BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 9 2012, pages 1274–1275
doi:10.1093/bioinformatics/bts134

Genome analysis Advance Access publication March 19, 2012

JETTA: junction and exon toolkits for transcriptome analysis

Junhee Seok[1,2], Weihong Xu[1], Hong Gao[1], Ronald W. Davis[1] and Wenzhong Xiao[1,3,*]

[1]Stanford Genome Technology Center, 855 California Street, Palo Alto, CA 94304, [2]Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305 and [3]Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: High-throughput genome-wide studies of alternatively spliced mRNA transcripts have become increasingly important in clinical research. Consequently, easy-to-use software tools are required to process data from these studies, for example, using exon and junction arrays. Here, we introduce JETTA, an integrated software package for the calculation of gene expression indices as well as the identification and visualization of alternative splicing events. We demonstrate the software using data of human liver and muscle samples hybridized on an exon–junction array.

Availability: JETTA and its demonstrations are freely available at http://igenomed.stanford.edu/~junhee/JETTA/index.html

Contacts: wxiao1@partners.org

Received on January 6, 2012; revised on February 22, 2012; accepted on March 13, 2012

1 INTRODUCTION

Recent developments of high-throughput technologies, such as exon–junction arrays (Clark *et al.*, 2007; Xu *et al.*, 2011) and RNA-Seq (Wang *et al.*, 2008), have extended transcriptome studies in biomedical research beyond the scope of gene expression. These genomic platforms enable the genome-wide measurements of alternative splicing (AS), the process by which individual exons of the same gene are spliced to produce different isoforms of mRNA transcripts, in clinical samples and animal disease models. This, in turn, requires computational algorithms and software tools for data analyses and visualization.

We have developed an integrated software package, JETTA, that provides a one-stop solution for gene expression and AS analyses of microarray data, from raw data files to visualization. The software provides many options for array normalization, probe selection, background correction, expression index computation, AS detection and data visualization. It can also be potentially utilized for AS detections from RNA-Seq data. Here, we describe JETTA in an analysis of human liver and muscle tissues assayed on a custom exon–junction GG-H array (Xu *et al.*, 2011).

2 SOFTWARE OVERVIEW

JETTA consists of two major modules: array calculator and AS analyzer (Fig. 1A). The array calculator supports commercial and custom exon and exon–junction arrays compatible with Affymetrix oligonucleotide array design. It calculates the expression indices of genes, exons and junctions from raw probe intensities. The AS analyzer detects and visualizes AS events between conditions. In addition, JETTA takes as input pre-calculated expression indices of individual samples measured by RNA-Seq and performs AS analyses between study groups.

The array calculator computes the expression indices through the steps of background correction, normalization and summarization. For background modeling, GCBIN (Clark *et al.*, 2007) and MAT (Kapur *et al.*, 2007) are included, both of which use linear models based on probe sequences to estimate the background level. Similarly, quantile (Irizarry *et al.*, 2003) and median-scaling normalization (Kapur *et al.*, 2007) are the two options for normalization. The processed probe signals are then summarized to gene-, exon- or junction-level expression levels using either Li–Wing fitting (Li and Wong, 2001), probe selection (Xing *et al.*, 2006) or median-polish algorithms (Irizarry *et al.*, 2003).

The AS analyzer takes in expression indices from either the output of the array calculator for microarray data or similar results calculated by other tools for RNA-Seq data such as rSeq (Jiang and Wong, 2009). The analyzer estimates AS signals that include Splicing Index, i.e. fold changes of each exon and junction normalized to the gene expression level (Clark *et al.*, 2002, 2007), and the corresponding *P*-values using the algorithms of detection above background (DABG), microarray analysis of differential splicing (MADS) or microarray detection of alternative splicing (MIDAS). DABG calculates the probability of the presence of each probe set based on the non-parametric distribution of background probe intensities (Clark *et al.*, 2007), and MADS (Xing *et al.*, 2008) and MIDAS (Affymetrix white paper: Alternative Transcript Analysis Methods for Exon Arrays v1.1) calculate the *P*-value of differential expression for each exon and junction relative to the corresponding gene expression levels between two conditions. For RNA-Seq data, Splicing Index and MIDAS are calculated, as DABG and MADS are specific to arrays.

The analyzer then detects AS events by allowing custom filtering on one or a combination of the provided AS signals. For example, to reduce false positive detections, it is helpful to require each detected event being supported by the signal of the exon as well as at least one of its connecting junctions when analyzing exon–junction arrays (Xu *et al.*, 2011).

JETTA also provides the visualization of AS signals (Fig. 1B). In addition, it is integrated with cisGenome Brower (Ji *et al.*, 2008) to allow the examination of raw signals of exons and junctions. The software is implemented as an R-package as well as standalone software with graphical user interface.

*To whom correspondence should be addressed.

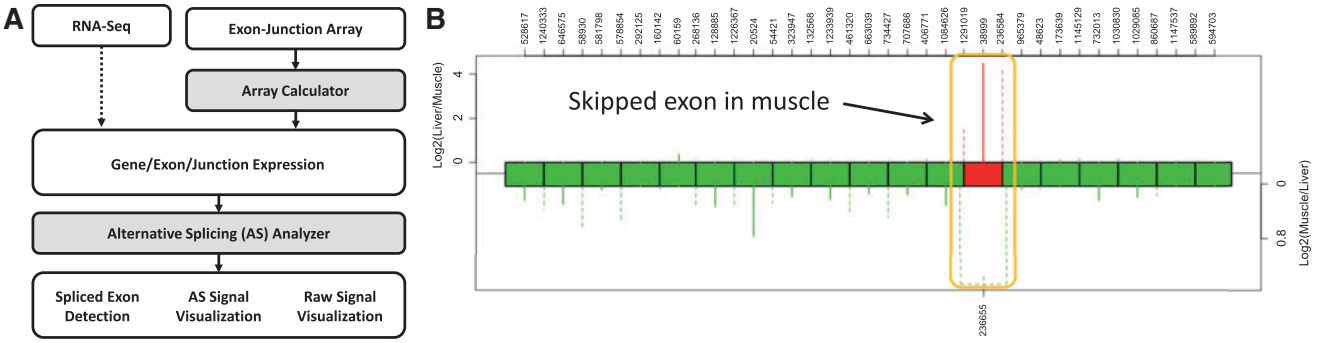

Fig. 1. (**A**) Modules of JETTA. (**B**) Visualization of AS events. Gene SLK is shown as an example. Exons are represented as blocks, and junctions are represented between exons or by bridging exons. Probeset IDs of exons (solid bars) and junctions (dotted bars) are labeled on the *x*-axis, and on the *y*-axis the corresponding logged fold changes are shown as upward and downward bars, respectively, so that the upward bars represent signals increased in liver and downward bars signals increased in muscle. The highlighted shows an AS, as supported by the increase of the exon and its two 'inclusion' junctions as well as a decrease of its 'skipping' junction in liver compared with muscle.

3 RESULTS AND DISCUSSION

JETTA is demonstrated using exon–junction array data of quadruplicates of human liver and muscle samples (Xu *et al*., 2011). Low-level data processing was performed through GCBIN correction, median-scaling normalization and median-polish summarization.

Alternatively, spliced exons between liver and muscle were then identified through the following steps. First, exon probe sets were selected with MIDAS *P*-values <0.01 and DABG *P*-values <0.01 (in at least one tissue), resulting in 13 150 candidate exons. Second, significant junction probe sets were identified that satisfied the same criteria of MIDAS and DABG *P*-values as above. Third, since exon probes alone sometimes are not sufficient for reliable analysis of splicing (Xing *et al.*, 2008), to increase the confidence of the findings, we determined among candidate exons those supported by at least one significant connecting junction that corroborated the AS signal of the exon. This gave a final result of 6461 exons in 2999 genes as alternatively spliced. As an example, Figure 1B shows the result of gene SLK.

To help experimental validations, we looked at 'skipped' exons present in one condition but absent in the other. Here, we considered only genes expressed higher than the overall median expression level and changed <2-fold between the conditions. Further, filtering by DABG *P*-values (>0.1 in the condition where the exon is less-expressed) and MADS *P*-values (<0.01) yielded 145 exons in 122 genes. Among them, GRANL1, VPS39 and SLK were tested with RT-PCR, and all were verified with large fold changes.

Similarly, we applied JETTA to an RNA-Seq dataset of the same tissue types (GSE26109) and detected 996 AS events, among which 52% were also identified by the array analysis (details described on JETTA webpage). JETTA detects alternatively spliced exons without decomposing the expression levels of previously known or reconstructed *de novo* transcript isoforms (Jiang *et al.*, 2009; Roberts *et al.*, 2011; Trapnell *et al.*, 2010; Wang *et al.*, 2003). It provides a complementary option for analysis of RNA-Seq data, especially for cross-comparing alternative spliced exons detected with results from the arrays.

In summary, JETTA is an integrated software package for expression and AS analyses and visualization. The software has been applied as a primary tool for the analyses of exon and junction arrays in an ongoing multicenter clinical study of severe trauma patients (Xu *et al.*, 2011). As high-throughput AS studies become prevalent in clinical research; we anticipate that JETTA will also be useful for the data analysis.

ACKNOWLEDGEMENTS

We thank Dr Wing H. Wong for discussions and helpful advice throughout JETTA development, and Shangping Feng and Luis Jevons for help on its implementation.

Funding: National Institutes of Health U54-GM062119, P01-HG000205 and R01-HG004634.

Conflict of Interest: none declared.

REFERENCES

Clark,T.A. *et al.* (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.

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

Irizarry,R. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

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

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

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.

Roberts,A. *et al.* (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–2329.

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.

Wang,E. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,H. *et al.* (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics,* **19** (Suppl. 1), i315–i322.

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

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

Xu,W. *et al.* (2011) Human transcriptome array for high-throughput clinical studies. *Proc. Natl Acad. Sci. USA*, **108**, 3707–3712.