Bioinformatics, 34(6), 2018, 1066–1068
doi: 10.1093/bioinformatics/btx690
Advance Access Publication Date: 27 October 2017
Applications Note

OXFORD

Gene expression

DIEGO: detection of differential alternative splicing using Aitchison's geometry

Gero Doose[1,2,3,†], Stephan H. Bernhart[1,2,†], Rabea Wagener[4] and Steve Hoffmann[1,2,5,*]

[1]Transcriptome Bioinformatics Group, Interdisciplinary Center for Bioinformatics, Leipzig University, 04107 Leipzig, [2]Chair of Bioinformatics, Faculty of Mathematics and Computer Science, Leipzig University, 04107 Leipzig, Germany, [3]ecSeq Bioinformatics, 04103 Leipzig, Germany, [4]Institute of Human Genetics, University of Ulm and University of Ulm Medical Center, 89081 Ulm, Germany and [5]Computational Biology Group, Leibniz Institute on Ageing - Fritz Lipmann Institute (FLI) and Friedrich-Schiller-University Jena, 07745 Jena, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Ziv Bar-Joseph

Received on May 23, 2017; revised on October 15, 2017; editorial decision on October 25, 2017; accepted on October 26, 2017

Abstract

Motivation: Alternative splicing is a biological process of fundamental importance in most eukaryotes. It plays a pivotal role in cell differentiation and gene regulation and has been associated with a number of different diseases. The widespread availability of RNA-Sequencing capacities allows an ever closer investigation of differentially expressed isoforms. However, most tools for differential alternative splicing (DAS) analysis do not take split reads, i.e. the most direct evidence for a splice event, into account. Here, we present DIEGO, a compositional data analysis method able to detect DAS between two sets of RNA-Seq samples based on split reads.
Results: The python tool DIEGO works without isoform annotations and is fast enough to analyze large experiments while being robust and accurate. We provide python and perl parsers for common formats.
Availability and implementation: The software is available at: www.bioinf.uni-leipzig.de/Software/DIEGO.
Contact: steve@bioinf.uni-leipzig.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The phenomenal complexity of transcripts is mainly enabled by alternative splicing, allowing cells to generate multiple different mRNAs from a single gene. It affects >95% of all human multi-exon genes, and differences in isoform usage, i.e. differential alternative splicing (DAS), contribute to many phenotypic differences, e.g. in diseases (Tazi *et al.*, 2009). One prominent example is the epidermal growth factor receptor (EGFR), where cancer cells often produce a splice variant that lacks exon 4—in contrast to the surrounding healthy tissue (Oltean and Bates, 2014). Thus, the detection of differential splice variants, e.g. in cancer versus control tissues, is of interest in many medical and biological research projects.

In the past years, different tools for the detection of DAS on the basis of RNA-seq data have been proposed [reviewed in (Liu *et al.*, 2014; Hayer *et al.*, 2015)]. These methods may roughly be divided into those that only work on existing isoform annotations and those that are able to detect changes of yet unknown isoforms, i.e. isoform resolution methods. The tool DEXSeq (Anders *et al.*, 2012) works with read counts for existing isoform annotations and uses a negative binomial model to detect alternative exon usage. This model has recently been extended to split read counts (Hartley and Mullikin, 2016). Alternatively, IUTA (Niu *et al.*, 2014) infers the isoform usage for two sets of samples and tests for differences under Aitchison's geometry (Aitchison, 1986). IUTA's application is also

limited to known isoforms. The most prominent isoform resolution method Cufflinks (Trapnell *et al.*, 2010), in contrast, is able to report new events by inferring isoform structure and transcript abundances. Cufflinks achieved good benchmarks under certain test conditions (Liu *et al.*, 2014). However, comparative studies also revealed that the performance of all methods varies considerably with the test scenario (Liu *et al.*, 2014; Hayer *et al.*, 2015). In several test cases, the agreement of the tools on DAS events was low. More importantly, for large scale projects with a multitude of samples, the runtime of DAS detection methods quickly becomes a serious bottleneck. Here, we present DIEGO (DIfferential altErnative splicinG detectiOn), a method that combines the simplicity of a count based approach with the ability to also report DAS of yet unknown isoforms. Based on split mapped RNA-seq reads, it is capable of rapidly analyzing even large groups of samples. Our approach also works on exon based read counts, as used by DEXSeq, to detect differential exon expression.

2 Materials and methods

2.1 Differential alternative splicing detection (DAS mode)

In a first step, all splice junctions (as inferred from split read alignments) with one or both splice sites within the boundaries of annotated genes are collected (see Supplementary Material). Junctions and genes with an insufficient split read coverage or not present in a minimum number of samples are discarded. Subsequently, for each gene the split read data is transformed into a compositional data space, i.e. the simplex space:

$$\mathcal{S}^n = \left\{ [x_1, \ldots, x_n] : x_i \geq 0 \text{ for } i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} x_i = 1 \right\}$$

where each vector consists of n components that correspond to n exon junctions of a gene. Note, that the sum of all components is constrained to 1. Thus, we transform the split read counts supporting the junctions of a gene to *fractions*. For each junction i of a gene we calculate

$$x_i = \frac{c_i}{\sum_{j \in \mathcal{D}} c_j}$$

where $\mathcal{D}$ is the set of all junctions for the given gene. After the transformation, we are now able to calculate the simplex center *cen*, roughly comparable to a mean in euclidean space, for each of the groups of samples under investigation. The distance between two centers then allows to directly measure abundance changes for each junction.

In the practical implementation of this approach, only junctions with a minimum absolute abundance change are considered in further statistical evaluation (default: 1.0). To detect significantly differential junctions, i.e. DAS events, DIEGO uses a Mann-Whitney U test. Therefore, the data is transformed back to the euclidean space. Furthermore, p-values obtained from this non-parametric test are subsequently corrected for multiple testing using the Benjamini-Hochberg Method. Finally, DIEGO reports the results in a csv—file.

2.2 Clustering and outlier detection

In order to detect outliers or identify sub-groups within a set of samples, our method allows clustering based on Aitchison's distance.

To do this, genes with the highest variance of exon junction expression are pre-selected, where the variance is defined as the average distance from each sample to the centre of all samples. For each sample pair, we average Aitchison's distances of all genes in this highly variable set to generate a distance matrix between the samples. This matrix is subjected to a hierarchical agglomerative clustering using the average linkage method. When run in clustering mode, the tool will generate a dendrogram that allows the user to easily spot similarly and dis-similarly spliced samples.

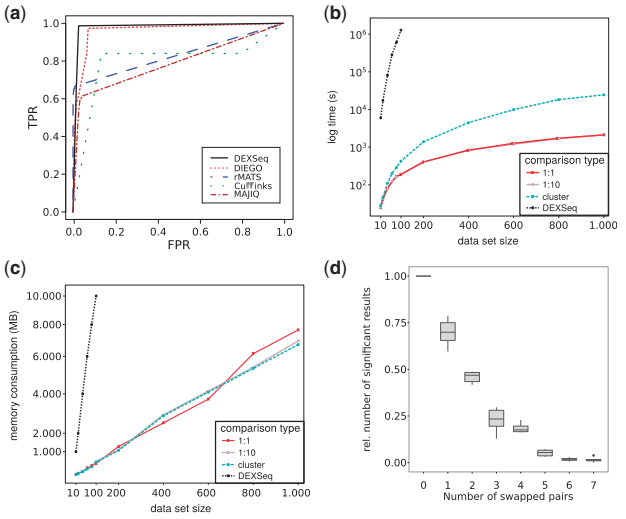

Fig. 1. Performance of DIEGO. (**a**) Receiver operator statistics for simulated data comparison of DIEGO (close dots) to DEXSeq (solid line), cufflinks (far dots), rMATSturbo (big dashes) and MAJIQ (dots/dashes) on an artificial dataset. (**b,c**) Time and memory consumption of DIEGO in clustering mode (blue) and DAS mode with different relative group sizes (solid line with dots 1:1, solid line 1:10 group size) compared to DEXSeq (black). (**d**) Effect of swapping samples between conditions on the number of significant results (Color version of this figure is available at *Bioinformatics* online.)

3 Performance evaluation

In simulation experiments, DIEGO performs comparable to DEXSeq regarding sensitivity and specificity, and outperforms rMATSturbo (Shen, 2014), MAJIQ (Vaquero-Garcia *et al.*, 2016) and Cufflinks (Fig. 1a). However, DIEGO clearly needs less time and memory compared to DEXSeq when using TCGA transcriptome data for an increasing number of samples (Fig. 1b and c). In order to get an idea on the stability of DIEGO with respect to wrong group assignments, we randomly chose a number of control/tumor sample pairs and swapped their assignments. As Figure 1d shows, DIEGO is quite stable against a small number of wrong assignments, while a higher number leads to a decrease of predicted differentially used splice junctions to only about 3% of the original value, indicating a high specificity when applied to random data. Analyzing DIEGO's predictions revealed a slight bias towards genes with a high number of splice sites to be predicted to contain DAS.

4 Conclusion

We present DIEGO, a fast and robust tool for detecting DAS in large RNA-Seq datasets. DIEGO includes parsers for standard splice aware aligners and TCGA's splice count files. The low time and memory consumption together with the relatively low false positive rate make DIEGO suited for the analysis of large RNA-Seq datasets with split read information.

Funding

This work has been supported by the German Federal Ministry of Education and Research (BMBF) (PTJ grant HNPCCSys 031 6065A, ICGC MMML-Seq 01KU1002A-J and ICGC-Data Mining 01KU1505-C, G). R.W. was supported by a Bausteinantrag of the Medical Faculty of University Ulm.

Conflict of Interest: none declared.

References

Aitchison,J. (1986) *The Statistical Analysis of Compositional Data, Monographs on Statistics and Applied Probability*. Chapman & Hall Ltd, London.

Anders,S. *et al*. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res*., **22**, 2008–2017.

Hartley,S.W. and Mullikin,J.C. (2016) Detection and visualization of differential splicing in RNA-seq data with junctionseq. *Nucleic Acids Res*., **44**, e127.

Hayer,K.E. *et al*. (2015) Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, **31**, 3938–3945.

Liu,R. *et al*. (2014) Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics*, **15**, 364.

Niu,L. *et al*. (2014) IUTA: a tool for effectively detecting differential isoform usage from RNA-Seq data. *BMC Genomics*, **15**, 862.

Oltean,S. and Bates,D.O. (2014) Hallmarks of alternative splicing in cancer. *Oncogene*, **33**, 5311–5318.

Shen,S. *et al*. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *PNAS*, **111**, E5593–E5601.

Tazi,J. *et al*. (2009) Alternative splicing and disease. *Biochim. Biophys. Acta*, **1792**, 14–26.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–515.

Vaquero-Garcia,J. *et al*. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, **5**, e11752.