Modeling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals

Maria Ruiz Ortega, Natanael Spisak, Thierry Mora,* and Aleksandra M. Walczak*
Laboratoire de physique de l'École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, 75005 Paris, France

Adaptive immunity's success relies on the extraordinary diversity of protein receptors on B and T cell membranes. Despite this diversity, the existence of public receptors shared by many individuals gives hope for developing population wide vaccines and therapeutics. Yet many of these public receptors are shared by chance. We present a statistical approach, defined in terms of a probabilistic V(D)J recombination model enhanced by a selection factor, that describes repertoire diversity and predicts with high accuracy the spectrum of repertoire overlap in healthy individuals. The model underestimates sharing between repertoires of individuals infected with SARS-CoV-2, suggesting strong antigen-driven convergent selection. We exploit this discrepancy to identify COVID-associated receptors, which we validate against datasets of receptors with known viral specificity. We study their properties in terms of sequence features and network organization, and use them to design an accurate diagnosis tool for predicting SARS-CoV-2 status from repertoire data.

I. INTRODUCTION

The unique antigen receptors expressed on the surface of B and T cells determine the set of epitopes that may be recognised during an immune response. Public sequences shared by many individuals have for a long time solicited potential therapeutic interest [1]. However, many of these receptors are shared by chance [2–5] and not necessarily due to prior antigenic experience. Here, by comparing healthy individuals and COVID-19 patients, we explore the interplay of chance and exposure on T-cell and B-cell receptor sharing.

B-cell receptors (BCR) are composed of a heavy and a light chain, and T cell receptors (TCR) are made of analogous beta and alpha chains. Each chain is formed via a DNA editing mechanism. This process, called V(D)J recombination, randomly splices genes together from germline segments (V, D and J for heavy and β chains, and V and J for light and α chains). In addition, a random number of base pairs are trimmed at the junctions between the segments, and random non-templated ones are inserted. The resulting junctional region makes up the hypervariable complementarity determining region 3 (CDR3), which is key for antigen specificity [6, 7]. After generation, receptors are selected through thymic selection for T cells [8], and central tolerance for B cells [9], to ensure proper receptor functionality and to avoid recognition of self-antigens.

Estimates of the number of possible combinations produced by V(D)J recombination range from 10^{20} [10] to 10^{61} [11] for TCR, and even more for BCR. These numbers far exceed the number of distinct antigen receptors, $\sim 10^8 - 10^{10}$ [12, 13], or even the total number of B and T cells in the human body, $\sim 10^{12}$ [14]. Despite this large difference, unrelated individuals share a considerable number of public sequences, as observed both in antigen-specific repertoires and in the whole repertoire as revealed by high-throughput sequencing [13, 15, 16].

While the sharing of public sequences was initially attributed to selection biases due to encountered antigens, it was then proposed that public sequences may be explained by convergent recombination [2]. The V(D)J recombination process generates biases that favor certain sequences over others, and those frequent sequences are likely to be found in many individuals. However convergent recombination itself is not enough to explain the amount of overlap between the TCR repertoires of unrelated individuals [5]. Convergent selection, which amplifies biases in sequence statistics, is further needed to explain it quantitatively. Convergent selection can either stem from peripheral selection triggered by encountered pathogens or avoiding self-antigens. Learning a single, sequence-independent correction factor allows one to predict the full spectrum of sharing observed among a large cohort of donors [5]. However, this approach ignores sequence-specific features of selection, and can only imperfectly predict which sequences are public or private. In addition, it has not been applied to B cells, for which there is no quantitative theory of repertoire sharing.

Here we develop a detailed framework to explain and quantitatively predict immune receptor sharing through both convergent recombination and convergent selection, as schematized in Fig. 1. We integrate selection pressures acting on receptor sequences [17, 18] into a statistical theory of repertoire sharing [5]. We apply this framework to successfully predict the sharing spectrum of TCR and BCR repertoires in human cohorts. By contrasting predictions in cohorts of healthy donors against patients recently infected with COVID-19, we identify lists of "overshared" sequences with putative specificity for SARS-CoV-2, which we validate against existing databases. Our approach not only predicts sharing among healthy individuals, but can also be used to identify antibodies and T-cell receptors of interest that are both specific and

*Corresponding authors. These authors contributed equally.

public, opening new avenues for applications of repertoire sequencing to diagnostics, precision medicine and vaccine design.

II. RESULTS

Predicting sharing in healthy BCR repertoires

We aim to quantify how convergent recombination and selection influence and shape the repertoire landscape. To that end, we examined the degree of sequence overlap across different BCR and TCR repertoires. We first combine all the unique CDR3 amino acid sequences of immunoglobulin heavy chains (IGH) or T cell β chains (TCRβ), which we refer to hereafter as “clonotypes,” from each individual of the respective cohorts. For each clonotype, we computed their *sharing number*, i.e. the number of individuals in which it appears. The sharing number offers a nuanced view in which clonotypes are neither purely “public” nor “private”, but may instead have different degrees of publicness. The “sharing spectrum” can be visualized by plotting the distribution of sharing numbers, i.e. the number of clonotypes corresponding to each sharing number (Fig. 1).

We investigated sharing in IGH repertoires. We analyzed high-throughput sequencing data from a cohort of 10 healthy donors totalling more than 3×10^8 IGH sequences, separated into IgM and IgG subpopulations [13]. The sharing number distribution of their IgM repertoires is shown in Fig. 2A (blue crosses). To explore the forces behind the existence of public sequences, we compared the distribution of sharing number in experimentally sampled datasets with theoretical predictions from previously described models for immune receptor sequences. Given a model $P(\sigma)$ for the probability of occurrence of any clonotype σ, we can predict the sharing spectrum either by simulating synthetic repertoires, or by using the technique of generating functions, which is computationally faster (see [5] and Methods).

The first model we consider describes the generation process in terms of the statistics of the V(D)J recombination process, denoted by $P_{\text{gen}}(\sigma)$. It is inferred from the nonproductive nucleotide sequences of the IgM repertoires of all donors, using the IGoR tool [19]. This model accounts for convergent recombination, which is the most evident driver of sharing. However, P_{gen} alone underestimates the amount of sharing (Fig. 2A, green curve). We hypothesized that this discrepancy could be explained by effects of selection that follow the generation process, which include functionality checks and central tolerance. To test this, we trained a selection model on top of P_{gen}, but this time on productive clonotypes, using the SONIA package [20]. Mathematically, the probability of a clonotype σ takes the form $P_{\text{post}}(\sigma) = Q(\sigma)P_{\text{gen}}(\sigma)$, where $Q(\sigma)$ is a sequence-dependent selection factor made up of a product of individual amino-acid choices at each position along the CDR3. $P_{\text{post}}(\sigma)$ characterizes a given CDR3 amino acid sequence not only by its probability of being stochastically generated, but also of having passed selection and belonging to the functional repertoire. The prediction of this model for the sharing spectrum shows excellent agreement with the data (Fig. 2A, red curve), implying that a combination of convergent recombination and functional selection is enough to explain the overlap between repertoires.

We extended this approach to the analysis of memory IGH repertoires, as captured by the IgG population composed of antigen-experienced cells (Fig. 2B). The level of sharing observed for these repertoires is even higher than for IgM (Fig. 2C). This may be a possible signature of the selection phase occurring during the transition from IgM to IgG, in which high-affinity B cells are positively selected in the germinal center [21]. To apply the same method as to IgM populations, we reconstructed the set of putative ancestral sequences of the IgG receptors. These ancestral sequences correspond to the unmutated sequences that cells carried prior to affinity maturation. Since they are rid of hypermutations, they are similar to naive sequences from inexperienced cells. However, they differ from them because they have passed an additional selection step, through their activation and initial recruitment to germinal centers. In practice, ancestral inference was done by dividing all nucleotide IGH sequences into clonal families, and by reconstructing the most likely naive progenitor of each family (see Methods). We then computed the sharing spectrum of the ancestral clonotypes across the same 10 donors (Fig. 2B, blue crosses). We learned an IgG-specific selection model P_{post} from the ancestral clonotypes, on top of the generative model learned above. Similarly to IgM, the model predicted sharing with high accuracy (Fig. 2B, red curve), while using the generation model P_{gen} alone underestimated it (green curve). Our statistical approach is thus able to capture the convergent selection not only of naive clonotypes, but also of memory precursor clonotypes.

Sharing in TCR repertoires depends on age and CMV status

Next, we applied our framework to TCR repertoires. We trained selection models from the combined productive TCRβ repertoire of 666 healthy donors, totalling $\approx 1.4 \times 10^7$ unique nucleotide sequences [16], on top of a $P_{\text{gen}}(\sigma)$ generation model trained on the nonproductive sequences from the same cohort. We tested two selection models: a SONIA model as before, and a more advanced model based on artificial neural networks implemented in the soNNia package [18]; training was done on a subsample of 10^6 amino acid sequences. By contrast to B cells, not only the generation model but also both models of selection substantially underestimated sharing (Fig. 2D, green and red curves). This discrepancy suggests that, while the selection model predicts clonotype frequency

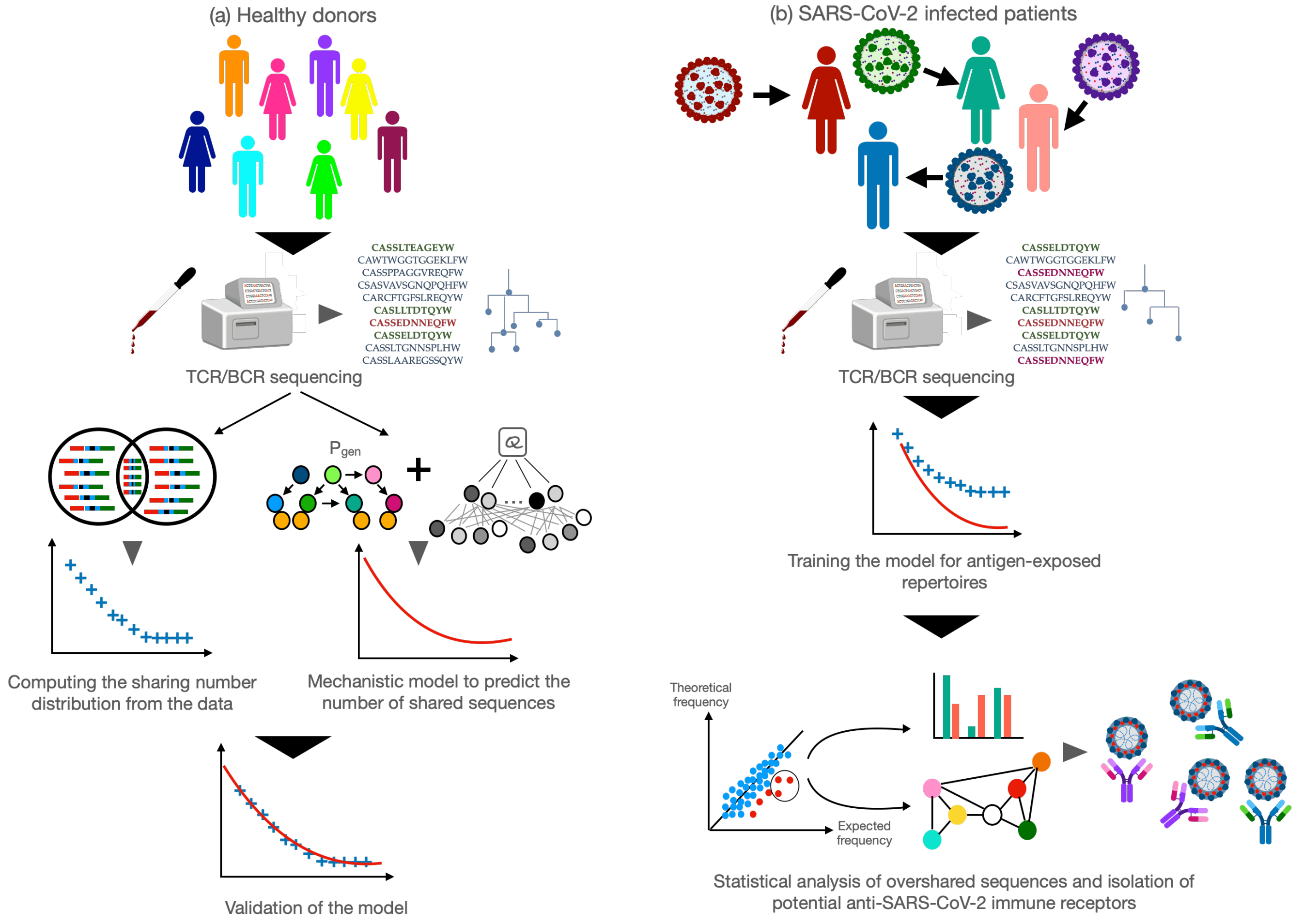

FIG. 1: **Schematic of the analysis pipeline.** (a) The BCR or TCR repertoires of a cohort of healthy donors are sequenced from blood samples. Raw reads are processed by removing duplicates and extracting the CDR3. For a given cohort, the distribution of sharing numbers m is obtained by counting the number of sequences that are found in exactly m individuals. In parallel, sequences are used to train a probabilistic generative model, in a two-step process: first learning the recombination model $P_{\rm gen}$ from nonproductive sequences; second learning a selection model Q to describe productive sequences. The model is validated by comparing its prediction for the distribution of sharing numbers with the data. (b) The repertoires of individuals currently or recently infected with SARS-CoV-2 are collected and processed as described for healthy donors. The distribution of sharing numbers is compared with the model prediction obtained previously, this time showing departures from the model, due to the enriched sharing of public COVID-associated clonotypes. These clonotypes are identified as those that are significantly more frequent in the cohort than predicted by the theory. The sequence features and organization of COVID-associated clonotypes is then analyzed and validated against databases of receptors with known SARS-CoV-2 specificity.

accurately [18], it misses important features of thymic selection, e.g. its dependence on HLA types which is not included in the model.

To overcome this issue, we introduce an additional step of convergent selection: we assume that only a fraction q of sequences sampled from $P_{\rm post}(\sigma)$ are selected. This fraction is picked at random but fixed once for all donors. The single parameter q is learned from fitting the relationship between the number of unique clonotypes and the number of unique nucleotide sequences, which depends on q but is not directly related to the sharing spectrum [5] (see Methods). Fitted values of q close to 1 mean that the selection model $P_{\rm post}$ captures convergent selection well, while smaller values of q imply that additional selection is at work. Accordingly, we found increasing values of q as the model gets more accurate: $q = 0.037$ for $P_{\rm gen}$ [5], $q = 0.072$ for SONIA, and $q = 0.472$ for soNNIa. The sharing predictions using this last value of q for soNNia is excellent (Fig. 2D, orange curve).

Ageing in TCR repertoires

More than a correction factor, q may be interpreted as the stringency of selection. The smaller is the q value, the less diverse will be our repertoire, and the more sharing

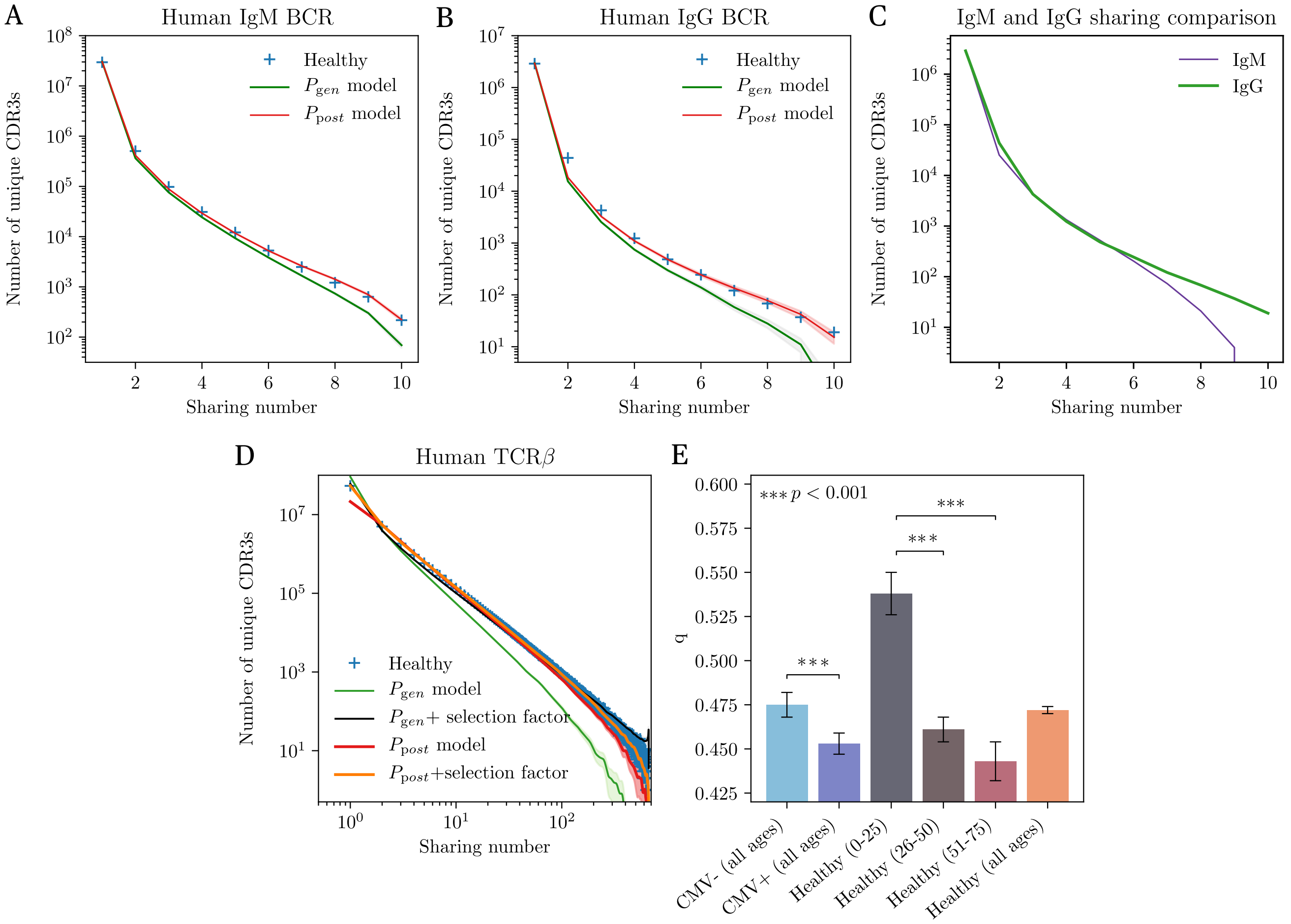

FIG. 2: **Sharing of BCR and TCR repertoires of healthy individuals.** (A) Distribution of the sharing number (the number of individuals in which a sequence is seen) of CDR3 amino-acid sequences of the heavy chains of IgM repertoires from 10 individuals. The prediction form the raw recombination model (P_{gen}, green line) underestimates sharing. Adding an *ad hoc* correction factor assuming a fraction q of sequences passing selection ($q = 0.759 \pm 0.001$) gives a good fit (not shown). The prediction from the generation and selection model (P_{post}, red line) reproduces the curve perfectly, with no need for a correction factor. (B) Distrubion of sharing number for the IgG repertoires of the same donors. The analysis is done on the naive ancestors of recontructed clonal lineages. The P_{gen} model is again inaccurate, requiring a correction factor $q = 0.490 \pm 0.005$, while the P_{post} works well. (C) Comparison of the sharing number distribution between two equal-size cohorts of IgM and IgG repertoires. IgG repertoires present a higher level of sharing, suggesting stronger convergent selection. (D) Distribution of the sharing number of CDR3 amino-acid sequences of TCRβ from 666 patients. Models predictions are shown for P_{gen} and P_{post}, with or without a correction factor q. The correction factor is $q = 0.037$ for P_{gen}, and $q = 0.472 \pm 0.002$ for P_{post}, indicating a better accuracy of the latter. In addition, the corrected P_{gen} model (black line) overestimates the number of sequences shared by all individuals relative to the data and to the corrected P_{post} prediction (orange line). (E) Value of the corrective factor q, intepreted as the inverse strength of convergent selection ($q = 1$: no selection; $q \ll 1$: strong selection), for different subgroups of the TCRβ cohort. We observe stronger selection in CMV positive individuals than in CMV negative ones, reflecting their common antigenic exposure. Convergent selection also substantially increases with age. Significance obtained with student's t test.

we expect. We used this interpretation to study repertoire evolution with age. Ageing is believed to be accompanied by a gradual loss of diversity in T cell receptor repertoires [22–24]. This phenomenon is due to a progressive deterioration of thymus functionality [25, 26] together with an expansion of memory T cell clones [27, 28] generated by former encounters with antigens. To study the influence of age on selection, we split the cohort from [16] into three ages subclasses (0-25, 26-50 and 51-75 years old), and inferred a different q factor in each group: $q_{0-25} = 0.538 \pm 0.012$, $q_{26-50} = 0.461 \pm 0.007$ and $q_{51-75} = 0.443 \pm 0.011$. The 0-25 group had a significantly less stringent selection factor than the other two age groups, consistent with repertoires becoming more and more convergent with age (Fig. 2E, grey, brown and maroon bars). We further checked that this effect was not due to the increased proportion with the age of donors who tested positive for Cytomegalovirus (CMV). CMV positive donors share more sequences than CMV negative donors (light and dark blue bars), because of enhanced convergent selection for CMV specificity among HLA-matched donors [16]. To control for this confounding effect, we used the Fisher-Yates shuffle algorithm to create 3 control groups with no age structure, but with the same respective proportions of CMV+ donors as in the original data. Applying the same pipeline to the control groups yielded statistically indistinguishable values of q, meaning that age and not CMV status is the main driver to convergent selection.

Convergent BCR sharing in COVID-19 donors

Having validated the predictive power of the SONIA model for BCR sharing in healthy donors, we asked whether it was still predictive for repertoires stimulated by a common disease such as COVID-19. We studied the IgG heavy-chain repertoires of 44 SARS-CoV-2-positive individuals (as confirmed by RT-qPCR test), obtained from three independent studies [29–31]. Fig. 3A shows that the model trained on healthy patients substantially underestimates sharing between the IgG repertoires of individuals with COVID-19. This enrichment of shared sequences among COVID-19 donors is strongly suggestive of convergent selection for sequences that have common SARS-CoV-2 epitope specificity.

To identify the putative clonotypes with SARS-CoV-2 specificity that are responsible for oversharing, we first estimated, for each shared clonotype σ, its frequency $P^{*}_{\rm data}(\sigma)$ in the COVID-19 cohort. This estimate is based on the number of donors in which it was found as well as its multiplicity (synonymous nucleotide variants) in each donor (see Methods); this information is gathered in the occurence vector $\vec{x} = x^{\sigma}_1, x^{\sigma}_2, ..., x^{\sigma}_M$ giving the number of nucleotide sequence variants coding for σ in each individual $1, \ldots, M$. We expect that number to generally correlate with the model prediction derived from healthy donors, $P_{\rm post}(\sigma)$, as shown in Fig. 3B. However, we also expect COVID-19 specific clonotypes to be overrepresented in $P_{\rm data}$ versus $P_{\rm post}$, and fall in the lower-right part of Fig. 3B. To identify those clonotypes, we first applied a constant multiplicative factor to $P_{\rm post}$ to correct for the systematic underestimate of $P_{\rm data}$ (see Methods). Next we call “COVID-associated” clonotypes whose estimate of $P_{\rm data}(\sigma)$ was significantly higher than the rescaled $P_{\rm post}$, with a false discovery rate of 10^{-4} (using Bayesian analysis, see Methods)

This procedure identified 6650 COVID-associated clonotypes. To validate their COVID specificity, we cross-checked this list against a database of antibodies with reported specificity for SARS-CoV-2, which we built from the literature (see Supplementary Material) [32–46]. Among them, 175 sequences ($\approx 2.6\%$) had at least a 90% Levenshtein similarity (one or two mismatches or gaps) in amino acid content with the CDR3 region of reported SARS-CoV-2 binding antibodies. To better understand the significance of this overlap, we carried out two control analyses. First, we followed the same pipeline to identify significantly overshared sequences in the IgG healthy cohort, and looked for sequence similarity of these overshared sequences in the assembled SARS-CoV-2 antibody database. We found no similar sequences, confirming the intuition that one should not find SARS-CoV-2 specific antibodies in the repertoires of healthy people sampled before the pandemic. Second, we constructed a mock antibody database of the same size as the original one, by taking random samples from the pool of healthy IgG sequences. Sequence matching among significant sequences in COVID-19 repertoires and this control database yielded an overlap of 2.7 ± 1.5 sequences that are a 90% similar by Levenshtein similarity. This result is also coherent since we don't expect all overshared sequences to be just due to SARS-CoV-2 infection.

We then asked whether the COVID-associated antibodies differed in the CDR3 sequence features from other antibodies found in healthy or COVID repertoires. In Fig. 3C and D show the comparison of V and J gene usage between COVID-associated and bulk repertoires. We observe a very diverse usage of V and J genes in COVID-associated sequences. While no V or J gene stands out as being COVID-specific, there are a few remarkable differences, such as the enrichment of IGHV3-30-3, IGHV4-31, IGHV1-24 and IGHV4-61, IGHJ6 in COVID-associated sequences, whereas IGHV3-23 and IGHJ4 are underrepresented. The CDR3 length distribution (Fig. 3E) shows a small but substantial bias towards longer sequences in COVID-19 repertoires (orange curve) compared to healthy repertoires (green curve), consistent with previous findings [29–31, 40].

We also examined the hypermutation rates in the three groups. COVID-associated sequences had lower hypermutation rates than healthy repertoires (Fig. S1). Since repertoires were taken during the infection, these B cells have expanded quickly but may not have had the time to accumulate mutations compared to mature memory B cells. However, decent neutralization power can be achieved without long affinity maturation [37], a feature already observed in immune responses to other respiratory viruses [47].

Finally, we wondered whether the degree of convergent selection was predictive of disease severity. To answer this question, we focused on the data from [30], which contains information about severity (mild, moderate, and severe) for each individual. To quantify convergence, we defined a selection factor $q_{\rm AB}$ for each pair of individuals A and B as the correction needed to exactly predict their repertoire overlap using the $P_{\rm post}$ model as a baseline. The lower $q_{\rm AB}$, the more convergent the repertoires of A and B are. The resulting pairwise selection factors

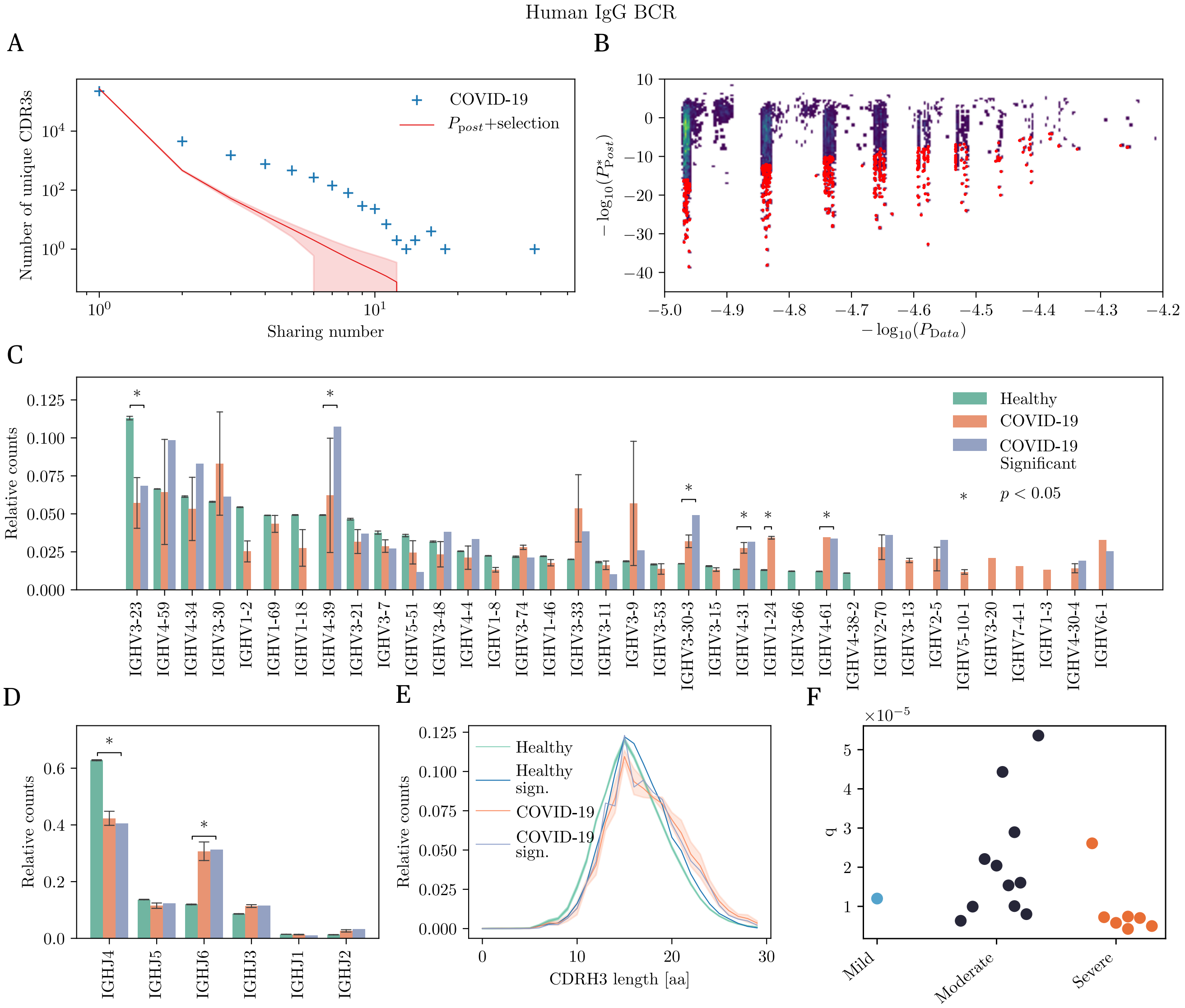

FIG. 3: **Identification and analysis of COVID-associated antibody heavy chains from significantly shared sequences.** (A) Sharing number distribution of IgG heavy chain CDR3 from SARS-CoV-2 positive individuals. This distribution is compared with the sharing expectation in healthy individuals, obtained using the P_{post} model. The discrepancy suggests the enrichment of these repertoires in SARS-CoV-2 specific antibodies. (B) Model prediction $P_{\text{post}}(\sigma)$ *vs.* empirical probability $P^*_{\text{data}}(\sigma)$ of all shared sequences. Red sequences are substantially more frequent in the data than predicted (posterior probability $< 10^{-4}$), and are predicted to be associated with SARS-CoV-2. (C) IGHV gene usage in healthy and COVID-19 repertoires, and among COVID-associated sequences. (D) IGHJ gene usage for the same groups (same color code). (E) CDRH3 length distribution averaged over individuals within each cohort. A control was added for sequences found to be significantly more shared than expected in healthy individuals, to control for the bias that shared sequences tend to have a shorter CDR3. (F) Convergent selection factor q learned from pairs of COVID-19 individuals in different severity group. Individuals with more severe symptoms seem to have a higher level of selection, although the difference is not statistically significant (Student's t test ($p = 0.12$)

(Fig. 3F) show increased convergent selection among individuals with severe forms of the disease, although the difference is not statistically significant due to the small number of donors.

Network analysis of COVID-specific repertoires

The CDR3 region is the major determinant of binding specificity of both BCR and TCR [48, 49]. We ex-

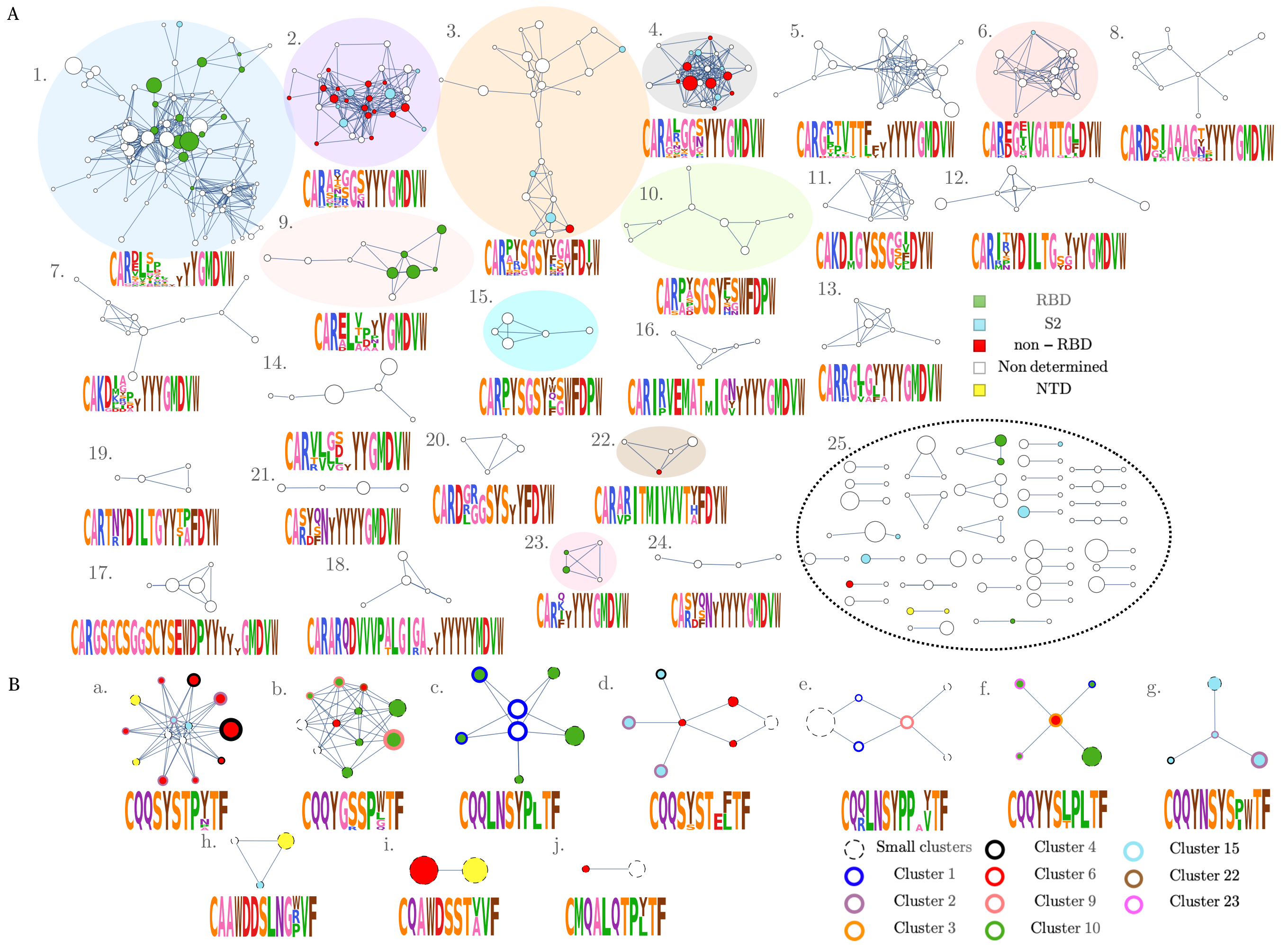

FIG. 4: **Networks analysis of IgG heavy chains associated with SARS-CoV-2.** (A) Each node represents a CDR3aa heavy chain clonotype. Edges connect clonotypes with two or fewer amino acid mismatches in their CDR3 region, and with the same V and J segments. Non-connected vertices are not shown. Colored nodes represent sequences that are at most one amino acid mismatch away from previously reported SARS-CoV-2 neutralizing antibodies, with the color indicating the region of the virus recognized by the antibody. (B) Network of light chains CDR3 amino acid sequences found in previous reports to be paired with a colored heavy chain of (A). The level of conservation is even more remarkable than for heavy chains.

pect that upon infection, repertoires will be enriched with CDR3 clonotypes with similar sequences recognizing predominant epitopes [50–54]. We looked for motifs among the COVID-associated clonotypes through a sequence-based network analysis. Such clustering analysis was previously used to visualize repertoires from individuals suffering from chronic diseases [55], giving a useful graphical representation of the structural differences among clonally expanded and healthy repertoires. We calculated Levenshtein distances between all pairs of CDR3 clonotypes and V, J gene identity. Clonotypes are considered adjacent if their distance is ≤ 2. We clustered clonotypes using single-linkage clustering from this adjacency matrix. The resulting cluster are shown in Fig. 4A, where each node is a clonotoype, and adjacent clonotypes are linked by edges. Sequence logos [56] under each cluster show that CDR3 sequences are well conserved within each group.

To assess the robustness and significance of the clusters, we designed a control cloneset obtained by pooling together the bulk IGH repertoires from all 44 COVID-19 patients. We built 5 random subsets of the same size as the COVID-associated subset, and repeated the network analysis on each. Note that while specificity clusters are still expected in bulk repertoires, they are likely to stand out less than in specific repertoires, where diversity is reduced and sequence similarity higher. Consistent with this, only 3.6±0.2% sequences of the control subsets belonged to non-singleton clusters, compared to 8.6% for the COVID-associated clonotypes. This implies that the network analysis captures sequence convergence of functional origin.

One hypothesis is that each cluster corresponds to distinct specificity group targeting a given epitope associated to SARS-CoV-2. To test this, in Fig. 4A we marked clonotypes found to be close to previously reported SARS-CoV-2 antibodies with a color that depends on which antigenic region of SARS-CoV-2 they recognize. In the largest cluster (labeled 1 in the figure), 13 out of 97 sequences (14%, versus $0.1 \pm 0.3\%$ for random clonotypes) are just one amino-acid away from antibodies recognizing the receptor binding site (RBD) of the S protein. Entry into human cell is initiated by interaction between the RBD and the cell surface receptor ACE2 [57–61] making RBD an immunodominant target for neutralizing antibodies [62]. Two clusters (3 and 4) include matches with antibodies interacting with other epitopes than RBD (non-RBD and S2 domain, which contain the fusion peptide). V and J gene usage is well conserved across clusters 2, 4 and 5 (IGHV3-33, IGHV3-30, IGHV3-30-3 and IGHJ6), while central positions of the CDR3s are enriched in different residues, consistent their distinct specificies.

We then asked whether there was an analogous structural selection acting on light chains. While we do not have light-chain repertoire datasets from COVID-19 donors, we can collect the light chains of all SARS-CoV-2-specific antibodies from the literature that matched (up to 2 aminoacid substitutions) our COVID-associated clonotypes in their CDR3. Applying the same clustering procedure as described above yields 10 different clusters, shown in Fig. 4B), with the color of the circle outline coding for the cluster membership of the heavy chain of that clone in A. For example, cluster (c) is exclusively associated to anti-RBD antibodies, and matches heavy-chain cluster (1). Cluster (g) represents a clear antigen-specific cluster with all components being antibodies acting on S2, but does not have a clear analog among heavy chain clusters, suggesting that specificity is primarily determined by the light chain. For light chains, the level of redundancy observed was even greater than for heavy chains. This was in part due to the fact that some SARS-CoV-2 confirmed antibodies from the literature differ in their CDR3 heavy chain, but have identical light chains.

Predicting sharing in T cell repertoires from COVID-19 patients

T cells play an important role in COVID-19 infections, and can persist for years even when IgG antibodies have become undetectable [63, 64]. We applied our sharing analysis to the TCRβ repertoires of 1414 donors with confirmed SARS-CoV-2, sampled at various timepoints following the peak of the disease, and totalling $1.6 \cdot 10^8$ reads [65].

Fig. 5A shows the sharing spectrum for all 1414 donors. We recall that, by contrast to IGH, a corrective selection factor q was needed to allow the statistical model P_{post} to correctly predict the sharing spectrum, even for healthy donors (Fig. 2C). This selection factor may be interpreted as quantifying the level of convergent selection among donors, as demonstrated for CMV+ donors (Fig. 2D). Inferring q on the COVID-19 cohort using the soNNia model yielded 0.452 ± 0.002, compared to $q = 0.472 \pm 0.006$ for the healthy cohort, suggesting that COVID-19 donors shared more TCR among them due their common infection. The corresponding prediction for the sharing spectrum is excellent (5A, orange curve).

While convergent selection is less evident than for B cell repertoires, highly clonal T-cell repertoires have been reported by several studies [40, 66]. Inspired by these results, we used the same approach as for B cells for identifying COVID-associated TCR. In Fig. 5B, we compare the soNNia-inferred probabilities $P_{\text{post}}(\sigma)$ with the empirical frequencies of the receptor sequences $P_{\text{data}}(\sigma)$. The potential responding clones marked in red were identified as those whose empirical frequency was unexpectly high relative to the model prediction, as described before for B cells. To focus on clonotypes with the largest significance, we picked the 0.1% clonotypes with the highest posterior probability of being COVID-associated, resulting in a list of 20,841 sequences. To validate the specificity of this list, we compared it to a list of clonotypes with know specificity to SARS-CoV-2 epitopes, as confirmed by a Multiplex Identification of Antigen-Specific (MIRA) assay [65], and found 3,722 (17%) overlapping clonotypes. By contrast, random lists of 20,841 sequences from healthy patients and from COVID-19 associated datasets (drawn from a biased distribution to reflect their sharing properties, see Methods) only shared 27 ± 5 and 193 ± 13 clonotypes with the MIRA assay, respectively. These much lower quantities demonstrate not only that the cohort of significantly shared sequences is enriched of SARS-CoV-2 specific clones but also indicates preexisting cross-reactive T cells recognizing SARS- CoV-2 epitope in previously unexposed individuals, probably related to other seasonal coronaviruses [66–68].

The analysis of sequence features revealed a slight increase of CDR3 length in COVID-associated versus generic clonotypes (Fig. 5D). A few V and J genes were enriched or depleted in COVID-associated TCR (Fig. 5E and F): TRBV5-1, TRBV20-1 and TRBV6-2, TRBJ1-1 are considerably underrepresented in the COVID-associated repertoire, while TRBV27, TRBJ2-1 and TRBJ1-2 are enriched, consistent with previous reports [40, 69].

To better understand the structure of the COVID-specific TCR repertoire, we next performed a network and clustering analysis of COVID-associated clonotypes. As before, we built a graph by putting edges between clonotypes with at most two differences in CDR3 amino-acid identity or in V and J gene usage. Clonotypes found in the MIRA dataset of [65] are colored according to their antigen specificity: yellow for the surface glycoprotein (S), red for the nucleocapsid phosphoprotein (N), brown for the transmembrane protein (M) and orange for the envelope protein (E); and the seven putative ac-

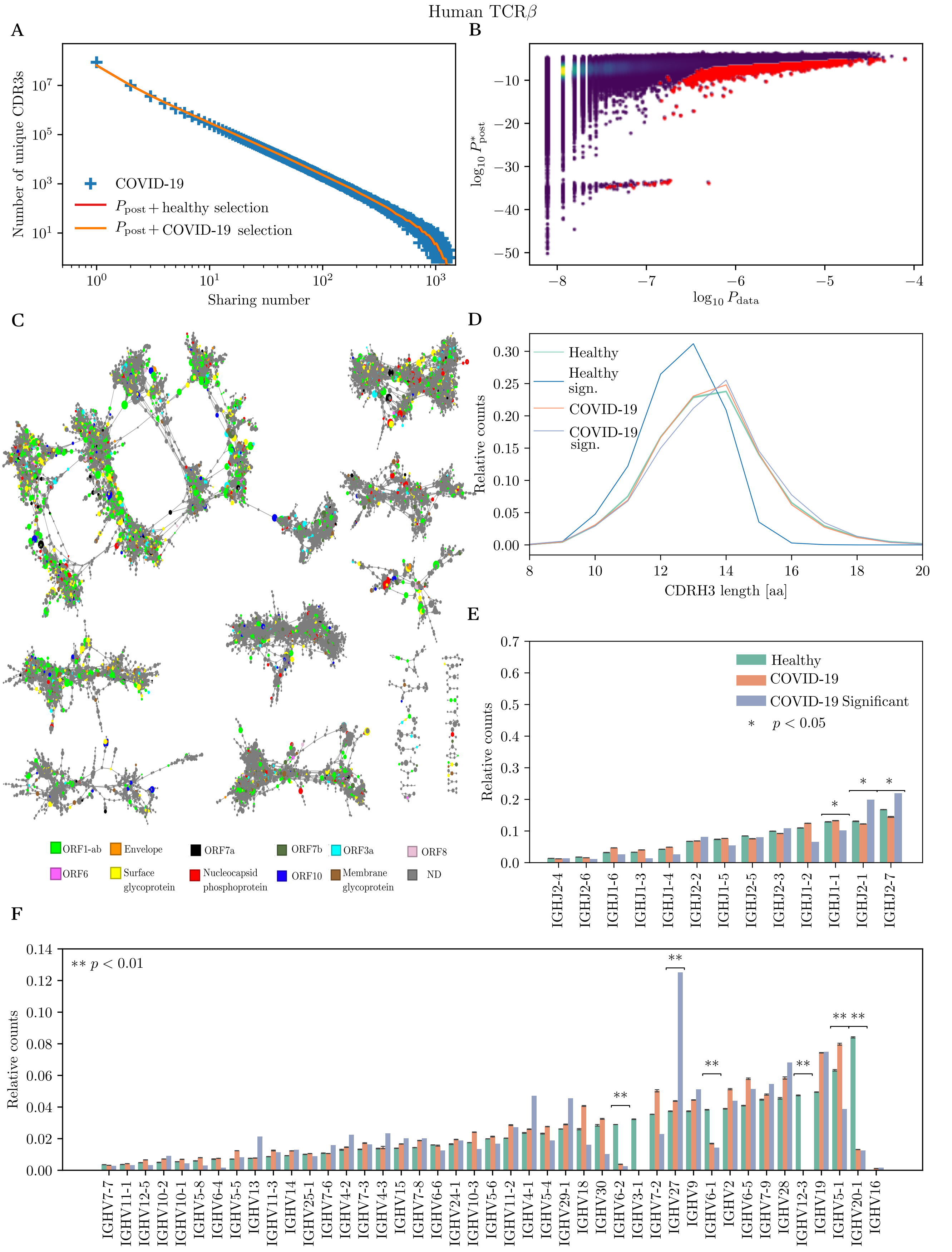
Human TCRβ
A
Number of unique CDR3s
Sharing number
COVID-19
$P_{\rm post}$ + healthy selection
$P_{\rm post}$ + COVID-19 selection
B
$\log_{10} P^*_{\rm post}$
$\log_{10} P_{\rm data}$
C
ORF1-ab
Envelope
ORF7a
ORF7b
ORF3a
ORF8
ORF6
Surface glycoprotein
Nucleocapsid phosphoprotein
ORF10
Membrane glycoprotein
ND
D
Relative counts
CDRH3 length [aa]
Healthy
Healthy sign.
COVID-19
COVID-19 sign.
E
Relative counts
Healthy
COVID-19
COVID-19 Significant
* $p < 0.05$
IGHJ2-4
IGHJ2-6
IGHJ1-6
IGHJ1-3
IGHJ1-4
IGHJ2-2
IGHJ1-5
IGHJ2-5
IGHJ2-3
IGHJ1-2
IGHJ1-1
IGHJ2-1
IGHJ2-7
F
Relative counts
** $p < 0.01$
IGHV7-7
IGHV11-1
IGHV12-5
IGHV10-2
IGHV10-1
IGHV5-8
IGHV6-4
IGHV5-5
IGHV13
IGHV11-3
IGHV14
IGHV25-1
IGHV7-6
IGHV4-2
IGHV7-3
IGHV4-3
IGHV15
IGHV7-8
IGHV6-6
IGHV24-1
IGHV10-3
IGHV5-6
IGHV11-2
IGHV4-1
IGHV5-4
IGHV29-1
IGHV18
IGHV30
IGHV6-2
IGHV3-1
IGHV7-2
IGHV27
IGHV9
IGHV6-1
IGHV2
IGHV6-5
IGHV7-9
IGHV28
IGHV12-3
IGHV19
IGHV5-1
IGHV20-1
IGHV16

FIG. 5: **Identification and analysis COVID-associated TCRβ.** (A) Sharing number distribution of CDR3 amino-acid TCRβ clonotypes from 1414 SARS-CoV-2 positive individuals. The model prediction from P_{post} (trained on healthy donors) is good, suggesting low convergent selection. (B) Model prediction P_{post} versus empirical frequency P^*_{data} for all shared CDR3 amino-acid clonotypes. Clonotypes marked in red are significantly more frequent in COVID-19 donors than expected (C) Network analysis of TCRβ amino acid CDR3s of COVID-associated clonotypes. Edge mark clonotypes with 2 mismatches or fewer. Vertices are colored according to the location of their recognized antigen according to the MIRA assay (see main text). (D) CDRH3 length distributions of generic and significantly overshared sequences. Sequences significantly overshared in healthy individuals are much shorter than sequences significantly overshared in COVID-19 individuals. (E) TRBJ gene usage in generic healthy and COVID repertoires, and in COVID-associated sequences. (F) TRBV gene usage for the same groups. Significance is obtained using Student's t test.

cessory proteins (ORF1ab (light green), ORF3a (light blue), ORF6 (magenta), ORF7a (black), ORF7b (dark green), ORF8 (light pink), and ORF10 (dark blue). The resulting network and clusters (Fig. 5G) paint a more complex picture than for B cells, with no clear sign of grouping by specificity. However, the level of clustering is significant: repeating the same graph analysis of the control set of clonotypes from healthy patients (corrected for their sharing properties as above, see Methods) yields a much lower level of clustering: $12.87 \pm 0.13\%$, versus 88% for the COVID-associated clonotypes. This means that, while clusters are not able to discriminate between different antigenic targets, selection for antigenic specificity still drives a local convergence of clonotypes.

Using public sequences for SARS-CoV-2 infection diagnosis

The public COVID-associated clonotypes identified above can be used as a bio-marker to detect exposure to current and past virus infections. Within our probabilistic framework, we developed a likelihood-ratio classification test based on comparing the likelihoods of occurrence of COVID-associated clonotypes in unexposed versus exposed individuals. These likelihoods were computed using $P_{\text{post}}(\sigma)$ for unexposed individuals, and the empirical clonotype distribution $P_{\text{data}}(\sigma)$ from COVID-19 donors for exposed individuals (see Methods for details).

We first applied this strategy to repertoires of IgG ancestors. We started from the previously identified list of 6650 clonotypes significantly overshared in COVID-19 patients. $P_{\text{data}}(\sigma)$ was estimated empirically from the repertoires of 20 COVID-19 donors reserved for training.

Plotting the distributions of log likelihood ratio scores in SARS-CoV-2 positive and negative individuals held out for testing (20 positive and 10 negative) shows a clear separation of the two groups across a threshold value of 0, indicating perfect discrimination by this score for this small cohort (Fig. 6A). This result implies that this score could be used for diagnostics of COVID-19.

We next applied the same method to TCRβ repertoires, using 700 SARS-CoV-2 positive individuals to train P_{data}, and keeping an equivalent number for testing. The raw log likelihood score using the 1,200 top COVID-associated clonotypes yielded poor discrimination power (Fig. S2). To improve performance, we trained a logistic regression classifier predicting COVID-19 status based on the presence or absence of top COVID-associated clonotypes as predictors. The weights of the model were learned using the same training set, and performance was evaluated on the remaining donors as testing set. Classification error was optimized by tuning the number of included clonotypes and regularization parameter to avoid overfitting (see Methods for details). The distribution of logistic regression scores (expressed as the probability of being positive) among the positive and negative cohorts of the testing set are shown in Fig. 6B. Setting a threshold at 0.5 gives 83% specificity and 92% precision. Varying the threshold yields the Receiver Operating Curve shown in Fig. 6C, which summarizes the classification performance for both methods (logistic and likelihood-ratio test). The performance of our method is lower than that reported in [65], where a larger cohort was used, and where some clonotypes leading to classification errors were filtered out by hand.

III. DISCUSSION

Our statistical model allowed us to characterize in detail how the generation and initial selection of repertoires could explain the observed spectrum of sharing among unrelated healthy donors. The methodology relies on two ingredients: a stochastic recombination model, $P_{\text{gen}}(\sigma)$, and a selection model, $P_{\text{post}} = Q(\sigma) \times P_{\text{gen}}(\sigma)$, where $Q(\sigma)$ is a sequence-specific selection factor. A previous attempt at predicting sharing in TCR [5] only considered P_{gen}. This model alone underestimates sharing, but applying a single corrective selection factor $0 < q \leq 1$ yielded good predictions. Our analysis shows that learning a proper selection model $Q(\sigma)$ from data allows for excellent predictions for BCR sharing, without the need for such a corrective factor. This suggests that for B cells, P_{post} accurately captures selection effects encoded in peripheral B cells, including central tolerance and affinity maturation. This success is surprising because we could have expected central tolerance to remove auto-reactive B cells, thereby reducing diversity and increasing sharing. This effect would be hard to capture by a linear models such as SONIA, which we used for $Q(\sigma)$.

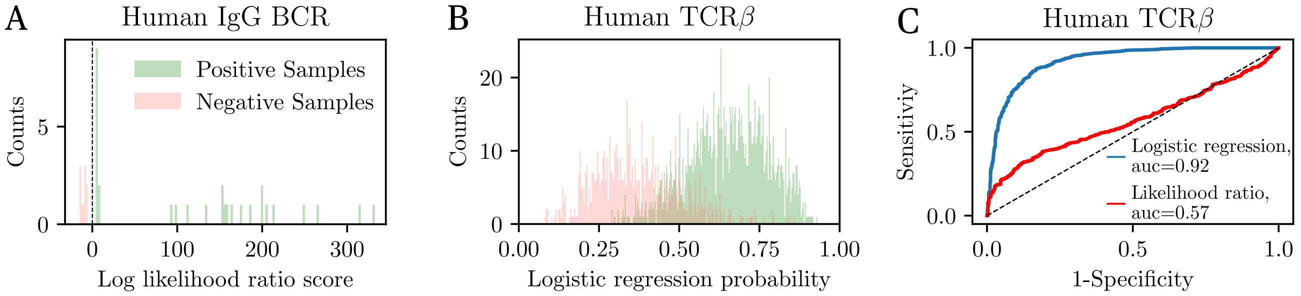

FIG. 6: **Repertoire-based SARS-CoV-2 diagnostics**. (A) Distribution of log likelihood ratios calculated from the IgG repertoires of 30 individuals (20 SARS-CoV-2 positive, and 10 negative). Positive values of the score imply likely SARS-CoV-2 positivity. The test perfectly separates positive and negative individuals. (B) Distribution of probabilities of SARS-CoV-2 positivity obtained from logistic regression on the presence or absence of clonotypes in TCRβ repertoires (1,000 individuals from the testing set; model trained on 1,000 individuals from the training set). Values over 1/2 indicate likely SARS-CoV-2 positivity. The distributions of healthy and COVID-19 individuals have little overlap (93% specificity). (C) ROC curve obtained by tuning the positivity threshold in (B). The AUC is the area under the curve (AUC=1: perfect discrimination; 1/2: no discrimination). ROC of the likelihood test is shown for comparison.

For T cells, the picture is more nuanced. Using P_{post} yielded much better predictions than P_{gen} alone, but a corrective factor q was still needed to get a perfect fit. The value of q is indicative of the level of selection not captured by the model but still necessary to explain sharing. Its increase from $q = 0.037$ when using P_{gen}, to $q = 0.472$ when using P_{post}, is a strong signature of model improvement. A possible hypothesis for why q is still needed is that negative thymic selection reduces diversity and increases sharing. Linear or even smooth non-linear models such as soNNia may not be able to capture such effects. Thymic selection also strongly depends on the HLA type, which determines what peptides may be presented to T-cells during selection, and this information is specific to each individual. One needs to go beyond universal models to make personnalized predictions of clonotype prevalence and sharing.

SARS-CoV-2 infection substantially affects the antibody repertoire [70–72] — around 0.15-0.8% of total memory B cells are spike-specific according to [73]. We expect to observe a pool of *overshared* BCR clonotypes among COVID-19 donors, corresponding to a convergent response. Our pipeline identifies such overshared, antigen-specific receptors. Since these antibodies are both COVID-specific and found in many donors, and they are likely to be elicited in many patients. Such identification of universal antibodies could help in the design of next-generation vaccines, or therapies based on neutralizing antibodies [32, 74–76].

Out of the 236,079 IgG ancestral clonotypes across all 43 COVID-19 donors collected from the literature, 20,113 (8.5%) were shared among at least two individuals. Within that pool, 6,650 were found to be significantly overshared, among which 175 had at least a 90% Levenshtein similarity with previously reported anti S-protein antibodies. The most shared IgG ancestral clonotype (CARGFDYW) was linked to the Kawasaki syndrome, an inflammatory disease that has been associated with COVID-19 [77]. Among significantly shared sequences, we also matched some clonotypes with previously described antibodies, such as the CV30 (CARDLDVSGGMDVW, IGHV3-53) in [43], the COVA2-07 (CAREAYGMDVW, IGHV3-53) in [38] and the COV2-2381 (CAAPYCSRTSCHDAFDIW, IGHV1-58) in [58]. All of them were reported to reach complete neutralization through competition with the ACE2 human protein for RBD binding, which is the main mechanism used by anti SARS-CoV-2 antibodies. Because RBD is likely to accumulate escape mutations, it may be useful to identify antibodies targeting more conserved epitopes of the S protein. The high similarity of the S2 subdomains between SARS-CoV-2 and SARS-CoV [58] suggests that targetting conserved epitopes may also allow for cross-neutralizing different coronaviruses [78].

Despite a common focus on humoral immunity against COVID-19, T cells play an essential role. The analysis of immunity in recovered SARS-CoV and MERS patients showed that the T-cell response is often more durable, remaining detectable up to 17 years after infection, even when antibodies are undetectable [79]. We applied the same pipeline for identifying overshared TCR clonotypes in a cohort of 1414 COVID-19 donors. While the level of convergent selection was much less striking than for BCR, the method was still able to detect a subset of 20,841 significantly shared sequences, 17 % of which were separately validated for SARS-CoV-2 specificity. We hypothesized several reasons for the absence of convergence in T-cell immune response. First, genetic differences in the HLA genes are known to influence the composition of TCR repertoires which can partially modify the individual response to pathogens [80]. The inclusion of HLA type in future models could improve the search and validation on antigen-specific clonotypes.

Our results are consistent with IgG repertoire diversity being reduced in patients with severe diagnostic, suggesting that this parameter could help predict the evolution of the disease [81, 82]. Relatedly, we used our selection parameter q to measure the difference in repertoire diversity of healthy donors from three age groups. We found significant differences between the youngest (0-25 y.o.) and the oldest (51-75 y.o.) healthy groups, implying greater T-cell diversity in younger individuals. It has been speculated that efficient protection against COVID-19 in children may arise from their highly diverse T-cell repertoire, while older adults may be at higher risk due to immunosenescence [83]. Further statistical analyses to study the relationship between age, repertoire diversity, and severity could help better predict disease evolution.

In this context of condition-associated sequence analysis, previous studies [65, 66, 84] yielded the identification of pre-existing SARS-CoV-2-reactive TCR clones in unexposed individuals. The role that they might play in asymptomatic or mild COVID-19 cases remains unclear. However, they could serve as a basis for thinking about vaccines elliciting a broader immune response. Since the responding B and T cell clones identified by this method repeatedly appear in different individuals, they constitute an interesting vaccine target since they could trigger a robust public response.

The clinical potential of this method also extends to the development of a B-cell or T-cell repertoire-based diagnostic tool to identify current or past infection from sequenced repertoires. The likelihood ratio test we propose here achieves perfect classification between SARS-CoV-2 positive and negative donors using the repertoires of IgG ancestors. However, the size of the testing cohort is very small (30 individuals), and the method should be validated on larger cohorts for clinical applications.

We speculate that the lower accuracy of the method applied to TCR is partly due to the HLA restriction of SARS-CoV-2 associated clonotypes, reducing their predictive power. However, we showed that the performance could be considerably improved by logistic regression, as was first proposed in [16, 65]. Predictive power might be further increased by including the HLA information of donors, and by using larger cohorts for training. Compared to the method described in [16], we used our trained generative model $P_{\text{post}}(\sigma)$ as a null model, rather than by directly recording clonotype frequencies in the healthy cohort. While this allows for estimating the probability of sequences that were never seen in the healthy cohort, its prediction may be inaccurate for certain sequences, affecting the accuracy of the score. Using a combination of probabilistic modeling for rare clonotypes and empirical frequencies for common ones may provide a intermediate method benefiting from both approaches.

Despite these challenges, developing TCR repertoire-based clinical tests is important and complementary to serological and BCR based tests, because we expect TCR repertoires to be more stable over time, and COVID-19 signatures to persist detectably in peripheral T cells longer than in antibody sera. More generally, our results are promising for the future of repertoire-based diagnostics. The method is general and applicable to any disease or condition that impacts immune repertoires. As more cohorts of donors with a common condition are screened for associated clonotypes, this approach would allow for doing several tests simultaneously from a single repertoire sample, including retrospectively as more tests are being designed.

IV. METHODS

Sequence preprocessing

IgM and IgG repertoires

The analysis of healthy BCR heavy-chain (IGH) repertoires was performed on several large datasets comprising a total of over 3×10^8 productive sequences from 10 healthy adult subjects [13]. The sequence isotype, IgM or IgG, was identified during primer removal and pairing of the two reads. We used the pre-processed data provided by the initial analysis of the dataset [13] with the exception of the two largest individuals (326651 and 326713). For these two donors we applied a custom pipeline to find high-quality out-of-frame transcripts required for inference of the generative model of V(D)J recombination. We aligned raw reads using pRESTO of the Immcantation environment [85] with a setup allowing for correcting for errors in UMIs and deal with insufficient UMI diversity. We grouped reads with the same UMI together, annotating each consensus sequence with a corresponding number of reads, which is a proxy for sequence quality. Pre-processed data were then aligned to V, D and J templates from IMGT [86] database using IgBlast [87].

We filtered the set of productive IgM sequences to keep only sequences with at most one mutation. Using a subset of non-productive, naive IgM sequences of high-quality (with at least 3 reads per consensus sequence), we used IGoR [19] to infer the statistics of the generative process, and to build a model $P_{\text{gen}}(\sigma)$ that can be used to generate synthetic sequences with no mutations, and free of selection effects that affect real productive sequences.

To include selection effects, we learned a sequence-specific selection factor $Q(\sigma)$ that modifies the probability of occurrences of sequences relative to the generation model. The resulting post-selection probability model reads: $P_{\text{post}}(\sigma) = Q(\sigma)P_{\text{gen}}(\sigma)$. The training of the function $Q(\sigma)$ was done using IgM productive sequences with at most 1 hypermutation using the SONIA software [20].

Clonal lineage reconstruction in IgG repertoires

Large fraction of the IgG repertoire has already undergone successive stages of differentiation and specialization after antigen activation, meaning that statistics describing naive IgM sequences features are no longer valid for IgG sequences. We incorporate these effects in a selection model inferred from the sequence features of the reconstructed naive progenitors of IgG sequences.

To reconstruct IgG clonal families and infer the ancestral sequences, we started by grouping together sequences with the same V and J genes and CDR3 length. Inside each group, we performed single linkage clustering of CDR3 sequences with a threshold of 90% identity [85]. Clusters obtained in this way correspond to clonal families, which we assume originate from a common naive unmutated progenitor. We reconstructed the CDR3 sequences of the naive ancestor by removing mutations from the germline segments of the sequence (using IMGT templates), and taking the consensus sequence for the non-templated segments.

An analogous procedure was used to process IgG heavy-chain antibody sequences from repertoires of 43 COVID-19 donors obtained from [29–31].

TCRβ repertoires

Healthy T cell β-chain repertoires were taken from a cohort of 666 donors [16], 641 of which were serotyped for cytomegalovirus (CMV): 352 subjects were CMV negative (CMV−) and 289 were CMV positive (CMV+), with a mean number of unique β chains $\approx 180{,}000$ per donor.

COVID-19 repertoires were obtained from a cohort of 1414 SARS-CoV-2 positive donors sequenced by Adaptive Biotechnologies, as described in [65]. Both datasets were already pre-processed.

We separated non-productive and productive sequences from healthy donors. We inferred a generative model from the nonproductive sequences using IGoR [19], and a selection model from the productive sequences using soNNia [18], an artificial neural network based extension of SONIA.

Analytic prediction of sharing from the density of probabilities

Given M repertoires sampled from M individuals, each with a number N_i of unique nucleotide sequences, $i = 1, \ldots, N$, the expected number of sequences sharing among exactly m individuals is given by the coefficients of the generating function [5]:

$$G(x, \{N_i\}) \approx \int_0^{+\infty} \rho(p) dp \prod_{i=1}^{n} \left[e^{-N_i p} + \left(1 - e^{-N_i p}\right) x \right], \tag{1}$$

where $\rho(p)dp$ is the total number of potential sequences whose probability falls between p and $p + dp$. Given a model distribution $P(\sigma)$, the integral in 1 is performed by Monte-Carlo: generate a large number of sequences from $P(\sigma)$, evaluate their probability $p = P(\sigma)$, and create a histogram of $\ln p$. Then approximate the integral over p by the method of trapezes.

The simplest model to consider is the generative model P_{gen} learned from nonproductive sequences, free of any selection. Then Eq. 1 is used with $P(\sigma) = P_{\text{gen}}(\sigma)/f$ calculated using OLGA [88], where f is the probability that the generative model produces a productive sequence (given by the IGoR model).

As a minimal to include thymic selection, and in particular negative selection that removes self-reactive receptors from the repertoire, we can introduce an *ad hoc* selection factor q representing the fraction of receptors that are not self-reactive. Then the probability distribution of non-self-reactive productive sequences is

$$P(\sigma) = \frac{P_{\text{gen}}(\sigma)}{fq}, \tag{2}$$

for a random (but fixed across individuals) fraction q of sequences, and $P(\sigma) = 0$ for all the others. This amounts to rescaling the distribution $\rho(p)$ in Eq. 1 by a factor $(fq)^{-1}$ in the p variable.

The stronger the selection (the lower q), the smaller the number of unique amino-acid sequences will be as a function of the number of unique nucleotide sequences. We can use that relation to fit the parameter q using least-square regression. The model prediction for the expected number of unique sequences among N random draws from p is given by:

$$M_0^{\text{nt}}(N, q) \approx \int_0^{+\infty} \rho(p) dp \left(1 - e^{-Np}\right), \tag{3}$$

where $\rho(p)$ is as before. We can compute that number for nucleotide and amino-acid sequences, $M_0^{\text{nt}}(N, q)$ and $M_0^{\text{aa}}(N, q)$, using the densities $\rho_{\text{nt}}(p)$ and $\rho_{\text{aa}}(p)$ corresponding to these two cases, as described before. The probability of generated sequences, both at the nucleotide and amino-acid level, is computed using OLGA [88] and Eq. 2. We then find the q that minimizes the total distance between each pair $\{M_i^{\text{nt}}, M_i^{\text{aa}}\}$ from the data to the parametric curve $M_0^{\text{nt}}(N)$ and $M_0^{\text{aa}}(N)$:

$$\begin{aligned} q^* = \arg\min_q \sum_{i=1}^{M} \min_{N_i} \Big[& \left(M_0^{\text{nt}}(N_i, q) - M_i^{\text{nt}}\right)^2 + \\ & + \left(M_0^{\text{aa}}(N_i, q) - M_i^{\text{aa}}\right)^2 \Big] \end{aligned} \tag{4}$$

Another way to include selection is to use the sequence-specific selection model $Q(\sigma)$ learned using SONIA or soNNia, and use Eq. 1 with $P(\sigma) = P_{\text{post}}(\sigma) = Q(\sigma) P_{\text{gen}}(\sigma)$ instead of $P_{\text{gen}}(\sigma)/(fq)$.

Finally, the two ways to model selection —sequence specific through $Q(\sigma)$, and ad hoc using q— may be combined by starting from P_{post} as a generative model, but assuming that only an unknown fraction q of sequences drawn from it actually survives. This leads to applying the same procedure as in the paragraph above, but with $P(\sigma) = P_{\rm post}(\sigma)/q$, and q inferred as in Eq. 4 but with predictions from the SONIA or soNNia model.

To define a pair-specific selection factor q_{AB} between two individuals A and B for the severity analysis, we fitted the prediction of Eq. 1 for the number of sequences shared between two individuals ($M = 2$), using $P(\sigma) = P_{\rm post}(\sigma)/q$. The fit was done using a grid search on q. This factor q_{AB} provides an inverse sharing index between A and B that is robust to the sample size.

Detecting significant levels of sharing among COVID-19 repertoires

To identify potential COVID-associated receptor clonotypes from oversharing in the cohort of infected individuals, we followed the approach developed in [89], by looking for clonotypes that are more shared in the COVID cohort than expected.

To estimate the baseline probability of CDR3 amino-acid sequences, we applied the SONIA (for B cells) or soNNia (for T cells) model to calculate $P_{\rm post}(\sigma)$.

We then computed the empirical frequency of sequences in the COVID-19 cohort, $P_{\rm data}(\sigma)$, from the counts x_i of CDR3 amino-acid clonotype σ in individual i, but in a probabilistic manner to represent our uncertainty. The probability of observing a receptor σ in repertoire i with $\approx N_i$ recombination events (taken to be almost equal to the number of unique nucleotide sequences) is binomially distributed, so that

$$\mathbb{P}(x_1, ..., x_n \mid P_{\rm data}) = \\ = \prod_{i=1}^{n} \binom{N_i}{x_i} \left[(1 - P_{\rm data})^{N_i}\right]^{x_i} \left[1 - (1 - P_{\rm data})^{N_i}\right]^{N_i - x_i}. \tag{5}$$

The maximum likelihood estimate of $P_{\rm data}(\sigma)$ reads:

$$P^*_{\rm data} = \arg\max_{P_{\rm data}} \mathbb{P}(x_1, ..., x_n \mid P_{\rm data}). \tag{6}$$

However in general we will work with the posterior distribution of $P_{\rm data}$, given by:

$$\rho(P_{\rm data} \mid x_1, ..., x_n) = \\ = \frac{\mathbb{P}(x_1, ..., x_n \mid P_{\rm data})}{\int_0^1 \mathbb{P}(x_1, ..., x_n \mid P_{\rm data}) dP_{\rm data}}, \tag{7}$$

where we have assumed a flat prior for $P_{\rm data}$.

In practice, we observe a systematic fold-difference between $P_{\rm post}$ and $P^*_{\rm data}$. To correct for this effect, we first rescaled $P_{\rm post}$, $P^*_{\rm post}(\sigma) = \alpha P_{\rm post}(\sigma)$. The correction factor α was fitted by minimizing $\sum_{\sigma \in \text{shared}} \left(\log P^*_{\rm data}(\sigma) - \log P_{\rm post}(\sigma) - \log \alpha\right)^2$.

To evaluate whether a sequence is more present than expected, we compute the posterior probability that its empirical frequency is larger than expected, $P_{\rm data} > P^*_{\rm post}$:

$$\mathbb{P}(P_{\rm data} < P^*_{\rm post}) = \int_0^{P^*_{\rm post}} \rho(P_{\rm data} \mid x_1, ..., x_n) dP_{\rm data}. \tag{8}$$

If this quantity is low, then we have confidence that the sequence is more shared than expected in the cohort, and is therefore called a COVID-associated clonotype. We used different thresholds depending on the context, often conservatively keeping the top sequences, rather than applying a fixed threshold.

Control cohort with matching $P_{\rm post}$ distribution

When comparing COVID-associated TCRβ to sequences from bulk repertoires, we observed differences in their $P_{\rm post}$ distribution reflecting the relationship between overlap and $P_{\rm post}$. To make comparisons without that possibly confounding factor, we built a healthy control by histogramming the $\log P_{\rm post}$ distribution of significantly shared sequences in COVID-19 patients into n equal-width bins. For each bin of width w and height h, we randomly drew clonotypes from healthy repertoires falling in that $P_{\rm post}$-bin, until we reached frequency h. The resulting dataset has the same distribution of $P_{\rm post}$ as SARS-CoV-2 sequences, but with sequences from the control repertoire.

Repertoire-based diagnosis

We used two methods to exploit information about the presence and absence of COVID-associated clonotypes in the repertoire to predict COVID-19 status: a likelihood ratio test, and logistic regression.

The likelihood ratio test evaluates the likelihood of the data under two competing hypotheses: infected with SARS-CoV-2 (H_1), or not infected (H_0). The log likelihood ratio score then reads $\mathcal{L} = \log(\mathbb{P}(\text{data}|H_1)/\mathbb{P}(\text{data}|H_0))$, with:

$$\mathbb{P}(\text{data}|H_0) = \prod_{\sigma \in i \cap S} (1 - e^{-N_i P^*_{\rm post}(\sigma)}) \prod_{\sigma \in S \setminus i} e^{-N_i P^*_{\rm post}(\sigma)}, \tag{9}$$

$$\mathbb{P}(\text{data}|H_1) = \prod_{\sigma \in i \cap S} (1 - e^{-N_i P^*_{\rm data}(\sigma)}) \prod_{\sigma \in S \setminus i} e^{-N_i P^*_{\rm data}(\sigma)}, \tag{10}$$

where $\sigma \in i \cap S$ means that the sequence is COVID-associated and is found in individual i, and $\sigma \in S \setminus i$ that it is COVID-associated but not present in i. We took the corrected model $P^*_{\rm post}$ for the null (uninfected) hypothesis, and the maximum likelihood estimator of the sequence frequency in COVID-19 donors, $P^*_{\rm data}$, for the infected hypothesis. To avoid overfitting, we learn $P_{\rm data}$ on a training set.

The set S can be adjusted by taking the n top clones (ranked from lowest to highest value of the score of Eq. 8), and letting n vary. Individuals with score $\mathcal{L} > c$ are called positives, and $\mathcal{L} < c$ negative, with c an adjustable threshold, which is set to achieve a given sensitivity on the training dataset.

For the purpose of TCR we turned to logistic regression, which uses a more general form of the score given by the logistic function:

$$\mathcal{L}^i_{\rm logit} = \log \frac{P(H_1|i)}{1-P(H_1|i)} = \beta_0 + \sum_{\sigma\in S} \beta_\sigma x^i_\sigma \qquad (11)$$

where $x^i_\sigma \in \{0,1\}$ denotes the absence of presence of sequence σ from S in individual i. Note that the likelihood ratio test is a particular case, with $\beta_\sigma = \log(e^{N_i P^*_{\rm data}(\sigma)} - 1) - \log(e^{N_i P^*_{\rm post}(\sigma)} - 1)$. By constrast, in logistic regression the parameters β_σ are trained on labeled data, i.e. repertoires of individuals with know COVID-19 status, denotes by $y_i = 1$ if the person was SARS-CoV-2 positive, and $y_i = 0$ otherwise. Logistic regression is performed by minimizing the cross-entropy over the β_σ:

$$\mathcal{H} = \sum_i \left[y_i \log P(H_1|i) + (1-y_i)\log(1-P(H_1|i))\right] + C^{-1}\sum_\sigma |\beta_\sigma|, \qquad (12)$$

where the last term is an ℓ_1 regularization. The parameters n (controling the size of S) and C were optimized to optimize performance (see Fig. S3). With $n = 7000$ and $C = 0.0005$, the logistic classifier achieved 83% specificity and 92% precision on held-out TCR. The dataset was separated between training and testing sets as described in the main text.

Acknowledgements

This work was supported by the Marie Skłodowska-Curie Actions H2020-MSCA-ITN-2017 program no 764698 from the European union, and by a European Research Council Consolidator grant no 724208.

[1] Li H, Ye C, Ji G, Han J (2012) Determinants of public t cell responses. *Cell Research* 22:33–42.

[2] Venturi V, et al. (2006) Sharing of t cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences* 103:18691–18696.

[3] Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public t-cell responses? *Nature Reviews Immunology* 8:231–238.

[4] Quigley MF, et al. (2010) Convergent recombination shapes the clonotypic landscape of the naïve t-cell repertoire. *Proceedings of the National Academy of Sciences* 107:19414–19419.

[5] Elhanati Y, Sethna Z, Callan Jr CG, Mora T, Walczak AM (2018) Predicting the spectrum of tcr repertoire sharing with a data-driven model of recombination. *Immunological Reviews* 284:167–179.

[6] Minervina A, Pogorelyy M, Mamedov I (2019) T-cell receptor and b-cell receptor repertoire profiling in adaptive immunity. *Transplant International* 32:1111–1123.

[7] Das J, Jayaprakash C (2018) *Systems immunology: An introduction to modeling methods for scientists* (CRC Press).

[8] Yates A (2014) Theories and quantification of thymic selection. *Frontiers in Immunology* 5:13.

[9] Nemazee D (2017) Mechanisms of central tolerance for b cells. *Nature Reviews Immunology* 17:281–294.

[10] Zarnitsyna V, Evavold B, Schoettle L, Blattman J, Antia R (2013) Estimating the diversity, completeness, and cross-reactivity of the t cell repertoire. *Frontiers in Immunology* 4:485.

[11] Mora T, Walczak AM (2016) Quantifying lymphocyte receptor diversity.

[12] Robins HS, et al. (2009) Comprehensive assessment of t-cell receptor β-chain diversity in $\alpha\beta$ t cells. *Blood, The Journal of the American Society of Hematology* 114:4099–4107.

[13] Briney B, Inderbitzin A, Joyce C, Burton DR (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 566:393–397.

[14] Arstila TP, et al. (1999) A direct estimate of the human &#x3b1;&#x3b2; t cell receptor diversity. *Science* 286:958–961.

[15] Madi A, et al. (2014) T-cell receptor repertoires share a restricted set of public and abundant cdr3 sequences that are associated with self-related immunity. *Genome research* 24 10:1603–12.

[16] Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nature Genetics* 49:659–665.

[17] Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences* 111:9875–9880.

[18] Isacchini G, Walczak AM, Mora T, Nourmohammad A (2021) Deep generative selection models of t and b cell receptor repertoires with sonnia. *Proceedings of the National Academy of Sciences* 118.

[19] Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with igor. *Nature Communications* 9:561.

[20] Sethna Z, et al. (2020) Population variability in the generation and selection of t-cell repertoires. *PLOS Computational Biology* 16:1–21.

[21] Sundling C, et al. (2021) Positive selection of igg+ over igm+ b cells in the germinal center reaction. *Immunity* 54:988–1001.e5.

[22] Egorov ES, et al. (2018) The changing landscape of naive t cell receptor repertoire with human aging. *Frontiers in Immunology* 9:1618.

[23] Britanova OV, et al. (2014) Age-related decrease in tcr repertoire diversity measured with deep and normalized sequence profiling. *The Journal of Immunology* 192:2689–2698.

[24] Vallejo AN (2006) Age-dependent alterations of the t cell repertoire and functional diversity of t cells of the aged. *Immunologic Research* 36:221–228.

[25] Chidgey A, Dudakov J, Seach N, Boyd R (2007) Impact of niche aging on thymic regeneration and immune reconstitution. *Seminars in Immunology* 19:331–340 Giving T Cells a Chance to Come Back.

[26] Palmer D (2013) The effect of age on thymic function. *Frontiers in Immunology* 4:316.

[27] Mayer A, Balasubramanian V, Walczak AM, Mora T (2019) How a well-adapting immune system remembers. *Proceedings of the National Academy of Sciences* 116:8815–8823.

[28] Nikolich-Žugich J (2008) Ageing and life-long maintenance of t-cell subsets in the face of latent persistent infections. *Nature Reviews Immunology* 8:512–522.

[29] Nielsen SC, et al. (2020) Human b cell clonal expansion and convergent antibody responses to sars-cov-2. *Cell Host & Microbe* 28:516 – 525.e5.

[30] Montague Z, et al. (2021) Dynamics of b cell repertoires and emergence of cross-reactive responses in patients with different severities of covid-19. *Cell Reports* 35:109173.

[31] Galson JD, et al. (2020) Deep sequencing of b cell receptor repertoires from covid-19 patients reveals strong convergent immune signatures. *Frontiers in Immunology* 11:3283.

[32] Cao Y, et al. (2020) Potent neutralizing antibodies against sars-cov-2 identified by high-throughput single-cell sequencing of convalescent patients' b cells. *Cell* 182:73–84.e16.

[33] Robbiani DF, et al. (2020) Convergent antibody responses to sars-cov-2 in convalescent individuals. *Nature* 584:437–442.

[34] Mor M, et al. (2021) Multi-clonal sars-cov-2 neutralization by antibodies isolated from severe covid-19 convalescent donors. *PLOS Pathogens* 17:1–26.

[35] Pinto D, et al. (2020) Cross-neutralization of sars-cov-2 by a human monoclonal sars-cov antibody. *Nature* 583:290–295.

[36] Raybould MIJ, Kovaltsuk A, Marks C, Deane CM (2020) Cov-abdab: the coronavirus antibody database. *Bioinformatics*.

[37] Zhou X, et al. (2021) Diverse immunoglobulin gene usage and convergent epitope targeting in neutralizing antibody responses to sars-cov-2. *Cell Reports* 35:109109.

[38] Brouwer PJM, et al. (2020) Potent neutralizing antibodies from covid-19 patients define multiple targets of vulnerability. *Science* 369:643–650.

[39] Barnes CO, et al. (2020) Structures of human antibodies bound to sars-cov-2 spike reveal common epitopes and recurrent features of antibodies. *Cell* 182:828–842.e16.

[40] Schultheiß C, et al. (2020) Next-generation sequencing of t and b cell receptor repertoires from covid-19 patients showed signatures associated with severity of disease. *Immunity* 53:442–455.e4.

[41] Goike J, et al. (2021) Synthetic repertoires derived from convalescent covid-19 patients enable discovery of sars-cov-2 neutralizing antibodies and a novel quaternary binding modality. *bioRxiv*.

[42] Stamatatos L, et al. (2021) mrna vaccination boosts cross-variant neutralizing antibodies elicited by sars-cov-2 infection. *Science* 372:1413–1418.

[43] Seydoux E, et al. (2020) Analysis of a sars-cov-2-infected individual reveals development of potent neutralizing antibodies with limited somatic mutation. *Immunity* 53:98–105.e5.

[44] Wang Z, et al. (2021) mrna vaccine-elicited antibodies to sars-cov-2 and circulating variants. *Nature* 592:616–622.

[45] Zhou Y, et al. (2021) Enhancement versus neutralization by sars-cov-2 antibodies from a convalescent donor associates with distinct epitopes on the rbd. *Cell Reports* 34:108699.

[46] Zost SJ, et al. (2020) Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the sars-cov-2 spike protein. *Nature Medicine* 26:1422–1427.

[47] Goodwin E, et al. (2018) Infants infected with respiratory syncytial virus generate potent neutralizing antibodies that lack somatic hypermutation. *Immunity* 48:339–349.e5.

[48] Rock EP, Sibbald PR, Davis MM, Chien YH (1994) Cdr3 length in antigen-specific immune receptors. *Journal of Experimental Medicine* 179:323–328.

[49] Saada R, Weinberger M, Shahaf G, Mehr R (2007) Models for antigen receptor gene rearrangement: Cdr3 length. *Immunology and cell biology* 85:323–332.

[50] Glanville J, et al. (2017) Identifying specificity groups in the t cell receptor repertoire. *Nature* 547:94–98.

[51] Dash P, et al. (2017) Quantifiable predictive features define epitope-specific t cell receptor repertoires. *Nature* 547:89–93.

[52] (2019) Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nature Communications* 10:1321.

[53] Madi A, et al. (2017) T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public cdr3 sequences. *eLife* 6:e22057.

[54] Pogorelyy MV, et al. (2019) Detecting t cell receptors involved in immune responses from single repertoire snapshots. *PLOS Biology* 17:1–13.

[55] Farmanbar A, Kneller R, Firouzi S (2019) Rna sequencing identifies clonal structure of t-cell repertoires in patients with adult t-cell leukemia/lymphoma. *npj Genomic Medicine* 4:10.

[56] Tareen A, Kinney JB (2019) Logomaker: beautiful sequence logos in Python. *Bioinformatics* 36:2272–2274.

[57] Shang J, et al. (2020) Structural basis of receptor recognition by sars-cov-2. *Nature* 581:221–224.

[58] Zhou P, et al. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579:270–273.

[59] Walls AC, et al. (2020) Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell* 181:281–292.e6.

[60] Tai W, et al. (2020) Characterization of the receptor-binding domain (rbd) of 2019 novel coronavirus: implication for development of rbd protein as a viral attachment inhibitor and vaccine. *Cellular & Molecular Immunology* 17:613–620.

[61] Scialo F, et al. (2020) Ace2: The major cell entry receptor

for sars-cov-2. *Lung* 198:867–877.

[62] Premkumar L, et al. (2020) The receptor-binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in sars-cov-2 patients. *Science Immunology* 5.

[63] Ng OW, et al. (2016) Memory t cell responses targeting the sars coronavirus persist up to 11 years post-infection. *Vaccine* 34:2008–2014.

[64] Tang F, et al. (2011) Lack of peripheral memory b cell responses in recovered patients with severe acute respiratory syndrome: A six-year follow-up study. *The Journal of Immunology* 186:7264–7268.

[65] Nolan S, et al. (2020) A large-scale database of t-cell receptor beta sequences and binding associations from natural and synthetic exposure to sars-cov-2. *Research Square*.

[66] Minervina AA, et al. (2021) Longitudinal high-throughput tcr repertoire profiling reveals the dynamics of t-cell memory formation after mild covid-19 infection. *eLife* 10:e63502.

[67] Braun J, et al. (2020) Sars-cov-2-reactive t cells in healthy donors and patients with covid-19. *Nature* 587:270–274.

[68] Mateus J, et al. (2020) Selective and cross-reactive sars-cov-2 t cell epitopes in unexposed humans. *Science* 370:89–94.

[69] Chang CM, et al. (2021) Profiling of t cell repertoire in sars-cov-2-infected covid-19 patients between mild disease and pneumonia. *Journal of Clinical Immunology* 41:1131–1145.

[70] Setliff I, et al. (2018) Multi-donor longitudinal antibody repertoire sequencing reveals the existence of public antibody clonotypes in hiv-1 infection. *Cell Host & Microbe* 23:845–854.e6.

[71] Parameswaran P, et al. (2013) Convergent antibody signatures in human dengue. *Cell Host & Microbe* 13:691–700.

[72] Truck J, et al. (2015) Identification of antigen-specific b cell receptor sequences using public repertoire analysis. *The Journal of Immunology* 194:252–261.

[73] Mathew D, et al. (2020) Deep immune profiling of covid-19 patients reveals distinct immunotypes with therapeutic implications. *Science* 369:eabc8511.

[74] Shi R, et al. (2020) A human neutralizing antibody targets the receptor-binding site of sars-cov-2. *Nature* 584:120–124.

[75] Liu STH, et al. (2020) Convalescent plasma treatment of severe covid-19: a propensity score–matched control study. *Nature Medicine* 26:1708–1713.

[76] Rajendran K, et al. (2020) Convalescent plasma transfusion for the treatment of covid-19: Systematic review. *Journal of medical virology* 92:1475–1483.

[77] Ko TM, et al. (2018) Immunoglobulin profiling identifies unique signatures in patients with Kawasaki disease during intravenous immunoglobulin treatment. *Human Molecular Genetics* 27:2671–2677.

[78] Ahmed SF, Quadeer AA, McKay MR (2020) Preliminary identification of potential vaccine targets for the covid-19 coronavirus (sars-cov-2) based on sars-cov immunological studies. *Viruses* 12.

[79] Zhao J, et al. (2017) Recovery from the middle east respiratory syndrome is associated with antibody and t cell responses. *Science Immunology* 2:eaan5393.

[80] Sharon E, et al. (2016) Genetic variation in mhc proteins is associated with t cell receptor expression biases. *Nature Genetics* 48:995–1002.

[81] Chang CM, et al. (2021) Profiling of t cell repertoire in sars-cov-2-infected covid-19 patients between mild disease and pneumonia. *Journal of Clinical Immunology* 41:1131–1145.

[82] Zheng HY, et al. (2020) Elevated exhaustion levels and reduced functional diversity of t cells in peripheral blood may predict severe progression in covid-19 patients. *Cellular & Molecular Immunology* 17:541–543.

[83] Shrotri M, et al. (2021) T cell response to sars-cov-2 infection in humans: A systematic review. *PLOS ONE* 16:1–21.

[84] Quiros-Fernandez I, Poorebrahim M, Fakhr E, Cid-Arregui A (2021) Immunogenic t cell epitopes of sars-cov-2 are recognized by circulating memory and na&#xef;ve cd8 t cells of unexposed individuals. *EBioMedicine* 72.

[85] Vander Heiden JA, et al. (2014) presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 30:1930–1932.

[86] Giudicelli V, et al. (2006) Imgt/ligm-db, the imgt® comprehensive database of immunoglobulin and t cell receptor nucleotide sequences. *Nucleic Acids Research* 34:D781–D784.

[87] Ye J, Ma N, Madden TL, Ostell JM (2013) Igblast: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research* 41:W34–W40.

[88] Sethna Z, Elhanati Y, Callan, Curtis G J, Walczak AM, Mora T (2019) OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* 35:2974–2981.

[89] Pogorelyy MV, et al. (2018) Method for identification of condition-associated public antigen receptor sequences. *eLife* 7:e33050.

Supplementary information

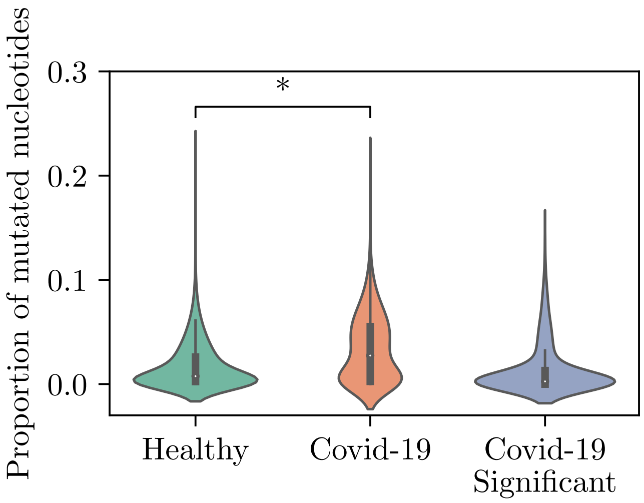

FIG. S1: Violin plots representing the proportion of mutated nucleotides in B cell receptor repertoires per individual in healthy and COVID-19 cohorts, as well as in significantly shared sequences. The dot represents the median SHM percentage. Student's t test: $*p < 10^{-5}$.

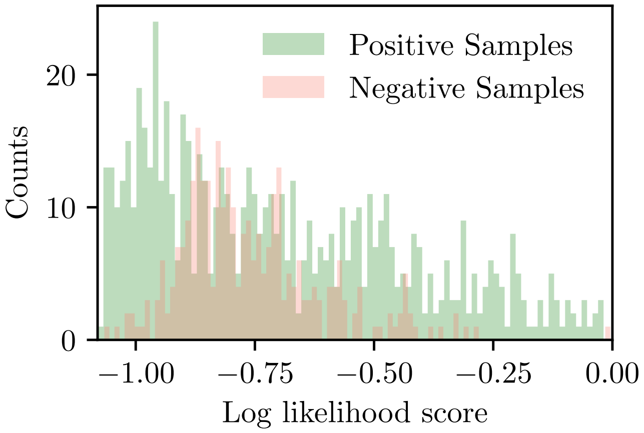

FIG. S2: Distribution of log likelihood ratio scores for 1000 T-cell repertoires belonging to the test cohort of 700 COVID-19 patients and 300 healthy individuals. The distribution of scores for each group completely overlap, leading to a classifier with a very poor accuracy.

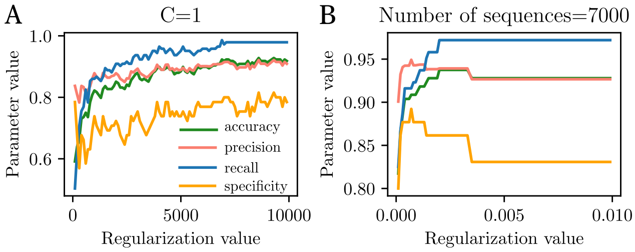

FIG. S3: Parameter selection for the logistic regression model applied to T cell repertoires. (A) Variation of the logistic regression performance as a function of the total number of sequences. In the absence of regularization, both precision and accuracy reach their maximum for n=7000 sequences. (B) For the optimal number of sequences found, the precision and specificity are maximized for C=0.0008.

TABLE S1: Table of top COVID19-associated BCR, containing their CDR3 amino acid sequence, V and J gene choices, P_{post} and P^*_{data} value, posterior probability of being associated (p), and index of reported antibody match from the curated database (index column in Table S3).

TABLE S2: Table of top COVID19-associated TCR, containing their CDR3 amino acid sequence, V and J gene choices, P_{post} and P^*_{data} value, posterior probability of being associated (p). All reported responding clones come from MIRA dataset.

TABLE S3: Summary table containing all reported antibodies extracted from the different studies (indicated in column *References*) used for sequence matching. When known, both heavy and light chain information are provided (CDR3 amino acid sequence, V and J gene choices) as well as the antigen target by the antibody.