PathwaySplice testing for RNA-seq datasets

Aimin Yan

2017-08-07

1 Introduction

This document gives an introduction to the use of the PathwaySplice R Bioconductor package. This package provides methods for (1) performing pathway analysis that ex-plicitly adjusts for the number of exons or junctions associated with each gene; (2) visualizing selection bias due to different number of exons or junctions for each gene and formally tests for presence of bias using logistic regression; (3) supporting gene sets based on the Gene Ontology terms, as well as more broadly defined gene sets (e.g. MSigDB) or user defined gene sets; (4) identifing the significant genes driving pathway significance and (5) organizing significant pathways with an enrichment map, where pathways with large number of overlapping genes are grouped together in a network graph.

Once installed, the PathwaySplice package can be easily loaded into R using:

```
library(PathwaySplice)
```

2 Use PathwaySplice

PathwaySplice uses the output from DEXSeq(Anders, Reyes, and Huber 2012) or JunctionSeq(Hartley and Mullikin 2016) as input basis. Assume that users have got the differnetial exon usage or differential splicing junctions usage results(res) from DEXSeq or JunctionSeq. Users need to use the internal function makeFeatureTable in PathwaySplice to convert res into a gene feature based table firstly:

```
featureBasedData <- PathwaySplice:::makeFeatureTable(res)
```

After users get genefeature based table(featureBasedData), users can use makeGeneTable function to convert it to a gene based table:
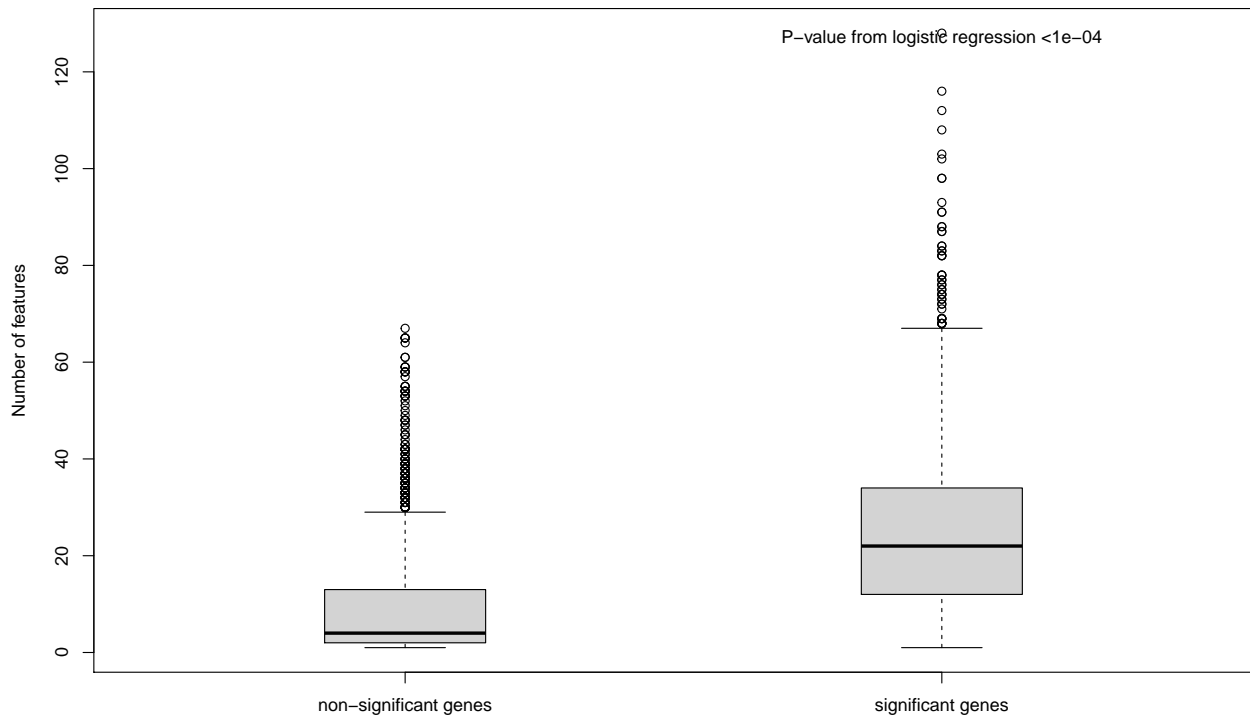
```
gene.based.table <- makeGeneTable(featureBasedData)
```

For a showcase, PathwaySplice supply a feature based data with this package, users can load this data directly:

```
data("featureBasedData")
gene.based.table <- makeGeneTable(featureBasedData)
```
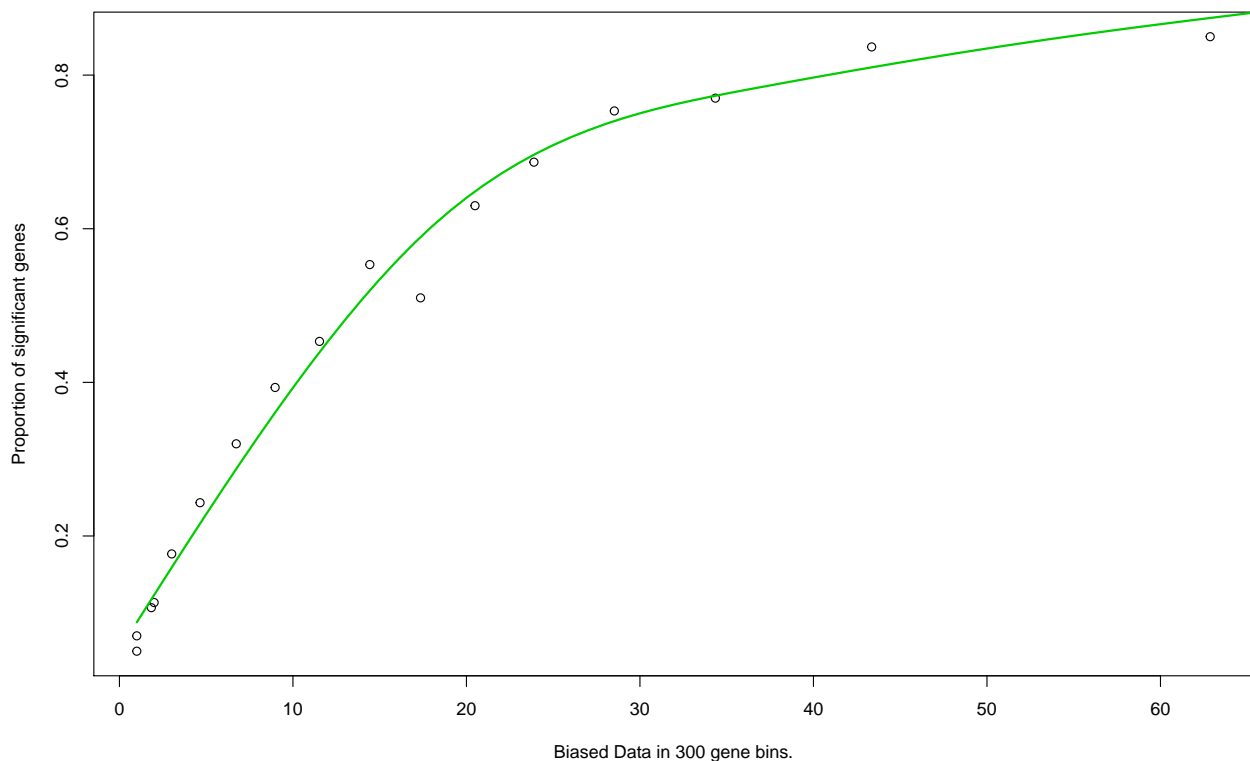
After this step, users can identify bias factor by the function lrTestBias

```
lrTestBias(gene.based.table,boxplot.width=0.3)
```

To perform analysis by adjusting number of feature, users can perform the following step:

```
res.adj <- runPathwaySplice(gene.based.table,genome='hg19',
                            id='ensGene',
                            test.cats=c('GO:BP'),
                            go.size.limit=c(5,30),method='Wallenius')
```


User can compare the output after adjutstment and the one before adjustment:

```
res.unadj <- runPathwaySplice(gene.based.table,genome='hg19',
                     id='ensGene',test.cats =c('GO:BP'),
                     go.size.limit = c(5, 30),
                     method='Hypergeometric')

compareResults(20, res.adj, res.unadj,
               gene.based.table,
               type.boxplot='Only3',
               output.dir="~/PathwaySplice_output")
```

Users can build up the enrichment network from the output of runPathwaySplice function by the following commands:

```
output.file.dir <- file.path("~/PathwaySplice_output","OutputEnmap")
enmap <- enrichmentMap(res.adj,n=3,
                     output.file.dir=output.file.dir,
                     similarity.threshold=0)
```

In the enrichment map, the size of the nodes indicates the number of significant genes within the pathway, which is also indicated by the number after ":". The color of the nodes indicates pathway significance, where smaller p-values correspond to dark red color. The thickness of the edges corresponds to JC similarity coefficient between the two pathways.

User can find several output files and a directory named as "OutputEnmap" under "~/PathwaySplice_output" directory. These files include:

"boxplot.html": To show the distributions of gene features associated with genes in significant gene sets before and after bias "adjustment".

"adjusted_unadjusted_overlap_venn.tiff": To visualize the overalp of genes in significant pathways before and after adjustment.

"adjustedOnly.csv": To list gene sets only in adjusted results.

"unadjustedOnly.csv": To list gene sets only in unadjusted results.

"results4venn.csv": To list 3 columns that corresponds to gene set names in adjustedOnly, unadjustedOnly and common area in venn plot.

In "OutputEnmap" directory, users can find "network.layout.for.cytoscape.gml", this file can be used as an input in cytoscape software(Shannon et al. 2003) to let users manipulate the networks.

3   Conclusion

In this tutorial, we go through basic steps in PathwaySplicing to process input from the outputs from DEXSeq or JunctionSeq, and to identify bias, and to perform gene set and pathway analysis, and to make enrichment network. Users can follow these steps to apply this Package on their RNA-Seq data as well.

4   Reference

Anders, Simon, Alejandro Reyes, and Wolfgang Huber. 2012. "Detecting differential usage of exons from RNA-seq data." *Genome Research* 22 (10): 2008–17. doi:10.1101/gr.133744.111.

Hartley, Stephen W., and James C. Mullikin. 2016. "Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq." *Nucleic Acids Research*, June. Oxford University Press, gkw501.

doi:10.1093/nar/gkw501.

Shannon, P., Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504. doi:10.1101/gr.1239303.