Differential expression analysis of RNA-seq data with the HTSDiff package

S. Balzergue, G. Rigaill, V. Brunaud, E. Blondet, A. Rau[1], O. Rogier, J. Caius,
C. Maugis-Rabusseau, L. Soubigou-Taconnat, S. Aubourg, C. Lurin, E. Delannoy,
and M.-L. Martin-Magniette.

[1]andrea.rau@jouy.inra.fr

HTSDiff version 1.0.5

Abstract

This vignette explains the use of the *HTSDiff* package for the differential analysis of RNA-seq data from two experimental conditions. A full presentation of the statistical method implemented in the *HTSDiff* package may be found in [1].

Contents

1 Introduction to HTSDiff

Numerous methods of differential analysis for RNA-seq data have been developed in recent years (e.g., [2, 3] among many others). Generally speaking, the majority of these methods rely on the same general principle as a typical microarray differential analysis: for each detected gene, a statistical test is performed to identify whether or not the normalized expression level in each condition is compatible with the per-gene null hypothesis of no difference in expression between the tested conditions (i.e., differences in expression may be attributed solely to technical and biological variability). As a small number of replicates are typically available in current RNA-seq experiments, a variety of approaches have been proposed to obtain appropriate estimates of this gene-specific variability by sharing information across the whole set of genes. In contrast to these test-based approaches, the *HTSDiff* package recasts the differential analysis as an unsupervised classification problem using a model-based clustering approach. In this way, the full gene population is jointly modeled to distinguish a limited number of gene groups, each behaving differently, and the resulting model is interpreted in terms of differential expression.

1.1 Citation

Please cite the appropriate articles when publishing results based on this software:

- Balzergue, D. *et al.* (2014) HTSDiff: A model-based clustering alternative to test-based methods in differential gene expression analyses by RNA-seq benchmarked on real and synthetic datasets (submitted).
  Introduced the use of a Poisson mixture model for differential analyses.

- Rau, A. *et al.* (2014) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models (submitted).
  Introduced the Poisson mixture model for RNA-seq data, described parameter initialization and estimation.

1.2 Quick start

A classic *HTSDiff* analysis might look like the following, where we assume there are four RNA-seq libraries in two conditions, and the counts are stored in tab-delimted text file:

```
> y <- read.table("counts.txt")
> conds <- c(1,1,2,2)
> mod <- HTSDiff(y, conds)
> DEresults <- mod$res
```

For the time being, *HTSDiff* has been defined only for two-group comparisons.

1.3 Description of HTSDiff model

We model the joint distribution of digital gene expression over all samples by a finite mixture of Poisson distributions, where samples are assumed to be conditionally independent given the components. In particular, let Y_{ijl} be the random variable corresponding to the digital gene expression measure (DGE) for gene i ($i = 1, \ldots, n$) of condition j ($j =$ 1,2) in biological replicate l ($l = 1, \ldots, r_j$), with y_{ijl} being the corresponding observed value of Y_{ijl}. We assume that the overall gene population is distributed under the following mixture:

$$f(\mathbf{y}; K, \mathbf{\Psi}_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \prod_{l=1}^{r_j} \mathcal{P}(y_{ijl}; \mu_{ijlk}), \tag{1}$$

where K is the number of clusters, $\mathbf{\Psi}_K = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\mu}')'$, $\boldsymbol{\mu}'$ contains all of the parameters in $\{\boldsymbol{\mu}_{ijlk}\}_{i,j,l,k}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ are the mixing proportions, with $\pi_k \in (0, 1)$ for all k and $\sum_{k=1}^{K} \pi_k = 1$, and $\mathcal{P}(\cdot)$ denotes the standard Poisson probability mass function.

The expectations of the Poisson distributions in Equation (1) are defined so that genes exhibiting similar expression patterns across samples are clustered together:

$$\mu_{ijlk} = w_i s_{jl} \lambda_{jk}$$

where $w_i = y_{i..}$ corresponds to the overall expression level of observation i (e.g., weakly to strongly expressed), and s_{jl} represents the normalized library size for replicate l of condition j, such that $\sum_{j,l} s_{jl} = 1$. We note that, similarly to test-based differential analysis methods such as *edgeR* and *DESeq2*, the normalization factors $\{s_{jl}\}_{j,l}$ are estimated from the data prior to fitting the model (for example, using the Trimmed Means of M-values method [4]) and are subsequently considered to be fixed. Finally, the unknown parameter vector $\boldsymbol{\lambda}_k = (\lambda_{1k}, \ldots, \lambda_{dk})$ corresponds to the clustering parameters that define the profiles of the genes in cluster k across all biological conditions (i.e., a measure of per-condition variability around the mean). The model parameters are estimated by an adapted Expectation-Maximization algorithm [5].

Based on an extensive simulation study and analyses of real data, a mixture of K=5 clusters was found to provide a satisfactory fit for most RNA-seq datasets. In *HTSDiff*, the first cluster is fixed to represent a set of non-differentially expressed genes by setting the value of $\boldsymbol{\lambda}_1 = (\lambda_{11}, \lambda_{21}) = 1$. After fitting the model, the remaining clusters are reorganized into two groups according to the absolute value of $\log_2(\lambda_{1k}/\lambda_{2k})$ for each cluster k; if this log-ratio is less than ϵ (fixed to 0.8 by default), the difference between the two conditions is judged to be sufficiently weak, and the cluster is considered to represent a group of non-differentially expressed genes. Finally, a gene is declared to be differentially expressed only if its conditional probability of non-differential expression is less than 10^{-8}.

2 Identifying differentially expressed genes with HTSDiff

2.1 Input data from Balzergue *et al.* (2014)

To illustrate the use of *HTSDiff* for the differential analysis of RNA-seq data in this vignette, we will work with the gene-level read counts from Balzergue *et al.* (2014) [1]. Briefly, in this study RNA-seq data were obtained from two biological replicates of both leaves and buds of wild-type Col0 *Arabidopsis thaliana*, which are known to have very different transcriptomic profiles. The same total RNA from each biological replicate was used, in parallel, for RNA-seq, microarray, and qRT-PCR analyses. The raw read counts and phenotypic labels for 28,094 genes are contained in a file called `counts.txt`.

We begin by loading the software packages and data, which are contained in the `initialDataset` object in the *HTSDiff* package (note that for this analysis, we will only make use of the original bud and leaf samples, contained in the columns labeled BF1, BF2, F1, and F2). In addition, we remove genes with 0 counts in all samples, leaving a total of 25,216 genes for the subsequent differential analysis.

```
> library(HTSDiff)
> data(initialDataset)
> y <- initialDataset[,c("BF1", "BF2", "F1", "F2")]
> ## Fix gene IDs as row names
> rownames(y) <- initialDataset[,1]
> head(y)
```

```
            BF1  BF2    F1   F2
AT1G01010   256  223   189  270
AT1G01020   465  393   428  333
AT1G01030   182  314    28   58
AT1G01040  1052 2193  2757 3145
AT1G01050  3872 2790  3461 2984
AT1G01060 11717  714 12108 1000
```

```
> conds <- c("BF","BF","F","F")
> y <- y[rowSums(y)>0,]
> dim(y)
```

```
[1] 25216     4
```

2.2 Inference

To perform a differential analysis using the *HTSDiff* package, we make use of the following code (which takes a few seconds to run):

```
> set.seed(12345)
> DEtest <- HTSDiff(counts=y, conds=conds)
```

By default, normalization for differences in library size is performed using the Trimmed Mean of M-values (TMM) procedure described in [4], although a variety of other methods (total count, upper quantile, median, and DESeq) are available.

The output of primary interest to most users of *HTSDiff* is the results data frame, contained in the object `res`:

```
> res <- DEtest$res
> head(res)
```

```
         id baseMean baseMeanA baseMeanB foldChange log2FoldChange tauDE tauNDE    DE
1 AT1G01010      238       211     264.6      1.255          0.328 3e-10  1e+00 FALSE
2 AT1G01020      409       377     440.9      1.168          0.225 3e-10  1e+00 FALSE
3 AT1G01030      136       222      49.4      0.223         -2.165 1e+00  2e-10  TRUE
4 AT1G01040     2432      1456    3408.5      2.341          1.227 1e+00  2e-10  TRUE
5 AT1G01050     3325      2919    3730.9      1.278          0.354 3e-10  1e+00 FALSE
6 AT1G01060     6475      5263    7687.7      1.461          0.547 3e-10  1e+00 FALSE
```

This object contains the gene IDs, base and condition-specific (normalized) means, fold-change, $\log_2$ fold-change, conditional probability of differential expression (`tauDE`), conditional probability of non-differential expression (`tauNDE`), and a binary vector (`DE`) indicating whether a gene is identified as differentially expressed (`TRUE`) or not. For these data, we find that a total of 8911 genes have been identified as differentially expressed using *HTSDiff*:

```
> table(res$DE)
```

```
FALSE  TRUE
16305  8911
```

In addition, the output of *HTSDiff* includes a variety of information regarding the fit of the underlying Poisson mixture model contained in the object `PMM`, including the estimates for $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$, the conditional probabilities of cluster membership for all genes and clusters (`probaPost`), the value of the log-likelihood and two penalized likelihood criteria (Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criterion), the type of algorithm used for parameter estimation, the number of iterations run in the EM algorithm, and the difference in log-likelihood values between the last and penultimate iterations of the algorithm:

```
> names(DEtest$PMM)

 [1] "lambda"          "pi"              "labels"          "probaPost"       "log.like"
 [6] "BIC"             "ICL"             "alg.type"        "lib.size"        "lib.type"
[11] "s"               "conds"           "iterations"      "logLikeDiff"     "model.selection"
```

The functionalities of the related *HTSCluster* package may be used to explore the PMM object in greater detail, including the summary and plot functions for the S3 class HTSCluster:

```
> summary(DEtest$PMM)

*************************************************
Number of clusters = 5
Model selection via NA
*************************************************
Cluster sizes:
Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5
     8751      4664      5206      3009      3586

Number of observations with MAP > 0.90 (% of total):
21316 (84.5%)

Number of observations with MAP > 0.90 per cluster (% of total per cluster):
 Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5
 7353      3474      4736      2549      3204
 (84.02%)  (74.49%)  (90.97%)  (84.71%)  (89.35%)

Lambda:
   Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5
BF         1      1.72      0.43      1.45      1.20
F          1      0.05      1.74      0.41      0.74

Pi:
Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5
     0.33      0.17      0.20      0.13      0.16
```

2.3 Comparison to other test-based methods

As a comparison, we also conduct a differential analysis with the test-based *edgeR* Bioconductor package [**?**], which makes use of a negative binomial model to test the per-gene null hypothesis of non-differential expression between conditions. The primary novelty of the *edgeR* package is the use of empirical Bayes methods to improve estimates of gene-specific biological variation, even when a small number of biological replicates are available.

The differential analysis with *edgeR* is run as follows:

```
> library(edgeR)
> y <- DGEList(counts=y, group=conds)
> y <- calcNormFactors(y)
> y <- estimateCommonDisp(y)
> y <- estimateTagwiseDisp(y)
> et <- exactTest(y)
> de <- decideTestsDGE(et, p=0.05, adjust="BH")
> summary(de)

   [,1]
-1  6340
0  15004
1   3872
```

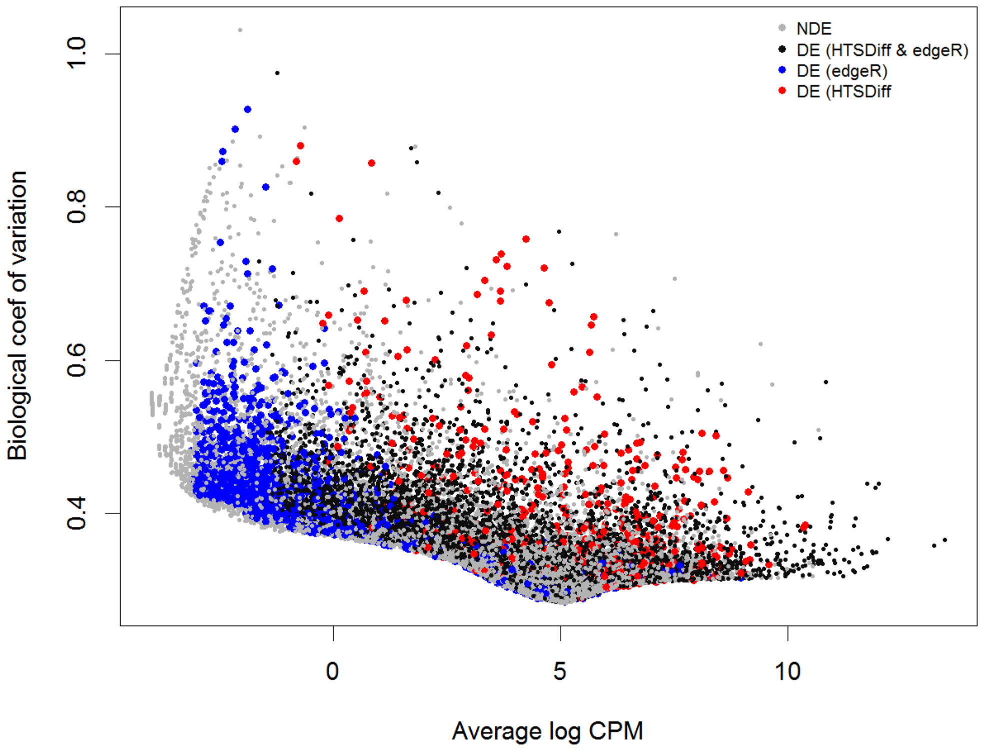
We note that a total of 10,212 genes (6340 underexpressed and 3872 overexpressed in buds with respect to leaves) are identified as significantly differentially expressed, after controlling the false discovery rate with the Benjamini-Hochberg method at 5%. We note that most of the genes identified as DE by *HTSDiff* (8352 out of 8911) are also identifed by *edgeR*, as shown in the table below.

```
> tab <- table(abs(de),res$DE)
> rownames(tab) <- c("NDE (edgeR)", "DE (edgeR)")
> colnames(tab) <- c("NDE (HTSDiff)", "DE (HTSDiff)")
> tab

              NDE (HTSDiff) DE (HTSDiff)
  NDE (edgeR)         14445          559
  DE (edgeR)           1860         8352
```

We also include a plot of the tagwise dispersions against the $\log_2$ counts per million (CPM) values; from this plot, it may be observed that genes identified as differentially expressed uniquely by *edgeR* tend to have fairly low biological coefficients of variation and small average CPM values, whereas those uniquely identified by *HTSDiff* tend to have larger biological coefficients of variation and average CPM values.

```
> A <- y$AveLogCPM
> disp <- getDispersion(y)
> colors <- ifelse(abs(de)==1 & res$DE==TRUE, "grey5", "grey70")
> colors <- ifelse(abs(de)==1 & res$DE==FALSE, "blue", colors)
> colors <- ifelse(abs(de)==0 & res$DE==TRUE, "red", colors)
> cex <- rep(0.4, length(abs(de)))
> cex <- ifelse(abs(de)==1 & res$DE==FALSE, 0.7, cex)
> cex <- ifelse(abs(de)==0 & res$DE==TRUE, 0.7, cex)
> plot(A, sqrt(disp), xlab="Average log CPM", ylab="Biological coef of variation",
+      type="n")
> points(A, sqrt(y$tagwise.dispersion), pch = 16, cex=cex, col=colors)
> legend("topright", c("NDE", "DE (HTSDiff & edgeR)", "DE (edgeR)", "DE (HTSDiff"),
+        col=c("grey70","grey5","blue", "red"), bty="n", pch=16, cex=0.7)
>
```



3 Session Info

```
> sessionInfo()

R version 3.1.1 (2014-07-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)

locale:
[1] LC_COLLATE=French_France.1252  LC_CTYPE=French_France.1252    LC_MONETARY=French_France.1252
[4] LC_NUMERIC=C                   LC_TIME=French_France.1252

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

other attached packages:
[1] edgeR_3.6.7      limma_3.20.8     HTSDiff_1.0.5    HTSCluster_2.0.5 capushe_1.0
[6] MASS_7.3-33      plotrix_3.5-7

loaded via a namespace (and not attached):
[1] poisson.glm.mix_1.2 tools_3.1.1
```

References

[1] S. Balzergue et al. HTSDiff: A model-based clustering alternative to test-based methods in differential gene expression analyses by RNA-seq benchmarked on real and synthetic datasets. *(submitted)*, 2014.

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

[3] M. D. Robinson et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinf.*, 26:139–140, 2010.

[4] A. Oshlack and M.J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(14), 2009.

[5] A. Rau et al. Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *(submitted)*, 2014.