Differential analysis of RNA-Seq data at the gene level using the DESeq2 package

Michael Love, Simon Anders, Wolfgang Huber

June 25, 2013

Contents

1 Introduction

In this lab, you will learn how to analyse a count table, such as arising from a summarised RNA-Seq experiment, for differentially expressed genes.

2 Input data

2.1 Experiment data

We read in a prepared *SummarizedExperiment*, which was generated from publicly available data from the article by Felix Haglund et al., "Evidence of a Functional Estrogen Receptor in Parathyroid Adenomas", J Clin Endocrin Metab, Sep 2012, http://www.ncbi.nlm.nih.gov/pubmed/23024189. Details on the generation of this object can be found in the vignette for the *parathyroidSE* package, http://bioconductor.org/packages/release/data/experiment/html/parathyroidSE.html.

The purpose of the experiment was to investigate the role of the estrogen receptor in parathyroid tumors. The investigators derived primary cultures of parathyroid adenoma cells from 4 patients. These primary cultures were treated with diarylpropionitrile (DPN), an estrogen receptor β agonist, or with 4-hydroxytamoxifen (OHT). RNA was extracted at 24 hours and 48 hours from cultures under treatment and control. The blocked design of the experiment allows for statistical analysis of the treatment effects while controlling for patient-to-patient variation.

We first load the *DESeq2* package and the data package *parathyroidSE*, which contains the example data set.

```
> library( "DESeq2" )
> library( "parathyroidSE" )
```

The `data` command loads a data object.

```
> data("parathyroidGenesSE")
```

The information in a *SummarizedExperiment* object can be accessed with accessor functions. For example, to see the actual data, i.e., here, the read counts, we use the `assay` function. (The `head` function restricts the output to the first few lines.)

```
> head( assay( parathyroidGenesSE ) )
```

```
                 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
ENSG00000000003  794 1064  444  952  518  855  414  365  278  1173   463   317
ENSG00000000005    4    1    2    3    3    1    0    1    0     0     0     0
ENSG00000000419  294  283  163  263  179  215  278  204  189   602   257   183
ENSG00000000457  156  185   93  144   75  122  228  172  116   421   182   122
ENSG00000000460  396  207  210  212  221  173  611  199  426  1390   287   417
ENSG00000000938    2    8    2    5    0    4   13   22    3    38    13    10
                [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
```

```
ENSG00000000003   986   424   305   390   587   713   957   346   433   403
ENSG00000000005     0     0     0     0     0     0     1     0     0     0
ENSG00000000419   588   275   263   281   406   568   764   287   260   250
ENSG00000000457   441   211   131   115   196   265   347   133   168   148
ENSG00000000460  1452   239   187   103   389   294   780   162    85   339
ENSG00000000938    26    13     7     3    10    18    15     7     8     7
                [,23] [,24] [,25] [,26] [,27]
ENSG00000000003   277   510   366   271   492
ENSG00000000005     0     0     0     0     0
ENSG00000000419   149   271   227   196   363
ENSG00000000457    84   183   136   118   195
ENSG00000000460    75   154   314   118   231
ENSG00000000938     5    13     8     7     8
```

In this count table, each row represents an Ensembl gene, each column a sequenced RNA library, and the values give the raw numbers of sequencing reads that were mapped to the respective gene in each library.

Question 1: For how many genes are there counts in this table?

We also have metadata on each of the samples (the "columns" of the count table):

```
> colData( parathyroidGenesSE )
```

```
DataFrame with 27 rows and 9 columns
         fileName         run experiment  patient treatment     time submission
    <BamFileList> <character>   <factor> <factor>  <factor> <factor>   <factor>
1        ########   SRR479052  SRX140503        1   Control      24h  SRA051611
2        ########   SRR479053  SRX140504        1   Control      48h  SRA051611
3        ########   SRR479054  SRX140505        1       DPN      24h  SRA051611
4        ########   SRR479055  SRX140506        1       DPN      48h  SRA051611
5        ########   SRR479056  SRX140507        1       OHT      24h  SRA051611
...           ...         ...        ...      ...       ...      ...        ...
23       ########   SRR479074  SRX140523        4       DPN      48h  SRA051611
24       ########   SRR479075  SRX140523        4       DPN      48h  SRA051611
25       ########   SRR479076  SRX140524        4       OHT      24h  SRA051611
26       ########   SRR479077  SRX140525        4       OHT      48h  SRA051611
27       ########   SRR479078  SRX140525        4       OHT      48h  SRA051611
        study    sample
     <factor>  <factor>
1   SRP012167 SRS308865
2   SRP012167 SRS308866
3   SRP012167 SRS308867
4   SRP012167 SRS308868
5   SRP012167 SRS308869
...       ...       ...
23  SRP012167 SRS308885
24  SRP012167 SRS308885
25  SRP012167 SRS308886
26  SRP012167 SRS308887
27  SRP012167 SRS308887
```

Question 2: What are the metadata for the genes (the "rows" of the count table)?

2.2 Collapsing technical replicates

There are a number of samples which were sequenced in multiple runs. For example, sample SRS308873 was sequenced twice. To see, we list the respective columns of the colData. (The use of as.data.frame forces R to show us the full list, not just the beginning and the end as before.)

```
> as.data.frame( colData(parathyroidGenesSE)[,c("sample","patient","treatment","time")] )
```

```
       sample patient treatment time
1   SRS308865       1   Control  24h
2   SRS308866       1   Control  48h
3   SRS308867       1       DPN  24h
4   SRS308868       1       DPN  48h
5   SRS308869       1       OHT  24h
6   SRS308870       1       OHT  48h
7   SRS308871       2   Control  24h
8   SRS308872       2   Control  48h
9   SRS308873       2       DPN  24h
10  SRS308873       2       DPN  24h
11  SRS308874       2       DPN  48h
12  SRS308875       2       OHT  24h
13  SRS308875       2       OHT  24h
14  SRS308876       2       OHT  48h
15  SRS308877       3   Control  24h
16  SRS308878       3   Control  48h
17  SRS308879       3       DPN  24h
18  SRS308880       3       DPN  48h
19  SRS308881       3       OHT  24h
20  SRS308882       3       OHT  48h
21  SRS308883       4   Control  48h
22  SRS308884       4       DPN  24h
23  SRS308885       4       DPN  48h
24  SRS308885       4       DPN  48h
25  SRS308886       4       OHT  24h
26  SRS308887       4       OHT  48h
27  SRS308887       4       OHT  48h
```

We recommend to first add together technical replicates (i.e., libraries derived from the same samples), such that we have one column per sample.

As is often the case, this preparatory step looks more complicated than the subsequent actual analysis. In fact, the following operations are not specific to *DESeq2*, but are specific preparations needed for this data set. To understand the general ideas of *DESeq2*, you could now skip to Section 3. What you will learn in the rest of this section is an example of a typical preparatory data manipulation task done with elementary R functions. Details on these can be found in general textbooks on R; also consider reading the help pages of the functions used.

We first use the function split to see which columns need to be collapsed.

```
> allColSamples <- colData(parathyroidGenesSE)$sample
> sp <- split( seq(along=allColSamples), colData(parathyroidGenesSE)$sample )
```

Using `sapply`, we loop over the elements of `sp`, which correspond to the distinct samples, construct subtables of the count table (i.e., `assay(parathyroidGenesSE)`) corresponding only to the current sample considered, and add up across rows if there is more than one column. The result of the `sapply` call is a new table, in which each column now corresponds to a different sample.

```
> countdata <- sapply(sp, function(columns)
+    rowSums( assay(parathyroidGenesSE)[,columns,drop=FALSE] ) )
> head(countdata)
```

```
                 SRS308865 SRS308866 SRS308867 SRS308868 SRS308869 SRS308870
ENSG00000000003       794      1064       444       952       518       855
ENSG00000000005         4         1         2         3         3         1
ENSG00000000419       294       283       163       263       179       215
ENSG00000000457       156       185        93       144        75       122
ENSG00000000460       396       207       210       212       221       173
ENSG00000000938         2         8         2         5         0         4
                 SRS308871 SRS308872 SRS308873 SRS308874 SRS308875 SRS308876
ENSG00000000003       414       365      1451       463      1303       424
ENSG00000000005         0         1         0         0         0         0
ENSG00000000419       278       204       791       257       771       275
ENSG00000000457       228       172       537       182       563       211
ENSG00000000460       611       199      1816       287      1869       239
ENSG00000000938        13        22        41        13        36        13
                 SRS308877 SRS308878 SRS308879 SRS308880 SRS308881 SRS308882
ENSG00000000003       305       390       587       713       957       346
ENSG00000000005         0         0         0         0         1         0
ENSG00000000419       263       281       406       568       764       287
ENSG00000000457       131       115       196       265       347       133
ENSG00000000460       187       103       389       294       780       162
ENSG00000000938         7         3        10        18        15         7
                 SRS308883 SRS308884 SRS308885 SRS308886 SRS308887
ENSG00000000003       433       403       787       366       763
ENSG00000000005         0         0         0         0         0
ENSG00000000419       260       250       420       227       559
ENSG00000000457       168       148       267       136       313
ENSG00000000460        85       339       229       314       349
ENSG00000000938         8         7        18         8        15
```

Novice users might find the preceding two code chunks difficult. Of course, there is a much easier way to add up the columns, namely by explicitly specifying the indices of the columns we want to use as is and the columns we want to add up, and using `cbind` to bind all the columns to a matrix:

```
> a <- assay(parathyroidGenesSE)
> countdata2 <- cbind( a[,1:8], a[,9]+a[,10], a[,11],
+    a[,12]+a[,13], a[,14:22], a[,23]+a[,24], a[,25], a[,26]+a[,27] )
> all( countdata == countdata2 )
```

```
[1] TRUE
```

While this is simpler to understand, it is more error-prone. Mistakes can easily happen when determining the column indices, and it is tedious to update the code if the input data changes, for

instance, if at a later time you would like to add more replicates to your data set. Hence, if you are a beginner in R and want to improve your R skills, try to understand how the `split` and the `sapply` calls above work, because only learning to master such expressions will give you the skills to make full use of R.

Having reduced our count data table to only one column per sample, we next need to subset the column metadata accordingly, as we now have less columns. We also now use the sample names as names for the column data rows:

```
> coldata <- colData(parathyroidGenesSE)[sapply(sp, `[`, 1),]
> rownames(coldata) <- coldata$sample
> coldata
```

```
DataFrame with 23 rows and 9 columns
             fileName         run experiment  patient treatment     time
          <BamFileList> <character>   <factor> <factor>  <factor> <factor>
SRS308865      ########   SRR479052  SRX140503        1   Control      24h
SRS308866      ########   SRR479053  SRX140504        1   Control      48h
SRS308867      ########   SRR479054  SRX140505        1       DPN      24h
SRS308868      ########   SRR479055  SRX140506        1       DPN      48h
SRS308869      ########   SRR479056  SRX140507        1       OHT      24h
...                 ...         ...        ...      ...       ...      ...
SRS308883      ########   SRR479072  SRX140521        4   Control      48h
SRS308884      ########   SRR479073  SRX140522        4       DPN      24h
SRS308885      ########   SRR479074  SRX140523        4       DPN      48h
SRS308886      ########   SRR479076  SRX140524        4       OHT      24h
SRS308887      ########   SRR479077  SRX140525        4       OHT      48h
          submission     study    sample
            <factor>  <factor>  <factor>
SRS308865  SRA051611 SRP012167 SRS308865
SRS308866  SRA051611 SRP012167 SRS308866
SRS308867  SRA051611 SRP012167 SRS308867
SRS308868  SRA051611 SRP012167 SRS308868
SRS308869  SRA051611 SRP012167 SRS308869
...              ...       ...       ...
SRS308883  SRA051611 SRP012167 SRS308883
SRS308884  SRA051611 SRP012167 SRS308884
SRS308885  SRA051611 SRP012167 SRS308885
SRS308886  SRA051611 SRP012167 SRS308886
SRS308887  SRA051611 SRP012167 SRS308887
```

Question 3: What do the quotation marks in the expression ‘`[`‘ do? What happens if you omit them?

To unclutter the output in the subsequent steps, we only keep the column data columns that we actually need for our analysis.

```
> coldata <- coldata[ , c( "patient", "treatment", "time" ) ]
> head( coldata )
```

```
DataFrame with 6 rows and 3 columns
           patient treatment     time
```

```
          <factor>  <factor> <factor>
SRS308865        1   Control      24h
SRS308866        1   Control      48h
SRS308867        1       DPN      24h
SRS308868        1       DPN      48h
SRS308869        1       OHT      24h
SRS308870        1       OHT      48h
```

Our *SummarizedExperiment* object also contains metadata on the rows, which we can simply keep unchanged:

```
> rowdata <- rowData(parathyroidGenesSE)
> rowdata
```

```
GRangesList of length 60620:
$ENSG00000000003
GRanges with 17 ranges and 2 metadata columns:
       seqnames               ranges strand   |   exon_id       exon_name
          <Rle>            <IRanges>  <Rle>   | <integer>     <character>
   [1]        X [99883667, 99884983]      -   |    653684 ENSE00001459322
   [2]        X [99885756, 99885863]      -   |    653685 ENSE00000868868
   [3]        X [99887482, 99887565]      -   |    653686 ENSE00000401072
   [4]        X [99887538, 99887565]      -   |    653687 ENSE00001849132
   [5]        X [99888402, 99888536]      -   |    653688 ENSE00002890912
   ...      ...                  ...    ... ...       ...             ...
  [13]        X [99890555, 99890743]      -   |    653696 ENSE00002799002
  [14]        X [99891188, 99891686]      -   |    653697 ENSE00001886883
  [15]        X [99891605, 99891803]      -   |    653698 ENSE00001855382
  [16]        X [99891790, 99892101]      -   |    653699 ENSE00001863395
  [17]        X [99894942, 99894988]      -   |    653700 ENSE00001828996

...
<60619 more elements>
---
seqlengths:
                1                2 ...            LRG_98            LRG_99
        249250621        243199373 ...             18750             13294
```

We now have all the ingredients to prepare our data object in a form that is suitable for analysis, namely:

- `countdata`: a table with the read counts, with technical replicates summed up,
- `coldata`: a table with metadata on the count table's columns, i.e., on the samples,
- `rowdata`: a table with metadata on the count table's rows, i.e., on the genes, and
- a design formula, which tells which factors in the column metadata table specify the experimental design and how these factors should be used in the analysis. We specify $\sim$ `patient + treatment`, which means that we want to test for the effect of treatment (the last factor), controlling for the effect of patient (the first factor). You can use R's formula notation to express any experimental design that can be described within an ANOVA-like framework.

To now construct the data object from the matrix of counts and the metadata table, we use:

```
> ddsFull <- DESeqDataSetFromMatrix(
+   countData = countdata,
+   colData = coldata,
+   design = ~ patient + treatment,
+   rowData = rowdata)
> ddsFull
```

```
class: DESeqDataSet
dim: 60620 23
exptData(0):
assays(1): counts
rownames(60620): ENSG00000000003 ENSG00000000005 ... LRG_98 LRG_99
rowData metadata column names(0):
colnames(23): SRS308865 SRS308866 ... SRS308886 SRS308887
colData names(3): patient treatment time
```

3 Running the DESeq2 pipeline

Here we will analyze a subset of the samples, namely those taken after 48 hours, with either control or DPN treatment, taking into account the multifactor design.

3.1 Preparing the data object for the analysis of interest

First we subset the relevant columns from the full dataset:

```
> dds <- ddsFull[ , colData(ddsFull)$treatment %in% c("Control","DPN") &
+                   colData(ddsFull)$time == "48h" ]
```

Sometimes it is necessary to "refactor" the factors, in case that levels have been dropped. (Here, for example, the treatment factor still contains the level "OHT", but no sample to this level.)

```
> dds$patient <- factor(dds$patient)
> dds$treatment <- factor(dds$treatment)
```

It will be convenient to make sure that `Control` is the *first* level in the treatment factor, so that the $\log_2$ fold changes are calculated as treatment over control. The function `relevel` achieves this:

```
> dds$treatment <- relevel( dds$treatment, "Control" )
```

A quick check whether we now have the right samples:

```
> colData(dds)
```

```
DataFrame with 8 rows and 3 columns
           patient treatment     time
          <factor>  <factor> <factor>
SRS308866        1   Control      48h
SRS308868        1       DPN      48h
SRS308872        2   Control      48h
SRS308874        2       DPN      48h
SRS308878        3   Control      48h
SRS308880        3       DPN      48h
SRS308883        4   Control      48h
SRS308885        4       DPN      48h
```

3.2 Running the pipeline

With the data object prepared, the *DESeq2* analysis can now be run with a single call to the function `DESeq`:

```
> dds <- DESeq(dds)
```

3.3 Inspecting the results table

The results for the last variable in the design formula, in our case the `treatment` variable, can be extracted using the `results` function.

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

```
DataFrame with 60620 rows and 5 columns
                 baseMean log2FoldChange     lfcSE    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000000003   622.683        -0.0216    0.0607   0.72174    0.9756
ENSG00000000005     0.678         0.0208    0.2043   0.91886    0.9896
ENSG00000000419   300.193        -0.0165    0.0764   0.82934    0.9800
ENSG00000000457   183.605        -0.0972    0.1085   0.37048    0.9340
ENSG00000000460   200.703         0.3517    0.1107   0.00149    0.0905
...                   ...            ...       ...       ...       ...
LRG_94                  0             NA        NA        NA        NA
LRG_96                  0             NA        NA        NA        NA
LRG_97                  0             NA        NA        NA        NA
LRG_98                  0             NA        NA        NA        NA
LRG_99                  0             NA        NA        NA        NA
```

As `res` is a `DataFrame` object, it carries metadata with information on the meaning of the columns:

```
> mcols(res)
```

```
DataFrame with 5 rows and 2 columns
          type                                        description
   <character>                                        <character>
1 intermediate                      the base mean over all rows
2      results log2 fold change (MAP): treatment DPN vs Control
3      results         standard error: treatment DPN vs Control
4      results              Wald test: treatment DPN vs Control
5      results     Wald test, BH adj.: treatment DPN vs Control
```

The first column, `baseMean`, is a just the average of the normalized count values, taken over all samples. The remaining four columns refer to a specific *contrast*, namely the comparison of the levels *DPN* versus *Control* of the factor variable *treatment*. See the help page for `results` (by typing `?results`) for information on how to obtain other contrasts.

The column `log2FoldChange` is the effect size estimate. It tells us how much the gene's expression seems to have changed due to treatment with DPN in comparison to control. This value is reported on a logarithmic scale to base 2: for example, a $\log_2$ fold change of 1.5 means that the gene's expression is increased by a factor of $2^{1.5} \approx 2.82$.

Of course, this estimate has an uncertainty associated with it, which is available in the column `lfcSE`, the standard error estimate for the log2 fold change estimate. We can also express the uncertainty of a particular effect size estimate as the result of a statistical test. The purpose of a test for differential expression is to test whether the data provides sufficient evidence to conclude that this value is really different from zero (and that the sign is correct). DESeq2 performs for each gene a *hypothesis test* to see whether evidence is sufficient to decide against the *null hypothesis* that there is no effect of the treatment on the gene and that the observed difference between treatment and control was merely caused by experimental variability (i. e., the type of variability that you can just as well expect between different samples in the same treatment group). As usual in statistics, the result of this test is reported as a *p value*, and it is found in the column `pvalue`. (Remember that a p value indicates the probability that a fold change as strong as the observed one, or even stronger, would be seen under the situation described by the null hypothesis.)

Finally, we note that a subset of the p values in `res` are `NA` ("not available"). This is `DESeq`'s way of reporting that all counts for this gene were zero, and hence not test was applied.

Question 4: How could you check to see if the `baseMean` is the mean of raw counts or the mean of normalized counts?

3.4 Multiple testing

Novices in high-throughput biology often assume that thresholding these p values at 0.05, as is often done in other settings, would be appropriate – but it is not. We briefly explain why:

There are 1957 genes with a p value below 0.05 among the 31523 genes, for which the test succeeded in reporting a p value:

```
> sum( res$pvalue < 0.05, na.rm=TRUE )
```

```
[1] 1957
```

```
> sum( is.na(res$pvalue) )
```

```
[1] 31523
```

Now, assume for a moment that the null hypothesis is true for all genes, i.e., no gene is affected by the treatment with DPN. Then, by the definition of *p value*, we expect up to 5% of the genes to have a p value below 0.05. This amounts to 1455 genes. If we just considered the list of genes with a p value below 0.05 as differentially expressed, this list should therefore be expected to contain up to $1455/1957 = 74\%$ false positives!

DESeq2 uses the so-called Benjamini-Hochberg (BH) adjustment; in brief, this method calculates for each gene an *adjusted p value* which answers the following question: if one called significant all genes with a p value less than or equal to this gene's p value threshold, what would be the fraction of false positives (the *false discovery rate*, FDR) among them (in the sense of the calculation outlined above)? These values, called the BH-adjusted p values, are given in the column `padj` of the `results` object.

Hence, if we consider a fraction of 10% false positives acceptable, we can consider all genes with an *adjusted* p value below 10%=0.1 as significant. How many such genes are there?

```
> sum( res$padj < 0.1, na.rm=TRUE )
```

```
[1] 505
```

We subset the results table to these genes and then sort it by the log2-fold-change estimate to get the significant genes with the strongest down-regulation

```
> resSig <- res[ which(res$padj < 0.1 ), ]
> head( resSig[ order( resSig$log2FoldChange ), ] )
```

```
DataFrame with 6 rows and 5 columns
                 baseMean log2FoldChange     lfcSE    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000180616      4.56         -1.097     0.330  8.83e-04  6.17e-02
ENSG00000183798      4.56         -1.035     0.324  1.41e-03  8.70e-02
ENSG00000230069      6.55         -1.026     0.321  1.39e-03  8.60e-02
ENSG00000170122     18.68         -1.005     0.276  2.76e-04  2.60e-02
ENSG00000146250     29.39         -0.956     0.211  5.83e-06  1.26e-03
ENSG00000163631    268.91         -0.940     0.106  9.45e-19  2.75e-15
```

and with the strongest upregulation

```
> tail( resSig[ order( resSig$log2FoldChange ), ] )
```

```
DataFrame with 6 rows and 5 columns
                 baseMean log2FoldChange     lfcSE    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000156414     137.0          0.733     0.139  1.38e-07  5.13e-05
ENSG00000236279      68.1          0.752     0.174  1.48e-05  2.63e-03
ENSG00000103257     169.7          0.771     0.146  1.35e-07  5.08e-05
ENSG00000165084      52.3          0.772     0.197  9.03e-05  1.14e-02
ENSG00000169174      30.3          0.893     0.253  4.26e-04  3.57e-02
ENSG00000128165      42.7          0.900     0.228  7.65e-05  9.84e-03
```

Question 5: What is the proportion of down- and up-regulation among the genes with adjusted p value less than 0.1?

3.5 Diagnostic plots

A so-called MA plot provides a useful overview for an experiment with a two-group comparison:

```
> plotMA(dds, ylim = c( -1.5, 1.5 ) )
```

The plot (Fig. 1) represents each gene with a dot. The x axis is the average expression over all samples, the y axis the $\log_2$ fold change between treatment and control. Genes with an adjusted p value below a threshold (here 0.1, the default) are shown in red.

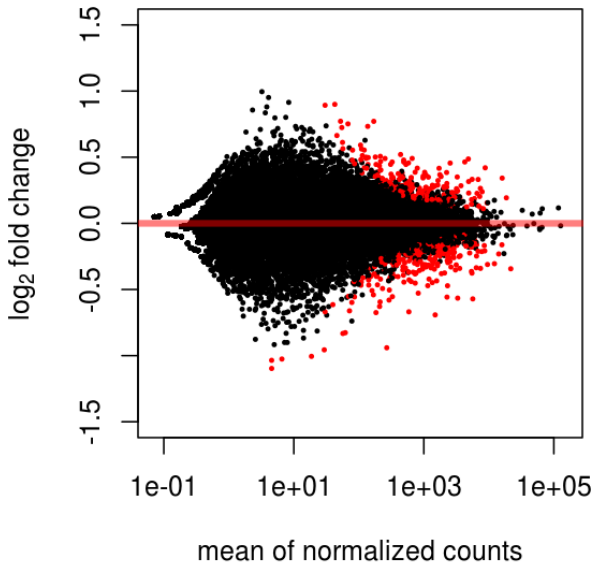

Figure 1: The MA-plot shows the $\log_2$ fold changes from the treatment over the mean of normalized counts, i.e. the average of counts normalized by size factor. The *DESeq2* package incorporates a prior on $\log_2$ fold changes, resulting in moderated estimates from genes with low counts and highly variable counts, as can be seen by the narrowing of spread of points on the left side of the plot.

This plot demonstrates that only genes with an average normalized count above 10 contain sufficient information to yield a significant call, and only above about 300 counts can smaller fold-changes become significant.

Also note *DESeq2*'s shrinkage estimation of log fold changes (LFCs): When count values are too low to allow an accurate estimate of the LFC, the value is "shrunken" towards zero to avoid that these values, which otherwise would frequently be unrealistically large, dominate the top-ranked log fold changes.

Whether a gene is called significant depends not only on its LFC but also on its within-group variability, which *DESeq2* quantifies as the *dispersion*. For strongly expressed genes, the dispersion can be understood as a squared coefficient of variation: a dispersion value of 0.01 means that the gene's expression tends to differ by typically $\sqrt{0.01} = 10\%$ between samples of the same treatment group. For weak genes, the Poisson noise is an additional source of noise, which is added to the dispersion.

The function `plotDispEsts` visualizes *DESeq2*'s dispersion estimates:

```
> plotDispEsts( dds )
```

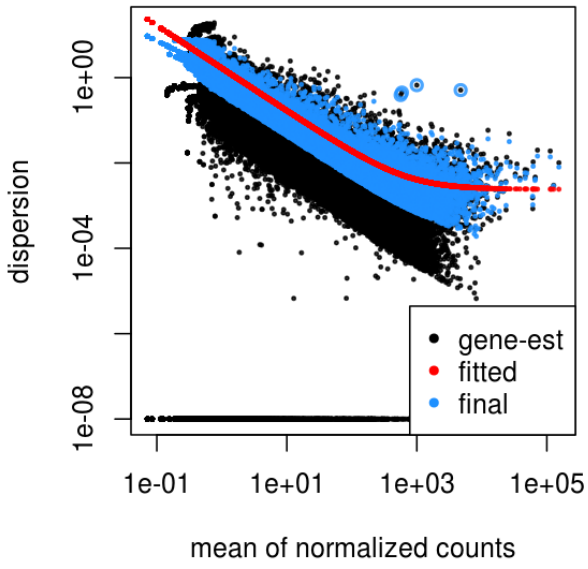

Figure 2: Plot of dispersion estimates. See text for details

The black dots are the dispersion estimates for each gene as obtained by considering the information from each gene separately. Unless one has many samples, these values fluctuate strongly around their true values. Therefore, we fit the red trend line, which shows the dispersions' dependence on the mean, and then shrink each gene's estimate towards the red line to obtain the final estimates (blue circles) that are then used in the hypothesis test.

Question 6: How could you change the MA-plot so as to color those genes with adjusted p-value less than 0.5 instead of 0.1?

Another useful diagnostic plot is the histogram of the p values (Fig. 3).

```
> hist( res$pvalue, breaks=100     )
```

Question 7: Revisit the discussion about p values and multiple testing in the previous section. Which part of the histogram is caused by genes that are called significant? And which part is caused by those that are truly significant? Why are there "spikes" at intermediate values?

4 Independent filtering

The MA plot (Figure 1) highlights an important property of RNA-Seq data. For weakly expressed genes, we have no chance of seeing differential expression, because the low read counts suffer from so

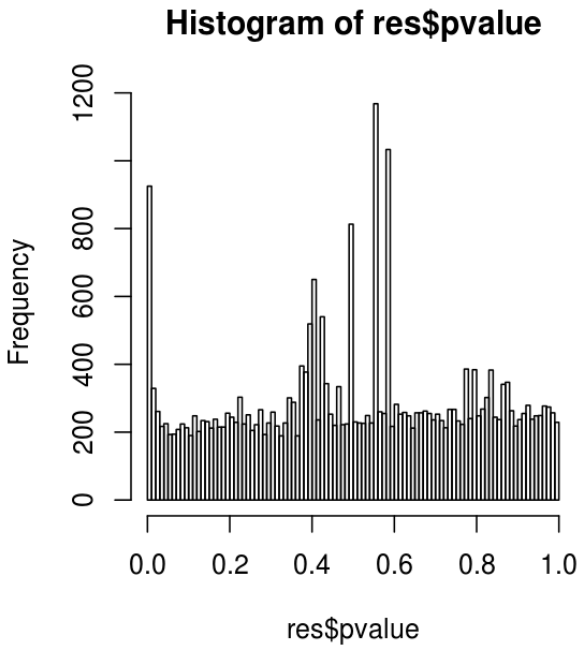

Figure 3: Histogram of the p values returned by the test for differential expression.

high Poisson noise that any biological effect is drowned in the uncertainties from the read counting. The MA plot suggests that for genes with less than one or two counts per sample, averaged over all samples, there is no real inferential power. We loose little if we filter out these genes:

```
> filterThreshold <- 2.0
> keep <- rowMeans( counts( dds, normalized=TRUE ) ) > filterThreshold
> table( keep )
```

```
keep
FALSE  TRUE
41503 19117
```

Note that none of the genes below the threshold had a significant adjusted p value

```
> min( res$padj[!keep], na.rm=TRUE )
```

```
[1] 0.421
```

At first sight, there may seem to be little benefit in filtering out these genes. After all, the test found them to be non-significant anyway. However, these genes have an influence on the multiple testing adjustment, whose performance improves if such genes are removed. Compare:

```
> table( p.adjust( res$pvalue, method="BH" ) < .1 )
```

```
FALSE  TRUE
28592   505
```

```
> table( p.adjust( res$pvalue[keep], method="BH" ) < .1 )
```

```
FALSE  TRUE
17927   631
```

By removing the weakly-expressed genes from the input to the FDR procedure, we have found more genes to be significant among those which we kept, and so improved the power of our test. This approach is known as *independent filtering*.

The term *independent* highlights an important caveat. Such filtering is permissible only if the filter criterion is independent of the actual test statistic [1]. Otherwise, the filtering would invalidate the test and consequently the assumptions of the BH procedure. This is why we filtered on the average over *all* samples: this filter is blind to the assignment of samples to the treatment and control group and hence independent.

Question 8: Redo the histogram as in Figure 3, now only using the genes that passed the filtering. What happened to the spikes at intermediate values?

4.1 Adding gene names

Our result table only uses Ensembl gene IDs, but gene names may be more informative. Bioconductor's annotation packages help with mapping various ID schemes to each other.

We load the annotation package *org.Hs.eg.db*:

```
> library( "org.Hs.eg.db" )
```

This is the organism annotation package ("org") for *Homo sapiens* ("Hs"), organized as an *AnnotationDbi* package ("db"), using Entrez Gene IDs ("eg") as primary key.

To get a list of all available key types, use

```
> cols(org.Hs.eg.db)
```

```
 [1] "ENTREZID"     "PFAM"         "IPI"          "PROSITE"      "ACCNUM"
 [6] "ALIAS"        "CHR"          "CHRLOC"       "CHRLOCEND"    "ENZYME"
[11] "MAP"          "PATH"         "PMID"         "REFSEQ"       "SYMBOL"
[16] "UNIGENE"      "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS" "GENENAME"
[21] "UNIPROT"      "GO"           "EVIDENCE"     "ONTOLOGY"     "GOALL"
[26] "EVIDENCEALL"  "ONTOLOGYALL"  "OMIM"         "UCSCKG"
```

Converting IDs with the native functions from the *AnnotationDbi* package is currently a bit cumbersome, so we provide the following convenience function (without explaining how exactly it works):

```
> convertIDs <- function( ids, fromKey, toKey, db, ifMultiple=c( "putNA", "useFirst" ) ) {
+    stopifnot( inherits( db, "AnnotationDb" ) )
+    ifMultiple <- match.arg( ifMultiple )
+    suppressWarnings( selRes <- AnnotationDbi::select(
+       db, keys=ids, keytype=fromKey, cols=c(fromKey,toKey) ) )
+    if( ifMultiple == "putNA" ) {
```

```
+       duplicatedIds <- selRes[ duplicated( selRes[,1] ), 1 ]
+       selRes <- selRes[ ! selRes[,1] %in% duplicatedIds, ] }
+    return( selRes[ match( ids, selRes[,1] ), 2 ] )
+ }
```

This function takes a list of IDs as first argument and their key type as the second argument. The third argument is the key type we want to convert to, the fourth is the *AnnotationDb* object to use. Finally, the last argument specifies what to do if one source ID maps to several target IDs: should the function return an NA or simply the first of the multiple IDs?

To convert the Ensembl IDs in the rownames of `res` to gene symbols and add them as a new column, we use:

```
> res$symbol <- convertIDs( row.names(res), "ENSEMBL", "SYMBOL", org.Hs.eg.db )
> res
```

```
DataFrame with 60620 rows and 6 columns
                 baseMean log2FoldChange     lfcSE    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000000003   622.683        -0.0216    0.0607   0.72174    0.9756
ENSG00000000005     0.678         0.0208    0.2043   0.91886    0.9896
ENSG00000000419   300.193        -0.0165    0.0764   0.82934    0.9800
ENSG00000000457   183.605        -0.0972    0.1085   0.37048    0.9340
ENSG00000000460   200.703         0.3517    0.1107   0.00149    0.0905
...                   ...            ...       ...       ...       ...
LRG_94                  0             NA        NA        NA        NA
LRG_96                  0             NA        NA        NA        NA
LRG_97                  0             NA        NA        NA        NA
LRG_98                  0             NA        NA        NA        NA
LRG_99                  0             NA        NA        NA        NA
                    symbol
                <character>
ENSG00000000003      TSPAN6
ENSG00000000005        TNMD
ENSG00000000419        DPM1
ENSG00000000457       SCYL3
ENSG00000000460    C1orf112
...                     ...
LRG_94                   NA
LRG_96                   NA
LRG_97                   NA
LRG_98                   NA
LRG_99                   NA
```

Finally, we note that you can easily save the results table in a CSV file, which you can then load with a spreadsheet program such as Excel:

```
> write.csv( as.data.frame(res), file="results.csv" )
```

5 Downstream analyses

A list of gene names is no final result. We demonstrate two possible further analysis steps.

5.1 Gene set enrichment analysis

Do the genes with a strong up- or down-regulation have something in common? We perform next a gene-set enrichment analysis (GSEA) to examine this question.

We use the gene sets in the Reactome database

```
> library( "reactome.db" )
```

This database works with Entrez IDs, so we add a column with such IDs, using our `convertIDs` function:

```
> res$entrez <- convertIDs( row.names(res), "ENSEMBL", "ENTREZID", org.Hs.eg.db )
```

Next, we subset the results table, `res`, to only those genes for which the Reactome database has data (i.e, whose Entrez ID we find in the respective key column of `reactome.db`) and for which the test gave a p value that was not NA.

```
> res2 <- res[ res$entrez %in% keys( reactome.db, "ENTREZID" ) &
+    !is.na( res$pvalue) , ]
> head(res2)
```

```
DataFrame with 6 rows and 7 columns
                 baseMean log2FoldChange     lfcSE    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000000419     300.2       -0.01646    0.0764     0.829     0.980
ENSG00000000938      12.1       -0.00527    0.3209     0.987     0.997
ENSG00000000971      19.8       -0.38114    0.2683     0.155     0.934
ENSG00000001084     323.3        0.04419    0.0935     0.637     0.953
ENSG00000001167     412.6       -0.05684    0.1018     0.576     0.934
ENSG00000001626      10.2        0.22318    0.3119     0.474     0.934
                    symbol      entrez
                <character> <character>
ENSG00000000419        DPM1        8813
ENSG00000000938         FGR        2268
ENSG00000000971         CFH        3075
ENSG00000001084        GCLC        2729
ENSG00000001167        NFYA        4800
ENSG00000001626        CFTR        1080
```

Using `select`, a function from *AnnotationDbi* for querying database objects, we get a table with the mapping from Entrez IDs to Reactome Path IDs

```
> reactomeTable <- AnnotationDbi::select( reactome.db, keys=res2$entrez,
+    keytype="ENTREZID", cols=c("ENTREZID","REACTOMEID") )
> head(reactomeTable)
```

```
  ENTREZID REACTOMEID
1     8813     162699
2     8813     163125
3     8813     392499
4     8813     446193
5     8813     446203
6     8813     446219
```

The next code chunk transforms this table into an *incidence matrix*. This is a boolean matrix with one row for each Reactome Path and one column for each gene in `res2`, which tells us which genes are members of which Reactome Paths. (If you want to understand how this chunk exactly works, read up about the `tapply` function.)

```
> incm <- do.call( rbind, with(reactomeTable, tapply(
+    ENTREZID, factor(REACTOMEID), function(x) res2$entrez %in% x ) ))
> colnames(incm) <- res2$entrez
> str(incm)
```

```
 logi [1:1439, 1:5296] FALSE FALSE FALSE FALSE FALSE FALSE ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:1439] "1059683" "109581" "109582" "109606" ...
  ..$ : chr [1:5296] "8813" "2268" "3075" "2729" ...
```

We remove all rows corresponding to Reactome Paths with less than 5 assigned genes.

```
> incm <- incm[ rowSums(incm) >= 5, ]
```

To test whether the genes in a Reactome Path behave in a special way in our experiment, we perform t-tests to see whether the average of the genes' $\log_2$ fold change values are different from zero. If so, we can say that our treatment tends to upregulate (or downregulate) the genes in the category. To facilitate the computations, we define a little helper function:

```
> testCategory <- function( reactomeID ) {
+   isMember <- incm[ reactomeID, ]
+   data.frame(
+      reactomeID = reactomeID,
+      numGenes = sum( isMember ),
+      avgLFC = mean( res2$log2FoldChange[isMember] ),
+      strength = sum( res2$log2FoldChange[isMember] ) / sqrt(sum(isMember)),
+      pvalue = t.test( res2$log2FoldChange[ isMember ] )$p.value,
+      reactomeName = reactomePATHID2NAME[[reactomeID]] ) }
```

The function can be called with a Reactome Path ID:

```
> testCategory("109581")
```

```
  reactomeID numGenes   avgLFC strength pvalue            reactomeName
1     109581      146 -0.00887   -0.107  0.307 Homo sapiens: Apoptosis
```

As you can see the function not only performs the t test and returns the p value but also lists other useful information such as the number of genes in the category, the average log fold change, a "strength" measure (see below) and the name with which Reactome describes the Path.

We call the function for all Paths in our incidence matrix and collect the results in a data frame:

```
> reactomeResult <- do.call( rbind, lapply( rownames(incm), testCategory ) )
```

As we performed many tests, we should again use a multiple testing adjustment.

```
> reactomeResult$padjust <- p.adjust( reactomeResult$pvalue, "BH" )
```

This is a list of Reactome Paths which are significantly differentially expressed in our comparison of DPN treatment with control, sorted according to sign and strength of the signal:

```
> reactomeResultSignif <- reactomeResult[ reactomeResult$padjust < 0.05, ]
> reactomeResultSignif[ order(reactomeResultSignif$strength), ]
```

```
     reactomeID numGenes  avgLFC strength   pvalue
547     2032785       10 -0.1216   -0.385 3.59e-04
1052      74159      136 -0.0318   -0.370 1.76e-04
1030      73857       95 -0.0340   -0.332 3.62e-04
435      189445       14  0.0949    0.355 2.75e-04
463      191273       22  0.0997    0.468 1.77e-04
534     2024096       21  0.1024    0.469 3.10e-04
759      392499      484  0.0218    0.481 4.37e-04
500      196849       52  0.0763    0.550 2.54e-04
501      196854       52  0.0763    0.550 2.54e-04
720      381070       47  0.0837    0.574 1.80e-04
128     1430728     1264  0.0183    0.649 6.03e-05
                                                          reactomeName padjust
547    Homo sapiens: YAP1- and WWTR1 (TAZ)-stimulated gene expression  0.0426
1052                                      Homo sapiens: Transcription  0.0426
1030                   Homo sapiens: RNA Polymerase II Transcription  0.0426
435                           Homo sapiens: Metabolism of porphyrins  0.0426
463                           Homo sapiens: Cholesterol biosynthesis  0.0426
534                                 Homo sapiens: HS-GAG degradation  0.0426
759                               Homo sapiens: Metabolism of proteins  0.0468
500  Homo sapiens: Metabolism of water-soluble vitamins and cofactors  0.0426
501                Homo sapiens: Metabolism of vitamins and cofactors  0.0426
720              Homo sapiens: Activation of Chaperones by IRE1alpha  0.0426
128                                          Homo sapiens: Metabolism  0.0426
```

Note that such lists need to be interpreted with care, and a grain of salt. Which of these categories make sense, given the biology of the experiment?

5.2 Nearest peak to a differentially expressed gene

The RNA-Seq experiment analyzed above provides a list of genes which have responded to a selective estrogen-receptor-beta agonist. We can investigate whether we find estrogen receptor binding sites in the vicinity of the gene with the highest fold induction. In order to match differentially expressed genes to other experiment data, we will use annotated binding sites of estrogen receptor alpha from the ENCODE project. It is not necessarily the case that these annotated binding sites are actually functional in the cell lines of the RNA-Seq experiment or biologically relevant as the alpha and beta subtypes are distinct proteins transcribed from different genes; here we only use these binding site data for demonstration purposes.

Let us consider a particular gene with a low p value. The `rowData` function provides us with all the information about the gene model; each of the exons is represented as a *GRanges*, and these are tied together as a *GRangesList*. We use the function `range` to extract the entire range of the gene, from the start of the left-most exon to the end of the right-most exon. This is all the information we need in order to find the nearest binding site.

```
> deGeneID <- "ENSG00000099194"
> res[deGeneID,]
```

```
DataFrame with 1 row and 7 columns
                 baseMean log2FoldChange     lfcSE    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000099194      8789           0.42    0.0209  7.27e-90  2.12e-85
                    symbol      entrez
                <character> <character>
ENSG00000099194         SCD        6319
```

```
> deGene <- range(rowData(dds[deGeneID,])[[1]])
> names(deGene) <- deGeneID
> deGene
```

```
GRanges with 1 range and 0 metadata columns:
                   seqnames                ranges strand
                      <Rle>             <IRanges>  <Rle>
  ENSG00000099194       10 [102106877, 102124591]      +
  ---
  seqlengths:
                   1                 2 ...            LRG_98            LRG_99
           249250621         243199373 ...             18750             13294
```

We would like to compare the location of this gene with the location of annotated estrogen receptor binding sites, provided by the UCSC Genome Browser. We must first alter the sequence name (the chromosome name) of the differentially expressed gene, as the Ensembl gene annotation does not use the "chr" prefix, which the UCSC chromosomes are annotated with. (Note that we ignore here another complication, which is that the Ensembl sequence "MT" corresponds to the UCSC's sequence "chrM".) We use the `paste0` function, which concatenates the character vectors provided without using any separating characters. We then create a range which is 10 Mb to the left and right of the start of the `deGene` object.

```
> as.character(seqnames(deGene))
```

```
[1] "10"
```

```
> ucscChrom <- paste0("chr",as.character(seqnames(deGene)))
> ucscRanges <- ranges(flank(deGene,width=10e6,both=TRUE))
> subsetRange <- GRanges(ucscChrom, ucscRanges)
> subsetRange
```

```
GRanges with 1 range and 0 metadata columns:
                  seqnames               ranges strand
                     <Rle>            <IRanges>  <Rle>
  ENSG00000099194    chr10 [92106877, 112106876]      *
  ---
  seqlengths:
   chr10
      NA
```

We now provide code which would download a track from the UCSC Genome Browser, in our case a track containing transcription factor binding sites obtained from ChIP-Seq experiments across various cell lines, generated by the ENCODE project.

The track names and table names must match a track name provided by the UCSC Genome Browser. For more information on these steps, see the detailed instructions in the vignette of the useful Bioconductor package *rtracklayer*.

```
> ##
> ## Please do not run this code if you do not have an internet connection,
> ## alternatively use the local file import in the next code chunk.
> ##
> library( "rtracklayer" )
> trackName <- "wgEncodeRegTfbsClusteredV2"
> tableName <- "wgEncodeRegTfbsClusteredV2"
> trFactor <- "ERalpha_a"
> mySession <- browserSession()
> ucscTable <- getTable(ucscTableQuery(mySession, track=trackName,
+                                      range=subsetRange, table=tableName,
+                                      name=trFactor))
```

Here we use a locally cached copy of `ucscTable`:

```
> ucscTableFile <- "localUcscTable.csv"
> ucscTable <- read.csv(ucscTableFile, stringsAsFactors=FALSE)
```

We now can use the downloaded table of annotated estrogen receptor peaks. Whether to use a cutoff on the provided peak scores at this step, or what scores cutoff to use, depends on your experience with the specific transcription factor and the ChIP-Seq experiments used to define these peaks. It often makes sense to visualize tracks in a genome browser in order to get a sense of the qualitative difference between peaks of different scores.

We create a *GRanges* object, `peaks`, from the table obtained from UCSC, and then we convert the chromosome names back to the Ensembl style using the global substitute function, `gsub`. Finally, we enforce that the sequence levels of the peaks match the sequence levels of the differential expressed gene, which is necessary for performing the nearest matching in the following code chunk.

```
> peaks <- with(ucscTable, GRanges(chrom, IRanges(chromStart, chromEnd),
+                                  score=score))
> seqlevels(peaks) <- gsub("chr(.+)","\\1",seqlevels(peaks))
> seqlevels(peaks) <- seqlevels(deGene)
```

Now we have two *GRanges* objects, defined over the same chromosomes, so we can use the `distanceToNearest` function from the package *GRanges*. This provides a *Hits* object, which contains the matches between the "query" and the "subject", the first and second arguments to the function, as well as the distance from the query to the subject. As we only have a single query, there should only be one nearest range in the subject. See the documentation via `?distanceToNearest` and `?Hits` for more information on the options for the this matching step.

```
> d2nearest <- distanceToNearest(deGene, peaks)
```

Question 9: What is the distance from the differentially expressed gene to all the peaks?

We can now examine the object `d2nearest`. This tells us that the nearest peak is 44 base pairs from the differential expressed gene.

```
> d2nearest
```

```
Hits of length 1
queryLength: 1
subjectLength: 168
  queryHits subjectHits  distance
   <integer>   <integer> <integer>
 1         1         118        44
```

The function `subjectHits` is used to extract the index of the closest hit in the `peaks` object.

```
> deGene
```

```
GRanges with 1 range and 0 metadata columns:
                   seqnames                ranges strand
                      <Rle>             <IRanges>  <Rle>
  ENSG00000099194       10 [102106877, 102124591]      +
  ---
  seqlengths:
                   1                2 ...            LRG_98            LRG_99
           249250621        243199373 ...             18750             13294
```

```
> peaks[subjectHits(d2nearest)]
```

```
GRanges with 1 range and 1 metadata column:
      seqnames                ranges strand |     score
         <Rle>             <IRanges>  <Rle> | <integer>
  [1]       10 [102124636, 102124912]      * |        76
  ---
  seqlengths:
                   1                2 ...            LRG_98            LRG_99
                  NA               NA ...                NA                NA
```

Is 44 base pairs unexpectedly close? Here we make a simple plot of the starting points of the peaks and gene along the chromosome, to get a sense of the distribution of peaks and how surprised we should be with the distance of the nearest. To identify the nearest peak, we construct a logical vector `peakNearest`, which can be used to change the y value and the color of the point corresponding to the nearest peak.

```
> plotRange <- start(deGene) + 1e6 * c(-1,1)
> peakNearest <- ( seq_along(peaks) == subjectHits(d2nearest) )
> plot(x=start(peaks), y=ifelse(peakNearest,.3,.2),
+      ylim=c(0,1), xlim=plotRange, pch='p',
+      col=ifelse(peakNearest,"red","grey60"),
+      yaxt="n", ylab="",
+      xlab=paste("2 Mb on chromosome",as.character(seqnames(deGene))))
> points(x=start(deGene),y=.8,pch='g')
```

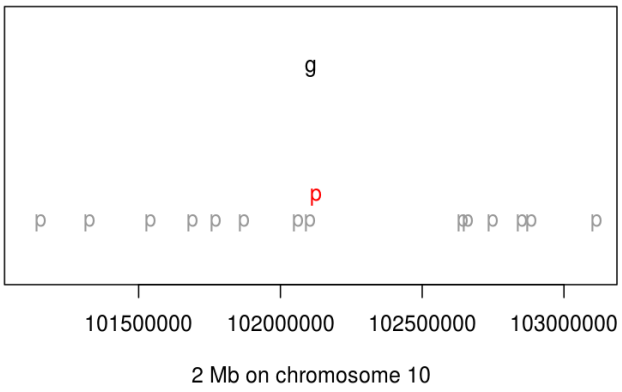

Figure 4: A 2 Mb genomic range showing the location of the differentially expressed gene (labelled 'g'), and the peaks (labelled 'p'). As there are only 14 peaks spread over 2 Mb, it is surprising to find a peak 44 base pairs away from the differentially expressed gene.

Again, the biological relevance of the distances between peaks and genes is another matter, especially considering the data are from different sources. An important consideration when investigating the distribution of distances between two sets of genomic features, is how the individual sets cluster along the genome.

Question 10: Are the peaks relatively uniformly distributed?

6 Working with rlog-transformed data

6.1 The rlog transform

Many common statistical methods for exploratory analysis of multidimensional data, especially methods for clustering and ordination (e. g., principal-component analysis and the like), work best for (at least approximately) homoskedastic data; this means that the variance of an observable (i.e.,

here, the expression strength of a gene) does not depend on the mean. In RNA-Seq data, however, variance grows with the mean. For example, if one performs PCA directly on a matrix of normalized read counts, the result typically depends only on the few most strongly expressed genes because they show the largest absolute differences between samples. A simple and often used strategy to avoid this is to take the logarithm of the normalized count values; however, now the genes with low counts tend to dominate the results because, due to the strong Poisson noise inherent to small count values, they show the strongest relative differences between samples.

As a solution, *DESeq2* offers the *regularized-logarithm transformation*, or *rlog* for short. For genes with high counts, the rlog transformation differs not much from an ordinary $\log_2$ transformation. For genes with lower counts, however, the values are shrunken towards the genes' averages across all samples. Using an empirical Bayesian prior in the form of a *ridge penality*, this is done such that the rlog-transformed data are approximately homoskedastic.

The function `rlogTransform` returns a *SummarizedExperiment* object which contains the rlog-transformed values in its *assay* slot:

```
> rld <- rlogTransformation(dds)
> head( assay(rld) )
```

To show the effect of the transformation, we plot the first sample against the second, first simply using the `log2` function (after adding 1, to avoid taking the log of zero), and then using the rlog-transformed values.

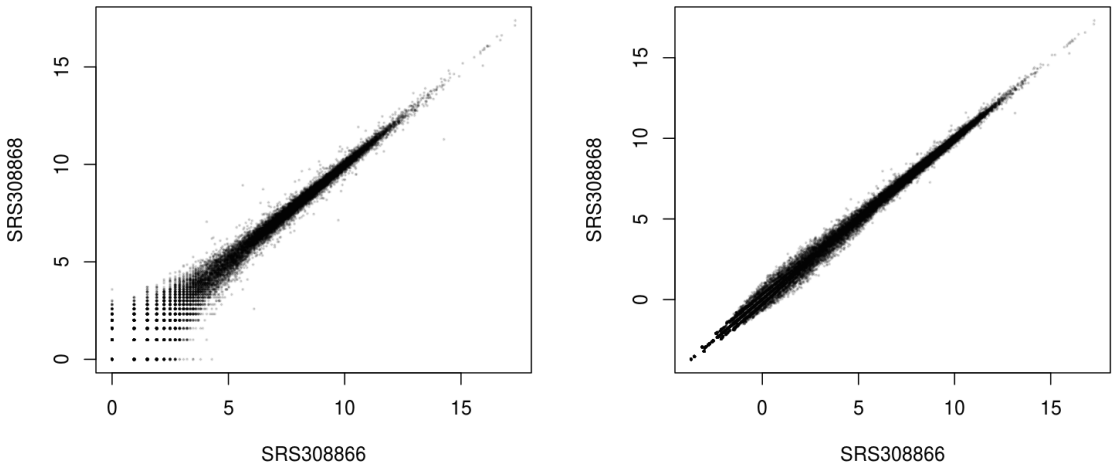

Figure 5: Scatter plot of sample 2 versus sample 1. Left: using an ordinary log_2 transformation. Right: Using the rlog transformation.

```
> par( mfrow = c( 1, 2 ) )
> plot( log2( 1+counts(dds, normalized=TRUE)[, 1:2] ), col="#00000020", pch=20, cex=0.3 )
> plot( assay(rld)[, 1:2], col="#00000020", pch=20, cex=0.3 )
```

Note that, in order to make it easier to see where several points are plotted on top of each other, we set the plotting color to a semi-transparent black (encoded as `#00000020`) and changed the points

to solid disks (`pch=20`) with reduced size (`cex=0.3`)[1].

In Figure 5, we can see how genes with low counts seem to be excessively variable on the ordinary logarithmic scale, while the rlog transform compresses differences for genes for which the data cannot provide good information anyway.

6.2 Sample distances

A useful first step in an RNA-Seq analysis is often to assess overall similarity between samples: Which samples are similar to each other, which are different? Does this fit to the expectation from the experiment's design?

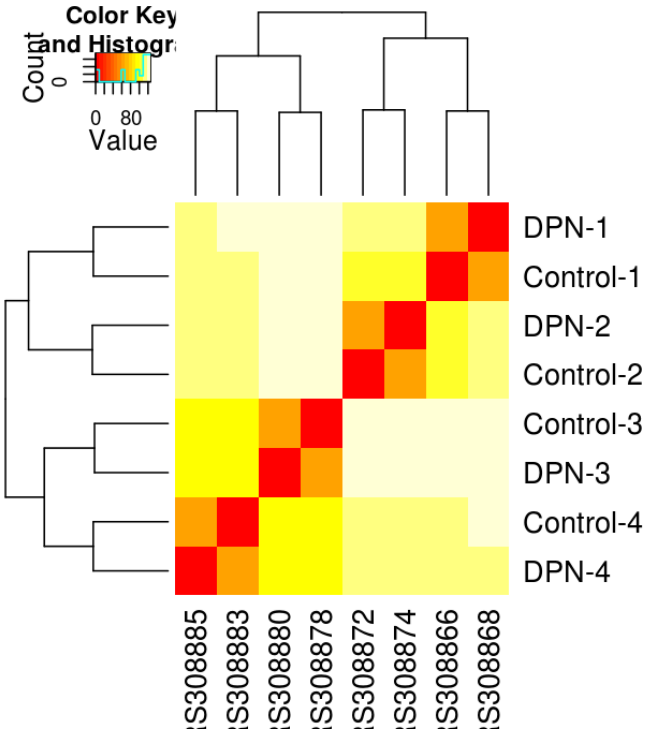

Figure 6: Heatmap of Euclidean sample distances after rlog transformation.

We use the R function `dist` to calculate the Euclidean distance between samples. To avoid that the distance measure is dominated by a few highly variable genes, and have a roughly equal contribution from all genes, we use it on the rlog-transformed data:

```
> sampleDists <- dist( t( assay(rld) ) )
> sampleDists
```

	SRS308866	SRS308868	SRS308872	SRS308874	SRS308878	SRS308880	SRS308883
SRS308868	64.0						
SRS308872	107.5	111.8					
SRS308874	106.8	108.9	65.2				
SRS308878	120.5	125.6	121.6	123.5			
SRS308880	118.0	122.3	121.7	121.8	63.2		
SRS308883	115.5	120.2	114.6	116.4	99.4	98.2	
SRS308885	110.8	113.6	111.5	111.6	99.9	95.2	64.0

[1]The function `heatscatter` from the package *LSD* offers a colourful alternative.

Note the use of the function `t` to transpose the data matrix. We need this because `dist` calculates distances between data *rows* and our samples constitute the columns.

We visualize the distances in a heatmap, using the function `heatmap.2` from the *gplots* package.

```
> sampleDistMatrix <- as.matrix( sampleDists )
> rownames(sampleDistMatrix) <- paste( colData(rld)$treatment,
+    colData(rld)$patient, sep="-" )
> library( "gplots" )
> heatmap.2( sampleDistMatrix, trace="none" )
```

Note that we have changed the row names of the distance matrix to contain treatment type and patient number instead of sample ID, so that we have all this information in view when looking at the heatmap (Fig. 6).

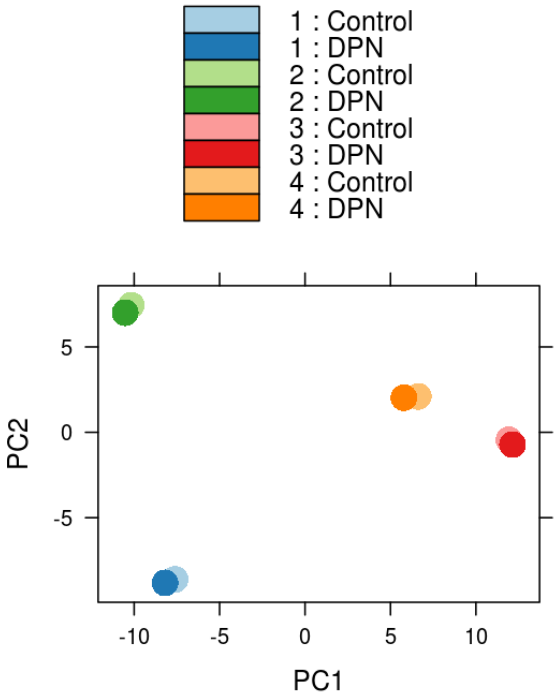

Figure 7: Principal components analysis (PCA) of samples after rlog transformation.

Question 11: Some people find the colour scheme used in Figure 6 ugly. Make a better version. *Hint:* Look at the sequential colour schemes in the *RColorBrewer* package and at the `colorRampPalette` function.

Another way to visualize sample-to-sample distances is a principal-components analysis (PCA). In this ordination method, the data points (i.e., here, the samples) are projected onto the 2D plane such that they spread out optimally (Fig. 7).

```
> print( plotPCA( rld, intgroup = c( "patient", "treatment") ) )
```

Here, we have used the function `plotPCA` which comes with *DESeq2*. The two terms specified as `intgroup` are column names from our sample data; they tell the function to use them to choose colours.

From both visualizations, we see that the differences between patients is much larger than the difference between treatment and control samples of the same patient. This shows why it was important to account for this paired design ("paired", because each treated sample is paired with one control sample from the *same* patient). We did so by using the design formula `!~ patient  treatment!` when setting up the data object in the beginning. Had we used an un-paired analysis, by specifying only `~ treatment`, we would not have found many hits, because then, the patient-to-patient differences would have drowned out any treatment effects.

Here, we have performed this sample distance analysis towards the end of our analysis. In practice, however, this is a step suitable to give a first overview on the data. Hence, one will typically carry out this analysis as one of the first steps in an analysis. To this end, you may also find the function `arrayQualityMetrics`, from the equinymous package, useful.

6.3 Gene clustering

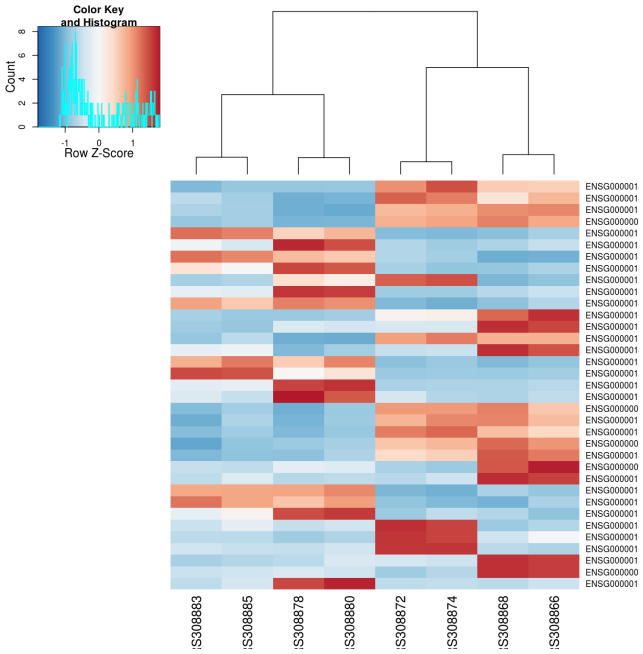

Figure 8: Heatmap with gene clustering.

In the heatmap of Fig. 6, the dendrogram at the side shows us a hierarchical clustering of the samples. Such a clustering can also be performed for the genes.

Since the clustering is only relevant for genes that actually carry signal, one usually carries it out only for a subset of most highly variable genes. Here, for demonstration, let us select the 35 genes with the highest variance across samples:

```
> library( "genefilter" )
> topVarGenes <- head( order( rowVars( assay(rld) ), decreasing=TRUE ), 35 )
```

The heatmap becomes more interesting if we do not look at absolute expression strength but rather at the amount by which each gene deviates in a specific sample from the gene's average across all samples. Hence, we center and scale each genes' values across samples, and plot a heatmap.

```
> heatmap.2( assay(rld)[ topVarGenes, ], scale="row",
+       trace="none", dendrogram="column",
+       col = colorRampPalette( rev(brewer.pal(9, "RdBu")) )(255))
```

We can now see (Fig. 8) blocks of genes which covary across patients. Often, such a heatmap is insightful, even though here, seeing these variations across patients is of limited value because we are rather interested in the effects between the two samples from each patient.

7 Advanced Questions

For these questions, we provide (and probably have) no solutions, advanced readers are encouraged to explore them.

1. *DESeq2* performs the shrinkage of the dispersion estimates by fitting a parametric curve on the mean of normalized counts (cf. Figure 2). However, one could argue that the biological variability of genes should not be a function of counts, but of counts per gene length (i. e., expression level), and that regression on that covariate should lead to a better fit. Write your own version of the `estimateDispersions` function to explore this question.

2. What is the contribution of UTR length variations to the between-replicates variability modelled by *DESeq2*? The read counting script (available in the vignette of *parathyroidSE*) uses all exons of the genes, which includes UTRs. Would detection power be increased –or would we preferentially detect different phenomena– if we left out UTRs from the counting (i. e. count reads that fall on coding exons only); or indeed, if we looked only at UTRs?

References

[1] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *PNAS*, 107(21):9546–9551, 2010.

8 Solutions

Answer 1:

```
> nrow(parathyroidGenesSE)
```

```
[1] 60620
```

Answer 2:

```
> rowData( parathyroidGenesSE )
```

```
GRangesList of length 60620:
$ENSG00000000003
GRanges with 17 ranges and 2 metadata columns:
       seqnames              ranges strand   |   exon_id       exon_name
          <Rle>           <IRanges>  <Rle>   | <integer>     <character>
   [1]        X [99883667, 99884983]      -   |    653684 ENSE00001459322
   [2]        X [99885756, 99885863]      -   |    653685 ENSE00000868868
   [3]        X [99887482, 99887565]      -   |    653686 ENSE00000401072
   [4]        X [99887538, 99887565]      -   |    653687 ENSE00001849132
   [5]        X [99888402, 99888536]      -   |    653688 ENSE00002890912
   ...      ...                 ...    ... ...       ...             ...
  [13]        X [99890555, 99890743]      -   |    653696 ENSE00002799002
  [14]        X [99891188, 99891686]      -   |    653697 ENSE00001886883
  [15]        X [99891605, 99891803]      -   |    653698 ENSE00001855382
  [16]        X [99891790, 99892101]      -   |    653699 ENSE00001863395
  [17]        X [99894942, 99894988]      -   |    653700 ENSE00001828996

...
<60619 more elements>
---
seqlengths:
                 1                 2 ...            LRG_98            LRG_99
         249250621         243199373 ...             18750             13294
```

Answer 3:

The function `sapply` expects an R function as its second argument. Here, we want to provide it with the function for vector subsetting (as in `a[1]`), and the name of this function is `[`. However, if we provide that name without the quotation marks, the R interpreter gets confused and complains about the unexpected symbol (try this out). Hence we need to quote the function name in our call to `sapply`.

Answer 4: The raw counts and normalized counts of a *DESeqDataSet* object are available via the accessor function `counts`, which has an argument `normalized`, which defaults to `FALSE`.

```
> all.equal(res$baseMean, rowMeans(counts(dds)))
```

```
[1] "Mean relative difference: 0.058"
```

```
> all.equal(res$baseMean, rowMeans(counts(dds,normalized=TRUE)))
```

```
[1] TRUE
```

Answer 5:

```
> table(sign(resSig$log2FoldChange))
```

```
 -1   1
280 225
```

Answer 6:

```
> plotMA(dds, pvalCutoff = 0.5, ylim = c( -1.5, 1.5) )
```

See Figure 9.

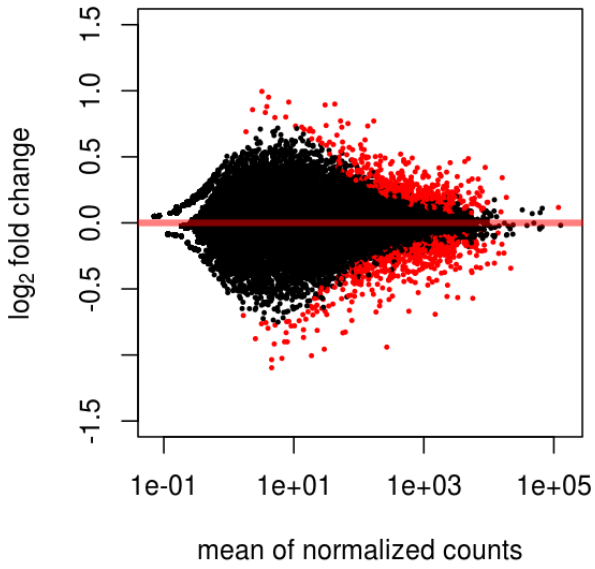

Figure 9: The MA-plot with red points indicating adjusted p value less than 0.5.

Answer 7: Genes that are not differentially expressed have p values that are approximately uniformly distributed between 0 and 1. This gives rise to the floor of bars of equal heights. The truly differentially expressed genes give rise to the tall bar(s) at the very left – but only to that part of the bars that raises above the uniform floor. Of course, we cannot know which of the genes in these tall bars are true ones and which are not. When only looking at the bars to the left of our chosen p value cut-off, the ratio of "floor" area to total area provides an estimate of the false discovery rate. This is a graphical way of understanding FDR.

The rule that p values from null cases are uniform is true only for continuous test statistics. However, for genes with low counts, the fact that we are working with integer counts becomes noticeable, and gives rise to the spikes at intermediate p values.

Answer 8: Run

```
> hist( res$pvalue[keep], breaks=100 )
```

See Figure 10. As explained before, the spikes were caused by genes with low counts. Having removed these, our p value histogram now looks smoother.

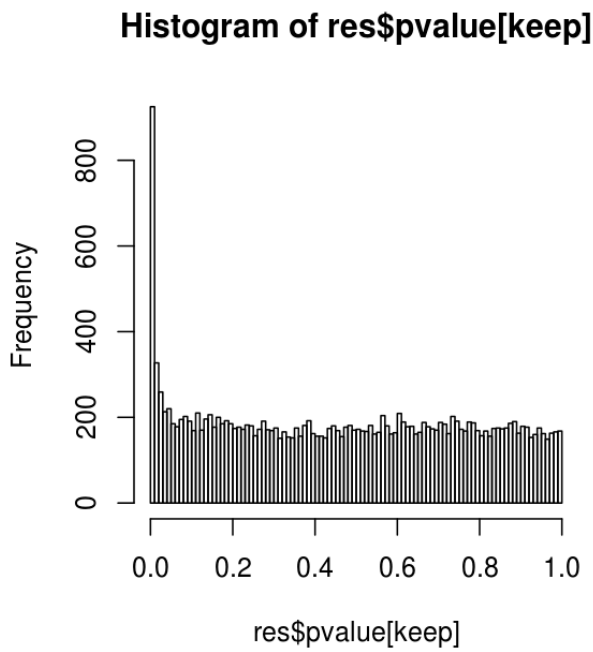

Figure 10: Histogram of the p values returned by the test for differential expression.

In this vignette, we have determined the value for `filterThreshold`, 2, by looking at Figure 1. More formal, automatable ways exist; if you are interested, please have a look at the vignette *Diagnostics for independent filtering* in the *genefilter* package.

Answer 9:

```
> distance(deGene, peaks)
```

```
  [1] 9475043 9429124 9425724 9416446 9378083 9377693 9292441 9147967 9052346
 [10] 8985894 8790050 8757482 8746704 8742783 8675290 8674776 8664975 8656382
 [19] 8482843 8223779 8221921 8135505 8001206 7455659 7085825 6942795 6925407
 [28] 6901309 6888085 6884994 6881071 6880702 6864672 6780251 6775780 6605483
 [37] 6599855 6596682 6588693 6583714 6571495 6420060 6351944 6351142 6329889
 [46] 6319659 6310496 6310171 6308777 6304329 6293603 6288349 6271335 6268792
 [55] 6262964 5922095 5213831 5117191 5058161 5057760 5038066 4950759 4846635
 [64] 4840466 4690188 4604926 4150395 4075276 3867680 3837677 3750148 3728194
 [73] 3713685 3672008 3545136 3489160 3483056 3482652 3374528 3145460 3009845
 [82] 3009471 2945580 2938224 2898739 2780411 2775152 2772192 2766539 2697775
 [91] 2134521 2099924 2093212 2080610 2079692 2078665 2077377 2074625 2052826
[100] 2050842 2044992 2044075 2043231 2032448 1972273 1970319 1864915 1843265
[109] 1397483  951762  778758  566968  415886  336977  235037   45750    1216
[118]      44  518646  536372  624809  727175  759664  988971 1236678 1267088
[127] 1482084 1528951 1573655 1600301 1752606 1755604 1767986 1810760 1842340
```

```
[136] 1962210 2003350 2039054 2048291 2137914 2293613 2342835 2345825 2347756
[145] 2399792 2400158 2404602 2409590 2489595 2791121 2803606 2806523 2866416
[154] 2867474 4107727 4108577 4109757 4112806 4115469 4116070 4198736 7246349
[163] 9538145 9592618 9593558 9704719 9860874 9909237
```

Answer 10:

We can answer this question by investigating the inter-peak distances. As all of our peaks are on the same chromosome, we just sort the peak starts and subtract the 2nd from the 1st, the 3rd from the 2nd, etc. Then we call the `summary` function which provides the mean and median. Note that the mean is constricted: it must be equal to the total span divided by the number of inter-peak distances. The median distance is about one quarter of the mean, so the peaks tend to cluster. You can also verify this by plotting the histogram of `peakDists`.

```
>   peakDists <- diff(sort(start(peaks)))
>   summary(peakDists)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    364    5110   32500  116000  103000 3050000
```

```
>   mean(peakDists)
```

```
[1] 116181
```

Answer 11:

```
> library("RColorBrewer")
> colours = colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
> heatmap.2( sampleDistMatrix, trace="none", col=colours)
```

See Figure 11.

9 Session Info

As last part of this document, we call the function `sessionInfo`, which reports the version numbers of R and all the packages used in this session. It is good practice to always keep such a record as it will help to trace down what has happened in case that an R script ceases to work because a package has been changed in a newer version.

```
R version 3.0.0 (2013-04-03)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
```

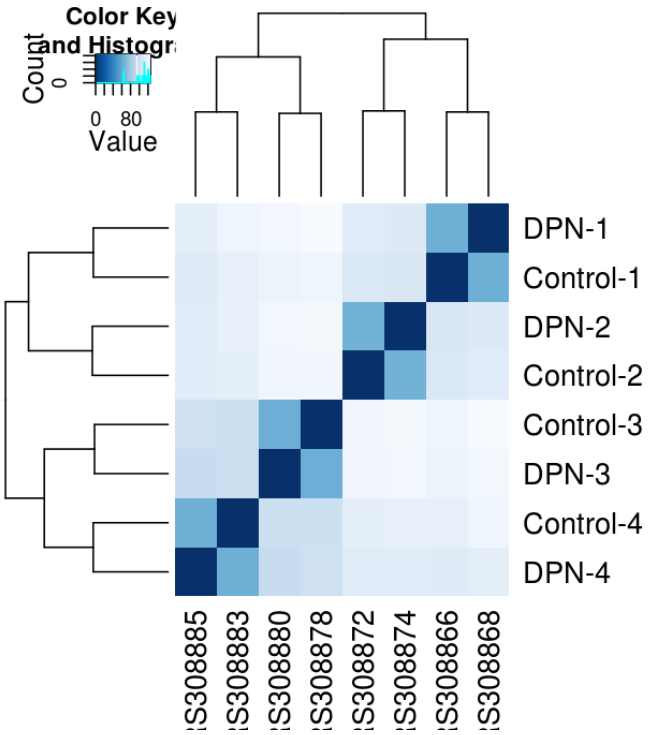

Figure 11: The same heatmap as in Figure 6 but with better colours.

```
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] genefilter_1.42.0     RColorBrewer_1.0-5    gplots_2.11.0.1
 [4] MASS_7.3-26           KernSmooth_2.23-10    caTools_1.14
 [7] gdata_2.12.0.2        gtools_2.7.1          reactome.db_1.44.0
[10] org.Hs.eg.db_2.9.0    RSQLite_0.11.4        DBI_0.2-7
[13] AnnotationDbi_1.22.5  parathyroidSE_0.99.5  DESeq2_1.0.16
[16] RcppArmadillo_0.3.820 Rcpp_0.10.3           lattice_0.20-15
[19] Biobase_2.20.0        GenomicRanges_1.12.4  IRanges_1.18.1
[22] BiocGenerics_0.6.0    cacheSweave_0.6-1     stashR_0.3-5
[25] filehash_2.2-1

loaded via a namespace (and not attached):
 [1] annotate_1.38.0   Biostrings_2.28.0 bitops_1.0-5      digest_0.6.3
 [5] locfit_1.5-9.1    Rsamtools_1.12.3  splines_3.0.0     stats4_3.0.0
 [9] survival_2.37-4   tools_3.0.0       XML_3.96-1.1      xtable_1.7-1
[13] zlibbioc_1.6.0
```