International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 2, No. 6 (2013): 285-292
Research Article
Open Access

ISSN: 2320-0634

Optimal Calculation of RNA-Seq Fold-Change Values

Charles D. Warden[1]*, Yate-Ching Yuan[2], and Xiwei Wu[1]

[1]*Integrative Genomics Core, Department of Molecular and Cellular Biology, City of Hope National Medical Center, Duarte, California*
[2]*Bioinformatics Core, Department of Molecular Medicine, City of Hope National Medical Center, Duarte, California*

*Corresponding author: Charles D. Warden; e-mail: cwarden@coh.org

Received: 25 October 2013 Accepted: 05 November 2013 Online: 15 November 2013

ABSTRACT

Biologists often use RNA-Sequencing (RNA-Seq) to identify a limited number of genes for subsequent validation, and one important factor for candidate gene selection is the fold-change in expression between two groups. However, RNA-Seq produces a wide range of read counts per gene, and genes with a low coverage of reads can produce artificially high fold-change values. In this paper, we present a solution to this problem: adding a factor between 0.01 and 1 to normalized expression values. This conclusion is based upon analysis of a large patient cohort of paired tumor and normal samples from patients with lung adenocarcinomas as well as a small, two-group cell line dataset. The optimal factor to add to normalized expression values is chosen based upon testing a range of factors on: the number of genes or transcripts whose expression is effectively censored (using three different alignment algorithms) and 2) the potential level of bias introduced by the factor (defined by comparing unadjusted gene lists). The robustness of these trends is also tested by comparing multiple mRNA quantification and differential expression algorithms. The relationship between RPKM cutoff and concordance between gene lists produced using different statistical methods can be complicated, but this study emphasizes that simple statistical analysis (amendable to the use of rounded RPKM values) at least provides equal quality results as popular algorithms for RNA-Seq differential expression.

Keywords: DEG = Differentially Expressed Gene; RNA-Seq = RNA-Sequencing; RPKM = Reads Per Kilobase per Million.

INTRODUCTION

RNA-Sequencing (RNA-Seq) is a powerful tool for quantifying gene expression as well as identifying alterative splicing and RNA-editing events [1-3]. There are a number of tools to calculate differentially expressed gene from RNA-Seq experiments [4, 5], many of which utilize RPKM (Reads Per Kilobase per Million [6]) normalized count values for gene and transcript abundance [7]. Biologists often use fold-change (the ratio of average expression between two groups) to prioritize genes of interest [8]. However, RNA-Seq experiments often have many genes with very low coverage with RPKM values that are very close to 0. So, the fold-change between two conditions can be high, even though the true expression levels may not differ greatly.

One solution to this problem is to define a minimum reliable threshold for RPKM values. If the test for differential expression utilizes a table of RPKM expression values, then adding that threshold to the exiting RPKM value will effectively censor all RPKM threshold below that value (whose values will now be rounded to that threshold). Many statistical tests (such as ANOVA) assume a normal distribution, which can be approximated using $\log_2$ RPKM expression values: in this case, $\log_2$(RPKM + cutoff) expression values would be used. However, the optimal threshold to choose is not obvious, and it is not clear how robust the use of a single threshold will be for various experiments. In order to provide a recommended range for RPKM thresholds, a large patient dataset with 134 samples of paired tumor and normal patients was used to compare gene lists defined using various RPKM cutoffs [9]. In order to test the robustness of observed trends, gene lists were produced using three aligners (TopHat [7], STAR [10], and Novoalign). Additionally, the TopHat alignment was tested using mRNA quantification tools

(Partek EM-Algorithm, cufflinks [7]) as well as four methods to define differentially expressed genes (ANOVA, cuffdiff [7], edgeR [4], and DESeq [5]). To be clear, edgeR and DESeq require the user to provide unnormalized read counts in order to estimate dispersion frequencies, so the goal of comparing different statistical tools was to compare genes defined using 2-way ANOVA with $\log_2$(RPKM + cutoff) values and other popular methods where a cutoff couldn't be implemented in this way. Likewise, cuffdiff does not simply use raw or normalized counts, so it is not possible to round RPKM (or, more precisely, FPKM) values for cuffdiff analysis. Cell line analysis was performed in addition to the large patient cohort comparison in order to test the robustness of the comparison of statistical analysis methods (and the impact of RPKM cutoff choice on the concordance between methods).

The results of these comparisons indicate that the optimal RPKM cutoff for most experiments should be between 0.01 and 1. Higher cutoffs should be used for more conservative gene lists and lower cutoffs should be used for larger gene lists. We find that the range of RPKM expression values is very similar regardless of alignment method, and the number of reads per experiment has the least influence on RPKM values if a RPKM cutoff of 0.1 is used. For the lung cancer cohort, using 2-way ANOVA (or 1-way ANOVA) on $\log_2$(RPKM + cutoff) expression values produces a gene list that is robust and meets or exceeds the proportion of overlapping genes for raw count based methods for RNA-Seq differential expression. The robustness of this strategy is further confirmed using cell line data with a simple study design (using 1-way ANOVA), where concordance of gene lists using different differential expression tools is less sensitive to the RPKM cutoff (used in Partek and sRAP, a novel open-source tool developed to implement the strategies discussed in this study).

MATERIALS AND METHODS

Alignment

For the lung cancer dataset, paired-end fastq files were downloaded from the Sequence Read Archive (ERP001058) [11]. Single-end alignments were performed using only the forward reads for each sample. TopHat (v.1.2.0, [7]) single-end alignments were produced using default parameters. TopHat paired-end alignments are not presented in this report because the run-time was unreasonably long. STAR (v.2.3.0, [10]) single-end and paired in alignments using default settings. Novoalign (v.2.07.05, http://www.novocraft.com) single-end and paired-end alignments were performed with a minimum information content of 16 and random assignment of ambiguous reads. As a technical note, Novocraft currently provides a different set of parameters for optimal RNA-Seq paired-end alignment, which might further improve performance; however, the parameters used in this study still show reasonably good concordance to the TopHat and STAR alignments (for the paired-end alignment). In all cases, reads were aligned to hg19.

For the cufflinks cell line dataset, paired-end fastq files were also downloaded from the SRA (for project SRP012607). Paired end alignments to hg19 were performed using TopHat (v.2.0.8 [12]). Both MiSeq and HiSeq datasets are analyzed and compared in this paper, but all subsequent analysis focuses specifically on the MiSeq samples.

mRNA Quantification

Unless otherwise specified, RPKM expression levels were calculated using Partek Genomics Suite™ (Partek, Inc., St. Louis, MO; verison 6.6). Transcript-based expression levels were calculated using an expectation-maximization algorithm similar to that implemented in Xing et al. [13]. A cutoff (of 0.01, 0.1, or 1) was then added to the RPKM expression values and then expression values underwent a $\log_2$ transformation. Baseline RPKM values were not analyzed on a $\log_2$ scale because genes with RPKM expression values of 0 would be undefined. Cufflinks (v.0.9.3, [7]) mRNA quantification of TopHat single-end reads was calculated using default settings. In both cases, RefSeq gene coordinates were used to define gene/transcript boundaries [14].

Differential Expression

For patient samples, differential expression was calculated using 2-way ANOVA using Partek Genomics Suite™ (Partek, Inc., St. Louis, MO; verison 6.6), where the two factors considered for analysis was group (tumor or normal) and patient ID (pairing tumor and normal samples, when both were available). Genes and transcripts were defined as differentially expressed if they showed a |fold-change| > 1.5 and false discovery rate (FDR) < 0.05. FDR values were calculated using the method of Benjamini and Hochberg [15] from the distribution of 2-way ANOVA p-values, and fold-change values were calculated on a linear scale using least-squares mean.

DESeq [5] and edgeR [4] provide two different strategies for differential expression: 1-factor negative binomial test and multi-factor generalized linear regression (still based upon a negative binomial distribution). For multi-factor analysis, both tumor and normal samples from the patients needed to be available for analysis, reducing the sample size from 134 to 100, and DESeq required the dispersion values to be calculated as if all replicates were for a single condition. These restrictions were not necessary for the 1-factor (e.g. tumor vs. normal) analysis. In both cases, genes were defined as differentially expressed with an FDR < 0.05. Cuffdiff [7] was also used to test for differential expression, but no genes could be defined with an FDR < 0.05 using 1-factor analysis (multi-factor analysis is not possible for cuffdiff). In all cases, TopHat single-end reads were used for analysis (quantified using cufflinks for cuffdiff and Partek EM for DESeq and edgeR).

Microarray Processing

Feature extraction files from GSE37704 were imported into Partek, and gProcessedSignal values were averaged by probe for gene expression analysis. Signal intensities were $\log_2$ transformed and quantile normalized. FDR values were calculated using the method of Benjamini and Hochberg [15] from the distribution of 1-way ANOVA p-values, and fold-change values were calculated on a linear scale using least-squares mean. Because only 28 probed meet the criteria of showing a |fold-change| > 1.5 and false discovery rate (FDR) < 0.05, microarray probes were defined as differentially expressed if they showed a |fold-change| > 1.5 and unadjusted p-value < 0.05. Differential expression of a single probe was enough to define a gene as differentially expressed.

sRAP Bioconductor Package

The strategies discussed in this paper have been implemented as part of the Simplified RNA-Seq Analysis Pipeline (sRAP), which is an R package that is available in Bioconductor [16]. sRAP normalizes RNA-Seq data by $\log_2$ transforming RPKM values with a rounding threshold (with a default setting of 0.1). Users can then define differentially expressed genes using fold-change cutoffs (default = 1.5), p-value cutoffs (calculated using ANOVA or linear regression, with ANOVA p-value < 0.05 as the default setting), and/or false discovery rate (Benjamini and Hochberg FDR, with a default setting of FDR < 0.05) values. The package also provides quality control metrics and functional enrichment via BD-Func [17], but those elements are not presented in this paper. Analysis presented in this paper was carried out using default parameters. Because $\log_2$ transformation is a required step in normalization for sRAP, RPKM values without a rounding threshold were estimated by adding an extremely small RPKM cutoff (1×10^{-45}). The sRAP package is available at http://www.bioconductor.org/packages/release /bioc/html/sRAP.html.

RESULTS AND DISCUSSION

RPKM Distributions are Similar for Most Aligners

Three different aligners (TopHat, Novoalign, and STAR) were used with both single-end and paired-end alignments (except for TopHat, where the paired end alignment was too time-consuming). The number of genes above various RKPM cutoffs was very similar regardless of alignment method (Figure 1). Because a large patient cohort was used for analysis, RKPM distributions can be compared not only for the overall population (Figure 1), but also for each individual sample. One important question is how the number of reads affects the distribution of RPKM values. To test this, correlations between number of genes above a RPKM cutoff and the total number of reads per sample were calculated for RPKM cutoff values of 1, 0.1, 0.01, and 0.0001 (which were selected based upon the changes in gene counts for RPKM value cutoffs at the population level, Figure 1). In most cases, high RPKM cutoffs (e.g. 1) showed a positive correlation with read counts (Figure 2). In other words, samples with more reads had a larger proportion of genes with RPKM cutoffs greater than 1, which makes intuitive sense. However, sufficiently low RPKM cutoffs (e.g. 0.01 and 0.001) caused a negative correlation between the number of genes above an RPKM cutoff and the total number of reads in the sample. There was no statistically significant correlation ($p < 0.05$) between the number reads per sample and the number of genes with RPKM cutoff greater than 0.1. All other things being equal, this indicates that trends observed for rounded RPKM values with an RPKM cutoff of 0.1 should be applicable regardless of the number of reads per sample.

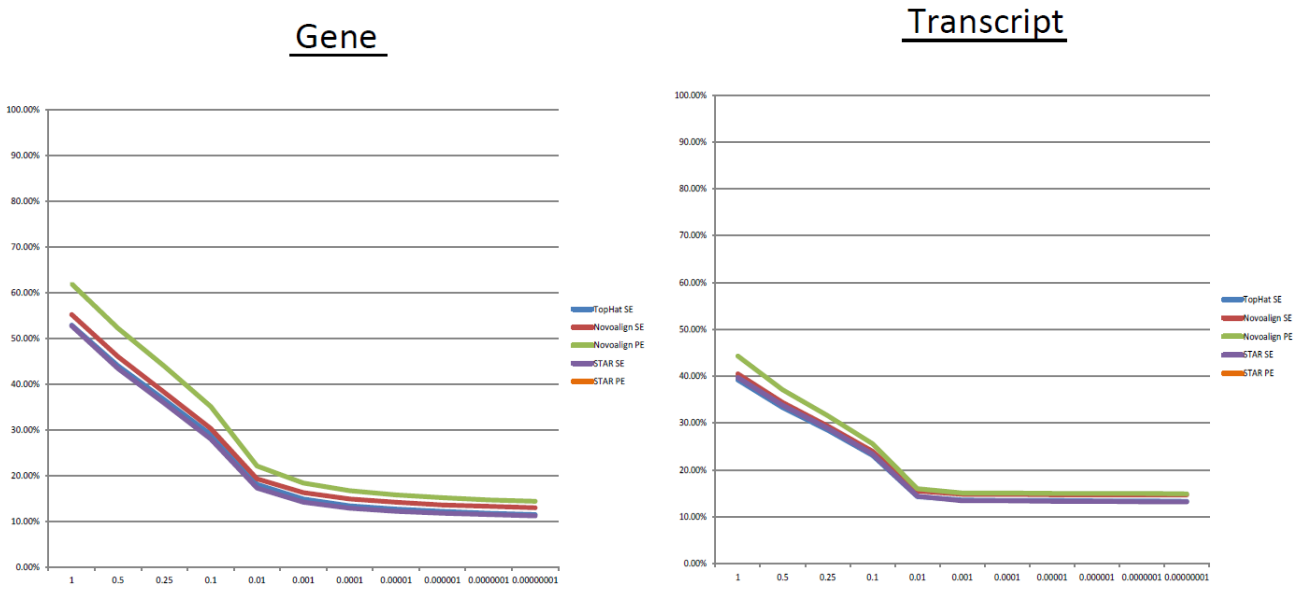

Figure 1. RPKM Distributions are Robust for Multiple Aligners. Reads were aligned using 5 different strategies (TopHat Single-End, STAR Single-End, STAR Paired-End, Novoalign Single-End, Novoalign Paired-End). RPKM normalized read counts (for gene-based and transcript-based quantification) were then pooled from all aligned samples, and the proportion of transcripts or genes above a certain threshold was measured. RKPM thresholds are shown along the x-axis and the proportion of genes / transcripts with signal above that threshold are shown on the y-axis. For both gene-based and transcript-based

quantification, the proportion of genes above various thresholds is the almost identical for all 5 alignment strategies. In particular, there are substantial diminishing returns when selecting thresholds less than 0.01

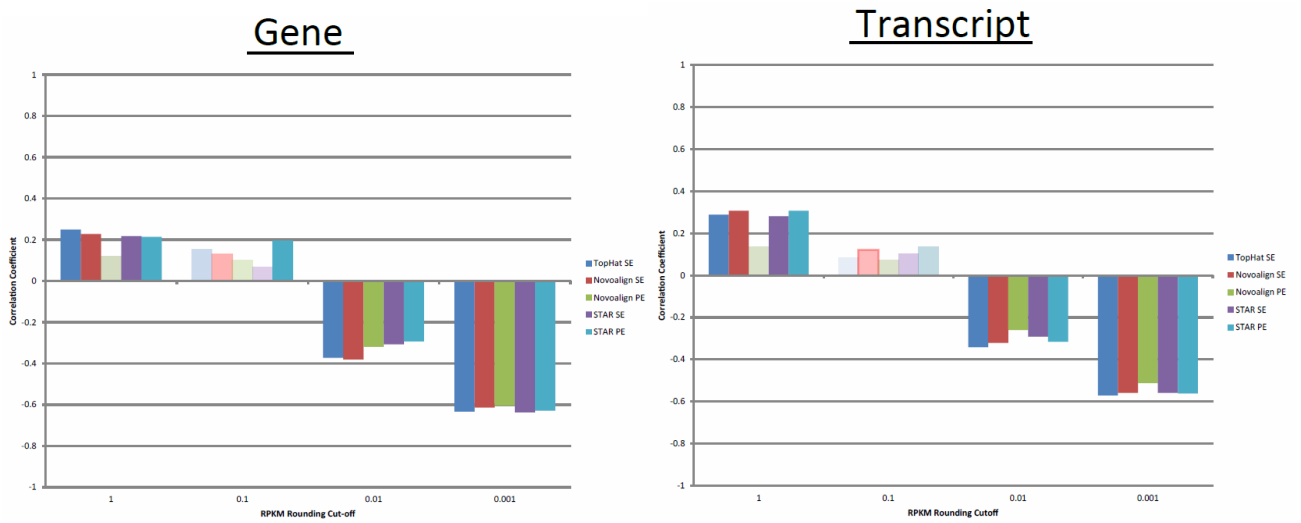

Figure 2. Influence of Read Count on RPKM Levels. Reads were aligned using 5 different strategies (TopHat Single-End, STAR Single-End, STAR Paired-End, Novoalign Single-End, Novoalign Paired-End). For each sample, the proportion of transcripts / genes above a certain RPKM normalized read counts, which is shown along the x-axis. Correlation coefficients between the total number of reads per sample and the number of genes above the given RPKM threshold were then calculated, which are plotted on the y-axis. Correlation coefficients with p-values > 0.05 are between gene / transcript count and total number of reads, while using a smaller cutoff leads partially transparent, while correlations with p-values < 0.05 are colored with darker shades of the colors specified in the figure legend. For all 5 alignment strategies, the total number of reads per sample had the least influence on the number of genes with RPKM > 0.1. Using a larger cutoff leads to a positive correlation to an increasingly negative correlation between gene / transcript count and the total number of reads.

Genome Alignment Only Modestly Affects Concordance of Differentially Expressed Genes

In order to better approximate a normal distribution, rounded RPKM values were $\log_2$ transformed, which affects both sample distributions (Figure S1) and gene distributions (Figure S2). Lists of differentially expressed genes were defined using 4 cutoffs: 1, 0.1, 0.01, and 0 (no rounding). Gene lists were also compared using different alignments (Figure 3). This comparison yielded 2 noteworthy results. First, the impact of RPKM cutoffs on gene lists was similar using most aligners. The only exception is the single-end Novoalign alignment. This is reasonable because Novoalign has special parameters for RNA-Seq data for paired-end but not single-end alignments. These results emphasize that Novoalign is not suitable for single-end RNA-Seq alignments (but paired-end alignments are OK) and this strategy of comparing alignments can reveal good-quality versus poor quality alignments.

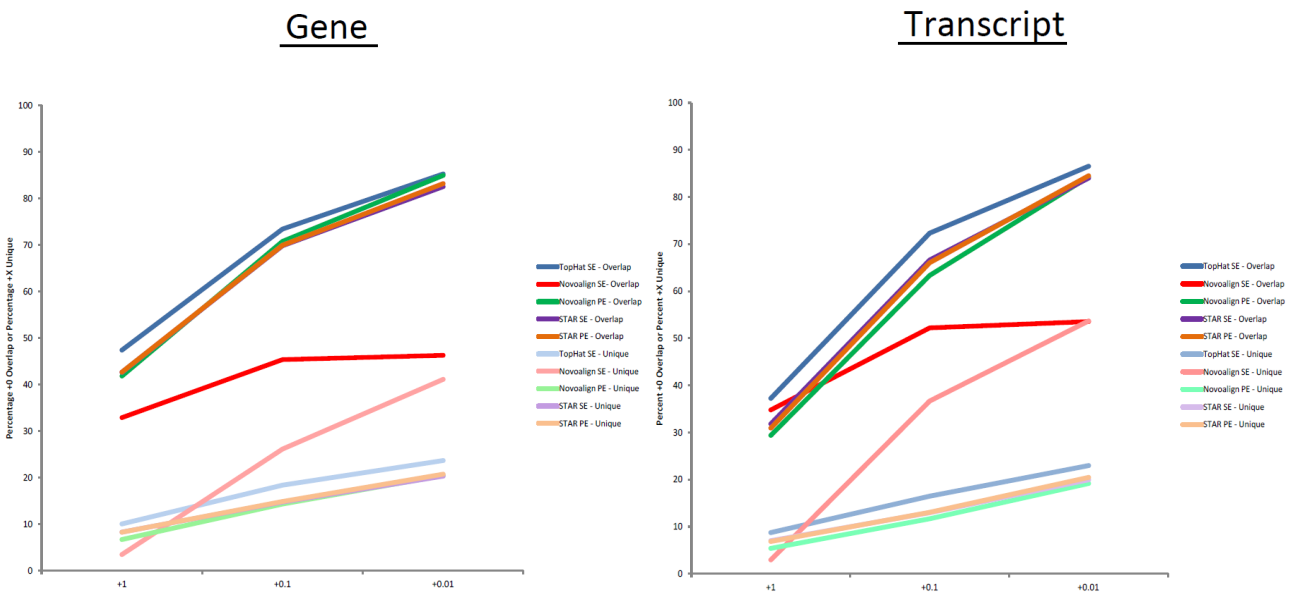

Figure 3. Bias Introduced by Rounding RPKM Values. The goal of rounding RPKM values is to help define more biologically-relevant fold-change values. However, choosing a high rounding threshold can remove useful genes from the resulting gene list and choosing a low rounding threshold can introduce new genes into the gene list that may be an artifact from rounding. Therefore, it is useful to compare gene lists produced from normal RPKM values to those defined using $\log_2$(RPKM + cutoff)

values. In this figure, gene and transcript lists for cutoffs of 1, 0.1, and 0.01 are compared to gene lists producing unadjusted RPKM expression values, using 5 alignment strategies. Choosing a cutoff of RPKM > 1 introduces the least amount of potential bias, but also reduces the size of the gene / transcript lists by more than 50%. The single-end Novoalign alignment is an outlier, probably because Novoalign only has parameters to optimize for mapping across splicing junctions for paired-end alignments. In all other cases, the majority of genes in the unadjusted PRKM gene list can be recovered using cutoffs of 0.1 or 0.01, with a modest increase in potential bias from rounding. In all cases, "overlap" is defined as the overlap between the gene list from a given cutoff and the gene list with no RPKM cutoff (+0): so, this overlap is independent of all other gene lists.

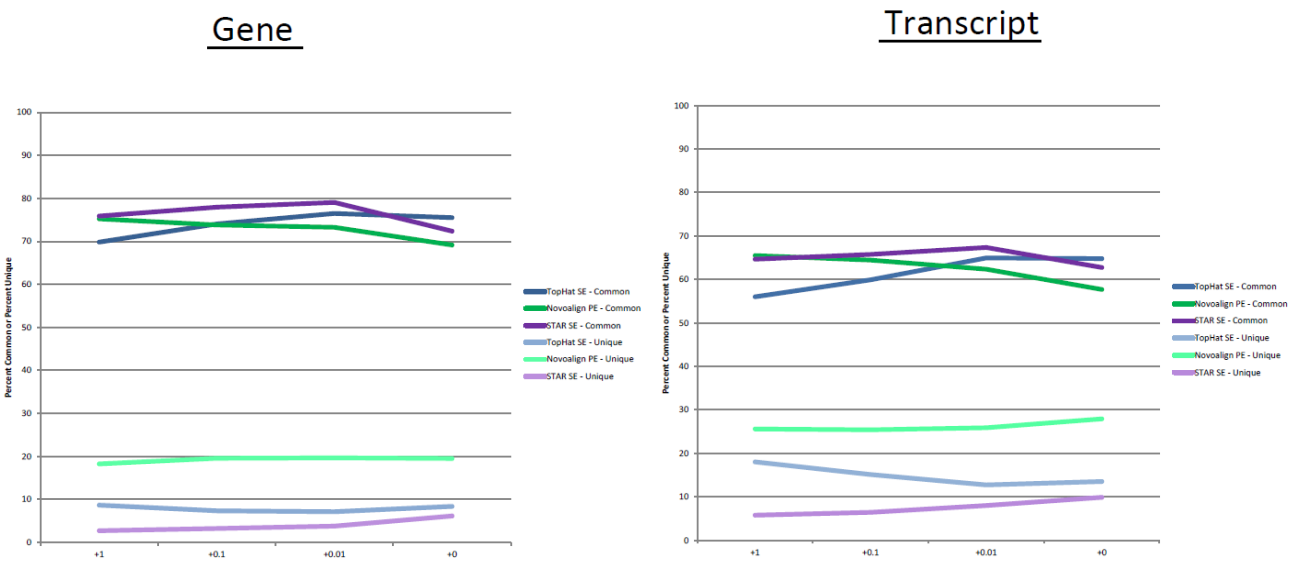

Figure 4. Gene Lists are Similar Regardless of Alignment Method. Three alignment methods (TopHat Single-End, STAR Single-End, Novoalign Paired-End) were tested to directly compare the impact of alignment method on production of gene lists. Novoalign paired-end alignment was used because the single-end alignment wasn't reliable. STAR single-end alignment was used without the paired-end alignment because the results were almost identical. The number of common and unique genes is roughly similar regardless of RPKM cutoff. In all cases, "common" refers to the set of genes (or transcripts) defined by all 3 alignments.

The second important conclusion is that higher RPKM cutoff values will result in smaller gene lists that are mostly (>90%) a subset of a gene list produced without rounding RPKM values (or performing a $\log_2$ transformation), whereas smaller RPKM cutoff values can result in gene lists that are more similar to the unrounded gene list but also contain more genes that were not identified prior to rounding and $\log_2$ transformation. Although the additional genes could be detected due to increased statistical power (due to a more normally distributed set of expression values), it is also possible that some of these genes are artifacts from rounding. Either way, rise is overlap is sharper than the rise in unique genes, so the benefits may outweigh the potential drawbacks. However, this decision can be left up to the individual analyst, and this figure importantly illustrates the quantitative relationship between RPKM cutoff and sensitivity.

The previous analysis studied how the relative RPKM cutoffs influence gene lists using various alignments, comparing rounded RPKM gene lists to unrounded RPKM gene lists. Alternatively, it can be useful to test if the alignment method affects the concordance of the gene lists (and if this concordance between alignments changes with RPKM cutoff). Unlike the relative analysis, the concordance between genes lists at various RPKM cutoffs (or for unrounded RPKM analysis) was fairly high (>70%, Figure 4). This emphasizes that the results from this study are not highly dependent on which of these three popular aligners is used.

Likewise, concordance of gene list was compared for two different mRNA quantification tools (cufflinks and Partek-EM). Gene-level lists were extremely similar using both programs (Figure 5), but there were more discrepancies for transcript-level lists. Although the concordance was still relatively high, this is one reason why all subsequent analysis will focus solely on gene-level analysis.

Algorithms Using Rounded RPKM Values Provide Robust Lists of Differentially Expressed Genes

The type of analysis presented in this paper is very common for microarray analysis, but RPKM values cannot be rounded for the most popular tools for RNA-Seq analysis. Namely, edgeR and DESeq require the user provide raw read counts for estimating dispersion values, and cuffdiff relies on the read alignment (and not just the RPKM values from cufflinks). Therefore, the broader applicability of the impact on RPKM rounding depends on the usefulness of the type of statistical analysis in the paper, and it is important to compare different algorithms for defining differentially expressed genes.

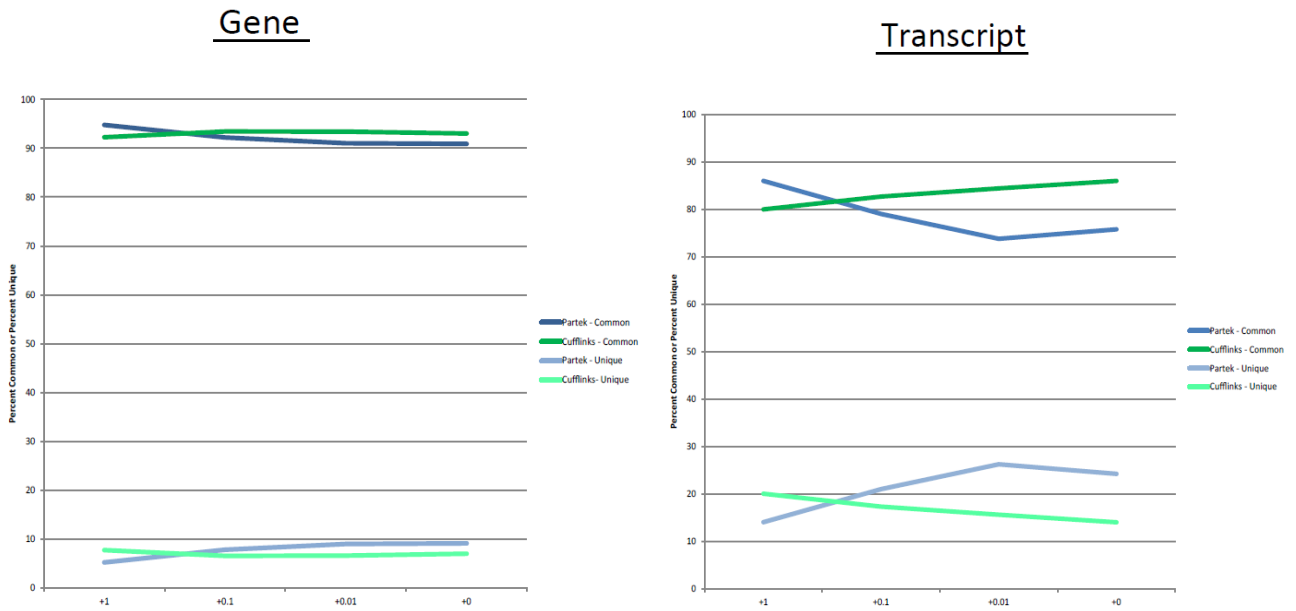

Figure 5. Concordance of Gene Lists Using Different mRNA Quantification Algorithms. Two methods of mRNA quantification (Partek EM algorithm and cufflinks) were testing and differentially expressed gene lists were created using various RPKM cutoffs. In both cases, "common" refers to genes (or transcripts) identified using both Partek-EM and cufflinks. **A.** Gene-level quantification is very similar regardless of mRNA quantification method or RPKM cutoff. **B.** Transcript level quantifications are relatively similar for both mRNA quantification tools, but the variation between programs is greater than gene-level quantification. The program with the greatest overlap depends upon the RPKM cutoff, so it is difficult to recommend an optimal tool

First, the lung cancer cohort was used to compare genes that were differentially expressed between paired normal and tumor samples. In all cases, genes were required to show a |fold-change| value > 1.5 and FDR < 0.05 in order to be considered differentially expressed. When possible, sample pairing was taking into consideration for the statistical analysis. This sort of multivariable modeling is not possible in cuffdiff, so simple tumor versus normal analysis was also conducted using Partek (2-way ANOVA with linear contrast), DESeq, and edgeR. No genes meet the criteria for differential expression for cuffdiff, so only gene lists from Partek, DESeq and edgeR could be compared. Importantly, the concordance between gene lists considerably varied based upon the differential expression algorithm and RPKM cutoff (Figure 6). For the simple 1-factor analysis, Partek always produced the gene list with the greatest number of genes represented in both the DESeq and edgeR gene lists. For the 2-factor (paired) analysis, the most robust gene list varied between Partek and DESeq, depending upon the RPKM cutoff used for rounding. In contrast, edgeR always showed the least amount of genes detected by all 3 algorithms and the largest number of genes detected by neither of the other two algorithms. This emphasizes that the analysis described in this paper is as good – if not better – for analysis of large patient cohorts, in comparison to these other popular tools.

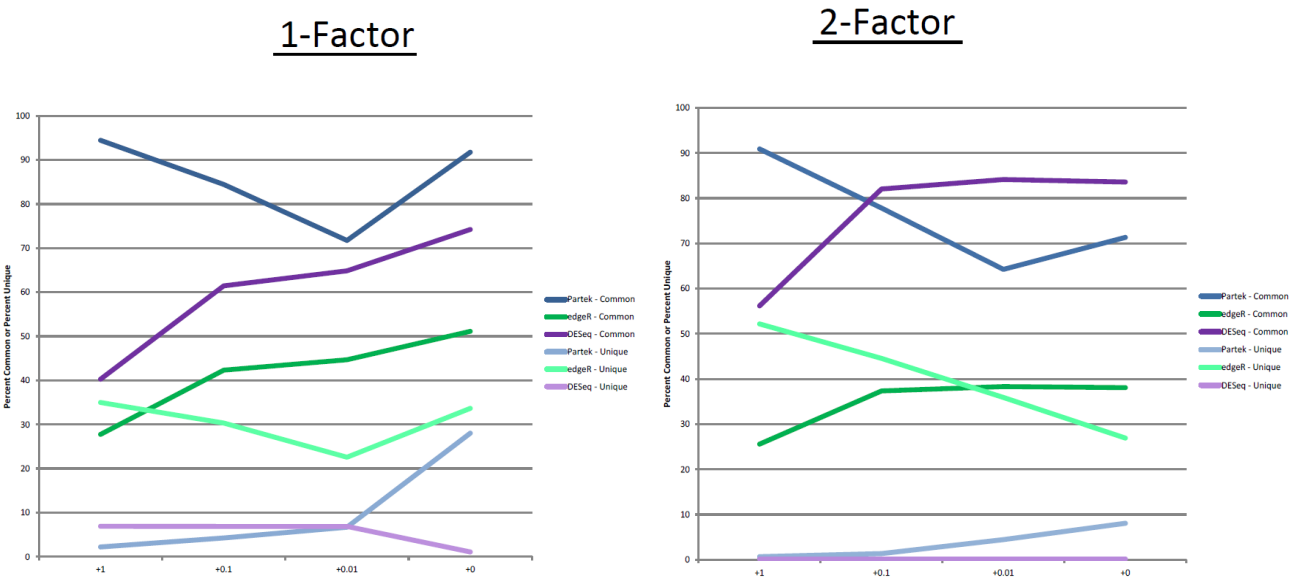

Figure 6. 2-way ANOVA of Rounded RPKM Values Provides Robust Gene Lists. Paired patient data for tumor and normal samples were compared, either with pairing information (2-Factor) or without pairing information (1-Factor) included in the

model for statistical analysis. The concordance of gene lists varies greatly depending upon the algorithm used for analysis. The algorithm with the greatest overlap depended upon the number of variables and RPKM cutoff (for analysis in Partek only), but edger always showed the least level of concordance. In all cases, the criterion for differential expression is |fold-change| > 1.5 and FDR < 0.05, and cuffdiff results could not be included in this plot because no genes met these criteria. In both cases, "common" refers to the list of genes identified using all 3 algorithms (Partek, edgeR, and DESeq).

Because Partek is a commercial program that requires users to purchase a yearly license (and all the other tools are open-source), it is important to emphasize that the Partek strategy is based upon standard statistical analysis that can be emulated with relative ease. To demonstrate this, a novel Bioconductor package was developed to assist users in reproducing the strategies discussed in this paper as part of the Simplified RNA-Seq Analysis Pipeline (sRAP) package, and the results of the sRAP package were compared to the Partek results. Using the same parameters, sRAP produced a much smaller list of differentially expressed genes (5770 genes for Partek versus 576 genes for sRAP, using an RPKM cutoff of 0.1), but the sRAP gene list was almost entirely a subset of the Partek gene list, as would be expected (Figure S3). This is likely to be due to slightly different implementations: for example, the Partek results may be more similar to the sRAP results when using a lower fold-change cutoff on sRAP. This is further emphasized by the subsequent cell line analysis.

Most benchmarks for RNA-Seq algorithms have been compared using cell line data. Although we consider the use of a large patient cohort to be a distinct advantage to this study, it is also useful to compare cell line analysis in order to emphasize the validity of the result. In other words, the previous analysis presented in Figure 6 implies that three of the most popular RNA-Seq analysis tools may not be optimal for RNA-Seq analysis, so it will help to demonstrate a comparison of rounded RKPM analysis (in Partek and sRAP) in a scenario that better illustrates the capabilities of these other popular tools. In fact, we conducted a comparison using the same cell line data published with the cufflinks/cuffdiff paper [7]. Both MiSeq and HiSeq data was produced for this paper, but the results were similar for both platforms (Figure S4). Therefore, only the MiSeq data was used for subsequent analysis, in order to simplify the presentation of results.

As expected, gene lists can now be defined using all 5 algorithms (Partek, sRAP, DESeq, edgeR, and cuffdiff), and the concordance between gene lists is now similar regardless of RPKM cutoff used for rounding in Partek and sRAP (Figure 7). The relative concordance between Partek, DESeq, and edgeR was roughly similar to the patient data (slightly better for Partek than DESeq, much worse for edgeR), but the Partek and sRAP gene lists are much more similar for the cell line data. It has been previously reported that DESeq is more conservative than edgeR [18], which is matches the results presented in this study. The percentage of genes identified with all 5 algorithms was similar for cuffdiff and DESeq, but cuffdiff listed more genes not identified by any of the other 4 algorithms.

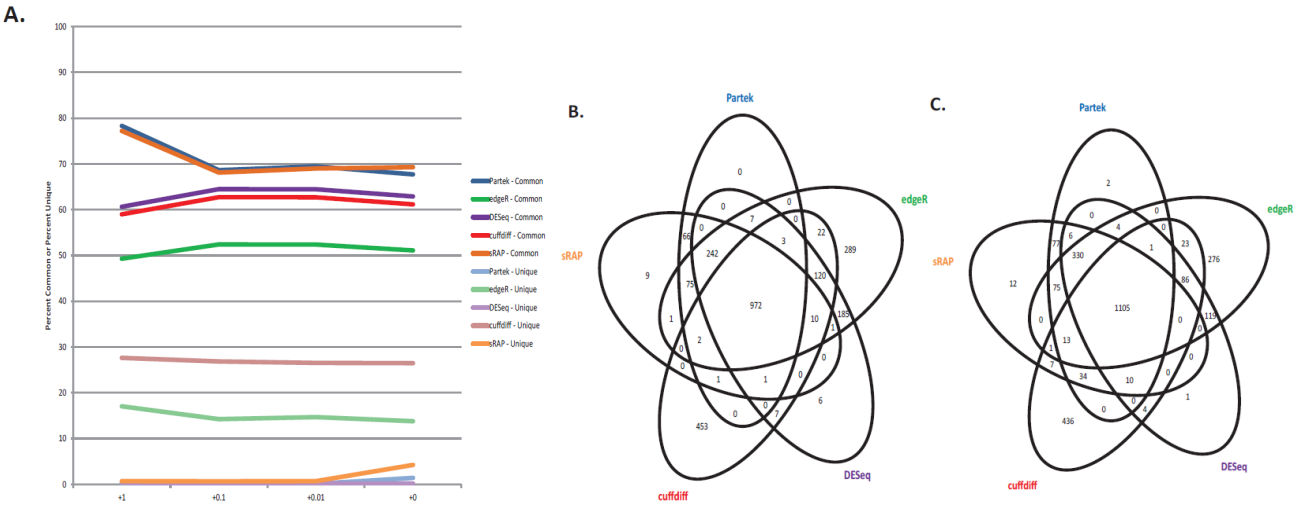

Figure 7. Partek and sRAP Provide the Most Robust Lists of Differentially Expressed Genes. Cell line data comparing two groups with triplicates was compared using 5 different algorithms (Partek, edgeR, DESeq, sRAP, and cuffdiff). RPKM cutoffs were varied for Partek and sRAP analysis. Unlike the patient dataset (where sRAP produced a small gene list and cuffdiff could not produce a gene list), all 5 algorithms provides lists of over 1000 differentially expressed genes. **A.** Concordance is more consistent across RPKM cut-offs for cell line compared to patient data, but both data types show significantly different levels of concordance depending upon which algorithm is used. Partek and sRAP contain the largest number of genes defined by all 5 algorithms. In this figure, "common" refers to gene identified using all 5 algorithms, and "unique" refers to genes only identified by 1/5 algorithms. **B.** Up-regulated gene overlap for all 5 algorithms. RPKM cutoff for Partek and sRAP is 0.1. **C.** Down-regulated gene overlap for all 5 algorithms. RPKM cutoff for Partek and sRAP is 0.1

All of the analysis presented so far indicates that rounded RPKM values provide robust gene lists, but this does not strictly describe the accuracy of the gene lists. For example, it is possible that programs like

edgeR (or cuffdiff, for the cell line data) had greater sensitivity and were still predicting valid, unique genes. The cell line RNA-Seq data also had corresponding microarray data, so the RNA-Seq and microarray results were compared. This resulted in less variability between algorithms, but there were some qualitative similarities to the robustness analysis (Figure S5): for example, the RNA-Seq genes lists produced by Partek and sRAP (and usually DESeq) contained the largest number of genes identified as differentially expressed in the microarray experiment. These results indicate that it is probably not reasonable to dismiss all of the uniquely identified genes as simply false positives, although it is possible that the uniquely identified genes still have some correlation with a false positive rate. The only conclusion that can be conservatively drawn from these results is that simple statistical methods (like those implemented in Partek and sRAP) are at least as good as these other popular algorithms, which is significant because these popular algorithms cannot use rounded RPKM values for statistical analysis.

CONCLUSION

This study illustrates how RPKM rounding will affect the size of differentially expressed genes, allowing analysts to pick the most suitable RPKM cutoff for their analysis. RPKM rounding has similar effects regardless of what aligner is used or what mRNA quantification tool is used to calculate RPKM (or FPKM, Fragments per Kilobase per Million reads, for cufflinks). In contrast, different algorithms for defining differentially expressed genes yielded results that substantially varied for different RPKM cutoffs for analysis of a large patient cohort (although RPKM cutoff was less significant when cell line data was compared).

These results are significant because they emphasize that strategies used for microarray analysis can also work as well as strategies that are unique for RNA-Seq analysis. This observation has been previously published [19], but no study has investigated the impact of rounding RPKM values on lists of differentially expressed genes. This is significant because fold-change values are commonly used by biologists to prioritize differentially expressed genes, and unrounded RPKM values can result in unreasonably large fold-change values (in low coverage genes). This study emphasizes that analysis using rounding RPKM values provides statistically sound results that are easy for biologists to interpret.

REFERENCES

1. Garber, M., et al., Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Meth, 2011. 8(6): p. 469-477.
2. Oshlack, A., M. Robinson, and M. Young, From RNA-seq reads to differential expression results. Genome Biology, 2010. 11(12): p. 220.
3. Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009. 10(1): p. 57-63.
4. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010. 26(1): p. 139-140.
5. Anders, S. and W. Huber, Differential expression analysis for sequence count data. Genome Biology, 2010. 11(10): p. R106.
6. Mortazavi, A., et al., Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth, 2008. 5(7): p. 621-628.
7. Trapnell, C., et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protocols, 2012. 7(3): p. 562-578.
8. Consortium, M., The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotech, 2006. 24(9): p. 1151-1161.
9. Seo, J.-S., et al., The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Research, 2012. 22(11): p. 2109-2119.
10. Dobin, A., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013. 29(1): p. 15-21.
11. Leinonen, R., H. Sugawara, and M. Shumway, The Sequence Read Archive. Nucleic Acids Research, 2011. 39(suppl 1): p. D19-D21.
12. Kim, D., et al., TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 2013. 14(4): p. R36.
13. Xing, Y., et al., An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Research, 2006. 34(10): p. 3150-3160.
14. Pruitt, K.D., T. Tatusova, and D.R. Maglott, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 2005. 33(suppl 1): p. D501-D504.
15. Benjamini, Y. and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 1995. 57(1): p. 289-300.
16. Gentleman, R., et al., Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 2004. 5(10): p. R80.
17. Warden, C.D., et al., BD-Func: a streamlined algorithm for predicting activation and inhibition of pathways. PeerJ, 2013. 1: p. e159.
18. Robles, J., et al., Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics, 2012. 13(1): p. 484.
19. Rapaport, F., et al., Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology, 2013. 14(9): p. R95.

Access Figures S1 – S5 from:
http://bioinfo.aizeonpublishers.net/content/2013/6/bioinfo285-292_FigS1-S5.pdf

© 2013; AIZEON Publishers; All Rights Reserved

This is an Open Access article distributed under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*****