Comparison of Differential Gene Expression Methodologies for RNA-sequencing Data

Darlene Olsen[1]
Norwich University[1], 158 Harmon Drive, Northfield, VT 05663

Abstract: Detection of differential expression in RNA-seq experiments offers an exciting opportunity in biomedical research to investigate the regulatory mechanisms that control cellular processes in organisms giving insight into diseases and developmental processes. However, due to the comprehensive nature of the data, there are many challenges in identifying lists of transcripts that are differentially expressed between two or more conditions. Several approaches have been developed but there is no consensus on the best approach. This research compares four different statistical methods for identifying differentially expressed genes in RNA-seq experiments.

Key Words: RNA-seq data, gene expression, differential expression, ROC analysis

1. Introduction

High-throughput sequencing technology has helped researchers obtain more accurate assessments of how cellular networks may be altered by disease. The introduction of RNA-seq has had a revolutionary impact on biomedical research by enabling researchers to investigate more complex aspects of the transcriptome (Lee et al., 2011; Guo et al., 2013). However, this rich genetic data is accompanied by complications at the analysis phase (Guo et al, 2013). Presently, researchers have not agreed on the best approach for analyzing RNA-seq data to determine the lists of transcripts that are differentially expressed between two or more conditions.

2. Background

2.1 RNA-seq experiments

In RNA-seq experiments the following process is used to identify differential expression of genes between two or more conditions. First, mRNA is isolated from each sample, randomly fragmented into small pieces, reverse transcribed into cDNA, amplified using a polymerase chain reaction (PCR), and finally sequenced into millions of *reads* (DNA bases) that are aligned with a reference genome, refer to Figure 1 (Li et al., 2012).

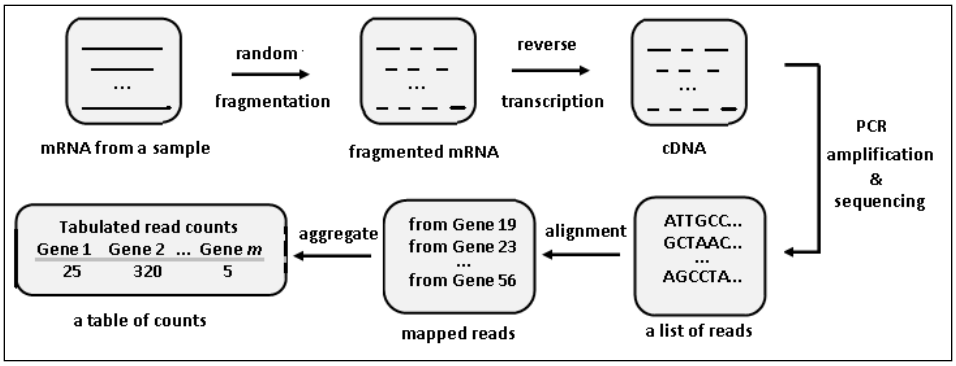

Figure 1. A representation of an RNA-seq experiment. Image adapted from Li et al., 2012.

The number of reads that match the reference is tabulated and statistical testing is used to decide if, for a given feature or gene, an observed difference in read counts for each sample is significant (Anders and Huber, 2010), as shown in Figure 2. RNA-seq has recently become the preferred method for gene expression profiling but the comprehensive nature of the data poses challenges at the analysis stage (Guo et al., 2013; McGettigan, 2013). Although there are RNA-seq statistical analysis tools available, there is an opportunity to improve RNA-seq analysis (Guo et al., 2013)

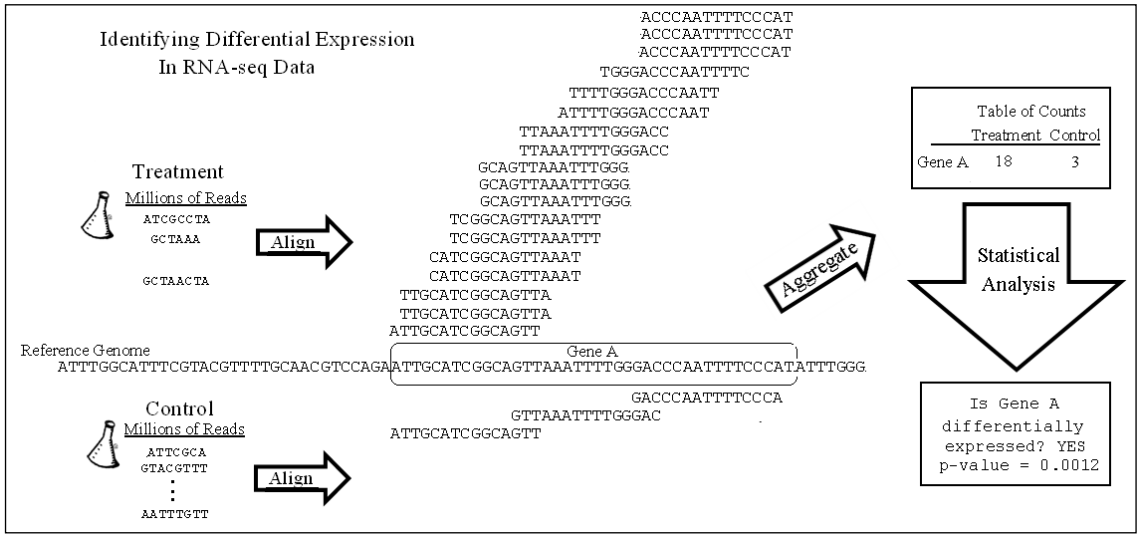

Figure 2. In this overview of RNA-seq differential expression analysis, the Treatment sample has a much larger count of reads for Gene A than in the Control sample, thus Gene A is said to be differentially expressed between the two samples. Image adapted from http://rafalab.jhsph.edu/688/lec/lecture-9-intro-to-seq.pdf.

2.2 Differential expression analysis in RNA-seq data

Two main components of the statistical analysis for identifying differential expression in RNA-seq data are parameter estimation for the fitted statistical model of the counts and testing for differential expression for each of the thousands of genes (Rapaport et. al, 2013). In order to detect differential expression of genes from the count table (or normalized count table), the counts are assumed to follow an underlying data distribution. Once the parameters of the underlying distribution of the counts have been estimated from the data, statistical testing is used to decide whether, for a given gene or feature, an

observed difference in read counts between conditions is significant (Anders and Huber, 2010). If there is a significant difference in the counts, that gene is said to be differentially expressed among the conditions. Since there are thousands of genes or transcripts that are being tested at once, multiple testing issues arise; therefore, each methodology uses a technique to control the False Discovery Rate (FDR) (Li et al., 2012). After a list of differentially expressed genes is generated, gene ontology enrichment tools can be used to gain biological insight into the regulatory mechanisms of the cell.

Typical approaches for modeling count data are the Poisson and the negative binomial distribution. Without modifications, both these distributions cannot address some of the biases in RNA-seq count data. Each methodology needs to account for the differences in library depth and gene length. In addition, each model adjusts parameter estimation to control for overdispersion in the data, greater amount of spread in the values than what would be expected in a statistical model.

The objective of this study is the comparison of the four methodologies of statistical analysis of RNA-seq data with a focus on the statistical methods used to approximate the parameters of the underlying distribution of the count data. The statistical test used to determine the p-value for detecting differential expression is characteristically a direct result of the method used to model the count data. While the focus will be on the estimation of parameters for the statistical model of the counts, at each stage of the analysis, both components will be examined in depth for the four methodologies.

3. Methods

3.1 Four methodologies for the analysis of differential expression

The four methodologies that are compared are PoissonSeq (Li et al., 2012), edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010) and Cuffdiff (Goff et al., 2012). These approaches are a subset of some of the most commonly used parametric methodologies (McGettigan, 2013).

PoissonSeq

- Models gene counts as a Poisson random variable.
- A goodness-of-fit estimate is used to define a gene set that is the least differentiated to compute the library normalization factor (Rapaport et. al., 2013).
- To account for overdispersion, the model transforms the read counts to account for the variability and uses a log-linear relationship to estimate the mean (Li et al., 2012).
- A score statistic based on the chi-square distribution is used for identifying differential expression.
- A novel estimation of the FDR based on permutations is used (Rapaport, 2013).

edgeR

- Models gene counts with the negative binomial distribution.
- Uses the Trimmed Means of M values (TMM), a scaling factor between two experiments after excluding genes with large counts or differences in gene expression (Rapaport et al., 2013).
- Assumes that the mean and variance are related by a dispersion factor to account for overdispersion (Rapaport et al., 2013; Robinson et al., 2010).

- The Fisher Exact Test adapted to the negative binomial distribution is used to detect differentially expressed genes. (Rapaport et al., 2013; Robinson et al., 2010).
- The FDR can be controlled with standard methods such as the Bonferroni correction or the Benjamini–Hochberg (BH) procedure.

DEseq

- Models gene counts with the negative binomial distribution.
- Uses scaling factor for a given sample by computing the median of the ratio, for each gene, of its read count over its geometric mean across all samples. (Rapaport et al., 2013).
- Assumes that the mean and variance are related by a dispersion factor to account for overdispersion (Rapaport et al., 2013; Robinson et al., 2010).
- The Fisher Exact Test adapted to the negative binomial distribution is used to detect DEGs (Rapaport et al., 2013; Robinson et al., 2010).
- The FDR controlled for by standard methods (Bonferroni correction or BH procedure).

Cuffdiff

- Models gene counts with the negative binomial distribution.
- Uses the Reads per Kilobase per Million reads (RPKM) to normalize the data (Rapaport et al., 2013).
- Uses a separate variance model for single-isoform and multi-isoform genes (Rapaport et al., 2013).
- Uses a test statistic based on the ratio of counts between two conditions that is shown to follow the normal distribution (Rapaport et al., 2013).
- The FDR controlled for by standard methods (Bonferroni correction or BH procedure).
- CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output (Goff et al., 2012).

3.2 Comparison of methodologies

All four methods are evaluated on real-world data**.** Sun et al. (2011) used RNA-seq to examine gene expression in 8 commonly used breast cancer cell lines. Studies have shown that ER+ and ER- breast cancers have very different gene expression profiles that may be useful in diagnosis and treatment. In this study, samples were differentiated by ER status where 4 were ER+ and 4 were ER-. The data are also available from the Gene Expression Omnibus database (GSE27003). (Sun et al., 2011)

Further analysis can be done using established benchmark datasets that contain cases with known outcomes to provide suitable evaluation measures (Vihinen, 2012). Data sets that have spiked-in synthetic oligonucleotides are mixed into samples, have known outcomes that yield a true measure of positive and false rates to perform an analysis to compare the performance of the four methods (Rapaport et al., 2013). Receiver operating characteristic curves (ROC) plot these two rates against each other for each method and if one ROC curve is above another or the area under the curve (AUC) is larger, this indicates the superiority of one method over the other (Auer and Doerge, 2010).

4. Conclusions

For the Sun et al. data, PoissonSeq and edgeR had the largest intersection of differentially expressed genes; whereas DESeq and Cuffdiff shared the least number of differentially expressed genes (Figure 3). There were many differences in the lists of differentially expressed genes among the four methodologies.

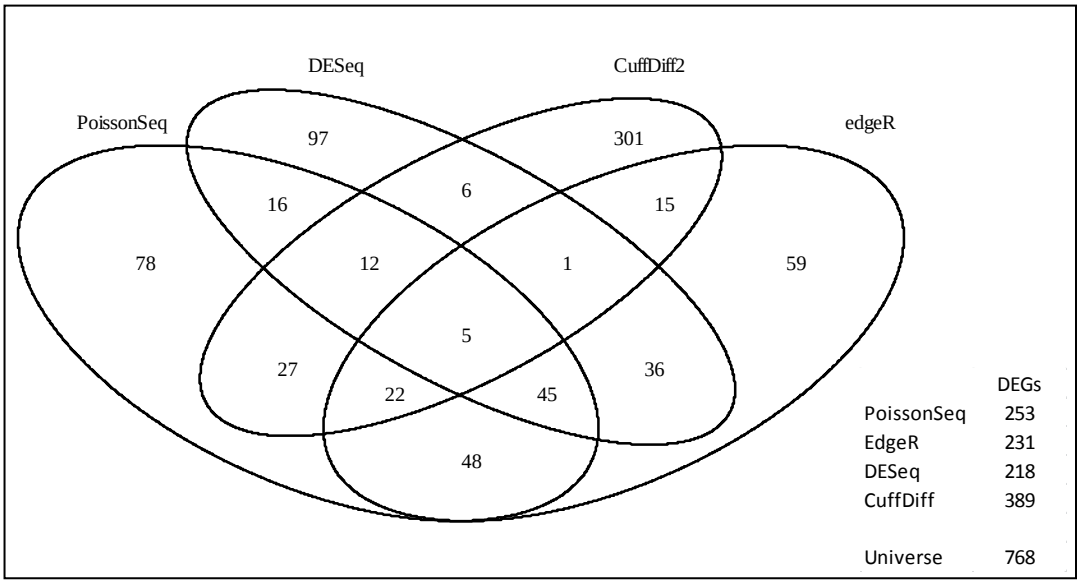

Figure 3. The intersection of the differentially expressed genes identified by the four different methodologies. The adjusted p-value cut-off was 0.05.

Rapaport et al. used the area under the receiver operating characteristic (ROC) curve (AUC) to evaluate the performance of the different techniques for the detection of differentially expressed genes. They found that edgeR and DESeq had a slight advantage in detecting differentially expressed genes reporting AUC for both methods 0.894; whereas, AUC for PoissonSeq was 0.878 and for Cuffdiff 0.865for an established benchmark dataset.

No single method is clearly superior for differential expression analysis, since each has particular strengths that may be suitable for specific RNA-Seq datasets (Rapaport et al., 2013). All four tools perform much better when there are biological or technical replicates available. Consistent with previous studies it suggests that biological replicates are a key factor for differential expression analysis in RNA-Seq datasets (Rapaport et al., 2013).

Acknowledgements

This project was also supported by the Vermont Genetics Network through Grant Number 2P20RR016462 from the INBRE Program of the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

References

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.

Auer, P. L. and Doerge R. W. (2010) Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, 185(2).

Goff, L., Trapnell C., and Kelley, D. (2012). cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.. R package version 2.4.1.

Guo et al. (2013) Evaluation of read count based RNAseq analysis methods. *BMC Genomics* 14(Suppl 8):S2.

Lee, J., Ji, Y., Liang, S., Cai, G. and Müller, P. (2011) On differential gene expression using RNA-Seq data. *Cancer Informatics*, vol. 10, pp. 205–215.

Li,J., Witten,D.M., Johnstone,I.M. and Tibshirani,R. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13, 523–538.

McGettigan, P. A. (2013) Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology*, 17(1): 4–11.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, M., Zumbo, P., Mason, C., Socci, N. and Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* 14 (9):R95.

Robinson, M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140.

Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, Carr JM, Khrebtukova I, Luo S, Zhang L, Schroth GP, Perez EA, Thompson EA. (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. PLoS ONE, 6

Vihinen, M (2012) How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* 13(Suppl 4):S2.