Published in final edited form as:

IEEE Glob Conf Signal Inf Process. 2014 December ; 2012: 1376–1379. doi:10.1109/GlobalSIP.2014.7032351.

The Impact of RNA-seq Alignment Pipeline on Detection of Differentially Expressed Genes

Cheng Yang,
Department of Biomedical Engineering, Georgia Institute of Technology, Emory University, and Peking University, Atlanta, Georgia, USA

Po-Yen Wu,
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

John H. Phan, and
Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA

May D. Wang*
Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA

May D. Wang: maywang@bme.gatech.edu

Abstract

RNA-seq data analysis pipelines are generally composed of sequence alignment, expression quantification, expression normalization, and differentially expressed gene (DEG) detection. Each step has numerous specific tools or algorithms, so we cannot explore all combinatorial pipelines and provide a comprehensive comparison of pipeline performance. To understand the mechanism of RNA-seq data analysis pipelines and provide some useful information for pipeline selection, we believe it is necessary to analyze the interactions among pipeline components. In this paper, by combining different alignment algorithms with the same quantification, normalization, and DEG detection tools, we construct nine RNA-seq pipelines to analyze the impact of RNA-seq alignment on downstream applications of gene expression estimates. Specifically, we find moderate linear correlation between the number of DEGs detected and the percentage of reads aligned with zero mismatch.

I. INTRODUCTION

Facilitated by next-generation sequencing (NGS) technology, high-throughput RNA sequencing (RNA-seq) interrogates the comprehensive profile of transcriptomes [1], enabling detailed identification of gene isoforms, translocation events, nucleotide variations, and post-transcriptional base modifications [2, 3].

*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.

A standard RNA-seq data analysis pipeline consists of (1) sequence read mapping, (2) expression quantification, (3) expression normalization, and (4) differentially expressed gene (DEG) detection, and each step has a considerable number bioinformatics tools. Since a pipeline consists of a sequence of the selected tools from each step, the combination of these tools provides a number of choices, yet raises the following question: Which pipeline should we use? Intuitively, the best pipeline would be composed of the best tool in each step. Researchers have conducted comparative analyses for the sequence alignment [4], expression quantification [5], expression normalization, and DEG detection [2] tools. The evaluation of the tools in a pipeline may be informative for pipeline selection. Based on this evaluation, we might select the most accurate alignment, quantification, normalization and DEG detection tools to construct a pipeline. However, the combination of the best tools does not ensure an accurate analysis result, especially when the performance of the tool is sample-related. For instance, Grant et al. [4] found that the base-level accuracy of alignment pipelines varies among samples. Until now, few studies systematically compared the performance of RNA-seq pipelines. Therefore, it remains uncertain whether the combination of best tools will produce a better-performing pipeline. To provide helpful information for pipeline selection and understand the mechanism of RNA-seq data analysis pipelines, we believe it is necessary to investigate the associations among the steps in RNA-seq pipelines. Once we know how the alignment step affects the final results (e.g., DEG detection), we can determine which alignment tool we should use and even estimate the number of DEGs with alignment metrics that can profile the alignment results.

In this paper, we analyze the impact of RNA-seq alignment pipeline on downstream applications of gene expression estimates, e.g., DEG detection. The rest of this paper is organized as follows. Section II introduces the experimental design and data analysis. Section III discusses the results and the potential impact of alignment on gene expression estimates. Finally, Section IV concludes our work.

II. METHODOLOGY

The workflow of this study is shown in Figure 1. To analyze the impact of alignment on gene expression estimates, we vary the alignment tools (Bowtie2 [6], BWA [7], GSNAP [8], Novoalign [9], and WHAM [10]) while using a fixed quantification tool (RSEM [11]), a normalization algorithm (trimmed mean of M-values normalization, TMM [12]), and a DEG detection tool (edgeR [13]).

A. Dataset

The dataset consists of SEQC samples A and B [14], which contain Stratagene's Universal Human Reference RNA and Ambion's Human Brain Reference RNA, respectively. The samples were sequenced with the Illumina HiSeq 2000 platform at three official sequencing sites, including the Beijing Genomics Institute (BGI), the Weill Cornell Medical College (CNL) and the Mayo Clinic (MAY). In this paper, we use only the data sequenced at BGI, which includes four replicates with around five million paired-end reads for each replicate. Each replicate has sixteen lanes, and we use the first two lanes.

B. Sequence Mapping and Expression Quantification

To analyze the impact of alignment on gene expression estimates, we vary the alignment tools, including Bowtie2, BWA, GSNAP, Novoalign and WHAM. For Bowtie2, GSNAP, Novoalign, and WHAM, we use two sequence alignment reporting strategies, single-hit and multiple-hit. Whereas single-hit aligners report only one location for a single read, multiple-hit aligners can report more than one location. BWA only reports single-hit alignments. We use the same reference genome (i.e., UCSC hg19) and the same genome annotation (i.e., AceView [15]) for all alignment pipelines. For gene expression quantification, we use RSEM with both the AceView transcriptome [15] and hg19 as reference genomes. The data generated from RSEM are in the form of gene counts.

C. Alignment Profiles

We characterize alignment profiles by using the percentage of reads aligned with zero and one mismatch as alignment metrics. Reads aligned with zero or one mismatch are more likely to account for gene expression estimates. We extract the percentage of reads aligned with no mismatch denoted as ZeroMismatchPercentage, and those with at most one mismatch denoted by OneMismatchPercentage. In addition, we count the number of reads aligned with single- or multiple-hit reporting. Since each sample has four replicates, we first compute the alignment metrics for each replicate, and then calculate the average as the alignment metrics of the sample.

D. DEG Detection Specificity

For gene expression estimates, evaluating every gene is not possible, especially when most genes have similar expression. As a result, we propose to use DEG detection as a downstream evaluation of gene expression estimates. We identify DEGs using the edgeR package in R. Before detecting DEGs, we use TMM (trimmed mean of M-values normalization) to normalize the data. Since each sample has four replicates (Replicates 1, 2, 3, and 4), we compare two replicates with the other two to detect DEGs (i.e., Replicates 1 and 2 vs. Replicates 3 and 4, Replicates 1 and 3 vs. Replicates 2 and 4, and Replicates 1 and 4 vs. Replicates 2 and 3). With various combinations, we have three groups, that is, we can get three DEG numbers for each sample. Because replicates come from the same sample, ideally the number of DEGs should be close to zero based on the assumption that the pipeline performs well. To capture and model this assumption, we define “DEG index” as “each pipeline’s total DEG number” to represent the pipeline’s quality. That is, for each pipeline, we add the three DEG numbers as its DEG index. The DEG index can quantify differences among pipelines. Meanwhile, the only variable in the comparison of pipelines is the alignment tool, which will be the only source of the discrepancy among the DEG indices of the pipelines. To investigate the effects of different DEG adjusted p-value thresholds on our observation, we detected DEGs with different thresholds (from $p = 0.01$ to 0.1). As larger adjusted p-value thresholds indicate looser constraints for DEGs, we expected more DEGs when we gradually increased the thresholds.

III. RESULTS AND DISCUSSION

Figures 2B and 2D show that, for most alignment tools for both Samples A and B, more than 60% of reads aligned with zero mismatch, and over 80% of reads aligned with zero or one mismatch, suggesting that the percentage of zero and one mismatch can cover the majority of reads in the alignment files. For both Samples A and B, alignment pipelines showed almost the same trend in ZeroMismatchPercentage and OneMismatchPercentage, suggesting that ZeroMismatchPercentage and OneMismatchPercentage in the alignment tools might be independent of the samples. We also verified that single-hit alignment pipelines only report one hit for each read; in contrast, multi-hit alignment pipelines can report several hits for some reads (Figures 2A and 2C).

Figures 3 and 4 show the key finding of our study: The DEG indices of RNA-seq pipelines have moderate linear correlation with the percentage of reads aligned with zero or one mismatch (ZeroMismatchPercentage and OneMismatchPercentage). Figures 3 and 4 show the impact of alignment pipelines on the DEG indices of Samples A and B, respectively. Note that single- and multiple-hit alignment strategies are distinctive. We use linear regression to measure their impact on DEG indices separately. For Sample A, both multiple-hit (blue boxes in Figure 3) and single-hit (red boxes in Figure 3) the DEG indices of alignment pipelines tended to decrease as ZeroMismatchPercentage increased. However, for the OneMismatchPercentage, the correlations between the DEG indices and the alignment pipelines were insignificant (Table I). As for Sample B (Figure 4 and Table II), both multi- and single-hit DEG indices of the alignment pipeline also had linear correlation with ZeroMismatchPercentage. Unlike that of Sample A, both multi- and single-hit DEG indices of the alignment pipeline exhibited a moderate linear correlation with OneMismatchPercentage in Sample B. This discrepancy might relate to the sample differences. Some sample-related metrics can also account for the impact of alignment pipelines on DEG index apart from the two metrics above. For Sample A, the sample-related metrics might fluctuate among results of alignment pipelines, while for Sample B, the other metrics may be consistent, which leads to that discrepancy. In addition, compared with single-hit alignment algorithms, ZeroMismatchPercentage of multiple-hit alignment algorithms have stronger linear impact on DEG index (Table I and II). Overall, our study discovered an alignment pipeline metric – ZeroMismatchPercentage – with moderate linear impact on gene expression estimation.

IV. CONCLUSION

We investigated the impact of alignment pipelines on gene expression estimates of RNA-seq pipelines. First, we constructed nine different RNA-seq pipelines by combining different alignment pipelines with the same quantification, normalization, and DEG detection tools. With these RNA-seq pipelines, we computed DEG indices for real datasets. Then, to profile alignment pipelines, we calculated the percentages of reads aligned with zero and one mismatch. Our study indicated that the ZeroMismatchPercentage of alignment pipelines had moderate linear impact on DEG index. Thus, we recommend constructing RNA-seq pipelines for DEG detection by choosing alignment tools that result in high ZeroMismatchPercentage. Although this preliminary study focused on two samples, nine

different pipelines, and two metrics we plan to include additional samples (i.e., SEQC samples C and D), pipelines, and metrics in a more comprehensive study.

Acknowledgments

This work was supported in part by grants from the National Institutes of Health (NHLBI 5U01HL080711, U54CA119338, 1RC2CA148265), the Georgia Cancer Coalition (Distinguished Cancer Scholar Award to Professor May D. Wang), the Children's Healthcare of Atlanta, Microsoft Research, the Hewlett-Packard Company, and China Scholarship Council (201306010292).

The authors would like to thank Mrs. Jane Chisholm for her valuable assistance.

References

1. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. Jun.2011 8:469–77. [PubMed: 21623353]
2. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013; 14:R95. [PubMed: 24020486]
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009; 10:57–63.
4. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics. Sep 15.2011 27:2518–28. [PubMed: 21775302]
5. Wu, P-Y.; Phan, JH.; Wang, MD. An Approach for Assessing RNA-seq Quantification Algorithms in Replication Studies. Genomic Signal Processing and Statistics (GENSIPS), 2013 IEEE International Workshop on; 2013; p. 15-18.
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9:357–359. [PubMed: 22388286]
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
8. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26:873–881. [PubMed: 20147302]
9. Novoalign. http://www.novocraft.com
10. Li Y, Patel JM, Terrell A. Wham: a high-throughput sequence alignment method. ACM Transactions on Database Systems (TODS). 2012; 37:28.
11. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011; 12:323. [PubMed: 21816040]
12. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]
13. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140. [PubMed: 19910308]
14. S. M.-I. Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotech. advance online publication, 08/24/online 2014.
15. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts. Genome biology. 2006; 7:S12.

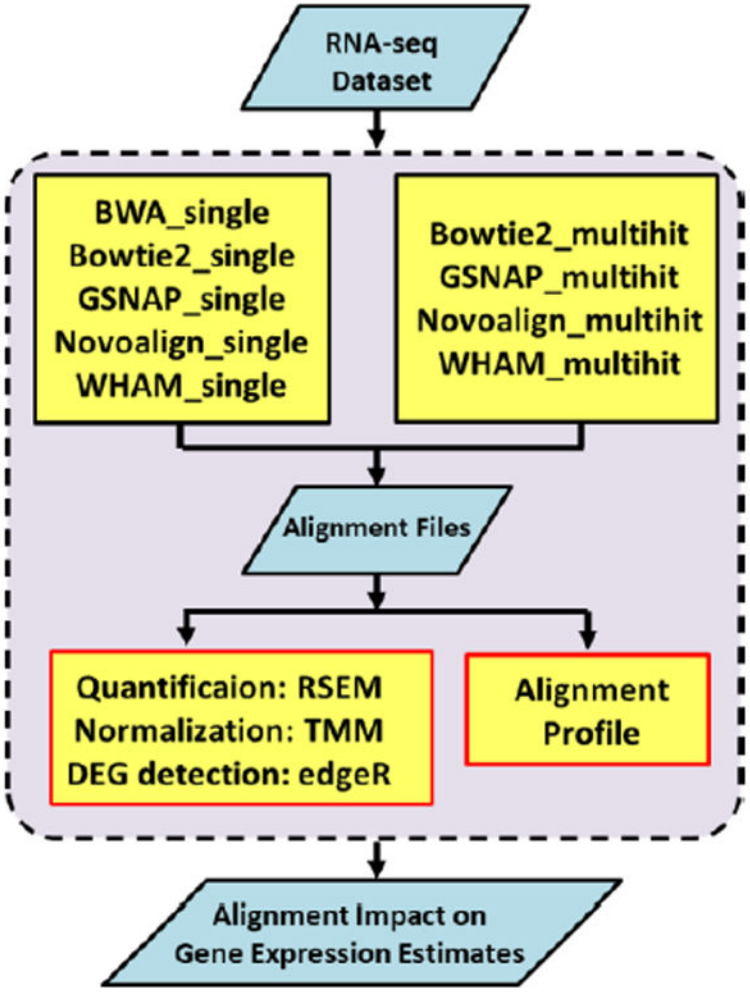

Figure 1.
The workflow for investigating the association between RNA-seq alignment profiles and gene expression estimates.

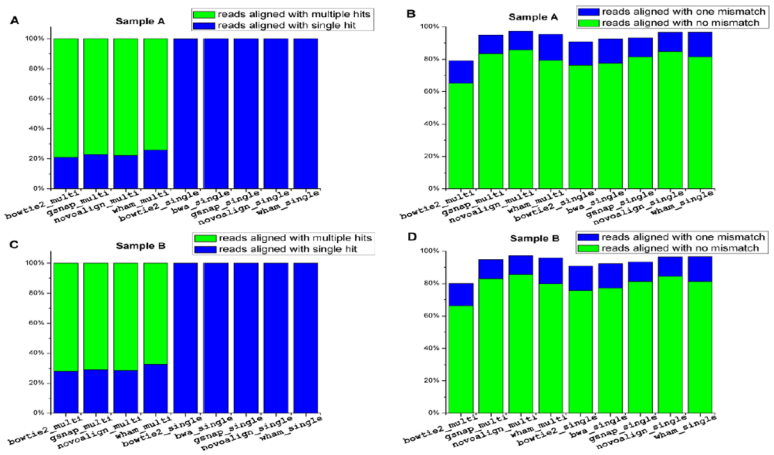

Figure 2.

Alignment profiles of Samples A and B, including percentage of reads aligned with single-hit or multiple-hit and percentage of reads aligned with zero or one mismatch.

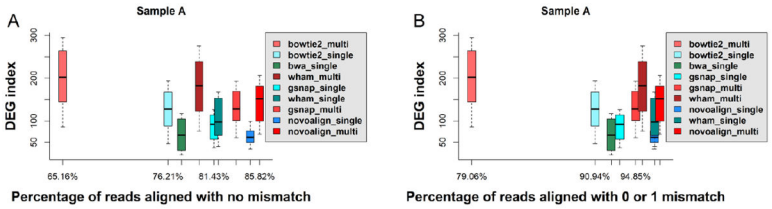

Figure 3.

The impact of alignment pipelines on gene expression estimation (Sample A). The boxplots illustrate the DEG indices of the alignment pipelines with their DEG adjusted p-value thresholds.

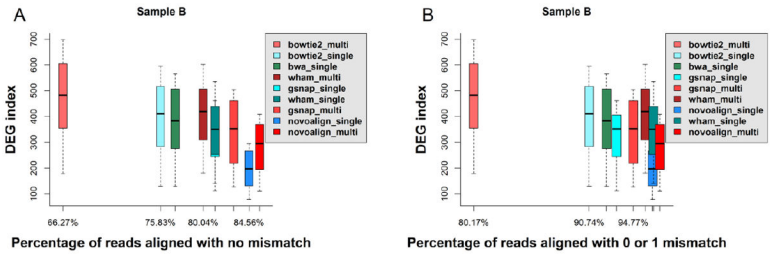

Figure 4.

The impact of alignment pipelines on gene expression estimation (Sample B). The boxplots illustrate the DEG indices of the alignment pipelines with their DEG adjusted p-value thresholds.

Table I

Correlation Coefficients of Sample A.

DEG adjusted p-value threshold	OneMismatchPercentage		ZeroMismatchPercentage	
	multiple-hit	***single-hit***	***multiple-hit***	***single-hit***
0.01	−0.7901	−0.0478	−0.8774	−0.0387
0.02	−0.7518	−0.2361	−0.8842	−0.2792
0.03	−0.8721	−0.2456	−0.9610 **	−0.2864
0.04	−0.8537	−0.2841	−0.9363 *	−0.3927
0.05	−0.7537	−0.3958	−0.8616	−0.4101
0.06	−0.6826	−0.5797	−0.8091	−0.7022
0.07	−0.7611	−0.5829	−0.8792	−0.7764
0.08	−0.7500	−0.3759	−0.8775	−0.6194
0.09	−0.7622	−0.3549	−0.8867	−0.5777
0.1	−0.7023	−0.3952	−0.8440	−0.6163

Significance codes:

** p-value < 0.05,

* p-value < 0.1.

Table II

Correlation Coefficients of Sample B.

DEG adjusted p-value threshold	OneMismatchPercentage		ZeroMismatchPercentage	
	multiple-hit	***single-hit***	***multiple-hit***	***single-hit***
0.01	−0.5828	−0.3632	−0.7507	−0.7059
0.02	−0.7194	−0.5530	−0.8585	−0.8246 *
0.03	−0.7689	−0.6980	−0.8940	−0.8754 *
0.04	−0.8104	−0.6447	−0.9208 *	−0.8109 *
0.05	−0.8026	−0.7132	−0.9142 *	−0.8697 *
0.06	−0.8234	−0.7764	−0.9175 *	−0.9187 **
0.07	−0.8689	−0.7473	−0.9452 *	−0.9363 **
0.08	−0.8546	−0.7435	−0.9315 *	−0.9475 **
0.09	−0.8499	−0.6872	−0.9345 *	−0.9246 **
0.1	−0.8091	−0.6547	−0.9118 *	−0.9034 **

Significance codes:

** p-value < 0.05,

* p-value < 0.1.