Prediction of tumor-specific splicing from somatic mutations as a source of neoantigen candidates

Franziska Lang[1], Patrick Sorn[1], Martin Suchan[1], Alina Henrich[1], Christian Albrecht[1], Nina Koehl[1], Aline Beicht[1], Pablo Riesgo-Ferreiro[1], Christoph Holtsträter[1], Barbara Schrörs[1], David Weber[1], Martin Löwer[1], Ugur Sahin[1,2,3,#], Jonas Ibn-Salem[1]

[1]TRON – Translational Oncology at the University Medical Center of Johannes Gutenberg University Mainz gGmbH, Mainz, Germany.

[2]BioNTech SE, Mainz, Germany.

[3]Johannes Gutenberg University Mainz, Mainz, Germany

[#]Corresponding author: Ugur Sahin, sahin@uni-mainz.de

Abstract

Background

Splicing is dysregulated in many tumors and may result in tumor-specific transcripts that can encode neoantigens, which are promising targets for cancer immunotherapy. Detecting tumor-specific splicing is challenging because many non-canonical splice junctions identified in tumor transcriptomes also appear in healthy tissues. However, somatic mutations can disrupt canonical splicing motifs or create novel ones leading to true tumor-specific targets.

Results

Here, we developed splice2neo to integrate the predicted splice effects from somatic mutations with splice junctions detected in tumor RNA-seq for individual cancer patients. Using splice2neo, we excluded canonical splice junctions and splice junctions from healthy tissue samples, annotated resulting transcript and peptide sequences, and integrated targeted re-quantification of supporting RNA-seq reads. By analyzing melanoma patient cohorts, we established a stringent detection rule to predict splice junctions as mutation-derived and tumor-specific targets. In an independent melanoma cohort, we identified 1.7 target splice junctions per tumor with an estimated false discovery rate of less than 5% and established tumor-specificity using additional healthy tissue samples. For individual examples of exon skipping events, we confirmed the expression in tumor-derived RNA by quantitative real-time PCR experiments. Most target splice junctions encoded at least one neoepitope candidate with predicted MHC-I or MHC-II binding. Compared to neoepitope candidates derived from non-synonymous point mutations, the splicing-derived neoepitope candidates had a lower self-similarity to corresponding wild-type peptides.

Conclusion

Identifying mutation-derived and tumor-specific splice junctions can lead to additional neoantigen candidates to expand the target repertoire for cancer immunotherapies.

Keywords

Alternative splicing, Tumor-specificity, Neoantigen candidates, Cancer, Individualized cancer immunotherapy

Background

Neoantigens are tumor-specific mutated gene products that are presented in form of neoepitopes on major histocompatibility complex (MHC) molecules to the immune system and that are recognized by CD4+ or CD8+ T cells [1]. Individualized cancer vaccines mediate successful anti-tumor responses by targeting these neoantigens [2–5]. Previous studies mainly focused on targeting neoantigens derived from single nucleotide variants (SNVs) as the most abundant mutation type. However, neoantigens can also derive from other mutation types such as short insertions and deletions (INDELs) or gene fusions which broaden the targetable neoantigen repertoire in particular in tumor types with low SNV burden [6, 7] and which were associated with the efficacy of immune checkpoint blockade [8, 9].

During mRNA splicing, multiple isoforms per gene may be created by removal of introns and joining of exons, increasing the functional diversity of the proteome [10]. Importantly, 60% of the alternative splicing events are variable between tissues [11], and inter-individual variability contributes to individual phenotypes [12]. The highly diverse nature of splicing was recently demonstrated by long-read sequencing of more than 70,000 novel transcripts in healthy tissues [13].

Alternative splicing is specifically dysregulated in many tumors, impacting protein function and contributing to tumor heterogeneity [14–17]. For instance, disrupting mutations or expression changes of genes encoding RNA splicing regulators such as SF3B1 alter splicing specifically in tumors [15–19]. Previous studies have shown that indeed novel peptides derived by alternative splicing in tumors can be presented by MHC and elicit cytotoxic $CD8^+$ T-cell responses in HLA-A24 transgenic mice [18], uveal melanoma [19], acute myeloid leukemia [20] and lung cancer [21]. Alternative splicing can result in frameshifts of the translational reading frame, leading to mutated peptide sequences highly dissimilar to the wild-type proteome. Such dissimilarity could increase the likelihood of T-cell recognition and make them excellent targets.

However, it is challenging to ensure the tumor-specificity of such splicing events to define them as true neoantigens and utilize them as safe and effective targets in individualized cancer vaccines. A common strategy is to focus on splice junctions that are absent or low expressed in a set of matched or unmatched healthy tissues [20, 22–24]. However, given the great diversity and stochasticity of alternative splicing across healthy tissue, more evidence is required to consider splice junctions as truly tumor-specific candidates for individualized cancer vaccines [25].

Somatic mutations – both intronic and exonic – can directly alter splicing by disrupting canonical splicing motifs or creating novel splicing motifs [26–29]. Notably, integrating somatic mutations with splicing events in whole-genome sequenced pan-cancer cohorts suggested that 34% of the intronic mutations near exon–intron boundaries may affect splicing [27]. Multiple computational tools predicting the effect of mutations on splicing and recent deep learning-based methods considerably improved splicing effect prediction [30, 31].

In contrast to splice junctions detected from RNA-seq alone, expressed splice junctions that are caused by somatic mutations can be considered as truly tumor-specific splice junctions. However, it is currently unclear whether such tumor-specific splice junctions can be predicted reliably from somatic mutations for individual cancer patients and how many splicing-derived neoantigen candidates qualify as promising targets for individualized cancer vaccines.

In this study, we integrated the potential effect of whole exome sequencing (WES) derived somatic mutations on splicing with splicing events detected from RNA-seq in order to predict splicing-derived neoantigen candidates. Using stringent detection rules to predict mutation-derived splice junctions, we identified a small but relevant set of tumor-specific splice junctions that can be interesting targets for individualized cancer vaccines.

Results

Splice2neo allows unified analysis of splice junctions from mutation effects and RNA-seq

We developed the R-package splice2neo to facilitate the integration of the effect of somatic mutations on splicing and splicing events detected in RNA-seq for the prediction of splicing-derived neoantigen candidates (Fig. 1A). Splice2neo converts the output of multiple mutation effect prediction tools (MMSplice [32], SpliceAI [33]) and splicing detection tools (SplAdder [34], LeafCutter [35]) in a unified splice junction format, defined by the genomic coordinates of the resulting splice junction positions. Splice2neo is compatible with other R and Bioconductor packages and can annotate splice junctions with the resulting transcript sequence, junction position in the transcript sequence, as well as, the resulting peptide sequence. The R-package is designed as a modular library and contains a range of

functionalities for customized junction analysis that can be carried out on a joined dataset of multiple samples or for individual samples.

Identification of mutation-retrieved splice junctions with support in matched tumor RNA-seq

Somatic SNVs and INDELs are truly tumor-specific as they are detected by comparing WES in tumor and matched normal samples. Such mutations can lead to a loss or gain of splicing donor or acceptor sequence motifs. We reasoned that splice junctions can be defined as tumor-specific targets (1) if they are caused by a somatic mutation, (2) if they are non-canonical and not detected in a large cohort of RNA-seq samples of unmatched healthy tissues [33], and (3) if they are detected in the RNA-seq of the patient's tumor sample. To develop sufficiently stringent detection criteria for such targets, we analyzed matched tumor and normal WES data and tumor RNA-seq data of 85 melanoma samples from two studies (VanAllen2015 [36], Riaz2017 [37]) with splice2neo as a discovery dataset (Fig. 1B, Fig. 2A).

First, we identified between 93 and 10,931 (median 963) somatic SNVs or INDELs in exonic and intronic regions per sample (Fig. 2A, upper panel). Then, we retrieved the potential effects of the identified somatic mutations on splicing with the deep learning-based tools MMSplice [32] and SpliceAI [33]. MMSplice was used to predict if a somatic mutations could cause exon skipping (ES) events and SpliceAI was used to predict all potential 3' (A3SS) or 5' (A5SS) alternative splice sites, ES events and intron retention (IR) events. Both tools report effect scores, which reflect the probability or the effect strength of a mutation to alter splicing. Initially, we did not apply cut-offs on these effect scores to ensure high sensitivity and allow unbiased optimization of cut-off values in downstream analysis. Therefore, all potential mutation effects on splicing from MMSplice and SpliceAI, including low scoring effects, were annotated with splice2neo and converted into a common splice junction format ("mutation-retrieved splice junction"), resulting in between 53 and 19,962 (median 1,588) mutation-retrieved splice junctions per sample (Fig. 2A, middle panel). Next, we removed canonical splice junctions in reference databases (GENCODE) and normal splice junctions previously detected in RNA-seq of 1,740 samples of 53 healthy tissues from the Gene and Tissue Expression (GTEx) atlas [33, 38]. After excluding canonical and normal splice junctions, we considered between 15 and 4,952 (median 430) junctions as novel mutation-retrieved splice junctions for further analysis (Fig. 2A, middle panel).

The reported effect scores from MMSplice and SpliceAI were low for most of the novel mutation-retrieved splice junctions (Additional file 1: Fig. S1A-B) and decreased with increasing distance of the mutation to the junction position (Additional file 1: Fig. S1C). Although SpliceAI and MMSplice effect scores were correlated for jointly identified ES splice

junctions (Pearson correlation R=-0.64), there were many splice junctions with divergent splice effect scores (Additional file 1: Fig. S1D).

Next, we identified splice junctions in matched tumor RNA-seq for the 85 melanoma samples using the RNA-seq-based tools LeafCutter [35] and SplAdder [34]. The number of detected splice junctions differed strongly between the VanAllen2015 and the Riaz2017cohort, which also showed strong differences in duplication rates in the sequencing data (Additional file 1: Fig. S1E-F). Overall, we identified between 16,826 and 110,419 splice junctions per patient, of which between 1,597 and 24,480 were novel (i.e. absent from GENCODE and GTEx) (Fig. 2A, lower panel). The percentage of novel splice junctions was higher among mutation-retrieved splice junctions (26%) than among the splice junctions retrieved from RNA-seq (15%) (Additional file 1: Fig. S1G).

Next, we analyzed whether mutation-retrieved junctions are expressed in tumor samples. We overlapped the novel mutation-retrieved splice junctions (total n = 59,210) with the novel RNA-seq retrieved splice junctions in RNA-seq (total n = 642,032). For 85 melanoma patients, we identified in total 122 mutation-retrieved splice junctions that were also identified by SplAdder or LeafCutter in RNA-seq (Fig. 2B). We reasoned that targeted re-mapping of RNA-seq reads to pre-defined junction sequences could lead to a higher sensitivity. We developed and implemented the tool EasyQuant and re-evaluated all novel mutation-retrieved splice junctions by quantifying RNA-seq reads supporting the splice junctions or read coverage of retained introns (Fig. 1B, Additional file 1: Fig. S2A). We observed at least one junction read (A3SS, A5SS, and ES) or a median read coverage > 0 (IR) using this targeted approach for 732 out of 59,210 mutation-retrieved splice junctions (Fig. 2B, Additional file 1: Fig. S2B-C). Re-quantified splice junctions and splice junctions identified by SplAdder or LeafCutter had a high overlap of 102 splice junctions (83% of the RNA-seq retrieved splice junctions) (Fig. 2B). Notably, mutation-retrieved junctions supported by RNA-seq were associated with better mutation effect scores from SpliceAI and MMSplice (Fig. 2C-D).

Together, most mutation-retrieved splice junctions have very low effect scores and are not supported by RNA-seq reads. For the identification of target splice junctions that are potentially tumor-specific, we focused on the subset of mutation-retrieved splice junctions with RNA-seq support (n=752) for further analysis.

Prediction of target splice junctions in melanoma samples

Next, we wanted to identify stringent detection rules to predict which mutation-retrieved splice junctions with RNA-seq support ("candidates") are caused by a somatic mutation. To estimate the false discovery rate (FDR), we performed a sample permutation analysis by

assessing the support of such candidates in RNA-seq data from all other tumor samples without the corresponding somatic mutation. (Fig. 3A). Here, we observed that half of the candidate splice junctions from ASS or ES events with RNA-seq support in the actual tumor sample were supported by RNA-seq reads in at least one independent tumor sample, resulting in a high estimated FDR of 0.50. This indicates that many of these candidate splice junctions can occur independently of the somatic mutation and, therefore, might not be tumor-specific and that more stringent filtering is required.

Consequently, we gradually filtered candidate splice junctions by thresholds on the mutation effect scores, and the re-quantification read support (Fig. 3B). First, we focused on splice junctions from ES and ASS events. We estimated the FDR with the sample permutation analysis and selected the set of filtering thresholds with lowest estimated FDR as optimal detection rule for target splice junctions. This optimal detection rule was defined by (1) SpliceAI score ≥ 0.35 or MMSplice score ≤ -0.35, and (2) identification by LeafCutter or SplAdder or re-quantification with at least 3 junction reads by EasyQuant (Fig. 3B-C). This detection rule resulted in an estimated FDR of 0.07 and identified 69 target splice junctions for 85 melanoma tumors in the discovery cohort (Additional file 2: Table S1). We evaluated the detection rule in an independent verification cohort of melanoma patients (Hugo2016 cohort [39]) and predicted 45 splice junctions for 27 patients with an estimated FDR of 0.04 (Additional file 1: Fig. S3A-B,Additional file 2: Table S1).

Next we focused on IR events. Here, no filter combination led to an estimated FDR lower than 0.50 in the discovery and verification cohort (Additional file 1: Fig. S4), and we, therefore, excluded IR events from further analysis.

Candidate splice junctions were previously filtered against reference splice junctions identified in GTEx samples. To estimate the tumor-specificity of target splice junctions, we compared all candidate splice junctions with splice junctions identified by Leafcutter or Spladder in an additional independent RNA-seq dataset of 141 samples from 49 healthy tissues [7]. Here, only 2 out of 69 (3%) and 3 out of 45 (7%) splice junctions predicted as targets in the discovery and verification set, respectively, were found in any healthy tissue sample. This was a strong reduction compared to splice junctions that were not predicted as targets (39% in discovery set and 40% in verification set) (Fig. 3D). We also used the 141 healthy tissue samples to examine the tumor-specificity of splice junction sets derived from tumor RNA-seq alone. As expected, 97% of the canonical and 84% of the normal RNA-seq derived splice junctions overlapped with the healthy tissue splice junctions. However, also among the novel RNA-seq derived splice junctions, which were already filtered against GTEx, 58% were also found in the independent dataset of healthy tissues (Fig. 3E). Together, this data shows that the here described approach of associating splicing with

somatic mutation effects using splice2neo and the detection rule leads to strong enrichment of tumor-specific splice junction targets.

Experimental confirmation of exon skipping junctions in tumor samples

Next, we experimentally confirmed predicted targets by qRT-PCR. We analyzed WES and RNA-seq data from eight formalin-fixed, paraffin-embedded (FFPE) tumor samples [7] (Additional file 1: Fig. S5A-B, Additional file 3: Table S2). For experimental confirmation, we focused solely on splice junctions from ES events, which lead to alternatively spliced transcripts that are distinguishable from the corresponding wild-type transcripts by product size. In total, we tested 21 ES events, of which only one was predicted as a target by the established detection rule (Additional file 1: Fig. S5B-C, Additional file 4: Table S3). This target splice junction (chr20:62606885-62608324:-) was predicted to be caused by a C to A substitution on chromosome 6 at position 62,607,186, disrupting an acceptor motif in the gene *SAMD10* (Sterile Alpha Motif Domain Containing 10) (Additional file 1: Fig. S5D). The targeted re-quantification of RNA-seq reads resulted in only two supporting junction reads, but LeafCutter detected the target splice junction in RNA-seq and the associated mutation resulted in the maximal SpliceAI score of 1.0 (Additional file 1: Fig. S5D-E, Additional file 4: Table S3), indicating a strong effect on splicing.

The expression of the target splice junction and two additional non-target junctions could be confirmed with qRT-PCR on RNA level (Additional file 1: Fig. S5E, Additional file 4: Table S3). Vice versa, the expression of 18 out of 20 mutation-retrieved splice junctions that were not classified as targets could also not be confirmed by qRT-PCR.

Splice junctions can generate neoantigen candidates

Next, we examined whether the predicted target splice junctions encode novel peptide sequences and therefore qualify as neoantigen candidates. Splice2neo annotates splice junctions based on their coordinates and reference transcripts with the resulting protein sequences. For in-frame junctions, we considered the peptide sequence (-/+13 amino acids) around the junction and potentially inserted amino acids. For frame-shift junctions, the peptide sequence was extended until the next stop codon. We refer to peptide sequences from predicted tumor-specific targets as “neoantigen candidates”. First, we wanted to estimate the number of neoantigen candidates per patient in the verification cohort. In total, 42 of the 45 predicted target splice junctions affected coding sequences and were translated into at least one peptide sequence, resulting in 45 neoantigen candidates for the 27 melanoma patients (Additional file 5: Table S4). Depending on the mutation load, we predicted between zero and eight neoantigen candidates per patient, which led to an average of 1.7 neoantigen candidates from alternative splicing per patient (Fig. 4A). We

found that the majority of neoantigen candidates in the verification cohort derived from an A3SS event (42%), followed by ES events (31%) (Fig. 4B). Furthermore, 62% of splicing-derived neoantigen candidates were generated by a frameshift. While also in-frame A3SS or A5SS events could generate neoantigen candidates longer than 26 amino acids if additionally amino acids were inserted, frame-shift splice junctions led generally to longer novel peptide sequences of up to 134 amino acids (Fig. 4C-D).

To examine whether the predicted neoantigen candidates from alternative splicing could be presented by MHC-I or MHC-II molecules, we predicted MHC binding features with NeoFox [40] using patient-specific HLA alleles. Most neoantigen candidates had excellent MHC binding abilities. We found that 58% of the neoantigen candidates from alternative splicing were predicted to generate at least one strongly binding MHC-I epitope (MHC-I binding rank < 0.5), and 47% of the neoantigen candidates were predicted to generate at least one strongly binding epitope for MHC-II (MHC-II binding rank < 2) (Fig. 4E). Neoantigen candidates from alternative splicing had comparable MHC-I and MHC-II binding ranks to neoantigen candidates derived from non-synonymous SNVs of the same cohort (Fig. 4F-G). However, the best-predicted neoepitope per neoantigen candidate from alternative splicing was less self-similar to the wild-type proteome [41] compared to the best-predicted neoepitope per neoantigen candidate from nonsynonymous SNVs (Fig. 4H), indicating a stronger potential for immunogenicity in the group of neoantigen candidates derived from alternative splicing.

Discussion

Disruption of canonical splicing in tumors can generate novel gene products that might be excellent targets for individualized cancer vaccines [6, 42]. While RNA-seq can detect thousands of novel splice junctions in individual tumor samples, it is challenging to ensure their specificity to tumor cells [25, 43]. One approach is to predict neoantigen candidates from alternative splicing by combining RNA-seq from tumor samples with RNA-seq from matched healthy tissue [22]. Given the diversity of splicing across tissues and cell types, it is questionable which healthy tissue is suitable as a control and if a single normal sample is sufficient. Alternatively, tumor-specificity might be defined by the absence of splice junctions from large healthy tissue sample collections of unrelated subjects [23, 24]. Databases of canonical reference splice junctions or large RNA-seq sample collections of no healthy tissues, such as GTEx [38], are valuable resources for constructing exclusion lists. However, it remains unclear if such data is suitable to capture splicing variations in rare cell types or conditions as well as individual splicing events caused, for example, by rare germline mutations.

In this study, we hypothesized that alternative splicing created by a loss or gain of canonical splicing sequence motifs by somatic mutations are eligible as tumor-specific targets and might be suitable neoantigen candidates for individualized cancer immunotherapy approaches. With splice2neo, we implemented an analysis approach to identify mutation-retrieved splice junctions with support in RNA-seq as candidates for potentially tumor-specific targets for individual cancer patients. Splice2neo combines the splicing effects from sequence motif alterations retrieved from deep learning-based tools with support in RNA-seq in matched tumor samples.

Multiple other tools were recently described to predict neoantigen candidates from alternative splicing. NeoSplice [22] and ASNEO [23] rely on a single RNA-seq-based method for splicing detection and are end-to-end pipelines from raw files to neoantigen candidates. IRIS [24] is a modular neoantigen prediction pipeline based on RNA-seq and supports customized pipelines. Regtools [28] provides functionalities to integrate DNA sequencing and RNA-seq data to identify potential splice-associated variants. DICAST [44] integrates several RNA-seq splicing tools for unified junction analysis. Both, Regtools and DICAST lack the transcript and protein sequence annotation for functional downstream analyses. In contrast to static end-to-end pipelines, splice2neo is designed as a modular library of multiple functionalities for customized splice junction analysis. It includes the unified integration of results from upstream tools, exclusion of canonical or normal junctions, and annotation with transcript and peptide sequences. The implementation of splice2neo as an R-package allows simple installation and compatibility with multiple other methods from the Bioconductor project [45, 46] for interactive analysis or integration into target identification pipelines. While we describe a defined analysis workflow in this study, splice2neo is not limited to the currently supported upstream tools but can be easily extended in the future to support other tools for alternative splicing detection or other event types, such as splicing junctions between exons and transposable elements [21, 47]. The here proposed approach works with data from single patients, which is relevant in the context of neoantigen prediction for individualized cancer vaccines. However, splice2neo might also be used to analyze larger cohorts of tumor samples, for instance, to identify and characterize shared splicing targets.

We initially considered all novel mutation-retrieved splice junctions that were supported by RNA-seq without any cut-off on splice effect scores. Only when we applied a stringent detection rule on the mutation effect scores and RNA-seq support, we were able to decrease the FDR and enrich for expressed splice junctions that could indeed be associated with somatic mutations. By applying this detection rule to an independent verification cohort of 27 melanoma patients, we showed that, dependent on the somatic mutation burden, on average 1.7 neoantigen candidates from alternative splicing could be predicted per patient with an estimated FDR of 0.04. We assume that the number of neoantigen candidates from somatic

mutation-derived splicing is not many magnitudes higher but still can be increased by technical advances. In the future, the computational analysis with splice2neo of larger tumor cohorts from more diverse entities and using personalized reference genomes may allow for inferring more sensitive detection rules, e.g., by a machine learning approach.

Studies predicting splicing-derived neoantigen candidates from tumor RNA-seq alone or in combination with matched normal RNA-seq reported markedly higher numbers of non-canonical splice junctions per patient in several tumor entities [17, 21, 22]. However, it remains unclear whether those splice junctions identified from RNA-seq alone are truly tumor-specific and qualify as safe and effective targets in individualized cancer vaccines.

While we could enrich for a small but highly valuable set of tumor-specific targets from A5SS, A3SS and ES events, we were not able to show that candidate splice junctions from IR events are potentially tumor-specific. Even with stringent cutoffs on the mutation scores and RNA-seq support, we observed high FDRs. IR detection can be generally challenging as RNA-seq reads might derive from unspliced pre-mRNA, repeat regions, or other transcripts [48, 49]. However, it was shown that somatic mutations indeed can cause IR events [29]. Furthermore, IRs were shown to generate MHC-I binding epitopes with mass spectrometry-based immunopeptidome analysis [50] and the load of non-canonical peptides derived from IRs correlated with favorable prognosis in pancreatic cancer [51]. These observations suggest that also IRs could contribute to the neoantigen repertoire. Improved computational tools for IR detection and long read sequencing might allow to better investigate splice junctions from IRs in the future [18, 48, 49, 52].

Here, we showed that the predicted target splice junctions from somatic mutations can generate neoantigen candidates. Although these neoantigen candidates might not be as abundant as neoantigens from non-synonymous SNVs, we show that they frequently cause frameshifts and lead to longer novel peptide sequences, leading to strong MHC binding neoepitope candidates that are dissimilar to the wild-type proteome. These features may characterize splicing-derived neoantigen candidates as promising targets for individualized cancer vaccines [1]. Future studies are required to examine if the identified neoantigen candidates indeed mount functional T-cell responses upon individualized cancer vaccines.

Conclusions

Splice2neo is a versatile tool for identifying and analyzing candidate splice junctions. Combining mutation effects on splicing and RNA-seq support with a stringent detection rule predicts tumor-specific splicing and neoantigen candidates that expand the individual target repertoire for cancer immunotherapies.

Methods

Samples and sequencing data

WES and RNA-seq from the tumor and WES from a corresponding normal sample was collected from three immune checkpoint blockade studies: VanAllen2015[36] (phs000452.v2.p1), Hugo2016 [39] (WES: SRP067938, SRP090294; RNA-seq: SRP070710) and Riaz2017 [37] (WES-seq: SRP095809; RNA-seq: SRP094781). For experimental confirmations, a small cohort of distinct tumor entities was analyzed ("FFPE cohort"). WES of tumor and normal samples were performed in technical replicates as described below. The RNA-seq data of these tumor samples was published previously [7] (EGAS00001004877). The RNA-seq data from a cohort of 141 samples from 49 healthy tissues was retrieved from a previous study (PRJNA764684) [7].

Whole exome sequencing

By addition of Xylene, 10 µm curls of FFPE tissue were deparaffinized, centrifuged and washed with ethanol. After deparaffinization, lysis was performed with proteinase K digestion and RNase A digestion subsequently. The lysate was transferred to a QIAamp MinElute silica column. After on-column washing steps, the DNA was eluted in low TE buffer.

By addition of mineral oil and incubation at 80°C, 10 µm curls of FFPE tissue per sample were deparaffinized. Deparaffinization was followed by lysis with proteinase K digestion and an RNase A digestion. After centrifugation, the aqueous phase of the two separated layers containing the DNA was transferred to the Maxwell RSC cartridge for purification using paramagnetic particles. Purified DNA was finally eluted in low TE buffer. Regardless of the extraction method, final quantity was assessed using NanoDrop and Qubit 3 fluorometer with dsDNA HS Assay Kit (Invitrogen).

Whole exome libraries were prepared in duplicates with an input of 100 ng genomic DNA from FFPE tissue samples. The genomic DNA was fragmented to a size range of 200 ± 50 bp using the Covaris S220 instrument. The fragmented DNA was end-repaired and adenylated followed by ligation of sequencing adaptor and the appropriate eight-nucleotide NEXTFLEX DNA barcode (Perkin Elmer) and pre-amplification of the library using the KAPA Hyper Prep kit (Roche). Subsequently, target regions were hybridized to biotinylated baits (SureSelectXT Human All Exon v6 and SureSelect XT Reagent kit, Agilent) and isolated using streptavidin-coated magnetic beads (Invitrogen). The post-capture library was then amplified with the KAPA Library Amplification kit (Roche). All intermediate and final library purification steps were carried out using AMPure XP beads (Beckman Coulter). Final library quantity and quality was assessed using the Qubit 3 fluorometer with Qubit dsDNA HS Assay Kit (Invitrogen) and the BioAnalyzer with High Sensitivity DNA Kit (Agilent).

All libraries were sequenced in paired-end mode (2 x 50 nt) on an Illumina NovaSeq 6000 instrument resulting in around 150 million distinct sequencing reads per library.

Somatic mutation calling

For WES read alignment and somatic mutation calling, we used the tronflow pipelines [53–55] implemented in Nextflow [56]. In short, reads from WES tumor and normal sequencing were aligned to the reference genome hg19 using bwa [57] (v0.7.17). Base score recalibration was performed with GATK4 [58] (v4.2.0.0) and duplicated reads were removed with Picard (GATK4 v.2.0.0). Single nucleotide variants (SNVs) and short insertions and deletions (INDELs) were detected with Mutect2 [59] (GATK4 v.2.0.0) without restriction to any target region and using the GnomAD [60] as a resource of germline mutations. Only mutation calls with PASS in the filter column were considered and calls from individual replicates were combined per sample.

Predicting the effect of somatic mutations on splicing

The effect of somatic mutations on splicing was predicted with SpliceAI [33] (v1.3.1) and MMSplice [32] (v2.1.1). For SpliceAI, the GENCODE V24 canonical annotation (grch37) included in the package was used. GENCODE v34lift37 was used as annotation for MMSplice. Both tools were run with default settings.

Identifying alternative splicing events from RNA-seq

RNA-seq reads were aligned to the hg19 reference genome with STAR [61] (v2.7.0a). The options `twopassMode Basic --outSAMstrandField intronMotif` were used to generate BAM files for the tool LeafCutter based on recommendation in the manual. Alternative splicing events were detected with SplAdder [34] (v3.0.2) and LeafCutter [35] (v0.2.9). Splicing events from exon skipping, alternative 3' or 5', intron retention and mutually exclusive exons were identified with SplAdder, ignoring mismatch information in the alignment file and using GENCODE (v34lift37) annotation. Splicing events were identified with LeafCutter annotation-free using the default settings. Here, the tool how_are_we_stranded_here [62] (v1.0.1) was used to computationally determine the strandedness of the RNA-seq and regtools [28] (v0.5.2) was used to extract junctions. The union of RNA-seq derived junctions was considered per sample if RNA-seq replicates were available.

Splice2neo: prediction of splice junctions and derived mutated transcript and peptide sequences

We developed the R-package splice2neo (https://github.com/TRON-Bioinformatics/splice2neo) to identify splice junctions defined by the chromosome (chr), the last genomic position of the left exon (start), the first base of the right exon (end) and the transcriptional direction (strand) in the format “chr:start-end:strand“ directly from the output of multiple tools. The R-package relies mainly on the functionalities of the Bioconductor packages GenomicRanges [63] and Biostrings [64]. Splice2neo converts the raw output from LeafCutter, SplAdder, MMSplice and SpliceAI into resulting splice junctions.

LeafCutter and SplAdder provide the genomic coordinates of the identified splicing events. Splice2neo transforms all events into the standardized junction format covering all potential splice junctions including canonical junctions.

SpliceAI predicts the probability of a mutation on splicing and provides the position of the alternative splicing events relative to the mutation. Table 1 describes the rules on how splice2neo creates splice junctions based on the predicted effect by SpliceAI. Mutations can be covered by more than one gene and the splice effect of a somatic mutation may differ between genes depending of the position of the somatic mutation within the respective gene. Splice2neo optionally considers only transcripts related to the gene annotated by SpliceAI while predicting the effect on splicing. In this work, gene annotation was considered while converting the effect of somatic mutations on splicing as predicted by SpliceAI.

MMSplice predicts the change on percent spliced in (PSI) for a given annotated exon. Splice2neo considers events with PSI score <= 0 as exon skipping events and uses the delta_logit_psi output as mutation effect score. Here, splice junctions are built by the end of the upstream exon and the start of the downstream exon.

Table 1: **Rules on how splice junctions are created based on the mutation effect predicted by SpliceAI**.

Change	Class	Left junction coordinate	Right junction coordinate
Donor loss	Intron retention	pos	pos + strand_offset
Donor loss	ES	Upstream_end	downstream_start
Donor gain	A5SS	pos	downstream_start
Acceptor loss	Intron retention	pos – strand_offset	pos
Acceptor loss	ES	Upstream_end	downstream_start
Acceptor gain	A3SS	Upstream_end	pos

The table describes the rules on how the left and right coordinate of the splice junction ("chr:start-end:strand") are determined based on the predicted effect from SpliceAI. Pos refers to the position that is predicted to be affected by the mutation by SpliceAI. Upstream_end and downstream_start refer to exon end or start coordinates. Splice junctions that relate to an intron retention event are defined by the rule `chr:pos-(pos+1):strand` and must cover an exon-intron boundary. The strand_offset is +1 for transcript on the positive strand and -1 otherwise.

Using splice2neo, we further annotated splice junctions with the resulting modified transcript and peptide sequences. Furthermore, splice junctions were annotated whether they are canonical using a database of canonical splice junctions built from GENCODE v34lift37 using the R-package GenomicFeatures [63] (v.1.46.5) and classified as "normal" if they were contained in a dataset of normal splice junctions previously identified by Jaganathan et al. with LeafCutter in 1,740 RNA-seq samples of 53 healthy tissues from GTEx [33].

The transcript context sequence covers by default 200-bp exonic sequence up- and downstream of a splice junction from an A3SS, A5SS or ES event ("cts_seq"). In case of intron retention events, the sequences covers the complete intronic sequence, flanked by 200-bp (default) exonic sequence up-and downstream of the intron retention event. The position of the splice junction in the sequence or the intron interval are given by "cts_junc_pos". A unique id ("cts_id") is given based on the context sequence and position as a XXH128 hash value.

Given the splice junction and coordinates of coding sequences (CDS) of reference transcripts, splice2neo annotates the resulting protein sequence for each affected transcript and the relevant peptide context sequence. In cases of in-frame events, this peptide context sequence covers junctions with potentially novel amino acid sequences (residue at the junction position or inserted novel residues) flanked with 13 wild-type amino acids up- and downstream. Frameshift peptides are flanked by an upstream wild-type region and are translated until the next stop codon. If splice junctions do not generate mutated gene products, no peptide context sequence is provided. This can be the case if junction does not affect the CDS, or the "mutated" gene product is only a truncated version of the wild-type CDS.

We used splice2neo v0.6.2 in this study.

Re-quantification of RNA-seq reads for mutation-retrieved splice junctions

To quantify the number of supporting RNA-seq reads for given splice junctions in a sensitive and targeted manner, we developed and applied EasyQuant (v0.4.0) https://github.com/TRON-Bioinformatics/easyquant. Analogous to the re-quantification for

gene fusions that we previously developed in EasyFuse [7], EasyQuant aligns reads to a context sequence ("cts_seq") constructed from the splice junction and calculates the reads supporting the junction (Supplemental Fig. 1). RNA-seq reads are mapped to the context sequence with STAR [61] with the following parameters: --outFilterMismatchNoverReadLmax 0.015 --alignEndsType EndToEnd --outFilterMultimapNmax -1 --outSAMattributes NH HI AS nM NM MD –scoreDelOpen -1000 --scoreInsOpen -1000 --scoreDelBase -2 –scoreInsBase -2. Next, the context sequence is divided into intervals based on the provided positions. For each interval, the number of reads within the interval is counted as well as reads overlapping interval end positions. Junctions from alternative splice sites or exon-skipping events are divided into two intervals. The end of the first interval represents the splice junction of interest. Reads covering the junction position by at least 10bp are considered to support the splice junction ("junction reads"). Context sequences for intron retentions are divided into three intervals, whereby the second interval is the retained intron. We summarized the coverage of the entire intron by the median number of reads per position in the interval. The output values from EasyQuant are defined in Table 2.

Table 2: Read quantification in EasyQuant per interval and input sequence.

Value	Definition
overlap_interval_end_reads	Number of reads that overlap interval end position by at least 10bp
span_interval_end_pairs	Number of read pair that span the interval end. Each read of the pair must be map on a different interval.
within_interval	Number of read that maps completely within the interval.
coverage_perc	Percent of interval positions that are covered by at least one read read.
coverage_mean	Mean number of mapped reads per interval position.
coverage_median	Median number of mapped reads per interval position.

If RNA-seq replicates were available, read counts and summary values were summed up per splice junction. To not overestimate intron retention read support, we considered here only potential intron retention events for which the retained intron region did not overlap with any exon of any other transcript.

False discovery rate estimation for prediction of target splice junctions

We aimed to derive a detection rule to predict splice junctions derived from somatic mutations as tumor-specific targets in a discovery cohort of 85 melanoma patients (VanAllen2015 [36], Riaz2017 [37]). The derived detection rule was confirmed in a verification cohort of 27 melanoma patients (Hugo2016 [39]).

To identify potential false positive splice junctions, we compared the RNA-seq support of novel mutation-retrieved splice junctions from the actual sample with the RNA-seq support from a different individual that does not has the corresponding mutation. The permutation analysis was performed among all samples per study (e.g. within VanAllen2015 and within Riaz2017 separately). Predicted splice junctions from somatic mutations were re-quantified with EasyQuant in the RNA-seq of all other independent tumor samples. Second, mutation-retrieved splice junctions were searched among the RNA-seq derived junctions identified in other biological independent samples.

A mutation-retrieved splice junction with RNA-seq support was labelled as false positive if there was support by re-quantification or SplAdder/LeafCutter derived splice junctions in at least one other independent sample. The estimated FDR was defined as the number of such false positives divided by all considered candidate splice junctions in the discovery or verification cohort.

To identify an appropriate detection rule, candidate splice junctions were gradually filtered by a grid of thresholds on the mutation effect scores from MMSplice and SpliceAI [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], and the re-quantification read support [1, 2, 3, 4]. The number of remaining splice junctions and the FDR was estimated for each threshold combination. The threshold combination with the lowest FDR in the discovery cohort was defined as the detection rule and remaining candidate splice junctions were defined as targets. This detection rule was applied to the verification cohort to estimate the resulting number of target splice junctions and the final estimated FDR.

Annotation of neoantigen candidates

MHC -I and –II genotypes were determined with HLA-HD [65] (v1.2.0.1) for each patient based on the WES-seq data of the normal sample. NeoFox [40] (v1.0.2) was used to annotate neoantigen candidates derived by target splice junctions with neoantigen features. Here, the predicted mutated peptide sequences without information about variant allele frequency or transcript expression was provided as input to NeoFox. Furthermore, MHC-I and -II genotypes of the patients were provided to the tool as input.

Confirmation of splice junction expression with qRT-PCR

Primers were designed with a custom R script as described previously [7]. In short, context sequences as predicted by splice2neo were used for primer design with Primer-BLAST [66]. Primer design was guided by the following restrictions: (1) no primer should align within 20 bp upstream and downstream of the junction; (2) amplicons must span the junction; and (3) amplicon size should not exceed 150 bp. Primer pairs aligning to off-target loci were removed.

Oligonucleotides were ordered at eurofins genomics as salt free, 0.01µM custom DNA oligos in plates. Stocks were diluted 1:10 with PCR-grade water prior to using them in PCR.

Splice junctions were validated in previously extracted total RNA from fresh frozen tissue samples [7]. The RNA was reverse transcribed into cDNA using TAKARA PrimerScript RT Reagent Kit with gDNA Eraser according to the manufacturer's manual. Each reaction contained 1 µg total RNA in a final volume of 60 µL. Reverse Transcriptase was replaced with water in no amplification controls.

qRT-PCR was performed on a BioRad CFX384 instrument using BioRad SsoAdvanced Universal SYBR Green Supermix. Reactions had an overall volume of 15 µL containing 7.5 µL BioRad SsoAdvanced Universal SYBR Green Supermix, 1 µL forward and 1µL reverse primer at 10µM concentration, 5.25 mL PCR-grade water and 1.25 µL cDNA. The thermal cycle protocol started with polymerase activation for 30 s at 98°C followed by 40 cycles of two-step PCR with denaturation for 10 s at 98°C and annealing/elongation for 30 s at 60°C. Finally, melting curve analysis was performed using the instruments standard protocol.

The PCR products were analyzed for the presence of amplicons matching the expected sizes using QIAGEN's QIAxcel Advanced capillary gel electrophoresis instrument with a 15 bp/600 bp alignment marker.

A custom R script was used to evaluate qRT-PCR and capillary gel electrophoresis results and to assign each tested amplicon as positive or negative given a defined set of rules. (1) The amplicon had to be detected in qRT-PCR with a Ct value below or equal to 35 and at least 5 cycles earlier compared to water and no amplification controls. (2) The amplicon product size had to match the expected amplicon size (allowing a maximum difference of 15% in amplicon size). (3) The amplicon associated peak had to cover at minimum 15% height relative to the total peak heights. Amplicons fulfilling the first two rules but with a relative height below 15% were manually inspected and classified manually.

Read support visualization with sashimi plots

A Sashimi plot (Additional File: Fig. S5D) was generated with the python package rmats2sashimiplot (v 2.0.4) (https://github.com/Xinglab/rmats2sashimiplot) using RNA-seq BAM files with coordinate and GENCODE (v34lift37) annotation. The original Sashimi plot was manually modified to include one further affected transcript while a non-affected transcript was removed from the plot.

Declarations

Ethics approval and consent to participate

The patient material of the FFPE cohort was collected as part of the RB_T002 research program (DRKS-ID: DRKS00011790). The studies were carried out in accordance with the Declaration of Helsinki and good clinical practice guidelines and with approval by the institutional review board or independent ethics committee of each participating site and the competent regulatory authorities.

Consent for publication

All patients provided written informed consent.

Availability of data and materials

RNA-seq data for the FFPE cohort [7] is available in EGA under accession number EGAS00001004877. Data for the Hugo cohort [39] is available in SRA (Sequence Read Archive) under accession numbers SRP067938, SRP090294 (WES-seq) and SRP070710 (RNA-seq). Data for the Riaz cohort [37] is available under accession numbers SRP095809 (WES-seq) and SRP094781 (RNA-seq). Data for the Van Allen cohort [36] is available in dbGap under accession number phs000452.v2.p1. RNA-seq for the cohort of healthy tissue is available under is in SRA under NCBI BioProject ID PRJNA764684. We acknowledge the authors and generators of these datasets and the grants that supported the studies.

The source code and documentation of splice2neo (https://github.com/TRON-Bioinformatics/splice2neo) and EasyQuant (https://github.com/TRON-Bioinformatics/easyquant) are available under open source licenses on GitHub.

Competing interests

U.S. is co-founder, chief executive officer and stock owner of BioNTech SE. The remaining authors declare no competing interests.

Funding

This work was supported by an ERC Advanced Grant to U.S. (ERC-AdG 789256).

Authors' contributions

Conceptualization: JI. Methodology: FL, JI. Software: FL, PS, PRF, JI. Data curation: FL, PS. Formal analysis: FL, JI, Investigation: FL, PS, MS, AH, CA, NK, AB, PRF, CH, DW, JI. Supervision: BS, DW, ML, US, JI. Visualization: FL, JI. Writing—original draft: FL, JI. Writing—reviewing and editing: FL, BS, DW, ML, US, JI. All authors read and approved the final manuscript.

Acknowledgements

We thank the RB_T002 research program (DRKS-ID: DRKS00011790) patients, from whom analyzed samples were obtained, and we thank the involved study site teams for their support and collaboration. We thank O. Akilli-Oeztuerk for support with biosampling. This work was supported by an ERC Advanced Grant to U.S. (ERC-AdG 789256). The authors further acknowledge the authors and generators of datasets used in this work and the grants that supported the studies.

Figure Legends

Fig. 1: splice2neo is a versatile tool for identifying and analyzing of splice junctions as a source of neoantigen candidates. (A) Overview of functionalities implemented in the splice2neo R package. Splice2neo formats the output of several splicing tools into a unified junction format “chr:start-end:strand“, can exclude canonical or normal splice junctions (e.g. from GENCODE or GTEx) and annotates altered transcript and peptide sequences. **(B)** Workflow to detect candidate splice junctions. The effect of somatic mutations on splicing was predicted with SpliceAI and MMSplice and expressed splicing events were detected with LeafCutter and SplAdder for a given tumor sample, followed by formatting with splice2eno into the unified splice junction format. Novel mutation and RNA-seq retrieved splice junctions were intersected to identify mutation-retrieved splice junctions with RNA-seq support as candidates. To expand the number of candidates, novel mutation-derived were re-quantified in tumor RNA-seq with EasyQuant in a targeted manner.

Fig. 2: Identification of mutation-retrieved splice junctions supported by RNA-seq in melanoma samples. (A) The number of somatic mutations, mutation-retrieved splice junctions and RNA-seq splice derived junctions per sample in the discovery cohort of 85 melanoma samples. **(B)** Overlap of mutation-retrieved splice junctions with those that were found in RNA-seq by SplAdder or LeafCutter and those that were re-quantified in RNA-seq with EasyQuant. Mutation-retrieved splice junctions with RNA-seq support (i.e. by SplAdder, LeafCutter or EasyQuant) were defined as candidate splice junctions. **(C-D)** The distribution of **(C)** MMSplice and **(D)** SpliceAI scores of candidate splice junctions with RNA-seq support and mutation-retrieved splice junctions without RNA-seq support.

Fig. 3: Prediction of tumor-specific target splice junctions based on mutation effect scores and RNA-seq support. (A) To calculate the amount of non-specific junctions and estimate a FDR, candidate junctions were compared to splice junctions identified by SplAdder/LeafCutter in other samples’ RNA-seq and their supporting reads were quantified in RNA-seq data of other samples with EasyQuant. **(B)** Candidate splice junctions from the

discovery set were gradually filtered by thresholds on the mutation effect scores from MMSplice and SpliceAI, and the re-quantification read support and the resulting estimated FDR and number of target splice junctions was determined. The set of filtering thresholds with lowest estimated FDR was selected as an optimal detection rule for target splice junctions and applied to verification cohort of 27 melanoma samples. **(C)** Target splice junctions were predicted from candidate junctions by the following detection rule: (1) restriction to ES and ASS events, (2) SpliceAI score ≥ 0.35 or MMSplice score ≤ -0.35, and (3) identification by LeafCutter/SplAdder or re-quantification with at least 3 junction reads by EasyQuant. **(D)** Overlap of the target and non-target splice junctions with splice junctions identified by SplAdder/LeafCutter in an additional RNA-seq dataset of 141 samples from 49 healthy tissues in the discovery and verification dataset. **(E)** The fraction of splice junctions derived from RNA-seq found in additional RNA-seq dataset of 141 samples from 49 healthy tissues in the discovery and verification dataset.

Fig. 4: Target splice junctions generate neoantigen candidates. (A) The number of neoantigen candidates per sample in the verification dataset. Inlet: Correlation of the neoantigen candidate count with the tumor mutation burden. **(B)** The fraction of neoantigen candidates from A3SS, A5SS and ES events. **(C)** The fraction of frameshift and in-frame neoantigen candidates. **(D)** The length distribution of frameshift and in-frame neoantigen candidates. **(E)** The fraction of strong or weak binding MHC-I and MHC-II neoantigen candidates.(strong: MHC-I binding rank < 0.5, MHC-II binding rank < 2, weak: 0.5 ≤ MHC-I binding rank < 2, 2 ≤ MHC-II binding rank < 10, no: MHC-I binding rank ≥ 2, MHC-II binding rank ≥ 10) **(F-G) (F)** MHC-I and **(G)** MHC-II binding rank of the best predicted neoepitope per neoantigen candidate for those from alternative splicing (AS) and SNVs. SNV-derived neoantigen candidates were retrieved as described in [67] **(H)** Self-similarity [41] of the best predicted neoepitope per neoantigen candidate for those from alternative splicing and those from SNVs.

References

1. Lang F, Schrörs B, Löwer M, Türeci Ö, Sahin U. Identification of neoantigens for individualized therapeutic cancer vaccines. Nat Rev Drug Discov. 2022;21:261–82. doi:10.1016/j.immuni.2013.07.012.
2. Sahin U, Derhovanessian E, Miller M, Kloke B-P, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017;547:222-226. doi:10.1038/nature23003.

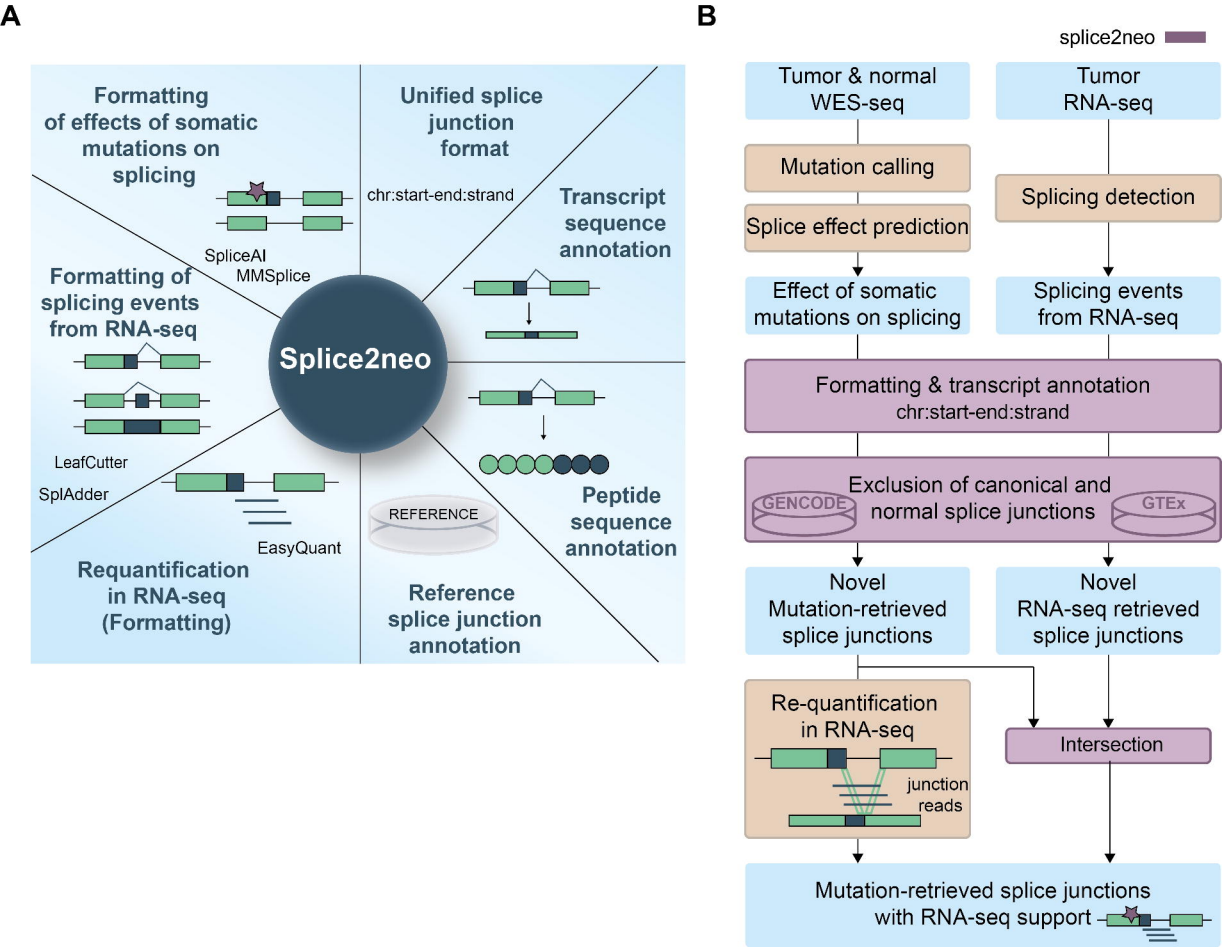
3. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017;547:217–221. doi:10.1038/nature22991.
4. Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature. 2019;565:234–9. doi:10.1038/s41586-018-0792-9.
5. Hilf N, Kuttruff-Coqui S, Frenzel K, Bukur V, Stevanović S, Gouttefangeas C, et al. Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature. 2019;565:240–5. doi:10.1038/s41586-018-0810-y.
6. Smith CC, Selitsky SR, Chai S, Armistead PM, Vincent BG, Serody JS. Alternative tumour-specific antigens. Nature Reviews Cancer. 2019;19:465–78. doi:10.1038/s41568-019-0162-4.
7. Weber D, Ibn-Salem J, Sorn P, Suchan M, Holtsträter C, Lahrmann U, et al. Accurate detection of tumor-specific gene fusions reveals strongly immunogenic personal neo-antigens. Nat. Biotechnol. 2022. doi:10.1038/s41587-022-01247-9.
8. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: A pan-cancer analysis. Lancet Oncol. 2017;18:1009–21. doi:10.1016/S1470-2045(17)30516-8.
9. Yang W, Lee K-W, Srivastava RM, Kuo F, Krishna C, Chowell D, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nat. Med. 2019;25:767–75. doi:10.1038/s41591-019-0434-2.
10. Rogalska ME, Vivori C, Valcárcel J. Regulation of pre-mRNA splicing: Roles in physiology and disease, and therapeutic prospects. Nat Rev Genet. 2023;24:251–69. doi:10.1038/s41588-019-0555-z.
11. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–6. doi:10.1038/nature07509.
12. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348:660–5. doi:10.1126/science.aaa0355.
13. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature. 2022;608:353–9. doi:10.1038/s41586-022-05035-y.
14. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The Functional Impact of Alternative Splicing in Cancer. Cell Reports. 2017;20:2215–26. doi:10.1016/j.celrep.2017.08.012.
15. Escobar-Hoyos L, Knorr K, Abdel-Wahab O. Aberrant RNA Splicing in Cancer. Annu Rev Cancer Biol. 2019;3:167–85. doi:10.1146/annurev-cancerbio-030617-050407.
16. Öther-Gee Pohl S, Myant KB. Alternative RNA splicing in tumour heterogeneity, plasticity and therapy. Dis Model Mech 2022. doi:10.1242/dmm.049233.
17. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell. 2018;34:211-224.e6. doi:10.7125/APAN.35.1.

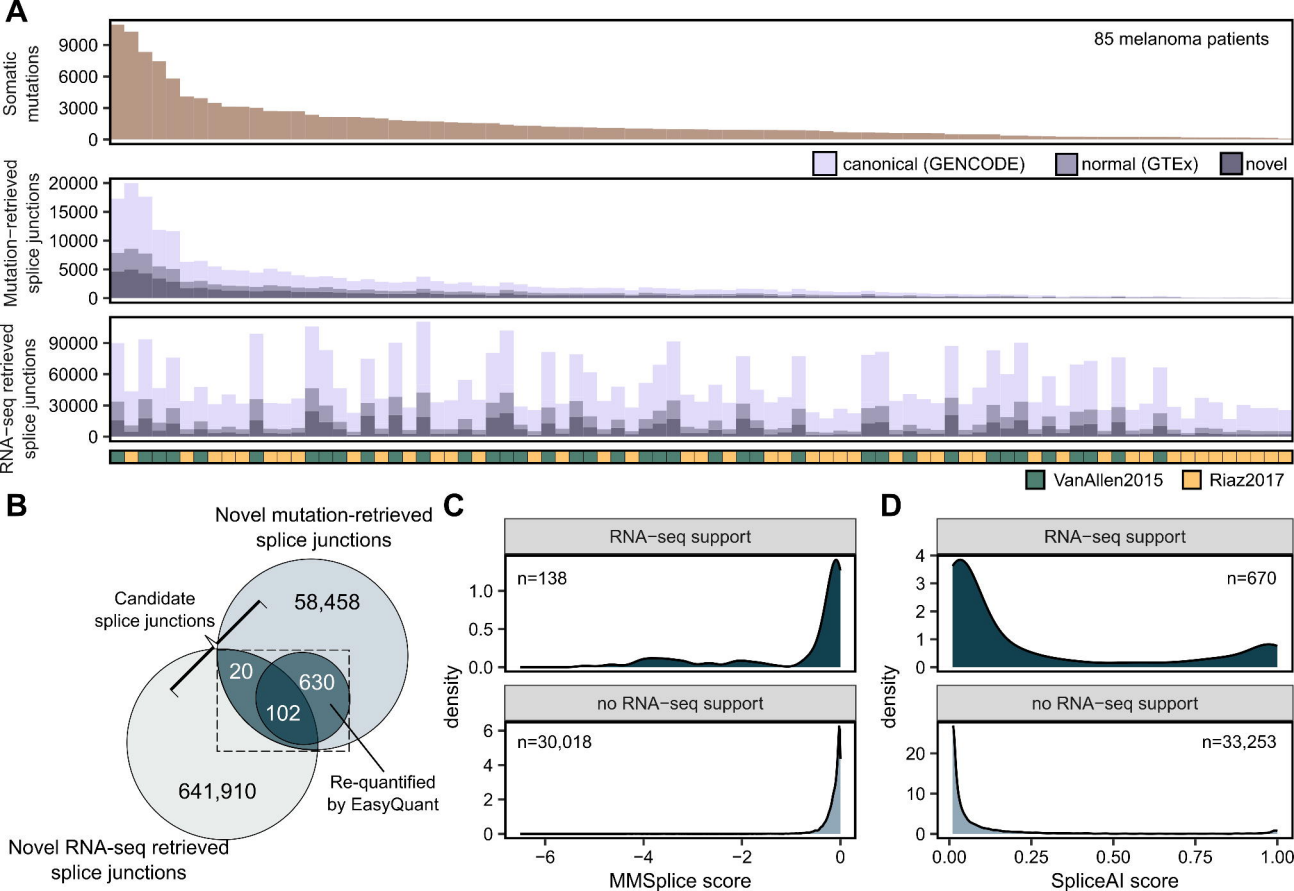
18. Oka M, Xu L, Suzuki T, Yoshikawa T, Sakamoto H, Uemura H, et al. Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer. Genome Biol. 2021;22:9. doi:10.1186/s13059-020-02240-8.
19. Bigot J, Lalanne AI, Lucibello F, Gueguen P, Houy A, Dayot S, et al. Splicing patterns in SF3B1 mutated uveal melanoma generate shared immunogenic tumor-specific neo-epitopes. Cancer Discov. 2021:1938–51. doi:10.1158/2159-8290.CD-20-0555.
20. Ehx G, Larouche J-D, Durette C, Laverdure J-P, Hesnard L, Vincent K, et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Immunity. 2021;54:737-752.e10. doi:10.1016/j.immuni.2021.03.001.
21. Merlotti A, Sadacca B, Arribas YA, Ngoma M, Burbage M, Goudot C, et al. Noncanonical splicing junctions between exons and transposable elements represent a source of immunogenic recurrent neo-antigens in patients with lung cancer. Sci Immunol. 2023;8:eabm6359. doi:10.1126/sciimmunol.abm6359.
22. Chai S, Smith CC, Kochar TK, Hunsucker SA, Beck W, Olsen KS, et al. NeoSplice: A bioinformatics method for prediction of splice variant neoantigens. Bioinform Adv. 2022;2:vbac032. doi:10.1093/bioadv/vbac032.
23. Zhang Z, Zhou C, Tang L, Gong Y, Wei Z, Zhang G, et al. ASNEO: Identification of personalized alternative splicing based neoantigens with RNA-seq. Aging (Albany NY). 2020;12:14633–48. doi:10.18632/aging.103516.
24. Pan Y, Phillips JW, Zhang BD, Noguchi M, Kutschera E, McLaughlin J, et al. IRIS: Discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing. Proc Natl Acad Sci U S A. 2023;120:e2221116120. doi:10.1073/pnas.2221116120.
25. David JK, Maden SK, Weeder BR, Thompson RF, Nellore A. Putatively cancer-specific exon-exon junctions are shared across patients and present in developmental and other non-cancer cells. NAR Cancer. 2020;2:zcaa001. doi:10.1093/narcan/zcaa001.
26. Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, et al. Systematic Analysis of Splice-Site-Creating Mutations in Cancer. Cell Reports. 2018;23:270-281.e3. doi:10.1016/j.celrep.2018.03.052.
27. PCAWG Transcriptome Core Group., Calabrese, Claudia, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578:129–36. doi:10.1038/s41586-020-1970-0.
28. Cotto KC, Feng Y-Y, Ramu A, Richters M, Freshour SL, Skidmore ZL, et al. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. Nat Commun. 2023;14:13. doi:10.1038/s41467-023-37266-6.
29. Jung H, Lee D, Lee J, Park D, Kim YJ, Park W-Y, et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. 2015;47:1242–8. doi:10.1038/ng.3414.
30. Riepe TV, Khan M, Roosing S, Cremers FPM, Hoen PAC 't. Benchmarking deep learning splice prediction tools using functional splice assays. Human Mutation. 2021;42:799–810. doi:10.1002/humu.24212.

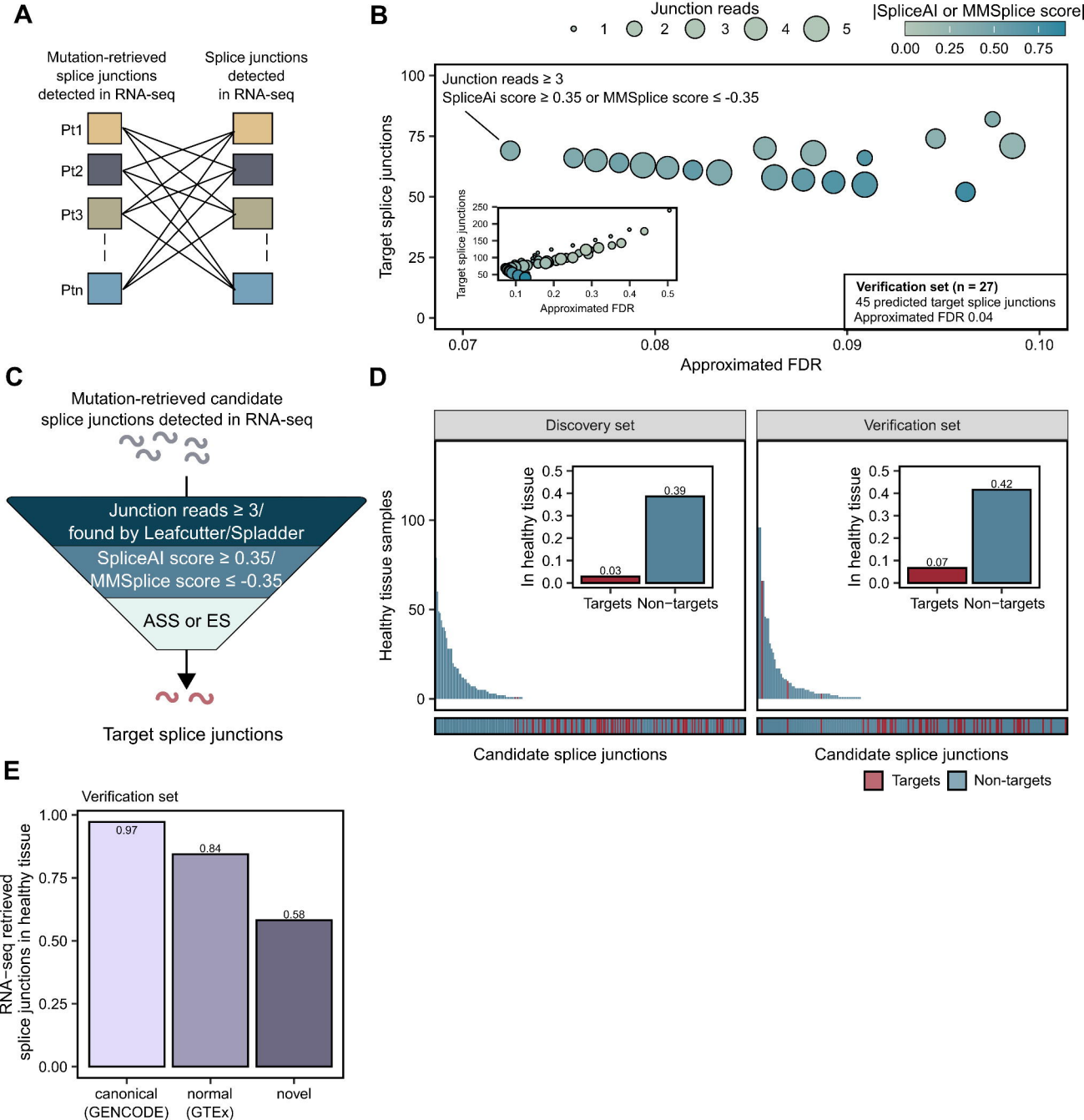
31. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet Med. 2020;22:1005–14. doi:10.1038/s41436-020-0766-9.
32. Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Avsec Ž, Gagneur J. MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20:48. doi:10.1186/s13059-019-1653-z.
33. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176:535-548.e24. doi:10.1016/j.cell.2018.12.015.
34. Kahles A, Ong CS, Zhong Y, Rätsch G. SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics. 2016;32:1840–7. doi:10.1093/bioinformatics/btw076.
35. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK. Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. 2018;50:151–8. doi:10.1038/s41588-017-0004-9.
36. van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015;350:207–11. doi:10.1126/science.aad0095.
37. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell. 2017;171:934-949.e16. doi:10.1016/j.cell.2017.09.028.
38. The GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348:648–60. doi:10.1126/science.1262110.
39. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2016;165:35–44. doi:10.1016/j.cell.2016.02.065.
40. Lang F, Ferreiro PR, Löwer M, Sahin U, Schrörs B. NeoFox: Annotating neoantigen candidates with neoantigen features. Bioinformatics 2021. doi:10.1093/bioinformatics/btab344.
41. Bjerregaard A-M, Nielsen M, Jurtz V, Barra CM, Hadrup SR, Szallasi Z, Eklund AC. An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. Front. Immunol. 2017;8:1566. doi:10.3389/fimmu.2017.01566.
42. Frankiw L, Baltimore D, Li G. Alternative mRNA splicing in cancer immunotherapy. Nat Rev Immunol. 2019;19:675–87. doi:10.1038/s41577-019-0195-7.
43. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18:437–51. doi:10.1038/nrm.2017.27.
44. Fenn A, Tsoy O, Faro T, Rößler FLM, Dietrich A, Kersting J, et al. Alternative splicing analysis benchmark with DICAST. NAR Genom Bioinform. 2023;5:lqad044. doi:10.1093/nargab/lqad044.
45. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80. doi:10.1182/blood-2003-09-3243.

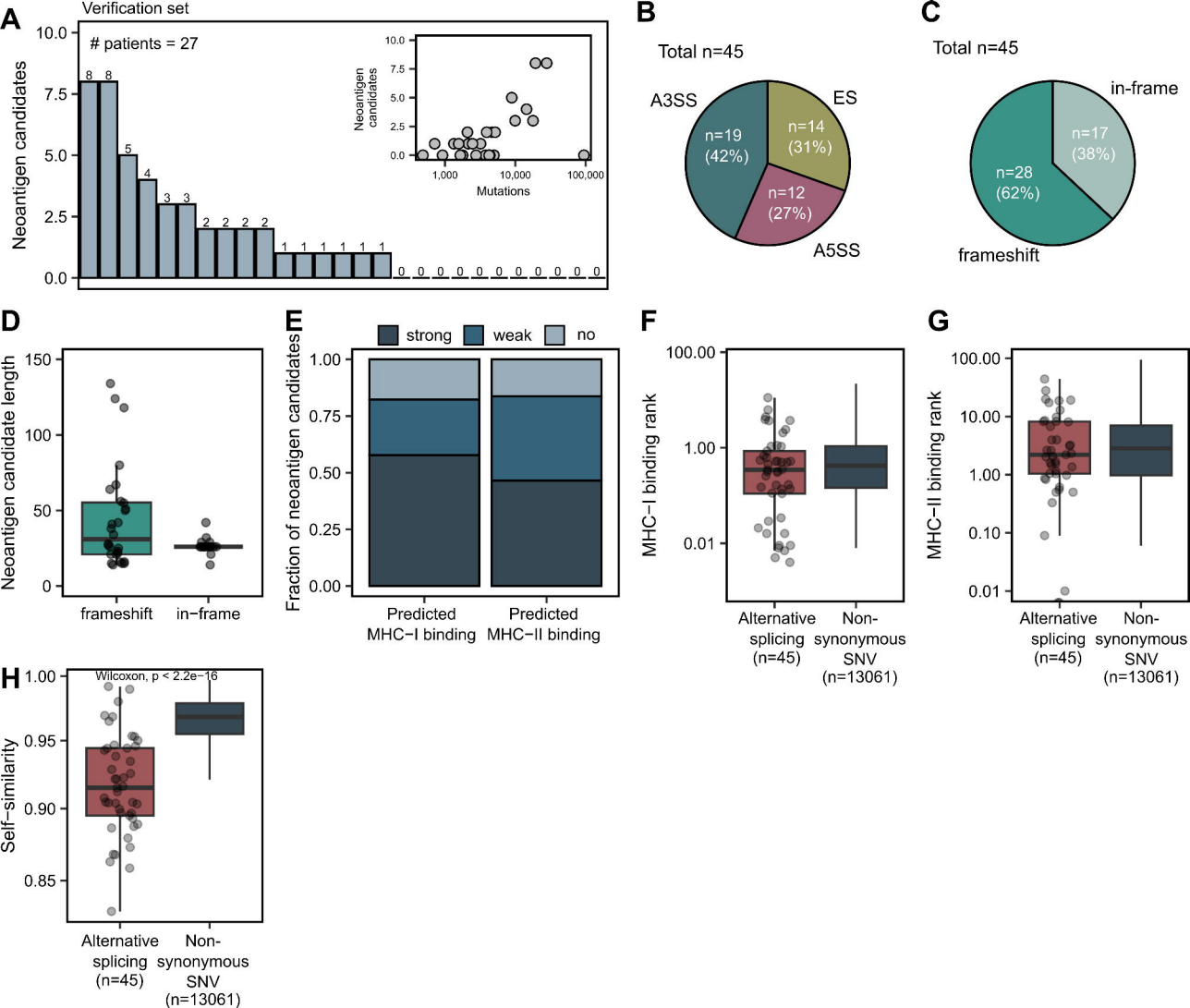
46. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21. doi:10.1038/nmeth.3252.
47. Burbage M, Rocañín-Arjó A, Baudon B, Arribas YA, Merlotti A, Rookhuizen DC, et al. Epigenetically controlled tumor antigens derived from splice junctions between exons and transposable elements. Sci Immunol. 2023;8:eabm6360. doi:10.1126/sciimmunol.abm6360.
48. Broseus L, Ritchie W. Challenges in detecting and quantifying intron retention from next generation sequencing data. Comput Struct Biotechnol J. 2020;18:501–8. doi:10.1016/j.csbj.2020.02.010.
49. Middleton R, Gao D, Thomas A, Singh B, Au A, Wong JJ-L, et al. IRFinder: Assessing the impact of intron retention on mammalian gene expression. Genome Biol. 2017;18:51. doi:10.1186/s13059-017-1184-4.
50. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. Nat. Biotechnol. 2018;36:1056–8. doi:10.1038/nbt.4239.
51. Dong C, Reiter JL, Dong E, Wang Y, Lee KP, Lu X, Liu Y. Intron-Retention Neoantigen Load Predicts Favorable Prognosis in Pancreatic Cancer. JCO Clin Cancer Inform. 2022;6:e2100124. doi:10.1200/CCI.21.00124.
52. Lorenzi C, Barriere S, Arnold K, Luco RF, Oldfield AJ, Ritchie W. IRFinder-S: A comprehensive suite to discover and explore intron retention. Genome Biol. 2021;22:307. doi:10.1186/s13059-021-02515-8.
53. Riesgo-Ferreiro P, Siek R, Muslu O, Mohr J. TRON-Bioinformatics/tronflow-alignment Release v1.5.0 (v1.5.0). Zenodo:10.5281/zenodo.6424860. doi:10.5281/zenodo.6424860.
54. Riesgo-Ferreiro P, Bresadola L. TRON-Bioinformatics/tronflow-bam-preprocessing: Release v1.7.3(v1.7.3). Zenodo:https://doi.org/10.5281/zenodo.6421395. doi:10.5281/zenodo.6421395.
55. Riesgo-Ferreiro P, Ritzel C. TRON-Bioinformatics/tronflow-mutect2: Release 1.2.0 (v1.2.0). Zenodo:https://doi.org/10.5281/zenodo.6249868. doi:10.5281/zenodo.6249868.
56. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 2017;35:316–9. doi:10.1038/nbt.3820.
57. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
58. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011;43:491–8. doi:10.1038/ng.806.
59. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019:Preprint at https://www.biorxiv.org/content/10.1101/861054v1. doi:10.1101/861054.
60. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43. doi:10.1038/s41586-020-2308-7.
61. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29:15–21. doi:10.1093/bioinformatics/bts635.

62. Signal B, Kahlke T. how_are_we_stranded_here: Quick determination of RNA-Seq strandedness. BMC Bioinformatics. 2022;23:49. doi:10.1186/s12859-022-04572-7.
63. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput. Biol. 2013;9:e1003118. doi:10.1371/journal.pcbi.1003118.
64. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.62.0, https://bioconductor.org/packages/Biostrings.
65. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Human Mutation. 2017;38:788–97. doi:10.1002/humu.23230.
66. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134. doi:10.1186/1471-2105-13-134.
67. Lang F, Sorn P, Schrörs B, Weber D, Kramer S, Sahin U, Löwer M. Multiple instance learning to predict immune checkpoint blockade efficacy using neoantigen candidates. bioRxiv 2022. doi:10.1101/2022.05.06.490587.


A
Formatting of effects of somatic mutations on splicing
SpliceAI
MMSplice
Unified splice junction format
chr:start-end:strand
Transcript sequence annotation
Formatting of splicing events from RNA-seq
LeafCutter
SplAdder
Splice2neo
Peptide sequence annotation
REFERENCE
EasyQuant
Requantification in RNA-seq (Formatting)
Reference splice junction annotation
B
splice2neo
Tumor & normal WES-seq
Tumor RNA-seq
Mutation calling
Splicing detection
Splice effect prediction
Effect of somatic mutations on splicing
Splicing events from RNA-seq
Formatting & transcript annotation
chr:start-end:strand
Exclusion of canonical and normal splice junctions
GENCODE
GTEx
Novel Mutation-retrieved splice junctions
Novel RNA-seq retrieved splice junctions
Re-quantification in RNA-seq
junction reads
Intersection
Mutation-retrieved splice junctions with RNA-seq support


A
85 melanoma patients
Somatic mutations
9000
6000
3000
0
canonical (GENCODE)
normal (GTEx)
novel
Mutation−retrieved splice junctions
20000
15000
10000
5000
0
RNA−seq retrieved splice junctions
90000
60000
30000
0
VanAllen2015
Riaz2017
B
Novel mutation-retrieved splice junctions
Candidate splice junctions
58,458
20
630
102
641,910
Re-quantified by EasyQuant
Novel RNA-seq retrieved splice junctions
C
RNA−seq support
n=138
no RNA−seq support
n=30,018
density
MMSplice score
D
RNA−seq support
n=670
no RNA−seq support
n=33,253
density
SpliceAI score


A
Mutation-retrieved splice junctions detected in RNA-seq
Splice junctions detected in RNA-seq
Pt1
Pt2
Pt3
Ptn
B
Junction reads
1
2
3
4
5
|SpliceAI or MMSplice score|
0.00
0.25
0.50
0.75
Junction reads ≥ 3
SpliceAi score ≥ 0.35 or MMSplice score ≤ -0.35
Target splice junctions
Approximated FDR
Verification set (n = 27)
45 predicted target splice junctions
Approximated FDR 0.04
C
Mutation-retrieved candidate splice junctions detected in RNA-seq
Junction reads ≥ 3/
found by Leafcutter/Spladder
SpliceAI score ≥ 0.35/
MMSplice score ≤ -0.35
ASS or ES
Target splice junctions
D
Discovery set
Verification set
Healthy tissue samples
In healthy tissue
0.03
0.39
0.07
0.42
Targets
Non−targets
Candidate splice junctions
Targets
Non-targets
E
Verification set
RNA−seq retrieved splice junctions in healthy tissue
0.97
0.84
0.58
canonical (GENCODE)
normal (GTEx)
novel


A
Verification set
patients = 27
Neoantigen candidates
Mutations
B
Total n=45
A3SS
ES
A5SS
n=19 (42%)
n=14 (31%)
n=12 (27%)
C
Total n=45
in-frame
frameshift
n=17 (38%)
n=28 (62%)
D
Neoantigen candidate length
frameshift
in-frame
E
strong
weak
no
Fraction of neoantigen candidates
Predicted MHC−I binding
Predicted MHC−II binding
F
MHC−I binding rank
Alternative splicing (n=45)
Non-synonymous SNV (n=13061)
G
MHC−II binding rank
Alternative splicing (n=45)
Non-synonymous SNV (n=13061)
H
Wilcoxon, p < 2.2e−16
Self−similarity
Alternative splicing (n=45)
Non-synonymous SNV (n=13061)