BIOINFORMATICS APPLICATIONS NOTE

Vol. 29 no. 2 2013, pages 273–274
doi:10.1093/bioinformatics/bts678

Gene expression

Advance Access publication November 21, 2012

Intron-centric estimation of alternative splicing from RNA-seq data

Dmitri D. Pervouchine[1,2,3,*], David G. Knowles[1,2] and Roderic Guigó[1,2]

[1]Centre de Regulació Genòmica (CRG) and [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain and [3]Moscow State University, 119992 Moscow, Russia

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Novel technologies brought in unprecedented amounts of high-throughput sequencing data along with great challenges in their analysis and interpretation. The percent-spliced-in (PSI, Ψ) metric estimates the incidence of single-exon–skipping events and can be computed directly by counting reads that align to known or predicted splice junctions. However, the majority of human splicing events are more complex than single-exon skipping.

Results: In this short report, we present a framework that generalizes the Ψ metric to arbitrary classes of splicing events. We change the view from exon centric to intron centric and split the value of Ψ into two indices, ψ_5 and ψ_3, measuring the rate of splicing at the 5′ and 3′ end of the intron, respectively. The advantage of having two separate indices is that they deconvolute two distinct elementary acts of the splicing reaction. The completeness of splicing index is decomposed in a similar way. This framework is implemented as `bam2ssj`, a BAM-file–processing pipeline for strand-specific counting of reads that align to splice junctions or overlap with splice sites. It can be used as a consistent protocol for quantifying splice junctions from RNA-seq data because no such standard procedure currently exists.

Availability: The C^{++} code of `bam2ssj` is open source and is available at https://github.com/pervouchine/bam2ssj

Contact: dp@crg.eu

Received on August 23, 2012; revised on October 29, 2012; accepted on November 15, 2012

1 INTRODUCTION

One major challenge in the analysis of high-throughput RNA sequencing data is to disentangle relative abundances of alternatively spliced transcripts. Many existing quantification methods do so by using considerations of likelihood, parsimony and optimality to obtain a consolidated view of cDNA fragments that map to a given transcriptional unit (Katz *et al.*, 2010; Montgomery *et al.*, 2010; Trapnell *et al.*, 2012). The advantage of such integrative approaches is that they provide robust estimators for transcript abundance by reducing sampling errors, as they effectively consider samples of larger size. In contrast, because all the reads from the same transcriptional unit are combined into one master model, there is no guarantee that the inclusion or exclusion of a specific exon is estimated independently of co-occurring splicing events (Katz *et al.*, 2010; Pan *et al.*, 2008).

The quantification of alternatively spliced isoforms based on the Ψ metric captures more accurately the local information related to splicing of each particular exon (Katz *et al.*, 2010). We follow Kakaradov *et al.* (2012) in considering only the reads that align to splice junctions (Fig. 1) and ignoring the reads that align to exon bodies (position-specific read counts are not considered). Ψ is defined as

$$\Psi = \frac{a+b}{a+b+2c} \tag{1}$$

where the factor of two in the denominator accounts for the fact that there are twice as many mappable positions for reads supporting exon inclusion as exon exclusion. Equation (1) defines an unbiased estimator for the fraction of mRNAs that represent the inclusion isoform under the assumption that splice-junction reads are distributed evenly. Ψ can also be derived from the expression values of whole isoforms, for instance, as the abundance of the inclusion isoform as the fraction of the total abundance. However, the non-uniform read coverage not only between but also within transcripts makes such estimates generally detrimental (Kakaradov *et al.*, 2012).

The Ψ metric can be generalized beyond the class of single-exon–skipping events by counting inclusion and exclusion reads regardless of exon adjacency (Fig. 1, dashed arcs). Although this definition helps to reduce the undercoverage bias by taking into account splice junctions that are not present in the reference annotation, it often assigns misleading values to Ψ metric, for instance, in the case of multiple-exon skipping, where the amount of support for exon exclusion does not reflect the true splicing rate of each individual intron.

2 APPROACH

In this work, we change the view from exon centric to intron centric. Each intron is defined uniquely by the combination of its 5′-splice site (D, donor) and 3′-splice site (A, acceptor). Denote by $n(D,A)$ the number of reads aligning to the splice junction spanning from D to A (Fig. 2) and define

$$\psi_5(D,A) = \frac{\mathrm{n}(D,A)}{\sum\limits_{A'} \mathrm{n}(D,A')} \text{ and } \psi_3(D,A) = \frac{\mathrm{n}(D,A)}{\sum\limits_{D'} \mathrm{n}(D',A)}, \tag{2}$$

where D' and A' run over all donor and acceptor sites, respectively, within the given genomic annotation set. Because A' could be A and D' could be D, both $\psi_5(D,A)$ and $\psi_3(D,A)$ are real numbers from 0 to 1. The value of $\psi_5(D,A)$ can be regarded as

[*] To whom correspondence should be addressed.

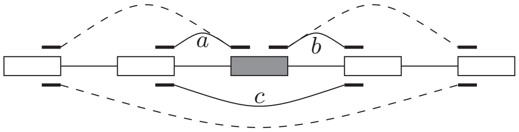

Fig. 1. The percent-spliced-in (PSI, Ψ) metric is defined as the number of reads supporting exon inclusion ($a + b$) as the fraction of the combined number of reads supporting inclusion and exclusion (c). The exon of interest is shown in gray. Only reads that span to the adjacent exons (solid arcs) account for Equation (1)

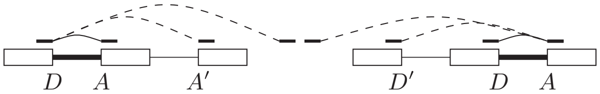

Fig. 2. Left: the 5′-splicing index, ψ_5, is the number of reads supporting the splicing event from D to A relative to the combined number of reads supporting splicing from D to any acceptor site A'. Right: the 3′-splicing index, ψ_3, is the number of reads supporting the splicing event from D to A relative to the combined number of reads supporting splicing from any donor site D' to A. The intron of interest is drawn thick

an estimator for the conditional probability of splicing from D to A, i.e. the fraction of transcripts in which the intron D to A is spliced, relative to the number of transcripts in which D is used as a splice site. Similarly, $\psi_3(D, A)$ is the relative frequency of D-to-A splicing with respect to the splicing events in which A is used.

In the particular case of single-exon skipping (Fig. 1), the values of Ψ, ψ_5 and ψ_3 are related as follows. Denote the upstream and downstream introns of the highlighted exon by (D_1, A_1) and (D_2, A_2), respectively. Let $\psi_5 = \psi_5(D_1, A_1)$ and $\psi_3 = \psi_3(D_2, A_2)$. Then, $\psi_5 = \frac{a}{a+c}$, $\psi_3 = \frac{b}{b+c}$ and $\Psi = \omega_5\psi_5 + \omega_3\psi_3$, where $\omega_5 = \frac{a+c}{a+b+2c}$ and $\omega_3 = \frac{b+c}{a+b+2c}$. Assuming uniform read coverage across the gene ($a \simeq b$), we get $\omega_5 \simeq \omega_3 \simeq \frac{1}{2}$ and, therefore,

$$\Psi \simeq \frac{\psi_5 + \psi_3}{2}. \qquad (3)$$

That is, in the particular case of single-exon skipping, the value of Ψ is equal to the average of ψ_5 and ψ_3 given that the read coverage is reasonably uniform. If a and b differ significantly, the contribution of ψ_5 and ψ_3 to Ψ is given by the weight factors ω_5 and ω_3.

Similarly, the completeness of splicing index (Tilgner *et al.*, 2012) is split into two indices, $\theta_5(D)$ and $\theta_3(A)$, where

$$\theta_5 = \frac{\sum\limits_{A'} \mathrm{n}(D, A')}{\sum\limits_{A'} \mathrm{n}(D, A') + \mathrm{n}(D)}, \quad \theta_3 = \frac{\sum\limits_{D'} \mathrm{n}(D', A)}{\sum\limits_{D'} \mathrm{n}(D', A) + \mathrm{n}(A)}, \qquad (4)$$

and $n(X)$ denotes the number of genomic reads (reads mapped uniquely to the genomic sequence) overlapping the splice site X. Note that θ_5 depends only on D and θ_3 depends only on A. The values of θ_5 and θ_3 are unbiased estimators for the absolute frequency of splice site usage, i.e. the proportion of transcripts in which D (or A) is used as a splice site, among all transcripts containing the splice site D (or A).

3 METHODS

To compute ψ_5, ψ_3, θ_5 and θ_3 for a given donor–acceptor pair, one needs to know five integers, $n(D, A)$, $\sum n(D, A')$, $\sum n(D', A)$, $n(D)$ and $n(A)$, of which only the first one depends on both D and A, while the rest have a single argument. We developed `bam2ssj`, a pipeline for counting these five integers directly from BAM input. `bam2ssj` is implemented in C^{++} and depends on SAMtools (Li *et al.*, 2009). The input consists of (i) a sorted BAM file containing reads that align uniquely to the genome or to splice junctions and (ii) a sorted GTF file containing the coordinates of exon boundaries. Each time the CIGAR string (Li *et al.*, 2009) contains xMyNzM, $x, z \geq 1$, the counter corresponding to the splice junction defined by yN is incremented. One mapped read may span several splice junctions and increment several counters. If the CIGAR string does not contain the xMyNzM pattern, the read is classified as genomic and increments $n(X)$ for every splice site X it overlaps. Position-specific counts (Kakaradov *et al.*, 2012) are implemented as a stand-alone utility that is not included in the current distribution. Importantly, `bam2ssj` counts reads that align to splice junctions in a strand-specific way, i.e. $n(D, A)$, $\sum n(D, A')$, $\sum n(D', A)$, $n(D)$ and $n(A)$ are reported for the correct (annotated) and incorrect (opposite to annotated) strand. We leave further processing of these counts by Equations (2)–(4) to the user.

4 RESULTS AND DISCUSSION

We validated `bam2ssj` by counting reads aligning to splice junctions in the whole-cell polyadenylated fraction of Cold Spring Harbor Long RNA-seq data (http://genome.ucsc.edu/ENCODE/). In total, 8 558 231 343 mapped reads were analyzed in 404 min ($\simeq$350 000 reads/sec). 1 184 553 724 reads align to splice junctions, of which $\simeq$1% align to the opposite strand. 1 699 718 327 reads overlap annotated splice junctions, of which $\simeq$5% map to the opposite strand. The values of $n(D, A)$ coincide with those reported by ENCODE in 98.9% of cases (1 163 251 008 reads); all discrepancies were due to the ambiguity of CIGAR translation in the mapper's output. Because RNA-seq data are increasingly processed into the compact BAM form, we propose that `bam2ssj` be used as a standard operating procedure for counting splice junction reads.

Funding: Grants BIO2011-26205 and CSD2007-00050 Consolider, Ministerio de Educación y Ciencia (Spain).

Conflict of Interest: none declared.

REFERENCES

Kakaradov,B. *et al.* (2012) Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. *BMC Bioinformatics*, **13 (Suppl. 6)**, S11.

Katz,Y. *et al.* (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Montgomery,S. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

Tilgner,H. *et al.* (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, **22**, 1616–1625.

Trapnell,C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.