BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 2 2010, pages 268–269
doi:10.1093/bioinformatics/btp643

Gene expression

MADS+: discovery of differential splicing events from Affymetrix exon junction array data

Shihao Shen[1], Claude C. Warzecha[2], Russ P. Carstens[2,3] and Yi Xing[1,4,5,*]

[1]Department of Biostatistics, University of Iowa, Iowa City, IA, [2]Cell and Molecular Biology Graduate Group, [3]Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, [4]Department of Internal Medicine and [5]Department of Biomedical Engineering, University of Iowa, Iowa City, IA, USA

Received on August 29, 2009; revised on October 20, 2009; accepted on November 10, 2009

Advance Access publication November 17, 2009

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The Affymetrix Human Exon Junction Array is a newly designed high-density exon-sensitive microarray for global analysis of alternative splicing. Contrary to the Affymetrix exon 1.0 array, which only contains four probes per exon and no probes for exon–exon junctions, this new junction array averages eight probes per probeset targeting all exons and exon–exon junctions observed in the human mRNA/EST transcripts, representing a significant increase in the probe density for alternative splicing events. Here, we present MADS+, a computational pipeline to detect differential splicing events from the Affymetrix exon junction array data. For each alternative splicing event, MADS+ evaluates the signals of probes targeting competing transcript isoforms to identify exons or splice sites with different levels of transcript inclusion between two sample groups. MADS+ is used routinely in our analysis of Affymetrix exon junction arrays and has a high accuracy in detecting differential splicing events. For example, in a study of the novel epithelial-specific splicing regulator ESRP1, MADS+ detects hundreds of exons whose inclusion levels are dependent on ESRP1, with a RT-PCR validation rate of 88.5% (153 validated out of 173 tested).

Availability: MADS+ scripts, documentations and annotation files are available at http://www.medicine.uiowa.edu/Labs/Xing/MADSplus/.

Contact: yi-xing@uiowa.edu

1 INTRODUCTION

Alternative splicing of precursor mRNAs is a prevalent regulatory mechanism in higher eukaryotes. Recent data suggest that over 90% of multi-exon human genes are alternatively spliced (Blencowe *et al.*, 2009). Use of splicing-sensitive microarrays is a popular approach for global profiling of alternative splicing events. One widely used commercial platform is the Affymetrix exon 1.0 array (Clark *et al.*, 2007). This array allocates on average four probes per exon for all transcript-confirmed and predicted exons in a mammalian genome, allowing for concurrent profiling of gene expression levels and alternative splicing events. Since its release, a variety of algorithms for splicing analysis using the exon 1.0 array have been proposed [for a summary of these algorithms, see (Laajala *et al.*, 2009)].

However, as a first-generation exon array platform, the design of the Affymetrix exon 1.0 array has its limitations. Most importantly, the array has only four probes per exon and no probes targeting exon–exon junctions. To address these limitations, Affymetrix recently released its second-generation exon array, the Human Exon Junction Array (HJAY), with a significantly improved probe design. This new array averages eight probes per probeset for 315 137 exons and 260 488 exon–exon junctions in the human genome, covering all alternative splicing events with mRNA/EST evidence in the UCSC/Ensembl databases (Yamamoto *et al.*, 2009). The increased probe density, inclusion of exon–exon junction probes and comprehensive coverage of known alternative splicing events make the HJAY a powerful platform for analysis of alternative splicing. Building on our previous work on the exon 1.0 array (Xing *et al.*, 2008), we have developed a pipeline, MADS+, for splicing analysis using the new Affymetrix exon junction array.

2 DESCRIPTIONS

MADS+ is implemented in Python/R and can be run on any Linux/Unix-like environment with installation of Python, R and required packages. An overview of the MADS+ pipeline is provided below:

Preprocessing: raw junction array CEL files are processed by the ProbeEffects program (Kapur *et al.*, 2007). A variety of background-correction and normalization methods are implemented in ProbeEffects. In particular, we employ a linear model with 80 parameters to correct for the background intensities of individual probes based on nucleotides at each position of a 25mer probe (Kapur *et al.*, 2007). The background-corrected intensities are used for downstream analysis.

Expression Index Calculation: for each gene, the background corrected and normalized intensities of its exon probes are used to derive an index for overall gene expression levels. The ProbeSelect program implements an iterative probe selection algorithm to select a subset of exon probes for each gene with highly correlated intensities across all samples (Xing *et al.*, 2006). These probes are treated as reliable indicators of overall gene expression levels and their probe intensities are fitted to the Li-Wong model to calculate expression indexes (Li *et al.*, 2001; Xing *et al.*, 2006).

Alternative Splicing Analysis: the alternative splicing analysis in MADS+ consists of several steps. First, for each probe in a

[*]To whom correspondence should be addressed.

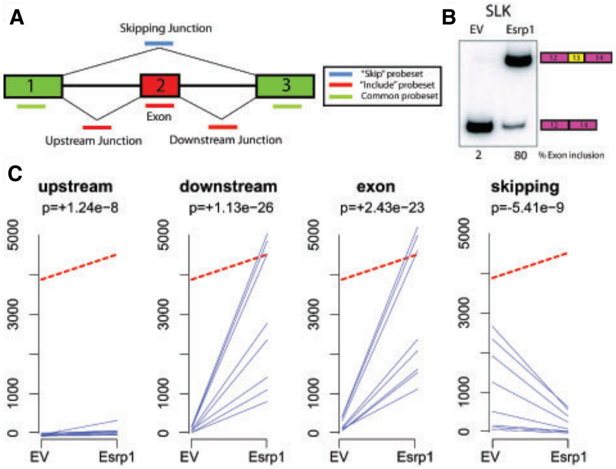

Fig. 1. Detection of a differentially spliced cassette exon using MADS+. (**A**) The HJAY probe design for cassette exons. Each probeset has eight probes. (**B**) Exon 13 of *SLK* is positively regulated by ESRP1. (**C**) The HJAY data indicate significantly elevated exon inclusion of *SLK* exon 13 in response to ESRP1 overexpression. In each panel, the dashed red line indicates the average gene expression levels of *SLK* in four replicate samples transduced with EV or ESRP1. The individual blue lines indicate the average background-corrected intensities of individual probes in a probeset. MADS+ *P*-value is shown below the name of each probeset. The '+' and '−' signs indicate the direction of change in splicing indexes from EV to ESRP1.

particular sample, we calculate a 'splicing index', represented as the background corrected probe intensity divided by the estimated gene expression index. Two separate one-sided *t*-tests are performed to test if the splicing index of a probe is substantially higher in one sample group over the other group. Then, for each probeset, the *P*-values of individual probes are summarized using Fisher's method to obtain a probeset-level *P*-value for differential splicing (Xing *et al.*, 2008). Next, MADS+ reads precompiled annotation files of alternative splicing events queried by the exon junction array. For each alternative splicing event, MADS+ evaluates the *P*-values of multiple probesets targeting exons and exon–exon junctions, requiring opposite trends for probesets targeting competing isoforms as evidence of differential splicing. For example, for each exon skipping event, we require that the 'skip' probeset and at least one of the three 'include' probesets (see Fig. 1A) have *P*-values below a predefined significance cutoff (default is 0.001), and that significant 'include' and 'skip' probesets display opposite directions of change in splicing index between the two sample groups. Finally, for candidate differential splicing events, MADS+ plots the background-corrected intensities of individual probes as well as the estimated gene expression indexes, allowing users to directly inspect the evidence for splicing differences between sample groups (see Fig. 1C for an example). It also produces annotated text output grouped according to the types of alternative splicing patterns (e.g. exon skipping, alternative splice site usage, mutually exclusive exon usage). This facilitates downstream analysis of identified differential splicing events.

3 EXAMPLES

The MADS+ pipeline has been used routinely in our exon-level analysis of the Affymetrix HJAY data. Figure 1 shows a candidate differential exon skipping event identified by MADS+. In this study, in order to identify splicing events regulated by the novel epithelial-specific splicing regulator ESRP1 (Warzecha *et al.*, 2009), a cDNA for mouse ESRP1 was ectopically expressed in the MDA-MB-231 mesenchymal human breast cancer cell line by viral transduction compared with empty vector (EV) control and profiled by the Affymetrix HJAY (four biological replicates each condition). As shown in Figure 1, exon 13 of the *SLK* gene was positively regulated by ESRP1. Its exon inclusion level was 2% after transduction by EV and 80% after transduction by ESRP1. From the HJAY data of this exon, we found significantly elevated splicing indexes of the exon inclusion probesets (upstream junction, downstream junction and exon) and significantly reduced splicing indexes of the exon skipping probeset in the ESRP1 group as compared with the EV group (Fig. 1C), which was validated by RT-PCR (Fig. 1B). In sum, among ESRP1-dependent exon skipping events identified by MADS+, we obtained a RT-PCR validation rate of 88.5% (153 exons validated out of 173 tested so far, Warzecha,C.C. *et al.*, unpublished data). This example demonstrates that MADS+ is a useful tool for global analysis and visualization of differential alternative splicing.

ACKNOWLEDGEMENTS

We wish to thank Wing Hung Wong for the discussions.

Funding: National Institutes of Health (grant R01-HG004634 to Y.X.; R01-GM088809 and R01-CA093769 to R.C.); Edward Mallinckrodt, Jr. Foundation (to Y.X.); University of Iowa Institute for Clinical and Translational Science (NIH grant UL1 RR024979).

Conflict of Interest: none declared.

REFERENCES

Blencowe,B.J. *et al.* (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.*, **23**, 1379–1386.

Clark,T.A. *et al.* (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, **8**, R64.

Kapur,K. *et al.* (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol.*, **8**, R82.

Laajala,E. *et al.* (2009) Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies. *Genome Biol.*, **10**, R77.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Warzecha,C.C. *et al.* (2009) ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol. Cell*, **33**, 591–601.

Xing,Y. *et al.* (2006) Probe selection and expression index computation of affymetrix exon arrays. *PLoS ONE*, **1**, e88.

Xing,Y. *et al.* (2008) MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA*, **14**, 1470–1479.

Yamamoto,M.L. *et al.* (2009) Alternative pre-mRNA splicing switches modulate gene expression in late erythropoiesis. *Blood*, **113**, 3363–3370.