# 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 is 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 cell 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.

*TO-DO: (MHC I CD8, MHC II CD4, tcr specificity,* $10^15$*, tcr generation) (Covid: proteome, epitopes) (Analysis of TCR repertoires: focus on network analysis and antigen specificity annotation, MIRA assays) (vaccines, first tcr studies) (Objectives)*

# 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 received the Ad26.COV2.S vaccine developed by Janssen Pharmaceutica during a clinical trial, and 5 subjects who received a placebo. Peripheral blood samples were collected post diagnosis or vaccination and immunosequencing of the CDR3 regions of human TCR$\beta$ 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 with 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 TCR$\beta$ 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 $CD8^+$ TCR$\beta$ sequences were obtained from Mayer-Blackwell et al. (2020). This sequences are proven to bind SARS-CoV-2 epitopes by Multiplex Identification of Receptor Antigen (MIRA) assays (Nolan et al., 2020) and are also enriched in bulk TCR$\beta$ repertoires of convalescent

individuals compared to healthy controls. For the present study, only TCRβ sequences with a strong evidence of HLA restriction (N = 1831) were taken into consideration.

### 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 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).

### 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 matching 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 proportion of the frequency of those SARS-CoV-2-specific TCRs.

### CDR3 pairwise distances

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 CPUs, 256 GB of RAM).

### Network analysis

In this analysis each unique CDR3 aminoacid sequence were considered 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. 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) and parallel v3.6.1 (R Core Team, 2019) were used. Networks were plotted with the ggraph v2.0.2 package (Pedersen, 2020). All other plots were generated with ggplot2 v3.3.2 (Wickham, 2016) and ggpubr v0.4.0 (Kassambara, 2020). 3D visualization of the SARS-CoV-2 spike protein (PDB ID: 6XR8) was generated with Protein Imager (Tomasello et al., 2020).

## Results

### TCR repertoire preliminary 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 is significantly lower in convalescent individuals compared to vaccinated and placebo ($p = 2.5 \cdot 10^{-9}$ and $p = 5.6 \cdot 10^{-4}$, two-sided Wilcoxon rank-sum test) and no significant differences were observed between vaccinated and placebo subjects (Fig. 1a).

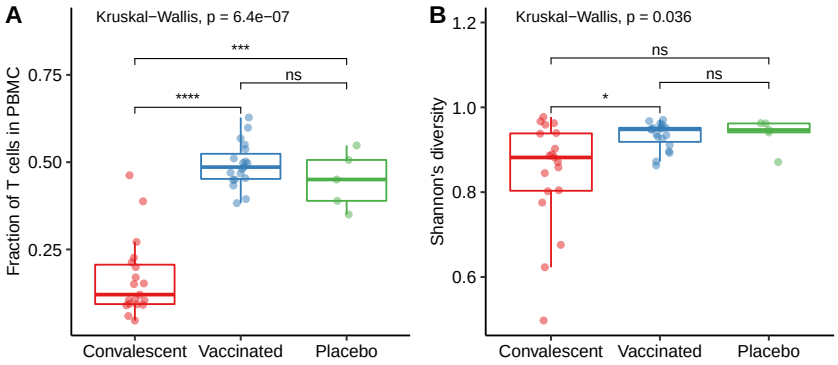



**Figure 1: TCR repertoire preliminary analysis. (a)** Fraction of T cells among peripheral blood mononuclear cells (PBMC). **(b)** Shannon diversity index. A value closer to 0 indicates the emergence of a few dominant clones, whereas it reaches its maximum when TCR frequencies are evenly distributed. 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).

Some convalescent patients TCRβ repertoires have a low Shannon diversity index (Fig. 1b), indicating that a few clones are expanded and possibly reflecting that these individuals had a recent adaptive immunity response, most likely to SARS-CoV-2 infection.

**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 unique 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).

Response to SARS-CoV-2 spike and non-spike was measured in terms of breadth (unique TCR sequences) and depth (frequency of those TCRs). Both convalescent and vaccinated subjects have a higher spike-specific response compared to placebos in termns of breadth and depth (Fig. 2a). 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 (Fig. 2b), 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 is directed towards structural proteins S and N (Fig. 3a). Although vaccinated subjects show some response to non-spike proteins, for most epitopes this signal is residual compared to the convalescent group and most likely due to unspecific annotations, since placebo recipients also show a minimal level of response. The localization of the 5 spike protein epitopes is shown in Fig. 3b. As opposed to B-cell epitopes, T-cell epitopes localization is not restricted to the protein surface since TCRs recognize antigens processed and presented in MHC molecules by human cells, and it shows in the spike protein 3D representation, where most of the epitopes are partially (S1, S3, S5) or completely (S2, S4) buried in the structue. S2 epitope is in fact the most widely recognized by convalescent (9/19) and vaccinated (12/19) individuals, and S4, the most hidden in the protein, is unrecognized by convalescent subjects and exclusively recognized by TCRs of 2 out of 19 vaccine recipients.

**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 (Fig.4a).

In addition, direct annotation with SARS-CoV-2-specific TCRs is not very precise, since there is some non-spike response in vaccine recipients, as well as response to some SARS-CoV-2 epitopes in healthy placebo recipients (Figs. 2b, 3a). Network analysis can help identifying those nodes that truly are SARS-CoV-2-specific.

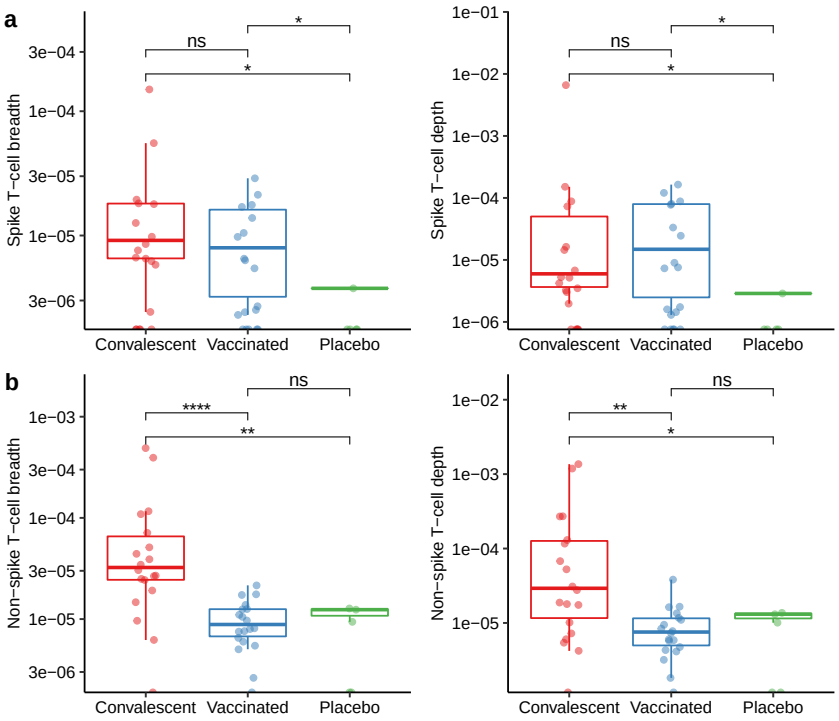



**Figure 2: SARS-CoV-2-specific TCRβ repertoire analysis. (a)** Spike-specific T-cell breadth and depth. **(b)** Non-spike-specific T-cell breadth and depth. Breadth is calculated as the fraction of unique TCR sequences specific to spike / non-spike proteins; depth is the relative frequency of those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

*TO-DO - Main points: - Vaccinated networks are richer than convalescent (more nodes and edges), reflecting lymphopenia - Non-spike nodes have more authority in convalescent networks - Non-spike nodes have more degree, loops and smaller distances in convalescent networks - PCA with the mentioned variables effectively separates true covid-specific nodes*

*Figures in the making: Fig 4b: covid similarity networks global metrics, Fig 4c: authority of spike / non-spike nodes, Fig 4d: degree, loops and distance histograms of spike / non-spike nodes, Fig 4e: PCA with variables in 4c,d that separate true covid-specific nodes*

# Bibliography


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

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

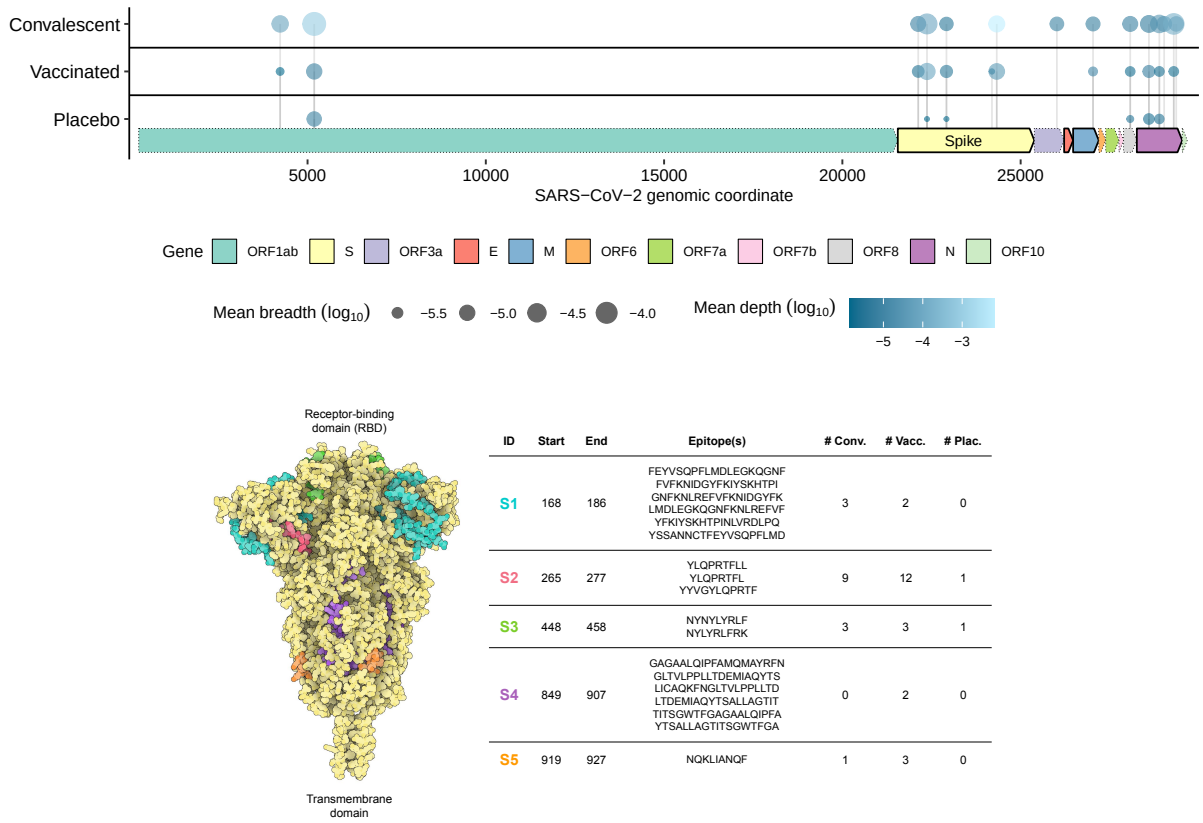
Pradyot Dash, Andrew J Fiore-Gartland, Tomer Hertz, George C Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi HO Nguyen, Katherine Kedzierska, et al. Quantifiable predictive features define epitope-specific t cell receptor repertoires. *Nature*, 547(7661):89–93, 2017.

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

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

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

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




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

**Figure 3: Epitopes recognized by T cells accross SARS-CoV-2 genome. (a)** Lollipop plot of the TCRβ SARS-CoV-2 specificity in convalescent, vaccinated and placebo individuals accross the coronavirus genome. Size of the dots indicate the mean breadth accross 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.


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

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

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

Sean Nolan, Marissa Vignali, Mark Klinger, Jennifer N Dines, Ian M Kaplan, Emily Svejnoha, Tracy Craft, Katie Boland, Mitch Pesesky, Rachel M Gittelman, et al. A large-scale database of T-cell receptor beta (TCR$\beta$) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Research square*, 2020.

Thomas Lin Pedersen. *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.

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/`.

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

Gianluca Tomasello, Ilaria Armenia, and Gianluca Molla. 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.

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

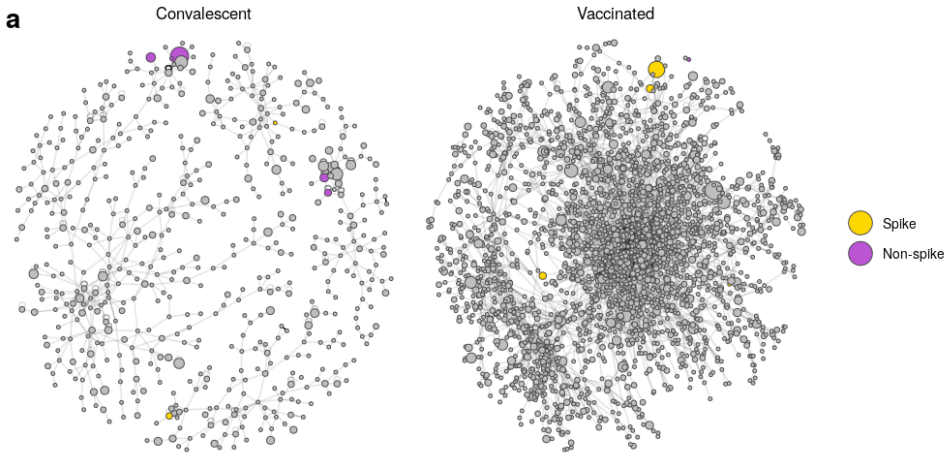



**Figure 4: SARS-CoV-2-specific TCRβ similarity networks. (a)** TCRβ similarity networks where each node represents a unique V gene + CDR3 aminoacid sequence combination. Only SARS-CoV-2-specific nodes (colored) and the nodes in their components (gray) are shown for convalescent subject 18 and vaccinated subject 12. Two nodes are connected by and edge if their tcrdist3 distance is $\leq 12$. 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).

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

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