A Review of Differential Gene Expression Software for mRNA sequencing

Lisa J. Cohen[1] and Jiying Li[2]

[1] Molecular, Cellular & Integrative Physiology Graduate Group
[2] Computer Science Graduate Program,
University of California, Davis

Abstract. RNAseq technology has become a standard toolkit in biological science due to its comprehensive RNA profiling ability and large decreases in sequencing costs in recent years. One of the main goals of RNAseq experiments is to identify differentially expressed genes in response to conditions of interest. Thus, the robustness and accuracy of RNAseq data is extremely critical to generate data-based hypothesis and draw conclusions. Here we compare six typical tools by applying them with a publicly available set of RNAseq data. Our results indicate that there are differences between each package, which could potentially provide researchers with some ideas and a reference to choose the most suitable package for their purposes. In addition, we correct a labeling mistake in the metadata associated with the original dataset through our analysis.

Keywords: differential gene expression, mRNA sequencing

Introduction

The field of transcriptomics research has progressed rapidly in the last decade with fast development of next-generation sequencing (NGS) technologies (Conesa et al. 2016). Sequencing of mRNA (RNAseq) has become the dominant approach for gene expression profiling and has almost replaced the conventional microarray technology because of cost and the opportunity to explore novel, unsolicited sequences in samples. The advance of RNAseq has significantly increased the ability to identify and quantify transcripts, which broadly enables its application in a wide range of research areas. However, in spite of the pervasive adoption of RNAseq as a standard method in the field, there is no consensus about a single optimal pipeline for running RNAseq data analysis. The complexity in RNAseq analysis arises from various perspectives, ranging from RNA sequencing depth to statistical analysis and other critical steps.

Differential Gene Expression

A standard differential gene expression experiment compares genes that are expressed in a treatment compared to a control group. In the case of sequencing, a typical RNAseq analysis for an organism with a well-annotated reference genome involves these steps: consideration/planning of experimental design, quality control of the raw reads, read alignment to an annotated reference genome, quantification of gene and transcript levels, visualization, differential gene expression, alternative splicing, functional analysis and eQTL mapping. Different challenges are associated with each step, and various methods and software packages were developed to deal with the challenges. For example, in reads alignment, TopHat and STAR are optional packages to map reads to the genome reference, and Bowtie is a useful tool for mapping reads to the transcriptome. In our project, we are mainly interested in software packages and methods applied in differentially expressed (DE) gene analysis, which is the core step in RNA-seq.

History of Software for Differential Gene Expression

Current software packages and pipelines developed for differential gene expression include methods based on negative binomial models, such as edgeR, DESeq and baySeq, non-parametric approach such as NOIseq and SAMseq, transformation of gene-level read counts for linear model with limma, as well as both gene-based and transcript-based detection methods called Cuffdiff2 and EBseq. The understanding of these packages is growing with continuous testing, and modifications have been incorporated in the packages constantly. It is generally helpful to visit back the history of the development of the packages to recognize their similarity and difference, as well as the challenges and solutions along the way.

Early microarray studies sometimes had few biological replicate samples. Differential expression is determined by the simplest statistics, such as fold-change. It became evident that considering the variability over replicates is essential, and advanced statistical testing procedures, such modified t-test has been adopted. Similarly, some early RNAseq studies only had single samples per group, and counts within the single sample were reported to fit well to Poisson distribution. As more availability of RNAseq samples, it is realized that the variability is higher than expected by Poisson distribution, which only uses a single parameters to explain variance, a phenomenon called overdispersion. This overdispersion problem makes Poisson distribution prone to false positive rate.

To deal with overdispersion and biological variability, methods based on negative binomial and beta negative binomial models were introduced, where the negative binomial distribution includes multiple parameters to explain the variance. As the detection of differentially expressed genes involves performing a large number of statistical tests, multiplicity needs to be taken into account when determining significance. The conventional correction of P values aims to control the family-wise error rate, however it is often too conservative in the context of biological studies. Later, false discovery rate (FDR) has become a common practice, which corrects the P values using the Benjamini and Hochberg (1995) method.

Recent studies have showed comparison of methods and pipelines in simulated and real data sets, such as (Kvam et al 2012) compared the performance of four related packages (edgeR, DESeq, baySeq and TSPM) on simulated data sets, Soneson and Delorenzi (2013) compared 11 approaches (edgeR, DESeq, baySeq, NBPSeq, TSPM, EBSeq, NOIseq, SAMseq, ShrinkSeq and two versions of limma) on simulated data sets. Nookaew et al. (2012) included in their comparison five packages (edgeR, DESeq, baySeq, NOIseq and Cuffdiff) on a real data set on *Saccharomyces cerevisiae*.

In our project, we present a systematic pipeline comparison of six software packages DESeq1, DESeq2, edgeR, PoissonSeq, EBSeq and limma. We focus on a very recent public datasets with a simple experiment setup of one condition group vs. one control group, and each group with 4 replicates (Hateley et al. 2016). Specifically, the RNA samples are extracted in *Drosophila melanogaster* in hypergravity study. Samples treated with hypergravity (3 times of gravity = G3) fly pupae were compared to samples in normal gravity condition (referred to hearafter as G1). These simple dataset allow us to focus on evaluating the different software packages with a relatively large numbers of replicates. The use of the public data sets also provides a reference to compare our result and guarantees future comparison of more packages.

Methods and Results

Data were obtained from the National Center for Biotechnology Information (NCBI), Sequence Read Archive (SRA): SRA study SRP073366, Bioproject PRJNA318586; GEO GSE80323. Data are from Drosophila melanogaster[Taxonomy ID: 7227] in the pupal life stage exposed to hypergravity at 3G force in a centrifuge vs. control conditions (Hateley et al 2016). This study was chosen because of its simple, balanced design (1 treatment vs. control) with 4 replicates per group and because *Drosophila melanogaster* has a well-annotated reference genome. After an initial PCA of our processed data, there appeared to be a problem with the group assignments for two samples (SRR3390484 = G3R3 and SRR3390478 = G1R1) in comparison to Figure 1. After contacting the authors to confirm, G1R1 and G3R3 were switched in our report here and corrected on NCBI by the authors.

The *Drosophila melanogaster* transcriptome (ensembl version r6) was indexed and reads were processed with the salmon quasi-mapping quantification software package to transcript level estimates (Patro et al. 2015). The salmon software has recently become a standard method in the field for quantifying RNAseq reads, which takes into consideration multi-mapping reads and the length of the reference transcripts into the estimate. Salmon is a similar package to Kallisto (Bray et al. 2016) that the authors used in Hateley et al. 2016. This probabilistic estimation method has recently replaced both previous methods of quantifying reads by alignment: HTSeq, which output raw counted reads aligning to the reference regardless of the length of the transcript, and cuffdiff, which did take into consideration the length of the transcript (Fragments Per Kilobase of

Table 1. Sample information according to the Sequence Read Archive (SRA) study SRP073366; Bioproject PRJNA318586; GEO GSE80323. The SRA has individual ID for Rn, Biosample, SampleName all for the same sample.

Treatment	Run	Biosample	SampleName
G1R1	SRR3390478	SAMN04858514	GSM2124263
G1R2	SRR3390479	SAMN04858515	GSM2124264
G1R3	SRR3390480	SAMN04858516	GSM2124265
G1R4	SRR3390481	SAMN04858517	GSM2124266
G3R1	SRR3390482	SAMN04858518	GSM2124267
G3R2	SRR3390483	SAMN04858519	GSM2124268
G3R3	SRR3390484	SAMN04858520	GSM2124269
G3R4	SRR3390485	SAMN04858521	GSM2124270

transcript per Million mapped reads or FPKM) but took a long time to quantify and was not compatible with downstream differential expression analysis software except its own in the tuxedo suite (cuffdiff, cummeRbund) developed for this purpose (Trapnell et al 2012). Quantification with either Kallisto or salmon is faster than any other quantification method currently available because the algorithms are based on probabilistic estimation. These programs use a one-step quasi-mapping approach to quantification, which bypasses the time-consuming requirement to first align to the reference transcriptome or genome then quantify reads in a separate step.

Biomart was used to convert Drosophila melanogaster transcript ID into gene names and gene ID (Durnick et al. 2009). Transcript level quantification resulted in 30,443 transcripts ('NumReads' output from salmon), which were then summarized at the gene level using the R package, tximport resulting in 13,868 genes (Soneson et al. 2016).

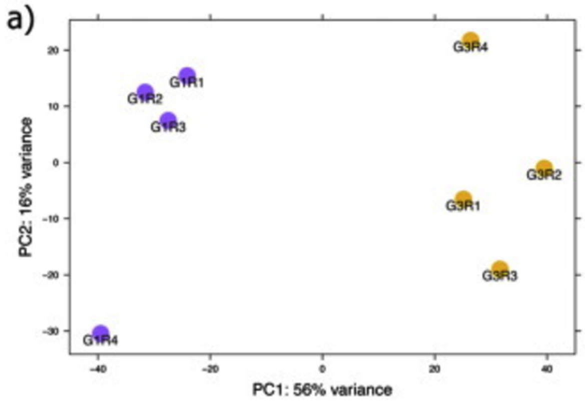


Figure 1. PCA from Hateley et al. (2016) Figure 1 indicates G3 samples (left purple dotes) are grouped together while G1 samples are grouped together (right yellow dots).

Software Packages

Here we provide brief summary of the software packages in this project. For more detailed description of the packages and the statistical models they apply, it is better to read the original publications and the related software package websites. Our R code and all files is available at Github

The same table of 13,868 genes were used as input for each package described below. Because the quantification output represents estimates, values are floating point decimals rather than integers. DESeq2 has a new function DESeqDataSetFromTximport() which will take output directly from tximport, whereas DESeq1 and the other packages used in this report did not. In these cases, values were rounded using the round() function in R.

After normalization and model fitting in each package, genes were considered ”significant” if padj¡0.05 and log2FC 1. Additionally, rows of genes were removed (filtered) if they had a mean of less than 1 across the row, or if all samples had 0 expression values.

edgeR

The package, edgeR computes differential expression using empirical Bayes estimation and negative binominal model. In particular Bayes estimation can moderate the degree of overdispersion across genes by considering gene expression level within each sample. As default, the Trimmed Mean of M-values (TMM) procedure is carried out to incorporate the effect of different sequencing depth between samples, where Benjamini and Hochberg (1995) procedure is used to control the false discovery rate (FDR).

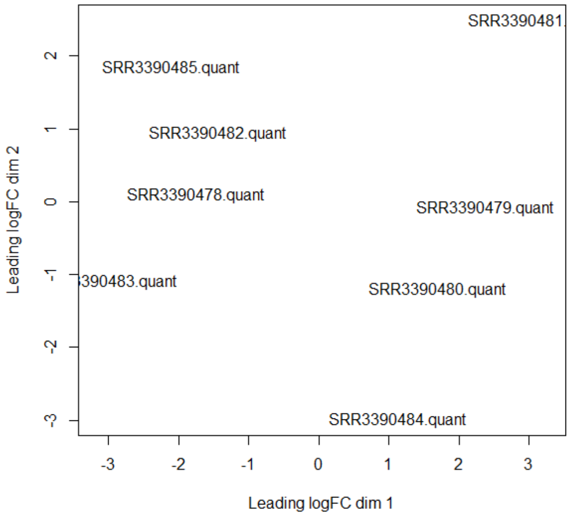


Figure 2. PCA after normalization with EdgeR

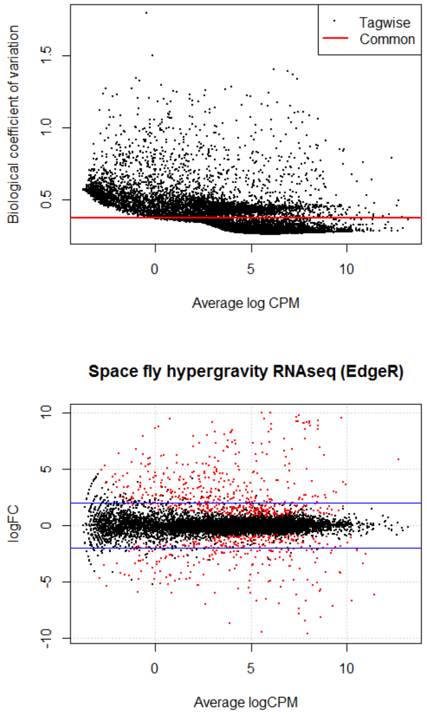


Figure 3. Dispersion (biological coefficient of variation, left) and MA plot (right) showing log2 fold change (FC) ratio vs. log2 mean expression on the x-axis of differentially expressed genes (red dots, padj<0.05) in EdgeR

The PCA in Figure 2 compared to the PCA from Hateley et al. 2016 (Figure 1) look similar in that the samples in group G1 are separate from group G3 along the PC1 axis. And the separation of samples along the PC2 axis are also similar.

DESeq

DESeq uses a negative binomial model, similar to edgeR. However, DESeq allows for a data-driven parameter estimation, modeling the observed relationship between the mean and variance when estimating dispersion. According to the user manual, this allows a balanced selection of differentially expressed genes throughout the dynamic range of the data (Anders and Huber 2010). Similar to edgeR, a scaling factor normalization procedure is carried out to consider different sequencing depths of each samples. The BenjaminiHochberg procedure is used to control the FDR (1995). In addition, DESeq has been developed to enable analysis of experiments with small numbers of replicates.

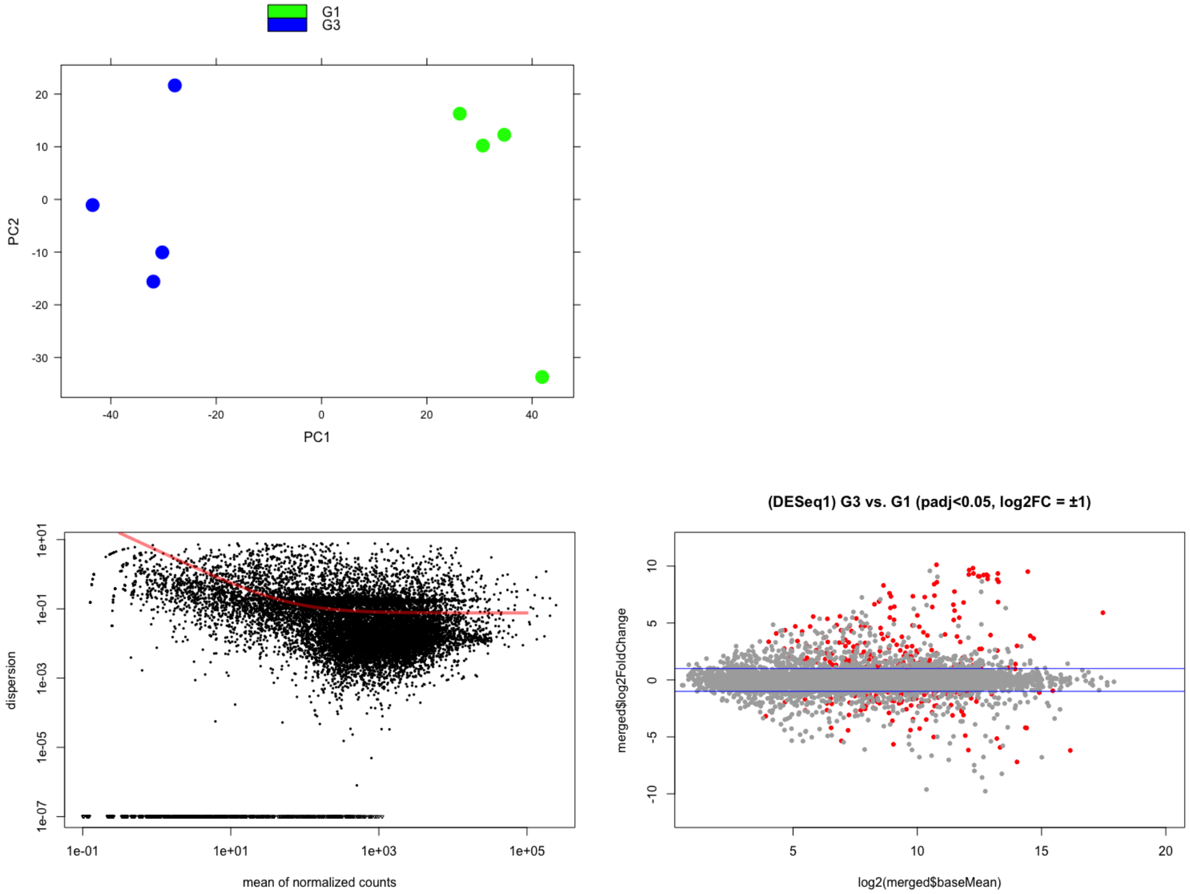


Figure 4. PCA (top), dispersion plot (bottom left), and MA plot (bottom right) showing log2 fold change (FC) ratio vs. log2 mean expression on the x-axis.

The PCA (Figure 4, top) appears to be similar to Hateley et al. (2016). Dispersion is a parameter that estimates variability within negative binomial distribution. The dispersion (Figure 4, bottom left) appears to be dense and extending below the line. This could lead to more false positives because the large group of data below the red line will be shifted up towards the fitted red line. DESeq has a tight control on false positives.

DESeq2

By far the most cited and widely used, the package, DESeq2 was developed by the same group as DESeq (Love, Huber and Anders 2014). The adjusted p-values from DESeq2 tend to be lower, The aim of developing DESeq2 was to restore statistical power lost with heavily controlling type-I control (false discovery) in DESeq. By fixing this, and allowing for shrinkage of dispersion below the mean, more differential expression is discovered.

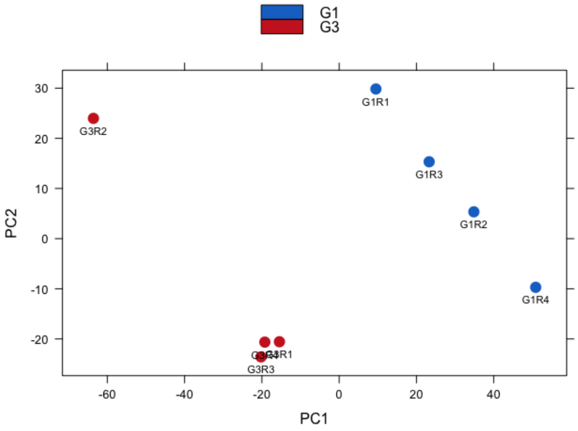


Figure 5. PCA with different pattern than observed in Hateley et al. 2014.

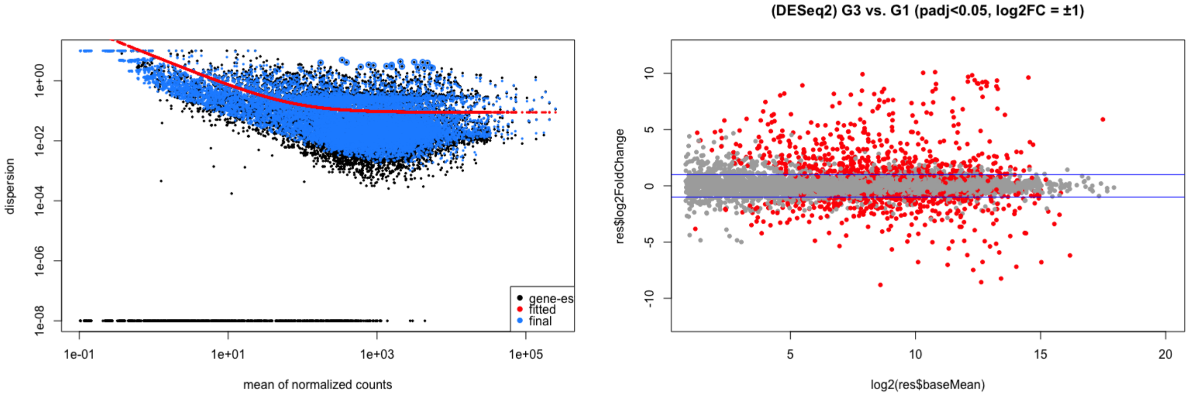


Figure 6. Dispersion, fitted (red line) and final shrinkage (blue) with DESeq2 (left). MA plot on right where red dots are padj$<$ 0.05 and blue lines are log2FC ±1.

While the separation of the samples is clear along PC1 between treatment (G3) and control (G1) replicates, the distances between samples in the G3 group along the PC2 axis are different (Figure 4). G3R2 is by itself while G3R3, G3R1 and G3R2 are tightly grouped together. There appears to be something strange going on with this PCA compared with the other packages that we can't figure out. Suggested to look and see which genes are most contributing to the variance of PC1 and PC2. While this pattern in the PCA appears to be specific to the DESeq2 package, it is curious that the authors of the study who generated the data (Hateley et al. 2016) did not observe this pattern when using DESeq2. We used DESeq2/1.12.4, while it is not clear what specific version of the package they used and perhaps our parameters are slightly different. DESeq2 undergoes development changes on a regular basis and default settings can be changed unbeknownst to users. For example, in 2014 the default setting of beta prior=FALSE in the main DESeq() function was changed to beta prior=TRUE, creating a major shift in the number of differentially expressed genes if users were not careful. It could be that there was a new normalization setting introduced to DESeq2 in our latest version compared to the version Hateley et al. 2016 used. Given the time constraints of this project, we were unable to investigate this further.

limma-voom

Limma, Linear Model for Microarray Analysis, is originally developed for analyzing microarray data, but the limma model has been extended for RNAseq analysis recently. According the user guide of limma, it is recommended to use TMM normalization in edgeR package to transform the normalized counts to logarithmic (base 2) scale and estimates their mean-variance relationship to determine a weight to each observation before linear model. Such transformation is called limma-voom. By default, the Benjamini-Hochberg procedure is used to estimate the FDR.

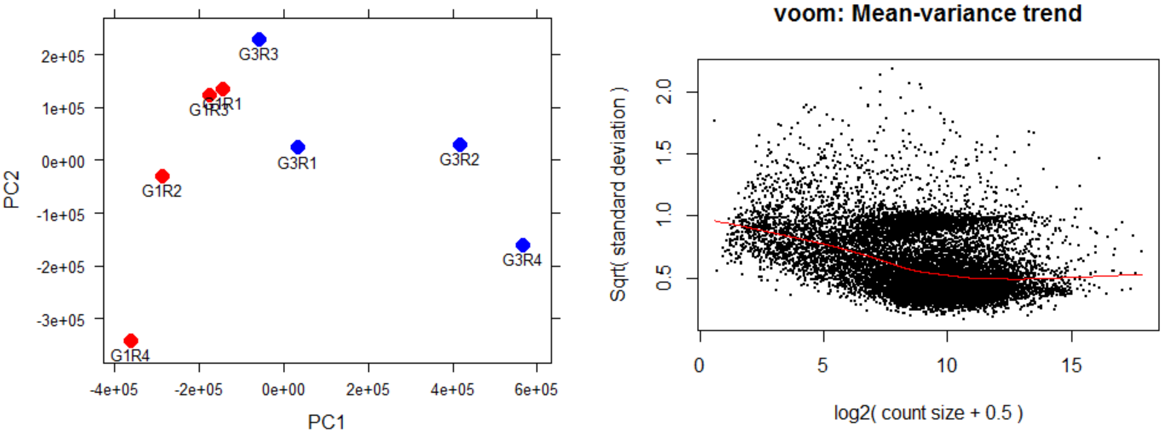


Figure 7. PCA (left) and mean variance trend (right).

Using a similar model to edgeR, limma produced almost exclusively significant results (>10k). Given the time constraints on this project, our code may not have been correct, so the numbers of genes for this package were not included in the final analysis (Figure 10).

EBSeq

EBSeq was mainly developed to identify differentially expressed isoforms, but it also has demonstrated a robust result in gene level analyses (Leng et al 2013). EBSeq estimates the posterior likelihoods of differential and equal expression by the aid of empirical Bayesian methods, and the underlying assumption is also negative binomial distribution. To account for the different sequencing depths, a median normalization procedure similar to DESeq is used as the default method. A Bayesian FDR estimate is provided.

The class of objects available to explore the output from the EBSeq package does not clearly include an output with P-values or adjusted p-values. The main feature of EBSeq is FC vs. Posterior FC, indicating a PPEE = posterior probability estimate (Figure 8). While EBSeq predicts the number of significant genes given an adjustable input padj threshold and outputs a list, the output table does not include these padj values. This is not useful for many users, although the FC vs. Posterior FC correlation does indicate outliers outside of the linear positive correlation (Figure 8, left).

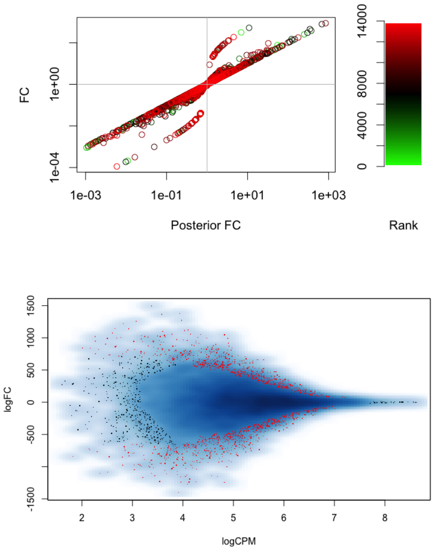


Figure 8. FC vs. Posterior FC (left) and MA plot (right) with log2FC vs. log expression (falsely labeled as ”logCPM”) where the blue smear contains expression.

PoissonSeq

The PoissonSeq package (Li et al. 2012) uses a goodness-of-fit estimate to define a gene set that is least differentiated between two conditions, which is then used to compute library normalization factors. Importantly, while most methods use standard approaches for multiple hypothesis correction, such as Benjamini-Hochberg, PoissonSeq implements a novel estimation of false discovery rate (FDR) for count data that is based on permutation. The permutation-based method generates a null distribution of the test statistic, rather than fully trusting a distribution Poisson distribution. The authors suggest that this method may be more robust to data that violate the assumptions of the distributions (either Poisson or negative binomial), which will then lead to more robust gene-gene correlation.

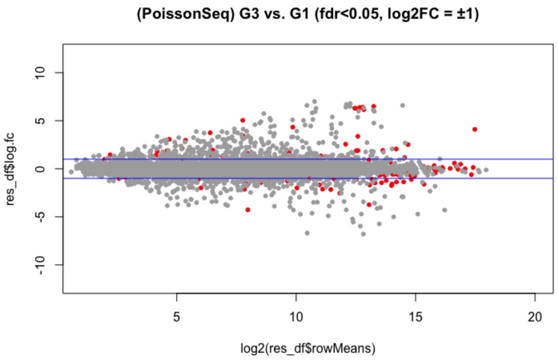


Figure 9. MA plot from PoissonSeq. Red dots are padj¡0.05 and blue lines are log2FC ±1.

The data structure output from PoissonSeq was not as intuitive to use as edgeR, DESeq, limma packages. The range of log2FC ratios vs. log2 mean expression are not as wide broad as seen in the other packages (Figure 9). The genes with padj< 0.05 also did not have as high log2FC values compared to the other packages.

Cufflinks and Cuffdiff

Cuffdiff 2 estimates expression at transcript level and controls for variability and mapping ambiguity by using a beta negative binomial model for fragment counts. Although Cuffdiff 2 enables to analyze signals at the transcript level, it reports differential expression also at the gene level and these gene level results can also be used as a basis for comparison with the other software packages. By default, Cuffdiff 2 uses a similar scaling factor procedure as DESeq to account for the different sequencing depths and the BenjaminiHochberg procedure to control the FDR. The Cuffdiff 2 method specifically addresses the uncertainties in counts owing to ambiguous reads that easily result in false differential expression calls of genes especially with several similar isoforms.

Discussion and Conclusions

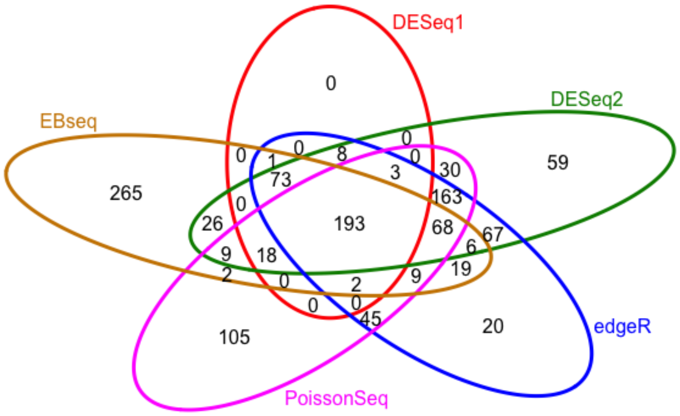
In conclusion, based on the results from this study, we are unable to recommend one consensus differential expression software package to use that will fit all RNAseq experiments. Each software package presented in this report has its own nuanced features that contribute to the results. Depending on the type of experiment and the reason for wanting to look at differentially expressed genes, investigators may choose one package over another. For example, if an experiment wants to look at many differentially expressed genes, edgeR or DESeq2 will be best. But if wanting to be conservative, DESeq1 might be best. EBSeq is robust for differential expression of isoforms. Alternatively, a combination of packages could be used, taking the overlap of significant genes identified by multiple packages. When presented with a new set of data for a differential expression analysis, we see here that perhaps confidence can be gained in trying multiple approaches to see what works best for the data set. When considering substantial downstream investment of time and financial resources, we realize that it is important to be conservative when selecting candidate genes that are differentially expressed. For those researchers who are not comfortable in the R programming environment, there are web-based solutions to differential expression, such as the new Degust platform offered by the Victoria Bioinformatics Consortium and the more veteran platform, Galaxy.

There appears to be a core set of genes that are overlapping between all packages. The overlap between main software packages are seen below in Figure 10. In the bottom venn diagram (Figure 10, bottom), EBseq and PoissonSeq were removed because they each had high numbers of genes unique to only those packages.

Hateley et al. 2016, who used Kallisto to quantify and DESeq2 for differential expression analysis, reported more than 1,100 genes to be differentially expressed with padj$<$ 0.05. We did not have time to examine the specific differences in output between our study here and the significant genes reported by Hateley et al. 2016. It would be interesting to look to see the identity of the genes overlapping.

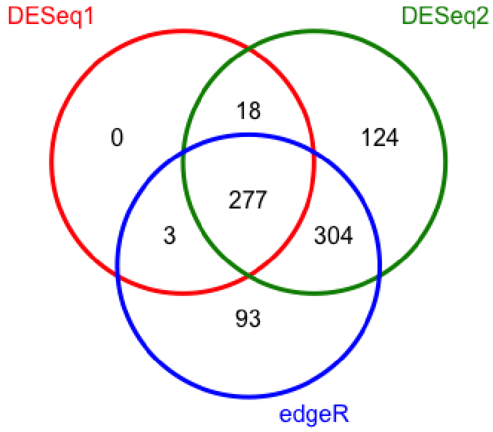
Based on our analyses (Figure 10, below), the center numbers of either 193 or 277 genes (depending on which overlapping packages) could be groups of genes to focus on in downstream studies. In the future, it may be interesting to develop a package to automatically coordinate a number of packages, using the code to run multiple packages in parallel (e.g. edgeR, DESeq, limma) then deliver the overlap of genes like we have here in Figure 10 as output to the user. While this is beyond the scope of this report, it is also important to validate these results experimentally and in the context of the annotations to see whether the genes that are predicted to be differentially expressed by these software packages make sense to the investigator.

Venn Diagram



Unique objects: All = 1191; S1 = 298; S2 = 723; S3 = 677; S4 = 647; S5 = 691

Venn Diagram



Unique objects: All = 819; S1 = 298; S2 = 723; S3 = 677

Figure 10. Venn diagraphs showing overlap between all software packages examined in this report (top) and the 3 main packages most heavily used and with the lowest numbers of unique genes specific for that package (bottom).

References

Anders and Huber. 2010 Differential expression analysis for sequence count data. Genome Biology, 11, pp. R106. doi: 10.1186/gb-2010-11-10-r106, http://genomebiology.com/2010/11/10/R106/.

Leng, N., J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould,R.M. Stewart, and C. Kendziorski. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments, Bioinformatics, 2013.

Love MI, Huber W and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, pp. 550. doi: 10.1186/s13059-014-0550-8.

Benjamini and Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 57, No. 1 (1995), pp. 289-300

Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhln M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res. 2012;40:1008497.

Hateley S, Hosamani R, Bhardwaj SR, Pachter L, Bhattacharya S. 2016. Transcriptomic response of Drosophila melanogaster pupae developed in hypergravity. Genomics. 2016 Sep 10. pii: S0888-7543(16)30088-X. doi: 10.1016/j.ygeno.2016.09.002.

Rob Patro, Geet Duggal, Carl Kingsford. Posted June 27, 2015. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. bioRxiv. doi: http://dx.doi.org/10.1101/021592

Nicolas L Bray, Harold Pimentel, Pll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology 34, 525527 (2016), doi:10.1038/nbt.3519

Soneson C and Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91.

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn and Lior Pachter. 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7, 562578 (2012) doi:10.1038/nprot.2012.016

Conesa et al 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Aug 26;17(1):181.

Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 2012;99:248-56.

Durinck S, Spellman P, Birney E and Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols, 4, pp. 11841191.

Soneson C, Love MI and Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2)

Li J1, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012 Jul;13(3):523-38. doi: 10.1093/biostatistics/kxr031. Epub 2011 Oct 14.