Identification and visualisation of differential isoform expression in RNA-seq time series

Supplementary Material

María José Nueda, Jordi Martorell-Marugan, Cristina Martí, Sonia Tarazona, Ana Conesa
September 3, 2017

1 Introduction

Alternative splicing (AS) is a common mechanism of higher eukaryotes to expand transcriptome complexity and functional diversity. The expression of alternative isoforms of many genes is a developmentally regulated process [1] and AS has been shown to occur as response to environmental cues [2]. Hence, there is an interest in studying the dynamics of AS. Deep sequencing methods currently used in transcriptomics research allow for the study of AS. Reads that map to splice junctions can be used to estimate splicing events, while several transcript reconstruction and quantification methods have been published that enable inference of isoform expression with different levels of accuracy [3]. More recently, long-read sequencing platforms, which allow transcript identification without the need of a reconstruction step, have become available to boost the study of isoform regulation [4]. While many algorithms have been developed for differential AS analysis most of these approaches target pair-wise comparisons i.e. [5] and [6] and have not yet developed specific models that integrate time course with differential splicing analysis. The DICESeq method was designed for a better estimation of isoform expression in time series, but does not implement a specific strategy for obtaining differentially expressed isoforms [7]. Topa and Honkela proposed to model time series as a Gaussian process and isoform levels as proportions over the total gene expression to then evaluate splicing as the change between time-dependent and time-independent models [8]. This approach requires large datasets to fit parameters and does not consider comparisons between multiple series, such as treatment and control.

Compared to genes or transcripts, the analysis of differential isoform expression in time course experiments poses a number of specific challenges. Different transcripts of the same gene may vary in their time trajectories and the analysis algorithm should be able to identify those genes where isoform profiles change differently in a significant manner, i.e. those genes that are differentially spliced across time. Additionally, joint visualisation of significant splicing changes is complicated by the fact that genes have different number of isoforms and hence data does not fit into the structure of traditional clustering, where the same number of data points is required for each feature. Therefore, novel clustering strategies should be envisioned to group genes expressing their isoforms in a similar fashion. Finally, transcripts of the same gene have frequently very
different expression levels, with one "major" isoform bearing most of the expression signal and alternative isoforms being lowerly expressed. Ideally, analysis approaches should be able to account for this characteristic and identify those cases where genes change their major isoform in the course of time.

maSigPro is an R package specifically designed for the analysis of multiple time course transcriptomics data. maSigPro fits an optimized polynomial linear model to describe the dynamics of gene expression in one or multiple experimental conditions and selects genes with significant model coefficients [9]. The package incorporates a clustering function to visualize genes with similar profiles. maSigPro was initially developed for microarrays and later updated to model count data [10]. In this paper we present Iso-maSigPro, a further adaptation to study differential isoform usage in time course RNA-seq experiments. We implement a new function to model differential splicing and combine this with differential transcript expression analysis to identify genes where isoforms change expression across time. Novel query and visualisation functions allow selecting genes with the strongest isoform switches and grouping genes with similar time-dependent AS patterns.

2 Methods

2.1 Model

The Generalized Linear Model (GLM) described in [10] to study the gene expression value $y_i$ at observation $i$, time $t_i$ and $s$ experimental conditions (i.e., treatments, tissues, strains, etc) identified by $s−1$ binary variables ($z_{1i},...,z_{si−1}$) can be written as follows (when considering $s=2$ and linear effects):

$$g(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 z_{1ij} + \beta_3 t_{ij} z_{1ij}$$

being $\mu_{ij} = E(y_{ij})$, $g$ a monotonic differentiable function called 'link function', which characterizes the GLM model, and $\beta_0, \beta_1, \beta_2, \beta_3$ the coefficients to estimate.

2.1.1 Model for Differentially Spliced Genes (DSG) across time:

For each multi-isoform gene two models are created, identifying $J$ isoforms with $J−1$ binary variables ($I_{1j},...,I_{J−1}$). The reference model, $M_0$, considers there exist only constant differences between isoforms and the global gene model, $M_1$, considers the possibility of a time vs condition vs isoform interaction. $M_0$ imposes parallel profiles to the different isoforms, in contrast $M_1$ allows modeling different profiles and hence captures the differential splicing cases. For instance, for a gene with two isoforms, two experimental conditions or series and linear effects:

$$M_0 : \quad g(\mu_{ij}) = \beta_0^0 + \beta_1^0 t_{ij} + \beta_2^0 z_{1ij} + \beta_3^0 t_{ij} z_{1ij} + \beta_4^0 I_{1j}$$

$$M_1 : \quad g(\mu_{ij}) = \beta_0^1 + \beta_1^1 t_{ij} + \beta_2^1 z_{1ij} + \beta_3^1 t_{ij} z_{1ij} + \beta_4^1 I_{1j} + \beta_5^1 t_{ij} I_{1j} + \beta_6^1 z_{1ij} I_{1j} + \beta_7^1 t_{ij} z_{1ij} I_{1j}$$

being $\mu_{ij} = E(y_{ij})$ the expected value for observation $i$ and isoform $j$.

To evaluate the statistical significance of the interaction, both models are compared for each gene. In GLMs hypothesis testing is based on the log-likelihood ratio statistic [11, 12].

$$2[l(\hat{\beta}_1) - l(\hat{\beta}_0)] \sim \chi^2_{p_1 - p_0}$$
where \( l(\hat{\beta}_1) \) is the maximized likelihood of the complete model with \( p_1 \) coefficients and \( l(\hat{\beta}_0) \) the likelihood of the reference model with \( p_0 \) parameters, being \( p_0 < p_1 \).

2.1.2 The Iso-maSigPro functions

Seven new functions have been added to the maSigPro package to enable analysis of differentially expressed isoforms. Figure 1 shows the analysis pipeline and novel Iso-maSigPro functions:

1. \texttt{IsoModel()} implements the DS models. Basically, data are first split between mono-isoform and multi-isoform genes and optionally low expressed alternative transcripts are filtered out. For each multi-isoform gene, \( M_0 \) and \( M_1 \) models are created using as \( f(t_{ij}, z_{ij}) \) the polynomial model obtained with the generic \texttt{make.design.matrix()} maSigPro function that best describes the experimental design. The comparison of both models gives as a result a FDR-corrected p-value of differential splicing.

2. Transcripts from significant DSGs are then subjected to regular Next-maSigPro analysis to detect Differentially Expressed Transcripts (DETs).

3. \texttt{IsoModel()} returns a list of DSGs together with the estimated models of associated isoforms to be used as input in \texttt{getDS()} function to obtain a selection of DSGs at a preestablished level of goodness of fit for each model.

4. Downstream analysis can be performed with functions \texttt{seeDS()}, \texttt{tableDS()}, \texttt{getDSPattern()}, \texttt{PodiumChange()} and \texttt{IsoPlot()}, that cluster, select and visualize patterns of isoform change.

Note that in this formulation, it is possible that a gene is called DSG but no significant DETs of that gene are found under the significance level, goodness of fit and multiple testing correction constraints of the regular maSigPro analysis.

\[ \text{IsoModel()} \quad \text{fits model for DS} \]
\[ \text{getDS()} \quad \text{selects significant DSGs and DETs} \quad \text{ IsoPlot()} \quad \text{shows expression of indicated DSG} \]
\[ \text{seeDS()} \quad \text{clusters DETs} \quad \text{PodiumChange()} \quad \text{finds DSGs with major isoform switch} \]
\[ \text{tableDS()} \quad \text{identifies cluster location of major and minor isoforms} \]
\[ \text{getDSPattern()} \quad \text{extracts genes with specific isoform clustering pattern} \]

Figure 1: Workflow for Iso-maSigPro analysis.

2.2 Visualization

Typically, maSigPro will cluster features according to their expression pattern in all experimental conditions. This option is still available for all differential transcripts regardless of their parent gene. The Iso-maSigPro framework allows for two additional
visualisation functions to study differential splicing results: differential splicing clustering and major isoform switch.

2.2.1 Differential splicing clustering with seeDS() and tableDS()

The clustering strategy implemented with these two functions aims to identify groups of DSGs with similar isoform expression patterns. First, seeDS() takes DETs - either from DSGs or in combination with DETs from single-transcript genes - and clusters them into \( k \) groups with any of the available maSigPro clustering approaches to define transcriptional patterns globally present in the data. Next, tableDS() identifies, for each DSG, the cluster(s) their DETs belong to and labels gene transcripts as major (here defined as the isoform with the highest total expression across conditions) or minor isoforms. This information is used to create a classification table that indicates the distribution of DETs of DSG across different clusters. By evaluating the classification table with the cluster profiles, the user can readily identify genes with strong or subtle expression differences among their set of isoforms.

2.2.2 PodiumChange()

This function returns DSGs that undergo a switch of their most expressed isoform during the time course. PodiumChange() can be applied taking into consideration only DETs or all isoforms of DSGs. This last option is interesting when the DSG has only one isoform called as DET. The function takes as input the result of getDS() and returns a list of genes with podium changes. The function can detect changes at any time point (eventual changes), for an indicated experimental condition or at specific subranges of time and experimental conditions. Finally an isoform-resolved expression profile graph of genes with podium changes can be plotted with the IsoPlot() function to reveal the switch among isoforms.

2.2.3 IsoPlot()

This function provides gene-level plots of the expression profiles of all transcripts in the input genes. Optionally, the user can choose to visualize all transcripts or only DETs of the selected genes. Typically, IsoPlot() will be used to inspect specific genes identified by the PodiumChange() or the tableDS() functions.

References

[1] Vuong, C.K. et al. (2016) The neurogenetics of alternative splicing. Nat. Rev. Neurosci., 17, 265-81.

[2] AlShareef, S. et al. (2017) Herboxidiene triggers splicing repression and abiotic stress responses in plants. BMC Genomics, 18(1), 260.

[3] Steijger, T. et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods, 10, 1177-1184.

[4] Au, K.F. et al. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. Proc. Natl. Acad. Sci. USA, 110, 50, E4821-30.
[5] Andres, S. et al. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research, 22*, 2008-2017.

[6] Matthew, E. et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res., 47*, e47.

[7] Huang, Y., and Sanguinetti, G. (2016) Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics, 32*, 2965-72.

[8] Topa, H. and Honkela, A. (2016) Analysis of differential splicing suggests different modes of short-term splicing regulation. *Bioinformatics, 32*, i147i155.

[9] Conesa, A. et al. (2006) maSigPro: a Method to Identify Significantly Differential Expression Profiles in Time-Course Microarray Experiments. *Bioinformatics, 22*, 9, 1096-1102.

[10] Nueda, M.J. et al. (2014) Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics, 30*, 18, 2598-602.

[11] McCullagh, P., and Nelder, J.A. (1989) Generalized linear models. Chapman & Hall/CRC, Boca Ratón, Florida, 2nd edition.

[12] Wood, S.N. (2006) Generalized additive models. An introduction with R. Chapman & Hall/CRC.