SplicingFactory – Splicing diversity analysis for transcriptome data

Péter Szikora[1], Tamás Pór[1], Endre Sebestyén[1*]

[1] 1st Department of Pathology and Experimental Cancer Research, Semmelweis University, Budapest, Hungary

[*] To whom correspondence should be addressed

Abstract

Summary: Alternative splicing contributes to the diversity of RNA found in biological samples. Current tools investigating patterns of alternative splicing check for coordinated changes in the expression or relative ratio of RNA isoforms. However, the molecular process of splicing is stochastic and changes in RNA isoform heterogeneity for a gene might arise between samples or conditions. Here we present a tool for the characterization and analysis of RNA heterogeneity using isoform level expression measurements.

Availability and implementation: The SplicingFactory package is freely available under the GPL-3.0 license from Bioconductor (https://bioconductor.org) for Windows, MacOS and Linux.

Contact: sebestyen.endre@med.semmelweis-univ.hu

Introduction

The mechanism of alternative splicing is well-known and described in most eukaryotic organisms (Lee and Rio, 2014). Alternative splicing expands the RNA repertoire of most genes, leading to changes in mRNA coding sequence or untranslated regions (UTRs). These changes might affect mRNA stability, localization or translation (Baralle and Giudice, 2017). Mis-splicing contributes to disease (Scotti and Swanson, 2015) and mutated splicing factors might act as oncoproteins or tumor-suppressors (Dvinge *et al.*, 2016).

RNA-sequencing experiments regularly investigate consistent alternative splicing or mRNA isoform changes between conditions (Van den Berge *et al.*, 2019). Most tools look for changes, where the ratio of mRNA isoforms or the presence of a specific alternative splicing event is coordinately increased or decreased. However, experimental evidence started to accumulate on the biological significance of gene expression (Eling *et al.*, 2019) and more interestingly, splicing variance (Wan and Larson, 2018). A number of methods are already developed to detect changes in expression variance (Eling *et al.*, 2019), but changes in splicing variance are not investigated regularly and only a few tools exist.

One of the first papers in this area used Shannon-entropy to characterize splicing variance, investigating cDNA and cDNA tag libraries in 27 cancer types (Ritchie *et al.*, 2008). In half of the cancers studied, they described a significant entropy gain compared to normal tissue. The RNentropy tool calculates Shannon-entropy for genes across samples to detect differential expression between any number of conditions (Zambelli *et al.*, 2018). Whippet uses Shannon-entropy to detect and quantify complex alternative splicing events (Sterne-Weiler *et al.*, 2018) and the authors describe that complex, high-entropy splicing events are conserved, tissue-regulated and more prevalent in various cancer types. SpliceHetero aims to

characterize spliceomic intra-tumor heterogeneity (sITH) from bulk tumor RNA-sequencing (Kim *et al.*, 2019). The authors used the Jensen - Shannon Divergence to characterize splice site usage differences between samples and found that increased sITH was correlated with cancer progression and worse survival. The Splice Expression Variation Analysis (SEVA) tool aims to model increased heterogeneity of splicing variants in cancer, using a rank-based multivariate statistics, comparing splice junction expression profiles between conditions (Afsari *et al.*, 2018). Finally, the sQTLseekeR R package analyzes associations between genotype information and transcript relative expression. Even though the main goal of sQTLseekeR is to detect splicing quantitative trait loci (sQTLs), it can also detect splicing variance QTLs (svQTLs) (Monlong *et al.*, 2014), where changes in splicing isoform diversity are associated to a genotype.

We developed the SplicingFactory R package, to facilitate the analysis of splicing isoform diversity in RNA-sequencing experiments, and investigate changes in diversity between conditions with a tool that integrates into the Bioconductor package ecosystem and uses standardized data structures.

Implementation

The splicing isoform diversity analysis works as a two-step process: 1) the package calculates a diversity value for each gene in each sample, using splicing isoform expression values, and 2) calculates differential diversity results between conditions. Diversity values from 1) can be used independently from step 2) for custom downstream analyses or visualizations.

Input data structure

The package can process R matrices and data frames with expression values, assay data from the SummarizedExperiment Bioconductor S3 object, data from the DGEList S4 object of edgeR (Robinson *et al.*, 2010), or the output of tximport (Soneson *et al.*, 2016). The package requires that samples are specified as columns and transcript level expression values are specified as rows. Additionally, it needs a vector of genes used to aggregate and analyze the splicing isoform level data, and a vector of sample categories used to calculate differential diversity. All of the data structures might contain RNA-sequencing read counts, RPKM, FPKM or TPM values. Any preliminary data normalization steps are the responsibility of users.

Diversity calculation

As a first step, the package calculates diversity values using the calculate_diversity() function for each gene and each sample. Multiple diversity measures are implemented, including the Shannon-entropy, Laplace-entropy, Gini-index, Simpson-index and the inverse Simpson-index.

Shannon-entropy is a classic measure of uncertainty in information theory, ranging from 0 to $\log_2$(isoform-number) for a gene. As the maximum value of Shannon-entropy depends on the number of splicing isoforms for a gene, we implemented a normalized Shannon-entropy, that ranges between 0 and 1. This makes it possible to compare entropy values of genes with different number of isoforms. A 0 Shannon-entropy means a single splicing isoform is expressed from a gene, while 1 means all isoforms are evenly expressed, with no dominant isoform. The package can also calculate Laplace-entropy, a Bayesian estimate of the Shannon-entropy, where a pseudo-count of 1 is added to the isoform categories for each sample.

The Gini-index originally intended to represent income inequality in economics. It ranges between 0 and 1, where 0 means complete equality, *i.e.* all isoforms have the same expression, while 1 means complete inequality, with only a single isoform being expressed. The Simpson-index is a measure of diversity originally used in ecology to quantify species diversity. A 0 Simpson-index means low diversity, *i.e.* one dominant isoform, while 1 means high diversity, where all isoforms have the same expression. In contrast, the inverse Simpson-index starts at value 1, and higher values mean greater isoform diversity.

The calculate_diversity() function returns a SummarizedExperiment object, that contains the gene level splicing diversity values, together with gene names, sample ids, and metadata information, including the method used, and if any normalization was applied. The function removes genes with a single isoform and adds an NA value for genes where the expression of all isoforms is 0 in a specific sample, and a meaningful diversity value is impossible to calculate.

Differential diversity calculation

Users can calculate splicing diversity changes between two conditions using the calculate_difference() function. Accepted input formats are R data frames or a SummarizedExperiment object. In the case of a data frame, gene names must be present in the first column, and splicing diversity values in all additional columns. Differences and $\log_2$ fold changes in diversity can be calculated using the mean or median values across conditions. The function returns the mean or median for both conditions, the difference of means or medians, and their $\log_2$ fold change.

Statistical significance of the changes can be assessed using a Wilcoxon-test or sample label shuffling and the function returns the p-values, together with the FDR corrected ones. It automatically excludes genes from the significance analysis, where some of the sample diversity values are missing and the total number of samples is insufficient for significance calculation.

Example analysis

To showcase the analysis steps, we processed a subset of the TCGA breast invasive carcinoma (BRCA) dataset (The Cancer Genome Atlas Network, 2012), that contains RSEM (Li and Dewey, 2011) based splicing isoform level expression estimates. The dataset was downloaded using the TCGAbiolinks package (Mounir *et al.*, 2019), and we selected 20 samples from the luminal A BRCA subtype together with the corresponding 20 normal tissue samples from the same patients. We selected 300 genes and their splicing isoforms from these samples. All steps of the data download, filtering and preprocessing are available as a separate R script in the package.

After importing the data, we calculated the Gini-index and the Shannon-entropy diversity measures for all genes and samples. As can be seen on **Figure 1A**, the two measures have opposing patterns as expected based on their definitions. Using the Shannon-entropy values, we calculated the diversity changes between the normal and tumor samples using means, and assessed their significance with the Wilcoxon-test. **Figure 1B** shows the results of this calculation, where genes with a significant change in Shannon-entropy are highlighted in yellow. In this dataset, we detected 18 significantly changing genes, using an absolute $\log_2$ fold-change > 1 and FDR < 0.05 as cutoffs. We selected the LRFFIP1 gene for visualization, where the mean Shannon-entropy difference was -0.22, the $\log_2$ fold-change was -0.79, and

the FDR was 0.000056. **Figure 1C** shows the Shannon-entropy values for the gene in each sample, grouped by condition. **Figure 1D** shows the % expression value of all LRFFIP1 splicing isoforms calculated from the originally downloaded RSEM-based data. A standard analysis workflow and several visualizations, together with the R code, are also available in the package vignette.

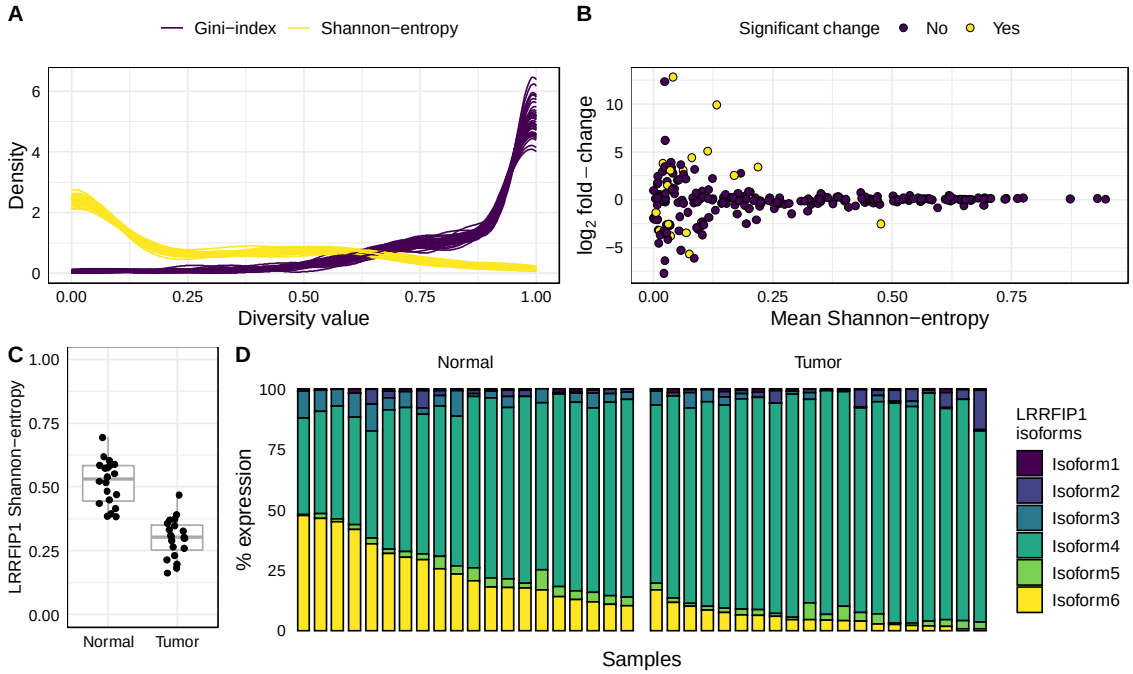

Figure 1: Visualization of various analysis results from the package using a set of normal and luminal A BRCA tumor samples from TCGA. **A**) Distribution of the Gini-index and Shannon-entropy calculated on the same set of samples. Separate lines mean a single sample. **B**) MA-plot of Shannon-entropy values. The x-axis shows the mean Shannon-entropy across samples, while the y-axis shows the $\log_2$ fold change of average Shannon-entropy values between tumor and normal samples. **C**) Box plot showing the Shannon-entropy values in the tumor and normal samples for the LRRFIP1 gene. **D**) Bar plot showing the % expression value of all transcript isoforms of the LRRFIP1 gene in the analyzed samples.

Conclusion

We have developed a package called SplicingFactory that enables the analysis of splicing isoform diversity in biological samples and between different conditions. We illustrated the functionality using publicly available TCGA data and detected a number of genes with changing splicing isoform diversity.

Future plans for development include adding functions to correct for possible biases from splicing isoform numbers or expression levels, normalize data before diversity calculation, use beta-regression for significance testing, and calculate alternative splicing diversity at the splicing event level.

Acknowledgements

We thank Alexa Szeifert for quality control analysis of RNA-sequencing datasets during initial development.

Funding information

This work was supported by the Semmelweis University Directorate of Innovation, Hungary [STIA_18_KF to ES] and the National Research, Development and Innovation Office, Hungary [FK-132666 to ES].

Conflict of interest

PS is an employee of Turbine Simulated Cell Technologies Ltd. TP is an employee of Covance Inc.

References

Afsari,B. *et al.* (2018) Splice Expression Variation Analysis (SEVA) for inter-tumor heterogeneity of gene isoform usage in cancer. *Bioinformatics*, **34**, 1859–1867.

Baralle,F.E. and Giudice,J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, **18**, 437–451.

Van den Berge,K. *et al.* (2019) RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.*, **2**, 139–173.

Dvinge,H. *et al.* (2016) RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer*, **16**, 413–430.

Eling,N. *et al.* (2019) Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.*

Kim,M. *et al.* (2019) SpliceHetero: An information theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq. *PLoS One*, **14**, e0223520.

Lee,Y. and Rio,D.C. (2014) Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.*, **84**, 150317182619002.

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.

Monlong,J. *et al.* (2014) Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.*, **5**, 4698.

Mounir,M. *et al.* (2019) New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEX. *PLoS Comput. Biol.*, **15**, e1006701.

Ritchie,W. *et al.* (2008) Entropy measures quantify global splicing disorders in cancer. *PLOS Comput. Biol.*, **4**, e1000011.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.

Scotti,M.M. and Swanson,M.S. (2015) RNA mis-splicing in disease. *Nat. Rev. Genet.*, **17**, 19–32.

Soneson,C. *et al.* (2016) Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521.

Sterne-Weiler,T. *et al.* (2018) Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Mol. Cell*, **72**, 187–200.

The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Wan,Y. and Larson,D.R. (2018) Splicing heterogeneity: separating signal from noise. *Genome Biol.*, **19**, 86.

Zambelli,F. *et al.* (2018) RNentropy: An entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Res.*, **46**, e46.