JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 8, 2009
© Mary Ann Liebert, Inc.
Pp. 1117–1140
DOI: 10.1089/cmb.2009.0108

Research Article

Complete Alternative Splicing Events Are Bubbles in Splicing Graphs

MICHAEL SAMMETH

ABSTRACT

Eukaryotic splicing structures are known to involve a high degree of alternative forms derived from a premature transcript by alternative splicing (AS). With the advent of new sequencing technologies, evidence for new splice forms becomes increasingly available—bit by bit revealing that the true splicing diversity of "AS events" often comprises more than two alternatives and therefore cannot be sufficiently described by pairwise comparisons as conducted in analyzes hitherto. Especially, I emphasize on "complete" AS events which include all hitherto known variants of a splicing variation. Challenges emerge from the richness of data (millions of transcripts) and artifacts introduced during the technical process of obtaining transcript sequences ("noise")—especially when dealing with single-read sequences known as expressed sequence tags (ESTs). Herein, I describe a novel method to efficiently predict AS events in different resolutions ("dimensions") from transcript annotations that allows for combination of fragmented EST data with full-length cDNAs and can cope with large datasets containing noise. At the doorstep of many new spliceforms becoming available by novel high-throughput sequencing technologies, the presented method helps to dynamically update AS databases. Applying this method to estimate the real complexity of alternative splicing, I found in human and murine annotations thousands of novel AS events that either have been disregarded or mischaracterized in earlier works. The growth of evidence for such events suggests that the number still keeps climbing. When considering complete events, the majority of exons that are observed as "mutually exclusive" in pairwise comparisons in fact involves at least one other alternative splice form that disagrees with their mutual exclusion. Similar observations also hold for the alternative skipping of two subsequent exons. Results suggest that the systematical analysis of complete AS events on large scale provides subtle insights in the mechanisms that drive (alternative) splicing.

Key words: exons, gene expression, genomes, introns, RNA.

1. INTRODUCTION

Alternative splicing (AS), a fundamental molecular process regulating eukaryotic gene expression, generates a substantial part of the proteome diversity (Lander et al., 2001) and is involved in numerous

Bioinformatics and Genomics, Centre for Genomic Regulation (CRG), Barcelona, Spain.

diseases (Kuyumcu-Martinez and Cooper, 2006; Lopez, 1998; Smith and Valcarcel, 2000), also in human. Over the recent years splicing variations for several organisms have been collected in various databases and several attempts to analyze the complexity of AS throughout different genomes have been undertaken (Grasso et al., 2004; Kim et al., 2007; Nagasaki et al., 2005; Sammeth et al., 2008; Yandell et al., 2006). Usually, splicing diversity is classified according to the observed pattern of exon-intron variation into structurally different events. However, all works hitherto focused on alternative splicing in a limited context by considering exclusively pairwise comparisons—either of a reference transcript to other splice forms, or by comparing the transcript data in an all-against-all fashion. Focusing on pairs of transcripts may—although describing the atomary elements of a pattern—separate events that actually together form more complex splicing structures. Imagine for instance the effort to reconstruct the mutual exclusion of 3 (or more) neighboring exons from a set of pairwise events. It has already been noticed that comparing transcripts one by one is not satisfactory for describing such complex variations and novel ways to deal with this shortcoming have been postulated (Zavolan and van Nimwegen, 2006).

In this work, I describe a technique to exhaustively and efficiently describe arbitrarily large splicing variations in *annotations*, i.e., transcript data aligned to the genome. With respect to the quality of the sequenced transcript data, such alignments are usually not free of artifacts and sequencing errors can lead to misalignments of the transcripts to the genome. Gaps introduced in the transcript sequence during alignment and misaligning nucleotides that are arbitrarily distributed to the left or right of an intron consequently lead to the observation of wrong or shifted introns and artificially suggest variation of the exon-intron structure (so-called *noise*). Another source of noise is due to technical difficulties in obtaining 5′- and 3′-complete transcript sequences, and especially single read sequences—typically not longer than 500 bases—without subsequent assembly usually represent fragments of transcribed genes called *expressed sequence tags* (ESTs). In contrast, ESTs that have been assembled to larger transcript sequences and contain part of an (hypothetical) open reading frame (i.e., a subsequence of the transcript that does not exhibit a stop codon in one of the 3 possible frames) are usually classified as *messenger RNAs* (mRNAs) although they often are still truncated parts of the real mRNA molecules. Based on these EST and mRNA data, curated reference transcripts are built, for example, the NCBI mRNA reference sequence collection (Pruitt et al., 2007), which are considered to be *full-length* but normally do not comprise all evidence for alternative splicing. Naturally there is much more EST evidence available than mRNA sequences and full-length transcripts are the minority; for example, Genbank (Benson et al., 2007) currently contains ~8 million human ESTs, ~260,000 mRNAs and ~25,000 RefSeq transcripts. Considering these—continuously growing—numbers and the advent of new sequencing technologies that already have been applied to explore transcript diversity (Ruan et al., 2007; Weber et al., 2007), the need for efficient methods to analyze huge annotation datasets becomes evident.

The rest of the article is organized as follows: In Section 3.2, I briefly introduce the problem of quantifying AS, which can be reduced to the problem of delineating single instances of such quanta, so-called AS events. Then I give a general definition of AS events that involve two or more alternatives (Section 2.1) and the terminology used for describing them, including a nomenclature to name structurally different classes (so-called patterns, Section 2.2). Next, in Section 3, I recapitulate splicing graphs—a data structure that can be inferred from transcript annotations. In these graphs, I describe the properties of *bubbles*, specific graph substructures that imply potential AS events (Section 3.1). I show how AS events are obtained from the bubble subgraphs (Section 3.2). In Section 4, I propose an efficient algorithm to exhaustively and non-redundantly extract all AS events reflected by a splicing graph—avoiding possible artifacts from the upstream annotation pipeline. In Section 5, I demonstrate that the time complexity of the method in praxis is dominated by the linear dependence on the number of transcripts in the annotation (Section 5.1) such that even the biggest annotation datasets currently can be analyzed within few hours. Although today's annotations of the human and the murine transcriptome exhibit primarily pairwise events, I find a substantial amount of AS events that have more than two alternatives and thus have either been disregarded or misjudged in analyzes hitherto (Section 5.2). However, a retrospective analysis of the Genbank submissions (Section 5.3) suggests that the discovery of novel AS events will continue. Further bioinformatical studies on different patterns which have been observed as "mutually exclusive exons" so far show clear differences in the lengths of intermediate introns (Section 5.4). Similarly, I conduct analyzes on the differential characteristics of events involving the skipping of 2 sequential exons (Section 5.5). In Section 6, I connect results from both large-scale studies with mechanisms proposed for special cases, highlighting the exploratory power of the described method.

2. EVENTS ARE UNITS FOR MEASURING AS

The increasing complex transcript evidence that becomes available from in-depth studies such as the FANTOM3 (functional annotation of mouse) project in mouse (Carninci et al., 2005) or the ENCODE (encyclopedia of DNA elements) in human (ENCODE Project Consortium, 2007) made the community reflect on the redefinition of terms that have so far grown consuetudinary in genetics. Going back to the 1970s, "alternative splicing" has historically been described as a general biological phenomenon for which later-on single instances have been delineated in the form of "events." Recently the spectrum of events has been extended from a limited set of *ad hoc* defined patterns to general comparisons (Nagasaki et al., 2005; Sammeth et al., 2008), however, still limited to the traditional pairwise transcript comparison. In this section, I will present formal descriptions for the terms used in this work. Subsequently, I introduce a general definition of AS events with 2 or more alternatives along with the concept of "complete" events.

2.1. Terminology of transcribed sequences

An RNA sequence aligned to the genome (hereafter called *transcript*, examples in Fig. 1A) can be described by a sequence of exon boundaries (i.e., *sites*) $rna = \langle s_i^{rna} \rangle_{i=1}^{n}, n \geq 2$ ordered by their position $pos(s_i^{rna})$ from 5′ to 3′. Genomic coordinates of sites that align to the negative strand are inverted $-pos(s_i^{rna})$ to preserve the 5′→3′ directionality when ordering them. Furthermore, a site $s = (pos(s), transcripts(s), class(s), type(s))$ is characterized by its functional classification $type(s) \in$ {donor, acceptor, start, end, 5′-truncation, 3′-truncation, root, leaf}, the set of supporting transcripts $transcripts(s)$ and their respective category $class(s) \in$ {RefSeq, mRNA, EST}. The types "start" and "end" refer to the presumptive transcription start respectively the poly-adenylation site of full-length transcripts, "5′-truncation" respectively "3′-truncation" correspondingly mark the truncated counterparts, and "root" and "leaf" specify two artificial sites that are necessary in the subsequently described technique (Section 3). The category $class(s)$ is assigned in a hierarchical manner, i.e., it is RefSeq if at least one RefSeq transcript is in $transcripts(s)$, if not it is mRNA if at least one of the mRNAs is supporting s, otherwise $class(s) = \text{EST}$. When investigating AS, I compare all sets of k transcripts $\{rna_i\}_{i=1}^{k}$ that overlap in the genomic region they align to on the same strand (a *locus*, Fig. 1A) and I distinguish differences observed in the exonic structure as *variants*.

Definition 1. *A "variant" p is a sequence of m sites* $\langle s_i^p \rangle_{i=1}^{m}$ *shared by a non-empty set of transcripts* $X_p = \bigcap_{i=1}^{m} transcripts(s_i^p)$. *I refer to* X_p *as the "partition" of supporting transcripts because different variants between* $pos(s_1)$ *and* $pos(s_m)$ *split the k transcripts of the corresponding locus into disjoint sets. A variant* $p = (s_1, s_m, X_p)$ *can equivalently be described by the delimiting sites* (s_1, s_m) *and its partition* X_p *because for each RNA* $\in X_p$ *there exists a subsequence* $\langle s_j^{rna} \rangle_{j=x}^{y} = \langle s_i^p \rangle_{i=1}^{m}$.

In analogy to the jargon used for the sites of a transcript, I say a variant p to *support* a certain site s iff $s \in \langle s_i^p \rangle_{i=1}^{m}$ and consequently $X_p \subseteq transcripts(s)$. Considering a set P of variants, I distinguish 3 types of sites according to their support by the variants.

Definition 2. *Given a set of* $|P|$ *compared variants, a site s can either be (i) a "common" site if it is supported by all variants* $X_p \subset transcripts(s) \; \forall \; p \in P$, *otherwise it is (ii) a "variable" site. With respect to the special role in the molecular mechanism, (iii) I further distinguish the subset of variable sites with* $type(s) \in \{donor, acceptor\}$ *and* $pos(s_1^{rna}) \leq pos(s) \leq pos(s_n^{rna}) \; \forall \; rna \in X_p, p \in P$ *as "alternative splice sites."*

If s is variable (Definition 2 ii), then naturally there is at least one of the $|P|$ variants not supporting s. The reason for a site s being supported by one variant p and being absent from another variant q can be either differential transcription start/poly-adenylation or alternative choices of the splicing machinery when processing the primary RNA molecule. The first two cases describe variable site, where s is absent in p simply because it is not transcribed, or it is cleaved off the primary transcript before involved in splicing, respectively. In order to serve as an alternative for the splicing machinery, s obviously has to be a splice site and additionally to be contained in the genomic region (i.e., the primary RNA molecules) of all transcripts that constitute the variants in P (Definition 2 iii).

AS events naturally involve $d \geq 2$ variants, where d is said the *dimension* of the event. To generally delineate AS events of dimension d in a locus, I extend an earlier described definition for *pairwise* $(d = 2)$

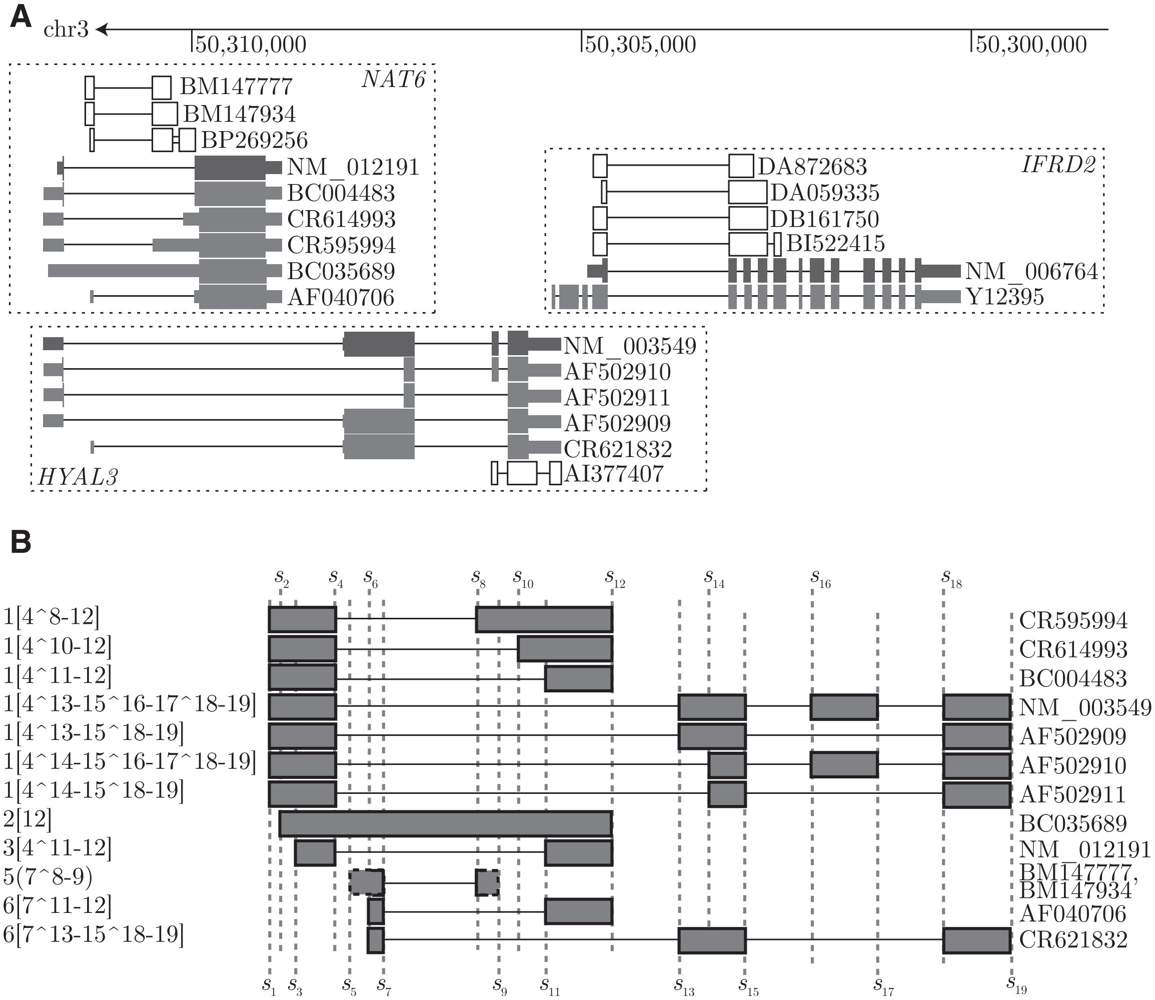


FIG. 1. (**A**) Splicing locus on the *Crick* strand of human (hg18) chromosome 3, containing the gene encoding a putative tumor suppressor, the so-called Fusion 2 protein (NAT6, N-acetyltransferase 6), the hyaluronoglucosaminidase 3 (HYAL3) gene and the gene for the interferon-related developmental regulator 2 (IFRD2). Annotations of RefSeq transcripts (dark gray), Genbank mRNAs (light gray) and some of the ESTs from dbEST with alternative splice forms (white) are depicted. Transcripts are shown in direction of transcription, from 5′ (left) to 3′ (right). Horizontal lines indicate introns that are spliced out between exons (boxes) and–where the annotation of an ORF is available–thick and thin box areas indicate coding respectively non-coding regions of the transcript. The double horizontal lines in the EST BP269256 show a gap in the alignment with the genomic sequence. (**B**) One complete AS event from the depicted locus with $d = 7$, 2 alternative splice sites (s_7, s_8), and a pattern 1[4^8−12], 1[4^10−12], 1[4^11−12], 1[4^13−15^16−17^18−19], 1[4^13−15^18−19], 1[4^13−15^18−19], 2[12], 3[4^11−12], 5(7^8−9), 6[7^11−12], 6[7^13−15^18−19]. Note that the presumptively truncated transcript extremities of the ESTs BM147777 and BM147934 are concatenated and extended to the longest exonic area before the common first donor/after the common last acceptor (Section 3). EST BP269256 describes an invalid variant due to the alignment error and EST AI377407 (respectively the rest of the transcripts that are not present) is not included in the event because it is not overlapping any of the alternative splice sites.

AS events (Sammeth et al., 2008). The basis for Definition 3 is to not restrict the assumption on the molecular mechanisms acting during the splicing process on a premature transcript *a priori* to either exon- or intron-definition and therefore to allow for possible interactions of parts of the splicing machinery across all exons and introns. Consequently, AS events are delimited at *common sites* of the d variants:

Definition 3. *An AS event of dimension $d \geq 2$ comprises $\{p_i\}_{i=1}^{d}$ different variants, such that* (*i*) *the first and the last site of each variant p_i are common sites or the first/last site of each variant is the first/last site of the supporting transcripts,* (*ii*) *any site besides common first/last sites of a variant p_i are variable sites, and* (*iii*) *amongst these variable sites there is at least one alternative splice site.*

Part (i) in Definition 3 determines the genomic region of an event that describes a certain splicing variation, which according to condition (ii) must be *minimal* with respect to the *d* compared variants. If this region is delimited to both ends by common sites, the event is said to be *internal*, otherwise it is an AS event that is possibly connected to differential transcription start/poly-adenylation. Requirement (iii) finally justifies the event to be an "AS event" for which the presence of alternative splice site(s) is indispensable. Parameter $d \geq 2$ determines the resolution when delineating AS events—so far exclusively pairwise events have been considered. If P includes all variants shown by the annotation in the corresponding genomic region, the event is said to be *complete* with respect to the given annotation (Fig. 1B). By this, complete events need to suffice Definition 3 and additionally require the number of variants included to be *maximal* (Definition 4).

Definition 4. *An AS event that contains d variants such that there exists no other variant* $q \notin \{p_i\}_1^d$ *that shares the boundaries with variants* p_i *is a "complete" AS event.*

2.2. A notation for the pattern of AS events with $d \geq 2$

For describing the morphology of AS events, I straightforwardly extend a general notation system we proposed earlier for the characterization of pairwise events (Sammeth et al., 2008). The notation allows for qualitative representation of the variable exon-intron structures, the *pattern*, and shows separately for all *d* variants in *P* the *variable* sites s_i by their relative position $1 \leq i \leq g$ within the event and a symbol identifying *type*(s_i). The number of variable sites *g* (the so-called *degree* of the event) is a key attribute of the polymorphism.

Symbols from $\Sigma = \{\frown, -, [,], (,)\}$ identify the role of a site as *type*(s_i) = donor "$\frown$", acceptor "$-$", start "[", end "]", 5′-truncation "(", or 3′-truncation ")". Because root and leaf sites are never present in $\langle s_i \rangle_{i=1}^g$ (Section 3), they do not require symbols for notation. Strings of alternating integers *i* and symbols for *type*(s_i) are concatenated for each variant *p* of the event, where "0" is used to denote the empty string. Imposing an order on the permutable variants in an event, the *d* strings are subsequently merged lexicographically—with commas separating the variants from each other—producing a string called the "AS code" of the event (Fig. 1B). By this, events generate identical AS codes only if they describe the same pattern.

3. SPLICING GRAPHS REPRESENT SPLICING STRUCTURES NONREDUNDANTLY

As in Sugnet et al., (2004), in this work a *splicing graph* $G(V, E)$ on a genomic locus is a directed acyclic graph (DAG) with each vertex $s \in V$ describing non-redundantly a site of the transcript annotation. Each edge $s \rightarrow t \in E$ consequently represents an exon (*type*(s) ∈ {start, acceptor}) or an intron (*type*(s) = donor) with transcript support *transcripts*($s \rightarrow t$) = *transcripts*(s) ∩ *transcripts*(t). In order to deal with alignment errors, introns with unusual splice site sequences are marked by *valid*($s \rightarrow t$) = **false**. In this work, introns have been considered as "trustworthy" (*valid*($s \rightarrow t$) = **true**) if they matched the sequence combinations (donor/acceptor) GT/AG, GC/AG, ATATC/AG, ATATC/AC, ATATC/AT, GTATC/AT or ATATC/AA as these constitute over 90% of the human introns—including introns spliced by the U12 spliceosome (Alioto, 2007). However, other criteria may be applied to distinguish introns from presumptive alignment errors, for instance the intron length or *class*(s) and *class*(t) of the delimiting sites.

Subsequently, V is completed by inserting two virtual sites, *root* = ($-\infty$, *T*, RefSeq, root) and *leaf* = ($+\infty$, *T*, RefSeq, leaf), which are connected to/from each transcription start/end site: $E \cup (root \rightarrow s) \cup (t \rightarrow leaf)$ for all $s, t \in V$ with *type*(s) = start and *type*(t) = end. As no computational method hitherto can confidently detect 5′- or 3′-complete transcripts, all edges $s_i \rightarrow t$ at the transcript extremities with *type*(s_i) = start, *class*(s_i) = EST are coalesced into one edge $s \rightarrow t, s = minarg_{s_i}(pos(s_i))$, and correspondingly all presumptively truncated edges $s \rightarrow t_j$ with *type*(t_j) = end, *class*(t_j) = EST into one edge $s \rightarrow t, t = maxarg_{t_j}(pos(t_j))$. Corresponding sites ($s$, t) are assigned *type*(s) = 5′-truncation and *type* (t) = 3′-truncation in order to highlight the unreliability of sequence borders that are exclusively supported by EST evidence. This step removes diversity artifacts in the exonic structure of ESTs that only differ from each other in the truncation point of the first/last exon. Additionally, transcription starts (respectively ends) s are replaced by acceptors (donors) if there is evidence for such in other transcript data at *pos*(s). Finally, adopting the technique described in Heber et al., (2002), vertices with $outdeg(s) = indeg(s) = 1$ are collapsed because they are uninformative with respect to the subsequently described technique (Lemma 1).

Lemma 1. *Vertices v with $indeg(v) = outdeg(v) = 1$ are uninformative for delimiting AS events in splicing graphs and can be collapsed without loss of generality.*

Proof. From the minimality criterion for the boundaries in Definition 3 (ii) can directly be deduced that for the boundaries s and t of an AS event holds $outdeg(s) > 1$ and $indeg(t) > 1$. Any vertex v with $indeg(v) = outdeg(v) = 1$ therefore cannot be the delimiting site of an event. Let v and w, $s \preceq u \prec v \prec w \preceq t$ be variable sites of the event and $u \rightarrow v$ be the in-edge and $v \rightarrow w$ be the out-edge of v. Then naturally $transcripts(u \rightarrow v) = transcripts(v) = transcripts(v \rightarrow w) = transcripts(u) \cap transcripts(w)$ holds and the corresponding partition of the variant going through v is equivalently described by a single edge $v \rightarrow w$. ■

On the remaining vertices $s \in V$ (Fig. 2A,B) a pre-order $\preceq$ is defined by extending the natural total order of their genomic position $pos(s)$ as follows:

Definition 5. *The preorder $\preceq$ on the sites $s \in V$ orders them by the total order on their genomic position $pos(s)$ from $5'$ to $3'$ and their type, s.t. $5'$ exon boundaries precede $3'$ exon boundaries annotated at the same genomic position $s \prec t$: $pos(s) < pos(t) \vee (pos(s) = pos(t) \wedge type(s) \in \{\text{start, acceptor}\})$.*

As described earlier, annotated transcription starts and poly-adenylation sites are replaced during graph construction, if the annotation shows splice site evidence at the same genomic position. Consequently, there exist no two vertices $s,t \in V$, $pos(s) = pos(t)$ with $type(s) = \text{start}$, $type(t) = \text{acceptor}$, nor with $type(s) = \text{end}$, $type(t) = \text{donor}$. Obviously, for all $s \in V$, $s \notin \{root, leaf\}$, it holds $root \prec s \prec leaf$.

3.1. Bubbles are subgraphs describing complete AS events

In G, a variant $p = (s, t, X_p)$ is a path $(s, \ldots, t)$ with a non-empty set of transcript support X_p (Definition 1), which excludes all paths in G describing splicing structures that have not been observed in nature. Subsequently, the variants between two vertices $s, t \in V$ are described by paths with s as common tail vertex, t as common head vertex and a set of non-empty partitions $\mathcal{X}_{s,t}$.

Observation 1. *For every pair of vertices $s \prec t$ with $transcripts(s) \cap transcripts(t) \neq \emptyset$ there exists at least one variant $p = (s, t, X_p)$ with $X_p \subseteq transcripts(s) \cap transcripts(t)$. An edge $u \rightarrow v$, $s \preceq u \prec v \preceq t$ with $(transcripts(u \rightarrow v) \cap X_p) \subsetneq X_p$ splits X_p into the partitions $X_{p'} = transcripts(u \rightarrow v) \cap X_p$ and $X_{p''} = X_p \setminus X_{p'}$. Decomposing X_p by iterative intersection with the transcript support of all edges between s and t results in the "partition set" $\mathcal{X}_{s,t}$ containing all partitions of the variants between s and t. Therefore, for the union of transcripts in the partitions of $\mathcal{X}_{s,t}$ naturally holds:*

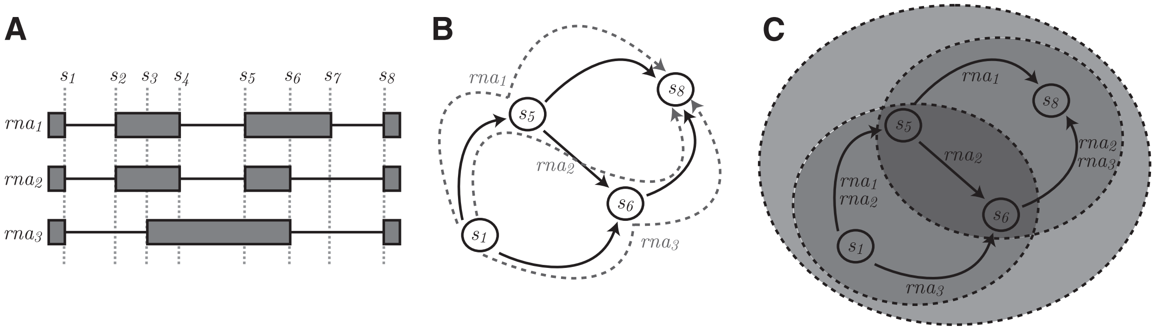


FIG. 2. (**A**) A cutoff from a locus showing $k = 3$ transcripts (rna_1, rna_2 and rna_3) and 8 sites $\langle s_1, \ldots, s_8 \rangle$. The exon-intron structure is shown schematically, i.e., exons (boxes) and introns (lines) are not drawn to scale. Different variants can be observed, for instance $(s_1, s_5, \{rna_1, rna_2\})$. (**B**) The corresponding splicing graph structure after contracting uninformative vertices. Dotted lines indicate the paths supported by single transcripts rna_1, rna_2 and rna_3. (**C**) Ovals highlight all 3 bubbles, that is $(s_1, s_6, \{rna_2\}, \{rna_3\})$, $(s_5, s_8, \{rna_1\}, \{rna_2\})$ and $(s_1, s_8, \{rna_1\}, \{rna_2\}, \{rna_3\})$. In contrast, there exists no bubble between s_5 and s_6 because they are connected by only a single variant (i.e., rna_2).

$$\bigcup_{rna \in X, X \in \mathcal{X}_{s,t}} (rna) = transcripts(s) \cap transcripts(t).$$

Figure 2C for instance shows 2 variants between s_1 and s_6, where $transcripts(s_1) \cap transcripts(s_6) = \{rna_1, rna_3\}$ is split by $transcripts(s_1 \rightarrow s_5) = \{rna_1, rna_2\}$, $transcripts(s_1 \rightarrow s_6) = \{rna_3\}$ and $transcripts(s_5 \rightarrow s_6) = \{rna_2\}$ into partition set $\mathcal{X}_{s_1,s_6} = \{\{rna_1\}, \{rna_2\}\}$. Clearly, $|\mathcal{X}_{s,t}| > 1$ shows that there exist different variants between s and t (Observation 1) and we call the subgraph of G spanned by all these variants a *bubble* (Fig. 2C).

Definition 6. *A "bubble" $(s, t, \mathcal{X}_{s,t})$ is a subgraph of G delimited by the vertices $s \prec t$ that comprises the maximal set of variants between s and t defined by the partition set $\mathcal{X}_{s,t}$, $|\mathcal{X}_{s,t}| \geq 2$* (maximality criterion for the number of variants) *s.t. there exists no $\mathcal{X}_{u,v}$ with $\mathcal{X}_{s,t} \subseteq \mathcal{X}_{u,v}$ for any $s \preceq u \prec v \preceq t$ with* $(s, t) \neq (u, v)$ (minimality criterion for the boundaries). *Vertex s is said to be the "source" and t to be the "sink" of the bubble.*

By Definition 6, bubbles are subgraphs that involve cycles in the undirected graph underlying G, but graph structures described earlier on DAGs[1] do not meet the attributes of a bubble.

Note that $(s, t, \mathcal{X}_{s,t})$ is equivalent to the notation of an AS event $\{p_i\}_{i=1}^d$ for $p_i = (s, t, X_{p_i})$, $\mathcal{X}_{s,t} = \bigcup_{p_i} (X_{p_i})$. It is straightforward to show that each bubble corresponds to a complete AS event (Definition 4). Furthermore, each arbitrary d-dimensional AS event according to Definition 3 is reflected by a combination of variants $\{X_{p_i}\}_{i=1}^d \subseteq \mathcal{X}_{s,t}$ in a bubble of G (Lemma 2).

Lemma 2. *For each AS event $\{p_i\}_{i=1}^d$ there exists a bubble $(s, t, \mathcal{X}_{s,t})$ with a combination of partitions $\{X_{p_i}\}_{i=1}^d \subseteq \mathcal{X}_{s,t}$ that describe the different variants of the event.*

Proof. Clearly, there exists a set of paths in G according to the variants of an AS event $\{p_i\}_{i=1}^d$. As by Definition 3, an AS event of dimension d is a set of sites that are not common to all of the d compared transcripts, flanked by splice sites (s, t) contained in all of them or the transcript start/end. Since the completion of G by (*root*, *leaf*) ensures common vertices also in the case of AS events that include variable transcript extremities, $\bigcap_{p_i} (X_{p_i}) = \{s, t\}$ holds for all AS events according to Definition 3 and consequently there exists a bubble $(s, t, \mathcal{X}_{s,t})$ with $\{X_{p_i}\}_{i=1}^d \subseteq \mathcal{X}_{s,t}$. ■

Moreover, the following can directly be deduced from Definition 6:

Corollary 1. *Bubbles can intersect in vertices and edges.*

Proof. Considering different splicing structures may involve common splice sites, exons or introns, it becomes obvious two variants (p_1, p_2) with $s_1^{p_1} \neq s_1^{p_2} \vee s_{m_1}^{p_1} \neq s_{m_2}^{p_2}$ can intersect in vertices or edges. Given additionally the variants p_3 and p_4 with $p_3 \cap p_1 = \{s_1^{p_1}, s_{m_1}^{p_1}\}$ and $p_4 \cap p_2 = \{s_1^{p_2}, s_{m_2}^{p_2}\}$, then ($p_1$, p_3) and on the other hand (p_2, p_4) are part of different bubbles. ■

Corollary 1 outlines the complexity of overlaps between bubbles. Figure 2C for instance shows 3 edge-intersecting bubbles. A more complex example is shown in Figure 3. It should be stressed that the various overlaps between bubbles spans a hierarchy on the underlying complete AS events that describes how these cascaded components form complex splicing structures. Theorem 1 however shows that there is a unique set of bubbles in G.

[1] "Blobs" described earlier (Gusfield and Bansal, 2005) in phylogenetic networks with recombination cycles are defined as subgraphs involving all edge-intersecting cycles of the underlying undirected graph—thus a blob may coincide with an isolated bubble, but comprises multiple edge-intersecting bubbles (Corrollary 1). Whereas, a single recombination cycle—a "gall" (Gusfield et al., 2004)—does not necessarily describe the complete bubble subgraph as by the maximality criterion for the number of variants in Definition 6 a bubble can contain more than one such cycle *iff* there are >2 variants between its source and sink. "Bulges" and "whirls" (Pevzner et al., 2004) have been described as short cycles with—respectively without—orientation, but may not be adequate as AS events can include subgraphs of considerable sizes—some of them even comprise the complete graph of a locus (Fig. 1).

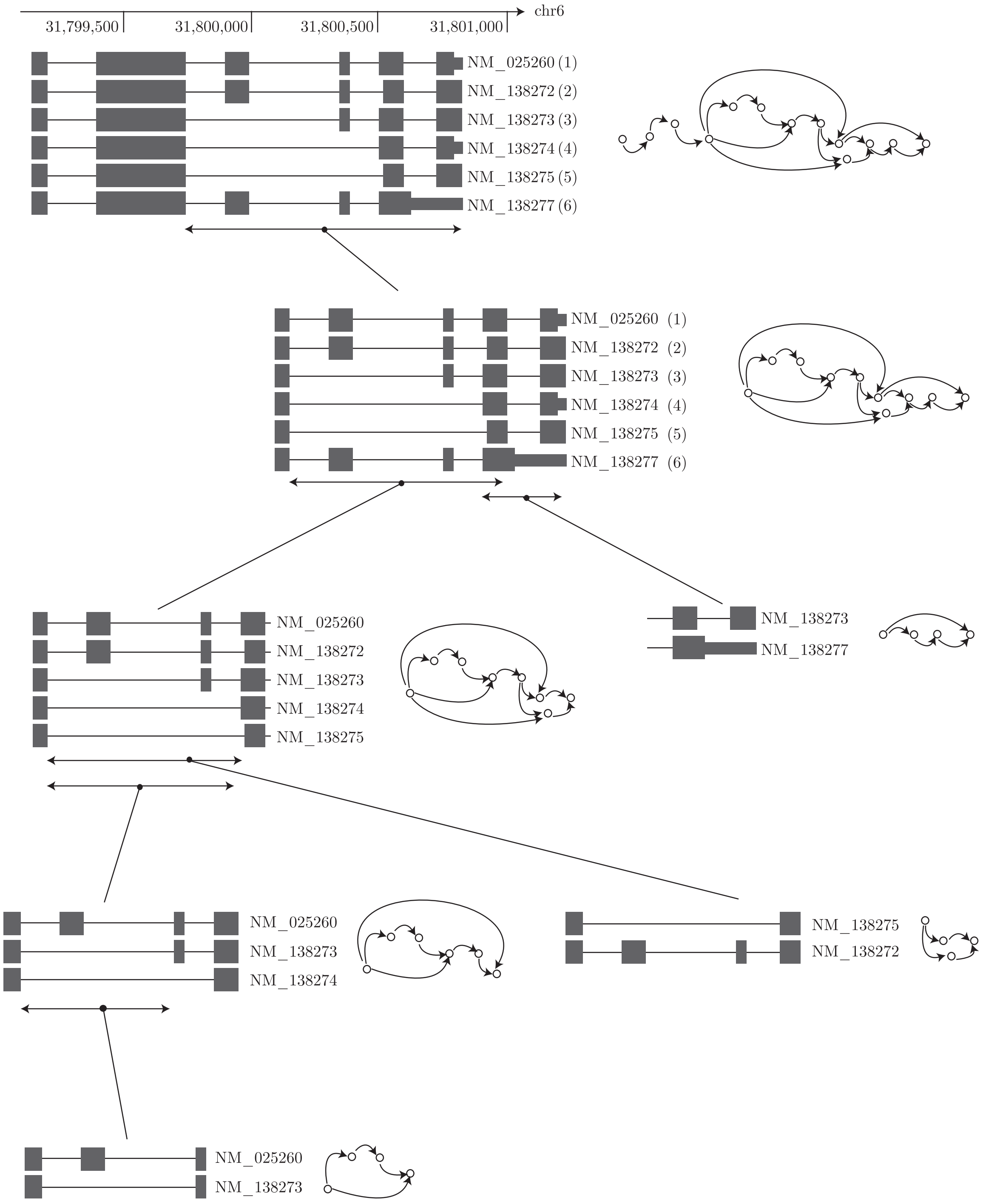


FIG. 3. The hierarchy complete AS events spanned by the transcripts of the RefSeq annotation in C6orf25, a gene of the immunoglobulin superfamily located in the human major histocompatibility complex (MHC class III) region. Each event is depicted by the corresponding transcript structure, and to the right the specific bubble subgraph of G is shown. The genomic overlap between events in the hierarchical cascade is indicated by double arrows.

Theorem 1. *The set of all bubbles contained in G is unique.*

Proof. Definition 6 implies that—given a pair of vertices (s, t), $s \prec t$—there exists none ($|\mathcal{X}_{s,t}| < 2$ or exactly one bubble (containing all $|\mathcal{X}_{s,t}|$ variants) between them. The complete set of bubbles as obtained from all tuples $(s, t) \in V$ is therefore unique and can be obtained by any arbitrary iteration over all $s \prec t$. ■

3.2. Not all variant combinations in bubbles necessarily form AS events

The preceding section has shown that each d-dimensional AS event is reflected by a variant combination in a bubble (Lemma 2). However, the mapping from AS events to d-tuple combinations of variants is *not injective*, hence not necessarily every combination of d variants in a bubble results in an AS event according to Definition 3. A bubble $(s, t, \mathcal{X}_{s,t})$ maximally harbors $\binom{|\mathcal{X}_{s,t}|}{d}$ events of dimension d, i.e., combinations $\{X_{p_i}\}_{i=1}^{d}$. Obviously, for $d > |\mathcal{X}_{s,t}|$ there exists none and for $d = |\mathcal{X}_{s,t}|$ there exists exactly one AS event, which is complete (Definition 4). However, Lemma 3 shows that in the case of $d < |\mathcal{X}_{s,t}|$ there can occur variant combinations that are all intersecting in one or more sites $u \notin \{s, t\}$ additionally to the source/sink and consequently do not describe AS events (Definition 3).

Lemma 3. *Some of the variants in a bubble $(s, t, \mathcal{X}_{s,t})$, $|\mathcal{X}_{s,t}| > 2$ may be intersecting in vertices $\{u, v\} \neq \{s, t\}$), or in edges. Such variants imply the presence of at least one other bubble $(u, v, \mathcal{X}_{u,v})$ with $s \preceq u \prec v \preceq t$. In this specific geometry of overlapping bubbles $(s, t, \mathcal{X}_{s,t})$ is said the "outer bubble" that contains the "inner bubble" $(u, v, \mathcal{X}_{u,v})$.*

Proof. Let the bubble $(s,\ t,\ \{X_{p1},\ X_{p2},\ X_{p3}\})$ contain 3 variants s.t. $p_1 \cap p_2 \cap p_3 = \{s,\ t\}$ and without violating Definition 6 $p_1 \cap p_2 = \{s,\ t,\ u\}$. By $p_1 \neq p_2$ and $s \prec u \prec t$ (Definition 5), p_1 and p_2 differ between s and u, or, between u and t (or in both parts). Correspondingly, there exists a bubble $(s, u, \mathcal{X}_{s,u})$, $\{X_{p_1}, X_{p_2}\} \subset \mathcal{X}_{s,u}$ and/or a bubble $(u, t, \mathcal{X}_{u,t})$, $\{X_{p_1}, X_{p_2}\} \subset \mathcal{X}_{u,t}$. The argumentation can straightforwardly be extended if p_1 and p_2 intersect in more than one vertex, also if these vertices are connected by edges. ■

Examples of outer and inner bubbles are shown in Figures 2C and 4A,B. Variants of inner bubbles are sub-paths of variants in outer bubbles and neither intersection in vertices (e.g., in Fig. 4B), nor intersection in edges (e.g., bubble $(s_1, s_6, \{rna_2\}, \{rna_3\})$ and bubble $(s_5, s_8, \{rna_1\}, \{rna_2\})$ in Fig. 2C) nor the order of the source/sink vertices $s \preceq u \prec v \preceq t$ (e.g., bubble $(s_2, s_4, \{rna_3\}, \{rna_4\})$ compared to either of the other bubbles in Fig. 4C) alone is sufficient for the geometry described in Lemma 3. It can further be shown that variants in inner bubbles are the reason for variant combinations of outer bubbles that violate condition (ii) of Definition 3 (Theorem 2).

Theorem 2. *Given an outer bubble $(s, t, \mathcal{X}_{s,t})$, for any combination $\{X_{p_i}\}_{i=1}^{d}, X_{p_i} \in \mathcal{X}_{s,t}$ that corresponds to a set of variants $\{p_1, \ldots, p_d\}$ intersecting in more vertices than s and t, there exists a combination of partitions $\{X_{p'_1}\}_{i=1}^{d}, X_{p'_i} \in \mathcal{X}_{u,v}$ of an inner bubble $(u, v, \mathcal{X}_{u,v})$ s.t. $X_{p_i} \subseteq X_{p'_i}$ and $\bigcap p'_i = \{u, v\}$.*

Proof. By Lemma 3, an outer bubble $(s, t, \mathcal{X}_{s,t})$ contains $i \geq 2$ variants p_i that in G are super-paths of variants p'_i of its inner bubble $(u, v, \mathcal{X}_{u,v})$. Because partition X_p is the intersection of the transcript support of all sites in p, $X_{p_i} \subseteq X_{p'_i}$ naturally holds. ■

For instance, outer bubble $(s_1,\ s_8, \{\{rna_1\},\ \{rna_2\},\ \{rna_3\}\})$ in Figure 2C gives rise to three variant combinations of $d = 2$. From these, combination of variants $(s_1,\ s_8,\ \{rna_1\})$ and $(s_1,\ s_8,\ \{rna_2\})$ does not

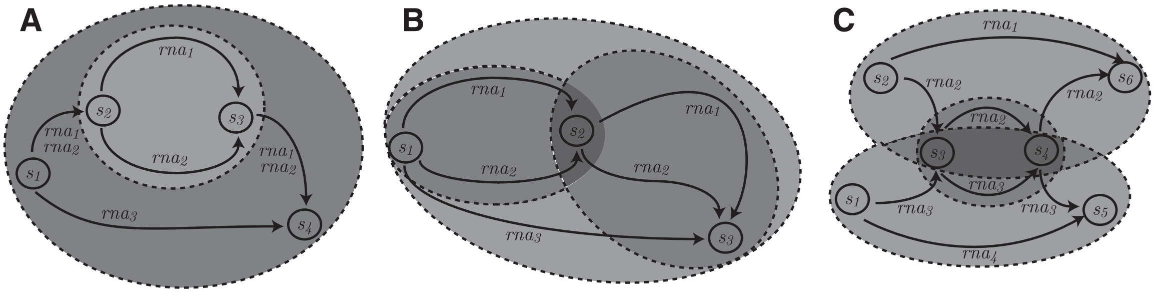


FIG. 4. Subgraphs showing different constellations of edge-intersecting bubbles with sites s_i numbered consecutively in $\preceq$. (**A**) The outer bubble $(s_1, s_4, \{rna_1\}, \{rna_2\}, \{rna_3\})$ contains the inner bubble $(s_2, s_3, \{rna_1\}, \{rna_2\})$. (**B**) The inner bubbles $(s_1, s_2, \{rna_1\}, \{rna_2\})$ and $(s_2, s_3, \{rna_1\}, \{rna_2\})$ are both contained in the outer bubble $(s_1, s_3, \{rna_1\}, \{rna_2\}, \{rna_3\})$. (**C**) Although bubble $(s_3, s_4, \{rna_2\}, \{rna_3\})$ edge-intersects with $(s_2, s_6, \{rna_1\}, \{rna_2\})$ and $(s_1, s_5, \{rna_3\}, \{rna_4\})$, it is not an inner bubble because neither all partitions in $\mathcal{X}_{s_2,s_6}$ nor in $\mathcal{X}_{s_1,s_5}$ are subsets of partitions in $\mathcal{X}_{s_3,s_4}$.

form an AS event because the variants intersect in $\{s_1, s_5, s_8\}$ (Definition 3). However, corresponding sub-paths of the variants $(s_5, s_8, \{rna_1\})$ and $(s_5, s_8, \{rna_2\})$ in the inner bubble do describe an AS event. Also, variants $(s_1, s_8, \{rna_2\})$ and $(s_1, s_8, \{rna_3\})$ are not constituting an AS event because they are super-paths of variants $(s_1, s_6, \{rna_2\})$ respectively $(s_1, s_6, \{rna_3\})$–which in turn represent an AS event. Exclusively variant combination $(s_1, s_8, \{rna_1\})$ and $(s_1, s_8, \{rna_3\})$ of bubble $(s_1, s_8, \{rna_1\}, \{rna_2\}, \{rna_3\})$ represents an AS event. Correspondingly, variants $(s_1, s_4, \{rna_2\})$ and $(s_1, s_4, \{rna_2\})$ in Figure 4A, and $(s_1, s_3, \{rna_1\})$ in combination with $(s_1, s_3, \{rna_2\})$ in Figure 4B do not describe AS events.

4. AN EXACT METHOD FOR THE EXHAUSTIVE EXTRACTION OF AS EVENTS FROM TRANSCRIPT ANNOTATIONS

I now present an algorithm to extract all events of dimension d from a splicing graph G. Initially one would try straightforwardly to iterate all sites in G in genomic order, collecting variants of bubbles $(s, t, \mathcal{X}_{s,t})$ starting at the source with $outdeg(s) > 1$. However, due to complex intersections of bubbles, the number of variants grows often exponentially when proceeding this way, and adequate measures to limit memory have to be taken.

4.1. An algorithm for retrieving AS events with arbitrary dimension and memory bounded to output size

Algorithm 1 iterates for all possible sinks of bubbles t in genomic order $\preceq$ (Definition 5) over preceding edges $s \rightarrow u$, $s \prec t$ in reverse genomic order $\succeq$, until the $5'$-most source vertex with $transcripts(s) \supseteq transcripts(t)$ is reached. This approach tackles the problem of exponentially growing paths by finding the bubbles in the graph structure one by one, limiting at any time the actual memory to the output size, i.e., the currently extracted bubble. During the inner iteration, $\mathcal{X}$ stores the current set of partitions which is constantly subdivided by INTERSECT() as new vertices s with $outdeg(s) \geq 2$ are iterated (Lemma 1 and Observation 1). Initially, $\mathcal{X} = \{transcripts(t)\}$ consists of a single partition. Collection $\mathcal{C}$ stores d-tuples of partitions that are excluded from generating AS events in outer outer bubbles because corresponding sub-paths have formed an AS event in an inner bubble (Theorem 2).

Algorithm 1 RetrieveASevents(V, E, d)

```
Input  : A DAG $G(V, E)$ and the dimension of the AS events $d$.
Output: All AS events of dimension $d$ reflected by $G$.
for $all\ vertices\ t \in V,\ indeg(t) \geq d\ (in\ genomic\ order \prec)$ do
  $\mathcal{X} \leftarrow \{transcripts(t)\}$
  $\mathcal{C} \leftarrow \emptyset$
  for $all\ in\text{-}edges\ v \rightarrow t$ do
    $\mathcal{C} \leftarrow transcripts(v \rightarrow t)$
  for $all\ vertices\ s \prec t,\ outdeg(s) \geq 2\ (in\ reverse\ genomic\ order \succ)$ do
    $\mathcal{X}_{s,t} \leftarrow \emptyset$
    for $all\ s \rightarrow u \in E$ do
      if $valid(s \rightarrow u)$ then
        $\mathcal{X}_{s,t} \leftarrow \mathcal{X}_{s,t} \cup$ INTERSECT($transcripts(s \rightarrow u),\ \mathcal{X}$)
      REMOVE($transcripts(s \rightarrow u), \mathcal{X}$)
    if $|\mathcal{X}_{s,t}| \geq d$ then
      EXTRACTEVENTS($s, t, \mathcal{X}_{s,t}, d, \mathcal{C}$)
    $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{X}_{s,t}$
    if $transcripts(t) \subseteq transcripts(s)$ then
      break
```

For any vertex s iterated in the inner loop, transcript support of valid out-edges $s \rightarrow u$ is intersected with the earlier found partitions in $\mathcal{X}$ to determine the new partition set $\mathcal{X}_{s,t}$ (Observation 1). Subsequently, EXTRACTEVENTS() outputs AS events as implied by all d-tuples in $\mathcal{X}_{s,t}$ that are not yet contained in $\mathcal{C}$ and for which the corresponding variants describe an alternative splice site according to condition (iii) in

Definition 3. The variant tuples of newly extracted AS events then are added to $\mathcal{C}$ in order to exclude them from generating AS events in further iterations of the inner loop.

4.2. *Correctness of the algorithm*

To assess the correctness of Algorithm 1 I firstly show that it recovers all variants between the source and the sink of the currently iterated bubble.

Lemma 4. *The partition set $\mathcal{X}_{s,t}$ between two vertices s and t generated by Algorithm 1 is complete and corresponds to the partition set of all different variants between s and t.*

Proof. Algorithm 1 initializes the partition set with *transcripts*(t) and recursively subdivides the partitions INTERSECT() with *transcripts*($s \rightarrow u$) as the inner loop proceeds–yielding the partitions of all variants between s and t as stated in Observation 1 which concludes the proof. ∎

By Theorem 2, variant combinations of outer bubbles that are super-paths of inner bubbles' variants do not give rise to AS events. To prove the correctness of Algorithm 1 it remains to be shown how such combinations are prevented:

Lemma 5. *Algorithm 1 does not consider variant combinations of outer bubbles with sub-paths that are part of an inner bubble.*

Proof. Given an outer bubble $(s, t, \mathcal{X}_{s,t})$ and an inner bubble $(u, v, \mathcal{X}_{u,v})$, two cases are distinguished. (i) $v = t$, e.g., bubble $(s_2, s_3, \mathcal{X}_{s_2,s_3})$ in Fig. 4B): by the iteration order over sink vertices in genomic order $\prec$ and over source vertices in reverse genomic order $\succ$, $(u, v, \mathcal{X}_{u,v})$ is iterated before $(s, t, \mathcal{X}_{s,t})$ and partition combinations are accordingly stored in $\mathcal{C}$. (ii) $v \succ t$: if $v \rightarrow t \in E$, the initialization of $\mathcal{C}$ with the transcript support of all in-edges of t prevents from variant combinations that are super-paths of variants in an inner bubble, as in bubble $(s_2, s_3, \mathcal{X}_{s_2,s_3})$ in Figure 4A; otherwise, for instance in bubble $(s_1, s_2, \mathcal{X}_{s_1,s_2})$ in Figure 4B, there exists a bubble $(v, t, \mathcal{X}_{v,t})$ as inner bubble of $(s, t, \mathcal{X}_{s,t})$ and the problem is reduced to case (i). ∎

Given Lemma 4 and Lemma 5, it can be shown that the set of bubbles found by Algorithm 1 is complete and non-redundant:

Theorem 3. *Algorithm 1 finds all bubbles in G exactly once and extracts AS events of a given dimension d that comply with Definition 3.*

Proof. The main double loop considers all possible boundaries ($s \prec t$) of bubbles. Every source/sink pair (s, t) is iterated exactly once, hence no bubble is found twice by the procedure. Lemma 4 demonstrates that Algorithm 1 retrieves all variants of a bubble between s and t, hence it can be concluded that all bubbles in G according to Theorem 1 are found. For each bubble $(s, t, \mathcal{X}_{s,t})$, all $\binom{|\mathcal{X}_{s,t}|}{d}$ variant combinations satisfying condition (i) of Definition 3 are considered. Lemma 5 shows that tuples contradicting condition (ii) are in every case excluded. Furthermore, a check for the presence of at least one alternative splice site in EXTRACTEVENTS() ensures that the extracted AS events fulfill condition (iii) of Definition 3. ∎

4.3. *Implementation and complexity estimation*

Algorithm 1 has been implemented as described in JAVA (compliance level 1.5), and the resulting executable is freely available from the webpage http://genome.crg.es/astalavista. Technical optimizations include the parallelization of operations in Algorithm 1 from program parts that handle the I/O. Parsing the input data and performing the clustering of genes into loci may take a considerable amount of time—given that the input file including EST data in human is nearly as big as the complete sequence of the human genome—as can be consumed by writing many AS events to disk. Time benchmarks with this parallelized program therefore reflect about the algorithmic time effort. Furthermore, operations on transcript sets (e.g., intersection, unity, comparison) are improved by encoding the respective transcript support in bit-arrays

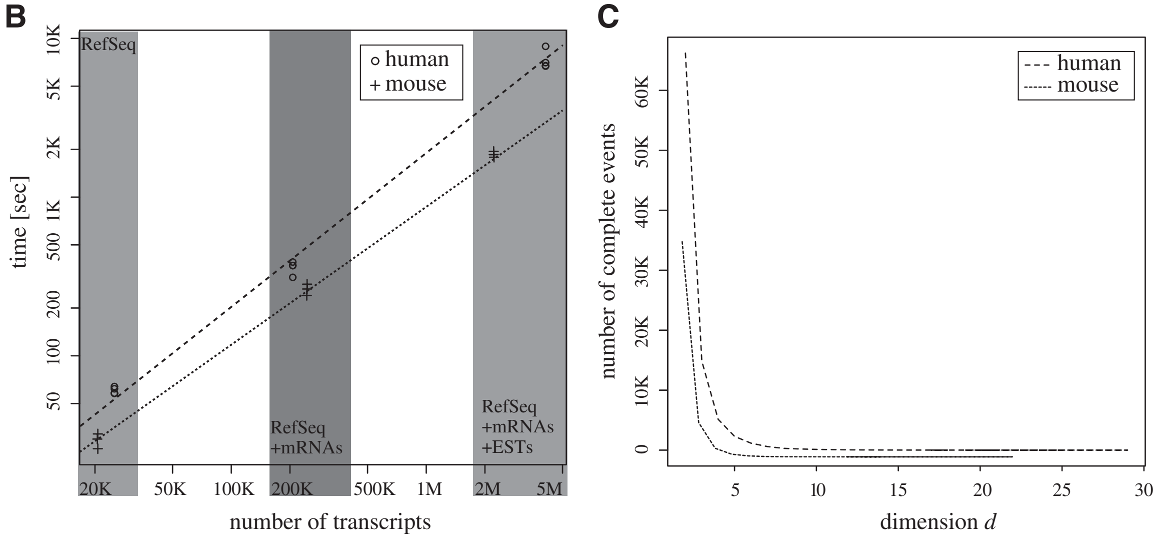
A

	human			mouse		
dataset	RefSeq	+mRNA	+ESTs	RefSeq	+mRNA	+ESTs
transcripts	25,161	206,779	4,093,918	20,618	244,882	2,219,200
time (reference)	0:00:30	1:10:24[a]	–[b]	0:00:14	1:19:58[c]	–[b]
events $d = 2$	9,477	34,603	134,875	2,638	19,127	52,508
time	0:01:04	0:05:30	2:28:36	0:00:30	0:04:23	0:32:02
events $d = 3$	7,173	17,294	97,193	2,669	4,839	15,896
time	0:00:58	0:05:10	1:51:34	0:00:26	0:04:00	0:30:54
events $d = 4$	21,620	43,476	153,292	9,321	10,570	17,025
time	0:01:02	0:05:12	1:57:22	0:00:26	0:04:21	00:27:30
complete events	6,493	26,412	91,117	1,787	16,184	42,198
time	0:00:58	0:05:10	1:55:12	0:00:32	0:04:42	0:29:53

[a] skipping 5 loci with > 1000 transcripts, i.e., a locus on chr2 (88,937,526–89,411,301) encoding parts of antibodies—mostly variable regions (2,645 transcripts), the locus on chr14 (21,180,949–22,090,938) encoding the T-cell receptor α chain (1,493 transcripts), a locus on chr14 (104,896,270–106,354,328) containing genes for several immunoglobulin heavy α, β and γ chains (9,766 transcripts), the locus on chr7 (141,647,256–142,210,559) coding for the T-cell receptor β chain (2,934 transcripts) and the locus on chr22 (20,710,462–21,595,078) encoding the immunoglobulin λ chain (1,931 transcripts).

[b] test run exceeded memory (>16GB) or time limits (5 days).

[c] skipping 9 loci with > 500 transcripts, i.e., 5 loci on chr6 (with 629, 866, 878, 1,167 respectively 1,198 transcripts) encoding parts of antibodies (IgG *kappa* chain, variable regions, etc. ...), the locus of the Trpm1 gene on chr7 (510 transcripts), the locus of the Eef2 gene on chr10 (548 transcripts), an immunoglobulin locus on chr12 (10,275 transcripts), and a locus with several olfactory receptor genes on chr14 (1,196 transcripts).



where each bit indicates whether a corresponding transcript of the locus is contained in the partition or not. Memory requirements of storing a bit for every transcript in the input are in practice negligible compared to the gained speed up. The current implementation runs optimally on a 64-bit system architecture with two or more CPUs, but performs nearly as effective on 32-bit systems with hardware resources that support parallel processes (e.g., dual core processors).

To estimate the complexity of Algorithm 1 consider that in each of the $\mathcal{O}(|V|)$ iterations of the outer loop, the transcripts of at most $\mathcal{O}(|E|)$ edges have to be compared to the transcripts in the current partition-set $\mathcal{X}$. Let *outdeg*(s) be the average out-degree of a vertex $s \in V$, $|$ *transcripts*$(s \rightarrow t)|$ be the average number of transcript support for an edge, and $|\mathcal{X}|$ be the average size of the partition set between two vertices. Assuming comparison of transcript sets in about constant time, complexity for retrieving all bubbles can be estimated by $\mathcal{O}(|V| \cdot |transcripts(s \rightarrow t)| \cdot |\mathcal{X}|)$ which, considering the reciprocal relation $|transcripts(s \rightarrow t)| \sim \frac{k}{|\mathcal{X}|}$, can be approximated by $\mathcal{O}(|V|k)$. Disregarding the update time of $\mathcal{C}$ and $\mathcal{X}$, and further estimating a constant time effort for the check for valid partition combinations in $\mathcal{C}$, time complexity of EXTRACTEVENTS() is determined by the output size $\sum_{(s,t)\in V}(|\mathcal{X}_{s,t}|^d)$ plus some additional overhead for iterating variants at the transcript extremities that are lacking an alternative splice site and therefore do not produce AS events. Space requirements for storing the splicing graph as given in Algorithm 1 are $\mathcal{O}(k(|V| + |E|))$, but can technically be reduced to $max(\mathcal{O}(k|V|), \mathcal{O}(k|E|))$ when exclusively labelling edges or equivalently vertices with the corresponding transcript support. Additionally, another $\mathcal{O}(|\mathcal{X}| \cdot |transcripts(s \rightarrow t)|) \sim \mathcal{O}(k)$ is required to store the current partition set in the double main loop of Algorithm 1.

5. COMPLETE AS EVENTS IN THE HUMAN AND MURINE ANNOTATIONS

For the subsequent analyses, I adopted the annotations of transcribed sequences in human and in mouse as downloaded from the UCSC Genome Browser (2009). These contain transcripts of the NCBI reference sequence database (RefSeq) (Pruitt et al., 2007), the mRNAs in GenBank (Benson et al., 2007), and ESTs from the GenBank subset called dbEST (Boguski et al., 1993) aligned to the genomes (reference sequence hg18 produced by the centers of the Human Genome Sequencing Consortium [2009], respectively, mm8 by the Mouse Genome Sequencing Consortium [2009]) using the program blat (Kent, 2002). These two organisms have the largest amount of EST data available in dbEST (8,134,045 ESTs for human, respectively 4,850,243 for mouse), of which I took the subset of ESTs that show signs of splicing—as specified by UCSC (i.e., "intronESTs"—a track containing spliced ESTs with $\geq$1 introns). Subsequently, Algorithm 1 has been applied iteratively for each locus in both datasets.

5.1. The information of mRNAs and ESTs about new splice forms is highly redundant

As a proof of principle, I applied the program to transcript annotations of different sizes (i.e., RefSeq, RefSeq + mRNAs and RefSeq + mRNAs + ESTs in human and mouse) and observed the time needed to explore the AS diversity found for different dimensions $d \in \{2, 3, 4\}$. Figure 5 summarizes the results, supplementary data files with the corresponding events are available from the program homepage at http://genome.crg.es/astalavista.

FIG. 5. (**A**) Benchmark of computational times for retrieving events of different dimensions *d* from the human and the murine RefSeq annotation, a dataset containing RefSeq transcripts and mRNAs ("+mRNA") and a dataset containing RefSeq, mRNAs and ESTs ("+ESTs"). The number of transcripts shows the size of each dataset. Additionally, the time (hh:mm:ss) of a reference method for pairwise AS event extraction is shown (Sammeth et al., 2008). (**B**) Relation between the number of transcripts in the input and the running time of the implementation of Algorithm 1 for the 3 datasets. Datapoints (circles for human and crosses for mouse) are derived from the extraction of events with dimension $d = 2$, $d = 3$ and $d = 4$ in the respective dataset. Packages from the R software (R Development Core Team, 2007) have been used when creating the plot. (**C**) Exponentially decreasing relation between dimension *d* of an event and the number of complete AS events with *d* variants in the human (dashed line) and the mouse (dotted line) transcriptome annotation.

Figure 5A shows characteristics of the running time for input datasets of different size, and also for varying event dimensions $d \in \{2, 3, 4\}$ or complete events. To give a comparison on the times measured, I extracted pairwise events using the implementation described for the AStalavista web server (Foissac and Sammeth, 2007) (i.e., "reference" in Fig. 5A). The time complexity of the reference is quadratic with respect to the number of transcripts in the input, as well as quadratic with the number of AS events found in each locus (Sammeth et al., 2008). Note that the number of events found by this reference method differs as (i) events are filtered for introns with non-canoncial (GT/AG) introns in the *variable* part of an AS event, and (ii) the method does not perform the confidence check for the transcription start/end sites of mRNAs or ESTs and therefore finds a magnitude of variations that stem from truncated transcript data. Therefore the times for runs of the reference method actually are not directly comparable, but it is clear that—although performing faster in the small RefSeq dataset of human and mouse, the reference is much slower then herein described technique when the size of the input/output size grows (i.e., when including mRNA data), and it is inapplicable to EST data. Note that the implementation of the method described herein deals with the biggest dataset (i.e., human RefSeq + mRNA + EST) in less than 2.5h, including problematic loci—some of them containing 10^4–10^5 annotated sequences like loci generating variety for the immunologic defense (e.g., Ig-chains, T-cell receptors)—which form large splicing graphs with a high degree of "invalid introns" that are observed due to rearrangements on DNA level. A considerable part of the running time is invested in scanning these loci for the comparatively little amount of AS events they contain. Still, the overall event retrieval rate is about 15 events/second–which grows to several hundred events per second when omitting problematic loci (data not shown). When retrieving AS events of higher dimension, the time needed for additional combinations theoretically grows exponentially with d (Section 4.3); however, obviously there is a compensation effect by the time gained when disregarding all bubbles with $|\mathcal{X}_{s,t}| < d$. Indeed, the number of complete AS events with d variants decreases in an exponential manner when considering higher dimensions d (Fig. 5C).

5.2. *More than a quarter of the AS events in human involve more than two variants*

The results of Section 5.1 show for human 24,904 new events—clustered in 6,945 different structures—which constitute >27% of all events and describe splicing variations comprising more than 2 alternative variants. Table 1 shows the patterns of the 50 most frequent internal events (Section 2.1) in human. Motivated by the large fraction of events with a true dimension $d > 2$, I set off to explore up to which degree pairwise transcript comparison conducted usually provides an adequate picture of the true splicing complexity. To this end, I compare for the human annotation the number of the 5 hitherto analyzed AS patterns (i.e., exon skipping, intron retention, alternative donor/acceptor sites and mutually exclusive exons) found in complete events with the number obtained by projecting all splicing variations to events of dimension $d = 2$ (Table 2).

As can be seen from Table 2, the fraction of splicing variations that are correctly described by a single pairwise event varies substantially amongst the different AS patterns. Whereas $\sim\frac{1}{4}$ of the pairwise AS events that describe retained introns or alternative donors/acceptors are part of variations between more than two transcripts, this holds >40% of the skipped exons. Strikingly, most (~94%) of the splicing variations that in pairwise transcript comparisons lead to the observation of "mutually exclusive exons" are in reality part of events with $d > 2$.

Figure 6 summarizes the most abundant complete AS events with $d > 2$. Obviously, the majority of alleged exon skipping events in reality is part of variations where another exon upstream or downstream is also alternatively included in some transcript evidence. These events are located in the tab "skipped exon" of Figure 6 and constitute in total 4,928 (~28%) of the "wrong" pairwise events. Another substantial part of $d = 2$ skipped exons (~26% = 4,425 events) indeed co-occurs with splice donor/acceptor variations in the upstream/downstream exon (intersecting area between tab "alternative donor/acceptor sites" and "skipped exon" in Fig. 6). The latter events also make up about half of the erroneously observed alt. donors/acceptors of which the rest mainly involves structures exhibiting more than one option for the donor/acceptor site (in total 3,646 events in Fig. 6). Retained introns involved in events with $d > 2$ often co-occur with variable donors/acceptors (886, that is nearly half of the retained introns observed when projecting to $d = 2$). Finally, 2,032 presumptive mutually exclusive exons are in AS events that show optional inclusion of 2 or 3 neighboring exons (Fig. 6). Section 5.4 further elaborates on interesting differences in properties between these patterns.

Table 1. 50 Most Frequent Patterns of Internal Complete Events Found in Human

Rank	*AS code*	*No. of events*	*Description*
1	0, 1−2^	24,547	skip 1 exon
2	1−, 2−	14,315	2 alternative acceptors
3	1^, 2^	13,647	2 alternative donors
4	0, 1^2−	5,990	retain 1 intron
5	0, 1−2^3−4^	2,023	skip 2 exons
6	0, 1−3^, 2−3^	1,478	skip 1 exon that has 2 possible acceptors
7	0, 1−2^3−4^, 3−4^	1,358	skip 2 exons or only the second one
8	0, 1−2^, 1−2^3−4^	1,106	skip 2 exons or only the first one
9	0, 1−2^, 1−3^	890	skip 1 exon that has 2 possible donors
10	0, 1−2^, 3−4^	849	include 0 or 1 of 2 alternative exons
11	1−, 2−, 3−	723	3 alternative acceptors
12	1^, 2^, 3^	681	3 alternative donors
13	1−2^, 1−2^3−4^, 3−4^	460	include 1 or 2 of 2 optional exons
14	1−2^3−, 3−, 4−	436	retain 1 exon sometimes if an alternative proximal acceptor downstream is used
15	0, 1−2^3−4^5−6^	395	skip 3 exons
16	1*2^, 3*4^	385	2 alternative first exons
17	0, 1−2^, 1−2^3−4^, 3−4^	377	include 0,1 or 2 of 2 optional exons
18	1^3 − 4^, 2^	358	skip 1 exon if an alternative proximal donor upstream is used
19	1^4−, 2^3−	338	for an intron combine upstream donor with downstram acceptor, downstream donor with upstream acceptor
20	0, 1^2−, 1^3−	331	retain 1 intron that has 2 possible acceptors
21	1−2^, 3−4^	327	2 mutually exclusive exons
22	0, 1^2−3^4−	314	retain 2 introns
23	1^, 2^3−4^	282	skip 1 exon if an alternative distal donor upstream is used
24	1−2^3−, 4−	277	skip 1 exon if an alternative distal acceptor downstream is used
25	1−2^3−, 1−2^4−, 3−	275	skip 1 exon sometimes if an alternative proximal acceptor downstream is used
26	1^, 1^3 − 4^, 2^	269	retain 1 exon sometimes if an alternative distal donor upstream is used
27	1^, 2^, 2^3 − 4^	265	retain 1 exon sometimes if an alternative proximal donor upstream is used
28	1^, 1^2−3^, 3^	250	2 alternative donors with an alternative intron in between
29	1−2^4−, 3−	230	skip 1 exon if an alternative proximal acceptor downstream is used
30	1−, 1−2^3−, 3−	226	2 alternative acceptors with an alternative intron in between
31	1−2^4−, 3−, 4−	224	retain 1 exon sometimes if an alternative distal acceptor downstream is used
32	1^3−4^, 2^, 2^3−4^	220	skip 1 exon sometimes if an alternative proximal donor upstream is used
33	1^, 1^3 − 4^, 2^3−4^	210	skip 1 exon sometimes if an alternative distal donor upstream is used
34	1^4−, 2^3−	193	for an intron combine upstream donor with upstream acceptor, downstream donor with downstream acceptor
35	0, 1−2^3−4^5−6^7−8^	178	skip 4 exons
36	0, 1−4^, 2−4^, 3−4^	161	skip 1 exon that has 3 possible acceptors
37	1−2^3−, 1−2^4−, 4−	158	skip 1 exon sometimes if an alternative distal acceptor downstream is used
38	1*2^4−, 3*	136	alternative transcription start in the first intron
39	1^3−, 1^4−, 2^3−	129	for 2 donors and 2 acceptors around an intron combine all but downstream donor with downstream acceptor
40	0, 1−2^3−4^5−6^, 3−4^5−6^	124	skip 3 exons or only the first one
41	1^3−, 2^3−, 2^4−	123	for 2 donors and 2 acceptors around an intron combine all but upstream donor with downstream acceptor
42	1^3−4;, 2;	122	alternative poly−adenylation site in the retained last intron
43	1−2;, 3−4;	89	2 alternative last exons

(*continued*)

Table 1. (Continued)

Rank	*AS code*	*No. of events*	*Description*
44	1^3–, 1^4–, 2^4–	87	for 2 donors and 2 acceptors around an intron combine all but downstream donor with upstream acceptor
45a	0, 1–2^3–4^, 1–2^3–4^5–6^	82	skip 3 exons or only the last one
45b	0, 1–2^, 3–4^, 5–6^	82	include 0 or 1 of 3 alternative exons
46	0, 1–2^3–4^5–6^7–8^9–10^	81	skip 5 exons
47	1^4–, 2^3–, 2^4–	80	for 2 donors and 2 acceptors around an intron combine all but upstream donor with upstream acceptor
48	0, 1^2–4^4–, 1^4–	79	skipped exon in retained intron
49	0, 1–2^3–4^5–6^7–8^9–10^11–12^	81	skip 6 exons

For each category, the number of single events is given and AS code for the pattern, along with a verbose description.

5.3. The number of complete events in the human and in the mouse annotation continues growing

I conducted a retrospective study to estimate up to which degree the spectrum of AS events we observe today has been available over the last years. Figure 7 depicts the growth of complete events with dimension 2, 3, 4, and 5 in Genbank since the 1990s, separately for human (A) and for mouse (B). First sequence submissions, that have been available as early as 1983, involved transcript evidence from other species (i.e., "xeno") that are disregarded in this study. The annotation growth of "native" sequences for both species is shown in Figure 7C, and Figure 7D depicts the rate of event discovery normalized according to the amount of sequence submissions.

In Figure 7A,B, the amount of novel events discovered shows a direct correlation to the number of sequences that have been submitted to Genbank in each year (Fig. 7C). Since about the year 2000, discovery rate seems to approach saturation at 10 events per 1000 sequences (Fig. 7D). As Genbank submissions are redundant, a drop in the discovery rate is expected by the time annotations are close to completion. Therefore current data suggests that the number of complete events will grow with about constant rate with additional submissions in the near future, but technological advances as high-throughput sequencing may lead more rapidly to a falling slope in discovery rate.

5.4. Introns between mutually exclusive exons are shorter than introns between two alternatively included exons

Amongst the variations that involve 2 alternatively included neighboring exons that are, 4 major groups that differ in structure can be distinguished: group 1—events that include strictly one of the exons (327 cases, ~6%); group 2—events that show inclusion of none or exclusively one of the alternative exons (849 cases, 15%); group 3—events that include one or both of them (460 cases, ~8%); and group 4—events that include one, both or none of the alternative exons (377 cases, ~7%). Usually, the picture of "mutually exclusive exons" fits the structure of group 1, however, one may still term the exons of events in group 2 as "mutually exclusive" since there is no transcript evidence including both of them.

Table 2. For Each of the Five types of AS Events Usually Considered in Literature, the Number of Corresponding Structures Found in Pairwise Events ($d = 2$) Is Shown in Comparison to the Fraction of Them Obtained When Considering Complete Events

Structure	*$d = 2$ events*	*Complete events*	*Fraction*
Skipped exon	42,054	24,547	58.4%
Alternative acceptors	19,382	14,315	73.9%
Alternative donors	17,727	13,647	77.0%
Retained intron	7,939	5,990	75.5%
Mutually exclusive exons	5,567	327	5.9%

As can be seen, the fraction of exons that are correctly observed as "mutually exclusive" in pairwise comparisons is especially low.

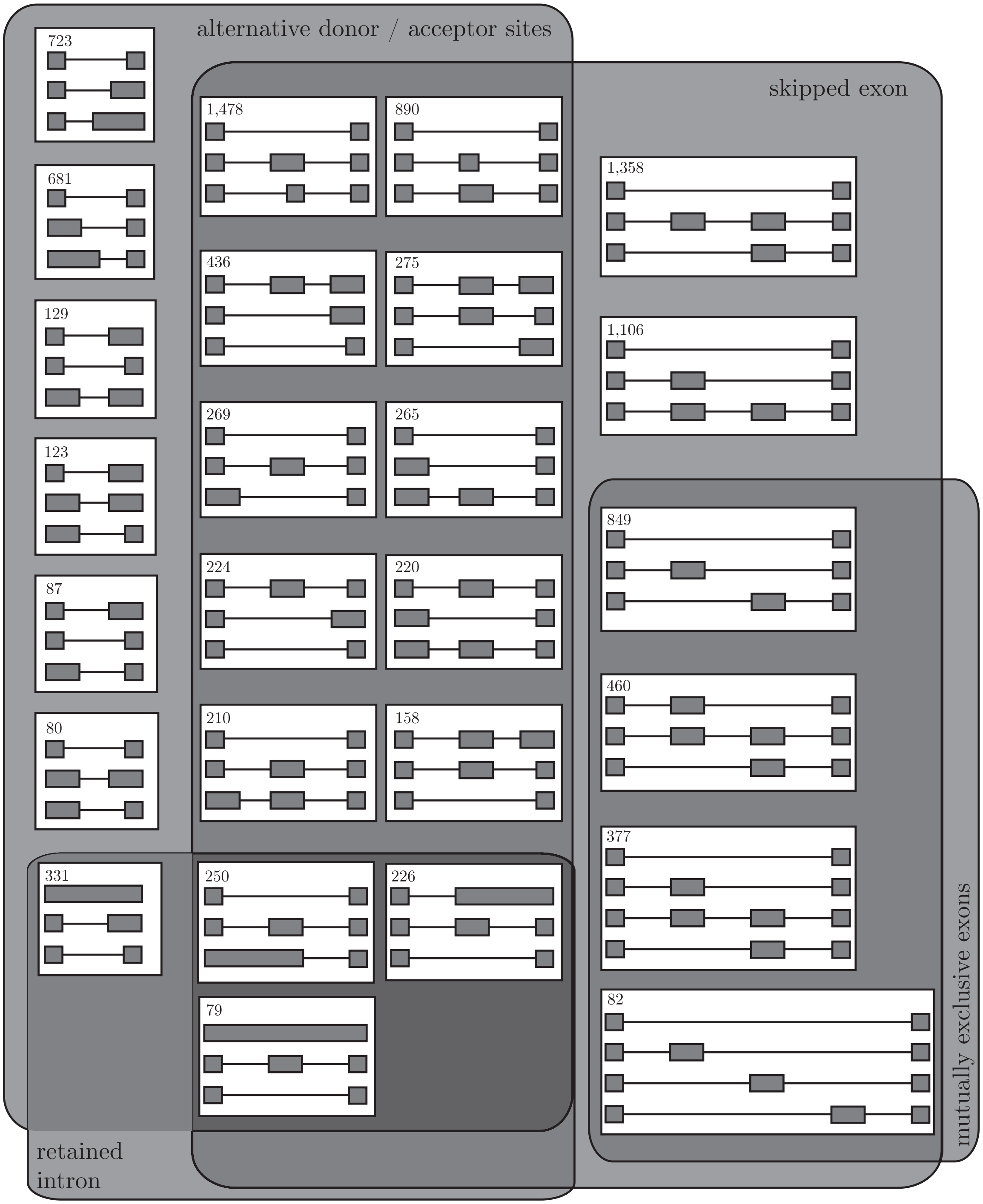


FIG. 6. AS events with $d > 2$ found in the 50 most abundant complete events of the human transcriptome. In pairwise projections these examples show multiple instances of alternative donors/accecptors, skipped exons, retained introns and mutually exclusive exons. Above each pictogram, the number of events with the corresponding structure is shown.

Figure 8 shows the distribution of intron lengths separately for the 3 introns of events in group 1 to 4 in coding regions of the human genome—which clearly differs: events of group 1 show significantly shorter 2nd introns (median 1,100 nt, p-value $\sim 5\mathrm{e}^{-7}$ two-sample Kolgomorov-Smirnov test) than the two flanking introns (median 1,929 nt). In contrast, middle introns in group 4 are only slightly shorter (median 2,059 *vs* 2,246 nt, p-value > 0.69). Events of group 2 exhibit 2nd introns that are a bit–but not significantly–longer

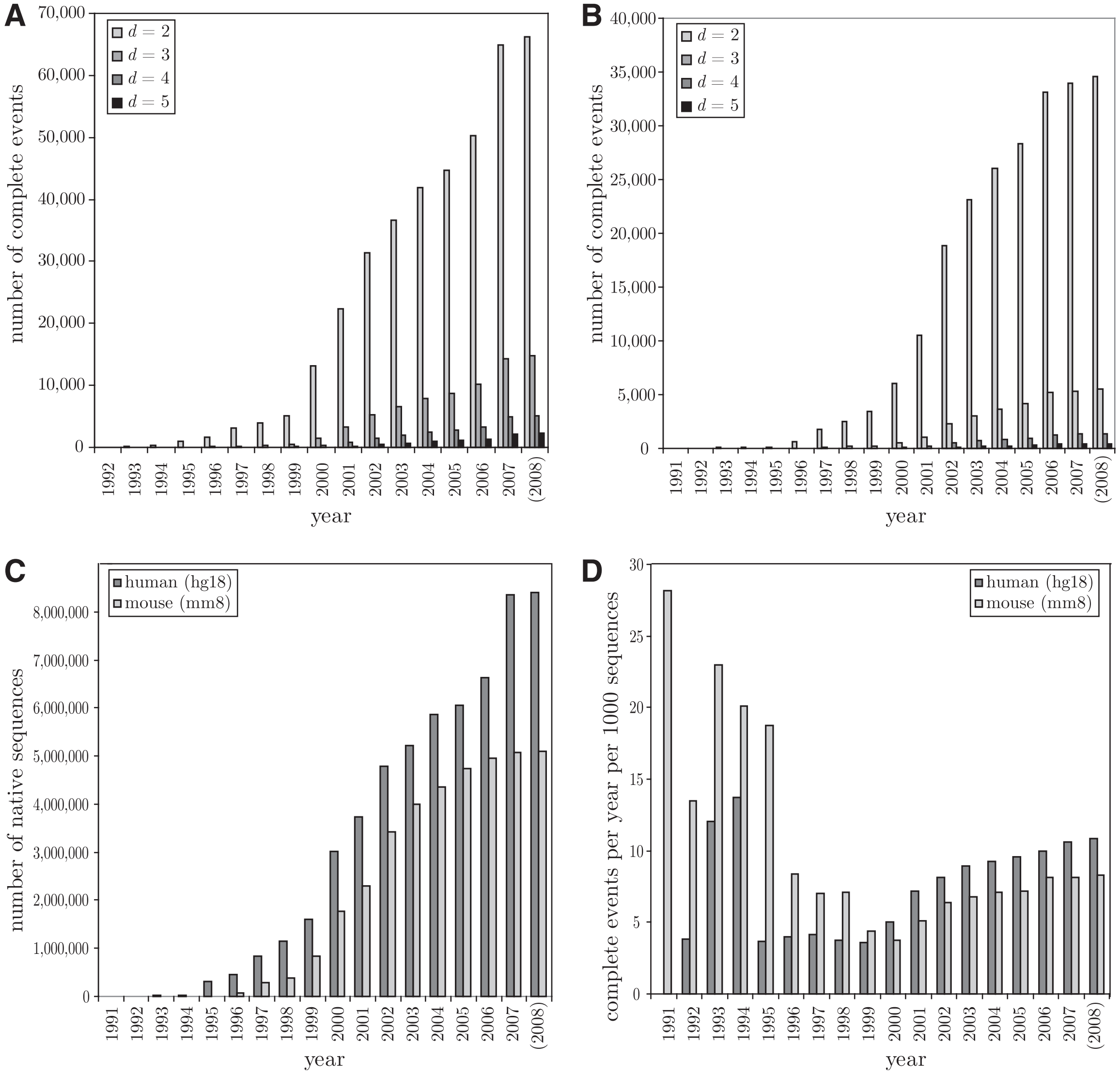


FIG. 7. Retrospective of annotations that havve been available in Genbank (including RefSeq and dbEST) over the last 15 years, for human and mouse. The growth of the number of complete events with dimensions 2, 3, 4, and 5 discovered in the human annotations (**A**) and in mouse (**B**). The number of complete events that has been available in each year is counted cumulatively. (**C**) The number of sequences that have been available from Genbank every year. (**D**) Discovery rate measured as number of complete events that can be detected in a 1000 sequences.

(median 2,897 *vs* 2,342 nt, p-value > 0.13), whereas group 3 contains clearly longer 2nd introns (3,288 *vs* 2,305 nt, p-value $\sim 1.3^{-4}$).

A deeper resolution on the differences provide the histograms of the length distributions for the middle introns in each of the 4 groups (Fig. 9). Group 1 shows >50% of events (i.e., in 91 events) with short 2nd introns (<1,500 nt, which is about the median length of introns in coding human transcripts) and ~1/3 (59 introns) very short introns (<500 nt). Also group 2 exhibits several (124) events with short and some (67) very short middle introns, still about twice as many as in group 3 and 4 (66 short and 33 very short introns). These observations support the hypothesis that events in group 1 and 2 (in contrast to events of group 3 and 4) may involve common molecular mechanisms to ensure mutual exclusion of the exons.

One could hypothesize that the common property of the former events could stem from a common molecular mechanism as described by steric blocking effects in the splicing process of mutually exclusive exons (Smith and Nadal-Ginard, 1989). However, additional analyzes have to be conducted to support this hypothesis as literature describes a wide spectrum of mechanisms that can lead to mutual exclusion of exons, e.g., the relative strengths of 5′ and 3′-splice sites (Kuo et al., 1991; Mullen et al., 1991; Zhuang

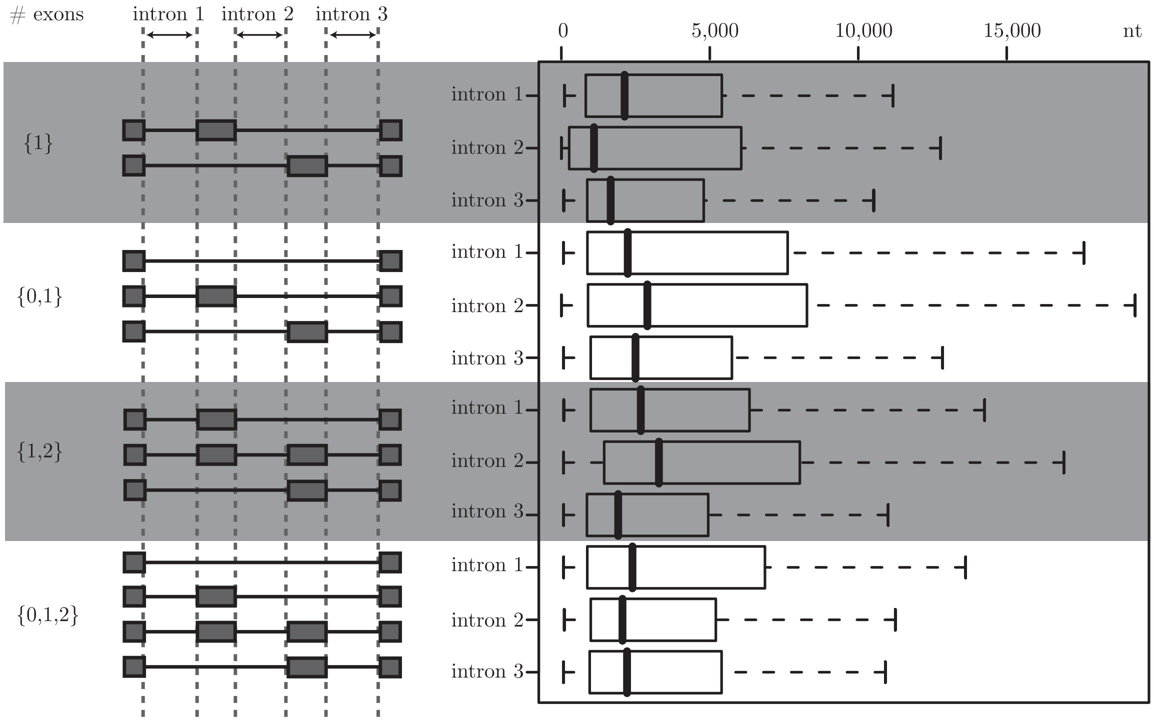


FIG. 8. Distribution of intron lengths occurring in AS events with two neighboring exons that are alternatively included in coding regions of the human genome: mutually exclusive exons (exactly one exon included), events that include one of the two exons or none (number of included exons = {0,1}), events that include one of the two exons or both, and events that include one of the two exons, none or both. Distributions are shown separately for the intron upstream of the first (i.e., the most 5′) exon ("intron 1"), the intron between the two optional exons ("intron 2") and the intron downstream of the second exon ("intron 3"). The figure has been generated involving the boxplot() function of the R package (R Development Core Team, 2007) where the boxes represent the length range for the 2nd and 3rd quartile of the distribution, separated by the median (bold vertical line). The 1st respectively the 4th quartile are shown as dashed lines.

et al., 1987), the pyrimidine content of a 3′-splice site (Fu et al., 1988; Mullen et al., 1991), the location and number of branchpoints (Gattoni et al., 1988; Goux-Pelletan et al., 1990; Helfman and Ricci, 1989; Helfman et al., 1990; Noble et al., 1987, 1988; Smith and Nadal-Ginard, 1989), branchpoint sequences (Mullen et al., 1991; Reed and Maniatis, 1988; Zhuang et al., 1989), intron sequences between branchpoint and 3′-splice site (Goux-Pelletan et al., 1990; Helfman et al., 1990; Libri et al., 1990), exon sequences (Black, 1991; Cooper and Ordahl, 1989; Hampson et al., 1989; Helfman et al., 1988; Libri et al., 1991; Mardon et al., 1987; Reed and Maniatis, 1986; Somasekhar and Mertz, 1985; Streuli and Saito, 1989) and *trans*-acting factors such as ASF or SF2 (Ge and Manley, 1990; Krainer et al., 1990).

5.5. The pattern of two consecutively skipped exons is determined by the strength of splice sites delimiting the exons and by the length of the intron between them

From the four largest groups of complete events I selected those that exhibit a variant skipping two subsequent exons: events that exclusively show the inclusion or the exclusion of both exons (Fig. 10A), events that additionally show the inclusion of exclusively the first (i.e., the upstream) exon (Fig. 10B), events that include only the second exon or both or none (Fig. 10C), and finally events that include both, either one or none of the two exons (Fig. 10D).

In each pattern, the distribution of lengths for both exons and each of the 6 (possible) introns has been visualized in a boxplot (Fig. 10, center column). The median of intron 1–3 varies in all patterns around 3kB, except for intron 2 of the pattern in Figure 10A, where the second intron is significantly shorter (median 973nt) than intron 1 and 3. Strikingly, reports on new therapies for muscular dystrophy show supporting evidence that one possible mechanism of double exon skipping is related to splicing kinetics and therefore controlled mainly by the distribution of intron lengths (Aartsma-Rus and van Ommen, 2007): these cases

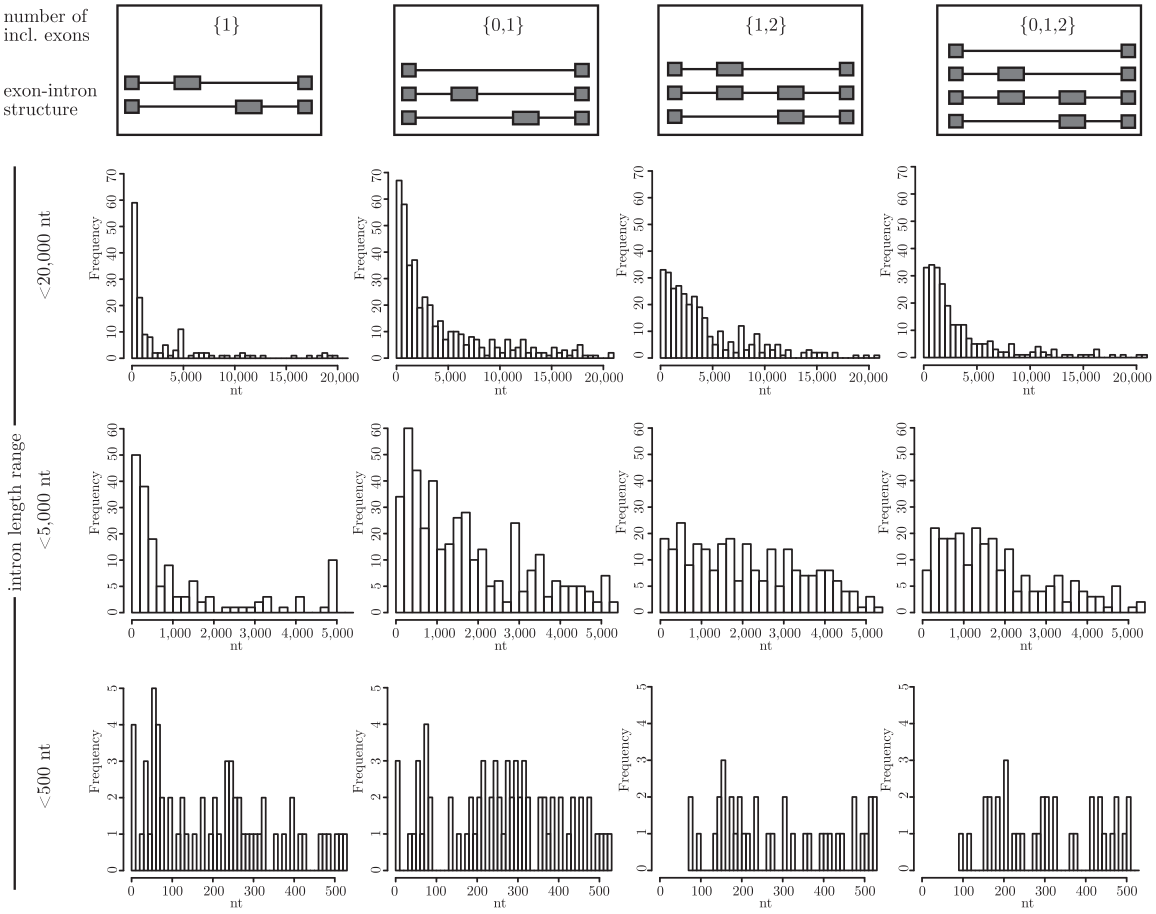


FIG. 9. Distribution of lengths exhibited by the 2nd intron of events that involve alternative neighboring exons, of which either strictly one is included (mutually exclusive exons), or one of them or none (number of included exons = {0,1}), one or both are included (number of included exons = {1,2}), or, none, one or both are included (number of included exons = {0,1,2}). Mutually exclusive exons show a substantial population of small and some very small middle introns that is partially present in events where none or one exon is included, but missing in the other variations.

report on a "relatively slow" splicing process of intron—1 due to its large size—in combination with the faster transcription of exon 1 and intron 2, which subsequently gets spliced out quickly. If intron 2 has been spliced first, intron 4 and 5 can no longer be spliced in the transcript and—either thermodynamically or by some regulation mechanism—intron 1 and 3 compete with intron 6 for determining the final splice structure.

Next (Fig. 10, right column), I compared the strengths of splice sites that delimit the exons and introns in each pattern. To this end, I measured the agreement of the splice site sequences with the human consensus by the the splice site models as implemented in the program geneID (Parra et al., 2000), including a separated weighting scheme for introns spliced by the U12 spliceosome (Alioto, 2007). The scoring is a log measure with the majority of sites yielding a score between (-2) and ($+2$). The analysis shows a clear asymmetry in the correlation of strengths between the splice sites delimiting the exons in pattern B and C: in Figure 10B the splice sites of the second exon are significantly weaker (median 0.89 *vs.* -0.42, e^{-15} unparametrical 2-sided Kolmogorov-Smirnov test) than the splice sites of the first one, whereas in Figure 10C an opposite correlation (median -0.38 *vs.* 0.71, $<e^{-13}$) is observed. Lastly, the pattern in Figure 10D exhibits no preferential exon or intron with weaker or stronger splice sites.

6. DISCUSSION

This work extends the defintion for AS events to the scope of $d \geq 2$ variants and describes an algorithm that exhaustively retrieves such events from large datasets, combining full-length cDNAs with highly

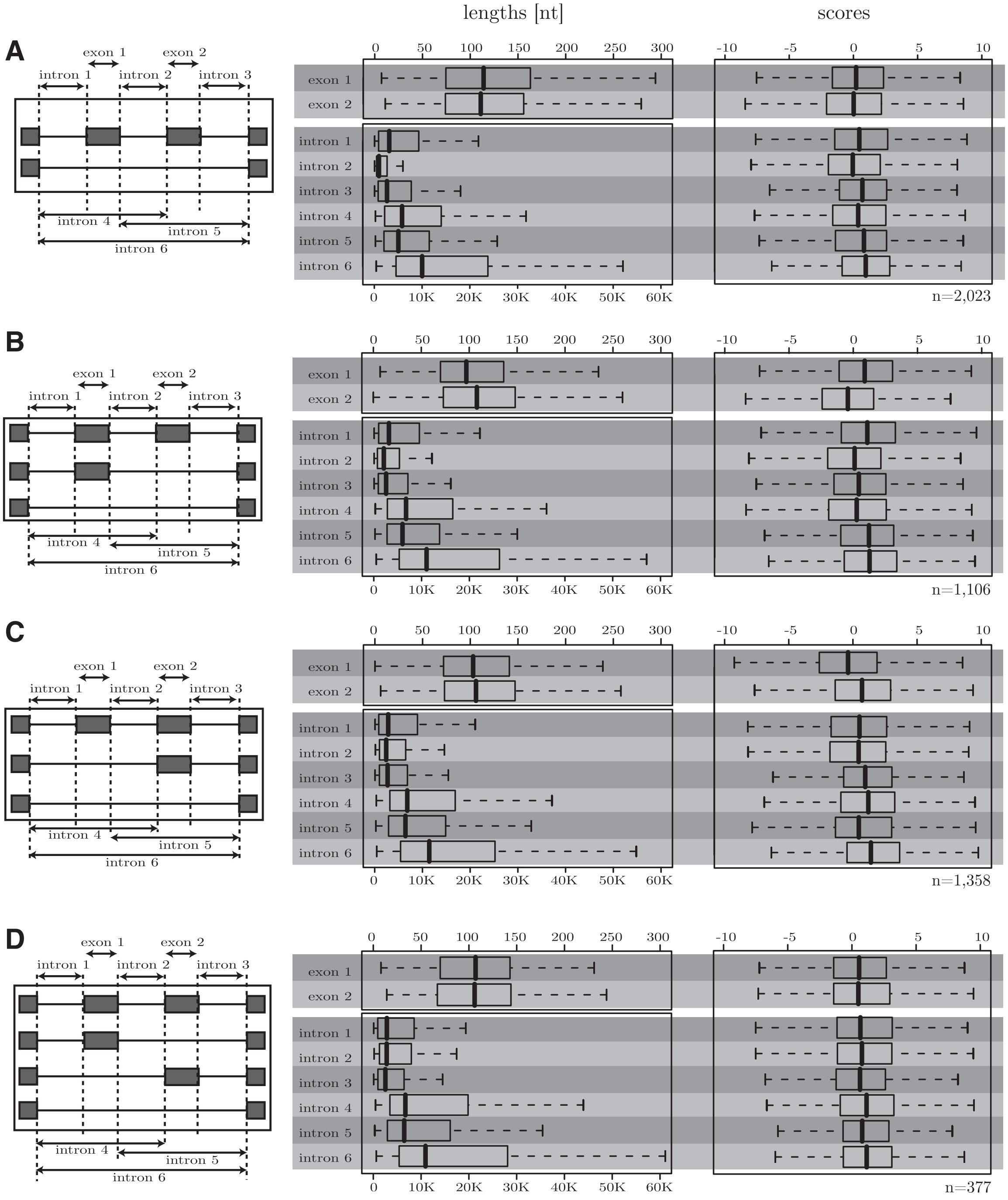


FIG. 10. Structurally different events including variants that support the skipping of two consecutive exons. (**A**) Complete events with $d = 2$ where both exons are either included, or skipped. (**B**) Complete events that show the inclusion of none, both exons, or only the more upstream one of them. (**C**) Complete events that describe splicing structure where none or both exons, or the only the more downstream exon is included. (**D**) Complete events that show variants including either none, both or either one of the variable exons.

fragmented ESTs. Due to an intrinsic clustering of transcripts into loci, no preliminary EST-clustering is required. Furthermore, by disregarding introns that appear to stem from technical artifacts in the transcript annotation, variations in the exon-intron structure are reduced to presumptive AS events. The method—comprising the steps of parsing the file with the annotated transcripts, clustering them into loci, retrieving the splice site sequence for all introns and extracting the events—copes with the biggest dataset of ∼4 million spliced transcripts in human in about 2½ hours and retrieves more than 90,000 *bona fide* AS events. A retrospective about the availability of AS events over the last 20 years suggests that—even in the rather well-established human annotation—we have still not reached completion and future annotations are expected to lead to the discovery of novel events. With the era of new sequencing technologies having just begun, ESTs can be produced rapidly (about a week from tissue harvesting to completion of DNA sequencing) and at very low costs (currently <$0.03 per EST using pyrosequencing), emphasizing the potential of the here-in presented method in high-throughput systems to explore rapidly splicing variations reflected by the transcriptome of different organisms, tissues, cells or cell states.

In human, the described method shows a plethora (>24,000) of AS events that involve actually more than two alternatives and that have hitherto not been described in their real complexity. Consequently ~¼ of the splicing variation has been ignored or miscategorized in previous works, for instance a study about patterns shows that only the minority of neighboring exons that are alternatively included are what fits the common understanding of "mutually exclusive exons." Each of the most frequent of those patterns exhibits unique characteristics in the length distribution of the introns located between the optional exons: many events that include exclusively one exon comprise (very) short such middle introns in contrast to events that alternatively include both exons. Also patterns that show the skipping of 2 consecutive exons identify markedly differing attributes that may be related to splicing kinetics (i.e., intron lengths and splice site strengths). Supported by case studies which describe molecular mechanisms with similar characteristics, hypotheses on the mechanism can be formulated, and thus systematical characterization of complete events can improve our current understanding of (alternative) splicing.

ACKNOWLEDGMENTS

I thank the whole genome bioinformatics laboratory group of the CRG, especially Sylvain Foissac, Sarah Djebali, Christoforos Nikolau, Hagen Tilgner and Julien Lagarde, for many discussions that contributed to this work; Tyler Alioto, for elucidating the intron-based perspective of splicing; and Oscar Gonzalez and Judith Flo, for assistance in hard- and software problems. Furthermore, I thank Gabriel Valiente and Roderic Guigó, co-authors of the conference version of this manuscript, for fruitful discussions that led to a significant improvement of the manuscript. During this work, I was funded by a postdoctoral fellowship by the German Academic Exchange Service (DAAD) and subsequently from the Spanish Ministry of Science (MEC).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Aartsma-Rus, A., and van Ommen, G.J. 2007. Antisense-mediated exon skipping: a versatile tool with therapeutic and research applications. *RNA* 13, 1609–1624.

Alioto, T. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* 35, D110–D115.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., et al. 2007. GenBank. *Nucleic Acids Res.* 35, D21–D25.

Black, D.L. 1991. Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes Dev.* 5, 389–402.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. 1993. dbEST—database for "expressed sequence tags." *Nat. Genet.* 4, 332–333.

Carninci, P., Kasukawa, T., Katayama, S., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Cooper, T.A., and Ordahl, C.P. 1989. Nucleotide substitutions within the cardiac troponin T alternative exon disrupt pre-mRNA alternative splicing. *Nucleic Acids Res.* 17, 7905–7921.

Consortium, T.E. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Foissac, S., and Sammeth, M. 2007. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* 35, W297–W299.

Fu, X.Y., Ge, H., and Manley, J.L. 1988. In vitro splicing of mutually exclusive exons from the chicken β-tropomyosin gene: role of the branch point location and very long pyrimidine stretch. *EMBO J.* 7, 809–817.

Gattoni, R., Schmitt, P., and Stevenin, J. 1988. In vitro splicing of adenovirus E1A transcripts: characterization of novel reactions and of multiple branch points abnormally far from the 3′ splice site. *Nucleic Acids Res.* 16, 2389–2409.

Ge, H., and Manley, J.L. 1990. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell* 62, 25–34.

Goux-Pelletan, M., Libri, D., d'Aubenton-Carafa, Y., et al. 1990. In vitro splicing of mutually exclusive exons from the chicken β-tropomyosin gene: role of the branch point location and very long pyrimidine stretch. *EMBO J.* 9, 241–249.

Grasso, C., Modrek, B., Xing, Y., et al. 2004. Genome-wide detection of alternative splicing in expressed sequences using partial order multiple sequence alignment graphs. *Pac. Symp. Biocomput.* 29–41.

Gusfield, D., and Bansal, V. 2005. A fundamental decomposition theorem for phylogenetic networks and incompatible characters. *Lect. Notes Bioinform.* 3500, 217–232.

Gusfield, D., Eddhu, S., and Langley, C. 2004. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinform. Comput. Biol.* 2, 173–213.

Hampson, R.K., La Follette, L., and Rottman, F.M. 1989. Alternative processing of bovine growth hormone mRNA is influenced by downstream exon sequences. *Mol. Cell Biol.* 9, 1604–1610.

Heber, S., Alekseyev, M., Sing-Hoi, S., and Pevzner, P. 2002. Splicing graphs and EST assembly problem. *Bioinformatics.* 18, 181–188.

Heber, S., Alekseyev, M., Sze, S.H., et al. 2002. Splicing graphs and EST assembly problem. *Bioinformatics* 18, S181–S188.

Helfman, D.M., and Ricci, W.M. 1989. Branch point selection in alternative splicing of tropomyosin pre-mRNAs. *Nucleic Acids Res.* 17, 5633–5650.

Helfman, D.M., Ricci, W.M., and Finn, L.A. 1988. Alternative splicing of tropomyosin pre-mRNAs in vitro and in vivo. *Genes Dev.* 2, 1627–1638.

Helfman, D.M., Roscigno, R.F., Mulligan, G.J., et al. 1990. Identification of two distinct intron elements involved in alternative splicing of beta-tropomyosin pre-mRNA. *Genes Dev.* 4, 98–110.

Human Genome Sequencing Consortium. 2009. Available at: http://genome.ucsc.edu/goldenPath/labs.html. Accessed July 1, 2009.

Kent, W.J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664.

Kim, E., Magen, A., and Ast, G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* 35, 125–131.

Krainer, A.R., Conway, G.C., and Kozak, D. 1990. The essential pre-mRNA splicing factor SF2 influences 5′ splice site selection by activating proximal sites. *Cell* 13, 35–42.

Kuo, H.C., Nasim, F.H., and Grabowski, P.J. 1991. Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. *Science* 251, 1045–1050.

Kuyumcu-Martinez, N.M., and Cooper, T.A. 2006. Misregulation of alternative splicing causes pathogenesis in myotonic dystrophy. *Prog. Mol. Subcell. Biol.* 44, 133–159.

Lander E.S., Linton, L.M., Birren, B., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Libri, D., Goux-Pelletan, M., Brody, E., et al. 1990. Exon as well as intron sequences are cis-regulating elements for the mutually exclusive alternative splicing of the beta tropomyosin gene. *Mol. Cell Biol.* 10, 5036–5046.

Libri, D., Piseri, A., and Fiszman, M.Y. 1991. Tissue-specific splicing in vitro of the beta-tropomyosin gene: dependence on an RNA secondary structure. *Science* 252, 1842–1845.

Lopez, A.J. 1998. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* 32, 279–305.

Mardon, H.J., Sebastio, G., and Baralle, F.E. 1987. A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucleic Acids Res.* 15, 7725–7733.

Mouse Genome Sequencing Consortium. 2009. Available at: www.ensembl.org/Mus_musculus/credits.html. Accessed July 1, 2009.

Mullen, M.P., Smith, C.W., Patton, J.G., et al. 1991. Alpha-tropomyosin mutually exclusive exon selection: competition between branchpoint/polypyrimidine tracts determines default exon choice. *Genes Dev.* 5, 642–655.

Nagasaki, H., Arita, M., Nishizawa, T., et al. 2005. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene* 364, 53–62.

Noble, J.C., Pan, Z.Q., Prives, C., et al. 1987. Splicing of SV40 early pre-mRNA to large T and small t mRNAs utilizes different patterns of lariat branch sites. *Cell* 50, 227–236.

Noble, J.C., Prives, C., and Manley, J.L. 1988. Alternative splicing of SV40 early pre-mRNA is determined by branch site selection. *Genes Dev.* 2, 1460–1475.

Parra, G., Blanco, E., and Guigó, R. 2000. GeneID in Drosophila. *Genome Res.* 10, 511–515.

Pevzner, P., Tang, H., and Tesler, G. 2004. De novo repeat classification and fragment assembly. *Genome Res.* 14, 1786–1796.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: www.R-project.org. Accessed July 1, 2009.

Reed, R., and Maniatis, T. 1986. A role for exon sequences and splice-site proximity in splice-site selection. *Cell* 46, 681–690.

Reed, R., and Maniatis, T. 1988. The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev.* 2, 1268–1276.

Ruan, Y., Ooi, H.S., Choo, S.W., et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* 17, 828–838.

Sammeth, M., Foissac, S., and Guigó, R. 2008. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.* 4, e1000147.

Smith, C.W., and Nadal-Ginard, B. 1989. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell* 56, 749–758.

Smith, C.W., and Valcárcel, J. 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* 25, 381–388.

Somasekhar, M.B., and Mertz, J.E. 1985. Exon mutations that affect the choice of splice sites used in processing the SV40 late transcripts. *Nucleic Acids Res.* 13, 5591–5609.

Streuli, M., and Saito, H. 1989. Regulation of tissue-specific alternative splicing: exon-specific cis-elements govern the splicing of leukocyte common antigen pre-mRNA. *EMBO J.* 8, 787–796.

Sugnet, C.W., Kent, W.J., Ares, M., et al. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* 66–77. Available at: www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd = Retrieve\&db = pubmed\&%dopt = Abstract\&list_uids = 14992493. Accessed July 1, 2009.

UCSC (University of California Santa Cruz) Genome Browser. 2009. Available at: http://genome.ucsc.edu. Accessed July 1, 2009.

Weber, A.P., Weber, K.L., Carr, K., et al. 2007. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol.* 144, 32–42.

Yandell, M., Mungall, C.J., Smith, C., et al. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput. Biol.* 2, e15.

Zavolan, M., and van Nimwegen, E. 2006. The types and prevalence of alternative splice forms. *Curr. Opin. Struct. Biol.* 16, 362–367.

Zhuang, Y., Leung, H., and Weiner, A.M. 1987. The natural 5′ splice site of simian virus 40 large T antigen can be improved by increasing the base complementarity to U1 RNA. *Mol. Cell Biol.* 7, 3018–3020.

Zhuang, Y.A., Goldstein, A.M., and Weiner, A.M. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proc. Natl. Acad. Sci. USA* 86, 2752–2756.

Address correspondence to:
Dr. Michael Sammeth
Bioinformatics and Genomics
Centre for Genomic Regulation (CRG)
Dr. Aiguader 88
08003 Barcelona, Catalunya, Spain

E-mail: micha@sammeth.net