Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2019 doi:10.20944/preprints201903.0124.v1

Strandedness during cDNA synthesis, the stranded parameter in htseq-count, and analysis of RNA-Seq data

Krishna A. Srinivasan, Suman K. Virdee, and Andrew G. McArthur*

M.G. DeGroote Institute for Infectious Disease Research, Department of Biochemistry and Biomedical Sciences, DeGroote School of Medicine, McMaster University, Hamilton, Ontario L8S 4K1, Canada

*To whom correspondence should be addressed.

Contact: mcarthua@mcmaster.ca

Abstract

RNA sequencing (RNA-Seq) is a complicated protocol, both in the laboratory in generation of data and at the computer in analysis of results. Several decisions during RNA-Seq library construction have important implications for analysis, most notably strandedness during complementary DNA (cDNA) library construction. Here we clarify bioinformatic decisions related to strandedness in both alignment of DNA sequencing reads to reference genomes and subsequent determination of transcript abundance.

1 Introduction

RNA sequencing (RNA-Seq) is a powerful method for quantifying the transcriptome and identifying differential gene expression (Mortazavi *et al*. 2008; Wang *et al*. 2009; Conesa *et al*. 2016). RNA-Seq produces a large amount of data; therefore software packages designed for data analysis are vital for the advancement of the field (Figure 1). Based on our work with RNA-Seq of zebrafish embryos and human cell lines, we identify a few key aspects of RNA-Seq analysis not consistently described in the literature.

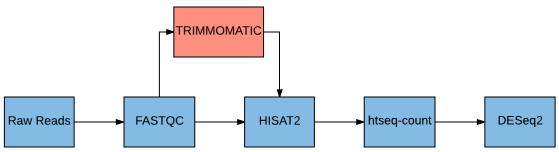

Figure 1. A standard RNA-Seq data analysis pipeline. FASTQC (Blankenberg *et al*. 2010) is the first tool in the pipeline and involves checking the quality of raw reads. Trimmomatic (Bolger *et al*. 2014), an optional step, may be used to trim low-quality bases if read quality is lower than a specified threshold. HISAT2 (Kim *et al*. 2015) is used to align reads to a reference genome. Compared with other read aligners, HISAT2 is more capable of dealing with reads that incorrectly map to pseudogenes. After alignment, htseq-count (Anders *et al*. 2015) identifies the number of reads that map to genes. DESeq2 (Love *et al*. 2014) uses the output of htseq-count to identify differentially expressed genes.

2 htseq-count and strandedness

Anders and colleagues developed HTSeq (Anders *et al*. 2015), a python framework for analyzing high throughput sequencing data, within which htseq-count is located. htseq-count is a commonly used tool in many data analysis pipelines (Conesa *et al*. 2016; Fonseca *et al*. 2014; Wu *et al*. 2013) which counts the number of sequencing reads mapping to genes after the read alignment step. Certain parameters in htseq-count are dependent upon the laboratory methods used; in particular, the 'stranded' parameter is dependent upon the complementary DNA (cDNA) library construction methodology. The three possible settings are 'Yes', 'No', and 'Reverse' and refer to strandedness during library construction. Strand-specific protocols involve preserving information on which strand the messenger RNA (mRNA) originated, while this information is lost in un-stranded protocols (Figure 2).

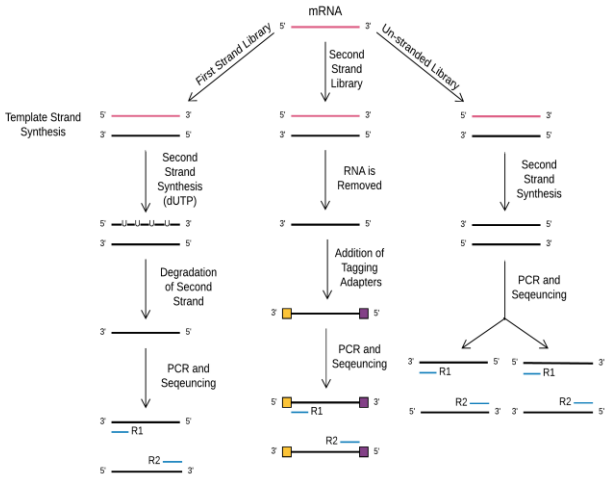

Figure 2. First strand, second strand and un-stranded cDNA library protocols. dUTP = deoxyuridine triphosphate, PCR = polymerase chain reaction, R1 = Read 1 (first read produced), R2 = Read 2 (second read produced).

For first strand synthesis, strand information is preserved by degrading the non-template strand and PCR amplifying only the template cDNA strand during library construction (the 'Reverse' option in htseq-count).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2019 doi:10.20944/preprints201903.0124.v1

Second strand protocols preserve strand information by using adapters to mark the template strand (the 'Yes' option in htseq-count, which is the default). When aligning reads from bi-directional Illumina sequencing of a first strand library, the second read produced will map to the DNA strand containing the coding sequence while the first read will map to the opposite strand. For a second strand cDNA library, the first read produced will map to the DNA strand containing the coding sequence while the second read will map to the opposite strand. The default 'Yes' stranded option in htseq-count may be incorrectly interpreted merely as a confirmation of having strand-specific data, instead of specifically indicating data from second strand library. Using stranded 'Yes' on a first strand cDNA library results in all read pairs that map to the reference genome in the correct orientation being discarded and reads that map in the incorrect orientation being counted as legitimate reads. Essentially, only antisense data is considered for transcript abundance measures. First strand cDNA libraries thus require the use of the stranded 'Reverse' option (Figure 3).

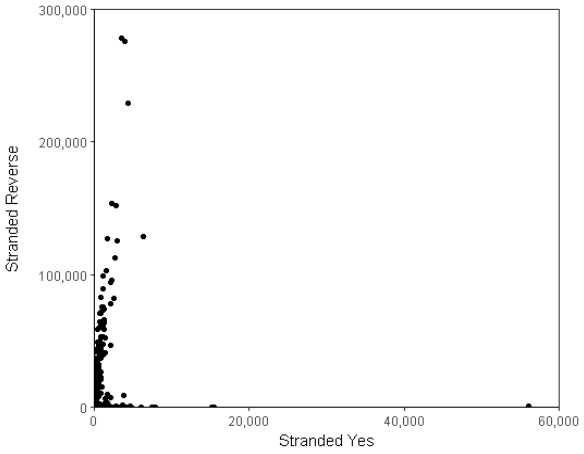

Figure 3. htseq-count read counts for zebrafish genes using stranded 'Reverse' vs. stranded 'Yes' with a first strand cDNA library, with each point representing a single gene. Stranded 'Reverse' produces larger counts for the majority of genes. In contrast, data is being unnecessarily removed using the 'Yes' option. Approximately 700,000 read pairs were mapped to genes using stranded 'Yes', while stranded 'Reverse' had 21 million read pairs, using an initial input of 25 million reads pairs.

3 Conclusions

Second strand cDNA libraries were more common when htseq-count was designed than now, resulting in the default 'Yes' option. However, first strand cDNA libraries with dUTP second strand marking, such as the NEBNext Ultra Directional RNA Prep Kit (New England Biolabs, Ipswich, MA, USA), are now more commonly used for RNA-seq (Levin *et al*. 2010) and require the 'Reverse' option. We also note that the upstream read alignment step performed by HISAT2 includes the 'rna-strandness' parameter (which add tags in the alignment file for read strandedness), as well the 'nofw' and 'norc' parameters (restricting read alignments to specific strands). The defaults are unstranded and no strand-based restrictions upon read alignment, respectively, but alternative settings could restrict the creation of read alignments for interpretation by htseq-count. The language used differs, where unstranded in HISAT2 corresponds to htseq-count's 'No', HISAT2's forward corresponds to htseq-count's 'Yes' (i.e. second strand cDNA libraries), and HISAT2's reverse corresponds to htseq-count's 'Reverse' (ie. first strand cDNA libraries). HISAT2's precurser Tophat (Kim *et al*. 2013) used FR-unstranded for 'No', FR-secondstrand for 'Yes', and FR-firststrand for 'Reverse'. As with any bioinformatics software, we recommend that users be aware of the relationship between software parameters and laboratory techniques, and use a stranded option that matches their cDNA library protocols to prevent unnecessary errors.

References

Anders, S. *et al*. (2015) HTseq — a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31, 166–169.

Blankenberg, D. *et al*. (2010) Manipulation of FASTQ data with galaxy. *Bioinformatics*, 26, 1783–1785.

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Conesa, A. *et al*. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17, 13-31.

Fonseca, N.A. *et al*. (2014) RNA-Seq gene profiling - A systematic empirical comparison. *PLoS ONE*, 9, e107026

Kim, D., Langmead, B. & Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12, 357–360.

Kim, D. *et al*. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14, R36.

Levin J.Z. *et al*. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*, 7, 709–715.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15, 550.

Mortazavi, A. *et al*. (2008) Mappingand quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5, 621-628.

Wang, Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10, 57–63.

Wu P-Y. *et al*. (2013) Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*, 14 Suppl 11:S8