Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W297–W299***
doi:10.1093/nar/gkm311

ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets

Sylvain Foissac* and Michael Sammeth

Centre de Regulació Genòmica, Barcelona, Spain

Received January 31, 2007; Revised April 6, 2007; Accepted April 14, 2007

ABSTRACT

In the process of establishing more and more complete annotations of eukaryotic genomes, a constantly growing number of alternative splicing (AS) events has been reported over the last decade. Consequently, the increasing transcript coverage also revealed the real complexity of some variations in the exon–intron structure between transcript variants and the need for computational tools to address 'complex' AS events. ASTALAVISTA (alternative splicing transcriptional landscape visualization tool) employs an intuitive and complete notation system to univocally identify such events. The method extracts AS events dynamically from custom gene annotations, classifies them into groups of common types and visualizes a comprehensive picture of the resulting AS landscape. Thus, ASTALAVISTA can characterize AS for whole transcriptome data from reference annotations (GENCODE, REFSEQ, ENSEMBL) as well as for genes selected by the user according to common functional/structural attributes of interest: http://genome.imim.es/astalavista

INTRODUCTION

Alternative splicing (AS) is a fundamental cellular process involved in eukaryotic gene expression (1–3). To decipher the molecular mechanisms responsible for AS, several computational studies have been presented over the last years, producing a considerable quantity of dedicated analyses and databases representing the transcript diversity resulting from AS events (4,5).

In contrast to the global transcript diversity, it is also possible to identify local variations observed in the exon–intron structures amongst the transcripts. For the investigation of the molecular mechanisms giving rise to AS, the separation of events in different classes has shown promise. Historically, four main AS events have been reported in literature: exon skipping, intron retention, alternative donor and acceptor splice sites. Obviously, the frequent observation of these 'simple' events is correlated with the fact that they involve the minimum number of variable splice sites (i.e. two).

However, current annotation datasets show a plethora of more complex variations that can be seen as variable combinations of simple events overlapping each other, indicating connections in regulation and function. Only recently efforts have been undertaken in order to properly identify and describe such 'complex' AS events. In previous work, 'bit matrices' identifying exonic and intronic segments of transcripts have been used to generally describe AS events (6). Alternatively, we have developed an intuitive notation system based on the relative position of alternatively used splice sites in order to univocally identify any possible AS event (Sammeth *et al.*, submitted).

In the process of investigating the phenomenon of AS, many web resources have already been made available (7–12). In general, these tools can be considered as AS-dedicated gene or genome browsers, suitable to access much information about each gene of interest but not convenient for comprehensive analyses of AS across genes. Moreover, none of them propose an exhaustive identification of AS events from custom input data.

Herein, we describe the ASTALAVISTA web server (alternative splicing transcriptional landscape visualization tool) that allows to dynamically identify, extract and display complex AS events from annotated genes. ASTALAVISTA gives the opportunity to investigate and compare types and distributions of the different AS events found in the input—whole genome annotations as well as user provided gene sets. To our knowledge, this is the first time a tool for the exhaustive extraction of AS events from custom datasets is made publicly available.

METHODS

ASTALAVISTA adopts a generic definition of AS events and a flexible notation system assigning a code based on

*To whom correspondence should be addressed. Tel: +34 933 160 166; Fax: +34 933 969 983; Email: sylvain.foissac@crg.es

the relative position of alternatively used splice sites. In brief, given a set of annotated transcripts, the method consists in first considering all pairwise comparisons between overlapping transcripts. A variation of the splicing structure is detected if some splice sites are not used in both transcripts. Then, according to the genomic coordinates, the relative order of the splice sites that are included in such variations is used to build a code describing the corresponding AS event. This approach overcomes limitations of methods focusing exclusively on simple events and circumvents the problem of chosing a reference transcript, defining a 'main' splice form to be compared with. The intrinsic transcript clustering prevents the method from being dependent on the assignment of transcripts to a certain gene name or locus. Furthermore, a redundancy filtering is applied in order to identify the list of unique AS events, regardless of how many transcript comparisons exhibit the same splicing variation. The genericity of the notation based on relative splice site positions enables to compare AS events across different genes, chromosomes or genomes. By this, events describing equal variations in the exon–intron structures are pooled in a common group. Finally, the distribution of AS events across these groups is used to depict the AS landscape in a dataset. More details about the method and the resulting 'AS code' are provided on the web site.

WEB SITE DESCRIPTION

Input: annotation datasets

ASTALAVISTA requires a set of transcripts with known exon–intron structure (e.g. from mRNAs or ESTs). As primary data source for custom transcript sets, the genomic positions of the exon boundaries for each transcript are provided using the gene transfer format (GTF). Each GTF line has nine required fields, of which the feature (e.g. 'exon', 'CDS', etc.), start and end coordinates on a chromosome (or a contig), the strand and an identifier for the corresponding transcript are used. Note that no gene identifiers are necessary due to the intrinsic clustering of transcripts into loci. To check the GTF format requirements, an example input is available on the web server. Optionally, if any protein coding region information is provided within the input (feature 'CDS'), transcripts and/or extracted AS events may be filtered according to the annotated CDSs. This straightforwardly allows to compare the AS landscapes of coding versus non-coding transcripts or of events localized in CDSs versus UTRs.

Alternatively, the user may also analyze the anatomy of AS as characterized by popular genomic human annotations (GENCODE, REFSEQ, ENSEMBL), or just provide any set of human genes. In the latter case, the list of genes can be specified by identifiers from various nomenclature systems, like REFSEQ mRNA IDs, SWISSPROT IDs, HUGO gene symbols, ENSEMBL transcript/gene/protein identifiers, etc. Therefore, the AS topology can be differentially assessed for custom datasets containing the respective genes of interest.

Output: landscape of AS events

From the provided annotation with respect to the specified options, the ASTALAVISTA protocol dynamically extracts AS events. As a summary, the main result page shows a list where each event type is depicted and its unique code in the relative-position notation is given. The list is ranked according to the occurrence (number or proportion) of the events. A graphical overview is provided in form of a pie diagram that displays the distribution of events across the groups, considering differentially each type of simple event and pooling the others in one group (Figure 1, left).

From the result summary page, the genomic coordinates of all AS events counted in a group are accessible by clicking on the corresponding list entry (Figure 1, top-right). For each AS event, the transcript identifiers and the variable splice sites giving rise to it are specified. Finally, each event is linked to the UCSC Genome Browser for further comparative analyses (Figure 1, bottom-right).

CONCLUSION AND PERSPECTIVE

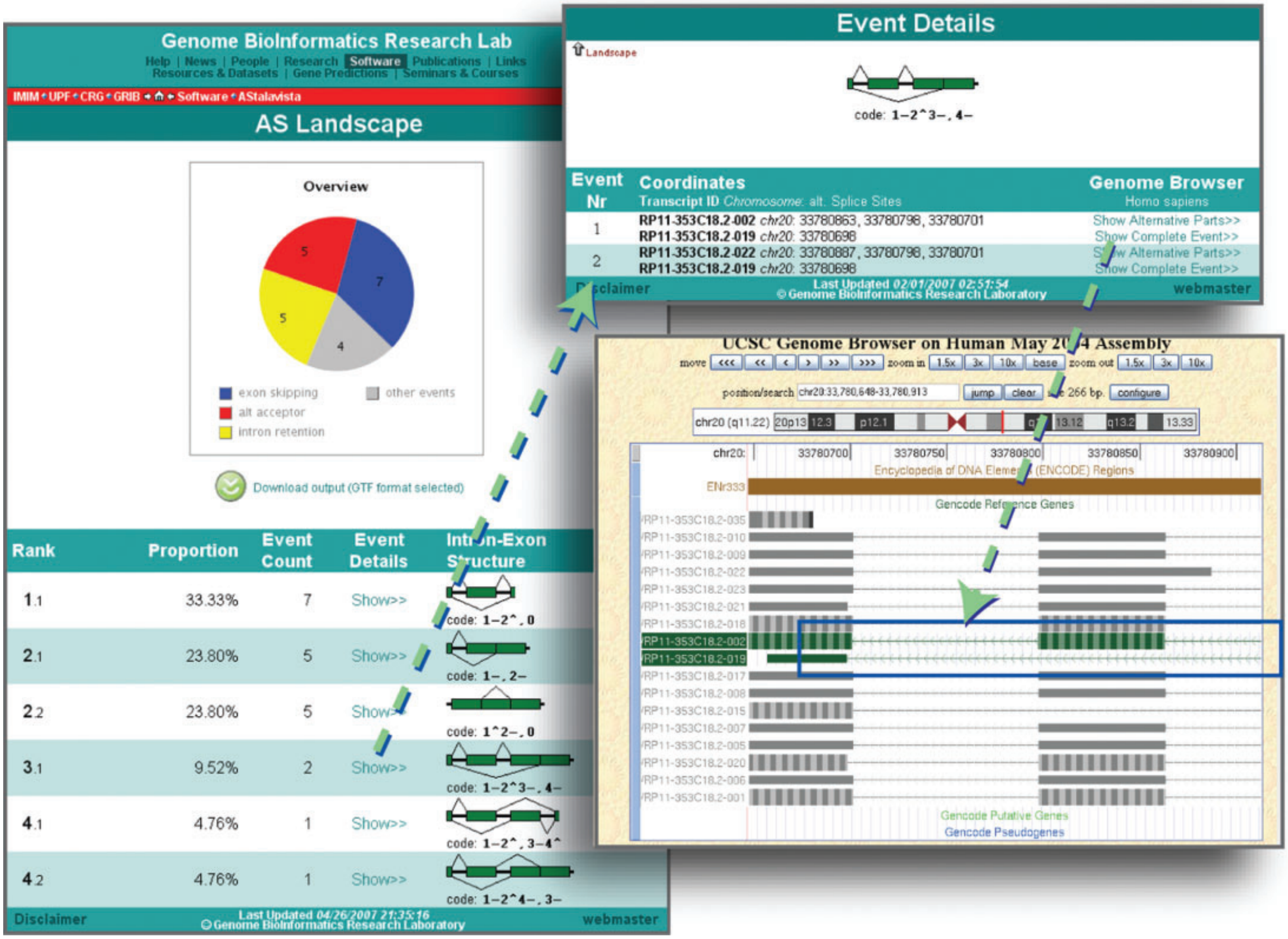
ASTALAVISTA is an explorative tool to exhaustively extract AS events reflected by a certain input dataset, to compare and to group them according to equal exon–intron structure variations. As a key feature, arbitrarily complex combinations of hitherto described AS events can be distinguished, either visually or by representation in a univocal notation system. The event-based model of AS permits to easily identify genes that involve the same type of event, e.g. alternative donors or double exon skipping. On the other hand, the comprehensive analysis of observed AS events provides a powerful tool for investigating correlations between differences in AS patterns and functional/structural features of genes, gene sets or complete genomes. In this concern, ASTALAVISTA can handle custom inputs according to any discriminatory criteria, e.g. common evolutionary conservation, pattern or intensity of expression, function/cellular localization of the gene product, etc.

Although the reference datasets currently provided on the server are dedicated to reference organisms, the generic ASTALAVISTA protocol is applicable to any genome, even if the sequencing/annotation process has not been completed. In the future, the web resource will be completed by reference annotations for more species.

ACKNOWLEDGEMENTS

The authors thank R. Guigò, E. Eyras and other members of the PRBB for inspiring and fruitful discussions. This work has been supported by a post-doctoral fellowship to M.S. Funding to pay the Open Access publication charge was provided by Centre de Regulacio Genomica, Catalunya.

Conflict of interest statement. None declared.



Genome BioInformatics Research Lab

Help | News | People | Research | Software | Publications | Links
Resources & Datasets | Gene Predictions | Seminars & Courses

IMIM • UPF • CRG • GRIB • ⌂ • Software • AStalavista

AS Landscape

Download output (GTF format selected)

Rank	Proportion	Event Count	Event Details	Intron-Exon Structure
1.1	33.33%	7	Show>>	code: 1-2^,0
2.1	23.80%	5	Show>>	code: 1-,2-
2.2	23.80%	5	Show>>	code: 1^2-,0
3.1	9.52%	2	Show>>	code: 1-2^3-,4-
4.1	4.76%	1	Show>>	code: 1-2^,3-4^
4.2	4.76%	1	Show>>	code: 1-2^4-,3-

Disclaimer Last Updated *04/26/2007 21:35:16* webmaster
© Genome BioInformatics Research Laboratory

Event Details

⇧Landscape

code: 1-2^3-,4-

Event Nr	Coordinates Transcript ID *Chromosome*: alt. Splice Sites	Genome Browser Homo sapiens
1	**RP11-353C18.2-002** *chr20*: 33780863, 33780798, 33780701 **RP11-353C18.2-019** *chr20*: 33780698	Show Alternative Parts>> Show Complete Event>>
2	**RP11-353C18.2-022** *chr20*: 33780887, 33780798, 33780701 **RP11-353C18.2-019** *chr20*: 33780698	Show Alternative Parts>> Show Complete Event>>

Disclaimer Last Updated *02/01/2007 02:51:54* webmaster
© Genome BioInformatics Research Laboratory

Figure 1. Analysis of the AS landscape in a sample dataset. The AS landscape is described by a list of AS events grouped according to equal variations in the exon–intron structure between transcripts (left). A schematic picture illustrates every type of event, specified by the respective code in the relative splice site position notation. The list is ranked according to the observed frequency of events, and as an overview, a pie diagram shows the resulting distribution. For each type of AS event, the enumeration of all genes/transcripts involved is provided, including the corresponding identifiers and genomic coordinates (top-right). The genomic positions are dynamically linked to the UCSC genome browser for further analysis (bottom-right).

REFERENCES

1. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
2. Lopez,A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, **32**, 279–305.
3. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
4. Florea,L. (2006) Bioinformatics of alternative splicing and its regulation. *Brief Bioinformatics*, **7**, 55–69.
5. Xing,Y. and Lee,C. (2006) Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, **7**, 499–509.
6. Nagasaki,H., Arita,M., Nishizawa,T., Suwa,M. and Gotoh,O. (2006) Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics*, **22**, 1211–1216.
7. Holste,D., Huo,G., Tung,V. and Burge,C.B. (2006) HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.*, **34**, D56–D62.
8. Bollina,D., Lee,B.T., Tan,T.W. and Ranganathan,S. (2006) ASGS: an alternative splicing graph web service. *Nucleic Acids Res.*, **34**, W444–W447.
9. Castrignano,T., Rizzi,R., Talamo,I.G., De Meo,P.D., Anselmo,A., Bonizzoni,P. and Pesole,G. (2006) ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization. *Nucleic Acids Res.*, **34**, W440–W443.
10. Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.*, **30**, 299–300.
11. Leipzig,J., Pevzner,P. and Heber,S. (2004) The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res.*, **32**, 3977–3983.
12. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.