Journal of Visualized Experiments

Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2.

--Manuscript Draft--

Article Type:	Invited Methods Collection - Author Produced Video
Manuscript Number:	JoVE62528R2
Full Title:	Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2.
Corresponding Author:	Yeqiang Liu CHINA
Corresponding Author's Institution:	
Corresponding Author E-Mail:	lyqdoctor@163.com
Order of Authors:	Shiyi Liu
	Zitao Wang
	Ronghui Zhu
	Feiyan Wang
	Yanxiang Cheng
	Yeqiang Liu
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$1200)
Please specify the section of the submitted manuscript.	Cancer Research
Please confirm that you have read and agree to the terms and conditions of the author license agreement that applies below:	I agree to the Author License Agreement
Please provide any comments to the journal here.	
Please indicate whether this article will be Standard Access or Open Access.	Standard Access ($1400)
Please confirm that you have read and agree to the terms and conditions of the video release that applies below:	I agree to the Video Release

TITLE:

Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2

AUTHORS:

Shiyi Liu[1*], Zitao Wang[1*], Ronghui Zhu[1], Feiyan Wang[2], Yanxiang Cheng[1], Yeqiang Liu[2]

[1]Department of Obstetrics and Gynecology, Renmin Hospital of Wuhan University, Wuhan, Hubei Province, P.R.C.

[2]Department of Pathology, Shanghai Skin Disease Hospital, Tongji University School of Medicine, Shanghai, Shanghai, P.R.C

Shiyi Liu and Zitao Wang contributed equally to this work.

Shiyi Liu, shiyiliu@whu.edu.cn
Zitao Wang, 2020283020254@edu.cn
Ronghui Zhu, 2020283020243@whu.edu.cn
Feiyan Wang, 496128208@qq.com

CORRESPONDING AUTHORS:

Yanxiang Cheng, yanxiangCheng@whu.edu.cn
Yeqiang Liu, 1500156@tongji.edu.cn

SUMMARY:

A detailed protocol of differential expression analysis methods for RNA sequencing was provided: limma, EdgeR, DESeq2.

ABSTRACT:

RNA sequencing (RNA-seq) is one of the most widely used technologies in transcriptomics as it can reveal the relationship between the genetic alteration and complex biological processes and has great value in diagnostics, prognostics, and therapeutics of tumors. Differential analysis of RNA-seq data is crucial to identify aberrant transcriptions, and limma, EdgeR and DESeq2 are efficient tools for differential analysis. However, RNA-seq differential analysis requires certain skills with R language and the ability to choose an appropriate method, which is lacking in the curriculum of medical education.

Herein, we provide the detailed protocol to identify DEGs between cholangiocarcinoma (CHOL) and normal tissues through limma, DESeq2 and EdgeR, respectively, and the results are shown in volcano plots and Venn diagrams. The three protocols of limma, DESeq2 and EdgeR are similar but have different steps among the processes of the analysis. For example, a linear model is used for statistics in limma, while the negative binomial distribution is used in edgeR and DESeq2. Additionally, the normalized RNA-seq count data is necessary for EdgeR and limma but is not necessary for DESeq2.

Here, we provide a detailed protocol for three differential analysis methods: limma, EdgeR and

DESeq2. The results of the three methods are partly overlapping. All three methods have their own advantages, and the choice of method only depends on the data.

INTRODUCTION:

RNA-sequencing (RNA-seq) is one of the most widely used technologies in transcriptomics with many advantages (e.g., high data reproducibility), and has dramatically increased our understanding of the functions and dynamics of complex biological processes[1,2]. Identification of aberrate transcripts under different biological context, which are also known as differentially expressed genes (DEGs), is a key step in RNA-seq analysis. RNA-seq makes it possible to get a deep understanding of pathogenesis related molecular mechanisms and biological functions. Therefore, differential analysis has been regarded as valuable for diagnostics, prognostics and therapeutics of tumors[3-5]. Currently, more open-source R/Bioconductor packages have been developed for RNA-seq differential expression analysis, particularly limma, DESeq2 and EdgeR[1,6,7]. However, differential analysis requires certain skills with R language and the ability to choose the appropriate method, which is lacking in the curriculum of medical education.

In this protocol, based on the cholangiocarcinoma (CHOL) RNA-seq count data extracted from The Cancer Genome Atlas (TCGA), three of the most known methods (limma[8], EdgeR[9] and DESeq2[10]) were carried out, respectively, by the R program[11] to identify the DEGs between CHOL and normal tissues. The three protocols of limma, EdgeR and DESeq2 are similar but have different steps among the processes of the analysis. For example, the normalized RNA-seq count data is necessary for EdgeR and limma[8, 9], whereas DESeq2 uses its own library discrepancies to correct data instead of normalization[10]. Furthermore, edgeR is specifically suitable for RNA-seq data, while the limma is used for microarrays and RNA-seq. A linear model is adopted by limma to assess the DEGs[12], while the statistics in edgeR are based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests[9].

In summary, we provide the detailed protocols of RNA-seq differential expression analysis by using limma, DESeq2 and EdgeR, respectively. By referring to this article, users can easily perform the RNA-seq differential analysis and choose the appropriate differential analysis methods for their data.

PROTOCOL:

NOTE: Open the R-studio program and load R file “DEGs.R”, the file can be acquired from Supplementary files/Scripts.

1. Downloading and pre-processing of data

1.1. Download the high-throughput sequencing (HTSeq) count data of cholangiocarcinoma (CHOL) from The Cancer Genome Atlas (TCGA). This step can be easily achieved by the following R code.

1.2. Click **Run** to install R packages.

1.3. Click **Run** to load R packages.

```
if(!requireNamespace("BiocManager", quietly=TRUE))
  + install.packages("BiocManager")
BiocManager::install(c("TCGAbiolinks", "SummarizedExperiment"))
```

1.4. Set the working directory.

```
library (TCGAbiolinks)
library(SummarizedExperiment)
setwd("C:/Users/LIUSHIYI/Desktop")
```

1.5. Choose the cancer type.

```
cancer <- "TCGA-CHOL"
```

1.6. Run the R code from the "GDCquery.R" file to download the data. The file "GDCquery.R" can be acquired from Supplementary files/Scripts:

```
source("Supplementary files/Scripts/GDCquery.R")
head(cnt)
##TCGA-3X-AAVA-01A-11R-A41I-07
##ENSG00000000003                4262
##ENSG00000000005                   1
##ENSG00000000419                1254
##ENSG00000000457                 699
##ENSG00000000460                 239
##ENSG00000000938                 334
```

NOTE: After execution, the CHOL HTSeq count data will be downloaded and named “cnt”, where rows represent ensemble gene IDs and columns represent sample IDs. Please notice the numbers at positions 14-15 in the sample IDs; numbers ranging from 01 to 09 indicate tumors and ranging from 10 to 19 indicate normal tissues.

1.2 Convert ensemble gene IDs to gene symbols.

1.2.1. Import the annotation file into R according to its storage path. The annotation file (gencode.v22.annotation.gtf) can be acquired from Supplementary files.

```
gtf_v22 <- rtracklayer::import('Supplementary files/gencode.v22.annotation.gtf')
```

1.2.2. Run the R code from the "gtf_v22.R" file, which can be acquired from Supplementary files/Scripts:

```
source("Supplementary files/Scripts/gtf_v22.R")
```

1.2.3. Apply the function "ann" to convert the ensemble gene IDs to gene symbols.

```
cnt=ann(cnt,gtf_v22)
```

1.3 Filtering low-expressed genes

1.3.1. Click **Run** to install the R package "edgeR".

```
BiocManager::install("edgeR")
```

1.3.2. Click **Run** to load the R package "edgeR".

```
library(edgeR)
```

1.3.3. Run the following R code to keep genes with counts per million (CPM) values greater than one in at least two samples.

```
keep <- rowSums(cpm(cnt)>1)>=2
cnt <- as.matrix(cnt[keep,])
```

NOTE: The counts per million (CPM) value is used instead of the read count to eliminate the deviation caused by different sequencing depths.

2. Differential expression analysis through “limma”

2.1. Click **Run** to install the R package "limma".

```
BiocManager::install("limma")
```

2.2. Click **Run** to load the R packages "limma", "edgeR".

```
library(limma)
library(edgeR)
```

2.3. Run the following R code to create the design matrix.

```
group <- substring(colnames(cnt),14,15) # Extract group information
group [group %in% "01"] <- "Cancer" # set ‘01’ as tumor tissue
group [group %in% "11"] <- "Normal" # set ‘11’ as normal tissue
group <- factor (group, levels = c("Normal","Cancer"))
```

2.3.1. Create the design matrix.

```
design <- model.matrix (~group)
rownames(design) <- colnames(cnt)
```

2.3.2. Create the DGEList object.

```
dge <- DGEList(counts = cnt, group = group)
```

2.3.3. Normalize the data.

```
dge <- calcNormFactors(dge, method = "TMM")
```

2.3.4. Run the following R code to perform the limma-trend method based differential expression analysis.

```
dge
##An object of class "DGEList"
##$counts
##TCGA-3X-AAVA-01A-11R-A41I-07
##TSPAN6                 4262
##DPM1                  1254
##SCYL3                 699
##C1orf112               239
##FGR                   334
```

2.3.5. Calculate the CPM value.

```
logdge <- cpm(dge, log=TRUE, prior.count=3)
```

2.3.6. Click **Run** to fit a linear model to predict the data or infer the relationship between variables.

```
fit <- lmFit (logdge, design)
```

2.3.7. Calculate the T value, F value and log-odds based on Bayesian.

```
fit <- eBayes(fit, trend=TRUE)
```

2.3.8. Extract the result table.

```
res_limma<- as.data.frame(topTable(fit,n=Inf))

head(res_limma)
##              logFC    AveExpr      t       P.Value      adj.P.Val       B
##RP11-252E2.2  -4.899493 -2.488589 -20.88052 2.386656e-25 4.931786e-21 47.28823
##BX842568.1    -4.347930 -2.595205 -20.14532 1.082759e-24 1.118706e-20 45.83656
##CTC-537E7.3   -5.154894 -2.143292 -19.59571 3.452354e-24 2.216114e-20 44.72001
##RP11-468N14.3 -6.532259 -2.029714 -19.49409 4.289807e-24 2.216114e-20 44.51056
##AP006216.5    -4.507051 -2.670915 -19.25649 7.153356e-24 2.956339e-20 44.01704
##RP11-669E14.4 -4.107204 -2.828311 -18.93246 1.448209e-23 4.987633e-20 43.33543
```

The result of differential expression analysis is saved in “res_limma”, which includes the gene id, log2 fold change value (logFC), the average log2 expression level of the gene in the experiment (AveExpr), the modified t statistic (t), relavent p value (P.Value), the false discovery rate (FDR) corrected p value (adj.P.Val) and the log-odds of differentially expressed genes (B)

NOTE: The function “calcNormFactors()” of the “edgeR” was used to normalize the data to eliminate the influence caused by sample preparation or library construction and sequencing. In the construction of design matrix, it is necessary to match experimental design (e.g., tissue type: normal or tumor tissues) to sample IDs of the matrix. limma-trend is suitable to data whose sequencing depth is the same, while limma-voom is suitable: (i) when the sample library size is different; (ii) data not normalized by TMM; (iii) there is a lot of “noise” in the data. A positive logFC means that gene is up-regulated in the experiment, while negative number means that gene is down-regulated.

2.3.9. Identify the DEGs.

```
res_limma$sig <- as.factor(
  ifelse(res_limma$adj.P.Val < 0.05 & abs(res_limma$logFC) > 2,
       ifelse(res_limma$logFC > 2 ,'up','down'),'not')) # The adj.p Value < 0.05 and the |log2FC| >= 2 are thresholds to identify the DEGs
summary(res_limma$sig)
##down   not    up
##1880 17341  1443
```

2.3.10. Output the result table to a file.

```
write.csv(res_limma, file = 'result_limma.csv')
```

2.3.11. Click **Run** to install the R package "ggplot2".

```
install.packages("ggplot2")
```

2.3.12. Click **Run** to load the R package "ggplot2".

```
library(ggplot2)
```

2.3.13. Run the R code from the "volcano.R" to create the volcano plot. The file "volcano.R" can be acquired from Supplementary files.

```
source("Supplementary files/Scripts/volcano.R")
volcano(res_limma,"logFC","adj.P.Val",2,0.05)
```

NOTE: Genes can be mapped to different positions according to their log2FC and adj-p values, the up regulated DEGs are colored in red, and the down-regulated DEGs are colored in green.

2.3.14. Click **Export** to save the volcano plot.

NOTE: The volcano plots can be generated and downloaded in different formats (e.g., pdf, TIFF, PNG, JPEG format). Genes can be mapped to different positions according to their log2FC and adj *p* values, the up-regulated DEGs (log2FC > 2, adj p < 0.05) are colored in red, and the down-regulated DEGs (log2FC < -2, adj p < 0.05) are colored in green, non-DEGs are colored in grey.

3. Differential expression analysis through “edgeR”

3.1. Click **Run** to load the R package "edgeR".

```
library(edgeR)
```

3.2. Run the following R code to create design matrix.

```
group <-substring(colnames(cnt),14,15)
group [group %in% "01"] <- "Cancer"
group [group %in% "11"] <- "Normal"
group=factor(group, levels = c("Normal","Cancer"))
```

```
design <-model.matrix(~group)
rownames(design) = colnames(cnt)

3.3. Click Run to create the DGEList object.
dge <- DGEList(counts=cnt)

3.4. Normalize the data.
dge <- calcNormFactors(dge, method = "TMM")

3.5. Click Run to estimate the dispersion of gene expression values.
dge <- estimateDisp(dge, design, robust = T)

3.6. Click Run to fit model to count data.
fit <- glmQLFit(dge, design)

3.7. Conduct a statistical test.
fit <- glmQLFTest(fit)

3.8. Extract the result table. The result is saved in “res_edgeR”, which includes the log fold change
value, log CPM, F, p value and FDR corrected p value.
res_edgeR=as.data.frame(topTags(fit, n=Inf))
head(res_edgeR)
##         logFC   logCPM       F       PValue         FDR
##GCDH  -3.299633 5.802700 458.5991 1.441773e-25 2.979280e-21
##MSMO1 -3.761400 7.521111 407.0416 1.730539e-24 1.787993e-20
##RCL1  -3.829504 5.319641 376.5043 8.652474e-24 5.516791e-20
##ADI1  -3.533664 8.211281 372.6671 1.067904e-23 5.516791e-20
##KCNN2 -5.583794 3.504017 358.6525 2.342106e-23 9.679455e-20
##GLUD1 -3.287447 8.738080 350.0344 3.848408e-23 1.194406e-19
# The result is saved in “res_edgeR”, which includes the log fold change value(logFC), log CPM, F,
p value and FDR corrected p value

3.9. Identify the DEGs.
res_edgeR$sig = as.factor(
  ifelse(res_edgeR$FDR < 0.05 & abs(res_edgeR$logFC) > 2,
        ifelse(res_edgeR$logFC > 2 ,'up','down'),'not'))
summary(res_edgeR$sig)
##down   not    up
##1578 15965  3121

3.10. Output the result table to a file.
write.csv(res_edgeR, file = 'res_edgeR.csv')

3.11. Create the volcano plot.
```

```
volcano(res_edgeR,"logFC","FDR",2,0.05)
```

3.12. Click **Export** to save the volcano plot.

4. Differential expression analysis through “DESeq2”

4.1. Click **Run** to install R packages "DESeq2".

```
BiocManager::install("DESeq2")
```

4.2. Click **Run** to load R packages "DESeq2".

```
library(DESeq2)
```

4.3. Run the following R code to determine the grouping factor.

```
group <-substring(colnames(cnt),14,15)
group [group %in% "01"] <- "Cancer"
group [group %in% "11"] <- "Normal"
group=factor(group, levels = c("Normal","Cancer"))
```

4.4. Create the DESeqDataSet object.

```
dds <-DESeqDataSetFromMatrix (cnt, DataFrame(group), design = ~group)
dds
##class: DESeqDataSet
##dim: 20664 45
##metadata(1): version
##assays(1): counts
##rownames(20664): TSPAN6 DPM1 ... RP11-274B21.13 LINC01144
##rowData names(0):
##colnames(45): TCGA-3X-AAVA-01A-11R-A41I-07 ...
##colData names(1): group
```

4.5. Perform the analysis.

```
dds <- DESeq(dds)
```

4.6. Generate the result table.

```
res_DESeq2 <- data.frame(results(dds))

head(res_DESeq2)
##          baseMean      log2FoldChange     lfcSE      stat       pvalue         padj
##TSPAN6    4704.9243     -0.8204515 0.3371667 -2.433370 1.495899e-02 2.760180e-02
##DPM1      1205.9087     -0.3692497 0.1202418 -3.070894 2.134191e-03 4.838281e-03
##SCYL3      954.9772      0.2652530 0.2476441  1.071106 2.841218e-01 3.629059e-01
##C1orf112   277.7756      0.7536911 0.2518929  2.992109 2.770575e-03 6.101584e-03
##FGR        345.8789     -0.6423198 0.3712729 -1.730047 8.362180e-02 1.266833e-01
##CFH      27982.3546     -3.8761382 0.5473363 -7.081823 1.422708e-12 1.673241e-11
```

NOTE: The result is saved in “res_DESeq2”, which includes the mean of the normalized read count (baseMean), log fold Change value(log2FoldChange), log fold change standard error (lfcSE), the Wald statistic (stat), original p value (pvalue) and corrected p value (padj)

4.7. Identify DEGs.

```
res_DESeq2$sig = as.factor(
  ifelse(res_DESeq2$padj < 0.05 & abs(res_DESeq2$log2FoldChange) > 2,
       ifelse(res_DESeq2$log2FoldChange > 2 ,'up','down'),'not'))
summary(res_DESeq2$sig)
##down   not    up
##1616 16110  2938
```

4.8. Output the result table to a file.

```
write.csv(res_DESeq2, file = 'res_DESeq2.csv')
```

4.9. Create the volcano plot.

```
volcano(res_DESeq2,"log2FoldChange","padj",2,0.05)
```

4.10. Click **Export** to save the volcano plot.

5. Venn diagram

5.1. Click **Run** to install the R package "VennDiagram".

```
install.packages("VennDiagram")
```

5.2. Click **Run** to load the R package "VennDiagram".

```
library (VennDiagram)
```

5.3. Make a Venn diagram of up regulated DEGs.

```
grid.newpage()
grid.draw(venn.diagram(list(Limma=rownames(res_limma[res_limma$sig=="up",]),
              edgeR=rownames(res_edgeR[res_edgeR$sig=="up",]),
              DESeq2=rownames(res_DESeq2[res_DESeq2$sig=="up",])),
            NULL,height = 3,width = 3,units = "in",
            col="black",lwd=0.3,fill=c("#FF6666","#FFFF00","#993366"),
            alpha=c(0.5, 0.5, 0.5),main = "Up-regulated DEGs"))
```

5.4. Click **Export** to save the Venn diagram.

5.5. Make a Venn diagram of down regulated DEGs.

```
grid.newpage()
grid.draw(venn.diagram(list(Limma=rownames(res_limma[res_limma$sig=="down",]),
              edgeR=rownames(res_edgeR[res_edgeR$sig=="down",]),
```

```
    DESeq2=rownames(res_DESeq2[res_DESeq2$sig=="down",])),
NULL,height = 3,width = 3,units = "in",
col="black",lwd=0.3,fill=c("#FF6666","#FFFF00","#993366"),
alpha=c(0.5, 0.5, 0.5),main = "Down-regulated DEGs"))
```

5.6. Click **Export** to save the Venn diagram.

REPRESENTATIVE RESULTS:

There are various approaches to visualize the result of differential expression analysis, among which the volcano plot and Venn diagram are particularly used. limma identified 3323 DEGs between the CHOL and normal tissues with the |logFC|≥2 and adj.P.Val <0.05 as thresholds, among which 1880 were down-regulated in CHOL tissues and 1443 were up-regulated (**Figure 1a**). Meanwhile, edgeR identified the 1578 down-regulated DEGs and 3121 up-regulated DEGs (**Figure 1b**); DESeq2 identified the 1616 down-regulated DEGs and 2938 up-regulated DEGs (**Figure 1c**). Comparing the results of these three methods, 1431 up-regulated DEGs and 1531 down-regulated DEGs were overlapped (**Figure 2**).

FIGURE AND TABLE LEGENDS:

Figure 1. Identification of differentially expressed genes (DEGs) between CHOL and normal tissues. (**a-c**) The volcano plots of all genes acquired by limma, edgeR and DESeq2, respectively, *adj p* value (-log10) is plotted against the fold change (log2), red points represent the up-regulated DEGs (adjusted *p* value<0.05 and log |FC|> 2) and the green points represent the down-regulated DEGs (adjusted *p* value< 0.05 and log |FC|< 2).

Figure 2. Venn diagrams show overlap among the results derived from the limma, edgeR and DESeq2.

DISCUSSION:

Abundant aberrate transcripts in cancers can be easily identified by RNA-seq differential analysis[5]. However, the application of RNA-seq differential expression analysis is often restricted as it requires certain skills with R language and the capacity to choose appropriate methods. To address this problem, we provide a detailed introduction to the three most known methods (limma, EdgeR and DESeq2) and tutorials for applying the RNA-seq differential expression analysis. This will facilitate the understanding of the similarities and differences across all three methods, enable the selection of a suitable method for individual data, and enable us to understand the complex dynamic biological processes.

Here, we present a detailed protocol for RNA-seq differential expression analysis through limma, edgeR and DESeq2 respectively, in five stages: (i) downloading and pre-processing of data, (ii-iv) differential expression analysis through limma, edgeR and DESeq2, respectively, (v) comparison of the results of these three methods through a Venn diagram.

The three methods have similar and different steps among the processes of the differential expression analysis. A linear model is used for statistics in limma, which is applicable for all gene

expression technologies, including microarrays, RNA-seq and quantitative PCR[8,13], while edgeR and DESeq2 implement a range of statistical methodologies based on the negative binomial distribution[9,10], and edgeR and DESeq2 are suitable for RNA-seq data. In addition, the normalized RNA-seq count data is necessary for EdgeR and limma, whereas DESeq2 uses its own library discrepancies to correct data instead of normalization and the data in DESeq2 must be an integer matrix. The normalization methods include TMM (trimmed mean of M-values), TMMwsp, RLE (relative log expression) and upperquartile, among which TMM is the most commonly used normalization method for RNA-seq data. The results of the three methods showed that DESeq2 and EdgeR obtain more DEGs than limma. The reason for this difference is that edgeR and DESeq2 are based on the negative binomial model, which contributes to large numbers of false positives. On the contrary, limma-voom only uses the variance function and does not show excessive false positives, as is the case with a variance stabilizing transformation followed by linear model analysis with limma[14-16].

All three methods have their own advantages, and the choice is just dependent on the type of data. For example, if there is microarray data, limma should be given with priority, but when it is the next-generation sequencing data, DESeq2 and EdgeR are preferred[9,10,17]. In summary, we provide here a detailed protocol for RNA-seq differential expression analysis with R packages limma, edgeR and DESeq2, respectively. The output results from the three methods are overlapping partly, and these differential methods have their respective advantages. Unfortunately, this protocol does not cover the technical details for other data types (e.g., microarray data) and methods (e.g., EBSeq)[18].

ACKNOWLEDGMENTS:

This work was supported by the National Natural Science Foundation of China (Grant No. 81860276) and Key Special Fund Projects of National Key R&D Program (Grant No. 2018YFC1003200).

DISCLOSURES:

The manuscript has not been published before and is not being considered for publication elsewhere. All authors have contributed to the creation of this manuscript for important intellectual content and read and approved the final manuscript. We declare there is no conflict of interest.

REFERENCES:

1. Tambonis, T., Boareto, M., Leite, V. B. P. Differential Expression Analysis in RNA-seq Data Using a Geometric Approach. *Journal of Computational Biology*. **25**, 1257-1265 (2018).
2. Wang, Z., Gerstein, M., Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*. **10**, 57-63 (2009).
3. Anders, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*. **8**, 1765-1786 (2013).
4. McDermaid, A., Monier, B., Zhao, J., Liu, B., Ma, Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in Bioinformatics*. **20**, 2044-2054 (2019).

5. Costa-Silva, J., Domingues, D., Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PloS One*. **12**, e0190152 (2017).
6. Law, C. W. et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*. **5** (2016).
7. Varet, H., Brillet-Guéguen, L., Coppée, J. Y., Dillies, M. A. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PloS One*. **11**, e0157022 (2016).
8. Ritchie, M. E.et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. **43**, e47 (2015).
9. Robinson, M. D., McCarthy, D. J., Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*. **26**, 139-140 (2010).
10. Love, M. I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. **15**, 550 (2014).
11. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*. **5**, R80 (2004).
12. Law, C. W., Chen, Y., Shi, W., Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. **15**, R29 (2014).
13. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*. **3** (2004).
14. Lund, S. P., Nettleton, D., McCarthy, D. J., Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*. **11**, (2012).
15. Reeb, P. D., Steibel, J. P. Evaluating statistical analysis models for RNA sequencing experiments. *Frontiers in Genetics*. **4**, 178 (2013).
16. Rocke, D. M. et al. Excess False Positive Rates in Methods for Differential Gene Expression Analysis using RNA-Seq Data. *bioRxiv* (2015).
17. Agarwal, A. et al. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC genomics*. **11**, 383 (2010).
18. Leng, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*. **29**, 1035-1043 (2013).

Figure 1

Click here to access/download;Figure;Figure 1.pdf

Figure 1

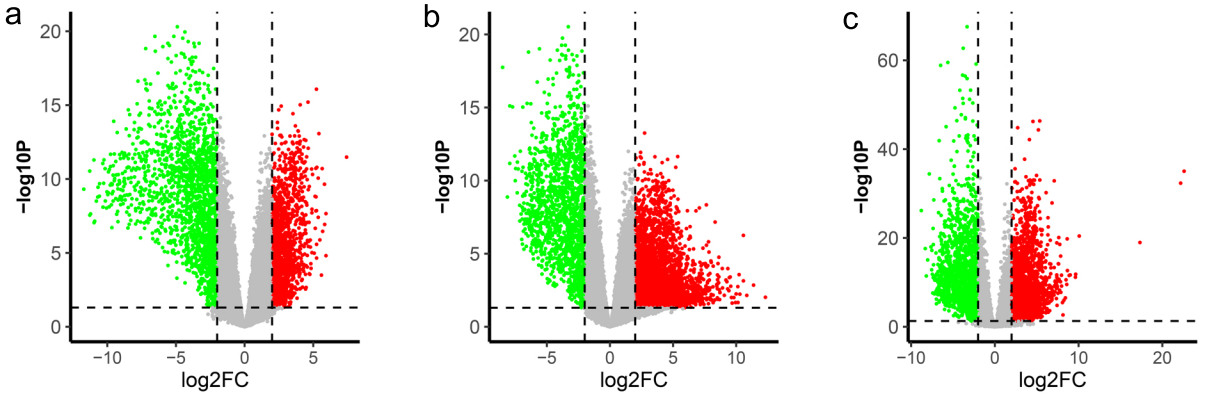
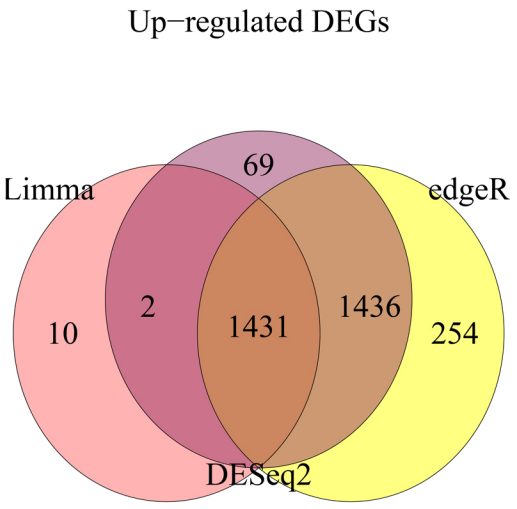
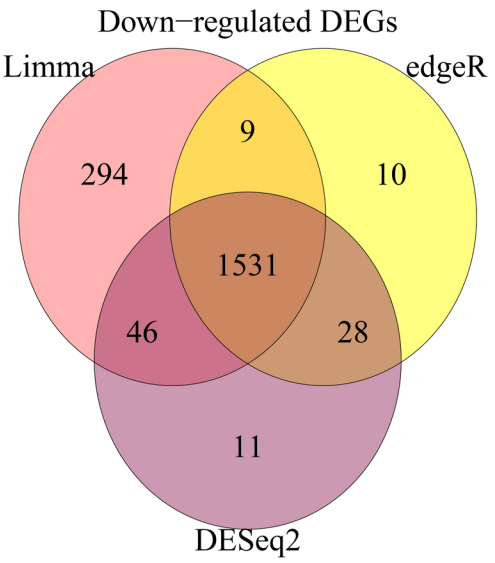


Figure 2





Reply to Reviewing Comments on the Revision

Manuscript ID: JoVE62528

Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2

Shiyi Liu[1], Zitao Wang[1], Ronghui Zhu[1], Feiyan Wang[2], Yanxiang Cheng[1], Yeqiang Liu[2]

[1] Renmin Hospital of Wuhan University,

[2] Tongji University School of Medicine

Explanation of this revision

First of all, we would like to express our sincere thanks to the editors and reviewers for their helpful comments and suggestions on our protocol and video. The explanation of the modifications as well as corrections in this revision are listed as follows (comment numbers are in point-by-point correspondence with the editor's and reviewers' comments).

We have carefully revised this protocol and re-recorded video by taking editor's and reviewers' comments into account. The changes are marked with yellow in the revised protocol (62528_R1).

Reply to Editorial and Production Comments

Changes to be made by the Author(s) regarding the written manuscript:

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues.

- Thank you very much for the remind. We have carefully checked and corrected the grammatical errors and typos in this protocol text.

2. Please move the coding text to a supplemental file and refer to them by name: for example, xx.py, etc. Use these references in the video as well.

- Thank you deeply for your comments. We have moved the coding text to the 'DEGs.R' file (Supplementary files/Scripts/DEGs.R), all other R files can be acquired from the 'Supplementary files/Scripts'.

3. Please ensure that all text in the protocol section is written in the imperative tense as if telling someone how to do the technique (e.g., "Do this," "Ensure that," etc.). The actions should be described in the imperative tense in complete sentences wherever possible. Avoid usage of phrases such as "could be," "should be," and "would be" throughout the Protocol.

- Thank you so much for your valuable suggestion. We have corrected that all text in the protocol section is written in the imperative tense. e.g., 'The numbers at positions 14-15 in the sample IDs should be noticed' have been revised to be 'Please notice the numbers at positions 14-15 in the sample IDs'. (Page 3, line 106)

4. The protocol text should contain the user input commands to run the scripts (run …) and not just the coding text.

- Indeed, this is a very good suggestion. We have added user input commands to run the scripts in protocol text (e.g., 'Click run to …', 'Click export to …').

Changes to be made by the Author(s) regarding the video:

1. Please increase the homogeneity between the video and the written manuscript. Ideally, the narration is a word for word reading of the written protocol.

- We have checked the manuscript and the video thoroughly. The video has been re-recorded with the text modified such that the homogeneity between them is increased. We ensure that the narration is a word for word reading of the protocol text.

2. When revising the video, please ensure that the video length Is under our 15 min limit.

- Thank you for the remind. The current video length is 07′ 54″, under 15 min limit.

3. Parts of the protocol narration are hard to make out. Please consider re-recording the voiceover to emphasize clarity. Please avoid large gaps of silence in the narration as well.

- Thanks for the suggestion. We have re-recorded the video and the voiceover to emphasize clarity. Large gaps of silence have been avoided in the narration. Please watch the video.

4. Title Cards

• In order to be consistent with JoVE's grammar style, consider the following changes:

please remove the asterisks from after author names

- Thank you. We have removed the asterisks after author names.

please delete the extra space after ""RNA,"" and the space before the colon, and add a comma after DESeq2

- We have deleted the extra space after 'RNA' and the space before the colon.

• consider swapping order of ""DESeq2"" and ""EdgeR"" in the title to match the order they appear in the interview statements and chapter order.

- We have swapped the order of 'DESeq2' and 'EdgeR' in the title, and the title is 'Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2'.

5. Video Content

• 3:47 - consider fading to white and fading back into the new screen for a clear but less jarring transition to 1.2

- The 'fading to white and fading back' have been applied to all transitions in

the re-recording video.

• 5:11, 8:54, 11:33, 13:52 - consider fading to white and fading back into the new screen for a less jarring transition between chapters

- The 'fading to white and fading back' have been applied to all transitions in the re-recording video.

• 15:19 - interview gets cut short. please consider letting it play amoment longer and fade out to be less jarring to viewers"

- Thanks for your advice. The interview part has been deleted in the re-recorded video due to the following reasons: (a) The video with the interview part is longer than 15 mins; (b) The content of the interview has been written in the introduction and discussion section of the protocol text.

6. Audio and Pacing

¤ At the following times, consider adding voiceover to explain what is being done here, or else increase the video speed or trim unimportant and previously explained details. Keep in mind that the final video will have to be under 15 minutes in total length before it is published:

• 2:33 - 2:57 - ""to install and load R packages ... [20 seconds of silence]""

• 2:57 - 3:16 - ""downloading the data .. [16 seconds of silence]"" Additionally, the narration here does not seem to be a full sentence, and does not come across as an instruction. JoVE videos are written in the imperative, to instruct viewers in how to best replicate the techniques shown.

• 4:17 - 4:46

• 4:55 - 5:02

• 5:36 - 5:50 - at 5:40 there seems to be another sentence fragment, at 5:50 we hear just ""construction of design matrix."" The narration is intended to describe what is being done on the screen.

• 5:53 - 6:09 [- 6:14] - at 6:09 we hear just ""normalization."" The narration is intended to describe what is being done on the screen.

• 4:55 - 5:02

• 8:08 - 8:26

• 11:10 - 11:18 ""saving"" here should be ""save,"" as JoVE videos are written in the imperative.

- We have re-recorded the video and the voiceover to (a) explain what is doing here, (b) eliminate the large gaps of silence (3s+) in the narration, (c) change the tense as imperative.

7. ▫ At the following time points, consider speeding up this section to reduce the total video duration and reduce the amount of silent video. JoVE makes sure there is instructional audio underneath relevant video. With this in mind, there should ideally not be 3+ seconds of silence in any given step.

• 6:50 - 6:56

• 7:09 - 7:17

• 7:31 - 7:38

• 7:52 - 8:00

- The 3s+ silences have been eliminated in the re-recording video.

• 9:15 - 9:22; at 9:22 - ""construction of design matrix as before"" is a sentence fragment and is not instructional. The narration should describe what a viewer should do to replicate the procedure shown.

- We have revised it as 'Run the following R code to create design matrix'.

• 9:25 - 9:44; at 9:44 - ""fitting model, statistics test"" is a sentence fragment and is not instructional. The narration should describe what a viewer should do to replicate the procedure shown.

• 9:47 - 10:03

• 12:02 - 12:20

• 12:57 - 13:05

- We have revised it as 'Click run to fit model to count data' and 'Conduct statistical test'.

• 13:32 - 13:38; at 13:38, ""saving"" here should be ""save,"" as JoVE videos function in the imperative.

• 13:41 - 13:48

• 14:17 - 14:22"

- We have revised it as 'Click "Export" to save the volcano plot' and 'Click "Export" to save the Venn diagram'.

8. • 7:59 - 8:08 audio is said very quickly/at an unnaturally fast speed, without any natural pauses. Consider re-recording this step. This seems to be the case for all spoken steps immediately following title cards".

- We have corrected it in re-recorded video.

9. • 14:36 - Consider extending the beginning and end of this interview, as the audio seems clipped (cannot hear first few words) at the beginning and end.

- The interview part has been deleted in the re-recorded video due to two reasons: (a) The video with the interview part is longer than 15 mins; (b) The content of the interview has been written in the introduction and discussion section of the protocol text.

Reply to Reviewer #1

We want to begin by thinking Reviewer #1 for writing that 'In general, the topic is a well-chosen one and is the requirement of several researchers and labs.' Our deepest gratitude also goes to you for your deep review of the manuscript. Your comments are all of great importance to our protocol. After this revision, we have written a point-by-point response letter to acknowledge your help and denote where we made revisions.

Reply to Comments.

Abstract:

Line # 31_Replace "diagnostic, prognostic and therapeutic" with "diagnostics, prognostics and therapeutics"

- Thank you very much. We have replaced it as '…having great value in diagnostics, prognostics and therapeutics of tumors.' (Page 1, line 33)

Line #32-35_ Kindly rephrase the sentence as the sentence in the current form looks very ambiguous.

- To make it clearer, we have rephrased the sentence as 'Differential analysis of RNA-seq data is crucial to identify the aberrant transcriptions, where the "limma", "EdgeR" and "DESeq2" are efficient tools for differential analysis. However, the RNA-seq differential analysis requires certain skills on R language and the capacity to choose appropriate method, which is lack in the curriculum of medical education'. (Page 1, line 34-37)

Line #38_ It is volcano plots and Venn diagrams.

-We have changed 'volcano and venn plots' to 'volcano plots and Venn diagrams' (Page 1, line 40)

Line #41-42_ Correct the sentence, "but "DESeq2" not"..??

-I we have corrected 'but "DESeq2" not' to be 'but is not necessary for "DESeq2".' (Page 1-2, line 43-44)

Line #43-46_ Kindly check and rephrase the sentences.

-We have carefully checked and revised the grammatical errors and typos in follow.

'In conclusion, we provide a detailed protocol for three differential analysis methods: limma, EdgeR and DESeq2. The results of the three methods are overlapping partly. All three methods have their own advantages, and the choice of method only depends on the data.' (Page 2, line 45-47)

Introduction:

Line #52_ change "....context, which also named differentially expressed genes (DEGs)" to " ...contexts, which are also known as differentially expressed genes (DEGs)".

-We have changed it to be as follow.

'…, which are also known as differentially expressed genes (DEGs), …' (Page 2, line 53-54)

Line #53-55_ Check the sentences for grammatical errors and rephrase.

- We have carefully checked and revised the grammatical errors and typos in this sentence. Now, it is better.

'It makes possible to get a deep insight of pathogenesis related molecular mechanisms and biological functions. Therefore, differential analysis has been regarded to be valuable for diagnostics, prognostics and therapeutics of tumors.' (Page 2, line 54-57)

Line #63_ Remove space between "proto cols".

-Done. (Page 2, line 63)

Protocol:

Change the heading to Downloading and pre-processing of data.

-The change has been made in the heading to be 'Downloading and pre-processing of data'. (Page 2, line 80)

Line #78_ Rephrase "realized" and complete the sentence in and unambiguous manner.

-We have changed the 'realized' to be 'achieved'. (Page 2, line 82)

Line #150_ Remove "RNA Seq". Just write "Differential expression analysis through 'limma'".

-We have removed the 'RNA Seq', the changes are listed as follows:

2 Differential expression analysis through "limma" (Page 3, line 131)

3 Differential expression analysis through "edgeR" (Page 5, line 208)

4 Differential expression analysis through "DESeq2" (Page 6, line 248)

Line #245_ Rephrase the sentence as "the volcano plots can be generated and downloaded in different formats".

-We have rephrased it. (Page 5, line 203)

Line #250_ Just write "Differential expression analysis through 'edgeR'".

- We have removed the 'RNA Seq'. (Page 5, line 208)

Discussion:

Line #451-452_ Correct the sentence "As abundant aberrate transcripts enriched in carcinogenesis, RNA-seq differential expression analysis plays a great role in RNA sequencing analysis(5)". It is ambiguous.

-We have corrected it as 'Abundant aberrate transcripts in cancers can be easily identified by RNA-seq differential analysis.' (Page 7, line 334-335)

Line #455_ Replace "its" with "their".

-Done. (Page 8, line 339)

Line #460-461_ Change the headings as suggested previously.

-We have corrected it. Please see the following revision.

'(i) downloading and pre-processing of data, (ii-iv) differential expression analysis through "limma", "edgeR" and "DESeq2", respectively, (v) comparison of the results of these three methods through Venn diagram.' (Page 8, line 343-345)

Line #468_ Kindly rephrase the sentence "...and "edgeR" and "DESeq2" are arisen for RNA-seq data".

-We have corrected it as 'and "edgeR" and "DESeq2" are suitable for RNA-seq data.'. (Page 8, line 350-351)

Reply to Reviewer #2

The authors would like to thank you for your deep review of the manuscript. Your time and efforts in handling our manuscript are much appreciated

Reply to Major Comments 1.

The three methods have been widely adopted in the literature, and there are already many papers that compare them and many other methods as well. The protocols listed in the manuscript are very basic R code, and these R codes are easily found in each of its own R package manual. The manuscript does not provide any new contribution.

- Even if the R codes of this protocol can be found online, the differential analysis requires certain skills on R language and the capacity to choose appropriate methods, which is lack in the curriculum of medical education.

In this protocol, the detailed processes of differential analysis are described. By referring to the step-by-step video and protocol, users can easily perform the RNA-seq differential analysis, and choose the appropriate differential analysis methods for their data.

The author concluded that three methods provides similar results based on Figure 2. However DESeq2 and edgeR have much more upregulated DEGs than limma. This is consistent with the literature that concerns about the extra false positives by DESeq2 and edgeR methods (references on this are listed in below).

1. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29(8): 1035-43.

2. Lund S, Nettleton D, McCarthy DJ, Smyth GK.(1, 2)differential expression in RNA-sequencing data using quasi-likelihood with shrunken dispersion estimates. Stat Appl Genet Mol Biol. 2012;11(5):. article 8.

3. Reeb PD, Steibel JP. Evaluating statistical analysis models for RNA sequencing experiments. Front Genet. 2013;4:178.

4. Rocke DM, Ruan L, Zhang Y, Gossett JJ, Durbin-Johnson B, Aviran S. Excess false positive rates in methods for differential gene expression analysis using RNA-Seq data. bioRxiv Preprint. 2015. http://dx.doi.org/10.1101/020784.

- Thank you very much for your suggestion. These articles have been cited and discussed in detail in the revised version.

'The results of three methods showed that "DESeq2" and "EdgeR" obtain more DEGs than "limma". The reason for this difference is that the "edgeR" and "DESeq2" are based on negative binomial model, which contribute to large numbers of false positives. On the contrary, limma-voom only use the variance function, do not show excessive false positives, as is also the case with a variance stabilizing transformation

followed by linear model analysis with limma(14-16). (Page 8, line 356-360)

……

The output results from the three methods are overlapping partly, and these differential methods have their respective advantages. Unfortunately, this protocol does not cover the technical details for other data types (e.g. microarray data) and methods (e.g. EBSeq)(18). (Page 8, line 365-368)

Thanks again for your suggestions.

Response to Editorial Comments

JoVE62528R1

"Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2"

- Changes to be made by the Author(s) regarding the written manuscript:

1. Please review the protocol text as I have edited it to fit our publication standard. We present the steps individually.

They are correct, thank you!

- Changes to be made by the Author(s) regarding the video:

1. Please include a results section in the video where Figure 1 and 2 are presented and discussed. This should have its own result title card.

We have added the result section with its own title card in video.

2. Please include a conclusion title card and a short interview section to conclude the video.

We have added the conclusion/interview section in the video.