# GOSJ manual

*Aimin Yan*

*May 13, 2016*

## Motivation:

Differential exons or splicing junctions usage analysis is very helpful for detecting alternative splicing events.

## Results:

We demonstrate this bias using published data from a study of differential splicing junction usage in Myelodysplastic syndromes(MDS) and a data set that we generated in a study of differential exons and splicing junctions usage in blood cancer. We show that the identified differential expressed genes through using differential usage of subfeatures(exons and/or splicing junctions) are biased by the number of subfeatures within gene, and subsequently lead to the bias in the subsequent Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) enrichment analysis. We suggest using an existing approach to adjust this bias. A R package(GOSJ) based on this approach is also developed to make this method be useful for other researchers. This GOSJ package is available on https://github.com/aiminy/GOSJ.git.

## Introduction

## RESULTS AND DISCUSSIONS

## Simulation studies based on a real data

We attemp to use simulation to establish the relationship between the proportion of DE and the number of subfeatures of genes. In this simulation setting, we sample the mean of the number of splicing junction, and use this mean to simulate an array A including splicing junction numbers with certain amounts. we convert this number into an array AAA between 0 and 1. Use the median of AAA array as the probability for being 1, we sample from [0,1] to get a list with certain length. This list is used as a gene set. For example, if this gene set has 30 genes, each gene has its splicing junction number, we will have 30 splicing junction number. After differential gene expression analysis, some genes are differentiallly expressed, and are labeled as 1,otherwise labeled as 0. Here we simulate two scenarios:

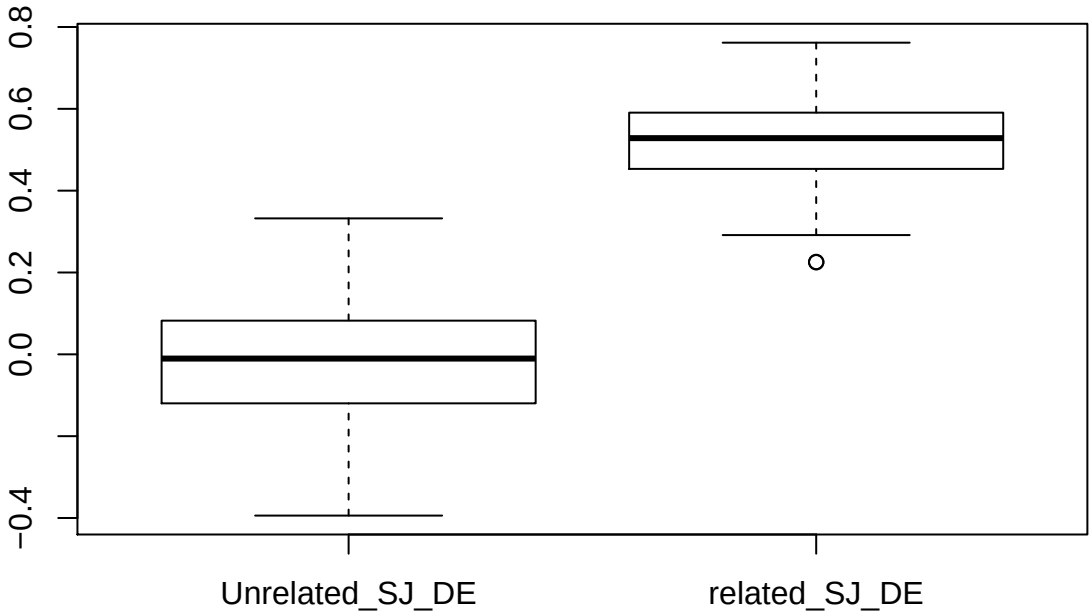
- Scenario1: the probability of being 1 is dependent on the median of AAA
- Scenario2: the probability of being 1 is not dependent on the median of AAA.

In each simulation,we generate 50 gene sets,and each gene set includes 30 genes. For each set, we calculate the probability of differentially expressed genes and the median of splicing junction numbers of genes. Here we use $i$ to indicat the number of gene sets in each simulation, so we can get:

The probabilities of differentially expressed genes of 50 gene sets: $P_1$,$P_2$,...,$P_i$,...,$P_{50}$

The median of splicing junctions of 50 gene sets: $F_1$,$F_2$,...,$F_i$,...,$F_{50}$

Based on this, the correlation coefficients between probabilites of differentially expressed genes and medians of splicing junction numbers of genes in each simulation can be obtained. We performed 100 simumations, correltion coefficients of 100 simulations are calcuated. The following plot shows the distribution of correlation coefficients between the probability of being 1 and the median number of splicing junctions in two scenarios for 100 simulations:




From the figure above, it is obvious: In Scenario1(Unrelated_SJ_DE), a gene is differentially expressed or not is not dependent on the number of splicing junctions, so the average correlation coefficients between the probability of being 1 and the median number of splicing junctions in this scenario is closed to 0. In Scenario2(related_SJ_DE), a gene is differentially expressed or not is dependent on the number of splicing junctions, so the average correlation coefficients between the probability of being 1 and the median number of splicing junctions in this scenario is far away from 0.

## Generate random data set based on the data set from blood cancer.

We have 13000 gene, each gene has certain number of subfeatures, in this smulation setting, we fix the number of of subfeatures for each gene as the real data, and perform the following permutations:

first , we permutate sample label

second, we permutate maping count

By this way, we attemp to identify the same enriched gene sets and pathways of GO and KEGG. If there are high consistencies between the resutls from ramdom data sets and that of real data sets, this indicate that the number of subfeatures of genes has cause bias on gene set enrichment analysis.

## Analysis on several real data sets

1. Reanalysis for a data set in Myelodysplastic syndromes(MDS) study
2. Analsysis for a data aset we obtained in blood cancer study

A data set including 6 samples is used for developing this approach. This 6 samples belong to two conditions, and there are 3 samples in each condition. The reads in each sample is aligned to the mouse mm10 reference genome using Tophat. Gene models is based on UCSC mm10 refSeq. We performed analysis based on 2 scenarios:

In the first scenario, we use gene-based counts to identify differentially expressed genes, then using this differnentially expressed gene list to identify the relationship between the proportion of DE and the possible bias factos such as gene length, number of exons and number splicing junctions.

In the second scenario, we apply JunctionSeq to subfeature-based counts to determine the $p$ value of differential usage of each subfeature of each gene, then calculate the genewise $p$ value based on the following method:

let $p_{il}$ be the $p$ value for $l$ subfeature(exon or splicing junction) of gene $i$, then gene wise $p$ value is calcualted using the following method:

$$P(at\ least\ 1\ type\ I\ error\ among\ m\ tests) = \frac{\sum_{i=1}^{M} 1-(1-\theta)^{n_i}}{|i : \exists p_{il} < \theta|}$$

Based on these genewsie p values to define the differentially expressed genes, then using this differnentially expressed genes list to identify the relationship between the propotion of DE and the possible bias factos such as gene length, number of exons and number splicing junctions. In this step, Ensemble-annotated exons and junction-spanning read counts are calculated, and a negative binomial model is applied to these counts to identify differential exon and junction usage while the overall gene expression changes are controlled.