Pan-cancer repository of validated natural and cryptic mRNA splicing mutations

Peter K. Rogan[1,3], Ben C. Shirley[3], Eliseos J. Mucaki[1], Joan H.M. Knoll[2,3]

University of Western Ontario Departments of Biochemistry[1] and Pathology and Laboratory Medicine[2], CytoGnomix Inc.[3] London, ON, Canada

Abstract

We present a major public resource of mRNA splicing mutations validated according to multiple lines of evidence of abnormal gene expression. Likely mutations present in all tumor types reported in the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) were identified based on the comparative strengths of splice sites in tumor versus normal genomes, and then validated by respectively comparing counts of splice junction spanning and abundance of transcript reads in RNA-Seq data from matched tissues and tumors lacking these mutations. The comprehensive resource features 341,486 of these validated mutations, the majority of which (69.9%) are not present in the Single Nucleotide Polymorphism Database (dbSNP 150). There are 131,347 unique mutations which weaken or abolish natural splice sites, and 222,071 mutations which strengthen cryptic splice sites (11,932 affect both simultaneously). 28,812 novel or rare flagged variants (with <1% population frequency in dbSNP) were observed in multiple tumor tissue types. Single variants or chromosome ranges can be queried using a Global Alliance for Genomics and Health (GA4GH)-compliant, web-based Beacon "Validated Splicing Mutations" either separately or in aggregate alongside other Beacons through the public Beacon Network (http://www.beacon-network.org/#/search?beacon=cytognomix), as well as through our website (http://validsplicemut.cytognomix.com/).

Introduction

Next generation sequencing continues to reveal large numbers of novel variants whose impact cannot be interpreted from curated variant databases, or through reviews of peer-reviewed biomedical literature[1]. This has created a largely unmet need for unequivocal sources of information regarding the molecular phenotypes and potential pathology of variants of unknown significance (VUS) in cancer genomes. Information theory (IT) has been proven to accurately predict impact of mutations on mRNA splicing, and has been used to interpret coding and non-coding mutations that alter mRNA splicing in both common and rare diseases[2,3]. We have described an IT-based framework for the interpretation and prioritization of non-coding variants of uncertain significance, which has been validated in multiple studies involving variants in patients with history or predisposition to heritable breast and/or ovarian cancer.[4]

The Cancer Genome Atlas (TCGA) Pan-Cancer Atlas is a comprehensive integrated genomic and transcriptomic resource containing data from >10,000 tumors across 33 different tumor types.[5] Here, we utilized IT-based tools for assessment of high quality sequenced variants in TCGA patients, as well as patients from tumor datasets provided by the International Cancer Genome Consortium (ICGC), for their potential impact on mRNA splicing.[6] The results of these genome-wide analyses are presented using an online internet resource (http://validsplicemut.cytognomix.com/; Figure 1 [below]) which can also be queried through the Beacon Network.[7]

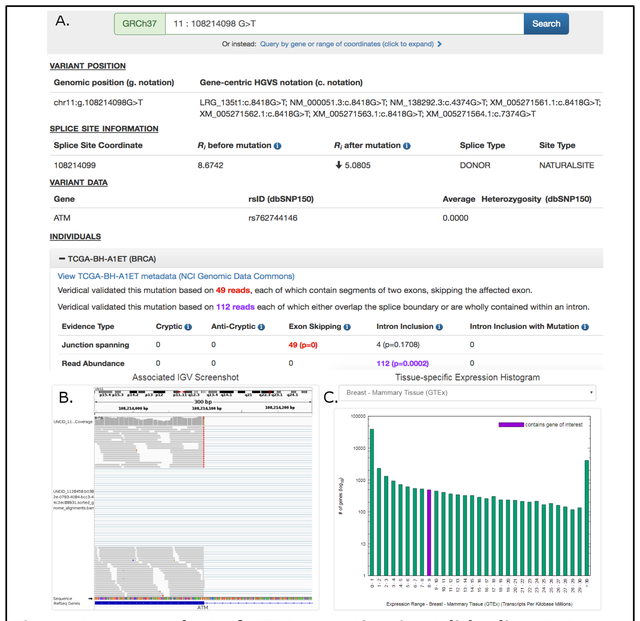


Figure 1. Screenshot of ATM mutation in ValidSpliceMut.
(**A**) Variant-specific and splice site-specific tabular results are presented under the headings "Splice Site Information" and "Variant Data". Results are organized by TCGA and ICGC sample IDs harboring the mutation within a series of expandable panels. Each panel consists of read counts and p-values by Veridical evidence type. Significant p-values (≤ 0.05) are bolded. (**B**) An integrative genome viewer (IGV) image showing alignment of expressed sequence reads. (**C**) A dynamically generated histogram presents expression levels of all genes for a selected normal tissue type.

Methods

TCGA and ICGC data acquisition and processing: Controlled-access data was obtained with permission from the Data Access Committee at NIH for TCGA and from ICGC. Patient RNA sequencing BAM files (tumor and normal, when available) and their associated VCF files were obtained from the CancerGenomeHub and Genomic Data Commons. ICGC data was downloaded through the Score client (v1.5.0). Additional VCFs from patients with congenital diseases were obtained from the Baylor Hopkins Center for Mendelian Genomics study (BH CMG; phs000711.v4.p1). Variants which did not pass quality control were not analyzed.

Information analysis and RNA-Seq validation of splicing variants: We used the Shannon Pipeline software (SP; applies IT to rapidly perform high-throughput, in silico prediction of variants impact on mRNA splicing) to analyze TCGA and ICGC variants (>168 million TCGA and >41 million ICGC variants) to evaluate their potential impact on splice site binding strength (change in information content, R_i, measured in bits). Veridical software analyzed the genomic variants by comparing the RNA-Seq alignment in the region surrounding the variant with the corresponding interval in control transcriptomes lacking the variant.[3] Veridical counts abnormally spliced reads in RNA-Seq data (molecular phenotypes such as cryptic site use, exon skipping, or intron inclusion), and determines the null hypothesis probability that the transformed read count corresponds to normal splicing ($p < 0.05$ is considered significant). Veridical-flagged mutations found in genes in the Catalogue Of Somatic Mutations In Cancer (COSMIC) Cancer Gene Census (CGC)[8] are highlighted, as are rare mutations also found in patients of congenital diseases from the BH CMG study (where literature supports the link of congenital disease to the affected gene)[9].

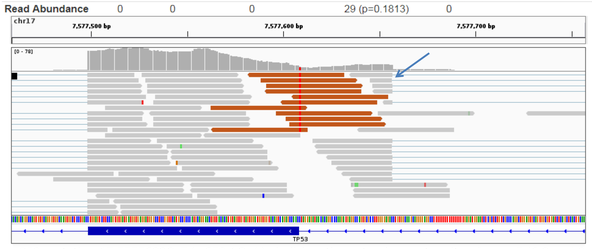
Development of the ValidSpliceMut database and Beacon: We created a publicly accessible Application Programming Interface (API) that can be utilized to programmatically query variants passing filter thresholds described above (https://beacon.cytognomix.com). It was built in accordance with the GA4GH Beacon v1.0.0 specification. We also developed the website ValidSpliceMut (Figure 1; below) to serve as a local interface to our Beacon, allowing users to manually search for a variant, by gene name or genome coordinate range.

Results

Examples of IT-predicted and RNAseq-validated mRNA Splicing Mutations in:

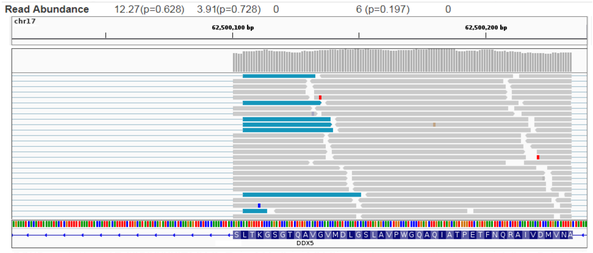
COSMIC Cancer Gene Census genes[8]

Evidence Type	Cryptic	Anti-Cryptic	Exon Skipping	Intron Inclusion	Intron Inclusion with Mutation
Junction spanning	0	0	0	12 (p=0.0002)	11 (p=0)
Read Abundance	0	0	0	29 (p=0.1813)	0



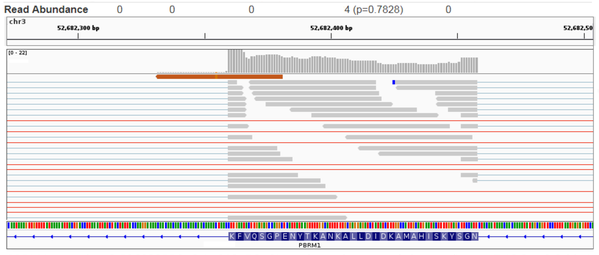
TP53 mutation 17:7577609C>T (c.673-1G>A) abolishes the natural acceptor site of exon 7 (6.0 > -4.9 bits), activating a cryptic site 49nt downstream (R_i = 5.2 bits; **blue** arrow). Veridical detects 11 reads with mutation which crosses the intron-exon junction (**brown**), and reports their number significant over controls (p≈0).

Evidence Type	Cryptic	Anti-Cryptic	Exon Skipping	Intron Inclusion	Intron Inclusion with Mutation
Junction spanning	66 (p=0)	0	0	2 (p=0.523)	2 (p=0.493)
Read Abundance	12.27(p=0.628)	3.91(p=0.728)	0	6 (p=0.197)	0



DDX5 mutation 17:62500098TACAG>T (c.441+2delACAG) creates a cryptic donor (-1.3 > 5.4 bits) within *DDX5* exon 4, which would lead to a 4nt deletion of exon 4 (**blue** reads). Veridical detects 66 reads which uses this cryptic donor site, and computes this number of reads significant over controls (p ≈ 0).

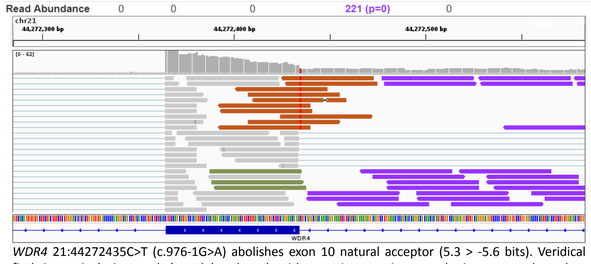
Evidence Type	Cryptic	Anti-Cryptic	Exon Skipping	Intron Inclusion	Intron Inclusion with Mutation
Junction spanning	0	0	9 (p=0)	2 (p=0.0784)	1(p=0)
Read Abundance	0	0	0	4 (p=0.7828)	0



PBRM1 mutation 3:52682355C>G (c.813+5G>C) weakens the natural donor of exon 7 (6.8 > 2.9 bits), which lead to a significant amount of increased exon skipping (**red** reads). Veridical detects 9 exon skipping reads over this region, and finds this number significant over controls (p ≈ 0).

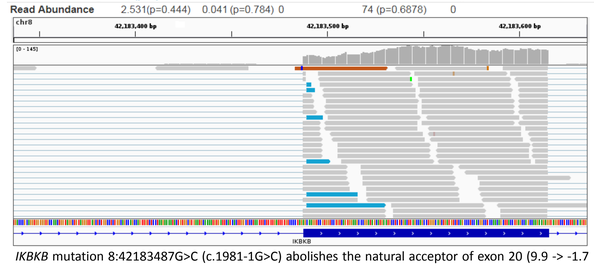
Baylor-Hopkins Centre for Medical Genomics study[9]

Evidence Type	Cryptic	Anti-Cryptic	Exon Skipping	Intron Inclusion	Intron Inclusion with Mutation
Junction spanning	0	0	0	12 (p=0.0046)	9 (p=0)
Read Abundance	0	0	0	221 (p=0)	0



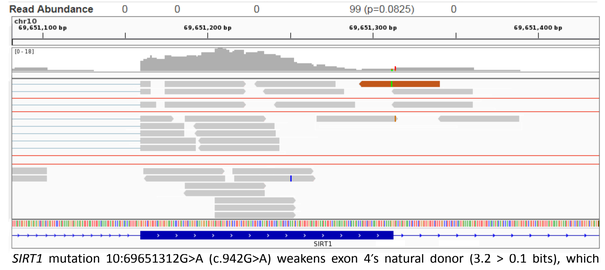
WDR4 21:44272435C>T (c.976-1G>A) abolishes exon 10 natural acceptor (5.3 > -5.6 bits). Veridical finds intron inclusion reads (**purple**) and reads with mutation crossing over the intron-exon boundary (**brown**) significant (p≈0). The mutation was found in LUAD patient TCGA-55-8085 as well as BH CMG patient BH2112_1 with https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0779-xsyndrome. *WDR4* has been linked to disorders with a microcephaly phenotype.

Evidence Type	Cryptic	Anti-Cryptic	Exon Skipping	Intron Inclusion	Intron Inclusion with Mutation
Junction spanning	10 (p=0)	0	0	1 (p=0.7447)	1 (p=0)
Read Abundance	2.531(p=0.444)	0.041 (p=0.784)	0	74 (p=0.6878)	0



IKBKB mutation 8:42183487G>C (c.1981-1G>C) abolishes the natural acceptor of exon 20 (9.9 -> -1.7 bits) while simultaneously strengthening a cryptic acceptor (2.0 > 10.8 bits), resulting in a 2nt deletion (**blue** reads). This mutation, identified in BLCA patient TCGA-FD-A3NA, was also found in a BH CMG patient with immunodeficiency (BH6169_1), which *IKBKB* has been linked to.

Evidence Type	Cryptic	Anti-Cryptic	Exon Skipping	Intron Inclusion	Intron Inclusion with Mutation
Junction spanning	0	0	4 (p=0.0032)	2 (p=0.1033)	1(p=0)
Read Abundance	0	0	0	99 (p=0.0825)	0



SIRT1 mutation 10:69651312G>A (c.942G>A) weakens exon 4's natural donor (3.2 > 0.1 bits), which resulted in a significant amount of exon skipping (**red** reads; p = 0.032). This mutation, found in SKCM patient TCGA-EE-A29V, was also found in BM CMG patient BH3973_1, with plantar lipomatosis. *SIRT1* has been linked to soft tissue tumors such as liposarcoma.

Table 1: Unique Flagged Variants from TCGA and ICGC Datasets

	Unique TCGA Variants / Genes	Unique ICGC Variants / Genes	Total
Total Variants (Passing QC)	> 168 million	> 41 million	>209 million
Met IT-Based Criteria	1,081,011 / 19,322	24,101 / 8,151	1,094,749 / 19,412
Weaken Natural Sites	373,770	11,662	380,852
Strengthen Cryptic Sites	744,848	12,547	752,472
Significant by Veridical (p < 0.05)	334,202 / 18,249	11,961 / 5,761	341,486 / 18,386
Flagged for Increased Exon Skipping	74,426	5,700	80,126
Flagged for Increased Cryptic Site Use	4,441	1,018	5,459
% without associated rsID	70.0%	42.4%	69.9%
Shared Between Multiple Tissues (Novel variants only)	9,810	246	10,125
Shared Between Multiple Tissues (Novel or <1% Average Heterozygosity)	27,581	942	28,813
Shared Molecular Phenotypes for Variants Identified in Multiple Tissues[a]	20,900 (>50%) 15,793 (>67%)	779 (>50%) 618 (>67%)	21,887 (>50%) 16,574 (>67%)

[a] **Variants (novel or <1% average heterozygosity) identified in multiple tumor types, flagged by Veridical for the same splicing category (i.e. junction spanning exon skipping) in more than half (>50%) or over two-thirds (>67%) of patients.**

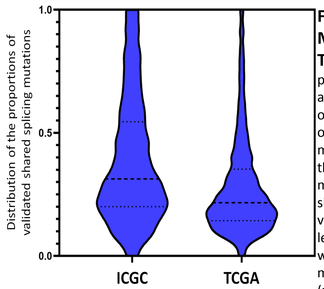


Figure 2. Census of Recurrent Splicing Mutations Present in Multiple ICGC and TCGA Patients. Predicted splicing mutations present in multiple tumors that cause splicing abnormalities were analyzed to determine the rate of validation. Violin plots indicate the distributions of the fraction of predicted and validated splicing mutations present in multiple patients relative to the total number of tumours carrying those mutations in TCGA and ICGA. To achieve statistical significance (95% C.I.), distributions of 1,379 variants shared by both datasets and present in at least 9 ICGC (left) and 24 TCGA (right) patients were compared. A higher overall proportion of mutations are validated in the ICGC dataset (average of 38.6% for ICGC and 27.8% for TCGA).

Table 2: Mutations Leading to Multiple Types of Aberrant Splicing

Gene	Splice Mutation	R_i (bits)	Tumor	Observed Splicing Event
SAP30BP	17:73702087G>A (c.661-1G>A) 17:73702091G>A (c.664G>A)	14.4>3.5 | -3.0>2.2 (Natural | Cryptic)	BLCA	Two closely situated mutations affect the natural acceptor of *SAP30BP* exon 10 (NM_013260). The first mutation abolishes the natural acceptor (increasing exon skipping and intron inclusion RNAseq reads), while the second creates a cryptic site 6nt downstream.
JUP	17:39919236T>C (c.1496A>G)	3.1>0.6 | 0.5>3.7 (Natural | Cryptic)	BLCA	Mutation within *JUP* exon 8 (NM_002230) simultaneously weakens a natural donor while creating a cryptic site. Exon skipping, intron inclusion and cryptic site use are all observed.
GAS5	1:173835706C>A (n.118-1G>T)	8.1>0.3 | 1.9>4.3 (Natural | Cryptic)	KIRC	The natural acceptor of *GAS5* exon 4 (NR_002578) is abolished, simultaneously strengthens a cryptic site 9nt downstr. Cryptic site is seen, along with exon skipping and intron inclusion (likely due to the abolished natural site).
CAPN8	1:223815845C>A (c.427-1G>T)	4.0>-3.7 | 1.4>3.8 (Natural | Cryptic)	LUAD	Simultaneously abolishes the natural acceptor of *CAPN8* exon 4 (NM_001143962) while strengthening a cryptic site 9nt downstr. Cryptic site use, exon skipping and intron inclusion is significantly increased in this patient.
SSH3	11:67074511A>T (c.465-2A>T)	6.4>-1.3 | 3.8>6.6 (Natural | Cryptic)	LUAD	*SSH3* exon 5 natural acceptor (NM_017857) is abolished, while a pre-existing cryptic site 9nt downstr. is strengthened. Exon skipping, total intron inclusion and cryptic use are seen.
TKTL1	X:153537696G>A (c.253-1G>A)	7.0>-3.9|-2.5>2.7 (Natural | Cryptic)	SKCM	The natural acceptor of *TKTL1* exon 3 (NM_012253) is abolished and a cryptic site 2nt downstream is created, leading to a deletion of 2nt. Cryptic site use, total exon skipping and intron inclusion are detected.
ELN	7:73474706G>C (c.1622G>C)	6.6>4.7 | -6.0>2.8 (Natural | Cryptic)	SKCM	The variant weakens *ELN* exon 25 (NM_000501) natural site while creating a cryptic site 3nt away. The cryptic site is observed, as is intron inclusion, however exon skipping is the predominant aberrant splicing isoform detected.
SPOP	17:47699430C>A (c.79-1G>T)	9.9>2.2 | -0.4>2.0 (Natural | Cryptic)	BLCA	The natural exon 4 acceptor of *SPOP* is weakened while an equivalently strengthened cryptic site 6nt downstr. is created. Cryptic site use, exon skipping and intron inclusion are seen.
SEC61G	7:54825289T>G (c.-6-2A>C)	7.7>-10.9|1.4>3.5 (Natural | Cryptic)	COAD	The natural acceptor of exon 2 (NM_014302) is abolished and a cryptic acceptor is created. Significant exon skipping, cryptic site use, and intron inclusion is detected by Veridical.
HLA-A	6:29912835G>C (c.1013-1G>C)	8.6>-3.1 | 1.9>4.3 (Natural | Cryptic)	CESC	The natural acceptor of exon 6 of *HLA-A* (M_002116) is abolished by this variant, and a pre-existing cryptic site 4nt downstream is strengthened. Exon skipping, total intron inclusion, and cryptic site use are detected.

Example mutations which significantly alter splicing in all ways measured by Veridical (i.e. cryptic splice site use, exon skipping, intron inclusion). Mutations are linked to their page on https://validsplicemut.cytognomix.com/.

Table 3: Validated Splicing Mutations in COSMIC Cancer Gene Census genes

Gene	Splice Mutation	R_i (bits)	Tumor	Observed Splicing Event
CASC5	15:40942786G>A (c.6212+5G>A)	4.8 > 1.7 (Natural Site)	AML	The natural donor site of *CASC5* exon 19 (NM_144508) is weakened, leading to a significant increase in intron inclusion.
DNMT3A	2:25467022A>G (c.1851+2T>C)	3.6 > -3.5 (Natural Site)	AML	The natural donor site of *DNMT3A* exon 15 (NM_022552) is abolished, resulting in a increase in total exon skipping and intron inclusion.
STAG2	X:123176495G>A (c.462G>A)	6.5 > 3.5 (Natural Site)	BLCA	The natural donor of *STAG2* exon 6 (NM_006603) is weakened, and a significant amount of exon 6 skipping is observed.
STAG2	X:123200024G>A (c.2097-1G>A)	19.5 > 8.6 (Natural Site)	BLCA	The natural acceptor of *STAG2* exon 21 (NM_006603) is weakened, resulting in a significant increase in exon 21 skipping.
ATM	11:108214098G>T (c.8418G>T)	8.7 > 5.1 (Natural Site)	BRCA	A natural donor site is weakened, leading to a significant increase in *ATM* exon 57 (NM_000051) skipping events. Some reads with mutation depict wildtype, leaky splicing.
BARD1	2:215645882A>T (c.716T>A)	0.9 > 3.1 (Cryptic Site)	BRCA	The mutation strengthens a cryptic site within *BARD1* exon 4 (NM_000465). Reads which use this activated cryptic site contain the mutation (one exception). Some reads with mutation depict wildtype, leaky splicing.
GATA3	10:8115701G>C (c.1048-1G>C)	0.9 > -10.7 (Natural Site)	BRCA	Mutation abolishes the natural acceptor of *GATA3* exon 6 (NM_002051). This increases the use of a pre-existing exonic cryptic splice site (4.2 > 5.6 bits; 8nt deletion) and significantly increases total intron inclusion.
TP53	17:7577609C>T (c.673-1G>A)	6.0 > -4.9 (Natural Site)	BRCA	A natural acceptor site is abolished, activating a cryptic site 49nt upstream (R_i =5.2 bits) of *TP53* exon 7 (NM_000546).
POLD1	19:50920353A>G (c.3119A>G)	8.6 > 6.1 (Natural Site)	COAD	The natural donor of *POLD1* exon 25 (NM_002691) is weakened, leading to a significant increase in overall exon skipping.
SMAD3	15:67482748C>G (c.1155-3C>G)	11.9>3.1|-4.0>7.7 (Natural | Cryptic)	COAD	Mutation weakens the natural acceptor of *SMAD3* exon 9 (NM_005902) and predicts a cryptic site that does not appear to be used. A significant number of intron inclusion reads are observed. A distant pre-existing cryptic acceptor (9.6 bits; 3598nt from natural acceptor) was observed.
PIK3R1	5:67591246A>G (c.936-2A>G)	7.5 > -7.3 (Natural Site)	GBM	The natural acceptor of *PIK3R1* exon 8 (NM_181504) is abolished, which promotes a significant increase in exon 8 skipping.
FAT1	4:187521515C>A (c.11641-1G>T)	5.3 > -2.4 (Natural Site)	HNSC	Natural acceptor of *FAT1* exon 22 (NM_005245) is abolished, resulting in intron inclusion, and use of 3 cryptic sites; 82nt upstream of acceptor, and 237nt and 234nt downstream from the natural acceptor.
TGFBR2	3:30729875G>A (c.1397-1G>A)	8.4 > -2.5 (Natural Site)	HNSC	*TGFBR2* exon 6 natural acceptor (NM_003242) is abolished, leading to 5 splicing events: intron inclusion, use of three cryptic sites (35nt exonic, 30nt and 972nt intronic), and exon 6 and 7 skipping.
PBRM1	3:52682355C>G (c.813+5G>C)	6.8 > 2.9 (Natural Site)	KIRC	The natural donor of *PBRM1* exon 8 (NM_018313) is weakened, which leads to a significant increase in exon 8 skipping.
PBRM1	3:52685756A>G (c.714+2T>C)	7.7 > 0.7 (Natural Site)	KIRC	The natural donor of *PBRM1* exon 7 (NM_018313) is abolished, resulting in a significant increase in exon skipping.
SETD2	3:47079269T>A (c.7239-2A>T)	9.8 > 2.1| 6.4 >9.0 (Natural | Cryptic)	KIRC	This mutation both significantly weakens the natural acceptor of *SETD2* exon 18 (NM_014159) and strengthens a 4nt exonic cryptic site.
RB1	13:49027249T>A (c.1814+2T>A)	4.9 >-13.7 (Natural Site)	LUAD	The natural donor of *RB1* exon 18 (NM_000321) is abolished, leading to a significant increase in both exon skipping and intron inclusion. All intron inclusion reads contain the mutation of interest.
RBM10	X:47006900G>T (c.17+3G>T)	7.8 > 4.1 (Natural Site)	LUAD	The natural donor of *RBM10* exon 2 (NM_005676) is weakened, leading to a significant increase in exon 2 skipping.
RBM10	X:47028898G>T (c.201+1G>T)	8.7 > -9.9 (Natural Site)	LUAD	*RBM10* exon 3 (NM_005676) natural donor is abolished. Reads which overlap the exon-intron junction are observed (with mutation). Use of a cryptic donor (61nt upstream; R_i =1.7 bits) is observed as well.
DDX5	17:62500098 TACAG>T (c.441+2delACAG)	-1.3 > 5.4 (Cryptic Site)	PRAD	The mutation creates a 5.4 bit cryptic donor within *DDX5* exon 4 (NM_004396), which would lead to a 4nt deletion of exon 4. Note that wildtype splicing is still the dominant isoform observed.
PTEN	10:89690802G>A (c.210-1G>A)	8.5 > -2.3 (Natural Site)	PRAD	The natural acceptor of *PTEN* exon 5 (NM_000314) is abolished, leading to an increased amount of exon 5 skipping.
NRAS	1:115258669A>G (c.111+2T>C)	8.1 > 1.1 (Natural Site)	SKCM	The mutation abolishes the natural donor of *NRAS* exon 2 (NM_002524), which promotes a significant increase in exon 2 skipping
PPP6C	9:127933364C>T (c.171G>A)	6.7 > 3.7 (Natural Site)	SKCM	The mutation weakens *PPP6C* exon 2 (NM_002721) natural donor, leading to increased intron inclusion. All reads crossing the splice junction have the mutation. An intronic cryptic site is also activated (110nt downstream).
PPP6C	9:127923119C>G (c.237+1G>C)	6.8 > -11.8 (Natural Site)	SKCM	This mutation abolishes the natural donor of *PPP6C* exon 3 (NM_002721), resulting in a significant increase in exon 3 skipping.
BAP1	3:52442512T>C (c.233A>G)	1.9 > 5.1 (Cryptic Site)	UVM	A pre-existing cryptic donor within *BAP1* exon 4 (NM_004656) is strengthened, leading to a significant increase in its use. This mutation leads to a 27nt deletion in the mutated exon 4 mRNA.

Example mutations which alter splicing in tumor-associated genes found in patients with these tumor types. Mutations are hyperlinked to their ValidSpliceMut Beacon page, which provides additional material such as IGV images of the RNAseq evidence for the regions of interest. GRCh37 coordinates are indicated.

Conclusions

We present the Validated Splicing Mutation Beacon and web resource (http://validsplicemut.cytognomix.com/) that consists of 341,486 unique mutations from TCGA and ICGC tumor patients predicted to alter splicing with RNA-Seq support. The majority of these mutations (69.9%) are not present in dbSNP 150. In total, 543 Tier 1 COSMIC CGC genes exhibited at least one Veridical-flagged variant present in the ValidSpliceMut database. This resource should significantly contribute to reducing the number of outstanding VUS in tumor (and possibly some germline) genomes, and substantially increases the number of functional variants with previously unappreciated consequences to mRNA splicing, in particular, those which activate cryptic splice sites.

References

1. Foley SB, Rios JJ, Mgbemena VE, *et al.* Use of Whole Genome Sequencing for Diagnosis and Discovery in the Cancer Genetics Clinic. *EBioMedicine*. 2015; 2(1):74–81. 26023681. **2.** Caminsky N, Mucaki EJ, Rogan PK. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis [version 1; referees: 2 approved]. *F1000Res*. 2014; 3:282. 25717368. **3.** Viner C, Dorman SN, Shirley BC, *et al.* Validation of predicted mRNA splicing mutations using high-throughput transcriptome data [version 2; referees: 4 approved]. *F1000Res*. 2014; 3:8. **4.** Mucaki EJ, Caminsky NG, Perri AM, *et al.* A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Med Genomics*. 2016; 9:19. **5.** Hoadley KA, Yau C, Hinoue T, *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell.* 2018; 173(2):291–304.e6. **6.** Shirley BC, Mucaki EJ, Rogan PK. Pan-cancer repository of validated natural and cryptic mRNA splicing mutations [version 2; peer review: 1 approved, 1 approved with reservations]. *F1000Research*. 2019; 7:1908. **7.** Global Alliance for Genomics and Health. GENOMICS: A federated ecosystem for sharing genomic, clinical data. *Science*. 2016; 352(6291):1278–1280. **8.** Tate JG, Bamford S, Jubb HC, *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*. 2019; 8;47(D1):D941-D947. **9.** Bamshad MJ, Shendure JA, Valle D *et al.* The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. *Am J Med Genet A*. 2012; 158A(7):1523-5.

Acknowledgements

SHARCNET · compute • calcul CANADA · NSERC CRSNG · CYTOGNOMIX