Investigation of Factors Affecting RNA-Seq Gene Expression Calls

Sahar Harati, John H. Phan, *IEEE Member*, and May D. Wang, *IEEE Member*

***Abstract*—RNA-seq enables quantification of the human transcriptome. Estimation of gene expression is a fundamental issue in the analysis of RNA-seq data. However, there is an inherent ambiguity in distinguishing between genes with very low expression and experimental or transcriptional noise. We conducted an exploratory investigation of some factors that may affect gene expression calls. We observed that the distribution of reads that map to exonic, intronic, and intergenic regions are distinct. These distributions may provide useful insights into the behavior of gene expression noise. Moreover, we observed that these distributions are qualitatively similar between two sequence mapping algorithms. Finally, we examined the relationship between gene length and gene expression calls, and observed that they are correlated. This preliminary investigation is important for RNA-seq gene expression analysis because it may lead to more effective algorithms for distinguishing between true gene expression and experimental or transcriptional noise.**

I. Introduction

RNA-seq has greatly improved the dynamic range of gene expression quantification, enabling the detection of very low and very high-expressed genes. However, accurate quantification of RNA-seq gene expression remains a challenge [1]. The random nature of RNA-seq (i.e., due to the random sampling of sequences) and the presence of experimental and/or transcriptional noise leads to an inherent ambiguity in distinguishing between noise and low-expression genes, i.e., "calling" gene expression [2, 3]. The presence of experimental noise complicates the detection of changes in low-expression genes, which may potentially be important disease biomarkers. Thus, we examined methods for identifying true gene expression calls and investigated factors in RNA-seq data analysis pipelines that may affect the detection of low-expression genes.

This work was supported in part by grants from the National Institutes of Health (NHLBI 5U01HL080711, Center of Cancer Nanotechnology Excellence U54CA119338, 1RC2CA148265), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to Professor May D. Wang), Microsoft Research, and Hewlett Packard.

S. Harati is a visiting scholar with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (e-mail: saharati2008@gmail.com).

J. H. Phan is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (e-mail: jhphan@gatech.edu).

M. D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (corresponding author, phone: 404-385-2954; fax: 404-385-0383; e-mail: maywang@bme.gatech.edu).

A simple method to distinguish between true gene expression calls and experimental noise involves thresholding on the total number of reads mapped to each gene. Genes with a total number of reads that is smaller than the threshold are deemed to be not expressed, and any reads that appear to originate from the gene are believed to be experimental noise. However, determining the appropriate threshold is challenging. Wagner et al. modeled gene expression distributions as a mixture of negative binomial and exponential distributions to represent functional expression and noise, respectively [4]. They observed that classifying expression levels with such a mixture model resulted in an empirical threshold of approximately 1 RPKM (reads per kilobase per million mapped reads [5]), which was consistent across a variety of RNA-seq datasets. Moreover, this result was in agreement with the results of a different approach by Hebenstreit et al. [2]. Hebenstreit et al. sought to differentiate between low and high expression genes by modeling the distribution of reads mapping to intronic and intergenic regions. Using these non-exonic distributions to determine a threshold, and using PCR to verify low-expression genes, they observed that these low-expression genes could likely be attributed to "leaky", but non-functional expression. Although these studies have established techniques for identifying true gene expression calls, the impact of such methods on RNA-seq applications is unknown. Moreover, it is unclear how RNA-seq data analysis pipelines affect gene expression calls.

We conducted a preliminary investigation of factors affecting RNA-seq gene expression calls. Using a method similar to that described by Hebenstreit et al., we empirically estimated the distribution of reads that mapped to exonic, intronic, and intergenic regions of the human genome, and observed distinct differences among these distributions (**Figure 1**). We then compared how these distributions changed when using different sequence mapping algorithms. Finally, we examined properties of genes such as length and number of exons, and observed that these properties are different in genes with a tendency to be expressed at levels indistinguishable from noise. A comprehensive investigation of these factors may be important for designing RNA-seq data analysis pipelines, improving the accuracy of gene expression estimation, and understanding transcriptional activity.

II. Methods

A. RNA-Seq Data and Sequence Mapping

We used an RNA-seq sample containing the Stratagene Universal Human Reference RNA (UHRR), sequenced using Illumina technology. The sample was sequenced to a depth of approximately 4-5 million paired-end reads with read

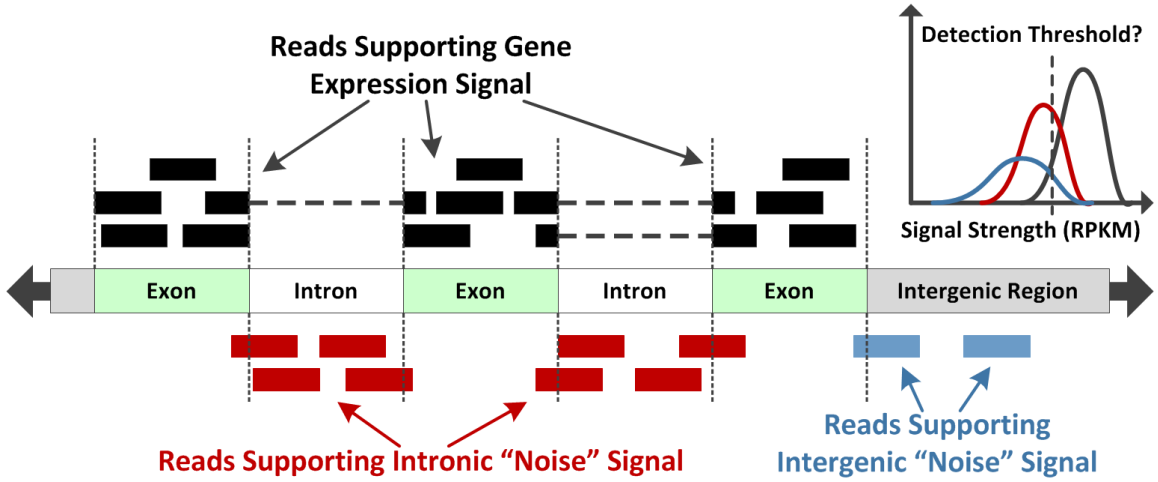

Figure 1. The distributions of RPKM-normalized read counts in the exon, intron, and intergenic regions produce distinct distributions. The properties of these distributions may be used to infer the level of experimental or transcriptional noise in an RNA-seq experiment.

length of 100 base pairs. We used BWA and TopHat to map sequences to the human genome (hg19) [6, 7]. We used a two-step alignment procedure for BWA. First, we mapped sequences to the AceView transcriptome [8], then we mapped remaining sequences to the human genome. Both mapping results (i.e., AceView transcriptome and human genome) were combined to produce the final mapping results [9]. For TopHat, we aligned all reads directly to the human genome while using the AceView transcriptome to guide the mapping of reads spanning exon junctions. Both RNA-seq pipelines produced BAM-formatted alignment files, which were used for subsequent gene expression quantification.

B. *Quantification of Gene Expression*

We used HTSeq to quantify gene expression as the number of reads that mapped to the exons of each gene in the AceView annotation [8, 10]. First, using SAMTools, we sorted the BAM-formatted alignment files by sequence read names [11]. Second, we used HTSeq with the GTF-formatted (i.e., General Transfer Format) AceView annotation and a sorted BAM file as input to count all reads that map completely within the exonic regions of genes. That is, a read was assigned to a gene only if the entire read was mapped within the exonic regions of the gene. This counting criterion is called "intersection-strict" in HTSeq. Reads that were only partially mapped to an exon, with the remainder mapping to intronic or intergenic regions, were not assigned to the gene. The assumption for this criterion is that reads partially mapped to introns or intergenic regions were more likely to be noise. Finally, we normalized the read counts for each gene using reads per kilobase per million mapped reads (RPKM) by estimating gene length as the sum of the lengths of all of the gene's exons [5].

C. *Quantification of "Noise" Expression*

We quantified RNA-seq noise using a method similar to that of Hebenstreit et al. [2]. Specifically, we used HTSeq to quantify reads mapping to intronic and intergenic regions in a manner similar to that of exonic regions. In order to achieve this, we created separate GTF-formatted annotations for introns and intergenic regions. The intron annotation file contains the start and end coordinates of all unique introns for each gene. The intergenic annotation file contains the start and end coordinates of all intergenic regions. In contrast to the "intersection-strict" HTSeq option used for exons, we used "intersection-nonempty" for both intron and intergenic counts. This criterion assigns a read to a gene's intronic region if that read partially or completely maps to one of the gene's introns. Similarly, it assigns a read to an intergenic region if that read partially or completely maps to the intergenic region. We normalized intronic read counts using RPKM by summing the number counts that mapped to all introns of each gene, then estimating intronic length as the sum of the lengths of all of the gene's introns. Intergenic RPKM was similarly computed, except that each intergenic feature contains only one contiguous region.

D. *Threshold Estimation for Gene Expression Calls*

We estimated the threshold for gene expression calls as the 90% quantile of intergenic RPKM values [2]. That is, the threshold is defined such that 10% of all intergenic regions have "noisy" RPKM expression values at or above the threshold; and the remaining 90% of intergenic regions have RPKM expression below the threshold. We then used this threshold as the criterion for detecting true gene expression signals vs. noisy signals. All genes with RPKM values below the threshold were deemed to be indistinguishable from noise.

III. Results and Discussion

A. *Distributions of Exonic, Intronic, and Intergenic Mapping are Distinct*

The AceView human transcriptome contains over 55,000 genes, including many experimental sequences. Thus, it is less conservative compared to transcriptome databases such as RefSeq [12]. Among these genes, over 38,000 contain multiple exons, i.e., these genes include intronic regions. Moreover, due to the overlap of some genes, only approximately 40,000 intergenic regions exist. **Table 1** lists the total number of exonic, intronic, and intergenic features, along with mapping statistics for the BWA and TopHat

TABLE 1. MAPPING STATISTICS FOR BWA AND TOPHAT PIPELINES

BWA Mapping			
	RPKM = 0	RPKM > 0	Total Features
Gene (Exon)	28,499	27,375	55,874
Intron	17,043	21,535	38,578
Intergenic	18,030	22,456	40,486
TopHat Mapping			
	RPKM = 0	RPKM > 0	Total Features
Gene (Exon)	29,666	26,208	55,874
Intron	17,184	21,394	38,578
Intergenic	18,313	22,173	40,486

mapping pipelines. Roughly half of all exonic, intronic, and intergenic features map to at least one read (i.e., RPKM > 0).

There is a clear difference among the distributions of reads mapping to exons, introns, and intergenic regions. **Figure 2** illustrates the three distributions for the BWA (**Figure 2A**) and TopHat (**Figure 2B**) mapping pipelines. As expected, reads are more likely to map to exons (black distribution) than to introns (red) or intergenic regions (blue). Moreover, reads are more likely to map to introns than to intergenic regions. That is, the distribution of intronic RPKM is slightly shifted in the positive direction compared to that of intergenic regions. This may be explained by gene splice variants in the intronic regions that have yet to be discovered.

We do not observe considerable qualitative differences between the BWA and TopHat mapping pipelines in terms of exon, intron, and intergenic region RPKM distributions. This may be due to the similarity of the underlying sequence mapping algorithm in both aligners, i.e., a Burrows-Wheeler transform-based algorithm. However, a more quantitative and comprehensive analysis is necessary to determine if the choice of analysis pipeline affects RNA-seq expression distributions.

B. Mapping Distributions May Be Informative for Gene Expression Calls

A convenient property of the decomposition into exon, intron, and intergenic region RPKM distributions is that we can estimate a "confidence" for true gene expression given a specific expression level. For example, we can observe that about 3000 genes are expressed with $\log_2$ RPKM of 1. In contrast, there are approximately 800 genes with intronic regions expressed at $\log_2$ RPKM of 1, and 700 intergenic regions with the same expression level. Assuming that reads mapping to intronic or intergenic regions truly represent experimental or transcriptional noise (i.e., we assume that our knowledge of the gene annotation is complete), we can estimate a "confidence" of 67% for true gene expression at a level of 1 RPKM since 3000/4500 features are expressed at a level of 1 RPKM. This confidence may be computed for all RPKM values.

Equipped with these distributions, we can find a suitable threshold for gene expression calls based on noise tolerance. For example, if we want a confidence of at least 50% for gene expression calls, we should choose a threshold such that the frequency of both the intergenic and intronic

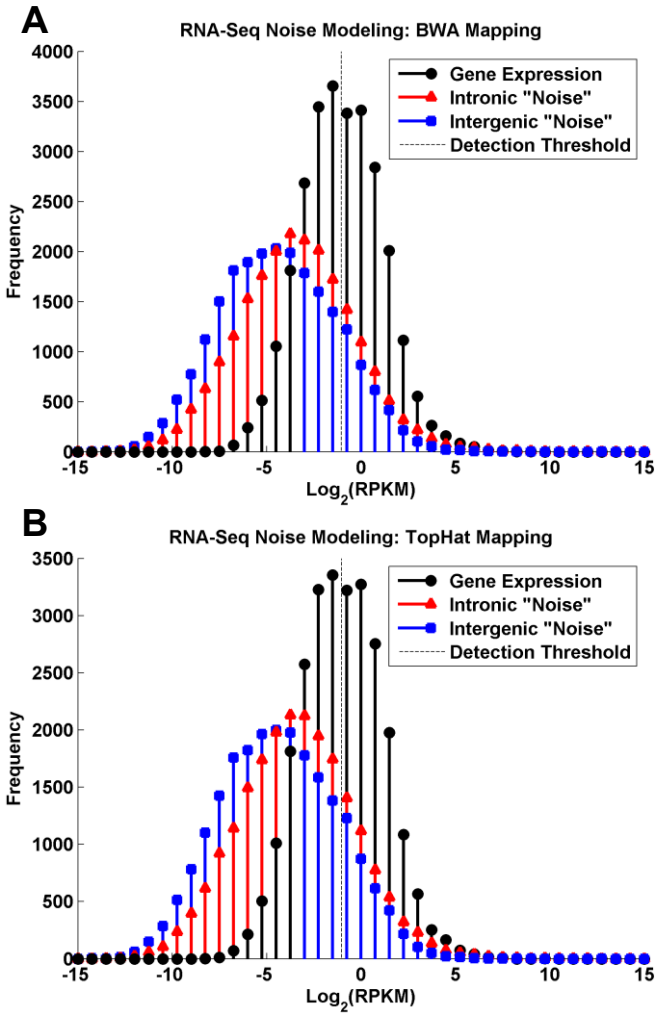

Figure 2. Distribution of true gene expression (i.e., exonic signal, black), intronic noise (red), and intergenic noise (blue) for (A) the BWA mapping pipeline and (B) the TopHat mapping pipeline. The dashed vertical line indicates the true gene expression calling threshold, determined as the 90% quantile of intergenic noise.

distributions is equal to that of the exonic distribution. Interestingly, this results in a similar threshold to that of the *ad hoc* method introduced by Hebenstreit et al. [2]. They used a 90% quantile of the intergenic distribution as the threshold, depicted by the dashed vertical line in **Figure 2**.

C. Properties of Genes Expressed Above and Below the Detection Threshold

We further characterize the nature of gene expression in the presence of experimental or transcriptional noise by examining gene properties such as length and number of exons that may be correlated with the threshold. Genes expressed above the detection threshold tend to be longer than genes expressed below the threshold (**Figure 3**). Although expression values have been normalized by gene length (i.e., using RPKM), this observed characteristic is likely due to the fact that short sequence reads are more likely to map to longer genes. Similarly, genes expressed above the threshold tend to contain a larger number of exons (**Figure 4**).

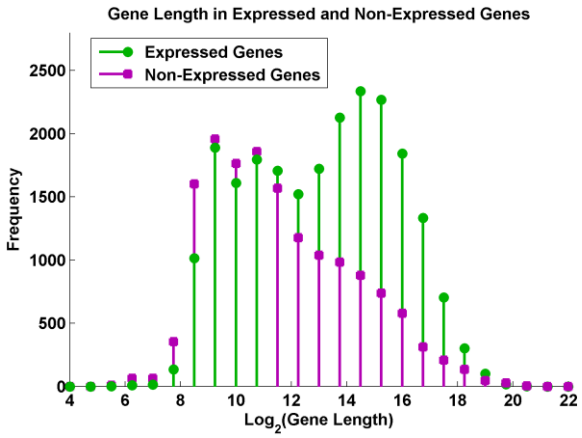

Figure 3. Distribution of gene length in expressed (green) and non-expressed (magenta) genes. Expressed genes tend to be longer.

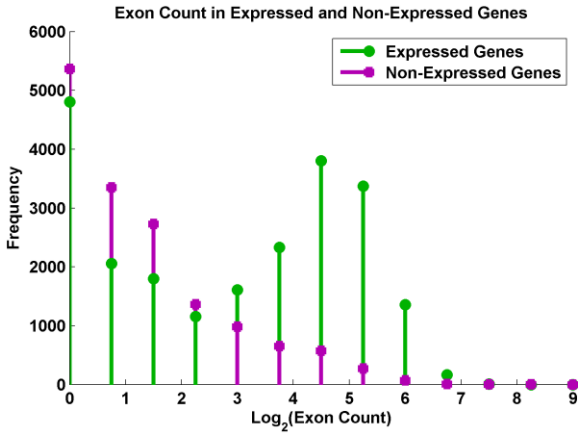

Figure 4. Distribution of exon count in expressed (green) and non-expressed (magenta) genes. Expressed genes tend to have more exons.

D. *Limitations and Future Investigations*

Although the experiments we conducted are limited, we observed some interesting characteristics of exonic, intronic, and intergenic mapping statistics, as well as correlations between expression "noise" and gene length. The results of this investigation may serve to guide future investigations. Specifically, future experiments may address the following limitations of this study. First, using only two different mapping pipelines, BWA and TopHat, we observed similar results. However, a comprehensive analysis of RNA-seq pipelines, including mapping, quantification, and normalization components should be examined to determine the effect of analysis pipeline on gene expression calls. Second, the specific choice of human genome annotation can largely impact downstream RNA-seq gene expression estimation [13]. Thus, a comprehensive analysis of the effect of genome annotation on the distributions of exonic, intronic, and intergenic region mapping is warranted. Third, we used only a single sample of one dataset in this study. Although Wagner et al. observed similar results across several datasets, a comprehensive analysis of various datasets from different sequencing platforms and with varying read depths may reveal other important factors for thresholding gene expression calls [4].

IV. Conclusion

We observed that the distributions of reads that map to exons, introns, and intergenic regions are distinct. Moreover, we can use these distributions to determine an approximate threshold for separating experimental or transcriptional noise from true gene expression. Such thresholding depends on assumptions about the genomic annotation. That is, we must assume that our knowledge of the genomic annotation is complete and that reads mapping to introns or intergenic regions are, in fact, the result of noise. Furthermore, we observed that two mapping pipelines, BWA and TopHat, produce very similar gene expression calling results. However, these pipelines are based on similar underlying algorithms. Finally, we observed that gene properties such as length and number of exons are correlated with the gene expression calling threshold. Overall, these preliminary results, and future investigations into gene expression noise, may be important in guiding us in the design of better RNA-seq experiments and data analysis pipelines to improve the accuracy of gene expression estimation.

References

[1] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics,* vol. 10, pp. 57-63, 2009.

[2] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann, "RNA sequencing reveals two major classes of gene expression levels in metazoan cells," *Molecular systems biology,* vol. 7, 2011.

[3] D. Hebenstreit, "Are gene loops the cause of transcriptional noise?," *Trends in Genetics,* vol. 29, pp. 333-338, 2013.

[4] G. P. Wagner, K. Kin, and V. J. Lynch, "A model based criterion for gene expression calls using RNA-seq data," *Theory in Biosciences,* vol. 132, pp. 159-164, 2013.

[5] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature methods,* vol. 5, pp. 621-628, 2008.

[6] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics,* vol. 25, pp. 1754-1760, 2009.

[7] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biol,* vol. 14, p. R36, 2013.

[8] D. Thierry-Mieg and J. Thierry-Mieg, "AceView: a comprehensive cDNA-supported gene and transcripts," *Genome biology,* vol. 7, p. S12, 2006.

[9] J. H. Phan, P.-Y. Wu, and M. D. Wang, "Improving the flexibility of RNA-Seq data analysis pipelines," in *Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on*, 2012, pp. 70-73.

[10] S. Anders. (2010). *HTSeq: Analysing high-throughput sequencing data with Python*. Available: http://www-huber.embl.de/users/anders/HTSeq/

[11] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and SAMtools," *Bioinformatics,* vol. 25, pp. 2078-2079, 2009.

[12] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research,* vol. 35, pp. D61-D65, 2007.

[13] P.-Y. Wu, J. H. Phan, and M. D. Wang, "Assessing the impact of human genome annotation choice on RNA-seq expression estimates," *BMC Bioinformatics,* vol. 14, p. S8, 2013.