HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2016 August 14.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2015 ; 2015: 6461–6464. doi:10.1109/EMBC.2015.7319872.

Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data

Ying Sha, **John H. Phan [IEEE Member]**, and **May D. Wang [IEEE Member]**

Abstract

We compare methods for filtering RNA-seq lowexpression genes and investigate the effect of filtering on detection of differentially expressed genes (DEGs). Although RNA-seq technology has improved the dynamic range of gene expression quantification, low-expression genes may be indistinguishable from sampling noise. The presence of noisy, low-expression genes can decrease the sensitivity of detecting DEGs. Thus, identification and filtering of these low-expression genes may improve DEG detection sensitivity. Using the SEQC benchmark dataset, we investigate the effect of different filtering methods on DEG detection sensitivity. Moreover, we investigate the effect of RNA-seq pipelines on optimal filtering thresholds. Results indicate that the filtering threshold that maximizes the total number of DEGs closely corresponds to the threshold that maximizes DEG detection sensitivity. Transcriptome reference annotation, expression quantification method, and DEG detection method are statistically significant RNA-seq pipeline factors that affect the optimal filtering threshold.

Section I

Introduction

RNA-seq enables quantitative profiling of gene expression with a high dynamic range. However, accurate quantification of expression still remains challenging. RNA-seq measurement errors are a direct result of the inherent random sampling process [1]. This measurement noise is more severe in low-expression genes. The filtering of low-expression genes is a common practice not only because it can increase our confidence in discovered differentially expressed genes (DEGs), but also because it can increase the number of total DEGs in one experiment [2].

DEG identification tools commonly suggest filtering of low-expression genes that have average counts below an empirical threshold [3] [4] [5] [6]. Other methods for determining filtering thresholds include estimation based on the distribution of intergenic reads per kilobase per million mapped reads (RPKM) values [7] and estimation based on External RNA Controls Consortium (ERCC) spike-in controls [8]. However, there is no systematic comparison of different filtering methods, and no investigation of how different RNA-seq pipelines affect these filtering methods. We specifically address these issues by evaluating 8 low-expression gene filtering methods using 48 RNA-seq pipelines to identify DEGs. We use the SEQC RNA-seq and qPCR datasets as benchmarks to measure true DEG detection [9]. This comprehensive investigation of different low-expression gene filtering methods and RNA-seq analysis pipelines can serve as a guide for researchers performing DEG identification with RNA-seq data.

Section II

Methods

A. RNA-SEQ and QPCR Data—We use the Sequencing Quality Control (SEQC) consortium RNA-seq and qPCR datasets [9]. Specifically, we use four replicates with eight lanes per replicate of samples A and B of the SEQC dataset that were sequenced at the Beijing Genomics Institute to identify differentially expressed genes. Samples A and B consist of the Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR), respectively. These samples also include ERCC spikeins. The transcriptome-wide qPCR dataset includes 20,801 genes measured from the same samples A and B.

B. RNA-SEQ Pipelines for Identification of Differentially Expressed Genes—A typical bioinformatics workflow for identifying DEGs in RNA-seq data includes mapping short reads, quantification of gene expression, normalization, and DEG identification. We map reads to either the Refseq or Ensembl transcriptome annotation using Tophat2 [10], Mapsplice [11], and Subread [12]. We use HTSeq [13] and featureCounts [14] for quantification of gene expression; and edgeR [3], DESeq/DESeq2 [4], [5] and Voom (limma) [6] for DEG identification. We do not treat normalization as an independent step since we use built-in normalization methods for each DEG identification package. In total, we implement 48 RNA-seq analysis pipelines for this study (Fig. 1).

C. Thresholding Methods for Filtering Low-Expression Genes—We investigate various low-expression gene filtering methods. Each filtering method is based on a statistic calculated from gene expression. We calculate the minimum (min), maximum (max), average (mean), and variance (sd) of raw read counts for each gene from the replicated samples. We also calculate average counts per million (CPM) and average RPKM. CPM is defined as read counts scaled by the number of sequenced fragments times one million [3]. We estimate gene length for RPKM [15] as the sum of the lengths of all of the gene's exons. CPM is equivalent to RPKM without length normalization.

In addition, we investigate other methods for determining filtering thresholds, including percentiles of intergenic distribution [7] and LODR (limit of detection ratio) introduced in erecdashboard [8]. LODR is derived from the analysis of external spike-in RNA control ratio mixtures, and is defined as the minimum count above which a gene with an absolute log fold-change signal has a 100% chance of obtaining a statistically significant adjusted p-value [8].

To facilitate comparison of different filtering methods, we transform the real values of various filtering thresholds into percentile-based thresholds. We also transform percentiles of the intergenic distribution into percentiles of RPKM, and transform LODR estimates into percentiles of average gene counts. We use the Benjamini-Hochberg method [16] after filtering genes to control the false discovery rate.

D. Evaluating Deg Detection Performance—We use the transcriptome-wide qPCR [9] data as a ground-truth to estimate DEG detection true positive rate (TPR) and positive predictive value (PPV) for each RNA-seq pipeline and each filtering method. qPCR genes

with a four-fold difference are viewed as true DEGs. We use four-fold instead of the twofold advocated in the MAQC study [17] to improve stringency. 3,974 out of 11,714 genes are DEGs. The total number of qPCR genes, 11,714, include only genes that exist in both the Ensembl and Refseq annotations. Filtering methods that achieve both high TPR and high PPV are regarded as good.

Section III

Results and Discussion

A. Filtering Increases Sensitivity and Precision of Deg Detection—The number of DEGs detected increases after filtering up to 20% of low-expression genes. However, the number of DEGs decreases if the filtering threshold is increased beyond 30%. As shown in Fig. 2A, by removing 15% of genes with lowest average read count, we are able to identify the maximum number of DEGs, which is 480 more DEGs than without filtering. We use Tophat2 with Refseq transcriptome reference, HTSeq quantification (intersection-strict), and edgeR DEG detection method to generate Fig. 2. This result agrees with the findings in [2]. However, those findings did not indicate whether newly discovered DEGs were “true” DEGs.

We use the transcriptome-wide qPCR dataset as a ground-truth to investigate the effect of filtering on DEG detection true positive rate and precision. Fig. 2B shows that the sensitivity, or true positive rate, of DEG detection also increases with appropriate filtering. Moreover, the precision of DEG detection increases with increasing low-expression gene filtering threshold (Fig. 2C).

B. Choosing an Optimal Filtering Threshold—Choosing an optimal filtering threshold consists of two steps: (1) choose an optimal filtering method, and (2) select an optimal threshold value. An important quality of an ideal filtering method is that it should be specific, i.e., genes with lower ranking of filter statistic should tend to be non-DEGs. For this reason, the minimum read count of a gene across samples is not a good filtering statistic. For example, using the minimum read count method, a gene that is significantly differentially expressed might be filtered because it is not expressed under one condition. LODR may not be an ideal filtering method since it is too strict and filters many true DEGs. Instead, LODR should only be used to determine if the sequencing depth is adequate for detecting genes of interest. Among the remaining filtering methods, the average read count could be considered ideal because it results in the highest F 1 score (i.e., combination of sensitivity and precision) compared with other methods while filtering less than 20% of genes.

Considering an optimal threshold for average gene counts, the percentile corresponding to the maximum TPR of DEG identification would be a good choice since it also corresponds to a reasonably high precision of DEG detection. However, in real applications, there will not be a validation dataset for us to compute both sensitivity and precision. Fortunately, as shown in Fig. 3B, the threshold chosen based on maximal number of DEGs and the threshold chosen based on maximal true positive rate are correlated. Thus, in applications without a ground-truth for DEGs, we can choose a filtering threshold that maximizes the number of discovered DEGs.

C. Optimal Filtering Thresholds Vary with RNA-SEQ Pipeline—We observe that the choice of RNA-seq analysis pipeline has a considerable impact on optimal filtering threshold. To further investigate which components of the workflow affect the thresholding values, we conduct analysis of variance (ANOV A). The choice of sequence mapping tool does not significantly affect the optimal filtering threshold (Table I). Among the rest of the analysis components, the choice of transcriptome annotation has the most significant effect on the specific thresholding values. The choice of DEG identification tool also has significant influence on specific thresholding values. Thus, there is no fixed thresholding value that can be applied to any specific analysis pipeline. We recommend to determine the optimal filtering threshold for each RNA-seq pipeline by maximizing the number of detected DEGs.

D. Intergenic Mapping Distribution May Not be a Good Filtering Method— Intergenic mapping distributions may be used to quantify RNA-seq experimental noise [7]. We quantify RNA-seq "noise" expression by aligning reads to intergenic regions, which we define as regions complementary to genic regions (including two flanking sequences of lkb at both ends) annotated by Aceview. We use the Aceview annotation because it is considered to be less conservative than other annotations such as Refseq [18]. However, this method highly depends on the completeness of genome annotation. For example, some intergenic regions may contain undiscovered genes that can inflate estimates of expression "noise".

As shown in Fig. 4, the distribution of exonic mapping is highly variable and depends on both the reference annotation (Fig. 4A and 4B) and RNA-seq pipeline (Fig. 4C and 4D). Thus, choosing filtering thresholds based on intergenic and exonic mapping distributions may also be highly variable. To avoid this variance, we suggest to avoid using intergenic mapping distributions to choose low-expression gene filtering thresholds for the purposes of DEG detection.

Section IV

Conclusion

Although we conducted limited experiments, we observed that filtering low-expression genes improved DEG detection sensitivity. A previous study demonstrated that independent filtering would increase the detection power of high throughput experiments [2]. We not only confirmed their results, but also used a transcriptome-wide qPCR dataset to validate that the sensitivity and the precision of DEGs also improved after filtering. In addition, we investigated the effect of pipeline choice on optimal filtering thresholds in order to establish a guideline. The results showed that there was no universal optimal filtering thresholds for all pipelines. However, from an empirical point of view, the optimal filtering threshold corresponds to a threshold that maximizes the number of DEGs.

Acknowledgement

The authors thank Po-Yen Wu and Dr. James Cheng for assisting in manuscript preparation.

References

[1]. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009; 10(1):57–63. [PubMed: 19015660]

[2]. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. U. S. A. 2010; 107:9546–9551. [PubMed: 20460310]

[3]. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. Jan.2010 26(1):139–40. [PubMed: 19910308]

[4]. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2012; 11(10):R106. [PubMed: 20979621]

[5]. Love MI, Anders S, Huber W. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. [PubMed: 25516281]

[6]. Law C, Chen Y, Shi W, Smyth G. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15(2):R29. [PubMed: 24485249]

[7]. Harati, S.; Phan, J.; Wang, M. Eng. Med. 2014. Investigation of factors affecting RNA-seq gene expression calls.

[8]. Munro, S. a; Lund, SP.; Pine, PS.; Binder, H.; Clevert, D-A.; Conesa, A.; Dopazo, J.; Fasold, M.; Hochreiter, S.; Hong, H.; Jafari, N.; Kreil, DP.; Labaj, PP.; Li, S.; Liao, Y.; Lin, SM.; Meehan, J.; Mason, CE.; Santoyo-Lopez, J.; Setterquist, R. a; Shi, L.; Shi, W.; Smyth, GK.; Stralis-Pavese, N.; Su, Z.; Tong, W.; Wang, C.; Wang, J.; Xu, J.; Ye, Z.; Yang, Y.; Yu, Y.; Salit, M. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat. Commun. Jan.2014 5:5125. [PubMed: 25254650]

[9]. SEQC. Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 2014

[10]. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. Apr.2013 14(4):R36. [PubMed: 23618408]

[11]. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm S. a, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. Oct.2010 38(18):e178. [PubMed: 20802226]

[12]. Liao Y, Smyth G, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013

[13]. Anders, S.; Pyl, PT.; Huber, W. HTSeq – A Python framework to work with high-throughput sequencing data HTSeq – A Python framework to work with high-throughput sequencing data. 2014. p. 0-5.

[14]. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–930. [PubMed: 24227677]

[15]. Mortazavi A, Williams B, McCue K. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008; 5(7):621–628. [PubMed: 18516045]

[16]. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B. 1995:289–300.

[17]. MAQC. Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.,". Nat. Biotechnol. 2006; 24(9): 1151–1161. [PubMed: 16964229]

[18]. Wu, P.; Phan, J.; Wang, M. BMC Bioinformatics. 2013. Assessing the impact of human genome annotation choice on RNA-seq expression estimates.

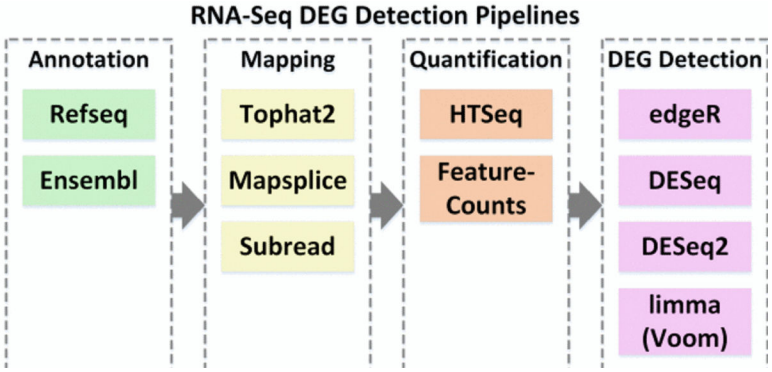

Figure 1.

RNA-seq DEG detection pipelines consist of transcriptome annotation, sequence mapping, expression quantification, and DEG detection method.

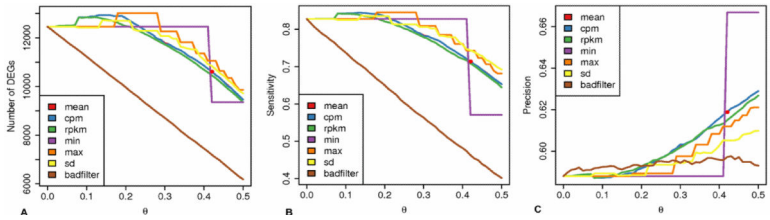

Figure 2.

Quantitative assessment of low-expression gene filtering methods. Using the Refseq-Tophat2-HTSeq-edgeR pipeline, we calculated (A) the number of DEGs, (B) the true positive rate (recall rate or sensitivity), and (C) the precision at FDR=0.1 as a function of filtering threshold, θ (percent of genes filtered), for different filtering methods. The red dots on the three graphs represent the percentile thresholds transformed from the LODR estimate.

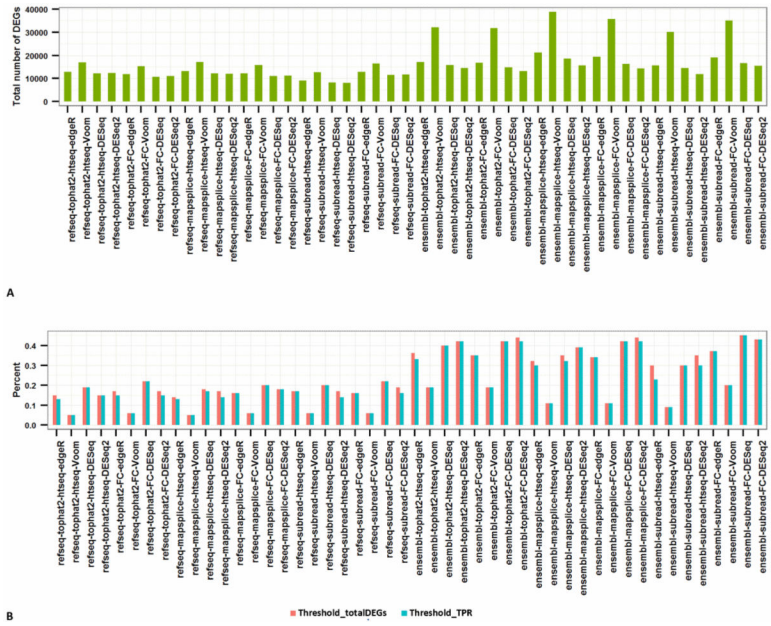

Figure 3.

(A) Total number of DEGs discovered using different pipelines. (B) Optimal filtering thresholds based on percentile of average count for different pipelines determined by maximum total number of DEGs and maximum TPR.

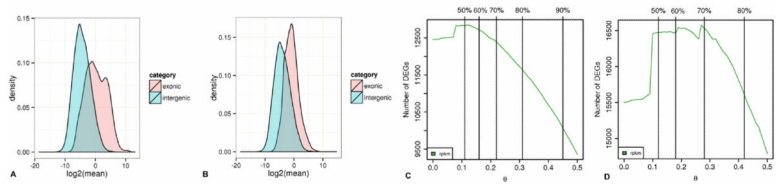

Figure 4.

Distribution of intergenic and exonic RNA-seq mapping. (A) Exonic and intergenic mapping based on the Refseq annotation. (B) Exonic and intergenic mapping based on the Ensembl annotation. (C) Thresholding determined by quantiles of intergenic mapping. The pipeline used was refseq-tophat2-htseq. (D) Thresholding determined by quantiles of intergenic mapping. The pipeline used was ensembl-tophat2-htseq.

Table I

Analysis of Variance in Optimal Filtering Threshold

Variable	Df	Sum Sq	F value	Pr(>F)
Annotation	1	0.369	247.509	**9.94E-19**
Mapping	2	0.003	1.049	0.3598151
Quantification	1	0.012	7.855	**0.0077769**
DEG Detection	3	0.296	66.141	**1.47E-15**
Residual	40	0.060	NA	NA