NOMAD2 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

NOMAD is a new, unsupervised, reference-free, and unifying algorithm that discovers regulated sequence variation through statistical analysis of *k*-mer composition in DNA or RNA sequencing experiments. It subsumes many application-specific algorithms, from splicing detection to RNA editing to applications in DNA-sequencing and beyond. Here, we introduce NOMAD2, a fast, scalable, and user-friendly implementation of NOMAD based on KMC, an efficient k-mer counting approach. The pipeline has minimal installation requirements, and can be executed with a single command. NOMAD2 enables efficient analysis of massive RNA-Seq datasets where it reveals novel biology, showcased by rapid analysis of 1,553 human muscle cells, the entire Cancer Cell Line Encyclopedia (671 cell lines, 5.7 TB) and a deep RNAseq study of Amyotrophic Lateral Sclerosis (ALS) with ~2 fold less computational resource and time than state of the art alignment methods. NOMAD2 enables reference-free biological discovery at unmatched scale and speed. By bypassing genome alignment, we provide examples of its new insights into RNA expression in normal and disease tissue, to introduce NOMAD2 to enable expansive biological discovery not previously possible.

Text

Regulated RNA transcript diversification can occur through myriad mechanisms that include alternative splicing, inclusion of sequences absent from the reference genome, RNA editing, and gene fusion expression. Massive RNA-seq datasets provide an unprecedented opportunity to discover regulated transcriptomic complexity in normal cells and its relationship to disease. For example, large corpuses like Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al. 2019), large scale single cell sequencing studies (Tabula Sapiens Consortium* et al. 2022) and deep bulk sequencing studies (e.g. in Amyotrophic Lateral Sclerosis, (Ma et al. 2022) provide opportunities for discovering pathogenic RNA. 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 by algorithmic bias.

Today, methods rely on direct alignment of reads to a reference genome (or transcriptome), creating biases and blindspots for biological discovery, including missing entire families of splicing events such as circRNA (Salzman et al. 2012). Alignment is particularly problematic when studying biological contexts that cannot be sufficiently modeled by a reference genome such as in tumors which can harbor complex structural variants. In other diseases such as neurodegeneration, or even in normal tissue, structural variation in a genome (Logsdon, Vollger, and Eichler 2020) can be transcribed into pathogenic RNA in Amyotrophic Lateral Sclerosis (ALS) (Ma et al. 2022; Gitler and Tsuiji 2016). Methods based on alignment to a genome lacking this variation will also miss detecting these sequences. Alignment is also computationally demanding, hindering the processing of massive datasets and presenting storage challenges as the size of the alignment output files (BAM format) can exceed the raw read inputs by a factor of two. Moreover, this step is only the first in

workflows that attempt to infer differential regulation of RNA isoforms or editing, among many other custom bioinformatic searches. After alignment, analysts must choose one of hundreds of custom bioinformatic algorithms to detect specific signals: RNA splicing, RNA editing, and V(D)J recombination all require separate and time consuming pipelines. These custom pipelines limit inference and significantly increase run time. Together, alignment and custom bioinformatics can censor reads that represent each of these biologically important events.

The scale of sequencing data is growing quickly both in sampling depths per experiment and the number of cells and/or conditions profiled in each study. Together, this demands a new approach for genomic discovery that is highly efficient and scalable, requiring minimal computational infrastructure.

To address these conceptual and technical challenges for studying sequence variation in RNA or DNA that is sample-dependent, we recently introduced NOMAD (Chaung et al. 2022). NOMAD leverages the observation that detecting sample-regulated sequence variation, such as alternative splicing, RNA editing, gene fusions, V(D)J, transposable element mobilization, allele-specific splicing, and genetic variation in a population, among many other regulated events can be unified–in theory and in practice. This is achieved with a simple model that analyzes *k*-mer composition of raw sequencing reads (Chaung et al. 2022). NOMAD finds constant sequences (anchors) that are followed by a set of sequences (targets) with sample-specific target variation and provides valid p-values (Chaung et al. 2022). NOMAD is reference-free, sidestepping the computational challenges associated with alignment and making it significantly faster and more efficient than alignment, and enabling discovery and statistical precision not currently available, even from pseudo-alignment (Bray et al. 2016). However, NOMAD's original implementation was unable to process large datasets, e.g. one we profile here, a study of ALS (550GB of raw FASTQ files, Methods).

Here, we introduce NOMAD's successor, NOMAD2 (https://github.com/refresh-bio/NOMAD) and illustrate a snapshot of its novel biological discovery in three massive datasets: 1,553 single cell RNA-seq samples (Tabula Sapiens Consortium* et al. 2022), 671 cancer cell lines (Ghandi et al. 2019), and deeply sequenced patient samples from ALS (Ma et al. 2022). NOMAD2 is implemented in 3 stages (Figure 1A). In stage 1, NOMAD2 records all *k*-mers (substrings of length *k*) containing anchor, target, and counts in each given sample separately with KMC (Deorowicz, Debudaj-Grabysz, and Grabowski 2013; Kokot, Dlugosz, and Deorowicz 2017). KMC and its alternatives, e.g. Jellyfish (Marçais and Kingsford 2011), DSK (Rizk, Lavenier, and Chikhi 2013) do not allow any gaps in the counted substrings, and so do not directly enable the generality of NOMAD's analytic framework. First, NOMAD2 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 (counts are updated appropriately) producing the records we request at this stage. The memory footprint of this stage is low, mainly thanks to the memory-frugal KMC and KMC-tools employed here.

In the second stage, NOMAD2 reads all the lists of records linearly to merge the anchors across samples and outputs a list of all anchors with associated summary statistics such as effect size (Methods). For each anchor the results are merged in a list of tuples (containing target, sample id, and count). Each list is used to construct a contingency table, a base data structure used to compute a statistically valid p-value, along with several measures (e.g. effect size) for downstream interpretability (Methods) (Chaung et al. 2022). This p-value is constructed using unsupervised optimization detailed in (Baharav, Tse, and Salzman, 2023).This step is also memory-frugal (Figure 1C, green area), because only data of a single anchor needs to reside in main memory, and is represented as a sparse matrix.

The third stage loads all anchors with their computed p-values into main memory to perform multiple testing correction, to yield an FDR controlled list of significant anchors (Benjamini and Yekutieli 2001). Memory consumption is low as there is now a single record for each anchor (its p-value) (Figure 1C, purple area). The majority of NOMAD2's code is parallelized (Methods).

Performance of NOMAD2 was evaluated on a machine equipped with AMD 3995WX 64-Cores CPU, 512 GB RAM, and 4 HDDs configured in RAID5 using two datasets ALS (Ma et al. 2022) (15 samples, 1.82 G reads, 550 GB FASTQ), and SS2 Muscle (Tabula Sapiens Consortium* et al. 2022) (1,553 cells, 3.69 G reads, 1.02 TB FASTQ). Experiments revealed that most of the crucial parameters have little impact on the running time/RAM usage/number of significant anchors detected, with the exception of moderate values of anchor length (those between 10 and 20, Figure 1B). Depending on the dataset, stage 1 or stage 2 is the most resource-demanding, where stage 3 requires < 10% of the total running time (Figure 1C). Increasing the number of threads decreases running time at the cost of increased memory consumption (Figure 1D). Memory usage and runtime grow linearly with the input size (Figure 1E), with the exception of RAM usage for ALS (Figure 1F). Comparing NOMAD2 with STAR on the ALS dataset, NOMAD completed computations at least 1.7x faster and consumed 4x less memory (Figure 1F). In all run-time comparisons between NOMAD2 measurements we compared NOMAD2 to the STAR alignment step alone which is typically less intensive than custom downstream bioinformatics analyses which are not required for NOMAD2.

Our previous analysis of single-cell data was limited by computational and statistical implementation in NOMAD (Dehghannasiri et al. 2022) permitting parallel analysis of only 400 cells. However, using NOMAD2, we could jointly analyze 1,553 SmartSeq2 (SS2) muscle cells from donor 2 (TSP2) in Tabula Sapiens data set (Tabula Sapiens Consortium* et al. 2022) and show biological insights enabled by inference at this scale. We ran NOMAD2 with lookahead distance 0 and concatenated anchor-target pairs which we call *extendors*, classifying anchors by the two most abundant targets per anchor as in (Dehghannasiri et al. 2022).

First, we show that NOMAD2 rapidly identifies candidate RNA editing de novo, including detecting potentially hyperedited events, filling a gap in existing bioinformatic tools. To identify anchors whose target diversity could be explained by the canonical A-to-I editing, we classified anchors as "mismatch" defined as cases where the two most abundant targets differ by single-base mismatches, has at least four uniquely mapping extendors to T2T each with >5% of anchor reads (Methods). The probability of such events under a model of sequencing error is vanishingly small (Bray et al. 2016) as each cell in the corpus is sampled from a donor and hence a common genetic background.

NOMAD2's analysis uncovers 512 representative anchors from 303 genes, where anchors' targets differ by single-base changes and are classified in some mismatch category (Figure 2A). We identified anchors where target diversity derives only from A<->G (or equivalently its reverse complement T<->C) changes. Such anchors are consistent with A-to-I editing (Nishikura 2016). Similarly, we assigned anchors whose target diversity can be explained by other base changes (e.g., A<->T, G<->C) to a "mismatch" category (Methods, Figure 2A, Suppl. Table 1). The A<->G category is ~10-fold more than other categories: A<->T is the next most prevalent category, potentially explained by other known mechanisms (Uzonyi et al. 2021). Total representative anchors when the most abundant targets differ by one base pair change are: 265 (A<->G), 22 (A <-> T), 19 (A<->C), and 4 (G<->C).

Anchors detected in the highest number of cells and categorized as A<->G mismatch where dominant targets differ in multiple positions (classified as >3 mismatches) map to *AGO2*, *GNB1*, and *TDRP*, detected in 40 (2.5%), 41 (2.5%), and 29 (1.8%) cells, respectively (Figure 2B). Findings in *AGO2* support our previous reports of editing which, like this analysis, also detected circRNA products from this anchor (Dehghannasiri et al. 2022). All mismatches in extendors assigned to *GNB1* are consistent with RNA editing based on BLAT to the T2T genome (BLAT) however have ambiguous mapping in hg38. The dominant target for *GNB1* had one position and the second, two positions consistent with editing. We also found five mismatch positions (including

one annotated SNP according to dbSNP155) classified as G<->A mismatched, in 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).

NOMAD2 also enables detection of annotated and novel alternative splicing. We used STAR alignment on NOMAD2 extendors to classify unannotated alternative splicing events defined by the junctional coordinates assigned by the two most abundant extendors. Anchors are classified as unannotated if at least one STAR-mapped junction was not annotated in the T2T human reference genome (Methods). NOMAD2 found 13,786 and 11,584 cell-regulated annotated and unannotated respectively in SS2 muscle cells. Anchors classified as unannotated alternative splicing and found in the highest number of cells were in *SNHG17*, a noncoding RNA with reported roles in tumorigenesis (Pan et al. 2020), and predicted to change the protein coding in *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) in 100 and 109 cells, respectively (Figure 2C).

To further showcase NOMAD2's discovery power, we analyzed unannotated alternative splicing in the entire 671 cell lines (5.7TB) from primary tumors in Cancer Cell Line Encyclopedia (CCLE), a comprehensive transcriptomic and phenotypic atlas, widely used for functional cancer biology (Ghandi et al. 2019). NOMAD2 took ~47 hours with 50 Gb memory (compared to ~80 hours with STAR using the same computing resources and memory and ~4TB of storage needed for BAM files). Separate algorithms have been used to profile mutations, splicing, and other events. However, 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). Yet, all of these events are 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 has shown many tumors acquire mutations in the spliceosome, and that activation of cryptic alternative splicing programs contributes to pathology (Liu and Rabadan 2021). Thus, it is of great importance to precisely detect cryptic splicing in tumors. 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). We used NOMAD2 to detect splicing events that were differentially regulated by cell line in the CCLE. NOMAD2 detected 6,697 genes categorized as being called due to differential alternative splicing; 5,755 genes had at least one unannotated alternative splicing (with >10% of reads). Previous analysis of CCLE determined that *MDM4*, a p53 regulator, had cell-line-specific alternative splicing (Ghandi et al. 2019; Dewaele et al. 2016). NOMAD2 recovers this positive control, skipping of exon 6, as well as detecting a novel junction not previously reported (Ghandi et al. 2019; Dewaele et al. 2016).

199 out of 5,755 genes for which NOAMD2 found unannotated alternative splicing in CCLE were cataloged in COSMIC database (Tate et al. 2019), a database of genes with known roles in tumorigenesis, implying an enrichment of COSMIC genes in NOMAD2 calls (exact binomial p-value = 1.92e-6, Methods). COSMIC genes *CTNNA2*, *NSD2*, *PTEN*, and *COL1A1* were the genes with the highest effect size among NOMAD calls with unannotated alternative splicing. *PTEN* is a classical tumor suppressor (Cristofano, Di Cristofano, and Pandolfi 2000). Loss of *PTEN* 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 represents novel splicing between exon 3 and a cryptic exon in the next intron (24 cell lines) and the other extendor represents annotated splicing between exon 3 and 4 but with two mismatches, one internal A to G in exon 4 and a T to A base pair change at the first base of exon 4 (Figure 2D). This establishes NOMAD'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.

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 2E). *RAB36* is a member of the RAS oncogene family and previous studies have shown significant association between the overexpression of this gene and poorer prognosis in leukemia (Wang et al. 2021). Similarly, NOMAD2 detected a splicing specific to leukemia and lymphoma (21 and 15 cell lines, respectively) in *SMIM14*, a small integral membrane protein (Figure 2F). We should note that this junction is not part of the T2T CAT+Liftoff annotation but was included in the NCBI refseq.

NOMAD2 discoveries extend to non-canonical transcriptional aberrations missed by STAR. To illustrate this, we considered anchors with high effect size where one extendor failed to be mapped by STAR but the other extendor could be assigned to a gene. We found an anchor in which the unmapped extendor had 75% of total anchor reads and the other extendor mapped to *SLC3A2* (Figure 2G). BLAT shows high expression of a sequence from *PDE1C* inserted into exon 5 *SLC3A2* in one upper aerodigestive tract carcinoma cell line (Figurew 2G). The *PDE1C* enzyme is thought to influence pathological vascular remodeling by regulating the stability of growth factor receptors (Beavo 1995). *SLC3A2* has previously been reported in a recurrent gene fusion in primary tumors (Shin et al. 2016), though the event detected we detected in CCLE is to our knowledge novel.

Finally, we note that NOMAD could not complete on the ALS data which profiles the effect of loss of the critical RNA binding protein TDP-43 from the nucleus in ALS patients (Ma et al. 2022). NOMAD2 extendors detect cryptic splicing in positive controls UNC13A, STMN2, and POLDIP3 recently experimentally validated to drive progression of the disease (Ma et al. 2022),. Beyond these examples, NOMAD2 predicted novel cryptic splicing events missed by other methods (Suppl. Table 2).

In summary, NOMAD is a highly efficient, general analytic framework for biological discovery. In addition to improving performance of common specialized bioinformatic tools (Dehghannasiri et al. 2022), 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 has been difficult or impossible to measure in large scale studies, and may be extensive in tumors. NOMAD2's scale enables a next-generation of massive genomic data analysis, including for essentially any RNA or DNA sequencing project.

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. NOMAD2 pipeline outline and resource consumption. (A) Outline of the NOMAD2 pipeline for toy data (2 input samples, anchor length 3, gap length 0, target length 3. (B) Change in anchor, gap, target length has a small impact on computation time, RAM usage, and the number of detected significant anchors. (C) Memory requirements depend on the stage and dataset. (D) Increase in the number of threads speeds up computation, but also increases RAM requirements. (E) Time and RAM usage increases linearly with an increasing number of input samples. (F) Run time increases linearly with an increasing number of input reads for a given number of samples, yet memory saturates and does not exceed 15GB.

Figure2. NOMAD2 enables rapid discovery at scale in scRNA-Seq and cancer cell lines. (A). Bar plot showing the number of anchors detected by NOMAD2 in SS2 muscle scRNA-Seq 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 found in the highest number of cells. For each gene, we show the multiple sequence alignment for the reference and the four most abundant extendors. The bar plots show the total read count for each extendor. Black arrows show the position of mismatches for each gene. (C) NOMAD2 identified pervasive unannotated alternative splicing in SS2 muscle cells. Genes *SNHG17*, *DIMT1*, and *CAMLG* were the genes with unannotated alternative splicing (involving one annotated (in green) and one unannotated junction (in red) found in the highest number of cells. Bar plots show the total read count for each junction. (D) 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 splice site) and was expressed specifically in upper aerodigestive carcinoma. (E) NOMAD2 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. (F) NOMAD2 detected a novel cryptic splicing for *SMIM14* with cancer-type-specific expression in lymphoma and leukaemia cell lines. (G) The anchor from *SLC3A2* had the highest fraction of unaligned reads by STAR (for Extendor1) which was expressed solely in upper aerodigestive carcinoma and includes exonic sequence from *SLC3A2* fused to *PDE1C*.

METHODS

Code availability

The NOMAD2 pipeline along with detailed instructions and test data are available through a GitHub repository: https://github.com/refresh-bio/NOMAD.

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).

Anchors filtering

To lower the memory requirements and computation time, several filters are applied in NOMAD2 to remove potential artifacts. The first pair of filters is applied in the first stage 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 stage, 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. NOMAD2 allows keeping only a user-defined number of most frequent targets per anchor (by default all targets are used).

Intermediate files format: SATC

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. 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 stage, each SATC file contains a single

sample id and is sorted w.r.t. anchor, and in a range of the same anchor w.r.t. targets. In consequence, 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.

Performance evaluation details

For the SS2 Muscle, KMC was run in RAM-only mode and 16 processes were assigned to the first stage (2 threads for each process), while for the ALS KMC was run in the default, disk mode, its RAM limit was set to 6GB, and 2 processes were used in the first stage (each with 16 threads). For the second stage, 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 (>1553 FASTQ files of average size 657MB), the most resource-demanding is the second stage, in the case of over a dozen large input samples (ALS) the first stage is roughly 74 % of the total computing time.

To measure the impact of the input data size we ran NOMAD2 on a number of subsets of SS2 Muscle samples (Figure 1E) and also on all samples of ALS subsampled to a given number of reads (Figure 1F). In both cases, running time grows linearly with the input size. In the case of SS2 Muscle, the memory usage also grows linearly, yet it is negligible in comparison with the input size. In the case of ALS, the memory grows with the number of input samples up to about 300 M reads and then it saturates. STAR was run with 32 threads, one sample at a time.

Merging SATC files

In the second stage, 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

NOMAD2 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 stage, 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 stage is sequential. To run NOMAD2 using p cores it is recommended to configure the number of processes for the second stage to p. On the first stage the decision should be based on the number of input samples. If there are no more than tens input samples, it is recommended to run a small number of first stage 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, NOMAD2 computes several measures for interpretability and processing of anchors, such as average target Hamming distance, average target edit distance, and target entropy. 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.

Splicing anchors

We performed a post facto step to find anchors whose target sequence diversity is due to alternative splicing. For each pair of anchor and target sequence, we first concatenate their sequences and then align it to the reference genome using STAR. An anchor is classified as a splicing anchor, if STAR reports a splice junction for at least one of the sequences corresponding to its top 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. For example, we assign a category "mismatch G<->A" to an anchor, if all its targets' sequences reduce to the same sequence when all G's in target sequences are changed 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 reverse complements of each other such as G<->A with T<->C and C<->A to T<->G, that can be collapsed when data is not strand specific. Hence, we report 4 mismatch categories in this paper as data we used is not strand specific.

To compute the probability of observing 4 targets >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 observations per anchor eg counts. 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 with detected significant anchor in CCLE, which was 40,133. Using this as the number of background genes, the null probability of a NOMAD-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 which results in null probability of 0.024 and exact binomial p-value of 1.92e-6.

Bibliography

Baharav, Tavor, David Tse, and Julia Salzman. 2023. "An Interpretable, Finite Sample Valid Alternative to Pearson's X2 for Scientific Discovery." *bioRxiv*.

Beavo, J. A. 1995. "Cyclic Nucleotide Phosphodiesterases: Functional Implications of Multiple Isoforms." *Physiological Reviews* 75 (4): 725–48.

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.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.
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.
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.
Gitler, Aaron D., and Hitomi Tsuiji. 2016. “There Has Been an Awakening: Emerging Mechanisms of C9orf72 Mutations in FTD/ALS.” *Brain Research* 1647 (September): 19–29.
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.
Logsdon, Glennis A., Mitchell R. Vollger, and Evan E. Eichler. 2020. “Long-Read Human Genome Sequencing and Its Applications.” *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-0236-x.
Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27 (6): 764–70.
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.
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.
Rizk, Guillaume, Dominique Lavenier, and Rayan Chikhi. 2013. “DSK: K-Mer Counting with Very Low Memory Usage.” *Bioinformatics* 29 (5): 652–53.
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.

Shin, Dong Hoon, Donghoon Lee, Dong Wan Hong, Seung Hyun Hong, Jung-Ah Hwang, Byung Il Lee, Hye Jin You, et al. 2016. "Oncogenic Function and Clinical Implications of SLC3A2-NRG1 Fusion in Invasive Mucinous Adenocarcinoma of the Lung." *Oncotarget* 7 (43): 69450–65.

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.

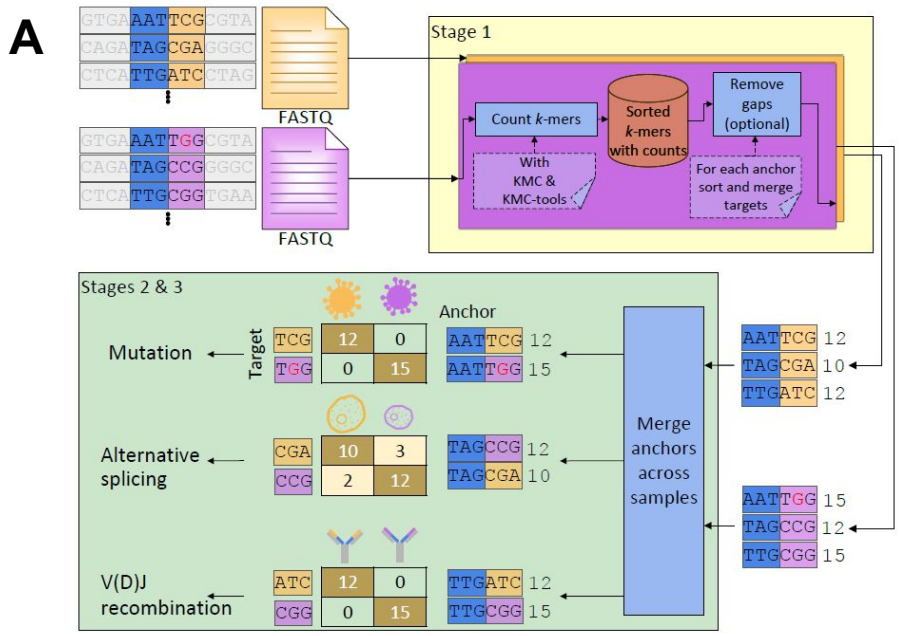
Tabula Sapiens Consortium*, Robert C. Jones, Jim Karkanias, Mark A. Krasnow, Angela Oliveira Pisco, Stephen R. Quake, Julia Salzman, et al. 2022. "The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans." *Science* 376 (6594): eabl4896.

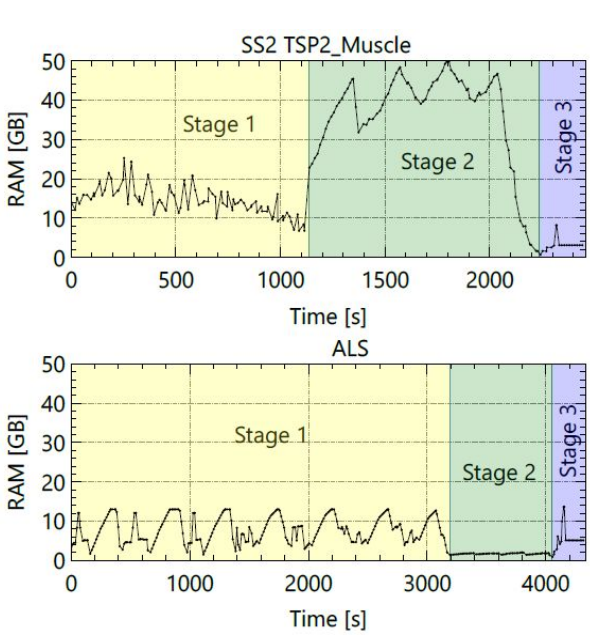
Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47.

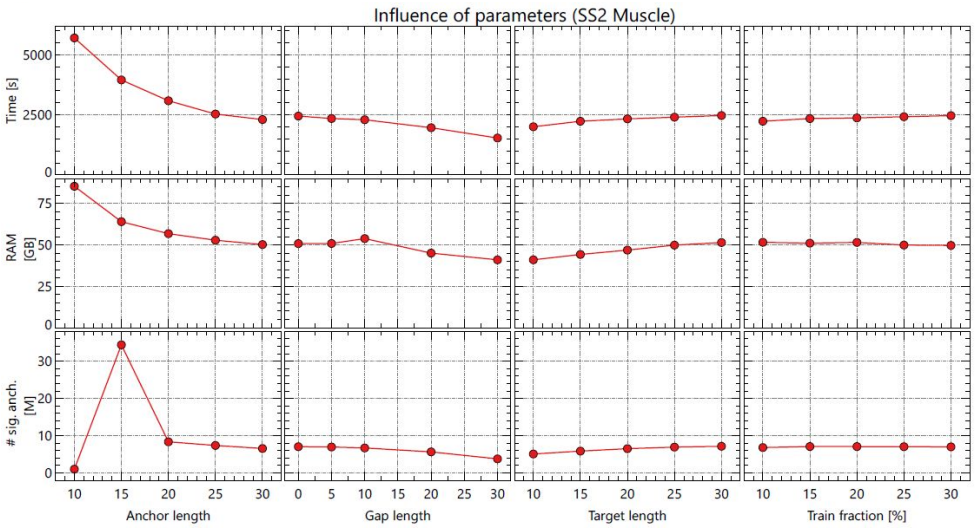
Uzonyi, Anna, Ronit Nir, Ofir Shliefer, Noam Stern-Ginossar, Yaron Antebi, Yonatan Stelzer, Erez Y. Levanon, and Schraga Schwartz. 2021. "Deciphering the Principles of the RNA Editing Code via Large-Scale Systematic Probing." *Molecular Cell* 81 (11): 2374–87.e3.

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

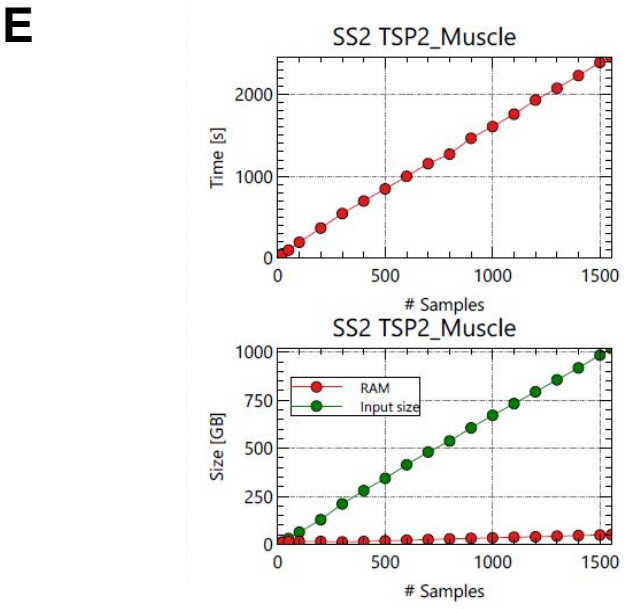






D

Time [s]				
stage 1 threads	stage 1 internal threads	stage 2 threads	Time [s]	RAM [GB]
8	1	8	6,568	17.17
16	1	16	3,603	28.55
16	2	32	2,447	49.91
32	2	64	1,637	100.36



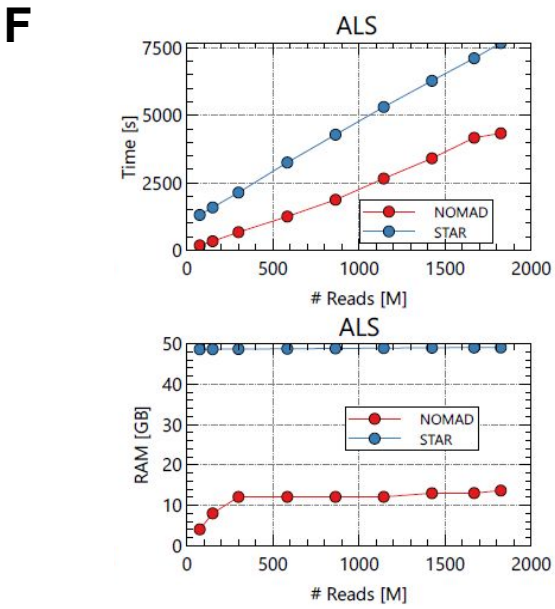
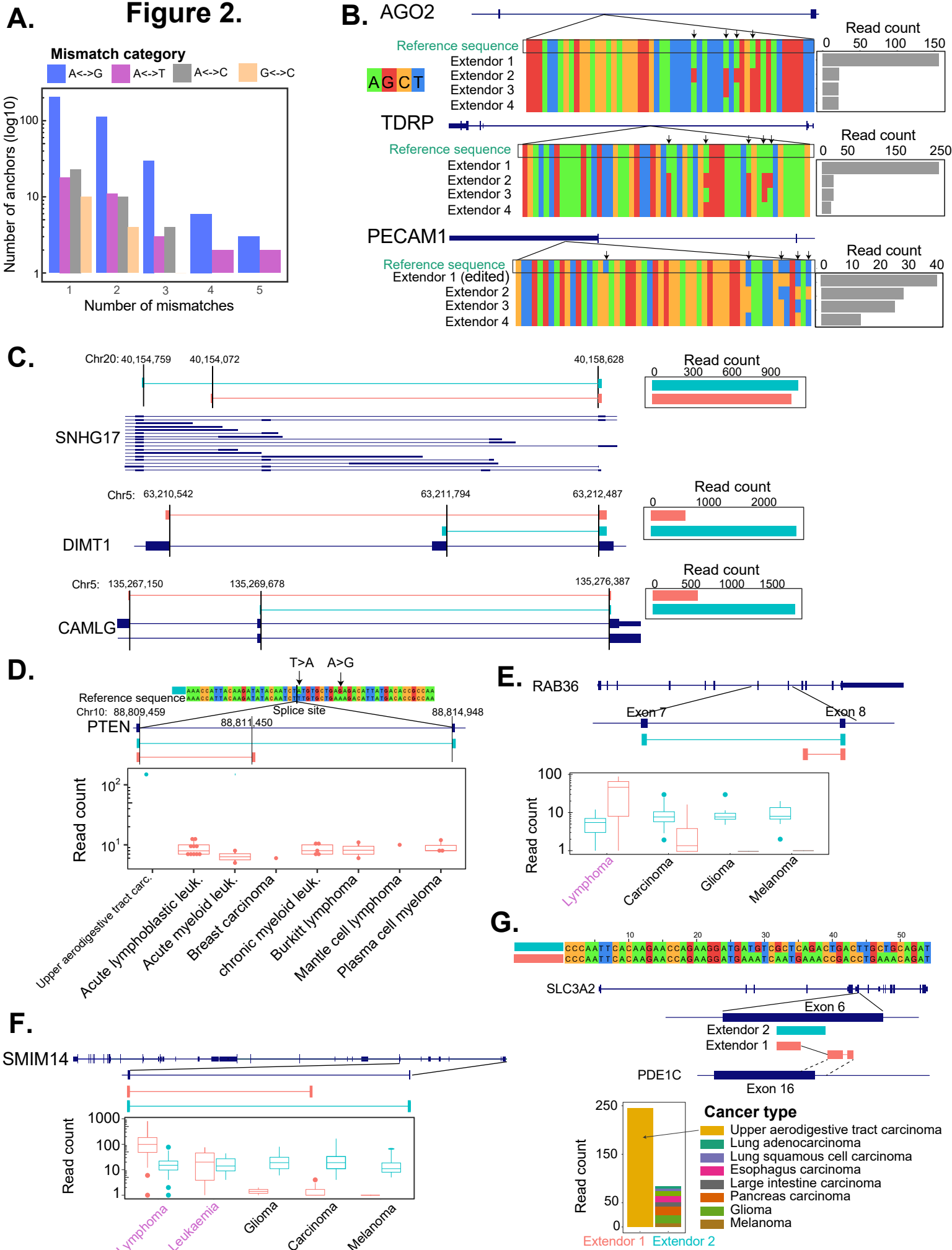


Figure 2.


A.
Mismatch category
A<->G
A<->T
A<->C
G<->C
Number of anchors (log10)
100
10
1
1
2
3
4
5
Number of mismatches
B.
AGO2
Reference sequence
Extendor 1
Extendor 2
Extendor 3
Extendor 4
A G C T
Read count
0
50
100
150
TDRP
Reference sequence
Extendor 1
Extendor 2
Extendor 3
Extendor 4
Read count
0
50
150
250
PECAM1
Reference sequence
Extendor 1 (edited)
Extendor 2
Extendor 3
Extendor 4
Read count
0
10
20
30
40
C.
Chr20: 40,154,759
40,154,072
40,158,628
SNHG17
Read count
0
300
600
900
Chr5: 63,210,542
63,211,794
63,212,487
DIMT1
Read count
0
1000
2000
Chr5: 135,267,150
135,269,678
135,276,387
CAMLG
Read count
0
500
1000
1500
D.
T>A
A>G
Reference sequence
Chr10: 88,809,459
Splice site
88,814,948
PTEN
88,811,450
Read count
10^2
10^1
Upper aerodigestive tract carc.
Acute lymphoblastic leuk.
Acute myeloid leuk.
Breast carcinoma
chronic myeloid leuk.
Burkitt lymphoma
Mantle cell lymphoma
Plasma cell myeloma
E.
RAB36
Exon 7
Exon 8
Read count
100
10
1
Lymphoma
Carcinoma
Glioma
Melanoma
G.
10
20
30
40
50
CCCAATTCACAAGAACCAGAAGGATGATGTCGCTCAGACTGACTTGCTGCAGAT
CCCAATTCACAAGAACCAGAAGGATGAAATCAATGAAACCGACCTGAAACAGAT
SLC3A2
Exon 6
Extendor 2
Extendor 1
PDE1C
Exon 16
Read count
250
150
50
0
Extendor 1
Extendor 2
Cancer type
Upper aerodigestive tract carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Esophagus carcinoma
Large intestine carcinoma
Pancreas carcinoma
Glioma
Melanoma
F.
SMIM14
Read count
1000
100
10
1
Lymphoma
Leukaemia
Glioma
Carcinoma
Melanoma