SPLASH2 provides ultra-efficient, scalable, and unsupervised discovery on raw sequencing reads

Marek Kokot[1,*], Roozbeh Dehghannasiri[2,3,*], Tavor Baharav[4], Julia Salzman[2,3,5,†], Sebastian Deorowicz[1,†]

Author affiliation

[1]Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland.
[2]Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA.
[3]Department of Biochemistry, Stanford University, Stanford, 94305, USA.
[4]Department of Electrical Engineering, Stanford University, Stanford, 94305, USA.
[5]Department of Statistics (by courtesy), Stanford University, Stanford, 94305, USA.

[†]Co-corresponding authors: julia.salzman@stanford.edu, sebastian.deorowicz@polsl.pl
*These authors contributed equally to this work.

Abstract

SPLASH is an unsupervised, reference-free, and unifying algorithm that discovers regulated sequence variation through statistical analysis of *k*-mer composition, subsuming many application-specific algorithms. Here, we introduce SPLASH2, a fast, scalable implementation of SPLASH based on an efficient *k*-mer counting approach. The pipeline has minimal installation requirements, and can be executed with a single command. SPLASH2 enables efficient analysis of massive datasets from a wide range of sequencing technologies and biological contexts at unmatched scale and speed, showcased by revealing new biology in rapid analysis of single-cell RNA-sequencing data from human muscle cells, and bulk RNA-seq from the entire Cancer Cell Line Encyclopedia (CCLE) and a study of Amyotrophic Lateral Sclerosis.

Text

Regulated RNA transcript diversification can occur through myriad mechanisms that include alternative splicing, inclusion of sequences absent from the reference genome and RNA editing. Massive RNA-seq datasets provide an unprecedented opportunity to discover regulated transcriptomic complexity in normal cells and its relationship to disease. Despite the promise of computational approaches to identify targetable transcript dysregulation, methods to analyze these datasets for discovering novel transcripts are limited by computational time and algorithmic biases such as in read alignment. Alignment could lead to missing entire families of splicing events such as circRNA (Salzman et al. 2012) and could censor reads representing important mechanisms particularly when studying biological contexts such as tumors that cannot be sufficiently modeled by a reference genome. Moreover, alignment is computationally demanding and presenting storage challenges as the size of the alignment output files can exceed the raw read inputs by a factor of two, hindering the processing of massive datasets. Moreover, this step is only the first step in workflows that aim to infer differential regulation of RNA variation. After alignment, user must choose one of hundreds of time-consuming bioinformatic algorithms to detect each specific signal such as RNA splicing, RNA editing, and V(D)J.

We recently introduced SPLASH (Chaung et al. 2022). SPLASH leverages a unified framework with statistical grounds to analyze *k*-mer composition of reads and detect sample-regulated RNA sequence variation. SPLASH at its core finds constant sequences (anchors) that are followed by a set of diverse sequences (targets) with sample-specific expression distribution and provides valid p-values (Chaung et al. 2022). SPLASH is reference-free, sidestepping the computational challenges associated with alignment and making it faster than other alignment-based methods, enabling discoveries that otherwise would have been impossible. However, SPLASH's original implementation was unable to process large datasets, such as one we profile here, a study of ALS (Ma et al. 2022) with 550GB of raw FASTQ files..

Here, we introduce SPLASH2 (https://github.com/refresh-bio/SPLASH) an ultra-fast, memory-efficient, and user-friendly implementation of SPLASH. The key initial step of SPLASH requires counting kmer tuples of anchors and targets. In the original SPLASH implementation, this is the most time-consuming and memory-intensive step, and is so intensive that SPLASH cannot complete on large corpus'. SPLASH2 resolves this by leveraging KMC ((Deorowicz, Debudaj-Grabysz, and Grabowski 2013; Kokot, Dlugosz, and Deorowicz 2017)), one of the fastest and most memory-efficient *k*-mer counters, capable of processing large-scale genomic datasets via memory optimization and architecture parallelization techniques. This engineering insight is leveraged in SPLASH2 which overcomes the key computational limitations of SPLASH. SPLASH2 is implemented in 3 main steps (Figure 1A). In step 1, SPLASH2 parses the input fastq files and counts anchor-target pairs in each sample separately; when gaps are desired between an anchor and its target sequences, they are subsequently removed as part of the parsing. The output of step 1 for each sample is a list of lexicographically sorted records containing anchor, target, and count. The target sequences and their counts found for each anchor across all samples are then merged in step 2, to build a contingency table for each unique anchor across all samples. The contingency table is used to compute a statistically-valid p-value using an unsupervised optimization (Chaung et al. 2022)(Baharav, Tse, and Salzman, n.d.) (Methods). This step is memory-frugal as the contingency table is represented by a sparse matrix and each contingency table is loaded in memory individually. Finally, multiple testing correction (Benjamini and Yekutieli 2001) is performed in step 3. To enhance the interpretability of SPLASH2's calls, we constructed *extendor* sequences by concatenating anchor-target pairs. To classify each anchor into distinct biologically-relevant categories, we used the alignment information of the extendors corresponding to its top two most abundant targets, analogous to the classification approach for SPLASH (Dehghannasiri et al. 2022).

We illustrate a snapshot of SPLASH2's potential for novel biological discovery using three massive datasets from different sequencing technologies and biological contexts: 1,553 human muscle single cells (3.7 billion reads) profiled with SmartSeq2 (SS2) from (Tabula Sapiens Consortium* et al. 2022), all 671 primary cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al. 2019), and an Amyotrophic Lateral Sclerosis (ALS) dataset (1.8 billion reads; (Ma et al. 2022). We characterized the time and memory requirement for each step of SPLASH2 using SS2 and ALS datasets. Step 2 took more time and memory when analyzing SS2 (many samples but fewer reads per sample), while Step 1 required more resources for the ALS datasets (few samples but deeper sequencing). Step3 accounted for < 10% of the total run time (Figure 1B). As our implementation supports multithreading, it offers decreased run time but at the cost of increased memory consumption (Figure 1C). Memory usage and runtime grow linearly with the input size (Figure 1D), with the exception of RAM usage for ALS (Figure 1E). Also, SPLASH2 was at least 1.7 times faster and consumed 4 times less memory (Figure 1E) compared to STAR (Dobin et al. 2013), which is even more striking given the fact that STAR alignment is only the first and relatively less intensive step in typical custom bioinformatics analyses. SPLASH2 detects cryptic splicing in positive controls *UNC13A*, *STMN2*, and *POLDIP3* experimentally validated to have functional roles in ALS; beyond these examples, SPLASH2 predicted novel cryptic splicing events missed by other methods (Ma et al. 2022). Our analysis revealed that varying different SPLASH's parameters (i.e., anchor, target, and gap length and fraction of training data) have little impact on the run time, RAM usage, and number of significant anchors, with the exception of moderate values of anchor length between 10 and 20 (Suppl. Figure 1).

Our previous analysis of single-cell data was limited by computational implementation in SPLASH (Dehghannasiri et al. 2022), only permitting parallel analysis of up to 400 cells. SPLASH2 empowers simultaneous analysis of all 1,553 muscle cells from donor 2 (TSP2) in Tabula Sapiens dataset (Tabula Sapiens Consortium* et al. 2022), enabling to new biological insights from inference at this scale.

First, we show that SPLASH2 rapidly identifies candidate RNA editing events *de novo*, including detecting potentially hyperedited events, filling a gap in existing bioinformatic tools (Suppl. Figure 2). To identify anchors whose target diversity could potentially be explained by the canonical A-to-I RNA editing, we

selected anchors where the two most abundant targets differ by single-base mismatches, has at least four uniquely mapping extendors to the human reference genome, each with >5% of anchor reads (Methods). The probability of such events under a model of sequencing error is vanishingly small as each cell in the corpus is sampled from a donor and hence a common genetic background.

SPLASH2's analysis uncovers 402 representative anchors from 303 genes, where anchors' targets differ by single-base changes and are classified in some mismatch category (Figure 2A, Suppl. Table 1). The A<->G (or equivalently its reverse complement T<->C) mismatch category had ~10-fold more anchors than other mismatch categories, consistent with canonical A-to-I editing (Nishikura 2016). (Methods, Figure 2A, Suppl. Table 1). : A<->T is the next most prevalent category, potentially explained by other known mechanisms (Uzonyi et al. 2021). Among the anchors that were categorized as A<->G mismatch with >3 mismatch positions in their dominant targets in the highest number of cells, anchors mapping to *AGO2, TDRP* and *PECAM1* were detected in 32 (2.1%), 15 (1%), and 12 (1%) cells, respectively (Figure 2B). Findings in *AGO2* support our previous reports of editing (Dehghannasiri et al. 2022). We found five mismatch positions (including one annotated SNP in the anchor sequence (dbSNP155) classified as G<->A in the 3' UTR of *PECAM1,* a regulator of endothelial junctional integrity (Privratsky and Newman 2014), where in three positions targets diverge consistent with A-to-I editing and cannot speculate about a causative mechanism for the other positions (Figure 2B).

SPLASH2 also enables detection of annotated and novel alternative splicing. Anchors were classified as unannotated alternative splicing if at least one STAR-mapped junction in the two most abundant extendors was not annotated in the human reference genome (Methods). In SS2 muscle cells, SPLASH2 identified 13,786 and 11,584 cell-regulated annotated and unannotated splicing anchors, respectively (Suppl. Table 2). Among unannotated alternative splicing anchors, anchors mapping to *SNHG17*, a noncoding RNA with reported roles in tumorigenesis (Pan et al. 2020), *CAMLG*, a calcium-modulating cyclophilin ligand, and *DIMT1*, an RNA-modifying enzyme with a role in ribosome biogenesis (Pan et al. 2020; Shen et al. 2021) were found in the highest number of cells (120, 118, and 112 cells, respectively) (Suppl. Figure 3).

To further showcase SPLASH2's discovery power, we analyzed unannotated alternative splicing in the entire 671 cell lines (5.7 TB) from primary tumors in Cancer Cell Line Encyclopedia (CCLE), a comprehensive transcriptomic dataset, widely used for functional cancer biology (Ghandi et al. 2019). SPLASH2 took ~47 hours with 50 Gb memory to process the entire corpus, while the STAR alignment alone (absent any custom downstream statistical or bioinformatic processing took) took ~80 hours (with same memory) and required ~4 TB of storage for the BAM files. Previous analyses of CCLE have entailed separate bioinformatic approaches to profile mutations, splicing, or other events. In addition to being fragmented, these alignment-first analyses are expected to suffer from poor sensitivity to unannotated splicing especially in cases where there are mutations in exonic sequence, high rates of mutation, or mapping to repetitive regions that are well-known sources of mapping failure (Castel et al. 2015), all of these events being critical to tumorigenesis and are likely to provide potentially excellent biomarkers and neoantigens (Stanley and Abdel-Wahab 2022; Quesada et al. 2011; Sveen et al. 2016).

Accumulating evidence suggests that a significant number of tumors acquire mutations in the spliceosome, leading to the activation of cryptic alternative splicing programs that contribute to pathology (Liu and Rabadan 2021). Thus, the accurate detection of cryptic splicing in tumors hold paramount importance. This problem is especially difficult because tumors also have a high mutational burden and accumulate structural variants missing from the reference genome, some of which are extensive (Korbel and Campbell 2013). SPLASH2's significant anchors included those mapping to 6,036 genes with evidence of differential alternative splicing (excluding intron retention), including 5,755 genes with at least one unannotated alternative splicing (with >10% of anchor reads). Previous analysis of CCLE determined that *MDM4*, a p53 regulator, had cell-line-specific alternative splicing involving skipping of exon 6 (Ghandi et al. 2019; Dewaele et al. 2016).

SPLASH2 rediscovered this positive control, as well as detecting a novel junction not previously reported. 199 genes with unannotated alternative splicing were cataloged in the COSMIC database (Tate et al. 2019), a database of genes with known roles in tumorigenesis, implying an enrichment of COSMIC genes in SPLASH2 calls (exact binomial p-value = 1.92e-6, Methods). COSMIC genes *CTNNA2*, *NSD2*, *PTEN*, and *COL1A1* had the highest effect size.

PTEN is a classic tumor suppressor (Cristofano, Di Cristofano, and Pandolfi 2000), and its loss has been linked to resistance to T cell immunotherapy (Peng et al. 2016). In addition, alternative splice variants of *PTEN* have been shown to phenocopy *PTEN* loss (Breuksch et al. 2018). One extendor for this gene (in 24 cell lines) represents novel splicing between exon 3 and a cryptic exon in the next intron and the other extendor represents annotated splicing between exons 3 and 4 but with two mismatches: an internal A to G in exon 4 and a T to A mismatch right at the first base of exon 4 (Figure 2C). This establishes SPLASH2's unique statistical power to jointly detect splicing and multiple genome mismatches to provide potentially new markers of *PTEN* loss, and more generally identify regulated splicing and sequence variation missing from the reference gnome. To our knowledge, this is the first report of such splicing in this critical regulatory gene.

We also detected a cancer-type-specific cryptic exon of *RAB36* in lymphoma cell lines (21/22 lymphoma cell lines with expression for the anchor) (Figure 2D). *RAB36* is a member of the RAS oncogene family and previous studies have shown significant association between the overexpression of this gene and poor prognosis in leukemia patients (Wang et al. 2021). Similarly, SPLASH2 detected a splicing specific to leukemia and lymphoma (21 and 15 cell lines, respectively) in *SMIM14*, a small integral membrane protein (Figure 2E). The junction found by SPLASH2 is not part of the T2T CAT+Liftoff annotation but was included in the NCBI refseq. Another example of cancer-type-specific regulation is intron retention in *CLIP3,* with significantly higher intron retention rate in glioma and melanoma compared to other cancer types (Suppl. Figure 4). Downregulation of *CLIP3* has been linked to radiotherapy resistance in glioblastomal stem-like cells (Kang et al. 2021), implying the importance of *CLIP3* variants to the prognosis of glioma patients. SPLASH2 was also able to rediscover *BCR-ABL1* gene fusion, a well-known biomarker and treatment targets for chronic myelogenous leukemia (CML) patients (Melo and Barnes 2007), where the extendor corresponding to the *BCR-ABL1* fusion junctions were the major isoform expressed in leukaemia cell lines (Figure 2F). This finding highlights the unique unifying and ultra-efficient ability of SPLASH2 to jointly detect fusions, cryptic splicing, intron retention and point mutations, all critical transcriptional changes in cancer.

SPLASH2 discoveries extend to other non-canonical transcriptional aberrations missed by STAR. To illustrate this, we considered anchors with high effect size where one of the top two extendors failed STAR mappi ng. This analysis yielded 13,465 anchors with variation in target sequences missed by reference alignment. Sequence expression in *SLC3A2* illustrates this bias of reference alignment: while the second abundant extendor mapped to *SLC3A2*, the most abundant extendor having 75% of total anchor reads and observed in an upper aerodigestive tract carcinoma cell line was unmapped. This extendor reflects a hypermutation in *SLC3A2,* highlighting the biases in current alignment-based methods in identifying one of the signatures of cancer genomes (Figure 2G). Among other genes with unaligned extendors was *HOOK3,* known to be altered in 0.1% of carcinoma cancers (André et al. 2017). The second extendor of this gene was unaligned by STAR and multiple sequence alignment revealed a 6-bp GCTATT insertion (Figure 2H). Together, these results establish SPLASH2 as an ultra-fast and unified approach for unbiased analysis of disease transcriptomes, emphasizing its ability as an alternative to reference-based approaches to unveil novel discoveries that would have otherwise been unattainable.

In summary, SPLASH2 is a highly efficient, general analytic framework for biological discovery. In addition to improving performance of common specialized bioinformatic tools, it discovers novel mutations, transcribed structural variation and post-transcriptional regulation that may be completely missed by alignment-based methods. To date, the scale of this missing variation, which may be extensive in tumors, has been

difficult or impossible to measure in large scale studies. SPLASH2's scale enables a next-generation of massive genomic data analysis, including for essentially any RNA or DNA sequencing protocle.

Acknowledgments:
We thank Aaron Gitler and Yi Zeng for suggestions to analyze the ALS data and useful discussion. J.S. is supported by the National Institute of General Medical Sciences grants R35 GM139517 and the Chan Zuckerberg Data Insights. TZB was funded in part by the Stanford Graduate Fellowship and the NSF GRFP.

Figures:

Figure 1. SPLASH2 pipeline outline and resource consumption. (A) Outline of the SPLASH2 pipeline for a toy data (2 input samples, anchor length 3, gap length 0, target length 3). Fastq files for each sample are parsed in Stage 1 using KMC and KMC-tools to extract anchor-target pairs and their counts. For each anchor, the targets and their counts across samples are merged in Stage 2 to construct a contingency table, which is then used to compute a p-value. Multiple testing correction is performed in Stage 3 and finally anchors with significant p-value are classified into biologically relevant events such as alternative splicing. (B) Real-time memory and run time requirements are shown for SS2 muscle cells (top plot) and ALS dataset (bottom plot). The computational burden of Stages 1 and 2 depends on the number and size of samples, larger deeply sequenced and an increased number of samples result in higher time and memory requirements for Stages 1 and 2, respectively. and edepend on the stepand dataset. (C) The table showing time and memory required for analyzing SS2 muscle cells for different number of threads. Multithreading can speed up computation, but at the cost of higher RAM requirements. (D) Time (top) and memory (bottom) requirement for different number of input SS2 muscle cells. Time and RAM usage increases linearly with thenumber of input samples. (E) Run time (top) and memory requirement (bottom) for different number of input reads (at fixed number of samples) using ALS dataset. Time increases linearly with the number of input reads, while memory saturates at 15GB. SPLASH2 was at least 1.7 times faster and 4 times more memory efficient than STAR.

Figure 2. SPLASH2 enables rapid discovery at scale in scRNA-Seq and cancer cell lines. (A). Bar plot showing the number of anchors detected by SPLASH2 in SS2 muscle cells for each mismatch category and number of mismatches (in total: 512 anchors from 303 genes). Consistent with the canonical RNA editing, G<->A category has the highest number of anchors for each number of mismatches. (B) Among genes with >3 mismatches, *AGO2*, *TDRP*, and *PECAM1* were identified in the largest number of cells. For each gene, we show the multiple sequence alignment ("Multiple Sequence Alignment" 2006) for the reference and the four most abundant extendors. The bar plots show the total read count for each extendor across all cells expressing it. Gene structures are based on CAT/liftoff annotation in T2T human genome reference assembly. Boxes and lines represent exons and introns, respectively, and the number inside each exon shows its order among all exons in the gene. (C) Simultaneous detection of mutations and a novel splice junction for *PTEN* (a known tumor suppressor) in CCLE. The annotated junction harbors two mutations, one internal and one at the splice site (in red) which was expressed in an upper aerodigestive tract carcinoma cell line. The box plots show the read counts for each extendor in each cell line of a specific tumor. (D) SPLASH2 detected a cryptic exon for *RAB36* in the intron between 7th and 8th exons which had higher expression than the canonical splicing in lymphoma cell lines. The pair of box plots show the read counts for each extendor in each cell line of a specific cancer type. The numbers next to each cancer type show the number of expressing cell lines. (E) SPLASH2 detected a novel cryptic splicing for *SMIM14* with higher expression in lymphoma and leukaemia cell lines. (F) SPLASH2 automatically rediscovered *BCR-ABL1* gene fusion via an anchor mapping to gene *ABL1* whose second extendor mapping to the fusion junction. Box plots of the read counts for each extendor show that Extendor 2 is solely expressed in leukaemia cell lines. (G) An anchor from *SLC3A2* had the highest fraction of unaligned reads by STAR which was expressed solely in upper aerodigestive tract carcinoma. Multiple sequence alignment of the reference sequence and a long 101-bp read for extendor 1 indicates that extendor 1

is due to hypermutation in this gene. (H) The second extendor of an anchor in *HOOK3*, which was unaligned by STAR, reflects a 6-bp insertion at the splice site of exon 16.

Suppl. Figure 1. Change SPLASH2 parameters such as anchor, gap, and target length and fraction of data used for training has minimal effect on run time, RAM usage, and the number of detected significant anchors.

Suppl. Figure 2. (A) Diagram showing how canonical RNA editing can lead to sequence diversity. Each extendor sequence corresponds to a specific set of A's in the reference sequence being edited to G in the RNA transcripts. (B) Toy example showing sour approach for identifying anchors corresponding to a specific RNA mismatch category.

Suppl. Figure 3. SPLASH2 analysis revealed widespread unannotated alternative splicing in SS2 muscle cells including in genes (A) *SNHG17*, (B) *DIMT1*, and (C) *CAMLG* that were found in the highest number of cells. For each gene, the unannotated and annotated junctions are shown in red and green, respectively. Bar plots show the total read count for each junction across all cells expressing it.

Suppl. Figure 4. Cancer-type-specific intron retention in *CLIP3,* with significantly higher intron retention rate in glioma and melanoma compared to other cancer types. Extendor 1 maps to the splice junction and extendor 2 reflects the intron retention.

Tables:

Table 1: RNA mismatches in SS2 muscle cells. List of anchors with evidence for RNA mismatches called by SPLASH2 in SS2 muscle cells.

Table 2: CCLE alternative splicing. List of alternative splicing anchors called by SPLASH2 in the CCLE dataset.

METHODS

Code availability

The SPLASH2 pipeline along with detailed instructions and test data are available through a GitHub repository: https://github.com/refresh-bio/SPLASH.

Data availability

The FASTQ files for the SS2 human muscle cells were downloaded from https://tabula-sapiens-portal.ds.czbiohub.org/. The ALS data was downloaded from the SRA database (SRP185789). The FASTQ files for the 671 cell lines in CCLE from primary tumors were downloaded from SRA database (SRP186687).

k-mer counting in SPLASH2

First, SPLASH2 enumerates all (a+g+t)-mers (a, g, t being anchor, gap, and target lengths, respectively). In the second step, these sequences are sorted lexicographically with KMC-tools (Deorowicz, Debudaj-Grabysz, and Grabowski 2013; Kokot, Dlugosz, and Deorowicz 2017). As a result, all occurrences of unique anchors are adjacent, which enables efficient gap removal and unique targets collapsing in the third step via a linear traversal over the (a+g+t)-mers which collects all targets for the current anchor (gaps are skipped). Then, targets are sorted and the unique targets merged and their counts are updated appropriately.

Intermediate files format

The pipeline is implemented as a set of separate applications. For communication, they store and load files in a SATC format. Each SATC file stores a list of records containing: **S**ample id **A**nchor **T**arget **C**ount. A SATC file starts with a header, which defines the number of bytes used to represent each record part, and also some other metadata. Records follow the header. The order of fields is as follows: sample id, anchor, target, count. Conceptually, for each field of a record, a minimal number of bytes is used; however, some additional techniques are applied to reduce the size of these files. After the first step, each SATC file contains a single sample id and is sorted with respect to the anchor, and in a range of the same anchor with respect to targets. Consequently, adjacent records share some number of most significant bytes. At least, the number of bytes used to represent sample id, but in many cases also the next bytes are identical. For this reason, we prepend one additional byte to each record. It contains a number of identical leading bytes with the previous record, then the identical bytes are not stored. During the reading, they are reconstructed based on the previous record. On top of this technique, we also apply zstd, a general-purpose compression algorithm to reduce the file sizes even more.

Merging SATC files

In the second step, for a single process, there are as many input SATC files, as the samples. Depending on the dataset, there may be hundreds of input samples. To merge them efficiently, we use a binary heap. For each unique anchor, we maintain a min-heap of size equal to the number of input samples. Another binary heap is used for filtering out the least frequent targets. This is only performed if the user defined the maximal number of most frequent target, n_t to be used for contingency table construction.

The architecture of the code and parallelization guidance

SPLASH2 is composed of a set of applications written in the C++ programming language accompanied by a single Python script to run them as separate processes. In the first step, we use KMC and KMC-tools which are parallel themselves. All the other binaries are single-threaded. In particular, the binary that merges bins and computes statistics in the second step is sequential. To run SPLASH2 using p cores, it is recommended to configure the number of processes for the second step to p. On the first step the decision should be based on the number of input samples. If the number of input samples is less than 10, it is recommended to run a small number of first step processes and assign more threads to each such process (for example for p=32, use 4 processes, each with 8 threads). In a case of a hundred (or more) input samples, it may be more beneficial to run more processes (for example 16 processes with 2 internal threads). Lower number of processes will usually lead to lower memory consumption.

Statistical methodology

By default, SPLASH2 computes several measures for interpretability and processing of anchors, such as average Hamming distance between target sequences, and target entropy quantifying the probability distribution of the targets. The optimization procedure for the p-value computation identifies latent structure in the matrix by splitting the counts matrix into train and test portions, computing 1-dimensional row and column embeddings (f and c) respectively (original problem is bi-convex so this is performed via alternating maximization). These f and c are then used, with the held-out test data, to yield a statistically valid p-value, and compute an effect size measure, which indicates how well the samples can be separated into two groups. Additional outputs such as an asymptotically-valid p-value are also available.

Anchors filtering

To lower the memory requirements and computation time, several filters are applied in SPLASH2 to remove potential artifacts. The first pair of filters is applied in the first step to discard all the anchors for a given sample where: a) its total count is not greater than 5, or b) the anchor contains a poly(A/C/G/T) string of length at least 8. In the second step, all the anchors fulfilling at least one of the following conditions are removed: a) the sum of counts across all samples and targets is not greater than 50, b) the number of unique targets is not greater

than 1, or c) the number of unique samples is not greater than 1. All the given values are defaults and can be redefined by the user. In some cases, the number of targets for a single anchor may be large, while only the most abundant ones are relevant. SPLASH2 allows keeping only a user-defined number of most frequent targets per anchor (by default all targets are used).

Hardware

Run time and memory consumption was evaluated on a machine equipped with AMD 3995WX 64-Cores CPU, 512 GB RAM, and 4 HDDs configured in RAID5.

Performance evaluation details

For the SS2 muscle cells, KMC was run in RAM-only mode and 16 processes were assigned to the first step (2 threads for each process), while for the ALS dataset, KMC was run in the default disk mode and its RAM limit was set to 6GB, and 2 processes were used in the first step (each with 16 threads). For the second step, we used 32 computing processes.

Increasing gap length slightly decreases all the measured values due to the lower number of existing (a+g+t)-mers: In SS2 muscle cells (1553 FASTQ files of average size 657MB), the most resource-demanding was the second step (Figure 1B), while in the case of few large input samples (i.e., ALS dataset) the first step was more resource demanding (Figure 1B).

To measure the impact of the input data size (both the number of samples and the size of each sample), we ran SPLASH2 on different numbers of SS2 muscle cells (Figure 1D) and also on all ALS samples when subsampled to a given number of reads (Figure 1E). In both cases, run time grows linearly with the input size. In the case of SS2 muscle cells, the memory usage also grows linearly, yet it is negligible compared to the input size. For ALS, the memory grows linearly with the number of input reads but then it saturates at about 300M reads. For measuring the run time and memory required by STAR, it was run with 32 threads, and one sample at a time, which is the same computation power used for SPLASH2 runs.

Splicing anchors

We performed a post facto step to find anchors whose target sequence diversity can be explained by alternative splicing. For each anchor, we first concatenate its sequence by the sequence of each one of its corresponding targets to obtain extendor sequences. We then align extendors to the human reference genome (T2T) using STAR. An anchor is classified as a splicing anchor, if STAR reports a splice junction for at least one of its two extendor sequences corresponding to the two most abundant targets.

RNA mismatch classification

If all the base pair substitutions between the targets of an anchor can be performed by a distinct mapping, we assign a mismatch category to the anchor (Suppl. Figure 2). For example, we assign a category “mismatch G<->A” to an anchor, if all its target sequences are reduced to a unique sequence by converting all G’s in target sequences to A. We assign 6 different categories: G<->A, T<->C, T<->A, C<->A, G<->C, and T<->G. Note that some of these categories are equivalent of each other and can be collapsed when sequencing data is not strand-specific such as G<->A with T<->C and C<->A to T<->G. We report 4 mismatch categories for SS2 muscle cells as SmartSeq2 sequencing is not strand specific.

To compute the probability of observing 4 targets with >5% abundance, under a biallelic mode, we use the following model. In a diploid genome, there are two alleles, and thus observing non-modified or edited RNA where the anchor is adjacent to a bi-allelic SNP should result in observing two targets with a high abundance. To compute the probability of observing 4 targets each with >5% abundance, we bound this probability by the probability that either allele has a second target of >5% abundance. We assume that there is only 1 allele and compute the probability of observing a target with > 5% abundance. There are 27*3 number of different 1 base pair changes from the dominant target which we assume to be an un-modified allele. The first order approximation of the probability of an observed target differing from the dominant allele is that the target has

one mismatch and if sequencing error is 1%, the probability of observing this target is less than 0.01/3. We then compute, with n.trials=27*3 to estimate the probability of observing one target of > 5% abundance of all anchor counts as follows:

```
> sum(rbinom(n.trials,N,(.01/3))/N>.05)/n.trials
```

Where N is the number of observations per anchor. Setting N = 100 to 1000 yields (up to computational precision) a value of 0.

Enrichment analysis of COSMIC genes

To compute the p-value for the enrichment of COSMIC genes in genes with unannotated alternative splicing in CCLE, we counted the number of unique gene names for which SPLASH2 detected significant anchor in CCLE, which was 40,133. Using this as the number of background genes, the null probability of a SPLASH2-called gene being a COSMIC gene is 736/40,133 = 0.0183. Considering this null probability, the exact binomial p-value for the enrichment of COSMIC genes in genes with unannotated alternative splicing is < 2.2e-16. However, to be more conservative we considered 30,000 as the number of background genes reflecting the approximate number of protein coding genes in the human genome giving the null probability of 0.024 and exact binomial p-value of 1.92e-6.

Bibliography

André, Fabrice, Monica Arnedos, Alexander S. Baras, José Baselga, Philippe L. Bedard, Michael F. Berger, Mariska Bierkens, et al. 2017. “AACR Project GENIE: Powering Precision Medicine through an International Consortium.” *Cancer Discovery* 7 (8): 818–31.

Baharav, Tavor, David Tse, and Julia Salzman. n.d. “An Interpretable, Finite Sample Valid Alternative to Pearson’s X2 for Scientific Discovery.” *In Preparation*.

Benjamini, Yoav, and Daniel Yekutieli. 2001. “The Control of the False Discovery Rate in Multiple Testing under Dependency.” *The Annals of Statistics* 29 (4): 1165–88.

Breuksch, Ines, Jonas Welter, Heide-Katharina Bauer, Thorsten Enklaar, Sebastian Frees, Joachim W. Thüroff, Annette Hasenburg, Dirk Prawitt, and Walburgis Brenner. 2018. “In Renal Cell Carcinoma the PTEN Splice Variant PTEN-Δ Shows Similar Function as the Tumor Suppressor PTEN Itself.” *Cell Communication and Signaling: CCS* 16 (1): 35.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” *Genome Biology* 16 (1): 195.

Chaung, Kaitlin, Tavor Z. Baharav, Ivan N. Zheludev, and Julia Salzman. 2022. “A Statistical Reference-Free Genomic Algorithm Subsumes Common Workflows and Enables Novel Discovery.” *bioRxiv : The Preprint Server for Biology*, July. https://doi.org/10.1101/2022.06.24.497555.

Cristofano, Antonio Di, Antonio Di Cristofano, and Pier Paolo Pandolfi. 2000. “The Multiple Roles of PTEN in Tumor Suppression.” *Cell*. https://doi.org/10.1016/s0092-8674(00)80674-1.

Dehghannasiri, Roozbeh, George Henderson, Rob Bierman, Kaitlin Chaung, Tavor Baharav, Peter Wang, and Julia Salzman. 2022. “Unsupervised Reference-Free Inference Reveals Unrecognized Regulated Transcriptomic Complexity in Human Single Cells.” *bioRxiv*. https://doi.org/10.1101/2022.12.06.519414.

Deorowicz, Sebastian, Agnieszka Debudaj-Grabysz, and Szymon Grabowski. 2013. “Disk-Based K-Mer Counting on a PC.” *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-160.

Dewaele, Michael, Tommaso Tabaglio, Karen Willekens, Marco Bezzi, Shun Xie Teo, Diana H. P. Low, Cheryl M. Koh, et al. 2016. “Antisense Oligonucleotide-Mediated MDM4 Exon 6 Skipping Impairs Tumor Growth.” *The Journal of Clinical Investigation* 126 (1): 68–84.

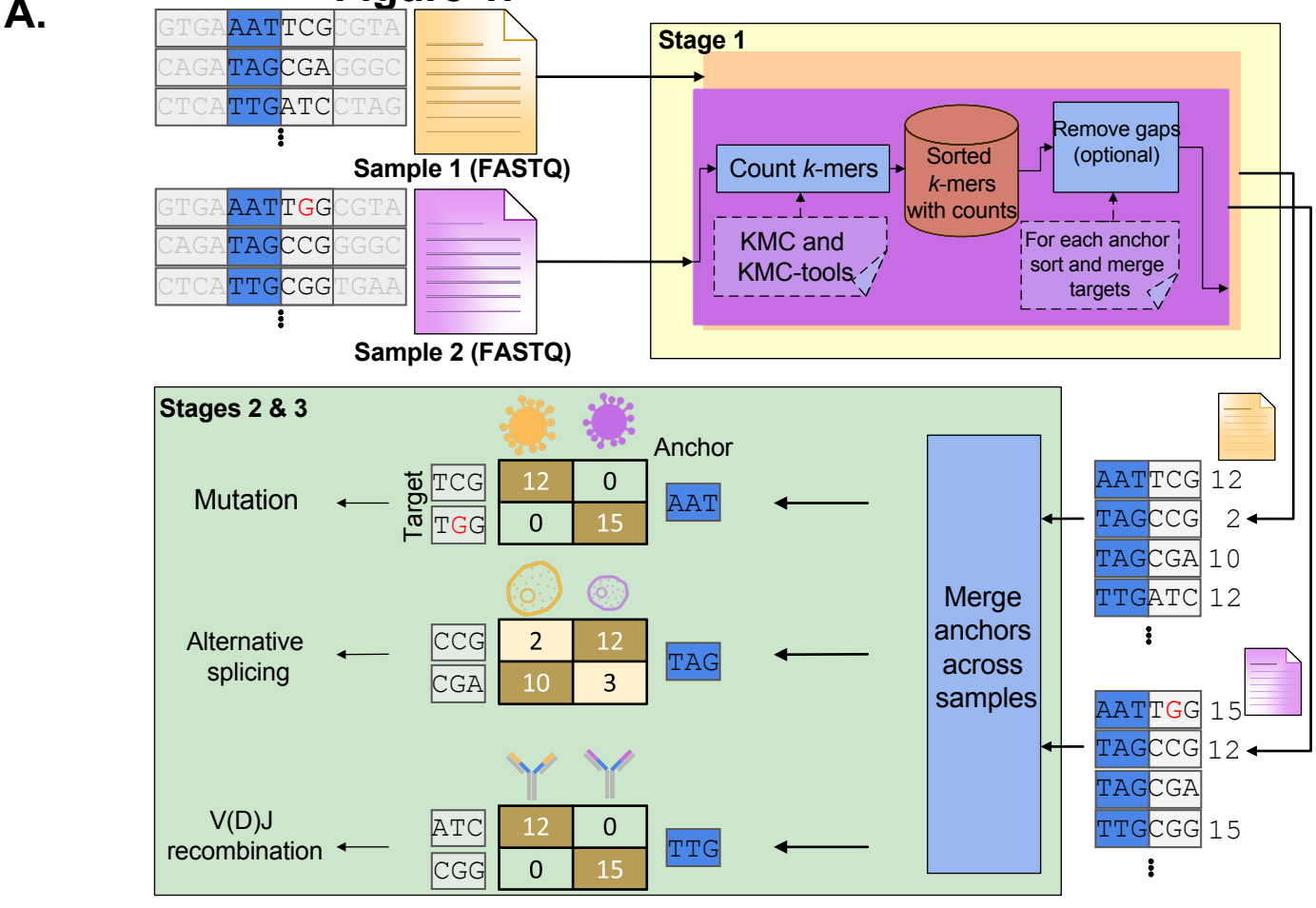
Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

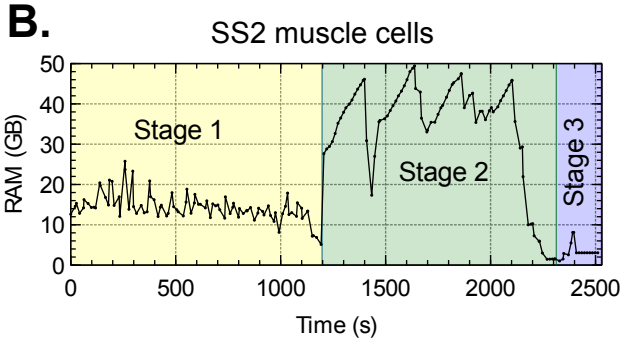
Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald 3rd, Jordi Barretina, et al. 2019. “Next-Generation Characterization of the Cancer Cell Line Encyclopedia.” *Nature* 569 (7757): 503–8.

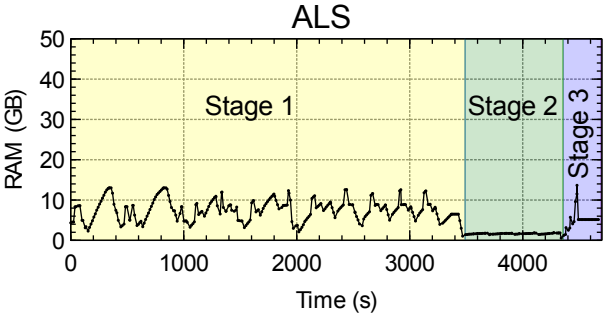
Kang, Hyunkoo, Sungmin Lee, Kyeongmin Kim, Jaewan Jeon, Seok-Gu Kang, Hyesook Youn, Hae Yu Kim, and Buhyun Youn. 2021. “Downregulated CLIP3 Induces Radioresistance by Enhancing Stemness and

Glycolytic Flux in Glioblastoma." *Journal of Experimental & Clinical Cancer Research: CR* 40 (1): 282.
Kokot, Marek, Maciej Dlugosz, and Sebastian Deorowicz. 2017. "KMC 3: Counting and Manipulating K-Mer Statistics." *Bioinformatics* 33 (17): 2759–61.
Korbel, Jan O., and Peter J. Campbell. 2013. "Criteria for Inference of Chromothripsis in Cancer Genomes." *Cell* 152 (6): 1226–36.
Liu, Zhaoqi, and Raul Rabadan. 2021. "Computing the Role of Alternative Splicing in Cancer." *Trends in Cancer Research* 7 (4): 347–58.
Ma, X. Rosa, X. Rosa Ma, Mercedes Prudencio, Yuka Koike, Sarat C. Vatsavayai, Garam Kim, Fred Harbinski, et al. 2022. "TDP-43 Represses Cryptic Exon Inclusion in the FTD–ALS Gene UNC13A." *Nature*. https://doi.org/10.1038/s41586-022-04424-7.
Melo, Junia V., and David J. Barnes. 2007. "Chronic Myeloid Leukaemia as a Model of Disease Evolution in Human Cancer." *Nature Reviews. Cancer* 7 (6): 441–53.
"Multiple Sequence Alignment." 2006. *Current Opinion in Structural Biology* 16 (3): 368–73.
Nishikura, Kazuko. 2016. "A-to-I Editing of Coding and Non-Coding RNAs by ADARs." *Nature Reviews. Molecular Cell Biology* 17 (2): 83–96.
Pan, Xuefeng, Zhiheng Guo, Yanyan Chen, Shu Zheng, Min Peng, Yi Yang, and Zhenpeng Wang. 2020. "STAT3-Induced lncRNA SNHG17 Exerts Oncogenic Effects on Ovarian Cancer through Regulating CDK6." *Molecular Therapy. Nucleic Acids* 22 (August): 38–49.
Peng, Weiyi, Jie Qing Chen, Chengwen Liu, Shruti Malu, Caitlin Creasy, Michael T. Tetzlaff, Chunyu Xu, et al. 2016. "Loss of PTEN Promotes Resistance to T Cell–Mediated Immunotherapy." *Cancer Discovery*. https://doi.org/10.1158/2159-8290.cd-15-0283.
Privratsky, Jamie R., and Peter J. Newman. 2014. "PECAM-1: Regulator of Endothelial Junctional Integrity." *Cell and Tissue Research* 355 (3): 607–19.
Quesada, Víctor, Laura Conde, Neus Villamor, Gonzalo R. Ordóñez, Pedro Jares, Laia Bassaganyas, Andrew J. Ramsay, et al. 2011. "Exome Sequencing Identifies Recurrent Mutations of the Splicing Factor SF3B1 Gene in Chronic Lymphocytic Leukemia." *Nature Genetics* 44 (1): 47–52.
Salzman, Julia, Charles Gawad, Peter Lincoln Wang, Norman Lacayo, and Patrick O. Brown. 2012. "Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types." *PloS One* 7 (2): e30733.
Shen, Hui, Yulia Gonskikh, Julian Stoute, and Kathy Fange Liu. 2021. "Human DIMT1 Generates N26,6A-Dimethylation–containing Small RNAs." *The Journal of Biological Chemistry* 297 (4). https://doi.org/10.1016/j.jbc.2021.101146.
Stanley, Robert F., and Omar Abdel-Wahab. 2022. "Dysregulation and Therapeutic Targeting of RNA Splicing in Cancer." *Nature Cancer*. https://doi.org/10.1038/s43018-022-00384-z.
Sveen, A., S. Kilpinen, A. Ruusulehto, R. A. Lothe, and R. I. Skotheim. 2016. "Aberrant RNA Splicing in Cancer; Expression Changes and Driver Mutations of Splicing Factor Genes." *Oncogene* 35 (19): 2413–27.
Tabula Sapiens Consortium*, Robert C. Jones, Jim Karkanias, Mark A. Krasnow, Angela Oliveira Pisco, Stephen R. Quake, Julia Salzman, et al. 2022. "The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans." *Science* 376 (6594): eabl4896.
Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47.
Uzonyi, Anna, Ronit Nir, Ofir Shliefer, Noam Stern-Ginossar, Yaron Antebi, Yonatan Stelzer, Erez Y. Levanon, and Schraga Schwartz. 2021. "Deciphering the Principles of the RNA Editing Code via Large-Scale Systematic Probing." *Molecular Cell* 81 (11): 2374–87.e3.
Wang, Xiaotao, Jie Xu, Baozhen Zhang, Ye Hou, Fan Song, Huijue Lyu, and Feng Yue. 2021. "Genome-Wide Detection of Enhancer-Hijacking Events from Chromatin Interaction Data in Rearranged Genomes." *Nature Methods* 18 (6): 661–68.

Figure 1.

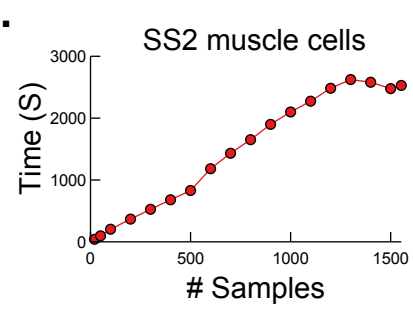


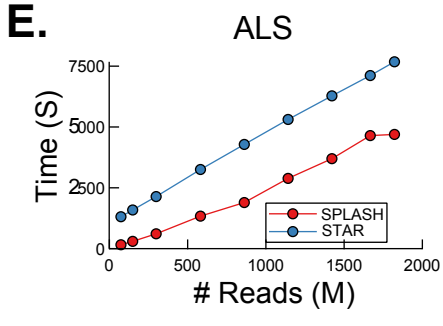


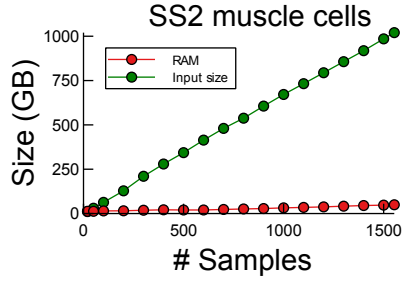


C.

# threads (stage1)	# internal threads (stage1)	# threads (stage2)	Time (s)	RAM (GB)
8	1	8	6,460	16.30
16	1	16	3,535	27.96
16	2	32	2,543	49.92
32	2	64	1,665	96.95







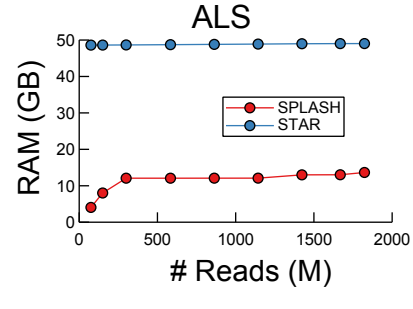
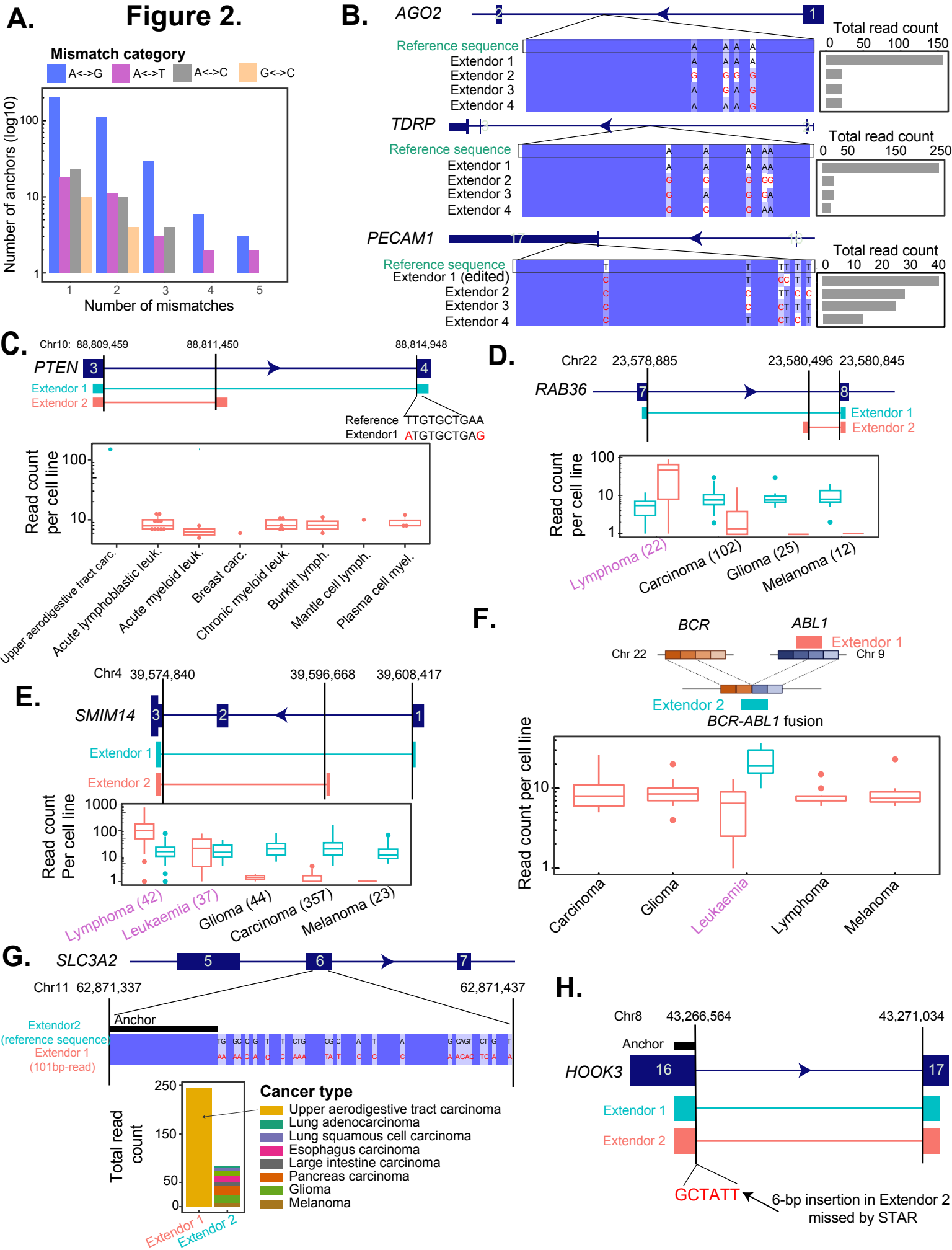
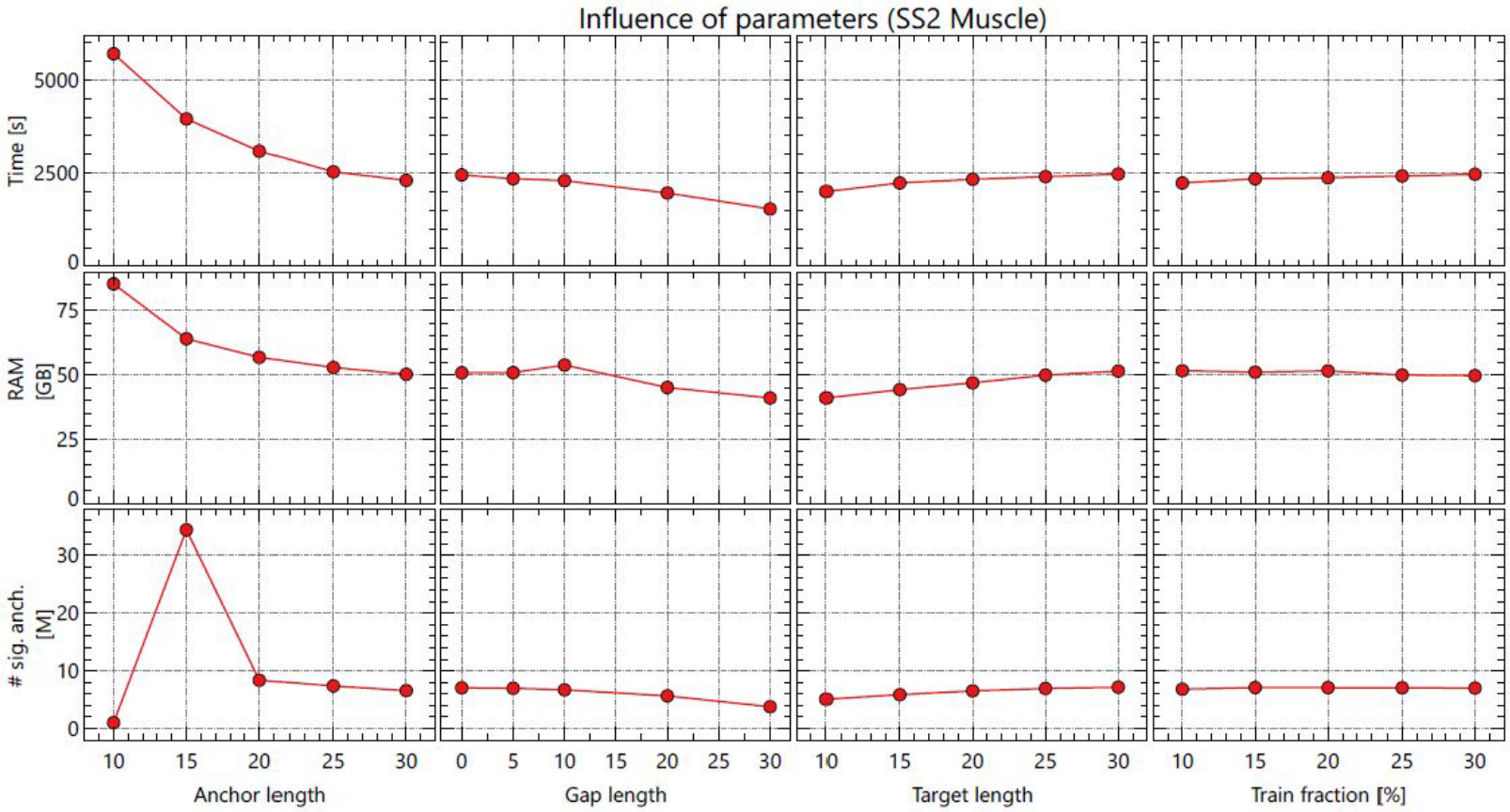


Figure 2.

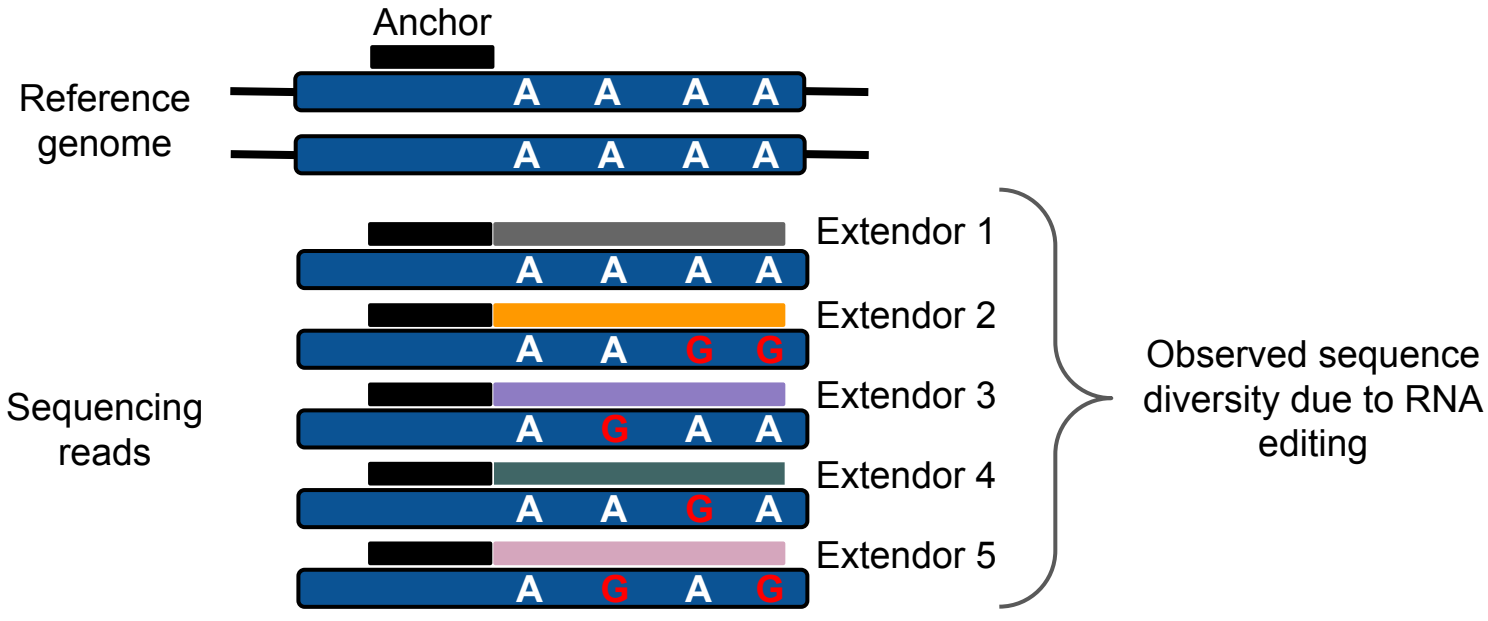

A.
Mismatch category
A<->G
A<->T
A<->C
G<->C
Number of anchors (log10)
Number of mismatches
B.
AGO2
Reference sequence
Extendor 1
Extendor 2
Extendor 3
Extendor 4
Total read count
TDRP
PECAM1
Extendor 1 (edited)
C.
Chr10: 88,809,459
88,811,450
88,814,948
PTEN
Extendor 1
Extendor 2
Reference TTGTGCTGAA
Extendor1 ATGTGCTGAG
Read count per cell line
Upper aerodigestive tract carc.
Acute lymphoblastic leuk.
Acute myeloid leuk.
Breast carc.
Chronic myeloid leuk.
Burkitt lymph.
Mantle cell lymph.
Plasma cell myel.
D.
Chr22
23,578,885
23,580,496
23,580,845
RAB36
Extendor 1
Extendor 2
Read count per cell line
Lymphoma (22)
Carcinoma (102)
Glioma (25)
Melanoma (12)
E.
Chr4
39,574,840
39,596,668
39,608,417
SMIM14
Extendor 1
Extendor 2
Read count Per cell line
Lymphoma (42)
Leukaemia (37)
Glioma (44)
Carcinoma (357)
Melanoma (23)
F.
BCR
ABL1
Chr 22
Chr 9
Extendor 1
Extendor 2
BCR-ABL1 fusion
Read count per cell line
Carcinoma
Glioma
Leukaemia
Lymphoma
Melanoma
G.
SLC3A2
Chr11
62,871,337
62,871,437
Anchor
Extendor2 (reference sequence)
Extendor 1 (101bp-read)
Total read count
Cancer type
Upper aerodigestive tract carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Esophagus carcinoma
Large intestine carcinoma
Pancreas carcinoma
Glioma
Melanoma
Extendor 1
Extendor 2
H.
Chr8
43,266,564
43,271,034
Anchor
HOOK3
Extendor 1
Extendor 2
GCTATT
6-bp insertion in Extendor 2 missed by STAR

Suppl. Figure 1

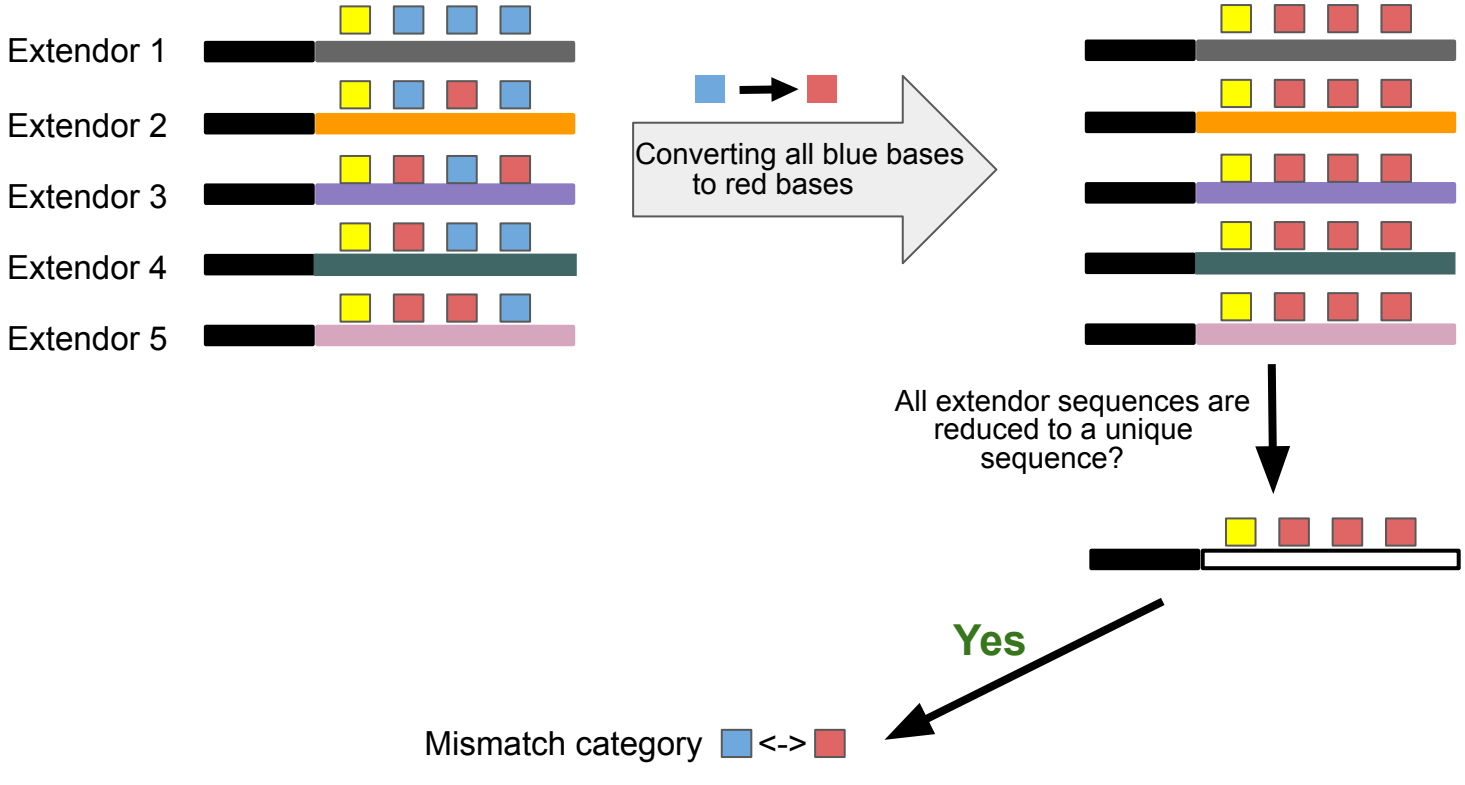


Suppl. Figure 2

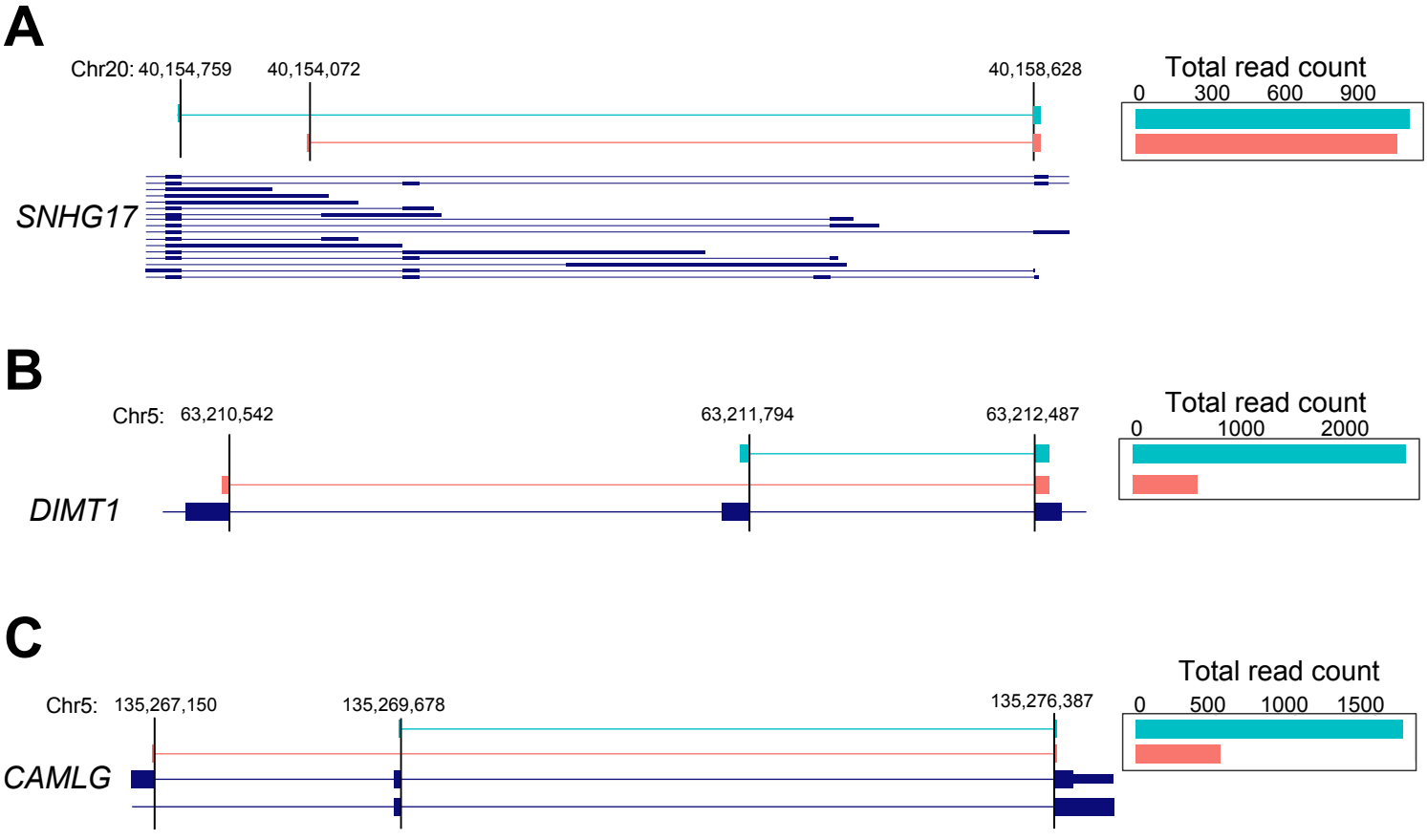
A



B

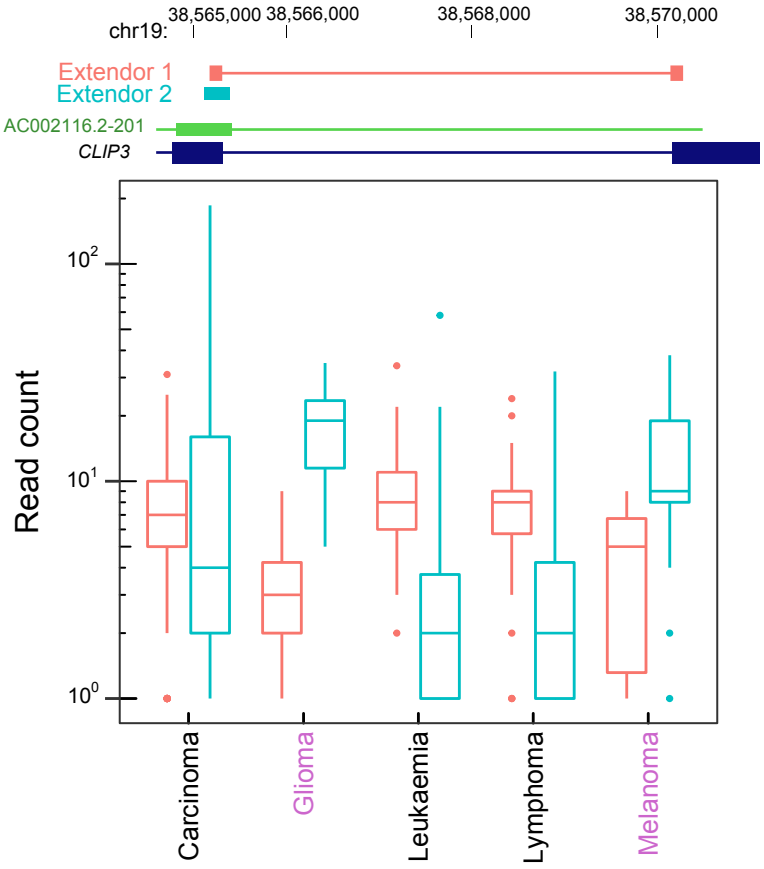
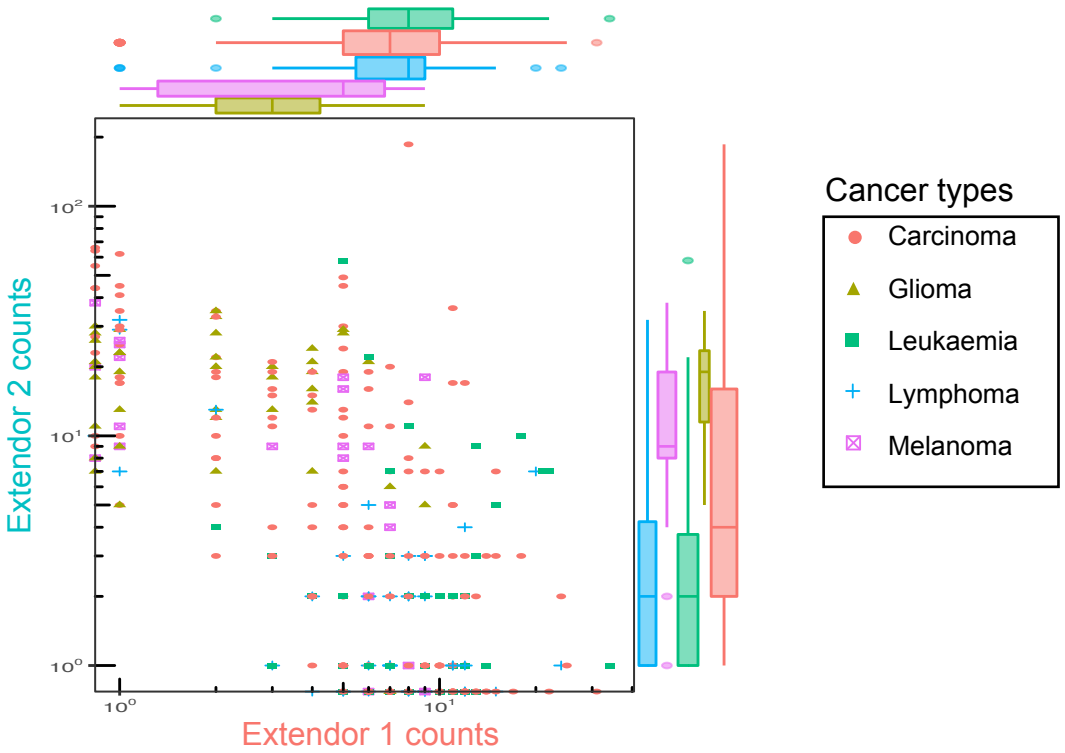


Suppl. Figure 3



Suppl. Figure 4

A



B