Differential quantification of alternative splicing events on spliced pangenome graphs

Simone Ciccolella[1,*], Davide Cozzi[1,*], Gianluca Della Vedova[1], Stephen Njuguna Kuria[2], Paola Bonizzoni[1,†], and Luca Denti[1,†]

[1] Department of Computer Science, University of Milan-Bicocca, Milan, Italy
[2] Department of Biochemistry and Biotechnology, Pwani University, Kilifi, Kenya
[*] Co-first authors. [†] Co-last authors.
✉ luca.denti@unimib.it

Abstract. Pangenomes are becoming a powerful frameworks to perform many bioinformatics analyses taking into account the genetic variability of a population, thus reducing the bias introduced by a single reference genome. With the wider diffusion of pangenomes, integrating genetic variability with transcriptome diversity is becoming a natural extension that demands specific methods for its exploration. In this work, we extend the notion of spliced pangenomes to that of *annotated spliced pangenomes*; this allows us to introduce a formal definition of Alternative Splicing (AS) events on a graph structure.
To investigate the usage of graph pangenomes for the quantification of AS events across conditions, we developed `pantas`, the first pangenomic method for differential analysis of AS events. A comparison with state-of-the-art linear reference-based approaches proves that `pantas` achieves competitive accuracy, making spliced pangenomes effective for conducting AS events quantification and opening future directions for the analysis of population-based transcriptomes. `pantas` is open-source and freely available at `github.com/algolab/pantas`.

Keywords: Alternative Splicing · Pangenomics · Sequence Analysis.

1 Introduction

Pangenomics is emerging as a new powerful computational framework for analyzing the genetic variability of a population without being negatively affected by the reference bias introduced when considering a single genome as a reference. The recent release of the first draft human pangenome reference [23] demonstrated that this richer reference can be used to perform many bioinformatics analyses, e.g., variant calling, with superior accuracy and precision, especially in structurally complex loci of the genome. In particular, pangenomes have also been showed to improve transcriptomic data analysis [32], thanks to the extension of the notion of pangenomes to that of spliced pangenomes (or pantranscriptomes). A spliced pangenome is obtained by enriching a pangenome, that is a graph structure representing the genetic diversity of multiple individuals of a population, with transcript variability coming from gene annotations. In [32], this complex but powerful graph structure is used to perform hapolotype-aware transcript quantification. A comparison with reference-based approaches proves that spliced pangenomes improve the accuracy of transcript quantification. Indeed, by incorporating both genetic variability and isoform diversity, spliced pangenomes allow to reduce the allelic bias and increase the number of reads aligned over heterozygous variations [32]. Apart from this seminal work, the usage of spliced pangenomes in the context of transcriptomic is fairly unexplored and their full potential is still under investigation. Similarly to pangenomics, pantranscriptomics needs suitable bioinformatics tools to fulfill its biological potential. Encouraged by the results in [32], we explore the adoption of spliced pangenomes for performing a classical transcriptomic task, namely the differential quantification of Alternative Splicing (AS) events across conditions.

Alternative splicing (AS) is a regulation mechanism which allows a single gene to code for multiple isoforms and hence express multiple proteins. Such a mechanism is the main contributing factor for the overwhelming complexity of transcriptomes in eukaryotes. AS is associated with different diseases [8,34,9] and it is also now clear that it plays a key role in various biological processes, like aging [7]. The advent of RNA-Sequencing technology (RNA-Seq) enabled the analysis of transcriptome and alternative splicing at an unprecedented speed and precision. A typical RNA-Seq analysis consists of comparing two conditions (e.g., control vs tumor) and checking for changes in terms of isoform abundances [36,19]: a change in the relative abundances of the isoforms of a gene is usually a strong evidence of differential splicing. An alternative approach consists in directly detecting and quantifying alternative splicing events. Instead of focusing on entire isoforms quantification, several approaches work at a finer-grained level and focus on exon-exon boundaries, also known as splice junctions, and check for changes in their usage. By analyzing splice junctions, these approaches can directly detect and quantify AS events providing more accurate results with respect to approaches based on transcript quantification [17]. Several tools have been proposed in the literature to differentially quantify AS events from RNA-Seq datasets [31,20,15,13,37,38,21,33].

Differently from state-of-the-art, where a single reference is considered and AS event quantification is modeled without taking into account the genetic variability of the population of the species under investigation, we propose to use a spliced pangenome and quantify AS events using this more powerful structure that inherently represents multiple individuals and gene annotations. To this aim, we introduce the notion of *annotated spliced pangenome*, we propose the first formalization of AS events on this richer structure, and we provide the first method, `pantas`, to perform AS events differential quantification on a spliced pangenome. Although the adoption of a graph structure is not new in the computational transcriptomics, where *splicing graphs* are commonly used to model isoform variability [4,5,29,20,15,13], no approach takes also into account the genetic variability of the population under investigation. It is indeed well known that genetic variations alter splicing pattern in many diseases [24,22]. However, establishing their impact on alternative splicing is still challenging [11]. Even though the adoption of spliced pangenomes may shed more light on the relationship between genetic variations and splicing, the current lack of efficient tools for pangenomic-based analysis of RNA-Seq data is hindering this kind of investigation. Our work here is a first step in closing this gap.

Experimental evaluation on simulated and real data from Drosophila and Human shows that spliced pangenomes can be effectively adopted to perform AS events quantification. A comparison with state-

of-the-art tools based on linear reference or splicing graph shows that AS event quantification from spliced pangenome provides competitive results. Our preliminary investigation paves the way to the adoption of spliced pangenomes for pantranscriptomic analysis, where genes are annotated w.r.t. a pangenome and not a single reference genome [1].

2 Methods

In this section, we first introduce the notion of *annotated spliced pangenome* and then we describe the approach based on this structure that we developed for detecting and quantifying AS events across RNA-Seq conditions.

2.1 Annotated spliced pangenome graph

A *variation graph*, as defined in [2], is a directed graph whose vertices correspond to portions of genomes and are labeled by nonempty strings, and edges represent consecutiveness between two such portions of genomes. A *spliced pangenome*, introduced in [32], is a variation graph with some distinguished walks: the walk R corresponding to the reference genome, the walks $\{H_1, H_2, \ldots, H_h\}$ representing the h input haplotypes, and the walks $\{T_1, T_2, \ldots, T_t\}$ representing t transcripts coming from a gene annotation. The arcs that belong to some transcript walk T_i, but do not belong to R, correspond to annotated *splice junctions*. These edges link two vertices of the graph that are not consecutive in the reference walk (due to the presence of intronic regions between the exons). Notice that, in a spliced pangenome, a vertex v that also belongs to at least a transcript walk T_i represents an exonic region of the reference genome. All other vertices represent intronic regions, intergenic regions, or alternative alleles. In this work, we will focus on a simpler version of this graph where no input haplotype is explicitly stored as a distinguished walk (*i.e.*, $h = 0$). Although this is a simplification of the original notion of spliced pangenome, it is sufficient for AS events detection and quantification since, as we will show in the following sections, AS events can be modeled using only transcript information.

An *annotated spliced pangenome* is a spliced pangenome whose vertices and arcs are annotated (or labeled) with some additional information that will be essential for formalizing and detecting AS events. An example of *annotated spliced pangenome* is given in Fig. 1. More comprehensive examples are given in Suppl. Fig. S1 and S2. More precisely, each vertex v of an annotated spliced pangenome is annotated with two information: the set $\mathcal{E}(v)$ of its exons and the set $\mathcal{T}(v)$ of its transcripts. We note that the a vertex can represent a single exonic region on the reference genome, but this exonic region may belong to multiple exons coming from different transcripts, e.g., due to an alternative form of an exon. For this reason, $\mathcal{E}(v)$ is a set of exons and not a single exon. To simplify the exposition, whenever $\mathcal{E}(v)$ or $\mathcal{T}(v)$ are empty, we say that the vertex v has no annotation — this can happen, e.g., for a vertex in an intergenic region. Moreover, each edge $e = \langle u, v \rangle$ of an annotated spliced pangenome that represents a splice junction is annotated with the set $\mathcal{T}(e)$ of the transcripts it belongs to. As for vertices, the same splice junction can be shared by many transcripts.

2.2 AS events quantification from annotated spliced pangenomes

Given an annotated spliced pangenome, the problem we want to tackle is detecting and quantifying the AS events supported by an input RNA-Seq dataset, comprising two conditions with optional replicates. To this aim, we consider as input of our problem the annotated spliced pangenome and the spliced alignments of the input samples computed against it. The expected output is the set of AS events (exon skipping, alternative acceptor, alternative donor, and intron retention) supported by the alignments. To solve this problem, we designed and developed `pantas`.

`pantas` starts by weighting all edges of the input graph based on how many times they are used by the input alignments of each replicate. This is necessary to compute the support of the AS events and

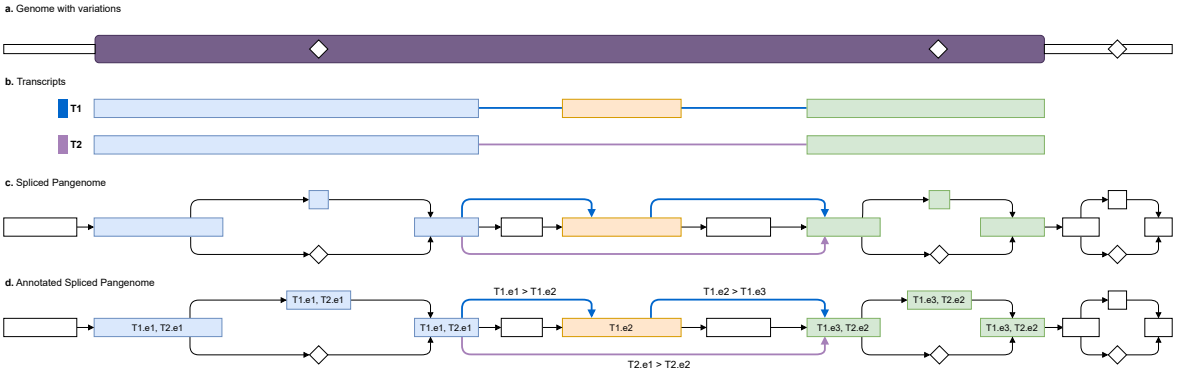

Fig. 1: Annotated spliced pangenome example. (a) Reference genome with a purple box representing a gene locus and white diamonds representing alternate alleles of variations. (b) The gene has two transcripts, namely T1 and T2, expressing an exon skipping event. (c) Spliced pangenome built from the reference genome, gene annation, and a set of variations. Colored vertices and edges belongs to the transcript walks and they represent portions of exons and splice junctions. White vertices are alternative alleles and introns portion that do not belong to the transcript walks. (d) Annotated spliced pangenome where each colored vertex (exon portion) is annotated by the transcripts and exons it belongs to and each colored edge (junction) is annotated with the junction information.

successively quantify them. Moreover, `pantas` augments the original graph with novel splice junctions (*i.e.*, new edges) supported by the alignments and not originally present in the annotated spliced pangenome. Since a novel splice junction can start or end inside a vertex (*e.g.*, due to an alternative 5' event), we need to precisely locate the novel AS events corresponding to that junction and store such locations — an alternative way to view this novel splice junction is to split a vertex into two. Then, by analyzing the annotated spliced pangenomes that has been augmented with read alignment information, `pantas` proceeds to detect the AS events.

From a high level point of view, AS events can be identified by locally comparing the splice junctions of two transcripts represented in the annotated spliced pangenome. An AS event is said to be *annotated* if the splice junctions of the two transcripts involved in the event are already annotated, *i.e.*, the edges representing the junctions are already present and annotated in the input graph. On the other hand, an event is said to be *novel* if the splice junctions of only one of the two transcripts is already annotated and the other junction is supported by read alignments only. We refer the reader to Section 2.3 for the formal description of how AS events can be detected from an annotated spliced pangenome.

After detecting the AS events from each replicate, `pantas` quantifies the events by combining the results obtained from each replicate. `pantas` represents each AS event as a pair of sets of edges, representing the two junctions sets (the first one from the constitutive isoform and the second one from the alternative isoform) involved in the event. By analyzing the support of the junctions in each replicate, `pantas` computes the Percent Spliced-In (ψ) of the event by computing the ratio between the support of the constitutive isoform and the alternative isoform [30]. Once each replicate has been analyzed, `pantas` performs differential quantification of the events supported by both conditions by computing the $\Delta\psi$ of each event as the difference between the absolute value of the ψ means in the two conditions.

Additionally, to simplify the downstream analysis of its results, `pantas` also surjects the positions of the edges involved in the events back to the reference genome. This is simply done by mapping the positions of the vertices linked by each edge from the graph space to the reference genome. We note that novel AS events need more care, since novel AS events usually involve novel splice sites and the novel splice junctions induced by the alignment cut a vertex of the graph in two parts. However, thanks

to the information stored when augmenting the annotated spliced pangenome graph with alignment information, this boils down to simply moving the splice site of some bases, depending on where the vertex was cut.

2.3 AS events detection

In this section we will provide the formal description of how AS events can be detected from an annotated spliced pangenome. We will first focus on annotated AS events and then we will formalize novel AS events. Differently from other approaches that model AS events as coordinates over a linear reference genome, we model the AS events directly on a graph structure and we provide a formal definition of each event in terms of graph elements (i.e., vertices and edges). The annotation introduced as distinctive property of annotated spliced pangenomes helps in this formalization.

Annotated AS events criteria To detect annotated alternative splicing events we consider every annotated junction independently as a potential event for which we check whether the following conditions apply. The definitions below are all referred to a splice junction $\langle a, b\rangle$ between two exons of a transcript $T_1 \in \mathcal{T}(a) \cap \mathcal{T}(b)$, and the junction $\langle a, b\rangle$ is covered by at least a user-specified number of reads (*i.e.*, it is supported in the sample). In the following, given a vertex v, we denote with $next(v)$ and $prev(v)$ the set of successors and predecessors of v, respectively.

Exon skipping We call an annotated exon skipping for the annotated splice junction $j = \langle a, b\rangle$ when there is a transcript T_2 (of the same gene) with an exon between the two vertices a and b — notice that $T_1 \neq T_2$ since the transcript T_1 uses the junction $\langle a, b\rangle$, therefore it cannot have any exon between a and b. Formally, there exists a transcript $T_2 \in \mathcal{T}(a) \cap \mathcal{T}(b) \setminus \mathcal{T}(j)$ with two exons $e_a \in \mathcal{E}(a)$, $e_b \in \mathcal{E}(b)$ that are not consecutive in t. Fig. 2.a shows an example in which $a = n_1$, $b = n_4$ and *T1.e2* is the skipped exon for the transcript *T2*.

Alternative 5' splicing site We call an annotated alternative 5' splicing site for the annotated junction $j = \langle a, b\rangle$ when there is another transcript (of the same gene) for which the first exon extends further towards the 3', but does not incorporate the second exon of the junction j. Formally, there exists a vertex $v \in next(a)$ and an exon $e \in \mathcal{E}(a) \cap \mathcal{E}(v) \setminus \mathcal{E}(b)$ (*i.e.*, the exon e covers both the vertex a and one of its successors v, but does not cover b) such that $\mathcal{T}(e) \cap \mathcal{T}(b) \neq \emptyset$. By construction, this implies that there is a transcript $T_2 \in \mathcal{T}(e)$, $T_1 \neq T_2$, that does not cover the junction j, but cover both exons joined by j. Fig. 2.b shows an example in which $a = n_1$, $b = n_3$ and there are vertices $(n_1, ..., n_2]$ would be extending exonic vertices of *T1.e1* but not of *T2.e1*.

Alternative 3' splicing site The definition of an annotated alternative 3' splicing site is symmetrical to that of 5' splicing sites. It suffices to identify a precedecessor $v \in prev(b)$ and an exon $e \in \mathcal{E}(b) \cap \mathcal{E}(v) \setminus \mathcal{E}(a)$ such that $\mathcal{T}(e) \cap \mathcal{T}(b) \neq \emptyset$. Fig. 2.c shows an example in which $a = n_1$, $b = n_3$ and $[n_2, ..., n_3)$ would be extending exonic vertices of *T1.e2* but not of *T2.e2*.

Intron retention We call an annotated intron retention for the annotated junction $j = \langle a, b\rangle$ when there is another transcript (of the same gene) for which a, b are in the same exon, i.e., there is a third exon spanning both exon a and b and including the intron in between them. Formally, there exist two edges $\langle a, u\rangle$ and $\langle v, b\rangle$ such that $\mathcal{E}(a) \cap \mathcal{E}(b) \cap \mathcal{E}(u) \cap \mathcal{E}(v) \neq \emptyset$. Notice that u and v are not necessarily distinct. Fig. 2.d shows an example in which $a = n_1$, $b = n_4$ which are part of exons *e1* and *e2* for transcript *T1* and of *e1* for *T2*.

Novel AS events criteria To detect novel alternative splicing events, we need to consider both annotated and novel splice junction edges (depending on the event type) since in this setting an AS

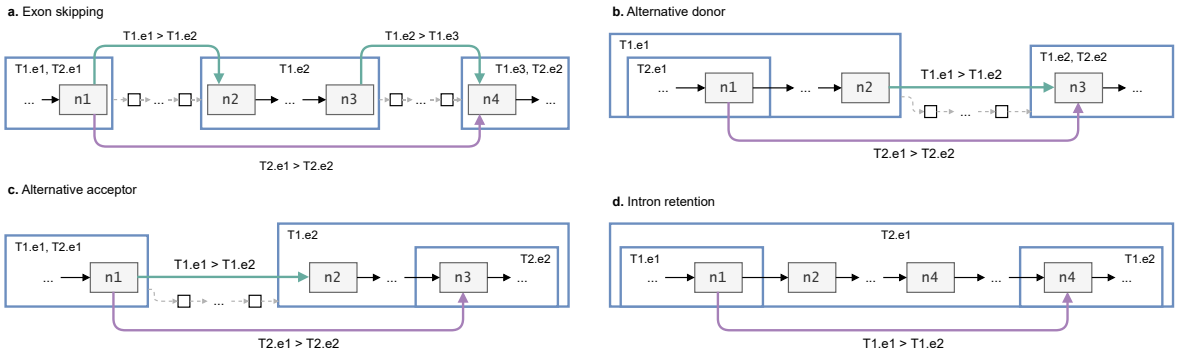

Fig. 2: All the annotated events expressed within the annotated spliced pangenome showing the different tags used, weights are omitted for readability; blue squares represent exons with their tags, green and purple edges are the annotated junctions with their tags, grey vertices are exonic and white are intronic. Exons are labeled with the transcript walks that are represented in the figure.

event occurs between the annotated junctions of a transcript and the novel junctions of another transcript. Therefore, for each AS event, we have to take into account two different scenarios, depending on which junctions are annotated and which are novel. For instance, if we have a novel exon skipping, it means that the novel exon junction is skipping an exon and the two junctions supporting the inclusion of the exon are already annotated. On the other hand, if the two inclusion junctions are novel (and the skipping junction is annotated), we have a cassette exon event. Therefore, we have to consider 8 different scenarios (2 per event). For ease of exposition, we decided to leave the formal definition of novel events as supplementary material (Suppl. Section S1 and Suppl. Fig. S3).

3 Experimental Evaluation

To evaluate `pantas` efficacy and correctness in quantifying AS events across conditions, we performed 3 experimental evaluations on simulated and real data. In the first experiment, we used simulated data to evaluate the correctness of `pantas` in detecting both annotated and novel AS events. In the second experimental evaluation, we considered a real RNA-Seq dataset from Drosophila Melanogaster and we evaluated the accuracy of `pantas` in differential quantifying annotated AS events. Finally, we evaluated the accuracy of `pantas` in quantifying RT-PCR validated AS events from a real human RNA-Seq dataset.

In all the considered scenarios, we constructed the annotated spliced pangenomes using a custom procedure based on the `vg` toolkit [18] and we computed the spliced alignments using the `mpmap` aligner [32]. This aligner was specifically introduced to align RNA-Seq to spliced pangenomes and to this aim it can align over novel splice junctions, that are edges not present in the input graph. This feature is essential for inferring novel AS events. More details on the input preparation can be found in Suppl. Section S2.

To put our results in perspective, we compared `pantas` with three state-of-the-art approaches, namely `rMATS` [31], `SUPPA2` [37], and `whippet` [33]. The former quantifies AS events by analyzing spliced alignments to a reference genome (e.g., computed with `STAR` [16]), the second by analyzing transcript quantification computed with `salmon` [28], and the latter by analyzing alignments to a custom graph, called Contiguous Splice Graph, which represents all non-overlapping exons coming from the gene annotation.

3.1 pantas event detection is correct and accurate

The goal of the first experimental evaluation was to assess the accuracy of pantas in detecting annotated and novel AS events. To do so, we simulated a RNA-Seq datasets using asimulator [26], a tool that simulates RNA-Seq datasets while introducing AS events. We considered the Drosophila Melanogaster reference and gene annotation (FlyBase r6.51 [35]) and the Drosophila Genetic Reference Panel (v2) [25]. To simulate a real-case scenario, where the sequenced sample is typically not present in the reference panel, we randomly selected a sample from the panel, we simulated reads from it, and then removed the considered sample from the reference panel. We used this reduced panel while preparing the input for pantas. We evaluated the accuracy of the 4 tools (pantas, rMATS, SUPPA2, and whippet) in terms of Precision, Recall, and F1-Measure computed by comparing the events reported by asimulator with the events reported by the tools. More details on the simulation settings can be found in Suppl. Section S3. Table 1 reports the results of this analysis. In the annotated event setting, pantas and rMATS resulted the most accurate tools for detecting AS events. Surprisingly, they achieved the same accuracy on exon skipping, rMATS resulted slightly more accurate on intron retention whereas pantas achieved very good results on alternative splice site events. On the other hand, whippet resulted the most accurate tool for detecting intron retention events. Both whippet and SUPPA2 accuracy was hampered by their low precision. We notice that in this analysis we were interested in evaluating the capability of each tool in detecting AS events and not in quantifying them. A post-filtering of the events, e.g., based on the statistical significance reported by the tools, would have improved the precision of the tools while lowering their recall. In the novel event scenario, instead, we could not include SUPPA2 and whippet. Indeed, SUPPA2 can detect only annotated events. On the other hand, whippet allows for augmenting the Contiguous Splice Graph with novel splice sites supported by read alignments, but we did not manage to run this augmentation successfully (it was able to detect only exon skipping events). In this setting, pantas achieved the highest recall for all event types and resulted the most accurate approach for detecting novel intron retetion and alternative splice site events. On exon skipping events, rMATS achieved a slightly higher precision. Surprisingly, pantas has been able to outperform rMATS on novel intron retentions, which are the hardest AS events to detect correctly [10]. Overall, the good accuracy achieved by pantas prove its correctness and the efficacy of its precise formalization of AS events on annotated spliced pangenomes.

3.2 pantas quantification shows good correlation with state-of-the-art

In the second part of our experiments, we evaluated the quality of pantas differential quantification of AS events from real RNA-Seq data. To this aim, we considered a recent study [3] that performed a genome wide transcriptome analysis of the ageing process in Drosophila Melanogaster (SRA BioProject ID: PRJNA718442). We considered the FlyBase (r6.51) genome and gene annotation and the Drosophila Genetic Reference Panel (v2). In this analysis, we considered only non-overlapping genes to allow for a more precise comparison. Indeed, overlapping genes increase the noise in the output of each tool and add further complexity in the AS quantification step since an AS event on a gene may be supported by reads not directly coming from that gene. Moreover, we performed a subsampling of the input RNA-Seq dataset and we randomly extracted 1/3 of the reads (approximately 8 750 000) from each replicate using seqtk. Indeed, we have not been able to consider the entire samples since, unexpectedly, vg mpmap was requiring more than 3 days to align a single replicate (using 32 threads) and this time requirement was making any analysis unfeasible. Unfortunately, vg mpmap is currently the only approach available to perform spliced alignment against a spliced pangenome but this does not restrict pantas usability, which has been devised and developed as a general approach that can potentially analyze the spliced alignments to a spliced pangenome computed by any spliced aligner. Overall, pantas (without taking into account the input preparation step) complete its differential analysis in less than 2 hours and required 16GB of RAM, making it a practical solution.

Since in this scenario we used real data without any wet-lab validation, we could not compare the results of the tools with a ground-truth but we performed an all-vs-all comparison of the $\Delta\psi$ provided

Annotated Events				
Tool	Type	Prec.	Rec.	F1
pantas	ES	**0.971**	0.988	**0.979**
	IR	0.964	0.936	0.950
	A3	**1.000**	**1.000**	**1.000**
	A5	**1.000**	**1.000**	**1.000**
rMATS	ES	**0.971**	0.988	**0.979**
	IR	0.929	0.989	0.958
	A3	0.928	0.988	0.957
	A5	0.949	0.974	0.961
whippet	ES	0.416	0.976	0.583
	IR	**0.982**	0.947	**0.964**
	A3	0.333	0.488	0.396
	A5	0.297	0.428	0.350
SUPPA2	ES	0.752	**1.000**	0.859
	IR	0.713	**1.000**	0.832
	A3	0.705	**1.000**	0.827
	A5	0.652	**1.000**	0.790

Novel Events				
Tool	Type	Prec.	Rec.	F1
pantas	ES	0.941	**0.958**	0.950
	IR	**0.624**	**0.895**	**0.736**
	A3	0.757	**0.771**	**0.764**
	A5	**0.844**	**0.750**	**0.794**
rMATS	ES	**0.964**	**0.958**	**0.961**
	IR	0.127	0.094	0.108
	A3	**0.801**	0.712	0.754
	A5	0.746	0.678	0.710

Table 1: Results on simulated data from Drosophila Melanogaster. Precision, recall, and F1-scores are computed by comparing `asimulator` truth with the output of each tool. Results are broken down by tool and by event type (ES: Exon Skipping, IR: Intron Retention, A3: Alternative 3', A5: Alternative 5').

by the 4 tools (`pantas`, `rMATS`, `whippet`, and `SUPPA2`). We considered only annotated events (in order to include all tools) and we removed all events without a strong evidence of differential change between the two conditions, i.e., events with $|\Delta\psi| \notin [0.05, 0.95]$. Moreover, since most approaches compute the significance of differential splicing by performing tests on read counts supporting the event junctions, we also filtered out from the resulting set of events all those events reported as *non statistically significant* by the tool. We considered all events reported by `rMATS` and `SUPPA2` with a p-value strictly lower than 0.05 and all events reported by `whippet` with a probability greater or equal to 0.9. Since `pantas` does not provide the statistical significance of an event yet, we used the read counts as a proxy and we considered only those events reported with a coverage of at least 5 alignments on any junction involved in the event. As shown in Fig. 3b, the resulting set of events considered in our analysis consists in 1022 events for `pantas`, 932 events for `rMATS`, 4317 events for `whippet`, and 523 events for `SUPPA2` (Suppl. Table S1 reports these numbers broken down by event type). Out of these events, 147 are shared among the 4 tools. When comparing the quantification ($\Delta\psi$) reported by the tools on this subset of events, we noticed a strong correlation (Fig. 3a). `pantas` achieved very high correlation with `rMATS` (Pearson correlation: 0.94) and `whippet` (0.926) while lower correlation with `SUPPA2` (0.833). Remarkably, `pantas`, `rMATS`, and `whippet` achieved higher correlation when compared among them but lower correlation when compared with `SUPPA2`. This was somehow expected since `SUPPA2` quantifies AS events starting from transcript quantification (based on k-mer analysis) and not from read alignment. As expected, considering all events reported by the tools, i.e., without any post-filtering on the statistical significance, leads to an increase in the number of events (Suppl. Fig. S5b and Suppl.Table S2) and a strong loss of correlation (Suppl. Fig. S5a). However, also in this scenario, `pantas` and `rMATS` have been able to achieve the best correlation (0.811). The good correlation achieved by `pantas` with other approaches based on read alignment proves that read alignment to spliced pangenome is effective and can be confidently used for AS event differential quantification, with the even larger advantage of not being negatively affected by allelic bias.

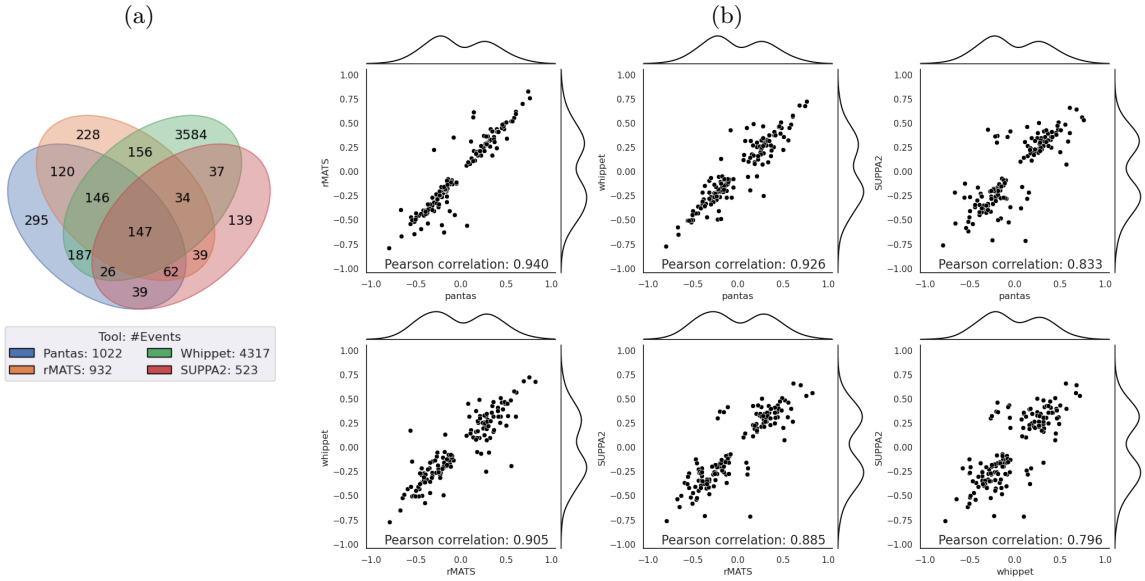

Fig. 3: Results on real data from Drosophila Melanogaster (significant events). (a) Venn diagram showing the number of significant AS events reported by each tool. (b) All-vs-all correlation plots of the $\Delta\psi$ reported by the considered tools for the 147 significant events shared among the tools. Detailed version where each point is color coded based on the event type is available at Suppl. Fig. S4.

3.3 `pantas` is effective in quantifying RT-PCR AS events

In the last part of our experimental evaluation, we considered a real human RNA-Seq dataset and we assessed the performance of `pantas` in detecting and quantifying RT-PCR validated AS events. Similarly to recent studies [37,15,33], we analyzed the RNA-Seq dataset provided by [6] (SRA BioProject ID: `PRJNA255099`) and we evaluated `pantas` accuracy in quantifying RT-PCR validated events. The RNA-Seq dataset consists of three replicates for two conditions (control condition versus double knockdown of the TRA2 splicing regulatory proteins, TRA2A and TRA2B). We considered the human genome and protein coding genes from Ensembl [27] (release 109) and the 1000 Human Project phase 3 VCF files [12].

Since the construction and indexing of the annotated spliced pangenome required more than one day and 240GB of RAM and aligning the 6 replicates required 1 week (using 32 threads), we designed an alternative procedure to build an annotated spliced pangenome restricted to a set of genes of interest and then align only a subset of the input RNA-Seq dataset. This procedure, detailed in Suppl. Section S4, is especially effective when the user is interested in analyzing a panel of genes, that is a common scenario in transcriptomics [14]. Using this alternative approach, we managed to reduce the times of the entire pipeline to less than 6 hours and the RAM requirements to less than 16GB, without degrading the AS events detection accuracy. We note that the most expensive steps of the entire pipeline are graph construction and indexing (1 hour and 20 minutes, 14GB of RAM) and read alignments (40 minutes per sample on average using 32 threads). All other tools (`rMATS`, `SUPPA2`, `whippet`) have been run on the original inputs, without any preprocessing.

We evaluated each tool accuracy in terms of correctly detected events w.r.t. the RT-PCR validated events and then we compared the predicted $\Delta\psi$ with the experimental $\Delta\psi$ provided by RT-PCR validation. Our evaluation focused on 77 RT-PCR validated exon skipping events (details on preprocessing of the truthset can be found in Suppl. Section S5). Fig. 4 reports the results of our analysis when considering only events reported as statistically significant by the tools (using the same criteria ap-

plied in Section 3.2) and with an absolute magnitude of change greater than 0.05 between the two conditions, i.e., events with $|\Delta\psi| \in [0.05, 0.95]$. Fig. S6 reports the same results obtained without any filtering. Regarding the number of events detected by each tool, `pantas` reported the highest number of significant events (64 out of 77) followed by `rMATS` (61), `SUPPA2` (44), and `whippet` (40). We note that `SUPPA2` and `whippet` reported less events since 14 out of the 77 RT-PCR events are novel events and, as previously shown, they cannot correctly detect novel events. Remarkably, only 25 events were reported as significant by all 4 tools, 3 events were only reported by `pantas` whereas 4 events were missed by our approach. The reason behind these results needs to be sought in the different way each tool computes the *significance* of an event. Indeed, as shown in Fig. S6, when considering all events (i.e., not only those reported as significant), each tool reported a higher number of events (as expected) and the number of events reported by all 4 tools increased to 46. When removing `whippet` from this analysis, the number of events reported by `pantas`, `rMATS`, and `SUPPA2` increased even more (64), showing good agreement among the three tools. Moreover, in this setting no event was reported by `pantas` only. However, 7 events remained undetected by `pantas`. By manually investigating both `STAR` and `vg mpmap` alignments over the loci of these events, we noticed that the two aligners show good agreement: the skipping junction of all events shows no support in the control replicates and very small support (1 or 2 alignments) in the knockdown replicates. A more polished statistical analysis of `pantas` results will allow to recover these events. Surprisingly, when considering significant events reported by the tools, 8 RT-PCR events were not reported by any tool. On the other hand, when considering all events reported by the tools, all RT-PCR events where detected, proving the complexity of correctly computing (and analyzing) the significance of differentially quantified AS events.

We then evaluated the correctness of the quantification provided by each tool by comparing the $\Delta\psi$ predicted by each tool with the experimental $\Delta\psi$ provided by the RT-PCR validation. As shown in Fig. 4, `whippet` quantification showed the best correlation with the RT-PCR expected quantification (Pearson correlation: 0.767) followed by `pantas` (0.692). However, `whippet` reported less events than `pantas` (40 against 64). Remarkably, although the distributions of the differences between the RT-PCR $\Delta\psi$ and the quantification provided by `pantas` and `rMATS` are very similar (with a 0.006 median difference in favor of `rMATS` and 0.005 average difference in favor of `pantas`), `rMATS` showed lower correlation than `pantas`. However, as previously reported [37], all tools struggle in achieving a good correlation with the RT-PCR quantification. The same trend can be observed when considering all events reported by the tools (Fig. S6). Remarkably, the correlation between `pantas` quantification and RT-PCR quantification increased and reached the correlation achieved by `whippet`. This indicates that most of the events that have not been considered as significant by `pantas` are actually well quantified, proving once more the need of improving the statistical significance analysis of `pantas` results. This will be the focus of future efforts.

Finally, as done in [37], we evaluated the false positive rate of each tool using 44 RT-PCR negative exon skipping events that did not show any significant change between the two conditions. `SUPPA2` and `pantas` reported the fewest false positives (28 and 31, respectively, out of 44 events) whereas `whippet` reported 32 false events and `rMATS` reported 37. When considering only events reported as significant by each tool, `pantas` reported 9 false significant events followed by `rMATS` (10 events). `SUPPA2` and `whippet` reported the lowest number of false significant events (2 and 4, respectively).

In conclusion, although `pantas` statistical validation is still not fully polished, the differential quantification computed from alignments to an annotated spliced pangenome is robust: `pantas` is able to retrieve more significant AS events without introducing more false calls. Overall, the results described in this section suggest a strong agreement between the different approaches for the differential quantification of AS events with no method clearly outperforming the others.

4 Conclusion

The recent adoption of a reference pangenome for analyzing genetic variability in humans opens the perspective of building a pantrascriptome that provides a complete picture of gene annotation in

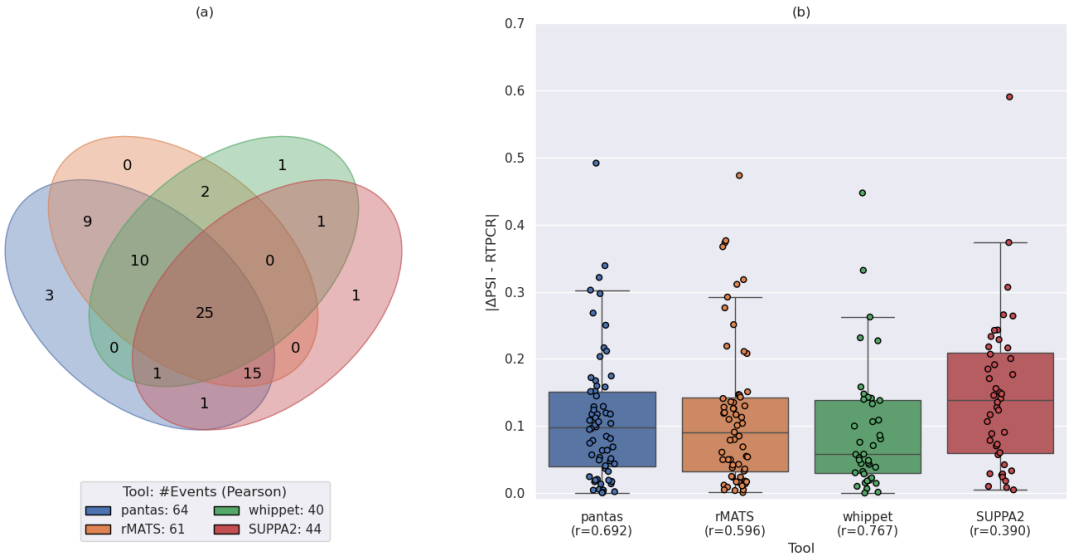

Fig. 4: (a) Venn diagram showing the number of significant AS events reported by each tool. The legend reports the total number of events reported by the tool. (b) Boxplot showing the distribution of the difference between the $\Delta\psi$ predicted by each tool and the $\Delta\psi$ provided by RT-PCR. The x axis labels report the tool name and the Pearson correlation (r) between the predicted $\Delta\psi$ and the RT-PCR $\Delta\psi$.

.

human population. However, moving from a reference-based gene annotation to a pangenome-based (or haplotype-aware) gene annotation, where genes are annotated w.r.t. multiple genomes as recently proposed in [1], requires to rethink tools for transcriptomic analysis, similarly as it has been done in computational graph pangenomics for analyzing genomic variants in a population.

In this paper, we advanced the investigation started in [32] of spliced pangenomes by proposing `pantas`, the first pangenomic approach for detecting and quantifying AS events across RNA-Seq conditions. Our approach is based on the novel notion of annotated spliced pangenome, introduced here as an enhanced version of a spliced pangenome, and on the precise formalization of AS events on this structure. Extensive experimental evaluation shows that `pantas` achieves comparable results with state-of-the-art methods based on a linear reference, while being able to potentially take into account the genetic variability of the population under investigation. In the experiments carried out in this paper we used a reference-based gene annotation (that is the one currently available), but `pantas` can also work with haplotype-aware gene annotations.

Future work will be devoted to enhance the statistical validation of `pantas` quantification. Moreover, we also planned to closely collaborate with the pangenomic community in order to reduce the high computational requirements for building and indexing a spliced pangenome as well as for mapping RNA-Seq reads to it. These are the inputs of `pantas` as well as of any future tool for the analysis of a pantranscriptome. Making these indispensable steps as efficient as possible will boost the effectiveness of spliced pangenomes and widen their applicability to more areas of transcriptomic.

References

1. Amaral, P., Carbonell-Sala, S., De La Vega, F.M., Faial, T., Frankish, A., Gingeras, T., Guigo, R., Harrow, J.L., Hatzigeorgiou, A.G., Johnson, R., et al.: The status of the human gene catalogue. Nature **622**(7981), 41–47 (2023)
2. Baaijens, J.A., Bonizzoni, P., Boucher, C., Della Vedova, G., Pirola, Y., Rizzi, R., Sirén, J.: Computational graph pangenomics: a tutorial on data structures and their applications. Natural Computing pp. 1–28 (2022)
3. Bajgiran, M., Azlan, A., Shamsuddin, S., Azzam, G., Halim, M.A.: Data on rna-seq analysis of drosophila melanogaster during ageing. Data in brief **38**, 107413 (2021)
4. Beretta, S., Bonizzoni, P., Della Vedova, G., Pirola, Y., Rizzi, R.: Modeling alternative splicing variants from rna-seq data with isoform graphs. Journal of Computational Biology **21**(1), 16–40 (2014)
5. Beretta, S., Bonizzoni, P., Denti, L., Previtali, M., Rizzi, R.: Mapping rna-seq data to a transcript graph via approximate pattern matching to a hypertext. In: Algorithms for Computational Biology: 4th International Conference, AlCoB 2017, Aveiro, Portugal, June 5-6, 2017, Proceedings 4. pp. 49–61. Springer (2017)
6. Best, A., James, K., Dalgliesh, C., Hong, E., Kheirolahi-Kouhestani, M., Curk, T., Xu, Y., Danilenko, M., Hussain, R., Keavney, B., et al.: Human tra2 proteins jointly control a chek1 splicing switch among alternative and constitutive target exons. Nature communications **5**(1), 4760 (2014)
7. Bhadra, M., Howell, P., Dutta, S., Heintz, C., Mair, W.B.: Alternative splicing in aging and longevity. Human genetics **139**, 357–369 (2020)
8. Biamonti, G., Amato, A., Belloni, E., Di Matteo, A., Infantino, L., Pradella, D., Ghigna, C.: Alternative splicing in alzheimer's disease. Aging clinical and experimental research **33**, 747–758 (2021)
9. Bonnal, S.C., López-Oreja, I., Valcárcel, J.: Roles and mechanisms of alternative splicing in cancer—implications for care. Nature reviews Clinical oncology **17**(8), 457–474 (2020)
10. Broseus, L., Ritchie, W.: Challenges in detecting and quantifying intron retention from next generation sequencing data. Computational and structural biotechnology journal **18**, 501–508 (2020)
11. Cheng, J., Nguyen, T.Y.D., Cygan, K.J., Çelik, M.H., Fairbrother, W.G., Gagneur, J., et al.: Mmsplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome biology **20**(1), 1–15 (2019)
12. Consortium, .G.P., et al.: A global reference for human genetic variation. Nature **526**(7571), 68 (2015)
13. Cozzi, D., Bonizzoni, P., Denti, L.: Esgq: Alternative splicing events quantification across conditions based on event splicing graphs pp. 242–249 (2023)
14. Denti, L., Pirola, Y., Previtali, M., Ceccato, T., Della Vedova, G., Rizzi, R., Bonizzoni, P.: Shark: fishing relevant reads in an rna-seq sample. Bioinformatics **37**(4), 464–472 (2021)
15. Denti, L., Rizzi, R., Beretta, S., Della Vedova, G., Previtali, M., Bonizzoni, P.: Asgal: aligning rna-seq data to a splicing graph to detect novel alternative splicing events. BMC bioinformatics **19**(1), 1–21 (2018)
16. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
17. Fenn, A., Tsoy, O., Faro, T., Rößler, F.L., Dietrich, A., Kersting, J., Louadi, Z., Lio, C.T., Völker, U., Baumbach, J., et al.: Alternative splicing analysis benchmark with dicast. NAR Genomics and Bioinformatics **5**(2), lqad044 (2023)
18. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al.: Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology **36**(9), 875–879 (2018)
19. Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.F., Hammond, S.M., Makowski, L., et al.: Diffsplice: the genome-wide detection of differential splicing events with rna-seq. Nucleic acids research **41**(2), e39–e39 (2013)
20. Kahles, A., Ong, C.S., Zhong, Y., Rätsch, G.: Spladder: identification, quantification and testing of alternative splicing events from rna-seq data. Bioinformatics **32**(12), 1840–1847 (2016)
21. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., Pritchard, J.K.: Annotation-free quantification of rna splicing using leafcutter. Nature genetics **50**(1), 151–158 (2018)
22. Li, Y.I., Van De Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., Pritchard, J.K.: Rna splicing is a primary link between genetic variation and disease. Science **352**(6285), 600–604 (2016)
23. Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al.: A draft human pangenome reference. Nature **617**(7960), 312–324 (2023)
24. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., Guigó, R.: Are splicing mutations the most frequent cause of hereditary disease? FEBS letters **579**(9), 1900–1903 (2005)

25. Mackay, T.F., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al.: The drosophila melanogaster genetic reference panel. Nature **482**(7384), 173–178 (2012)
26. Manz, Q., Tsoy, O., Fenn, A., Baumbach, J., Völker, U., List, M., Kacprowski, T.: Asimulator: splice-aware rna-seq data simulation. Bioinformatics **37**(18), 3008–3010 (2021)
27. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al.: Ensembl 2023. Nucleic acids research **51**(D1), D933–D941 (2023)
28. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C.: Salmon provides fast and bias-aware quantification of transcript expression. Nature methods **14**(4), 417–419 (2017)
29. Rogers, M.F., Thomas, J., Reddy, A.S., Ben-Hur, A.: Splicegrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data. Genome biology **13**(1), 1–17 (2012)
30. Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., Hubner, N.: Alternative splicing signatures in rna-seq data: percent spliced in (psi). Current protocols in human genetics **87**(1), 11–16 (2015)
31. Shen, S., Park, J.W., Lu, Z.x., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., Xing, Y.: rmats: robust and flexible detection of differential alternative splicing from replicate rna-seq data. Proceedings of the National Academy of Sciences **111**(51), E5593–E5601 (2014)
32. Sibbesen, J.A., Eizenga, J.M., Novak, A.M., Sirén, J., Chang, X., Garrison, E., Paten, B.: Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nature Methods pp. 1–9 (2023)
33. Sterne-Weiler, T., Weatheritt, R.J., Best, A.J., Ha, K.C., Blencowe, B.J.: Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. Molecular Cell **72**(1), 187–200.e6 (2018)
34. Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R., Skotheim, R.: Aberrant rna splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene **35**(19), 2413–2427 (2016)
35. Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V., et al.: Flybase 2.0: the next generation. Nucleic acids research **47**(D1), D759–D765 (2019)
36. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L.: Differential analysis of gene regulation at transcript resolution with rna-seq. Nature biotechnology **31**(1), 46–53 (2013)
37. Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., Eyras, E.: Suppa2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome biology **19**, 1–11 (2018)
38. Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., Gonzalez-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., Barash, Y.: A new view of transcriptome complexity and regulation through the lens of local splicing variations. elife **5**, e11752 (2016)