A Comprehensive RNA-Seq pipeline includes meta-analysis, interactivity and automatic reporting

Giulio Spinozzi[1*], Valentina Tini[1], Laura Mincarelli[2], Brunangelo Falini[1] and Maria Paola Martelli[1*]

[1]University of Perugia, Department of Medicine, Section of Hematology

[2]Earlham Institute, Norwich Research Park

*Email of Corresponding authors: giulio.spinozzi@unipg.it, maria.martelli@unipg.it

Abstract

There are many methods available for each phase of the RNA-Seq analysis and each of them uses different algorithms. It is therefore useful to identify a pipeline that combines the best tools in terms of time and results. For this purpose, we compared five different pipelines, obtained by combining the most used tools in RNA-Seq analysis. Using RNA-Seq data on samples of different Acute Myeloid Leukemia (AML) cell lines, we compared five pipelines from the alignment to the differential expression analysis (DEA). For each one we evaluated the peak of RAM and time and then compared the differentially expressed genes identified by each pipeline. It emerged that the pipeline with shorter times, lower consumption of RAM and more reliable results, is that which involves the use of *HISAT2* for alignment, *featureCounts* for quantification and *edgeR* for differential analysis. Finally, we developed an automated pipeline that recurs by default to the cited pipeline, but it also allows to choose between different tools. In addition, the pipeline makes a final meta-analysis that includes a Gene Ontology and Pathway analysis. The results can be viewed in an interactive *Shiny App* and exported in a report (pdf, word or html formats).

Introduction

Large-scale expression analysis is an important tool for RNA analysis, but there are many different approaches and techniques for studying differential gene expression under different conditions. In particular, sequencing techniques are becoming the method of choice in the transcriptome analysis. Even within the RNA-Seq, however, it is possible to resort to many different approaches. There are numerous aligners as well as different software for quantification and it is becoming increasingly important to identify a unique pipeline for differential analysis that knows how to choose the best approaches to obtain precise results in a short time.

Previously, standard pipelines have been defined for RNA-Seq analysis, such as the pipeline using *TopHat2* for alignment and *Cufflinks* and *cummeRbund* for quantification and differential analysis (Trapnell et al., 2012), or its most recent evolution, which recurs to *HISAT2* for alignment, *StringTie* for quantification and *Ballgown* for differential analysis (Pertea et al., 2016). There have also been works that have shown how using different pipelines and algorithms leads to different performances in terms of time, memory and results. In a recent work (Germain et al., 2016), for example, it has been highlighted how the quantification methods using a statistical approach are better in terms of estimation of absolute abundance, while the methods that use read count prove more reliable in comparisons between different samples.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27317v2 | CC BY 4.0 Open Access | rec: 5 Nov 2018, publ: 5 Nov 2018

It is therefore interesting not only to identify a pipeline that is more efficient than the others, but also to create an easy-to-use tool that allows applying this pipeline and any subsequent analysis on a sample set.

Methods

The data we analyzed came from RNA-Seq experiments performed on two different AML cell lines with NPM1 mutation: OCI-AML3 (Quentmeier et al., 2005) and IMS-M2 (Chi et al., 2010). In both cases the treatment conditions were compared with the conditions without treatment. For each condition the experiment was done in triplicate. The kit used for the preparation of the sample was the truSeq RNA (Illumina, 2011), while the sequencer used for the sequencing was HiSeq 2500 by Illumina (Illumina, 2015), in rapid run and with a flow cell. Sequencing occurred in paired-end and using two lanes for sample. The two lanes corresponding to the same sample and to the same read have been merged into a single file before the alignment phase.

An initial quality analysis was performed on FastQ files using *FastQC* software (Andrews, 2010) and a contaminant genome evaluation using *FastQ-Screen* (Andrews, 2011). We then removed the PhiX genome and the ribosomal genome by identifying sequences through alignment on samples with bwa. The five compared pipelines were the following: *TopHat2* (Kim et al., 2013), *Cufflinks* (Trapnell et al., 2010), *cummeRbund* (Goff et al., 2012); *HISAT2* (Kim et al., 2015), *StringTie* (Pertea et al., 2015), *Ballgown* (Fu et al., 2017); *HISAT2*, *featureCounts* (Liao et al., 2014), *DESeq2* (Love et al., 2014); *HISAT2*, *featureCounts*, *edgeR* (Robinson et al., 2010); *kallisto* (Bray et al., 2016), *sleuth* (Pimentel).

Results

For the five pipelines, we analyzed the time taken for the various processes and the peaks of memory used. The data below are for the analysis of the sample treated using 4 threads for each process. The times were then multiplied by the six samples to obtain estimates of the total hours for each pipeline.

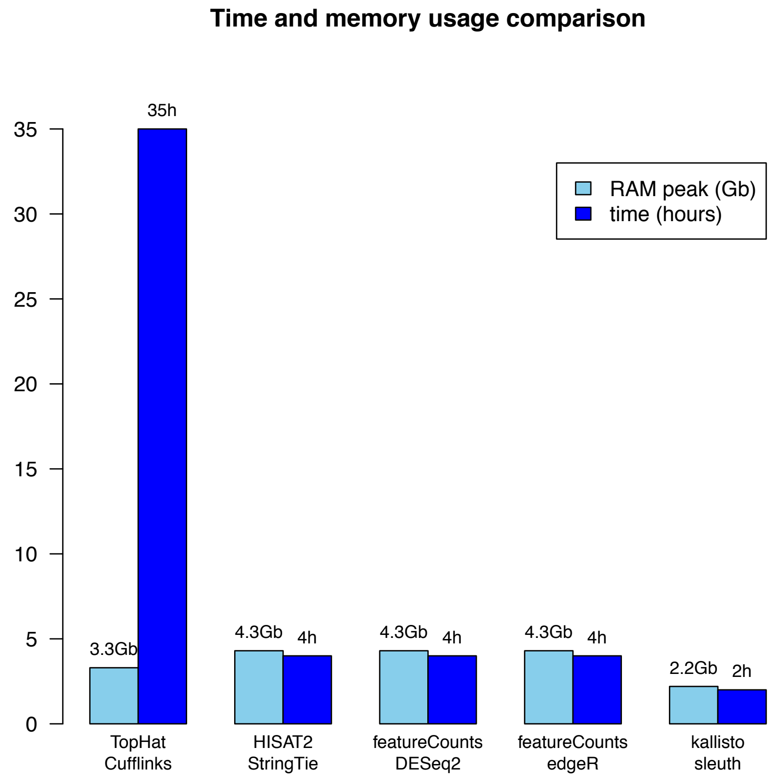


Figure 1. Histogram of the times and the RAM memory peaks reached during the RNA-Seq analysis for the five pipelines.

As well as comparing time and RAM memory performance, we then compared the pipelines through the results obtained after the differential analysis. The Venn diagram below shows the genes differentially expressed and with an absolute value of Fold Change greater than 1.5.

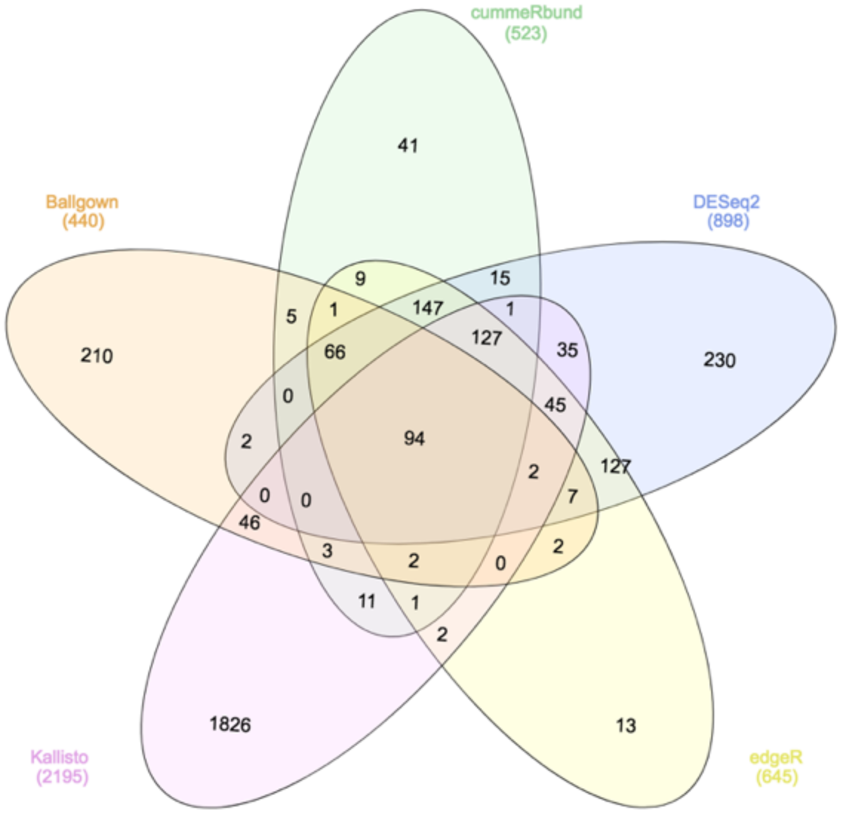


Figure 2. Venn diagram of genes with p-value lower than 0.05 and absolute value of log2-Fold Change higher than 1.5 common to the five different methods of analysis used.

By first evaluating the pipelines in terms of time consumption and memory usage peaks achieved, it was found that the most expensive step is in all cases the alignment, except in the *Kallisto* pipeline, where higher RAM peaks are reached in the differential analysis. Comparing the three different aligners to which we have recourse, it is clear that the slowest is TopHat2, which requires more than five hours for sample for alignment only. With *HISAT2* it comes down to about 15 minutes for sample, followed by the longest conversion steps of the SAM file in BAM, sorting and indexing, which increase the time for a sample up to 40 minutes. *Kallisto*, on the other hand, takes about 20 minutes for sample for pseudo-alignment and quantification. As for the consumption of memory, the highest are with *HISAT2*, with 4.3Gb, then down to 3.3Gb with *TopHat2* up to 1.4Gb with *Kallisto*, which, using an alignment on the transcriptome, requires consumption of memory lower than the alignments made on the genome. As for the quantification methods, instead, *featureCounts* with its read count requires slightly shorter times than the quantification with the statistical approach of *Cufflinks* and *StringTie*. The consumption of memory, however, never exceed a few hundred megabytes. The differential analysis in R, finally, required in all cases only a few minutes or a few seconds and did not have much influence in the overall time.

In terms of memory usage, all pipelines are quite similar, with the highest peaks concentrating in the initial alignment phase. Interestingly, however, note that all the peaks are kept below 5Gb, so the analysis in all cases can be carried forward even on a normal PC.

Regarding the results of the differential expression compared between the five pipelines, looking only at the number of genes that are significantly expressed compared with the control, *Kallisto* provides the largest number, with over 2,000 genes, followed by *DESeq2* with 898 genes. Looking at the results of *Kallisto*, however, it emerges that more than 80% of the identified genes are unique to the pseudo-alignment method and are not supported by the other methods. This suggests that they are therefore false positives. The pipeline that has instead identified the least number of genes differentially expressed is that of *HISAT2-StringTie-Ballgown,* with less than 500 genes. Even in this case, however, half of the identified genes is unique to this method and therefore allows us to suppose that these are also false positives. Better results are obtained with *DESeq2*, which has about 500 significant genes also confirmed by other methods, but another 200 are not in common. *edgeR* gives results similar to DESeq2, but, thanks to more stringent statistical methods, the number of significant genes is reduced to 500 and almost all of them are in common with the other methods. It therefore seems that using the statistical approach of *edgeR* rather than *DESeq2* allows to eliminate most of the false positives. Finally, even with the *TopHat2-Cufflinks-cummeRbund* pipeline good results are obtained from the differential analysis, with about 500 genes, most of which in common with the other methods.

In the light of these results, it emerges that, although *Kallisto* is the method that requires shorter times, the large number of potential false positives makes it unreliable compared to other pipelines. *TopHat2* and *HISAT2-featureCounts-edgeR* pipelines are the ones that give better results in differential analysis, as most of the genes they identify are in common with other methods. In terms of time, however, the *edgeR* pipeline is better than that of *TopHat2* as it requires almost a tenth of the time.

The *HISAT2-featureCounts-edgeR* pipeline is therefore the best of the five pipelines used because it takes a short time, the quantification method (the read count) is the simplest and fastest, and the differential analysis gives the best results, with a low number of potential false positives.

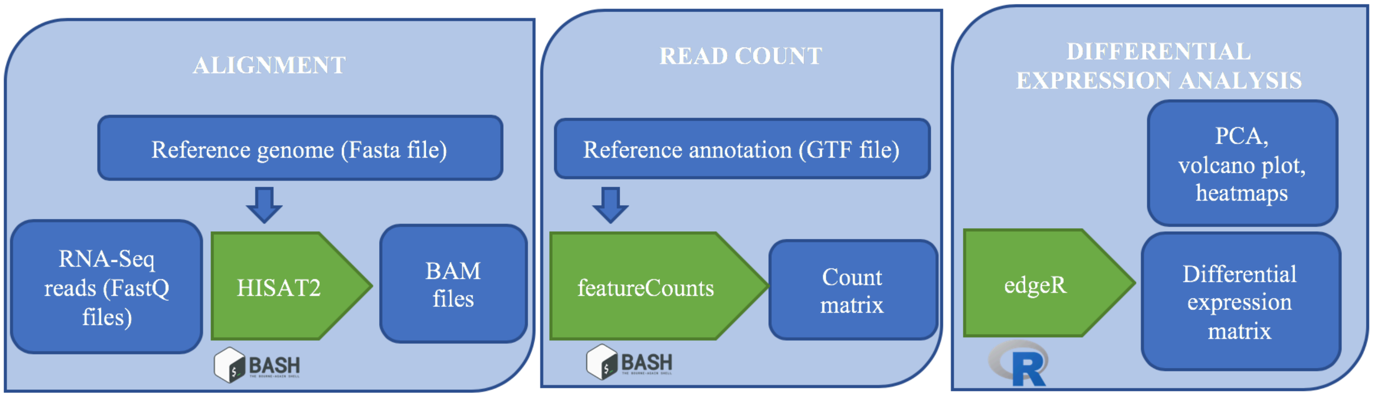

ALIGNMENT
Reference genome (Fasta file)
RNA-Seq reads (FastQ files)
HISAT2
BAM files
BASH
READ COUNT
Reference annotation (GTF file)
featureCounts
Count matrix
BASH
DIFFERENTIAL EXPRESSION ANALYSIS
PCA, volcano plot, heatmaps
edgeR
Differential expression matrix
R

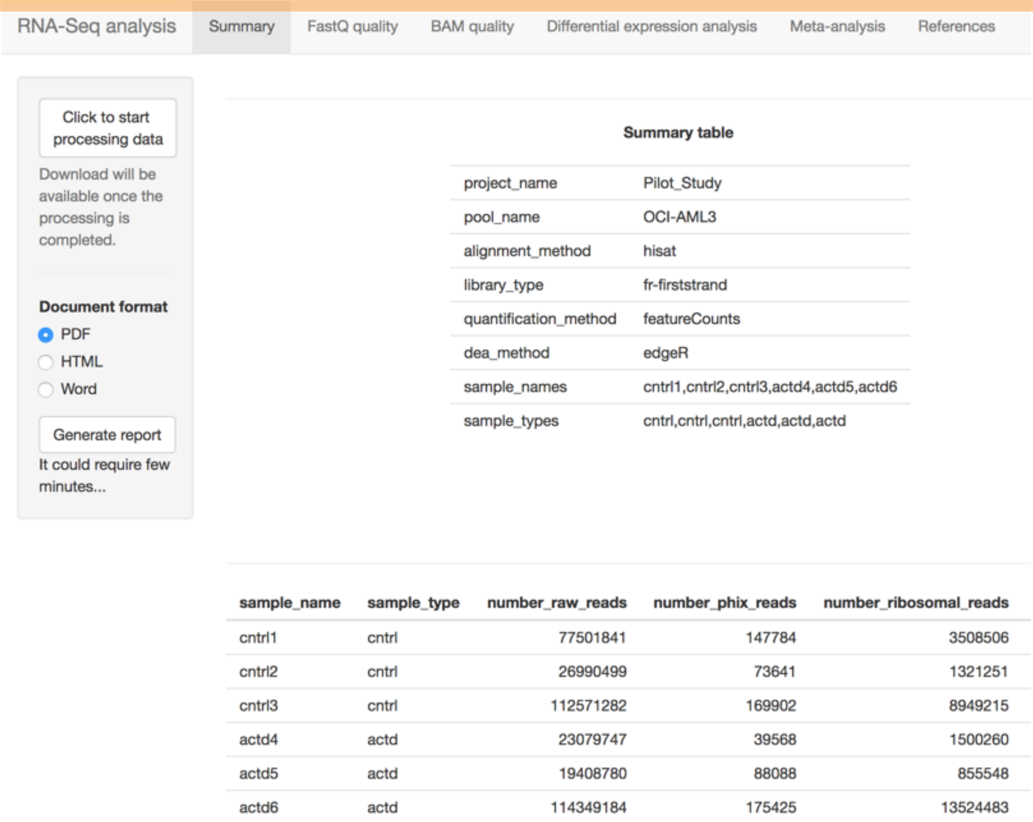
Figure 3. Workflow of HISAT2-featureCounts-edgeR *pipeline.*

In this pipeline we use *HISAT2*, an alignment software that uses indexing schemes based on the Burrows-Wheeler transform and on the Ferragina-Manzini index (FM) using two types of alignment indexing: an FM index on the whole genome to anchor the various alignments and numerous local FM indexes for a rapid extension of these alignments. Starting from the pre-processed FastQ files and a Fasta file of the reference genome, BAM files are obtained for each sample. Then on the aligned BAM files are performed the quantification with *featureCounts*, which uses the method of read count. In addition to the BAM files, *featureCounts* requires an annotation file (GTF file) of the reference genome. The matrixes produced as output of *featureCounts* become the input for *edgeR*. Before the differential analysis, the genes were filtered to keep those that have at least one count per million in at least two samples. After filtering, the data were normalized through the TMM method (Trimmed Mean of M values) (Robinson et al., 2010). The model used is a Generalized Linear Model (GLM), which represents an extension of the simplest linear model. Each gene is fitted through a negative binomial distribution. Differential Expression Analysis is performed by a likelihood ratio test. It consists in comparing the logarithm of the calculated likelihood for two different models, one of which is the one obtained under the null hypothesis. The value obtained is compared with the corresponding probability distribution and thus the p-value is obtained, which is then corrected by the FDR method. Outputs consist of a matrix with differential expression results, a PCA, a volcano plot, a top genes heatmap, and a sample distance heatmap.

Once we identified the most efficient pipeline, we made it the default choice in the automated pipeline we developed (https://github.com/giuliospinozzi/creo_pipelines). The application can be used by command line or by using a graphical interface and allows to choose between different methods of alignment, quantification and differential expression analysis. In particular, it makes quality control, pre-processing, alignment, transcript quantification and differential expression analysis on BAM files. Given the input files and the working directory, the pipeline is completely automated. First, quality control on FastQ files is performed with *FastQC* e *FastQ-Screen*. *FastQC* makes quality control and creates one report for sample. *FastQ-Screen* estimates approximately the percentage of reads that can be mapped on genomes other than human, like ribosomal genome, PhiX genome and mouse genome. This allows to evaluate the presence of contaminating genomes. Pre-processing follows quality control: the reads are aligned on PhiX genome and ribosomal genome to eliminate contaminations. Alignment can be performed with *TopHat2* or *HISAT2*; in the first case quantification is performed with *Cufflinks* and DEA with *cummeRbund*, in the second case quantification is performed with *featureCounts* and DEA with *DESeq2* or *edgeR*. A second intermediate quality control analysis is also performed on the aligned BAM files with some of the *RSeQC* scripts and in particular: *inner_distance, junction_annotation, junction_saturation, bam_stat, read_distribution*.

It is possible to perform an optional meta-analysis on the results. It consists in *Gene Ontology* enrichment analysis and *KEGG Pathway* enrichment analysis on the differentially expressed genes (with absolute Fold Change value higher than 1.5 and adjusted p-value lower than 0.05). The meta-analysis part has been developed exclusively for the human genome at the moment, although the rest of the pipeline can also work for different genomes. One of the future goals is to expand the genomes available also for meta-analysis.

Finally, the results obtained and saved in the appropriate folders can be viewed in an interactive *Shiny App* (Chang et al., 2018), from which you can also download a report with all the results. The advantage of show the results in this form is that, once the Shiny App is launched, it is intuitive and easy to use even for those who are not familiar with computer science. The *Shiny App* shows the results of the RNA-Seq analysis divided into a series of tabs for each phase: the summary tab contains two tables that show the initial setting parameters and details about pre-processing on the FastQ files; the FastQ quality tab contains the *FastQC* and *FastQ-Screen* outputs; the BAM quality tab contains the *RSeQC* outputs obtained from the quality analysis on the aligned files; the differential expression analysis tab contains a result table and a series of plots and in particular a PCA, a volcano plot, a heatmap of the 35 genes with greater variance, a heatmap showing the distances between the samples; the Meta-analysis tab is divided into two sub-tab, one for GO analysis and the other for Pathway analysis, both containing a result table, a series of dot-plots and interactive network plots.



Download will be available once the processing is completed.

● PDF

It could require few minutes...

Summary table

project_name	Pilot_Study
pool_name	OCI-AML3
alignment_method	hisat
library_type	fr-firststrand
quantification_method	featureCounts
dea_method	edgeR
sample_names	cntrl1,cntrl2,cntrl3,actd4,actd5,actd6
sample_types	cntrl,cntrl,cntrl,actd,actd,actd

sample_name	sample_type	number_raw_reads	number_phix_reads	number_ribosomal_reads
cntrl1	cntrl	77501841	147784	3508506
cntrl2	cntrl	26990499	73641	1321251
cntrl3	cntrl	112571282	169902	8949215
actd4	actd	23079747	39568	1500260
actd5	actd	19408780	88088	855548
actd6	actd	114349184	175425	13524483

Figure 4. Screenshot of Shiny App that shows analysis results.

The aim of this project, in addition to expanding the choice options available within the pipeline, is to make the pipeline available to the entire institute through a centralized platform and take advantage of its ease of use (both for the *GUI* specifically created for analysis and for the *Shiny App*).

Acknowledgements

The project is funded by the European Research Council (ERC) to Maria Paola Martelli (ContraNPM1AML) and by CINECA (HP10CZ1LLU) to Giulio Spinozzi.

References

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Protocols, 7, 562-578.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protocols, 11, 1650-1667.

Germain PL, Vitriolo A, Adamo A, Laise P, Das V, Testa G. (2016). RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. Nucleic Acids Res., 44, 5054-5067

Quentmeier H, Martelli MP, Dirks WG, Bolli N, Liso A, MacLeod RAF, Nicoletti I, Mannucci R, Pucciarini A, Bigerna B, Martelli MF, Mecucci C, Drexler HG, Falini B. (2005). Cell line OCI/AML3 bears exon-12 NPM gene mutation-A and cytoplasmic expression of nucleophosmin. Leukemia, 19, 1760–1767

Chi HT, Vu HA, Iwasaki R, Nagamura F, Tojo A, Watanabe T, Sato Y. (2010). Detection of exon 12 type A mutation of NPM1 gene in IMS-M2 cell line. Leuk. Res., 34, 261–262.

Illumina. (2011). TruSeqTM RNA and DNA Sample Preparation Kits v2. Data Sheet Illumina® Seq., 1–4

Illumina. (2015). HiSeq® 2500 Sequencing System. Specif. Sheet Illumina® Seq., 1–4

Andrews S. (2010). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Andrews S. (2011). FastQ Screen. http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol., 14

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515

Goff LA, Trapnell C, Kelley D. (2012). cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R Packag. Version 2.2

Kim D, Langmead B, Salzberg SL. (2015) HISAT: A fast spliced aligner with low memory requirements. Nat. Methods, 12, 357–360

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol., 33, 290–295

Fu J, Frazee AC, Collado-Torres L, Jaffe AE, Leek JT. (2017). ballgown: Flexible, isoform-level differential expression analysis. R Packag. version 2.10.0

Liao Y, Smyth GK, Shi W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30, 923–930

Love MI, Huber W, Anders S, Lönnstedt I, Speed T, Salzberg S, et al. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., 15, 550

Robinson MD, McCarthy DJ, Smyth GK. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–40

Bray NL, Pimentel H, Melsted P, Pachter L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol., 34, 525–527

Pimentel H. sleuth: Tools for investigating RNA-Seq. R Packag. version 0.29.0

Robinson M, Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol., 11, R25

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. (2018). shiny: Web Application Framework for R. R package version 1.1.0.