metaRNASeq: Differential meta-analysis of RNA-seq data

Guillemette Marot, Florence Jaffrézic, Andrea Rau

Modified: February 26, 2013. Compiled: February 26, 2014

Abstract

This vignette illustrates the use of the *metaRNASeq* package to combine data from multiple RNA-seq experiments. Based both on simulated and real publicly available data, it also explains the way the p-value data provided in the package have been obtained.

Contents

1 Introduction

High-throughput sequencing (HTS) data, such as RNA-sequencing (RNA-seq) data, are increasingly used to conduct differential analyses, in which gene-by-gene statistical tests are performed in order to identify genes whose expression levels show systematic covariation with a particular condition, such as a treatment or phenotype of interest. Due to their large cost, however, only few biological replicates are often considered in each experiment leading to a low detection power of differentially expressed genes. For this reason,

analyzing data arising from several experiments studying the same question can be a useful way to increase detection power for the identification of differentially expressed genes.

The *metaRNASeq* package implements two p-value combination techniques (inverse normal and Fisher methods); see [4] for additional details. There are two fundamental assumptions behind the use of these p-value combination procedures: first, that p-values have been obtained the same way for each experiment (i.e., using the same model and test); and second, that they follow a uniform distribution under the null hypothesis. In this vignette, we illustrate these p-value combination techniques after obtaining p-values for differential expression in each individual experiment using the *DESeq2* Bioconductor package [1]. Count data are simulated using the `sim.function` provided in the *metaRNASeq* package; see section 2 for additional detail.

2 Simulation study

To begin, we load the necessary packages and simulation parameters:

```
> library(metaRNASeq)
> library(DESeq2)
> data(param)
> dim(param)

[1] 26408     3

> data(dispFuncs)
```

These simulation parameters include the following information:

- `param`: Matrix of dimension (26408 $\times$ 3) containing mean expression in each of two conditions (here, labeled "condition 1" and "condition 2") and a logical vector indicating the presence or absence of differential expression for each of 26,408 genes
- `dispFuncs`: List of length 2, where each list is a vector containing the two estimated coefficients (α_0 and α_1) for the gamma-family generalized linear model (GLM) fit by *DESeq* (version 1.8.3) describing the mean-dispersion relationship for each of the two real datasets considered in [4]. These regressions represent the typical relationship between mean expression values μ and dispersions α in each dataset, where the coefficients α_0 and α_1 are found to parameterize the fit as $\alpha = \alpha_0 + \alpha_1/\mu$.

These parameters were calculated on real data sets from two human melanoma cell lines [5], corresponding to two different studies performed for the same cell line comparison, with two biological replicates per cell line in the first and three per cell line in the second. These data are presented in greater detail in [5] and [2], and are freely available in the Supplementary Materials of the latter.

Once parameters are loaded, we simulate data. We use the `set.seed` function to obtain reproducible results.

```
> set.seed(123)
> matsim <- sim.function(param = param, dispFuncs = dispFuncs)
> sim.conds <- colnames(matsim)
> rownames(matsim) <- paste("tag", 1:dim(matsim)[1],sep="")
> dim(matsim)

[1] 26408    16
```

The simulated matrix data contains 26,408 genes and 4 replicates per condition per study. It is possible to change the number of replicates in each study using either the `nrep` argument or the `classes` argument. Using `nrep` simulates the same number of replicates per condition per study. In order to simulate an unbalanced design, the `classes` argument may be used. For example, setting

```
classes = list(c(1,2,1,1,2,1,1,2),c(1,1,1,2,2,2,2))
```

leads to 5 and 3 replicates in each condition for the first study, and 3 and 4 replicates in each condition in the second.

3 Individual analyses of the two simulated data sets

Before performing a combination of p-values from each study, it is necessary to perform a differential analysis of the individual studies (using the same method). In the following example, we make use of the *DESeq2* package to obtain p-values for differential analyses of each study independently; however, we note that other differential analysis methods (e.g., *edgeR* or *baySeq*) could be used prior to the meta analysis.

3.1 Differential analysis of each individual study with DESeq2

Inputs to DEseq2 methods can be extracted with `extractfromsim` for each individual study whose name appears in the column names of `matsim`, see the following example for study1 and study2.

```
> colnames(matsim)

 [1] "study1cond1" "study1cond1" "study1cond1" "study1cond1"
 [5] "study1cond2" "study1cond2" "study1cond2" "study1cond2"
 [9] "study2cond1" "study2cond1" "study2cond1" "study2cond1"
[13] "study2cond2" "study2cond2" "study2cond2" "study2cond2"
```

```
> simstudy1 <- extractfromsim(matsim,"study1")
> head(simstudy1$study)

     rep1 rep2 rep3 rep4 rep5 rep6 rep7 rep8
tag1  338  401  428  565  476  545  407  367
tag2  919  849 1397 1541  917 1268 1596 1020
tag3  127  166  235  276  133  206  238  127
tag4  224  353  426  252  881  717  889  808
tag5    4    4    8    6    9    5   10    9
tag6  108   61   39   22  158   97   16  107

> simstudy1$pheno

      study condition
rep1 study1 untreated
rep2 study1 untreated
rep3 study1 untreated
rep4 study1 untreated
rep5 study1   treated
rep6 study1   treated
rep7 study1   treated
rep8 study1   treated

> simstudy2 <- extractfromsim(matsim,"study2")
```

Differential analyses for each study are then easily performed using the `DESeqDataSetFromMatrix` method.

```
> dds1 <- DESeqDataSetFromMatrix(countData = simstudy1$study,
+  colData = simstudy1$pheno,design = ~ condition)
> res1 <- results(DESeq(dds1))
> dds2 <- DESeqDataSetFromMatrix(countData = simstudy2$study,
+   colData = simstudy2$pheno,design = ~ condition)
> res2 <- results(DESeq(dds2))
```

We recommand to store both p-value and Fold Change results in lists in order to perform meta-analysis and keep track of the potential conflicts (see section 5)

```
> rawpval <- list("pval1"=res1[["pvalue"]],"pval2"=res2[["pvalue"]])
> FC <- list("FC1"=res1[["log2FoldChange"]],"FC2"=res2[["log2FoldChange"]])
```

Differentially expressed genes in each individual study can also be marked in a matrix `DE`:

```
> adjpval <- list("adjpval1"=res1[["padj"]],"adjpval2"=res2[["padj"]])
> DE <- mapply(adjpval, FUN=function(x) ifelse(x <= 0.05, 1, 0))
> colnames(DE)=c("DEstudy1","DEstudy2")
```

Since the proposed p-value combination techniques rely on the assumption that p-values follow a uniform distribution under the null hypothesis, it is necesary to check that the histograms of raw-pvalues reflect that assumption:

```
> par(mfrow = c(1,2))
> hist(rawpval[[1]], breaks=100, col="grey", main="Study 1", xlab="Raw p-values")
> hist(rawpval[[2]], breaks=100, col="grey", main="Study 2", xlab="Raw p-values")
```

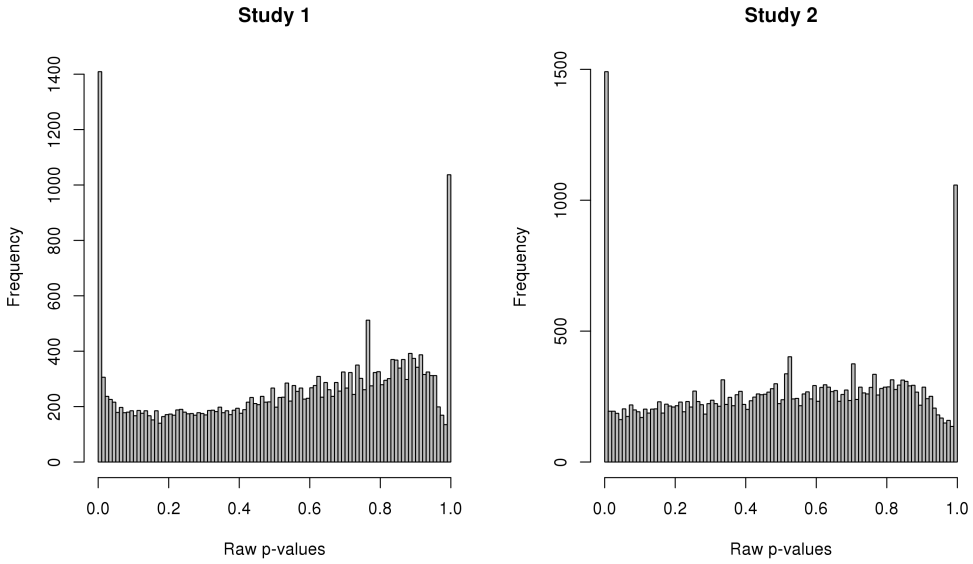

Figure 1: Histograms of raw p-values for each of the individual differential analyses performed using the *DESeq2* package.

The peak near 0 corresponds to differentially expressed genes, no other peak should appear. If another peak appears (for example like in this example where a peak is observed for p-values close to 1), then we suggest to use HTSFilter (see the following section).

3.2 Using HTSFilter to validate the p-value uniform distribution assumption

Genes with very low values of expression often lead to an enrichment of p-values close to 1 as they take on discrete values; as such genes are unlikely to display evidence for

differential expression, it has been proposed to apply *independent filtering* to filter these genes [3]. In addition, the application of such a filter typically removes those genes contributing to a peak of p-values close to 1, leading to a distribution of p-values under the null hypothesis more closely following a uniform distribution. As the proposed p-value combination techniques rely on this assumption, it is sometimes necessary to independently filter genes with very low read counts. For this purpose, we recommand the use of the *HTSFilter* package, see [3] for more details; note that we apply the filter in *HTSFilter* to each study individually after estimating library sizes and per-gene dispersion parameters.

Once the data are filtered, we use the *DESeq* package to perform differential analyses of each of the two individual datasets. The following function `resDESeq1study` is a wrapper of the main functions of the data filter in *HTSFilter* and differential analysis in *DESeq*, selecting the appropriate columns in the simulated data set for each study. The following two steps could be replaced by direct uses of the *HTSFilter* and *DESeq* packages and concatenation of results in one list (see `resDESeq.alt`).

```
> library(DESeq)
> library(HTSFilter)
> resDESeq1study <- function(studyname, alldata, cond1totest="cond1",
+     cond2totest="cond2", fitType = "parametric") {
+   study <- alldata[,grep(studyname,colnames(alldata))]
+   studyconds <- gsub(studyname,"",colnames(study))
+   colnames(study) <- paste(studyconds,1:dim(study)[2],sep=".")
+   cds <- newCountDataSet(study, studyconds)
+   cds <- estimateSizeFactors(cds)
+   cds <- estimateDispersions(cds, method="pooled", fitType=fitType)
+   ## Filter using Jaccard index for each study
+   filter <- HTSFilter(cds, plot=FALSE)
+   cds.filter <- filter$filteredData
+   on.index <- which(filter$on == 1)
+   cat("# genes passing filter", studyname, ":", dim(cds.filter)[1], "\n")
+   res <- as.data.frame(matrix(NA, nrow = nrow(cds), ncol=ncol(cds)))
+   nbT <- nbinomTest(cds.filter, cond1totest, cond2totest)
+   colnames(res) <- colnames(nbT)
+   res[on.index,] <- nbT
+   res
+ }
```

The wrapper can be applied simultaneously to the two studies with the use of the function `lapply`:

```
> studies <- c("study1", "study2")
> resDESeq <- lapply(studies,
+   FUN=function(x) resDESeq1study(x, alldata=matsim))
```

```
# genes passing filter study1 : 14044
# genes passing filter study2 : 13839
```

Note that `resDESeq` can be created directly from two or more *DEseq* results called `res.study1`, `res.study2`, ...:

```
resDESeq.alt <- list(res.study1,res.study2)
```

Since only p-values are necessary to perform meta-analysis, we keep them in lists called `rawpval` for raw p-values and `adjpval` for p-values adjusted to correct for multiple testing (e.g., to control the false discovery rate at 5% using the Benjamini-Hochberg method).

```
> rawpval <- lapply(resDESeq, FUN=function(res) res$pval)
> adjpval <- lapply(resDESeq, FUN=function(res) res$padj)
> DE <- mapply(adjpval, FUN=function(x) ifelse(x <= 0.05, 1, 0))
> colnames(DE)=paste("DE",studies,sep=".")
```

`DE` returns a matrix with 1 for genes identified as differentially expressed and 0 otherwise (one column per study). To confirm that the raw p-values under the null hypothesis are roughly uniformly distributed, we may also inspect histograms of the raw p-values from each of the individual differential analyses (see Figure˜2):

```
> par(mfrow = c(1,2))
> hist(rawpval[[1]], breaks=100, col="grey", main="Study 1",
+   xlab="Raw p-values")
> hist(rawpval[[2]], breaks=100, col="grey", main="Study 2",
+   xlab="Raw p-values")
```

4 Use of p-value combination techniques

The code in this section may be used independently from the previous section if p-values from each study have been obtained using the same differential analysis test between the different studies. Vectors of p-values must have the same length; `rawpval` is a list (or data.frame) containing the vectors of raw p-values obtained from the individual differential analyses of each study.

The p-value combination using the Fisher method may be performed with the `fishercomb` function, and the subsequent p-values obtained from the meta-analysis may be examined (Figure˜3, left):

```
> fishcomb <- fishercomb(rawpval, BHth = 0.05)
> hist(fishcomb$rawpval, breaks=100, col="grey", main="Fisher method",
+   xlab = "Raw p-values (meta-analysis)")
```

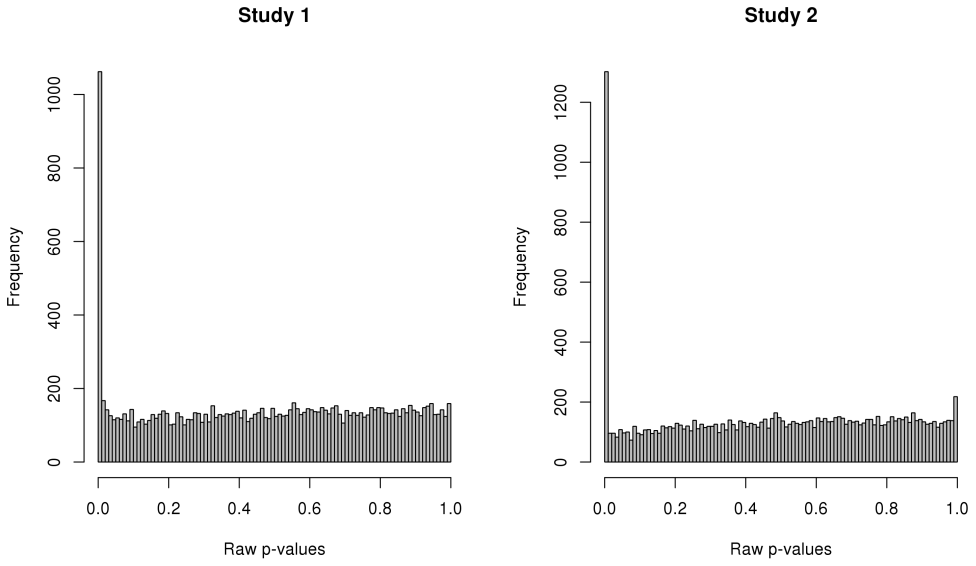

Figure 2: Histograms of raw p-values for each of the individual differential analyses performed using the *HTSFilter* and *DESeq* packages.

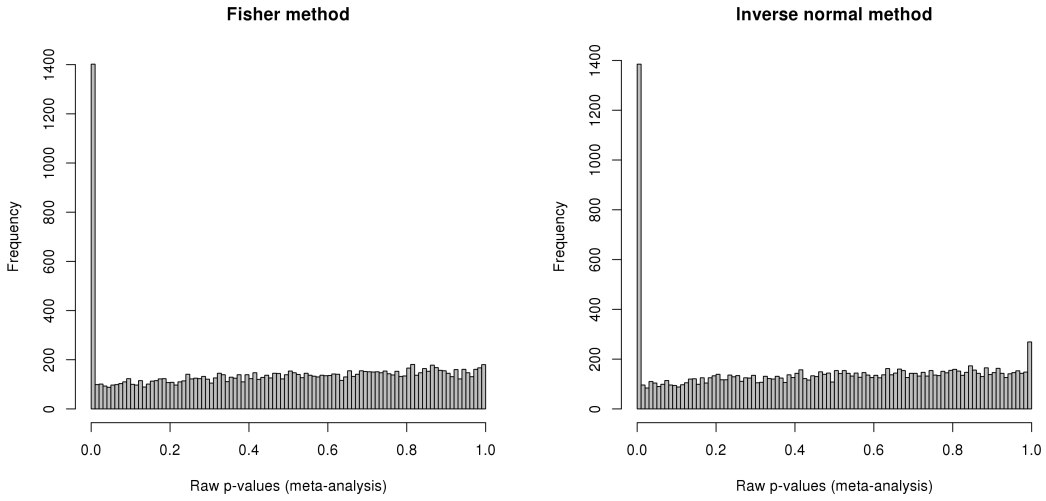

Figure 3: (Left) Histogram of raw p-values obtained after a meta-analysis of all studies, with p-value combination performed using the Fisher method. (Right) Histogram of raw p-values obtained after a meta-analysis of all studies, with p-value combination performed using the inverse normal method.

The use of the inverse normal combination technique requires the choice of a weight for each study. In this example, we choose `nrep=8`, since 8 replicates had been simulated in each study. As before, we may examine a histogram of the subsequent p-values obtained from the meta-analysis (Figure˜3, right).

```
> invnormcomb <- invnorm(rawpval,nrep=c(8,8), BHth = 0.05)
> hist(invnormcomb$rawpval, breaks=100, col="grey",
+   main="Inverse normal method",
+   xlab = "Raw p-values (meta-analysis)")
```

Finally, we suggest summarizing the results of the individual differential analyses as well as the differential meta-analysis (using the Fisher and inverse normal methods) in a data.frame:

```
> DEresults <- data.frame(DE,
+   "DE.fishercomb"=ifelse(fishcomb$adjpval<=0.05,1,0),
+   "DE.invnorm"=ifelse(invnormcomb$adjpval<=0.05,1,0))
> head(DEresults)
```

```
  DE.study1 DE.study2 DE.fishercomb DE.invnorm
1         0         0             0          0
2         0         0             0          0
3         0         0             0          0
4         1         1             1          1
5        NA        NA            NA         NA
6         0         0             0          0
```

5 Treatment of conflicts in differential expression

As pointed out in [4], it is not possible to directly avoid conflicts between over- and under- expressed genes in separate studies that appear in differential meta-analyses of RNA-seq data. We thus advise checking that individual studies identify differential expression in the same direction (i.e., if in one study, a gene is identified as differentially over-expressed in condition 1 as compared to condition 2, it should not be identified as under-expressed in condition 1 as compared to condition 2 in a second study). Genes displaying contradictory differential expression in separate studies should be removed from the list of genes identified as differentially expressed via meta-analysis.

We build a matrix `signsFC` gathering all signs of fold changes from individual studies.

```
> signsFC <- mapply(FC, FUN=function(x) sign(x))
> sumsigns <- apply(signsFC,1,sum)
> commonsgnFC <- ifelse(abs(sumsigns)==dim(signsFC)[2], sign(sumsigns),0)
```

The vector `commonsgnFC` will return a value of 1 if the gene has a positive $\log_2$ fold change in all studies, -1 if the gene has a negative $\log_2$ fold change in all studies, and 0 if contradictory $\log_2$ fold changes are observed across studies (i.e., positive in one and negative in the other). By examining the elements of `commonsgnFC`, it is thus possible to identify genes displaying contradictory differential expression among studies.

```
> unionDE <- unique(c(fishcomb$DEindices,invnormcomb$DEindices))
> FC.selecDE <- data.frame(DEresults[unionDE,],do.call(cbind,FC)[unionDE,],
+   signFC=commonsgnFC[unionDE], DE=param$DE[unionDE])
> keepDE <- FC.selecDE[which(abs(FC.selecDE$signFC)==1),]
> conflictDE <- FC.selecDE[which(FC.selecDE$signFC == 0),]
> dim(FC.selecDE)

[1] 1356    8

> dim(keepDE)

[1] 1183    8

> dim(conflictDE)

[1] 173   8

> head(keepDE)

   DE.study1 DE.study2 DE.fishercomb DE.invnorm        FC1
4          1         1             1          1  1.2690129
11         1         1             1          1 -0.9249095
22         1         1             1          1 -1.0931078
36         1         1             1          1 -2.6692474
55         0         1             1          1 -0.4072206
59         1         1             1          1  1.0937474
          FC2 signFC   DE
4   2.0235564      1 TRUE
11 -0.5442498     -1 TRUE
22 -0.9702876     -1 TRUE
36 -2.6921846     -1 TRUE
55 -1.0288618     -1 TRUE
59  1.3215246      1 TRUE
```

Note that out of all the conflicts, 147 represented genes were simulated to be truly differentially expressed.

```
> table(conflictDE$DE)

FALSE  TRUE
   26   147
```

6 Session Info

```
> sessionInfo()

R version 3.0.3 beta (2014-02-25 r65077)
Platform: x86_64-unknown-linux-gnu (64-bit)

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

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

other attached packages:
 [1] HTSFilter_1.2.1         DESeq_1.14.0
 [3] lattice_0.20-24         locfit_1.5-9.1
 [5] Biobase_2.22.0          DESeq2_1.2.10
 [7] RcppArmadillo_0.4.000.4 Rcpp_0.11.0
 [9] GenomicRanges_1.14.4    XVector_0.2.0
[11] IRanges_1.20.6          BiocGenerics_0.8.0
[13] metaRNASeq_0.4

loaded via a namespace (and not attached):
 [1] AnnotationDbi_1.24.0 DBI_0.2-7
 [3] RColorBrewer_1.0-5   RSQLite_0.11.4
 [5] XML_3.98-1.1         annotate_1.40.0
 [7] edgeR_3.4.2          genefilter_1.44.0
 [9] geneplotter_1.40.0   grid_3.0.3
[11] limma_3.18.13        splines_3.0.3
[13] stats4_3.0.3         survival_2.37-7
[15] tools_3.0.3          xtable_1.7-1
```

References

[1] S.˜Anders and W.˜Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(R106):1–28, 2010.

[2] M.-A. Dillies, A.˜Rau, J.˜Aubert, C.˜Hennequet-Antier, M.˜Jeanmougin, N.˜Servant, C.˜Keime, G.˜Marot, D.˜Castel, J.˜Estelle, G.˜Guernec, B.˜Jagla, L.˜Jouneau, D.˜Laloë, C.˜Le˜Gall, B.˜Schaëffer, S.˜Le˜Crom, and F.˜Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 2012. doi: 10.1093/bib/bbs046.

[3] A.˜Rau, M.˜Gallopin, G.˜Celeux, and F.˜Jaffrézic. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, 2013.

[4] A.˜Rau, G.˜Marot, and F.˜Jaffrézic. Differential meta-analysis of RNA-seq data from multiple studies. *ArXiv e-prints*, page 1306.3636, 2013.

[5] T.˜Strub, S.˜Giuliano, T.˜Ye, C.˜Bonet, C.˜Keime, D.˜Kobi, S.˜Le˜Gras, M.˜Cormont, R.˜Ballotti, C.˜Bertolotto, and I.˜Davidson. Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma. *Oncogene*, 30:2319–2332, 2011.