Supplementary Materials for

A statistical reference-free genomic algorithm subsumes common workflows and enables novel discovery

Kaitlin Chaung[1]†, Tavor Z. Baharav[2]†, Ivan N. Zheludev[3], Julia Salzman[1,3,4,*]

Correspondence to: julia.salzman@stanford.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S4
Captions for Data S1 to S4

Other Supplementary Materials for this manuscript include the following:

Data S1 to S4

1. Protein domain analysis
2. Significant anchors
3. Additional summary tables
4. Anchor genome annotations

Materials and Methods

Anchor preprocessing

Following notation in (*13*), anchors and targets are defined as contiguous subsequences of length *k* positioned at a distance *R*=max(0, (*L* - 2 * *k*) / 2) apart (rounded), where *L* is the length of the first read processed in the dataset. If *L*=100 and *k*=27, then *R*=23. Anchor sequences can be extracted as adjacent, disjoint sequences or as tiled sequences that begin at a fixed step size. For this manuscript NOMAD was run with 4M reads per FASTQ file, anchor sequences tiled by 5bp, and *k*=27. To satisfy the independence assumption for computing p-values in the NOMAD statistics, only one read is used if the sequencing data is paired end; for this manuscript, we use read 1. Extracted anchor and target sequences are then counted for each sample with the UNIX command, `sort | uniq -c`, and anchor-target counts are then collected across all samples for restratification by the anchor sequence. This stratification step allows for user control over parallelization. To reduce the number of hypotheses tested and required to correct for, we proceed with p-value calculation only for anchors with more than 50 total counts across all samples. We further discard anchors that have only one unique target, anchors that appear in only 1 sample, and (anchor, sample) pairs that have fewer than 6 counts. Finally, we retain only anchors having more than 30 total counts after above thresholds were applied removals. This approach efficiently constructs sample by target counts matrices for each anchor. We note that for a fixed number of anchor-target pairs, under alternatives such as differential exon skipping, NOMAD analysis for larger choices of R have provably higher power than smaller choices, following the style of analysis in (*39*).

NOMAD p-values

While contingency tables have been widely analyzed in the statistics community (*40–42*), to our knowledge no existing tests provide closed form, finite-sample valid statistical inference with desired power for the application at hand (Supplement). We develop a test statistic S that has power to detect sample-dependent sequence diversity and is designed to have low power when there are a few outlying samples with low counts as follows. First, we randomly construct a function f, which maps each target independently to {0,1}. We then compute the mean value of targets with respect to this function. Next, we compute the mean within each sample of this function. Then, we construct our anchor-sample score for sample j, S_j, as a scaled version of the difference between these two. Finally, we construct our test statistic S as the weighted sum of these S_j, with weights c_j (which denote class-identity in the two-group case with metadata and are chosen randomly without metadata, see below). In the below equations, $D_{j,k}$ denotes the sequence of the k-th target observed for the j-th sample.

$$\hat{\mu} = \frac{1}{M}\sum_{j,k} f(D_{j,k})$$

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} f(D_{j,k})$$

$$S_j = \sqrt{n_j}(\hat{\mu}_j - \hat{\mu})$$

$$S = \sum_{j=1}^{p} c_j S_j$$

This allows us to construct statistically valid p-values as:

$$P = 2\exp\left(-\frac{2(1-a)^2 S^2}{\sum_{j:n_j>0} c_j^2}\right) + 2\exp\left(-\frac{2a^2 M S^2}{\left(\sum_j c_j\sqrt{n_j}\right)^2}\right) \quad \text{with} \quad a = \left(1 + \sqrt{\frac{M\sum_j c_j^2}{\left(\sum_j c_j\sqrt{n_j}\right)^2}}\right)^{-1}$$

by applying Hoeffding's inequality on these sums of independent random variables (under the null). The derivation is detailed in the Supplement.

This statistic is computed for *K* different random choices of f, and in the case where sample group metadata is not available, jointly for each of the *L* random choices of c. We call the random choice of c_j's "random c's" below. The choice of f and c that minimize the p-value are reported, and are used for computing the p-value of this anchor. To yield valid p-values we apply Bonferroni correction over the *L*K* multiple hypotheses tested (just *K* when sample metadata is used and randomization on c is not performed). Then, to determine the significant anchors, we apply BY correction (BH with positive dependence) to the list of p-values for each anchor, yielding valid FDR controlled q-values reported throughout the manuscript (*43*).

$$Q_{(i)}^{\mathrm{BY}} = \min\left(\min_{j\geq i}\frac{m(\log m + 1)p_{(j)}}{j}, 1\right)$$

NOMAD Effect size

NOMAD provides a measure of effect size when the c_j's used are +/- 1, to allow for prioritization of anchors with large inter-sample differences in target distributions. Effect size is calculated based on the split c and function f that yield the most significant NOMAD p-value. Fixing these, the effect size is the absolute value of the difference between the mean function value over targets (with respect to f) across those samples with $c_j = +1$ denoted A_+, and the mean over targets (with respect to f) across those samples with $c_j = -1$ denoted A_-.

$$\left|\frac{1}{\sum_{j\in A_+} n_j}\sum_{j\in A_+} n_j\hat{\mu}_j - \frac{1}{\sum_{j\in A_-} n_j}\sum_{j\in A_-} n_j\hat{\mu}_j\right|$$

This effect size has natural relations to a simple 2 group alternative hypothesis. It can also be shown to relate to the total variation distance between the empirical target distributions of the two groups. These connections are discussed further in the Supplement.

Consensus sequences

A consensus sequence is built for each significant anchor for the sequence downstream of the anchor sample. A separate consensus is built for each sample by aggregating all reads from this sample that contain the given anchor. Then, NOMAD constructs the consensus as the plurality vote of all these reads; concretely, the consensus at basepair i is the plurality vote of all reads that contain the anchor, i basepairs after the anchor appears in the read (a read does not vote for consensus base i if it has terminated within i basepairs after the anchor appeared). The consensus base as well as the fraction agreement with this base among the reads is recorded.

The consensus sequences can be used for both splice site discovery and other applications, such as identifying point mutations and highly diversifying sequences, e.g. V(D)J rearrangements. The statistical properties of consensus building make it an appealing candidate for use in short read sequencing (*44*), and may have information theoretic justification in *de novo* assembly (*15*) (Supplement).

To provide intuition regarding the error correcting capabilities of the consensus, consider a simple probabilistic model where our reads from a sample all come from the same underlying sequence. In this case, under the substitution only error model, we have that the probability that our consensus for n reads makes a mistake at a given location i under independent sequencing error rate epsilon (substitution only) is at most

$$\mathbb{P}(\text{error at basepair i}) \leq \sum_{k \geq n/2}^{n} \binom{n}{k} \epsilon^k (1-\epsilon)^{n-k} \leq \frac{n}{2} \binom{n}{n/2} \epsilon^{n/2}$$

We can see that even for n=10, this probability is less than 1.3E-7 for a given basepair, which can be union-bounded over the length of the consensus to yield a vanishingly small probability of error. Thus, for a properly aligned read, if a basepair differs between the consensus and reference it is almost certainly a SNP.

Element annotations

To identify false positive sequences or contextualize mobile genetic elements, anchors and targets are aligned with bowtie2 to a set of indices, corresponding to databases of sequencing artifacts, transposable elements, and mobile genetic elements. In these alignments, using bowtie2, the best hit is reported, relative to an order of priority (*13*). The reference used, in order of priority, are: UniVec, Illumina adapters, Escherichia phage phiX174, Rfam (*45*), Dfam (*46*), TnCentral (*47*), ACLAME (*48*), ICEberg (*49*), CRISPR direct repeats (*50*), ITSoneDB (*51*), ITS2 (*52*), WBcel235, TAIR10, grch38_1kgmaj. To perform these annotations, bowtie2 indices were built from the respective reference fastas, using bowtie2-build with default parameters.

Anchors and targets were then aligned to each index, using bowtie2-align with default parameters. For each sequence, we report the alignment to the reference and the position of that alignment for each reference in the prespecified set. Anchors and targets, and their respective element annotations, are reported in the element annotation summary files.

Genome annotations

Anchor, target, and consensus sequences can be aligned to reference genomes and transcriptomes, to provide information about the location of sequences relative to genomic elements.

For significant anchors, target, and consensus sequences, we report information regarding the anchor, target, and consensus sequences' alignments to both a reference genome and transcriptome in the genome annotation summary files (table S4). All alignments reported below are run in two modes in parallel: bowtie2 end-to-end mode (the bowtie2 default parameters) and bowtie2 local mode (`-local`, in addition to the bowtie2 default parameters); the following columns are prefixed with "end_to_end" or "local", for end-to-end mode and local mode, respectively.

To report alignments to the transcriptome, the sequences are aligned to the reference transcriptome with bowtie2, with `-k 1`, in addition to the above parameters, to report a maximum of one alignment per sequence. If there is a transcriptome alignment, we report the alignment to the reference and the MAPQ score of the alignment.

To report alignments to the genome, the sequences are aligned to the reference genome, with the same parameters above. If there is a genome alignment, we report the alignment to the reference, the strand of the alignment, and the alignment MAPQ score. To lend further context, we report any annotated gene intersection to the reference genome alignment, by first converting the genome alignments to BED format and then using `bedtools intersect` on the genome alignments BED file and a BED file of gene annotations (for this manuscript, we use hg38 RefSeq); for each sequence, we report the list of distinct gene intersections per sequence genome alignment. For each sequence genome alignment, we also report its distances to the nearest annotated exon junctions, by using `bedtools closest` with the sequence genome alignments and BED files of annotated exon start and end coordinates. To report distances of a sequence genome alignment to the nearest upstream exon starts and nearest upstream exon starts, we report the closest exon start and exon end, respectively, with the `bedtools closest` parameters `-D ref -id -t first`. To report distances of a sequence genome alignment to the nearest downstream exon starts and nearest downstream exon starts, we report the closest exon start and exon end, respectively, with the `bedtools closest` parameters `-D ref -iu -t first`.

NOMAD protein profiles

For each set of enriched anchors, homology-based annotation was attempted against an annotated protein database, the Pfam (*18*). For each dataset, up to 1000 of the most significant anchors (q-value < 0.01) were retained for the following analysis: we first generated a substring of each downstream consensus by appending each consensus nucleotide assuming both conditions were met: a minimum observation count of 10 and a minimum agreement fraction of 0.8, until whichever metric first exhibited two consecutive failures at which point no further nucleotide was added. A limit of 1000 anchors was used due to computational constraints from HMMer3 (see below). Anchors that did not have any consensus nucleotides appended were kept as is. An extended anchor was generated for each experiment in which an anchor was found. Each extended anchor was then stored in a final concatenated multi FASTA file with unique seqID headers for each experiment's extended anchors.

The number of matched anchors used for NOMAD and control analysis per dataset are as follows: 201 high effect size anchors in SARS-CoV-2 from South Africa, 252 high effect size anchors in SARS-CoV-2 from France; 1000 anchors were used for rotavirus, human T cells, human B cells, *Microcebus* natural killer T cells, and *Microcebus* B cells.

To assess these extended anchors for protein homology, this concatenated FASTA file was then translated in all six frames using the standard translation table using seqkit (*53*) prior to using hmmsearch from the HMMer3 package (*54*) to assess each resulting amino acid sequences against the Pfam35 profile Hidden Markov Model (pHMM) database.

All hits to the Pfam database were then binned at different E-value orders of magnitude and plotted. In each case, control assessments were performed by repeating the extension and homology searches against an equivalent number of control anchors, selected as the most frequent anchors from that dataset.

Lastly it is worth noting that while only counts of the best scoring Pfam hits were assessed in this study, other information is also produced by HMMer3. In particular, relative alignment positions are given for each hit which could be used to more finely pinpoint the precise locus at which sequence diversification is detected.

We note that while the number of input anchors for NOMAD and control sets are matched, it is possible to have more control protein domains in the resulting barplots, as only high E-value hits to Pfam are reported in the visualizations.

Control analysis

To construct control anchor lists based on abundance, we considered all anchors input to NOMAD and counted their abundance, collapsing counts across targets. That is, an anchor receives a count determined by the number of times it appears at an offset of 5 in the read up to position R- max(0,R/2-2*k) where R is the length of the read, summed over all targets. The 1000 most abundant anchors were output as the control set. For analysis comparing control to NOMAD anchors, min(|NOMAD anchor list|,1000) most abundant anchors from the control set were used and the same number of NOMAD anchors were used, sorted by p-value.

SARS-CoV2 analysis

The SARS-CoV2 datasets used in this manuscript were analyzed with NOMAD's unsupervised mode (no sample metadata provided). To identify high effect size anchors, a threshold of `effect_size_randCjs` > 0.5 was used (table S2). The Wuhan variant reference genome was downloaded from NCBI, assembly NC_045512.2. The Omicron and Delta mutation variants were downloaded as FASTA from the UCSC track browser in June 2022, with the following parameters: clade 'Viruses', assembly NC_045512.2, genome 'SARS-CoV-2', group 'Variations and Repeats', track 'Variants of Concert', and table 'Omicron Nuc Muts (variantNucMuts_B_1_1_529)' and 'Delta Nuc Muts (variantNucMutsV2_B_1_617_2)'.

Variant genomes were downloaded in FASTA file format, and bowtie indices were built from these FASTA files, using default parameters. To determine alignment of anchors to the Wuhan genome, anchor sequences were converted to FASTA format and aligned to the Wuhan bowtie index with bowtie (default parameters). After mapping of NOMAD anchors, the number of control anchors were chosen to match the number of anchors mapped by bowtie to report comparable numbers.

Mutation consistency to the Omicron and Delta variants was reported as follows. For each anchor mapping to the Wuhan reference in the positive strand, an anchor at position x is called mutation-consistent if there is an annotated variant between positions $x+k+D$ and $x+R+2*D$, where D=max(0, (L - 2 * k) / 2), L is the length of the first read processed in the dataset, and the factor of 2 reflects the bowtie convention of reporting the left-most base in the alignment. The reciprocal logic was used to define mutation-consistency for anchors mapping to the negative strand – e.g. a mutation had to occur between positions $x-(R+D)$ and x. In total, we report: a) number of anchors mapping with bowtie default parameters to the Wuhan reference; b) number and fraction of mutation-consistent anchors as described above.

Viral protein profile analysis

In influenza, NOMAD's most frequently hit profiles were Actin (62 hits), and GTP_EFTU (23 hits), and the influenza-derived Hemagglutinin (17 hits), consistent with virus-induced alternative splicing of Actin (*55*) and EF-Tu, further elucidating these proteins' roles during infection (*37*, *56*) (no such hits were found in the control). Similarly, in a study of metagenomics of rotavirus breakthrough cases, NOMAD protein profile analysis prioritized domains known to be involved in host immune suppression.

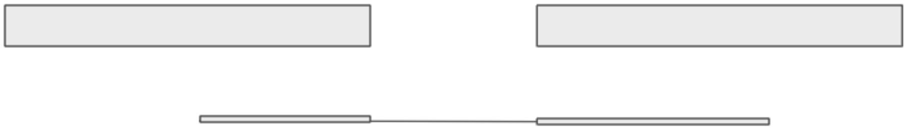
In rotavirus, the most enriched domain in NOMAD compared to control was the rotavirus VP3 (Rotavirus_VP3, 76 NOMAD hits vs 9 control hits), a viral protein known to be involved in host immune suppression (*57*), and the rotavirus NSP3 (Rota_NSP3, 87 NOMAD vs 35 control hits), a viral protein involved in subverting the host translation machinery (*58*), both proteins that might be expected to be under constant selection given their intimate host interaction.

Identifying cell-type specific isoforms in SS2

In the analysis of HLCA SS2 data, we utilize "isoform detection conditions" for alternative isoform detection. These conditions select for (anchor, target) pairs that map exclusively to the human genome, anchors with at least one split-mapping consensus sequence, mu_lev > 5, and M > 100. mu_lev is defined as the average target distance from the most abundant target as measured by Levenstein distance. To identify anchors and targets that map exclusively to the human genome, we included anchors and targets that had exactly one element annotation, where that one element annotation must be grch38_1kgmaj. To identify anchors with at least one split-mapping consensus, we selected anchors that had at least one consensus sequence with at least 2 called exons. The conditions on Levenshtein distance, designed to require significant across-target sequence diversity, significantly reduced anchors analyzed (excluding many SNP-like effects). We further restricted to anchors with M > 100, to account for the lower cell numbers in macrophage cells; note that the user can perform inference with a lower M requirement, based on input data. These isoform detection parameters were used to identify the SS2 examples discussed in this manuscript, MYL12. For HLA discussion, gene names were called using consensus_gene_mode.

Splice junction calls

To identify exon coordinates for reporting annotations in this manuscript, consensus sequences are mapped with STAR aligner (default settings) (*59*). Gapped alignments are extracted and their coordinates are annotated with known splice junction coordinates using 'bedtools bamtobed --split'; each resulting contiguously mapping segment is called a "called exon" (see below). From each consensus sequence, called exons are generated as start and end sites of each contiguously mapped sequence in the spliced alignment. These 'called exons' are then stratified as start sites and end sites. Note that the extremal positions of all called exons would not be expected to coincide with a splice boundary (see below); "called exon" boundaries would coincide with an exon boundary if they are completely internal to the set of called exon coordinates. Each start and end site of each called exon is intersected with an annotation file of known exon coordinates; it receives a value of 0 if the site is annotated, and 1 if it is annotated as alternative. The original consensus sequence and the reported alignment of the consensus sequence are also reported. Gene names for each consensus are assigned by bedtools intersect with gene annotations (hg38 RefSeq for human data by default), possibly resulting in multiple gene names per consensus.



Caption: Example of how spliced reads are converted to "called exons" (bottom) and are compared to annotated exons (top); right most and leftmost boundaries of called exons are not expected to coincide with annotated exon boundaries and are excluded from analysis of concordance between consensus called-exons and annotations.

HLA analysis in HLCA

NOMAD summary files were processed by restricting to anchors aligning to the human genome , and having at least 1 target with this characteristic. Further, mu_lev had to exceed 1.5. For HLA discussion, gene names were called using consensus_gene_mode.

B,T Cell Transcriptome Annotations

To determine the most frequent transcriptome annotation for a dataset, all significant anchors were mapped to the human transcriptome (GRCh38, Gencode) with bowtie2, using default parameters and `-k 1` to report at most one alignment per anchor. Then, the bowtie2 transcript hits are aggregated by counting over anchors. The transcript hits with the highest counts over all anchors were reported.

Further immune cell protein domain analysis

In human B and T cells, NOMAD blindly rediscovered the high degree of single-cell variability in the immunoglobulin (IG) in B cells: this locus was most highly ranked by anchor counts per transcript (Fig. 4E). In B cells, NOMAD anchor counts were highest in genes IGKV3-11, IGKV3D-20, IGKV3D-11, and IGKC, the first three being variable regions of the B cell receptor (Fig. 4E).

Parallel analysis of T cells showed similar rediscovery and extension of known biology: HLA-B, RAP1B, TRAV26-2, and TRBV20-1 were the highest-ranked transcripts in T cells measured by anchor counts. HLA-B is a major histocompatibility (MHC) class I receptor known to be expressed in T cells, and TRAV26-2 and TRBV20-1 are variable regions of the T cell receptor. T cell expression of HLA-B alleles has been correlated with T cell response to HIV (*60*, *61*). Fig. 4E shows many other genes known to be rearranged by V(D)J were also recovered. In the control sets for both B and T cells, enriched genes were unrelated to immune functions (Fig. 4E, Fig. S2G,H).

HLA-B (Fig. 4E) is the most densely hit transcript in T cells. Mapping assembled consensuses shows two dominant alleles: one perfectly matches a reference allele, the other has 4 polymorphisms all corresponding with positions of known SNPs. NOMAD statistically identifies T cell variation in the expression of these two alleles, some T cells having only detectable expression of one but not the other (p< 4,6E-24) (*62*). Other HLA alleles called by NOMAD, including HLA-F, have similar patterns of variation in allele-specific expression (Supplement).

NOMAD comparison to BASIC analysis in lemur spleen B cells

To compare performance, we first ran BASIC on the lemur spleen B cells, with the following additional parameters: -a. We then ran NOMAD on cells where BASIC failed to identify the light chain variable gene family, by selecting cells annotated as "No BCR light chain" from the BASIC output. From the NOMAD output, we identified anchors which mapped to the IGL gene by bowtie; to do this, we used the command `grep IGL  “$file”`, where “$file” corresponds to the NOMAD anchor genome annotations output file. This resulted in the following 5 anchors: CCTCAGAGGAGGGCGGGAACAGCGTGA, CTCGGTCACTCTGTTCCCGCCCTCCTC, GCCCCCTCGGTCACTCTGTTCCCGCCC, GGGCGGGAACAGCGTGACCGAGGGGGC, TCACTCTGTTCCCGCCCTCCTCTGAGG.

We then fetched the consensus sequences associated with the above IGL-mapping anchors, and converted those consensus sequences into FASTA format. We ran the following command on that FASTA file (denoted by “$fasta”): `blastn -outfmt "$fmt" -query "$fasta" -remote -db nt -evalue 0.1 -task blastn -dust no -word_size 24 -reward 1 -penalty -3 -max_target_seqs 200`, where $fmt corresponds to "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore sseqid sgi sacc slen staxids stitle".

From this BLAST output, we checked that light chain variable regions were identified, via grep for the term "light chain variable", yielding 60 sequences. Each cell could have at most 5 contributions to this number, and thus at least 12 cells (conservatively) had NOMAD-identified partial light chain variable sequences.

Timing for SS2

Because code was run on a server with dynamic memory, we report summary statistics as follows. For the steps parallelized by FASTQ file, such as anchor and target retrieval, total time for dataset run, as reported by Nextflow, was parsed per cell. Thus, the average time per cell is reported. For the steps parallelized by 64 files (q-value calculations), total extracted times were summed and divided by number of cells. For steps that consisted of aggregating files, total run time was divided by number of cells. Thus, the total time and memory should be multiplied by the total number of cells to achieve an estimate of the pipeline time for this dataset.

Laptop analysis details

Laptop specs:
An Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz (launched in 2015)
2 cores, total of 4 threads, 3 of which NOMAD was allowed to use.
8 GB DDR3 RAM
SODIMM DDR3 Synchronous 1600 MHz (0.6 ns)
Dataset:
10 files of SS2 T cell reads
43,870,027 reads total

Figure data

Protein graphics from https://pdb101.rcsb.org/browse/coronavirus.
Virus graphics from https://thenounproject.com/icon/virus-2198681.
Nasal swab graphics from https://thenounproject.com/icon/swab-3826339.
Person graphics from https://thenounproject.com/icon/person-1218528.
Flower graphics from https://thenounproject.com/icon/flower-3580625/.

Microscope graphics from https://thenounproject.com/icon/microscope-5000952/.
Bacteria graphics from https://thenounproject.com/icon/bacteria-3594201/.
Cell graphics from https://thenounproject.com/icon/cell-1529259/.
MiSeq graphic from Bioicons, DBCLS.
Cell graphics from Bioicons, Servier.

Supplementary Text

Generality of NOMAD

In this work we focused our experimental results on identifying changes in viral strains and specific examples of RNA-seq analysis. NOMAD's probabilistic formulation extends much further however, and subsumes a broad range of problems. Many other tasks, some described below, can also be framed under this unifying probabilistic formulation. Thus, NOMAD provides an efficient and general solution to disparate problems in genomics.

We outline examples of NOMAD's predicted application in various biological contexts, highlighting the anchors that would be flagged as significant:

- RNA splicing, even if not alternative or regulated, can be detected by comparing DNA-seq and RNA-seq
  - Examples of predicted significant anchors: sequences upstream of spliced or edited sequences including circular, linear, or gene fusions
- RNA editing can be detected by comparing RNA-seq and DNA-seq
  - Examples of predicted significant anchors: sequences preceding edited sites
- Liquid biopsy – reference free detection of SNPs, centromeric and telomeric expansions with mutations
  - Examples of predicted significant anchors: sequences in telomeres (resp. centromeres) preceding telomeric (resp. centromeric) sequence variants or chromosomal ends (telomeres) in cancer-specific chromosomal fragments
- Detecting MHC allelic diversity
  - Examples of predicted significant anchors: sequences flanking MHC allelic variants
- Detecting disease-specific or person-specific mutations and structural variation in DNA
  - Examples of predicted significant anchors: sequences preceding structural variants or mutations
- Cancer genomics eg. BCR-ABL fusions and other events
  - Examples of predicted significant anchors: sequences preceding fusion breakpoints
- Transposon or retrotransposon insertions or mobile DNA/RNA
  - Examples of predicted significant anchors: (retro)transposon arms or boundaries of mobile elements
- Adaptation
  - Examples of predicted significant anchors: sequences flanking regions of DNA with time-dependent variation

- Novel virus' and bacteria; emerging resistance to human immunity or drugs
    - Examples of predicted significant anchors: sequences flanking rapidly evolving or recombined RNA/DNA
- Alternative 3' UTR use
    - Examples of predicted significant anchors: 3' sequences with targets including both the poly(A) or poly(U), or adaptors in cases of libraries prepared by adaptor ligation versus downstream transcript sequence
- Hi-C or any proximity ligation
    - Examples of predicted significant anchors: for Hi-C, DNA sequences with differential proximity to genomic loci as a function of sample; similarly, for other proximity ligation anchors would be predicted when the represented element has differential localization with other elements
- Finding combinatorially controlled genes e.g. V(D)J
    - Examples of predicted significant anchors sequences in the constant, D, J, or V domains

Generality of NOMAD anchor, target and consensus construction

NOMAD can function on any biological sequence and does not need anchor-target pairs to take the form of gapped kmers, and can take very general forms. One example is (XXY)^m where X is a base in the anchor and Y in the target, to identify sequences such as in known diversity generating retroelements (*63*), or ones with synonymous amino acid changes. X and Y could also be amino acid sequences or other discrete variables considered in molecular biology. NOMAD consensus building can be developed into statistical *de novo* assemblies, including mobile genetic elements with and without circular topologies. Much more general forms of anchor-target pairs (or tensors) can be defined and analyzed, including other univariate or multivariate hash functions on targets or sample identity. NOMAD can also be further developed to analyze higher dimensional relationships between anchors, where inference can be performed on tensors across anchors, targets, and samples. Similarly, hash functions can be optimized under natural maximization criterion, which is the subject of concurrent work. The hash functions can also be generalized to yield new new statistics, optimizing power against different alternatives.

Statistical Inference

In this section we discuss the statistics underlying our p-value computation. As discussed, detecting deviations from the global null, where the probability of observing a given target *k*-mer *t* *L* bases downstream of an anchor *a* is the same across samples, can be mapped to a statistical test on counts matrices (contingency tables).

Probabilistic model

Formally, we study the null model posed below.

Null model:

Conditional on anchor a, each target is sampled independently from a common vector of (unknown) target probabilities not depending on the sample.

Despite its rich history, the field of statistical inference for contingency tables still has many open problems (*40*). The field's primary focus has been on either small contingency tables (2x2, e.g. Fisher's exact test(*64*)), high counts settings where a chi-square test yields asymptotically valid p-values, or computationally intensive Markov-Chain Monte-Carlo (MCMC) methods. None of these approaches are simultaneously efficient and provide closed form, finite-sample valid statistical inference with desired power for the application setting at hand.

We note that even though we are not aware of directly applicable results, it may be theoretically possible to obtain finite-sample-valid p-values using likelihood ratio tests or a chi-squared statistic. However, even if this were possible, it would not allow for the modularity of our proposed method, where we can a) weight target discrepancies differently as a function of their sequences, to allow for power against different alternatives, b) reweight each sample's contribution to normalize for unequal sequencing depths, and c) offer biological interpretability in the form of cluster detection and target partitioning. Overall, the statistics we develop for NOMAD are extremely flexible. Ongoing work is focused on further optimizing this general procedure, including application specific tuning of the functions f and robustification of the statistic against biological and technical noise.

Test intuition

From a more linear algebraic perspective, the intuition for the power of our test can be captured as follows; any test will reduce to computing a scalar valued test statistic from the contingency table, and determining whether this is above or below a rejection threshold. Restricting to linear statistics for simplicity, this corresponds to a hyperplane in the contingency table space (T x p, targets x samples). Informally, this means that our statistic loses information; it is taking a T x p matrix, projecting it down to 1 dimensional space, and thresholding, yielding a significant null space, and causing our test statistic to lose power in these directions: for any fixed projection, is has no power against many alternatives. Thus, we make 2 modifications: firstly, we utilize random projections, to ensure that we do not deterministically miss certain alternatives (fixed random seed programmatically for reproducibility). Secondly, we use several random projections in the computation of our test statistic, taking the minimum p-value over each of these directions, trading off between the probability of missing a true positive and the correction factor required.

One natural choice of f is constructed to capture the intuition that target diversity is most interesting when target sequences are highly divergent. To define f, i) targets are ranked by abundance; ii) the i-th target is assigned a scalar value measuring its minimum distance (such as Hamming, Levenstein) to all more abundant targets. Note that in order to ensure that this inference is statistically valid, we need to split the data and measure abundance on a subset of data that we do not use for downstream processing (to avoid data snooping). This function has some power to identify sample-dependent splicing, but little power to discriminate SNPs in targets. This is because, as these scores will be aggregated over the targets of a given sample, we see that in this example all samples that express the primary isoform will have an average target function value close to 0, whereas the alternatively spliced samples will have large target function values. However, such a function f has a major drawback; it is not able to fully utilize the dynamic range of this function. Since our procedure is scale invariant it suffices to consider f bounded between 0 and 1, and so we need to normalize by the maximum value of f that can be observed, which is k=27. This can be problematic, as seen by an example where the spliced target is a distance of 5 away, leaving its value at 5/27 instead of 1. To this end, we instead appeal to the probabilistic nature of our problem, and utilize several independent random functions f. That is to say, each random function f we utilize assigns a value of 0 or 1 independently to each target, fully utilizing the available dynamic range, and extending our detection power beyond SNPs.

p-value computation

NOMAD's p-value computation is performed independently on each anchor, and so statistical inference can be performed in parallel across all anchors. Our test statistic is based on a linear combination of row and column counts, giving valid FDR-controlled q-values by classical concentration inequalities and multiple hypothesis correction (Fig. S1A). To formalize our notation, we define $D_{j,k}$ as the sequence identity of the k-th target observed for the j-th sample. This ordering with respect to k that we assign is for analysis purposes only, it has no relation to the order in which targets are observed in the actual FASTQ files (can be thought of as randomly permuting the order in which we observe the targets). Appealing to the null model, we have that each $D_{j,k}$ is then an independent draw from the common target distribution.

To construct our p-value, we first estimate the expectation (unconditional on sample identity) of $f(D_{j,k})$ as $\hat{\mu}$ by aggregating all our data. Next, we aggregate these $f(D_{j,k})$ across only sample j to compute $\hat{\mu}_j$, constructing S_j as the difference between the these two, normalizing by $\sqrt{n_j}$ to ensure that each S_j will have essentially constant variance (up to the correlation between $\hat{\mu}, \hat{\mu}_j$). This is performed as below:

$$\hat{\mu} = \frac{1}{M} \sum_{j,k} f(D_{j,k})$$

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} f(D_{j,k})$$

$$S_j = \sqrt{n_j}(\hat{\mu}_j - \hat{\mu})$$

$$S = \sum_{j=1}^{p} c_j S_j$$

We see that S_j is a signed measure of how different the target distribution of sample j is from the table average, when viewed under the expectation with respect to f. This function f is critical to obtain good statistical guarantees, and the choice of f determines the direction of statistical power, such as power to detect SNPs versus alternative splicing or other events. In this work we design a general probabilistic solution, utilizing several random functions f which take value 0 or 1 on targets, independently and with equal probability. In order to increase the probability that NOMAD identifies anchors with significant variation, several (K=10 by default) random functions are utilized for each anchor, though more may be desired depending on the application.

After constructing these signed anchor-sample scores, they need to be reduced to a scalar valued test-statistic. Consider first the case where we are given sample metadata, i.e. we know that our samples come from two groups, and we want our test to detect whether the target distribution differs between the two groups. One natural way of performing such a test is to first aggregate the anchor-sample scores over each group, and then compute the difference between these group aggregates.

We formalize this by assigning a scalar c_j to each sample, where in this two group comparison with metadata c_j = +/- 1 encodes the sample's identity, and construct the anchor statistic S as the inner product between the vector of c_j's and the anchor-sample scores. This statistic will have high expected magnitude if there is significant variation in target distribution between the two groups.

In many biologically important applications however, cell-type metadata is not available. In these cases, NOMAD detects heterogeneity within a dataset by performing several (L=50 by default) random splits of the samples into two groups . For each of these L splits NOMAD assigns c_j= +/-1 independently and with equal probability for each sample, computes the test statistic for each split, and selects the split yielding the smallest p-value.

We now investigate the statistical properties of S. First, observe that S has mean 0 under the null hypothesis. This allows us to bound the probability that the random variable S is larger than our observed anchor statistic as follows. Since f and c are fixed, and are independent of the data, we have that since $f(D_{j,k})$ are bounded between 0 and 1 we can apply Hoeffding's inequality for bounded random variables. Defining μ as the expectation with respect to the common underlying distribution of $f(D_{j,k})$ (unknown), we center our random variables by subtracting the sample mean $\hat{\mu}$, our estimate of the true mean μ. Standard bounds can now be applied to decompose this deviation probability into two intuitive terms:

1) the probability that the statistic $\tilde{S}$, constructed with additional knowledge of the true μ, is large

$$\tilde{S} = \sum_j c_j \left(\hat{\mu}_j - \mu\right)$$

2) the probability that $\hat{\mu}$ is far from μ.

Following this approach, we have that

$$
\begin{aligned}
&\mathbb{P}\left(|S| \geq \epsilon\right) \\
&\quad= \mathbb{P}\left(\left|\sum_{j,k} c_j \frac{f(D_{j,k}) - \hat{\mu}}{\sqrt{n_j}}\right| \geq \epsilon\right) \\
&\quad= \mathbb{P}\left(\left|\sum_{j,k} c_j \frac{f(D_{j,k}) - \mu}{\sqrt{n_j}} + (\mu - \hat{\mu}) \sum_j c_j \sqrt{n_j}\right| \geq \epsilon\right) \\
&\quad\leq \min_{a \in (0,1)} \mathbb{P}\left(\left|\sum_{j,k} c_j \frac{f(D_{j,k}) - \mu}{\sqrt{n_j}}\right| \geq (1-a)\epsilon\right) + \mathbb{P}\left(\left|(\mu - \hat{\mu}) \sum_j c_j \sqrt{n_j}\right| \geq a\epsilon\right) \\
&\quad\overset{(a)}{=} \min_{a \in (0,1)} \mathbb{P}\left(\left|\sum_{j,k} \frac{c_j}{\sqrt{n_j}} (f(D_{j,k}) - \mu)\right| \geq (1-a)\epsilon\right) + \mathbb{P}\left(\left|\frac{1}{M} \sum_{j,k} f(D_{j,k}) - \mu\right| \geq \frac{a\epsilon}{\left|\sum_j c_j \sqrt{n_j}\right|}\right) \\
&\quad\overset{(b)}{\leq} \min_{a \in (0,1)} 2\exp\left(-\frac{(1-a)^2 \epsilon^2}{2\sum_{j,k} \frac{c_j^2}{4n_j}}\right) + 2\exp\left(-\frac{\frac{a^2 M^2 \epsilon^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}}{2M\frac{1}{4}}\right) \\
&\quad= \min_{a \in (0,1)} 2\exp\left(-\frac{2(1-a)^2 \epsilon^2}{\sum_{j: n_j > 0} c_j^2}\right) + 2\exp\left(-\frac{2a^2 M \epsilon^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}\right).
\end{aligned}
$$

where (a) comes from the assumption that the sum in the denominator of the second term is nonzero, as otherwise this second term is 0 and we can essentially set a=0. (b) utilizes Hoeffding's inequality on each of these two terms. We can easily optimize this bound over a to within a factor of two of optimum by equating the two terms (as one is increasing in a and the other is decreasing), which is achieved when

$$
a = \left(1 + \sqrt{\frac{M \sum_{j: n_j > 0} c_j^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}}\right)^{-1}
$$

Thus, for an observed value of our test statistic S, we construct NOMAD’s statistically valid p-values as

$$
P = 2\exp\left(-\frac{2(1-a)^2 S^2}{\sum_{j: n_j > 0} c_j^2}\right) + 2\exp\left(-\frac{2a^2 M S^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}\right) \quad \text{with} \quad a = \left(1 + \sqrt{\frac{M \sum_j c_j^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}}\right)^{-1}
$$

q-value computation

Our q-values are computed using Benjamini Yekutieli correction (*43*) as

$$
Q_{(i)}^{\mathrm{BY}} = \min\left(\min_{j \geq i} \frac{m(\log m + 1) p_{(j)}}{j}, 1\right)
$$

which enables NOMAD to control the false discovery rate of the reported significant anchors.

Note that, in the case of A anchors, we can construct a strictly more powerful statistical procedure by modifying how we correct for multiple hypotheses. Instead of first applying Bonferroni correction over the L*K hypotheses (different c and f configurations) then BY correcting the A aggregate hypotheses, we can directly apply BY correction to all A*L*K individual hypotheses. This procedure will still be FDR controlled, and will yield at least as many discovered anchors. For clarity here, however, we apply Bonferroni correction to yield valid p-values for each anchor individually.

Effect size

NOMAD provides a measure of effect size when the c_j's used are +/- 1, to allow for prioritization of anchors with fewer counts but large inter-sample differences in target distributions. Effect size is calculated based on the split c and function f that yield the most significant NOMAD p-value. Fixing these, the effect size is computed as the difference between the mean over targets with respect to f across those samples with c = +1, and the mean over targets (with respect to f) across those samples with c = -1. This effect size is bounded between 0 and 1, with 0 indicating no effect (target distributions are identical when aggregated within each group), and 1 indicating disjoint supports. Defining A_+ as the set of j where $c_j>0$, and A_- as the set of j where $c_j < 0$ (generalizing beyond the case of c_j=+/-1), this is formally computed as:

$$\left|\frac{1}{\sum_{j\in A_+} n_j}\sum_{j\in A_+} n_j\hat{\mu}_j - \frac{1}{\sum_{j\in A_-} n_j}\sum_{j\in A_-} n_j\hat{\mu}_j\right|$$

In this simple case of c_j=+/-1 and {0,1} valued f, this is simply a projection of the T x p table to a 2x2 table. Even considering more general f, there is an easy to understand alternative that NOMAD is designed to have power against. The effect size should be thought of under the alternative hypothesis where the columns follow multinomial distributions with probability vector p_1 or probability vector p_2, depending on the group identity c_j. The effect size we compute can be thought of in this scenario as measuring the difference between the expectation of f under p_1 and p_2. In the case of maximizing the effect size over all possible {0,1}-valued f, the effect size will be equal to the total variation distance between the empirical distributions of the group c_j = +1 and c_j=-1. Thus, the effect size will be 1 if and only if the two sample groups partition targets into 2 disjoint sets on which the function f takes opposite values, as to be expected from the total variation distance interpretation (Fig. S1B). This f will place a value of 1 on targets where the empirical frequency of the +1 group $p_{1,t}$ is larger than that of the -1 group $p_{2,t}$. Since p_1 and p_2 are probability distributions, this ends up being exactly the total variation distance between them (i.e. half the vector ell-1 distance). Note that we can also consider a signed variant of this effect size measurement, where if we restrict ourselves to the same c and f for several anchors, the effect size sign gives us additional information about the direction of the effect.

Ability to operate without metadata

As discussed, NOMAD can be run without any metadata. For the HLCA dataset, when run on the two donors without metadata, NOMAD calls 6287 anchors (2269 genes) as opposed to the 3439 anchors (1384 genes) called with metadata for donor 1. Filtering for genes hit by more than two anchors, NOMAD's metadata free approach calls >94% of the genes called by the metadata-based approach (Fig. S3A). For donor 2, NOMAD calls 5619 anchors (1844) genes without any metadata as opposed to the 3775 anchors (1125) genes called with metadata. Filtering for genes hit by more than two anchors, NOMAD's metadata free approach calls >90% of the genes called by the metadata-based approach, increasing to >94% for those genes hit by at least 3 anchors.

p-value computation for scatterplots depicting target fraction abundance

p-values are constructed as follows: first, we compute p, the average occurrence of target 1 for this anchor (sum of counts of observations of target 1 divided by the total number of observations). Then, for all possible n_j, we compute 1% and 99% quantiles (confidence bounds) for a binomial distribution with n_j trials and heads probability p. If the fraction of target 1 in each sample was independent of sample identity, and were indeed binomially distributed, then each sample would have at least a 98% probability of falling within this confidence interval. Thus, we compute our test statistic X as the number of samples that fall outside of the [1,99] quantiles, and compute as our p-value the probability that a binomial random variable with n = number of samples and p = .02 is at least as large as X.

While intuitive, the above analysis is loose. Firstly, since binomials are discrete distributions, we will rarely be able to compute exact 1% and 99% quantiles. Thus, the probability that for any given n_j a sample will fall outside of the [1,99] quantiles, which we denote p_j, is almost always substantially less than .02. The true distribution of X is then poisson binomial, with this vector of probabilities (all at most .02), one for each sample. However, as this p-value is numerically difficult to compute, we bound this p-value as the probability that a binomial random variable with n = number of samples with $p_j>0$ and $p=\max_j p_j \leq .02$ is greater than our observed test statistic.

Hypergeometric p-value computation

p-values for protein domain analysis were generated using a hypergeometric test. For a given domain, we construct the 2x2 contingency table, where the first row is the number of NOMAD hits for this domain, followed by the total number of NOMAD hits not in this domain. The second row is the mirror of this for control, where the first entry is the number of control hits for this domain, followed by the total number of control hits not in this domain. Then, a one-sided p-value is computed using Fisher's exact test, which is identically a hypergeometric test. Then, we apply Bonferroni correction for the total number of protein domains expressed by either NOMAD or control, to yield the stated p-values.

S1

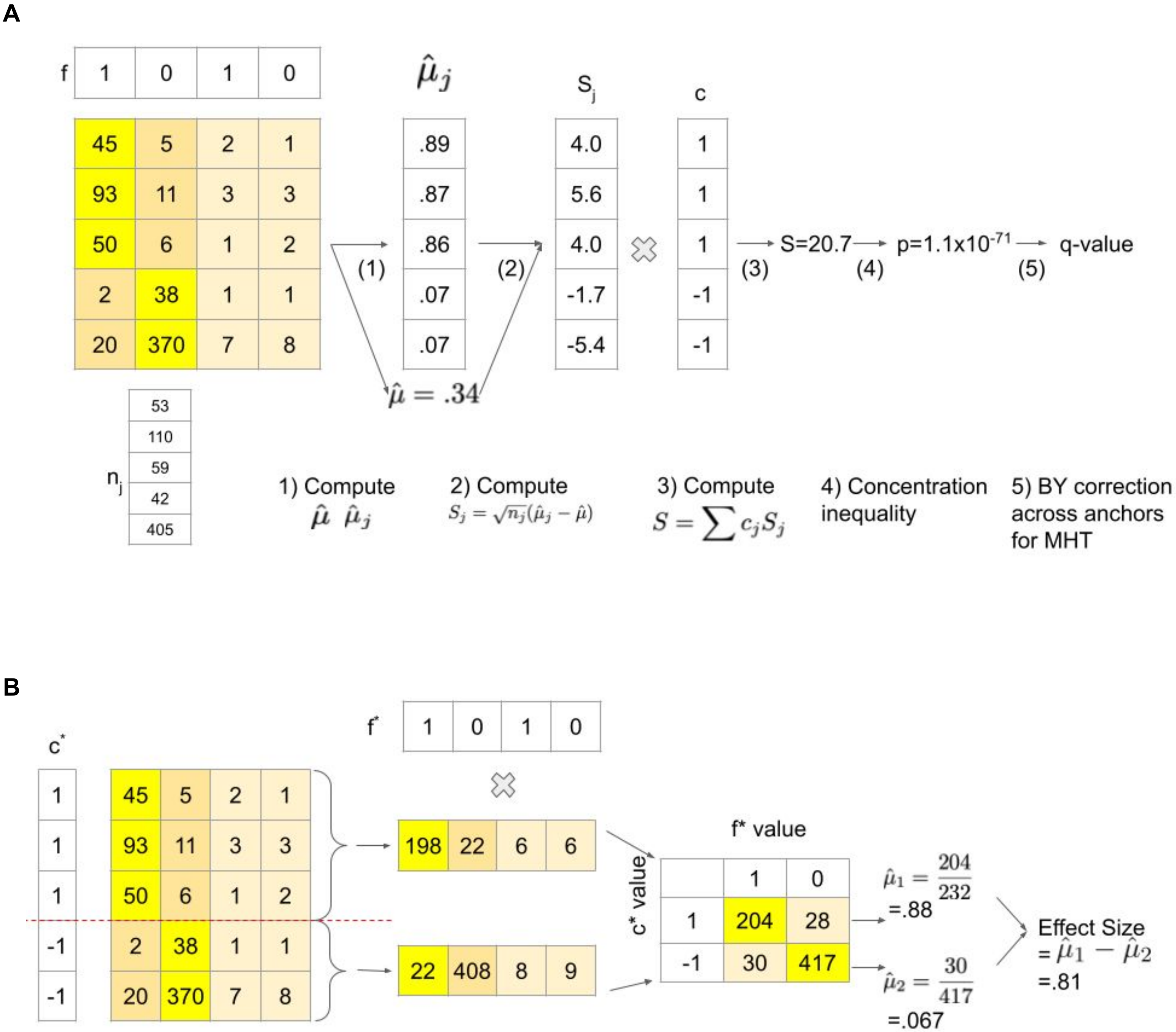
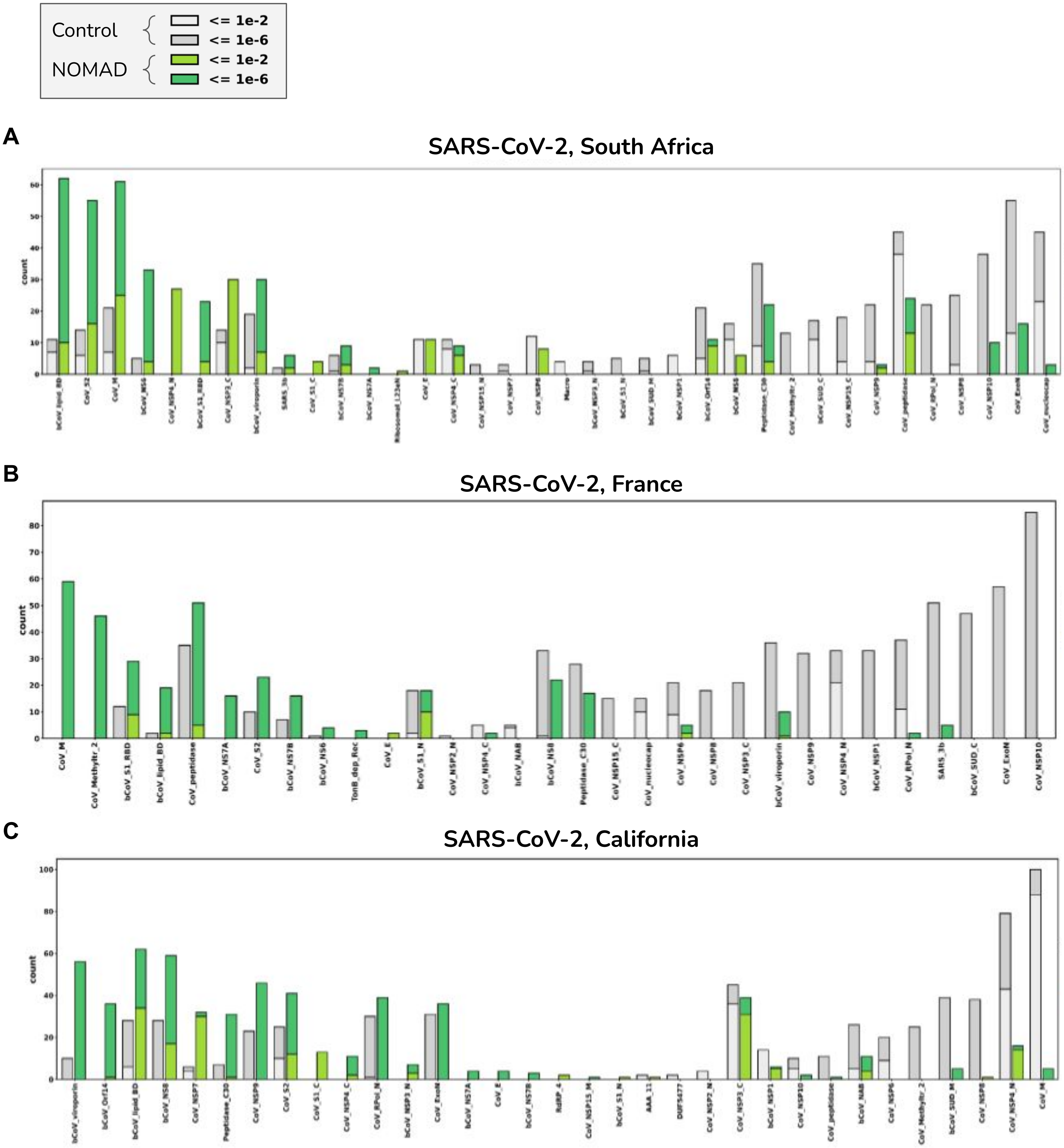

A
f 1 0 1 0
45 5 2 1
93 11 3 3
50 6 1 2
2 38 1 1
20 370 7 8
n_j 53 110 59 42 405
$\hat{\mu}_j$
.89 .87 .86 .07 .07
$\hat{\mu} = .34$
S_j
4.0 5.6 4.0 -1.7 -5.4
c
1 1 1 -1 -1
(1)
(2)
(3)
(4)
(5)
S=20.7
$p=1.1x10^{-71}$
q-value
1) Compute $\hat{\mu}$ $\hat{\mu}_j$
2) Compute $S_j = \sqrt{n_j}(\hat{\mu}_j - \hat{\mu})$
3) Compute $S = \sum c_j S_j$
4) Concentration inequality
5) BY correction across anchors for MHT
B
c^*
1 1 1 -1 -1
f^* 1 0 1 0
45 5 2 1
93 11 3 3
50 6 1 2
2 38 1 1
20 370 7 8
198 22 6 6
22 408 8 9
f* value
c* value
1 0
1 204 28
-1 30 417
$\hat{\mu}_1 = \frac{204}{232}$ =.88
$\hat{\mu}_2 = \frac{30}{417}$ =.067
Effect Size $= \hat{\mu}_1 - \hat{\mu}_2$ =.81

Fig. S1: NOMAD Statistics.

A. p-value computation for NOMAD. Contingency table transposed for visual convenience (rows are samples and columns are targets). Starting with a samples by targets counts matrix, NOMAD utilizes one (or several) functions f mapping targets to values within [0,1]. The mean with respect to f is taken over the targets in each row j to yield $\hat{\mu}_j$, and an estimate for the mean over all target observations of f is taken, yielding $\hat{\mu}$. The anchor-sample scores S_j are then constructed as the difference between the row mean $\hat{\mu}_j$ and the overall mean $\hat{\mu}$, and is scaled by $\sqrt{n_j}$. These anchor-sample scores are weighted by c_j in [-1,1] and summed to yield the anchor statistic S. Finally, a p-value is computed utilizing classical concentration inequalities, which we correct for multiple hypothesis testing (with dependence) by constructing q-values using Benjamini-Yekutieli, a variant of BH testing which corrects for arbitrary dependence.

B. Effect size computation for NOMAD. Effect size is calculated based on the random split c and random function f that yielded the most significant NOMAD p-value. Fixing these, the effect size is computed as the difference between the mean across targets (with respect to f) across those samples with $c_j = +1$, and the mean across targets (with respect to f) across those samples with $c_j = -1$. This should be thought of as studying an alternative where samples from c_j=+1 have targets that are independent and identically distributed with mean (under f) of μ_1, and samples with c_j=-1 have targets that are independent and identically distributed with mean (under f) of μ_2. The total effect size is estimated as $\mu_1 - \mu_2$.

S2


Control
<= 1e-2
<= 1e-6
NOMAD
<= 1e-2
<= 1e-6
A
SARS-CoV-2, South Africa
count
0
10
20
30
40
50
60
bCoV_lipid_BD
CoV_S2
CoV_M
bCoV_NS6
CoV_NSP4_N
bCoV_S1_RBD
CoV_NSP3_C
bCoV_viroporin
SARS_3b
CoV_S1_C
bCoV_NS7B
bCoV_NS7A
Ribosomal_L23eN
CoV_E
CoV_NSP4_C
CoV_NSP15_N
CoV_NSP7
CoV_NSP6
Macro
bCoV_NSP3_N
bCoV_S1_N
bCoV_SUD_M
bCoV_NSP1
bCoV_Orf14
bCoV_NS8
Peptidase_C30
CoV_Methyltr_2
bCoV_SUD_C
CoV_NSP15_C
CoV_NSP9
CoV_peptidase
CoV_RPol_N
CoV_NSP8
CoV_NSP10
CoV_ExoN
CoV_nucleocap
B
SARS-CoV-2, France
count
0
10
20
30
40
50
60
70
80
CoV_M
CoV_Methyltr_2
bCoV_S1_RBD
bCoV_lipid_BD
CoV_peptidase
bCoV_NS7A
CoV_S2
bCoV_NS7B
bCoV_NS6
TonB_dep_Rec
CoV_E
bCoV_S1_N
CoV_NSP2_N
CoV_NSP4_C
bCoV_NAB
bCoV_NS8
Peptidase_C30
CoV_NSP15_C
CoV_nucleocap
CoV_NSP6
CoV_NSP8
CoV_NSP3_C
bCoV_viroporin
CoV_NSP9
CoV_NSP4_N
bCoV_NSP1
CoV_RPol_N
SARS_3b
bCoV_SUD_C
CoV_ExoN
CoV_NSP10
C
SARS-CoV-2, California
count
0
20
40
60
80
100
bCoV_viroporin
bCoV_Orf14
bCoV_lipid_BD
bCoV_NS8
CoV_NSP7
Peptidase_C30
CoV_NSP9
CoV_S2
CoV_S1_C
CoV_NSP4_C
CoV_RPol_N
bCoV_NSP3_N
CoV_ExoN
bCoV_NS7A
CoV_E
bCoV_NS7B
RdRP_4
CoV_NSP15_M
bCoV_S1_N
AAA_11
DUF5477
CoV_NSP2_N
CoV_NSP3_C
bCoV_NSP1
CoV_NSP10
CoV_peptidase
bCoV_NAB
CoV_NSP6
CoV_Methyltr_2
bCoV_SUD_M
CoV_NSP8
CoV_NSP4_N
CoV_M

S2

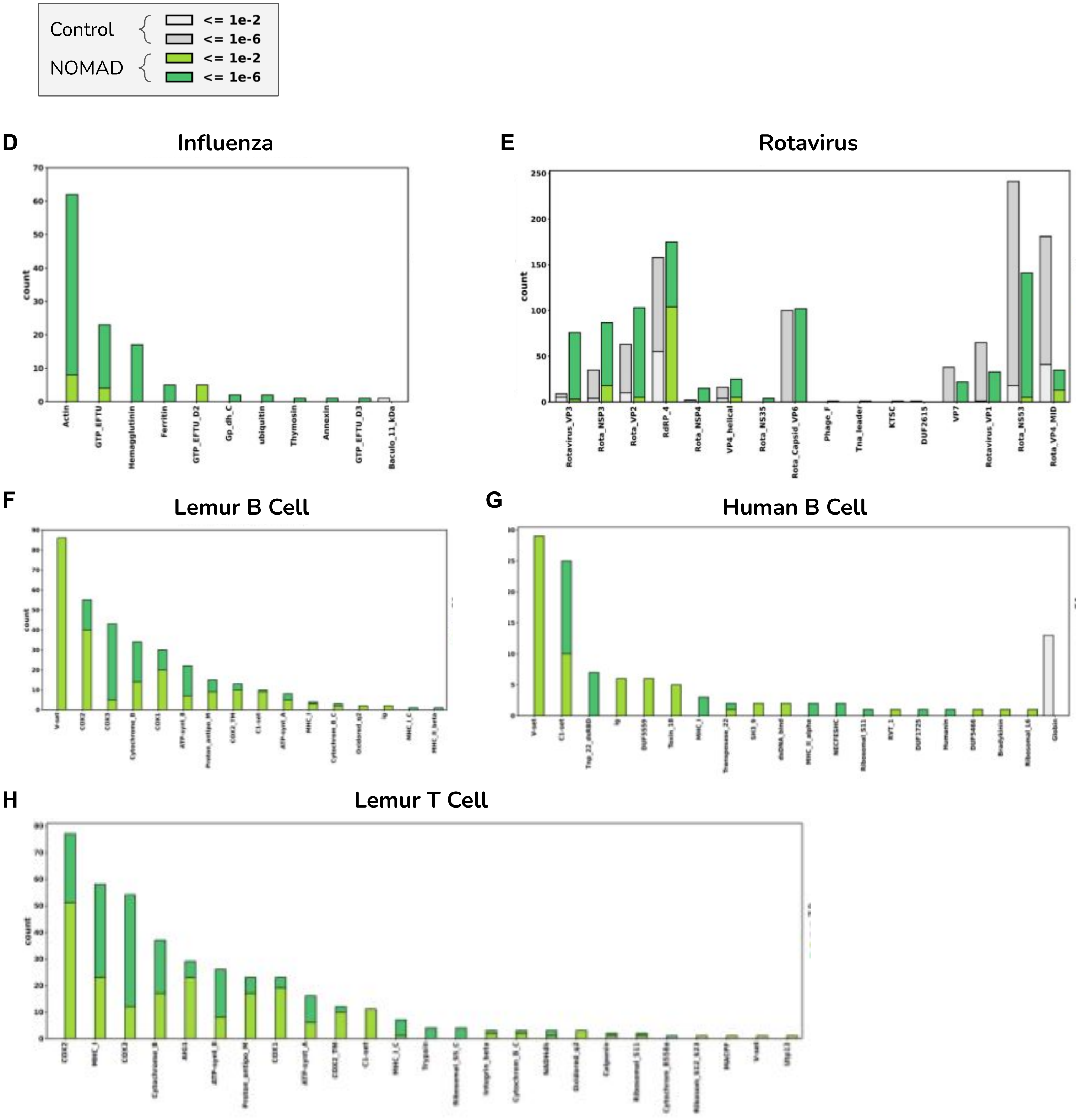

Control
<= 1e-2
<= 1e-6
NOMAD
<= 1e-2
<= 1e-6
D
Influenza
count
0
10
20
30
40
50
60
70
Actin
GTP_EFTU
Hemagglutinin
Ferritin
GTP_EFTU_D2
Gp_dh_C
ubiquitin
Thymosin
Annexin
GTP_EFTU_D3
Baculo_11_kDa
E
Rotavirus
count
0
50
100
150
200
250
Rotavirus_VP3
Rota_NSP3
Rota_VP2
RdRP_4
Rota_NSP4
VP4_helical
Rota_NS35
Rota_Capsid_VP6
Phage_F
Tna_leader
KTSC
DUF2615
VP7
Rotavirus_VP1
Rota_NS53
Rota_VP4_MID
F
Lemur B Cell
count
0
10
20
30
40
50
60
70
80
90
V-set
COX2
COX3
Cytochrome_B
COX1
ATP-synt_8
Proton_antipo_M
COX2_TM
C1-set
ATP-synt_A
MHC_I
Cytochrom_B_C
Oxidored_q2
Ig
MHC_I_C
MHC_II_beta
G
Human B Cell
0
5
10
15
20
25
30
V-set
C1-set
Trp_22_dsRBD
Ig
DUF5559
Toxin_18
MHC_I
Transposase_22
SH3_9
dsDNA_bind
MHC_II_alpha
NECFESHC
Ribosomal_S11
RVT_1
DUF1725
Humanin
DUF5466
Bradykinin
Ribosomal_L6
Globin
H
Lemur T Cell
count
0
10
20
30
40
50
60
70
80
COX2
MHC_I
COX3
Cytochrome_B
JIG1
ATP-synt_8
Proton_antipo_M
COX1
ATP-synt_A
COX2_TM
C1-set
MHC_I_C
Trypsin
Ribosomal_S5_C
Integrin_beta
Cytochrom_B_C
NADH4
Oxidored_q2
Calponin
Ribosomal_S11
Cytochrom_B558a
Ribosom_S12_S23
MACPF
V-set
Ubp13

Fig. S2: NOMAD protein profile hits to the Pfam database (greens) and control (greys); ordered by enrichment in NOMAD hits compared to control; all NOMAD anchors were used as input, without effect size filters.

A. Protein profile analysis of NOMAD significant anchors from the original South African genomic surveillance study (SRP348159) that identified the Omicron strain during the period 2021-11-14 to 2021-11-23 (*19*)
B. Protein profile analysis of NOMAD significant anchors from France data, Oropharyngeal swabs from patients with SARS-CoV-2 from 2021-12-6 to 2022-2-27 in France (SRP365166), a period of known Omicron-Delta coinfection (*17*).
C. Protein profile analysis of NOMAD significant anchors from California data (SRR15881549), before viral strain divergence in the spike had been reported (*22*) serving as a negative control.
D. Protein profile analysis of NOMAD significant anchors from influenza-A data (SRP294571).
E. Protein profile analysis of NOMAD significant anchors from rotavirus breakthrough cases (SRP328899).
F. Protein profile analysis of NOMAD significant anchors from *Microcebus* spleen B cells, from the Tabula Microcebus consortium.
G. Protein profile analysis of NOMAD significant anchors from human T cells from donor 1, from the Tabula Sapiens consortium.
H. Protein profile analysis of NOMAD significant anchors from *Microcebus* natural killer T cells from the Tabula Microcebus consortium.

S3

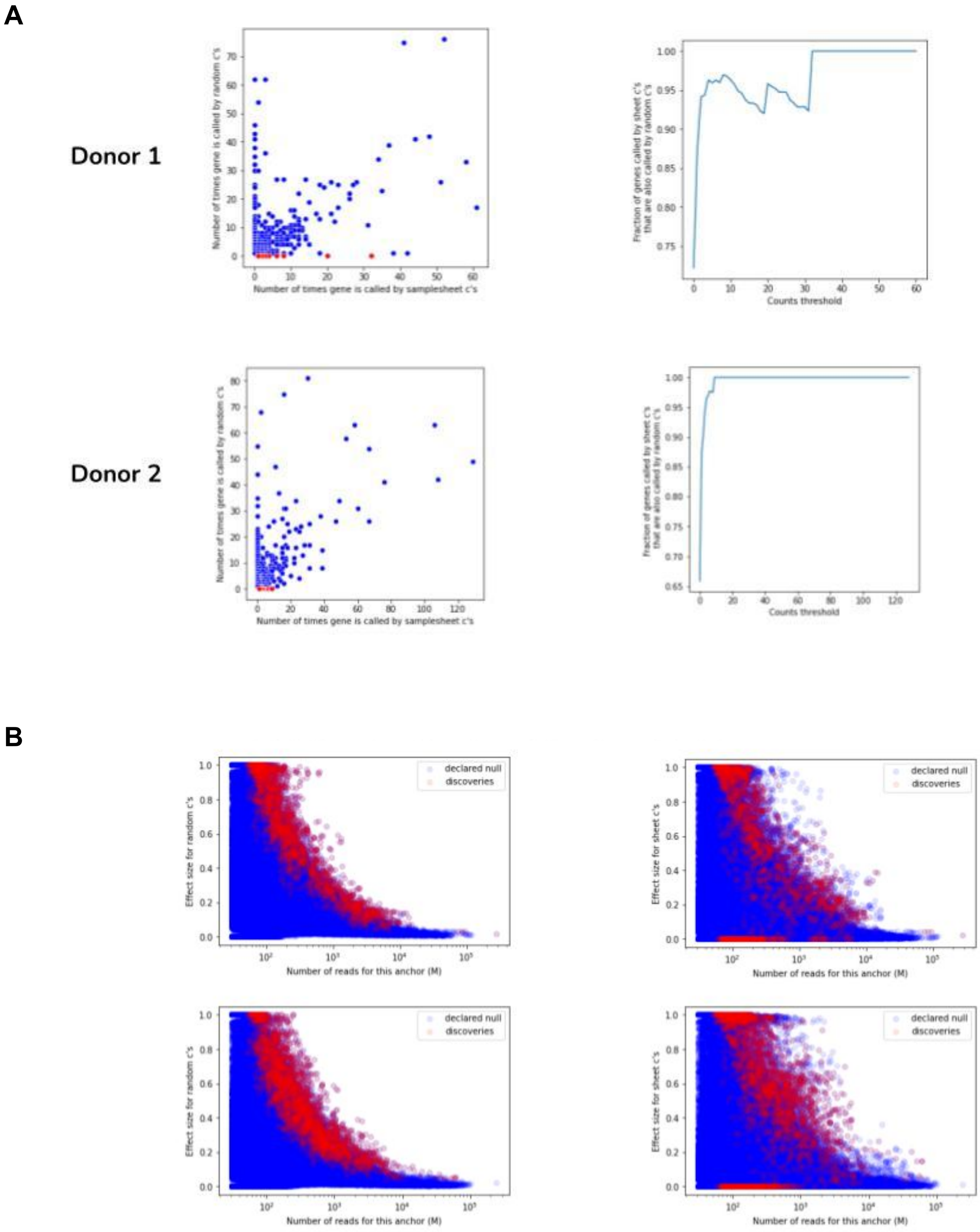

A
Donor 1
Number of times gene is called by random c's
Number of times gene is called by samplesheet c's
Fraction of genes called by sheet c's that are also called by random c's
Counts threshold
Donor 2
Number of times gene is called by random c's
Number of times gene is called by samplesheet c's
Fraction of genes called by sheet c's that are also called by random c's
Counts threshold
B
Effect size for random c's
Number of reads for this anchor (M)
declared null
discoveries
Effect size for sheet c's
Number of reads for this anchor (M)
declared null
discoveries
Effect size for random c's
Number of reads for this anchor (M)
declared null
discoveries
Effect size for sheet c's
Number of reads for this anchor (M)
declared null
discoveries

Fig. S3: Analysis of significant anchors in HLCA.

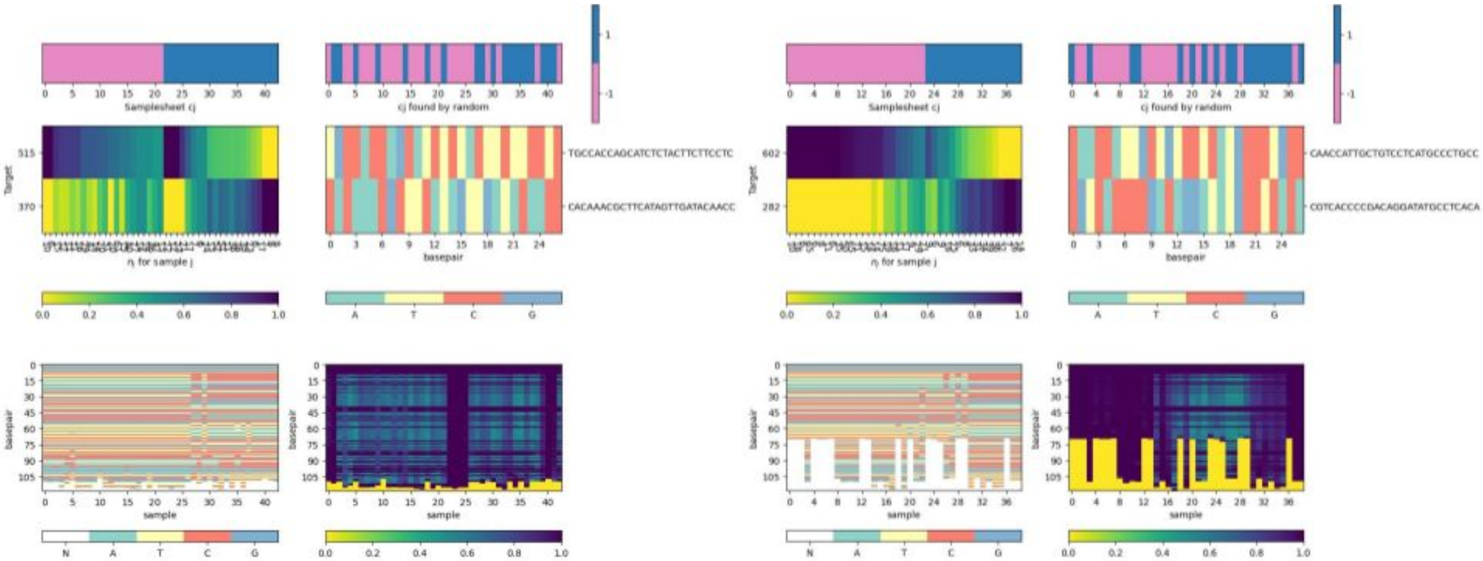
A. Random c's can recover samplesheet c's. For the HLCA dataset, of the 3439 anchors (1384 genes) called by the input metadata (samplesheet c's) in donor 1 (BY correction, alpha=.05), we have that 72% of the genes called were also called by NOMAD's selection of random c's (6287 called by anchors by random c's, 2268 genes). Left plot indicates for each gene (dot) how many times it was called by samplesheet c's vs random c's. Red dots indicate those genes not called by random c's. On the right plot we have the fraction of genes that are called at least x times by samplesheet c's that are also called by random c's. We see that for x=2 (i.e. all genes hit by at least 2 anchors), random c's call >94% of those genes called by samplesheet c's.

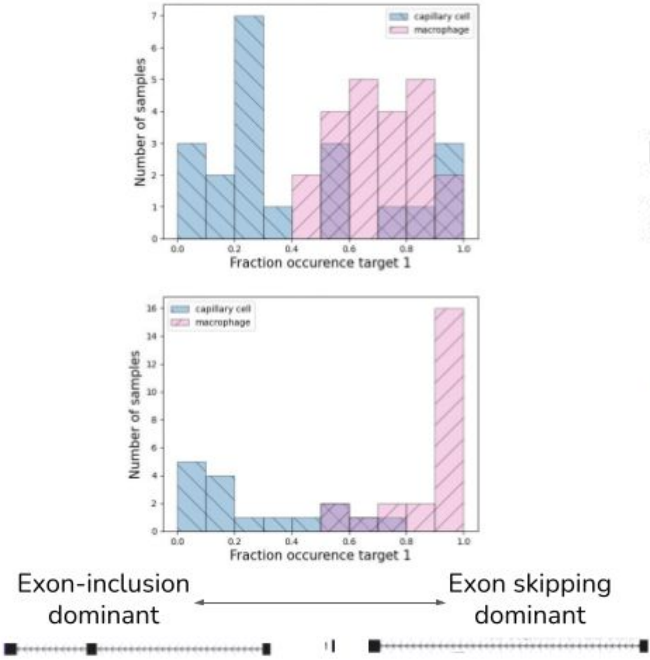
   For donor 2 similar results are observed, with 3775 (5619) anchors from samplesheet c's and 1125 (1844) genes for samplesheet c's (random c's) respectively. >90% of samplesheet c discoveries for x=2, >94% for x=3.

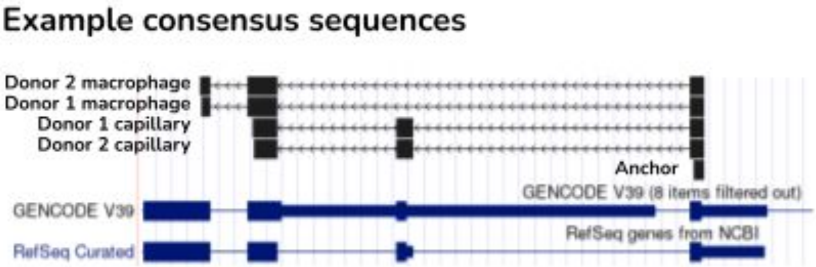
B. Effect size plotted against number of reads for HLCA dataset for donor 1 (top row) and donor 2 (bottom row), macrophage (left) and capillary cells (right).

S4

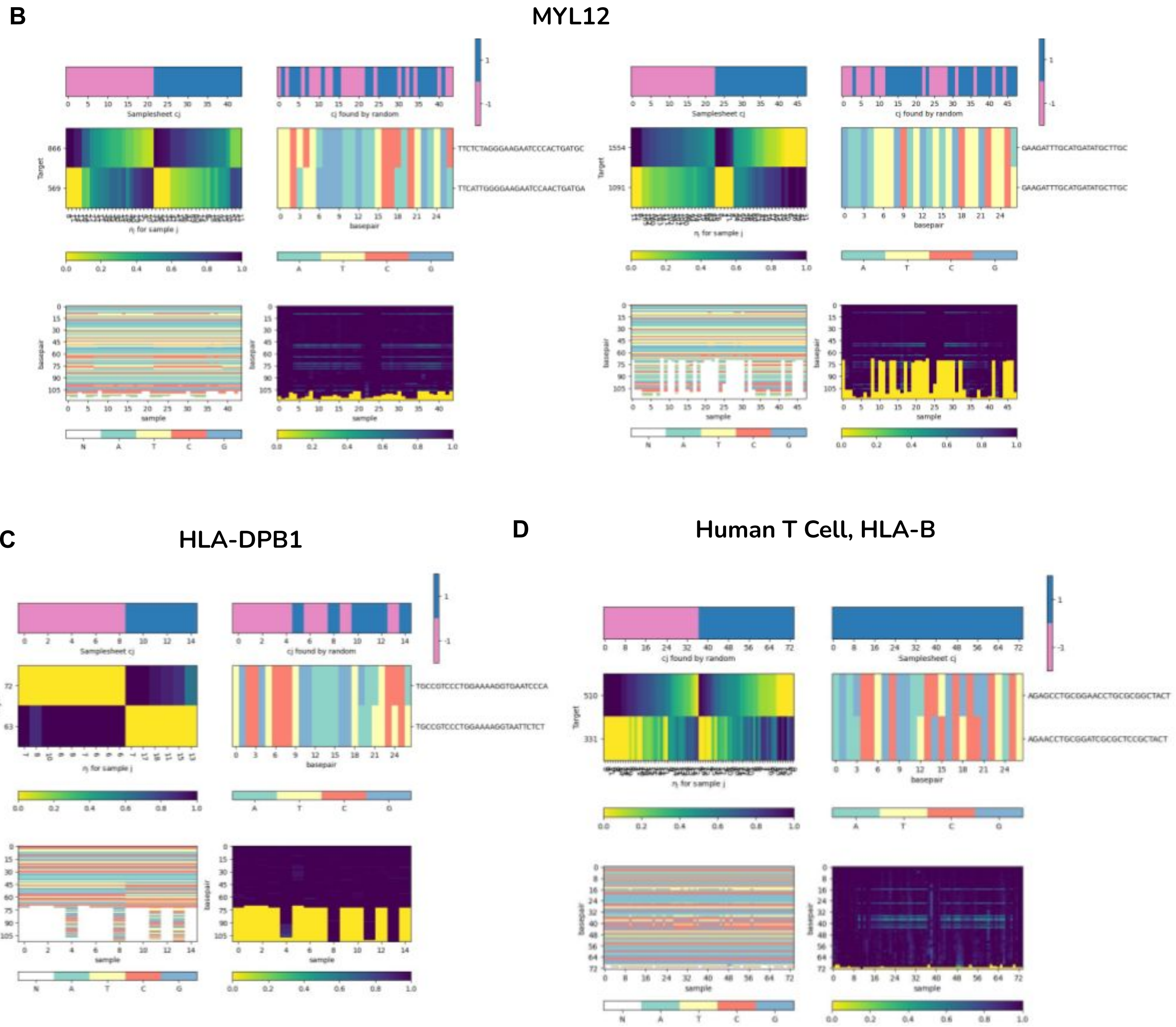
A **MYL6**


Samplesheet cj
cj found by random
1
-1
Target
515
370
602
282
TGCCACCAGCATCTCTACTTCTTCCTC
CACAAACGCTTCATAGTTGATACAACC
CAACCATTGCTGTCCTCATGCCCTGCC
CGTCACCCCGACAGGATATGCCTCACA
r_j for sample j
basepair
0.0
0.2
0.4
0.6
0.8
1.0
A
T
C
G
N
sample


capillary cell
macrophage
Number of samples
Fraction occurence target 1
Exon-inclusion dominant
Exon skipping dominant


Example consensus sequences
Donor 2 macrophage
Donor 1 macrophage
Donor 1 capillary
Donor 2 capillary
Anchor
GENCODE V39 (8 items filtered out)
GENCODE V39
RefSeq genes from NCBI
RefSeq Curated

S4



B MYL12

C HLA-DPB1

D Human T Cell, HLA-B

Fig. S4: Sample heatmaps.

A. NOMAD detects anchors in MYL6, a positive control. Q value of 1.4E-8 for donor 1, 5.8E-41 for donor 2. Consensus split-read mapping shows capillary cells dominantly include and macrophage cells skip exon the exon in MYL6.

Heatmaps show the complete data for the called anchors. Each set of heatmaps is for one anchor sequence. The primary plot is the center left one, which shows the samples x targets contingency table. Each column represents a sample, and each row represents a unique target. The color indicates what fraction of the sample's (column's) targets come from the target corresponding to that row. The x-ticks correspond to n_j, the number of times the anchor was observed in this sample. The y-ticks indicate the number of times this target appeared (following this anchor), and the targets are sorted by abundance. The two top plots indicate the c_j's used; when metadata c_j's are available (from the samplesheet), they will be in the upper left, and the optimizing random c_j's will be in the upper right.

The middle left plot is used to visualize the targets that follow this anchor. Each row represents a target (sequence given in y-tick) corresponding to the row to the left of it in the contingency table. The columns are base pair positions along the sequence of each target. Each nucleotide is color-coded, to show the similarity of the targets (e.g. to indicate whether they differ by a SNP, deletion, alternative splicing, etc).

The two bottom plots relate to the consensus sequences. The lower left plot shows the nucleotide sequence (same color scheme as the center right one for the targets). Each column corresponds to the consensus sequence for the sample of the same column above it in the contingency table. The rows are base pair positions along each consensus. These consensus sequences are variable length, and a value of 0 (yellow color) on the bottom of a sequence indicates that the consensus has ended. The bottom right plot shows the fraction agreement per nucleotide within a sample with its consensus sequence. We can see that for samples where only one isoform / SNP is expressed the consensus stays near 100%, while for samples with a diverse set of targets the consensus is less uniform.

B. MYL6

C. MYL12

D. HLA-DPB1

E. Human T cell, HLA-B

Data S1: Protein domain analysis

For SARS-CoV-2 datasets, we use significant NOMAD anchors meeting the effect size requirement of <0.8 as input anchors; for remaining datasets, up to the top 1000 significant NOMAD anchors are used as input anchors. For all datasets, we match the number of control anchors to NOMAD anchors, taking the most abundant anchors. Input anchors were assessed for protein homology against the Pfam database. The resulting 'raw' .tblout outputs were then processed, keeping the best hit (based on E-value) per each initial anchor, and any hits with an E-value better than 0.01 were parsed into an *_nomad.Pfam (or *_control.Pfam) file used for subsequent plotting.

Data S2: Significant anchors

Tables containing significant anchors, anchor statistics, and C_j used for each sample.

Data S3 : Additional summary tables

Tables containing significant anchors, their targets, anchor statistics, anchor and target reverse complement information, highest priority element annotations for anchors and targets, anchors annotations, and consensus annotations.

Data S4: Anchor genome annotations

Tables containing significant anchors, and their genome and transcriptome annotations.