

RESEARCH ARTICLE

Open Access

The landscape of coding RNA editing events in pediatric cancer



Ji Wen^{1,2†}, Michael Rusch^{2†}, Samuel W. Brady^{2†}, Ying Shao², Michael N. Edmonson², Timothy I. Shaw², Brent B. Powers², Liqing Tian², John Easton², Charles G. Mullighan¹, Tanja Gruber³, David Ellison¹ and Jinghui Zhang^{2*} 

Abstract

Background: RNA editing leads to post-transcriptional variation in protein sequences and has important biological implications. We sought to elucidate the landscape of RNA editing events across pediatric cancers.

Methods: Using RNA-Seq data mapped by a pipeline designed to minimize mapping ambiguity, we investigated RNA editing in 711 pediatric cancers from the St. Jude/Washington University Pediatric Cancer Genome Project focusing on coding variants which can potentially increase protein sequence diversity. We combined de novo detection using paired tumor DNA-RNA data with analysis of known RNA editing sites.

Results: We identified 722 unique RNA editing sites in coding regions across pediatric cancers, 70% of which were nonsynonymous recoding variants. Nearly all editing sites represented the canonical A-to-I ($n = 706$) or C-to-U sites ($n = 14$). RNA editing was enriched in brain tumors compared to other cancers, including editing of glutamate receptors and ion channels involved in neurotransmitter signaling. RNA editing profiles of each pediatric cancer subtype resembled those of the corresponding normal tissue profiled by the Genotype-Tissue Expression (GTEx) project.

Conclusions: In this first comprehensive analysis of RNA editing events in pediatric cancer, we found that the RNA editing profile of each cancer subtype is similar to its normal tissue of origin. Tumor-specific RNA editing events were not identified indicating that successful immunotherapeutic targeting of RNA-edited peptides in pediatric cancer should rely on increased antigen presentation on tumor cells compared to normal but not on tumor-specific RNA editing per se.

Keywords: RNA editing, Pediatric cancer, Genomics, Immunotherapy

Background

Post-transcriptional modification of RNA sequences, termed RNA editing, occurs in many species [1]. In humans, two canonical editing types have been identified: adenosine to inosine (A-to-I) editing mediated by the adenosine deaminase acting on RNA (ADAR)

enzyme family [2] and cytosine to uracil (C-to-U) editing induced by apolipoprotein B mRNA editing (APOBEC) enzymes [3]. While most editing occurs in Alu repeats [4, 5], RNA editing can also affect protein coding regions [5, 6]. RNA-Seq analysis can identify these editing events through comparison with DNA sequencing, including whole genome (WGS) or whole exome (WES) sequencing, by identifying variants present in RNA but not in DNA [7]. Multiple studies have used RNA-Seq data from The Cancer Genome Atlas (TCGA) to analyze RNA editing events in adult solid tumors and their effects on

* Correspondence: Jinghui.Zhang@StJude.org

[†]Ji Wen, Michael Rusch and Samuel W. Brady contributed equally to this work.

²Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

cancer viability, invasiveness, drug sensitivity, and patient survival [8, 9]. However, RNA editing has not been investigated in pediatric malignancies.

Despite clearly established mechanisms for canonical RNA editing, there is a lack of consensus regarding its prevalence. Some consider RNA editing to be a rare event across the transcriptome [10–13], while others consider it widespread [14, 15] perhaps confounded by mistaken inclusion of technical artifacts [11–13]. For example, one study comparing RNA and DNA sequencing of B cell lines, primary skin fibroblasts and cerebral cortex reported abundant exonic RNA editing, including many noncanonical events [14]. However, reanalysis of the same data showed that RNA editing was less frequent [10, 16], and most of the previously reported [14] editing sites were likely the result of faithful transcription of pseudogenes that share high homology with the canonical genes. There is also disagreement regarding the prevalence of non-canonical (non-A-to-I, non-C-to-U) RNA editing [10, 17–19].

These controversies highlight the importance of accurate RNA-Seq mapping to the human transcriptome [20], which can reduce false-positive RNA editing events and increase the sensitivity for detection of true events. RNA-Seq mapping algorithms were initially designed to ascertain gene expression levels but were not optimized for detecting RNA editing events. For example, when sequencing reads end near splice junctions or true RNA editing events, soft-clipped mappings may be produced, hampering detection of editing events. It is also difficult to ensure accurate mapping to paralogs or expressed pseudogenes. To reduce mapping artifacts in RNA-Seq and thus improve the detection of RNA editing, we have developed an alignment pipeline, StrongArm, which performs competitive mapping with multiple aligners to multiple reference databases to resolve ambiguity by applying knowledge-guided rules. These rules were designed to reduce the error rate and bias from a single aligner, especially near error-prone splice junctions and in paralogous regions. This competitive mapping approach, initially designed for detecting complex gene fusion events in ependymoma [21], has the potential for systematically removing false-positive RNA editing calls caused by a variety of sources of error.

We applied StrongArm along with a post-processing pipeline to identify novel and known single nucleotide variant (SNV) RNA editing events in protein-coding regions that show differences in matched RNA-Seq and DNA sequencing (WGS or WES) from 711 pediatric cancer samples from the St. Jude/Washington University Pediatric Cancer Genome Project (PCGP) [22]. As low-quantity RNA editing events may be difficult to detect *de novo*, we also analyzed known RNA editing sites reported in the RADAR [5] database, a well-curated RNA

editing resource, as done by others [8]. In all, we identified 722 RNA editing sites in coding regions across pediatric cancers, including 584 known and 138 novel editing sites. We observed an enrichment of RNA editing in pediatric brain tumors, including in genes involved in neurotransmitter signaling. We compared pediatric cancer RNA editing profiles to normal tissues from the GTEx project and found that the coding RNA editing profile of each pediatric cancer type largely resembles that of its corresponding normal tissue. This suggests that RNA editing is rarely tumor-specific in pediatric cancer but is largely related to the tissue of origin. Together, these results present a comprehensive analysis of RNA editing in pediatric cancer, yielding novel RNA editing events whose biological function should be investigated in future studies.

Methods

Sample collection

Details regarding RNA and DNA isolation from PCGP samples have been reported in a series of PCGP-related papers [22–31]. All samples have also been published previously and the raw data of the entire PCGP cohort can be accessed via the St. Jude Cloud Genomics Platform (<https://pecan.stjude.cloud/permalink/rnaediting>) [32].

RNA-Seq

PolyA-enriched mRNA-seq of PCGP samples was performed using the Illumina TruSeq V2 RNA library preparation kit, with a starting input of 1 µg of total RNA according to manufacturer's protocol. The number of cycles of library amplification was reduced to 10 cycles to reduce PCR duplicates. The resulting data files were converted to FASTQ files using CASAVA 1.8.2. All reads were 101 bp in length. For discovery of RNA editing events, we initially used 717 pediatric cancer samples' RNA-Seq from the PCGP, including only samples for which tumor-normal DNA-Seq (WGS or WES) was also available (Supplementary Fig. 1). The novel editing sites thus identified, along with known RNA editing events from RADAR, were then analyzed in 954 PCGP samples with RNA-Seq, whether or not DNA-Seq was available. Finally, these 954 samples were filtered to remove samples for which RNA was isolated outside of St. Jude Children's Research Hospital, due to batch effects in RNA variant allele fractions (VAFs), and relapsed samples and other samples were also filtered out such that only one diagnosis sample per patient was included in the final 711 samples (Supplementary Fig. 1).

Whole exome and whole genome sequencing

WES was performed using the Illumina TruSeq Exome Library Prep Kit with 1 µg of genomic DNA input using the manufacturer's protocol. The WGS was performed

for paired tumor and normal genomes to >30-fold coverage as described previously [33]. WES and WGS were performed on the Illumina HiSeq 2000 using a paired sequencing (2×100 cycles).

RNA editing validation

Validation of RNA editing was performed by deep amplicon sequencing on the MiSeq platform. Flanking Primers were designed using Batch Primer3 [34] with local batch automation and parameter modification. Optimal amplicon sizes ranged from approximately 120 bp–200 bp for use in downstream library construction. Total RNA was reverse transcript to cDNA using BioRad iScript cDNA Synthesis Kit. PCR was performed using AmpliTaq Gold 360 master mix (Applied BioSystems), 10 μ M of each primer, and 10 ng of genomic DNA and cDNA using the following parameters: 95 °C for 10 min, 95 °C for 30 s, 58 °C for 30 s, 72 °C for 30 s for 35 cycles, 72 °C for 5 min, and storage at 4 °C. All amplicons were quality-checked on a 2% agarose E-gel (Invitrogen). Amplicons were pooled and purified using Agencourt Ampure XP Beads. DNA libraries were created from pooled amplicons using the Nextflex DNA kit (Bio Scientific), following the manufacturer's instructions. Libraries were normalized for sequencing on Illumina MiSeq platform by using a 2×150 paired-end version 2 sequencing kit.

StrongArm mapping pipeline

StrongArm accepts reads stored in unaligned BAM format [35], which may be generated from FASTQ using the FastqToSam function of Picard (<http://picard.sourceforge.net>). We used several input BAMs for each sample, each with two million reads. The whole workflow of StrongArm is illustrated in Fig. 1A. All mapping of PCGP RNA-Seq data was done using the GRCh37 reference genome.

In the first phase, reads were aligned using five combinations of mapper and database. The aligners we used include BWA [36] and STAR [37]. We used the reference genome database and three custom databases. Each sequence in the custom databases is built by selecting a set of exons from a particular annotation source (Table 1). The sequences corresponding to these exons from the forward strand of the reference genome are concatenated to make the custom sequence. Information about the set of exons used is stored in the sequence name. After mapping to the custom databases, each mapping is translated into genomic coordinates using the information from the sequence name, and the reference annotation. The combinations of mapper and database used are listed in Table 1.

After initial mapping, the pipeline chooses a record for each read from among the five available. The best record

is chosen according to the following algorithm: (1) mapped records are always better than unmapped; (2) among mapped records, those with more matches are always better than those with fewer; and (3) among mapped records with the same number of matches, records with fewer indels are always better than those with more.

At this stage, there may still be several records tied for the best. If there are multiple candidates, and the reads are paired, then the best pair is chosen based on the following rules: (1) pairs on the same chromosome are always better than pairs on different chromosomes; (2) pairs that are closer (on a log₁₀ scale, with integer granularity) are then preferred; (3) pairs in forward-reverse orientation are then preferred; (4) pairs from the same mapper and database combination are finally preferred; and (5) if ties remain, then the choice is made based on the priority of the mapper and database combination used (Table 1).

Next, the individual aligned BAM files are sorted and merged, duplicates are marked, and the files are indexed using Picard tools SortSam, MergeSamFiles, and MarkDuplicates. The resulting unrefined BAM file is usable, and may be sufficient for some analyses. This completes the alignment phase of the mapping.

The second phase of the mapping is for refinement. First, records of interest are extracted using SAMExtractUnmapped in Bambino [39]. The records extracted include unmapped reads and reads with soft-clipping, indels, or high quality mismatches. The resulting files are converted to FASTA, large files are split, and small files are batched as an optimization.

These reads are then aligned to the reference genome using sim4 [40]. If the mate read is aligned, then the sim4 search is restricted to the 100 kb region on the side of the mate read that would be expected to achieve forward-reverse orientation. If the mate read is not aligned, then the sim4 search is restricted to 100 kb on either side of the original mapping position.

If the sim4 mapping is better than the original, then it is used in place of the original. The algorithm determines if the sim4 alignment is an improvement as follows: (1) if the read was originally unmapped, then it is an improvement; (2) the alignments are scored using +1 for alignment, -1 for gap open, -1 for gap extend, and +5 for any splice of length < 100 kb, and if the scores are unequal, then improvement is determined based on the scoring; (3) if the reads were previously on different chromosomes, but sim4 places them on the same chromosome, then it is an improvement; (4) if the reads were previously not in forward-reverse orientation, but sim4 places them in forward-reverse, then it is an improvement; and (5) otherwise, it is not an improvement.

At this point, the pipeline also soft-clips the alignment of poly-A tails, which may contain one or more spurious

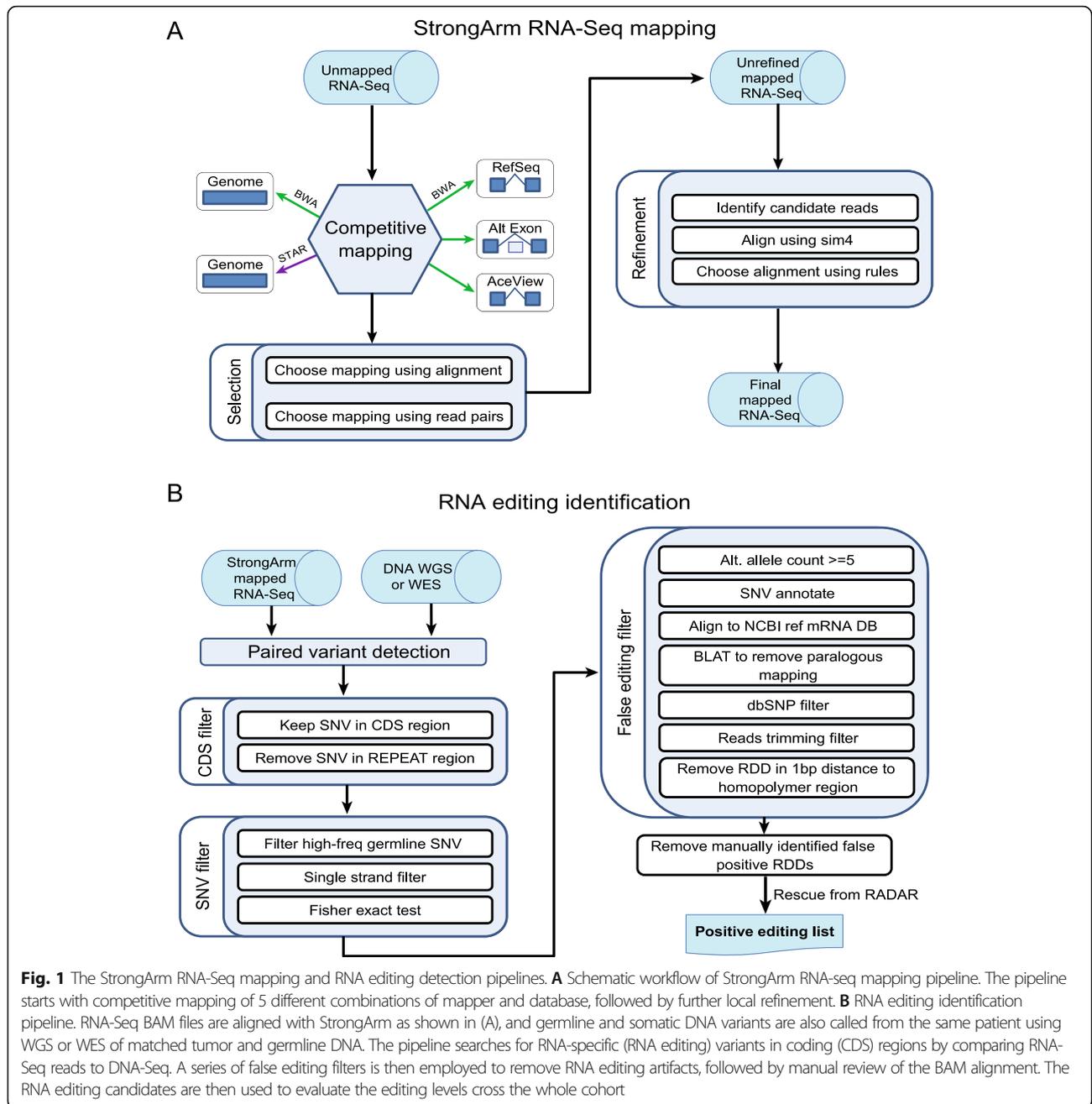


Fig. 1 The StrongArm RNA-Seq mapping and RNA editing detection pipelines. **A** Schematic workflow of StrongArm RNA-seq mapping pipeline. The pipeline starts with competitive mapping of 5 different combinations of mapper and database, followed by further local refinement. **B** RNA editing identification pipeline. RNA-Seq BAM files are aligned with StrongArm as shown in (A), and germline and somatic DNA variants are also called from the same patient using WGS or WES of matched tumor and germline DNA. The pipeline searches for RNA-specific (RNA editing) variants in coding (CDS) regions by comparing RNA-Seq reads to DNA-Seq. A series of false editing filters is then employed to remove RNA editing artifacts, followed by manual review of the BAM alignment. The RNA editing candidates are then used to evaluate the editing levels cross the whole cohort

Table 1 The five mapper and database combinations used by StrongArm

Priority	Mapper	Database
1	BWA	RefSeq : Every RefSeq transcript found in UCSC's refFlat table [38]; useful for finding canonical splicing.
2	BWA	RefSeq alternate exons : Fragments of RefSeq transcripts formed by choosing an ordered subset of exons from each gene that contains a single pair of adjacent exons which are not adjacent in any annotation, and with 100 bp of sequence on either side of the event; useful for finding alternate splicing.
3	BWA	AceView : Every AceView transcript found in UCSC's assembly table; useful for various other known splice forms.
4	BWA	Whole genome : the genome reference sequence (no translation is performed); useful for unspliced reads and some structural variation events.
5	STAR	Whole genome (STAR) : the genome reference sequence (no translation is performed); useful for some novel exons and structural variation events.

splices to poly-A runs in the reference genome. The new individual aligned BAM files are sorted and merged, duplicates are marked, and the files are indexed, as for the unrefined BAM, to create the final BAM file.

TopHat and STAR

StrongArm alignment was compared to TopHat and STAR alignment for 15 PCGP samples' RNA-seq data. TopHat version 2.1.1 using Bowtie version 2.1.0 was run in both default mode with parameter “-m 2”, and annotation-based mode with parameter “-m 2 --transcriptome-index GENCODE.v19.index”. STAR mapping was performed with STAR version 2.7.1a in two-pass mode. The GRCh37 reference genome was used. The RNA editing VAF of the 722 sites analyzed in the study were determined using the Ace2.ReadReport utility within Bambino [39] for the samples aligned with StrongArm, TopHat, and STAR.

Post-processing for detecting RNA editing events

After mapping is performed with StrongArm, the RNA editing detection pipeline starts with variant calling using Bambino [39] followed by filters to limit analysis to SNVs within gene coding regions and to remove false editing events due to germline variants, paralogous mapping, and homopolymer regions. The resultant RNA editing candidates were further curated by manual review. To rescue low-confidence editing events for which editing was not detected de novo, we reviewed the evidence of coding editing sites that were included in the RADAR database with ≥ 3 mutant reads in ≥ 2 PCGP samples (a total of 384 RNA editing events were rescued). The RADAR [5] version 2 database (hg19) was used to compare our identified editing events with RADAR. We also determined whether RNA editing events we discovered de novo were already present in DARNED or REDiportal, in addition to RADAR. For this, DARNED [6] hg19 editing sites were downloaded in February 2021 from <https://darned.ucc.ie/download/>. REDiportal [41] hg19 editing sites were downloaded February 2021 from <http://srv00.recas.ba.infn.it/atlas/download.html>. The RNA editing VAF of the 722 sites were calculated using the Ace2.ReadReport utility within Bambino [39].

Normal sample analysis using GTEx

GTEx RNA-Seq files were previously downloaded from dbGaP accession phs000424 and reads were previously mapped using STAR version 2.7.1a in two-pass mode to hg38 for another study. To avoid remapping these 5454 BAM files to hg19 using StrongArm for this study (due to the storage resources and computational throughput this would require) we instead used the existing hg38-aligned files and analyzed the RNA editing levels for

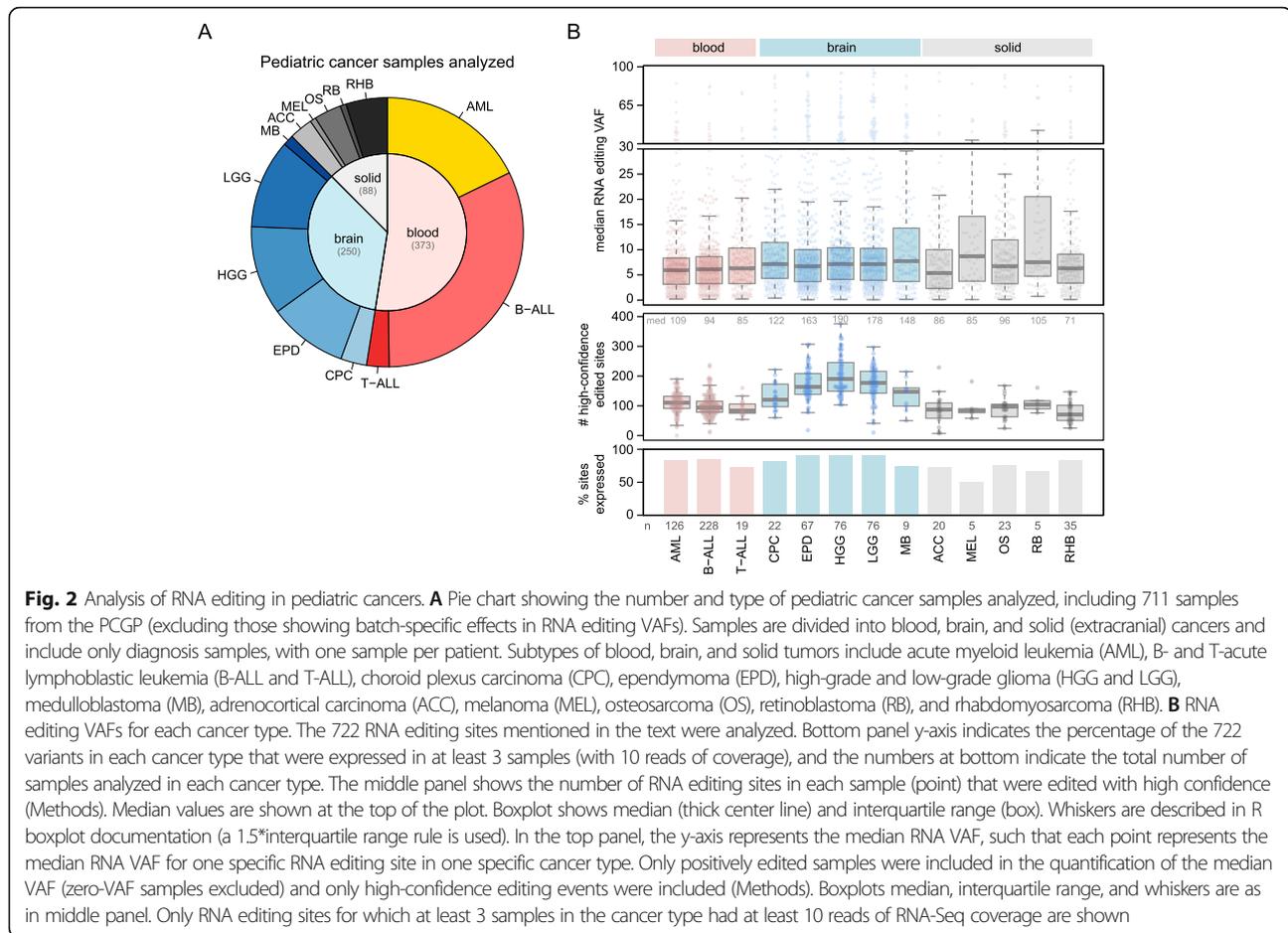
each of 722 RNA editing sites identified in pediatric cancer (converted to hg38 coordinates), using the Ace2.ReadReport utility in Bambino [39]. To verify that RNA editing results would have been similar had we re-mapped all GTEx samples to StrongArm with hg19, we correlated RNA editing in 18 GTEx samples re-mapped with StrongArm (hg19) vs. the existing STAR (hg38) alignments. In each of these 18 samples, the Pearson r correlations for total read coverage across the 722 sites when comparing the two approaches were $r > 0.97$ and the mutant read correlations were $r > 0.99$, indicating similar results between the two approaches. Sample tissue type annotations were obtained from <https://github.com/ucscGenomeBrowser/kent/blob/master/src/hg/makeDb/outside/gtexHub/metadata/sraToSample.tab>.

High-confidence RNA editing events

In Fig. 2B boxplots, only high-confidence RNA editing events were shown, and high confidence was defined as follows. If an editing site had read coverage of greater than 100, at least 3 mutant reads were required to consider the site edited with high confidence. For sites with 20–99 reads of coverage, at least 2 mutant reads were required. Finally, for sites with less than 20 reads of coverage, only 1 mutant read was required to consider the site high confidence. These thresholds were based on analysis of adjacent control sites that were within 2 base pairs upstream or downstream of the 722 RNA editing sites in 15 PCGP samples' RNA-Seq data, to quantify the background error mutation rate and thus determine what number of mutant reads indicated true-positive editing. Of the 722 RNA editing sites, 569 had a suitable adjacent control site within 2 base pairs upstream or downstream of the editing site that was of the same reference allele (e.g. A for an A > G variant) as the actual editing site. We evaluated the mutation error rate using these 569 adjacent control sites and found that among RNA editing sites with less than 20 reads of coverage, 94% of 1-mutant read variants at RNA editing sites were true-positives. However, below 100 reads of coverage, only 79% of 1-mutant read variants at RNA editing sites were true-positives, and thus 2 mutant reads were required when 20–100 reads of coverage were available (leading to 90% true-positives). Above 100 reads of coverage, sites with 3 mutant reads gave over 89% true positives, whereas 2 mutant reads gave only 76% true positives, leading to 3 mutant reads required above 100 reads of coverage.

Neoepitope prediction

Neoepitope prediction was performed using neoepiscope [42] run on all non-silent RNA editing events identified, using the chromosome, position, reference allele, and alternate allele of each editing event as input. Predictions



were made for 15 common HLA haplotypes reported in the literature [43, 44].

Results

Competitive RNA-Seq mapping using StrongArm

Current RNA-Seq alignment tools have varying properties that can affect the detection of RNA editing. For example, STAR has a high mapping rate because it permits incomplete alignments which increase the sensitivity for detecting RNA editing events, but the fraction of fully-mapped reads is lower than other tools which can lead to false-negatives. Annotation-based TopHat2 has the opposite characteristics [45]. As expression of pseudogenes and paralogs is common across the human genome, ambiguity in RNA-Seq mapping can lead to erroneous RNA editing calls [10, 16], whereas mapping of DNA sequencing reads can rely on unique intronic sequences to produce the correct mapping. Moreover, aligners like STAR will occasionally soft-clip a read when the read's end spans a splice junction (Supplementary Fig. 2A) or contains bona fide nucleotide variants (e.g. RNA editing or genetic variants). For RNA editing

analysis, this can lead to missing true RNA editing events (Supplementary Fig. 2B).

To facilitate the detection of RNA editing and other variants using RNA-Seq data, we designed an exhaustive mapping pipeline called StrongArm. This pipeline selects the mapping location of a read-pair based on mapping to five different reference database/mapper combinations, from which the best mapping is chosen (Fig. 1A). To analyze the pipeline's sensitivity, we compared StrongArm results with two popular aligners, STAR and TopHat2, using RNA-Seq data from 15 pediatric cancer samples. StrongArm and STAR had higher mapping rates than TopHat2 or annotation-based TopHat2 (Supplementary Fig. 3A), with the algorithmic trade-off of more soft-clipped reads with StrongArm and STAR (Supplementary Fig. 3B). Compared to STAR, StrongArm was able to map more reads at full length without soft-clipping (Supplementary Fig. 3B, C), largely due to improved mapping around splice junctions and non-reference genomic locations (Supplementary Fig. 2). StrongArm alignment also led to significantly more RNA editing sites being evaluable (which we defined by at least 10 reads of coverage at the editing site, as done

by others [8]) than STAR (Supplementary Fig. 4A), although when coverage was evaluable with both tools the RNA editing VAFs were comparable (Supplementary Fig. 4B).

Identification of coding RNA editing events in pediatric cancers

We analyzed RNA editing in 954 pediatric cancer samples from PCGP (later filtered to 711 as described in Methods; Fig. 2A, Supplementary Fig. 1) using RNA-Seq data mapped with StrongArm (Fig. 1A). We first identified RNA-specific single-nucleotide variants (SNVs) from the aligned RNA-seq reads using Bambino [39], followed by applying multiple filters to remove false-positives (Fig. 1B, Methods). The pipeline eliminates germline and somatic DNA variants from consideration using matched tumor-normal WGS or WES (available for 717 cases) and public single nucleotide polymorphism (SNP) data (Fig. 1B, Methods). We focused on variants in coding regions for this analysis.

To remove additional false positive hits, we performed manual curation by visually inspecting the alignment of each variant [39] (Fig. 1B, final step) to remove the remaining false positives in the following categories. (1) Co-occurrence of SNPs with paralogous variants. For example, a false-positive RNA editing call in the mono-exonic *GLUD2* gene was in fact a SNP (rs9421572) in its multi-exonic paralog *GLUD1* near an exon-intron boundary (described in Supplementary Fig. 5A). This error was due to the gap-open penalty incurred by a splice junction in *GLUD1* resulting in a preference for an unspliced region in *GLUD2* during mapping. (2) Variants at the 3' side of homopolymers on the antisense strand, suggesting an error introduced during reverse transcription (Supplementary Fig. 5B). (3) Imperfect genome annotation. For example, some purported RNA editing events could be accounted for by alternative splicing to exons not included in the transcript model during mapping, thus leading to spurious variant calls, as in the case of *PHB2* (Supplementary Fig. 5C). Manual curation is an important step as automated analysis identified 334 to 1530 (731 on average, after removing somatic SNVs) putative RNA editing events per sample based on de novo variant calling. However, 95 to 99% (99% on average) of these were recognized to be artifacts during manual review. A total of 340 unique editing sites were detected in the PCGP cohort.

Combining these editing sites detected de novo with those present in the RADAR database (Methods), we identified 722 unique RNA editing sites in coding regions post manual curation (Supplementary Table 1), including 706 canonical A-to-I events and 14 canonical C-to-U events. Approximately 70% of these were non-synonymous missense (498), nonsense (1), or stop-loss

(8) variants while the remaining 30% were synonymous. Further, 584 of 722 (81%) were previously reported in RNA editing databases (e.g. RADAR [5], DARNED [6], and/or REDportal [41]). Of the remaining 138 sites not found in these databases, 90 represented a new variant affecting a gene known to have editing events in these databases (e.g. missense but at a different amino acid), while 48 sites represent the first reported RNA editing event affecting protein-coding (e.g. missense, nonsense, or stop-loss) of the gene. We selected 10 editing sites and performed experimental validation by deep amplicon sequencing using paired DNA/RNA samples in five leukemias with varying RNA VAFs; all variants were confirmed to be present exclusively in RNA samples (Supplementary Table 2).

We then analyzed the prevalence of RNA editing on these 722 sites across the major subtypes of pediatric cancer, including blood, brain, and solid (extracranial) cancers (Fig. 2A). Brain cancers had more editing events than other cancers (Fig. 2B, middle panel), with a median of 190 positive RNA editing events per sample in high-grade glioma (HGG) and 178 in low-grade glioma (LGG), as compared to blood cancers (from median 85 in T-ALL to 109 in AML) and solid tumors (from median 71 in rhabdomyosarcoma (RHB) to 105 in retinoblastoma (RB)). RNA editing VAFs were low overall in edited transcripts, with similar VAFs across cancer types (Fig. 2B, top, which shows median VAFs at approximately 0.05 for each RNA editing site in each cancer).

Comparison of RNA editing between pediatric cancers and normal tissue

We next asked whether the 722 RNA editing sites identified in pediatric cancer were also found in normal tissues by analyzing 5454 RNA-Seq samples from GTEx. Only 7 of the 722 RNA editing events were tumor-specific, as the remainder could be found in one or more normal tissues (Fig. 3, Supplementary Table 3); 6 of these 7 sites were only edited in a few cancer samples, while the 7th (a *MEX3C* A74A silent variant) was edited specifically in blood cancers (Supplementary Tables 3–4). We observed RNA editing events which were enriched in both normal and leukemic blood samples (Fig. 3, top box), and events enriched in both normal and cancerous brain samples (Fig. 3, bottom box). This suggests that RNA editing in pediatric cancer is related to the tissue of origin, rather than tumor-specific effects. Some tissue specificity of RNA editing was related to tissue-specific expression of the edited gene, rather than enrichment of editing per se (blue in Fig. 3 indicates lack of expression; see also specific variants discussed below).

Figure 4 shows the tissue-specific profiles of example RNA editing events with patterns of interest, including ubiquitous editing with ubiquitous expression; tissue-



specific editing with tissue-specific expression; and tissue-specific editing with ubiquitous expression. For example, some RNA editing events were ubiquitous across tissue types in genes with ubiquitous expression, such as previously reported *NEIL1* K242R [46] editing (Fig. 4A). Some were enriched in normal blood and leukemic samples due to tissue-specific expression of the gene, including *IL12RB1* R356G (Fig. 4B), detected previously in normal immune cells [47]. Glutamate receptors, including *GRIK1*, *GRM4*, and *GRIA2*, and other receptors involved in neurotransmission were also enriched in both normal and malignant brain samples (Fig. 4C), consistent with studies in normal neural tissue [48]. RNA editing in various calcium and other ion-binding proteins, which can

also affect neurotransmission, were likewise enriched in brain samples (Fig. 4D) as expected [5, 6, 49]. The above tissue-specific effects were related to increased expression of the edited gene itself, not necessarily increased RNA editing. By contrast, a few genes including *METTL10* and *PDCD7* had widespread expression across most tissue types, but brain-enriched editing (Fig. 4E), suggesting an increase in editing itself in brain tissue. While the above editing sites are previously reported, a few novel sites showed tissue-specific editing, such as the brain (glioma)-enriched *LRP4* gene (Fig. 4F). Finally, some RNA editing sites showed depletion in a specific tissue type, such as the lack of *IGFBP7* editing in normal and leukemic blood samples (Fig. 4G).

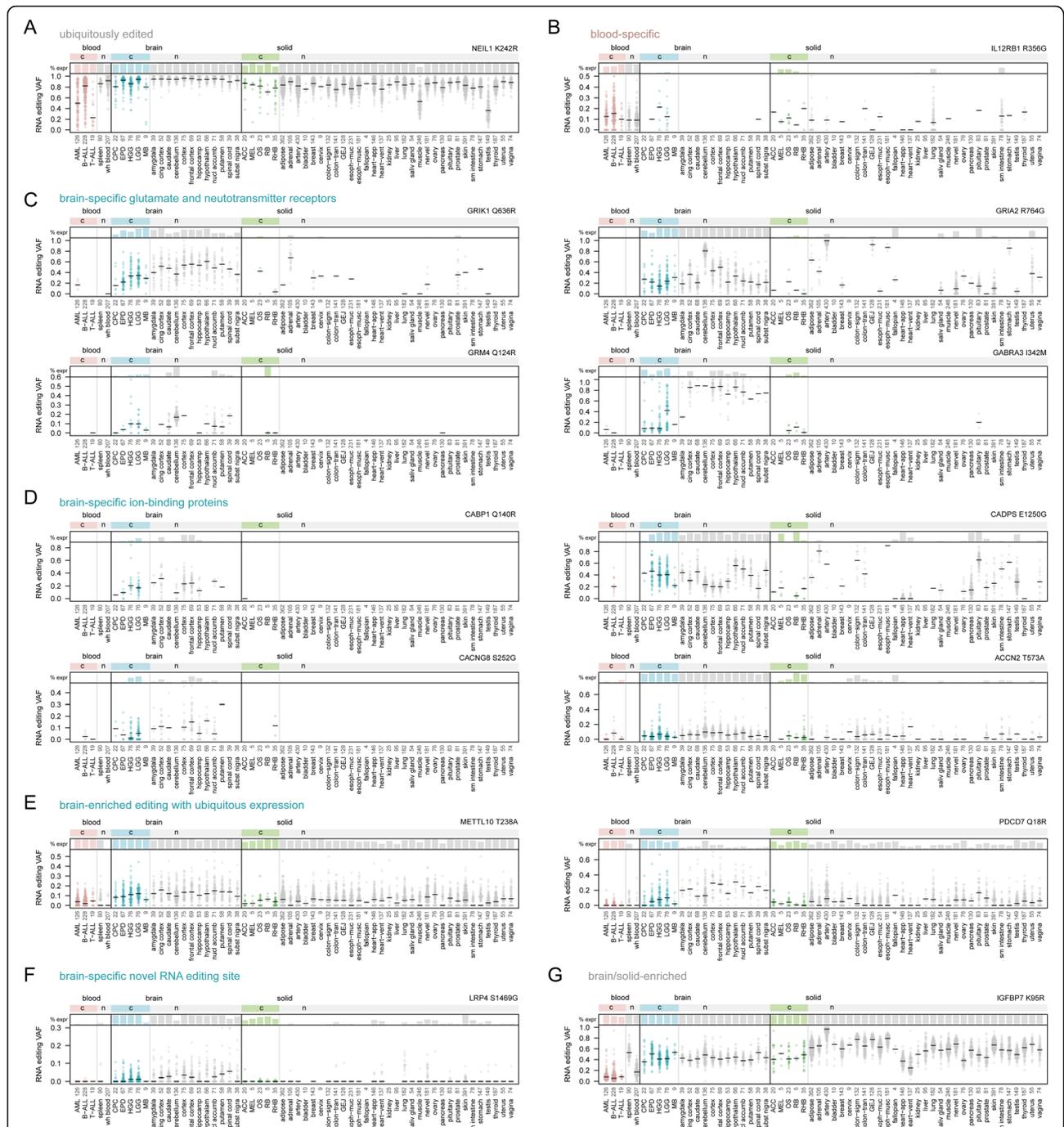


Fig. 4 Tissue specificity of selected RNA editing events. Plots are shown for selected RNA editing events taken from among the 722 sites identified in the study. Bottom y-axis for each graph represents the RNA editing VAF for the variant noted. Each point represents one cancer or normal sample, and horizontal black bars represent the median for each cancer or normal tissue type. The top portion of each graph indicates the percentage of samples in which the editing site is expressed (“% expr”) with at least 10 reads of sequencing coverage. VAFs are only shown in the bottom panels for samples meeting this criterion. Samples are divided into cancer (c) or normal (n) tissue types as in Fig. 3 (see Supplementary Table 6 for cancer type abbreviation definitions). The RNA editing site is shown at the top-right of each graph, expressed as the gene and amino acid change caused by the editing event. Each panel highlights an editing event with a specific pattern of interest, including (A) a ubiquitously edited site, (B) a site both edited and expressed primarily in blood cells, (C and D) sites both expressed and edited preferentially in the brain, (E) sites with ubiquitous expression and editing enrichment in brain, (F) a brain-enriched editing site not reported in RADAR, DARNED, or REDportal, unlike the others in this figure, and (G) a site edited preferentially in solid and brain tissue but not edited in most blood tissues

Previously reported editing events in *GRIA2*, *AZINI1*, and *COG3* are thought to promote adult tumor progression based on functional cell viability assays [8]. We also detected these events in multiple pediatric cancer types, suggesting they may promote pediatric cancer progression as well, although they are also present in multiple normal tissues (Supplementary Tables 3–4).

Association between RNA editing and splicing

RNA splicing and editing can occur co-transcriptionally and RNA editing may alter splicing [50, 51]. This could lead to different editing levels between mRNA and pre-mRNA, which can be analyzed by comparing editing levels among spliced mRNA and unspliced (intron-retaining) pre-mRNA. While we used polyA-enriched RNA-Seq data to analyze PCGP samples, we nonetheless observed substantial intronic reads in many genes, indicating the presence of some pre-mRNA with which this analysis could be performed.

To assess the spliced mRNA editing level (rather than the overall RNA editing level used in our previous analyses), we first identified the reads for which the read or its mate pair were mapped to splice junctions. These “spliced reads” were considered as a signal of mRNA which were then used to calculate the (spliced) mRNA editing level (Fig. 5A). The difference between the overall editing level (VAF) and mRNA editing level (VAF) were compared using a Wilcoxon Rank Sum test for each editing site.

As shown in Fig. 5B, since polyA-enriched RNA-Seq filters out many immature RNAs, most RNA editing sites do not show significant differences between the overall RNA editing level and the spliced mRNA (junction-corrected) editing level. However, several editing sites showed a significant difference. For example, *NARF* (Fig. 5C, top-left) had a lower editing level in mRNA, but higher overall editing level, suggesting a relationship between editing and splicing. This could be because unedited *NARF* pre-mRNAs are more prone to splicing, or because edited *NARF* mRNAs are more prone to degradation. By contrast, *SORBS1* (Fig. 5C, bottom-left) has more editing in mRNA, possibly because edited *SORBS1* pre-mRNA is more prone to splicing.

Because the junction-corrected mRNA level is calculated from polyA-enriched RNA-Seq in these cases, the real difference between mRNA and pre-mRNA might be more apparent if total RNA library preparation was used. Therefore, we compared a subset of samples sequenced by both total stranded and polyA-enriched RNA-Seq. The *GRIA2* gene, for example, had moderately (though significantly) different editing between mRNA and pre-mRNA using polyA-enriched RNA-Seq (Fig. 5B; Fig. 5C, top-right). However, in total stranded RNA-Seq the difference was much more apparent (Fig. 5C, bottom-right),

indicating that more splicing-editing correlations may exist than our polyA-focused analysis suggests. Because RNA editing relies on local RNA secondary structures, it is possible that editing preferentially targets spliced transcripts in some genes, and unspliced in others, based on secondary structure. Moreover, editing can create cryptic GT-AG splicing sites to affect splicing output and efficacy. Thus, the interplay between splicing and RNA editing is likely to be gene-specific, consistent with our results (Fig. 5C).

Discussion

We used a competitive mapping approach to improve paralog mapping and mitigate alignment errors when reads span splice junctions for discovery of RNA editing. This study is the first comprehensive analysis of RNA editing in pediatric cancers, which has been previously studied in adult cancers [8, 9]. We have noted both known and novel RNA editing sites whose biological function could be investigated in future studies.

It has been proposed that RNA editing may lead to neoepitopes that may be targeted by anti-cancer immunotherapies [52]. Indeed, we used neoepiscopes [42] to run neoepitope prediction on peptides resulting from RNA editing events to identify peptides predicted to be presented effectively by 15 common HLA class I haplotypes [43, 44]; many of these peptides were predicted to bind effectively to these HLA alleles and thus would theoretically generate epitopes that could be recognized by T cells (Supplementary Table 5). However, our results indicate that, in pediatric cancer, RNA editing events are not specifically enriched in tumors, which would make immunotherapeutic targeting difficult. Rather, nearly all RNA editing events we detected in pediatric cancer were also present in one or more normal tissues; indeed, the profile of each pediatric cancer type essentially matched that of its normal tissue of origin. This suggests that immunotherapeutic efforts focused on identifying tumor-specific RNA editing will be difficult to implement in pediatric cancer. However, if RNA-edited genes are overexpressed in a specific cancer type compared to normal, it may be possible to target these editing events immunotherapeutically with good therapeutic index. For example, the *CD6* gene is overexpressed in a subset of leukemias, with expression of over 100 transcripts per million (TPM) in 10% of acute myeloid leukemia (AML), 18% of T-lineage acute lymphoblastic leukemia (T-ALL), and 2% of B-ALL, but virtually no expression in other cancerous tissues (Supplementary Fig. 6A). *CD6* undergoes RNA editing leading to an S52G missense variant (Supplementary Tables 1, 3, 4), and thus targeting this peptide immunotherapeutically may have therapeutic potential. *CD6* is also overexpressed in some normal blood and small intestinal tissues from GTEx (Supplementary Fig. 6B) concomitant with S52G

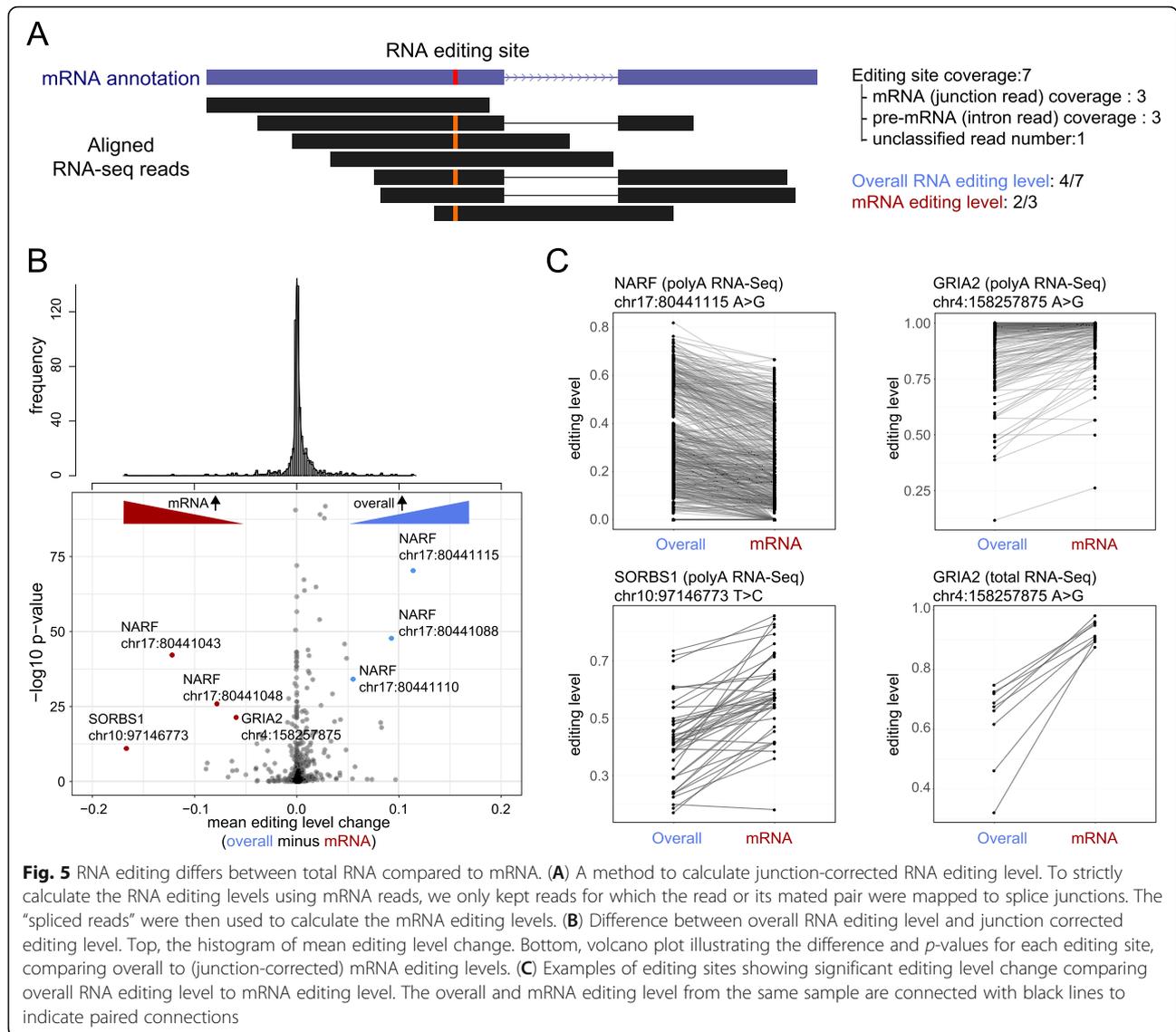


Fig. 5 RNA editing differs between total RNA compared to mRNA. **(A)** A method to calculate junction-corrected RNA editing level. To strictly calculate the RNA editing levels using mRNA reads, we only kept reads for which the read or its mated pair were mapped to splice junctions. The “spliced reads” were then used to calculate the mRNA editing levels. **(B)** Difference between overall RNA editing level and junction corrected editing level. Top, the histogram of mean editing level change. Bottom, volcano plot illustrating the difference and p -values for each editing site, comparing overall to (junction-corrected) mRNA editing levels. **(C)** Examples of editing sites showing significant editing level change comparing overall RNA editing level to mRNA editing level. The overall and mRNA editing level from the same sample are connected with black lines to indicate paired connections

editing (Supplementary Tables 3, 4), but this may be acceptable given that transient immunosuppression or gastrointestinal issues may be an acceptable toxicity of anticancer therapy.

Further, if certain RNA editing events are preferentially presented via HLA molecules on tumor cells in preference to normal cells, these may represent valid targets (despite RNA editing in both normal and tumor cells) as shown for *CCNI* editing in some adult cancers [53], an editing event which we also observed in all pediatric cancer types analyzed (Supplementary Tables 3–4). While beyond the scope of this study, it would be of interest for future studies to test whether any of the RNA editing events we identified are preferentially HLA-loaded on tumor cells compared to normal cells to identify potential therapeutic targets.

Conclusions

These findings indicate that the RNA editing profile of each pediatric cancer type is similar to its corresponding normal tissue of origin. Thus, the somatic mutations present in pediatric cancers do not appear to promote tumor-specific RNA editing events which can be immuno-therapeutically targeted. Rather, RNA editing profiles in pediatric cancer likely result from the transcriptional state the cancer inherits from the original normal tissue. However, RNA editing events occurring in cancers with overexpression of the edited transcript may provide therapeutic targets, which merits further study. These results also provide a map of the coding RNA editing events across pediatric cancers, and identify novel RNA editing events whose function should be explored in future studies.

Abbreviations

ACC: Adrenocortical carcinoma; ADAR: Adenosine deaminase acting on RNA; AML: Acute myeloid leukemia; APOBEC: Apolipoprotein B mRNA editing enzymes; B-ALL: B-lineage acute lymphoblastic leukemia; CDS: Coding regions; CPC: Choroid plexus carcinoma; EPD: Ependymoma; GTEX: Genotype-Tissue Expression project; HGG: High-grade glioma; LGG: Low-grade glioma; MB: Medulloblastoma; MEL: Melanoma; OS: Osteosarcoma; PCGP: St. Jude/Washington University Pediatric Cancer Genome Project; RB: Retinoblastoma; RHB: Rhabdomyosarcoma; SNP: Single nucleotide polymorphism; SNV: Single nucleotide variant; T-ALL: T-lineage acute lymphoblastic leukemia; TCGA: The Cancer Genome Atlas; TPM: Transcripts per million; VAF: Variant allele fraction; WES: Whole-exome sequencing; WGS: Whole-genome sequencing

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12885-021-08956-5>.

Additional file 1: Supplementary Fig. 1. RNA editing analysis workflow and samples included. Workflow showing the discovery process and the criteria for sample inclusion at each step. To discover de novo potentially novel RNA editing sites (top), SNV calling was performed on samples with tumor RNA-Seq, and germline plus tumor DNA-Seq (either WGS or WES), including 717 samples (including 716 pediatric cancer samples from the PCGP and one leukemia cell line, Nalm6). After filtering and discovering RNA editing events in these 717 samples as shown in Fig. 1B, the presence of each RNA editing event discovered de novo or rescued from RADAR ($n = 722$ sites total) was analyzed in 954 pediatric cancer RNA-Seq samples from the PCGP, whether or not DNA-Seq was available (middle). Finally, certain samples with batch effects in RNA editing VAFs and relapsed or duplicate samples were filtered out, resulting in 711 pediatric cancer diagnosis samples with one diagnosis sample per patient (bottom). This set of 711 samples was the primary sample set shown in analyses in this study. **Supplementary Fig. 2.** Examples comparing alignment between StrongArm vs. STAR mapping. (A) Splice junction-adjacent reads that failed mapping by STAR (blue) but mapped by StrongArm (red) as viewed in the BAM alignment. This shows a view of one sample's BAM files aligned by the two tools, with each row representing one read, and part of the *COPA* gene is shown. Soft-clipped read regions have a darker gray appearance. Gray arrows (>) indicate that part of the read is aligned elsewhere (to another exon). Non-reference sites are shown in red, including an RNA editing site (T > C). Only reads containing the edited allele are shown. (B) Read with RNA editing at the end that failed mapping (soft-clipped) by STAR but mapped by StrongArm. Coloring and other features are as in (A), except that the *MRPS27* gene is shown. This shows a view of one sample's BAM files aligned by the two tools. Only reads containing the edited allele (G > A, red) are shown. **Supplementary Fig. 3.** Mapping rate and soft-clipping comparison between StrongArm and other tools. (A) Mapping rate comparison between StrongArm, STAR, TopHat2 and annotation-based TopHat2. The mapping rate of the same sample in different aligners is connected by a dotted line. Y-axis represents the percent of mapped reads, and 15 selected pediatric cancer samples from PCGP are analyzed. (B) The percentage of reads with soft-clipped nucleotides in each aligner. The same sample is connected by a dotted line. TopHat2 does not map soft-clipped reads. (C) The percentage of reads (y-axis) with different numbers of soft-clipped bases (x-axis). The unclipped reads (0) are separated due to different axis scale. This compares STAR (blue points) and StrongArm (red points) and does not include TopHat2, since TopHat2 does not map soft-clipped reads. Each point represents one of the 15 PCGP samples shown in (A) and (B). Boxplot shows median (thick center line) and interquartile range (box). Whiskers are described in R boxplot documentation (a 1.5*interquartile range rule is used). **Supplementary Fig. 4.** Comparison of RNA editing detection between STAR- and StrongArm-mapped RNA-Seq data. This analysis includes 15 PCGP samples' RNA-Seq data mapped by both StrongArm and STAR (the same samples used for Supplementary Fig. 3). (A) More RNA editing sites are evaluable by StrongArm than by STAR mapping. Boxplot shows the number of RNA editing sites (among the 722 RNA editing sites evaluated in the study) that are informative (at least

10 reads of coverage) only by STAR or only by StrongArm mapping in each sample. Boxplot shows median (thick center line) and interquartile range (box). Whiskers are described in R boxplot documentation (a 1.5*interquartile range rule is used). (B) High concordance of RNA editing VAFs derived from STAR-mapped and StrongArm-mapped BAM files, when analyzing RNA editing sites with at least 10 reads of coverage with both STAR and StrongArm. See table at top-left of each graph to see the number of variants falling into this category (the "Both (shown)" category). Each graph shows one leukemia patient. Each point represents a single RNA editing event positioned by its RNA editing VAF by StrongArm (x-axis) or by STAR (y-axis). Dotted line represents the identity line ($x = y$). r value is by Pearson correlation. **Supplementary Fig. 5.** False-positive RNA editing examples requiring manual removal. (A) Example false RNA editing due to the presence of a germline SNP in one of two paralogs. *GLUD1* and *GLUD2* are paralogs; *GLUD1* has introns while *GLUD2* is non-exonic. The coding regions (CDS) have 98% identity between the two genes. Left, in patients with *GLUD1* SNP rs9421572 (a SNP which is near a splice junction in exon 7), DNA (WGS) reads will correctly map the SNP to *GLUD1* instead of *GLUD2*, as the intronic region of reads containing the SNP will map uniquely to *GLUD1* but not to intron-less *GLUD2*. Right, in RNA-Seq, by contrast, neither *GLUD1* nor *GLUD2* mRNAs contain introns, and therefore the mapping will prefer *GLUD2* since mapping that does not require the read to span a splice junction is preferred, and *GLUD2* lacks introns making it preferred. The *GLUD1* SNP is therefore aberrantly considered to be a *GLUD2* RNA editing event. Such events must be removed by manual curation. (B) Example of false RNA editing due to the presence of a homopolymer genomic region using an example in the *RNF19A* gene. False-positive editing events frequently occurred within one base position of homopolymer sequences on the 3' side of the homopolymer along the antisense strand, suggesting the error was introduced during reverse transcription. Such events must be removed by manual curation. (C) Example of false RNA editing due to mapping to the wrong splice variant. The *PHB2* gene includes a very small (or "nano") exon in one of the UCSC transcripts (uc021qug.1). However, none of the RefSeq transcripts used for mapping includes this exon; therefore, reads which include the nano exon are instead incorrectly mapped to NM_001144831 which lacks the nano exon. Thus, RNA-Seq reads for many samples, including the example adrenocortical cancer sample SJACT001, incorrectly map to NM_001144831 and appear to have an RNA editing event in a region that should in fact map to the nano exon. **Supplementary Fig. 6.** An example of outlier expression of an RNA-edited gene in a specific tissue type. (A) Boxplot showing *CD6* expression in TPM across PCGP pediatric cancer tissues. Parentheses indicate the number of samples analyzed. See Fig. 2 legend for cancer type abbreviations. Each blue point represents one cancer sample. Outlier samples, as determined by the R "boxplot" function, are also outlined in red. Boxplot shows median (thick center line) and interquartile range (box). Whiskers are described in R boxplot documentation (a 1.5*interquartile range rule is used). (B) As in (A), except that the expression of *CD6* is shown in normal GTEX tissues. Due to different RNA-Seq library preparation and sequencing methods, the data in panels (A) and (B) may not be directly comparable to one another.

Additional file 2: Supplementary Table 1. 722 RNA editing sites identified in pediatric cancer samples. **Supplementary Table 2.** Deep amplicon sequencing of RNA validation of 10 RNA editing events. **Supplementary Table 3.** Percentages of cancer and normal tissue samples with RNA editing at each site. **Supplementary Table 4.** VAFs of RNA editing events in each sample. **Supplementary Table 5.** Neopeptide predictions for 722 RNA editing sites. **Supplementary Table 6.** Abbreviations for cancer and normal tissue types.

Acknowledgements

Not applicable.

Authors' contributions

JZ conceived the study. JW, MR, SWB, MNE, TIS, BBP, and LT performed bioinformatic analysis. YS performed experimental analysis under the direction of JE. CGM provided leukemia samples for validation experiments. TG helped to curate initial pipeline development including providing

samples. JZ and DE supervised the study. SWB, JW, MR and JZ wrote the manuscript.

Funding

This research was also supported by Cancer Center Support Grant P30CA021765 from the National Institutes of Health and in part by the American Lebanese Syrian Associated Charities (ALSAC). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Raw sequencing data for all cancer samples are available on St. Jude Cloud [32] (<https://pecan.stjude.cloud/permalink/maediting>) and BAM files are available on EGA at accessions referenced in past PCGP-related studies [22–31] (EGAD00001000261, EGAD00001001864, EGAD00001000260, EGAD00001000161, EGAD00001000259, EGAD00001000164, EGAD00001000160, EGAD00001000163, EGAD00001000268, EGAD00001000162, EGAD00001000085, EGAD00001000165, EGAD00001000135, EGAD00001000159). GTEx data are available on dbGaP under accession phs000424.v8.p2.

Declarations

Ethics approval and consent to participate

PCGP cancer and matched normal samples were obtained under written informed consent which was provided by a parent or guardian of each child, or by patients themselves if 18 years of age or older. This study was approved by the institutional review board at St. Jude Children's Research Hospital.

Consent for publication

Not applicable.

Competing interests

CGM has received consulting and speaking fees from Illumina and Amgen, and research support from Loxo Oncology, Pfizer and Abbvie, and holds stock in Amgen.

Author details

¹Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA. ²Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA. ³Department of Pediatrics, Stanford University, Palo Alto, California 94305, USA.

Received: 8 August 2021 Accepted: 2 November 2021

Published online: 17 November 2021

References

- Yablonovitch AL, Deng P, Jacobson D, Li JB. The evolution and adaptation of A-to-I RNA editing. *Zhang J. PLoS Genet.* 2017;13(11):e1007064. <https://doi.org/10.1371/journal.pgen.1007064>.
- Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem.* 2010;79(1):321–49. <https://doi.org/10.1146/annurev-biochem-060208-105251>.
- Blanc V, Davidson NO. APOBEC-1-mediated RNA editing. *Wiley Interdiscip Rev Syst Biol Med.* 2010;2:594–602.
- Bazak L, Levanon EY, Eisenberg E. Genome-wide analysis of Alu editability. *Nucleic Acids Res.* 2014;42(11):6876–84. <https://doi.org/10.1093/nar/gku414>.
- Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 2014;42(13):i. <https://doi.org/10.1093/nar/gku004>.
- Kiran A, Baranov PV. DARNED: a Database of RNA editing in humans. *Bioinformatics.* 2010;26:1772–6.
- Gassner FJ, Zaborsky N, Buchumenski I, Levanon EY, Gatterbauer M, Schubert M, et al. RNA editing contributes to epitranscriptome diversity in chronic lymphocytic leukemia. *Leukemia.* 2021;35:1053–63.
- Han L, Diao L, Yu S, Xu X, Li J, Zhang R, et al. The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell.* 2015;28(4):515–28. <https://doi.org/10.1016/j.ccell.2015.08.013>.
- Peng X, Xu X, Wang Y, Hawke DH, Yu S, Han L, et al. A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer. *Cancer Cell.* 2018;33:817–828.e7.
- Kleinman CL, Adoue V, Majewski J. RNA editing of protein sequences: a rare event in human transcriptomes. *RNA.* 2012;18(9):1586–96. <https://doi.org/10.1261/rna.033233.112>.
- Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science.* 2012;335:1302.
- Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science.* 2012;335:1302.
- Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science.* 2012;335:1302.
- Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science.* 2011;333:53–8.
- Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* 2014;24:365–76.
- Schrider DR, Gout J-F, Hahn MW. Very Few RNA and DNA Sequence Differences in the Human Transcriptome. Awadalla P, editor. *PLoS One.* 2011;6:e25842, 10, DOI: <https://doi.org/10.1371/journal.pone.0025842>.
- Piskol R, Peng Z, Wang J, Li JB. Lack of evidence for existence of noncanonical RNA editing. *Nat Biotechnol.* 2013;31(1):19–20. <https://doi.org/10.1038/nbt.2472>.
- Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods.* 2012;9(6):579–81. <https://doi.org/10.1038/nmeth.1982>.
- Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, et al. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. *Cell Rep.* 2014;6(5):906–15. <https://doi.org/10.1016/j.celrep.2014.01.037>.
- Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet.* 2012;13(1):36–46. <https://doi.org/10.1038/nrg3117>.
- Parker M, Mohankumar KM, PUNCHIHEWA C, WEINLICH R, DALTON JD, LI Y, et al. C11orf95-RELA fusions drive oncogenic NF- κ B signalling in ependymoma. *Nature.* 2014;506(7489):451–5. <https://doi.org/10.1038/nature13109>.
- Downing JR, Wilson RK, Zhang J, Mardis ER, Pui C-H, Ding L, et al. The Pediatric Cancer Genome Project. *Nature Genetics.* 2012;44:619–22.
- Chen X, Stewart E, Shelat AA, Qu C, Bahrami A, Hatley M, et al. Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer Cell.* 2013;24(6):710–24. <https://doi.org/10.1016/j.ccr.2013.11.002>.
- Chen X, Bahrami A, Pappo A, Easton J, Dalton J, Hedlund E, et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep.* 2014;7(11):104–12. <https://doi.org/10.1016/j.celrep.2014.03.003>.
- Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline mutations in predisposition genes in pediatric Cancer. *N Engl J Med.* 2015;373(24):2336–46. <https://doi.org/10.1056/NEJMoa1508054>.
- Roberts KG, Li Y, Payne-Turner D, Harvey RC, Yang Y-L, Pei D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N Engl J Med.* 2014;371:1005–15.
- Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature.* 2012;481(7380):157–63. <https://doi.org/10.1038/nature10725>.
- Faber ZJ, Chen X, Gedman AL, Boggs K, Cheng J, Ma J, et al. The genomic landscape of core-binding factor acute myeloid leukemias. *Nat Genet.* 2016;48(12):1551–6. <https://doi.org/10.1038/ng.3709>.
- Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature.* 2017;547(7663):311–7. <https://doi.org/10.1038/nature22973>.
- Liu Y, Easton J, Shao Y, Maciaszek J, Wang Z, Wilkinson MR, et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat Publ Gr.* 2017;49(8):1211–8. <https://doi.org/10.1038/ng.3909>.
- Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet.* 2019;51(2):296–307. <https://doi.org/10.1038/s41588-018-0315-5>.
- McLeod C, Gout AM, Zhou X, Thrasher A, Rahbarinia D, Brady SW, et al. St. Jude Cloud-a Pediatric Cancer Genomic Data Sharing Ecosystem. *Cancer Discov.* 2021;11:1082–99.
- Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, et al. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *N Engl J Med.* 2009;361:1058–66.

34. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics*. 2007;23(10):1289–91. <https://doi.org/10.1093/bioinformatics/btm091>.
35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
36. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 2009;25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
37. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
38. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res*. 2013;41(D1):D64–9. <https://doi.org/10.1093/nar/gks1048>.
39. Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics*. 2011;27(6):865–6. <https://doi.org/10.1093/bioinformatics/btr032>.
40. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*. 1998;8(9):967–74. <https://doi.org/10.1101/gr.8.9.967>.
41. Mansi L, Tangaro MA, Lo Giudice C, Flati T, Kopel E, Schaffer AA, et al. REDportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. *Nucleic Acids Res*. 2021;49:D1012–9.
42. Wood MA, Nguyen A, Struck AJ, Ellrott K, Nellore A, Thompson RF. Neopepscope improves neoepitope prediction with multivariant phasing. *Bioinformatics*. 2020;36:713–20.
43. Bardi MS, Jarduli LR, Jorge AJ, Camargo RBOG, Carneiro FP, Gelinski JR, et al. HLA-A, B and DRB1 allele and haplotype frequencies in volunteer bone marrow donors from the north of Parana State. *Rev Bras Hematol Hemoter*. 2012;34(1):25–30. <https://doi.org/10.5581/1516-8484.20120010>.
44. Jawdat D, Uyar FA, Alaskar A, Müller CR, Hajeer A. HLA-A, -B, -C, -DRB1, -DQB1, and -DPB1 Allele and Haplotype Frequencies of 28,927 Saudi Stem Cell Donors Typed by Next-Generation Sequencing. *Front Immunol*. 2020;0:2257.
45. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rättsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*. 2013;10(12):1185–91. <https://doi.org/10.1038/nmeth.2722>.
46. Yeo J, Goodman RA, Schirle NT, David SS, Beal PA. RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. *Proc Natl Acad Sci U S A*. 2010;107(48):20715–9. <https://doi.org/10.1073/pnas.1009231107>.
47. Turner AJ, Aggarwal P, Miller HE, Waukau J, Routes JM, Broeckel U, et al. The introduction of RNA-DNA differences underlies interindividual variation in the human IL12RB1 mRNA repertoire. *Proc Natl Acad Sci U S A*. 2015; 112(50):15414–9. <https://doi.org/10.1073/pnas.1515978112>.
48. Seeburg PH, Higuchi M, Sprengel R. RNA editing of brain glutamate receptor channels: Mechanism and physiology. *Brain Res Rev*. 1998;26:217–29. [https://doi.org/10.1016/S0165-0173\(97\)00062-3](https://doi.org/10.1016/S0165-0173(97)00062-3).
49. Lo Giudice C, Tangaro MA, Pesole G, Picardi E. Investigating RNA editing in deep transcriptome datasets with REDtools and REDportal. *Nat Protoc*. 2020;15(3):1098–131. <https://doi.org/10.1038/s41596-019-0279-7>.
50. Hsiao YHE, Bahn JH, Yang Y, Lin X, Tran S, Yang EW, et al. RNA editing in nascent RNA affects pre-mRNA splicing. *Genome Res*. 2018;28:812–23.
51. Laurencikiene J, Källman AM, Fong N, Bentley DL, Öhman M. RNA editing and alternative splicing: the importance of co-transcriptional coordination. *EMBO Rep*. 2006;7:303–7.
52. Zhou C, Zhu C, Liu Q. Toward in silico Identification of Tumor Neoantigens in Immunotherapy. *Trends Mol Med*. 2019;25(11):980–92. <https://doi.org/10.1016/j.molmed.2019.08.001>.
53. Zhang M, Fritsche J, Roszik J, Williams LJ, Peng X, Chiu Y, et al. RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat Commun*. 2018;9(1):1–10. <https://doi.org/10.1038/s41467-018-06405-9>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

