

Chapter 24

Analysis of Alternative Splicing Events in Custom Gene Datasets by AStalavista

Sylvain Foissac and Michael Sammeth

Abstract

Alternative splicing (AS) is a eukaryotic principle to derive more than one RNA product from transcribed genes by removing distinct subsets of introns from a premature polymer. We know today that this process is highly regulated and makes up a large part of the differences between species, cell types, and states. The key to compare AS across different genes or organisms is to tokenize the AS phenomenon into atomary units, so-called AS events. These events then usually are grouped by common patterns to investigate the underlying molecular mechanisms that drive their regulation. However, attempts to decompose loci with AS observations into events are often hampered by applying a limited set of a priori defined event patterns which are not capable to describe all AS configurations and therefore cannot decompose the phenomenon exhaustively.

In this chapter, we describe working scenarios of AStalavista, a computational method that reports all AS events reflected by transcript annotations. We show how to practically employ AStalavista to study AS variation in complex transcriptomes, as characterized by the human GENCODE annotation. Our examples demonstrate how the inherent and universal AStalavista paradigm allows for an automatic delineation of AS events in custom gene datasets. Additionally, we sketch an example of an AStalavista use case including next-generation sequencing data (RNA-Seq) to enrich the landscape of discovered AS events.

Key words Gene expression, RNA processing, Alternative splicing, AS event, Definition of alternative splicing, Transcriptome annotation, RNA-seq, Splicing nomenclature, AS code, Bioinformatics

1 Introduction

Alternative splicing (AS) is the mechanism enabling eukaryotic cells to define and to regulate the set of introns that is removed from nascent RNA molecules during transcription. AS plays a key role in gene expression and transcriptome diversity, and the phenomenon still remains to be fully characterized [1]. One way to investigate AS is to compare and to characterize the RNA content of cell samples (e.g., from different experimental conditions, tissues, treatments, etc.) in order to identify variations between transcripts and consequently the underlying AS events. On the way to achieve efficient and user-friendly tools, computational methods of

AS analysis are constantly coping with advances in transcriptomics, in particular, recent innovations in the field of so-called next-generation sequencing (NGS) technologies, which require high-throughput pipelines to analyze large data volumes.

AStalavista (alternative splicing and transcriptional landscape visualization tool altogether) provides a computational method and software application for the automatic characterization of AS landscapes in custom transcriptome data [2]. The method automatically identifies AS events by comparing all transcripts of a given annotation file and reporting an exhaustive yet nonredundant list of the resulting variations. To allow for the generic characterization of—possibly hitherto undiscovered—morphologies of AS events, so-called AS patterns, AStalavista employs a generic nomenclature, the alternative splicing code (the “AS code”).

In this chapter, we present the AStalavista method and its notation to describe AS events and their corresponding patterns. Employing publicly available data, we demonstrate the simplicity and universality of the approach for custom gene datasets: we first provide a concrete example of an AStalavista run on the transcript annotation of the complete human genome, and then we use the method to study transcriptome variation depicted by a NGS experiment of the ENCODE project.

2 Materials

2.1 *Installing AStalavista*

The AStalavista package can be accessed from the website

<http://astalavista.sammeth.net>

where also the source code is available from the “Download” section. To install, just unzip/untar the bundle once downloaded, for instance, by the command

```
tar xzf AStalavista-3.2.tgz
```

The command creates a folder named “AStalavista” (with the corresponding version number) that contains all the files from the tarball: documentation, license information, and program files. AStalavista is executed by wrapper scripts located in the “bin” subfolder within the installation directory, either “astalavista.bat” (for Windows platforms) or “astalavista” (for UNIX-based operating systems, including OSX).

The “--help” flag provides a brief description of the software, its version, and the available options: flags to handle verbosity, parallelization, output behavior, and available tools (“--list-tools”; see below).

```
AStalavista-3.2/bin/AStalavista--elp
AStalavista v3.2 (Flux Library: 1.22)
-----Documentation & Issue Tracker-----
```

```

Flux Wiki (Docs): http://sammeth.net/confluence
Flux JIRA (Bugs): http://sammeth.net/jira
Please feel free to create an account in the
public
JIRA and reports any bugs or feature requests.
-----
The Flux library comes with a set of tools.
You can switch tools with the -t option. The
general options
change the behaviour of all the packaged tools.
General Flux Options:
...
List of available tools
...

```

2.2 Obtaining the GENCODE Annotation

The full GENCODE transcriptome annotation dataset [3] can be downloaded from the project's ftp site by the command:

```

wget
ftp://ftp.sanger.ac.uk/pub/gencode/
release\_12/gencode.v12.annotation.gtf.gz ./
and subsequently, the downloaded gzip archive is decompressed by
gunzip Gencode.v17.annotation.gtf.gz

```

2.3 Retrieving RNA-Seq Alignments

In our NGS examples, we employed publicly available RNA-Seq data from the ENCODE project [4] as downloaded from the CRG's (Centre for Genomic Regulation, Barcelona) RNA-Seq Dashboard:

http://genome.crg.es/encode_rna_dashboard/hg19/

The CRG dashboard allows navigating hundreds of RNA-Seq experiments produced by the ENCODE consortium to profile the human transcriptome of different cell types, subcellular compartments, and RNA populations. Employing the STAR mapper [5], reads of some of the datasets have been split-mapped to the human genome in order to identify the position of the introns. For our example, we picked one of these mapping files downloaded from the URL:

http://genome.crg.es/~jlagarde/encode/pre-DCC/wgEncodeCshlLongRnaSeq/20130218_promocell_batches1-2_minus_batch1sFASTQ/SID38226_SID38227_SJ.bed

The aligned RNA-Seq data has been obtained from long RNAs that have been extracted from HPAEC (Human Pulmonary Artery Endothelial Cells) and sequenced on an Illumina HiSeq 2000 machine using a strand-specific protocol. However, the ENCODE dashboard provides data from many other experimental conditions, which could be combined or compared with this example dataset.

3 The AStalavista Method

3.1 AS Event

Definition and Nomenclature

Computer-aided comparison of transcriptome data is a natural way to analyze AS, and many studies have followed that approach [6]. In order to make conclusions on the effects of AS, the idea is to identify in the first place common differences in the exon-intron structure of processed RNA molecules. This can be straightforwardly achieved by comparing the transcript sequences with each other, either directly or by using their localization within the context of the corresponding genomic sequence, if available. The “mapping” of transcripts onto a genomic reference allows to identify the position of all exons and introns, producing what we call a transcriptome annotation.

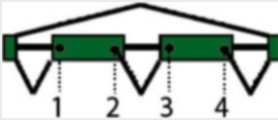
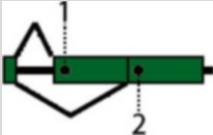
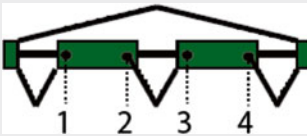
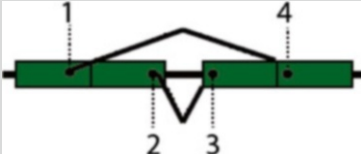
3.1.1 AS Events

Given an annotation, each transcript can be uniquely identified by the type and the genomic position of all of its exonic boundaries. Here we consider three types of exonic boundaries, we refer to as “sites”: the transcription start site (TSS) at the 5′ extremity of the transcript, the cleavage site (CVS) at the 3′-end of the transcript—often coupled with the polyadenylation site—and splice sites that delineate introns in the premature transcript; splice sites are furthermore categorized as *donor* or *acceptor* sites according to their localization at the 5′- or at the 3′-end of an intron, respectively. Alternative splicing occurs when a splice site is used in the splicing process of one transcript but not of another transcript of the same locus. An AS event intuitively comprises one or multiple of such alternative splice sites; moreover, since transcription and splicing are known to co-occur and interact in eukaryotic cells [7, 8], some AS events also include alternative TSSs or CVSs in addition to alternative splice site(s).

More precisely, when comparing the exon boundaries of two or more transcripts from a genomic locus, the AStalavista approach first identifies *constitutive* (splice) sites as exon boundaries employed by all transcripts that overlap the corresponding genomic position; consequently, sites that are not used by all transcripts of the locus are considered as *variable*. In a further step, AStalavista delineates AS events as a maximal sequence of consecutive variable sites, with at least one alternative splice site. Based on this definition, AS events either contain exclusively alternative splice sites delimited by common sites at both ends (so-called internal AS events) or they extend until the transcript extremities and contain variable TSSs and/or CVSs.

By this generic event definition of AStalavista, resulting events can describe variations between two or more different transcript structures. The number of transcripts that are simultaneously compared naturally plays an important role in the process of assessing whether a certain site is variable or constitutive. For the sake of convenience, we will consider here pairwise AS events only, but we

Table 1
Examples of distinct AS events with the corresponding AS patterns

Traditional name	Exon-intron structure	AS code	Example
A Exon skipping		0, 1-2 [^]	PTBP1 polypyrimidine tract-binding protein 1, ninth exon
B Alternative acceptor		1-, 2-	PTBP1 polypyrimidine tract-binding protein 1, ninth exon
C		0, 1-2 [^] 3-4 [^]	FBLIM1 filamin-binding LIM protein 1, fourth + fifth exon
D		1 [^] 4-, 2 [^] 3-	ZWINT ZW10 interacting kinetochore protein, sixth – seventh exon

would like to point interested readers to the concept of complete alternative splicing events for a more advanced characterization of AS loci [9].

Table 1 shows some AS patterns along with examples as they are observed in the GENCODE annotation of the human transcriptome. The alternative exon 9 in the PTBP1 (polypyrimidine tract binding protein 1) gene has been reported to be entirely skipped (“exon skipping,” A) or included with different acceptor flanks (“alternative acceptor,” B). One of the multiple possibilities to process the pre-mRNA of the FBLIM1 (Filamin-binding LIM protein 1) gene (C) excludes simultaneously the subsequent exons 4 (188 nt) and 5 (103 nt), such that overall the reading frame is not disturbed (292 nt corresponds to 97aa). (D) Similarly, variations on either side of the sixth intron in the ZWINT locus regulate the optional inclusion of 47 amino acids into the resulting protein. As we will show in the next section, the complete set of AS events can be automatically identified and reported by the AStalavista software [10].

3.1.2 The AS Code

AS events are commonly classified in different categories based on the observed differences in the exon-intron structure between the respective transcripts. Typically, most studies employ the terms

“exon skipping,” “alternative donor/acceptor,” “intron retention,” and (sometimes) “mutually exclusive exons” to describe AS patterns. However, a nomenclature built exclusively by this set of predefined terms allows to describe only a couple of AS configurations, which due to their simplicity are observed quite frequently but constitute only a small fraction of the spectrum of all event patterns found in complete transcriptome annotations. An exhaustive characterization of the AS landscape from arbitrary transcript structures requires a more flexible notation that permits to identify all events that share the same “topography.”

For this reason, AStalavista uses a generic notation system that can univocally describe any AS event. This nomenclature assigns an “AS code” to every sequence of alternative splice sites, such that events with qualitatively identical variations in their exon-intron structure receive the same AS code. This AS code represents each variable site by a number and a symbol, according to its relative position within the event and to its type. More precisely, the code of a certain AS event is generated as follows: first, all variable sites that are included in the event are sorted according to their position in the directionality of transcription, from 5' to 3'. Each site is assigned a number according to this order (1, 2, 3, etc.) and a symbol based on its type: “[“for TSS, ”]” for CVS, “^” for donor, and “-” for acceptor. Then, sites from the same transcript are represented next to each other by consecutive pairs of such number-symbol tuples (in the corresponding 5' → 3' order), forming a string that represents the AS event. In the absence of variable sites in one of the transcripts of the event (e.g., to describe the string for a transcript skipping an exon), the digit “0” is employed as a corresponding string. The AS code then is a comma-separated concatenation of thus obtained strings, one from each transcript variant, ordered by their numbers. Some examples for AS events along with their corresponding AS code are shown in Table 1.

3.2 Processing AS Events from Transcriptome Annotations: The GENCODE Example

The ENCODE project intends to establish a repertoire of functional elements in the human genome [11]. To this aim, a tremendous amount of experimental data has been (is being) produced and analyzed by hundreds of scientists worldwide. As part of the results from this effort, the GENCODE annotation provides the community with an expertized and thoroughly curated set of genes and transcript positions in the genomic sequence. The GENCODE annotation describes the exon-intron structure of thousands of alternative transcripts and therefore allows us to profile AS extensively and accurately. Therefore, as a first example for the application, we will employ AStalavista to characterize the AS landscape of the entire human transcriptome as described by the GENCODE annotation.

3.2.1 Running AStalavista on the GENCODE Annotation

The AStalavista program package is subdivided in different tools; a list of available tools can be obtained by requesting help via the corresponding command option

```
./bin/AStalavista -list-tools
```

which by the current version (AStalavista-3.2) outputs the summary:

```
List of available tools
scorer - Splice site scorer
sortBED - Sort a BED file. If no output
file is specified, result is printed to standard out
sortGTF - Sort a GTF file. If no output file
is specified, result is printed to standard out
asta - AStalavista event retriever
subsetter - Extract a random subset of lines
from a file
```

The tool “asta” performs the extraction of alternative splicing, for which additional parameters and their default values in turn are obtained by

```
./bin/AStalavista -t asta -h
```

Parameters can either be provided on the command line or in a separate parameter file (specified by flag “-p”). A mandatory parameter is the path to the file with the annotation that is to be analyzed (parameter “IN_FILE”, or command line parameter “-i”). As an example, an AStalavista run on the v12 of the GENCODE transcript collection is executed by

```
./bin/AStalavista -t asta -i Gencode.v12.annotation.gtf
```

AStalavista employs all exons (i.e., lines identified by the feature “exon” in the third column) from the gtf file, which are to be sorted by their genomic positions and require a “transcript_id” identifier for correct gene clustering. If the file is not sorted, AStalavista will automatically sort the file before processing.

As a default setting, AStalavista reports only alternative splicing events [10] that are not linked to variable transcription initiation and/or cleavage sites, so-called internal AS events (ASI). External AS events (ASE) that include TSSs and/or CVSs can also be included in the list of reported event types (value ASE for the parameter EVENTS or command line flag “-e”):

```
./bin/AStalavista -t asta -i Gencode.v12.annotation.gtf -e ASI,ASE
```

Per default, AStalavista further on shows exclusively pairwise AS events, i.e., events with exactly two variants (dimension 2); events of a higher dimension are projected to events between all

pairs of their variants. To change the order of dimension for the reported AS events, parameter `EVENTS_DIMENSION` (flag `-d`) has to be set to a corresponding integer value or “-1” for outputting the so-called complete AS events [9]:

```
./bin/AStalavista -t asta -i Gencode.v12.annotation.gtf -d -1
```

3.2.2 Interpreting the Results

AStalavista produces from the input annotation in gtf format a file with events that are equally stored as gtf features: start (column 4) and end (column 5) of the event are the genomic coordinates of the delimiting common sites [10], respectively, the first/last variable site in the case of ASE events; the nature of these delimiting sites is detailed in the “flanks” attribute, which adopts the same “AS code” notation as used for “splice chain” and “structure” (see below):

```
chr1      Gencode_v12      as_event
12227    12721    .      +      .      tran-
script_id "ENST00000518655.2,ENST00000456328.2/
ENST00000515242.2"; gene_id "chr1:11869-14412W";
flanks "12227^,12721^"; structure "1-,2-"; splice_
chain "12595-,12613-"; dimension "2_2"; degree "4";
```

The attribute “degree” summarizes the number of variable sites in the event, which is naturally increasing with longer events or events of more variants. Additionally, the attribute “dimension” provides information about the reported event with respect to the underlying complete event: the dimensionality is denoted as a string of the form “ X_Y ” where X is the number of variants in the reported event and Y is the number of events in the corresponding complete event. All events with $X=Y$ are complete, and those with $X<Y$ are not ($X>Y$ is not possible by definition).

The “transcript_id” attribute describes a comma-separated list of the transcript identifiers for each of the variant structures of the event; if there is more than one transcript supporting a certain variant of the event, the corresponding identifiers are furthermore separated by a “/” separator. The attribute “splice_chain” describes the genomic coordinates of each variable site in the event, with the variants in the same comma-separated order as the identifiers of “transcript_id”.

As introduced previously, the AS code nomenclature specifies after the genomic coordinate of each site also the site type by a certain symbol: “^” denotes a splice donor, “.” a splice acceptor, and “[“a transcription start and”]” a cleavage site. Together with the name of the chromosome/contig, the value of “splice_chain” can be considered as a unique identifier of the event, because it describes every event univocally by the morphology of its sites as localized by corresponding genomic coordinates. Therefore, also events delimited by the same flanks (start/end tuples) can still differ by their splice chains.

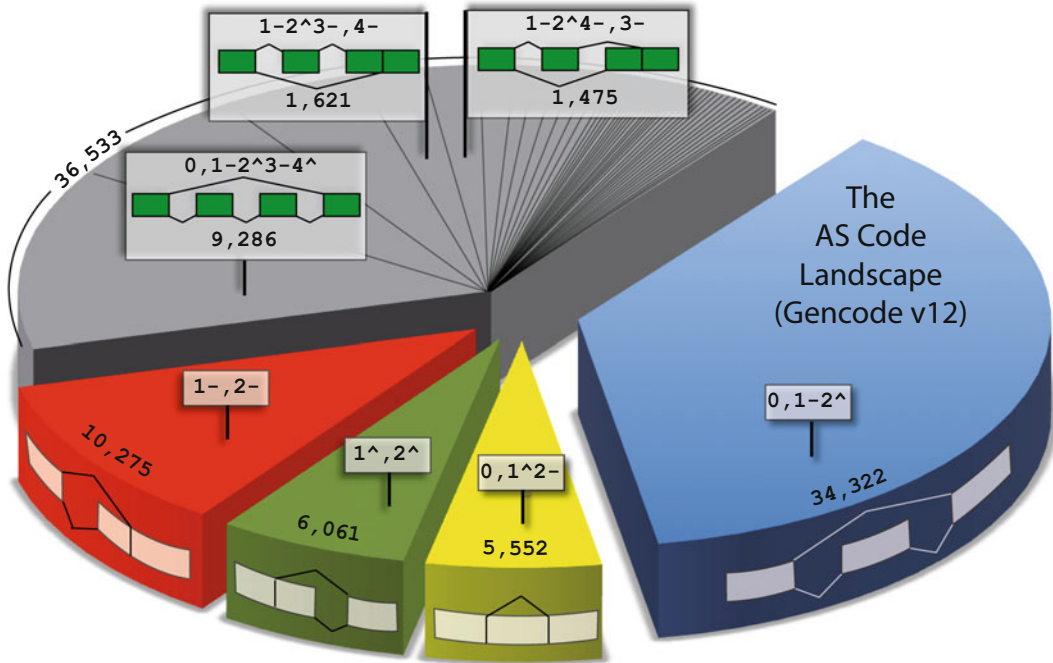


Fig. 1 AS landscape of internal events of the Gencode (v12) annotation

By replacing genomic coordinates reported for the “splice_chain” by relative positions of the corresponding sites within the event, the attribute “structure” abstracts events to their respective patterns that are named by the corresponding AS code; examples for AS codes of common events are: “0,1-2[^]” for exon skipping, “1[^],2[^]” for alternative donors, “1-,2-” for alternative acceptors, “0,1[^]2-” for intron retention, and “1-2[^],3-4[^]” for mutually exclusive exons.

Figure 1 depicts all *internal* AS events found in the Gencode (v12) transcriptome annotation, grouped by their respective AS code. Altogether, the landscape of different AS morphologies reveals that traditional events (blue, yellow, green, and red slices) in our days hardly make up 37 % of the spectrum spanned by all AS patterns—the majority of events (gray slice) cluster into 813 different, more complex events. Note that, depending on the position of the event, it might further be possible to introduce alternative TSSs or CVSs in several AS shapes to further amplify the landscape (*external* AS events).

The result of a default AStalavista runs as shown in Fig. 1, for example, GENCODE annotation retrieves 92,743 internal AS events (ASI) that cluster in 817 different patterns (respectively 670,104 events clustering in 17,741 patterns when including ASE events). Expectedly, the most abundant event patterns are of degree 2 which corresponds to the minimum number of variable

sites required for an AS event: “0,1-2[^]” (exon skipping) ranks first, “1-,2-” (alt. acceptors) ranks second, “1[^],2[^]” (alternate donors) ranks third, and “0,1[^]2-” (intron retention) ranks fourth most abundant pattern. The reasons for the differences between these patterns lie in splicing mechanism and constraints of coding sequences, which constitute the main part of messenger RNA transcripts [10].

Although the event complexity specified by “degree” usually anticorrelates with the observed abundance, there are exceptions. For instance, the skipping of multiple exons is mechanistically facilitated and therefore more often observed than other events of the corresponding degree: “0,1-2[^]3-4[^]” ranks before “1[^],2[^]” and “0,1[^]2-”, “0,1-2[^]3-4[^]5-6[^]” and also before many other patterns of degree 4, etc. Moreover, only ~47 % of pairwise events found on the GENCODE annotation are complete; from another point of view, that means that the true complexity of more than half of the events is underestimated when looking at them exclusively by pairwise variant comparisons.

3.3 Discovery of Novel AS Events by RNA-Seq Experiments

3.3.1 Employing AStalavista for NGS Data

AStalavista can also analyze NGS data from RNA sequencing experiments (RNA-Seq). The main conceptual difference between transcriptome annotations, like GENCODE and RNA-Seq data, is that the latter only provide partial information on the transcriptomes: RNA-Seq reads are limited to parts of expressed genes, because current NGS technologies still cannot reproduce sequences of entire RNA molecules and transcripts thus are fragmented before sequencing. Although RNA-Seq reads correspond to fragments and cannot be considered as full-length transcripts, they can provide valuable information about AS by potentially novel splice junctions they harbor: if the exon-exon junction of a spliced transcript is captured in a read, the position of the corresponding intron can be revealed by aligning the read sequence on the genomic reference.

Dedicated NGS alignment methods—as implemented in software like GEM [12] or STAR [5]—allow to generate such spliced alignments where introns correspond to gaps flanked by exonic regions. These experimentally derived evidences about splicing events can be analyzed by AStalavista, either separately or in combination with a reference annotation, to identify novel alternative splicing events that are not captured by the reference gene set. An interesting and original aspect of the latter approach is that no transcript assembly is required for the analysis; indeed, AStalavista considers reads as independent pieces of information and does not take any assumptions which of them might originate from the same and which of them have been obtained from independent RNA molecules, respectively. In the next section, we employ this strategy to illustrate how AStalavista can enrich a repertoire of AS events by processing RNA-Seq data.

3.3.2 Processing RNA-Seq Data

We converted the format of the bed file with mapped reads downloaded earlier from the CRG dashboard, and produced an AStalavista-compatible gtf file by the commands:

```
bed=SID38226_SID38227_SJ.bed; gtf=${bed%.bed}.gtf; cat $bed | awk -v bed=$bed -v OFS="\t" '{print $1,bed,"exon",$2,$2,$5,$6,".", "gene_id \"SJ\"NR\"\""; transcript_id \"SJ\"NR\"\"";};print $1,bed,"exon",$3,$3,$5,$6,".", "gene_id \"SJ\"NR\"\""; transcript_id \"SJ\"NR\"\"";}'>$gtf
```

Finally, we concatenated thus obtained gtf representation of the NGS mappings with the GENCODE annotation from our previous example:

```
cat Gencode.v12.annotation.gtf $gtf>Gencode_rnaseq.gtf
```

The joint dataset then was processed with AStalavista in order to compare correspondingly obtained AS events with the results derived from the GENCODE annotation alone:

```
./bin/AStalavista -t asta -i Gencode_rnaseq.gtf
```

3.3.3 Identifying Novel AS Events

After processing the NGS-enriched GENCODE annotation with AStalavista, we characterized events involving RNA-Seq reads that were not detected using the annotation alone. We found 342 such novel internal AS events (ASI). Figure 2 shows that the distribution of patterns among these additionally found events follows closely the one described by GENCODE events, albeit some rank positions are permuted: evidence by RNA-Seq split mappings indicates that among the simple AS patterns, alternative exon boundaries (89 donors and 79 acceptors events) are most underestimated by the current GENCODE annotation, followed by alternatively skipped exons (53 novel alternative exons); furthermore, RNA-Seq data pinpoints 177 novel complex events of 34 distinct patterns (Fig. 2).

Figure 2 also summarizes the number and nature of AS events that are revealed by RNA-Seq introns; many describe the skipping of (multiple) exons: 53 “0,1-2^” (1 exon), 21 “0,1-2^3-4^” (2 exons), 11 “0,1-2^3-4^5-6^” (3 exons), etc. Albeit our analysis focuses on a small set of high-confidence events, we observe nearly half of the exons that predicted to be skipped (20 out of 53) exhibit a frame-preserving length, i.e., the size of the exon is a multiple of 3. These events include, for instance, the third exon in the TMCO1 (transmembrane domain protein) locus, where also additional EST (Expressed Sequenced Tags) evidence (BP308568) and complementary computational gene models (ENST00000476173) support the skipping of the corresponding 60 nt exon; as another example, we find NGS evidence for the skipping of the 36 nt long exon number 10 in the TRIP (thyroid hormone receptor interacting protein) gene, which so far has not been discovered by EST evidence.

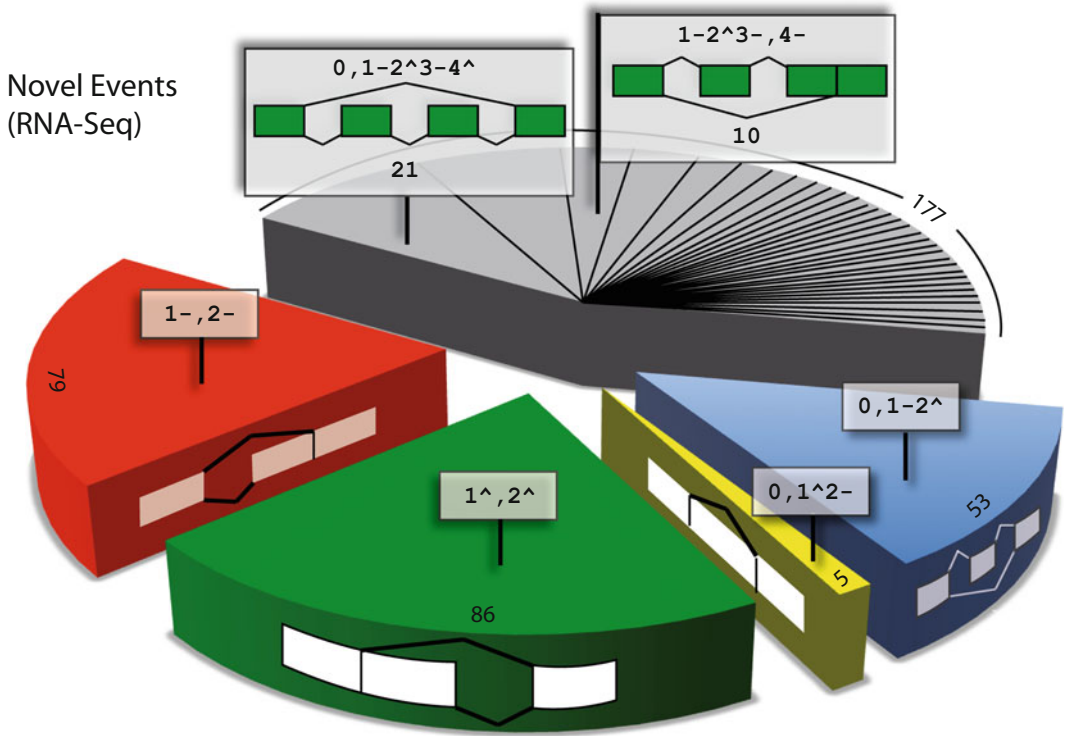


Fig. 2 Distribution of novel AS events found by additional RNA-Seq evidence

As in other analyses that employ NGS data, potential artifacts might impair the results, and some of the events might originate from errors in the sequencing or the alignment process. As in the case of reference annotations, AStalavista processes the exons and introns provided by NGS data and all variations in the exon-intron structure are reported. Reassuringly, however, we find among these novel events several traces of constraints created by coding sequences, for instance, a nonnegligible proportion of exons that are predicted to be skipped by RNA-Seq evidence would not break the coding frame. Also, there is independent evidence from cDNA and EST data that supports some of the NGS-derived events. Finally, it is worth noting that most of the NGS evidence for exon skipping predicts alternative exons longer than 100 bp, which are unlikely to originate from misalignments or gaps in the alignment process.

Certainly, those AS events can be subject to further ranking and/or filtering procedures, for instance, by considering their overlap with the annotated coding sequence, by taking the number of reads supporting each of the splice junctions in the event and their corresponding alignment quality (e.g., gaps) into account, etc. However, the example demonstrates that already a straightforward application of AStalavista to different RNA-Seq datasets

allows to analyze cell-type or condition-specific transcriptomes, as by adding the split-mapped reads to a reference annotation and comparing the correspondingly obtained AS events for novel configurations.

An interesting specificity of the approach is that—unlike most of other RNA-Seq data processing tools—AStalavista does not attempt to assemble the reads into longer contigs nor even into transcribed fragments of the genomic space (the so-called transfrags). Although transfrag-based methods have been producing better results in terms of accuracy along the years, transcriptome reconstruction from RNA-Seq data is still a highly challenging task in the field [3]. Whether methods use a genomic reference, as in our example, or whether they rather compare reads with each other (“de novo” assembly), the risk of merging some unrelated reads to a common contig cannot be completely avoided [13], generating some artificial hybrid that does not exist in vivo. Moreover, transcriptome assembly methods are prone to errors when dealing with repeated sequences and/or incomplete data, thus processing RNA-Seq data by AStalavista right after the read mapping step improves the general reliability of the analysis. Under the assumption that genomic read alignments are usually a more reliable source of information than transcript structures reconstructed upon them, AStalavista is not concluding that reads with a common subsequence originate from identical transcripts but rather focuses on splicing differences they (might) harbor: by the atomary definition of delineating variations strictly within the common genomic space ensures that any of the reported AS events is supported by two real sequences.

Another advantage of the AStalavista approach is its efficiency in terms of processing time. The example shown in this chapter took about 5 min on a 2 GHz Intel system, occupying <1 GB of RAM, which makes it suitable for high-throughput processing. The rather simplistic examples we provide here can straightforwardly be extended to the application of AStalavista for processing additional datasets, e.g., for comparing different reference annotations, transcriptomes from different cell compartments, and/or different experimental conditions or data preprocessing strategies.

References

1. Kornblihtt AR, Schor IE, Alló M (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* 14:153–165
2. Foissac S, Sammeth M (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res* 35:W297
3. Harrow J, Frankish A, Gonzalez JM (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22:1760
4. Dunham I, Birney E, Lajoie BR, Sanyal A (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57

5. Dobin A, Davis CA, Schlesinger F (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15
6. Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Res* 9:1288–1293. doi:[10.1101/gr.9.12.1288](https://doi.org/10.1101/gr.9.12.1288)
7. Tilgner H, Knowles DG, Johnson R (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
8. Boireau S, Maiuri P, Basyuk E, de la Mata M, Knezevich A, Pradet-Balade B, Bäcker V, Kornblihtt A, Marcello A, Bertrand E (2007) The transcriptional cycle of HIV-1 in real-time and live cells. *J Cell Biol* 179:291–304. doi:[10.1083/jcb.200706018](https://doi.org/10.1083/jcb.200706018)
9. Sammeth M (2009) Complete alternative splicing events are bubbles in splicing graphs. *J Comput Biol* 16:1117
10. Sammeth M, Foissac S, Guigó R (2008) A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol* 4:e1000147. doi:[10.1371/journal.pcbi.1000147](https://doi.org/10.1371/journal.pcbi.1000147)
11. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A et al (2012) Landscape of transcription in human cells. *Nature* 489:101–108. doi:[10.1038/nature11233](https://doi.org/10.1038/nature11233)
12. Marco-Sola S, Sammeth M, Guigó R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 9:1185. doi:[10.1038/nmeth.2221](https://doi.org/10.1038/nmeth.2221)
13. Mundry M, Bornberg-Bauer E (2012) Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PLoS One* 7:e31410