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

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

Affiliations:

[1]Department of Biomedical Data Science, Stanford University; Stanford, 94305, USA.

[2]Department of Electrical Engineering, Stanford University; Stanford, 94305, USA.

[3]Department of Biochemistry, Stanford University; Stanford, 94305, USA.

[4]Department of Statistics (by courtesy), Stanford University; Stanford, 94305, USA.

*Corresponding author. Email: julia.salzman@stanford.edu

† Co-first authors

Abstract: We introduce a probabilistic model that enables study of myriad, disparate and fundamental problems in genome science and expands the scope of inference currently possible. Our model formulates an unrecognized unifying goal of many biological studies – to discover sample-specific sequence diversification – and subsumes many application-specific models. With it, we develop a novel algorithm, NOMAD, that performs valid statistical inference on raw reads, completely bypassing references and sample metadata. NOMAD's reference-free approach enables data-scientifically driven discovery with previously unattainable generality, illustrated with de novo prediction of adaptation in SARS-CoV-2, novel single-cell resolved, cell-type-specific isoform expression, including in the major histocompatibility complex, and de novo identification of V(D)J recombination. NOMAD is a unifying, provably valid and highly efficient algorithmic solution that enables expansive discovery.

One-Sentence Summary: We present a unifying formulation of disparate genomic problems and design an efficient, reference-free solution.

Introduction

Sequence diversification – mutation, reassortment or rearrangement of nucleic acids – is fundamental to evolution and adaptation across the tree of life. Diversification of pathogen genomes enables host range expansion, and as such host-interacting genes are under intense selective pressure (*1*). Sequence diversity in these genes is sample-dependent in hosts and in time as dominant strains emerge. In jawed vertebrates, V(D)J recombination and somatic hypermutation generate sequence diversity during the adaptive immune response that varies across cells, and is thus sample-dependent. Sequence diversification in the transcriptome takes the form of regulated RNA-isoform expression. This diversification enables varied phenotypes from the same reference genome including expression programs that enable cell specialization: cell-type specific splicing yields sample-dependent sequence diversity, samples being cells or cell-types. These examples are snapshots of sequence diversification which is core to wide-ranging biological functions, including adaptation, with applications in disparate fields ranging from plant biology to ecological metagenomics, among others (Fig. 1A, Supplement).

In this work, we identify the conceptually uniting biological goal in the above problems: to determine sequences with sample-dependent diversity. Biological studies today typically provide deep sampling of the nucleic acid composition, for example through RNA-seq or DNA-seq. In principle, this data provides an opportunity to identify sample-dependent sequence diversity with powerful and efficient statistical models. A timely example is sequencing of SARS-CoV-2: sequences in the spike glycoprotein are diversified as strains compete, for example during the emergence of the omicron variant (*2*). Patient samples collected in late 2021 include sampling of Delta and Omicron strains, thus containing sample-dependent sequence diversity in regions that differentiate the strains, including but not limited to ones in the spike protein coding region. Statistical approaches should be able to capture this diversity without a reference genome.

Today, genomic algorithms to detect sample-specific diversity are highly specialized to each application and lack conceptual unification, creating several issues. First, biological inference is limited if a workflow is chosen that does not have power to detect significant signals in the data. For example, if the workflow does not map transposable element insertions, none will be found, as these variants are missing from references.

Second, the first step in existing approaches almost always requires the use of references through alignment or their construction *de novo*. In human genomics, workflows beginning with reference alignment miss important variation absent from assemblies, even with pangenomic approaches. It is well-known that genetic variants associated with ancestry of under-studied populations are poorly represented in databases and result in health disparities (*3*). In disease genomics (*4*), sequences of pathogenic cells may be missing from reference genomes, or no reference genome exists, as in genomically unstable tumors (*5*). Many species do not have reference genomes even when they in principle could be attained due to logistical and computational overheads. In these cases, alignment-based methods we call "reference-first" approaches fail. In viral surveillance, reference-first approaches are even more problematic (*6*). Viral reference genomes cannot capture the complexity of viral quasispecies (*7*) or the vast extent of polymorphism (*8*). New viral assemblies are constantly being added to reference databases (*9*, *10*). In the microbial world, pre-specifying a set of reference genomes is infeasible due to its inherent rapid genomic changes. References also cannot capture insertional diversity of mobile elements, which have significant phenotypic and clinical impact (*11*) and are only partially cataloged in references (*12*).

Third, current approaches have severe technical limitations. Statistical analysis in reference-first approaches is conditional on the output of the noisy alignment step, making it difficult or impossible to provide valid statistical significance levels in downstream analyses, such as differential testing or expression analysis. When available, computationally intensive resampling is required.

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

In this work we first show that detecting sample-dependent sequence diversification can be formulated probabilistically. Second, we reduce this to a statistical test on raw sequencing read data (e.g. FASTQ files). Third, we implement the test in a highly efficient algorithmic workflow, providing novel discovery across disparate biological disciplines (Fig. 1B,C).

To begin, we define an "anchor" k-mer in a read, and say the anchor has sample-dependent diversity if the distribution of k-mers starting R basepairs downstream of it (called "targets") depends on the sample (Fig. 1E) (*13*). Inference can be performed for much more general constructions of anchors and targets: any tuple of disjoint subsequences from DNA, RNA or protein sequence data can be analyzed in this framework (Supplement).

This formulation unifies many fundamental problems in genome science. It allows us to develop a novel statistics-first approach, NOMAD (Novel multi-Omics Massive-scale Analysis and Discovery), that is reference-free and operates directly on raw sequencing data. It is an extremely computationally efficient algorithm to detect sample-dependent sequence diversification, through the use of a novel statistic of independent interest that provides closed form p-values (Methods). NOMAD makes all predictions blind to references and annotations, though they can be optionally used for *post-facto* interpretation. This makes NOMAD fundamentally different from existing methods. To illustrate, as a special case NOMAD is detection of differential isoform expression. In this domain, the closest approach to NOMAD is Kallisto (*14*) which requires a reference transcriptome and statistical resampling for inference, and is further challenged to provide exact quantification for more than a handful of paralogous genes and isoforms. Unlike NOMAD, Kallisto cannot discover spliced isoforms *de novo.*

NOMAD's calls are "significant anchors": sequences *a* where, given observing *a* in a read, the conditional distribution of observing a target sequence *t* a distance *R* downstream of *a* is sample-dependent (Fig. 1E). NOMAD is by default an unsupervised algorithm that does not require any sample identity metadata. It finds approximate best splits of data into two groups (Methods), or it can use user-defined groups if desired. Anchors are reported with an effect size in [0,1], a measure of target distribution difference between sample groups: 0 if the groups have no difference in target distributions and increasing to 1 when the target distributions of the two groups are disjoint. NOMAD has multiple major technical innovations: 1) a parallelizable, fully containerized, and computationally efficient approach to parse FASTQ files into contingency tables, 2) novel statistical analysis of the derived tables, using concentration inequalities to derive closed form p-values, 3) a micro-assembly-based consensus sequence representing the dominant error-corrected sequence, similar to (*15*, *16*), downstream of the anchor for post-facto interpretation and identification of SNPs, indels or isoforms, to name a few (Fig. 1D). If post anchor-identification inference is desired, the consensus, rather than raw reads, is aligned. This reduces the number of reads to align by ~1000x in real data.

Together, NOMAD's theoretical development yields an extremely computationally efficient implementation. We ran NOMAD on a 2015 Intel laptop with an Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz processor, generating significance calls for single cell RNA-seq totaling over 43 million reads in only 1 hour 45 min. When performed on a compute cluster, the same analysis is completed in an average of 2.28 minutes with 750 MB of memory for 4 million reads, a dramatic speed up over existing methods for *de novo* splicing detection and significance calls.

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

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

To test this, we analyzed Oropharyngeal swabs from patients with SARS-CoV-2 from 2021-12-6 to 2022-2-27 in France, a period of known Omicron-Delta coinfection (*17*). We ran NOMAD and analyzed anchors with effect size >.5, as high effects are predicted if samples can be approximately partitioned by strain, though results are similar without this threshold (fig. S2). For each anchor, we assigned a protein domain label based on *in silico* translation of its consensus sequence (*18*). The protein domain with best mapping to the Pfam database is assigned to the anchor, producing a set of "NOMAD protein profiles" (Methods, table S1), and is compared to matched controls (Methods). NOMAD protein profiles are significantly different from controls (p=1.1E-12, chi-squared test, Fig. 2B). The most NOMAD-enriched domains are the receptor binding domain of the betacoronavirus spike glycoprotein (7 NOMAD vs 0 control hits, p=2.9E-4 hypergeometric p-value, corrected,) and Orf7A a transmembrane protein that inhibits host antiviral response (6 NOMAD vs 0 control hits, p=1.6E-3 hypergeometric p-value, corrected). All analysis is blind to the data origin (SARS-CoV-2).

We further analyzed patient samples from the original South African genomic surveillance study that identified the Omicron strain during the period 2021-11-14 to 2021-11-23 (*19*), again without metadata, reference genomes and directly on input FASTQ files (Methods). NOMAD protein profiles were significantly different from controls (chi-squared test, p=2.5E-39, Fig. 2B). NOMAD-enriched domains in France and South African are highly consistent: the top four domains in both datasets are permutations of each other. The most NOMAD-enriched domains versus controls were the betacoronavirus S2 subunit of the spike protein involved in eliciting the human antibody response (*20*) (23 NOMAD vs 2 control hits, p=2.9E-6 hypergeometric p-value, corrected), the matrix glycoprotein which interacts with the spike (20 NOMAD vs 0 control hits, p=8.4E-8 hypergeometric p-value, corrected), and the receptor binding domain of the spike protein and to which human antibodies have been detected (*21*) (20 NOMAD vs 0 control hits, p=8.4E-8 hypergeometric p-value, corrected). All domains are biologically predicted to be under strong selective pressure; NOMAD discovers this *de novo*.

We further aligned NOMAD anchors with effect size >.5 to the Wuhan reference strain to test if NOMAD anchors rediscovered known strain-defining mutations. We defined a mutation-consistent anchor as one consistent with detecting an Omicron or Delta variant mutation (Supplement). NOMAD anchors are significantly enriched for being mutation-consistent: in the French data, of the uniquely mapped anchors (Wuhan reference), 30.5% (44/144) of NOMAD's calls were mutation consistent versus 7.6% (6/78) in the control (p=4.0E-5, hypergeometric test). For the South African data, 67.9% (89/131) of NOMAD's called anchors were mutation consistent vs 20% (12/60) for the control (p=4.4E-10, hypergeometric test).

Examples of mutation-consistent anchors are presented in Fig. 2 including in the membrane (Fig. 2D, top) and spike protein (Fig. 2D, middle and bottom). Differences between consensus sequences and Wuhan reference illustrate NOMAD's unsupervised rediscovery of annotated strain-specific variants, including co-detection of a deletion and a mutation (Fig. 2D). While the Wuhan reference genome was used for *post-facto* interpretation, no alignment or sample metadata was used to generate NOMAD's calls, only to interpret them (Fig. 2). NOMAD consensuses also extend discovery (Fig. 2D), identifying strain-specific variants beyond the target: both are annotated omicron variants (Fig. 2D). NOMAD's statistical approach automatically links discovered variants within patients *de novo*: one consensus contains the omicron Variant of Concern (VOC) T22882G; a second consensus has a single VOC T22917G identified in Omicron strains BA.4 and BA.5 in May of 2022, 3 months after the analyzed samples were collected; a third consensus contains the VOC as well as the VOC G22898A; a single further consensus shows no mutations, consistent with Delta infection. Together, this

suggests that mutations in BA.4 and BA.5 were circulating well before the VOC was called in May 2022.

We further analyzed 499 samples collected in California (2020) before viral strain divergence in the spike had been reported (*22*) as a negative control. No enrichment of NOMAD protein profiles related to the spike or Orf7a domains were observed (fig. S2B), supporting the idea that NOMAD calls are not false positives. To explore the generality of NOMAD for reference-free discovery in other viral infections, we additionally ran NOMAD on a study of influenza-A and of rotavirus breakthrough cases (Fig. S2A,B, Methods). NOMAD Protein profile analysis showed enrichment in domains involved in viral suppression of the host response and regulated alternative splicing (Supplement). Together, this suggests that NOMAD analysis could aid in viral surveillance, including detecting emergence of variants of concern directly from short read sequencing, bypassing a requirement for reference genomes and without manual scrutiny of individual samples or their assembled genomes (*23*).

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

NOMAD is a general algorithm to discover sample-dependent sequence diversification in disparate applications including RNA expression and beyond. To illustrate the former, we ran NOMAD without any parameter tuning on single cell RNA-seq datasets, testing if it could perform the fundamental but previously distinct tasks of identifying regulated expression of paralogous genes, alternative splicing and V(D)J recombination (Fig. 3A).

First, we tested if NOMAD discovers alternatively spliced genes in single cell RNA-seq (Smart-seq 2) of human macrophage versus capillary lung cells, chosen because they have a recently established positive control of alternative splicing, MYL6, a subunit of the myosin light chain (*24*). NOMAD rediscovered MYL6 and made new discoveries not reported in the literature. For example, we discovered reproducible cell-type specific regulation of MYL12 isoforms, MYL12A and MYL12B. Like MYL6, MYL12 is a subunit of the myosin light chain. In humans (as in many species) two paralogous genes, MYL12A and MYL12B, sharing >95% nucleotide identity in the coding region, are located in tandem on chromosome 18 (Fig. 3B). Reference-first algorithms fail to quantify differential expression of MYL12A and MYL12B due to mapping ambiguity. NOMAD automatically detects targets that unambiguously distinguish the two paralogous isoforms, and demonstrates their clear differential regulation in capillary cells and macrophages (Fig. 3B). We confirmed MYL12 isoform specificity in pairwise comparisons of the same cell types in two further independent single cell sequencing studies of primary cells from the same cell-types (Supplement). MYL12 was recently discovered to mediate allergic inflammation by interacting with CD69 (*25*); while today little is known about differential functions of the two MYL12 paralogs, the distinct roles of highly similar actin paralogs provides a precedent (*26*, *27*).

NOMAD also called reproducible cell-type specific allelic expression and splicing in the major histocompatibility (MHC) locus (Fig. 3C), the most polymorphic region of the human genome which carries many significant disease risk associations (*28*). Despite its central importance in human immunity and complex disease, allotypes are difficult to quantify, and statistical methods to reliably distinguish them do not exist. NOMAD finds (i) allele-specific expression of HLA-DRB within cell types and (ii) cell-type specific splicing, predicted to change the amino acid and 3' UTR sequence of HLA-DPA1 (Methods, Fig. 3D). These empirical results bear out a snapshot of the theoretical prediction that NOMAD's design gives it high statistical power to simultaneously identify isoform expression variation and allelic expression, including that missed by existing algorithms (table S3).

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

We next tested if NOMAD could identify T cell receptor (TCR) and B cell receptor (BCR) variants. B and T cell receptors are generated through V(D)J recombination and somatic hypermutation, yielding sequences that are absent from any reference genome and cannot be cataloged comprehensively due to their diversity (>10^{12}) (*29*). Existing methods to identify V(D)J rearrangement require specialized workflows that depend on receptor annotations and alignment and thus fail when the loci are unannotated (*30*, *31*). In many organisms, including *Microcebus murinus*, the mouse lemur, T cell receptor loci are incompletely unannotated as they must be manually curated (*32*)).

We tested if NOMAD could identify TCR and BCR rearrangements in the absence of annotations on 111 natural killer T and 289 B cells isolated from the spleen of two mouse lemurs (*Microcebus murinus*) profiled by Smart-Seq2 (SS2) (*32*), and performed the same analysis on a random choice of 50 naive B cells from the peripheral blood and 128 CD4+ T cells from two human donors profiled with SS2 (*33*) for comparison (Fig. 4A, Methods).

NOMAD protein profiles (Fig. 4B), which are blind to the sample's biological origin and do not require any reference genome, revealed that NOMAD's most frequent hits in lemur B cell were IG-like domains resembling the antibody variable domain (86 hits), and COX2 (55 hits) a subunit of cytochrome c oxidase, known to be activated in the inflammatory response (*34*). NOMAD's top hits for Lemur T cell were COX2 and MHC_I (77 and 58 hits, respectively). Similar results were obtained for the human samples (Fig. 4B). They include novel predictions of cell-type specific allelic expression of HLA-B in T cells (Fig. 4D, Supplement) where NOMAD found statistical evidence that cells preferentially express a single allele ($p< 4{,}6E-24$); consensus analysis shows SNPs in HLA-B are concordant with known positions of polymorphism.

We further predicted that BCR and TCR rearrangements would also be discovered by investigating the transcripts most hit by NOMAD anchors. We mapped NOMAD-called lemur B and T cell anchors to an approximation of its transcriptome: that from humans which diverged from lemur ~60-75 million years ago (*35*). Lemur B cell anchors most frequently hit the immunoglobulin light and kappa variable regions; lemur T cell anchors most frequently hit the HLA and T cell receptor family genes (Methods, Fig. 4C). Similar results are found in human B and T cells (Fig. 4E, Supplement). Transcripts with the most hits in the control were unrelated to immune function. To further illustrate NOMAD's power, consider its comparison to existing pipelines. They cannot be run without the annotation for the lemur TCR locus; for assembling BCR sequences, pipelines e.g. BASIC (*31*) cannot always identify V(D)J rearrangement, including in some cells profiled in the lemur dataset (*32*). We selected the 35 B/plasma cells where BASIC could not programmatically identify variable gene families on the light chain variable region. NOMAD automatically identified anchors mapping to the IGLV locus, with consensus sequences that BLAST to the light chain variable region (Supplement). Together, this shows that NOMAD identifies sequences with adaptive immune function including V(D)J in both B and T cells *de novo*, using either no reference genome (protein profile analysis) or only an annotation guidepost from a related organism (human). In addition to being simple and unifying, NOMAD can extend discovery compared to custom pipelines.

Conclusion

We have shown that problems from disparate subfields of genomic data-science are unified in their goal to discover sample-dependent sequence diversification; NOMAD is a statistics-first algorithm that formulates and efficiently solves this task, with great generality.

We provided a snapshot of NOMAD's discoveries in disparate areas in genome science. In SARS-CoV-2 patient samples during the emergence of the omicron variant, it finds the spike

protein is highly enriched for sequence diversification, bypassing genome alignment completely. NOMAD provides evidence that Variants of Concern can be detected well before they are flagged as such or added to curated databases. This points to a broader impact for NOMAD in viral and other genomic surveillance, since emerging pathogens will likely have sequence diversification missing from any reference.

NOMAD finds novel cell-type specific isoform expression in homologous genes missed by reference-guided approaches, such as in MYL12A/B and in the MHC (HLA) locus, even in a small sample of single cell data. Highly polymorphic and multicopy human genes have been recalcitrant to current genomic analyses and are critical to susceptibility to infectious and complex diseases, e.g. the MHC. NOMAD could shed new light on other polymorphic loci including non-coding RNAs, e.g. spliceosomal variants (*36*, *37*). In addition, NOMAD unifies detection of many other examples of transcriptional regulation: intron retention, alternative linear splicing, allele-specific splicing, gene fusions, and circular RNA. Further, NOMAD prioritizes V(D)J recombination as the most sample-specific sequencing diversifying process in B and T cells of both human and mouse lemur, where inference in lemur is made using only an approximate genomic reference (human) which diverged from lemur ~60 million years ago.

Disparate data – DNA and protein sequence, or any "-omics" experiment, from Hi-C to spatial transcriptomics – can be analyzed in the NOMAD framework. NOMAD may also be impactful in analysis of plants, microbes, and mobile elements, including transposable elements and retrotransposons, which are far less well annotated, and are so diverse that references may never capture them.

NOMAD illustrates the power of statistics-first genomic analysis with optional use of references for *post-facto* interpretation. It translates the field's "reference-first" approach to "statistics-first", performing direct statistical hypothesis tests on raw sequencing data, enabled by its probabilistic modeling of raw reads rather than of alignment outputs. By design, NOMAD is highly efficient: it will enable direct, large-scale study of sample-dependent sequence diversification, completely bypassing the need for references or assemblies. NOMAD promises data-driven biological study previously impossible.

Limitations of the study

Naturally, some problems cannot be formulated in the manner posed, such as cases where the estimand is RNA or DNA abundance. However, the problems that can be addressed using this formulation span diverse fields which are of great current importance (Supplement), including those previously discussed. Further, NOMAD's statistical test can be applied to tables of gene expression (including as measured by *k*-mers) by samples.

References and Notes

1. L. Yang, M. Emerman, H. S. Malik, R. N. McLaughlin, Retrocopying expands the functional repertoire of APOBEC3 antiviral proteins in primates. *eLife*. **9** (2020), doi:10.7554/eLife.58436.
2. L. Wang, G. Cheng, Sequence analysis of the emerging SARS-CoV-2 variant Omicron in South Africa. *J. Med. Virol.* **94**, 1728–1733 (2022).
3. K. M. West, E. Blacksher, W. Burke, Genomics, health disparities, and missed opportunities for the nation's research agenda. *JAMA*. **317**, 1831–1832 (2017).
4. T. Wang, L. Antonacci-Fulton, K. Howe, H. A. Lawson, J. K. Lucas, A. M. Phillippy, A. B. Popejoy, M. Asri, C. Carson, M. J. P. Chaisson, X. Chang, R. Cook-Deegan, A. L. Felsenfeld, R. S. Fulton, E. P. Garrison, N. A. Garrison, T. A. Graves-Lindsay, H. Ji, E. E. Kenny, B. A. Koenig, Human Pangenome Reference Consortium, The Human Pangenome Project: a global resource to map genomic diversity. *Nature*. **604**, 437–446 (2022).
5. C.-Z. Zhang, A. Spektor, H. Cornils, J. M. Francis, E. K. Jackson, S. Liu, M. Meyerson, D. Pellman, Chromothripsis from DNA damage in micronuclei. *Nature*. **522**, 179–184 (2015).
6. The Nucleic Acid Observatory Consortium, A Global Nucleic Acid Observatory for Biodefense and Planetary Health. *arXiv* (2021), doi:10.48550/arxiv.2108.02678.
7. K. Kirkegaard, N. J. van Buuren, R. Mateo, My Cousin, My Enemy: quasispecies suppression of drug resistance. *Curr. Opin. Virol.* **20**, 106–111 (2016).
8. D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, H. Chang, The Architecture of SARS-CoV-2 Transcriptome. *Cell*. **181**, 914-921.e10 (2020).
9. R. C. Edgar, J. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, B. Al-Shayeb, J. F. Banfield, M. de la Peña, A. Korobeynikov, R. Chikhi, A. Babaian, Petabase-scale sequence alignment catalyses viral discovery. *Nature*. **602**, 142–147 (2022).
10. A. A. Zayed, J. M. Wainaina, G. Dominguez-Huerta, E. Pelletier, J. Guo, M. Mohssen, F. Tian, A. A. Pratama, B. Bolduc, O. Zablocki, D. Cronin, L. Solden, E. Delage, A. Alberti, J.-M. Aury, Q. Carradec, C. da Silva, K. Labadie, J. Poulain, H.-J. Ruscheweyh, P. Wincker, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science*. **376**, 156–162 (2022).
11. D. R. Evans, M. P. Griffith, A. J. Sundermann, K. A. Shutt, M. I. Saul, M. M. Mustapha, J. W. Marsh, V. S. Cooper, L. H. Harrison, D. Van Tyne, Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *eLife*. **9** (2020), doi:10.7554/eLife.53886.
12. R. J. Wright, A. M. Comeau, M. G. I. Langille, From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *BioRxiv* (2022), doi:10.1101/2022.04.27.489753.
13. J. Abante, P. L. Wang, J. Salzman, DIVE: a reference-free statistical approach to diversity-generating and mobile genetic element discovery. *BioRxiv* (2022), doi:10.1101/2022.06.13.495703.
14. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

15. A. Motahari, K. Ramchandran, D. Tse, N. Ma, in *2013 IEEE International Symposium on Information Theory* (IEEE, 2013), pp. 1640–1644.

16. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. **27**, 2957–2963 (2011).

17. A. Bal, B. Simon, G. Destras, R. Chalvignac, Q. Semanas, A. Oblette, G. Queromes, R. Fanget, H. Regue, F. Morfin, M. Valette, B. Lina, L. Josset, Detection and prevalence of SARS-CoV-2 co-infections during the Omicron variant circulation, France, December 2021 - February 2022. *medRxiv* (2022), doi:10.1101/2022.03.24.22272871.

18. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

19. R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J. Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, T. de Oliveira, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*. **603**, 679–686 (2022).

20. C. M. Poh, G. Carissimo, B. Wang, S. N. Amrun, C. Y.-P. Lee, R. S.-L. Chee, S.-W. Fong, N. K.-W. Yeo, W.-H. Lee, A. Torres-Ruesta, Y.-S. Leo, M. I.-C. Chen, S.-Y. Tan, L. Y. A. Chai, S. Kalimuddin, S. S. G. Kheng, S.-Y. Thien, B. E. Young, D. C. Lye, B. J. Hanson, L. F. P. Ng, Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients. *Nat. Commun.* **11**, 2806 (2020).

21. P. Jörrißen, P. Schütz, M. Weiand, R. Vollenberg, I. M. Schrempf, K. Ochs, C. Frömmel, P.-R. Tepasse, H. Schmidt, A. Zibert, Antibody Response to SARS-CoV-2 Membrane Protein in Patients of the Acute and Convalescent Phase of COVID-19. *Front. Immunol.* **12**, 679841 (2021).

22. J. E. Gorzynski, H. N. De Jong, D. Amar, C. R. Hughes, A. Ioannidis, R. Bierman, D. Liu, Y. Tanigawa, A. Kistler, J. Kamm, J. Kim, L. Cappello, N. F. Neff, S. Rubinacci, O. Delaneau, M. J. Shoura, K. Seo, A. Kirillova, A. Raja, S. Sutton, V. N. Parikh, High-throughput SARS-CoV-2 and host genome sequencing from single nasopharyngeal swabs. *medRxiv* (2020), doi:10.1101/2020.07.27.20163147.

23. D. Jacot, T. Pillonel, G. Greub, C. Bertelli, Assessment of SARS-CoV-2 Genome Sequencing: Quality Criteria and Low-Frequency Variants. *J. Clin. Microbiol.* **59**, e0094421 (2021).

24. J. E. Olivieri, R. Dehghannasiri, P. L. Wang, S. Jang, A. de Morree, S. Y. Tan, J. Ming, A. Ruohao Wu, Tabula Sapiens Consortium, S. R. Quake, M. A. Krasnow, J. Salzman, RNA splicing programs define tissue compartments and cell types at single-cell resolution. *eLife*. **10** (2021), doi:10.7554/eLife.70692.

25. K. Hayashizaki, M. Y. Kimura, K. Tokoyoda, H. Hosokawa, K. Shinoda, K. Hirahara, T. Ichikawa, A. Onodera, A. Hanazawa, C. Iwamura, J. Kakuta, K. Muramoto, S. Motohashi, D. J. Tumes, T. Iinuma, H. Yamamoto, Y. Ikehara, Y. Okamoto, T. Nakayama, Myosin light chains 9 and 12 are functional ligands for CD69 that regulate airway inflammation. *Sci. Immunol.* **1**, eaaf9154 (2016).

26. P. Vedula, S. Kurosaka, N. A. Leu, Y. I. Wolf, S. A. Shabalina, J. Wang, S. Sterling, D. W. Dong, A. Kashina, Diverse functions of homologous actin isoforms are defined by their

nucleotide, rather than their amino acid sequence. *eLife*. **6** (2017), doi:10.7554/eLife.31661.

27. B. J. Perrin, J. M. Ervasti, The actin gene family: function follows isoform. *Cytoskeleton (Hoboken)*. **67**, 630–634 (2010).

28. V. Matzaraki, V. Kumar, C. Wijmenga, A. Zhernakova, The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).

29. B. Briney, A. Inderbitzin, C. Joyce, D. R. Burton, Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*. **566**, 393–397 (2019).

30. I. Lindeman, G. Emerton, L. Mamanova, O. Snir, K. Polanski, S.-W. Qiao, L. M. Sollid, S. A. Teichmann, M. J. T. Stubbington, BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods*. **15**, 563–565 (2018).

31. S. Canzar, K. E. Neu, Q. Tang, P. C. Wilson, A. A. Khan, BASIC: BCR assembly from single cells. *Bioinformatics*. **33**, 425–427 (2017).

32. The Tabula Microcebus Consortium, C. Ezran, S. Liu, S. Chang, J. Ming, O. Botvinnik, L. Penland, A. Tarashansky, A. de Morree, K. J. Travaglini, K. Hasegawa, H. Sin, R. Sit, J. Okamoto, R. Sinha, Y. Zhang, C. J. Karanewsky, J. L. Pendleton, M. Morri, M. Perret, M. A. Krasnow, Tabula Microcebus: A transcriptomic cell atlas of mouse lemur, an emerging primate model organism. *BioRxiv* (2021), doi:10.1101/2021.12.12.469460.

33. Tabula Sapiens Consortium*, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco, S. R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown, W. Harper, M. Hemenez, R. Ponnusamy, A. Salehi, B. A. Sanagavarapu, E. Spallino, K. A. Aaron, W. Concepcion, J. M. Gardner, B. Kelly, et al., The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*. **376**, eabl4896 (2022).

34. A. L. Groeger, C. Cipollina, M. P. Cole, S. R. Woodcock, G. Bonacci, T. K. Rudolph, V. Rudolph, B. A. Freeman, F. J. Schopfer, Cyclooxygenase-2 generates anti-inflammatory mediators from omega-3 fatty acids. *Nat. Chem. Biol.* **6**, 433–441 (2010).

35. C. Ezran, C. J. Karanewsky, J. L. Pendleton, A. Sholtz, M. R. Krasnow, J. Willick, A. Razafindrakoto, S. Zohdy, M. A. Albertelli, M. A. Krasnow, The mouse lemur, a genetic model organism for primate biology, behavior, and health. *Genetics*. **206**, 651–664 (2017).

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

37. S.-M. Kuo, C.-J. Chen, S.-C. Chang, T.-J. Liu, Y.-H. Chen, S.-Y. Huang, S.-R. Shih, Inhibition of Avian Influenza A Virus Replication in Human Cells by Host Restriction Factor TUFM Is Correlated with Autophagy. *MBio*. **8** (2017), doi:10.1128/mBio.00481-17.

38. P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

39. J. Salzman, H. Jiang, W. H. Wong, Statistical Modeling of RNA-Seq Data. *Stat. Sci.* **26** (2011), doi:10.1214/10-STS343.

40. A. Agresti, A Survey of Exact Inference for Contingency Tables. *Stat. Sci.* **7**, 131–153 (1992).

41. P. Diaconis, B. Sturmfels, Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** (1998), doi:10.1214/aos/1030563990.

42. Y. Chen, P. Diaconis, S. P. Holmes, J. S. Liu, Sequential monte carlo methods for statistical analysis of tables. *J. Am. Stat. Assoc.* **100**, 109–120 (2005).

43. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188 (2001).

44. A. Ståhlberg, P. M. Krzyzanowski, J. B. Jackson, M. Egyud, L. Stein, T. E. Godfrey, Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Res.* **44**, e105 (2016).

45. I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, E. Rivas, S. R. Eddy, R. D. Finn, A. Bateman, A. I. Petrov, Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).

46. J. Storer, R. Hubley, J. Rosen, T. J. Wheeler, A. F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA*. **12**, 2 (2021).

47. K. Ross, A. M. Varani, E. Snesrud, H. Huang, D. O. Alvarenga, J. Zhang, C. Wu, P. McGann, M. Chandler, Tncentral: a prokaryotic transposable element database and web portal for transposon analysis. *MBio*. **12**, e0206021 (2021).

48. R. Leplae, A. Hebrant, S. J. Wodak, A. Toussaint, ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* **32**, D45-9 (2004).

49. D. Bi, Z. Xu, E. M. Harrison, C. Tai, Y. Wei, X. He, S. Jia, Z. Deng, K. Rajakumar, H.-Y. Ou, ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res.* **40**, D621-6 (2012).

50. D. Couvin, A. Bernheim, C. Toffano-Nioche, M. Touchon, J. Michalik, B. Néron, E. P. C. Rocha, G. Vergnaud, D. Gautheret, C. Pourcel, CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).

51. M. Santamaria, B. Fosso, F. Licciulli, B. Balech, I. Larini, G. Grillo, G. De Caro, S. Liuni, G. Pesole, ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences. *Nucleic Acids Res.* **46**, D127–D132 (2018).

52. C. Selig, M. Wolf, T. Müller, T. Dandekar, J. Schultz, The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Res.* **36**, D377-80 (2008).

53. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE*. **11**, e0163962 (2016).

54. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. **11**, 431 (2010).

55. M. G. Thompson, M. Dittmar, M. J. Mallory, P. Bhat, M. B. Ferretti, B. M. Fontoura, S. Cherry, K. W. Lynch, Viral-induced alternative splicing of host genes promotes influenza replication. *eLife*. **9** (2020), doi:10.7554/eLife.55500.

56. X. Sun, G. R. Whittaker, Role of the actin cytoskeleton during influenza virus internalization into polarized epithelial cells. *Cell. Microbiol.* **9**, 1672–1682 (2007).

57. Y. Song, N. Feng, L. Sanchez-Tacuba, L. L. Yasukawa, L. Ren, R. H. Silverman, S. Ding, H. B. Greenberg, Reverse genetics reveals a role of rotavirus VP3 phosphodiesterase activity in inhibiting rnase L signaling and contributing to intestinal viral replication in vivo. *J. Virol.* **94** (2020), doi:10.1128/JVI.01952-19.

58. M. Gratia, E. Sarot, P. Vende, A. Charpilienne, C. H. Baron, M. Duarte, S. Pyronnet, D. Poncet, Rotavirus NSP3 Is a Translational Surrogate of the Poly(A) Binding Protein-Poly(A) Complex. *J. Virol.* **89**, 8773–8782 (2015).

59. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

60. P. Kiepiela, A. J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale, S. Chetty, P. Rathnavalu, C. Moore, K. J. Pfafferott, L. Hilton, P. Zimbwa, S. Moore, T. Allen, C. Brander, M. M. Addo, M. Altfeld, I. James, S. Mallal, M. Bunce, L. D. Barber, P. J. R. Goulder, Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature*. **432**, 769–775 (2004).

61. S. Elahi, W. L. Dinges, N. Lejarcegui, K. J. Laing, A. C. Collier, D. M. Koelle, M. J. McElrath, H. Horton, Protective HIV-specific CD8+ T cells evade Treg cell suppression. *Nat. Med.* **17**, 989–995 (2011).

62. J. M. Francis, D. Leistritz-Edwards, A. Dunn, C. Tarr, J. Lehman, C. Dempsey, A. Hamel, V. Rayon, G. Liu, Y. Wang, M. Wille, M. Durkin, K. Hadley, A. Sheena, B. Roscoe, M. Ng, G. Rockwell, M. Manto, E. Gienger, J. Nickerson, D. C. Pregibon, Allelic variation in class I HLA determines CD8+ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2. *Sci. Immunol.* **7**, eabk3070 (2022).

63. B. Medhekar, J. F. Miller, Diversity-generating retroelements. *Curr. Opin. Microbiol.* **10**, 388–395 (2007).

64. R. A. Fisher, On the Interpretation of X2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*. **85**, 87 (1922).

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

Funding:

Stanford University Discovery Innovation Award (J.S.)

National Institute of General Medical Sciences grant R35 GM139517 (J.S.)
National Science Foundation Faculty Early Career Development Program Award MCB1552196 (J.S.)
National Science Foundation Graduate Research Fellowship Program (T.Z.B.)
Stanford Graduate Fellowship (T.Z.B.)

Author contributions

K.C. designed and implemented the pipeline. T.Z.B. designed, developed, and analyzed the statistics. I.N.Z. developed the protein domain analysis. J.S. designed and developed the statistics, and conceptualized and supervised the project. All authors analyzed data and wrote the manuscript.

Competing interests

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

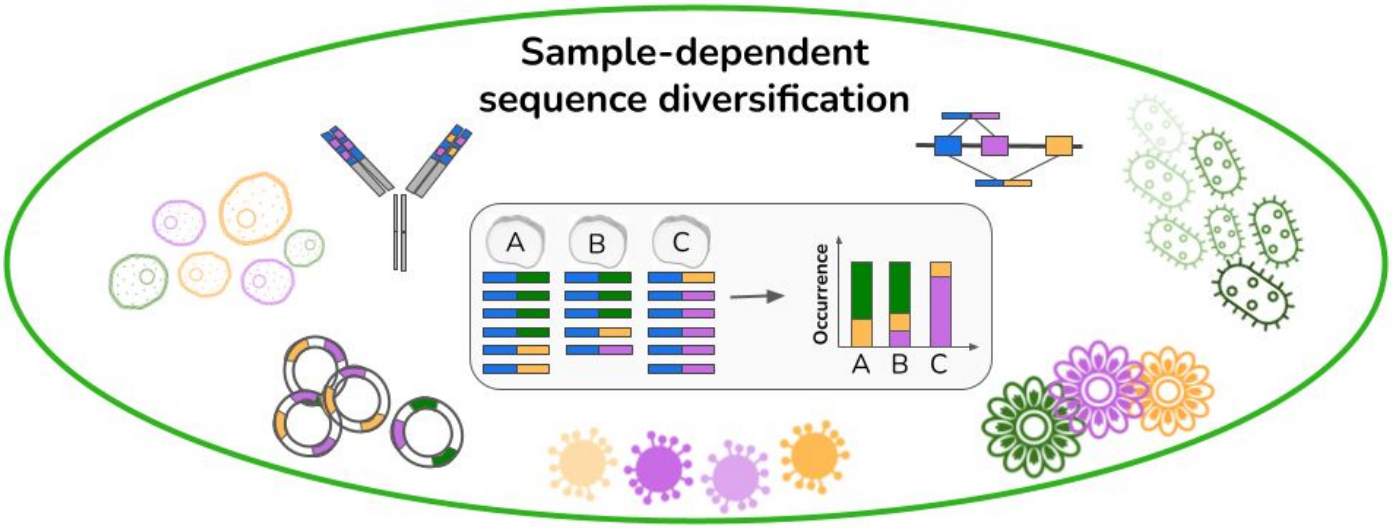
Data and materials availability

The code used in this work is available as a fully-containerized Nextflow pipeline (*38*) at https://github.com/kaitlinchaung/nomad.
The human lung scRNA-seq data used here is accessible through the European Genome-phenome Archive (accession number: EGAS00001004344); FASTQ files from donor 1 and donor 2 were used. The FASTQ files for the Tabula Sapiens data (both 10X Chromium and Smart-seq2) were downloaded from https://tabula-sapiens-portal.ds.czbiohub.org/; B cells were used from donor 1 and T cells from donor 2. The mouse lemur single-cell RNA-seq data used in this study was generated as part of the Tabula Microcebus consortium; the FASTQ files were downloaded from https://tabula-microcebus.ds.czbiohub.org. Viral data was downloaded from the NCBI: SARS-CoV-2 from France (SRP365166), SARS-CoV-2 from South Africa (SRP348159), 2020 SARS-CoV-2 from California (SRR15881549), influenza (SRP294571), and rotavirus (SRP328899).
The sample sheets used as pipeline input, including individual sample SRA accession numbers, for all analyses are uploaded to pipeline GitHub repository. Similarly, scripts to perform supplemental analysis can be found on the pipeline repository.

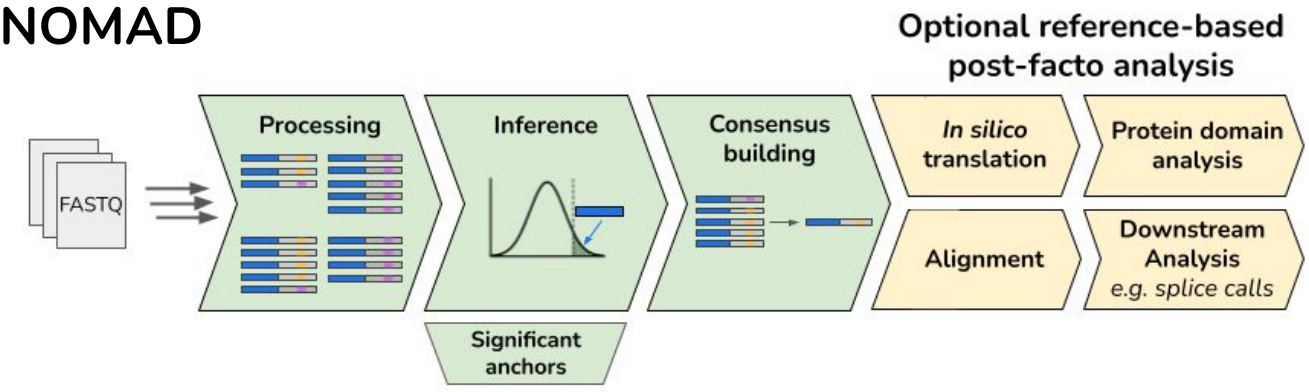
Supplementary Materials

Materials and Methods
Supplementary Text
Figs. S1 to S4
Data S1 to S4
References (37-64)

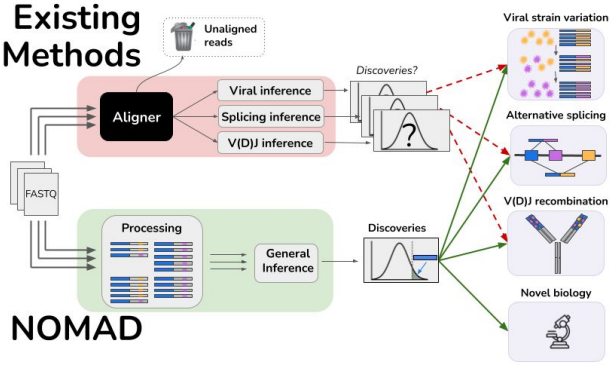
A



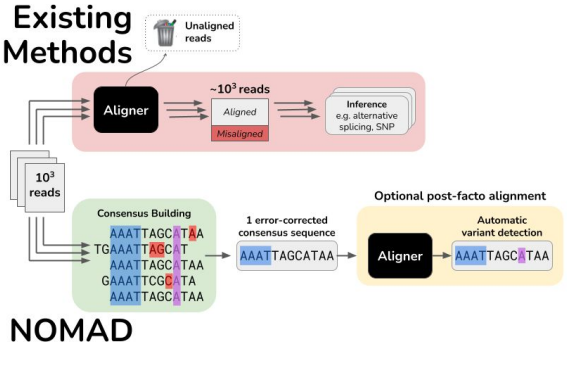
B



C



D



E

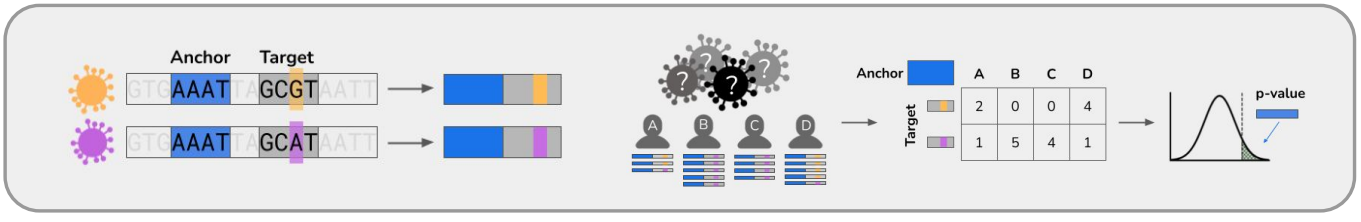


Figure 1: Overview of sample-dependent sequence diversity and NOMAD model and pipeline.

A. **Biological generality of sample-dependent sequence diversification.** The study of sample-dependent sequence diversification unifies problems in disparate areas of genomics which are currently studied with application-specific models and algorithms. Viral genome mutations, alternative splicing, and V(D)J recombination all fit under this framework, where sequence diversification depends on the sample (through cell-type or infection strain type). Myriad problems in plant genomics, metagenomics and biological adaptation are subsumed by this framework.

B. **Overview of NOMAD pipeline.** NOMAD takes as input raw FASTQ files for any number of samples >1 and processes them in parallel, counting (anchor, target) pairs per sample. NOMAD performs inference on these aggregated counts, outputting statistically significant anchors. For each significant anchor, a denoised per-sample consensus sequence is built (Fig. 1D). NOMAD also enables optional reference-based post-facto analysis. If a reference genome is available, NOMAD can align the consensus sequences to the reference, enabling denoised downstream analysis (e.g. SNPs, indels, or splice calls). In silico translation of consensuses can optionally be used to study relationships of anchors to protein domains by mapping to databases such as Pfam (Methods).

C. **Overview of NOMAD versus existing workflows.** Existing workflows (red) discard low-quality reads during FASTQ processing and alignment, only then performing statistical testing after algorithmic bias is introduced; p-values are then not unconditionally valid. Further, for every desired inferential task, a different inference pipeline must be used. NOMAD (green) performs direct statistical inference on raw FASTQ reads, bypassing alignment and enabling data-scientifically driven discovery. Due to its generality, NOMAD can simultaneously detect myriad biological examples of sample-dependent sequence diversification.

D. **NOMAD consensus building.** NOMAD constructs a per-sample consensus sequence for every significant anchor by taking all reads in which the anchor (blue) appears, and recording plurality votes for each nucleotide, denoising reads while preserving the true variant; sequencing errors in red and biological mutations in purple. Existing approaches require alignment of all reads to a reference prior to error correction, requiring orders of magnitude more computation, discarding reads in both processing and alignment, and potentially making erroneous alignments due to sequencing error. They further require inferential steps, e.g. to detect if there is a SNP or alternatively spliced variant.

E. **Example construction of NOMAD anchor, target pairs.** A stylized expository example of viral surveillance: 4 individuals A-D are infected with one of two variants (orange and purple), differing by a single basepair (orange and purple). NOMAD anchor k-mers are blue (k=4), followed by a lookahead distance of L=2, and the corresponding k-mer targets. Given sequencing reads from the 4 individuals as shown, NOMAD generates a target by sample contingency table for this blue anchor, and computes a p-value to test if this anchor has sample-dependent sequence diversity.

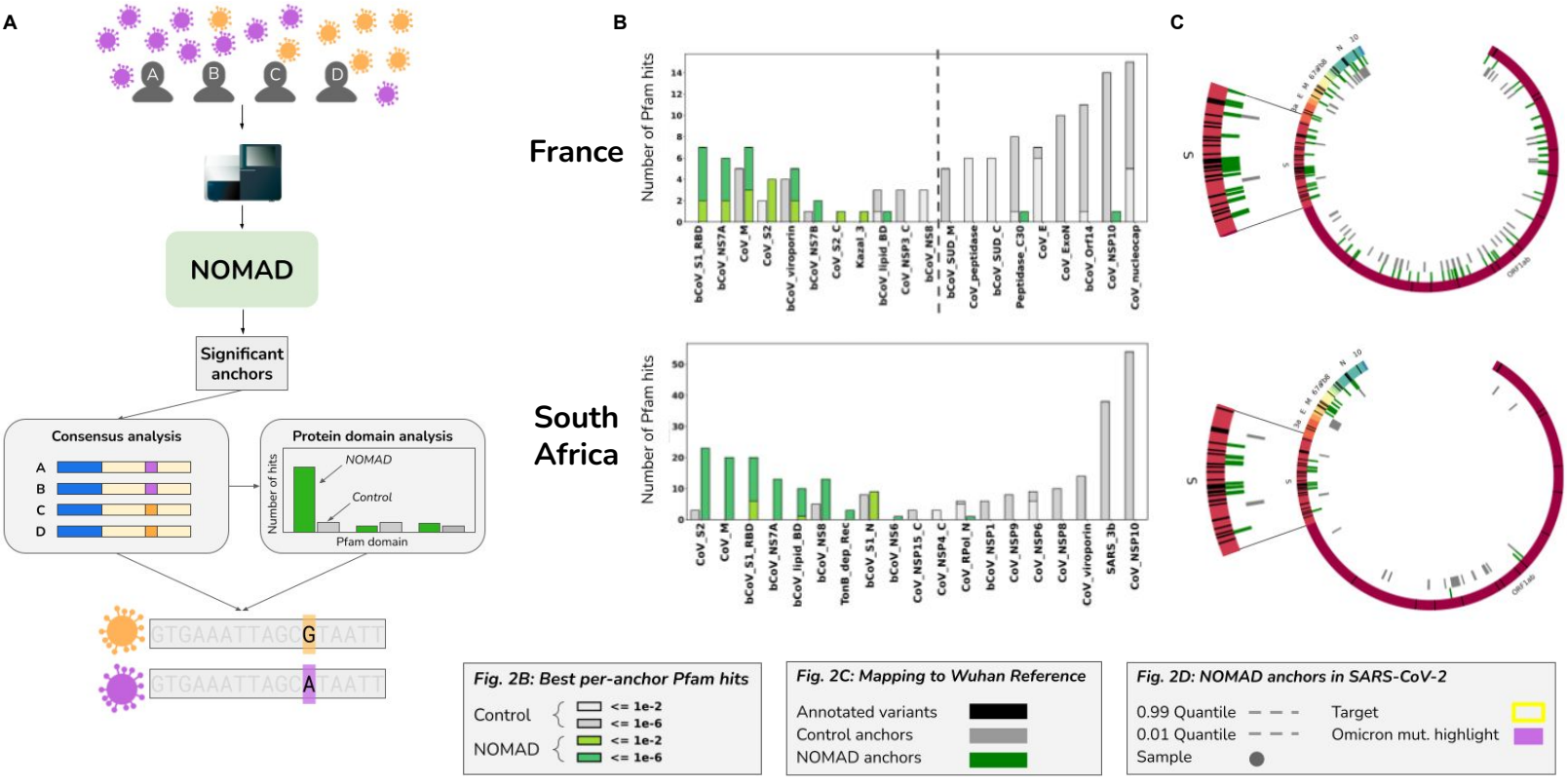


Fig. 2B: Best per-anchor Pfam hits

Control <= 1e-2 <= 1e-6
NOMAD <= 1e-2 <= 1e-6

Fig. 2C: Mapping to Wuhan Reference

Annotated variants
Control anchors
NOMAD anchors

Fig. 2D: NOMAD anchors in SARS-CoV-2

0.99 Quantile
0.01 Quantile
Sample
Target
Omicron mut. highlight

D

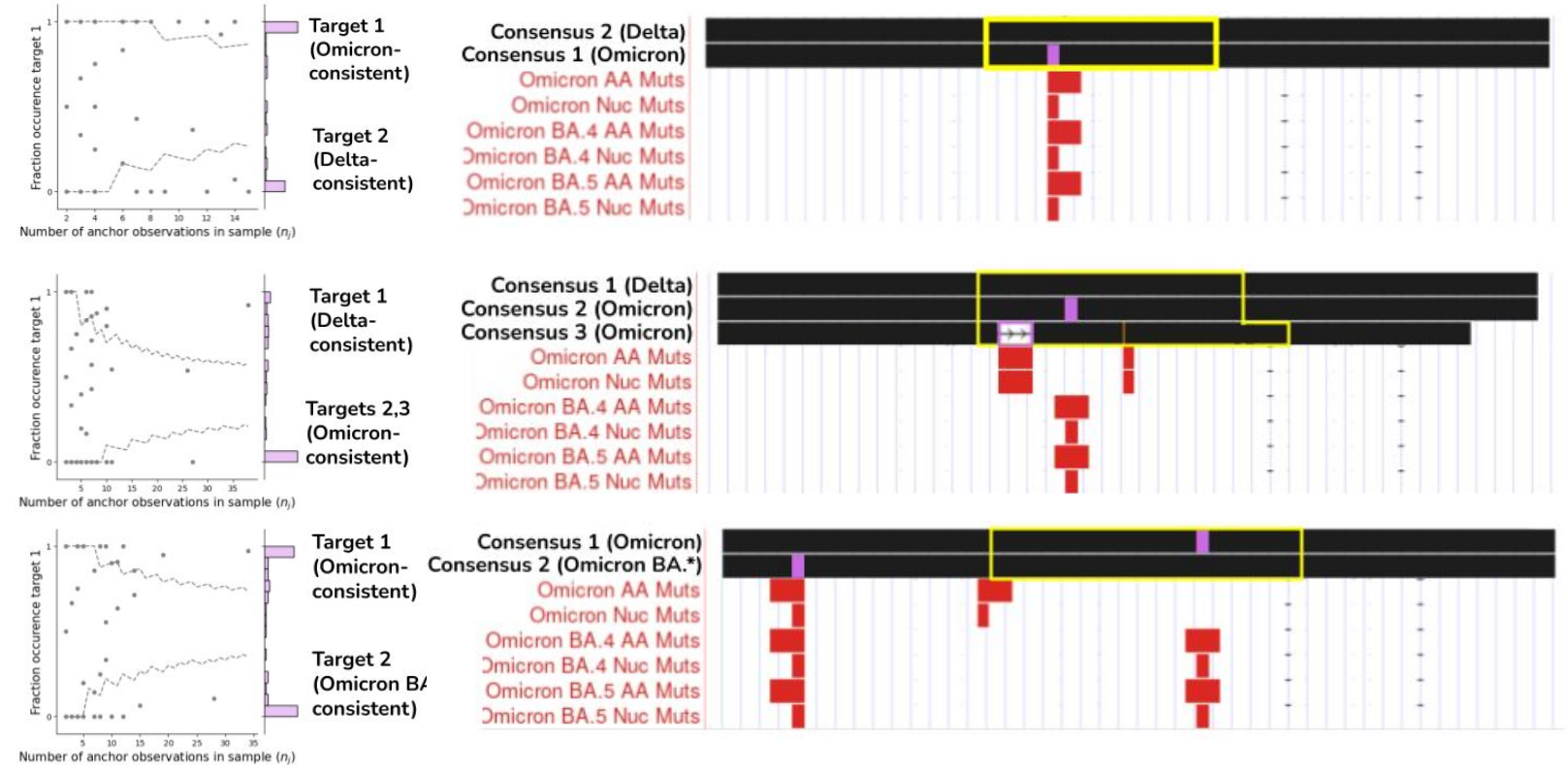


Figure 2: NOMAD analysis of SARS-CoV-2 data.

A. **Stylized example representing NOMAD workflow for viral data.** Patients with varying viral strains are sampled; two representative strains with differentiating mutations are depicted in orange and purple. NOMAD is run on raw FASTQs generated from sequencing patient samples. Significant anchors are called without a reference genome or clinical metadata. Optional post-facto analysis quantifies domain enrichment via in silico translation of consensus sequences derived from NOMAD-called anchors versus controls. Consensuses can also be used to call variants *de novo* and can be compared to annotated variants e.g. in SARS-CoV-2, Omicron.

B. **NOMAD protein profile analysis of SARS-CoV-2.** NOMAD SARS-CoV-2 protein profile hits (anchor effect size >.5) to the Pfam database (greens) and control (greys) for France and South Africa datasets; ordered by enrichment in NOMAD hits compared to control showing large distributional differences (chi-squared test p-values France: < 1.1E-12, SA: <2.5E-39). Spike protein domains are highly enriched in the NOMAD versus control. In the France data, the most NOMAD-enriched domain is the betacoronavirus S1 receptor binding domain (hypergeometric p=2.9E-4, corrected) followed by Orf7A (hypergeometric p=1.6E-3, corrected), known to directly interact with the host innate immune defense. In the South Africa data, the most enriched NOMAD profiles are CoV S2 (p=2.9E-6) and the coronavirus membrane protein (p=8.4E-8). Plots were truncated for clarity of presentation as indicated by dashed grey lines (Fig. S2A, B).

C. **NOMAD anchors are enriched near annotated variants of concern.** NOMAD anchors (effect size >.5) for SARS-CoV2 mapping to the Wuhan reference (NC_045512) show enrichment near variants of concern. SARS-CoV2 genome depicted with annotated ORFs and lines depicting positions of variants of concern (VOC) annotated as Omicron and Delta variants. No control anchor maps to spike or other areas of VOC density except in N (nucleocapsid).

D. **NOMAD consensuses identify variants of concern *de novo*.** Examples of NOMAD-detected anchors in SARS-CoV2 (France data). Scatterplots (left) show the fraction of each sample's observed fraction of target 1 (the most abundant target) for three representative anchors, binomial confidence intervals: (.01,.99), p=empirical fraction occurrence of target 1 (Supplement). y-axis shows histogram of the fraction occurrence of target 1. Mutations (right)found in the targets are highlighted in purple, BLAT shows single nucleotide mutations match known Omicron mutations. Binomial p-values of 6.8E-8, 3.1E-7, and 6.7E-15 respectively (Methods). The anchor in (top) maps to the coronavirus membrane protein; anchors in (middle and bottom) map to the spike protein. One sample (out of 26) depicted in the bottom plot has a consensus mapping perfectly to the Wuhan reference; 3 other consensuses contain annotated Omicron mutations, some designated as VOC in May of 2022, 3 months after these samples were collected.

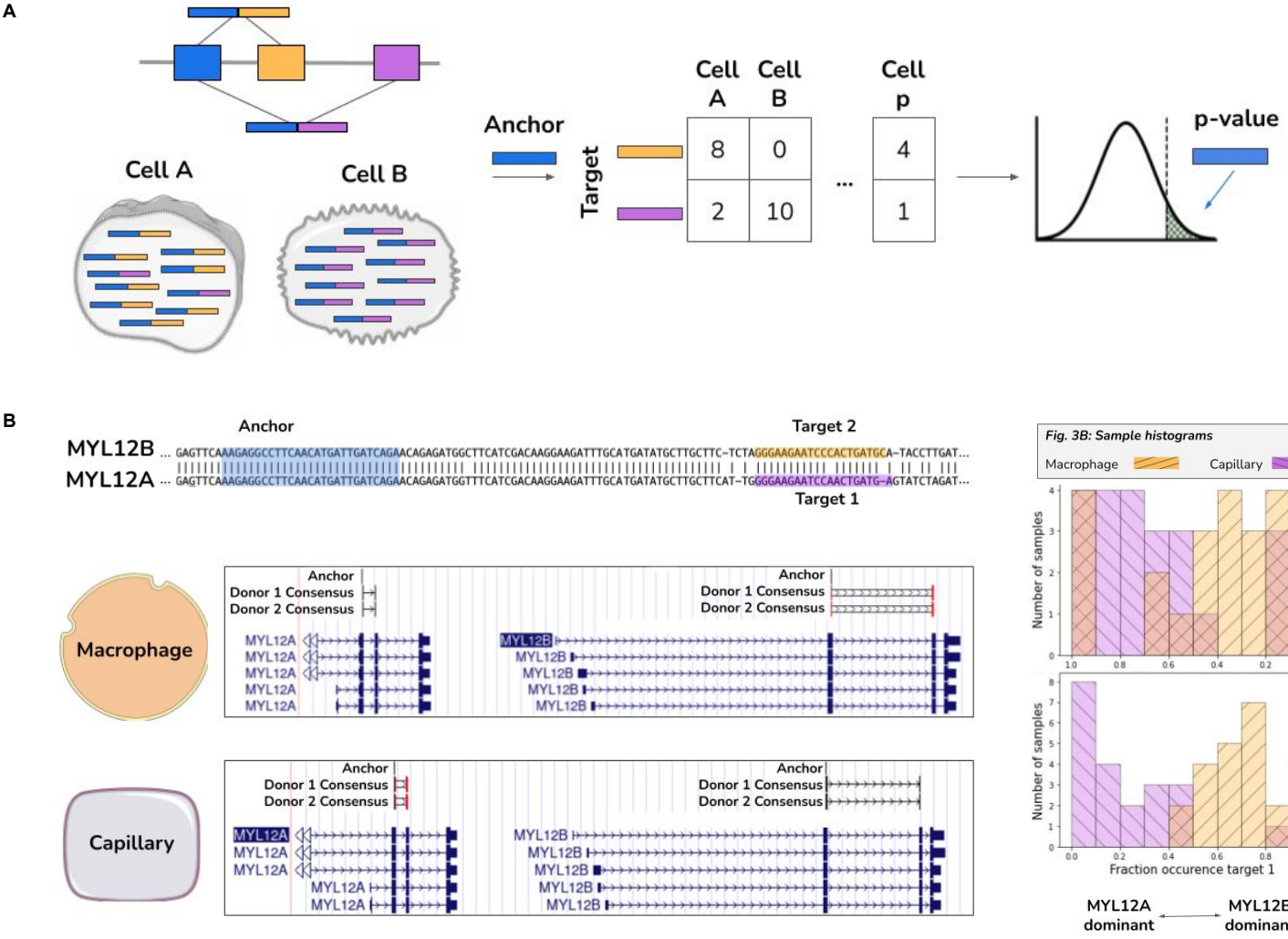


Figure 3: Detection of differentially regulated alternative splicing and isoforms from single cell RNA-seq.

A. **Stylized diagram depicting differentially regulated alternative splicing detection with 3 exons and 2 isoforms with NOMAD.** Isoform 1 consists of exon 1 (blue) and exon 2 (orange), and is predominantly expressed in cell A. Isoform 2 consists of exon 1 (blue) and exon 3 (purple) and is primarily expressed in cell B.. An anchor sequence in exon 1 (blue), then generates target sequences in exon 2 (orange) or exon 3 (purple). Counts are used to generate a contingency table, and NOMAD's statistical inference detects this differentially regulated alternative splicing.

B. **Detection of differential regulation of MYL12A/B isoforms.** (top-left) Shared anchor (q-value 2.5E-8, donor 1, 2.3E-42 for donor 2) highlighted in yellow, maps *post fact* to both MYL12 isoforms, highlighting the power of NOMAD inference: MYL12A and MYL12B isoforms share >95% nucleotide identity in coding regions. (bottom-left) NOMAD's approach automatically detects target and consensus sequences that unambiguously distinguish the two isoforms. (right) In both donors, NOMAD reveals differential regulation of MYL12A and MYL12B in capillary cells (MYL12A dominant) and macrophages (MYL12B dominant).

C. **NOMAD identifies single-cell-type regulated expression of HLA-DRB1 alleles.** NOMAD shared anchor, q-value of 4.0E-10 for donor 1, 1.2E-4 for donor 2. Scatter plots show cell-type regulation of different HLA-DRB1 alleles not explained by a null binomial sampling model $p<2E-16$ for donor 1, 5.6E-8 for donor 2, finite sample confidence intervals depicted in gray (Methods). Each (donor, cell-type) pair has a dominant target, per-cell fractions represented as "fraction target 1" in scatterplots, and a dominant consensus mapping to the HLA-DRB1 3' UTR (multiway alignment); donor 1 capillary consensus contains an insertion and deletion.

D. **Cell-type specific splicing of HLA-DPA1 in capillary versus macrophage cells.** Anchor q-value: 7.9E-22. Detected targets are consistent with macrophages exclusively expressing the short splice isoform which excises a portion of the ORF and changes the 3' UTR compared to the dominant splice isoform in capillary cells; splice variants found *de novo* by NOMAD consensuses. Binomial hypothesis test as in D for cell-type target expression depicted in scatter plots (binomial $p<2.8E-14$).

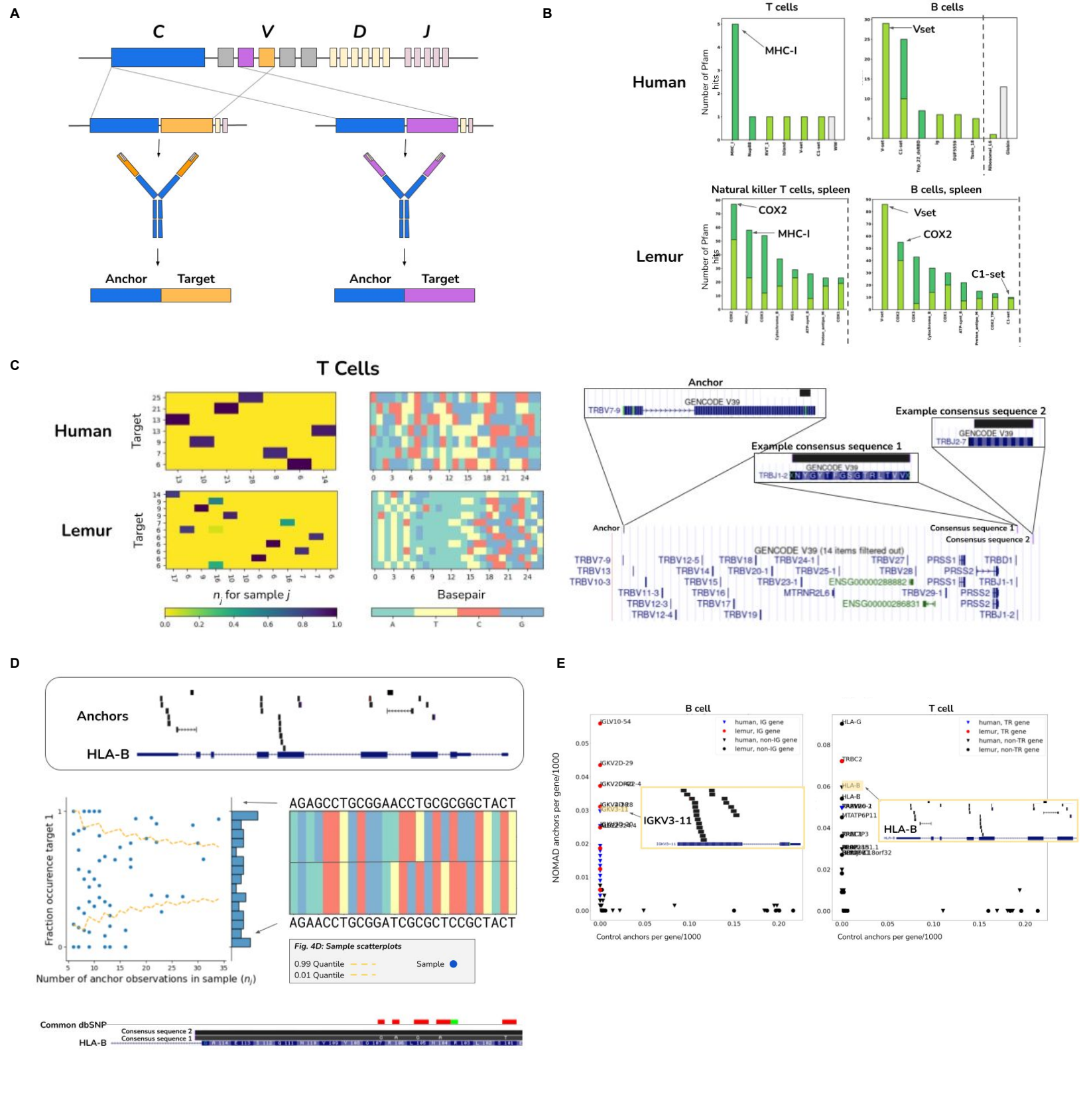


Figure 4: Unsupervised identification of V(D)J recombination in human and lemur immune cells.

A. **Stylized diagram depicting NOMAD detection of V(D)J recombination, with example variable regions in the heavy chain.** An anchor sequence in the constant region (blue), generates target sequences (orange and purple) during V(D)J recombination, in which immunoglobulins may receive different gene segments during rearrangement. NOMAD is able to rediscover and detect these recombination events by prioritizing sample-specific TCR and BCR variants.

B. **Unsupervised NOMAD protein profile analysis shows MHC and immunoglobulin variable regions are enriched in B and T cells.** NOMAD recovers domains known to be diversified in adaptive immune cells, bypassing any genome reference or alignment; control hits computed from the most abundant anchors have no such enrichment. In B cells, hits in the V set, IG like domains resembling the antibody variable region, are at a relatively high E-value, as predicted by protein diversification generated during V(D)J, making matching to reference domains imperfect. The third most hit domain is Tnp_22_dsRBD, a double stranded RNA binding domain, suggesting potential activation of LINE elements in B cells. COX2, known to be involved in immune response, is highly ranked in both lemur T and B cells. Plots were truncated for clarity of presentation as indicated by dashed grey line (Fig. S2F-H).

C. **NOMAD detects combinatorial expression of T cell receptors in immune cells *de novo*.** In human T cells (right), we show a NOMAD anchor in the TRVB7-9 gene, and two example consensuses which map to disjoint J segments, TRBJ1-2 and TRBJ2-7. Histograms of this anchor depict combinatorial single-cell (columns) by target (row) expression of targets detected by NOMAD. Histogram for lemur T cells depicted similarly; lemur T cell anchor maps to the human gene TBC1D14.

D. **NOMAD detects cell-type and allele-specific expression of HLA-B and HLA-B alleles *de novo*.** NOMAD-annotated anchors are enriched in HLA-B (top Fig. 4.D.1). Sample scatterplot (middle)Fig. 4.D.2 shows that T cells have allelic-specific expression of HLA-B, not explicable by low sampling depth (binomial test as in Fig. 3d,e described in Methods, $p < 4.6E-24$). Fig. 4.D.3: HLA-B sequence variants are identified *de novo* by the consensus approach (bottom), including allele-specific expression of two HLA-B variants, one annotated in the genome reference, the other with 5 SNPs coinciding with annotated SNPs.

E. **NOMAD analysis of lemur and human B (left) and T (right) cells recovers B, T cell receptors and HLA loci as most densely hit loci**. Human genes are depicted as triangles; lemur as circles. *Post facto* alignments show variable regions in the kappa light chain in human B cells are most densely hit by NOMAD anchors and absent from controls; in T cells, the HLA loci and TRB including its constant and variable region are most densely hit, which are absent from controls. x-axis indicates the fraction of the 1000 control anchors (most abundant anchors) that map to the named transcript, y-axis indicates the fraction of NOMAD's 1000 most significant anchors that map to the named transcript. Each inset depicts anchor density alignment in the IGKV region (left) and HLA-B in CD4+ T cells (top right) and TRBC-2 (bottom right), showing these regions are densely hit.

Supplementary Materials for

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

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

Correspondence to: julia.salzman@stanford.edu

This PDF file includes:

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

Other Supplementary Materials for this manuscript include the following:

Data S1 to S4

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

Materials and Methods

Anchor preprocessing

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

NOMAD p-values

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

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

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

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

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

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

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

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

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

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

NOMAD Effect size

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

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

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

Consensus sequences

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

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

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

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

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

Element annotations

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

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

Genome annotations

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

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

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

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

NOMAD protein profiles

For each set of enriched anchors, homology-based annotation was attempted against an annotated protein database, the Pfam (*18*). For each dataset, up to 1000 of the most significant anchors (q-value < 0.01) were retained for the following analysis: we first generated a substring

of each downstream consensus by appending each consensus nucleotide assuming both conditions were met: a minimum observation count of 10 and a minimum agreement fraction of 0.8, until whichever metric first exhibited two consecutive failures at which point no further nucleotide was added. A limit of 1000 anchors was used due to computational constraints from HMMer3 (see below). Anchors that did not have any consensus nucleotides appended were kept as is. An extended anchor was generated for each experiment in which an anchor was found. Each extended anchor was then stored in a final concatenated multi FASTA file with unique seqID headers for each experiment's extended anchors.

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

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

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

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

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

Control analysis

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

SARS-CoV2 analysis

The SARS-CoV2 datasets used in this manuscript were analyzed with NOMAD's unsupervised mode (no sample metadata provided). To identify high effect size anchors, a

threshold of `effect_size_randCjs` > 0.5 was used (table S2). The Wuhan variant reference genome was downloaded from NCBI, assembly NC_045512.2. The Omicron and Delta mutation variants were downloaded as FASTA from the UCSC track browser in June 2022, with the following parameters: clade 'Viruses', assembly NC_045512.2, genome 'SARS-CoV-2', group 'Variations and Repeats', track 'Variants of Concert', and table 'Omicron Nuc Muts (variantNucMuts_B_1_1_529)' and 'Delta Nuc Muts (variantNucMutsV2_B_1_617_2)'.

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

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

Viral protein profile analysis

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

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

Identifying cell-type specific isoforms in SS2

In the analysis of HLCA SS2 data, we utilize "isoform detection conditions" for alternative isoform detection. These conditions select for (anchor, target) pairs that map exclusively to the human genome, anchors with at least one split-mapping consensus sequence, mu_lev > 5, and M > 100. mu_lev is defined as the average target distance from the most

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

Splice junction calls

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

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

HLA analysis in HLCA

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

B,T Cell Transcriptome Annotations

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

Further immune cell protein domain analysis

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

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

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

NOMAD comparison to BASIC analysis in lemur spleen B cells

To compare performance, we first ran BASIC on the lemur spleen B cells, with the following additional parameters: -a. We then ran NOMAD on cells where BASIC failed to identify the light chain variable gene family, by selecting cells annotated as "No BCR light chain" from the BASIC output. From the NOMAD output, we identified anchors which mapped

to the IGL gene by bowtie; to do this, we used the command `grep IGL "$file"`, where "$file" corresponds to the NOMAD anchor genome annotations output file. This resulted in the following 5 anchors: CCTCAGAGGAGGGCGGGAACAGCGTGA, CTCGGTCACTCTGTTCCCGCCCTCCTC, GCCCCCTCGGTCACTCTGTTCCCGCCC, GGGCGGGAACAGCGTGACCGAGGGGGC, TCACTCTGTTCCCGCCCTCCTCTGAGG.

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

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

Timing for SS2

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

Laptop analysis details

Laptop specs:

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

Dataset:

10 files of SS2 T cell reads
43,870,027 reads total

Figure data

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

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

Supplementary Text

Generality of NOMAD

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

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

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

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

Generality of NOMAD anchor, target and consensus construction

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

Statistical Inference

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

Probabilistic model

Formally, we study the null model posed below.

Null model:

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

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

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

Test intuition

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

One natural choice of f is constructed to capture the intuition that target diversity is most interesting when target sequences are highly divergent. To define f, i) targets are ranked by abundance; ii) the i-th target is assigned a scalar value measuring its minimum distance (such as Hamming, Levenstein) to all more abundant targets. Note that in order to ensure that this inference is statistically valid, we need to split the data and measure abundance on a subset of data that we do not use for downstream processing (to avoid data snooping). This function has

some power to identify sample-dependent splicing, but little power to discriminate SNPs in targets. This is because, as these scores will be aggregated over the targets of a given sample, we see that in this example all samples that express the primary isoform will have an average target function value close to 0, whereas the alternatively spliced samples will have large target function values. However, such a function f has a major drawback; it is not able to fully utilize the dynamic range of this function. Since our procedure is scale invariant it suffices to consider f bounded between 0 and 1, and so we need to normalize by the maximum value of f that can be observed, which is k=27. This can be problematic, as seen by an example where the spliced target is a distance of 5 away, leaving its value at 5/27 instead of 1. To this end, we instead appeal to the probabilistic nature of our problem, and utilize several independent random functions f. That is to say, each random function f we utilize assigns a value of 0 or 1 independently to each target, fully utilizing the available dynamic range, and extending our detection power beyond SNPs.

p-value computation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Following this approach, we have that

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

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

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

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

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

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

Effect size

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

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

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

Ability to operate without metadata

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

p-value computation for scatterplots depicting target fraction abundance

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

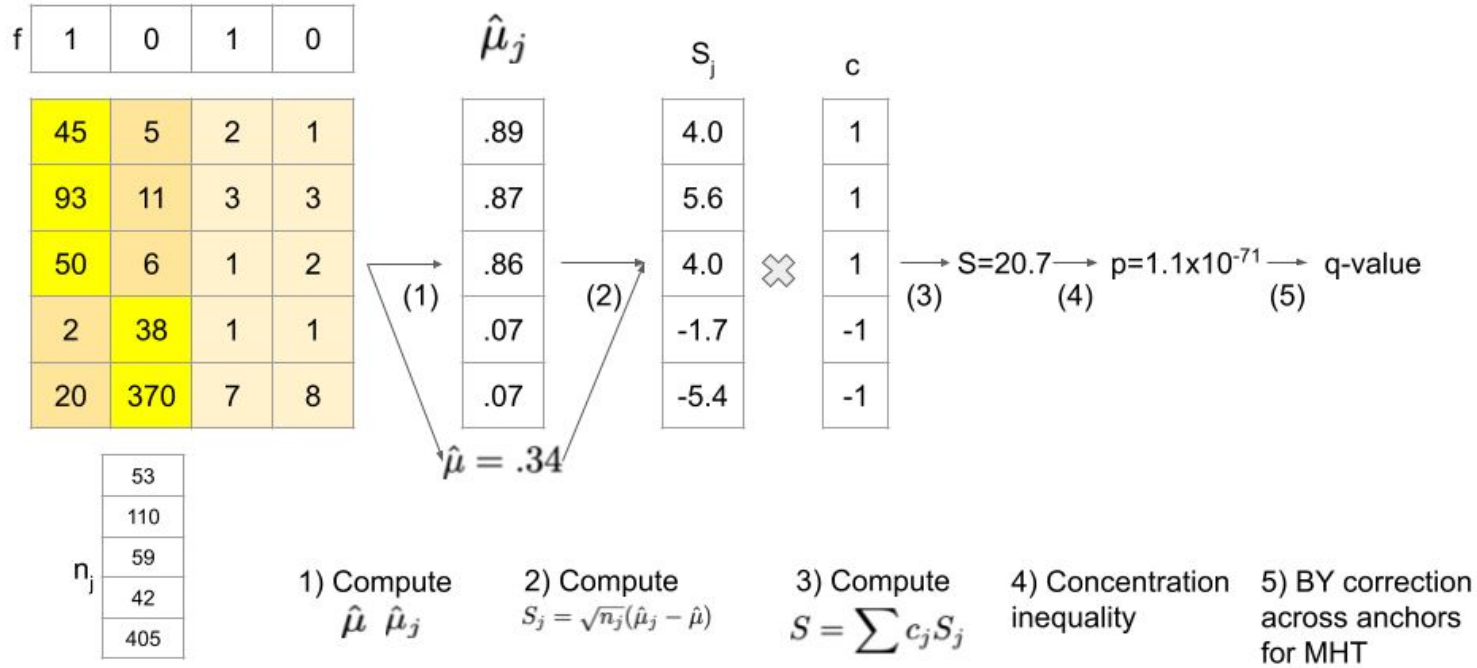
While intuitive, the above analysis is loose. Firstly, since binomials are discrete distributions, we will rarely be able to compute exact 1% and 99% quantiles. Thus, the probability that for any given n_j a sample will fall outside of the [1,99] quantiles, which we denote p_j, is almost always substantially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

Hypergeometric p-value computation

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

S1

A



B

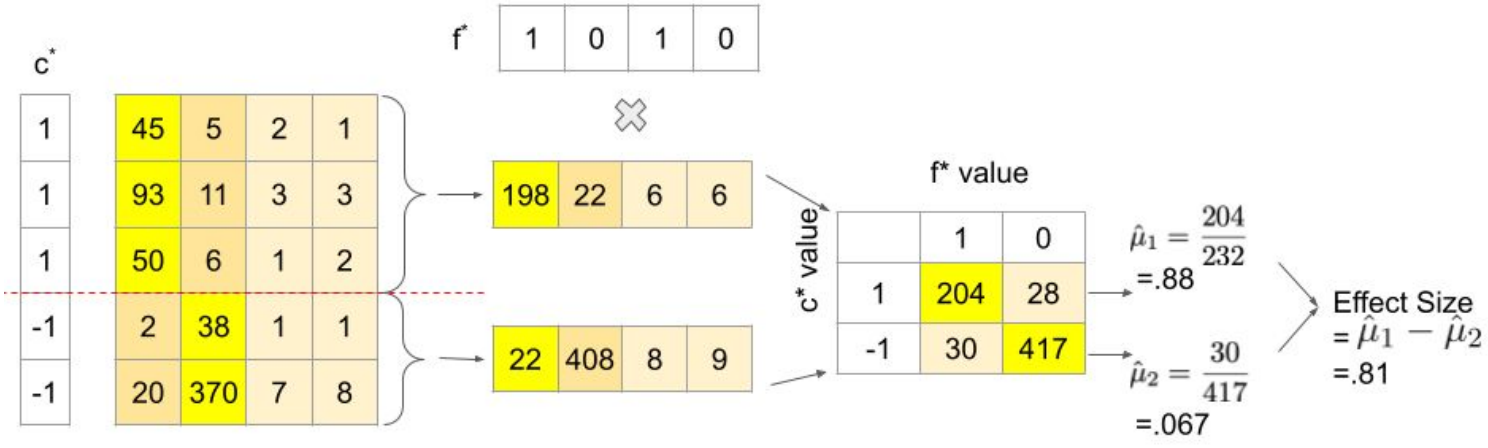
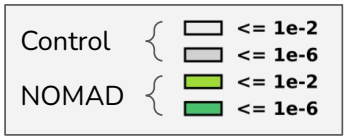


Fig. S1: NOMAD Statistics.

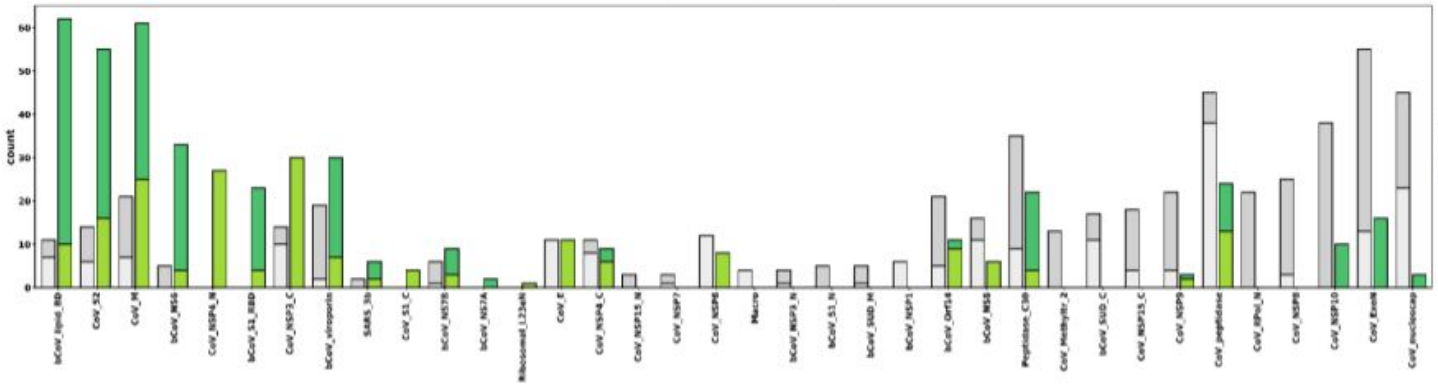
A. p-value computation for NOMAD. Contingency table transposed for visual convenience (rows are samples and columns are targets). Starting with a samples by targets counts matrix, NOMAD utilizes one (or several) functions f mapping targets to values within [0,1]. The mean with respect to f is taken over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

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

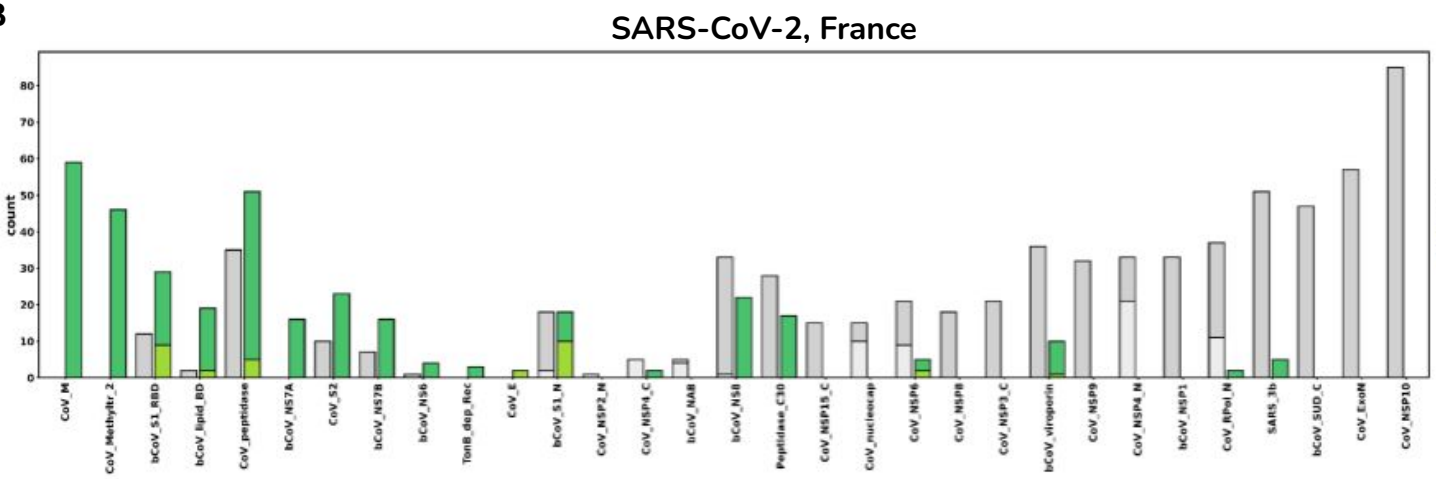
S2



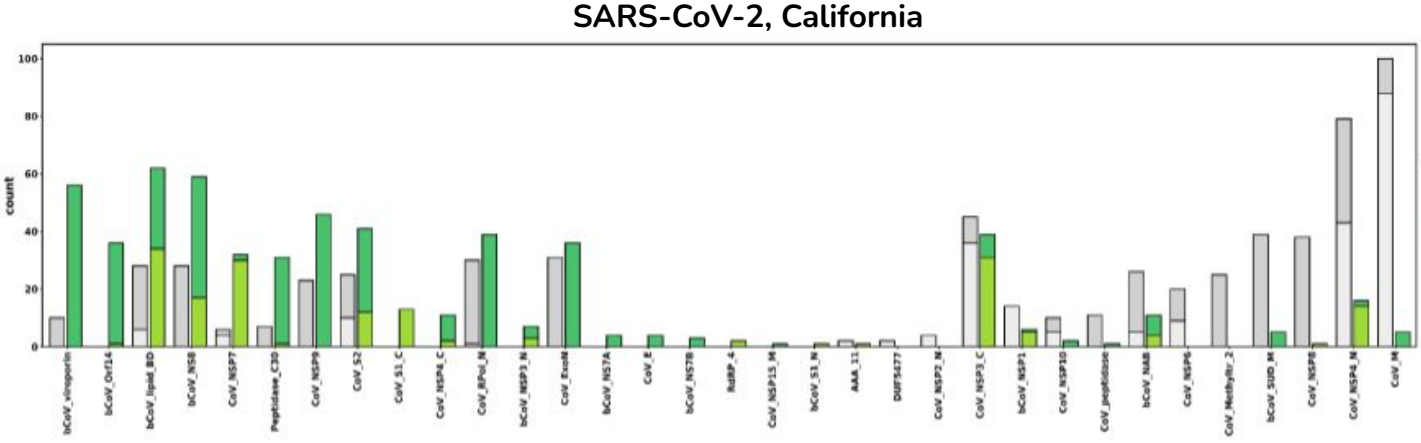
A



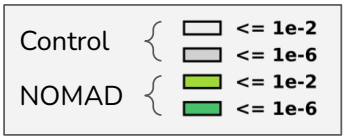
B



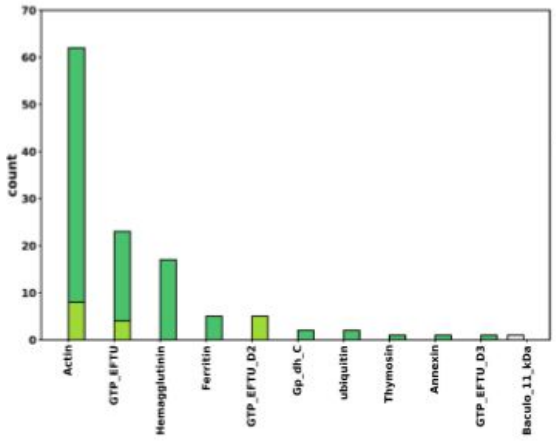
C



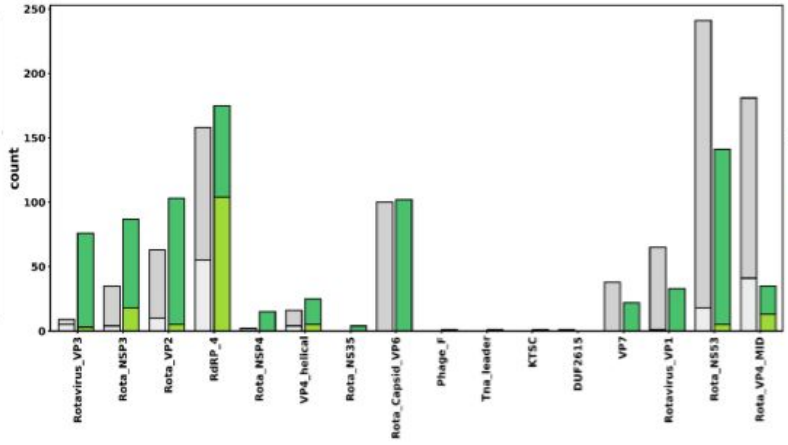
S2



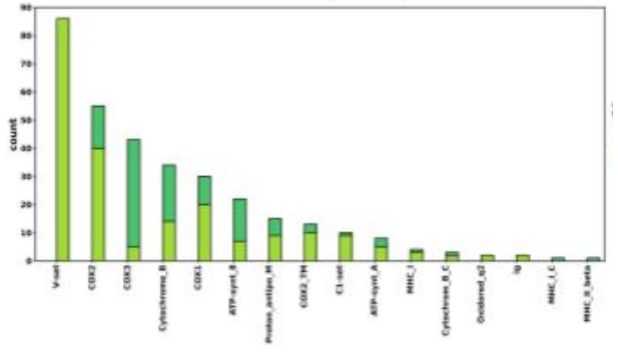
D Influenza



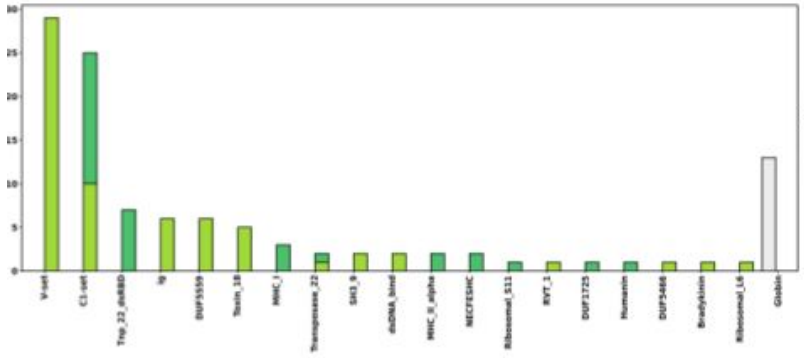
E Rotavirus



F Lemur B Cell



G Human B Cell



H Lemur T Cell

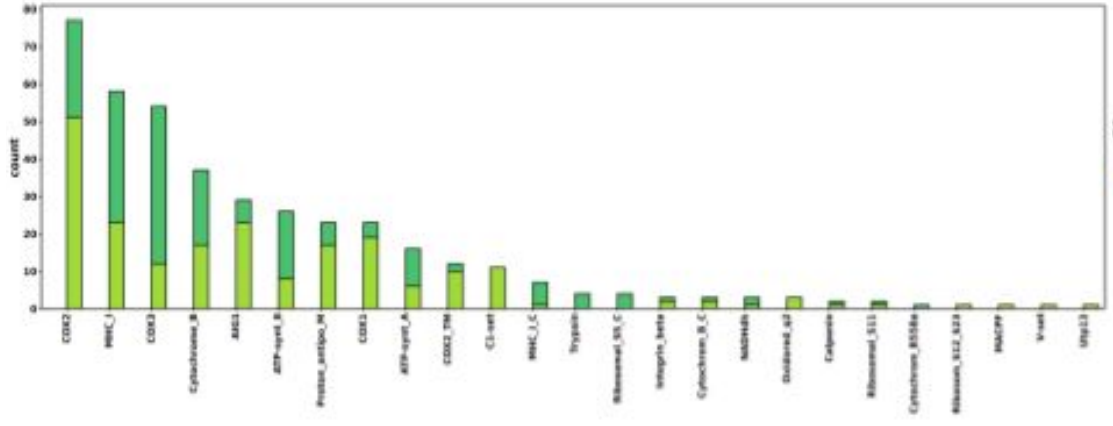


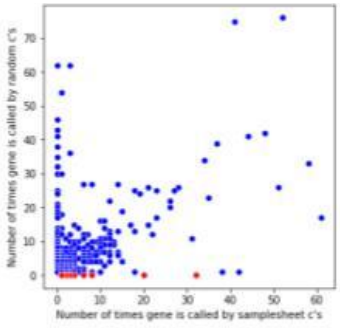
Fig. S2: NOMAD protein profile hits to the Pfam database (greens) and control (greys); ordered by enrichment in NOMAD hits compared to control; all NOMAD anchors were used as input, without effect size filters.

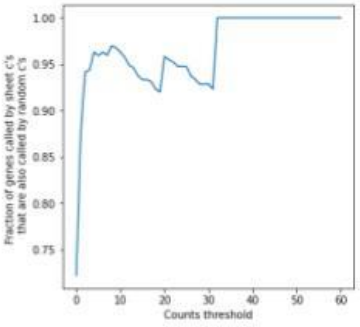
A. Protein profile analysis of NOMAD significant anchors from the original South African genomic surveillance study (SRP348159) that identified the Omicron strain during the period 2021-11-14 to 2021-11-23 (*19*)
B. Protein profile analysis of NOMAD significant anchors from France data, Oropharyngeal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

S3

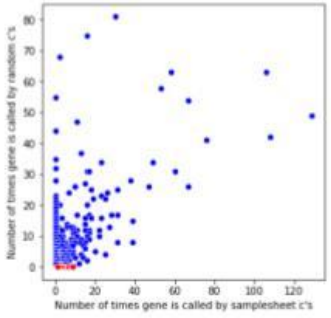
A

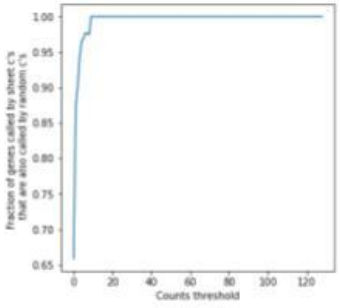
Donor 1



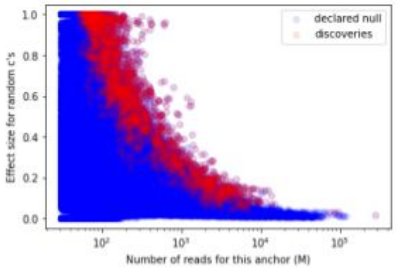


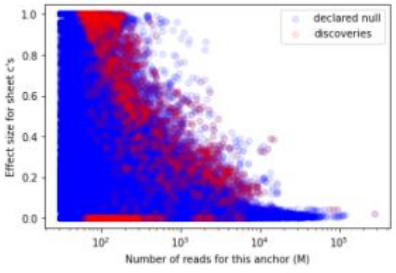
Donor 2

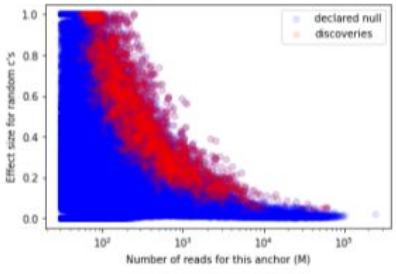




B







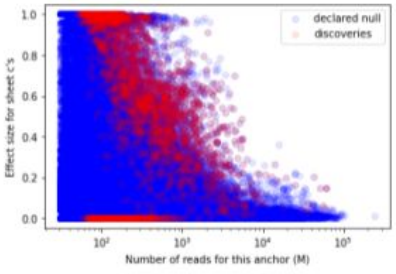


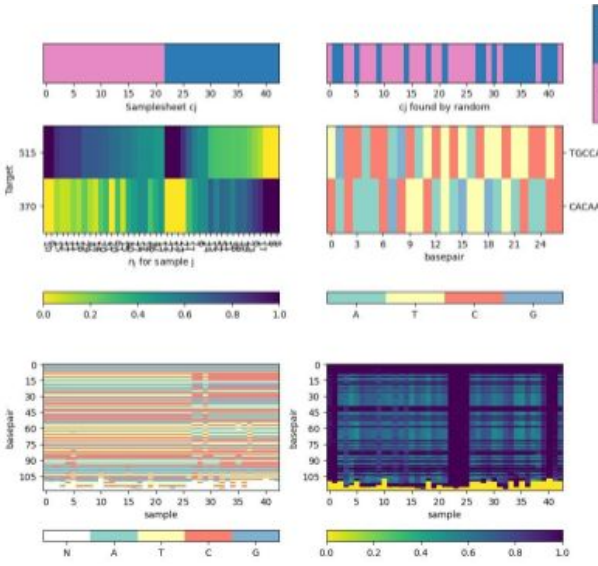
Fig. S3: Analysis of significant anchors in HLCA.

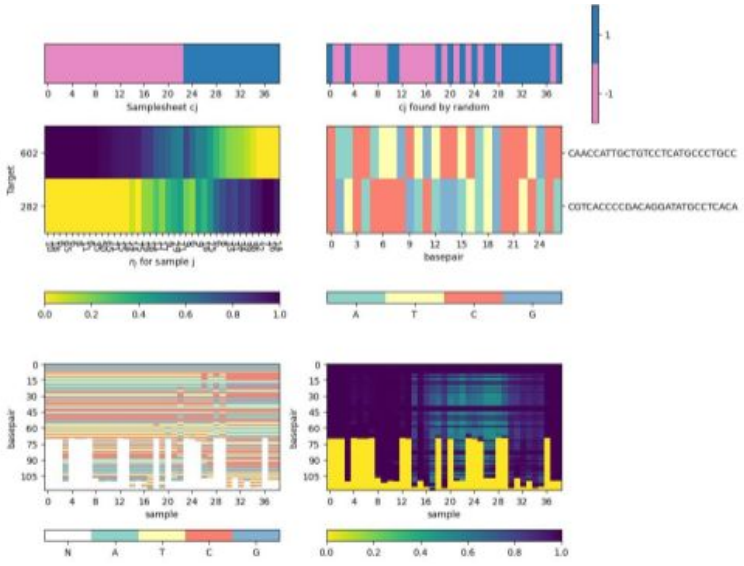
A. Random c's can recover samplesheet c's. For the HLCA dataset, of the 3439 anchors (1384 genes) called by the input metadata (samplesheet c's) in donor 1 (BY correction, alpha=.05), we have that 72% of the genes called were also called by NOMAD's selection of random c's (6287 called by anchors by random c's, 2268 genes). Left plot indicates for each gene (dot) how many times it was called by samplesheet c's vs random c's. Red dots indicate those genes not called by random c's. On the right plot we have the fraction of genes that are called at least x times by samplesheet c's that are also called by random c's. We see that for x=2 (i.e. all genes hit by at least 2 anchors), random c's call >94% of those genes called by samplesheet c's.
   For donor 2 similar results are observed, with 3775 (5619) anchors from samplesheet c's and 1125 (1844) genes for samplesheet c's (random c's) respectively. >90% of samplesheet c discoveries for x=2, >94% for x=3.
B. Effect size plotted against number of reads for HLCA dataset for donor 1 (top row) and donor 2 (bottom row), macrophage (left) and capillary cells (right).

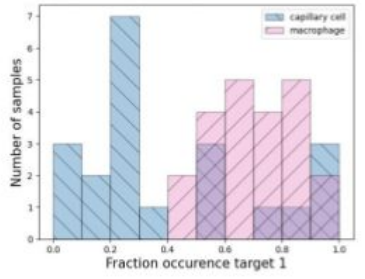
S4

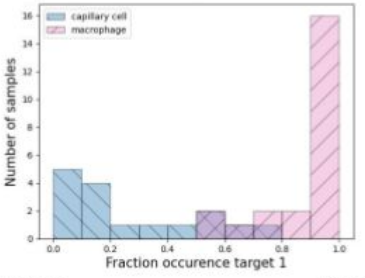
A

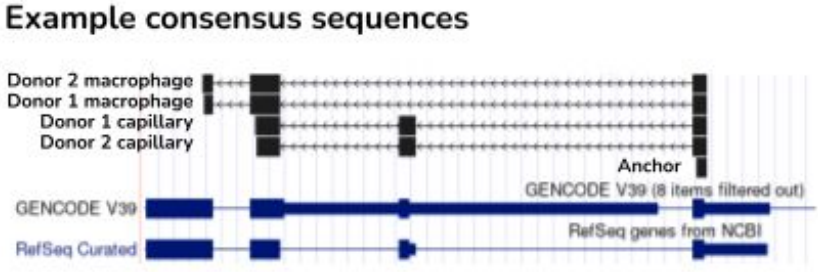
MYL6





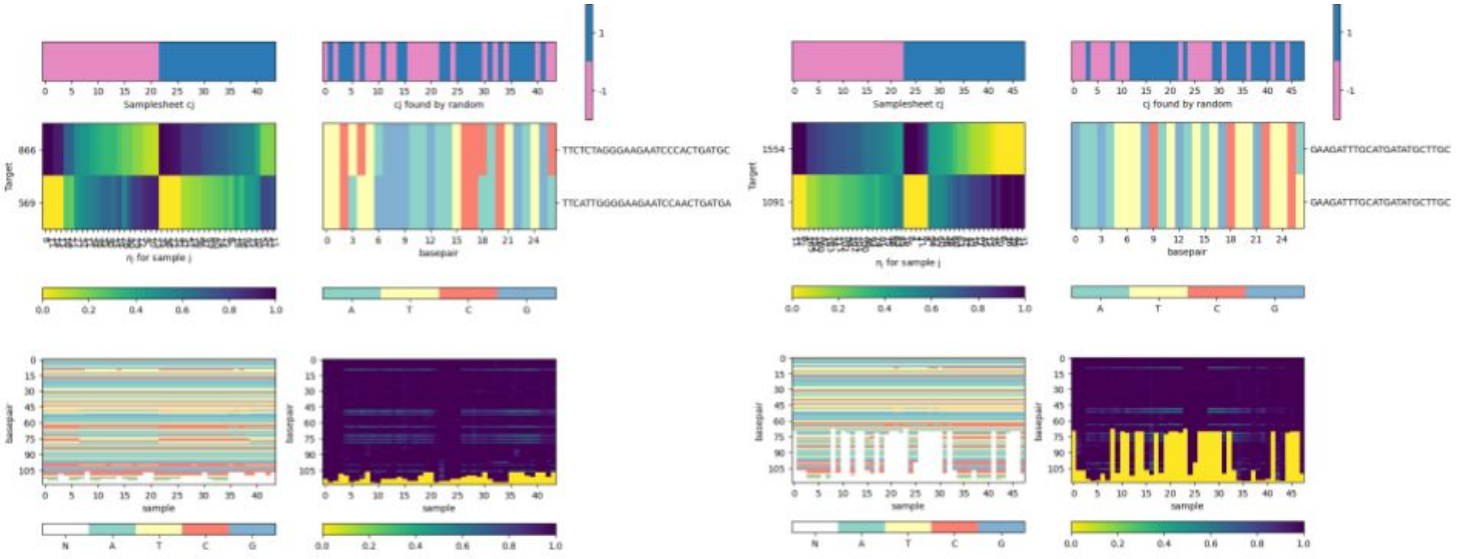






S4

B MYL12



C HLA-DPB1

D Human T Cell, HLA-B

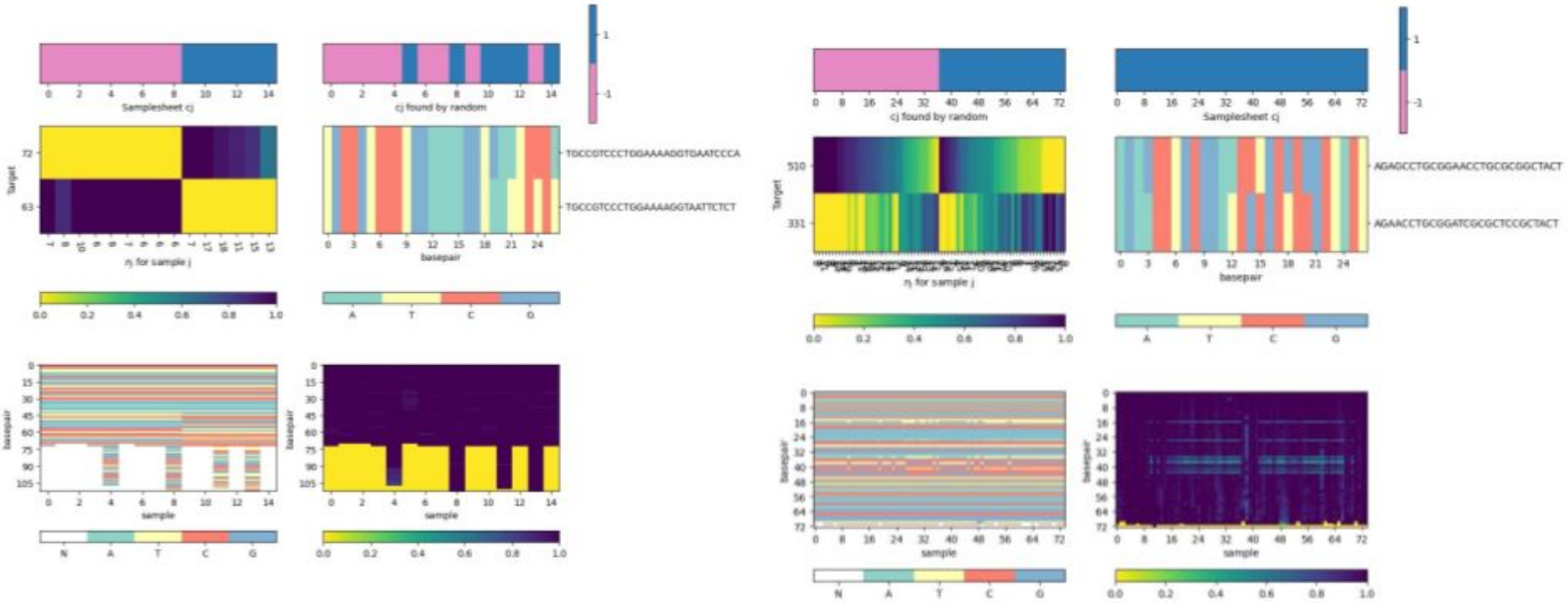


Fig. S4: Sample heatmaps.

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

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

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

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

B. MYL6
C. MYL12
D. HLA-DPB1
E. Human T cell, HLA-B

Data S1: Protein domain analysis

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

Data S2: Significant anchors

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

Data S3 : Additional summary tables

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

Data S4: Anchor genome annotations

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