Changes to ValidSpliceMut Database, Version 3

While performing post-publication analysis of the ValidSpliceMut database, we noticed that several novel mutation clusters in the same individuals were identified in the TCGA-OV dataset (>10). These patterns are unlikely to have arisen according to typical mutation signatures that have been reported in cancer (Kucab et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16).

Upon further investigation, we determined that many the variants called in TCGA-OV datasets downloaded directly from the US NCI GDC repository (in VCF format) frequently contained sequence reads spanning intron/exon boundaries with clusters of sequence variants misaligned to intronic sequences of the reference genome (hg19). BLAST analysis of the intronic portion of these reads demonstrated that they aligned precisely with the adjacent exon (in a few instances, with rare cryptic exons occurring within the same intron). Had these VCF-derived variants been true genomic mutations, the reads from which they were derived should not have spanned the intron/exon boundary, nor contained sequences from the adjacent exon sequences. They did not meet the definition of genomic reads as expected from DNAseq, but rather junction-spanning mRNA reads, according to the Veridical definition. The origin of these discrepancies appears to be a lack of stringency in alignment algorithms that incorporated the RNASeq data and subsequent contamination of the BAM files from which the VCF variant calls were made. In addition to the erroneous reads, these exons also had reads that were properly aligned over the junction to the neighboring exon. This issue led to Veridical (our software that identifies mutation-directed splicing changes in RNAseq) misinterpreting these reads and erroneously implicating any mutation affecting the exon as causing junction-spanning intron inclusion.

Upon further investigation, we realized that this issue was compounded by the fact that RNAseq-based variant calls (VCF files) were included in the Veridical analysis of TCGA-OV, as well as to a smaller extent, TCGA-ESCA and TCGA-STAD datasets. This would normally not be a major issue, as RNAseq-derived variants within exons were found in many cases to be real (and cause splicing mutations). However, the previously described issue with alignment in TCGA-OV exacerbated the issue, as these misalignments were misinterpreted by the variant-calling software as true (QC passing) mutations. The TCGA-ESCA and TCGA-STAD datasets also contained limited RNA-seq data which lead to the database containing RNAseq-exclusive called mutations (not called in DNAseq). However, only 76 TCGA-ESCA and 292 TCGA-STAD ValidSpliceMut entries were associated with variants exclusively called using RNAseq data. The number of RNASeq derived variants in TCGA-OV was much higher (N=104,920). Thus, RNAseq BAM files of TCGA-ESCA and TCGA-STAD patients lack the alignment errors observed in TCGA-OV.

We developed an algorithm and software to identify and evaluate these false positive, miscalled "variant" clusters from the Shannon Pipeline output from our original TCGA-OV analysis. The algorithm identified regions of high variant density in the intronic sequences adjacent to an exon (any junction with > 3 mutations within the first 10nt of the intron). Once identified, the sequences of all potential neighboring exons of interest were extracted (via SAMTools) and their sequences were aligned to the positions of mutations in the cluster. If the alternative nucleotides of each mutation matched the reference sequence of the neighboring exon with at least 50% identity, this indicated a likely misalignment of RNA-Seq derived read(s) in the original tumour genomic sequence. In such cases, the ValidSpliceMut entry (if present) for the mutation in this particular tumour was eliminated. The 50% threshold was chosen due to an investigation of multiple regions with 40-60% similarity to the next exon, whereby the lower value was determined to be caused by an association to a neighboring exon that was not present in the RefSeq transcript data used. If there was a <50% sequence identity match to the next exon, the likelihood that the cluster arose naturally is increased, especially if they are no gaps in mutations within the cluster. Therefore, we only eliminate clusters of variants if they have at least one gap between two or more mutations.

While we initially performed the mutation clustering analysis using a 10nt intronic window, the resulting report would only contain called "mutations" within that 10nt window. However, these misaligned clusters can cause multiple instances of mutations that can extend 30-40nt based on the lengths of the sequenced reads responsible for these apparent changes. Therefore, we performed a second analysis in which the window length was expanded to 50nt. Thus, if a cluster was identified within the 10nt window, we eliminated all mutations associated with the same splice site indicated within the 50nt window. Note that there are some clusters in certain tumours that were exclusively identified in the second analysis but not the initial 10nt window. If these extended clusters were a 100% match to the next exon, we filtered out these mutations as well. When these filters are applied, the number of ValidSpliceMut entries decreased from 802,701 to 734,662, while the total number of unique mutations is reduced from 341,486 to 309,848 (N=31,638 unique mutations removed total). Additional statistics can be seen in the table below.

	Entries in DB Assoc. with RNAseq-only Variants [% of Tissue Type]	Removed Entries in DB Assoc. with RNAseq-only Variants	Unique RNAseq-only Variants in DB [% of Tissue Type]	Removed Unique RNAseq-only Variants in DB
TCGA-ESCA	76 [0.2%]	47 [0.1%]	74 [0.4%]	47 [0.2%]
TCGA-STAD	292 [1.0%]	191 [0.6%]	287 [1.4%]	187 [0.9%]
TCGA-OV	104,920 [63.4%]	67,801 [41.0%]	48,688 [55.2%]	31,592 [35.8%]

DB – ValidSpliceMut database; "RNAseq-only variants" are variants that were not identified in DNAseq, but exclusively via RNAseq BAM files for these tumor patients; % are based on the total number of entries / unique mutations in ValidSpliceMut for each tissue type, which are as follows (# of entries / # of unique mutations): TCGA-ESCA – 33,085 / 19,361; TCGA-STAD: 30,229 / 20,245; TCGA-OV: 165,363 / 88,136

Mutations called exclusively from RNASeq data in ValidSpliceMut will now be indicated with an additional, separate field in the database. It will be annotated on the results webpage generated from user queries of the database.

Please note that we intend to completely repeat the Shannon pipeline and Veridical analysis of the TCGA-OV dataset excluding RNAseq-based VCF files. We performed and released the current update, as we felt that it was important to correct the database immediately for false positive splicing mutations due to mis-mapping of RNA-Seq derived reads prior to a complete reanalysis of this dataset.