Bioinformatics, 31(13), 2015, 2222–2224
doi: 10.1093/bioinformatics/btv119
Advance Access Publication Date: 24 February 2015
Applications Note

OXFORD

Gene expression

rSeqNP: a non-parametric approach for detecting differential expression and splicing from RNA-Seq data

Yang Shi[1,2], Arul M. Chinnaiyan[2,3,4,5,6] and Hui Jiang[1,6,*]

[1]Department of Biostatistics, [2]Michigan Center for Translational Pathology, [3]Department of Pathology, [4]Comprehensive Cancer Center, [5]Howard Hughes Medical Institute and [6]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on July 22, 2014; revised on February 1, 2015; accepted on February 19, 2015

Abstract

Summary: High-throughput sequencing of transcriptomes (RNA-Seq) has become a powerful tool to study gene expression. Here we present an R package, rSeqNP, which implements a non-parametric approach to test for differential expression and splicing from RNA-Seq data. rSeqNP uses permutation tests to access statistical significance and can be applied to a variety of experimental designs. By combining information across isoforms, rSeqNP is able to detect more differentially expressed or spliced genes from RNA-Seq data.
Availability and implementation: The R package with its source code and documentation are freely available at http://www-personal.umich.edu/~jianghui/rseqnp/.
Contact: jianghui@umich.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput sequencing of transcriptomes (RNA-Seq) is a widely used approach to study gene expression (Mortazavi *et al.*, 2008). Many statistical approaches have been developed to characterize gene expression variation across RNA-Seq experiments, and many of them are designed for testing differential expression (DE) of genes without considering their alternative spliced isoforms. For a comprehensive review, see Rapaport *et al.* (2013). Several recent studies have shown that directly applying the DE approach for detecting differential splicing (DS) may lead to erroneous results, because those approaches do not incorporate the complexity induced by isoform expression estimation for genes with multiple isoforms (Leng *et al.*, 2013; Trapnell *et al.*, 2013). To this end, several approaches were recently developed to detect DE at isoform level (Glaus *et al.*, 2012; Leng *et al.*, 2013; Trapnell *et al.*, 2013; Vardhanabhuti *et al.*, 2013). However, there are two remaining issues: (i) many existing approaches only compare between two biological conditions (such as normal versus diseased) and their usages for complex experimental designs are thus limited, and (ii) most existing approaches assume parametric distributions (Poisson or negative binomial) for observed read counts which, although can achieve good performance when the distributional assumptions hold, may have severely deteriorated performance should the distributional assumptions be violated, and that is often the case especially for large sample size RNA-Seq data where outliers usually exist (Li and Tibshirani, 2013).

Here we present rSeqNP, a non-parametric approach for testing DE and DS from RNA-Seq data. rSeqNP extends a non-parametric approach for detecting DE (Li and Tibshirani, 2013) and aims at detecting both DE and DS. rSeqNP can be used with a variety of RNA-Seq experimental designs, including those with two (unpaired or paired) or multiple biological conditions, and those with quantitative or survival outcomes.

2 Methods

Data preprocessing: Before applying rSeqNP, the raw sequence reads need to be processed to obtain the expression estimates of all

Table 1. Non-parametric statistics used by rSeqNP

Study design	Test statistic
Two condition comparison	Wilcoxon rank-sum statistic
Paired two condition comparison	Wilcoxon singed-rank statistic
Multiple condition comparison	Kruskal-Wallis statistic
Quantitative outcomes	Spearman's rank correlation coefficient
Survival outcomes	Score statistic of the Cox proportional hazard model

Table 2. Numbers of genes identified by different programs

	rSeqNP (unpaired)	rSeqNP (paired)	EBSeq (unpaired)	Cuffdiff (unpaired)
Gene.DE	3346	**4544**	2514	14
Isoform.DE[a]	2792 (2933)	**4163** (**4453**)	3279 (4323)	26 (31)
GDS	4122	**6050**	–	–

[a]Number of genes (Number of isoforms) that are detected.

the genes and their isoforms for each sample in the RNA-Seq study. This can be done using software tools like rSeq (Jiang and Wong, 2009), RSEM (Li and Dewey, 2011) and Cuffdiff (Trapnell *et al.*, 2013).

Testing DE *of genes and isoforms:* Using estimated expression values as input, rSeqNP tests for DE of genes and isoforms using non-parametric statistics that are constructed based on ranks of expression values. Table 1 summarizes the test statistics that are used in various study designs of RNA-Seq experiments. See Supplementary Section S1 for details. rSeqNP also reports *p*-values from these tests and false discovery rates (FDR) based on the Benjamini–Hochberg (BH) procedure.

Testing DE *and DS of genes jointly:* For each gene, rSeqNP also computes an overall gene-level differential score (*GDS*) based on the statistics used in testing the DE of the isoforms. Suppose that a gene has J distinct isoforms, and T_j is the statistic for testing the DE of the jth isoform (e.g. for two condition comparison, T_j is the Wilcoxon rank-sum statistic), the *GDS* is computed as $GDS = \sum_{j=1}^{J} T_j^2$. The *GDS* captures both DE and DS of the gene. For genes with a given number of isoforms, larger *GDS* indicates stronger evidence of DE and DS. The *GDS* incorporates information from all the isoforms of the gene, and therefore is more comprehensive in detecting differentially expressed and spliced genes than simply detecting genes that contain differentially expressed isoforms.

Estimating P-values and FDR for GDS: Since the null distribution of the *GDS* is unknown, rSeqNP implements a permutation plug-in method to estimate the *p*-values and FDRs (Li and Tibshirani, 2013). See Supplementary Section S2 for details.

3 Results

Using simulation studies, we find that our proposed approach has well controlled type I error rate, and achieves good statistical power for moderate sample sizes and effect sizes. See Supplementary Section S3 for details. Here we apply rSeqNP to a real RNA-Seq dataset and compare it with EBSeq (Leng *et al.*, 2013) and Cuffdiff (Trapnell *et al.*, 2013), which are two existing approaches for detecting DE of genes and isoforms from RNA-Seq data. The dataset was generated from paired-end RNA-Seq experiments performed on prostate cancer samples and matched benign samples from 14 Chinese prostate cancer patients (Ren *et al.*, 2012). Since neither EBSeq nor Cuffdiff can handle paired two-group comparison, we run all three programs on the dataset under the setting of unpaired two-group comparison, i.e. treat cancer and benign samples as two distinct groups, as well as run rSeqNP under the setting of paired two-group comparison. For preprocessing, we use the programs suggested by each of the three approaches to quantify expression values for all the genes and isoforms: RSEM (Li and Dewey, 2011) for EBSeq, rSeq (Jiang and Wong, 2009) for rSeqNP and the integrated quantification program for Cuffdiff. Nevertheless, the three programs for quantification show very similar results (see Supplementary Section S4).

Table 2 summarizes the numbers of differentially expressed and spliced genes identified by each program. See Supplementary Section S4 for detailed steps and additional results of the analysis. We use FDR ≤ 0.05 for rSeqNP and Cuffdiff, and posterior probability of being differentially expressed (PPDE) ≥ 0.95 for EBSeq to call a differential event. PPDE is the metric reported by EBSeq and PPDE ≥ 0.95 corresponds to controlling FDR at 5%. As expected, when treating the data as two distinct groups, rSeqNP detects fewer differentially expressed genes and isoforms, but by applying the *GDS*, more differentially expressed or spliced genes are detected. Furthermore, when accounting for the paired two group nature of the data, rSeqNP detects even more differential events. We find that Cuffdiff detects a much smaller number of differentially expressed or spliced genes, which is consistent with report from another study (Seyednasrollah *et al.*, 2013).

4 Conclusion

We present rSeqNP, a non-parametric approach for detecting differentially expressed and spliced genes from RNA-Seq data. It is flexible in handling various types of experimental designs. It is worth mentioning that, as pointed out by the reviewer, our method relies on expression estimates for genes and isoforms reported from upstream programs, which are typically based on parametric approaches. The major limitation of rSeqNP is that its power is relatively low for small sample size RNA-Seq data. In simulation studies, we show that the power is decent with five or more samples in each group for two-group comparison. If the sample size is even smaller, parametric approaches would be preferred.

Funding

National Institute of Health (U01CA111275 and 1UM1HG006508 to A.M.C.). A.M.C. is also supported by the Alfred A. Taubman Institute, the American Cancer Society, and the Howard Hughes Medical Institute. H.J. is supported by University of Michigan (startup grant).

Conflict of Interest: none declared.

References

Glaus,P. *et al.* (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, **28**, 1721–1728.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Leng,N. *et al.* (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.*, **22**, 519–536.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.

Ren,S. *et al.* (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.*, **22**, 806–821.

Seyednasrollah,F. *et al.* (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinf.*, **16**, 59–70.

Trapnell,C. *et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.

Vardhanabhuti,S. *et al.* (2013) A hierarchical bayesian model for estimating and inferring differential isoform expression for multi-sample RNA-Seq data. *Stat. Biosci.*, **5**, 119–137.