TRABAJO FIN DE MÁSTER

MÁSTER EN BIOINFORMÁTICA

SARS-CoV-2-specific T cell receptors after disease and vaccination

Autora

María Soledad Benítez Cantos

Directores

Carlos Cano Gutiérrez

Juana María Vivo Molina

Facultad de Biología

Universidad de Murcia

—

Murcia, septiembre de 2021

Contents

Abstract

Immunization against SARS-CoV-2 following infection or vaccination has been extensively studied from the perspective of antibody responses. However, T cells also play an important role in protection against COVID-19, conferring in some cases a more durable immunization. Signals of past and present infections are encoded in the set of up to 10^{10} different T cell receptors (TCRs) that bind to antigens to trigger an immune response. A tremendous effort has been made for *in vitro* identification of SARS-CoV-2-specific TCR sequences, and in 2021 the first TCR repertoire studies in disease and vaccination have been published. Nevertheless, these analyses are often shallow and centered only in SARS-CoV-2 TCRs, disregarding the information from the rest of the repertoire. In the present study, published TCR repertoires of 19 COVID-19 convalescent, 19 Ad26.COV2.S (Janssen) vaccinated and 5 placebo recipient individuals have been re-analyzed with a different SARS-CoV-2-specific TCRs dataset and from a sequence similarity perspective. Convalescent and vaccinated cohorts exhibited a similar TCR response to the viral spike protein –the single antigen in the vaccine– in terms of breadth, depth and epitope location, whereas response to non-spike antigens was significantly higher in convalescent subjects compared with vaccinated and placebo cohorts. Furthermore, sequence similarity network analysis revealed that putatively unspecific TCRs (i.e. any SARS-CoV-2 TCRs in placebos and non-spike TCRs in vaccinated) are less connected in the repertoire network, suggesting that they may be cross-reactive to other antigens. These findings aim to broaden the understanding of how SARS-CoV-2 exposure shapes the TCR repertoire in disease and vaccination.

Resumen

La inmunización frente a SARS-CoV-2 tras la infección o la vacunación ha sido ampliamente estudiada desde la perspectiva de las respuestas de anticuerpos. Sin embargo, las células T también tienen un papel importante en la protección contra la COVID-19, confiriendo en algunos casos una inmunidad más duradera. Las señales de infecciones pasadas y presentes quedan reflejadas en el conjunto de hasta 10^{10} receptores de células T (TCRs) distintos que se unen a antígenos para desencadenar una respuesta inmunitaria. Se han realizado grandes esfuerzos para la identificación *in vitro* de secuencias de TCR específicas de SARS-CoV-2, y en 2021 se han publicado los primeros estudios de repertorios de TCR tras la enfermedad y la vacunación. No obstante, estos análisis suelen ser someros y se centran exclusivamente en los TCRs asociados a SARS-CoV-2, despreciando la información del resto del repertorio. En el presente estudio, repertorios de TCR públicos de 19 individuos convalecientes de COVID-19, 19 individuos que recibieron la vacuna Ad26.COV2.S (Janssen) y 5 individuos que recibieron un placebo han sido analizados de nuevo con un conjunto distinto de TCRs asociados a SARS-CoV-2 y con un enfoque de similitud de secuencia. Los individuos convalecientes y vacunados exhibieron una respuesta similar de TCR ante la proteína *spike* –el único antígeno presente en la vacuna– en términos de amplitud, profundidad y localización de epítopos, mientras que la respuesta al

resto de antígenos virales fue significativamente más alta en los sujetos convalecientes en comparación con las cohortes de vacunados y placebos. Además, el análisis de redes de similitud de secuencia reveló que aquellos TCRs putativamente inespecíficos (cualquier TCR asociado a SARS-CoV-2 en placebos y TCRs asociados a proteínas distintas a *spike* en vacunados) están menos conectados en la red del repertorio, sugiriendo que pueden presentar reactividad cruzada con otros antígenos. Estos hallazgos pretenden ampliar el conocimiento sobre cómo la exposición a SARS-CoV-2 influye en la composición del repertorio de TCR tras la enfermedad y la vacunación.

Introduction

Coronavirus disease 2019 (COVID-19), caused by the novel human pathogen severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a highly transmissible disease that has resulted in a widespread global pandemic (Hu et al., 2021). The understanding of the immunology of COVID-19 has rapidly evolved since early 2020, with a focus on vaccine development. From December 2020 to June 2021, 7 different vaccines have been listed for World Health Organization (WHO) Emergency Use Listing. As of 30 August 2021, a total of 5,019,907,027 vaccine doses have been administered worldwide (WHO, 2021).

The adaptive immune system is key for a successful response to most viral infections. It is composed by three main elements: B cells, which produce antibodies, $CD4^+$ T cells with helper and effector functionalities and $CD8^+$ T cells that kill infected cells. The activaction of these cells relies on the recognition of foreign antigenic proteins. Neutralizing antibodies bind to regions of viral antigens (called epitopes) located in the protein surface and aim to block the attachment of the virus to the human host cell, thus preventing cell infection. Most current vaccines aim to produce an antibody response, but although it is critical for virus neutralization and disease control, B cell responses to SARS-CoV-2 have limited duration and breadth (Sauer and Harris, 2020). The role of T cells in COVID-19 infection and their importance in vaccines are gaining interest among the scientific community since T cells are major mediators of long-term memory and persist much longer than antibodies (Harris and Sauer, 2021). The importance of T cells is further supported by the T cell lymphopenia (low lymphocyte counts in peripheral blood) upon COVID-19 infection that correlates with disease severity (Liu et al., 2020).

The T cell receptors (TCR), located on the cellular membrane surface, are the T cells equivalent of B cell receptors (a membrane-bound version of antibodies). Unlike antibodies, these receptors are not capable of direct binding to a viral protein, but they require that it has been previously processed either by infected cells or by antigen presenting cells. These cells then display the antigenic epitopes on their major histocompatibility complex (MHC) surface membrane molecules, and the TCR binds to both the MHC and the epitope before its activation. $CD8^+$ T cells recognize epitopes presented by MHC class I molecules, whereas $CD4^+$ T cells bind to epitopes in MHC class II molecules (Murphy and Weaver, 2016) (Fig. 1a).

To ensure an adaptive immunity response to any pathogen, the T cell pool of an individual should contain billions of different clones based on the sequence of their TCR, in an attempt to cover the vast range of possible foreign antigens. Since it would be inefficient to encode such a large number of different TCR genes in a genome, different TCR sequences are generated by somatic recombination of variable (V), diversity (D) and joining (J) gene segments (Fig. 1b). In the case of TCRβ chains, 1 V segment out of 65, 1 J segment out of 13 and 1 D segment out of 2 are randomly selected and joined in the T cell genome during its maduration in the thymus. This somatic recombination process is not clean, as nucleotides can be additioned or deleted in the segment junctions, thus contributing to sequence variability. This highly variable region of the TCR is called complementarity-determining region 3 (CDR3) and it encodes the

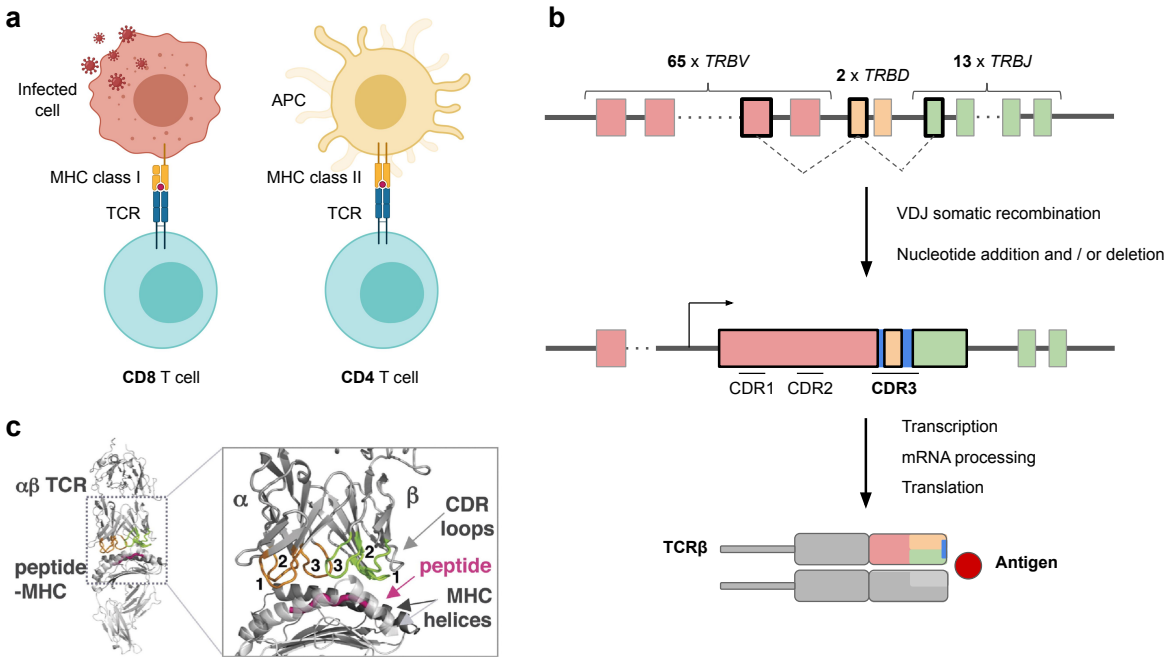


Figure 1: Basis of adaptive T cell immunity. (a) $CD8^+$ recognize epitopes presented in MHC class I molecules by infected cells, whereas $CD4^+$ T cells bind to epitopes presented in MHC class II molecules by antigen presenting cells (APCs). Figure created using BioRender (https://biorender.com/). **(b)** A simplified scheme representing the process of VDJ somatic recombination in T cells genome. One of each V, D and J genetic segments is selected randomly and joined in the genome, with nucleotide additions and / or deletions, generating the CDR3 region. **(c)** A 3D representation of TCR-peptide-MHC complex (left) and a closeup of the interface (right) showing the CDR1 and 2 TCR loops contacting the MHC, while the genetically recombined CDR3 contact the antigenic peptide bound to the MHC. Figure taken from Garcia and Adams (2005)

TCR protein region in direct contact with the antigen (Fig. 1c). This is why the identity of a T lymphocyte or clonotype is usually defined by its CDR3 amino acid sequence and sometimes also by its V segment, which contains CDR1 and 2, responsible of MHC binding (Garcia and Adams, 2005).

VDJ recombination of the different TCR genes could theoretically generate between 10^{15} and 10^{20} TCR chains. Despite this, the actual diversity present in an adult human is estimated at around 10^9-10^{10} different clonotypes, implying that TCR development is subject to different generation probabilities (Lythe et al., 2016). Due to this massive level of sequence variation, TCR repertoire analysis poses a series of challenges, from the wet lab techniques to amplify the TCR sequences, to the bioinformatic analysis of the complex datasets generated.

The TCR repertoire records past and ongoing immune responses, and can be regarded as an ultimate example of a high-dimensional and intimately personalized biomarker of the adaptive immunity landscape of an individual, beign the hallmark of precision medicine of the future (Heather et al., 2018). Yet this powerful information is difficult to interpret by itself due to the following reasons: a) TCR repertoires, although highly diverse, have little overlap across individuals; b) an individual's repertoire is a complex compendium of immune states; c) there may be few and low-frequency instances of antigen-specific TCRs; d) an antigen may be recognized by different TCRs, and some TCRs may be cross-reactive to unrelated antigens (many-to-many relationship); and e) information can be retrieved from multiple spaces of the TCR repertoire (frequencies distributions, sequence similarity and

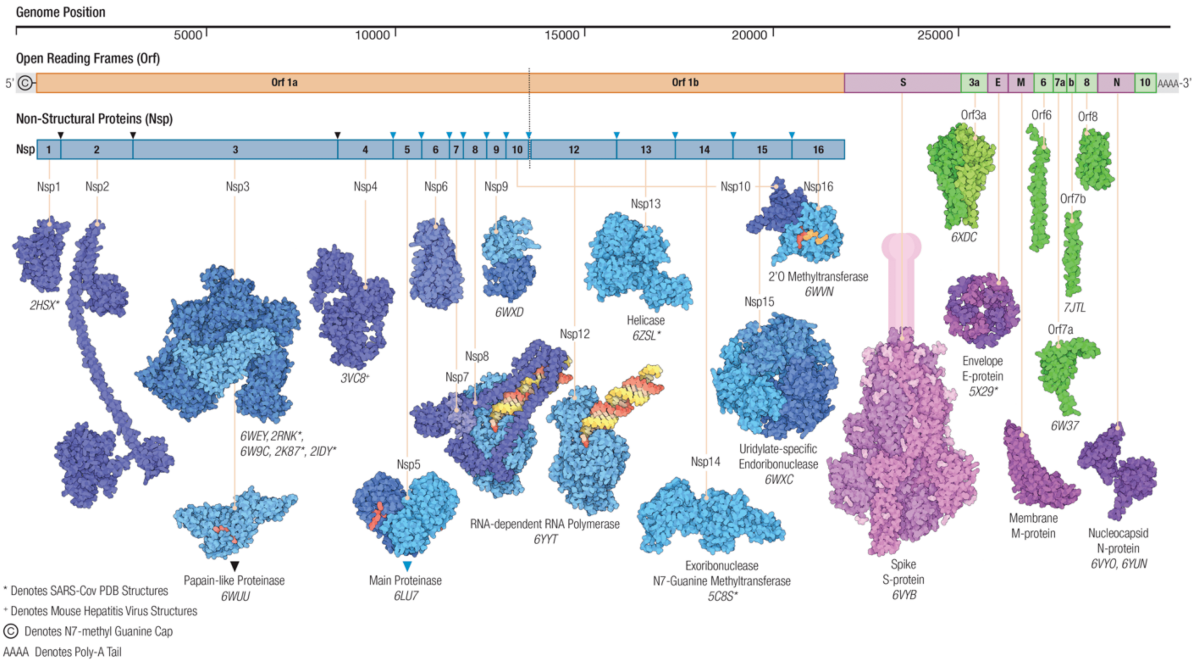


Figure 2: Architecture of the SARS-CoV-2 genome and proteome. Virion structural proteins spike (S), envelope (E), membrane (M) and nucleocapsid (N) are showed in purple. Current vaccines only elicit immunity to spike protein. Non-structural proteins are in shades of blue and open reading frame (Orf) proteins appear in green. Figure taken from Lubin et al. (2020)

3D-structural space). Thus, TCR repertoire data highly benefits from data analysis techniques such as machine learning or network analysis (Greiff et al., 2020).

In an attempt to shed light in TCR biomarker discovery, researchers have used *in vitro* antigen-enrichment of TCR repertoires, such as peptide-MHC tetramer sorting assays or Multiplexed Identification of T cell Receptor Antigen specificity assays (MIRA). In brief, antigen-specific TCR discovery with this techniques consists in exposing T cells from a blood sample to known epitopes fixed in a reagent and then sequence the TCRs of the responding T cells (Klinger et al., 2015). Then, this antigen-specific TCRs can be used to screen bulk TCR repertoires in a cohort of interest, but the constraints of high inter-individual diversity and TCR cross-reactivity may still make it difficult to find these TCRs in the population (Mayer-Blackwell et al., 2020).

For many pathologies the major impediment to their study from an immunological repertoires perspective is the lack of these *in-vitro*-determined antigen-specific TCRs (Greiff et al., 2020), but the scenario is somewhat better for COVID-19. In July 2020 Adaptive Biotechnologies and Microsoft publicly released data of MIRA assays including more than 160,000 SARS-CoV-2-specific TCRs recognizing 269 different SARS-CoV-2 epitopes (Nolan et al., 2020). These epitopes covered most of SARS-CoV-2 proteins, shown in Fig. 2.

Several articles about SARS-CoV-2 vaccines eliciting robust antibody responses against SARS-CoV-2 spike (S) protein were published in high impact journals during their early phase clinical trials (Sahin et al., 2020; Stephenson et al., 2021; Ewer et al., 2021) and later came the first preprints analyzing the immunogenicity of vaccines from a T cell and TCR repertoire perspective (Minervina et al., 2021; Alter et al., 2021; Swanson et al., 2021). However, these studies tend to analyze TCR repertoires in a very narrow manner, limiting their analysis to measure the frequencies of the SARS-CoV-2-specific TCRs

found in a repertoire by an exact sequence match, and therefore probably underestimating the TCR response in COVID-19.

The present study aims to analyze in depth the published TCR repertoires of COVID-19 convalescent individuals, Ad26.COV2.S vaccine (developed by Janssen Pharmaceutica) recipient individuals and placebo recipient individuals to fully characterize how viral and vaccine immunization alter the complex landscape of TCR repertoires in health and disease.

Material and Methods

TCR data

This study is based on public data from three previous works (Alter et al., 2021; Nolan et al., 2020; Mayer-Blackwell et al., 2020).

The dataset used for TCR repertoire analysis (Alter et al., 2021) includes samples from 32 individuals: 8 convalescent from COVID-19, 19 who have received the Ad26.COV2.S vaccine developed by Janssen Pharmaceutica during a clinical trial, and 5 subjects who have received a placebo. Peripheral blood samples were collected post diagnosis or vaccination, and immunosequencing of the CDR3 regions of human TCRβ chains was performed with the immunoSEQ Assay (Adaptive Biotechnologies). Data was accessed on July 2021 via Adaptive Biotechnologies immuneACCESS® database (immuneACCESS® DOI: https://doi.org/10.21417/GA2021N).

To match the sample size of vaccinated individuals with data generated by the same procedure, 11 TCR repertoire samples from COVID-19-convalescent subjects were randomly selected from the COVID-19-HUniv12Oct dataset on Adaptive Biotechnologies ImmuneCODE™ project (Nolan et al., 2020). The full dataset contains TCRβ repertoires from 193 convalescent patients whose blood sample was collected at the Hospital Univesitario 12 de Octubre (Madrid, Spain). Data was accessed on Aug 2021 via Adaptive Biotechnologies immuneACCESS® database (immuneACCESS® DOI: https://doi.org/10.21417/ADPT2020COVID, ImmuneCODE-COVID-Release-002).

SARS-CoV-2-specific $CD8^+$ TCRβ sequences were obtained from Mayer-Blackwell et al. (2020). This sequences are proven to bind SARS-CoV-2 epitopes by MIRA assays (Nolan et al., 2020) and are also enriched in bulk TCR repertoires of convalescent individuals compared to healthy controls. For the present study, only TCR sequences with a strong evidence of HLA restriction (N = 1831) were taken into consideration.

Preliminary TCR repertoire analysis

T cell proportion estimated by TCR sequencing, as well as Simpson clonality indices for all repertoires were available in Adaptive Biotechnologies samples metadata. Simpson clonality is calculated for a repertoire as the square root of Simpson's diversity index for all unique TCRs:

$$\text{Simpson clonality} = \sqrt{\sum_{i}^{N} p_i^2}$$

where p_i the relative frequency of TCR i in a sample with N unique TCRs.

SARS-CoV-2-specific CD4⁺ TCRs discovery

While SARS-CoV-2-specific CD4⁺ have been used to annotate TCR repertoires in previous studies (Alter et al., 2021; Gittelman et al., 2021), those enriched and high-reliable datasets are not currently public. ImmuneCODE™ project contains an unenriched dataset of 6809 CD4⁺ TCRs that bind 49 different SARS-CoV-2 epitopes presented by class II MHC molecules in MIRA assays. Data was accessed on Aug 2021 via Adaptive Biotechnologies immuneACCESS® database (immuneACCESS® DOI: https://doi.org/10.21417/ADPT2020COVID, ImmuneCODE-COVID-Release-002).

These TCRs were further screened for enrichment compared to a background of healthy individuals repertoires in order to remove TCRs that may be highly public or cross-reactive to common antigens. 64 TCRs were selected to annotate the repertoires, in addition to the CD8⁺ dataset. The enrichment analysis was performed with tcrdist3 Python toolkit (Docker image v0.1.9) (Mayer-Blackwell et al., 2020; Dash et al., 2017), following the same meta-clonotype discovery pipeline employed for SARS-CoV-2 CD8⁺ TCR discovery as in Mayer-Blackwell et al. (2020).

In brief, firstly this pipeline groups all the *in vitro* determined SARS-CoV-2-specific TCRs by the recognized epitope, which result in sets of similar TCR sequences. Then, for each TCR, its sequence distance with the rest of TCRs in the group is calculated, as well as with a control set of TCRs with matched V and J gene frequencies. An optimal distance radius is selected based on the propotion of SARS-CoV-2-specific TCRs and control TCRs included within it, always trying to minimize the latter below a given threshold. If this optimal radius is found for a given SARS-CoV-2-specific TCR, it is considered enriched, along with the rest of SARS-CoV-2-specific TCRs within the radius.

Measurement of T cell response to SARS-CoV-2

The 43 TCRβ repertoires were annotated for antigen-specificity with the SARS-CoV-2-specific TCRs (CD4⁺ and CD8⁺) by exact matching of CDR3 aminoacid sequence and V gene. The SARS-CoV-2 response of each individual to spike and non-spike proteins was measured in terms of breadth, defined as the proportion of distinct TCRs recognizing certain protein among all the unique sequences in a repertoire, and in terms of depth, which is the relative frequency of those SARS-CoV-2-specific TCRs.

Sequence similarity network analysis

To assess TCR sequence similarities in a repertoire, pairwise distances between all CDR3 sequences in a given sample were computed with tcrdist3 (Mayer-Blackwell et al., 2020; Dash et al., 2017), which implements a custom distance metric based on BLOSUM62 substitution matrix to account for similar aminoacid substitutions, and applies different weights depending on the importance of every CDR3

position in antigen binding. Total runtime was 103 hours with parallel processing (40 threads, 2.50 GHz, 256 GB of RAM).

Undirected graphs were built considering each unique V gene + CDR3 aminoacid sequence combination as a node. An edge was built between two nodes if their pairwise distance was $\leq$ 12. The reason behind this threshold is that 12 is the greatest possible distance between two CDR3 with one mismatch according to tcrdist3 algorithm. When the same V gene + CDR3 aminoacid sequence was encoded by more than one distinct TCR nucleotide sequence, a self-loop was added to the node. Edge weights (inverse distance between two nodes) were taken into account for hub score calulation. Networks were built and analyzed with R igraph package v1.2.6 (Csardi and Nepusz, 2006).

Data analysis and visualization

All plots and analyses were carried in R 3.6.1 (R Core Team, 2019). For data analysis, the packages dplyr v1.0.2 (Wickham et al., 2020), tidyr v1.1.2 (Wickham, 2020), rstatix v0.7.0 (Kassambara, 2021), factoextra v1.0.5 (Kassambara and Mundt, 2017) and parallel v3.6.1 (R Core Team, 2019) were used. Networks were plotted with the ggraph v2.0.2 package (Pedersen, 2020). SARS-CoV-2 genome schematic representation was made with gggenes v0.4.1 package (Wilkins, 2020). 3D visualization of the SARS-CoV-2 spike protein (PDB ID: 6XR8) was generated with Protein Imager (Tomasello et al., 2020). All other plots were generated with ggplot2 v3.3.2 (Wickham, 2016) and ggpubr v0.4.0 (Kassambara, 2020).

Processed data, code for generating the plots and the scripts to run tcrdist3 are available on https://github.com/marisolbc/masters_thesis_analyses.

Results

Preliminary TCR repertoire analysis

One of the clinical characteristics of SARS-CoV-2-infected patients, lymphopenia, can be observed from a TCR repertoire analysis perspective. In TCR repertoire sequencing from a peripheral blood sample, both the number of nucleated cells and total T cells can be estimated by the amplification of reference gene primers. The fraction of T cells was significantly lower in convalescent individuals (C) compared to vaccinated (V) and placebo (P) ($p = 2.5 \cdot 10^{-9}$ and $p = 5.6 \cdot 10^{-4}$, two-sided Wilcoxon rank-sum test) and no significant differences were observed between vaccinated and placebo subjects ($median_C = 0.12$, $IQR_C = 0.11$; $median_V = 0.49$, $IQR_V = 0.07$; $median_P = 0.45$, $IQR_P = 0.12$) (Fig. 3a). Some convalescent patients TCRβ repertoires had a higher clonality ($median_C = 0.12$, $IQR_C = 0.14$; $median_V = 0.05$, $IQR_V = 0.03$; $median_P = 0.05$, $IQR_P = 0.02$) (Fig. 3b), indicating that a few clones were expanded and possibly reflecting that these individuals have had a recent adaptive immunity response, most likely to SARS-CoV-2 infection.

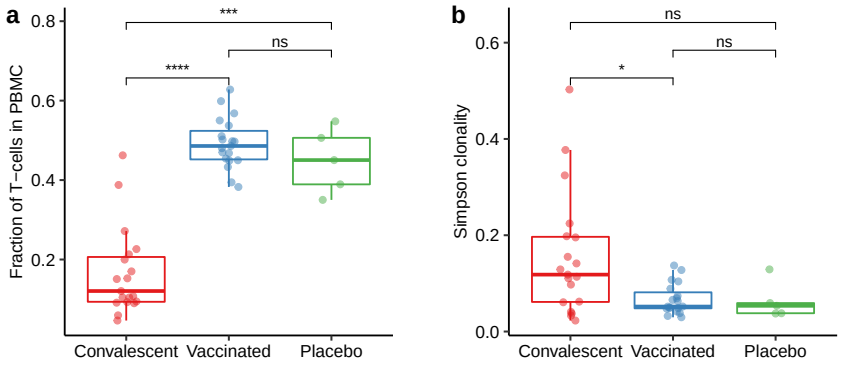


Figure 3: TCR repertoire preliminary analysis. **(a)** Fraction of T cells among peripheral blood mononuclear cells (PBMC). **(b)** Simpson clonality index. A value closer to 1 indicates the emergence of a few dominant clones, whereas it reaches its minimum when TCR frequencies are evenly distributed. Statistical significance was determined by two-sided Wilcoxon rank-sum tests. N = 43 independent samples (19 SARS-CoV-2 convalescent individuals, 19 Ad26.COV2.S vaccine recipients, 5 placebo recipients).

Breadth and depth of SARS-CoV-2-specific T cell response

To evaluate the magnitude of the T cell response to SARS-CoV-2 after disease and vaccination, TCR repertoires of convalescent, vaccinated and placebo recipients individuals were annotated with $CD8^+$ and $CD4^+$ TCR datasets that had previously been determined to be SARS-CoV-2-specific and screened for enrichment compared to a background of healthy individuals repertoires in order to remove TCRs that may be unspecific (i.e. cross-reactive to common antigens) (See Material and Methods). Among the 43 repertoires analyzed there were 11,604,850 distinct TCR sequences (V gene + CDR3 aminoacid sequence) of which only 284 were SARS-CoV-2-specific. Out of these annotated 284 TCRs, 51 were public (i.e. present in more than one individual).

Responses to SARS-CoV-2 spike and non-spike were measured in terms of breadth (unique TCR sequences) and depth (frequency of those TCRs). Both convalescent and vaccinated subjects had a higher and undistinguishable spike-specific response compared to placebos in terms of breadth and depth (breadth: $median_C = 6.64 \cdot 10^{-6}$, $IQR_C = 1.4 \cdot 10^{-5}$; $median_V = 5.45 \cdot 10^{-6}$, $IQR_V = 1.1 \cdot 10^{-5}$; $median_P = 0$, $IQR_P = 0$; depth: $median_C = 4.21 \cdot 10^{-6}$, $IQR_C = 1.44 \cdot 10^{-5}$; $median_V = 7.32 \cdot 10^{-5}$, $IQR_V = 5.47 \cdot 10^{-5}$; $median_P = 0$, $IQR_P = 0$) (Fig. 4a). By contrast, breadth and depth of non-spike TCRs were significantly higher in convalescent individuals versus vaccinated and placebos, and there were no significant differences between the latter two (breadth: $median_C = 3.04 \cdot 10^{-5}$, $IQR_C = 3.96 \cdot 10^{-5}$; $median_V = 8.09 \cdot 10^{-6}$, $IQR_V = 6.27 \cdot 10^{-6}$; $median_P = 9.4 \cdot 10^{-6}$, $IQR_P = 1.24 \cdot 10^{-5}$; depth: $median_C = 2.76 \cdot 10^{-5}$, $IQR_C = 1.15 \cdot 10^{-4}$; $median_V = 7.32 \cdot 10^{-6}$, $IQR_V = 6.82 \cdot 10^{-6}$; $median_P = 1 \cdot 10^{-5}$, $IQR_P = 1.31 \cdot 10^{-5}$) (Fig. 4b), as expected because Ad26.COV2.S vaccine only carry the spike antigen.

An in-depth analysis of the genomic localization of SARS-CoV-2 epitopes recognized by TCRs revealed that most of the T cell immune response in convalescent individuals was directed towards structural proteins S and N (Fig. 5a). Although vaccinated subjects showed some response to non-spike proteins, for most epitopes this signal was less broad and deep compared to the convalescent group and most likely due to unspecific annotations or cross-reactivity, since placebo recipients also showed a minimal level of response. The localization of the 5 spike protein epitopes is shown in Fig. 5b. As opposed to B-

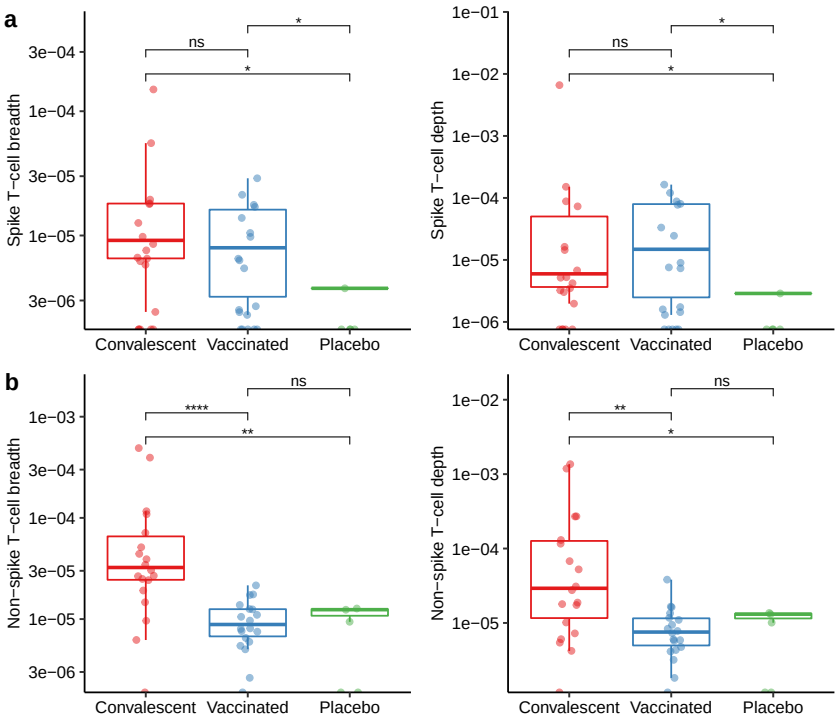
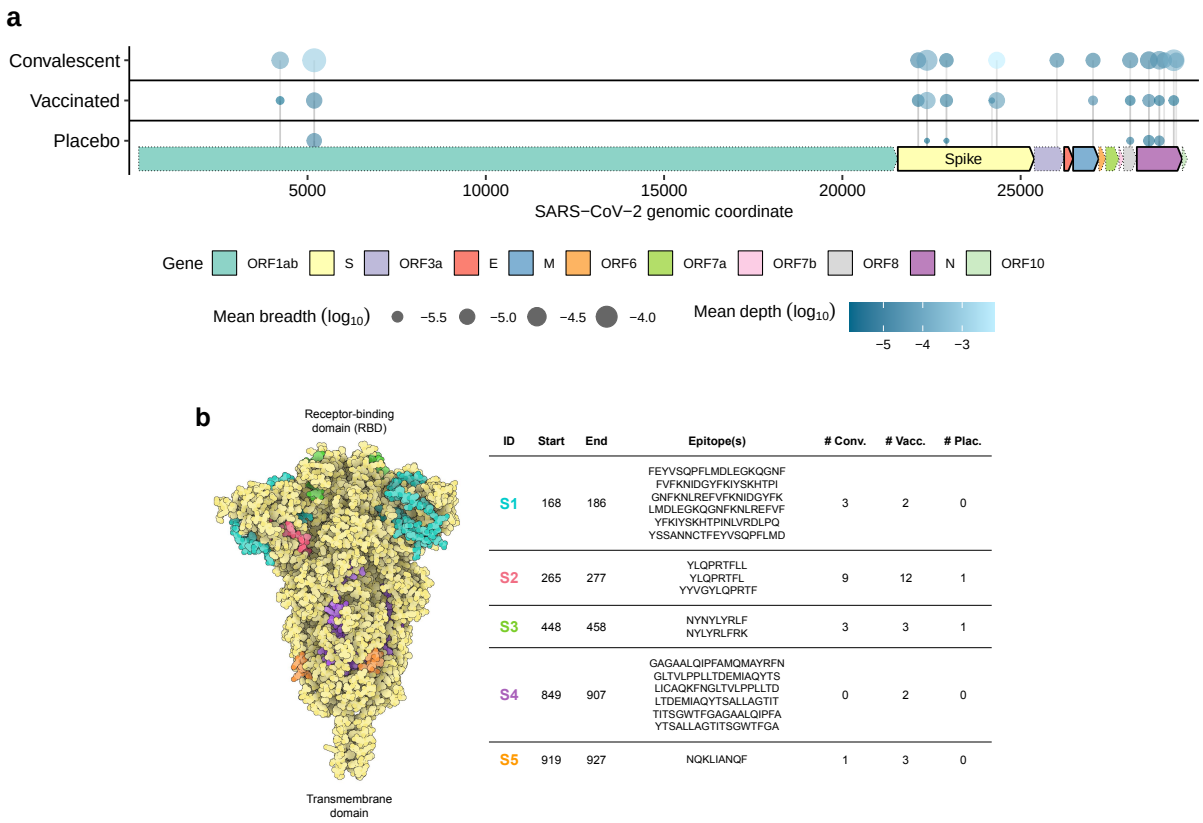


Figure 4: SARS-CoV-2-specific TCRβ repertoire analysis. (a) Spike-specific T cell breadth and depth. **(b)** Non-spike-specific T cell breadth and depth. Breadth was calculated as the fraction of unique TCR sequences specific to spike / non-spike proteins; depth is the relative frequency of those specific TCRs in the repertoire. Statistical significance was determined by two-sided Wilcoxon rank-sum tests on the original values (logarithmic transformation is only applied for visualization purposes). N = 43 independent samples (19 SARS-CoV-2 convalescent individuals, 19 Ad26.COV2.S vaccine recipients, 5 placebo recipients)

cell epitopes, T cell epitopes localization was not restricted to the protein surface since TCRs recognize antigens processed and presented in MHC molecules by human cells, and it was shown in the spike protein 3D representation, where most of the epitopes were partially (S1, S3, S5) or completely (S2, S4) buried in the structure. S2 epitope was in fact the most widely recognized by convalescent (9/19) and vaccinated (12/19) individuals, and S4, the most hidden in the protein, was unrecognized by convalescent subjects and exclusively recognized by TCRs of 2 out of 19 vaccine recipients.

TCR similarity network analysis

The landscape of TCR repertoires is vast and complex. A simple antigen specificity annotation by V gene and CDR3 aminoacid sequence exact match, although informative and straightforward, can underestimate the magnitude of the T cell response. In order to capture the SARS-CoV-2-specific TCR repertoire architecture, graphs representing networks of similar TCRs were generated for the 43 TCR repertoires. Fig. 6a shows two SARS-CoV-2-specific TCR sequence similarity networks of one convalescent and one vaccinated individual. A SARS-CoV-2-specific TCR similarity network is defined as all the components (subgraphs in which any two nodes are connected to each other by paths) that include at least one SARS-CoV-2-specific node (yellow, purple). Spike-specific nodes are present in both graphs, while non-spike-specific nodes are restricted to the convalescent individual, as expected. In both networks, some SARS-CoV-2-specific nodes had a high relative frequency, represented by node size in the graph. As for the networks architecture, the most striking difference was the network



ID	Start	End	Epitope(s)	# Conv.	# Vacc.	# Plac.
S1	168	186	FEYVSQPFLMDLEGKQGNF FVFKNIDGYFKIYSKHTPI GNFKNLREFVFKNIDGYFK LMDLEGKQGNFKNLREFVF YFKIYSKHTPINLVRDLPQ YSSANNCTFEYVSQPFLMD	3	2	0
S2	265	277	YLQPRTFLL YLQPRTFL YYVGYLQPRTF	9	12	1
S3	448	458	NYNYLYRLF NYLYRLFRK	3	3	1
S4	849	907	GAGAALQIPFAMQMAYRFN GLTVLPPLLTDEMIAQYTS LICAQKFNGLTVLPPLLTD LTDEMIAQYTSALLAGTIT TITSGWTFGAGAALQIPFA YTSALLAGTITSGWTFGA	0	2	0
S5	919	927	NQKLIANQF	1	3	0

Figure 5: Epitopes recognized by T cells across SARS-CoV-2 genome. (a) Lollipop plot of the TCRβ SARS-CoV-2 specificity in convalescent, vaccinated and placebo individuals across the coronavirus genome. Size of the dots indicate the mean breadth across all samples in a group and color scale indicates the mean depth of the response. Genes outlined with solid lines are structural (S, E, M and N), whereas dotted ones encode non-structural proteins. **(b)** Localization and characteristics of the 5 SARS-CoV-2 spike epitopes recognized by TCRs in this study. Epitopes appear colored in the three monomers of a spike protein 3D representation (PDB: 6XR8, side view). Start, End: epitope protein coordinates (1-based); # Conv., # Vacc., # Plac.: number of individuals in a group with TCRs specific to that epitope.

size, with the convalescent graph having less edges and nodes that the vaccinated one. This was also observed at the group level, and furthermore, convalescent SARS-CoV-2-specific TCRs networks had a lower average degree (number of edges per node) compared to vaccine recipient individuals (Fig. 6b).

In this work, it has been noted that direct annotation with SARS-CoV-2-specific TCRs is not very precise, since there was some non-spike response in vaccine recipients, as well as response to some SARS-CoV-2 epitopes in healthy placebo recipients (Figs. 4b, 5a). Network analysis can help identifying those nodes that truly are SARS-CoV-2-specific versus those that may be cross-reactive. It was expected for spike-specific TCRs to have more importance in convalescent and vaccinated subjects networks compared to placebo, and for non-spike-specific TCRs to be relevant only in convalescent individuals. Important nodes in network science are known as hubs, or nodes with a number of edges exceeding the average, and the network local property that measures this quality is the hub score, which takes into account how well connected is a node itself, and those nodes that it is connected to. In convalescent and placebo subjects, there were no significant differences in hub scores distributions of spike-specific and non-spike specific nodes, beign the hub scores very low in the placebo recipient SARS-CoV-2-specific

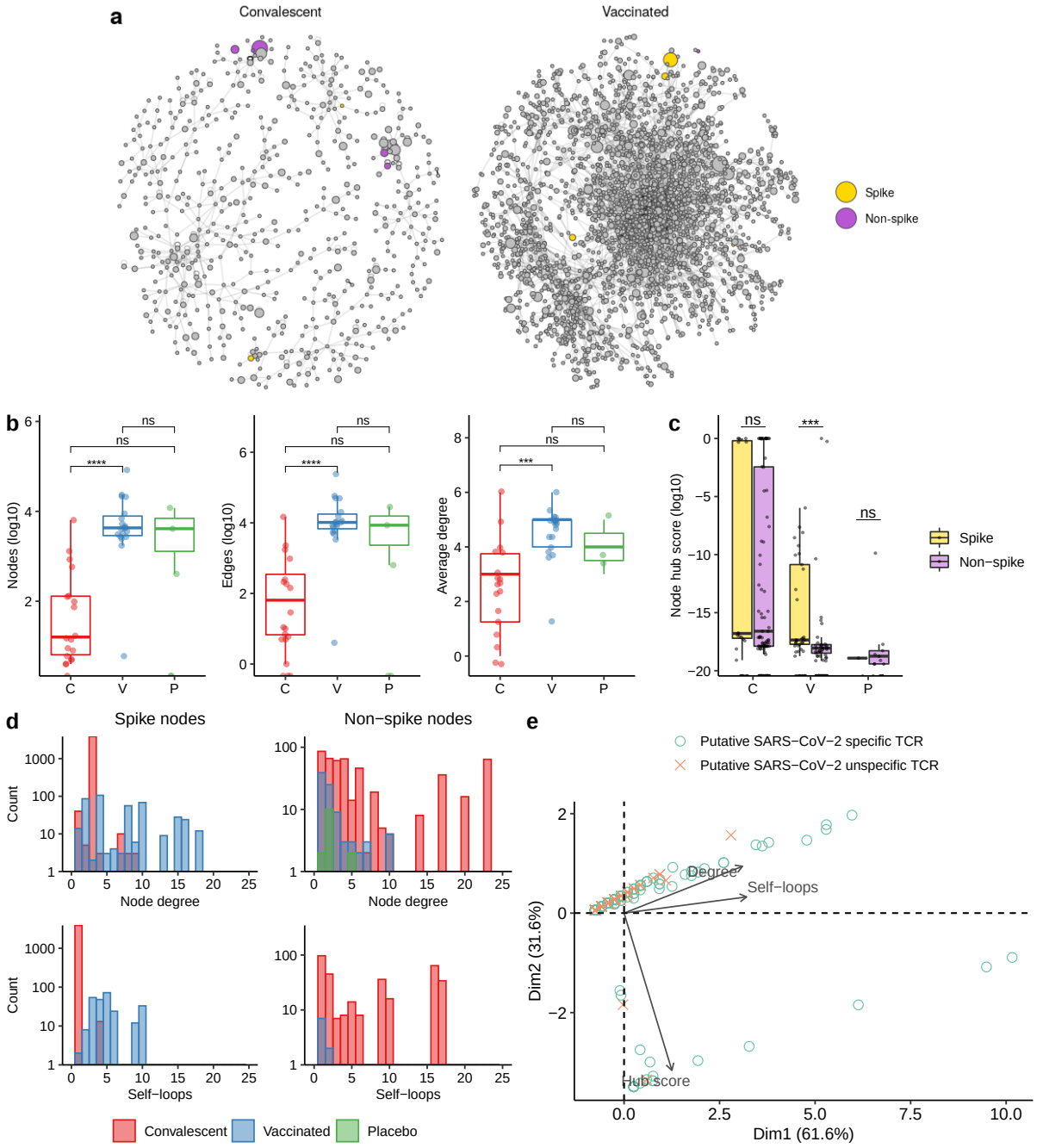


Figure 6: SARS-CoV-2-specific TCRβ similarity networks. (a) TCRβ similarity networks as undirected graphs, where each node represents a unique V gene + CDR3 aminoacid sequence combination. Only SARS-CoV-2-specific nodes (colored) and the nodes in their components (gray) are shown for convalescent subject 18 and vaccinated subject 12. Two nodes are connected by and edge if their tcrdist3 distance is ≤ 12. Loops (self-edges that start and end at the same node) represent additional unique nucleotide sequences encoding the same V gene + amino acid sequence (i.e., convergence). Node size represents the TCR frequency. **(b)** Distribution of network global properties (number of nodes, edges and average degree) by group. **(c)** Spike and non-spike nodes hub score distributions by group. **(d)** Histograms of SARS-CoV-2-specific nodes degree and self-loops. Y axis represent node counts taking their frequency into account. **(e)** Putative SARS-CoV-2-specific and unnspecific biplot based on network local metrics. A node was putatively considered SARS-CoV-2-specific if it was a spike node in a convalescent or vaccinated network, or if it was a non-spike node in a convalescent network. Statistical significance was determined by two-sided Wilcoxon rank-sum tests on the original values (logarithmic transformation is only applied for visualization purposes). N = 43 independent samples (19 SARS-CoV-2 convalescent individuals, 19 Ad26.COV2.S vaccine recipients, 5 placebo recipients). C: convalescent, V: vaccinated, P: placebo.

nodes. However, hub scores were significantly greater for spike-specific nodes in vaccinated individuals compared to non-spike-specific nodes ($p = 2.37 \cdot 10\text{-}4$, two-sided Wilcoxon rank-sum test; spike nodes median = $1.87 \cdot 10^{-18}$, IQR = $6.53 \cdot 10^{-15}$; non-spike nodes median = 0, IQR = $6.34 \cdot 10^{-19}$) (Fig. 6c). This suggests that TCRs annotated as non-spike-specific in vaccine recipient subjects or any SARS-CoV-2-specific TCR in the placebo group do not have an important role in their TCR repertoires and may be SARS-CoV-2-unspecific and cross-reactive.

In addition, the spike-specific nodes with highest degrees (more connections) were from vaccinated subjects, whereas the highly connected non-spike-specific nodes belonged to convalescent individuals (Fig. 6d, top panels). The same trend was observed for self-loops, which represent TCR convergence, i.e. additional nucleotide TCR sequences encoding the same V gene + amino acid sequence (Fig. 6d, bottom panels). A principal component analysis with the aforementioned network local properties facilitate the differentiation between putative SARS-CoV-2-specific TCRs (spike-specific detected in convalescent and vaccinated individuals and non-spike-specific detected in convalescent subjects) and putative SARS-CoV-2-unspecific TCRs (those detected in placebo repertoires and non-spike-specific in vaccine recipients) (Fig 6e). Nevertheless, due to the constantly changing nature of the TCR repertoires, after a time period following antigen exposure, some putative SARS-CoV-2-specific TCRs can be undistinguishable from putative unspecific ones in terms of hub score, degree and self-loops, since they become less important in the adaptive immunity landscape once the infection is cleared.

Discussion

The current study presents an in-depth characterization of TCR repertoires after COVID-19 disease and vaccination. Analyzing these repertoires from the frequency space revealed low proportions of T cells and less TCR diversity (more clonality) in convalescent individuals compared to healthy vaccine and placebo recipients, which was consistent with the clinical characteristics of the COVID-19 disease: lymphopenia and antigen-specific clonal expansions that reduce the repertoire size and diversity.

Adding 11 convalescent individual samples to match the sample size of the vaccine recipients cohort increased the statistical power of the comparisons between both groups, and both convalescent cohorts TCR repertoires beign amplified by the same assay minimized potential batch effects. This strategy could not be applied to placebo, since there are no further published data, but an alternative would be to add a cohort of unexposed healthy controls, on which there are many TCR repertoires published in immuneACCESS database.

Response to SARS-CoV-2 was measured in terms of breadth and depth, where immunization to the spike protein in convalescent and vaccinated cohorts and a high response to non-spike antigens exclusive to convalescent subjects were observed, acknowledging that infection provides immunization against several virus proteins, while vaccines only elicit immunization against the single antigen they carry (the spike protein). The results were consistent with those shown in the data source publication (Alter et al., 2021), even though a different dataset of SARS-CoV-2-specific TCRs was used to annotate the bulk TCR repertoires. In Alter et al. (2021), authors claim to have used a TCR dataset determined to be SARS-CoV-2-specific and enriched in subjects with natural infection relative to placebos curated by Snyder

et al. (2020) from public MIRA datasets (Nolan et al., 2020). However, this study is still a preprint and the data have not yet been made publicly available.

The workaround was to use an enriched SARS-CoV-2-specific TCR dataset generated by Mayer-Blackwell et al. (2020) from the same MIRA data in the immuneCODE database, but with a different method than Snyder et al. (2020). Because $CD8^+$ MIRA dataset was made publicly available before the $CD4^+$ dataset, the authors have only curated and published the enriched dataset of SARS-CoV-2-specific $CD8^+$ TCRs to this date. However, as both the input data (MIRA unenriched $CD4^+$ TCRs) and the code were public when the present work was being made, the enrichment analysis could be performed to extract 64 new SARS-CoV-2-specific enriched TCRs.

In-depth analysis of the location of the epitopes recognized by TCRs showed that, in COVID-19 convalescent individuals, epitopes of S and N proteins are immunodominant amongst coronavirus antigens, as in the case of antibodies immune response (Amrun et al., 2020). It was also observed that in some cases, in natural infection non-spike responses are stronger, particularly to epitopes of N protein, as it has been previously reported (Grifoni et al., 2020; Snyder et al., 2020). The response to multiple epitopes of the same antigen may provide long-lasting protection to SARS-CoV-2, specially with the emergence of new variants. Response to spike protein of convalescent and vaccine recipient individuals was undistinguishable in terms of breadth and depth, but also 4 out of 5 spike epitopes were recognized by TCRs in both cohorts, remarking the similarity of infection and vaccine immunization. The only epitope exclusive to vaccine recipients (S4) was the most buried in the protein 3D structure, raising the question of whether antigenic processing of the whole virus versus the isolated S protein can generate different epitopes.

Studying TCR repertoires only from the frequency space (in terms of clonal expansions and diveristy) or from the sequence space simply looking for the presence or absence of certain sequences ignores the huge inter-individual TCR variability and may reveal only a tiny fraction of the true underlying signal of an immune response in the repertoire. To get a fuller and unexplored picture of TCR response in COVID-19 disease and vaccination, TCR repertoires were studied from a sequence similarity perspective. Given the vast diversity of sequences in a TCR repertoire, pairwise distances calculation can become a computationally expensive task. Currently, there are several existing tools for TCR sequence similarity analysis (Huang et al., 2020; Pogorelyy et al., 2019; Mayer-Blackwell et al., 2020). tcrdist3, although it is one of the most time-consuming algorithms, was chosen among all the alternatives for having a biochemically informed distance metric that accounts for sequence similarities from an aminoacid properties perspective.

This allowed to build networks of TCR sequence similarity that revealed that spike-specific TCRs in convalescent and vaccinated cohorts acted as network hubs, beign connected to several similar sequences and often to itself, as the TCR generation process can converge and generate exact same CDR3 aminoacid sequences with different nucleotide sequences. Non-spike-specific TCRs were present in the three cohorts of this study, but the network analysis perspective showed that they had low hub scores in vaccinated and placebo repertoires, indicating that, as it would be expected, neither of them had experienced an immune response against SARS-CoV-2 non-spike antigens, but may have

cross-reactive TCRs. The presence of putatively cross-reactive TCRs in the SARS-CoV-2-specific dataset used for annotation denotes that the enrichment of MIRA datasets can be further fine-tuned. It has been found that the origin of these cross-reactive TCRs may be previous exposure to other common cold coronaviruses (Minervina et al., 2021).

Although the present work provides a deeper understanding of SARS-CoV-2 immunization in disease and vaccination, one of its limitations relies on the repertoires antigen-specificity annotation process. Responses to SARS-CoV-2 epitopes could be studied with a sequence similarity network approach if and only if at least one SARS-CoV-2-specific TCR was found in the repertoire via a V gene and CDR3 aminoacid sequence exact match. This ignores the presence in the repertoire of TCRs highly similar (but not exact matches) to those in the SARS-CoV-2-specific dataset. One way to improve this work would consist of screening the repertoires with a fuzzy match approach and establishing a tcrdist3 distance threshold, but note that this would also increase the chances of finding unspecific TCRs, that can be subsequently evaluated according to its node metrics in the network analysis.

Another limitation to further extend this analysis is the lack of vaccine recipients TCR repertoire data to track the immune response at the population level. Vaccine developers are gradually publishing repertoire data, but currently the cohorts are very small. As these datasets become available, characteristics of TCR responses to vaccines based on different technologies (viral vector, mRNA) could be compared. Also, the identification of novel SARS-CoV-2-specific TCRs could benefit from more TCR repertories of convalescent individuals bearing different MHC alleles combination, since the MHC alleles determine which epitopes are preferentially presented and thus which TCRs are going to be expanded in the repertoire.

All in all, the great effort of the scientific community in SARS-CoV-2-specific immune repertoire data generation has quickly made possible to track certain TCRs as indicators of past and ongoing SARS-CoV-2 immunization, which may be detected in blood at least two months past exposure (Gittelman et al., 2021). Thanks to the boosted research in COVID-19 immunology, we are closer than ever to defining TCRs as biomarkers, and these advances set a precedent of best practices for studying TCR repertoires in other infections and diseases.

Conclusions

The curation of published SARS-CoV-2-specific TCRs datasets allowed the measurement of the response to SARS-CoV-2 antigens in disease and vaccination, showing that COVID-19 convalescent subjects are immunized against several viral antigens, while vaccine immunization is restricted to the spike protein. Sequence similarity network analysis has the potential to discern between true TCR responses and unspecific cross-reactivities, but its relevance should be further studied. The titanic effort of the scientific community towards the understanding of SARS-CoV-2 immunology will continue to provide useful immune repertoire data to establish accurate SARS-CoV-2 immunization TCR

biomarkers.

Bibliography

Alter, Galit, Yu, Jingyou, Liu, Jinyan, Chandrashekar, Abishek, Borducchi, Erica N, Tostanoski, Lisa H, McMahan, Katherine, Jacob-Dolan, Catherine, Martinez, David R, Chang, Aiquan, et al. Immunogenicity of Ad26.COV2.S vaccine against SARS-CoV-2 variants in humans. *Nature*, pages 1–5, 2021.

Amrun, Siti Naqiah, Lee, Cheryl Yi-Pin, Lee, Bernett, Fong, Siew-Wai, Young, Barnaby Edward, Chee, Rhonda Sin-Ling, Yeo, Nicholas Kim-Wah, Torres-Ruesta, Anthony, Carissimo, Guillaume, Poh, Chek Meng, et al. Linear B-cell epitopes in the spike and nucleocapsid proteins as markers of SARS-CoV-2 exposure and disease severity. *EBioMedicine*, 58:102911, 2020.

Csardi, Gabor and Nepusz, Tamas. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL https://igraph.org.

Dash, Pradyot, Fiore-Gartland, Andrew J, Hertz, Tomer, Wang, George C, Sharma, Shalini, Souquette, Aisha, Crawford, Jeremy Chase, Clemens, E Bridie, Nguyen, Thi HO, Kedzierska, Katherine, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, 2017.

Ewer, Katie J, Barrett, Jordan R, Belij-Rammerstorfer, Sandra, Sharpe, Hannah, Makinson, Rebecca, Morter, Richard, Flaxman, Amy, Wright, Daniel, Bellamy, Duncan, Bittaye, Mustapha, et al. T cell and antibody responses induced by a single dose of ChAdOx1 nCoV-19 (AZD1222) vaccine in a phase 1/2 clinical trial. *Nature medicine*, 27(2):270–278, 2021.

Garcia, K Christopher and Adams, Erin J. How the T cell receptor sees antigen—a structural view. *Cell*, 122(3):333–336, 2005.

Gittelman, Rachel M, Lavezzo, Enrico, Snyder, Thomas M, Zahid, H Jabran, Elyanow, Rebecca, Dalai, Sudeb, Kirsch, Ilan, Baldo, Lance, Manuto, Laura, Franchin, Elisa, et al. Diagnosis and tracking of SARS-CoV-2 infection by T-Cell receptor sequencing. *medRxiv*, pages 2020–11, 2021.

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

Grifoni, Alba, Weiskopf, Daniela, Ramirez, Sydney I, Mateus, Jose, Dan, Jennifer M, Moderbacher, Carolyn Rydyznski, Rawlings, Stephen A, Sutherland, Aaron, Premkumar, Lakshmanane, Jadi, Ramesh S, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell*, 181(7):1489–1501, 2020.

Harris, Tim and Sauer, Karsten. Are T cell repertoires useful as diagnostics for SARS-CoV-2 infection? *Expert Review of Molecular Diagnostics*, 21(2):137–139, 2021.

Heather, James M, Ismail, Mazlina, Oakes, Theres, and Chain, Benny. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in bioinformatics*, 19(4):554–565, 2018.

Hu, Ben, Guo, Hua, Zhou, Peng, and Shi, Zheng-Li. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, 19(3):141–154, 2021.

Huang, Huang, Wang, Chunlin, Rubelt, Florian, Scriba, Thomas J, and Davis, Mark M. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nature biotechnology*, 38(10):1194–1202, 2020.

Kassambara, Alboukadel. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL https://CRAN.R-project.org/package=ggpubr. R package version 0.4.0.

Kassambara, Alboukadel. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2021. URL https://CRAN.R-project.org/package=rstatix. R package version 0.7.0.

Kassambara, Alboukadel and Mundt, Fabian. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017. URL https://CRAN.R-project.org/package=factoextra. R package version 1.0.5.

Klinger, Mark, Pepin, Francois, Wilkins, Jen, Asbury, Thomas, Wittkop, Tobias, Zheng, Jianbiao, Moorhead, Martin, and Faham, Malek. Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One*, 10(10):e0141561, 2015.

Liu, Jing, Li, Sumeng, Liu, Jia, Liang, Boyun, Wang, Xiaobei, Wang, Hua, Li, Wei, Tong, Qiaoxia, Yi, Jianhua, Zhao, Lei, et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine*, 55:102763, 2020.

Lubin, Joseph H, Zardecki, Christine, Dolan, Elliott M, Lu, Changpeng, Shen, Zhuofan, Dutta, Shuchismita, Westbrook, John D, Hudson, Brian P, Goodsell, David S, Williams, Jonathan K, et al. Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first six months of the COVID-19 pandemic. *bioRxiv*, 2020.

Lythe, Grant, Callard, Robin E, Hoare, Rollo L, and Molina-París, Carmen. How many TCR clonotypes does a body maintain? *Journal of theoretical biology*, 389:214–224, 2016.

Mayer-Blackwell, Koshlan, Schattgen, Stefan, Cohen-Lavi, Liel, Crawford, Jeremy Chase, Souquette, Aisha, Gaevert, Jessica A, Hertz, Tomer, Thomas, Paul G, Bradley, Philip, and Fiore-Gartland, Andrew. TCR meta-clonotypes for biomarker discovery with tcrdist3: quantification of public, HLA-restricted TCR biomarkers of SARS-CoV-2 infection. *bioRxiv*, 2020.

Minervina, Anastasia A, Pogorelyy, Mikhail V, Kirk, Allison M, Allen, Emma Kaitlynn, Allison, Kim J, Lin, Chung-Yang, Brice, David C, Zhu, Xun, Vegesana, Kasi, Wu, Gang, et al. Convergent epitope-specific T cell responses after SARS-CoV-2 infection and vaccination. *medRxiv*, 2021.

Murphy, K.M. and Weaver, C. *Janeway's Immunobiology*. Garland Science, 2016. ISBN 9780815345503.

Nolan, Sean, Vignali, Marissa, Klinger, Mark, Dines, Jennifer N, Kaplan, Ian M, Svejnoha, Emily, Craft, Tracy, Boland, Katie, Pesesky, Mitch, Gittelman, Rachel M, 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. *Research Square*, 2020.

Pedersen, Thomas Lin. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*, 2020. URL https://CRAN.R-project.org/package=ggraph. R package version 2.0.2.

Pogorelyy, Mikhail V, Minervina, Anastasia A, Shugay, Mikhail, Chudakov, Dmitriy M, Lebedev, Yuri B, Mora, Thierry, and Walczak, Aleksandra M. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biology*, 17(6):e3000314, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

Sahin, Ugur, Muik, Alexander, Derhovanessian, Evelyna, Vogler, Isabel, Kranz, Lena M, Vormehr, Mathias, Baum, Alina, Pascal, Kristen, Quandt, Jasmin, Maurus, Daniel, et al. COVID-19 vaccine BNT162b1 elicits human antibody and TH 1 T cell responses. *Nature*, 586(7830):594–599, 2020.

Sauer, Karsten and Harris, Tim. An effective COVID-19 vaccine needs to engage T cells. *Frontiers in Immunology*, 11, 2020.

Snyder, Thomas M, Gittelman, Rachel M, Klinger, Mark, May, Damon H, Osborne, Edward J, Taniguchi, Ruth, Zahid, H Jabran, Kaplan, Ian M, Dines, Jennifer N, Noakes, Matthew N, et al. Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. *MedRxiv*, 2020.

Stephenson, Kathryn E, Le Gars, Mathieu, Sadoff, Jerald, De Groot, Anne Marit, Heerwegh, Dirk, Truyers, Carla, Atyeo, Caroline, Loos, Carolin, Chandrashekar, Abishek, McMahan, Katherine, et al. Immunogenicity of the ad26. cov2. s vaccine for covid-19. *Jama*, 325(15):1535–1544, 2021.

Swanson, Phillip A, Padilla, Marcelino, Hoyland, Wesley, McGlinchey, Kelly, Fields, Paul A, Bibi, Sagida, Faust, Saul N, McDermott, Adrian B, Lambe, Teresa, Pollard, Andrew J, et al. T-cell mediated immunity after AZD1222 vaccination: A polyfunctional spike-specific Th1 response with a diverse TCR repertoire. *medRxiv*, 2021.

Tomasello, Gianluca, Armenia, Ilaria, and Molla, Gianluca. The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics*, 36(9):2909–2911, 2020.

WHO. WHO coronavirus (COVID-19) dashboard. https://covid19.who.int/, 2021. Accessed: 2021-08-31.

Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wickham, Hadley. *tidyr: Tidy Messy Data*, 2020. URL https://CRAN.R-project.org/package=tidyr. R package version 1.1.2.

Wickham, Hadley, François, Romain, Henry, Lionel, and Müller, Kirill. *dplyr: A Grammar of Data Manipulation*, 2020. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.2.

Wilkins, David. *gggenes: Draw Gene Arrow Maps in 'ggplot2'*, 2020. URL https://CRAN.R-project.org/package=gggenes. R package version 0.4.1.