Title

The Genetics of Splicing in Neuroblastoma

Journal

Cancer Discovery, 5(4)

ISSN

2159-8274

Authors

Chen, Justin

Hackett, Christopher S

Zhang, Shile

et al.

Publication Date

2015-04-01

DOI

10.1158/2159-8290.cd-14-0892

Peer reviewed

HHS Public Access

Author manuscript

Cancer Discov. Author manuscript; available in PMC 2015 October 01.

Published in final edited form as:

Cancer Discov. 2015 April ; 5(4): 380–395. doi:10.1158/2159-8290.CD-14-0892.

The genetics of splicing in neuroblastoma

Justin Chen[1,2], **Christopher S. Hackett**[2], **Shile Zhang**[3,4], **Young K. Song**[4], **Robert J.A. Bell**[1,5], **Annette M. Molinaro**[2,5,7], **David A. Quigley**[5,6], **Allan Balmain**[5], **Jun S. Song**[7,8], **Joseph F. Costello**[5], **W. Clay Gustafson**[6], **Terry Van Dyke**[9], **Pui-Yan Kwok**[10,11,12], **Javed Khan**[4], and **William A. Weiss**[2,6]

[1]Biomedical Sciences Graduate Program, University of California, San Francisco, CA 94158 USA

[2]Department of Neurology and Neurosurgery, University of California, San Francisco, CA 94158 USA

[3]Program in Bioinformatics, Boston University, Boston, MA, USA

[4]Oncogenomics Section, Pediatric Oncology Branch, National Cancer Institute, Bethesda, MD 20892, USA

[5]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA 94158 USA

[6]Institute for Cancer Research, Oslo, Norway

[7]Department of Epidemiology and Biostatistics, [6]Pediatrics, University of California, San Francisco, CA 94158 USA

[8]Departments of Bioengineering and Physics, University of Illinois, Urbana-Champaign, Urbana, IL 61801, USA

[9]Mouse Cancer Genetics Program, Center for Advanced Preclinical Research, National Cancer Institute, Frederick, MD 21702, USA

[10]Institute for Human Genetics, University of California, San Francisco, CA 94158 USA

[11]Dermatology, University of California, San Francisco, CA 94158 USA

[12]Cardiovascular Research Institute, University of California, San Francisco, CA 94158 USA

Abstract

Regulation of mRNA splicing, a critical and tightly regulated cellular function, underlies the majority of proteomic diversity, and is frequently disrupted in disease. Using an integrative

Disclosure Declaration

The authors do not have any conflicts of interest.

Data Access

Backcross array data and genotyping data have been archived in the NCBI Gene Expression Omnibus (Accession: GSE55248).

Author Contributions

JC performed the majority of the experiments and data analysis. CSH harvested tissues, genotyped the mice, and assisted with expression profiling. SZ, YKS, and JK performed NB RNA-Seq. YKS, TVD, and JK provided support for expression array data generation. DQ and AB assisted with e/sQTL analysis. PYK provided support for genotyping. RJAB, JSS, and JFC provided support for TCGA WGS analysis. AMM provided statistical assistance. WCG helped design experiments and prepare the manuscript. WAW supervised the project. JC and WAW designed experiments and wrote the paper.

genomics approach, we combined both genome and exon level transcriptome data in two somatic tissues (cerebella and peripheral ganglia) from a transgenic mouse model of neuroblastoma, a tumor that arises from peripheral neural crest. Here we describe splicing quantitative trait loci (sQTL) associated with differential splicing across the genome that we use to identify genes with previously unknown functions within the splicing pathway and to define *de novo* intronic splicing motifs that influence splicing from hundreds of bases away. Our results show that these splicing motifs represent sites for functional recurrent mutations and highlight novel candidate genes in human cancers, including childhood neuroblastoma.

Keywords

Alternative splicing; neuroblastoma; sQTL; MYC; FUBP1

Introduction

Alternative splicing of mRNA is a highly conserved process that is subject to both genetic regulation and heritability (1). Some of these regulatory systems act in *cis* within the primary sequence of the pre-mRNA transcript, while others act in *trans* via genetically distant factors recruited to the splice site (2). Alternative splicing may be particularly important to cancer, as the unique cancer environment selects for novel splice isoforms that promote tumor growth, metastasis, or response to treatment (3). Additionally, recurrent somatic mutations in known splicing factors, including *U2AF1* (4) and *SF3B1* (5), implicate functional contributions of this pathway in cancer and have led to interest in these factors as targets for cancer therapy (6).

Neuroblastoma is the most common cancer of infancy and the most common extracranial solid tumor of childhood. Extensive whole-genome and whole exome studies, including sequencing analyses of over 300 tumors, have identified point mutations in genetic drivers of neuroblastoma (*MYCN*, *ALK*, *PHOX2B*, *ARID1A/B*, and *ATRX)* in only a minority of patients (7). Genome-wide association studies of high-risk neuroblastoma, however, have identified a robust signal at the *BARD1* locus where the risk allele is associated with functional effects of *BARD1* splicing (8). Coupled with the identification of differences in splicing between stage 1 and stage 4 disease in neuroblastoma patients (9), alternative splicing has the potential to be a major contributor to this disease.

We used an integrative genomics approach to survey alternative splicing in neuroblastoma, combining both genome and transcriptome data into a single analysis. Linkage mapping, by identifying associations between genotypes and phenotypes in a genetically-controlled cohort, can identify genomic regions with functional importance. This type of approach, when adapted to high-throughput technologies and used to query expression quantitative trait loci (eQTL), represents a powerful tool to discover genetic mechanisms governing gene expression (10). We applied an extension of this concept, a splicing quantitative trait locus (sQTL) analysis (11–16), in a defined backcrossed mouse system using a genetically engineered model of neuroblastoma (17). By comparing two somatic neural tissues, our sQTL analysis uncovered a complex genome-wide splicing landscape, including the

identification of novel *trans*-acting splicing-related genes. Coupled with available whole-genome sequencing data of the parental strains, we also identified novel candidate intronic splicing motifs. We found that these motifs serve as sites for recurrent somatic mutations in human cancer that lead to functional changes in alternative splicing. We also identify a strain-specific triplet splicing preference within *FUBP1* that leads to upregulation of MYC, with functional consequences in human neuroblastoma.

Results

sQTL Distribute Throughout the Genome

FVB/NJ mice transgenic for *TH-MYCN* were backcrossed to wild-type 129/SVJ mice, and the N1 generation (n=102) was profiled on Affymetrix Exon Arrays and genotyped at 349 SNP and microsatellite markers. We identified 1664 and 1751 sQTL (defined here as a paired alternative splicing event associated with a marker, as markers may have multiple associations - see Methods) in cerebellum (CB) and superior cervical ganglia (SCG), representing peripheral neural crest- and brain-derived tissues, respectively (Figures 1A and 1B, 5% false detection rate). The low density of our genotyping panel reflects the controlled genetic heterogeneity of our backcrossed cohort and was not intended to identify causative polymorphisms. Instead, the resulting genetic map allowed us to distinguish splicing events with local *cis* effects from those with distal *trans* effects. The majority of sQTL was within 50 Mb of the spliced transcript, and thus defined to be *cis* (90.3% in CB and 92.5% in SCG, Supplemental Tables 1 and 2). Of these *cis*-sQTL, 874 reflected similar splicing events in both CB and SCG whereas only 5 *trans*-sQTL (mapping at least 50 Mb away from the spliced transcript or to a different chromosome) overlapped between tissues (Figure 1C, Supplemental Table 3). The polymorphisms of both of these types of sQTL associate with strain-specific alternative splicing events, suggesting different possible functional roles. *Trans*-sQTL could represent polymorphisms within genes that influence the splicing pathway, which can then affect the splicing of other transcripts derived from a distal genomic region. Conversely, *cis*-sQTL represent polymorphisms at or near the gene being spliced that can directly alter canonical splice sites or create novel transcriptional start sites, cryptic splice sites, and polyadenylation signals. These *cis*-sQTL polymorphisms could also reside in functional regions where they can affect the ability of intronic splicing enhancers (ISEs), intronic splicing silencers (ISSs), and their exonic counterparts, ESEs/ESSs to recruit accessory splicing factors to mediate alternative splicing.

Trans-sQTL Identify Genes that Influence Splicing

Trans-eQTL have become powerful tools to identify transcription factors (18) and other functional elements (19), as these function in *trans* to regulate transcription or translation of other genes. We therefore looked at putative *trans*-sQTL to identify novel regulators that could affect splicing of distant genes. If master regulators of splicing were subject to genetic control, we would expect to find a high degree of co-localization, whereby the majority of *trans*-sQTL (with alternatively spliced exons spread throughout the genome) would associate with just a few genomic loci. Instead, we found that t*rans*-sQTL were scattered throughout the genome (Figure 2A), with little overlap at any individual genomic locus. We observed a maximum of eight co-localizing *trans*-sQTL and only a few “hotspots,” or

individual loci linked to several splicing events (≥4, Figure 2B). We focused on two of these hotspots to simplify our efforts to validate genes that influence splicing. Specifically, we focused on the 5 *trans*-sQTL linked to the marker rs29347557 on chromosome 10 in SCG and the 8 *trans*-sQTL linked to the marker rs33477935 on the X chromosome in CB.

Candidate genes in these regions were identified by examining differential transcript expression between SCG and CB (Figure 2B). Given the overwhelming tissue-specific nature of the *trans*-sQTL, we hypothesized their origin to be from tissue-specific regulation of putative splicing-related genes in these regions. The minimal overlapping 95% confidence interval for sQTL mapping to rs29347557 in SCG encompassed the range between markers rs13480474 and rs38621064 on chromosome 10. This region spans over 13 Mb and contains 66 known genes, of which 30 were differentially expressed between CB and SCG (Supplemental Table 4). Almost all of these genes had known functions. Among these, *Sf3b5* encodes a splicing factor subunit (20) and was the only gene known to function within the splicing pathway, supporting the idea that differentially expressed splicing machinery would reside in these loci (Figure 2B). The sQTL mapping to rs33477935 in CB possessed 95% confidence intervals that minimally overlapped from rs33478059 to rs13483805 on the X chromosome. This region spans over 77 Mb and contains 489 known genes, 123 of which were differentially expressed (Supplemental Table 5). Two of these genes are known splicing factors, *Hnrnpa3* (21), and *Rbm10* (22). Others, such as *Rbmx,* are also implicated in this pathway. We identified putative candidate genes *Rbmx2,* which contains an RNA-binding motif, and *Ddx26b,* which contains a DEAD-box domain. Both of these genes are relatively uncharacterized (Figure 2B).

To examine the role these genes could play within the splicing pathway and to extend these results into a human setting, we used siRNA to knock down *DDX26B* and *RBMX2* in HEK293T cells (Supplemental Figure S1A) and looked for splicing differences in the orthologous human genes associated with the co-localized *trans*-sQTL, by RT-PCR. We looked specifically for alternative products by eliminating amplification of the canonical isoform. Cloning and sequencing of RT-PCR products from five of these eight targets did not support any changes in splicing patterns (*Ccnb2, Cdc123, Fbxl13, Ndufv2, Saps2,* data not shown), whereas *C5* (orthologous to *AI182371*) and *FBXW12* (orthologous to *Fbxw15*) indicated changes in splicing following knockdown (Supplemental Table 4 and Supplemental Figure S1B – we were unable to investigate the final target due to the lack of annotation). While it appeared that knockdown of *DDX26B* and *RBMX2* both affected the RT-PCR profile of *C5,* we examined all potential alternative products and discarded those arising from non-specific priming. Specifically in the presence of siRNA directed towards *RBMX2,* we identified a novel splice isoform of *C5* lacking six internal exons (Supplemental Figure S1C). Similarly, a previously undescribed splice isoform of *FBXW12* lacking an internal exon was expressed only when *DDX26B* was knocked down (Supplemental Figure S1D). These isoforms were undetected in the presence of non-targeting control siRNA. These data support functional roles for *DDX26B* and *RBMX2* within the splicing pathway. We did not identify a statistically significant difference in *Ddx26b* or *Rbmx2* expression between the parental mouse strains (Supplemental Figure S1E), suggesting that the difference in their abilities to affect splicing does not lie at the transcriptional level.

Cis-sQTL Identify Strain-Specific Isoforms

In both CB and SCG, a *cis*-sQTL identified strain-specific isoform expression of fifth exon within the Astroactin 2 (*Astn2*) gene. This was associated with a LOD score of 65.1 at rs13477756 (Supplemental Figures S2A and B), the highest observed in CB. *Astn2* may function in neuronal migration, and exonic deletions have been identified in patients with schizophrenia (23). RT-PCR on cDNA derived from CB in both parental strains confirmed that 129/SvJ mice specifically expressed a transcript resulting from exon skipping of the fifth exon (Supplemental Figures S2C and D).

We examined whole genome sequence information for parental FVB/NJ (24) and 129/SvJ (25) at the *Astn2* locus, which revealed a lack of coverage throughout exon 5 specifically in 129/SvJ, leading us to hypothesize that this alternative isoform could result from structural variation between the two strains. Indeed, a ~30 kb deletion present in 129/SvJ, but not in FVB/NJ, co-localized with *ASTN2* and encompassed the entirety of exon 5 (Supplemental Figure S2E). After excluding the X chromosome due to the mixed source for the 129/SvJ WGS, we further identified 1,324 somatic copy number variations (SCNVs) in 129/SvJ mice and 1,802 SCNVs in FVB/NJ compared to the reference genome (Supplemental Figure S3A and S3B). Fifty-five of these co-localized with *cis*-sQTL-identified alternative splicing events in 129/SvJ and fifty-eight co-localized in FVB/NJ, with 38 shared between strains (Supplemental Table 6). Expression of the alternative isoform in 129/SvJ is thus a result from genomic structural variation closely linked to the rs13477756 marker. Despite this alternative mechanism, this sensitive detection of strain-specific exon usage in *Astn2* by a SNP marker residing over 2 MB away from the gene itself is a clear proof of principle for our sQTL method.

The highest LOD score observed in SCG (LOD=73.9) belonged to the *FUSE*-Binding Protein 1 (*Fubp1*) gene (Figure 3A). *Fubp1* is a transcription factor that binds the Far Upstream Element (*FUSE*), thereby regulating transcription of *MYC* (26,27). Given the identification of cancer-associated somatic mutations in *FUBP1* (28) and the known importance of MYC proteins in neuroblastoma (29), we further explored *Fubp1*. Although array data indicated a strain-specific splicing difference at the fifth exon (Figure 3B), complete skipping of this exon was not detected by RT-PCR (Supplemental Figure S4). We therefore explored other models through which splicing could modulate *Fubp1* activity. Exon 5 within *Fubp1* is a known site of triplet NAGNAG splicing (30), which occurs at the terminal end of 3' splice sites. This repetition includes the highly conserved AG dinucleotide that defines the intron-exon boundary, resulting in alternative 3' splice-sites that differ by three bases that are referred to by their respective intron lengths. For *Fubp1*, this splicing event incorporates a serine at position 97 in the proximal isoform (shorter intron, $Fubp1^{97S}$) that is absent in the distal (longer intron, $Fubp1^{97-}$) (Figure 3C). Expression analyses of these isoforms across several tissue types suggests that this splicing event is not a stochastic process, but rather subject to an as-yet-undefined regulatory mechanism.

Gene-level expression datasets supported tumor suppressive effects of *FUBP1* in human neuroblastoma, as higher expression correlated with better survival in all stages (31)

($p=1.48\times10^{-4}$, $FDR=2.70\times10^{-3}$, Supplemental Figure S5A) and event-free survival in a *MYCN* non-amplified cohort (32) ($p=7.48\times10^{-8}$, $FDR=6.51\times10^{-6}$, Supplemental Figure S5B). Since these data do not have sufficient resolution to evaluate triplet splicing of exon 5, we next analyzed RNA-seq data to quantify $FUBP1^{97-}$:$FUBP1^{97S}$ ratios in human neuroblastoma. Given the inverse relationship between MYC and MYCN (33), the poor survival of neuroblastoma patients expressing MYC (34), and the known interaction between *MYC* and *FUBP1*, we looked specifically at *MYCN* non-amplified patients. Median event-free survival was 669 days in patients who expressed higher $FUBP1^{97-}$:$FUBP1^{97S}$ ratios as compared to 966 days for those with low ratios ($p = 0.0486$, Figure 3D). This indicated that higher $FUBP1^{97-}$ expression relative to $FUBP1^{97S}$ was correlated with poor event-free survival in *MYCN* non-amplified disease, consistent with recent evidence that triplet splicing is subject to regulation and has functional consequences (30,35). We next examined isoform-specific effects in human *MYCN* non-amplified neuroblastoma cell lines. We transduced lentivirus with V5-tagged constructs of $FUBP1^{97S}$ or $FUBP1^{97-}$ cDNA into SHEP and SK-N-AS cell lines. Overexpression of $FUBP1^{97-}$, but not $FUBP1^{97S}$, led to higher MYC levels (Figure 3E), consistent with deregulation of triplet splicing within exon 5 of *FUBP1* having a functional effect.

Splicing Motifs Serve as Sites of Recurrent Mutations in Cancer

Our genotyping marker set was not designed to identify functional polymorphisms. Rather, markers were selected to map chromosomal regions to either parental strain. As *cis*-sQTL represent differences in splicing driven by local genetic polymorphisms, we compared whole genome sequence information for FVB/NJ (24) and 129/SvJ (25) surrounding alternatively spliced exons possessing *cis*-sQTL in both tissues, identifying recurrent intronic decamer motifs that spanned the SNP. Enrichment analysis of these sequences identified 22 unique motifs (E-Value < 0.05, Supplemental Figure S6). Seven were similar to previously described motifs, with six known to bind splicing factors (Supplemental Figure S7, FDR < 0.05) (36). Thus, our method identified motifs associated with regulation of alternative splicing.

Collectively, these motifs matched 0.07% of all possible decamer sequence permutations. Despite this low rate, our position-agnostic approach identified an abundance of splicing motifs in both mouse and human introns (94.7% of murine introns and 95.2% of human introns harbored at least one of these splicing motifs), leading us to hypothesize that these splicing motifs could be functionally conserved in humans, regardless of their relative intronic positions. It has also been estimated that as much as 87% of genes have conserved splicing patterns between mouse and human orthologs (37). To see if these splicing motifs could represent functional sites commonly mutated in human cancer, we queried whole genome sequencing of 40 neuroblastoma tumors with matched normal sequence (38). Three candidate genes, *DYNLRB1*, *CGB7*, and *MOXD2P*, were identified as having an enrichment of somatic mutations compared to germline variation in intronic splicing motifs, each occurring in 2 samples (5% of the cohort, Figures 4A–C, respectively). For *DYNLRB1* and the chr7:141943454 splicing motif mutation in *MOXD2P*, these mutations resulted in a stronger match to a splicing motif by replacing a base found infrequently (or not at all) to a base more commonly associated with the motif. In the case of *CGB7* and the

chr7:141944631 splicing motif mutation in *MOXD2P*, the motif is effectively destroyed by substituting in a base that is not normally found.

The low incidence of recurrent genes with splicing motif mutations was unsurprising, given the low mutation rate in neuroblastoma (7). We similarly analyzed whole-genome sequencing from 42 high-risk glioblastoma (GBM) tumors, each also with a paired normal sample (39,40), and found a higher rate of recurrence. We identified three candidate genes, *LOC100505811*, *TRAM2-AS1*, and *NPIPA1* with enrichment of somatic mutations compared to germline variants in intronic splicing motifs in at least three samples (7.1% –11.1% of the cohort, Supplemental Figures S8A–C, respectively). Eleven additional genes were recurrent in at least 2 samples (4.7% of the cohort, Table 1).

Splicing Motif Mutations Lead to Functional Changes in Alternative Splicing

With RNA-seq data available for 25 of the 42 samples, we analyzed the TCGA GBM dataset further. Each of the recurrent mutations identified in GBM occurred at highly conserved positions within the splicing motif and either substituted a nucleotide that was not normally observed at that position (*LOC100505811* and *NPIPA1*), or replaced a nucleotide with the conserved nucleotide (*TRAM2-AS1*), creating a match to the splicing motif. To see if these mutations changed the splicing of these genes, we profiled normalized exon expression for the exons immediately proximal to the introns with splicing motif mutations. While the trend in differences of normalized exon expression between the two exons in *LOC100505811* was not statistically significant (Supplemental Figure S8A), the splicing motif mutations in *TRAM2-AS1* and *NPIPA1* were associated with a significant decrease in normalized exon expression for the exons downstream of the intron (reduction in means by 12.87%, $p = 0.02248$, and 27.62%, $p = 0.03037$, Supplemental Figures S8B and C, respectively). Thus, these mutations are located in sites correlated with changes in alternative splicing.

We continued to examine the neuroblastoma splicing motif mutations in *MOXD2P* and *CGB7*, along with the *NPIPA1* (chr16:15040359) and *TRAM2-AS1* splicing motif mutations identified in GBM. We cloned these introns containing the mutations into the pGINT splicing reporter where the introns separated the eGFP coding sequence into two exons, and we used qPCR to empirically assess canonical splicing efficiencies in SHEP human neuroblastoma cells after a 16 hour transfection. We compared introns containing the wild-type allele, the splicing motif mutation, and a control mutation that converted the AG dinucleotide at the 3’ splice site to a GG dinucleotide. We hypothesized that introduction of the mutant allele in the splicing motif would alter the ability of the cells to efficiently splice the eGFP exons, with the control mutation eliminating canonical splicing altogether.

The *MOXD2P* splicing motif mutations at chr7:141943454 (A>C) and chr7:141944631 (C>T) are located 47 bp and 149 bp from the nearest intron-exon boundary, respectively. These mutant alleles led to 3-fold and 2-fold increases in splicing efficiency, respectively, when compared to the wild-type alleles (Figures 5A and 5B, $p=0.0317$ and $p=0.0015$). The *CGB7* splicing motif mutation at chr19:49558080 (G>C), however, did not lead to a statistically significant change in splicing efficiency, despite residing only 18 bp away from the nearest intron-exon boundary (Figure 5C).

The G>A mutation at chr16:15040359 within *NPIPA1* resides 520 bp away from the nearest intron-exon boundary in an intron approximately 3 kb in total length. This mutation led to a 14% increase of canonical splicing efficiency when compared to the wild-type allele (Figure 5D, p=0.0478). Similarly, the T>G mutation at chr6:52445058 within *TRAM2-AS1* is located 650 bp away from the nearest intron-exon boundary in a 3.3 kb intron. The mutant allele led to a decrease in splicing efficiency, resulting in 56% of the splicing efficiency of wild-type allele (Figure 5E, p=0.0375). All splicing reporters with a splice-site mutation had undetectable eGFP expression. These results confirm that splicing motifs, even when located beyond immediate exon-intron boundaries, can function to influence alternative splicing. It should be noted that enhancement or reduction in splicing efficiency was not associated directionally with either gain-of-function or loss-of-function mutations; this can be attributed to the creation of a match to a splicing motif could increase the affinity for either splicing silencers or splicing enhancers. Similarly, a mutation that destroys a match to a splicing motif could decrease the affinity for either silencers or enhancers, leading to unpredictable splicing effects.

Recurrent Splicing Motif Mutations Identify Novel Genes of Interest in Neuroblastoma

Given the genome-wide somatic mutation rates in these WGS cohorts (2.685/Mb and 6.082/Mb for neuroblastoma and GBM, respectively), the nucleotide-level recurrence of splicing motif mutations suggests a high degree of selection whereby these altered splicing profiles contribute to the development of disease. Three genes (*NPIPA1*, *DYNLRB1*, and *AIMP1*) had enriched splicing motif mutations in both neuroblastoma and GBM (Table 1), but these genes lacked research establishing their involvement in either disease and highlighted the need for further investigation.

While NPIPA1, nuclear pore complex interacting protein family--member A1, has an unknown function, both *DYNLRB1* and *AIMP1* encode proteins implicated in the pathogenesis of other cancers. DYNLRB1, dynein light chain roadblock-type 1, interacts with TGFβ signaling and is mutated in human ovarian, colorectal, and gastric cancers (41). Interestingly, in ovarian cancer, novel splice isoforms have been identified, the products of which result in a reduction in TGFβ signaling (42). Functional studies also implicate a role for DYNLRB1 in cell migration. AIMP1, aminoacyl tRNA synthetase complex-interacting multifunctional protein 1, is an intracellular protein that is subject to proteolytibc cleavage when present at the cell surface. Upon cleavage, the released 22 kDa peptide induces apoptosis, antagonizes angiogenesis, and can stimulate an inflammatory response in the tumor microenvironment. In contrast, at high intracellular concentrations, AIMP1 induces cell migration (43).

Although we can examine the genomic effects of these mutations with regards to splicing motifs, we cannot categorically define these as gain-of- or loss-of-function at the gene or protein level. As the results from our splicing reporter assays suggest, both the creation of and destruction of splicing motifs could facilitate and impair splicing, depending on the binding of splicing enhancers or splicing silencers. The introns containing the splicing motif mutations in *DYNLRB1* and *AIMP1* spanned over 8 kb and 10 kb, respectively, making them unamenable to *in vitro* splicing reporter assays. Additionally, any specific effects on splicing

are not necessarily predictive of downstream gene function, as alternatively spliced variants could have no function or are immediately degraded by nonsense mediated decay (possibly leading to no effect or haploinsufficiency), act as a dominant negative, or even show constitutive activity.

To investigate the relevance of these genes in neuroblastoma, we examined gene expression and survival across two publically available neuroblastoma datasets containing either 87 untreated primary neuroblastoma tumors across all stages (31) or 102 untreated MYCN non-amplified primary neuroblastoma tumors (32). Across both datasets, higher *DYNLRB1* expression correlated with better overall survival and event-free survival ($p=8.40\times10^{-3}$, $FDR=9.97\times10^{-3}$, and $p=4.49\times10^{-5}$, $FDR=6.83\times10^{-4}$, respectively, Figures 6A and 6B), suggesting a tumor suppressive effect, whereas higher *AIMP1* expression correlated with worse overall survival and event-free survival ($p=5.62\times10^{-4}$, FDR=0.0157, and $p=3.16\times10^{-4}$, $FDR=2.49\times10^{-3}$, respectively, Figures 6C and 6D), indicating an oncogenic effect. Higher *NPIPA1* expression was associated with better event-free survival in neuroblastoma as well, statistically significant only in the *MYCN* non-amplified dataset ($p=2.10\times10^{-3}$, $FDR=3.49\times10^{-3}$, Figure 6E; p=N.S., Figure 6F). These data suggest that splicing motif mutations undergo selection in genes with clinical relevance and can be useful to identify novel targets in disease.

Discussion

In this study we leveraged a mouse genetic system to decode and map genetic regulation of splicing. We identified over 2500 putative sQTL, consistent with alternative splicing being subject to a greater degree of genetic control (16) than indicated by the earliest studies (<400 sQTL) (11–14). Furthermore, by using the TH-MYCN neuroblastoma mouse model, we were able to examine splicing in somatic tissue during the course of neuroblastoma development and oncogenic overexpression. Although several splicing factors are known direct targets of MYC, we have observed no changes in transgene expression between the parental strains, suggesting genetic differences between strains are the primary drivers of the sQTL. As the conservation of splicing between mouse and human has been estimated to be up to 87% (37), we looked to see if our results could have relevance in a human setting.

With the growing appreciation that alternative splicing can play functional roles in cancer, splicing factors have garnered considerable interest as potential targets for therapy (6). In this regard, our analysis enabled us to focus on only a few genes in each region. Although this analysis was limited by relatively large 95% confidence intervals stemming from the backcrossed nature of our cohort, we were nevertheless able to utilize this approach to identify critical loci harboring genes previously unknown to affect splicing. We did not find a high degree of *trans*-sQTL co-localization, sites that might be indicative of splicing master regulators. This could reflect a rarity of such *trans*-sQTL overall, or might alternatively be attributed to a limitation within our system, as polymorphisms in such a critical component of the splicing process might not be tolerated in either strain. Our analysis reveals an abundance of *trans*-acting genes that influence splicing and act in a limited gene-to-gene scope in a highly tissue specific fashion. While the genes we have identified could act as bone fide splicing factors and bind directly to RNA or the spliceosome, they could also act

indirectly by affecting the expression or activity of other splicing factors. Nevertheless, these data suggest a greater complexity to the splicing pathway than what has been already defined by mass spectrometry experiments focused on the spliceosome (44). It should be noted that our *trans*-sQTL analysis is not comprehensive; by specifically examining genes within the confidence interval that are differentially expressed between tissues, we were not able to identify genes with activities that are modified by post-translational modifications and microRNAs that could regulate the expression of other splicing factors.

FUBP1 has been reported to be a tumor suppressor (28) and an oncogene (45) with the ability to both activate and repress MYC transcription (46). The strain-specific splicing we observe suggests distinct functions for alternatively spliced *FUBP1* isoforms. Of the two triplet splicing products, FUBP1^{97-} led to an increase of MYC whereas FUBP1^{97S} reduced MYC levels. Our RNA-seq-based survival analysis (Figure 3D) represented an interim analysis of the TARGET program and is currently under-powered to examine the relationship between isoform expression and disease stage. These data raise the possibility that *FUBP1* mutations observed in cancers like oligodendroglioma are not strictly loss of function, but may act in a dominant negative fashion akin to FUBP1^{97-}. Despite the relationship between MYC and FUBP1 and the critical importance of MYC proteins in neuroblastoma tumorigenesis (29,34), extensive genomic sequence analysis in neuroblastoma tumors has yet to show recurrent mutations in *FUBP1* or *MYC*. In the absence of recurrent mutation, our analysis reveals splicing differences in *FUBP1*, not detectable by conventional expression analysis, which associate with risk.

While lacking in whole-exome sequencing analyses, intronic mutations in key regulatory motifs are potentially just as disruptive as mutations in coding regions. Our analysis identified novel splicing motifs that reside along the length of the intron. Previous splicing analyses have focused on exonic splicing motifs, using methods such as analyzing evolutionary conservation or systematic evolution of ligands by exponential enrichment (SELEX) on known and purified splicing factors. Until now, the most definitive analysis of intronic splicing enhancers utilized a splicing reporter construct to test the effects of random decamers inserted into the intron (47). Our unbiased analysis is the first of which we are aware to identify recurrent intronic splicing motifs *in vivo*.

We identified and established relevance for these splicing motifs by analyzing WGS from two different human tumor types, identifying novel candidate genes with functional implications in the process. The TCGA WGS cohort provided us with additional insights into the effects of these mutations, as a portion of the samples were accompanied by RNA-seq data. Despite this, not all samples had both transcriptomic and genomic data available, limiting our analysis of normalized exon expression for the *NPIPA1* splicing motif mutation to just three tumors. Nevertheless, four out of five *in vitro* splicing reporters containing these introns with splicing motif mutations confirm functional significance, showing that even single nucleotide changes hundreds of bases away from the nearest intron-exon boundary could impact mRNA splicing. These splicing motifs may function to bind accessory splicing factors to promote exon skipping, enhance exon inclusion, or even drive truncated transcripts through intron-retention or alternative polyadenylation. They may also alter splicing by changing the transcript's secondary structure; without highly accurate RNA

folding predictions, this analysis remains challenging and uncertain. It is also possible that these splicing motifs reside in non-coding RNA, which could in turn regulate expression of splicing factors.

The recurrent somatic mutations we identified in neuroblastoma and GBM might be easily dismissed due to their intronic nature, but our analysis extends understanding of the cancer genome by identifying functional ramifications for these intronic sites. The somatic mutations we identify within splicing motifs represent just one avenue whereby cancers may benefit from the deregulation of these gene products. Our analysis also broadens the applicability of these data beyond the limited individuals who possess these specific mutations by identifying novel candidate tumor suppressors and oncogenes in neuroblastoma.

Methods

Generation of a heterogeneous cohort backcrossed mice

Mice were obtained from The Jackson Laboratory (Bar Harbor, ME) and were housed and treated following UCSF IACUC guidelines. FVB/NJ mice transgenic for *TH-MYCN* were bred to wild-type 129/SvJ mice. Transgenic animals from the resulting F1 generation were identical genetically, possessing an allele for each gene from both parental strains. F1 mice were then backcrossed to 129/SvJ mice to generate the N1 generation used in this analysis. Superior cervical ganglia (SCG) and cerebellum (CB) were surgically isolated and snap-frozen in liquid nitrogen (n=102).

Genotyping

DNA was isolated from spleen tissue using a proteinase K lysis followed by phenol chloroform extraction. Microsatellite marker genotyping was carried out by the Marshfield Clinic (Marshfield, WI), and CIDR (Baltimore, MD). SNP genotyping was performed using template-directed primer extension with fluorescence polarization detection (FP-TDI, Acycloprime II, Perkin Elmer, Waltham, MA) and SNPStream 48-plex (Beckman Coulter, Brea, CA). Markers and map positions are shown in Supplemental Table 7. The marker set had an average spacing of 8 MB genome-wide (excluding the high density of markers on chromosome 10).

RNA Isolation

RNA was isolated from SCG using the RNeasy kit (QIAGEN, Valencia, CA). All other RNA isolations utilized Trizol (Invitrogen/Life Technologies, Grand Island, NY) for phase-separation prior to purification with the RNeasy kit.

Expression Arrays

1 μg of RNA was used as a starting template for RiboMinus rRNA subtraction (Invitrogen/Life Technologies) followed by the ST labeling protocol (Affymetrix, Santa Clara, CA). Labeled samples were hybridized to Affymetrix Mouse Exon 1.0 arrays. Arrays were normalized using RMA in the XPS Bioconductor package within R at both the exon level

and the transcript level using core probes. Differential expression between tissue types at the transcript level was examined using Significance Analysis of Microarrays (SAM) (48).

sQTL Analysis

Exon expression was normalized to gene expression by calculating normalized exon expression, a ratio of exon expression to gene expression (both values determined using the XPS Bioconductor package). sQTL were calculated using normalized exon expression as a quantitative trait in eQTL software as previously described (49). Briefly, linkage between normalized exon expression and loci was assessed by linear regression with genome-wide significance determined using an FDR < 0.05. Due to the ability of exonic polymorphisms to function as ESEs or ESSs, we did not exclude exons that harbored SNPs between the two strains. *Trans*-sQTL were drawn using Circos (50). LOD scores and related confidence intervals were drawn and calculated using the R-QTL package within R.

Mouse WGS Analysis

FVB/NJ and 129/SvJ whole-genome sequencing data were downloaded from the ENA (Accession: ERP000687) and SRA (Accession: SRX205921), respectively, as raw FASTQ files. Reads were mapped to the reference genome (mm9) using Bowtie2 (v. 2.1.0) using the "—fast-local" preset. Genotypes were called using the GATK UnifiedGenotyper package (v. 2.4–9). SNPs were compared using the VCFtools package (v. 1.10). Consensus sequences for both strains were generated using the samtools mpileup function. Copy number was assessed using FREEC (v. 6.3) (51) with a 5kb window at 1 kb intervals and a breakpoint threshold of 0.4. Results from the X chromosome in 129/SvJ were discarded as WGS was performed on a mixed pool of 6 mice.

Splicing Motifs

Spliced exons were identified as possessing a *cis*-sQTL in both tissues, and the direction of splicing in each strain was noted. 19-mer sequences surrounding polymorphic nucleotides and their positions relative to the spliced exon were acquired in the direction of transcription from consensus FASTA sequences generated from strain-specific WGS. Sequences were trimmed to exclude exonic bases, and regions where either strain returned a poly(N) motif indicative of a lack of sequencing coverage were discarded. Extracted sequences were binned based on the direction of the associated splicing and their origin (5' intron or 3' intron).

MEME (52) was used to identify recurrent motifs with a width of at least 10 bases in these bins using a background GC content of 42%. Motifs with an E-Value < 0.05 and derived from at least 20 sequences were reported as significant. Redundant motifs were identified as those that shared a Pearson correlation coefficient > 0.60, and the representative motif was chosen with the lowest E-Value. TOMTOM (52) was used to compare sets of identified motifs at an FDR < 0.05 using Pearson correlation coefficients and requiring a minimum overlap of five bases. Cumulative matching of splicing motifs were determined by generating a multi-FASTA file consisting of 1,048,576 possible decamer permuations. 743, or 0.07% of these sequences matched the splicing motifs when compared to this file using MAST (52) at a sequence p-value threshold of 0.0001. Introns containing splicing motif

matches were calculated by extracting unique RefSeq introns from the UCSC hg19 build (n=195,956) and mm9 assemblies of FVB/NJ and 129/SvJ (n=184,637) created in this study. Splicing motif matches were determined using FIMO (52) at a p-value of 0.0001. The number of unique intron matches found in FVB/NJ (174,842) and 129/SvJ (174,746) were averaged when representing the percentage of murine introns containing at least a single splicing motif.

Motif Analysis of Whole-Genome Sequencing

Primary neuroblastoma and patient matched normal DNA WGS data for 40 patients from the St. Jude Children's Hospital – Washington University Pediatric Cancer Genome Project were downloaded from the European Genome-Phenome Archive (Dataset ID: EGAD00001000135). Primary GBM and patient matched normal DNA WGS data for 42 patients were downloaded from The Cancer Genome Atlas project's Cancer Genomics Hub (www.cghub.ucsc.edu, dbGAP Study Accession number: phs000178) (39,40).

Single-nucleotide variants (SNVs) were detected with MuTect, a Bayesian framework for the detection of somatic mutations (53). Somatic and germline SNVs were filtered according to MuTect defaults (germline variants were kept at a LOD(N) threshold of 2.3). SNVs for TCGA GBM samples aligned to hg18 were converted to hg19 coordinates using the UCSC liftover tool (54). For technical reasons, one GBM sample only had SNV calls from chromosomes 1–7, and another GBM sample only had SNV calls from chromosomes 1–12. The overall somatic mutation rates were calculated by dividing the total number of somatic mutations by the total number of "callable bases" identified by MuTect. The GBM samples with missing SNV calls were omitted from the calculation of somatic mutation rates.

Intronic sequences were extracted from the hg19 reference genome encompassing 9 bp upstream and downstream of the identified in the 5'-3' direction of transcription. Reference sequences in addition to sequences containing the alternative allele were analyzed for splicing motif occurrences using MAST at a sequence p-value threshold of 0.0001, recording the best motif match to either the reference or alternative sequence as determined by the lowest sequence p-value.

GBM RNA-Seq Analysis

Open-access level 3 RNA-Seq data for 25 samples that overlapped with the 42 WGS GBM samples were downloaded through the TCGA Data Portal (55). In the absence of transcript expression levels, normalized exon expression was calculated by taking the ratio of individual exon expression (RPKM) to the maximally expressed exon within the gene. Significance was determined by Student's t-Test.

Neuroblastoma RNA-Seq and Survival Analysis

RNA-Seq was performed on samples collected for the TARGET program (currently unpublished) and represent an interim analysis of these data. The cohort consisted of all high risk and stage 4S tumors (n=20), and included stage 3 tumors (n=2) that had similar tumor aggressiveness to stage 4.

Total RNA was extracted from the fresh frozen tumor samples by the Trizol/RNeasy kit protocol described previously (56). Whole transcriptome libraries for Illumina HiSeq 2000 were prepared according to the TruSeq RNA protocol, in which poly-A mRNA is purified from total RNA at the initial step. Two indexed samples were pooled and sequenced on an Illumina HiSeq 2000 with 100-bp paired end. Quality of RNA and library was assessed by an Agilent BioAnalyzer.

100-bp paired end reads were first aligned to the reference human genome (hg19) using spliced read mapper Tophat 2.0.8 (57). We then used Cufflinks 2.1.1 (58) for gene abundance estimation. Abundances in fragments per kilobase of exon per million fragments mapped (FPKM) were calculated for each annotated RefSeq gene. Kaplan-Meier analysis was performed using GraphPad Prism. High and low distal:proximal isoform ratios were determined by a median split. Significance was determined by the Log-rank test.

Splicing Reporters

The region of interest from *NPIPA1* was PCR-amplified using AccuPrime Taq High Fidelity (Invitrogen/Life Technologies) from genomic DNA obtained from HEK293T cells using gNPIPA1, gTRAM2-AS1, gMOXD2P, gCGB7 Cloning-F and Cloning-R primers (Integrated DNA Technologies (IDT), Coralville, Iowa, Supplemental Table 8) and TOPO-cloned into the PCR2.1-TOPO vector (Invitrogen/Life Technologies). The pGINT fluorescent splicing reporter was obtain from Addgene (Cambridge, MA, Plasmid 24217) and used as a template for PCR amplification of the GFP exons using the GFPEX1-F/R and GFPEX2-F/R primers (IDT) and Phusion High-Fidelity PCR Master Mix (Thermo Scientific) according to manufacturer's instructions. GFP fragments containing specific introns were generated via PCR using the GFPEX1-F, GFPEX2-R, and the appropriate Intron-F/R primer pair (IDT). The GFP fragment containing the second MOXD2P intron (containing the chr7:141944631 mutation) was generated through PCR amplification of the MOXD2P-2 gBLOCK (IDT) and GFP Exon 2, using the GFPEX1-F, GFPEX2-R and the Phusion High-Fidelity PCR Master Mix. The GFP inserts were cloned into the NotI and XhoI restriction sites in the pGINT plasmid, replacing the modified adenovirus intron, using the Quick Ligation Kit (New England Biolabs, Ipswich, MA) according to manufacturer's instructions. Mutant alleles were obtained using the Quichange II XL Site-Directed Mutagenesis Kit (Stratagene/Agilent Technologies, Santa Clara, CA) and the appropriate SDM-F/R primers (IDT). Plasmids were verified using Sanger sequencing (Quintara Biosciences, Richmond, CA).

1 μg of each splicing reporter was transfected into SHEP cells with 2 μl of Lipofectamine 2000 (Invitrogen/Life Technologies) in individual wells of 24-well plates according to the manufacturer's instructions. RNA was harvested after 16 hours using Trizol (Invitrogen/Life Technologies) and the RNeasy mini kit (Qiagen). 500 ng of RNA was used as input for cDNA synthesis. Gene expression on 5× diluted cDNA was measured through quantitative real-time PCR using 5' nuclease assays, using the GINT and NeoR qPCR primers/probes (IDT, Supplemental Table 8) and the KAPA PROBE FAST qPCR Kit (Kapa Biosciences, Boston, MA). Reactions were run in triplicate on an ABI 7900HT Fast Real-Time PCR system (Applied Biosystems/Life Technologies) and analyzed using SDS 2.3 software.

Splicing efficiency was calculated using the ΔΔCT method, comparing GFP expression to the expression of the Neomycin resistance gene included on the pGINT plasmid to control for plasmid copy number.

Tissue Culture

All cell lines were obtained from the UCSF Cell Culture Facility but not reauthenticated by the authors. SHEP cells were grown in RPMI-1640 supplemented with 10% FBS. SK-N-AS and HEK293T cells were grown in DMEM supplemented with 10% FBS, non-essential amino acids, and sodium pyruvate. All cell lines were maintained at 37 Degrees with 5% CO_2 and tested for mycoplasma contamination by PCR.

Lentiviral Transduction

Human *FUBP1*$^{97-}$ cDNA was obtained from the IMAGE consortium (Accession: BC017247) and PCR-cloned into the pENTR-D/TOPO gateway vector (Invitrogen/Life Technologies) using the FUBP1 Cloning-F and FUBP1 Cloning-R primers (IDT, Supplemental Table 8) to add the directional TOPO motif and exclude a stop codon. This entry clone was used as a substrate to create an *FUBP1*^{97S} entry vector using the Quickchange II XL Site-Directed Mutagenesis Kit (Stratagene/Agilent Technologies) and Site-Directed Mutagenesis Primers FUBP197S SDM-F/FUBP197S SDM-R (IDT, Supplemental Table 8). Plasmid sequences were verified using Sanger sequencing (Quintara Biosciences) Lentiviral constructs were made by LR recombination (Invitrogen/Life Technologies) with the pLenti 6.3 destination vector (Invitrogen/Life Technologies). Virus was packaged in HEK293T cells using the ViraPower lentiviral packaging plasmid mix and Lipofectamine 2000 (Invitrogen/Life Technologies) over the course of 72 hours. Viral supernatant was harvested and filtered through a 0.45 μM syringe filter and used to directly transduce neuroblastoma cells for 24 hours. Stably transduced cells were selected with 10 μg/ml blasticidin (Invitrogen/Life Technologies).

Neuroblastoma Expression Profiling and Survival Analysis

Survival data were obtained from two datasets on the R2: microarray analysis and visualization platform (59). The Versteeg dataset (31) (GEO Accession ID: GSE16476) spanned all neuroblastoma stages and included 88 samples profiled on Affymetrix U133P2 expression arrays. The Seeger dataset (32) (GEO Accession ID: GSE3446) included 102 primary untreated neuroblastoma tumors without *MYCN* gene amplification profiled on Affymetrix HG-U133A and HG-U133B platforms. Expression data for both datasets were calculated using the MAS 5.0 algorithm. *FUBP1*, *DYNLRB1*, *AIMP1*, and *NPIPA1* expression were assayed using the following probes: 214093_s_at, 217918_at, 202542_s_at, and 214870_x_at. High and low expression groups were separated by scanning along the cohort with a minimum size of 8. Significance was assessed by the Log-Rank test. The false discovery rate was determined by the associated Q-Value.

Western Blotting

Cell lysate was harvested using Cell Lysis Buffer (Cell Signaling Technology, Danvers, MA) supplemented with protease inhibitor (Roche, Basel, Switzerland) and 1% SDS.

Protein was quantitated with the BCA assay kit (Pierce Biotechnology, Rockford, IL). Equal amounts of total protein were loaded and run on 4%–12% SDS-polyacrylamide gels (Invitrogen/Life Technologies) and transferred to PVDF membranes using the iBlot (Invitrogen/Life Technologies). After blocking (1 hour RT, 5% non-fat milk in TBS-T), membranes were incubated overnight (4 degrees, 5% BSA in TBS-T) with V5-specific antibody (1:5000, Invitrogen/Life Technologies), GAPDH-specific antibody (1:10000, Millipore, Billerica, MA), or C-MYC-specific antibody (XRP, 1:1000, Cell Signaling Technology). Antibodies were detected with HRP-linked mouse or rabbit (Calbiochem/Millipore) secondary antibodies followed by enhanced chemilluminescence (Amersham/GE, Piscataway, NJ).

cDNA Synthesis and RT-PCR

cDNA synthesis was performed using SuperScript VILO MasterMix (Invitrogen/Life Technologies) according to the manufacturer's instructions. 1 μg of total RNA was used as starting material in the reverse transcriptase reaction.

RT-PCR to determine retention of *FUBP1* Exon 5 and validate the alternative *ASTN2* isoform was performed using GoTaq Green MasterMix (Promega, Madison, WI) in 25 ul reactions according to the manufacturer's instructions using 1 μL of cDNA and FUBP1-1F/FUBP1-7R or Astn2-F/R primers, respectively (IDT, Supplemental Table 8). PCR products were run on a 1.5% agarose gel and Sanger sequenced (Quintara Biosciences). Sequence analysis and chromatogram generation was performed with Geneious (Auckland, New Zealand).

Quantitative-RT-PCR was performed using 1 μL of a 1:10 dilution of cDNA using the KAPA SYBR FAST qPCR Kit (Kapa Biosystems) according to manufacturer's instructions and the appropriate primer sets (IDT, Supplemental Table 8). Reactions were run in triplicate on an ABI 7900HT Fast Real-Time PCR system (Applied Biosystems/Life Technologies) and analyzed using SDS 2.3 software.

Knockdown of Putative Trans-sQTL Splicing-Related Genes

siRNA targeting human *DDX26B* and *RBMX2* were purchased as ON-TARGETplus SMARTpools (Dharmacon/Thermo Scientific Bio, Pittsburgh, PA). ON-TARGETplus Non-Targeting siRNA #1 (D-001810-01-05) was used as a negative control. 140 pmol of siRNA was transfected into HEK293T cells using Lipofectamine 2000 for a period of 72 hours, after which RNA was harvested using Trizol.

cDNA was synthesized and RT-PCR was performed using primers C5 RT-F/RT-R and FBXW12 RT-F/RT-R (IDT, Supplemental Table 8). The targeted amplicons for *C5* and *FBXW12* spanned exons 4–21 and exons 4–10, respectively. Products were run on a 1.5% agarose gel, and bands indicating the molecular weight of the full-length isoform was excised. The remaining gel products were extracted using the Gel-Extraction kit (Qiagen) and re-amplified using the same primers. The final products were analyzed on a 2% agarose gel. Differentially expressed products were gel-extracted and TOPO cloned into PCR2.1-TOPO vectors (Invitrogen/Life Technologies). Products were identified by Sanger

sequencing (Quintara Biosciences) using M13 Forward and M13 Reverse primers and matched to the genome using BLAT (60).

Genomic Coordinates

All genomic coordinates are representative of either mm9 or hg19 reference builds.

Statistical Analysis

All statistics were performed using the R Project for Statistical Computing unless otherwise noted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank M. Huang, L. McHenry, E. Simonds, S. Ilkhanizadeh, and D. Levens for reagents, experimental advice and helpful discussions.

Grant Support

WAW was supported by NIH grants CA176287, CA82103, CA102321, CA148699, CA159859, CA081403; and the Katie Dougherty, Pediatric Brain Tumor, St. Baldricks, and Samuel G. Waxman Foundations, and a CureSearch Grand Challenge Award. RJAB and JSS were supported by NIH R01CA163336 and the Sontag Foundation Distinguished Scientist Award. Computational analysis was performed using the HDFCC TIPCC, supported by NCI 5P30CA082103-15.

Abbreviations

sQTL	Splicing quantitative trait locus
SCG	Superior cervical ganglia
CB	Cerebellum
NEE	Normalized exon expression
GV	Germline variant
SM	Somatic mutation

References

1. Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, et al. Heritability of alternative splicing in the human genome. Genome Res. 2007; 17:1210–1218. [PubMed: 17671095]
2. Garcia-Blanco, M.; Baraniak, A.; Lasda, E.; Baraniak, E. Nat Biotechnol. Vol. 22. Nature Publishing Group; 2004. Alternative splicing in disease and therapy; p. 535-546.
3. Chen, J.; Weiss, WA. Oncogene. Macmillan Publishers Limited; 2014. Alternative splicing in cancer: implications for biology and therapy.
4. Brooks, AN.; Choi, PS.; de Waal, L.; Sharifnia, T.; Imielinski, M.; Saksena, G., et al. A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events. In: Ast, G., editor. PLoS One. Vol. 9. Public Library of Science; 2014. p. e87361

5. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011; 478:64–69. [PubMed: 21909114]
6. Bonnal, S.; Vigevani, L.; Valcárcel, J.; Vigevani, J. Nat Rev Drug Discov. Vol. 11. Nature Publishing Group; 2012. The spliceosome as a target of novel antitumour drugs; p. 847-859.
7. Lawrence, MS.; Stojanov, P.; Polak, P.; Kryukov, GV.; Cibulskis, K.; Sivachenko, A., et al. Nature. Vol. 499. Nature Publishing Group, a division of Macmillan Publishers Limited; 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes; p. 214-218.All Rights Reserved.
8. Bosse K, Diskin S, Cole K, Wood A, Schnepp R, Norris G, et al. Common Variation at BARD1 Results in the Expression of an Oncogenic Isoform That Influences Neuroblastoma Susceptibility and Oncogenicity. Cancer Res. 2012; 72:2068–2078. [PubMed: 22350409]
9. Guo, X.; Branch, P.; Chen, Q-R.; Song, Y.; Wei, J., et al. SAIC-Frederick. BMC Med Genomics. Vol. 4. BioMed Central Ltd; 2011. Exon array analysis reveals neuroblastoma tumors have distinct alternative splicing patterns according to stage and MYCN amplification status; p. 35
10. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science. 2002; 296:752–755. (80−). [PubMed: 11923494]
11. Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
12. Kwan, T.; Benovoy, D.; Dias, C.; Gurd, S.; Provencher, C.; Beaulieu, P., et al. Nat Genet. Vol. 40. Nature Publishing Group; 2008. Genome-wide analysis of transcript isoform variation in humans; p. 225-231.
13. Heinzen, E.; Ge, D.; Cronin, K.; Maia, J.; Shianna, K.; Gabriel, W., et al. PLOS Biol. Vol. 6. Public Library of Science; 2008. Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits; p. e1000001
14. Fraser, H.; Xie, X.; Fraser, H.; Xie, X. Genome Res. Cold Spring Harbor Lab; 2009. Common polymorphic transcript variation in human disease.
15. Montgomery S, Sammeth M, Gutierrez-Arcelus M, Lach R, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
16. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014; 24:14–24. [PubMed: 24092820]
17. Weiss WA, Aldape K, Mohapatra G, Feuerstein BG, Bishop JM. Targeted expression of MYCN causes neuroblastoma in transgenic mice. EMBO J. 1997; 16:2985–2995. [PubMed: 9214616]
18. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, Kong A, et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet. 2011; 43:561–564. [PubMed: 21572415]
19. Westra, H-J.; Peters, MJ.; Esko, T.; Yaghootkar, H.; Schurmann, C.; Kettunen, J., et al. Nat Genet. Vol. 45. Nature Publishing Group, a division of Macmillan Publishers Limited; 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations; p. 1238-1243.All Rights Reserved.
20. Will, CL. EMBO J. Vol. 21. EMBO Press; 2002. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein; p. 4978-4988.
21. Jurica, MS.; Licklider, LJ.; Gygi, SR.; Grigorieff, N.; Moore, MJ. RNA. Vol. 8. Cold Spring Harbor Lab; 2002. Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis; p. 426-439.
22. Inoue A, Yamamoto N, Kimura M, Nishio K, Yamane H, Nakajima K. RBM10 regulates alternative splicing. FEBS Lett. 2014
23. Vrijenhoek, T.; Buizer-Voskamp, J.; Stelt, I.; Strengman, E.; Sabatti, C.; Kessel, A., et al. Am J Hum Genet. Vol. 83. Elsevier; 2008. Recurrent CNVs Disrupt Three Candidate Genes in Schizophrenia Patients; p. 504-510.

24. Wong, K.; Bumpstead, S.; Weyden, L.; Reinholdt, L.; Wilming, L.; Adams, D., et al. Genome Biol. Vol. 13. BioMed Central Ltd; 2012. Sequencing and characterization of the FVB/NJ mouse genome; p. R72
25. Wartman L, Larson D, Xiang Z, Ding L, Chen K, Lin L, et al. Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression. J Clin Invest. 2011
26. Duncan, R.; Bazar, L.; Michelotti, G.; Tomonaga, T.; Krutzsch, H.; Avigan, M., et al. Genes Dev. Vol. 8. Cold Spring Harbor Lab; 1994. A sequence-specific, single-strand binding protein activates the far upstream element of c-myc and defines a new DNA-binding motif; p. 465-480.
27. Braddock DT, Louis JM, Baber JL, Levens D, Clore GM. Structure and dynamics of KH domains from FBP bound to single-stranded DNA. Nature. 2002; 415:1–6.
28. Bettegowda C, Agrawal N, Jiao Y, Sausen M, Wood LD, Hruban RH, et al. Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science. 2011; 333:1453–1455. (80−). [PubMed: 21817013]
29. Huang M, Weiss WA. Neuroblastoma and MYCN. Cold Spring Harb Perspect Med. 2013; 3:a014415. [PubMed: 24086065]
30. Bradley, R.; Merkin, J.; Lambert, N.; Burge, C.; Cambridge, M. PLOS Biol. Vol. 10. Public Library of Science; 2012. Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution; p. e1001229
31. Molenaar JJ, Koster J, Zwijnenburg DA, Sluis P, Valentijn LJ, Ploeg I, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature. 2012; 483:589–593. [PubMed: 22367537]
32. Asgharzadeh S, Pique-Regi R, Sposto R, Wang H, Yang Y, Shimada H, et al. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. J Natl Cancer Inst. 2006; 98:1193–1203. [PubMed: 16954472]
33. Breit S, Schwab M. Suppression of MYC by high expression of NMYC in human neuroblastoma cells. J Neurosci Res. 1989; 24:21–28. [PubMed: 2810395]
34. Wang LL, Suganuma R, Ikegaki N, Tang X, Naranjo A, McGrady P, et al. Neuroblastoma of undifferentiated subtype, prognostic significance of prominent nucleolar formation, and MYC/MYCN protein expression: a report from the Children's Oncology Group. Cancer. 2013; 119:3718–3726. [PubMed: 23901000]
35. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet. 2004; 36:1255–1257. [PubMed: 15516930]
36. Ray, D.; Kazan, H.; Cook, KB.; Weirauch, MT.; Najafabadi, HS.; Li, X., et al. Nature. Vol. 499. Nature Publishing Group, a division of Macmillan Publishers Limited; 2013. A compendium of RNA-binding motifs for decoding gene regulation; p. 172-177.All Rights Reserved.
37. Zambelli F, Pavesi G, Gissi C, Horner DS, Pesole G. Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics. 2010; 11:534. [PubMed: 20920313]
38. Cheung, N-K.; Zhang, J.; Lu, C.; Parker, M.; Bahrami, A.; Tickoo, S., et al. JAMA. Vol. 307. American Medical Association; 2012. Association of Age at Diagnosis and Genetic Mutations in Patients With Neuroblastoma; p. 1062
39. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013; 155:462–477. [PubMed: 24120142]
40. TCGA Network TCGAR. Nature. Vol. 455. Macmillan Publishers Limited; 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways; p. 1061-1068.All rights reserved
41. Tang Q, Staub CM, Gao G, Jin Q, Wang Z, Ding W, et al. A novel transforming growth factor-beta receptor-interacting protein that is also a light chain of the motor protein dynein. Mol Biol Cell. 2002; 13:4484–4496. [PubMed: 12475967]
42. Ding W, Tang Q, Espina V, Liotta LA, Mauger DT, Mulder KM. A transforming growth factor-beta receptor-interacting protein frequently mutated in human ovarian cancer. Cancer Res. 2005; 65:6526–6533. [PubMed: 16061631]

43. Schwarz, MA.; Thornton, J.; Xu, H.; Awasthi, N.; Schwarz, RE. Cell proliferation and migration are modulated by Cdk-1-phosphorylated endothelial-monocyte activating polypeptide II. In: Singh, SR., editor. PLoS One. Vol. 7. Public Library of Science; 2012. p. e33101
44. Zhou Z, Licklider L, Gygi S, Reed R. Comprehensive proteomic analysis of the human spliceosome. Nature. 2002; 419:182–185. [PubMed: 12226669]
45. Malz M, Weber A, Singer S, Riehmer V, Bissinger M, Riener M-O, et al. Overexpression of far upstream element binding proteins: a mechanism regulating proliferation and migration in liver cancer cells. Hepatology. 2009; 50:1130–1139. [PubMed: 19585652]
46. Chung H-J, Liu J, Dundr M, Nie Z, Sanford S, Levens D. FBPs are calibrated molecular tools to adjust gene expression. Mol Cell Biol. 2006; 26:1–15.
47. Wang Y, Ma M, Xiao X, Wang Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat Struct Mol Biol. 2012; 19:1044–1052. [PubMed: 22983564]
48. Tusher G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci. 2001; 98:5116–5121. [PubMed: 11309499]
49. Quigley D, To M, Pérez-Losada J, Pelorosso F, Mao J-H, Nagase H, et al. Genetic architecture of mouse skin inflammation and tumour susceptibility. Nature. 2009; 458:505–508. [PubMed: 19136944]
50. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]
51. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012; 28:423–425. [PubMed: 22155870]
52. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37:W202–W208. [PubMed: 19458158]
53. Cibulskis, K.; Lawrence, MS.; Carter, SL.; Sivachenko, A.; Jaffe, D.; Sougnez, C., et al. Nat Biotechnol. Vol. 31. Nature Publishing Group, a division of Macmillan Publishers Limited; 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples; p. 213-219.All Rights Reserved.
54. UCSC LiftOver Tool. [Internet]. Available from: http://genome.ucsc.edu/cgi-bin/hgLiftOver.
55. TCGA Data Portal. [Internet]. Available from: https://tcga-data.nci.nih.gov/tcga/
56. Wei, J.; Khan, J. Purification of total RNA from mammalian cells and tissues. In: Bowtell, D.; Sambrook, J., editors. DNA microarrays a Mol clonging Man. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2002. p. 110-119.
57. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562–578. [PubMed: 22383036]
58. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
59. R2: microarray analysis and visualization platform. [Internet]. Available from: http://r2.amc.nl.
60. UCSC BLAT. [Internet]. Available from: https://genome.ucsc.edu/cgi-bin/hgBlat.

Statement of Significance

Somatic mutations with predictable downstream effects are largely relegated to coding regions, which comprise less than 2% of the human genome. Using an unbiased, *in vivo* analysis of a mouse model of neuroblastoma, we have identified intronic splicing motifs that translate into sites for recurrent somatic mutations in human cancers.

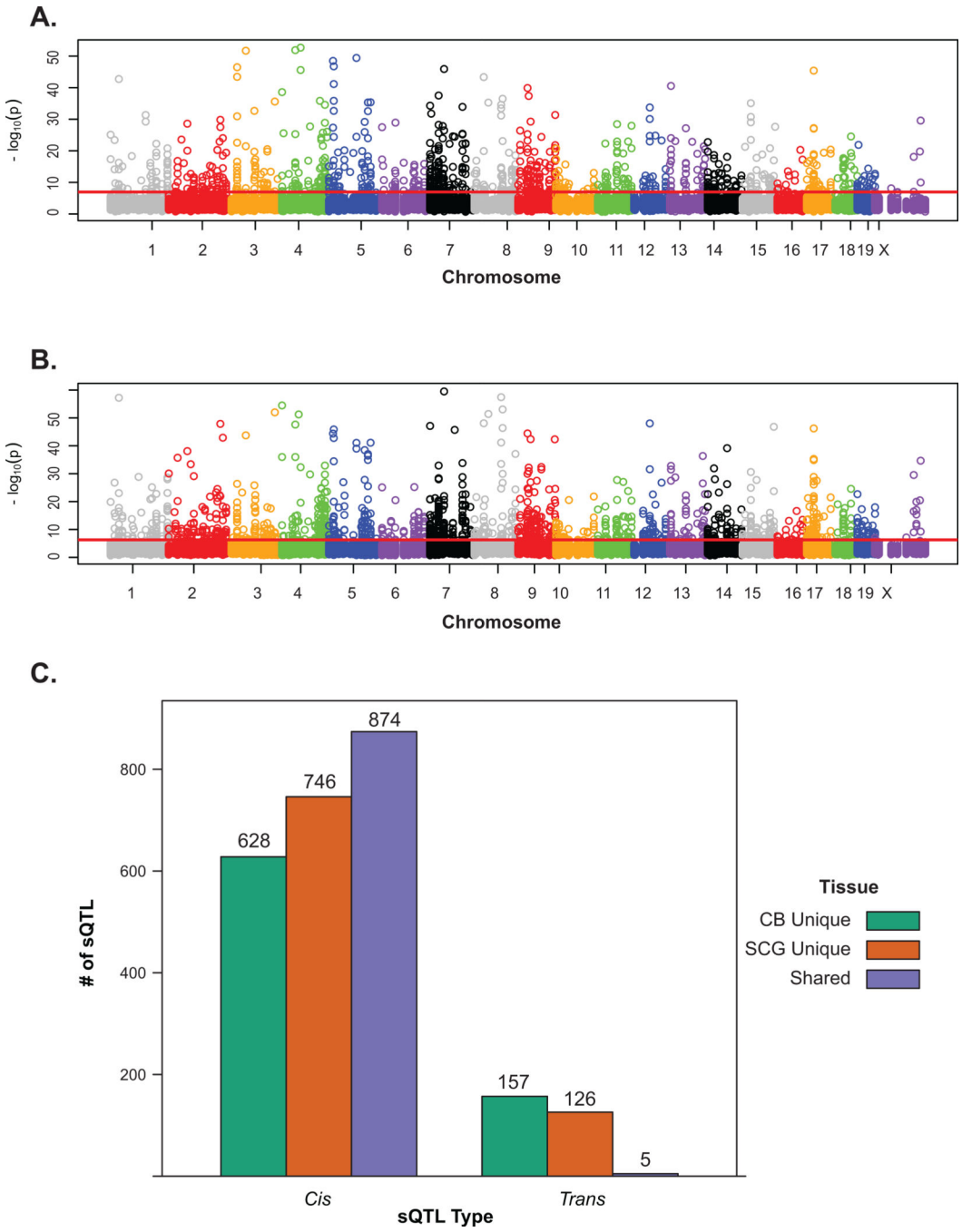

Figure 1. Genomic distribution of sQTL

sQTL are distributed throughout the genome similarly in CB (A) and SCG (B). The x-axis indicates the location of each SNP, and the y-axis indicates the level of significance of an associated sQTL. The horizontal red line is drawn to mark an estimation of genome-wide significance at an FDR < 0.05.

C) *Cis*-sQTL were more abundant than *trans*-sQTL, and a majority were shared between both tissues. In contrast, *trans*-sQTL were largely tissue specific.

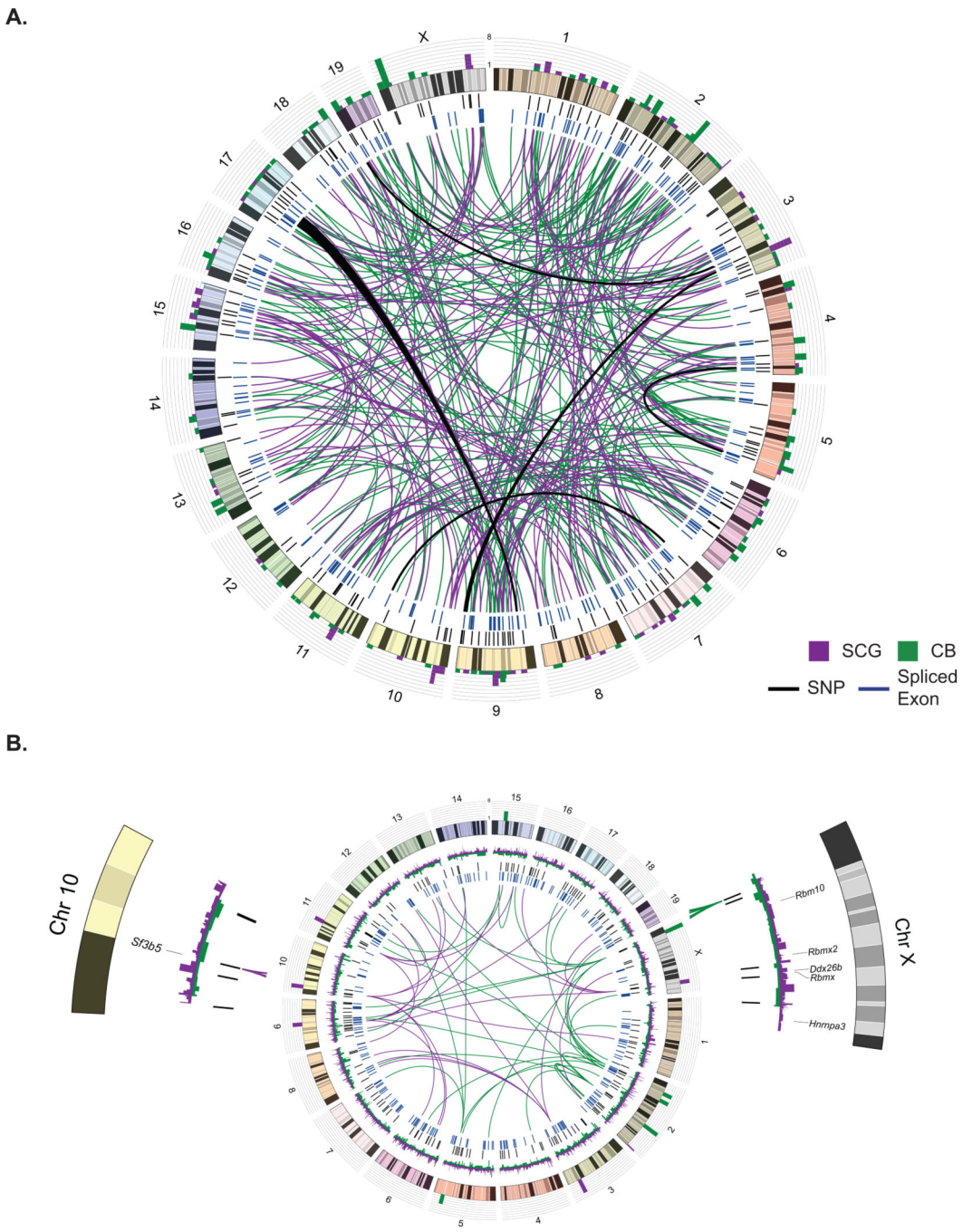

FIgure 2. *Trans*-sQTL in the cerebellum and SCG

A. Circos plot of 162 cerebella and 131 SCG *trans*-sQTL identified at an FDR < 0.05. Mouse chromosomes are plotted on the outer ring, with locations of SNPs with sQTL indicated in black on the first inner ring. Locations of spliced exons are identified in blue on the second inner ring. Each t*rans*-sQTL (paired association between SNP and alternatively spliced exon) is depicted by a line that links the exon and the SNP in SCG (purple), and cerebellum (green), with sQTL shared between both tissues drawn in black. The histogram

outside of the chromosomes indicates the number of sQTL (on a scale of 1–8) that co-localize at any given genomic locus.

B. Circos plot of sQTL hotspots, where 4 or more co-localized. An additional data track between the chromosome ideograms and SNP loci indicates differentially expressed genes at the transcript level (purple = higher expression in SCG, green = higher expression in CB). Five SCG-specific *trans*-sQTL mapped to rs29347557 on chromosome 10. This region included differential expression of *SF3B5*, encoding a splicing factor subunit. Eight CB-specific *trans*-sQTL mapped to rs33477935 on the X chromosome. Shown are the locations of the candidate genes that are differentially expressed at this locus.

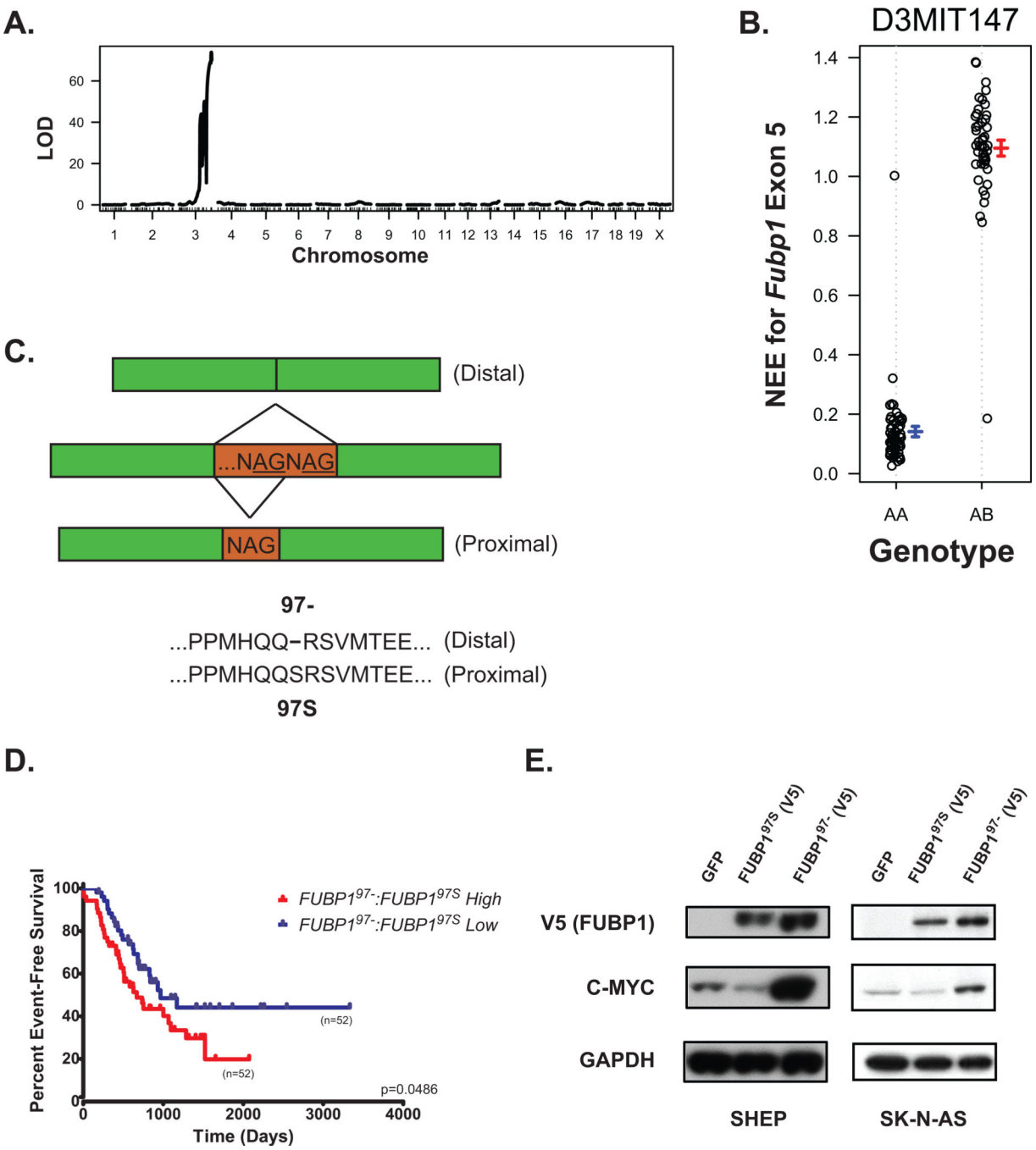

Figure 3. *Fubp1* possesses a *cis*-sQTL

A. The sQTL for *Fubp1* has a LOD score of 73.9 on chromosome 3 where the gene is located, indicating a *cis* effect.

B. Normalized Exon Expression (NEE) levels for *Fubp1* Exon 5 show loss of expression associated with the homozygous 129/SvJ allele.

C. Exon 5 of *FUBP1* is a site of triplet splicing. The distal isoform leads to incorporation of a serine at position 97 (*FUBP1*^{97S}) whereas the proximal isoform lacks the serine (*FUBP1*$^{97-}$).

D. *MYCN* non-amplified neuroblastoma patients with high $FUBP1^{97-}$:$FUBP1^{97S}$ ratios (red, n=52) had a reduced event free survival when compared to those with low $FUBP1^{97-}$:$FUBP1^{97S}$ ratios (blue, n=52, log-rank p = 0.0486). There was not a significant difference in survival across all neuroblastoma patients (data not shown, n =134).
E. Western Blot of human neuroblastoma cell lines transduced with GFP control, $FUBP1^{97S}$, or $FUBP1^{97-}$ lentivirus tagged with V5. MYC is upregulated in both SHEP and SK-N-AS lines in the presence of high $FUBP1^{97-}$ levels.

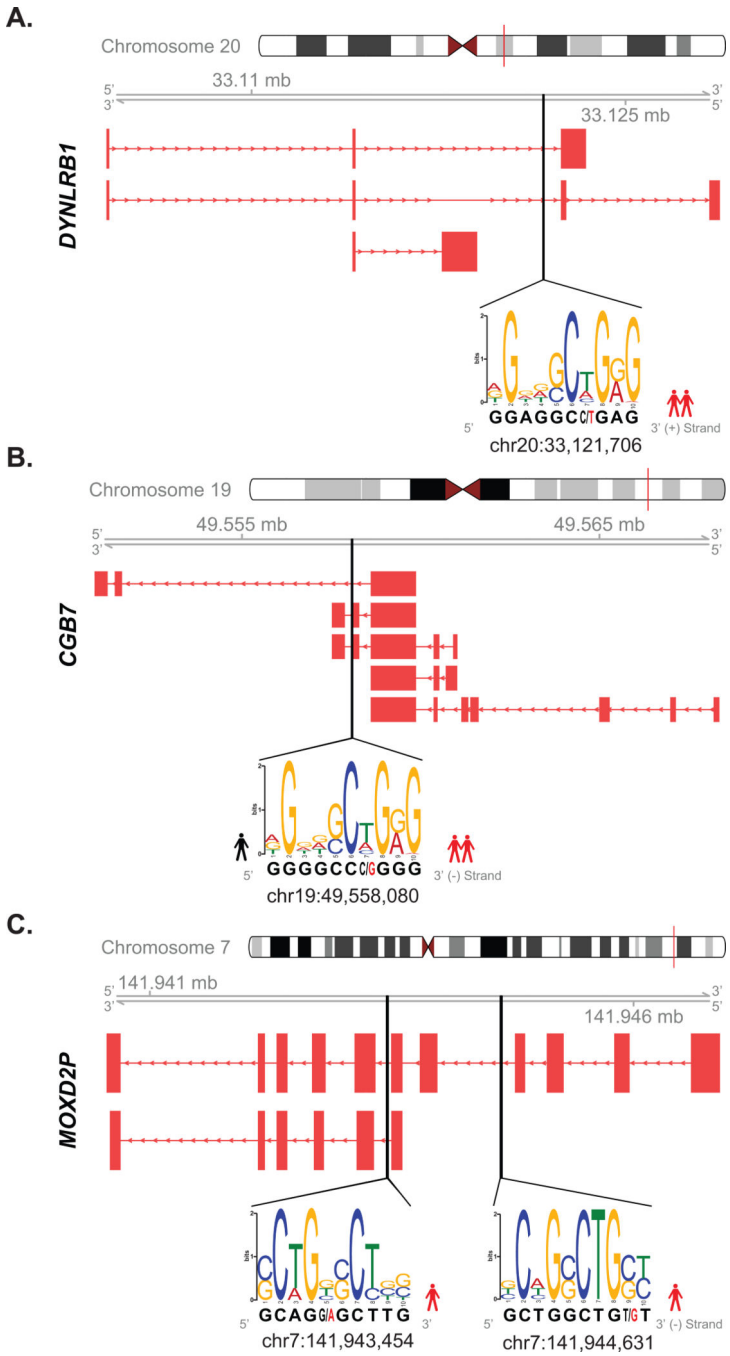

Figure 4. Recurrent somatic mutations occur in intronic splicing motifs in neuroblastoma

Analysis of 40 neuroblastoma samples reveals enrichment of somatic mutations in splicing motifs in *DYNLRB1* (A), *CGB7* (B), and *MOXD2P* (C). Germline variants (GVs) and somatic mutations (SMs) were identified by MuTect. MAST was used to examine the sequence immediately surrounding both intronic GVs and intronic SMs for a match to any of the 22 splicing motifs (sequence-level match $p < 0.0001$). The sequence logo of the splicing motif is drawn linked to the physical genomic location of the somatic mutation (black line). The total height of each nucleotide position is the information content in bits

and represents the level of conservation for that position. The height of each nucleotide letter represents the ratio that they are found to occupy that position. The gene structures of known isoforms are depicted in red with arrows indicating the direction of transcription. The reference sequence (black) is given directly beneath the sequence logo with the position and nucleotide of the mutant allele shown in red. Red silhouettes indicate the number of tumor samples with that particular mutation. A black silhouette indicates the allele was found in a normal sample.

A. Two neuroblastoma samples (5%) had a recurrent C>T mutation within *DYNLRB1* at chr20:33,121,706. This mutation created a stronger match to a splicing motif by substituting a base found infrequently to a base more commonly associated with the motif.

B. Two tumor samples had a recurrent G>C mutation at chr19:49,558,080 within *CGB7*; another neuroblastoma germline sample was also found to be G/C heterozygous at the same position. The splicing motif matched here was the same as observed in the *DYNLRB1* mutation, and when analyzed in the orientation of the corresponding transcript, this mutation effectively destroyed the match to the motif.

C. Two neuroblastoma samples were identified with distinct splicing motif mutations at chr7:141,944,631 (C>T) and chr7:141,943,454 (A>C) in *MOXD2P*. When analyzed in the orientation of the corresponding transcript, the mutation at chr7:141,944,631 destroyed a splicing motif match whereas the chr7:141,943,454 mutation resulted in creating a match to the splicing motif.

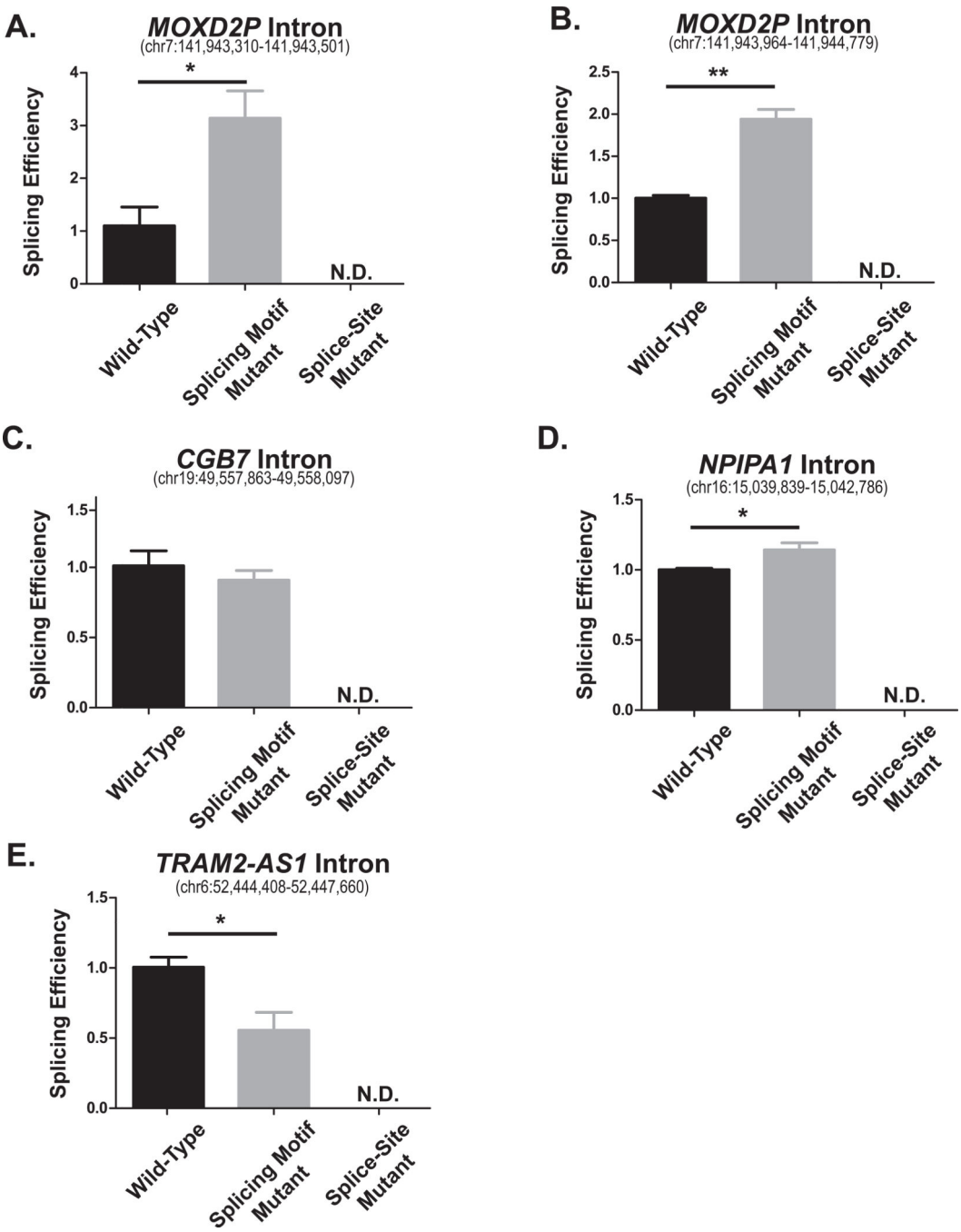

Figure 5. Recurrent somatic mutations in intronic splicing motifs affect splicing efficiency

A. The mutant splicing motif allele within *MOXD2P* at chr7:141943454 led to >3-fold increase in splicing efficiency (213.8%) compared to the wild-type allele ($p = 0.0317$).

B. The mutant splicing motif allele within *MOXD2P* at chr7:141944631 led to an increase in splicing efficiency (93.9%) compared to the wild-type allele ($p = 0.0015$).

C. No changes were detected between the mutant and wild-type splicing motif allele in the *CGB7* intron.

D. The mutant splicing motif allele within *NPIPA1* at chr16:15040359 led to a 14% increase in splicing efficiency compared to the wild-type allele (p=0.0478).
E. The mutant splicing motif allele at chr6:52445058 within *TRAM2-AS1* resulted in a decrease to 56% splicing efficiency compared to the wild-type allele (p=0.0375).
Splicing efficiency was measured as the relative expression of the canonical eGFP transcript compared to the expression of plasmid Neomycin resistance marker. Shown are representative means from at least 3 independent experiments. Error bars represent S.E.M. N.D. = Not Detected. Significance was assessed by Student's t test.

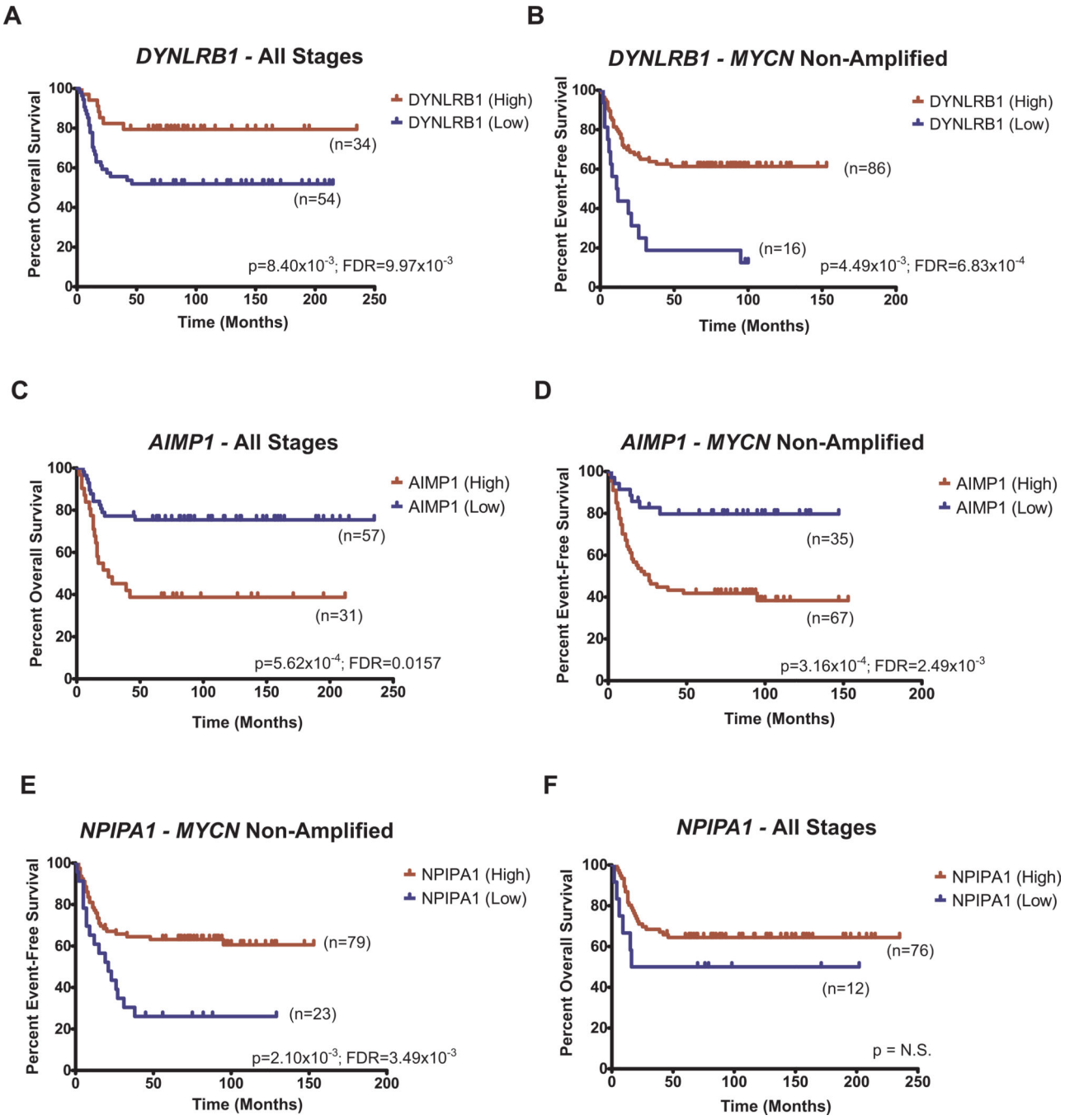

Figure 6. *DYNLRB1, AIMP1*, and *NPIPA1* expression correlate with survival in neuroblastoma

Kaplan-Meier analysis of genes containing recurrent splicing motif mutations in two datasets, one containing all stages of neuroblastoma and the other containing only *MYCN* non-amplified neuroblastoma, reveals low expression of *DYNLRB1* correlates with poor overall and event-free survival compared to high expression (A and B, $p=8.40\times10^{-3}$ and $p=4.49\times10^{-3}$). High *AIMP1* expression correlates with poor overall and event-free survival compared to low expression in both datasets (C and D, $p=5.62\times10^{-4}$ and $p=3.16\times10^{-4}$). Low expression of *NPIPA1* correlates with poor event-free survival compared to high

expression, but is only significant in the *MYCN* non-amplified cohort (E and F, $p=2.10\times10^{-3}$ and p=N.S.) False detection rates (FDR) are given for the identification of the high and low expression groups. Significance was assessed by the log-rank test.

Table 1

Recurrent genes with enriched splicing motif mutations

Enrichment was identified by taking the ratio of samples with somatic mutations (SMs) within splicing motifs to samples with germline variants (GVs) within splicing motifs.

Gene Name	GBM Splicing Motif GVs	GBM Splicing Motif SMs	Neuroblastoma Splicing Motif GVs	Neuroblastoma Splicing Motif SMs
MOXD2P	0	0	0	1-chr7:141944631 C>T 1-chr7:141943454 A>C
CGB7	0	0	1-chr19:49558080 G/C	2-chr19:49558080 G>C
DYNLRB1	1-chr20:33104296 C/G	2-chr20:33121706 C>T	0	2-chr20:33121706 C>T
NPIPA1	1-chr16:15040359 G/A	4-chr16:15040359 G>A 1-chr16:15031765 G>C	0	1-chr16:15040359 G>A
LOC100505811	0	3 chr5:117618623 C>T	0	0
TRAM2-AS1	0	3 chr6:52445058 T>G	0	0
AIMP1	0	2-chr4:107260426 G>T	0	1-chr4:107260426 G>T
PRDM12	0	2-chr9:133545135 G>A	0	0
TIGD7	0	2-chr16:3353982 A>G	0	0
FBXO45	0	2-chr3:196310237 C>T	0	0
SLC4A1	0	2-chr17:42342896 C>A	0	0
ARHGAP29	1-chr1:94677742 G/C	2-chr1:94678897 G>A	0	0
LOC100128006	1-chr17:12681207 G/A	1-chr17:12679656 T>G 1-chr17:12673734 C>T	0	0
LOC440896	1-chr9:69180408 G/A	1-chr9:69179798 T>C 1-chr9:69180519 C>G	0	0
STXBP5	1-chr6:147648825 G/C	2-chr6:147693778 T>C	0	0
TNNI2	1-chr11:1860346 T/G	1-chr11:1860346 T>G 1-chr11:1862029 G>A	0	0