DATA NOTE

Pan-cancer repository of validated natural and cryptic mRNA splicing mutations [version 1; peer review: 1 approved, 1 approved with reservations]

Ben C. Shirley[1*], Eliseos J. Mucaki[2*], Peter K. Rogan[1-4]

[1]CytoGnomix Inc., London, Ontario, N5X 3X5, Canada
[2]Biochemistry, University of Western Ontario, London, Ontario, N6A 2C1, Canada
[3]Computer Science, University of Western Ontario, London, Ontario, N6A 2C1, Canada
[4]Oncology, University of Western Ontario, London, Ontario, N6A 2C1, Canada

[*] Equal contributors

v1 **First published:** 07 Dec 2018, **7**:1908 (https://doi.org/10.12688/f1000research.17204.1)

Latest published: 20 Mar 2019, **7**:1908 (https://doi.org/10.12688/f1000research.17204.2)

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) 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 351,423 of these validated mutations, the majority of which (69.1%) are not present in the Single Nucleotide Polymorphism Database (dbSNP 150). There are 117,951 unique mutations which weaken or abolish natural splice sites, and 244,415 mutations which strengthen cryptic splice sites (10,943 affect both simultaneously). 27,803 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 (https://validsplicemut.cytognomix.com/).

Keywords

RNA Splice Sites, Single Nucleotide Polymorphism, Genome, Mutation, Chromosomes, Neoplasms, Information Theory, Next generation sequencing, validation

Open Peer Review

Reviewer Status ✓ ?

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 20 Mar 2019	✓ report	
version 1 published 07 Dec 2018	✓ report	? report

1 **Emanuele Buratti**, International Centre for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy

2 **Francesca D. Ciccarelli**, Francis Crick Institute, London, UK
King's College London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Peter K. Rogan (progan@uwo.ca)

Author roles: Shirley BC: Data Curation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Mucaki EJ**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Rogan PK**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: PKR cofounded and BCS is an employee of CytoGnomix Inc., which hosts the interactive webpage described in this study. CytoGnomix markets subscriptions to and services based on the software that generated the ValidSpliceMut database. EJM has no conflict of interest.

Grant information: PKR is supported by The Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2015-06290], Canadian Foundation for Innovation, Canada Research Chairs, and CytoGnomix. Compute Canada and Shared Hierarchical Academic Research Computing Network (SHARCNET) provided high performance computing and storage facilities.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Shirley BC, Mucaki EJ and Rogan PK. **Pan-cancer repository of validated natural and cryptic mRNA splicing mutations [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2018, **7**:1908 (https://doi.org/10.12688/f1000research.17204.1)

First published: 07 Dec 2018, **7**:1908 (https://doi.org/10.12688/f1000research.17204.1)

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, such sources are critically needed to assist in distinguishing driver mutations from overwhelming numbers of bystander mutations. VUS classification criteria highlight the limitations in genome interpretation due to ambiguous variant interpretation. Of the 458,899 variant submissions in NCBI's ClinVar database with clinical interpretations, nearly half (n=221,271) are VUS (as of November 5th 2018). Only 10,784 variants in ClinVar have been documented to affect mRNA splicing at splice donor or acceptor sites, with 1,063 of these being classified as VUS, and cryptic mRNA splicing mutations are not explicitly described. The current ACMG criteria[2] for variant pathogenicity prevent clinical classification of most VUS. Functional evidence that VUS either disrupt or abolish expression of genes has been sought to improve classification and provide insight into the roles, if any, of individual VUS in predisposing or causing disease. We present a comprehensive data repository for a relatively common mutation type (cis-acting variants that alter mRNA splicing). Mutations are predicted with information theory-based analyses[3], and supported with functional evidence that variants in tumor genomes are specifically associated with abnormally spliced mRNAs that are infrequent or absent in transciptomes lacking these variants[4].

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[3,5–15]. 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 novel variants in patients with history or predisposition to heritable breast and/or ovarian cancer[11–15].

The Cancer Genome Atlas (TCGA) Pan-Cancer Atlas (PCA) is a comprehensive integrated genomic and transcriptomic resource containing data from >10,000 tumors across 33 different tumor types[16]. Here, we utilized IT-based tools for assessment of high quality sequenced variants in TCGA patients for their potential impact on mRNA splicing. The accuracy of predicted mutations was evaluated with an algorithm we previously developed that compares transcripts from individuals carrying these variants with others lacking them. The results of these genome-wide analyses are presented using an online resource which can be queried through the Beacon Network[17].

Methods

TCGA data acquisition and processing

Controlled-access data was obtained with permission from the Data Access Committee at NIH for TCGA and from the International Cancer Genomics Consortium. Patient RNA sequencing BAM files (tumor and normal, when available) and their associated VCF files (GRCh37) were initially obtained from the CancerGenomeHub (CGhub). Files were later downloaded through Genomic Data Commons using the GDC Data Transfer Tool (version 1.3.0), as CGhub was decommissioned mid-project. Variants in VCF files which did not pass quality control (QC) were not analyzed.

Information analysis and RNA-Seq validation of splicing variants

We used the *Shannon Pipeline* software (which applies IT to rapidly perform high-throughput, *in silico* prediction of the impacts of variants on mRNA splicing)[18] to analyze all QC-passing variants in VCFs from TCGA (>168 million variants) to evaluate their potential impact on splice site binding strength (changes in information content, R_i, measured in bits). Variants which were predicted to strengthen known natural sites or weaken cryptic splice sites were excluded from all subsequent analyses.

To validate the potential impact of Shannon Pipeline-flagged mutations, *Veridical* software analyzed genomic variants (including insertions and deletions) by comparing the RNA-Seq alignment in the region surrounding the variant with the corresponding interval in control transcriptomes (normal and tumor tissue of the same type) lacking the variant[4,19]. Veridical: a) counts abnormally spliced reads in RNA-Seq data (categorized as: cryptic site use, exon skipping, or intron inclusion [containing or adjacent to the flagged mutation]), b) applies the Yeo-Johnson transformation to these results, and c) determines the null hypothesis probability (p-value) that the transformed read count corresponds to normal splicing. In tumor types where normal controls were not available, a set of RNA-Seq datasets from 100 different normal tissues from TCGA were used (e.g. a combination of 5 tissue types: BRCA, BLCA, LUAD, KIRC, PRAD). Veridical results that were not significant for a particular variant (p-value > 0.05 for all of the splicing categories) were not further analyzed. After analysis, Veridical validated 351,423 unique mutations for their direct impact on mRNA splicing (Table 1). The Shannon pipeline-flagged and Veridical-filtered results were combined into a single large table (*Dataset 1*[20]), the source data for the ValidSpliceMut SQL database and the associated Beacon application.

Development of the ValidSpliceMut database and Beacon

We created a publicly accessible Application Programming Interface (API) (https://beacon.cytognomix.com) that can be utilized to programmatically query variants passing filter thresholds described above (*Dataset 1*[20]). It was built in accordance with the GA4GH Beacon v1.0.0 specification, which describes a Representational State Transfer (REST) API for genetic data sharing. A Beacon accepts queries using an HTTP request and returns JavaScript Object Notation (JSON). Our Beacon implementation is coded in PHP 7.0 and utilizes a MySQL database (version: 5.7.24) with indexes applied to variant ID, chromosome, and coordinate fields (GRCh37). The returned JSON object reports whether the variant was found within our Beacon dataset as well as metadata including splice site coordinate, splice type, site type, the IT-based measures $R_{i,initial}$

Table 1. Unique Flagged Variants by TCGA Tumor Tissue Type*.

TCGA-ACC	**TCGA-BLCA**	**TCGA-BRCA**	**TCGA-CESC**	**TCGA-CHOL**	**TCGA-COAD**	**TCGA-DLBC**
1776	10,100	27,507	26,710	10,410	9600	6497
TCGA-ESCA	**TCGA-GBM**	**TCGA-HNSC**	**TCGA-KICH**	**TCGA-KIRC**	**TCGA-KIRP**	**TCGA-LAML**
12,856	1156	2834	27,340	6733	4747	20,770
TCGA-LGG	**TCGA-LIHC**	**TCGA-LUAD**	**TCGA-LUSC**	**TCGA-MESO**	**TCGA-OV**	**TCGA-PAAD**
1432	14,981	18,618	2667	284	95,193	1593
TCGA-PCPG	**TCGA-PRAD**	**TCGA-READ**	**TCGA-SARC**	**TCGA-SKCM**	**TCGA-STAD**	**TCGA-TGCT**
90	997	5104	21,107	12,707	19,761	464
TCGA-THCA	**TCGA-THYM**	**TCGA-UCEC**	**TCGA-UCS**	**TCGA-UVM**		
57,610	17,063	29,076	11,044	2501		

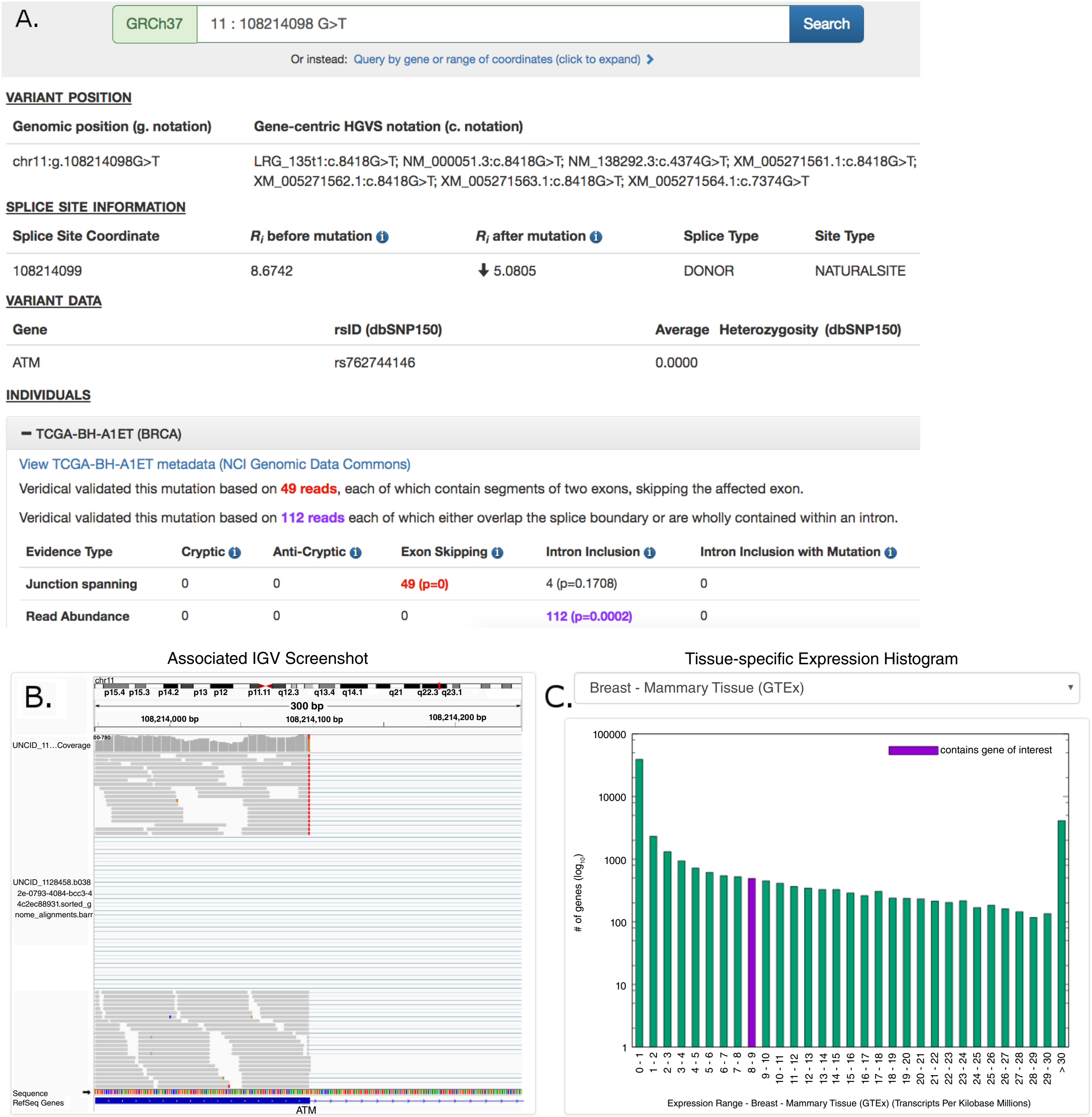
*The number of Veridical-flagged mutations in each The Cancer Genome Atlas (TCGA) cancer data set. Variants shared between multiple tissue types are counted for each category. Variant and RNA-Seq data were provided by The Cancer Genome Atlas Pan-Cancer Analysis Project[16].

and $R_{i,final}$, affected individual IDs, tumor type, Veridical evidence by type annotated with significance level, and, if known, the corresponding rsID with its average heterozygosity (dbSNP 150). The metadata for each variant sent to the Beacon Network is a concise subset of available results in our database. It includes the first relevant database entry, meaning that if the variant exists within multiple individuals only the first will contribute fields to the metadata. However, among this metadata is a hyperlink to our local website containing results for any remaining tumors.

We developed the website ValidSpliceMut (Figure 1) to serve as a local interface to our Beacon, allowing users to manually search for a variant, by gene name or genome coordinate range. ValidSpliceMut automatically queries our Beacon, and formats the results of the search, if any. This website provides a complete view of variants, including Veridical-based evidence on all data related to every affected individual. If a variant is associated with multiple splice sites, the user is presented with a brief overview of all affected sites and must select a desired site to continue. To obtain the coordinate of the queried variant in gene-centric notation, a link is provided which queries the Mutalyzer API and generates coordinates for all available transcripts. ValidSpliceMut only reports transcripts for the gene affected by the variant.

A results page presents variant-specific data in tabular format and an expandable list of panels describing the affected individuals. Each of these panels contains Veridical output in tabular format for the selected tumor, a link to the tumor metadata at US National Cancer Institute (by querying the GDC API to obtain a UUID which is used to construct a link to the GDC data portal), an Integrative Genome Viewer (IGV) screenshot containing the variant (IGV screenshots are available for selected variants, see below), and a histogram which presents the expression levels of the variant-containing gene compared to all other gene expression levels across a selected normal tissue type (created dynamically using gnuplot 5.0). The tissue expression data is provided by GTEx (downloaded on 10/22/18). However, several TCGA tumor types did not have a GTEx equivalent (CHOL, DLBC, MESO, READ, SARC, THYM and UVM). The GNF Expression Atlas 2[21] was downloaded from the UCSC Genome Browser and was used for expression data for both lymph nodes (DLBC) and the thymus (THYM). For the remaining tissues, expression data from the following studies were obtained from the Genome Expression Omnibus (GEO): GSE76297 (CHOL), GSE2549 (MESO), GSE15781 (READ) GSE44426 (SARC), and GSE44295 (UVM).

To generate IGV images presented on the webpage, a bash script was written to automatically load the RNA-Seq BAM file of a patient with a mutation of interest into IGV, set the viewing window within the region of interest (300nt window, centered on the variant), sorted to bring reads containing the variant of interest to the top of the screen (to increase chance of visualizing mutant splice form), followed by a screen capture. The generation and storage of IGV images for all patient-mutation pairs would be prohibitive due to limitations in time and server space requirements. Therefore IGV images showing evidence of splicing abnormalities were generated *only* for patient-mutation pairs which met the most stringent criteria: the mutation was required to be flagged for junction-spanning cryptic site use, exon skipping, or intron inclusion (with mutation); the flagged category must include 5 or more reads in this category; if the variant is present in the dbSNP database (release 150), the frequency was required to be < 1% of the population; and the Veridical results, in which the mutations flagged were required to exhibit $p \leq 0.01$ for at least one form of evidence of a splicing abnormality. In some cases, the splicing event observed by Veridical may not be present within the image window as the automated procedure used to create these images does not present all evidential sequence reads due to limitations on the number of reads that are shown. Additionally, reads appearing as exon skipping may instead indicate a pre-existing cryptic site outside of the viewing window (see Table 2; *FAT1*: g.187521515C>A [c.11641-1G>T] and *SMAD3*:g.67482748C>G [c.1155-3C>G]).



11 : 108214098 G>T

Or instead: Query by gene or range of coordinates (click to expand) ›

VARIANT POSITION

Genomic position (g. notation)	Gene-centric HGVS notation (c. notation)
chr11:g.108214098G>T	LRG_135t1:c.8418G>T; NM_000051.3:c.8418G>T; NM_138292.3:c.4374G>T; XM_005271561.1:c.8418G>T; XM_005271562.1:c.8418G>T; XM_005271563.1:c.8418G>T; XM_005271564.1:c.7374G>T

SPLICE SITE INFORMATION

Splice Site Coordinate	R_i before mutation ℹ	R_i after mutation ℹ	Splice Type	Site Type
108214099	8.6742	↓ 5.0805	DONOR	NATURALSITE

VARIANT DATA

Gene	rsID (dbSNP150)	Average Heterozygosity (dbSNP150)
ATM	rs762744146	0.0000

INDIVIDUALS

━ TCGA-BH-A1ET (BRCA)

View TCGA-BH-A1ET metadata (NCI Genomic Data Commons)

Veridical validated this mutation based on **49 reads**, each of which contain segments of two exons, skipping the affected exon.

Veridical validated this mutation based on **112 reads** each of which either overlap the splice boundary or are wholly contained within an intron.

Evidence Type	Cryptic ℹ	Anti-Cryptic ℹ	Exon Skipping ℹ	Intron Inclusion ℹ	Intron Inclusion with Mutation ℹ
Junction spanning	0	0	**49 (p=0)**	4 (p=0.1708)	0
Read Abundance	0	0	0	**112 (p=0.0002)**	0

Associated IGV Screenshot

chr11
p15.4 p15.3 p14.2 p13 p12 p11.11 q12.3 q13.4 q14.1 q21 q22.3 q23.1
300 bp
108,214,000 bp
108,214,100 bp
108,214,200 bp
UNCID_11...Coverage
00-780
UNCID_1128458.b038 2e-0793-4084-bcc3-4 4c2ec88931.sorted_g nome_alignments.barr
Sequence →
RefSeq Genes
ATM

Tissue-specific Expression Histogram

Figure 1. Screenshot of *ATM*:g.108214098G>T Results Provided By ValidSpliceMut Website. (**A**) The 'Variant Position' heading displays the variant of interest in g. notation, and provides a link which queries the Mutalyzer API to obtain the variant coordinate in a gene-centric c. mutation format. Variant-specific and splice site-specific tabular results are presented under the headings "Splice Site Information" and "Variant Data". Results are organized by TCGA sample IDs harboring the mutation within a series of expandable panels. A link is provided to patient tumor metadata on the GDC data portal. Each panel consists of read counts and p-values by Veridical evidence type. Significant p-values (≤ 0.05) are highlighted in bold. Evidence types deemed "strongly corroborating" (Viner *et al.* 2014) are color coded and correspond to the dynamically generated text appearing above the table. (**B**) An integrative genome viewer (IGV) image showing alignment of expressed sequence reads. IGV screenshots are provided only for mutations present <1% of population (in dbSNP 150), with ≥ 5 junction-spanning reads, and are highly significant ($p < 0.01$) for cryptic splicing, exon skipping, and/or intron inclusion with mutation. A specific IGV screenshot for this sample captures the region surrounding the mutation. Here, several RNA-Seq reads show skipping of the affected exon. (**C**) A dynamically generated histogram presents expression levels of all genes for a selected normal tissue type. Genes are grouped into bins based on expression level, denoted on the x-axis. The number of genes present in each bin is shown on the y-axis ($\log_{10}$ scale). The histogram key indicates the expression range of the variant-containing gene. Tissue type can be changed via a drop-down list.

Table 2. Validated Splicing Mutations in COSMIC Cancer Gene Census genes in TCGA tumor genomes.

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_144508.4) 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_022552.4) is abolished, resulting in a significant 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_006603.4) 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_006603.4) 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_000051.3) skipping events. Some reads with mutation are involved in wildtype splicing (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_000465.2). Reads which use activated cryptic site contain the mutation (one exception). Some reads with mutation are involved in wildtype splicing (leaky splicing).
GATA3	10:8115701G>C (c.1048-1G>C)	0.9 > -10.7 (Natural Site)	BRCA	The mutation abolishes the natural acceptor of *GATA3* exon 6 (NM_002051.2). This both increases the use of a pre-existing exonic cryptic splice site (4.2 > 5.6 bits; leads to an 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_000546.5).
POLD1	19:50920353A>G (c.3119A>G)	8.6 > 6.1 (Natural Site)	COAD	The natural donor of *POLD1* exon 25 (NM_002691.3) is weakened, leading to a significant increase in total exon skipping.
SMAD3	15:67482748C>G (c.1155-3C>G)	11.9>3.1|-4.0 > 7.7 (Natural | Cryptic)	COAD	This mutation both weakens the natural acceptor of *SMAD3* exon 9 (NM_005902.3) and creates a cryptic site (does not appear to be used). A significant amount intron inclusion reads are observed. Use of a distant pre-existing cryptic acceptor (9.6 bits; 3598nt from natural acceptor) was.
PIK3R1	5:67591246A>G (c.936-2A>G)	7.5 > -7.3 (Natural Site)	GBM	The natural acceptor of *PIK3R1* exon 8 (NM_181504.3) 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	The natural acceptor of *FAT1* exon 22 (NM_005245.3) is abolished, resulting in both intron inclusion (total intron inclusion and the use of a 2.3 bit cryptic site 82nt upstream of natural acceptor) and use of two exonic cryptic sites (237nt and 234nt from natural acceptor; Ri=1.0 bits and -0.2 bits, respectively).
TGFBR2	3:30729875G>A (c.1397-1G>A)	8.4 > -2.5 (Natural Site)	HNSC	*TGFBR2* exon 6 natural acceptor (NM_003242.5) is abolished, leading to multiple splicing events: intron inclusion, use of three cryptic sites (35nt exonic [R_i=3.7 bits], 30nt and 972nt intronic [R_i=0.4 bits and 11.2 bits, respectively]), and exon 6 and 7 skipping (uses a novel exon ~55kb downstream of exon 7).
PBRM1	3:52682355C>G (c.813+5G>C)	6.8 > 2.9 (Natural Site)	KIRC	The natural donor of *PBRM1* exon 8 (NM_018313.4) 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_018313.4) is abolished, resulting in a significant increase in total 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_014159.6) while strengthening a 4nt exonic cryptic site, which is used.
RB1	13:49027249T>A (c.1814+2T>A)	4.9 >-13.7 (Natural Site)	LUAD	The natural donor of *RB1* exon 18 (NM_000321.2) 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_005676.4) is weakened, leading to a significant increase in exon 2 skipping.

Gene	Splice Mutation	R_i (bits)	Tumor	Observed Splicing Event
RBM10	X:47028898G>T (c.201+1G>T)	8.7 > -9.9 (Natural Site)	LUAD	*RBM10* exon 3 (NM_005676.4) natural donor is abolished. RNAseq reads which overlap the exon-intron junction are observed (all reads contain mutation). Use of cryptic donor (61nt upstream of donor; 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_004396.3), 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_000314.4) is abolished, leading to an increased amount of total 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_002524.4), 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_002721.4) natural donor, leading to increased intron inclusion. All reads which cross the junction contain the mutation. A intronic cryptic site is also activated (110nt downstr.; R_i=6.4 bits).
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_002721.4), resulting in a significant increase in exon 3 skipping.
BAP1	3:52442512T>C (c.233A>G)	1.9 > 5.1 (Cryptic Site)	UVM	A cryptic donor within *BAP1* exon 4 (NM_004656.3) is strengthened, leading to a significant increase in its use. Its use leads to a 27 nt deletion of exon 4.

Example mutations which alter splicing in tumor-associated genes found in patients with the same tumor type. Mutations are linked to their page on https://validsplicemut.cytognomix.com/, which provides additional material such as RNAseq images of the regions of interest. GRCh37 coordinates provided.

Dataset validation and discussion

We have derived a GA4GH-standardized, searchable web resource for a large set of validated mRNA splicing variants present in diverse tumor types. All variants passing QC in TCGA cancer patients were analyzed with the Shannon pipeline[18]. This revealed that 1,297,242 variants were predicted to have significant impacts on normal mRNA splicing (347,549 natural and 985,112 cryptic splice sites; 35,419 affecting both types). Subsequent RNA-Seq analysis with Veridical[4] provided evidence of abnormal gene expression specifically associated with a subset of these variant(s), identifying 351,423 unique mutations. Results are searchable through either the Beacon Network, or our publicly-accessible webpage.

Our results contrast with another TCGA study that investigated alternative mRNA splicing[22] and demonstrated a limited set of non-constitutive exon-exon junctions attributable to cis-acting splicing mutations (n = 32). The 2,736 novel or rare variants that we report which specifically activate cryptic splicing (significant 'junction-spanning cryptic site use' reads found by Veridical), exceed the number reported in another study that analyzed all available TCGA tumor transcriptomes (n=1,964)[23].

Validated variants (which we define as mutations) were also tallied by tumor tissue type in our study (Table 1). 33.6% of unique mutations (n=117,951) significantly weaken natural splice sites, while 69.6% (n=244,415) strengthen novel or pre-existing cryptic sites. 242,983 mutations (69%) are absent from dbSNP 150. 73,975 mutations (21%) are present in <1% of the population, of which 27,803 of these (and those not present in dbSNP) were present in multiple tumor types. Valid mutations lacking rsIDs represent either novel or recently observed variants. This low level of dbSNP saturation is consistent with the idea that many currently unknown mRNA splicing mutations may yet be discovered through additional sequencing studies.

In Table 2, we highlight a subset of validated splicing mutations (n=25) which were identified in known driver genes implicated in the COSMIC (Catalogue Of Somatic Mutations In Cancer) Cancer Gene Census catalog (CGC)[24]. These mutations are associated with either increased exon skipping, intron inclusion, and/or cryptic site use. Mutations in Table 2 are hyperlinked to the ValidSpliceMut webpage which provides additional information, including expression evidence supporting predictions made by the Shannon pipeline.

Many mutations generated multiple types of abnormal read evidence present in mis-spliced transcripts. Interestingly, a subset of mutations (n=28) produced evidence for every type of abnormal splicing reported by Veridical. *Dataset 2*[25] (see Data Availability) describes 11 representative mutations that simultaneously increase exon skipping, intron inclusion, and activate (or significantly increase utilization of) a strengthened cryptic site. In all but one instance, the mutation weakens the natural site while simultaneously strengthening a nearby cryptic site. The one exception involves the gene *SAP30BP*, where simultaneously occurring mutations in the same read (in linkage disequilibrium; separated by 4 nucleotides) independently cause two separate splicing changes: g.73702087G>A (c.661-1G>A; abolishes the natural acceptor of exon 10) and g.73702091G>A (c.664G>A; creates a weak cryptic acceptor site). The combined splicing impact of these variants is significant exon skipping, intron inclusion, and use of the activated cryptic site.

Because of the requirement for expression validation, this resource presents a set of splicing abnormalities in which we have the highest confidence. We anticipate that some correct predictions of the Shannon pipeline may have not been validated by Veridical due to the limitations of mRNA detection; for example, either low expression of the gene harboring the mutation or nonsense-mediated decay of the corresponding transcript could be consistent with the effects of a valid splicing mutation, but in the absence of a sufficient number of abnormal reads, the mutation could not be confirmed. Furthermore, at the time that the current analysis was performed, the available Shannon pipeline version did not report regulatory splicing variants adjacent to constitutive and cryptic splice sites which influence exon definition. Due to the substantial processing required for the complete TCGA dataset, the present analysis does not incorporate the effects of these variants on exon definition, which we have modeled by IT[6]; it does not predict the relative abundance of leaky, natural and cryptic isoforms, though such information might be inferred from the expression data on each tumor. The current version of Shannon pipeline does integrate predictions of splicing regulatory sequences and accounts for relative abundance of mRNA isoforms by exon definition, and is available through the MutationForecaster system.

The Validated Splicing Mutation resource should substantially 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, activation of cryptic splice sites. In our previous study[19], a subset of the TCGA breast cancer patient data was evaluated with IT-based tools, identifying 988 mutations as significantly altering normal splicing by Veridical (19% of total mutations flagged by IT). This database greatly expands the size of the repository. Here, a higher ratio of rare or novel mutations have been validated by Veridical (24% of total mutations were flagged by IT). The higher yield found could be related to the same mutation being present in multiple samples from the same tumor type and other tumor tissues, which would be expected to increase the probability of observing abnormally expressed splice forms for the mutation.

bioRxiv

An earlier version this article is available from bioRxiv: https://doi.org/10.1101/474452[26]

Software availability

Archieved code and scripts used as part of this study are available from Zenodo,

Zenodo: Validated Splicing Mutations Beacon API http://doi.org/10.5281/zenodo.1579898[27]

Zenodo: Validated Splicing Mutations Website http://doi.org/10.5281/zenodo.1579822[28]

Zenodo: Expression Data Processing, Histogram input generation and IGV Bash Script Generating Programs http://doi.org/10.5281/zenodo.1582421[29]

Data availability

Zenodo: **Dataset 1. Validated natural and cryptic mRNA splicing mutations.** Source data computed by the Shannon pipeline and Veridical, displayed on the ValidSpliceMut website (https://validsplicemut.cytognomix.com/). DOI: http://doi.org/10.5281/zenodo.1488211[20]

Zenodo: **Dataset 2. Mutations which lead to multiple types of aberrant splicing.** Representative set of mutations which significantly alter splicing in all evidence types analyzed by Veridical (i.e. cryptic splice site use, exon skipping, intron inclusion). Mutations are linked to their page on https://validsplicemut.cytognomix.com/, which provides additional material such as RNA-Seq images of the regions of interest. DOI: https://dx.doi.org/10.5281/zenodo.1489941[25]

Consent

Controlled-access TCGA sequence data was accessed with permission from NCBI (dbGaP Project #988: "Predicting common genetic variants that alter the splicing of human gene transcripts"; Approval Number #13930-11; PI: PK Rogan) and the International Cancer Genome Consortium (ICGC Project #DACO-1056047; "Validation of mutations that alter gene expression").

Grant information

PKR is supported by The Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2015-06290], Canadian Foundation for Innovation, Canada Research Chairs, and CytoGnomix. Compute Canada and Shared Hierarchical Academic Research Computing Network (SHARCNET) provided high performance computing and storage facilities.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We acknowledge Coby Viner, Stephanie Dorman, William J. Phillips and Ujani Hazra for their contributions to the early stages of this project. We are grateful to Max Barkley, Milan Panik and Miro Cupak (DNAStack) for their assistance in integrating our ValidSpliceMut Beacon into the GA4GH network.

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.
   PubMed Abstract | Publisher Full Text | Free Full Text
2. Richards S, Aziz N, Bale S, *et al.*: **Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.** *Genet Med.* 2015; **17**(5): 405–424.
   PubMed Abstract | Publisher Full Text | Free Full Text
3. 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.
   PubMed Abstract | Publisher Full Text | Free Full Text
4. 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.
   PubMed Abstract | Publisher Full Text | Free Full Text
5. Mucaki EJ, Ainsworth P, Rogan PK: **Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants.** *Hum Mutat.* 2011; **32**(7): 735–742.
   PubMed Abstract | Publisher Full Text
6. Mucaki EJ, Shirley BC, Rogan PK: **Prediction of mutant mRNA splice isoforms by information theory-based exon definition.** *Hum Mutat.* 2013; **34**(4): 557–565.
   PubMed Abstract | Publisher Full Text
7. Rogan PK, Svojanovsky S, Leeder JS: **Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations.** *Pharmacogenetics.* 2003; **13**(4): 207–218.
   PubMed Abstract
8. Rogan PK, Schneider TD: **Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites.** *Hum Mutat.* 1995; **6**(1): 74–76.
   PubMed Abstract | Publisher Full Text
9. Rogan PK, Faux BM, Schneider TD: **Information analysis of human splice site mutations.** *Hum Mutat.* 1998; **12**(3): 153–171.
   PubMed Abstract | Publisher Full Text
10. Peterlongo P, Catucci I, Colombo M, *et al.*: ***FANCM* c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor.** *Hum Mol Genet.* 2015; **24**(18): 5345–5355.
    PubMed Abstract | Publisher Full Text | Free Full Text
11. 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.
    PubMed Abstract | Publisher Full Text | Free Full Text
12. Caminsky NG, Mucaki EJ, Perri AM, *et al.*: **Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known *BRCA* Mutations.** *Hum Mutat.* 2016; **37**(7): 640–652.
    PubMed Abstract | Publisher Full Text
13. Yang XR, Devi BCR, Sung H, *et al.*: **Prevalence and spectrum of germline rare variants in *BRCA1*/*2* and *PALB2* among breast cancer cases in Sarawak, Malaysia.** *Breast Cancer Res Treat.* 2017; **165**(3): 687–697.
    PubMed Abstract | Publisher Full Text
14. Dos Santos ES, Caputo SM, Castera L, *et al.*: **Assessment of the functional impact of germline *BRCA1*/*2* variants located in non-coding regions in families with breast and/or ovarian cancer predisposition.** *Breast Cancer Res Treat.* 2018; **168**(2): 311–325.
    PubMed Abstract | Publisher Full Text
15. Burke LJ, Sevcik J, Gambino G, *et al.*: ***BRCA1* and *BRCA2 5′* noncoding region variants identified in breast cancer patients alter promoter activity and protein binding.** *Hum Mutat.* 2018; **39**(12): 2025–2039.
    PubMed Abstract | Publisher Full Text
16. 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.
    PubMed Abstract | Publisher Full Text | Free Full Text
17. Global Alliance for Genomics and Health: **GENOMICS. A federated ecosystem for sharing genomic, clinical data.** *Science.* 2016; **352**(6291): 1278–1280.
    PubMed Abstract | Publisher Full Text
18. Shirley BC, Mucaki EJ, Whitehead T, *et al.*: **Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences.** *Genomics Proteomics Bioinformatics.* 2013; **11**(2): 77–85.
    PubMed Abstract | Publisher Full Text | Free Full Text
19. Dorman SN, Viner C, Rogan PK: **Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer.** *Sci Rep.* 2014; **4**: 7063.
    PubMed Abstract | Publisher Full Text | Free Full Text
20. Mucaki EJ, Shirley BC, Rogan PK: **Dataset 1. Validated natural and cryptic mRNA splicing mutations [Data set].** *Zenodo.* 2018.
    http://www.doi.org/10.5281/zenodo.1488211
21. Su AI, Wiltshire T, Batalov S, *et al.*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A.* 2004; **101**(16): 6062–6067.
    PubMed Abstract | Publisher Full Text | Free Full Text
22. Kahles A, Lehmann KV, Toussaint NC, *et al.*: **Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients.** *Cancer Cell.* 2018; **34**(2): 211–224.e6.
    PubMed Abstract | Publisher Full Text
23. Jayasinghe RG, Cao S, Gao Q, *et al.*: **Systematic Analysis of Splice-Site-Creating Mutations in Cancer.** *Cell Rep.* 2018; **23**(1): 270–281.e3.
    PubMed Abstract | Publisher Full Text | Free Full Text
24. Futreal PA, Coin L, Marshall M, *et al.*: **A census of human cancer genes.** *Nat Rev Cancer.* 2004; **4**(3): 177–183.
    PubMed Abstract | Publisher Full Text | Free Full Text
25. Mucaki EJ, Shirley BC, Rogan PK: **Dataset 2. Mutations which lead to multiple types of aberrant splicing.** *Zenodo.* 2018.
    http://www.doi.org/10.5281/zenodo.1489941
26. Shirley BC, Mucaki EJ, Rogan PK: **Pan-Cancer Repository of Validated Natural and Cryptic mRNA Splicing Mutations.** *bioRxiv.* 2018; 474452.
    Publisher Full Text
27. Shirley BC, Mucaki EJ, Rogan PK: **Validated Splicing Mutations Beacon API (Version 1.0.0).** *Zenodo.* 2018.
    http://www.doi.org/10.5281/zenodo.1579898
28. Shirley BC, Mucaki EJ, Rogan PK: **Validated Splicing Mutations Website (Version 1.0.0).** *Zenodo.* 2018.
    http://www.doi.org/10.5281/zenodo.1579822
29. Mucaki EJ, Shirley BC, Rogan PK: **Expression Data Processing, Histogram input generation and IGV Bash Script Generating Programs.** *Zenodo.* 2018.
    http://www.doi.org/10.5281/zenodo.1582421

Open Peer Review

Current Peer Review Status: ✓ ?

Version 1

Reviewer Report 07 January 2019

https://doi.org/10.5256/f1000research.18813.r41665

? **Francesca D. Ciccarelli**
Cancer Systems Biology Laboratory, Francis Crick Institute, London, UK

The paper entitled "Pan-cancer repository of validated natural and cryptic mRNA splicing mutations" by Shirley, Mucaki and Rogan describes an extensive analysis of pancancer somatic variants in samples of the TCGA dataset to identify mutations that affect splicing. To this aim, the authors combine the methods that they previously developed to predict mutations affecting splicing, with a validation of their effect on matched mRNA data from TCGA.

The study is technically sound and follows a line of investigation that has been a long standing interest of the authors. Despite this, I have a number of comments that hopefully will help strengthen the study:

1. The authors write that their IT-based framework to predict slicing variants "has been validated in multiple studies" and they refer to numerous papers. However, in all of them they act as co-authors, showing that their method is mostly used by themselves and their collaborators. This is not necessarily a problem, but it would certainly strengthen the study if the authors would perform a comparative assessment of their performance with other available methods to predict splicing mutations, for example those in dbNSFP. This will provide a less biased interpretation of the final results.
2. Somehow related to the previous point, the authors mention that their results "contrast with another TCGA study that investigated alternative mRNA splicing". In my opinion this point should be further explored: what are the main differences and what is the extent of overlap in concordant predictions? What are the possible reasons for these differences? This is important because the cited paper in Cancer Cell analysed the same dataset of mutations.
3. The authors notice that the number of variants which activate cryptic splicing exceed the number reported in a recently published study in Cell Reports. Similarly to before: what is the extent of overlap between the two datasets? Stating that a dataset is bigger than another one is not necessarily an indication that it is better.
4. The authors validate ~27% of predicted splicing variants using the mRNA data (351k validated of the 1.2M predicted). This is a surprisingly low fraction. Later in the manuscript, the authors briefly discuss about the possible reasons of such a discrepancy. One of them is the possible occurrence of nonsense mediate decay which will not confirm the mutations because no or very few reads will be detected. However, as the authors acknowledge, the absence of supporting reads only in

mutated individuals as compared to the presence of reads in WT sample would be a strong indication of the effective role of these mutations on splicing. This can be quantified from the same RNAseq data and in my opinion should be done.

5. In general, the authors seem to exclude that their prediction method could lead to false positives. Rather they justify the poor overlap with limitations of mRNA detection. If this is the case, this should be quantified and probably a comparison with other prediction methods could help.
6. Of the >351k mutations with an effect on splicing supported by RNA data, only 35 affect CGC genes. Is this only a subset of mutations affecting driver genes or is it the complete list? In the former case, I would suggest that the authors provide the full list as supplementary data. In the latter case, the authors should discuss the implication of such a low number. Considering that there are >700 CGC genes, does it mean that aberrant slicing is very rarely a driver event? Is the overwhelming majority of splicing variants passenger?

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational cancer genomics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 Mar 2019

Peter Rogan, University of Western Ontario, London, Canada

We thank the reviewer for their valuable comments. Our responses follow:

1. The authors write that their IT-based framework to predict slicing variants "has been validated in multiple studies" and they refer to numerous papers. However, in all of them they act as co-authors, showing that their method is mostly used by themselves and their collaborators. This is not necessarily a problem, but it would certainly strengthen the study if the authors would perform a comparative assessment of their performance with other available methods to predict splicing mutations, for example those in dbNSFP. This will provide a less biased interpretation of the final results.

Response: We have previously compared our mutation prediction methods with others. Mucaki *et al.* Hum. Mut. 34:556-65 (2013) showed that IT-based single mutation and exon definition methods performed as well or better than MaxEntScan and Human Splice Finder websites. MaxEntScan computes relative entropy which is similar to IT, except it applies a correction for local base composition (which does not measure free energy, in contrast with IT: J. Theor. Biol. 201:87-92, 1989). Human Splice Finder does not measure changes in binding affinity and its basis is *ad hoc*. We also reviewed all articles (300+) which have used IT-based tools to predict changes in splicing (Caminsky *et al*., F1000Research 3:282, 2014). This reference covers the vast majority of studies that used IT-based bioinformatic tools for mutation analysis and, compared results obtained using these tools with other available software. The cited studies included a large proportion of mutations that we, ourselves, did not coauthor, or were analyzed by others, removing an obvious source of bias.

Regarding potential bias in our results, the IT-based position weight matrices (iPWMs) of splice recognition sites that we derived and use are based on a comprehensive set of splice sites spanning all known coding genes (see appendix of Rogan *et al.* Pharmacogenetics & Genomics 13(4):207-18, 2003). Other bioinformatic methods for splice site detection are based on many fewer splice sites for PWMs and are much more likely to be subject to bias based on how those sites were chosen. Also, the determination of information content in natural and mutated splice sites obeys the second law of thermodynamics (Schneider, J. Theor. Biol. 189:427-41, 1997); information contents have been formally proven to be related to binding affinities of splice site to splicesomes and splicing factors. MaxEntScan differs from IT because it applies a correction for local base composition, which is energetically and biochemically irrelevant to binding site affinity, and is therefore, biased.

As comparisons of IT with other methods have been covered previously, and the F1000Research Data Note article format is intended to specifically present results using the new resource we describe, in our opinion, reevaluation of other algorithms would not add anything of value to this manuscript.

2. Somehow related to the previous point, the authors mention that their results "contrast with another TCGA study that investigated alternative mRNA splicing". In my opinion this point should be further explored: what are the main differences and what is the extent of overlap in concordant predictions? What are the possible reasons for these differences? This is important because the cited paper in Cancer Cell analysed the same dataset of mutations.

Response: The paper being referred to in the reviewer's comment is Kahles *et al.*, Cancer Cell., 34(2):211–224.e62018. This paper reports: "In a joint analysis of *cis* and *trans* associations with 50% prior on each type, we identified 32 *cis*- and seven *trans*-sQTL (Bonferroni corrected p < 0.05)." Information regarding these cis- sQTLs can be found here: https://api.gdc.cancer.gov/data/3920a044-6874-4049-8010-55f0922243b7.

However, the document that provided does not indicate that these are actual splicing mutations. The document contains SNP coordinates, but not sequence changes. An assessment of known rsIDs at these coordinates only accounted for only 7 of the variants in our database. In fact, none of the substitutions (nor the rsIDs) in the article could be related back to changes in mRNA splice strength. It is not possible to comparison the results presented in this paper to those in Kahles *et al*. (2018).

3. The authors notice that the number of variants which activate cryptic splicing exceed the number reported in a recently published study in Cell Reports. Similarly to before: what is the extent of

overlap between the two datasets? Stating that a dataset is bigger than another one is not necessarily an indication that it is better.

Response: Jayasinghe *et al.* Cell Rep.;23(1):270-281.e3 (2018) identified 2056 variants in TCGA patients which activated cryptic splice sites, of which 1964 were confirmed by manual review of RNAseq data using the Integrated Genome Viewer (IGV). The data provided (Supplementary Table S1; "Passed Manual Review" tab) was sufficient to perform a comparison with our results. We scanned their manually reviewed variants using the Shannon Pipeline (SP). Despite reporting 1964 manually reviewed mutations, 50 mutations are shared between multiple patients (there are 1914 unique mutations total). 1510 of these mutations were found to alter at least one mRNA splice site (natural or cryptic), of which 1176 met our SP filtering criteria (either decreased natural site strength or cryptic splice sites strengthened by ≥2 bits and exceeding the strength of the nearest natural site of the same polarity).

We considered the possibility that splicing mutations in Jayasinghe *et al.* that were not flagged by SP could have instead altered the strength of splicing regulatory factor binding sites (SRFs). The high-throughput IT-based variant analysis tools needed to address this question were not available at the time the TCGA genomic data were processed. We recently introduced a new version of SP which is capable of analyzing variants impacting both constitutive splice sites and SRFs (including SRSF1, SRSF2, SRSF5, SRSF6, hnRNPA1, ELAVL1, PTB and TIA1). Upon analysis of the variants in Jayasinghe *et al.* for IT-derived changes in natural, cryptic and SRF binding site strengths, 1746 of the 1914 (91.2%) were found to be significant. It is conceivable that remaining unclassified variants may affect binding by splicing factors for which we have not yet derived iPWMs (i.e. SRSF7).

Of the 1176 variants meeting our filtering criteria, 824 were flagged by Veridical (70.1%). Interestingly, 27 Veridical-flagged mutations alter the same genomic coordinate in different tumours and another 64 affect the adjacent genomic coordinate within the same splice site in other individuals. We further investigated the remaining mutations (those evaluated, but not flagged) to better understand the discrepancy. In approximately 10% of cases, Veridical did not find any alternative validating cryptic splicing events in the region that contrasted with the read distribution in the set of control transcriptomes. For example, chr9:35389842G>C in TCGA-CN-5370 was expected to abolish the natural acceptor site of *UNC13B* exon 24, however the read counts for intron inclusion in the RNAseq data were too sparse to be deemed significant (p=0.33).

Jayasinghe *et al.* also found mutations in TCGA patients that were not evaluated in our study. This could occur for either because the RNASeq BAM file for a particular TCGA patient failed to download, or the "key" file that associates the BAM file to its TCGA name was incomplete in some tumors (i.e. TCGA-CG-4436; TCGA-STAD), due to both the BAM file name and header lacking this information. This would not impact the accuracy of the data that is present in our database or that we report, only the level of concordance between our results and those of Jayasinghe *et al*.

When comparing this dataset with our own, we discovered an instance where discrepant RNAseq data for the same tumour in the same TCGA patient led to different IT results. Originally, Veridical did not find a significant splicing change in *POLR1B* exon 14 at the +1 position, a G>A mutation g.113331138G>A, present in patient TCGA-Z6-AAPN in the ESCA dataset. When we manually reviewed the RNAseq BAM file used in this analysis, alternate or exon skipped splice forms were not observed, confirming the results reported from Veridical. In fact, TCGA deposited two separate BAM files containing RNAseq data for this same patient:

“TCGA-Z6-AAPN-01A-11R-A406-31_rnaseq.bam” and “UNCID_2681450.b17c4505-8a84-4cdd-8782-fbc456deb2a6.sorted_genome_alignments.bam”. The latter BAM file shows evidence of both predicted exon skipping and activation of a cryptic acceptor site 12 nt downstream of exon 14 of *POLR1B*. Comparison of SNPs between these BAM files in other genes did not show any evidence of sample switching or contamination. Although this issue does not appear to be widespread in the TCGA dataset, such discrepancies exceed the scope of our study, and rightfully should be addressed by TCGA.

Nevertheless, this discovery did prompt us to re-analyze the TCGA-ESCA variants through Veridical using the second set of BAM files, and we have now included those results to ValidSpliceMut. In this new set, the previously mentioned *POLR1B* mutation is deemed significant due to increased exon skipping (86 reads showing exon skipping; p=0.0000).

4. The authors validate ~27% of predicted splicing variants using the mRNA data (351k validated of the 1.2M predicted). This is a surprisingly low fraction. Later in the manuscript, the authors briefly discuss about the possible reasons of such a discrepancy. One of them is the possible occurrence of nonsense mediate decay which will not confirm the mutations because no or very few reads will be detected. However, as the authors acknowledge, the absence of supporting reads only in mutated individuals as compared to the presence of reads in WT sample would be a strong indication of the effective role of these mutations on splicing. This can be quantified from the same RNAseq data and in my opinion should be done.

In the revision of this paper, the fraction of validated predicted splicing variants in the ValidSpliceMut database increased from 27% to 31%, as a result of a series of improvements in the software used for processing. SP was significantly upgraded over the course of this project,. The previous iteration would incorrectly report information changes at pre-existing splice sites adjacent to certain mutations. These sites were characterized by genomic coordinates that were altered by insertions or deletions (indel), regardless of whether the site overlapped or included the sequence change. In such instances, some altered natural splice sites could be designated as cryptic sites. SP also reported changes at cryptic acceptor and donor sites in the first and last exons of a gene, respectively, which were not likely to have a meaningful impact on splicing. Therefore, all datasets processed prior to this upgrade were reanalyzed for indel variants. The TCGA ESCA dataset was reprocessed with a second set of RNAseq BAM files (see above response to point 3), which increased the fraction of flagged mutations for that tumor type. Finally, we processed an additional 7 tumor datasets from ICGC and included the validated mutations in our primary beacon database. The statistics of mutations, their distributions and support have been updated in the present version of the manuscript.

From our perspective, the proportion of variants validated is not “a surprisingly low fraction". The results reported here are consistent with our previous published studies (Dorman *et al.*, Sci. Rep. 4: 7063, 2014). Aside from NMD, as demonstrated below, some mutations that significantly alter splice site strength may not have been flagged by Veridical as a consequence of low levels of expression of the gene in the tumor (or controls) itself. Furthermore, Veridical cannot make an accurate assessment of the region of interest in control samples if these lack sufficient read abundance levels to determine the probability (p-value) of observing expression in the mutation-containing vs control samples. Also, our analysis did not take into account other impacts of the variant on sequences that influence exon definition, such as binding to splicing regulatory factors or mRNA secondary structure. In our response to point 3, we described a discrepancy in BAM file sources, which could also lead to a lower fraction of confirmed variants. Finally, miscalled variants (despite the stringent quality control criteria applied to select variants) could contribute to the fraction of variants not supported by Veridical analysis. Such technical artifacts have been shown to be quite common in exome sequencing in areas of the genome characterized by low mappability (from Shi *et al.*, Cell Rep. 2018 Nov 6;25(6):1446-1457, 2018): "Examination of the genomic locations of mutations revealed that 41.1% of the artifactual somatic mutations occurred in regions of low mappability compared with only 6.4% for the validated somatic heterogeneous mutations."

As indicated in our response to point 3, the previous version of this paper only evaluated variants for their impact on constitutional mRNA splice sites and cryptic sites, and excluded impacts of mutations at splicing regulatory factor binding sites (SRFs). The scope and time required to assess SRFs precluded the reanalysis of all datasets for such changes. However, to address the issue raised by the reviewer, we evaluated the degree to which ignoring SRFs would affect the overall discovery of splicing-related variants. The updated version of the Shannon Pipeline (SP) with this capability was used to examine 1050 mRNA splicing variants that have been demonstrated to affect exon recognition (Cheung *et al.*, Mol Cell. 2019 Jan 3;73(1):183-194.e8). These splicing variants were experimentally validated using a high throughput, multiplexed splicing minigene reporter assay in that study. SP reported a change in splicing and/or SRF binding strength for 1017 of these 1050 mutations (96.9%; where change in SRF strength was ≥ 3 bits). After accounting for SRF location (e.g. exonic TIA1 sites were eliminated, since these have not been proven to have splicing effects; see Table 1 of Caminsky *et al.*, F1000Res.;3:282, 2014, for a full description of each SRF), the number of flagged variants was reduced to 940 (89.5%). Based on changes in constitutive splice site strength alone, 447 variants were flagged (435 weaken natural sites, 14 strengthen cryptic sites to a level exceeding the nearest natural site). Therefore, 46.3% of the constitutive mutations at natural or cryptic splice sites were also flagged by SP. This suggests that the lower predictive accuracy of SP in our original submission was, in part, due to the limitations in its ability to detect pathogenic mutations in SRF binding sites.

We addressed the reviewer's suggestion to compare expression of the same gene in tumours with Veridical-validated mutations with other tumors with SP-predicted mutations in the same gene that were not experimentally validated. To perform the analysis, we obtained pre-processed mRNA expression data from the same RNAseq sources of TCGA patients from cBioPortal (www.cbioportal.com; provisional datasets were used, which contained largest number of patients for each tumor type). We extracted these gene expression values with a software program we wrote that determined transcripts per million (TPM) for each gene containing a SP-flagged variant. Expression values for gene present multiple times (due to the presence of multiple splice isoforms) were averaged for the particular tissue from which they were derived.

We separated mRNA expression values for each gene in TCGA patients into the Veridical-flagged vs. non-validated SP-predicted mutation categories, and performed a Student's t-test on the two groups. The expression values of 58.2% of genes were statistically distinct with 90% confidence; >2 patients per category per gene). With at least 10 patients per category, the number of statistically different genes increased to 69.3%. Among these genes, *patients with Veridical-flagged variants had higher overall gene expression in 99.7% of cases*. These inherent differences in expression suggest that the failure to validate predicted mutations may be related to little or no expression of these genes in tumour and/or control samples, rather than to accuracy of IT prediction methods. Non-sense mediated decay could be responsible for the decreased expression of these mutated genes in the tumour genomes that carry them, or the failure to validate could be related to low levels of expression of these genes in the particular tissues from which the tumours were derived. This analysis is now described in the revised manuscript ('Dataset validation and discussion,' para. 8).

Note: The gene expression data from cBioPortal had some limitations: 4196 genes containing variants flagged by SP are not present in the mRNA expression datasets, though the vast majority occurred in non-coding RNAs (i.e. 145 microRNAs, 194 LINC RNAs) or other uncharacterized RNAs (e.g. 324 'LOC' RNAs). Furthermore, certain TCGA patients that we analyzed were not available in cBioPortal among the available expression datasets (2 TCGA-BRCA, 18 TCGA-COAD, 19 TCGA-GBM, 1 TCGA-HNSC, 1 TCGA-LUAD, 119 TCGA-OV, 4 TCGA-READ, 4 TCGA-STAD, 4 TCGA-THCA, and 7 TCGA-UCEC patients). Nevertheless, sufficient data were obtained for the analysis that we carried out.

Furthermore, this analysis has other caveats, especially regarding the accuracy of RNAseq quantification between samples, even upon normalization of the data. The vast majority of TCGA tumor samples do not have a matched normal counterpart, and many TCGA tumor datasets (e.g. TCGA-LAML) do not provide any normal control RNAseq data at all. While we could perform an analysis in which the tumor is compared to a set of normals for the same tissue, variation in expression between different individuals would obfuscate evidence of any apparent NMD caused by the mutation. A 50% reduction of expression (or less) may not be observable under these conditions.

5. In general, the authors seem to exclude that their prediction method could lead to false positives. Rather they justify the poor overlap with limitations of mRNA detection. If this is the case, this should be quantified and probably a comparison with other prediction methods could help.

Response: We have quantified RNA expression in tumors for mutations that were validated by Veridical vs those which were not (see response to point 4). Regarding false positives: An extensive comparison of Information Theory to other bioinformatic programs which evaluate variants for splicing impact (MaxEntScan and Human Splice Finder) has been performed (Mucaki *et al.* 2013; Caminsky et al. 2014; see response to point 1). False positives are extremely rare because strength of binding sites (in bits) is directly related to their binding affinities; we have demonstrated that unused cryptic splice sites in the vicinity of natural splice sites are significantly weaker. Based on our experience and published analyses of a genome-wide site of binding sites (Rogan *et al.* 2003), such decoy splice sites are nearly always at least 4 bits (2^4 or 16 fold) weaker than sites that are actually recognized by spliceosomes.

The predictive accuracy of the IT methodology for detecting expression-validated mutations was determined to be 87.9% (762 of 867 variants from 122 different publications; changes to SRFs were included in this variant dataset; Caminsky *et al.* 2014). This value is similar to the predictions made to those of Cheung *et al.* (2019), where we predicted splice site and/or SRFs changes to 89.5% mutations experimentally validated to cause exon definition events.

The performance of IT-based methods for predicting splicing mutations has been well established over the past two and a half decades. Re-evaluation of its accuracy is not necessary, and this issue is, at best, only tangentially relevant, to the purpose of presenting the resource described in a Data Note article.

6. Of the >351k mutations with an effect on splicing supported by RNA data, only 35 affect CGC

genes. Is this only a subset of mutations affecting driver genes or is it the complete list? In the former case, I would suggest that the authors provide the full list as supplementary data. In the latter case, the authors should discuss the implication of such a low number. Considering that there are >700 CGC genes, does it mean that aberrant splicing is very rarely a driver event? Is the overwhelming majority of splicing variants passenger?

Response: There are 25 CGC genes indicated in Table 2, however these were never intended to be interpreted to be a complete list of CGC genes with Veridical-flagged mutations. The table has been renamed to indicate that these are a set of representative mutations. In the previous version of this paper, the number of variants ("n=25") did not indicate the total number of CGC genes. This has been removed and replaced with the actual number of CGC splicing mutations that were validated:

"In Table 2, we highlight a subset of validated splicing mutations which were identified in known driver genes implicated in the COSMIC (Catalogue Of Somatic Mutations In Cancer) Cancer Gene Census catalog (CGC) [27]. In total, 543 "Tier 1" CGC genes have at least one Veridical-flagged variant present in the ValidSpliceMut database."

Competing Interests: Not applicable.

Reviewer Report 12 December 2018

https://doi.org/10.5256/f1000research.18813.r41666

✓ **Emanuele Buratti**

International Centre for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy

The increasing amount of sequencing data that is being generated in many biological systems has represented a real challenge to researchers in terms of trying to link individual changes to a particular biological process. The attempt, described in this work, to use IT approaches to evaluate the potential biological significance can significantly contribute to fill this gap. The laboratory of Peter Rogan has a long standing and internationally prominent role in addressing the possible consequences of sequence variants on the pre-mRNA splicing process especially with regards to its connection with human disease. The ValidSpliceMut developed in this work presents a user friendly interface that allows users to manually search for a variant (by gene name or genome coordinate range) and obtain information with regards to its possible effect on splicing. This will greatly help to better appreciate the functionality of Variants of Unknown Significance that are currently abundant genomic and transcriptomic Atlases.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I have more than twenty years experience in the investigation of pre-mRNA splicing processes and especially their potential connection with a variety of human diseases, both monogenic (Cystic Fibrosis, Pompe Disease, Neurofibromatosis) and polygenic (Amyotrophic Lateral Sclerosis, Frontotemporal Dementia). I am the author of more than 160 research papers in peer-reviewed publications and of several articles in scientific books on these subjects (orcid.org/0000-0002-1356-9074)

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.