BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1777–1779
doi:10.1093/bioinformatics/btu090

Gene expression

Advance Access publication February 17, 2014

SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing

Xi Wang[1,2] and Murray J. Cairns[1,2,3,*]

[1]School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, University of Newcastle, Newcastle, [2]Hunter Medical Research Institute, New Lambton, and [3]Schizophrenia Research Institute, Sydney, NSW, Australia

Associate Editor: Martin Bishop

ABSTRACT

Summary: SeqGSEA is an open-source Bioconductor package for the functional integration of differential expression and splicing analysis in RNA-Seq data. SeqGSEA implements an analysis pipeline, which first computes differential splicing and differential expression scores, followed by integrating them into a per-gene score that quantifies each gene's association with a phenotype of interest, and finally executes gene set enrichment analysis in a cutoff-free manner to achieve biological insights. SeqGSEA accounts for biological variability and determines the statistical significance of gene pathways and networks using subject permutation, and thus requires at least five samples per group. Real applications show that SeqGSEA detects more biologically meaningful gene sets without biases toward long or highly expressed genes. SeqGSEA can be set up to run in parallel to reduce the analysis time.

Availability and implementation: The SeqGSEA package with a vignette is available at http://bioconductor.org/packages/release/bioc/html/SeqGSEA.html.

Contact: Murray.Carins@newcastle.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 1, 2013; revised on November 19, 2013; accepted on February 7, 2014

1 INTRODUCTION

By enabling simultaneous quantification of abnormal expression and aberrant splicing, high-throughput transcriptome sequencing or RNA-Seq is providing unprecedented insight into disease and condition-associated transcriptome alterations. Although these complex patterns of transcriptional activity have a profound influence in the pathophysiology of disease, we need tools that are capable of integrating their biological influences to determine their functional significance. Functional annotation of independently differentially expressed genes and their pathway analysis is ideal for investigating highly controlled interventions in clonal or syngeneic systems with no underlying heterogeneity. However, when comparing conditions in outbred populations, particularly for complex phenotypes or diseases, biologically significant effects are often distributed in pathways or networks such that the influence of individual genes is redundant and untraceable. Arguably the best alternative in thi+situation is to consider related gene sets using gene set enrichment analysis (GSEA) to identify disturbances at the pathway and network levels that have biologically associated function (Goeman and Buhlmann, 2007).

Although there are existing tools for performing GSEA on conventional gene expression data, there are currently no tools for implementing GSEA with RNA-Seq data that can efficiently incorporate gene expression and alternative splicing (AS). For example, GOSeq (Young *et al*., 2010) cannot account for AS in its enrichment analysis. In higher primates, the majority of genes undergo AS; especially, in humans more than 90% genes are alternatively spliced (Wang *et al*., 2008). Recent evidence suggests that splicing and transcription largely take place almost simultaneously (Djebali *et al*., 2012), and both result in the alteration of mRNA composition, stability, localization and translation. In this Applications Note, we introduce a new software system for implementing a recently proposed enrichment analysis approach for RNA-Seq data, combining both overall differential expression (DE) and differential splicing (DS) (Wang and Cairns, 2013). We also demonstrate the utility of this approach-designated SeqGSEA, to maximize the yield of biological insight from RNA-Seq data without any biases toward long or highly expressed genes, in a number of data sets.

2 IMPLEMENTATION

In well-annotated organisms, such as the human and mouse, the first step of RNA-Seq analysis is usually to map short reads to the reference genome with splice-tolerant mappers, such as Tophat, to yield BAM/SAM-formatted outputs. Given a set of gene annotation, SeqGSEA uses a Python script based on the HTSeq package (http://www-huber.embl.de/users/anders/HTSeq) to count reads mapped to individual exons of each gene. Based on the read counts, SeqGSEA uses the DSGseq method (Wang *et al*., 2013) to quantify transcript-level differences resulting from AS between samples of different phenotypes. Similarly, based on read counts summed up for each gene, SeqGSEA imports DESeq method (Anders and Huber, 2010) to quantify overall expression differences for each gene. As the statistical metrics given by DSGseq and DESeq for differential expression and differential splicing are not necessarily comparable, SeqGSEA uses a normalization step to aid integration. The normalization is performed against background metric distributions achieved by a number (say, 1000) of permutated datasets. We call the normalized metrics the DE and DS scores. To integrate DE and DS scores, two strategies can be selected, including a weighted linear

*To whom correspondence should be addressed.

combination and a rank-based integration. The rank-based strategy has a tendency to generate slightly more overrepresented gene sets than the linear combination. (See Supplementary Note S1 for more details on DE and DS integration.) We term the integrated scores as gene scores, representing each gene's association with the phenotype studied. Finally, SeqGSEA uses the state-of-the-art GSEA method (Subramanian *et al*., 2005) for functional enrichment analysis. The GSEA method then computes normalized enrichment scores (NESs) for each gene set against the permutation background, and calls statistical significance based on the observed NESs compared with those on the permutated datasets. SeqGSEA ignores the direction of gene expression changes in its analysis, as the significance of direction in DS is ambiguous, and the significance of direction in DE can also be variable, as interacting gene pairs are often inversely regulated. An overview of SeqGSEA's analysis workflow is shown in Supplementary Figure S1.

The SeqGSEA package includes a built-in capacity to generate well-designed figures for users to easily present the results of analysis, and well-formatted output text files to archive analysis outcomes with relevant information. For example, Supplementary Figure S2a shows the distribution of NESs compared with those on the permutated datasets, from which it is clear that a few gene sets are significantly overrepresented in the studied dataset. Another example in Supplementary Figure S2b shows the relationship between NESs and *P*-values, false discovery rates and family-wise error rates, providing a quick and direct overview of SeqGSEA's analysis results. For each significantly overrepresented gene set, functions are also available to generate detailed gene set-specific plots (Supplementary Fig. S3).

SeqGSEA is implemented in R, and has been deposited in Bioconductor. With a regular update cycle, SeqGSEA will keep updating by taking users' comments and implementing new functions that make the package more comprehensive. To reduce analysis running time, SeqGSEA was made multi-thread compatible so that users can easily run SeqGSEA with a specified number of multiple cores.

3 APPLICATION EXAMPLES

We applied the SeqGSEA package to a prostate cancer RNA-Seq dataset, downloaded from the NCBI SRA database with accession number SRP002628. The data were generated from 20 prostate cancer samples and 10 matched benign samples by Illumina GAII, ~22.2 million reads on average per sample. We mapped the reads to the human reference genome using Tophat, resulting in BAM files. Read counts at the exon level were computed using Python scripts based on HTSeq released with the package. DE and DS analyses were then conducted, followed by DE and DS score integration and gene set analysis with the GSEA algorithm. The analysis scripts are exemplified in the package vignette (released with the Bioconductor package; section "An analysis example") and provided in Supplementary Note S2 with corresponding mathematical explanation. With linear combination of DE and DS equally contributed, SeqGSEA resulted in 16 overrepresented gene sets at false discovery rate 0.05 (Supplementary Fig. S2b).

Comparison with alternative methods. We have compared SeqGSEA with alternative analysis pipelines, which can only account for DE and DS separately, on the prostate cancer dataset as well as two schizophrenia transcriptome sequencing datasets and shown that more gene sets with high biological relevancy can be detected by SeqGSEA (Wang and Cairns, 2013). A brief summary of this comparison study is provided in Supplementary Note S3. Moreover, while the existing GOSeq strategy compensates for biased selection of long and highly expressed genes (Young *et al*., 2010), we have not observed any bias in SeqGSEA's gene scores against gene length or read counts (Supplementary Note S4).

Typical running time. We have recorded the running time with the cancer data on the MsigDB collection 5 (c5, 1454 gene sets). Using eight cores of Intel Sandy Bridge E5-2600 CPUs, it took <16 h to finish. If a gene set collection contains more gene sets, it takes slightly more time for the whole procedure. We summarized running time of the cancer data and two schizophrenia datasets across two gene set collections in Supplementary Table S1.

4 DISCUSSION

SeqGSEA provides an RNA-Seq data analysis pipeline for functional integration of DE and DS, gathering multiple analysis steps together into a Bioconductor package, which enables bioinformaticians and biologists to interpret their own RNA-Seq data into biological insight, quickly and easily. The advantage of SeqGSEA over existing functional analysis pipelines includes (i) taking into account sequencing bias and biological variation with negative binomial models in DESeq and DSGseq, (ii) efficiently accounting of both differential expression and splicing, (iii) permutation-based gene scores securing enrichment analysis without bias toward gene sets of long and highly expressed genes and (iv) cutoff-free property that facilitates detection of subtle accumulating alteration in gene activity.

This package can also be used for DE- or DS-only GSEA by setting the weighting parameter in DE/DS score combination to 1 or 0. DE-only GSEA is particularly useful for organisms with few or none AS, such as prokaryotes. Therefore, SeqGSEA can be applied to any organisms and not just higher eukaryotes.

SeqGSEA performs statistical significance assessment based on subject permutation, and thus requires a moderate samples size. Greater than five samples per group meets the least requirement for running SeqGSEA. More importantly, enough sample size is critical to secure accurate analysis results because of the sample heterogeneity, particularly when studying human samples, which are generally subject to much larger biological variation than cell lines or model organisms (Oberg *et al*., 2012).

Funding: This study was supported by the Schizophrenia Research Institute, utilizing funding from NSW Health and the Henderson Foundation; X.W. is supported by a National Health and Medical Research Council project grant (631057) and M.C. is supported by an M.C. Ainsworth Research Fellowship in Epigenetics. Computational infrastructure used in this work was provided by Intersect Australia Ltd.

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Djebali,S. *et al*. (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Oberg,A.L. *et al*. (2012) Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC Genomics*, **13**, 304.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Wang,X. and Cairns,M.J. (2013) Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinformatics*, **14** (**Suppl. 5**), S16.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,W. *et al.* (2013) Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*, **518**, 164–170.

Young,M.D. *et al.* (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, **11**, R14.