Differential analysis of count data – the DESeq2 package

Michael Love[1*], Simon Anders[2], Wolfgang Huber[2]

[1] Department of Biostatistics, Dana Farber Cancer Institute and
Harvard School of Public Health, Boston, US;
[2] European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[*]michaelisaiahlove (at) gmail.com

May 1, 2015

Abstract

A basic task in the analysis of count data from RNA-Seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package *DESeq2* provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions[1]. This vignette explains the use of the package and demonstrates typical workflows. An RNA-Seq workflow on the Bioconductor website: http://www.bioconductor.org/help/workflows/rnaseqGene/ (formerly the Beginner's Guide PDF), covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files.

DESeq2 version: 1.8.1

If you use *DESeq2* in published research, please cite:

M. I. Love, W. Huber, S. Anders: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biology* 2014, **15**:550.
http://dx.doi.org/10.1186/s13059-014-0550-8

[1]Other *Bioconductor* packages with similar aims are *edgeR*, *baySeq*, *DSS* and *limma*.

Contents

1 Standard workflow 4
1.1 Quick start . . . 4
1.2 Input data . . . 4
1.2.1 Why raw counts? . . . 4
1.2.2 *SummarizedExperiment* input . . . 4
1.2.3 Count matrix input . . . 5
1.2.4 *HTSeq* input . . . 7
1.2.5 Note on factor levels . . . 8
1.2.6 Collapsing technical replicates . . . 8
1.2.7 About the pasilla dataset . . . 8
1.3 Differential expression analysis . . . 8
1.4 Exploring and exporting results . . . 10
1.4.1 MA-plot . . . 10
1.4.2 Plot counts . . . 11
1.4.3 More information on results columns . . . 12
1.4.4 Exporting results to HTML or CSV files . . . 13
1.5 Multi-factor designs . . . 14

2 Data transformations and visualization 16
2.1 Count data transformations . . . 16
2.1.1 Blind dispersion estimation . . . 16
2.1.2 Extracting transformed values . . . 17
2.1.3 Regularized log transformation . . . 17
2.1.4 Variance stabilizing transformation . . . 17
2.1.5 Effects of transformations on the variance . . . 19
2.2 Data quality assessment by sample clustering and visualization . . . 19
2.2.1 Heatmap of the count matrix . . . 19
2.2.2 Heatmap of the sample-to-sample distances . . . 21
2.2.3 Principal component plot of the samples . . . 21

3 Variations to the standard workflow 23
3.1 Wald test individual steps . . . 23
3.2 Contrasts . . . 23
3.3 Interactions . . . 23
3.4 Time-series experiments . . . 25
3.5 Likelihood ratio test . . . 25
3.6 Approach to count outliers . . . 26
3.7 Dispersion plot and fitting alternatives . . . 27
3.7.1 Local or mean dispersion fit . . . 27
3.7.2 Supply a custom dispersion fit . . . 27
3.8 Independent filtering of results . . . 28
3.9 Tests of log2 fold change above or below a threshold . . . 29
3.10 Access to all calculated values . . . 31
3.11 Sample-/gene-dependent normalization factors . . . 33
3.12 Model matrix not full rank . . . 34
3.12.1 Linear combinations . . . 34

3.12.2 Levels without samples . . . 36

4 Theory behind DESeq2 39
4.1 The DESeq2 model . . . 39
4.2 Changes compared to the *DESeq* package . . . 39
4.3 Methods changes since the 2014 DESeq2 paper . . . 40
4.4 Count outlier detection . . . 40
4.5 Contrasts . . . 40
4.6 Expanded model matrices . . . 41
4.7 Independent filtering and multiple testing . . . 41
4.7.1 Filtering criteria . . . 42
4.7.2 Why does it work? . . . 42
4.7.3 Diagnostic plots for multiple testing . . . 43

5 Frequently asked questions 45
5.1 How can I get support for DESeq2? . . . 45
5.2 Why are some p values set to `NA`? . . . 46
5.3 How can I get unfiltered DESeq results? . . . 46
5.4 How do I use the variance stabilized or rlog transformed data for differential testing? . . . 46
5.5 Can I use DESeq2 to analyze paired samples? . . . 46
5.6 Can I run DESeq2 to contrast the levels of 100 groups? . . . 46
5.7 Can I use DESeq2 to analyze a dataset without replicates? . . . 47
5.8 How can I include a continuous covariate in the design formula? . . . 47
5.9 What are the exact steps performed by `DESeq()`? . . . 47

6 Acknowledgments 47

7 Session Info 47

1 Standard workflow

1.1 Quick start

Here we show the most basic steps for a differential expression analysis. These steps imply you have a *SummarizedExperiment* object `se` with columns `batch` and `condition` in `colData(se)`, and that `condition` has levels "untreat" and "treat".

```
dds <- DESeqDataSet(se, ~ batch + condition)
dds <- DESeq(dds)
res <- results(dds, contrast=c("condition","treat","untreat"))
```

If you have a count matrix and sample information table, the first line would instead use `DESeqDataSetFromMatrix`, shown below.

1.2 Input data

1.2.1 Why raw counts?

As input, the *DESeq2* package expects count data as obtained, e. g., from RNA-Seq or another high-throughput sequencing experiment, in the form of a matrix of integer values. The value in the i-th row and the j-th column of the matrix tells how many reads have been mapped to gene i in sample j. Analogously, for other types of assays, the rows of the matrix might correspond e. g. to binding regions (with ChIP-Seq) or peptide sequences (with quantitative mass spectrometry).

The count values must be raw counts of sequencing reads. This is important for *DESeq2*'s statistical model [1] to hold, as only the actual counts allow assessing the measurement precision correctly. Hence, please do not supply other quantities, such as (rounded) normalized counts, or counts of covered base pairs – this will only lead to nonsensical results.

1.2.2 SummarizedExperiment input

The class used by the *DESeq2* package to store the read counts is *DESeqDataSet* which extends the *SummarizedExperiment* class of the *GenomicRanges* package. This facilitates preparation steps and also downstream exploration of results. For counting aligned reads in genes, the `summarizeOverlaps` function of *GenomicAlignments* with `mode="Union"` is encouraged, resulting in a *SummarizedExperiment* object.

An example of the steps to produce a *SummarizedExperiment* can be found in an RNA-Seq workflow on the Bioconductor website: http://www.bioconductor.org/help/workflows/rnaseqGene/ and in the vignette for the data package *airway*. Here we load the *SummarizedExperiment* from that package in order to build a *DESeqDataSet*.

```
library("airway")
data("airway")
se <- airway
```

A *DESeqDataSet* object must have an associated design formula. The design formula expresses the variables which will be used in modeling. The formula should be a tilde ($\sim$) followed by the variables with plus signs

between them (it will be coerced into an *formula* if it is not already). An intercept is included, representing the base mean of counts. The design can be changed later, however then all differential analysis steps should be repeated, as the design formula is used to estimate the dispersions and to estimate the log2 fold changes of the model. The constructor function below shows the generation of a *DESeqDataSet* from a *SummarizedExperiment* `se`.

Note: In order to benefit from the default settings of the package, you should put the variable of interest at the end of the formula and make sure the control level is the first level.

```
library("DESeq2")
ddsSE <- DESeqDataSet(se, design = ~ cell + dex)
ddsSE
```

```
## class: DESeqDataSet
## dim: 64102 8
## exptData(1): ''
## assays(1): counts
## rownames(64102): ENSG00000000003 ENSG00000000005 ... LRG_98 LRG_99
## rowRanges metadata column names(0):
## colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
## colData names(9): SampleName cell ... Sample BioSample
```

1.2.3 Count matrix input

Alternatively, the function `DESeqDataSetFromMatrix` can be used if you already have a matrix of read counts prepared. For this function you should provide the counts matrix, the column information as a *DataFrame* or *data.frame* and the design formula. To demonstate the use of this function, we first load the `pasillaGenes` data object, to pull out the count matrix, `countData`, and sample information, `colData`.

```
library("pasilla")
library("Biobase")
data("pasillaGenes")
countData <- counts(pasillaGenes)
colData <- pData(pasillaGenes)[,c("condition","type")]
```

`countData` should be a matrix of read counts, where columns correspond to different samples. `colData` should be a *data.frame* or *DataFrame* of information about the samples, where the rows directly match the columns of `countData`.

```
head(countData)
```

```
##             treated1fb treated2fb treated3fb untreated1fb untreated2fb
## FBgn0000003          0          0          1            0            0
## FBgn0000008         78         46         43           47           89
## FBgn0000014          2          0          0            0            0
## FBgn0000015          1          0          1            0            1
## FBgn0000017       3187       1672       1859         2445         4615
## FBgn0000018        369        150        176          288          383
##             untreated3fb untreated4fb
## FBgn0000003            0            0
```

```
## FBgn0000008           53           27
## FBgn0000014            1            0
## FBgn0000015            1            2
## FBgn0000017         2063         1711
## FBgn0000018          135          174
```

```
head(colData)
```

```
##              condition        type
## treated1fb     treated single-read
## treated2fb     treated  paired-end
## treated3fb     treated  paired-end
## untreated1fb untreated single-read
## untreated2fb untreated single-read
## untreated3fb untreated  paired-end
```

These two objects can be read into R from flat files using base R functions such as `read.csv` or `read.delim`. For *HTSeq* count files, see the dedicated input function below. If you have used the `featureCounts` function in the *Rsubread* package, the matrix of read counts can be directly provided from the `"counts"` element in the list output. With the count matrix, `countData`, and the sample information, `colData`, we can construct a *DESeqDataSet*:

```
dds <- DESeqDataSetFromMatrix(countData = countData,
                              colData = colData,
                              design = ~ condition)
dds
```

```
## class: DESeqDataSet
## dim: 14470 7
## exptData(0):
## assays(1): counts
## rownames(14470): FBgn0000003 FBgn0000008 ... FBgn0261574 FBgn0261575
## rowRanges metadata column names(0):
## colnames(7): treated1fb treated2fb ... untreated3fb untreated4fb
## colData names(2): condition type
```

If you have additional feature data, it can be added to the *DESeqDataSet* by adding to the metadata columns of a newly constructed object. (Here we add redundant data just for demonstration, as the gene names are already the rownames of the `dds`.)

```
featureData <- data.frame(gene=rownames(pasillaGenes))
(mcols(dds) <- DataFrame(mcols(dds), featureData))
```

```
## DataFrame with 14470 rows and 1 column
##              gene
##          <factor>
## 1     FBgn0000003
## 2     FBgn0000008
## 3     FBgn0000014
## 4     FBgn0000015
## 5     FBgn0000017
```

```
## ...           ...
## 14466 FBgn0261571
## 14467 FBgn0261572
## 14468 FBgn0261573
## 14469 FBgn0261574
## 14470 FBgn0261575
```

1.2.4 HTSeq input

You can use the function `DESeqDataSetFromHTSeqCount` if you have `htseq-count` from the *HTSeq* python package[2]. For an example of using the python scripts, see the *pasilla* data package. First you will want to specify a variable which points to the directory in which the *HTSeq* output files are located.

```
directory <- "/path/to/your/files/"
```

However, for demonstration purposes only, the following line of code points to the directory for the demo *HTSeq* output files packages for the *pasilla* package.

```
directory <- system.file("extdata", package="pasilla", mustWork=TRUE)
```

We specify which files to read in using `list.files`, and select those files which contain the string `"treated"` using `grep`. The `sub` function is used to chop up the sample filename to obtain the condition status, or you might alternatively read in a phenotypic table using `read.table`.

```
sampleFiles <- grep("treated",list.files(directory),value=TRUE)
sampleCondition <- sub("(.*treated).*","\\1",sampleFiles)
sampleTable <- data.frame(sampleName = sampleFiles,
                          fileName = sampleFiles,
                          condition = sampleCondition)
ddsHTSeq <- DESeqDataSetFromHTSeqCount(sampleTable = sampleTable,
                                       directory = directory,
                                       design= ~ condition)
ddsHTSeq
```

```
## class: DESeqDataSet
## dim: 70463 7
## exptData(0):
## assays(1): counts
## rownames(70463): FBgn0000003:001 FBgn0000008:001 ... FBgn0261575:001
##   FBgn0261575:002
## rowRanges metadata column names(0):
## colnames(7): treated1fb.txt treated2fb.txt ... untreated3fb.txt
##   untreated4fb.txt
## colData names(1): condition
```

[2] available from http://www-huber.embl.de/users/anders/HTSeq, described in [2]

1.2.5 Note on factor levels

In the three examples above, we applied the function `factor` to the column of interest in `colData`, supplying a character vector of levels. It is important to supply levels (otherwise the levels are chosen in alphabetical order) and to put the "control" or "untreated" level as the first element ("reference level"), so that the log2 fold changes produced by default will be the expected comparison against the reference level, that is $\log_2(\text{treated}/\text{untreated})$. An *R* function for easily changing the reference level is `relevel`. An example of setting the reference level of a factor with `relevel` is:

```
dds$condition <- relevel(dds$condition, "untreated")
```

In addition, when subsetting the columns of a *DESeqDataSet*, i.e., when removing certain samples from the analysis, it is possible that all the samples for one or more levels of a variable in the design formula are removed. In this case, the `droplevels` function can be used to remove those levels which do not have samples in the current *DESeqDataSet*:

```
dds$condition <- droplevels(dds$condition)
```

1.2.6 Collapsing technical replicates

DESeq2 provides a function `collapseReplicates` which can assist in combining the counts from technical replicates into single columns. See the manual page for an example of the use of `collapseReplicates`.

1.2.7 About the pasilla dataset

We continue with the *pasilla* data constructed from the count matrix method above. This data set is from an experiment on *Drosophila melanogaster* cell cultures and investigated the effect of RNAi knock-down of the splicing factor *pasilla* [3]. The detailed transcript of the production of the *pasilla* data is provided in the vignette of the data package *pasilla*.

1.3 Differential expression analysis

The standard differential expression analysis steps are wrapped into a single function, `DESeq`. The steps of this function are described in Section 4.1 and in the manual page for `?DESeq`. The individual sub-functions which are called by `DESeq` are still available, described in Section 3.1.

Results tables are generated using the function `results`, which extracts a results table with log2 fold changes, p values and adjusted p values. With no arguments to `results`, the results will be for the last variable in the design formula, and if this is a factor, the comparison will be the last level of this variable over the first level. Details about the comparison are printed to the console. The text, `condition treated vs untreated`, tells you that the estimates are of the logarithmic fold change $\log_2(\text{treated}/\text{untreated})$.

```
dds <- DESeq(dds)
res <- results(dds)
res
```

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
```

```
## DataFrame with 14470 rows and 6 columns
##              baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##             <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0000003     0.159         0.0346    0.0461    0.7487     0.454        NA
## FBgn0000008    52.226         0.0197    0.2093    0.0942     0.925     0.988
## FBgn0000014     0.390         0.0118    0.0622    0.1901     0.849        NA
## FBgn0000015     0.905        -0.0429    0.1054   -0.4076     0.684        NA
## FBgn0000017  2358.243        -0.2554    0.1198   -2.1317     0.033     0.230
## ...               ...            ...       ...       ...       ...       ...
## FBgn0261571  9.45e-02        0.02949    0.0441    0.6690     0.503        NA
## FBgn0261572  3.24e+00       -0.20258    0.1710   -1.1846     0.236        NA
## FBgn0261573  1.40e+03       -0.00229    0.1329   -0.0172     0.986     0.997
## FBgn0261574  2.72e+03        0.02538    0.1669    0.1521     0.879     0.981
## FBgn0261575  3.36e+00        0.08990    0.1632    0.5510     0.582        NA
```

These steps should take less than 30 seconds for most analyses. For experiments with many samples (e.g. 100 samples), one can take advantage of parallelized computation. Both of the above functions have an argument `parallel` which if set to `TRUE` can be used to distribute computation across cores specified by the `register` function of *BiocParallel*. For example, the following chunk (not evaluated here), would register 4 cores, and then the two functions above, with `parallel=TRUE`, would split computation over these cores.

```
library("BiocParallel")
register(MulticoreParam(4))
```

We can order our results table by the smallest adjusted p value:

```
resOrdered <- res[order(res$padj),]
head(resOrdered)
```

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 6 rows and 6 columns
##              baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##             <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0039155       453          -3.71     0.160     -23.2 4.01e-119 3.11e-115
## FBgn0029167      2165          -2.08     0.104     -20.1  6.68e-90  2.59e-86
## FBgn0035085       367          -2.23     0.137     -16.3  1.89e-59  4.87e-56
## FBgn0029896       258          -2.21     0.159     -13.9  5.85e-44  1.13e-40
## FBgn0034736       118          -2.57     0.185     -13.9  8.07e-44  1.25e-40
## FBgn0040091       611          -1.43     0.120     -11.9  1.11e-32  1.44e-29
```

We can summarize some basic tallies using the `summary` function.

```
summary(res)
```

```
##
## out of 11836 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)     : 394, 3.3%
## LFC < 0 (down)   : 409, 3.5%
```

```
## outliers [1]     : 15, 0.13%
## low counts [2]   : 4084, 35%
## (mean count < 9.2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

The `results` function contains a number of arguments to customize the results table which is generated. Note that the `results` function automatically performs independent filtering based on the mean of normalized counts for each gene, optimizing the number of genes which will have an adjusted p value below a given FDR cutoff, `alpha`, By default the argument `alpha` is set to 0.1. If the adjusted p value cutoff will be a value other than 0.1, `alpha` should be set to that value. Independent filtering will be discussed further in Section 3.8.

If a multi-factor design is used, or if the variable in the design formula has more than two levels, the `contrast` argument of `results` can be used to extract different comparisons from the *DESeqDataSet* returned by `DESeq`. Multi-factor designs are discussed further in Section 1.5, and the use of the `contrast` argument is dicussed in Section 3.2.

For advanced users, note that all the values calculated by the *DESeq2* package are stored in the *DESeqDataSet* object, and access to these values is discussed in Section 3.10.

1.4 Exploring and exporting results

1.4.1 MA-plot

In *DESeq2*, the function `plotMA` shows the log2 fold changes attributable to a given variable over the mean of normalized counts. Points will be colored red if the adjusted p value is less than 0.1. Points which fall out of the window are plotted as open triangles pointing either up or down.

```
plotMA(res, main="DESeq2", ylim=c(-2,2))
```

The MA-plot of log2 fold changes returned by *DESeq2* allows us to see how the shrinkage of fold changes works for genes with low counts. You can still obtain results tables which include the "unshrunken" log2 fold changes (for a simple comparison, the ratio of the mean normalized counts in the two groups). A column `lfcMLE` with the unshrunken maximum likelihood estimate (MLE) for the log2 fold change will be added with an additional argument to `results`:

```
resMLE <- results(dds, addMLE=TRUE)
head(resMLE, 4)
```

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 4 rows and 7 columns
##               baseMean log2FoldChange    lfcMLE     lfcSE      stat    pvalue
##              <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0000003     0.159         0.0346    3.2929    0.0461    0.7487     0.454
## FBgn0000008    52.226         0.0197    0.0281    0.2093    0.0942     0.925
## FBgn0000014     0.390         0.0118    0.6239    0.0622    0.1901     0.849
## FBgn0000015     0.905        -0.0429   -0.8135    0.1054   -0.4076     0.684
##                  padj
```

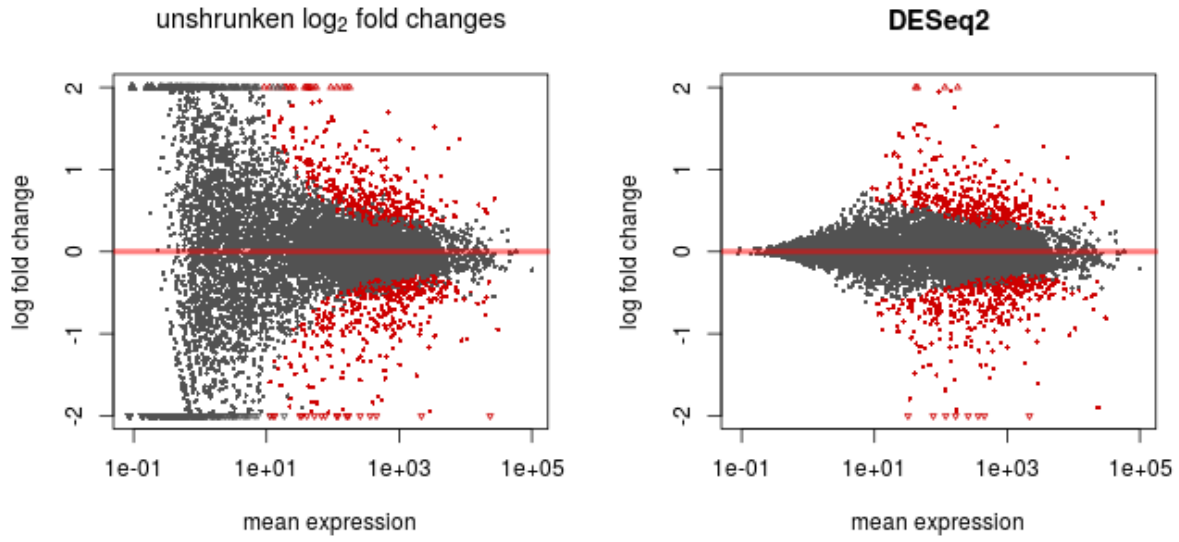

Figure 1: **MA-plot.** These plots show the log2 fold changes from the treatment over the mean of normalized counts, i.e. the average of counts normalized by size factors. The left plot shows the "unshrunken" log2 fold changes, while the right plot, produced by the code above, shows the shrinkage of log2 fold changes resulting from the incorporation of zero-centered normal prior. The shrinkage is greater for the log2 fold change estimates from genes with low counts and high dispersion, as can be seen by the narrowing of spread of leftmost points in the right plot.

```
##             <numeric>
## FBgn0000003        NA
## FBgn0000008     0.988
## FBgn0000014        NA
## FBgn0000015        NA
```

After calling `plotMA`, one can use the function `identify` to interactively detect the row number of individual genes by clicking on the plot.

```
identify(res$baseMean, res$log2FoldChange)
```

1.4.2 Plot counts

It can also be useful to examine the counts of reads for a single gene across the groups. A simple function for making this plot is `plotCounts`, which normalizes counts by sequencing depth and adds a pseudocount of $\frac{1}{2}$ to allow for log scale plotting. The counts are grouped by the variables in `intgroup`, where more than one variable can be specified. Here we specify the gene which had the smallest p value from the results table created above. You can select the gene to plot by rowname or by numeric index.

```
plotCounts(dds, gene=which.min(res$padj), intgroup="condition")
```

For customized plotting, an argument `returnData` specifies that the function should only return a *data.frame* for plotting with `ggplot`.

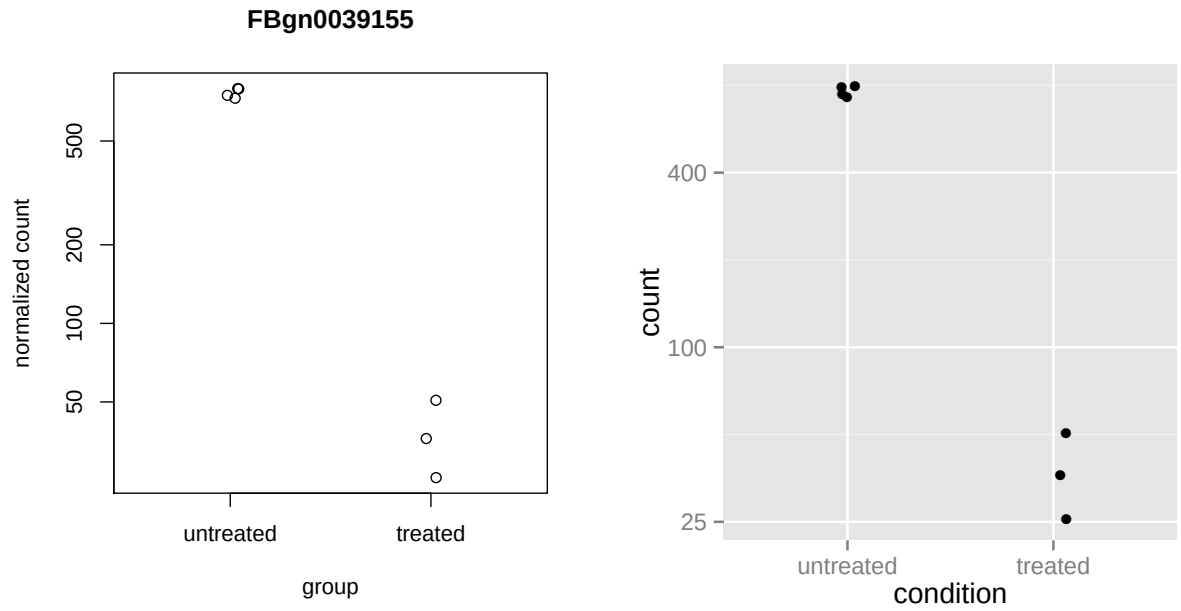

Figure 2: **Plot of counts for one gene.** The plot of normalized counts (plus a pseudocount of $\frac{1}{2}$) either made using the `plotCounts` function (left) or using another plotting library (right, using *ggplot2*).

```
d <- plotCounts(dds, gene=which.min(res$padj), intgroup="condition",
                returnData=TRUE)
library("ggplot2")
ggplot(d, aes(x=condition, y=count)) +
  geom_point(position=position_jitter(w=0.1,h=0)) +
  scale_y_log10(breaks=c(25,100,400))
```

1.4.3 More information on results columns

Information about which variables and tests were used can be found by calling the function `mcols` on the results object.

```
mcols(res)$description
```

```
## [1] "mean of normalized counts for all samples"
## [2] "log2 fold change (MAP): condition treated vs untreated"
## [3] "standard error: condition treated vs untreated"
## [4] "Wald statistic: condition treated vs untreated"
## [5] "Wald test p-value: condition treated vs untreated"
## [6] "BH adjusted p-values"
```

For a particular gene, a log2 fold change of -1 for `condition treated vs untreated` means that the treatment induces a change in observed expression level of $2^{-1} = 0.5$ compared to the untreated condition. If the variable of interest is continuous-valued, then the reported log2 fold change is per unit of change of that variable.

Note: some values in the results table can be set to NA, for either one of the following reasons:

1. If within a row, all samples have zero counts, the baseMean column will be zero, and the log2 fold change estimates, p value and adjusted p value will all be set to NA.
2. If a row contains a sample with an extreme count outlier then the p value and adjusted p value are set to NA. These outlier counts are detected by Cook's distance. Customization of this outlier filtering and description of functionality for replacement of outlier counts and refitting is described in Section 3.6,
3. If a row is filtered by automatic independent filtering, based on low mean normalized count, then only the adjusted p value is set to NA. Description and customization of independent filtering is described in Section 3.8.

The column of log2FoldChange for the default workflow will contain shrunken estimates of fold change as mentioned above. It is possible to add a column to the results table – without rerunning the analysis – which contains the unshrunken, or maximum likelihood estimates (MLE), log2 fold changes. This will add the column lfcMLE directly after log2FoldChange.

```
head(results(dds, addMLE=TRUE),4)
```

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 4 rows and 7 columns
##               baseMean log2FoldChange    lfcMLE     lfcSE      stat    pvalue
##              <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0000003     0.159         0.0346    3.2929    0.0461    0.7487     0.454
## FBgn0000008    52.226         0.0197    0.0281    0.2093    0.0942     0.925
## FBgn0000014     0.390         0.0118    0.6239    0.0622    0.1901     0.849
## FBgn0000015     0.905        -0.0429   -0.8135    0.1054   -0.4076     0.684
##                  padj
##             <numeric>
## FBgn0000003        NA
## FBgn0000008     0.988
## FBgn0000014        NA
## FBgn0000015        NA
```

1.4.4 Exporting results to HTML or CSV files

An HTML report of the results with plots and sortable/filterable columns can be exported using the *ReportingTools* package on a *DESeqDataSet* that has been processed by the DESeq function. For a code example, see the "RNA-seq differential expression" vignette at the *ReportingTools* page, or the manual page for the publish method for the *DESeqDataSet* class.

A plain-text file of the results can be exported using the base *R* functions write.csv or write.delim. We suggest using a descriptive file name indicating the variable and levels which were tested.

```
write.csv(as.data.frame(resOrdered),
          file="condition_treated_results.csv")
```

Exporting only the results which pass an adjusted p value threshold can be accomplished with the subset function, followed by the write.csv function.

```
resSig <- subset(resOrdered, padj < 0.1)
resSig
```

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 803 rows and 6 columns
##              baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##             <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0039155       453          -3.71     0.160     -23.2 4.01e-119 3.11e-115
## FBgn0029167      2165          -2.08     0.104     -20.1  6.68e-90  2.59e-86
## FBgn0035085       367          -2.23     0.137     -16.3  1.89e-59  4.87e-56
## FBgn0029896       258          -2.21     0.159     -13.9  5.85e-44  1.13e-40
## FBgn0034736       118          -2.57     0.185     -13.9  8.07e-44  1.25e-40
## ...               ...            ...       ...       ...       ...       ...
## FBgn0000633      14.6          0.587     0.229      2.57    0.0103    0.0993
## FBgn0003862    2076.3         -0.400     0.156     -2.56    0.0103    0.0996
## FBgn0028694     960.2          0.313     0.122      2.57    0.0103    0.0996
## FBgn0038349    1275.4         -0.311     0.121     -2.56    0.0103    0.0996
## FBgn0243512    1135.0         -0.384     0.150     -2.56    0.0103    0.0996
```

1.5 Multi-factor designs

Experiments with more than one factor influencing the counts can be analyzed using model formula including the additional variables. The data in the *pasilla* package have a condition of interest (the column `condition`), as well as information on the type of sequencing which was performed (the column `type`), as we can see below:

```
colData(dds)
```

```
## DataFrame with 7 rows and 3 columns
##              condition        type sizeFactor
##               <factor>    <factor>  <numeric>
## treated1fb     treated single-read      1.512
## treated2fb     treated  paired-end      0.784
## treated3fb     treated  paired-end      0.896
## untreated1fb untreated single-read      1.050
## untreated2fb untreated single-read      1.659
## untreated3fb untreated  paired-end      0.712
## untreated4fb untreated  paired-end      0.784
```

We create a copy of the *DESeqDataSet*, so that we can rerun the analysis using a multi-factor design.

```
ddsMF <- dds
```

We can account for the different types of sequencing, and get a clearer picture of the differences attributable to the treatment. As `condition` is the variable of interest, we put it at the end of the formula. Thus the `results` function will by default pull the `condition` results unless `contrast` or `name` arguments are specified. Then we can re-run `DESeq`:

```
design(ddsMF) <- formula(~ type + condition)
ddsMF <- DESeq(ddsMF)
```

Again, we access the results using the `results` function.

```
resMF <- results(ddsMF)
head(resMF)
```

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 6 rows and 6 columns
##               baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##              <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0000003     0.159        0.03270    0.0437    0.7478    0.4546        NA
## FBgn0000008    52.226        0.01219    0.2078    0.0587    0.9532     0.988
## FBgn0000014     0.390        0.00971    0.0563    0.1724    0.8631        NA
## FBgn0000015     0.905       -0.03567    0.0932   -0.3828    0.7019        NA
## FBgn0000017  2358.243       -0.25675    0.1100   -2.3331    0.0196     0.134
## FBgn0000018   221.242       -0.06669    0.1417   -0.4706    0.6380     0.884
```

It is also possible to retrieve the log2 fold changes, p values and adjusted p values of the `type` variable. The `contrast` argument of the function `results` takes a character vector of length three: the name of the variable, the name of the factor level for the numerator of the log2 ratio, and the name of the factor level for the denominator. The `contrast` argument can also take other forms, as described in the help page for `results` and in Section 3.2.

```
resMFType <- results(ddsMF, contrast=c("type","single-read","paired-end"))
head(resMFType)
```

```
## log2 fold change (MAP): type single-read vs paired-end
## Wald test p-value: type single-read vs paired-end
## DataFrame with 6 rows and 6 columns
##               baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##              <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0000003     0.159       -0.02392    0.0381   -0.6280    0.5300        NA
## FBgn0000008    52.226       -0.06233    0.1974   -0.3158    0.7521     0.885
## FBgn0000014     0.390        0.00570    0.0490    0.1163    0.9074        NA
## FBgn0000015     0.905       -0.05581    0.0815   -0.6845    0.4937        NA
## FBgn0000017  2358.243        0.00938    0.1088    0.0862    0.9313     0.973
## FBgn0000018   221.242        0.26861    0.1383    1.9420    0.0521     0.210
```

If the variable is continuous or an interaction term (see Section 3.3) then the results can be extracted using the `name` argument to `results`, where the name is one of elements returned by `resultsNames(dds)`.

2 Data transformations and visualization

2.1 Count data transformations

In order to test for differential expression, we operate on raw counts and use discrete distributions as described in the previous Section 1.3. However for other downstream analyses – e.g. for visualization or clustering – it might be useful to work with transformed versions of the count data.

Maybe the most obvious choice of transformation is the logarithm. Since count values for a gene can be zero in some conditions (and non-zero in others), some advocate the use of *pseudocounts*, i. e. transformations of the form

$$y = \log_2(n+1) \quad \text{or more generally,} \quad y = \log_2(n+n_0), \tag{1}$$

where n represents the count values and n_0 is a positive constant.

In this section, we discuss two alternative approaches that offer more theoretical justification and a rational way of choosing the parameter equivalent to n_0 above. The *regularized logarithm* or *rlog* incorporates a prior on the sample differences [1], and the other uses the concept of variance stabilizing transformations (VST) [4, 5, 6]. Both transformations produce transformed data on the $\log_2$ scale which has been normalized with respect to library size.

The point of these two transformations, the *rlog* and the VST, is to remove the dependence of the variance on the mean, particularly the high variance of the logarithm of count data when the mean is low. Both *rlog* and VST use the experiment-wide trend of variance over mean, in order to transform the data to remove the experiment-wide trend. Note that we do not require or desire that all the genes have *exactly* the same variance after transformation. Indeed, in Figure 4 below, you will see that after the transformations the genes with the same mean do not have exactly the same standard deviations, but that the experiment-wide trend has flattened. It is those genes with row variance above the trend which will allow us to cluster samples into interesting groups.

2.1.1 Blind dispersion estimation

The two functions, `rlog` and `varianceStabilizingTransformation`, have an argument `blind`, for whether the transformation should be blind to the sample information specified by the design formula. When `blind` equals `TRUE` (the default), the functions will re-estimate the dispersions using only an intercept (design formula ~ 1). This setting should be used in order to compare samples in a manner wholly unbiased by the information about experimental groups, for example to perform sample QA (quality assurance) as demonstrated below.

However, blind dispersion estimation is not the appropriate choice if one expects that many or the majority of genes (rows) will have large differences in counts which are explanable by the experimental design, and one wishes to tranform the data for downstream analysis. In this case, using blind dispersion estimation will lead to large estimates of dispersion, as it attributes differences due to experimental design as unwanted "noise", and shrinks the tranformed values towards each other. By setting `blind` to `FALSE`, the dispersions already estimated will be used to perform transformations, or if not present, they will be estimated using the current design formula. Note that only the fitted dispersion estimates from mean-dispersion trend line is used in the transformation. So setting `blind` to `FALSE` is still mostly unbiased by the information about the samples.

2.1.2 Extracting transformed values

The two functions return an object of class *DESeqTransform* which is a subclass of *SummarizedExperiment*. The `assay` function is used to extract the matrix of normalized values:

```
rld <- rlog(dds)
vsd <- varianceStabilizingTransformation(dds)
rlogMat <- assay(rld)
vstMat <- assay(vsd)
```

Note that if you have many samples, and the `rlog` function is taking too long, there is an argument `fast=TRUE`, which will perform an approximation of the rlog: instead of shrinking the samples independently, the function will find the optimal amount of shrinkage for each gene given the mean counts. The optimization is performed on the same likelihood of the data as the original rlog. The speed-up for a dataset with 100 samples is around 15x.

2.1.3 Regularized log transformation

The function `rlog`, stands for *regularized log*, transforming the original count data to the log2 scale by fitting a model with a term for each sample and a prior distribution on the coefficients which is estimated from the data. This is the same kind of shrinkage (sometimes referred to as regularization, or moderation) of log fold changes used by the `DESeq` and `nbinomWaldTest`, as seen in Figure 1. The resulting data contains elements defined as:

$$\log_2(q_{ij}) = \beta_{i0} + \beta_{ij}$$

where q_{ij} is a parameter proportional to the expected true concentration of fragments for gene i and sample j (see Section 4.1), β_{i0} is an intercept which does not undergo shrinkage, and β_{ij} is the sample-specific effect which is shrunk toward zero based on the dispersion-mean trend over the entire dataset. The trend typically captures high dispersions for low counts, and therefore these genes exhibit higher shrinkage from the`rlog`.

Note that, as q_{ij} represents the part of the mean value μ_{ij} after the size factor s_j has been divided out, it is clear that the rlog transformation inherently accounts for differences in sequencing depth. Without priors, this design matrix would lead to a non-unique solution, however the addition of a prior on non-intercept betas allows for a unique solution to be found. The regularized log transformation is preferable to the variance stabilizing transformation if the size factors vary widely.

2.1.4 Variance stabilizing transformation

Above, we used a parametric fit for the dispersion. In this case, the closed-form expression for the variance stabilizing transformation is used by `varianceStabilizingTransformation`, which is derived in the file `vst.pdf`, that is distributed in the package alongside this vignette. If a local fit is used (option `fitType="locfit"` to `estimateDispersions`) a numerical integration is used instead.

The resulting variance stabilizing transformation is shown in Figure 3. The code that produces the figure is hidden from this vignette for the sake of brevity, but can be seen in the `.Rnw` or `.R` source file. Note that the vertical axis in such plots is the square root of the variance over all samples, so including the variance due to the experimental conditions. While a flat curve of the square root of variance over the mean may seem like

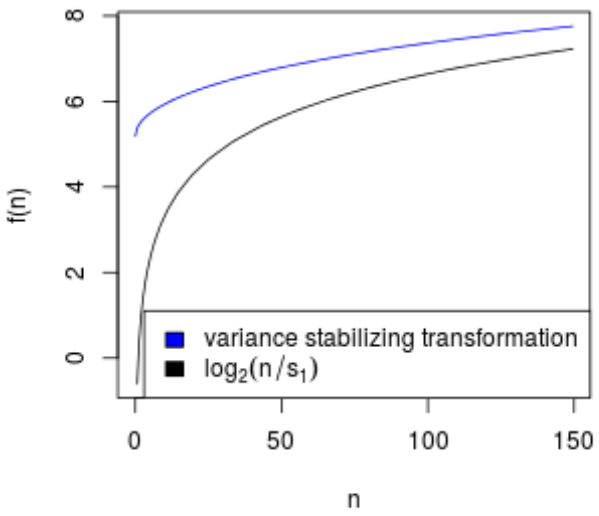

Figure 3: **VST and log2.** Graphs of the variance stabilizing transformation for sample 1, in blue, and of the transformation $f(n) = \log_2(n/s_1)$, in black. n are the counts and s_1 is the size factor for the first sample.

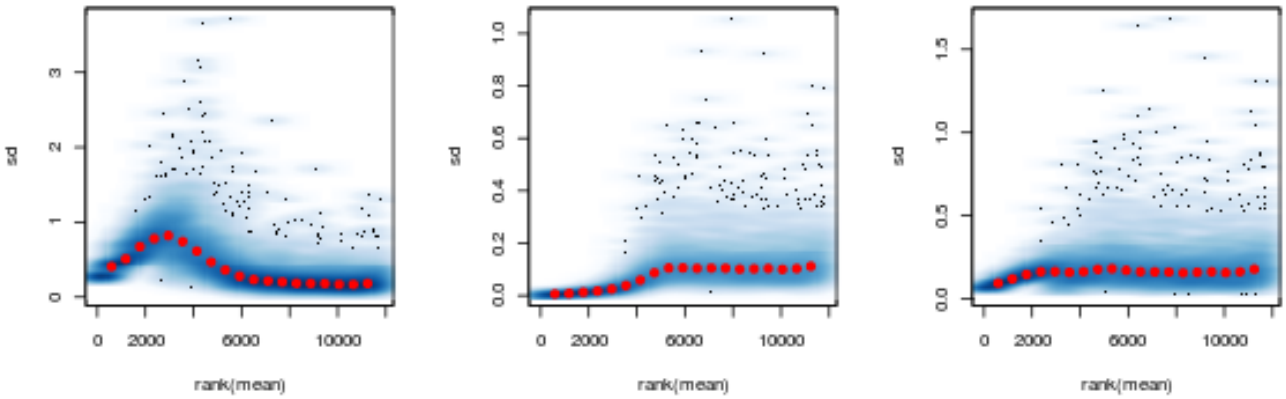

Figure 4: **Standard deviation over mean.** Per-gene standard deviation (taken across samples), against the rank of the mean, for the shifted logarithm $\log_2(n+1)$ (left), the regularized log transformation (center) and the variance stabilizing transformation (right).

the goal of such transformations, this may be unreasonable in the case of datasets with many true differences due to the experimental conditions.

2.1.5 Effects of transformations on the variance

Figure 4 plots the standard deviation of the transformed data, across samples, against the mean, using the shifted logarithm transformation (1), the regularized log transformation and the variance stabilizing transformation. The shifted logarithm has elevated standard deviation in the lower count range, and the regularized log to a lesser extent, while for the variance stabilized data the standard deviation is roughly constant along the whole dynamic range.

```
library("vsn")
par(mfrow=c(1,3))
notAllZero <- (rowSums(counts(dds))>0)
meanSdPlot(log2(counts(dds,normalized=TRUE)[notAllZero,] + 1))
meanSdPlot(assay(rld[notAllZero,]))
meanSdPlot(assay(vsd[notAllZero,]))
```

2.2 Data quality assessment by sample clustering and visualization

Data quality assessment and quality control (i. e. the removal of insufficiently good data) are essential steps of any data analysis. These steps should typically be performed very early in the analysis of a new data set, preceding or in parallel to the differential expression testing.

We define the term *quality* as *fitness for purpose*[3]. Our purpose is the detection of differentially expressed genes, and we are looking in particular for samples whose experimental treatment suffered from an anormality that renders the data points obtained from these particular samples detrimental to our purpose.

2.2.1 Heatmap of the count matrix

To explore a count matrix, it is often instructive to look at it as a heatmap. Below we show how to produce such a heatmap from the raw and transformed data.

```
library("RColorBrewer")
library("gplots")
select <- order(rowMeans(counts(dds,normalized=TRUE)),decreasing=TRUE)[1:30]
hmcol <- colorRampPalette(brewer.pal(9, "GnBu"))(100)
```

```
heatmap.2(counts(dds,normalized=TRUE)[select,], col = hmcol,
          Rowv = FALSE, Colv = FALSE, scale="none",
          dendrogram="none", trace="none", margin=c(10,6))
```

```
heatmap.2(assay(rld)[select,], col = hmcol,
          Rowv = FALSE, Colv = FALSE, scale="none",
          dendrogram="none", trace="none", margin=c(10, 6))
```

```
heatmap.2(assay(vsd)[select,], col = hmcol,
          Rowv = FALSE, Colv = FALSE, scale="none",
          dendrogram="none", trace="none", margin=c(10, 6))
```

[3] http://en.wikipedia.org/wiki/Quality_%28business%29

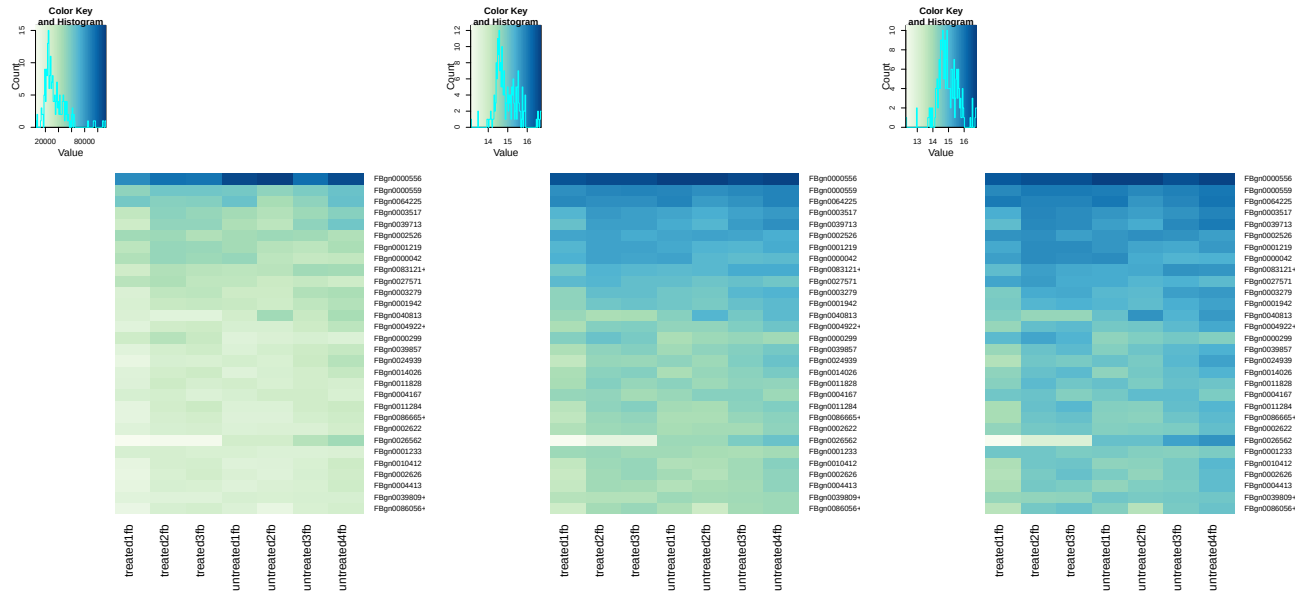

Figure 5: Heatmaps showing the expression data of the 30 most highly expressed genes. The data is of raw counts (left), from regularized log transformation (center) and from variance stabilizing transformation (right).

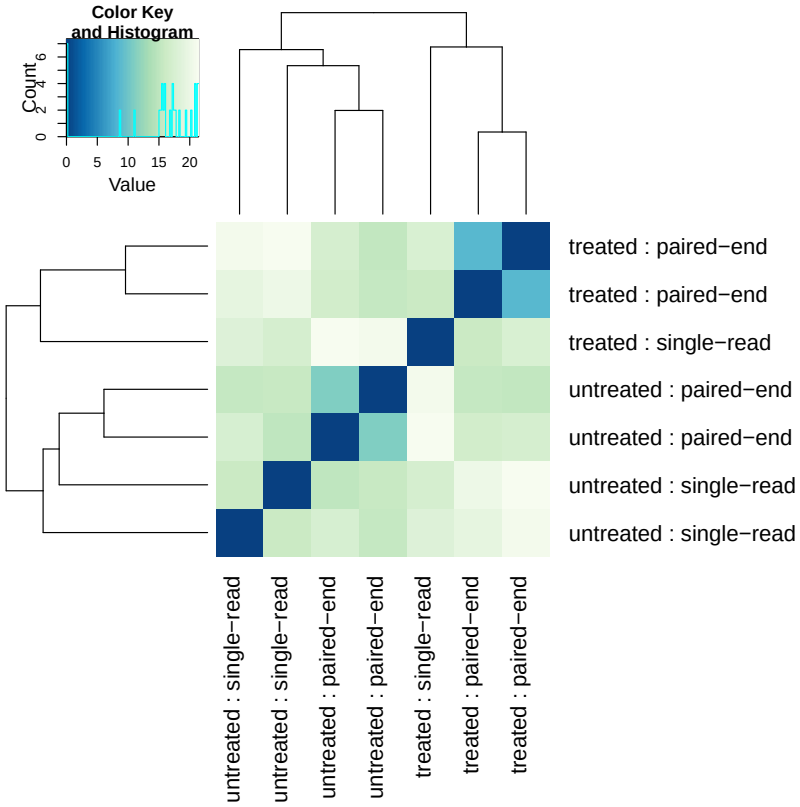

Figure 6: **Sample-to-sample distances.** Heatmap showing the Euclidean distances between the samples as calculated from the regularized log transformation.

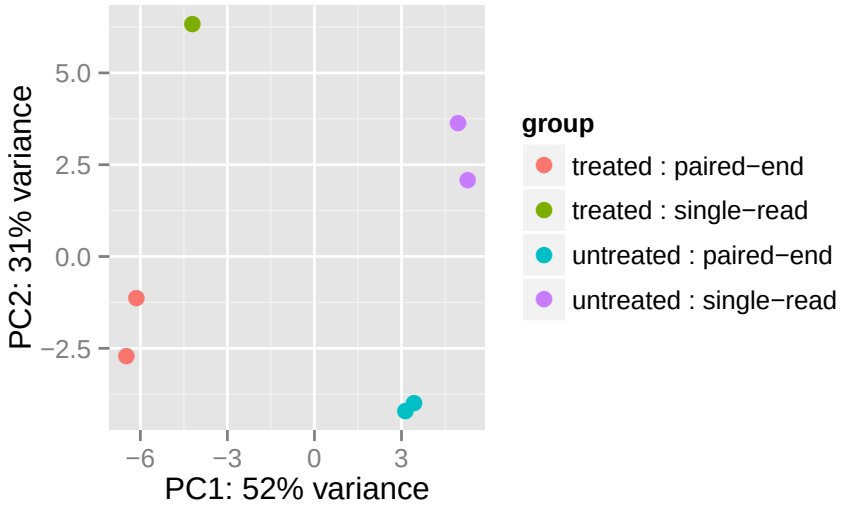

Figure 7: **PCA plot.** PCA plot. The 7 samples shown in the 2D plane spanned by their first two principal components. This type of plot is useful for visualizing the overall effect of experimental covariates and batch effects.

2.2.2 Heatmap of the sample-to-sample distances

Another use of the transformed data is sample clustering. Here, we apply the `dist` function to the transpose of the transformed count matrix to get sample-to-sample distances. We could alternatively use the variance stabilized transformation here.

```
distsRL <- dist(t(assay(rld)))
```

A heatmap of this distance matrix gives us an overview over similarities and dissimilarities between samples (Figure 6): We have to provide a hierarchical clustering `hc` to the `heatmap.2` function based on the sample distances, or else the `heatmap.2` function would calculate a clustering based on the distances between the rows/columns of the distance matrix.

```
mat <- as.matrix(distsRL)
rownames(mat) <- colnames(mat) <- with(colData(dds),
                                       paste(condition, type, sep=" : "))
hc <- hclust(distsRL)
heatmap.2(mat, Rowv=as.dendrogram(hc),
          symm=TRUE, trace="none",
          col = rev(hmcol), margin=c(13, 13))
```

2.2.3 Principal component plot of the samples

Related to the distance matrix of Section 2.2.2 is the PCA plot of the samples, which we obtain as follows (Figure 7).

```
plotPCA(rld, intgroup=c("condition", "type"))
```

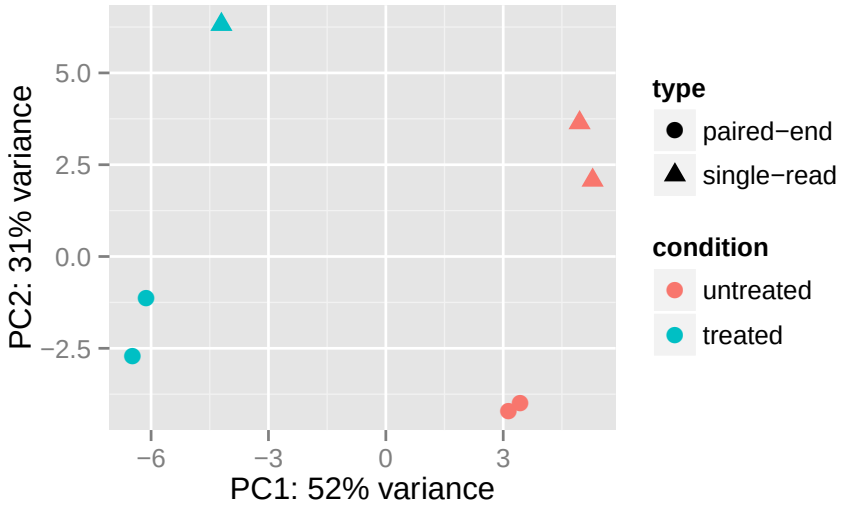

Figure 8: **PCA plot.** PCA plot customized using the *ggplot2* library.

It is also possible to customize the PCA plot using the ggplot function.

```
data <- plotPCA(rld, intgroup=c("condition", "type"), returnData=TRUE)
percentVar <- round(100 * attr(data, "percentVar"))
ggplot(data, aes(PC1, PC2, color=condition, shape=type)) +
  geom_point(size=3) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance"))
```

3 Variations to the standard workflow

3.1 Wald test individual steps

The function `DESeq` runs the following functions in order:

```
dds <- estimateSizeFactors(dds)
dds <- estimateDispersions(dds)
dds <- nbinomWaldTest(dds)
```

3.2 Contrasts

A contrast is a linear combination of estimated log2 fold changes, which can be used to test if differences between groups are equal to zero. The simplest use case for contrasts is an experimental design containing a factor with three levels, say A, B and C. Contrasts enable the user to generate results for all 3 possible differences: log2 fold change of B vs A, of C vs A, and of C vs B (the other three possible pairs will simply have $-1\times$ the log2 fold changes of these three).

In order to fit models with "shrunken" log2 fold changes in a manner which is independent to the choice of reference level, *DESeq2* uses "expanded model matrices", described further in Section 4.6. The expanded model matrices include a coefficient for each level of the factors in addition to an intercept. The `contrast` argument of `results` function is again used to extract test results of log2 fold changes of interest.

Log2 fold changes can also be added and subtracted by providing a `list` to the `contrast` argument with two elements: the names of the log2 fold changes to add, and the names of the log2 fold changes to subtract. Alternatively, a numeric vector of the length of `resultsNames(dds)` can be provided, for manually specifying the linear combination of terms. Demonstrations of the use of contrasts for various designs can be found in the examples section of the help page for the `results` function. The formula that is used to generate the contrasts can be found in Section 4.5.

3.3 Interactions

Interaction terms can be added to the design formula, in order to test, for example, if the log2 fold change attributable to a given condition is different based on a second variable, for example if the treatment effect of a drug is differs based on another grouping variable like species. Interactions are specified in R formula using a colon, `:`, between the two variables names. Demonstrations of extracting results for many different kinds of interaction model are shown in the examples section of the help page for `results`. When interations are present in the design, and a prior is used for the log2 fold changes, then only the interaction terms are moderated, and the main effect terms are not moderated. See `?nbinomWaldTest` for more details.

Note that if the comparisons of interest are, for example, the effect of a condition in different sets of samples, many users find that a simpler design which likewise allows for a set-specific condition effect is the following: first combine the factors of interest into a single factor with all combinations of the original factors, and then use a design with just this factor. For example, if we have two factors `set` and `condition`, and we want to test for the set-specific condition effects in all sets:

```
dds$group <- factor(paste0(dds$set, dds$condition))
design(dds) <- ~ group
```

```
dds <- DESeq(dds)
resultsNames(dds)
results(dds, contrast=c("group", "1B", "1A"))
```

If one wants to test whether the condition effect is *different* across sets, then a design with interactions makes these comparisons easier to test. It is important to remember that interactions are *additional terms* (addition in the log scale), beyond the *main effect* terms. If the *main effect* of a condition on counts is a log2 fold change of 2, and the *interaction term* for the condition in a particular set Y is a log2 fold change of 0.5, then the log2 fold change for the condition for set Y is $2 + 0.5 = 2.5$. Testing the interation term alone tests if the condition effect in set Y is *different* than in the other set(s). Testing the *sum* of the main effect and the interaction term tests if there is a condition effect in set Y. This is shown in more detail in the examples section of the help page for `results`.

Designs with factors with only two levels and an interaction term present a special case to the normal *DESeq2* methods, because we change the type of model matrix used in the fitting, using what we call R's "standard model matrix". For all other designs, we use an "expanded model matrix" which fits a term for each level in the model, so there are no reference levels. The reason to use a standard model matrix for this special case of factors with two levels and an interaction term is that it is desirable to have only a single term to test for a significant interaction between the variables. Furthermore, the expanded design matrices are not necessary in this case for their purpose (for other designs, they provide symmetric shrinkage regardless of the reference level).

This does create a difference in the interpretation of the contrasts extracted with `results` for designs with interactions: for expanded design matrices, the main effect is an average effect across all sets; while for standard design matrices, the main effect is only for the reference level of the other factors. We explain in more detail below:

Suppose we have an interaction design of $\sim$ `set + condition + set:condition`.

Expanded model matrices The following code produces a comparison of B over A *across all sets*.

```
results(dds, contrast=c("condition","B","A"))
```

Standard model matrices The above results table now produces a comparison of B over A only for the *reference level* of `set`. Let us suppose the reference level of set is X, and the second set is Y. So for standard model matrices, the above results table gives the condition effect for set X.

Below we explain how to interpret the interaction terms for both kinds of model matrix.

Expanded model matrices The following code would give the *interaction effect* for condition in set Y, which tests if the condition effect is different in set Y than in the other sets.

```
results(dds, contrast=list("setY.conditionB","setY.conditionA"))
```

And the following would give the condition effect in set Y (the main effect and the interaction terms added together):

```
results(dds, contrast=list(c("conditionB","setY.conditionB"),
                           c("conditionA","setY.conditionA")))
```

Standard model matrices Here, we have only a single interaction term, `setY.conditionB`. This term represents the additional effect of condition for set Y compared to the main effect of condition in set X. By symmetry, it also represents the additional effect of set in condition B compared to the main effect of set in

condition A. Testing if the condition effect is different between the two sets is accomplished by extracting only this interaction term:

```
results(dds, name="setY.conditionB")
```

The effect of condition in set Y is the above interaction term added to the main effect of condition B vs A in set X:

```
results(dds, contrast=list(c("conditionB","setY.conditionB")))
```

For standard model matrices, to obtain the "average condition effect" for both sets, we can use a numeric contrast, assigning a 1 to the main effect `conditionB` and a 0.5 to the interaction effect `setY.conditionB`, as they appear in `resultsNames(dds)`. This can be understood as taking us halfway from the condition effect in set X to the condition effect in set Y:

```
resultsNames(dds)
results(dds, contrast=c(0,0,1,0.5))
```

You can use the following lines of code to find out the model matrix type and to extract the model matrix which was used by `DESeq`. For models with shrinkage on log2 fold changes (`betaPrior=TRUE`), the type will be `"expanded"` unless the design has factors with only two levels and an interaction term is included, in which case it will be `"standard"`.

```
attr(dds, "modelMatrixType")
attr(dds, "modelMatrix")
```

3.4 Time-series experiments

There are a number of ways to analyze time-series experiments, depending on the biological question of interest. In order to test for any differences over multiple time points, once can use a design including the time factor, and then test using the likelihood ratio test as described in Section 3.5, where the time factor is removed in the reduced formula. For a control and treatment time series, one can use a design formula containing the condition factor, the time factor, and the interaction of the two. In this case, using the likelihood ratio test with a reduced model which does not contain the interaction terms will test whether the condition induces a change in gene expression at any time point after the reference level time point (time 0). An example of the later analysis is provided in an RNA-Seq workflow on the Bioconductor website: http://www.bioconductor.org/help/workflows/rnaseqGene/.

3.5 Likelihood ratio test

DESeq2 offers two kinds of hypothesis tests: the Wald test, where we use the estimated standard error of a log2 fold change to test if it is equal to zero, and the likelihood ratio test (LRT). The LRT examines two models for the counts, a *full* model with a certain number of terms and a *reduced* model, in which some of the terms of the *full* model are removed. The test determines if the increased likelihood of the data using the extra terms in the *full* model is more than expected if those extra terms are truly zero.

The LRT is therefore useful for testing multiple terms at once, for example testing 3 or more levels of a factor at once, or all interactions between two variables. The LRT for count data is conceptually similar to an analysis of variance (ANOVA) calculation in linear regression, except that in the case of the Negative Binomial GLM,

we use an analysis of deviance (ANODEV), where the *deviance* captures the difference in likelihood between a full and a reduced model.

The likelihood ratio test can be specified using the `test` argument to `DESeq`, which substitutes `nbinomWaldTest` with `nbinomLRT`. In this case, the user needs to provide a reduced formula, e.g. one in which a number of terms from `design(dds)` are removed. The degrees of freedom for the test is obtained from the difference between the number of parameters in the two models.

3.6 Approach to count outliers

RNA-Seq data sometimes contain isolated instances of very large counts that are apparently unrelated to the experimental or study design, and which may be considered outliers. There are many reasons why outliers can arise, including rare technical or experimental artifacts, read mapping problems in the case of genetically differing samples, and genuine, but rare biological events. In many cases, users appear primarily interested in genes that show a consistent behavior, and this is the reason why by default, genes that are affected by such outliers are set aside by *DESeq2*, or if there are sufficient samples, outlier counts are replaced for model fitting. These two behaviors are described below.

The `DESeq` function (and `nbinomWaldTest`/`nbinomLRT` functions) calculates, for every gene and for every sample, a diagnostic test for outliers called *Cook's distance*. Cook's distance is a measure of how much a single sample is influencing the fitted coefficients for a gene, and a large value of Cook's distance is intended to indicate an outlier count. The Cook's distances are stored as a matrix available in `assays(dds)[["cooks"]]`.

The `results` function automatically flags genes which contain a Cook's distance above a cutoff for samples which have 3 or more replicates. The p values and adjusted p values for these genes are set to `NA`. At least 3 replicates are required for flagging, as it is difficult to judge which sample might be an outlier with only 2 replicates.

With many degrees of freedom – i. e., many more samples than number of parameters to be estimated – it is undesirable to remove entire genes from the analysis just because their data include a single count outlier. When there are 7 or more replicates for a given sample, the `DESeq` function will automatically replace counts with large Cook's distance with the trimmed mean over all samples, scaled up by the size factor or normalization factor for that sample. This approach is conservative, it will not lead to false positives, as it replaces the outlier value with the value predicted by the null hypothesis.

The default Cook's distance cutoff for the two behaviors described above depends on the sample size and number of parameters to be estimated. The default is to use the 99% quantile of the $F(p, m-p)$ distribution (with p the number of parameters including the intercept and m number of samples). The default for gene flagging can be modified using the `cooksCutoff` argument to the `results` function. The gene flagging functionality can be disabled by setting `cooksCutoff` to `FALSE` or `Inf`. The automatic outlier replacement performed by `DESeq` can be disabled by setting the `minReplicatesForReplace` argument to `Inf`.

`DESeq` automatically replaces outliers if there are sufficient replicates and a row contains a count with very high Cook's distance. `DESeq` preserves the original counts in `counts(dds)` saving the replacement counts as a matrix named `replaceCounts` in `assays(dds)`.

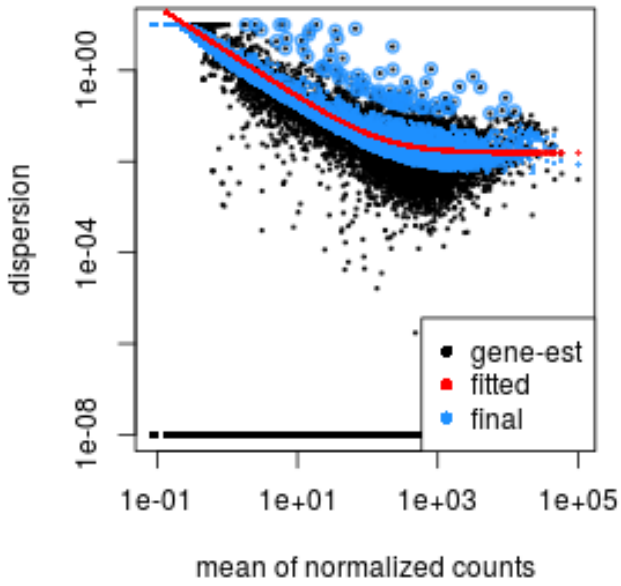

Figure 9: **Dispersion plot.** The dispersion estimate plot shows the gene-wise estimates (black), the fitted values (red), and the final maximum *a posteriori* estimates used in testing (blue).

3.7 Dispersion plot and fitting alternatives

Plotting the dispersion estimates is a useful diagnostic. The dispersion plot in Figure 9 is typical, with the final estimates shrunk from the gene-wise estimates towards the fitted estimates. Some gene-wise estimates are flagged as outliers and not shrunk towards the fitted value, (this outlier detection is described in the man page for `estimateDispersionsMAP`). The amount of shrinkage can be more or less than seen here, depending on the sample size, the number of coefficients, the row mean and the variability of the gene-wise estimates.

```
plotDispEsts(dds)
```

3.7.1 Local or mean dispersion fit

A local smoothed dispersion fit is automatically substitited in the case that the parametric curve doesn't fit the observed dispersion mean relationship. This can be prespecified by providing the argument `fitType="local"` to either `DESeq` or `estimateDispersions`. Additionally, using the mean of gene-wise disperion estimates as the fitted value can be specified by providing the argument `fitType="mean"`.

3.7.2 Supply a custom dispersion fit

Any fitted values can be provided during dispersion estimation, using the lower-level functions described in the manual page for `estimateDispersionsGeneEst`. In the code chunk below, we store the gene-wise estimates

which were already calculated and saved in the metadata column `dispGeneEst`. Then we calculate the median value of the dispersion estimates above a threshold, and save these values as the fitted dispersions, using the replacement function for `dispersionFunction`. In the last line, the function `estimateDispersionsMAP`, uses the fitted dispersions to generate maximum *a posteriori* (MAP) estimates of dispersion.

```
ddsCustom <- dds
useForMedian <- mcols(ddsCustom)$dispGeneEst > 1e-7
medianDisp <- median(mcols(ddsCustom)$dispGeneEst[useForMedian],na.rm=TRUE)
dispersionFunction(ddsCustom) <- function(mu) medianDisp
ddsCustom <- estimateDispersionsMAP(ddsCustom)
```

3.8 Independent filtering of results

The `results` function of the *DESeq2* package performs independent filtering by default using the mean of normalized counts as a filter statistic. A threshold on the filter statistic is found which optimizes the number of adjusted p values lower than a significance level `alpha` (we use the standard variable name for significance level, though it is unrelated to the dispersion parameter α). The theory behind independent filtering is discussed in greater detail in Section 4.7. The adjusted p values for the genes which do not pass the filter threshold are set to `NA`.

The independent filtering is performed using the `filtered_p` function of the *genefilter* package, and all of the arguments of `filtered_p` can be passed to the `results` function. The filter threshold value and the number of rejections at each quantile of the filter statistic are available as attributes of the object returned by `results`. For example, we can visualize the optimization by plotting the `filterNumRej` attribute of the results object, as seen in Figure 10. Note that if the maximum number of rejections is very small such that the line of rejections over filter threshold appears noisy, the expected false discovery rate might not be held exactly for this small set.

```
attr(res,"filterThreshold")
```

```
## 46.5%
##   9.2
```

```
plot(attr(res,"filterNumRej"),type="b",
     ylab="number of rejections")
```

Independent filtering can be turned off by setting `independentFiltering` to `FALSE`.

```
resNoFilt <- results(dds, independentFiltering=FALSE)
addmargins(table(filtering=(res$padj < .1), noFiltering=(resNoFilt$padj < .1)))
```

```
##          noFiltering
## filtering FALSE TRUE  Sum
##     FALSE  6934    0 6934
##     TRUE    108  695  803
##     Sum    7042  695 7737
```

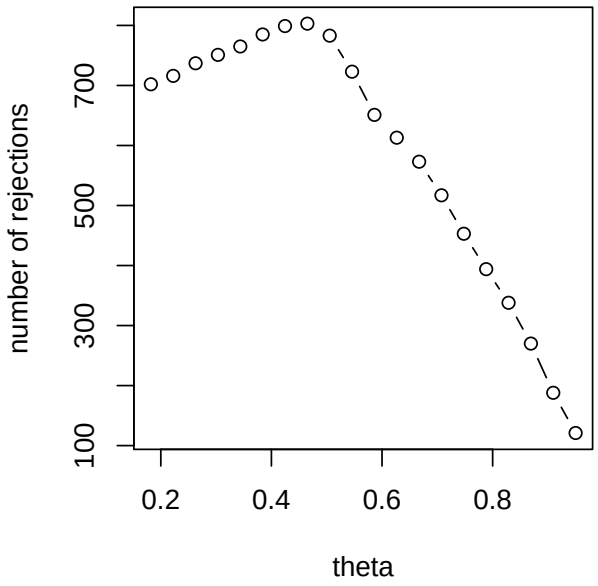

Figure 10: **Independent filtering.** The `results` function maximizes the number of rejections (adjusted p value less than a significance level), over theta, the quantiles of a filtering statistic (in this case, the mean of normalized counts).

3.9 Tests of log2 fold change above or below a threshold

It is also possible to provide thresholds for constructing Wald tests of significance. Two arguments to the `results` function allow for threshold-based Wald tests: `lfcThreshold`, which takes a numeric of a non-negative threshold value, and `altHypothesis`, which specifies the kind of test. Note that the *alternative hypothesis* is specified by the user, i.e. those genes which the user is interested in finding, and the test provides p values for the null hypothesis, the complement of the set defined by the alternative. The `altHypothesis` argument can take one of the following four values, where β is the log2 fold change specified by the `name` argument:

- `greaterAbs` - $|\beta| > \text{lfcThreshold}$ - tests are two-tailed
- `lessAbs` - $|\beta| < \text{lfcThreshold}$ - p values are the maximum of the upper and lower tests
- `greater` - $\beta > \text{lfcThreshold}$
- `less` - $\beta < -\text{lfcThreshold}$

The test `altHypothesis="lessAbs"` requires that the user have run `DESeq` with the argument `betaPrior=FALSE`. To understand the reason for this requirement, consider that during hypothesis testing, the null hypothesis is favored unless the data provide strong evidence to reject the null. For this test, including a zero-centered prior on log fold change would favor the alternative hypothesis, shrinking log fold changes toward zero. Removing the prior on log fold changes for tests of small log fold change allows for detection of only those genes where the data alone provides evidence against the null.

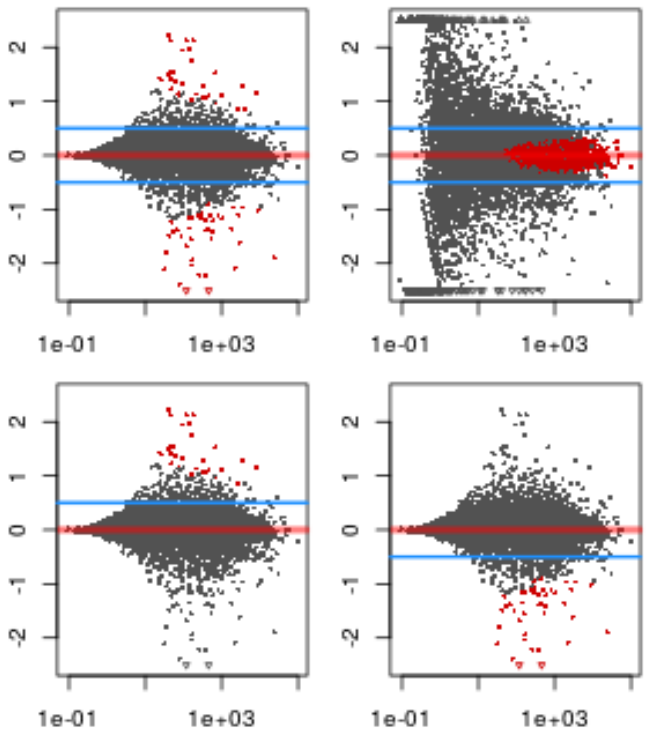

Figure 11: **MA-plots of tests of log2 fold change with respect to a threshold value.** Going left to right across rows, the tests are for `altHypothesis = "greaterAbs"`, `"lessAbs"`, `"greater"`, and `"less"`.

The four possible values of `altHypothesis` are demonstrated in the following code and visually by MA-plots in Figure 11. First we run DESeq and specify `betaPrior=FALSE` in order to demonstrate `altHypothesis="lessAbs"`.

```
ddsNoPrior <- DESeq(dds, betaPrior=FALSE)
```

In order to produce results tables for the following tests, the same arguments (except `ylim`) would be provided to the `results` function.

```
par(mfrow=c(2,2),mar=c(2,2,1,1))
yl <- c(-2.5,2.5)

resGA <- results(dds, lfcThreshold=.5, altHypothesis="greaterAbs")
resLA <- results(ddsNoPrior, lfcThreshold=.5, altHypothesis="lessAbs")
resG <- results(dds, lfcThreshold=.5, altHypothesis="greater")
resL <- results(dds, lfcThreshold=.5, altHypothesis="less")

plotMA(resGA, ylim=yl)
abline(h=c(-.5,.5),col="dodgerblue",lwd=2)
plotMA(resLA, ylim=yl)
abline(h=c(-.5,.5),col="dodgerblue",lwd=2)
plotMA(resG, ylim=yl)
abline(h=.5,col="dodgerblue",lwd=2)
plotMA(resL, ylim=yl)
abline(h=-.5,col="dodgerblue",lwd=2)
```

3.10 Access to all calculated values

All row-wise calculated values (intermediate dispersion calculations, coefficients, standard errors, etc.) are stored in the *DESeqDataSet* object, e.g. `dds` in this vignette. These values are accessible by calling `mcols` on `dds`. Descriptions of the columns are accessible by two calls to `mcols`.

```
mcols(dds,use.names=TRUE)[1:4,1:4]
```

```
## DataFrame with 4 rows and 4 columns
##                    gene  baseMean   baseVar   allZero
##                <factor> <numeric> <numeric> <logical>
## FBgn0000003 FBgn0000003     0.159     0.178     FALSE
## FBgn0000008 FBgn0000008    52.226   154.611     FALSE
## FBgn0000014 FBgn0000014     0.390     0.444     FALSE
## FBgn0000015 FBgn0000015     0.905     0.799     FALSE
```

```
# here using substr() only for display purposes
substr(names(mcols(dds)),1,10)
```

```
##  [1] "gene"       "baseMean"   "baseVar"    "allZero"    "dispGeneEs"
##  [6] "dispFit"    "dispersion" "dispIter"   "dispOutlie" "dispMAP"
## [11] "Intercept"  "conditionu" "conditiont" "SE_Interce" "SE_conditi"
## [16] "SE_conditi" "MLE_Interc" "MLE_condit" "WaldStatis" "WaldStatis"
## [21] "WaldStatis" "WaldPvalue" "WaldPvalue" "WaldPvalue" "betaConv"
## [26] "betaIter"   "deviance"   "maxCooks"
```

```
mcols(mcols(dds), use.names=TRUE)[1:4,]
```

```
## DataFrame with 4 rows and 2 columns
##                  type                                     description
##           <character>                                     <character>
## gene            input
## baseMean  intermediate     mean of normalized counts for all samples
## baseVar   intermediate variance of normalized counts for all samples
## allZero   intermediate                all counts for a gene are zero
```

The mean values $\mu_{ij} = s_j q_{ij}$ and the Cook's distances for each gene and sample are stored as matrices in the assays slot:

```
head(assays(dds)[["mu"]])
```

```
##             treated1fb treated2fb treated3fb untreated1fb untreated2fb
## FBgn0000003      0.232      0.121      0.138        0.158        0.249
## FBgn0000008     79.292     41.141     46.989       54.328       85.815
## FBgn0000014      0.593      0.308      0.352        0.409        0.646
## FBgn0000015      1.287      0.668      0.762        0.921        1.454
## FBgn0000017   3208.117   1664.554   1901.137     2659.781     4201.345
## FBgn0000018    322.080    167.113    190.865      240.407      379.743
##             untreated3fb untreated4fb
## FBgn0000003        0.107        0.118
## FBgn0000008       36.828       40.552
## FBgn0000014        0.277        0.305
```

```
## FBgn0000015        0.624        0.687
## FBgn0000017     1803.025     1985.333
## FBgn0000018      162.968      179.446
```

```
head(assays(dds)[["cooks"]])
```

```
##             treated1fb treated2fb treated3fb untreated1fb untreated2fb
## FBgn0000003   0.101057   0.040231    1.91947       0.0348     0.054763
## FBgn0000008   0.001619   0.052975    0.03061       0.0522     0.005116
## FBgn0000014   1.462284   0.100402    0.12300       0.0931     0.181793
## FBgn0000015   0.026910   0.168946    0.02097       0.1773     0.031421
## FBgn0000017   0.000413   0.000182    0.00451       0.0361     0.054716
## FBgn0000018   0.209919   0.078047    0.04838       0.2022     0.000436
##             untreated3fb untreated4fb
## FBgn0000003       0.0236      0.02601
## FBgn0000008       0.4135      0.25905
## FBgn0000014       0.3425      0.05883
## FBgn0000015       0.0362      0.42199
## FBgn0000017       0.1121      0.10367
## FBgn0000018       0.1291      0.00422
```

The dispersions α_i can be accessed with the `dispersions` function.

```
head(dispersions(dds))
```

```
## [1] 10.0000  0.0538  6.3982  1.7320  0.0133  0.0220
```

```
# which is the same as
head(mcols(dds)$dispersion)
```

```
## [1] 10.0000  0.0538  6.3982  1.7320  0.0133  0.0220
```

The size factors s_j are accessible via `sizeFactors`:

```
sizeFactors(dds)
```

```
##   treated1fb   treated2fb   treated3fb untreated1fb untreated2fb untreated3fb
##        1.512        0.784        0.896        1.050        1.659        0.712
## untreated4fb
##        0.784
```

For advanced users, we also include a convenience function `coef` for extracting the matrix of coefficients $[\beta_{ir}]$ for all genes i and parameters r, as in the formula in Section 4.1. This function can also return a matrix of standard errors, see `?coef`. The columns of this matrix correspond to the effects returned by `resultsNames`. Note that the `results` function is best for building results tables with p values and adjusted p values.

```
head(coef(dds))
```

```
##             Intercept conditionuntreated conditiontreated
## FBgn0000003    -2.719           -0.01728          0.01727
## FBgn0000008     5.703           -0.00986          0.00986
## FBgn0000014    -1.355           -0.00591          0.00591
## FBgn0000015    -0.211            0.02147         -0.02147
```

```
## FBgn0000017    11.179           0.12768         -0.12768
## FBgn0000018     7.787           0.05192         -0.05192
```

The beta prior variance σ_r^2 is stored as an attribute of the *DESeqDataSet*:

```
attr(dds, "betaPriorVar")
```

```
##          Intercept conditionuntreated   conditiontreated
##           1.00e+06           1.05e-01           1.05e-01
```

The dispersion prior variance σ_d^2 is stored as an attribute of the dispersion function:

```
dispersionFunction(dds)
```

```
## function (q)
## coefs[1] + coefs[2]/q
## <environment: 0x96445a0>
## attr(,"coefficients")
## asymptDisp  extraPois
##     0.0154     2.5652
## attr(,"fitType")
## [1] "parametric"
## attr(,"varLogDispEsts")
## [1] 0.961
## attr(,"dispPriorVar")
## [1] 0.47
```

```
attr(dispersionFunction(dds), "dispPriorVar")
```

```
## [1] 0.47
```

3.11 Sample-/gene-dependent normalization factors

In some experiments, there might be gene-dependent dependencies which vary across samples. For instance, GC-content bias or length bias might vary across samples coming from different labs or processed at different times. We use the terms "normalization factors" for a gene $\times$ sample matrix, and "size factors" for a single number per sample. Incorporating normalization factors, the mean parameter μ_{ij} from Section 4.1 becomes:

$$\mu_{ij} = NF_{ij} q_{ij}$$

with normalization factor matrix NF having the same dimensions as the counts matrix K. This matrix can be incorporated as shown below. We recommend providing a matrix with row-wise geometric means of 1, so that the mean of normalized counts for a gene is close to the mean of the unnormalized counts. This can be accomplished by dividing out the current row geometric means.

```
normFactors <- normFactors / exp(rowMeans(log(normFactors)))
normalizationFactors(dds) <- normFactors
```

These steps then replace `estimateSizeFactors` in the steps described in Section 3.1. Normalization factors, if present, will always be used in the place of size factors.

The methods provided by the *cqn* or *EDASeq* packages can help correct for GC or length biases. They both describe in their vignettes how to create matrices which can be used by *DESeq2*. From the formula above, we see that normalization factors should be on the scale of the counts, like size factors, and unlike offsets which are typically on the scale of the predictors (i.e. the logarithmic scale for the negative binomial GLM). At the time of writing, the transformation from the matrices provided by these packages should be:

```
cqnOffset <- cqnObject$glm.offset
cqnNormFactors <- exp(cqnOffset)
EDASeqNormFactors <- exp(-1 * EDASeqOffset)
```

3.12 Model matrix not full rank

While most experimental designs run easily using design formula, some design formulas can cause problems and result in the `DESeq` function returning an error with the text: "the model matrix is not full rank, so the model cannot be fit as specified." There are two reasons for this problem, either one or more columns in the model matrix are linear combinations of other columns, or there are levels of factors or combinations of levels of multiple factors which are missing samples. We address these problems below and discuss possible solutions:

3.12.1 Linear combinations

The simplest case of the linear combination, or linear dependency problem, is if two variables contain exactly the same information, such as in the following sample table. The software cannot fit an effect for `batch` and `condition`, because they produce identical columns in the model matrix. This is also referred to as "confounding". A unique solution of coefficients (the β_i in the following Section 4.1) is not possible.

```
##   batch condition
## 1     1         A
## 2     1         A
## 3     2         B
## 4     2         B
```

Another situation which will cause problems is when the variables are not identical, but one variable can be formed by the combination of other factor levels. In the following example, the effect of batch 2 vs 1 cannot be fit because it is identical to a column in the model matrix which represents the condition C vs A effect.

```
##   batch condition
## 1     1         A
## 2     1         A
## 3     1         B
## 4     1         B
## 5     2         C
## 6     2         C
```

In both of these cases, the batch effect cannot be fit and must be removed from the model formula. The options are either to assume there is no batch effect (which we know is highly unlikely given the literature on batch effects in sequencing datasets) or to repeat the experiment and properly balance the conditions across batches, for example:

```
##   batch condition
## 1     1         A
## 2     1         B
## 3     1         C
## 4     2         A
## 5     2         B
## 6     2         C
```

Finally, there is the case of an experiment with grouped individuals, where a group-specific effect of some treatment is sought. This design requires a bit of refactoring in order to fit the model terms. Here we use a trick from the *edgeR* user guide, from the section "Comparisons Both Between and Within Subjects". An example of such an experimental design is:

```
(coldata <- data.frame(grp=factor(rep(c("X","Y"),each=4)),
                       ind=factor(rep(1:4,each=2)),
                       cnd=factor(rep(c("A","B"),4))))
```

```
##   grp ind cnd
## 1   X   1   A
## 2   X   1   B
## 3   X   2   A
## 4   X   2   B
## 5   Y   3   A
## 6   Y   3   B
## 7   Y   4   A
## 8   Y   4   B
```

If we try to analyze with a formula such as, $\sim$ `ind + grp*cnd`, we will obtain an error, because the effect for group is a linear combination of the individuals.

However, the following steps allow for an analysis of group-specific condition effects, while controlling for differences in individual. For object construction, use a dummy design, such as $\sim$ `1`. Then add a column `ind.n` which distinguishes the individuals "nested" within a group. Here, we add this column to coldata, but in practice you would add this column to `dds`.

```
coldata$ind.n <- factor(rep(rep(1:2,each=2),2))
coldata
```

```
##   grp ind cnd ind.n
## 1   X   1   A     1
## 2   X   1   B     1
## 3   X   2   A     2
## 4   X   2   B     2
## 5   Y   3   A     1
## 6   Y   3   B     1
## 7   Y   4   A     2
## 8   Y   4   B     2
```

Now we can reassign our *DESeqDataSet* a design of $\sim$ `grp + grp:ind.n + grp:cnd`, before we call `DESeq`. This new design will result in the following model matrix:

```
model.matrix(~ grp + grp:ind.n + grp:cnd, coldata)
```

```
##   (Intercept) grpY grpX:ind.n2 grpY:ind.n2 grpX:cndB grpY:cndB
## 1           1    0           0           0         0         0
## 2           1    0           0           0         1         0
## 3           1    0           1           0         0         0
## 4           1    0           1           0         1         0
## 5           1    1           0           0         0         0
## 6           1    1           0           0         0         1
## 7           1    1           0           1         0         0
## 8           1    1           0           1         0         1
## attr(,"assign")
## [1] 0 1 2 2 3 3
## attr(,"contrasts")
## attr(,"contrasts")$grp
## [1] "contr.treatment"
##
## attr(,"contrasts")$ind.n
## [1] "contr.treatment"
##
## attr(,"contrasts")$cnd
## [1] "contr.treatment"
```

where the terms `grpX.cndB` and `grpY.cndB` give the group-specific condition effects.

3.12.2 Levels without samples

The base R function for creating model matrices will produce a column of zeros if a level is missing from a factor or a combination of levels is missing from an interaction of factors. The solution to the first case is to call `droplevels` on the column, which will remove levels without samples. This was shown in the beginning of this vignette.

The second case is also solvable, by manually editing the model matrix, and then providing this to `DESeq`. Here we construct an example dataset to illustrate:

```
group <- factor(rep(1:3,each=6))
condition <- factor(rep(rep(c("A","B","C"),each=2),3))
(d <- data.frame(group, condition)[-c(17,18),])
```

```
##    group condition
## 1      1         A
## 2      1         A
## 3      1         B
## 4      1         B
## 5      1         C
## 6      1         C
## 7      2         A
## 8      2         A
## 9      2         B
```

```
## 10     2         B
## 11     2         C
## 12     2         C
## 13     3         A
## 14     3         A
## 15     3         B
## 16     3         B
```

Note that if we try to estimate all interaction terms, we introduce a column with all zeros, as there are no condition C samples for group 3. (Here, `unname` is used to display the matrix concisely.)

```
m1 <- model.matrix(~ condition*group, d)
colnames(m1)
```

```
## [1] "(Intercept)"       "conditionB"        "conditionC"
## [4] "group2"            "group3"            "conditionB:group2"
## [7] "conditionC:group2" "conditionB:group3" "conditionC:group3"
```

```
unname(m1)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##  [1,]    1    0    0    0    0    0    0    0    0
##  [2,]    1    0    0    0    0    0    0    0    0
##  [3,]    1    1    0    0    0    0    0    0    0
##  [4,]    1    1    0    0    0    0    0    0    0
##  [5,]    1    0    1    0    0    0    0    0    0
##  [6,]    1    0    1    0    0    0    0    0    0
##  [7,]    1    0    0    1    0    0    0    0    0
##  [8,]    1    0    0    1    0    0    0    0    0
##  [9,]    1    1    0    1    0    1    0    0    0
## [10,]    1    1    0    1    0    1    0    0    0
## [11,]    1    0    1    1    0    0    1    0    0
## [12,]    1    0    1    1    0    0    1    0    0
## [13,]    1    0    0    0    1    0    0    0    0
## [14,]    1    0    0    0    1    0    0    0    0
## [15,]    1    1    0    0    1    0    0    1    0
## [16,]    1    1    0    0    1    0    0    1    0
## attr(,"assign")
## [1] 0 1 1 2 2 3 3 3 3
## attr(,"contrasts")
## attr(,"contrasts")$condition
## [1] "contr.treatment"
##
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

We can remove this column like so:

```
m1 <- m1[,-9]
unname(m1)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
##  [1,]    1    0    0    0    0    0    0    0
##  [2,]    1    0    0    0    0    0    0    0
##  [3,]    1    1    0    0    0    0    0    0
##  [4,]    1    1    0    0    0    0    0    0
##  [5,]    1    0    1    0    0    0    0    0
##  [6,]    1    0    1    0    0    0    0    0
##  [7,]    1    0    0    1    0    0    0    0
##  [8,]    1    0    0    1    0    0    0    0
##  [9,]    1    1    0    1    0    1    0    0
## [10,]    1    1    0    1    0    1    0    0
## [11,]    1    0    1    1    0    0    1    0
## [12,]    1    0    1    1    0    0    1    0
## [13,]    1    0    0    0    1    0    0    0
## [14,]    1    0    0    0    1    0    0    0
## [15,]    1    1    0    0    1    0    0    1
## [16,]    1    1    0    0    1    0    0    1
```

Now this matrix `m1` can be provided to the `full` argument of `DESeq`. For a likelihood ratio test of interactions, a model matrix using a reduced design such as $\sim$ `condition + group` can be given to the `reduced` argument. Wald tests can also be generated instead of the likelihood ratio test, but for user-supplied model matrices, the argument `betaPrior` must be set to `FALSE`.

4 Theory behind DESeq2

4.1 The DESeq2 model

The *DESeq2* model and all the steps taken in the software are described in detail in our pre-print [1], and we include the formula and descriptions in this section as well. The differential expression analysis in *DESeq2* uses a generalized linear model of the form:

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$
$$\mu_{ij} = s_j q_{ij}$$
$$\log_2(q_{ij}) = x_{j\cdot}\beta_i$$

where counts K_{ij} for gene i, sample j are modeled using a negative binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j[4] and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i give the log2 fold changes for gene i for each column of the model matrix X.

By default these log2 fold changes are the maximum *a posteriori* estimates after incorporating a zero-centered Normal prior – in the software referrred to as a β-prior – hence *DESeq2* provides "moderated" log2 fold change estimates. Dispersions are estimated using expected mean values from the maximum likelihood estimate of log2 fold changes, and optimizing the Cox-Reid adjusted profile likelihood, as first implemented for RNA-Seq data in *edgeR* [7, 8]. The steps performed by the `DESeq` function are documented in its manual page; briefly, they are:

1. estimation of size factors s_j by `estimateSizeFactors`
2. estimation of dispersion α_i by `estimateDispersions`
3. negative binomial GLM fitting for β_i and Wald statistics by `nbinomWaldTest`

For access to all the values calculated during these steps, see Section 3.10

4.2 Changes compared to the DESeq package

The main changes in the package *DESeq2*, compared to the (older) version *DESeq*, are as follows:

- *SummarizedExperiment* is used as the superclass for storage of input data, intermediate calculations and results.
- Maximum *a posteriori* estimation of GLM coefficients incorporating a zero-centered Normal prior with variance estimated from data (equivalent to Tikhonov/ridge regularization). This adjustment has little effect on genes with high counts, yet it helps to moderate the otherwise large variance in log2 fold change estimates for genes with low counts or highly variable counts.
- Maximum *a posteriori* estimation of dispersion replaces the `sharingMode` options `fit-only` or `maximum` of the previous version of the package. This is similar to the dispersion estimation methods of DSS [9].
- All estimation and inference is based on the generalized linear model, which includes the two condition case (previously the *exact test* was used).
- The Wald test for significance of GLM coefficients is provided as the default inference method, with the likelihood ratio test of the previous version still available.

[4] The model can be generalized to use sample- **and** gene-dependent normalization factors, see Appendix 3.11.

- It is possible to provide a matrix of sample-/gene-dependent normalization factors (Section 3.11).
- Automatic independent filtering on the mean of normalized counts (Section 4.7).
- Automatic outlier detection and handling (Section 4.4).

4.3 Methods changes since the 2014 DESeq2 paper

- For the calculation of the beta prior variance, instead of simply matching the empirical quantile to the quantile of a Normal distribution, we use the weighted quantile function of the *Hmisc* package. The weighting is described in the man page for `nbinomWaldTest`. The weights used are the inverse of the expected variance of log counts (as used in the diagonals of the matrix W in the GLM). The effect is that the estimated prior variance is robust against noisy estimates of log fold change from genes with very small counts.

4.4 Count outlier detection

DESeq2 relies on the negative binomial distribution to make estimates and perform statistical inference on differences. While the negative binomial is versatile in having a mean and dispersion parameter, extreme counts in individual samples might not fit well to the negative binomial. For this reason, we perform automatic detection of count outliers. We use Cook's distance, which is a measure of how much the fitted coefficients would change if an individual sample were removed [10]. For more on the implementation of Cook's distance see Section 3.6 and the manual page for the `results` function. Below we plot the maximum value of Cook's distance for each row over the rank of the test statistic to justify its use as a filtering criterion.

```
W <- res$stat
maxCooks <- apply(assays(dds)[["cooks"]],1,max)
idx <- !is.na(W)
plot(rank(W[idx]), maxCooks[idx], xlab="rank of Wald statistic",
     ylab="maximum Cook's distance per gene",
     ylim=c(0,5), cex=.4, col=rgb(0,0,0,.3))
m <- ncol(dds)
p <- 3
abline(h=qf(.99, p, m - p))
```

4.5 Contrasts

Contrasts can be calculated for a *DESeqDataSet* object for which the GLM coefficients have already been fit using the Wald test steps (`DESeq` with `test="Wald"` or using `nbinomWaldTest`). The vector of coefficients β is left multiplied by the contrast vector c to form the numerator of the test statistic. The denominator is formed by multiplying the covariance matrix Σ for the coefficients on either side by the contrast vector c. The square root of this product is an estimate of the standard error for the contrast. The contrast statistic is then compared to a normal distribution as are the Wald statistics for the *DESeq2* package.

$$W = \frac{c^t \beta}{\sqrt{c^t \Sigma c}}$$

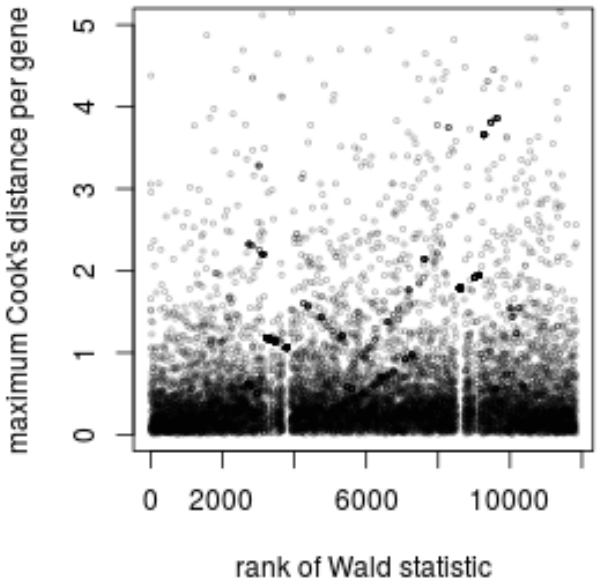

Figure 12: **Cook's distance.** Plot of the maximum Cook's distance per gene over the rank of the Wald statistics for the condition. The two regions with small Cook's distances are genes with a single count in one sample. The horizontal line is the default cutoff used for 7 samples and 3 estimated parameters.

4.6 Expanded model matrices

As mentioned in Section 3.2, *DESeq2* uses "expanded model matrices" with the log2 fold change prior, in order to produce log2 fold change estimates and test results which are independent of the choice of reference level. These model matrices differ from the standard model matrices, in that they have an indicator column (and therefore a coefficient) for each level of factors in the design formula in addition to an intercept. Expanded model matrices are not used without the log2 fold change prior or in the case of designs with 2 level factors and an interaction term.

These matrices are therefore not full rank, but a coefficient vector β_i can still be found due to the zero-centered prior on non-intercept coefficients. The prior variance for the log2 fold changes is calculated by first generating maximum likelihood estimates for a standard model matrix. The prior variance for each level of a factor is then set as the average of the mean squared maximum likelihood estimates for each level and every possible contrast, such that that this prior value will be reference-level-independent. The `contrast` argument of the `results` function is again used in order to generate comparisons of interest.

4.7 Independent filtering and multiple testing

4.7.1 Filtering criteria

The goal of independent filtering is to filter out those tests from the procedure that have no, or little chance of showing significant evidence, without even looking at their test statistic. Typically, this results in increased detection power at the same experiment-wide type I error. Here, we measure experiment-wide type I error in terms of the false discovery rate.

A good choice for a filtering criterion is one that

1. is statistically independent from the test statistic under the null hypothesis,
2. is correlated with the test statistic under the alternative, and
3. does not notably change the dependence structure –if there is any– between the tests that pass the filter, compared to the dependence structure between the tests before filtering.

The benefit from filtering relies on property 2, and we will explore it further in Section 4.7.2. Its statistical validity relies on property 1 – which is simple to formally prove for many combinations of filter criteria with test statistics– and 3, which is less easy to theoretically imply from first principles, but rarely a problem in practice. We refer to [11] for further discussion of this topic.

A simple filtering criterion readily available in the results object is the mean of normalized counts irrespective of biological condition (Figure 13). Genes with very low counts are not likely to see significant differences typically due to high dispersion. For example, we can plot the $-\log_{10} p$ values from all genes over the normalized mean counts.

```
plot(res$baseMean+1, -log10(res$pvalue),
     log="x", xlab="mean of normalized counts",
     ylab=expression(-log[10](pvalue)),
     ylim=c(0,30),
     cex=.4, col=rgb(0,0,0,.3))
```

4.7.2 Why does it work?

Consider the p value histogram in Figure 14. It shows how the filtering ameliorates the multiple testing problem – and thus the severity of a multiple testing adjustment – by removing a background set of hypotheses whose p values are distributed more or less uniformly in $[0, 1]$.

```
use <- res$baseMean > attr(res,"filterThreshold")
table(use)
```

```
## use
## FALSE  TRUE
##  6728  7742
```

```
h1 <- hist(res$pvalue[!use], breaks=0:50/50, plot=FALSE)
h2 <- hist(res$pvalue[use], breaks=0:50/50, plot=FALSE)
colori <- c(`do not pass`="khaki", `pass`="powderblue")
```

```
barplot(height = rbind(h1$counts, h2$counts), beside = FALSE,
        col = colori, space = 0, main = "", ylab="frequency")
text(x = c(0, length(h1$counts)), y = 0, label = paste(c(0,1)),
     adj = c(0.5,1.7), xpd=NA)
```

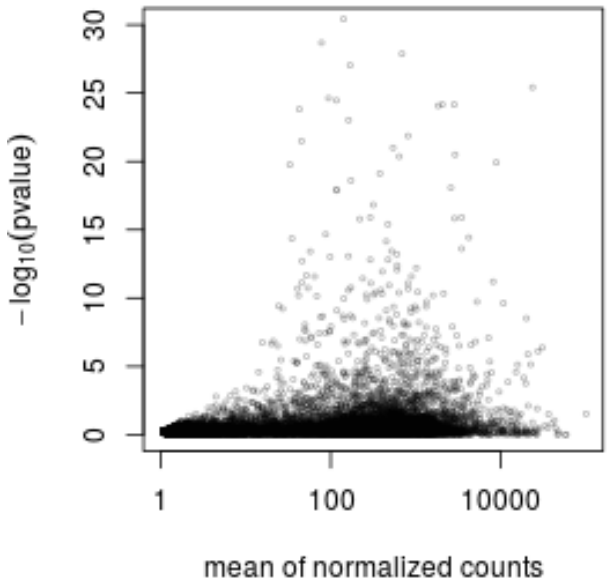

Figure 13: **Mean counts as a filter statistic.** The mean of normalized counts provides an independent statistic for filtering the tests. It is independent because the information about the variables in the design formula is not used. By filtering out genes which fall on the left side of the plot, the majority of the low p values are kept.

```
legend("topright", fill=rev(colori), legend=rev(names(colori)))
```

4.7.3 Diagnostic plots for multiple testing

The Benjamini-Hochberg multiple testing adjustment procedure [12] has a simple graphical illustration, which we produce in the following code chunk. Its result is shown in the left panel of Figure 15.

```
resFilt <- res[use & !is.na(res$pvalue),]
orderInPlot <- order(resFilt$pvalue)
showInPlot <- (resFilt$pvalue[orderInPlot] <= 0.08)
alpha <- 0.1
```

```
plot(seq(along=which(showInPlot)), resFilt$pvalue[orderInPlot][showInPlot],
     pch=".", xlab = expression(rank(p[i])), ylab=expression(p[i]))
abline(a=0, b=alpha/length(resFilt$pvalue), col="red3", lwd=2)
```

Schweder and Spjøtvoll [13] suggested a diagnostic plot of the observed p-values which permits estimation of the fraction of true null hypotheses. For a series of hypothesis tests $H_1, \ldots, H_m$ with p-values p_i, they suggested plotting

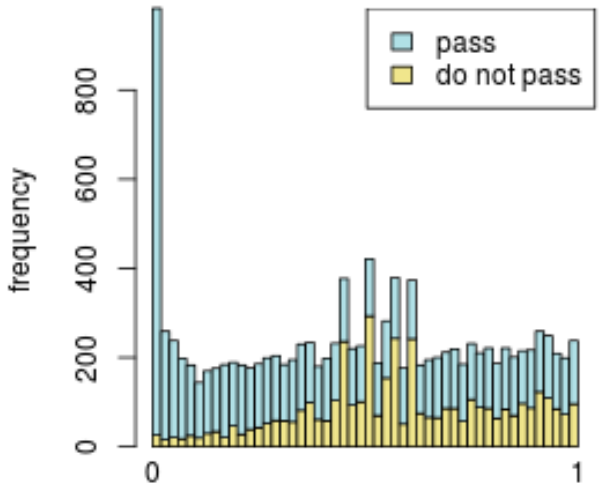

Figure 14: **Histogram of p values for all tests.** The area shaded in blue indicates the subset of those that pass the filtering, the area in khaki those that do not pass.

$$(1-p_i, N(p_i)) \text{ for } i \in 1,\ldots,m, \tag{2}$$

where $N(p)$ is the number of p-values greater than p. An application of this diagnostic plot to `resFilt$pvalue` is shown in the right panel of Figure 15. When all null hypotheses are true, the p-values are each uniformly distributed in $[0,1]$, Consequently, the cumulative distribution function of $(p_1,\ldots,p_m)$ is expected to be close to the line $F(t)=t$. By symmetry, the same applies to $(1-p_1,\ldots,1-p_m)$. When (without loss of generality) the first m_0 null hypotheses are true and the other $m-m_0$ are false, the cumulative distribution function of $(1-p_1,\ldots,1-p_{m_0})$ is again expected to be close to the line $F_0(t)=t$. The cumulative distribution function of $(1-p_{m_0+1},\ldots,1-p_m)$, on the other hand, is expected to be close to a function $F_1(t)$ which stays below F_0 but shows a steep increase towards 1 as t approaches 1. In practice, we do not know which of the null hypotheses are true, so we can only observe a mixture whose cumulative distribution function is expected to be close to

$$F(t) = \frac{m_0}{m}F_0(t) + \frac{m-m_0}{m}F_1(t). \tag{3}$$

Such a situation is shown in the right panel of Figure 15. If $F_1(t)/F_0(t)$ is small for small t, then the mixture fraction $\frac{m_0}{m}$ can be estimated by fitting a line to the left-hand portion of the plot, and then noting its height on the right. Such a fit is shown by the red line in the right panel of Figure 15.

```
plot(1-resFilt$pvalue[orderInPlot],
     (length(resFilt$pvalue)-1):0, pch=".",
```

```
     xlab=expression(1-p[i]), ylab=expression(N(p[i])))
abline(a=0, slope, col="red3", lwd=2)
```

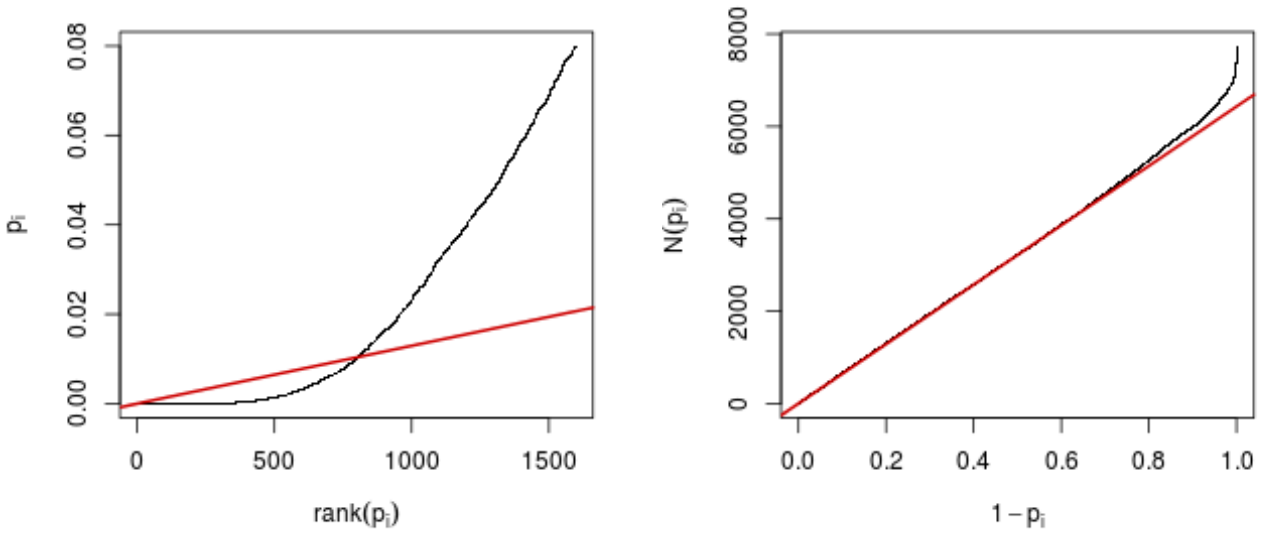

Figure 15: *Left:* illustration of the Benjamini-Hochberg multiple testing adjustment procedure [12]. The black line shows the p-values (y-axis) versus their rank (x-axis), starting with the smallest p-value from the left, then the second smallest, and so on. Only the first 1597 p-values are shown. The red line is a straight line with slope α/n, where $n = 7737$ is the number of tests, and $\alpha = 0.1$ is a target false discovery rate (FDR). FDR is controlled at the value α if the genes are selected that lie to the left of the rightmost intersection between the red and black lines: here, this results in 803 genes. *Right:* Schweder and Spjøtvoll plot, as described in the text. For both of these plots, the p-values `resFilt$pvalues` from Section 4.7.1 were used as a starting point. Analogously, one can produce these types of plots for any set of p-values, for instance those from the previous sections.

5 Frequently asked questions

5.1 How can I get support for DESeq2?

We welcome questions about our software, and want to ensure that we eliminate issues if and when they appear. We have a few requests to optimize the process:

- all questions should take place on the Bioconductor support site: https://support.bioconductor.org, which serves as a repository of questions and answers. This helps to save the developers' time in responding to similar questions. Make sure to tag your post with "deseq2". It is often very helpful in addition to describe the aim of your experiment.
- before posting, first search the Bioconductor support site mentioned above for past threads which might have answered your question.

- if you have a question about the behavior of a function, read the sections of the manual page for this function by typing a question mark and the function name, e.g. `?results`. We spend a lot of time documenting individual functions and the exact steps that the software is performing.
- include all of your R code, especially the creation of the *DESeqDataSet* and the design formula. Include complete warning or error messages, and conclude your message with the full output of `sessionInfo()`.
- if possible, include the output of `as.data.frame(colData(dds))`, so that we can have a sense of the experimental setup. If this contains confidential information, you can replace the levels of those factors using `levels()`.

5.2 Why are some p values set to NA?

See the details in Section 1.4.3.

5.3 How can I get unfiltered DESeq results?

Users can obtain unfiltered GLM results, i.e. without outlier removal or independent filtering with the following call:

```
dds <- DESeq(dds, minReplicatesForReplace=Inf)
res <- results(dds, cooksCutoff=FALSE, independentFiltering=FALSE)
```

In this case, the only p values set to `NA` are those from genes with all counts equal to zero.

5.4 How do I use the variance stabilized or rlog transformed data for differential testing?

The variance stabilizing and rlog transformations are provided for applications other than differential testing, for example clustering of samples or other machine learning applications. For differential testing we recommend the `DESeq` function applied to raw counts as outlined in Section 1.3.

5.5 Can I use DESeq2 to analyze paired samples?

Yes, you should use a multi-factor design which includes the sample information as a term in the design formula. This will account for differences between the samples while estimating the effect due to the condition. The condition of interest should go at the end of the design formula. See Section 1.5.

5.6 Can I run DESeq2 to contrast the levels of 100 groups?

DESeq2 will work with any kind of design specified using the R formula. We enourage users to consider exploratory data analysis such as principal components analysis as described in Section 2.2.3, rather than performing statistical testing of all combinations of dozens of groups.

As a speed concern with fitting very large models, note that each additional level of a factor in the design formula adds another parameter to the GLM which is fit by *DESeq2*. Users might consider first removing genes with very few reads, e.g. genes with row sum of 1, as this will speed up the fitting procedure.

5.7 Can I use DESeq2 to analyze a dataset without replicates?

If a *DESeqDataSet* is provided with an experimental design without replicates, a message is printed, that the samples are treated as replicates for estimation of dispersion. More details can be found in the manual page for `?DESeq`.

5.8 How can I include a continuous covariate in the design formula?

Continuous covariates can be included in the design formula in the same manner as factorial covariates. Continuous covariates might make sense in certain experiments, where a constant fold change might be expected for each unit of the covariate. However, in many cases, more meaningful results can be obtained by cutting continuous covariates into a factor defined over a small number of bins (e.g. 3-5). In this way, the average effect of each group is controlled for, regardless of the trend over the continuous covariates. In R, *numeric* vectors can be converted into *factors* using the function `cut`.

5.9 What are the exact steps performed by `DESeq()`?

See the manual page for `DESeq`, which links to the subfunctions which are called in order, where complete details are listed.

6 Acknowledgments

We have benefited in the development of *DESeq2* from the help and feedback of many individuals, including but not limited to: The Bionconductor Core Team, Alejandro Reyes, Andrzej Oles, Aleksandra Pekowska, Felix Klein, Vince Carey, Devon Ryan, Steve Lianoglou, Jessica Larson, Christina Chaivorapol, Pan Du, Richard Bourgon, Willem Talloen, Elin Videvall, Hanneke van Deutekom, Todd Burwell, Jesse Rowley, Igor Dolgalev, Stephen Turner, Ryan C Thompson, Tyr Wiesner-Hanks, Konrad Rudolph, David Robinson.

7 Session Info

- R version 3.2.0 (2015-04-16), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.28.0, BiocGenerics 0.14.0, DESeq2 1.8.1, GenomeInfoDb 1.4.0, GenomicRanges 1.20.3, IRanges 2.2.1, RColorBrewer 1.1-2, Rcpp 0.11.5, RcppArmadillo 0.5.000.0, S4Vectors 0.6.0, airway 0.102.0, ggplot2 1.0.1, gplots 2.16.0, knitr 1.10, pasilla 0.8.0, vsn 3.36.0
- Loaded via a namespace (and not attached): AnnotationDbi 1.30.1, BiocInstaller 1.18.1, BiocParallel 1.2.1, BiocStyle 1.6.0, DBI 0.3.1, DESeq 1.20.0, Formula 1.2-1, Hmisc 3.16-0, KernSmooth 2.23-14, MASS 7.3-40, RSQLite 1.0.0, XML 3.98-1.1, XVector 0.8.0, acepack 1.3-3.3, affy 1.46.0, affyio 1.36.0, annotate 1.46.0, bitops 1.0-6, caTools 1.17.1, cluster 2.0.1, codetools 0.2-11, colorspace 1.2-6, digest 0.6.8, evaluate 0.7, foreign 0.8-63, formatR 1.2, futile.logger 1.4.1, futile.options 1.0.0, gdata 2.13.3, genefilter 1.50.0, geneplotter 1.46.0, grid 3.2.0, gridExtra 0.9.1,

gtable 0.1.2, gtools 3.4.2, highr 0.5, labeling 0.3, lambda.r 1.1.7, lattice 0.20-31, latticeExtra 0.6-26, limma 3.24.3, locfit 1.5-9.1, magrittr 1.5, munsell 0.4.2, nnet 7.3-9, plyr 1.8.2, preprocessCore 1.30.0, proto 0.3-10, reshape2 1.4.1, rpart 4.1-9, scales 0.2.4, splines 3.2.0, stringi 0.4-1, stringr 1.0.0, survival 2.38-1, tools 3.2.0, xtable 1.7-4, zlibbioc 1.14.0

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*, 15:550, 2014. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

[2] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq – A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 2014. URL: http://dx.doi.org/10.1093/bioinformatics/btu638.

[3] A. N. Brooks, L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner, and B. R. Graveley. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, pages 193–202, 2011. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.108662.110, doi:10.1101/gr.108662.110.

[4] Robert Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83:394–405, 1988.

[5] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1):Article 3, 2003.

[6] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. URL: http://genomebiology.com/2010/11/10/R106.

[7] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49(1):1–39, 1987. URL: http://www.jstor.org/stable/2345476.

[8] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40:4288–4297, January 2012. URL: http://www.ncbi.nlm.nih.gov/pubmed/22287627, doi:10.1093/nar/gks042.

[9] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, September 2012. URL: http://dx.doi.org/10.1093/biostatistics/kxs033, doi:10.1093/biostatistics/kxs033.

[10] R. Dennis Cook. Detection of Influential Observation in Linear Regression. *Technometrics*, February 1977.

[11] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *PNAS*, 107(21):9546–9551, 2010. URL: http://www.pnas.org/content/107/21/9546.long.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[13] T. Schweder and E. Spjotvoll. Plots of P-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502, 1982. doi:10.1093/biomet/69.3.493.