frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 19 April 2023
DOI 10.3389/fimmu.2023.1031914

OPEN ACCESS

EDITED BY
Pei-Hui Wang,
Shandong University, China

REVIEWED BY
Valentyn Oksenych,
University of Oslo, Norway
Youpeng Fan,
Southwest University, China

*CORRESPONDENCE
Gur Yaari
gur.yaari@biu.ac.il

SPECIALTY SECTION
This article was submitted to Viral Immunology, a section of the journal Frontiers in Immunology

RECEIVED 30 August 2022
ACCEPTED 22 March 2023
PUBLISHED 19 April 2023

CITATION
Safra M, Tamari Z, Polak P, Shiber S, Matan M, Karameh H, Helviz Y, Levy-Barda A, Yahalom V, Peretz A, Ben-Chetrit E, Brenner B, Tuller T, Gal-Tanamy M and Yaari G (2023) Altered somatic hypermutation patterns in COVID-19 patients classifies disease severity. *Front. Immunol.* 14:1031914. doi: 10.3389/fimmu.2023.1031914

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[6], Yigal Helviz[7], Adva Levy-Barda[8], Vered Yahalom[9], Avi Peretz[5,10], Eli Ben-Chetrit[11], Baruch Brenner[4,12], Tamir Tuller[13], Meital Gal-Tanamy[10] 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]Jesselson Integrated Heart Center, Shaare Zedek Medical Center, Hebrew University School of Medicine, Jerusalem, Israel, [7]Intensive Care Unit, Shaare Zedek Medical Center, Hebrew University School of Medicine, Jerusalem, Israel, [8]Biobank, Department of Pathology, Rabin Medical Center-Belinson Campus, Petah Tikva, Israel, [9]Blood Services and Apheresis Institute, Rabin Medical Center, Petah Tikva, Israel, [10]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel, [11]Infectious Diseases Unit, Shaare Zedek Medical Center, Hebrew University School of Medicine, Jerusalem, Israel, [12]Institute of Oncology, 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

Introduction: The success of the human body in fighting SARS-CoV2 infection relies on lymphocytes and their antigen receptors. Identifying and characterizing clinically relevant receptors is of utmost importance.

Methods: 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-CoV2 compared with uninfected controls.

Results: 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.

Discussion: 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-CoV2 is adaptive immune receptor repertoire sequencing (AIRR-seq) (2–4), 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-CoV2 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, 14), cytomegalovirus (15), and others (16). 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, 17, 18). 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 (19–22). Recently, for example, a new approach to detect SARS-CoV2 infection by TCR sequencing has been FDA approved for clinical use (21).

B cell development involves three major steps: V(D)J recombination, affinity maturation, and class switch recombination. V(D)J recombination is the process by which B cells generate a diverse array of receptors (BCRs). This process involves a random selection and rearrangement of gene segments called variable (V), diversity (D) and joining (J). The recombination of these segments leads to the creation of a diverse array of receptors that can respond to a wide range of pathogens (23, 24). 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. SHM is a mechanism by which B cells can rapidly diversify the antigen-binding regions of their receptors. During SHM, different enzymatic pathways orchestrate together to introduce mutations specifically in the genomic regions encoding the BCR (25). These mutations can result in altered affinity towards antigens. The repeated cycles of SHM and affinity-dependent selection lead to the generation of high-affinity B cells capable of recognizing and responding to diverse antigens. While selection depends on better binding, the SHM mechanism is independent of pathogen affinity. Extensive investigations have been devoted to understanding the SHM mechanism (26–29), but to the best of our knowledge, no connection of a specific infection to a specific SHM pathway or pattern was made. The mature B cell can, after activation, undergo class switch recombination. This allows mature B cells to switch the isotype of their heavy chain, leading to the production of different classes of secreted BCRs (antibodies) with different effector functions. Following V(D)J recombination, affinity maturation, and class switching, the antigen-specific B cell can become a memory B cell, i.e., a long-lived B cell that retains a "memory" of previous encounters with antigens, allowing for a quicker and more effective response upon re-exposure. It can also become a plasmablast, which is a quickly dividing B cell that secretes antibodies, and later on become a plasma cell, which is a fully mature B cell that secretes large amounts of antibodies (23).

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 (22). 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-CoV2 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-CoV2 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.

Library preparation

Bulk: Ig repertoires were bulk sequenced according to the method described in detail in (30). Briefly, PBMCs were purified using Lymphoprep (Axis Shield), according to the manufacturers' instructions. RNA was extracted using a Direct-zol RNA miniprep kit (Zymo Research, R2050) according to the manufacturer's instructions. RNA was reverse-transcribed using an oligo dT primer. An adaptor sequence was added to the 3' end, which contains a universal priming site and a 17-nucleotide unique molecular identifier. Products were purified, followed by PCR using primers targeting the different BCR isotypes and the universal adaptor. PCR products were then purified using AMPure XP beads. A second PCR was performed to add the Illumina P5 adaptor to the constant region end, and a sample-indexed P7 adaptor to the universal adaptor. Final products were purified, quantified with a TapeStation (Agilent Genomics), and pooled in equimolar proportions, followed by 2×300 paired-end sequencing with a 20% PhiX spike on the Illumina MiSeq platform according to the manufacturer's recommendations. 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 (14), and four from dataset (31).

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 (32), and aligned to IMGT IGHV/D/J genes (33) 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 (34). All P values were calculated using Wilcox test and adjusted using the Benjamini-Hochberg procedure (35).

Generating an SHM model

A 5-mer SHM model was built using the function createTargetingModel from the shazam R package (29), 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 (WRC/GTW and WA/TW 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 (36) 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

10.3389/fimmu.2023.1031914

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 (37). 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 (38). 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 Figure 1 and in Figure S1 for heavy chains, and for kappa and lambda light chains in Figures S2 and S3.

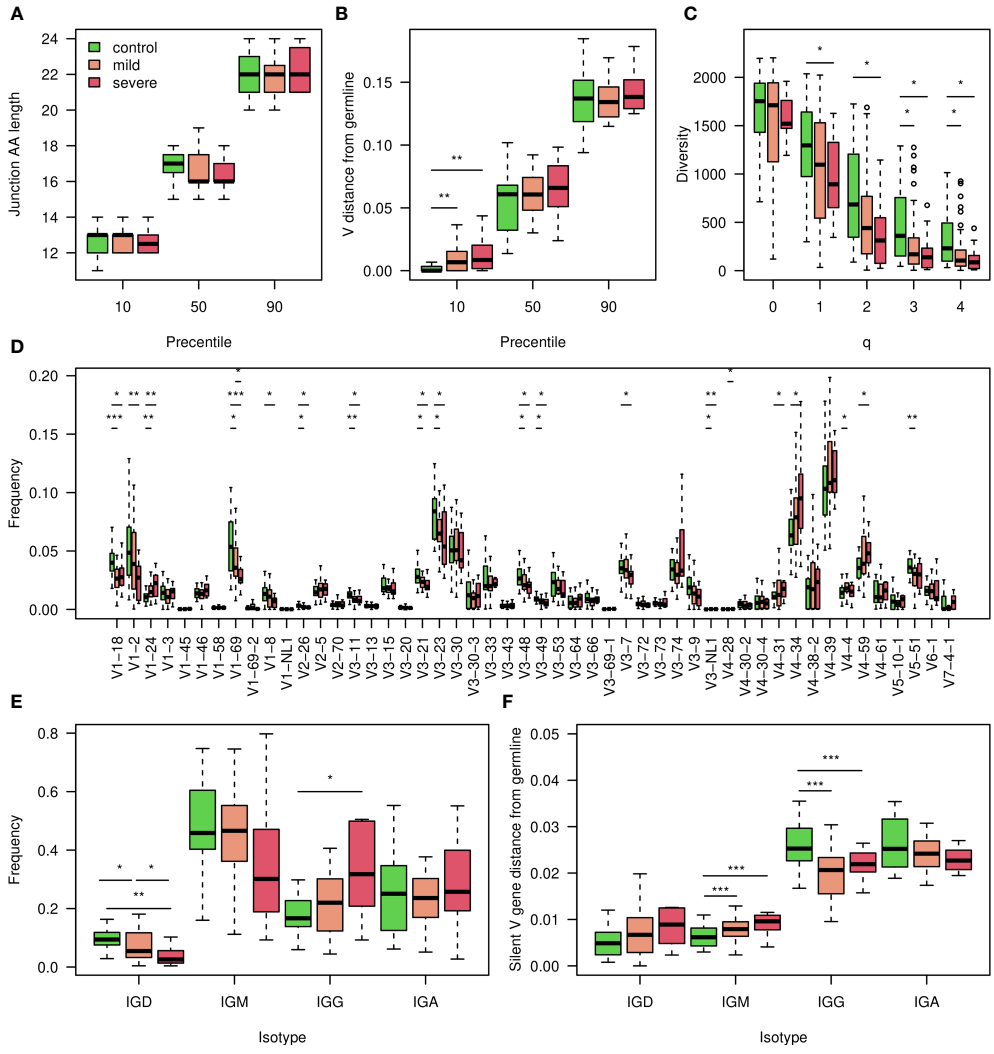

FIGURE 1
Characterization of the COVID-19 heavy chain BCR cohort. **(A)** 10,50 and 90 percentiles of AA CDR3 length in individuals with corona at indicated severity and controls. **(B)** 10,50 and 90 percentiles of V gene distances from germline in COVID-19 infected individuals at indicated severity 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.

10.3389/fimmu.2023.1031914

As expected, the diversity of BCR clones is significantly lower in COVID-19 patients compared with controls (Figure 1C). No significant difference was observed in CDR3 AA length (Figure 1A), and only slight increase was seen inV gene mutation distribution (Figure 1B). For many V genes we observed significantly reduced usage in COVID-19 patients (Figure 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 (39, 40), 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 (Figure 1E), and those of IGA remained unchanged. We also measured silent mutability frequencies for each isotype (Figure 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 (Figure 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 (Figure 2A). The test set resulted in an F1 score of above 0.85 (Figure 2B). We then extracted the coefficient used by the algorithm, corresponding to the contribution of each AA frequency to the classification of the disease (Figure 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 (41). 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.

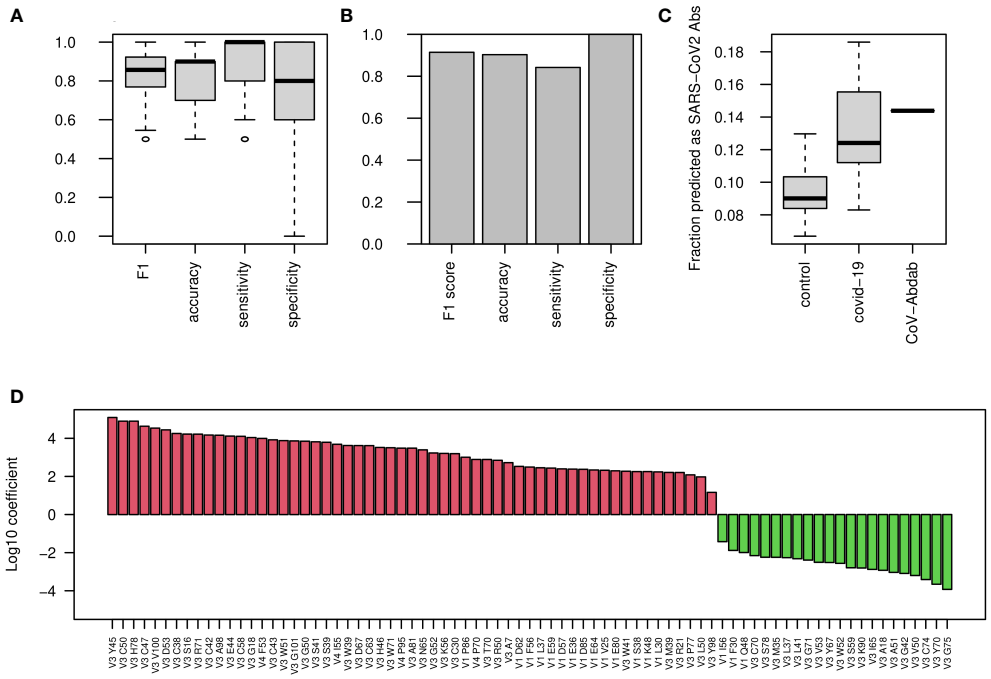

FIGURE 2
COVID-19 classification using AA frequencies at all V gene positions. **(A)** Boxplots showing the F1 score, accuracy, sensitivity, and specificity for COVID-19 classification by AA frequency at each position in each V family. Shown are values calculated for 50 random splits to train and validation 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, B)**.

10.3389/fimmu.2023.1031914

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 (9). 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 (Figure 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 Figure 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 (Figure 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),

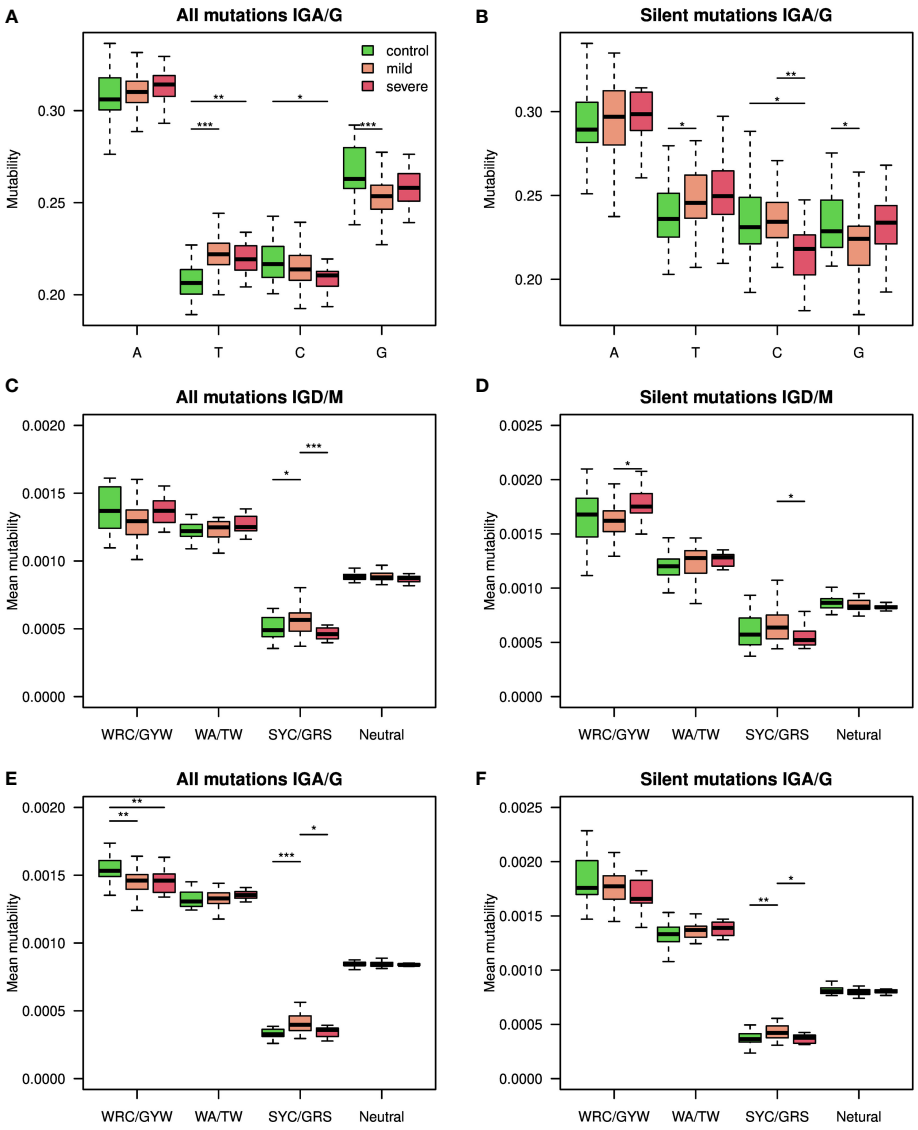

FIGURE 3
Silent and replacement mutability in SHM single base mutability, 5-mers hot-spots and cold-spots. **(A)** A single base mutability model was built based on IGA/G isotypes of COVID-19 patients and controls. Shown are boxplots representing the normalized sum of single base mutability. **(B)** 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.

10.3389/fimmu.2023.1031914

but increases in Adenine and Thymine (A and T). The same results were obtained when considering silent mutations only (Figure 3B). Five main pathways are responsible for introducing mutations during SHM (12). Three introduce mutations in C and G, and the other two involve the low fidelity DNA polymerase pol η, 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 (26). In general, two highly mutated hot-spot motifs are commonly observed in SHM. One is WRC/GYW (where W = {A, T}, Y = {C, T} R = {G, A}, and the mutated position is underlined), and the other is WA/TW. In addition, SYC/GRS (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 (Figures 3C, E). The most significant changes between the different groups are a decrease in the WRC/GYW site and an increase in SYC/GRS 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. Figures 3D, F shows the resulting mutability scores for the same sequence motifs. The observed pattern resembles the one observed in Figures 3C, 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 (Figure S4). Moreover, using SHM matrices based only on a specific V family resulted in a much lower signal (Figure 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 (Figures 4A, S5, S6A, S8), and one using silent mutations only (Figures 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 ~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 (Figure S7A, B). A model based on the combination of light and heavy chains does not obtain better results than using the heavy chain only (Figure 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 (Figure 4C).

All patterns with non-zero coefficients have much higher mutability frequencies in mild patients compared with severe patients (Figure 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

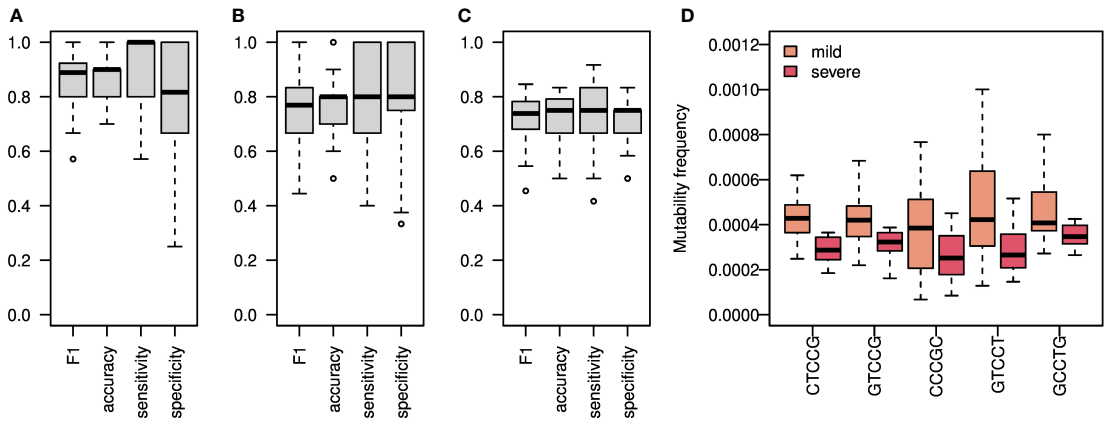

FIGURE 4
SHM Heavy chain enables classification of both SARS-CoV2 infection and COVID-19 severity. **(A)** An ML algorithm was trained on the substitutions matrix of the 5-mer SHM model, which was created for the IGA/G isotypes. Boxplots representing F1 score, accuracy, specificity, and sensitivity of 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.

10.3389/fimmu.2023.1031914

account only one representative from each clone. As shown in Figure 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 (Figure 5A. Histograms summarizing the sizes and numbers of samples having at least one representation in the clones can be found in Figures S9A, 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 (Figure 5B). Even lower results were obtained when training the algorithm to count the frequencies of shared clones between samples (Figure 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 (Figure 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.

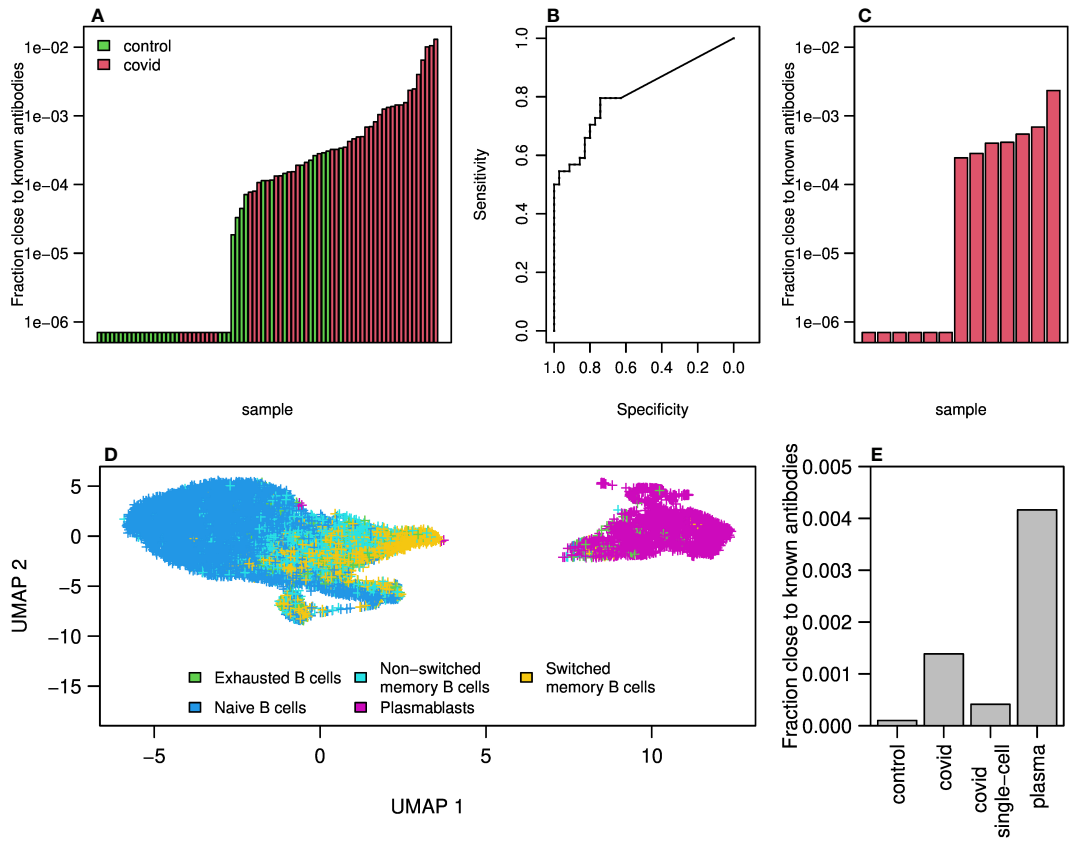

FIGURE 5
Clones of antibodies in our sequencing close to known COVID-19 antibodies from CoV-AbDab database. **(A)** Sum of frequencies of clones (same V and J genes and 85 % similarity in AA of CDR3) close to known COVID-19 antibodies (from CoV-AbDab data base) in COVID-19 patients and controls. **(B)** 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 SingIR. **(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.

We then applied the SingleR R package to classify cell types by single cell expression profiles. Two-dimensional UMAP reduced plots are shown in Figure 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 Figure 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 WRC/GYW hot-spots mutability scores decrease upon COVID-19 infection, and SYC/GRS cold-spots increase (Figure 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 WRC/GYW hot-spots in plasmablasts (0.00168) compared with the other B cell types (0.00178), and an increase in the mutability scores of the SYC/GRS cold-spots (0.0003 and 0.0002, respectively).

Discussion

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

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

In this study we collected and sequenced the BCR repertoires of 51 SARS-CoV2 infected individuals as well as 28 control ones. We do not have information about the SARS-CoV2 strains in which patients were infected by, 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. Here, in contrast to previous reports (22), we were able to stratify COVID-19 patients and healthy individuals based on shared clusters of BCR sequences. The moderate classification results of such approach led us to explore different sets of features that turned out to be more informative. AA frequencies at all V gene positions served as a basis for an ML model that produced a high F1 score ($\sim$ 85 %) in classifying COVID-19 infection.

The patterns of AA alterations in BCRs arise during the process of affinity maturation, that includes two iterative processes, namely SHM and affinity-dependent selection. These patterns can stem from the antibodies against SARS-CoV2 or from overall altered SHM mechanism in COVID-19 patients.

An important question that may arise when inspecting the presented approach is whether it is specific to COVID-19, or perhaps it simply detects general signals related to an adaptive 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 (12, 14, 18), 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. To test this, we applied our algorithm to data from $\sim$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 (29, 44). Yet, except a recent publication about Crohn's disease (45), 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 bindingSARS-CoV2 are very close to the germline (6, 46–48). 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 (Figure 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 particular, 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 pol η.

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/neTW 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 (49), 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% (Figure 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 (Figure 2C). As such, the features (detailed in Figure 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.

Data availability statement

The data presented in the study are deposited in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) repository, accession number PRJNA839749.

Ethics statement

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. The patients/participants provided their written informed consent to participate in this study.

Author contributions

GY, MG-T, 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, EB-C, and BB collected the samples from COVID-19 patients; A-LB and VY collected the samples from healthy volunteers; MS, PP, and GY wrote the manuscript; all authors edited the manuscript. All authors contributed to the article and approved the submitted version.

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), the Bar Ilan Data Science Institute and Israeli Council for Higher Education grant, and the

European Union's Horizon 2020 research and innovation program (825821).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1031914/full#supplementary-material

References

1. Cascella M, Rajnik M, Aleem A, Dulebohn SC, Di Napoli R. *Features, evaluation, and treatment of coronavirus (COVID-19)*. Treasure Island, FL, USA: Statpearls (2022).

2. Sakharkar M, Rappazzo CG, Wieland-Alter WF, Hsieh C-L, Wrapp D, Esterman ES, et al. Prolonged evolution of the human b cell response to SARS-CoV-2 infection. *Sci Immunol* (2021) 6:eabg6916. doi: 10.1126/sciimmunol.abg6916

3. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, et al. Next-generation sequencing of t and b cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. *Immunity* (2020) 53:442–55. doi: 10.1016/j.immuni.2020.06.024

4. Sokal A, Chappert P, Barba-Spaeth G, Roeser A, Fourati S, Azzaoui I, et al. Maturation and persistence of the anti-SARS-CoV-2 memory b cell response. *Cell* (2021) 184:1201. doi: 10.1016/j.cell.2021.01.050

5. Galson JD, Schaetzle S, Bashford-Rogers RJ, Raybould MI, Kovaltsuk A, Kilpatrick GJ, et al. Deep sequencing of b cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front Immunol* (2020) 11:3283. doi: 10.3389/fimmu.2020.605170

6. Kreer C, Zehner M, Weber T, Ercanoglu MS, Gieselmann L, Rohde C, et al. Longitudinal isolation of potent near-germline SARS-CoV-2-neutralizing antibodies from COVID-19 patients. *Cell* (2020) 182:843–54. doi: 10.1016/j.cell.2020.06.044

7. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, et al. Ireceptor: A platform for querying and analyzing antibody/b-cell and t-cell receptor repertoire data across federated repositories. *Immunol Rev* (2018) 284:24–41. doi: 10.1111/imr.12666

8. Olsen TH, Boyles F, Deane CM. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci* (2022) 31:141–6. doi: 10.1002/pro.4205

9. Raybould MI, Kovaltsuk A, Marks C, Deane CM. Cov-abdab: the coronavirus antibody database. *Bioinformatics* (2020) 37:734–5. doi: 10.1093/bioinformatics/btaa739

10. Greiff V, Yaari G, Cowell L. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr Opin Syst Biol* (2020) 24:109–19. doi: 10.1016/j.coisb.2020.10.010

11. Foers AD, Shoukat MS, Welsh OE, Donovan K, Petry R, Evans SC, et al. Classification of intestinal t-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. *J Pathol* (2021) 253:279–91. doi: 10.1002/path.5592

12. Shemesh O, Polak P, Lundin KE, Sollid LM, Yaari G. Machine learning analysis of naïve b-cell receptor repertoires stratifies celiac disease patients and controls. *Front Immunol* (2021) 12:627813. doi: 10.3389/fimmu.2021.627813

13. Carter JA, Preall JB, Grigaityte K, Goldfless SJ, Jeffery E, Briggs AW, et al. Single t cell sequencing demonstrates the functional role of αβ tcr pairing in cell lineage and antigen specificity. *Front Immunol* (2019) 10:1516. doi: 10.3389/fimmu.2019.01516

14. Eliyahu S, Sharabi O, Elmedvi S, Timor R, Davidovich A, Vigneault F, et al. Antibody repertoire analysis of hepatitis c virus infections identifies immune signatures associated with spontaneous clearance. *Front Immunol* (2018) 9:3004. doi: 10.3389/fimmu.2018.03004

15. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nat Genet* (2017) 49:659–65. doi: 10.1038/ng.3822

16. Arnaout R, Luning Prak N, Schwab N, Rubelt F. The future of blood testing is the immunome. *Front Immunol* (2021) 12:228. doi: 10.3389/fimmu.2021.626793

17. Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, et al. The immuneml ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Mach Intell* (2021) 3:936–44. doi: 10.1038/s42256-021-00413-z

18. Ostrovsky-Berman M, Frankel B, Polak P, Yaari G. Immune2vec: Embedding b/t cell receptor sequences in rn using natural language processing. *Front Immunol* (2021) 12:2706. doi: 10.3389/fimmu.2021.680687

19. Dalai SC, Dines JN, Snyder TM, Gittelman RM, Eerkes T, Vaney P, et al. Clinical validation of a novel t-cell receptor sequencing assay for identification of recent or prior severe acute respiratory syndrome coronavirus 2 infection. *Clin Infect Dis* (2022) 75 (12):2079–087. doi: 10.1101/2021.01.06.21249345

20. Elyanow R, Snyder TM, Dalai SC, Gittelman RM, Boonyaratanakornkit J, Wald A, et al. T-Cell receptor sequencing identifies prior SARS-CoV-2 infection and correlates with neutralizing antibodies and disease severity. *JCI insight* (2022) 7(10). doi: 10.1101/2021.03.19.21251426

21. Gittelman RM, Lavezzo E, Snyder TM, Zahid HJ. Longitudinal analysis of T cell receptor repertoires reveals shared patterns of antigen-specific response to SARS-CoV-2 infection. *JCI Insight* (2022) 7(10):e151849. doi: 10.1172/jci.insight.151849

22. Shoukat MS, Foers AD, Woodmansey S, Evans SC, Fowler A, Soilleux EJ. Use of machine learning to identify a t cell response to SARS-CoV-2. *Cell Rep Med* (2021) 2:100192. doi: 10.1016/j.xcrm.2021.100192

23. Chi X, Li Y, Qiu X. V (d) j recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology* (2020) 160:233–47. doi: 10.1111/imm.13176

24. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. *J Allergy Clin Immunol* (2013) 131:959–71. doi: 10.1016/j.jaci.2013.01.046

25. Pilzecker B, Jacobs H. Mutating for good: Dna damage responses during somatic hypermutation. *Front Immunol* (2019) 10:438. doi: 10.3389/fimmu.2019.00438

26. MacCarthy T, Kalis SL, Roa S, Pham P, Goodman MF, Scharff MD, et al. V-Region mutation *in vitro*, *in vivo*, and in silico reveal the importance of the enzymatic properties of aid and the sequence environment. *Proc Natl Acad Sci* (2009) 106:8629–34. doi: 10.1073/pnas.0903803106

27. Schramm CA, Douek DC. Beyond hot spots: biases in antibody somatic hypermutation and implications for vaccine design. *Front Immunol* (2018) 9:1876. doi: 10.3389/fimmu.2018.01876

28. Spisak N, Walczak AM, Mora T. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Res* (2020) 48:10702–12. doi: 10.1093/nar/gkaa825

29. Yaari G, Vander Heiden J, Uduman M, Gadala-Maria D, Gupta N, Stern JN, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* (2013) 4:358. doi: 10.3389/fimmu.2013.00358

30. Turchaninova M, Davydov A, Britanova O, Shugay M, Bikos V, Egorov E, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) 11:1599–616. doi: 10.1038/nprot.2016.093

31. Vander Heiden JA, Stathopoulos P, Zhou JQ, Chen L, Gilbert TJ, Bolen CR, et al. Dysregulation of b cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *J Immunol* (2017) 198:1460–73. doi: 10.4049/jimmunol.1601415

32. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA, et al. Presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30:1930–2. doi: 10.1093/bioinformatics/btu138

33. Brochet X, Lefranc M-P, Giudicelli V. Imgt/v-quest: the highly customized and integrated system for ig and tr standardized vj and vdj sequence analysis. *Nucleic Acids Res* (2008) 36:W503–8. doi: 10.1093/nar/gkn316

34. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) 31:3356–8. doi: 10.1093/bioinformatics/btv359

35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat society: Ser B (Methodological)* (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

36. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. R package caret: Classification and regression training. (2019).

37. Hao Y, Hao S, Andersen-Nissen E, Mauck WMIII, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell* (2021) 184:3573–87. doi: 10.1016/j.cell.2021.04.048

38. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* (2019) 20:163–72. doi: 10.1038/s41590-018-0276-y

39. Kotagiri P, Mescia F, Rae WM, Bergamaschi L, Tuong ZK, Turner L, et al. B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. *Cell Rep* (2022) 38:110393. doi: 10.1016/j.celrep.2022.110393

40. He B, Liu S, Wang Y, Xu M, Cai W, Liu J, 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 targeted Ther* (2021) 6:1–12. doi: 10.1038/s41392-021-00610-7

41. Guo Y, Chen K, Kwong PD, Shapiro L, Sheng Z. Cab-rep: a database of curated antibody repertoires for exploring antibody diversity and predicting antibody prevalence. *Front Immunol* (2019) 10:2365. doi: 10.3389/fimmu.2019.02365

42. Nielsen SC, Yang F, Jackson KJ, Hoh RA, Röltgen K, Jean GH, et al. Human b cell clonal expansion and convergent antibody responses to SARS-CoV-2. *Cell Host Microbe* (2020) 28:516–25. doi: 10.1016/j.chom.2020.09.002

43. Wang Y, Yuan M, Lv H, Peng J, Wilson IA, Wu NC. A large-scale systematic survey reveals recurring molecular features of public antibody responses to SARS-CoV-2. *Immunity* (2022) 55(6):1105–17. doi: 10.1101/2021.11.26.470157

44. Odegard VH, Schatz DG. Targeting of somatic hypermutation. *Nat Rev Immunol* (2006) 6:573–83. doi: 10.1038/nri1896

45. Safra M, Werner L, Polak P, Peres A, Salamon N, Schvimer M, et al. A somatic hypermutation-based machine learning model stratifies individuals with crohn's disease and controls. *Genome Res* (2022) 33(1):71–9. doi: 10.1101/gr.276683.122

46. Ehling RA, Weber CR, Mason DM, Friedensohn S, Wagner B, Bieberich F, et al. SARS-CoV-2 reactive and neutralizing antibodies discovered by single-cell sequencing of plasma cells and mammalian display. *Cell Rep* (2022) 38:110242. doi: 10.1016/j.celrep.2021.110242

47. Mor M, Werbner M, Alter J, Safra M, Chomsky E, Lee JC, et al. Multi-clonal SARS-CoV-2 neutralization by antibodies isolated from severe COVID-19 convalescent donors. *PLoS Pathogens* (2021) 17(2):e1009165. doi: 10.1101/2020.10.06.323634

48. Pan Y, Du J, Liu J, Wu H, Gui F, Zhang N, et al. Screening of potent neutralizing antibodies against SARS-CoV-2 using convalescent patients-derived phage-display libraries. *Cell Discov* (2021) 7:1–19. doi: 10.1038/s41421-021-00295-w

49. Zhou J, Kleinstein S. Position-dependent differential targeting of somatic hypermutation. *J Immunol* (2020) 205(12):3468–79. doi: 10.4049/jimmunol.2000496