Machine learning assisted analysis on TCR profiling data from COVID-19-convalescent and healthy individuals unveils cross-reactivity between SARS-CoV-2 and a wide spectrum of pathogens and other diseases

Georgios K. Georgakilas[1,2,+], Achilleas P. Galanopoulos[1,3,+], Zafeiris Tsinaris[1], Maria Kyritsi[1], Varvara Mouchtouri[1], Matthaios Speletas[3,#,*] and Christos Hadjichristodoulou[1,#].

[1] Laboratory of Hygiene and Epidemiology, Faculty of Medicine, University of Thessaly, Greece.

[2] Laboratory of Genetics, Department of Biology, University of Patras, Greece.

[3] Department of Immunology & Histocompatibility, Faculty of Medicine, University of Thessaly, Greece.

[+,#] Equal contribution.

[*] To whom correspondence should be addressed.

Abstract

During the last two years, the emergence of SARS-CoV-2 has led to millions of deaths worldwide, with a devastating socio-economic impact on a global scale. The scientific community's focus has recently shifted towards the association of the T cell immunological repertoire with COVID-19 progression and severity, by utilising T cell receptor sequencing (TCR-Seq) assays. The Multiplexed Identification of T cell Receptor Antigen (MIRA) dataset, which is a subset of the immunoACCESS© study, provides thousands of TCRs that can specifically recognize SARS-CoV-2 epitopes. Our study proposes a novel Machine Learning (ML) assisted approach for analysing TCR-Seq data from the antigens' point of view, with the ability to accurately distinguish between COVID-19-convalescent and healthy individuals in the case of MIRA dataset. Most SARS-CoV-2 antigens were found to exhibit equal levels of recognition by MIRA TCRs in both convalescent and healthy cohorts, leading to the assumption of putative cross-reactivity between SARS-CoV-2 and other infectious agents. This hypothesis was validated by combining MIRA with other public TCR profiling repositories that host assays and sequencing data concerning a plethora of pathogens. Our study provides evidence regarding the cross-reactivity between SARS-CoV-2 and a wide spectrum of pathogens and diseases, with *M. tuberculosis* and Influenza virus exhibiting the highest levels of cross-reactivity. These results can potentially shift the emphasis of immunological studies towards an increased application of TCR profiling assays that have the potential to uncover key mechanisms of cell-mediated immune response against pathogens and diseases.

Introduction

Since SARS-CoV-2 was initially reported in Wuhan, China, there have been 515 million confirmed cases and 6.2 million deaths worldwide as of May 2022, according to the Johns Hopkins Coronavirus Resource Centre[1]. Individuals infected with SARS-CoV-2 exhibit a wide spectrum of responses, from asymptomatic to requiring admission to an intensive care unit[2]. Both the research

community and pharmaceutical industry have been rigorously studying COVID-19 epidemic patterns[3] and implications of SARS-CoV-2 infection[4], aiming to identify novel prognostic and therapeutic avenues[5–7], while also displaying a massive effort to bring a plethora of vaccination schemes to the public within a very limited timeframe[8].

During the past year, there has been a shift in published research highlighting the need to better understand the T cell immunological profile association with COVID-19 progression and severity[9–15]. T cell immunity appears to be a much more sensitive indicator of past infections in comparison with antibody response. High-throughput methods approaching T cell response can be informative by correlating concepts of clonal depth, breadth and dynamics with symptoms and disease severity[16]. Furthermore, previous studies associated with other viruses such as Middle East Respiratory Syndrome (MERS) and SARS-CoV-1 indicate that coronavirus-specific T cells appear to have long term persistence[17,18]. The same phenomenon also seems to take place in SARS-CoV-2 biology[19,20]. These observations could shape the hypothesis of cross-reactivity between different pathogens, where past infection or vaccination could be protective through long-lived T cell clones. The cross-reactivity phenomenon between SARS-CoV-2 and other coronaviruses has been reported in the literature. Neutralising antibodies isolated from memory-B cells of a SARS-CoV-1 infected individual have been described to react with SARS-CoV-2 surface glycoprotein[21]. In addition, cross-reactive T cells recognising SARS-CoV-2 seem to be acquired during previous infections by other human coronaviruses in 20% to 50% of unexposed individuals around the world[22–27]. Those preexisting cells may affect the clinical manifestations of COVID-19 infection.

The adaptive immune response to pathogenic infections is largely dependent on the CD4+ and CD8+ T cell subfamilies[28]. Upon activation, CD8+ T cells are able to exterminate infected cells and form the long-term memory T cell subpopulation. Conversely, CD4+ T cells control the function of myeloid cells, support CD8+ response and play a key role in the selection of antigen-specific B cells which contribute to the host organism's neutralising antibody arsenal. T cell receptors (TCRs) are proteins localised on the surface of T cells that are products of recombined genomic sequences during the T cell developmental process. The uniqueness of each TCR sequence essentially controls the T cells' specificity. TCRs recognize peptides presented by the major histocompatibility complex (MHC) on the surface of most cell types (MHC class-I recognized by CD8-expressing T cytotoxic cells), or on the surface of antigen-presenting cells (APC) (MHC class-II recognized by CD4-expressing T helper cells). The ability of TCRs to recognize more than one peptide-MHC structure defines cross-reactivity[29]. The cross-reactivity of T cells is considered as one arm of the well-described heterologous immunity, namely the immunity that can develop towards one pathogen after exposure to non-identical pathogens[30]. The other arm concerns the bystander activation of T cells, that can be caused by independent activation via released cytokines, or by low-affinity recognition of pMHC[29,31]. Heterologous immunity has been well established in viral infections in several animal models, but also in human viral infections where cross-reactivity of T cells could potentially influence the protection or severity of virus-associated immunopathology[32–34].

T cell cross-reactivity, however, comes at a cost. Pathogen-induced autoimmune disorders may also result from cross-reactive T cells, following the immune system's initial reaction to the pathogen. This phenomenon has been termed molecular mimicry, where peptides derived from pathogens can activate autoreactive T cells due to structural similarity between pathogenic and self-peptides, causing autoimmune diseases or accelerating a previously initiated autoimmune process[35–37]. The mechanics of T cells cross-reactivity involves changes in complementarity-determining region (CDR) loop conformation, altered TCR docking on the pMHC, flexible changes in pMHC and structural degeneracy[29,38–40]. The binding of TCR with peptides differs concerning their affinity, since the bound peptides could be considerably different in their chemistry[38,39]. In this

context, several previous studies have reported the presence of cross-reactivity among different viruses, or epitopes derived from the same pathogen[41–43].

Recent evidence in the literature highlights[11–13,15] an effort of the research community to explore the TCR repertoire in the context of several levels of COVID-19 infection severity, by utilising data from high-throughput TCR-Sequencing (TCR-Seq) assays. The majority of work focuses on developing computational methods to unveil differences and similarities between healthy and infected subjects related to the TCR repertoire diversity, CDR3 length distribution and the V and J gene segment preference[11–14]. Other studies have attempted to combine the aforementioned TCR-related statistics with Machine Learning (ML), aiming to provide predictive tools to distinguish healthy and infected subjects[44,45]. To our knowledge however, no studies exist in the literature which attempt to approach this field from the viewpoint of the extent each SARS-CoV-2 antigen is recognized by the TCR repertoire of COVID-19 infected and non-infected individuals.

In this study, we propose a novel ML-based approach for analysing TCR repertoires derived from the Multiplexed Identification of T cell Receptor Antigen (MIRA) specificity assay[46] (Fig. 1A and 1B). Such data have been recently generated through academic partnerships with the industrial sector, and were released in the form of freely accessible databases such as immunoACCESS©[15]. This resource includes the immunoSEQ dataset of sequenced TCRb repertoires from COVID-19 exposed, infected or recovered individuals who participate in the Immune Response Action to COVID-19 Events study, as well as thousands of patients' blood samples collected by international institutions globally. The immunoACCESS© repository also includes the MIRA dataset which is complementary to immunoSEQ; the repository catalogues TCRb sequences and TCR specific information about the peptide's interaction in amino acid level, as well as the targeting epitope molecule it comes from. Our approach (Fig. 1B) focused on the MIRA dataset and utilised, for the first time, the level at which each SARS-CoV-2 antigen is recognized by the available TCRs in each sample, to train several ML algorithms that can distinguish samples from COVID-19-convalescent and healthy (no known exposure) cohorts with over 85% accuracy. The module for highlighting the importance of each antigen revealed that the TCR clones recognizing ORF7b, nucleocapsid phosphoprotein as well as ORF1ab, ORF10, ORF3a and membrane glycoprotein to a lesser degree, play a key role for the classification task. Additionally, all MIRA TCRs, regardless of their assigned cohort, were further analysed to determine their potential for recognizing epitopes originating from pathogens other than SARS-CoV-2 (Fig. 1C). To this end, data from public TCR databases were processed to unveil significant cross-reactivity between SARS-CoV-2 and multiple pathogens and other diseases, with *M. tuberculosis* and Influenza virus being the most cross-reactive.

Results

MIRA dataset exploration unveils differentially recognized SARS-CoV-2 antigens between convalescent and healthy samples

The MIRA dataset includes 144 samples (experiments) from the immunoACCESS study[15] that are divided into 5 cohorts: 1) COVID-19-convalescent, 2) healthy (no known exposure), 3) COVID-19-acute, 4) COVID-19-non-acute and 5) COVID-19-exposed (Fig. 2A). Most cohorts have a balanced male-to-female ratio, however there are 26 samples from unknown gender, of which the majority (N=21) are reported as COVID-19-convalescent cohort. The COVID-19-acute, -non-acute and -exposed cohorts were removed from all subsequent analyses due to low sample numbers. Initial analysis regarding the normalised number of unique TCRs in each sample revealed there is no statistically significant difference between the healthy and convalescent cohorts (Fig. 2B).

Most existing studies focus on TCR properties such as the underlying VJ rearranged sequences, CDR3 sequences, CDR3 size and clonal depth/diversity[11–14]. This information has been frequently used to characterise the TCR repertoire of immune response against COVID-19 and to train ML algorithms that are able to distinguish between samples of distinct cohorts[44,45]. In this study, a different approach was adopted. Rather than using the aforementioned TCR-related data, we exploited the MIRA dataset[15], and the connection between TCRs and the SARS-CoV-2 epitopes to generate an 11-dimensional vector representing the level at which each SARS-CoV-2 antigen is recognised by the TCRs in each sample (Fig. 1B and 2C).

The results of this approach revealed that even in samples from the healthy cohort, all SARS-CoV-2 antigens are recognized by TCRs to some extent, suggesting either previous unreported COVID-19 infection of subjects in the healthy cohort, or putative cross-reactivity between SARS-CoV-2 and other pathogens. Interestingly, the ORF1ab and surface glycoprotein are the two antigens recognized by TCRs with the highest overall clonal depth, although the difference in clonal expansion between the two cohorts is not statistically significant. More importantly, the nucleocapsid phosphoprotein, ORF7b, ORF10, ORF8 and envelope exhibit statistically significant differences in the number of TCRs recognizing their epitopes between the two cohorts. However, the ORF10, ORF8 and envelope proteins are recognized by TCRs with low clonal depth. The projection of these samples on the principal component analysis (PCA) space (Fig. 2D) has led to the assumption that this approach could be used to develop a ML-based framework to accurately distinguish between samples from the two cohorts (Fig. 1B).

Explainable ML highlights key SARS-CoV-2 antigens for classifying samples into the convalescent and healthy MIRA cohorts

The strategy of modelling the MIRA dataset involved a repeated process (N=20) of separately splitting samples from healthy and convalescent cohorts into training and test sets. At each split, 5 models based on Gaussian Naive Bayes (GaussianNB), Decision Trees (DT), K-Nearest Neighbours (KNN), Random Forests (RF) and Support Vector Machines (SVM) were trained and evaluated (Fig. 1B).

Using a prediction score cut-off of 0.5 enabled the extraction of performance metrics on each test set (Fig. 3A). SVM was the overall best performing algorithm with a median performance of at least 0.75 in all metrics. Notably, the SVM algorithm exhibits 0.856 median balanced accuracy, 0.896 precision, 0.962 sensitivity, 0.75 specificity and 0.9 negative predictive value (NPV). To observe SVM's performance across the whole spectrum of prediction score thresholds, an incremental cut-off was applied and all metrics were calculated at each step (Fig. 3B).

To assess the importance of each feature, a repeated (N=50) feature value permutation process was applied for all ML algorithms (Fig. 3C). Overall, the most important features are ORF7b, nucleocapsid phosphoprotein as well as ORF1ab, ORF10, ORF3a as well as membrane glycoprotein to a lesser extent. After selecting only the most important features for each algorithm, the training and evaluation process was repeated. This resulted in slightly improved performance for GaussianNB, KNN and SVM algorithms, but not for DT and RF, as expected, considering their innate ability to readily rely only on important features (Fig. 3D).

Exploratory analysis of MIRA TCRs unveils cross-reactivity between SARS-CoV-2 and other pathogens and diseases

MIRA TCRs exhibit diverse frequency of occurrence and clonal expansion levels (Fig. 4A). In general, the most frequent clonotypes in the dataset present low mean expansion after being

triggered with the MIRA assay, while the least frequent clonotypes are associated with the highest expansion levels.

The analysis focused on the six most common clonotypes with frequencies of greater than 11% of the total number of subjects; CASSIRSSYEQYF+V19-01+J02-07, CASSLAGAYEQYF+TCRBV05-01+TCRBJ02-07, CASSLSAPQETQYF+TCRBV27-01+TCRBJ02-05, CASSLSSPQETQYF+TCRBV27-01+ TCRBJ02-05, CASSDRGPNQPQHF+TCRBV27-01+TCRBJ01-05 and CASSDRGPTDTQYF+ TCRBV27-01+TCRBJ02-03, that were found in 23%, 15.04%, 13.27%, 12.38%, 11.5% and 11.5% of total MIRA subjects, respectively (Fig. 4A, marked with arrows). The distribution of the clonal expansion level in samples belonging to the two cohorts was also calculated, not unveiling any statistically significant differential expansion between the two cohorts (Fig. 4B, based on Mann-Whitney; the statistical test could not be performed for some of the TCRs, denoted with *p*-val N/A). For every related sample, the number of times each clonotype appears in the corresponding sample(s) was divided by the total number of the sample's clonotypes.

Calculation of cohort distribution took place in each clonotype's subgroup compared to the whole sample with Fisher's exact test (Fig. 4C). We observed a significant difference in the case of the most frequent clonotype CASSIRSSYEQYF+V19-01+J02-07 (*p*-value = 0.0019) and the fourth most common clonotype CASSLSSPQETQYF+V27-01+J02-05 (*p*-value = 0.038). Specifically, CASSIRSSYEQYF+V19-01+J02-07 clonotype is detected in a sample's subpopulation where the majority of subjects are characterised as healthy. In contrast, CASSLSSPQETQYF+V27-01+J02-05 clonotype is detected only in convalescent subjects.

Additionally, the SARS-CoV-2 antigen targets of the six most common clonotypes were identified. CASSIRSSYEQYF+TCRBV19-01+TCRBJ02-07 interacts with surface glycoprotein and ORF1ab, CASSLAGAYEQYF+TCRBV05-01+TCRBJ02-07 recognises the nucleocapsid phosphoprotein, CASSLSAPQETQYF+TCRBV27-01+TCRBJ02-05 recognises ORF1ab and envelope, CASSLSSPQETQYF+TCRBV27-01+TCRBJ02-05, CASSDRGPNQPQHF+TCRBV27-01+TCRBJ01-05 and CASSDRGPTDTQYF+TCRBV27-01+TCRBJ02-03 interact with ORF1ab. Public databases such as McPAS, TCR3d and VDJdb were used to query the cross-reactive components of the aforementioned TCRs. CASSIRSSYEQYF+TCRBV19-01+TCRBJ02-07 was also found to interact with Influenza virus and Epstein-Barr Virus (EBV). The description of this specific clonotype as part of the immune response against Influenza virus has been previously described[47]. These results verified the suspicions of cross-reactivity that surfaced through observations based on results of the initial analysis presented in Figure 2C. Figure 4D highlights the most common TCR's interaction sites on surface glycoprotein from SARS-CoV-2 and matrix protein 1 (M1) from Influenza A virus. The same information is also depicted in the form of a circular plot in Figure 4E. The remaining five most common clonotypes (CASSLAGAYEQYF+TCRBV05-01+TCRBJ02-07, CASSLSAPQETQYF+TCRBV27-01+TCRBJ02-05, CASSLSSPQETQYF+TCRBV27-01+ TCRBJ02-05, CASSDRGPNQPQHF+TCRBV27-01+TCRBJ01-05 and CASSDRGPTDTQYF+ TCRBV27-01+TCRBJ02-03) were not found in VDJdb, McPAS or TCR3d to interact with other pathogens.

To further assess the cross-reactive properties of all TCRs in MIRA dataset, the epitopes in MIRA as well as epitopes in the three aforementioned public TCR databases were used to match unique MIRA CDR3 sequences with antigens from SARS-CoV-2 and other pathogens and diseases (Fig. 5A, Supplementary Tables 1-4). In general, some CDR3 sequences have the ability to recognize epitopes from multiple antigens. Therefore, the number of SARS-CoV-2 antigen connections with other pathogens' and diseases' antigens might not be equal to the number of unique MIRA CDR3 sequences.

As shown in Figure 5A and Supplementary Figure 1, the cross-reactivity phenomenon is widespread and links SARS-CoV-2 to a plethora of pathogens and other diseases that can be grouped into three major categories (Supplementary Table 4). The first category includes *M. tuberculosis* and viruses such as Influenza virus, Cytomegalovirus (CMV), EBV, Human Immunodeficiency Virus (HIV), Hepatitis C Virus (HCV), Yellow Fever Virus (YFV), Dengue Virus (DENV) and Human T-lymphotropic Virus type 1 (HTLV-1). From 2,136 CDR3 sequences that are common between the MIRA and McPAS, TCR3d and VDJdb repositories, 1,792 (83.9%) have the ability to recognize epitopes from SARS-CoV-2 and members of the first category (Supplementary Tables 1 and 4). The second category consists of malignancies and malignancy-related agents such as Melanoma, Breast Cancer and Neoantigens. Roughly 9.9% (211 out of 2,136) of the CDR3 sequences that are common between MIRA and the three public TCR profiling databases, recognize epitopes from the second category (Supplementary Tables 1 and 4). On the other hand, 6.2% (133 out of 2,136) of the CDR3 sequences recognize epitopes from the third category, which reflects auto-immune states and disorders arising from external stimuli such as Celiac Disease, Inflammatory Bowel Disease (IBD), Diabetes Type 1, Psoriatic Arthritis, Allergy and Toxic Epidermal Necrolysis (Supplementary Tables 1 and 4).

The most cross-reactive partner of SARS-CoV-2 was found to be *M. tuberculosis*, with 747 CDR3 sequences able to recognise epitopes from both pathogens (Figure 5A, Supplementary Table 4). These CDR3 sequences were isolated from CD8+ T cell TCRs in MIRA. However, in McPAS, TCR3d and VDJdb, 680 out of the 747 CDR3 sequences are reported to originate from CD4+ and 67 from CD8+ T cell TCRs. Specifically, 271 unique CDR3 sequences (240 from CD4+ and 31 from CD8+ T cells) interacting with ORF1ab and 127 (113 from CD4+ and 14 from CD8+ T cells) with surface glycoprotein that can also interact with *M. tuberculosis* antigens (Supplementary Tables 1 and 2).

Influenza virus exhibits the second highest number of CDR3 sequences (498 out of 2,136) that are cross-reactive with SARS-CoV-2 (Supplementary Table 4). We observed the majority of CDR3 sequences originate from CD8+ T cell TCRs (487 out of 498) according to the public TCR profiling databases, in contrast to the case of *M. tuberculosis*. Notably, 144 unique CDR3 sequences (1 from CD4+ and 143 from CD8+ T cells) recognizing Influenza virus also interact with surface glycoprotein and 123 with ORF1ab (5 from CD4+ and 118 from CD8+ T cells).

To have a complete view of the cross-reactivity phenomenon, we further generated a circular plot depicting the “cross-talk” between distinct antigen regions through the recognition by MIRA CDR3 sequences (Fig. 5B, Supplementary Table 3). To ensure concise visualisation, a subset with the most cross-reactive pathogens, as depicted in Figure 5A, was selected for generating the plot including *M. tuberculosis*, Influenza virus (A subtype), CMV, EBV and HIV. Each scaled segment in the circle represents an antigen and every antigen is coloured based on the pathogen it belongs to. The links connecting antigen pairs correspond to the “cross-talk” between the antigens’ segments through their ability to be recognized by a specific TCR in MIRA. Although *M. tuberculosis* is the most cross-reactive partner of SARS-CoV-2, the matching epitope information is missing from McPAS, TCR3d and VDJdb for the majority of CDR3 sequences. Thus, the *M. tuberculosis* connections in Figure 5B are severely limited. The same phenomenon was also observed for the other pathogens, but to a lesser degree.

Discussion

Over the last two years, the global impact of COVID-19 on healthcare[48] and the socio-economic[49] field has been devastating. The response of both the scientific community and pharmaceutical industry was swift and decisive in exploring the biological aspect of SARS-CoV-2 and its

pathological implications, as well as delivering pharmaceutical products that could assist in restraining COVID-19. During the past year, an observed shift in the literature was evident, highlighting the T cell immunological profile characterization in the framework of COVID-19 progression and severity[9–15]. Collaborations between academia and industry resulted in the publication of immunological datasets from studies with thousands of subjects, such as the immunoACCESS© resource[15]. Specifically, the MIRA dataset provides access to thousands of TCR clonotypes that can specifically recognize SARS-CoV-2 epitopes (Fig. 1A).

Our strategy is a novel ML-oriented TCR profiling assay analytic approach, which can highlight the targeted antigens of the immune response against pathogens and diseases (Fig. 1B and 3). During the last two decades we have experienced an abundance of breakthroughs in biotechnology that facilitated the dawn of the big data era for biology. We believe the scientific community should emphasise the development of efficient and accurate computational approaches, to exploit the wealth of information embedded in the ever-increasing volume of biomedical data. ML can be the ideal substrate for combining data from heterogeneous sources of information, while unveiling higher-order and more abstract connections between the underlying mechanisms of biological phenomena and the environment. In the context of COVID-19 related research and other infectious diseases, ML could be used for combining epidemiological surveillance with data from immunoassays (TCR-Seq and others), genomics, transcriptomics and even metagenomics. Such approaches can provide a solid foundation for understanding the entanglement of genetic factors and the environment, as well as their implication on the progression of pandemics, for example, within different populations.

One key observation in our study is that CD8+ T cells in the MIRA dataset with the ability to recognize epitopes from ORF1ab and surface glycoprotein exhibit similar levels of clonal expansion between the two cohorts and present the highest clonal expansion levels in the healthy cohort (Fig. 2C); this suggests either previously unreported COVID-19 exposure or putative cross-reactivity between SARS-CoV-2 and other pathogens. The former hypothesis could not be verified by any means. Therefore, we proceeded exploring whether the MIRA TCRs exhibit cross-reactive properties that can be a product of an immune response against both SARS-CoV-2 and other pathogens and diseases.

Our analysis was based on the CDR3 sequences and the corresponding epitopes that are common between the MIRA dataset and McPAS, TCR3d and VDJdb repositories. The results unveiled widespread cross-reactivity that link SARS-CoV-2 to a plethora of pathogens and other diseases (Fig. 5), which can be grouped into three major categories: a) *M. tuberculosis* and viruses including Influenza virus, CMV, EBV and HIV among others, b) malignancies and malignancy-related agents, and c) auto-immune states and disorders. Interestingly, the majority of CDR3 sequences that target pathogens in the first category originate from CD8+ T Cell TCRs, according to McPAS, TCR3d and VDJdb, with *M. tuberculosis* being the exception. The association of BCG vaccine with CD4+ T cell response against *M. tuberculosis* has been previously reported in literature[50]. In contrast to the first category, we observed the exact opposite pattern for CDR3 sequences that are associated with pathological states from the second and third categories, since they are derived mostly from CD4+ T cell TCRs.

M. tuberculosis was found to exhibit the highest levels of T cell cross-reactivity with SARS-CoV-2. These results are in accordance with published epidemiological studies conducted prior to SARS-CoV-2 vaccine implementation, that have suggested a negative association between incidence, morbidity and mortality of COVID-19 and national Bacille Calmette-Guérin (BCG) vaccination programs. Specifically, in countries where national BCG vaccination has been implemented, lower numbers COVID-19 cases and related deaths have been recorded[51–53]. A study conducted by Escobar et al.[54] found a 10.4% reduction in mortality from COVID-19 for every 10% increase in a

country's BCG index. Discovered in 1921, even today the function of BCG vaccine remains obscure to a certain extent. The BCG vaccine contains attenuated Mycobacterium bovis, and induces humoral and adaptive immunity, activating both non-specific and cross-reactive immune responses in the host[55,56] against a variety of infectious (viruses, bacteria, fungi and parasites) and non-infectious agents. Epigenetic and metabolic reprogramming of innate immune cells, known as trained immunity, is considered responsible for these protective effects[57–60].

Influenza virus exhibits the second highest number of CDR3 sequences that are cross-reactive with SARS-CoV-2. This type of cross-reactivity could be attributed to the seasonal vaccination against Influenza virus and relevant exposure of a large portion of the population to this virus. The role of protective immunity induced by the polyvalent influenza virus vaccine (against Influenza A virus and/or Influenza B virus subtypes), and the likelihood of COVID-19 has been previously examined[61–63]; meanwhile others explored this association from a clinical manifestation and disease outcome perspective[64]. Although the precise pathophysiological mechanisms underlying this association require further investigation, three main theories have been put forward. The first theory relates to antigenic mimicry which results in clonal activation and lymphocytes proliferation[65]. Depending on each individual's HLAs, only a limited number of epitopes can be recognised, and those are the immunodominant ones. A second theory of trained immunity has also been proposed as the mechanism behind these beneficial heterologous effects of vaccines[66]. Influenza virus vaccination acts as a non-specific exciter of our immune response[67]. Debisarun et al. found that rather than binding, cytokines broaden T cell responses against SARS-CoV-2[68]. Salem et al. suggested flu-induced bystander immune response as a probable protective mechanism[69].

Information related to T cell cross-reactivity between SARS-CoV-2 and the remaining viruses from the first category of cross-reactive pathogens is scarce in the literature. Cellular cross-reactivity against EBV and SARS-CoV-2 has not been studied to date. However, there have been reports associating the clinical manifestation of COVID-19 with reactivation of EBV infection and correlating it with severe disease progression, thus underpinning a possible entanglement[70–72]. Cross-reactivity between SARS-CoV-2 and HIV has also been reported in studies describing false positive HIV results in COVID-19 patients. Such cases were associated with antibody cross-reaction during immunoassay screening tests, while cross-reactive CDR3 regions were detected between the two viruses[73,74]. Sequence analysis has shown that HIV and SARS-CoV proteins share common motifs that shape a degree of homology[75]. In addition, studying similar antigenic features could help in engineering super antibodies that neutralise different pathogens[76].

Evidence in the literature regarding the connection between SARS-CoV-2 with cancer at various levels remains extremely limited. Most related studies examine the immune response to SARS-CoV-2 in patients with cancer[77]. However, the cross-reactivity phenomenon between SARS-CoV-2 and malignancy-related antigens has not been previously reported. Conversely, since COVID-19 pandemic was declared in 2020, numerous reports in the literature have linked the immune response against SARS-CoV-2 proteins with self-antigens, thus unveiling putative COVID-19 implications for autoimmune disorders including immune thrombocytopenic purpura, Guillian-Barrė syndrome and subvariants, antiphospholipid antibodies and lupus anticoagulant, Kawasaki and multisystem inflammatory syndrome in children[78–81]. In general, the emergence of autoimmunity after viral infections involving EBV, CMV, HTLV-1, herpes and hepatitis virus among others has been thoroughly described in the literature[82].

The observations in this study are of multilevel significance, ranging from indirect protection from severe COVID-19 infection based on vaccination against and/or previous exposure to Influenza viruses, *M. tuberculosis* and other pathogens, to the association of SARS-CoV-2 related TCRs with malignancies and autoimmune disorders. However, there are several limitations related to this

study. The MIRA dataset includes a very limited number of samples associated with COVID-19-acute, COVID-19-non-acute and COVID-19-exposed cohorts, thus prohibiting any statistical or ML analysis to potentially connect TCR profile irregularities with disease severity. We did however manage to make statistically significant assumptions using samples derived from COVID-19-convalescent and healthy cohorts, although ideally the number of these samples should be higher. Another limitation relates to the availability of HLA allele information connecting TCR clonotypes and epitopes. The cross-reactivity analysis focused only on the CDR3 amino acid sequence comparison between MIRA dataset and other databases, due to the unavailability of the relevant HLA information for most TCR clonotypes. Therefore, it should be noted that the cross-reactions described here could only take place in individuals with specific HLA alleles, enabling the presentation of relative epitopes to potential long-lived memory T cells developed during previous infection and/or vaccination. Additionally, the extent of cross-reactivity is delimited by the inherent data bias in McPAS, TCR3d and VDJdb, stemming from the scientific community's focus on specific pathogenic cases.

We believe that our study provides a novel ML-based computation framework for analysing TCR-Seq datasets and systematically highlights the breadth and depth of "cross-talk" between antigens from different pathogens, a phenomenon that may also exhibit therapeutic implications, especially in the COVID-19 context. It is our view that the scientific community should accelerate the effort of generating TCR profiling data without omitting the relevant HLA information, and include assays based on as many human pathologies as possible. ML can play a pivotal role in combining such data with multiomics and epidemiological surveillance, to build intelligent infrastructures that can be an invaluable asset in fully understanding the underlying immune response complexity.

Methods

Data collection and pre-processing

CD8+ TCRs that are able to bind to SARS-CoV-2 epitopes were retrieved from the immuneACCESS© database[15] (Fig. 1A). These SARS-CoV-2 specific TCRs are part of the MIRA dataset which is based on 144 samples (experiments) obtained from cohorts with exposed subjects and healthy controls (Fig. 1B and 2A). It should be noted that for certain subjects, more than one sample is available in MIRA. Specifically, 90 samples originate from COVID-19-convalescent subjects, 39 from healthy (no known exposure), 4 from COVID-19-acute, 8 from COVID-19-non-acute, and 3 from COVID-19-exposed. Samples from the COVID-19-acute, COVID-19-non-acute, and COVID-19-exposed cohorts were excluded from the analyses presented herein, due to their limited number. TCR sequences were initially filtered to keep CDR3 regions delimited by a conserved cysteine at the start and a conserved phenylalanine or tryptophan at the end (anchors of CDR3 region). Unproductive CDR3 segments and sequences containing special characters not corresponding to amino acids (X, #, *, etc.) were also excluded. MIRA was further filtered to keep information associated only with functional V genes, removing information related to pseudogenes and ORFs according to the immunogenetics information system (IMGT)[83]. Clonotypes with ambiguous V gene family members (denoted with X) were also removed. The analysis focused on the remaining 130,072 (120,128 unique) TCRs detected in the studied cohorts (28 healthy and 85 convalescent subjects). The unique number of TCRs per 1,000 TCRs in each sample is depicted in Figure 2B.

All remaining TCRs were further analysed from the SARS-CoV-2 antigens' point of view. For every sample, each TCR was assigned to an antigen category (N=11), based on its epitope recognition ability (Fig. 1B and 2C). Some TCRs were able to recognize epitopes from different antigens. For

these cases, the assignment to the corresponding antigens was weighted based on the number of antigens. The number of TCRs per antigen was normalised based on the total number of TCRs per sample. This approach enabled the representation of each sample by an 11-dimensional vector and facilitated the aggregation of a dataset used to build several ML models that can classify samples into the convalescent and healthy categories and explore the underlying biology (Fig. 1B).

Data catalogues derived from three public TCR databases with immunogenetic information were downloaded to investigate the MIRA dataset's TCR involvement in immune response during other infections. Pathology-associated TCR database McPAS[84] is a manually curated dataset of TCR sequences associated with various pathological manifestations, containing information about the T cell type, organ or tissue antigen target and related MHC molecules (version 4/1/2022). TCR3d[85] is a structural repertoire database including experimentally determined TCR structures and complexes. It also contains TCR sequences and related antigenic peptides and MHC molecules. We used a data frame derived from TCR3d focused on TCRb CDR3 sequence level and the association with viruses (version 13/1/2022). VDJdb[86] is a database containing TCR sequences, their cognate antigens and related MHC molecules (version 22/3/2022). All three aforementioned databases were filtered to keep information about immune response in human species and CDR3 sequences associated with the beta chain of TCR; the MIRA filtering strategy described earlier in this section, was also applied here. Additionally, in the case of VDJdb, CDR3 sequences with zero confidence score were removed from downstream analyses. As stated in VDJdb's documentation, the higher the score the more confidence in antigen specificity annotation of a given TCR clonotype. Some VDJdb sequences were processed during fixing steps according to IMGT nomenclature and were included in our analysis. Furthermore, the McPAS sequences associated with antigen identification method id "3" were removed in accordance with the database's recommendation for confidence in the accuracy of the data.

All protein sequences were downloaded from UniProt[87] and the cross-reactivity exploration was achieved with custom Python scripts and Circos[88]. The statistical, dimensionality reduction, ML and feature importance analyses were performed with in-house developed software based on Python's scipy and scikit-learn as well as R's dplyr, plyr, GLDEX, TSDT, stats, stringr and ggplot2 libraries.

Machine learning model training and feature importance estimation

All 90 COVID-19-convalescent samples were labelled as positives, and the 39 healthy (no known exposure) samples as negatives (Fig. 1B and 2A). Each sample consists of an 11-dimensional vector of normalised values, for every SARS-CoV-2 antigen, that represents the percentage of TCRs recognizing the antigen's epitopes from the total amount of SARS-CoV-2-specific TCRs in the sample (Fig. 1B and 2C). Visualisation of the data in the principal component space is depicted in Figure 2C. Both positive and negative samples were randomly divided into training and test sets based on a 7:3 ratio. This process was repeated 20 times to generate an equal amount of training/test set combinations and control for any bias that could result from the splitting process (Fig. 1B).

These sets were used to train and evaluate a total of 100 models, based on popular ML algorithmic families (Fig. 1B): GaussianNB, DT, KNN, RF and SVM. The hyperparameters of each model were tuned based on a grid search approach, and balanced accuracy was the target metric for choosing the best performing model. For GaussianNB and the variance smoothing parameter, the grid search was run on the values 1e-9, 1e-8 and 1e-7. For DT, the benchmarked values for maximum depth were 10, 30 and 90 with the maximum features parameter was set to None. In the case of KNN, the k-neighbours parameter values were 2, 5 and 10. The algorithmic options for selecting nearest neighbours were ball_tree, kde_tree and brute, and the distance metric parameter values were 1 (manhattan) and 2 (euclidean). For RF, the maximum depth values were 10, 30 and 90; the

number of estimators were 10, 50 and 100. The maximum features parameter was set to None. The SVM kernel was set to radial basis function and the different 'C' parameter values were 0.1, 1, 10 and 100. The gamma values were 1, 0.1, 0.01 and 0.001. All models were trained based on a 10-fold cross validation scheme repeated 10 times. Subsequently, the best performing model was evaluated on its designated test set.

This approach resulted in the calculation of performance metrics such as balanced accuracy, precision, sensitivity, specificity and NPV (Fig. 3A and 3B). Since 20 models were trained and benchmarked for each ML algorithm, all performance plots depict the metrics' score distributions from all test sets, providing hints of putative training/test split bias and data heterogeneity. Additionally, the importance of each feature was estimated by repeated (N=50) random shuffles of single feature values, to assess the decrease or increase of the models' performance (Fig. 3C). For each ML algorithm, features with a median score above 0.01 were selected for a second round of training and evaluation, following the previously aforementioned strategy (Fig. 3D).

Identification of TCRs that recognize epitopes from antigens of SARS-CoV-2 and other pathogens and diseases

Several definitions of the clonotype concept exist as various studies approach it with different immunogenetic characteristics. The MIRA dataset contains TCRb sequences targeting specific SARS-CoV-2 epitopes and the TCR bioidentity is described by CDR3 amino acid sequence, V and J genes. In this analysis, every clonotype consists of sequences characterised by the same V gene family member, the same J gene family member and the same CDR3 sequence in amino acid level. Clonal expansion of every clonotype in each sample (experiment) was calculated by counting the times it appears divided by the total MIRA clonotypes detected. Hence, clonal expansion was defined as a measure of the T cell proportion expressing a specific TCRb sequence. The mean clonal expansion of each clonotype was assessed by calculating the mean of expansion values from all subjects (Fig. 4A).

The six most frequent clonotypes were further analysed from the viewpoint of antigen recognition, clonal expansion range and clinical impact. The level of clonal expansion was also calculated separately for the convalescent and healthy cohorts (Fig. 4B). Additionally, each clonotype was characterised for the clinical cohort distribution and statistical significance was also determined with Fisher's exact test (Fig. 4C).

The cross-reactivity phenomenon between SARS-CoV-2 and other pathogens was also examined. We used the public TCR databases McPAS[84], TCR3d[85] and VDJdb[86] to confirm the existence of different pathogens and other diseases associated with clonotypes targeting SARS-CoV-2 epitopes in the MIRA dataset. Quantitative analysis was conducted calculating the number of unique CDR3 sequences associated with each pathology, to capture the number of putative cross-reactive TCRs (Fig. 5A). In certain cases, a single CDR3 sequence is able to recognize epitopes from multiple antigens. Thus, the number of SARS-CoV-2 antigen connections with other pathogens' and diseases' antigens is not equal to the number of unique MIRA CDR3 sequences (Fig. 5 & Supplementary Table 4). This approach was applied twice, once by screening all CDR3 sequences from the aforementioned databases (derived from CD8+ and CD4+ T cells) and once by targeting only CD8+ T cells, to examine any potential functionality bias (Supplementary Fig. 1). Most databases' information is related to Influenza virus, *M. tuberculosis*, EBV, CMV and HIV. Thus, the analysis was carried out using as a frame of reference the epitopes on the amino acid level and aiming at how antigens are recognised by TCRs in the context of SARS-CoV-2 infection or relevant pathogens. To identify locations crucial for the interaction with cross-reactive CDR3 regions, the epitopes were aligned to the pathogenic proteins they derive from (Fig. 4D, 4E and

5B). This was the first step to characterise the specific domains and functionality of antigens recognised by the same CDR3 sequences.

It should be noted that these public databases include information for both CD4+ and CD8+ T cells and a data bias exists due to the numerous study results associated with specific pathogenic cases. To briefly mention the most notable cases, there are 16,162 CDR3 sequences associated with *M. tuberculosis*, 3,639 with Influenza virus, 2,663 with CMV, 1,583 with HIV and 1,437 with EBV. In the case of TCRs derived from CD8+ T cells only, a different data bias was observed in the public catalogues. For example, 3,479 CDR3 sequences were observed to be associated with Influenza virus, 1,183 with *M. tuberculosis*, 2,658 with CMV, 1,334 with EBV and 1,237 with HIV.

The 3-dimensional view of M1 and surface glycoprotein antigenic molecules of Influenza virus and SARS-CoV-2 respectively, was generated to highlight the epitope locations of interaction with the most common MIRA clonotype (Fig. 4D). The antigenic epitopes were aligned on reference sequences using Clustal algorithm and the 3-dimensional structure was generated with Jmol, within Jalview software based on 1AA7 (A and B chain view) and 6X29 (A chain view) Protein Data Bank[89] entries for M1 and surface glycoprotein, respectively.

Data availability

Supplementary data are provided with this paper.

Acknowledgements

We wish to acknowledge Ms Lemonia Anagnostopoulos for proofreading the manuscript.

The study was funded by own resources of the Laboratory of Hygiene and Epidemiology of the Medical School of the University of Thessaly.

Author information

Affiliations

Department of Immunology & Histocompatibility, Faculty of Medicine, University of Thessaly, 41500, Larissa, Greece

Achilleas P. Galanopoulos & Matthaios Speletas

Laboratory of Hygiene and Epidemiology, Faculty of Medicine, University of Thessaly, 41222, Larissa, Greece

Georgios K. Georgakilas, Achilleas P. Galanopoulos, Zafeiris Tsinaris, Maria Kyritsi, Varvara Mouchtouri & Christos Hadjichristodoulou

Laboratory of Genetics, Department of Biology, University of Patras, 26500, Campus Rio, Patras, Greece

Georgios K. Georgakilas

Contributions

G.K.G. and A.P.G. designed the study under M.S.'s and C.H.'s supervision. G.K.G. performed the Machine Learning analysis, cross-reactivity plots and prepared the figures. A.P.G. performed the MIRA and external databases TCR/CDR3 data filtering and the cross-reactivity analysis. G.K.G. and A.P.G. wrote the paper with the assistance of Z.T., M.K., V.M., M.S. and C.H. Study supervision was conducted by M.S. and C.H.

Corresponding authors

Correspondence to Matthaios Speletas.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. JHCRC. John Hopkins Coronavirus Resource Center. *Johns Hopkins University & Medicine*.
2. Guan, W.-J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
3. Velavan, T. P. & Meyer, C. G. The COVID-19 epidemic. *Trop. Med. Int. Health* **25**, 278–280 (2020).
4. Lopez-Leon, S. *et al.* More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Sci. Rep.* **11**, 16144 (2021).
5. Whitley, R. Molnupiravir - A step toward orally bioavailable therapies for covid-19. *N. Engl. J. Med.* (2021) doi:10.1056/NEJMe2117814.
6. Gupta, A. *et al.* Early Treatment for Covid-19 with SARS-CoV-2 Neutralizing Antibody Sotrovimab. *N. Engl. J. Med.* **385**, 1941–1950 (2021).
7. Gottlieb, R. L. *et al.* Early Remdesivir to Prevent Progression to Severe Covid-19 in Outpatients. *N. Engl. J. Med.* **386**, 305–315 (2022).
8. Tregoning, J. S., Flight, K. E., Higham, S. L., Wang, Z. & Pierce, B. F. Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. *Nat. Rev. Immunol.* **21**, 626–636 (2021).
9. Minervina, A. A. *et al.* Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *Elife* **10**, (2021).
10. Hanna, S. J. *et al.* T cell phenotypes in COVID-19 - a living review. *Oxf Open Immunol* **2**, (2020).
11. Shomuradova, A. S. *et al.* SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. *Immunity* **53**, 1245–1257.e5 (2020).
12. Chang, C.-M. *et al.* Profiling of T Cell Repertoire in SARS-CoV-2-Infected COVID-19 Patients Between Mild Disease and Pneumonia. *J. Clin. Immunol.* **41**, 1131–1145 (2021).
13. Li, L. *et al.* T Cell Immunity Evaluation and Immunodominant Epitope T Cell Receptor Identification of Severe Acute Respiratory Syndrome Coronavirus 2 Spike Glycoprotein in COVID-19 Convalescent Patients. *Front Cell Dev Biol* **9**, 696662 (2021).
14. Wang, P. *et al.* Comprehensive analysis of TCR repertoire in COVID-19 using single cell sequencing. *Genomics* **113**, 456–462 (2021).

15. Nolan, S. *et al.* A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq* (2020) doi:10.21203/rs.3.rs-51964/v1.
16. Gittelman, R. M. *et al.* Diagnosis and tracking of past SARS-CoV-2 infection in a large study of Vo'. *Italy through T-cell receptor sequencing. medRxiv* **9**, 2020 (2020).
17. Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K. & Perlman, S. Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *J. Virol.* **88**, 11034–11044 (2014).
18. Zhao, J. *et al.* Recovery from the Middle East respiratory syndrome is associated with antibody and T-cell responses. *Sci Immunol* **2**, (2017).
19. Gallais, F. *et al.* Intrafamilial exposure to SARS-CoV-2 induces cellular immune response without seroconversion. medRxiv 2020.06. 21.20132449 [Preprint]. 22 June 2020.
20. Thieme, C. *et al.* The SARS-COV-2 T-Cell Immunity is Directed Against the Spike, Membrane, and Nucleocapsid Protein and Associated with COVID 19 Severity. (2020) doi:10.2139/ssrn.3606763.
21. Pinto, D. *et al.* Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **583**, 290–295 (2020).
22. Rydyznski Moderbacher, C. *et al.* Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. *Cell* **183**, 996–1012.e19 (2020).
23. Sekine, T. *et al.* Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. *Cell* **183**, 158–168.e14 (2020).
24. Le Bert, N. *et al.* SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* **584**, 457–462 (2020).
25. Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* **370**, 89–94 (2020).
26. Sette, A. & Crotty, S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. *Nat. Rev. Immunol.* **20**, 457–458 (2020).
27. Braun, J. *et al.* SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* **587**, 270–274 (2020).
28. Pennock, N. D. *et al.* T cell responses: naive to memory and everything in between. *Adv. Physiol. Educ.* **37**, 273–283 (2013).
29. Petrova, G., Ferrante, A. & Gorski, J. Cross-Reactivity of T Cells and Its Role in the Immune System. *CRI* **32**, (2012).
30. Welsh, R. M., Che, J. W., Brehm, M. A. & Selin, L. K. Heterologous immunity between viruses. *Immunol. Rev.* **235**, 244–266 (2010).
31. Bangs, S. C. *et al.* Human CD4+ memory T cells are preferential targets for bystander activation and apoptosis. *J. Immunol.* **182**, 1962–1971 (2009).
32. Selin, L. K., Varga, S. M., Wong, I. C. & Welsh, R. M. Protective heterologous antiviral immunity and enhanced immunopathogenesis mediated by memory T cell populations. *J. Exp. Med.* **188**, 1705–1715 (1998).
33. Urbani, S. *et al.* Heterologous T cell immunity in severe hepatitis C virus infection. *J. Exp. Med.* **201**, 675–680 (2005).
34. Sharma, S. & Thomas, P. G. The two faces of heterologous immunity: protection or immunopathology. *J. Leukoc. Biol.* **95**, 405–416 (2014).
35. Macdonald, W. A. *et al.* T cell allorecognition via molecular mimicry. *Immunity* **31**, 897–908 (2009).
36. Wooldridge, L. *et al.* A single autoimmune T cell receptor recognizes more than a million different peptides. *J. Biol. Chem.* **287**, 1168–1177 (2012).
37. Christen, U. *et al.* A viral epitope that mimics a self antigen can accelerate but not initiate autoimmune diabetes. *J. Clin. Invest.* **114**, 1290–1298 (2004).

38. Reiser, J.-B. *et al.* CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nat. Immunol.* **4**, 241–247 (2003).
39. Ding, Y. H., Baker, B. M., Garboczi, D. N., Biddison, W. E. & Wiley, D. C. Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. *Immunity* **11**, 45–56 (1999).
40. Borbulevych, O. Y. *et al.* T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility. *Immunity* **31**, 885–896 (2009).
41. Cornberg, M. *et al.* CD8 T cell cross-reactivity networks mediate heterologous immunity in human EBV and murine vaccinia virus infections. *J. Immunol.* **184**, 2825–2838 (2010).
42. Haanen, J. B. A. G., Wolkers, M. C., Kruisbeek, A. M. & Schumacher, T. N. M. Selective Expansion of Cross-Reactive Cd8+ Memory T Cells by Viral Variants. *J. Exp. Med.* **190**, 1319–1328 (1999).
43. Clute, S. C. *et al.* Broad cross-reactive TCR repertoires recognizing dissimilar Epstein-Barr and influenza A virus epitopes. *J. Immunol.* **185**, 6753–6764 (2010).
44. Sidhom, J.-W. & Baras, A. S. Deep learning identifies antigenic determinants of severe SARS-CoV-2 infection within T-cell repertoires. *Sci. Rep.* **11**, 14275 (2021).
45. Shoukat, M. S. *et al.* Use of machine learning to identify a T cell response to SARS-CoV-2. *Cell Rep Med* **2**, 100192 (2021).
46. Klinger, M. *et al.* Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PLoS One* **10**, e0141561 (2015).
47. Sant, S. *et al.* Single-Cell Approach to Influenza-Specific CD8+ T Cell Receptor Repertoires Across Different Age Groups, Tissues, and Following Influenza Virus Infection. *Front. Immunol.* **9**, 1453 (2018).
48. Kaye, A. D. *et al.* Economic impact of COVID-19 pandemic on healthcare facilities and systems: International perspectives. *Best Pract. Res. Clin. Anaesthesiol.* **35**, 293–306 (2021).
49. Nundy, S., Ghosh, A., Mesloub, A., Albaqawy, G. A. & Alnaim, M. M. Impact of COVID-19 pandemic on socio-economic, energy-environment and transport sector globally and sustainable development goal (SDG). *J. Clean. Prod.* **312**, 127705 (2021).
50. Jasenosky, L. D., Scriba, T. J., Hanekom, W. A. & Goldfeld, A. E. T cells and adaptive immunity to Mycobacterium tuberculosis in humans. *Immunol. Rev.* **264**, 74–87 (2015).
51. Miller, A. *et al.* Correlation between universal BCG vaccination policy and reduced mortality for COVID-19. *bioRxiv* (2020) doi:10.1101/2020.03.24.20042937.
52. Berg, M. K., Yu, Q., Salvador, C. E., Melani, I. & Kitayama, S. Mandated Bacillus Calmette-Guérin (BCG) vaccination predicts flattened curves for the spread of COVID-19. *Sci Adv* **6**, eabc1463 (2020).
53. Charoenlap, S., Piromsopa, K. & Charoenlap, C. Potential role of Bacillus Calmette-Guérin (BCG) vaccination in COVID-19 pandemic mortality: Epidemiological and Immunological aspects. *Asian Pac. J. Allergy Immunol.* **38**, 150–161 (2020).
54. Escobar, L. E., Molina-Cruz, A. & Barillas-Mury, C. BCG vaccine protection from severe coronavirus disease 2019 (COVID-19). *Proc. Natl. Acad. Sci. U. S. A.* **117**, 17720–17726 (2020).
55. Moorlag, S. J. C. F. M., Arts, R. J. W., van Crevel, R. & Netea, M. G. Non-specific effects of BCG vaccine on viral infections. *Clin. Microbiol. Infect.* **25**, 1473–1478 (2019).
56. Uthayakumar, D. *et al.* Non-specific Effects of Vaccines Illustrated Through the BCG Example: From Observations to Demonstrations. *Front. Immunol.* **9**, 2869 (2018).
57. Saeed, S. *et al.* Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* **345**, 1251086 (2014).
58. Mehta, S. & Jeffrey, K. L. Beyond receptors and signaling: epigenetic factors in the regulation of innate immunity. *Immunol. Cell Biol.* **93**, 233–244 (2015).
59. Arts, R. J. W. *et al.* BCG Vaccination Protects against Experimental Viral Infection in Humans

through the Induction of Cytokines Associated with Trained Immunity. *Cell Host Microbe* **23**, 89–100.e5 (2018).

60. Moulson, A. J. & Av-Gay, Y. BCG immunomodulation: From the 'hygiene hypothesis' to COVID-19. *Immunobiology* **226**, 152052 (2021).
61. Jehi, L. *et al.* Individualizing Risk Prediction for Positive Coronavirus Disease 2019 Testing: Results From 11,672 Patients. *Chest* **158**, 1364–1375 (2020).
62. Noale, M. *et al.* The Association between Influenza and Pneumococcal Vaccinations and SARS-Cov-2 Infection: Data from the EPICOVID19 Web-Based Survey. *Vaccines (Basel)* **8**, (2020).
63. Pawlowski, C. *et al.* Exploratory analysis of immunization records highlights decreased SARS-CoV-2 rates in individuals with recent non-COVID-19 vaccinations. *Sci. Rep.* **11**, 4741 (2021).
64. Fink, G. *et al.* Inactivated trivalent influenza vaccine is associated with lower mortality among Covid-19 patients in Brazil. medRxiv 2020.06. 29.20142505. *DOI: https://doi. org/10. 1101/2020. 06* **29**, (2020).
65. Cohen, I. R. Antigenic mimicry, clonal selection and autoimmunity. *J. Autoimmun.* **16**, 337–340 (2001).
66. Netea, M. G. *et al.* Defining trained immunity and its role in health and disease. *Nat. Rev. Immunol.* **20**, 375–388 (2020).
67. Eldanasory, O. A., Rabaan, A. A. & Al-Tawfiq, J. A. Can influenza vaccine modify COVID-19 clinical course? *Travel Med. Infect. Dis.* **37**, 101872 (2020).
68. Pallikkuth, S., Williams, E., Pahwa, R., Hoffer, M. & Pahwa, S. Association of Flu specific and SARS-CoV-2 specific CD4 T cell responses in SARS-CoV-2 infected asymptomatic heath care workers. *Vaccine* **39**, 6019–6024 (2021).
69. Salem, M. L. & El-Hennawy, D. The possible beneficial adjuvant effect of influenza vaccine to minimize the severity of COVID-19. *Med. Hypotheses* (2020).
70. Lehner, G. F. *et al.* Correlation of interleukin-6 with Epstein–Barr virus levels in COVID-19. *Crit. Care* **24**, 1–3 (2020).
71. Nadeem, A., Suresh, K., Awais, H. & Waseem, S. Epstein-Barr Virus Coinfection in COVID-19. *J Investig Med High Impact Case Rep* **9**, 23247096211040626 (2021).
72. Vigón, L. *et al.* Impaired Antibody-Dependent Cellular Cytotoxicity in a Spanish Cohort of Patients With COVID-19 Admitted to the ICU. *Front. Immunol.* **12**, 742631 (2021).
73. Tan, S. S., Chew, K. L., Saw, S., Jureen, R. & Sethi, S. Cross-reactivity of SARS-CoV-2 with HIV chemiluminescent assay leading to false-positive results. *Journal of clinical pathology* vol. 74 614 (2021).
74. Salih, R. Q. *et al.* False-positive HIV in a patient with SARS-CoV-2 infection; a case report. *Ann Med Surg (Lond)* **71**, 103027 (2021).
75. Kliger, Y. & Levanon, E. Y. Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy. *BMC Microbiol.* **3**, 20 (2003).
76. Mishra, N. *et al.* Cross-neutralization of SARS-CoV-2 by HIV-1 specific broadly neutralizing antibodies and polyclonal plasma. *PLoS Pathog.* **17**, e1009958 (2021).
77. Fendler, A. *et al.* Functional antibody and T cell immunity following SARS-CoV-2 infection, including by variants of concern, in patients with cancer: the CAPTURE study. *Nat Cancer* **2**, 1321–1337 (2021).
78. Mehandru, S. & Merad, M. Pathological sequelae of long-haul COVID. *Nat. Immunol.* **23**, 194–202 (2022).
79. Ehrenfeld, M. *et al.* Covid-19 and autoimmunity. *Autoimmun. Rev.* **19**, 102597 (2020).
80. Galeotti, C. & Bayry, J. Autoimmune and inflammatory diseases following COVID-19. *Nature reviews. Rheumatology* vol. 16 413–414 (2020).
81. Dotan, A. *et al.* The SARS-CoV-2 as an instrumental trigger of autoimmunity. *Autoimmun. Rev.* **20**, 102792 (2021).
82. Barzilai, O., Ram, M. & Shoenfeld, Y. Viral infection can induce the production of

autoantibodies. *Curr. Opin. Rheumatol.* **19**, 636–643 (2007).
83. Lefranc, M.-P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2014).
84. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
85. Gowthaman, R. & Pierce, B. G. TCR3d: The T cell receptor structural repertoire database. *Bioinformatics* **35**, 5323–5325 (2019).
86. Bagaev, D. V. *et al.* VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* **48**, D1057–D1062 (2020).
87. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
88. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
89. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

Figure legends

Figure 1. Overview of this study. (A) Outline of the Multiplexed Identification of T cell Receptor Antigen (MIRA) assay and the corresponding dataset available from the immunoACCESS© project web resource[15]. (B) Analytic steps in this study, regarding the novel utilisation of the MIRA dataset for training Machine Learning algorithms that can highlight important SARS-CoV-2 antigens for distinguishing samples between healthy and COVID-19-convalescent cohorts. (C) Strategy for exploring T cell cross-reactivity between SARS-CoV-2 and other pathogens and diseases.

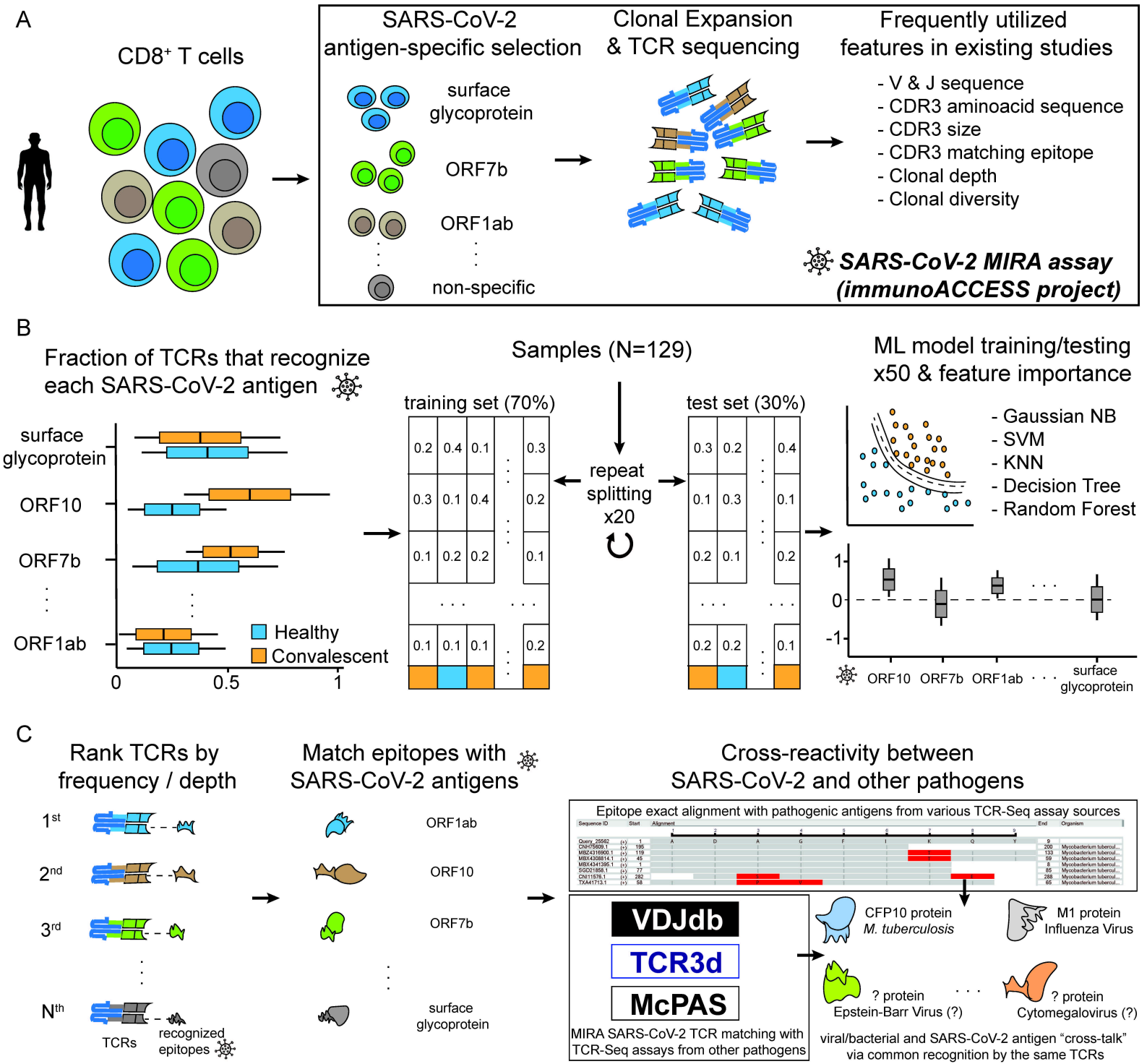
Figure 2. Exploratory analysis of the Multiplexed Identification of T cell Receptor Antigen (MIRA) dataset. (A) Number of samples in each MIRA cohort. (B) Per sample normalised number of unique T cell receptors (TCRs) in the healthy and convalescent cohorts. (C) Per sample normalised number of TCRs that recognise each SARS-CoV-2 antigen in the healthy and convalescent cohorts. (D) Projection of healthy and convalescent samples on the principal component analysis (PCA) space. Healthy and convalescent distributions in (B) and (C) were compared with the Mann-Whitney test.

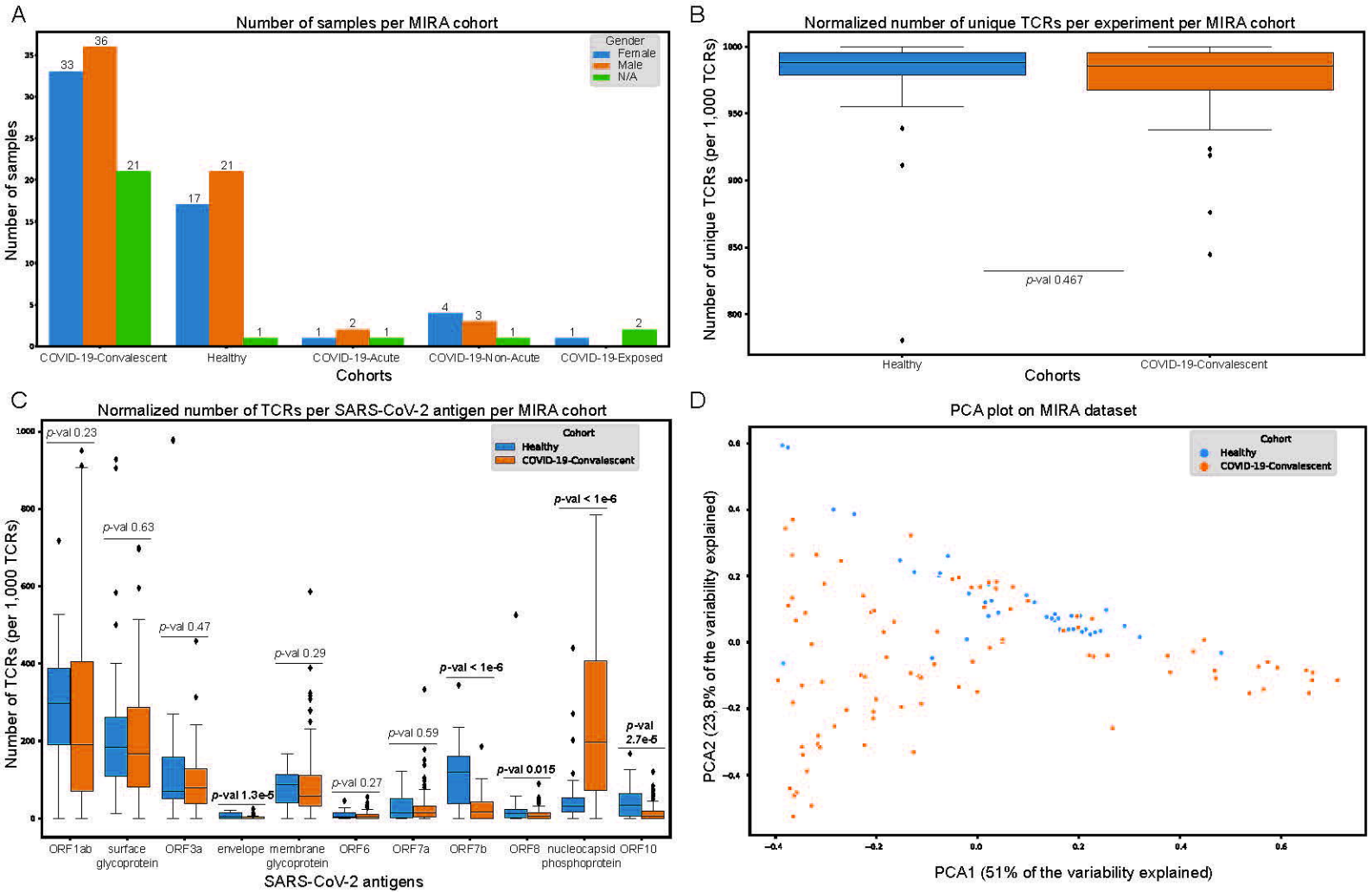
Figure 3. Evaluation of Machine Learning (ML) algorithms trained on the healthy and convalescent cohorts in the Multiplexed Identification of T cell Receptor Antigen (MIRA) dataset. (A) Balanced accuracy, precision, sensitivity, specificity and negative predictive value (NPV) of each algorithm after selecting a prediction score cut-off of 0.5. (B) Support Vector Machines (SVM) performance on multiple prediction score cut-offs. (C) Feature importance score after 50 permutations on all 20 randomly generated test sets. (D) ML algorithms' performance after selecting only the important features for each algorithm and retraining.

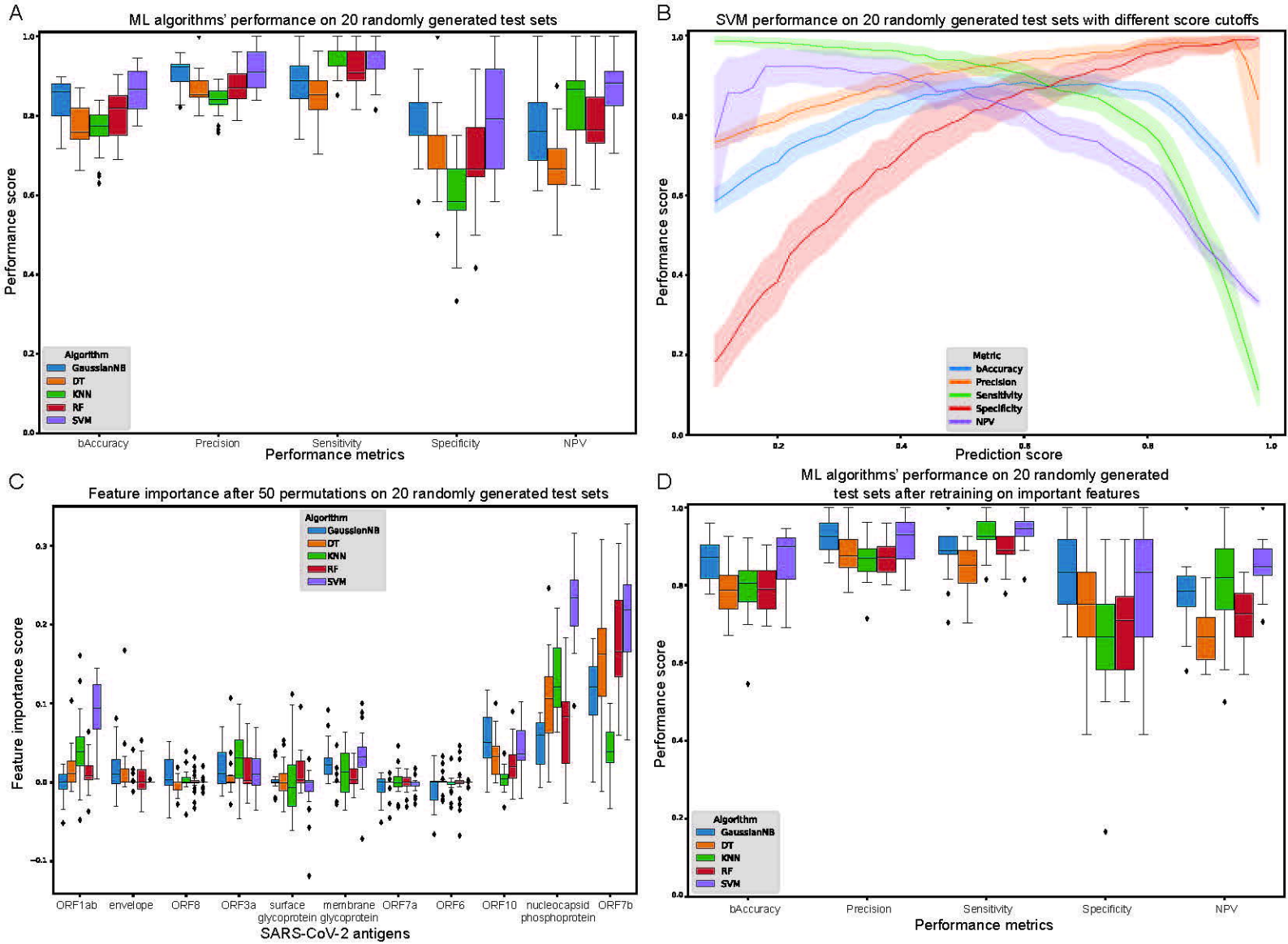
Figure 4. Exploration of the most common Multiplexed Identification of T cell Receptor Antigen (MIRA) T cell receptors (TCRs) in terms of clonal expansion and cross-reactivity. (A) Occurrence frequency and clonal expansion of all MIRA TCRs. The arrows point to the 6 most common TCRs (present in at least 11.5% of total number of subjects) that were further analysed in terms of clonal expansion in the two cohorts (B). There was no statistically significant differential expansion detected between the two cohorts based on Mann-Whitney test; however, the statistical test could not be performed for some TCRs (denoted as *p*-val N/A). (C) Cohort distribution of the 1st and 4th most common MIRA TCRs that were found to be enriched in either cohort after applying Fisher's exact test. (D) Secondary structure of surface glycoprotein (SARS-CoV-2) and Matrix protein 1 (M1) with cross-reactive sections, based on the most common MIRA TCR, highlighted with red

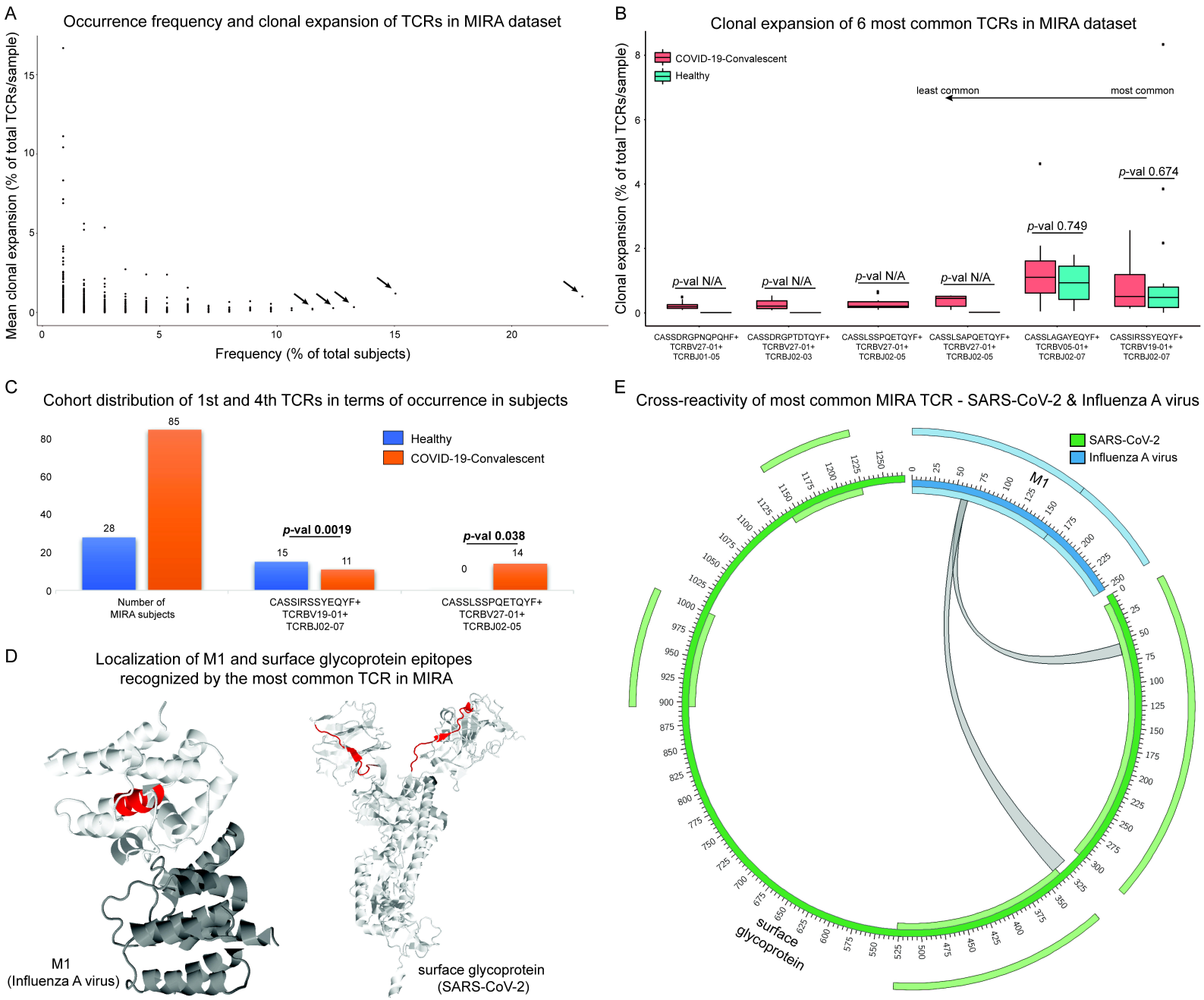
colour. 1AA7 (A and B chain view) and 6X29 (A chain view) Protein Data Bank[89] entries were used for M1 and surface glycoprotein, respectively. (E) Circular plot, as an alternative view of (D), depicting the cross-reactive property of the most common MIRA TCR that recognizes epitopes from surface glycoprotein (SARS-CoV-2) and M1 (Influenza A virus). The inner and outer light-colored tracks represent the annotated domains.

Figure 5. Cross-reactivity analysis of all Multiplexed Identification of T cell Receptor Antigen (MIRA) T cell receptors (TCRs). (A) Heatmap of unique MIRA complementarity-determining region 3 (CDR3) counts that exhibit cross-reactivity between SARS-CoV-2 (x-axis) and other pathogens and diseases (y-axis). (B) Circular plot that depicts the cross-reactivity of MIRA CDR3 regions between antigens that originate from SARS-CoV-2 and a selected subset of pathogens from (A). The inner and outer light-coloured tracks represent the annotated protein domains. Each connection represents the ability of a single CDR3 region to recognize a part of a SARS-CoV-2 antigen and a part of another pathogen's protein. The connections are coloured based on their corresponding non-SARS-CoV-2 pathogens. NP, SG, EP and MG stand for nucleocapsid phosphoprotein, surface glycoprotein, envelope and membrane glycoprotein.

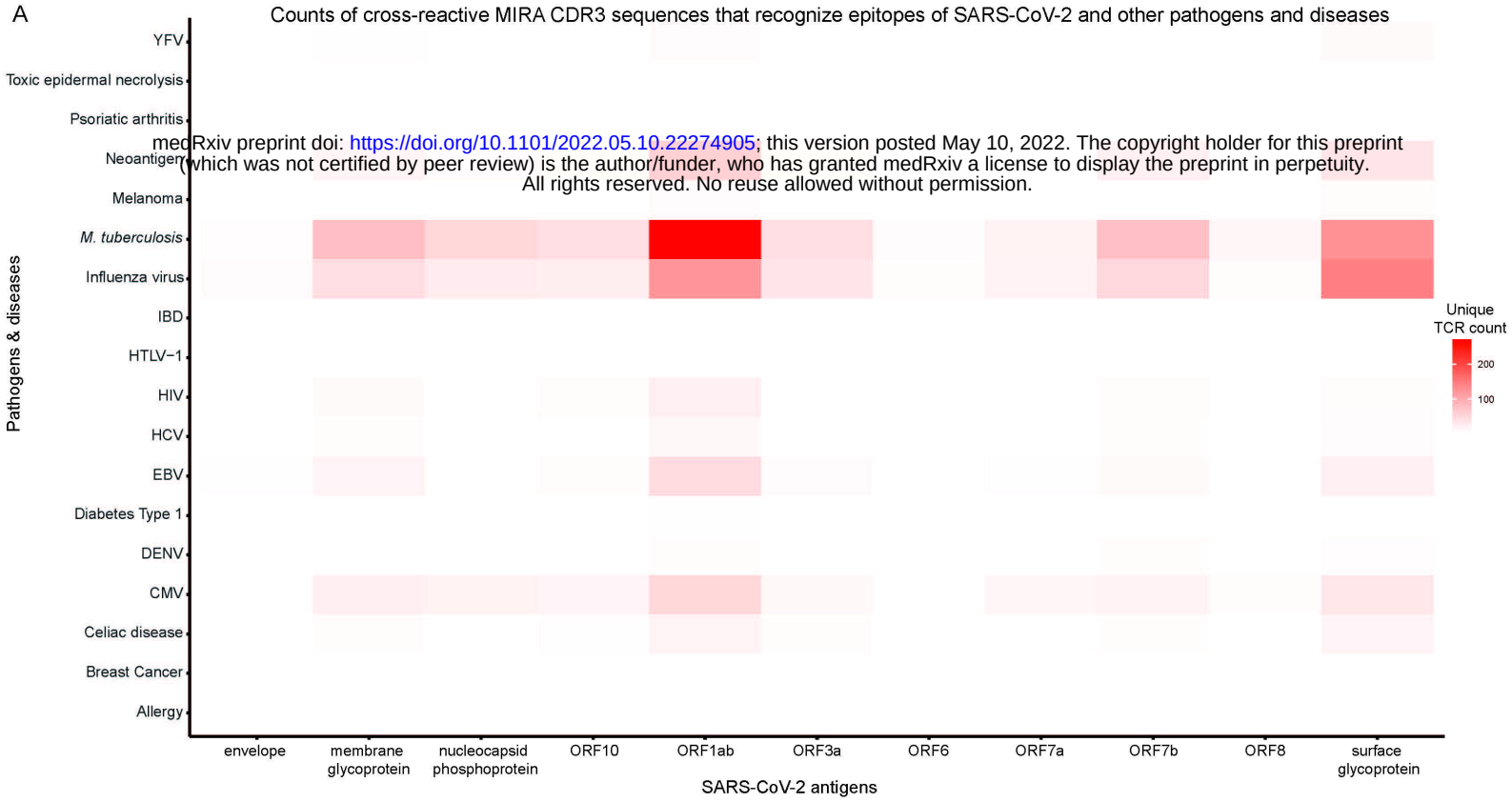

A
CD8+ T cells
SARS-CoV-2 antigen-specific selection
surface glycoprotein
ORF7b
ORF1ab
non-specific
Clonal Expansion & TCR sequencing
Frequently utilized features in existing studies
- V & J sequence
- CDR3 aminoacid sequence
- CDR3 size
- CDR3 matching epitope
- Clonal depth
- Clonal diversity
SARS-CoV-2 MIRA assay (immunoACCESS project)
B
Fraction of TCRs that recognize each SARS-CoV-2 antigen
surface glycoprotein
ORF10
ORF7b
ORF1ab
Healthy
Convalescent
0
0.5
1
Samples (N=129)
training set (70%)
repeat splitting x20
test set (30%)
ML model training/testing x50 & feature importance
- Gaussian NB
- SVM
- KNN
- Decision Tree
- Random Forest
ORF10
ORF7b
ORF1ab
surface glycoprotein
C
Rank TCRs by frequency / depth
1st
2nd
3rd
Nth
TCRs
recognized epitopes
Match epitopes with SARS-CoV-2 antigens
ORF1ab
ORF10
ORF7b
surface glycoprotein
Cross-reactivity between SARS-CoV-2 and other pathogens
Epitope exact alignment with pathogenic antigens from various TCR-Seq assay sources
VDJdb
TCR3d
McPAS
MIRA SARS-CoV-2 TCR matching with TCR-Seq assays from other pathogens
CFP10 protein
M. tuberculosis
M1 protein
Influenza Virus
? protein
Epstein-Barr Virus (?)
? protein
Cytomegalovirus (?)
viral/bacterial and SARS-CoV-2 antigen "cross-talk" via common recognition by the same TCRs


A
Number of samples per MIRA cohort
Gender
Female
Male
N/A
Number of samples
33
36
21
17
21
1
1
2
1
4
3
1
1
2
COVID-19-Convalescent
Healthy
COVID-19-Acute
COVID-19-Non-Acute
COVID-19-Exposed
Cohorts
B
Normalized number of unique TCRs per experiment per MIRA cohort
Number of unique TCRs (per 1,000 TCRs)
p-val 0.467
Healthy
COVID-19-Convalescent
Cohorts
C
Normalized number of TCRs per SARS-CoV-2 antigen per MIRA cohort
Cohort
Healthy
COVID-19-Convalescent
Number of TCRs (per 1,000 TCRs)
p-val 0.23
p-val 0.63
p-val 0.47
p-val 1.3e-5
p-val 0.29
p-val 0.27
p-val 0.59
p-val < 1e-6
p-val 0.015
p-val < 1e-6
p-val 2.7e-5
ORF1ab
surface glycoprotein
ORF3a
envelope
membrane glycoprotein
ORF6
ORF7a
ORF7b
ORF8
nucleocapsid phosphoprotein
ORF10
SARS-CoV-2 antigens
D
PCA plot on MIRA dataset
Cohort
Healthy
COVID-19-Convalescent
PCA2 (23.8% of the variability explained)
PCA1 (51% of the variability explained)


A
ML algorithms' performance on 20 randomly generated test sets
Performance score
Algorithm
GaussianNB
DT
KNN
RF
SVM
bAccuracy
Precision
Sensitivity
Specificity
NPV
Performance metrics
B
SVM performance on 20 randomly generated test sets with different score cutoffs
Performance score
Metric
bAccuracy
Precision
Sensitivity
Specificity
NPV
Prediction score
C
Feature importance after 50 permutations on 20 randomly generated test sets
Feature importance score
Algorithm
GaussianNB
DT
KNN
RF
SVM
ORF1ab
envelope
ORF8
ORF3a
surface glycoprotein
membrane glycoprotein
ORF7a
ORF6
ORF10
nucleocapsid phosphoprotein
ORF7b
SARS-CoV-2 antigens
D
ML algorithms' performance on 20 randomly generated test sets after retraining on important features
Performance score
Algorithm
GaussianNB
DT
KNN
RF
SVM
bAccuracy
Precision
Sensitivity
Specificity
NPV
Performance metrics


A
Occurrence frequency and clonal expansion of TCRs in MIRA dataset
Mean clonal expansion (% of total TCRs/sample)
Frequency (% of total subjects)
B
Clonal expansion of 6 most common TCRs in MIRA dataset
COVID-19-Convalescent
Healthy
least common
most common
p-val N/A
p-val N/A
p-val N/A
p-val N/A
p-val 0.749
p-val 0.674
CASSDRGPNQPQHF+ TCRBV27-01+ TCRBJ01-05
CASSDRGPTDTQYF+ TCRBV27-01+ TCRBJ02-03
CASSLSSPQETQYF+ TCRBV27-01+ TCRBJ02-05
CASSLSAPQETQYF+ TCRBV27-01+ TCRBJ02-05
CASSLAGAYEQYF+ TCRBV05-01+ TCRBJ02-07
CASSIRSSYEQYF+ TCRBV19-01+ TCRBJ02-07
C
Cohort distribution of 1st and 4th TCRs in terms of occurrence in subjects
Healthy
COVID-19-Convalescent
28
85
p-val 0.0019
15
11
p-val 0.038
0
14
Number of MIRA subjects
CASSIRSSYEQYF+ TCRBV19-01+ TCRBJ02-07
CASSLSSPQETQYF+ TCRBV27-01+ TCRBJ02-05
D
Localization of M1 and surface glycoprotein epitopes recognized by the most common TCR in MIRA
M1 (Influenza A virus)
surface glycoprotein (SARS-CoV-2)
E
Cross-reactivity of most common MIRA TCR - SARS-CoV-2 & Influenza A virus
SARS-CoV-2
Influenza A virus
M1
surface glycoprotein

A



B