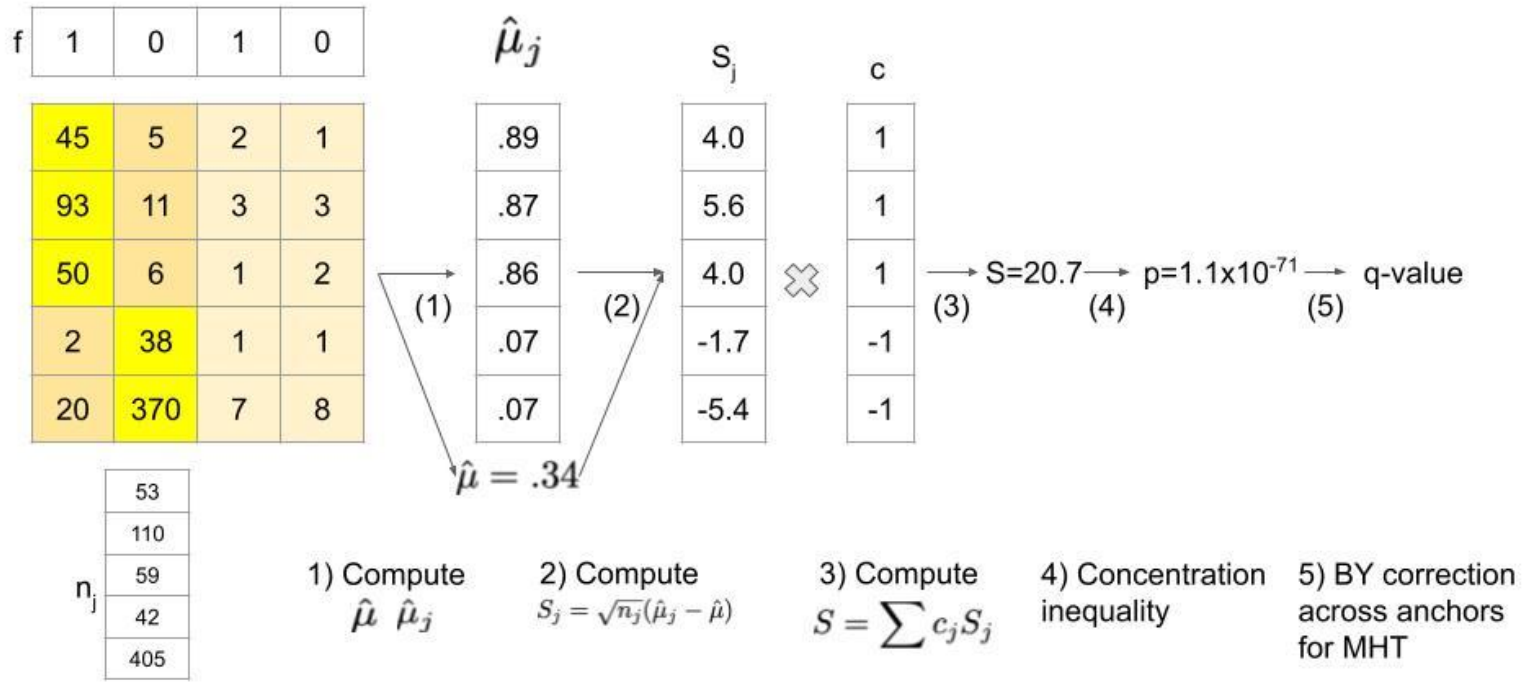
Supplementary Materials

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

A) P value computation



B) Effect size computation

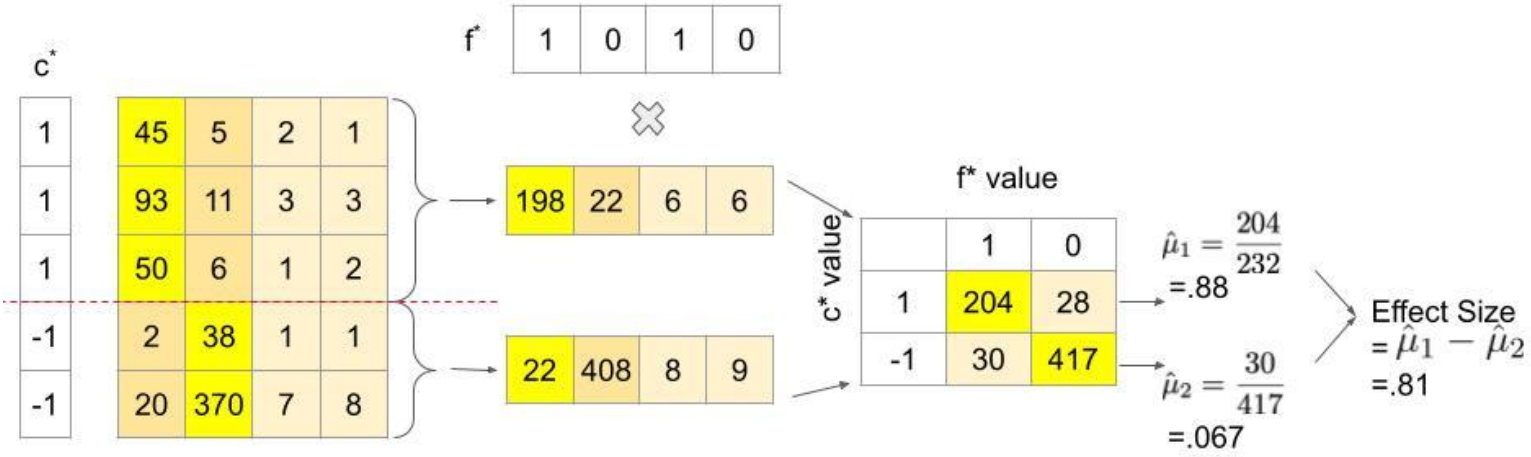
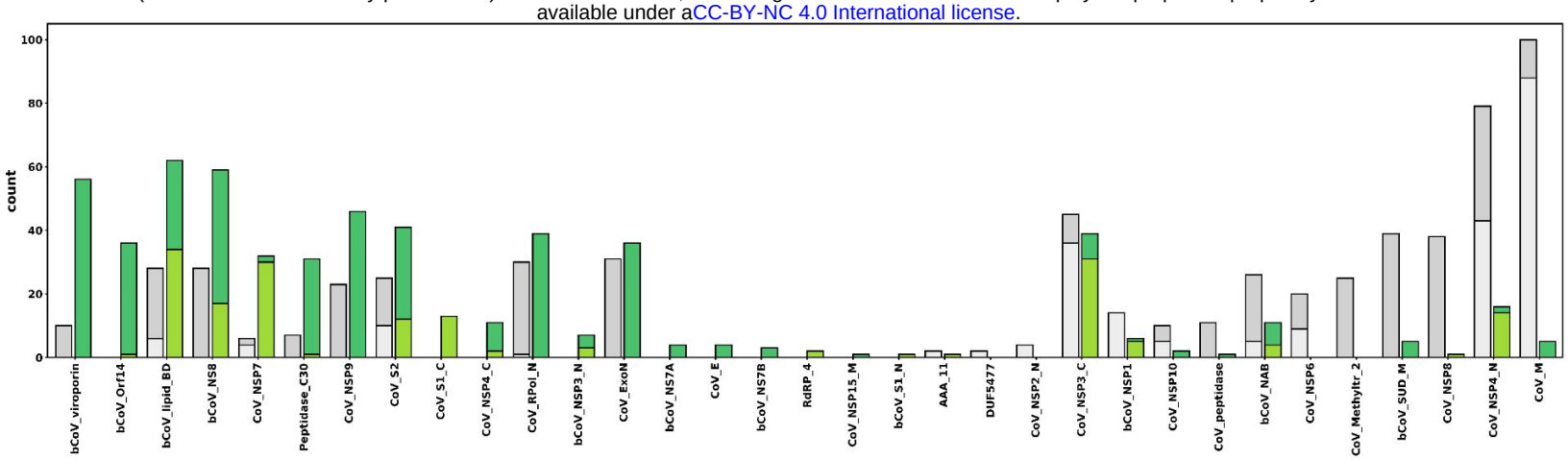


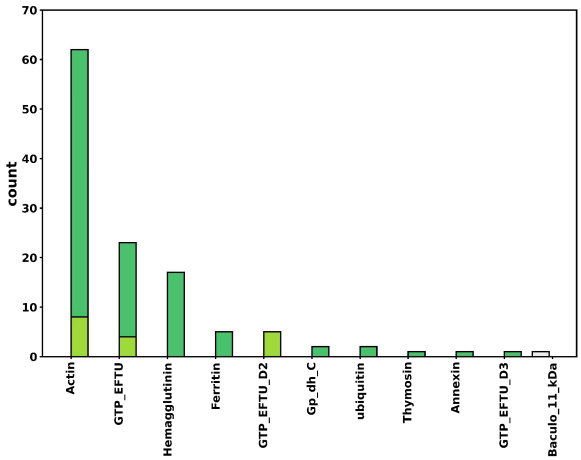
Figure S1

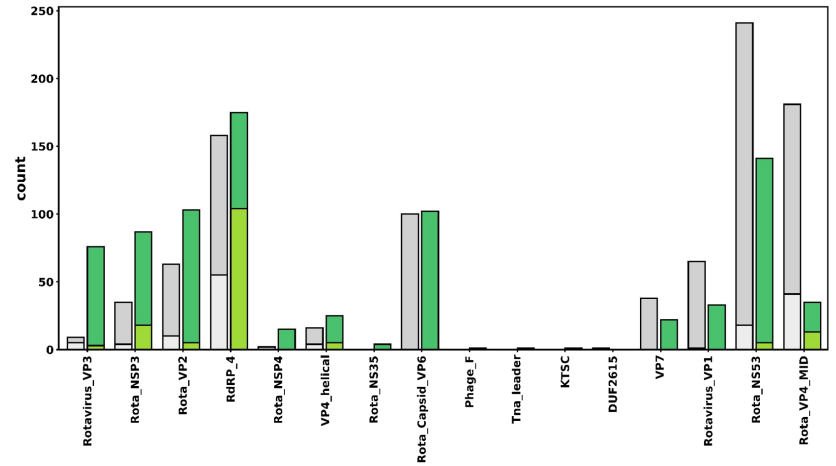
A. p-value computation for NOMAD. Contingency table transposed for visual convenience (rows are samples and columns are targets). Starting with a samples by targets counts matrix, NOMAD utilizes one (or several) functions f mapping targets to values within [0,1]. The mean with respect to f is taken over 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.

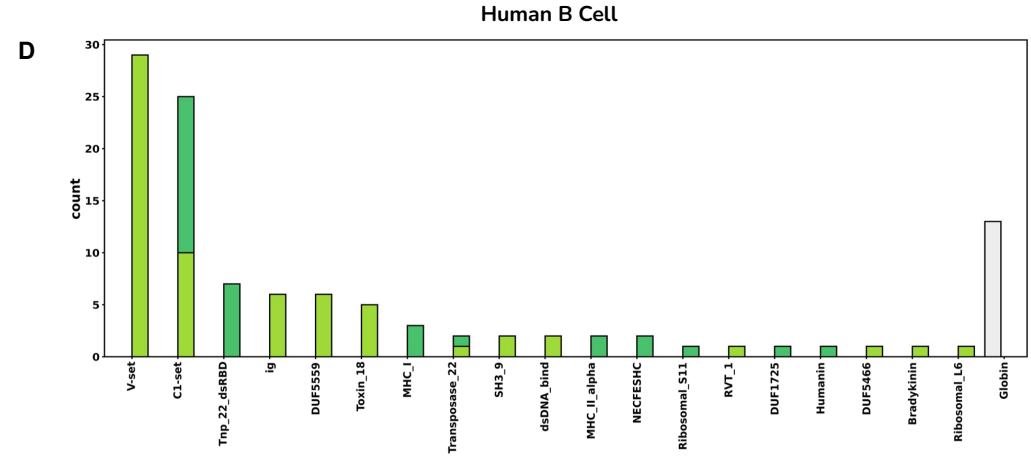
B. Effect size computation for NOMAD. Effect size is calculated based on the random split c and random function f that yielded the most significant NOMAD p-value. Fixing these, the effect size is computed as the difference between the mean across targets (with respect to f) across those samples with 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 $\mu_1 - \mu_2$.

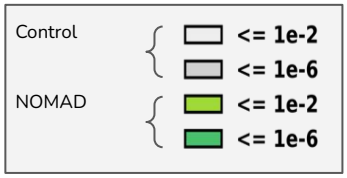
S2

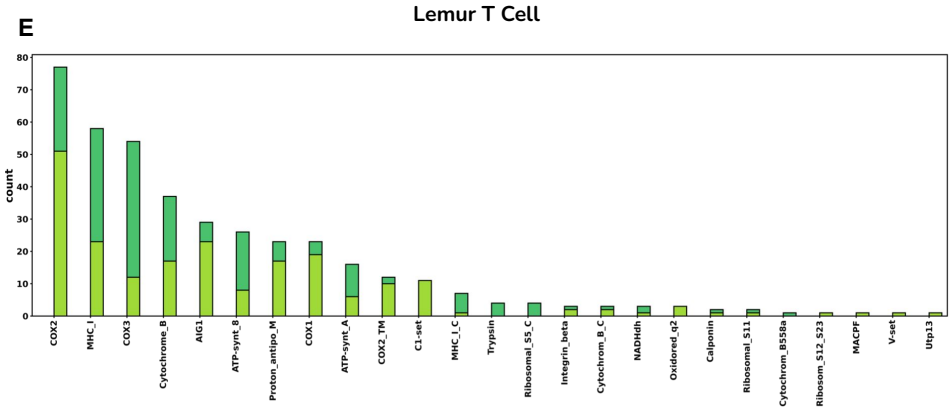












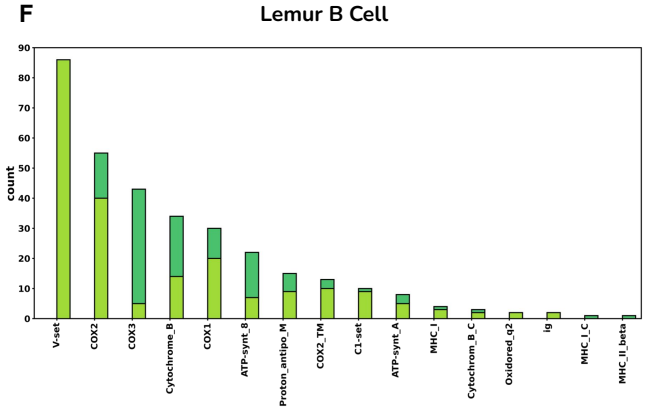


Figure S2

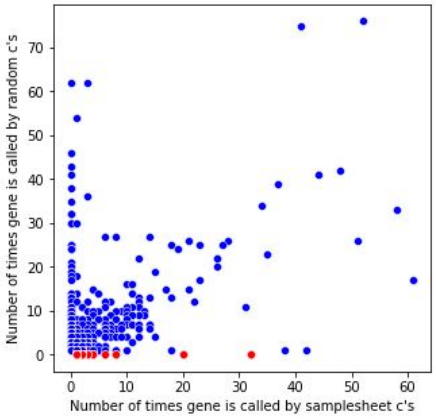
NOMAD protein profile hits to the Pfam database (greens) and control (greys); ordered by enrichment in NOMAD hits compared to control; all NOMAD anchors were used as input, without effect size filters.

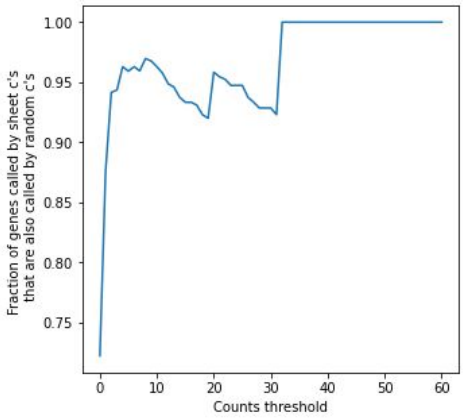
A. Protein profile analysis of NOMAD significant anchors from California data (SRR15881549), before viral strain divergence in the spike had been reported (Gorzynski *et al.*, 2020) serving as a negative control.
B. Protein profile analysis of NOMAD significant anchors from influenza-A data (SRP294571).
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

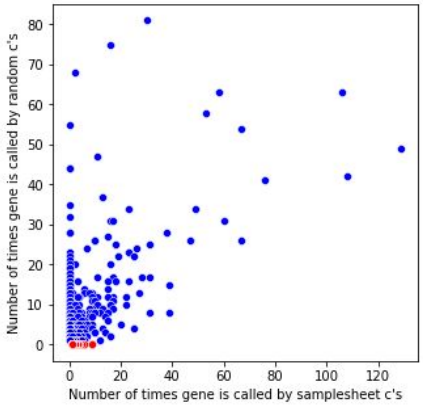
A

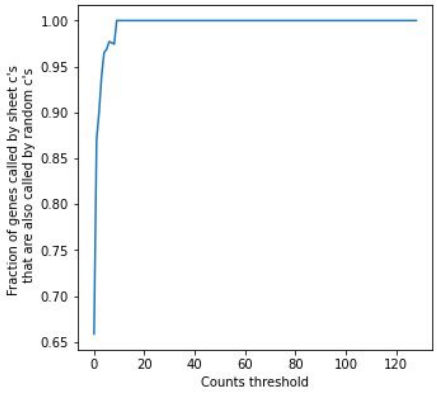
Donor 1



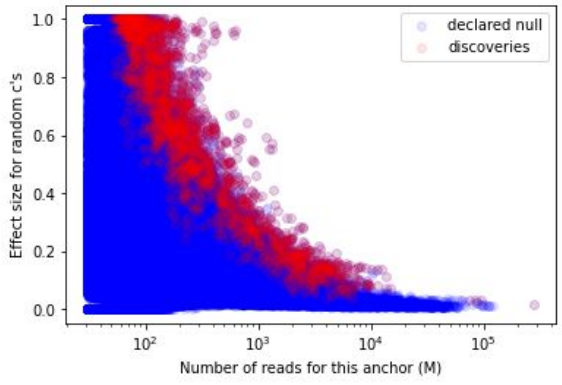


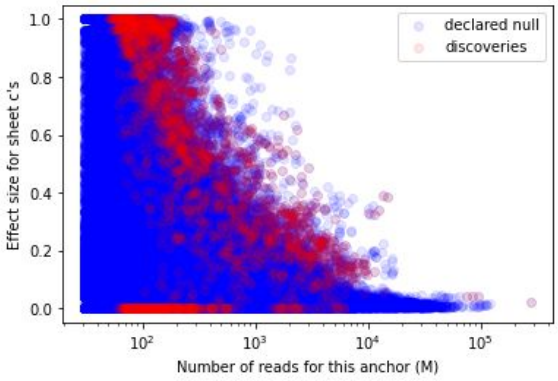
Donor 2

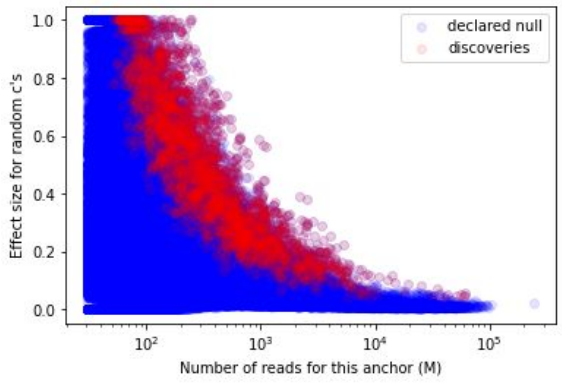




B







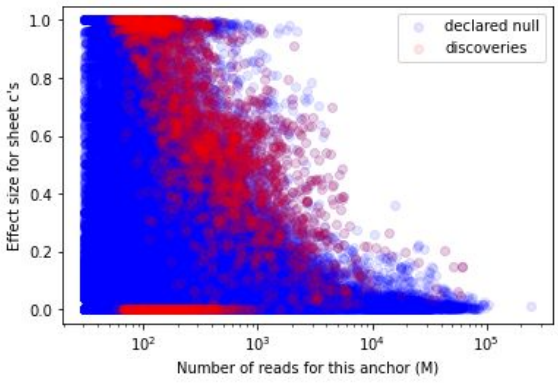
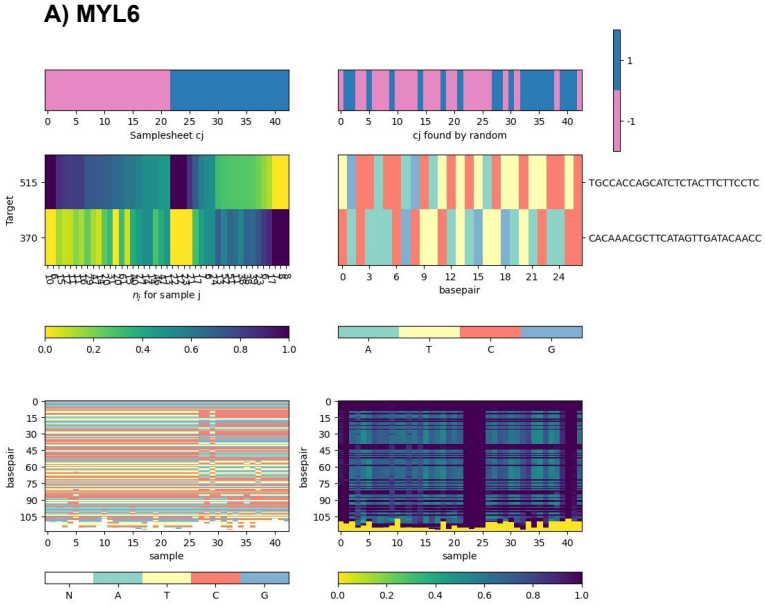
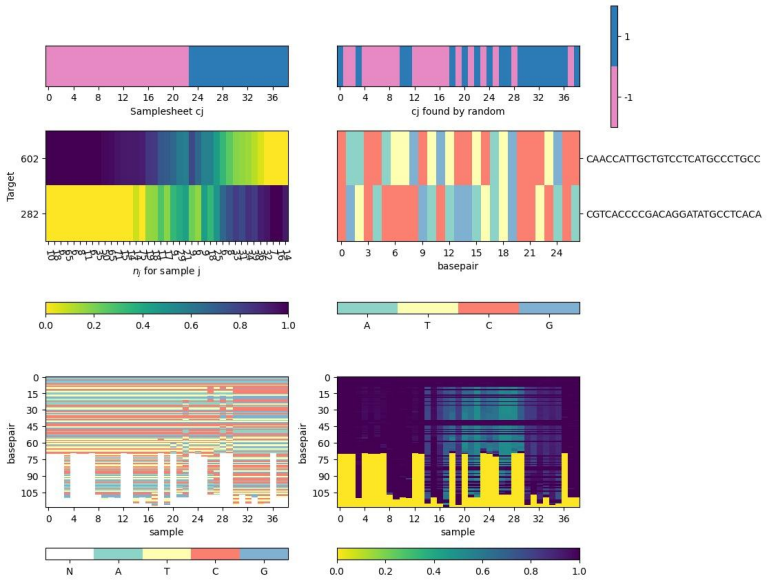


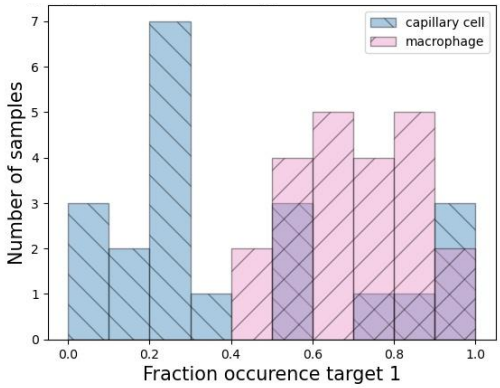
Figure S3

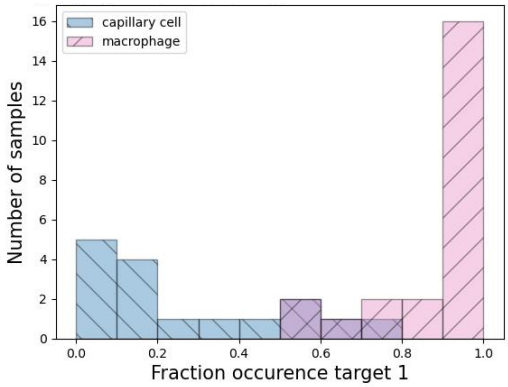
A. Random c's can recover samplesheet c's. For the HLCA dataset, of the 3439 anchors (1384 genes) called by the input metadata (samplesheet c's) in donor 1 (BY correction, alpha=.05), we have that 72% of the genes called were also called by NOMAD's selection of random c's (6287 called by anchors by random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

A) MYL6

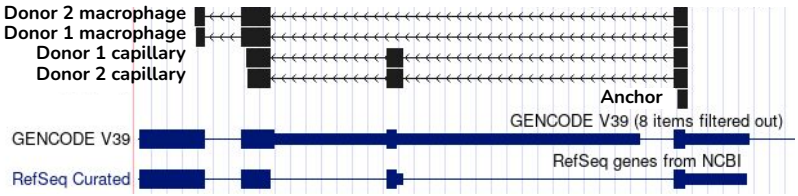








Example consensus sequences

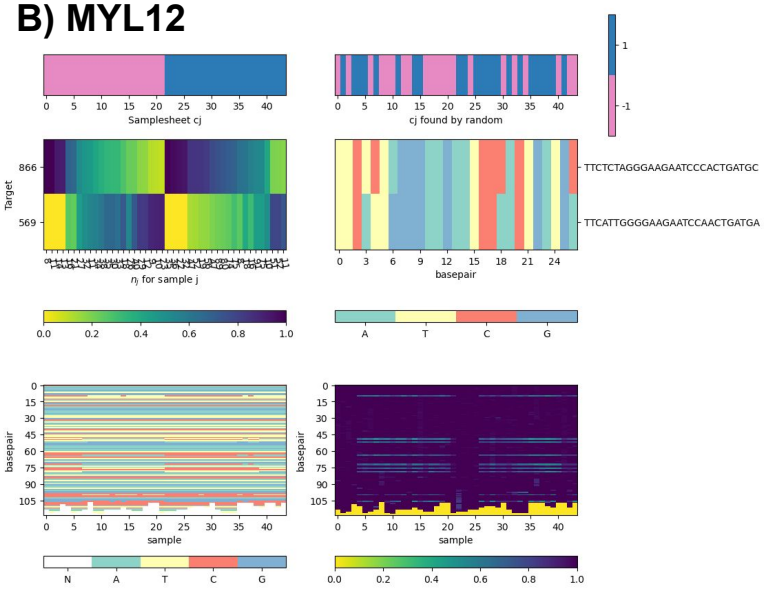


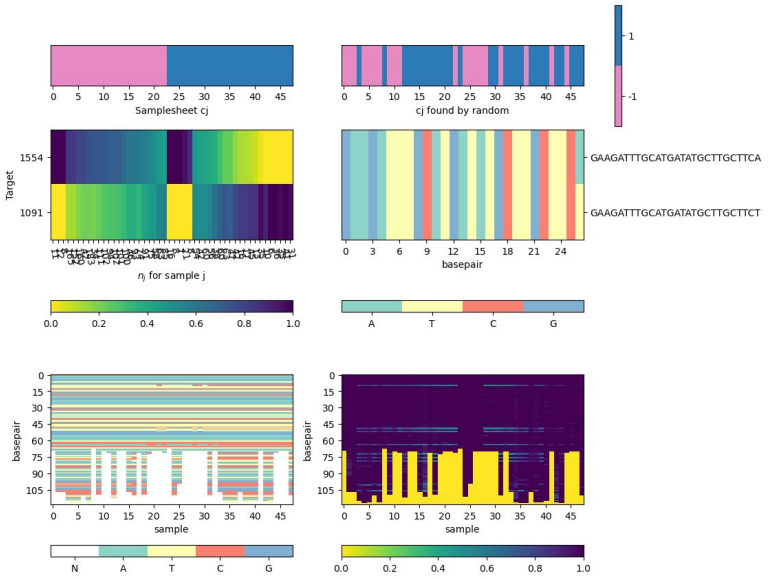
Exon-inclusion dominant ⟷ Exon skipping dominant

S4

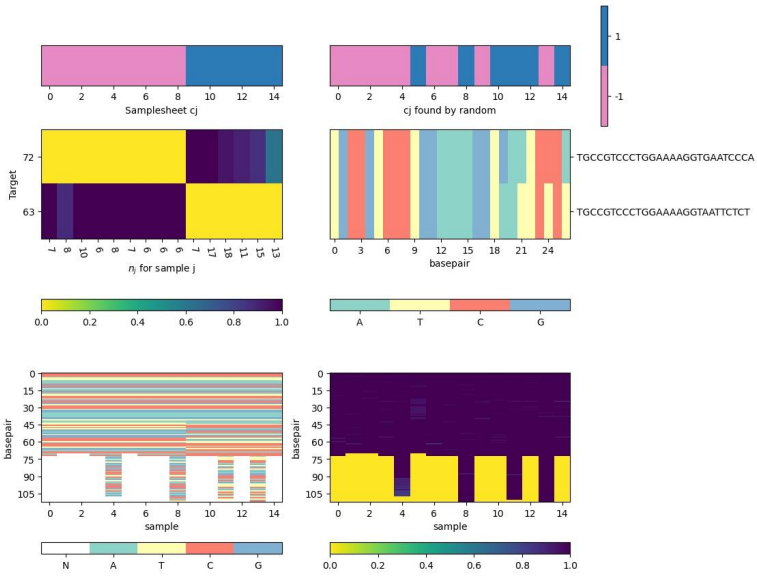
D

B) MYL12





C) HLA-DPB1



D) Human T Cell, HLA-B

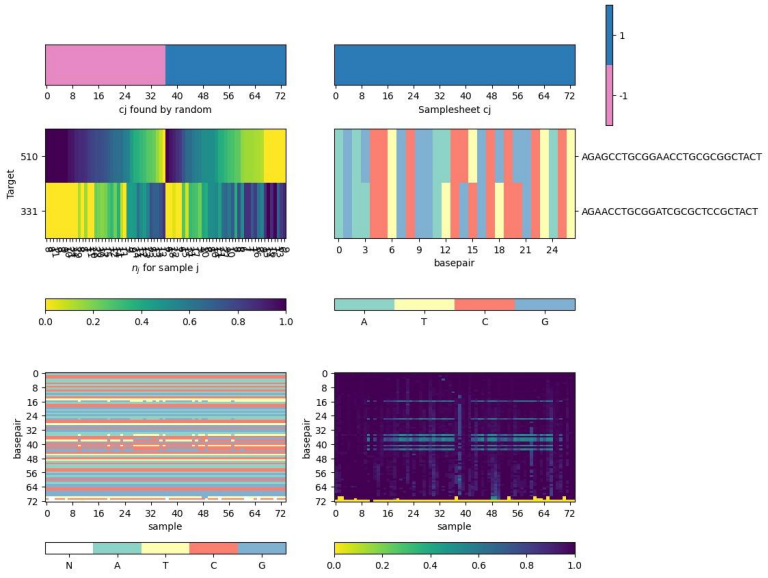


Figure S4

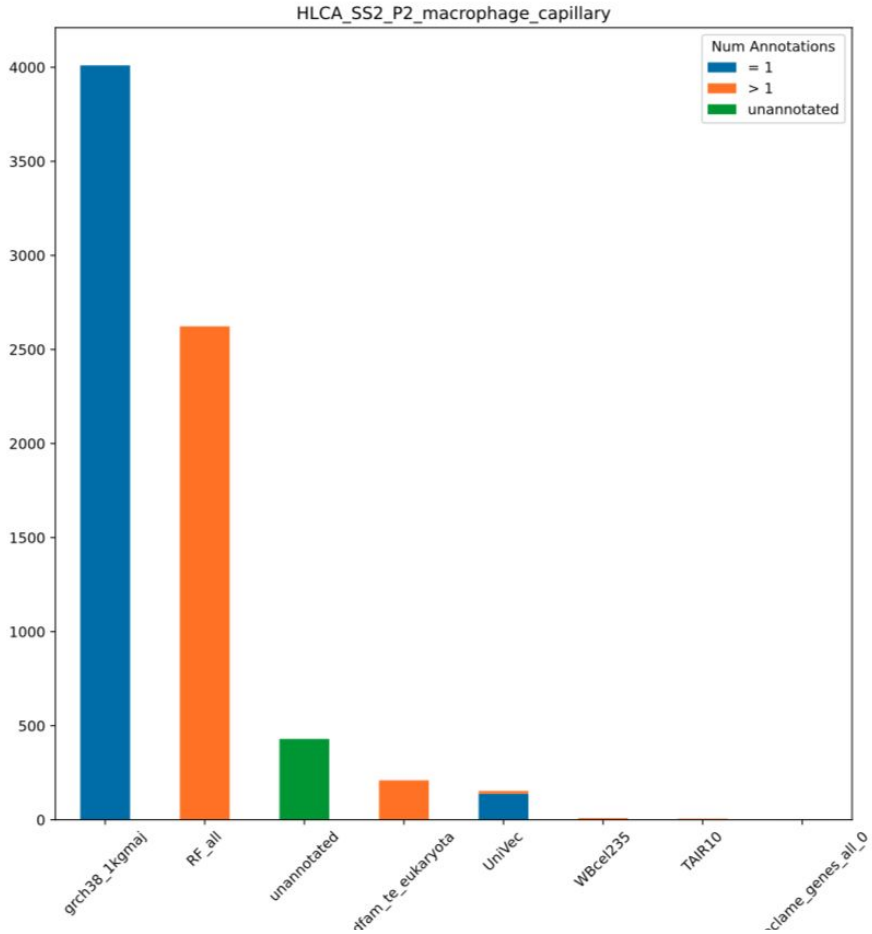
Heatmaps show the complete data for the called anchors. Each set of heatmaps is for one anchor sequence. The primary plot is the center left one, which shows the samples x targets contingency table. Each column represents a sample, and each row represents a unique target. The color indicates what fraction 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

A



B

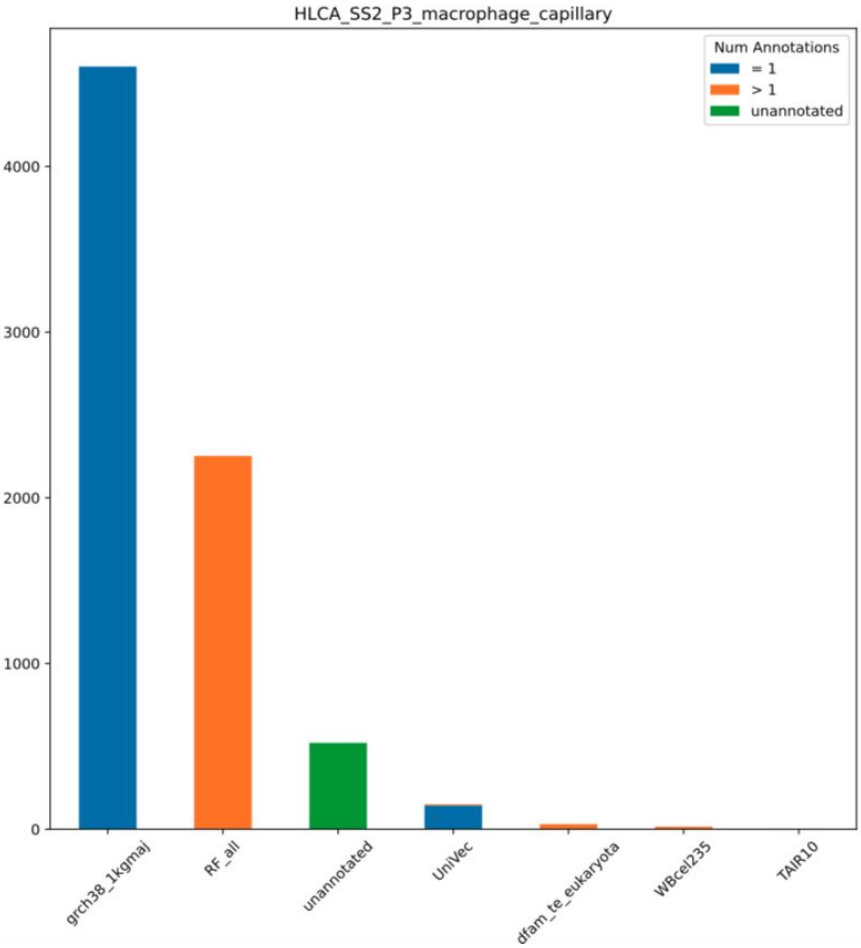
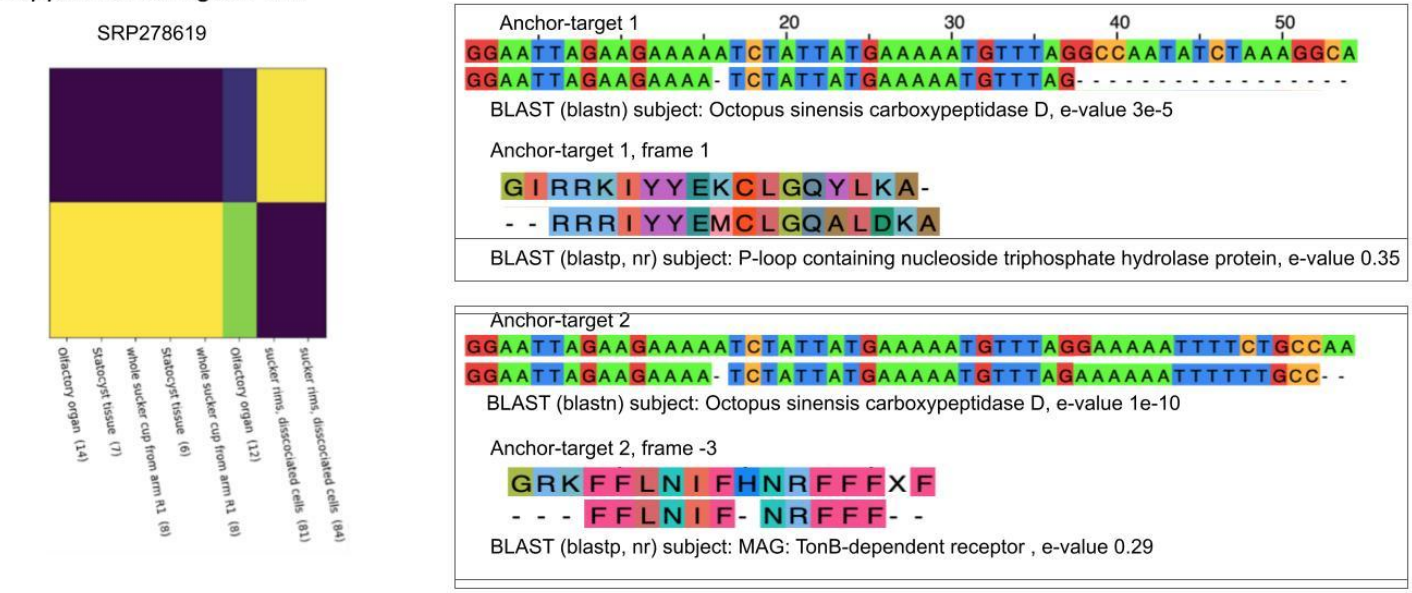


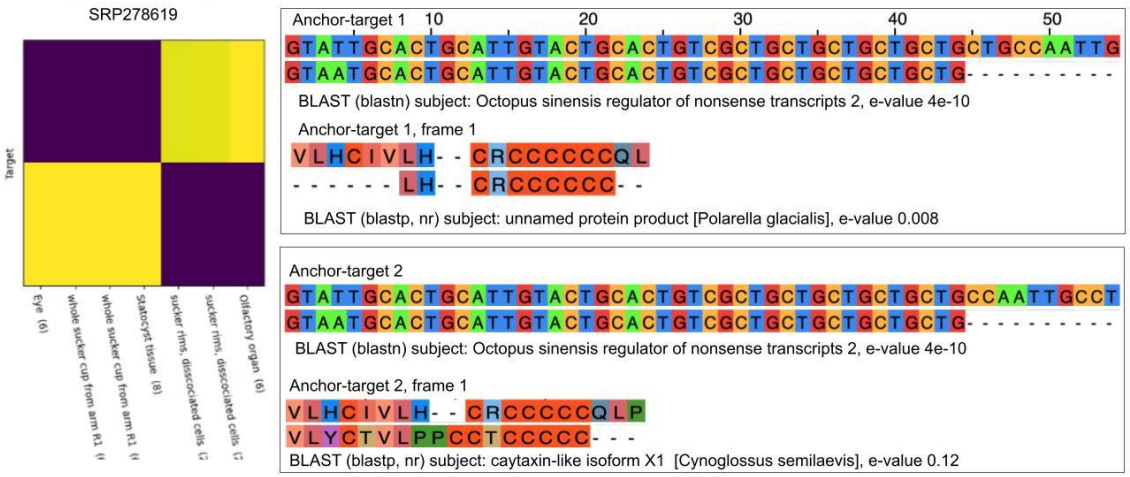
Fig. S5: Element Annotation Bar Plots

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.

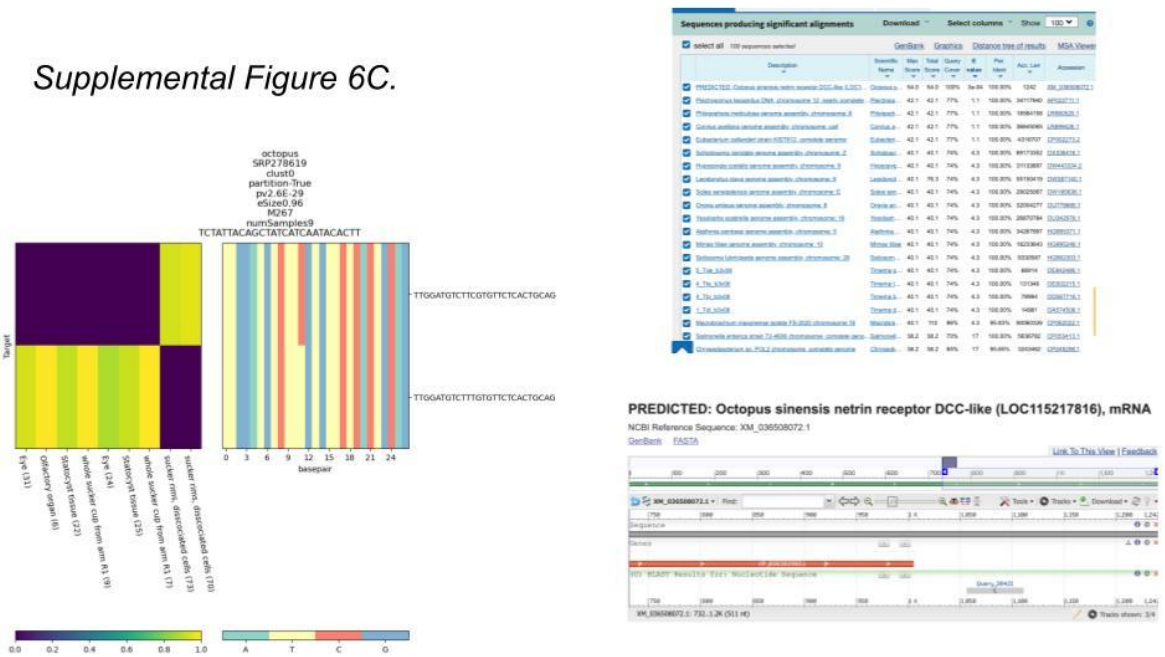


Fig. S6A: Predicted Octopus sinsensis carboxypeptidase (LOC115224523)

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.

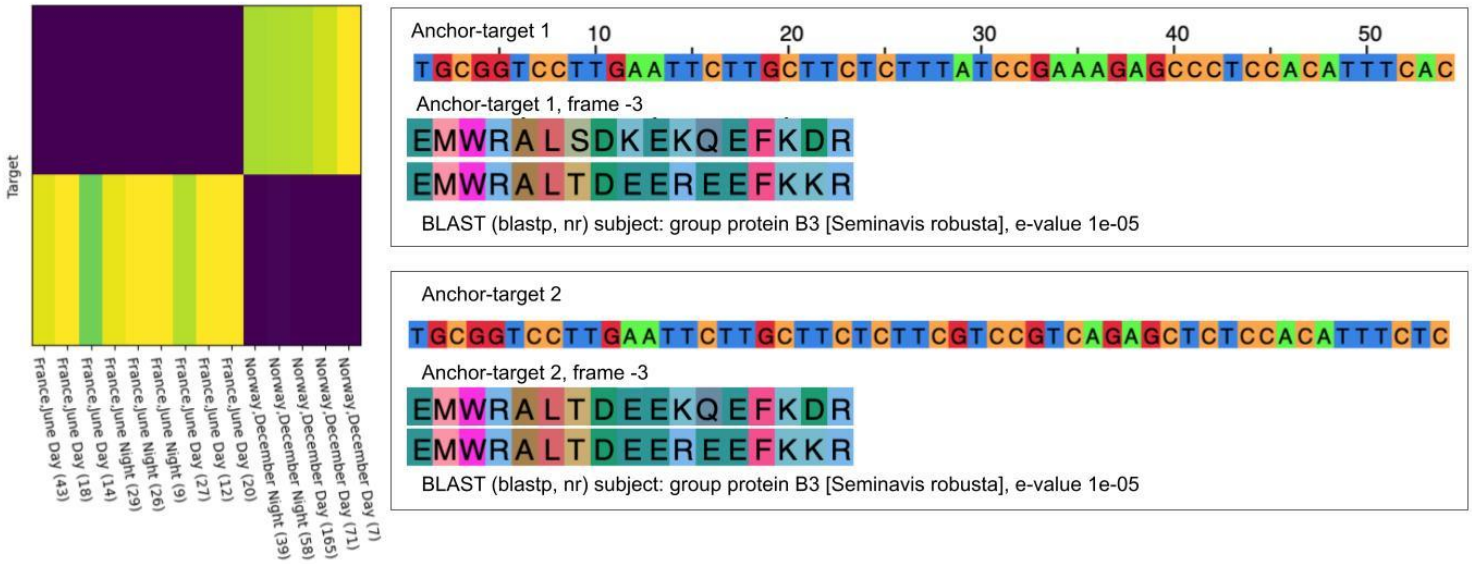
Fig. S6B: Predicted Octopus sinensis regulator of nonsense transcripts 2 (LOC115223858)

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.

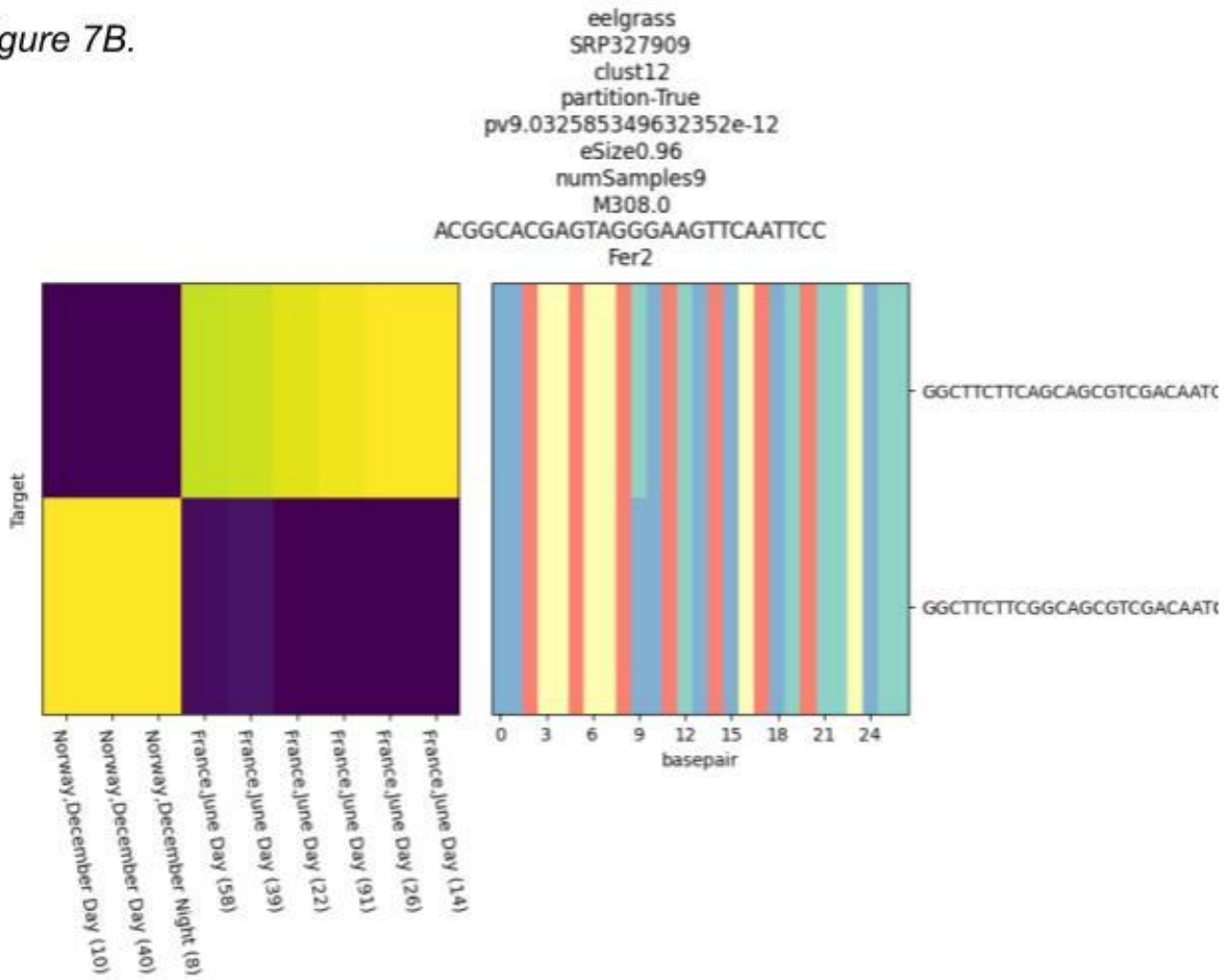
Fig. S6C: Predicted Octopus sinensis netrin receptor DCC-like (LOC115217816)

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.

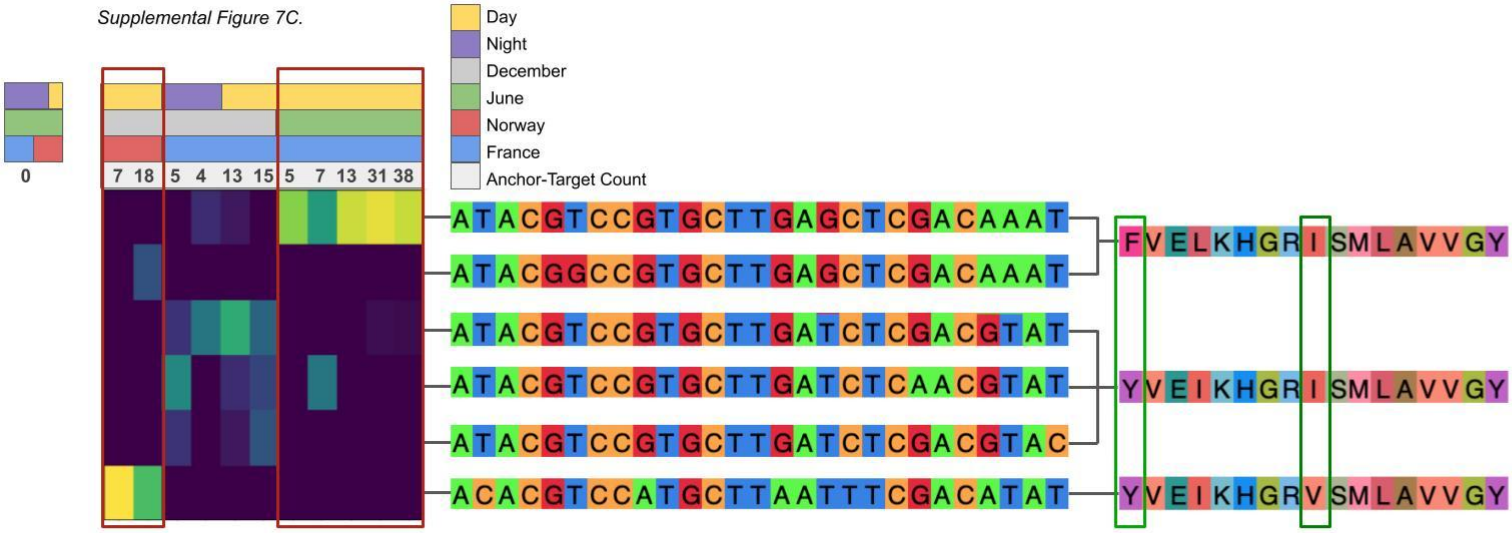
SRP327909 *Supplemental figure 7A.*



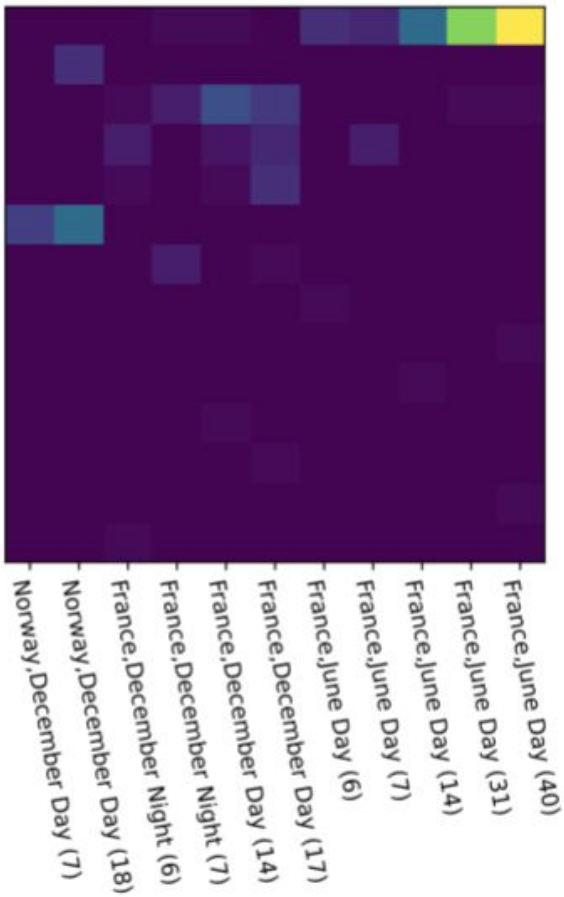
Supplemental Figure 7B.



Supplemental Figure 7C.



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 $\tilde{S}$, 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.

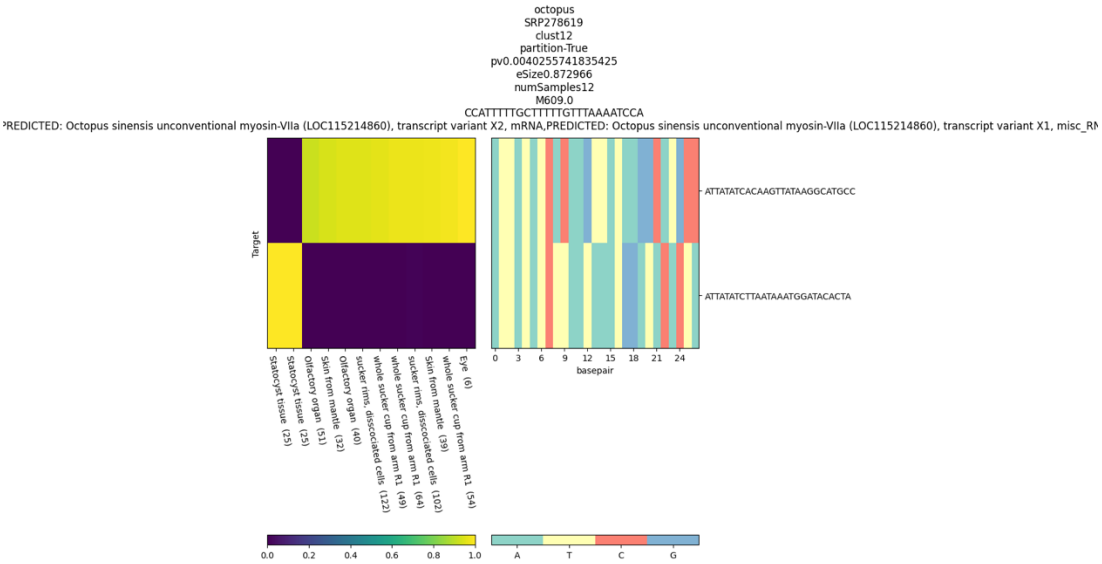
Figure data

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
```

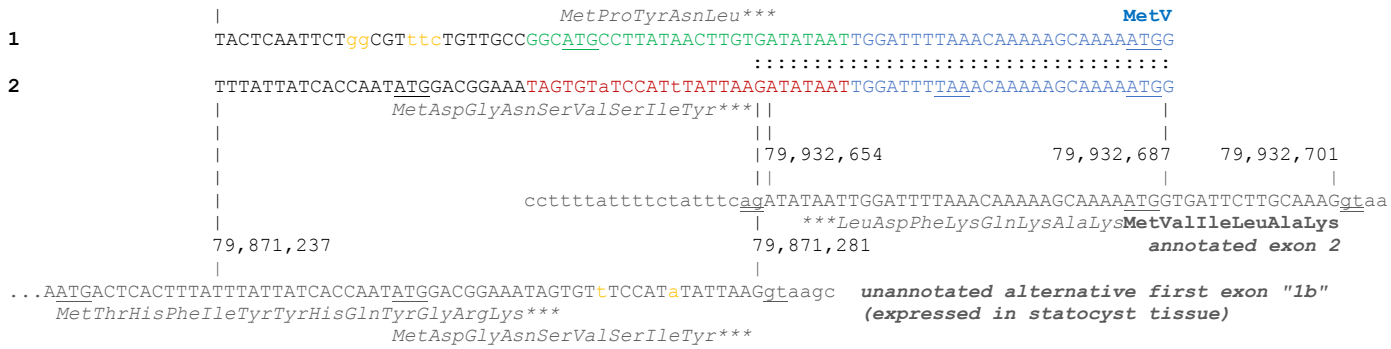
reverse-complements:

```
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
                  |                                                    |
```



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         TTCTTGCAAAGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  63
                 |||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  9744916   TTCTTGC--AGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  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          TTCTTGCAAAGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  63
                  |||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  79975491   TTCTTGC--AGGAGATCATGTGTGGCTGGAGCCAAAGACCAAAGAAGAATTTTCTGTGG  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.

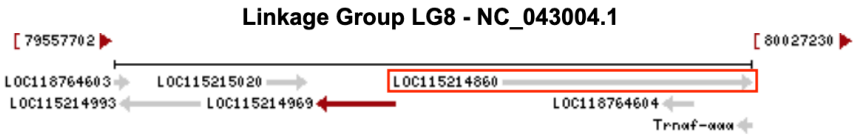
Similarly for *O. bimaculoides* "myosin-VIIa" genes:

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.

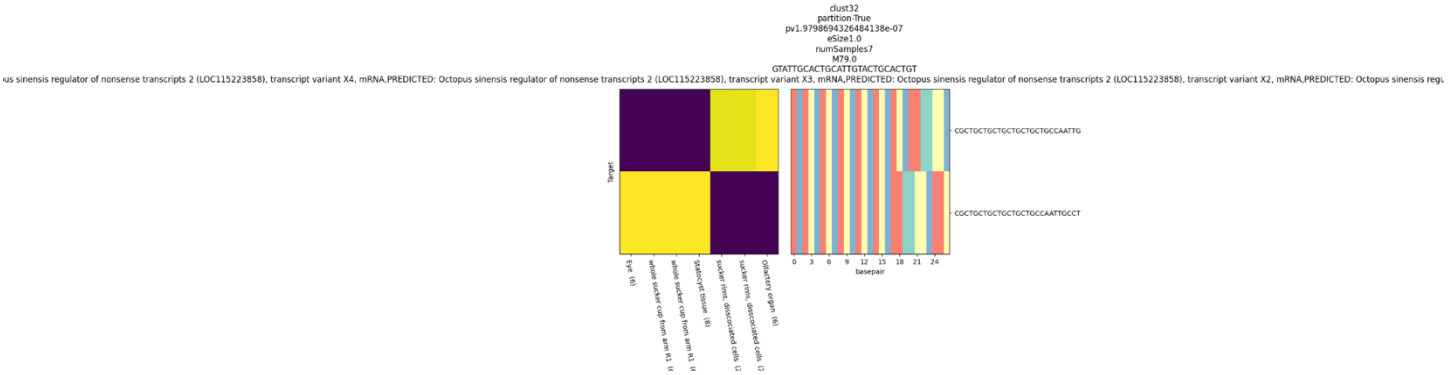
The *O. bimaculoides* gene on chromosome 8 appears to have a complete set of domains, including motor domain and two sets of the repeat domains. However, we know that it is lacks the 5' two exons present in *O. sinensis* myosin-VIIa LOC115214860.

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.

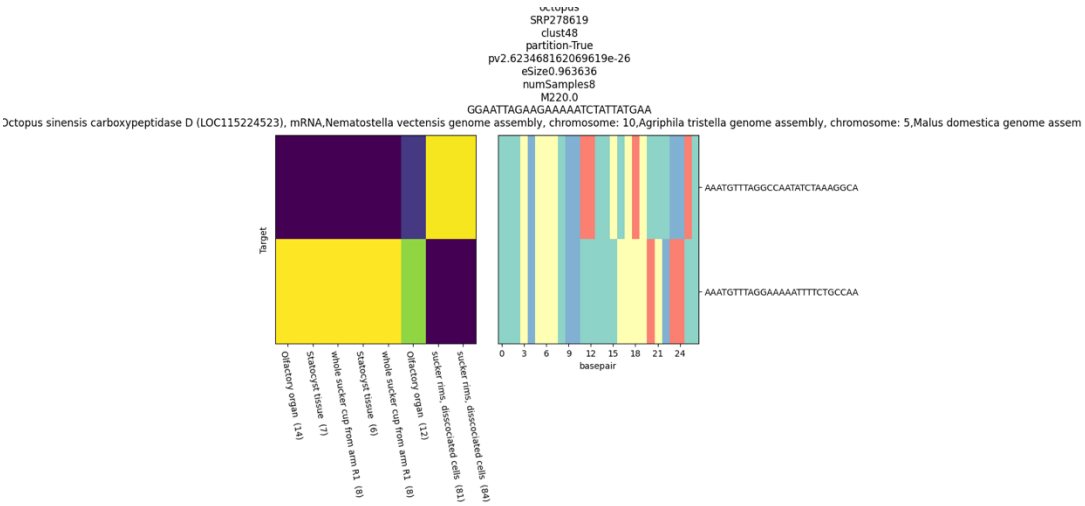
Alignment of transcript variants from each: *bimaculoides* = XM_014915650.2, *sinensis* = XM_036513028.1 shows very high identity throughout the coding sequence, but abrupt divergence from just before the stop codons (marked in red). The match to anchor-target is shown in blue background.

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_ AGCAAAGGTGACAGTGAGGTTGCTCCTCCTCCACCACCACCACCACCACAAGATAAGATA
     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

AT**1** GGAATTAGAAGAAAAATCTATTATGAAAAATGTTTAGGCCAATATCTAAAGGCA
AT**2** GGAATTAGAAGAAAAATCTATTATGAAAAATGTTTAGGAAAAATTTTCTGCCAA

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

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

AT**1**rc TGCCTTTAGATATTGGCCTAAACATTTTTCATAATAGATTTTTCTTCTAATTCC
AT**2**rc TTGGCAGAAAATTTTTTCCTAAACATTTTTTCATAATAGATTTTTTCTTCTAATTCC

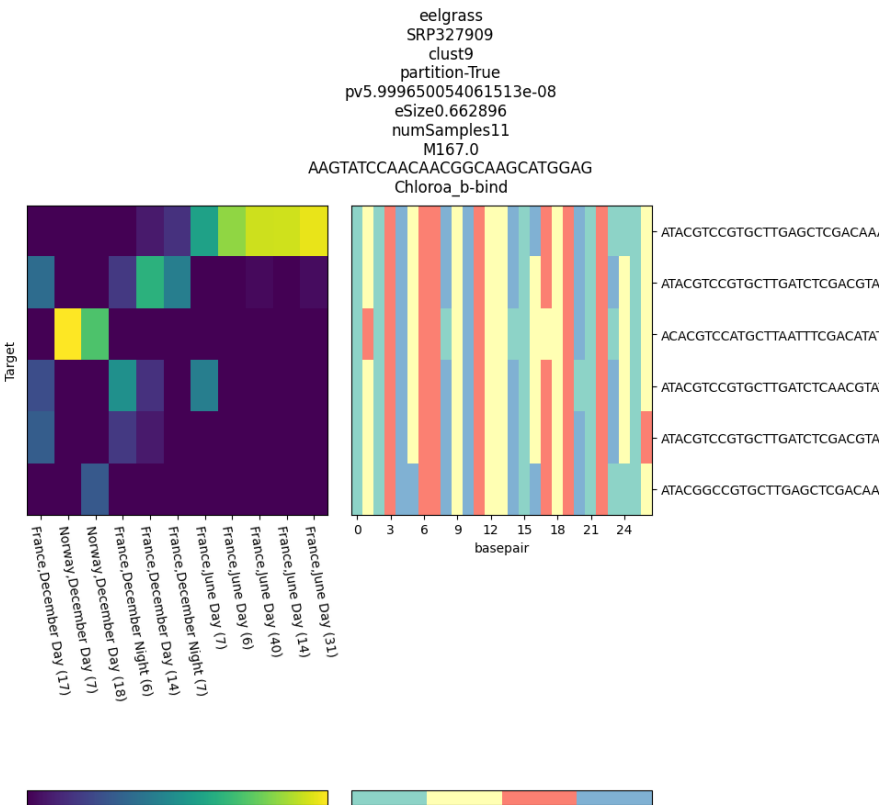
BLAST only finds these sequences in the *O. sinensis* genome, but **not** in the *O. bimaculoides* genome.

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_1, 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.

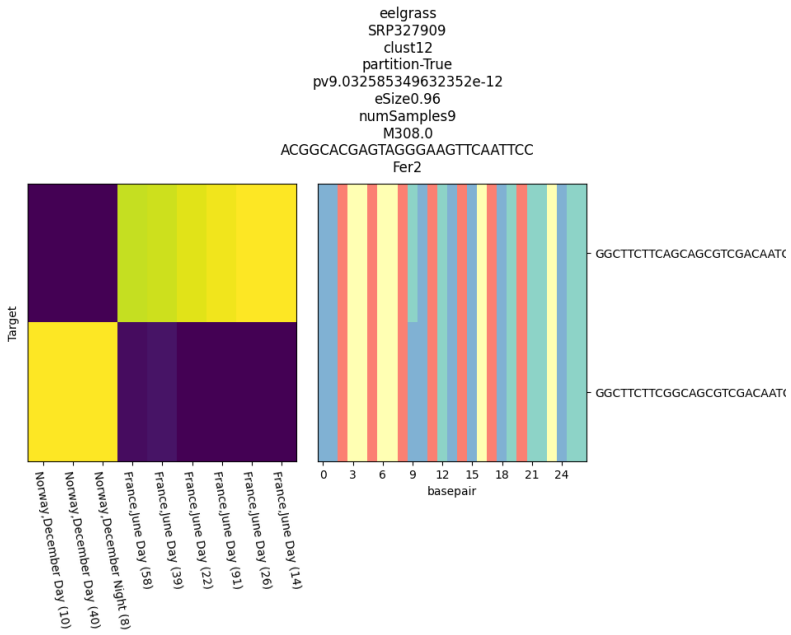
Top BLASTP hits (elided second hits are additional hits on species already shown). These and additonal top hits are **diatoms**.

```
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**.

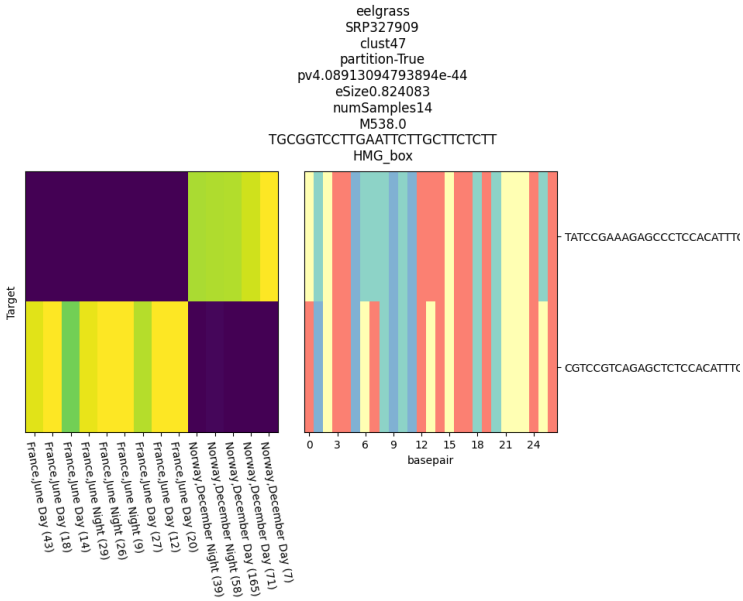
ferredoxin [Thalassiosira oceanica]

```
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.

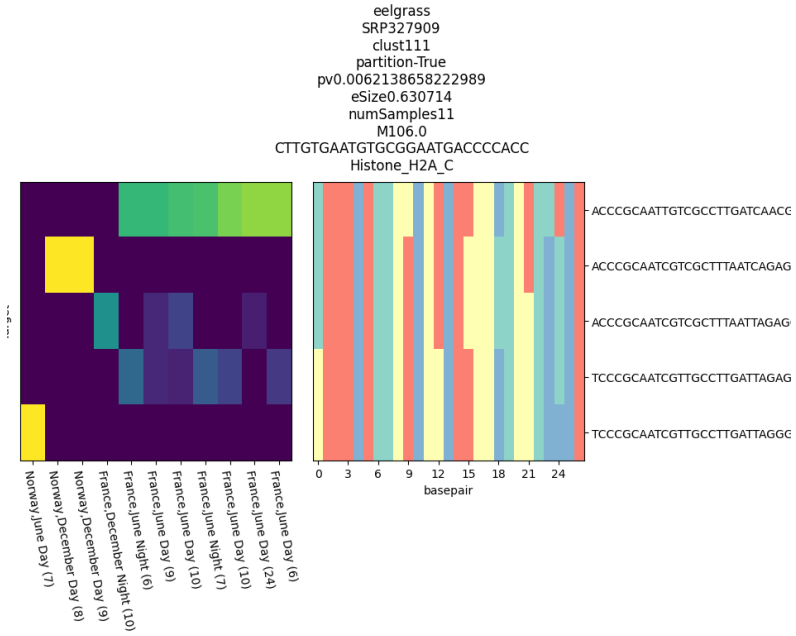
InterPro search of VKEDDPDLTFGGVGKKLGEMWRALSDKEKQEFKDR gives hit to IPR009071 / Pfam PF00505 = "HMG (high mobility group) box", covering entire sequence except last residue.

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.

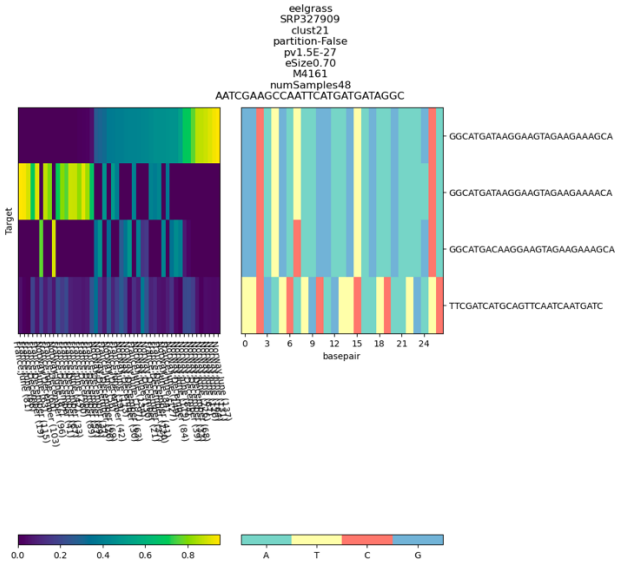
histone H2A isoform 1 [Phaeodactylum tricornutum CCAP 1055/1]

```
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
ATATGTTACAAAAAAAAAATAGCAAAAAATTTAAGCTTATTAAGAAGAATGAAAGCCCGAAGAGGAGTTTT
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
NNYEQDLTSV**LIQSGAFAFFYFLIMPPIIMNWL**RLRWYKRKLFET**YLQFMFVFLFFPGILLWAPFINF**RRLPRDPTMKHPWSTPRDSST

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

>intron-retention
MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLHLGVLLTSIADEPAFAVTGS
NNYEQDLTS**VLIQSGAFAFFYFLIMPV**YNCKVILT

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.