Disease diagnostics using machine learning of immune receptors

Maxim E. Zaslavsky[1], Nikhil Ram-Mohan[2], Joel M. Guthridge[3], Joan T. Merrill[3], Jason D. Goldman[4,5], Ji-Yeun Lee[6], Krishna M. Roskin[7,8], Charlotte Cunningham-Rundles[9], M. Anthony Moody[10,11,12], Barton F. Haynes[11,12,13], Benjamin A. Pinsky[6,14], James R. Heath[15,16], Judith A. James[3], Samuel Yang[2], Catherine A. Blish[14], Robert Tibshirani[17,18], Anshul Kundaje[1,19,*,§], Scott D. Boyd[6,*,§]

[1] Department of Computer Science, Stanford University
[2] Department of Emergency Medicine, Stanford University
[3] Department of Medicine, Arthritis and Clinical Immunology, Oklahoma Medical Research Foundation
[4] Swedish Center for Research and Innovation, Swedish Medical Center
[5] Division of Allergy and Infectious Diseases, University of Washington
[6] Department of Pathology, Stanford University
[7] Department of Pediatrics, University of Cincinnati, College of Medicine
[8] Divisions of Biomedical Informatics and Immunobiology, Cincinnati Children's Hospital Medical Center
[9] Icahn School of Medicine at Mount Sinai
[10] Department of Pediatrics, Duke University
[11] Duke Human Vaccine Institute, Duke University
[12] Department of Immunology, Duke University
[13] Department of Medicine, Duke University
[14] Department of Medicine, Stanford University
[15] Institute for Systems Biology
[16] Department of Bioengineering, University of Washington
[17] Department of Statistics, Stanford University
[18] Department of Biomedical Data Science, Stanford University
[19] Department of Genetics, Stanford University
[§] These authors contributed equally
* Correspondence: akundaje@stanford.edu (A.K); publications_scott_boyd@stanford.edu (S.D.B)

Abstract

Clinical diagnoses rely on a wide variety of laboratory tests and imaging studies, interpreted alongside physical examination and documentation of symptoms and patient history. However, the tools of diagnosis make little use of the immune system's internal record of specific disease exposures encoded by the antigen-specific receptors of memory B cells and T cells. We have combined extensive receptor sequence datasets with three different machine learning representations of the contents of immune repertoires to develop an interpretive framework, *MAchine Learning for Immunological Diagnosis (Mal-ID)*, that screens for multiple illnesses simultaneously. This approach can already reliably distinguish a wide range of disease states, including specific acute or chronic infections, and autoimmune or immunodeficiency disorders, and could contribute to identifying new infectious diseases as they emerge. Importantly, many features of the model of immune receptor sequences are human-interpretable. They independently recapitulate known biology of the responses to infection by SARS-CoV-2 or HIV, and reveal common features of autoreactive immune receptor repertoires, indicating that machine learning on immune repertoires can yield new immunological knowledge.

Main text

Modern medical diagnosis relies heavily on laboratory testing for cellular or molecular abnormalities in specimens from a patient, or the presence of pathogenic microorganisms[1,2]. For autoimmune disorders like lupus or multiple sclerosis, diagnosis via a combination of clinical or imaging abnormalities, detection of autoantibodies, and exclusion of other conditions is a lengthy process that can delay treatment[3,4]. Evolution has provided vertebrate animals with immune systems that carry out molecular surveillance for abnormal exposures, using B cells and T cells expressing diverse, randomly generated antigen receptors. In response to viruses, vaccines, and other exposures, the repertoire of B and T cell receptors changes in composition, due to clonal expansion of stimulated cells, introduction of additional somatic mutations into B cell receptor genes, and selection processes that further reshape the immune cell populations. In disregulated immunity, self-reactive lymphocytes can also clonally proliferate and cause immunological pathologies.

Being able to interpret the specificities encoded in a patient's adaptive immune system could allow assessment for many infectious diseases at once, as well as offering insight into autoimmune reactions[5–7]. Tracking immune receptor repertoires has already proved useful in diagnosing lymphocyte malignancies and monitoring cancer treatment responses[8,9]. But other clinically relevant information in immune repertoires in the wide variety of infectious disease, autoimmunity, and immunodeficiency disorders has remained largely inaccessible. At issue is the high variability of immune receptor genes due to somatic rearrangement. To overcome this challenge, we hypothesized that the B cell immune response to each disease exhibits distinct and systematic patterns across people that can be identified with machine learning techniques for biological sequence data.

Using systematically collected datasets of B cell receptor (BCR) heavy chain (IgH) sequences from peripheral blood, we identify the presence of infectious and immunological diseases by developing and combining three machine learning models (**Figure 1**). Previous modeling attempts relied on nearly identical receptor sequence segments as a sign that two people have the same disease[10–13]. Instead, we group individuals by inferring broader functional similarities in their immune receptors. We also detect other shared characteristics of immune responses: the extent of class switching of antibody constant regions, degree of somatic mutation diversification of BCR repertoires, and effects of selection in distorting quantitative features like immunoglobulin heavy chain complementarity determining region 3 (CDR-H3) length. The machine learning process distinguishes healthy from diseased individuals, viral infections from autoimmune or immunodeficiency conditions, and different pathogen infections from each other — without prior knowledge of pathogenesis.

This approach also generates interpretable rankings for disease-specific sequences, revealing that the classifier recapitulates independently discovered biological facts, including identifying neutralizing antibodies to SARS-CoV-2. It also pinpoints V(D)J rearrangement gene segment and receptor sequence length patterns in different diseases that are less well characterized in the literature.

Integrated repertoire models of disease states

Even in patients with an active infectious disease, only a fraction of immune receptors may be devoted to the responsible pathogen. To determine an individual's immune status from BCR sequencing, a diagnostic algorithm must sift through hundreds of thousands of unique sequences to identify the rare specific ones[14]. Candidate disease-specific BCRs can be highly variable across individuals. Exact T cell receptor sequence or motif matches have previously been used to identify individuals exposed to the same disease[10,15], but B cell receptors show greater sequence diversity due to somatic hypermutation during B cell stimulation.

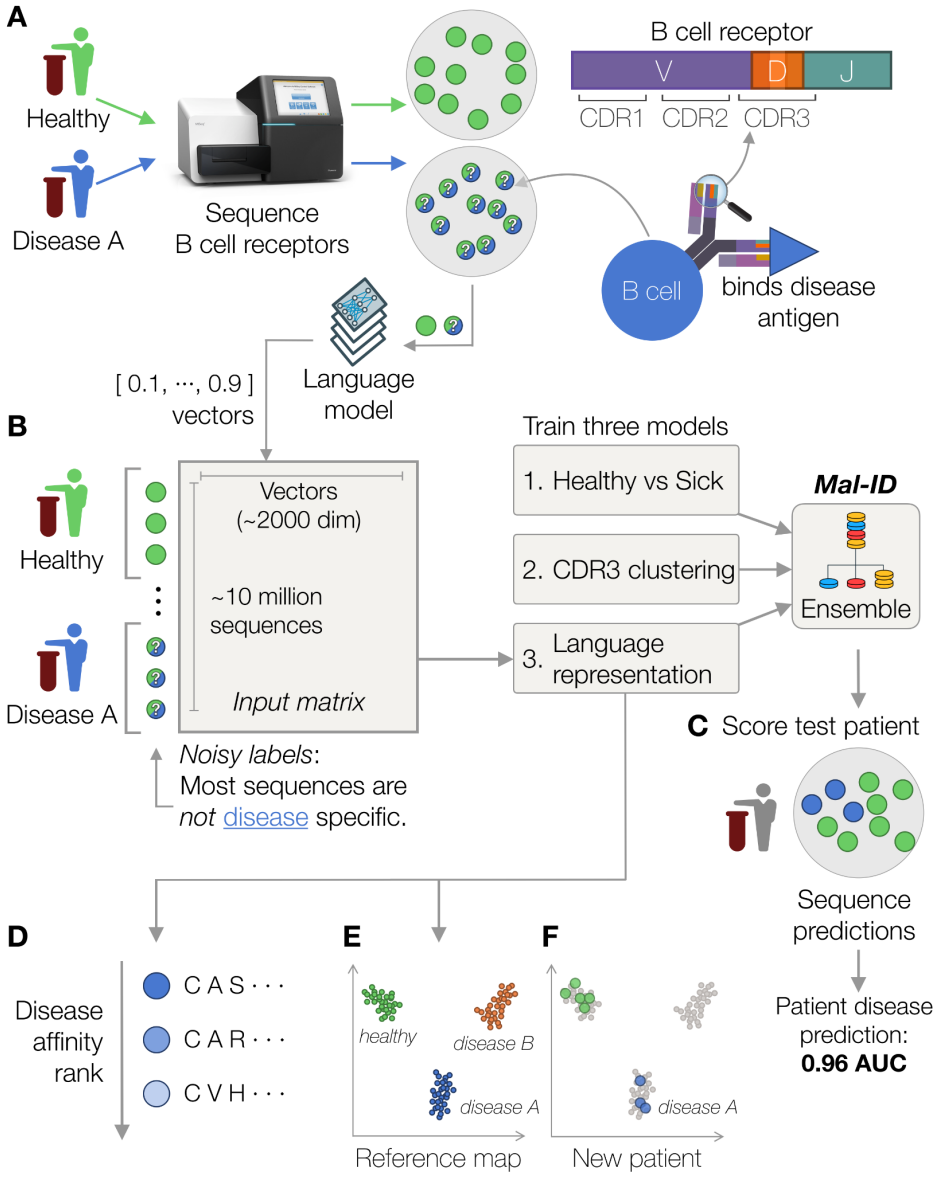


Figure 1: MAchine Learning for Immunological Diagnosis (*Mal-ID*) framework schematic.
(a) Data collection: we sequenced B cell receptor repertoires from individuals at many disease states. Each B cell is one V(D)J gene rearrangement.
(b) We extracted sequence features with an immune language model. Each raw sequence was converted into a numerical vector. The language model can group related sequences between people. Then we trained an ensemble of machine learning models to predict disease from immune receptor sequence features.
(c) We applied the ensemble of models to predict disease status of held-out test individuals.
(d-f) We interpreted the language representation sequence model:
(d) We found suspected antigen-specific immune receptors by ranking sequences according to their predicted disease specificity.
(e) We constructed a reference map of immune receptor sequences. Each point is one sequence.
(f) We visualized a held-out test patient's immune status over time by overlaying their sequences on the reference map. The immune response to disease is visible in blue.

Here, we use a combination of three models to improve recognition of distinct kinds of disease states, and to identify similar receptor sequences selected for binding to disease-related antigens. Each classifier model extracts different aspects of immune repertoires (**Figure 1B**). The first model predicts whether a person's IgH repertoire differs from those seen in healthy control individuals without recent immune system stimulation, providing a means of detecting active immune response — such as those in illness or vaccination. The second predictor identifies groups of highly similar sequences across individuals. The third classifier evaluates a broader proxy for functional similarity, rather than direct sequence identity, to find more loosely related immune receptors with common antigen targets. The three models are then blended into a final prediction of immune status. The final trained program accepts an individual's collection of IgH sequences from peripheral blood B cells as input, and returns a prediction of the probability the person has each disease on record (**Figure 1C**).

We applied this approach to cohorts of patients with diagnoses of Covid-19 (n=75), HIV (n=96)[11], Systemic Lupus Erythematosus (SLE, n=17), and Common Variable Immunodeficiency (CVID, n=13)[16], and healthy controls (n=154) — 355 individuals in total. We combined new datasets with ones previously reported, all collected with a standardized protocol of IgH amplicon sequencing from peripheral blood B cells. The non-

Covid-19 cohorts were collected before the emergence of SARS-CoV-2. To evaluate whether our proposed strategy can generalize to new immune repertoires, patients were strictly separated into training and testing sets; we trained separate models for each cross-validation group and report averaged classification performance. As described below, we also tested and excluded the possibility that demographic differences between cohorts could explain diagnosis accuracy. Details of the three models follow:

Global immune status prediction: The first machine learning model predicts whether an individual's IgH repertoire composition differs from healthy blood donors. Previously, we identified class-switched IgH sequences with low somatic mutation (SHM) frequencies in acute Ebola or Covid-19 cases, consistent with naive B cells recently having class-switched during the response to infection[17–19]. We trained a lasso linear model with each isotype's proportion of mutated sequences and IGHV gene/IGHJ gene counts as features. These features may also represent BCR repertoire changes accumulated in chronic conditions. Certain IGHV gene segments may be more prevalent among antigen-responding V(D)J rearrangements than the general population of immune receptors. As antigen-binding B cells become clonally expanded, the distribution of IGHV gene usage across the repertoire can change[17,18]. Other groups have piloted immune status classification using deviations in V(D)J recombination gene segment usage from healthy baseline[20,21]. The classifier distinguished 201 Covid-19, HIV, SLE, and CVID patients from 154 healthy individuals with an area under the Receiver Operating Characteristic curve (AUC) score of 0.91. AUC is the likelihood the model ranks a randomly-chosen positive example over a negative example — representing whether the classifier tends to assign high probability to the correct class and low probability to incorrect classes[22]. (Other performance metrics are provided in **Supplementary Table 1**.)

Convergent clustering of antigen-specific sequences by edit distance: The second classifier detects highly similar CDR3 amino acid sequences shared between individuals with the same diagnosis, as we and others have previously reported[11,15,23]. The CDR-H3s are the highly variable heavy chain regions that often determine B cell receptor binding specificity. We clustered CDR-H3 sequences with the same IGHV gene, IGHJ gene, and CDR-H3 length, and at least 85% sequence identity — allowing for B cell receptor variability created by somatic hypermutation. As in previous work[11], we selected clusters enriched for sequences from subjects with a particular disease. These clusters represent exemplar sequences predictive of a specific disease across individuals. A new sample's sequences can then be assigned to nearby clusters. To predict immune status, we used the number of disease-specific clusters each specimen falls into as features in a lasso linear model. This CDR-H3 sequence clustering approach predicted patient disease state with 0.82 AUC — when it was able to make a prediction. 1.4% of samples did not have any sequences fall into the clonal parameter and edit distance criteria that defined the clusters. The classifier abstained from making any prediction for these challenging samples.

Language model feature extraction from B cell receptor sequences: Amino acid edit distance may not be an optimal measure of receptor similarity. CDR-H3 sequences encode complex three-dimensional structures, and small sequence changes can cause important structural changes, while different structures with divergent primary amino acid sequences can bind the same target antigen[24,25]. While disease-associated receptors may have lexically dissimilar sequences, they may still share the function of binding to the same target. The third classifier in our framework aims to map primary amino acid sequences into a lower-dimensional space that better captures functional similarities, not just the lexical proximity represented by edit distance. Extending beyond prior research using amino acid biochemical properties, like charge and polarity, rather than edit distance alone to find receptor groups[26–29], we extracted a putative functional representation of B cell receptors. To do so, we used UniRep, one of many self-supervised protein language models shown to learn functional properties for prediction tasks with an approach borrowed from natural language processing[30,31].

Much like words are the building blocks arranged by grammatical rules to convey meaning, protein sequences are built from amino acids composed in an order compatible with polypeptide chain folding and assuming a structure that can carry out functions, like binding to another molecule or catalyzing a chemical reaction. UniRep was trained to predict randomly masked amino acids using the unmasked amino acids in the remaining sequence context of each protein. This requires learning the short and long-range relationships between different regions of a sequence, analogous to learning natural language phrases and grammar rules to anticipate the next word in a sentence. To achieve its task, the UniRep recurrent neural network compresses each sequence into an internal, low-dimensional embedding, capturing traits that allow accurate reconstruction. If the final model can successfully un-mask protein sequences, the compression and uncompression has extracted fundamental features that summarize the input sequences. UniRep's internal representation was shown to encode fundamental properties like structural classes[30].

UniRep was originally trained on over 20 million proteins from many organisms[30]. We hypothesized that by creating a version specialized for immune receptor proteins, we might obtain improved representations for IgH repertoire classification. We continued UniRep's training procedure, but this time to better reconstruct masked B cell receptor sequences (**Methods**). While prior autoencoder models have enabled classification of clusters of similar sequences[32,33], our fine-tuned language model approach combines knowledge of global patterns in proteins from many domains of life, with the specific intricacies of BCR variation. For the disease classification task, we used the low-dimensional embedding learned by the fine-tuned language model to transform each IgH sequence into a 1900-dimensional numerical feature vector, regardless of the IgH sequence length. Then we trained a lasso linear model that maps receptor sequence vectors to disease labels. By aggregating each sequence's predicted class probabilities using a trimmed mean calculation, we arrived at patient-level predictions of specific disease exposures. We chose the trimmed mean because it is a central estimate robust to noisy contamination by the many background, non-disease-related sequences that are part of any repertoire. The language model embedding strategy achieved 0.84 AUC classification performance with no abstentions, demonstrating that the language model learns the syntax of immune receptor sequences, despite their enormous diversity. Because this classifier starts with a predictor for individual receptors, then aggregates sequence calls into a patient-level prediction[26,34], it allows interpretation of which sequences matter most for prediction of each disease — a major advantage over black-box methods. Below, we confirmed that sequences prioritized by our predictor are enriched for disease-specific and antigen-specific IgH.

Ensemble: Finally, we combined all three classifiers — the "healthy or not" global repertoire composition, CDR-H3 sequence clustering, and language model representation strategies — into an ensemble immune status predictor (**Supplementary Figure 1**). We call this adaptive immune receptor analysis framework *MAchine Learning for Immunological Diagnosis (Mal-ID*). By blending the probabilistic outputs from multiple classifiers trained with different strategies, the meta-model exploits each predictor's strengths and can resolve mistakes[35]. (As with the other models, we trained a separate meta-model for each cross-validation group.)

This ensemble approach distinguished five specific disease states in 414 specimens from 355 individuals with 0.96 AUC, compared to 0.82 AUC achieved by the above CDR-H3 clustering model alone (**Supplementary Table 1**). When the ensemble model made a final decision for a given specimen by selecting the disease state (class) with highest predicted probability, the model achieved 85% accuracy across all held-out test sets (**Figure 2A**). In practice, diagnostic sensitivity, the precise threshold on the predicted probability of each disease state, can be tuned to disease prevalence and the desired tradeoff between precision and recall.

The 85% of correct classifications included specimens from all disease types, even unusual immune states like CVID. The CVID cohort was limited to challenging patients measured to have moderate levels of immune

disruption[16], and the HIV cohort included individuals with widely varying breadth of neutralizing antibodies. Of the 15% of misclassified repertoires, 1.4% were abstentions carried over from the CDR-H3 sequence clustering component of the model ensemble. In the remaining ~14% of classification mistakes, the ensemble model tended to have low confidence in its prediction (**Figure 2B**). Allowing the strategy to abstain from inconclusive predictions is important to make the diagnostic robust for challenging real-world cases.

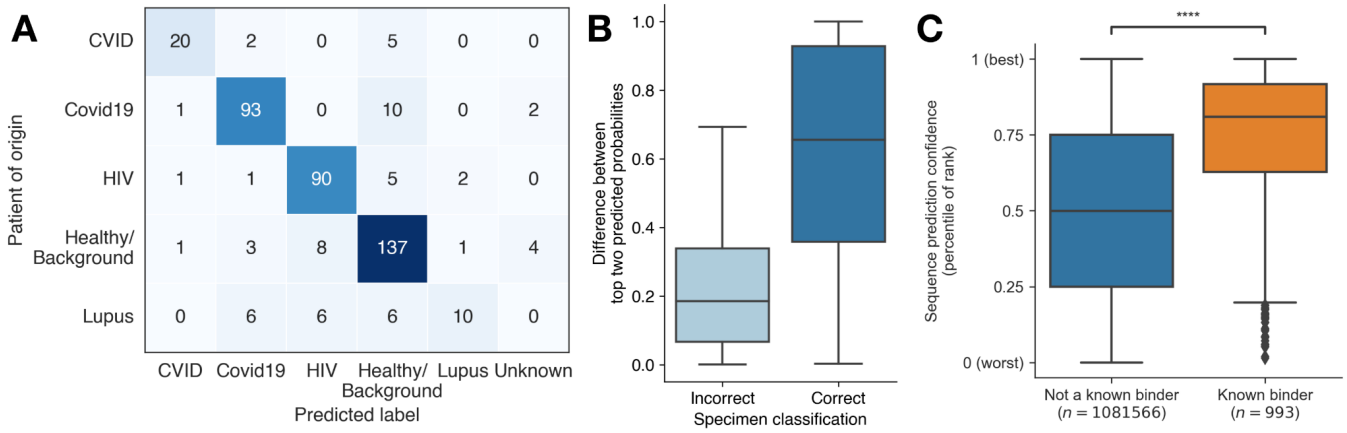


Figure 2: **(a)** Disease classification performance on held-out test data by the ensemble of three machine learning models of B cell repertoires. **(b)** Model prediction confidence for correct versus incorrect predictions, as measured by the difference between the top two predicted class probabilities. A higher difference implies that the model is more certain in its decision to predict the winning disease label, whereas a low difference suggests that the top two possible predictions were a toss-up. **(c)** Sequences in our data convergent to known SARS-CoV-2 binders (orange) were ranked significantly higher than other sequences by the model (one-sided Wilcoxon rank-sum test, U-statistic = 2.770e8, p <= 1e-153). Non-overlapping sequences (blue) may include additional SARS-CoV-2 binders not yet identified in the literature.

Limited impact of age, sex, and race on classification

To study whether extraneous covariates were confounding our results, we investigated whether we could distinguish age, sex, or ancestry of healthy immune receptor repertoires. By training new classifiers to predict these variables, we found that the sex of a healthy individual could not accurately be determined from IgH sequences (**Supplementary Table 2, Supplementary Figure 3C**). However, IgH sequences did carry a weak signal of ancestry, with 0.78 AUC predictive power (**Supplementary Figure 3D**). This signal may also have been increased because many individuals with African ancestry in the cohorts we studied live in Africa and have potentially different environmental exposures.

Healthy IgH sequence repertoires also carried a signal of age. Previous studies have tracked immune aging in gene expression, cytokine levels, and immune cell type frequencies[36,37]. We also see a modest effect of aging through immune receptor sequences. When we dichotomized age as under or over 50 years old to cast this continuous variable as a classification problem, the prediction model achieved 0.70 AUC (**Supplementary Figure 3B**). However, the signatures of age detected by the classifier may correspond to different background or environmental exposures for people over 50 versus younger individuals. For instance, circulating influenza virus types have changed following successive pandemics. The first influenza strain someone is exposed to generates a bias in their influenza responses thereafter, likely by forming memory B cell pools with specificities related to early virus exposures[38]. When we instead divided age into groups by decade, the model achieved only 0.60 AUC and abstained from prediction on almost a fifth of individuals (**Supplementary Figure 3A**). This worse performance suggests that more granular aging differences are challenging to disentangle at the

sequence level with the number of participants, the age ranges, and the cell sampling and sequencing depth in this study. Also, our study was restricted to somatically hypermutated IgD/IgM and class switched IgG/IgA isotypes, reflecting the populations of B cells that are shaped by antigenic stimulation and selection. Studying naive B cells may reveal additional age, sex, or ancestry effects.

We also investigated whether subtle demographic differences between cohorts drove our disease classification results. For example, the age medians and ranges of the cohorts were: HIV (median 31 years, range 19-64); SLE (median 37 years, range 21-71); healthy controls (median 45 years, range 17-87); Covid-19 (median 48 years, range 21-90); CVID (median 50 years, range 23-76). The percentage of females in each cohort was 46% (CVID), 47% (healthy controls), 52% (Covid-19), 64% (HIV), and 93% (SLE). The prevalence of females in our SLE cohort matches general epidemiology[39]. The ancestries and geographical locations of participants also differed between cohorts. Most notably, at least 89% of individuals with HIV were from Africa[11]. 84% of individuals known to have Hispanic/Latino ancestry were in the Covid-19 cohort, and 74% of Caucasians were healthy controls.

To show that demographic metadata are insufficient to predict disease in our dataset, we attempted to predict disease state from age, sex, or ancestry alone, without using sequence patterns at all. No classifier exceeded 0.78 AUC, far lower than the 0.946 AUC when we retrained the *Mal-ID* sequence prediction ensemble on individuals with known demographics only (**Supplementary Table 3, Supplementary Figure 3E-H**). Finally, we also retrained the ensemble disease classification meta-model with age, sex, and ancestry effects regressed out from the ensemble feature matrix. After this correction, classification performance on the 273 individuals with full demographic information available dropped slightly from 0.946 AUC to 0.891 AUC (**Supplementary Table 3, Supplementary Figure 3I-J**). The small decrement in performance after decorrelating sequence features from demographic covariates suggests that age, sex, and ancestry effects have, at most, a modest impact on disease classification.

Language model recapitulates immunological knowledge

We designed our machine learning framework to identify biologically interpretable features of the immunological conditions we studied, not just provide a black box classifier. To assess the ties between the accurate machine learning classification and known biology, we examined which sequences contributed most to predictions of each disease. For example, we ranked all sequences from Covid-19 patients by the predicted probability of their relationship to SARS-CoV-2 immune response using the *Mal-ID* language model classifier. In discriminating between different diseases, sequences highly prioritized for Covid-19 prediction included IGHV gene segments seen in independently isolated antibodies with strong SARS-CoV-2 binding. For example, IGHV3-53, IGHV3-9, and IGHV2-70 have been implicated in spike protein receptor-binding domain binding[40,41], and were highly ranked (**Figure 3B**). So was IGHV1-24, found in N-terminal domain-directed antibodies[42].

We observed a similar pattern for HIV rankings. IGHV4-34, an IGHV gene previously described in HIV-specific B cell responses — with unusually high somatic hypermutation frequencies in individuals producing broadly-neutralizing antibodies[11] — was ranked highly by the model (**Figure 3C**). IGHV4-38-2 was also highly ranked for HIV prediction, and was prevalent among HIV-specific B cells in another analysis[43]. However, IGHV4-38-2 gene usage is significantly more common in African populations in our data (**Supplementary Figure 4**), as in prior literature[44]. The model may have especially prioritized the IGHV4-38-2 gene because our HIV cohort is predominantly of African ancestry. Other IGHV genes flagged by the model are not stratified by ancestry (**Supplementary Figure 4**).

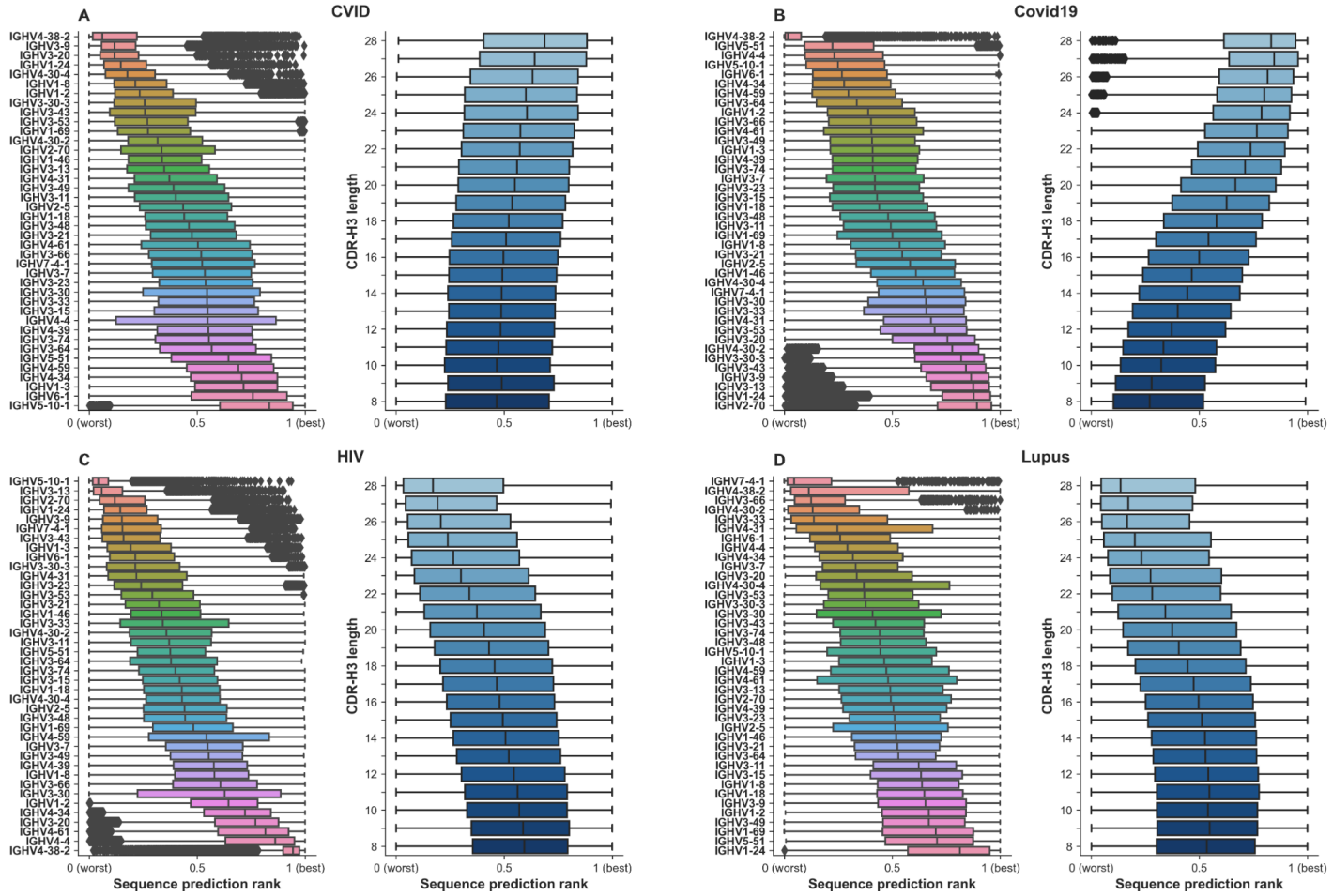


Figure 3: Disease patient-originating sequences ranked by predicted disease class probability and grouped by IGHV gene or CDR-H3 length.

The model's prioritization of IGHV5-51 for SLE prediction matches prior data[45]. But in this classification, previously reported IGHV genes, like IGHV4-34, IGHV4-39, and IGHV4-59[46], were not distinct for SLE prediction (**Figure 3D**). IGHV4-34, in particular, has been reported to be expressed at higher frequencies in SLE patients[47]. The language model classifier made greater use of IGHV4-34 for identifying HIV and CVID infected individuals, instead. The predictor's low ranking for IGHV4-34 as a SLE-defining characteristic may seem at odds with its importance in the literature, but this discrepancy may reflect differences between secreted antibody protein concentrations and the peripheral blood B cell IGHV gene usage, and that increased usage of a particular IGHV gene segment may be seen in several clinical conditions. Clinical tests for the presence of IGHV4-34 antibodies have 70% sensitivity and 95% specificity for SLE[47]. In other words, seven of ten SLE patients are expected to have IGHV4-34 serum antibodies, while only five in 100 healthy individuals may have them[47]. Because IGHV4-34 antibodies are notable for their inherent autoreactivity, are expressed at decreased frequencies in healthy people, but are not necessarily a specific marker of lupus, they may be a general marker of several immune disorders[11,48,49]. Similar reasons may explain why IGHV4-39 and IGHV4-59, two other genes highlighted in prior SLE literature[46], were ranked highly for HIV and CVID prediction. While any one of these V genes on their own is insufficient to diagnose lupus, a clearer picture emerges in combination with the other IGHV genes and specific sequence patterns prioritized by the machine learning model.

The IGHV5-51 gene prioritized for SLE was also highly ranked for CVID prediction (**Figure 3A**). Surprisingly, the *Mal-ID* language model predictor assigned very high CVID ranking to IGHV5-10-1, which is part of a haplotype variant that can replace IGHV3-9 and IGHV1-8[44]. The IGHV5-10-1 gene has also been associated with other conditions with autoimmunity, such as autoimmune hepatitis and celiac disease, where IGHV5-51 has also been implicated[50,51]. Together with the increased usage of IGHV5-10-1 in CVID patients, this may reflect predisposition to CVID related to germline variation in the immunoglobulin heavy chain locus.

The sequence model's rankings also favored certain CDR-H3 lengths. This was notable, because there is no direct input into the model of raw CDR-H3 sequences or their length; all UniRep embedding vectors provided as input to the model have identical sizes, regardless of original sequence lengths. Shorter CDR-H3 lengths were favored by the model for the chronic diseases SLE and HIV (**Figure 3C-D**), consistent with our previous observations about selection for B cell receptors with shorter CDR-H3 segments in HIV[11]. On the other hand, sequences with longer CDR-H3 lengths were favored by the sequence model for Covid-19 class prediction (**Figure 3B**). These prioritized sequences could reflect B cell clones recently derived from naive B cells that have not yet had time to undergo selection, which favors shorter CDR-H3 lengths in memory B cells.

Isotype usage varied by person and across disease cohorts (**Supplementary Figure 5**). To prevent isotype sampling artifacts from driving disease predictions, we designed the sequence model to apply balanced weights to all isotypes. As a result, all isotypes were included among model-prioritized sequences for prediction of each disease (**Supplementary Figure 6**). For Covid-19 prediction, IgG sequences played a slightly bigger role than other isotypes, as may be expected in this infectious disease[18,52–54]. The other models used in the *Mal-ID* ensemble were also designed not to be influenced by isotype sampling amounts. The repertoire composition model quantifies each isotype group separately, and the convergent clustering approach is blind to isotype information. To be sure that differences in isotype proportions between patient cohorts were not sufficient to predict disease, we also trained a separate model to predict disease from a sample's isotype balance alone — with no sequence information provided. The isotype-proportions model achieved only 0.69 AUC, far lower than *Mal-ID*'s 0.96 AUC disease classification performance. Therefore, *Mal-ID*'s classification is robust to data artifacts like isotype proportions.

Language model identifies SARS-CoV-2 binders

Only a small minority of peripheral blood B cell receptor sequences from Covid-19 patients are directly related to the antigen-specific immune response to SARS-CoV-2. Other naive and memory B cells continue to circulate even during acute illness[55]. The 0.96 AUC performance suggests that the ensemble model addresses this "needle in the haystack" issue. We inspected the sequences selected by our language model classifier to assess how important sequences are prioritized.

We matched Covid-19 patient sequences to their nearest neighbors in the CoV-AbDab database of antibodies known to bind the virus[56], allowing up to 15% amino acid mismatches in CDR-H3 arising from somatic hypermutation and other sources of clonal diversity. The dataset of sequences known to bind to SARS-CoV-2 was collected by orthogonal experimental methods, such as direct isolation of B cells that bind the SARS-CoV-2 receptor binding domain (RBD), followed by BCR sequencing[57]. Unlike our global repertoire scans of a limited number of patients, the external database includes larger source cohorts, meaning it may contain more Covid-19 response types than our dataset. The database is also biased towards potential therapeutic antibodies identified by isolating spike antigen-specific B cells.

Despite these differences, we found sequences in our data with high sequence identity (shared IGHV and IGHJ genes, and over 85% amino acid identity in CDR-H3) to known SARS-CoV-2 antigen-specific antibodies — comprising over 10% of known binding antibodies in the CoV-AbDab database, covering all major epitopes and IGHV genes (**Supplementary Figure 2A-B**). 64% of the matching sequences in our dataset were IgG sequences, followed by IgD/M (19%) and IgA (7%), with the final 10% seen in multiple isotypes. This IgG dominance pattern reflects how sequences that class switched to IgG were stimulated by antigen, and is consistent with the isotype relationships examined above. As a negative control, we repeated the process with sequences from healthy subjects. Healthy subject-originating sequences matched 6.8% of Cov-AbDab clusters in total. Nearly 92% of the matched healthy control sequences came from the IgD/M isotypes, most likely representing naive B cells that could mount a response should SARS-CoV-2 enter the body. Matches were rare: 0.09% of unique Covid-19 patient sequences from our dataset matched any CoV-AbDab cluster, along with 0.008% of unique healthy subject sequences. This order of magnitude difference is expected because antigenic stimulation of the IgG population in Covid-19 patients results in clonal expansion and diversification through somatic hypermutation.

In support of the biological relevance of *Mal-ID*'s top-ranked sequences, many were independently validated to bind SARS-CoV-2. Covid-19 patient-originating sequences that overlapped with this known-binder database were assigned significantly higher ranks by the prediction model (**Figure 2C, Supplementary Figure 2C-D**). 84% of matches occurred in the top half of ranked sequences. These binding relationships were not known to the classifier at training time, and the CoV-AbDab sequences were not used in the model. The concordance between automatically-prioritized sequences and experimentally-validated, disease-specific sequences from separate cohorts suggests the language model classifier learned meaningful rules that recapitulate biological knowledge gained during the extraordinary international research effort in response to the COVID-19 pandemic. Additionally, sequences targeting the SARS-CoV-2 RBD made up a smaller portion of the overlapping datasets than they comprise of the full CoV-AbDab database (**Supplementary Figure 2B**). This difference suggests that while RBD-targeting receptors have been prioritized in experimental studies[58], the immune response for other epitopes is of equal or greater importance for disease classification.

Repertoire progression between disease states

To evaluate further disease-specific insights from *Mal-ID*, we developed a novel immune repertoire visualization to convey disease status at a glance. From the training set, we created a reference two-dimensional UMAP layout using receptors that the *Mal-ID* language model classifier learned to confidently separate into distinct groups by immune state (**Figure 4A**). Since this supervised UMAP is conditioned on the disease labels assigned to sequences, any visual distortions created by the reduction into two dimensions are less likely to bias against the disease classes.

We overlaid sequences that were held out from the training set onto the reference UMAP visualization. For example, we visualized the *Mal-ID* language model's interpretation of therapeutic monoclonal antibodies against SARS-CoV-2[59]. Six of eight antibodies were projected into an immune receptor region specific for Covid-19 response (**Supplementary Figure 7**). Since *Mal-ID* was trained on sequence pseudo-labels created from each individual's global disease type, we do not expect perfect sequence-level classification, despite successful repertoire-level disease predictions. One of the mis-projected antibodies (bebtelovimab) had low-confidence sequence predictions (less than 1% difference between the top two class predicted probabilities), and is distinct from the others because it maintained efficacy against circulating Omicron variants of concern[60].

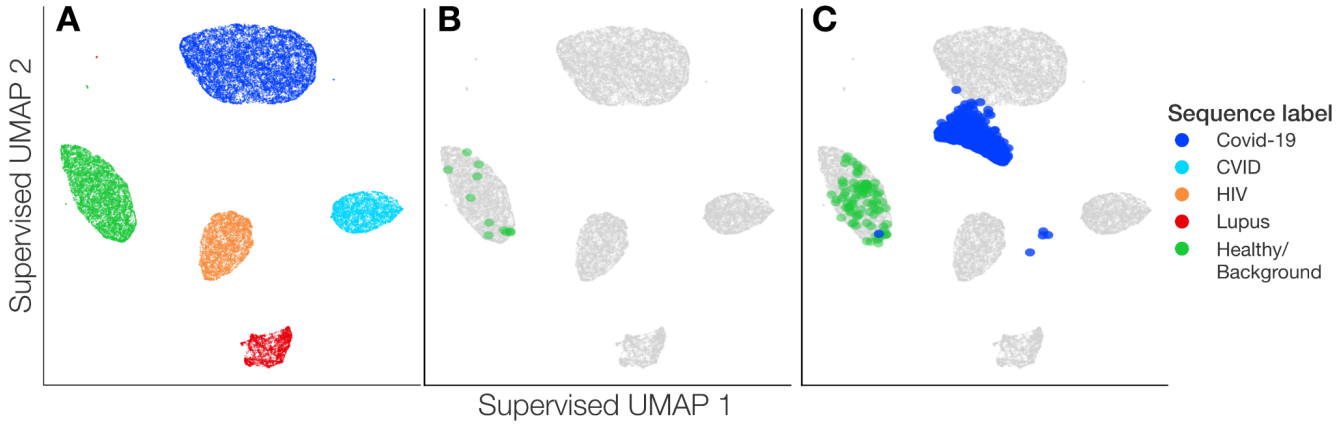


Figure 4: **(a)** Reference UMAP of disease-specific B cell receptors from many patients. Each point is a sequence, colored by the predicted identity of that sequence. Receptors are arranged in disease-specific clusters.
(b-c) Separate Covid-19 patient repertoire overlaid on the reference UMAP, sampled at **(b)** day 5 (Healthy repertoire) and **(c)** day 12 (Covid-19 response emerging) of symptoms.

With the same visualization technique, we can project repeated samples from a held-out test set patient onto the reference map, enabling immune repertoire composition monitoring over time. Unlike the therapeutic mAb visualization of only eight sequences in total, patient repertoires contain a multitude of high-confidence and low-confidence sequences for disease prediction. Sequences with low probability predictions by *Mal-ID* can be excluded to focus the visualization on BCRs likely to be disease-specific. As one example Covid-19 patient's infection and immune responses progressed, the visualization revealed their collection of immune receptors shifting from the Healthy/Background region (green) at day 5 after onset of symptoms, into a Covid-19 area (blue) by day 12 (**Figure 4B-C**).

Discussion

Pathogenic exposures shape the immune system's collection of antigen-specific adaptive immune receptors, forming a record of past and present illnesses. The pathogenic immune responses of autoimmune diseases and the impaired immunity of immunodeficiency conditions are also associated with distinctive alterations in the receptors expressed by B cells and T cells. We applied a three-part machine learning analysis framework to well-characterized disease datasets derived from over 350 individuals with five distinct immunological states, classifying immune responses of study participants with performance of 0.96 AUC. Faced with highly diverse repertoires containing hundreds of thousands of unique sequences, the *Mal-ID* ensemble of classifiers learned disease-specific patterns and chose meaningful sequences for prediction of viral infections, an autoimmune disorder, and an immune deficiency. These signatures of disease and specific pathogens overrode more modest differences detectable between individuals differing by sex, age, or ancestry.

More importantly, the model's interpretability enables testing hypotheses about antigen-specific human immune cell receptors in different illnesses. One key innovation in this study is the development of a language model that translates immune receptors into compressed vectors capturing important properties. The disease predictor is not a black box; we can trace its decisions back to the original sequences. We assessed which sequences drove predictions of each disease, in an effort to identify and validate antigen-specific receptors. The highest priority sequences for Covid-19 disease prediction matched virus-binding sequences previously identified from other patients in a variety of studies spanning different laboratories and countries[56]. We also visualized immune response changes during disease progression, highlighting the potential of longitudinal

immune repertoire monitoring. The language model we developed for B cell receptor sequences could be applied to further tasks beyond identifying disease exposures.

This approach should also be amenable to detecting evidence of more distant prior exposures to pathogens or other immune stimuli, using data from memory immune cell receptor repertoires. Conventional serology tests may only be positive for recent infections or vaccinations, should antibody levels wane soon after exposure. Memory B and T cells can be longer-lived, so an immune repertoire deconvolution strategy may detect distant exposures in individuals who have become seronegative. Further analysis of past exposures could shed light on why some patients are more susceptible to conditions such as the lingering post-infection symptoms of "long Covid", or could help test hypotheses implicating prior viral infections in the initiation of autoimmune diseases[61,62]. It is possible that the model will fail to detect disease exposures in the distant past if very low frequencies of specific B or T cell receptors are present at the time of testing, but such negative findings could have clinical relevance: memory B cell frequencies that are too low to detect may correlate with susceptibility to re-infection.

While the proof of concept in this study provides promising results, it is limited to five immune states and cohorts of only several hundred individuals. The *Mal-ID* framework appears to capture fundamental principles of immune responses, but additional testing in separate clinical cohorts will be needed to further assess its generalizability. The biological validation against known SARS-CoV-2 binders already revealed some limits to the model's grasp of the ultra-high-dimensional receptor repertoire space: 16% of sequences that matched the external database had language-model-assigned ranks under 50% (**Figure 2C**). Since our specificity group detection approach is focused on common response patterns shared across individuals, it is less likely to be able to interpret immune responses unique to a single individual.

With further data collection and additional testing of classification performance on independent held-out cohorts, we believe this immune repertoire analysis strategy could contribute to disease definition and diagnosis, and refining understanding of patterns of immune response such as autoreactivity shared across different pathologies. We anticipate extending this approach to other autoimmune conditions, immunological treatment complications like transplantation rejection, and less well understood conditions suspected to have an immunological basis, like chronic fatigue syndrome. This analysis technique may be able to predict which patients respond to immuno-oncology checkpoint blockade therapy and illuminate the basis for low response rates. In elderly individuals, this technique could identify those with more severe immunological aging and greater risk of severe infectious illnesses. Finally, in future pandemics, this approach could provide useful knowledge about novel pathogen exposures by highlighting patient repertoires that do not match any known disease, and potentially narrowing down the family of viruses to which a new pathogen belongs, such as an influenza virus rather than a coronavirus. Since antigen-specific antibodies are highly ranked by the *Mal-ID* model, it may contribute to developing novel monoclonal antibodies with therapeutic utility in future infectious disease outbreaks.

Online Methods

B cell repertoire sequencing

We assembled immune receptor repertoires from 75 Covid-19, 96 chronic HIV-1, 17 Systemic Lupus Erythematosus (SLE), and 13 Common Variable Immunodeficiency (CVID) patients, along with 154 healthy controls. We excluded mild Covid-19 cases and samples prior to seroconversion. Additionally, we only included SLE patients who had multiple auto-antibodies detected. These filters limited model training data to peak-disease specimens to improve our chances of learning patterns for the disease-specific minority of receptor sequences, which we can then search for in more challenging mild cases.

However, we wanted to avoid creating an artificially simple classification problem from filtering to trivially separable immune states. To this end, we restricted our CVID cohort to “group II”: only patients without an extreme reduction in isotype-switched B cells, as defined in our previous work, were included[16]. We reasoned that these intermediate CVID cases would be challenging to differentiate from other immune states, whereas CVID “group I” patients may be easy to identify as extreme outliers. (Additionally, we had insufficient group I patients for a thorough analysis.) For a similar reason, our HIV cohort included patients regardless of whether they generated broadly neutralizing antibodies to HIV. Had we instead restricted our analysis to HIV-infected individuals who produce broadly neutralizing antibodies, we may have created a more-easily separable HIV class, due to the unusual characteristics of those antibodies[11].

Across these diverse immune states, over 10 million B cell receptors were sampled, PCR amplified with immunoglobulin gene primers, and sequenced as previously described[11]. Briefly, we amplified each immunoglobulin heavy chain isotype in a separate PCR reaction using random hexamer-primed cDNA templates, and performed paired-end Illumina MiSeq sequencing. To reduce the potential for batch effects, data collection followed a consistent protocol. We annotated V, D, and J gene segments with IgBLAST v1.3.0, keeping productive rearrangements only[63]. Using IgBLAST’s identification of mutated nucleotides, we calculated the fraction of the IGHV gene segment that was mutated in any particular sequence; this is the somatic hypermutation rate (SHM) of that B cell receptor. We restricted our dataset to CDR-H3 segments with eight or more amino acids; otherwise the CDR-H3 clustering method below might group short but unrelated sequences.

Then we grouped nearly identical sequences within the same person into clones. For each individual, we grouped all nucleotide sequences from all specimens (including specimens at different timepoints) across all isotypes, and ran single-linkage clustering, requiring clustered sequences to have matching IGHV genes, IGHJ genes, and CDR-H3 lengths, and at least 90% CDR-H3 sequence identity by string substitution distance[11]. We kept only class-switched IgG or IgA isotype sequences, and non-class-switched but still antigen-experienced IgD or IgM sequences with at least 1% SHM. By restricting the IgD and IgM isotypes to somatically hypermutated BCRs only, we ignored any unmutated cells that had not been stimulated by an antigen and were irrelevant for disease classification. The selected non-naive IgD and IgM receptor sequences were combined into an IgM/D group. Finally, we deduplicated the dataset. For each sample from a patient, we kept one copy of each clone per isotype — choosing the sequence with the highest number of RNA reads. On average, any two patients had less than 0.001% sequence overlap, underscoring the enormous diversity of B cell receptor sequences.

We divided individuals into three stratified cross-validation folds, each split into a training set and a test set (**Supplementary Figure 8**). Stratified cross-validation preserved the global imbalanced disease class

distribution in each fold. We also carved out a validation set from each training set, to use for several tasks described below: language model fine-tuning, classifier hyperparameter optimization, and ensemble metamodel training. In all folds, we observed less than 0.1% of sequences shared between any pair of the train, validation, and test sets. Since any single repertoire contains many clonally related sequences, but is very distinct from other people's immune receptors, we made sure to place all sequences from an individual person into only the training, validation, or the test set, rather than dividing a patient's sequences across the three groups. Otherwise, the prediction strategies evaluated here could appear to perform better than they actually would on brand-new patients. Given the chance to see part of someone's repertoire in the training procedure, a prediction strategy would have an easier time of scoring other sequences from the same person in a held-out set. Had we not avoided this pitfall, models may also have overfit to the particularities of training patients.

Healthy versus Sick classifier using overall repertoire composition features

We created IgG, IgA, and IgM/D summary feature vectors for each specimen. For a specimen's subset of sequences belonging to each isotype, we first calculated the median sequence somatic hypermutation rate, and the proportion of sequences that are somatically hypermutated (with at least 1% SHM). We also tallied IGHV gene and IGHJ gene usage, counting each clone once. To account for different total clone counts across samples, we normalized total counts to sum to one per specimen. Then we log-transformed and Z-scored the matrix representing how counts are distributed across V-J gene pairs. Finally, we performed a PCA to reduce the count matrix to fifteen dimensions. All transformations were computed on each training set and applied to the corresponding test set.

We fit a binary lasso linear model with L1 regularization on the 51-dimensional (17 x 3 isotypes) feature vectors from each specimen to predict whether a specimen was from a healthy control or from an actively ill patient[64]. Features were standardized to zero mean and unit variance. We repeated this feature engineering and model training procedure on each cross-validation fold separately, then combined results from all test folds. All analyses were performed and plotted with python v3.7.10, numpy v1.19.1, pandas v1.2.5, scipy v1.6.2, scikit-learn v1.0, umap-learn v0.5.1, matplotlib v3.3.2, and seaborn v0.11.2.

Disease classifier by clustering CDR-H3 sequences with edit distance

We performed single-linkage clustering on CDR-H3 sequences from B cells with identical IGHV genes, IGHJ genes, and CDR-H3 lengths, as described previously[11]. Nearest-neighbor clusters were iteratively merged if all cross-cluster pairs had high sequence identity (at least 85%), as measured by string substitution distance.

Filter to disease-specific clusters: We kept clusters with sequences from three or more individuals, as long as at least 80% of those individuals were positive for some disease. For each remaining predictive cluster, we created a cluster centroid — a single consensus sequence. Recall that each cluster member is a clone from which only the most abundant sequence was sampled. Rather than having each cluster member contribute equally to the consensus centroid sequence, contributions at each position were weighted by clone size: the number of unique BCR sequences originally part of each clone.

Compute feature vector for each specimen: Sequences from a sample were then matched to these predictive cluster centroids. In order to be assigned, a sequence must have the same IGHV gene, IGHJ gene, and CDR-H3 length as the candidate cluster, and must have at least 85% sequence identity with the consensus sequence representing the cluster's centroid. After assigning sequences to clusters, we counted cluster

memberships across all sequences from each specimen. We found these cluster memberships for training set samples, then computed a feature vector for each specimen. A specimen's score for a particular disease was defined as the number of disease-predictive clusters into which some sequences from the specimen were matched. This featurization captures the presence or absence of convergent immunoglobulin sequences, without regard for isotypes.

Fit and evaluate model: Features were standardized, then used to fit a linear logistic regression with L1 regularization and balanced class weights (inversely proportional to input class frequencies), using scikit-learn v1.0. The featurization and model were fitted on each training set and applied to the corresponding test set. Predicted labels from all test sets were then concatenated for accuracy evaluation.

We abstained from prediction if a sample had no sequences fall into a predictive cluster. Abstentions hurt accuracy scores, but were not included in the AUC calculation, since no predicted class probabilities are available for abstained samples. Fewer than 1.5% of samples resulted in abstention (**Supplementary Table 1**).

Language model representations for immune sequences

We combined the CDR-H1, CDR-H2, and CDR-H3 segments of each receptor sequence, then embedded the concatenated amino acid strings with the UniRep neural network, using the jax-unirep v2.1.0 implementation[65]. A final 1900-dimensional vector representation was calculated by averaging UniRep's hidden state over the original protein's length dimension[30].

To embed sequences, we used either the default UniRep weights, or weights fine-tuned on a subset of each cross-validation fold's training set — yielding a total of three fine-tuned models, one per fold. We chose the weights that minimized cross-entropy loss on a subset of the held-out validation set. (We used subsets here due to computational resource constraints.)

The fine-tuning procedure was unsupervised. Besides the raw CDR1+2+3 sequence, no disease or other class labels were provided during fine-tuning. As a result, the fine-tuned language models are specialized to B cell receptor patterns, but not hyper-specialized to the disease classification problem. They can be applied to other immune sequence prediction tasks.

Disease classifier using language model embeddings

The analysis pipeline for classifying disease with language model embeddings of sequences is complex, but necessarily so because it aggregates individual sequence data to generate patient-level predictions.

Sequence-level disease classifier: First, we trained lasso models to map sequences to disease labels — one model per fold. As input data, we used UniRep embeddings (standardized to zero mean and unit variance), along with categorical dummy variables representing the IGHV gene and isotype of each sequence.

Making predictions for individual sequences before aggregating to a patient-level prediction has interpretation benefits, but the two-stage approach introduces a new challenge. The available ground truth data associates *patients*, not sequences, with disease states. We do not know which of their sequences are truly disease related. To train the individual-sequence-level model, we provided noisy sequence labels derived from patient global immune status. But this transfer creates very noisy labels: even at the peak-disease timepoints in our

dataset, disease-specific immune receptor patterns nevertheless represent just a small subset of a patient's vast immune receptor repertoire. Our approach must account for unreliable sequence labels and choose the right subset of sequences to make a patient-level decision.

We used highly regularized statistical models equipped to withstand the noisy training labels created by transferring patient labels to the sequence-level prediction task. Models were trained with the scikit-learn v1.0 implementation of logistic regression with lasso regularization and multinomial loss, using balanced class weights and default hyperparameters. The lasso's L1 penalty encouraged sparsity among the ~2000 input features[64]. Because isotype use varies from person to person, we trained the sequence-level model with isotype weights to account for this imbalance.

Aggregate sequence predictions to specimen prediction: Since we have no true sequence labels, we cannot evaluate classification performance for the sequence-level classifier. Instead, we aggregated sequence predictions into a patient specimen-level prediction. Using the predicted disease class probabilities for each sequence belonging to a specimen, we computed the trimmed mean for each class across the sequences. That is, we removed the top and bottom 10% of outlying scores, then computed the mean of the remainder, weighing sequences inversely proportional to their isotype's overall usage in the specimen. (This way, minority isotype signal is not drowned out.) Then we renormalized disease class probabilities to sum to one for each specimen.

Tune class decision thresholds: To complete this specimen-level classifier based on aggregating sequence predictions, we tuned class decision thresholds against the held-out validation set. Specifically, we reweighted class probabilities to optimize the Matthews correlation coefficient, a classification performance metric that is meaningful even under class imbalance[66]. Before applying class weights, the winning label for each specimen was chosen based on the class with highest predicted probability. If a class then had its probabilities reweighted by ⅕, for example, the model must be five times more confident to choose that class label. Importantly, these weights were applied only in the choice of a final predicted label for each specimen. This procedure affected the confusion matrix, accuracy, and other metrics based on predicted labels, but the AUC did not change.

Evaluate classifier: Finally, we evaluated the specimen sequence-prediction-aggregating predictor on the test set. Each test specimen's sequences were scored, then combined with a trimmed mean as above. The resulting disease class probabilities for each specimen were reweighted by the global class weights found above, to arrive at final predicted specimen labels. Ground truth specimen disease status is known, so we could evaluate classification performance here, unlike at the sequence-level prediction stage.

Ensemble metamodel

After training healthy-vs-sick, CDR-H3 clustering, and language model embedding and aggregation models on each fold's training set, we combined the classifiers with an ensemble strategy. For each fold, we ran all trained base classifiers on the validation set, and concatenated the resulting predicted class probability vectors from each base model. We carried over any specimen abstentions from the CDR-H3 clustering model (the other models do not abstain). Finally, we trained a multiclass linear support vector machine (SVM) to map the combined predicted probability vectors to validation set specimen disease labels. The model was trained in a "one-vs-rest" fashion. We evaluated this metamodel on the held-out test set.

Predicting demographic information from healthy subject repertoires

We repeated the above process to predict age, sex, or ancestry instead of disease. Input data was limited to healthy controls to avoid learning any disease-specific patterns. To cast this as a classification problem, age was discretized either into deciles or as a binary “under 50 years old” / “50 or older” variable. Notably, the oldest individual among the healthy control participants was 87 years old, with only five people over 70. Therefore, our data do not assess IgH repertoire changes at more extreme older ages. We excluded any healthy individuals over 70 years old, as well as those under 20 years old, from the analysis.

For each of the three tasks, we trained a sequence-level lasso-regularized linear model starting from language model embeddings, and aggregated sequence predictions to specimen predictions as above. Similarly, we trained a CDR-H3 edit distance-based clustering model for each task, and an overall repertoire composition model. In this case, the repertoire composition model did not predict whether a specimen was overall healthy or sick; instead it predicted the age, sex, or ancestry class directly (we optimized multinomial loss). Finally, the three trained models were combined into a new ensemble meta-model for each prediction problem. We note that we did not explicitly introduce data from allelic variant typing in germline IGHV, IGHD, or IGHJ gene segments into our models, but such data could be expected to increase detection of ancestry in such datasets.

Evaluating predictive power of potential demographic confounding variables

We retrained the entire *Mal-ID* disease-prediction set of models on the subset of individuals with known age, sex, and ancestry. (As above, we excluded any individuals over 70 years old or under 20 years old.) Additionally, we regressed out those demographic variables from the feature matrix used as input to the ensemble step. Specifically, we fit a linear regression for each column of the feature matrix, to predict the column’s values from age, sex, and ancestry. The feature matrix column was then replaced by the fitted model’s residuals. This procedure orthogonalizes or decorrelates the metamodel’s feature matrix from age, sex, and ancestry effects. We regressed out covariates at the metamodel stage because it is a specimen-level, not sequence-level model, and age/sex/ancestry demographic information is tied to specimens rather than sequences.

Separately, we also trained models to predict disease from either age, sex, or ancestry information encoded as categorical dummy variables. Here, no sequence information was provided as input. The best-performing model in each case ranged from a linear SVM, to a linear logistic regression model with elastic net regularization, to a random forest model.

Model ranking of disease-specific sequences

In each test set, we scored Covid-19 patient-originating sequences with the sequence-level classifier based on language model embeddings. Predicted Covid-19 class probabilities were combined for all sequences across folds. Some sequences were seen in multiple people, appearing in more than one test fold and thus receiving a different predicted probability from each fold’s model. We deduplicated these sequences by choosing the copy with highest predicted disease class probability, to capture just how disease-related the sequence could be. Then sequences were ranked by their predicted probability, and ranks were rescaled from 0 to 1 (highest original probability). We repeated this process for other diseases.

Using these ranked sequence lists, we examined the relationship between rank and sequence properties like CDR-H3 length, isotype, and IGHV gene segment. For the IGHV gene usage comparison (**Figure 3**), IGHV

genes with very low prevalence were removed. To set a prevalence threshold, we found the greatest proportion each IGHV gene ever comprises of any cohort, and took the median of these proportions (**Supplementary Figure 9**). IGHV genes that failed to make up at least 0.5% of any disease cohort were removed. Specifically, this half of IGHV genes was filtered out: V1-45, V1-58, V1-68, V1-f, V1/OR15-1, V1/OR15-2, V1/OR15-3, V1/OR15-4, V1/OR15-5, V2-26, V2-70D, V3-16, V3-19, V3-22, V3-35, V3-38, V3-43D, V3-47, V3-52, V3-64D, V3-71, V3-72, V3-73, V3-NL1, V3-d, V3-h, V3/OR16-10, V3/OR16-13, V3/OR16-8, V3/OR16-9, V4-28, V4-55, V4/OR15-8, V5-78, V7-81, VH1-17P, VH1-67P, VH3-41P, VH3-60P, VH3-65P, and VH7-27P. Most IGHV genes remaining after this filter had consistent, balanced prevalence across cohorts (**Supplementary Figure 10**).

Overlap with database of known SARS-CoV-2 binders

We downloaded the Jan 31, 2022 version of CoV-AbDab[56], filtering to antibody sequences known to bind to SARS-CoV-2 (including weak binders). Further, we selected sequences from human patients or human antibody libraries, and removed any IGHV genes that were never present in our dataset, as these sequences would never be matched. We clustered the remaining SARS-CoV-2 binders from CoV-AbDab with identical IGHV gene, IGHJ gene, and CDR-H3 lengths and at least 95% sequence identity. Several related sequences were combined and replaced by a consensus sequence.

Then we found overlapping sequences between our dataset and CoV-AbDab. First, in case of sequences in our dataset originating from different isotypes but sharing the same IGHV gene, IGHJ gene, and CDR-H3 sequence, we kept the copy with highest predicted Covid-19 probability, in order to assess the strength of a sequence's relationship to the disease. Then we assigned each sequence originating from a Covid-19 patient in our dataset (from any isotype) to the nearest CoV-AbDab cluster centroid, as long as they had the same IGHV gene, IGHJ gene, and CDR-H3 length and at least 85% sequence identity. Iterating over sequences in model-ranked order, starting with highest confidence sequences, we counted the cumulative number of matches to a cluster in the known binder database. Finally, we calculated enrichment of these observed counts versus expected hits if sequences were ordered at random. The number of draws to sample a certain number of known binders without replacement from a pool of sequences follows the negative hypergeometric distribution. With N total sequences containing n<N known binders, a new known-binder is expected to be found with every $\frac{N+1}{n+1} = \frac{N-n}{n+1} + 1$ sequence draws[67,68].

Repertoire visualization

For each receptor, the lasso sequence model gives predicted class logits, which are proportional to the dot product of the embedded sequence vector and the model coefficients. In other words, this linear transformation applies the coefficients as weights on the input features, creating a sequences-by-classes matrix. To create a 2D visualization, we then ran UMAP on the per-disease-state logits for each sequence. We provided sequence labels as supervision to the UMAP so they are less likely to be distorted in the layout[69].

We created a reference UMAP for each fold using a subset of training set sequences likely to be related to each disease state (or healthy). We selected this subset of sequences with the following filters:

First, to form the subset of sequences for a particular disease class, we only considered sequences that originated from a patient with that disease. Otherwise, the sequence could not plausibly be related to that disease. It would not make sense for a Covid-19 representative sequence to come from an HIV patient, for example.

Second, the lasso sequence model's prediction for this sequence must match the disease class, as well. After all, we are constructing a reference layout of disease-specific sequences, so we should only include sequences the model has classified into the disease class. Similarly, we only consider sequences from the healthy class that originated from a healthy subject and are predicted to belong to that class.

Third, we excluded sequences whose predictions were close calls. We wish to avoid these borderline sequences in the construction of the reference map, especially because of the high label noise (described earlier). Therefore, we filtered potential sequences to those with predicted disease class probability at least 0.2 greater than probabilities predicted for any other class.

Finally, we sorted the remaining candidate sequences for each disease by their predicted probability of belonging to that disease state, and kept the top 20% to create a succinct pool of reference sequences for each class. The per-class logits for only these sequences were used to construct a UMAP.

Once the UMAP was constructed, we projected held-out sequences into the layout. First, we overlaid therapeutic monoclonal antibodies onto the 2D map. Their sequences were found via Thera-SAbDab[59] and annotated with IgBLAST[63]. We computed supervised embeddings (per-class logits) for each sequence using the sequence-level lasso model, and applied the trained UMAP transformation, producing 2D coordinates for each antibody.

Second, we overlaid a held-out test patient's sequences on the UMAP, applying the same process to a subset of the patient repertoire's sequences predicted to be disease-specific. We used the model and UMAP transformations belonging to the fold where the patient was in the held-out test set. The patient's repertoire was filtered to sequences whose predicted labels match the overall specimen prediction by the ensemble metamodel, or sequences predicted to be Healthy/Background. As a result, the visualization included both the healthy and disease related components of this patient's B cell repertoire. We further filtered sequences to those with confident model predictions: we chose sequences having top predicted class probability at least 0.1 greater than the next highest class probability. All sequences remaining after these filtering steps were sorted by their predicted class probability. We kept the top 20% of the sorted list across Healthy/Background and the overall specimen predicted label class.

Data and code availability

Data and code will be deposited online.

Ethics statement

The use of data was approved by Stanford University IRBs #48973 and #55689, as well as institutional approvals at local sites.

Author contributions

M.E.Z., A.K., and S.D.B. conceived the study. N.R.M., J.M.G., J.T.M., J.D.G., C.C.R., M.A.M, B.F.H., B.A.P., J.R.H., J.A.J., S.Y., and C.A.B. provided blood samples, as well as clinical and demographic data annotation and analysis. J.Y.L. prepared and sequenced samples. K.M.R. designed the data warehouse. M.E.Z. and R.T.

performed the computational analysis. M.E.Z., R.T., A.K., and S.D.B. wrote the manuscript with inputs from all authors.

Funding

S.D.B. was partially supported by NIH/NIAID grants R01AI130398, R01AI127877, U19AI057229, U54CA260518, and a philanthropic gift from an anonymous donor. M.E.Z. was supported by the National Science Foundation Graduate Research Fellowship. R.T. was supported by the National Institutes of Health (5R01 EB001988-16) and the National Science Foundation (19 DMS1208164). B.F.H and M.A.M. were supported by the NIH, NIAID, Division of AIDS Center for HIV/AIDS Vaccine Immunology-Immunogen Discovery (UM-1 AI100645) and the Consortia for HIV/AIDS Vaccine Development (UM1 AI144371). C.C.R. was supported by the National Institutes of Health, AI 101093, AI-086037, AI-48693, and the David S Gottesman Immunology Chair. A.K. was partially supported by the Stanford School of Medicine COVID19 Research Fund. S.Y was supported by NIH/NIAID grants R01AI153133, R01AI137272, and 3U19AI057229–17W1 COVID SUPP 2 and a philanthropic gift from Eva Grove.

Acknowledgments

We thank Akshay Balsubramani and members of the Kundaje and Boyd labs for helpful discussions. We also thank the Stanford Covid-19 Biobank Study Group's members: Rosen Mann, Anita Visweswaran, Thanmayi Ranganath, Jonasel Roque, Monali Manohar, Hena Naz Din, Komal Kumar, Kathryn Jee, Brigit Noon, Jill Anderson, Bethany Fay, Donald Schreiber, Nancy Zhao, Rosemary Vergara, Julia McKechnie, Aaron Wilk, Lauren de la Parte, Kathleen Whittle Dantzler, Maureen Ty, Nimish Kathale, Arjun Rustagi, Giovanny Martinez-Colon, Geoff Ivison, Ruoxi Pi, Maddie Lee, Rachel Brewer, Taylor Hollis, Andrea Baird, Michele Ugur, Drina Bogusch, Georgie Nahass, Kazim Haider, Kim Quyen Thi Tran, Laura Simpson, Michal Tal, Iris Chang, Evan Do, Andrea Fernandes, Allie Lee, Neera Ahuja, Theo Snow, James Krempski.

Declaration of Interests

M.E.Z., R.T., A.K., and S.D.B. are co-inventors on a patent application related to this manuscript. S.D.B. has consulted for Regeneron, Sanofi, Novartis, and Janssen on topics unrelated to this study and owns stock in AbCellera Biologics. A.K. is scientific co-founder of Ravel Biotechnology Inc., is on the scientific advisory board of PatchBio Inc., SerImmune Inc., AINovo Inc., TensorBio Inc. and OpenTargets, is a consultant with Illumina Inc. and owns shares in DeepGenomics Inc., Immunai Inc., and Freenome Inc. C.A.B. reports compensation for consulting and/or SAB membership from Catamaran Bio, DeepCell Inc., Immunebridge, Sangamo Therapeutics, and Revelation Biosciences on topics unrelated to this study. J.D.G. has consulted for Eli Lilly and Gilead and reports research support from Eli Lilly, Gilead, Regeneron, Merck, and collaborative services agreements with Adaptive Biotechnologies, Monogram Biosciences and Labcorp (outside of this study). Other co-authors declare that they have no competing interests.

Supplementary Information

Prediction task	Strategy	Accuracy	ROC AUC	AUPRC	Abstention rate
Healthy or sick (with any disease)	Global repertoire statistics	84.8%	0.910	0.792	0%
Specific disease	Raw CDR-H3 sequence clustering (state of the art) [11]	43.2%	0.823	0.786	1.4%
Specific disease	Language model embedding and classification	67.1%	0.843	0.832	0%
Specific disease	Ensemble of all models (Figure 2A)	84.5%	**0.961**	**0.947**	1.4%

Supplementary Table 1: Average cross-validated test set performance on 414 specimens. AUPRC stands for area under the precision-recall curve.

Input	Prediction target	Model strategy	Accuracy	AUC
Sequences from 121 healthy individuals	Age (20-30, 30-40, …, 60-70)	Language model	33%	0.60
		Full ensemble	22% (19.8% abstentions)	0.60
	Age (under 50, 50 or older)	Language model	63%	0.66
		Full ensemble	57% (4.1% abstentions)	0.70
	Sex	Language model	59%	0.47
		Full ensemble	49% (2.5% abstentions)	0.59
	Ancestry	Language model	70%	0.77
		Full ensemble	56% (3.3% abstentions)	0.78

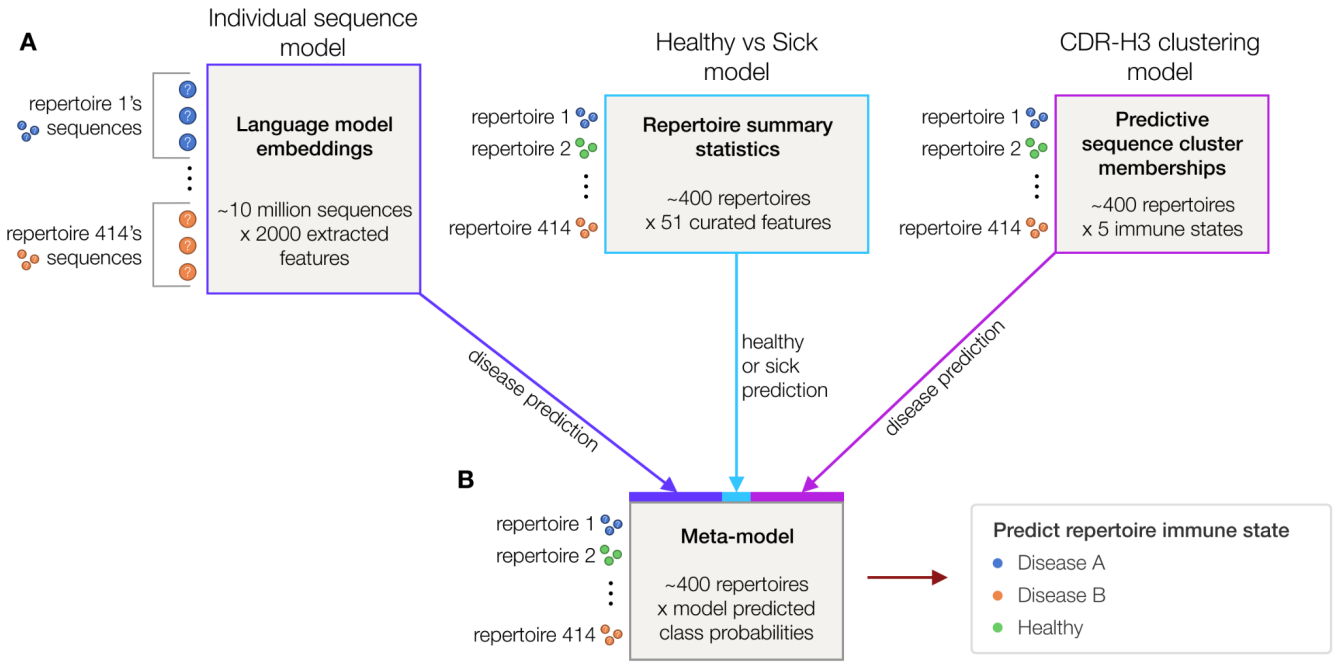
Supplementary Table 2: Model performance for predicting age, sex, and ancestry of healthy individuals, using the language model classifier alone or the full *Mal-ID* ensemble architecture. To cast age as a classification problem, the continuous variable was discretized either into deciles or at a 50-year threshold.

Input	**Prediction target**	**Accuracy**	**AUC**
Age (20-30, 30-40, …, 70-80)	Disease (304 samples from 273 individuals with known age, sex, and ancestry)	43%	0.60
Age (under 50, 50 or older)		46%	0.56
Sex		37%	0.56
Ancestry		66%	0.78
Sequences (*Mal-ID* ensemble model)		80.6% (1.0% abstentions)	0.946
Sequences (*Mal-ID* ensemble model, with age, sex, and ancestry regressed out)		73.4% (1.0% abstentions)	0.891
Sequences (*Mal-ID* ensemble model)	**Disease (full dataset: 414 samples from 355 individuals)** (Figure 2A)	**84.5%** (1.4% abstentions)	**0.961**

Supplementary Table 3: Classification results for disease prediction from age, sex, or ancestry alone without any sequence features, compared to disease classification from sequence using the *Mal-ID* architecture.

Supplementary Figure 1: Classification pipeline.

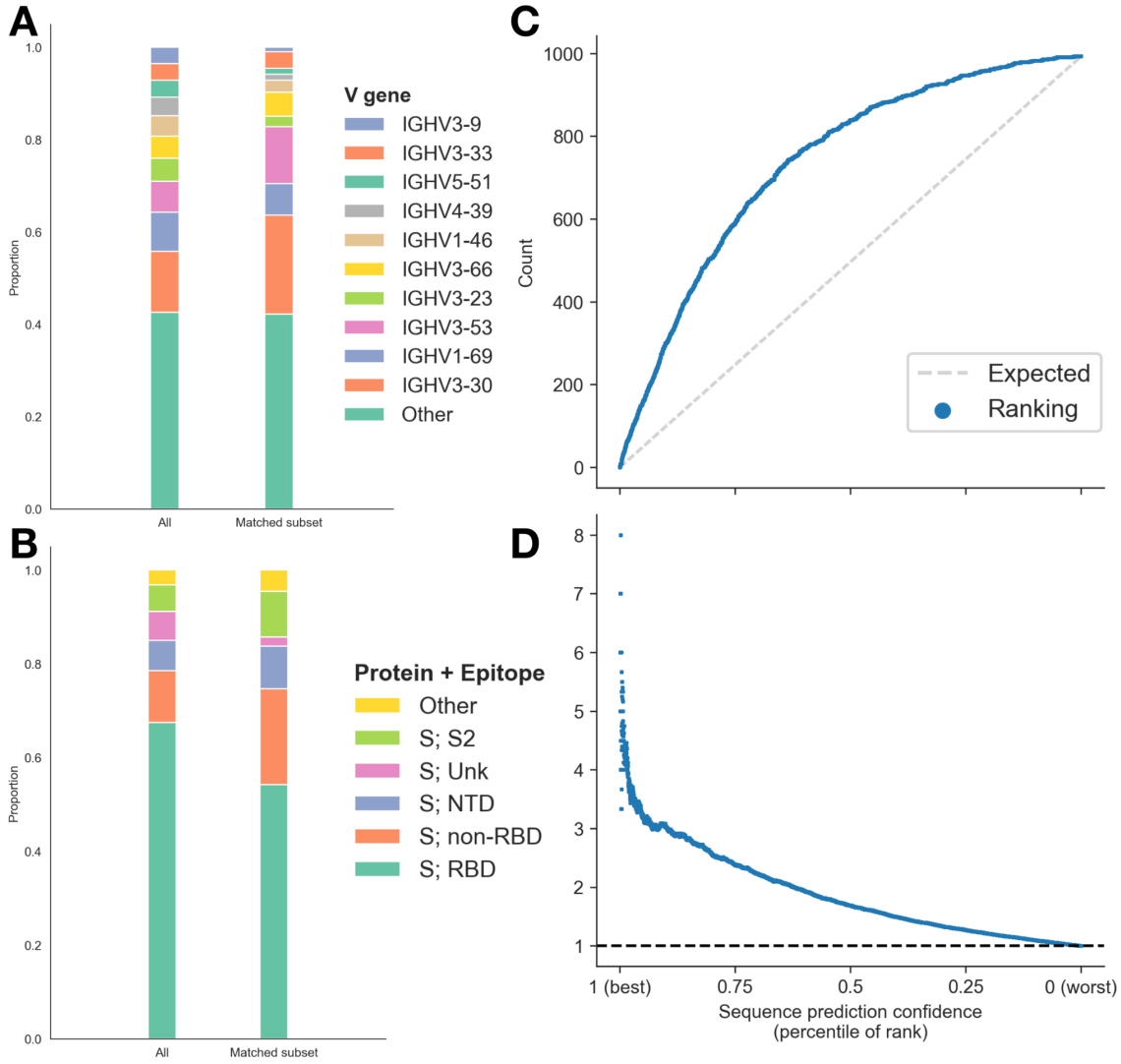
(a) Stage one: fit three models on training set.
(b) Stage two: fit meta-model on validation set to ensemble the three inner models.



Supplementary Figure 2: Language model identification of SARS-CoV-2 binders.

(a) IGHV gene usage in the entire external database of known SARS-CoV-2 binding antibody sequences, versus IGHV gene usage in the subset also found in our independent cohorts.
(b) Epitope specificities of the entire external database of known SARS-CoV-2 binding antibody sequences, versus epitope specificities for the subset also found in our independent cohorts.

(c) Cumulative count of how many sequences matched the external database of known binders, when sequences are processed in the order of their predicted model confidence, versus at random.
(d) Enrichment for cumulative number of sequence matches to the external known binder database when sequences are processed by model rank order versus random order.

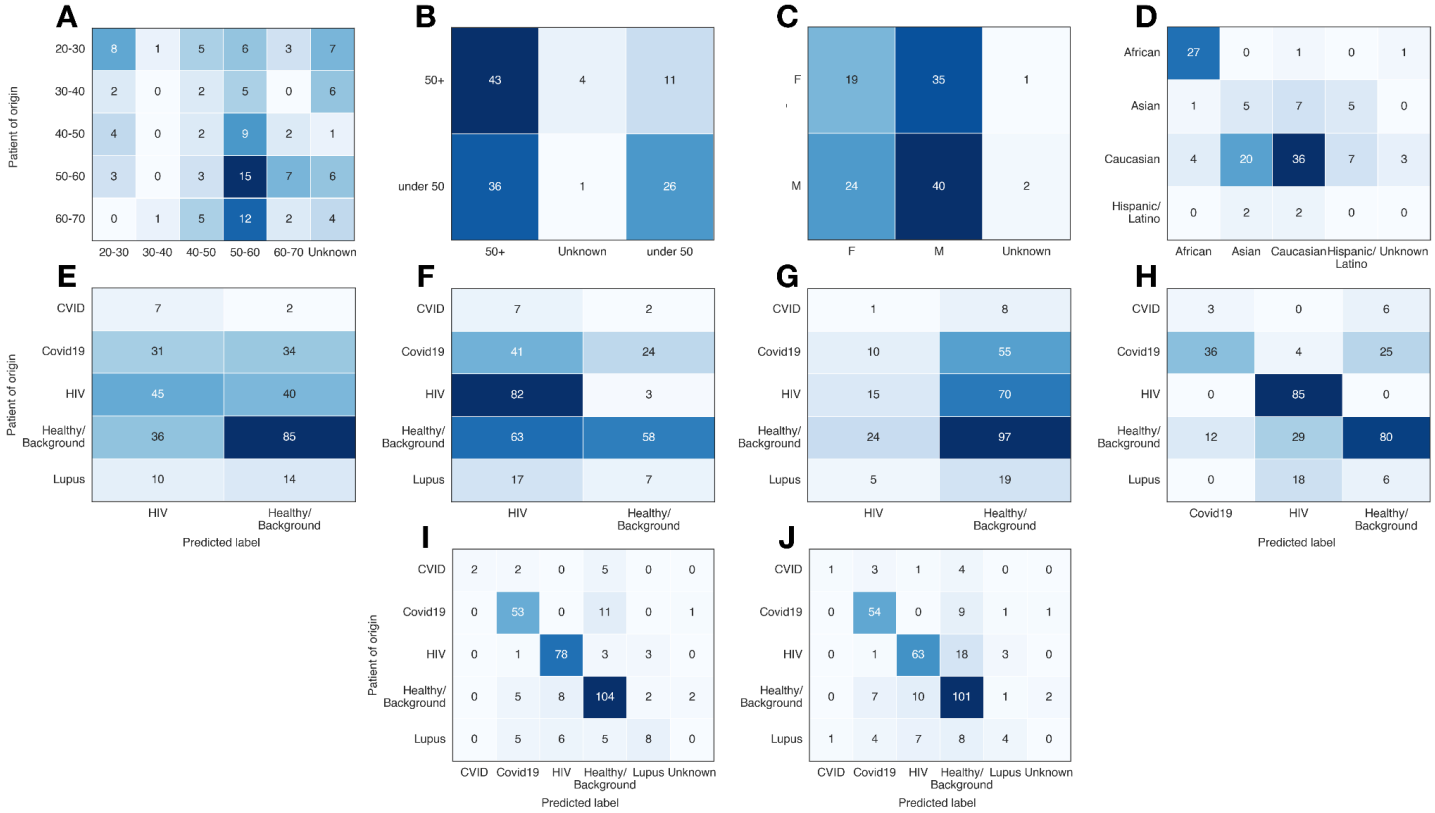


Supplementary Figure 3

(a-d) *Mal-ID* metamodel classification performance for predicting **(a)** age (deciles), **(b)** age (under 50 or 50+), **(c)** sex, or **(d)** ancestry from sequence patterns in healthy individuals.

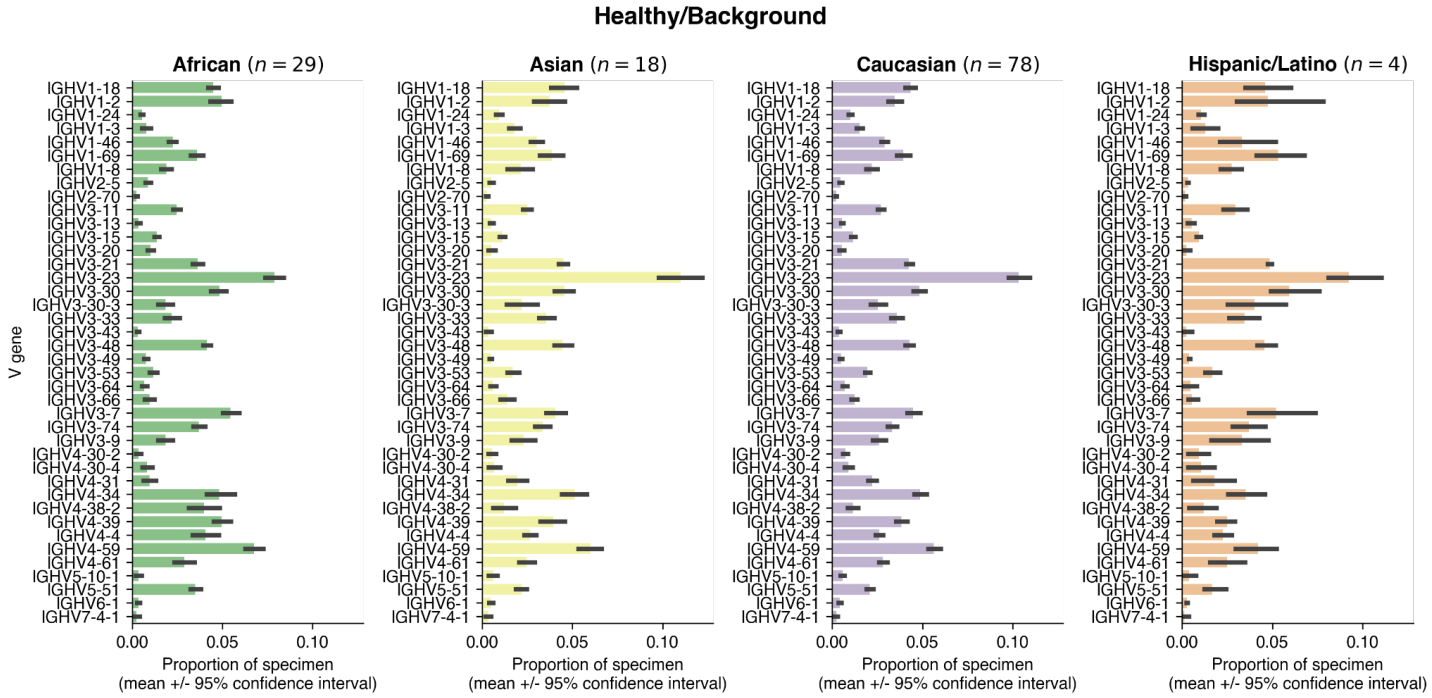
(e-h) Disease prediction performance from **(e)** age (deciles), **(f)** age (under 50 or 50+), **(g)** sex, or **(h)** ancestry alone, without any sequence information.

(i-j) Metamodel classification performance for predicting disease from sequence patterns in individuals with known age, sex, and ancestry, in **(i)** standard *Mal-ID* process, or **(j)** with age, sex, and ancestry regressed out of metamodel feature matrix.



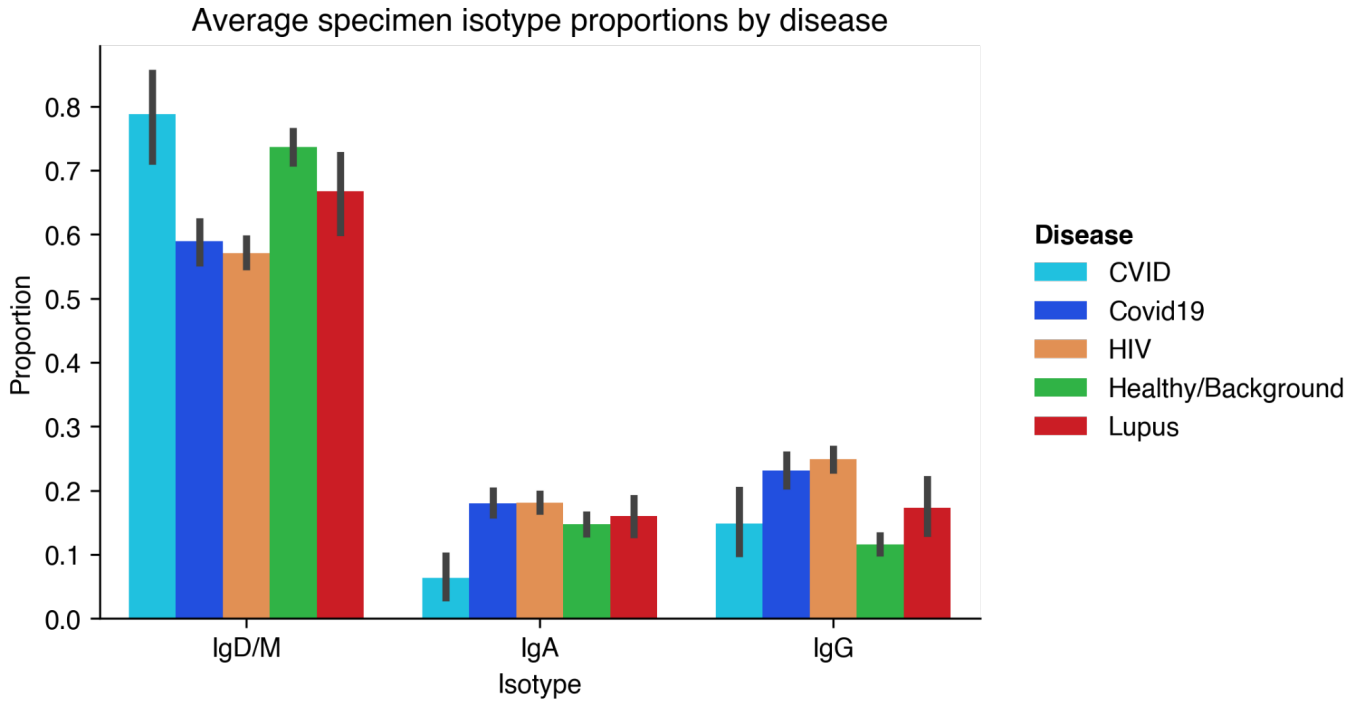
Supplementary Figure 4

IGHV gene use proportions in healthy control specimens, stratified by ancestry. Average and 95% confidence interval plotted.



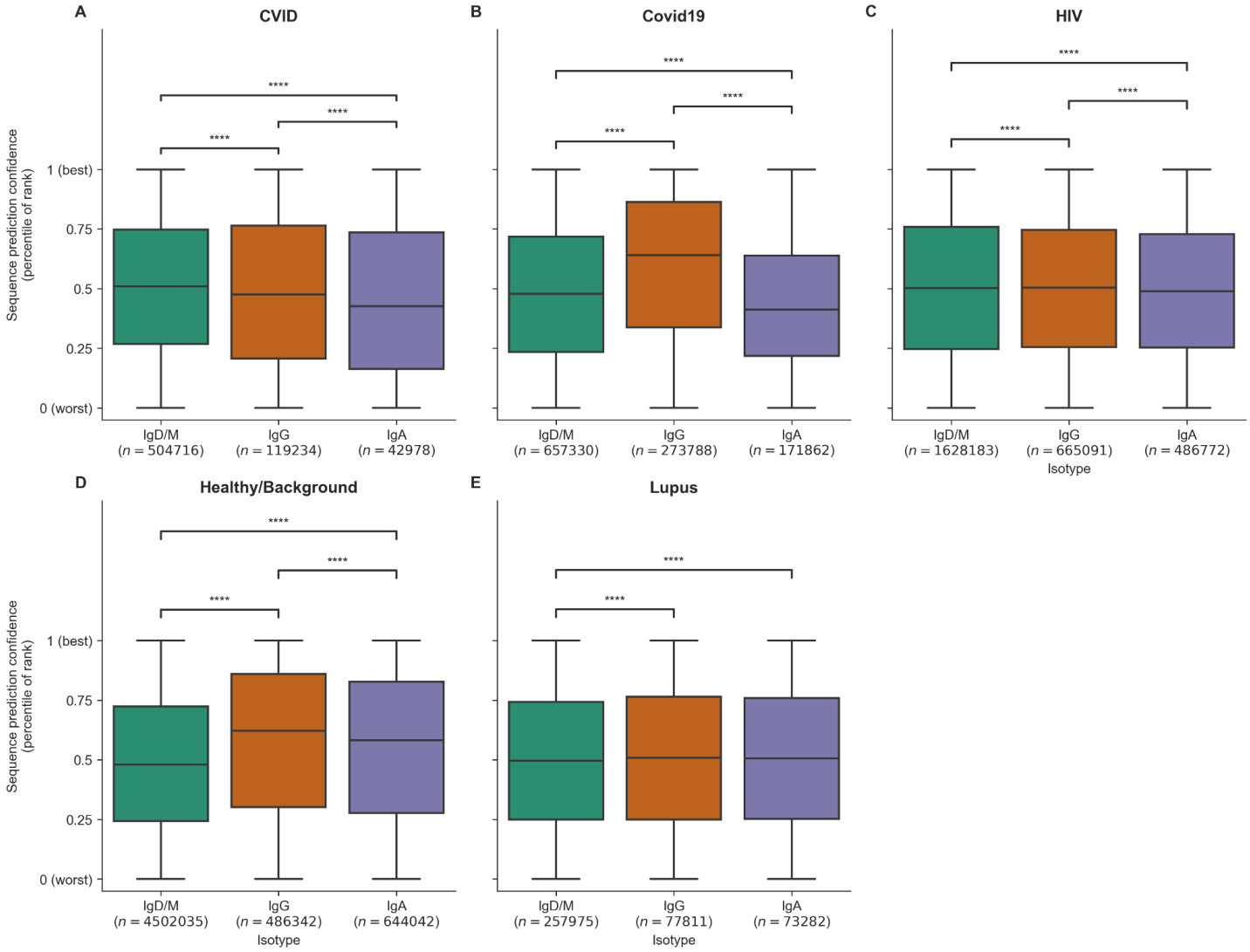
Supplementary Figure 5

Specimen isotype proportions per disease. Average and 95% confidence interval plotted.



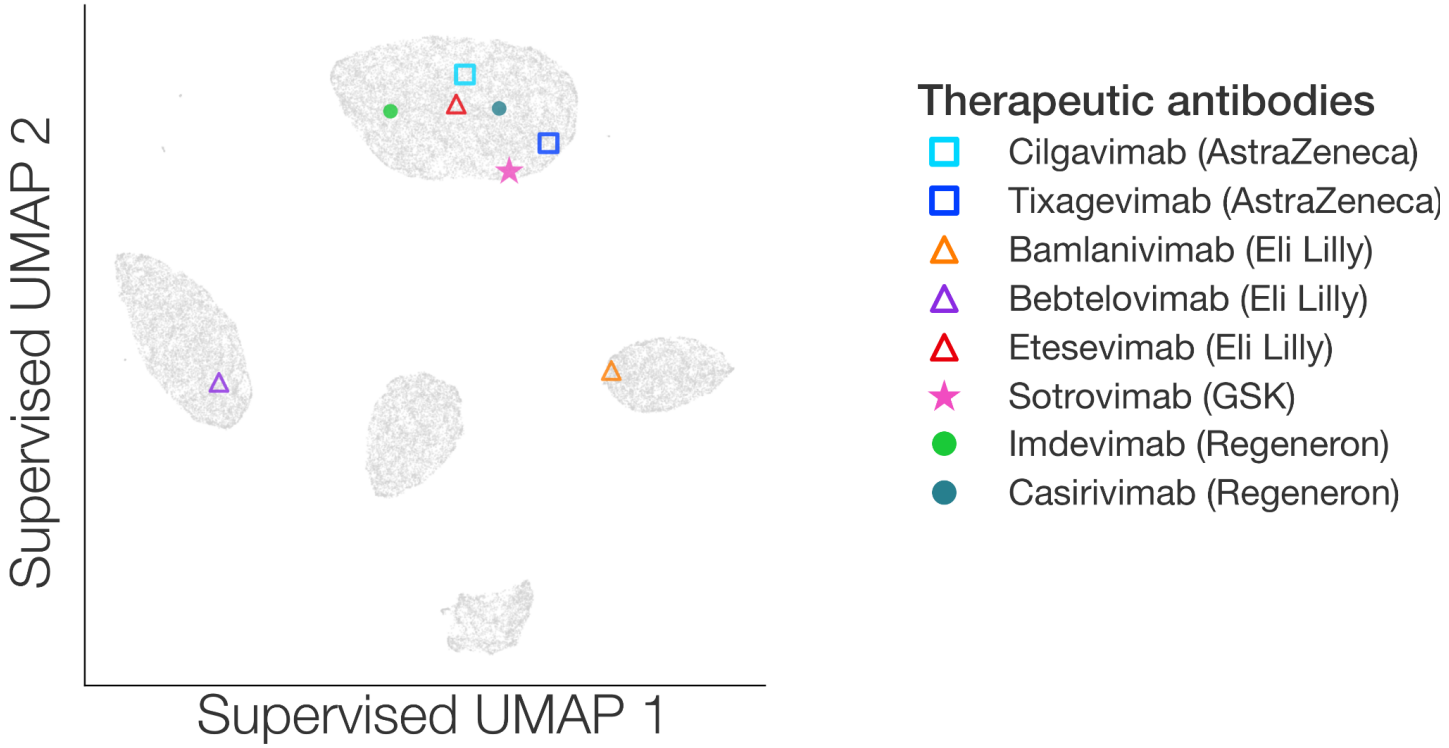
Supplementary Figure 6

Disease patient-originating sequences ranked by predicted disease class probability and grouped by isotype. Significance was tested for each isotype pair in each panel. **** means p <= 1e-4 by two-sided Wilcoxon rank-sum test, with Bonferroni multiple hypothesis testing correction across all tests in all panels.



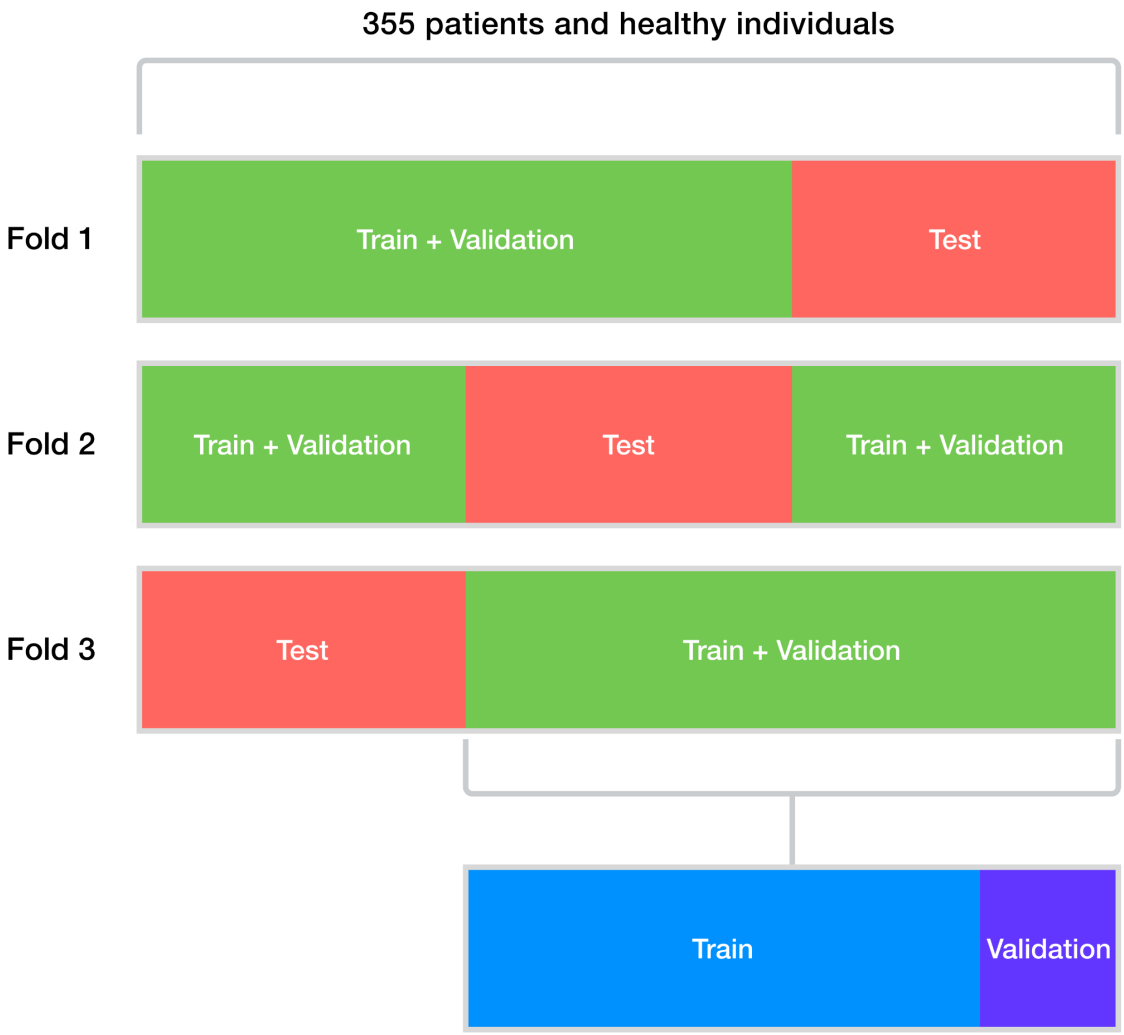
Supplementary Figure 7

Therapeutic monoclonal antibody sequences projected onto the reference UMAP (**Figure 4A**).



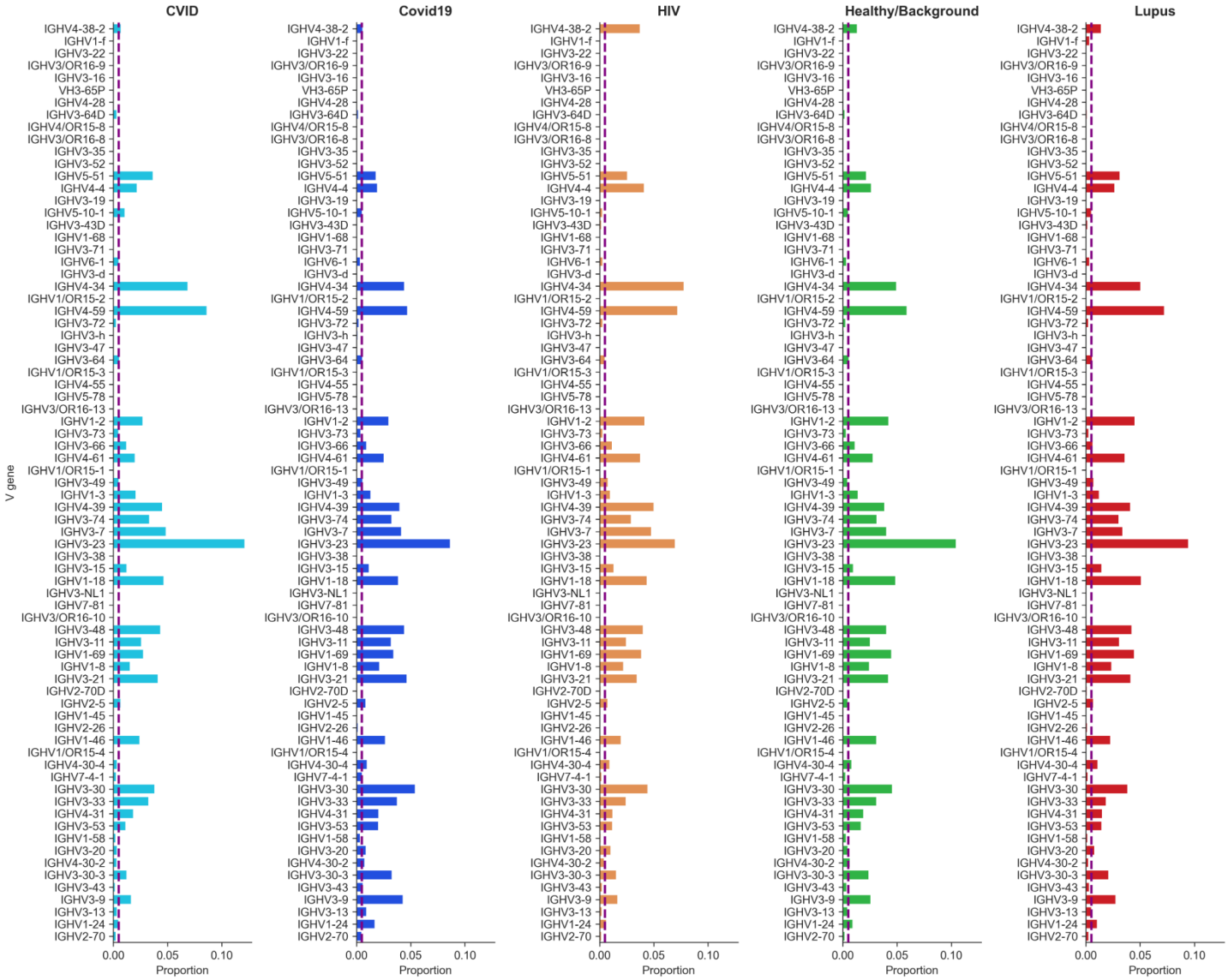
Supplementary Figure 8

Schematic of cross validation strategy.



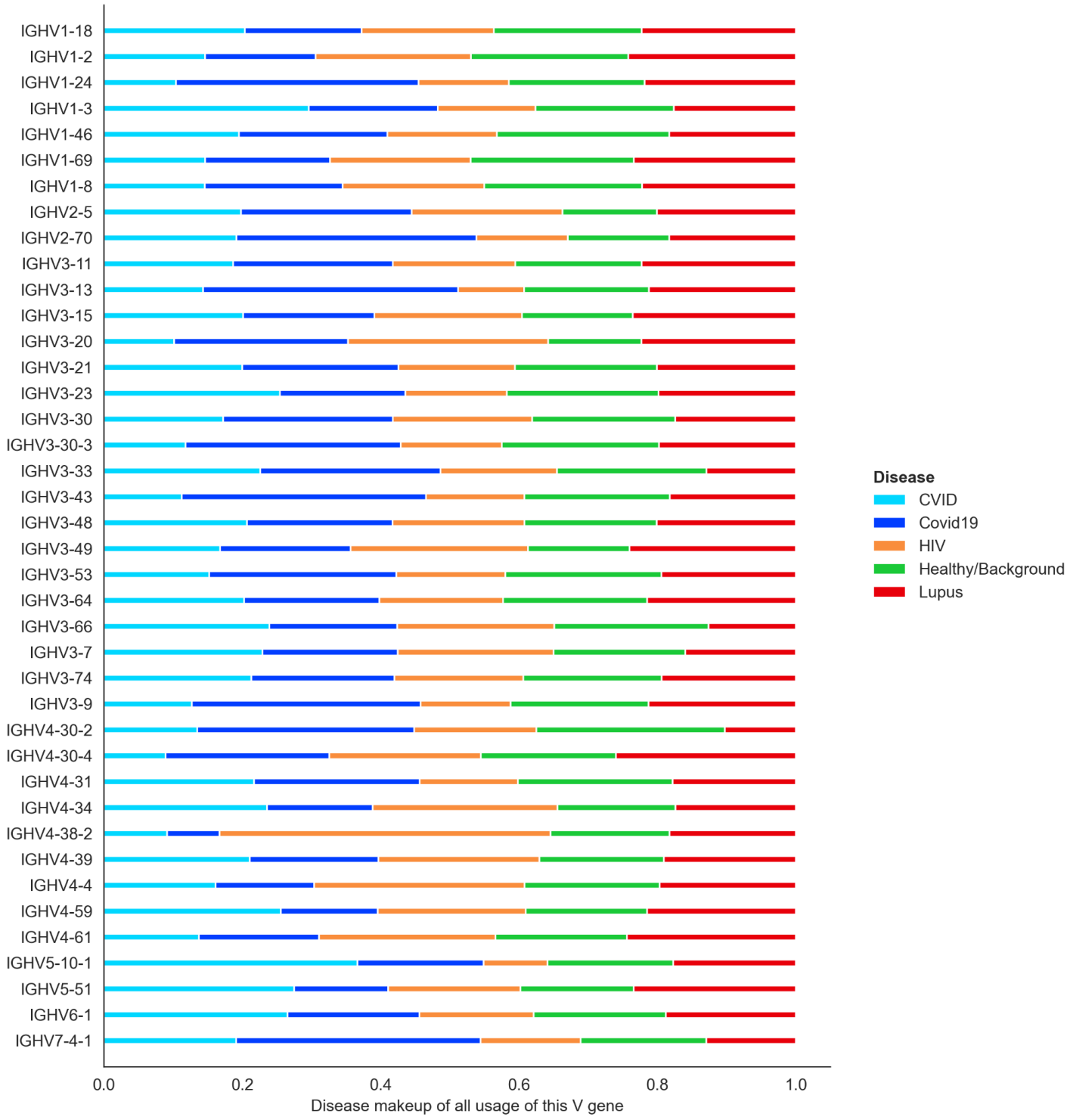
Supplementary Figure 9

IGHV gene proportions in each cohort. We also calculated the highest proportion each IGHV gene represents of any disease cohort, and plotted the median of these proportions (overlaid dashed line). Rare IGHV genes that did not exceed the purple dashed line in at least one disease were then filtered out.



Supplementary Figure 10

Stacked bar plots representing how prevalent each IGHV gene is by disease, after filtering out rare IGHV genes.



References

1. Charlton CL, Babady E, Ginocchio CC, Hatchette TF, Jerris RC, Li Y, Loeffelholz M, McCarter YS, Miller MB, Novak-Weekley S, Schuetz AN, Tang Y-W, Widen R, Drews SJ. Practical Guidance for Clinical Microbiology Laboratories: Viruses Causing Acute Respiratory Tract Infections. Clin Microbiol Rev [Internet]. 2019 Jan;32(1). Available from: http://dx.doi.org/10.1128/CMR.00042-18 PMCID: PMC6302358

2. Hurt CB, Nelson JAE, Hightow-Weidman LB, Miller WC. Selecting an HIV Test: A Narrative Review for Clinicians and Researchers. Sex Transm Dis. 2017 Dec;44(12):739–746. PMCID: PMC5718364

3. Milo R, Miller A. Revised diagnostic criteria of multiple sclerosis. Autoimmun Rev. 2014 Apr;13(4-5):518–524. PMID: 24424194

4. Fava A, Petri M. Systemic lupus erythematosus: Diagnosis and clinical management. J Autoimmun. 2019 Jan;96:1–13. PMCID: PMC6310637

5. Nielsen SCA, Boyd SD. Human adaptive immune receptor repertoire analysis-Past, present, and future. Immunol Rev. Wiley; 2018 Jul;284(1):9–23. PMID: 29944765

6. Boyd SD, Crowe JE Jr. Deep sequencing and human antibody repertoire analysis. Curr Opin Immunol. 2016 Jun;40:103–109. PMCID: PMC5203765

7. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. Front Immunol. 2018 Feb 21;9:224. PMCID: PMC5826328

8. van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, García-Sanz R, van Krieken JHJM, Droese J, González D, Bastard C, White HE, Spaargaren M, González M, Parreira A, Smith JL, Morgan GJ, Kneba M, Macintyre EA. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia. 2003 Dec;17(12):2257–2317. PMID: 14671650

9. Ching T, Duncan ME, Newman-Eerkes T, McWhorter MME, Tracy JM, Steen MS, Brown RP, Venkatasubbarao S, Akers NK, Vignali M, Moorhead ME, Watson D, Emerson RO, Mann TP, Cimler BM, Swatkowski PL, Kirsch IR, Sang C, Robins HS, Howie B, Sherwood A. Analytical evaluation of the clonoSEQ Assay for establishing measurable (minimal) residual disease in acute lymphoblastic leukemia, chronic lymphocytic leukemia, and multiple myeloma. BMC Cancer. Springer Science and Business Media LLC; 2020 Jun 30;20(1):612. PMCID: PMC7325652

10. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, Rieder M, Robins HS. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet. 2017 May;49(5):659–665. PMID: 28369038

11. Roskin KM, Jackson KJL, Lee J-Y, Hoh RA, Joshi SA, Hwang K-K, Bonsignori M, Pedroza-Pacheco I, Liao H-X, Moody MA, Fire AZ, Borrow P, Haynes BF, Boyd SD. Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. Nat Immunol. 2020 Feb;21(2):199–209. PMCID: PMC7223457

12. Yang F, Nielsen SCA, Hoh RA, Röltgen K, Wirz OF, Haraguchi E, Jean GH, Lee J-Y, Pham TD, Jackson KJL, Roskin KM, Liu Y, Nguyen K, Ohgami RS, Osborne EM, Nadeau KC, Niemann CU, Parsonnet J, Boyd SD. Shared B cell memory to coronaviruses and other pathogens varies in human age groups and tissues. Science. 2021 May 14;372(6543):738–741. PMCID: PMC8139427

13. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF, Walter EB, Liao H-X, Albrecht RA, García-Sastre A, Chaparro-Riggers J, Rajpal A, Pons J, Simen BB, Hanczaruk B, Dekker CL, Laserson J, Koller D, Davis MM, Fire AZ, Boyd SD. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe. 2014 Jul 9;16(1):105–114. PMCID: PMC4158033

14. Greiff V, Yaari G, Cowell LG. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. Current Opinion in Systems Biology. 2020 Dec 1;24:109–119.

15. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, Davis MM. Identifying specificity groups in the T cell receptor repertoire. Nature. 2017 Jul 6;547(7661):94–98. PMCID: PMC5794212

16. Roskin KM, Simchoni N, Liu Y, Lee J-Y, Seo K, Hoh RA, Pham T, Park JH, Furman D, Dekker CL, Davis MM, James JA, Nadeau KC, Cunningham-Rundles C, Boyd SD. IgH sequences in common variable immune deficiency reveal altered B cell development and selection. Sci Transl Med. 2015 Aug 26;7(302):302ra135. PMCID: PMC4584259

17. Davis CW, Jackson KJL, McElroy AK, Halfmann P, Huang J, Chennareddy C, Piper AE, Leung Y, Albariño CG, Crozier I, Ellebedy AH, Sidney J, Sette A, Yu T, Nielsen SCA, Goff AJ, Spiropoulou CF, Saphire EO, Cavet G, Kawaoka Y, Mehta AK, Glass PJ, Boyd SD, Ahmed R. Longitudinal Analysis of the Human B Cell Response to Ebola Virus Infection. Cell. 2019 May 30;177(6):1566–1582.e17. PMCID: PMC6908968

18. Nielsen SCA, Yang F, Jackson KJL, Hoh RA, Röltgen K, Jean GH, Stevens BA, Lee J-Y, Rustagi A, Rogers AJ, Powell AE, Hunter M, Najeeb J, Otrelo-Cardoso AR, Yost KE, Daniel B, Nadeau KC, Chang HY, Satpathy AT, Jardetzky TS, Kim PS, Wang TT, Pinsky BA, Blish CA, Boyd SD. Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. Cell Host Microbe. 2020 Oct 7;28(4):516–525.e5. PMCID: PMC7470783

19. Gaebler C, Wang Z, Lorenzi JCC, Muecksch F, Finkin S, Tokuyama M, Cho A, Jankovic M, Schaefer-Babajew D, Oliveira TY, Cipolla M, Viant C, Barnes CO, Bram Y, Breton G, Hägglöf T, Mendoza P, Hurley A, Turroja M, Gordon K, Millard KG, Ramos V, Schmidt F, Weisblum Y, Jha D, Tankelevich M, Martinez-Delgado G, Yee J, Patel R, Dizon J, Unson-O'Brien C, Shimeliovich I, Robbiani DF, Zhao Z, Gazumyan A, Schwartz RE, Hatziioannou T, Bjorkman PJ, Mehandru S, Bieniasz PD, Caskey M, Nussenzweig MC. Evolution of antibody immunity to SARS-CoV-2. Nature. 2021 Mar;591(7851):639–644. PMCID: PMC8221082

20. Sevy AM, Soto C, Bombardi RG, Meiler J, Crowe JE Jr. Immune repertoire fingerprinting by principal component analysis reveals shared features in subject groups with common exposures. BMC Bioinformatics. 2019 Dec 4;20(1):629. PMCID: PMC6894320

21. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. BMC Bioinformatics. 2017 Mar 7;18(1):155. PMCID: PMC5340033

22. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Mach Learn. 2001 Nov 1;45(2):171–186.

23. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017 Jul 6;547(7661):89–93. PMCID: PMC5616171

24. Marks C, Deane CM. How repertoire data are changing antibody science. J Biol Chem. 2020 Jul 17;295(29):9823–9837. PMCID: PMC7380193

25. Wu NC, Yuan M, Liu H, Lee C-CD, Zhu X, Bangaru S, Torres JL, Caniels TG, Brouwer PJM, van Gils MJ,

Sanders RW, Ward AB, Wilson IA. An Alternative Binding Mode of IGHV3-53 Antibodies to the SARS-CoV-2 Receptor Binding Domain. Cell Rep. 2020 Oct 20;33(3):108274. PMCID: PMC7522650

26. Ostmeyer J, Christley S, Toby IT, Cowell LG. Biophysicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue. Cancer Res. 2019 Apr 1;79(7):1671–1680. PMCID: PMC6445742

27. Konishi H, Komura D, Katoh H, Atsumi S, Koda H, Yamamoto A, Seto Y, Fukayama M, Yamaguchi R, Imoto S, Others. Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning. BMC Bioinformatics. Springer; 2019;20(1):1–11.

28. Beshnova D, Ye J, Onabolu O, Moon B, Zheng W, Fu Y-X, Brugarolas J, Lea J, Li B. De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. Sci Transl Med [Internet]. 2020 Aug 19;12(557). Available from: http://dx.doi.org/10.1126/scitranslmed.aaz3738 PMCID: PMC7887928

29. Shemesh O, Polak P, Lundin KEA, Sollid LM, Yaari G. Machine Learning Analysis of Naïve B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. Front Immunol. 2021 Mar 10;12:627813. PMCID: PMC8006302

30. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019 Dec;16(12):1315–1322. PMCID: PMC7067682

31. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A [Internet]. 2021 Apr 13;118(15). Available from: http://dx.doi.org/10.1073/pnas.2016239118 PMCID: PMC8053943

32. Friedensohn S, Neumeier D, Khan TA, Csepregi L, Parola C, de Vries ARG, Erlach L, Mason DM, Reddy ST. Convergent selection in antibody repertoires is revealed by deep learning [Internet]. Available from: http://dx.doi.org/10.1101/2020.02.25.965673

33. Sidhom J-W, Larman HB, Pardoll DM, Baras AS. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. Nat Commun. 2021 Mar 11;12(1):1605. PMCID: PMC7952906

34. Widrich M, Schäfl B, Pavlović M, Ramsauer H, Gruber L, Holzleitner M, Brandstetter J, Sandve GK, Greiff V, Hochreiter S, Klambauer G. Modern Hopfield Networks and Attention for Immune Repertoire Classification [Internet]. Available from: http://dx.doi.org/10.1101/2020.04.12.038158

35. Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov. Wiley; 2018 Jul;8(4):e1249.

36. Alpert A, Pickman Y, Leipold M, Rosenberg-Hasson Y, Ji X, Gaujoux R, Rabani H, Starosvetsky E, Kveler K, Schaffert S, Furman D, Caspi O, Rosenschein U, Khatri P, Dekker CL, Maecker HT, Davis MM, Shen-Orr SS. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. Nat Med. 2019 Mar;25(3):487–495. PMCID: PMC6686855

37. Sayed N, Huang Y, Nguyen K, Krejciova-Rajaniemi Z, Grawe AP, Gao T, Tibshirani R, Hastie T, Alpert A, Cui L, Kuznetsova T, Rosenberg-Hasson Y, Ostan R, Monti D, Lehallier B, Shen-Orr SS, Maecker HT, Dekker CL, Wyss-Coray T, Franceschi C, Jojic V, Haddad F, Montoya JG, Wu JC, Davis MM, Furman D. An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging. Nat Aging. 2021 Jul;1:598–615. PMCID: PMC8654267

38. Gostic KM, Ambrose M, Worobey M, Lloyd-Smith JO. Potent protection against H5N1 and H7N9 influenza

via childhood hemagglutinin imprinting. Science. 2016 Nov 11;354(6313):722–726. PMCID: PMC5134739

39. Weckerle CE, Niewold TB. The unexplained female predominance of systemic lupus erythematosus: clues from genetic and cytokine studies. Clin Rev Allergy Immunol. 2011 Feb;40(1):42–49. PMCID: PMC2891868

40. Yuan M, Liu H, Wu NC, Lee C-CD, Zhu X, Zhao F, Huang D, Yu W, Hua Y, Tien H, Rogers TF, Landais E, Sok D, Jardine JG, Burton DR, Wilson IA. Structural basis of a shared antibody response to SARS-CoV-2. Science. 2020 Aug 28;369(6507):1119–1123. PMCID: PMC7402627

41. Kim C, Ryu D-K, Lee J, Kim Y-I, Seo J-M, Kim Y-G, Jeong J-H, Kim M, Kim J-I, Kim P, Bae JS, Shim EY, Lee MS, Kim MS, Noh H, Park G-S, Park JS, Son D, An Y, Lee JN, Kwon K-S, Lee J-Y, Lee H, Yang J-S, Kim K-C, Kim SS, Woo H-M, Kim J-W, Park M-S, Yu K-M, Kim S-M, Kim E-H, Park S-J, Jeong ST, Yu CH, Song Y, Gu SH, Oh H, Koo B-S, Hong JJ, Ryu C-M, Park WB, Oh M-D, Choi YK, Lee S-Y. A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein. Nat Commun. 2021 Jan 12;12(1):288. PMCID: PMC7803729

42. Cerutti G, Guo Y, Zhou T, Gorman J, Lee M, Rapp M, Reddem ER, Yu J, Bahna F, Bimela J, Huang Y, Katsamba PS, Liu L, Nair MS, Rawi R, Olia AS, Wang P, Zhang B, Chuang G-Y, Ho DD, Sheng Z, Kwong PD, Shapiro L. Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. Cell Host Microbe. 2021 May 12;29(5):819–833.e7. PMCID: PMC7953435

43. Townsley SM, Donofrio GC, Jian N, Leggat DJ, Dussupt V, Mendez-Rivera L, Eller LA, Cofer L, Choe M, Ehrenberg PK, Geretz A, Gift S, Grande R, Lee A, Peterson C, Piechowiak MB, Slike BM, Tran U, Joyce MG, Georgiev IS, Rolland M, Thomas R, Tovanabutra S, Doria-Rose NA, Polonis VR, Mascola JR, McDermott AB, Michael NL, Robb ML, Krebs SJ. B cell engagement with HIV-1 founder virus envelope predicts development of broadly neutralizing antibodies. Cell Host Microbe. 2021 Apr 14;29(4):564–578.e9. PMCID: PMC8245051

44. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, Wilson RK, Holt RA, Eichler EE, Breden F. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. Am J Hum Genet. 2013 Apr 4;92(4):530–546.

45. Fraser NLW, Rowley G, Field M, Stott DI. The VH gene repertoire of splenic B cells and somatic hypermutation in systemic lupus erythematosus. Arthritis Res Ther. 2003 Feb 3;5(2):R114–21. PMCID: PMC165037

46. Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, Lee JC, Thomas DC, Flint SM, Kellam P, Jayne DRW, Lyons PA, Smith KGC. Analysis of the B cell receptor repertoire in six immune-mediated diseases. Nature. 2019 Oct;574(7776):122–126. PMCID: PMC6795535

47. Pugh-Bernard AE, Silverman GJ, Cappione AJ, Villano ME, Ryan DH, Insel RA, Sanz I. Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. J Clin Invest. 2001 Oct;108(7):1061–1070. PMCID: PMC200949

48. Cappione A 3rd, Anolik JH, Pugh-Bernard A, Barnard J, Dutcher P, Silverman G, Sanz I. Germinal center exclusion of autoreactive B cells is defective in human systemic lupus erythematosus. J Clin Invest. 2005 Nov;115(11):3205–3216. PMCID: PMC1242189

49. Reed JH, Jackson J, Christ D, Goodnow CC. Clonal redemption of autoantibodies by somatic hypermutation away from self-reactivity during human immunization. J Exp Med. 2016 Jun 27;213(7):1255–1265. PMCID: PMC4925023

50. Schultheiß C, Simnica D, Willscher E, Oberle A, Fanchi L, Bonzanni N, Wildner NH, Schulze Zur Wiesch J, Weiler-Normann C, Lohse AW, Binder M. Next-generation immunosequencing reveals pathological T-

cell architecture in autoimmune hepatitis. Hepatology. Wiley; 2021 Apr;73(4):1436–1448. PMID: 32692457

51. Snir O, Mesin L, Gidoni M, Lundin KEA, Yaari G, Sollid LM. Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. J Immunol. 2015 Jun 15;194(12):5703–5712. PMID: 25972486

52. Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, Grifoni A, Ramirez SI, Haupt S, Frazier A, Nakao C, Rayaprolu V, Rawlings SA, Peters B, Krammer F, Simon V, Saphire EO, Smith DM, Weiskopf D, Sette A, Crotty S. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. Science [Internet]. 2021 Feb 5;371(6529). Available from: http://dx.doi.org/10.1126/science.abf4063 PMCID: PMC7919858

53. Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, Wu JE, Alanio C, Kuri-Cervantes L, Pampena MB, D'Andrea K, Manne S, Chen Z, Huang YJ, Reilly JP, Weisman AR, Ittner CAG, Kuthuru O, Dougherty J, Nzingha K, Han N, Kim J, Pattekar A, Goodwin EC, Anderson EM, Weirick ME, Gouma S, Arevalo CP, Bolton MJ, Chen F, Lacey SF, Ramage H, Cherry S, Hensley SE, Apostolidis SA, Huang AC, Vella LA, UPenn COVID Processing Unit, Betts MR, Meyer NJ, Wherry EJ. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science [Internet]. 2020 Sep 4;369(6508). Available from: http://dx.doi.org/10.1126/science.abc8511 PMCID: PMC7402624

54. Robbiani DF, Gaebler C, Muecksch F, Lorenzi JCC, Wang Z, Cho A, Agudelo M, Barnes CO, Gazumyan A, Finkin S, Hägglöf T, Oliveira TY, Viant C, Hurley A, Hoffmann H-H, Millard KG, Kost RG, Cipolla M, Gordon K, Bianchini F, Chen ST, Ramos V, Patel R, Dizon J, Shimeliovich I, Mendoza P, Hartweger H, Nogueira L, Pack M, Horowitz J, Schmidt F, Weisblum Y, Michailidis E, Ashbrook AW, Waltari E, Pak JE, Huey-Tubman KE, Koranda N, Hoffman PR, West AP Jr, Rice CM, Hatziioannou T, Bjorkman PJ, Bieniasz PD, Caskey M, Nussenzweig MC. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. Nature. 2020 Aug;584(7821):437–442. PMCID: PMC7442695

55. Kreer C, Zehner M, Weber T, Ercanoglu MS, Gieselmann L, Rohde C, Halwe S, Korenkov M, Schommers P, Vanshylla K, Di Cristanziano V, Janicki H, Brinker R, Ashurov A, Krähling V, Kupke A, Cohen-Dvashi H, Koch M, Eckert JM, Lederer S, Pfeifer N, Wolf T, Vehreschild MJGT, Wendtner C, Diskin R, Gruell H, Becker S, Klein F. Longitudinal Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19 Patients. Cell. 2020 Sep 17;182(6):1663–1673. PMCID: PMC7497397

56. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody database. Bioinformatics. 2021 May 5;37(5):734–735. PMCID: PMC7558925

57. Zost SJ, Gilchuk P, Chen RE, Case JB, Reidy JX, Trivette A, Nargi RS, Sutton RE, Suryadevara N, Chen EC, Binshtein E, Shrihari S, Ostrowski M, Chu HY, Didier JE, MacRenaris KW, Jones T, Day S, Myers L, Eun-Hyung Lee F, Nguyen DC, Sanz I, Martinez DR, Rothlauf PW, Bloyet L-M, Whelan SPJ, Baric RS, Thackray LB, Diamond MS, Carnahan RH, Crowe JE Jr. Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. Nat Med. 2020 Sep;26(9):1422–1427. PMCID: PMC8194108

58. Corti D, Purcell LA, Snell G, Veesler D. Tackling COVID-19 with neutralizing monoclonal antibodies. Cell. 2021 Aug 19;184(17):4593–4595. PMCID: PMC8375150

59. Raybould MIJ, Marks C, Lewis AP, Shi J, Bujotzek A, Taddese B, Deane CM. Thera-SAbDab: the Therapeutic Structural Antibody Database. Nucleic Acids Res. 2020 Jan 8;48(D1):D383–D388. PMCID: PMC6943036

60. Westendorf K, Žentelis S, Wang L, Foster D, Vaillancourt P, Wiggin M, Lovett E, van der Lee R, Hendle J, Pustilnik A, Michael Sauder J, Kraft L, Hwang Y, Siegel RW, Chen J, Heinz BA, Higgs RE, Kallewaard NL, Jepson K, Goya R, Smith MA, Collins DW, Pellacani D, Xiang P, de Puyraimond V, Ricicova M, Devorkin L, Pritchard C, O'Neill A, Dalal K, Panwar P, Dhupar H, Garces FA, Cohen CA, Dye JM, Huie KE, Badger

CV, Kobasa D, Audet J, Freitas JJ, Hassanali S, Hughes I, Munoz L, Palma HC, Ramamurthy B, Cross RW, Geisbert TW, Menacherry V, Lokugamage K, Borisevich V, Lanz I, Anderson L, Sipahimalani P, Corbett KS, Yang ES, Zhang Y, Shi W, Zhou T, Choe M, Misasi J, Kwong PD, Sullivan NJ, Graham BS, Fernandez TL, Hansen CL, Falconer E, Mascola JR, Jones BE, Barnhart BC. LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants [Internet]. Available from: http://dx.doi.org/10.1101/2021.04.30.442182

61. Su Y, Yuan D, Chen DG, Ng RH, Wang K, Choi J, Li S, Hong S, Zhang R, Xie J, Kornilov SA, Scherler K, Pavlovitch-Bedzyk AJ, Dong S, Lausted C, Lee I, Fallen S, Dai CL, Baloni P, Smith B, Duvvuri VR, Anderson KG, Li J, Yang F, Duncombe CJ, McCulloch DJ, Rostomily C, Troisch P, Zhou J, Mackay S, DeGottardi Q, May DH, Taniguchi R, Gittelman RM, Klinger M, Snyder TM, Roper R, Wojciechowska G, Murray K, Edmark R, Evans S, Jones L, Zhou Y, Rowen L, Liu R, Chour W, Algren HA, Berrington WR, Wallick JA, Cochran RA, Micikas ME, Petropoulos CJ, Cole HR, Fischer TD, Wei W, Hoon DSB, Price ND, Subramanian N, Hill JA, Hadlock J, Magis AT, Ribas A, Lanier LL, Boyd SD, Bluestone JA, Chu H, Hood L, Gottardo R, Greenberg PD, Davis MM, Goldman JD, Heath JR. Multiple Early Factors Anticipate Post-Acute COVID-19 Sequelae. Cell [Internet]. 2022 Jan 25; Available from: https://www.sciencedirect.com/science/article/pii/S0092867422000721

62. Bjornevik K, Cortese M, Healy BC, Kuhle J, Mina MJ, Leng Y, Elledge SJ, Niebuhr DW, Scher AI, Munger KL, Ascherio A. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science. 2022 Jan 21;375(6578):296–301. PMID: 35025605

63. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W34–40. PMCID: PMC3692102

64. Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY; 2001.

65. Ma EJ, Kummer A. Reimplementing Unirep in JAX [Internet]. bioRxiv. 2020. p. 2020.05.11.088344. Available from: https://www.biorxiv.org/content/biorxiv/early/2020/05/13/2020.05.11.088344

66. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Jan 2;21(1):6. PMCID: PMC6941312

67. Schuster EF, Sype WR. On the negative hypergeometric distribution. Internat J Math Ed Sci Tech. Taylor & Francis; 1987 May 1;18(3):453–459.

68. Johnson NL, Kotz S, Kemp AW. Univariate Discrete Distributions. John Wiley & Sons; 2005.

69. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv [stat.ML]. 2018. Available from: http://arxiv.org/abs/1802.03426