HTSFilter: An independent data-based filter for replicated high-throughput transcriptome sequencing experiments

A. Rau[a], **M. Gallopin**[a], **G. Celeux**[b] and **F. Jaffrézic**[a]

[a]UMR 1313 GABI
INRA
Jouy-en-Josas, France 78352
{andrea.rau, melina.gallopin, florence.jaffrezic}@jouy.inra.fr

[b]Inria Saclay – Île-de-France
Orsay, France 91405
gilles.celeux@math.u-psud.fr

Mots clefs : Independent filter, gene expression, RNA-seq, differential analysis.

Over the past five years, next-generation high-throughput sequencing (HTS) technology has become an essential tool for genomic and transcriptomic studies. In particular, the use of HTS technology to directly sequence the transcriptome, known as RNA sequencing (RNA-seq), has revolutionized the study of gene expression by opening the door to a wide range of novel applications. Unlike microarray data, RNA-seq data represent highly heterogeneous counts for genomic regions of interest (typically genes), and often exhibit zero-inflation and a large amount of overdispersion among biological replicates. As such, a great deal of methodological research has recently focused on appropriate normalization and analysis techniques that are adapted to the characteristics of RNA-seq data, particularly for the study of differential expression among experimental conditions.

Because a large number of hypothesis tests (typically in the thousands or tens of thousands) are performed for gene-by-gene differential analyses, stringent false discovery rate control is required at the expense of the power of an experiment to detect truly differentially expressed (DE) genes. To reduce this impact, data filters are often used in order to identify and remove genes which appear to generate an uninformative signal and have little chance of showing significant evidence of differential expression; only hypotheses corresponding to genes that pass the filter are subsequently tested, which in turn tempers the correction needed to adjust for multiple testing. However, in practice an arbitrary filtering threshold is typically fixed for RNA-seq data without accounting for the overall sequencing depth or variability of a given experiment, and little attention is paid to its impact on the downstream analysis.

In this work, we propose a Bioconductor package, `HTSFilter`, that implements a data-driven method to identify an appropriate filtering threshold for replicated RNA-seq data [1]. The main idea underlying this method is to identify the threshold that maximizes the filtering similarity among biological 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). More precisely, we denote the observed read counts for all genes in sample j as $\mathbf{y}_j = (y_{gj})$, where $\mathcal{C}(j)$ is the experimental condition of sample j. After binarizing the data for a fixed filtering threshold s (1 if $y_{gj} > s$ and 0 otherwise), the Jaccard

		Sample j: Normalized counts $> s$	Sample j: Normalized counts $\leq s$
Sample j': Normalized counts $> s$		a	b
Sample j': Normalized counts $\leq s$		c	d

Table 1: Constants used to calculate the Jaccard index defined in Equation (1).

similarity between two biological replicates may be defined as follows:

$$J_s(\mathbf{y}_j, \mathbf{y}_{j'}) = \frac{a}{a+b+c}, \tag{1}$$

where a, b, and c are defined in Table 1. Because multiple replicates and conditions are typically available in HTS experiments, we extend the definition of the pairwise Jaccard index in (1) to a global Jaccard index by averaging the indices calculated over all pairs in each condition:

$$J_s^\star(\mathbf{y}) = \text{mean}\left\{J_s(\mathbf{y}_j, \mathbf{y}_{j'}) : j < j' \text{ and } \mathcal{C}(j) = \mathcal{C}(j')\right\}. \tag{2}$$

Finally, we identify the threshold $s^\star$ that yields the largest possible global Jaccard index:

$$s^\star = \underset{s}{\text{argmax}}\, J_s^\star(\mathbf{y}).$$

In practice, in the `HTSFilter` package we calculate the value of the global Jaccard index in (2) for a fixed set of threshold values and fit a loess curve through the set of points; the value of $s^\star$ is subsequently set to be the maximum of these fitted values.

In comparisons with alternative data filters regularly used in practice, we have demonstrated the effectiveness of our proposed method to correctly filter weakly expressed genes, leading to increased detection power for moderately to highly expressed genes. Interestingly, this data-driven threshold varies among experiments, highlighting the interest of the method proposed here. The `HTSFilter` package is compatible with a variety of data classes and analysis pipelines that have been proposed for RNA-seq data, including `matrix` and `data.frame` objects, the S4 class `CountDataSet` in the `DESeq` pipeline [2], and the S3 class `DGEList` in the `edgeR` pipeline [3]. A package vignette distributed with the `HTSFilter` package describes the use of the filtering method within each of these pipelines.

Références

[1] Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013) Independent data-based filtering for replicated high-throughput transcriptome sequencing experiments (submitted).
[2] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, 11(R106):1-28.
[3] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139-140.