Published in final edited form as:

IEEE Int Workshop Genomic Signal Process Stat. 2013 November ; 2013: 15–18. doi:10.1109/GENSIPS.2013.6735918.

An Approach for Assessing RNA-seq Quantification Algorithms in Replication Studies

Po-Yen Wu,

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA, pwu33@gatech.edu

John H. Phan, and

Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA, jhphan@gatech.edu

May D. Wang

Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA, maywang@bme.gatech.edu

Abstract

One way to gain a more comprehensive picture of the complex function of a cell is to study the transcriptome. A promising technology for studying the transcriptome is RNA sequencing, an application of which is to quantify elements in the transcriptome and to link quantitative observations to biology. Although numerous quantification algorithms are publicly available, no method of systematically assessing these algorithms has been developed. To meet the need for such an assessment, we present an approach that includes (1) simulated and real datasets, (2) three alignment strategies, and (3) six quantification algorithms. Examining the normalized root-mean-square error, the percentage error of the coefficient of variation, and the distribution of the coefficient of variation, we found that quantification algorithms with the input of sequence alignment reported in the transcriptomic coordinate usually performed better in terms of the multiple metrics proposed in this study.

I. INTRODUCTION

A major branch of next-generation sequencing technology is RNA sequencing (RNA-seq), deemed a revolutionary tool for transcriptomic research [1]. Featuring its base pair resolution and high throughput, RNA-seq has prompted research that identifies and examines comprehensive transcriptomic profiles, novel transcriptomic elements, gene regulation mechanisms, and gene structures (e.g., exon-intron boundaries and alternative splicing patterns).

A single RNA-seq run can produce millions of sequence reads. Focusing on RNA-seq transcriptome profiling, a standard data analysis pipeline includes sequence reads alignment,

*Corresponding Author: Contact information for the corresponding author: maywang@bme.gatech.edu, Phone: 404-385-2954, Fax: 404-894-4243, Address: Suite 4106, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332, USA.

expression quantification and normalization, and various downstream inferences. Expression quantification algorithms attempt to uniquely assign sequence reads to transcripts or genes. However, this is a challenging process since (1) alternative spliced isoforms of a gene share exons and (2) sequence reads may map to multiple loci due to high similarity among some genomic regions [2]. These challenges result in read assignment uncertainty.

To address these challenges, researchers have developed a number of quantification algorithms. Several (e.g., HTSeq [3] and BEDTools [4]) simplify the problem by counting the number of sequence reads aligned to a targeted gene with the predefined gene model, and others (e.g., Cufflinks [5] and RSEM [6]), which potentially resolve the multi-mapping issue, build upon the Poisson model and probabilistically assign sequence reads to transcripts or genes. Quantification algorithms can be classified into two categories in terms of the input information from the sequence alignment. Some require sequence alignment to be reported in genomic coordinates while others require sequence alignment to be reported in transcriptomic coordinates.

Because of the various characteristics of quantification algorithms, we propose an approach for systematically assessing the performance of these algorithms using both simulated and real datasets. The remaining sections are organized as follows: Section II summarizes experimental design and data analysis steps, Section III presents evaluation metrics for both simulated and real datasets, and finally, Section IV concludes our findings from this study.

II. METHODOLOGY

The workflow of this study is shown in Figure 1. Data sources include one simulated dataset and two real datasets. We align the sequence reads of these datasets to the UCSC hg19 genome using TopHat [7] either with or without external genome annotation information (GTF file) and to the RefSeq transcriptome using Bowtie [8]; then we quantify the sequence alignments reported in the genomic coordinates by Cufflinks [5], HTSeq [3], and MISO [9] and those reported in the transcriptomic coordinates by RSEM [6], eXpress [10], and MMSEQ [11]. We compute gene and transcript read counts from the simulated dataset as the ground truth to investigate the performance of the expression estimation of various quantification algorithms.

A. Dataset

To create the simulated dataset, we use Flux Simulator [12] to produce a gene expression profile and then apply the same expression profile to generate three technical replicates with the read length of 2x100 base pairs (bp) for each sequencing depth (10, 50, and 100 million read pairs). In Flux Simulator, multiple library preparation and sequencing steps introduce variations among the technical replicates.

We also download two publicly available datasets from the NCBI SRA repository. The first dataset, which contains three thrombin-treated samples and two control samples, studied the effect of thrombin on endothelial function (SRA accession: SRP008482 [13]). Using the Illumina HiSeq 2000 platform, the authors sequenced these samples, each of which has around 50 million 2x100 bp read pairs. The second dataset, which contains four treatments,

each with two replicates, investigated the off-target effect of EGFP siRNA and pro-siRNA in the HeLa-d1EGFP cell line (SRA accession: SRP018552 [14]). Using the Illumina Genome Analyzer II platform, the authors sequenced these samples, each with about 25 million 50 bp single-ended reads.

B. Sequence Alignment

To prepare necessary input information for sequence alignment, we download the hg19 reference genome and the RefSeq GTF file, which annotates 23,801 genes and 42,671 transcripts, from UCSC Genome Browser. We identify three possible sequence alignment scenarios that vary according to the presence of external GTF information and the reporting strategies of alignment tools (i.e., reporting aligned sequence reads in either genomic or transcriptomic coordinates). In our study, we include two sequence alignment tools, TopHat and Bowtie. TopHat reports alignments only in genomic coordinates. Without providing external GTF information, it uses a spliced technique to directly align sequence reads to the reference genome (notation: TopHat G). However, if external GTF information is available, we restrict TopHat so that it aligns sequence reads only to transcriptome sequences extracted from the genome (notation: TopHat T). We use Bowtie to produce alignment files with transcriptomic coordinate reporting (notation: Bowtie).

C. Expression Quantification and Normalization

We apply Cufflinks, HTSeq, and MISO to quantify gene and transcript expression for the sequence alignment reported in genomic coordinates. The Cuffdiff 2 program in the Cufflinks package produces gene and transcript expression estimates in terms of read count and FPKM [15] (i.e., fragments per kilobase of exon per million fragments mapped) estimates. The htseq-count program in the HTSeq package generates read count estimates only for genes, while MISO provides read count estimates for both genes and transcripts. For sequence alignment reported in transcriptomic coordinates, we use RSEM, eXpress, and MMSEQ to quantify gene and transcript expression. These three quantification algorithms are able to quantify both genes and transcripts and produce both read count and FPKM estimates. In addition, RSEM provides in-house TPM (i.e., transcripts per million) estimates. For read count estimates, we apply the trimmed mean of M-values normalization method (TMM) [16] to eliminate the effect of normalization when computing some assessment metrics.

D. Assessment Metrics

To evaluate the performance of quantification algorithms, we propose two assessment metrics that incorporate ground truth from the simulated dataset and one assessment metric for the two real datasets. For the simulated dataset, since the ground truth (the true gene and transcript counts) is available, we first compute the normalized root-mean-square error (RMSE) between estimated counts and true counts given a sequencing depth and a quantification strategy. Figures 2 and 3 include nine and seven quantification strategies for gene- and transcript-level analyses, respectively. Since RMSE is not scale-invariant, to adjust the sequencing depth effect, we divide RMSE by 10 for 100M cases, by 5 for 50M cases, and by 1 for 10M cases. The error bar for each case results from the inclusion of three replicates. Using TMM-normalized expression estimates, we calculate the gene-wise

coefficient of variation (CV) across all replicates for the simulated dataset that serves as the ground truth measurement for replicate variation. We examine the percentage error when comparing the gene-wise CV from various quantification strategies with that from the simulated dataset. We count the number of genes or transcripts with percentage errors of the CV between 0% and 5%, 0% and 10%, 0% and 15%, and so on up to 100%. The same technique applies to the transcript-wise CV analysis.

Both real datasets have technical replicates for each condition, such as a treatment or control. Assuming that the variation between technical replicates is negligible, the gene-wise CV should be small for accurate quantification strategies. Accounting for normalized expression estimates, such as FPKM, TMM, and TPM, we include 15 and 13 quantification strategies for gene- and transcript-level analyses, respectively. For each quantification strategy, we compute the condition-wise CV for each gene and then summarize these CVs into the gene-wise CV. Box plots in Figures 4 and 5 show that the distribution of gene-wise CVs varies according to the quantification strategy. The same technique applies to the transcript-wise CV analysis.

III. RESULTS AND DISCUSSION

As described in Section II.D, we proposed three assessment metrics to investigate the performance of quantification algorithms. Figures 2A and 3A show that normalized RMSE varies according to the quantification strategy and the sequencing depth. A lower normalized RMSE indicates smaller deviation between estimated and true expression. Note that the high sequencing depth (e.g., 100M read pairs) does not improve significantly in the RMSE metric. Figures 2B and 3B demonstrate that the number of genes/transcripts falling within predefined percentage errors of the CV differs according to the quantification strategy. Curves closer to the upper-left corner of the figure indicate a closer measure of variation to the ground truth. Figures 4 and 5 use box plots to illustrate that the distribution of the gene-wise or transcript-wise CV also varies according to the quantification strategy in the real RNA-seq datasets. Lower CV indicates a smaller variation among the technical replicates.

Our results suggest that genome alignment with external GTF information (TopHat T) results in smaller deviations and lower CVs than that without GTF information (TopHat G). This observation holds for Cufflinks, HTSeq, and MISO with the three metrics, shown in Figures 2 to 5. Without external GTF information, TopHat searches for candidate mapping locations in the entire genome, which may increase the odds that a sequence read aligns to an incorrect location. When focusing on expression-based applications, if the GTF file is available, we recommend using GTF-guided alignment to speed up the computational process and increase the expression estimation accuracy.

Another key finding is that quantification algorithms based on transcriptome alignment tend to result in smaller deviations than genome alignment when using simulated expression as the reference, with eXpress being an exception (Figures 2A and 3A). The analysis of real RNA-seq datasets shows that transcriptome alignment also results in a comparable or lower CV across technical replicates than genome alignment (Figures 4 and 5). The difference becomes more significant in the transcript-wise CV analysis. However, for a percentage

error of the CV, quantification algorithms based on genome alignment, such as Cufflinks and HTSeq, outperform all other quantification algorithms, with RSEM being tied with the Cufflinks TopHat G quantification strategy (Figures 2B and 3B). Based on these observations, we infer that quantification algorithms based on genome alignment may cause greater absolute deviation but maintain relative variation among technical replicates. Even though the observations are inconsistent, quantification algorithms based on transcriptome alignment generally perform better in terms of the three metrics, with RSEM outperforming others in all cases. The potential reason is that transcriptome alignment preserves direct mapping information about each transcript while genome alignment does not. Thus, while genome alignment requires additional effort to assign reads to one of the transcripts of a gene, transcriptome alignment requires only the identification of chimeric mappings in multi-mapping cases.

Our study identified several outlying cases. For one, with three evaluation metrics, the MISO package performed worse (i.e., a greater deviation and a higher CV) in most of the cases, which may be due to MISO ignoring too much information such as (1) read pairs mapped to the same strand and (2) reads that have no paired mate. Moreover, Bayesian-based algorithms (e.g., MISO and eXpress) tended to introduce higher variation in expression estimates. We hypothesize that Bayesian-based approaches heavily depend on the prior distribution. Thus, if algorithms cannot converge within a predefined number of iterations, the expression estimates may be only a suboptimal solution. This hypothesis is partially confirmed by the high sensitivity of eXpress to the “forgetting factors” discussed in [10].

IV. CONCLUSION

We proposed an approach that includes three alignment strategies and six quantification algorithms for assessing RNA-seq quantification algorithms in replication studies using both simulated and real RNA-seq datasets. By examining multiple metrics, we found that the TopHat T alignment strategy always outperforms the TopHat G alignment strategy. In addition, quantification algorithms using sequence alignment reported in transcriptomic coordinates usually resulted in a smaller deviation and a lower CV, with eXpress being an exception. Furthermore, RSEM consistently performed better compared with other quantification algorithms. Our approach is useful for comprehensively assessing new RNA-seq quantification algorithms. Based on our results, we suggest using quantification algorithms with transcriptome alignment. If only genome alignment is possible and external GTF information is available, incorporating GTF information yields higher performance. To reinforce our observations in this study, we plan to include more real datasets in future studies.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009 Jan.10:57–63. [PubMed: 19015660]
2. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods.* 2011 Jun.8:469–477. [PubMed: 21623353]
3. Anders S. HTSeq: Analysing high-throughput sequencing data with Python.

4. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010 Mar 15;26:841–842. [PubMed: 20110278]
5. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 2010 May;28:511-U174.
6. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics*. 2011 Aug 4;12.
7. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009 May 1;25:1105–1111. [PubMed: 19289445]
8. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009; 10:R25. [PubMed: 19261174]
9. Katz Y, Wang ET, Airolidi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*. 2010 Dec;7:1009-U101. [PubMed: 21057496]
10. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*. 2013 Jan;10:71-U99. [PubMed: 23160280]
11. Turro E, Su SY, Goncalves A, Coin LJM, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*. 2011; 12
12. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, Sammeth M. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*. 2012 Nov;40:10073–10083. [PubMed: 22962361]
13. Zhang LQ, Cheranova D, Gibson M, Ding SH, Heruth DP, Fang DY, Ye SQ. RNA-seq Reveals Novel Transcriptome of Genes and Their Isoforms in Human Pulmonary Microvascular Endothelial Cells Treated with Thrombin. *PLoS One*. 2012 Feb 16;7.
14. Huang LF, Jin JM, Deighan P, Kiner E, McReynolds L, Lieberman J. Efficient and specific gene knockdown by small interfering RNAs produced in bacteria. *Nature Biotechnology*. 2013 Apr; 31:350.
15. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008 Jul;5:621–628. [PubMed: 18516045]
16. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010; 11

Figure 1.
The experimental design and workflow for assessing the performance of RNA-seq quantification algorithms.

Figure 2.

Gene-level performance metrics for simulated RNA-seq data, including (A) the root-mean-square error and (B) the percentage error of the coefficient of variation.

Figure 3.

Transcript-level performance metrics for simulated RNA-seq data. Metrics are the same as those in Figure 2.

Figure 4.

The distribution of the coefficient of variation varies according to the quantification strategy for both gene- and transcript-level expression estimates using the SRP008482 dataset.

Figure 5.

The distribution of the coefficient of variation varies according to the quantification strategy for both gene- and transcript-level expression estimates using the SRP018552 dataset.