recount (DER analyses)

Kai Kammers and Shannon Ellis

June 09, 2016

Contents

Here is an example of how to download and analyze a `RangedSummarizedExperiment` object with the gene counts with SRA study id SRP032789.

We first load the required packages.

```
## load libraries
library('recount')
library('SummarizedExperiment')
library('limma')
library('edgeR')
library('qvalue')
library('topGO')
library('matrixStats')
library('RSkittleBrewer')
library('derfinder')
library('BiocParallel')
library('GenomicRanges')

## set colors
trop <- RSkittleBrewer('tropical')[c(1, 2)]
```

Expressed regions

```
chrs <- paste0('chr', c(1:22, 'X', 'Y'))
bp <- SerialParam() ## Change if you have access to more cores

if(!file.exists('regions_SRP032789.Rdata')) {
    regions_list <- bplapply(chrs, function(chr) {
        regs <- expressed_regions('SRP032789', chr, cutoff = 5L,
            maxClusterGap = 3000L, verbose = FALSE)
        return(regs)
    }, BPPARAM = bp)
    names(regions_list) <- chrs
    regions <- unlist(GRangesList(regions_list))

    ## Save the regions
    save(regions, regions_list, file = 'regions_SRP032789.Rdata')
} else {
    load('regions_SRP032789.Rdata')
}
```

```
## Compute coverage matrix for study SRP032789, only for chromosome 22
if(!file.exists('covMat_SRP032789.Rdata')) {
    covMat <- bplapply(chrs, function(chr) {
        coverageMatrix <- coverage_matrix('SRP032789', chr,
            regions_list[[chr]], verboseLoad = FALSE)
        return(coverageMatrix)
    }, BPPARAM = bp)
    covMat <- do.call(rbind, covMat)

    ## Round the coverage matrix to integers
    covMat <- round(covMat, 0)
    save(covMat, file = 'covMat_SRP032789.Rdata')
} else {
    load('covMat_SRP032789.Rdata')
}
```

```
## download phenotype data from
## http://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP032789
pheno <- read.table('SraRunTable_SRP032789.txt', sep = '\t',
    header=TRUE,
    stringsAsFactors = FALSE)

## check ordering of samples
pheno <- pheno[pheno$Run_s %in% colnames(covMat), ]
identical(pheno$Run_s, colnames(covMat))
```

```
## [1] FALSE
```

```
## obain correct order for pheno data
pheno <- pheno[match(colnames(covMat), pheno$Run_s), ]
identical(pheno$Run_s, colnames(covMat))
```

```
## [1] TRUE
```

```
head(cbind(pheno$Run_s, colnames(covMat)))
```

```
##      [,1]         [,2]
## [1,] "SRR1027171" "SRR1027171"
## [2,] "SRR1027173" "SRR1027173"
## [3,] "SRR1027174" "SRR1027174"
## [4,] "SRR1027175" "SRR1027175"
## [5,] "SRR1027176" "SRR1027176"
## [6,] "SRR1027177" "SRR1027177"
```

```
## find grouping information
group <- pheno$tumor_type_s
table(group)
```

```
## group
## HER2 Positive Breast Tumor      Non-TNBC Breast Tumor
##                          5                          6
##    Normal Breast Organoids          TNBC Breast Tumor
##                          3                          6
```

```
## subset data to HER2 and TNBC type
covMat_filt <- covMat[, group %in% c('HER2 Positive Breast Tumor', 'TNBC Breast Tumor')]
```

```
group <- group[group %in% c('HER2 Positive Breast Tumor', 'TNBC Breast Tumor')]
dim(covMat_filt)
```

```
## [1] 328481     11
```

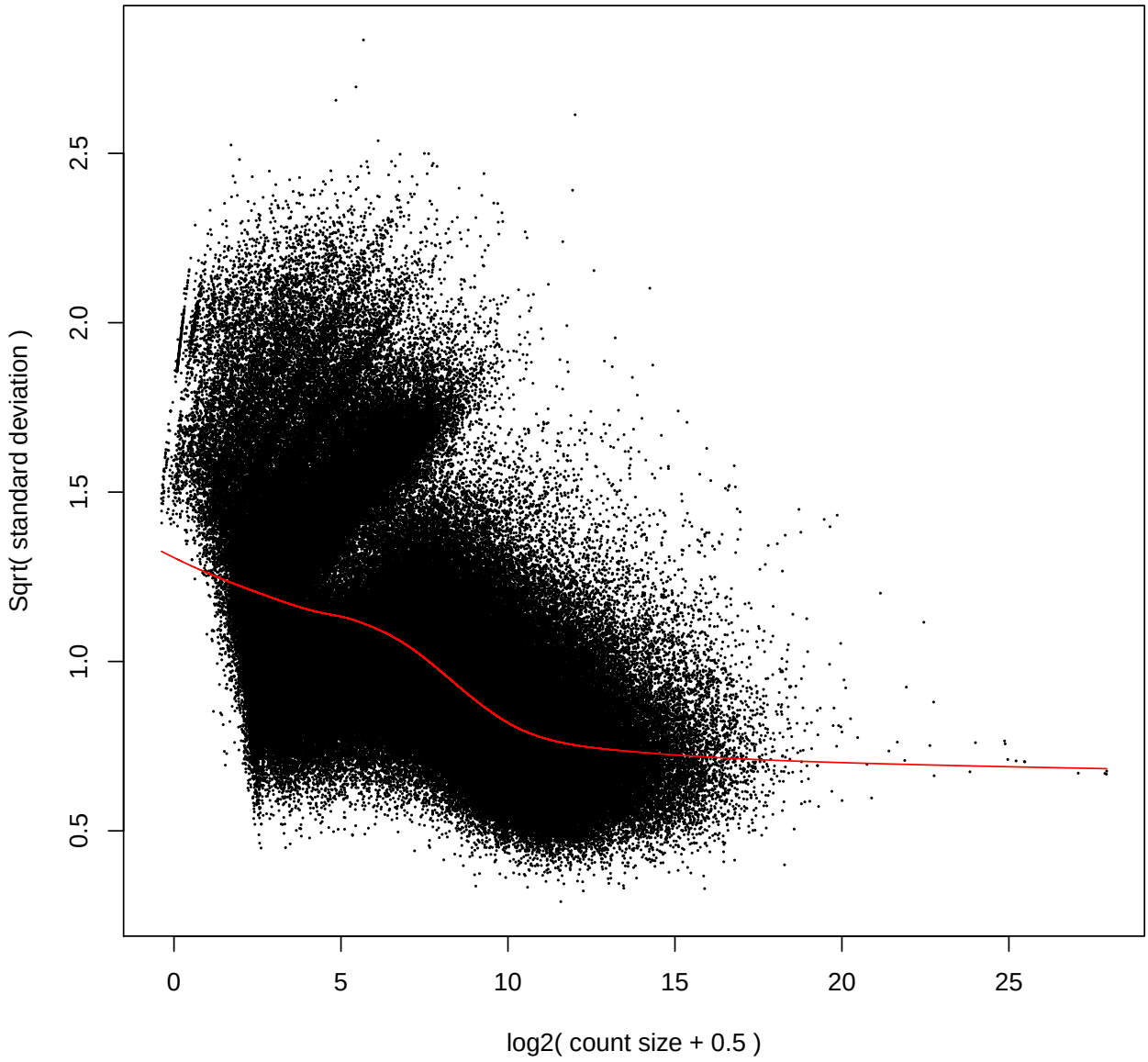
```
## filter count matrix
counts <- covMat_filt
filter <- apply(counts, 1, function(x) mean(x) > 5)
counts <- counts[filter, ]
dim(counts)
```

```
## [1] 291247     11
```

```
design <- model.matrix(~ group)
design
```

```
##    (Intercept) groupTNBC Breast Tumor
## 1            1                      1
## 2            1                      1
## 3            1                      1
## 4            1                      1
## 5            1                      1
## 6            1                      0
## 7            1                      0
## 8            1                      0
## 9            1                      0
## 10           1                      0
## 11           1                      1
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

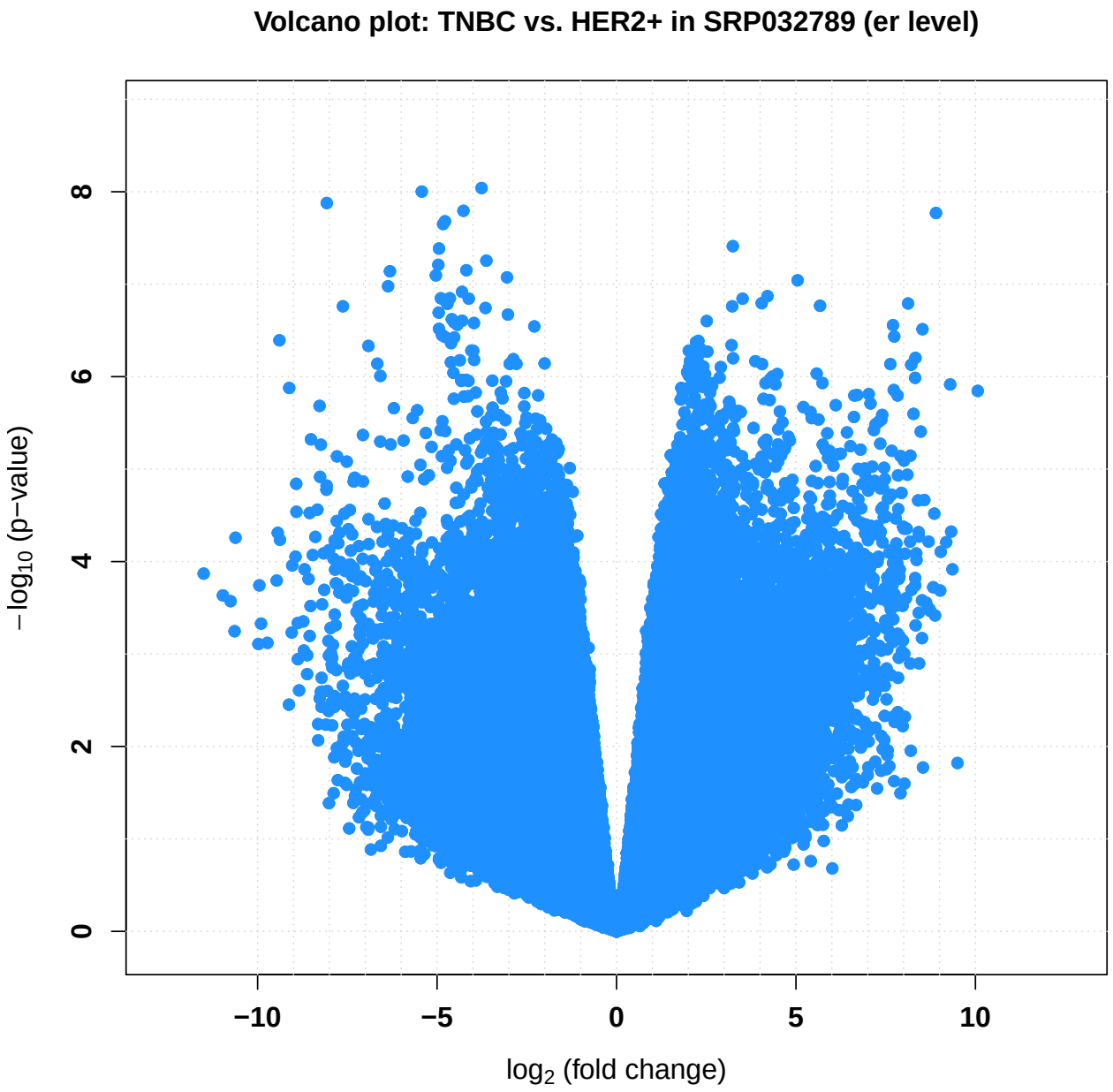
```
dge <- DGEList(counts = counts)
dge <- calcNormFactors(dge)
v <- voom(dge, design,plot = TRUE)
```




```
fit <- lmFit(v, design)
fit <- eBayes(fit)
log2FC <- fit$coefficients[, 2]
p.mod <- fit$p.value[, 2]
q.mod <- qvalue(p.mod)$q

## Volcano plot
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
rx2 <- c(-1, 1) * 1.1 * max(abs(log2FC))
ry2 <- c(-0.1, max(-log10(p.mod))) * 1.1
plot(log2FC, -log10(p.mod),
     pch = 19, xlim = rx2, ylim = ry2, col = trop[2],
     xlab = bquote(paste(log[2], ' (fold change)')),
     ylab = bquote(paste(-log[10], ' (p-value)')))
abline(v = seq(-10, 10, 1), col = 'lightgray', lty = 'dotted')
```

```
abline(h = seq(0, 23, 1), col = 'lightgray', lty = 'dotted')
points(log2FC, -log10(p.mod), pch = 19, col = trop[2])
title('Volcano plot: TNBC vs. HER2+ in SRP032789 (er level)')
```



Reproducibility

This analysis report was made possible thanks to:

- R (R Core Team, 2016)
- *BiocParallel* (Morgan, Obenchain, Lang, and Thompson, 2016)
- *BiocStyle* (Oleś, Morgan, and Huber, 2016)
- *derfinder* (Collado-Torres, Nellore, Frazee, Wilks, et al., 2016)
- *devtools* (Wickham and Chang, 2016)

- *edgeR* (Robinson, McCarthy, and Smyth, 2010)
- *GenomicRanges* (Lawrence, Huber, Pagès, Aboyoun, et al., 2013)
- *knitcitations* (Boettiger, 2015)
- *matrixStats* (Bengtsson, 2016)
- *recount* (Collado-Torres and Leek, 2016)
- *rmarkdown* (Allaire, Cheng, Xie, McPherson, et al., 2016)
- *RSkittleBrewer* (Frazee, 2016)
- *SummarizedExperiment* (Morgan, Obenchain, Hester, and Pagès, 2016)
- *topGO* (Alexa and Rahnenfuhrer, 2016)
- *limma* (Law, Chen, Shi, and Smyth, 2014)

Bibliography file

[1] A. Alexa and J. Rahnenfuhrer. topGO: Enrichment Analysis for Gene Ontology. R package version 2.25.0. 2016.

[2] J. Allaire, J. Cheng, Y. Xie, J. McPherson, et al. rmarkdown: Dynamic Documents for R. R package version 0.9.6. 2016. URL: https://CRAN.R-project.org/package=rmarkdown.

[3] H. Bengtsson. matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors). R package version 0.50.2. 2016. URL: https://CRAN.R-project.org/package=matrixStats.

[4] C. Boettiger. knitcitations: Citations for 'Knitr' Markdown Files. R package version 1.0.7. 2015. URL: https://CRAN.R-project.org/package=knitcitations.

[5] L. Collado-Torres and J. T. Leek. recount: Explore and download data from the recount project. R package version 0.99.10. 2016. URL: https://github.com/leekgroup/recount.

[6] L. Collado-Torres, A. Nellore, A. C. Frazee, C. Wilks, et al. "Flexible expressed region analysis for RNA-seq with derfinder". In: bioRxiv (2016). DOI: 10.1101/015370. URL: http://biorxiv.org/content/early/2016/05/07/015370.

[7] A. Frazee. RSkittleBrewer: Fun with R Colors. R package version 1.1. 2016. URL: https://github.com/alyssafrazee/RSkittlel

[8] C. Law, Y. Chen, W. Shi and G. Smyth. "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts". In: Genome Biology 15 (2014), p. R29.

[9] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, et al. "Software for Computing and Annotating Genomic Ranges". In: PLoS Computational Biology 9 (8 2013). DOI: 10.1371/journal.pcbi.1003118. URL: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118}.

[10] M. Morgan, V. Obenchain, J. Hester and H. Pagès. SummarizedExperiment: SummarizedExperiment container. R package version 1.3.4. 2016.

[11] M. Morgan, V. Obenchain, M. Lang and R. Thompson. BiocParallel: Bioconductor facilities for parallel evaluation. R package version 1.7.2. 2016. URL: https://github.com/Bioconductor/BiocParallel.

[12] A. Oleś, M. Morgan and W. Huber. BiocStyle: Standard styles for vignettes and other Bioconductor documents. R package version 2.1.6. 2016. URL: https://github.com/Bioconductor/BiocStyle.

[13] 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/.

[14] M. D. Robinson, D. J. McCarthy and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: Bioinformatics 26 (2010), pp. -1.

[15] H. Wickham and W. Chang. devtools: Tools to Make Developing R Packages Easier. R package version 1.11.1. 2016. URL: https://CRAN.R-project.org/package=devtools.

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

```
## Time difference of 3.168565 mins
```

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

```
## 2016-06-13 18:19:21
```

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

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

```
##  setting  value
##  version  R version 3.3.0 RC (2016-05-01 r70572)
##  system   x86_64, darwin13.4.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2016-06-13
```

```
## Packages -----------------------------------------------------------------------------------------
```

```
##  package              * version  date       source
##  acepack                1.3-3.3  2014-11-24 CRAN (R 3.3.0)
##  AnnotationDbi        * 1.35.3   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         * 0.19.1   2016-06-11 Bioconductor
##  BiocParallel         * 1.7.2    2016-05-20 Bioconductor
##  BiocStyle            * 2.1.6    2016-06-11 Bioconductor
##  biomaRt                2.29.2   2016-05-30 Bioconductor
##  Biostrings             2.41.2   2016-06-08 Bioconductor
##  bitops                 1.0-6    2013-08-17 CRAN (R 3.3.0)
##  BSgenome               1.41.0   2016-05-05 Bioconductor
##  bumphunter             1.13.0   2016-05-05 Bioconductor
##  chron                  2.3-47   2015-06-24 CRAN (R 3.3.0)
##  cluster                2.0.4    2016-04-18 CRAN (R 3.3.0)
##  codetools              0.2-14   2015-07-15 CRAN (R 3.3.0)
##  colorout             * 1.1-2    2016-05-05 Github (jalvesaq/colorout@6538970)
##  colorspace             1.2-6    2015-03-11 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            * 1.7.8    2016-06-08 Bioconductor
##  derfinderHelper        1.7.3    2016-05-20 Bioconductor
##  devtools               1.11.1   2016-04-21 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)
##  edgeR                * 3.15.0   2016-05-27 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 3.3.0)
##  formatR                1.4      2016-05-09 CRAN (R 3.3.0)
##  Formula                1.2-1    2015-04-07 CRAN (R 3.3.0)
##  GenomeInfoDb         * 1.9.1    2016-05-13 Bioconductor
##  GenomicAlignments      1.9.2    2016-06-13 Bioconductor
##  GenomicFeatures        1.25.12  2016-05-21 Bioconductor
```

```
##  GenomicFiles           1.9.11   2016-06-03 Bioconductor
##  GenomicRanges        * 1.25.4   2016-06-10 Bioconductor
##  ggplot2                2.1.0    2016-03-01 CRAN (R 3.3.0)
##  GO.db                * 3.3.0    2016-05-05 Bioconductor
##  graph                * 1.51.0   2016-05-05 Bioconductor
##  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                   1.1.0    2016-01-28 CRAN (R 3.3.0)
##  IRanges              * 2.7.6    2016-06-10 Bioconductor
##  iterators              1.0.8    2015-10-13 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 3.3.0)
##  latticeExtra           0.6-28   2016-02-09 CRAN (R 3.3.0)
##  limma                * 3.29.7   2016-06-13 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 3.3.0)
##  matrixStats          * 0.50.2   2016-04-24 CRAN (R 3.3.0)
##  memoise                1.0.0    2016-01-29 CRAN (R 3.3.0)
##  munsell                0.4.3    2016-02-13 CRAN (R 3.3.0)
##  nnet                   7.3-12   2016-02-02 CRAN (R 3.3.0)
##  pkgmaker               0.22     2014-05-14 CRAN (R 3.3.0)
##  plyr                   1.8.3    2015-06-12 CRAN (R 3.3.0)
##  qvalue               * 2.5.2    2016-05-20 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                   0.12.5   2016-05-14 CRAN (R 3.3.0)
##  RCurl                  1.95-4.8 2016-03-01 CRAN (R 3.3.0)
##  recount              * 0.99.10  2016-06-12 Github (leekgroup/recount@7a7ea73)
##  RefManageR             0.10.13  2016-04-04 CRAN (R 3.3.0)
##  registry               0.3      2015-07-08 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            * 0.9.6    2016-05-01 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 3.3.0)
##  Rsamtools              1.25.0   2016-05-05 Bioconductor
##  RSkittleBrewer       * 1.1      2016-06-13 Github (alyssafrazee/RSkittleBrewer@230d1d0)
##  RSQLite                1.0.0    2014-10-25 CRAN (R 3.3.0)
##  rstudioapi             0.5      2016-01-24 CRAN (R 3.3.0)
##  rtracklayer            1.33.5   2016-06-13 Bioconductor
##  S4Vectors            * 0.11.4   2016-06-11 Bioconductor
##  scales                 0.4.0    2016-02-26 CRAN (R 3.3.0)
##  SparseM              * 1.7      2015-08-15 CRAN (R 3.3.0)
##  stringi                1.0-1    2015-10-22 CRAN (R 3.3.0)
##  stringr                1.0.0    2015-04-30 CRAN (R 3.3.0)
##  SummarizedExperiment * 1.3.4    2016-06-10 Bioconductor
##  survival               2.39-4   2016-05-11 CRAN (R 3.3.0)
##  topGO                * 2.25.0   2016-05-05 Bioconductor
##  VariantAnnotation      1.19.2   2016-06-07 Bioconductor
```

```
##  withr                1.0.1    2016-02-04 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              0.13.0   2016-05-05 Bioconductor
##  yaml                 2.1.13   2014-06-12 CRAN (R 3.3.0)
##  zlibbioc             1.19.0   2016-05-05 Bioconductor
```