Disease diagnostics using machine learning of immune receptors

Maxim E. Zaslavsky[1], Erin Craig[2], Jackson K. Michuda[2], Nikhil Ram-Mohan[3], Ji-Yeun Lee[4], Khoa D. Nguyen[4], Ramona A. Hoh[4], Tho D. Pham[4,5], Ella S. Parsons[6], Susan R. Macwana[7], Wade DeJager[7], Krishna M. Roskin[8,9], Charlotte Cunningham-Rundles[10], M. Anthony Moody[11,12,13], Barton F. Haynes[12,13,14], Jason D. Goldman[15,16], James R. Heath[17,18], Imelda Balboni[19], Paul J Utz[20], Kari C. Nadeau[6,20], Benjamin A. Pinsky[4,20], Catherine A. Blish[20], Joan T. Merrill[7], Joel M. Guthridge[7], Judith A. James[7], Samuel Yang[3], Robert Tibshirani[2,21], Anshul Kundaje[1,22,*,§], Scott D. Boyd[4,6,*,§]

[1] Department of Computer Science, Stanford University, Stanford, CA, USA

[2] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[3] Department of Emergency Medicine, Stanford University, Stanford, CA, USA

[4] Department of Pathology, Stanford University, Stanford, CA, USA

[5] Stanford Blood Center, Stanford, CA, USA

[6] Sean N. Parker Center for Allergy and Asthma Research, Stanford University, Stanford, CA, USA

[7] Department of Arthritis and Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA

[8] Department of Pediatrics, University of Cincinnati, College of Medicine, Cincinnati, OH, USA

[9] Divisions of Biomedical Informatics and Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[10] Icahn School of Medicine at Mount Sinai, New York, NY, USA

[11] Department of Pediatrics, Duke University, Durham, NC, USA

[12] Duke Human Vaccine Institute, Duke University, Durham, NC, USA

[13] Department of Immunology, Duke University, Durham, NC, USA

[14] Department of Medicine, Duke University, Durham, NC, USA

[15] Swedish Center for Research and Innovation, Swedish Medical Center, Seattle, WA, USA

[16] Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, USA

[17] Institute for Systems Biology, Seattle, WA, USA

[18] Department of Bioengineering, University of Washington, Seattle, WA, USA

[19] Department of Pediatrics, Stanford University, Stanford, CA, USA

[20] Department of Medicine, Stanford University, Stanford, CA, USA

[21] Department of Statistics, Stanford University, Stanford, CA, USA

[22] Department of Genetics, Stanford University, Stanford, CA, USA

[§] 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 findings and the patient's history and symptoms. Currently, the tools of diagnosis make limited 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, and there has been little integration of the combined information from B cell and T cell receptor sequences. Here, we analyze extensive receptor sequence datasets with three different machine learning representations of immune receptor repertoires to develop an interpretive framework, *MAchine Learning for Immunological Diagnosis (Mal-ID)*, that screens for multiple illnesses simultaneously. This approach is effective in identifying a variety of disease states, including acute and chronic infections and autoimmune disorders. It is able to do so even when there are other differences present in the immune repertoires, such as between pediatric or adult patient groups. 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 and HIV, provide evidence of receptor antigen specificity, and reveal common features of autoreactive immune receptor repertoires, indicating that machine learning on immune repertoires can yield new immunological knowledge. This framework could be useful in identifying immune responses to new infectious diseases as they emerge.

Main text

Modern medical diagnosis relies heavily on laboratory testing for cellular or molecular abnormalities in specimens from a patient, such as the presence of pathogenic microorganisms[1,2]. For autoimmune disorders like lupus or multiple sclerosis, diagnosis via a combination of patient history, physical examination, imaging observations, detection of autoantibodies and exclusion of other conditions can be a lengthy process[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 stimuli the repertoire of B and T cell receptors changes in composition by clonal expansion of antigen-specific cells, introduction of additional somatic mutations into B cell receptor genes, and selection processes that further reshape lymphocyte populations. In dysregulated 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 simultaneous assessment for many infectious and autoimmune diseases[5–7]. Tracking immune receptor repertoires has already proved useful in diagnosing lymphocyte malignancies and monitoring cancer treatment responses[8,9], and shows promise in the context of antibody-mediated pathologies[10]. Challenges in this field are the low frequency of antigen-specific BCRs and TCRs in many patients with acute infectious or autoimmune diseases, and the high complexity and diversity of immune receptor genes due to somatic gene rearrangement during lymphocyte development and somatic hypermutation after antigen stimulation of B cells[6,11]. Differences in sample types, timing, experimental protocols for sequence library preparation and the necessity of controlling for demographic and epidemiological factors may also influence the data[12]. Further limitations have been the relatively small sizes of human cohorts from which BCR and TCR sequence data have been collected, and incomplete knowledge about the relative importance of B cell compared to T cell responses in various immunological conditions. Some prior investigations of disease or vaccination-related immune repertoires have attempted to identify highly similar receptor sequences or subsequences in people with the same exposures[13–21], or represented receptor sequences with alternative encodings of amino acid biochemical properties such as charge and polarity to find receptor groups[22–25]. Learned representations of either TCR or

BCR sequences with language models and variational autoencoders are also candidates for immune state classification or for functional purposes such as therapeutic antibody optimization[26–34]. Additionally, probabilistic models of V(D)J recombination and selection processes have been proposed to improve interpretation of the stochastic nature of immune receptor generation and expansion in response to antigenic stimuli[35,36]. Despite these advances, it is still unclear to what extent immune repertoire sequence data are sufficient for generalized and accurate infectious or immunological disease classification in humans.

To overcome these challenges, we have developed *MAchine Learning for Immunological Diagnosis (Mal-ID),* which combines three machine learning representations applied to both B cell receptor (BCR) and T cell receptor (TCR) repertoires (**Figure 1**) to identify the presence of infectious or immunological diseases in patients. *Mal-ID* relies on several biologically informed representations of BCR and TCR data, from overview summary metrics of receptor populations to focused analysis of the key antigen-binding loops CDR1, CDR2, and CDR3 (complementarity regions 1, 2 and 3) with sequence distance measures and protein language modeling. We apply *Mal-ID* to systematically collected datasets of 14.3 million BCR heavy chain (IgH) clones and 19.2 million TCR beta chain (TRB) clones from peripheral blood samples of 461 individuals, as well as external datasets collected with different library preparation and sequencing protocols. *Mal-ID* distinguishes healthy from diseased individuals, viral infections from autoimmune conditions, and different infections from each other, without prior knowledge of pathogenesis or of which sequences are antigen specific. Importantly, this approach also generates interpretable rankings for disease-associated sequences, recapitulating independently discovered biological facts, including identifying SARS-CoV-2-specific antibodies and T cells.

Integrated repertoire models of disease states

Mal-ID uses a combination of three models per gene locus (BCR heavy chain, IgH; and TCR beta chain, TRB) to improve recognition of distinct kinds of disease states, and to identify candidate receptor sequences of lymphocytes stimulated by disease-related antigens. Each classifier model extracts different aspects of immune repertoires (**Figure 1b**). The first model uses variable gene IGHV or TRBV gene segment frequencies and IGHV mutation rates across a person’s IgH repertoire. The second predictor identifies groups of highly similar sequences across individuals. The third classifier evaluates a broader proxy for functional similarity based on protein language modeling, rather than direct sequence identity, to find more loosely related immune receptors with potential common antigen targets. We train disease predictors with each representation. The three BCR and three TCR models are then blended into a final prediction of immune status. The final trained program accepts an individual's collection of sequences from peripheral blood B and T cells as input, and returns a prediction of the probability the person has each disease on record (**Figure 1c**).

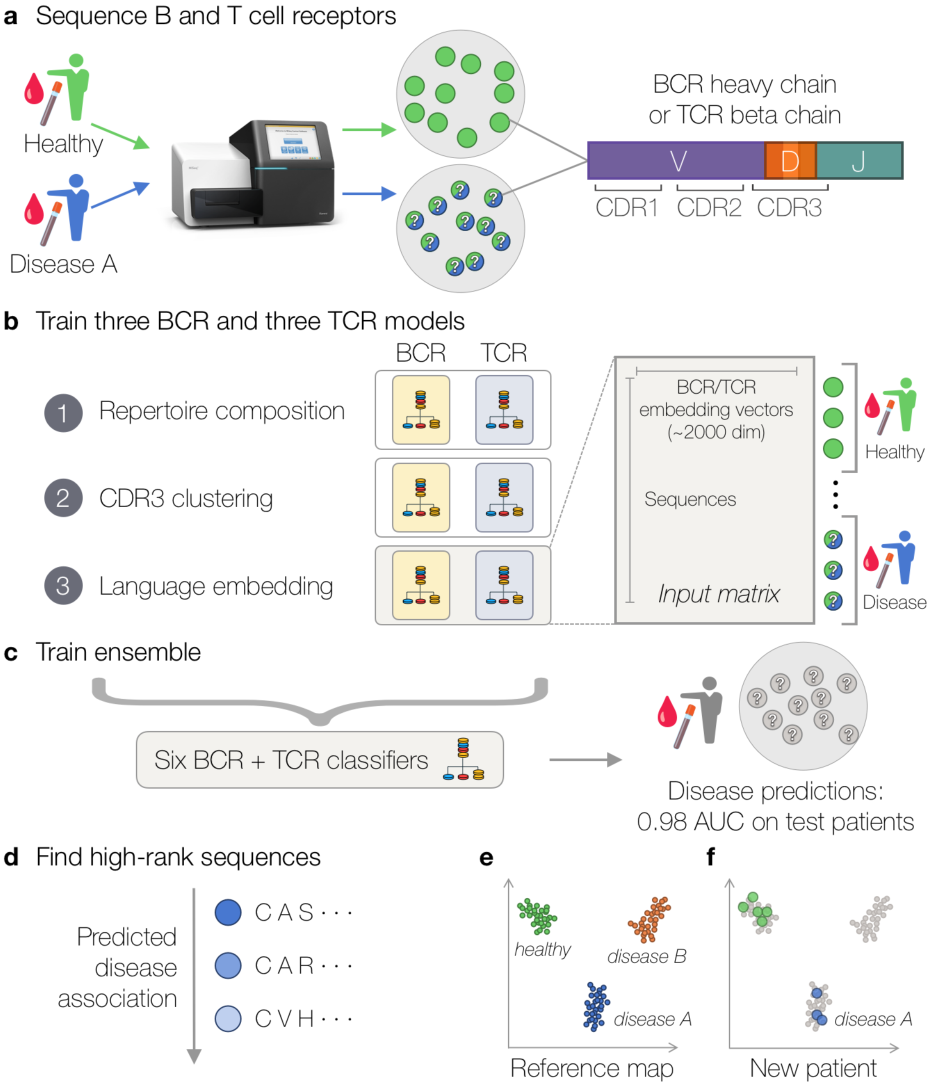


Figure 1: MAchine Learning for Immunological Diagnosis (*Mal-ID*) framework.

a, B and T cell receptor gene repertoires are amplified and sequenced from blood samples of individuals with different disease states. Question marks indicate that most sequences from patients are not disease specific.

b, Machine learning models are trained to predict disease using several immune repertoire feature representations. These include sequence feature extraction using language models fine-tuned to BCR heavy chain or TCR beta chain patterns. The language model feature extraction converts each amino acid sequence into a numerical vector.

c, An ensemble disease predictor is trained using the three BCR and three TCR base models. The combined model predicts disease status of held-out test individuals.

d, Suspected antigen-specific immune receptors are identified by ranking sequences according to their predicted disease association.

e, A reference map of immune receptor sequences is constructed Each point is one sequence.

f, Visualizing a held-out test patient's immune status by overlaying their sequences on the reference map. The immune response to disease A is visible in blue.

We applied this approach to cohorts of patients with diagnoses of Covid-19 (n=63), HIV (n=95)[14], and Systemic Lupus Erythematosus (SLE, n=86), and healthy controls (n=217), with 461 individuals in total (**Supplementary Table 1**). We combined new datasets with ones previously reported, all generated with a standardized sequencing protocol to minimize batch effects (**Methods**). The non-Covid-19 cohort samples 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 three training, validation, and testing sets, with each person falling into one test set (**Supplementary Figure 1**). Some patients had multiple samples; all were grouped together for the cross-validation divisions. We trained separate models for each cross-validation fold and report averaged classification performance. As described below, we also tested and excluded the possibility that demographic differences between cohorts could explain diagnosis accuracy.

Model 1: Overall repertoire composition. The first machine learning model uses an individual's IgH or TRB repertoire composition to predict disease status. Prior studies have reported immune status classification using deviations in B cell or T cell V(D)J recombination gene segment usage from healthy individuals[19,37,38]. Certain V gene segments may be more prevalent among antigen-responding V(D)J rearrangements than in the population of immune receptors in naive lymphocytes, and increase in frequency as antigen-specific cells become clonally expanded[39,40]. We previously identified class-switched IgH sequences with low somatic mutation (SHM) frequencies as prominent features of acute infection with Ebola virus or SARS-CoV-2, consistent with naive B cells recently having class-switched during the primary response to infection[39–41]. V gene usage changes and other repertoire changes have also been described in chronic infectious or immunological conditions[10,14]. We trained a lasso linear model with V/J gene counts in TRB and IgH data, and somatic hypermutation rate in IGHV, as features.

Model 2: 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, an approach we and others have previously reported[14,17,18]. The CDR3s are the highly variable regions of IgH and TRB that often determine antigen binding specificity. For each locus, we clustered CDR3 sequences with the same V gene, J gene, and CDR3 length that had high sequence identity, allowing for some variability created by somatic hypermutation in B cell receptors. A new sample's sequences can then be assigned to nearby clusters with the same constraints. We selected clusters enriched for sequences from subjects with a particular disease, using Fisher's exact test and setting a significance threshold based on cross-validation with data derived from different individuals. These clusters represent candidate sequences predictive of a specific disease across individuals. We assigned each sample's sequences to these predictive clusters. For each sample, we counted how many clusters associated with each disease were matched, and used these counts as features in a lasso linear model to predict immune status.

Model 3: Language model feature extraction from B and T cell receptor sequences. Immune receptor 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[42,43]. Disease-associated receptors may have apparently dissimilar sequences by edit distance but share the function of binding to the same target. Using language models fine-tuned on BCR and TCR sequences, the third classifier in our framework aims to map primary amino acid sequences into a lower-dimensional space with the potential to capture functional similarities, beyond sequence similarity represented by edit distance. We extracted a putative functional representation of BCRs and TCRs with 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[44,45]. Much as words are the building blocks arranged by grammatical rules to convey meaning, protein sequences are built from amino acids joined in an order compatible with polypeptide chain folding and assuming a structure that can carry out functions such as 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 short and long-range relationships between different sequence regions, analogous to learning natural language phrases and grammar rules to anticipate the next word in a sentence. 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 predict masked portions of protein sequences, the compression and uncompression has extracted fundamental features that summarize the input sequences.

UniRep's internal representation, trained on over 20 million proteins from many organisms[44], was shown to encode fundamental properties like structural classes[44].

To create a language model specialized for immune receptor proteins, we continued UniRep's training procedure separately for masked IgH or TRB sequences for each cross-validation fold (**Methods**). Prior autoencoder models have enabled classification of clusters of similar sequences[32,34]; notably, an advantage of our fine-tuning using BCR and TCR sequences of a language model based initially on global patterns in proteins from many domains of life is that the final model retains high performance on UniRep's original training data while showing improved prediction of BCR and TCR amino acid sequences, suggesting it combines global and domain-specific protein rules (**Supplementary Figure 2**). For disease classification, the low-dimensional embedding learned by the BCR or TCR fine-tuned language model transformed each sequence into a 1900-dimensional numerical feature vector, regardless of sequence length. We then trained a lasso linear model to map receptor sequence vectors to disease labels. Aggregating each sequence's predicted class probabilities using a trimmed mean, we obtained patient-level predictions of specific disease states. The trimmed mean was robust to noise in the model in the form of rare sequences with extremely high or low disease association probabilities, but other central estimates perform similarly for classification accuracy (**Supplementary Table 2**). Because this classifier starts with a predictor for individual receptors, then aggregates sequence calls into a patient-level prediction[22,33], it allows interpretation of which sequences matter most for prediction of each disease. Below, we confirmed that sequences prioritized by our predictor are enriched for disease-specific B and T cells, demonstrating that the language model learns the syntax of immune receptor sequences, in spite of their enormous diversity.

Ensemble: Finally, we combined all three classifiers (global repertoire composition, CDR3 sequence clustering, and language model embedding) for IgH and three for TRB into the final *Mal-ID* ensemble predictor of disease (**Supplementary Figure 3**). Blending the probabilistic outputs from multiple classifiers trained with different strategies, the metamodel exploits each predictor's strengths and can resolve mistakes[46]. As with the individual component models in *Mal-ID*, we trained a separate metamodel for each cross-validation group, maintaining strict separation of each individual's data into training, validation or test datasets.

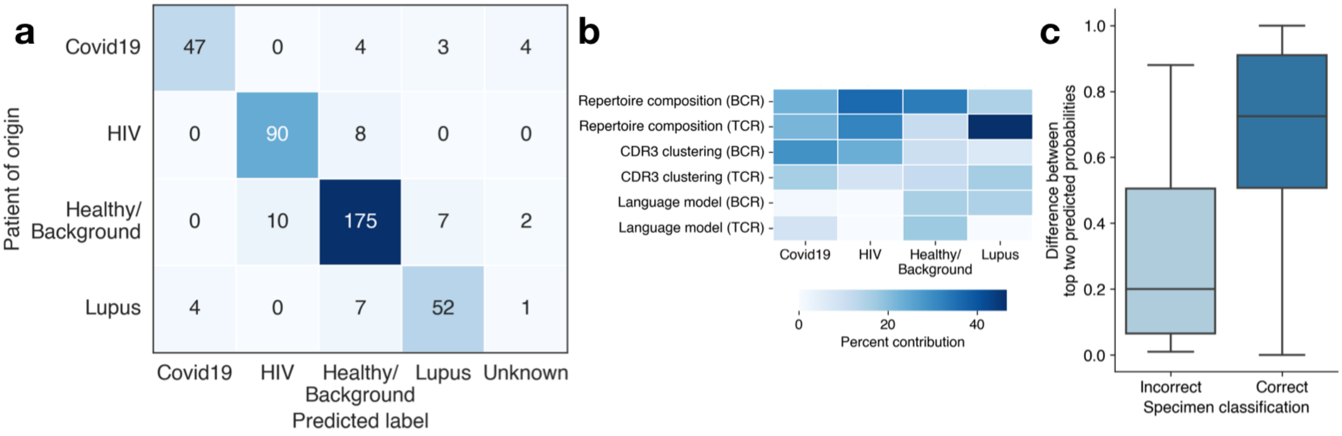


Figure 2: *Mal-ID* classifies disease using IgH and TRB sequences. **a**, Disease classification performance on held-out test data by the ensemble (random forest) of three B cell repertoire and three T cell repertoire machine learning models, combined over all cross-validation folds. **b**, Ensemble model (elastic net logistic regression fit on global fold) feature contributions for predicting each class, summarized by featurization method and whether the features were extracted from BCR or TCR information. To determine "percent contribution", feature coefficients were converted to absolute values, summed by featurization method (such as BCR repertoire composition classifier predicted probability-derived features), and divided by the sum of all coefficients. **c**, Difference of probabilities of the top two predicted classes for correct versus incorrect ensemble model (random forest) predictions. 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. Results were combined across all cross-validation folds.

This ensemble approach distinguished four specific disease states in 414 paired BCR and TCR samples from 410 individuals with an area under the Receiver Operating Characteristic curve (AUC) score of 0.98 (**Figure 2a**). In comparison, the previously reported CDR3 clustering model, with parallels to many convergent sequence discovery approaches in the literature, achieves only 0.93 AUC for BCR and 0.89 AUC for TCR. 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[47]. Other performance metrics are provided in **Supplementary Table 3**. Performance was also consistent across different types of ensemble models: a random forest metamodel achieves the highest accuracy, but an alternative metamodel using elastic net logistic regression performs similarly (**Supplementary Figure 4**). To achieve the significantly higher 0.98 AUC in the ensemble approach, all modeling strategies contributed to varying degrees depending on gene locus and disease, highlighting different strengths of BCR and TCR repertoire feature associations with each disease (**Figure 2b, Supplementary Figure 4**). The combined BCR+TCR metamodel outperforms BCR-only or TCR-only versions, highlighting the benefit of integrating signals from both B cell and T cell populations when such data is available (**Supplementary Table 3**). The *Mal-ID* ensemble model achieves 88.6% accuracy across all held-out test sets (**Figure 2a**). Of the 11.4% of misclassified repertoires, 1.7% were samples that did not have any sequences belonging to Model 2 CDR3 clusters. The CDR3 clustering component of the metamodel abstained from making any prediction for these challenging samples. In the remaining ~10% of classification mistakes, the ensemble model predictions failed to identify a clear winning label (**Figure 2c**). Allowing the strategy to abstain from inconclusive predictions is important for diagnostic robustness with challenging real-world cases. 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.

While cross-validation mitigates the risk of overfitting, we wanted to assess whether *Mal-ID* would generalize to new data from other sources. We fit a final model on all the data, which we call the “global fold”, to distinguish from the three cross-validation folds (**Supplementary Figure 1**). Then we downloaded Covid-19 patient and healthy donor repertoires from other BCR or TCR studies with similar cDNA sequencing protocols. In four external cohorts, two with only BCR sequences and the other two with only TCR sequences, *Mal-ID* predicted disease type with 100% and 86% accuracy, respectively (**Supplementary Table 4**). In both cases, the AUC is over 0.99, suggesting that the TCR accuracy of 86% may be improved by tuning the decision thresholds for choosing predicted labels based on the different base rates of disease in these outside data with only Covid-19 patients and healthy donors present, given that the AUC summarizes over all choices of probability thresholds for class label selection[48]. This ability to generalize to new datasets provides additional evidence that *Mal-ID* learns true biological disease-related signals, and that *Mal-ID* performs well when only BCR or only TCR data are available, rather than the preferred data including both receptor types. *Mal-ID* could also be fine-tuned to generalize to datasets from the many other sequencing protocols used by different laboratories, such as the genomic DNA-templated and normalized clone count data from Adaptive Biotechnologies[12], to address the differences in V gene usage (**Supplementary Figure 5a**).

Limited impact of age, sex, and race on classification

Besides diseases, patient demographics also shape the immune repertoire[49–52]. To study the degree to which extraneous covariates were confounding our disease classification results, we investigated whether we could distinguish age, sex, or ancestry of healthy individuals based on their immune receptor repertoire data. By training new classifiers to predict these variables, we found that the sex of a healthy individual could not accurately be determined from IgH or TRB sequences (**Supplementary Table 5**). However, sequences did carry a weak signal potentially related to ancestry, with 0.75 AUC predictive power. Ancestry separation is visible in IGHV and TRBV gene usage (**Supplementary Figure 5b**). Contributions to this signal may include germline TRB and IGHV locus differences, shaping of TCR repertoires by HLA alleles that differ between ancestry groups, and different environmental exposures in the African ancestry individuals living in Africa in the data[53,54]. In the full *Mal-ID* disease classification setting, the T cell model components had less accuracy in distinguishing HIV patients and healthy controls from this African cohort, though the corresponding IgH repertoires were distinct (**Supplementary Figure 6**), highlighting the advantage of incorporating both BCR and TCR information with an ensemble metamodel.

Previous studies have tracked age-related changes in gene expression, cytokine levels, and immune cell type frequencies[55,56]. We observe a modest signal of age in healthy IgH and TRB sequence repertoires. When we dichotomized age as under or over 50 years old to cast this continuous variable as a classification problem, the prediction model achieved 0.75 AUC (**Supplementary Table 5**). The signatures of age detected by the classifier may correspond to different historical infectious disease or environmental exposures for people over 50 versus younger individuals, such as imprinting effects on memory B cell and T cell pools related to different childhood virus exposures, as in the case of influenza viruses[57]. However, the Model 2 component in this age prediction model abstains on a high number of samples: 13% of repertoires had no sequences fall into age-associated CDR3 clusters. The AUC measure does not reflect this classification deficiency, because abstained samples have no predicted class probabilities and cannot be included in the computation of metrics that use predicted probabilities. On the other hand, every abstention hurts the accuracy metric: each one counts as a prediction error, so that the accuracy of predicting “under 50” versus “over 50” was 58.8%.

We also observed that V gene usage shows more defined age separation in TRB data than in IgH, particularly for pediatric compared to adult samples (**Supplementary Figure 5c**). The *Mal-ID* architecture can distinguish individuals under eighteen from those eighteen or older (78% accuracy including 17% abstentions, or 0.99 AUC not counting abstentions; **Supplementary Table 5**). Despite the substantial differences between the repertoires of adults and children that can be detected with this approach, age effects did not seem to interfere with disease classification, because *Mal-ID* distinguished pediatric lupus patients from healthy children (**Supplementary Figure 7**). Adult lupus patients were the most challenging to classify, with many predicted to be healthy individuals instead (**Supplementary Figure 7**), potentially reflecting the subset of patients with well-controlled disease in response to treatment[10]. More granular aging differences proved challenging to disentangle at the sequence level with the number of participants, age ranges, and cell sampling and sequencing depth in this study. When we divided age into groups by decade, the age prediction model achieved 37% accuracy and abstained from prediction on 18% of samples (0.70 AUC if not counting the abstentions). We restricted *Mal-ID*’s scope to somatically hypermutated IgD/IgM and class switched IgG/IgA isotypes, reflecting the populations of B cells that are most likely to be shaped by antigenic stimulation and selection. Studying naive B cells may reveal additional age, sex, or ancestry effects. The high abstention rates observed for Model 2 also suggest that finding convergent clusters of age-associated CDR3 sequences may be unrealistic, whereas a scan for global repertoire changes, like Model 1, may be better suited to demographic prediction tasks.

We further tested whether demographic differences between disease cohorts drove our classification results. For example, the age medians and ranges of the cohorts were: SLE (median 18 years, range 7-71); HIV (median 31 years, range 19-64); healthy controls (median 34.5 years, range 8-81); Covid-19 (median 48 years, range 21-88) (**Supplementary Table 1**). The percentage of females in each cohort was 50% (healthy controls), 52% (Covid-19), 64% (HIV), and 85% (SLE). The prevalence of females in our SLE cohort is consistent with general epidemiology for this disease[58]. The ancestries and geographical locations of participants also differed between cohorts. Notably, 89% of individuals with HIV were from Africa[14]. To address the extent to which demographic metadata could contribute to disease prediction in our current datasets, we attempted to predict disease state from age, sex, and ancestry alone, without using sequence data at all. The best disease classification AUC values were 0.70, 0.58, and 0.79 with only age, sex, or ancestry features, respectively. Combining all demographic features for a demographics-only classifier achieved an AUC of 0.86, substantially lower than the AUC of 0.98 when we retrained the *Mal-ID* sequence prediction ensemble with demographic covariates included as features, underscoring the disease signal we extract from BCR and TCR sequences (**Supplementary Table 6, Supplementary Figure 8a-b**). This demographics-only classifier also only achieved 0.77 and 0.68 AUC on the BCR and TCR external validation cohorts, respectively, compared to the >0.99 AUC performance of the standard *Mal-ID* model. As an additional version of this test, we also retrained the disease classification metamodel with age, sex, and ancestry effects regressed out from the ensemble feature matrix. After this correction, classification performance on the individuals with full demographic information available dropped slightly from 0.98 AUC to 0.97 AUC (**Supplementary Table 6, Supplementary Figure 8c**). 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. 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 BCR and TCR classifiers based on language model embeddings. In discriminating between different diseases, sequences highly prioritized for Covid-19 prediction used IGHV gene segments seen in independently isolated antibodies that bind SARS-CoV-2 spike antigen: IGHV3-30-3, IGHV3-9, and IGHV2-70[59–61] (**Figure 3b**). Similarly, IGHV1-24, found in a prominent class of N-terminal domain-directed antibodies, was highly ranked[62].

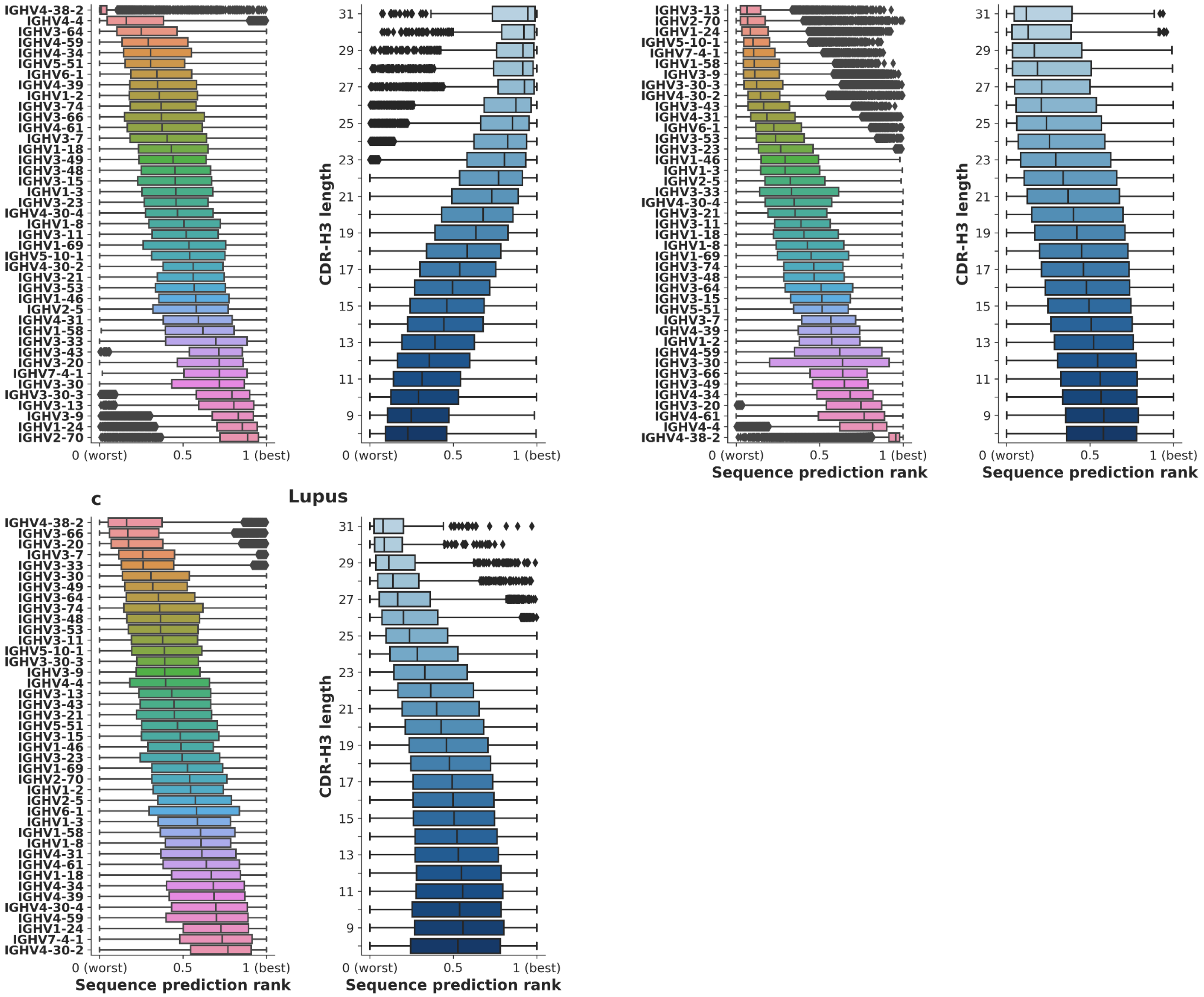


Figure 3: Disease patient-originating IgH sequences, ranked by predicted disease class probability, show high ranks for IGHV genes known to be disease-associated and for CDR-H3 length patterns reflecting selection. **a**, Covid-19 class prediction rankings; **b**, HIV class prediction rankings; **c**, SLE class prediction rankings. Ranks range from 0 (lowest disease association) to 1 (highest disease association). For each distribution, the box ranges from the 25th to 75th percentile, with the median marked. Whiskers extend to 1.5 times the interquartile range, and outlier points on the extremes are plotted individually.

The model's prioritization of IGHV4-34, IGHV4-39, and IGHV4-59 for SLE prediction (**Figure 3c**) also matches prior reports of higher frequency expression of these gene segments in SLE patients[10,63]. 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[14], was ranked highly for HIV classification by the model (**Figure 3b**). The IGHV4-38-2 V gene was also highly ranked for HIV prediction, consistent with its reported use in HIV-specific B cells in another analysis[64]; however, this is a case where this gene segment is more common in the IgH germline loci of African populations[65], underscoring the detectable but not decisive impact of demographic factors on immune repertoire data (**Supplementary Figure 9**). Other IGHV genes flagged by the model are not stratified by ancestry (**Supplementary Figure 9**). As expected from genetic variation in the alleles of HLA proteins that restrict TCR binding, some TRBV genes were also stratified by ancestry (**Supplementary Figure 9**). TRBV10-2, TRBV24-1, and TRBV25-1, all gene segments enriched in African healthy controls, were the top three highly ranked TRBV gene groups for classifying our predominantly African HIV cohort (**Supplementary Figure 10b**).

The sequence model's rankings also favored certain CDR3 lengths, one of the major features in immunoglobulin and TCR gene rearrangements affected by selection, despite no direct input of sequence length into the model. Shorter IgH CDR3 segments were favored for the chronic diseases SLE and HIV (**Figure 3b-c**), consistent with reported selection patterns in HIV[14], but longer CDR3s were favored for Covid-19 prediction (**Figure 3a**). These prioritized sequences could reflect clones recently derived from naive B cells that have not yet undergone extensive selection that would favor shorter CDR3 lengths in antigen-experienced B cells. TCR rankings follow the same pattern, except for SLE, where longer CDR3 sequences are favored (**Supplementary Figure 10c**).

B cell isotype usage varied by person and across disease cohorts (**Supplementary Figure 11**). To prevent isotype sampling artifacts from driving disease predictions, we designed the sequence model to apply balanced weights to all major isotypes (without separate weighting of subisotypes of IgG and IgA). As a result, all isotypes were included among model-prioritized sequences for prediction of each disease (**Supplementary Figure 12**). For Covid-19 prediction, IgG sequences played a slightly bigger role than other isotypes, as expected by the prominence of IgG-expression in antigen-specific B cells in this infectious disease[40,66–68]. The other models used in the *Mal-ID* ensemble were also designed not to be influenced by isotype sampling variation. 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 0.70 AUC, compared to *Mal-ID*'s 0.98 AUC disease classification performance.

Language model identifies SARS-CoV-2 binders

Only a small minority of peripheral blood B and T cell receptor sequences from Covid-19 patients are directly related to the antigen-specific immune response to SARS-CoV-2. Other naive and memory T and B cells continue to circulate even during acute illness[69,70]. The 0.98 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 applied the language model component of *Mal-ID* to IgH and TRB sequences downloaded from public databases of SARS-CoV-2 specific receptors[71,72] 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[73]. We calculated a Covid-19 class probability for each known binder sequence and for sequences from held-out healthy donors in our dataset.

The prediction model assigned significantly higher ranks to known-binder sequences compared to healthy donor sequences (**Figure 4**). When viewed as how well we discover known binders with *Mal-ID* rankings, we achieve AUCs of 0.74 (IgH) and 0.59 (TRB). 80% and 63% of known binders scored in the top half of ranked IgH and TRB sequences, respectively, while 53% and 29% of known IgH and TRB binders were in the top 20% of ranks. These binding relationships were not known to the classifier at training time, and the binding sequence databases were not used to train the model. The high ranking of 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.

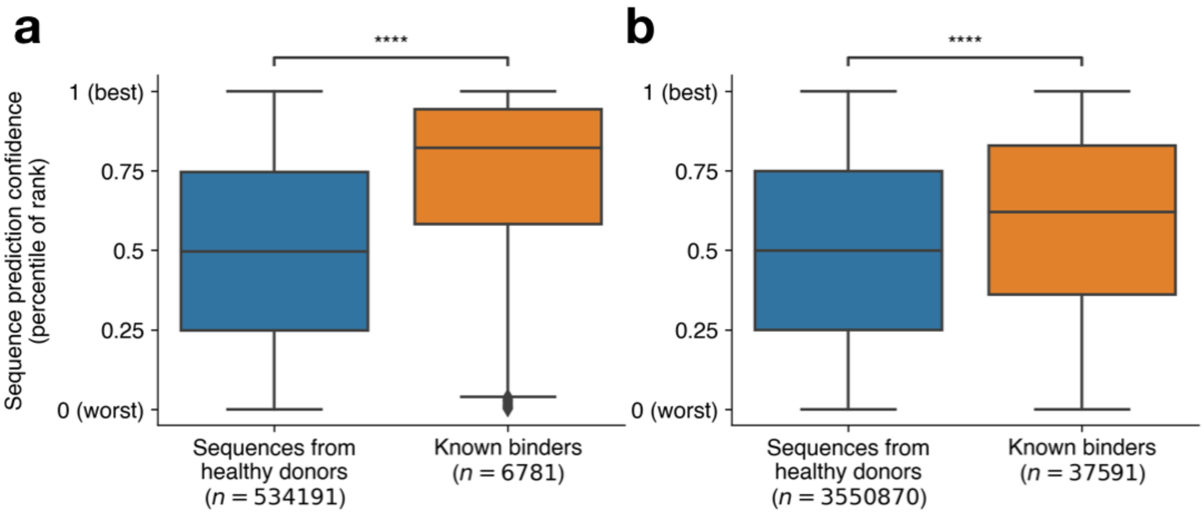


Figure 4: Sequences validated to be specific for SARS-CoV-2 (orange) were ranked significantly higher than healthy donor sequences by *Mal-ID*'s language embedding classifier model (one-sided Wilcoxon rank-sum test). One cross-validation fold shown. **a**, IgH sequences: U-statistic = 2.7e9, p ~ 0; **b**, TRB sequences: U-statistic = 7.8e10, p ~ 0.

We compared these known-binder discovery results to an alternative strategy of calculating the distance from each known binding sequence to the nearest Covid-19 associated cluster identified by the CDR3 clustering model. Ranking sequences by distance to Covid-19 predictive CDR3 clusters does not enable discovery of known binders: the resulting AUCs were 0.54 (IgH) and 0.49 (TRB) (**Supplementary Figure 13**). Only 16.5% of IgH and 3.9% of TRB sequences scored 0.5 or higher on the rank scale ranging from 0 (worst) to 1 (best). The vast majority of sequences had infinite distance (i.e. a rank of 0) from any Covid-19 associated cluster: there were no selected clusters with the same clonal lineage parameters (V gene, J gene, and CDR3 length). This result suggests that the *Mal-ID* language model approach is better suited for discovery of known binders than the CDR3 clustering strategy.

To further study the sequences most highly ranked by *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 Uniform Manifold Approximation and Projection (UMAP) layout using IgH or TRB receptors that the *Mal-ID* language model classifier learned to confidently separate into distinct groups by immune state (**Figure 5a**). Since these supervised UMAPs are 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. Then we overlaid patient repertoires that were held out from the training set onto the reference UMAP visualization. Covid-19, HIV, and SLE patient repertoires all contained sequences that were predicted to be highly associated with the disease in question and that were projected onto the disease-specific regions of the IgH or TRB reference map (**Figure 5b-d**).

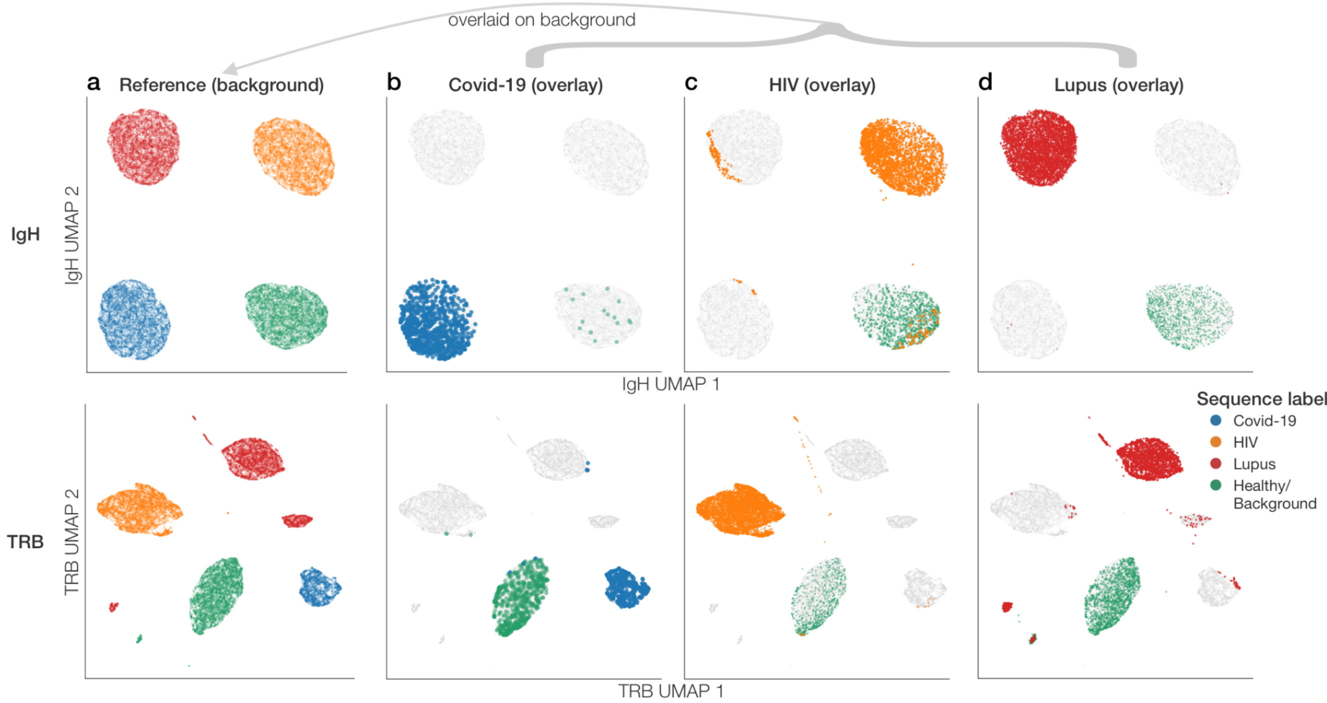


Figure 5: Individual IgH (top row) and TRB (bottom row) receptor sequences can be visualized by predicted disease association to interpret the disease status of a person's collective immune repertoire. **a**, Reference UMAPs were created from the most disease-associated B and T cell receptors from many patients. Each point is a sequence, colored by the predicted identity of that sequence. Receptors are arranged in disease-specific clusters. **b-d**, IgH and TRB repertoires from a Covid-19 patient (**b**), an HIV patient (**c**), and an SLE patient (**d**) were projected onto the reference maps. The foreground points are again colored by predicted disease specificity of each overlaid sequence, with the reference maps shown in the background in gray.

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 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 461 individuals with four distinct immunological states, classifying immune responses of study participants with performance of 0.98 AUC by leveraging both B and T cell signals in the 410 individuals from whom both data types were available. The evaluation strategy ensured that there was never a situation where a model was trained on data from a patient and then evaluated on other data from the same person. 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 and an autoimmune disorder. These signatures of disease and specific pathogens overrode more modest differences detectable between individuals differing by sex, age, or ancestry. *Mal-ID* additionally generalizes to data from other laboratories and experimental protocols.

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 trio of methods to extract signal from B and T cell repertoires, fusing aggregate repertoire composition properties with detection of important sequence groups and with a language model interpretation of individual sequences. Integrating these models outperforms them individually and suggests that they capture different patterns. We also observed that combining data from BCR and TCR repertoires provides more accurate classification than either receptor type alone, potentially reflecting variation in the roles of B cell and T cell responses in different diseases. The disease predictor is not a black box; we can trace the decisions in the language model component back to the original sequences by ranking sequences according to predicted disease association. This language model classifier ranking allows discovery of more sequences independently known to be disease-associated than can be discovered with other approaches like CDR3 clustering. We also visualized immune repertoires in disease, highlighting the potential of monitoring for disease-associated receptor sequences. Notably, labels on individual sequences are not required to train these models.

The model assemblage we developed in *Mal-ID* for B and T cell receptor sequences could be applied to tasks beyond identifying disease exposures. Our initial goal has been to classify current acute or chronic diseases, to ensure relevance for the care of individual patients, but this approach should also be amenable for other purposes, such as detecting evidence of more distant prior exposures to pathogens or other immune stimuli in memory cell immune cell receptor repertoires. Conventional serology tests may only be positive for recent infections or vaccinations, as a result of antibody levels waning after exposure. Memory B and T cells can be long-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[74,75]. 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 four immune states and cohorts of only several hundred individuals. The *Mal-ID* framework appears to capture fundamental principles of immune responses, and it appears to generalize to separate clinical cohorts, but additional testing will be needed to further assess its generalizability to many other immunological states. Model predictions are affected by different experimental protocols and sequencing platforms, which is to be expected given the prior literature on systematic variation across platforms in V gene use measurements, which are a part of the *Mal-ID* classification scheme. We believe the *Mal-ID* repertoire composition classifier can be extended to additional repertoire sequencing technologies by training on more disparate datasets, or by down-weighting the importance of the repertoire composition features in the disease prediction metamodel, to rely less on precise V gene usage patterns. The biological validation against known SARS-CoV-2 binders also revealed some limits to the model’s grasp of the ultra-high-dimensional receptor repertoire space. For example, 20% of IgH and 37% of TRB sequences from the external databases had language-model-assigned ranks under 50% (**Figure 4**). Since our specificity group detection approach is anchored in the concept of common response patterns shared across individuals, it is less likely to be able to interpret truly idiosyncratic immune receptors unique to a single individual. The enrichment for higher ranks among TRB known binders may be lower than for IgH because the interactions between TCR and genetically diverse HLA molecules during T cell stimulation introduce additional potential differences between cohorts, and activation of T cells upon peptide stimulation in culture may result in some non-specific bystander clone activation. Further, unlike the IgH classification, the

TCR analyses do not exclude naive T cells that could contain low frequencies of SARS-CoV-2 specific clones in unexposed individuals.

Extended to further datasets and clinical cohorts at population scale, this immune repertoire analysis strategy offers a strategy for disease definition refinement and diagnosis, as well as improving understanding of immune response features such as autoreactivity that are 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.

Methods

B and T cell repertoire sequencing

We assembled immune receptor repertoires from 63 Covid-19, 95 chronic HIV-1, and 86 Systemic Lupus Erythematosus (SLE) patients, along with 217 healthy controls. Among Covid-19 patients, we excluded mild cases, samples prior to seroconversion, and patients known to be immunosuppressed. These filters limited model training data to peak-disease samples to improve our chances of learning patterns for the disease-specific minority of receptor sequences. However, we wanted to avoid creating an artificially simple classification problem from filtering to trivially separable immune states. To this end, 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[14].

Across these diverse immune states, over 14.3 million B and 19.2 million T cell receptor clones were sampled, PCR amplified with immunoglobulin and T cell receptor gene primers, and sequenced as previously described[14,49]. Briefly, we amplified T cell receptor beta chains and each immunoglobulin heavy chain isotype in separate PCR reactions 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[76]. 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 heavy chain. On the other hand, T cell receptors are known not to exhibit somatic hypermutation and to have CDR1β and CDR2β regions that match the germline sequence. Accordingly, we used TCR CDR1β and CDR2β annotations from IMGT reference TRBV gene germline sequences. We also restricted our dataset to CDR-H3 and CDR3β segments with eight or more amino acids; otherwise the edit distance clustering method below might group short but unrelated sequences.

We grouped nearly identical sequences within the same person into clones. To do so, for each individual, we grouped all nucleotide sequences from all samples (including samples at different timepoints) across all isotypes, and ran single-linkage hierarchical clustering to infer clonal lineages, iteratively merging sequence clusters from the same individual with matching IGHV/TRBV genes, IGHJ/TRBJ genes, and CDR-H3/CDR3β lengths, and with any cross-cluster pairs having at least 95% CDR3β sequence identity by string substitution distance, or at least 90% CDR-H3 identity, which allows for BCR somatic hypermutation[14].

Among BCR sequences, 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. Similarly, we kept one copy of each TCRβ clone. Any samples with fewer than 100 IgG, 100 IgA, and 500 IgD/M clones, or with fewer than 500 TRB clones, were rejected. On average, any two patients had 0.0004% IgH and 0.169% TRB sequence overlap, underscoring the enormous diversity of T cell receptor and especially B cell receptor sequences.

Cross-validation

We divided individuals into three stratified cross-validation folds, each split into a training set and a test set (**Supplementary Figure 1**). Each individual was assigned to one test set. The splits were respected across the training of the complete *Mal-ID* pipeline. 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. For example, while the repertoire classification, CDR3 clustering, and language model base classifiers are trained on the training set, the ensemble model is trained on the validation set, and then evaluated on the test set (**Supplementary Figure 3**). This happens separately for each fold; in other words, one collection of models is trained using fold 1's training, validation, and test sets, then a separate set of models is trained using fold 2's training, validation, and test sets, and so on. On average in any fold, we observed 0.05% of IgH and 4.8% of TRB 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 been overfitted to the particularities of training patients. For the minority of individuals with multiple samples, we accordingly made sure that, in each cross-validation fold, all samples from the same person were grouped together into one of the training, validation, or test sets, as opposed to being spread across multiple sets.

Finally, for the purpose of external cohort validation, we repeated the model training procedures with a "global" fold designed to incorporate all the data, by having only a training set and a validation set but no test set (**Supplementary Figure 1**). Repertoires from these independent studies are used in place of the test set at evaluation time.

Evaluation metrics

Models were trained with the scikit-learn implementations of logistic regression (with multinomial loss and default regularization strength hyperparameter $\lambda = 1/n$, where n is the number of training sequences), random forests (with 100 trees), and support vector machines (in "each class versus the rest" mode, with linear kernel and default regularization strength hyperparameter C=1.0), all with prevalence-balanced class weights. Predicted labels from all test sets were concatenated for global accuracy evaluation. Performance metrics that take predicted class probabilities as input, including ROC AUC and auPRC, were computed separately for each fold, because probabilities may be on different scales in each fold and should not be combined for a global AUC or auPRC score. We report multiclass AUC and auPRC calculated in a one-versus-one fashion, taking the class size-weighted average of the binary AUCs/auPRCs calculated for each pair of classes, allowing each class a turn to be the positive class in the pair. All analyses were performed and plotted with software versions *python v3.9.13, numpy v1.22.0, pandas v1.4.3, scipy v1.8.1, scikit-learn v1.1.1, python-glmnet v2.2.1, jax v0.3.14, umap-learn v0.5.3, matplotlib v3.5.2*, and *seaborn v0.11.2*.

Model 1: Disease classifier using overall BCR or TCR repertoire composition features

For each sample, we created IgG, IgA, IgM/D, and TRB summary feature vectors by tallying IGHV/TRBV gene and IGHJ/TRBJ gene usage, counting each clone once. We ranked IGHV or TRBV genes by training set prevalence and excluded the bottom half, to avoid overfitting to minute differences in rare V gene proportions between cohorts. To account for different total clone counts across samples, we normalized total counts to sum to one per sample. Then we log-transformed and Z-scored (i.e. subtracted the mean and divided by the standard deviation, to achieve zero mean and unit variance) 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. In addition, for each sample's subset of BCR sequences belonging to each isotype, we calculated the median sequence somatic hypermutation rate and the proportion of sequences that are somatically hypermutated (with at least 1% SHM). Only BCRs have somatic hypermutation, so we did not include mutation rate features of TCRs. In total, we arrived at 51 features across IgG, IgA, and IgM/D (fifteen count matrix principal components and two mutation rate features per isotype) for the IgH repertoire composition model, and 15 features for the TRB repertoire composition model.

We fit separate lasso logistic regression linear models with L1 regularization on the 51-dimensional (17 x 3 isotypes) BCR and 15-dimensional TCR feature vectors from each sample to predict disease. Features were standardized to zero mean and unit variance. We repeated this feature engineering and model training procedure on each cross-validation fold separately.

Model 2: Disease classifier by clustering CDR-H3 sequences with edit distance

We performed single-linkage clustering on CDR3β sequences from T cells with identical TRBV genes, TRBJ genes, and CDR3β lengths, and separately on CDR-H3 sequences from B cells with identical IGHV genes, IGHJ genes, and CDR-H3 lengths, as described previously[14]. Nearest-neighbor clusters were iteratively merged if any cross-cluster pairs had high sequence identity: at least 90% for CDR3β or 85% for CDR-H3, allowing for somatic hypermutation in B cells, as measured by string substitution distance (normalized Hamming distance).

Filter to BCR and TCR disease-specific enriched clusters: For each sequence cluster found in a cross-validation fold's training set, we performed a Fisher's exact test using a two-by-two contingency table denoting how many unique people have a particular disease and have some receptor sequences fall into the cluster. In other words, each cluster's *p*-value from the Fisher's exact test denotes the cluster's enrichment for a particular disease. This approach is consistent with prior work that selects a set of disease-specific enriched sequences, then counts exact matches to this sequence set in new samples[13]. Given a *p*-value threshold, the full list of training set clusters was filtered to clusters specific for each disease type. We performed all the following featurization and model fitting steps for *p*-values ranging from 0.0005 to 0.05, then selected the *p*-value that led to the highest validation set performance as measured by the Matthews correlation coefficient (MCC) score, a classification performance metric that is well-suited to imbalanced datasets[77]. The final chosen *p*-values differed depending on the cross-validation fold and the receptor type (i.e. BCR or TCR).

Compute BCR and TCR cluster membership feature vectors for each sample: For each selected enriched 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 or TCR sequences originally part of each clone. Sequences from a sample were then matched to these predictive cluster centroids. In order to be assigned, a sequence must have the same IGHV/TRBV gene, IGHJ/TRBJ gene, and CDR-H3/CDR3β length as the candidate cluster, and must have at least 85% (BCR) or 90% (TCR) 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 sample. Cluster membership counts were arranged as a feature vector for each sample: a sample's count for a particular disease was defined as the number of disease-enriched clusters into which some sequences from the sample were matched. This featurization captures the presence or absence of convergent T cell receptor or immunoglobulin sequences (separated by locus, but without regard for IgH isotypes).

Fit and evaluate model for each locus: Features were standardized, then used to fit separate BCR and TCR linear logistic regression models with L1 regularization and balanced class weights (inversely proportional to input class frequencies). The featurizations and models were fitted on each training set and applied to the corresponding test set.

We abstained from prediction if a sample had no sequences fall into a predictive cluster; this indicated no evidence was found for any particular class. Abstentions hurt accuracy and MCC scores, but were not included in the AUC calculation, since no predicted class probabilities are available for abstained samples. Fewer than 2.5% of samples resulted in abstention (**Supplementary Table 3**).

Language model representations for immune receptors

We combined the CDR-H1/CDR1β, CDR-H2/CDR2β, and CDR-H3/CDR3β 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[78]. A final 1900-dimensional vector representation was calculated by averaging UniRep's hidden state over the original protein's length dimension[44].

To embed sequences, we used weights fine-tuned on a subset of each cross-validation fold's training set, yielding a total of six fine-tuned models: one per fold and gene locus. We chose the weights that minimized cross-entropy loss on a subset of the held-out validation set. For example, we fine-tuned UniRep on fold 1's BCR training set until reaching minimal cross-entropy loss on fold 1's BCR 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 or T cell receptor patterns, but not hyper-specialized to the disease classification problem. They can be applied to other immune sequence prediction tasks.

Model 3: 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.

Train sequence-level disease classifier: First, we trained lasso classification models to map sequences to disease labels — one model per fold and per locus. As input data, we used fine-tuned UniRep embeddings (standardized to zero mean and unit variance), along with categorical dummy variables representing the IGHV gene and isotype of each BCR sequence or the TRBV gene of each TCR 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. The lasso's L1 penalty encouraged sparsity among the ~2000 input features[79]. Because isotype use varies from person to person, we trained the sequence-level BCR model with isotype weights to account for this imbalance.

Aggregate sequence predictions to sample prediction: Since we have no true sequence labels, we cannot evaluate classification performance for the sequence-level classifier. Instead, we aggregated BCR or TCR sequence predictions into a patient sample-level prediction, by the following procedure.

Given a sample with *n* sequences, each of which has *k* predicted probabilities (one predicted probability for each of the *k* disease classes), in the form of a *[n x k]* matrix:

For each class among the *k* classes:

1. Sort the *n* sequence-level predicted probabilities in ascending order. This represents a list of each sequence's predicted probability of belonging to the disease class in question.

2. Trim the top and bottom 10% of sequence probabilities. This means that we will remove the 10% of sequence-level probabilities that have the lowest predicted probability and the 10% of sequence-level probabilities that have the highest predicted probability.

3. Calculate a weighted mean of the trimmed probabilities. In other words, we calculate the average probability of the remaining sequence-level probabilities, where the weight of each probability is inversely proportional to its isotype prevalence. (This way, minority isotype signal is not drowned out.)

Re-normalize the trimmed weighted mean probabilities to sum to 1. This means that we will divide each probability by the sum of all probabilities, so that the probabilities add up to 1.

This procedure gives the final *k*-dimensional predicted disease class probabilities vector for the sample. The vector contains the predicted probability of each disease class for the given sample.

Evaluate classifier: Finally, we evaluated the sequence-prediction-aggregating predictor on the test set. Each test sample's sequences were scored then combined with a trimmed mean to arrive at final predicted sample labels. Ground truth sample disease status is known, so we could evaluate classification performance here, unlike at the sequence-level prediction stage.

Ensemble metamodel

After training repertoire composition, CDR3 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 sample abstentions from the CDR3 clustering model (the other models do not abstain). Finally, we trained a random forest classification metamodel to map the combined predicted probability vectors to validation set sample disease labels. We evaluated this metamodel on the held-out test set. To evaluate feature contributions to predictions of each immune state class, we also fit an alternate elastic net logistic regression metamodel with the same input features. To arrive at a meaningful set of coefficients from the elastic net regularization's coefficient shrinkage and feature selection[80], we tuned the regularization strength hyperparameter with internal cross-validation using the *glmnet* library, again with multinomial loss and balanced class weights. This internal cross-validation also respected participant labels in the splits, as in the overall cross-validation design above. All variants of the ensemble metamodel perform similarly (**Supplementary Figure 4**).

Batch effect evaluation

Having integrated many datasets in this study, we sought to test whether our disease classification performance was driven by technical differences between batches of library preparation or sequencing instrument run. It would be expected in any study of human cohorts to identify some batch effects, given the difficulty of collecting identical samples in identical manner, at identical severity and timepoints, from patients suffering from diseases that appear in different populations at different frequencies. Notably, the IgH data collected for individual participants in this study were typically based on multiple Illumina MiSeq sequencer runs, and were combined prior to analysis. Many of our sequencing run batches included only one disease type, but batches that included both diseased and healthy controls from the same population permitted accurate classification of the disease or healthy state, for example, with classification of HIV-infected patients and healthy controls that were sequenced together in the same batch, or SLE patients and healthy controls sequenced in the same batch.

Acknowledging that there were biological differences between many sequencing batches that were enriched for a particular disease state, and that several sequencer runs were performed for some sample sets, we evaluated the potential impact of these batch differences using the language model embeddings of BCR and TCR repertoires from the disease types found in multiple batches: Covid-19 patients, SLE patients, and healthy donors. We applied the kBET batch effect metric from the single cell sequencing literature[81,82]. kBET measures whether cells from many batches are well-mixed by comparing the batch label distribution among each cell's neighbors to the global distribution. In place of cells described by gene expression vectors, we have sequences described by language model embedding features. We measured kBET for every disease in every test set fold and in both BCR and TCR data. For example, we constructed a k-nearest neighbors graph (k = 50) with all BCR sequences from Covid-19 patients in test fold 1. We performed chi-squared tests for the difference between the batch label distribution among each sequence's 50 nearest neighbors and the expected distribution from the total number of sequences belonging to each batch in the entire graph. After multiple hypothesis correction with a significance threshold of p=0.05, we measured the number of sequences for which we could reject the null hypothesis that the local neighborhood batch distribution is the same as the global batch distribution. Aggregating these results by disease across gene loci and folds, we see that the null hypothesis is rejected for only 17.1% of sequences on average, suggesting that the sequence data in the graph are well mixed according to batch (**Supplementary Table 7**). The average rejection rate is higher for Covid-19 BCR sequences at 34.1%, which may be influenced by disease severity differences between cohorts (**Supplementary Table 1**). Time point differences between batches may also have an effect on kBET metrics for acute diseases like Covid-19. At earlier time points, Covid-19 patient repertoires may include more healthy background sequences, leading to a different batch overlap graph in comparison to how batches compare after clonal expansion of Covid-19 responding sequences. Overall, these results suggest that most sequences have well-mixed batch proportions amongst their nearest neighbors.

Validation on external cohorts

The best test of whether our model has learned true biological signal as opposed to batch effects is whether our model generalizes to unseen data from other cohorts. For the purposes of evaluating external cohorts, rather than using models trained on our cross-validation divisions of the data, we trained a set of “global” models incorporating all *Mal-ID* data without holding out a test set (**Supplementary Figure 1**). This included training “global” fine-tuned BCR and TCR language models. To train the ensemble metamodel, we still held out a validation set, with a ratio of training set to validation set size equivalent to the ratio used in the cross-validation regime.

We downloaded data from other BCR and TCR Covid-19 patient and healthy donor repertoire studies with cDNA sequencing[51,52,83–86]. For the acute Covid-19 cases, we selected peak timepoint samples at least two weeks after symptom onset, after which time we would expect seroconversion[40]. We reprocessed sequences through the same version of IgBLAST and IgBLAST reference data as used for the primary *Mal-ID* cohorts, to ensure consistent gene nomenclature. (This was not possible for the Britanova et al. datasets[51,52] because the raw sequences were unavailable, so we used their gene calls and confirmed the naming was consistent with our training data, especially for indistinguishable TRBV genes TRBV6-2/6-3 and TRBV12-3/12-4.) As with the core *Mal-ID* cohort, we filled in TCR CDR1β and CDR2β sequences using TRBV reference sequences downloaded from IMGT. We embedded productive CDR1+2+3 sequences with the global fine-tuned language models, then processed the downloaded repertoires through the entire *Mal-ID* model architecture.

For comparison, we repeated this analysis by downloading Covid-19 patient and healthy donor TCR repertoire data collected with the Adaptive Biotechnologies sequencing protocol[13,72], which we reprocessed with the same IgBLAST version as above, for consistency. We filtered to acute Covid-19 cases sampled between 11 and 21 days after symptom onset with no recorded immunosuppression, cancer, autoimmune disease, or other comorbidities. The number of healthy control repertoires was very large, so we sampled the same number of healthy samples as the total number of selected Covid-19 samples.

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. Only one healthy control individual was over 80 years old, therefore our data do not assess repertoire changes at more extreme older ages. We excluded the healthy individual over 80 years old from the analysis.

For each of the three demographic prediction tasks, we trained the full BCR+TCR *Mal-ID* architecture on all cross-validation folds. We note that we did not explicitly introduce data from allelic variant typing in germline IGHV, IGHD, or IGHJ gene segments or in HLA genes 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 80 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 sample-level, not sequence-level model, and age/sex/ancestry demographic information is tied to samples 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. Finally, we trained metamodels with both demographic features and sequence features, along with interaction terms between the demographic and sequence features to allow for interaction effects. Comparing the performance of these models to the demographics-only models shows the added value of adding sequence information.

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/CDR3β length, isotype, and IGHV/TRBV gene segment. For the V gene usage comparison (**Figure 3**), V genes with very low prevalence were removed. To set a prevalence threshold, we found the greatest proportion each V gene ever comprises of any cohort, and took the median of these proportions (**Supplementary Figure 14**). The following rare IGHV and TRBV genes were filtered out (half of the totals): IGHV1-45, IGHV1-68, IGHV1-69D, IGHV1-f, IGHV1/OR15-1, IGHV1/OR15-2, IGHV1/OR15-3, IGHV1/OR15-4, IGHV2-10, IGHV2-26, IGHV2-70D, IGHV3-16, IGHV3-19, IGHV3-22, IGHV3-35, IGHV3-38, IGHV3-43D, IGHV3-47, IGHV3-52, IGHV3-64D, IGHV3-71, IGHV3-72, IGHV3-73, IGHV3-NL1, IGHV3-d, IGHV3-h, IGHV3/OR16-10, IGHV3/OR16-13, IGHV3/OR16-8, IGHV3/OR16-9, IGHV4-28, IGHV4-55, IGHV4/OR15-8, IGHV5-78, IGHV7-81, VH1-17P, VH1-67P, VH3-41P, VH3-60P, VH3-65P, VH7-27P; TRBV10-1, TRBV11-1, TRBV11-3, TRBV12-2, TRBV12-5, TRBV13, TRBV15, TRBV16, TRBV17, TRBV20/OR9-2, TRBV26, TRBV27, TRBV29/OR9-2, TRBV3-1, TRBV3-2, TRBV4-1, TRBV4-2, TRBV4-3, TRBV5-3, TRBV5-7, TRBV5-8, TRBV6-4, TRBV6-7, TRBV6-8, TRBV6-9, TRBV7-1, TRBV7-4, TRBV7-7. Most V genes remaining after this filter had consistent, balanced prevalence across cohorts (**Supplementary Figure 15**).

Model ranking of known SARS-CoV-2 binder sequences

We downloaded the July 26, 2022 version of CoV-AbDab[71], and reprocessed these B cell receptor heavy chain sequences through the same version of IgBLAST as used for our primary cohorts to ensure consistent V gene nomenclature. We filtered to antibody sequences known to bind to SARS-CoV-2 (including weak binders), and selected sequences from human patients or human antibody libraries. 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, using single linkage clustering as in the pipeline for our primary cohorts. As a result, several related sequences were combined and replaced by a consensus sequence.

Similarly, we downloaded the ImmuneCode MIRA database[72], version 002.1, and reprocessed these T cell receptor beta chain sequences with our pipeline's standard IgBLAST version for consistent V gene nomenclature. By the same logic as above, we filtered to productive sequences from patients with acute Covid-19, and also to only the TRBV genes present in our dataset, as any others would not be compatible with the sequence model, which uses V gene segment identity as a feature. Among the remaining SARS-CoV-2 associated sequences, we deduplicated those with identical TRBV genes, TRBJ genes, and CDR3β sequences.

We calculated the probability that each sequence was associated with the Covid-19 class, using a single cross-validation fold's sequence model (since probabilities may not necessarily be comparable across folds). Since isotype designations were not available in the CoV-AbDab dataset, we scored each CoV-AbDab sequence with all possible isotype settings and kept the version with highest predicted Covid-19 probability, in order to assess the strength of a sequence's relationship to the disease. Then we scored healthy donor sequences from the held-out test set of the same cross-validation fold, ensuring that they were not used to train the model. The Covid-19 class probabilities were converted to ranks, and then we calculated AUC scores using model rankings versus which BCR or TCR sequences matched the external databases.

We also generated sequence rankings with the CDR3 clustering model from the same cross-validation fold for comparison. For each known binder and healthy donor CDR3 sequence, we computed the Hamming distance to its nearest Covid-19 associated cluster centroid with the same V gene, J gene, and CDR3 length (because the model only forms clusters among sequences with these matching clonal parameters). These distances were ranked, assigning highest rank (1.0) to the shortest distance, consistent with the previous analysis. However, many query sequences were infinitely far from any Covid-19 associated cluster centroid. That is, when selecting a list of clusters predictive of the Covid-19 class, the CDR3 clustering model did not choose any clusters with the same V genes, J genes, and CDR3 lengths as these query sequences. Accordingly, they were assigned the worst rank (0.0), indicating these sequences displayed no evidence of disease association according to the clustering model. We computed AUCs of rank versus known binder identity as in the prior analysis.

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 (i.e. per-class) logits for each sequence. We provided sequence labels as supervision to the UMAP so they are less likely to be distorted in the layout[87].

We created a reference UMAP for each fold and each locus 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 that results from imputing sequence labels from patient disease status (as described earlier). Therefore, we filtered potential sequences to those with predicted disease class probability at least 0.05 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 30% to create a succinct pool of reference sequences for each class. We subselected 10,000 of these selected sequences for each class, to arrive at a uniform number of “reference” (i.e. very class associated) sequences for each class (i.e. for each disease and for the healthy 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. For a given held-out test patient, we computed supervised embeddings (per-class logits) for each sequence using the sequence-level lasso model, and applied the trained UMAP transformation to produce 2D coordinates, using 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 sample 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 sorted the remaining sequences by their predicted class probability, and kept the top 30% of the sorted list across Healthy/Background and the overall sample predicted label class.

Data availability

Data will be deposited online.

Code availability

Code will be deposited online.

Ethics statement

The use of data was approved by Stanford University IRBs #13952, #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., T.D.P., E.S.P., C.C.R., M.A.M, B.F.H., J.D.G., J.R.H., I.B., P.J.U., K.C.N., B.A.P., C.A.B., J.T.M., J.M.G., J.A.J., and S.Y. provided blood samples, as well as clinical and demographic data annotation and analysis. J.Y.L., K.D.N., and R.A.H. prepared and sequenced samples. K.M.R. designed the data warehouse. M.E.Z., E.C., J.K.M., and R.T. performed the computational analysis. M.E.Z., R.T., A.K., and S.D.B. wrote the manuscript with input from all authors.

Funding

S.D.B. was partially supported by NIH/NIAID grants R01AI130398, R01AI127877, U19AI057229, U54CA260518, U19AI167903 and a philanthropic gift from an anonymous donor. M.E.Z. was supported by the National Science Foundation Graduate Research Fellowship and the Stanford Bio-X Bowes Graduate Student 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. C.A.B. was supported by the Burroughs Wellcome Fund Investigators in the Pathogenesis of Infectious Diseases 1016687 and U19 AI057229. S.R.M., W.D., J.M.G., J.T.M. and J.A.J. were partially supported by NIH/NIAMS AR073750 and NIH/NIAID UM1AI144292.

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: Elizabeth J. Zudock, Marjan M. Hashemi, Kristel C. Tjandra, Jennifer A. Newberry, James V. Quinn, 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, was 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, Gilead, GSK, and Karius, 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). R.T is a consultant for Genentech. J.A.J. has served as a consultant for AbbVie, Janssen, Novartis, and GlaxoSmithKline. J.A.J. also has unrelated patents through the Oklahoma Medical Research Foundation which the foundation has licensed to Progentec Biosciences, LLC. J.T.M has served as a consultant for AbbVie, Alexion, Alumis, Amgen, AstraZeneca, Aurinia, Bristol Myers Squibb, EMD Serono, Genentech, Gilead, GlaxoSmithKline, Lilly, Merck, Pfizer, Provention, Remegen, Sanofi, UCB, and Zenas, and reports research support from AstraZeneca, Bristol Myers Squibb, and GlaxoSmithKline (outside of this study). Other co-authors declare that they have no competing interests.

Supplementary Information

Immune state	Cohort	Sample type	Patient and clone counts	Demographics
Acute Covid-19	Hospital inpatients, ranging from 7 to 37 days after symptom onset	Whole blood RNA (Paxgene tubes)	48 patients (31% in ICU) 48 samples 403562 IgH clones 654000 TRB clones	58% Hispanic/Latino, 17% Asian, 17% Caucasian, 2% African, 6% unknown Median age 44.5 years old; range 21 to 86 52% female
	Hospital inpatients, CoV2+ IgG seroconverted, ranging from 9 to 35 days after symptom onset[40]	PBMC RNA	10 patients (70% in ICU) 10 samples 256655 IgH clones 193568 TRB clones	Ethnicities unknown Median age 65 years old; range 36 to 88 60% female
	Hospital inpatients, ranging from 8 to 37 days after symptom onset (BCR only)	PBMC RNA	5 patients 5 samples 276076 IgH clones	40% Caucasian, 20% African, 20% Asian, 20% unknown Median age 57 years old; range 26 to 73 40% female
Lupus	Adult lupus (BCR only)	PBMC RNA	23 patients (69% have multiple autoantibodies; 22% nephritis, 35% no nephritis, 43% unknown nephritis status) 34 samples 520355 IgH clones	52% Caucasian, 39% African, 9% unknown Median age 36 years old; range 21 to 71 (with two unknown) 95% female (not counting 2 patients of unknown sex)
	Pediatric lupus, untreated	Whole blood RNA (Tempus tubes)	43 patients (53% have nephritis) 43 samples 2256194 IgH clones 2362725 TRB clones	35% Asian, 28% Caucasian, 28% Hispanic/Latino, 7% African, 2% unknown Median age 13 years old; range 7 to 18 74% female
	Adult lupus	PBMC RNA	15 patients 16 samples 296828 IgH clones 520543 TRB clones	80% Caucasian, 7% African, 7% Asian, 7% Hispanic/Latino Median age 42 years old; range 21 to 68 93% female
	Adult lupus	Whole blood RNA (Paxgene tubes)	5 patients 5 samples 286755 IgH clones 740123 TRB clones	60% African, 20% Asian, 20% Caucasian Median age 46 years old; range 34 to 51 100% female

Immune state	**Cohort**	**Sample type**	**Patient and clone counts**	**Demographics**
HIV-1	Primary cohort[14]	PBMC RNA	95 patients (47% make broadly neutralizing Abs) 98 samples 2762764 IgH clones 3164681 TRB clones	89% African, 11% unknown Median age 31 years old; range 19 to 64 64% female
Healthy donors	Primary adult cohort[88]	PBMC RNA	102 healthy donors 102 samples 4740876 IgH clones 5803482 TRB clones	70% Caucasian, 24% Asian, 5% Hispanic/Latino, 1% African, 1% unknown Median age 51.5 years old; range 17 to 81 43% female
	HIV negative[14]	PBMC RNA	43 healthy donors 43 samples 832374 IgH clones 1472515 TRB clones	65% African, 35% unknown Median age 27 years old; range 20 to 51 51% female
	Lupus negative (BCR only)	PBMC RNA	23 healthy donors 27 samples 365431 IgH clones	52% African, 43% Caucasian, 4% unknown Median age 42.5 years old; range 24 to 70 (with one unknown) 86% female (not counting 1 individual of unknown sex)
	Lupus negative	PBMC RNA	4 healthy donors 4 samples 125576 IgH clones 107635 TRB clones	All Caucasian Median age 49 years old; range 33 to 67 75% female
	Lupus negative	Whole blood RNA (Paxgene tubes)	2 healthy donors 2 samples 117351 IgH clones 377830 TRB clones	50% Caucasian, 50% African Median age 47.5 years old; range 47 to 48 0% female
	Pediatric control cohort	PBMC RNA	43 healthy donors 43 samples 1134937 IgH clones 3834725 TRB clones	51% Caucasian, 19% Asian, 2% Hispanic/Latino, 28% unknown Median age 13 years old; range 8 to 18 49% female

Supplementary Table 1: Cohort and batch info for 461 individuals with a total of 480 samples. 414 of the 480 samples had both BCR and TCR sequencing performed, representing 410 of the total 461 individuals. The remainder only underwent BCR IgH sequencing.

Strategy applied to predicted class probability vectors for all sequences in a sample	**BCR ROC AUC**	**TCR ROC AUC**
Trimmed mean from top and bottom (2.5%, 5%, 10% trimming)	0.842 +/- 0.015	0.885 +/- 0.015
Trimmed mean from bottom only (2.5%, 5%, 10% trimming)	0.858 +/- 0.010	0.885 +/- 0.014
Mean (untrimmed)	0.862 +/- 0.010	0.883 +/- 0.015
Weighted median	0.855 +/- 0.010	0.885 +/- 0.018
Entropy threshold (1.2, 1.3)	0.846 +/- 0.019	0.731 +/- 0.076

Supplementary Table 2: Validation set performance of various aggregation strategies of Model 3 predictions for individual sequences to predictions of an entire repertoire, showing that many approaches perform similarly.

We report average and standard deviation across three folds for the following strategies: trimming by different amounts ranging from 2.5% to 10%, trimming only from the bottom end of the probability distribution, not trimming at all (i.e. taking a standard mean), computing a weighted median (the weights being adjustments for isotype proportions, as described in **Methods**), and using entropy thresholds to exclude close call sequences (those who have roughly equal predicted probabilities for all classes) from aggregation. Note that an entropy threshold of 1.4 or higher would apply no filtering to four-class predicted probability vectors.

Strategy	**Locus**	**Accuracy**	**ROC AUC**	**auPRC**	**Abstention rate**
Global repertoire statistics (Model 1)	BCR	81.2%	0.939	0.938	0%
	TCR	76.1%	0.940	0.927	0%
CDR3 sequence clustering (Model 2)	BCR	74.4%	0.926	0.927	2.3%
	TCR	70.1%	0.885	0.879	0.2%
Language model embedding and classification (Model 3)	BCR	68.8%	0.829 (0.856 if allowed 2.3% abstention)	0.835 (0.857 if allowed 2.3% abstention)	0%
	TCR	71.0%	0.881 (0.883 if allowed 0.2% abstention)	0.857 (0.858 if allowed 0.2% abstention)	0%
Ensemble of all models (random forest)	BCR	83.1%	0.959	0.954	2.3%
	TCR	77.3%	0.947	0.939	0.2%
	BCR + TCR **(Figure 2a)**	88.6%	0.981	0.976	1.7%

Supplementary Table 3: Average cross-validated test set performance on 480 BCR samples, 414 TCR samples, or 414 BCR + TCR samples. auPRC stands for area under the precision-recall curve.

Abstentions hurt accuracy scores (they count as incorrect predictions), but are not included in the calculation of probability-based metrics ROC AUC and auPRC, because no predicted class probabilities are generated for abstained samples. For a fairer comparison of models 2 and 3, we also calculated how much model 3's ROC AUC might increase if model 3 was allowed the same number of abstentions as model 2, by post-hoc excluding the 2.3% or 0.2% worst model 3 predictions in the BCR and TCR cases, respectively.

Locus	Covid-19 cohort	Healthy donor cohort	Accuracy	ROC AUC	auPRC	Abstention rate
BCR	7 samples from Kim et al, 2021[83]	6 healthy samples from Briney et al, 2019[84]	100%	1.0	1.0	0%
TCR	17 samples from Shomuradova et al, 2020[85]	39 healthy samples from Britanova et al, 2014 and 2016[51,52]	85.7%	0.995	0.998	0%

Supplementary Table 4: External validation cohort performance using BCR-only or TCR-only random forest metamodels.

Input	Prediction target	Accuracy	ROC AUC
BCR+TCR sequence features from 165 healthy samples	Sex	47.3% (7.3% abstentions)	0.546
	Ancestry	51.5% (4.2% abstentions)	0.752
	Age (<20, 20-30, …, 70-80)	37.0% (17.6% abstentions)	0.696
	Age (under 50, 50 or older)	58.8% (13.3% abstentions)	0.748
BCR+TCR sequence features from 109 healthy samples	Age (under 18, 18 or older)	78.0% (17.4% abstentions)	0.989

Supplementary Table 5: Model performance for predicting age, sex, and ancestry of healthy individuals with known demographics, retraining the full *Mal-ID* BCR+TCR ensemble architecture for each task. To cast age as a classification problem, the continuous variable was discretized either into deciles or at a 50-year threshold. We report held-out test set performance, averaged over three cross-validation folds, from the model architecture (random forest, lasso logistic regression, or linear support vector machine) with highest ROC AUC. Abstentions hurt accuracy scores (they count as incorrect predictions), but are not included in the calculation of the probability-based AUC metric, because no predicted class probabilities are generated for abstained samples.

The pediatric vs adult age classification is reported for two cross-validation folds, not three as for the other analyses. One cross-validation fold was removed because the BCR CDR3 clustering component (Model 2) abstained on enough of the fold's validation set that only examples from the "over 18" class remained for training a metamodel. This absence of "under 18" samples in one fold stems from two design decisions. First, the validation set includes fewer samples than the train or test sets, and it gets even smaller after filtering to healthy donors only for this analysis. Second, we use the same cross-validation splits for all analyses; they were designed to split diseases evenly, not ages.

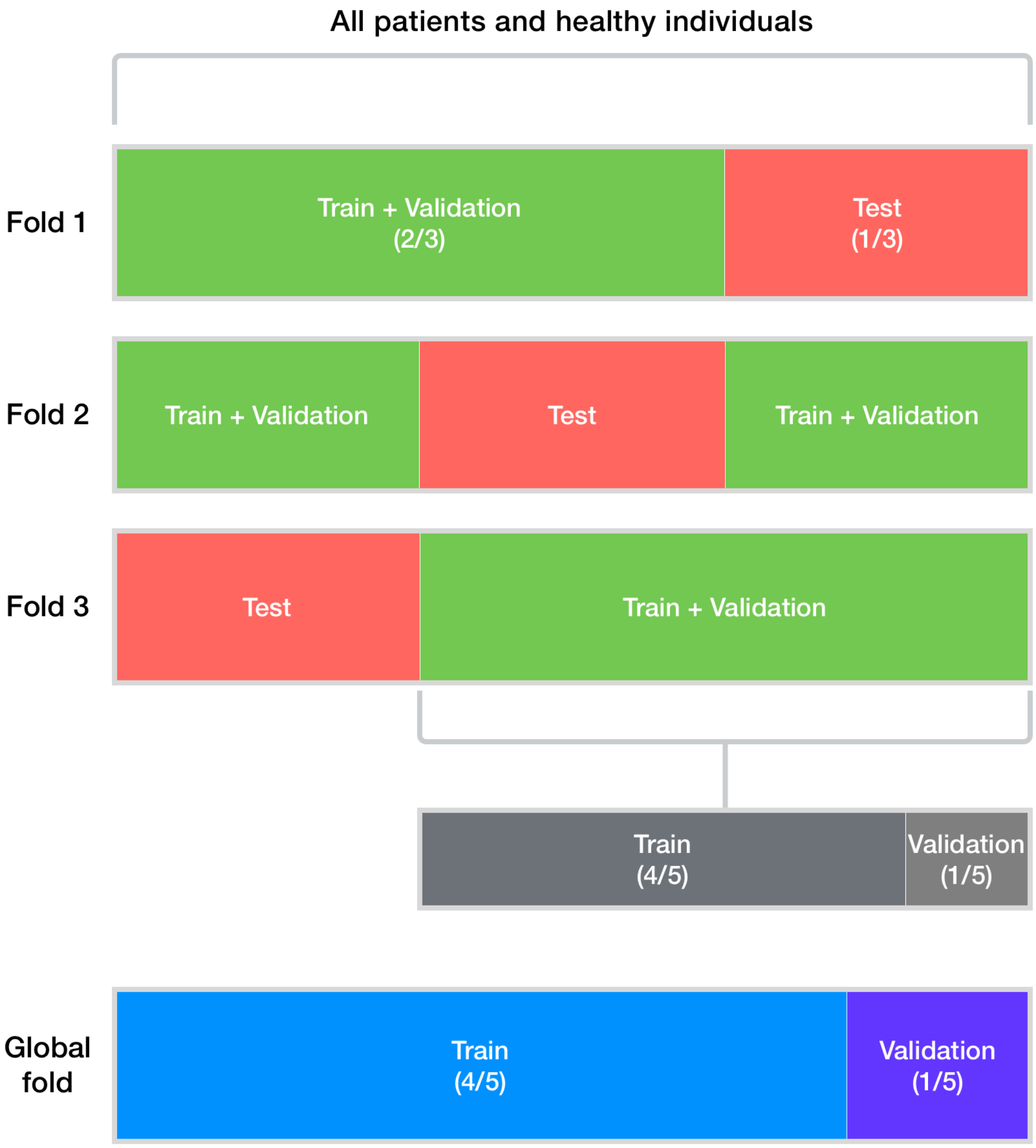
Input	Prediction target	Accuracy	ROC AUC
Age	Disease (358 BCR+TCR samples from individuals with known age, sex, and ancestry)	46.6%	0.704
Sex		33.2%	0.579
Ancestry		56.4%	0.785
Age, sex, ancestry		66.5%	0.856
BCR + TCR sequence features, age, sex, ancestry, and interaction terms between sequence and demographic features		86.6% (1.7% abstentions)	0.980
BCR + TCR sequences features with age, sex, and ancestry regressed out		84.1% (1.7% abstentions)	0.969

Supplementary Table 6: Classification results for disease prediction with demographics-aware variants of the *Mal-ID* random forest ensemble model. (When age is incorporated as a feature, it is treated as a continuous variable.) We report held-out test set performance averaged over three cross-validation folds. Abstentions hurt accuracy scores (they count as incorrect predictions), but are not included in the calculation of the probability-based AUC metric, because no predicted class probabilities are generated for abstained samples.

Immune state	**BCR**	**TCR**
Covid-19	0.341 +/- 0.040	0.050 +/- 0.052
SLE	0.170 +/- 0.078	0.171 +/- 0.032
Healthy/Background	0.148 +/- 0.033	0.143 +/- 0.041

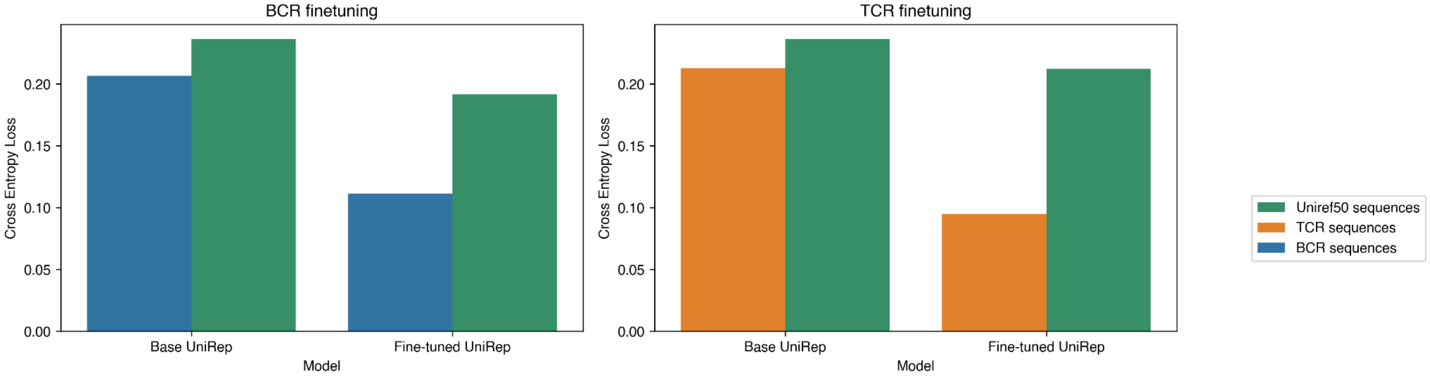
Supplementary Table 7: kBET batch effect measurement of average rejection rate of the null hypothesis that the batch distribution in a sequence’s local neighborhood is the same as the global batch distribution (reporting average +/- standard deviation across 3 folds). Values closer to 0 indicate the null hypothesis is rarely rejected and suggest the batches are well mixed.

Supplementary Figure 1



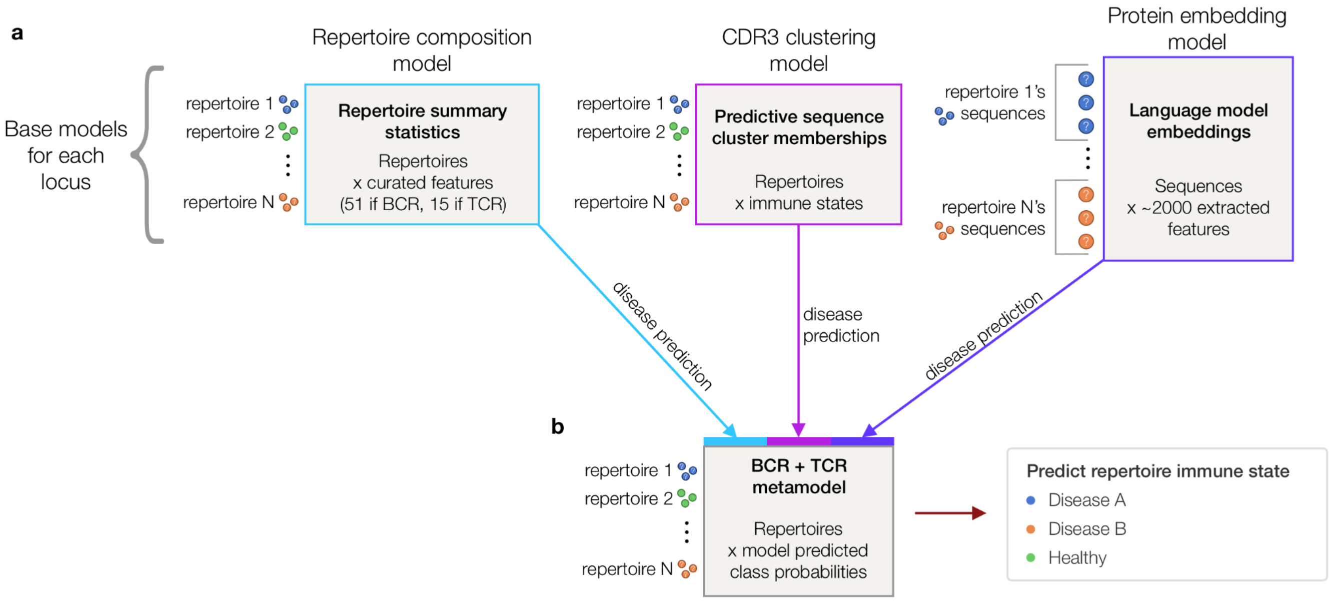
Schematic of cross-validation strategy. In each of three folds, individuals are divided into a train, validation, and test set; that all sequences from an individual are only in the train, only in the validation, or only in the test set. We also created a “global fold” to train a final model on the entire dataset, for downstream evaluation on independent cohorts.

Supplementary Figure 2



Fine tuning the UniRep language model on BCR heavy chain and TCR beta chain sequences led to a reduction in cross entropy loss (i.e. improved performance) on the BCR and TCR datasets, respectively, without causing an increase (i.e. without hurting performance) on the original UniRep training dataset, called UniRef50[89]. Here, we show the result of BCR or TCR fine-tuning for the "global" fold in the *Mal-ID* cross-validation strategy, with 20 bootstrap samples of 1000 UniRef50 sequences and 1000 *Mal-ID* global fold validation set sequences. Extraneous proteins (longer than 2000 amino acids or containing X, B, Z, or J amino acids) were removed from UniRef50, as in the original UniRep publication[44]. This result demonstrates that fine-tuning preserves knowledge of global protein patterns learned by base UniRep, i.e. no catastrophic forgetting occurs.

Supplementary Figure 3

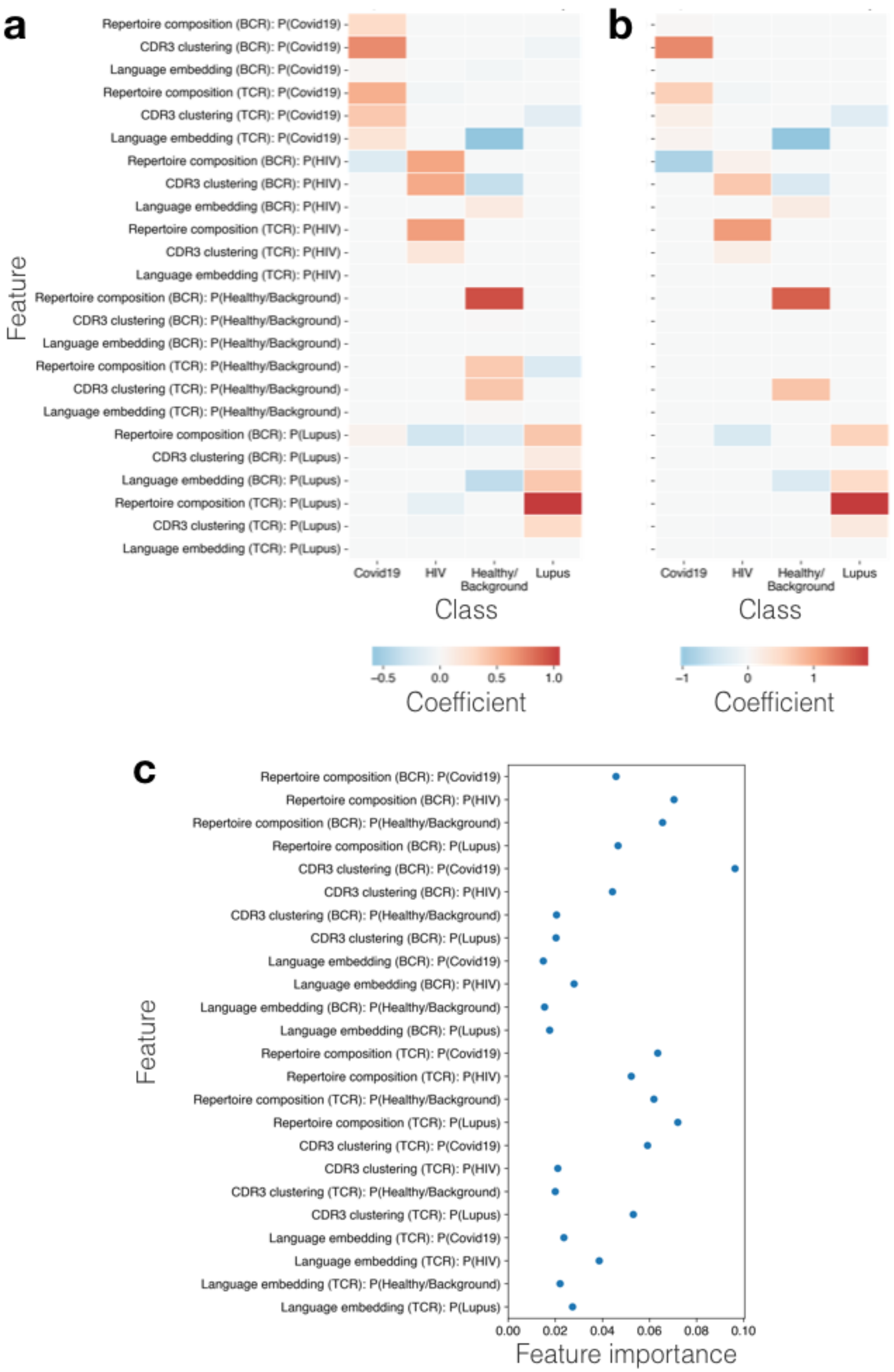


The *Mal-ID* classification pipeline for disease prediction (or other prediction tasks) has two stages.

a, stage one: we fit three models per locus (i.e. three IgH and three TRB models) on a cross-validation fold's training set.

b, stage two: we fit a metamodel on the validation set to ensemble the three inner models per locus.

Supplementary Figure 4



Supplementary Figure 4, continued

The *Mal-ID* ensemble model's feature importances for disease classification suggests that all feature extraction approaches contribute, but that immune signals are spread between B and T cell repertoires in different ways depending on the disease type.

We show feature importances for the "global fold" (i.e. for the final model fit with the full dataset), in three different versions of the ensemble model:

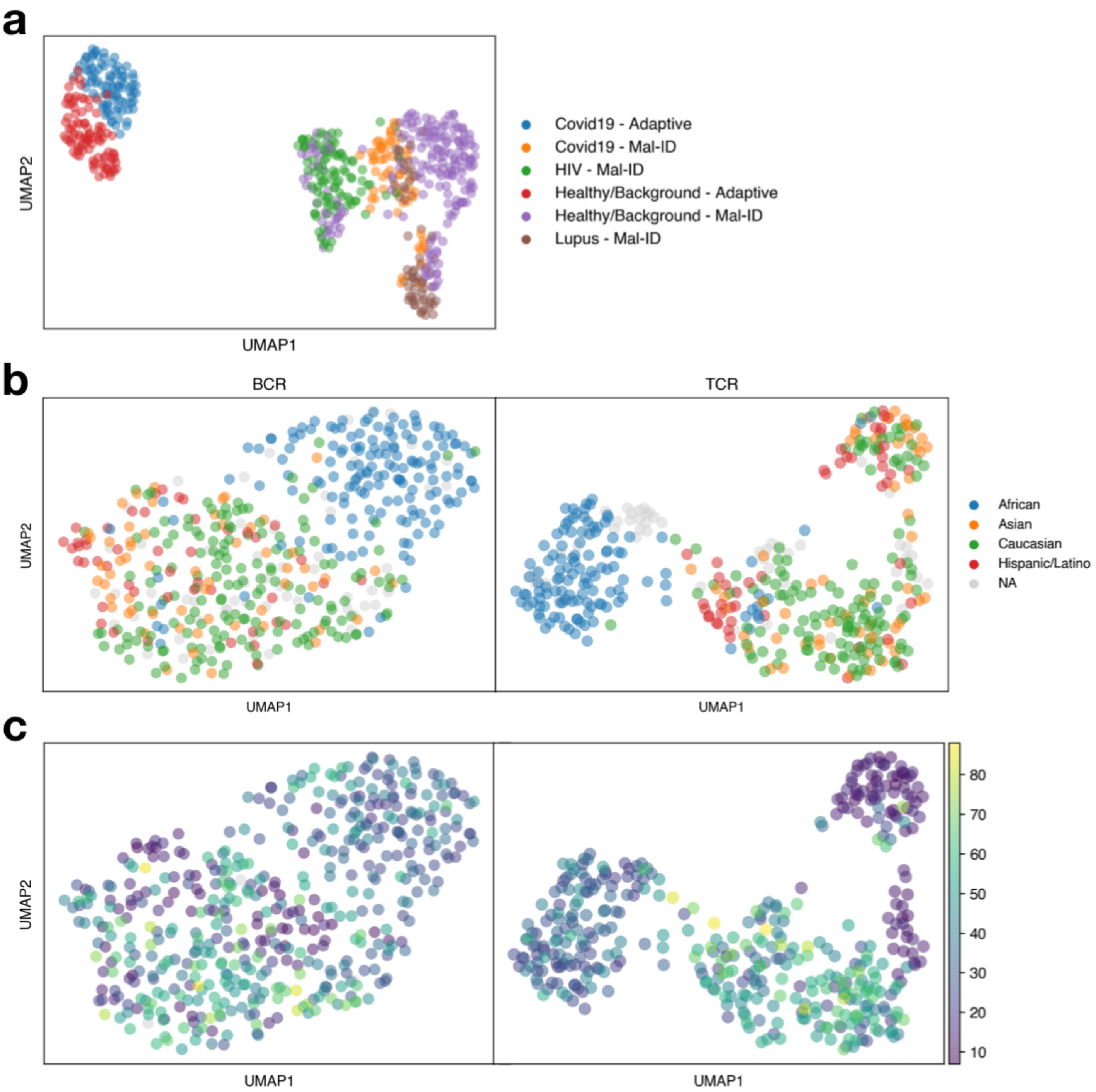
a, elastic net logistic regression (AUC 0.982 +/- 0.005 across 3 cross-validation folds);

b, lasso logistic regression (AUC 0.983 +/- 0.005);

c, random forest (AUC 0.981 +/- 0.013), which does not delineate feature contribution to each class.

Each feature is named for the *Mal-ID* subcomponent it originated from. For example, "*Repertoire composition (BCR): P(Covid19)*" is the feature coefficient for the BCR IgH repertoire composition model's predicted probability for the Covid-19 class (all base model predicted class probabilities were concatenated to form the input to the ensemble model). The random forest (**c**), unlike the other models, does not have feature contributions delineated by target class; instead the plot reflects how much each feature contributes to the overall classification task across all immune states.

Supplementary Figure 5

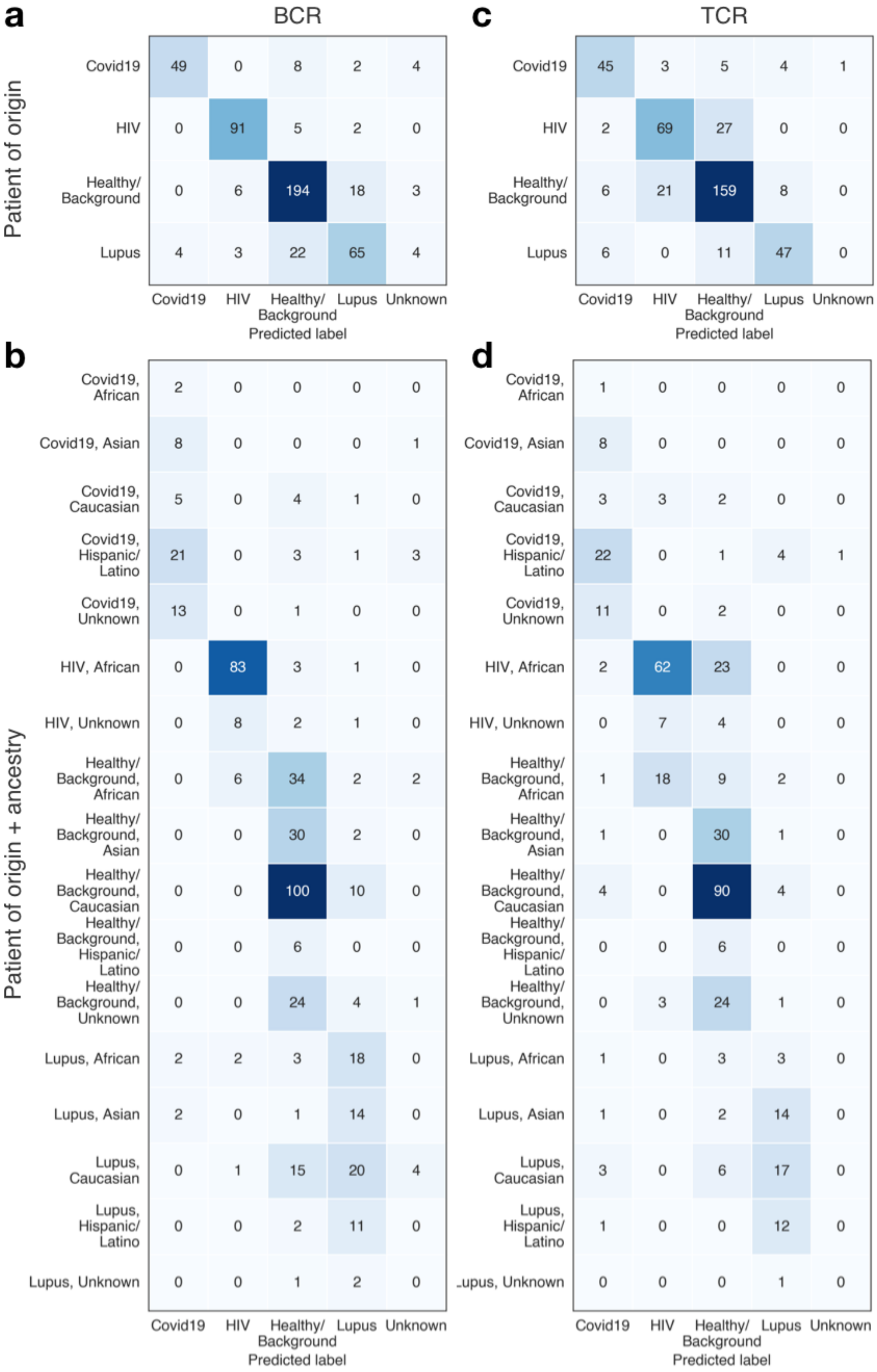


Supplementary Figure 5, continued

a, External cohorts from sequencing strategies different from the cDNA sequencing approach, such as Adaptive sequencing[13,72], have different V gene usage than the *Mal-ID* dataset. A UMAP of TRBV gene use proportions by sample (excluding rare V genes, to avoid disproportionate effects from minute differences in their proportions) shows that Adaptive cohort V gene use is systematically different from our cohorts.

b-c, V gene usage proportions of IgH (left panels) and TRB (right panels) repertoires in the *Mal-ID* dataset, visualized with UMAP and colored by ancestry (**b**) or age (**c**), show that demographic traits are related to V gene usage trends. (Rare V genes are again excluded.)

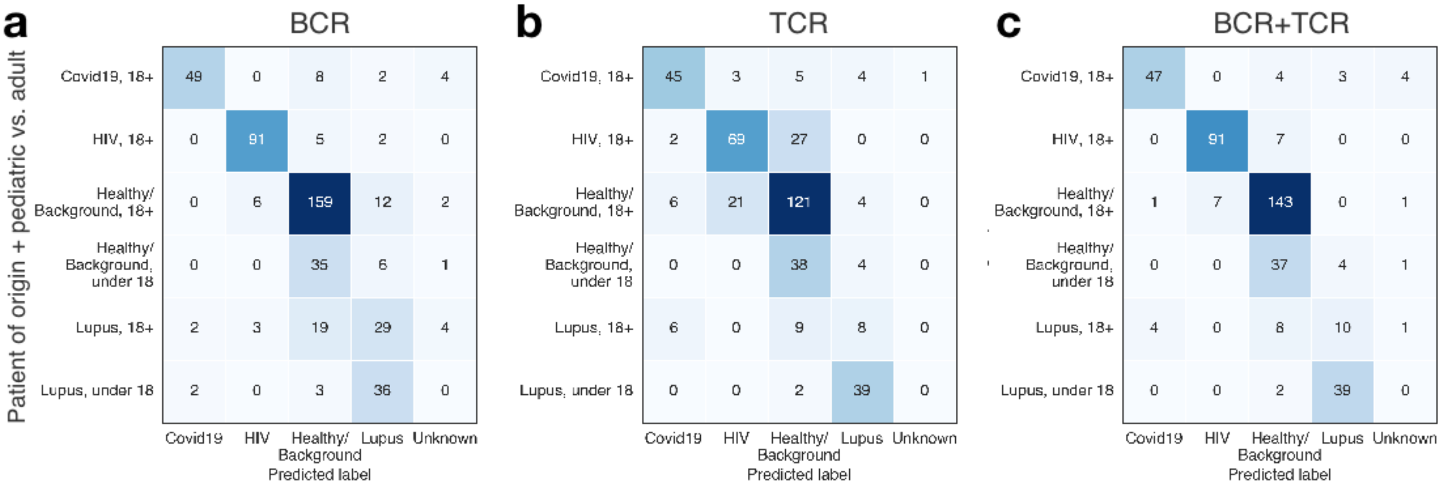
Supplementary Figure 6



Supplementary Figure 6, continued

BCR-only (**a-b**) and TCR-only (**c-d**) ensemble models show differences in disease classification. Delineating by the ground truth disease status and ancestry of each sample (**b, d**) shows that the "Healthy/Background - African" cohort, a healthy control group corresponding to the HIV cohort and whose members are predominantly African and live in Africa, is misclassified as HIV by the TCR model, but not by the BCR model. (The BCR and TCR metamodels have a different total number of samples due to BCR-only cohorts.)

Supplementary Figure 7

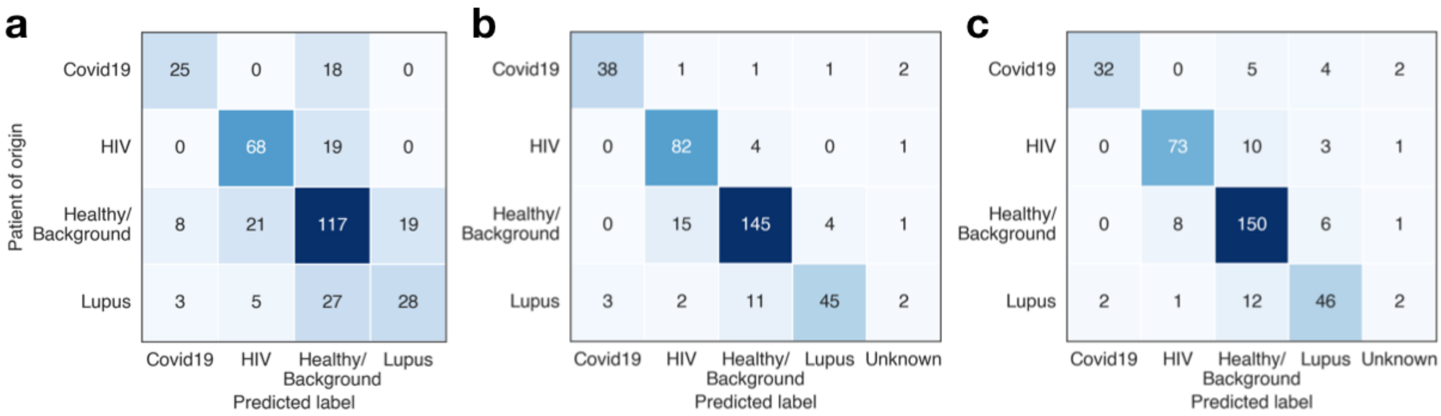


Metamodel classification performance, delineated by the ground truth disease status and age of each sample, shows that *Mal-ID* successfully differentiates between pediatric samples of different immune states.

a, BCR-only metamodel; **b**, TCR-only metamodel; **c**, BCR + TCR metamodel.

(The BCR and TCR metamodels have a different total number of samples due to BCR-only cohorts.)

Supplementary Figure 8



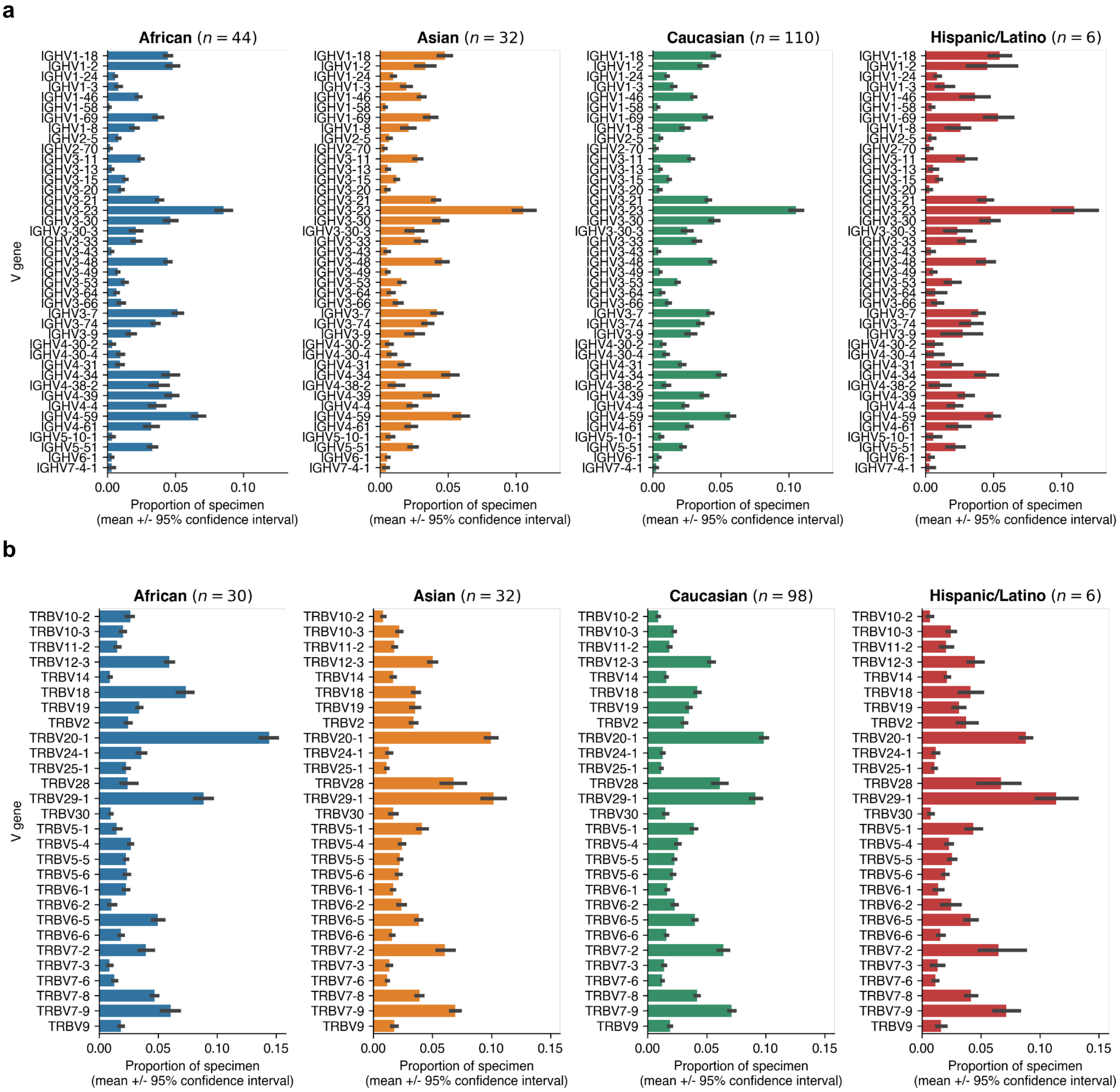
Demographic covariates have limited impact on disease classification.

a, Metamodel classification performance using only age, sex, and ancestry features, without any sequence features.

b, Metamodel classification performance using age, sex, and ancestry demographic features, along with sequence features, and interaction terms between these two sets of features.

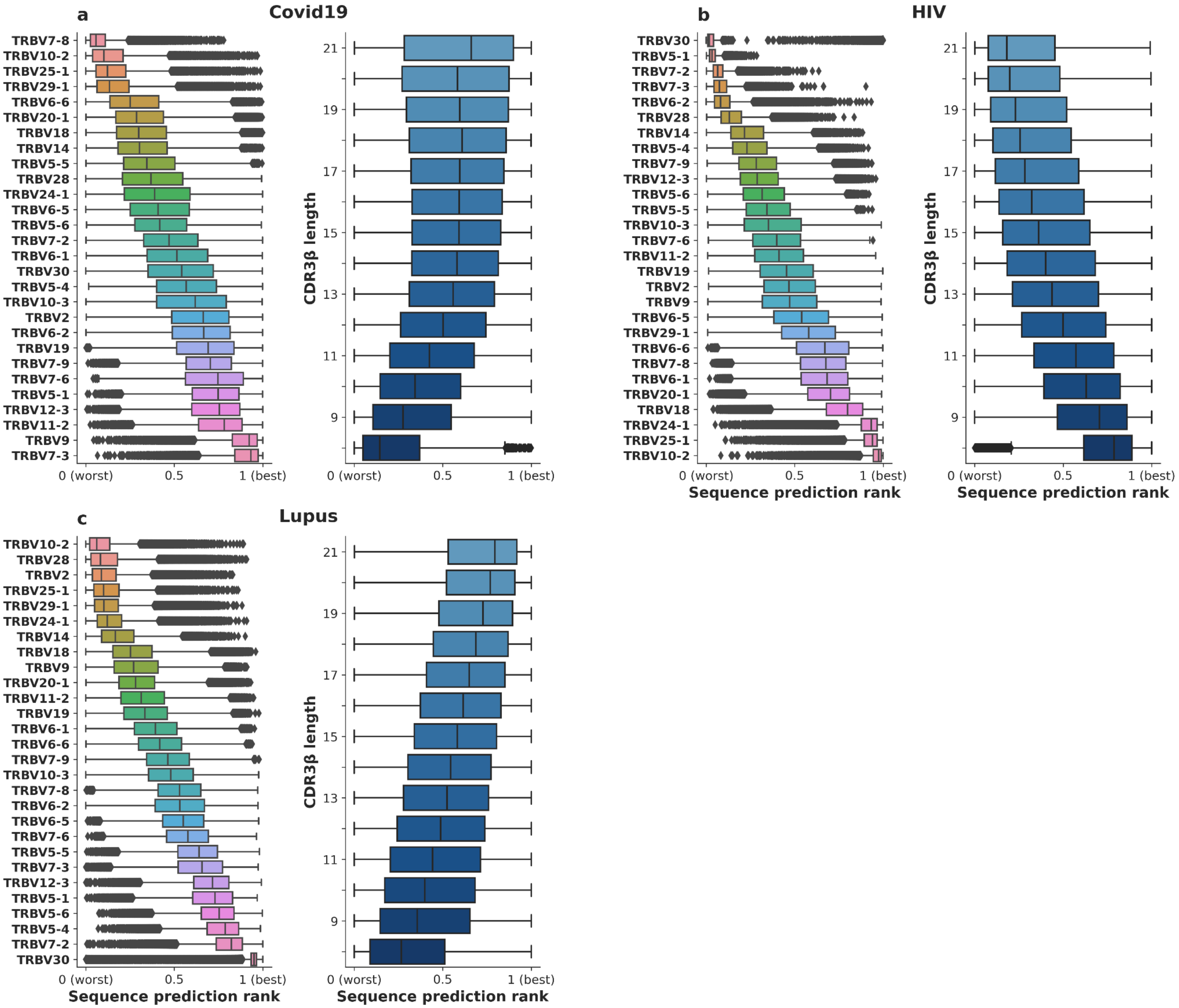
c, Metamodel classification performance using sequence features only, with age, sex, and ancestry regressed out.

Supplementary Figure 9



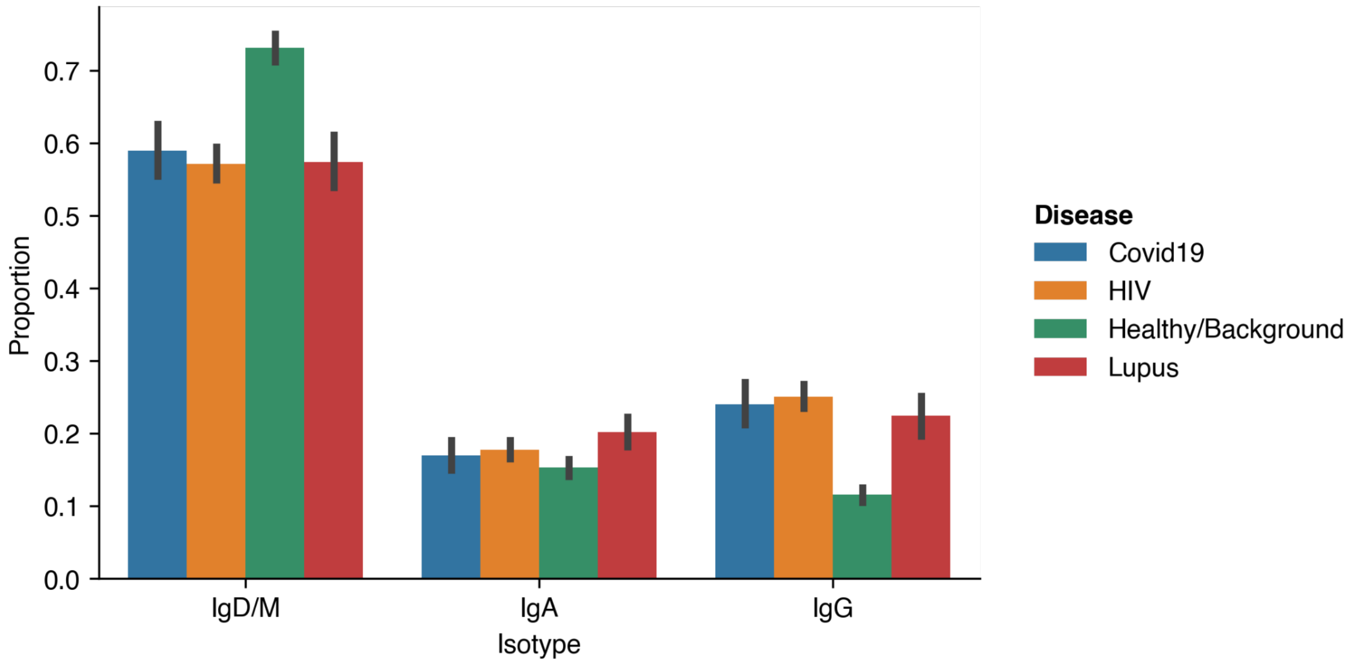
IGHV or TRBV gene use proportions in healthy control samples, stratified by ancestry, suggest that some V gene usage is related to ancestry. Average and 95% confidence interval plotted. **a,** BCR (note higher sample sizes due to presence of BCR-only cohorts). **b,** TCR.

Supplementary Figure 10



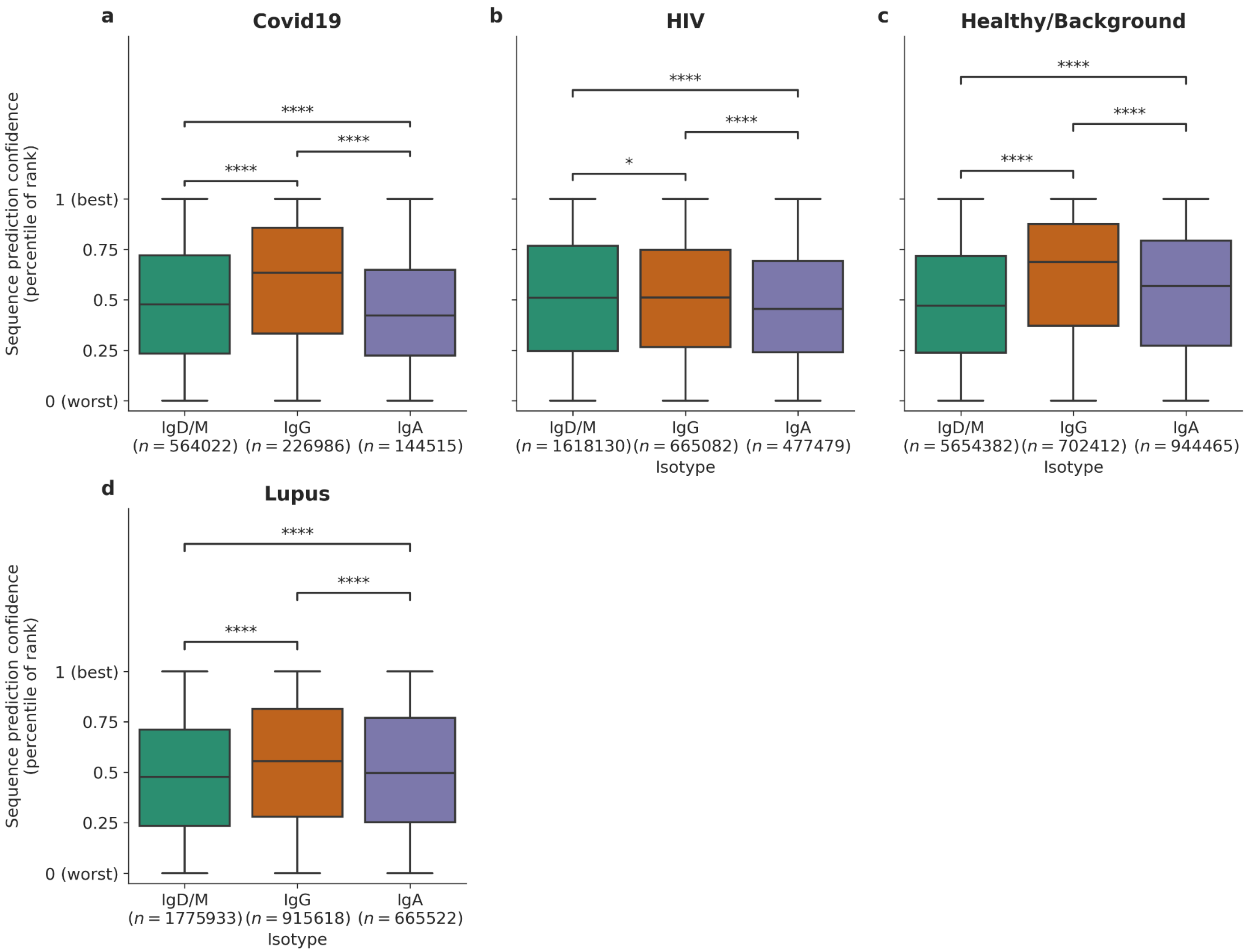
Disease patient-originating TRB sequences, ranked by predicted disease class probability, show high ranks for certain TRBV genes and for certain CDR3 length patterns reflecting selection.

Supplementary Figure 11



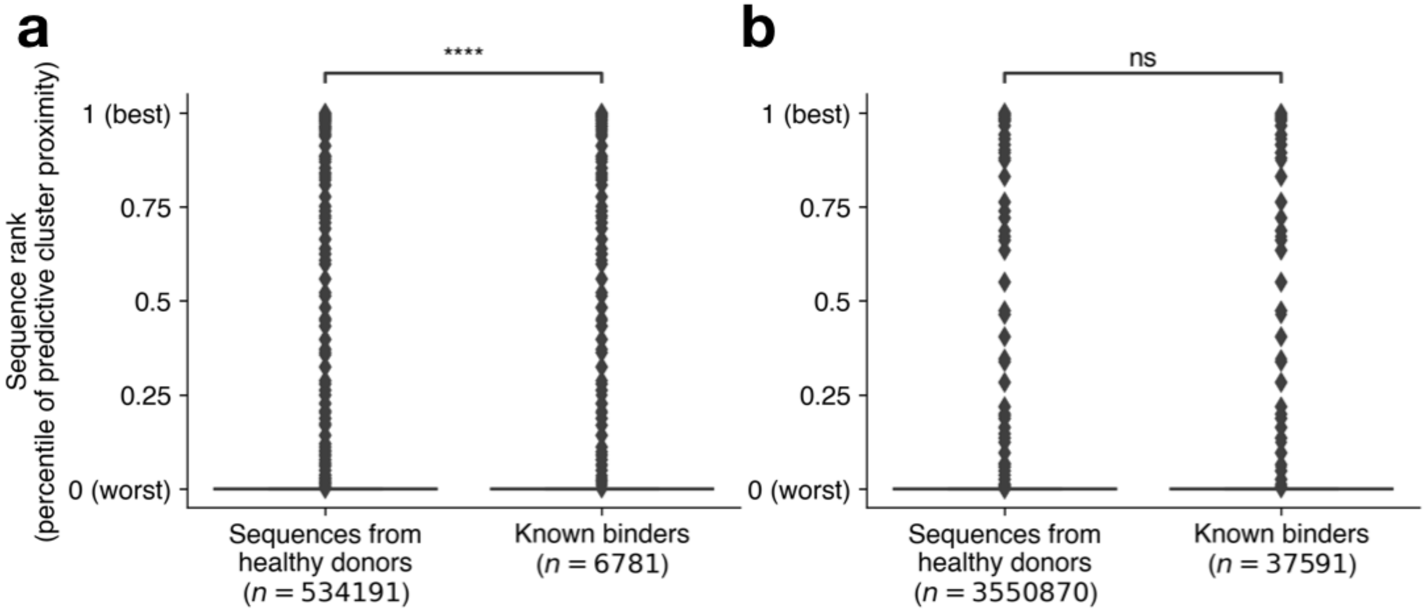
Average isotype proportions per sample present in the data (with 95% confidence interval shown, as well) are different between immune states. Differences in isotype proportions are technical artifacts and are corrected for in our analysis scheme to ensure that the models do not learn disease classification based on isotype proportion.

Supplementary Figure 12



Disease patient-originating sequences, ranked by predicted disease class probability and grouped by isotype, show subtle favoring of particular isotypes for predicting each disease. Significance was tested for each isotype pair in each panel. * means p <= 0.05 and **** 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 13



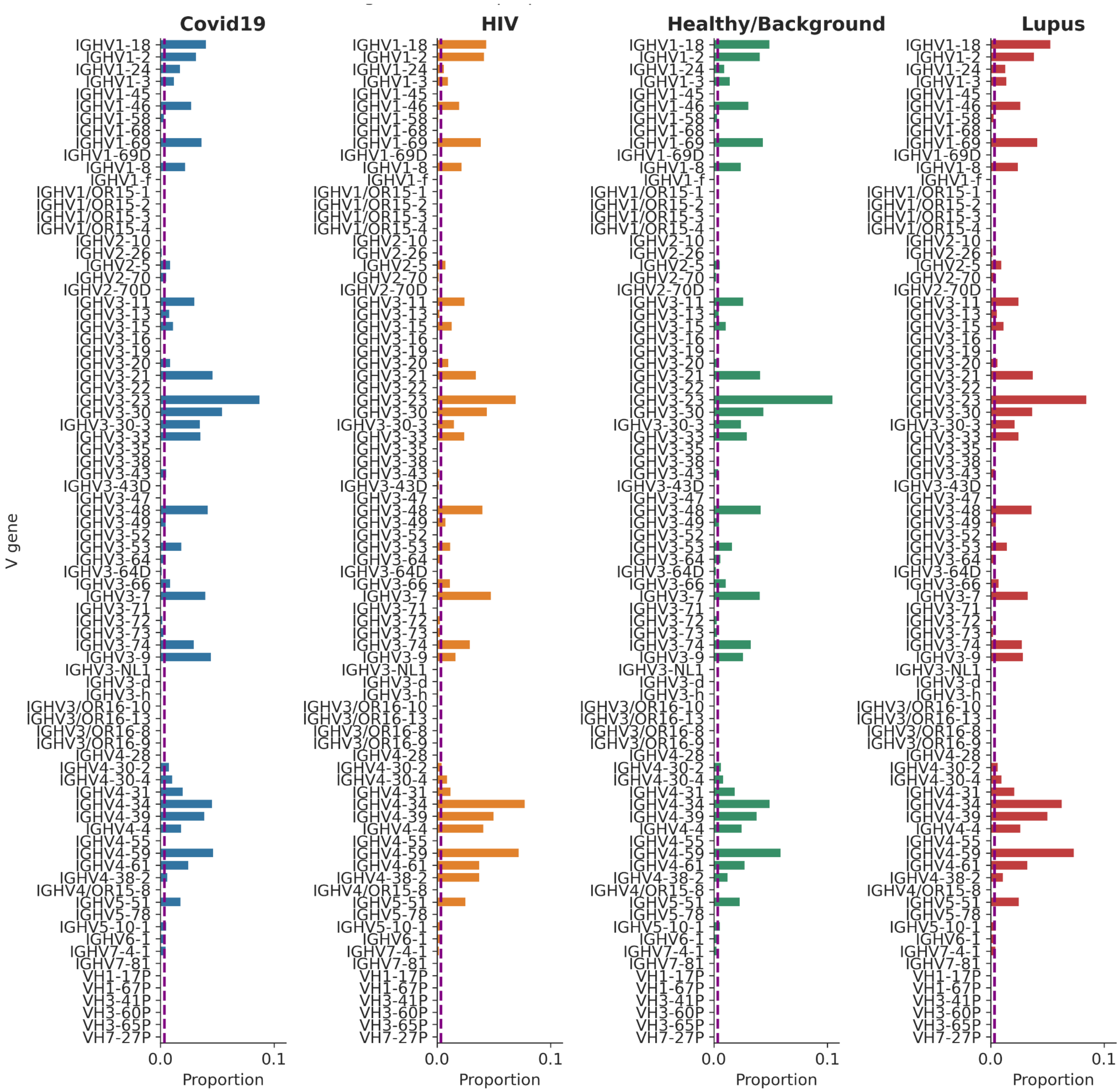
Validated SARS-CoV-2 associated sequences and healthy donor sequences are not well separated when ranked by distance to nearest Covid-19 associated cluster found by the CDR3 clustering model. One cross-validation fold is shown, along with a one-sided Wilcoxon rank-sum test for increased ranks among known binder sequences.

a, IgH sequences: U-statistic = 1.9e9, p < 1e-52. **b,** TRB sequences: U-statistic = 6.5e10, p = 1.0.

High rank (ranging up to 1.0) indicates high proximity to Covid-19 associated sequences. The model finds clusters among sequences with the same V gene, J gene, and CDR3 length. Therefore, if a query sequence has a V gene, J gene, and CDR3 length for which there are no Covid-19 associated clusters, then it is considered to have infinite distance from a disease-predictive cluster and zero (worst) rank. The vast majority of sequences have zero rank, as a result.

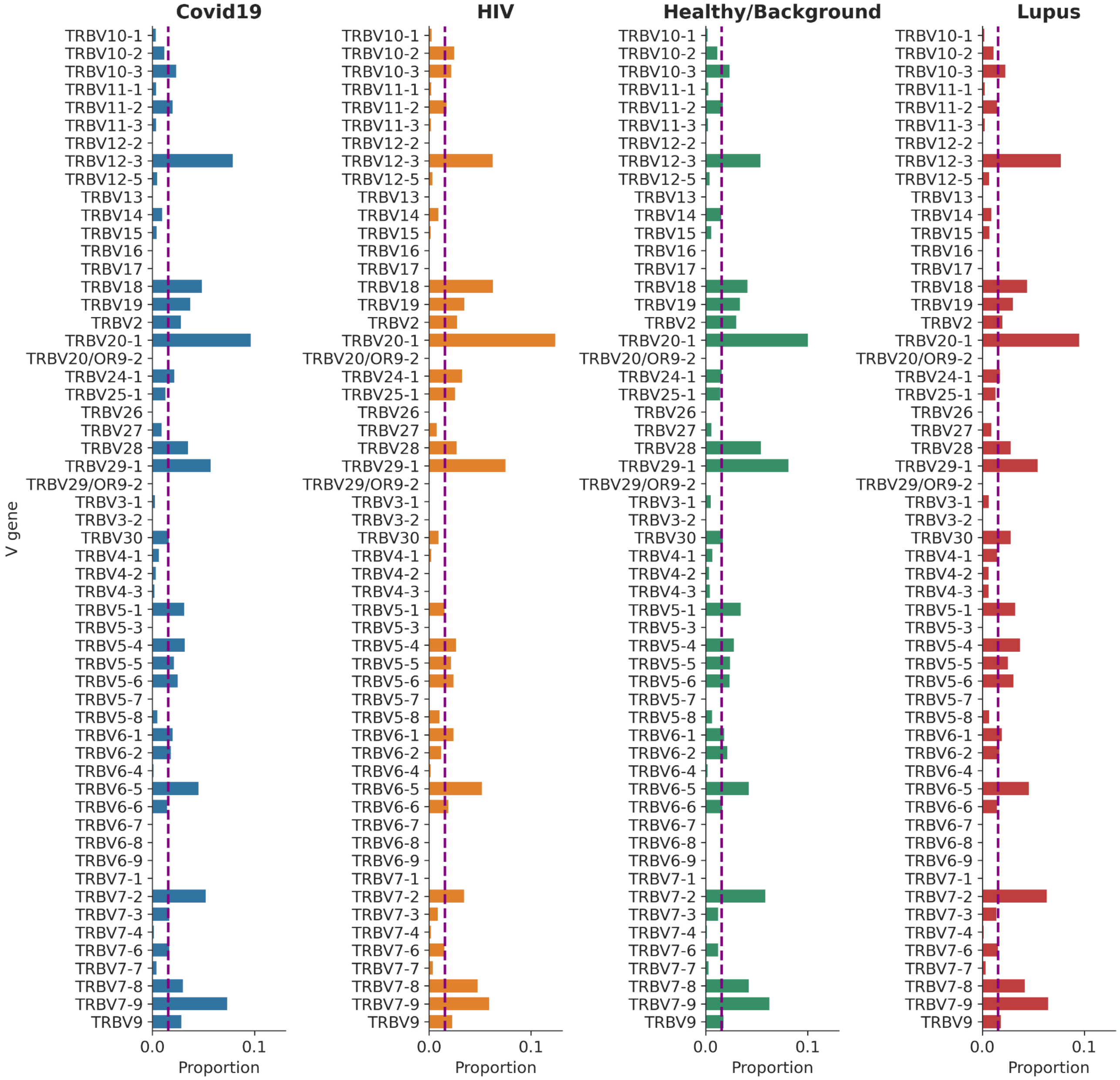
Supplementary Figure 14

a



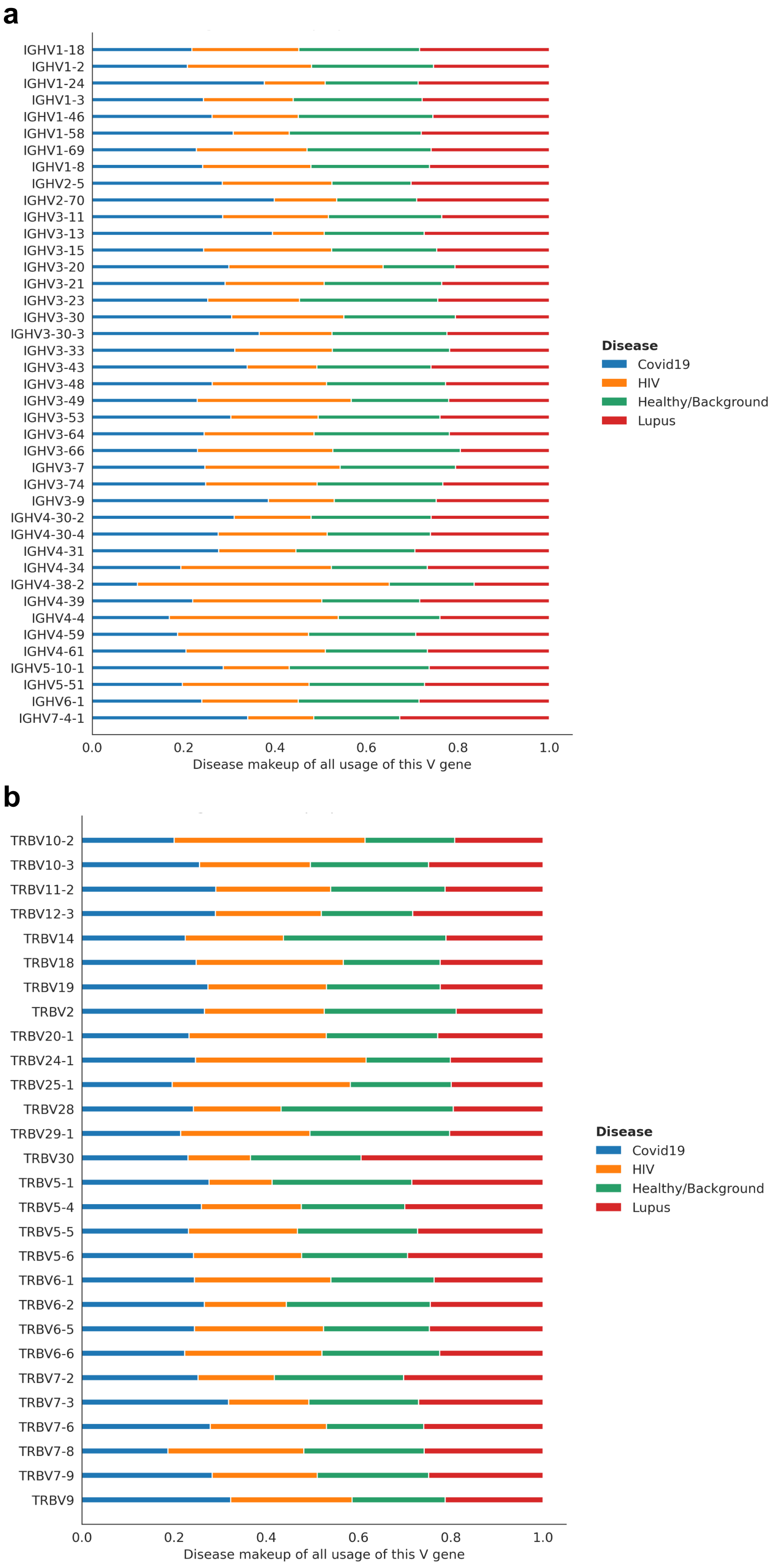
Supplementary Figure 14, continued

b



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

Supplementary Figure 15



Stacked bar plots representing how prevalent each IGHV and TRBV gene is by disease, after filtering out rare V genes. **a**, IGHV; **b**, TRBV.

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 YW, Widen R, Drews SJ. Practical Guidance for Clinical Microbiology Laboratories: Viruses Causing Acute Respiratory Tract Infections. Clin Microbiol Rev. 2019 Jan;32(1). 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. 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

11. 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.

12. Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, Uddin I, Ismail M, Oakes T, Chain B, Eugster A, Kashofer K, Rainer PP, Darko S, Ransier A, Douek DC, Klatzmann D, Mariotti-Ferrandiz E. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nat Biotechnol. 2021 Feb;39(2):236–245. PMCID: 5772558

13. 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

14. Roskin KM, Jackson KJL, Lee JY, Hoh RA, Joshi SA, Hwang KK, Bonsignori M, Pedroza-Pacheco I, Liao HX, 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

15. Yang F, Nielsen SCA, Hoh RA, Röltgen K, Wirz OF, Haraguchi E, Jean GH, Lee JY, 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

16. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF, Walter EB, Liao HX, 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

17. 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

18. 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

19. Liu X, Zhang W, Zhao M, Fu L, Liu L, Wu J, Luo S, Wang L, Wang Z, Lin L, Liu Y, Wang S, Yang Y, Luo L, Jiang J, Wang X, Tan Y, Li T, Zhu B, Zhao Y, Gao X, Wan Z, Huang C, Fang M, Li Q, Peng H, Liao X, Chen J, Li F, Ling G, Zhao H, Luo H, Xiang Z, Liao J, Liu Y, Yin H, Long H, Wu H, Yang H, Wang J, Lu Q. T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. Ann Rheum Dis. 2019 Aug;78(8):1070–1078. PMID: 31101603

20. Zhang H, Liu L, Zhang J, Chen J, Ye J, Shukla S, Qiao J, Zhan X, Chen H, Wu CJ, Fu YX, Li B. Investigation of antigen-specific T-cell receptor clusters in human cancers. Clin Cancer Res. American Association for Cancer Research (AACR); 2020 Mar 15;26(6):1359–1371. PMID: 31831563

21. Chronister WD, Crinklaw A, Mahajan S, Vita R, Koşaloğlu-Yalçın Z, Yan Z, Greenbaum JA, Jessen LE, Nielsen M, Christley S, Cowell LG, Sette A, Peters B. TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. Front Immunol. 2021 Mar 11;12:640725. PMCID: PMC7991084

22. 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

23. 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.

24. Beshnova D, Ye J, Onabolu O, Moon B, Zheng W, Fu YX, Brugarolas J, Lea J, Li B. De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. Sci Transl Med. 2020 Aug 19;12(557). PMCID: PMC7887928

25. 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

26. Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. Patterns (N Y). 2022 Jul 8;3(7):100513. PMCID: PMC9278498

27. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv [q-bio.BM]. 2021.

28. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinformatics Advances. 2022;2(1).

29. Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning. Patterns (N Y). 2022 Feb 11;3(2):100406. PMCID: PMC8848015

30. Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, Svozil D, Bitton DA. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. MAbs. 2022 Jan-Dec;14(1):2020203. PMCID: PMC8837241

31. Wu K, Yost KE, Daniel B, Belk JA, Xia Y, Egawa T, Satpathy A, Chang HY, Zou J. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. bioRxiv. 2021. p. 2021.11.18.469186.

32. Sidhom JW, 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

33. 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. bioRxiv. 2020. p. 2020.04.12.038158.

34. 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. bioRxiv. 2020. p. 2020.02.25.965673.

35. Sethna Z, Isacchini G, Dupic T, Mora T, Walczak AM, Elhanati Y. Population variability in the generation and selection of T-cell repertoires. PLoS Comput Biol. 2020 Dec;16(12):e1008394. PMCID: PMC7725366

36. Isacchini G, Walczak AM, Mora T, Nourmohammad A. Deep generative selection models of T and B cell receptor repertoires with soNNia. Proc Natl Acad Sci U S A. 2021 Apr 6;118(14). PMCID: PMC8040596

37. 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

38. 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

39. 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

40. Nielsen SCA, Yang F, Jackson KJL, Hoh RA, Röltgen K, Jean GH, Stevens BA, Lee JY, 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

41. 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

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

43. Wu NC, Yuan M, Liu H, Lee CCD, 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

44. 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

45. 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. 2021 Apr 13;118(15). PMCID: PMC8053943

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

47. 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.

48. Alexandari A, Kundaje A, Shrikumar A. Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation. In: Iii HD, Singh A, editors. Proceedings of the 37th International Conference on Machine Learning. PMLR; 13--18 Jul 2020. p. 222–232.

49. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, Olshen RA, Weyand CM, Boyd SD, Goronzy JJ. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci U S A. 2014 Sep 9;111(36):13139–13144. PMCID: PMC4246948

50. Geursen A, Skinner MA, Townsend LA, Perko LK, Farmiloe SJ, Peake JS, Simpson IJ, Fraser JD, Tan PL. Population study of T cell receptor V beta gene usage in peripheral blood lymphocytes: differences in ethnic groups. Clin Exp Immunol. 1993 Oct;94(1):201–207. PMCID: PMC1534351

51. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, Bolotin DA, Lukyanov S, Bogdanova EA, Mamedov IZ, Lebedev YB, Chudakov DM. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J Immunol. 2014 Mar 15;192(6):2689–2698. PMID: 24510963

52. Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, Mamedov IZ, Pogorelyy MV, Bolotin DA, Izraelson M, Davydov AN, Egorov ES, Kasatskaya SA, Rebrikov DV, Lukyanov S, Chudakov DM. Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. J Immunol. 2016 Jun 15;196(12):5005–5013. PMID: 27183615

53. Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK. Genetic variation in MHC proteins is

associated with T cell receptor expression biases. Nat Genet. 2016 Sep;48(9):995–1002. PMCID: PMC5010864

54. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, Deikus G, Auckland K, Eichler EE, Marasco WA, Sebra R, Sharp AJ, Smith ML, Bashir A, Watson CT. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. Front Immunol. 2020 Sep 23;11:2136. PMCID: PMC7539625

55. 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

56. 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

57. 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

58. 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

59. Yuan M, Liu H, Wu NC, Lee CCD, 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

60. Kim C, Ryu DK, Lee J, Kim YI, Seo JM, Kim YG, Jeong JH, Kim M, Kim JI, Kim P, Bae JS, Shim EY, Lee MS, Kim MS, Noh H, Park GS, Park JS, Son D, An Y, Lee JN, Kwon KS, Lee JY, Lee H, Yang JS, Kim KC, Kim SS, Woo HM, Kim JW, Park MS, Yu KM, Kim SM, Kim EH, Park SJ, Jeong ST, Yu CH, Song Y, Gu SH, Oh H, Koo BS, Hong JJ, Ryu CM, Park WB, Oh MD, Choi YK, Lee SY. A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein. Nat Commun. 2021 Jan 12;12(1):288. PMCID: PMC7803729

61. Brouwer PJM, Caniels TG, van der Straten K, Snitselaar JL, Aldon Y, Bangaru S, Torres JL, Okba NMA, Claireaux M, Kerster G, Bentlage AEH, van Haaren MM, Guerra D, Burger JA, Schermer EE, Verheul KD, van der Velde N, van der Kooi A, van Schooten J, van Breemen MJ, Bijl TPL, Sliepen K, Aartse A, Derking R, Bontjer I, Kootstra NA, Wiersinga WJ, Vidarsson G, Haagmans BL, Ward AB, de Bree GJ, Sanders RW, van Gils MJ. Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. Science. 2020 Aug 7;369(6504):643–650. PMCID: PMC7299281

62. 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 GY, 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

63. 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

64. 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

65. 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.

66. 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. 2021 Feb 5;371(6529). PMCID: PMC7919858

67. 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. 2020 Sep 4;369(6508). PMCID: PMC7402624

68. 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 HH, 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

69. 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

70. Braun J, Loyal L, Frentsch M, Wendisch D, Georg P, Kurth F, Hippenstiel S, Dingeldey M, Kruse B, Fauchere F, Baysal E, Mangold M, Henze L, Lauster R, Mall MA, Beyer K, Röhmel J, Voigt S, Schmitz J, Miltenyi S, Demuth I, Müller MA, Hocke A, Witzenrath M, Suttorp N, Kern F, Reimer U, Wenschuh H, Drosten C, Corman VM, Giesecke-Thiel C, Sander LE, Thiel A. SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. Nature. 2020 Nov;587(7833):270–274. PMCID: 6988269

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

72. Nolan S, Vignali M, Klinger M, Dines JN, Kaplan IM, Svejnoha E, Craft T, Boland K, Pesesky M, Gittelman RM, Snyder TM, Gooley CJ, Semprini S, Cerchione C, Mazza M, Delmonte OM, Dobbs K, Carreño-Tarragona G, Barrio S, Sambri V, Martinelli G, Goldman JD, Heath JR, Notarangelo LD, Carlson JM, Martinez-Lopez J, Robins HS. A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. Res Sq. 2020 Aug 4; PMCID: PMC7418738

73. 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 LM, 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

74. 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. 2022 Jan 25;

75. 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

76. 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

77. 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

78. Ma EJ, Kummer A. Reimplementing Unirep in JAX. bioRxiv. 2020. p. 2020.05.11.088344.

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

80. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. Wiley; 2005 Apr;67(2):301–320.

81. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019 Jan;16(1):43–49. PMCID: 5112579

82. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, Theis FJ. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022 Jan;19(1):41–50. PMCID: PMC8748196

83. Kim SI, Noh J, Kim S, Choi Y, Yoo DK, Lee Y, Lee H, Jung J, Kang CK, Song KH, Choe PG, Kim HB, Kim ES, Kim NJ, Seong MW, Park WB, Oh MD, Kwon S, Chung J. Stereotypic neutralizing VH antibodies against SARS-CoV-2 spike protein receptor binding domain in patients with COVID-19 and healthy individuals. Sci Transl Med. 2021 Jan 27;13(578). PMCID: PMC7875332

84. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019 Feb;566(7744):393–397. PMCID: PMC6411386

85. Shomuradova AS, Vagida MS, Sheetikov SA, Zornikova KV, Kiryukhin D, Titov A, Peshkova IO, Khmelevskaya A, Dianov DV, Malasheva M, Shmelev A, Serdyuk Y, Bagaev DV, Pivnyuk A, Shcherbinin DS, Maleeva AV, Shakirova NT, Pilunov A, Malko DB, Khamaganova EG, Biderman B, Ivanov A, Shugay M, Efimov GA. SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. Immunity. 2020 Dec 15;53(6):1245-1257.e5. PMCID: PMC7664363

86. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, Knoetze N, Breden FMW, Christley S,

Scott JK, Cowell LG, Breden F. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev. 2018 Jul;284(1):24–41. PMCID: PMC6344122

87. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [stat.ML]. 2018.

88. Nielsen SCA, Roskin KM, Jackson KJL, Joshi SA, Nejad P, Lee JY, Wagar LE, Pham TD, Hoh RA, Nguyen KD, Tsunemoto HY, Patel SB, Tibshirani R, Ley C, Davis MM, Parsonnet J, Boyd SD. Shaping of infant B cell receptor repertoires by environmental factors and infectious disease. Sci Transl Med. 2019 Feb 27;11(481). PMCID: PMC6733608

89. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015 Mar 15;31(6):926–932. PMCID: PMC4375400