A statistical reference-free algorithm subsumes and generalizes common genomic sequence analysis and uncovers novel biological regulation

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

Affiliations:

[1]Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA.
[2]Department of Electrical Engineering, Stanford University, Stanford, 94305, USA.
[3]Department of Biochemistry, Stanford University, Stanford, 94305, USA.
[4]Department of Statistics (by courtesy), Stanford University, Stanford, 94305, USA.
*Corresponding Author. Email: julia.salzman@stanford.edu
† Co-first authors

Summary: We show that myriad, disparate mechanisms that diversify genomes and transcriptomes can be captured by a unifying principle: sample-dependent sequence variation. This variation occurs in both RNA and DNA and functions to regulate transcript expression and adaptation. Using this insight, we develop a novel highly efficient algorithm – NOMAD – that performs inference on raw reads without any genomic reference or sample metadata. NOMAD unifies data-scientifically driven discovery with previously unattainable speed and generality. Examples include SARS-CoV-2, humans, and non-model animals and plants with both bulk and single cell RNA-sequencing data. A snapshot of its novel discoveries include missing variants in SARS-CoV-2, gene regulation in diatoms epiphytic to eelgrass, an oceanic plant critical to the carbon cycle and significantly impacted by climate change, and in octopus where it identifies isoform regulation in genes missing from the reference. NOMAD is a new unifying approach to sequence analysis that enables expansive discovery.

One-sentence summary: We present a unifying, reference-free formulation of disparate genomic problems bypassing reference genomes.

Introduction

Sequence variation – mutation, reassortment or rearrangement of nucleic acids – is fundamental to regulating gene expression, and to evolution and adaptation across the tree of life. Consider the sequence composition of two recently diverged genomes, such as the emergence of two viral strains. In a typical example, the vast majority of the two emergent strains' genomic sequence is shared, and the strains only differ in sequence at a few strain-defining locations. These sequence differences could be point

mutations, small or large insertions or deletions, or other rearrangements, including those as a consequence of mobile elements. This toy example illustrates a fundamental feature of a set of genomes that have shared ancestral genetic material and have since undergone selection. Their genomes have regions of sample-dependent sequence diversification: a shared identical sequence neighbors one that is diverse between samples. This feature characterizes genomes in much more generality than two viral strains. It includes any number of polymorphic members of a population including where genetic variation associated with disease or phenotype, mobile elements themselves (Abante, Wang and Salzman, 2022) or their genomic targets, loci that are rearranged or hypermutated (Medhekar and Miller, 2007) within an organism such as V(D)J recombination and somatic hypermutation, among many others.

Now consider the sequence composition of an alternatively spliced RNA transcript with a cassette exon, exon 2, that is differentially included in different cells (samples). Again, these transcripts have the property that the sequence of exon 1 will be a constant common sequence to all transcripts, but the sequence composition downstream of it will differ depending on if exon 2 is skipped: if it is, the sequence downstream will represent exon 3, and if not, exon 2. Thus, the set of sequences of these transcripts fit within the sequence signal that NOMAD will detect.

Again, this feature of a constant sequence juxtaposed next to one with a sequence composition that is sample-dependent characterizes a regulated transcriptome more generally, and can detect any sequences sample-specifically impacted by RNA editing, RNA splicing, alternative polyadenylation or allele-specific expression. Moreover, many other genomic problems across different ‘omic measurement designs can be mapped to a feature NOMAD will detect: one where a constant sequence flanks a set of variable sequences with sequence-dependent biological function (see Supplement).

Mapping the detection of disparate biological problems to a common model is both conceptually appealing but also allows us to design a single algorithm that solves disparate genomic problems including those described above, called NOMAD. NOMAD 1) provides a single unifying solution to disparate problems in genomics; 2) enables statistical inference for sequences of candidate biological function that bypasses reference genomes completely, operating directly on raw reads; 3) prioritizes and identifies biologically regulated sequences enabling biologists to focus on sequences with differential regulation across samples. Here we explain the concepts behind NOMAD and apply them to biological examples where reference genomes are well curated and domains where they are not. In both cases, NOMAD expands the scope of inference possible today and promises to be of broad use for genomic discovery in most domains where it is applied. Practitioners of genomics are familiar with how time consuming, and many times ad hoc, these procedures can be. Three main problems with the existing paradigm in genomics are highlighted below.

The first problem is fundamental: specifying and relying on a reference genome can be highly limiting for biological inference, and generating and or aligning to a reference is foundational in the field of genomics today. Alignment-based methods struggle when sequences can map to multiple or repetitive locations, and or reference genomes are incorrect or incomplete (Shi *et al.*, 2022; Zhao, Shi and Pollard, 2022). These regions comprise ~54 % of the human genome (Nurk *et al.*, 2022) and are sometimes among the most important to analyze. This means users must supervise running and make subjective interpretation at many steps, including subjective decisions about which statistical tests to perform on the output: eg. whether to run a statistical test for detecting SNPs, or alternative splicing, or some other test. This creates several major impediments to discovery.

In human genomics, workflows beginning with reference alignment miss important variation absent from assemblies, even with pangenomic approaches. We call methods that begin with alignment, “reference-first”. It is well-known that genetic variants associated with ancestries of under-studied populations are poorly represented in databases (Sherman *et al.*, 2019), resulting in health disparities (West, Blacksher and Burke, 2017). In disease genomics (Wang *et al.*, 2022), sequences of pathogenic cells may be missing from reference genomes, or no reference genome exists, as in genomically unstable tumors (C.-Z. Zhang *et al.*, 2015). Many species do not have reference genomes even when they in principle could be attained due to logistical and computational overheads. In these situations, alignment-first approaches fail. Practitioners of human genomics can see more detailed and further examples of the insights from NOMAD’s application to human single cell RNA-seq data as well as its improved precision compared to state of the art algorithms (Dehghannasiri *et al.*, 2022).

In viral surveillance, reference-first approaches are even more problematic (The Nucleic Acid Observatory Consortium, 2021). Viral reference genomes cannot capture the complexity of viral quasispecies (Kirkegaard, van Buuren and Mateo, 2016) or the vast extent of polymorphism (Kim *et al.*, 2020). New viral assemblies are constantly being added to reference databases (Edgar *et al.*, 2022; Zayed *et al.*, 2022). In the microbial world, pre-specifying a set of reference genomes is infeasible due to its inherent rapid genomic changes. References also cannot capture insertional diversity of mobile elements, which have significant phenotypic and clinical impact (Evans *et al.*, 2020) and are only partially cataloged in references (Wright, Comeau and Langille, 2022). Finally, reliance on mapping is also statistically problematic as it reduces raw, observed sequence data to a set of positions in the genome, and introduces errors both due to intrinsically randomized procedure of modern mappers (Langmead *et al.*, 2009; Dobin *et al.*, 2013) and at the very least, and computational procedure introduces some uncertainty.

A second issue with the existing paradigm in genomics is that the practitioner must choose a specific bioinformatic procedure: in an RNA-seq experiment, is the

analyst interested in V(D)J recombination, RNA editing or splicing (or others)? No current single algorithm can detect all events. Consider transposable elements, known to drive evolution and cause an unascertained number of human diseases (Pascarella *et al.*, 2022). A custom and multi-step workflow is required to detect these insertions because they are highly polymorphic, and the algorithm must address the issues of multi-mapping and the high degree of repetitive sequence (Shi *et al.*, 2022; Zhao, Shi and Pollard, 2022). Similarly, if V(D)J recombination is present in a sample, but a custom workflow is not specified to detect these rearrangements, they will not be reported. Further examples abound in plant and microbial genomics (Michael and VanBuren, 2020).

A final issue with genomic algorithms today is that the statistical inference they provide is based on summaries generated from alignment, which necessarily introduce noise. Technically, statistical inference is conditional on aligners, resulting in poorly calibrated p-values. More than a purely statistical issue, this problem can significantly impact biological discovery. Because statistical analysis in reference-first approaches is conditional on the output of the noisy alignment step, complicated null distributions for test statistics must be derived, which is sometimes infeasible in closed-form. It is then difficult and in some cases impossible (Chung and Romano, 2013) to provide valid statistical significance levels in downstream analyses, such as in differential expression analysis. When null distributions are not available, computationally intensive resampling is required, increasing the computational burden by orders of magnitude. Even when employed, these procedures report discoveries based on a nominal FDR, typically based on permutation distributions, there is no guarantee that these values are accurate. In fact have been repeatedly shown to be wildly inaccurate, sometimes yielding an FDR of .5 when an FDR of .05 was desired (Romano, Sesia and Candès, 2019). Together, these three points, along with the extensive manual time modern bioinformatics consumes, there is a convincing argument to attempt to develop an alternative to the way genomics is approached today.

NOMAD makes important inroads in overcoming these three significant problems in genomics. Moreover, it presents a unified way to address one of the foremost challenges for biological discovery in genomics: to identify sample-dependent sequence variation, the variation most likely to be regulated and functional. NOMAD both recovers such true positive variation in diverse biological contexts, from viral genomes to human transcripts, and discovers new biology. We illustrate a downstream computational procedure for interpretation and novel discovery on NOMAD's outputs in a set of selected examples ranging from highly scrutinized (SARS-CoV-2 and human) to studies eelgrass, an ocean-dwelling plant of great importance to carbon cycling which have comparatively little genomic study (Jueterbock *et al.*, 2021), and of *Octopus bimaculoides*, which has a recently assembled genome where we use NOMAD to illuminate transcriptional regulation missing from the reference. As a snapshot of these

new discoveries, NOMAD predicts un-annotated genomic variants in the well-studied delta and Omicron strains of SARS-CoV-2, uncovers novel cell-type regulation of gene paralogs and cell-type specific expression of alleles in the major histocompatibility locus. In mouse lemur, NOMAD discovers T cell receptor rearrangements impossible to identify with other tools. Finally, we illustrate further NOMAD's novel regulatory biological discovery using reads with no partial mapping to reference genomes in *Octopus bimaculoides* and the eelgrass *Zostera marina*. While we present specific examples in this paper, NOMAD's approach is general and should be of wide interest for theorists and practitioners across biological disciplines.

NOMAD is a statistics-first approach to identify sample-dependent sequence variation

We first show that multitudes of genomic analyses reduce to detecting sample-dependent sequence variation, and admit natural probabilistic formulations. History suggests that formulating scientific problems mathematically can radically improve the precision of inference, discovery and computational efficiency (ref. eg population genetics or compressive sensing) (Donnelly and Tavaré, 1995; Candes and Wakin, 2008). Thus, approaching genomics through this fundamental probabilistic formulation is naturally expected to yield more general, efficient and powerful solutions. Here, we leverage the probabilistic formulation above to derive a statistical test on raw sequencing read data (e.g. FASTQ files) to discover sample-dependent sequence variation completely bypassing references or ontology. We implement the test in a highly efficient algorithmic workflow, providing novel discovery across disparate biological disciplines (Fig. 1B,C).

To begin, we define an "anchor" *k*-mer in a read, and define an anchor to have sample-dependent diversity if the distribution of *k*-mers starting *R* base pairs downstream of it (called "targets") depends on the sample (Fig. 1E). This definition can be generalized to different constructions of anchors and targets, and can be applied to diverse data sources including DNA, RNA and protein sequence. The length of the anchor and target (k) are general. For intuition, suppose k is long enough an anchor of length k is found in a unique position in a genome, though this is not necessary and in fact sometimes not desired. In this case, suppose the anchor is found in a sequence adjacent to one which has a strain-specific mutation (Fig. 1C). Then, target counts following this anchor quantify the sequence variants, up to errors in sampling such as introduced in library preparation or sequencing. NOMAD's null states that each sample has the same distribution of variants immediately following this anchor. This framework detects sample-specific sequence variation including a broad range of biological features from mutations or indels (an anchor immediately preceding these variants, as in Fig. 1C), differentially regulated alternative splicing (an anchor in the constitutive exon), V(D)J recombination, RNA editing and many other features, for example

CRISPR repeats (Supplement, work in preparation). While anchors were chosen of length k=27 for biological inference in this manuscript, genomic inference is similar under a range of parameter choices (Methods, Kokot et al, in preparation).

The above is a unifying formulation that allows expansive numbers problems to be studied (Supplement) and allows us to develop a novel statistics-first approach, NOMAD (Novel multi-Omics Massive-scale Analysis and Discovery), that is reference-free and operates directly on raw sequencing data. It is an extremely computationally efficient algorithm to detect sample-dependent sequence variation, through the use of a novel statistic of independent interest that provides closed form p-values, meaning no permutation, resampling or numerical solutions to complex likelihoods are required for significance tests (Methods). Arguably, the most important reason NOMAD can reveal new biology is that all sequences predicted to have sample-dependent diversity are called without access to any reference and annotation information, though they can be optionally used for *post-facto* interpretation. This key property makes NOMAD fundamentally different from existing methods which critically rely on reference mapping prior to statistical inference. To illustrate, as a special case NOMAD automatically detects differential isoform expression. In this domain, the closest approach to NOMAD is Kallisto (Bray *et al.*, 2016). However, Kallisto still requires a reference transcriptome and statistical resampling for inference, and is further challenged to provide accurate expression estimates for more than a handful of paralogous genes and isoforms due to model limitations, specifically parameter identifiability. Further, unlike NOMAD, Kallisto cannot discover spliced isoforms *de novo.*

Summary of NOMAD's technical innovation

NOMAD is by default an unsupervised algorithm that does not require any sample identity metadata, though user-defined metadata can be optionally used (Methods). Significance testing is detailed in the Methods. At a high level, NOMAD tests whether sequences have evidence of sample-dependent sequence variation: an anchor is called significant if our statistical test rejects the hypothesis that for a sequence a, the conditional distribution of observing a target sequence t a distance R downstream of a is sample-independent (Fig. 1E). Significance testing is performed by finding approximate best splits of the samples into two groups (Methods). If a user desires, the test can be restricted to user-defined groups. Anchors are reported with an effect size taking values in [0,1]. This value measures how well partitioned target sequences are between sample groups: 0 if the groups have no difference in target distributions and increasing to 1 when the target distributions of the two groups are disjoint. NOMAD has significant power in detecting two-group type deviations, e.g. mutations between one group of Omicron samples and one group of Delta samples, but also has substantial power against more complex alternatives that involve combinatorial expression, as we show

with detection of V(D)J recombination. The statistical basis for this power is discussed in greater detail in a separate statistical work (Baharav, Tse and Salzman, 2023).

NOMAD has multiple major technical innovations: 1) a parallelizable, fully containerized, and computationally efficient approach to parse FASTQ files into tables that summarize joint kmer composition, 2) a novel statistical analysis of the derived tables, using concentration inequalities to derive closed form p-values, 3) a micro-assembly-based consensus sequence representing the dominant error-corrected sequence, similar to (Magoč and Salzberg, 2011; Motahari *et al.*, 2013; Edgar, 2016), downstream of the anchor for post-facto interpretation and identification of SNPs, indels or isoforms, to name a few (Fig. 1D). If post anchor-identification inference is desired, the consensus, rather than the raw reads, is aligned. This reduces the number of sequences to align by ~1000x in real data.

NOMAD's novel approach yields an extremely computationally efficient implementation without code optimization. We ran NOMAD on a 2015 Intel laptop with an Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz processor, generating significance calls for single cell RNA-seq totaling over 10 million reads in only 1 hour 45 min. When performed on a compute cluster, the same analysis is completed in an average of 22.8 minutes with 750 MB of memory for 10 million reads. NOMAD provides a dramatic speed up over existing methods for *de novo* splicing detection and significance calls because it reduces the number of reads needed for optional alignment by orders of magnitude. Further, NOMAD bypasses resampling-based approaches for significance testing, such as permutation testing, which create further computational overhead.

NOMAD discovers sequence variation in proteins at the host-viral interface without a genomic reference

We first show that NOMAD automatically detects viral strain evolution without any reference genome, or knowledge of the sample origin. Existing approaches are computationally-intensive, require genome assemblies and rely on heuristics. It is still difficult to call strain variation with them, and impossible to know how many strains are currently missed, though estimates suggest orders of magnitude remain to be discovered (Edgar *et al.*, 2022).

Yet, emergent viral threats or variants of concern, e.g. within SARS-CoV-2, will necessarily be absent from reference databases. Because virus' genomes are under selective pressure to diversify when infecting a host, NOMAD should prioritize anchors near genome sequences that are under selection, in theory, based purely on their statistical features: sequences flanking variants that distinguish strains have consistent sample-dependent sequence diversity. When NOMAD is run on patients infected with varying relative titers of Omicron and Delta strains of SARS-CoV-2, significant anchors are expected to be called adjacent to strain-specific mutations (Fig. 2A); they should be, and as we show are, discoverable without any knowledge of a reference. NOMAD

rediscovers known variants without a reference genome, overcoming significant challenges in the field of viral genomics.

To test this prediction, we analyzed oropharyngeal swabs from patients with SARS-CoV-2 from 2021-12-6 to 2022-2-27 in France, a period of known Omicron-Delta coinfection (Bal *et al.*, 2022). We ran NOMAD and analyzed anchors with high degree of target partitioning by sample measured by a per-anchor quantity we call effect size: high effect sizes are predicted if samples can be approximately partitioned by strain, although results are similar without this threshold (here, effect size>0.5 was used). NOMAD's statistical test identifies biologically significant anchors, as shown by comparing NOMAD's significant anchors against a set of control anchors, which are generated as the most abundant anchor sequences (Methods). We perform protein domain analysis independent of any reference genome to illustrate reference-genome-free interpretation of NOMAD's calls. For each anchor, we assigned a protein domain label based on *in silico* translation of its consensus sequence (Mistry *et al.*, 2021). The protein domain with the best mapping to the Pfam database is assigned to the anchor, producing a set of "protein profiles" (Methods, Supplementary Table 1). NOMAD protein profiles are significantly different from controls (p=3.7E-7, chi-squared test, Fig. 2B). The most NOMAD-enriched domains are the receptor binding domain of the betacoronavirus spike glycoprotein (7 NOMAD vs 0 control hits, p=2.3E-2 hypergeometric p-value, corrected) and Orf7A a transmembrane protein that inhibits host antiviral response (6 NOMAD vs 0 control hits, p=6.4E-2 hypergeometric p-value, corrected). All analysis is blind to the data origin (SARS-CoV-2).

We further analyzed patient samples from the original South African genomic surveillance study that identified the Omicron strain during the period 2021-11-14 to 2021-11-23 (Viana *et al.*, 2022), again without metadata or reference genomes and directly on input FASTQ files (Methods). NOMAD protein profiles were significantly different from controls (chi-squared test, p=2.5E-39, Fig. 2B). NOMAD-enriched domains in France and South African are highly consistent: the top four domains in both datasets are permutations of each other. The most NOMAD-enriched domains versus controls were the betacoronavirus S2 subunit of the spike protein involved in eliciting the human antibody response (Poh *et al.*, 2020) (23 NOMAD vs 2 control hits, p=2.9E-6 hypergeometric p-value, corrected), the matrix glycoprotein which interacts with the spike (20 NOMAD vs 0 control hits, p=8.4E-8 hypergeometric p-value, corrected), and the receptor binding domain of the spike protein to which human antibodies have been detected (Jörrißen *et al.*, 2021) (20 NOMAD vs 0 control hits, p=8.4E-8 hypergeometric p-value, corrected). All domains are biologically predicted to be under strong selective pressure, and thus exhibit sample-specific sequence diversity; NOMAD discovers this *de novo*.

We aligned NOMAD anchors with high effect size (Supplement) to the Wuhan, Delta, and Omicron BA.1 and BA.2 reference strains to test if NOMAD anchors

rediscovered known strain-defining mutations. We defined an anchor to recover a strain defining mutation if any of its 2 targets (>5% of total anchor counts) mapped to different strains (Supplement). NOMAD anchors are significantly enriched for being mutation-consistent: in the French data, 252 anchors are significant and have effect size >.5. Of the 83 anchors that also map to a reference and have a target mapping to a reference, 69.9% (58/83) of NOMAD's calls were strain defining versus 5.6% (14/252) in the control (p=1.3E-31, hypergeometric test). Anchors that are not classified as strain-defining include examples of unannotated mutations (Fig). For the South African data, 201 anchors are significant and have high effect size. 26% (34/130) of NOMAD's called anchors were mutation consistent vs 3.0% (6/201) for the control (p=3.1E-10, hypergeometric test). This is due to the low prevalence of Omicron in the dataset, where we see that further restricting the set of anchors to those occurring in at least 5 samples, 93% (14/15) of the anchors are strain-defining.

Examples of strain-defining anchors detected by NOMAD are presented in Fig. 2 where metadata is used post-facto to visualize sample expression of each strain-defining target. Examples include anchors in the membrane (Fig. 2C, top) and spike protein (Fig. 2C, middle and bottom). Differences between consensus sequences and Wuhan reference illustrate NOMAD's unsupervised rediscovery of annotated strain-specific variation, including co-detection of a deletion and a mutation (Fig. 2C). While the Wuhan reference genome was used for *post-facto* interpretation, for emphasis: no alignment or sample metadata was used to generate NOMAD's calls, only to interpret them (Fig. 2). NOMAD consensuses also extend discovery (Fig. 2C), identifying strain-specific variation beyond the target: both are annotated Omicron variants (Fig. 2C). NOMAD's statistical approach automatically links discovered variants within patients *de novo*: one consensus contains the Omicron Variant of Concern (VOC) T22882G; a second consensus has a single VOC T22917G identified in Omicron strains BA.4 and BA.5 in May of 2022, 3 months after the analyzed samples were collected; a third consensus contains the VOC as well as the VOC G22898A; a single further consensus shows no mutations, consistent with Delta infection. Together, this suggests that mutations in BA.4 and BA.5 were circulating well before the VOC was called in May 2022. This implies that the VOC could have been automatically detected by NOMAD before it was present in reference annotations. A large fraction of called genetic variants exist outside of currently annotated SARS-CoV-2 variants, a subject of future work (Supplement). One such example is a 6 basepair deletion in the spike protein whose presence is perfectly predicted by whether a patient (sample) was infected with Delta or Omicron (Fig. 2D).

We further analyzed 499 samples collected in California (2020) before viral strain divergence in the spike had been reported (Gorzynski *et al.*, 2020) as a negative control. No enrichment of NOMAD protein profiles related to the spike or Orf7a domains were observed (Supplementary Fig. 2B), supporting the idea that NOMAD calls are not

false positives. To explore the generality of NOMAD for reference-free discovery in other viral infections, we additionally ran NOMAD on a study of influenza-A and of rotavirus breakthrough cases (Supplemental Fig. 2A,B, Methods). NOMAD Protein profile analysis showed enrichment in domains involved in viral suppression of the host response and regulated alternative splicing (Supplement). Together, these datasets suggest that NOMAD analysis could aid in viral surveillance, including detecting emergence of variants of concern directly from short read sequencing (Jacot *et al.*, 2021).

NOMAD discovers paralog-specific expression in single cell RNA-seq

NOMAD is a general algorithm to discover sample-dependent sequence variation in disparate applications including RNA expression and beyond. To illustrate the former, we ran NOMAD without any parameter tuning on single cell RNA-seq datasets, testing if it could perform the fundamental but previously distinct tasks of identifying regulated expression of paralogous genes and alternative splicing (Fig. 3A). Each of these tasks is difficult, and each currently constitute sub-disciplines within bioinformatics. Algorithms to perform these tasks are time intensive to run and require significant intellectual resources for their application-specific design.

First, we tested if NOMAD discovers alternatively spliced genes in single cell RNA-seq (Smart-seq 2) data from human macrophage versus capillary lung cells, chosen because they have a recently established positive control of alternative splicing, MYL6, a subunit of the myosin light chain (Olivieri *et al.*, 2021). NOMAD rediscovered MYL6 and made new, reproducible discoveries not reported in the literature (Supplemental Table 3). For example, we discovered reproducible cell-type specific regulation of MYL12 isoforms, MYL12A and MYL12B. Like MYL6, MYL12 is a subunit of the myosin light chain. In humans (as in many species) two paralogous genes, MYL12A and MYL12B, sharing >95% nucleotide identity in the coding region, are located in tandem on chromosome 18 (Fig. 3B). Reference-first algorithms fail to quantify differential expression of MYL12A and MYL12B due to mapping ambiguity. NOMAD automatically detects targets that unambiguously distinguish the two paralogous isoforms, and demonstrates their clear differential regulation in capillary cells and macrophages (Fig. 3B). We confirmed MYL12 isoform specificity in pairwise comparisons of the same cell types in two further independent single cell sequencing studies of primary cells from the same cell-types (Supplement). MYL12 was recently discovered to mediate allergic inflammation by interacting with CD69 (Hayashizaki *et al.*, 2016); while today little is known about differential functions of the two MYL12 paralogs, the distinct roles of highly similar actin paralogs provides a precedent (Perrin and Ervasti, 2010; Vedula *et al.*, 2017).

Another significant challenge in human genomics is analyzing variation in the major histocompatibility (MHC) locus, including cell-type specific expression. The MHC

is one of the most polymorphic regions of the human genome and carries many significant disease risk associations (Matzaraki *et al.*, 2017). Despite its central importance in human immunity and complex disease, allotypes are difficult to quantify and have not and perhaps cannot be exhaustively annotated. Statistical methods to reliably distinguish allotype expression at single-cell level do not exist, while analysis of bulk data requires custom algorithms. Because the locus is polymorphic, standard RNA-seq alignment based methods consider reads from the MHC to be multimapped and may discard them (Dobin *et al.*, 2013; Bray *et al.*, 2016). NOMAD calls novel reproducible cell-type specific allelic expression and splicing in the MHC, discovering allele-specific expression of HLA-DRB1 (Fig. 3C) where macrophage and capillary cells preferentially express different alleles in donors 1 and 2, differing by an insertion. Across donors, the allele expressed is similar within the same cell-type. In another example of NOMAD's findings in the MHC, NOMAD predicts novel cell-type specific splicing which would change the amino acid and 3' UTR sequence of HLA-DPA1 in donor 2 (Methods, Fig. 3D).

These empirical results bear out a snapshot of NOMAD's discovery power. They validate the theoretical prediction that NOMAD has high statistical power to simultaneously identify isoform expression variation and allelic expression, including that missed by existing algorithms (Supplementary Table 3). Finally, while in this manuscript we focus on anchors that have aligned to the human genome, NOMAD makes many predictions of cell-type specific RNA expression that include sequences that map to repetitive elements or do not map to the human reference (Supplemental Methods). In donor 1, 53% (in donor 2 61%), of 4010 (resp. 4603) anchors map to the human genome and no other reference; 35% (resp. 30%) map to both the Rfam and human genome (Supplemental Figure 5); 6% and 7% have no map to any reference used for annotation which includes repetitive and mobile elements (Supplemental Methods). As an example, 9 and 18 anchors (donor 1 and 2 respectively) BLAST to MHC alleles in the NCBI database (Supplementary Table 3). Together, this data suggests that NOMAD enables automatic discovery of biologically important sequences outside of the human reference.

Unsupervised discovery of B, T cell receptor diversity from single-cell RNA-seq

B and T cell receptors are generated through V(D)J recombination and somatic hypermutation, yielding sequences that are absent from any reference genome and cannot be cataloged comprehensively due to their diversity (>10^{12} sequences) (Briney *et al.*, 2019). Detecting and quantifying V(D)J variation is critical to understanding the natural immune response from virus' to tumors (Gee *et al.*, 2018; Cao *et al.*, 2020; Grant *et al.*, 2021), (Gee *et al.*, 2018; Grant *et al.*, 2021), and can be used to engineer drugs. Existing methods to identify V(D)J rearrangement require specialized workflows that depend on receptor annotations and alignment; they thus fail when the loci are

unannotated or when sequences have variation with high distance from annotated references (Canzar *et al.*, 2017; Lindeman *et al.*, 2018). In many organisms, T cell receptor loci are incompletely unannotated as they must be manually curated (The Tabula Microcebus Consortium *et al.*, 2021). Further, V(D)J recombination will be missed in standard RNA-seq workflows. In summary, it remains difficult and time consuming to identify T cell receptor (TCR) and B cell receptor (BCR) variants, and algorithms still suffer from blindspots due to heuristics embedded in their pipelines.

To illustrate the power of NOMAD's general and reference-free statistical approach, we tested if NOMAD could identify TCR and BCR rearrangements in the absence of annotations. We ran NOMAD with identical parameters used in analysis of other data on 111 natural killer T and 289 B cells isolated from the spleen of two mouse lemurs (*Microcebus murinus*) profiled by Smart-Seq2 (SS2) (The Tabula Microcebus Consortium *et al.*, 2021), and performed the same analysis on a random choice of 50 naive B cells from the peripheral blood and 128 CD4+ T cells from two human donors profiled with SS2 (Tabula Sapiens Consortium* *et al.*, 2022) for comparison (Fig. 4A, Methods). For emphasis, no sample metadata was used; NOMAD was not provided a reference genome or any knowledge of the samples' biological origin or annotation. NOMAD's novel statistical test has power to detect combinatorial target expression where each sample has a unique or nearly unique target deriving from recombination and hypermutation.

NOMAD protein profiles (Fig. 4B) revealed that the most frequent hits in lemur B cell were IG-like domains resembling the antibody variable domain (86 hits), and COX2 (55 hits) a subunit of cytochrome c oxidase, known to be activated in the inflammatory response (Groeger *et al.*, 2010). NOMAD's top hits for Lemur T cell were COX2 and MHC_I (77 and 58 hits, respectively). Similar results were obtained for the human samples (Fig. 4B): even without pre-specifying a search for V(D)J rearrangement in B and T cells, NOMAD reidentified these loci. In addition, NOMAD generated novel predictions of cell-type specific allelic expression of HLA-B in T cells (Supplement) find statistical evidence that cells preferentially express a single allele of HLA-B ($p < 4{,}6E\text{-}24$); consensus analysis shows SNPs found by NOMAD de novo in HLA-B are in fact concordant with known positions of polymorphism (Fig. 4D).

We further predicted that BCR and TCR rearrangements would also be discovered by investigating the transcripts most hit by NOMAD anchors. We mapped NOMAD-called lemur B and T cell anchors to an approximation of its transcriptome: that from humans which diverged from lemur ~60-75 million years ago (Ezran *et al.*, 2017). Lemur B cell anchors most frequently hit the immunoglobulin light and kappa variable regions; lemur T cell anchors most frequently hit the HLA and T cell receptor family genes (Methods, Fig. 4C). Similar results are found in human B and T cells (Fig. 4E, Supplement). Transcripts with the most hits in the control were unrelated to immune function, showing that NOMAD identifies a signal not explained by k-mer abundance. To

illustrate NOMAD's statistical power and capacity for discovery, consider its comparison to existing pipelines. They cannot be run without the annotation for the lemur TCR locus; for assembling BCR sequences, pipelines e.g. BASIC (Canzar *et al.*, 2017) cannot always identify V(D)J rearrangement, including in some cells profiled in the lemur dataset (The Tabula Microcebus Consortium *et al.*, 2021). We selected the 35 B / plasma cells where BASIC could not programmatically identify variable gene families on the light chain variable region. NOMAD automatically identified anchors mapping to the IGLV locus, with consensus sequences that BLAST to the light chain variable region (Supplement). Together, NOMAD identifies sequences with adaptive immune function including V(D)J in both B and T cells *de novo*, using either no reference genome (protein profile analysis) or only an annotation guidepost from a related organism (human). In addition to being simple and unifying, NOMAD can extend discovery beyond what is available today with custom pipelines.

NOMAD uncovers novel biology in octopus and eelgrass

We applied NOMAD to two understudied domains of life to illustrate the breadth and generality of NOMAD's discovery potential as well as its broad potential interest to biologists. First, we analyzed transcriptomic data from the nervous system of the non-model organism *Octopus bimaculoides* to illustrate transcript regulation missed by reference-first analyses. Second, we analyzed transcriptomic data from an ecological study of the seagrass species *Zostera marina* which is critical to carbon cycling and has a rapidly changing habitat due to climate change (Wilson and Lotze, 2019). While adaptation to climate change is critical, the mechanisms underlying its adaptation to dynamic temperatures and daylight ranges are unknown. We outline a general workflow that shows new discoveries made by NOMAD which would have been missed if the available reference had been used (Fig. 5a). Thus, while conventional mapping strategies are possible for these non-model organisms, they limit inference and discovery. We ran NOMAD in unsupervised mode with a lookahead distance of 0 for ease of interpretation (Methods). We mapped concatenated anchor-target pairs to available references, which include the reference genome of the species under study, as well as mobile genetic elements, adapter sequences and RFAM, among others, using both bowtie2 and STAR. We selected anchors where none of the anchors or anchor-target pairs mapped to the reference genome, even partially (Methods). While many follow-up analyses are possible after this stringent selection step, we concentrated on anchor-targets which mapped to protein domains in Pfam, again to showcase interpretable discoveries that would be missed by conventional genome alignment strategies (Fig. 5a).

First, we used NOMAD to identify tissue-specific transcriptional regulation that would be missed by genome alignment, using RNA-Seq data from a single California two-spot octopus *(Octopus bimaculoides)* generated in a study on chemotactile

sensation (van Giesen et al. 2020; personal communication). For each of the 254,066 significant anchors, we aligned the anchor concatenated to its most abundant target (Methods), 17,366 (6.8%) did not Bowtie2-align to any database, 26,199 (10.3%) did not STAR-align to the available reference. STAR alignments include sequences that are partially mapped through soft-clipping as well as mismatches and indels. To illustrate missing regulatory biology from genome alignments, we focused on a highly conservative list of 8,908 anchors which had at most one target aligning to a reference with either Bowtie2 or STAR (Methods). We used BLAST to query NOMAD-called anchor-targets that were completely unaligned by any method (Methods); matches to other *Octopus* species provide evidence that the assembled *O. bimaculoides* genome (Albertin *et al.*, 2015) is incomplete.

Several sequences had BLAST matches to *Octopus sinensis* (Methods, Table 6), a species closely related to *O. bimaculoides*. One example maps to an unconventional Myosin VIIa (LOC115214860), also called MYO7A in humans; mutations in MYO7A cause Usher syndrome, leading to deafness and blindness (Wu *et al.*, 2011). The anchor and both targets (as well as their extended consensuses) map to the *O. sinensis* genome. The more abundant target 1 and anchor also map to the 5' end of an annotated myosin-VIIa transcript (XM_036505518.1), and represent the first and second exons, respectively. The more restricted target 2, expressed only in statocyst tissue, represents a novel alternative first exon (Figure 5B). The matches of target 1 and 2 consensus sequences in the *O. sinensis* genome are imperfect (but ≥92% identity), whereas they have perfect matches in the *O. bimaculoides* genome. The anchor has a perfect match in *O. sinensis* but is not found in the *O. bimaculoides* genome (Supplemental Text). We believe the annotated myo-VIIa gene and transcript in *O. bimaculoides* (LOC106880717; XM_052969897.1) are incomplete based on our evidence from RNA-Seq, as they lack sequences corresponding to target 1 and the anchor. While the *O. sinensis* gene model is more complete in this regard, it is likely that it is incorrect in other regards (Supplemental Text). The restriction of target 2 expression to statocyst tissue is intriguing given MYO7A's association with Usher syndrome, as octopus statocyst is responsive to sound (Solé *et al.*, 2013; Y. Zhang *et al.*, 2015).

We highlight several other anchor-target pairs that do not map to the *O. bimaculoides* genome or transcriptome yet do map to annotated transcripts in *O. sinensis*: carboxypeptidase D, Upf2, and netrin (Supplemental Text). An anchor and its two targets map to the 3' UTR of carboxypeptidase D; target 2 is a close match to the annotated transcript, whereas target 1 contains a 13 nt deletion with respect to target 2 and the genome. The two variants have surprisingly dichotomous expression: target 1 is expressed in sucker rims, the olfactory organ, and statocyst tissue; only target 2 is all other tissues sampled. A second anchor and its two targets map to the 3' UTR of Upf2 ("regulator of nonsense transcripts 2"), involved in nonsense-mediated decay. Target 1 and 2 differ only in the number of CAG repeats -- either six or five, respectively, and

show mutually exclusive expression in the tissues sampled. For both carboxypeptidase D and Upf2, *O. bimaculoides* has an annotated transcript yet the anchor or target sequences are missing, suggesting the *O. bimaculoides* annotation is incomplete. A third anchor-target set matches netrin, a protein involved in axon-guidance: an anchor-target pair differentiated by a single T to C variation, consistent with canonical RNA-editing was found having isoform usage specific to either eye, whole sucker cup, statocyst tissue, and olfactory organ, or to sucker rims (Supplemental Table 6, Supp. Figure 6C).

In this analysis, we focus on the strongest evidence that NOMAD discovers biology that is completely missed by using a reference genome. We have also not investigated novel proteins or transcripts among the thousands of anchors that are unmapped, and are not partially mapped to the Octopus genome by STAR or BLAST. Further, we only include discussion of examples where all anchor-targets are not mapped by STAR, and conjecture that relaxing this will reveal additional examples of transcript variation missed by reference-first methods. We anticipate that investigating these events in octopus, and in other organisms, will enable researchers to prioritize tissue-specific and tissue-regulated splicing and editing, among other (post-)transcriptional regulatory events.

To further demonstrate new biological discoveries from NOMAD, we analyzed RNA-seq data from a completely different domain of life: a seagrass species *Zostera marina*, also called eelgrass, which is under intense selection due to climate change and is critical to carbon sequestration in the ocean (Röhr *et al.*, 2018). A reference genome for this diploid organism and its companion chloroplast and mitochondrial genome were assembled in 2016 (Olsen et al., 2016) and 2021 (Ma et al., 2021). Eelgrass is critical to global ocean ecosystems (Yu *et al.*, 2020, 2022). As global oceans warm, eelgrass has established habitats in low-daylight arctic conditions, a requirement for the species as global temperatures force it to acclimate to high latitudes. Genetic and post-transcriptional adaptation that allow eelgrass to survive at low and high latitudes with greatly varying dynamic ranges of daylight exposure are unknown (Rock and Daru, 2021), but ecological studies have shown that eelgrass is extensively colonized by epiphytic diatoms (Jacobs and Noten, 1980) which have significant impacts on global carbon pumps (Tréguer *et al.*, 2018), and regulate seagrass biomass and growth (Jacobs and Noten, 1980; Prazukin *et al.*, 2022).

We ran NOMAD on a matched collection of RNA-seq from eelgrass shoots collected in coastal regions of France, a mediterranean climate, and of Norway, a near-arctic climate, sampled in summer and winter and during day and night (Methods). Of the most abundant target per anchor, 33855/254991 (13.2%) anchor-target pairs failed alignment to any database (Bowtie2), 29,714 (11.6%) failed STAR alignment to a *Zostera marina* nuclear, chloroplast, and mitochondrial index, and 14,680 (5.7%) failed alignment by both methods.

In-silico translation and Pfam search of anchor-target pairs failing primary sequence alignment yielded hits to 118 protein domains (e value <.05, Supplemental Table 7). The largest enrichment of NOMAD-called Pfam domains versus controls hit Chlorophyll A-B binding protein domain, part of the light-harvesting complex of photosystem II (Fig. 5C). We selected an anchor with a large effect size (0.66) mapping to this domain (Fig. 5D). The most- and third-most abundant anchor-targets had respective best BLAST (blastn) hits with 92% sequence identity to transcripts covering 46/54 and 52/54 bases, to two diatoms (*Fistulifera solaris* and *Epithemia pelagica*, resp.). The second-and fourth-most abundant anchor-targets had marginally better BLAST hits with > 94% sequence identity to two different diatom genomes (*Epithemia pelagica* and *Phaeodactylum tricornutum*, respectively). The distribution of anchor-target counts is affected by all three main variables: geographic location, time of year, and time of day. The most frequently observed target dominates samples from France in June, daytime. This anchor is only detected in Norway December day samples. Targets detected only in France December samples, are more abundant in samples collected during day: all targets from sampling at night have fewer than 5 counts. Together, this suggests the hypothesis that diatom gene expression is impacted by circadian cycles– and upregulated during moderate daylight– and also that the abundance of the host diatom varies seasonally and by location.

Other domains enriched in NOMAD anchor-targets include those annotated as silicon transport, a known biochemical pathway in diatoms (Shrestha and Hildebrand, 2015), and HMG (high mobliy group) box protein domains (Supp. Fig. 7A). In investigated anchors with Pfam hits to Chlorophyll A-B binding (Fig. 5D), HMG box, and 2Fe-2S iron-sulfur cluster binding domain (Supp Figure 7A, Supp Figure 7B) the top hits from Protein BLAST are to diatom species. The two anchor-targets annotated as HMG-box had best BLAST (blastp) hits to a single diatom species with e-values of 0.007 and 1e-05. This suggests that these discoveries by NOMAD derive from co-associated diatoms (most likely species not specifically represented in the current database), rather than eelgrass per se, an inference not possible by restricting to reads mapping to the *Zostera marina* reference genome. For each example, target variation is strongly associated with collection location and or season (Fig. 5D). By bypassing reference alignments, NOMAD identifies differentially expressed sequence variation outside of the *Zostera* genome – likely within the associated epiphyte population. This sequence variation (Fig. 5D) tracks with sample metadata suggesting that genes and proteins expressed by diatoms may have important ecological and or regulatory features missed by past analyses. This highlights that NOMAD can automatically identify regulated sequence expression of co-associated species, whether known or unknown.

While we focused above on NOMAD's discoveries that fail to map to the reference genome, NOMAD also provides insight and joint inference for anchor-targets

mapping to the eelgrass reference genome. One example (Fig. 5E) is a gene encoding the NdhL subunit of chloroplast NADPH dehydrogenase complex (NDH), involved in cyclic electron transport (Laughlin, Savage and Davies, 2020). NOMAD detects a SNP discriminating samples collected in Norway from those in France and additionally, intron retention (Supplemental Text) that is regulated in June compared to December, completely discriminating these times in samples from Norway. This form of splicing regulation will be interesting for further study as modulation of NDH function may affect photosynthetic efficiency and oxidative stress under varying light conditions (M. Ma *et al.*, 2021). Together, these highlighted examples show biology found by NOMAD would be missed by conventional pipelines and illustrate NOMAD's ability to identify sequences completely missing from the genome, or unmappable, revealing new biology of tissue-specific, environmental, and metagenomic regulation. The analytic approach followed here is general and can be followed for any organism and can identify multiple mechanisms of regulation including editing to splicing and sequences missing from the reference genome.

Conclusion

Genomic analysis today is performed with complex computer scientific workflows in a highly domain-specific manner. Here, we have shown that a simple, direct analytic framework with transparent statistics can unify the approaches taken in disparate subfields of genomic data-science and enable novel discovery missed in current workflows. We have achieved this by stating their goal to discover sample-dependent sequence variation in a unified probabilistic model, and used this model to develop NOMAD. NOMAD is an efficient statistics-first algorithm that formulates and efficiently solves tasks in genome science with great generality.

We provided a snapshot of NOMAD's discoveries in currently siloed subfields of genome science. When run on SARS-CoV-2 patient samples during the emergence of the Omicron variant, NOMAD finds that the spike protein is highly enriched for sequence variation, bypassing genome alignment completely. NOMAD provides evidence that Variants of Concern can be detected well before they are flagged as such or added to curated databases. This points to a broader impact for NOMAD in viral and other genomic surveillance, since emerging pathogens will likely have sequence variation missing from any reference.

NOMAD finds novel cell-type specific isoform expression in homologous genes missed by reference-guided approaches, such as in MYL12A/B and in the MHC (HLA) locus, even in a small sample of single cell data. Highly polymorphic and multicopy human genes have been recalcitrant to current genomic analyses and are critical to susceptibility to infectious and complex diseases, e.g. the MHC. NOMAD could shed new light on other polymorphic loci including non-coding RNAs, e.g. spliceosomal variants (Kuo *et al.*, 2017; Buen Abad Najar, Yosef and Lareau, 2019). In addition,

NOMAD unifies detection of many other examples of transcriptional regulation: intron retention, alternative linear splicing, allele-specific splicing, gene fusions, and circular RNA. Further, NOMAD prioritizes V(D)J recombination as the most sample-specific sequencing diversifying process in B and T cells of both human and mouse lemur, where inference in lemur is made using only an approximate genomic reference (human) which diverged from lemur ~60 million years ago.

Disparate data – DNA and protein sequence, or any "-omics" experiment, from Hi-C to spatial transcriptomics – can be analyzed through NOMAD's framework ("Generality of NOMAD", Supplement). NOMAD may also be impactful in analysis of plants, microbes, and mobile elements, including transposable elements and retrotransposons, which are far less well annotated, and are so diverse that references may never capture them. Beyond these discoveries, our work also shows that NOMAD provides a more efficient and powerful solution in domains where competing methods exist, such as the detection of differentially-regulated alternative-splicing.

NOMAD illustrates the power of statistics-first genomic analysis, where references and annotations are only optionally used for *post-facto* interpretation. References are valuable for interpretation of sequence data, including for example, in curating variants associated with disease, such as for genome-wide association studies. However, to be useful, references need not be used to filter raw data by reducing it to a set of positions and variants with respect to the reference alignment. This step itself introduces statistical bias in quantification, blindspots and errors. NOMAD provides a new approach to the use of references by translating the field's "reference-first" approach to "statistics-first", performing direct statistical hypothesis tests on raw sequencing data, enabled by its probabilistic modeling of raw reads rather than of alignment outputs. By design, NOMAD is highly efficient: it enables direct, large-scale study of sample-dependent sequence variation, completely bypassing the need for references or assemblies. NOMAD promises data-driven biological study with scope and power previously impossible.

Limitations of the study

Naturally, some problems cannot be easily formulated in the manner posed, such as cases where quantification of sample-specific RNA or DNA abundance alone is desired (e.g. differential gene expression analysis). However, the problems that can be addressed using this formulation span diverse fields which are of great current importance (Supplement), including those previously discussed. Further, NOMAD's statistical test can be applied to any count matrices, including tables of gene expression by samples.

Acknowledgements

We thank Elisabeth Meyer for extensive edits, figure help, and assistance with exposition; Arjun Rustagi for extensive discussion and assistance in interpretation of viral biology; Roozbeh Dehghannasiri for selecting and collating data from HLCA/TSP/Tabula Microcebus analysis, running the SpliZ, and feedback on presentation; the Tabula Microcebus Consortium (Camille Ezran and Hannah Frank) for data sharing and discussion of B and T cell receptor detection algorithms; Jessica Klein for extensive figure assistance; Michael Swift for discussion of B and T cell receptor detection algorithms; Julia Olivieri for feedback on presentation and detailed comments on the manuscript; Andy Fire for useful discussions and feedback on the manuscript; Robert Bierman for figure assistance and laptop timing; Sebastian Deorwicz and Marek Kokot for early access to code, and Aaron Straight, Catherine Blish, Peter Kim, and all members of the Salzman lab for feedback and comments.

Funding
J.S. is supported by the Stanford University Discovery Innovation Award, National Institute of General Medical Sciences grant nos. R35 GM139517 and the National Science Foundation Faculty Early Career Development Program Award no. MCB1552196. T.Z.B. is supported by the National Science Foundation Graduate Research Fellowship Program and the Stanford Graduate Fellowship.

Author contributions
K.C. designed and implemented the pipeline. T.Z.B. designed, developed, and analyzed the statistics. G.H., P.L. W. provided NOMAD data analysis. I.N.Z. developed the protein domain analysis. J.S. designed and developed the statistics, and conceptualized and supervised the project. All authors analyzed data and wrote the manuscript.

Competing interests
K.C., T.Z.B. and J.S. are inventors on provisional patents related to this work.
The authors declare no other competing interests.

Data and materials availability
The code used in this work is available as a fully-containerized Nextflow pipeline (Di Tommaso *et al.*, 2017) at https://github.com/salzmanlab/nomad, commit 1b73949d.

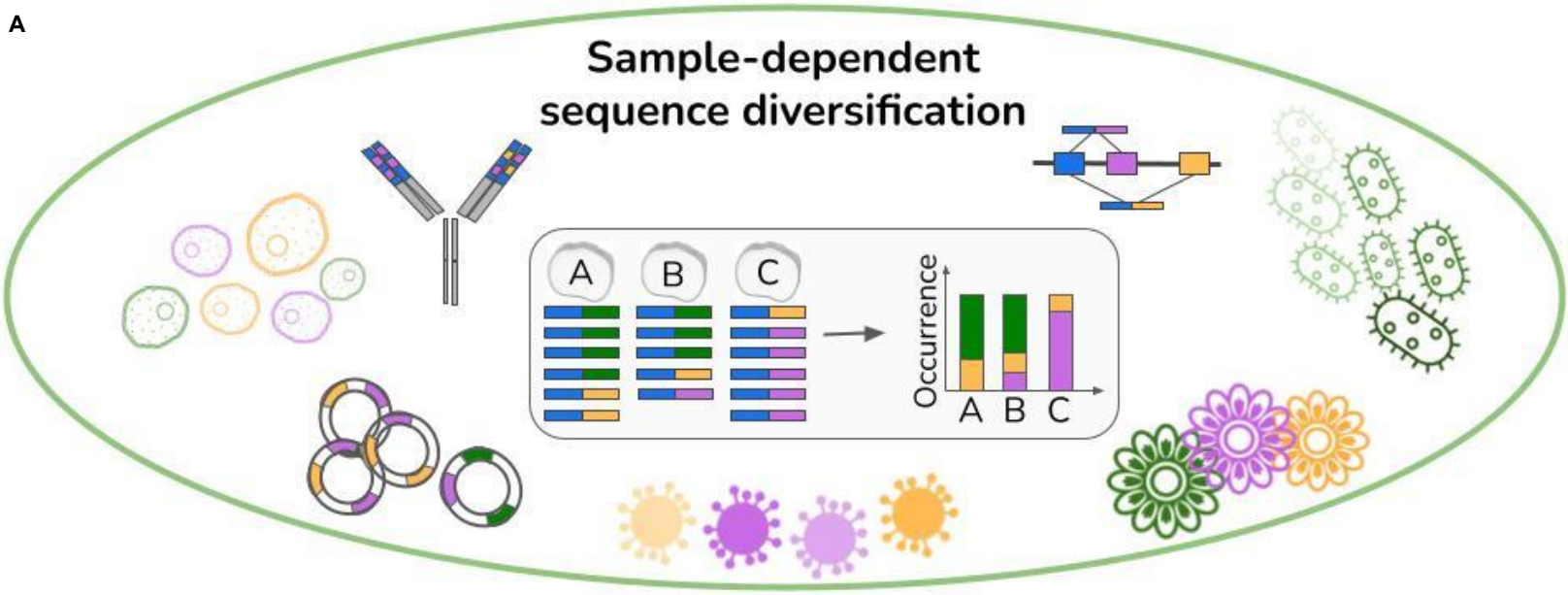
The human lung scRNA-seq data used here is accessible through the European Genome-phenome Archive (accession number: EGAS00001004344); FASTQ files from donor 1 and donor 2 were used. The FASTQ files for the Tabula Sapiens data (both 10X Chromium and Smart-seq2) were downloaded from https://tabula-sapiens-portal.ds.czbiohub.org/; B cells were used from donor 1 and T cells from donor 2. The mouse lemur single-cell RNA-seq data used in this study was generated as part of the Tabula Microcebus consortium; the FASTQ files were

downloaded from https://tabula-microcebus.ds.czbiohub.org. Viral data was downloaded from the NCBI: SARS-CoV-2 from France (SRP365166), SARS-CoV-2 from South Africa (SRP348159), 2020 SARS-CoV-2 from California (SRR15881549), influenza (SRP294571), and rotavirus (SRP328899).

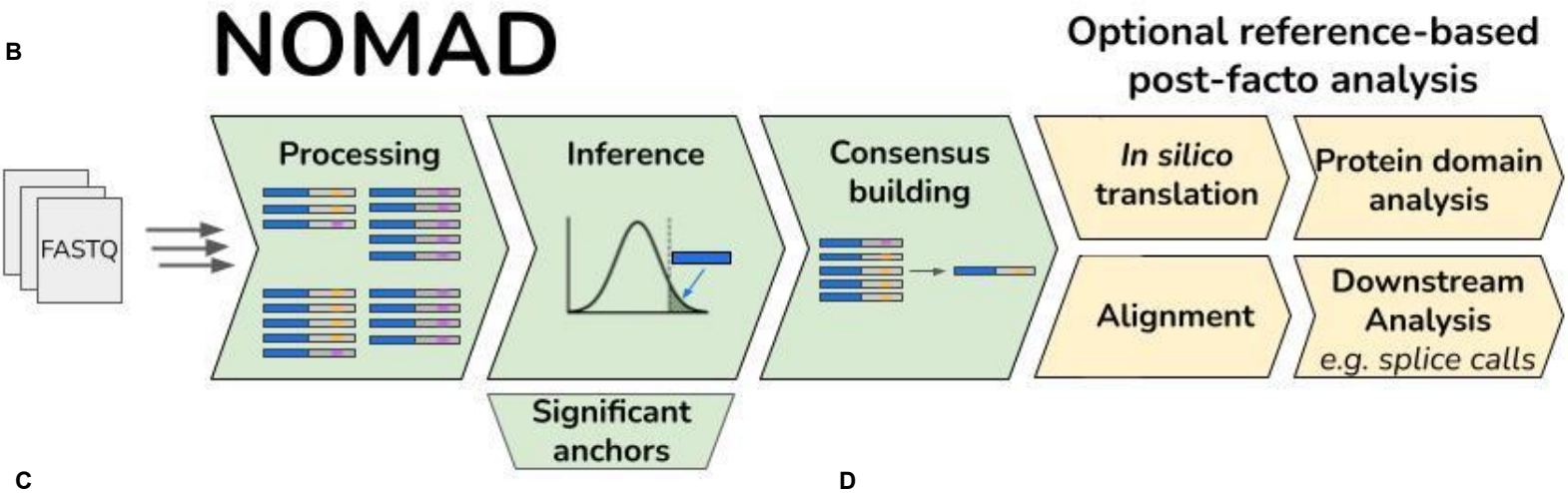
The sample sheets used as pipeline input, including individual sample SRA accession numbers, for all analyses are uploaded to pipeline GitHub repository. Similarly, scripts to perform supplemental analysis can be found on the pipeline repository.

FASTP v0.23.2 was installed on 2/15/23 using bioconda. R-NOMAD v0.3.9 was installed on 2/23/23, commit 5dafdc8.

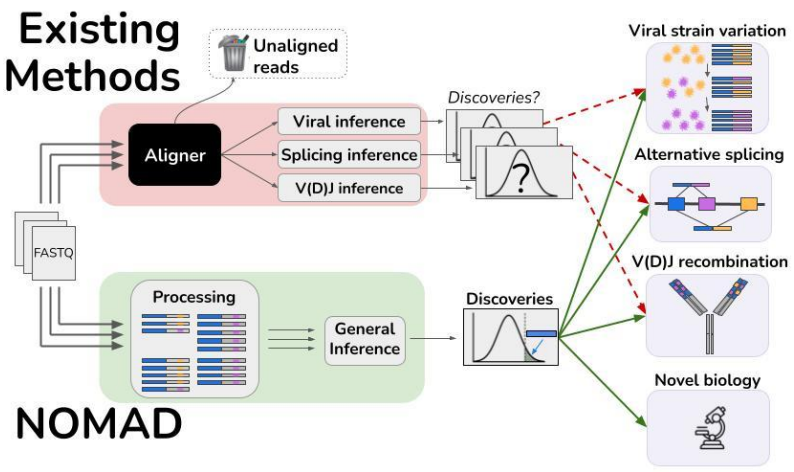
A


Sample-dependent sequence diversification
A
B
C
Occurrence
A B C

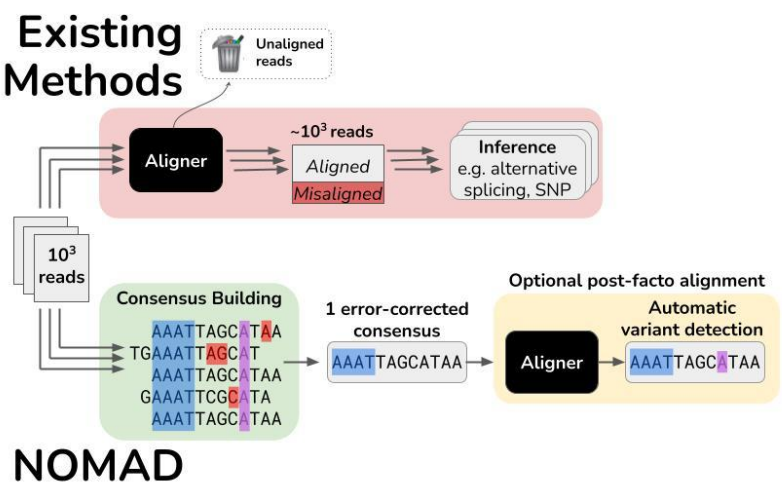
B


NOMAD
Optional reference-based post-facto analysis
FASTQ
Processing
Inference
Consensus building
In silico translation
Protein domain analysis
Alignment
Downstream Analysis
e.g. splice calls
Significant anchors

C


Existing Methods
Unaligned reads
Aligner
Viral inference
Splicing inference
V(D)J inference
Discoveries?
FASTQ
Processing
General Inference
Discoveries
Viral strain variation
Alternative splicing
V(D)J recombination
Novel biology
NOMAD

D


Existing Methods
Unaligned reads
Aligner
~10³ reads
Aligned
Misaligned
Inference
e.g. alternative splicing, SNP
10³ reads
Consensus Building
AAATTAGCATAA
TGAAATTAGCAT
AAATTAGCATAA
GAAATTCGCATA
AAATTAGCATAA
1 error-corrected consensus
AAATTAGCATAA
Optional post-facto alignment
Automatic variant detection
Aligner
AAATTAGCATAA
NOMAD

E

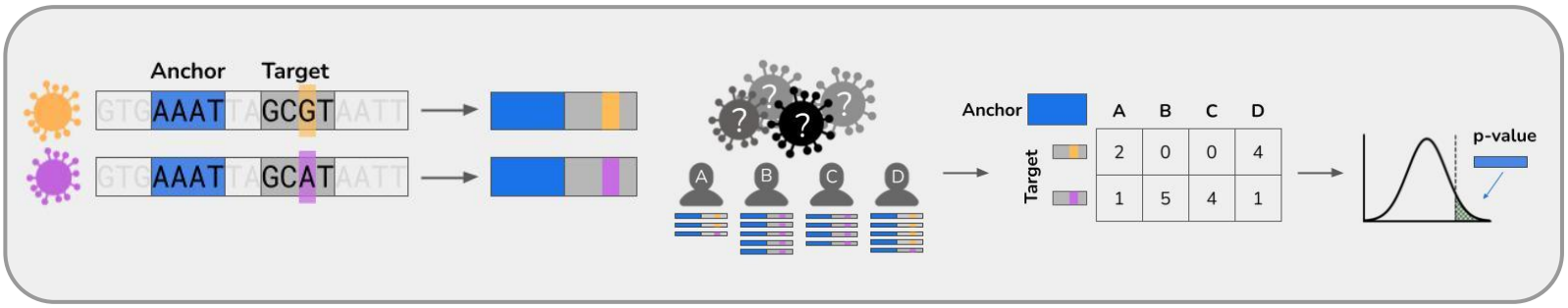

Anchor
Target
AAAT
GCGT
AAAT
GCAT
A
B
C
D
Anchor
Target
A B C D
2 0 0 4
1 5 4 1
p-value

Figure 1

A. **Generality of NOMAD model.** The study of sample-dependent sequence variation unifies problems in disparate areas of genomics which are currently studied with application-specific models and algorithms. Viral genome mutations, alternative splicing, and V(D)J recombination all fit under this framework, where sequence variation depends on the sample (through cell-type or infection strain type). Myriad problems in plant genomics, metagenomics and biological adaptation are subsumed by this framework.
B. **Overview of NOMAD pipeline.** NOMAD takes as input raw FASTQ files for any number of samples >1 and processes them in parallel, counting (anchor, target) pairs per sample. NOMAD performs inference on these aggregated counts, outputting statistically significant anchors. For each significant anchor, a denoised per-sample consensus sequence is built (Fig. 1D). NOMAD also enables optional reference-based post-facto analysis. If a reference genome is available, NOMAD can align the consensus sequences to the reference, enabling denoised downstream analysis (e.g. SNPs, indels, or splice calls). In silico translation of consensuses can optionally be used to study relationships of anchors to protein domains by mapping to databases such as Pfam (Methods).
C. **Overview of NOMAD (green) versus existing workflows (red)**. Existing workflows (red) discard low-quality reads during FASTQ processing and alignment, only then performing statistical testing after algorithmic bias is introduced; p-values are then not unconditionally valid. Further, for every desired inferential task, a different inference pipeline must be used. NOMAD (green) performs direct statistical inference on raw FASTQ reads, bypassing alignment and enabling data-scientifically driven discovery. Due to its generality, NOMAD can simultaneously detect myriad biological examples of sample-dependent sequence variation.
D. **NOMAD consensus building**. NOMAD constructs a per-sample consensus sequence for every significant anchor by taking all reads in which the anchor (blue) appears, and recording plurality votes for each nucleotide, denoising reads while preserving the true variant; sequencing errors in red and biological mutations in purple. Existing approaches require alignment of all reads to a reference prior to error correction, requiring orders of magnitude more computation, discarding reads in both processing and alignment, and potentially making erroneous alignments due to sequencing error. They further require inferential steps, e.g. to detect if there is a SNP or alternatively spliced variant.
E. **Construction of NOMAD anchor, target pairs.** A stylized expository example of viral surveillance: 4 individuals A-D are infected with one of two variants (orange and purple), differing by a single basepair (orange and purple). NOMAD anchor *k*-mers are blue (k=4), followed by a lookahead distance of L=2, and the corresponding *k*-mer targets. Given sequencing reads from the 4 individuals as

shown, NOMAD generates a target by sample contingency table for this blue anchor, and computes a p-value to test if this anchor has sample-dependent sequence diversity.

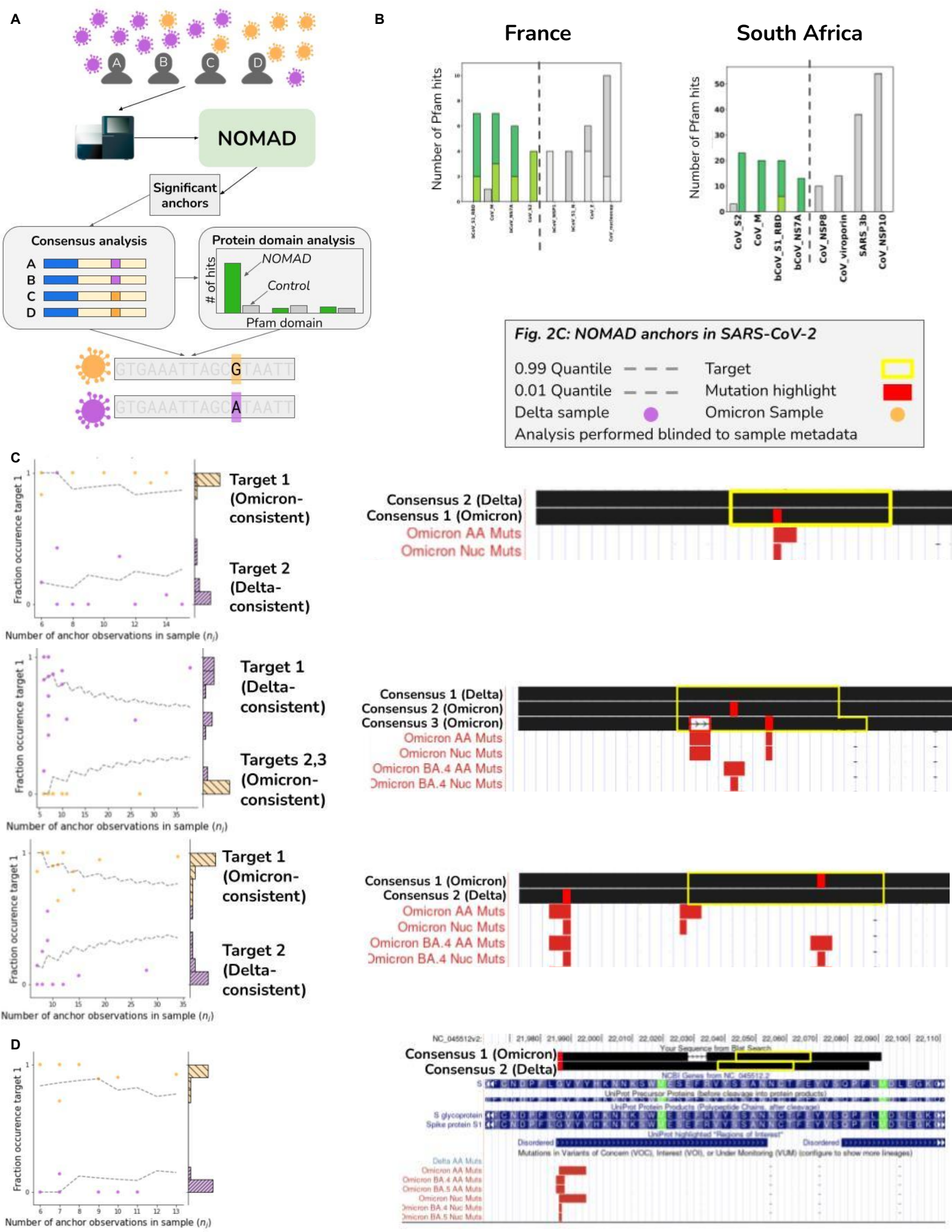

A
A
B
C
D
NOMAD
Significant anchors
Consensus analysis
A
B
C
D
Protein domain analysis
# of hits
NOMAD
Control
Pfam domain
GTGAAATTAGCGTAATT
G
GTGAAATTAGCATAATT
A
B
France
South Africa
Number of Pfam hits
CoV_S2
CoV_M
bCoV_S1_RBD
bCoV_NS7A
CoV_NSP8
CoV_viroporin
SARS_3b
CoV_NSP10
Fig. 2C: NOMAD anchors in SARS-CoV-2
0.99 Quantile
0.01 Quantile
Delta sample
Target
Mutation highlight
Omicron Sample
Analysis performed blinded to sample metadata
C
Fraction occurence target 1
Number of anchor observations in sample (n_j)
Target 1 (Omicron-consistent)
Target 2 (Delta-consistent)
Consensus 2 (Delta)
Consensus 1 (Omicron)
Omicron AA Muts
Omicron Nuc Muts
Target 1 (Delta-consistent)
Targets 2,3 (Omicron-consistent)
Consensus 1 (Delta)
Consensus 2 (Omicron)
Consensus 3 (Omicron)
Omicron BA.4 AA Muts
Omicron BA.4 Nuc Muts
Target 1 (Omicron-consistent)
Target 2 (Delta-consistent)
Consensus 1 (Omicron)
Consensus 2 (Delta)
D
Fraction occurence target 1
Number of anchor observations in sample (n_j)
NC_045512v2:
Consensus 1 (Omicron)
Consensus 2 (Delta)
Your Sequence from Blast Search
NCBI Genes from NC_045512.2
UniProt Precursor Proteins (before cleavage into protein products)
UniProt Protein Products (Polypeptide Chains, after cleavage)
S glycoprotein
Spike protein S1
UniProt highlighted "Regions of Interest"
Disordered
Mutations in Variants of Concern (VOC), Interest (VOI), or Under Monitoring (VUM) (configure to show more lineages)
Delta AA Muts
Omicron AA Muts
Omicron BA.4 AA Muts
Omicron BA.5 AA Muts
Omicron Nuc Muts
Omicron BA.4 Nuc Muts
Omicron BA.5 Nuc Muts

Figure 2

A. **Stylized example representing NOMAD workflow for viral strain identification.** Patients with varying viral strains are sampled; two representative strains with differentiating mutations are depicted in orange and purple. NOMAD is run on raw FASTQs generated from sequencing patient samples. Significant anchors are called without a reference genome or clinical metadata. Optional post-facto analysis quantifies domain enrichment via in silico translation of consensus sequences derived from NOMAD-called anchors versus controls. Consensuses can also be used to call variants *de novo* and can be compared to annotated variants e.g. in SARS-CoV-2, Omicron.
B. NOMAD SARS-CoV-2 protein profile hits (anchor effect size >.5) to the Pfam database (greens) and control (greys) for France and South Africa datasets; ordered by enrichment in NOMAD hits compared to control showing large distributional differences (chi-squared test p-values France: < 3.7E-7, SA: <2.5E-39). Spike protein domains are highly enriched in NOMAD calls versus controls. In the France data, the most NOMAD-enriched domain is the betacoronavirus S1 receptor binding domain (hypergeometric p=2.3E-2, corrected) followed by Orf7A (hypergeometric p=6.4E-2, corrected), known to directly interact with the host innate immune defense. In the South Africa data, the most enriched NOMAD profiles are CoV S2 (p=2.9E-6) and the coronavirus membrane protein (p=8.4E-8).
C. Examples of NOMAD-detected anchors in SARS-CoV-2 (France data) classified as strain-defining. Scatterplots (left) show the fraction of each sample's observed fraction of target 1 (the most abundant target) for three representative anchors, binomial confidence intervals: (.01,.99), p=empirical fraction occurrence of target 1 (Supplement). y-axis shows histogram of the fraction occurrence of target 1. Mutations (right) found in the targets are highlighted in purple, BLAT shows single nucleotide mutations match known Omicron mutations. Binomial p-values of 1.6E-21, 1.0E-13, and 1.8E-16 respectively (Methods). These uncorrected p-values, and are not used for any significance testing, but rather to quantify the visually separated nature of the plots. The anchor in (top) maps to the coronavirus membrane protein; anchors in (middle and bottom) map to the spike protein. One sample (out of 26) depicted in the bottom plot has a consensus mapping perfectly to the Wuhan reference; 3 other consensuses contain annotated Omicron mutations, some designated as VOC in May of 2022, 3 months after these samples were collected.
D. Example of a NOMAD-detected anchor in SARS-CoV-2 that is not classified as strain-defining. Same format as Fig. 2C, showing a 6 basepair deletion in the spike protein that only occurs in patients infected with Omicron. Binomial p-value of 1.3E-19. The highlighted deletion is not annotated as a VOC yet it is present only in samples classified as Omicron.

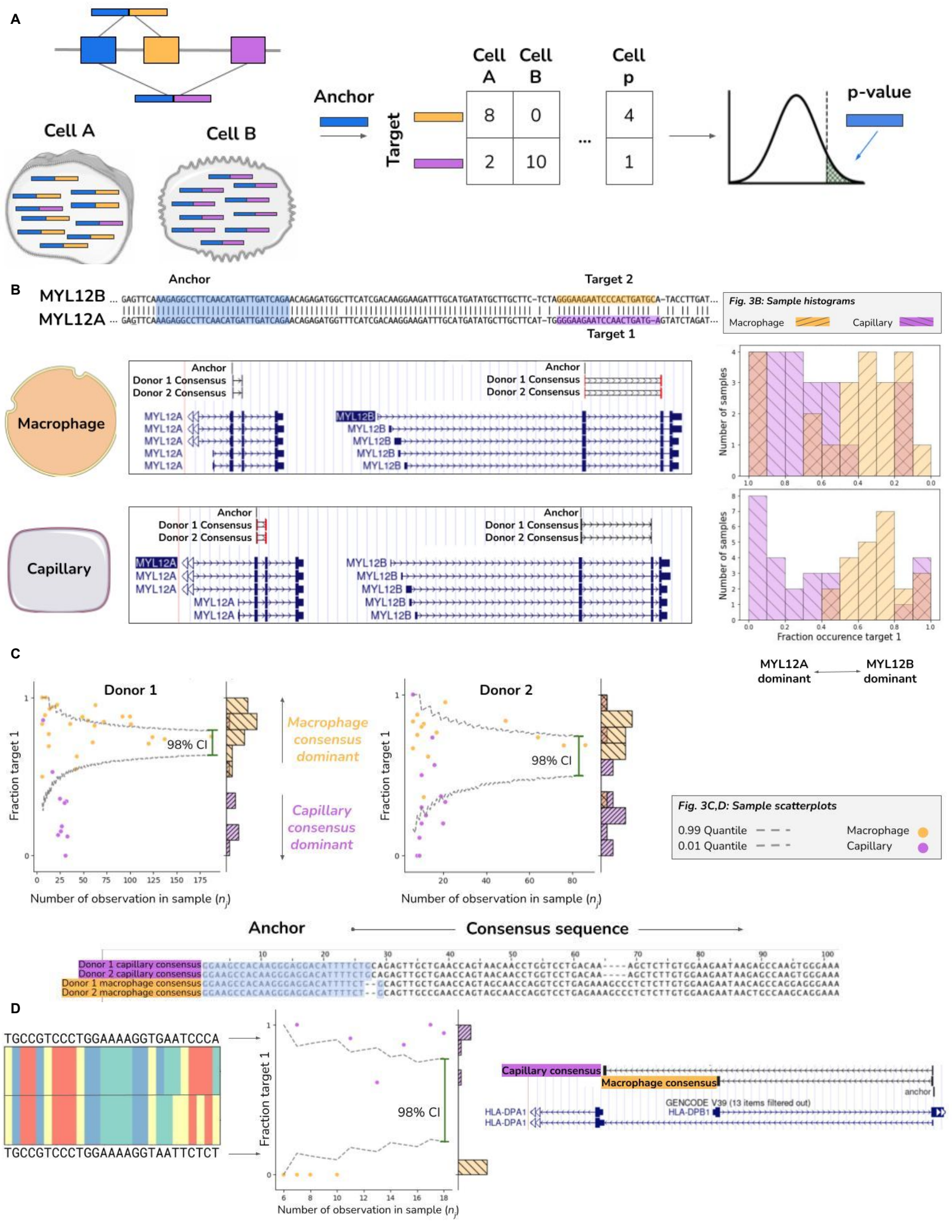

A
Cell A
Cell B
Anchor
Target
Cell A
Cell B
Cell p
8
0
4
2
10
1
...
p-value
B
Anchor
Target 2
MYL12B
MYL12A
Target 1
Fig. 3B: Sample histograms
Macrophage
Capillary
Macrophage
Capillary
Anchor
Donor 1 Consensus
Donor 2 Consensus
MYL12A
MYL12B
Number of samples
Fraction occurence target 1
MYL12A dominant
MYL12B dominant
C
Donor 1
Donor 2
98% CI
Fraction target 1
Number of observation in sample (n_j)
Macrophage consensus dominant
Capillary consensus dominant
Fig. 3C,D: Sample scatterplots
0.99 Quantile
0.01 Quantile
Macrophage
Capillary
Anchor
Consensus sequence
Donor 1 capillary consensus
Donor 2 capillary consensus
Donor 1 macrophage consensus
Donor 2 macrophage consensus
D
TGCCGTCCCTGGAAAAGGTGAATCCCA
TGCCGTCCCTGGAAAAGGTAATTCTCT
Capillary consensus
Macrophage consensus
anchor
GENCODE V39 (13 items filtered out)
HLA-DPA1
HLA-DPB1

Figure 3

A. Toy diagram depicting differentially regulated alternative splicing detection with 3 exons and 2 isoforms. Isoform 1 consists of exon 1 (blue) and exon 2 (orange), and is predominantly expressed in cell A. Isoform 2 consists of exon 1 (blue) and exon 3 (purple) and is primarily expressed in cell B.. An anchor sequence in exon 1 (blue), then generates target sequences in exon 2 (orange) or exon 3 (purple). Counts are used to generate a contingency table, and NOMAD's statistical inference detects this differentially regulated alternative splicing.
B. Detection of differential regulation of MYL12A/B isoforms. (top-left) Shared anchor (q-value 2.5E-8, donor 1, 2.3E-42 for donor 2) highlighted in yellow, maps *post fact* to both MYL12 isoforms, highlighting the power of NOMAD inference: MYL12A and MYL12B isoforms share >95% nucleotide identity in coding regions. (bottom-left) NOMAD's approach automatically detects target and consensus sequences that unambiguously distinguish the two isoforms. (right) In both donors, NOMAD reveals differential regulation of MYL12A and MYL12B in capillary cells (MYL12A dominant) and macrophages (MYL12B dominant).
C. NOMAD-identified anchor mapping to HLA-DRB1 (shared anchor, q-value of 4.0E-10 for donor 1, 1.2E-4 for donor 2). Scatter plots show cell-type regulation of different HLA-DRB1 alleles not explained by a null binomial sampling model ($p<2E-16$) for donor 1, ($p<5.6E-8$) for donor 2, finite sample confidence intervals depicted in gray (Methods). Each (donor, cell-type) pair has a dominant target, per-cell fractions represented as "fraction target 1" in scatterplots, and a dominant consensus mapping to the HLA-DRB1 3' UTR (multiway alignment); donor 1 capillary consensus contains an insertion and deletion.
D. Donor 1 specific splice variant of HLA-DPA1. Anchor q-value: 7.9E-22. Detected targets are consistent with macrophages exclusively expressing the short splice isoform which excises a portion of the ORF and changes the amino acid composition and 3' UTR compared to the dominant splice isoform in capillary cells; splice variants found *de novo* by NOMAD consensuses. Binomial hypothesis test as in 3C for cell-type specific target expression depicted in scatter plots (binomial $p<2.8E-14$).

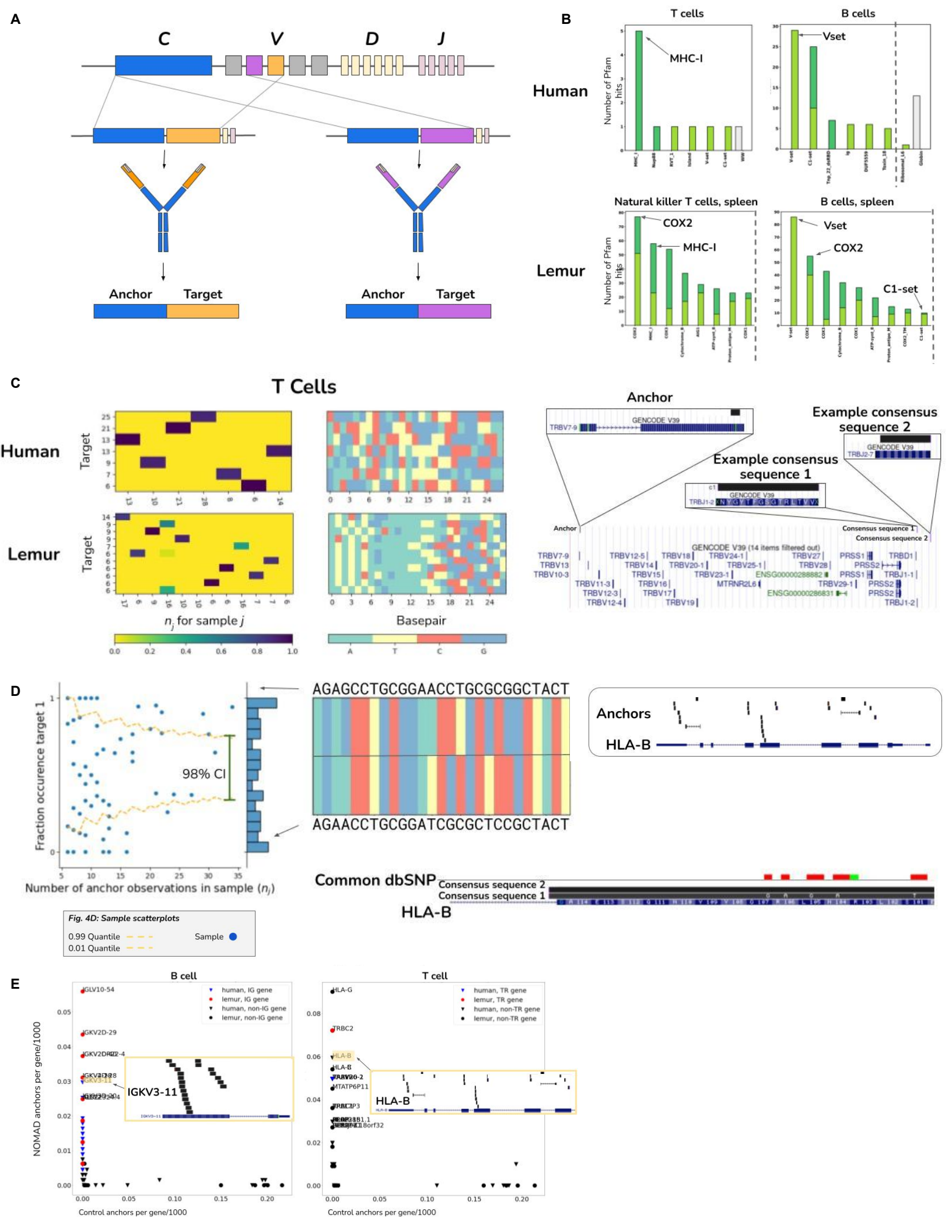

A
C
V
D
J
Anchor
Target
Anchor
Target
B
Human
T cells
B cells
MHC-I
Vset
Number of Pfam hits
Lemur
Natural killer T cells, spleen
B cells, spleen
COX2
MHC-I
Vset
COX2
C1-set
C
T Cells
Human
Lemur
Target
Target
n_j for sample j
Basepair
A
T
C
G
Anchor
GENCODE V39
Example consensus sequence 2
Example consensus sequence 1
Consensus sequence 1
Consensus sequence 2
GENCODE V39 (14 items filtered out)
D
Fraction occurence target 1
98% CI
Number of anchor observations in sample (n_j)
AGAGCCTGCGGAACCTGCGCGGCTACT
AGAACCTGCGGATCGCGCTCCGCTACT
Anchors
HLA-B
Common dbSNP
Consensus sequence 2
Consensus sequence 1
HLA-B
Fig. 4D: Sample scatterplots
0.99 Quantile
0.01 Quantile
Sample
E
B cell
T cell
human, IG gene
lemur, IG gene
human, non-IG gene
lemur, non-IG gene
human, TR gene
lemur, TR gene
human, non-TR gene
lemur, non-TR gene
IGKV3-11
HLA-B
NOMAD anchors per gene/1000
Control anchors per gene/1000
Control anchors per gene/1000

Figure 4

A. Toy diagram depicting V(D)J recombination, specifically of different variable regions in the heavy chain. An anchor sequence in the constant region (blue), generates target sequences (orange and purple) during V(D)J recombination, in which immunoglobulins may receive different gene segments during rearrangement. NOMAD is able to rediscover and detect these recombination events by prioritizing sample-specific TCR and BCR variants.
B. Analysis of lemur and human B (left) and T (right) cells. Human genes are depicted as triangles; lemur as circles. *Post facto* alignments show variable regions in the kappa light chain in human B cells are most densely hit by NOMAD anchors and absent from controls; in T cells, the HLA loci and TRB including its constant and variable region are most densely hit, which are absent from controls. x-axis indicates the fraction of the 1000 control anchors (most abundant anchors) that map to the named transcript, y-axis indicates the fraction of NOMAD's 1000 most significant anchors that map to the named transcript. Each inset depicts anchor density alignment in the IGKV region (left) and HLA-B in CD4+ T cells (top right) and TRBC-2 (bottom right), showing these regions are densely hit.
C. In human T cells (right), we show a NOMAD anchor in the TRVB7-9 gene, and two example consensuses which map to disjoint J segments, TRBJ1-2 and TRBJ2-7. Histograms of this anchor depict combinatorial single-cell (columns) by target (row) expression of targets detected by NOMAD. Histogram for lemur T cells depicted similarly; lemur T cell anchor maps to the human gene TBC1D14.
D. NOMAD-annotated anchors are enriched in HLA-B (top Fig. 4.D.1). Sample scatterplot (middle)Fig. 4.D.2 shows that T cells have allelic-specific expression of HLA-B, not explicable by low sampling depth (binomial test as in Fig. 3d,e described in Methods, $p < 4.6E-24$). Fig. 4.D.3: HLA-B sequence variants are identified *de novo* by the consensus approach (bottom), including allele-specific expression of two HLA-B variants, one annotated in the genome reference, the other with 5 SNPs coinciding with annotated SNPS.
E. NOMAD protein profile analysis shows that NOMAD recovers domains known to be diversified in adaptive immune cells, bypassing any genome reference or alignment; control hits computed from the most abundant anchors have no such enrichment. In B cells, hits in the V set, IG like domains resembling the antibody variable region, are at a relatively high E-value, as predicted by protein diversification generated during V(D)J, making matching to reference domains imperfect. The third most hit domain is Tnp_22_dsRBD, a double stranded RNA binding domain, suggesting potential activation of LINE elements in B cells. COX2, known to be involved in immune response, is highly ranked in both lemur T and B cells. Plots were truncated for clarity of presentation as indicated by dashed grey line (Fig. S2 D-F).

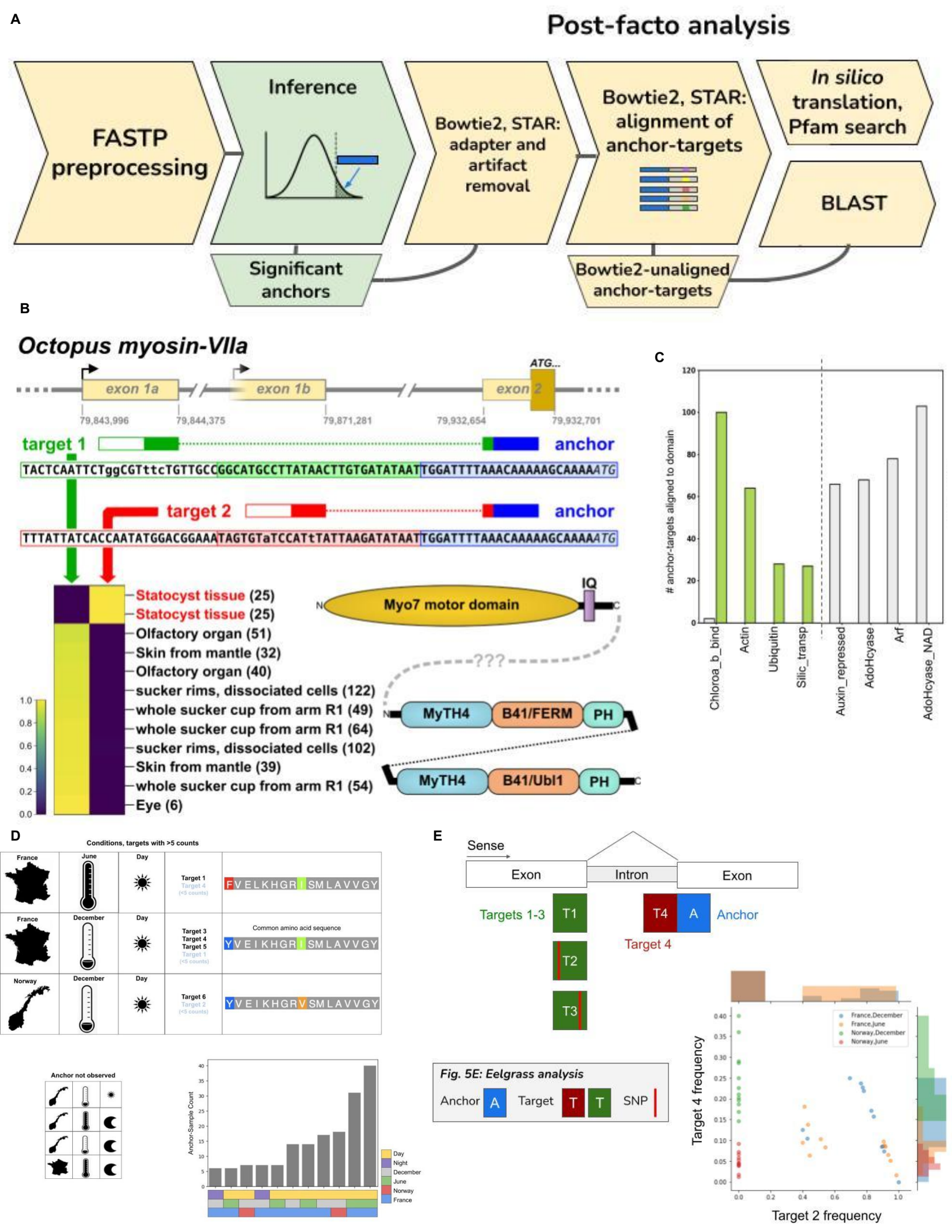

A
Post-facto analysis
FASTP preprocessing
Inference
Bowtie2, STAR: adapter and artifact removal
Bowtie2, STAR: alignment of anchor-targets
In silico translation, Pfam search
BLAST
Significant anchors
Bowtie2-unaligned anchor-targets
B
Octopus myosin-VIIa
exon 1a
exon 1b
exon 2
ATG...
79,843,996
79,844,375
79,871,281
79,932,654
79,932,701
target 1
anchor
TACTCAATTCTggCGTttcTGTTGCCGGCATGCCTTATAACTTGTGATATAATTGGATTTTAAACAAAAAGCAAAAATG
target 2
anchor
TTTATTATCACCAATATGGACGGAAATAGTGTaTCCATtTATTAAGATATAATTGGATTTTAAACAAAAAGCAAAAATG
Statocyst tissue (25)
Statocyst tissue (25)
Olfactory organ (51)
Skin from mantle (32)
Olfactory organ (40)
sucker rims, dissociated cells (122)
whole sucker cup from arm R1 (49)
whole sucker cup from arm R1 (64)
sucker rims, dissociated cells (102)
Skin from mantle (39)
whole sucker cup from arm R1 (54)
Eye (6)
Myo7 motor domain
IQ
???
MyTH4
B41/FERM
PH
MyTH4
B41/Ubl1
PH
C
anchor-targets aligned to domain
Chloroa_b_bind
Actin
Ubiquitin
Silic_transp
Auxin_repressed
AdoHcyase
Arf
AdoHcyase_NAD
D
Conditions, targets with >5 counts
France
June
Day
Target 1
Target 4
(<5 counts)
FVELKHGRISMLAVVGY
France
December
Day
Target 3
Target 4
Target 5
Target 1
(<5 counts)
Common amino acid sequence
YVEIKHGRISMLAVVGY
Norway
December
Day
Target 6
Target 2
(<5 counts)
YVEIKHGRVSMLAVVGY
Anchor not observed
Anchor-Sample Count
Day
Night
December
June
Norway
France
E
Sense
Exon
Intron
Exon
Targets 1-3
T1
T2
T3
T4
A
Anchor
Target 4
Fig. 5E: Eelgrass analysis
Anchor
A
Target
T
T
SNP
France,December
France,June
Norway,December
Norway,June
Target 4 frequency
Target 2 frequency

Figure 5:

A. To showcase novel biology, anchors were concatenated with their adjacent targets (lookahead distance 0) and aligned with STAR and bowtie to reference genomes. Anchors where at most one anchor-target aligned by any method were selected for BLAST or in silico translation and Pfam mapping.
B. Anchor-targets have BLAST hits to *O. sinensis* chromosome LG8 (NC_043004.1) but fail alignment to the *O. bimaculoides* genome, are shown as reverse-complements, to give the mRNA sense-strand. The anchor is depicted in blue and targets are depicted in red or green; white boxes are sequences extending beyond the targets (from raw read data). Bases in lowercase are positions where the *O. bimaculoides* RNA-Seq data differs from the *O. sinensis* reference. Target 2 represents an unannotated alternative first exon, expressed in statocyst tissue; its 5' extent is unknown so is grayed-out. The genomic coordinates shown are for *O. sinensis* chromosome LG8 (NC_043004.1); the sequence of exon 2 is missing from the *O. bimaculoides* genome assembly (whereas exon 1a and 1b are present but unannotated). The matching *O. sinensis* gene (LOC115214860) encodes a protein with only the motor and IQ domains. Adjacent to it in head-to-head orientation is a gene also annotated as myosin-VIIa (LOC115214969) which lacks a motor domain but has two repeats of other domains typical of myosin-VIIa. B41 = Band 4.1 homolog; FERM = (Four.1 protein, Ezrin, Radixin, Moesin) domain 1, F1 sub-domain; Ubl1 = first ubiquitin-like (Ubl) domain located at the N-terminus of coronavirus SARS-CoV non-structural protein 3 (Nsp3) and related proteins; PH = Pleckstrin homology-like domain.
C. In eelgrass SRP327909, the 4 domains with the most anchor-targets mapping to the domain (e-value <= 0.01) among NOMAD calls (green) and controls (gray). Highest NOMAD counts are in Chloroa_b_bind, part of the photosystem II light harvesting complex.
D. (L) Conditions in which anchor-targets were observed at least 5 times illustrate condition-specific target expression in daylight conditions. Targets expressed in distinct conditions produce distinct substitutions in the protein sequences, and in France December Day, 3 distinct targets produce a shared amino acid sequence. Amino acid sequences have best Pfam hits to Chlorophyll A-B binding protein (all e-values < 2.3e-07). BLAST (blastp) AA sequence 1 has 100% query cover and 100% identity to a transcript in 3 diatom species, AA sequence 2 has 100% query cover and 100% identity to a transcript in 4 diatom species, and AA sequence 3 has a single substitution to a transcript in 4 diatom species (100% query cover, 94.12% identity). BLAST (blastn) anchor-targets (AT) all best hit diatoms: AT1 best hits F. solaris (92% query cover, 92% identity), AT2 best hits E. pelagica (92% query cover, 94.23% identity), AT3 best hits E. pelagica (96% query cover, 92.31% identity), AT4 best hits 4 P. tricornutum (94% query cover, 96.08% identity), AT5 best hits E. pelagica and P. tricornutum (94% query cover,

94.12% identity), and AT6 best hits F. solaris (96% query cover, 92.31% identity). (R): anchor-sample counts (barplot) are highest in samples harvested during the day.

E. In eelgrass, NOMAD identifies season-specific splicing and location-specific polymorphism in anchor-targets. Pfam analysis assigns the sequences to NADH dehydrogenase transmembrane subunit. The top 3 targets in terms of abundance BLAST to splice junctions, and differ from one another by SNPs in two positions. The 4th most abundant target represents intron retention. In this anchor, only samples from France express target 2, partitioning perfectly by country. Within the samples from Norway, the collection month (December or June) can be perfectly predicted by fractional expression of target 4. Full sequences and additional discussion are presented in the supplemental text.

Materials and Methods

Anchor preprocessing and parameter choices

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, to reduce computational burden. For this manuscript NOMAD was run with 1M 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 FASTQ files from read 1; for HLCA datasets, both read 1 and read 2 were used. 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 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. Then, we retain only anchors having more than 30 total counts after the above thresholds were applied. 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, larger choices of R have provably higher power than smaller choices, following the style of analysis in (Salzman, Jiang and Wong, 2011). We give examples of how choices of *k*, R, and tiling length impact results in France SARS-CoV-2 data as follows, showing that NOMAD yields similar results for a range of parameter choices. Default parameters shown in bold: we tested K = [25, **27**, 30]; Tile = [3, **5**, 7]; Lookahead = [0, 15, **23**]. For K = 25, 94.4% of anchors with default parameters contain at least one of the K=25 anchors as a substring. For K = 30, 93.8% of anchors with K=30 contain at least one of the anchors with default parameters a substring. For tile size of 3, 85% of the anchors from the default run can be found in the significant anchors of tile size of 3. For tile size of 7, 85% of the anchors from the default run can be found in the significant anchors of tile size of 5. For lookahead distance of 0, 37% of the anchors from the default run can be found in the significant anchors of tile size of 3; for lookahead distance of 15, 76% of the anchors from the default run can be found in the significant anchors. Overall, as tile size decreases, anchor calls increase (4715, 5522, 7891 for [7,5,3] respectively). As k varies, anchor calls stay essentially the same (5875, 5522, and 5958 for k=[25,27,30] respectively). Finally, for lookahead distance, the total number of calls decrease as lookahead distance increases (13239, 8295, 5522 for [0,15,23] respectively).

NOMAD p-values

While contingency tables have been widely analyzed in the statistics community (Agresti, 1992; Diaconis and Sturmfels, 1998; Chen *et al.*, 2005), 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, an anchor-sample score is constructed for sample j, S_j, as a scaled version of the difference between these two. Finally, the test statistic S is computed 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, and M denotes the total number of observations of this anchor.

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

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

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

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

Statistically valid p-values are computed 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, here with K=10 and L=50. 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 Benjamini-Yekutieli (BY) correction (Benjamini Hochberg (BH) with hypotheses with positive dependence) to the list of p-values (for each anchor), yielding valid FDR controlled q-values reported throughout the manuscript (Benjamini and Yekutieli, 2001) implemented with the statsmodels.api.stats.multipletests functionality in python.

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 (Ståhlberg *et al.*, 2016), and may have information theoretic justification in *de novo* assembly (Motahari *et al.*, 2013) (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 we can union-bound 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 (Abante, Wang and Salzman, 2022). The reference used are: UniVec, Illumina adapters, grass carp GCA_019924925 (Wu *et al.*, 2022), Escherichia phage phiX174, Rfam (Kalvari *et al.*, 2021), Dfam (Storer *et al.*, 2021), TnCentral (Ross *et al.*, 2021), ACLAME (Leplae *et al.*, 2004), ICEberg (Bi *et al.*, 2012), CRISPR direct repeats (Couvin *et al.*, 2018), ITSoneDB (Santamaria *et al.*, 2018), ITS2 (Selig *et al.*, 2008), WBcel235, TAIR10, grch38_1kgmaj, Octopus bimaculoides (Albertin *et al.*, 2015) , Octopus bimaculoides transcriptome (van Giesen *et al.*, 2020) , Zostera marina (X. Ma *et al.*, 2021), and Zostera marina chloroplast and mitochondrion (Olsen et al., 2016). Grass carp and other indices unrelated to the biological study are used as controls. 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 (Mistry *et al.*, 2021). 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 (.5) anchors in SARS-CoV-2 from South Africa, 252 high effect size (.5) anchors in SARS-CoV-2 from France; 1000 anchors (no effect size filter) 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 (Shen *et al.*, 2016) prior to using hmmsearch from the HMMer3 package (Johnson, Eddy and Portugaly, 2010) 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 variation 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-CoV-2 analysis

The SARS-CoV-2 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 Concern’, 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.

To determine what NOMAD calls (and control anchors) were strain defining we perform the following. To generate NOMAD’s calls, we filter for anchors that are significant (with a BY corrected p-value less than .05) and have large effect size (> .5), yielding a list of N NOMAD-called anchors. Control anchors are generated by taking the N anchors with the highest counts. For each of these anchors we construct their target x sample contingency table, first filtering out all anchors with fewer than 30 counts, only 1 unique target, or only 1 unique sample, and filtering out all samples with 5 or fewer counts. Then, we discard all targets that constitute less than 5% of the remaining counts for that anchor. These the remaining anchors and targets are then bowtie aligned to an index comprised of the Wuhan reference genome (NC_045512v2), and Delta (OK091006.1), Omicron BA.1 (OX315743.1), and Omicron BA.2 (OX315675.1) assemblies. For this alignment, options -a --best --strata were used. Then, for each set of anchors (NOMAD calls, and controls), the list is filtered to only anchors that align to one of the reference assemblies, further requiring that each anchor have at least one target that aligns to a reference assembly. Then for each anchor, we declare it to be strain defining if, for any of the reference assemblies, it has at least one target that maps to it and one target that does not.

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 (Thompson *et al.*, 2020) and EF-Tu, further elucidating these proteins' roles during infection (Sun and Whittaker, 2007; Kuo *et al.*, 2017) (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 (Song *et al.*, 2020), and the rotavirus NSP3 (Rota_NSP3, 87 NOMAD vs 35 control hits), a viral protein involved in subverting the host translation machinery (Gratia *et al.*, 2015), 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) (Dobin *et al.*, 2013). 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 (Kiepiela *et al.*, 2004; Elahi *et al.*, 2011). 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. S2 E,F).

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 ($p < 4{,}6E\text{-}24$) (Francis *et al.*, 2022). 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)

Laptop analysis dataset:

Ten B and T cells from donor 2 blood sequenced by Smart-Seq2 were used for the laptop benchmarking. These files totalled 43,870,027 reads, averaging 4.3M reads per cell. The fastq files for the Tabula Sapiens data were downloaded from https://tabula-sapiens-portal.ds.czbiohub.org/.

Files used:
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A13_S73_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A18_S78_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A19_S79_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A21_S81_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A3_S63_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A5_S65_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A6_S66_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A8_S68_R1_001.fastq.gz

TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A9_S69_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_B10_S94_R1_001.fastq.gz

Data and NOMAD runs for Zostera marina and Octopus bimaculoides._

Data was downloaded from SRP327909 (eelgrass), and SRP278619 (Octopus), using nf-core fetchngs run in default mode (Ewels *et al.*, 2020) and preprocessed with FASTP (Chen *et al.*, 2018) run in default mode to mitigate false positive calls due to adapter concatenation. NOMAD was run with gap length=0, anchor_unique_targets_threshold=1, anchor_count_threshold=50, anchor_samples_threshold=1, anchor_sample_counts_threshold=5, and excluding anchor-targets containing poly A / C / G / T run of length 8. 500 pairs of random c and f were chosen. Fastp v0.3.9 was installed on 2/23/23 using bioconda.

Supplemental Tables 6 and 7 – Anchor-target Bowtie and Pfam annotations

Anchor-targets are selected from the NOMAD calls if they do not contain a homopolymer length > 5, if effect size > 0.1, and if the corrected p value < 0.01. Element annotations are run as described on anchor-targets and all hits are reported; anchor-targets mapping to UniVec, Illumina adapters, or to the Ctenopharyngodon idella (grass carp which we used as a control as it contains many artifactual Illumina adapters, personal communication, GCA_019924925) genome are removed. Element annotations are run for anchors; anchor-targets are removed if the anchor maps to UniVec, Illumina adapters, or to the grass carp genome. Anchor-targets with no element annotation are in-silico translated in six frames (-3,-2,-1,1,2,3) to amino acid sequence and submitted to HMMer search of the Pfam database. Element annotations, NOMAD statistics, and the best Pfam hit by full sequence e-value across the 6 translation frames is reported for each of the selected anchor targets. Anchor-targets are aligned with STAR 2.7.5 to a reference index generated from a FASTA of grass carp and COVID-19; anchor-targets for which STAR reports a hit are removed. Anchor-targets are aligned with STAR 2.7.5 to a reference index generated from either the Octopus bimaculoides reference assembly or the Zostera marina nuclear, mitochondrial, and chloroplast genomic FASTA. Element annotations was run using NOMAD commit ID 728066b.

BLAST of anchor-targets in Octopus and Eelgrass

We selected anchors with no more than 1 target mapping with either STAR to the reference genome or Bowtie2 element annotations, thus cases where no sample-specific diversity would be detected if a reference genome were used. A target was selected if its fraction exceeded .5 and its anchor effect size exceeded .9. The 1808 anchor-target pairs satisfying this criterion were BLASTed and sequences mapping to

an accession annotated as Octopus sinesis were further analyzed. BLAST hits were merged into table X with an indicator variable for whether the sequence was queried. X/Y had no BLAST hit. Selected anchor-targets are submitted to BLAST with parameters:

-db nt -evalue 0.1 -task blastn -dust no -word_size 24 -reward 1 -penalty -3 -max_target_seqs 4

Generation of Figure 6C.

Anchors are ordered by descending number of observations; the top 200,000 are concatenated with their anchor-targets and submitted to Pfam; these anchor-targets are defined as controls. The best Pfam hit by full sequence e-value is retained for each anchor-target. For each domain, the top 4 domains hit by anchor-targets in Supplemental table 7 and in the controls are plotted, where bar height is the number of anchor-targets that hit the domain with a full-sequence e-value < 0.01.

Generation of contingency table heatmaps.

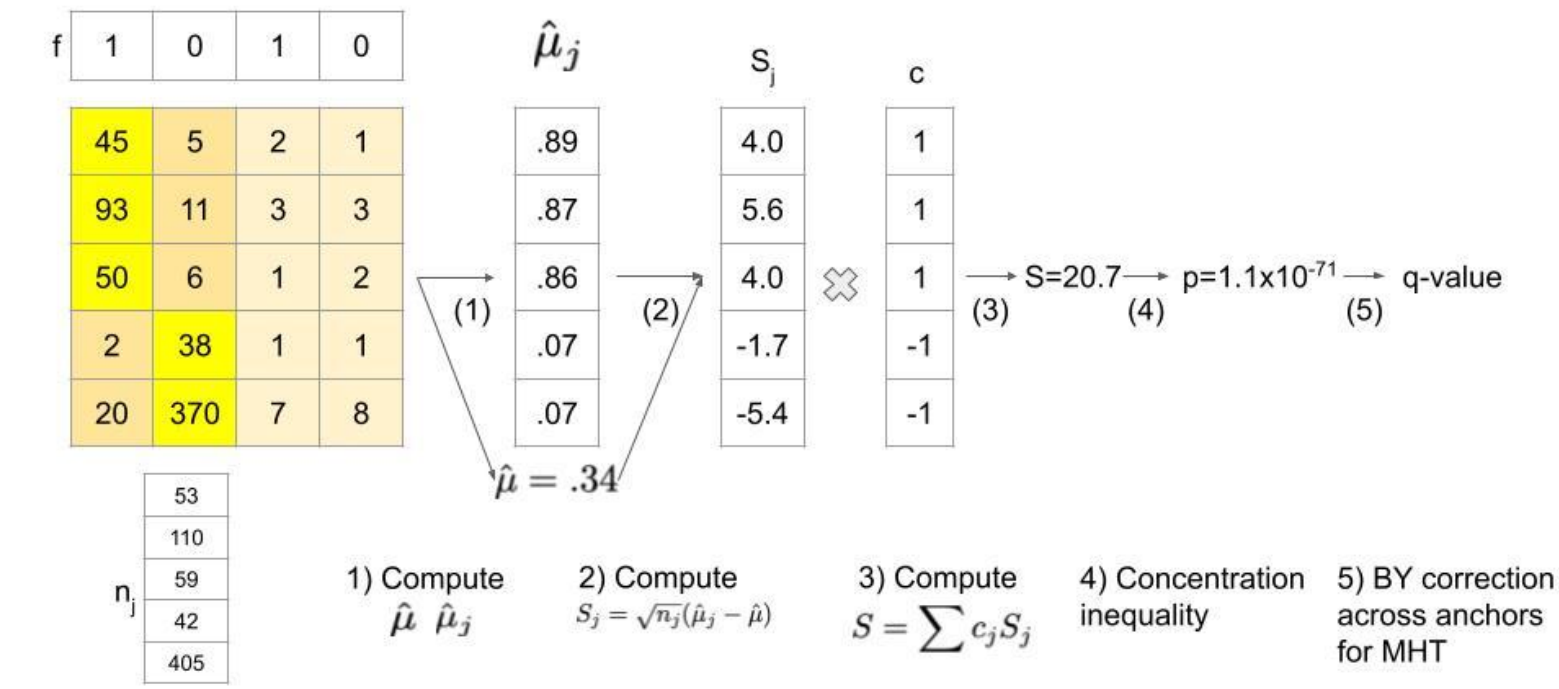
To plot the anchor-target heatmaps, we exclude targets with low counts. Concretely, we filter out targets that occur fewer than 5 times, have less than 5% of the total counts of that anchor, and retain at most the top 10 targets, while ensuring that at least 2 targets are plotted. Then, all samples with fewer than 5 counts are discarded.

Supplementary Materials

Figs. S1 to S7
Supplementary Text
Tables S1 to S7
References

S1

A) P value computation



B) Effect size computation

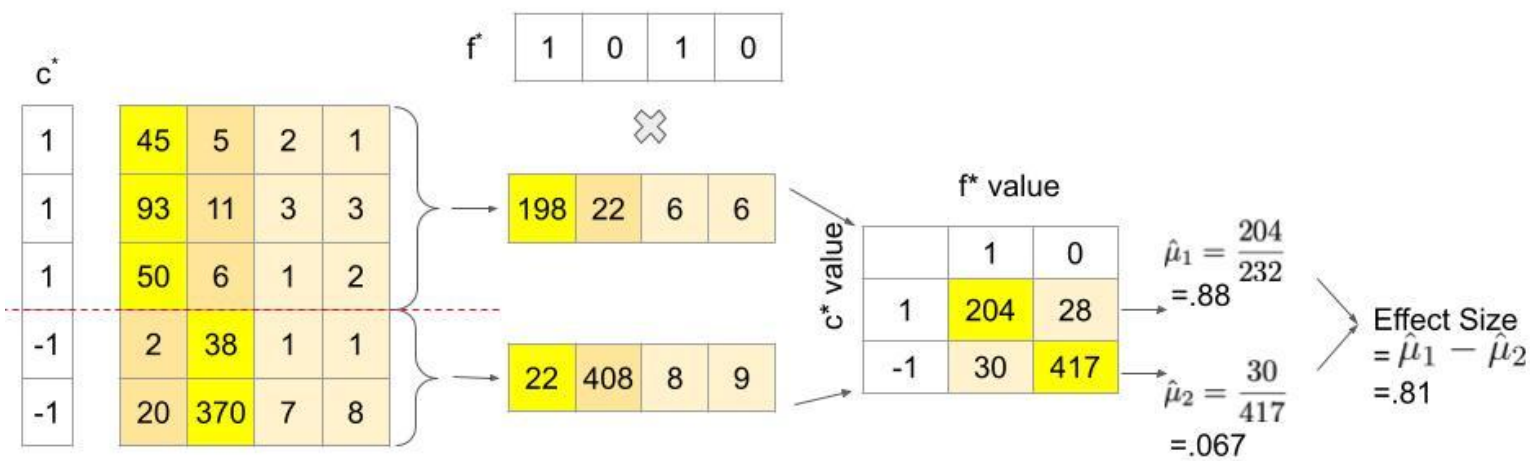


Figure S1

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 $\mu_1 - \mu_2$.

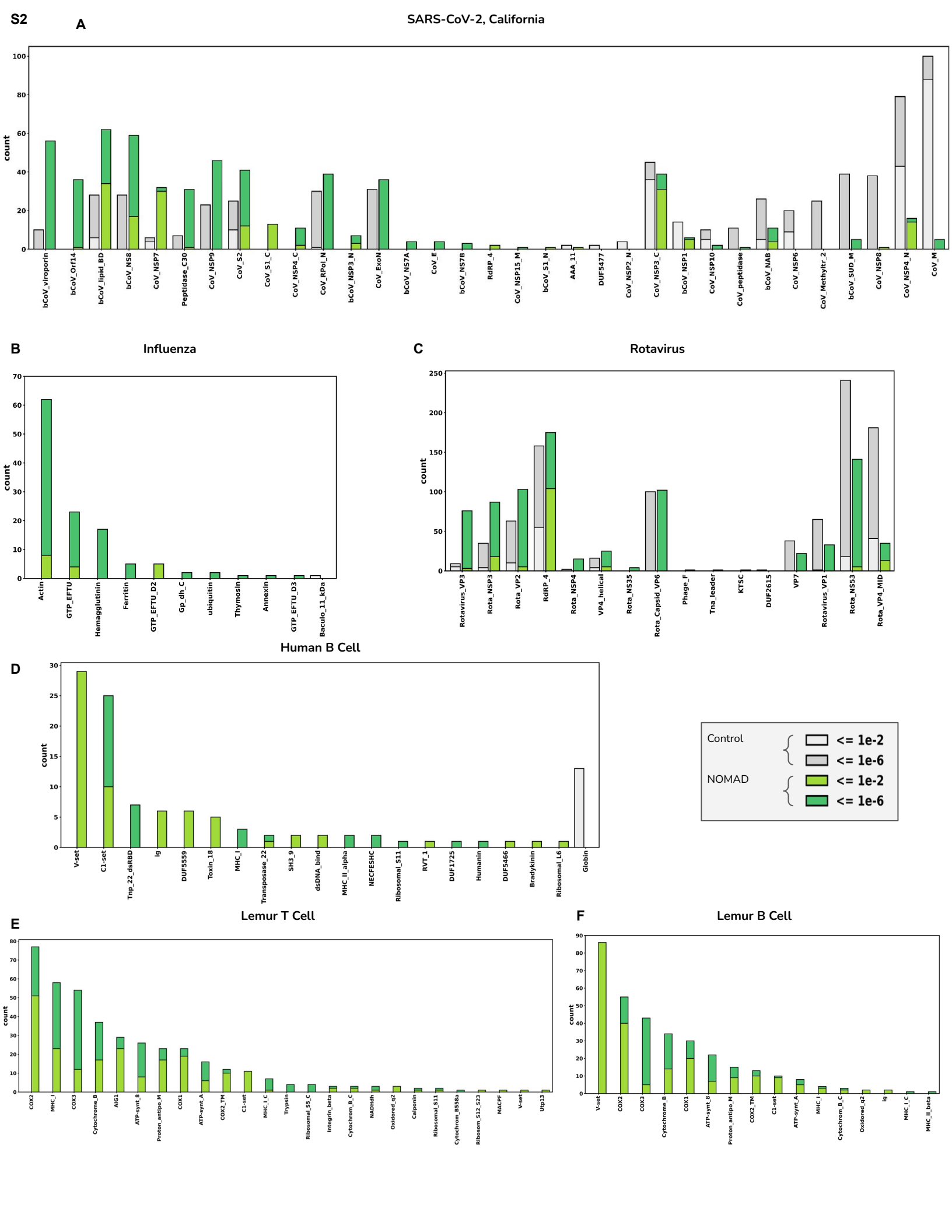

S2
A
SARS-CoV-2, California
count
bCoV_viroporin
bCoV_Orf14
bCoV_lipid_BD
bCoV_NS8
CoV_NSP7
Peptidase_C30
CoV_NSP9
CoV_S2
CoV_S1_C
CoV_NSP4_C
CoV_RPol_N
bCoV_NSP3_N
CoV_ExoN
bCoV_NS7A
CoV_E
bCoV_NS7B
RdRP_4
CoV_NSP15_M
bCoV_S1_N
AAA_11
DUF5477
CoV_NSP2_N
CoV_NSP3_C
bCoV_NSP1
CoV_NSP10
CoV_peptidase
bCoV_NAB
CoV_NSP6
CoV_Methyltr_2
bCoV_SUD_M
CoV_NSP8
CoV_NSP4_N
CoV_M
B
Influenza
count
Actin
GTP_EFTU
Hemagglutinin
Ferritin
GTP_EFTU_D2
Gp_dh_C
ubiquitin
Thymosin
Annexin
GTP_EFTU_D3
Baculo_11_kDa
C
Rotavirus
count
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
D
Human B Cell
count
V-set
C1-set
Tnp_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
Control
<= 1e-2
<= 1e-6
NOMAD
<= 1e-2
<= 1e-6
E
Lemur T Cell
count
COX2
MHC_I
COX3
Cytochrome_B
AIG1
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
NADHdh
Oxidored_q2
Calponin
Ribosomal_S11
Cytochrom_B558a
Riboson_S12_S23
MACPF
V-set
Utp13
F
Lemur B Cell
count
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

Figure S2

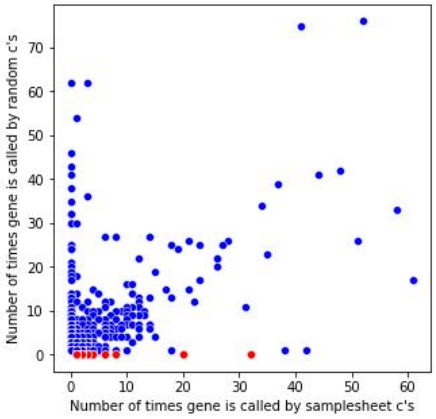
NOMAD protein profile hits to the Pfam database (greens) and control (greys); ordered by enrichment in NOMAD hits compared to control; all NOMAD anchors were used as input, without effect size filters.

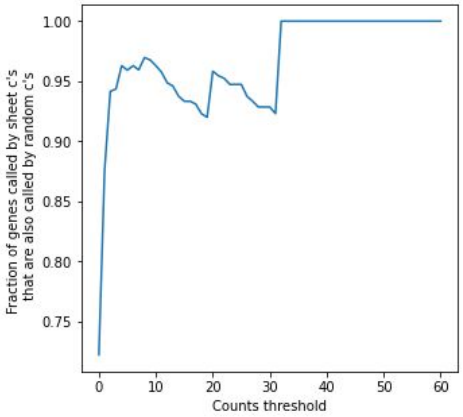
A. Protein profile analysis of NOMAD significant anchors from California data (SRR15881549), before viral strain divergence in the spike had been reported (Gorzynski *et al.*, 2020) serving as a negative control.
B. Protein profile analysis of NOMAD significant anchors from influenza-A data (SRP294571).
C. Protein profile analysis of NOMAD significant anchors from rotavirus breakthrough cases (SRP328899).
D. Protein profile analysis of NOMAD significant anchors from *Microcebus* spleen B cells, from the Tabula Microcebus consortium.
E. Protein profile analysis of NOMAD significant anchors from human T cells from donor 1, from the Tabula Sapiens consortium.
F. Protein profile analysis of NOMAD significant anchors from *Microcebus* natural killer T cells from the Tabula Microcebus consortium.

S3

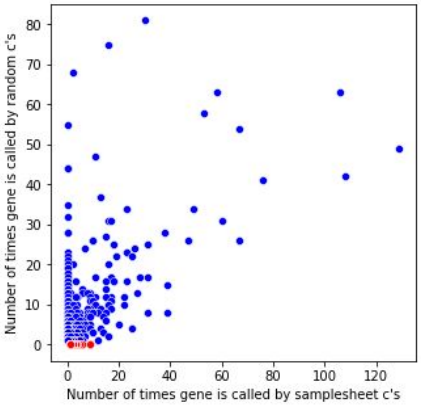
A

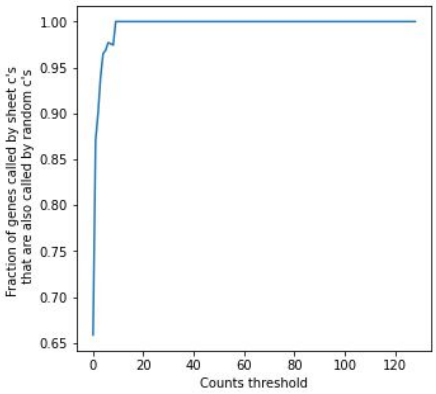
Donor 1



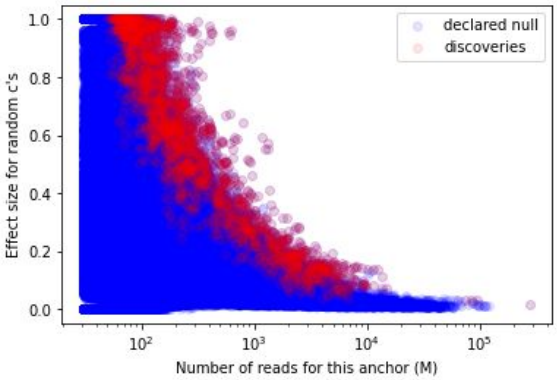


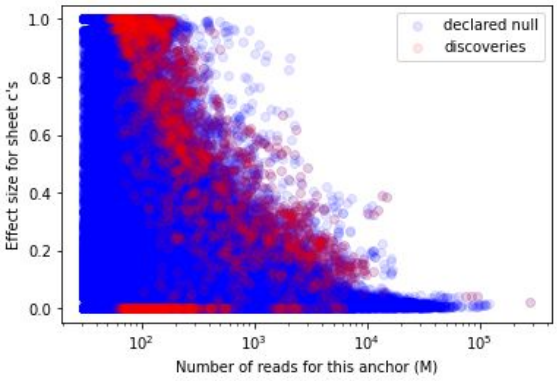
Donor 2

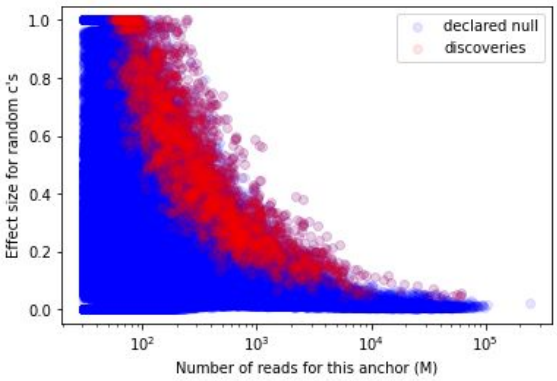




B







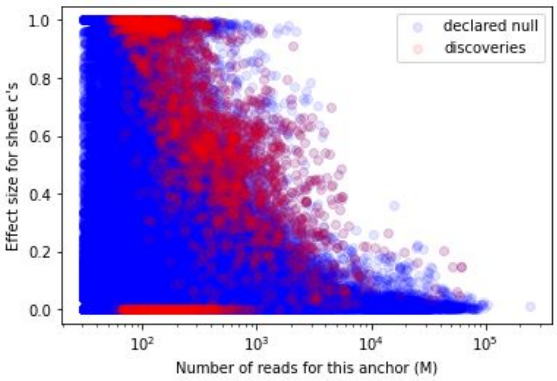
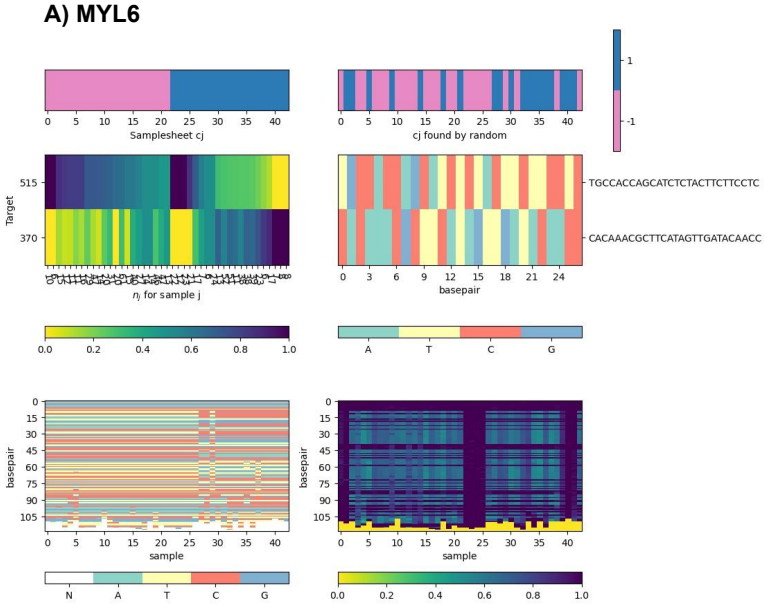


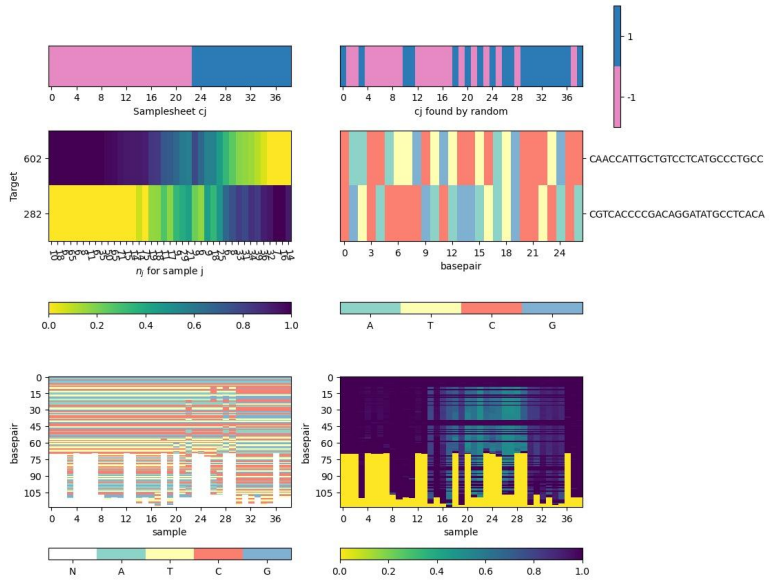
Figure S3

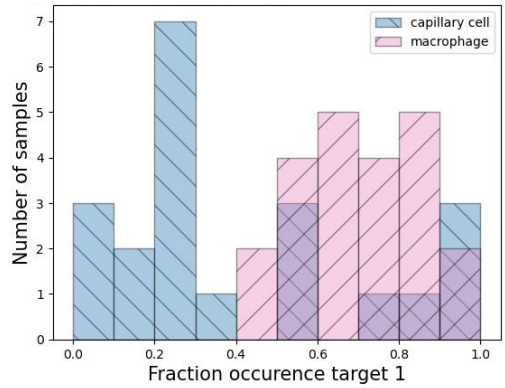
A. Random c's can recover samplesheet c's. For the HLCA dataset, of the 3439 anchors (1384 genes) called by the input metadata (samplesheet c's) in donor 1 (BY correction, alpha=.05), we have that 72% of the genes called were also called by NOMAD's selection of random c's (6287 called by anchors by random 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).

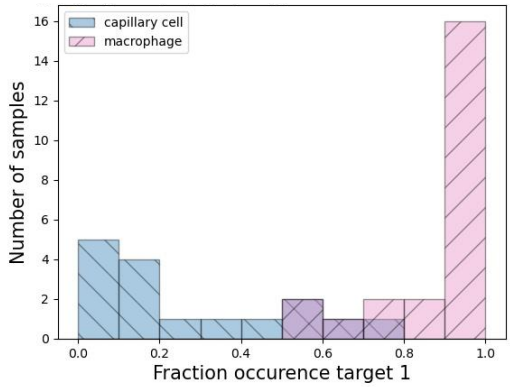
S4

A) MYL6

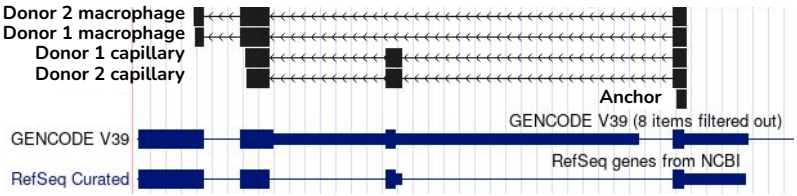








Example consensus sequences

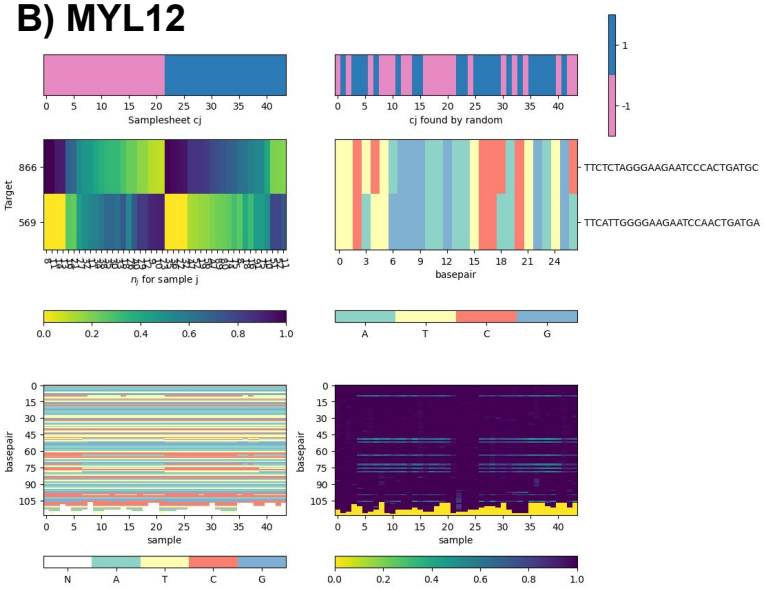


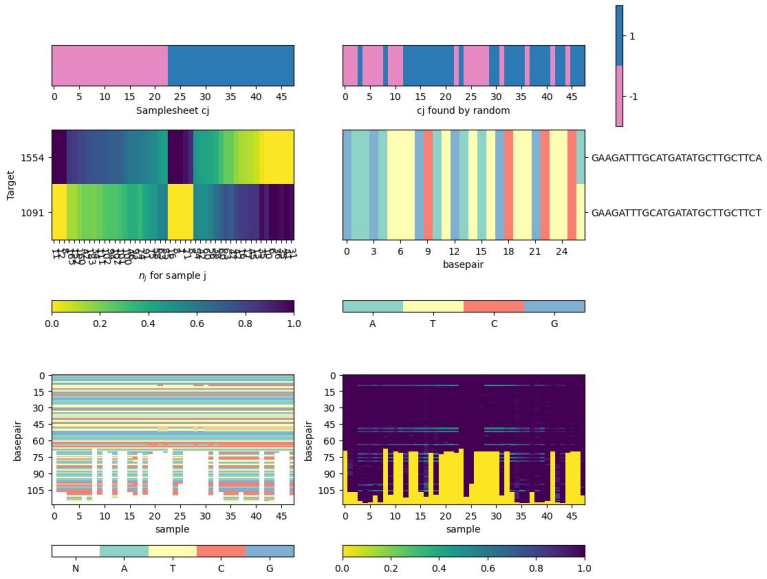
Exon-inclusion dominant ⟷ Exon skipping dominant

S4

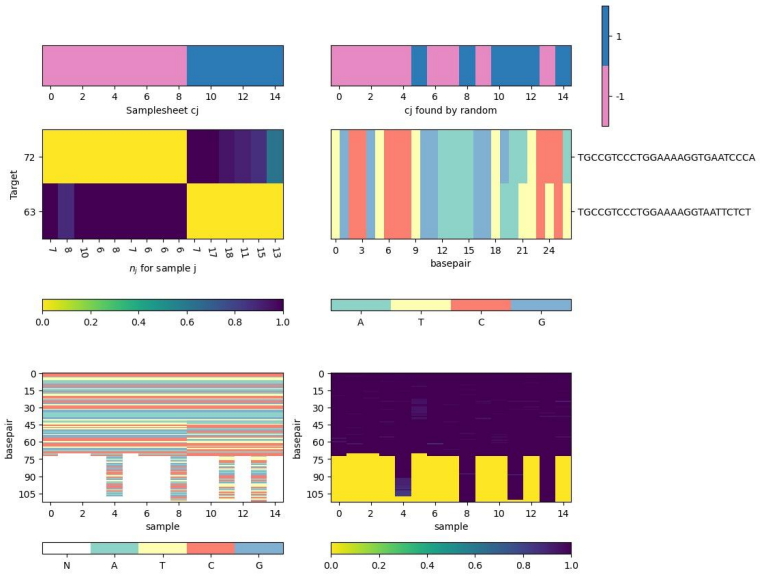
D

B) MYL12





C) HLA-DPB1



D) Human T Cell, HLA-B

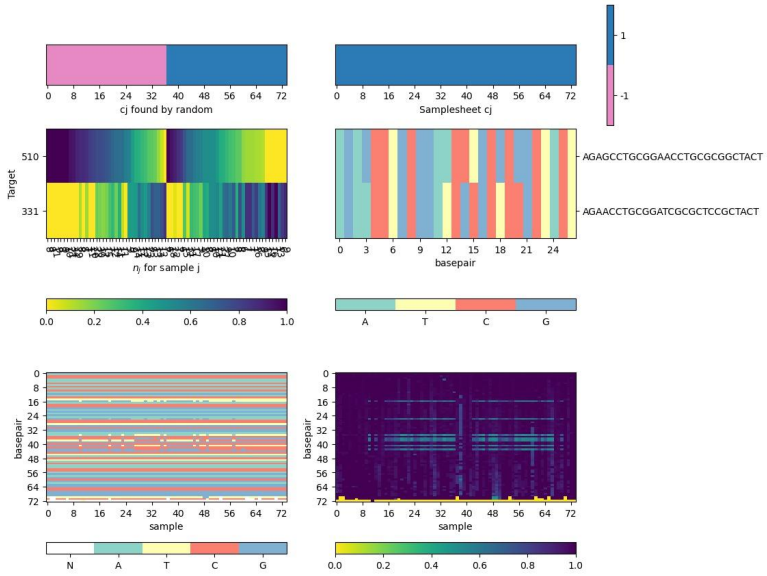


Figure S4

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 cj's used; when samplesheet cjs are available, they will be in the upper left, and the optimizing random cjs will be in the upper right.

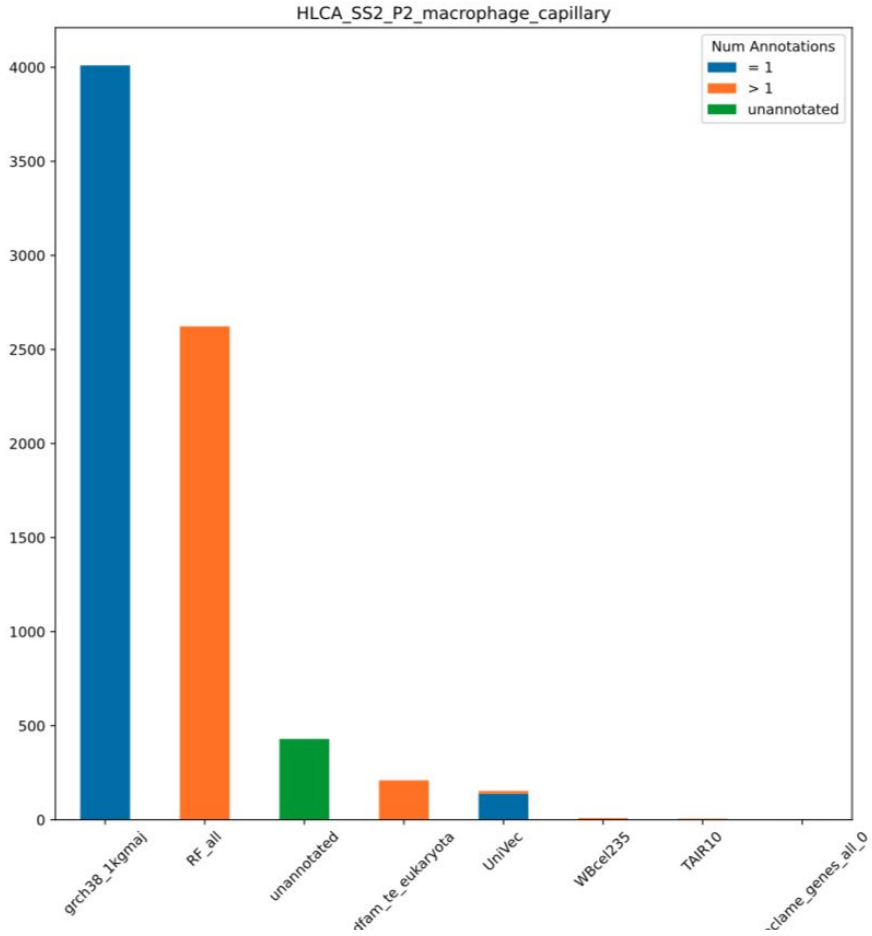
The middle left plot is used to visualize the targets that follow this anchor. Each row represents a target (sequence given in y-tick) corresponding to the row to the left of it in the contingency table. The columns are base pair positions along the sequence of each target. Each nucleotide is color-coded, 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 -1 (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.

1. MYL6
2. MYL12
3. HLA-DPB1
4. Human T cell, HLA-B

S5

A


HLCA_SS2_P2_macrophage_capillary
Num Annotations
= 1
> 1
unannotated
4000
3500
3000
2500
2000
1500
1000
500
0
grch38_1kgmaj
RF_all
unannotated
dfam_te_eukaryota
UniVec
WBcel235
TAIR10
clame_genes_all_0

B

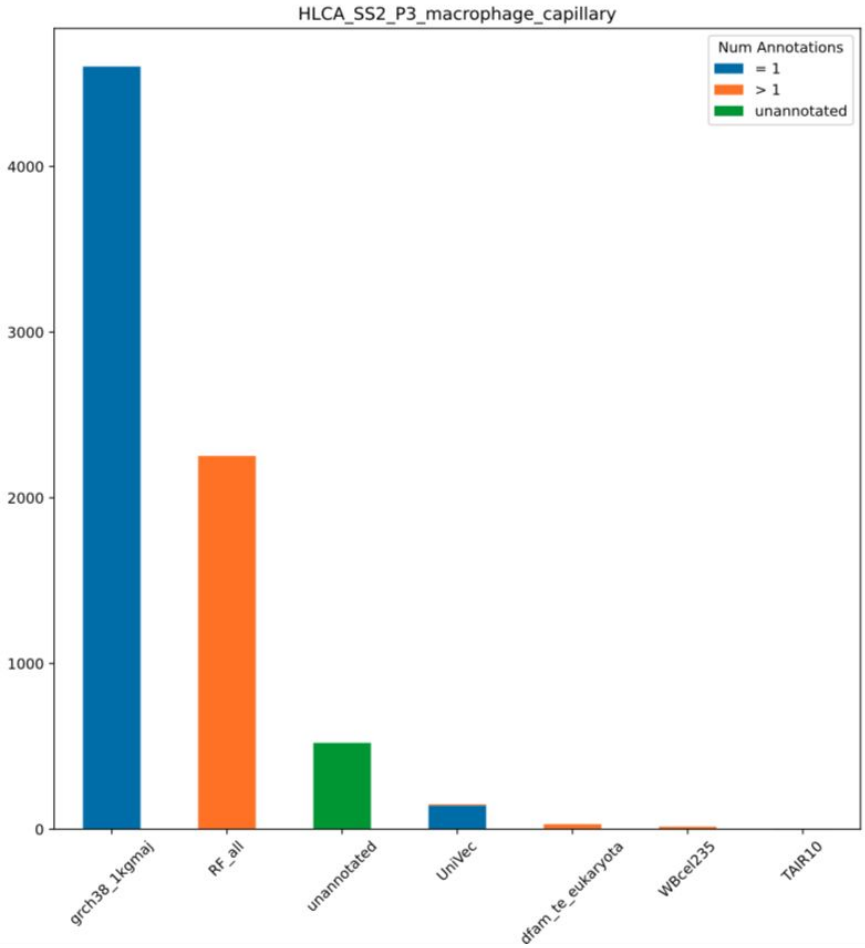
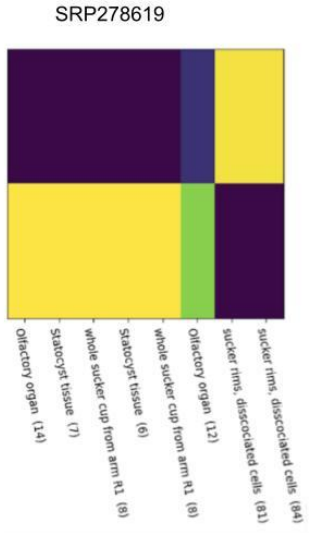
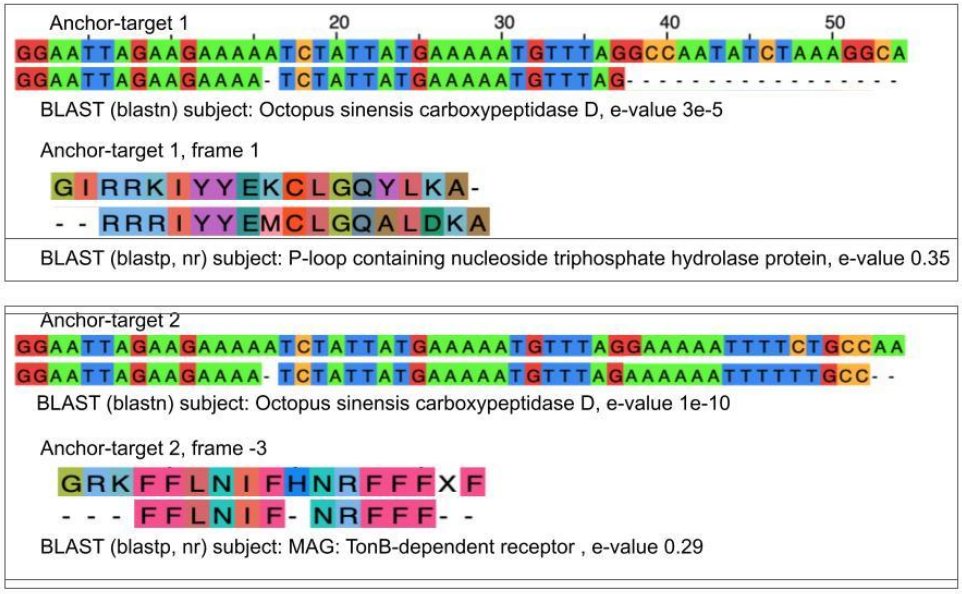

HLCA_SS2_P3_macrophage_capillary
Num Annotations
= 1
> 1
unannotated
4000
3000
2000
1000
0
grch38_1kgmaj
RF_all
unannotated
UniVec
dfam_te_eukaryota
WBcel235
TAIR10

Fig. S5: Element Annotation Bar Plots

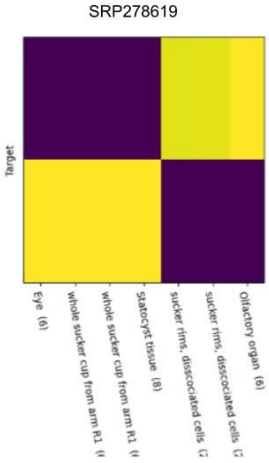
Element annotation bar plots were created with the additional summary files from NOMAD output. To quantify the most frequently occurring element annotation per anchor, the `anchor_top_ann` column was used. For each anchor, the `anchor_num_ann` column was used to quantify the distribution of anchors with exactly one, more than one, or no element annotations for each `anchor_top_ann` unique value.

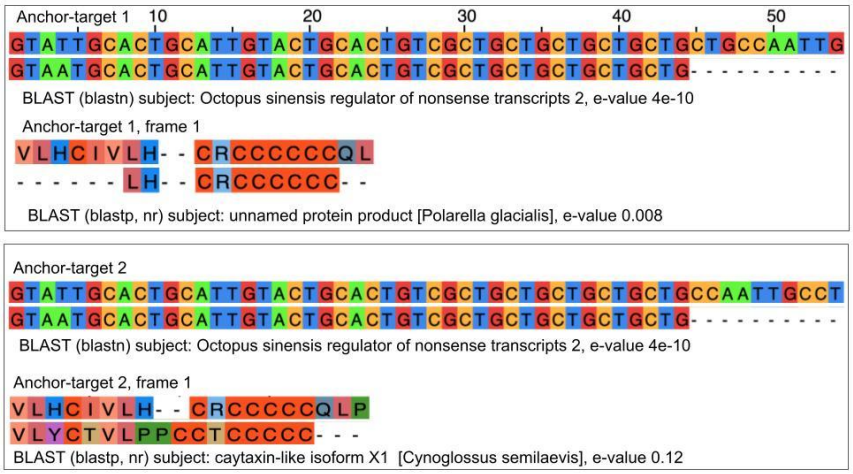
Supplemental figure 6A.



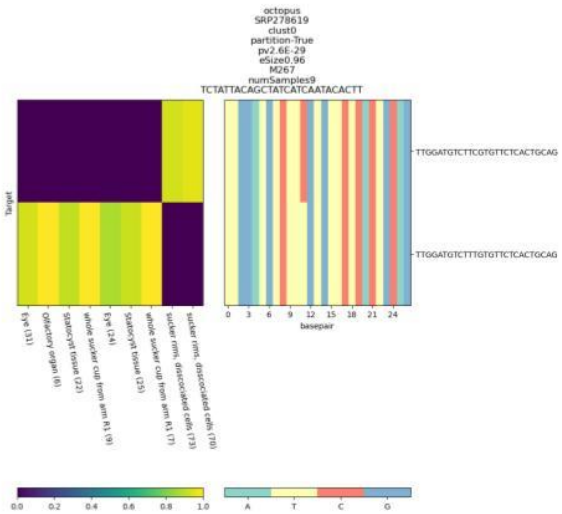


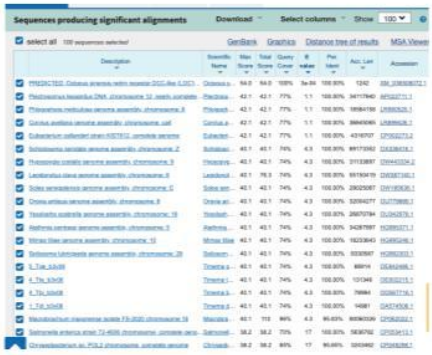
Supplemental figure 6B.





Supplemental Figure 6C.





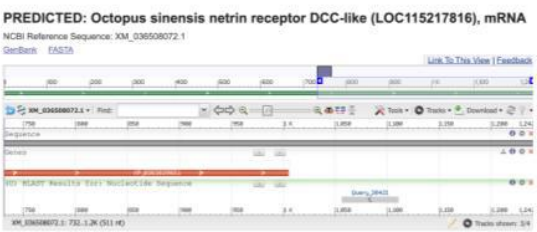


Fig. S6A: Predicted Octopus sinsensis carboxypeptidase (LOC115224523)

Anchor-targets 1 and 2 have differentiating nucleotide composition between positions 39 and 54. Anchor-targets 1 and 2 have a best BLAST (blastn) hit to a common transcript in the predicted Octopus sinensis carboxypeptidase D (LOC115224523), with e-values of 3e-5 and 1e-10 respectively. Anchor-target 1's best hit has 97% identity with a length 37 transcript; this 37mer is a substring of the 52mer to which anchor-target 2 has a 94% identity match. Anchor-targets 1 and 2 have no reported Pfam hits after in-silico translation to 6 frames of amino-acid sequence. The best BLAST (blastp, nr) result for in-silico translated anchor-target 1 is a frame 1 hit with 88% query cover, 76% identity, and e-value of 0.35 to P-loop containing nucleoside triphosphate hydrolase protein [Aspergillus leporis]. For anchor-target 2, the best BLAST (blastp) result is in antisense frame 3, with 70% query cover, 91.67% identity, and e-value of 0.29 to MAG: TonB-dependent receptor [Cryomorphaceae bacterium MED-G14]. In SRP*619, sucker rims use only anchor-target 1, olfactory organ uses both anchor-target 2, and statocyst tissue and whole sucker cup use only anchor-target 2. In PRNJA*, subesophageal brain use only anchor-target 1, while ova, testes, and skin use anchor-targets 1 and 2.

Fig. S6B: Predicted Octopus sinensis regulator of nonsense transcripts 2 (LOC115223858)

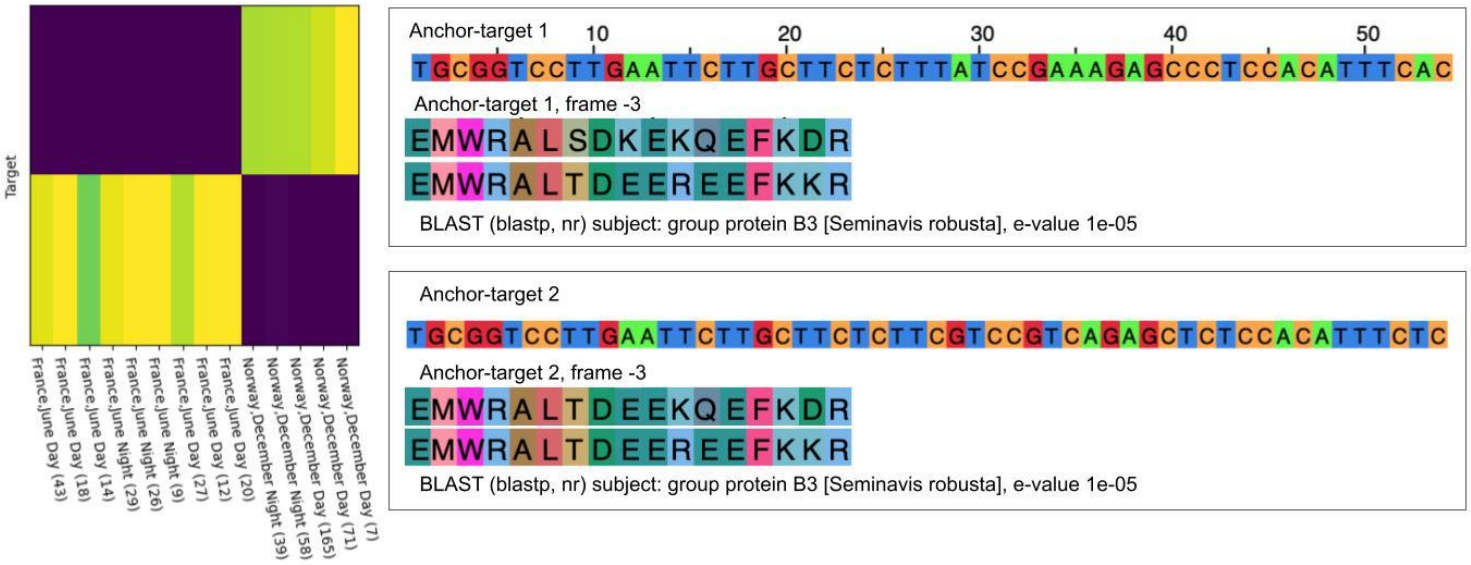
Inclusion of a single CTG repeat differentiates anchor-target 1 from anchor-target 2 in positions 46-54. Anchor-targets 1 and 2 have a best BLAST (blastn) hit to a common transcript in the predicted Octopus sinensis regulator of nonsense transcripts 2 (LOC115223858), both with e-values of 4e-10. Anchor-targets 1 and 2 both have 98% identity to the length 44 BLAST subject transcript. Anchor-targets 1 and 2 have no Pfam hits after in-silico translation and Pfam search. Anchor-target 1's best blastp (nr) hit is in frame 1, with 55% query cover, 100% identity, and e-value of 0.008 to unnamed protein product [Polarella glacialis]. Anchor-target 2's best blastp (nr) hit is in frame 1, with 83% query cover, 70.59% identity, and an e-value 0f 0.12 to caytaxin-like isoform X1 [Cynoglossus semilaevis]. Usage of anchor-target 1 is exclusive to the olfactory organ and sucker rim; usage of anchor-target 2 is exclusive to the eye, whole sucker cup from arm, and statocyst tissue.

Fig. S6C: Predicted Octopus sinensis netrin receptor DCC-like (LOC115217816)

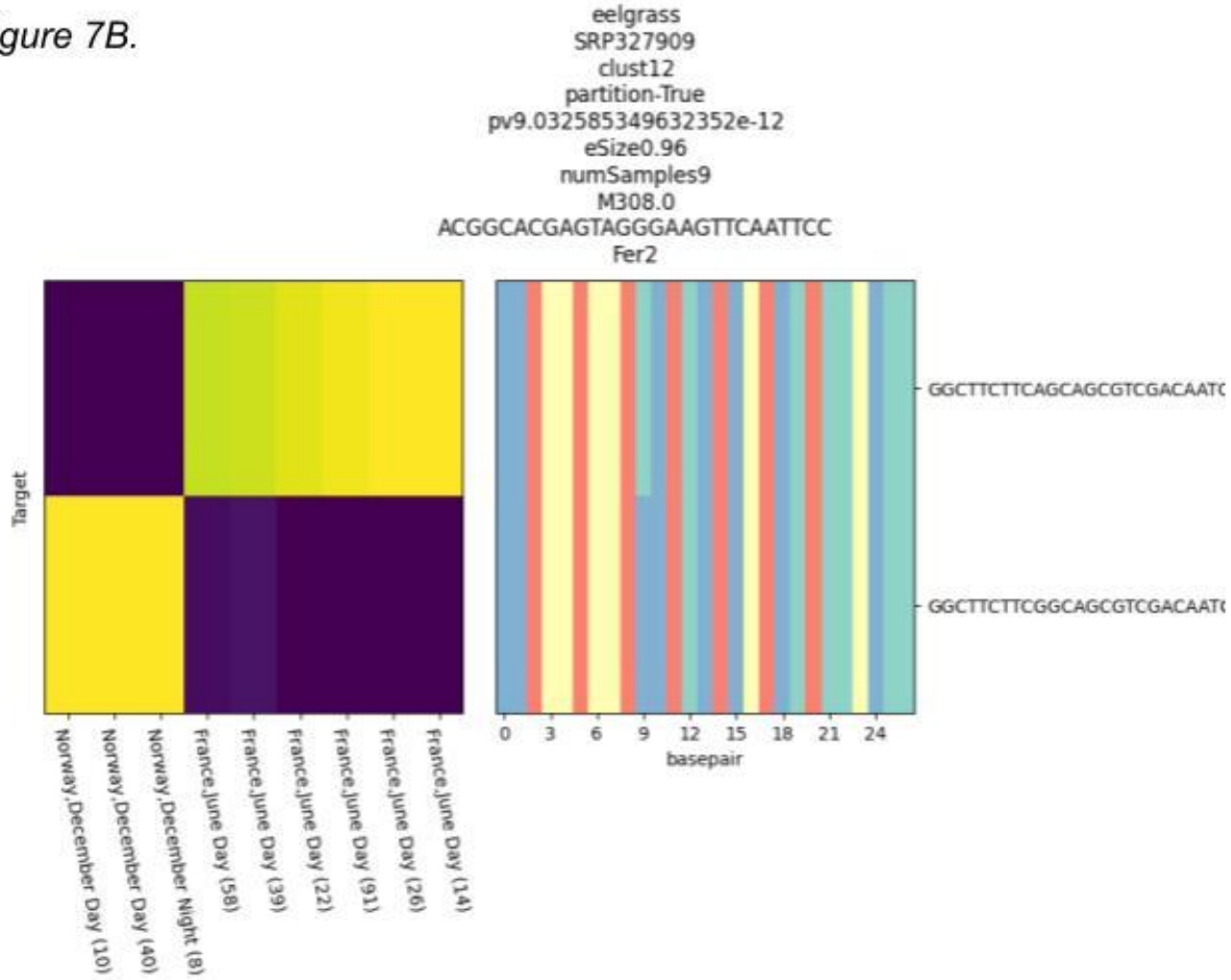
Sample-target heatmap and target nucleotide composition heatmaps are shown with statistics from NOMAD. BLAST output is shown at right. Target 1 is observed only in FASTQs annotated as sucker rims, dissociated cells, while target 2 is observed in eye, olfactory organ, whole sucker cup from arm, and statocyst tissue.

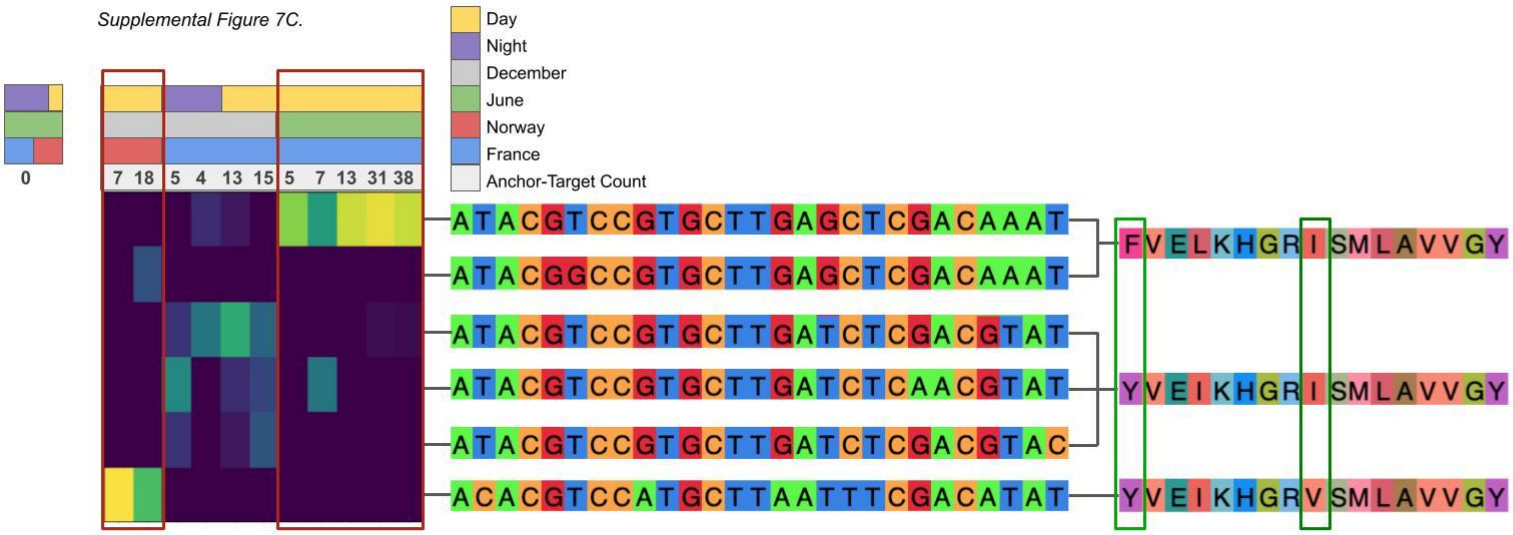
S7

SRP327909 *Supplemental figure 7A.*

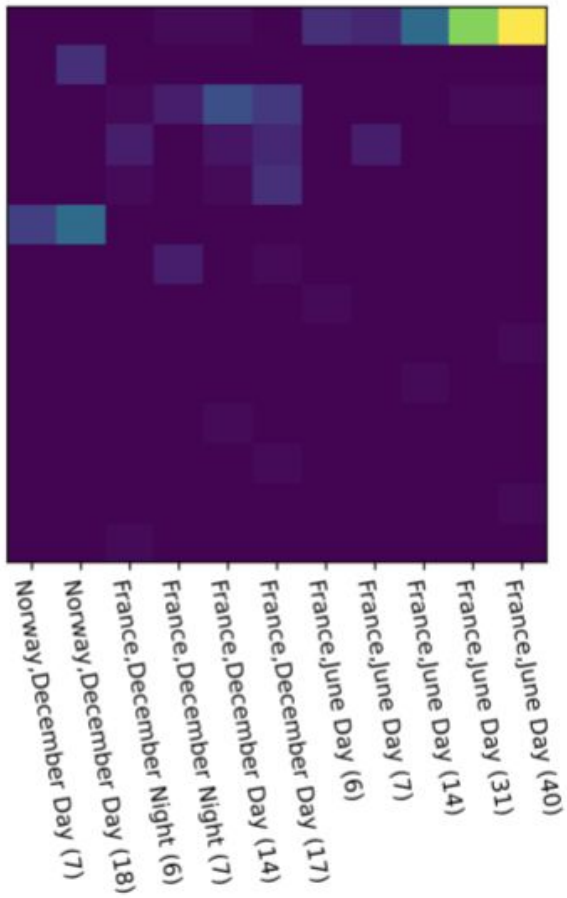


Supplemental Figure 7B.





Supplemental Figure 7D.




```
[[ 0   0   0   1   1   0   5   4  13  30  37]
 [ 0   5   0   0   0   0   0   0   0   0   0]
 [ 0   0   1   3   9   6   0   0   0   1   1]
 [ 0   0   3   0   2   4   0   3   0   0   0]
 [ 0   0   1   0   1   5   0   0   0   0   0]
 [ 7  13   0   0   0   0   0   0   0   0   0]
 [ 0   0   0   3   0   1   0   0   0   0   0]
 [ 0   0   0   0   0   0   1   0   0   0   0]
 [ 0   0   0   0   0   0   0   0   0   0   1]
 [ 0   0   0   0   0   0   0   0   1   0   0]
 [ 0   0   0   0   1   0   0   0   0   0   0]
 [ 0   0   0   0   0   1   0   0   0   0   0]
 [ 0   0   0   0   0   0   0   0   0   0   1]
 [ 0   0   1   0   0   0   0   0   0   0   0]]
```

Fig. S7A: HMG-Box

Anchor-targets 1 and 2 are differentiated by contrasting dinucleotides at positions 28-29 and positions 34-35, as well as single-nucleotide variations at positions 41 and 53. Anchor-targets 1 and 2 have no BLAST (blastn) hits. Six-frame in silico translation and Pfam search yields best hits for anchor-targets 1 and 2 to HMG-Box (PF00505.22) with respective e-values of 3.3e-5 and 2.7e-5. Anchor-target 1's best blastp (nr) hit is in frame -3, with 100% query cover, 70.59% identity, and an e-value of 0.007 group protein B3 [Seminavis robusta]. Anchor-target 2's best blastp (nr) hit is in frame -3, with 100% query cover, 82.35% identity, and e-value of 1e-05 to the same amino acid sequence as anchor-target 1, in group protein B3 [Seminavis robusta]. Usage of anchor-target 1 is specific to samples collected in Rovika, Norway in December, and usage of anchor-target 2 is specific to samples collected in Montpellier, France, in June.

Fig. S7B: Fer2

Sample-target heatmap and target nucleotide composition heatmaps are shown with statistics from NOMAD. Observations of target 1 are specific to France in June at daytime; observations of target 2 are specific to Norway in December, both at night and day.

Fig. S7C: Chlorophyll A-B binding protein

Fraction of anchor-targets (rows) shown in each FASTQ (columns) with bars to indicate whether the sample was collected in day or night, December or June, and Norway or France. Multiple sequence alignment of targets corresponding to rows and of amino acids corresponding to targets translated in frame -2. Amino acid sequences have best Pfam hits to Chlorophyll A-B binding protein with a worst e-value of 2.3e-07. Searching with BLAST (blastp) AA sequence 1 has 100% query cover and 100% identity to a transcript in 3 diatom species, AA sequence 2 has 100% query cover and 100% identity to a transcript in 4 diatom species, and AA sequence 3 has a single substitution to a transcript in 4 diatom species (100% query cover, 94.12% identity). Using BLAST (blastn) AT1 had a best hit to a F. solaris (92% query cover, 92% identity), AT2 to E. pelagica (92% query cover, 94.23% identity), AT3 to E. pelagica (96% query cover, 92.31% identity), AT4 to 4 P. tricornutum (94% query cover, 96.08% identity), AT5 to E. pelagica and P. tricornutum (94% query cover, 94.12% identity), and AT6 to F. solaris (96% query cover, 92.31% identity).

Fig. S7D: Chlorophyll A-B binding protein; raw target-sample counts.

A heatmap illustrating counts of anchor-target observations (rows) in each sample where the anchor was observed(columns). At right, the counts matrix used to generate this heatmap.

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 adapters in cases of libraries prepared by adapter 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 (Medhekar and Miller, 2007), 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 (Agresti, 1992). The field's primary focus has been on either small contingency tables (2x2, e.g. Fisher's exact test(Fisher, 1922)), 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). Under the null model, each $D_{j,k}$ is then an independent draw from the common target distribution.

NOMAD test statistics are closely related to existing statistical tests which will be explored in work in preparation. To construct p-values, we first estimate the expectation (unconditional on sample identity) of $f(D_{j,k})$ as $\hat{\mu}$ by collapsing across samples. Next, we aggregate $f(D_{j,k})$ 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 and standard terms:

1) the probability that the statistic $\widetilde{S}$, constructed with unavailable 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

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

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

q-values are computed using Benjamini Yekutieli correction (Benjamini and Yekutieli, 2001) as

$$Q_i^{\mathrm{BY}} = \min\left(\min_{j \geq i} \frac{c(m) p_{(j)}}{j}, 1\right) \quad \text{where} \quad c(m) = \sum_{i=1}^{m} \frac{1}{i}$$

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

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.

NOMAD runs 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

We provide p-values to quantify the visually striking nature of the plots depicting fraction abundance of target 1. Under a null model, where all samples are expressing this target with the same probability, the number of times each sample expresses target 1 is binomial(n_j,p), for common p. As seen from the plots, many samples exhibit highly deviating occurrences (number of observations of target 1 that are far from the expected pn_j. The p-values we provide to this effect are not used in any NOMAD discovery or analysis, and are just used to quantify the visuals.

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.

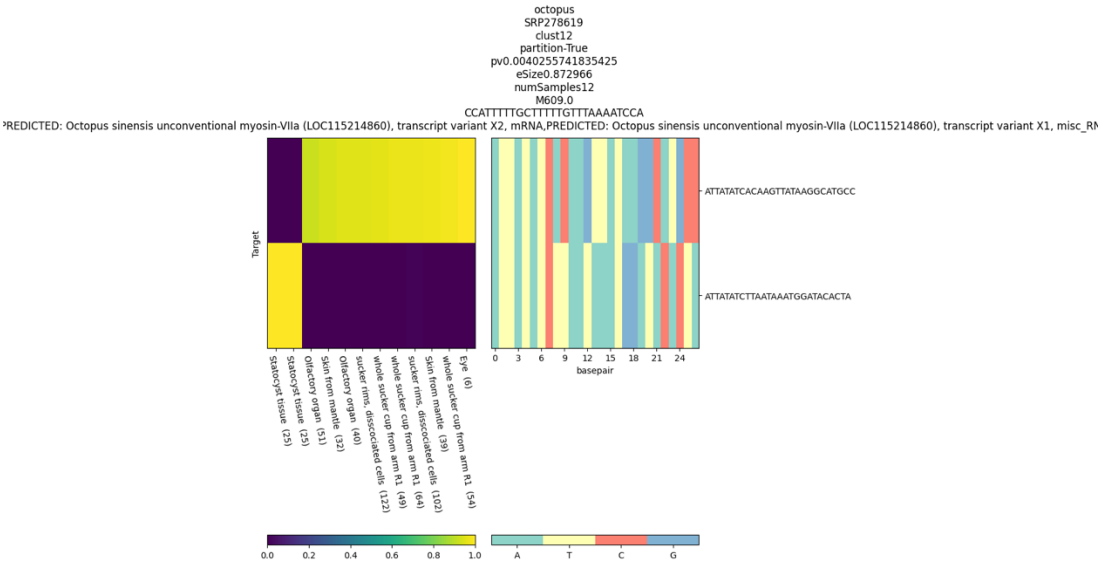
Figure data

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.

Octopus and Zostera (eelgrass) anchor-target alignments

Octopus bimaculoides assembly on NCBI: name = ASM634580v1 accession = GCF_006345805.1
Octopus sinensis assembly on NCBI: name = ASM634580v1 accession = GCF_006345805.1

Octopus bimaculoides/sinensis *myosin-VIIa (Myo7a)*



anchor in blue, targets in green/red.

```
anchor   CCATTTTTGCTTTTTGTTTAAAATCCA
T1       ATTATATCACAAGTTATAAGGCATGCC
T2       ATTATATCTTAATAAATGGATACACTA
```

anchor+target:

```
AT1      CCATTTTTGCTTTTTGTTTAAAATCCAATTATATCACAAGTTATAAGGCATGCC
         ||||||||||||||||||||||||||||||||||||
AT2      CCATTTTTGCTTTTTGTTTAAAATCCAATTATATCTTAATAAATGGATACACTA
```

reverse-complements:

```
AT1      GGCATGCCTTATAACTTGTGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                            ||||||||||||||||||||||||||||||||||||
AT2      TAGTGTATCCATTTATTAAGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
```

extended anchor-targets:

anchor in blue, target in red, extended consensus in black.

```
1 CCATTTTTGCTTTTTGTTTAAAATCCAATTATATCACAAGTTATAAGGCATGCCGGCAACAGAAACGCCAGAATTGAGTA
  ||||||||||||||||||||||||||||||||||||
2 CCATTTTTGCTTTTTGTTTAAAATCCAATTATATCTTAATAAATGGATACACTATTTCCGTCCATATTGGTGATAATAAA
```

reverse-complements:

```
1 TACTCAATTCTGGCGTTTCTGTTGCCGGCATGCCTTATAACTTGTGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                                                  ||||||||||||||||||||||||||||||||||||
2 TTTATTATCACCAATATGGACGGAAATAGTGTATCCATTTATTAAGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
```

Alignment of extended anchor targets as reverse-complements (gives the sense strand) to *O. sinensis* genome.

Numbers are genome coordinates for *Octopus sinensis* chromosome LG8 = NC_043004.1 "Octopus vulgaris isolate Ov201803 linkage group LG8, ASM634580v1". The gene model is LOC115214860.

```
<<< start of first exon (not shown) = 79,843,996
...CATCCCGGATTTTCTACTCAATTCTacCGTct-TGTTGCCGGCATGCCTTATAACTTGTGgtaggta  annotated first exon = "1a"
                |                                                |
                79,844,331                                       79,844,375
                |                                                |
```

```
                 |                          MetProTyrAsnLeu***                                  MetV
1                TACTCAATTCTggCGTttcTGTTGCCGGCATGCCTTATAACTTGTGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                                                                  ::::::::::::::::::::::::::::::::::
2                TTTATTATCACCAATATGGACGGAAATAGTGTaTCCATtTATTAAGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                 |              MetAspGlyAsnSerValSerIleTyr***||                                 |
                 |                                            ||                                 |
                 |                                            |79,932,654              79,932,687    79,932,701
                 |                                            ||                                 |             |
                 |                        ccttttattttctatttcagATATAATTGGATTTTAAACAAAAAGCAAAAATGGTGATTCTTGCAAAGgtaa
                 |                                            |   ***LeuAspPheLysGlnLysAlaLysMetValIleLeuAlaLys
                 79,871,237                                   79,871,281                        annotated exon 2
                 |                                            |
...AATGACTCACTTTATTTATTATCACCAATATGGACGGAAATAGTGTtTCCATaTATTAAGgtaagc  unannotated alternative first exon "1b"
    MetThrHisPheIleTyrTyrHisGlnTyrGlyArgLys***                        (expressed in statocyst tissue)
                                 MetAspGlyAsnSerValSerIleTyr***
```

anchor is in blue; target 1 in green; target 2 in red.
gray = genomic sequence from *O. sinensis*, showing splice signals.
splice dinucleotides double-underlined.
lowercase-orange = SNP differences between *O. sinensis* and anchor-targets (*O. bimaculoides*)
ATGs single-underlined, and translations shown. All upstream ATGs have downstream stop codons shortly after, and the annotated start codon has an upstream stop shortly before. Thus the alternative first exons do not introduce additional protein sequence at the N-terminus.

The *O. bimaculoides* genome assembly contains exon 1a and 1b (though not annotated as such) but not exon 2:

BLAST of extended **anchor-target #1** (reverse-complement) against *O. bimaculoides* genome gives a partial match:

```
Octopus bimaculoides isolate UCB-OBI-ISO-001 chromosome 8, ASM119413v2
Sequence ID: NC_068988.1      Length: 97793173      Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
85.1 bits(93)  5e-15   48/49(98%)     0/49(0%)       Plus/Plus
        Features:
        401698 bp at 5' side: glypican-3
        80171 bp at 3' side: myosin-viia
Query  1        TACTCAATTCTGGCGTTTCTGTTGCCGGCATGCCTTATAACTTGTGATA  49
                ||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  9633645  TACTCAATTCTGGCGTTTCTGTTGCCGGCATGCCTTATAACTTGTGGTA  9633693
```

• On the View Browser, this match lies within XR_008264717.1, annotated as "lncRNA", "uncharacterized LOC128248543". The next gene downstream is an annotated myosin-VIIa gene (LOC106880717; XM_052969897.1). Both genes are on the plus-strand and so LOC128248543 and LOC106880717 could form a single transcriptional unit.

BLAST of extended **anchor-target #2** (reverse-complement) against *O. bimaculoides* genome gives a partial match:

```
Octopus bimaculoides isolate UCB-OBI-ISO-001 chromosome 8, ASM119413v2
Sequence ID: NC_068988.1      Length: 97793173      Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
85.1 bits(93)  5e-15   48/49(98%)     0/49(0%)       Plus/Plus
        Features:
        426544 bp at 5' side: glypican-3
        55325 bp at 3' side: myosin-viia
Query  1        TTTATTATCACCAATATGGACGGAAATAGTGTATCCATTTATTAAGATA  49
                ||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  9658491  TTTATTATCACCAATATGGACGGAAATAGTGTATCCATTTATTAAGGTA  9658539
```

• There are no annotated genome features in the region of this match. It lies between the above match of target #1 and the downstream annotated myosin-VIIa gene (LOC106880717). This matches their arrangement in the *O. sinensis* myosin-VIIa gene (LOC115214860).

The common part of the anchor-targets = GATATAATTGGATTTTAAACAAAAAGCAAAAATGG , which lies in *O. sinensis* exon 2, is the part not matched in the two searches above.

BLAST of this common part against the *O. bimaculoides* genome gives gives ***no match***.
• Thus sequence corresponding to exon 2 is missing from the *O. bimaculoides* genome assembly.

Inspection of the annotated myosin-VIIa transcripts on chromosome 8 / LG8 for *O. bimaculoides* and *O. sinensis* shows that they have different first coding exons (exon 2 in *O. sinensis*).

```
XM_052969897 (bimaculoides): MPQQYFHKTPEQYYCSDITNNAVPKQHTDRYLHLL GDHVWLEPKTKEEFSVAIGARVKFTESGRVLVVDDDGK...
XM_036505518 (sinensis):                                 MVILAK GDHVWLEPKTKEEFSVAIGAKVKFTESGRVLVVDDDGK...
```

Examination of reads for sequence preceding the anchor gives this consensus (in black):

```
TCCAGACTCTGTAAATTTCACCCGGGCACCAATGGCCACAGAAAATTCTTCTTTGGTCTTTGGCTCCAGCCACACATGATCTCCCTTTGCAAGAATCA
CCATTTTTGCTTTTTGTTTAAAATCCA
```

reverse-complement:

```
TGGATTTTAAACAAAAAGCAAAAATGG
TGATTCTTGCAAAGGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGGCCATTGGTGCCCGGGTGAAATTTACAGAGTCTGGA
```

This sequence lies in exon 3 of *O. sinensis*, and the annotated exon 2 of *O. bimaculoides*, and represents the start of the shared protein sequence between them, as shown above. This proves that the exons 1a, 1b, 2, and 3 we observe for myosin-VIIa (based on *O. sinensis*) are indeed connected in *O. bimaculoides*.

Query with pre-sequence consensus (reverse-complement) in the *O. bimaculoides* genome:
(sequence in blue comes from preceding exon.) This gives an identical match.

```
Octopus bimaculoides isolate UCB-OBI-ISO-001 chromosome 8, ASM119413v2
Sequence ID: NC_068988.1      Length: 97793173      Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
163 bits(88)   9e-39   93/95(98%)     2/95(2%)       Plus/Plus
        Features:
        myosin-viia
Query  4        TTCTTGCAAAGGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  63
                |||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  9744916  TTCTTGC--AGGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  9744973

Query  64       CCATTGGTGCCCGGGTGAAATTTACAGAGTCTGGA  98
                ||||||||||||||||||||||||||||||||||||
Sbjct  9744974  CCATTGGTGCCCGGGTGAAATTTACAGAGTCTGGA  9745008
```

Query with pre-sequence consensus (reverse-complement) in the *O. sinensis* genome:
(sequence in blue comes from preceding exon 2.) This has two mismatches (in red).

```
Octopus vulgaris isolate Ov201803 linkage group LG8, ASM634580v1
Sequence ID: NC_043004.1      Length: 107269306     Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
152 bits(82)   2e-35   91/95(96%)     2/95(2%)       Plus/Plus
        Features:
        unconventional myosin-viia
Query  4         TTCTTGCAAAGGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  63
                 |||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  79975491  TTCTTGC--AGGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  79975548

Query  64        CCATTGGTGCCCGGGTGAAATTTACAGAGTCTGGA  98
                 ||||||||||||  |||||||||||||||||||||
Sbjct  79975549  CCATTGGTGCCAAGGTGAAATTTACAGAGTCTGGA  79975583
```

survey of the myosin-VIIa homologs in *O. sinensis* and *O. bimaculoides*

Search of the NCBI Gene database for "octopus sinensis myosin-VIIa" finds six genes, table below.

Name/Gene ID	Description	Location	a.a.	transcripts, protein domains
LOC115224051	unconventional myosin-VIIa	Chromosome LG24, NC_043020.1 (10520014..10643911)	920	one variant. Has **Motor_domain** and IQ domain.
LOC115215798	myosin-VIIa	Chromosome LG9, NC_043005.1 (49378971..49480887)	1,239	two variants, same a.a. size; for X2, longer RNA. **No motor domain**. Has [MyTH4, B41/FERM1_F1_Myosin-VII,

Name/Gene ID	Description	Location	a.a.	transcripts, protein domains
				FERM_C1_MyoVII], [MyTH4, B41/Ubl1_cv_Nsp3_N-like, PH-like].
LOC115214860	unconventional myosin-VIIa	Chromosome LG8, NC_043004.1 (79843996..80027230)	855	one variant. Has **MYSc_Myo7 motor** and IQ domain.
LOC115214969	myosin-VIIa	Chromosome LG8, NC_043004.1 (79707148..79764229, complement)	1,237	one variant. **No motor domain**. Has [MyTH4, B41/FERM1_F1_Myosin-VII, PH-like, FERM_C1_MyoVII], SH3, [MyTH4, B4/FERM2_F1_Myosin-VII, PH-like, FERM_C2_MyoVII].
LOC115229165	unconventional myosin-VIIa-like	NW_021832531.1 (136771..143156)	---	only RNA is labeled as a pseudogene
LOC115217708	unconventional myosin-X	Chromosome LG11, NC_043007.1 (70820204..71380859)	2,422	variant X1, the longest. Has **MYSc_Myo22 motor** and [MyTH4, B41/FERM_F1_DdMyo7_like, PH-like], [MyTH4, B41/FERM_F1_DdMyo7_like, PH-like].

Red is the Myo7a that matches our anchor-targets. The two genes on chromosome LG8 have pale-orange background. Slash "/" means the domains overlap. Square brackets "[]" highlight repeat structure.

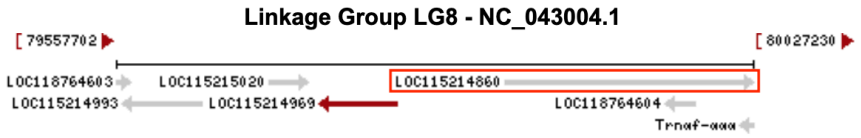
Similarly for *O. bimaculoides* "myosin-VIIa" genes:

Name/Gene ID	Description	Location	a.a.	transcripts, protein domains
LOC106872589	myosin-VIIa	Chromosome 10, NC_068990.1 (55111848..55211343, complement) Alias: OCBIM_22021541mg	1,240	two variants, same a.a. size; this is X2, longer RNA. **No motor domain**. Has [MyTH4, B41/FERM1_F1_Myosin-VII, PH-like], [MyTH4, B41/FERM2_F1_Myosin-VII, PH-like].
LOC106880717	myosin-VIIa	Chromosome 8, NC_068988.1 (9713762..9821311) Alias: OCBIM_22005746mg	2,128	one variant. Has **MYSc_Myo7 motor** domain, [MyTH4, B41/FERM_F1, PH-like, FERM_C1_MyoVII], SH3, [MyTH4, B4/FERM2_F1_Myosin-VII, PH-like, FERM_C2_MyoVII]

The gene of interest on chromosome 8 has pale-orange background.

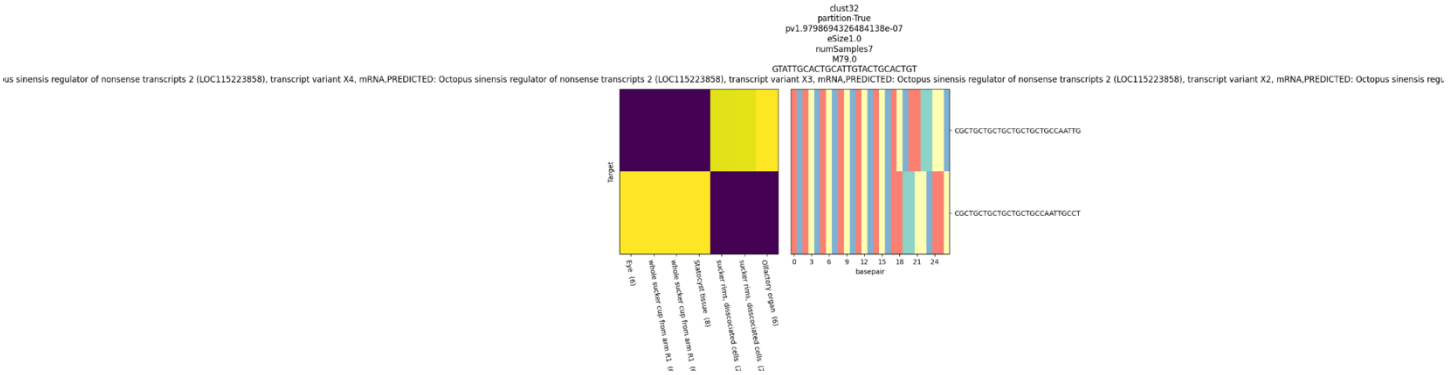
The *O. bimaculoides* gene on chromosome 8 appears to have a complete set of domains, including motor domain and two sets of the repeat domains. However, we know that it is lacks the 5' two exons present in *O. sinensis* myosin-VIIa LOC115214860.

For *O. sinensis*, there are two myosin-VIIa genes on chromosome LG8, LOC115214860 and LOC115214969, adjacent but in head-to-head orientation:



LOC115214860 contains the motor domain, and LOC115214969 contains two sets of the repeat domains. Thus when combined they have the exact same domain content as the single *O. bimaculoides* gene. Therefore, we suspect that LOC115214860 may be improperly oriented in the *O. sinensis* assembly: if inverted, it would then be the same orientation as LOC115214969 and they could form a single transcriptional unit.

Octopus bimaculoides/sinensis *Upf2*



anchor-targets:

anchor in blue, targets in red.

```
A    GTATTGCACTGCATTGTACTGCACTGT
T1   CGCTGCTGCTGCTGCTGCTGCCAATTG
T2   CGCTGCTGCTGCTGCTGCCAATTGCCT

AT1  GTATTGCACTGCATTGTACTGCACTGTCGCTGCTGCTGCTGCTGCTGCCAATTG
AT2  GTATTGCACTGCATTGTACTGCACTGTCGCTGCTGCTGCTGCTG---CCAATTGCCT
```

reverse-complements: (this is the sense strand)

```
AT1rc     CAATTGGCAGCAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCAATAC
AT2rc  AGGCAATTGG---CAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCAATAC
```

BLASTN of AT2rc against nr/nt gives the same partial match to four *O. sinensis* transcript variants, here showing one of them. This match lies in the 3'-UTR.

"regulator of nonsense transcripts 2" is also known as Upf2, a subunit involved in nonsense-mediated decay.

```
PREDICTED: Octopus sinensis regulator of nonsense transcripts 2 (LOC115223858), transcript variant X1, mRNA
Sequence ID: XM_036513028.1   Length: 5101   Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
76.1 bits(83)  4e-10   43/44(98%)     0/44(0%)       Plus/Plus
Query  11    CAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCAATAC  54
             ||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  4813  CAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCATTAC  4856
```

Further inspection shows this more extensive partial match:

```
          CAATTGGCAG-CAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCAATAC          anchor-target1 rev-comp
       AGGCAATTGG----CAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCAATAC          anchor-target2 rev-comp
       |||||| |||    ||||||||||||||||||||||||||||||||||||||||||| |||
CAGCAAAAGGCAA-TGGCGGCCAGCAGCAGCAGCAGCGACAGTGCAGTACAATGCAGTGCATTACCAGTGCAGT  sinensis X1
```

BLASTN of AT2rc against the *O. sinensis* genome finds the same hit as in the transcript above.

BLASTN of AT2rc against the *O. bimaculoides* genome gives only short hits.

The gene model for *O. sinensis* Upf2 is LOC115223858.

The gene model for *O. bimaculoides* Upf2 is LOC106869790.

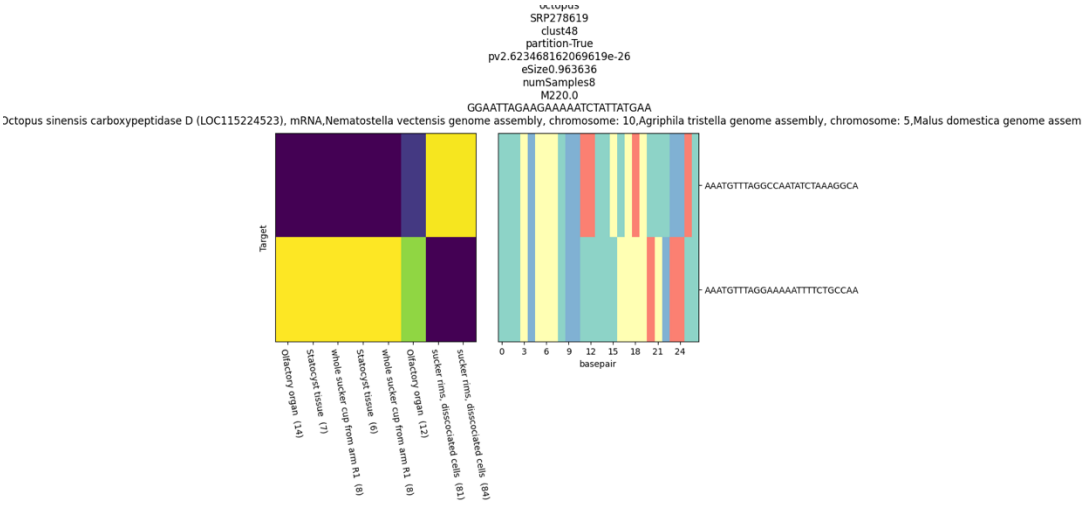
Alignment of transcript variants from each: *bimaculoides* = XM_014915650.2, *sinensis* = XM_036513028.1 shows very high identity throughout the coding sequence, but abrupt divergence from just before the stop codons (marked in red). The match to anchor-target is shown in blue background.

The *O. bimaculoides* genome assembly appears to be missing sequence corresponding to what is found in *O. sinensis*, and this may have impacted automated transcript annotations in *bimaculoides*.

```
            3690      3700      3710      3720      3730
bimac_ ATCAACACCCCAAGGGTGCTCCTGATGCAGACTTGATCTTCGGATCAAAG----------
       :::::::::::::::::::::::::::::::::::::::::::::::::
sinen_ ATCAACACCCCAAGGGTGCTCCTGATGCAGACTTGATCTTCGGATCAAAGGAGCGTTGAA
            4610      4620      4630      4640      4650      4660


bimac_ --------------TGAG-------------------------------------TTGT
                     ::::                                     ::::
sinen_ GGAACTATCAAGCTTGAGGAAAGTCTACGACTACTCGTCCAGGCCAACAACGGTTATTGT
            4670      4680      4690      4700      4710      4720

    3740      3750      3760      3770
bimac_ CACTGCTTCCCATACTTCAGCTGAATTTATTTAAC-----------------GTG-----
         ::: :: ::       :::::: :: : : :::                 :::
sinen_ GCCTGTTTGCC-------AGCTGACTTCACTAAACCGATATTTCTGGCTGTAGTGGCGGC
```

```
            4730              4740      4750      4760      4770

       3780
bimac_ -ATG-------------------------------------------------------
        :::
sinen_ AATGCAGAGCAGTTAGAGCAGCAAAAGGCAATGGCGGCCAGCAGCAGCAGCAGCGACAGT
         4780      4790      4800      4810      4820      4830

                       3790       3800
bimac_ --------ATGTGTTCTCAG-GCATCTCCA---CACT-----------------------
                  :::      ::: ::::  :::   :: :
sinen_ GCAGTACAATG------CAGTGCATTACCAGTGCAGTGCTGCGTTGTGCTGTGTGTGTGT
         4840            4850      4860      4870      4880

         3810
bimac_ -TGTGTATG--------------------------------------------------
        ::::::::
sinen_ GTGTGTATGTGTGTGTTGTGCCAACCAGCAAACACGGCCTTTTACTGAAAAAGCAGCCAA
     4890      4900      4910      4920      4930      4940

                                                            3820
bimac_ ----------------------------------------------CAAG-TAAGA--
                                                     :::: :::::
sinen_ AGCAAAGGTGACAGTGAGGTTGCTCCTCCTCCACCACCACCACCACCAAGATAAGATA
     4950      4960      4970      4980      4990      5000


bimac_ ------------------------------------------------------------

sinen_ TAATATATACACACATCACATACAGACATATAGACACACCCCCTCCCTCTTTCTCCTCTC
     5010      5020      5030      5040      5050      5060


bimac_ --AAA
         :::
sinen_ ACAAA
     5070
```

Octopus bimaculoides/sinensis *carboxypeptidase-D (CPD)*



anchor-targets:

anchor in blue, targets in red.

```
A    GGAATTAGAAGAAAAATCTATTATGAA
T1   AAATGTTTAGGCCAATATCTAAAGGCA
T2   AAATGTTTAGGAAAAATTTTCTGCCAA

AT1  GGAATTAGAAGAAAAATCTATTATGAAAAATGTTTAGGCCAATATCTAAAGGCA
AT2  GGAATTAGAAGAAAAATCTATTATGAAAAATGTTTAGGAAAAATTTTCTGCCAA
```

I note that AT2 in particular has a repeat structure, underlined.

reverse-complements: (this will turn out to be the sense strand)

```
AT1rc TGCCTTTAGATATTGGCCTAAACATTTTTCATAATAGATTTTTCTTCTAATTCC
AT2rc TTGGCAGAAAATTTTTCCTAAACATTTTTCATAATAGATTTTTCTTCTAATTCC
```

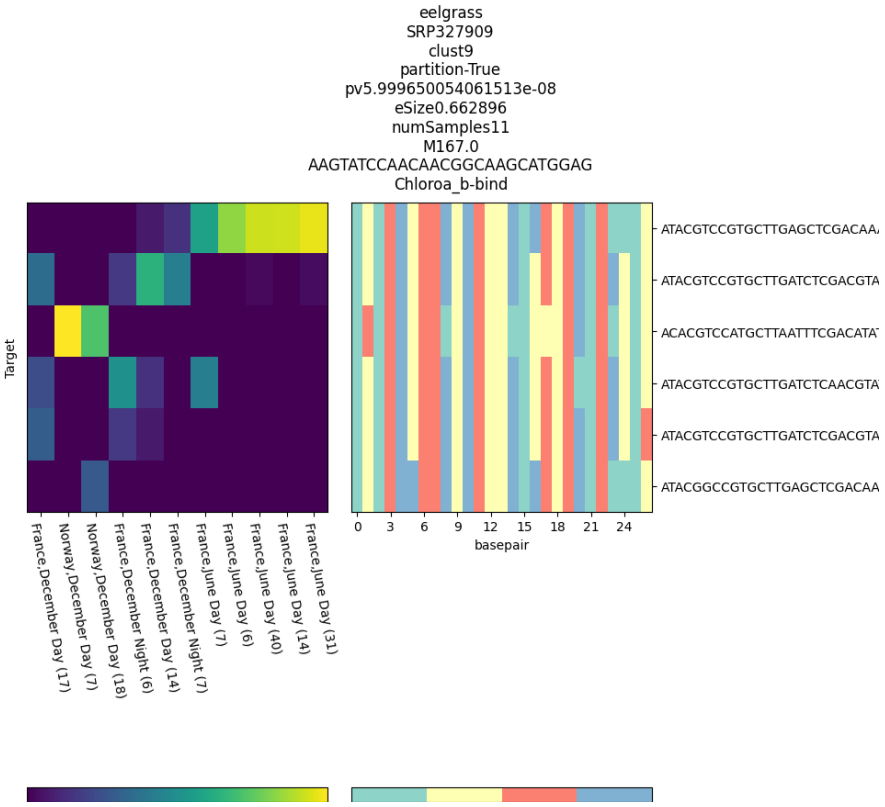
BLAST only finds these sequences in the *O. sinensis* genome, but **not** in the *O. bimaculoides* genome.

The hits are in *O. sinensis* chromosome LG25 (ASM634580v1) = accession NC_043021.1, and lie within the 3' UTR of a carboxypeptidase D gene (there is another on chromosome LG16).

AT1 represents a 13 nt **deletion** relative to the genome and AT2.
(The minus-strand of the genome and reverse-complements of the anchor-targets are shown, giving the mRNA sense strand.)

```
AT1rc                    TGCCTTTAGATATTGG-------------CCTAAACATTTTTCATAATAGATTTTTCTTCTAATTCC
                         ||||||||||||||||              |||||||||||||||||||||||||| |||||||||||
ACACATGATTTGCCGGTGCCATTTTGCCTTTAGATATTGGGCAAAAAATTTTTTCTAAACATTTTTCATAATAGATTTT-CTTCTAATTCCTCATTTTGCACCCCCCCAC
                                        ||||||||||| ||||| |||||||||||||||||||||||| |||||||||||
AT2rc                                   TTGGCAGAAAA-TTTTTCCTAAACATTTTTCATAATAGATTTTTCTTCTAATTCC
```

Zostera marina *fucoxanthin chlorophyll a/c protein (domain Chloroa_b-bind)*



anchor in blue, target in red, extended consensus in black.

AT1 extended

```
AAGTATCCAACAACGGCAAGCATGGAG
ATACGTCCGTGCTTGAGCTCGACAAAT
CGGAGGCGATCGAAGGTTTCTTGATCTCCATCGGCAACCAAACCAAGAGGATCGAAGAATCCAAGAGGAGGTTGAGCACCCA
```

protein translation in frame -3:

```
GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY
```

BLASTN of anchor-target-consensus gives no hits in the *Z. marina* genome.
BLASTN of anchor-target-consensus against nr/nt: top distinct species (multiple hits for each species; three examples below) are all **diatoms**. Although high-scoring, they are only ~80% identity, so the true species that AT1 comes from is not in the database.
The genome assemblies are not annotated with gene models.

```
Epithemia pelagica genome assembly, chromosome: 12
Sequence ID: OX337239.1       Length: 3305232       Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
```

```
133 bits(147)  1e-26   111/136(82%)   0/136(0%)      Plus/Minus
Query  1        AAGTATCCAACAACGGCAAGCATGGAGATACGTCCGTGCTTGAGCTCGACAAATCGGAGG  60
                |||||||||||||||  | | |||  |||| || |||||||||| ||||||  | ||||||
Sbjct  1537005  AAGTATCCAACAACACCCAACATACAGATGCGGCCGTGCTTGATCTCGACGTAACGGAGG  1536946

Query  61       CGATCGAAGGTTTCTTGATCTCCATCGGCAACCAAACCAAGAGGATCGAAGAATCCAAGA  120
                || |||||  ||||||| || |||||||||||  |  ||||| ||||||||||| ||||||
Sbjct  1536945  CGGTCGAACTTTTCTTGGTCACCATCGGCAACAAGGCCAAGGGGATCGAAGAAGCCAAGA  1536886

Query  121      GGAGGTTGAGCACCCA  136
                |||| ||| || ||||
Sbjct  1536885  GGAGCTTGGGCGCCCA  1536870
...
Fistulifera solaris JPCC DA0580 DNA, chromosome 4_l, complete sequence
Sequence ID: AP026133.1       Length: 1362389       Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
131 bits(144)  3e-26   108/132(82%)   0/132(0%)      Plus/Plus
Query  3       GTATCCAACAACGGCAAGCATGGAGATACGTCCGTGCTTGAGCTCGACAAATCGGAGGCG  62
               |||||| |||||||||||||||||||||| |||||||| |  || || ||||||||||||
Sbjct  396744  GTATCCGACAACGGCAAGCATGGAGATACGGCCGTGCTTCAATTCAACGAATCGGAGGCG  396803

Query  63      ATCGAAGGTTTCTTGATCTCCATCGGCAACCAAACCAAGAGGATCGAAGAATCCAAGAGG  122
                |||||  | ||||| || |||||||  || |  || || |||||||| |||||||| ||
Sbjct  396804  GTCGAACTTCTCTTGGTCGCCATCGGTGACAAGGCCGAGGGGATCGAAAAATCCAAGCGG  396863

Query  123     AGGTTGAGCACC  134
                ||||| |||||
Sbjct  396864  CGGTTGGGCACC  396875
...
Phaeodactylum tricornutum isolate CCAP 1055/1 genome assembly, chromosome: 2
Sequence ID: OU594943.1       Length: 2499861       Number of Matches: 4
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
126 bits(139)  1e-24   107/132(81%)   0/132(0%)      Plus/Minus
Query  3        GTATCCAACAACGGCAAGCATGGAGATACGTCCGTGCTTGAGCTCGACAAATCGGAGGCG  62
                ||||||||||| ||||||||||| |||||||| ||||||| ||| ||  | ||||| ||
Sbjct  2367238  GTATCCAACAACAGCAAGCATGGAAATACGTCCATGCTTGATCTCAACGTAACGGAGACG  2367179

Query  63       ATCGAAGGTTTCTTGATCTCCATCGGCAACCAAACCAAGAGGATCGAAGAATCCAAGAGG  122
                 |||||  | || || || || ||||| || |  || || |||||||||||||||| ||
Sbjct  2367178  GTCGAACTTCTCCTGGTCACCGTCGGCGACAAGGCCGAGGGGATCGAAGAATCCAAGCGG  2367119

Query  123      AGGTTGAGCACC  134
                ||| ||||||||
Sbjct  2367118  AGGCTGAGCACC  2367107
```

InterPro search of `GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY` gives hit to IPR022796 / Pfam PF00504 = "Chlorophyll A-B binding protein" that covers all the sequence except first residue.

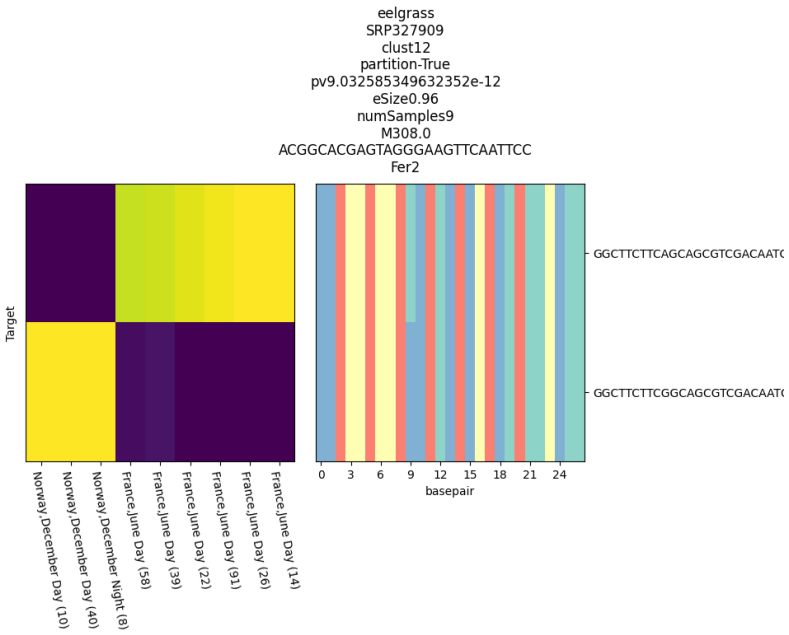
Top BLASTP hits (elided second hits are additional hits on species already shown). These and additonal top hits are **diatoms**.

```
protein fucoxanthin chlorophyll a/c protein [Phaeodactylum tricornutum CCAP 1055/1]
Sequence ID: XP_002184619.1   Length: 197    Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
89.7 bits(221)        1e-20   Compositional matrix adjust.  42/44(95%)     43/44(97%)     0/44(0%)
Query  1   GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY  44
           GAQPPLGFFDPLGLVADGDQE FDRLR+VELKHGRISMLAVVGY
Sbjct  38  GAQPPLGFFDPLGLVADGDQEKFDRLRYVELKHGRISMLAVVGY  81
...
hypothetical protein FisN_2Hh208 [Fistulifera solaris]
Sequence ID: GAX26374.1       Length: 169    Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
88.6 bits(218)        2e-20   Compositional matrix adjust.  42/44(95%)     42/44(95%)     0/44(0%)
Query  1   GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY  44
           GAQPPLGFFDPLGLV DGDQE FDRLRFVELKHGRISMLAVVGY
Sbjct  36  GAQPPLGFFDPLGLVTDGDQEKFDRLRFVELKHGRISMLAVVGY  79
...
Chain 19, FCP-F [Chaetoceros gracilis]
Sequence ID: 6JLU_19  Length: 166    Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
88.2 bits(217)        3e-20   Compositional matrix adjust.  41/44(93%)     43/44(97%)     0/44(0%)
Query  1   GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY  44
           GAQPPLGFFDPLGLVADGDQE FDRLR+VE+KHGRISMLAVVGY
...
```

truncated fucoxanthin chlorophyll a/c-binding protein precursor [Cylindrotheca fusiformis]

```
Sequence ID: AAN08829.1L      ength: 167     Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
86.3 bits(212)         1e-19   Compositional matrix adjust.  40/44(91%)     41/44(93%)     0/44(0%)
Query  1   GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY  44
           GAQPPLGFFDPLGLVADGDQE FDRLR+VELKHGRI ML VVGY
Sbjct  36  GAQPPLGFFDPLGLVADGDQEKFDRLRYVELKHGRICMLGVVGY  79
```

Zostera marina *ferredoxin (domain Fer2)*



anchor in blue, target in red, extended consensus in black.

AT1 extended

ACGGCACGAGTAGGGAAGTTCAATTCC
GGCTTCTTCAGCAGCGTCGACAATGAA
GACGTCGTCGGCACACTCGATGGTTTCATCGATGCCTTCTTCCTCCGAGATGAGCTTGACAGAATATCCGAGAGAGGTCGGCG

protein translation in frame -3:

PTSLGYSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR

BLASTN of anchor-target-consensus gives no hits in the *Z. marina* genome.
BLASTN of anchor-target-consensus against nr/nt: no full-length hits.

InterPro search of PTSLGYSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR gives hit to IPR001041 / Pfam PF00111 = "2Fe-2S iron-sulfur cluster binding domain", matching EEGIDETIECADDVFIVDAAEEAGIELPYSCR.

Top four BLASTP hits -- all species are **diatoms**.

ferredoxin [Thalassiosira oceanica]

```
Sequence ID: EJK54785.1       Length: 125    Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
73.9 bits(180)         3e-15   Compositional matrix adjust.  35/44(80%)     39/44(88%)     0/44(0%)
Query  2   TSLGYSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR  45
           TSL YSVK+ +EEEGID T ECADDVFIVDAAEE G++LPYSCR
Sbjct  26  TSLDYSVKVFNEEEGIDATFECADDVFIVDAAEEEGVDLPYSCR  69
```

ferredoxin [Schizostauron trachyderma]

```
Sequence ID: UDP55462.1       Length: 99     Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
68.6 bits(166)         2e-13   Compositional matrix adjust.  30/40(75%)     36/40(90%)     0/40(0%)
Query  6   YSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR  45
           Y VKL+SEE+GID TI+C DDVF++DAAEE G+ELPYSCR
Sbjct  4   YKVKLLSEEQGIDTTIDCNDDVFVLDAAEEQGVELPYSCR  43
```

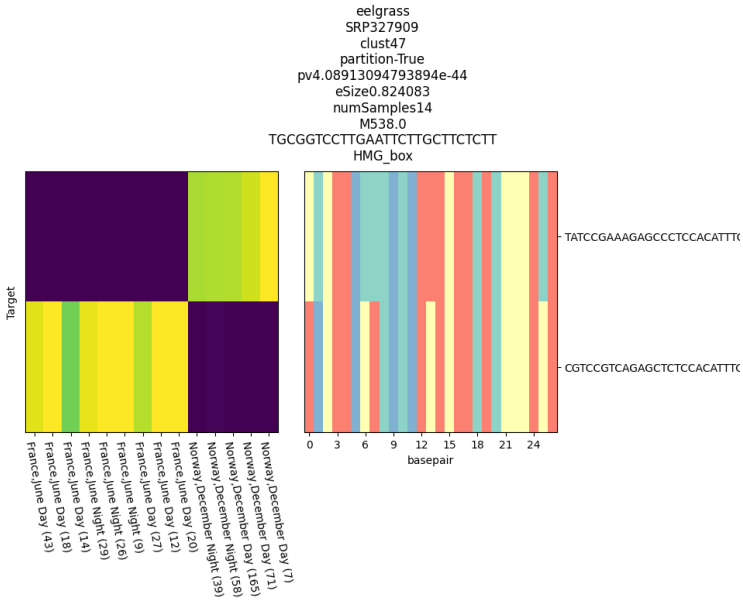
ferredoxin [Skeletonema grevillei]

```
Sequence ID: YP_010201387.1   Length: 99     Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
68.2 bits(165)         3e-13   Compositional matrix adjust.  30/42(71%)     38/42(90%)     0/42(0%)
Query  4   LGYSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR  45
           + Y+V LISEE GI+ TIEC+DDVF++DAAEE+GI+LPYSCR
Sbjct  2   VNYNVTLISEEHGINSTIECSDDVFVLDAAEESGIDLPYSCR  43
```

ferredoxin [Cerataulina daemon]

```
Sequence ID: YP_009093291.1   Length: 100    Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
68.2 bits(165)         3e-13   Compositional matrix adjust.  30/40(75%)     36/40(90%)     0/40(0%)
Query  6   YSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR  45
           Y VKL+S+E GID TI+C+DDVFI+DAAEE GI+LPYSCR
Sbjct  4   YKVKLVSDEHGIDTTIDCSDDVFILDAAEEQGIDLPYSCR  43
```

Zostera marina *HMG-box protein (domain HMG_box)*



anchor in blue, target in red, extended consensus in black.

AT1 extended

TGCGGTCCTTGAATTCTTGCTTCTCTT
TATCCGAAAGAGCCCTCCACATTTCAC
CAAGCTTCTTTCCTACACCACCAAAGGTCAAATCAGGATCGTCTTCCTTCAC

protein translation in frame -1:

VKEDDPDLTFGGVGKKLGEMWRALSDKEKQEFKDR

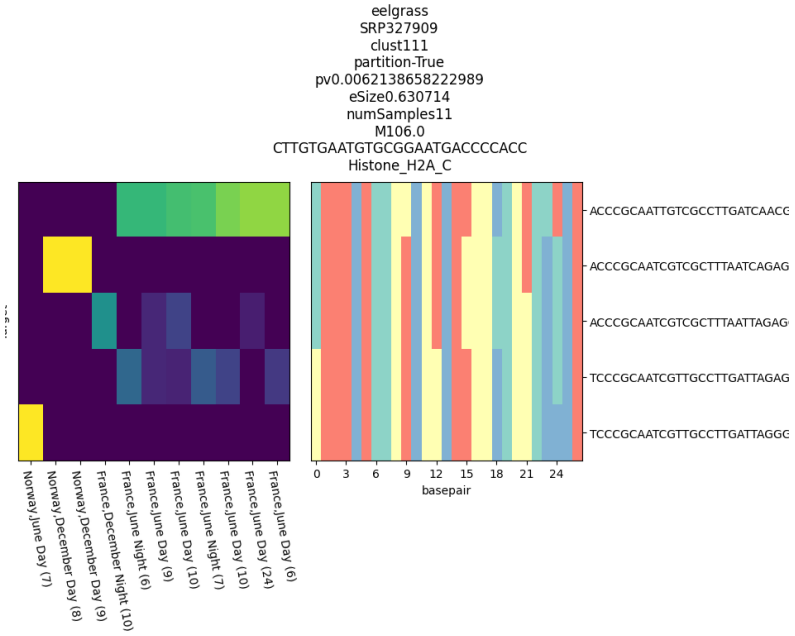
BLASTN of anchor-target-consensus gives no hits in the *Z. marina* genome.
BLASTN of anchor-target-consensus against nr/nt: no full-length hits.

InterPro search of VKEDDPDLTFGGVGKKLGEMWRALSDKEKQEFKDR gives hit to IPR009071 / Pfam PF00505 = "HMG (high mobility group) box", covering entire sequence except last residue.

Top two BLASTP hits -- both species are **diatoms**. (Third-ranking hit is a fungus.)

```
group protein B3 [Seminavis robusta]
Sequence ID: CAB9513894.1     Length: 75     Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
56.2 bits(134)         6e-09   Compositional matrix adjust.  24/35(69%)     33/35(94%)     0/35(0%)
Query  1   VKEDDPDLTFGGVGKKLGEMWRALSDKEKQEFKDR  35
           VKE++P++TFG +GKKLGEMWRAL+D+E++EFK R
Sbjct  39  VKEENPEITFGQMGKKLGEMWRALTDEEREEFKKR  73

HMG high mobility group box-containing protein [Nitzschia inconspicua]
Sequence ID: KAG7340638.1     Length: 79     Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
53.9 bits(128)         6e-08   Compositional matrix adjust.  24/34(71%)     29/34(85%)     0/34(0%)
Query  1   VKEDDPDLTFGGVGKKLGEMWRALSDKEKQEFKD  34
           +KE+ PDLTFGGVGKKLGEMWRAL +K K+ +K
Sbjct  40  LKEEHPDLTFGGVGKKLGEMWRALDEKTKENYKS  73
```

Zostera marina *histone H2A (domain Histone_H2A_C)*



anchor in blue, target in red, extended consensus in black.

AT1 extended

CTTGTGAATGTGCGGAATGACCCCACC

ACCCGCAATTGTCGCCTTGATCAACGC

ATCTAATTCTTCGTCTCCACGAATCGCCAATTGTAAATGACGCGGTGTAATTCGTTTGACCTTGAGA

protein translation in frame -2:

LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK

BLASTN of anchor-target-consensus gives no hits in the *Z. marina* genome.

BLASTN of anchor-target-consensus against nr/nt: top hit is a **diatom**.

```
Phaeodactylum tricornutum isolate CCAP 1055/1 genome assembly, chromosome: 2
Sequence ID: OU594943.1       Length: 2499861        Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Identities     Gaps    Strand
138 bits(152)  2e-28   103/121(85%)   0/121(0%)      Plus/Minus
Query  1       CTTGTGAATGTGCGGAATGACCCCACCACCCGCAATTGTCGCCTTGATCAACGCATCTAA  60
               ||||||||||||| ||||| || ||||| ||||||||||| |||||||| | ||| || ||
Sbjct  607034  CTTGTGAATGTGGGGAATCACACCACCTCCCGCAATTGTGGCCTTGATGAGCGCGTCCAA  606975

Query  61      TTCTTCGTCTCCACGAATCGCCAATTGTAAATGACGCGGTGTAATTCGTTTGACCTTGAG  120
```

```
                 ||| ||||| |||||||| |||||||| ||||| || |||||||| || || ||||||||
Sbjct  606974  TTCCTCGTCACCACGAATGGCCAATTGCAAATGCCGGGGTGTAATACGCTTCACCTTGAG  606915

Query  121     A  121
               |
Sbjct  606914  A  606914
```

InterPro search of `LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK` gives hit to IPR032454 / Pfam PF16211 = "C-terminus of histone H2A", covering `EELDALIKATIAGGGVIPHIHK`. InterPro also gives a longer match of the entire sequence, to Panther PTHR23430 = "HISTONE H2A".

BLASTP of `LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK` gives a large number of 100% identical hits, over a wide range species. When restricted to an arbitrary **diatom** (*Phaeodactylum tricornutum*), there is also a perfect match. By comparison, restriction to *Zostera marina* gives a strong but non-perfect match.

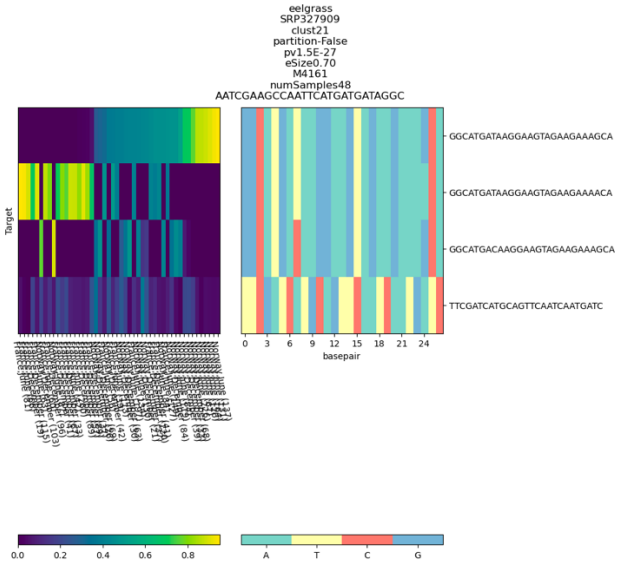
histone H2A isoform 1 [Phaeodactylum tricornutum CCAP 1055/1]

```
Sequence ID: XP_002181345.1Length: 138Number of Matches: 1
    See 1 more title(s) See all Identical Proteins(IPG)
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
82.0 bits(201)         7e-23   Compositional matrix adjust.  40/40(100%)    40/40(100%)    0/40(0%)
Query  1    LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK  40
            LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK
Sbjct  84   LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK  123
```

Histone H2A [Zostera marina]

```
Sequence ID: KMZ73852.1       Length: 134    Number of Matches: 1
Alignment statistics for match #1 Score      Expect  Method  Identities     Positives      Gaps
79.3 bits(194)         1e-21   Compositional matrix adjust.  38/40(95%)     38/40(95%)     0/40(0%)
Query  1    LKVKRITPRHLQLAIRGDEELDALIKATIAGGGVIPHIHK  40
            LKVKRITPRHLQLAIRGDEELD LIK TIAGGGVIPHIHK
Sbjct  86   LKVKRITPRHLQLAIRGDEELDTLIKGTIAGGGVIPHIHK  125
```

Zostera marina *NdhL (L subunit of NADPH dehydrogenase complex)*




```
>anchor
AATCGAAGCCAATTCATGATGATAGGC
>target1
GGCATGATAAGGAAGTAGAAGAAAGCA
>target2
GGCATGATAAGGAAGTAGAAGAAAACA
>target3
```

```
GGCATGACAAGGAAGTAGAAGAAAGCA
>target4
TTCGATCATGCAGTTCAATCAATGATC
```

BLAST against Z. marina genome in NCBI, all hit in the same 466,922 bp scaffold_137, all are on minus strand.

```
Zostera marina strain Finnish scaffold_137, whole genome shotgun sequence
Sequence ID: LFYR01000468.1   Length: 466922
```

anchor:

```
Query  1       AATCGAAGCCAATTCATGATGATAGGC  27
               ||||||||||||||||||||||||||
Sbjct  167261  AATCGAAGCCAATTCATGATGATAGGC  167235
```

target1: (queried with anchor+target1)

```
Query  27      CGGCATGATAAGGAAGTAGAAGAAAGCA  54
               ||||||||||||||||||||||||||||
Sbjct  167136  CGGCATGATAAGGAAGTAGAAGAAAGCA  167109
```

target4: (queried with anchor+target4)

```
Query  1       AATCGAAGCCAATTCATGATGATAGGCTTCGATCATGCAGTTCAATCAATGATC  54
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  167261  AATCGAAGCCAATTCATGATGATAGGCTTCGATCATGCAGTTCAATCAATGATC  167208
```

the longer consensus for target4 (called ">3") from reads:

```
AATCGAAGCCAATTCATGATGATAGGCTTCGATCATGCAGTTCAATCAATGATC
AATCGAAGCCAATTCATGATGATAGGCTTCGATCATGCAGTTCAATCAATGATCAATCAGTACTGTTCGCAGTTGTAAAG
```

Looking at genomic context (plus-strand is shown, so above sequences are now reverse-complements), we see that **target1+anchor** is a splice junction, whereas **target4+anchor** is most likely intron retention. Translations, including upstream and downstream, match the annotated protein (see below).
Note that target4+anchor has **multiple stop codons** in this reading frame (and in fact, has stops in all reading frames). So if it is translated, it would result in a truncated protein.

```
>LFYR01000468.1:167000-168000 Zostera marina strain Finnish scaffold_137, whole genome
shotgun sequence
GTGATATTATTAGTATTTTTTAGATTGCGGATGAACCAGCGTTTGCTGTGACCGGAAGCAATAACTATGA

GCAAGATTTGACTTCAGTGCTTATACAATCAGGAGCATTTGCTTTCTTCTACTTCCTTATCATGCCGgta
                          ...SerGlyAlaPheAlaPhePheTyrPheLeuIleMetProVal

tataattgcaaagtgatacttacataataattatttcattgactttacaactgcgaacagtactgattga
TyrAsnCysLysValIleLeuThr******LeuPheHis***LeuTyrAsnCysGluGlnTyr***LeuI

tcattgattgaactgcatgatcgaagCCTATCATCATGAATTGGCTTCGATTGAGATGGTACAAGCGCAA
leIleAsp***ThrAla***SerLysProIleIleMetAsnTrpLeuArgLeuArgTrpTyr...

ATTATTCGAGACGTATTTACAGTTCATGTTTGTATTCCTCTTTTTTCCTGGGTAAGCGAAGTAGTTATAA

GTACATCAACATAGATCGAAGGAAATATATCATTTGAATATTTGAGAATTTATTCCAGGATTTTATTGTG
GGCGCCATTCATCAACTTTAGGAGACTCCCCAGAGATCCGACGATGAAGCACCCTTGGTCTACCCCTCGA
GATTCATCGACTTAATTAATTACTACATTGTATTTAATTTAATTGTGAATTTAATTATATATCAAATCAT
ATATGTTACAAAAAAAAATAGCAAAAAATTTAAGCTTATTAAGAAGAATGAAAGCCCGAAGAGGAGTTTT
CTTTAGGTCAACATCAGAAGTTACAAAATGAAGCGTGGGCTGAAAATGAGACGCCTGCCCCACCTGAGCT
GGAAAACTTCTCACTCTCTAGGAAAAGAGACGATCGACAGATAAGGAAACATTCACCATCGGTGGTACAG
TTATCGGTGCATGACACCGAAGAAGTTCTCCATATGAGAGGTTTGTCTAACCGATCCCGCCGCTGGAGAG
GGTTGCACTAGCTGCTGAGACTGACGGAGATTAAAAGGGGTATCTAAAATAACTGAAGGGCAAGAAGTAA
GGACCAAAACCTTAAAAAAGAAATCAAAAGAAAGTCAAGGTACGTACCATCATCAATGCTGCGTGCATTA
CATACTGCGAGGGGAATAAGT
```

This region of the *Z. marina* genome is annotated with a protein feature. Oddly, there is no corresponding transcript annotated.
protein accession = KMZ74005.1 name = "hypothetical protein ZOSMA_137G00200 [Zostera marina]"

```
>KMZ74005.1 hypothetical protein ZOSMA_137G00200 [Zostera marina]
```

MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLH
LGVLLTSIADEPAFAVTGSNNYEQDLTSVLIQSGAFAFFYFLIMPPIIMNWLRLRWYKRKLFETYLQFMF
VFLFFPGILLWAPFINFRRLPRDPTMKHPWSTPRDSST

Within the protein entry, it gives the nucleotide coordinates that construct this coding sequence:

```
CDS          1..178
             /locus_tag="ZOSMA_137G00200"
             /coded_by="join(LFYR01000468.1:166629..166775,
             LFYR01000468.1:166859..166942,
             LFYR01000468.1:167023..167136,
             LFYR01000468.1:167236..167330,
             LFYR01000468.1:167408..167504)"
```

protein sequence of the intron-retention variant found by NOMAD:

```
>intron-retention
MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLH
LGVLLTSIADEPAFAVTGSNNYEQDLTSVLIQSGAFAFFYFLIMPVYNCKVILT
```

InterPro search on the entire protein finds IPR019654 / Pfam PF10716 domain = "NAD(P)H-quinone oxidoreductase subunit L", as well as longer PANTHER domain PTHR36727, "NAD(P)H-QUINONE OXIDOREDUCTASE SUBUNIT L, CHLOROPLASTIC" and also predicts transmembrane regions (based on TMHMM). Below is a schematic, transmembrane regions in **bold-red**, Pfam NdhL domain underlined, PANTHER NdhL domain in blue-background.

```
>KMZ74005.1 hypothetical protein ZOSMA_137G00200 [Zostera marina]
MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLHLGVLLTSIADEPAFAVTGS
NNYEQDLTSVLIQSGAFAFFYFLIMPPIIMNWLRLRWYKRKLFETYLQFMFVFLFFPGILLWAPFINFRRLPRDPTMKHPWSTPRDSST
```

For the intron-retention variant, InterPro only finds the PANTHER NDhL domain. It has only one transmembrane domain.

```
>intron-retention
MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLHLGVLLTSIADEPAFAVTGS
NNYEQDLTSVLIQSGAFAFFYFLIMPVYNCKVILT
```

Supplementary Tables

Supplementary Tables 1 : Protein domain analysis
For SARS-CoV-2 datasets, we use significant NOMAD anchors meeting the effect size requirement of >.5 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.

Supplementary Tables 2: Significant anchors
Tables containing significant anchors, anchor statistics, and C_j used for each sample.

Supplementary Tables 3 : 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.

Supplementary Tables 4: Anchor genome annotations
Tables containing significant anchors, and their genome and transcriptome annotations.

Supplementary Tables 5: BLAST results
Tables containing BLAST results for unannotated anchors.

Supplementary Tables 6: Octopus results
Tables containing results for Octopus called anchors.

Supplementary Tables 7: Eelgrass results
Tables containing results for Eelgrass called anchors.

Bibliography

Abante, J., Wang, P.L. and Salzman, J. (2022) “DIVE: a reference-free statistical approach to diversity-generating and mobile genetic element discovery,” *BioRxiv* [Preprint]. doi:10.1101/2022.06.13.495703.

Agresti, A. (1992) “A Survey of Exact Inference for Contingency Tables,” *Statistical Science*, 7(1), pp. 131–153. doi:10.1214/ss/1177011454.

Albertin, C.B. *et al.* (2015) “The octopus genome and the evolution of cephalopod neural and morphological novelties.,” *Nature*, 524(7564), pp. 220–224. doi:10.1038/nature14668.

Baharav, T.Z., Tse, D. and Salzman, J. (2023) “An interpretable, finite sample valid alternative to Pearson’s X2 for scientific discovery.” In preparation.

Bal, A. *et al.* (2022) “Detection and prevalence of SARS-CoV-2 co-infections during the Omicron variant circulation, France, December 2021 - February 2022,” *medRxiv* [Preprint]. doi:10.1101/2022.03.24.22272871.

Benjamini, Y. and Yekutieli, D. (2001) “The control of the false discovery rate in multiple testing under dependency,” *The Annals of Statistics*, 29(4), pp. 1165–1188. doi:10.1214/aos/1013699998.

Bi, D. *et al.* (2012) “ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria.,” *Nucleic Acids Research*, 40(Database issue), pp. D621-6. doi:10.1093/nar/gkr846.

Bray, N.L. *et al.* (2016) “Near-optimal probabilistic RNA-seq quantification.,” *Nature Biotechnology*, 34(5), pp. 525–527. doi:10.1038/nbt.3519.

Briney, B. *et al.* (2019) “Commonality despite exceptional diversity in the baseline human antibody repertoire.,” *Nature*, 566(7744), pp. 393–397. doi:10.1038/s41586-019-0879-y.

Buen Abad Najar, C.F., Yosef, N. and Lareau, L.F. (2019) “Coverage-dependent bias creates the appearance of binary splicing in single cells,” *BioRxiv* [Preprint]. doi:10.1101/2019.12.19.883256.

Candes, E.J. and Wakin, M.B. (2008) “An Introduction To Compressive Sampling,” *IEEE signal processing magazine*, 25(2), pp. 21–30. doi:10.1109/MSP.2007.914731.

Canzar, S. *et al.* (2017) “BASIC: BCR assembly from single cells.,” *Bioinformatics*, 33(3), pp. 425–427. doi:10.1093/bioinformatics/btw631.

Cao, Y. *et al.* (2020) “Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients’ B Cells.,” *Cell*, 182(1), pp. 73-84.e16. doi:10.1016/j.cell.2020.05.025.

Chen, S. *et al.* (2018) “fastp: an ultra-fast all-in-one FASTQ preprocessor.,” *Bioinformatics*, 34(17), pp. i884–i890. doi:10.1093/bioinformatics/bty560.

Chen, Y. *et al.* (2005) “Sequential monte carlo methods for statistical analysis of tables,” *Journal of the American Statistical Association*, 100(469), pp. 109–120. doi:10.1198/016214504000001303.

Chung, E. and Romano, J.P. (2013) “Exact and asymptotically robust permutation tests,” *The Annals of Statistics*, 41(2), pp. 484–507. doi:10.1214/13-AOS1090.

Couvin, D. *et al.* (2018) “CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins.,” *Nucleic Acids Research*, 46(W1), pp. W246–W251. doi:10.1093/nar/gky425.

Dehghannasiri, R. *et al.* (2022) “Unsupervised reference-free inference reveals unrecognized regulated transcriptomic complexity in human single cells,” *BioRxiv* [Preprint]. doi:10.1101/2022.12.06.519414.

Diaconis, P. and Sturmfels, B. (1998) “Algebraic algorithms for sampling from conditional distributions,” *The Annals of Statistics*, 26(1). doi:10.1214/aos/1030563990.

Di Tommaso, P. *et al.* (2017) “Nextflow enables reproducible computational workflows.,” *Nature Biotechnology*, 35(4), pp. 316–319. doi:10.1038/nbt.3820.

Dobin, A. *et al.* (2013) “STAR: ultrafast universal RNA-seq aligner.,” *Bioinformatics*, 29(1), pp. 15–21. doi:10.1093/bioinformatics/bts635.

Donnelly, P. and Tavaré, S. (1995) “Coalescents and genealogical structure under neutrality.,” *Annual Review of Genetics*, 29, pp. 401–421. doi:10.1146/annurev.ge.29.120195.002153.

Edgar, R.C. (2016) “UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing,” *BioRxiv* [Preprint]. doi:10.1101/081257.

Edgar, R.C. *et al.* (2022) “Petabase-scale sequence alignment catalyses viral discovery.,” *Nature*, 602(7895), pp. 142–147. doi:10.1038/s41586-021-04332-2.

Elahi, S. *et al.* (2011) “Protective HIV-specific CD8+ T cells evade Treg cell suppression.,” *Nature Medicine*, 17(8), pp. 989–995. doi:10.1038/nm.2422.

Evans, D.R. *et al.* (2020) “Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital.,” *eLife*, 9. doi:10.7554/eLife.53886.

Ewels, P.A. *et al.* (2020) “The nf-core framework for community-curated bioinformatics pipelines.,” *Nature Biotechnology*, 38(3), pp. 276–278. doi:10.1038/s41587-020-0439-x.

Ezran, C. *et al.* (2017) “The mouse lemur, a genetic model organism for primate biology, behavior, and health.,” *Genetics*, 206(2), pp. 651–664.

doi:10.1534/genetics.116.199448.

Fisher, R.A. (1922) “On the Interpretation of X2 from Contingency Tables, and the Calculation of P,” *Journal of the Royal Statistical Society*, 85(1), p. 87. doi:10.2307/2340521.

Francis, J.M. *et al.* (2022) “Allelic variation in class I HLA determines CD8+ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2.,” *Science Immunology*, 7(67), p. eabk3070. doi:10.1126/sciimmunol.abk3070.

Gee, M.H. *et al.* (2018) “Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes.,” *Cell*, 172(3), pp. 549-563.e16. doi:10.1016/j.cell.2017.11.043.

van Giesen, L. *et al.* (2020) “Molecular basis of chemotactile sensation in octopus.,” *Cell*, 183(3), pp. 594-604.e14. doi:10.1016/j.cell.2020.09.008.

Gorzynski, J.E. *et al.* (2020) “High-throughput SARS-CoV-2 and host genome sequencing from single nasopharyngeal swabs.,” *medRxiv* [Preprint]. doi:10.1101/2020.07.27.20163147.

Grant, R.A. *et al.* (2021) “Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia.,” *Nature*, 590(7847), pp. 635–641. doi:10.1038/s41586-020-03148-w.

Gratia, M. *et al.* (2015) “Rotavirus NSP3 Is a Translational Surrogate of the Poly(A) Binding Protein-Poly(A) Complex.,” *Journal of Virology*, 89(17), pp. 8773–8782. doi:10.1128/JVI.01402-15.

Groeger, A.L. *et al.* (2010) “Cyclooxygenase-2 generates anti-inflammatory mediators from omega-3 fatty acids.,” *Nature Chemical Biology*, 6(6), pp. 433–441. doi:10.1038/nchembio.367.

Hayashizaki, K. *et al.* (2016) “Myosin light chains 9 and 12 are functional ligands for CD69 that regulate airway inflammation.,” *Science Immunology*, 1(3), p. eaaf9154. doi:10.1126/sciimmunol.aaf9154.

Jacobs, R.P.W.M. and Noten, T.M.P.A. (1980) “The annual pattern of the diatoms in the epiphyton of eelgrass (Zostera marina L.) at Roscoff, France,” *Aquatic Botany*, 8, pp. 355–370. doi:10.1016/0304-3770(80)90065-0.

Jacot, D. *et al.* (2021) “Assessment of SARS-CoV-2 Genome Sequencing: Quality Criteria and Low-Frequency Variants.,” *Journal of Clinical Microbiology*, 59(10), p. e0094421. doi:10.1128/JCM.00944-21.

Johnson, L.S., Eddy, S.R. and Portugaly, E. (2010) “Hidden Markov model speed heuristic and iterative HMM search procedure.,” *BMC Bioinformatics*, 11, p. 431. doi:10.1186/1471-2105-11-431.

Jörrißen, P. *et al.* (2021) “Antibody Response to SARS-CoV-2 Membrane Protein in Patients of the Acute and Convalescent Phase of COVID-19.,” *Frontiers in Immunology*, 12, p. 679841. doi:10.3389/fimmu.2021.679841.

Jueterbock, A. *et al.* (2021) “Adaptation of temperate seagrass to arctic light relies on seasonal acclimatization of carbon capture and metabolism.,” *Frontiers in plant science*, 12, p. 745855. doi:10.3389/fpls.2021.745855.

Kalvari, I. *et al.* (2021) “Rfam 14: expanded coverage of metagenomic, viral and microRNA families.,” *Nucleic Acids Research*, 49(D1), pp. D192–D200. doi:10.1093/nar/gkaa1047.

Kiepiela, P. *et al.* (2004) “Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA.,” *Nature*, 432(7018), pp. 769–775. doi:10.1038/nature03113.

Kim, D. *et al.* (2020) “The Architecture of SARS-CoV-2 Transcriptome.,” *Cell*, 181(4), pp. 914-921.e10. doi:10.1016/j.cell.2020.04.011.

Kirkegaard, K., van Buuren, N.J. and Mateo, R. (2016) “My Cousin, My Enemy: quasispecies suppression of drug resistance.,” *Current opinion in virology*, 20, pp. 106–111. doi:10.1016/j.coviro.2016.09.011.

Kuo, S.-M. *et al.* (2017) “Inhibition of Avian Influenza A Virus Replication in Human Cells by Host Restriction Factor TUFM Is Correlated with Autophagy.,” *mBio*, 8(3). doi:10.1128/mBio.00481-17.

Langmead, B. *et al.* (2009) “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.,” *Genome Biology*, 10(3), p. R25. doi:10.1186/gb-2009-10-3-r25.

Laughlin, T.G., Savage, D.F. and Davies, K.M. (2020) “Recent advances on the structure and function of NDH-1: The complex I of oxygenic photosynthesis.,” *Biochimica et biophysica acta. Bioenergetics*, 1861(11), p. 148254. doi:10.1016/j.bbabio.2020.148254.

Leplae, R. *et al.* (2004) “ACLAME: a CLAssification of Mobile genetic Elements.,” *Nucleic Acids Research*, 32(Database issue), pp. D45-9. doi:10.1093/nar/gkh084.

Lindeman, I. *et al.* (2018) “BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq.,” *Nature Methods*, 15(8), pp. 563–565. doi:10.1038/s41592-018-0082-3.

Ma, M. *et al.* (2021) “The significance of chloroplast NAD(P)H dehydrogenase complex and its dependent cyclic electron transport in photosynthesis.,” *Frontiers in plant science*, 12, p. 661863. doi:10.3389/fpls.2021.661863.

Ma, X. *et al.* (2021) “Improved chromosome-level genome assembly and annotation of

the seagrass, Zostera marina (eelgrass).," *F1000Research*, 10, p. 289. doi:10.12688/f1000research.38156.1.

Magoč, T. and Salzberg, S.L. (2011) "FLASH: fast length adjustment of short reads to improve genome assemblies.," *Bioinformatics*, 27(21), pp. 2957–2963. doi:10.1093/bioinformatics/btr507.

Matzaraki, V. *et al.* (2017) "The MHC locus and genetic susceptibility to autoimmune and infectious diseases.," *Genome Biology*, 18(1), p. 76. doi:10.1186/s13059-017-1207-1.

Medhekar, B. and Miller, J.F. (2007) "Diversity-generating retroelements.," *Current Opinion in Microbiology*, 10(4), pp. 388–395. doi:10.1016/j.mib.2007.06.004.

Michael, T.P. and VanBuren, R. (2020) "Building near-complete plant genomes.," *Current Opinion in Plant Biology*, 54, pp. 26–33. doi:10.1016/j.pbi.2019.12.009.

Mistry, J. *et al.* (2021) "Pfam: The protein families database in 2021.," *Nucleic Acids Research*, 49(D1), pp. D412–D419. doi:10.1093/nar/gkaa913.

Motahari, A. *et al.* (2013) "Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads," in *2013 IEEE International Symposium on Information Theory*. *2013 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 1640–1644. doi:10.1109/ISIT.2013.6620505.

Nurk, S. *et al.* (2022) "The complete sequence of a human genome.," *Science*, 376(6588), pp. 44–53. doi:10.1126/science.abj6987.

Olivieri, J.E. *et al.* (2021) "RNA splicing programs define tissue compartments and cell types at single-cell resolution.," *eLife*, 10. doi:10.7554/eLife.70692.

Pascarella, G. *et al.* (2022) "Recombination of repeat elements generates somatic complexity in human genomes.," *Cell*, 185(16), pp. 3025-3040.e6. doi:10.1016/j.cell.2022.06.032.

Perrin, B.J. and Ervasti, J.M. (2010) "The actin gene family: function follows isoform.," *Cytoskeleton*, 67(10), pp. 630–634. doi:10.1002/cm.20475.

Poh, C.M. *et al.* (2020) "Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients.," *Nature Communications*, 11(1), p. 2806. doi:10.1038/s41467-020-16638-2.

Prazukin, A.V. *et al.* (2022) "Vertical distribution of epiphytic diatoms in relation to the eelgrass Zostera noltii canopy biomass and height," *Aquatic Botany*, 176, p. 103466. doi:10.1016/j.aquabot.2021.103466.

Rock, B.M. and Daru, B.H. (2021) "Impediments to understanding seagrasses' response to global change," *Frontiers in Marine Science*, 8. doi:10.3389/fmars.2021.608867.

Röhr, M.E. *et al.* (2018) “Blue carbon storage capacity of temperate eelgrass (*Zostera marina*) meadows,” *Global Biogeochemical Cycles* [Preprint]. doi:10.1029/2018GB005941.

Romano, Y., Sesia, M. and Candès, E. (2019) “Deep Knockoffs,” *Journal of the American Statistical Association*, pp. 1–27. doi:10.1080/01621459.2019.1660174.

Ross, K. *et al.* (2021) “Tncentral: a prokaryotic transposable element database and web portal for transposon analysis.,” *mBio*, 12(5), p. e0206021. doi:10.1128/mBio.02060-21.

Salzman, J., Jiang, H. and Wong, W.H. (2011) “Statistical Modeling of RNA-Seq Data.,” *Statistical Science*, 26(1). doi:10.1214/10-STS343.

Santamaria, M. *et al.* (2018) “ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences.,” *Nucleic Acids Research*, 46(D1), pp. D127–D132. doi:10.1093/nar/gkx855.

Selig, C. *et al.* (2008) “The ITS2 Database II: homology modelling RNA structure for molecular systematics.,” *Nucleic Acids Research*, 36(Database issue), pp. D377-80. doi:10.1093/nar/gkm827.

Shen, W. *et al.* (2016) “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.,” *Plos One*, 11(10), p. e0163962. doi:10.1371/journal.pone.0163962.

Sherman, R.M. *et al.* (2019) “Assembly of a pan-genome from deep sequencing of 910 humans of African descent.,” *Nature Genetics*, 51(1), pp. 30–35. doi:10.1038/s41588-018-0273-y.

Shi, Z.J. *et al.* (2022) “Fast and accurate metagenotyping of the human gut microbiome with GT-Pro.,” *Nature Biotechnology*, 40(4), pp. 507–516. doi:10.1038/s41587-021-01102-3.

Shrestha, R.P. and Hildebrand, M. (2015) “Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses.,” *Eukaryotic Cell*, 14(1), pp. 29–40. doi:10.1128/EC.00209-14.

Solé, M. *et al.* (2013) “Ultrastructural damage of Loligo vulgaris and Illex coindetii statocysts after low frequency sound exposure.,” *Plos One*, 8(10), p. e78825. doi:10.1371/journal.pone.0078825.

Song, Y. *et al.* (2020) “Reverse genetics reveals a role of rotavirus VP3 phosphodiesterase activity in inhibiting rnase L signaling and contributing to intestinal viral replication in vivo.,” *Journal of Virology*, 94(9). doi:10.1128/JVI.01952-19.

Ståhlberg, A. *et al.* (2016) “Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing.,” *Nucleic Acids Research*, 44(11), p. e105. doi:10.1093/nar/gkw224.

Storer, J. *et al.* (2021) “The Dfam community resource of transposable element families,

sequence models, and genome annotations.," *Mobile DNA*, 12(1), p. 2. doi:10.1186/s13100-020-00230-y.

Sun, X. and Whittaker, G.R. (2007) "Role of the actin cytoskeleton during influenza virus internalization into polarized epithelial cells.," *Cellular Microbiology*, 9(7), pp. 1672–1682. doi:10.1111/j.1462-5822.2007.00900.x.

Tabula Sapiens Consortium* *et al.* (2022) "The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.," *Science*, 376(6594), p. eabl4896. doi:10.1126/science.abl4896.

The Nucleic Acid Observatory Consortium (2021) "A Global Nucleic Acid Observatory for Biodefense and Planetary Health," *arXiv* [Preprint]. doi:10.48550/arxiv.2108.02678.

The Tabula Microcebus Consortium *et al.* (2021) "Tabula Microcebus: A transcriptomic cell atlas of mouse lemur, an emerging primate model organism," *BioRxiv* [Preprint]. doi:10.1101/2021.12.12.469460.

Thompson, M.G. *et al.* (2020) "Viral-induced alternative splicing of host genes promotes influenza replication.," *eLife*, 9. doi:10.7554/eLife.55500.

Tréguer, P. *et al.* (2018) "Influence of diatom diversity on the ocean biological carbon pump," *Nature Geoscience*, 11(1), pp. 27–37. doi:10.1038/s41561-017-0028-x.

Vedula, P. *et al.* (2017) "Diverse functions of homologous actin isoforms are defined by their nucleotide, rather than their amino acid sequence.," *eLife*, 6. doi:10.7554/eLife.31661.

Viana, R. *et al.* (2022) "Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa.," *Nature*, 603(7902), pp. 679–686. doi:10.1038/s41586-022-04411-y.

Wang, T. *et al.* (2022) "The Human Pangenome Project: a global resource to map genomic diversity.," *Nature*, 604(7906), pp. 437–446. doi:10.1038/s41586-022-04601-8.

West, K.M., Blacksher, E. and Burke, W. (2017) "Genomics, health disparities, and missed opportunities for the nation's research agenda.," *The Journal of the American Medical Association*, 317(18), pp. 1831–1832. doi:10.1001/jama.2017.3096.

Wilson, K.L. and Lotze, H.K. (2019) "Climate change projections reveal range shifts of eelgrass Zostera marina in the Northwest Atlantic," *Marine Ecology Progress Series*, 620, pp. 47–62. doi:10.3354/meps12973.

Wright, R.J., Comeau, A.M. and Langille, M.G.I. (2022) "From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools," *BioRxiv* [Preprint]. doi:10.1101/2022.04.27.489753.

Wu, C.-S. *et al.* (2022) "Chromosome-level genome assembly of grass carp (Ctenopharyngodon idella) provides insights into its genome evolution.," *BMC Genomics*, 23(1), p. 271. doi:10.1186/s12864-022-08503-x.

Wu, L. *et al.* (2011) “Structure of MyTH4-FERM domains in myosin VIIa tail bound to cargo.,” *Science*, 331(6018), pp. 757–760. doi:10.1126/science.1198848.

Yu, L. *et al.* (2020) “Somatic genetic drift and multilevel selection in a clonal seagrass.,” *Nature Ecology & Evolution*, 4(7), pp. 952–962. doi:10.1038/s41559-020-1196-4.

Yu, L. *et al.* (2022) “Ocean currents drive the worldwide colonization of the most widespread marine plant, eelgrass (*Zostera marina*),” *BioRxiv* [Preprint]. doi:10.1101/2022.12.10.519859.

Zayed, A.A. *et al.* (2022) “Cryptic and abundant marine viruses at the evolutionary origins of Earth’s RNA virome.,” *Science*, 376(6589), pp. 156–162. doi:10.1126/science.abm5847.

Zhang, C.-Z. *et al.* (2015) “Chromothripsis from DNA damage in micronuclei.,” *Nature*, 522(7555), pp. 179–184. doi:10.1038/nature14493.

Zhang, Y. *et al.* (2015) “Hearing characteristics of cephalopods: modeling and environmental impact study.,” *Integrative zoology*, 10(1), pp. 141–151. doi:10.1111/1749-4877.12104.

Zhao, C., Shi, Z.J. and Pollard, K.S. (2022) “Pitfalls of genotyping microbial communities with rapidly growing genome collections,” *BioRxiv* [Preprint]. doi:10.1101/2022.06.30.498336.