Supplemental Information

Document S1. Figures S1 to S7; Notes S1 to S7.
Data S1. Protein domain profiling output tables.
Data S2. Tables for SARS-CoV-2 and single-cell analyses: significant anchors, anchor statistics, and c_j's used for each anchor.
Data S3. Additional summary tables for macrophage and capillary single-cell analyses: significant anchors, their targets, anchor statistics, anchor and target reverse complement information, highest priority element annotations for anchors and targets, anchors annotations, and consensus annotations.
Data S4. Tables for human and lemur immune single-cell analyses: significant anchors, and their genome and transcriptome annotations.
Data S5. Artificial genome sequences with defining mutations for SARS-CoV-2 Delta, Omicron BA.1, and Omicron BA.2 strains.
Data S6. Table for *Octopus bimaculoides* analysis: significant anchors, their targets, anchor statistics, STAR mapping annotations, Pfam results, and BLAST results.
Data S7. Table for *Octopus bimaculoides* analysis: significant anchors, their targets, anchor statistics, STAR mapping annotations, Pfam results, and BLAST results.
Data S8. SARS-CoV-2 metadata.

Document S1

Figure S1

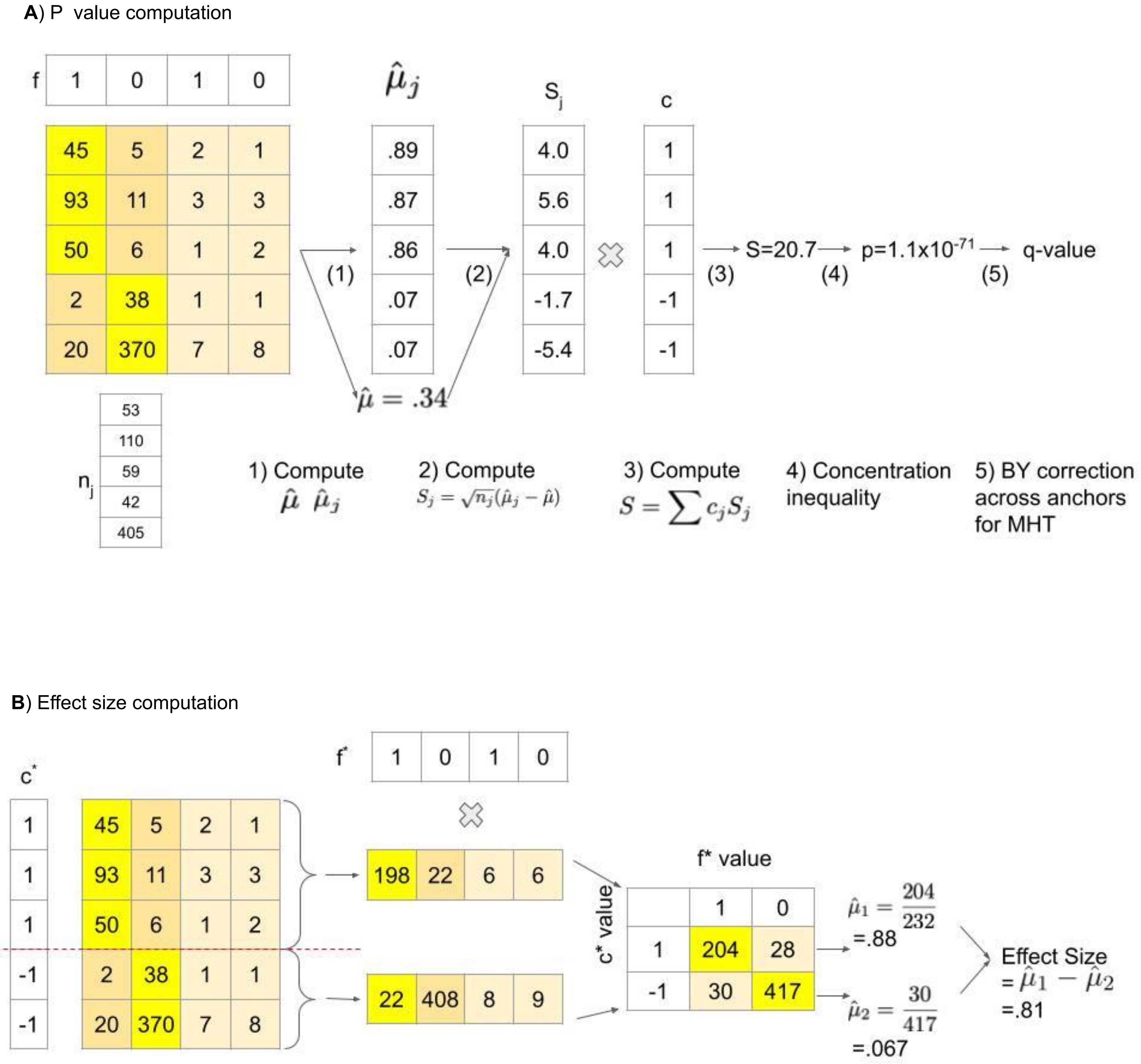

A) P value computation
f
1 0 1 0
45 5 2 1
93 11 3 3
50 6 1 2
2 38 1 1
20 370 7 8
n_j
53
110
59
42
405
$\hat{\mu}_j$
.89
.87
.86
.07
.07
$\hat{\mu} = .34$
S_j
4.0
5.6
4.0
-1.7
-5.4
c
1
1
1
-1
-1
(1)
(2)
(3)
S=20.7
(4)
$p=1.1x10^{-71}$
(5)
q-value
1) Compute $\hat{\mu}$ $\hat{\mu}_j$
2) Compute $S_j =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

Figure S2

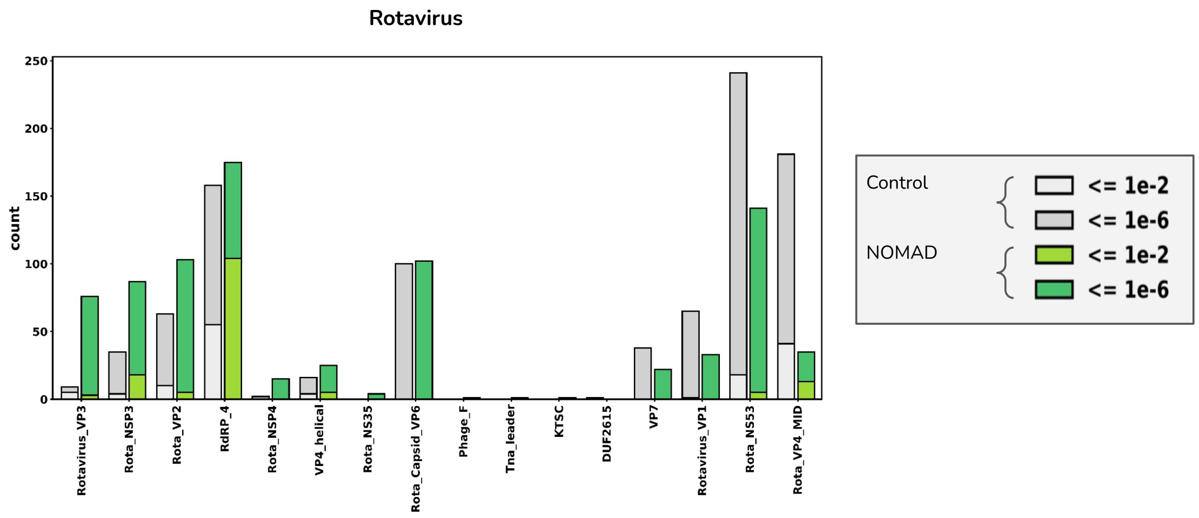


Figure S3

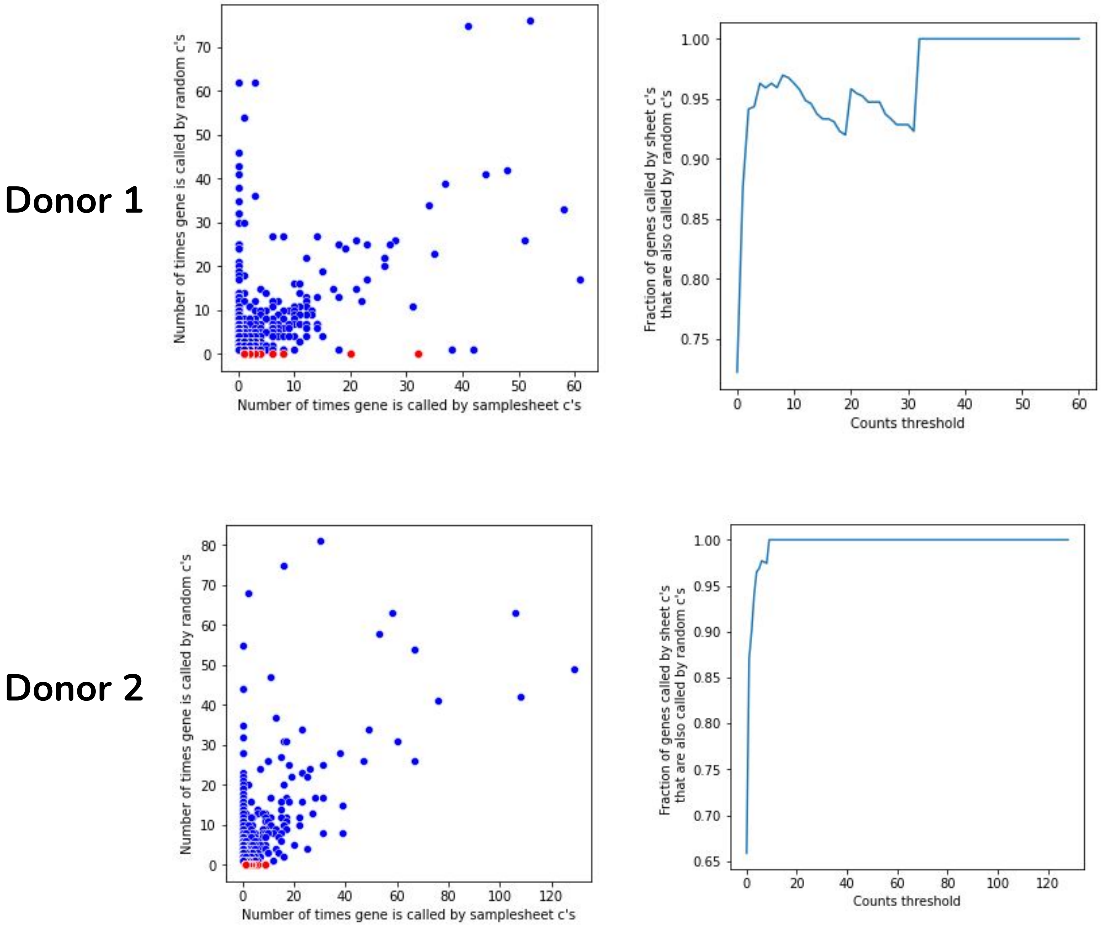


Figure S4

A. MYL6

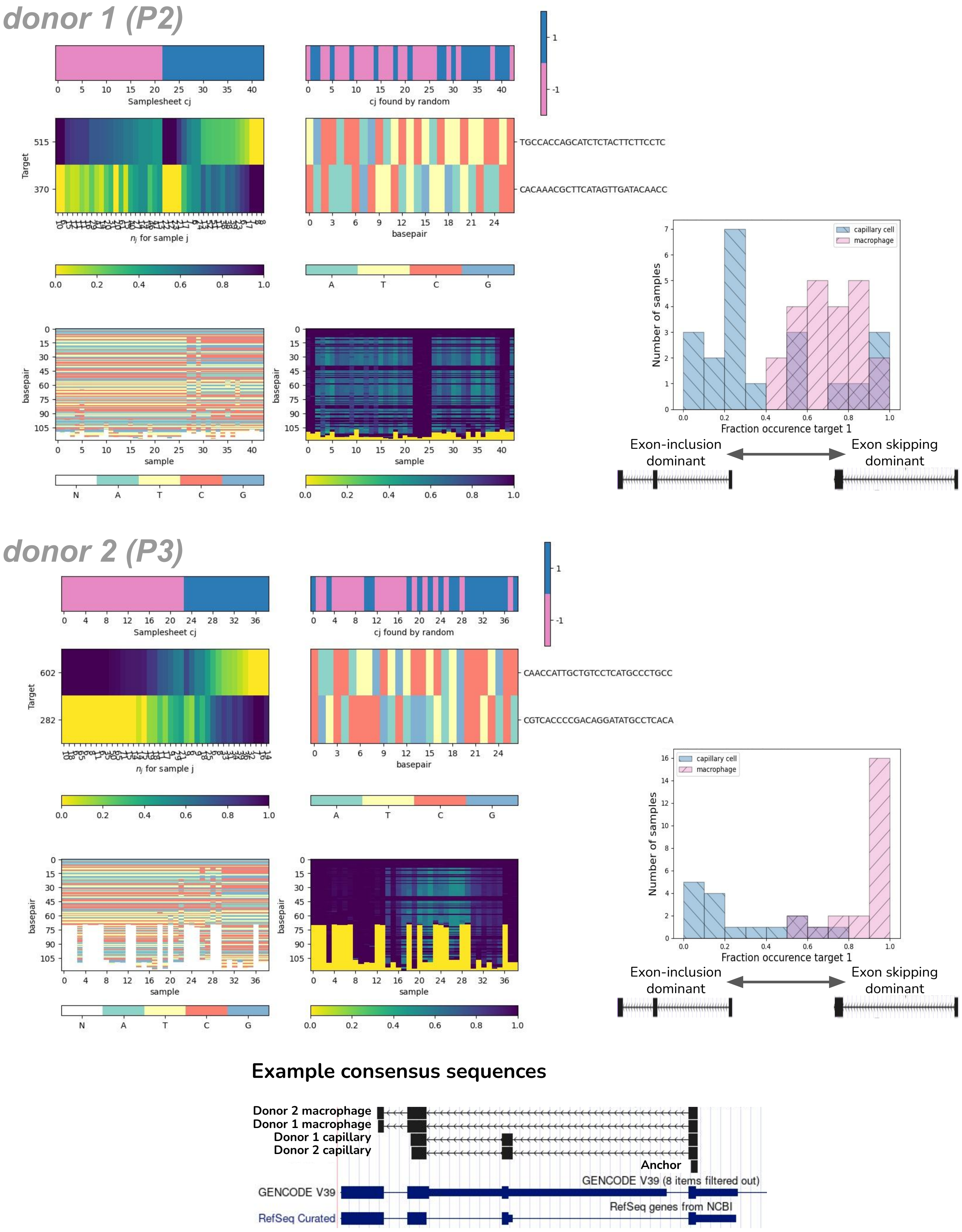
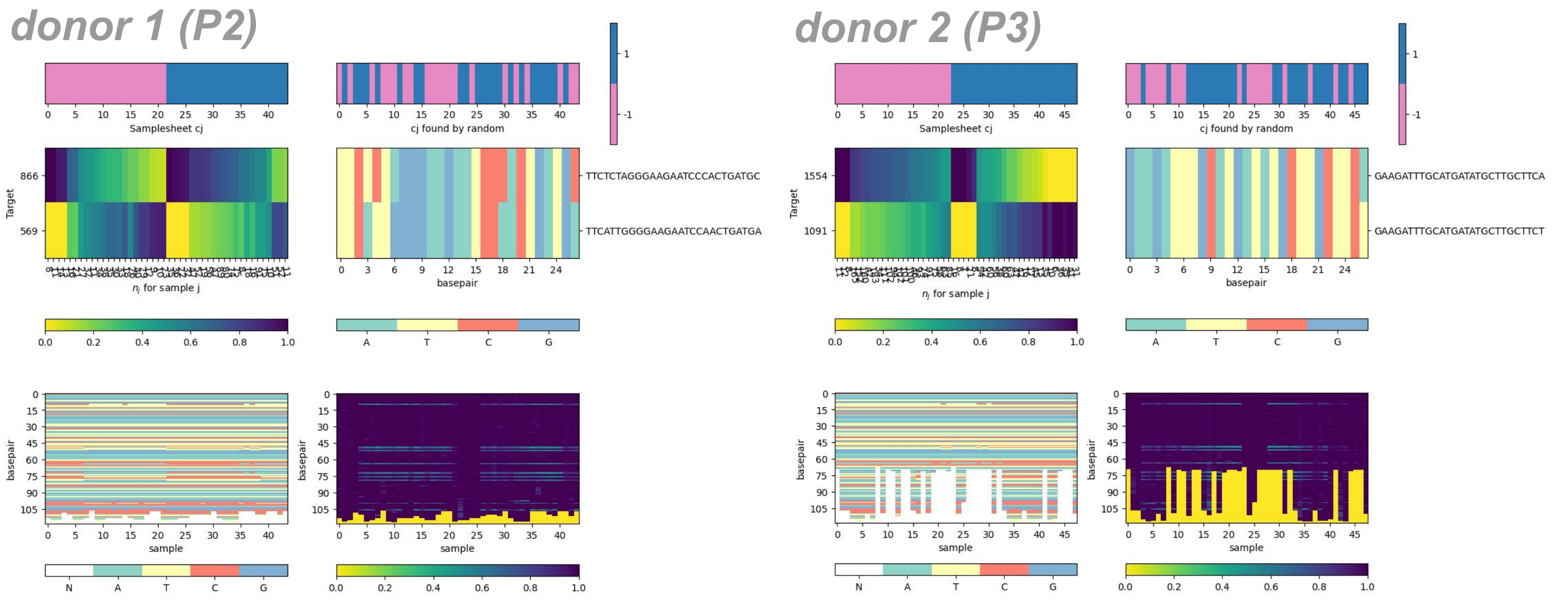
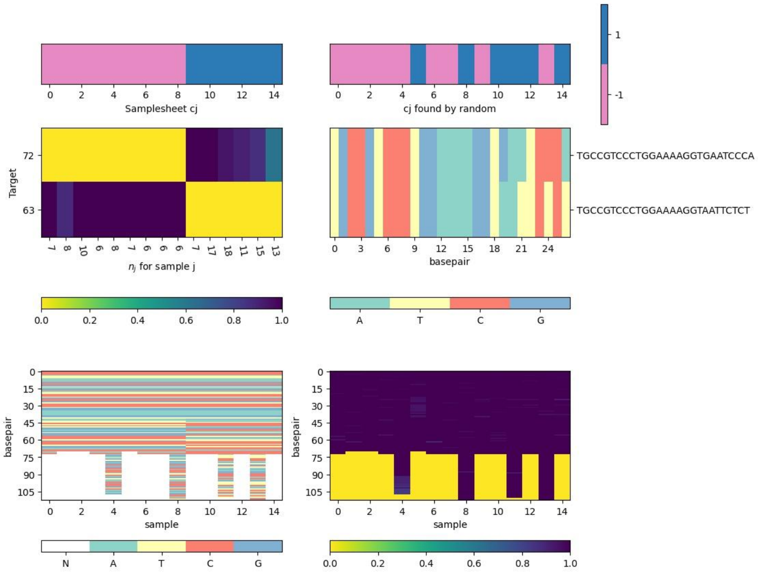


Figure S4

B. MYL12



C. HLA-DPB1



D. HLA-B, human T cell

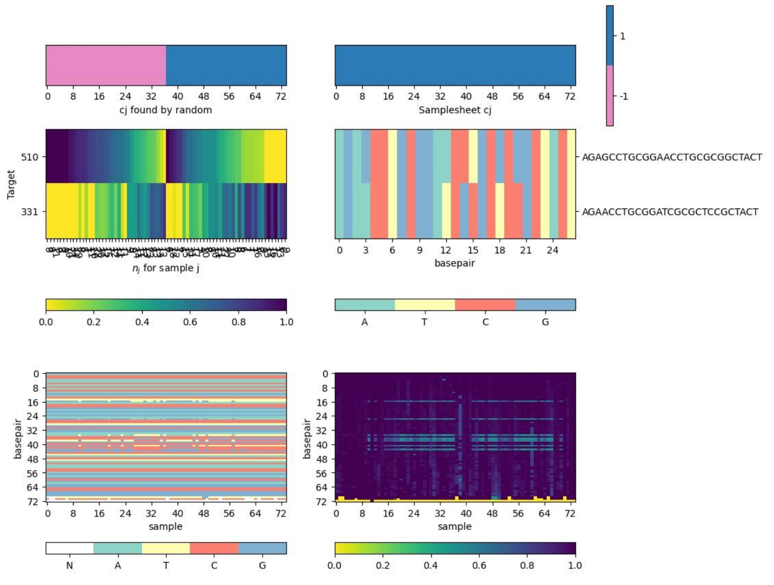


Figure S5

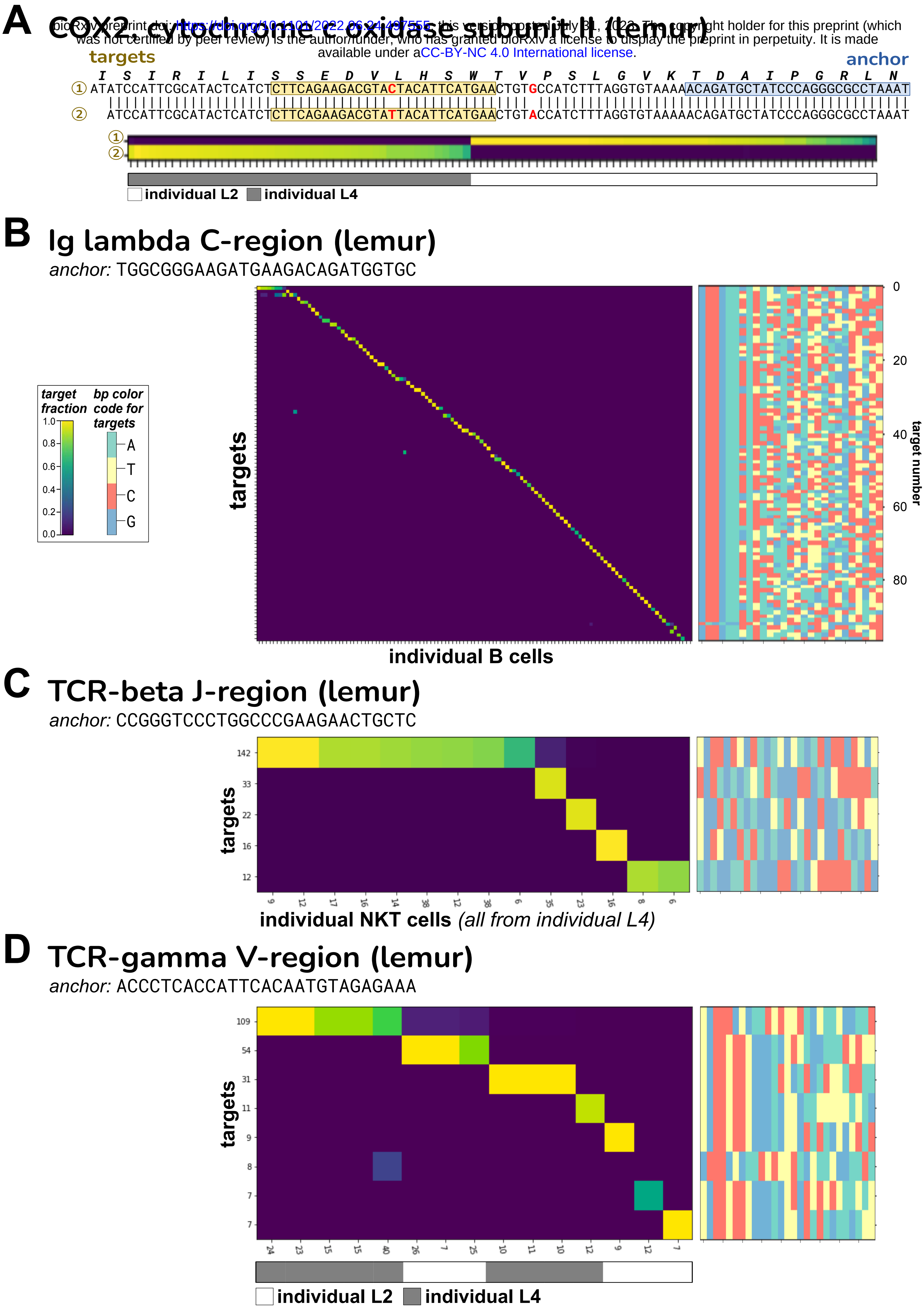
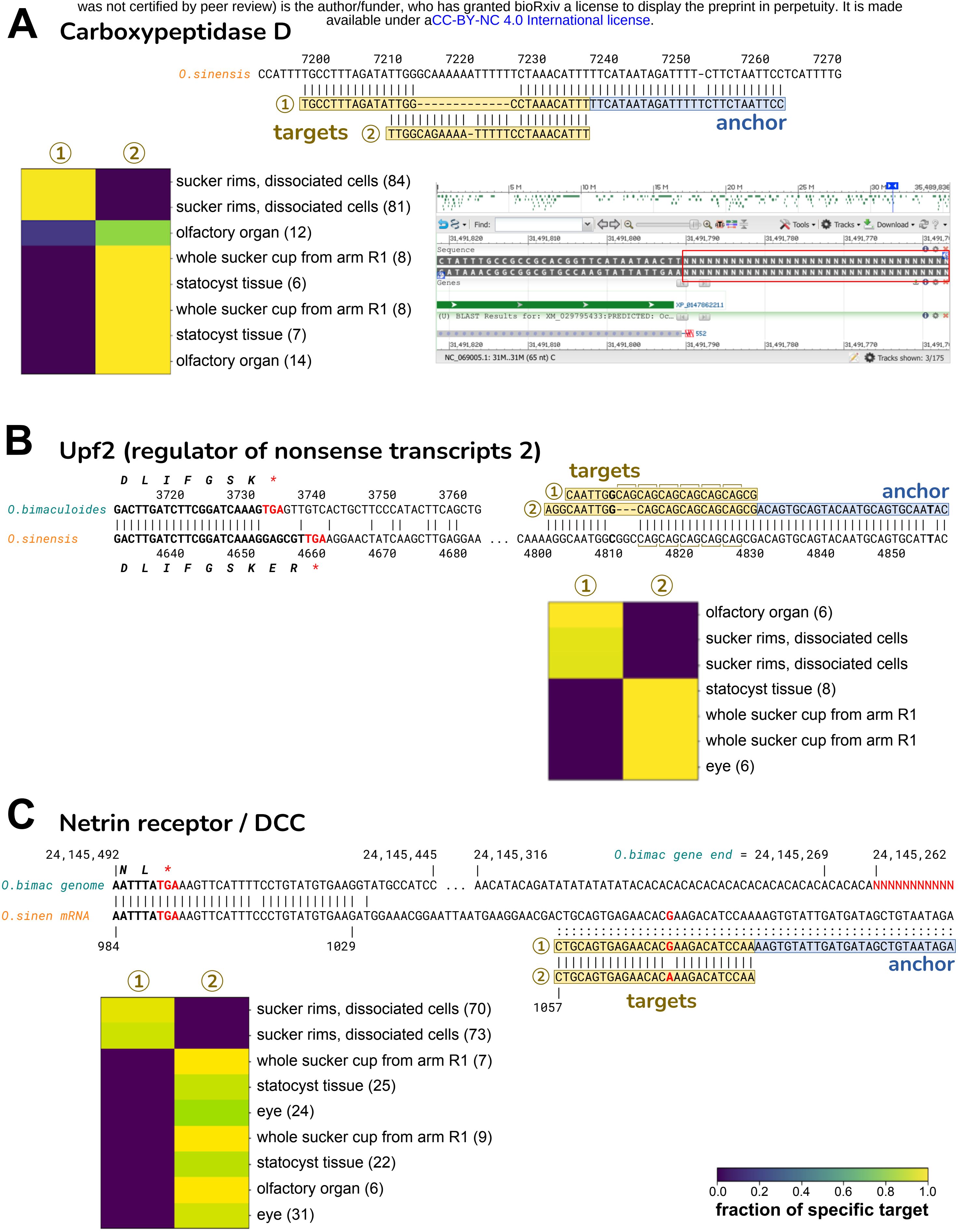

A
COX2, cytochrome c oxidase subunit II (lemur)
targets
anchor
I S I R I L I S S E D V L H S W T V P S L G V K T D A I P G R L N
① ATATCCATTCGCATACTCATCTCTTCAGAAGACGTACTACATTCATGAACTGTGCCATCTTTAGGTGTAAAAACAGATGCTATCCCAGGGCGCCTAAAT
② ATCCATTCGCATACTCATCTCTTCAGAAGACGTATTACATTCATGAACTGTACCATCTTTAGGTGTAAAAACAGATGCTATCCCAGGGCGCCTAAAT
①
②
individual L2
individual L4
B
Ig lambda C-region (lemur)
anchor: TGGCGGGAAGATGAAGACAGATGGTGC
target fraction
1.0
0.8
0.6
0.4
0.2
0.0
bp color code for targets
A
T
C
G
targets
individual B cells
0
20
40
60
80
target number
C
TCR-beta J-region (lemur)
anchor: CCGGGTCCCTGGCCCGAAGAACTGCTC
142
33
22
16
12
targets
9
12
17
16
14
38
12
38
6
35
23
16
8
6
individual NKT cells (all from individual L4)
D
TCR-gamma V-region (lemur)
anchor: ACCCTCACCATTCACAATGTAGAGAAA
109
54
31
11
9
8
7
7
targets
24
23
15
15
40
26
7
25
10
11
10
12
9
12
7
individual L2
individual L4

Figure S6

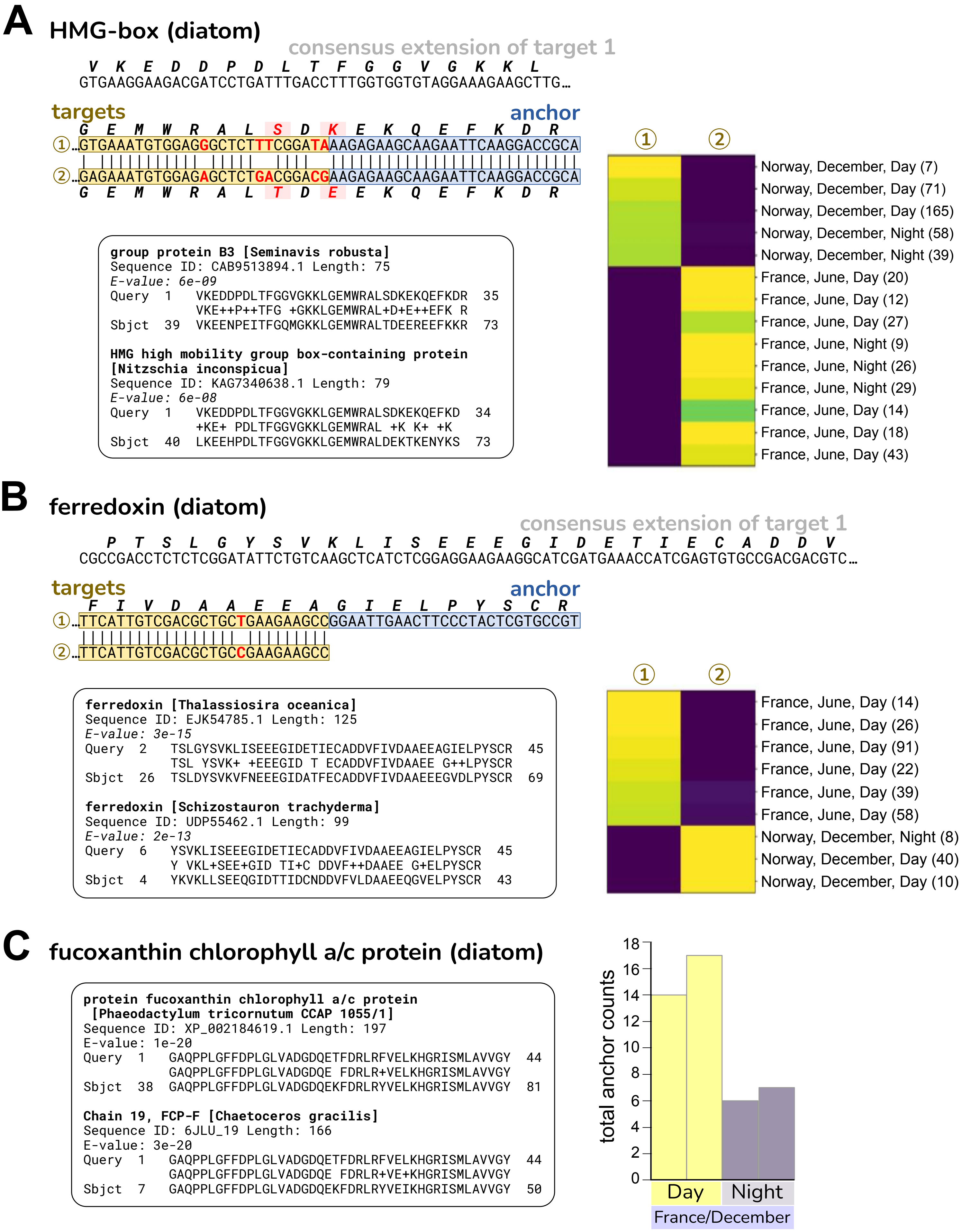


A Carboxypeptidase D

B Upf2 (regulator of nonsense transcripts 2)

C Netrin receptor / DCC

Figure S7



```
group protein B3 [Seminavis robusta]
Sequence ID: CAB9513894.1 Length: 75
E-value: 6e-09
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
E-value: 6e-08
Query  1    VKEDDPDLTFGGVGKKLGEMWRALSDKEKQEFKD  34
            +KE+ PDLTFGGVGKKLGEMWRAL +K K+ +K
Sbjct  40   LKEEHPDLTFGGVGKKLGEMWRALDEKTKENYKS  73
```

```
ferredoxin [Thalassiosira oceanica]
Sequence ID: EJK54785.1 Length: 125
E-value: 3e-15
Query  2   TSLGYSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR  45
           TSL YSVK+ +EEEGID T ECADDVFIVDAAEE G++LPYSCR
Sbjct  26  TSLDYSVKVFNEEEGIDATFECADDVFIVDAAEEEGVDLPYSCR  69

ferredoxin [Schizostauron trachyderma]
Sequence ID: UDP55462.1 Length: 99
E-value: 2e-13
Query  6   YSVKLISEEEGIDETIECADDVFIVDAAEEAGIELPYSCR  45
           Y VKL+SEE+GID TI+C DDVF++DAAEE G+ELPYSCR
Sbjct  4   YKVKLLSEEQGIDTTIDCNDDVFVLDAAEEQGVELPYSCR  43
```

```
protein fucoxanthin chlorophyll a/c protein
 [Phaeodactylum tricornutum CCAP 1055/1]
Sequence ID: XP_002184619.1 Length: 197
E-value: 1e-20
Query  1   GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY  44
           GAQPPLGFFDPLGLVADGDQE FDRLR+VELKHGRISMLAVVGY
Sbjct  38  GAQPPLGFFDPLGLVADGDQEKFDRLRYVELKHGRISMLAVVGY  81

Chain 19, FCP-F [Chaetoceros gracilis]
Sequence ID: 6JLU_19 Length: 166
E-value: 3e-20
Query  1   GAQPPLGFFDPLGLVADGDQETFDRLRFVELKHGRISMLAVVGY  44
           GAQPPLGFFDPLGLVADGDQE FDRLR+VE+KHGRISMLAVVGY
Sbjct  7   GAQPPLGFFDPLGLVADGDQEKFDRLRYVEIKHGRISMLAVVGY  50
```

Supplemental Figure Legends

Figure S1. SPLASH computations.

A. *p*-value computation for SPLASH for user specified vectors *f* and *c*. Contingency table transposed for visual convenience (rows are samples and columns are targets). Starting with a samples by targets counts matrix, SPLASH 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 Benjamini Hochberg testing which corrects for arbitrary dependence.

B. Effect size computation for SPLASH. Reported effect size is calculated based on the random split *c* and random function *f* that yielded the most significant SPLASH *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 $\mu_1 - \mu_2$.

Figure S2. Rotavirus protein domain profiling.

We performed SPLASH protein domain profiling in a dataset of virally enriched samples from breakthrough infections in patients that had been vaccinated against rotavirus; nearly all did have rotavirus infection, but some also had other coinfections[17]. The top domains were rotavirus VP3 (Rotavirus_VP3, 76 SPLASH hits vs 9 control hits) followed by rotavirus NSP3 (Rota_NSP3, 87 SPLASH vs 35 control hits), indicating that these rotavirus genes have sequences that vary significantly among these patients; they have roles in blocking host innate immunity[18], and so may be under selective pressure for variation.

Figure S3. Effectiveness of SPLASH randomized sample splitting.

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 SPLASH'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.

Figure S4. Additional details for single cell analyses.

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 × targets contingency table. Each column represents a sample, and each row represents a unique target. The color indicates what fraction of the sample's (column's) targets come from the target corresponding to that row. The *x*-ticks correspond to n_j, the number of times the anchor was observed in this sample. The *y*-ticks indicate the number of times this target appeared (following this anchor), and the targets are sorted by abundance. The two top plots indicate the c_j's used; when samplesheet c_js are available, they will be in the upper left, and the optimizing random c_js 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.

Panels A-C are macrophage and capillary cells from human lung (HLCA dataset). In the samplesheet c_js metadata heatmap (upper left) and histogram plots for panel A, pink is macrophage and blue is capillary cell.

- **A. MYL6.** SPLASH rediscovers alternative splicing in MYL6: capillary cells express more of the exon-inclusion variant, while macrophages express predominantly the exon-skipped variant. At the bottom is a UCSC Genome Browser screenshot, showing BLAT mapping of macrophage and capillary consensus sequences which fully capture the inclusion or exclusion of the short alternative exon.
- **B. MYL12.** This data is partly presented in Figure 3A.
- **C. HLA-DPB1.** This data is partly presented in Figure 3C.
- **D. HLA-B (human T cells, Tabula Sapiens).** This data is partly presented in Figure 3D. In this case, all cells have the same metadata label (CD4+ T cell) so only the optimizing random c_js heatmap is relevant and so it is shown in the upper left. The best random c_js found divide the cells into two groups that differ in their relative distribution of targets (note that other c_js might give an even larger difference), indicating that HLA-B expression is not uniform across this set of cells.

Figure S5. Lemur COX2 allelic detection, additional Ig/TCR anchors.

For panels B-D, target × sample heatmaps are shown on the left (samples are individual cells), and bp color-maps on the right. The latter encode the target sequences with different colors for each type of base, and are only provided to give a quick visual impression of the sequence variability. The small numbers at the sides of the heatmaps are summed target counts (over rows or columns).

- **A. COX2.** SPLASH detects numerous anchors that have targets that correspond precisely to the identities of the two lemur individuals represented in the set of cells analyzed. One example is an anchor in COX2. The alignment of two consensuses (both from NKT cells) shows that targets 1 and 2 differ at a single silent position (C vs T) and another silent SNP is also present outside the target region (G vs A); these positions are highlighted in red. Consensuses 1 and 2 align perfectly to different lemur mitochondrial genome accessions (NC_028718.1 and KR911907.1). Targets 1 and 2 are found exclusively in individuals L4 and L2, respectively.
- **B. Ig-lambda C-region.** Depicted is a lemur anchor that maps to the 5’ end of the human Ig-lambda C3 segment. It has 97 different targets, which lie within the hypervariable CDR3 region. Nearly all targets are expressed clonotypically (i.e., cells do not share targets), except for targets 1 and 3 (at the top of the heatmap).

C. **Ig-beta J-region.** We analyzed natural killer T (NKT) cells from lemur. In humans and mice, a subset of NKT cells express stereotypical or shared TCR genes; Figure 4B shows an example anchor in TCR-alpha. Here we show an anchor that maps to a human TCR-beta J-region (J2-1); targets reside in the V-region. A number of the NKT cells express a shared TCR-beta target (seen in the top row of the heatmap). All of these cells derive from individual L4.

D. **Ig-gamma V-region.** There is evidence for shared TCR-gamma sequences in NKT cells as well. This anchor maps to a human TCR-gamma V-region (TRGV9); the targets lie at the V-J junction. Individual L4 has two different shared targets (rows 1 and 3) while individual L2 shows three cells expressing a shared target (row 2). Other NKT cells express unique targets, however.

Figure S6. *O. bimaculoides* 3' UTR anchors show tissue-specific expression. In the heatmap , the parenthetical numbers are summed anchor counts.

A. **Carboxypeptidase D (CPD).** The anchor and targets align to the 3' UTR of the *O. sinensis* CPD mRNA (XM_029795433.2), but are not found in the *O. bimaculoides* genome assembly. The NCBI Browser screenshot at lower-right shows that the 3' UTR of the *O. bimaculoides* CPD gene (LOC106880679, Ch.25) is entirely missing from the genome: right after the end of the coding region, a run of Ns begins (red box). (There is a second *O. bimaculoides* CPD gene on Ch.16, LOC106873734, but has much lower identity to the sole *O. sinensis* CPD.) Target 2 is identical to *O. sinensis* except for two mismatches; target 1 has a 12-nt deletion relative to target 2. Target 1 is only expressed in dissociated cells from sucker rims, and at a low level in one olfactory organ sample. All other tissues express only target 2.

B. **Upf2 (regulator of nonsense transcripts 2).** The alignment of Upf2 mRNAs from *O. bimaculoides* (XM_014915650.2) and *O. sinensis* (XM_036513028.1) shows that they diverge just before the stop codon, so the 3' UTRs are unrelated. Our anchor-targets from *O. bimaculoides* map only to *O. sinensis* Upf2; neither the anchor-targets nor the *O. sinensis* 3' UTR match anywhere in the *O. bimaculoides* genome. The alignment also shows the downstream portion of the *O. sinensis* 3' UTR where the anchor-targets map. The targets differ in the number of tandem repeats of CAG: target 1 and 2 have six and five repeats, respectively. Target 1 is expressed in dissociated cells from sucker rims, and in olfactory organ; the other tissues express target 2.

C. **Netrin receptor/DCC.** The alignment of the *O. bimaculoides* genome (gene LOC106883766) and *O. sinensis* mRNA (XM_036508072.1) shows that the two diverge shortly after the stop codon. The *O. bimaculoides* gene ends in dinucleotide repeats just before the genome becomes a run of Ns (in red). Our anchor-targets from *O. bimaculoides* map only to *O. sinensis* netrin receptor 3' UTR (also shown in the alignment); neither the anchor-targets nor the *O. sinensis* 3' UTR match anywhere in the *O. bimaculoides* genome. The targets differ at a single nucleotide: target 1 and 2 have G and A, respectively; *O. sinensis* has a G in this position. If the *O. bimaculoides* genome encodes A, then target 1 is consistent with A-to-I RNA editing (inosine read as G during reverse transcription). The majority of tissues express target 2 only, while target 1 is only expressed in dissociated cells of sucker rims.

Figure S7. Diatom anchors in eelgrass samples show variation with location/season or Day vs Night.

A. **HMG (high mobility group) box domain.** The two targets show several nucleotide differences that result in two coding differences. The translation of the consensus sequence has its best two protein BLAST matches to HMG box proteins from diatom species, shown in the inset. Its best Pfam match is also HMG_box. Target 1 is found only in Norway/December samples, while target 2 is found only in France/June samples; both targets are found in both Day and Night samples.

B. **Ferredoxin.** The two targets show a silent single nucleotide polymorphism. The translation of the consensus sequence has its best protein BLAST matches to ferredoxin from several diatom species, the top two are shown in the inset. Its best Pfam match is also ferredoxin (Fer2 = 2Fe-2S iron-sulfur cluster binding domain). Target 1 is found only in France/June samples, while target 2 is found only in Norway/December samples.

C. **Fucoxanthin chlorophyll a/c protein (FCP).** This anchor and its targets are also presented in Figure 5C. At left, the translation of the consensus sequence has its best protein BLAST matches to several diatom species, two that are named as FCP are shown in the inset. The amino acid identity for *Phaeodactylum tricornutum* is 42/44 (95%). The consensus also BLASTs to the *P. tricornutum* genome, nucleotide identity 107/132 (81%) (not shown); this level of mismatch suggests the true origin species is not in the NCBI database. At right, histogram shows total anchor counts for Night are ~60% lower than for Day, indicating circadian regulation of this gene. All are samples from France in December (the only location/season in which this anchor had both Day and Night representation).

Note S1. Generality of SPLASH.

In this work we focused our experimental results on identifying changes in viral strains and specific examples of RNA-seq analysis. SPLASH'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, SPLASH provides an efficient and general solution to disparate problems in genomics. We outline examples of SPLASH'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 e.g. BCR-ABL fusions
    - 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 SPLASH anchor, target and consensus construction

SPLASH can function on any biological sequence and does not need anchor-target pairs to take the form of gapped *k*-mers, and can take very general forms. For example, one could consider schemes that respect triplet codons: $[X_1X_2Y_1][X_3X_4Y_2][X_5X_6Y_3]$... where X_i are bases in the anchor and Y_i are bases in the target, this would focus specifically on variation in the wobble position, the fastest to diverge; similar schemes might be appropriate for mechanisms with known patterns of diversity, such as diversity generating retroelements[79]. X and Y could also be amino acid sequences or other discrete variables considered in molecular biology. SPLASH 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. SPLASH can also be further developed to analyze higher dimensional relationships between anchors, where statistical 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.

Note S2. SPLASH statistical inference.

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*, *R* 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 [80]. The field's primary focus has been on either small contingency tables (2×2, e.g. Fisher's exact test [81]), high counts settings where a chi-square test yields asymptotically valid *p*-values, or computationally intensive Markov-Chain Monte-Carlo (MCMC) methods. While contingency tables have been widely analyzed in the statistics community [80,82,83], to our knowledge no existing tests provide closed form, finite-sample valid statistical inference with desired power for the application 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 SPLASH 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

The test statistic used by SPLASH can be thought of as using a vector *f* to partition the rows (targets) of the contingency table into two groups, assigning them 1 and 0 respectively. Then, for each sample, we compute the expected value of its target (i.e. what fraction of the targets from this sample were assigned a 1). We construct a per sample score as the difference between its target expected value and the global target expected value (with respect to the target distribution across all samples), and scale this difference by the square root of the number of observations of this anchor. A vector *c* assigns each sample a scalar value, and the final test statistic is computed as the *c*-weighted sum of these per-sample scores. Due to the linear structure of the test statistic, finite-sample *p*-value bounds are possible through classical concentration inequalities.

For fixed vectors *f* and *c*, there are many alternatives that SPLASH would not have power against. Thus, to detect different alternatives, SPLASH utilizes several randomly chosen *f* and *c*, applying Bonferroni correction to the result. In subsequent work, more sophisticated (optimization-based) approaches to computing improved *f* and *c* have been developed, leveraging a linear algebraic perspective on the test statistic. Additional generalizations of *f* and *c* may be of interest.

p-value computation

SPLASH'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 false discovery rate (FDR) controlled *q*-values by classical concentration inequalities and multiple hypothesis correction (Figure 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.

SPLASH's test statistic poses many exciting research directions, which we explore in a separate statistical work[55]. 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

SPLASH 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 $c_j = \pm 1$ 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, SPLASH 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 SPLASH assigns $c_j = \pm 1$ independently and with equal probability for each sample, computes the test statistic for each split, and selects the split yielding the smallest p-value.

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

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

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

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

Following this approach, we have that

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

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 SPLASH's finite-sample valid *p*-values as

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

q-value computation

q-values are computed using Benjamini Yekutieli correction [58] as

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

which enables SPLASH to control the FDR of the reported significant anchors.

Effect size

SPLASH 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 SPLASH *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 $c_j = \pm 1$), 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 $c_j = \pm 1$ and {0,1} valued *f*, this is simply a projection of the $T \times p$ table to a 2×2 table. Even considering more general *f*, there is an easy to understand alternative that SPLASH 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 (Figure 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 ℓ_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.

Note S3. SPLASH is robust to parameter choices and effective without metadata.

SPLASH is robust to parameter choices

We give examples of how choices of *k*, *R*, and tiling length impact results in France SARS-CoV-2 data as follows, showing that SPLASH yields similar results for a range of parameter choices. Default parameters shown in bold: we tested *k* = [25, **27**, 30]; Tile = [3, **5**, 7]; Lookahead = [0, 15, **23**]. For *k* = 25, 94.4% of anchors with default parameters contain at least one of the K=25 anchors as a substring. For *k* = 30, 93.8% of anchors with *k* = 30 contain at least one of the anchors with default parameters a substring. For tile size of 3, 85% of the anchors from the default run can be found in the significant anchors of tile size of 3. For tile size of 7, 85% of the anchors from the default run can be found in the significant anchors of tile size of 5. For lookahead distance of 0, 37% of the anchors from the default run can be found in the significant anchors of tile size of 3; for lookahead distance of 15, 76% of the anchors from the default run can be found in the significant anchors. Overall, as tile size decreases, anchor calls increase (4715, 5522, 7891 for [7, 5, 3] respectively). As *k* varies, anchor calls stay essentially the same (5875, 5522, and 5958 for *k* = [25, 27, 30] respectively). Finally, for lookahead distance, the total number of calls decrease as lookahead distance increases (13239, 8295, 5522 for *R* = [0, 15, 23] respectively).

SPLASH is effective without metadata

As discussed, SPLASH can be run without any metadata. For the HLCA dataset, when run on the two donors without metadata, SPLASH 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, SPLASH's metadata free approach calls >94% of the genes called by the metadata-based approach (Figure S3). For donor 2, SPLASH 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, SPLASH'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.

Note S4. Lemur lambda light chain and surrogate light chains found by SPLASH.

As detailed in Methods, we attempted to identify light chain variable regions in lemur B cells where the BASIC pipeline[38] could not. We were successful in identifying a full variable region by extension (using `grep` of raw reads) of an initial seed consensus sequence found by SPLASH. Below we give the sequence identified, and the NCBI IgBlast report for it. IgBlast uses human Ig reference sequences (lemur Ig regions are not well annotated) so it is uncertain which differences are due to lemur and which to hypermutation, however there appears to be a high mutational load in this variable region, which may be why BASIC could not identify it.

This is the cell that had a full V-region:

cell: MAA001400_B109012_I1_S193

```
>cons1-MAA001475_B112525_O18_S354_R1_001
GGCCTTGGGCTGACCTAGGACGGTGAGCTGGGTCCCTCTGCCGAAGACAAACATCGACTGAGGCTCAGACCAA

>cell_1_lambda_VJ_assembled
GTGAGTCCCCAGGAACCAGAGCTCACAGGAGCCTCCACCATGGCCAGGGCTCCTGTCCTC
GTCCCTCTCCTCGCTCTCTGCTCAGGGTCCTGGGCACAGTCTGGACTCACCCAGGAAGCC
TTGGTGTCAGGGTCTGTGGGACACAAGGTCACCCTGTCCTGCGCTGGACACAGCAACAGT
GTTGGTTCATTTGGGGTGGACTGGTGCCAGCAGGTTCCTGGTGGTGCCCCCAAAACTGTG
ATGCTCGGGACAACTCGGCCCTCAGGGATCCCCGATCGCTTCTCCGGCTCCAGGTCTGGG
AACACGGCCTCTTTGACCATCTCGGACCTCCCGGACCTCCAGCCGGAGGACGAGGCTGAC
TATTACTGTTCAACTTGGGACAGAACCCTGCGTGCTCATGTGTTCGGCGGTGGGACCAAG
GTGATCGTCGTAGGTCAGCCCAAGGCCGCCCCCTCGGTCACGCTGTTCCCGCCCTCCTC
```

Protein translation:

```
>cell_1_lambda_VJ_assembled
MARAPVLVPLLALCSGSWAQSGLTQEALVSGSVGHKVTLSCAGHSNSVGSFGVDWCQQVP
GGAPKTVMLGTTRPSGIPDRFSGSRSGNTASLTISDLPDLQPEDEADYYCSTWDRTLRAH
VFGGGTKVIVVGQPKAAPSVTLFPPS
```

Below is the NCBI IgBlast report for the full V-region (FASTA sequence given at the end.) Coding differences from germline gene are shown in magenta.

Top V gene match	Top J gene match	Chain type	stop codon	V-J frame	Productive	Strand	V frame shift
IGLV2-14*05	IGLJ6*01	VL	No	In-frame	Yes	+	No

V region end	V-J junction	J region start
TACTG	TTCAACTTGGGACAGAACCCTGCGTGCTC	ATGTG

```
                                      <--------------------------------FR1-IMGT---------------------------------><----------CDR1
                                       Q  S  G  L  T  Q  E  A  L  V  S  G  S  V  G  H  K  V  T  L  S  C  A  G  H  S     N  S  V
                    Query_1      97   CAGTCTGGACTCACCCAGGAAGCCTTGGTGTCAGGGTCTGTGGGACACAAGGTCACCCTGTCCTGCGCTGGACACAGCA---ACAGTGTT  183
V  69.1% (192/278)  IGLV2-14*05  1    .......CC..G..T...CCT....CC.....T......CCT.....GTC.A.....A.C......A.....AC.....GTG..GT..G.  90
                                       Q  S  A  L  T  Q  P  A  S  V  S  G  S  P  G  Q  S  I  T  I  S  C  T  G  T  S  S  D  V  G
V  68.7% (191/278)  IGLV2-18*01  1    .......CC..G..T...CCTC...CC.....C......CCT.....GTCA......A.C......A.....AC.....---GTGAC...  87
V  68.7% (191/278)  IGLV2-18*02  1    .......CC..G..T...CCTC...CC.....C......CCT.....GTCA......A.C......A.....AC.....---GTGAC...  87

                                      -IMGT---------><--------------------FR2-IMGT---------------------><CDR2-IM><--------------
                                       G  S  F     G  V  D  W  C  Q  Q  V  P  G  G  A  P  K  T  V  M  L  G     T  T  R  P  S  G
                    Query_1      184  GGTTCATTTG---GGGTGGACTGGTGCCAGCAGGTTCCTGGTGGTGCCCCCAAAACTGTGATGCTCGGG---ACAACTCGGCCCTCAGGG  267
V  69.1% (192/278)  IGLV2-14*05  91   ....ATAACT---AT..CTC.....A...A...CAC..A..CAAA.........CTCA....TTAT.A.GTC.GT.A.............  177
                                       G  Y  N     Y  V  S  W  Y  Q  Q  H  P  G  K  A  P  K  L  M  I  Y  E  V  S  N  R  P  S  G
V  68.7% (191/278)  IGLV2-18*01  88   ...AGT.A.AACC.T..CTC.....A.......CCC..A..CACA.........CTCA....TTAT.A.GTC.GT.A.............  177
V  68.7% (191/278)  IGLV2-18*02  88   ...AGT.A.AACC.T..CTC.....A.......CCC..A..CACA.........CTCA....TTAT.A.GTC.GT.A.............  177

                                      ---------------------------------------FR3-IMGT-------------------------------------------
                                       I  P  D  R  F  S  G  S  R  S  G  N  T  A  S  L  T  I  S  D  L  P  D  L  Q  P  E  D  E  A
                    Query_1      268  ATCCCCGATCGCTTCTCCGGCTCCAGGTCTGGGAACACGGCCTCTTTGACCATCTCGGACCTCCCGGACCTCCAGCCGGAGGACGAGGCT  357
V  69.1% (192/278)  IGLV2-14*05  178  G....T...........T.......A......C...........CC.........---------.T.GG......G.T............  258
                                       V  P  D  R  F  S  G  S  K  S  G  N  T  A  S  L  T  I           S  G  L  Q  A  E  D  E  A
V  68.7% (191/278)  IGLV2-18*01  178  G....T...........T..G....A......C...........CC.........---------.T.GG......G.T............  258
V  68.7% (191/278)  IGLV2-18*02  178  G....T...........T..G....A......C...........CC.........---------.T.GG......G.T............  258

                                      -----------><-----------CDR3-IMGT-----------><--------FR4-IMGT--------->
                                       D  Y  Y  C  S  T  W  D  R  T  L  R  A  H  V  F  G  G  G  T  K  V  I  V  V
                    Query_1      358  GACTATTACTGTTCAACTTGGGACAGAACCCTGCGTGCTCATGTGTTCGGCGGTGGGACCAAGGTGATCGTCGTAG  433
V  69.1% (192/278)  IGLV2-14*05  259  ..T........-----------------------------------------------------------------  269
                                       D  Y  Y
V  68.7% (191/278)  IGLV2-18*01  259  ..T........-----------------------------------------------------------------  269
V  68.7% (191/278)  IGLV2-18*02  259  ..T........-----------------------------------------------------------------  269
J  90.6% (29/32)    IGLJ6*01     3    ----------------------------------------...........A....C..........C....----  34
J  88.2% (30/34)    IGLJ3*02     5    ------------------------------------------...........A.........C...C....C...  38
J  87.1% (27/31)    IGLJ2*01     8    ---------------------------------------------........A.........C...C....C...  38
```

Here are the BLAST matches of consensuses from two cells to the unique region of IGLL1/5, one of the surrogate light chains. This suggests that these two cells have not yet rearranged their light chain.

cell: MAA001475_B112525_C4_S52

```
>cons1-MAA001475_B112525_C4_S52_R1_001
GGCCTTGGGCTGACCTGCTCCACGGGATCCGCGGCACTGGACCGGCTGCTTCCCGCCCGGGCTCCGGGGCC
```

Consensus sequence above is the original strand, but BLAST was done with its reverse-complement, the sense strand (this is the Query). The alignment indicates a 1-nt insertion in the consensus relative to IGLL5 (which would cause a frameshift).

```
PREDICTED: Microcebus murinus immunoglobulin lambda-like polypeptide 5 (LOC105882024),
transcript variant X2, mRNA
Sequence ID: XM_012784097.2   Length: 821   Number of Matches: 1   Range 1: 193 to 250
Score         Expect  Identities    Gaps       Strand
91.6 bits(49) 1e-14   56/59(95%)    1/59(1%)   Plus/Plus
             G  P  G  A  R  A  G  S  S  R  S  S  A  A  D  P  V  E  Q
              A  P  E  P  G  R  E  A  A  G  P  V  P  R  I  P  W  S  R
Query  1    GGCCCCGGAGCCCGGGCGGGAAGCAGCCGGTCCAGTGCCGCGGATCCCGTGGAGCAGGT  59
            ||||| |||||||||||||||||||||||||||||  ||||||||||||||||||||||
Sbjct  193  GGCCCTGGAGCCCGGGCGGGAAGCAGCCGGTCCAG-CCCGCGGATCCCGTGGAGCAGGT  250
             G  P  G  A  R  A  G  S  S  R  S  S   P  R  I  P  W  S  R
```

Protein comparison of the mouse lemur IGLL5 above to human IGLL5 (where domains are identified) shows that the above region clearly lies in the IGLL5 N-terminal unique region. The matching part above is shown in blue.

```
Query  12   LRLGKGQVGCDAPK--GPGPRLRWPLLLLGLAVGTHGFLSSTEAPRSRAPGPGARAGSSR  69
            +R   GQVGC+ P+  GPGPR RWPLLLLGLA+  HG L    AP+S  P PGA  GSSR
Sbjct  1    MRPKTGQVGCETPEELGPGPRQRWPLLLLGLAMVAHGLLRPMVAPQSGDPDPGASVGSSR  60

Query  70   SSPRIPWSRFLLQPSPRGAGARCWPRGFWSEPQSLWYIFGRGTQLTILGQPKAAPSVTLF  129
            SS R  W R LLQPSP+ A  RCWPRGFWSEPQSL Y+FG GT++T+LGQPKA P+VTLF
Sbjct  61   SSLRSLWGRLLLQPSPQRADPRCWPRGFWSEPQSLCYVFGTGTKVTVLGQPKANPTVTLF  120
                                                 J-region         C-region

Query  130  PPSSEELQANKATLVCLMSDFYPGAVSVAWKADGSAVTQGVETTQASKQSNGKYAASSYL  189
            PPSSEELQANKATLVCL+SDFYPGAV+VAWKADGS V  GVETT+ SKQSN KYAASSYL
Sbjct  121  PPSSEELQANKATLVCLISDFYPGAVTVAWKADGSPVKAGVETTKPSKQSNNKYAASSYL  180

Query  190  SLSPAQWKAGGRFSCQVTHEGSTVEKTVAPAECA  223
            SL+P QWK+   +SCQVTHEGSTVEKTVAP EC+
Sbjct  181  SLTPEQWKSHRSYSCQVTHEGSTVEKTVAPTECS  214
```

cell: MAA001475_B112525_O18_S354

```
>cons1-MAA001475_B112525_O18_S354_R1_001
GGCCTTGGGCTGACCTAGGACGGTGAGCTGGGTCCCTCTGCCGAAGACAAACATCGACTGAGGCTCAGACCAA
```

Consensus sequence above is the original strand, but BLAST was done with its reverse-complement, the sense strand (this is the Query). Here the match is to human

IGLL5. The consensus includes regions upstream of the J-region, so part of the IGLL5-unique region.

```
Homo sapiens immunoglobulin lambda like polypeptide 5 (IGLL5), transcript variant 1, mRNA
Sequence ID: NM_001178126.2   Length: 1300   Number of Matches: 1  Range 1: 502 to 577
Alignment statistics for match #1 Score     Expect Identities     Gaps    Strand
86.9 bits(95) 5e-13 65/76(86%)       3/76(3%)       Plus/Plus
              W  S  E  P  Q  S  M  F  -  V  F  G  R  G  T  Q  L  T  V  L  G  Q  P  K  A
Query  1    TTGGTCTGAGCCTCAGTCGATGTTT---GTCTTCGGCAGAGGGACCCAGCTCACCGTCCTAGGTCAGCCCAAGGCC  73
            ||||||||||||||||||  ||| |   |||||||| |  |||||| || ||||||||||||||||||||||||||
Sbjct  502  TTGGTCTGAGCCTCAGTCACTGTGTTATGTCTTCGGAACTGGGACCAAGGTCACCGTCCTAGGTCAGCCCAAGGCC  577
              W  S  E  P  Q  S  L  C  Y  V  F  G  T  G  T  K  V  T  V  L  G  Q  P  K  A
                                        J-region
```

Note S5. Octopus and eelgrass analyses, additional notes.

Octopus bimaculoides Myo-VIIa

We give more detailed alignment information for Myo-VIIa anchor and targets, to document the unannotated alternative first exon “1b” (that is expressed specifically in statocyst tissue) and the absence of sequence matching exon 2 in the *O. bimaculoides* genome. Note that we do not have information on the start of exon “1b”. We use reverse-complements of the anchor and targets (the sense strand).

Alignment to O. sinensis genome.

```
<<< start of first exon (not shown) = 79,843,996
...CATCCCGGATTTTCTACTCAATTCTacCGTct-TGTTGCCGGCATGCCTTATAACTTGTGgtaggta annotated first exon = "1a"
                 |                                            |
                 79,844,331                                   79,844,375
                 |                                            |
                 |                            MetProTyrAsnLeu***                             MetV
consensus_1      TACTCAATTCTGGCGTTTCTGTTGCCGGCATGCCTTATAACTTGTGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                                                              :::::::::::::::::::::::::::::::::::
consensus_2      TTTATTATCACCAATATGGACGGAAATAGTGTATCCATTTATTAAGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                 |              MetAspGlyAsnSerValSerIleTyr***||                                |
                 |                                            ||                                |
                 |                                            |79,932,654              79,932,687    79,932,701
                 |                                            ||                                |             |
                 |                         cctttattttctatttcagATATAATTGGATTTTAAACAAAAAGCAAAAATGGTGATTCTTGCAAAGgtaa
                 |                                            |   ***LeuAspPheLysGlnLysAlaLysMetValIleLeuAlaLys
                 79,871,237                                   79,871,281                       annotated exon 2
                 |                                            |
...AATGACTCACTTTATTTATTATCACCAATATGGACGGAAATAGTGTtTCCATaTATTAAGgtaagc  unannotated alternative first exon "1b"
    MetThrHisPheIleTyrTyrHisGlnTyrGlyArgLys***                         (expressed in statocyst tissue)
                                MetAspGlyAsnSerValSerIleTyr***
```

anchor is in blue; target 1 in green; target 2 in red.

brown = genomic sequence from *O. sinensis*, showing splice signals. Splice dinucleotides and ATGs underlined.

lowercase-orange = SNP differences in *O. sinensis* from anchor-targets (*O. bimaculoides*)

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.

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

Alignment to *O. bimaculoides* genome.

There is no full perfect match for the relevant exon 2 portion nor for the anchor (underlined sequence) (ATATAATTGGATTTTAAACAAAAAGCAAAAATGG) in *O. bimaculoides*. Sequences matching exons 1a and 1b are present in the *O. bimaculoides* genome. The splice donor for *O. sinensis* myoVIIa exon 1a matches perfectly to an *O. bimaculoides* splice donor in an annotated noncoding RNA XR_008264717.1 (gene LOC128248543) located upstream of the annotated *O. bimaculoides* myoVIIa transcript.

```
9,633,631        9,633,645                                    9,633,690
   |             |                                            |
...CATCCCGGATTTTCTACTCAATTCTGGCGTTTCTGTTGCCGGCATGCCTTATAACTTGTGgtaagtat O.bimaculoides splice donor, XR_008264717.1
...CATCCCGGATTTTCTACTCAATTCTacCGTct-TGTTGCCGGCATGCCTTATAACTTGTGgtaggta  O.sinensis annotated first exon = "1a"
                 |                                            |
                 |                            MetProTyrAsnLeu***                             MetV
consensus_1      TACTCAATTCTGGCGTTTCTGTTGCCGGCATGCCTTATAACTTGTGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                                                              :::::::::::::::::::::::::::::::::::
consensus_2      TTTATTATCACCAATATGGACGGAAATAGTGTATCCATTTATTAAGATATAATTGGATTTTAAACAAAAAGCAAAAATGG
                 |              MetAspGlyAsnSerValSerIleTyr***||
                 |                                            |exon 2 no match in O. bimaculoides
                 |                                            |
                 9,658,490                                    9,658,536
                 |                                            |
...AATGACTCACTTTATTTATTATCACCAATATGGACGGAAATAGTGTATCCATTTATTAAGgtaagc  unannotated alternative first exon "1b"
    MetThrHisPheIleTyrTyrHisGlnTyrGlyArgLys***                         (expressed in statocyst tissue)
                                MetAspGlyAsnSerValSerIleTyr***
```

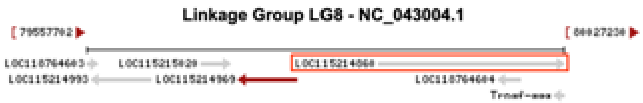
Numbers are genome coordinates for *O. bimaculoides* chromosome 8 = NC_068988.1 "Octopus bimaculoides isolate UCB-OBI-ISO-001 chromosome 8, ASM119413v2, whole genome shotgun sequence". The *O. bimaculoides* myoVIIa gene model is LOC106880717.

There appears to be an assembly issue for the *O. sinensis* Myo-VIIa gene on chromosome LG8 that matches our anchor-targets. There are several *O. sinensis* genes annotated with myosin-VIIa in the genome. We list them below, including their protein domain content.

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, 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.

Neither of the two myo-VIIa genes on Ch. LG8 has a full complement of protein domains: LOC115214860 has the N-terminal myosin motor domain, but lacks the tail domains. LOC115214969 has all the tail domains but lacks the motor domain. The two genes are adjacent on the chromosome, but are in head-to-head orientation, as shown in this graphic screenshot from NCBI (LOC115214860 is the red arrow, LOC115214969 has been marked with a red box):



If LOC115214860 was inverted, then all domains would be present in the correct linear order.

Zostera marina (eelgrass) NADPH quinone oxidoreductase subunit L (NdhL) intron retention

We have confirmed the intron retention event by sequence extension of target 4 from raw reads to reach the end of exon 2 (data not shown).

We show here the predicted translation of the intron retention isoform of NdhL (target 4 of Figure 5D). It causes a frameshift and termination shortly after the end of exon 3 (where the other targets are located). The anchor is in exon 4.

anchor is in blue. target4 is in red. target1 is in green, target2 and target3 are aligned underneath target1, differences in magenta.

Intron sequence is shown in lowercase, splice dinucleotides underlined. Reverse-complements of anchor-targets are shown.

```
>LFYR01000468.1:167070-167279 Zostera marina strain Finnish scaffold_137, whole genome shotgun
sequence
GCAAGATTTGACTTCAGTGCTTATACAATCAGGAGCATTTGCTTTCTTCTACTTCCTTATCATGCCGgta
                         ...SerGlyAlaPheAlaPhePheTyrPheLeuIleMetProVal
target2                                TGTTTTCTTCTACTTCCTTATCATGCC
                                        ValPhePheTyrPheLeuIleMetPro
target3                                TGCTTTCTTCTACTTCCTTGTCATGCC
                                        AlaPhePheTyrPheLeuValMetPro

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

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

The frameshift and termination occur within the second transmembrane domain of the protein. (There is a third transmembrane domain, but it is not predicted by all programs.) The topology below was predicted by CCTOP (https://cctop.ttk.hu/)

blue = cytoplasmic; **red = transmembrane;** green = extracellular. The InterPro NdhL domain is underlined.

```
>full-length_NdhL
MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLHLGVLLTSIA
DEPAFAVTGSNNYEQDLTSVLIQSGAFAFFYFLIMPPIIMNWLRLRWYKRKLFETYLQFMFVFLFFPGILLWAPFINFR
RLPRDPTMKHPWSTPRDSST
>intron-retention_NdhL
MTHLLLPLPSKVTGAFNHREWSCHRVPHPVSSAQRTRPLISASISKTKKINGRLMCNIESSKATNSTLLHLGVLLTSIA
DEPAFAVTGSNNYEQDLTSVLIQSGAFAFFYFLIMPVYNCKVILT
```

Note S6. SPLASH runs on a laptop.

Computational benchmarking for SPLASH

SPLASH is computationally much more efficient than other approaches, due to its use of k-mers rather than reference alignment, and its closed-form statistics obviating compute-intensive significance testing. SPLASH is implemented as a fully containerized and parallelized workflow that requires only the FASTQ read files and no parameter tuning by the user. We ran SPLASH on a 2015 Intel laptop with an Intel® Core™ i7-6500U CPU @ 2.50GHz processor, generating significance calls for single cell RNA-seq totaling over 10 million reads in only 1 hour 45 min. When performed on a compute cluster, the same analysis is completed in an average of 22.8 minutes with 750 MB of memory for 10 million reads.

Timing for SS2

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

Laptop analysis details

Laptop specs:
An Intel® Core™ i7-6500U CPU @ 2.50GHz (launched in 2015)
2 cores, total of 4 threads, 3 of which SPLASH was allowed to use.
8 GB DDR3 RAM
SODIMM DDR3 Synchronous 1600 MHz (0.6 ns)

Laptop analysis dataset:

Ten B and T cells from donor 2 blood sequenced by Smart-Seq2 were used for the laptop benchmarking. These files totalled 43,870,027 reads, averaging 4.3M reads per cell. The fastq files for the Tabula Sapiens data were downloaded from https://tabula-sapiens-portal.ds.czbiohub.org/. Files used:
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A13_S73_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A18_S78_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A19_S79_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A21_S81_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A3_S63_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A5_S65_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A6_S66_R1_001.fastq.gz

TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A8_S68_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_A9_S69_R1_001.fastq.gz
TSP2_Blood_NA_SS2_B114581_B133053_Lymphocytes_B10_S94_R1_001.fastq.gz

Note S7. Anchor and target sequences, *q*-values, and binomial *p*-values.

Binomial *p*-value bound computation for plots depicting target fraction abundance

We provide *p*-values to quantify the visually striking nature of the plots depicting fraction abundance per a specific target (target 1 by default). 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 SPLASH 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 Bin(m,q)≥ X, where with m = number of samples and q = .02.

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 q_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 Bin(m,q')≥X, where m = number of samples with q'=max$_j$ q_j , where $q' \leq$.02.

Anchor and target sequences, with *q*-values and binomial *p*-values

Targets are numbered by decreasing abundance, unless otherwise stated.

q-values are the BY-corrected *p*-values output by SPLASH, as detailed in Note S2. Binomial *p*-value calculations are described above, and are with respect to target 1, unless otherwise stated.

SARS-CoV-2 mutation K417N (Figure 2A)

q-value: 9.4e-05
binomial *p*-value: 6.4e-07

```
>anchor
ATTCATTTGTAATTAGAGGTGATGAAG
>target_1_Delta
ACTGGAAAGATTGCTGATTATAATTAT
>target_2_K417N_Omicron
ACTGGAAATATTGCTGATTATAATTAT
```

SARS-CoV-2 mutations V213G, NL211I, R214REPE (Figure 2B)

q-value: 8.3e-08
binomial *p*-value: 1e-13

```
>anchor
TTTAAGAATATTGATGGTTATTTTAAA
>target_1_Delta
TAATTTAGTGCGTGATCTCCCTCAGGG
>target_2_V213G_BA.2
TAATTTAGGGCGTGATCTCCCTCAGGG
>target_3_NL211I-R214REPE_BA.1
TATAGTGCGTGAGCCAGAAGATCTCCC
```

SARS-CoV-2 mutations P681R, N679K, P681H (Figure 2C)

q-value: 1.2e-04
binomial *p*-value: 4.9e-12
(reverse-complements are shown in Figure 1C)

```
>anchor
GTGACATAGTGTAGGCAATGATGGATT
>target_1_P681R_Delta (abundance order = 1)
CGACGAGAATTAGTCTGAGTCTGATAA
>target_2_P681R-Q677H (abundance order = 3)
CGACGAGAATTAGTATGAGTCTGATAA
>target_3_P681R-Q677H (abundance order = 4)
CGACGAGAATTAGTGTGAGTCTGATAA
>target_4_N679K-P681H_Omicron (abundance order = 2)
CGATGAGACTTAGTCTGAGTCTGATAA
```

MYL12A / MYL12B (Figure 3A, S4B)

P2 *q*-value: 2.5e-08

P2 binomial *p*-value: 9.9e-37 (with respect to target 2)
P3 *q*-value: 2.3E-42
P3 binomial *p*-value: 2.2e-45 (with respect to target 1)

```
>P2_anchor
AAGAGGCCTTCAACATGATTGATCAGA
>P2_target_2_MYL12A
TTCATTGGGGAAGAATCCAACTGATGA
>P2_target_1_MYL12B
TTCTCTAGGGAAGAATCCCACTGATGC
>P2_consensus_MYL12A_macrophage
ACAGAGATGGTTTCATCGACAAGGAAGATTTGCATGATATGCTTGCTTCATTGGGGAAGAATCCAACTG
ATGAGTATCTAGATGCCATGATGAATGAGGCTCCAGGCCCCATCAATT
>P2_consensus_MYL12B_capillary
ACAGAGATGGCTTCATCGACAAGGAAGATTTGCATGATATGCTTGCTTCTCTAGGGAAGAATCCCACTG
ATGCATACCTTGATGCCATGATGAATGAGGCCCCAGGGCCCATCA
>P3_anchor
AAGAGGCCTTCAACATGATTGATCAGA
>P3_target_1_MYL12A
GAAGATTTGCATGATATGCTTGCTTCA
>P3_target_2_MYL12B
GAAGATTTGCATGATATGCTTGCTTCT
>p3_consensus_MYL12A_macrophage
ACAGAGATGGTTTCATCGACAAGGAAGATTTGCATGATATGCTTGCTTCATTGGGGAAGAATCCAACTG
ATGAGTATCTAGATGCCATGATGAATGAGGCTCCAGGCC
>p3_cons_MYL12B_capillary
ACAGAGATGGCTTCATCGACAAGGAAGATTTGCATGATATGCTTGCTTCTCTAGGGAAGAATCCCACTG
ATGCATACCTTGATGCCATGATGAATGAGGCCCCAGGGCCCATCAATTT
```

HLA-DRB1 / HLA-DRB4 (Figure 3B)
P2 *q*-value: 4.0e-10
P2 binomial *p*-value: 2e-17
P3 *q*-value: 1.2e-4
P2 binomial *p*-value: 1.6e-08
(reverse-complements are shown in Figure 3B)

```
>P2_anchor
GGAAGCCACAAGGGAGGACATTTTCTG
>P2_target_1_DRB1
GTGGAAGAATAACTGCCAAGCAGGAAA
>P2_target_2_DRB4
GGAAGAATAAGAGCCAAGTGGGAAAGC
>P2_consensus_DRB1_macrophage
GGAAGCCACAAGGGAGGACATTTTCTG
CAGTTGCCGAACCAGTAGCAACCAGGTCCTGAGAAAGCCCTCTCTTGTGGAAGAATAACTGCCAAGCAG
GAAAGCTTTTCATTCTGCAAAGCTGGGACAGAAGGTTCTTCCTTGAATGT
```

```
>P2_consensus_DRB4_capillary
CAGAGTTGCTGAACCAGTAACAACCTGGTCCTGACAAAGCTCTTGTGGAAGAATAAGAGCCAAGTGGGA
AAGCTTTTCATCTTGCAAAGCTGGGGCAGAAGGTTCTTCCTTGAATGT
>P3_anchor   (same sequence as P2_anchor)
GGAAGCCACAAGGGAGGACATTTTCTG
>P3_target_1_DRB1
AGGTCCTGAGAAAGCCCTCTCTTGTGG
>P3_target_3_DRB4
CCTGGTCCTGACAAAGCTCTTGTGGAA
>P3_consensus_DRB1_macrophage
CAGTTGCTGAACCAGTAGCAACCAGGTCCTGAGAAAGCCCTCTCTTGTGGAAGAATAACAGCCAGGAGG
GAAAGCTTTTCATCCTGCAAAGCTGGGGCAGAAAGTTCTTCT
>P3_consensus_DRB4_capillary
GGAAGCCACAAGGGAGGACATTTTCTG
CAGAGTTGCTGAACCAGTAACAACCTGGTCCTGACAAAGCTCTTGTGGAAGAATAAGAGCCAAGTGGGA
AAGCCTTTCATCTTGCAAAGCTGGGGCAGAAGGTTCTTCCTTGA
```

HLA-DPA1 / HLA-DPB1 (Figure 3C, S4C)

P3 *q*-value: 7.9e-22

P3 binomial *p*-value: 9.15e-18

(anchor as given here is sense strand for DPA1, antisense strand for DPB1)

```
>P3_anchor
AGATGTATCTCTCCAGGAAGCGCTGTG
>P3_target_1_DPA1
TGCCGTCCCTGGAAAAGGTGAATCCCA
>P3_target_2_DPB1
TGCCGTCCCTGGAAAAGGTAATTCTCT
>P3_consensus_DPB1_macrophage
TCCCATTAAACGCGTAGCATTCCTGCCGTCCCTGGAAAAGGTAATTCTCTGGAGTGGCCCTGCCCTGGA
CCACAGATGTGAGCAGCACCATCAGTAACGCCGTCAGAGCCACT
>P3_consensus_DPA1_capillary
TCCCATTAAACGCGTAGCATTCCTGCCGTCCCTGGAAAAGGTGAATCCCAGCCATGCTGATTCCTCTCC
ACCCATTTCCAGTGCTAGAGGCCCACAGTTTCAGTCTCATCTGC
```

HLA-B (Figure 3D, S4D)

q-value: 2.7e-05

binomial *p*-value: 1.7e-25

```
>anchor
TTGGGACCGGAACACACAGATCTTCAA
>target_1_HLA-B
AGAGCCTGCGGAACCTGCGCGGCTACT
>target_2_HLA-B
AGAACCTGCGGATCGCGCTCCGCTACT
>consensus_1_HLA-B
```

```
TTGGGACCGGAACACACAGATCTTCAAGACCAACACACAGACTGACCGAGAGAGCCTGCGGAACCTGCG
CGGCTACTACAACCAGAGCGAGGCCGGGTC
>consensus_2_HLA-B
TTGGGACCGGAACACACAGATCTTCAAGACCAACACACAGACTTACCGAGAGAACCTGCGGATCGCGCT
CCGCTACTACAACCAGAGCGAGGCCGGGTC
```

human Ig-kappa C-region (Figure 4B)

q-value = 1.6E-35

```
>anchor
TGGCGGGAAGATGAAGACAGATGGTGC
>Targ0
GCTTGGTCCCCTGGCCAAAAGTCCCGG
>Targ1
GCTTGGTCCCCTGGCCAAAAGGGCTAC
>Targ2
GCTTGGTCCCCTGGCCAAAAGTGTACG
>Targ3
CCTTGGTCCCTCCGCCGAAAGAAGGTG
>Targ4
GCTTGGTCCCCTGGCCAAAAGTGTCGT
>Targ5
GCTTGGTCCCCTGGCCAAAAGTGCCCG
>Targ6
CTTTGGTCCCAGGGCCGAAAGTGAATA
>Targ7
CCTTGGTCCCTTGGCCGAACGTCCACC
```

human TCR-alpha C-region (Figure 4B)

q-value = 3.4E-5

```
>anchor
GTACACGGCAGGGTCAGGGTTCTGGAT
>Targ1
TGCCTTTGCCGAAGTTGAGTGCATACC
>Targ2
TCCCTGATCCAAAGATTATCTTGGAAG
>Targ3
TGCCTGTCCCAAAGGTGAGTTTGTTTC
>Targ4
TCCCAGCGCCCCAGATTAACTGATAGT
>Targ5
TCCCCCTTGCAAAGAGCAGCTTCTGGC
>Targ6
TTCCTCCTCCAAAAGTTAGCTTGTTGC
>Targ7
```

```
TCCCTGTCCCAAAATAGAACTGGTTAC
>Targ8
TTCCTCTTCCAAAGTATAGCCTCCCCA
>Targ9
TTCCCTTTCCAAAGACCAGCTTTTCAG
>Targ10
TTCCCTGTCCGAAGATAAGCTTTCCTC
>Targ11
TCCCTGCTCCAAAGCGCATGTCATTGT
>Targ12
TTCCCTTCCCAAAGATCAGAGCAGTTC
>Targ13
TCCCAGATCCAAAGTAAAATTTGTTGA
>Targ14
TCCCTTGCCCAAAGATTAGTTTGCCTG
>Targ15
TTCCTCTTCCAAATGTAGGTATGTAGC
>Targ16
TTCCATCTCCAAACATGAGTCTGGCAT
>Targ17
TTCCACTCCCAAAAGTAAGTGCTCTCC
>Targ18
TTCCTTTTCCAAATGTCAGTTTATAGT
>Targ19
TGCCTGTTCCAAAGATGTATTTGTAGG
>Targ20
TTCCAGTTCCAAAGGTAACTTTCTGGT
>Targ21
TCCCTTGTCCAAATGTCAGCTTTCCAT
>Targ22
TCCCCTTCCCGAAAGTGAGTTGGTAAC
>Targ23
TGCCAGTTCCAAAGATGAGCTTGTTTG
```

lemur Ig-heavy V-region (Figure 4B)

q-value = 1.3E-11

```
>anchor
AGCCTGGGGGGTCCCTGAGACTCTCCT
>Targ0
AGTGACTACTACATGAGCTGGGTCCGC
>Targ1
AGCAGCTATGGGATGAACTGGGTCCGC
>Targ2
AGCAACTACTGGATGAGCTGGGTCCGC
>Targ3
```

```
AAGAACTATGAGATAAACTGGGTCCGC
>Targ4
AGCAGCTACTACATGCACTGGGTCCGC
>Targ5
AGCAGCTACGATATGAACTGGGTCCGA
>Targ6
AGTGACTACTACATGAACTGGGTCCGC
>Targ7
AGCAGCCATGGAATGCACTGGGTCCGC
>Targ8
AGCAGCTACGATATGAACTGGGTCCGC
>Targ9
AGCAGCTATGATATGCATTGGGTCCGC
>Targ10
AGTGACCACCACATGAGCTGGGTCCGC
>Targ11
GATGACTACCTCATGCACTGGATCCGC
>Targ12
AGCAGCTATGCCATGAGCTGGGTCCGC
>Targ13
AGTAGTTACTGGATGAACTGGGTCCGC
>Targ14
GATTACTATGGCATGAACTGGGTCCGC
>Targ15
ACCAATTTTGGGATGAACTGGGTCCGC
>Targ16
AGCAGCTATGGGATGCACTGGGTCCGC
>Targ17
ACCAGTTATGGGATGAACTGGGTCCGC
```

lemur TCR-alpha C-region (Figure 4B)

q-value = 4.1E-7

```
>anchor
TCAGCTGGTACACGGCGGGGTCAGGGT
>Targ0
AGTCTGGTCCCTGCTCCAAAGCGCAGA
>Targ1
AGCCTGGTCCCTGCTCCAAAAATCAAC
>Targ2
AGCAGAGTGCCAGTCCCAAAGATGAGC
>Targ3
ACGGTGGTTCCTTTCCCAAAGATCAAC
>Targ4
AGTTGGGTGCCAGTTCCAAACACGGGT
>Targ5
```

```
AACTGGGTCCCGGATCCAAAGGTCAGT
>Targ6
AGTTGTGTCCCTTTTCCAAAGGTGACT
>Targ7
AGTTTGGTCCCAGATCCAAAGTAAAAT
>Targ8
AATCTGGTCCCAGTCCCAAAGATGAGC
>Targ9
AGTCTGGTCCCTGATCCAAAGATTAGC
```

Octopus bimaculoides Myo-VIIa (Figure 5A)

q-value = 4.0e-03

(reverse-complements shown in Figure 5A)

```
>anchor
CCATTTTTGCTTTTTGTTTAAAATCCA
>target_1
ATTATATCACAAGTTATAAGGCATGCC
>target_2
ATTATATCTTAATAAATGGATACACTA
```

fucoxanthin chlorophyll a/c protein, diatom (Figure 5C)

q-value = 6.0e-08

(reverse-complements shown in Figure 5C)

```
>anchor
AAGTATCCAACAACGGCAAGCATGGAG
>target_1 (abundance order = 1)
ATACGTCCGTGCTTGAGCTCGACAAAT
>target_2 (abundance order = 6)
ATACGGCCGTGCTTGAGCTCGACAAAT
>target_3 (abundance order = 2)
ATACGTCCGTGCTTGATCTCGACGTAT
>target_4 (abundance order = 4)
ATACGTCCGTGCTTGATCTCAACGTAT
>target_5 (abundance order = 5)
ATACGTCCGTGCTTGATCTCGACGTAC
>target_6 (abundance order = 3)
ACACGTCCATGCTTAATTTCGACATAT
```

Zostera marina NADPH quinone oxidoreductase subunit L (NdhL) (Figure 5D)

q-value = 6.5e-56

(reverse-complements shown in Figure 5D)

```
>anchor
AATCGAAGCCAATTCATGATGATAGGC
>target1
```

```
GGCATGATAAGGAAGTAGAAGAAAGCA
>target2
GGCATGATAAGGAAGTAGAAGAAAACA
>target3
GGCATGACAAGGAAGTAGAAGAAAGCA
>target4
TTCGATCATGCAGTTCAATCAATGATC
```

human MYL6 (Figure S4A)

```
>anchor
AAGGTCCTCAGCCATTCAGCACCATGC
>P2_consensus1_macrophage
GGACGAGCTCTTCATAGTTGATACAACCATTGCTGTCCTCATGCCCTGCCACCAGCATCTCTACTTCTT
CCTCTGTCATCTTCTCACCCAGTGTGACAAGAACATGCCGGATTTC
>P2_consensus2_capillary
GGACGAGCTCCGCCCCATGGGCCCGTCACCCCGACAGGATATGCCTCACAAACGCTTCATAGTTGATAC
AACCATTGCTGTCCTCATGCCCTGCCACCAGCATCTCTACTTCTTCC
>P2_target1
TGCCACCAGCATCTCTACTTCTTCCTC
>P2_target2
CACAAACGCTTCATAGTTGATACAACC
>P3_consensus_macrophage
AAGGTCCTCAGCCATTCAGCACCATGCGGACGAGCTCTTCATAGTTGATACAACCATTGCTGTCCTCAT
GCCCTGCCACCAGCATCTCTACTTCTTCCTCTGTCATCTTCTCACCCAGTGTGACAAGAACATGCCGGA
>P3_consensus_capillary
AAGGTCCTCAGCCATTCAGCACCATGCGGACGAGCTCCGCCCCATGGGCCCGTCACCCCGACAGGATAT
GCCTCACAAACGCTTCATAGTTGATACAACCATTGCTGTCCTCATGCCCTGCCACCAGCATCTCTACTT
CTTCCT
```

mouse lemur COX2 (cytochrome c oxidase subunit II) (Figure S5A)

(reverse-complements are shown in Figure S5A)

```
>anchor
ATTTAGGCGCCCTGGGATAGCATCTGT
>target_1
TTCATGAATGTAGTACGTCTTCTGAAG
>target_2
TTCATGAATGTAATACGTCTTCTGAAG
```

lemur IGLC3 with 97 targets (Figure S5B)

```
>anchor
ACCGAGGGGGCGGCCTTGGGCTGACCT
>Targ0
GCCGAACACCCCAGTGCCACCACTCCT
>Targ1
```

```
GCCGAAGATATGACCACTCAGGCTGTC
>Targ2
GCCGAACACATGATTGTAGCTGCCATC
>Targ3
GCCGAATACATTAACACCACTGTTGTC
>Targ4
GCCGAACACATAACCATATGAATCACC
>Targ5
GCCGAACACACCACCACTGCTGTCCCC
>Targ6
GCCGAACACATTAACACCACCGTCCCA
>Targ7
GCCGAATACAGCACTGTTGTGCCACAC
>Targ8
GCCGAAGATATAAGTGTTCCTGCCCGC
>Targ9
GCCGAACACACCAACACCACTGCTGTC
>Targ10
GCCGAACACACCAACACCAGTTTCCCA
>Targ11
GCCGAAGATAACACCACTGTTGTCCCA
>Targ12
GCCGAACACACTGTAGCTGCCATCATA
>Targ13
GCCGAACACATAACCATATGAACCACC
>Targ14
GCCGAAGATATACTGAATGCTGCTCCC
>Targ15
GCCGAAGATATAAGTATTAGAGCTGCC
>Targ16
GCCGAACACCCGAGCATCAAGACTGCT
>Targ17
GCCGAATACATAAGCACTCAGGCTTTT
>Targ18
GCCGAACACCCGACCATTCAGGCTGCT
>Targ19
GCCGAATACATAAGTGCCACTGTTGGC
>Targ20
GCCGAAGATATACGCACTCAGGCTACT
>Targ21
GCCGAACACCTGACCACTCAGGCTACT
>Targ22
GCCGAACACACCAACACCACTGTTGTC
>Targ23
GCCGAACACCCAACTAGCACTGGCATC
```

```
>Targ24
GCCGAACACACCAGCACGTAGGCTGCT
>Targ25
GCCGAACACATGACCACTCAGGCTACT
>Targ26
GCCGAACACATGAGCACTCAGGCTTCT
>Targ27
GCCGAACACCCGACTGTAGCTGCCATC
>Targ28
GCCGAAGATATTAACACCACTGTTGTC
>Targ29
GCCGAAGATATCACTCAGGCTACTGTC
>Targ30
GCCGAACACCCAACTCTTAGAGCTGCC
>Targ31
GCCGAACACATCAGCACTGTTGTGCCA
>Targ32
GCCGAACACAAGATTGTAGCTGCCATC
>Targ33
GCCGAACACATAACTCTTAGAGCTGCC
>Targ34
GCCGAACACCCCAGTGCCACCACTCTT
>Targ35
GCCGAACACATCACCACTCAGGCTACT
>Targ36
GCCGAACACCCTGCTGTCATAGGACTG
>Targ37
GCCGAACACCCAATTAACACCACTGCT
>Targ38
GCCGAACACCCAAGCATCAAGACTGGT
>Targ39
GCCGAACACACGAGCATCAAGACTGCT
>Targ40
GCCGAACACCCAACCATATGAATCACC
>Targ41
GCCGAACACACCATGACCACTCAGGCT
>Targ42
GCCGAACACACCATAGTTTCCATAACC
>Targ43
GCCGAACACCGCATTAAGACTGCTGTC
>Targ44
GCCGAAGATATACTGGTTGCTGAACCA
>Targ45
GCCGAACACACCATGAGTACCAGTGCT
>Targ46
```

```
GCCGAATACATGACCACTCAGGCTGTC
>Targ47
GCCGAACACACCATCAAGACTGCTGTC
>Targ48
GCCGAAGATATAAGTGCCGCTGCCCGC
>Targ49
GCCGAACACATGACCACTCAGGCTTCT
>Targ50
GCCGAACACACCAGCATCAAGACTGCT
>Targ51
GCCGAAGATATAAGTGTTGCTGCCCGC
>Targ52
GCCGAACACCCAAGCATCAAGACTGCT
>Targ53
GCCGAACACACCATGACTCAGGCTGCT
>Targ54
GCCGAACACCCAAACACCACTGTTGTC
>Targ55
GCCGAACACATGAGCACTCAGGCTACT
>Targ56
GCCGAACAGACCACTCAGGCTACTATC
>Targ57
GCCGAAGATATACCCATATGAACCACC
>Targ58
GCCGAAGATATGACCACTCAGGCTACT
>Targ59
GCCGAACACCCAACCATATGAACCACC
>Targ60
GCCGAATACATAATTGTAGCTGTCATC
>Targ61
GCCGAACACACCACCACTCAGGCTGTC
>Targ62
GCCGAACACAAAATTAACACCACTGCT
>Targ63
GCCGAACACAGCACGCAGACTGCTGTC
>Targ64
GCCGAACACCCAAGTGCCGCTGCCCGC
>Targ65
GCCGAACACCCAGCACTGTTGTGCCAC
>Targ66
GCCGAAAACATAAGTCTTAGACCTGCC
>Targ67
GCCGAAGATATACGTATCAAGACTGCT
>Targ68
GCCGAAGATATTGTTTTCACTAACCCA
```

```
>Targ69
GCCGAAGATAGCACTGTTGTGCCACAC
>Targ70
GCCGAACACACGAGCACCCAGACTACT
>Targ71
GCCGAATACATGACCATTCAGGCTGCT
>Targ72
GCCGAATATATAACTCTTAGAACTGCC
>Targ73
GCCGAACACAAAACGGTTGCTGAACCA
>Targ74
GCCGAACATCCAACTCTTAGAGCTGCC
>Targ75
GCCGAACACCCAAGTCTTAGAGCTGCC
>Targ76
GCCGAACACATGACTGTAGCTGTCATC
>Targ77
GCCGAACACCCAATGGTTGCTGAACCA
>Targ78
GCCGAACACCCAAAGTGCCGCTGCCCG
>Targ79
GCCGAACACACCAGTCTTAGAGCTGCC
>Targ80
GCCGAAGATATTAACACCAGTTTCCCA
>Targ81
GCCGAACACACTGTAGCTGTCATCATA
>Targ82
GCCGAATACAAATGGTTGCTGAACCAC
>Targ83
GCCGAACACCCTATTAACACCACTGCT
>Targ84
GCCGAACACAGCATCAAGACTGCTGTC
>Targ85
GCCGAATACATAATCAAGACTGCTGTC
>Targ86
GCCGAACACACCACTCAGGCTACTATC
>Targ87
GCCGAAGATAGCATGAGTACCAGTATT
>Targ88
GCCGAAGATAAGACCACTCAGGCTACT
>Targ89
GCCGAACACAATAGCTGCCATCATAAG
>Targ90
GCCGAACACCTGATTGTAGCTGTCATC
>Targ91
```

```
GCCGAACACAAGACTAACACTGTCATC
>Targ92
CAGAGGCCTGTGTCCACCTGGGGAGCC
>Targ93
GCCGAACACACCTAGAGCTGCCATTCC
>Targ94
GCCGAATACATTAACACCACTGCTGTC
>Targ95
GCCGAATACATAATTGTAGCTGCCATC
>Targ96
GCCGAAGACAAACATCGACTGAGGCTC
```

lemur TCR-beta J-region (Figure S5C)

```
>anchor
CCGGGTCCCTGGCCCGAAGAACTGCTC
>Targ0
TGCCGCTGCAGATGTAGACGCCGCTGT
>Targ1
CGCAGAGATACAGGGCCGAGTCCCCCA
>Targ2
TGGCACAGAGGTACGTGGCGGAGTCTT
>Targ3
TGCTGGCACAGAGGTACGTGGCAGAGT
>Targ4
AGAGGAACAGGGCCGAGTCCCCCAGCG
```

lemur TCR-gamma V-region (Figure S5D)

```
>anchor
ACCCTCACCATTCACAATGTAGAGAAA
>Targ0
TGCCCGTGAACTCTTCAGTAATGGAAC
>Targ1
TGCCTCCTGGGAGTCTAGGAAACTCTT
>Targ2
TGCCTCCTGGGACTGACGACTTACCAA
>Targ3
TGCCTCCTGGGAGTTGAATTTTTATAG
>Targ4
TGCCTCCTGGGAGTTGCACAGTGTCAC
>Targ5
GCCCGTGAACTCTTCAGTAATGGAACA
>Targ6
TGCCTCCTGGGAGTCGCTCTCTAATAT
>Targ7
TGCCTCCTGGGAGTTGCACAGAAGATT
```

Octopus bimaculoides **carboxypeptidase D (Figure S6A)**

(reverse-complements are shown in Figure S6A)

```
>anchor
GGAATTAGAAGAAAAATCTATTATGAA
>target_1
AAATGTTTAGGCCAATATCTAAAGGCA
>target_2
AAATGTTTAGGAAAAATTTTCTGCCAA
```

Octopus bimaculoides **Upf2 (regulator of nonsense transcripts 2) (Figure S6B)**

(reverse-complements are shown in Figure S6B)

```
>anchor
GTATTGCACTGCATTGTACTGCACTGT
>target_1
CGCTGCTGCTGCTGCTGCTGCCAATTG
>target_2
CGCTGCTGCTGCTGCTGCCAATTGCCT
```

Octopus bimaculoides **netrin receptor / DCC (Figure S6C)**

(reverse-complements are shown in Figure S6C)

```
>anchor
TCTATTACAGCTATCATCAATACACTT
>target_1
TTGGATGTCTTCGTGTTCTCACTGCAG
>target_2
TTGGATGTCTTTGTGTTCTCACTGCAG
```

HMG-box (diatom) (Figure S7A)

(reverse-complements are shown in Figure S7A)

```
>anchor
TGCGGTCCTTGAATTCTTGCTTCTCTT
>target_1
TATCCGAAAGAGCCCTCCACATTTCAC
>target_2
CGTCCGTCAGAGCTCTCCACATTTCTC
```

ferredoxin (diatom) (Figure S7B)

(reverse-complements are shown in Figure S7B)

```
>anchor
ACGGCACGAGTAGGGAAGTTCAATTCC
>target_1
GGCTTCTTCAGCAGCGTCGACAATGAA
```

```
>target_2
GGCTTCTTCGGCAGCGTCGACAATGAA
```