**HTSFilter**: Data-based filtering for replicated transcriptome sequencing experiments

Andrea Rau, Mélina Gallopin, Gilles Celeux, Florence Jaffrézic

Modified: October 11, 2016. Compiled: April 24, 2017

**Abstract**

This vignette illustrates the use of the HTSFilter package to filter replicated data from transcriptome sequencing experiments (e.g., RNA sequencing data) for a variety of different data classes: matrix, data.frame, the S3 classes associated with the edgeR package (DGEExact and DGELRT), and the S4 class associated with the DESeq2 package (DESeqDataSet).

**Contents**

1 Introduction 1
2 Input data 3
3 matrix and data.frame classes 4
4 edgeR package pipeline 6
   4.1 S3 class DGEExact 6
   4.2 S3 class DGELRT 8
5 DESeq2 package pipeline: S4 class DESeqDataSet 9
6 Alternative normalization using EDAseq 10
7 Session Info 11

1 Introduction

High-throughput sequencing (HTS) data, such as RNA-sequencing (RNA-seq) data, are increasingly used to conduct differential analyses, in which statistical tests are performed for each biological feature (e.g., a gene, transcript, exon) in order to identify those whose expression levels show systematic covariation with a particular condition, such as a treatment or phenotype of interest. For the remainder.
of this vignette, we will focus on gene-level differential analyses, although these methods may also be applied to differential analyses of (count-based measures of) transcript- or exon-level expression.

Because hypothesis tests are performed for gene-by-gene differential analyses, the obtained \( p \)-values must be adjusted to correct for multiple testing. However, procedures to adjust \( p \)-values to control the number of detected false positives often lead to a loss of power to detect truly differentially expressed (DE) genes due to the large number of hypothesis tests performed. To reduce the impact of such procedures, independent data filters are often used to identify and remove genes that appear to generate an uninformative signal [1]; this in turn moderates the correction needed to adjust for multiple testing. For independent filtering methods for microarray data, see for example the genefilter Bioconductor package [2].

The HTSFilter package implements a novel data-based filtering procedure based on the calculation of a similarity index among biological replicates for read counts arising from replicated transcriptome sequencing (RNA-seq) data. This technique provides an intuitive data-driven way to filter high-throughput transcriptome sequencing data and to effectively remove genes with low, constant expression levels without incorrectly removing those that would otherwise have been identified as DE. The two fundamental assumptions of the filter implemented in the HTSFilter package are as follows:

1. Biological replicates are present for each experimental condition, and
2. Data can be appropriately normalized (scaled) to correct for systematic inter-sample biases.

Assuming these conditions hold, HTSFilter implements a method to identify a filtering threshold that maximizes the filtering similarity among replicates, that is, one where most genes tend to either have normalized counts less than or equal to the cutoff in all samples (i.e., filtered genes) or greater than the cutoff in all samples (i.e., non-filtered genes). This filtering similarity is defined using the global Jaccard index, that is, the average Jaccard index calculated between pairs of replicates within each experimental condition; see Rau et al. (2013) [3] for more details.

For more information about between-sample normalization strategies, see [4]; in particular, strategies for normalizing data with differences in library size and composition may be found in [5] and [6], and strategies for normalizing data exhibiting sample-specific biases due to GC content may be found in [7] and [8]. Within the HTSFilter package, the Trimmed Means of M-values (TMM) [6] and DESeq [5] normalization strategies may be used prior to calculating an appropriate data-based filter. If an alternative normalization strategy is needed or desired, the normalization may be applied prior to filtering the data with `normalization="none"` in the HTSFilter function; see Section 6 for an example.

The HTSFilter package is able to accommodate unnormalized or normalized replicated count data in the form of a matrix or data.frame (in which each row corresponds to a biological feature and each column to a biological sample), one of the S3 classes associated with the edgeR package (DGEList, DGEEExact, DGEGLM, and DGEGLRT), or DESeqDataSet (the S4 class associated with the DESeq2 package), as illustrated in the following sections.

Finally, we note that the filtering method implemented in the HTSFilter package is designed to filter transcriptome sequencing, and not microarray, data; in particular, the proposed filter is effective for data with features that take on values over a large order of magnitude and with a subset of features exhibiting small levels of expression across samples (see, for example, Figure 1). In this vignette, we illustrate its use on count-based measures of gene expression, although its use is not strictly limited to
Figure 1: Histogram of log transformed counts from the Sultan et al. data [9], illustrating the large number of genes with very small counts as well as the large heterogeneity in counts observed.

discrete data.

2 Input data

For the purposes of this vignette, we make use of data from a study of sex-specific expression of liver cells in human and the DESeq and edgeR packages for differential analysis. Sultan et al. [9] obtained a high-throughput sequencing data (using a 1G Illumina Genome Analyzer sequencing machine) from a human embryonic kidney and a B cell line, with two biological replicates each. The raw read counts and phenotype tables were obtained from the ReCount online resource [10].

To begin, we load the HTSFilter package, and attach the gene-level count data contained in sultan:

```r
> library(HTSFilter)
> library(edgeR)
> library(DESeq2)
> data("sultan")
> hist(log(exprs(sultan)+1), col="grey", breaks=25, main="",
```
The unfiltered data contain 9010 genes in four samples (two replicates per condition).

### 3 matrix and data.frame classes

To filter high-throughput sequencing data in the form of a *matrix* or *data.frame*, we first access the expression data, contained in `exprs(sultan)`, and create a vector identifying the condition labels for each of the samples via the `pData` Biobase function. We then filter the data using the `HTSFilter` function, specifying that the number of tested thresholds be only 25 (`s.len=25`) rather than the default value of 100 to reduce computation time for this example. Note that as it is unspecified, the default normalization method is used for filtering the data, namely the Trimmed Mean of M-values (TMM) method of Robinson and Oshlack [6]. To use the DESeq normalization method [5], `normalization="DESeq"` may be specified.

```r
> mat <- exprs(sultan)
> conds <- as.character(pData(sultan)$cell.line)
> ## Only 25 tested thresholds to reduce computation time
> filter <- HTSFilter(mat, conds, s.min=1, s.max=200, s.len=25)
> mat <- filter$filteredData
> dim(mat)
[1] 4995  4

> dim(filter$removedData)
[1] 4015  4
```

For this example, we find a data-based threshold equal to 11.764; genes with normalized values less than this threshold in all samples are filtered from subsequent analyses. The proposed filter thus removes 4015 genes from further analyses, leaving 4995 genes.

We note that an important part of the filter proposed in the `HTSFilter` package is a check of the behavior of the global similarity index calculated over a range of threshold values, and in particular, to verify that a reasonable maximum value is reached for the global similarity index over the range of tested threshold values (see Figure 2); the maximum possible value for the global Jaccard index is nearly 1.
HTSFilter: Data-based filtering for replicated transcriptome sequencing experiments

Figure 2: Global Jaccard index for the sultan data calculated for a variety of threshold values after TMM normalization \cite{6}, with a loess curve (blue line) superposed and data-based threshold values (red cross and red dotted line) equal to 11.764.

To illustrate the importance of this check, we attempt to re-apply the proposed filter to the previously filtered data (in practice, of course, this would be nonsensical):

```r
> par(mfrow = c(1,2), mar = c(4,4,2,2))
> filter.2 <- HTSFilter(mat, conds, s.len=25)
> dim(filter.2$removedData)
[1] 0 4
> hist(log(filter.2$filteredData+1), col="grey", breaks=25, main="", +   xlab="Log(counts+1)"
```

In the lefthand panel of Figure 3, we note a plateau of large global Jaccard index values for thresholds less than 2, with a decrease thereafter; this corresponds to filtering no genes, unsurprising given that genes with low, constant levels of expression have already been filtered from the analysis (see the righthand panel of Figure 3).
HTSFilter: Data-based filtering for replicated transcriptome sequencing experiments

Figure 3: (left) Global Jaccard index for the sultan data calculated for a variety of threshold values after TMM normalization [6], with a loess curve (blue line) superposed and data-based threshold values (red cross and red dotted line) equal to 11.764. (right) Global Jaccard index for the previously filtered sultan data, with loess curve (blue line) superposed as before.

4 edgeR package pipeline

We next illustrate the use of HTSFilter within the edgeR pipeline for differential analysis (S3 classes DGEList, DGEExact, DGEGLM, or DGELRT). For the purposes of this vignette, we will consider the S3 classes DGEExact and DGELRT. The former is the class containing the results of the differential expression analysis between two groups of count libraries (resulting from a call to the function exactTest in edgeR); the latter is the class containing the results of a generalized linear model (GLM)-based differential analysis (resulting from a call to the function glmLRT in edgeR). Although the filter may be applied earlier in the edgeR pipeline (i.e., to objects of class DGEList or DGEGLM), we do not recommend doing so, as parameter estimation makes use of counts adjusted using a quantile-to-quantile method (pseudo-counts).

4.1 S3 class DGEExact

We first coerce the data into the appropriate class with the function DGEList, where the group variable is set to contain a vector of condition labels for each of the samples. Next, after calculating normalizing factors to scale library sizes (calcNormFactors), we estimate common and tagwise dispersion parameters using estimateDisp (using the quantile conditional likelihood for each gene) and obtain differential analysis results using exactTest. Finally, we apply the filter using the HTSFilter function, again specifying that the number of tested thresholds be only 25 (s.len=25) rather than the default value of 100. Note that as it is unspecified, the default normalization method is used for filtering the
data, namely the Trimmed Mean of M-values (TMM) method [6]; alternative normalization, including "pseudo.counts" for the quantile-to-quantile adjusted counts used for parameter estimation, may also be specified. We suppress the plot of the global Jaccard index using plot = FALSE, as it is identical to that shown in Figure 2.

```r
> dge <- DGEList(counts=exprs(sultan), group=conds)
> dge <- calcNormFactors(dge)
> dge <- estimateDisp(dge)
Design matrix not provided. Switch to the classic mode.
> et <- exactTest(dge)
> et <- HTSFilter(et, DGEList=dge, s.len=25, plot=FALSE)$filteredData
> dim(et)
[1] 4995  3
> class(et)
[1] "DGEExact"
attr(,"package")
[1] "edgeR"
> topTags(et)
Comparison of groups: Ramos B cell-HEK293T
  logFC   logCPM     PValue
ENSG00000100721  14.399287  11.53328  0.000000e+00
ENSG00000133124 -14.394458  11.56789  0.000000e+00
ENSG00000105369  11.925779  11.23367  0.000000e+00
ENSG00000135144  10.921892  11.05896  2.600818e-309
ENSG00000110777  13.850554  10.97744  2.014098e-302
ENSG00000111348  13.797192  10.92306  5.226492e-295
ENSG00000118308  13.494132  10.61473  9.238774e-270
ENSG00000177606 -0.9731885  10.81415  9.092173e-263
ENSG000001012124 10.171666  10.29617  3.995433e-247
ENSG00000149418  10.899611  10.19056  6.189743e-239
  FDR
ENSG00000100721  0.000000e+00
ENSG00000133124  0.000000e+00
ENSG00000105369  0.000000e+00
ENSG00000135144  3.247771e-306
ENSG00000110777  2.012083e-302
ENSG00000111348  5.226492e-295
ENSG00000118308  9.238774e-270
ENSG00000177606  9.092173e-261
ENSG000001012124 2.217465e-244
ENSG00000149418  3.091777e-236
```
Note that the filtered data are of the class \texttt{DGEExact}, allowing for a call to the \texttt{topTags} function.

\begin{verbatim}
> topTags(et)

Comparison of groups: Ramos B cell-HEK293T

\begin{tabular}{llll}
  & logFC & logCPM & PValue \\
ENSG00000100721 & 14.399287 & 11.53328 & 0.000000e+00 \\
enSG00000133124 & -14.394458 & 11.56789 & 0.000000e+00 \\
ENSG00000105369 & 11.925779 & 11.23367 & 0.000000e+00 \\
ENSG00000135144 & 10.921892 & 11.05896 & 2.600818e-309 \\
ENSG00000110777 & 13.850554 & 10.97744 & 2.014098e-302 \\
ENSG00000111348 & 13.797192 & 10.92306 & 6.278069e-298 \\
ENSG00000118308 & 13.494132 & 10.61473 & 1.294723e-272 \\
ENSG00000177606 & -9.731885 & 10.81415 & 1.456204e-263 \\
ENSG0000012124 & 10.171666 & 10.29617 & 3.995433e-247 \\
ENSG00000149418 & 10.899611 & 10.19056 & 6.189743e-239 \\

FDR \\
ENSG00000100721 & 0.000000e+00 \\
enSG00000133124 & 0.000000e+00 \\
ENSG00000105369 & 0.000000e+00 \\
ENSG00000135144 & 3.247771e-306 \\
ENSG00000110777 & 2.012083e-299 \\
ENSG00000111348 & 5.226492e-295 \\
ENSG00000118308 & 9.238774e-270 \\
ENSG00000177606 & 9.092173e-261 \\
ENSG0000012124 & 2.217465e-244 \\
ENSG00000149418 & 3.091777e-236 \\
\end{tabular}
\end{verbatim}

\subsection{S3 class \texttt{DGELRT}}

We follow the same steps as the previous example, where the \texttt{estimateDisp} function is now used to obtain per-gene dispersion parameter estimates using the adjusted profile loglikelihood, the \texttt{glmFit} function is used to fit a negative binomial generalized log-linear model to the read counts for each gene, and the \texttt{glmLRT} function is used to conduct likelihood ratio tests for one or more coefficients in the GLM. The output of \texttt{glmLRT} is an S3 object of class \texttt{DGELRT} and contains the GLM differential analysis results. As before, we apply the filter using the \texttt{HTSFilter} function, again suppressing the plot of the global Jaccard index using \texttt{plot = FALSE}, as it is identical to that shown in Figure 2.

\begin{verbatim}
> design <- model.matrix(~conds)
> dge <- DGEList(counts=exprs(sultan), group=conds)
> dge <- calcNormFactors(dge)
> dge <- estimateDisp(dge, design)
> fit <- glmFit(dge,design)
> lrt <- glmLRT(fit,coef=2)
> lrt <- HTSFilter(lrt, DGEGLM=fit, s.len=25, plot=FALSE)$filteredData
\end{verbatim}
\textit{HTSFilter}: Data-based filtering for replicated transcriptome sequencing experiments

\begin{verbatim}
> dim(lrt)
[1] 4995  4
> class(lrt)
[1] "DGEList"
attr(,"package")
[1] "edgeR"
\end{verbatim}

Note that the filtered data are of the class \textit{DGEList}, allowing for a call to the \textit{topTags} function.

\begin{verbatim}
> topTags(lrt)

Coefficient: condsRamos B cell

\begin{tabular}{llllll}
  & logFC & logCPM & LR & PValue & FDR \\
ENSG00000100721 &  14.39 &   11.53 & 1711.21 & 0.000000e+00 & 0.000000e+00 \\
ENSG00000133124 & -14.39 &   11.56 & 1554.76 & 0.000000e+00 & 0.000000e+00 \\
ENSG00000110777 &  13.85 &   10.97 & 1487.28 & 0.000000e+00 & 0.000000e+00 \\
ENSG00000105369 &  11.92 &   11.23 & 1576.31 & 0.000000e+00 & 0.000000e+00 \\
ENSG00000135144 &  10.92 &   11.06 & 1495.59 & 0.000000e+00 & 0.000000e+00 \\
ENSG00000111348 &  13.79 &   10.92 & 1465.78 & 1.067182e-320 & 8.884288e-318 \\
ENSG00000118308 &  13.49 &   10.61 & 1345.24 & 1.666173e-294 & 1.188933e-291 \\
ENSG00000177606 &  -9.73 &   10.82 &  1222.98 & 6.192499e-268 & 3.864288e-265 \\
ENSG0000012124 &  10.17 &   10.29 &  1199.88 & 6.468477e-263 & 5.570841e-260 \\
ENSG00000149418 &  10.89 &   10.19 &  1171.19 & 1.152845e-256 & 5.708411e-254 \\
\end{tabular}
\end{verbatim}

\section{DESeq2 package pipeline: S4 class \texttt{DESeqDataSet}}

The \textit{HTSFilter} package allows for a straightforward integration within the \textit{DESeq2} analysis pipeline, most notably allowing for $p$-values to be adjusted only for those genes passing the filter. Note that \textit{DESeq2} now implements an independent filtering procedure by default in the \textit{results} function; this filter is a potential alternative filtering technique and does not need to be used in addition to the one included in \textit{HTSFilter}. In fact, each filter is targeting the same weakly expressed genes to be filtered from the analysis. As such, if the user wishes to make use of \textit{HTSFilter} within the \textit{DESeq2} pipeline,
the argument independentFiltering=FALSE should be used when calling the results function in DESeq2.

To illustrate the application of a filter for high-throughput sequencing data in the form of a DESeqDataSet (the class used within the DESeq2 pipeline for differential analysis), we coerce sultan into an object of the class DESeqDataSet using the function DESeqDataSetFromMatrix. Once again, we specify that the number of tested thresholds be only 25 (s.len=25) rather than the default value of 100 to reduce computation time. For objects in the form of a DESeqDataSet, the default normalization strategy is "DESeq", although alternative normalization strategies may also be used.

```r
> dds <- DESeqDataSetFromMatrix(countData = exprs(sultan),
+    colData = data.frame(cell.line = conds),
+    design = ~ cell.line)
> dds <- DESeq(dds)
> filter <- HTSFilter(dds, s.len=25, plot=FALSE)$filteredData
> class(filter)
> dim(filter)
> res <- results(filter, independentFiltering=FALSE)
> head(res)
```

The filtered data remain an object of class DESeqDataSet, and subsequent functions from DESeq2 (such as the results summary function results) may be called directly upon it.

### 6 Alternative normalization using EDAseq

As a final example, we illustrate the use of the HTSFilter package with an alternative normalization strategy, namely the full quantile normalization method in the EDASeq package; such a step may be useful when the TMM or DESeq normalization methods are not appropriate for a given dataset. Once again, we create a new object of the appropriate class with the function newSeqExpressionSet and normalize data using the betweenLaneNormalization function (with which="full") in EDASeq.

```r
> library(EDASeq)
> ses <- newSeqExpressionSet(exprs(sultan),
+    phenoData=pData(sultan))
> ses.norm <- betweenLaneNormalization(ses, which="full")
```

Subsequently, HTSFilter is applied to the normalized data (again using s.len=25), and the normalization method is set to norm="none". We may then make use of the on vector in the results, which identifies filtered and unfiltered genes (respectively) with 0 and 1, to identify rows in the original data matrix to be retained.

```r
> filter <- HTSFilter(counts(ses.norm), conds, s.len=25, norm="none",
+    plot=FALSE)
> head(filter$on)
> table(filter$on)
```
7 Session Info

> sessionInfo()

R version 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.2 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so

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

attached base packages:
[1] parallel stats4 stats graphics grDevices
[6] utils datasets methods base

other attached packages:
[1] DESeq2_1.16.0 SummarizedExperiment_1.6.0
[3] DelayedArray_0.2.0 matrixStats_0.52.2
[5] Biobase_2.36.0 GenomicRanges_1.28.0
[7] GenomeInfoDb_1.12.0 IRanges_2.10.0
[9] S4Vectors_0.14.0 BiocGenerics_0.22.0
[11] edgeR_3.18.0 limma_3.32.0
[13] HTSFilter_1.16.0

loaded via a namespace (and not attached):
[1] Rcpp_0.12.10 locfit_1.5-9.1
[3] lattice_0.20-35 rprojroot_1.2
[5] digest_0.6.12 plyr_1.8.4
[7] backports_1.0.5 acepack_1.4.1
[9] RSQLite_1.1-2 DESeq_1.28.0
[11] evaluate_0.10 ggplot2_2.2.1
[13] zlibbioc_1.22.0 lazyeval_0.2.0
[15] data.table_1.10.4 annotate_1.54.0
[17] rpart_4.1-11 Matrix_1.2-9
[19] checkmate_1.8.2 rmarkdown_1.4
[21] splines_3.4.0 BiocParallel_1.10.0
References


