

# Comparison of Recount with GTEx

*Andrew E. Jaffe and Kasper D. Hansen*

*July 18, 2016*

## Contents

<b>1</b>	<b>Overview</b>	<b>1</b>
<b>2</b>	<b>Dependencies</b>	<b>1</b>
2.1	R packages	1
2.2	Data objects	2
<b>3</b>	<b>Mapping GTEx annotation</b>	<b>3</b>
<b>4</b>	<b>Comparison</b>	<b>3</b>
<b>5</b>	<b>Differential expression</b>	<b>7</b>
<b>6</b>	<b>Reproducibility</b>	<b>9</b>

## 1 Overview

This document compares GTEx data release v6 to Recount. The main issue addressed in this document is mapping up genes and samples between the two datasets. The annotations are different:

- GTEx uses Gencode v19 mapped to hg19.
- Recount uses Gencode v25 mapped to hg38, specifically this GFF3 file.

## 2 Dependencies

### 2.1 R packages

```
library('ballgown')
library('coop')
library('org.Hs.eg.db')
library('readr')
library('recount')
library('rtracklayer')
library('stringr')
library('SummarizedExperiment')
library('limma')
library('edgeR')
```

## 2.2 Data objects

### 2.2.1 From Recount

### 2.2.2 From GTEx website

We have downloaded the annotation GTF files as well as the raw gene count matrix from the GTEx portal.

```
if(all(file.exists('gencode.v19.genes.patched_contigs.gtf', 'GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_reads.gct.gz')) {
  dataPath <- '.'
} else {
  dataPath <- "/dcs01/ajaffe/GTEX/V6" # wherever data was downloaded
}

gtexGtf <- import(file.path(dataPath, "gencode.v19.genes.patched_contigs.gtf"))
gtexData <- read_tsv(file.path(dataPath,
  "GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_reads.gct.gz"),
  skip = 2, progress = FALSE)

## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Name = col_character(),
##   Description = col_character()
## )

## See spec(...) for full column specifications.

## Warning: 5 parsing failures.
##   row                  col          expected actual
## 3564 GTEx-UJMC-1926-SM-3GADS no trailing characters e+05
## 28086 GTEx-XYKS-2726-SM-4E3IC no trailing characters e+05
## 28519 GTEx-133LE-1926-SM-5N9FV no trailing characters e+05
## 33344 GTEx-ZZPU-0326-SM-5N9BJ no trailing characters e+05
## 39997 GTEx-132NY-2726-SM-5PNY2 no trailing characters e+05

gtexCounts <- as.data.frame(gtexData[, 3:ncol(gtexData)])
rownames(gtexCounts) <- gtexData$Name
rm(gtexData)
```

### 2.2.3 From elsewhere

These are the Rail-RNA processed samples

```
if(!file.exists(file.path('SRP012682', 'rse_gene.Rdata'))) {
  download_study('SRP012682')
}
load('SRP012682/rse_gene.Rdata')
gtexPd <- colData(rse_gene)
```

Let's match everything up.

```
mm <- match(colnames(gtexCounts), gtexPd$sampid)
gtexCounts <- gtexCounts[, !is.na(mm)]
```

```
gtexPd <- gtexPd[mm[!is.na(mm)], ]
rse_gene <- rse_gene[, mm[!is.na(mm)]]
```

### 3 Mapping GTEx annotation

We map between version by using Ensembl gene IDs.

```
## filter counts
geneMatch <- match(ballgown:::ss(rownames(gtexCounts), "\\\\"), 
                     ballgown:::ss(rowData(rse_gene)$gene_id, "\\\\")) 
gtexCounts <- gtexCounts[!is.na(geneMatch),]
rse_gene <- rse_gene[geneMatch[!is.na(geneMatch)],]
## filter map
gtexMap <- gtexGtf[!duplicated(gtexGtf$gene_id)]
names(gtexMap) <- gtexMap$gene_id
gtexMap <- gtexMap[rownames(gtexCounts)]
gtexMap$EnsemblGeneID = ballgown:::ss(names(gtexMap), "\\\\")

## Number of genes:
nrow(gtexCounts)
```

```
## [1] 51491
```

Let's load data from Recount.

```
rse_gene <- scale_counts(rse_gene)
recountCounts <- assays(rse_gene)$counts
recountMap <- rowRanges(rse_gene)
stopifnot(all(colnames(recountCounts) == rownames(gtexPd)))
```

### 4 Comparison

```
gtexCounts <- as.matrix(gtexCounts)
ind <- which(colSums(is.na(gtexCounts)) == 0)
gtexCounts2 <- log2(sweep(gtexCounts, MARGIN = 2, FUN = "/",
                           colSums(gtexCounts) / (4 * 10^7)) + 1)[, ind]
recountCounts2 <- log2(recountCounts[, ind]+1)
gtexPd2 <- gtexPd[ind, ]

normCors <- sapply(seq_len(nrow(gtexCounts2)),
                   function(ii) pcor(gtexCounts2[ii, ], recountCounts2[ii,]))
summary(normCors)

##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.    NA's
## -0.3011   0.7233   0.9420   0.8105   0.9858   1.0000    688

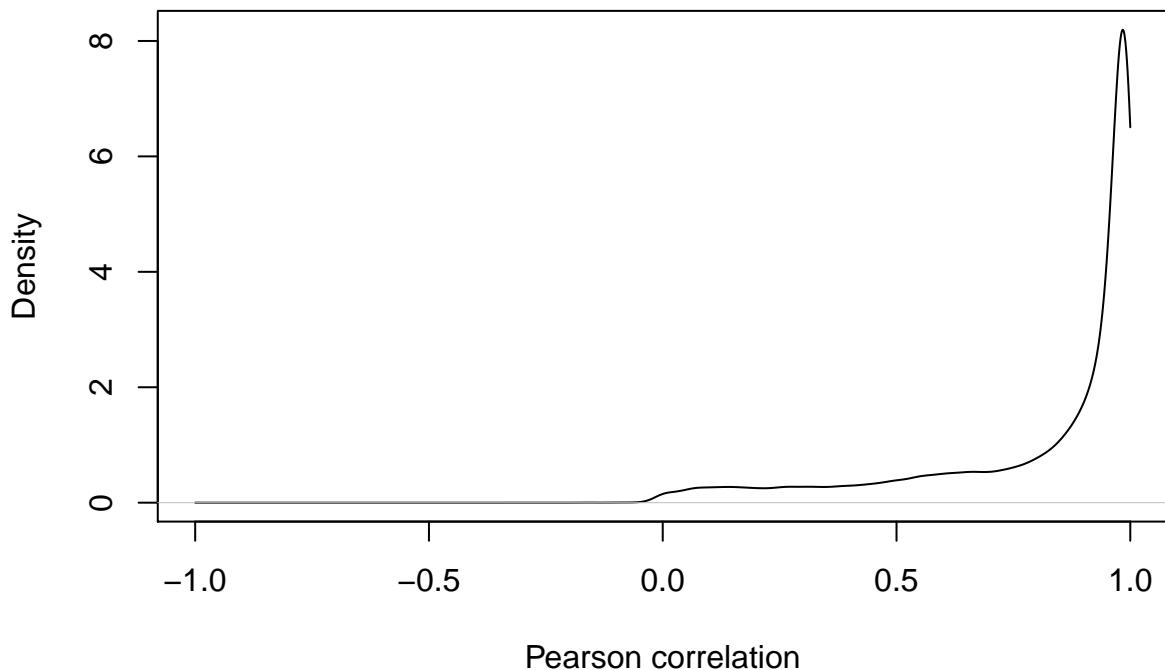
sum(normCors <= 0.95, na.rm = TRUE)

## [1] 26591
sum(normCors <= 0.80, na.rm = TRUE)

## [1] 15179
```

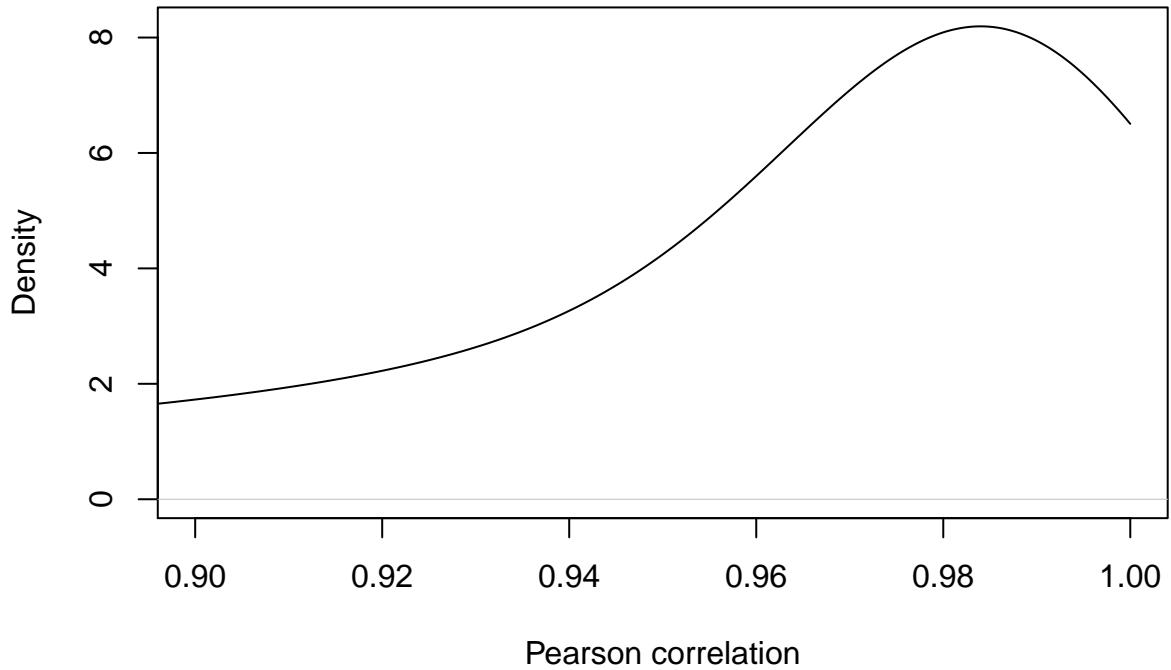
```
mean(normCors >= 0.99, na.rm = TRUE)  
## [1] 0.173848  
dens <- density(normCors, from = -1, to = 1, na.rm = TRUE, n = 4096)  
plot(dens, xlab = "Pearson correlation",  
     main = "Size-scaled counts")
```

Size-scaled counts



```
plot(dens, xlab = "Pearson correlation",  
     main = "Size-scaled counts", xlim = c(0.9,1))
```

## Size-scaled counts



```
ind = which(gtexMap$gene_type == "protein_coding")
## Number of protein coding genes:
length(ind)

## [1] 18998

normCors_coding <- sapply(seq_len(nrow(gtexCounts2[ind,])),
  function(ii) pcor(gtexCounts2[ind[ii], ], recountCounts2[ind[ii], ]))
summary(normCors_coding)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.    NA's
## -0.2345  0.9710  0.9874  0.9510  0.9932  0.9985       8

sum(normCors_coding <= 0.95, na.rm = TRUE)

## [1] 3246

sum(normCors_coding <= 0.80, na.rm = TRUE)

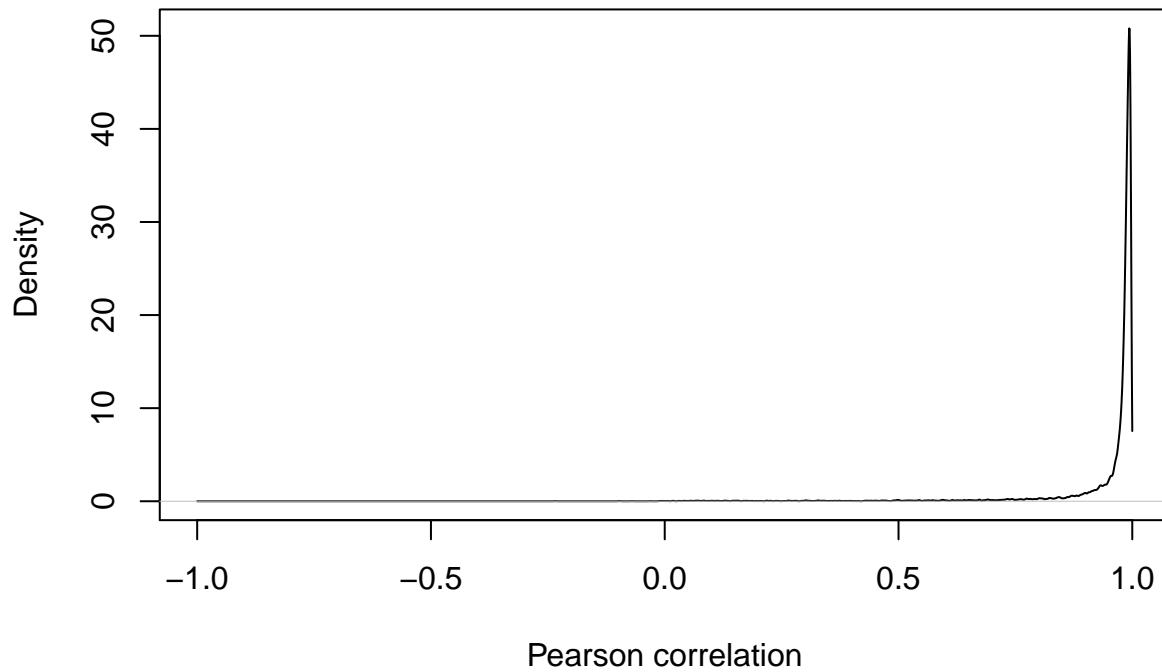
## [1] 1098

mean(normCors_coding >= 0.99, na.rm = TRUE)

## [1] 0.3988415

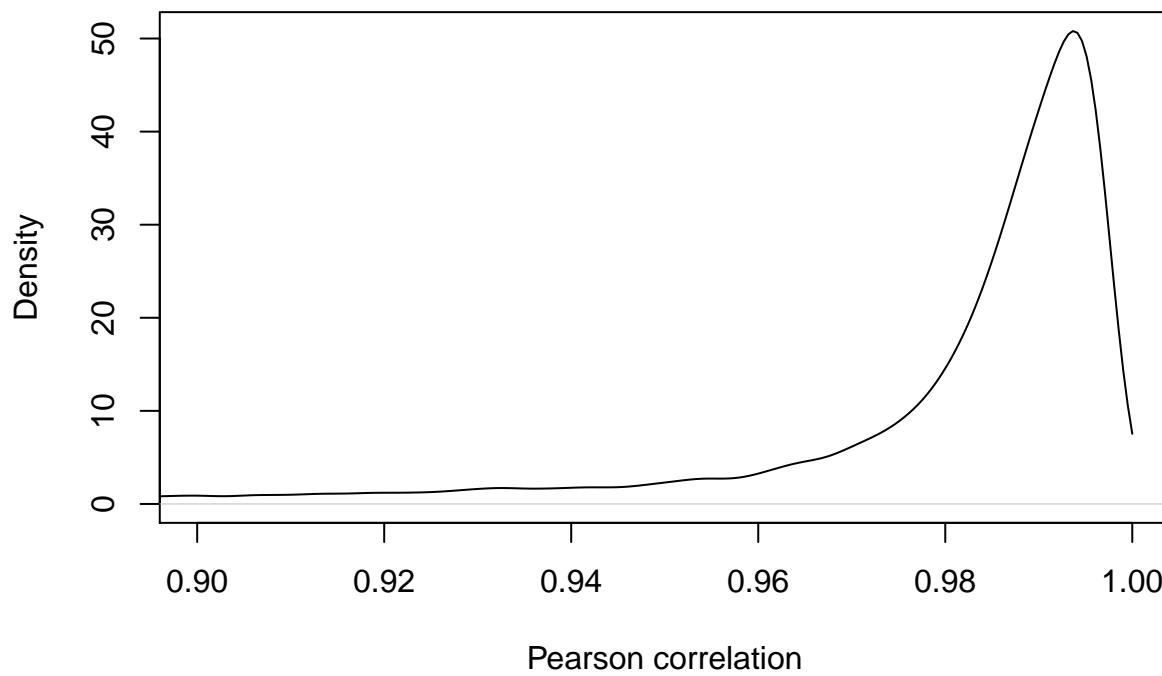
dens <- density(normCors_coding, from = -1, to = 1, na.rm = TRUE, n = 4096)
plot(dens, xlab = "Pearson correlation",
  main = "Size-scaled counts (Protein Coding)")
```

### Size-scaled counts (Protein Coding)



```
plot(dens, xlab = "Pearson correlation",
      main = "Size-scaled counts (Protein Coding)", xlim = c(0.9,1))
```

### Size-scaled counts (Protein Coding)



## 5 Differential expression

Between colon and blood

```
indTissue <- c(which(gtexPd2$smts == "Colon"),
  which(gtexPd2$smtsd == "Whole Blood"))
gtexPd2_sub <- gtexPd2[indTissue, ]
recountCounts2_sub <- recountCounts2[, indTissue]
gtexCounts2_sub <- gtexCounts2[, indTissue]
design <- model.matrix(~ smts , data = gtexPd2_sub)
```

Using recount:

```
dge_recount <- DGEList(counts = recountCounts2_sub)
dge_recount <- calcNormFactors(dge_recount)
v_recount <- voom(dge_recount, design, plot=FALSE)
fit_recount <- lmFit(v_recount, design)
eb_recount <- ebayes(fit_recount)
out_recount <- data.frame(log2FC = fit_recount$coef[, 2],
  tstat = eb_recount$t[, 2], pvalue = eb_recount$p[, 2])
colnames(out_recount) <- paste0(colnames(out_recount), "_recount")
```

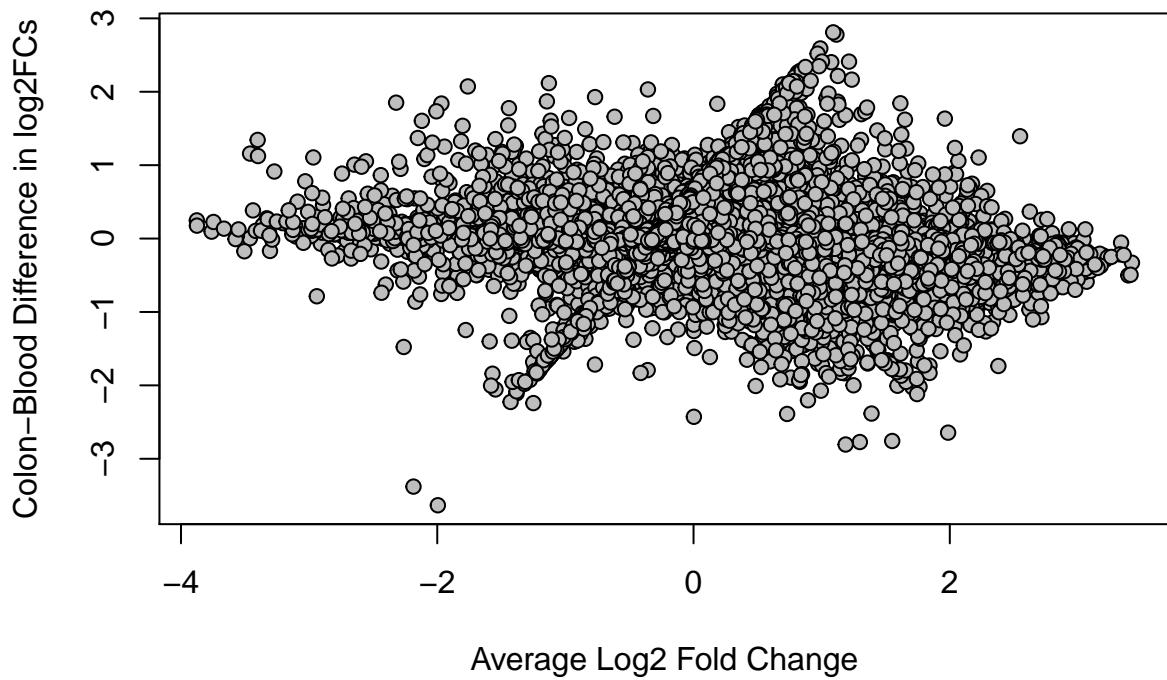
And using original counts:

```
dge_gtex <- DGEList(counts = gtexCounts2_sub)
dge_gtex <- calcNormFactors(dge_gtex)
v_gtex <- voom(dge_gtex, design, plot=FALSE)
fit_gtex <- lmFit(v_gtex, design)
eb_gtex <- ebayes(fit_gtex)
out_gtex <- data.frame(log2FC = fit_gtex$coef[, 2],
  tstat = eb_gtex$t[, 2], pvalue = eb_gtex$p[, 2])
colnames(out_gtex) <- paste0(colnames(out_gtex), "_gtex")
```

Compare:

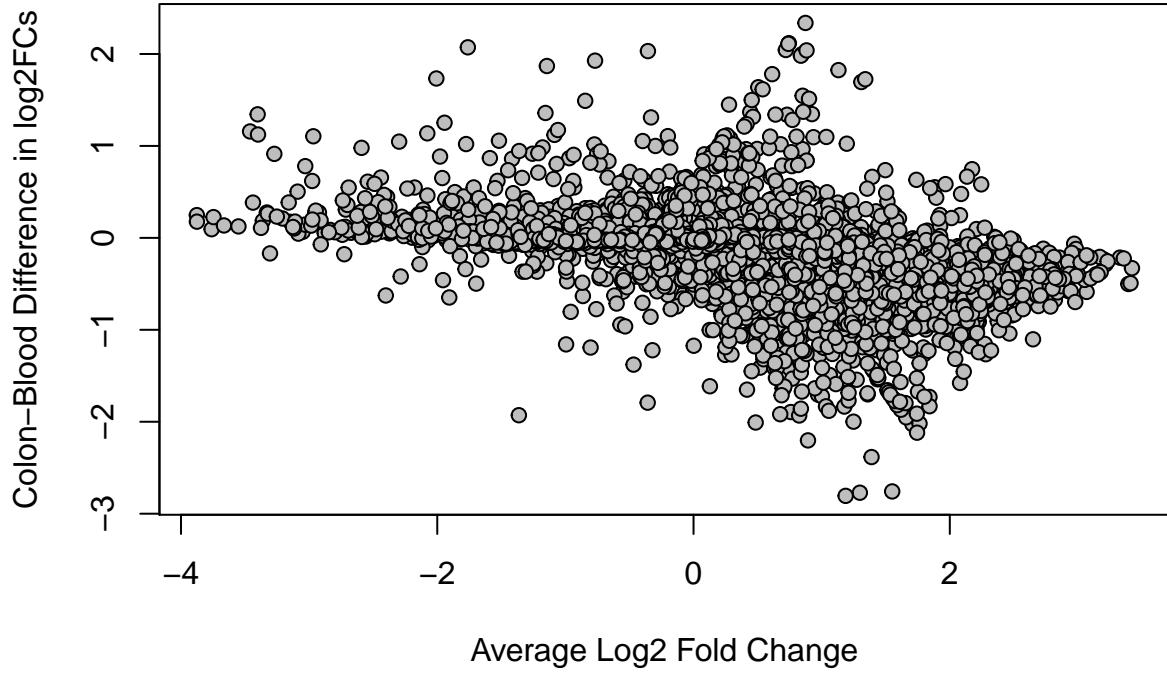
```
M <- out_recount$log2FC_recount - out_gtex$log2FC_gtex
A <- ( out_recount$log2FC_recount + out_gtex$log2FC_gtex)/2
plot(M ~ A, xlab="Average Log2 Fold Change",
  ylab="Colon-Blood Difference in log2FCs",
  pch = 21, bg="grey", main = "All Genes")
```

### All Genes



```
plot(M ~ A, xlab="Average Log2 Fold Change", subset=ind,
     ylab="Colon-Blood Difference in log2FCs",
     pch = 21, bg="grey", main = "Protein Coding Genes")
```

### Protein Coding Genes



```

## Genes with M changes greater than 2


|       |              |
|-------|--------------|
| table | (abs(M) > 2) |
|-------|--------------|



##
## FALSE TRUE
## 51428    63


|       |        |                                    |
|-------|--------|------------------------------------|
| round | (table | (abs(M) > 2) / length(M) * 100, 3) |
|-------|--------|------------------------------------|



##
## FALSE TRUE
## 99.878  0.122


|       |                   |
|-------|-------------------|
| table | (abs(M[ind]) > 2) |
|-------|-------------------|



##
## FALSE TRUE
## 18982    16


|       |        |                                           |
|-------|--------|-------------------------------------------|
| round | (table | (abs(M[ind]) > 2) / length(ind) * 100, 3) |
|-------|--------|-------------------------------------------|



##
## FALSE TRUE
## 99.916  0.084

```

The R-squared is 0.8395661 for all 51491 genes and 0.916911 for all 18998 protein coding genes.

## 6 Reproducibility

This analysis report was made possible thanks to:

- R (R Core Team, 2016)
- *ballgown* (Fu, Frazee, Collado-Torres, Jaffe, et al., 2016)
- *BiocStyle* (Oleś, Morgan, and Huber, 2016)
- *coop* (Schmidt, 2016)
- *devtools* (Wickham and Chang, 2016)
- *edgeR* (McCarthy, J., Chen, Yunshun, et al., 2012)
- *knitcitations* (Boettiger, 2015)
- *org.Hs.eg.db* (Carlson, 2016)
- *readr* (Wickham, Hester, and Francois, 2016)
- *recount* (Collado-Torres, Nellore, Kammers, Ellis, et al., 2016)
- *rmarkdown* (Allaire, Cheng, Xie, McPherson, et al., 2016)
- *rtracklayer* (Lawrence, Gentleman, and Carey, 2009)
- *stringr* (Wickham, 2016)
- *SummarizedExperiment* (Morgan, Obenchain, Hester, and Pagès, 2016)

Bibliography file

- [1] J. Allaire, J. Cheng, Y. Xie, J. McPherson, et al. rmarkdown: Dynamic Documents for R. R package version 1.3. 2016. URL: <https://CRAN.R-project.org/package=rmarkdown>.
- [2] C. Boettiger. knitcitations: Citations for ‘Knitr’ Markdown Files. R package version 1.0.7. 2015. URL: <https://CRAN.R-project.org/package=knitcitations>.
- [3] M. Carlson. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.4.0. 2016.
- [4] L. Collado-Torres, A. Nellore, K. Kammers, S. E. Ellis, et al. “recount: A large-scale resource of analysis-ready RNA-seq expression data”. In: bioRxiv (2016). DOI: 10.1101/068478. URL:

<http://biorxiv.org/content/early/2016/08/08/068478>.

- [5] J. Fu, A. C. Frazee, L. Collado-Torres, A. E. Jaffe, et al. ballgown: Flexible, isoform-level differential expression analysis. R package version 2.6.0. 2016.
- [6] M. Lawrence, R. Gentleman and V. Carey. “rtracklayer: an R package for interfacing with genome browsers”. In: Bioinformatics 25 (2009), pp. 1841-1842. DOI: 10.1093/bioinformatics/btp328. URL: <http://bioinformatics.oxfordjournals.org/content/25/14/1841.abstract>.
- [7] McCarthy, D. J., Chen, Yunshun, et al. “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. In: Nucleic Acids Research 40.10 (2012), pp. -9.
- [8] M. Morgan, V. Obenchain, J. Hester and H. Pagès. SummarizedExperiment: SummarizedExperiment container. R package version 1.4.0. 2016.
- [9] A. Oleś, M. Morgan and W. Huber. BiocStyle: Standard styles for vignettes and other Bioconductor documents. R package version 2.2.1. 2016. URL: <https://github.com/Bioconductor/BiocStyle>.
- [10] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2016. URL: <https://www.R-project.org/>.
- [11] D. Schmidt. Co-Operation: Fast Correlation, Covariance, and Cosine Similarity. R package version 0.6-0. 2016. URL: <https://cran.r-project.org/package=coop>.
- [12] H. Wickham. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. 2016. URL: <https://CRAN.R-project.org/package=stringr>.
- [13] H. Wickham and W. Chang. devtools: Tools to Make Developing R Packages Easier. R package version 1.12.0. 2016. URL: <https://CRAN.R-project.org/package=devtools>.
- [14] H. Wickham, J. Hester and R. Francois. readr: Read Tabular Data. R package version 1.0.0. 2016. URL: <https://CRAN.R-project.org/package=readr>.

```
## Time spent creating this report:  
diff(c(timestart, Sys.time()))  
  
## Time difference of 14.07494 mins  
## Date this report was generated  
message(Sys.time())  
  
## 2017-02-07 15:23:50  
## Reproducibility info  
options(width = 120)  
devtools::session_info()  
  
## Session info -----  
  
## setting value  
## version R version 3.3.2 RC (2016-10-26 r71594)  
## system x86_64, darwin13.4.0  
## ui X11  
## language (EN)  
## collate en_US.UTF-8  
## tz America/New_York  
## date 2017-02-07  
  
## Packages -----  
## package * version date source  
## acepack 1.4.1 2016-10-29 CRAN (R 3.3.0)  
## annotate 1.52.1 2016-12-22 Bioconductor
```

```

## AnnotationDbi          * 1.36.2   2017-01-30 Bioconductor
## assertthat              0.1       2013-12-06 cran (@0.1)
## backports                1.0.5     2017-01-18 CRAN (R 3.3.2)
## ballgown                 * 2.6.0     2016-10-18 Bioconductor
## base64enc                  0.1-3     2015-07-28 cran (@0.1-3)
## bibtex                     0.4.0     2014-12-31 CRAN (R 3.3.0)
## Biobase                   * 2.34.0     2016-10-18 Bioconductor
## BiocGenerics               * 0.20.0     2016-10-18 Bioconductor
## BiocParallel                1.8.1     2016-10-30 Bioconductor
## BiocStyle                  * 2.2.1     2016-11-24 Bioconductor
## biomaRt                    2.30.0     2016-10-18 Bioconductor
## Biostrings                  2.42.1     2016-12-01 Bioconductor
## bitops                      1.0-6      2013-08-17 cran (@1.0-6)
## BSgenome                    1.42.0     2016-10-18 Bioconductor
## bumphunter                  1.14.0     2016-10-18 Bioconductor
## checkmate                   1.8.2      2016-11-02 CRAN (R 3.3.1)
## cluster                      2.0.5     2016-10-08 CRAN (R 3.3.2)
## codetools                   0.2-15     2016-10-05 CRAN (R 3.3.2)
## colorout                     * 1.1-2     2016-10-19 Github (jalvesaq/colorout@6d84420)
## colorspace                   1.3-2      2016-12-14 CRAN (R 3.3.2)
## coop                        * 0.6-0     2016-12-13 CRAN (R 3.3.2)
## data.table                  1.10.4     2017-02-01 CRAN (R 3.3.2)
## DBI                          0.5-1      2016-09-10 cran (@0.5-1)
## derfinder                   1.8.0      2016-10-18 Bioconductor
## derfinderHelper              1.8.0      2016-10-18 Bioconductor
## devtools                     1.12.0     2016-06-24 CRAN (R 3.3.0)
## digest                       0.6.12     2017-01-27 CRAN (R 3.3.2)
## doRNG                        1.6        2014-03-07 CRAN (R 3.3.0)
## downloader                  0.4        2015-07-09 CRAN (R 3.3.0)
## edgeR                        * 3.16.5     2016-12-15 Bioconductor
## evaluate                     0.10       2016-10-11 cran (@0.10)
## foreach                      1.4.3      2015-10-13 CRAN (R 3.3.0)
## foreign                      0.8-67     2016-09-13 CRAN (R 3.3.2)
## Formula                      1.2-1      2015-04-07 CRAN (R 3.3.0)
## genefilter                   1.56.0     2016-10-18 Bioconductor
## GenomeInfoDb                 * 1.10.2     2016-12-29 Bioconductor
## GenomicAlignments             1.10.0     2016-10-18 Bioconductor
## GenomicFeatures                1.26.2     2016-12-17 Bioconductor
## GenomicFiles                  1.10.3     2016-10-21 Bioconductor
## GenomicRanges                 * 1.26.2     2017-01-02 Bioconductor
## GEOquery                     2.40.0     2016-10-18 Bioconductor
## ggplot2                      2.2.1      2016-12-30 CRAN (R 3.3.2)
## gridExtra                     2.2.1      2016-02-29 CRAN (R 3.3.0)
## gtable                       0.2.0      2016-02-26 CRAN (R 3.3.0)
## Hmisc                         4.0-2      2016-12-31 CRAN (R 3.3.2)
## htmlTable                     1.9        2017-01-26 CRAN (R 3.3.2)
## htmltools                     0.3.5      2016-03-21 cran (@0.3.5)
## htmlwidgets                   0.8        2016-11-09 CRAN (R 3.3.2)
## httr                           1.2.1      2016-07-03 CRAN (R 3.3.0)
## IRanges                        * 2.8.1     2016-11-08 Bioconductor
## iterators                     1.0.8      2015-10-13 CRAN (R 3.3.0)
## jsonlite                      1.2        2016-12-31 CRAN (R 3.3.2)
## knitrCitations                * 1.0.7      2015-10-28 CRAN (R 3.3.0)
## knitr                         1.15.1     2016-11-22 CRAN (R 3.3.2)

```

```

## lattice          0.20-34  2016-09-06 CRAN (R 3.3.2)
## latticeExtra    0.6-28   2016-02-09 CRAN (R 3.3.0)
## lazyeval         0.2.0    2016-06-12 cran (@0.2.0)
## limma           * 3.30.10 2017-02-03 Bioconductor
## locfit          1.5-9.1  2013-04-20 CRAN (R 3.3.0)
## lubridate        1.6.0    2016-09-13 CRAN (R 3.3.0)
## magrittr         1.5      2014-11-22 cran (@1.5)
## Matrix           1.2-8   2017-01-20 CRAN (R 3.3.2)
## matrixStats      0.51.0   2016-10-09 CRAN (R 3.3.0)
## memoise          1.0.0    2016-01-29 CRAN (R 3.3.0)
## mgcv             1.8-16   2016-11-07 CRAN (R 3.3.0)
## munsell          0.4.3    2016-02-13 cran (@0.4.3)
## nlme             3.1-131  2017-02-06 CRAN (R 3.3.2)
## nnet              7.3-12   2016-02-02 CRAN (R 3.3.2)
## org.Hs.eg.db     * 3.4.0   2016-10-19 Bioconductor
## pkgmaker         0.22     2014-05-14 CRAN (R 3.3.0)
## plyr              1.8.4    2016-06-08 cran (@1.8.4)
## qvalue            2.6.0    2016-10-18 Bioconductor
## R6                2.2.0    2016-10-05 CRAN (R 3.3.0)
## RColorBrewer     1.1-2    2014-12-07 cran (@1.1-2)
## Rcpp              0.12.9   2017-01-14 CRAN (R 3.3.2)
## RCurl             1.95-4.8 2016-03-01 cran (@1.95-4.)
## readr             * 1.0.0   2016-08-03 CRAN (R 3.3.0)
## recount            * 1.0.9   2017-02-07 Bioconductor
## RefManageR        0.13.1   2016-11-13 CRAN (R 3.3.2)
## registry          0.3      2015-07-08 CRAN (R 3.3.0)
## rentrez            1.0.4    2016-10-26 CRAN (R 3.3.0)
## reshape2           1.4.2    2016-10-22 CRAN (R 3.3.0)
## RJSONIO            1.3-0    2014-07-28 cran (@1.3-0)
## rmarkdown           * 1.3     2016-12-21 CRAN (R 3.3.2)
## rngtools           1.2.4    2014-03-06 CRAN (R 3.3.0)
## rpart              4.1-10   2015-06-29 CRAN (R 3.3.2)
## rprojroot          1.2      2017-01-16 CRAN (R 3.3.2)
## Rsamtools          1.26.1   2016-10-22 Bioconductor
## RSQLite             1.1-2    2017-01-08 CRAN (R 3.3.2)
## rtracklayer        * 1.34.1  2016-11-02 Bioconductor
## S4Vectors          * 0.12.1  2016-12-01 Bioconductor
## scales              0.4.1    2016-11-09 CRAN (R 3.3.2)
## stringi             1.1.2    2016-10-01 cran (@1.1.2)
## stringr             * 1.1.0   2016-08-19 cran (@1.1.0)
## SummarizedExperiment * 1.4.0   2016-10-18 Bioconductor
## survival            2.40-1   2016-10-30 CRAN (R 3.3.0)
## sva                 3.22.0   2016-10-18 Bioconductor
## tibble              1.2      2016-08-26 cran (@1.2)
## VariantAnnotation  1.20.2   2016-12-01 Bioconductor
## withr               1.0.2    2016-06-20 CRAN (R 3.3.0)
## XML                 3.98-1.5 2016-11-10 CRAN (R 3.3.2)
## xtable              1.8-2    2016-02-05 CRAN (R 3.3.0)
## XVector             0.14.0   2016-10-18 Bioconductor
## yaml                2.1.14   2016-11-12 CRAN (R 3.3.2)
## zlibbioc            1.20.0   2016-10-18 Bioconductor

```