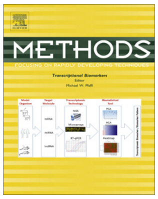
Methods xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



In silico to *in vivo* splicing analysis using splicing code models

Matthew R. Gazzara [a,b], Jorge Vaquero-Garcia [a,c], Kristen W. Lynch [b], Yoseph Barash [a,c,*]

[a] *Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[b] *Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[c] *Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Alternative splicing
Splicing code
BIN1
AVISPA
Centronuclear myopathy (CNM)
Myotonic dystrophy (DM)

ABSTRACT

With the growing appreciation of RNA splicing's role in gene regulation, development, and disease, researchers from diverse fields find themselves investigating exons of interest. Commonly, researchers are interested in knowing if an exon is alternatively spliced, if it is differentially included in specific tissues or in developmental stages, and what regulatory elements control its inclusion. An important step towards the ability to perform such analysis *in silico* was made with the development of computational splicing code models. Aimed as a practical how-to guide, we demonstrate how researchers can now use these code models to analyze a gene of interest, focusing on *Bin1* as a case study. Bridging integrator 1 (BIN1) is a nucleocytoplasmic adaptor protein known to be functionally regulated through alternative splicing in a tissue-specific manner. Specific *Bin1* isoforms have been associated with muscular diseases and cancers, making the study of its splicing regulation of wide interest. Using AVISPA, a recently released web tool based on splicing code models, we show that many *Bin1* tissue-dependent isoforms are correctly predicted, along with many of its known regulators. We review the best practices and constraints of using the tool, demonstrate how AVISPA is used to generate high confidence novel regulatory hypotheses, and experimentally validate predicted regulators of *Bin1* alternative splicing.

1. Introduction

Splicing, the removal of introns and precise joining of exons, is an essential step in the biogenesis of mature eukaryotic mRNA. High-throughput studies across multiple tissues show that the inclusion of exonic regions in mature mRNAs can greatly vary, with ~95% of human multi-exon genes undergoing alternative splicing (AS) [1,2]. The increased complexity of the transcriptome has several consequences. First, it serves to expand the proteome by allowing the same gene to produce mRNA isoforms that differ in coding sequence [3]. Additionally, AS can influence the fate of mRNA transcripts, either by the introduction of premature stop codons, which marks the transcript for nonsense-mediated decay [4], or by altering untranslated regions, which influences the presence of elements involved in transcript stability, translation efficiency, and localization [5]. Recently, AS was also shown to regulate the biogenesis of miRNAs that span exon–intron boundaries in primary transcripts [6]. Highlighting the importance of splicing and its regulation, studies estimate that anywhere from 15 to 50 percent of disease-causing mutations affect splicing [7].

Alternative splicing's key role in post-transcriptional control of gene expression and its pervasiveness motivated much work to elucidate the mechanisms of AS regulation. Besides identifying spliceosome components and their interactions with the core splicing signals, decades of research resulted in the identification of many *cis*- and *trans*-acting elements involved in pre-mRNA splicing [for reviews see 8,9]. These include features such as splice site strength [10], local secondary structure [11], and splicing regulatory elements (SREs) which interact with RNA binding proteins (RBPs) to enhance or repress exon inclusion [12].

The role of AS in gene expression and disease state has also led to much interest in the broader community in mapping AS regulatory elements controlling exons of interest. Researchers became interested in identifying splicing defects due to mutations, tracing putative regulators, and understanding how exon inclusion levels change across cellular conditions. Consequently, tools were developed to identify some of the elements affecting splicing outcome. For example, a number of tools were created that search for splice sites and branch points and score how well they bind core spliceosomal components [13–15]. Tools are also available for basic motif searches for putative SREs or RBP binding sites [16,17], and some allow for scoring of core splicing signals as well [18,19].

Abbreviations: AS, alternative splicing; AVISPA, advanced visualization of splicing prediction analysis; BIN1, bridging integrator 1; CNM, centronuclear myopathy; CNS, central nervous system; DM, myotonic dystrophy; N2a, Neuro2-a; NFE, normalized feature effect; PTB, polypyrimidine tract binding protein 1; QKI, Quaking homolog KH domain RNA binding protein; RBP, RNA binding protein; SRE, splicing regulatory element.

* Corresponding author at: Department of Genetics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA. Fax: +1 (215) 573 3111.

E-mail address: yosephb@mail.med.upenn.edu (Y. Barash).

1046-2023/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2013.11.006

The decades of research into splicing regulation revealed splicing to be a highly complex process, involving many regulatory elements interacting in a context specific manner. This observation motivated researchers to move from descriptive tools that give a "parts list" to a predictive splicing "code", as a set of probabilistic rules that would predict splicing outcome directly from genomic sequence, given the cellular context [12]. Consequently, machine learning techniques were applied to high throughput, exonic level, expression data to develop such probabilistic code models [20]. Using over a thousand putative regulatory elements such as sequence motifs, RNA structure, and conservation, these algorithms were able to give accurate predictions for changes in exon inclusion levels across four main mouse tissue groups: central nervous system (CNS), muscle, digestive, and embryonic vs. adult tissues. Briefly, given a putative alternative cassette exon, these algorithms first compute the values for the many putative regulatory features extracted from the exon and its flanking regions depicted in Fig. 1a. They then use these values along with the cellular context (e.g., CNS tissue) to predict the splicing outcome (e.g., "increased exon inclusion in the brain"). These new algorithms thus offered a framework for performing predictive splicing analysis. Indeed, initial work demonstrated the splicing code models were able to identify novel splicing changes affecting functional domains in disease-associated genes, recapitulate much of the previous results about regulatory elements, and identify novel regulatory elements that were experimentally verified [20]. Follow up research also demonstrated that splicing code models originally derived for mouse were successfully applied to human, chicken and frog with a high degree of overlap in underlying regulatory features [21]. However, these works were limited to sets of previously identified alternative cassette exons, and thus could not offer researchers splicing analysis for general usage.

AVISPA (http://avispa.biociphers.org) is a recently released web tool aimed to make splicing code models accessible for general usage [22]. Implemented as a Galaxy server to support iterative updates of new datasets and improved models [23], AVISPA is designed as a user friendly front end to splicing code models. It allows researchers to analyze an exon of interest to get predictions of whether the exon is alternatively spliced, whether it is expected to exhibit tissue-dependent inclusion, and to identify putative regulatory elements. Fig. 1b provides a brief overview of AVISPA's input, processing, and output. The user's query exon of interest is first mapped to the genome and a set of RNA related features is extracted. Next, an ensemble of over 5000 splicing code models is used to derive predictions of the splicing outcome. The exon is first scored for how likely it is to be an alternatively spliced cassette exon. If confidence in alternative splicing passes a significance threshold, the exon's feature set is then run through a second set of models that predict whether the exon will be differentially included in specific tissues. Importantly, beyond the prediction of splicing outcome, AVISPA also reports the query's feature enrichment and the effects of the *in silico* removal of regulatory motifs on the splicing prediction [22]. In all, AVISPA offers those without a computational background, or even those outside of the splicing field, the ability to interrogate current splicing code models to gain insights on the splicing profile and regulation of exons in genes of interest.

This paper serves as a "how-to guide" for *in silico* splicing analysis, focusing on how to use the AVISPA web tool. Before delving into analysis details, it is important for potential users to first note some limitations of AVISPA's current implementation. First, AVISPA does not predict whole transcript structure but rather local changes in exon inclusion levels under different conditions. Second, it only supports cassette exons. While cassette exons are the most common form of alternative splicing in mammals [1], many other forms are known, such as 3′ and 5′ splice site variations, but are not yet supported. Third, AVISPA only supports predictions for differential splicing in the four main tissue groups listed before. Finally, it does not predict absolute exon inclusion levels. This means that instead of predicting, for example, "40% exon inclusion in brain and 20% in most other tissues" it offers predictions for "increased inclusion in the brain". Other more technical constraints of AVISPA's implementation are discussed as part of the analysis case described below. Addressing these limitations is an ongoing effort. Nonetheless, as we illustrate below, AVISPA can be effectively applied for *in silico* splicing analysis of genes of interest. Moreover, it is important to note that none of the limitations described above are inherent to *in silico* splicing analysis, and thus can be expected to be improved upon as updates are introduced into AVISPA.

Here we illustrate how one can use AVISPA to carry out *in silico* splicing analysis on a gene of interest, *Bin1*. Bridging integrator 1 (BIN1) is a nucleocytoplasmic adaptor protein known to be functionally regulated through alternative splicing in a tissue-specific manner [24]. The mouse version of *Bin1* is similar in structure and organization to the human gene and both play a role in muscle cell differentiation [25,26]. Moreover, a muscle specific isoform is essential for membrane curvature and T-tubule biogenesis in skeletal muscle and splicing misregulation of this exon has been associated with the muscle disorders myotonic dystrophy (DM) and centronuclear myopathy (CNM) [27,28]. Additionally, *BIN1* has

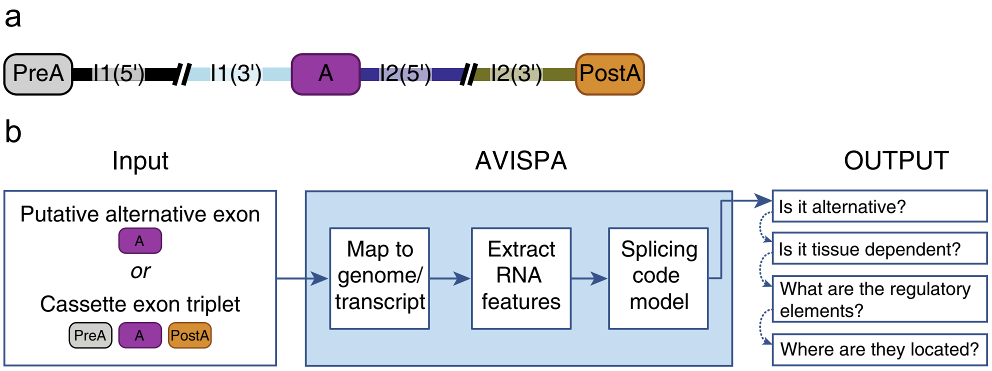


Fig. 1. Overview of AVISPA. (a) The genomic regions of a cassette exon triplet used for analysis. These include the putative alternative exon ("A"), the two flanking exons upstream ("PreA") and downstream ("PostA"), and the intronic regions proximal to these exons. (b) Queries are input as a single exon or as a triplet (preferred) and first mapped to a transcript database or the genome. A matched query has RNA features extracted that are then run through an ensemble of splicing codes to predict if the exon is alternative and if it is differentially spliced in certain tissues. Graphical summary files indicate what are the putative regulatory elements controlling the exon's inclusion, and where they are located.

Please cite this article in press as: M.R. Gazzara et al., Methods (2013), http://dx.doi.org/10.1016/j.ymeth.2013.11.006

features of a tumor suppressor and *BIN1* missplicing is associated with many human cancers due to a loss of its inhibitory interaction with the oncogenic transcription factor, Myc [29,30]. Fig. 2 illustrates some of the more well described splicing patterns of *BIN1* and select protein isoforms analyzed in this paper. The figure also helps to illustrate some of the limitations of AVISPA as prediction for the complex alternative splicing event involving exons 13–16 is currently not supported. Nonetheless, the tissue-specific patterns and disease association of this gene make a detailed understanding of its splicing regulation particularly useful.

While some splicing regulatory elements of the exons of this gene have been described, it is likely that the full picture is far from complete. For example, alternative splicing of exons 7 and 13 of *Bin1* were only recently implicated in a network of exons regulated by the splicing factors Quaking (QKI) and polypyrimidine tract-binding protein (PTB) in myoblast cells [31]. Using a splicing event centered on exon 13 as a case study (hereafter, "triplet 12.13.17", Fig. 2), we describe how to use AVISPA and accurately interpret the output, while highlighting some current limitations and precautions that should be considered. Extending this analysis to the other exons of *Bin1*, we rediscover experimentally verified features of disease-associated alternative splicing events and suggest high-confidence predictions for novel regulatory effects. Finally, we experimentally validate predicted CNS splicing regulators *in vivo* using the mouse Neuro2-a (N2a) neuroblastoma cell line.

2. Methods

2.1. In silico splicing analysis using AVISPA

The event coordinates provided by Hall et al. [31] were used to identify the query exon as *Bin1* exon 13 flanked by exons 12 and 17. Coordinates for this exon and its flanking regions were downloaded from the UCSC genome browser and input into AVISPA. Running an exon prediction task, or a query, in AVISPA requires first uploading the query as genomic coordinates (BED6 format) or sequence (FASTA format). A query consists of either the putative cassette exon of interest ("A") alone or a cassette exon triplet, which specifies both the exon upstream ("C1" or "PreA") and downstream ("C2" or "PostA") of the query exon (Fig. 1a). If an exon triplet is specified, the name for each element must follow a specific format with the region added as a suffix (e.g., "Bin1_12.13.17_C2" or "Bin1_12.13.17_PostA"). Since AVISPA's analysis is based on the mm10 mouse genome assembly, the original data files of Hall et al. [31] were converted using the Lift-Over tool integrated within AVISPA's Galaxy server.

When considering the query event one wishes to analyze, it is important to keep in mind certain length restrictions that could lead to an error. AVISPA does not analyze triplets containing micro exons less than 10 nucleotides (nt) long, exons larger than 5000 nt, or introns less than 25 nt. Additionally, only the exon proximal portions of introns are analyzed (i.e., 300 nt from each exon/intron boundary). For example, while triplet 12.13.17 does not violate any length restrictions, the region between exon 13 and 17 is quite large (>5300 nt) and spans multiple introns and exons. This means only the first 300 nt of the intron downstream of exon 13 (the I2(5′) region) and the last 300 nt of the intron upstream of exon 17 (the I2(3′) region) are analyzed. While this region normally captures the vast majority of known splicing regulatory elements, any regulatory features located deep within these introns are not considered in the predictions and will not be reported. In contrast, if an intron is shorter than 600 nt the entire intron is analyzed and features that fall in the overlap of two intronic regions are analyzed for their effects in both regions.

It is also crucial to avoid misuse of query matching that might lead to spurious results. After submitting, a query is first matched to an internal database (DB) of known cassette exons, a DB of known transcripts, or the reference genome, in that order [22]. Thus, specifying an entire exon triplet in coordinates is the safest option to produce a mapping, as the coordinates specify the exact

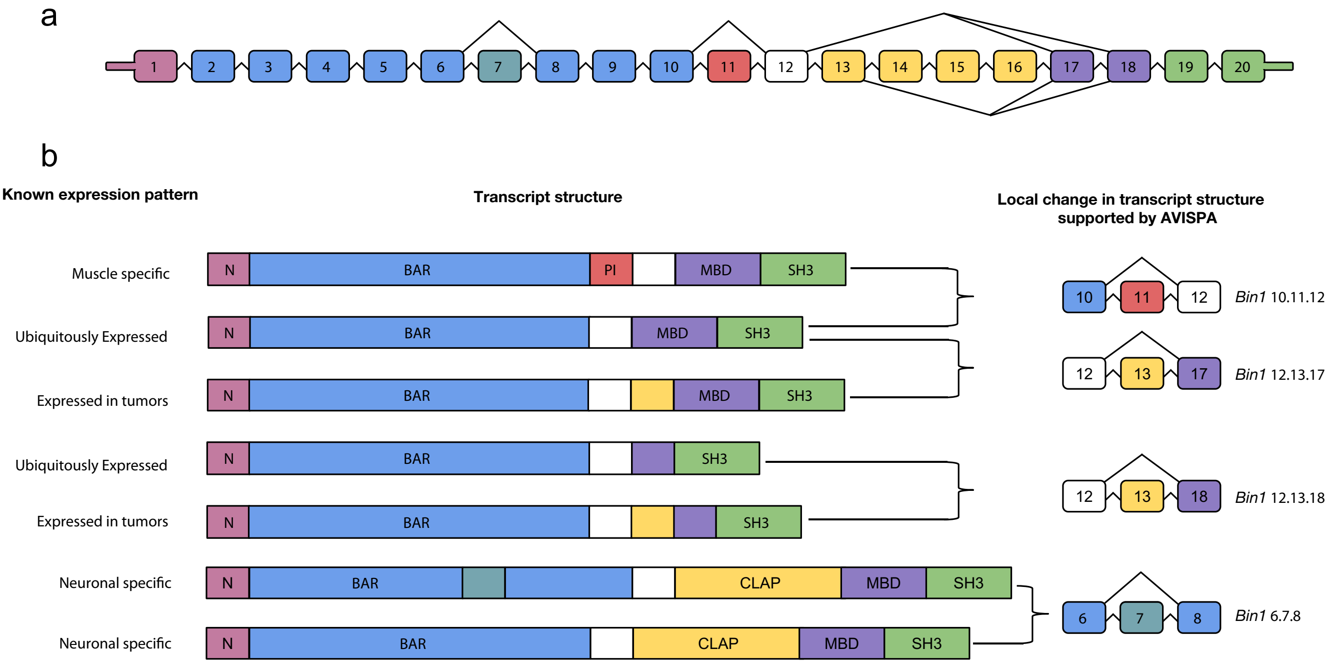


Fig. 2. *BIN1* alternative splicing. (a) A splice graph for several human *BIN1* alternative splicing events that are conserved in mouse. (b) Specific isoforms with their associated expression pattern (left) and local transcript variation analyzed using AVISPA (right). Notice the isoforms at the bottom include a complex, multi-exon alternative splicing event that is not modeled in AVISPA. BIN1 protein domains include: N,N-terminal amphipathic; BAR, BIN1-amphysin-Rvs167 domain; PI, phosphoinositide-binding domain; MBD, Myc-binding domain; SH3, Src homology 3 domain (adapted from Fugier et al. [27]).

location in the genome. By contrast, submitting a single exon query may be simpler for users to execute, but has the potential to match multiple events in AVISPA's DB (e.g., the query exon being present in transcripts with different flanking exons, as in the case of *Bin1* triplets 12.13.17 and 12.13.18). In such cases AVISPA will inform the user that the query exon has multiple matches in the DB and offer to upload BED files for these matches that the user could then execute. It is important to note that AVISPA always uses the *matched* sequence in its DB. This allows users to submit even a partial sequence of a single exon, and still get AVISPA to match it and perform splicing analysis for it. Another consequence of using the matched sequence is that incorrectly defined exon boundaries and SNVs will be ignored. Fig. 3a shows AVISPA's output mapping a sequence matched to *Bin1* triplet 12.13.17. In this case, the user submission included a wrong exonic start position and a fabricated SNP mismatch, highlighted in green and red, respectively (Fig. 3a). However, in some cases users may have *bona fide* alternative 3′ or 5′ splice site definitions for some of the query's exons that do not match the ones in AVISPA's DB. In such cases, users can submit the triplet coordinates and check the "use coordinates as is" option. Using this option, the specified coordinates are matched directly to the genome, ignoring the known splice sites. The output will be based on the user defined coordinates, though the DB matched coordinates are reported as well. Care should be taken when using triplet coordinates to specify a query. Using this option allows users to submit genomic regions such as introns and untranslated regions without any error indication. While useful for exploring unannotated exons, misuse of this feature may produce highly inaccurate or meaningless output since AVISPA was not trained on these types of sequences.

Once a query's information has been uploaded, the user has several options for the splicing analysis execution. Generally, AVISPA's pipeline involves two prediction stages. The first stage determines whether the query exon is alternatively or constitutively spliced, and the second determines whether the exon is differentially included in specific tissues [22]. Before execution, the user can specify the significance threshold for each prediction stage by using either a false positive rate (FPR) or relative rank value, calculated by comparing the query to a labeled set of exons the code model was trained on. It is important to note that if a query does not pass the given threshold for the first alternative versus constitutive prediction step, the tissue-specific predictions are not executed in order to save computational costs. However, if there is experimental evidence for the query exon being alternative, the user may specify this in a checkbox during submission. Selecting this option sets a permissive threshold (FPR = 0.5) for the first stage prediction, resulting in over a 90% chance that a known alternative exon will proceed to the second, tissue-dependent splicing analysis.

After execution, AVISPA's graphical summary of the *in silico* splicing analysis is available either through the web tool's interface, or as downloadable html files. The summary files include splicing prediction confidence, assessments of enriched regulatory features in the query, and evaluations of the *in silico* removal of regulatory *cis* elements on predictions. Extracts from these summary files for two of *Bin1* splicing events are discussed in the results section.

2.2. In vivo experimental validation

2.2.1. Cell culture and transfection

For *in vivo* validation of AVISPA's CNS predictions, Neuro-2a (N2a) cells (ATCC, CCL-131) were cultured according to the manufacturer's recommendations in DMEM (Cellgro 10-013) supplemented with 10% heat-inactivated fetal bovine serum (FBS) (Gibco 16000-044). To deplete splicing factors that AVISPA predicts to be relevant to our events, cells were transiently transfected using Lipofectamine 2000 (Life Technologies 11668019), according to manufacturer's recommendations, with siRNA targeting *Ptbp1*, *Qk*, or green fluorescent protein (GFP) control (Thermo Scientific Dharmacon M-042865-01, M-042676-01, or P-002048-01, respectively). Protein depletion was confirmed from whole cell lysates by Western blot using antibodies for QKI (Bethyl Laboratories, Inc. A300-183A) or PTB (Calbiochem NA63), with hnRNP L (Abcam ab6106) as a loading control.

2.2.2. RT-PCR analysis

RNA was isolated 48 h following transfection using RNA-Bee (Tel-Test, Inc. Cs-105B) following the manufacturer's protocol. Reverse transcription-PCR (RT-PCR) was performed as previously described in detail [32] using sequence specific primers for *Bin1* triplet 12.13.17 (Forward: 5′-GCTGCTACCCCTGAGATCAGAGTG; Reverse: 5′-GTTGCTTCACTGGCTGCTGTTCCC) and triplet 6.7.8 (Forward: 5′-AGCTGGTGGACTATGACAGTGCCC; Reverse: 5′-CGCG ATGCTCTGGAACGTGTTGAC). Gels were quantified by densitometry through the use of a Typhoon PhosphorImager (Amersham Biosciences). Percent inclusion was calculated as the percent of isoforms including the variable exon over the total isoforms relevant for each exon triplet analyzed.

3. Results and discussion

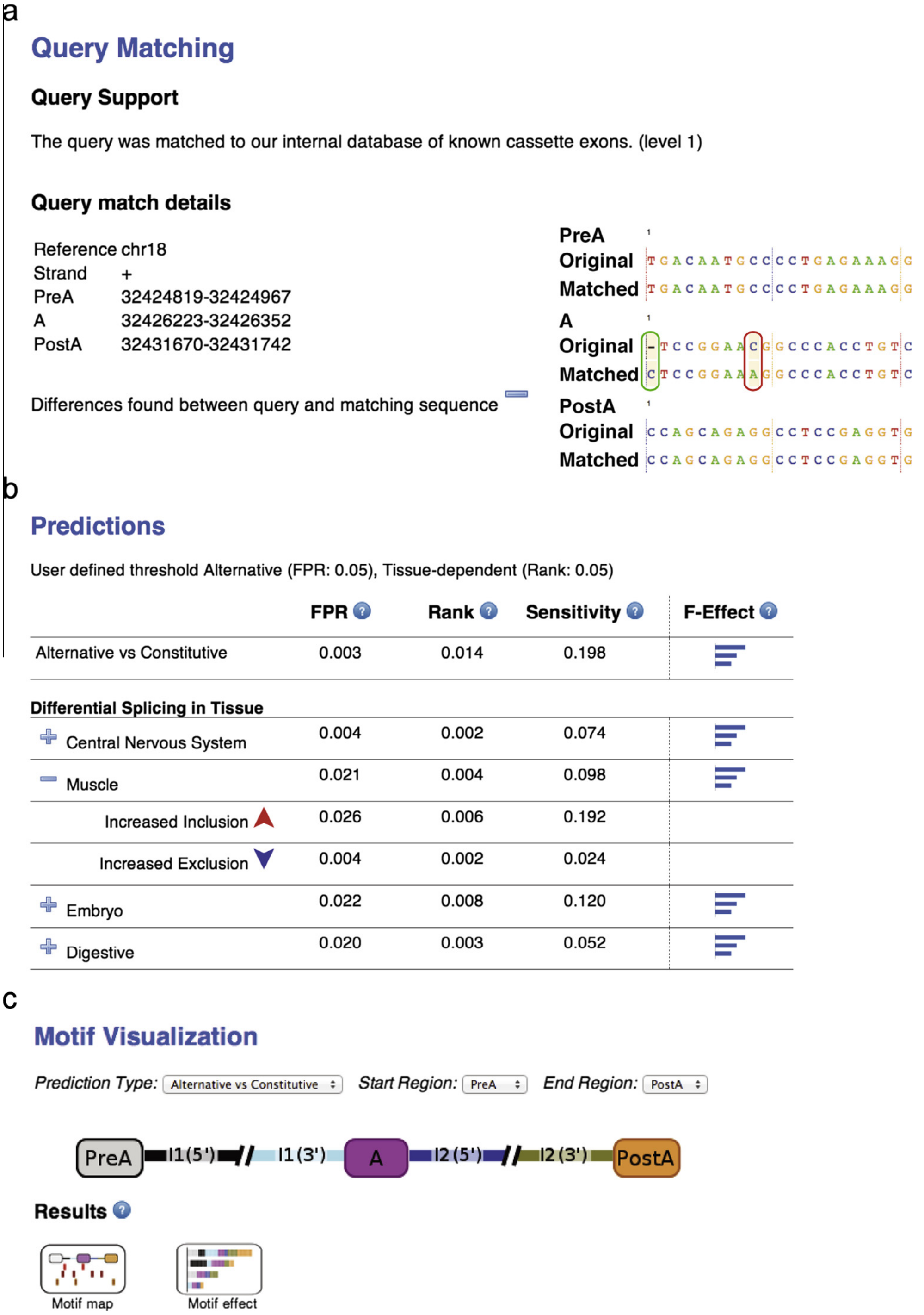
A query for *Bin1* triplet 12.13.17 was uploaded in the form of a triplet BED as described above and AVISPA was executed with the default parameters (FPR = 0.05 for alternative vs. constitutive splicing and rank value = 0.05 for tissue-specific predictions). To gain further insights on *Bin1* splicing patterns, the same process was repeated for all of *Bin1* exons found in RefSeq and AVISPA's transcript DB.

3.1. Interpretation of AVISPA output for cancer-associated Bin1 triplet 12.13.17

3.1.1. Analysis of alternative vs. constitutive and tissue-specific predictions

Fig. 3 highlights the three sections of the query summary for triplet 12.13.17 that is used to navigate through the various splicing predictions that AVISPA provides. We see from the query matching section of the output (Fig. 3a) that this exon was matched to AVISPA's DB of cassette exons (level 1), indicating transcript-based support for this exon being alternatively spliced. Query match details indicate the coordinates of the matched event used for predictions, and we see there are differences between the query and the matched sequence. As mentioned before, it is the matched sequence that is analyzed, not the input which had a mislabeled exonic start position and a fabricated SNP.

The results of the splicing predictions are presented as a table and indicate that, based on the specified thresholds, this exon is both alternatively spliced and differentially included in all four tissue groups (Fig. 3b). Notably, passing the significance threshold for multiple tissues can occur for multiple reasons. Obviously, the more permissive the significance threshold used, the higher the chance more tissue-dependent predictions will pass it. Also, some regulatory features are shared between models for different tissue-dependent splicing and their occurrence will push for higher scores in all prediction tasks. Some of those features may be general regulation indicators, such as high intronic conservation around the alternative exon, while others represent factors known to operate in different tissues, such as FOX-1/2 binding in both muscle and brain tissues [33]. It is therefore important to also compare the relative confidence in the predictions and the chance of error. For example, the summary shown in Fig. 3b indicates a false positive



Query Matching

Query Support

The query was matched to our internal database of known cassette exons. (level 1)

Query match details

Reference	chr18
Strand	+
PreA	32424819-32424967
A	32426223-32426352
PostA	32431670-32431742

Differences found between query and matching sequence −

```
PreA      1
Original  TGACAATGCCCCTGAGAAAGG
Matched   TGACAATGCCCCTGAGAAAGG
A         1
Original  -TCCGGAACGGCCCACCTGTC
Matched   CTCCGGAAAGGCCCACCTGTC
PostA     1
Original  CCAGCAGAGGCCTCCGAGGTG
Matched   CCAGCAGAGGCCTCCGAGGTG
```

Predictions

User defined threshold Alternative (FPR: 0.05), Tissue-dependent (Rank: 0.05)

	FPR ?	Rank ?	Sensitivity ?	F-Effect ?
Alternative vs Constitutive	0.003	0.014	0.198	≡
Differential Splicing in Tissue				
+ Central Nervous System	0.004	0.002	0.074	≡
− Muscle	0.021	0.004	0.098	≡
Increased Inclusion ▲	0.026	0.006	0.192	
Increased Exclusion ▼	0.004	0.002	0.024	
+ Embryo	0.022	0.008	0.120	≡
+ Digestive	0.020	0.003	0.052	≡

Motif Visualization

Prediction Type: Alternative vs Constitutive ⬍ *Start Region:* PreA ⬍ *End Region:* PostA ⬍

Results ?

Fig. 3. Query matching and predictions summary for *Bin1* triplet 12.13.17. (a) The query matching section provides information on which database the query was matched against (known events, known transcripts, or the genome), the matched sequence location, and any discrepancies between the query and the matched sequence. Here, a different start position (green circle) and a single nucleotide discrepancy (red circle) were detected. (b) The predictions table lists for every prediction task (rows) the matching false positive rate (FPR), relative rank, and sensitivity associated with the query's score. Only predictions that pass the user defined threshold are listed. Predictions for differential inclusion are denoted with an up blue arrow, and for differential exclusion with a down red arrow. Here, the muscle dependent splicing predictions were expanded for illustrative purposes (details for other predictions are found by clicking the "+" in the output). In this case AVISPA's predictions were more confident in muscle dependent differential exclusion, though predictions for changes in both directions passed the user defined threshold. Additionally, links to the *in silico* feature analysis (as in Figs. 4b and 6a) for each prediction are available as clickable bar charts in the final column. (c) The final portion of the output summary allows users to visualize predicted regulatory motifs for each prediction by mapping them to the UCSC genome browser ("Motif map" icon, as in Figs. 4c and 6b) or representing NFE values (see main text) as stacked bar charts colored by regional location ("Motif effect" icon, as in Fig. 4a). Users can specify which splicing prediction they wish to visualize and apply regional restrictions by using the drop down menus.

rate of 0.003 for the alternative versus constitutive prediction step. This means that, based on comparisons between the query and the negative set used to train the model, the probability of falsely rejecting the null hypothesis (i.e., classifying this exon as alternative when it is actually constitutive) is 0.3% (Fig. 3b, first row). The rank values for the the tissue-dependent splicing predictions indicates that the most confident prediction is for differential inclusion in CNS where only 0.2% of the samples in the reference set achieve a score as high as this *Bin1* exon 13 triplet (Fig. 3b, rank = 0.002).

In addition to predicting whether the query is differentially spliced in these tissue groups, AVISPA also supplies predictions for whether the exon exhibits increased inclusion or increased exclusion in these tissues. This information about the polarity of the tissue-dependent splicing changes is provided for each tissue group by clicking on the "+" sign in the splicing prediction table, and here we show muscle as an example (Fig. 3b). In muscle AVISPA makes predictions that pass the given significance threshold for both increased inclusion (rank = 0.006) and increased exclusion (rank = 0.002). Dual high confidence predictions such as this indicate that the exon has features that strongly suggest there is a splicing change in muscle, but the direction of this change is less clear.

Similar bi-directional predictions are given for CNS, embryo, and digestive tissues, but the relative magnitudes of these predictions point to an interesting hypothesis where the exon is differentially included during development but then excluded in adult tissues (increased inclusion rank = 0.002 for embryo; increased exclusion rank = 0.002 for CNS, muscle, and digestive tissues). Such a splicing pattern would be in line with reports that missplicing of the *BIN1* gene, leading to inclusion of exon 13, is associated with human cancers (Fig. 2, *Bin1* 12.13.17 and *Bin1* 12.13.18). Specifically, several works have shown that when exon 13 is included in this way in adult tissues in both human and mouse cell lines, BIN1 no longer interacts with Myc to inhibit oncogenic activity [24,29,30,34,35]. We note that this same tissue-specific preference for inclusion in embryo and exclusion in adult tissues is observed with the splicing prediction for the *Bin1* 12.13.18 triplet (data not shown).

3.1.2. CNS regulatory feature analysis

In addition to splicing predictions, AVISPA provides the user with information about numerous regulatory features used to make these predictions. This information can help guide users to which *cis* and *trans* elements may control the inclusion of their exon of interest. The first type of information is the relative enrichment of these features in the query compared to several reference groups of exons, such as alternative or constitutive exons. This information is available to users as "Feature Effect" files for each prediction made, found by clicking on the bar charts in the rightmost column (Fig. 3b). Here we use the outputs for the CNS-specific predictions as an example.

Various features are represented as a table with a heat map coloring scheme to highlight those that have particularly high or low values for their respective groups (Fig. 4b). Each row represents a feature and each column compares these features of the *Bin1* exon 13 triplet to a reference set. For example, we see that the position of the first AG dinucleotide upstream of exon 13 scores relatively high, particularly when compared to the constitutive set where it scores higher than ~98.4% of this set (Fig. 4b, fourth column). Similarly, this feature scores higher than ~95.8% of the alternative set (Fig. 4b, third column). The fact that this feature scores higher than a greater percentage of the constitutive set could be expected because more distant first AG dinucleotides have been associated with more distant branch point sequences and alternative splicing [15]. We also note the high scores for secondary structure free regions downstream of exon 13 (I2(5′)) and upstream of exon 17 (I2(3′)). Because many splicing regulatory proteins bind RNA in a single strand state [36], these values indicate a higher likelihood that RBPs can bind to these regions and regulate exon inclusion.

For putative regulatory sequence motifs, AVISPA also includes a normalized feature effect (NFE) score, summarized using a bar chart. For each splicing prediction that passes the user defined threshold, a matching NFE summary file is produced. Briefly, the feature effect (FE) is computed by comparing the predictions for the "wild type" and an *in silico* "mutant" where the regulatory motif has been removed. The NFE score is then computed by normalizing each individual feature effect by the total effect of all motifs evaluated [22]. Given the many possible regulatory elements surrounding a query exon, the NFE score can thus help narrow down the list of regulatory candidates to ones that are more likely to affect the exon. However, since the scores are normalized per query and are based on differences in prediction confidence, users should be careful not to associate NFE scores with a measure of exon inclusion levels.

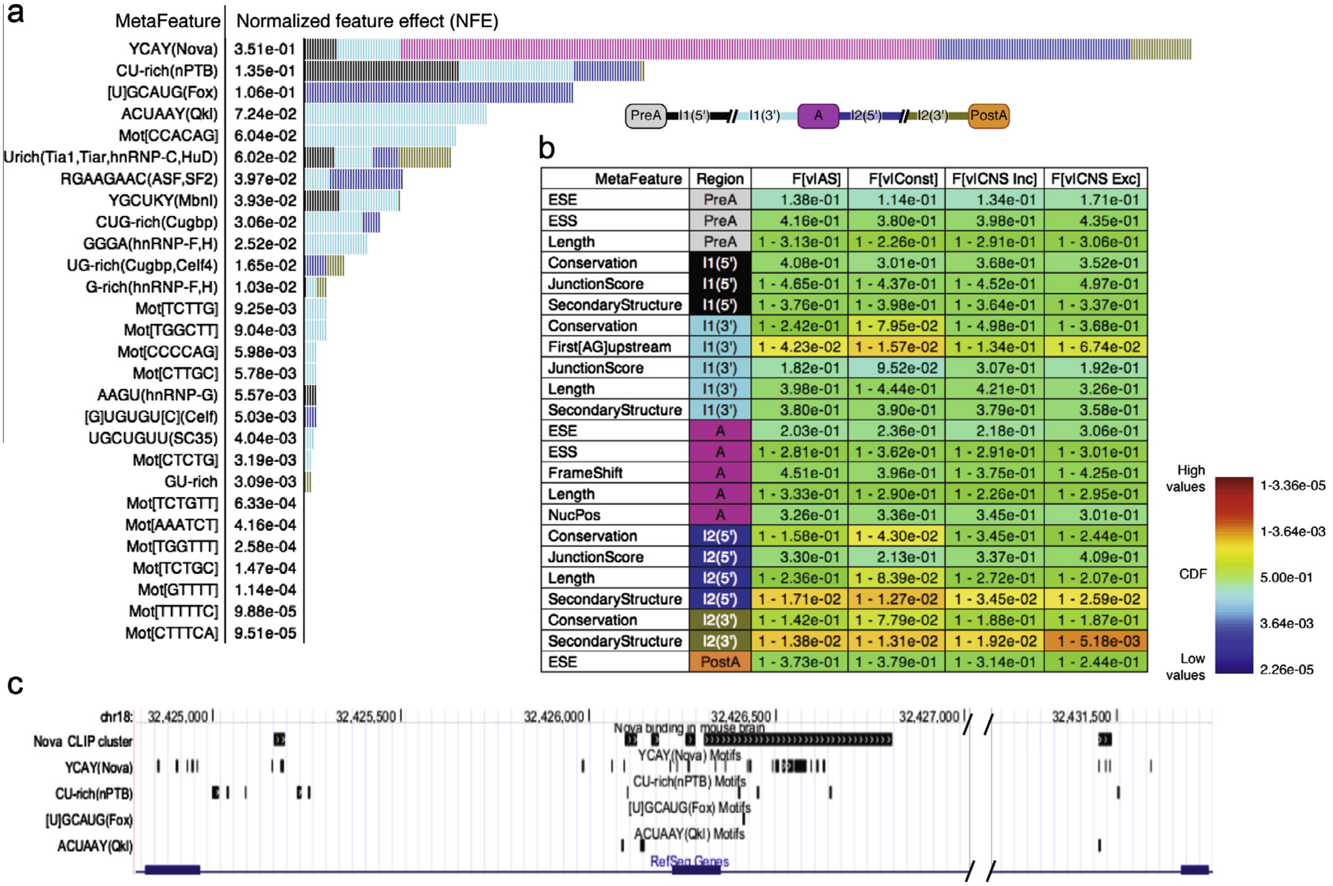
Fig. 4a represents the NFE scores of regulatory motifs that may affect CNS *Bin1* exon 13 as a stacked bar chart where the colors represent the region in which a motif was found. This chart is generated for each splicing prediction and is linked to AVISPA's summary page. Selecting the prediction type in the Motif Visualization section (e.g., CNS) and clicking the "Motif Effect" icon (Fig. 3c) displays the chart. AVISPA predicts motifs known to bind the splicing factors NOVA, PTB, FOX-1/2, and QKI to be particularly important regulators of CNS-specific alternative splicing of *Bin1* triplet 12.13.17 (Fig. 4a).

A number of these predictions are consistent with previously described results. Utilizing another aspect of AVISPA's visualization output, the top four motifs were mapped to the UCSC genome browser by selecting the "Motif Map" icon (Fig. 3c), along with a custom track of UV cross-linking and immunoprecipitation (CLIP) tag clusters for NOVA binding in mouse brain [37] (Fig. 4c). The combined genome browser view shows that numerous AVISPA predicted, CNS-relevant NOVA motifs correspond to actual *in vivo* binding sites (Fig. 4c, top two tracks). The two CLIP binding sites within the regions considered that are not predicted by AVISPA lack the YCAY NOVA consensus motif. Additionally, NOVA, the topped rank motif by NFE (Fig. 4a), was shown to be an important regulator of exon 13 splicing in mouse brains [37]. Similarly, Muscleblind-like protein (MBNL) ranks eighth overall by NFE (Fig. 4a) and has previously been observed to have a modest effect on splicing of this event (~5% change) in MBNL1 knockout mouse brains compared with controls [38]. Finally, we note the appearance of the SF2/ASF (also known as SRSF1) motif which ranks 7th overall in CNS by NFE score (Fig 4a). AVISPA similarly predicts this splicing factor to be influential in muscle (ranks 9th) and embryonic tissues (ranks 13th) (data not shown). Interestingly, slight SF2/ASF overexpression was sufficient to transform immortal rodent fibroblasts, in part due to increased exon 13 inclusion of the *Bin1* event described here [35]. This result thus highlights the potential role of other putative regulatory factors described here to maintain BIN1 tumor suppressor activity.

AVISPA's CNS predictions also suggest potential novel regulators. FOX-1/2 ranks third by NFE score (Fig. 4a) and is an important muscle and neuronal specific splicing factor but has not, to our knowledge, been shown to regulate *Bin1*. Previous work has shown that triplet 12.13.17 is under PTB and QKI regulation in the mouse proliferating myoblast cell line (C2C12) by microarray [31], but the neuronal regulation of this event has yet to be examined. AVISPA predicts these two RBPs to be particularly influential for this alternative event in CNS tissues where PTB ranks second overall by NFE and QKI ranks fourth (Fig. 4a). We chose to validate these two novel CNS predictions *in vivo* using a mouse neuroblastoma cell line.

3.1.3. In vivo validation of AVISPA predictions of PTB and QKI regulation

To experimentally validate the predicted CNS regulation of our event of interest by PTB and QKI, we performed low cycle RT-PCR on RNA extracted from N2a cells depleted for these splicing factors by siRNA transfection and compared them to control RNA from cells transfected with siRNA for GFP. Protein depletion was



MetaFeature	Normalized feature effect (NFE)
YCAY(Nova)	3.51e-01
CU-rich(nPTB)	1.35e-01
[U]GCAUG(Fox)	1.06e-01
ACUAAY(Qkl)	7.24e-02
Mot[CCACAG]	6.04e-02
Urich(Tia1,Tiar,hnRNP-C,HuD)	6.02e-02
RGAAGAAC(ASF,SF2)	3.97e-02
YGCUKY(Mbnl)	3.93e-02
CUG-rich(Cugbp)	3.06e-02
GGGA(hnRNP-F,H)	2.52e-02
UG-rich(Cugbp,Celf4)	1.65e-02
G-rich(hnRNP-F,H)	1.03e-02
Mot[TCTTG]	9.25e-03
Mot[TGGCTT]	9.04e-03
Mot[CCCCAG]	5.98e-03
Mot[CTTGC]	5.78e-03
AAGU(hnRNP-G)	5.57e-03
[G]UGUGU[C](Celf)	5.03e-03
UGCUGUU(SC35)	4.04e-03
Mot[CTCTG]	3.19e-03
GU-rich	3.09e-03
Mot[TCTGTT]	6.33e-04
Mot[AAATCT]	4.16e-04
Mot[TGGTTT]	2.58e-04
Mot[TCTGC]	1.47e-04
Mot[GTTTT]	1.14e-04
Mot[TTTTTC]	9.88e-05
Mot[CTTTCA]	9.51e-05

MetaFeature	Region	F[vIAS]	F[vIConst]	F[vICNS Inc]	F[vICNS Exc]
ESE	PreA	1.38e-01	1.14e-01	1.34e-01	1.71e-01
ESS	PreA	4.16e-01	3.80e-01	3.98e-01	4.35e-01
Length	PreA	1 - 3.13e-01	1 - 2.26e-01	1 - 2.91e-01	1 - 3.06e-01
Conservation	I1(5')	4.08e-01	3.01e-01	3.68e-01	3.52e-01
JunctionScore	I1(5')	1 - 4.65e-01	1 - 4.37e-01	1 - 4.52e-01	4.97e-01
SecondaryStructure	I1(5')	1 - 3.76e-01	1 - 3.98e-01	1 - 3.64e-01	1 - 3.37e-01
Conservation	I1(3')	1 - 2.42e-01	1 - 7.95e-02	1 - 4.98e-01	1 - 3.68e-01
First[AG]upstream	I1(3')	1 - 4.23e-02	1 - 1.57e-02	1 - 1.34e-01	1 - 6.74e-02
JunctionScore	I1(3')	1.82e-01	9.52e-02	3.07e-01	1.92e-01
Length	I1(3')	3.98e-01	1 - 4.44e-01	4.21e-01	3.26e-01
SecondaryStructure	I1(3')	3.80e-01	3.90e-01	3.79e-01	3.58e-01
ESE	A	2.03e-01	2.36e-01	2.18e-01	3.06e-01
ESS	A	1 - 2.81e-01	1 - 3.62e-01	1 - 2.91e-01	1 - 3.01e-01
FrameShift	A	4.51e-01	3.96e-01	1 - 3.75e-01	1 - 4.25e-01
Length	A	1 - 3.33e-01	1 - 2.90e-01	1 - 2.26e-01	1 - 2.95e-01
NucPos	A	3.26e-01	3.36e-01	3.45e-01	3.01e-01
Conservation	I2(5')	1 - 1.58e-01	1 - 4.30e-02	1 - 3.45e-01	1 - 2.44e-01
JunctionScore	I2(5')	3.30e-01	2.13e-01	3.37e-01	4.09e-01
Length	I2(5')	1 - 2.36e-01	1 - 8.39e-02	1 - 2.72e-01	1 - 2.07e-01
SecondaryStructure	I2(5')	1 - 1.71e-02	1 - 1.27e-02	1 - 3.45e-02	1 - 2.59e-02
Conservation	I2(3')	1 - 1.42e-01	1 - 7.79e-02	1 - 1.88e-01	1 - 1.87e-01
SecondaryStructure	I2(3')	1 - 1.38e-02	1 - 1.31e-02	1 - 1.92e-02	1 - 5.18e-03
ESE	PostA	1 - 3.73e-01	1 - 3.79e-01	1 - 3.14e-01	1 - 2.44e-01

Fig. 4. *In silico* feature analysis of *Bin1* triplet 12.13.17 in CNS. (a) Top scoring motif effects represented as a stacked bar chart. Each name in the first column ("MetaFeatures") corresponds to a motif and RBP known to bind it. The second column provides a summed normalized feature effect (NFE) value for all occurrences of that motif in the seven genomic regions analyzed and the stacked bars are a visual representation of these values where each bar corresponding to NFE of 10^{-3}. Colors correspond to the region in which the motif was found as in the legend provided. (b) Enrichment of non-motif features. Each feature (row) in the query is compared against several reference sets (columns) indicated by the header (AS, alternative set; Const, constitutive set; CNS Inc, CNS inclusion set; and CNS Exc, CNS exclusion set). The table entries correspond to the relative rank of the query's feature value compared to the reference set of each column. Enrichment is visually represented as a heat map where red is associated with high values and blue is associated with low values. The region containing each feature is indicated in the second column, with colors matching the provided regional legend. (c) Motif map of top four scoring motifs on the UCSC genome browser along with a custom track containing NOVA CLIP binding clusters from mouse brain [37] to show overlap between AVISPA predicted NOVA motifs and *in vivo* binding sites.

confirmed by western blot (Fig. 5a). Upon PTB knockdown (KD), we observed a striking decrease in exon 13 inclusion, with an average differential of 41% compared to controls (i.e., 85–44% inclusion, Fig. 5b, $p = 3.3 \times 10^{-6}$, two tailed *t*-test). While the difference was not as pronounced upon QKI KD, we observed a consistent decrease in exon inclusion, averaging 12% (Fig. 5b, $p = 0.01$, two tailed *t*-test). These data suggest that both PTB and QKI act to promote exon 13 inclusion, in the context of the exon 12.13.17 triplet.

AVISPA similarly predicted PTB and QKI to be important regulators of the neuronal specific inclusion of exon 7 (triplet 6.7.8, Fig. 2) where PTB ranks first and QKI ranks fifth by NFE scores (data not shown). RT-PCR results validate these CNS predictions and show that, upon PTB or QKI KD, exon 7 inclusion increases on average by 22% or 34%, respectively (Fig. 5b, $p = 2.1 \times 10^{-5}$, two tailed *t*-test for PTB KD; $p = 4.5 \times 10^{-7}$, two tailed *t*-test for QKI KD). In all, these data suggest that PTB and QKI normally act to repress exon 7 inclusion and promote exon 13 inclusion in N2a cells and highlight AVISPA's ability to suggest novel splicing regulators of exons of interest for experimental investigation.

3.2. Analysis of muscular disease-associated splicing event suggests novel regulators

As mentioned previously, the misregulation of the muscle-specific inclusion of *Bin1* exon 11 has been associated with muscle

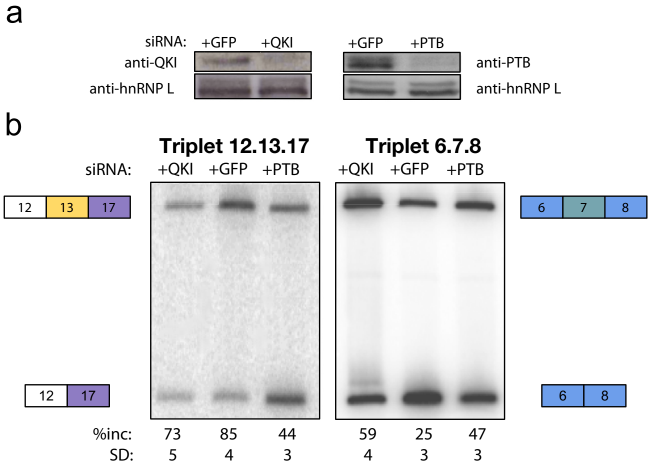
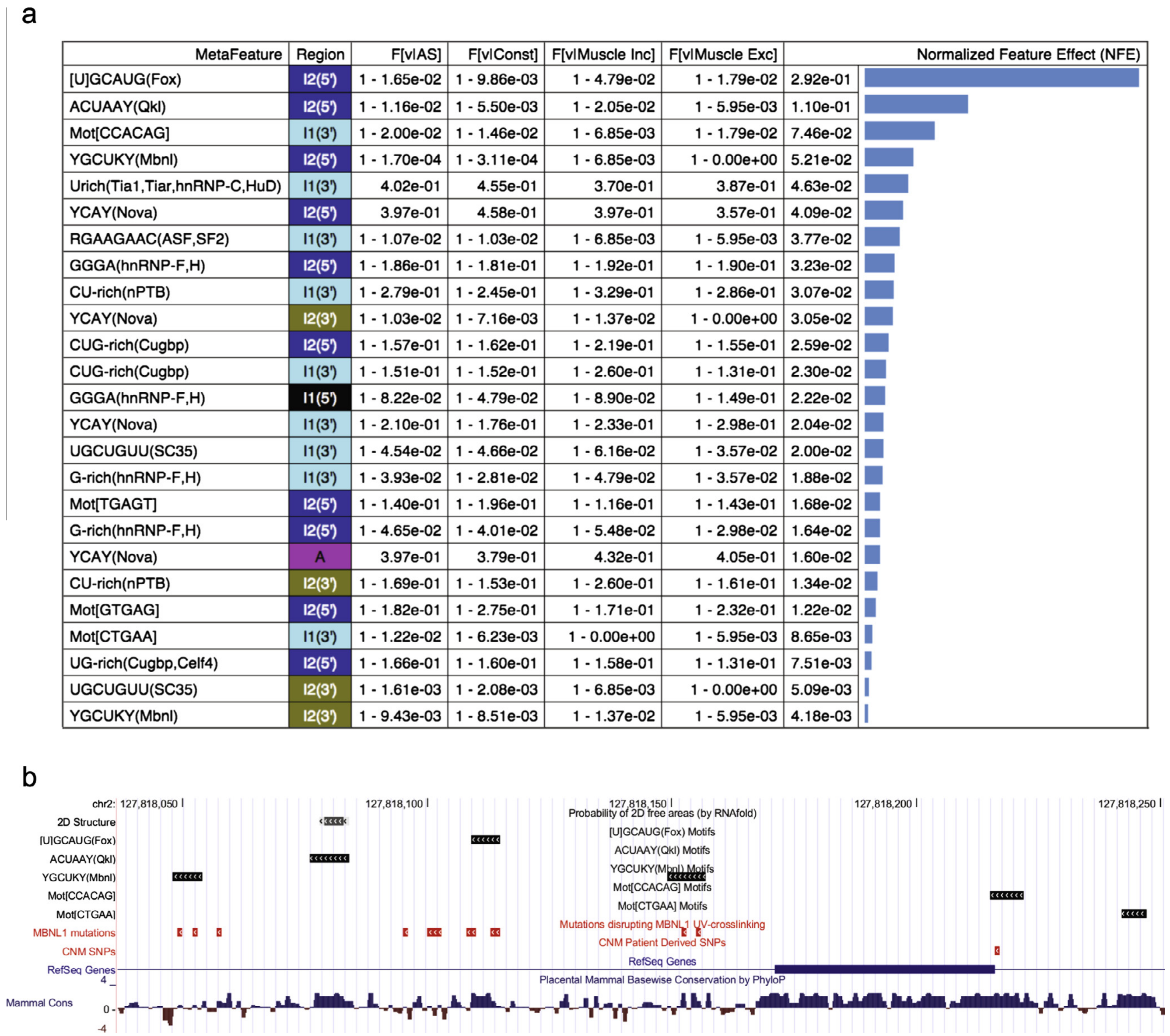


Fig. 5. *In vivo* validation of QKI and PTB regulation in N2a cells. (a) Western blots confirm siRNA knockdown of QKI and PTB when compared to controls transfected with GFP siRNA. Antibodies for hnRNP-L were used as loading controls. (b) RT-PCR validations of predicted CNS splicing regulation with QKI or PTB depletion in N2a cells for triplet 12.13.17 (left) and 6.7.8 (right). Representation of isoform structure is presented next to each gel and correspond to Fig. 2. Percent inclusion (%inc) and standard deviation (SD) represent data from at least four replicates.



a

MetaFeature	Region	F[vIAS]	F[vIConst]	F[vIMuscle Inc]	F[vIMuscle Exc]	Normalized Feature Effect (NFE)
[U]GCAUG(Fox)	I2(5')	1 - 1.65e-02	1 - 9.86e-03	1 - 4.79e-02	1 - 1.79e-02	2.92e-01
ACUAAY(Qkl)	I2(5')	1 - 1.16e-02	1 - 5.50e-03	1 - 2.05e-02	1 - 5.95e-03	1.10e-01
Mot[CCACAG]	I1(3')	1 - 2.00e-02	1 - 1.46e-02	1 - 6.85e-03	1 - 1.79e-02	7.46e-02
YGCUKY(Mbnl)	I2(5')	1 - 1.70e-04	1 - 3.11e-04	1 - 6.85e-03	1 - 0.00e+00	5.21e-02
Urich(Tia1,Tiar,hnRNP-C,HuD)	I1(3')	4.02e-01	4.55e-01	3.70e-01	3.87e-01	4.63e-02
YCAY(Nova)	I2(5')	3.97e-01	4.58e-01	3.97e-01	3.57e-01	4.09e-02
RGAAGAAC(ASF,SF2)	I1(3')	1 - 1.07e-02	1 - 1.03e-02	1 - 6.85e-03	1 - 5.95e-03	3.77e-02
GGGA(hnRNP-F,H)	I2(5')	1 - 1.86e-01	1 - 1.81e-01	1 - 1.92e-01	1 - 1.90e-01	3.23e-02
CU-rich(nPTB)	I1(3')	1 - 2.79e-01	1 - 2.45e-01	1 - 3.29e-01	1 - 2.86e-01	3.07e-02
YCAY(Nova)	I2(3')	1 - 1.03e-02	1 - 7.16e-03	1 - 1.37e-02	1 - 0.00e+00	3.05e-02
CUG-rich(Cugbp)	I2(5')	1 - 1.57e-01	1 - 1.62e-01	1 - 2.19e-01	1 - 1.55e-01	2.59e-02
CUG-rich(Cugbp)	I1(3')	1 - 1.51e-01	1 - 1.52e-01	1 - 2.60e-01	1 - 1.31e-01	2.30e-02
GGGA(hnRNP-F,H)	I1(5')	1 - 8.22e-02	1 - 4.79e-02	1 - 8.90e-02	1 - 1.49e-01	2.22e-02
YCAY(Nova)	I1(3')	1 - 2.10e-01	1 - 1.76e-01	1 - 2.33e-01	1 - 2.98e-01	2.04e-02
UGCUGUU(SC35)	I1(3')	1 - 4.54e-02	1 - 4.66e-02	1 - 6.16e-02	1 - 3.57e-02	2.00e-02
G-rich(hnRNP-F,H)	I1(3')	1 - 3.93e-02	1 - 2.81e-02	1 - 4.79e-02	1 - 3.57e-02	1.88e-02
Mot[TGAGT]	I2(5')	1 - 1.40e-01	1 - 1.96e-01	1 - 1.16e-01	1 - 1.43e-01	1.68e-02
G-rich(hnRNP-F,H)	I2(5')	1 - 4.65e-02	1 - 4.01e-02	1 - 5.48e-02	1 - 2.98e-02	1.64e-02
YCAY(Nova)	A	3.97e-01	3.79e-01	4.32e-01	4.05e-01	1.60e-02
CU-rich(nPTB)	I2(3')	1 - 1.69e-01	1 - 1.53e-01	1 - 2.60e-01	1 - 1.61e-01	1.34e-02
Mot[GTGAG]	I2(5')	1 - 1.82e-01	1 - 2.75e-01	1 - 1.71e-01	1 - 2.32e-01	1.22e-02
Mot[CTGAA]	I1(3')	1 - 1.22e-02	1 - 6.23e-03	1 - 0.00e+00	1 - 5.95e-03	8.65e-03
UG-rich(Cugbp,Celf4)	I2(5')	1 - 1.66e-01	1 - 1.60e-01	1 - 1.58e-01	1 - 1.31e-01	7.51e-03
UGCUGUU(SC35)	I2(3')	1 - 1.61e-03	1 - 2.08e-03	1 - 6.85e-03	1 - 0.00e+00	5.09e-03
YGCUKY(Mbnl)	I2(3')	1 - 9.43e-03	1 - 8.51e-03	1 - 1.37e-02	1 - 5.95e-03	4.18e-03

Fig. 6. *In silico* feature analysis of *BIN1* triplet 10.11.12 in muscle. (a) Top 25 scoring individual motifs predicted to affect muscle specific splicing predictions, separated by region. This table provides motif enrichment values compared to reference sets, as in Fig. 4b, and NFE values for each of these individual motifs, as in Fig. 4a. (b) A portion of the motif map proximal to exon 11 on the UCSC genome browser with select top-scoring motifs of interest, secondary structure, and conservation. Predicted MBNL motifs show overlap with experimentally derived mutations shown to affect MBNL1 UV-crosslinking to an exon 11 *BIN1* minigene [27]. Additionally, we map a CNM patient derived disease SNP that induces exon skipping in humans (bottom track) or activates an upstream cryptic 3′ss (not shown) in canines with IMGD [28], both of which overlap AVISPA predicted motifs CCACAG and CTGAA, respectively.

disorders [27,28], making it an intriguing example to analyze (Fig. 2). Although earlier studies found evidence of exon 11 inclusion in differentiating mouse muscle cells and similar transcript structure to human *BIN1* [25,26], there is a lack of transcriptional evidence for this exon's inclusion. Therefore, for the analysis presented here the human sequence was used instead. Notably, applying a mouse splicing code to human exons has already been shown to produce accurate predictions for human exons [21]. Admittedly, AVISPA's human exons analysis is still in beta at this time, but *Bin1* exon 11 analysis serves well to illustrate AVISPA's capabilities for potential users.

AVISPA accurately predicts exon 11 to be alternatively spliced (FPR = 0.015), in line with known transcript structures (Fig. 2). AVISPA's output summary prediction table (similar to Fig. 3b, omitted for this event for brevity) reports tissue-dependent splicing for all major tissue groups pass the default threshold (Rank <0.05), with the best rank (i.e., most confident prediction) given to differential splicing in muscle (0.005), followed by digestive (0.009) embryo (0.016) and CNS (0.023). Although the predicted direction of these changes was somewhat ambiguous, the relative confidence in the muscle predictions agrees with the literature, suggesting increased inclusion of this muscle-specific, PI domain encoding exon (increased inclusion rank = 0.004 vs. exclusion rank = 0.013).

The feature analysis for this event not only recapitulates some of the known regulatory mechanisms, but also suggests novel putative regulators of this disease-associated splicing event. In DM, expanded CUG and CCUG repeats act to sequester the splicing factor Muscleblind-like-1 (MBNL1), which was shown to be an important promoter of exon 11 inclusion in a *BIN1* minigene reporter in muscle through overexpression and siRNA knockdown of MBNL1 [27]. AVISPA predicts MBNL1 to be a regulator of this exon where it ranks fourth by NFE (Fig. 6a). Strikingly, the MBNL1 binding sites that AVISPA predicts to be influential map to two of three non-overlapping segments shown by Fugier et al. [27] to bind MBNL1 by UV cross-linking. Additionally, we observe overlap of the predicted motifs with four UGC disrupting mutations that together contribute to decreased MBNL1 binding and a loss of splicing responsiveness upon MBNL1 overexpression or depletion (Fig. 6b). AVISPA does not predict MBNL binding at those experimental mutation sites that do not contain a motif similar to the consensus (YGCUKY). MBNL1 CLIP-seq data in mouse muscle and proliferating myoblasts (C2C12 cell line) does not show binding within these same regions, which highlights AVISPA's power to suggest regulators that may be overlooked when using high-throughput experimental assays in isolation.

In another study focusing on CNM, a patient derived homozygous mutation at the 3′ splice site (3′ss) of exon 11 (G>A at the final nt of the intron) resulted in exon skipping in humans. Similarly, a 3′ss mutation (A>G at the second to last nt) that activates an upstream cryptic 3′ss was found to be the cause of the canine Inherited Myopathy of Great Danes (IMGD) in five affected dogs [28]. While AVISPA currently does not support alternative 3′ and 5′ splice site predictions, it does identify motifs related to these events as affecting its splicing predictions. Specifically, AVISPA's predictions include a CCACAG motif, which maps to the annotated 3′ss of exon 11, and a CTGAA motif mapped to the cryptic alternative 3′ss utilized in IMGD dogs (Fig. 6b). Previous reports show that competing splice sites can affect splicing potential [39], suggesting that the conserved yet cryptic alternative 3′ss may be functionally relevant in this event.

The analysis of exon 11 also suggests novel regulatory elements that may control this disease-associated event. The top scoring motif in muscle is UGCAUG, a motif known to bind the splicing factor FOX-1/2 (Fig. 6a). Here, the motif is found downstream of exon 11 in a relatively highly conserved region (Fig. 6b). FOX-1/2 is a brain and muscle specific splicing factor and it has been shown to more commonly promote exon inclusion in these tissues when binding downstream of an alternative exon [33]. In addition, AVISPA's analysis also suggests QKI as a putative regulator of this exon in muscle where it ranks second on an individual motif level (Fig. 6a). Here QKI is predicted to bind to an ACUAAC motif, a perfect match to its consensus, in a highly conserved region. Additionally, this motif falls in a region predicted to be secondary structure free, suggesting this motif exists in an accessible, single stranded form (Fig. 6b). This result is in line with recent analysis showing QKI can promote exon inclusion when bound downstream of exons in muscle and that it regulates splicing of other *Bin1* exons [31].

In summary, AVISPA analysis of *Bin1* exon 11 is not only consistent with previous published results, but also offers QKI and FOX-1/2 as important novel regulators of the exon inclusion levels. AVISPA's predictions thus serve as hypotheses that can be experimentally tested to provide a more complete picture of the regulation of this essential splicing event in muscle.

4. Conclusions

Several decades of research have revealed the prevalence of alternative splicing throughout the genome, its importance for post-transcriptional control of gene expression, and many *cis*- and *trans*-acting elements that regulate this complex process. Much progress has been made more recently in creating predictive splicing code models that consider how these regulatory elements may combine to create splicing outcomes that differ across cellular contexts. Here we have shown how these splicing code models can now be applied, via the AVISPA web tool, by researchers studying RNA biogenesis, development, and diseases with a possible splicing defect component. AVISPA allows researchers to analyze exons from any gene of interest. Its graphical summary allows users to test whether a given exon matches known cassette events and transcripts in AVISPA's DB; assess confidence in the exon being alternatively spliced and differentially included in specific tissues; identify enriched regulatory features such as secondary structure free regions; assess the effect regulatory motifs exert on splicing predictions; and map these motifs to the genome browser. Using *Bin1* as a case study we supplied a detailed how-to guide for splicing analysis, highlighting important cautions and limitations of AVISPA's current implementation that users should consider. We have shown that AVISPA's *in silico* splicing analysis of *Bin1* are in line with many experimental results and suggest novel regulators of disease-associated splicing events. Such predictions offer users specific hypotheses that can then be experimentally validated, for example by using RNAi depletion of splicing factors known to bind the identified regulatory motifs. Using this strategy, we validated QKI and PTB as novel CNS regulators of the cancer-related *Bin1* triplet 12.13.17 triplet and the neuronal-specific inclusion of exon 7 in triplet 6.7.8. We hope the illustrative analysis presented here will help guide analysis and discoveries pertaining to transcriptome complexity, RNA regulation, and human disease.

Acknowledgements

We thank members of the Barash lab for helpful comments and discussions regarding the manuscript and members of the Lynch lab for advice on experimental protocols.

References

[1] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Nature 456 (7221) (2008) 470–476.
[2] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Nat. Genet. 40 (12) (2008) 1413–1415.
[3] T.W. Nilsen, B.R. Graveley, Nature 463 (7280) (2010) 457–463.
[4] O. Isken, L.E. Maquat, Nat. Rev. Genet. 9 (9) (2008) 699–712.
[5] A. Kalsotra, T.A. Cooper, Nat. Rev. Genet. 12 (10) (2011) 715–729.
[6] Z. Melamed, A. Levy, R. Ashwal-Fluss, G. Lev-Maor, K. Mekahel, N. Atias, S. Gilad, R. Sharan, C. Levy, S. Kadener, G. Ast, Mol. Cell 50 (6) (2013) 869–881.
[7] G.-S. Wang, T.A. Cooper, Nat. Rev. Genet. 8 (10) (2007) 749–761.
[8] M. Chen, J.L. Manley, Nat. Rev. Mol. Cell. Biol. 10 (11) (2009) 741–754.
[9] D.D. Licatalosi, R.B. Darnell, Nat. Rev. Genet. 11 (1) (2010) 75–87.
[10] G. Yeo, C.B. Burge, J. Comput. Biol. 11 (2004) 377–394.
[11] M. Hiller, Z. Zhang, R. Backofen, S. Stamm, PLoS Genet. 3 (11) (2007) e204.
[12] Z. Wang, C.B. Burge, RNA 14 (5) (2008) 802–813.
[13] R.I. Dogan, L. Getoor, W.J. Wilbur, S.M. Mount, Nucl. Acids Res. 35 (Suppl. 2) (2007) W285–W291.
[14] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, C. Béroud, Nucl. Acids Res. 37 (9) (2009) e67.
[15] A. Corvelo, M. Hallegger, C.W.J. Smith, E. Eyras, PLoS Comput. Biol. 6 (11) (2010) e1001016.
[16] X.H.-F. Zhang, T. Kangsamaksin, M.S.P. Chao, J.K. Banerjee, L.A. Chasin, Mol. Cell. Biol. 25 (16) (2005) 7323–7332.
[17] L. Cartegni, J. Wang, Z. Zhu, M.Q. Zhang, A.R. Krainer, Nucl. Acids Res. 31 (13) (2003) 3568–3571.
[18] S. Schwartz, E. Hall, G. Ast, Nucl. Acids Res. 37 (2) (2009) W189–W192.
[19] F. Piva, M. Giulietti, A.B. Burini, G. Principato, Hum. Mutat. 33 (1) (2012) 81–85.
[20] Y. Barash, J.A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B.J. Blencowe, B.J. Frey, Nature 465 (7294) (2010) 53–59.
[21] N.L. Barbosa-Morais, M. Irimia, Q. Pan, H.Y. Xiong, S. Gueroussov, L.J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C.M. Misquitta-Ali, M.D. Wilson, P.M. Kim, D.T. Odom, B.J. Frey, B.J. Blencowe, Science 338 (6114) (2012) 1587–1593.
[22] Y. Barash, J. Vaquero-Garcia, J. González-Vallinas, H.Y. Xiong, W. Gao, L.J. Lee, B.J. Frey, Genome Biol. 14 (2013) R114.
[23] J. Goecks, A. Nekrutenko, J. Taylor, Genome Biol. 8 (2010) R86.
[24] R. Wechsler-Reya, D. Sakamuro, J. Zhang, J. Duhadaway, G.C. Prendergast, J. Biol. Chem. 272 (50) (1997) 31453–31458.
[25] N.C. Mao, E. Steingrimsson, J. DuHadaway, W. Wasserman, J.C. Ruiz, N.G. Copeland, N.A. Jenkins, G.C. Prendergast, Genomics 56 (1) (1999) 51–58.
[26] R.J. Wechsler-Reya, K.J. Elliott, G.C. Prendergast, Mol. Cell. Biol. 18 (1) (1998) 566–575.
[27] C. Fugier, A.F. Klein, C. Hammer, S. Vassilopoulos, Y. Ivarsson, A. Toussaint, V. Tosch, A. Vignaud, A. Ferry, N. Messaddeq, Y. Kokunai, R. Tsuburaya, P. de la Grange, D. Dembele, V. Francois, G. Precigout, C. Boulade-Ladame, M.-C. Hummel, A. Lopez de Munain, N. Sergeant, A. Laquerrière, C. Thibault, F.

Deryckere, D. Auboeuf, L. Garcia, P. Zimmermann, B. Udd, B. Schoser, M.P. Takahashi, I. Nishino, G. Bassez, J. Laporte, D. Furling, N. Charlet-Berguerand, Nat. Med. 17 (6) (2011) 720–725.

[28] J. Böhm, N. Vasli, M. Maurer, B. Cowling, G.D. Shelton, W. Kress, A. Toussaint, I. Prokic, U. Schara, T.J. Anderson, J. Weis, L. Tiret, J. Laporte, PLoS Genet. 9 (6) (2013) e1003430.

[29] K. Ge, J. DuHadaway, W. Du, M. Herlyn, U. Rodeck, G.C. Prendergast, PNAS 96 (17) (1999) 9689–9694.

[30] Q. Xu, C. Lee, Nucl. Acids Res. 31 (19) (2003) 5635–5643.

[31] M.P. Hall, R.J. Nagel, W.S. Fagg, L. Shiue, M.S. Cline, R.J. Perriman, J.P. Donohue, M. Ares Jr., RNA 19 (5) (2013) 627–638.

[32] C. Rothrock, B. Cannon, B. Hahm, K.W. Lynch, Mol. Cell 12 (5) (2003) 1317–1324.

[33] C. Zhang, Z. Zhang, J. Castle, S. Sun, J. Johnson, A.R. Krainer, M.Q. Zhang, Genes Dev. 22 (18) (2008) 2550–2563.

[34] T. Tajiri, X. Liu, P.M. Thompson, S. Tanaka, S. Suita, H. Zhao, J.M. Maris, G.C. Prendergast, M.D. Hogarty, Clin Cancer Res. 9 (9) (2003) 3345–3355.

[35] R. Karni, E. de Stanchina, S.W. Lowe, R. Sinha, D. Mu, A.R. Krainer, Nat. Struct. Mol. Biol. 14 (3) (2007) 185–193.

[36] D. Ray, H. Kazan, K.B. Cook, M.T. Weirauch, H.S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L.H. Matzat, R.K. Dale, S.A. Smith, C.A. Yarosh, S.M. Kelly, B. Nabet, D. Mecenas, W. Li, R.S. Laishram, M. Qiao, H.D. Lipshitz, F. Piano, A.H. Corbett, R.P. Carstens, B.J. Frey, R.A. Anderson, K.W. Lynch, L.O.F. Penalva, E.P. Lei, A.G. Fraser, B.J. Blencowe, Q.D. Morris, T.R. Hughes, Nature 499 (7457) (2013) 172–177.

[37] C. Zhang, M.A. Frias, A. Mele, M. Ruggiu, T. Eom, C.B. Marney, H. Wang, D.D. Licatalosi, J.J. Fak, R.B. Darnell, Science 329 (5990) (2010) 439–443.

[38] E.T. Wang, N.A.L. Cody, S. Jog, M. Biancolella, T.T. Wang, D.J. Treacy, S. Luo, G.P. Schroth, D.E. Housman, S. Reddy, E. Lécuyer, C.B. Burge, Cell 150 (4) (2012) 710–724.

[39] M.J. Hicks, W.F. Mueller, P.J. Shepard, K.J. Hertel, Mol. Cell. Biol. 30 (8) (2010) 1878–1886.