recount (gene and exon analyses)

Kai Kammers and Shannon Ellis

July 22, 2016

Contents

Included is an example of how to download and analyze expression data from SRA study SRP032798. The data come from human breast cancer samples, and we compare the transcriptomes of TNBC samples (triple negative breast cancer) and HER2-positive breast cancer samples (breast cancer type that tests positive for a protein called human epidermal growth factor receptor 2). Code here demonstrates how to carry out differential expression analyses on gene, exon, junction, and differential expressed region (DER) levels within a single study using `limma` and `voom`. We test for concordance among the results of each analysis and demonstrate how to carry out gene ontology analysis using `topGO` to characterize top hits from differential expression analyses.

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')
```

Gene level analysis

We first download the project of interest (SRP032798), obtaining expression data for the study of interest. We obtain summaries of the number of samples and genes included using `colData()` and `rowData()`, respectively.

```
## Find the project of interest (SRP032789), e.g. with parts of the abstract
project_info <- abstract_search('To define the digital transcriptome of three breast cancer')

## Explore information
project_info
```

```
##     number_samples species
## 865             20   human
##
## 865 Goal: To define the digital transcriptome of three breast cancer subtypes (TNBC, Non-TNBC, and H
##       project
## 865 SRP032789
```

```
## Browse the project at SRA
browse_study(project_info$project)

## Download the gene level RangedSummarizedExperiment data
if(!file.exists(file.path('SRP032789', 'rse_gene.Rdata'))) {
    download_study(project_info$project)
}

## Load the data
load(file.path(project_info$project, 'rse_gene.Rdata'))
rse_gene
```

```
## class: RangedSummarizedExperiment
## dim: 23779 20
## metadata(0):
## assays(1): counts
## rownames(23779): 1 10 ... 9994 9997
## rowData names(3): gene_id bp_length symbol
## colnames(20): SRR1027171 SRR1027173 ... SRR1027190 SRR1027172
## colData names(21): project sample ... title characteristics
```

```
## This is the phenotype data provided by the recount project
colData(rse_gene)
```

```
## DataFrame with 20 rows and 21 columns
##                project      sample  experiment         run
##            <character> <character> <character> <character>
## SRR1027171   SRP032789   SRS500214   SRX374850  SRR1027171
## SRR1027173   SRP032789   SRS500216   SRX374852  SRR1027173
## SRR1027174   SRP032789   SRS500217   SRX374853  SRR1027174
## SRR1027175   SRP032789   SRS500218   SRX374854  SRR1027175
## SRR1027176   SRP032789   SRS500219   SRX374855  SRR1027176
## ...                ...         ...         ...         ...
## SRR1027187   SRP032789   SRS500230   SRX374866  SRR1027187
## SRR1027188   SRP032789   SRS500231   SRX374867  SRR1027188
## SRR1027189   SRP032789   SRS500232   SRX374868  SRR1027189
## SRR1027190   SRP032789   SRS500233   SRX374869  SRR1027190
## SRR1027172   SRP032789   SRS500215   SRX374851  SRR1027172
##            read_count_as_reported_by_sra reads_downloaded
```

```
##                               <integer>        <integer>
## SRR1027171                     88869444         88869444
## SRR1027173                    107812596        107812596
## SRR1027174                     98563260         98563260
## SRR1027175                     91327892         91327892
## SRR1027176                     96513572         96513572
## ...                                 ...              ...
## SRR1027187                     75260678         75260678
## SRR1027188                     65709192         65709192
## SRR1027189                     65801392         65801392
## SRR1027190                     74356276         74356276
## SRR1027172                     80986440         58902122
##            proportion_of_reads_reported_by_sra_downloaded paired_end
##                                                 <numeric>  <logical>
## SRR1027171                                              1       TRUE
## SRR1027173                                              1       TRUE
## SRR1027174                                              1       TRUE
## SRR1027175                                              1       TRUE
## SRR1027176                                              1       TRUE
## ...                                                   ...        ...
## SRR1027187                                      1.0000000       TRUE
## SRR1027188                                      1.0000000       TRUE
## SRR1027189                                      1.0000000       TRUE
## SRR1027190                                      1.0000000       TRUE
## SRR1027172                                      0.7273084       TRUE
##            sra_misreported_paired_end mapped_read_count        auc
##                             <logical>         <integer>  <numeric>
## SRR1027171                       FALSE          86949307 5082692127
## SRR1027173                       FALSE         104337779 6077034329
## SRR1027174                       FALSE          95271238 5504462845
## SRR1027175                       FALSE          88820239 5150234117
## SRR1027176                       FALSE          93464650 5416681912
## ...                                ...               ...        ...
## SRR1027187                       FALSE          64697612 3567078255
## SRR1027188                       FALSE          65278500 4856453823
## SRR1027189                       FALSE          65328289 4858587600
## SRR1027190                       FALSE          73911898 5501089036
## SRR1027172                       FALSE          57523391 3351013968
##            sharq_beta_tissue sharq_beta_cell_type
##                  <character>          <character>
## SRR1027171           breast                  esc
## SRR1027173           breast                  esc
## SRR1027174           breast                  esc
## SRR1027175           breast                  esc
## SRR1027176           breast                  esc
## ...                     ...                  ...
## SRR1027187           breast                  esc
## SRR1027188           breast                  esc
## SRR1027189           breast                  esc
## SRR1027190           breast                  esc
## SRR1027172           breast                  esc
##            biosample_submission_date biosample_publication_date
##                          <character>                <character>
## SRR1027171   2013-11-07T12:40:22.203    2013-11-08T01:11:17.160
```

```
## SRR1027173   2013-11-07T12:40:32.283    2013-11-08T01:11:14.827
## SRR1027174   2013-11-07T12:40:28.283    2013-11-08T01:11:52.283
## SRR1027175   2013-11-07T12:40:34.343    2013-11-08T01:11:15.963
## SRR1027176   2013-11-07T12:40:36.303    2013-11-08T01:11:46.430
## ...                              ...                        ...
## SRR1027187   2013-11-07T12:40:56.180    2013-11-08T01:11:29.587
## SRR1027188   2013-11-07T12:40:58.170    2013-11-08T01:12:06.660
## SRR1027189   2013-11-07T12:40:20.227    2013-11-08T01:11:33.080
## SRR1027190   2013-11-07T12:40:18.090    2013-11-08T01:12:11.320
## SRR1027172   2013-11-07T12:40:26.217    2013-11-08T01:11:45.250
##              biosample_update_date avg_read_length geo_accession
##                        <character>       <integer>   <character>
## SRR1027171 2014-03-07T16:09:38.542             120    GSM1261016
## SRR1027173 2014-03-07T16:09:38.698             120    GSM1261018
## SRR1027174 2014-03-07T16:09:38.637             120    GSM1261019
## SRR1027175 2014-03-07T16:09:38.731             120    GSM1261020
## SRR1027176 2014-03-07T16:09:38.768             120    GSM1261021
## ...                              ...             ...           ...
## SRR1027187 2014-03-07T16:09:39.093             120    GSM1261032
## SRR1027188 2014-03-07T16:09:39.130             150    GSM1261033
## SRR1027189 2014-03-07T16:09:38.498             150    GSM1261034
## SRR1027190 2014-03-07T16:09:38.469             150    GSM1261035
## SRR1027172 2014-03-07T16:09:38.604              87    GSM1261017
##              bigwig_file       title
##              <character> <character>
## SRR1027171 SRR1027171.bw       TNBC1
## SRR1027173 SRR1027173.bw       TNBC3
## SRR1027174 SRR1027174.bw       TNBC4
## SRR1027175 SRR1027175.bw       TNBC5
## SRR1027176 SRR1027176.bw       TNBC6
## ...                  ...         ...
## SRR1027187 SRR1027187.bw      HER2-5
## SRR1027188 SRR1027188.bw        NBS1
## SRR1027189 SRR1027189.bw        NBS2
## SRR1027190 SRR1027190.bw        NBS3
## SRR1027172 SRR1027172.bw       TNBC2
##                                   characteristics
##                                   <CharacterList>
## SRR1027171          tumor type: TNBC Breast Tumor
## SRR1027173          tumor type: TNBC Breast Tumor
## SRR1027174          tumor type: TNBC Breast Tumor
## SRR1027175          tumor type: TNBC Breast Tumor
## SRR1027176          tumor type: TNBC Breast Tumor
## ...                                           ...
## SRR1027187 tumor type: HER2 Positive Breast Tumor
## SRR1027188    tumor type: Normal Breast Organoids
## SRR1027189    tumor type: Normal Breast Organoids
## SRR1027190    tumor type: Normal Breast Organoids
## SRR1027172          tumor type: TNBC Breast Tumor
```

```
## At the gene level, the row data includes the names of the genes and
## the sum of the reduced exons widths, which can be used for taking into
## account the gene length.
rowData(rse_gene)
```

```
## DataFrame with 23779 rows and 3 columns
##           gene_id bp_length      symbol
##       <character> <integer> <character>
## 1               1      4027        A1BG
## 2              10      1317        NAT2
## 3             100      1532         ADA
## 4            1000      4473        CDH2
## 5       100008589      5071     RNA28S5
## ...           ...       ...         ...
## 23775        9991      8234       PTBP3
## 23776        9992       803       KCNE2
## 23777        9993      4882       DGCR2
## 23778        9994      6763    CASP8AP2
## 23779        9997      1393        SCO2
```

Downloaded count data are first scaled to take into account differing coverage between samples. Phenotype data (`pheno`) are obtained and ordered to match the sample order of the gene expression data (`rse_gene`). Only those samples that are HER2-positive or TNBC are included for analysis. Prior to differential gene expression analysis, count data are obtained in matrix format and then filtered to only include those genes with greater than five average normalized counts across all samples.

```
## Scale counts by taking into account the total coverage per sample
rse <- scale_counts(rse_gene)

## Download additional 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)

## Obtain correct order for pheno data
pheno <- pheno[match(rse$run, pheno$Run_s), ]
identical(pheno$Run_s, rse$run)
```

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

```
head(cbind(pheno$Run_s, rse$run))
```

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

```
## Obtain grouping information
colData(rse)$group <- pheno$tumor_type_s
table(colData(rse)$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 types
rse <- rse[, rse$group %in% c('HER2 Positive Breast Tumor',
                              'TNBC Breast Tumor')]

## Save filtered rse object
rse_gene_filt <- rse

## Obtain count matrix
counts <- assays(rse_gene_filt)$counts

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

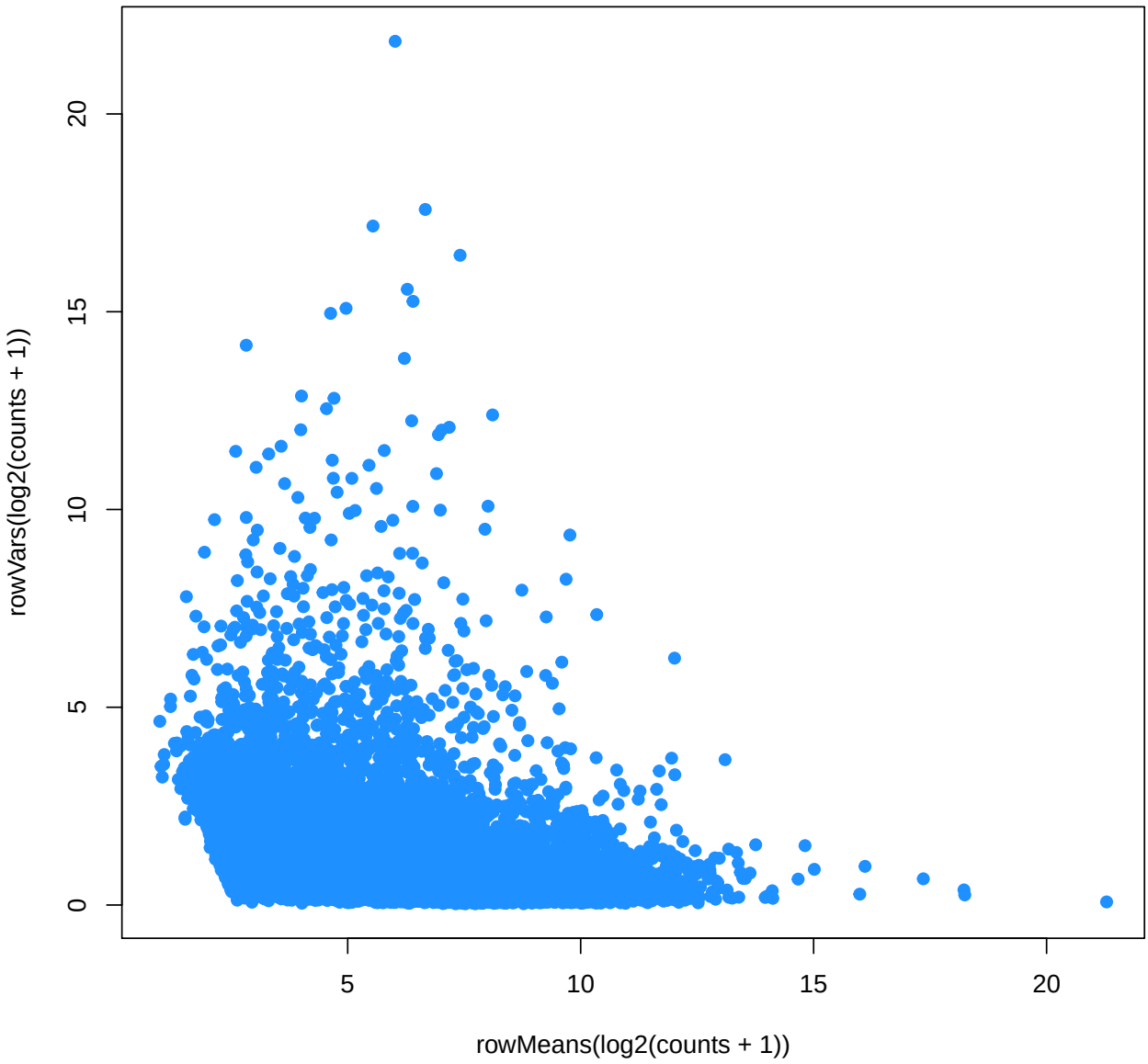
```
## [1] 17874    11
```

```
## Save for gene, exon and junction comparisons
counts_gene <- counts
counts_gene[1:5, 1:5]
```

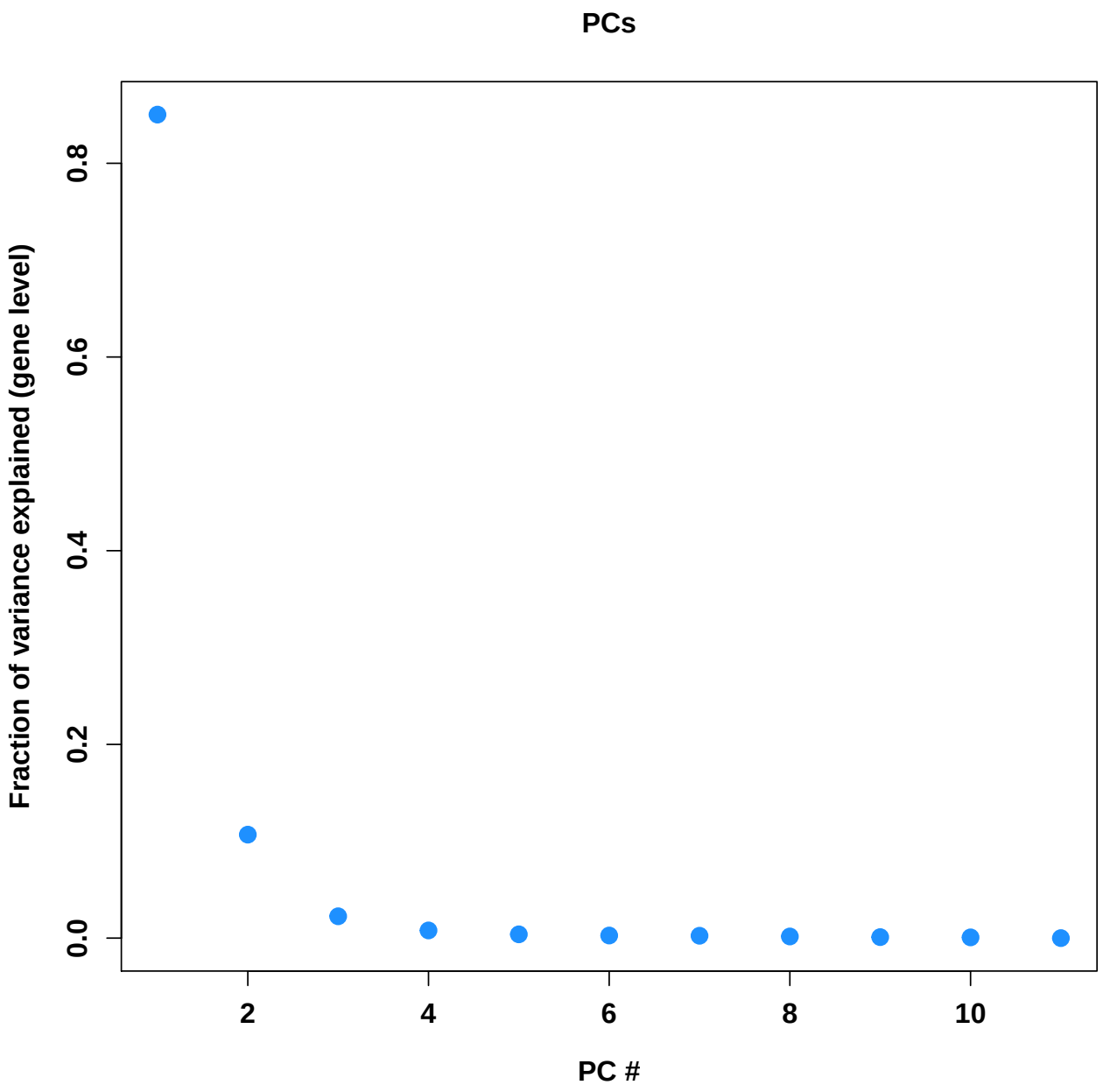
```
##           SRR1027171 SRR1027173 SRR1027174 SRR1027175 SRR1027176
## 1                 66         46         42         64         65
## 100               97         22         73         90        109
## 1000              29        183        450        160        153
## 100008589    1842832    2085085    3122352    3082593    2862316
## 100009676         92         50        123        119         99
```

To get a better sense of the data, we plot the mean-variance relationship for each gene. Similarly, we run principal component analysis (PCA) to identify any sample outliers within the data. We assess the variance explained by each of the first 11 PCs as well as visualize the relationship of each sample in the first two PCs.

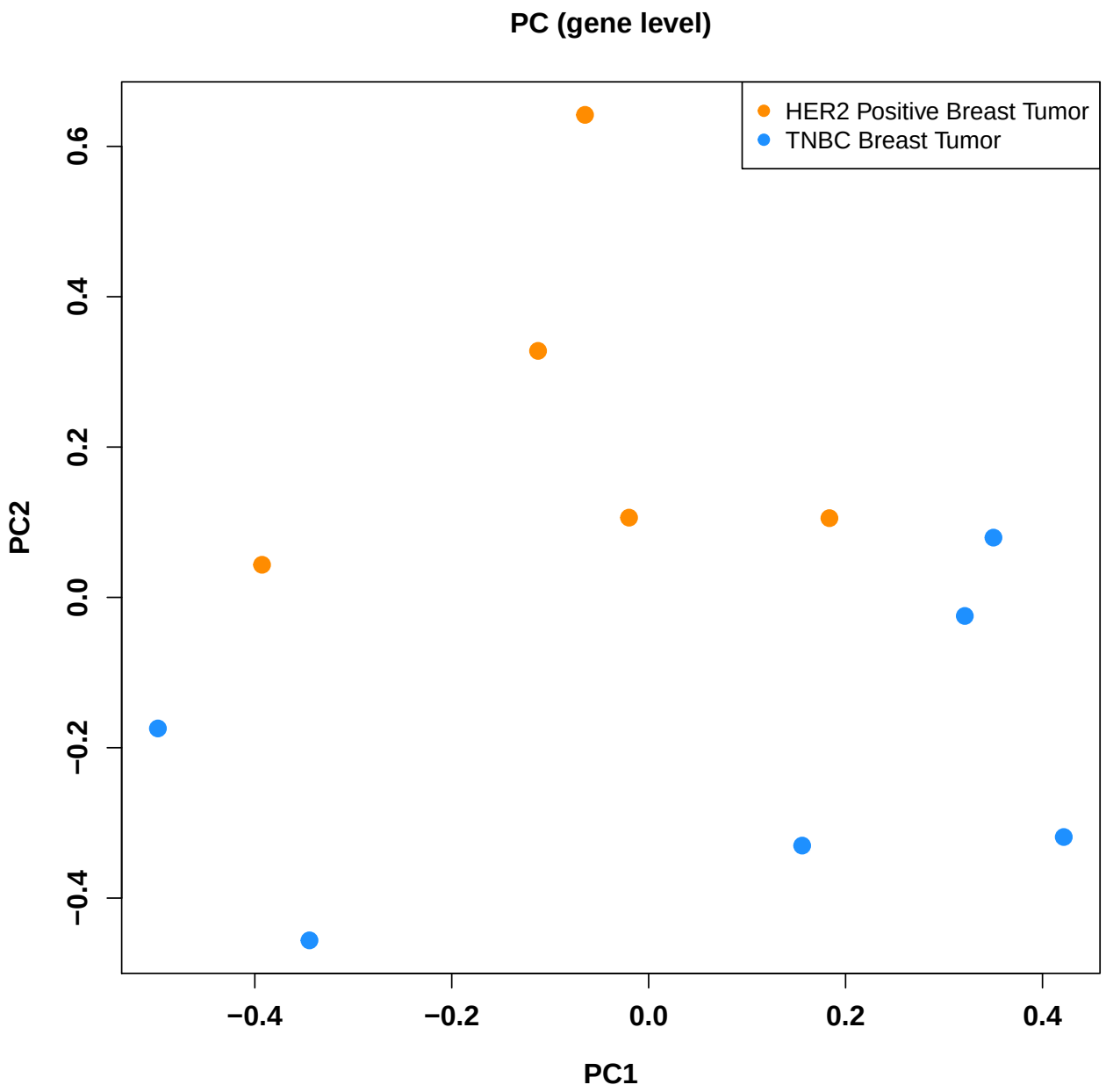
```
## Set colors
trop <- RSkittleBrewer('tropical')[c(1, 2)]
cols <- as.numeric(as.factor(rse$group))

## Look at mean variance relationship
plot(rowMeans(log2(counts + 1)), rowVars(log2(counts + 1)),
     pch = 19, col = trop[2])
```




```
## Calculate PCs with svd function
expr.pca <- svd(counts - rowMeans(counts))

## Plot PCs
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(expr.pca$d^2 / sum(expr.pca$d^2), pch = 19, col = trop[2], cex = 1.5,
     ylab = 'Fraction of variance explained (gene level)', xlab = 'PC #',
     main = 'PCs')
```




```
## Plot PC1 vs. PC2
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(expr.pca$v[, 1], expr.pca$v[, 2], pch = 19, col = trop[cols], cex = 1.5,
     xlab = 'PC1', ylab = 'PC2',
     main = 'PC (gene level)')
legend('topright', pch = 19, col = trop[c(1, 2)],
       names(summary(as.factor(rse$group))))
```



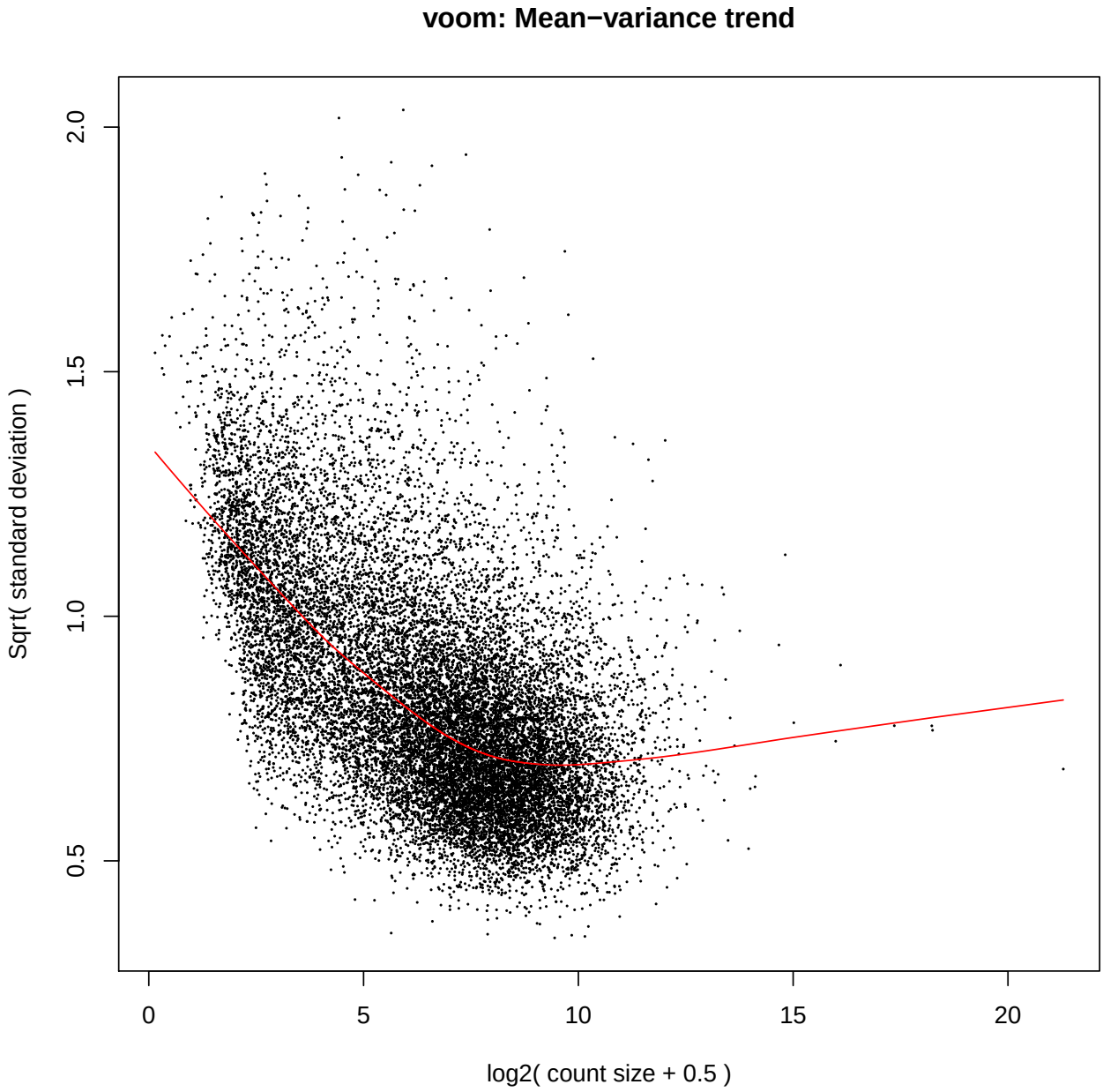
Having determined there are no sample outliers in these data, we carry out differential gene expression analysis. Differential gene expression between TNBC and HER2-positive samples are determined using `limma` and `voom`. Differentially expressed genes are visualized using a volcano plot to compare the effect size of the differential expression [as measured by the $log_2(fold - change)$ in expression] and its significance [$-log_10(p - value)$].

```
## Perform differential expression analysis with limma-voom
design <- model.matrix(~ rse$group)
design
```

```
##    (Intercept) rse$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")$`rse$group`
## [1] "contr.treatment"
```

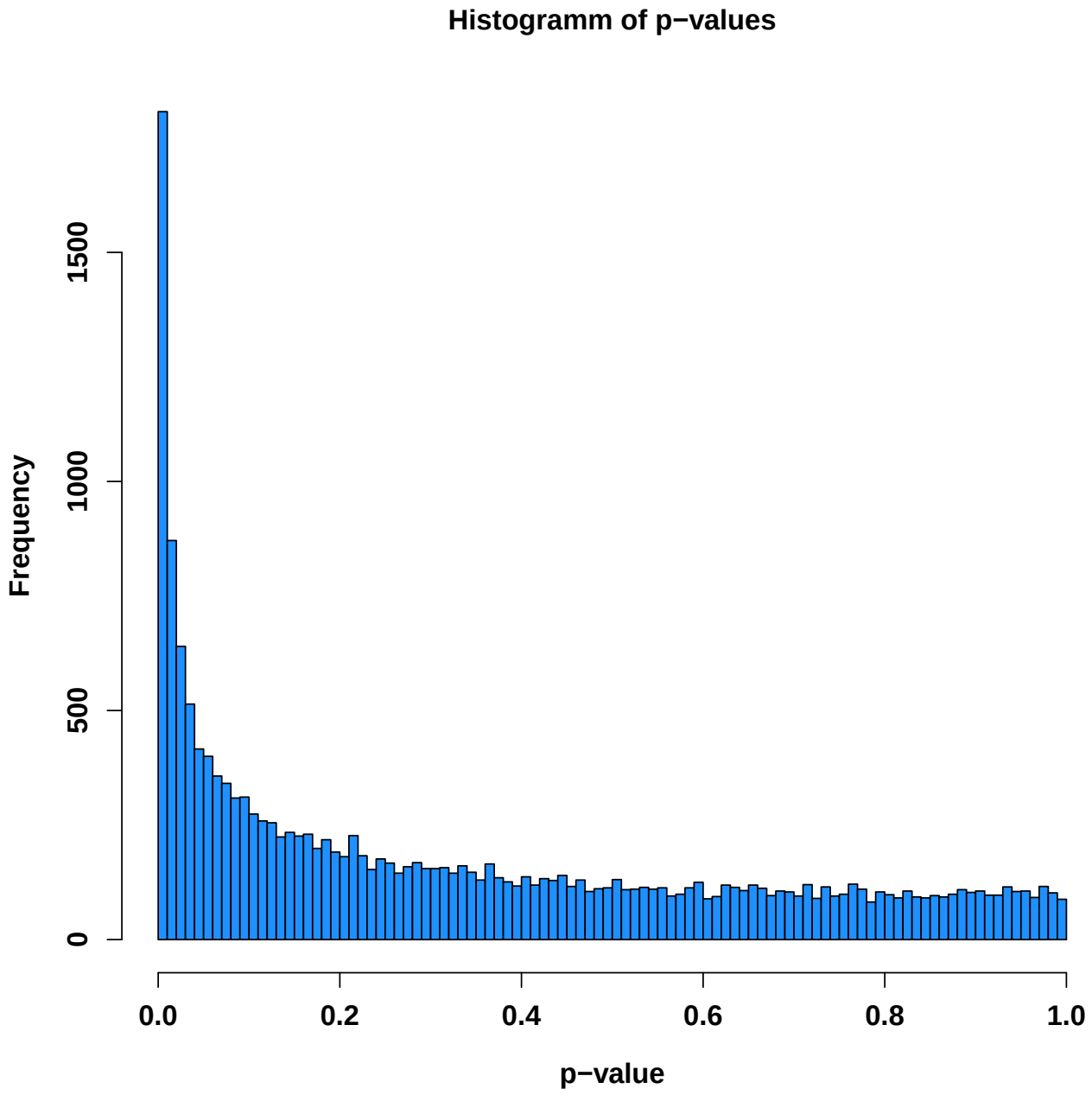
```
dge <- DGEList(counts = counts)
dge <- calcNormFactors(dge)
v <- voom(dge, design, plot = TRUE)
```




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

## Determine the number of genes differentially expressed at q<0.05
sum(res_gene$q.mod < 0.05)
```

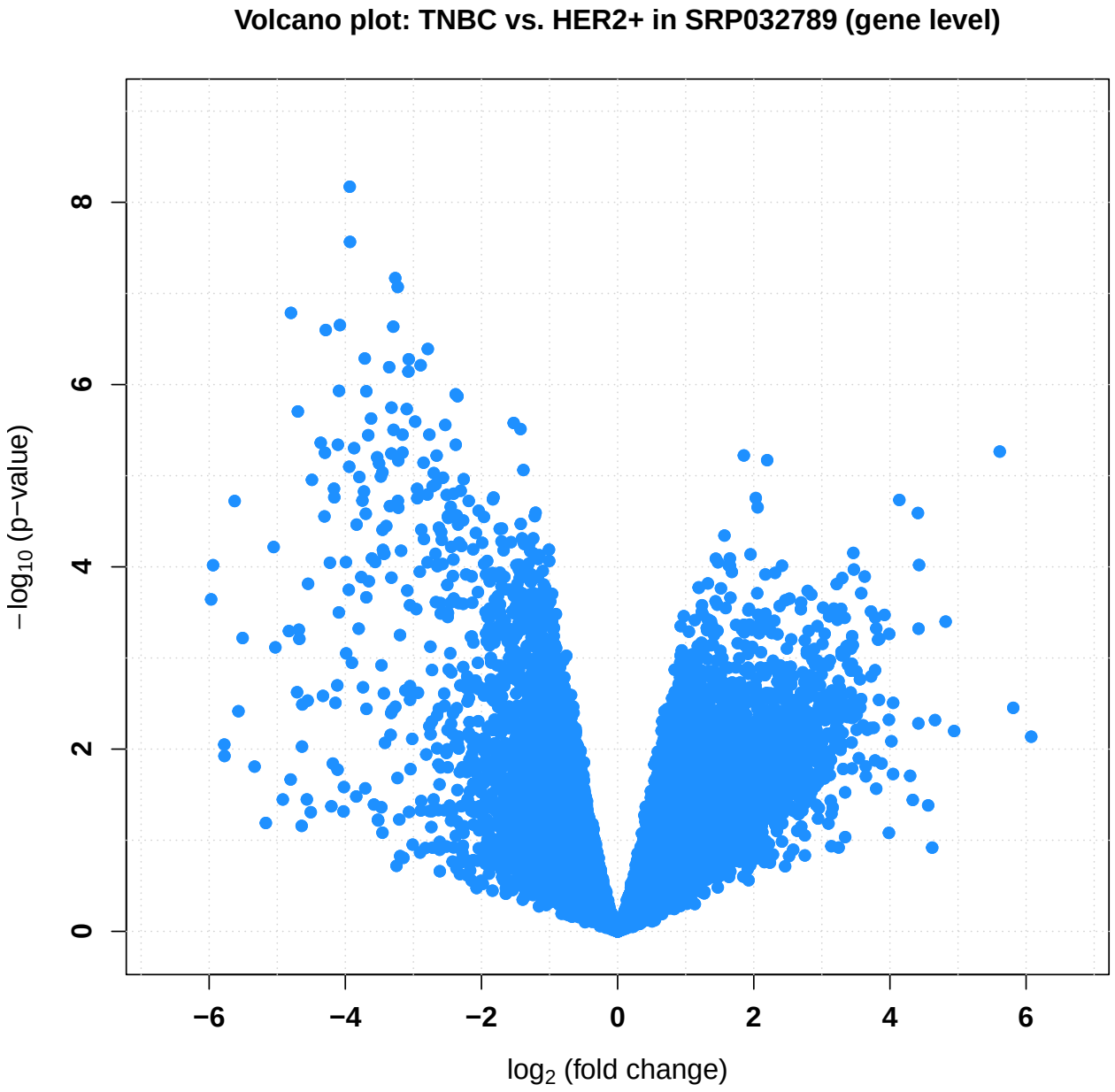
```
## [1] 1611
```

```
## Histogram of p-values
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
hist(p.mod, col = trop[2], xlab = 'p-value',
     main = 'Histogramm of p-values', breaks = 100)
```




```
## 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 (gene level)')
```



Gene set enrichment analysis

To get a better understanding of those genes showing differential gene expression, we utilize `topGO`, a gene set analysis library. Genes included in this analysis are those reaching a q-value cutoff less than 0.05.

```
names(q.mod) <- rownames(counts)
interesting <- function(x) x < 0.05
```

After determining which genes to include for analysis, `topGO` objects are generated and the enrichment tests are run. The Kolomogorov-Smirnov (`ks`) test is used to test for distributional differences. Here, we ask whether each GO group is "enriched" for differentially expressed (`q.mod < 0.05`) genes. Equivalently, we are testing whether the p-value distributions are the same for genes in and outside of each gene ontology. We run tests on the "biological processes" ontology.

```
topgoobjBP <- new('topGOdata',
    description = 'biological process',
    ontology = 'BP', allGenes = q.mod, geneSelectionFun = interesting,
    annotationFun = annFUN.org, mapping = 'org.Hs.eg.db', ID = 'entrez')
```

```
##
## Building most specific GOs .....

##  ( 10647 GO terms found. )

##
## Build GO DAG topology ..........

##  ( 14443 GO terms and 34755 relations. )

##
## Annotating nodes ...............

##  ( 13762 genes annotated to the GO terms. )
```

```
bptest <- runTest(topgoobjBP, algorithm = 'weight01', statistic = 'ks')
```

```
##
##           -- Weight01 Algorithm --
##
##       the algorithm is scoring 14443 nontrivial nodes
##       parameters:
##           test statistic: ks
##           score order: increasing

##
##   Level 20:  1 nodes to be scored    (0 eliminated genes)

##
##   Level 19:  7 nodes to be scored    (0 eliminated genes)

##
##   Level 18:  20 nodes to be scored   (1 eliminated genes)

##
##   Level 17:  40 nodes to be scored   (30 eliminated genes)

##
##   Level 16:  130 nodes to be scored  (91 eliminated genes)
```

```
##
##   Level 15:  266 nodes to be scored  (179 eliminated genes)

##
##   Level 14:  520 nodes to be scored  (552 eliminated genes)

##
##   Level 13:  925 nodes to be scored  (1159 eliminated genes)

##
##   Level 12:  1336 nodes to be scored (2451 eliminated genes)

##
##   Level 11:  1602 nodes to be scored (4326 eliminated genes)

##
##   Level 10:  1878 nodes to be scored (6122 eliminated genes)

##
##   Level 9:   1967 nodes to be scored (8287 eliminated genes)

##
##   Level 8:   1855 nodes to be scored (9989 eliminated genes)

##
##   Level 7:   1641 nodes to be scored (11002 eliminated genes)

##
##   Level 6:   1187 nodes to be scored (11986 eliminated genes)

##
##   Level 5:   679 nodes to be scored  (12493 eliminated genes)

##
##   Level 4:   293 nodes to be scored  (13053 eliminated genes)

##
##   Level 3:   74 nodes to be scored   (13234 eliminated genes)

##
##   Level 2:   21 nodes to be scored   (13397 eliminated genes)

##
##   Level 1:   1 nodes to be scored    (13475 eliminated genes)
```

```
bptest
```

```
##
## Description: biological process
## Ontology: BP
## 'weight01' algorithm with the 'ks' test
## 14443 GO terms scored: 50 terms with p < 0.01
## Annotation data:
##     Annotated genes: 13762
##     Significant genes: 1131
##     Min. no. of genes annotated to a GO: 1
##     Nontrivial nodes: 14443
```

```
bpres_gene <- GenTable(topgoobjBP, pval = bptest,
                       topNodes = length(bptest@score), numChar = 100)
head(bpres_gene, n = 10)
```

```
##         GO.ID
## 1  GO:0008589
## 2  GO:0016579
## 3  GO:0030049
## 4  GO:0007050
## 5  GO:0006355
## 6  GO:0070933
## 7  GO:0071557
## 8  GO:0010606
## 9  GO:0001816
## 10 GO:0050673
##                                                                    Term
## 1                        regulation of smoothened signaling pathway
## 2                                     protein deubiquitination
## 3                                      muscle filament sliding
## 4                                            cell cycle arrest
## 5                        regulation of transcription, DNA-templated
## 6                                     histone H4 deacetylation
## 7                                   histone H3-K27 demethylation
## 8  positive regulation of cytoplasmic mRNA processing body assembly
## 9                                           cytokine production
## 10                                 epithelial cell proliferation
##    Annotated Significant Expected    pval
## 1         61           9     5.01 0.00012
## 2        101          14     8.30 0.00032
## 3         30           7     2.47 0.00042
## 4        237          26    19.48 0.00066
## 5       3062         298   251.64 0.00083
## 6         10           0     0.82 0.00084
## 7          4           3     0.33 0.00125
## 8          6           3     0.49 0.00212
## 9        543          25    44.63 0.00214
## 10       318          35    26.13 0.00237
```

Exon level analysis

As above, we are interested here in differential expression. However, rather than summarizing across genes, this analysis will look for differential expression at the exon level. In this analysis, we include all exons that

map to the previous filtered genes and again carry out differential expression analysis using `limma` and `voom`.

Here, we download data from the same project as above (SRP032798); however, this time, we are interested in obtaining the exon level data.

```
## Find a project of interest (SRP032789)
project_info <- abstract_search('To define the digital transcriptome of three breast cancer')
project_info
```

```
##     number_samples species
## 865             20   human
##
## 865 Goal: To define the digital transcriptome of three breast cancer subtypes (TNBC, Non-TNBC, and H
##       project
## 865 SRP032789
```

```
## Browse the project at SRA
browse_study(project_info$project)

## Download the exon level RangedSummarizedExperiment data
if(!file.exists(file.path('SRP032789', 'rse_exon.Rdata'))) {
    download_study(project_info$project, type = 'rse-exon')
}

## Load the data
load(file.path(project_info$project, 'rse_exon.Rdata'))
rse_exon
```

```
## class: RangedSummarizedExperiment
## dim: 226117 20
## metadata(0):
## assays(1): counts
## rownames(226117): 1 1 ... 9997 9997
## rowData names(0):
## colnames(20): SRR1027171 SRR1027173 ... SRR1027190 SRR1027172
## colData names(21): project sample ... title characteristics
```

```
## This is the sample phenotype data provided by the recount project
colData(rse_exon)
```

```
## DataFrame with 20 rows and 21 columns
##                project      sample  experiment         run
##            <character> <character> <character> <character>
## SRR1027171   SRP032789   SRS500214   SRX374850  SRR1027171
## SRR1027173   SRP032789   SRS500216   SRX374852  SRR1027173
## SRR1027174   SRP032789   SRS500217   SRX374853  SRR1027174
## SRR1027175   SRP032789   SRS500218   SRX374854  SRR1027175
## SRR1027176   SRP032789   SRS500219   SRX374855  SRR1027176
## ...                ...         ...         ...         ...
## SRR1027187   SRP032789   SRS500230   SRX374866  SRR1027187
## SRR1027188   SRP032789   SRS500231   SRX374867  SRR1027188
## SRR1027189   SRP032789   SRS500232   SRX374868  SRR1027189
## SRR1027190   SRP032789   SRS500233   SRX374869  SRR1027190
```

```
## SRR1027172   SRP032789   SRS500215   SRX374851  SRR1027172
##            read_count_as_reported_by_sra reads_downloaded
##                                <integer>        <integer>
## SRR1027171                      88869444         88869444
## SRR1027173                     107812596        107812596
## SRR1027174                      98563260         98563260
## SRR1027175                      91327892         91327892
## SRR1027176                      96513572         96513572
## ...                                  ...              ...
## SRR1027187                      75260678         75260678
## SRR1027188                      65709192         65709192
## SRR1027189                      65801392         65801392
## SRR1027190                      74356276         74356276
## SRR1027172                      80986440         58902122
##            proportion_of_reads_reported_by_sra_downloaded paired_end
##                                                 <numeric>  <logical>
## SRR1027171                                              1       TRUE
## SRR1027173                                              1       TRUE
## SRR1027174                                              1       TRUE
## SRR1027175                                              1       TRUE
## SRR1027176                                              1       TRUE
## ...                                                   ...        ...
## SRR1027187                                      1.0000000       TRUE
## SRR1027188                                      1.0000000       TRUE
## SRR1027189                                      1.0000000       TRUE
## SRR1027190                                      1.0000000       TRUE
## SRR1027172                                      0.7273084       TRUE
##            sra_misreported_paired_end mapped_read_count        auc
##                             <logical>         <integer>  <numeric>
## SRR1027171                       FALSE          86949307 5082692127
## SRR1027173                       FALSE         104337779 6077034329
## SRR1027174                       FALSE          95271238 5504462845
## SRR1027175                       FALSE          88820239 5150234117
## SRR1027176                       FALSE          93464650 5416681912
## ...                                ...               ...        ...
## SRR1027187                       FALSE          64697612 3567078255
## SRR1027188                       FALSE          65278500 4856453823
## SRR1027189                       FALSE          65328289 4858587600
## SRR1027190                       FALSE          73911898 5501089036
## SRR1027172                       FALSE          57523391 3351013968
##            sharq_beta_tissue sharq_beta_cell_type
##                  <character>          <character>
## SRR1027171            breast                  esc
## SRR1027173            breast                  esc
## SRR1027174            breast                  esc
## SRR1027175            breast                  esc
## SRR1027176            breast                  esc
## ...                      ...                  ...
## SRR1027187            breast                  esc
## SRR1027188            breast                  esc
## SRR1027189            breast                  esc
## SRR1027190            breast                  esc
## SRR1027172            breast                  esc
##            biosample_submission_date biosample_publication_date
```

```
##                          <character>                <character>
## SRR1027171   2013-11-07T12:40:22.203    2013-11-08T01:11:17.160
## SRR1027173   2013-11-07T12:40:32.283    2013-11-08T01:11:14.827
## SRR1027174   2013-11-07T12:40:28.283    2013-11-08T01:11:52.283
## SRR1027175   2013-11-07T12:40:34.343    2013-11-08T01:11:15.963
## SRR1027176   2013-11-07T12:40:36.303    2013-11-08T01:11:46.430
## ...                              ...                        ...
## SRR1027187   2013-11-07T12:40:56.180    2013-11-08T01:11:29.587
## SRR1027188   2013-11-07T12:40:58.170    2013-11-08T01:12:06.660
## SRR1027189   2013-11-07T12:40:20.227    2013-11-08T01:11:33.080
## SRR1027190   2013-11-07T12:40:18.090    2013-11-08T01:12:11.320
## SRR1027172   2013-11-07T12:40:26.217    2013-11-08T01:11:45.250
##              biosample_update_date avg_read_length geo_accession
##                        <character>       <integer>   <character>
## SRR1027171 2014-03-07T16:09:38.542             120    GSM1261016
## SRR1027173 2014-03-07T16:09:38.698             120    GSM1261018
## SRR1027174 2014-03-07T16:09:38.637             120    GSM1261019
## SRR1027175 2014-03-07T16:09:38.731             120    GSM1261020
## SRR1027176 2014-03-07T16:09:38.768             120    GSM1261021
## ...                             ...             ...           ...
## SRR1027187 2014-03-07T16:09:39.093             120    GSM1261032
## SRR1027188 2014-03-07T16:09:39.130             150    GSM1261033
## SRR1027189 2014-03-07T16:09:38.498             150    GSM1261034
## SRR1027190 2014-03-07T16:09:38.469             150    GSM1261035
## SRR1027172 2014-03-07T16:09:38.604              87    GSM1261017
##              bigwig_file       title
##              <character> <character>
## SRR1027171 SRR1027171.bw       TNBC1
## SRR1027173 SRR1027173.bw       TNBC3
## SRR1027174 SRR1027174.bw       TNBC4
## SRR1027175 SRR1027175.bw       TNBC5
## SRR1027176 SRR1027176.bw       TNBC6
## ...                   ...         ...
## SRR1027187 SRR1027187.bw      HER2-5
## SRR1027188 SRR1027188.bw        NBS1
## SRR1027189 SRR1027189.bw        NBS2
## SRR1027190 SRR1027190.bw        NBS3
## SRR1027172 SRR1027172.bw       TNBC2
##                                    characteristics
##                                    <CharacterList>
## SRR1027171          tumor type: TNBC Breast Tumor
## SRR1027173          tumor type: TNBC Breast Tumor
## SRR1027174          tumor type: TNBC Breast Tumor
## SRR1027175          tumor type: TNBC Breast Tumor
## SRR1027176          tumor type: TNBC Breast Tumor
## ...                                           ...
## SRR1027187 tumor type: HER2 Positive Breast Tumor
## SRR1027188    tumor type: Normal Breast Organoids
## SRR1027189    tumor type: Normal Breast Organoids
## SRR1027190    tumor type: Normal Breast Organoids
## SRR1027172          tumor type: TNBC Breast Tumor
```

As above, downloaded count data are first scaled to take into account differing coverage between samples. The same phenotype data (`pheno`) are used and again ordered to match the sample order of the expression

data (rse_exon). Only those samples that are HER2-positive or TNBC are included for analysis. Prior to differential exon expression analysis, count data are obtained in matrix format and then filtered to only include exons within genes that had been analyzed previously.

```
## Scale counts by taking into account the total coverage per sample
rse <- scale_counts(rse_exon)

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

## Obtain correct order for pheno data
pheno <- pheno[match(rse$run, pheno$Run_s), ]
identical(pheno$Run_s, rse$run)
```

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

```
head(cbind(pheno$Run_s, rse$run))
```

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

```
## Obtain grouping information
colData(rse)$group <- pheno$tumor_type_s
table(colData(rse)$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 types
rse <- rse[, rse$group %in% c('HER2 Positive Breast Tumor',
                              'TNBC Breast Tumor')]

## Save filtered rse object
rse_exon_filt <- rse
rse_exon_filt
```

```
## class: RangedSummarizedExperiment
## dim: 226117 11
## metadata(0):
## assays(1): counts
```

```
## rownames(226117): 1 1 ... 9997 9997
## rowData names(0):
## colnames(11): SRR1027171 SRR1027173 ... SRR1027187 SRR1027172
## colData names(22): project sample ... characteristics group
```

```
## Obtain count matrix
counts <- assays(rse_exon_filt)$counts
dim(counts)
```

```
## [1] 226117     11
```

```
## Filter count matrix (keep exons that are in filtered gene counts matrix)
filter <- rownames(counts) %in% rownames(counts_gene)
counts <- counts[filter, ]
dim(counts)
```

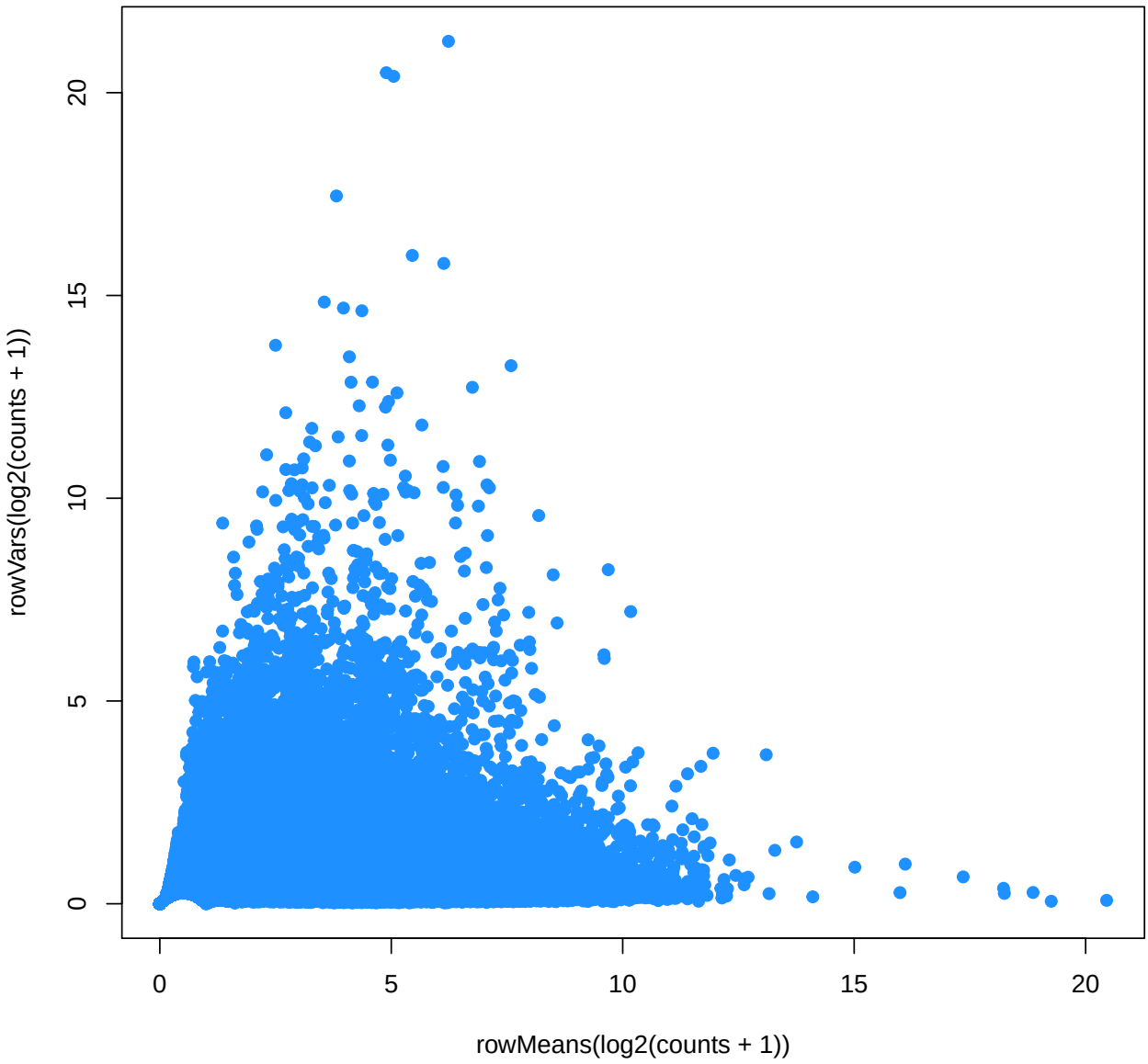
```
## [1] 204559     11
```

```
## Save for gene, exon and junction comparisons
counts_exon <- counts
counts_exon[1:5, 1:5]
```

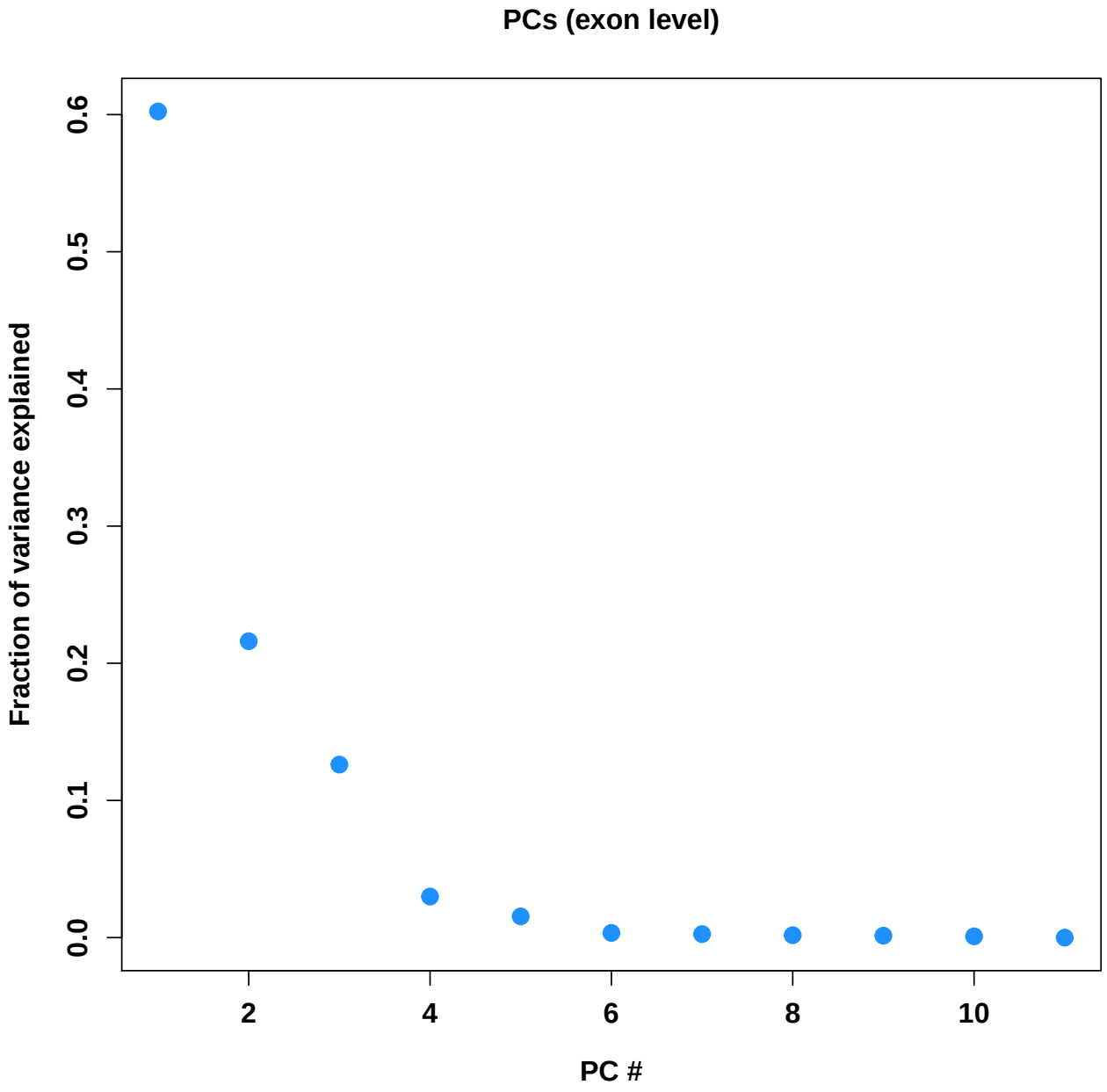
```
##   SRR1027171 SRR1027173 SRR1027174 SRR1027175 SRR1027176
## 1         10          6          8         15         10
## 1         10          9         17         23         19
## 1          7          3          1          3          6
## 1         14         13          4          6         15
## 1          3          2          1          1          2
```

As above, to get a better sense of the data, we assess the mean-variance relationship for each exon. Similarly, we run principal component analysis (PCA) to identify any sample outliers within the data. We assess the variance explained by each of the first 11 PCs as well as visualize the relationship of each sample in the first two PCs.

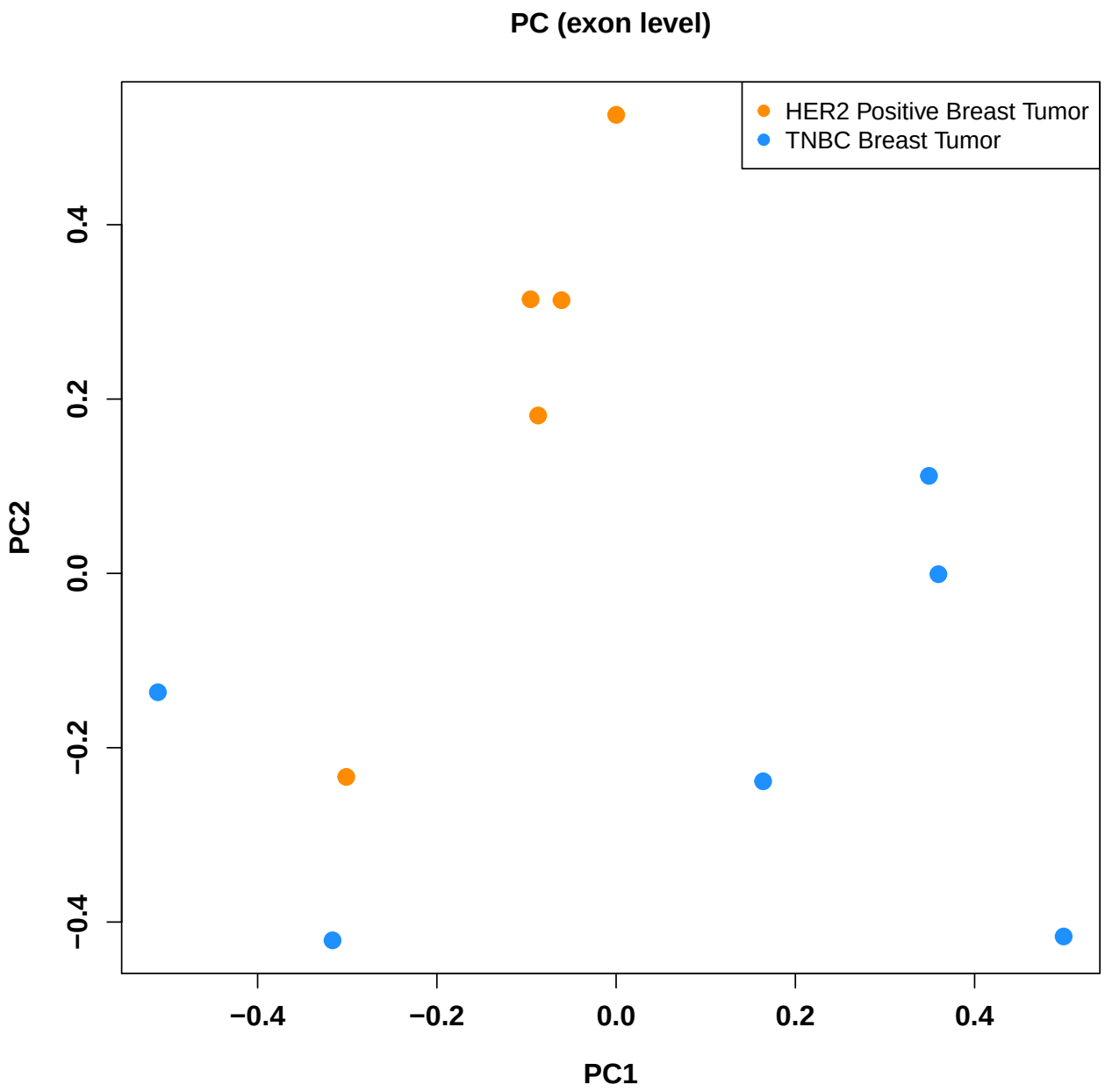
```
## Set colors
trop <- RSkittleBrewer('tropical')[c(1, 2)]
cols <- as.numeric(as.factor(rse$group))

## Look at mean variance relationship
plot(rowMeans(log2(counts + 1)), rowVars(log2(counts + 1)),
     pch = 19, col = trop[2])
```




```
## Calculate PCs with svd function
expr.pca <- svd(counts - rowMeans(counts))

## Plot PCs
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(expr.pca$d^2 / sum(expr.pca$d^2), pch = 19, col = trop[2], cex = 1.5,
     ylab = 'Fraction of variance explained', xlab = 'PC #',
     main = 'PCs (exon level)')
```




```
## Plot PC1 vs. PC2
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(expr.pca$v[, 1], expr.pca$v[, 2], pch = 19, col = trop[cols], cex = 1.5,
     xlab = 'PC1', ylab = 'PC2',
     main = 'PC (exon level)')
legend('topright', pch = 19, col = trop[c(1, 2)],
       names(summary(as.factor(rse$group))))
```



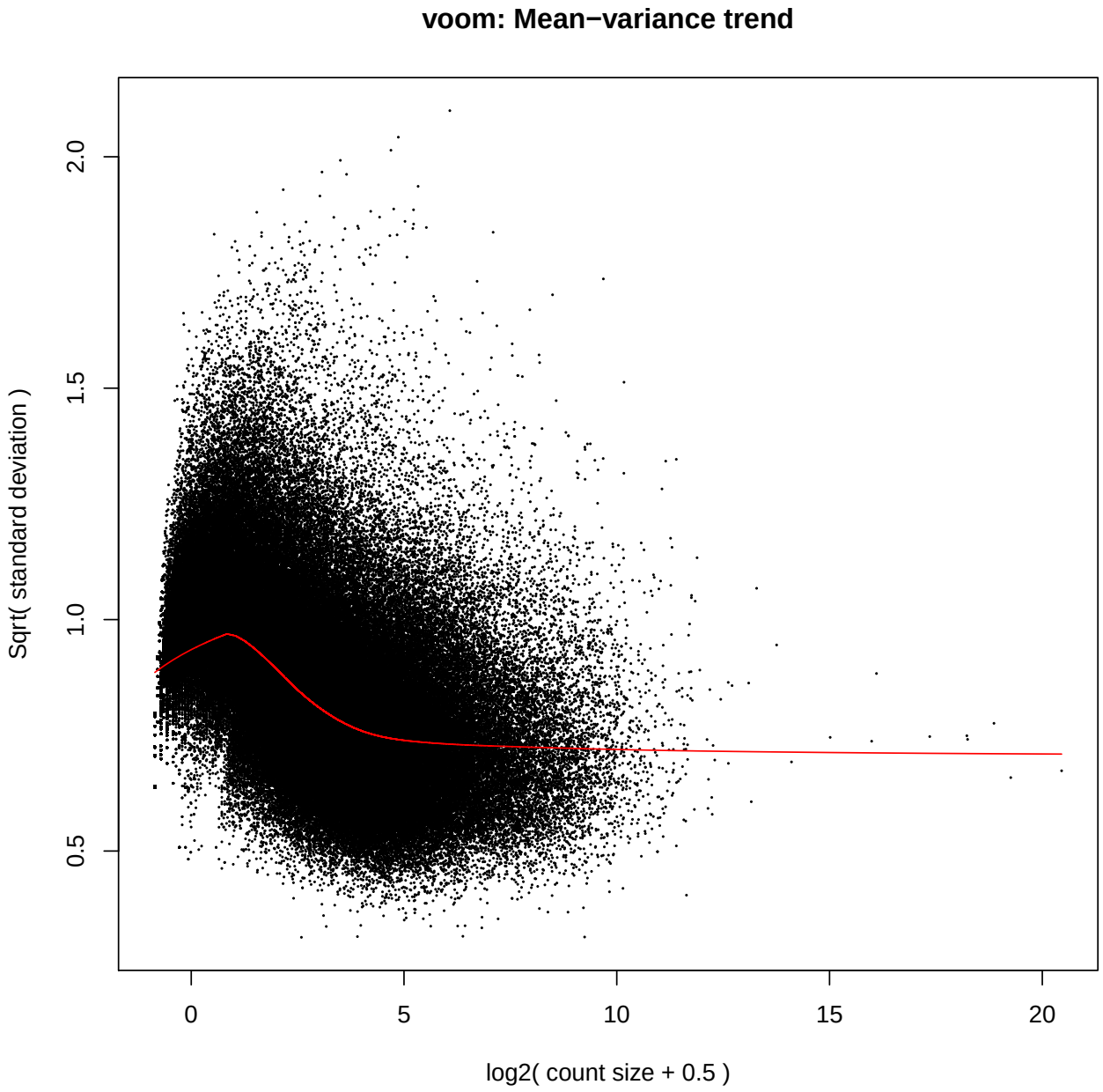
Again, differential expression analysis is carried out using `limma` and `voom`; however, this time at the exon, rather than gene, level. Data are again visualized using a volcano plot to assess the strength [$log_2(fold-change)$ in expression] and its significance [$-log_10(p-value)$].for each exon.

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

```
##    (Intercept) rse$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")$`rse$group`
## [1] "contr.treatment"
```

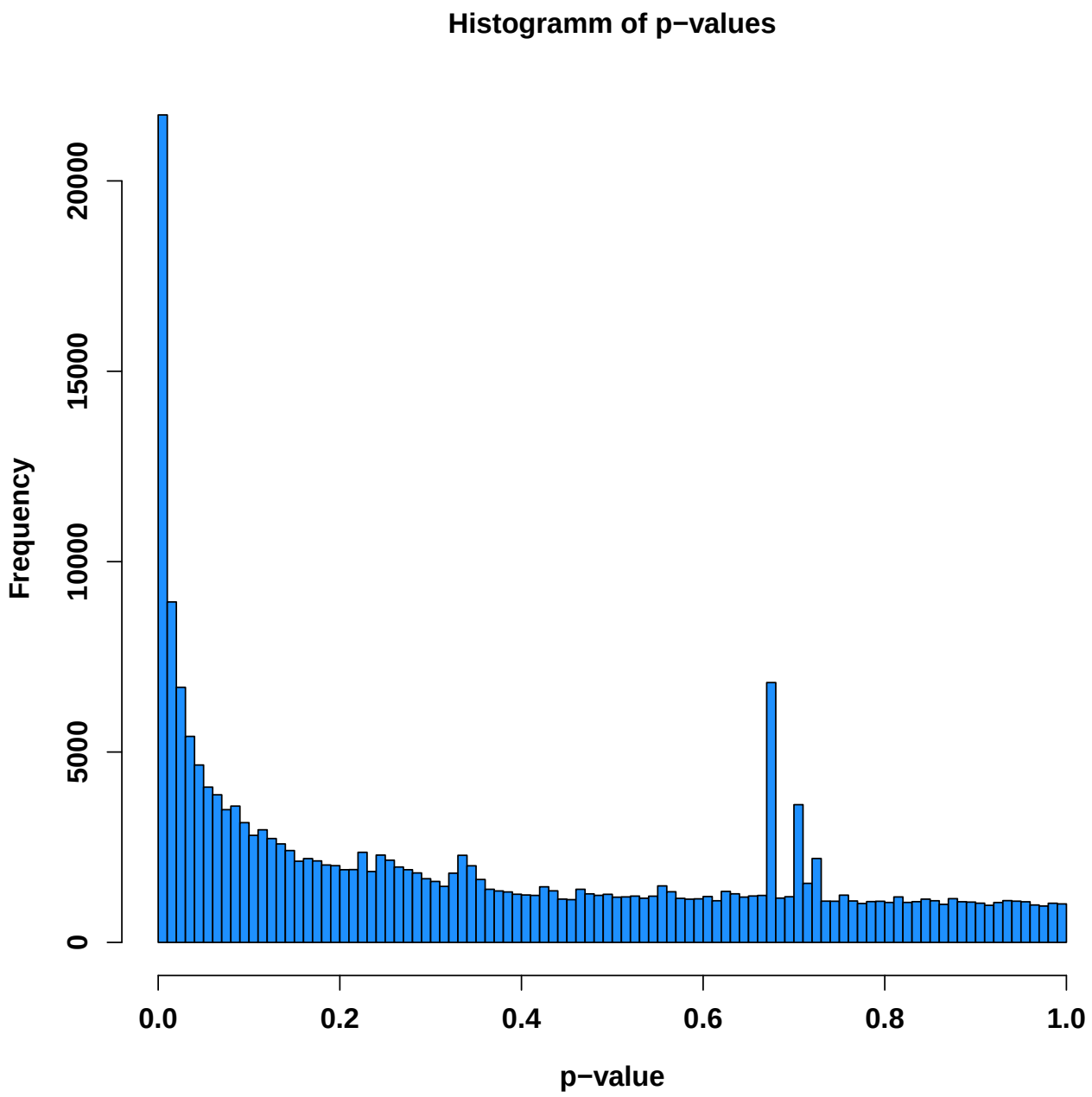
```
dge <- DGEList(counts = counts)
dge <- calcNormFactors(dge)
v <- voom(dge, design,plot = TRUE)
```



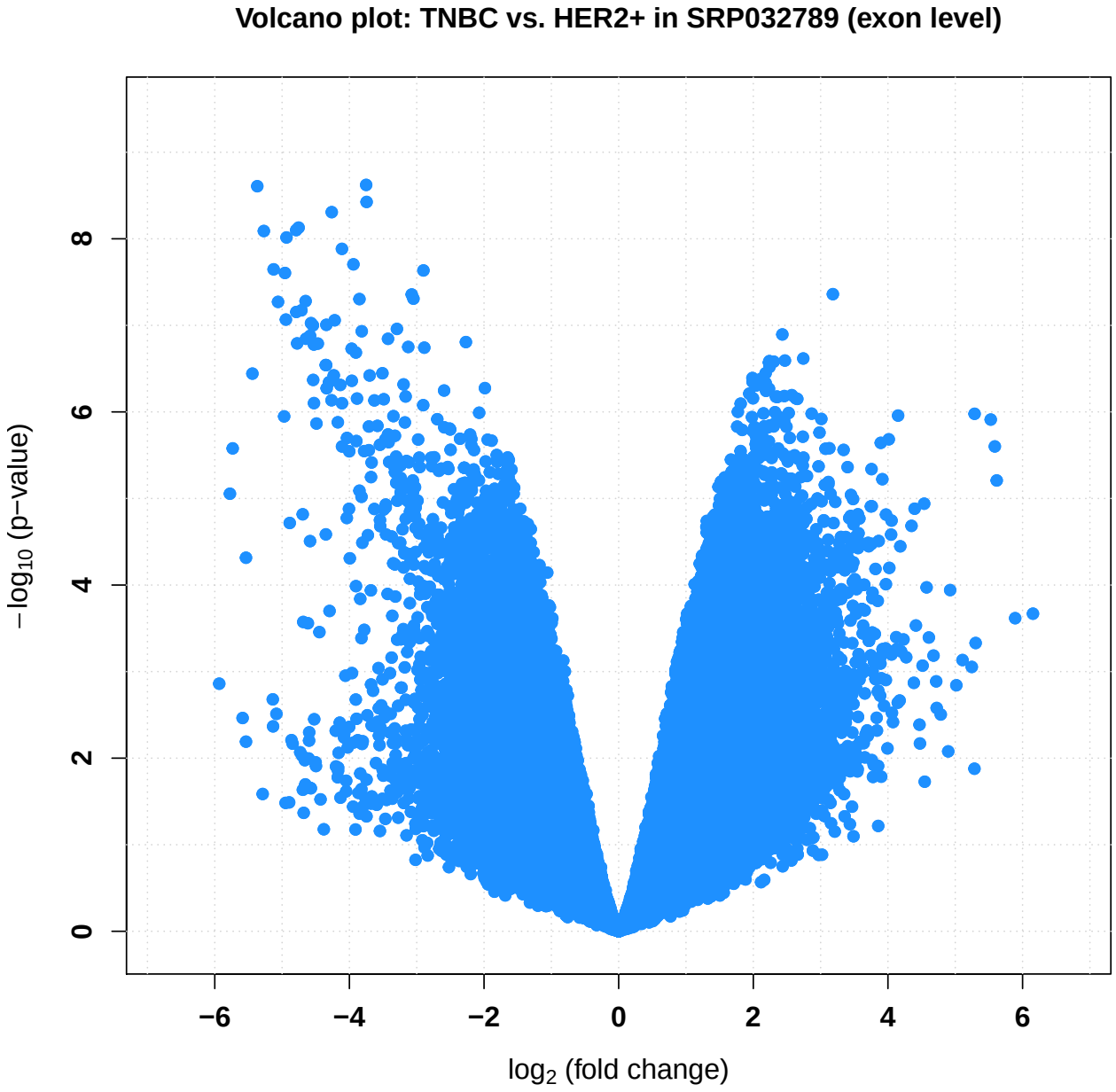
```
fit <- lmFit(v, design)
fit <- eBayes(fit)
log2FC <- fit$coefficients[, 2]
p.mod <- fit$p.value[, 2]
q.mod <- qvalue(p.mod)$q
res_exon <- data.frame(log2FC, p.mod, q.mod)


## Determine the number of exons differentially expressed at q<0.05
sum(res_exon$q.mod < 0.05)
```

```
## [1] 23647
```

```
## Histogram of p-values
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
hist(p.mod, col = trop[2], xlab = 'p-value',
     main = 'Histogramm of p-values', breaks = 100)
```




```
## 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 (exon level)')
```



Junction level analysis

As above, we are interested here in differential expression. However, rather than summarizing across genes, this analysis will look for differential expression at the junction level. In this analysis, we include all junctions that map to the previous filtered genes and again carry out differential expression analysis using `limma` and `voom`.

Here, we download data from the same project as above (SRP032798); however, this time, we are interested in obtaining the junction level data.

```
## Find a project of interest (SRP032789)
project_info <- abstract_search('To define the digital transcriptome of three breast cancer')
project_info
```

```
##     number_samples species
## 865             20   human
##
## 865 Goal: To define the digital transcriptome of three breast cancer subtypes (TNBC, Non-TNBC, and H
##       project
## 865 SRP032789
```

```
## Browse the project at SRA
browse_study(project_info$project)

## Download the exon level RangedSummarizedExperiment data
if(!file.exists(file.path('SRP032789', 'rse_jx.Rdata'))) {
    download_study(project_info$project, type = 'rse-jx')
}

## Load the data
load(file.path(project_info$project, 'rse_jx.Rdata'))
rse_jx
```

```
## class: RangedSummarizedExperiment
## dim: 672203 20
## metadata(0):
## assays(1): counts
## rownames: NULL
## rowData names(8): junction_id found_junction_gencode_v24 ...
##   symbol class
## colnames(20): SRR1027171 SRR1027173 ... SRR1027190 SRR1027172
## colData names(21): project sample ... title characteristics
```

```
## This is the sample phenotype data provided by the recount project
colData(rse_jx)
```

```
## DataFrame with 20 rows and 21 columns
##                project      sample  experiment         run
##            <character> <character> <character> <character>
## SRR1027171   SRP032789   SRS500214   SRX374850  SRR1027171
## SRR1027173   SRP032789   SRS500216   SRX374852  SRR1027173
## SRR1027174   SRP032789   SRS500217   SRX374853  SRR1027174
## SRR1027175   SRP032789   SRS500218   SRX374854  SRR1027175
## SRR1027176   SRP032789   SRS500219   SRX374855  SRR1027176
## ...               ...         ...         ...         ...
## SRR1027187   SRP032789   SRS500230   SRX374866  SRR1027187
## SRR1027188   SRP032789   SRS500231   SRX374867  SRR1027188
## SRR1027189   SRP032789   SRS500232   SRX374868  SRR1027189
## SRR1027190   SRP032789   SRS500233   SRX374869  SRR1027190
## SRR1027172   SRP032789   SRS500215   SRX374851  SRR1027172
##            read_count_as_reported_by_sra reads_downloaded
##                                <integer>        <integer>
## SRR1027171                      88869444         88869444
## SRR1027173                     107812596        107812596
## SRR1027174                      98563260         98563260
## SRR1027175                      91327892         91327892
## SRR1027176                      96513572         96513572
```

```
## ...                                ...             ...
## SRR1027187                     75260678         75260678
## SRR1027188                     65709192         65709192
## SRR1027189                     65801392         65801392
## SRR1027190                     74356276         74356276
## SRR1027172                     80986440         58902122
##            proportion_of_reads_reported_by_sra_downloaded paired_end
##                                                 <numeric>  <logical>
## SRR1027171                                              1       TRUE
## SRR1027173                                              1       TRUE
## SRR1027174                                              1       TRUE
## SRR1027175                                              1       TRUE
## SRR1027176                                              1       TRUE
## ...                                                   ...        ...
## SRR1027187                                      1.0000000       TRUE
## SRR1027188                                      1.0000000       TRUE
## SRR1027189                                      1.0000000       TRUE
## SRR1027190                                      1.0000000       TRUE
## SRR1027172                                      0.7273084       TRUE
##            sra_misreported_paired_end mapped_read_count        auc
##                             <logical>         <integer>  <numeric>
## SRR1027171                      FALSE          86949307 5082692127
## SRR1027173                      FALSE         104337779 6077034329
## SRR1027174                      FALSE          95271238 5504462845
## SRR1027175                      FALSE          88820239 5150234117
## SRR1027176                      FALSE          93464650 5416681912
## ...                               ...               ...        ...
## SRR1027187                      FALSE          64697612 3567078255
## SRR1027188                      FALSE          65278500 4856453823
## SRR1027189                      FALSE          65328289 4858587600
## SRR1027190                      FALSE          73911898 5501089036
## SRR1027172                      FALSE          57523391 3351013968
##            sharq_beta_tissue sharq_beta_cell_type
##                  <character>          <character>
## SRR1027171            breast                  esc
## SRR1027173            breast                  esc
## SRR1027174            breast                  esc
## SRR1027175            breast                  esc
## SRR1027176            breast                  esc
## ...                      ...                  ...
## SRR1027187            breast                  esc
## SRR1027188            breast                  esc
## SRR1027189            breast                  esc
## SRR1027190            breast                  esc
## SRR1027172            breast                  esc
##            biosample_submission_date biosample_publication_date
##                          <character>                <character>
## SRR1027171   2013-11-07T12:40:22.203    2013-11-08T01:11:17.160
## SRR1027173   2013-11-07T12:40:32.283    2013-11-08T01:11:14.827
## SRR1027174   2013-11-07T12:40:28.283    2013-11-08T01:11:52.283
## SRR1027175   2013-11-07T12:40:34.343    2013-11-08T01:11:15.963
## SRR1027176   2013-11-07T12:40:36.303    2013-11-08T01:11:46.430
## ...                             ...                        ...
## SRR1027187   2013-11-07T12:40:56.180    2013-11-08T01:11:29.587
```

```
## SRR1027188   2013-11-07T12:40:58.170    2013-11-08T01:12:06.660
## SRR1027189   2013-11-07T12:40:20.227    2013-11-08T01:11:33.080
## SRR1027190   2013-11-07T12:40:18.090    2013-11-08T01:12:11.320
## SRR1027172   2013-11-07T12:40:26.217    2013-11-08T01:11:45.250
##              biosample_update_date avg_read_length geo_accession
##                        <character>       <integer>   <character>
## SRR1027171 2014-03-07T16:09:38.542             120    GSM1261016
## SRR1027173 2014-03-07T16:09:38.698             120    GSM1261018
## SRR1027174 2014-03-07T16:09:38.637             120    GSM1261019
## SRR1027175 2014-03-07T16:09:38.731             120    GSM1261020
## SRR1027176 2014-03-07T16:09:38.768             120    GSM1261021
## ...                            ...             ...           ...
## SRR1027187 2014-03-07T16:09:39.093             120    GSM1261032
## SRR1027188 2014-03-07T16:09:39.130             150    GSM1261033
## SRR1027189 2014-03-07T16:09:38.498             150    GSM1261034
## SRR1027190 2014-03-07T16:09:38.469             150    GSM1261035
## SRR1027172 2014-03-07T16:09:38.604              87    GSM1261017
##              bigwig_file       title
##              <character> <character>
## SRR1027171 SRR1027171.bw       TNBC1
## SRR1027173 SRR1027173.bw       TNBC3
## SRR1027174 SRR1027174.bw       TNBC4
## SRR1027175 SRR1027175.bw       TNBC5
## SRR1027176 SRR1027176.bw       TNBC6
## ...                   ...         ...
## SRR1027187 SRR1027187.bw      HER2-5
## SRR1027188 SRR1027188.bw        NBS1
## SRR1027189 SRR1027189.bw        NBS2
## SRR1027190 SRR1027190.bw        NBS3
## SRR1027172 SRR1027172.bw       TNBC2
##                                   characteristics
##                                   <CharacterList>
## SRR1027171          tumor type: TNBC Breast Tumor
## SRR1027173          tumor type: TNBC Breast Tumor
## SRR1027174          tumor type: TNBC Breast Tumor
## SRR1027175          tumor type: TNBC Breast Tumor
## SRR1027176          tumor type: TNBC Breast Tumor
## ...                                           ...
## SRR1027187 tumor type: HER2 Positive Breast Tumor
## SRR1027188    tumor type: Normal Breast Organoids
## SRR1027189    tumor type: Normal Breast Organoids
## SRR1027190    tumor type: Normal Breast Organoids
## SRR1027172          tumor type: TNBC Breast Tumor
```

As above, downloaded count data are first scaled to take into account differing coverage between samples. The same phenotype data (`pheno`) are used and again ordered to match the sample order of the expression data (`rse_jx`). Only those samples that are HER2-positive or TNBC are included for analysis. Prior to differential exon expression analysis, count data are obtained in matrix format and then filtered to only include junction within genes that had been analyzed previously.

```
## Scale counts by taking into account the total coverage per sample
rse <- scale_counts(rse_jx, by = 'mapped_reads', round = FALSE)

## Download pheno data from
```

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

## Obtain correct order for pheno data
pheno <- pheno[match(rse$run, pheno$Run_s), ]
identical(pheno$Run_s, rse$run)
```

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

```
head(cbind(pheno$Run_s, rse$run))
```

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

```
## Obtain grouping information
colData(rse)$group <- pheno$tumor_type_s
table(colData(rse)$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 types
rse <- rse[, rse$group %in% c('HER2 Positive Breast Tumor',
                              'TNBC Breast Tumor')]

## Save filtered rse object
rse_jx_filt <- rse
rse_jx_filt
```

```
## class: RangedSummarizedExperiment
## dim: 672203 11
## metadata(0):
## assays(1): counts
## rownames: NULL
## rowData names(8): junction_id found_junction_gencode_v24 ...
##   symbol class
## colnames(11): SRR1027171 SRR1027173 ... SRR1027187 SRR1027172
## colData names(22): project sample ... characteristics group
```

```
## Obtain count matrix
counts <- assays(rse_jx_filt)$counts
dim(counts)
```

```
## [1] 672203     11
```

```
##### Start: Obtain geneIDs for juctions
## Obtain geneIDs
gene_id <- rownames(counts_gene)

## Save number of genes that a junctions maps to
## We will exclude non-unique junctions later
num_genes <- lapply(rowData(rse_jx_filt)$gene_id_proposed, function(x) length(x))
num_genes <- unlist(num_genes)

## Save only the first gene_id
jx_gene_id <- lapply(rowData(rse_jx_filt)$gene_id, function(x) x[1])
jx_gene_id <- unlist(jx_gene_id)

## There are NAs: not every junctions is annotated
jx_gene_id[1:100]
```

```
##   [1] NA       NA       "653635" NA       NA       "653635" NA
##   [8] NA       NA       NA       NA       NA       NA       NA
##  [15] NA       NA       NA       NA       NA       NA       NA
##  [22] NA       NA       NA       NA       NA       "653635" NA
##  [29] NA       NA       NA       NA       NA       NA       NA
##  [36] NA       NA       NA       NA       NA       NA       NA
##  [43] NA       NA       NA       NA       NA       NA       NA
##  [50] NA       NA       "653635" NA       NA       NA       NA
##  [57] NA       NA       NA       NA       NA       NA       NA
##  [64] NA       NA       NA       NA       NA       NA       NA
##  [71] NA       NA       NA       NA       NA       NA       NA
##  [78] NA       NA       NA       NA       NA       NA       NA
##  [85] NA       NA       NA       NA       NA       NA       NA
##  [92] NA       NA       NA       NA       NA       NA       NA
##  [99] NA       NA
```

```
## Compare lengths
length(jx_gene_id) == dim(counts)[1]
```

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

```
## Find non-unique mapping junctions
double_jx <- which(num_genes >1)

## Check non-unique mapping junctions
rowData(rse_jx_filt)[double_jx, 'gene_id']
```

```
## CharacterList of length 3411
## [[1]] <NA>
```

```
## [[2]] <NA>
## [[3]] <NA>
## [[4]] <NA>
## [[5]] <NA>
## [[6]] <NA>
## [[7]] <NA>
## [[8]] <NA>
## [[9]] <NA>
## [[10]] <NA>
## ...
## <3401 more elements>
```

```
## Set non-unique mapping junctions to "NA" in
jx_gene_id[double_jx] <- NA

rownames(counts) <- jx_gene_id
##### End: Obtain geneIDs for juctions


## Filter count matrix (keep exons that are in filtered gene counts matrix)
filter <- rownames(counts) %in% rownames(counts_gene)
counts <- counts[filter, ]
dim(counts)
```

```
## [1] 187013     11
```

```
## Since we only look at a subset of samples, there are many junctions with zero counts
## We remove them
counts <- counts[apply(counts, 1, sum) > 0, ]
dim(counts)
```

```
## [1] 171193     11
```

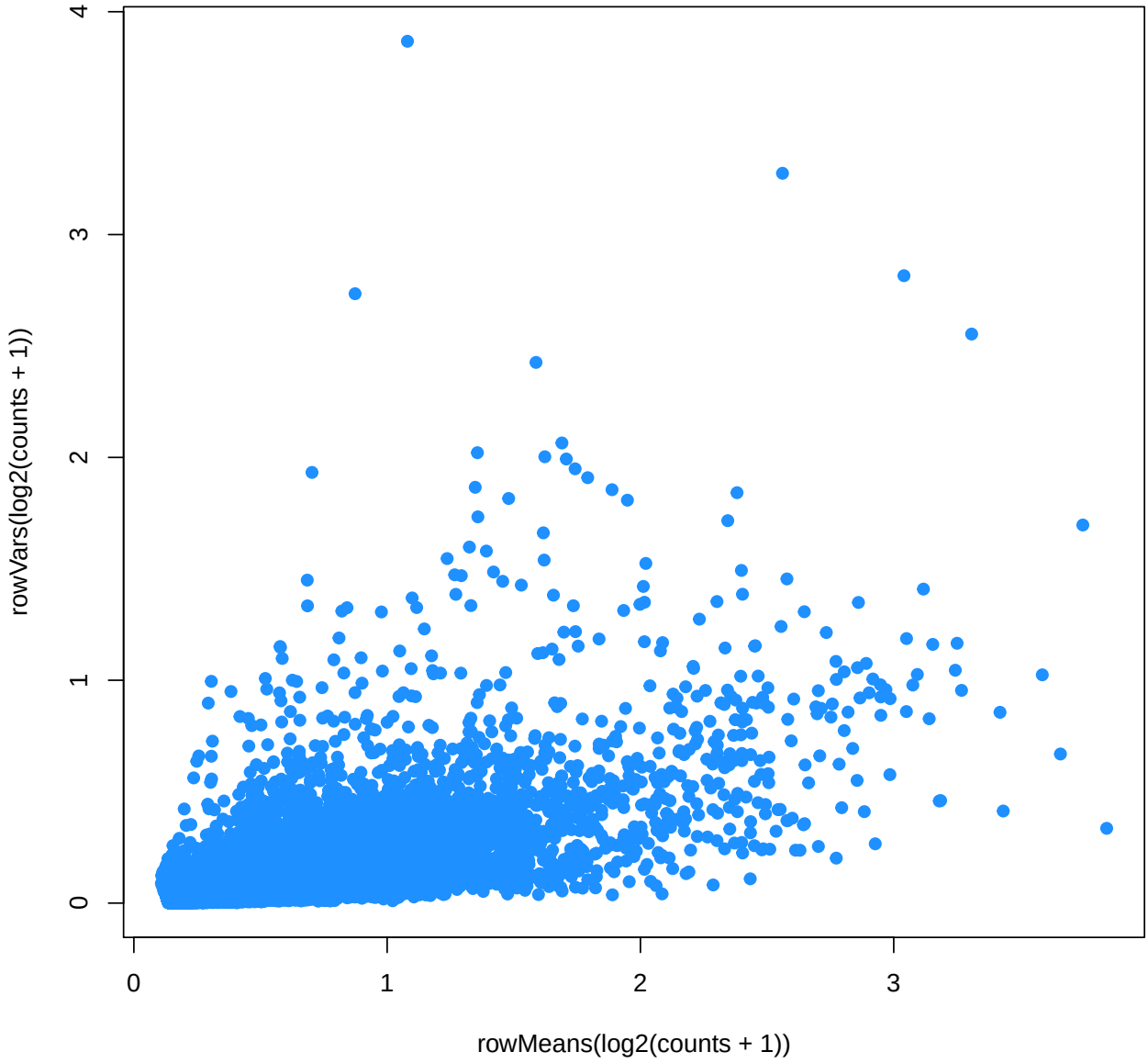
```
## Remove junctions with low counts across samples
counts <- counts[rowMeans(counts) > 0.1, ]

## Save for gene, exon and junction comparisons
counts_jx <- counts
counts_jx[1:10, ]
```

```
##        SRR1027171 SRR1027173 SRR1027174 SRR1027175 SRR1027176 SRR1027183
## 26155  0.20446141 0.10649173  0.1399513  0.1125870 0.43985733 0.00000000
## 26155  0.10223070 0.04259669  0.1516139  0.1125870 0.13076839 0.05579103
## 26155  0.14056722 0.07454421  0.1166261  0.1501159 0.16643250 0.08368654
## 26155  0.25557676 0.10649173  0.2099269  0.2501932 0.24964875 0.07438803
## 26155  0.10223070 0.02129835  0.2682400  0.2501932 0.26153679 0.09298504
## 57801  0.11500954 0.03194752  0.1516139  0.1876449 0.10699232 0.05579103
## 9636   0.60060539 0.03194752  0.1632765  0.3627802 0.07132822 0.18597009
## 375790 0.03833651 0.11714091  0.1516139  0.2251739 0.14265643 0.05579103
## 375790 0.24279792 0.21298347  0.1516139  0.3002319 0.11888036 0.12088056
## 375790 0.19168257 0.12779008  0.1166261  0.1876449 0.26153679 0.06508953
##        SRR1027184 SRR1027185  SRR1027186 SRR1027187 SRR1027172
```

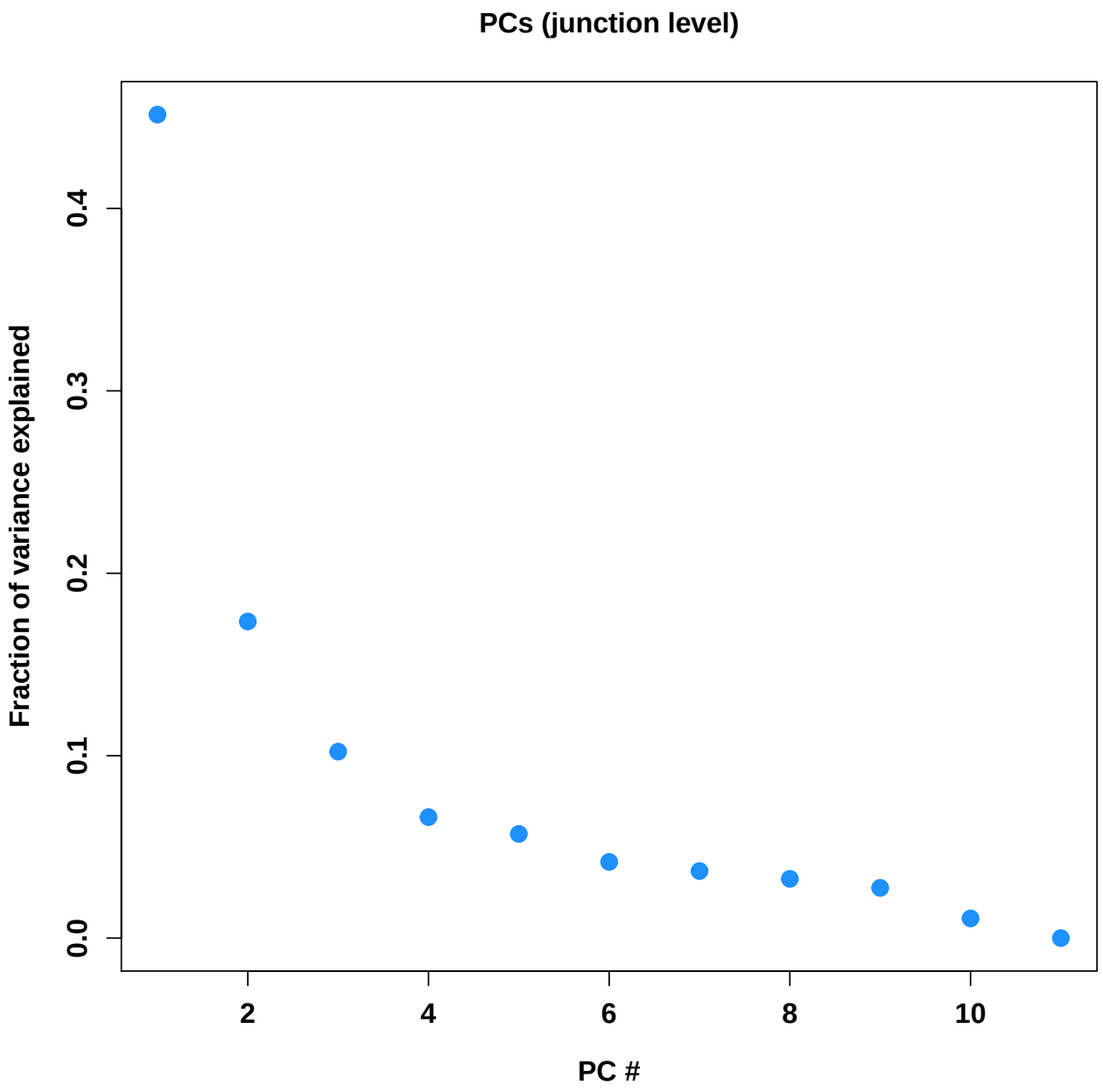
```
## 26155  0.02473931 0.02160652 0.009604733 0.08586956  0.5144759
## 26155  0.08246437 0.07562282 0.028814200 0.17173912  0.2204897
## 26155  0.06597150 0.11883586 0.048023666 0.12021739  0.3674828
## 26155  0.18966806 0.19445869 0.057628399 0.20608695  0.5512242
## 26155  0.04123219 0.19445869 0.115256798 0.08586956  0.4777276
## 57801  0.02473931 0.08642608 0.009604733 0.06869565  0.3307345
## 9636   0.04123219 0.15124564 0.067233132 0.18891303  0.5144759
## 375790 0.09071081 0.20526195 0.163280464 0.17173912  0.3674828
## 375790 0.07421794 0.10803260 0.038418933 0.20608695  0.5144759
## 375790 0.08246437 0.10803260 0.019209466 0.29195651  0.5512242
```

As above, to get a better sense of the data, we assess the mean-variance relationship for each juction. Similarly, we run principal component analysis (PCA) to identify any sample outliers within the data. We assess the variance explained by each of the first 11 PCs as well as visualize the relationship of each sample in the first two PCs.

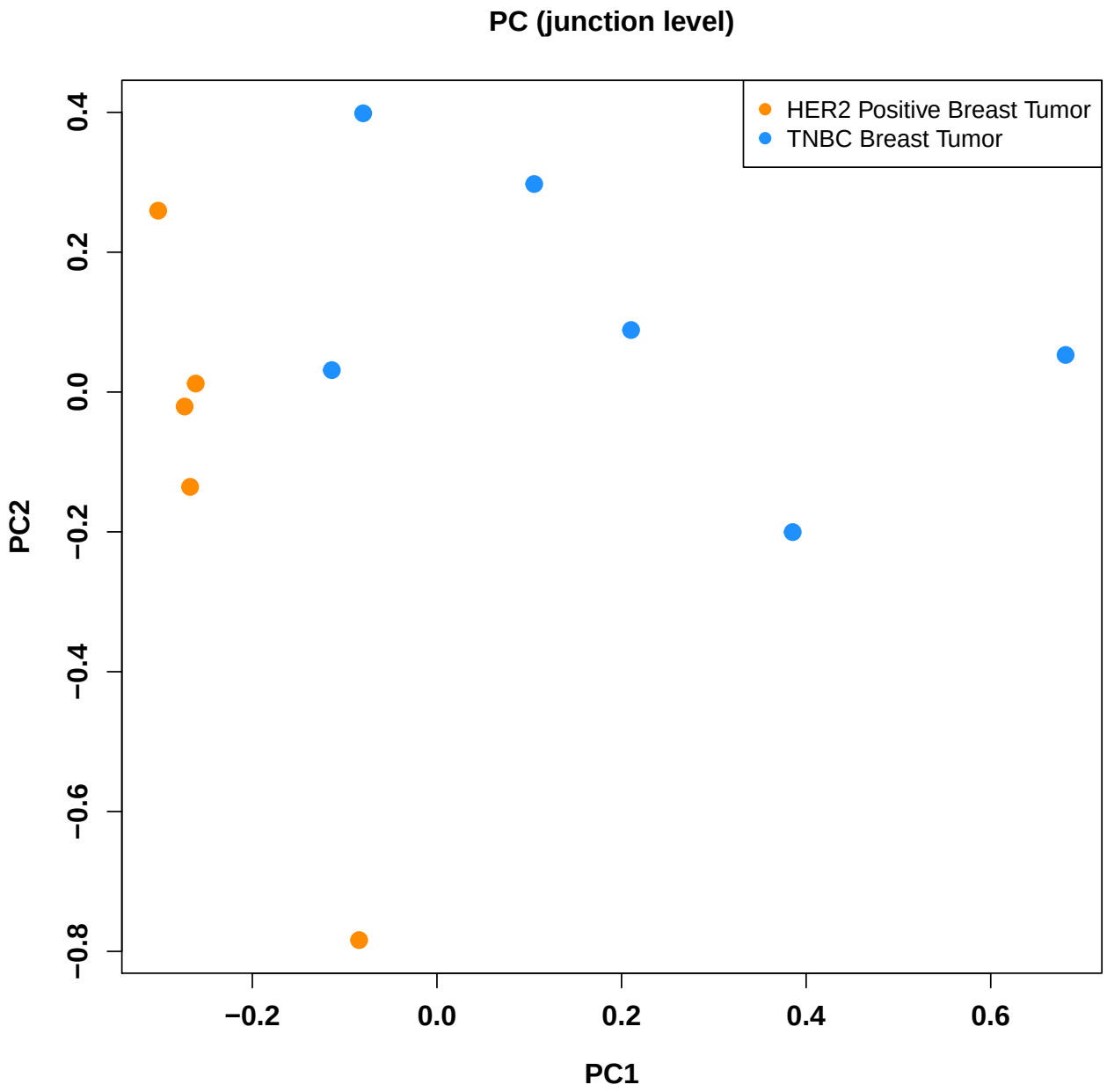
```
## Set colors
trop <- RSkittleBrewer('tropical')[c(1, 2)]
cols <- as.numeric(as.factor(rse$group))

## Look at mean variance relationship
plot(rowMeans(log2(counts + 1)), rowVars(log2(counts + 1)),
     pch = 19, col = trop[2])
```




```
## Calculate PCs with svd function
expr.pca <- svd(counts - rowMeans(counts))

## Plot PCs
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(expr.pca$d^2 / sum(expr.pca$d^2), pch = 19, col = trop[2], cex = 1.5,
     ylab = 'Fraction of variance explained', xlab = 'PC #',
     main = 'PCs (junction level)')
```




```
## Plot PC1 vs. PC2
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(expr.pca$v[, 1], expr.pca$v[, 2], pch = 19, col = trop[cols], cex = 1.5,
     xlab = 'PC1', ylab = 'PC2',
     main = 'PC (junction level)')
legend('topright', pch = 19, col = trop[c(1, 2)],
       names(summary(as.factor(rse$group))))
```



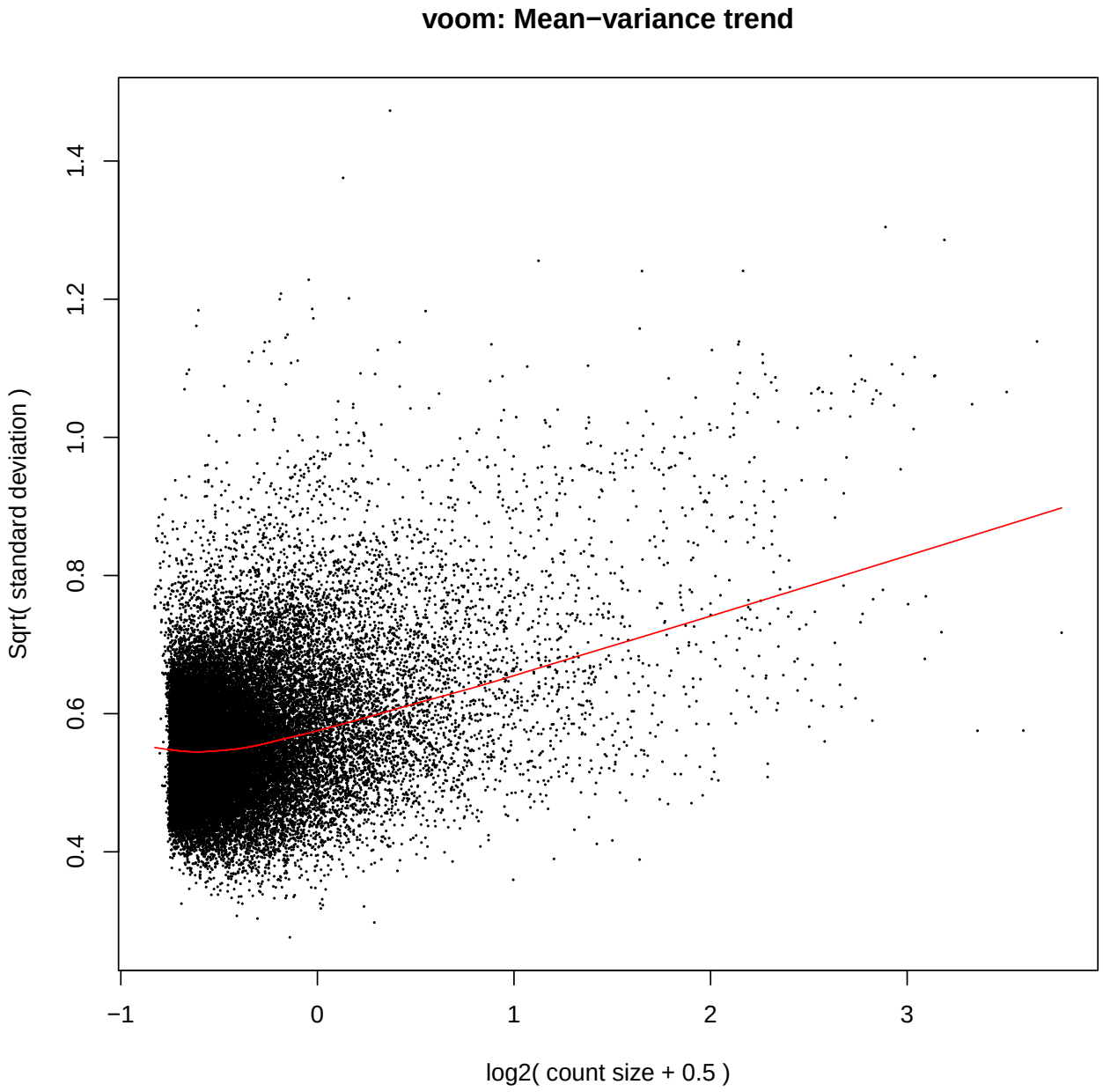
Again, differential expression analysis is carried out using `limma` and `voom`; however, this time at the junction, rather than gene, level. Data are again visualized using a volcano plot to assess the strength [$log_2(fold - change)$ in expression] and its significance [$-log_10(p - value)$] for each junction.

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

```
##    (Intercept) rse$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")$`rse$group`
## [1] "contr.treatment"
```

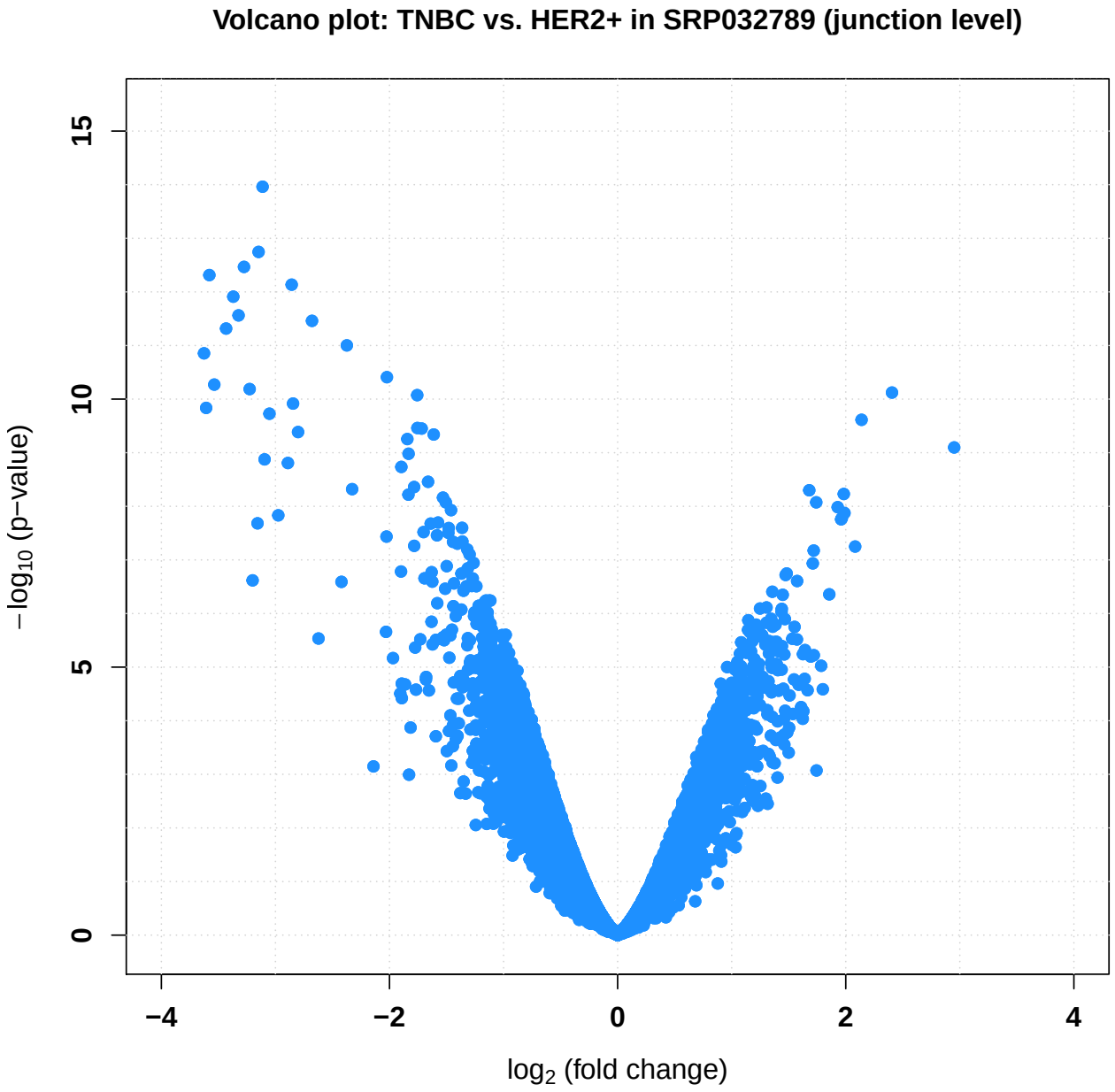
```
dge <- DGEList(counts = counts)
dge <- calcNormFactors(dge)
v <- voom(dge, design,plot = TRUE)
```



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

## Determine the number of exons differentially expressed at q<0.05
sum(res_jx$q.mod < 0.05)
```

```
## [1] 19805
```

```
## 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, 2356, 1), col = 'lightgray', lty = 'dotted')
points(log2FC, -log10(p.mod), pch = 19, col = trop[2])
title('Volcano plot: TNBC vs. HER2+ in SRP032789 (junction level)')
```



Comparison of gene, exon, junction, and DER results

To compare findings at the gene, exon, junction, and DER level, we obtained a single exon level [or junction level or DER level] p-value for each gene included at the gene level analysis. To do this, we utilized Simes' rule, such that for each gene included in the gene level analysis, the p-values for exons [or junctions or DERs] within that gene were extracted and sorted. Each exon level [or junction level or DER level] p-value is then multiplied by the number of exons [or junctions or DERs] present within the gene. For each exon [or junction or DER] (1,2...n), this quantity is divided by that exon's rank [or junction's rank or DER's rank] (where 1=most significant exon [or junction or DER] and n=least significant). The minimum value from this calcultion is assigned as the exon level [or junction level or DER level] p-value at each gene. DER results are loaded from the DER analysis report that is described and rendered in `recount_DER_SRP032789.*`

```
## Obtain geneIDs
gene_id <- unique(rownames(counts_exon))

## Calculate p-values for genes with Simes' rule
p_exon_gene <- NULL
for(i in seq_len(length(gene_id))){
    p_exon <- res_exon$p.mod[rownames(counts_exon) %in% gene_id[i]]
    p_exon <- sort(p_exon)
    p_exon_simes <- NULL
    for(j in 1:length(p_exon)){
        p_exon_simes[j] <- length(p_exon) * p_exon[j] / j
    }
    p_exon_gene[i] <- min(p_exon_simes)
}
names(p_exon_gene) <- gene_id
q_exon_gene <- qvalue(p_exon_gene)$q

## Determine the number of 'gene level exons' differentially expressed q < 0.05
sum(q_exon_gene < 0.05)
```

```
## [1] 7935
```

```
## As above, 'topGO' can be utilized to assign biological function to
## differentially expressed exons.

## Gene set analysis (p-values of genes derived with Simes' rule from exon p-values)
interesting <- function(x) x < 0.05

topgoobjBP <- new('topGOdata',
    description = 'biological process',
    ontology = 'BP', allGenes = q_exon_gene, geneSelectionFun = interesting,
    annotationFun = annFUN.org, mapping = 'org.Hs.eg.db', ID = 'entrez')
```

```
##
## Building most specific GOs .....

##  ( 10647 GO terms found. )

##
## Build GO DAG topology ..........

##  ( 14443 GO terms and 34755 relations. )

##
## Annotating nodes ...............

##  ( 13762 genes annotated to the GO terms. )
```

```
bptest <- runTest(topgoobjBP, algorithm = 'weight01', statistic = 'ks')
```

```
##
##           -- Weight01 Algorithm --
##
##       the algorithm is scoring 14443 nontrivial nodes
##       parameters:
##           test statistic: ks
##           score order: increasing

##
##   Level 20:  1 nodes to be scored    (0 eliminated genes)

##
##   Level 19:  7 nodes to be scored    (0 eliminated genes)

##
##   Level 18:  20 nodes to be scored   (1 eliminated genes)

##
##   Level 17:  40 nodes to be scored   (30 eliminated genes)

##
##   Level 16:  130 nodes to be scored  (91 eliminated genes)

##
##   Level 15:  266 nodes to be scored  (179 eliminated genes)

##
##   Level 14:  520 nodes to be scored  (552 eliminated genes)

##
##   Level 13:  925 nodes to be scored  (1159 eliminated genes)

##
##   Level 12:  1336 nodes to be scored (2451 eliminated genes)

##
##   Level 11:  1602 nodes to be scored (4326 eliminated genes)

##
##   Level 10:  1878 nodes to be scored (6122 eliminated genes)

##
##   Level 9:   1967 nodes to be scored (8287 eliminated genes)

##
##   Level 8:   1855 nodes to be scored (9989 eliminated genes)

##
##   Level 7:   1641 nodes to be scored (11002 eliminated genes)
```

```
##
##   Level 6:   1187 nodes to be scored (11986 eliminated genes)

##
##   Level 5:   679 nodes to be scored  (12493 eliminated genes)

##
##   Level 4:   293 nodes to be scored  (13053 eliminated genes)

##
##   Level 3:   74 nodes to be scored   (13234 eliminated genes)

##
##   Level 2:   21 nodes to be scored   (13397 eliminated genes)

##
##   Level 1:   1 nodes to be scored    (13475 eliminated genes)
```

```
bptest
```

```
##
## Description: biological process
## Ontology: BP
## 'weight01' algorithm with the 'ks' test
## 14443 GO terms scored: 77 terms with p < 0.01
## Annotation data:
##     Annotated genes: 13762
##     Significant genes: 6194
##     Min. no. of genes annotated to a GO: 1
##     Nontrivial nodes: 14443
```

```
bpres_exon <- GenTable(topgoobjBP, pval = bptest,
                       topNodes = length(bptest@score), numChar = 100)
head(bpres_exon, n = 10)
```

```
##         GO.ID                                                          Term
## 1  GO:0031124                                       mRNA 3'-end processing
## 2  GO:0051493                      regulation of cytoskeleton organization
## 3  GO:0000398                               mRNA splicing, via spliceosome
## 4  GO:0033120                          positive regulation of RNA splicing
## 5  GO:0007049                                                    cell cycle
## 6  GO:0006886                                intracellular protein transport
## 7  GO:0006369               termination of RNA polymerase II transcription
## 8  GO:1903507 negative regulation of nucleic acid-templated transcription
## 9  GO:0008286                             insulin receptor signaling pathway
## 10 GO:0048025       negative regulation of mRNA splicing, via spliceosome
##    Annotated Significant Expected    pval
## 1         80          61    36.01 5.3e-06
## 2        380         183   171.03 5.4e-06
## 3        264         175   118.82 1.9e-05
## 4         23          20    10.35 2.7e-05
```

```
## 5       1650         810   742.63 5.1e-05
## 6        937         482   421.72 0.00020
## 7         47          35    21.15 0.00025
## 8       1072         488   482.49 0.00045
## 9        302         156   135.92 0.00051
## 10        20          15     9.00 0.00060
```

```
## Obtain geneIDs
gene_id <- unique(rownames(counts_jx))

## Calculate p-values for genes with Simes' rule
p_jx_gene <- NULL
for(i in seq_len(length(gene_id))){
    p_jx <- res_jx$p.mod[rownames(counts_jx) %in% gene_id[i]]
    p_jx <- sort(p_jx)
    p_jx_simes <- NULL
    for(j in 1:length(p_jx)){
        p_jx_simes[j] <- length(p_jx) * p_jx[j] / j
    }
    p_jx_gene[i] <- min(p_jx_simes)
}
names(p_jx_gene) <- gene_id


## Determine the number of 'gene leveljunction' differentially expressed q < 0.05
q_jx_gene <- qvalue(p_jx_gene)$q

## Determine the number of 'gene level junctions' differentially expressed q < 0.05
sum(q_jx_gene < 0.05)
```

```
## [1] 4379
```

```
## As above, 'topGO' can be utilized to assign biological function to
## differentially expressed exons.

## Gene set analysis (p-values of genes derived with Simes' rule from exon p-values)
interesting <- function(x) x < 0.05

topgoobjBP <- new('topGOdata',
    description = 'biological process',
    ontology = 'BP', allGenes = q_jx_gene, geneSelectionFun = interesting,
    annotationFun = annFUN.org, mapping = 'org.Hs.eg.db', ID = 'entrez')
```

```
##
## Building most specific GOs .....

##  ( 7987 GO terms found. )

##
## Build GO DAG topology ..........

##  ( 11849 GO terms and 28345 relations. )
```

```
##
## Annotating nodes ...............

##  ( 6396 genes annotated to the GO terms. )
```

```
bptest <- runTest(topgoobjBP, algorithm = 'weight01', statistic = 'ks')
```

```
##
##           -- Weight01 Algorithm --
##
##       the algorithm is scoring 11849 nontrivial nodes
##       parameters:
##           test statistic: ks
##           score order: increasing

##
##   Level 20:  1 nodes to be scored    (0 eliminated genes)

##
##   Level 19:  5 nodes to be scored    (0 eliminated genes)

##
##   Level 18:  12 nodes to be scored   (1 eliminated genes)

##
##   Level 17:  21 nodes to be scored   (11 eliminated genes)

##
##   Level 16:  85 nodes to be scored   (37 eliminated genes)

##
##   Level 15:  190 nodes to be scored  (71 eliminated genes)

##
##   Level 14:  389 nodes to be scored  (246 eliminated genes)

##
##   Level 13:  667 nodes to be scored  (570 eliminated genes)

##
##   Level 12:  1017 nodes to be scored (1271 eliminated genes)

##
##   Level 11:  1275 nodes to be scored (2159 eliminated genes)

##
##   Level 10:  1544 nodes to be scored (3089 eliminated genes)

##
##   Level 9:   1626 nodes to be scored (3995 eliminated genes)
```

```
##
##   Level 8:   1575 nodes to be scored (4847 eliminated genes)

##
##   Level 7:   1425 nodes to be scored (5327 eliminated genes)

##
##   Level 6:   1039 nodes to be scored (5749 eliminated genes)

##
##   Level 5:   609 nodes to be scored  (5954 eliminated genes)

##
##   Level 4:   273 nodes to be scored  (6132 eliminated genes)

##
##   Level 3:   74 nodes to be scored   (6210 eliminated genes)

##
##   Level 2:   21 nodes to be scored   (6267 eliminated genes)

##
##   Level 1:   1 nodes to be scored    (6289 eliminated genes)
```

```
bptest
```

```
##
## Description: biological process
## Ontology: BP
## 'weight01' algorithm with the 'ks' test
## 11849 GO terms scored: 68 terms with p < 0.01
## Annotation data:
##     Annotated genes: 6396
##     Significant genes: 4050
##     Min. no. of genes annotated to a GO: 1
##     Nontrivial nodes: 11849
```

```
bpres_jx <- GenTable(topgoobjBP, pval = bptest,
                     topNodes = length(bptest@score), numChar = 100)
head(bpres_jx, n = 10)
```

```
##         GO.ID
## 1  GO:0006614
## 2  GO:0006415
## 3  GO:0016259
## 4  GO:0006414
## 5  GO:0019083
## 6  GO:0000184
## 7  GO:0006413
## 8  GO:0048205
## 9  GO:0016032
```

```
## 10 GO:0002479
##                                                                                                          Term
## 1                                       SRP-dependent cotranslational protein targeting to membrane
## 2                                                                            translational termination
## 3                                                                     selenocysteine metabolic process
## 4                                                                            translational elongation
## 5                                                                                 viral transcription
## 6                              nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
## 7                                                                            translational initiation
## 8                                                                       COPI coating of Golgi vesicle
## 9                                                                                       viral process
## 10 antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
##    Annotated Significant Expected    pval
## 1         90          76    56.99 4.4e-17
## 2        143         108    90.55 7.0e-15
## 3         70          61    44.32 9.8e-15
## 4        162         121   102.58 2.7e-14
## 5        143         111    90.55 1.8e-13
## 6         94          80    59.52 9.8e-13
## 7        217         165   137.41 5.5e-12
## 8         13          13     8.23 5.5e-05
## 9        557         405   352.70 5.8e-05
## 10        53          43    33.56 0.00017
```

```r
## Load p-values from DER anaysis
load(file.path(project_info$project, 'AnnotatedDERs.Rdata'))
p.mod <- annotatedDERs

## Obtain geneIDs
gene_id <- unique(names(p.mod))

## Calculate p-values for genes with Simes' rule
p_DER_gene <- NULL
for(i in seq_len(length(gene_id))){
    p_DER <- p.mod[names(p.mod) %in% gene_id[i]]
    p_DER <- sort(p_DER)
    p_DER_simes <- NULL
    for(j in 1:length(p_DER)){
        p_DER_simes[j] <- length(p_DER) * p_DER[j] / j
    }
    p_DER_gene[i] <- min(p_DER_simes)
}
names(p_DER_gene) <- gene_id
q_DER_gene <- qvalue(p_DER_gene)$q

## Determine the number of 'gene level DERs' differentially expressed q < 0.05
sum(q_DER_gene < 0.05)
```

```
## [1] 6824
```

```r
## As above, 'topGO' can be utilized to assign biological function to
## differentially expressed DERs.
```

```
## Gene set analysis (p-values of genes derived with Simes' rule from DER p-values)
interesting <- function(x) x < 0.05

topgoobjBP <- new('topGOdata',
    description = 'biological process',
    ontology = 'BP', allGenes = q_DER_gene, geneSelectionFun = interesting,
    annotationFun = annFUN.org, mapping = 'org.Hs.eg.db', ID = 'entrez')
```

```
##
## Building most specific GOs .....

##  ( 9963 GO terms found. )

##
## Build GO DAG topology ..........

##  ( 13811 GO terms and 33233 relations. )

##
## Annotating nodes ...............

##  ( 11049 genes annotated to the GO terms. )
```

```
bptest <- runTest(topgoobjBP, algorithm = 'weight01', statistic = 'ks')
```

```
##
##           -- Weight01 Algorithm --
##
##       the algorithm is scoring 13811 nontrivial nodes
##       parameters:
##           test statistic: ks
##           score order: increasing

##
##   Level 20:  1 nodes to be scored    (0 eliminated genes)

##
##   Level 19:  6 nodes to be scored    (0 eliminated genes)

##
##   Level 18:  18 nodes to be scored   (1 eliminated genes)

##
##   Level 17:  39 nodes to be scored   (19 eliminated genes)

##
##   Level 16:  121 nodes to be scored  (61 eliminated genes)

##
##   Level 15:  255 nodes to be scored  (137 eliminated genes)
```

```
##
##   Level 14:  498 nodes to be scored  (450 eliminated genes)

##
##   Level 13:  868 nodes to be scored  (976 eliminated genes)

##
##   Level 12:  1278 nodes to be scored (2082 eliminated genes)

##
##   Level 11:  1531 nodes to be scored (3630 eliminated genes)

##
##   Level 10:  1792 nodes to be scored (5116 eliminated genes)

##
##   Level 9:   1863 nodes to be scored (6821 eliminated genes)

##
##   Level 8:   1771 nodes to be scored (8222 eliminated genes)

##
##   Level 7:   1576 nodes to be scored (9027 eliminated genes)

##
##   Level 6:   1153 nodes to be scored (9783 eliminated genes)

##
##   Level 5:   661 nodes to be scored  (10157 eliminated genes)

##
##   Level 4:   285 nodes to be scored  (10510 eliminated genes)

##
##   Level 3:   73 nodes to be scored   (10647 eliminated genes)

##
##   Level 2:   21 nodes to be scored   (10777 eliminated genes)

##
##   Level 1:   1 nodes to be scored    (10833 eliminated genes)
```

```
bptest
```

```
##
## Description: biological process
## Ontology: BP
## 'weight01' algorithm with the 'ks' test
## 13811 GO terms scored: 59 terms with p < 0.01
## Annotation data:
##     Annotated genes: 11049
##     Significant genes: 5739
##     Min. no. of genes annotated to a GO: 1
##     Nontrivial nodes: 13811
```

```
bpres_DER <- GenTable(topgoobjBP, pval = bptest,
                      topNodes = length(bptest@score), numChar = 100)
head(bpres_DER, n = 10)
```

```
##         GO.ID
## 1  GO:0031124
## 2  GO:0033120
## 3  GO:0000398
## 4  GO:0051493
## 5  GO:0016032
## 6  GO:0010606
## 7  GO:0006886
## 8  GO:0002026
## 9  GO:0006369
## 10 GO:0000244
##                                                                  Term
## 1                                          mRNA 3'-end processing
## 2                                 positive regulation of RNA splicing
## 3                                      mRNA splicing, via spliceosome
## 4                               regulation of cytoskeleton organization
## 5                                                        viral process
## 6  positive regulation of cytoplasmic mRNA processing body assembly
## 7                                      intracellular protein transport
## 8                        regulation of the force of heart contraction
## 9                       termination of RNA polymerase II transcription
## 10                             spliceosomal tri-snRNP complex assembly
##    Annotated Significant Expected    pval
## 1         75          57    38.96 6.2e-05
## 2         21          17    10.91 0.00018
## 3        254         180   131.93 0.00027
## 4        335         195   174.00 0.00029
## 5        692         413   359.43 0.00080
## 6          6           6     3.12 0.00081
## 7        862         498   447.73 0.00091
## 8         19          13     9.87 0.00098
## 9         42          30    21.82 0.00115
## 10        11          11     5.71 0.00138
```

To determine the concordance between the gene level and (exon, junction, DER) level analyses, the top hits (as determined by p-value) are compared. Results are plotted such that the points falling along the identity line would indicate complete agreement between the top hits of each analysis.

```
## Set colors
trop <- RSkittleBrewer('tropical')[c(1, 2, 3)]

## Obtain and sort p-values for genes
p.mod1 <- res_gene$p.mod
names(p.mod1) <- rownames(res_gene)
p.mod1.sort <- p.mod1[order(p.mod1)]

## Obtain and sort p-values for genes derived from exons
p.mod2 <- p_exon_gene
p.mod2.sort <- p.mod2[order(p.mod2)]
```

```
## Obtain and sort p-values for genes derived from junctions
p.mod3 <- p_jx_gene
p.mod3.sort <- p.mod3[order(p.mod3)]

## Obtain and sort p-values for genes derived from DER
p.mod4 <- p_DER_gene
p.mod4.sort <- p.mod4[order(p.mod4)]


## Overlap of features:
    ## gene level and exon level
table(names(p.mod1.sort) %in% names(p.mod2.sort))
```

```
##
##  TRUE
## 17874
```

```
    ## gene level and junction level
table(names(p.mod1.sort) %in% names(p.mod3.sort))
```

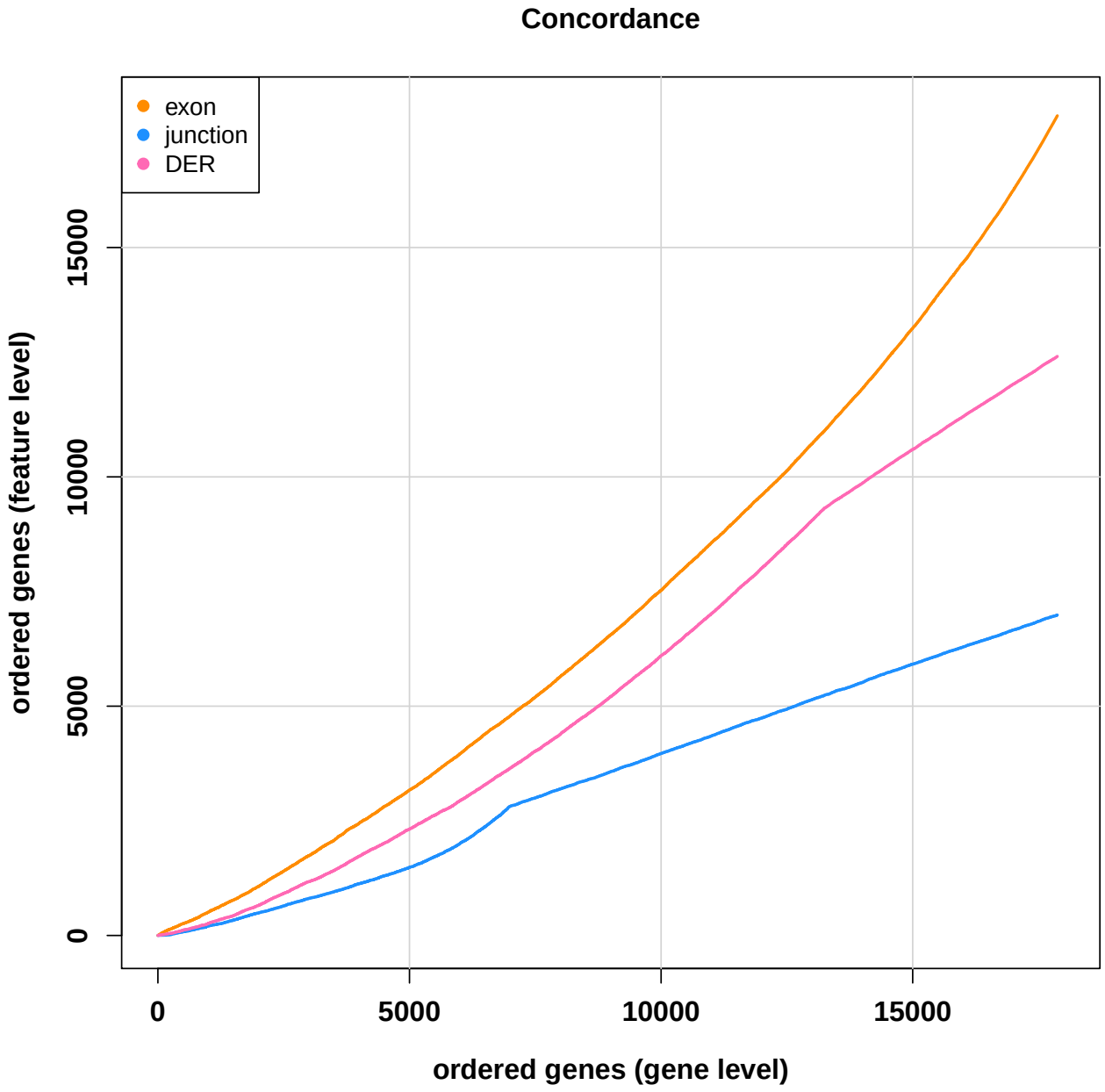
```
##
## FALSE  TRUE
## 10885  6989
```

```
    ## gene level and DER level
table(names(p.mod1.sort) %in% names(p.mod4.sort))
```

```
##
## FALSE  TRUE
##  5246 12628
```

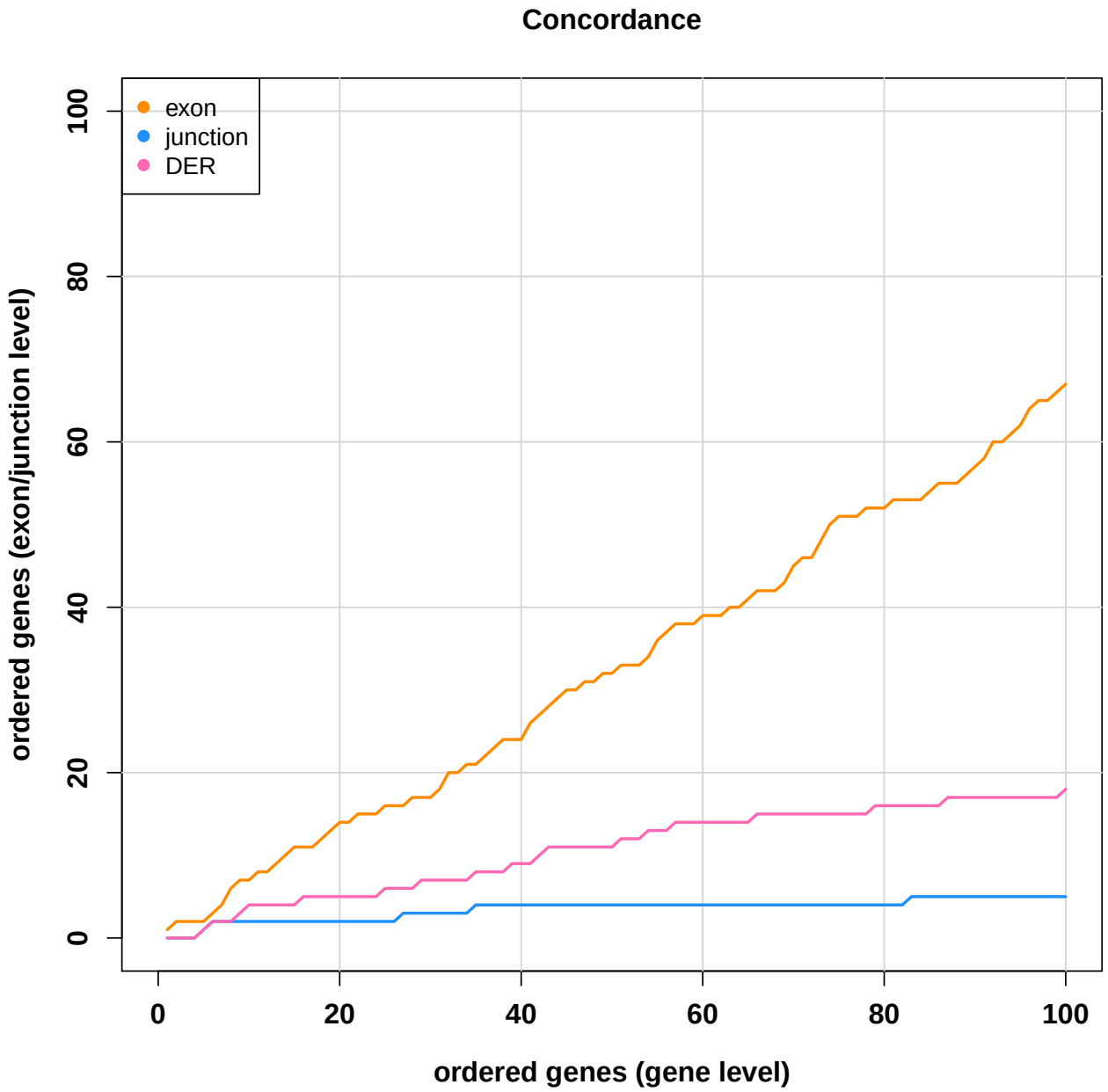
```
conc_exon <- NULL
conc_jx <- NULL
conc_DER <- NULL
for(i in seq_len(length(p.mod1.sort))) {
    conc_exon[i] <- sum(names(p.mod1.sort)[1:i] %in% names(p.mod2.sort)[1:i])
    conc_jx[i] <- sum(names(p.mod1.sort)[1:i] %in% names(p.mod3.sort)[1:i])
    conc_DER[i] <- sum(names(p.mod1.sort)[1:i] %in% names(p.mod4.sort)[1:i])
}


## All genes
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(seq(1:length(p.mod1.sort)), conc_exon,
     type = 'l', las = 0,
     xlim = c(0, 18000),
     ylim = c(0, 18000),
     xlab = 'ordered genes (gene level)',
     ylab = 'ordered genes (feature level)',
     main = 'Concordance')
for(k in 1:3){
```

```
    abline(v = k * 5000, cex = 0.5, col = 'lightgrey')
    abline(h = k * 5000, cex = 0.5, col = 'lightgrey')
}
points(seq(1:length(p.mod1.sort)), conc_jx, type = 'l', lwd = 2, col = trop[2])
lines(seq(1:length(p.mod1.sort)), conc_exon, lwd = 2,  col = trop[1])
lines(seq(1:length(p.mod1.sort)), conc_DER, lwd = 2,  col = trop[3])
legend('topleft', pch = 19, col = trop[c(1, 2, 3)], c("exon", "junction", "DER"))
```

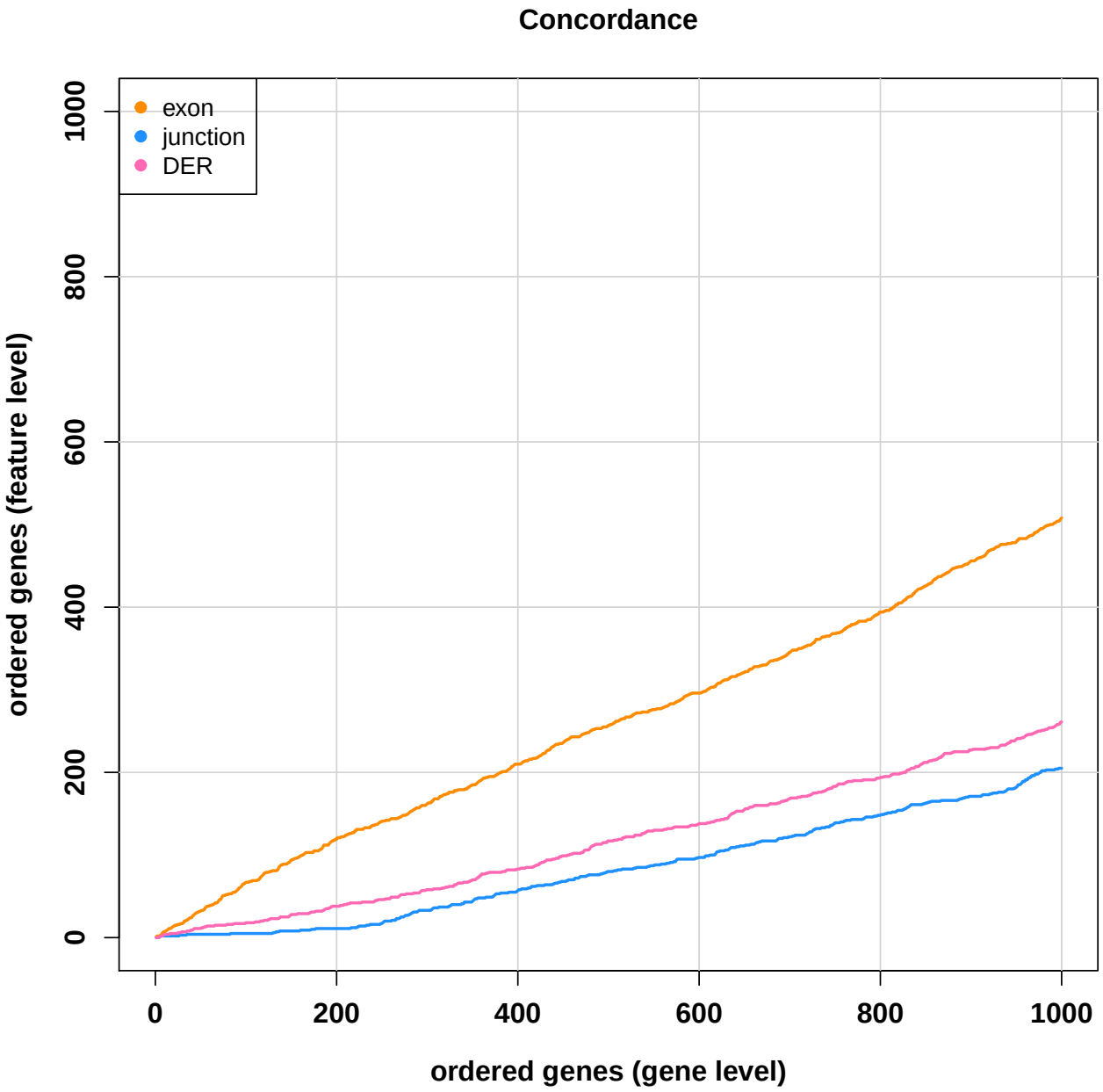



```
## Top 100 genes
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(seq(1:length(p.mod1.sort[1:100])), conc_exon[1:100],
     type = 'l',
     xlim = c(0, 100),
     ylim = c(0, 100),
```

```
     xlab = 'ordered genes (gene level)',
     ylab = 'ordered genes (exon/junction level)',
     main = 'Concordance')
for(k in 1:5){
    abline(v = k * 20, cex = 0.5, col = 'lightgrey')
    abline(h = k * 20, cex = 0.5, col = 'lightgrey')
}
points(seq(1:length(p.mod1.sort[1:100])), conc_jx[1:100], type = 'l', lwd = 2, col = trop[2])
lines(seq(1:length(p.mod1.sort[1:100])), conc_exon[1:100], lwd = 2,  col = trop[1] )
lines(seq(1:length(p.mod1.sort[1:100])), conc_DER[1:100], lwd = 2,  col = trop[3])
legend('topleft', pch = 19, col = trop[c(1, 2, 3)], c("exon", "junction", "DER"))
```



```
## Top 1,000 genes
par(font.lab = 2, cex.lab = 1.2, font.axis = 2, cex.axis = 1.2)
plot(seq(1:length(p.mod1.sort[1:1000])), conc_exon[1:1000],
     type = 'l',
     xlim = c(0, 1000),
     ylim = c(0, 1000),
     xlab = 'ordered genes (gene level)',
     ylab = 'ordered genes (feature level)',
     main = 'Concordance')
for(k in 1:5){
    abline(v = k * 200, cex = 0.5, col = 'lightgrey')
    abline(h = k * 200, cex = 0.5, col = 'lightgrey')
}
points(seq(1:length(p.mod1.sort[1:1000])), conc_jx[1:1000], type = 'l', lwd = 2, col = trop[2])
lines(seq(1:length(p.mod1.sort[1:1000])), conc_exon[1:1000], lwd = 2,  col = trop[1] )
lines(seq(1:length(p.mod1.sort[1:1000])), conc_DER[1:1000], lwd = 2,  col = trop[3])
legend('topleft', pch = 19, col = trop[c(1, 2, 3)], c("exon", "junction", "DER"))
```



Concordance can also be calculated looking at the gene ontology (GO) groups identified from the gene and exon level analyses. Again, we plot the agreement between the two analyses such that complete agreement between the two analyses would fall along the identity line.

Reproducibility

This analysis report was made possible thanks to:

- R (Collado-Torres, Nellore, and Leek, 2016)
- *BiocStyle* (R Core Team, 2016)
- *derfinder* (Alexa and Rahnenfuhrer, 2016)
- *devtools* (Allaire, Cheng, Xie, McPherson, et al., 2016)
- *edgeR* (with contributions from Andrew J. Bass, Dabney, and Robinson, 2015)

- *knitcitations* (Bengtsson, 2016)
- *matrixStats* (Boettiger, 2016)
- *qvalue* (Collado-Torres, Nellore, Frazee, Wilks, et al., 2016)
- *recount* (Frazee, 2016)
- *rmarkdown* (Morgan, Obenchain, Hester, and Pagès, 2016)
- *RSkittleBrewer* (Oleś, Morgan, and Huber, 2016)
- *SummarizedExperiment* (Robinson, McCarthy, and Smyth, 2010)
- *topGO* (Law, Chen, Shi, and Smyth, 2014)
- *limma* (Wickham and Chang, 2016)

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 1.0. 2016. URL: https://CRAN.R-project.org/package=rmarkdown.

[3] J. D. S. with contributions from Andrew J. Bass, A. Dabney and D. Robinson. qvalue: Q-value estimation for false discovery rate control. R package version 2.5.2. 2015. URL: http://github.com/jdstorey/qvalue.

[4] 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.

[5] C. Boettiger. knitcitations: Citations for 'Knitr' Markdown Files. R package version 1.0.7.1. 2016. URL: https://github.com/cboettig/knitcitations.

[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] L. Collado-Torres, A. Nellore and J. T. Leek. recount: Explore and download data from the recount project. R package version 0.99.0. 2016. URL: https://github.com/leekgroup/recount.

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

[9] 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.

[10] M. Morgan, V. Obenchain, J. Hester and H. Pagès. SummarizedExperiment: SummarizedExperiment container. R package version 1.3.7. 2016.

[11] 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.

[12] 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/.

[13] 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.

[14] 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.

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

```
## Time difference of 16.84014 mins
```

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

```
## 2016-07-22 18:28:26
```

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

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

```
##  setting  value
##  version  R version 3.3.0 (2016-05-03)
##  system   x86_64, darwin13.4.0
##  ui       RStudio (0.99.902)
##  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  2013-05-03 CRAN (R 3.1.0)
##  AnnotationDbi        * 1.35.4   2016-07-08 Bioconductor
##  bibtex                 0.4.0    2014-12-31 CRAN (R 3.3.0)
##  Biobase              * 2.33.0   2016-05-27 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 3.2.0)
##  cluster                2.0.4    2016-04-18 CRAN (R 3.3.0)
##  codetools              0.2-14   2015-07-15 CRAN (R 3.3.0)
##  colorspace             1.2-6    2015-03-11 CRAN (R 3.1.3)
##  data.table             1.9.6    2015-09-19 CRAN (R 3.2.0)
##  DBI                    0.4-1    2016-05-08 CRAN (R 3.2.5)
##  derfinder            * 1.7.9    2016-06-20 Bioconductor
##  derfinderHelper        1.7.3    2016-05-27 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)
##  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 3.2.0)
##  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 3.1.3)
##  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
##  ggplot2                2.1.0    2016-03-01 CRAN (R 3.2.4)
##  GO.db                * 3.3.0    2016-05-25 Bioconductor
##  graph                * 1.51.0   2016-05-27 Bioconductor
##  gridExtra              2.2.1    2016-02-29 CRAN (R 3.2.4)
##  gtable                 0.2.0    2016-02-26 CRAN (R 3.2.3)
##  Hmisc                  3.17-4   2016-05-02 CRAN (R 3.2.5)
##  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 3.2.0)
##  knitcitations        * 1.0.7.1  2016-06-06 Github (cboettig/knitcitations@76c128a)
##  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 3.2.3)
##  limma                * 3.29.16  2016-07-21 Bioconductor
##  locfit                 1.5-9.1  2013-04-20 CRAN (R 3.1.0)
##  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 3.2.5)
##  memoise                1.0.0    2016-01-29 CRAN (R 3.2.3)
##  munsell                0.4.3    2016-02-13 CRAN (R 3.2.3)
##  nnet                   7.3-12   2016-02-02 CRAN (R 3.3.0)
##  org.Hs.eg.db         * 3.3.0    2016-05-25 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 3.2.3)
##  RColorBrewer           1.1-2    2014-12-07 CRAN (R 3.1.2)
##  Rcpp                   0.12.6   2016-07-19 CRAN (R 3.3.0)
##  RCurl                  1.95-4.8 2016-03-01 CRAN (R 3.2.4)
##  recount              * 0.99.0   2016-06-06 Github (leekgroup/recount@bfa4681)
##  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 3.1.2)
##  RJSONIO                1.3-0    2014-07-28 CRAN (R 3.2.0)
##  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 3.3.0)
##  Rsamtools              1.25.0   2016-05-27 Bioconductor
##  RSkittleBrewer       * 1.1      2016-06-07 Github (alyssafrazee/RSkittleBrewer@230d1d0)
##  RSQLite                1.0.0    2014-10-25 CRAN (R 3.1.2)
##  rstudioapi             0.6      2016-06-27 CRAN (R 3.3.0)
##  rtracklayer            1.33.10  2016-07-14 Bioconductor
##  S4Vectors            * 0.11.10  2016-07-21 Bioconductor
##  scales                 0.4.0    2016-02-26 CRAN (R 3.2.3)
##  SparseM              * 1.7      2015-08-15 CRAN (R 3.2.0)
##  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)
##  topGO                * 2.25.0   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 3.2.4)
##  xtable               1.8-2    2016-02-05 CRAN (R 3.2.3)
##  XVector              0.13.6   2016-07-08 Bioconductor
##  yaml                 2.1.13   2014-06-12 CRAN (R 3.3.0)
##  zlibbioc             1.19.0   2016-05-27 Bioconductor
```