Statistical analysis of RNA-Seq pseudocounts

Nima Hejazi

2016 Oct 31 (Mon), 22:29:03

Abstract

We present the results of performing linear modeling to determine differential expression of genes in the RNA-seq sample based on the method of `limma-voom`. Both a simple and full model were used to analyze differential expression of genes, and several visualizations of the analytic results are given, including two volcano plots as well as a heatmap of differential expression across sample groups.

Contents

I. Introduction

The statistical analysis procedure documented here uses the `limma-voom` method for performing linear modeling with RNA-seq pseudocounts produced by using the (pseudo)alignment algorithm of Bray *et al.* (2016) with the `Kallisto` software. Since the alignment procedure differs slightly from that used in standard alignment software (*e.g.*, `bowtie`), the counts are not integers; please consult the documentation of these tools if there are concerns about the method employed. After importing the transcript quantification results produced by the `Kallisto` aligner, the `tximport` R package was used to perform gene-level summarization. Following this, after performing elementary filtering on the gene-level quantification results, the method of `limma-voom` was employed to analyze the (pseudo)counts with two different modeling procedures: (1) the first using a simple design matrix containing an intercept term and a term for fetal vs. adult samples; and (2) the second using a more complex design matrix containing an intercept term, a term for the sample type (fetal vs. adult), a term for the sex of the sample, a term for the sample age, a term for the race of the sample, and a term for the quality of the sample (as measured by RIN).

Below, we present the results of applying the `limma-voom` modeling procedure, including tables of the top 10 genes showing differential expression (based on the Benjamini-Hochberg FDR), as well as several visualizations produced from the results of the statistical analysis described above.

II. Methodology and Results

The linear modeling procedure of `limma-voom` was invoked (see code below), and the tables of the **top 10 genes** for each modeling paradigm are given below:

```
# fit linear models to each gene using voom with simple design matrix
vfit_simple <- limma::lmFit(v_simple)
vfit_simple <- limma::eBayes(vfit_simple)
tt1 <- limma::topTable(vfit_simple,
                       coef = which(colnames(design_simple) == "type"),
                       adjust.method = "BH", number = Inf,
                       sort.by = "none", confint = TRUE)
```

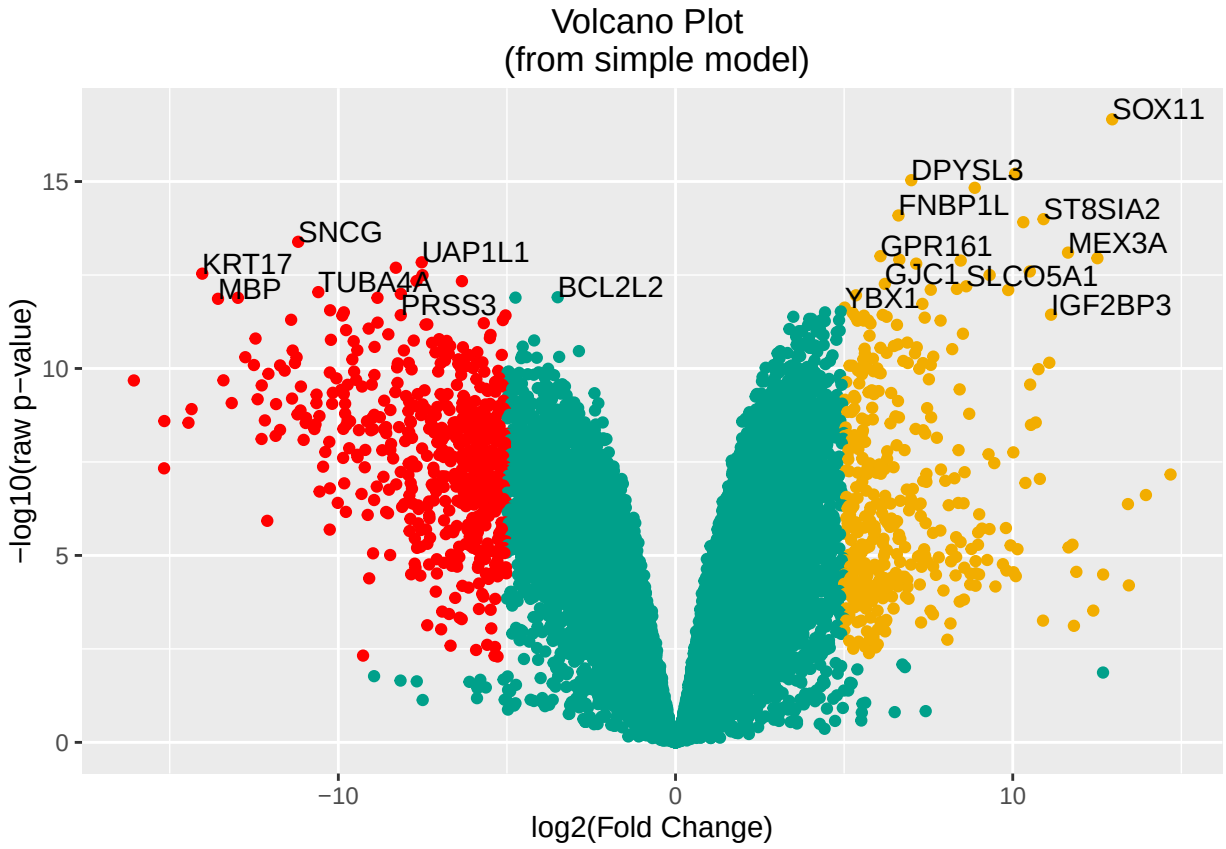
geneID	lowerCI	FoldChange	upperCI	pvalue	fdrBH
SOX11	5449	7914	11496	2.17e-17	3.893e-13
DPYSL3	97.77	127.3	165.7	9.21e-16	5.507e-12
SLA	743.1	1075	1555	6.174e-16	5.507e-12
FBN3	331.7	469	663.1	1.473e-15	6.604e-12
FNBP1L	73.05	97.79	130.9	8.098e-15	2.906e-11
ST8SIA2	1182	1928	3145	1.013e-14	3.03e-11
DCX	795.8	1272	2032	1.224e-14	3.137e-11
SNCG	0.0002457	0.0004281	0.0007457	4.091e-14	9.174e-11
MEX3A	1740	3189	5846	7.971e-14	1.589e-10
GPR161	48.96	67.52	93.11	9.897e-14	1.776e-10

```
# fit linear models to each gene using voom with the full design matrix
vfit_full <- limma::lmFit(v_full)
vfit_full <- limma::eBayes(vfit_full)
tt2 <- limma::topTable(vfit_full,
                       coef = which(colnames(design_full) == "type"),
                       adjust.method = "BH", number = Inf,
                       sort.by = "none", confint = TRUE)
```

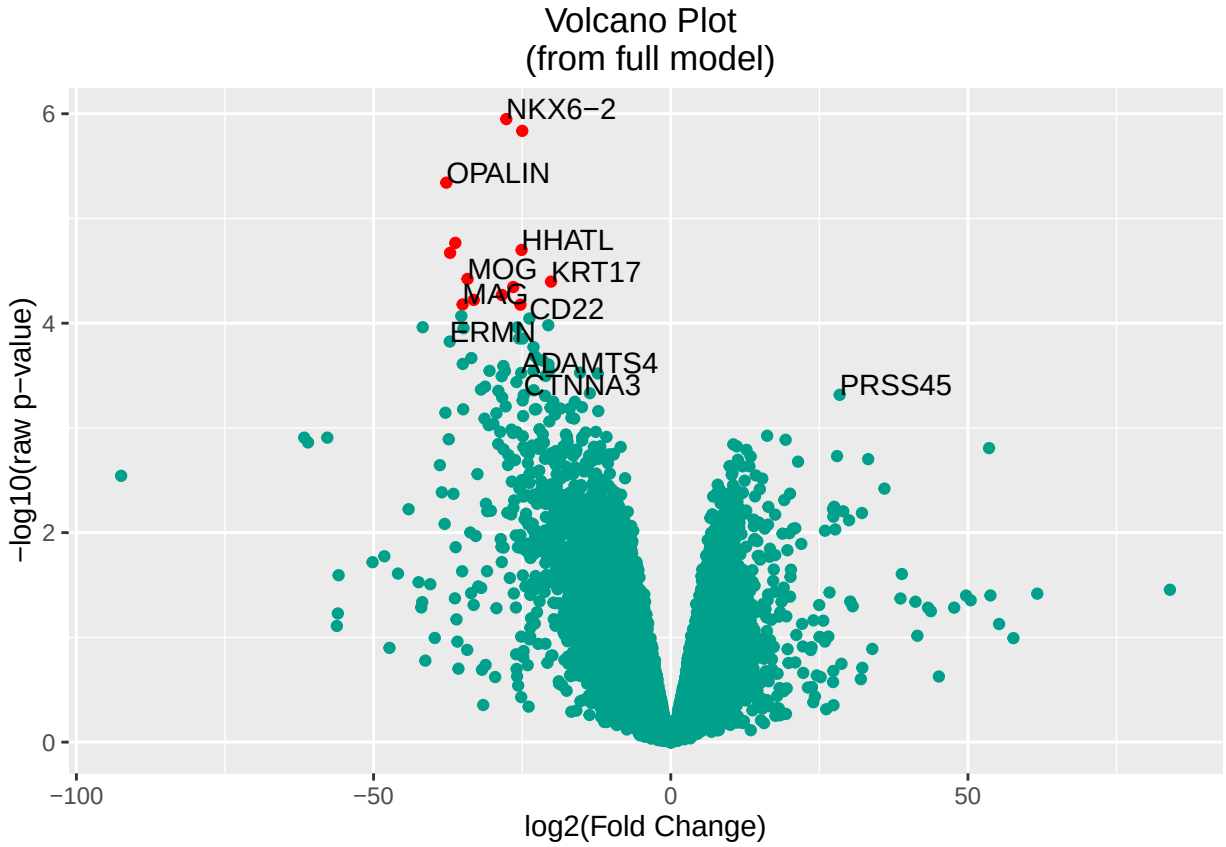
geneID	lowerCI	FoldChange	upperCI	pvalue	fdrBH
NKX6-2	7.436e-11	4.675e-09	2.939e-07	1.124e-06	0.01307
TMEM235	6.422e-10	3.004e-08	1.405e-06	1.457e-06	0.01307
OPALIN	5.932e-15	4.287e-12	3.098e-09	4.55e-06	0.02721
HHATL	1.58e-10	2.782e-08	4.9e-06	1.998e-05	0.06366
HOXD1	7.971e-15	1.224e-11	1.88e-08	1.712e-05	0.06366
KLK6	3.004e-15	6.668e-12	1.48e-08	2.129e-05	0.06366
KRT17	9.585e-09	8.638e-07	7.784e-05	4.007e-05	0.08986
MOG	2.58e-14	5.092e-11	1.005e-07	3.778e-05	0.08986
HSD11B1	2.593e-11	1.052e-08	4.271e-06	4.513e-05	0.08997
FA2H	3.899e-12	2.796e-09	2.005e-06	5.378e-05	0.09177

III. Data Visualization

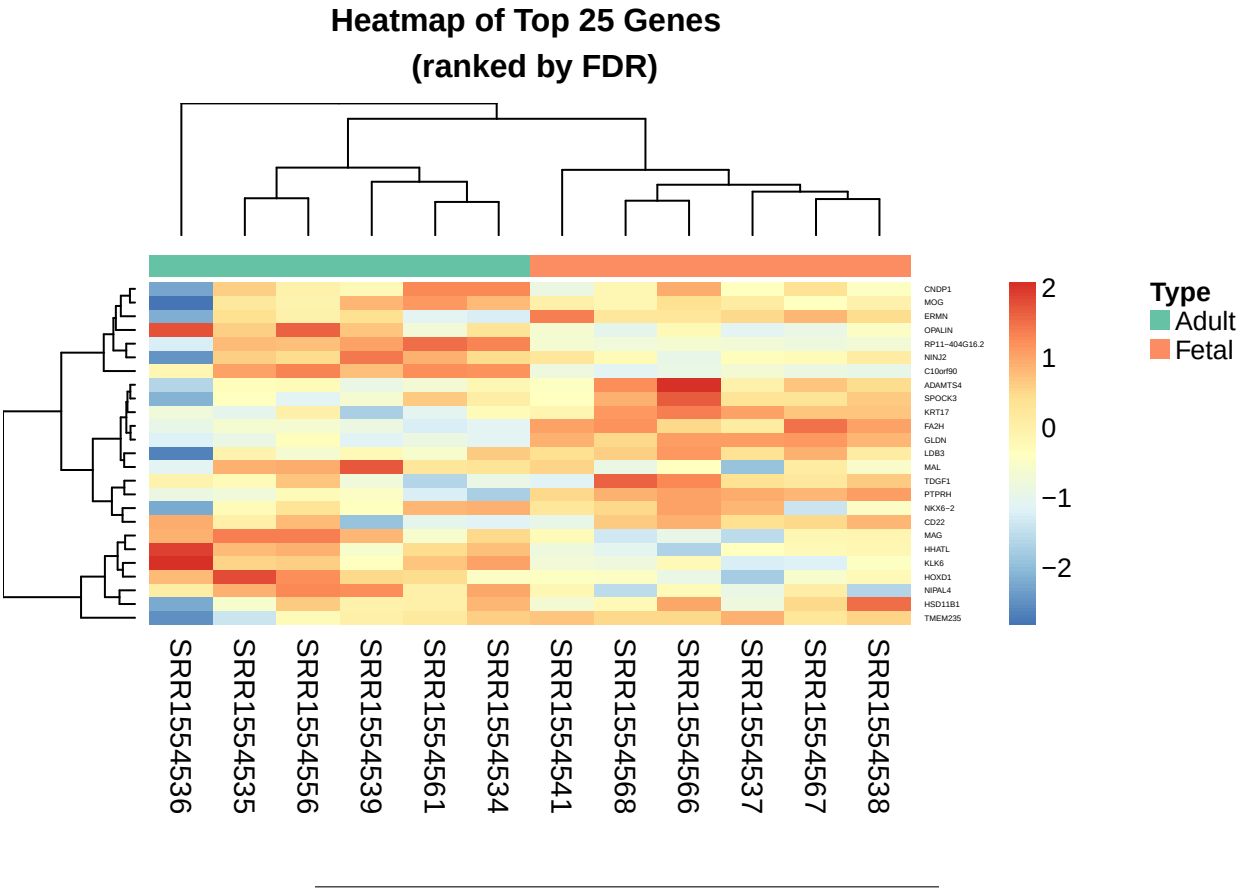
The volcano plot below displays the differential expression (as measured by fold change) against the transformed p-values, for the simple linear model containing terms for an intercept and the sample type (adult vs. fetal). Those genes for which the log10 fold change is above **5** and the Benjamini-Hochberg adjusted p-value is below **0.01** are denoted by the warmer colors (*i.e.*, red, yellow) and the gene symbols for the *top 50 genes* (as ranked by the Benjamini-Hochberg FDR) are displayed on the plot.



The volcano plot below displays the differential expression (as measured by fold change) against the transformed p-values, for the full linear model containing terms for an intercept, the sample type (adult vs. fetal), the sex of the sample, the sample age, the race of the sample, and the quality of the sample (as measured by RIN). Those genes for which the log10 fold change is above **5** and the Benjamini-Hochberg adjusted p-value is below **0.10** are denoted by the warmer colors (*i.e.*, red, yellow) and the gene symbols for the *top 50 genes* (as ranked by the Benjamini-Hochberg FDR) are displayed on the plot.



The heatmap below displays the normalized expression results of the top 25 genes with the highest differential expression across the groupings of the samples, with hierarchical clustering performed across both the 12 samples and the top 25 genes. From the plot, it appears that those genes showing heightened differential expression in the fetal group display lowered expression in the adult samples.



IV. Reproducibility Notice

What follows is the *session information* associated with the R session in which this report was compiled:

R version 3.3.1 (2016-06-21)

Platform: x86_64-apple-darwin16.0.0 (64-bit)

locale: en_US.UTF-8||en_US.UTF-8||en_US.UTF-8||C||en_US.UTF-8||en_US.UTF-8

attached base packages: *grid*, *stats4*, *parallel*, *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *RColorBrewer(v.1.1-2)*, *NMF(v.0.20.6)*, *synchronicity(v.1.1.9.1)*, *bigmemory(v.4.5.19)*, *bigmemory.sri(v.0.1.3)*, *cluster(v.2.0.5)*, *rngtools(v.1.2.4)*, *pkgmaker(v.0.22)*, *registry(v.0.3)*, *wesanderson(v.0.3.2)*, *ggbiplot(v.0.55)*, *scales(v.0.4.0)*, *plyr(v.1.8.4)*, *reshape2(v.1.4.2)*, *limma(v.3.30.0)*, *EnsDb.Hsapiens.v79(v.1.1.0)*, *ensembldb(v.1.6.0)*, *GenomicFeatures(v.1.26.0)*, *AnnotationDbi(v.1.36.0)*, *Biobase(v.2.34.0)*, *GenomicRanges(v.1.26.1)*, *GenomeInfoDb(v.1.10.0)*, *IRanges(v.2.8.0)*, *S4Vectors(v.0.12.0)*, *BiocGenerics(v.0.20.0)*, *tximport(v.1.2.0)*, *readr(v.1.0.0)*, *dtplyr(v.0.0.1)*, *data.table(v.1.9.7)*, *nima(v.0.3.5)*, *tibble(v.1.2)*, *devtools(v.1.12.0)*, *ggplot2(v.2.1.0)*, *dplyr(v.0.5.0)* and *colorout(v.1.1-2)*

loaded via a namespace (and not attached): *httr(v.1.2.1)*, *foreach(v.1.4.3)*, *AnnotationHub(v.2.6.0)*, *splines(v.3.3.1)*, *shiny(v.0.14.1)*, *assertthat(v.0.1)*, *interactiveDisplayBase(v.1.12.0)*, *pander(v.0.6.0)*, *Rsamtools(v.1.26.1)*, *yaml(v.2.1.13)*, *RSQLite(v.1.0.0)*, *lattice(v.0.20-34)*, *digest(v.0.6.10)*, *XVector(v.0.14.0)*, *colorspace(v.1.2-7)*, *htmltools(v.0.3.5)*, *httpuv(v.1.3.3)*, *Matrix(v.1.2-7.1)*, *XML(v.3.98-1.4)*, *biomaRt(v.2.30.0)*, *zlibbioc(v.1.20.0)*, *xtable(v.1.8-2)*, *BiocParallel(v.1.8.0)*, *withr(v.1.0.2)*, *SummarizedExperiment(v.1.4.0)*, *lazyeval(v.0.2.0)*, *survival(v.2.39-5)*, *magrittr(v.1.5)*, *mime(v.0.5)*, *memoise(v.1.0.0)*, *evaluate(v.0.10)*, *doParallel(v.1.0.10)*, *ggthemes(v.3.2.0)*, *BiocInstaller(v.1.24.0)*, *tools(v.3.3.1)*, *gridBase(v.0.4-7)*, *formatR(v.1.4)*, *stringr(v.1.1.0)*, *munsell(v.0.4.3)*, *Biostrings(v.2.42.0)*, *RCurl(v.1.95-4.8)*, *iterators(v.1.0.8)*, *labeling(v.0.3)*, *bitops(v.1.0-6)*, *rmarkdown(v.1.1)*, *gtable(v.0.2.0)*, *codetools(v.0.2-15)*, *DBI(v.0.5-1)*, *R6(v.2.2.0)*, *GenomicAlignments(v.1.10.0)*, *gridExtra(v.2.2.1)*, *knitr(v.1.14)*, *rtracklayer(v.1.34.0)*, *stringi(v.1.1.2)* and *Rcpp(v.0.12.7)*

V. References

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): 1–17.

Robles, José A, Sumaira E Qureshi, Stuart J Stephen, Susan R Wilson, Conrad J Burden, and Jennifer M Taylor. 2012. "Efficient Experimental Design and Analysis Strategies for the Detection of Differential Expression Using RNA-Sequencing." *BMC Genomics* 13 (1). BioMed Central: 1.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1). bepress: Article–3.

Soneson, Charlotte, and Mauro Delorenzi. 2013. "A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data." *BMC Bioinformatics* 14 (1). BioMed Central: 1.