Characterization of SARS-CoV-2 public CD4+ αβ T cell clonotypes through reverse epitope discovery

Elisa Rosati[1]*, Mikhail V. Pogorelyy[2]*, Anastasia A. Minervina[2]*, Alexander Scheffold[3], Andre Franke[1], Petra Bacher[1,3#] and Paul G. Thomas[2#+]

[1]Institute of Clinical Molecular Biology, Christian-Albrecht University of Kiel, Kiel, Germany.

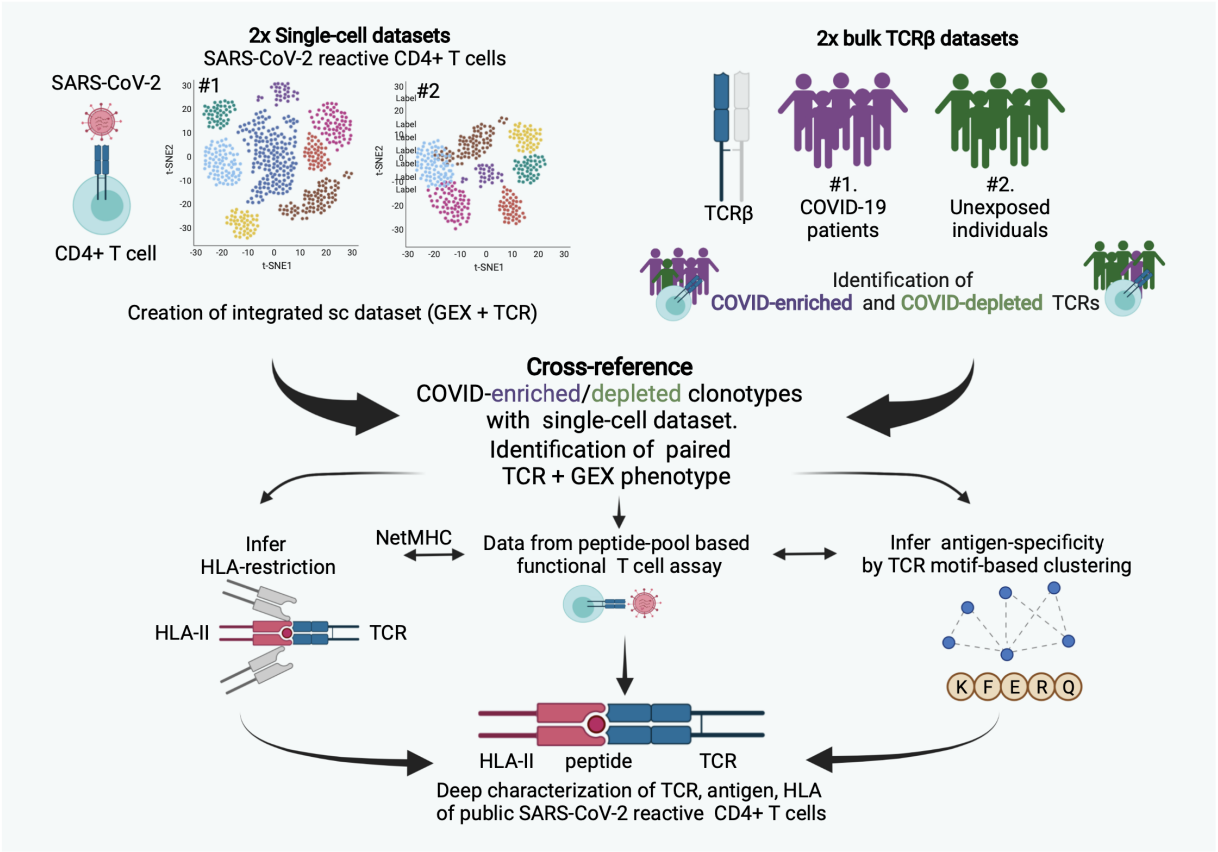
[2]St. Jude Children's Research Hospital, Department of Immunology, Memphis, TN, 38103.

[3]Institute of Immunology, Christian-Albrecht University of Kiel, Kiel, Germany

*Shared first authorship

[#]Shared last authorship

[+]Corresponding author. Email: paul.thomas@stjude.org



Highlights

- Identification of highly public CD4+ T cell responses to SARS-CoV-2
- Systematic prediction of exact immunogenic HLA class II epitopes for CD4+ T cell response
- Methodological framework for reverse epitope discovery, which can be applied to other disease contexts and may provide essential insights for future studies and clinical applications

Abstract

The amount of scientific data and level of public sharing produced as a consequence of the COVID-19 pandemic, as well as the speed at which these data were produced, far exceeds any previous effort against a specific disease condition. This unprecedented situation allows for development and application of new research approaches. One of the major technical hurdles in immunology is the characterization of HLA-antigen-T cell receptor (TCR) specificities. Most approaches aim to identify reactive T cells starting from known antigens using functional assays. However, the need for a reverse approach identifying the antigen specificity of orphan TCRs is increasing.

Utilizing large public single-cell gene expression and TCR datasets, we identified highly public $CD4^+$ T cell responses to SARS-CoV-2, covering >75% of the analysed population. We performed an integrative meta-analysis to deeply characterize these clonotypes by TCR sequence, gene expression, HLA-restriction, and antigen-specificity, identifying strong and public $CD4^+$ immunodominant responses with confirmed specificity. $CD4^+$ COVID-enriched clonotypes show T follicular helper functional features, while clonotypes depleted in SARS-CoV-2 individuals preferentially had a central memory phenotype. In total we identify more than 1200 highly public CD4+ T cell clonotypes reactive to SARS-CoV-2. TCR similarity analysis showed six prominent TCR clusters, for which we predicted both HLA-restriction and cognate SARS-CoV-2 immunodominant epitopes. To validate our predictions we used an independent cohort of TCR repertoires before and after vaccination with *ChAdOx1,* a replication-deficient simian adenovirus-vectored vaccine*,* encoding the SARS-CoV-2 spike protein. We find statistically significant enrichment of the predicted spike-reactive TCRs after vaccination with *ChAdOx1,* while the frequency of TCRs specific to other SARS-CoV-2 proteins remains stable. Thus, the CD4-associated TCR repertoire differentiates vaccination from natural infection.

In conclusion, our study presents a novel reverse epitope discovery approach that can be used to infer HLA- and antigen-specificity of orphan TCRs in any context, such as viral infections, antitumor immune responses, or autoimmune disease.

Introduction

The worldwide scientific effort to overcome the COVID-19 pandemic led to the generation of an extraordinarily large amount of publicly available data describing the human immune response to SARS-CoV-2. A lot of these studies point to the importance of a T cell response in resolving COVID-19, as well as providing long-term protection against the variants of SARS-CoV-2 (Dan et al., 2021; Geers et al., 2021; Rydyznski Moderbacher et al., 2020; Sekine et al., 2020; Tarke et al., 2021a). One of the increasingly popular ways to study the complexity of the T cell response is T cell receptor repertoire sequencing (Mukhopadhyay, 2021). The T cell receptor (TCR) is a heterodimer of alpha and beta chains, both of which are formed in a semi-random DNA recombination process resulting in a unique repertoire in each individual (Dupic et al., 2021). However, even individual diverse T cell repertoires can show similar features after encountering the same antigens. In particular, response to immunodominant epitopes triggers large clonal expansions, and TCRs recognising such epitopes frequently have highly similar sequences (Dash et al., 2017; Glanville et al., 2017; Pogorelyy et al., 2019). Thus, analysis of TCR repertoires could shed light on the differences and commonalities of the T cell immune response among different individuals, on the identity of the most immunogenic antigens, and may provide targets for development of diagnostic tools as well as therapeutic treatments such as adoptive T cell transfers (NLM, NCT04762186 clinical trial). As a proof of principle, identification of T cell clonotypes reactive to SARS-CoV-2 antigens has eventually led to the development of a diagnostic test for SARS-CoV-2 authorized for emergency use by the FDA (Dalai et al., 2021).

In order to identify TCR repertoire signatures related to COVID-19, multiple groups utilized bulk TCR repertoire sequencing (Minervina et al., 2021a; Niu et al., 2020; Nolan et al., 2020; Schultheiß et al., 2020; Shomuradova et al., 2020; Snyder et al., 2020), that quantitatively measures frequencies of large numbers of unpaired TCRalpha or TCRbeta clonotypes, as well as single cell TCR sequencing techniques (Bacher et al., 2020; Bernardes et al., 2020; Kusnadi et al., 2021; Liao et al., 2020; Lu et al., 2021; Meckiff et al., 2020; Wen et al., 2020; Xu et al., 2020; Zhang et al., 2020), that produce many fewer, but paired, alpha/beta TCR sequences. One of the major challenges of such approaches is that, even at the peak of the immune response to

SARS-CoV-2 infection, only a small fraction of total peripheral T cells recognize viral epitopes. Hence, many studies rely on T cell antigen-enrichment methodologies, such as MHC-multimer-staining or peptide stimulation with subsequent enrichment for activated cells (e.g. ARTE assay, AIM assay, MIRA assay) (Altman et al., 1996; Bacher et al., 2013; Klinger et al., 2015; Reiss et al., 2017). These approaches increase the number of SARS-CoV-2 specific TCRs detected in each sample and help to identify immunodominant epitopes (reviewed in Grifoni et al. 2021). Stimulating T cells with peptide libraries is the most frequent approach used for SARS-CoV-2 epitope-discovery (Braun et al., 2020; Grifoni et al., 2020; Lu et al., 2021; Mateus et al., 2020; Nelde et al., 2021; Peng, Yanchun et al., 2020; Tarke et al., 2021b). Here, we focus on public clonotypes (clones found in multiple individuals), which provide the necessary power for a robust statistical analysis and, in addition, hold the highest potential for further, population-wide applications. We propose a reverse epitope discovery technique, which, instead of starting from the epitopes to identify reactive T cells, utilizes TCR repertoires as the means to find immunodominant responses in an unbiased manner. We performed a comprehensive TCR meta-analysis of publicly available single cell and bulk TCR repertoire datasets and identified SARS-CoV-2 reactive TCRs with complete TCR alpha and beta chain information and inferred their HLA restriction. Moreover, through TCR clustering based on sequence similarity, we were able to identify several prominent alpha/beta TCR motifs and predict their antigen specificity.

Results and discussion

In this report, we jointly analysed **(1)** two published single-cell datasets of SARS-CoV-2-reactive $CD4^+$ T cells, identified based on CD154+ up-regulation after peptide pool stimulation and including gene expression and TCR information for a total of 59 individuals, of which 49 were COVID-19 patients and 10 healthy controls (Bacher et al., 2020; Meckiff et al., 2020) and **(2)** the largest published bulk TCRbeta datasets of 786 healthy individuals (Emerson et al., 2017) and 1414 COVID-19 patients (Snyder et al., 2020), including TCRs with known specificity for certain SARS-CoV-2 peptide pools (Nolan et al., 2020) and **(3)** a published bulk TCR dataset before and after SARS-CoV-2 vaccination with *ChAdOx1* (Swanson et al., 2021), which we used as validation for our findings.

In order to find public CD4$^+$ T cell responses to SARS-CoV-2 infection we first merged two publicly available single-cell datasets of CD4$^+$ SARS-CoV-2-reactive T cells (Bacher et al., 2020; Meckiff et al., 2020). Both datasets were obtained using the same antigen-reactive T cell enrichment assay (ARTE-assay), based on CD154+ up-regulation on antigen-reactive CD4+ T cells (Bacher et al., 2013). Of note, differently from Bacher et al. the antigenic pool used in Meckiff et al. only includes peptides from the spike protein (without N-terminal domain) and the membrane glycoprotein. The combined dataset contained 125,258 cells that passed quality control steps, which resulted in 13 functional clusters after unsupervised analysis (**Figure 1A**). Cluster phenotypes were defined using markers of cell populations used in the original publications (Bacher et al., 2020; Meckiff et al., 2020). In particular, we found clusters corresponding to helper follicular T (TfH) cells (clusters 1-2), helper type 1 T (Th1) cells (cluster 3), transitional Tfh/Tcm T cells (cluster 4), central memory T (Tcm) cells (clusters 5), Th17 phenotypes (clusters 6-7), effector memory T cells (cluster 8), type I IFN-signature T cells (clusters 9-10), cytotoxic T cells (clusters 11-12), and cycling T cells (cluster 13) (**Figure 1A-B, Supplementary Table 1**). Using this extended dataset, we also confirmed previous findings of a significant enrichment of TfH T cells (Clusters 1 and 2) in COVID-19 patients in comparison to healthy controls (**Figure 1C-D**). Interestingly, the abundance of the two main TfH T cell subsets was significantly different in hospitalized versus non-hospitalized COVID-19 patients (**Figure 1D**). In particular, cluster 2 was enriched in severe disease, and expressed higher cytotoxic markers such as *CCL3*, *CCL4, CCL5*, *XCL1*, *XCL2*, *GZMB* and *GNLY.* In contrast, cluster 1 expressed higher levels of *IL2*, *CD69* and genes of the *TNF* family, and was nominally enriched in mild-disease patients as indeed previously shown by Meckiff *et al*. (**Figure 1D, Supplementary Table 1**).

In order to select TCRs corresponding to the most public CD4$^+$ T cell responses, we next searched TCRbeta sequences from the combined single cell dataset in the TCR repertoires from a large cohort of COVID-19 patients (Snyder et al., 2020), as well as from pre-pandemic COVID19-naive controls (Emerson et al., 2017). We identified TCRs shared among individuals and strongly associated with SARS-CoV-2 infection, as defined by presence in significantly more patients than controls (COVID-enriched TCRs) (Fisher's exact test), which are reported in

Supplementary Table 2. We also identified TCRs which were significantly decreased in COVID-19 patients as compared to the healthy population (COVID-depleted TCRs), reported in **Supplementary Table 3** (**Figure 2A)**. Notably, when mapping COVID-enriched TCRs to the single-cell RNAseq data, these significantly accumulated in TfH T cell clusters (clusters 1,2,4) while COVID-depleted TCRs rather accumulated in effector memory subpopulations (clusters 6,7,8) (**Figure 2B**). The effect size (log2-fold enrichment) for COVID-depleted TCRs was much smaller than for COVID-enriched TCRs (**Figure 2A).** Moreover, COVID-depleted TCRs were present in a large fraction of donors from both the control and the COVID-19 cohorts (**Figure 2C)**, e.g., the majority of COVID-depleted TCRs (465 out of 594) were simultaneously found in >100 controls and >100 COVID-19 patients, while only 24 of 1248 COVID-enriched clonotypes had the same level of publicity. The number of unique T cell clones in a subset of the analyzed COVID-19 patients was low in comparison to healthy controls, potentially due to COVID-19 associated lymphopenia. This could lead to small, yet significant, underrepresentation of highly public clonotypes in the COVID-19 cohort. We therefore focused on the COVID-enriched TCR clonotypes for further analysis, because the occurrence pattern and phenotype of this group is consistent with expansion of T cell clones specific for SARS-CoV-2 antigens.

Next, we aimed to identify TCRs with highly similar sequence motifs and thus likely to have the same antigen specificity (Dash et al., 2017; Glanville et al., 2017). We used TCRdist (Dash et al., 2017) to identify highly similar TCR sequences among TCRs enriched in COVID-19 and indeed found several prominent TCR clusters (**Figure 2D**). We also inferred the HLA-restriction for most of these TCRs, using the procedure suggested in (Minervina et al., 2021a), see **Figure 2E** for the representative Manhattan plot (and **Supplementary Table 2** for predicted HLA-restrictions).

TCRdist uses both TCRalpha and TCRbeta amino acid sequences to compute the distance between clonotypes. However, analysis of the resulting sequence motifs demonstrate unequal contribution of TCR alpha and beta chains in different clusters. For example, cluster 3 is largely defined by a conserved beta chain motif, allowing for diverse alpha chains, while in cluster 2 there is an almost invariant alpha chain paired with a set of very diverse TCRbeta chains. In a few other large clusters both TCR chains show strongly conserved amino acid motifs (**Figure**

2D). These differences could be potentially explained by the variable number of contacts of TCRalpha/TCRbeta chains with the antigenic peptide and MHC. Thus alpha-driven, beta-driven and alpha/beta driven motifs are interesting targets for solving TCR-pMHC ternary structures. To predict the potential antigen-reactivity of the TCR clusters we next cross-referenced the TCRbeta sequences of COVID-enriched clonotypes with the MIRA MHC-II dataset. This database contains specificity information of certain $CD4^+$ TCRbeta sequences towards peptide pools of SARS-CoV-2 proteins. Specifically, in the MIRA MHC-II experiment, TCR-antigen specificity was tested against 56 peptide pools containing 1 to 6 overlapping peptide 19mers each, and spanning over the membrane (M), nucleocapsid (N) and spike (S) SARS-CoV-2 proteins. Our set of COVID-enriched TCRs mapped to 22 peptide pools, with most matches to $M_{149\text{-}191}$, $N_{46\text{-}96}$ and $S_{743\text{-}801}$ pools. The number of TCRs matching into the MIRA database was significantly larger in the COVID-enriched TCRs (34%, 428 out of 1248) as compared to the COVID-depleted TCRs (13%, 80 out of 594) suggesting that the former group is indeed more enriched for common COVID-reactive clonotypes. Most TCRs belonging to the same TCRdist cluster were assigned to the same peptide pool from the MIRA MHC-II dataset.

Inference of cognate peptide pools and HLA-restriction allowed us to use NetMHC to predict specific antigenic peptides within the pool and also to validate our prediction. In fact, the predicted antigenic peptide and HLA restriction for TCRdist cluster 2 and cluster 5 exactly match the experimental results from (Mudd et al., 2021) and (Lu et al., 2021) respectively, thus supporting the validity of our methodology.

Finally, to validate the set of COVID-enriched $CD4^+$ clonotypes using an independent dataset, we used a large collection of TCRbeta repertoires before and after SARS-CoV-2 vaccination with *ChAdOx1* (Swanson et al., 2021), a replication-deficient simian adenovirus-vectored vaccine, encoding the SARS-CoV-2 spike protein. For each TCR repertoire from the pre-vaccination (day 0) or post-vaccination time point (day 28) we calculate the fraction of unique TCRbeta clonotypes matching TCRbeta sequences from each of our largest antigen-specific TCR clusters (**Figure 2D**). We find significant enrichment of the predicted spike-reactive TCRs of clusters #2 and #5 after vaccination with *ChAdOx1,* while the frequency of TCR clusters reactive to the membrane or nucleocapsid proteins remain unchanged (**Figure**

2G). This result serves as an independent validation of our approach and, at the same time, shows how the TCR clusters we identified may be potentially used to identify SARS-CoV-2 epitope-specific TCRs both in the context of vaccination and natural infection.

Conclusions

In conclusion, our study identified 1248 paired TCR clonotypes potentially specific to highly immunogenic epitopes from SARS-CoV-2. Many of these TCRs remain orphan of their epitopes. However, we identified and inferred antigen-specificity for 428 TCRs with the aid of the MIRA dataset (Nolan et al., 2020). We also inferred HLA-restriction for most of these TCRs. TCR-HLA pairings were also validated based on NetMHC binding predictions and functional experiments in other publications (Lu et al., 2021; Mudd et al., 2021) The resulting set of highly characterized public TCRs reactive to SARS-CoV-2 covers more than 76 percent of individuals from Snyder et al, i.e. at least 20 unique COVID-enriched TCR sequences per individual were found. The broad coverage of HLA haplotypes in our study also provides means to further investigate immunodominant responses in a genetically diverse population.

No immunodominant SARS-CoV-2 epitopes had yet been reported for most HLA class II alleles, especially if underrepresented in European populations. Furthermore, the high publicity of the characterized clonotypes makes them promising candidates for further studies on CD4+ T cell immunity against SARS-CoV-2 as well as for immunotherapeutic applications aimed at utilizing highly specific, immunodominant, T cell responses in the context of precision and personalized medicine (Müller et al., 2021; O'Reilly et al., 2016; Qian et al., 2018).

Study limitations

The described method of reverse epitope discovery also has limitations that we want to highlight here. The method is strongly focused on public T cell responses which, on one side, make the findings significant and applicable to a larger fraction of the population. On the other side, immunodominant responses may be also driven by private clonotypes or clonotypes without an identifiable motif cluster, which would not be detected by our method (or that would not appear among the most interesting hits). Although this is a limitation of our approach, it may be a field to expand to in order to identify more T cell responses. Similarly, the publicity of HLA alleles is

also a limiting factor of our method, as clonotypes recognizing peptides presented on rare HLA alleles would be hard to detect.

Another limitation lies in the availability of TCR antigen-specificity information from public databases and resources, such as VDJdb (Bagaev et al., 2020) and the MIRA dataset (Nolan et al., 2020). The expansion of these resources is of utmost importance for target identification of orphan TCRs. The strong general interest in the COVID-19 pandemic led to an unprecedented production of high amounts of TCR repertoire data, which is until now unmatched for other antigens or diseases.

Despite these limitations, we think that the reverse epitope discovery approach already proved itself valuable for identifying both cross-reactive (Minervina et al., 2021b) and immunodominant responses to SARS-CoV-2 (Mudd et al., 2021) and it holds potential for application in other disease contexts.

Methods

Utilized public data

Single-cell data of SARS-CoV-2 reactive $CD154^+$ T cells were obtained from (Bacher et al., 2020) and (Meckiff et al., 2020). Bulk TCR data of healthy individuals and COVID-19 patients were the datasets used in (Emerson et al., 2017) and (Snyder et al., 2020), respectively.

Single-cell datasets integration and filtering

The preprocessing of the scRNAseq data was performed with the 10x Genomics' Cell Ranger software v3.1.0 using the human genome reference GRCh38 v3.0.0 for the mappings. The resulting raw feature-barcode matrix files were analyzed with the R package Seurat v.3.2.0 (Butler et al., 2018). Thereby, all genes with a detected expression in less than 0.1% of the non-empty cells were excluded. Moreover, TCR genes were not considered for further analyses to avoid functional clustering of cells based on TCR information. To minimize the number of doublets, empty cells, and cells with a low-quality transcriptome, only cells harboring between

400 and 3000 RNA features and less than 5% mitochondrial RNA were selected for further processing.

TCR information was integrated into the metadata of the Seurat object after filtering of cells containing more than 2 TCR alpha or 2 TCRbeta chains. After merging of Seurat metadata and TCR information, cells without TCR information were excluded from further analysis. Afterwards, data were log-normalized and scaled based on all genes. After performing a PCA dimensionality reduction (40 dimensions) with the RunPCA function, the expression values were corrected for batch effects caused by different sources of the data , sample preparation batches and sequencing run batches using the R package Harmony v1.0 (Korsunsky et al., 2019). In the final steps, the Uniform Manifold Approximation and Projection (UMAP) dimensional reduction was performed with the RunUMAP function using 40 dimensions, a shared nearest neighbor graph was created with the FindNeighbors method, and the clusters identification was performed with a resolution of 0.4 using theFindClusters function. 13 clusters were identified. Cluster marker genes were determined using FindMarkers with the MAST method (Finak et al., 2015).

COVID-19 TCR association using bulk TCR public datasets

To identify public TCRbeta clonotypes we used two large datasets, one of COVID patients (n=1414, Snyder et al. 2020) and one of healthy subjects sampled pre pandemic (n=786, Emerson et al. 2017). For each TCRbeta from combined single cell TCRseq dataset we calculate number of unique donors from both bulk TCRbeta repertoire cohorts sharing it (a TCRbeta is considered shared if both CDR3 amino acid sequence and V segment family match, as suggested in (DeWitt et al., 2018)). Next, we use a two-sided Fisher exact test with Benjamini-Hochberg multiple-testing correction to identify sequences overrepresented (i.e. found in more donors) in either cohort (adjusted p-value<0.05 is used as significance threshold).

Identification of motifs in TCR amino acid sequences using TCRdist

We used the TCRdist implementation in *conga* python package to calculate pairwise TCRdist between unique abTCR sequences and plot sequence logos for TCR motifs (Schattgen et al., 2021). We define TCR motifs as connected components on the TCR similarity network, where

each node is a unique alpha/betaTCR clonotype, and edge connects them if distance is less than 120 TCRdist units. To filter TCR chimeras and other artifacts occuring during 10x Genomics sequencing resulting in rare spurious connections between motif clusters, we deleted top 1% of nodes and vertices by network betweenness centrality values. *igraph R* package was used to manipulate similarity networks (Csardi and Nepusz, 2006), *gephi* was used for network layout and visualisation (Jacomy et al., 2014).

HLA specificity imputation from TCR data

For each donor from (Emerson et al., 2017; Snyder et al., 2020) we use HLA-types inferred in (Minervina et al., 2021a). Then for each TCRbeta significantly enriched in COVID cohort we do a one-sided Fisher exact test with Benjamini-Hochberg multiple testing correction to check if given TCRbeta co-occurs with each of HLA-alleles.

Prediction of COVID-enriched TCR specificity

We mapped TCRbeta chain sequences from aggregated single cell dataset to peptide-pool specific TCRbeta clonotypes for MIRA class II dataset (release 002.1) allowing for one amino acid mismatch in CDR3 amino acid sequence. Next we selected six large clusters on the TCRdist similarity network with distinct MIRA peptide pool assignments, calculated consensus MIRA pool and HLA-restriction within a cluster and used NetMHCIIpan-4.0 (Reynisson et al., 2020) to predict the antigenic peptide within peptide pool.

Statistical analysis

Statistical analysis was performed in R version 4.0.2. Wilcoxon rank-sum test (Mann-Whitney U test) was used to compare the proportion of cells in each Seurat functional cluster between healthy controls and COVID-19 patients as well as between severe and mild COVID-19 cases. Fisher exact test was used to compare the number of COVID-depleted and COVID-enriched clonotypes being part of each functional Seurat cluster. Multiple testing correction was performed using the Benjamini-Hochberg procedure. Ns not significant, * $p<0.5$, ** $p<0.01$, *** $p<0.001$

Data and code availability

All analysed data are publicly available. In detail, single-cell data can be found under SRA accession numbers SRP293741 (Bacher et al., 2020) and SRP267404 (Meckiff et al., 2020). Bulk TCR repertoire data from COVID infected subjects, and MIRA Class II dataset (release 002.1) (Snyder et al., 2020) are publicly available from ImmuneAccess database (https://clients.adaptivebiotech.com/pub/covid-2020), as well as healthy control data from (Emerson et al., 2017) (https://clients.adaptivebiotech.com/pub/emerson-2017-natgen) and *ChAdOx1* immunized cohort (https://doi.org/10.21417/PAS2021STM). Utilized scripts for the analysis of the merged single-cell dataset are available on the GitHub page https://github.com/pogorely/reverse_epitope_discovery

Acknowledgements

This research was supported by the German Research Foundation (DFG) under Germany's Excellence Strategy—EXC 2167-390884018 Precision Medicine in Chronic Inflammation to P.B., A.F., and A.S.; DFG 433038070 to P.B., A.F., and A.S.; DFG 4096610003 to E.R; by a COVID-19 research grant from the Land Schleswig-Holstein, DIO002/ CoVispecT to P.B. and A.S. This research was partially supported by R01AI136514 (P.G.T).

Author contributions

Conceptualization: E.R, M.V.P, A.A.M; Analysis: E.R, M.V.P, A.A.M.; Visualization: A.A.M, E.R; Resources: P.T, P.B., A.S, A.F.; Supervision and coordination: P.B. and P.T. Writing original draft: E.R, M.V.P., A.A.M. All authors provided discussion, participated in revising the manuscript, and agreed to the final version.

Declaration of interest

P.B. and A.S. are consultants of Miltenyi Biotec, who own IP rights concerning parts of the ARTE technology. P.G.T has consulted and/or received honoraria and travel support from Illumina, Johnson and Johnson, and 10X Genomics. P.G.T. serves on the Scientific Advisory

Board of Immunoscape and Cytoagents. The authors have applied for patents covering some aspects of these studies.

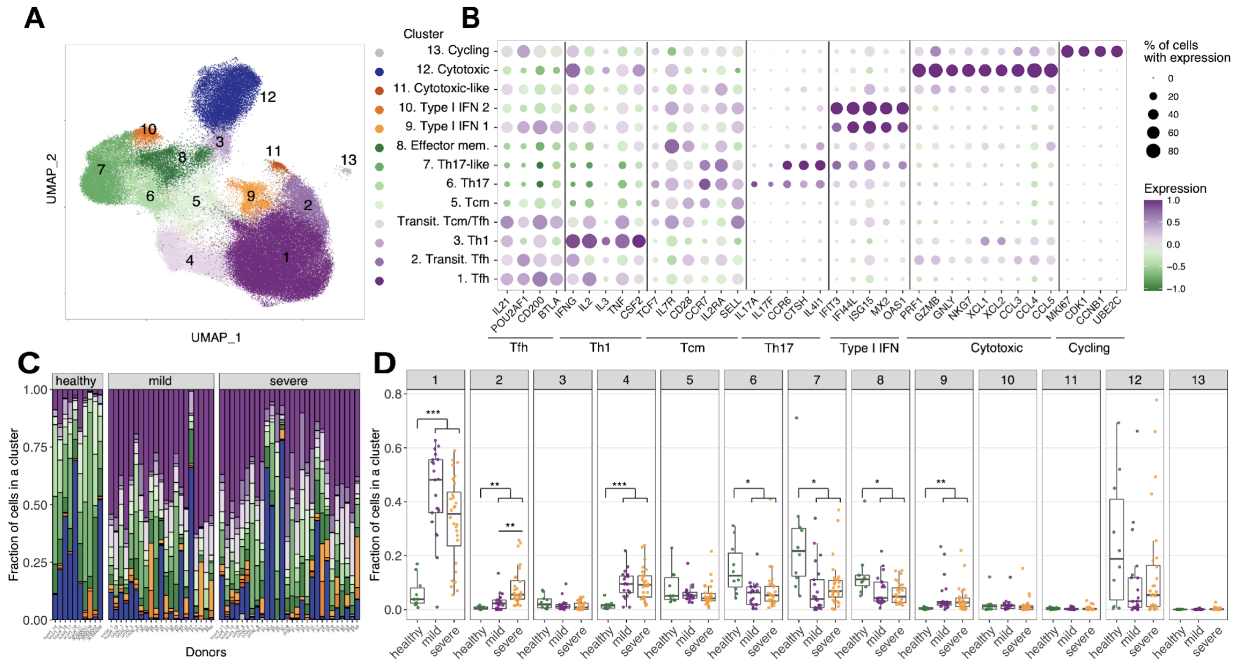


Figure 1: A. UMAP of single cells for merged (Bacher et al., 2020; Meckiff et al., 2020) datasets of SARS-CoV-2 antigen-enriched CD4 T cells based. Colors show clusters of cells with distinct gene expression profiles **B**. Differentially expressed genes in each GEX cluster. **C.** Distribution of cells between GEX clusters is plotted for each donor. Healthy donors less Tfh cells (populations 1 and 2) **D.** Boxplots showing cell proportion distribution among functional clusters for each patient (Mann-Whitney U test, Bonferroni multiple comparison correction).

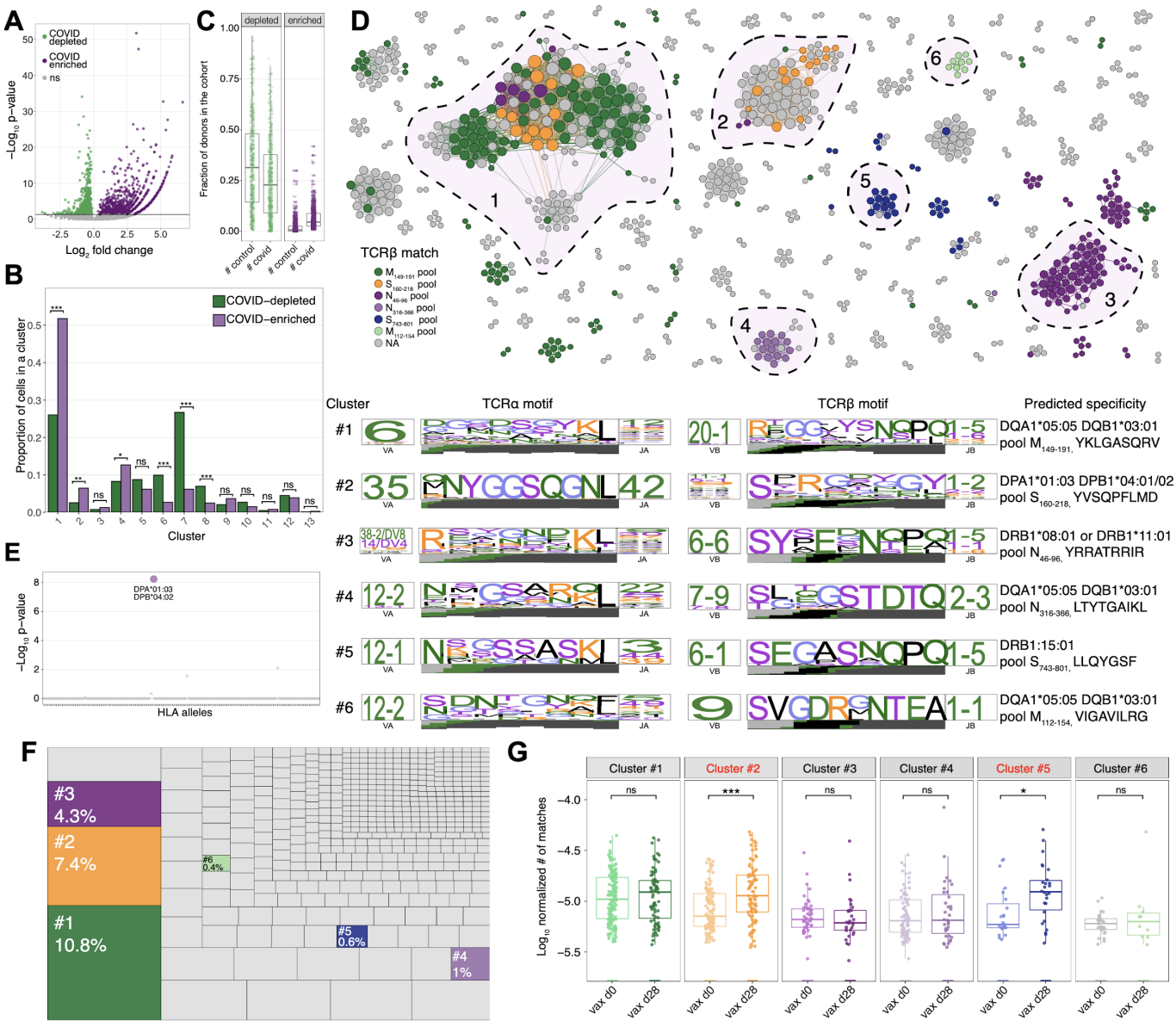


Figure 2: A. Volcano plot shows enrichment of TCRbeta chains identified in merged scTCRseq dataset in large (n=1414) collection of repertoires from COVID patients ((Snyder et al., 2020), purple) in comparison to the healthy donor cohort from Emerson et al (n=786) (x-axis) vs p-value (y-axis). **B**. Barplot showing the distribution of COVID-enriched (purple) and COVID-depleted (green) TCR clonotypes in GEX clusters. Fisher exact test was used for the comparison, with Bonferroni multiple comparison correction. **C.** The boxplots show a fraction of donors from healthy and COVID-19 cohorts sharing significantly COVID-depleted (green) and

COVID-enriched (purple) clonotypes. **D.** Similarity network of COVID-associated public TCR clonotypes. Each vertex represents a TCR alpha/beta clonotype, edges connect vertices with <120 TCRdist units. Colors show predicted SARS-CoV-2 peptide pool from MIRA class II dataset. Bottom: TCRdist logos for the most prominent clonotype clusters with predicted peptide specificity and HLA-restriction. **E.** Manhattan plot for association of representative cluster #2 clonotype with various HLA-types. **F.** A treemap showing the fraction of repertoire occupied by clonotypes from prominent TCR similarity clusters from (c). **G.** Occurrence of TCRbeta from 6 large clusters from (c) before and after SARS-CoV-2 vaccination with ChAdOx1 vaccine. Significantly more TCRs from spike-specific clusters #2 and #5 are found after vaccination (one sided Wilcoxon rank sum test with Benjamini-Hochberg multiple testing correction).

Supplementary files:

Supplementary Table 1: Differentially expressed genes for each functional Seurat cluster.

Supplementary Table 2: Unique alpha/beta TCR clonotypes enriched in COVID-19 cohort

Supplementary Table 3: Unique alpha/beta TCR clonotypes enriched in healthy control cohort

References

Altman, J.D., Moss, P.A.H., Goulder, P.J.R., Barouch, D.H., McHeyzer-Williams, M.G., Bell, J.I., McMichael, A.J., and Davis, M.M. (1996). Phenotypic Analysis of Antigen-Specific T Lymphocytes. Science *274*, 94–96.

Bacher, P., Schink, C., Teutschbein, J., Kniemeyer, O., Assenmacher, M., Brakhage, A.A., and Scheffold, A. (2013). Antigen-Reactive T Cell Enrichment for Direct, High-Resolution Analysis of the Human Naive and Memory Th Cell Repertoire. J. Immunol. *190*, 3967–3976.

Bacher, P., Rosati, E., Esser, D., Martini, G.R., Saggau, C., Schiminsky, E., Dargvainiene, J., Schröder, I., Wieters, I., Khodamoradi, Y., et al. (2020). Low-Avidity CD4+ T Cell Responses to SARS-CoV-2 in Unexposed Individuals and Humans with Severe COVID-19. Immunity *53*, 1258-1271.e5.

Bagaev, D.V., Vroomans, R.M.A., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E.S., Zvyagin, I.V., et al. (2020). VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res. *48*, D1057–D1062.

Bernardes, J.P., Mishra, N., Tran, F., Bahmer, T., Best, L., Blase, J.I., Bordoni, D., Franzenburg, J., Geisen, U., Josephs-Spaulding, J., et al. (2020). Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19. Immunity *53*, 1296-1314.e9.

Braun, J., Loyal, L., Frentsch, M., Wendisch, D., Georg, P., Kurth, F., Hippenstiel, S., Dingeldey, M., Kruse, B., Fauchere, F., et al. (2020). SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. Nature *587*, 270–274.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal *Complex Systems*, 1695.

Dalai, S.C., Dines, J.N., Snyder, T.M., Gittelman, R.M., Eerkes, T., Vaney, P., Howard, S., Akers, K., Skewis, L., Monteforte, A., et al. (2021). Clinical Validation of a Novel T-cell Receptor Sequencing Assay for Identification of Recent or Prior SARS-CoV-2 Infection (Infectious Diseases (except HIV/AIDS)).

Dan, J.M., Mateus, J., Kato, Y., Hastie, K.M., Yu, E.D., Faliti, C.E., Grifoni, A., Ramirez, S.I., Haupt, S., Frazier, A., et al. (2021). Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. Science *371*, eabf4063.

Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature *547*, 89–93.

DeWitt, W.S., Smith, A., Schoch, G., Hansen, J.A., Matsen, F.A., and Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. ELife *7*, e38358.

Dupic, T., Bensouda Koraichi, M., Minervina, A.A., Pogorelyy, M.V., Mora, T., and Walczak, A.M. (2021). Immune fingerprinting through repertoire similarity. PLOS Genet. *17*, e1009301.

Emerson, R.O., DeWitt, W.S., Vignali, M., Gravley, J., Hu, J.K., Osborne, E.J., Desmarais, C., Klinger, M., Carlson, C.S., Hansen, J.A., et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat. Genet. *49*, 659–665.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278.

Geers, D., Shamier, M.C., Bogers, S., den Hartog, G., Gommers, L., Nieuwkoop, N.N., Schmitz, K.S., Rijsbergen, L.C., van Osch, J.A.T., Dijkhuizen, E., et al. (2021). SARS-CoV-2 variants of concern partially escape humoral but not T-cell responses in COVID-19 convalescent donors and vaccinees. Sci. Immunol. *6*, eabj1750.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. Nature *547*, 94–98.

Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings, S.A., Sutherland, A., Premkumar, L., Jadi, R.S., et al. (2020). Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell *181*, 1489-1501.e15.

Grifoni, A., Sidney, J., Vita, R., Peters, B., Crotty, S., Weiskopf, D., and Sette, A. (2021). SARS-CoV-2 Human T cell Epitopes: adaptive immune response against COVID-19. Cell Host Microbe S1931312821002389.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE *9*, e98679.

Klinger, M., Pepin, F., Wilkins, J., Asbury, T., Wittkop, T., Zheng, J., Moorhead, M., and Faham, M. (2015). Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. PLOS ONE *10*, e0141561.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

Kusnadi, A., Ramírez-Suástegui, C., Fajardo, V., Chee, S.J., Meckiff, B.J., Simon, H., Pelosi, E., Seumois, G., Ay, F., Vijayanand, P., et al. (2021). Severely ill COVID-19 patients display impaired exhaustion features in SARS-CoV-2-reactive CD8 $^{+}$ T cells. Sci. Immunol. *6*, eabe4782.

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat. Med. *26*, 842–844.

Lu, X., Hosono, Y., Nagae, M., Ishizuka, S., Ishikawa, E., Motooka, D., Ozaki, Y., Sax, N., Maeda, Y., Kato, Y., et al. (2021). Identification of conserved SARS-CoV-2 spike epitopes that expand public cTfh clonotypes in mild COVID-19 patients. J. Exp. Med. *218*, e20211327.

Mateus, J., Grifoni, A., Tarke, A., Sidney, J., Ramirez, S.I., Dan, J.M., Burger, Z.C., Rawlings, S.A., Smith, D.M., Phillips, E., et al. (2020). Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science *370*, 89–94.

Meckiff, B.J., Ramírez-Suástegui, C., Fajardo, V., Chee, S.J., Kusnadi, A., Simon, H., Eschweiler, S., Grifoni, A., Pelosi, E., Weiskopf, D., et al. (2020). Imbalance of Regulatory and Cytotoxic SARS-CoV-2-Reactive CD4+ T Cells in COVID-19. Cell *183*, 1340-1353.e16.

Minervina, A.A., Komech, E.A., Titov, A., Bensouda Koraichi, M., Rosati, E., Mamedov, I.Z., Franke, A., Efimov, G.A., Chudakov, D.M., Mora, T., et al. (2021a). Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. ELife *10*, e63502.

Minervina, A.A., Pogorelyy, M.V., Kirk, A.M., Allen, E.K., Allison, K.J., Lin, C.-Y., Brice, D.C., Zhu, X., Vegesana, K., Wu, G., et al. (2021b). Convergent epitope-specific T cell responses after SARS-CoV-2 infection and vaccination (Infectious Diseases (except HIV/AIDS)).

Mudd, P.A., Minervina, A.A., Pogorelyy, M.V., Turner, J.S., Kim, W., Kalaidina, E., Petersen, J., Schmitz, A.J., Lei, T., Haile, A., et al. (2021). SARS-CoV-2 mRNA vaccination elicits robust and persistent T follicular helper cell response in humans (Immunology).

Mukhopadhyay, M. (2021). Diving into the TCR repertoire. Nat. Methods *18*, 30–30.

Müller, T.R., Jarosch, S., Hammel, M., Leube, J., Grassmann, S., Bernard, B., Effenberger, M., Andrä, I., Chaudhry, M.Z., Käuferle, T., et al. (2021). Targeted T cell receptor gene editing provides predictable T cell product function for immunotherapy. Cell Rep. Med. *2*, 100374.

Nelde, A., Bilich, T., Heitmann, J.S., Maringer, Y., Salih, H.R., Roerden, M., Lübke, M., Bauer, J., Rieth, J., Wacker, M., et al. (2021). SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. Nat. Immunol. *22*, 74–85.

Niu, X., Li, S., Li, P., Pan, W., Wang, Q., Feng, Y., Mo, X., Yan, Q., Ye, X., Luo, J., et al. (2020). Longitudinal Analysis of T and B Cell Receptor Repertoire Transcripts Reveal Dynamic Immune Response in COVID-19 Patients. Front. Immunol. *11*, 582010.

Nolan, S., Vignali, M., Klinger, M., Dines, J.N., Kaplan, I.M., Svejnoha, E., Craft, T., Boland, K., Pesesky, M., Gittelman, R.M., et al. (2020). A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. (In Review).

O'Reilly, R.J., Prockop, S., Hasan, A.N., Koehne, G., and Doubrovina, E. (2016). Virus-specific T-cell banks for "off the shelf" adoptive therapy of refractory infections. Bone Marrow Transplant. *51*, 1163–1172.

Peng, Yanchun, Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., Dejnirattisai, W., Rostron, T., Supasa, P., Liu, C., et al. (2020). Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. Nat. Immunol. *21*, 1336–1345.

Pogorelyy, M.V., Minervina, A.A., Shugay, M., Chudakov, D.M., Lebedev, Y.B., Mora, T., and Walczak, A.M. (2019). Detecting T cell receptors involved in immune responses from single repertoire snapshots. PLOS Biol. *17*, e3000314.

Qian, C., Wang, Y., Reppel, L., D'aveni, M., Campidelli, A., Decot, V., and Bensoussan, D. (2018). Viral-specific T-cell transfer from HSCT donor for the treatment of viral infections or diseases after HSCT. Bone Marrow Transplant. *53*, 114–122.

Reiss, S., Baxter, A.E., Cirelli, K.M., Dan, J.M., Morou, A., Daigneault, A., Brassard, N., Silvestri, G., Routy, J.-P., Havenar-Daughton, C., et al. (2017). Comparative analysis of activation induced marker (AIM) assays for sensitive identification of antigen-specific CD4 T cells. PLOS ONE *12*, e0186998.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.

Rydyznski Moderbacher, C., Ramirez, S.I., Dan, J.M., Grifoni, A., Hastie, K.M., Weiskopf, D., Belanger, S., Abbott, R.K., Kim, C., Choi, J., et al. (2020). Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. Cell *183*, 996-1012.e19.

Schattgen, S.A., Guion, K., Crawford, J.C., Souquette, A., Barrio, A.M., Stubbington, M.J.T., Thomas, P.G., and Bradley, P. (2021). Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). Nat. Biotechnol.

Schultheiß, C., Paschold, L., Simnica, D., Mohme, M., Willscher, E., von Wenserski, L., Scholz, R., Wieters, I., Dahlke, C., Tolosa, E., et al. (2020). Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. Immunity *53*, 442-455.e4.

Sekine, T., Perez-Potti, A., Rivera-Ballesteros, O., Strålin, K., Gorin, J.-B., Olsson, A., Llewellyn-Lacey, S., Kamal, H., Bogdanovic, G., Muschiol, S., et al. (2020). Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. Cell *183*, 158-168.e14.

Shomuradova, A.S., Vagida, M.S., Sheetikov, S.A., Zornikova, K.V., Kiryukhin, D., Titov, A., Peshkova, I.O., Khmelevskaya, A., Dianov, D.V., Malasheva, M., et al. (2020). SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. Immunity *53*, 1245-1257.e5.

Snyder, T.M., Gittelman, R.M., Klinger, M., May, D.H., Osborne, E.J., Taniguchi, R., Zahid, H.J., Kaplan, I.M., Dines, J.N., Noakes, M.T., et al. (2020). Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels (Infectious Diseases (except HIV/AIDS)).

Swanson, P.A., Padilla, M., Hoyland, W., McGlinchey, K., Fields, P.A., Bibi, S., Faust, S.N., McDermott, A.B., Lambe, T., Pollard, A.J., et al. (2021). AZD1222/ChAdOx1 nCoV-19 vaccination induces a polyfunctional spike protein-specific Th1 response with a diverse TCR repertoire. Sci. Transl. Med. eabj7211.

Tarke, A., Sidney, J., Methot, N., Yu, E.D., Zhang, Y., Dan, J.M., Goodwin, B., Rubiro, P., Sutherland, A., Wang, E., et al. (2021a). Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. Cell Rep. Med. *2*, 100355.

Tarke, A., Sidney, J., Kidd, C.K., Dan, J.M., Ramirez, S.I., Yu, E.D., Mateus, J., da Silva Antunes, R., Moore, E., Rubiro, P., et al. (2021b). Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Rep. Med. *2*, 100204.

Wen, W., Su, W., Tang, H., Le, W., Zhang, X., Zheng, Y., Liu, X., Xie, L., Li, J., Ye, J., et al. (2020). Immune cell profiling of COVID-19 patients in the recovery stageby single-cell sequencing. Cell Discov. *6*, 31.

Xu, G., Qi, F., Li, H., Yang, Q., Wang, H., Wang, X., Liu, X., Zhao, J., Liao, X., Liu, Y., et al. (2020). The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing. Cell Discov. *6*, 73.

Zhang, J.-Y., Wang, X.-M., Xing, X., Xu, Z., Zhang, C., Song, J.-W., Fan, X., Xia, P., Fu, J.-L., Wang, S.-Y., et al. (2020). Single-cell landscape of immunological responses in patients with COVID-19. Nat. Immunol. *21*, 1107–1118.