OUTRIDER - OUTlier in RNA-Seq fInDER

***Felix Brechtmann*[1], *Christian Mertes*[1], *Agne Matuseviciute*[1], *Vicente Yepez*[1,2], *Julien Gagneur*[1,2]**

[1] Technical University Munich, Department of Informatics, Munich, Germany
[2] Quantitative Biosciences Munich, Gene Center, Ludwig-Maximilians Universität München, Munich, Germany

May 1, 2024

Abstract

In the field of diagnostics of rare diseases, RNA-seq is emerging as an important and complementary tool for whole exome and whole genome sequencing. *OUTRIDER* is a framework that detects aberrant gene expression within a group of samples. It uses the negative binomial distribution which is fitted for each gene over all samples. We additionally provide an autoencoder, which automatically controls for co-variation before fitting. After fitting, each sample can be tested for aberrantly expressed genes. Furthermore, *OUTRIDER* provides functionality to easily filter unexpressed genes, to analyse the data as well as to visualize the results.

If you use *OUTRIDER* in published research, please cite:

Brechtmann F*, Mertes C*, Matuseviciute A*, Yepez V, Avsec Z, Herzog M, Bader D M, Prokisch H, Gagneur J; **OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data**; *AJHG*; 2018; DOI: https://doi.org/10.1016/j.ajhg.2018.10.025

Contents

1 Introduction

OUTRIDER (OUTlier in RNA-seq fInDER) is a tool for finding aberrantly expressed genes in RNA-seq samples. It does so by fitting a negative binomial model to RNA-seq read counts, correcting for variations in sequencing depth and apparent co-variations across samples. Read counts that significantly deviate from the distribution are detected as outliers. *OUTRIDER* makes use of an autoencoder to control automatically for confounders within the data. A scheme of this approach is given in Figure 1.

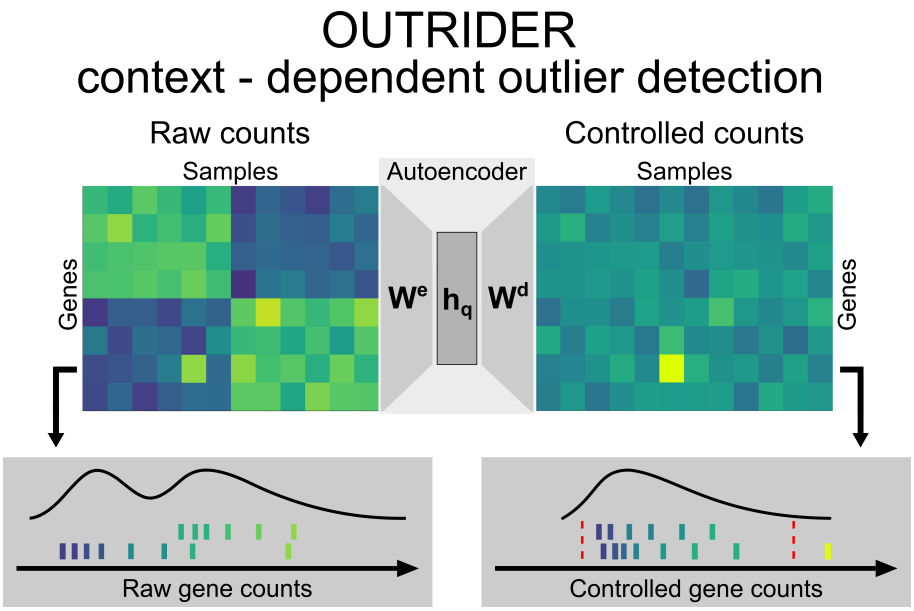


Figure 1: Context-dependent outlier detection. The algorithm identifies gene expression outliers whose read counts are significantly aberrant given the co-variations typically observed across genes in an RNA sequencing data set. This is illustrated by a read count (left panel, fifth column, second row from the bottom) that is exceptionally high in the context of correlated samples (left six samples) but not in absolute terms for this given gene. To capture commonly seen biological and technical contexts, an autoencoder models co-variations in an unsupervised fashion and predicts read count expectations. By comparing the earlier mentioned read count with these context-dependent expectations, it is revealed as exceptionally high (right panel). The lower panels illustrate the distribution of read counts before and after applying the correction for the relevant gene. The red dotted lines depict significance cutoffs.

Differential gene expression analysis from RNA-seq data is well-established. The packages *DESeq2*[1] or *edgeR*[2] provide effective workflows and preprocessing steps to perform differential gene expression analysis. However, these methods aim at detecting significant differences between groups of samples. In contrast, *OUTRIDER* aims at detecting outliers within a given population. A scheme of this difference is given in figure 2.

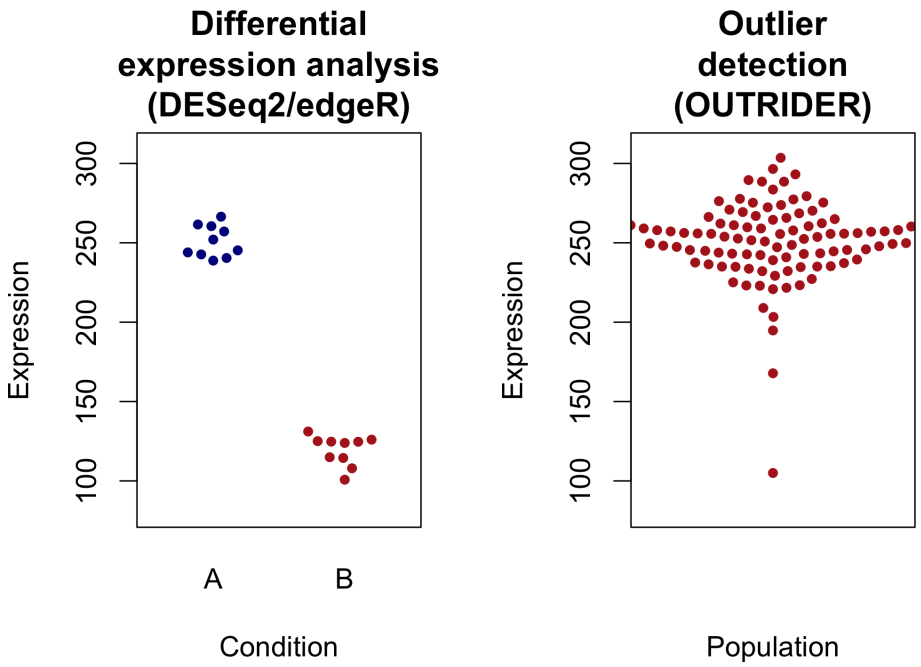


Figure 2: Scheme of workflow differences. Differences between differential gene expression analysis and outlier detection.

2 Prerequisites

To get started on the preprocessing step, we recommend to read the introductions of *DESeq2*[1], *edgeR*[2] or the RNA-seq workflow from Bioconductor: *rnaseqGene*. In brief, one usually starts with the raw FASTQ files from the RNA sequencing run. Those are then aligned to a given reference genome. At the time of writing (October 2018), we recommend the STAR aligner[3]. After obtaining the aligned BAM files, one can map the reads to exons or genes of a GTF annotation file using HT-seq or by using `summarizedOverlaps` from *GenomicAlignments*. The resulting count table can then be loaded into the *OUTRIDER* package as we will describe below.

3 A quick tour

Here we assume that we already have a count table and no additional preprocessing needs to be done. We can start and obtain results with 3 commands. First, create an *OutriderDataSet* from a count table or a Summarized Experiment object. Second, run the full pipeline using the command `OUTRIDER`. In the third and last step the results table is extracted from the *OutriderDataSet* with the `results` function. Furthermore, analysis plots that are described in section 5 can be directly created from the *OutriderDataSet* object.

```
library(OUTRIDER)
```

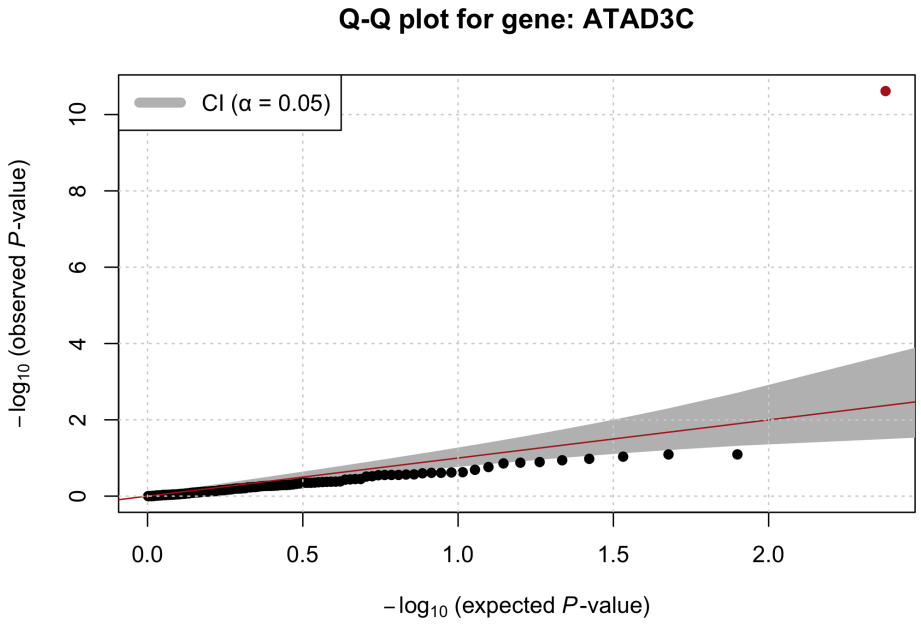
```
# get data
ctsFile <- system.file('extdata', 'KremerNBaderSmall.tsv',
        package='OUTRIDER')
ctsTable <- read.table(ctsFile, check.names=FALSE)
ods <- OutriderDataSet(countData=ctsTable)

# filter out non expressed genes
ods <- filterExpression(ods, minCounts=TRUE, filterGenes=TRUE)

# run full OUTRIDER pipeline (control, fit model, calculate P-values)
ods <- OUTRIDER(ods)
```

```
## [1] "Wed May  1 01:58:59 2024: Initial PCA loss: 4.73997327486604"
## [1] "Wed May  1 01:59:01 2024: Iteration: 1 loss: 4.19354069417627"
## [1] "Wed May  1 01:59:03 2024: Iteration: 2 loss: 4.17555743183664"
## [1] "Wed May  1 01:59:05 2024: Iteration: 3 loss: 4.16656490809821"
## [1] "Wed May  1 01:59:06 2024: Iteration: 4 loss: 4.1612420443096"
## [1] "Wed May  1 01:59:07 2024: Iteration: 5 loss: 4.15793603698054"
## [1] "Wed May  1 01:59:08 2024: Iteration: 6 loss: 4.15557302727805"
## [1] "Wed May  1 01:59:09 2024: Iteration: 7 loss: 4.15389125937973"
## [1] "Wed May  1 01:59:10 2024: Iteration: 8 loss: 4.15228926127285"
## [1] "Wed May  1 01:59:11 2024: Iteration: 9 loss: 4.15119080776361"
## [1] "Wed May  1 01:59:12 2024: Iteration: 10 loss: 4.15007090742281"
## [1] "Wed May  1 01:59:12 2024: Iteration: 11 loss: 4.14924416489662"
## [1] "Wed May  1 01:59:13 2024: Iteration: 12 loss: 4.1487151214388"
## [1] "Wed May  1 01:59:14 2024: Iteration: 13 loss: 4.14784235003435"
## [1] "Wed May  1 01:59:14 2024: Iteration: 14 loss: 4.14743207189597"
## [1] "Wed May  1 01:59:15 2024: Iteration: 15 loss: 4.1473639411029"
## Time difference of 15.26398 secs
## [1] "Wed May  1 01:59:15 2024: 15 Final nb-AE loss: 4.1473639411029"
```

```
# results (only significant)
res <- results(ods)
head(res)
```

```
##    geneID sampleID       pValue      padjust zScore  l2fc rawcounts
##    <char>   <char>        <num>        <num>  <num> <num>     <int>
## 1: ATAD3C  MUC1360 2.425660e-11 1.349316e-07   5.29  1.88       948
## 2: NBPF15  MUC1351 7.905889e-10 4.397790e-06   5.76  0.77      7591
## 3:  MSTO1  MUC1367 3.889718e-09 2.163724e-05  -6.23 -0.81       761
## 4:  HDAC1  MUC1350 1.845115e-08 1.026378e-04  -5.90 -0.78      2215
## 5:  DCAF6  MUC1374 6.912880e-08 3.845411e-04  -5.68 -0.62      2348
## 6: NBPF16  MUC1351 2.764418e-07 7.688782e-04   4.81  0.67      4014
##    meanRawcounts normcounts meanCorrected  theta aberrant AberrantBySample
##            <num>      <num>         <num>  <num>   <lgcl>            <num>
## 1:         82.29     248.24         67.32  16.52     TRUE                1
```

```
## 2:       4224.88    7010.84       4121.18 112.55     TRUE               2
## 3:       1327.87     728.83       1276.05 152.21     TRUE               1
## 4:       3805.56    2126.88       3648.97 136.16     TRUE               1
## 5:       4869.53    3077.43       4724.33 195.09     TRUE               1
## 6:       2459.90    3813.01       2402.34 108.08     TRUE               2
##    AberrantByGene padj_rank
##             <num>     <num>
## 1:              1         1
## 2:              1         1
## 3:              1         1
## 4:              1         1
## 5:              1         1
## 6:              1         2

# example of a Q-Q plot for the most significant outlier
plotQQ(ods, res[1, geneID])
```



4 An *OUTRIDER* analysis in detail

Apart from running the full pipeline using the single wrapper function `OUTRIDER`, the analysis can also be run step by step. The wrapper function does not include any preprocessing functions. Discarding non expressed genes or samples failing quality measurements should be done manually before running the `OUTRIDER` function or starting the analysis pipeline.

In this section we will explain the analysis functions step by step.

For this tutorial we will use the rare disease data set from Kremer *et al.*[4]. For testing purposes, this package contains a small subset of it.

4.1 OutriderDataSet

To use *OUTRIDER* create an *OutriderDataSet*, which derives from a RangedSummarizedExperiment object. The *OutriderDataSet*can be created by supplying a count matrix and optional sample annotation matrices. Alternatively, an existing Summarized experiment object from other Biocnductor backages can be used.

```
# small testing data set
odsSmall <- makeExampleOutriderDataSet(dataset="Kremer")

# full data set from Kremer et al.
baseURL <- paste0("https://static-content.springer.com/esm/",
        "art%3A10.1038%2Fncomms15824/MediaObjects/")
count_URL <- paste0(baseURL, "41467_2017_BFncomms15824_MOESM390_ESM.txt")
anno_URL  <- paste0(baseURL, "41467_2017_BFncomms15824_MOESM397_ESM.txt")

ctsTable <- read.table(count_URL, sep="\t")
annoTable <- read.table(anno_URL, sep="\t", header=TRUE)
annoTable$sampleID <- annoTable$RNA_ID

# create OutriderDataSet object
ods <- OutriderDataSet(countData=ctsTable, colData=annoTable)
```

4.2 Preprocessing

It is recommended to preprocess the data before fitting. Our model requires that for every gene at least one sample has a non-zero count and that we observe at least one read for every 100 samples. Therefore, all genes that are not expressed must be discarded.

We provide the function `filterExpression` to remove genes that have low FPKM (Fragments Per Kilobase of transcript per Million mapped reads) expression values. The needed annotation to estimate FPKM values from the counts should be the same as for the counting. Here, we normalize by the total exon length of a gene. To do so the joint lenght of all exons needs to be provided. When providing a gtf, gff or TxDb object to the `filterExpression`, we extract this information automatically. But therfore the geneID's of the count table and the gtf need to match.

By default the cutoff is set to an FPKM value of one and only the filtered *OutriderDataSet* object is returned. If required, the FPKM values can be stored in the *OutriderDataSet* object and the full object can be returned to visualize the distribution of reads before and after filtering.

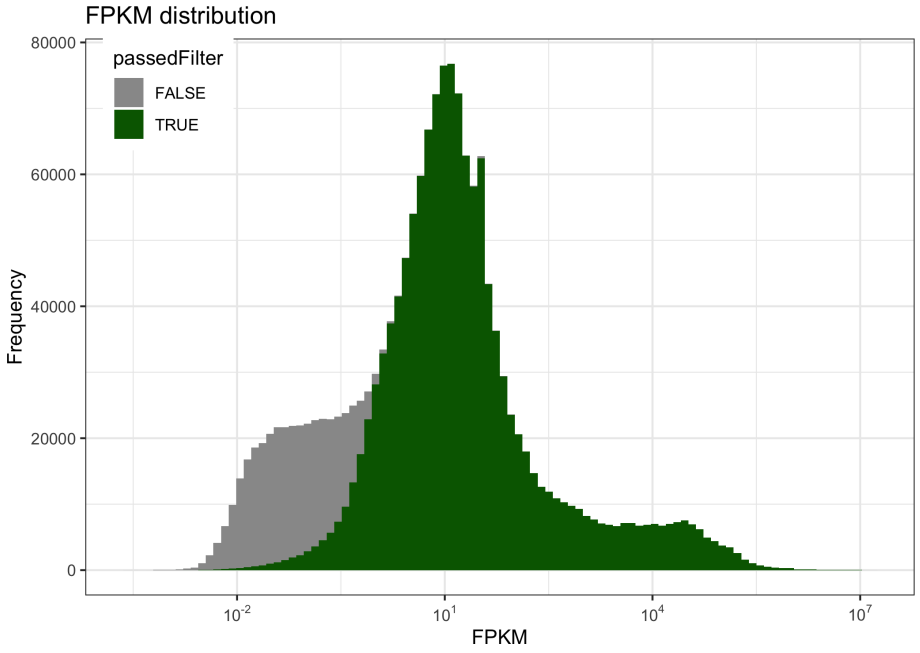
```
# get annotation
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(org.Hs.eg.db)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
map <- select(org.Hs.eg.db, keys=keys(txdb, keytype = "GENEID"),
        keytype="ENTREZID", columns=c("SYMBOL"))
```

However, the `TxDb.Hsapiens.UCSC.hg19.knownGene` contains only well annotated genes. This annotation will miss a lot of genes captured by RNA-seq. To include all predicted annotations as well as non-coding RNAs please download the txdb object from our homepage[1] or create it yourself from the UCSC website[2,3].

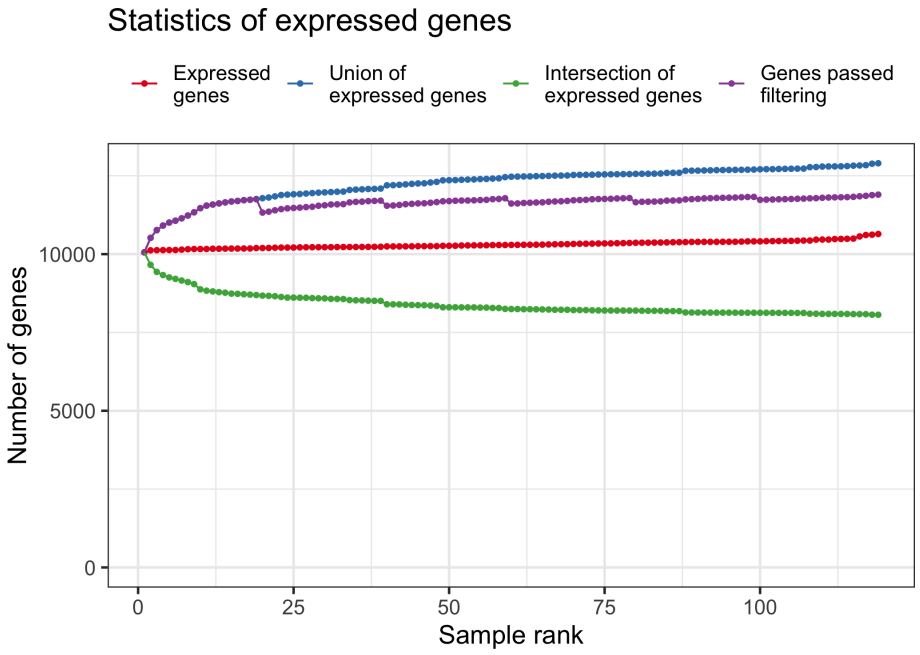
```
try({
    library(RMariaDB)
    library(AnnotationDbi)
    con <- dbConnect(MariaDB(), host='genome-mysql.cse.ucsc.edu',
            dbname="hg19", user='genome')
    map <- dbGetQuery(con, 'select kgId AS TXNAME, geneSymbol from kgXref')

    txdbUrl <- paste0("https://cmm.in.tum.de/public/",
            "paper/mitoMultiOmics/ucsc.knownGenes.db")
    download.file(txdbUrl, "ucsc.knownGenes.db")
    txdb <- loadDb("ucsc.knownGenes.db")

})
```

```
# calculate FPKM values and label not expressed genes
ods <- filterExpression(ods, txdb, mapping=map,
        filterGenes=FALSE, savefpkm=TRUE)

# display the FPKM distribution of counts.
plotFPKM(ods)
```

[1] https://cmm.in.tum.de/public/paper/mitoMultiOmics/ucsc.knownGenes.db
[2] https://genome.ucsc.edu/cgi-bin/hgTables
[3] http://genomewiki.ucsc.edu/index.php/Genes_in_gtf_or_gff_format




```
# display gene filter summary statistics
plotExpressedGenes(ods)
```

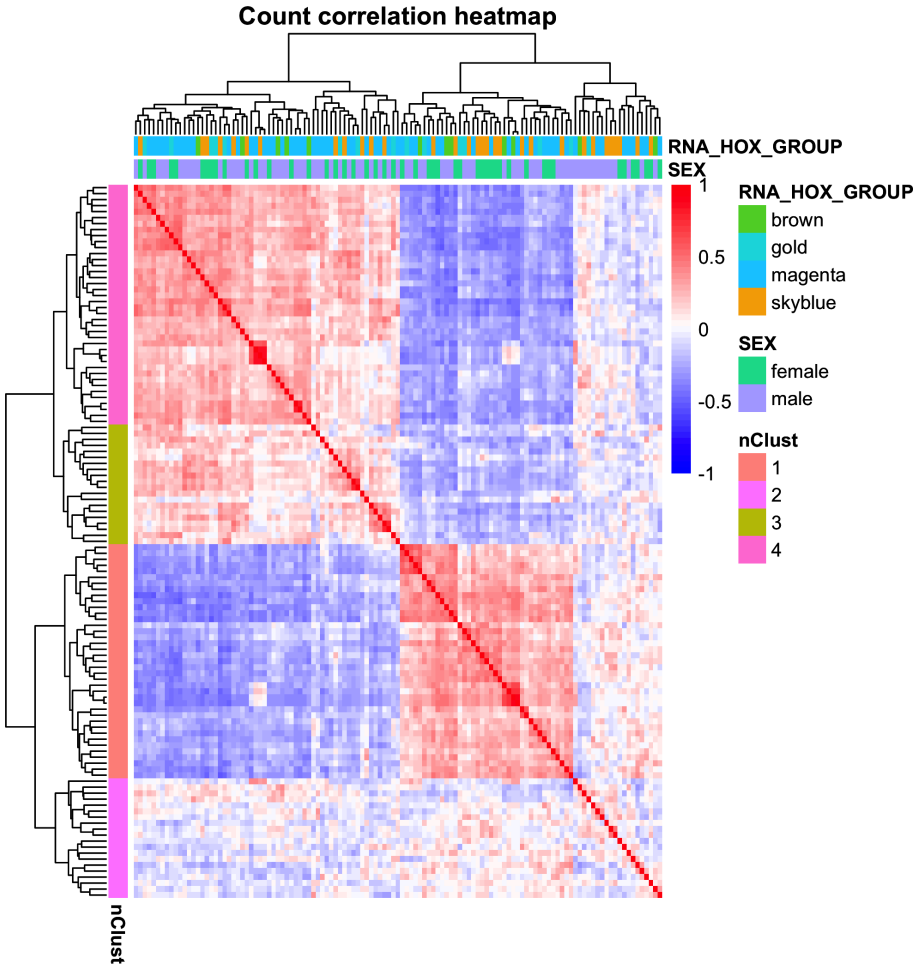



```
# do the actual subsetting based on the filtering labels
ods <- ods[mcols(ods)$passedFilter,]
```

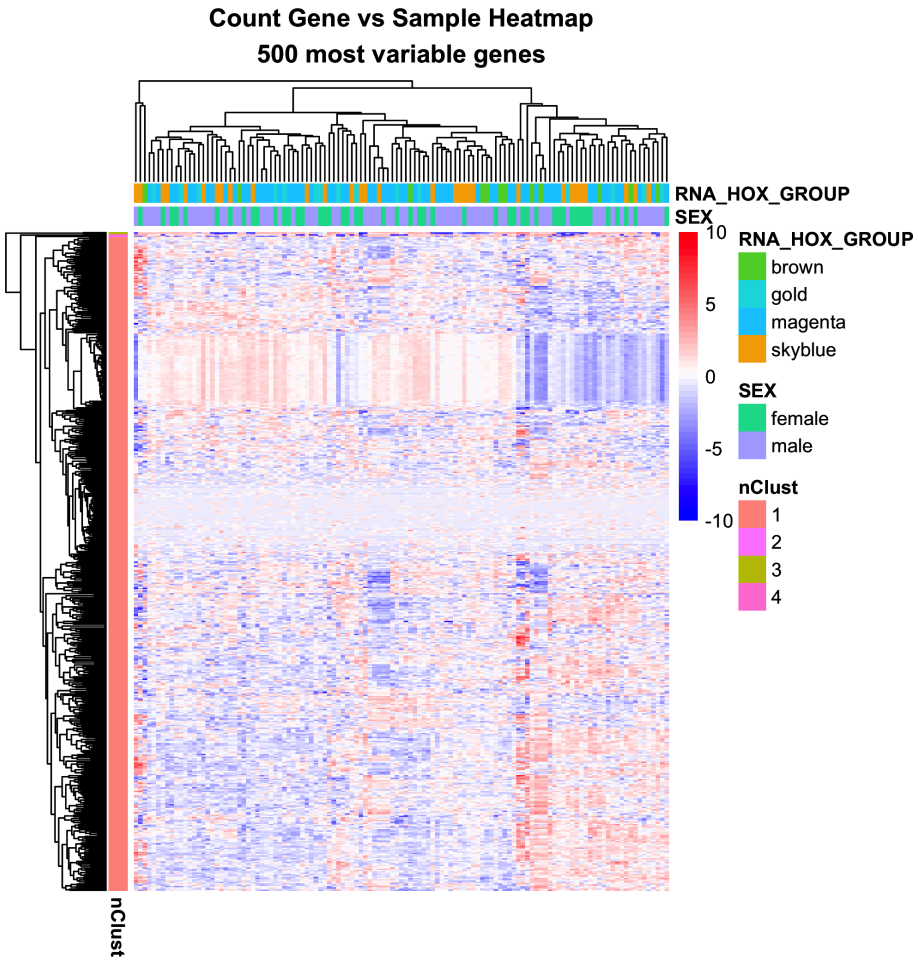
4.3 Controlling for Confounders

The next step in any analysis workflow is to visualize the correlations between samples. In most RNA-seq experiments correlations between the samples can be observed. These are often due to technical confounders (e.g. sequencing batch) or biological confounders (e.g. sex, age). These confounders can adversely affect the detection of aberrant features. Therefore, we provide options to control for them.

```
# Heatmap of the sample correlation
# it can also annotate the clusters resulting from the dendrogram
ods <- plotCountCorHeatmap(ods, colGroups=c("SEX", "RNA_HOX_GROUP"),
        normalized=FALSE, nRowCluster=4)
```




```
# Heatmap of the gene/sample expression
ods <- plotCountGeneSampleHeatmap(ods, colGroups=c("SEX", "RNA_HOX_GROUP"),
        normalized=FALSE, nRowCluster=4)
```

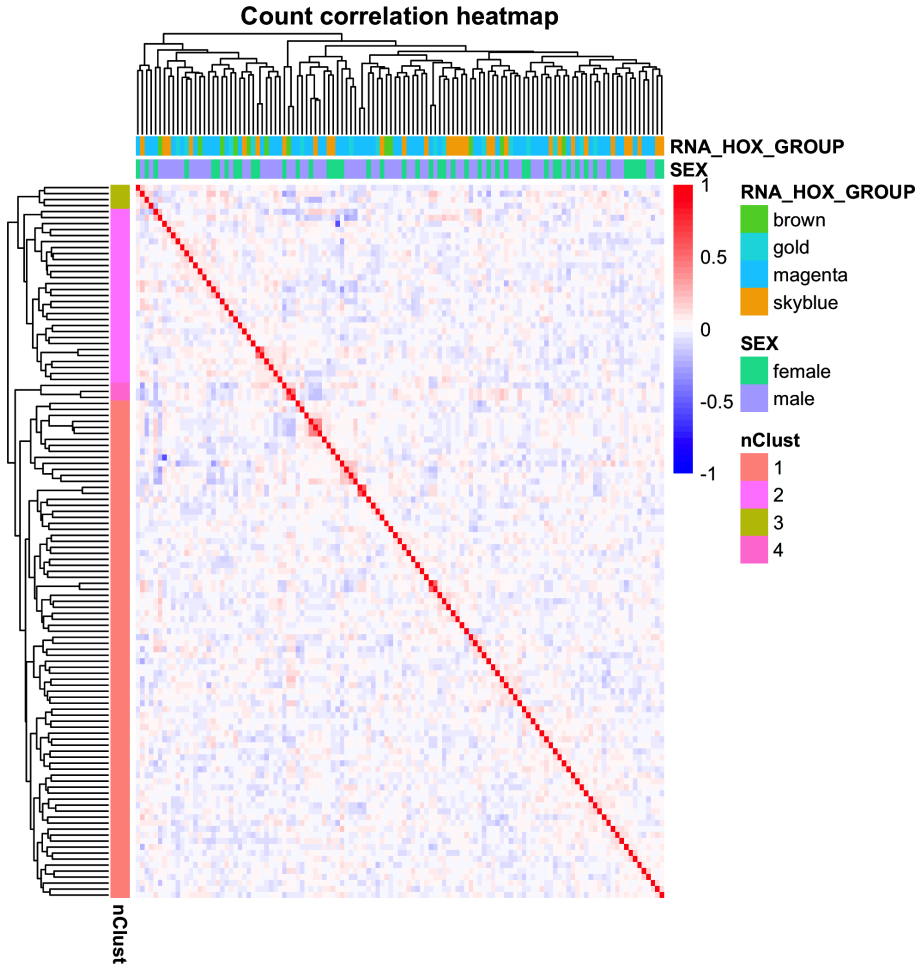


We have different ways to control for confounders present in the data. The first and standard way is to calculate the `sizeFactors` as done in *DESeq2*[1].

Additionally, the `controlForConfounders` function calls an autoencoder that automatically controls for confounders present in the data. Therefore an encoding dimension q needs to be set or the default value 20 is used. The optimal value of q can be determined using the `findEncodingDim` function. After controlling for confounders, the heatmap should be plotted again. If it worked, no batches should be present and the correlations between samples should be reduced and close to zero.

```
# automatically control for confounders
# we use only 3 iterations to make the vignette faster. The default is 15.
ods <- estimateSizeFactors(ods)
ods <- controlForConfounders(ods, q=21, iterations=3)

## [1] "Wed May  1 01:59:32 2024: Initial PCA loss: 5.95086157341829"
## [1] "Wed May  1 02:00:03 2024: Iteration: 1 loss: 5.37003868338894"
## [1] "Wed May  1 02:00:18 2024: Iteration: 2 loss: 5.35604952994249"
## [1] "Wed May  1 02:00:30 2024: Iteration: 3 loss: 5.34989239860339"
## Time difference of 54.75963 secs
```

```
## [1] "Wed May  1 02:00:30 2024: 3 Final nb-AE loss: 5.34989239860339"

# Heatmap of the sample correlation after controlling
ods <- plotCountCorHeatmap(ods, normalized=TRUE,
        colGroups=c("SEX", "RNA_HOX_GROUP"))
```



Alternatively, other methods can be used to control for confounders. In addition to the *autoencoder*, we implemented a PCA based approach. The PCA implementation can be utilized by setting `implementation="pca"`. Also PEER can be used together with the OUTRIDER framework. A detailed description on how to do this can be found in section 6.1. Furthermore, any other method can be used by providing the `normalizationFactor` matrix. This matrix must be computed beforehand using the appropriate method. Its purpose is to normalize for technical effects or control for additional expression patterns.

4.4 Finding the right encoding dimension q

In the previous section, we fixed the encoding dimension $q = 21$. But having the right encoding dimension is crucial in finding outliers in the data. On the one hand, if q is too big the autoencoder will learn the identity matrix and will overfit the data. On the other hand, if q is too small the autoencoder cannot learn the necessary covariates existing in the data. Therefore, it is recommended for any new dataset to estimate the optimal encoding dimension to gain the best performance. With the function `findEncodingDim` one can find the optimal encoding dimension. To this end, we artificially introduce corrupted counts randomly into the dataset and monitor the performance calling those corrupted counts. The optimal dimension q is then selected as the dimension maximizing the area under the precision-recall curve for identifying corrupted counts.

```
# find the optimal encoding dimension q
ods <- findEncodingDim(ods)

# visualize the hyper parameter optimization
plotEncDimSearch(ods)
```

Since this function runs a full OUTRIDER fit for a range of encoding dimensions, it is quite CPU intensive, but can increase the overall performance of the autoencoder and is recommended for any data set. If q is not provided by the user, it will be estimated based on the number of samples.

4.4.1 Excluding samples from the autoencoder fit

Since OUTRIDER expects that each sample within the population is independent of all others, replicates could mask effects specific to this sample. This is also true if trios are present in the data, where the parents can be seen as biological replicates. Here, we recommend to exclude the sample of interest or the replicates from the fitting. Later on, for all samples P-values are calculated.

In this rare disease data set we know that two samples (MUC1344 and MUC1365) have the same defect. To exclude one or both of them, we can use the `sampleExclusionMask` function.

```
# set exclusion mask
sampleExclusionMask(ods) <- FALSE
sampleExclusionMask(ods[,"MUC1365"]) <- TRUE

# check which samples are excluded from the autoencoder fit
sampleExclusionMask(ods)
```

```
##   35834   57415   61695   61982   65937   66623   69245   69248   69456
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
##   70038   70041   72748   74123   74172   76619   76620   76621   76622
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
```

```
##   76623   76624   76625   76626   76627   76628   76629   76630   76631
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
##   76632   76633   76635   76636   76637   76638 MUC0486 MUC0487 MUC0488
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC0489 MUC0490 MUC0491 MUC1342 MUC1343 MUC1344 MUC1345 MUC1346 MUC1347
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1348 MUC1349 MUC1350 MUC1351 MUC1352 MUC1354 MUC1355 MUC1357 MUC1358
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1359 MUC1360 MUC1361 MUC1362 MUC1363 MUC1364 MUC1365 MUC1367 MUC1368
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE    TRUE   FALSE   FALSE
## MUC1369 MUC1370 MUC1371 MUC1372 MUC1373 MUC1374 MUC1375 MUC1376 MUC1377
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1378 MUC1379 MUC1380 MUC1381 MUC1382 MUC1383 MUC1384 MUC1390 MUC1391
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1392 MUC1393 MUC1394 MUC1395 MUC1396 MUC1397 MUC1398 MUC1400 MUC1401
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1402 MUC1403 MUC1404 MUC1405 MUC1407 MUC1408 MUC1409 MUC1410 MUC1411
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1412 MUC1413 MUC1414 MUC1415 MUC1416 MUC1417 MUC1418 MUC1419 MUC1420
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1421 MUC1422 MUC1423 MUC1424 MUC1425 MUC1426 MUC1427 MUC1428 MUC1429
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## MUC1436 MUC1437
##   FALSE   FALSE
```

4.5 Fitting the negative binomial model

The fit of the negative binomial model is done during the autoencoder fitting. This step is only needed if alternative methods to control the data are used. To fit the dispersion and the mean, the `fit` function is applied to the *OutriderDataSet*.

```
# fit the model when alternative methods where used in the control step
ods <- fit(ods)
hist(theta(ods))
```

4.6 P-value calculation

After determining the fit parameters, two-sided P-values are computed using the following equation:

$$p_{ij} = 2 \cdot min \left\{ \frac{1}{2}, \sum_{0}^{k_{ij}} NB(\mu_{ij}, \theta_i), 1 - \sum_{0}^{k_{ij-1}} NB(\mu_{ij}, \theta_i) \right\}, \quad (1)$$

where the $\frac{1}{2}$ term handles the case of both terms exceeding 0.5, which can happen due to the discrete nature of counts. Here μ_{ij} are computed as the product of the fitted correction values from the autoencoder and the fitted mean adjustements. If required a one-sided test can be performed using the argument `alternative` and specifying 'less' or 'greater' depending on the research question. Multiple testing correction is done across all genes in a per-sample fashion using Benjamini-Yekutieli's false discovery rate method[5]. Alternatively, all adjustment methods supported by `p.adjust` can be used via the `method` argument.

```
# compute P-values (nominal and adjusted)
ods <- computePvalues(ods, alternative="two.sided", method="BY")
```

4.7 Z-score calculation

The Z-scores on the log transformed counts can be used for visualization, filtering, and ranking of samples. By running the `computeZscores` function, the Z-scores are computed and stored in the *OutriderDataSet* object. The Z-scores are calculated using:

$$z_{ij} = \frac{l_{ij} - \mu_j^l}{\sigma_j^l} \qquad (2)$$

$$l_{ij} = \log_2 (\frac{k_{ij} + 1}{c_{ij} + 1}),$$

where μ_j^l is the mean and σ_j^l the standard deviation of gene j and l_{ij} is the log transformed count after correction for confounders.

```
# compute the Z-scores
ods <- computeZscores(ods)
```

5 Results

The *OUTRIDER* package offers multiple ways to display the results. It creates a results table containing all the values computed during the analysis. Furthermore, it offers various plot functions that guide the user through the analysis.

5.1 Results table

The `results` function gathers all the previously computed values and combines them into one table.

```
# get results (default only significant, padj < 0.05)
res <- results(ods)
head(res)
```

```
##     geneID sampleID       pValue      padjust zScore  l2fc rawcounts
##     <char>   <char>        <num>        <num>  <num> <num>     <int>
## 1:  NUDT12    65937 1.714844e-22 2.021918e-17 -10.34 -7.97         0
## 2:   STAG2  MUC0490 3.750099e-21 4.421623e-16  -9.73 -1.87       622
## 3:  TALDO1  MUC1427 7.290799e-18 8.596350e-13  -9.43 -3.28       482
## 4:  MCOLN1  MUC1361 1.079028e-15 1.272248e-10  -8.80 -2.41       150
## 5:  RETSAT  MUC1374 1.175468e-15 1.385957e-10  -8.94 -3.64        76
## 6: CSNK2A1  MUC1358 2.794734e-15 1.647591e-10  -8.12 -0.67      2060
##    meanRawcounts normcounts meanCorrected  theta aberrant AberrantBySample
##            <num>      <num>         <num>  <num>   <lgcl>            <num>
## 1:        470.22       0.00        453.43  19.28     TRUE                2
## 2:       4184.49    1108.88       4066.04  83.41     TRUE                6
## 3:       4699.52     469.42       4556.79  27.85     TRUE                1
## 4:        912.70     157.77        841.06  41.91     TRUE                1
## 5:       1959.22     148.40       1869.81  21.82     TRUE                2
## 6:       3277.50    2017.48       3208.50 384.89     TRUE                5
##    AberrantByGene padj_rank
##             <num>     <num>
## 1:              1       1.0
## 2:              1       1.0
## 3:              1       1.0
## 4:              1       1.0
## 5:              1       1.0
## 6:              1       1.5
```

```
dim(res)
```

```
## [1] 245  15
```

```
# setting a different significance level and filtering by Z-scores
res <- results(ods, padjCutoff=0.1, zScoreCutoff=2)
head(res)
```

```
##     geneID sampleID       pValue      padjust zScore  l2fc rawcounts
##     <char>   <char>        <num>        <num>  <num> <num>     <int>
## 1:  NUDT12    65937 1.714844e-22 2.021918e-17 -10.34 -7.97         0
## 2:   STAG2  MUC0490 3.750099e-21 4.421623e-16  -9.73 -1.87       622
## 3:  TALDO1  MUC1427 7.290799e-18 8.596350e-13  -9.43 -3.28       482
## 4:  MCOLN1  MUC1361 1.079028e-15 1.272248e-10  -8.80 -2.41       150
## 5:  RETSAT  MUC1374 1.175468e-15 1.385957e-10  -8.94 -3.64        76
## 6: CSNK2A1  MUC1358 2.794734e-15 1.647591e-10  -8.12 -0.67      2060
##    meanRawcounts normcounts meanCorrected  theta aberrant AberrantBySample
##            <num>      <num>         <num>  <num>   <lgcl>            <num>
## 1:        470.22       0.00        453.43  19.28     TRUE                2
## 2:       4184.49    1108.88       4066.04  83.41     TRUE                6
## 3:       4699.52     469.42       4556.79  27.85     TRUE                1
## 4:        912.70     157.77        841.06  41.91     TRUE                1
```

```
## 5:       1959.22     148.40       1869.81  21.82     TRUE              2
## 6:       3277.50    2017.48       3208.50 384.89     TRUE              5
##    AberrantByGene padj_rank
##             <num>     <num>
## 1:              1       1.0
## 2:              1       1.0
## 3:              1       1.0
## 4:              1       1.0
## 5:              1       1.0
## 6:              1       1.5
```

```
dim(res)
```

```
## [1] 313  15
```

In details the table contains:

- sampleID / geneID: The gene or sample ID as provided by the user, e.g. rowData(ods) and colData(ods) respectively.
- pValue / padjust: The nominal P-value and the FDR corrected P-value indicating the outlier status. Find more details at computePvalues.
- zScore / l2fc: The z score and $\log_2$ fold change as computed by computeZscores.
- rawcounts: The observed read counts.
- normcounts: The expected count given the fitted autoencoder model for the given gene-sample combination.
- meanRawcounts / meanCorrected: For this gene, the mean of the observed or expected counts, respectively, given the fitted autoencoder model.
- theta: The dispersion parameter of the NB distribution for the given gene.
- aberrant: The outlier status of this event: TRUE or FALSE.
- AberrantBySample / AberrantByGene: Number of outliers for the given sample or gene, respectively.
- padj rank: Rank of this outlier event within the given sample.
- FDR set: The subset-name used for the P-value computation.

5.2 Number of aberrant genes per sample

One quantity of interest is the number of aberrantly expressed genes per sample. This can be displayed using the plotting function `plotAberrantPerSample`. Alternatively, the function `aberrant` can be used to identify aberrant events,

which can be summed by sample or gene using the paramter `by`. These numbers depend on the cutoffs, which can be specified in both functions (`padjCutoff` and `zScoreCutoff`).

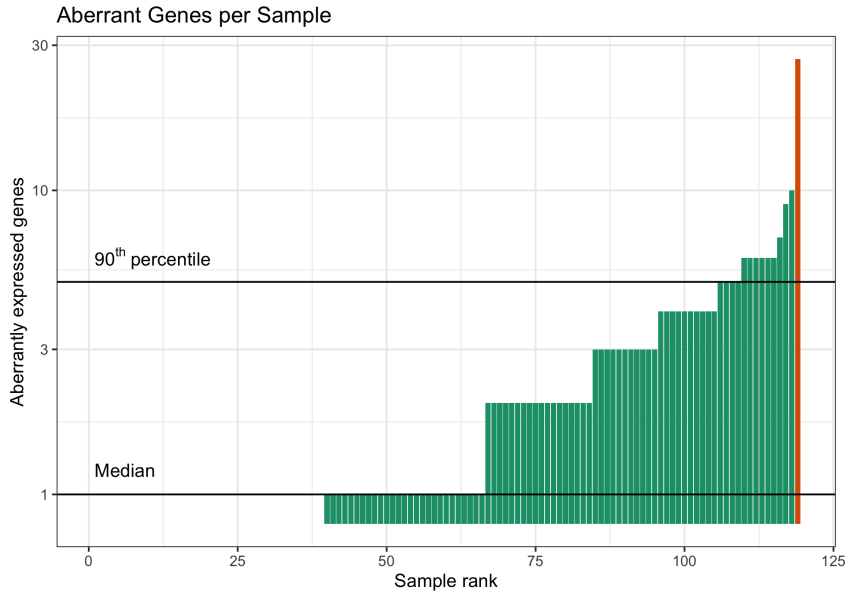
```
# number of aberrant genes per sample
tail(sort(aberrant(ods, by="sample")))
```

```
## MUC1367 MUC1381 MUC1368 MUC1363 MUC1364   76633
##       6       6       7       9      10      27
```

```
tail(sort(aberrant(ods, by="gene", zScoreCutoff=1)))
```

```
##        ZFAT      DNAJC3 SLMO2-ATP5E        NXT2      RPS4Y1        PRKY
##           2           2           2           2           2           2
```

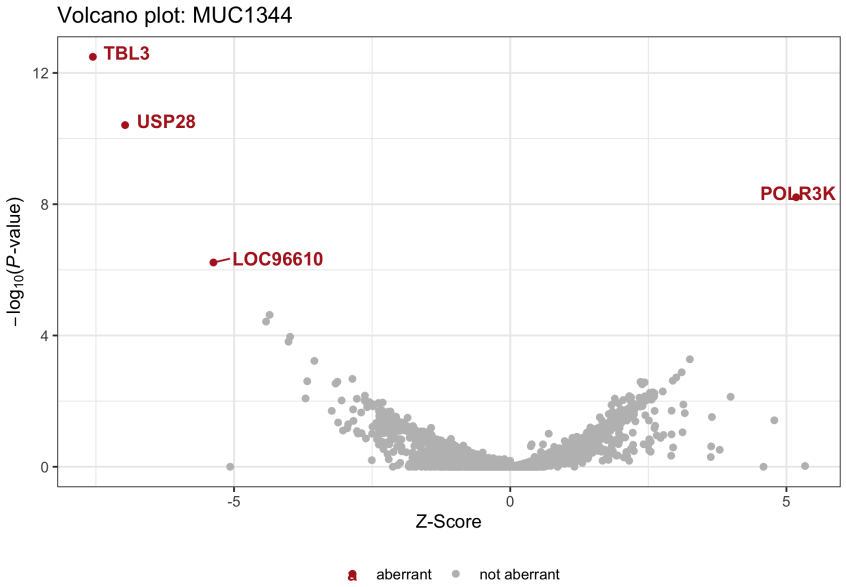
```
# plot the aberrant events per sample
plotAberrantPerSample(ods, padjCutoff=0.05)
```



5.3 Volcano plots

To view the distribution of P-values on a sample level, volcano plots can be displayed. Most of the plots make use of the *plotly* framework to create interactive plots. To only use basic R functionality from *graphics* the `basePlot` argument can be set to `TRUE`.

```
# MUC1344 is a diagnosed sample from Kremer et al.
plotVolcano(ods, "MUC1344", basePlot=TRUE)
```

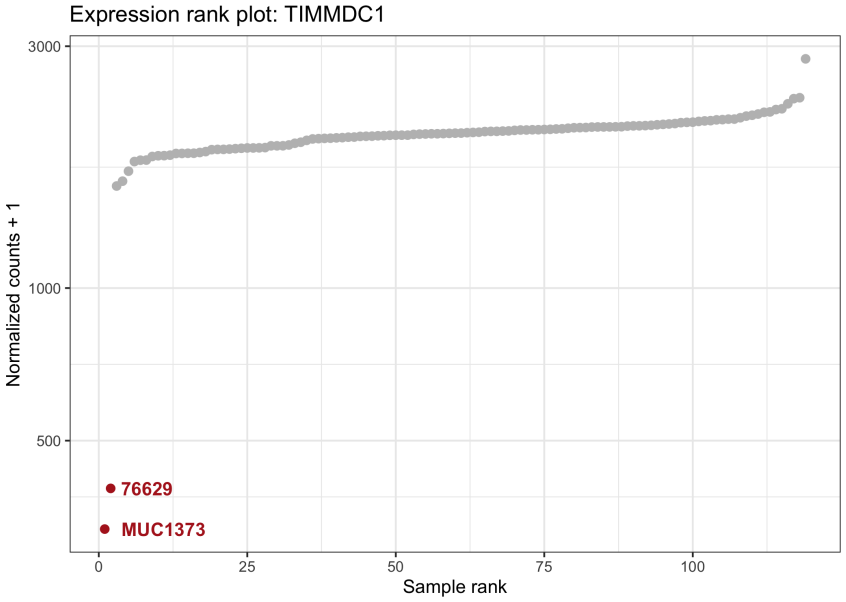


5.4 Gene level plots

Additionally, we include two plots at the gene level. `plotExpressionRank` plots the counts in ascending order. By default, the controlled counts are plotted. To plot raw counts, the argument `normalized` can be set to `FALSE`.

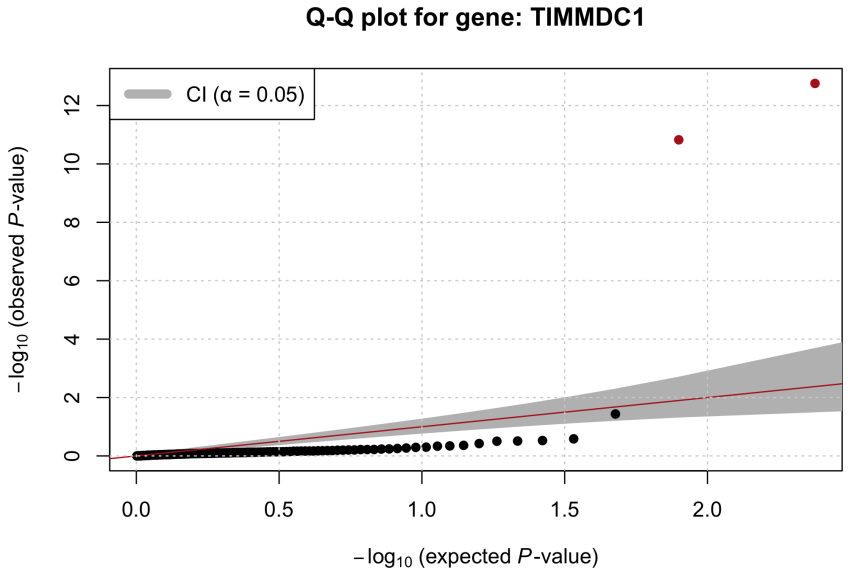
When using the *plotly* framework for plotting, all computed values are displayed for each data point. The user can access this information by hovering over each data point with the mouse.

```
# expression rank of a gene with outlier events
plotExpressionRank(ods, "TIMMDC1", basePlot=TRUE)
```



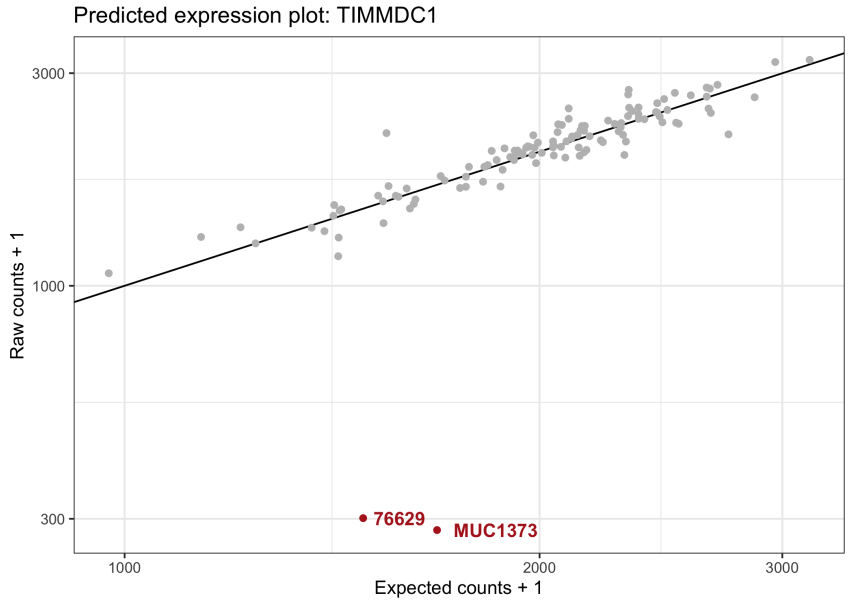
The quantile-quantile plot can be used to see whether the fit converged well. In presence of an outlier, it can happen that most of the points end up below the confidence band. This is fine and indicates that we have conservative P-values for the other points. Here is an example with two outliers:

```
## QQ-plot for a given gene
plotQQ(ods, "TIMMDC1")
```



Since we do test how fare the observed count is away from the exprected expression level, it is also helpful to visualize the predictions against the observed counts.

```
## Observed versus expected gene expression
plotExpectedVsObservedCounts(ods, "TIMMDC1", basePlot=TRUE)
```



6 Additional features

6.1 Using PEER to control for confounders

PEER[6] is a well known tool to control for unknown effects in RNA-seq data. PEER is only available through the *peer* GitHub repository. The R source code can be downloaded form here: https://github.com/downloads/PMBio/peer/R_peer_source_1.3.tgz. The installation of the package has to be done manually by the user. After the installation one can use the following function to control for confounders with PEER.

```
#'
#' PEER implementation
#'
peer <- function(ods, maxFactors=NA, maxItr=1000){

    # check for PEER
    if(!require(peer)){
        stop("Please install the 'peer' package from GitHub to use this ",
                "functionality.")
    }

    # default and recommendation by PEER: min(0.25*n, 100)
    if(is.na(maxFactors)){
        maxFactors <- min(as.integer(0.25* ncol(ods)), 100)
    }
```

```
    # log counts
    logCts <- log2(t(t(counts(ods)+1)/sizeFactors(ods)))

    # prepare PEER model
    model <- PEER()
    PEER_setNmax_iterations(model, maxItr)
    PEER_setNk(model, maxFactors)
    PEER_setPhenoMean(model, logCts)
    PEER_setAdd_mean(model, TRUE)

    # run fullpeer pipeline
    PEER_update(model)

    # extract PEER data
    peerResiduals <- PEER_getResiduals(model)
    peerMean <- t(t(2^(logCts - peerResiduals)) * sizeFactors(ods))

    # save model in object
    normalizationFactors(ods) <- pmax(peerMean, 1E-8)
    metadata(ods)[["PEER_model"]] <- list(
            alpha     = PEER_getAlpha(model),
            residuals = PEER_getResiduals(model),
            W         = PEER_getW(model))

    return(ods)
}
```

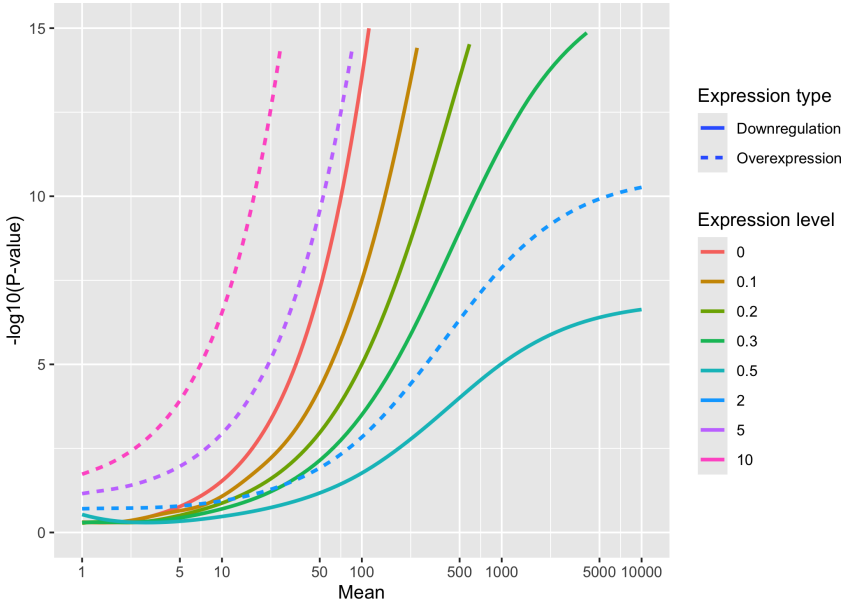
With the function above we can run the full OUTRIDER pipeline as follows:

```
# Control for confounders with PEER
ods <- estimateSizeFactors(ods)
ods <- peer(ods)
ods <- fit(ods)
ods <- computeZscores(ods, peerResiduals=TRUE)
ods <- computePvalues(ods)

# Heatmap of the sample correlation after controlling
ods <- plotCountCorHeatmap(ods, normalized=TRUE)
```

6.2 Power anaylsis

We provide the `plotPowerAnalysis` function to show, what kind of changes can be significant depending on the mean count.

```
## P-values versus Mean Count
plotPowerAnalysis(ods)
```



Here, we see that it is only for sufficiently high expressed genes possible, to obtain significant P-values, especially for the downregulation cases.

References

[1] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, dec 2014. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8, doi:10.1186/s13059-014-0550-8.

[2] Mark D Robinson, Davis J. McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40, jan 2010. URL: https://doi.org/10.1093/bioinformatics/btp616, doi:10.1093/bioinformatics/btp616.

[3] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. URL: https://doi.org/10.1093/bioinformatics/bts635, doi:10.1093/bioinformatics/bts635.

[4] Laura S Kremer, Daniel M Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, Eliška Koňaříková, Birgit Repp, Gabi Kas-

tenmüller, Jerzy Adamski, Peter Lichtner, Christoph Leonhardt, Benoit Funalot, Alice Donati, Valeria Tiranti, Anne Lombes, Claude Jardel, Dieter Gläser, Robert W Taylor, Daniele Ghezzi, Johannes A Mayr, Agnes Rötig, Peter Freisinger, Felix Distelmaier, Tim M Strom, Thomas Meitinger, Julien Gagneur, and Holger Prokisch. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications*, 8:15824, jun 2017. URL: https://www.nature.com/articles/ncomms15824.pdf, doi:10.1038/ncomms15824.

[5] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. URL: https://projecteuclid.org/euclid.aos/1013699998, arXiv:0801.1095, doi:10.1214/aos/1013699998.

[6] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012. doi:10.1038/nprot.2011.457.

Session info

Here is the output of `sessionInfo()` on the system on which this document was compiled:

```
## R version 4.4.0 alpha (2024-03-27 r86216)
## Platform: aarch64-apple-darwin20
## Running under: macOS Ventura 13.6.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] org.Hs.eg.db_3.19.0
##  [2] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
##  [3] beeswarm_0.4.0
```

```
##  [4] OUTRIDER_1.22.0
##  [5] data.table_1.15.4
##  [6] SummarizedExperiment_1.34.0
##  [7] MatrixGenerics_1.16.0
##  [8] matrixStats_1.2.0
##  [9] GenomicFeatures_1.56.0
## [10] AnnotationDbi_1.66.0
## [11] Biobase_2.64.0
## [12] GenomicRanges_1.56.0
## [13] GenomeInfoDb_1.40.0
## [14] IRanges_2.38.0
## [15] S4Vectors_0.42.0
## [16] BiocGenerics_0.50.0
## [17] BiocParallel_1.38.0
## [18] knitr_1.45
##
## loaded via a namespace (and not attached):
##   [1] RColorBrewer_1.1-3       jsonlite_1.8.8
##   [3] magrittr_2.0.3           magick_2.8.3
##   [5] farver_2.1.1             rmarkdown_2.26
##   [7] BiocIO_1.14.0            zlibbioc_1.50.0
##   [9] vctrs_0.6.5              memoise_2.0.1
##  [11] Rsamtools_2.20.0         RCurl_1.98-1.14
##  [13] PRROC_1.3.1              webshot_0.5.5
##  [15] htmltools_0.5.8          S4Arrays_1.4.0
##  [17] progress_1.2.3           curl_5.2.1
##  [19] SparseArray_1.4.0        htmlwidgets_1.6.4
##  [21] plyr_1.8.9               httr2_1.0.0
##  [23] plotly_4.10.4            cachem_1.0.8
##  [25] GenomicAlignments_1.40.0 lifecycle_1.0.4
##  [27] iterators_1.0.14         pkgconfig_2.0.3
##  [29] Matrix_1.7-0             R6_2.5.1
##  [31] fastmap_1.1.1            GenomeInfoDbData_1.2.12
##  [33] digest_0.6.35            pcaMethods_1.96.0
##  [35] colorspace_2.1-0         DESeq2_1.44.0
##  [37] RSQLite_2.3.5            seriation_1.5.4
##  [39] labeling_0.4.3           filelock_1.0.3
##  [41] fansi_1.0.6              mgcv_1.9-1
##  [43] httr_1.4.7               abind_1.4-5
##  [45] compiler_4.4.0           withr_3.0.0
##  [47] bit64_4.0.5              backports_1.4.1
##  [49] viridis_0.6.5            DBI_1.2.2
##  [51] heatmaply_1.5.0          dendextend_1.17.1
##  [53] highr_0.10               biomaRt_2.60.0
##  [55] rappdirs_0.3.3           DelayedArray_0.30.0
```

```
##  [57] rjson_0.2.21             tools_4.4.0
##  [59] glue_1.7.0               restfulr_0.0.15
##  [61] nlme_3.1-164             grid_4.4.0
##  [63] checkmate_2.3.1          reshape2_1.4.4
##  [65] generics_0.1.3           gtable_0.3.4
##  [67] ca_0.71.1                tidyr_1.3.1
##  [69] hms_1.1.3                xml2_1.3.6
##  [71] utf8_1.2.4               XVector_0.44.0
##  [73] ggrepel_0.9.5            foreach_1.5.2
##  [75] pillar_1.9.0             stringr_1.5.1
##  [77] splines_4.4.0            dplyr_1.1.4
##  [79] BBmisc_1.13              BiocFileCache_2.12.0
##  [81] lattice_0.22-6           rtracklayer_1.64.0
##  [83] bit_4.0.5                tidyselect_1.2.1
##  [85] registry_0.5-1           locfit_1.5-9.9
##  [87] Biostrings_2.72.0        gridExtra_2.3
##  [89] xfun_0.43                pheatmap_1.0.12
##  [91] stringi_1.8.3            UCSC.utils_1.0.0
##  [93] lazyeval_0.2.2           yaml_2.3.8
##  [95] evaluate_0.23            codetools_0.2-19
##  [97] tibble_3.2.1             BiocManager_1.30.22
##  [99] cli_3.6.2                munsell_0.5.0
## [101] Rcpp_1.0.12              dbplyr_2.5.0
## [103] png_0.1-8                XML_3.99-0.16.1
## [105] parallel_4.4.0           ggplot2_3.5.0
## [107] assertthat_0.2.1         blob_1.2.4
## [109] prettyunits_1.2.0        bitops_1.0-7
## [111] txdbmaker_1.0.0          viridisLite_0.4.2
## [113] scales_1.3.0             purrr_1.0.2
## [115] crayon_1.5.2             BiocStyle_2.32.0
## [117] rlang_1.1.3              KEGGREST_1.44.0
## [119] TSP_1.2-4
```