Published in final edited form as:

IEEE Int Workshop Genomic Signal Process Stat. 2012 December ; 2012: 70–73. doi:10.1109/GENSIPS.2012.6507729.

Improving the Flexibility of RNA-Seq Data Analysis Pipelines

John H. Phan,

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

Po-Yen Wu, and

Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA,
USA, leoboki@gmail.com

May D. Wang*

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

Abstract

Accurate quantification of gene or isoform expression with RNA-Seq depends on complete knowledge of the transcriptome. Because a complete genomic annotation does not yet exist, novel isoform discovery is an important component of the RNA-Seq quantification process. Thus, a typical RNA-Seq pipeline includes a transcriptome mapping step to quantify known genes and isoforms, and a reference genome mapping step to discover new genes and isoforms. Several tools implement this approach, but are limited in that they force the use of a single mapping algorithm at both the transcriptome and reference genome mapping stages. The choice of mapping algorithm could affect quantification accuracy on a per-dataset basis. Thus, we describe a method that enables the merging of transcriptome and reference genome mapping stages provided that they conform to the standard SAM/BAM format. This procedure could potentially improve the accuracy of gene or isoform quantification by increasing flexibility when selecting RNA-Seq data analysis pipelines. We demonstrate an example of a flexible RNA-Seq pipeline by assessing its potential for novel isoform discovery and by validating its quantification performance using qRT-PCR.

I. INTRODUCTION

RNA-Seq, an application of next-generation sequencing (NGS), quantifies mRNA expression by mapping sequenced mRNA fragments, or “reads” to a transcriptome or reference genome [1]. Transcriptome references contain the sequences of gene isoforms (i.e., gene splice variants) assembled from exons defined in various genomic annotations. Mapping reads directly to a reference genome may not identify transcriptome sequences that span introns. However, direct reference genome mapping enables identification of previously unknown exons and novel splice junctions, possibly leading to discovery of novel isoforms. Many genomic annotations exist, owing to the fact that a complete transcriptome definition

*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.

does not yet exist. Thus, many gene isoforms may still be undefined, leading to erroneous RNA-Seq quantification in two situations. First, reads from an unknown gene isoform may be incorrectly assigned to another gene isoform, resulting in incorrect estimation of isoform abundance [2]. Second, reads from an unknown gene isoform may be ignored, leading to incorrect estimation of gene expression. Fig. 1 is an example of the importance of mapping reads to a reference genome. The gray bars in Fig. 1 represent reads mapping to the transcriptome consisting of known exons (light blue bars). Mapping of reads to a reference genome, including introns (white bars), enables the discovery of exon variants (dark blue bar) as well as novel splice junctions (red bar). Detection of such novel exon variants and novel isoforms is important for accurate estimation of gene expression and isoform abundance.

We describe a general procedure for mapping reads to a transcriptome as well as to a reference genome in order to capture the information provided in the transcriptome annotation while enabling discovery of novel isoforms. This procedure is common among a number of tools including TopHat, OSA, and GSNAP [3–5]. However, these tools are limited in that the transcriptome and reference genome mapping stages are tied to specific mapping algorithms. Recent studies have shown that different mapping tools can produce highly variable results due to the fact that each tool incorporates a different set of algorithmic features [6]. Because of the diversity of NGS data acquisition platforms for RNA-Seq and the diversity of biological problems, the ability to choose algorithmic features at each stage of mapping is important for RNA-Seq quantification. Our mapping procedure enables flexibility in choosing mapping algorithms separately for each stage, while still merging the mapping results of these stages, in order to optimize RNA-Seq quantification on a per-application or per-dataset basis.

II. METHODS

A. RNA-Seq Data

We use an RNA-Seq dataset from the Sequence Read Archive (SRA) (accession number: SRP008482) to evaluate multiple RNA-Seq pipelines. This dataset includes three samples of human pulmonary microvascular endothelial cells (HMVEC-L) treated with thrombin for 6 hours; and two control samples [7]. Samples were acquired using the Illumina HiScanSQ to produce approximately 50 million read pairs per sample with read lengths of 101 base pairs. The dataset includes qRT-PCR validation results for only three genes, which we use for validation of RNA-Seq quantification.

B. RNA-Seq Mapping Pipelines

We compare two standard RNA-Seq mapping pipelines to the flexible RNA-Seq mapping pipeline by mapping and quantifying samples from the public dataset and examining mapping statistics as well as accuracy of differential gene expression. Specifically, we use TopHat (v2.0.4) and OSA for the standard pipelines, which map reads to a reference genome and transcriptome as illustrated in Fig. 2. Reads are mapped to a reference genome using a transcriptome annotation as guidance with regard to known splice junctions. For comparison, we use the flexible spliced mapping pipeline with Bowtie2 [8]. The flexible

pipeline, first, maps reads to a transcriptome reference, and then maps previously unmapped reads to a reference genome, allowing flexibility in selecting mapping tools (Fig. 3). These two stages are finally merged into one mapping result. For example, we can use (1) Bowtie2 for the transcriptome mapping stage because of its computational efficiency, and (2) TopHat for the reference genome stage because of its ability to identify potentially novel splice isoforms [4]. In this initial study, we use Bowtie2 for both stages as a simple example to show that it is still capable of identifying potentially novel isoforms.

1) Transcriptome and Human Genome References—For both the standard and flexible pipelines, we use the UCSC Human Genome (hg19) reference that includes sequences of the main chromosomes, unmapped contigs, and random contigs, which belong to chromosomes, but are unmapped [9]. We use a transcriptome annotation from Ensembl (GRCh37, release 68) [10]. The annotation file (GTF) was modified and the transcriptome sequence was extracted from the reference genome as follows. First, the sequence names in the annotation file were translated from the NCBI to the UCSC naming convention (e.g., ‘1’ => ‘chr1’ and ‘X’ to ‘chrX’). Second, the annotation file was simplified by removing all additional fields except those describing ‘transcript_id’, ‘gene_id’, or ‘gene_name’ (using the Cufflinks gffread utility) [11]. Third, all non-exon rows were removed. Finally, the transcriptome sequences were extracted from the reference genome using the TopHat gtf_to_fasta utility [4].

2) Translation of Mapping Coordinates—In order to combine the transcriptome mapping results with the reference genome mapping results in the flexible pipeline, coordinates of reads mapping to the transcriptome must be translated to genome coordinates. Translation of this mapping from transcriptome coordinates to genome coordinates requires (1) translation of the mapping coordinate and (2) translation of the CIGAR code to indicate the presence of introns. Fig. 4 is an example of such a translation. The example read maps at a coordinate of 2900 base pairs (bp) relative to the beginning of the transcript and spans three exons (i.e., exons 18, 19, and 20). Moreover, the CIGAR (Concise Idiosyncratic Gapped Alignment Report) code of 101M indicates that all 101 bases of the read match (or mis-match) the template with no insertions or deletions. After translation, the mapping coordinate becomes 67626173 and the new CIGAR code becomes (33M 5673N 65M 3706N 3M), which includes two intron (‘N’) regions. All valid paired-end mappings to the transcriptome are translated to genomic coordinates and merged with all valid paired-end mappings to the genome, resulting in a SAM/BAM file that can be quantified with Cufflinks [11].

3) Quantification and Normalization—We quantify gene and isoform expression using Cufflinks [11]. Cufflinks can handle multiple-mapped reads (i.e., secondary alignments) using the “NH” tag in SAM/BAM files. Bowtie2 and TopHat alignment files are sorted by alignment coordinate and include the NH tag for identifying multiple-mapped reads. However, it is necessary to further process the OSA alignment files by adding the NH tags and sorting by alignment coordinate. We use the annotation-guided mode in Cufflinks to enable quantification of novel isoforms while retaining existing information from the known genome annotation. Cufflinks uses Fragment Per Kilobase of transcript per Million mapped

reads (FPKM) to normalize expression values. Gene and isoform expression values are comparable between samples after FPKM normalization [12].

C. Evaluation Metrics

We evaluate and compare the flexible and standard RNA-Seq spliced mapping pipelines by examining the ability of each pipeline to detect potentially novel isoforms and measuring quantification accuracy. For each pipeline, we count the number of isoforms detected by Cufflinks that are candidates for novel isoforms. The total number of novel isoforms roughly corresponds to the amount of reads mapping to un-annotated regions. Thus, these reads were assigned in the reference genome mapping stage. From this data, we can infer the ability of each pipeline to identify potentially novel gene isoforms. However, these isoforms must be verified prior to making any conclusions about the ability of a pipeline to identify novel isoforms.

The qRT-PCR information included in the RNA-Seq dataset, described in Section IIA, enables evaluation of the pipelines by comparing quantification results to a ground truth [7]. Assuming that the qRT-PCR validation of three genes (i.e., CELF1, FANCD2, and TRAF1) can be treated as a ground truth for gene expression, we evaluate the accuracy of the estimated fold-change of these genes for each RNA-Seq pipeline.

III. RESULTS AND DISCUSSION

Despite the fact that the flexible RNA-Seq mapping pipeline using Bowtie2 is not able to detect novel splice junctions as depicted in Fig. 1, this flexible pipeline detects slightly more potentially novel isoforms (i.e., isoforms that do not exactly match existing isoforms in the Ensembl transcriptome annotation) compared to the standard TopHat and OSA pipelines (Fig. 5). This trend is consistent for all five samples in the dataset. Although we do not know which of the candidate novel isoforms is a result of mapping error, it is interesting to note the variations that can occur from small changes in mapping strategy (e.g., the flexible Bowtie2 is very similar to the TopHat pipeline in that TopHat also uses Bowtie2).

Definitive validations of RNA-Seq pipeline performance should also depend on external data from methods such as qRT-PCR. Evaluation of the RNA-Seq mapping pipelines using the qRT-PCR ground truth shows variability among pipelines (Fig. 6). Among the three genes for which qRT-PCR data are available (i.e., CELF1, FANCD2, and TRAF1), RNA-Seq log₂ fold-change estimates for CELF1 and FANCD2 are close to that of qRT-PCR. TRAF1, on the other hand, exhibits much lower RNA-Seq estimated fold-change compared to qRT-PCR. One reason for this is the use of Cufflinks for quantification of expression values. Cufflinks estimates isoform abundance using transcript assembly [11]. Thus, if cufflinks cannot assemble transcripts that exactly match the gene annotation, it reports expression values for regions similar to the existing transcripts. For the CELF1 and FANCD2 genes, reported genomic regions are very close to the annotated genes. However, the closest reported match to the TRAF1 gene is longer by approximately 4000 bases, resulting in a bias in the FPKM estimate of TRAF1 expression. Considering only CELF1 and FANCD2, the average absolute deviation of RNA-Seq fold-change from qRT-PCR fold-change (i.e., absolute value of the difference between RNA-Seq fold-change and qRT-PCR fold-change) for the

Bowtie2, TopHat, and OSA pipelines are 0.28, 0.30, and 0.43, respectively. Thus, the flexible RNA-Seq pipeline using Bowtie2 appears to more accurately quantify differential gene expression. However, additional datasets (e.g., with more validated genes), quantification methods (e.g., other than Cufflinks), and mappers should be considered before making final conclusions.

IV. CONCLUSION

RNA-Seq quantification by mapping, first, to a transcriptome, then to a reference genome, is necessary for (1) estimation of gene and isoform expression and (2) discovery of novel genes and isoforms. Although multiple tools are capable of this two-stage quantification procedure, these tools are typically not flexible in terms of choice of mapping algorithm. We described a procedure to combine transcriptome and reference genome mapping results from different mapping tools, provided that these tools produce standard SAM/BAM output. Using an RNA-Seq dataset with qRT-PCR validation information, we compared a flexible RNA-Seq pipeline that uses Bowtie2 in the transcriptome and reference genome stages to two standard mapping tools: OSA and TopHat. Results suggest that the flexible Bowtie2 pipeline identifies more candidate novel isoforms and is slightly more accurate than TopHat and OSA in terms of estimating gene fold-change (i.e., with respect to qRT-PCR results). The analysis in this study only serves as an example of the potential of using this flexible RNA-Seq pipeline for optimizing sequence mapping on a per-dataset or per-application basis. An expanded study comparing more mapping tools and using more datasets is necessary to further explore the benefits of flexible RNA-Seq pipelines.

Acknowledgments

This research has been supported by grants from the Parker H. Petit Institute for Bioengineering and Bioscience (IBB), National Institutes of Health (NIH) (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), National Cancer Institute (NCI), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Georgia Research Alliance, Hewlett-Packard, and Microsoft Research.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; 10:57–63.
2. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*. 2009; 25:1026–1032. [PubMed: 19244387]
3. Hu J, Ge H, Newman M, Liu K. OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics*. 2012; 28:1933–1934. [PubMed: 22592379]
4. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25:1105–1111. [PubMed: 19289445]
5. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010; 26:873–881. [PubMed: 20147302]
6. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*. 2011; 27:2518–2528. [PubMed: 21775302]
7. Zhang LQ, Cheranova D, Gibson M, Ding S, Heruth DP, Fang D, Ye SQ. RNA-seq Reveals Novel Transcriptome of Genes and Their Isoforms in Human Pulmonary Microvascular Endothelial Cells Treated with Thrombin. *PloS one*. 2012; 7:e31229. [PubMed: 22359579]

8. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012
9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
10. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S. Ensembl 2012. *Nucleic acids research*. 2012; 40:D84–D90. [PubMed: 22086963]
11. 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; 28:511–515.
12. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008; 5:621–628. [PubMed: 18516045]

Fig. 1.
Discovery of Novel Gene Isoforms using RNA-Seq

Fig. 2.
Standard RNA-Seq Spliced Mapping Pipeline

Fig. 3.
Flexible RNA-Seq Spliced Mapping Pipeline

Fig. 4.
Translation of Transcriptome Mapping Coordinates

Fig. 5.
Number of Candidate Novel Isoforms Detected by Cufflinks Varies by RNA-Seq Pipeline

Fig. 6.
Comparison of RNA-Seq Mapping Pipelines to qRT-PCR Ground Truth