Jongchul Seon

RNA Count example 1

Load

biological replicates C57BL/6J (10) <-> DBA/2J (11) are hybridized in each experiment and lane.

experiment number and lane.number are different.There will be some variances among these.

We are going to use limma package to compare the means of gene expression values for two groups of replicates for a given gene

```
library(devtools)
```

```
## WARNING: Rtools is required to build R packages, but is not
currently installed.
##
## Please download and install Rtools 3.3 from http://cran.r-
project.org/bin/windows/Rtools/ and then run find_rtools().
```

```
library(Biobase)
```

```
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
##
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs
##
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, as.vector, cbind,
##     colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
##     grep, grepl, intersect, is.unsorted, lapply, lengths, Map,
##     mapply, match, mget, order, paste, pmax, pmax.int, pmin,
##     pmin.int, Position, rank, rbind, Reduce, rownames, sapply,
##     setdiff, sort, table, tapply, union, unique, unlist, unsplit
```

```
##
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

library(goseq)

## Loading required package: BiasedUrn
## Loading required package: geneLenDataBase
## Loading required package: DBI

library(limma)

##
## Attaching package: 'limma'
##
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA

library(genefilter)

##
## Attaching package: 'genefilter'
##
## The following object is masked from 'package:base':
##
##     anyNA

#### library(DESeq2)



con =url("http://bowtie-
bio.sourceforge.net/recount/ExpressionSets/bottomly_eset.RData")
load(file=con)
close(con)
bot = bottomly.eset
pdata_bot=pData(bot)
fdata_bot = featureData(bot)
edata = exprs(bot)

sum(is.na(edata))

## [1] 0

dim(edata)

## [1] 36536    21

dim(pdata_bot)
```

```
## [1] 21  5
```

```
dim(fdata_bot)
```

```
##   featureNames featureColumns
##          36536              1
```

```
head(edata)
```

```
##                     SRX033480 SRX033488 SRX033481 SRX033489 SRX033482
## ENSMUSG00000000001       369       744       287       769       348
## ENSMUSG00000000003         0         0         0         0         0
## ENSMUSG00000000028         0         1         0         1         1
## ENSMUSG00000000031         0         0         0         0         0
## ENSMUSG00000000037         0         1         1         5         0
## ENSMUSG00000000049         0         1         0         1         0
##                     SRX033490 SRX033483 SRX033476 SRX033478 SRX033479
## ENSMUSG00000000001       803       433       469       585       321
## ENSMUSG00000000003         0         0         0         0         0
## ENSMUSG00000000028         1         0         7         6         1
## ENSMUSG00000000031         0         0         0         0         0
## ENSMUSG00000000037         4         0         0         0         0
## ENSMUSG00000000049         0         0         0         0         0
##                     SRX033472 SRX033473 SRX033474 SRX033475 SRX033491
## ENSMUSG00000000001       301       461       309       374       781
## ENSMUSG00000000003         0         0         0         0         0
## ENSMUSG00000000028         1         1         1         1         1
## ENSMUSG00000000031         0         0         0         0         0
## ENSMUSG00000000037         4         1         1         0         1
## ENSMUSG00000000049         0         0         0         0         0
##                     SRX033484 SRX033492 SRX033485 SRX033493 SRX033486
## ENSMUSG00000000001       555       820       294       758       419
## ENSMUSG00000000003         0         0         0         0         0
## ENSMUSG00000000028         2         1         1         4         1
## ENSMUSG00000000031         0         0         0         0         0
## ENSMUSG00000000037         2         1         1         1         1
## ENSMUSG00000000049         0         0         0         0         0
##                     SRX033494
## ENSMUSG00000000001       857
## ENSMUSG00000000003         0
## ENSMUSG00000000028         5
## ENSMUSG00000000031         0
## ENSMUSG00000000037         2
## ENSMUSG00000000049         0
```

```
head(pdata_bot)
```

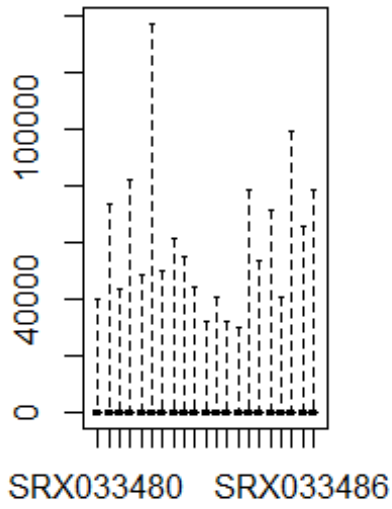
```
##           sample.id num.tech.reps   strain experiment.number
lane.number
## SRX033480 SRX033480             1 C57BL/6J                 6
1
## SRX033488 SRX033488             1 C57BL/6J                 7
```

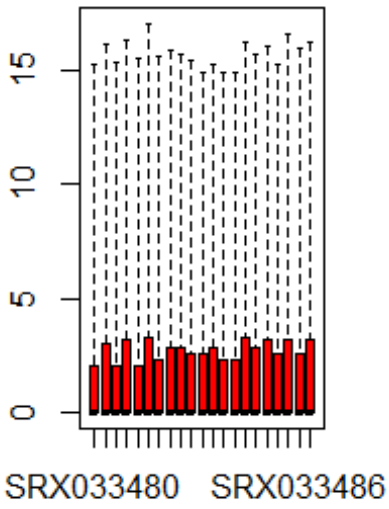
```
1
## SRX033481 SRX033481             1 C57BL/6J                 6
2
## SRX033489 SRX033489             1 C57BL/6J                 7
2
## SRX033482 SRX033482             1 C57BL/6J                 6
3
## SRX033490 SRX033490             1 C57BL/6J                 7
3

unique(pdata_bot[,1])

##  [1] SRX033480 SRX033488 SRX033481 SRX033489 SRX033482 SRX033490
SRX033483
##  [8] SRX033476 SRX033478 SRX033479 SRX033472 SRX033473 SRX033474
SRX033475
## [15] SRX033491 SRX033484 SRX033492 SRX033485 SRX033493 SRX033486
SRX033494
## 21 Levels: SRX033472 SRX033473 SRX033474 SRX033475 SRX033476 ...
SRX033494
```

Genes whose average counts are over than 5 are selected and log(2) transformed.

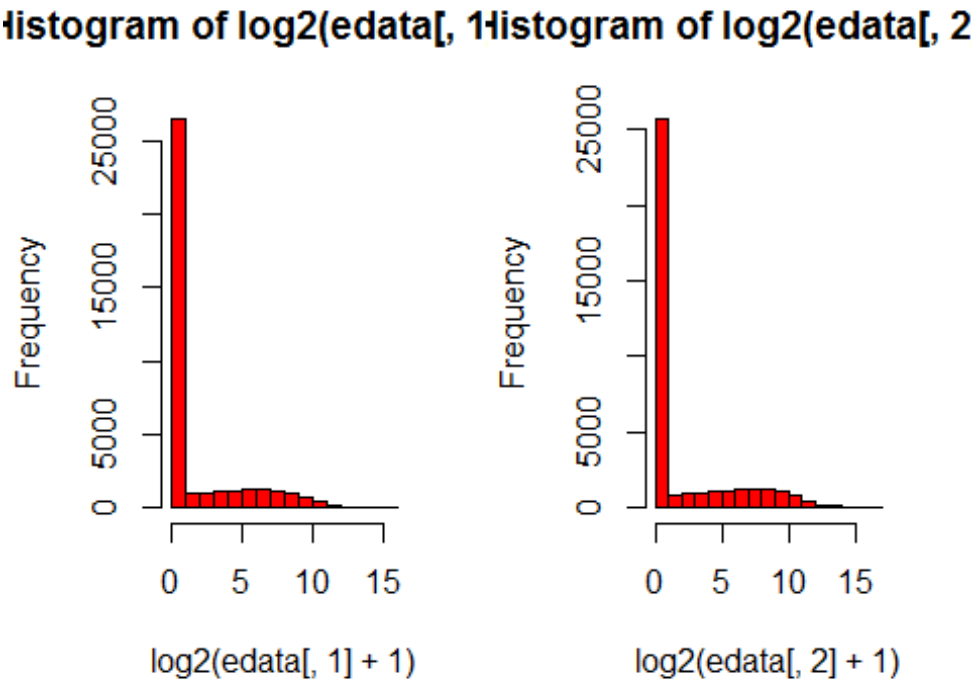
A common pre-processing technique is to remove features that don't have much data

```
par(mfrow=c(1,2))

boxplot(edata,col=2,range=0)
boxplot(log2(edata+1),col=2,range=0)
```






```
hist(log2(edata[,1]+1),col=2)
hist(log2(edata[,2]+1),col=2)
```



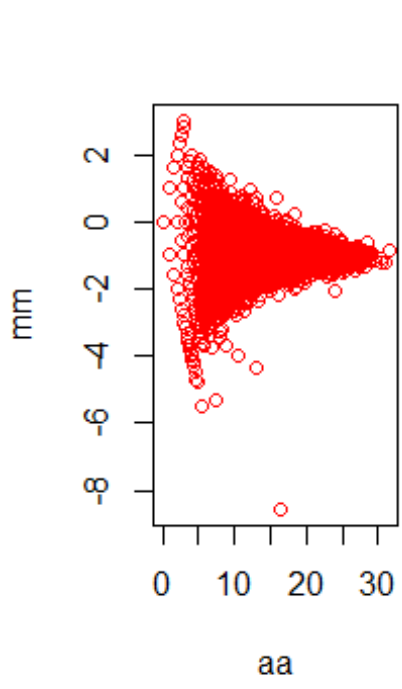
```
mm = log2(edata[,1]+1) - log2(edata[,2]+1)
aa = log2(edata[,1]+1) + log2(edata[,2]+1)
plot(aa,mm,col=2)


fdata_bot = fdata_bot[rowMeans(edata) > 5]
edata = edata[rowMeans(edata) > 5, ]
edata = log2(as.matrix(edata) + 1)

hist(edata,,col=2,range=0)
```

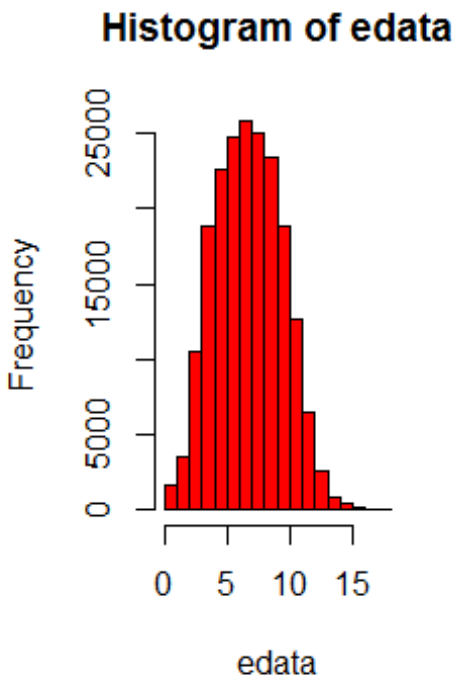
```
## Warning in plot.window(xlim, ylim, "", ...): "range" is not a
graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab =
ylab, ...):
## "range" is not a graphical parameter
```

```
## Warning in axis(1, ...): "range" is not a graphical parameter
```

```
## Warning in axis(2, ...): "range" is not a graphical parameter
```





NCBI Build 37, mm9. check out the paper which describes the data, Evaluating Gene Expression in C57BL/6J and DBA/2JMouse Striatum Using RNA-Seq and Microarrays. http://www.ncbi.nlm.nih.gov/pubmed?term=21455293.

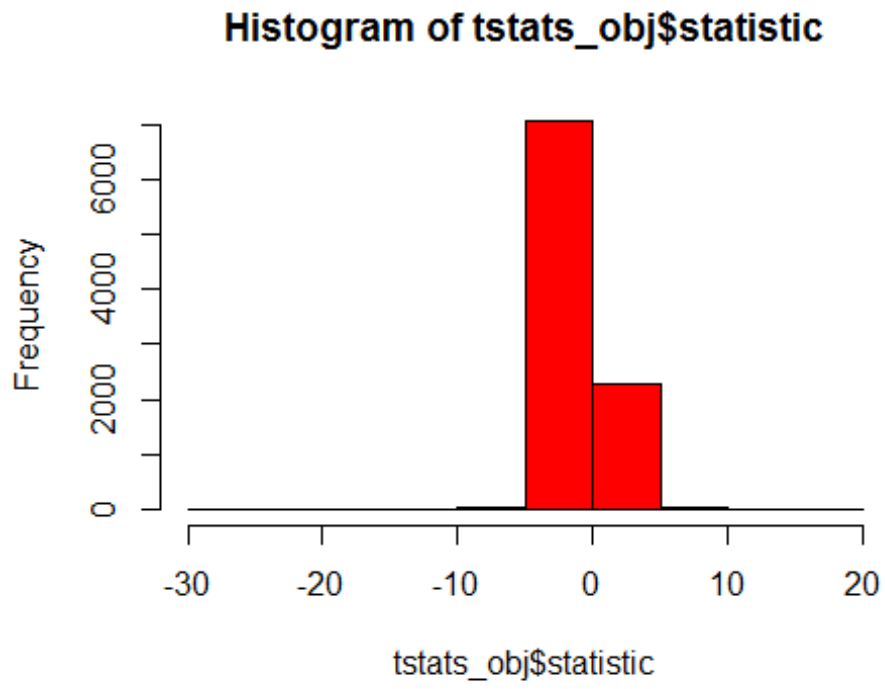
```
class(supportedGenomes())

## [1] "data.frame"

isS4(supportedGenomes())

## [1] FALSE

head(supportedGenomes())

##        db species      date                                name
## 1    hg38   Human Dec. 2013  Genome Reference Consortium GRCh38
## 2    hg19   Human Feb. 2009  Genome Reference Consortium GRCh37
## 3    hg18   Human Mar. 2006                     NCBI Build 36.1
## 4    hg17   Human  May 2004                       NCBI Build 35
## 5    hg16   Human Jul. 2003                       NCBI Build 34
## 6 vicPac2  Alpaca Mar. 2013 Broad Institute Vicugna_pacos-2.0.1
##
AvailableGeneIDs
## 1

## 2
ccdsGene,ensGene,exoniphy,geneSymbol,knownGene,nscanGene,refGene,xenoRe
fGene
## 3
acembly,acescan,ccdsGene,ensGene,exoniphy,geneSymbol,geneid,genscan,kno
wnGene,knownGeneOld3,refGene,sgpGene,sibGene,xenoRefGene
## 4
acembly,acescan,ccdsGene,ensGene,exoniphy,geneSymbol,geneid,genscan,kno
wnGene,refGene,sgpGene,vegaGene,vegaPseudoGene,xenoRefGene
## 5
acembly,ensGene,exoniphy,geneSymbol,geneid,genscan,knownGene,refGene,sg
pGene
## 6

species <- supportedGenomes()[,2]
## species

## source("http://www.bioconductor.org/biocLite.R")
## biocLite("org.Mm.eg.db")


#### species[species %in% "Mouse"]
```

Using limma package to find differently expressed genes between two strands.

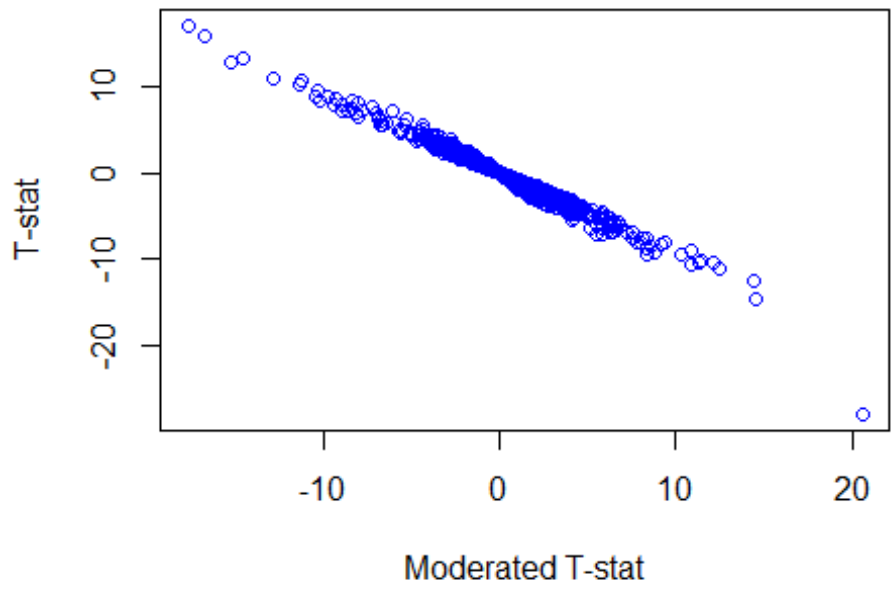
```
mod = model.matrix(~ pdata_bot$strain)
fit_limma = lmFit(edata,mod)
ebayes_limma = eBayes(fit_limma)
```

limma dose moderated t-statistics. Moderated t-statistics lead to pvalues in the same way that ordinary t-statistics do except that the degrees of freedom are increased, reflecting the greater reliability associated with the smoothed standard errors.

```
par(mfrow=c(1,1))

tstats_obj = rowttests(edata,pdata_bot$strain)
##names(tstats_obj)
hist(tstats_obj$statistic,col=2)
```




```
plot(ebayes_limma$t[,2],tstats_obj$statistic,col=4,
     xlab="Moderated T-stat",ylab="T-stat")
```



Bonferroni and Benjamini-Hochberg FDR correction with p.adjust

You can use the `p.adjust` function to get "multiple testing corrected" p-values which you can then use to control error rates.

223 genes are differently expressesed between two strands at the 5% FDR level using Benjamini-Hochberg correction.

```
limma_pvals =
topTable(ebayes_limma,adjust.method="BH",sort.by="none",number=dim(edat
a)[1])

## Removing intercept from test coefficients

dim(limma_pvals)

## [1] 9431    6

sum(limma_pvals$adj.P.Val < 0.05)

## [1] 223

genes <- limma_pvals$adj.P.Val < 0.05

DE <- featureNames(fdata_bot)[genes]
```

```
length(DE)
```

```
## [1] 223
```

Gene Set analysis using goseq package

using database of 'Gene Ontology Consortium' to do gene set analysis

```
genes = as.integer(limma_pvals$adj.P.Val < 0.05)
not_na = !is.na(genes)
names(genes) = rownames(edata)
##names(genes)

genes = genes[not_na]

head(supportedGenomes(),n=12)[,1:4]
```

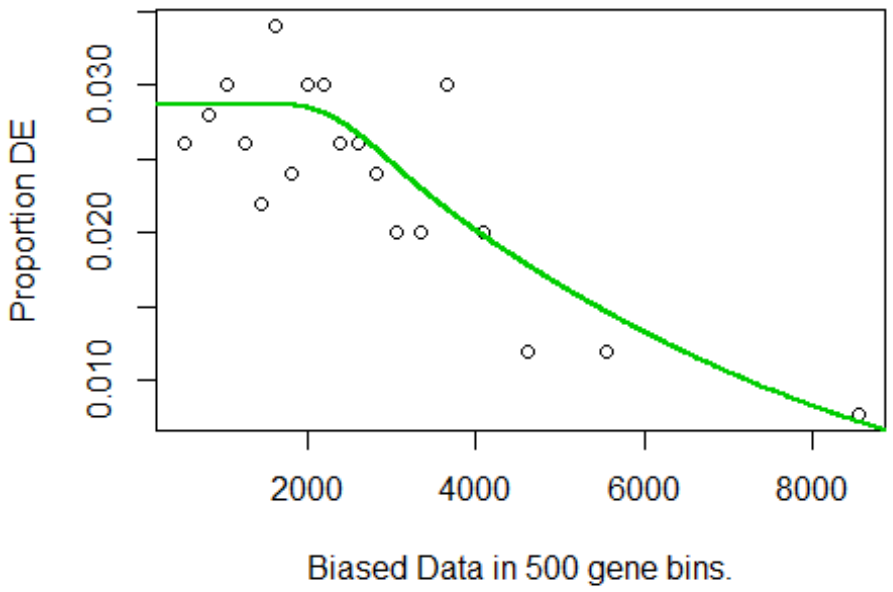
```
##         db   species      date
name
## 1     hg38     Human Dec. 2013       Genome Reference Consortium
GRCh38
## 2     hg19     Human Feb. 2009       Genome Reference Consortium
GRCh37
## 3     hg18     Human Mar. 2006                          NCBI Build
36.1
## 4     hg17     Human  May 2004                            NCBI Build
35
## 5     hg16     Human Jul. 2003                            NCBI Build
34
## 6  vicPac2    Alpaca Mar. 2013      Broad Institute Vicugna_pacos-
2.0.1
## 7  vicPac1    Alpaca Jul. 2008                Broad Institute
VicPac1.0
## 8  dasNov3 Armadillo Dec. 2011                  Broad Institute
DasNov3
## 9  otoGar3  Bushbaby Mar. 2011                  Broad Institute
OtoGar3
## 10 papHam1    Baboon Nov. 2008 Baylor College of Medicine HGSC
Pham_1.0
## 11 papAnu2    Baboon Mar. 2012      Baylor College of Medicine
Panu_2.0
## 12 felCat5       Cat Sep. 2011                    ICGSC Felis_catus-
6.2
```

```
??nullp
```

```
## starting httpd help server ... done
```

```
pwf=nullp(genes,"mm9","ensGene")
```

```
## Loading mm9 length data...
```




```
head(pwf)

##                     DEgenes bias.data        pwf
## ENSMUSG00000000001       0      3213 0.02367823
## ENSMUSG00000000056       0      4405 0.01859498
## ENSMUSG00000000058       0       976 0.02871612
## ENSMUSG00000000078       0      4221 0.01929329
## ENSMUSG00000000088       0       628 0.02871612
## ENSMUSG00000000093       0      3569 0.02199846

GO.MF=goseq(pwf,"mm9","ensGene",test.cats=c("GO:MF"))

## Fetching GO annotations...
## Loading required package: AnnotationDbi
## Loading required package: stats4
## Loading required package: IRanges
## Loading required package: S4Vectors
##
## For 526 genes, we could not find any categories. These genes will be
excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see
documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
## 'select()' returned 1:1 mapping between keys and columns
```

```
head(GO.MF)
```

```
##        category over_represented_pvalue under_represented_pvalue
## 753  GO:0004888            2.248999e-06                0.9999994
## 770  GO:0004930            9.539111e-06                0.9999978
## 744  GO:0004872            1.113570e-05                0.9999964
## 1226 GO:0008528            1.183815e-05                0.9999986
## 156  GO:0001653            1.545060e-05                0.9999981
## 2755 GO:0038023            1.552570e-05                0.9999952
##      numDEInCat numInCat                                      term
## 753          23      385   transmembrane signaling receptor activity
## 770          16      211         G-protein coupled receptor activity
## 744          25      482                             receptor activity
## 1226          9       65 G-protein coupled peptide receptor activity
## 156           9       67                   peptide receptor activity
## 2755         23      432                signaling receptor activity
##      ontology
## 753        MF
## 770        MF
## 744        MF
## 1226       MF
## 156        MF
## 2755       MF
```

```
GO.MF[1:10,]
```

```
##        category over_represented_pvalue under_represented_pvalue
## 753  GO:0004888            2.248999e-06                0.9999994
## 770  GO:0004930            9.539111e-06                0.9999978
## 744  GO:0004872            1.113570e-05                0.9999964
## 1226 GO:0008528            1.183815e-05                0.9999986
## 156  GO:0001653            1.545060e-05                0.9999981
## 2755 GO:0038023            1.552570e-05                0.9999952
## 3505 GO:0060089            7.464065e-05                0.9999717
## 743  GO:0004871            1.185824e-04                0.9999555
## 2436 GO:0033040            2.244481e-04                1.0000000
## 781  GO:0004944            4.757635e-04                1.0000000
##      numDEInCat numInCat                                      term
## 753          23      385   transmembrane signaling receptor activity
## 770          16      211         G-protein coupled receptor activity
## 744          25      482                             receptor activity
## 1226          9       65 G-protein coupled peptide receptor activity
## 156           9       67                   peptide receptor activity
## 2755         23      432                signaling receptor activity
## 3505         27      604             molecular transducer activity
## 743          25      555                signal transducer activity
## 2436          2        2              sour taste receptor activity
## 781           2        2         C5a anaphylatoxin receptor activity
##      ontology
## 753        MF
## 770        MF
```

```
## 744        MF
## 1226       MF
## 156        MF
## 2755       MF
## 3505       MF
## 743        MF
## 2436       MF
## 781        MF
```

```
devtools::session_info()
```

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

##  setting  value
##  version  R version 3.2.2 (2015-08-14)
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  English_United States.1252
##  tz       Asia/Seoul
##  date     2015-12-18

## Packages
-------------------------------------------------------------------

##  package              * version  date       source
##  annotate               1.48.0   2015-10-14 Bioconductor
##  AnnotationDbi        * 1.32.2   2015-12-09 Bioconductor
##  BiasedUrn            * 1.06.1   2013-12-29 CRAN (R 3.2.2)
##  Biobase              * 2.30.0   2015-10-14 Bioconductor
##  BiocGenerics         * 0.16.1   2015-11-06 Bioconductor
##  BiocParallel           1.4.3    2015-12-18 Bioconductor
##  biomaRt                2.26.1   2015-11-23 Bioconductor
##  Biostrings             2.38.2   2015-11-21 Bioconductor
##  bitops                 1.0-6    2013-08-17 CRAN (R 3.2.2)
##  DBI                  * 0.3.1    2014-09-24 CRAN (R 3.2.2)
##  devtools             * 1.9.1    2015-09-11 CRAN (R 3.2.2)
##  digest                 0.6.8    2014-12-31 CRAN (R 3.2.2)
##  evaluate               0.8      2015-09-18 CRAN (R 3.2.2)
##  formatR                1.2.1    2015-09-18 CRAN (R 3.2.2)
##  futile.logger          1.4.1    2015-04-20 CRAN (R 3.2.2)
##  futile.options         1.0.0    2010-04-06 CRAN (R 3.2.2)
##  genefilter           * 1.52.0   2015-10-14 Bioconductor
##  geneLenDataBase      * 1.6.0    2015-10-27 Bioconductor
##  GenomeInfoDb           1.6.1    2015-11-03 Bioconductor
##  GenomicAlignments      1.6.1    2015-10-22 Bioconductor
##  GenomicFeatures        1.22.7   2015-12-18 Bioconductor
##  GenomicRanges          1.22.2   2015-12-12 Bioconductor
##  GO.db                  3.2.2    2015-11-18 Bioconductor
##  goseq                * 1.22.0   2015-10-14 Bioconductor
```

```
##  htmltools              0.2.6    2014-09-08 CRAN (R 3.2.2)
##  IRanges              * 2.4.6    2015-12-12 Bioconductor
##  knitr                  1.11     2015-08-14 CRAN (R 3.2.2)
##  lambda.r               1.1.7    2015-03-20 CRAN (R 3.2.2)
##  lattice                0.20-33  2015-07-14 CRAN (R 3.2.2)
##  limma                * 3.26.3   2015-11-16 Bioconductor
##  magrittr               1.5      2014-11-22 CRAN (R 3.2.2)
##  Matrix                 1.2-3    2015-11-28 CRAN (R 3.2.2)
##  memoise                0.2.1    2014-04-22 CRAN (R 3.2.2)
##  mgcv                   1.8-10   2015-12-12 CRAN (R 3.2.3)
##  nlme                   3.1-122  2015-08-19 CRAN (R 3.2.2)
##  org.Mm.eg.db         * 3.2.3    2015-11-24 Bioconductor
##  RCurl                  1.95-4.7 2015-06-30 CRAN (R 3.2.2)
##  rmarkdown              0.8.1    2015-10-10 CRAN (R 3.2.2)
##  Rsamtools              1.22.0   2015-10-14 Bioconductor
##  RSQLite              * 1.0.0    2014-10-25 CRAN (R 3.2.2)
##  rtracklayer            1.30.1   2015-10-22 Bioconductor
##  S4Vectors            * 0.8.5    2015-12-11 Bioconductor
##  stringi                1.0-1    2015-10-22 CRAN (R 3.2.2)
##  stringr                1.0.0    2015-04-30 CRAN (R 3.2.2)
##  SummarizedExperiment   1.0.1    2015-11-06 Bioconductor
##  survival               2.38-3   2015-07-02 CRAN (R 3.2.2)
##  XML                    3.98-1.3 2015-06-30 CRAN (R 3.2.2)
##  xtable                 1.8-0    2015-11-02 CRAN (R 3.2.2)
##  XVector                0.10.0   2015-10-14 Bioconductor
##  yaml                   2.1.13   2014-06-12 CRAN (R 3.2.2)
##  zlibbioc               1.16.0   2015-10-14 Bioconductor
```