurate [2], [3]. In terms of the market share, almost two thirds of NGS equipments are from llumina [4]. The inexpensive productivity of a large volume of sequence data is the primary advantage of Illumina sequencing.

However, it has higher error rates and much shorter read lengths than traditional Sanger sequencing has. The short read length can be tailored by merging the paired-end reads generated from the same amplicon [5]–[7]. The higher error rates can be compensated for by filtering out or correcting errors. For any biological analysis, properly handling erroneous reads is crucial for ensuring the correctness of downstream genomic analysis.

Several studies reported position-specific and sequence-specific effects and other reasons for Illumina errors. Miscalling more frequently occurs during the first and last cycles [8]. GC rich regions can be miscalled more often than the others by A to C and C to G [9]. Some other studies showed the effects of filtering methods such as B-tail trimming [10] or quality-score-based end-trimming to a uniform length [11]. Some paired-end read-merging methods choose the base which has the higher quality score as the right base, when mismatches occur [12], [13].

This work was supported by the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2011-0009963 to S.Y.).

S. Kwon is with the Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea.

S. Park is with the School of Electrical Engineering, Korea University, Seoul 136-713, Korea.

B. Lee is with the Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea.

S. Yoon is with the Department of Electrical and Computer Engineering, Bioinformatcs Institute, and the Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-744, Korea.

```
@SRR069027.1 HWUSI-EAS1661_9323_FC619KG:7:1:1128:17890 length=55
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGC
+SRR069027.1 HWUSI-EAS1661_9323_FC619KG:7:1:1128:17890 length=55
CCCCCCCBCCCCCCCCCCACC@CCCCCCC?CBCBCCDCC:C?@C@:<BC=7C###
```

Fig. 1. Example FASTQ file

In this paper, we focus on the quality score and perform an in-depth analysis of its effects in various aspects. In addition, we show possible filtering methods and their effects with respect to different standards and present several points that should be considered carefully when correcting a base by merging.

II. Background

When we use raw sequencing reads, we do not know the exact error locations, but for each base in a read, we know its error estimate Q. It is also known as the *quality score* and is defined as $Q = -10\log_{10}(p)$, where p represents the probability that the base calling is incorrect. In normal sequencing results, Q ranges between 2 and 40.

The nucleotide sequence and its associated quality scores are usually stored in the FASTQ format [14]. The quality score of a base is encoded with a single character whose ASCII value minus 33 corresponds to the quality score. For instance, character 'C' represents the quality score of 34 and '#' 2. Fig. 1 shows an example.

Illumina sequencers can easily generate multi-million reads. Many reads are replicates since the sequencing coverage and the duplication rate are high. Discarding erroneous data is thus effective to get reliable results and widely used in practice. In this paper, we examine the following filtering methods and their effects.

First, end trimming refers to a technique to remove all the bases following a predetermined position in a given read. The rationale behind this technique is that the probability of error increases substantially as we go to the end of a read in the Illumina sequencing technology. Second, read filtering is to discard a read from downstream analysis if the average quality score of the bases in that read is greater than a threshold. Third, base substitution means replacing an erroneous base by 'N' (wildcard symbol for nucleotides) in case the quality score of the base is low.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

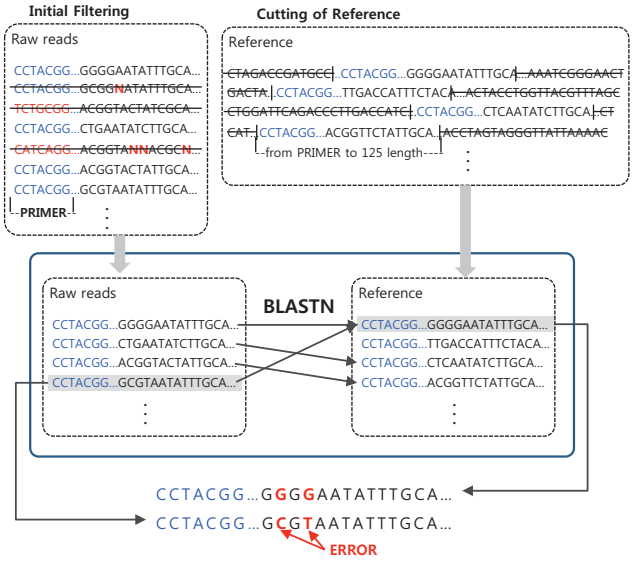

Fig. 2. Analysis Process. The top left represents filtering of raw reads which have the wrong primer or 'N' base. The top right represents cutting of the reference to the same length of reads from the primer. The middle represents the matching process of reads to references with BLASTN. The bottom shows the process comparing one of the reads and its reference.

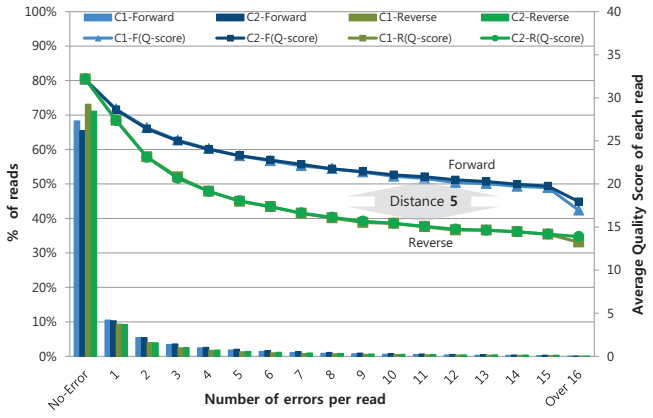

Fig. 3. Percentage distribution of reads and average quality score according to the number of errors for each read. The bar graphs represent the distribution of reads and the line graphs the average quality score of each read. A series of blue dots represent the forward side and green ones represent the reverse side.

III. Materials and Methods

A. Data

We used a public artificial microbial dataset [15]. The total number of raw reads is over 4 million. The length of each read is 125 nt with four different types: C1-forward/reverse and C2-forward/reverse. The total number of the reference reads is 90. The primer sequences are CCTACGGGAGGCAGCAG for forward and ATTACCGCGGCTGCTGG for reverse.

B. Analysis

Fig. 2 shows the analysis process. First, we prepare the filtered data set. Raw reads which have the wrong primer or base 'N' are filtered out. Out of each reference, we select only 125 bases starting from the primer. For the analysis in the reverse direction, reference data should be reversed and complemented due to the nature of paired-end reads.

To find the best matching reference for each read, we execute BLASTN [16] with the default options. Through this process, we postulate the most related reference as the error-free sequence. Some inappropriate reads in the result (*e.g.*, 'No HIT' or erroneous starting base) are also filtered. All reads and references start in the primer region. The indel error is not taken into account in this paper because Illumina sequencers have more substitution-type miscalls than indels [17]. We then compare each read to the reference, and each mismatching base is considered an error.

C. Evaluation metrics

For the end trimming technique, we define the ratio of error in total (RET) for base position x as

$$RET = \frac{\text{\# erroneous bases at } x \text{ in all reads}}{\text{\# total number of errors in all reads}} \quad (1)$$

For the read filtering method, we define RET for the average score threshold q as

$$RET = \frac{\text{\# errors in reads with average quality score } \lfloor q \rfloor}{\text{\# total number of errors in all reads}} \quad (2)$$

the ratio of error in read (RER) as

$$RER = \frac{\text{\# errors in reads with average quality score } \lfloor q \rfloor}{\text{\# total number of bases in such reads}} \quad (3)$$

and the probability of erroneous read (PER) as

$$PER = \frac{\text{\# erroneous reads with average quality score } \lfloor q \rfloor}{\text{\# all reads with average quality score } \lfloor q \rfloor} \quad (4)$$

For the base substitution, we define the probability of erroneous base (PEB) for the base with quality score r as

$$PEB = \frac{\text{\# erroneous bases with quality score } r}{\text{\# all bases with quality score } r} \quad (5)$$

and the RET as

$$RET = \frac{\text{\# erroneous bases with quality score } r}{\text{\# total number of errors in all reads}} \quad (6)$$

IV. Experimental Results

A. Interrelation between the quality score and the number of errors

Fig. 3 shows that the more errors in a read the lower the average quality score. Errorless reads compose a large percentage (about 70% in the whole), so most of the reads can be thought of as reliable data without errors. The distribution of reads is sharply lowered as the number of errors increase, so 90% of the reads have less than 5 errors. When we analyze based on the direction of the reads, the reverse side shows a higher percentage of errorless reads than the forward side does (forward: 66.9%; reverse: 72.1%), which means that the reverse side has more reliable data than the forward side has.

The most striking observation from this result is that the reverse side has a lower quality score than the forward side has in the same situation. If there are at least one or more

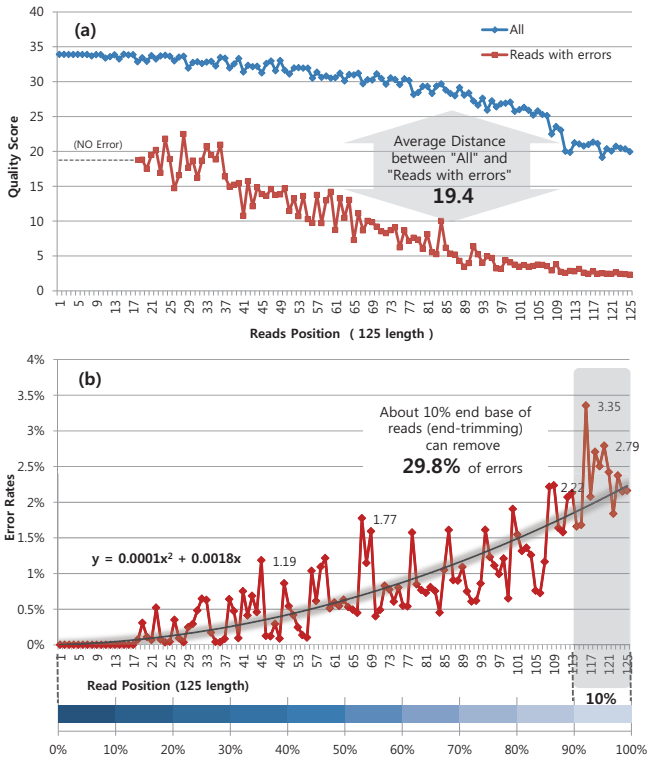

Fig. 4. Quality score and error rates according to the reads base position. (a) This chart shows the average quality score of each base position. The blue line represents the average quality score of all reads and the red line the score only if the base position has an error. (b) This chart shows actual error rates according to the read position. The red line represents the ratio of error in total (RET). We propose the blurred gray line as a secondary polynomial which matches well the real error. The horizontal bar on the bottom represents the distribution of data.

errors, the quality score of the reverse side is 5 points less than the quality score of the forward side on average. This observation may be considered when we merge paired-end reads. Most merging methods compare the quality scores when the two sides do not have the same base (called mismatch) and select the base that has the higher quality score over the other. However, the reverse side usually has a lower quality score, so compensation values must be applied.

B. Analysis of the positional influence

In Fig. 4 no error was observed at the first 17 bases because we filtered the reads out using the primer. This analysis shows that the quality score decreases as the base position moves to the end of read. This is because of Illumina-specific miscalling features such as cycle-dependent variations of the cross-talk, declining intensities and phasing [8].

The quality score, only in the case of an error, also decreases towards the end of read, maintaining the distance score 19.4 in the whole reads. Thus, if a base of the front part is miscalled, the quality score may not be the lowest score but be the proportionally lowest value around the base position. When we merge and correct the paired-end reads, we should consider the decreasing feature if the two sides are highly overlapped.

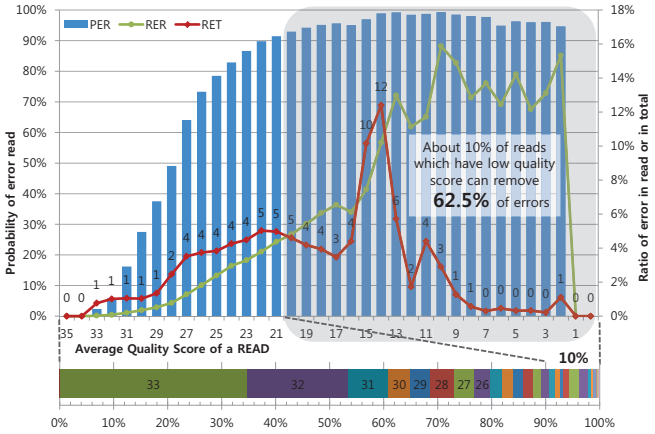

Fig. 5. Probability of erroneous reads and the error rates of the whole data according to the average quality score of a read. The bar graph represents the probability of a erroneous read (PER) and the red line the ratio of the errors in total (RET). The green line represents the ratio of the error in a read (RER) and the horizonal bar the distribution of data which have the same average quality score.

The actual error rates of Illumina reads increase towards the end of reads, showing a similar distribution to the proposed secondary polynomial. With this polynomial, we can expect the error rates if a read has a longer length. However, the error rates do not form smooth curves, probably due to the fact that we used a small amount of reference data. A particular location, which has a higher ratio of specific bases, seems to produce an effect usually known as sequence-specific error. However, it is true that the error rates increase toward the end of reads. If we filter out the errors by trimming the later part of reads across-the-board, up to approximately 10% of reads out of the whole (*e.g.*, in this case from 113 to 125 base positions), we could remove roughly 29.8% of errors.

C. Quality-score-based read filtering

Fig. 5 shows that the lower the average quality score, the higher the probability of erroneous reads. The reads which get quality scores of over 30 points on average have many errorless reads and have low error counts. Their proportions are approximately 65% out of the whole. Thus, most reads have very good quality scores and have few errors. Once the score of a read is under 20 points, the read has over 90% probability of being an erroneous read. If we filter out errors by eliminating the reads that have lowest average quality score, up to 10% of reads out of the whole (*e.g.*, in this case from 20 to 0), we could remove about 62.5% of all the errors.

D. Handling individual bases

Fig. 6 shows that the error probabilities of the bases tend to increase until the quality score becomes 4. When we focus on the quality score of 4, the error proportion is very low (0.001%). However, most of the bases are erroneous (with 67% probability). The quality score of 2 (the lowest value) gives the highest error rates in total mostly because of the

TABLE I

COMPARISON OF FILTERING METHODS

Method	End trimming	Read filtering	Base substitution
Filtering criterion	Base position	Average quality score of a read	Quality score of a base
Filtering effect (%) (10% threshold)	29.8	62.5	84.9
Filtering principle	Remove end part to the same length	Remove reads with low average quality score	Replace bases with low quality scores by 'N'.
Remarks	Easy to apply but limited filtering effects	Need to examine all the quality scores for averaging	Handling 'N' may cause inconvenience in downstream analysis

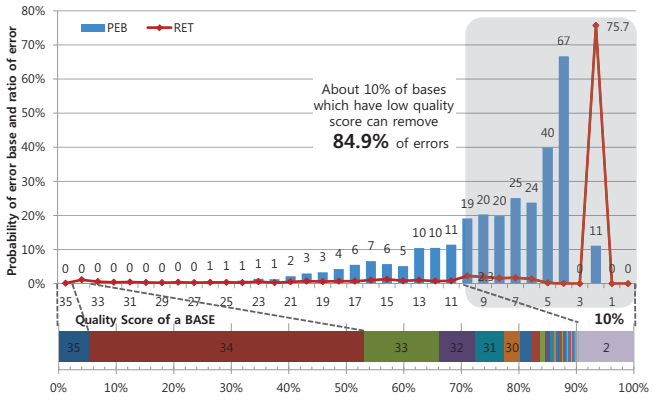

Fig. 6. Probability of erroneous bases and the error rates out of the whole according to the quality score of a base. The bar graph represents the probability that a base is an error (PEB) and the line the ratio of the error in total (RET). The horizontal bar represents the distribution of data, which have the same quality scores.

error proportion (9.48%). However, the probability is lower than when the quality score is 4. Whereas most cases with the quality score 2 (about 89%) are errorless bases, most cases with the quality score of 4 are erroneous bases.

To minimize the loss of raw data and to maximize the filtering effect, we should handle the most effective quality score of 4, or in a wider range (from 10 to 4). If we filter out the bases that have the lowest quality scores, up to 10% of bases out of the whole (*e.g.*, in this case from 10 to 0), we could remove roughly 84.9% of the errors. Unlike end-trimming or eliminating a read, the base can be handled by substituting the base with 'N' because the error is individually dispersed in the whole length of the read although most of the errors are located in the end.

V. CONCLUSION

In this paper, we carried out in-depth analysis of Illumina sequencing data and confirmed that called bases and the associated quality scores are closely related. We also compared three types of error handling methods listed in Table 1. Each technique has its own advantages and disadvantages and should be selected depending on the specific characteristic of the data set under experiment.

REFERENCES

[1] M. Metzker, "Sequencing technologies?the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2009.

[2] M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers," *BMC genomics*, vol. 13, no. 1, p. 341, 2012.

[3] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, "Comparison of next-generation sequencing systems," *Journal of Biomedicine and Biotechnology*, vol. 2012, 2012.

[4] GenomeWeb. http://www.genomeweb.com.

[5] G. Gloor, R. Hummelen, J. Macklaim, R. Dickson, A. Fernandes, R. MacPhee, and G. Reid, "Microbiome profiling by illumina sequencing of combinatorial sequence-tagged pcr products," *PLoS One*, vol. 5, no. 10, p. e15406, 2010.

[6] S. Rodrigue, A. Materna, S. Timberlake, M. Blackburn, R. Malmstrom, E. Alm, and S. Chisholm, "Unlocking short read sequencing for metagenomics," *PLoS One*, vol. 5, no. 7, p. e11840, 2010.

[7] H. Zhou, D. Li, N. Tam, X. Jiang, H. Zhang, H. Sheng, J. Qin, X. Liu, and F. Zou, "Bipes, a cost-effective high-throughput method for assessing microbial diversity," *The ISME Journal*, vol. 5, no. 4, pp. 741–749, 2010.

[8] M. Kircher, U. Stenzel, J. Kelso *et al.*, "Improved base calling for the illumina genome analyzer using machine learning strategies," *Genome Biol*, vol. 10, no. 8, p. R83, 2009.

[9] J. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput dna sequencing," *Nucleic acids research*, vol. 36, no. 16, pp. e105–e105, 2008.

[10] A. Minoche, J. Dohm, and H. Himmelbauer, "Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems," *Genome Biol*, vol. 12, no. 11, p. R112, 2011.

[11] V. Kunin, A. Engelbrektson, H. Ochman, and P. Hugenholtz, "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates," *Environmental Microbiology*, vol. 12, no. 1, pp. 118–123, 2009.

[12] T. Magoč and S. Salzberg, "Flash: fast length adjustment of short reads to improve genome assemblies," *Bioinformatics*, vol. 27, no. 21, pp. 2957–2963, 2011.

[13] B. Liu, J. Yuan, S. Yiu, Z. Li, Y. Xie, Y. Chen, Y. Shi, H. Zhang, Y. Li, T. Lam *et al.*, "Cope: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly," *Bioinformatics*, vol. 28, no. 22, pp. 2870–2874, 2012.

[14] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants," *Nucleic acids research*, vol. 38, no. 6, pp. 1767–1771, 2010.

[15] A. Bartram, M. Lynch, J. Stearns, G. Moreno-Hagelsieb, and J. Neufeld, "Generation of multimillion-sequence 16s rrna gene libraries from complex microbial communities by assembling paired-end illumina reads," *Applied and environmental microbiology*, vol. 77, no. 11, pp. 3846–3852, 2011.

[16] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[17] S. Hoffmann, C. Otto, S. Kurtz, C. Sharma, P. Khaitovich, J. Vogel, P. Stadler, and J. Hackermüller, "Fast mapping of short sequences with mismatches, insertions and deletions using index structures," *PLoS computational biology*, vol. 5, no. 9, p. e1000502, 2009.