Comparison of Recount with GTEx"

Andrew E. Jaffe and Kasper D. Hansen

July 18, 2016

Contents

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 UCSC knownGene as represented by the `TxDb.Hsapiens.UCSC.hg38.knownGene` package, mapped to hg38.

Dependencies

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')
```

Data objects

From Recount

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_ge
    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)
```

```
## 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 <- gtexData[, 3:ncol(gtexData)]
rownames(gtexCounts) <- gtexData$Name
```

From elsewhere

These are the the Rail-RNA processed samples

```
# For when the data is available via recount:
## download_study('SRP012682')
## load('SRP012682/rse_gene.Rdata')

# Will remove this line in the future and uncomment the above
if(file.exists('SRP012682/rse_gene.Rdata')) {
    load('SRP012682/rse_gene.Rdata')
} else {
    load('/dcl01/leek/data/recount-website/rse/rse_gtex/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)]]
```

Mapping GTEx annotation

We map between version by using ENTREZ gene ids. The Recount representation is already using ENTREZ gene ids, but we need to map GTEx data to ENTREZ.

```
gtexMap <- gtexGtf[!duplicated(gtexGtf$gene_id)]
names(gtexMap) <- gtexMap$gene_id
gtexMap <- gtexMap[rownames(gtexCounts)]
stopifnot(all(rownames(gtexMap) == rownames(gtexCounts)))
gtexMap$EnsemblGeneID <- ballgown:::ss(gtexMap$gene_id, "\\.")
eid2ens <- select(org.Hs.eg.db, gtexMap$EnsemblGeneID,
    "ENTREZID", "ENSEMBL")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
eid2ens <- CharacterList(split(eid2ens$ENTREZID, eid2ens$ENSEMBL))
gtexMap$EntrezID <- eid2ens[gtexMap$EnsemblGeneID]
table(elementNROWS(gtexMap$EntrezID))
```

```
##
##     1     2     3     4     5     6     7     8
## 56001   278    28     3     2     2     3     1
```

```
table(gtexMap$gene_type, elementNROWS(gtexMap$EntrezID) == 1)
```

```
##
##                             FALSE  TRUE
##   3prime_overlapping_ncrna      0    20
##   antisense                    20  5207
##   IG_C_gene                     0    13
##   IG_C_pseudogene               0     9
##   IG_D_gene                     0    37
##   IG_J_gene                     0    16
##   IG_J_pseudogene               0     3
##   IG_V_gene                     0   138
##   IG_V_pseudogene               0   185
##   lincRNA                      32  6907
##   miRNA                        29  2938
##   misc_RNA                      0  2026
##   Mt_rRNA                       0     2
##   Mt_tRNA                       0    17
##   polymorphic_pseudogene        0    40
##   processed_transcript          7   456
##   protein_coding              213 19431
##   pseudogene                   15 13553
##   rRNA                          0   527
##   sense_intronic                1   724
##   sense_overlapping             0   197
##   snoRNA                        0  1446
##   snRNA                         0  1899
##   TR_C_gene                     0     5
##   TR_D_gene                     0     3
##   TR_J_gene                     0    74
##   TR_J_pseudogene               0     4
##   TR_V_gene                     0    97
##   TR_V_pseudogene               0    27
```

Basically, only protein coding genes have a ENTREZ id. We keep only the protein coding genes which are uniquely mapped to an ENTREZ id.

```
eidIndex <- which(elementNROWS(gtexMap$EntrezID)==1 &
    gtexMap$gene_type == "protein_coding")
gtexMap <- gtexMap[eidIndex,]
gtexCounts <- gtexCounts[eidIndex,]
gtexMap$EntrezID <- sapply(gtexMap$EntrezID,"[", 1)
```

Now, we still have multiple Ensembl gene ids mapping to the same ENTREZ id. We drop those as well

```
dropIDs <- gtexMap$EntrezID[duplicated(gtexMap$EntrezID)]
keepIdx <- which(!gtexMap$EntrezID %in% dropIDs)
gtexCounts <- gtexCounts[keepIdx, ]
gtexMap <- gtexMap[keepIdx, ]
rownames(gtexCounts) <- names(gtexMap) <- gtexMap$EntrezID
dim(gtexCounts)
```

```
## [1] 18382  8551
```

Let's load data from Recount.

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

## match by gene
geneMatch <- match(recountMap$gene_id, gtexMap$EntrezID )
recountCounts <- recountCounts[!is.na(geneMatch), ]
recountMap <- recountMap[!is.na(geneMatch)]

gtexMap <- gtexMap[geneMatch[!is.na(geneMatch)], ]
gtexCounts <- gtexCounts[geneMatch[!is.na(geneMatch)], ]
stopifnot(all(rownames(gtexCounts) == rownames(recountCounts)))

# save(gtexCounts, gtexMap, gtexPd, recountCounts, recountMap, file =
#     "GTEx_websiteAndRecount_EntrezMatched_17559_vs_8551_reproduced.Rdata",
#     compress = TRUE)
```

Comparison

```
# load("GTEx_websiteAndRecount_EntrezMatched_17559_vs_8551_reproduced.Rdata")
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
```

```
## -0.1040  0.9319  0.9647  0.9386  0.9838  0.9958       9
```

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

```
## [1] 6515
```

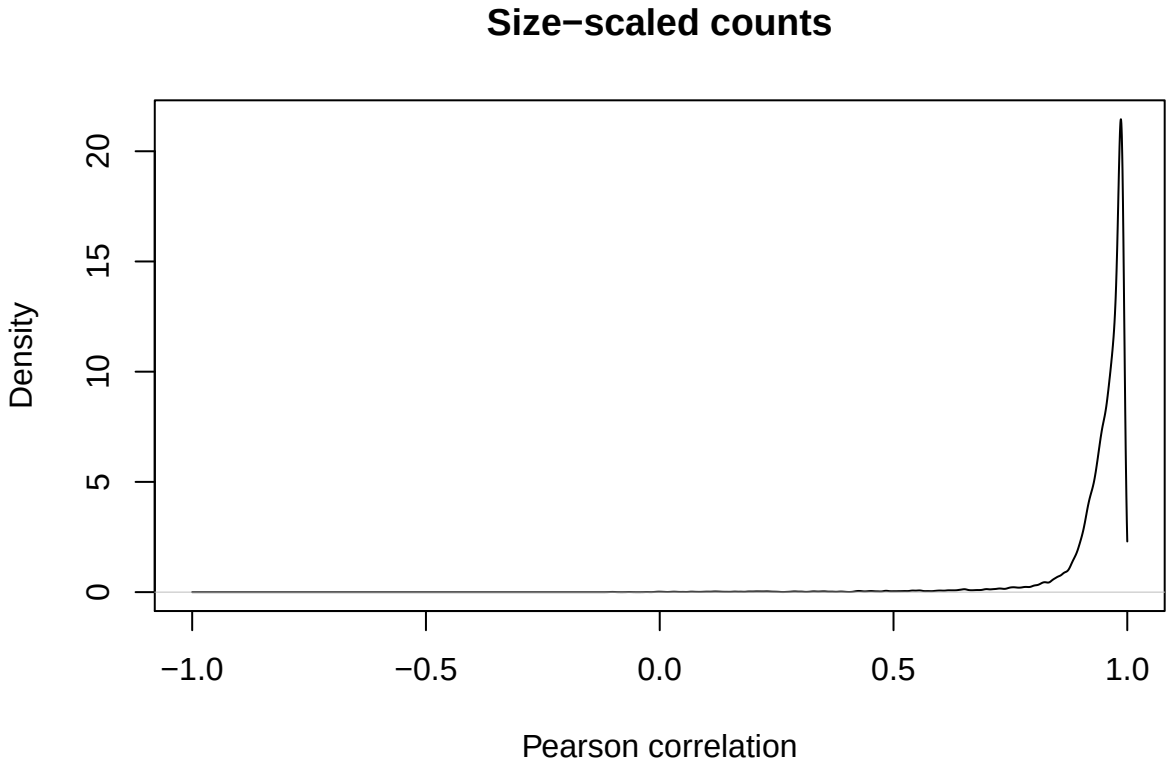
```
sum(normCors <= 0.80, na.rm = TRUE)
```

```
## [1] 799
```

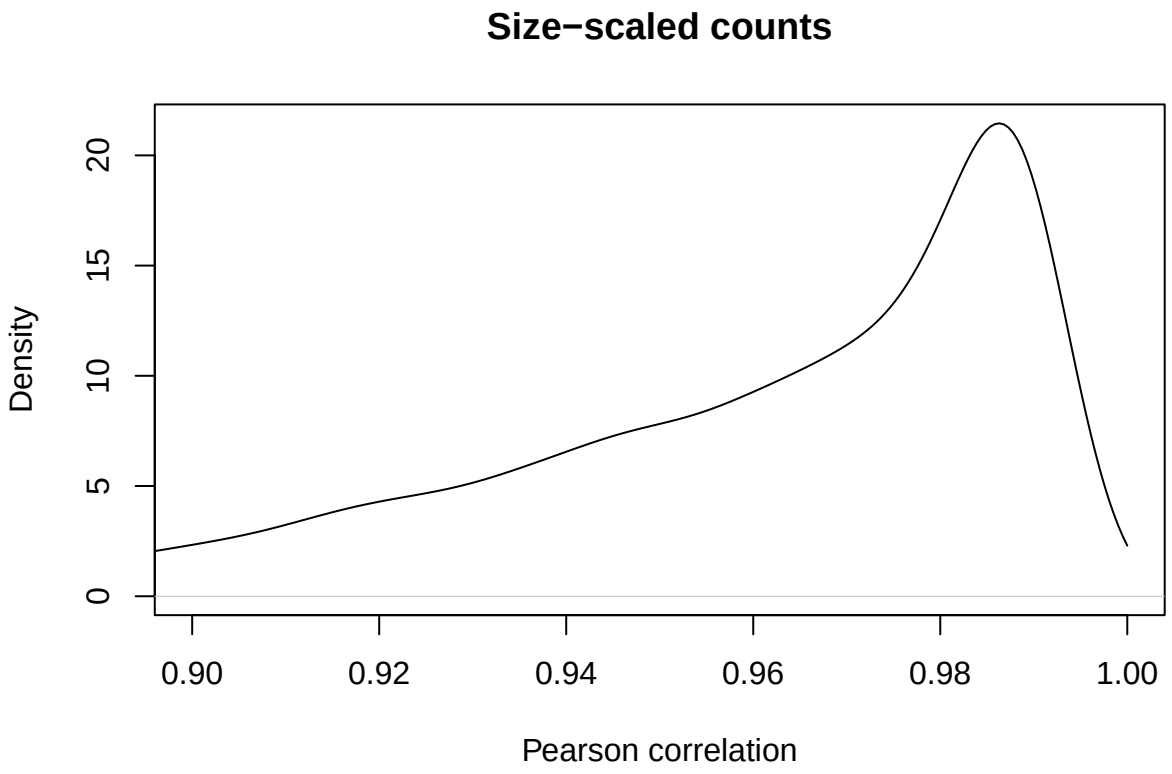
```
mean(normCors >= 0.99, na.rm = TRUE)
```

```
## [1] 0.07703704
```

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




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



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")
```

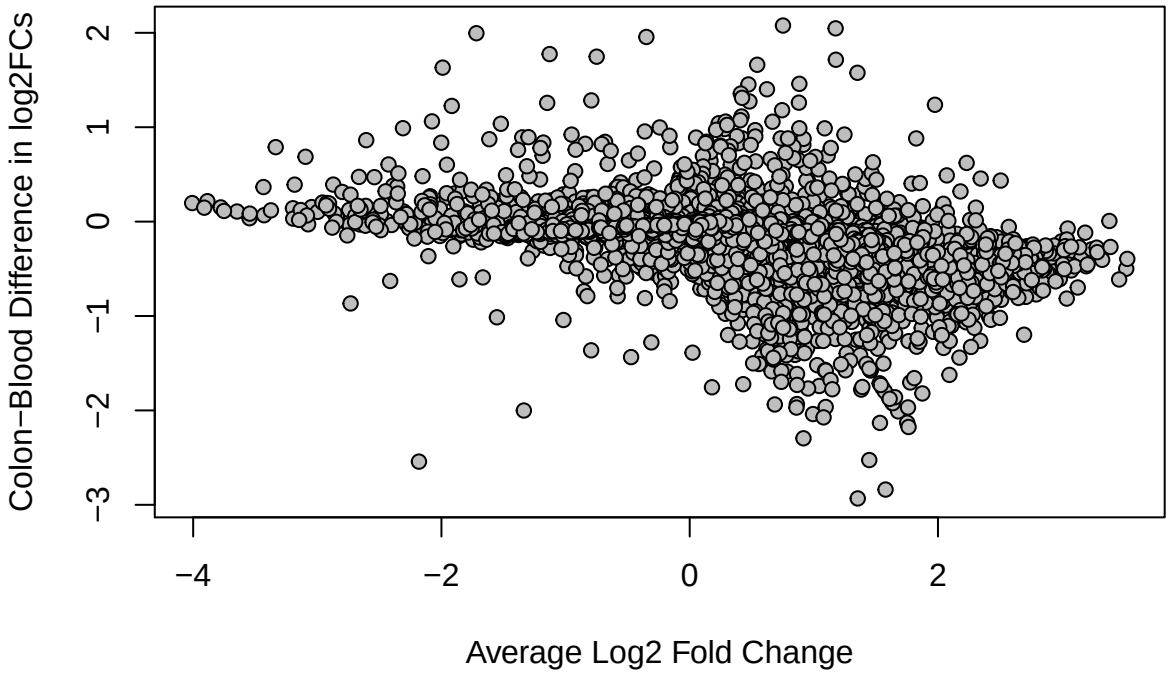
And using original counts:

```
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")
```



The R-squared is 0.9285357

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 and Francois, 2015)
- *recount* (Collado-Torres and Leek, 2016)
- *rmarkdown* (Allaire, Cheng, Xie, McPherson, et al., 2016)
- *rtracklayer* (Lawrence, Gentleman, and Carey, 2009)
- *stringr* (Wickham, 2015)
- *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 1.0. 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 3.3.0. 2016.

[4] L. Collado-Torres and J. T. Leek. recount: Explore and download data from the recount project. R package version 0.99.22. 2016. URL: https://github.com/leekgroup/recount.

[5] J. Fu, A. C. Frazee, L. Collado-Torres, A. E. Jaffe, et al. ballgown: Flexible, isoform-level differential expression analysis. R package version 2.5.2. 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 1.3.7. 2016.

[9] A. Oleś, M. Morgan and W. Huber. BiocStyle: Standard styles for vignettes and other Bioconductor documents. R package version 2.1.19. 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 0.4-0. 2016. URL: https://cran.r-project.org/package=coop.

[12] H. Wickham. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0. 2015. 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 and R. Francois. readr: Read Tabular Data. R package version 0.2.2. 2015. URL: https://CRAN.R-project.org/package=readr.

```
## Time spent creating this report:
diff(c(timestart, Sys.time()))
```

```
## Time difference of 7.348175 mins
```

```
## Date this report was generated
message(Sys.time())
```

```
## 2016-07-22 17:53:08
```

```
## Reproducibility info
options(width = 120)
devtools::session_info()
```

```
## Session info -------------------------------------------------------------------------------------------
```

```
##  setting  value
##  version  R version 3.3.1 (2016-06-21)
##  system   x86_64, darwin13.4.0
##  ui       X11
```

```
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2016-07-22

## Packages -----------------------------------------------------------------------------------------
##  package              * version  date       source
##  acepack                1.3-3.3  2014-11-24 CRAN (R 3.3.0)
##  annotate               1.51.0   2016-05-05 Bioconductor
##  AnnotationDbi        * 1.35.4   2016-07-08 Bioconductor
##  ballgown             * 2.5.2    2016-05-27 Bioconductor
##  bibtex                 0.4.0    2014-12-31 CRAN (R 3.3.0)
##  Biobase              * 2.33.0   2016-05-05 Bioconductor
##  BiocGenerics         * 0.19.2   2016-07-08 Bioconductor
##  BiocParallel           1.7.5    2016-07-21 Bioconductor
##  BiocStyle            * 2.1.19   2016-07-11 Bioconductor
##  biomaRt                2.29.2   2016-05-30 Bioconductor
##  Biostrings             2.41.4   2016-06-17 Bioconductor
##  bitops                 1.0-6    2013-08-17 CRAN (R 3.3.0)
##  BSgenome               1.41.2   2016-06-17 Bioconductor
##  bumphunter             1.13.1   2016-07-11 Bioconductor
##  chron                  2.3-47   2015-06-24 CRAN (R 3.3.0)
##  cluster                2.0.4    2016-04-18 CRAN (R 3.3.1)
##  codetools              0.2-14   2015-07-15 CRAN (R 3.3.1)
##  colorout             * 1.1-2    2016-05-05 Github (jalvesaq/colorout@6538970)
##  colorspace             1.2-6    2015-03-11 CRAN (R 3.3.0)
##  coop                 * 0.4-0    2016-04-05 CRAN (R 3.3.0)
##  data.table             1.9.6    2015-09-19 CRAN (R 3.3.0)
##  DBI                    0.4-1    2016-05-08 CRAN (R 3.3.0)
##  derfinder              1.7.9    2016-07-11 Bioconductor
##  derfinderHelper        1.7.3    2016-05-20 Bioconductor
##  devtools               1.12.0   2016-06-24 CRAN (R 3.3.0)
##  digest                 0.6.9    2016-01-08 CRAN (R 3.3.0)
##  doRNG                  1.6      2014-03-07 CRAN (R 3.3.0)
##  downloader             0.4      2015-07-09 CRAN (R 3.3.0)
##  edgeR                * 3.15.2   2016-07-08 Bioconductor
##  evaluate               0.9      2016-04-29 CRAN (R 3.3.0)
##  foreach                1.4.3    2015-10-13 CRAN (R 3.3.0)
##  foreign                0.8-66   2015-08-19 CRAN (R 3.3.1)
##  formatR                1.4      2016-05-09 CRAN (R 3.3.0)
##  Formula                1.2-1    2015-04-07 CRAN (R 3.3.0)
##  genefilter             1.55.2   2016-05-27 Bioconductor
##  GenomeInfoDb         * 1.9.4    2016-07-14 Bioconductor
##  GenomicAlignments      1.9.6    2016-07-17 Bioconductor
##  GenomicFeatures        1.25.15  2016-07-08 Bioconductor
##  GenomicFiles           1.9.11   2016-06-03 Bioconductor
##  GenomicRanges        * 1.25.9   2016-06-26 Bioconductor
##  GEOquery               2.39.3   2016-05-20 Bioconductor
##  ggplot2                2.1.0    2016-03-01 CRAN (R 3.3.0)
##  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                  3.17-4   2016-05-02 CRAN (R 3.3.0)
##  htmltools              0.3.5    2016-03-21 CRAN (R 3.3.0)
##  httr                   1.2.1    2016-07-03 CRAN (R 3.3.0)
```

```
##  IRanges              * 2.7.11   2016-06-22 Bioconductor
##  iterators              1.0.8    2015-10-13 CRAN (R 3.3.0)
##  jsonlite               1.0      2016-07-01 CRAN (R 3.3.0)
##  knitcitations        * 1.0.7    2015-10-28 CRAN (R 3.3.0)
##  knitr                  1.13     2016-05-09 CRAN (R 3.3.0)
##  lattice                0.20-33  2015-07-14 CRAN (R 3.3.1)
##  latticeExtra           0.6-28   2016-02-09 CRAN (R 3.3.0)
##  limma                * 3.29.16  2016-07-21 Bioconductor
##  locfit                 1.5-9.1  2013-04-20 CRAN (R 3.3.0)
##  lubridate              1.5.6    2016-04-06 CRAN (R 3.3.0)
##  magrittr               1.5      2014-11-22 CRAN (R 3.3.0)
##  Matrix                 1.2-6    2016-05-02 CRAN (R 3.3.1)
##  matrixStats            0.50.2   2016-04-24 CRAN (R 3.3.0)
##  memoise                1.0.0    2016-01-29 CRAN (R 3.3.0)
##  mgcv                   1.8-13   2016-07-21 CRAN (R 3.3.0)
##  munsell                0.4.3    2016-02-13 CRAN (R 3.3.0)
##  nlme                   3.1-128  2016-05-10 CRAN (R 3.3.1)
##  nnet                   7.3-12   2016-02-02 CRAN (R 3.3.1)
##  org.Hs.eg.db         * 3.3.0    2016-07-22 Bioconductor
##  pkgmaker               0.22     2014-05-14 CRAN (R 3.3.0)
##  plyr                   1.8.4    2016-06-08 CRAN (R 3.3.0)
##  qvalue                 2.5.2    2016-05-27 Bioconductor
##  R6                     2.1.2    2016-01-26 CRAN (R 3.3.0)
##  RColorBrewer           1.1-2    2014-12-07 CRAN (R 3.3.0)
##  Rcpp                   0.12.6   2016-07-19 CRAN (R 3.3.0)
##  RCurl                  1.95-4.8 2016-03-01 CRAN (R 3.3.0)
##  readr                * 0.2.2    2015-10-22 CRAN (R 3.3.0)
##  recount              * 0.99.22  2016-07-14 Bioconductor
##  RefManageR             0.10.13  2016-04-04 CRAN (R 3.3.0)
##  registry               0.3      2015-07-08 CRAN (R 3.3.0)
##  rentrez                1.0.2    2016-04-21 CRAN (R 3.3.0)
##  reshape2               1.4.1    2014-12-06 CRAN (R 3.3.0)
##  RJSONIO                1.3-0    2014-07-28 CRAN (R 3.3.0)
##  rmarkdown            * 1.0      2016-07-08 CRAN (R 3.3.0)
##  rngtools               1.2.4    2014-03-06 CRAN (R 3.3.0)
##  rpart                  4.1-10   2015-06-29 CRAN (R 3.3.1)
##  Rsamtools              1.25.0   2016-05-05 Bioconductor
##  RSQLite                1.0.0    2014-10-25 CRAN (R 3.3.0)
##  rstudioapi             0.6      2016-06-27 CRAN (R 3.3.0)
##  rtracklayer          * 1.33.10  2016-07-14 Github (Bioconductor-mirror/rtracklayer@136a2bd)
##  S4Vectors            * 0.11.10  2016-07-21 Bioconductor
##  scales                 0.4.0    2016-02-26 CRAN (R 3.3.0)
##  stringi                1.1.1    2016-05-27 CRAN (R 3.3.0)
##  stringr              * 1.0.0    2015-04-30 CRAN (R 3.3.0)
##  SummarizedExperiment * 1.3.7    2016-07-08 Bioconductor
##  survival               2.39-5   2016-06-26 CRAN (R 3.3.0)
##  sva                    3.21.0   2016-05-27 Bioconductor
##  VariantAnnotation      1.19.8   2016-07-12 Bioconductor
##  withr                  1.0.2    2016-06-20 CRAN (R 3.3.0)
##  XML                    3.98-1.4 2016-03-01 CRAN (R 3.3.0)
##  xtable                 1.8-2    2016-02-05 CRAN (R 3.3.0)
##  XVector                0.13.6   2016-07-08 Bioconductor
##  yaml                   2.1.13   2014-06-12 CRAN (R 3.3.0)
##  zlibbioc               1.19.0   2016-05-05 Bioconductor
```