Altered somatic hypermutation patterns in COVID-19 patients classifies disease severity

Modi Safra[1,2], Zvi Tamari[1,2], Pazit Polak[1,2], Shachaf Shiber[3,4], Moshe Matan[5], Hani Karameh[8], Yigal Helviz[9], Adva Levy-Barda[11], Vered Yahalom[4,12], Avi Peretz[5,6], Eli Ben-Chetrit[7], Baruch Brenner[4,10], Tamir Tuller[13], Meital Gal-Tanamy[6], and Gur Yaari[1,2,*]

[1]Bio-engineering, Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel
[2]Bar Ilan Institute of Nanotechnologies and Advanced Materials, Bar Ilan University, Ramat Gan, Israel
[3]Emergency Department, Rabin Medical Center- Belinson campus, Petah Tikva, Israel
[4]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
[5]Clinical Microbiology Laboratory, Baruch Padeh Medical Center, Poriya, Israel
[6]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel
[7]Infectious Diseases Unit, Shaare Zedek Medical Center, Hebrew University School of Medicine, Jerusalem, Israel
[8]Jesselson Integrated Heart Center, Shaare Zedek Medical Center, Hebrew University School of Medicine, Jerusalem, Israel
[9]Intensive Care Unit, Shaare Zedek Medical Center, Hebrew University School of Medicine, Jerusalem, Israel
[10]Institute of Oncology, Rabin Medical Center- Belinson campus, Petah Tikva, Israel
[11]Biobank, Department of pathology, Rabin Medical Center- Belinson campus, Petah Tikva, Israel
[12]Blood Services & Apheresis Institute Director, Rabin Medical Center- Belinson campus, Petah Tikva, Israel
[13]Department of Biomedical Engineering and The Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
[*]Correspondence: Gur Yaari, gur.yaari@biu.ac.il

December 20, 2022

Abstract

The success of the human body in fighting SARS-CoV-2 infection relies on lymphocytes and their antigen receptors. Identifying and characterizing clinically relevant receptors is of utmost importance. We report here the application of a machine learning approach, utilizing B cell receptor repertoire sequencing data from severely and mildly infected individuals with SARS-CoV-2 compared with uninfected controls. In contrast to previous studies, our approach successfully stratifies non-infected from infected individuals, as well as disease level of severity. The features that drive this classification are based on somatic hypermutation patterns, and point to alterations in the somatic hypermutation process in COVID-19 patients. These features may be used to build and adapt therapeutic strategies to COVID-19, in particular to quantitatively assess potential diagnostic and therapeutic antibodies. These results constitute a proof of concept for future epidemiological challenges.

Keywords: machine learning, BCR, AIRR-seq, COVID-19,somatic hypermutation, B cell

Background

Despite the unprecedented speed of vaccine development against SARS-CoV2, the virus continues to undergo changes that cause repeated waves of COVID-19 morbidity worldwide, with increasing infectivity. Risk factors such as age (> 60) and preexisting medical conditions can predict to some extent whether an individual will become severely ill or not, but the prediction is not very accurate. The early phase of infection results in direct tissue damage, followed by a late phase when the infected cells trigger an immune response, by recruitment of immune cells that release cytokines (reviewed in [1]). In severe patients, this may result in a "cytokine storm" and a systemic inflammatory response. Many individuals do not respond well enough to the vaccine, either because of old age or immune impairments. Thus, there is an ongoing search for anti-viral therapies and passive vaccines, as well as research into the basic mechanisms related to the virus and immunity towards it.

One useful path to investigate the immunity towards SARS-CoV-2 is adaptive immune receptor repertoire sequencing (AIRR-seq) [2, 3, 4], revealing noticeable changes in affected individuals in many arms of the immune system [5, 6]. Millions of B and T cell receptor

(BCR and TCR, respectively) sequences from hundreds of individuals have been shared in public archives such as iReceptor [7] and OAS [8]. Thousands of individual antibody sequences validated as targeting and neutralizing SARS-CoV-2 have been published in datasets such as CoV-AbDab [9].

In the past few years, several studies have used AIRR-seq data to train machine learning (ML) algorithms to classify individuals who carry diseases [10], including celiac [11, 12], hepatitis C virus infection [13], cytomegalovirus [14], and others [15]. Finding the connection between AIRR-seq data and health states is a highly challenging task, because of the massive volume of AIRR-seq datasets that can include tens of millions of sequences that dilute the disease-specific biological signals. Another difficulty is our inability to determine to which antigen(s) each receptor can bind based solely on the receptor sequence. New methods to identify relevant repertoire features are continuously developed [10, 16, 17]. Besides the diagnostic and prognostic potential, such features can be critical in teaching us about the mechanisms behind the disease and the successful immune response towards it. Thus far, the vast majority of efforts to classify the health state or severity of COVID-19 have relied on TCR data [18, 19, 20, 21]. Recently, for example, a new approach to detect SARS-CoV-2 infection by TCR sequencing has been FDA approved for clinical use [20].

B cells undergo affinity maturation after pathogen encounter, to further adapt to the specific pathogen. Affinity maturation includes iterative cycles of somatic hypermutation (SHM) and affinity dependent selection. While selection depends on better binding, the SHM mechanism is independent of pathogen affinity. During SHM, different enzymatic pathways orchestrate together to introduce mutations specifically in the genomic regions encoding the antibody [22]. Extensive investigations have been devoted to understanding the SHM mechanism [23, 24, 25, 26], but to the best of our knowledge, no connection of a specific infection to a specific SHM pathway or pattern was made. The use of BCR sequencing is considered more difficult than TCR, because of SHM and higher diversity in the complementary determining region 3 (CDR3). It has been reported that BCR sequencing data cannot be used to classify individuals with COVID-19 [21]. Nevertheless, BCR data may be more informative than TCR in some cases, as BCRs undergo affinity maturation to adapt to each pathogen.

Here, using bulk and single cell BCR sequencing data, we successfully classify SARS-CoV-

2 infected vs. naive individuals, as well as determine disease severity. Compared with the traditional sequence similarity clustering based approach, we obtain better classifications by considering SHM pattern changes in SARS-CoV-2 infected individuals. SHM specific patterns connected to decreased severity, as well as important amino acid (AA) composition in SARS-CoV2 antibodies, were identified.

Methods

Collection of samples

The repertoires composing the dataset were collected at three medical centers. IRB approval numbers: Rabin (Beilinson) Medical Center, 0256-20-RMC; Baruch Padeh Medical Center, 0037-20-POR; Shaare Zedek Medical Center, 0303-20-SZMC. 28 samples of controls were collected, as well as 39 mild patients with COVID-19 and 12 severely infected patients. Patients' data can be found in Table S1. We do not have information about the SARS-CoV2 strains, but they are almost certain to be the original strain (before Alpha (B.1.1.7)). All samples were collected between April and early November 2020, and the earliest documented variant strains, as well as the earliest vaccines, arrived in Israel in late December 2020.

Library preparation

Bulk: Ig repertoires were bulk sequenced according to the method described in [27]. All controls as well as 32 COVID-19 patients were sequenced for both heavy and light chains. These were used as the train/validation groups for the ML algorithms. For the rest of the patients, only heavy chains were sequenced, and served as the test group. 13 more controls for the test group were added from previously published datasets. Nine controls from dataset [28], and four from dataset [29].

Single cell: PBMCs from 13 individuals were prepared from fresh 5ml blood samples, and frozen according to the manufacturer's instruction of the "Fresh Frozen Human Peripheral Blood Mononuclear Cells for Single Cell RNA Sequencing" protocol, document number CG00039 Rev D, 10X Genomics. Patients' data can be found in Table S2. We do not have information about the SARS-CoV2 strains, as these tests were not routinely performed at that

time (January-February 2021). Patients were not vaccinated. Libraries were prepared according to the manufacturer's instruction of the "Chromium Next GEM Single Cell 5' Reagent Kit v2 (Dual Index)" protocol, document number CG000331 Rev A, 10X Genomics. Libraries were pooled, mixed with 1% PhiX, and sequenced on an Illumina NovaSeq twice using an SP and an S1 kits.

Data processing and statistics

FASTA files were generated using the PRESTO pipeline [30], and aligned to IMGT IGHV/D/J genes [31] using the VDJbase pipeline. Only sequences which started at the first 30 bases of the V gene were included. Isotype frequencies, V, D, J and combinations of V & J gene usage and CDR3 AAs 3-mers, as well as CDR3 AA lengths and V gene identities were calculated using a custom-designed R script (see data and code availability section). The same script also calculated the frequencies of BCR clusters (sharing the same V and J genes and junction AA length). Diversity was calculated using the alphaDiversity function from the Alakazam R package [32]. All P values were calculated using Wilcox test and adjusted using the Benjamini-Hochberg procedure [33].

Generating an SHM model

A 5-mer SHM model was built using the function createTargetingModel from the shazam R package [23], once for silent mutations only and once for both silent and replacement mutations. To create these metrics for one representative from each clone, we used the collapseClones function from the same package. For each repertoire, substitutions, mutability, and targeting values were collapsed into a single table. Tables from all repertoires were collapsed into a single table. The tables enable both training ML algorithms and calculating mean mutability in specific sites (WR<u>C</u>/<u>G</u>TW and W<u>A</u>/<u>T</u>W hot-spots, the SYC/GRS cold-spot and all other sites). The table was also used to calculate single base mean mutability levels in all repertoires. The single base mutability was calculated as the average of all 5-mers with the same base in the middle.

Training and estimation of ML algorithms

50 random splits to train and validation groups were made in order to estimate the F1 score, accuracy, sensitivity, and specificity of each model. Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET) using the caret R package [34] were trained on tables containing data from the repertoires. Feature selection was done using t-test calculations between frequencies in the different groups in the train subset only. Only features with P value below a certain threshold were selected. The algorithm was then trained on the selected data, and classifications were made for the validation groups. F1 score, accuracy, sensitivity, and specificity were calculated for each random split.

COVID-19 classification using AA frequencies at all V gene positions

Frequencies of each AA along 103 positions (according to the IMGT numbering) in each V gene family were calculated for all repertoires. The train/validation samples were used to train the same algorithm as explained above, and to estimate the F1 score, accuracy, sensitivity, and specificity of the algorithm. The validation group was used to estimate the parameters of the algorithm on unseen data. Coefficients of the algorithm were extracted and enabled to calculate scores for single antibodies. If a certain AA was present in the sequence, it received a frequency of 1. Otherwise, it received a frequency of 0. This equation was used to calculate scores for all antibodies in all repertoires, as well as scores for known COVID-19 antibodies from the CoV-AbDab database.

Single cell data analysis

Single cell data was analyzed using cell-ranger 6.0.1 with output of both VDJ recombination and gene expression data. Cell-ranger output was then manipulated using the Seurat R package [35]. Cells with more than 5% mitochondrial gene expression were removed. Data was normalized, and PCA and UMAP on the top 10 PCAs were done using standard Seurat functions. Cell identity was determined using the SingleR R package against a sorted dataset from the celldex R package [36]. Barcodes of VDJ data and gene expression data were matched using R.

Results

BCR gene usage cannot classify SARS-CoV2 infection

To assess changes in BCR repertoires of COVID-19 patients, we collected 79 blood samples and sequenced their BCR repertoires. Samples were split to three groups: uninfected individuals, mildly and severely COVID-19 infected patients. For each group we characterized several whole repertoire features, such as CDR3 AA length distribution, V gene mutation distribution, clonal diversity, V, D, J and combination of V and J gene usage. We also calculated frequencies of BCR clusters (same V and J gene as well as same CDR3 AA length). These measurements are shown in Fig. 1 and in Fig. S1 for heavy chains, and for kappa and lambda light chains in Figs. S2 and S3. As expected, the diversity of BCR clones is significantly lower in COVID-19 patients compared with controls (Fig.1C). No significant difference was observed in CDR3 AA length (Fig.1A), and only slight increase was seen in V gene mutation distribution (Fig.1B). For many V genes we observed significantly reduced usage in COVID-19 patients (Fig.1D). Three exceptions are IGHV4-34, IGHV4-39 and IGHV4-59 that demonstrate increased usage upon infection, which is further increased in severe patients compared with mild ones. These results support previously published COVID-19 data [37, 38], and suggest that antibodies against SARS-CoV2 mainly comprise those genes. To further validate these conclusions, we tried to build ML classifiers based on V, V & J gene usage, or V & J gene usage and 85% similarity in the CDR3 AAs. However, these models yielded less than 70% accuracy, suggesting low impact of V or V & J gene usage on the response to SARS-CoV2 infection.

We explored further whole repertoire features, and compared isotype frequencies between the different groups. While we observed a reduction in the frequencies of IGD and IGM upon SARS-CoV2 infection, the levels of IGG increased (Fig.1E), and those of IGA remained unchanged. We also measured silent mutability frequencies for each isotype (Fig. 1F). These measurements avoid changes which are caused by antibodies selective pressure. In contrast to the IGG and IGA class switched isotypes, in which mutability upon infection is reduced, in IGD and IGM mutability is increased. In severe patients, the IGD and IGM mutability was even higher (Fig.1F).

BCR V gene AA composition successfully classifies SARS-CoV2 infection and may reveal important features of antibodies against the virus

We continued exploring classification approaches to stratify COVID-19 patients and uninfected individuals. To this end, we explored AA frequencies along the V gene, aggregated by V gene family. We generated a table with 10,300 columns, counting AA frequencies along 103 V gene positions (aligned according to IMGT numbering), for the 5 most highly used V gene families (IGHV1-5). Using this approach we obtained a high F1 score of more than 0.85, and similar levels of accuracy, sensitivity,and specificity (Fig.2A). The test set resulted in an F1 score of above 0.85 (Fig. 2B). We then extracted the coefficient used by the algorithm, corresponding to the contribution of each AA frequency to the classification of the disease (Fig. 2D).

To further validate that these changes are unique to COVID-19 patients, we downloaded a dataset of more than 450 repertoires from cAb-rep data collection [39]. These data include repertoire sequencing results from a wide variety of clinical conditions such as Hepatitis B virus infection, vaccinations against Hepatitis B virus and influenza, and several autoimmune diseases. Applying our algorithm to these data to classify COVID-19 infection resulted in a false positive rate of only 6%, indicating that our classification is specific to COVID-19 infection.

These results were obtained for the repertoire level, and we sought to test their applicability to the single BCR sequence level. For this, we transferred the features selected for the repertoire level model, i.e., AA frequencies along the V gene families, to calculate a score for single BCR sequences. We calculated such scores for a list of more than 5,000 known antibodies against SARS-CoV2 from the CoV-AbDab database [40]. The scores of the known antibodies were higher than those came from whole repertoires of control patients as well as most of the COVID-19 infected repertoires (Fig. 2C), suggesting that these coefficients are meaningful not only for the repertoire level, but also for single BCR sequences. Our attempts to classify the severity of COVID-19 using this method were not successful, so for this purpose, we explored other sets of features. The coefficients of the algorithm can be seen in Fig. 2D.

Mutation bias in class-switched B cells of COVID-19 patients

As reduced levels of overall BCR mutability were seen upon SARS-CoV2 infection only in the class switched isotypes (Fig 1F), we quantified single base mutability patterns in these isotypes. As seen in figure 3A, the mean relative mutability is reduced in COVID-19 patients at Cytosine and Guanine (C and G), but increases in Adenine and Thymine (A and T). The same results were obtained when considering silent mutations only (Fig. 3B). Five main pathways are responsible for introducing mutations during SHM [41]. Three introduce mutations in C and G, and the other two involve the low fidelity DNA polymerase polη, which mutates A and T. The significant differences in mutability observed in COVID-19 patients suggest altered activity of those arms. To further investigate SHM in SARS-CoV2 infection, we applied a commonly used 5-mers SHM mutability model [23]. In general, two highly mutated hot-spot motifs are commonly observed in SHM. One is WR$\underline{\text{C}}$/$\underline{\text{G}}$YW (where W = {A, T}, Y = {C, T} R = {G, A}, and the mutated position is underlined), and the other is W$\underline{\text{A}}$/$\underline{\text{T}}$W. In addition, SY$\underline{\text{C}}$/$\underline{\text{G}}$RS (where S = {C, G}), is considered as a cold-spot sequence motif. We first built a 5-mer mutability model based on both silent and replacement mutations. Such a model combines the effects of SHM and antigen-driven selection. We divided the 5-mers to those occurring in the two hot-spots, in the cold-spot, and in all other neutral sites, and show their levels for IGD/IGM and for IGA/IGG (Fig. 3C and E). The most significant changes between the different groups are a decrease in the WR$\underline{\text{C}}$/$\underline{\text{G}}$YW site and an increase in SY$\underline{\text{C}}$/$\underline{\text{G}}$RS in IGA/IGG of COVID-19 patients. This increase is not seen in severely infected patients.

To understand whether these patterns stem from SHM or from antigen-driven selection, we built another model, taking only silent mutations into consideration. Fig. 3D and F shows the resulting mutability scores for the same sequence motifs. The observed pattern resembles the one observed in Fig. 3C and E, suggesting that the alteration between the groups results from altered SHM characteristics. To avoid the effect of clonal expansion on mutability calculations, we repeated all calculations, taking into account only one representative from each clone. Similar results were obtained using this approach (Fig. S4). Moreover, using SHM matrices based only on a specific V family resulted in a much lower signal (Fig. S5F). Importantly, the mentioned SHM patterns reflect the relative likelihood for each mutation pattern and do not indicate the overall mutability level.

Silent SHM patterns classify SARS-CoV2 infection and severity

To estimate the level of connection between changes in SHM patterns and SARS-CoV2 infection, we tried again to build a classifier of samples' origin. We built two models, one using all mutations (Fig. 4A, S5, S6A and S8), and one using silent mutations only (Fig. 4B, S6B). Taking all mutations into account, we obtained an F1 score of over 0.85, as well as accuracy, sensitivity, and specificity values. Taking only silent mutations into account, we obtained a slightly lower result of $\sim$ 0.8 F1 score and accuracy. These results strengthen our hypothesis that the differences between the repertoires emerge mainly from SHM itself and not from antigen-driven selection. Using only light chain sequences for the mutability model reaches much lower results, as expected (Fig. S7A and B). A model based on the combination of light and heavy chains does not obtain better results than using the heavy chain only (Fig. S8).

Next, we tried to classify COVID-19 severity using SHM patterns. Since the mutability in the cold-spot motif changes the most between severe and mild patients, we built a model using mutability scores of this cold-spot only. We obtained an F1 score and accuracy of about 0.75 in severity classifications (Fig. 4C).

All patterns with non-zero coefficients have much higher mutability frequencies in mild patients compared with severe patients ((Fig. 4D). Again, to avoid the effect of clonal expansion and selective pressure on the inferred mutability model, we repeated the mutability model inference taking into account only one representative from each clone. As shown in Fig. S5, the results were comparable to those obtained using all sequences.

Known SARS-CoV2 antibodies are enriched in plasmablasts from COVID-19 patients

We thought to find in our sequencing data, antibodies that may be related to the known COVID-19 antibodies. As mentioned above, during the COVID-19 pandemic a new database summarizing all known SARS-CoV2 antibodies was published, containing more than 5,000 antibody AA sequences of both heavy and light chains. For each of our repertoires, we calculated and summarized the frequencies of sequences that are similar to known antibodies. We defined similar antibodies by 85% identity in the CDR3 AAs, and the same V and J genes. As expected, the frequencies of similar to known antibodies in COVID-19 patients were higher than those in

control individuals (Fig. 5A. Histograms summarizing the sizes and numbers of samples having at least one representation in the clones can be found in Fig. S9A and B). Using the sum of frequencies of similar to known COVID-19 clones, we reached an accuracy of above 70% in repertoire classification and an AUC of 0.81 (Fig. 5B). Even lower results were obtained when training the algorithm to count the frequencies of shared clones between samples (Fig. S10). Although significant, this result is lower than that achieved by considering mutations along the V gene.

To further explore the similarity to known antibodies, we performed 10X Genomics single cell sequencing including V(D)J and gene expression, on blood samples from additional 13 mild COVID-19 patients. Using single cell sequencing data enables matching of heavy and light chains, which cannot be done with bulk sequencing. Moreover, single cell sequencing provides the ability to identify cell type using gene expression signatures. We found similar to known antibodies in 7 out of the 13 repertoires. The frequencies were overall lower compared with those seen in the bulk RNA sequencing cohort (Fig. 5C). This could be due to the differences in sequencing methods, or because in the single cell cohort the patients were diagnosed on average more recently than the bulk cohort and thus may have had lower levels of SARS-CoV2 specific antibodies.

We then applied the SingleR R package to classify cell types by single cell expression profiles. Two-dimensional UMAP reduced plots are shown in Fig. 5D, demonstrating a distinct cluster of plasmablasts. We summarized the frequency of known SARS-CoV2 clusters in bulk sequenced COVID-19 patients, bulk controls, single cell unsorted data, and single cell plasmablasts only. As shown in Fig. 5E, COVID-19 patients show enriched levels of similarity to known SARS-CoV2 antibody compared with controls. Single cells show higher levels than controls but lower than bulk, as discussed above. Among plasmablasts of COVID-19 patients, we see the highest frequency of known antibody clusters, indicating a stereotypical response to SARS-CoV2. Lastly, to validate our observation that WR<u>C</u>/<u>G</u>YW hot-spots mutability scores decrease upon COVID-19 infection, and SY<u>C</u>/<u>G</u>RS cold-spots increase (Fig. 3), we split the single cell data into plasmablasts vs. all other B cell types. We built a mutability SHM matrix for each of these subsets, and indeed found a reduction in the mutability scores of WR<u>C</u>/<u>G</u>YW hot-spots in plasmablasts (0.00168) compared with the other B cell types (0.00178), and an

315 increase in the mutability scores of the SYC/GRS cold-spots (0.0003 and 0.0002, respectively).

316 Discussion

317 The COVID-19 pandemic, caused by evolving variants of SARS-CoV2, has infected a large
318 proportion of the population worldwide. Antibodies play a critical role in eliminating the virus
319 from the body. Serological tests are routinely used to estimate immunity of individuals against
320 SARS-CoV2, convalescent plasma donations were used to treat severely ill COVID-19 patients,
321 and many monoclonal antibodies were developed as candidate passive vaccinations.

322 Although the pandemic has caused a huge health and economic burden, it brought several
323 important advantages for biomedical research. With so many researchers and funding oppor-
324 tunities focusing on a single topic, the pandemic facilitated both broad and profound analyses
325 of the virus and the immune responses towards it. During the past two and a half years,
326 thousands of COVID-19 binding/neutralizing antibodies have been published and deposited
327 in public datasets[42, 43]. This huge amount of data facilitates finding BCR sequences that
328 are similar to known antibody sequences, and searching for common features. Such features
329 may be used in the clinic for diagnosis of the disease, but in the case of COVID-19 there are
330 easier, faster and cheaper ways to do that. Much more importantly, it can teach us about the
331 development of the immune response towards the virus.

332 Here, in contrast to previous reports[21], we were able to stratify COVID-19 patients and
333 healthy individuals based on shared clusters of BCR sequences. The moderate classification
334 results of such approach led us to explore different sets of features that turned out to be more
335 informative. AA frequencies at all V gene positions served as a basis for an ML model that
336 produced a high F1 score ($\sim 85\%$) in classifying COVID-19 infection.

337 The patterns of AA alterations in BCRs arise during the process of affinity maturation, that
338 includes two iterative processes, namely SHM and affinity-dependent selection. These patterns
339 can stem from the antibodies against SARS-COV2 or from overall altered SHM mechanism in
340 COVID-19 patients.

341 An important question that may arise when inspecting the presented approach is whether
342 it is specific to COVID-19, or perhaps it simply detects general signals related to an adaptive
343 immune response towards a new pathogen. We believe that the presented approach is specific

to COVID-19 because: 1. The signal does not disappear when choosing a single representative per clone, which eliminates the effect of general clonal expansion. 2. The signal is based on an SHM pattern, which is subject to an antigen-specific affinity maturation. 3. Our lab has a lot of experience in ML-based classification of different clinical conditions[44, 17, 28], and for each condition the features identified by the algorithm as the most essential for classification were different. SHM patterns have never been previously identified as a feature, as far as we know (but see our recent publication [45]). To test this, we applied our algorithm to data from ∼450 samples, including infection with Hepatitis B virus, vaccinations against Hepatitis B virus and influenza, and several autoimmune diseases. 94% of these repertoires were classified as healthy, indicating that our algorithm does not classify any neo-response as COVID-19.

Extensive research has been devoted to study SHM mechanisms affecting other regions in the antibody besides the CDR3[46, 23]. Yet, this knowledge has not been used for disease classifications, nor for improving antibody engineering. We sought to follow the SHM machinery during SARS-CoV2 infection, starting with the whole repertoire level. It is well established that antibodies binding SARS-CoV2 are very close to the germline[47, 5, 48, 49]. Surprisingly, even at the repertoire level, we detected a decrease in mutability of IGG BCRs. To explore whether the AA frequency-based signal results from alterations in SHM or affinity dependent selection, we followed the mutability rates of silent mutations only. These mutations are not subjected to affinity dependent selection pressure, thus reflecting changes in the machinery of SHM. We found that most SHM changes upon SARS-CoV2 infection were observed even when counting only silent mutations, which are not subject to affinity selection, suggesting dramatic changes in the SHM machinery upon SARS-CoV2 infection. To further pinpoint the effects on the SHM machinery, we repeated the calculations taking only one representative from each clone into account, thereby abolishing the effect of clonal expansion (Fig. S5). This step slightly reduced the F1 score, in a non-significant way. The fact that eliminating the effect of clonal expansion on our findings did not abolish the differences suggests that there are true changes in the SHM machinery. Moreover, the moderate performance reduction when taking only one representative per clone, hints that the SHM changes during SARS-CoV2 infection may be further enhanced by clonal expansion, potentially aiding the battle with the virus.

Many pathways are involved in the introduction of mutations to BCR sequences. In par-

ticular, two common SHM hot-spots, WRC/GYW and WA/TW, are affected by two different pathways. While mutations in WRC/GYW motifs are mediated by the activation induced deaminase, mutability at WA/TW motifs also involve the low fidelity DNA polymerase polη.

In the class switched IGA and IGG isotypes, we observed decreased mutability levels with increasing severity of COVID-19 at WRC/GYW motifs, and increased mutability at WA/TW sites. Again, these changes were observed even when counting silent mutations only, further supporting an impact of the virus on the SHM introduction mechanism. The reduced mutability in WRC/GYW motifs and the mildly increased mutability in WA/TW motifs may hint that AID levels could be decreased upon COVID-19 infection. This possibility will need to be validated in future studies. Another future direction is to test for possible SHM positional effects. The presence of such an effect was lately suggested [50], and it will be very interesting to inspect whether this is relevant to our results.

Another specific SHM target is the cold-spot SYC/GRS. Surprisingly, we found an increase in mutability rates of this cold-spot in COVID-19 repertoires. Moreover, this increase was not observed in severely infected patients, suggesting that this mechanism may be critical for production of efficient antibodies and thereby for prevention of severe illness.

Building on our success in classifying patients from healthy individuals, we sought to develop an ML-based algorithm to classify disease severity. This could have important clinical outcomes, since medications and passive vaccines now exist that can prevent deterioration if diagnosed individuals are treated rapidly. However, these treatments have side effects and are not given to the wide population. Prediction of disease severity by the known risk factors is highly inaccurate, and there are currently no other means to classify severity. Using mutability patterns from silent mutations only, we estimate our ability to classify COVID-19 severity at approximately 75%(Fig. 4C). The known risk factors to develop severe COVID-19 are mostly preexisting conditions such as older age, hypertension, obesity, diabetes. Here, we suggest another risk biomarker that involves basic features of the adaptive immune system. Many more steps are needed to enable prediction of COVID-19 infection and severity based on BCR sequencing data. We provide here a first step towards it.

AA frequency patterns along the V genes at the whole repertoire level is a sufficient feature for relatively good classification of COVID-19. Looking at the identity of AA along the V gene

of a single BCR sequence may reveal its affinity towards the virus. To explore the connection between the new BCR repertoire data generated here and known SARS-CoV2 antibody sequences we took a two way approach. Building on the hypothesis that the whole repertoire level signal responsible for the classification stems from individual SARS-CoV2-specific antibodies generated during the infection, we derived a single sequence score based on the repertoire classification signal. Although sequences with high scores are scarce in both healthy and COVID-19 repertoires, their prevalence in the CoV-abDab data is significantly higher (Fig. 2C). As such, the features (detailed in Fig. 2D) may be used for more rational antibody design towards the virus. In addition, we explored the presence of similar sequences to the validated CoV-abDab antibodies in both bulk and in single cell sequenced repertoires. We found a higher fraction of sequences with high similarity to known antibodies in COVID-19 patients compared with controls. This can also be used for successful classification of the repertoires. Notably, a group of COVID-19 patients had no similar antibodies to those in the list, suggesting that despite the massive efforts so far, the list is incomplete. On the other hand, in some control samples we found few sequences similar to known antibodies. These antibodies may provide a basis for protection from COVID-19 symptoms or complications to individuals who carry them.

Declarations

Ethics approval and consent to participate

The repertoires composing the dataset were collected at three medical centers. IRB approval numbers: Rabin (Beilinson) Medical Center, 0256-20-RMC; Baruch Padeh Medical Center, 0037-20-POR; Shaare Zedek Medical Center, 0303-20-SZMC. All participants received an explanation about the study from a medical doctor, and signed an informed consent form.

Consent for publication

Not applicable.

Availability of data and code

Our sequencing data will be available on NCBI upon publication, under BioProject PRJNA839749. All code will be available on github.

Competing interests

The authors declare that they have no competing interests.

Funding

We thank the Israeli Ministry of Science grant 3-16909, the Israeli Science Foundation grant 3768/19, the United States–Israel Binational Science Foundation (2017253), and the European Union's Horizon 2020 research and innovation program (825821). The contents of this document are the sole responsibility of the iReceptor Plus Consortium and can under no circumstances be regarded as reflecting the position of the European Union.

Authors' contributions

GY, MGT, and TT conceived the research; GY supervised the work; MS performed the computational analyses; ZT prepared and sequenced the BCR libraries; PP coordinated between all the parties and transferred the samples from the hospitals to the lab at Bar Ilan University; SS, MM, HK, YH, AP, EBC, BB collected the samples from COVID-19 patients; ALB, VY collected the samples from healthy volunteers; MS, PP, GY wrote the manuscript; all authors edited the manuscript and approved it for publication.

Acknowledgements

Not applicable.

References

[1] Marco Cascella, Michael Rajnik, Abdul Aleem, Scott C Dulebohn, and Raffaela Di Napoli. Features, evaluation, and treatment of coronavirus (covid-19). *Statpearls [internet]*, 2022.

[2] Christoph Schultheiß, Lisa Paschold, Donjete Simnica, Malte Mohme, Edith Willscher, Lisa von Wenserski, Rebekka Scholz, Imke Wieters, Christine Dahlke, Eva Tolosa, et al. Next-generation sequencing of t and b cell receptor repertoires from covid-19 patients showed signatures associated with severity of disease. *Immunity*, 53(2):442–455, 2020.

[3] Aurélien Sokal, Pascal Chappert, Giovanna Barba-Spaeth, Anais Roeser, Slim Fourati, Imane Azzaoui, Alexis Vandenberghe, Ignacio Fernandez, Annalisa Meola, Magali Bouvier-Alias, et al. Maturation and persistence of the anti-sars-cov-2 memory b cell response. *Cell*, 184(5):1201–1213, 2021.

[4] Mrunal Sakharkar, C Garrett Rappazzo, Wendy F Wieland-Alter, Ching-Lin Hsieh, Daniel Wrapp, Emma S Esterman, Chengzi I Kaku, Anna Z Wec, James C Geoghegan, Jason S McLellan, et al. Prolonged evolution of the human b cell response to sars-cov-2 infection. *Science immunology*, 6(56), 2021.

[5] Christoph Kreer, Matthias Zehner, Timm Weber, Meryem S Ercanoglu, Lutz Gieselmann, Cornelius Rohde, Sandro Halwe, Michael Korenkov, Philipp Schommers, Kanika Vanshylla, et al. Longitudinal isolation of potent near-germline sars-cov-2-neutralizing antibodies from covid-19 patients. *Cell*, 182(4):843–854, 2020.

[6] Jacob D Galson, Sebastian Schaetzle, Rachael JM Bashford-Rogers, Matthew IJ Raybould, Aleksandr Kovaltsuk, Gavin J Kilpatrick, Ralph Minter, Donna K Finch, Jorge Dias, Louisa K James, et al. Deep sequencing of b cell receptor repertoires from covid-19 patients reveals strong convergent immune signatures. *Frontiers in immunology*, page 3283, 2020.

[7] Brian D Corrie, Nishanth Marthandan, Bojan Zimonja, Jerome Jaglale, Yang Zhou, Emily Barr, Nicole Knoetze, Frances MW Breden, Scott Christley, Jamie K Scott, et al. ireceptor: A platform for querying and analyzing antibody/b-cell and t-cell receptor repertoire data across federated repositories. *Immunological reviews*, 284(1):24–41, 2018.

[8] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022.

[9] Matthew IJ Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. Cov-abdab: the coronavirus antibody database. *BioRxiv*, 2020.

[10] Victor Greiff, Gur Yaari, and Lindsay Cowell. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, 2020.

[11] Or Shemesh, Pazit Polak, Knut E. A. Lundin, Ludvig M. Sollid, and Gur Yaari. Machine learning analysis of naïve b-cell receptor repertoires stratifies celiac disease patients and controls. *Frontiers in Immunology*, 12: 633, 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.627813. URL `https://www.frontiersin.org/article/10.3389/fimmu.2021.627813`.

[12] Andrew D Foers, M Saad Shoukat, Oliver E Welsh, Killian Donovan, Russell Petry, Shelley C Evans, Michael EB FitzPatrick, Nadine Collins, Paul Klenerman, Anna Fowler, et al. Classification of intestinal t-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. *The Journal of pathology*, 253(3):279–291, 2021.

[13] Jason A Carter, Jonathan B Preall, Kristina Grigaityte, Stephen J Goldfless, Eric Jeffery, Adrian W Briggs, Francois Vigneault, and Gurinder S Atwal. Single t cell sequencing demonstrates the functional role of $\alpha\beta$ tcr pairing in cell lineage and antigen specificity. *Frontiers in immunology*, 10:1516, 2019.

[14] Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nature genetics*, 49(5):659–665, 2017.

[15] Ramy Arnaout, Nina Luning Prak, Nicholas Schwab, and Florian Rubelt. The future of blood testing is the immunome. *Frontiers in Immunology*, 12:228, 2021.

[16] Milena Pavlović, Lonneke Scheffer, Keshav Motwani, Chakravarthi Kanduri, Radmila Kompova, Nikolay Vazov, Knut Waagan, Fabian LM Bernal, Alexandre Almeida Costa,

Brian Corrie, et al. The immuneml ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nature Machine Intelligence*, 3(11):936–944, 2021.

[17] Miri Ostrovsky-Berman, Boaz Frankel, Pazit Polak, and Gur Yaari. Immune2vec: Embedding b/t cell receptor sequences in r^n using natural language processing. *Frontiers in immunology*, page 2706, 2021.

[18] Rebecca Elyanow, Thomas M. Snyder, Sudeb C. Dalai, Rachel M. Gittelman, Jim Boonyaratanakornkit, Anna Wald, Stacy Selke, Mark H. Wener, Chihiro Morishima, Alex L. Greninger, Michael R. Holbrook, Ian M. Kaplan, H. Jabran Zahid, Jonathan M. Carlson, Lance Baldo, Thomas Manley, Harlan S. Robins, and David M. Koelle. T-cell receptor sequencing identifies prior sars-cov-2 infection and correlates with neutralizing antibody titers and disease severity. *medRxiv*, 2021. doi: 10.1101/2021.03.19.21251426. URL `https://www.medrxiv.org/content/early/2021/03/22/2021.03.19.21251426`.

[19] Rachel M. Gittelman, Enrico Lavezzo, Thomas M. Snyder, H. Jabran Zahid, Rebecca Elyanow, Sudeb Dalai, Ilan Kirsch, Lance Baldo, Laura Manuto, Elisa Franchin, Claudia Del Vecchio, Monia Pacenti, Caterina Boldrin, Margherita Cattai, Francesca Saluzzo, Andrea Padoan, Mario Plebani, Fabio Simeoni, Jessica Bordini, Nicola I. Lorè, Dejan Lazarevic, Daniela M. Cirillo, Paolo Ghia, Stefano Toppo, Jonathan M. Carlson, Harlan S. Robins, Giovanni Tonon, and Andrea Crisanti. Diagnosis and tracking of sars-cov-2 infection by t-cell receptor sequencing. *medRxiv*, 2021. doi: 10.1101/2020.11.09.20228023. URL `https://www.medrxiv.org/content/early/2021/02/10/2020.11.09.20228023`.

[20] Sudeb C Dalai, Jennifer N Dines, Thomas M Snyder, Rachel M Gittelman, Tera Eerkes, Pashmi Vaney, Sally Howard, Kipp Akers, Lynell Skewis, Anthony Monteforte, et al. Clinical validation of a novel t-cell receptor sequencing assay for identification of recent or prior sars-cov-2 infection. *medRxiv*, 2021. doi: https://doi.org/10.1101/2021.01.06.21249345.

[21] M Saad Shoukat, Andrew D Foers, Stephen Woodmansey, Shelley C Evans, Anna Fowler, and Elizabeth J Soilleux. Use of machine learning to identify a t cell response to sars-cov-2. *Cell Reports Medicine*, 2(2):100192, 2021.

[22] Bas Pilzecker and Heinz Jacobs. Mutating for good: Dna damage responses during somatic hypermutation. *Frontiers in immunology*, 10:438, 2019.

[23] Gur Yaari, Jason Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel NH Stern, Kevin O'Connor, David Hafler, Uri Laserson, Francois Vigneault, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in immunology*, 4:358, 2013.

[24] Chaim A Schramm and Daniel C Douek. Beyond hot spots: biases in antibody somatic hypermutation and implications for vaccine design. *Frontiers in immunology*, page 1876, 2018.

[25] Natanael Spisak, Aleksandra M Walczak, and Thierry Mora. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic acids research*, 48(19):10702–10712, 2020.

[26] Thomas MacCarthy, Susan L Kalis, Sergio Roa, Phuong Pham, Myron F Goodman, Matthew D Scharff, and Aviv Bergman. V-region mutation in vitro, in vivo, and in silico reveal the importance of the enzymatic properties of aid and the sequence environment. *Proceedings of the National Academy of Sciences*, 106(21):8629–8634, 2009.

[27] MA Turchaninova, A Davydov, OV Britanova, Mikhail Shugay, Vasileios Bikos, ES Egorov, VI Kirgizova, EM Merzlyak, DB Staroverov, DA Bolotin, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nature protocols*, 11(9):1599–1616, 2016.

[28] Sivan Eliyahu, Oz Sharabi, Shiri Elmedvi, Reut Timor, Ateret Davidovich, Francois Vigneault, Chris Clouser, Ronen Hope, Assy Nimer, Marius Braun, et al. Antibody repertoire analysis of hepatitis c virus infections identifies immune signatures associated with spontaneous clearance. *Frontiers in immunology*, 9:3004, 2018.

[29] Jason A Vander Heiden, Panos Stathopoulos, Julian Q Zhou, Luan Chen, Tamara J Gilbert, Christopher R Bolen, Richard J Barohn, Mazen M Dimachkie, Emma Ciafaloni, Teresa J Broering, et al. Dysregulation of b cell repertoire formation in myasthenia gravis

patients revealed through deep sequencing. *The Journal of Immunology*, 198(4):1460–1473, 2017.

[30] Jason A Vander Heiden, Gur Yaari, Mohamed Uduman, Joel NH Stern, Kevin C O'Connor, David A Hafler, Francois Vigneault, and Steven H Kleinstein. presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, 30(13):1930–1932, 2014.

[31] Xavier Brochet, Marie-Paule Lefranc, and Véronique Giudicelli. Imgt/v-quest: the highly customized and integrated system for ig and tr standardized vj and vdj sequence analysis. *Nucleic acids research*, 36(suppl_2):W503–W508, 2008.

[32] Namita T Gupta, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H Kleinstein. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20):3356–3358, 2015.

[33] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[34] M Kuhn, J Wing, S Weston, A Williams, C Keefer, A Engelhardt, T Cooper, Z Mayer, B Kenkel, and M Benesty. R package caret: Classification and regression training, 2019.

[35] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[36] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

[37] Bing He, Shuning Liu, Yuanyuan Wang, Mengxin Xu, Wei Cai, Jia Liu, Wendi Bai, Shupei Ye, Yong Ma, Hengrui Hu, et al. Rapid isolation and immune profiling of sars-cov-2 specific memory b cell in convalescent covid-19 patients via libra-seq. *Signal transduction and targeted therapy*, 6(1):1–12, 2021.

[38] Prasanti Kotagiri, Federica Mescia, William M Rae, Laura Bergamaschi, Zewen K Tuong, Lorinda Turner, Kelvin Hunter, Pehuén P Gerber, Myra Hosmillo, Christoph Hess, et al. B cell receptor repertoire kinetics after sars-cov-2 infection and vaccination. *Cell reports*, 38(7):110393, 2022.

[39] Yicheng Guo, Kevin Chen, Peter D Kwong, Lawrence Shapiro, and Zizhang Sheng. cab-rep: a database of curated antibody repertoires for exploring antibody diversity and predicting antibody prevalence. *Frontiers in immunology*, 10:2365, 2019.

[40] Matthew I J Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, 37(5): 734–735, 08 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa739. URL `https://doi.org/10.1093/bioinformatics/btaa739`.

[41] Bas Pilzecker and Heinz Jacobs. Mutating for good: Dna damage responses during somatic hypermutation. *Frontiers in Immunology*, 10, 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.00438. URL `https://www.frontiersin.org/article/10.3389/fimmu.2019.00438`.

[42] Sandra CA Nielsen, Fan Yang, Katherine JL Jackson, Ramona A Hoh, Katharina Röltgen, Grace H Jean, Bryan A Stevens, Ji-Yeun Lee, Arjun Rustagi, Angela J Rogers, et al. Human b cell clonal expansion and convergent antibody responses to sars-cov-2. *Cell host & microbe*, 28(4):516–525, 2020.

[43] Yiquan Wang, Meng Yuan, Huibin Lv, Jian Peng, Ian A Wilson, and Nicholas C Wu. A large-scale systematic survey reveals recurring molecular features of public antibody responses to sars-cov-2. *Immunity*, 2022.

[44] Or Shemesh, Pazit Polak, Knut EA Lundin, Ludvig M Sollid, and Gur Yaari. Machine learning analysis of naïve b-cell receptor repertoires stratifies celiac disease patients and controls. *Frontiers in immunology*, page 633, 2021.

[45] Modi Safra, Lael Werner, Pazit Polak, Ayelet Peres, Naomi Salamon, Michael Schvimer, Batia Weiss, Iris Barshack, Dror S Shouval, and Gur Yaari. A somatic hypermutation-

based machine learning model stratifies individuals with crohn's disease and controls. *Genome Research*, pages gr–276683, 2022.

[46] Valerie H Odegard and David G Schatz. Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6(8):573–583, 2006.

[47] Michael Mor, Michal Werbner, Joel Alter, Modi Safra, Elad Chomsky, Smadar Hada-Neeman, Ksenia Polonsky, Cameron J Nowell, Alex E Clark, Anna Roitburd-Berman, et al. Multi-clonal live sars-cov-2 in vitro neutralization by antibodies isolated from severe covid-19 convalescent donors. *BioRxiv*, 2020.

[48] Yongbing Pan, Jianhui Du, Jia Liu, Hai Wu, Fang Gui, Nan Zhang, Xiaojie Deng, Gang Song, Yufeng Li, Jia Lu, et al. Screening of potent neutralizing antibodies against sars-cov-2 using convalescent patients-derived phage-display libraries. *Cell Discovery*, 7(1): 1–19, 2021.

[49] Roy A Ehling, Cédric R Weber, Derek M Mason, Simon Friedensohn, Bastian Wagner, Florian Bieberich, Edo Kapetanovic, Rodrigo Vazquez-Lombardi, Raphaël B Di Roberto, Kai-Lin Hong, et al. Sars-cov-2 reactive and neutralizing antibodies discovered by single-cell sequencing of plasma cells and mammalian display. *Cell reports*, 38(3):110242, 2022.

[50] JQ Zhou and SH Kleinstein. Position-dependent differential targeting of somatic hypermutation. the journal of immunology. *ji*, 2000496, 2020.

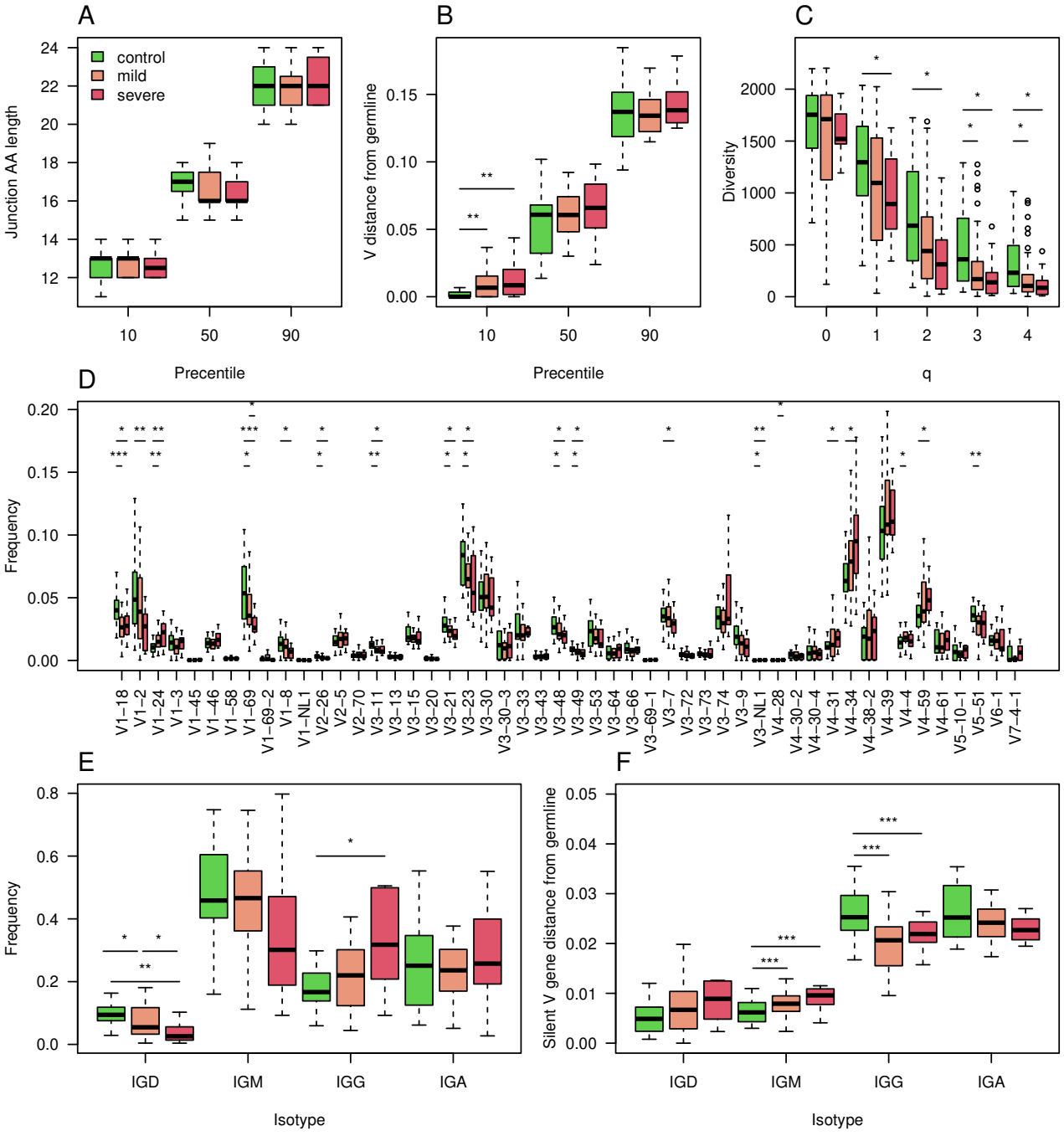


Figure 1: **Characterization of the COVID-19 heavy chain BCR cohort** A. 10,50 and 90 percentiles of AA CDR3 length in individuals with corona at indicated severity and controls. B. 10,50 and 90 percentiles of V gene distances from germline in COVID-19 infected individuals at indicated severity and controls.C. Boxplot showing calculated Hill diversity indexes upon different q values between individuals infected by COVID-19 at indicated severity and controls. D. Boxplots showing V gene usage in individuals infected by COVID-19 at indicated severity and controls, shown top 50's mean frequencies. E. Boxplots showing the isotype frequencies in individuals infected by COVID-19 at indicated severity and controls. F. Boxplots showing silent mutations' frequencies along the V gene in different isotypes of individuals infected by COVID-19 at indicated severity and controls . In the whole figure, * marks P value less than 0.05. ** marks P value less than 0.01 and *** marks P value less than 0.001.

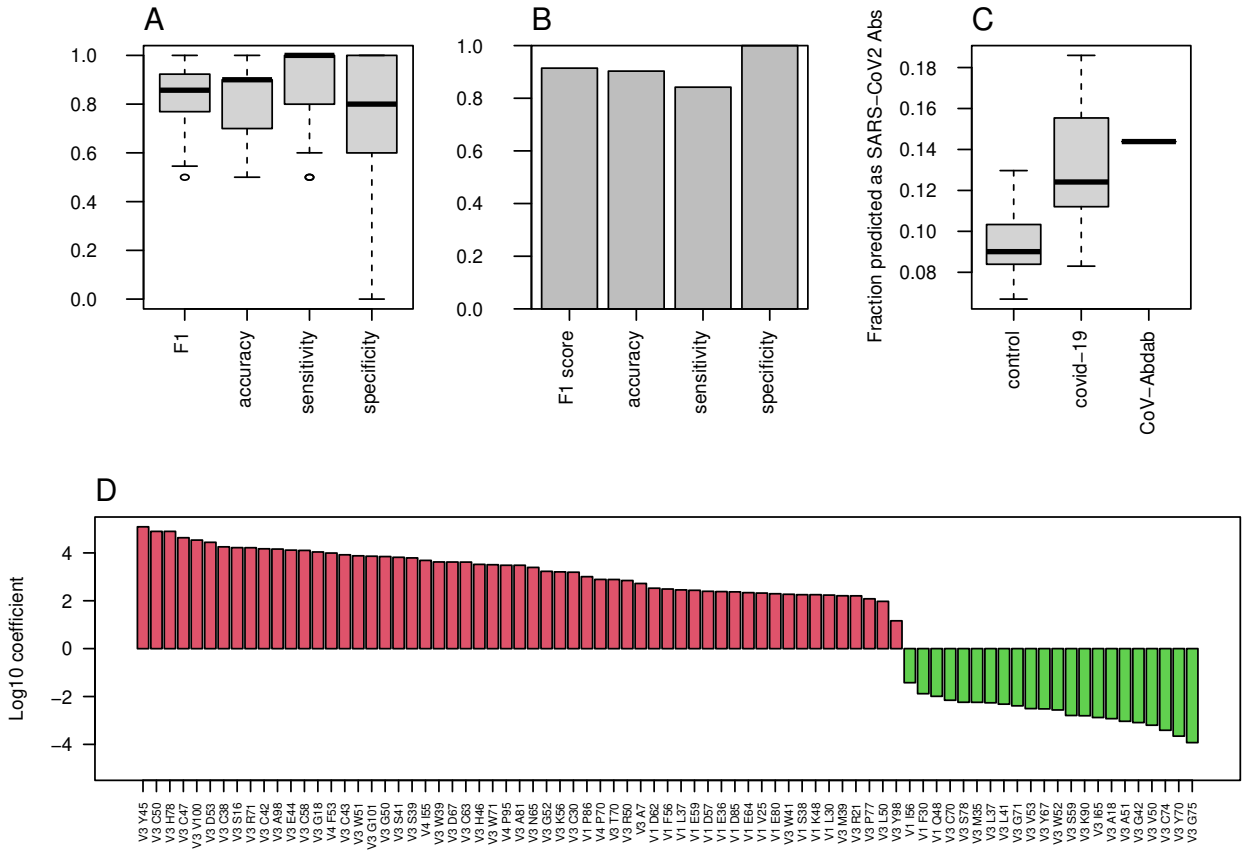


Figure 2: **COVID-19 classification using AA frequencies at all V gene positions** A. Boxplots showing the F1 score, accuracy, sensitivity, and specificity for COVID-19 classification by AA frequency at each position in each V family. Shown are values calculated for 50 random splits to train and validation groups. B. Bar plots showing the indicated scores on the external test group. C. COVID-19 single antibody scores were calculated using the coefficients of the algorithm described in panel A. Boxplos showing the fraction of antibody sequences with scores above 0 in control and COVID-19 infected repertoires, as well as in CoV-AbDab COVID-19 antibodies, are shown. D. Log10 coefficients of the algorithm described in A and B.

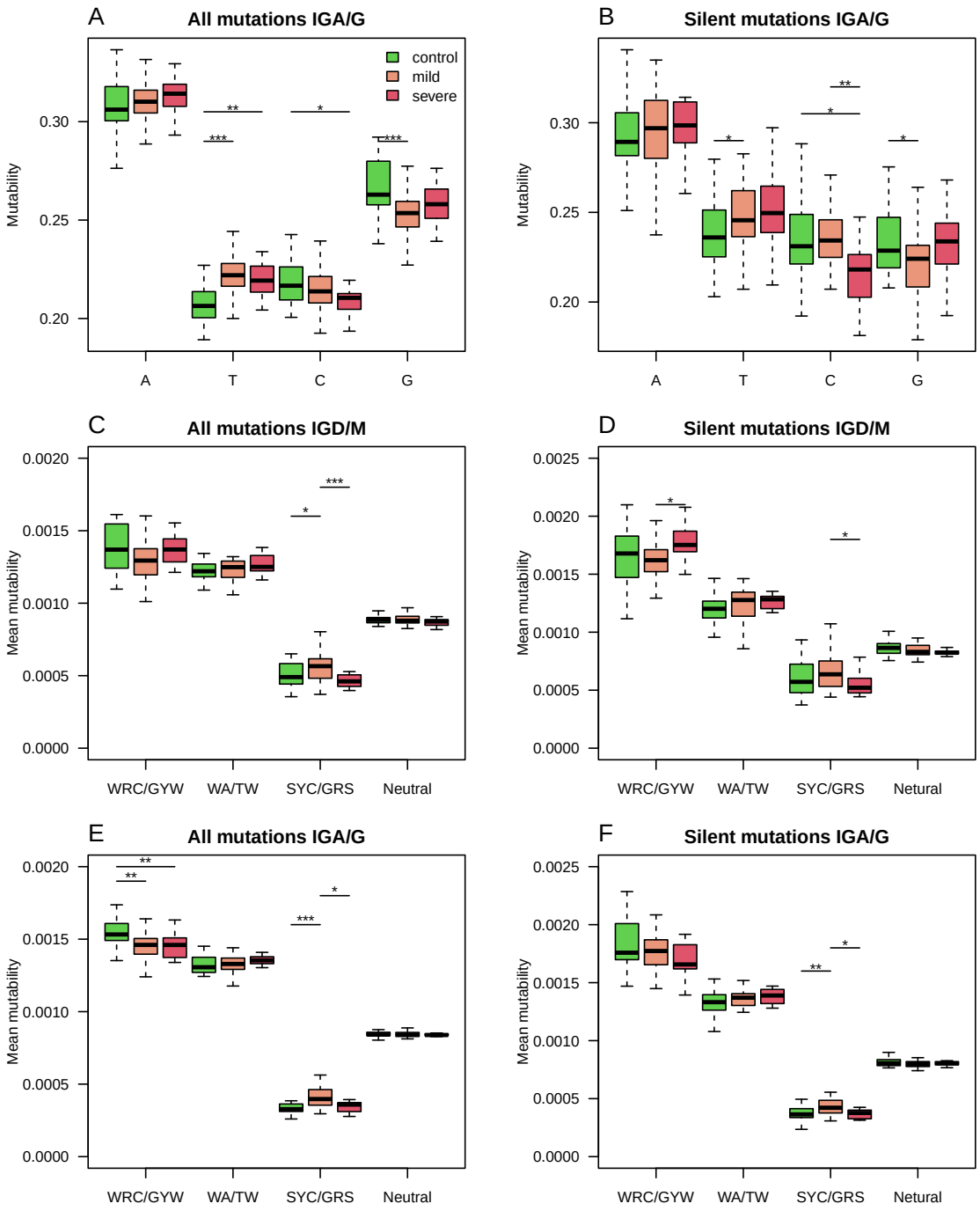


Figure 3: **Silent and replacement mutability in SHM single base mutability, 5-mers hot-spots and cold-spots**

A. A single base mutability model was built based on IGA/G isotypes of COVID-19 patients and controls. Shown are boxplots representing the normalized sum of single base mutability. B. The same plot as in A but for silent mutations only. C-D. An 5-mer SHM model based on both silent and replacement mutations in C, or silent only mutations in D, was built using the IGD and IGM isotypes of COVID-19 patients at different severity levels and controls. Shown mutability of the two known SHM hot-spots, SHM cold-spots, and the rest of the sites. E-F. An 5-mer SHM model based on both silent and replacement mutations in E, or silent only mutations in F, was built using the IGA and IGG isotypes of COVID-19 patients at different severity levels and controls. Shown mutability of the two known SHM hot-spots, SHM cold-spots, and the rest of the sites. In the whole figure, * marks P value less than 0.05. ** marks P value less than 0.01 and *** marks P value less than 0.001.

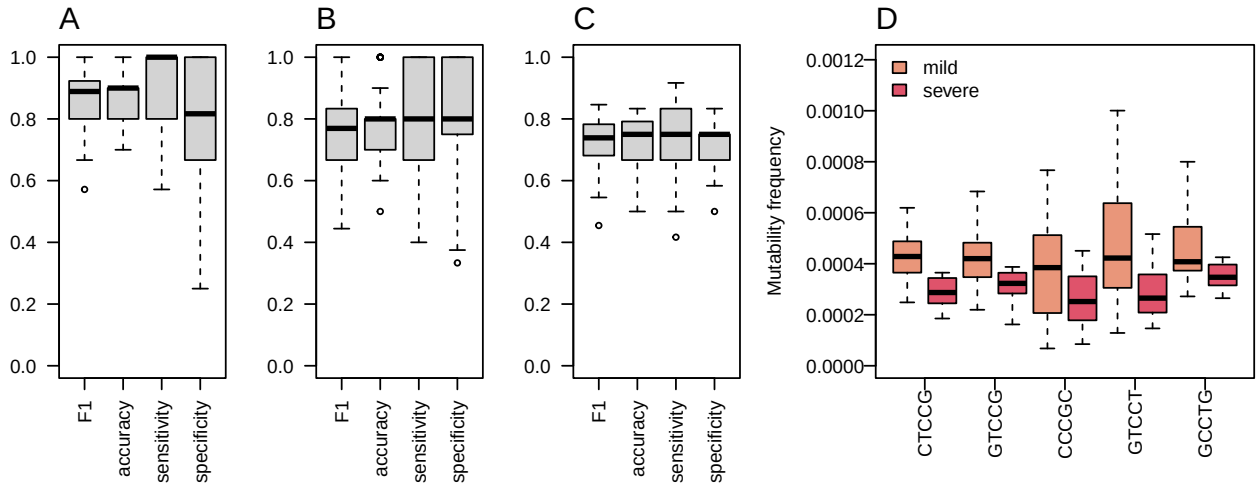


Figure 4: **SHM Heavy chain enables classification of both SARS-CoV2 infection and COVID-19 severity**

A. An ML algorithm was trained on the substitutions matrix of the 5-mer SHM model, which was created for the IGA/G isotypes. Boxplots representing F1 score, accuracy, specificity, and sensitivity of 50 random splits to train and test groups are shown. B. The same algorithm as in A was trained on silent mutations only. Shown are Boxplots representing the F1 score, accuracy, specificity, and sensitivity of 50 random splits to train and test groups. C. Boxplots showing F1 score, accuracy, specificity, and sensitivity of 20 leave-one-out cross validation of severity classification. Each leave-one-out was on 12 severe COVID-19 patients and 12 randomly selected mild COVID-19 patients. The ML algorithm was trained on the mutability matrix of the SHM cold-spots in these groups. D. Frequency of mutability in mild and severe individuals with COVID-19. Boxplots of frequencies of repeating coefficients of the algorithm explained in C are shown.

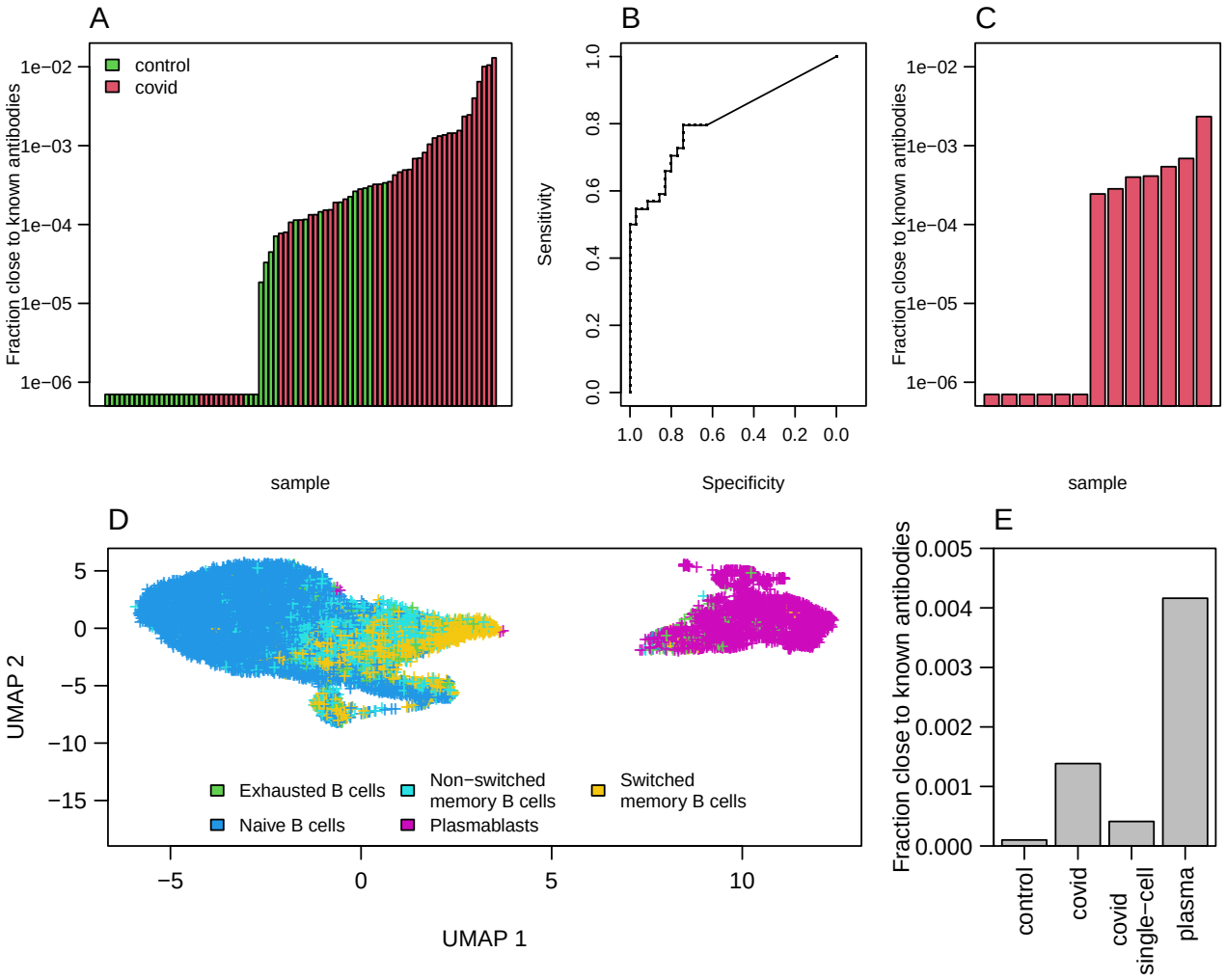
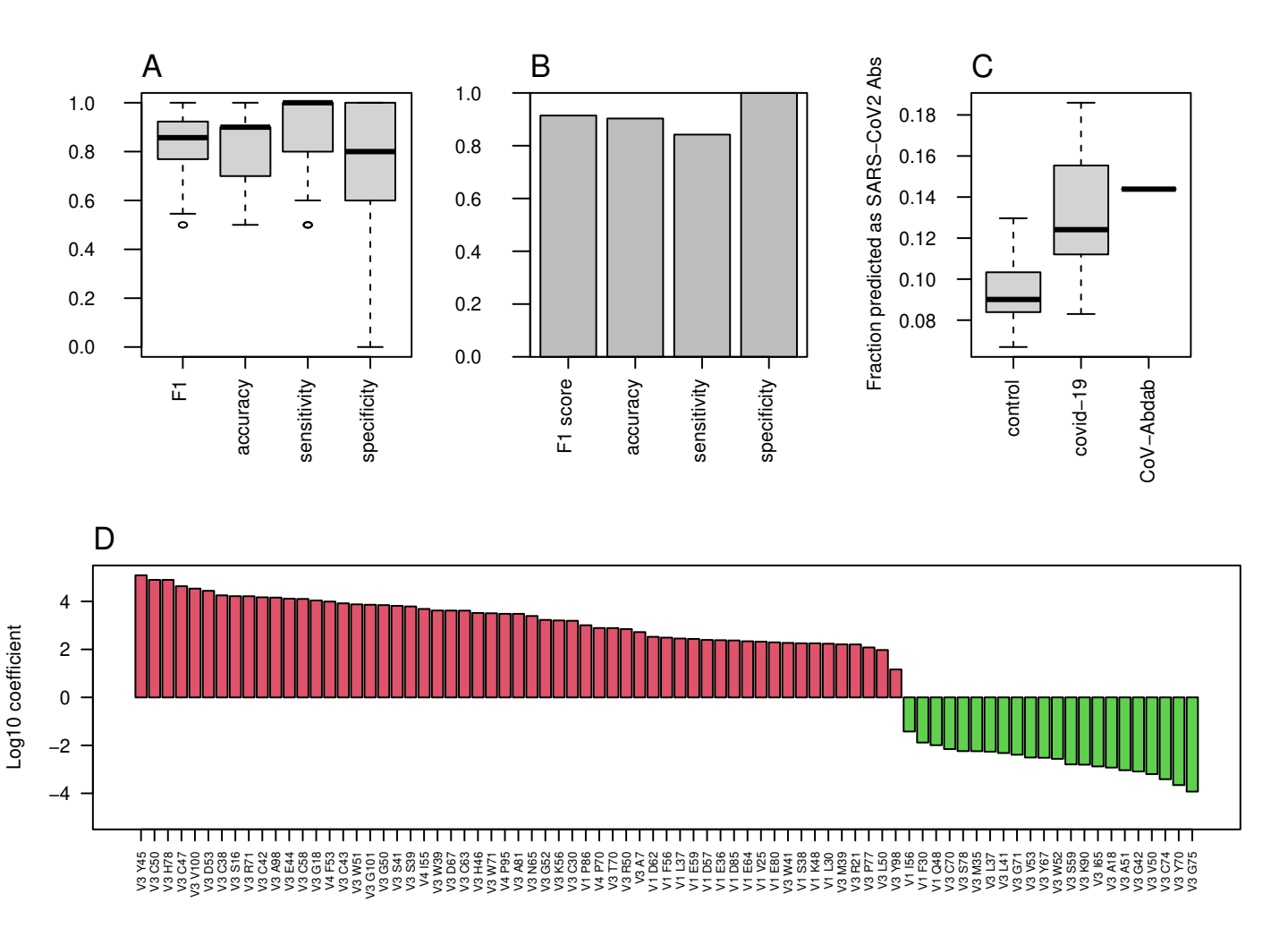
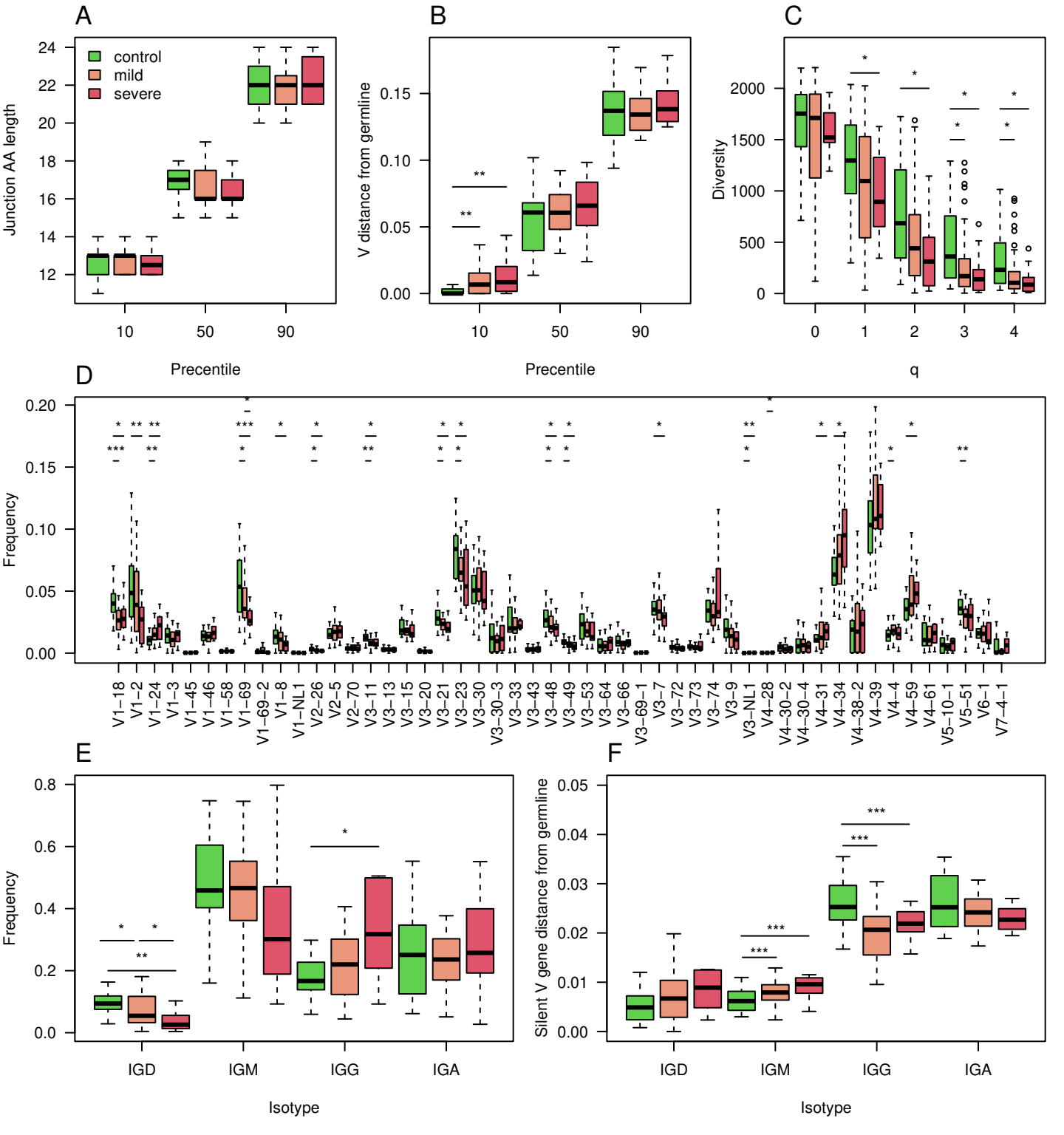
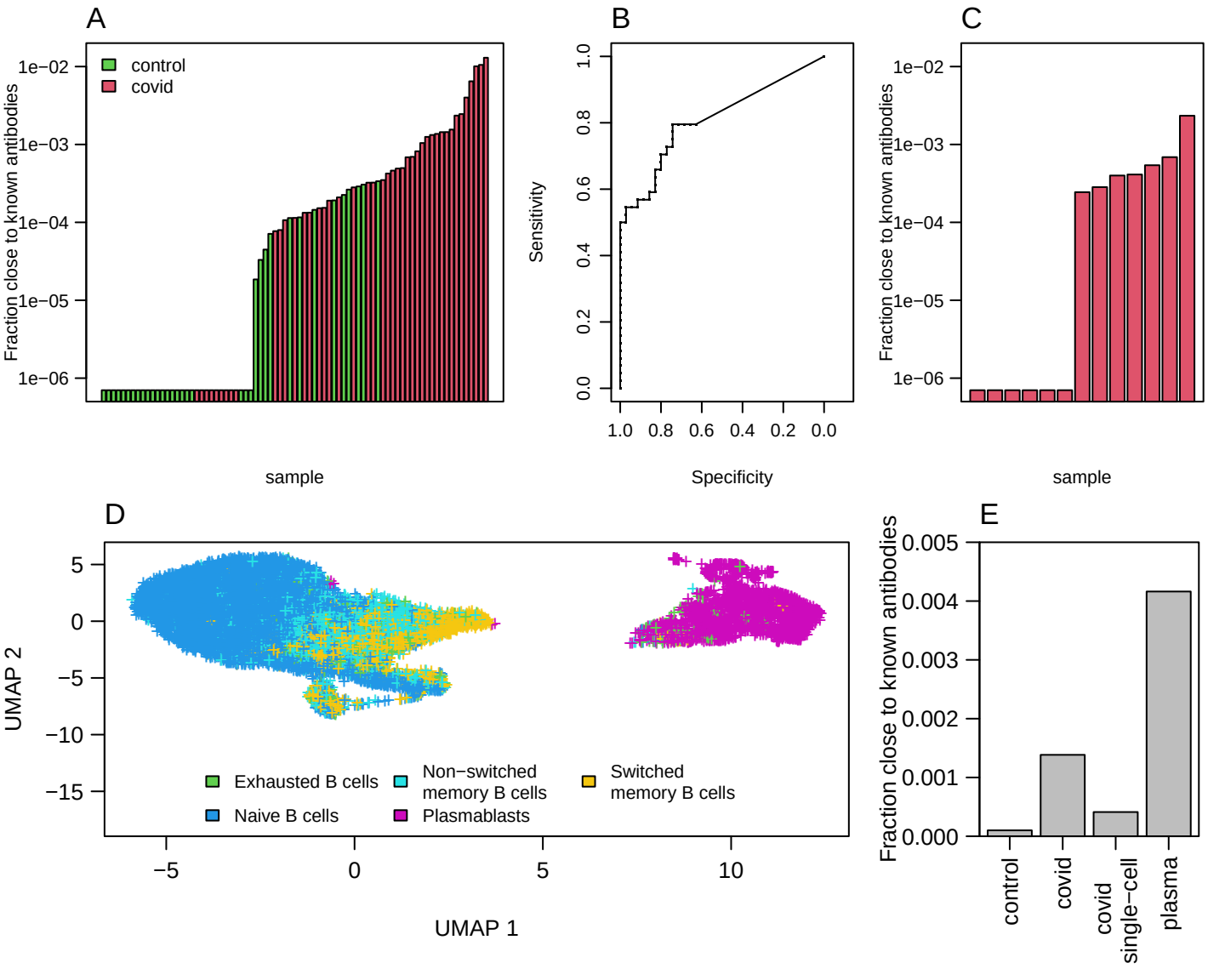


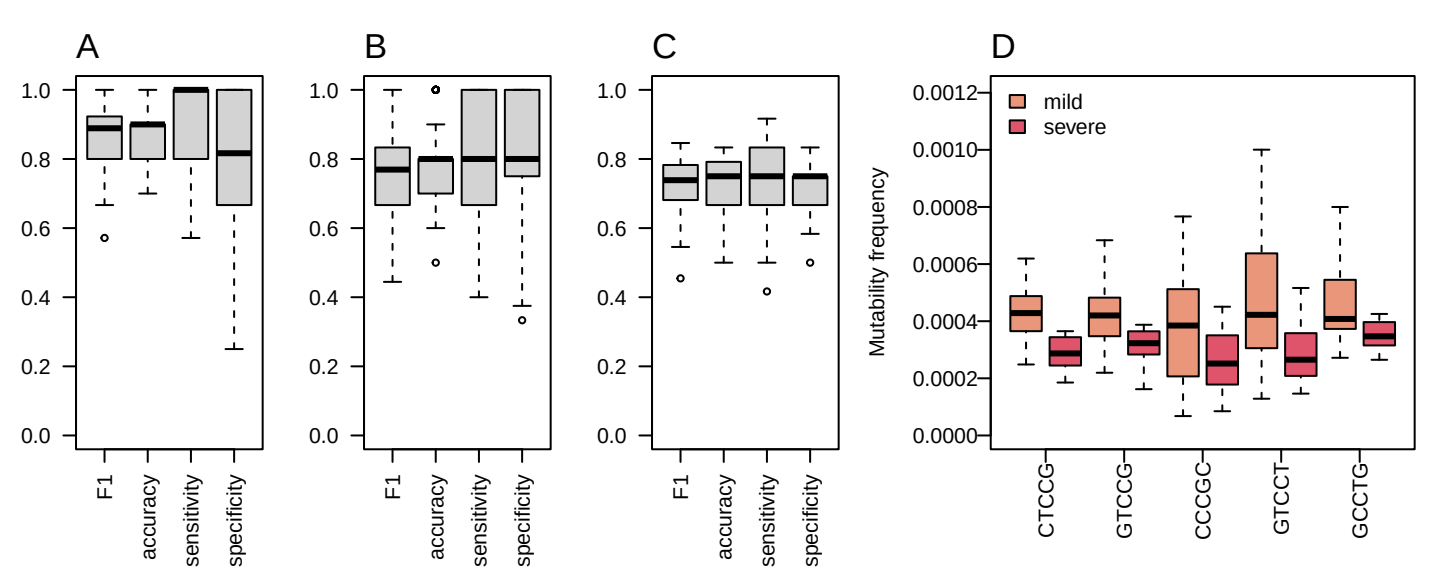
Figure 5: **Clones of antibodies in our sequencing close to known COVID-19 antibodies from CoV-AbDab database.**

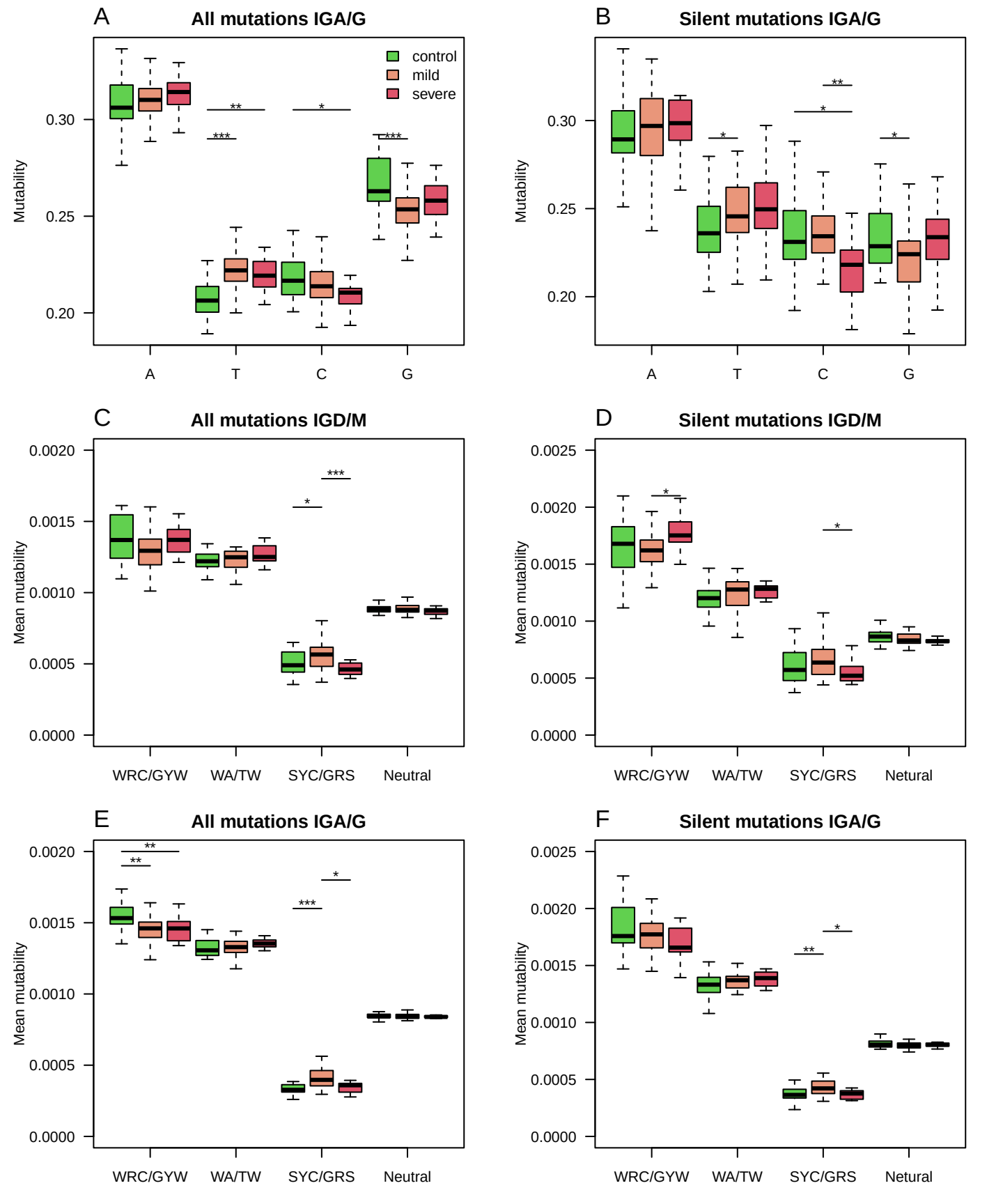
A. Sum of frequencies of clones (same V and J genes and 85% similarity in AA of CDR3) close to known COVID-19 antibodies (from CoV-AbDab data base) in COVID-19 patients and controls. B. ROC curve summarizing the results shown in A. C. Sum frequencies of clones close to COVID-19 antibodies in 13 single cell COVID-19 patients data. D. UMAP on gene expressions of B cells isolated from 13 patients showing differences between naive, memory and plasmablast cells. Cell type identification was done using SinglR. E. Sum of frequencies of antibodies close to known COVID-19 antibodies in bulk sequencing of COVID-19 patients and control as well as in sequences from single cell sequences of COVID-19 patients and in cells identified as plasmablast cells.


A
1.0
0.8
0.6
0.4
0.2
0.0
F1
accuracy
sensitivity
specificity
B
1.0
0.8
0.6
0.4
0.2
0.0
F1 score
accuracy
sensitivity
specificity
C
Fraction predicted as SARS−CoV2 Abs
0.18
0.16
0.14
0.12
0.10
0.08
control
covid−19
CoV−Abdab
D
Log10 coefficient
4
2
0
−2
−4
V3 Y45
V3 C50
V3 H78
V3 C47
V3 V100
V3 D53
V3 C38
V3 S16
V3 R71
V3 C42
V3 A98
V3 E44
V3 C58
V3 G18
V4 F53
V3 C43
V3 W51
V3 G101
V3 G50
V3 S41
V3 S39
V4 I55
V3 W99
V3 D67
V3 C63
V3 H46
V3 W71
V4 P95
V3 A81
V3 N65
V3 G52
V3 K56
V3 C30
V1 P86
V4 P70
V3 T70
V3 R50
V3 A7
V1 D62
V1 F56
V1 L37
V1 E59
V1 D57
V1 E36
V1 D85
V1 E64
V1 V25
V1 E80
V3 W41
V1 S38
V1 K48
V1 L30
V3 M39
V3 R21
V3 P77
V3 L50
V3 Y98
V1 I56
V1 F30
V1 Q48
V3 C70
V3 S78
V3 M35
V3 L37
V3 L41
V3 G71
V3 V53
V3 Y67
V3 W52
V3 S59
V3 K90
V3 I65
V3 A18
V3 A51
V3 G42
V3 V50
V3 C74
V3 Y70
V3 G75


A
B
C
control
mild
severe
Junction AA length
V distance from germline
Diversity
Precentile
Precentile
q
D
Frequency
V1−18
V1−2
V1−24
V1−3
V1−45
V1−46
V1−58
V1−69
V1−69−2
V1−8
V1−NL1
V2−26
V2−5
V2−70
V3−11
V3−13
V3−15
V3−20
V3−21
V3−23
V3−30
V3−30−3
V3−33
V3−43
V3−48
V3−49
V3−53
V3−64
V3−66
V3−69−1
V3−7
V3−72
V3−73
V3−74
V3−9
V3−NL1
V4−28
V4−30−2
V4−30−4
V4−31
V4−34
V4−38−2
V4−39
V4−4
V4−59
V4−61
V5−10−1
V5−51
V6−1
V7−4−1
E
F
Frequency
Silent V gene distance from germline
IGD
IGM
IGG
IGA
Isotype
Isotype


A
control
covid
Fraction close to known antibodies
1e−02
1e−03
1e−04
1e−05
1e−06
sample
B
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
C
Fraction close to known antibodies
sample
D
UMAP 2
UMAP 1
Exhausted B cells
Naive B cells
Non−switched memory B cells
Plasmablasts
Switched memory B cells
E
Fraction close to known antibodies
0.005
0.004
0.003
0.002
0.001
0.000
control
covid
covid single−cell
plasma


A
B
C
D
1.0
0.8
0.6
0.4
0.2
0.0
F1
accuracy
sensitivity
specificity
Mutability frequency
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
mild
severe
CTCCG
GTCCG
CCCGC
GTCCT
GCCTG


A
All mutations IGA/G
control
mild
severe
Mutability
0.30
0.25
0.20
**
*
***
***
A
T
C
G
B
Silent mutations IGA/G
Mutability
0.30
0.25
0.20
**
*
*
*
A
T
C
G
C
All mutations IGD/M
Mean mutability
0.0020
0.0015
0.0010
0.0005
0.0000
*
***
WRC/GYW
WA/TW
SYC/GRS
Neutral
D
Silent mutations IGD/M
Mean mutability
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
*
*
WRC/GYW
WA/TW
SYC/GRS
Netural
E
All mutations IGA/G
Mean mutability
0.0020
0.0015
0.0010
0.0005
0.0000
**
**
***
*
WRC/GYW
WA/TW
SYC/GRS
Neutral
F
Silent mutations IGA/G
Mean mutability
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
**
*
WRC/GYW
WA/TW
SYC/GRS
Netural