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

Benchmarking RNA-Seq Quantification Tools

Raghu Chandramohan, Po-Yen Wu, *Student Member, IEEE*,John H. Phan, *Member, IEEE*, and May D. Wang, *Senior Member, IEEE*

***Abstract*—RNA-Seq, a deep sequencing technique, promises to be a potential successor to microarraysfor studying the transcriptome. One of many aspects of transcriptomics that are of interest to researchers is gene expression estimation. With rapid development in RNA-Seq, there are numerous tools available to estimate gene expression, each producing different results.However, we do not know which of these tools produces the most accurate gene expression estimates. In this study we have addressed this issue using Cufflinks, IsoEM, HTSeq, and RSEM to quantify RNA-Seq expression profiles. Comparing results of these quantification tools, we observe that RNA-Seq relative expression estimates correlate with RT-qPCR measurements in the range of 0.85 to 0.89, with HTSeq exhibiting the highest correlation. But, in terms of root-mean-square deviation of RNA-Seq relative expression estimates from RT-qPCR measurements, we find HTSeq to produce the greatest deviation. Therefore, we conclude that, though Cufflinks, RSEM, and IsoEM might not correlate as well as HTSeq with RT-qPCR measurements, they may produce expression values with higher accuracy.**

I. Introduction

Transcriptomic research has been prevalent in the past two decades. Data derived from RNA-Seq, a specialized protocol using deep sequencing technology, has been useful in analyzing the transcriptomein recent years. Earlier technologies used to study the transcriptome include probe-based sequencing methods and hybridization-based microarray methods. RNA-Seq offers distinct advantages over these methods. For example, prior knowledge of existing genomic sequences is not needed to detect transcripts. RNA-Seq also distinctly reveals sequence variations due to its low background noise[1, 2] and minimal cross-hybridization errors. The throughput with which the entire transcriptome canbe studied with RNA-Seq cannot be matched by any other technology at present.

One of many applications of RNA-Seq is the study of gene expression profiles. In an RNA-Seq experiment, expression profiles are indirectly inferred from sequence coverage, or the number of sequence reads that align to a particular region of the transcriptome. Numerous tools have been developed to quantify the expression profiles. However, it is not clear which bioinformatics tool is the most accurate for RNA-Seq expression quantification.

RNA-Seq quantification is challenging due to data properties such as effectivegene length and read length [3]. Moreover, effective normalization is necessary to compare inter-sample expression profiles [1,4,5]. Usually, quantification tools can effectively estimate gene expression. However, isoform expression estimation (for alternatively spliced genes) is more difficult as reads from isoforms of a gene can map to common exonic regions, increasing the complexity to identify the origin of the read with respect to an isoform [6]. Thus, the source of these ambiguously mapped reads is difficult to infer. A simple way to deal with ambiguously mapped reads is to simply discard them and keep only the uniquely mapped reads. More complex quantification tools attempt to resolve ambiguously mapped reads by using maximum likelihood estimation [7-9]. However, maximum likelihood estimatescan be inaccurate for low-expressing genes. In such case, Bayesian approaches may be more reliable [6].

To provide a guideline for selecting RNA-Seq quantification tools, we compare variations in expression estimates when different tools are applied in a typical RNA-Seq pipeline and investigate the cause of these variations.

II. Methodology

We examine RNA-Seq quantification tools within a typical workflow where reads generated by the sequencing machine are first mapped to areferenceassembly. A single alignment tool is applied since our objective is to assessvariations in gene expression estimates among different quantification tools. Alignment outputs undergo various preprocessing stages to conform tothe requirements of each quantification tool. The quantification tools then estimate gene expression and/or is form expression(Figure 1).

A. Datasets

Single-readIllumina data was downloaded from the publicly available NCBI SRA repository (accession numbers: SRX003926 and SRX003927). SRX003926contains mixed human brain sample (sample A) with four technical replicates andSRX003927contains mixed human cell lines (sample B) with three technical replicates. These datasets were used in the MAQC (MicroArray Quality Control) project[10] and were

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.

R. Chandramohan is with the School of Biology, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: rchandramohan7@gatech.edu).

P.-Y. Wu is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: pwu33@gatech.edu).

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).

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

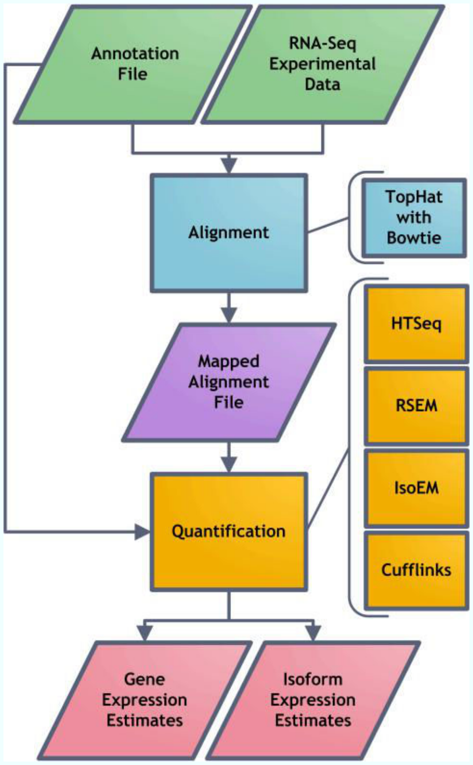

Figure 1.Expression quantification workflow.

chosen because of the availability of corresponding TaqMan RT-qPCR measurements with four replicates for sample A and four replicates for sample B(GSE5350 in NCBI's Gene Expression Omnibus repository). The reads in this dataset were generated by Illumina's first generationhigh-throughput sequencing platform, the Illumina Genome Analyzer. The reads are 36bp in length with a library fragment length of approximately 200.

The reference assembly used for this study was downloaded from Ensembl(GRCh37 release 67) along with the GTF (gene transfer format) file for the corresponding reference assembly. The GTF file was processed to contain information only on the 24 main chromosomes (chromosomes 1-22, X, and Y) and the mitochondrial DNA.The first column entries in the GTF file must match the chromosome names in the genome fasta file as they are matched in some of the quantification tools like Cufflinks.

B. *Sequence Alignment*

We used TopHat (v.2.0.6),a spliced alignment tool, to align the RNA-Seq reads to the Ensemble reference genome assembly [11].We ran TopHat using the ultrafast short read mapping program Bowtie[12].On an average 68% of the reads aligned to the reference.

C. *Expression Quantification*

Cufflinks estimates expression profiles using a statistical model in which the probability of observing each fragment is a linear function of the expression level of the transcripts from which it could have originated. In the case of paired-end reads, Cufflinks, like most quantification tools we are studying, makes use of fragment length distribution for more accurate assignment of a fragment to the transcript[9].

Preprocessing the data for Cufflinks is not necessary because TopHat produces alignment files as required by Cufflinks (v.2.0.2). Cufflinks requires a GTF file and a BAM file as inputs for quantification and outputs the gene-level and isoform-level expression estimates. The FPKM (fragments per kilobase of exon per million fragments mapped) normalization method is applied.

HTSeq uses a naive count-based approach for expressionestimation. The htseq-count script allows the user to choose how reads assigned to the corresponding gene from a list of three modes. These modes correspond to the overlap of features in the alignment:"union", "intersection-strict", and "intersection-nonempty".

HTSeq(v.0.5.3p9) uses a SAM file for quantification along with the GTF file (in case the alignment is present in BAM format, use samtools (v.0.1.18) to convert it to SAM format). The parameters for HTSeq were modified to conform to our dataset. The mode used in this study is the "intersection-nonempty" mode and the default strand-specific assay flag was turned off. HTSeq outputs counts of only reads aligned to genes but not the counts of reads involved in a particular gene's isoforms, i.e.,the gene is considered to be a union of all exons.

RSEM (RNA-Seq by Expectation Maximization. v.1.2.1) can be used to estimate expression levels of genes and their isoforms using two scripts: rsem-prepare-reference and rsem-calculate-expression[7]. The first step involves running the rsem-prepare-reference script, which essentially parses the genome fasta file into transcripts as specified in the GTF file. The output of this script is a reference file which is used to run the rsem-calculate-expression script for estimating expression levels. The rsem-calculate-expression script is responsible for aligning the reads to the transcripts and also for estimating the expression levels. The script uses Bowtie[12] to perform the alignment (Bowtie parameters were matched to that of our reference alignments).Once the read and the reference file, prepared from running rsem-prepare-reference, are supplied to the script,we obtain the expression estimates of genes and their isoforms in terms of count, TPM, and FPKM.

IsoEM, like RSEM, is based on the expectation maximization algorithm. This two-step algorithm uses the weights calculated considering the insert size distribution, base quality scores, and strand informationto calculate the expected number of reads that correspond to a particular genomic region. This quantifier was designed to estimate expressions from reads aligned to the transcriptome[8].

IsoEM (v.1.1.1) has two mandatory arguments: the GTF annotation file and library fragment length distribution. IsoEM requires the SAM alignment file to be sorted according to the read name (i.e. first column of the SAM file). The output of gene and isoform expression estimates is normalized using FPKM.

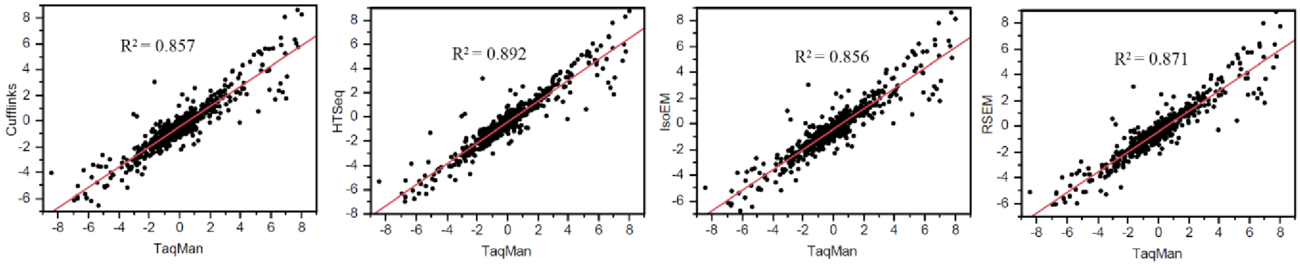

Figure 2. Correlation between relative gene expression estimates of TaqMan RT-qPCR and RNA-Seq FPKM measurementsfrom different quantification tools.

D. Evaluation Metrics

To maintain consistency, FPKM was selected as the normalization method for all RNA-Seq quantification tools in the study.The raw RNA-Seq (FPKM) and RT-qPCR measurements are both in a linear scale.

To evaluate RNA-Seq quantification tools, we compare RNA-Seq relative gene expression measurements to that of RT-qPCR. For each RNA-Seq quantification tool and for RT-qPCR, we averaged the gene expression estimates for each sample across technical replicates. Then we $\log_2$ transformed the ratio of the sample A mean to the sample B mean. The relative gene expression estimate is

$$LR_{gene} = log_2\left(\frac{\sum_{i=1}^{4} A_{i,gene}/4}{\sum_{i=1}^{4} B_{i,gene}\ /4}\right) \qquad (1)$$

where $A_{i,gene}$ is the gene expression estimate of replicate i from sample A and $B_{i,gene}$ is the gene expression estimate of replicate i from sample B.

To understand the relationship between the RT-qPCR values and the RNA-Seq FPKM values, we performed a bivariate analysis. First, the relative gene expression estimates for each RNA-Seq quantification was tabulated. Another table with the RT-qPCR relative gene expression estimates was tabulated as well. The tables were joined based on the gene name. To compare commonly detected genes, if the gene expression was zero for all the replicates in a sample for any quantification tool, then such gene was not considered for the comparison. This resulted in 531 commonly expressed genes.

Relative expression values were plotted in a scatter plot pair-wise: RT-qPCR vs. each RNA-Seq quantification tool.Alinear line was fit to calculate the coefficient of determination (R^2). Since we applied an ordinary least squares regression model to fit the data, the R^2 is equivalent to the square of the Pearson correlation coefficient.

To further understand the relationship between the relative expression estimates from the RT-qPCR and the RNA-Seq data, we calculated the root-mean-square deviation (RMSD).RMSD for each RNA-Seq quantification tool was calculated pair-wise against relative RT-qPCR gene expression estimates as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}\left(LR_{RNA-Seq,i} - LR_{RT-qPCR,i}\right)^2}{N}} \qquad (2)$$

where $LR_{RNA\text{-}Seq,i}$ is the relative gene expression estimate of gene i from an RNA-Seq tool and $LR_{RT\text{-}qPCR,i}$ is the relative gene expression estimate of gene i from RT-qPCR.

III. Results and Discussion

A. RT-qPCR vs. RNA-Seq

From the bivariate analysis (Figure2), we observe that HTSeq relative expression estimates have the highest correlation with RT-qPCR values. Inferring from R^2 values, wecan also see that the performance of each tool is fairly consistent in terms of estimating relative expression values. From this analysis, we conclude that there is very little difference in using one quantification tool over another. However, the root-mean-square deviation of HTSeq is larger than that of the other quantification tools(all of which seem to have similar RMSD values as shownin Figure 3).

B. Gene and Isoform Detection

In our annotation file we have 53,893 genes and 182,921 isoforms. Counting the number of expression estimates that are not zero, we observe that Cufflinks, HTSeq, IsoEM, and RSEM on average were able to detect 28259, 20721, 22035, and 22961 genes respectively (Figure 4). The isoform detection plot also has the same trend with Cufflinks detecting the most isoforms (Figure 5). Cufflinks, RSEM, and IsoEM identify on average 127819, 57446, 54440 isoforms, respectively.

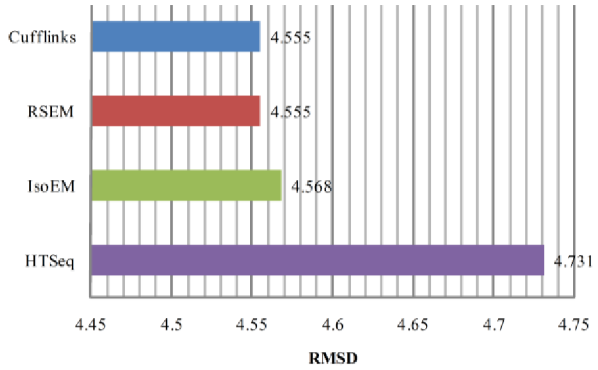

Figure 3. Root-mean-square deviation (RMSD) of relative gene expression estimates of RNA-Seq quantification tools against TaqMan RT-qPCR.

C. Computational Cost

IsoEM is the fastest quantification tool in the study (Table 1). Using a Java platform, it manages to make the most of the available resources by occupying all available CPUs.For Cufflinks and RSEM, it is also possible to speed up the quantification process with the multi-thread setting. HTSeq does not support multi-thread computing.

TABLE 1 COMPUTATION TIME FOR EXPRESSION ESTIMATION

	Cufflinks	HTSeq	RSEM	IsoEM
Run Time (1 core)	173m 15s	4m 2s	49m 13s	-
Run Time (15 cores)	15m 21s	-	12m 53s	0m 34s

IV. CONCLUSION

In this study, we assessed four commonly used RNA-Seq quantification tools. By comparing relative expression estimates, we observe that all tools are highly correlated with Taqman RT-qPCR values, which are considered to be the current "goldstandard" assay. Among these tools, HTSeq has the highest correlation with R^2=0.89. But we also find that HTSeq exhibits the highest deviation from RT-qPCR when we perform RMSD analysis in terms of relative expression levels.

HTSeq is a fast and easy to handle tool and its results correlate well as it has a better linear fit with RT-qPCR expression. However, we observe that there is a lateral shift in the relative expression estimates of HTSeq asinferred from Figure 3. Though Cufflinks, RSEM, and IsoEM might not correlate as well as HTSeq with RT-qPCR expression values, they may provide the user with more accurate expression values.Because of its computational efficiency, HTSeq can be used as a tool for preliminary data analysis or for quick assessment of relative expression estimates. We also observe that Cufflinks consistently detected more genes and isoforms than any other tool used in the study.

V.FUTURE WORK

RSEM, IsoEM, and Cufflinks include bias correction options. For future experiments, it would be interesting to investigate whether enabling bias correction would affect the performance of these tools in terms of correlation and RMSD when using RT-qPCR as the reference. Future benchmark studies that include more quantification tools and more robust performance metrics may provide further guidance for selecting RNA-Seq data analysis pipelines.

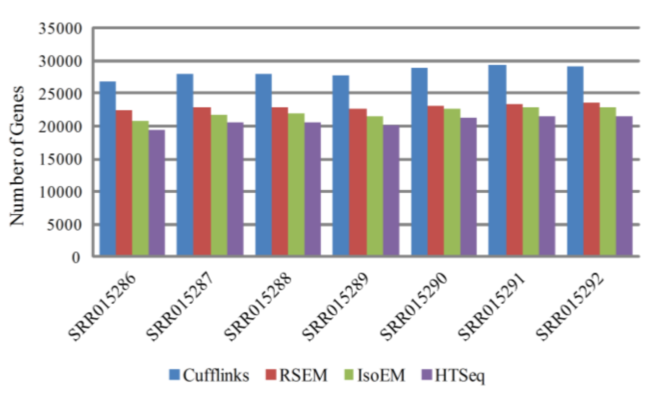

Figure 4. Number of genes detected in each sample by different quantification tools.

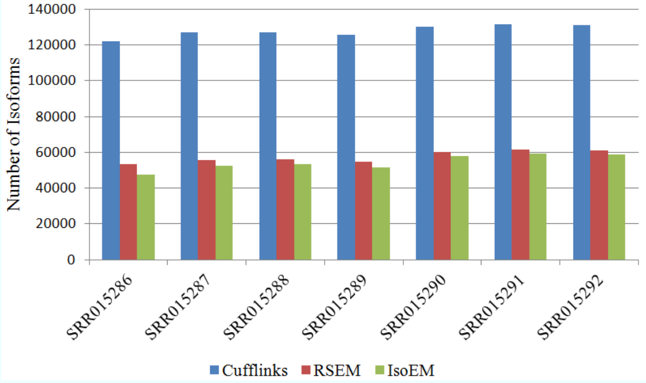

Figure 5. Number of isoforms detected in each sample by different quantification tools.

REFERENCES

[1] 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.

[2] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome research,* vol. 18, pp. 1509-1517, 2008.

[3] S. Lee, C. H. Seo, B. Lim, J. O. Yang, J. Oh, M. Kim, S. Lee, B. Lee, C. Kang, and S. Lee, "Accurate quantification of transcriptome from RNA-Seq data by effective length normalization," *Nucleic acids research,* vol. 39, pp. e9-e9, 2011.

[4] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science,* vol. 320, pp. 1344-1349, 2008.

[5] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature,* vol. 456, pp. 470-476, 2008.

[6] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature methods,* vol. 8, pp. 469-477, 2011.

[7] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics,* vol. 12, p. 323, 2011.

[8] M. Nicolae, S. Mangul, I. Măndoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from RNA-Seq data," *Algorithms in Bioinformatics,* pp. 202-214, 2010.

[9] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature biotechnology,* vol. 28, pp. 511-515, 2010.

[10] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, and K. Y. Lee, "The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature biotechnology,* vol. 24, pp. 1151-1161, 2006.

[11] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics,* vol. 25, pp. 1105-1111, 2009.

[12] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol,* vol. 10, p. R25, 2009.