PathwaySplice testing for RNA-seq datasets

Aimin Yan

2017-09-19

1 Introduction

In alternative splicing ananlysis of RNASeq data, one popular approach is to first identify gene features (e.g. exons or junctions) significantly associated with splicing using methods such as DEXSeq (ref) or JunctionSeq (ref), and then perform pathway analysis based on the list of genes associated with the significant gene features.

A major challenge is that without explicit adjustment, pathways analysis would be biased toward categories that include genes with a large number of gene features, because these genes are more likely to be selected as "significant genes" in pathway analysis.

PathwaySplice is an R package that provides methods for (1) performing pathway analysis that explicitly 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.

2 Using PathwaySplice

- After installation, the PathwaySplice package can be loaded into R using:

```
library(PathwaySplice)
```

The input file of PathwaySplice are p-values for multiple gene features associated with each gene. This information can be obtained from DEXSeq(Anders, Reyes, and Huber 2012) or JunctionSeq(Hartley and Mullikin 2016) output files. As an example, PathwaySplice include a feature based dataset within the package, users can load this data directly:

```
data("featureBasedData")
head (featureBasedData)
```

```
       geneID countbinID     pvalue
```

1 ENSG00000279928 E003 0.19792636 2 ENSG00000279457 E002 0.55363528 3 ENSG00000279457 E003 0.12308986 4 ENSG00000279457 E004 0.11887268 5 ENSG00000279457 E005 0.03981720 6 ENSG00000279457 E006 0.07570489

Next the makeGeneTable function can be used to convert it to a gene based table:

```
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")
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 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.

"network.layout.for.cytoscape.gml": The 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.