The Junction Usage Model (JUM): A method for comprehensive annotation-free differential analysis of tissue-specific global alternative pre-mRNA splicing patterns

Qingqing Wang[1,2,3] and Donald C. Rio[1,2,3 *]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA.

[2]Center for RNA Systems Biology (CRSB), University of California, Berkeley, California 94720, USA.

[3]California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, California 94720, USA.

*Correspondence to: Donald Rio, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. E-mail: don_rio@berkeley.edu

Abstract

Alternative pre-mRNA splicing (AS) generates exceptionally diverse transcriptome and proteome profiles that critically affect eukaryotic gene expression in different tissues, developmental stages and disease. However, current efforts to evaluate tissue-specific AS patterns rely completely or partially on an annotated libraries of known gene transcripts, which hinders the analysis of AS patterns that are novel or specific to the cell/tissue or for non- or poorly annotated genomes. To tackle this problem, we describe a method called the Junction Usage Model (JUM) that offers a *de novo* approach to analyze tissue-specific AS profiles without any prior knowledge of the transcriptome. JUM exclusively uses RNA-seq reads mapped to splice junctions to construct statistical models and to accurately quantify AS changes, and then faithfully reconstructs the detected splice junctions into AS patterns based on their unique topological features. Compared to other recent methods, we found that JUM consistently identified true novel tissue-specific AS events that could not be identified by other methods, and further rejected false positive and/or misclassified AS events. In summary, JUM provides a new framework and software that enables the thorough investigation of the dynamic and tissue-specific AS regulation in a wide range of cells, tissues and organisms.

Introduction

Alternative pre-mRNA splicing (AS) is a major gene regulatory mechanism that greatly expands proteomic diversity and serves as a crucial determinant of cell fate and identity. More than 95% of human gene transcripts undergo AS that enables one single gene locus to produce multiple, and usually functionally distinct pre-mRNA and protein isoforms[1,2]. AS is under complex regulation using a large constellation of RNA-binding proteins that interact with *cis*-acting RNA elements embedded in nuclear pre-mRNA sequences and determine the AS of pre-mRNAs in cell- and tissue-specific patterns[3,4]. Specific AS mRNA isoforms are closely associated with distinct cellular states and these different mRNAs affect almost every aspect of cellular function, including proliferation, differentiation, apoptosis and migration[1,5,6]. Furthermore, mutations that result in aberrant AS are a major source for human diseases, such as cancer, immune disorders and neurodegeneration[7-9]. Given the role of defects in pre-mRNA splicing in human disease, a thorough and comprehensive evaluation of static global AS profiles, as well as the dynamic patterns of AS during development, differentiation and in specific tissue types will be critical to understand cellular disease states and facilitate the development of screening and therapeutic strategies to diagnose, treat and prevent many diseases linked to defects in AS. Examination of tissue-specific RNA-seq data sets has already revealed exceptionally diverse and dynamic features of AS that are tissue- and developmental stage-specific[1,5]. However, given this complexity the thorough and systematic quantification and analysis of cellular AS profiles among a complex array of tissues or cell types remains a major unsolved challenge in the bioinformatics of gene expression.

Recent technical advances in short-read high-throughput Illumina transcriptome sequencing (RNA-seq) provides powerful tools with which to investigate AS at the genome-wide

scale, but at the same time presents a formidable computational challenge to accurately quantify global AS changes from raw RNA-seq data. Previously, a number of computational software tools and algorithms have been developed for this purpose[7,10-18], and although they apply different statistical models, all of these tools rely either completely or partially on a developer- or user-defined pre-annotated library of known AS events or an incompletely annotated transcriptome. These splice junction annotation libraries can be generated using two different approaches: 1) combining known mRNA transcript isoforms and AS patterns from publicly available databases, such as the UCSC genome browser and Ensembl; and 2) assembling the transcriptome from RNA-seq data in a *de novo* way. For the first approach, the major caveat is that it restricts AS analysis to only previously known or annotated AS events. Many recent RNA-seq experiments showed that a plethora of novel and functional AS patterns are constantly found in biological samples and using pre-built libraries of known AS events underestimates the complexity of AS in these samples[2,19,20] (Fig. 1a). Another caveat is that a fixed library of annotated AS events (usually compiled from all available tissues and cell type mRNAs from the organism in public databases) neglect the dynamic and tissue-specific characteristics of AS. In many cases, an annotated splicing event in an AS library can present a distinct pattern in the specific sample under study that is different from the annotated pattern[21] (Fig. 1b). In this scenario, using the fixed, annotated patterns in the library will hinder analysis of the actual AS patterns in the specific RNA sample. Some recent methods (such as rMATS[15] and MAJIQ[18]) are equipped with a "*de novo*" working mode, in which the previously annotated AS events are supplemented with novel splice junction-implicated AS events in the sample profiled by the software. However, the annotated library is still the primary source for AS events, which directly affects the accuracy and comprehensiveness of AS analysis and can mislead the detection of

novel AS events by the software. The second approach without an annotated library of splice junctions on the other hand, can identify novel AS isoforms in the sample and extend AS analysis to novel, sample-specific AS events. However, although there have been a few tools that offer re-assembling transcriptomes from RNA-seq data using probabilistic models[22-26], a precise and deterministic *ab initio* assembly of transcriptomes from shotgun RNA sequencing remains a big challenge in the field. Especially for genes that produce multiple transcripts with complex AS patterns, it is very difficult to accurately identify all splicing isoforms simply from short Illumina sequencing reads, and the difficulty in transcriptome assembly directly affects the accuracy of downstream AS analysis. Considering the caveats and difficulties described above, there is an urgent need for development of tools that can perform accurate, comprehensive, and tissue-specific global AS analysis that do not depend on prior knowledge of annotated AS libraries or transcriptome.

Here, we present a new method called the Junction Usage Model (JUM) that performs *de novo* differential analysis of global AS patterns completely independent of a priori knowledge of AS event libraries or transcriptome annotations. JUM utilizes sequence reads that are directly mapped over splice junctions to quantify and analyze AS patterns and through evaluating the "usage" of splice junctions providing a complete and accurate evaluation of global AS patterns that are specific to the biological sample(s) under study. JUM furthermore faithfully re-translates the analyzed junctions into AS patterns based on the unique topological features of each pattern. Specifically, JUM is equipped with stringent statistical criteria to accurately analyze the AS pattern of intron retention, an understudied AS category. JUM provides a new approach to study the extraordinarily diverse global cellular AS transcriptome profiles and the dynamic regulation of AS, and can be readily applied to a wide range of RNA samples for differential AS analysis

with high specificity, accuracy and sensitivity.

Results

JUM utilizes short sequence reads spanning splice junctions for AS analysis and defines AS structures as the basic quantification unit

Many currently available methods for AS analysis utilize Illumina sequence reads that are mapped to a full or partial AS isoform to quantify the levels of that isoform, or reads mapped to every exon in the gene transcripts to evaluate if a specific exon is more included or excluded[10-12,14] (Supplemental Fig. 1). JUM, however, exclusively uses sequence reads mapped over splice junctions for AS quantification (Fig. 2a), as these reads provide the most direct evidence for the splicing of the corresponding intron, and the number of reads mapped to a splice junction is an unambiguous measure for the level of splicing.

From there, JUM defines an AS structure as the basic quantification unit for AS analysis. An AS structure is a set of splice junctions that share the same start genomic coordinate or the same ending coordinate, with each splice junction in an AS structure defined as a sub-AS-junction (Fig. 2a,2b). The AS structures are the essential elements that form the conventionally recognized AS patterns: alternative 5' splice site (A5'SS), alternative 3' splice site (A3'SS), cassette exon (CE) and mutually exclusive exons (MXE). For example, an A5'SS or A3'SS event is composed of one AS structure with two sub-AS-junctions (Fig. 2b); a CE event is composed of two AS structures, each with two sub-AS-junctions (Fig. 2c); a MXE event with two mutually exclusive exons is composed of two AS structures, each with two sub-AS-junctions (Fig. 2d). After the profiling of all AS structures, JUM counts sequence reads that are mapped to each sub-AS-junction in these AS structures under each biological condition, and defines the

read count as the "usage" of a sub-AS-junction in the corresponding AS structure under that condition.

JUM performs global differential AS analysis by quantifying the differential usage of sub-AS-junctions in AS structures upon various biological conditions

To quantify global AS profile changes, JUM compares the usage of every profiled sub-AS-junction in its corresponding AS structure between the control sample and a specific biological condition, and identifies all AS structures that contain sub-AS-junctions with differential usage (Fig. 3a). To do this, JUM utilizes multiple biological replicates to build robust statistical power directly from raw read counts mapped to each sub-AS-junction. JUM models the total number of reads that map to a sub-AS-junction as negative binomial distribution (Fig. 3b, Eq. 1). Negative binomial distributions have been widely applied in high-throughput sequencing data analysis to model read counts, as these models nicely depict the over-dispersion phenomenon observed in next-gen sequencing experiments[11,27-30]. In negative binomial distributions, the variance among biological replicates is dependent on the mean through a parameter that describes dispersion (Fig. 3b, Eq. 2). To infer the dispersion parameter, JUM applies a similar empirical Bayes approach as described[28-31]. JUM first estimates a dispersion parameter for each sub-AS-junction with Cox-Reid-adjusted maximum likelihood. JUM then fits a mean-variance function for all sub-AS-junctions from all AS structures on their average normalized count values. Finally, JUM shrinks the dispersion parameter for each individual sub-AS-junction towards the fitted value depending on how close the real dispersion tends to be to the fitted value and the replicate sample size[28-31].

To evaluate if a biological condition significantly changes the usage of a sub-AS-junction in the corresponding AS structure, JUM adapts a generalized linear model (GLM) approach as described[11,30,32], so that two GLM models are fitted and tested for each sub-AS-junction in the corresponding AS structure[11] (Fig. 3c). The basal model evaluates the effect from the following three elements to the usage of the sub-AS-junction: the basal expression level of the AS structure of the corresponding gene (α_i^s, Fig. 3c, Eq. 4), the fraction of sequence reads that mapped to each sub-AS-junction from the total number of reads mapped to the AS structure (α_{ij}^E, Fig. 3c, Eq. 4), as well as the overall change of basal expression of the AS structure upon a biological condition ($\alpha_{i\varepsilon_k}^C$, Fig. 3c, Eq. 4). The effect model evaluates an additional influence imposed on the usage of the sub-AS-junction by a biological condition ($\alpha_{ij\varepsilon_k}^C$, Fig. 3c, Eq. 3). The fitting of the effect and basal model are compared and a χ^2 likelihood-ratio test performed[11] so as to test if a biological condition causes significant differential usage of a sub-AS-junction in the corresponding AS structure.

JUM faithfully re-constructs AS structures into conventionally recognized AS patterns without priori knowledge of the transcriptome or a previous AS event annotation

AS structures with differential usage of sub-AS-junctions directly reflect changes in splicing patterns, but the concept of AS structures is abstract, compared to the simple and fixed five conventional categories of AS patterns (A5'SS, A3'SS, CE, MXE, and intron retention—IR) widely recognized by RNA biologists. Thus, after global differential AS analysis using AS structures, JUM re-constructs the profiled AS structures into the conventional categories of AS patterns, but does so in a *de novo* way that is completely independent of any a prior knowledge

of the transcriptome annotation nor the use of annotated libraries of AS events. As a result, the JUM output reports a list of AS events that are specific to the biological sample in each of the AS pattern categories, and also identifies those AS events whose splicing are significantly altered under a biological condition. To do this, JUM first converts each AS pattern into graphs by converting exons into nodes and representing splice junctions as arcs that connect to the exon nodes. Based on the unique topological features of the graphs representing each AS pattern and employing tools from graph theory, JUM then translates AS structures into AS patterns. JUM defines a frequency parameter S_I for each sub-AS-junction as the number of AS structures containing that specific sub-AS-junction. It can be proven that a given sub-AS-junction can only be included in up to two AS structures (S_I can only be 1 or 2), based on the definition of AS structures.

For the AS pattern of A5SS or A3SS, the graph for these two AS patterns are asymmetric, and is composed of only one AS structure with sub-AS-junction S_I value all equal to 1 (Fig. 4a and 4b). For the AS pattern of CE, the graph for CE is symmetric, and is composed of two AS structures, each AS structure containing two sub-AS-junctions with S_I value 1 and 2, respectively (Fig. 4c); extra quality control criterion here includes tiled sequence read support over the entire cassette exon region. For the AS pattern of MXE with n mutually exclusive exons, the graph for MXE is composed of one pair of A5SS- or A3SS-like AS structures, each has n sub-AS-junction with S_I value all equal to 1 (Fig. 4d). In this case, extra quality control criteria include: coordinates of MXE exons meet condition $a_i < b_i < a_{i+1}$, where $i = 1,..,n$ (Fig. 4d) and tiled sequence read support over the entire regions of all mutually exclusive exons. Based on the unique graphical symmetry of the AS structure composition in each AS pattern,

JUM searches for sets of AS structures that match each AS pattern and bundle them together as one AS event under the corresponding AS pattern category (for the detailed algorithm, see Online Methods).

It should be noted that JUM defines an additional AS pattern category called "composite AS", which describes an AS event that is a coherent combination of several conventionally recognized AS patterns. Such complex AS events are usually not included in currently available AS computational analysis tools, but can be naturally found extensively in physiological RNA samples from Drosophila, rodents and humans (Fig. 4e), further illustrating the complexity and diversity of AS in specific tissues. For example, JUM observes an AS event in the transcripts of gene Eif-4E in Drosophila male heads that is a combination of CE, A5'SS and A3'SS patterns and can not be simply decomposed into any of the three categories (Fig. 4e). JUM is able to recognize such complex AS patterns and quantify their AS changes, because the topological features for a composite AS pattern are a linear combination of the corresponding conventional AS patterns that form the composite AS pattern. For example, a composite AS pattern event that is a combination of an A5'SS and CE is composed of four AS structures, three of them have sub-AS-junction S_I value all equal to 2 and one AS structure has a sub-AS-junction S_I value equal to 1 with the rest S_I value equal to 2 (Fig. 4e).

JUM performs stringent differential AS analysis on intron retention

Compared to other AS patterns, intron retention (IR) has been an under-investigated category but nevertheless a key AS mechanism. Many IR events in eukaryotes have been shown to play crucial roles for the normal functioning of the organism. For example, tissue-specific IR of the Drosophila P-element transposase transcripts enables the restriction of transposon activity only

in the germ line tissues[33,34]. Recently, a bioinformatics study reported that widespread retained intron was associated with various cancer types compared to normal tissues[35]. Increased IR has also been shown to be associated with the pluripotent state of stem cells[36]. For analyzing IR, two most commonly applied methods to quantify intron-retained Mrna isoforms is to either use the sum of sequence reads mapped to the upstream exon-intron boundary and the downstream intron-exon boundary, or to use reads mapped to the intronic region (Fig. 5a). The intron-exclusion isoform is usually quantified by using the reads that mapped to the splice junction that resulted from the splicing of the intron (Fig. 5a). A major caveat for these methods of measurement of intron retention is that many other AS patterns can be mistaken for IR, especially if using a pre-built annotated AS event library. Many introns listed in a pre-built AS library that are classified as intron retention can actually be due to more complicated AS within the intronic region in a specific sample under study (Fig. 5b,c,d), the most common cases being CE or MXE exons residing within the intron (Fig. 5b,c,d) or an A5'SS/A3'SS event at the edge of the intron (Fig. 5d). In situations like these, sequence reads mapped to the intronic region can in fact come from exonic reads for CE or MXE exons and reads mapped to exon-intron or intron-exon boundaries can actually come from A5'SS or A3'SS events, but these reads are mistakenly used as support for intron-retained isoforms.

To avoid false-positive calls of IR as described above, JUM applies a stringent three-criteria strategy to perform differential AS analysis on IR (Fig. 5e). JUM first profiles for all of the valid splice junctions from the RNA-seq data. For each splice junction and the corresponding intron, JUM counts the number of sequence reads mapped to the upstream exon-intron boundary (N_1), the number of reads mapped to the splice junction (N_2) and the number of reads mapped to the downstream intron-exon boundary (N_3) (Fig. 5e). JUM then defines two AS structures for

each intron: N_1 versus N_2, as well as N_3 versus N_2. For an intron to be classified as significantly changed IR, both AS structures must be differentially "used" with the same trend (N_1 significantly differing from N_2 and N_3 differing from N_2, with the same trend). These two criteria are set to avoid an A5'SS or A3'SS event to be mistaken as IR (Fig. 5e). Finally, JUM requires sequence reads mapped to the intronic region to be approximately uniformly distributed all across the intron, in order to confirm the retention of the whole intron (Fig. 5e). This criterion is used to prevent a CE or MXE event to be mistaken as IR, as reads mapped to CE or MXE exons in the intron will present a "spikey" read distribution compared to the whole intronic region.

We tested the performance of JUM on IR analysis on different sample types. A comparison of JUM and MISO[10] in analyzing IR events in Drosophila male fly head RNA samples showed that JUM significantly deduced the false positive rate of IR identification compared to MISO (Fig. 5b,c,d and Supplemental Table 1, Supplemental Fig. 2). Moreover, Dvinge et al.[35] recently used MISO and a pre-annotated human AS event library to compare the AS profiles of patient tumor and matched normal tissues in The Cancer Genome Atlas (TCGA) database and reported that extensive retained intron is a featured and highly elevated pattern of splicing observed in many different cancer types[35]. To test if elevated intron retention is indeed associated with cancer, we used JUM to analyze the global AS patterns in colon cancer patient tumor versus matched normal colon tissue RNA-seq datasets from the TCGA database. Colon cancer was chosen for JUM analysis because it is one of the cancer types reported in Dvinge et al.[35] to have the most elevated intron retention. In addition, almost no biological replicates were provided for the tumor and matched normal tissue sample for each patient in TCGA. Dvinge et al. used the median of the number of increased retained IR events across all patients as a

measurement for evaluating IR in cancer. This measurement does not account for biological variance among cancer patients and thus does not provide high statistical significance for evaluating increased IR in cancer (Supplemental Fig. 3). The importance of integrating biological variation has been emphasized by numerous previous studies[11,29,30,37,38] in order to draw meaningful and accurate conclusions about a biological process with certainty. JUM is capable of incorporating biological variance across patients to build robust statistics. We used JUM to analyze five male colon cancer patients that are of similar ages (60-68 years old), tumor type, vital states and with matched tumor and normal tissue samples sequenced from the same platform (Supplemental Table 2). Interestingly, in contrast to previous analyses[35], JUM does not identify significantly elevated or widespread intron retention events in the human tumor samples. Only a total of 98 IR events are significantly changed in colon tumor versus normal tissues (Supplemental Table 1; Supplemental Fig. 4). Among them, 59 IR events showed an increase of the retained intron isoform and 39 IR events showed an increase of the intron exclusive isoform (Supplemental Fig. 4). We also used JUM to analyze six female colon cancer patients that are picked using the same standards, and obtained similar results (Supplemental Table 3). From here, we conclude that with the heterogeneity of tumor, widespread IR is not associated with the colon cancer transcriptome.

JUM provides a thorough and accurate differential analysis of tissue-specific global AS profile changes

Experimental validation of JUM analysis: We tested the performance of JUM on various types of RNA-seq datasets derived from RNA samples spanning Drosophila, mouse and human tissues and cell lines. JUM identifies many novel and functionally important AS events in these samples

that can be directly linked to the phenotypes observed in the biological treatment or tissue-type in these studies[20,21]. Importantly, JUM-predicted significant AS pattern changes are validated using experimental methods. For one study in mouse embryonic cortical neurons, we randomly picked ten JUM-predicted AS targets of the splicing factor PQBP1 and two JUM-predicted non-AS targets for verification, and all 12 AS events were validated by RT-PCR[20]. Among them, the significant AS changes of the *Ncam1* transcripts upon PQBP1 knockdown was only discovered through JUM but not other tools in use, and this AS change in *Ncam1* is functionally associated to the dendritic outgrowth defect observed in PQBP1-perturbed neurons[20] (Fig. 1a). For another study in male fruit fly heads, the significant AS changes of the *fruitless* mRNA transcript isoforms that are directly linked to abnormal male courtship behavior in a Drosophila strain were captured exclusively using JUM analysis, but not other annotation-library-based AS analysis software in comparison, and validated using RT-PCR[21].

JUM performance compared to other methods: To further evaluate the performance of JUM in differential AS analysis, we compared JUM to two recently developed AS analysis tools, MISO[10] and rMATS[15], in analyzing the Drosophila head transcriptomes between the wildtype and a male-courtship defect strain[21]. MISO completely depends on a developer-provided annotated library of AS events[10] while rMATS, although also depends on a user-provided annotated library of known gene transcripts, offers a "*de novo*" mode to detect novel AS events in the sample[15]. We compared JUM with MISO and rMATS (with the *de novo* mode) (Supplemental Table 5).

We first checked if MISO and rMATS also detected the 26 JUM-identified and RNA-seq experiment-validated significantly changed AS events in Drosophila male head that are closely

associated with the courtship behavior defect phenotype[21] (Fig. 6a; Supplemental Table 6). Among them, we found that the majority of AS events (14 events, 56%) are exclusively identified by JUM, due to the fact that neither MISO nor rMATS recognizes these novel, Drosophila male head sample-specific AS events in the annotated AS library nor by the *de novo* detection mode (Fig. 6a; Supplemental Table 6); 4 AS events (16%) are identified by MISO and rMATS also and interestingly, all of these events are cassette exon patterns, the most well studied and annotated AS pattern type (Fig. 6a; Supplemental Table 6); 1 AS event is identified by rMATS also but not MISO and 1 by MISO also but not rMATS (4%); 1 AS event is identified by rMATS and MISO also, respectively, however with only part of the correctly annotated coordinates (4%) (Supplemental Table 6). These results suggest that JUM is capable of identifying true novel, tissue-specific AS events that could not be recognized by annotation-based or partially annotation-dependent techniques (even with a *de novo* working mode).

We then took the top 30 most significantly changed AS events identified by rMATS in the most well-studied and annotated AS pattern category, cassette exon, and tested if JUM can identify these CE events as well. We found that among them, only 5 CE events are also identified by JUM as CE events (17%) (Fig. 6b, Supplemental Table 7). However, for the rest 25 CE events identified by rMATs, the majority (21 events, 70%) are in fact mis-classified as CE events in the fly heads. JUM identified them as significant AS events but re-classified them correctly as composite AS patterns (Fig. 6b; Supplemental Table 7); 4 CE events (17%) are not identified in JUM, because JUM did not include the involved junctions for downstream analysis in the first place, due to a custom-set quality control setting for filtering valid splice junctions (Online Methods). These events are identified by JUM once the setting is adjusted. These results suggest that JUM efficiently rejects false positive AS events suggested by other annotation-

dependent techniques and re-classifies mis-annotated AS events to the correct category based on the actual AS pattern in the tissue.

Discussion

As a major mechanism for eukaryotic gene regulation, AS generates exceptional diversity. Different tissues, even sub-cellular populations within a given tissue or organ possess their own distinct AS profiles and these profiles are dynamically changed over different temporal stages of development and cellular activities. Such transcriptome diversity and dynamics of AS patterns impose a major challenge for computational tools to quantify and compare AS profiles from RNA-seq data. Many of the currently available AS analysis software tools employ the strategy of using pre-built annotated libraries of AS event derived from previous EST or RNA-seq data. This strategy greatly facilitates downstream analysis, but at the same time fails to address, detect and classify the diversity of AS, because different RNA samples can present their own unique AS profiles and many novel AS events. Some splicing analysis tools can use a *de novo* built transcriptome annotation in order to focus AS analysis on sample-specific AS patterns. However, the difficulty in reconstructing accurate transcriptome annotations from shotgun sequencing RNA-seq data affects the quality of downstream AS mRNA isoform analysis. JUM, on the other hand, uses a completely novel approach for analyzing tissue-specific AS patterns. Instead of relying on priori knowledge of annotated transcriptomes, JUM identifies AS patterns present in the specific RNA samples *de novo* utilizing the unique topological features of each AS pattern classification. JUM then provides differential quantitative analysis of these AS events by modeling short-read sequence reads mapped to splice junctions by integrating biological variance across replicates and statistically testing the effects imposed on

splice junction usage from distinct biological conditions. The approach that JUM uses not only provides a thorough and statistically robust investigation of the diverse and dynamic AS patterns specific to a given biological sample, but also eliminates the labor-intensive computational effort required to build pre-annotated AS event libraries or transcriptome re-assembly.

Another novel feature of JUM lies in its stringent statistical parameters to analyze IR events. IR is a crucial AS regulatory mechanism for gene control and is a clear, yet understudied AS pattern, compared to the other AS categories. The complexity of AS and the composition and structure of IR makes it easy to mistake other AS types as IR events (Fig. 5). JUM designs a well-developed program to analyze IR specifically, and provides a reliable measure to evaluate the importance of IR in cellular activities and diseased cellular states.

JUM presents a totally novel and statistically rigorous approach to address, evaluate, quantitate and classify the complex and diverse patterns of AS profiles in eukaryotic transcriptomes. We are confident that this new approach will provide new and important insights to the dynamic regulation of AS and gene expression. Our results indicate novel isoform detection, quantification and classification of transcripts from Drosophila head, mouse neurons and human cancer genome RNA samples. These initial applications already indicate that JUM will be the method of choice going forward to analyze complex pre-mRNA splicing patterns at the transcriptome-wide level, particularly in complex tissue types that are already known to generate extremely diverse mRNA isoform profiles, such as gonads (testes and ovaries), pluripotent stem cells and a variety of neuronal cell types and nervous system tissues. Finally, the JUM algorithm should be useful for detecting, mapping and analyzing the biogenesis of a novel pattern of non-coding RNAs that are circular in structure. These newly discovered RNA species are thought to play a role in microRNA regulation[39].

Online Methods

JUM package distribution

A user-friendly version of the JUM package has been deposited on GitHub. The codes are written in perl and bash shell scripts.

RNA-seq data

Raw RNA-seq data (FASTQ format) for mouse embryonic cortical neurons and Drosophila male fly heads described in the paper are derived as previously described[20,21]. Human colon tumor and matched normal tissue poly-A selected RNA-seq data (in BAM format) are acquired from the TCGA database. To avoid technical and sampling bias brought by factors other than the cancerous state, five male colon cancer patients and six female colon cancer patients are picked that have matched tumor and normal tissue samples sequenced with the same platform/center and sequencing read length, with vital state “alive”, tumor type “primary” and within a certain age range. A detailed description of the patient tumor and normal samples used in this study are listed in Supplemental Table 1. The downloaded BAM files are transformed back to FASTQ format by using the SamToFastq function in PICARD tools before analysis. The FASTQ data are then mapped to the human genome hg38 as described below. The sequencing read mapping results are summarized in Supplemental Table 3 for each patient.

RNA-seq data preparation for JUM

RNA-seq reads are mapped to the human (hg38), mouse (mm9) and Drosophila (dm3) genomes respectively using STAR[40] in the 2-pass mode, as instructed in the STAR manual. Only unique mapped reads are kept in the output for JUM analysis.

Junction filtering in JUM analysis

The current version of JUM filters for valid junctions to be included in the downstream analysis as those identified from all samples (all replicates for both control and treated samples, for example) with a minimum of a user-defined number of reads mapped to it. We used 5 reads for all the analysis presented in this paper.

Algorithm to construct AS patterns from profiled AS structures

We first profile all AS structures from the RNA-seq data and calculate the S_I value for each sub-AS-junction in these AS structures. Two AS structures are defined as "linked", if they share one specific sub-AS-junction, and a "path" is drawn between the two AS structures. Under this definition, a "loop" of AS structures are searched in the whole pool of AS structures, with every AS structure in the loop linked to one other by a path. Each profiled loop of AS structures is corresponding to an AS pattern, and is allocated to each AS pattern category based on the features of the sub-AS-junction S_I value distributions.

Acknowledgements

We thank Gideon Dreyfuss and Kate Abruzzi for helpful critiques. We thank Yeon Lee for help testing the user-friendly version of the JUM package. The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This work

was supported by NIH R01GM097352 and NIH R35GM118121 (D. Rio, PI) and by the NIH Center for RNA Systems Biology at U.C., Berkeley (P50GM102706; J. Cate, PI). Q.W. is supported by the Arnold O. Beckman Postdoctoral Fellowship. The authors declare no competing financial interests.

References

1 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).
2 Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457-463, doi:10.1038/nature08909 (2010).
3 Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-718, doi:10.1016/j.cell.2009.02.009 (2009).
4 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**, 689-701, doi:10.1038/nrg3778 (2014).
5 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-1415, doi:10.1038/ng.259 (2008).
6 Shkreta, L. *et al.* Cancer-Associated Perturbations in Alternative Pre-messenger RNA Splicing. *Cancer Treat Res* **158**, 41-94, doi:10.1007/978-3-642-31659-3_3 (2013).
7 Li, Y. *et al.* RNA splicing is a primary link between genetic variation and disease.pdf. *Science* **352**, 600-604, doi:10.1126/science.aad9417 (2016).
8 Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. RNA splicing factors as oncoproteins and tumour suppressors.pdf. *Nat Rev Cancer* **16**, 413-430, doi:0.1038/nrc.2016.51 (2016).
9 Taylor, J. P., Brown, R. H., Jr. & Cleveland, D. W. Decoding ALS: from genes to mechanism. *Nature* **539**, 197-206, doi:10.1038/nature20413 (2016).
10 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015, doi:10.1038/nmeth.1528 (2010).
11 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017, doi:10.1101/gr.133744.111 (2012).
12 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53, doi:10.1038/nbt.2450 (2013).
13 Hu, Y. *et al.* DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res* **41**, e39, doi:10.1093/nar/gks1026 (2013).
14 Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res* **21**, 193-202, doi:10.1101/gr.108662.110 (2011).

15 Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593-5601, doi:10.1073/pnas.1419161111 (2014).
16 Vitting-Seerup, K., Porse, B. T., Sandelin, A. & Waage, J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* **15**, 81, doi:10.1186/1471-2105-15-81 (2014).
17 Aschoff, M. *et al.* SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* **29**, 1141-1148, doi:10.1093/bioinformatics/btt101 (2013).
18 Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, e11752, doi:10.7554/eLife.11752 (2016).
19 Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448, doi:10.1038/nature13424 (2014).
20 Wang, Q., Moore, M. J., Adelmant, G., Marto, J. A. & Silver, P. A. PQBP1, a factor linked to intellectual disability, affects alternative splicing associated with neurite outgrowth. *Genes Dev* **27**, 615-626, doi:10.1101/gad.212308.112 (2013).
21 Wang, Q. *et al.* The PSI-U1 snRNP interaction regulates male mating behavior in Drosophila. *Proc Natl Acad Sci U S A* **113**, 5269-5274, doi:10.1073/pnas.1600936113 (2016).
22 Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660-1666, doi:10.1093/bioinformatics/btu077 (2014).
23 Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092, doi:10.1093/bioinformatics/bts094 (2012).
24 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 (2011).
25 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512, doi:10.1038/nprot.2013.084 (2013).
26 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
27 Lu, J., Tomfohr, J. K. & Kepler, T. B. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6**, 165, doi:10.1186/1471-2105-6-165 (2005).
28 Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881-2887, doi:10.1093/bioinformatics/btm453 (2007).
29 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
30 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
31 Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**, 321-332, doi:10.1093/biostatistics/kxm030 (2008).

32 McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288-4297, doi:10.1093/nar/gks042 (2012).

33 Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem* **84**, 291-323, doi:10.1146/annurev-biochem-060614-034316 (2015).

34 Majumdar, S. & Rio, D. C. P Transposable Elements in Drosophila and other Eukaryotic Organisms. *Microbiol Spectr* **3**, MDNA3-0004-2014, doi:10.1128/microbiolspec.MDNA3-0004-2014 (2015).

35 Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med* **7**, 45, doi:10.1186/s13073-015-0168-9 (2015).

36 Solana, J. *et al.* Conserved functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative splicing in planarians. *Elife* **5**, doi:10.7554/eLife.16797 (2016).

37 Baggerly, K. A., Deng, L., Morris, J. S. & Aldaz, C. M. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19**, 1477-1483 (2003).

38 Hansen, K. D., Wu, Z., Irizarry, R. A. & Leek, J. T. Sequencing technology does not eliminate biological variability. *Nat Biotechnol* **29**, 572-573, doi:10.1038/nbt.1910 (2011).

39 Panda, A. C., Grammatikakis, I., Munk, R., Gorospe, M. & Abdelmohsen, K. Emerging roles and context of circular RNAs. *Wiley Interdiscip Rev RNA*, doi:10.1002/wrna.1386 (2016).

40 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

41 Doherty, P., Moolenaar, C. E., Ashton, S. V., Michalides, R. J. & Walsh, F. S. The VASE exon downregulates the neurite growth-promoting activity of NCAM 140. *Nature* **356**, 791-793, doi:10.1038/356791a0 (1992).

Figure legends

Figure 1. Software that relies on pre-annotated AS splice junction libraries cannot detect or accurately quantitate novel AS events or distinct AS patterns specific to different tissue samples. (a) A novel and functionally important AS event of the *Ncam1* gene transcript in mouse embryonic cortical neurons, that is not included in publically available mouse transcriptome annotation. Genomic structure of Ncam1 is shown. Blue rectangles are exons and lines between exons introns. Genomic annotation on top is from the RefSsq database, while the annotation on the bottom is the actual transcriptome profile inferred from RNA-seq data of mouse embryonic cortical neurons[20]. RNA-seq tracks from control and splicing factor PQBP1 knockdown samples are shown in the middle, with red arc between exons representing level of sequence reads mapped to the corresponding splice junction. The height of an arc shows the relative depth of read counts mapped to that splice junction. The VASE exon is a novel alternatively spliced cassette exon[41]. The VASE exon-included isoform encodes a Ncam1 protein that inhibits neurite outgrowth, while the VASE exon-excluded isoform encodes a Ncam1 protein that promotes neurite outgrowth[41]. The shRNA knockdown of splicing factor PQBP1 results in elevated inclusion of the VASE exon-containing isoform and is linked to the dendritic outgrowth defects observed in PQBP1 knockdown neurons[20]. AS analysis methods that depend on pre-built AS annotation libraries missed this crucial exon that can be targeted by splicing factors to change the function of Ncam1 and neurite outgrowth in neurons[20]. (b) Two distinct AS patterns of the Drosophila *fruitless* gene mRNAs expressed in male Drosophila head tissue and the Drosophila Schineider-2 (S2) cell line[21]. RNA-seq data read density tracks derived from both tissue types are shown, with arcs representing splice junctions that link a common 5' exon to the three alternative last exons, each corresponding to the *fruitless* isoform fru-A, fru-B and fru-C, respectively. The

relative levels of the fru-A, fru-B and fru-C isoforms determine normal male fly courtship behavior[21]. In Drosophila male heads, all three isoforms are present. However, in the Drosophila S2 tissue culture cell line only fru-B and fru-C mRNA isoforms are expressed, together with an additional isoform (fru-i) that uses an alternative polyadenylation signal downstream of the common 5' exon present in the fru-B and fru-C mRNA isoforms. AS analysis software tools that rely on a fixed annotated AS splice junction library can not detect and accurately quantitate the distinct fruitless mRNA isoform distributions present in these two different types of Drosophila RNA samples.

Figure 2. JUM exclusively uses and quantitates sequence reads mapped to splice junctions and defines AS structures as the basic quantification unit for global AS pattern analysis. (a) JUM uses RNA-seq reads mapped to splice junctions for AS quantification. Green rectangles indicate exons and lines introns. Green and blue short lines represent reads that mapped to splice junctions connecting exons, which are the most direct evidence for the existence and quantitative assessment of a given splice junction. JUM defines the start coordinate of a splice junction as the 5' initiation site (5'IS) and the end coordinate of a splice junction as the 3' ending site (3'ES). An AS "structure" is defined as a set of junctions that share the same 5'IS or the same 3'ES. Each splice junction in an AS structure is defined as a sub-AS-junction. (b,c,d) AS structures are the basic element that comprise all conventionally recognized AS patterns.

Figure 3. JUM performs differential AS analysis by evaluating the differential usage of each sub-AS-junction found in AS structures. (a) The flow chart of the procedure JUM uses to perform and quantitate differential AS analysis. (b) JUM models the sequence reads that map to

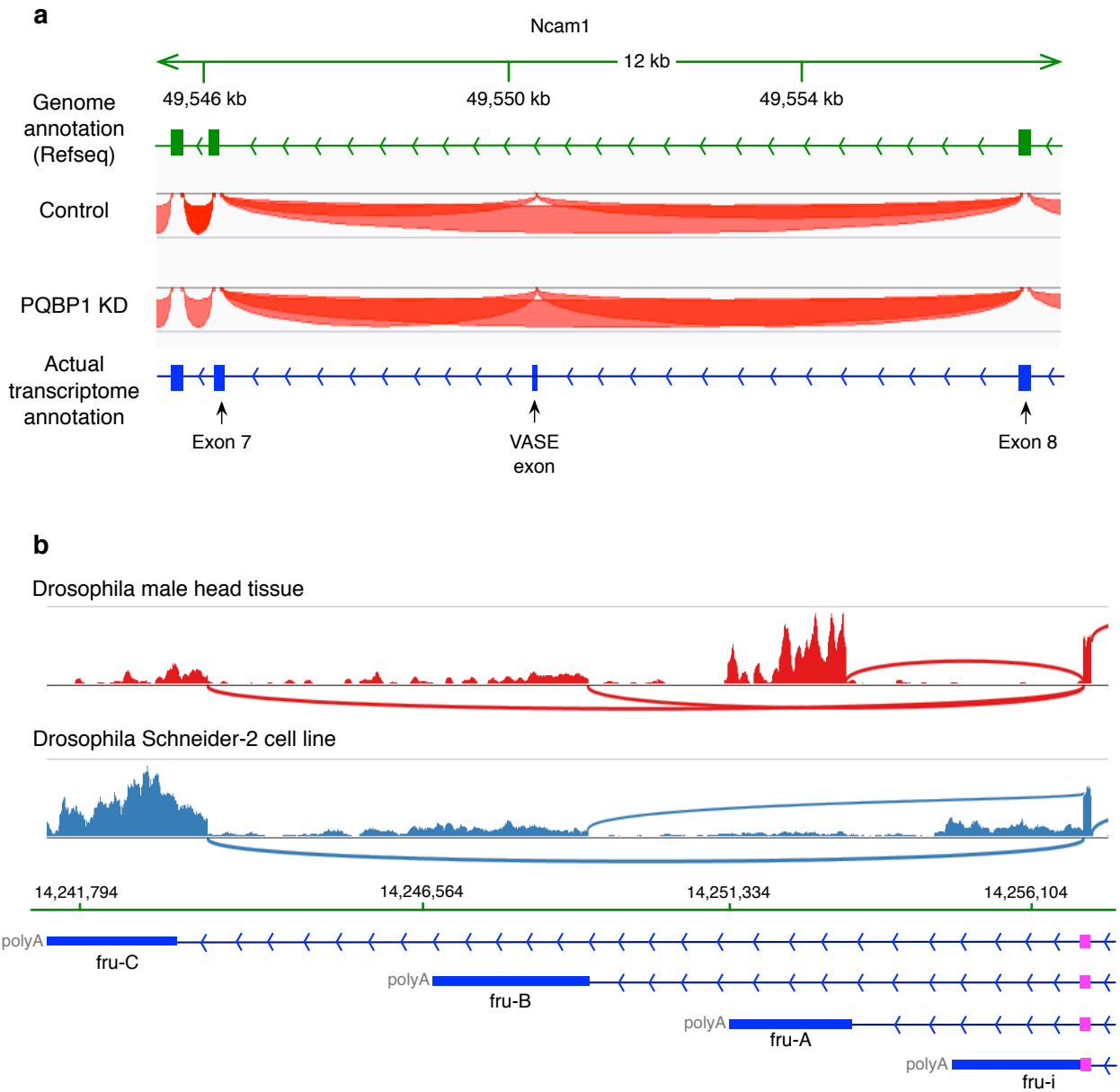
a sub-AS-junction as negative binomial distribution. (c) JUM fits two generalized linear models to evaluate the influence of a given biological condition on the usage of a specific sub-AS-junction.

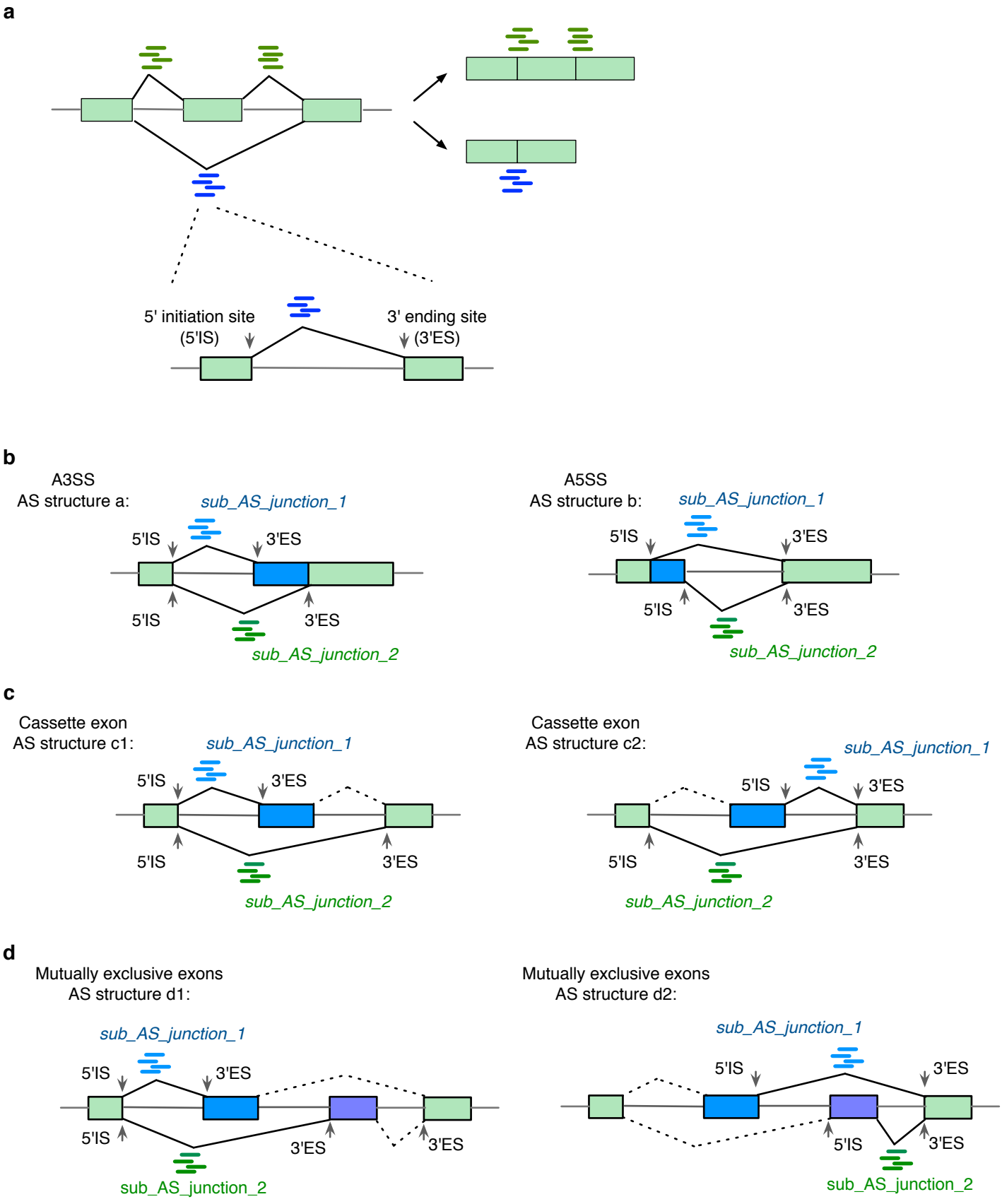
Figure 4. JUM constructs AS structures into conventionally recognized categories of AS patterns based on the unique topological features of each AS pattern type. (a, b, c, d) The topological features of AS patterns: Cassette exons (a), mutually exclusive exons (b), alternative 5' splice sites (c), and alternative 3' splice sites (d) represented by graphs and the frequency parameter S_I of sub-AS-junctions. (e) JUM defines an additional AS pattern category—the "composite AS" which is a more complex combination of several conventionally recognized AS patterns. An example for such a complex "composite" AS pattern is shown in the bottom panel for the eIF-4E gene transcripts found in Drosophila male head tissue RNA-seq samples[21]. Arcs represent splice junctions that connect different exons.

Figure 5. JUM applies stringent criteria for detection, quantitation and analysis of intron retention events that dramatically reduces false positive rates compared to software relying on pre-annotated splice junction libraries. (a) The most commonly used quantification parameters for intron retention. RNA-seq reads spanning intron-exon or exon-intron boundaries are represented by short green or blue lines, respectively. Short purple lines represent sequence reads mapped to intronic regions. Short red lines represent sequence reads mapped to the splice junction to the corresponding intron. (b,c,d) The commonly applied strategies in AS analysis software mis-classify other AS patterns as intron retention. Three MISO-reported[10] significantly changed intron retention events were shown that actually correspond to mutually exclusive exons

splicing events (b), alternative promoters (c), cassette exons mixed with alternative 3' splice sites (d) from Drosophila male head tissue in a comparison of control wildtype fly strain and a transgenic fly strain that expresses the truncated PSI protein[20]. The start and end points of the retained intron events reported by MISO are denoted by red arrows. Arcs represent splice junctions identified from the RNA-seq data. Exon coverage from RNA-seq data is also shown in blue. (e) The three criteria that JUM uses to analyze intron retention, in order to reduce false positive intron retention calls. Short blue and green lines represent reads mapped to the exon-intron or intron-exon boundaries, respectively. Short red lines represent sequence reads mapped to the splice junctions. Short purple reads represent sequence reads mapped to the intronic regions and are required to be approximately uniformly distributed all across the entire intronic region of the retained intron.

Figure 6. Comparison of JUM performance to rMATs and MISO. (a) Comparison of JUM, rMATS and MISO in detecting significantly changed, male courtship defect phenotype-associated AS events in Drosophila male head RNA-seq samples. (b) Comparison of JUM and rMATS in detecting rMATS-identified top 30 most significantly changed CE events between wildtype Drosophila head sample and a male courtship defect strain.


a
Ncam1
12 kb
49,546 kb
49,550 kb
49,554 kb
Genome annotation (Refseq)
Control
PQBP1 KD
Actual transcriptome annotation
Exon 7
VASE exon
Exon 8
b
Drosophila male head tissue
Drosophila Schneider-2 cell line
14,241,794
14,246,564
14,251,334
14,256,104
polyA
fru-C
polyA
fru-B
polyA
fru-A
polyA
fru-i


a
5' initiation site
(5'IS)
3' ending site
(3'ES)
b
A3SS
AS structure a:
sub_AS_junction_1
5'IS
3'ES
5'IS
3'ES
sub_AS_junction_2
A5SS
AS structure b:
sub_AS_junction_1
5'IS
3'ES
5'IS
3'ES
sub_AS_junction_2
c
Cassette exon
AS structure c1:
sub_AS_junction_1
5'IS
3'ES
5'IS
3'ES
sub_AS_junction_2
Cassette exon
AS structure c2:
sub_AS_junction_1
5'IS
3'ES
5'IS
3'ES
sub_AS_junction_2
d
Mutually exclusive exons
AS structure d1:
sub_AS_junction_1
5'IS
3'ES
5'IS
3'ES
3'ES
sub_AS_junction_2
Mutually exclusive exons
AS structure d2:
sub_AS_junction_1
5'IS
3'ES
5'IS
3'ES
sub_AS_junction_2

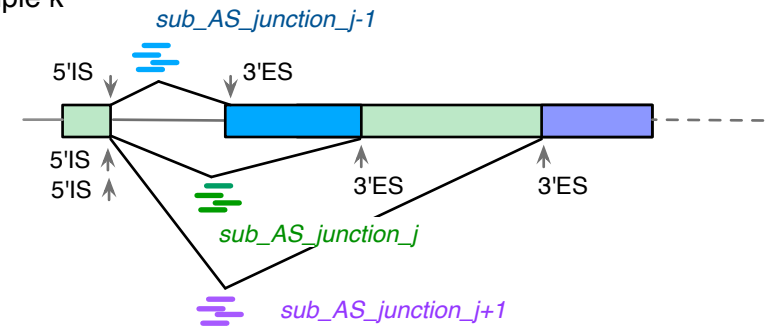
a

Profile all AS structures from RNA-seq samples → Count # of reads that are mapped to each sub-AS-junctions and calculate the "usage" of each sub-AS-junction of all AS structures → Identify AS structures where the "usage" of a sub-AS-junction is significantly changed under a biological condition

b

RNA-Seq sample k

AS structure *i*



Model reads mapped to sub_AS_junction_*j* as negative binomial distribution:

$$Y_{ijk} \sim NB(\mu_{ijk}, \phi_{ij}) \quad \text{Eq. 1}$$

$$Var(Y_{ijk}) = \mu_{ijk}(1 + \mu_{ijk}\phi_{ij}) \quad \text{Eq. 2}$$

c

Fitting and test two generalized linear models for each sub-AS-junction to evaluate if a biological condition significantly changes the usage of a sub-AS-junction in an AS structure

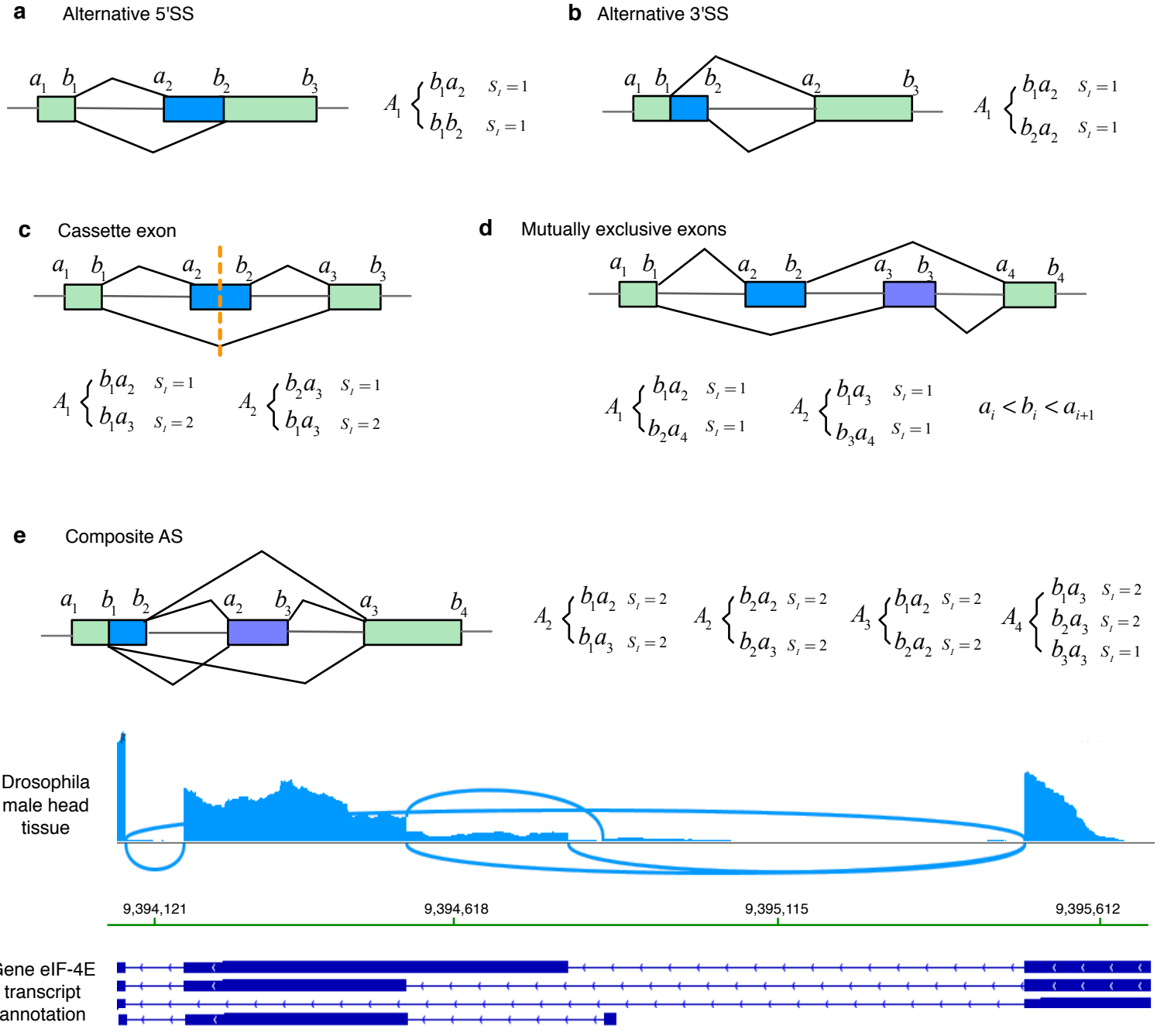
The effect model

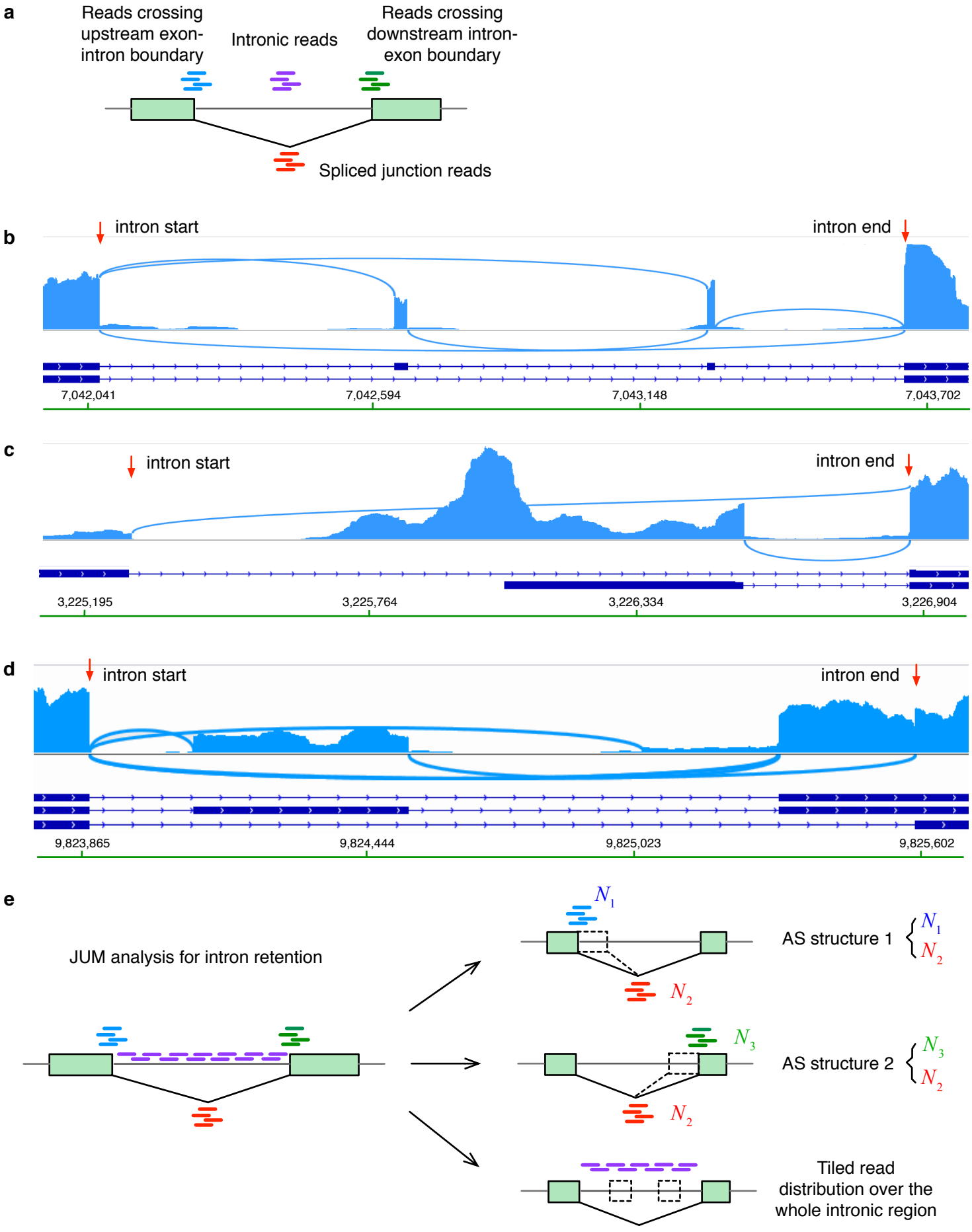
$$\log \mu_{ijk} = \alpha_i^S + \alpha_{ij}^E + \alpha_{i\varepsilon_k}^C + \boxed{\alpha_{ij\varepsilon_k}^C} \quad \text{Eq. 3}$$

The basal model

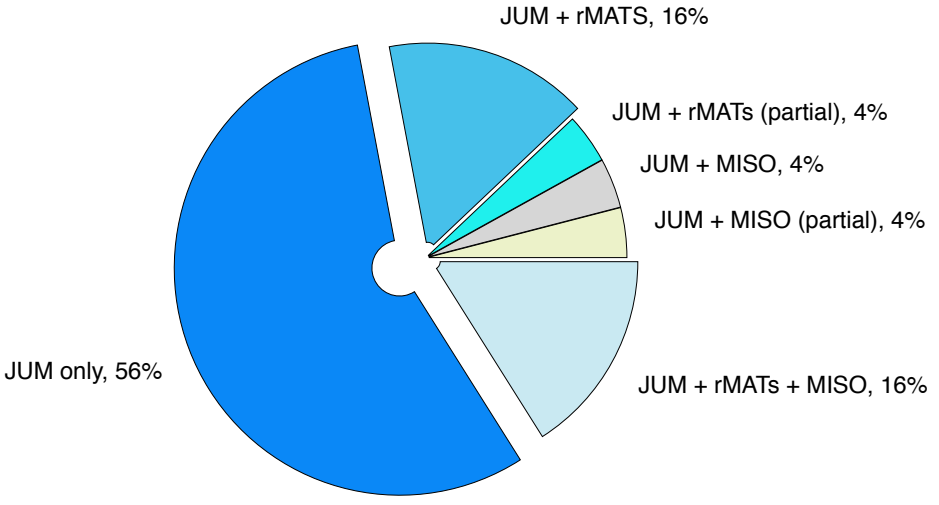
$$\log \mu_{ijk} = \alpha_i^S + \alpha_{ij}^E + \alpha_{i\varepsilon_k}^C \quad \text{Eq. 4}$$

$\alpha_{ij\varepsilon_k}^C$: The effect that biological condition ε imposes on reads mapped to sub_AS_junction_*j* of AS structure *i*

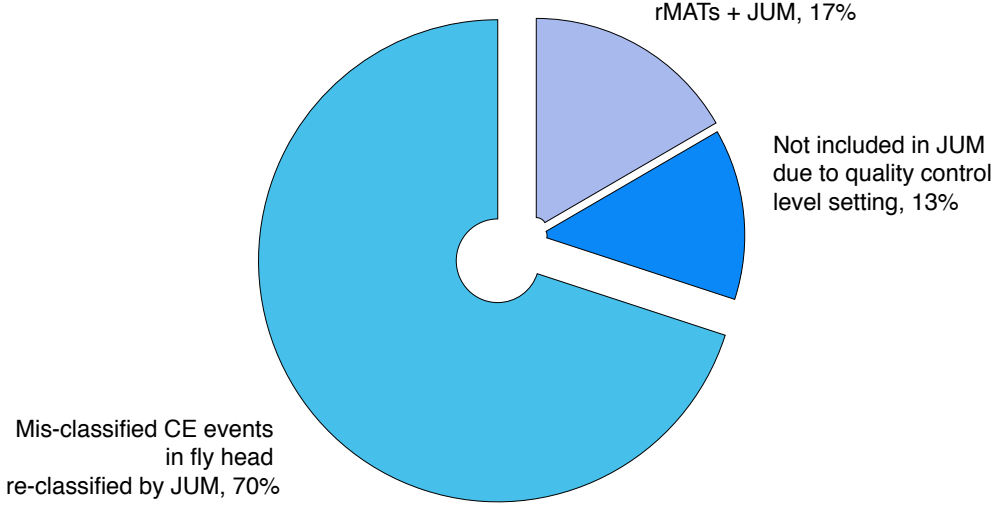

a
Alternative 5'SS
b
Alternative 3'SS
c
Cassette exon
d
Mutually exclusive exons
e
Composite AS
Drosophila male head tissue
9,394,121
9,394,618
9,395,115
9,395,612
Gene eIF-4E transcript annotation


a
Reads crossing upstream exon-intron boundary
Intronic reads
Reads crossing downstream intron-exon boundary
Spliced junction reads
b
intron start
intron end
7,042,041
7,042,594
7,043,148
7,043,702
c
intron start
intron end
3,225,195
3,225,764
3,226,334
3,226,904
d
intron start
intron end
9,823,865
9,824,444
9,825,023
9,825,602
e
JUM analysis for intron retention
N_1
N_2
AS structure 1
N_1
N_2
N_3
N_2
AS structure 2
N_3
N_2
Tiled read distribution over the whole intronic region

a


JUM + rMATS, 16%
JUM + rMATs (partial), 4%
JUM + MISO, 4%
JUM + MISO (partial), 4%
JUM + rMATs + MISO, 16%
JUM only, 56%

b


rMATs + JUM, 17%
Not included in JUM due to quality control level setting, 13%
Mis-classified CE events in fly head re-classified by JUM, 70%