The comparison of limma and DESeq2 in gene analysis

Yihan Tong¹

¹College of Innovation and Experiment, Northwest A&F University, China

Abstract. Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. With the development of techniques, many methods to analyze the differentially expressed (DE) genes have emerged, especially the downstream analysis approaches, such as limma, DESeq2, and edgeR. However, it is unclear whether using different methods leads to different results. This article has compared the results gained from DESeq2 and limma when conducting downstream analysis for RNA sequencing data. Evidently, the number of genes they found is different from each other. DESeq2 found more genes than limma. But more than 90% of the genes detected by the two methods are overlapped, which means both methods are reliable. If precise results are needed, limma has a better ability to find the accurate DE genes. In the end, we analyzed the reason of the difference and summarized when it is better to use limma than DESeq2.

1 Introduction

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. The study of differences in gene-expression patterns is one of the most promising approaches for understanding mechanisms of differentiation and development [1]. Currently, biomedical research attempts to explain the mechanisms by which a particular disease develops. For this reason, gene expression studies have been proven to be a great resource [2]. Differentially-expressed genes can be valuable to pinpoint candidate biomarkers, therapeutic targets and gene signatures for diagnostics [3]. For instance, analyzing differentially-expressed genes is widely used in pharmaceutical and clinical research, such as tumor diagnose [4] and cancer prognosis [5], which will help human to search and solve many diseases problems.

Since the 20th century, many techniques to gain differentially-expressed genes have emerged, such as RT-qPCR [6], Microarrays [7] and RNA sequencing [8]. However, when we get the differentially-expressed genes, we always need to handle with those random data using a lot of statistical methods. The most common method we use is the packages in R, such as limma [9], DESeq2 [10] and edge R [11]. These three R packages are most frequently used to analyze data of differentially-expressed genes, but whether or not different packages to deal with the same data will affect the results is uncertain. In this article, we focus on the differences between limma and DESeq2, especially when they deal with the same set of genes. We find that although we used the same data, we gained different results. The results indicated that different gene expressions vary from different statistic methods, and also suggest that we can choose different packages or

bind them together to analyze the data according to the need of the experiment.

2 Materials and Methods

2.1 Specimen and Data Collection

Our data comes from an article written by Sheridan et al. in 2015 from BMC Cancer. The data used here are available through GEO Series accession number GSE63310.

2.2 Data Analysis

In this article, we mainly talk about the difference between DESeq2 (version 1.30.0) and limma (version 3.20.5). Data were TMM normalized [12] and transformed into log2 counts per million. Linear models with observational-level weights [13] were fitted to obtain average expression values for each gene in each sample type and moderated t-statistics were used to assess differential expression between populations [14] using the DESeq2 and limma packages. Then, we used FDR<0.01 to filter the genes. Where applicable, the Student's t-test was used, with $p < 0.05$ considered statistically significant.

2.3 Results Visualization

We divide our results into up regulated and down regulated genes, then we gain a bar graph of the total number of differentially expressed genes by Micro Office and the Venn diagrams of different samples by Draw Venn Diagram

* Corresponding author: tongyihan720@163.com

(<http://bioinformatics.psb.ugent.be/webtools/Venn/>).

2.4 Gene annotation

The second column in our DGE list is named genes, which is used to store the gene information correlated to the counting matrix. So, we use R package *Mus.musculus* (version 1.3.1) to annotate every DE gene according to their ENTREZ ID.

3 Result and Discussion

To identify whether using limma and DESeq2 gains the same results when analyze the differentially expressed (DE) genes, we chose the sample from an article written by Sheridan et al. in 2015 from BMC Cancer, whose GEO Series accession number is GSE63310. We know little about the molecular regulators that orchestrate stem cell renewal, proliferation and differentiation along the mammary epithelial hierarchy. To learn more about the mechanism, this experiment uses a large-scale pooled RNAi screen in primary mouse mammary stem cell (MaSC)-enriched basal cells using 1295 shRNAs against genes principally involved in transcriptional regulation and revealed a number of novel genes that influence the activity or survival of mammary stem and/or progenitor cells. This experiment concludes three groups of cells, which are basal, liminal progenitor (LP), and mature luminal (ML). The samples were sequenced by Illumina HiSeq2 2000 and then got the 100bp single-end sequencing fragment.

According to the flows of different packages, in limma, 100 bp single-ended reads were compared to mouse reference genome (MM10), while DESeq2 were not. This is not a significant difference in DE analysis

process, since it only had a slight influence on the results. Then, limma used *featureCounts* to calculate the count of each gene before TMM was used to normalize the genes, and the results were transferred to log2 counts per million. Finally, the average expression value of each gene in each sample was fitted by the linear model of level weight observation using limma, while DESeq2 did differentially expressed genes analysis only by using DESeq in package DESeq2. Genes were sequenced by FDR and log2-fold-change in both packages.

We described both DESeq2 and limma workflows for analyzing RNA-seq data, using gene-level counts as input, after samples organization, gene annotation, data scale conversion, low expression gene deletion, normalization, P-value adjustment (P value<0.05) in both two methods, then we obtained DE genes and a list of gene signatures.

We sorted the results into groups Basal vs LP and Basal vs ML, which included down regulated genes and up regulated genes respectively. We used limma and DESeq2 respectively to get the results of differentially expressed genes. Trough package limma, we totally got 16,624 genes, while DESeq2 got 21,166 genes in all, which had 4,542 genes more than the former. As for the total number of differentially expressed genes using different packages when we analyze up and down regulated genes, we discovered that DESeq2 usually got more genes than limma. However, we observed that when it comes to down regulated genes in group Basal vs ML, limma found slightly more genes than DESeq2, with 4927 genes and 4898 genes respectively, and the extra number took up only 29 more genes, 0.1% of all, which means that we can roughly ignore the difference and regard them finding the same number of genes. (Fig.1)

Fig.1 Total number of differentially expressed genes using different packages.

Fig.2 The intersection and complement of Venn diagrams of total genes using different packages in Basal vs LP and Basal vs ML.

Next, we go deeper to look at the results. As like the total number of DE genes, it is noticeable DESeq2 got more unique genes while limma got more uniquely up regulated genes in Basal vs ML.(Fig.2). (A) The up and down regulated genes that are shared and unique in results got from limma and DESeq2 respectively. No matter in up regulated genes or down regulated genes results, most of the gene results are shared in both packages. For instance, as for the group of basal vs LP, there are 4566 genes and 4319 genes overlapped in up regulated and down regulated genes respectively. However, the unique results from limma are 100 genes or so less than the unique genes found by DESeq2. For example, in up regulated genes, unique genes which are got from limma are in number of 298, 111 genes less than those got from DESeq2. (B)The two Venn diagrams show the results in group of basal vs ML. The amount of uniquely up regulated genes found by DESeq2 shows significantly high numbers than that found by limma in up regulated genes. The particularly down regulated genes share a similar amount both in the results got from limma and DESeq2 severally, with only 29 more genes found by limma.

We found those unique genes in our data and compared to the original results used to draw the Venn diagram, interestingly, almost every group of results has almost 25% of genes whose p-value >0.04 as well as 20% of genes who have a low p-value (p-value <0.01) (Supplementary Material). In fact, those who has a high

p-value or low p-value can be hard to tell whether this is a special function in DESeq2 or limma, because those results are vague since they are at the edge of the ability that the program can distinct, but except this, there are still almost 50% genes having a medium p-value. According to the number of those genes, it's not difficult to tell that DESeq2 still gets more results than limma.

According to our results, since most genes (over 90%) are overlapped between DESeq2 and limma, we can tell that both packages are reliable. Plus, it is obvious that limma is more sensitive to the data. There are several reasons for that: limma has the ability to incorporate quantitative weights into all levels of the statistical analysis, from normalization to linear modelling and gene set testing [15]. This ability allows increasing power to detect differentially expressed genes, and having a model-based approach avoids the need for ad hoc decisions about which observations or samples to filter out.[16] Secondly, batch effect has been removed when using limma, while DESeq2 only detects the differentially expressed genes, this can be the most significant reason leading to this result. Lastly, limma accepts a DGEList object from the edgeR package. The read counts are processed by the voom function in limma to calculate the weight of genes to adjust the means and variance to filter the results, while DESeq2 doesn't do this.

4 Conclusion

In this article, we figure out that although limma and DESeq2 are quite frequently used in gene expression analysis, they still have many differences. For instance, DESeq2 uses the non-linear model while limma uses the linear model. In addition, different mechanisms lead to different number of results, such as batch effect removal, which can be the main reason for this result. Overall, there are not too much difference between two packages. More than 90% of the genes detected in each group were overlapped between two methods, which means that each method is quite reliable when doing the downstream analysis. In conclusion, if a rough result is acceptable, both packages are suitable, but if a precise result is required, limma is recommended.

Reference

1. Kozian DH, Kirschbaum BJ. Comparative gene-expression analysis. *Trends Biotechnol*, **17(2)**: 73-78. (1999)
2. Segundo-Val IS, Sanz-Lozano CS. Introduction to the Gene Expression Analysis. *Methods Mol Biol*. 1434:29-43. (2016)
3. Rodriguez-Esteban R, Jiang X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med Genomics*. **10(1)**:59. (2017)
4. Bismeyer T, van der Velden BHM, Canisius S, Lips EH, Loo CE, Viergever MA, Wesseling J, Gilhuijs KGA, Wessels LFA. Radiogenomic Analysis of Breast Cancer by Linking MRI Phenotypes with Tumor Gene Expression. *Radiology*. **296(2)**:277-287 (2020)
5. Wang L, Luo X, Cheng C, Amos CI, Cai G, Xiao F. A gene expression-based immune signature for lung adenocarcinoma prognosis. *Cancer Immunol Immunother*. **69(9)**:1881-1890. (2020)
6. Rocha DJ, Santos CS, Pacheco LG. Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. *Antonie Van Leeuwenhoek*. **108(3)**:685-93. (2015)
7. Hung JH, Weng Z. Analysis of Microarray and RNA-seq Expression Profiling Data. *Cold Spring Harb Protoc*. 2017(3). (2017)
8. Ramsköld D, Kavak E, Sandberg R. How to analyze gene expression using RNA-sequencing data. *Methods Mol Biol*. **802**:259-274. (2012)
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. **43(7)**:e47. (2015)
10. Varet H, Brillet-Guéguen L, Coppée JY, Dillies MA. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS One*. **11(6)**:e0157022. doi: 10.1371/journal.pone.0157022. (2016)
11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26(1)**:139-140. (2010)
12. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11(3)**:R25. (2010)
13. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. **15(2)**:R29. (2014)
14. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 3:Article3.(2004)
15. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. **43(7)**:e47. (2015)
16. Ritchie ME, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, Smyth GK. Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics*. 7:261. (2006)