JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 8, 2017
Mary Ann Liebert, Inc.
Pp. 746–755
DOI: 10.1089/cmb.2017.0021

Zseq: An Approach for Preprocessing Next-Generation Sequencing Data

ABEDALRHMAN ALKHATEEB and LUIS RUEDA

ABSTRACT

Next-generation sequencing technology generates a huge number of reads (short sequences), which contain a vast amount of genomic data. The sequencing process, however, comes with artifacts. Preprocessing of sequences is mandatory for further downstream analysis. We present Zseq, a linear method that identifies the most informative genomic sequences and reduces the number of biased sequences, sequence duplications, and ambiguous nucleotides. Zseq finds the complexity of the sequences by counting the number of unique *k*-mers in each sequence as its corresponding score and also takes into the account other factors such as ambiguous nucleotides or high GC-content percentage in *k*-mers. Based on a *z*-score threshold, Zseq sweeps through the sequences again and filters those with a z-score less than the user-defined threshold.

Zseq algorithm is able to provide a better mapping rate; it reduces the number of ambiguous bases significantly in comparison with other methods. Evaluation of the filtered reads has been conducted by aligning the reads and assembling the transcripts using the reference genome as well as *de novo* assembly. The assembled transcripts show a better discriminative ability to separate cancer and normal samples in comparison with another state-of-the-art method. Moreover, *de novo* assembled transcripts from the reads filtered by Zseq have longer genomic sequences than other tested methods. Estimating the threshold of the cutoff point is introduced using labeling rules with optimistic results.

Keywords: machine learning, next-generation sequencing, preprocessing, RNA-SEQ analysis.

1. INTRODUCTION

In the last decade, next-generation sequencing (NGS) technology has evolved rapidly, reducing the cost of genome sequencing and influencing the progression of cancer research and other fields. The main purpose of NGS studies is to find clues to gene and protein structures and functions in the sequenced reads.

School of Computer Science, University of Windsor, Windsor, Canada.

However, this advanced technology can also produce unexpected artifacts (Waszak et al., 2014; Lavezzo et al., 2016). Some of these artifacts come from cDNA library preparation; those are repetitive low-complex regions that appear in the sequenced reads (Mackinnon et al., 2009). High GC content is also a common bias due to cDNA library preparation, while GC content tends to last more in the preparation process (Yakovchuk et al., 2006). GC-content bias in reads is also known to aggravate genome assembly, and hence it may result in poor genome assembly. Nevertheless, the sequencing procedure itself can produce low-complex repetitive regions such as a sequence of ambiguous nucleotides. In general, it is not clear to what extent GC-content bias affects genome assembly (Chen et al., 2013).

A low-complexity sequence of nucleotides has highly biased distribution of nucleotides in a way that makes the sequence less diverse of unique *k*-mers of nucleotides. The lower the complexity of a sequence, the more likely that the sequence will be mapped to different parts of the genome. In other words, when we process low-complex sequences, there is less chance that we can align it to a specific part of the genome uniquely. This low level of certainty regarding the real position of a sequence makes it less desirable to be used.

Poly A/Poly T is a chain of A or T, used to prime the three and five sites in a genome sequence during cDNA library preparation (Brown, 2012). Poly A/T sequences may cause bias in the reads. The intronic Poly A/T tails tend to splice out rather than staying between coding exons (Zhao et al., 2014). The GC content represents the ratio of a G-C pair in the genome sequence. The stop codons show a significantly high ratio of A-T nucleotides (Wuitschick and Karrer, 1999), while coding codons have a higher GC content (Pozzoli et al., 2008). The GC content of a gene plays an important role in carrying the genetic information. The GC content of the human genome varies among different chromosomes. However, the average GC content of the human genome is 41% (Vogel, 1997). The representation of A+T sequences can

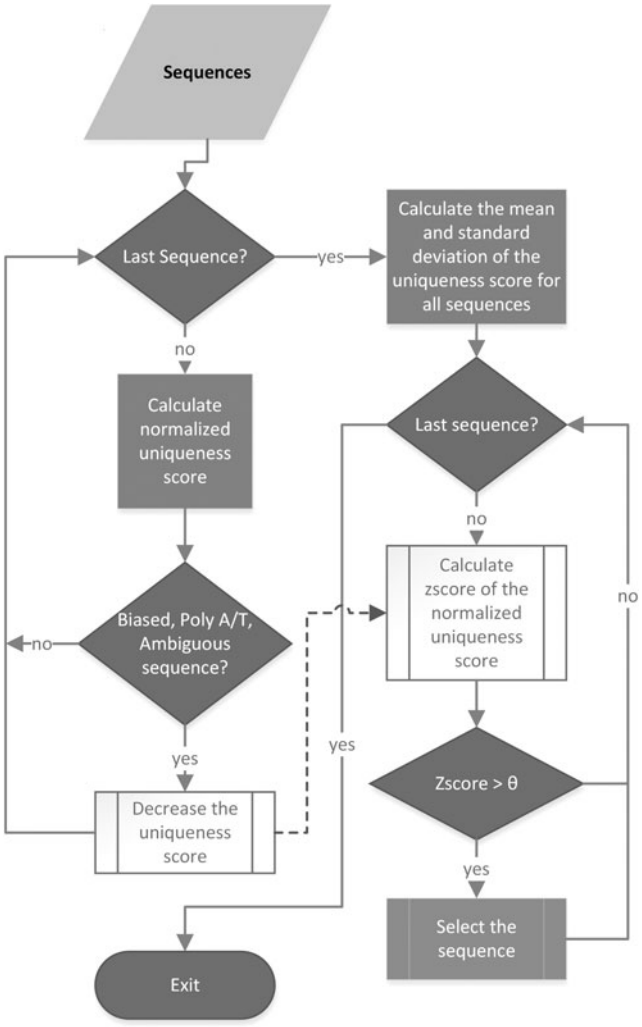

FIG. 1. Schematic representation of the process for filtering reads using the Zseq method.

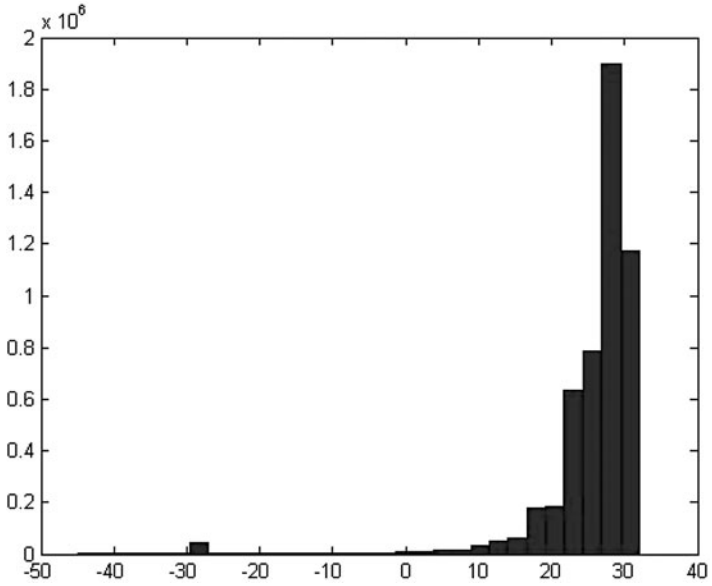

FIG. 2. Distribution of the normalized uniqueness scores for all reads in sample (SRR202054) ($\mu = 25.8169$, $\sigma = 7.1681$).

be significantly lower, because in the preparation of a standard library, a gel slice is used and heated up to 50°C, thereby increasing the bias of the GC content (Quail et al., 2008).

There are different techniques that try to remove those sequences with low-complex patterns from samples. Morgulis et al. (2006) presented the symmetric DUST method, which masks low-complex regions in a sequence to overcome context sensitivity in calculating the complexity score. Schmieder and Edwards (2011) proposed two methods to evaluate the sequence complexity. The first method is based on entropy as a measure. The second method, which is a variant of the DUST algorithm based on BLAST search, filters out the low-complex score sequences. Both methods consider each triplet of nucleotides as a word.

One of the downsides of the previous methods is that they focus only on the complexity of the sequences. This can be misleading in some cases due to the highly biased nature of the sequences. In this article, we propose a novel method called Zseq, which decreases the uniqueness score of highly biased regions, thereby filtering highly biased sequences and low-complex sequences.

2. METHODS

The *z*-score measurement has been used in different applications in bioinformatics (Cheadle et al., 2003; Margulies et al., 2005). Chopping sequence into *k*-mers is an essential technique in read assembly. We

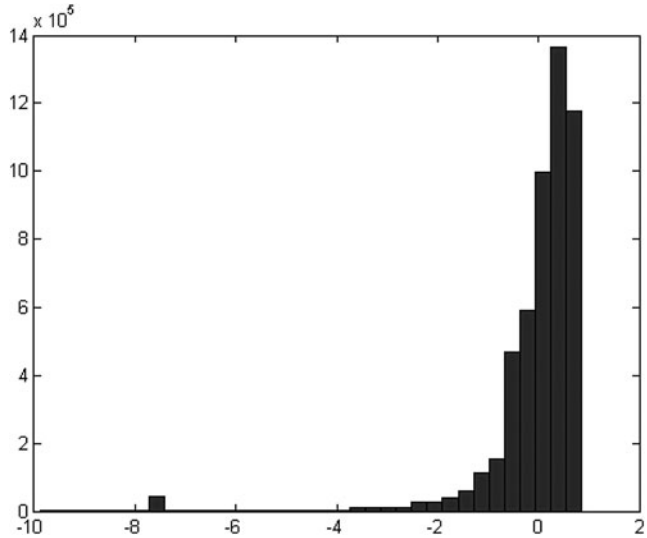

FIG. 3. Distribution of the *z*-scores of the normalized uniqueness scores corresponding to each read for sample (SRR202054).

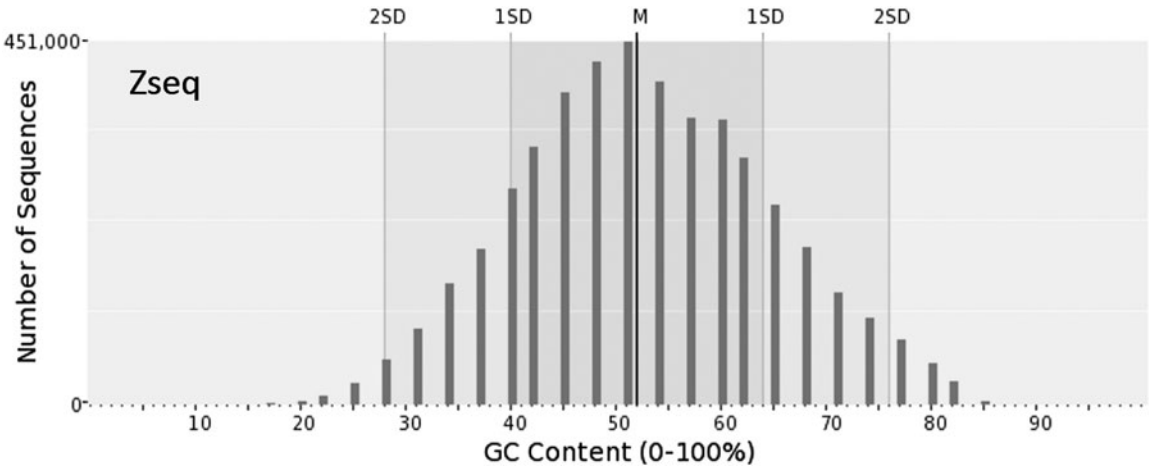

FIG. 4. Percentage of GC content for all filtered reads using the Zseq histogram with $\mu = 52.63\%$ and $\sigma = 12.08\%$.

present the Zseq algorithm that uses the z-score measurement based on uniqueness scores of all reads. The uniqueness score is the normalized number of unique k-mers in each read that takes low-complex regions into account. Figure 1 depicts the process of finding reads with improved quality. Each module is explained in detail in the next few paragraphs.

In the first step, Zseq scans all the reads and calculates the uniqueness score for all reads. The uniqueness score corresponding to each read is equal to the number of unique k-mers in that read. Zseq considers the default k-mer size, w, as 4-mers, which makes the vocabulary of four nucleotides (A,T,C,G) to be $4^4 = 256$ words. As the long reads may contain thousands of nucleotides, the 3-mer size is not sufficient to measure the complexity of the reads. This is because a 3-mer word can exist many times in the same read without being considered as unique, even when it is associated with different nucleotides each time. Zseq excludes the 5-mers of the low-complex/biased artifacts, such as ambiguous bases (N), PolyA/T, and GC content, from being unique by decreasing the unique score of the reads by one for each $2w$ to reduce the chances of selecting this sequence later. The uniqueness score of each read is then normalized by dividing it by the length of the read. The normalized uniqueness scores of all reads are stored in a vector with the same order of the read in the input file.

Figure 2 shows the distribution of the normalized uniqueness scores for all reads for sample SRR202054 from the prostate cancer data set used in the study of Kim et al. (2011). The x-axis shows the normalized uniqueness scores, while the y-axis shows the number of reads. As shown in the figure, the penalized sequences have a very small score down to −30. These are sequences that have been generated using reads that contain long PolyA/T sequences, very high GC content, or very high number of ambiguous nucleotides (N).

In the next step, Zseq calculates the mean and standard deviation for the normalized uniqueness scores. The mean of the normalized uniqueness scores of all reads is calculated in the first loop. The variance is

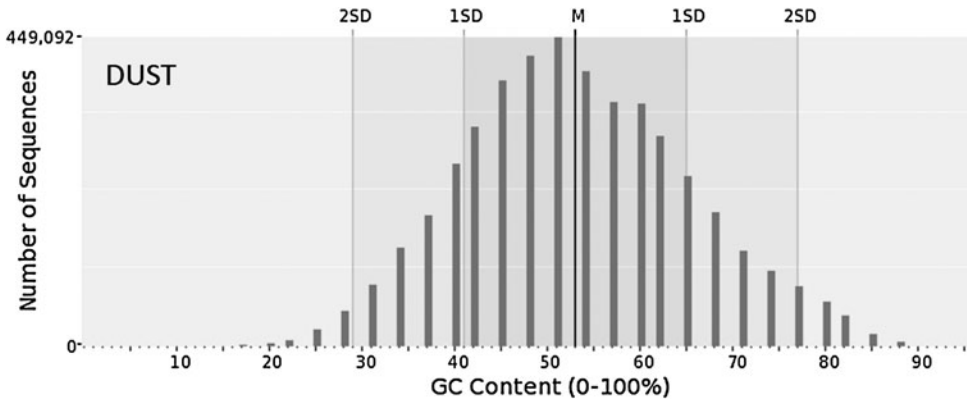

FIG. 5. Percentage of GC content for all filtered reads using the DUST histogram with $\mu = 53.09\%$ and $\sigma = 12.36\%$.

Table 1. Comparison of the Results of Applying Zseq on Samples from the Prostate Cancer Data Set as a Result of Applying DUST on the Same Samples

	Original			*Zseq*			*DUST*		
Sample number	*Occurrences of N*	*Mean GC content (%)*	*Mapping rate (%)*	*Occurrences of N*	*Mean GC content (%)*	*Mapping rate (%)*	*Occurrences of N*	*Mean GC content (%)*	*Mapping rate (%)*
SRR202054	40,690	52.82 ± 14.06	91.50	11,135	52.61 ± 12.20	93.00	19,177	52.89 ± 12.33	92.80
SRR202055	42,965	53.01 ± 13.74	91.20	9336	52.48 ± 12.10	92.40	19,470	52.91 ± 12.38	92.10
SRR202056	40,243	52.94 ± 13.99	91.40	10,721	52.67 ± 12.22	92.80	18,336	52.95 ± 12.36	92.60
SRR202057	42,630	52.94 ± 13.94	91.30	10,403	52.65 ± 12.22	92.60	20,018	52.93 ± 12.36	92.40
SRR202058	16,643	53.12 ± 14.03	91.00	14,023	52.63 ± 12.08	92.40	16,198	53.09 ± 12.36	92.30
SRR202059	17,741	52.56 ± 13.88	90.70	14,042	52.18 ± 12.02	92.00	17,091	52.61 ± 12.28	91.90
SRR202060	19,958	53.44 ± 13.98	90.90	13,775	53.23 ± 12.09	92.40	17,281	53.51 ± 12.21	92.30
SRR202061	2156	50.06 ± 11.50	77.00	1849	48.87 ± 9.96	79.20	2100	49.95 ± 11.12	77.90
SRR202062	5837	52.81 ± 13.64	69.10	5122	52.69 ± 11.77	71.30	5466	52.91 ± 11.84	71.30

also calculated linearly using a naive algorithm to reduce the cost of this step. The standard deviation is calculated from the variance of the vector of the normalized uniqueness scores.

Next, for each normalized uniqueness score, we calculate the z-score using the mean, μ, and the standard deviation, σ, as follows:

$$z = (s - \mu)/\sigma. \tag{1}$$

The z-score represents how many standard deviations the normalized uniqueness score of the read is away from the mean μ for all normalized uniqueness scores. In other words, if a read has a z-score of 0, it means that the read has the normalized uniqueness score of μ, while a z-score of value 1 means that the normalized uniqueness score is away exactly one standard deviation from the μ. Figure 3 shows the z-scores for all reads in the sample (SRR202054), where the x-axis is the z-score of the normalized uniqueness scores, while the y-axis indicates how many reads a particular z-score has in the sample.

Finally, the user-adjustable threshold θ is used to determine whether or not to select the reads, if the z-score of the normalized uniqueness score of the reads is greater than or equal to θ, the read will be selected; otherwise, it will be filtered out.

2.1. Estimating the cutoff point

A data-driven method based on the labeling rules is used to filter out the reads with low uniqueness score. The method automatically determines the cutoff point c to compensate θ in the histogram of reads uniqueness scores and removes those reads whose uniqueness score is less than c. The labeling rules model calculates the rst quartile q1 and third quartile q3 using mean and standard deviation, both of which are in the rst loop through the reads. The cutoff point is calculated as follows:

$$c = q1 - g(q3 - q1), \tag{2}$$

where g is the g-factor that can be calculated as follows:

$$g = (h - q1)/h, \tag{3}$$

with h being the highest value in the histogram of reads' uniqueness scores. After calculating the cutoff point c, the method sweeps again throughout the reads and selects those that have *uniquenessscore* $>= c$.

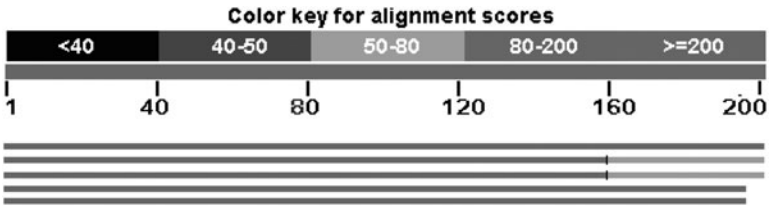

FIG. 6. Biologically meaningful human genomic sequences found using BLAST. *De novo* assembled transcripts using original reads.

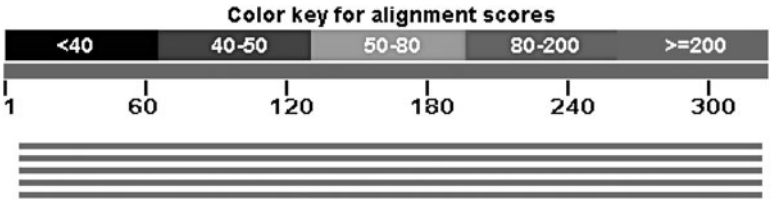

FIG. 7. Biologically meaningful human genomic sequences found using BLAST. *De novo* assembled transcripts using reads filtered by DUST.

3. RESULTS

In our experiments, we used the prostate cancer data set utilized in the study by Kim et al. (2011). The data set is publicly available in NCBI Gene Expression Omnibus (GEO) under Accession No. GSE29155. It contains 11 samples in total, where 7 of them belong to tumor tissues and the remaining 4 samples are benign. We measured the GC content and the number of ambiguous bases of the outcomes of each method, and then aligned the results of both methods to the human genome using Tophat2 as the alignment method (Kim et al., 2013).

DUST takes a value that ranges from 0 and 100 as the complexity threshold, while Zseq takes a z-score value as a complexity threshold, which shows how many standard deviations the normalized uniqueness score of the read is away from the mean. For the DUST method, we chose the value 5 as the threshold, which means that the value of the complexity of the read has to be greater than or equal to 5 to be selected; otherwise, DUST will ignore the read. For Zseq, we have chosen -1.5 as the value of the threshold, which makes the read good to be selected if the z-score of that read is greater than or equal to -1.5. The reason behind selecting these two thresholds is that both methods filter almost the same number of reads in each sample. The filtered reads using Zseq have less GC content than the filtered reads using DUST. It also has smaller standard deviation, which makes the reads centered more around the mean than DUST. Figures 4 and 5 show the GC-content distributions for both methods applied on the same sample set (SRR202058).

Zseq shows a slight improvement in reducing the GC content, mapping rate, and mapping time, while dropping the number of ambiguous bases drastically in comparison with DUST. Table 1 shows that the number of ambiguous bases, N, in the filtered reads using Zseq has drastically decreased compared with the ambiguous bases that have been filtered out using DUST in all samples. For example, the number of occurrences of N in sample SRR202054 for filtered reads by DUST is 19,177, while there are only 11,135 filtered reads using Zseq for the same sample. The results indicate that Zseq slightly shrunk the GC-content percentage distribution and reduced the mean of the GC-content percentage. For sample SRR202055, the mean of the GC content is $52.48\% \pm 12.10\%$ using Zseq, which is less than the $52.91\% \pm 12.38\%$ obtained

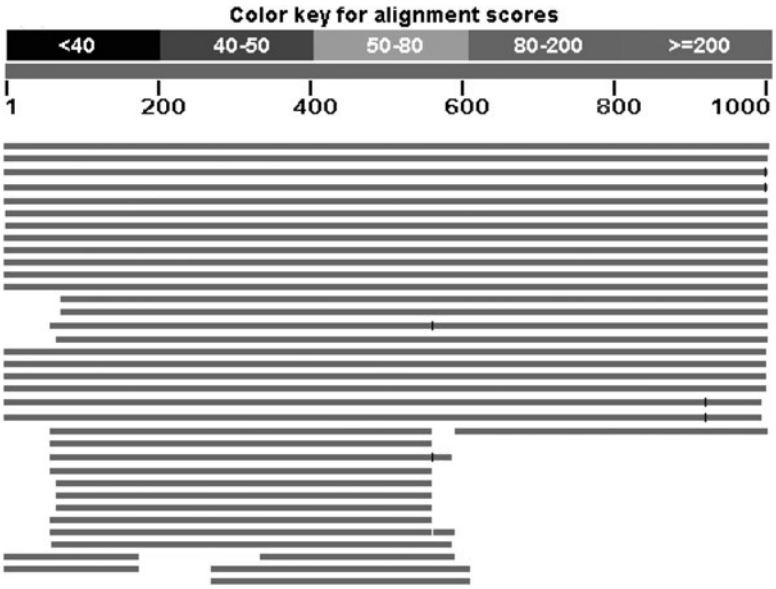

FIG. 8. Biologically meaningful human genomic sequences found using BLAST. *De novo* assembled transcripts using reads filtered by Zseq.

Table 2. Average Mapping Rate of Transcripts Using the Data Set Generated by the Original Reads, Reads Filtered by DUST, and Reads Filtered by Zseq

Original	*DUST*	*Zseq*
88.90%	90.10%	90.40%

using the DUST method. Zseq also shows better mapping alignment for the filtered reads than DUST for most of the samples. For example, in sample SRR202061, the reads filtered by Zseq have 79.20% mapping rate, which is greater than 77.90% mapping rate for reads filtered by DUST, the only exception is sample SRR202062, which shows a similar mapping rate of 71.30% for both DUST and Zseq.

3.1. De novo sequence validation

Using Trinity *de novo* assembler (Grabherr et al., 2011), transcripts have been reconstructed for the original reads of sample SRR202058, reads that have been filtered by DUST and reads that have been filtered by Zseq. In the next step, all three sets of constructed transcripts were evaluated by searching the assembled transcripts with the human genome sequences using BLAST (Altschul et al., 1997). The set of the reconstructed transcript using the filtered reads by Zseq contains a higher number of long sequences in comparison with the other two sets. Figures 6, 7, and 8 show the meaningful sequences for each set. Some of the sequences, which were built using the reads filtered by Zseq, have a length of 1000 bp or more along with a high alignment score, while the sequence length is slightly more than 300 bp using the reads filtered by DUST and 200 bp for the original reads without filtering.

3.2. Machine learning validation

In another experiment, we used an independent data set containing 12 samples (six tumors and six matched normal) (Kannan et al., 2011). Using these samples, three data sets were generated, one from the original reads, one by applying DUST on the reads, and the third one by applying Zseq on the reads for all samples. In the next step, all reads corresponding to each data set have been aligned to human genome hg19 using Tophat2 (Kim et al., 2013) and Cufflinks assembler (Trapnell et al., 2012) with default parameters to assemble the transcripts to the human genome and estimate their abundance, which is measured by FPKM value (fragments per kilo bases of exons for per million mapped reads). Table 2 shows the average mapping rate of reads filtered by each method.

Each generated data set using filtered reads has 43,497 features (transcripts) with FPKM values. Also, each of the 12 samples was labeled as *cancer* or *matched benign*. The FPKM value equals 0 if the transcript has not been presented in that sample. We measured the number of transcripts that can individually separate all cancer samples from normal samples perfectly, with 100% accuracy. In other words, we want to compute the number of transcripts generated using filtered reads by each method, in such a way that the FPKM values corresponding to cancer samples can be separated from those of FPKM of normal samples. Figure 9 depicts two transcripts; transcript *a* has clearly separable FPKM values, while in transcript *b*, the FPKM values cannot be separated accurately.

Table 3 shows the number of transcripts that contain separable FPKM values. These results indicate that applying Zseq influences the alignment tool and assembler to quantify more meaningful transcripts that can discriminate cancer and normal samples in comparison with the DUST method and original reads.

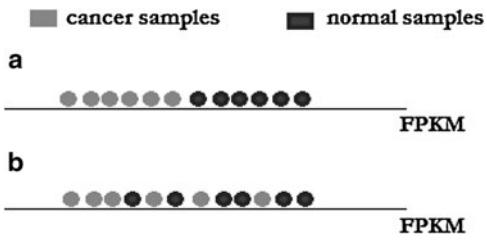

FIG. 9. An example of two transcripts, one with separable FPKM values **(a)**, and other transcript with inseparable FPKM values **(b)**.

Table 3. The Number of Discriminative Transcripts For Each of the Three Data Sets

Data set	*No. of discriminative transcripts*
Original	167
Filtered by DUST	159
Filtered by Zseq	231

Moreover, using chi2 (Liu and Setiono, 1995) statistical test on the 231 discriminative transcripts from Zseq data set, the NM_001145410 transcript corresponding to NONO gene was the most significant transcript among all other transcripts in all three data sets. NONO is known to regulate in different types of cancers such as breast and prostate cancer (Traish et al., 1997; Ishiguro et al., 2003). Next, a support vector machine (SVM) with linear kernel was applied on the three data sets using this transcript as feature. SVM is a supervised learning machine that tries to find an optimal separating hyperplane between classes (Cortes and Vapnik, 1995). Using a leave-two-out cross-validation scheme, the classification returns 100% accuracy for the Zseq data set, 91.66% for the DUST data set, while it was down to 83.33% in the original read data set.

3.3. Result of estimated cutoff point Zseq

Result of estimated cutoff point Zseq as shown in Tables 4 and 5 suggested that the method does not find the optimal point. The result of Zseq on the prostate cancer data set using the threshold $\theta = -1.5$ in the previous section outperformed the result of the EC-Zseq. Despite having a better mapping rate, EC-Zseq falls short in mean GC content to Zseq with θ, in a number of ambiguous nucleotide measurements comparing to DUST and Zseq with θ, and in a number of decisive transcripts comparing to Zseq with θ. However, EC-Zseq still shows a better result than the original data set or preprocessing the data set using the DUST method.

4. CONCLUSION

We have presented a novel method for filtering the reads that reduce the number of biased, duplicate, or ambiguous sequences. Our method finds the complexity of the sequences by assigning a unique score to each read. Using a user-defined threshold, the user can filter the reads with a score less than the threshold. Applying the proposed method on real samples shows that the Zseq algorithm is statistically sound and provides a better mapping rate, while it significantly reduces the number of ambiguous bases in comparison

Table 4. Some Artifact Measurements of Prostate Cancer Samples That Were Preprocessed By Ec-Zseq

	EC-Zseq		
Sample number	*Occurrences of N*	*Mean GC content (%)*	*Mapping rate (%)*
SRR202054	33,124	52.71 ± 13.40	93.40
SRR202055	27,890	52.91 ± 12.76	93.30
SRR202056	34,453	52.82 ± 13.07	93.50
SRR202057	30,321	52.68 ± 12.52	93.40
SRR202058	14,760	52.87 ± 13.43	92.90
SRR202059	15,203	52.18 ± 12.62	92.80
SRR202060	16,704	53.31 ± 12.09	92.70
SRR202061	1926	49.11 ± 10.62	79.70
SRR202062	5484	532.70 ± 12.47	72.10

Table 5. The Number of Decisive Transcripts for the Data Set That Was Preprocessed By Ec-Zseq

Data set	*No. of decisive transcripts*
Preprocessed by EC-Zseq	222

with other state-of-the-art methods. Estimating the cutoff point using Labeling rules shows a good result. However, it is not the optimal. The Zseq method is publicly available and can be accessed using the following link: http://sourceforge.net/projects/zseq.

ACKNOWLEDGMENT

This work has been partially supported by NSERC, the Natural Science and Engineering Research Council of Canada.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Brown, T. *Introduction to Genetics: A Molecular Approach*. Garland Science, 2012. ISBN 9780815365099. Available at: URL http://books.google.ca/books?id=TsvKPQAACAAJ. Last viewed on Jan. 20, 2017.

Cheadle, C., Vawter, M.P., Freed, W.J., et al. 2003. Analysis of microarray data using z score transformation. *J. Mol. Diagn*. 5, 73–81.

Chen, Y.-C., Liu, T., Yu, C.-H., et al. 2013. Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS One* 8, e62856.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learn*. 20, 273–297.

Grabherr, M.G., Haas, B.J., Yassour, M., et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. 29, 644–652.

Ishiguro, H., Uemura, H., Fujinami, K., et al. 2003. 55 kDa nuclear matrix protein (nmt55) mRNA is expressed in human prostate cancer tissue and is associated with the androgen receptor. *Int. J. Cancer*. 105, 26–32.

Kannan, K., Wang, L., Wang, J., et al. 2011. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl Acad. Sci. U. S. A.* 108, 9172–9177.

Kim, D., Pertea, G., Trapnell, C., et al. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14, R36.

Kim, J.H., Dhanasekaran, S.M., Prensner, J.R., et al. 2011. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Res*. 21, 1028–1041.

Lavezzo, E., Barzon, L., Toppo, S., et al. 2016. Third generation sequencing technologies applied to diagnostic microbiology: Benefits and challenges in applications and data analysis. *Expert Rev. Mol. Diagn*. 16, 1011–1023.

Liu, H., and Setiono, R. 1995. Chi2: Feature selection and discretization of numeric attributes. Presented at 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, IEEE Computer Society, Herndon, VA, USA. pp. 388–388.

Mackinnon, M.J., Li, J., Mok, S., et al. 2009. Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. *PLoS Pathog*. 5, e1000644.

Margulies, M., Egholm, M., Altman, W.E., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

Morgulis, A., Gertz, E.M., Schäffer, A.A., et al. 2006. A fast and symmetric dust implementation to mask low-complexity DNA sequences. *J. Comput. Biol*. 13, 1028–1040.

Pozzoli, U., Menozzi, G., Fumagalli, M., et al. 2008. Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol. Biol*. 8, 99.

Quail, M.A., Kozarewa, I., Smith, F., et al. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*. 5, 1005–1010.

Schmieder, R., and Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864.

Traish, A.M., Huang, Y.-H., Ashba, J., et al. 1997. Loss of expression of a 55 kDa nuclear protein (nmt55) in estrogen receptor-negative human breast cancer. *Diagn. Mol. Pathol*. 6, 209–221.

Trapnell, C., Roberts, A., Goff, L., et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc*. 7, 562–578.

Vogel, F. 1997. *Vogel and Motulsky's Human Genetics: Problems and Approaches*, Volume 878. Springer: London, New York.

Waszak, S.M., Kilpinen, H., Gschwind, A.R., et al. 2014. Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics* 30, 165–171.

Wuitschick, J., and Karrer, K. 1999. Analysis of genomic G+C content, codon usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*. *J. Eukaryot. Microbiol*. 46, 239–247.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M.D. 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*. 34, 564–574.

Zhao, Z., Wu, X.,Kumar, P.K., et al. 2014. Bioinformatics analysis of alternative polyadenylation in green alga *Chlamydomonas reinhardtii* using transcriptome sequences from three different sequencing platforms. *G3* 4, 871–883.

Address correspondence to:
Prof. Luis Rueda
School of Computer Science
University of Windsor
401 Sunset Avenue
Windsor ON N9B 3P4
Canada

E-mail: alkhate@uwindsor.ca