# GOSJ: An R package for adjusting bias from gene structure on identifying differentially expressed genes in RNA-Seq data analysis

*Aimin Yan*

*May 13, 2016*

## Motivation:

Differential exons or splicing junctions usage analysis is very helpful for detecting alternative splicing events. There are several studies that researchers attempt to undertand alternative splicing events by performing differential exons or/and splicing junctions usage analysis. In these analysis, the differentially expressed gene are derived based on differential exon or/and splicing junction usage analysis, then subsequent gene set or/and pathway analysis are performed based on differentlly expressed gene list. However, it is observed that the derived differentially expressed gene list from differential exon or/and splicing junction analysis is biased by the number of exons or/ and splicing junctions within gene, and cause further bias in subsequent get set and pathway analysis.

## 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 obtained in a study of differential exons and splicing junctions usage in blood cancer. We show that the identified differentially expressed genes through using differential usage of gene features(exons and/or splicing junctions) are biased by the number of features within gene, and subsequently lead to the bias in following Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) enrichment analysis. We develop an approach to adjust this bias. A R package(GOSJ) based on this approach is also implemented to make this method be available for other researchers. This GOSJ package is accessable on https://github.com/aiminy/GOSJ.git.

## Introduction

RNA-Seq data has been extensively used to measure expression of transcripts, and to investigate alternative splicing, allele specific expression and RNA editing. Several methods have been developed to study alternative splicing. One of these methods is to focus on studying differential feature(exons or/and splicing junction) usage to understand alternative splicing. In these studies, the counts of features are obtained using the annotated gene model(see the following Figure), and are used to calculte usage of these features through the following method:

$$Usage\ of\ feature\ i = \frac{counts\ of\ feature\ i}{counts\ of\ other\ features\ excluding\ feature\ i}$$

Based on the usage of exons and splicing junctions, differential usage between two conditions are identified. The differentially expressed genes are derived from the differential usage of exons and splicing junctions within genes, subsequent gene sets or/and pathway analysis are performed using this derived differentlly expressed genes. It is obserserved that identifying differentially expressed genes using this approach is biased by the number of exons and splicing junctions.

There are several previous studies that showed using RNA-Seq data to study differentially expressed genes and subsequent gene sets or pathways analysis could be biased by read counts of transcripts and length

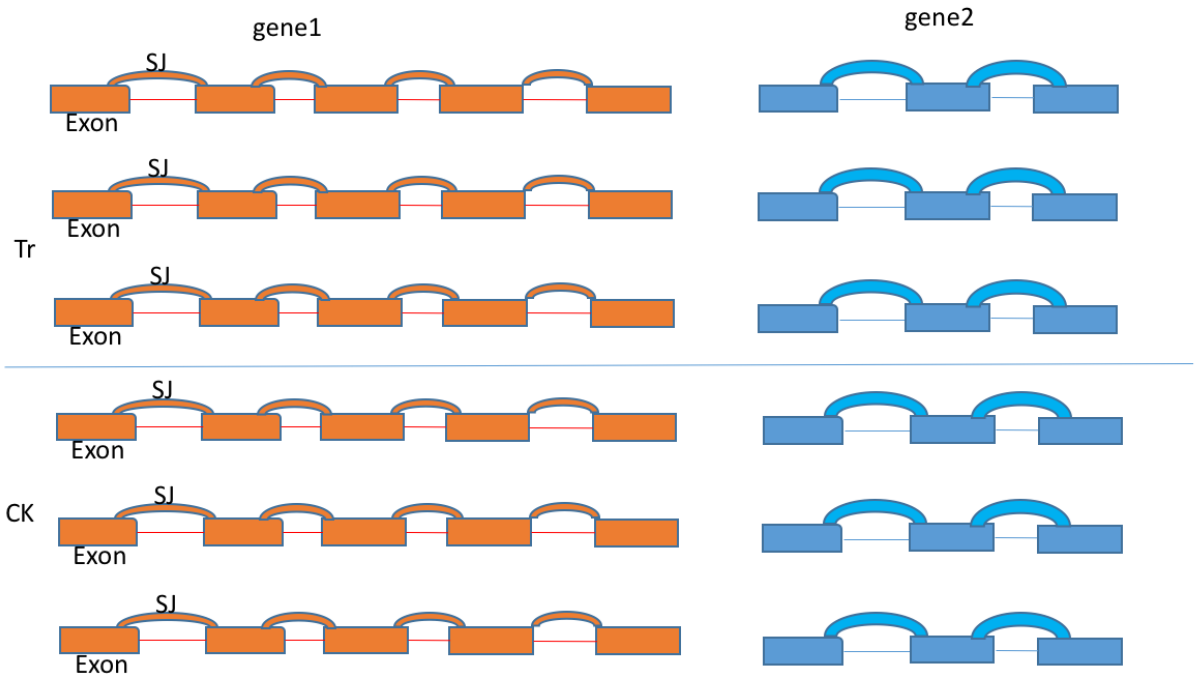


Figure 1: Bias from gene structure

of transcripts. Several methods are also developed to correct these biases either in the step of identifing differentially genes or in the step of performing gene sets enrichment analysis(Gao et al. 2011). A well known R package for correcting bias in gene set enrichment analysis, GoSeq, is available(Young et al. 2010). A logistic regression based approach is also developed for adjusting length bias of transcirpt in Gene Ontology enrichment analysis(Mi et al. 2012).

Similar situations are also observed in analyzing genome-wide methylation data. Several studies found that number of methylation within genes could cause a bias to identify diffenertially methylated genes. An R package is also available for correcting this bias(Phipson, Maksimovic, and Oshlack 2016).

However, bias on identifying differentilly expressed genes due to different gene structure and its corrections have not been addressed in detail. In this paper, we explore to identify the bias factor on RNA-Seq data related to gene structure, we demonstrate that by identifying right bais factors could lead to the better adjustment on DE analysis and the following gene GO term and pathway enrichment analysis. A R package based on this analysis procedure is also developed.

# RESULTS AND DISCUSSIONS

## Analysis on several real data sets

working flow

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 of splicing junctions.

gene_feature_DE.tiff pwFeatureFeature.pdf pwfFeatureGL.pdf pwfGeneFeature.pdf pwfGeneGL.pdf

gene_feature

GeneGL

GeneFeature

FeatureGL

FeatureFeature

In the second scenario,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. In this step,the $p$ value of differential usage of each subfeature of each gene are determined,and used to calculate the genewise $p$ value based on the following method described in DEXSeq R package(Anders, Reyes, and Huber 2012):

let $p_{il}$ be the $p$ value for $l$ subfeature(exon or splicing junction) of gene $i$, then genewise $p$ 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, we set 0.05 as threshold to define differentlly expressed gene list. We firstly applied logistic regression to identify the relatioship between the differentlly expressed gene list and possible bias factors such as the number splicing junction and exons.

Secondly, we use these differnentially expressed gene list to identify the relationship between the propotion of DE in gene sets and the possible bias factos such as gene length, number of exons and number splicing junctions.

Bias from sj and GL

Bias from sj and exons

- Correction for GO term enrichment analysis
- Correction for KEGG pawhay anlaysis.
- Correction for hallmark gene sets

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

Using the idea in (Geeleher et al. 2013), we explore to fix gene structure in the data set in blood cancer, and use permutation to generate more data sets to demostrate the bias from gene structure. We have about 13000 genes, each gene has certain number of subfeatures, in this permutation setting, we fix the number of subfeatures for each gene as the real data, and perform the following permutations:

- Permutating sample label
- Permutating mapping counts

By this way, we explore to identify the same enriched gene sets and pathways of GO and KEGG. If there are high consistencies between the resutls from permutated data sets and that of real data sets, this indicates that the number of subfeatures of genes biases gene set or/and pathway analysis.

## Conclusion

In summary, our work identified the potential bias in RNA-Seq data analysis when using differential exons or/and splicing junction usage to represent differential expressed gene, and using these differential expressed genes to perform subsequent gene set and pathway analysis. We futher sugguest a bias correction approach that could provide a more accutate gene set and pathway analsis. We implemented this method into an R package, and we believe that this R packcage could help other researchers when they perform similar type of RNA-Seq data analysis.

## 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.

Gao, Liyan, Zhide Fang, Kui Zhang, Degui Zhi, and Xiangqin Cui. 2011. "Length bias correction for RNA-seq data in gene set analyses." *Bioinformatics (Oxford, England)* 27 (5): 662–9. doi:10.1093/bioinformatics/btr005.

Geeleher, Paul, Lori Hartnett, Laurance J Egan, Aaron Golden, Raja Affendi Raja Ali, and Cathal Seoighe. 2013. "Gene-set analysis is severely biased when applied to genome-wide methylation data." *Bioinformatics (Oxford, England)* 29 (15): 1851–7. doi:10.1093/bioinformatics/btt311.

Mi, Gu, Yanming Di, Sarah Emerson, Jason S Cumbie, and Jeff H Chang. 2012. "Length bias correction in

gene ontology enrichment analysis using logistic regression." *PloS One* 7 (10). Public Library of Science: e46128. doi:10.1371/journal.pone.0046128.

Phipson, Belinda, Jovana Maksimovic, and Alicia Oshlack. 2016. "missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform." *Bioinformatics (Oxford, England)* 32 (2): 286–8. doi:10.1093/bioinformatics/btv560.

Young, Matthew D, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. 2010. "Gene ontology analysis for RNA-seq: accounting for selection bias." *Genome Biology* 11 (2). BioMed Central: R14. doi:10.1186/gb-2010-11-2-r14.