Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

Short technical report

SLaP mapper: A webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes

Michael Fiebig[a], Eva Gluenz[a], Mark Carrington[b], Steven Kelly[c,*]

[a] Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK
[b] Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK
[c] Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

ARTICLE INFO

Article history:
Received 14 June 2014
Received in revised form 30 July 2014
Accepted 31 July 2014
Available online 8 August 2014

Keywords:
RNA splicing
Trypanosoma
Leishmania
Polyadenylation
RNA-Seq
Kinetoplastida

ABSTRACT

The Kinetoplastida are a diverse and globally distributed class of free-living and parasitic single-celled eukaryotes that collectively cause a significant burden on human health and welfare. In kinetoplastids individual genes do not have promoters, but rather all genes are arranged downstream of a small number of RNA polymerase II transcription initiation sites and are thus transcribed in polycistronic gene clusters. Production of individual mRNAs from this continuous transcript occurs co-transcriptionally by trans-splicing of a ~39 nucleotide capped RNA and subsequent polyadenylation of the upstream mRNA. SLaP mapper (**S**pliced-**L**eader **a**nd **P**olyadenylation mapper) is a fully automated web-service for identification, quantitation and gene-assignment of both spliced-leader and polyadenylation addition sites in Kinetoplastid genomes. SLaP mapper only requires raw read data from paired-end Illumina RNAseq and performs all read processing, mapping, quality control, quantification, and analysis in a fully automated pipeline. To provide usage examples and estimates of the quantity of sequence data required we use RNAseq obtained from two different library preparations from both *Trypanosoma brucei* and *Leishmania mexicana* to show the number of expected reads that are obtained from each preparation type. SLaP mapper is an easy to use, platform independent webserver that is freely available for use at http://www.stevekellylab.com/software/slap. Example files are provided on the website.

1. Introduction

The Kinetoplastida are a diverse and globally distributed group of free-living and parasitic single-celled eukaryotes. In kinetoplastids messenger RNAs are produced by co-transcriptional processing of continuously transcribed polycistronic gene clusters [1]. Co-transcriptional processing occurs via trans-splicing of a ~39 nucleotide 5′-capped spliced-leader sequence and 3′ polyadenylation of the upstream gene [2–5]. Trans-splicing occurs predominantly at AG dinucleotides, however no canonical nucleotide motif has been identified for polyadenylation sites. Moreover, the AAUAAA motif found at polyadenylation sites in most other eukaryotes [6] is not present in kinetoplastids [7].

Several tools have been developed that predict trans-splice acceptor and polyadenylation sites [8–10], however, these tools do not predict relative site usage. Current sequencing technology now makes it possible to determine these sites empirically on a genome-wide scale and quantify the extent to which different trans-splice and polyadenylation sites are used. RNA-sequencing studies of *Trypanosoma brucei* and *Leishmania major* have already begun to reveal the large extent to which individual genes can harbour multiple trans-splice and polyadenylation sites [11,12]. To capitalise on this technological advancement, and enable widespread analysis of trans-splice and polyadenylation sites within the community, we have developed a fully automated web-service called SLaP Mapper. This server only requires raw read data obtained by paired-end Illumina RNASeq, and uses this raw read data to identify and quantify trans-splice-acceptor and polyadenylation sites genome-wide.

2. Materials and methods

2.1. Differences in library construction

Typical libraries generated from random hexamer primed cDNA are suitable for identification of splice acceptor sites (Table 1). However, the extent to which polyadenylation sites are discovered

* Corresponding author. Tel.: +44 01865 275123.
E-mail address: steven.kelly@plants.ox.ac.uk (S. Kelly).

http://dx.doi.org/10.1016/j.molbiopara.2014.07.012
0166-6851/© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

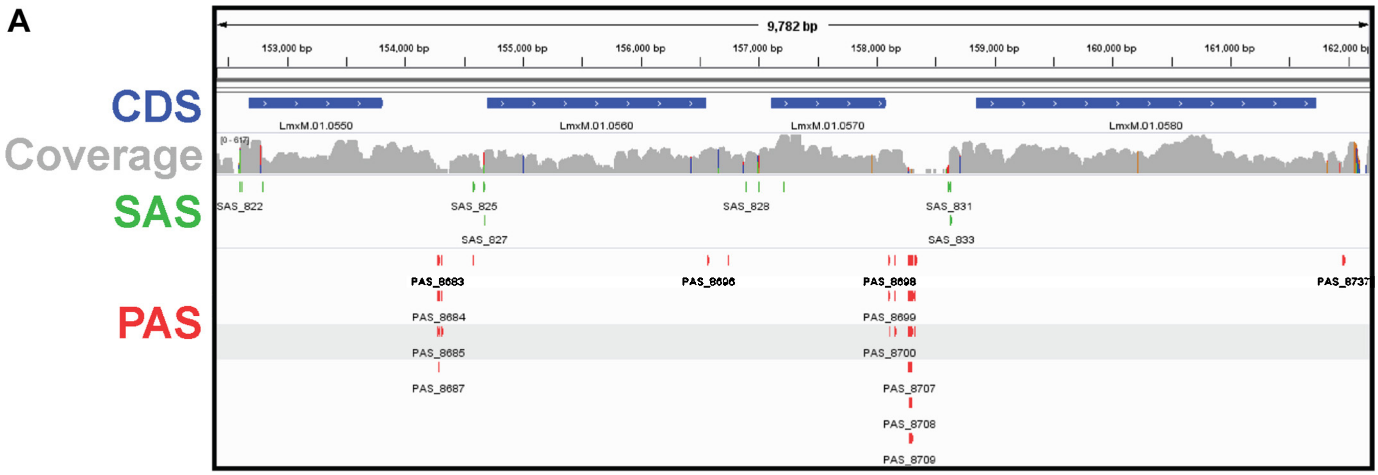
Table 1
The number of reads observed using different library preparation methods in two different species. *L. mexicana* based on 3 independent biological replicates. *T. brucei* based on 2 independent biological replicates. Numbers in brackets indicate one standard deviation.

Species	Library type	Poly(A) reads per million reads	Trans-splice reads per million reads
L. mexicana	T15VN	34 784 (2100)	4343 (300)
L. mexicana	Random primed	5673 (730)	79 773 (10 000)
T. brucei	T15VN	167 864 (8000)	4782 (250)
T. brucei	Random primed	701 (50)	76 563 (4800)

depends on the library preparation protocol that is used (Table 1). To generate a library enriched for poly(A) containing reads it is recommended that the first strand cDNA synthesis reaction is primed with a 5′-T15VN-3′ oligonucleotide (V = A, G or C; N = T, A, G or C) [11], followed by second strand synthesis with random hexamer primers.

2.2. Algorithm overview

SLaP mapper uses pre-built indices for the currently available kinetoplastid genomes. The user must supply raw sequence reads in gzipped fastq format (phred encoding offset is automatically detected). Due to FTP limitations SLaP mapper can only accept individual files less than 2 GB. Read files larger then 2 GB can be analysed using SLAP mapper by splitting these files into pieces each smaller than 2 GB then combining the results files. Once read files have been uploaded, reads containing a putative poly(A) tail or spliced-leader sequence are identified, the spliced-leader or poly(A) tail is removed from the read and the rest of the read is mapped to the user specified genome. Each putative identified splice-acceptor site is checked to confirm that it does not contain the spliced-leader sequence and *bona fide* splice-acceptor sites are assigned to their nearest directionally appropriate coding sequence (CDS). Similarly, putative poly(A) addition sites are checked that they do not encode runs of A residues and *bona fide* sites are assigned to their nearest directionally appropriate CDS. When the analysis is complete the results are emailed to the user in tab-delimited text, BED and GFF file formats so that the results are easily viewed in spreadsheet editors or viewed on commonly used genome browsers such as the Integrative genomics viewer [13] (Fig. 1A). The results files contain the position of the observed site, its dinucleotide (for trans-splice sites only), its strand, its



B

Chromosome	Position	Dinucleotide	Site-strand	Closest CDS	CDS strand	Distance to CDS	Status	Count
LmxM.01	152602	AG	+	LmxM.01.0550	+	82	OK	154
LmxM.01	152617	AG	+	LmxM.01.0550	+	67	OK	2
LmxM.01	152796	GG	+	LmxM.01.0550	+	0	Internal	1
LmxM.01	154583	AG	+	LmxM.01.0560	+	120	OK	2
LmxM.01	154674	AG	+	LmxM.01.0560	+	29	OK	19
LmxM.01	154679	AG	+	LmxM.01.0560	+	24	OK	3
LmxM.01	156898	AG	+	LmxM.01.0570	+	211	OK	1
LmxM.01	157003	AG	+	LmxM.01.0570	+	106	OK	31
LmxM.01	157214	AG	+	LmxM.01.0570	+	0	Internal	1
LmxM.01	158612	AG	+	LmxM.01.0580	+	243	OK	21
LmxM.01	158625	AG	+	LmxM.01.0580	+	230	OK	1
LmxM.01	158631	AG	+	LmxM.01.0580	+	224	OK	4

Fig. 1. (A) Screen shot of SLaP mapper results as visualised on the IGV genome browser. Four data tracks are shown. CDS are the gene models from V6 of the *L. mexicana* genome. Coverage is the from raw RNAseq reads mapped to the *L. mexicana* V6 genome (the coloured lines in the coverage plot indicate single nucleotide polymorphisms between the genome reference and the strain used for RNAseq). SAS are the splice acceptor addition sites identified by SLaP mapper. PAS are the polyadenylation addition sites identified by SLaP mapper. (B) The corresponding entries in the SLaP mapper results file for all SAS sites shown in A. The poly(A) results for the 55 sites shown in part A are not listed for space reasons. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

occurrence (i.e. the number of mapped reads) and the nearest directionally appropriate CDS (Fig. 1B). A summary of the mapping and filtration processes and the settings used to perform these steps is also provided in the results package that is emailed to the user.

SLaP mapper is an analysis pipeline and uses a number of freely available programs and custom written Perl scripts. The algorithm proceeds in five phases.

(1) **Read quality control**. This step uses the Trimmomatic read processing tool [14] to remove known Illumina adaptor sequences and to trim reads based on quality scores. At this step read-pairs are also assessed for overlapping segments and those reads that overlap in the centre are joined using the fastq-join utility [15].

(2) **Identification, preparation and mapping of spliced-leader and poly(A) containing reads**. Here all reads are treated individually and scanned for the presence of poly(A) tails or the appropriate species-specific spliced-leader sequence. A spliced-leader containing read is defined as a read containing at least 12 nucleotides of the 3′ end of the spliced-leader sequence, this minimum length is changeable by the user. A poly(A) containing read is defined as a read which ends in 5 or more A residues (reads are treated as un-stranded and scanned on both strands). The minimum poly(A) length is also specifiable by the user. Spliced-leader reads are split at the splice junction and the non-spliced-leader part of the read is mapped to the selected reference genome. Similarly the poly(A) read is split at the run of A residues and the non-poly(A) tail part of the read is mapped to the selected reference genome. Read mapping is performed using bowtie2 [16].

(3) **Filtering of putative splice-acceptor and poly(A) reads**. Mapped putative splice-junction reads are checked to ensure that the location in the genome does not encode the 12 bases of the splice acceptor. Similarly mapped putative poly(A) tail reads are checked to ensure that the location in the genome does not contain an analogous run of A residues as was present in the read. Only *bona fide* splice-junction and polyadenylation addition sites are retained for further analysis. The option to disable this poly(A) site filter is provided on the webserver.

(4) **Assigning sites to genes**. Once reads have been mapped the location of the sites is recorded and they are assigned to CDS according to the following rules. Trans-splice sites are assigned to a CDS if they occur on the same strand as the CDS and downstream of the stop codon of the preceding gene and upstream of the stop codon of the gene in question. Polyadenylation sites are assigned to a CDS if they lie on the same strand as the CDS, downstream of the start codon of the CDS and upstream of the start codon of the next downstream CDS. Splice leader addition sites and poly(A) sites that occur within CDS are assigned to the CDS in which they reside. It should be noted here that trypanosomatid genomes contain a number of stable transcripts lacking CDSs that occur between true mRNAs. Thus it is possible that some sites belonging to non-coding transcripts may be incorrectly annotated to CDSs by this method. For this reason a separate file is also provided that lists the identified sites without assigning them to CDS.

(5) **Quantification**. Sites are quantified as the number of reads which uniquely map to each site location.

3. Discussion and conclusions

SLaP mapper is a simple to use resource that enables users to identify and quantify trans-splice and polyadenylation sites in kinetoplastid genomes. It is the only such software of its kind and it requires only a web-browser and no specialised knowledge of any programming environment. The user only need select the appropriate species and upload unprocessed read files. We describe the expected number of informative reads per-million reads that are obtained using two different library preparation protocols in two different species (Table 1). This shows that relatively little sequence data (<10 million reads) is required to provide a comprehensive genome-wide analysis of site usage.

Recent RNA-sequencing studies of *T. brucei* and *L. major* have revealed that many genes can harbour multiple alternative processing sites [11,12]. While the functional significance of these sites has yet to be determined on a genome wide scale, it is likely that some of these alternative sites are important to the regulation and/or function of the final transcript. For example, alternative use of two different spliced-leader addition sites in *T. brucei* facilitates the dual localisation of an isoleucyl-tRNA synthetase [17]. In this case the alternative processing sites either include or exclude a mitochondrial localisation signal from the N-terminus of the final polypeptide. SLaP mapper can be readily used to detect such alternative processing sites for transcripts (for example see Fig. 1).

In addition to providing a resource that will facilitate the annotation of novel kinetoplastid genomes, this server can also be used to quantify differences in splice-acceptor and polyadenylation site usage across a range of species. This is useful for comparative gene expression studies and in the analysis of post-transcriptional processing of kinetoplastid mRNA. Future releases of SLaP mapper will include more kinetoplastid genomes as they become available.

Acknowledgements

MF is supported by the Wellcome Trust. SK is a Leverhulme Trust Early Career Fellow. EG is a Royal Society University Research Fellow. MC is supported by the Wellcome trust 085256/Z/08/Z. The authors would like to thank the two anonymous Referees for their advice on the manuscript.

References

[1] Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. Science 2005;309:436–42.

[2] Parsons M, Nelson RG, Watkins KP, Agabian N. Trypanosome mRNAs share a common 5′ spliced leader sequence. Cell 1984;38:309–16.

[3] Ullu E, Matthews KR, Tschudi C. Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. Mol Cell Biol 1993;13:720–5.

[4] Sutton RE, Boothroyd JC. Evidence for trans splicing in trypanosomes. Cell 1986;47:527–35.

[5] LeBowitz JH, Smith HQ, Rusche L, Beverley SM. Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev 1993;7:996–1007.

[6] Wickens M. How the messenger got its tail: addition of poly(A) in the nucleus. Trends Biochem Sci 1990;15:277–81.

[7] Schurch N, Hehl A, Vassella E, Braun R, Roditi I. Accurate polyadenylation of procyclin mRNAs in *Trypanosoma brucei* is determined by pyrimidine-rich elements in the intergenic regions. Mol Cell Biol 1994;14:3668–75.

[8] Kelly S, Wickstead B, Maini PK, Gull K. Ab initio identification of novel regulatory elements in the genome of *Trypanosoma brucei* by Bayesian inference on sequence segmentation. PLoS ONE 2011;6:e25666.

[9] Siegel TN, Tan KS, Cross GA. Systematic study of sequence motifs for RNA trans splicing in *Trypanosoma brucei*. Mol Cell Biol 2005;25:9586–94.

[10] Gopal S, Awadalla S, Gaasterland T, Cross GA. A computational investigation of kinetoplastid trans-splicing. Genome Biol 2005;6:R95.

[11] Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi, C, et al. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 2010;6.

[12] Rastrojo A, Carrasco-Ramiro F, Martin D, Crespillo A, Reguera RM, Aguado C, et al. The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. BMC Genomics 2013;14:223.

[13] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.

[14] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014.

[15] Aronsky E. Command-line tools for processing biological sequencing data; 2011 http://codegooglecom/p/ea-utils

[16] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.

[17] Rettig J, Wang Y, Schneider A, Ochsenreiter T. Dual targeting of isoleucyl-tRNA synthetase in *Trypanosoma brucei* is mediated through alternative trans-splicing. Nucleic Acids Res 2012;40:1299–306.