bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tracking SARS-CoV-2 T cells with epitope-T-cell receptor recognition models

Pieter Meysman[1,2,✉], **Anna Postovskaya**[1,2], **Nicolas De Neuter**[1,2], **Benson Ogunjimi**[2,3,4,5], **and Kris Laukens**[1,2]

[1]ADREM data lab, University of Antwerp
[2]Antwerp Unit for Data Analysis and Computation in Immunology and Sequencing (AUDACIS), Antwerp
[3]Antwerp Center for Translational Immunology and Virology (ACTIV), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp
[4]Centre for Health Economics Research & Modeling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp
[5]Department of Paediatrics, Antwerp University Hospital

Much is still not understood about the human adaptive immune response to SARS-CoV-2, the causative agent of COVID-19. In this paper, we demonstrate the use of machine learning to classify SARS-CoV-2 epitope specific T-cell clonotypes in T-cell receptor (TCR) sequencing data. We apply these models to public TCR data and show how they can be used to study T-cell longitudinal profiles in COVID-19 patients to characterize how the adaptive immune system reacts to the SARS-CoV-2 virus. Our findings confirm prior knowledge that SARS-CoV-2 reactive T-cell diversity increases over the course of disease progression. However our results show a difference between those T cells that react to epitope unique to SARS-CoV-2, which show a more prominent increase, and those T cells that react to epitopes common to other coronaviruses, which begin at a higher baseline.

SARS-CoV-2 | COVID-19 | T-cell receptor | Machine learning
Correspondence: ***pieter.meysman@uantwerpen.be***

Introduction

The emergence of a novel coronavirus in 2019, termed SARS-CoV-2, has led to the most prominent global pandemic in recent history. Infection by SARS-CoV-2 manifests as COVID-19, a disease of which symptoms and severity greatly vary, and which has caused substantial loss of life all over the world. Characterizing the immune response against this novel virus has become a top priority, in the hope that novel insights can lead to new treatment plans or can aid in vaccine development.

Special attention is now being paid to T-cell response in particular, which was already determined an important factor in long-term immunity against coronaviruses during SARS and MERS outbreaks (1–3). Moreover, SARS-specific T cells are still found in individuals 17 years later and demonstrate robust cross-reactivity against SARS-CoV-2 suggesting the possibility of long-term protection for SARS-CoV-2 (4).

Current findings indicate that T cells play a key role in susceptibility to and severity of the ongoing COVID-19 pandemic as well. In particular, disease severity has been found to be linked with a sub-optimal or excessive T-cell response (5). Similarly, while most of the patients present with high antibody titers (6), frequently they do not provide necessary virus neutralization (7). Moreover, some studies have reported that around 30% of patients recover with very low levels of neutralizing antibodies or none at all (8). Most of the patients, however, developed CD8+ T cell response (9), with cells predominantly showing activation signals (10). Activated CD4+ T cells were also detected in the majority of patients, sometimes even reaching 100% in convalescent patients (9). Interestingly, 83% of patients were found to have CD4+ T cells specific to the SARS-CoV-2 spike protein in one study (11), and the magnitude of antibody response was correlated to the CD4+ T cell with the same protein specificity in another (9). Surprisingly, T cells specific to the SARS-CoV-2 epitopes were even found in a high proportion of unexposed individuals (4, 9, 11), which was later demonstrated to be due to cross-reactivity with common cold coronaviruses (12) and might explain pre-existing protection from the SARS-CoV-2 in some individuals. Altogether those findings support a key role for T cells in the immune defense against SARS-CoV-2 and are important considerations for vaccine development.

A multitude of computational approaches to discover T-cell epitopes for rational SARS-CoV-2 vaccine design have been performed using different perspectives: genetic similarity between SARS and SARS-CoV-2 (9, 13–15), previous knowledge about SARS immunogenic epitopes (13), de novo epitope prediction with respect to affinity to MHC estimated utilizing structural biology (16, 17) or machine learning (13, 14, 17), and MHC distribution across populations (14). Different approaches resulted in various numbers of potential targets in different groups. For instance, Lee et al. reported 28 SARS-CoV-2 peptides identical to those of SARS (13), and Fast et al. yielded 405 de novo candidate T-cell epitopes (17). For some predictions, experimental validation was later performed by different groups (18).

One key technique that facilitates insights into the makeup of an individual's T-cell repertoire is high-throughput T-cell receptor sequencing. Several TCR sequencing studies have now been performed in COVID19 patients and revealed some characteristics of the T-cell receptor repertoire during COVID-19. One study, for example, found clusters of T cells tied to disease severity, which predominantly comprised public clonotypes (19). Another group identified statistically enriched public TCR sequences from a large number of repertoires to distinguish SARS-CoV-2 positive from healthy individuals (20). In addition, several studies have determined SARS-CoV-2 specific TCRs and distinct CDR3 motifs down to the individual epitope level (18, 20). However, one key

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

downside of T-cell receptor sequencing is the high diversity of TCR sequences across individuals. Only a handful of TCR sequences can be found across different individuals. These so-called 'public' clones can therefore be tracked across individuals and stored in databases for reference. However, TCR repertoires consist mostly of individual-specific sequences, which due to their individual nature cannot be expected to be found in a database. Recently, computational methods have been developed to convert epitope-TCR pairing data to predictive models (21) that generalize the epitope-TCR specificity determinants. Such models can be used to screen TCR repertoires to find additional potential epitope-specific TCRs that are not contained in any database (22). These models are based on the concept that TCRs targeting the same epitope tend to have similar amino acid sequences (23). In this study, we create such prediction models for SARS-CoV-2 epitopes and apply them to track epitope-specificity over time.

Methods

Epitope-TCR data. A collection was established of experimentally validated TCR-epitope pairs by combining two primary sources:

- The VDJdb database, which contained tetramer-derived data from Shomuradova et al (18). Accessed on the 26th of May, 2020.
- The ImmuneCODE collection from Adaptive Technologies and Microsoft, which contained pairs derived through MIRA assay (20). Accessed on the 25th of June, 2020.

For all extracted pairs, several data curation steps were performed. All pairs matching more than one possible SARS-CoV-2 epitope were removed for training data. Only valid TCR sequences that could be matched to standard IMGT were kept. Where needed, a limit was placed on 5000 unique TCRs, which were then selected randomly.

Longitudinal TCR data. Public repertoire data was retrieved through the iReceptor gateway (24) in the week of the 13th of July, 2020. In particular, TCR data was extracted of those studies that had longitudinal tracking of COVID-19 patients, namely from Minervina et al. (25) and Schultheiß et al (19). Only those TCRs that occur at a frequency of at least 1 in 100 000 were retained, to compensate for the different sequencing depths between studies. Meta data was made uniform so that the time points are annotated by days after onset of symptoms.

Protein and sequence data. Protein sequence data for 119 Nidovirales species, which included the human and non-human SARS viruses along with other coronaviruses and single-strand RNA viruses, were downloaded from the Corona OMA Orthology Database (26). In this manner, the used protein amino acid sequences for SARS-CoV-2 corresponds to Genbank accession GCA_009858895.3, and the protein sequences for SARS-CoV to GCA_000864885.1. Epitopes were matched to all proteins for all 119 species with an exact match, as the degree of variation allowed in the epitope space while retaining TCR recognition is still an unsolved question. Matches across all species for each epitope were tallied, and the annotation for SARS-CoV-2 was retained. Sequence identity between proteins was established using a pairwise protein BLAST.

Model training and application. For the machine learning part of this study, we made use of the TCRex framework (22). We trained models for all epitopes that had more than 30 distinct TCRs. Only those models that had an AUC ROC higher than 0.7 and a AUC PR higher than 0.35 in a cross-validated setting were retained, as per the default TCRex criteria. The models were then applied to full TCR repertoires, where a match was defined as a probable epitope-specific TCR if the score is higher than 0.9 and the BPR is lower than 1e-4. For normalization, reported hits were divided by the unique TCR repertoire size.

Results

Recognition model performance. In total, 47 distinct epitope TCRex models could be trained for SARS-CoV-2. An overview of all models and their performance can be found in Table S1. This is almost as many models as are available for all non-SARS-CoV-2 epitopes combined (49), indicating the vast amount of data that has been generated in the past few months compared to what has been collected for all prior pathogens and diseases. 24 of these epitopes match the SARS-CoV-2 replicase protein coded by ORF1ab, 16 match the SARS-CoV-2 spike protein encoded by ORF2 and the final 7 are distributed across the remaining proteins. In addition, 19 of the epitopes are unique to SARS-CoV-2 in our data set of 119 Nidovirales species. As can been seen in figure 1, this does not seem to be evenly distributed across the protein of origin. 9 out of the 16 epitopes derived from the spike protein are unique to SARS-CoV-2, whereas only 6 out of 24 are unique for the ORF1ab replicase protein. As previously reported (27), the spike protein of SARS-CoV-2 shares 76% amino acid sequence identity to that of SARS-CoV. In contrast, the replicase protein has 86% sequence identity between SARS-CoV-2 and SARS-CoV. It is well known that viruses accumulate mutations to avoid potential immunogenic epitopes (28, 29), and the same process may be playing a role here.

As we are integrating models from different resources and diverse experimental methods, we wished to confirm if this data is comparable. Interestingly, one epitope has both tetramer (315 TCRs) and MIRA data (366 TCRs), namely YLQPRTFLL (YLQ). However in the case of the MIRA data, the TCRs were not uniquely assigned to this epitope, but all were assigned to the trio YLQPRTFL,YLQPRTFLL, and YYVGYLQPRTF. These 366 TCRs were thus excluded from the training data. Thus the tetramer-based YLQ model can be applied on the MIRA YLQ data as an independent model. In this manner, TCRex predicted 81 putative YLQ-reactive T-cells in the YLQ MIRA data out of 366 TCRs. Note that

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

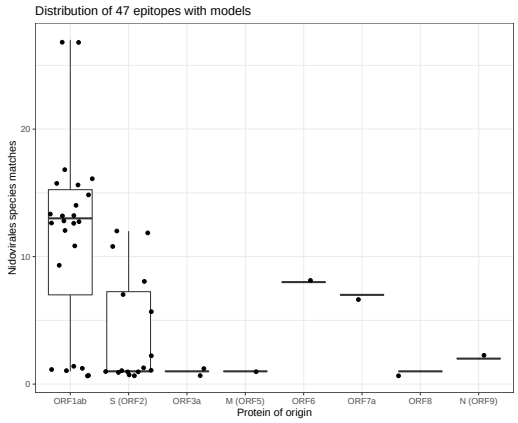

Fig. 1. Distribution of the 47 epitopes for which models could be created across SARS-CoV-2 proteins (x-axis) and 119 Nidovirales species.

only 35 TCRs matched between the two datasets based on CDR3 sequence (not accounting for V/J genes), showing that TCRex is able to extrapolate from found TCR patterns. This number of TCRex predictions was assigned an enrichment P-value of 6.44e-246 based on the built-in binomial test. No other epitopes present in TCRex (including the 49 non-SARS-CoV-2 models) were predicted to have a single TCR target within this data set. Thus the data is comparable and the models can be used without respect of their origin.

Longitudinal tracking. Once established, these models can be applied to any TCR repertoire data and thus can be used to study putative SARS-CoV-2 reactive T cells in the currently available COVID-19 data. A large TCR data set was made available by Schultheiß et al (19), which featured longitudinal samples from both patients with active disease and those that have recovered. As can be seen in Figure 2, the percentage of predicted SARS-CoV-2 reactive TCRs increases as time goes on after on-set of symptoms (Spearman rho = 0.36, P-value = 0.0022). This is both due to an expansion of distinct SARS-CoV-2 TCR sequences and a contraction of the remainder of the TCR repertoire. This matches findings from prior studies investigating T-cell immunity, which have observed that SARS-CoV-2 specific T-cell immunity mounts as disease progresses, alongside an absolute decrease in T-cell population size (30, 31). In addition, the increase is irrespective of final disease outcome.

As can be seen in Figure 3, there seems to be a difference between those TCRs that are unique to SARS-CoV-2 versus those that are not. The fraction of TCRs that are predicted to match unique SARS-CoV-2 epitopes shows a similar increase as was found for all epitopes (Spearman rho = 0.36, P-value = 0.0021). While that for the epitopes occurring across coronaviruses have a markedly lower increase signal that is not significant (Spearman rho = 0.23, P-value = 0.057). Indeed cross-reactive TCRs predicted to target epitopes not unique to SARS-CoV-2 start at a higher level, and only seem to gradually increase as the infection progresses. This may be in line with prior reports of existing cross-reactive T cells in uninfected patients (9, 11, 32).

The number of SARS-CoV-2 TCRs do decrease once the pa-

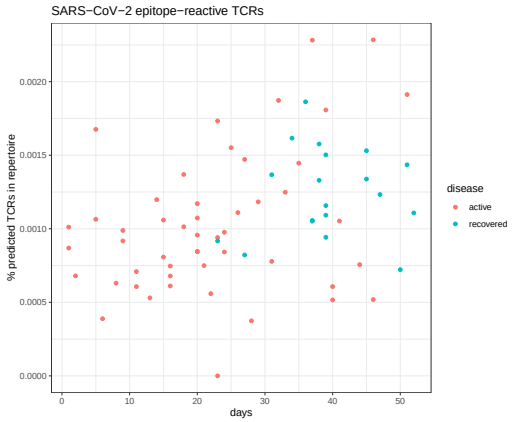

Fig. 2. SARS-CoV-2 predicted TCRs plotted against time on the data from Schultheiß et al.

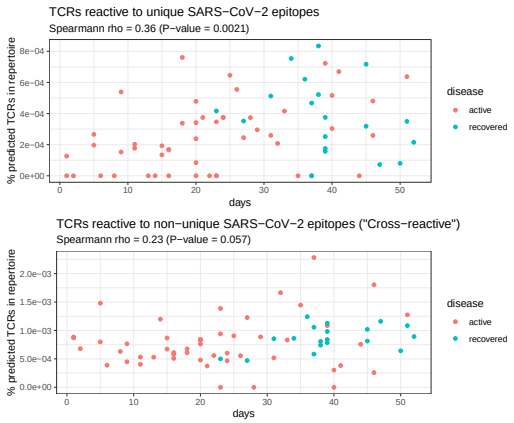

Fig. 3. Time profile of the frequency of SARS-CoV-2 predicted TCRs, on the data from Schultheiß et al. Top: predicted TCRs for epitopes unique to SARS-CoV-2. Bottom: predicted TCRs for epitopes not unique to SARS-CoV-2, which can be seen as 'cross-reactive' across species.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

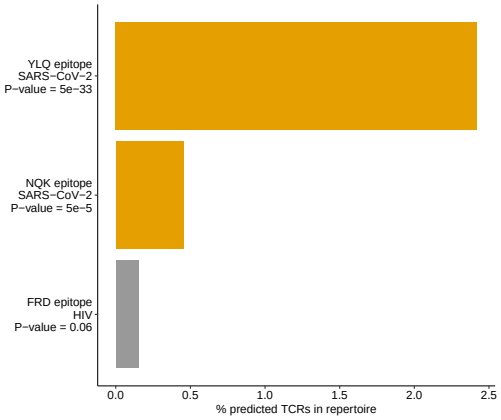

Fig. 4. Predicted TCRs in the contracted TCRs identified in donor M on the data from Minervina et al. The x-axis shows the frequency of predicted TCRs for each epitope. Yellow bar color denotes significant enrichment of epitope-reactive TCRs against the default TCR background according to TCRex criteria.

tient enters recovery. An example can be seen in Figure S1. This can also be seen based on the TCR data from Minervina et al., where both donors were sampled after symptoms had disappeared as per Figure S2. T cells contracted over time points within a single individual were considered as associated with SARS-CoV-2. A list of 661 and 372 contracted TCRs (one set for each donor) were identified using EdgeR in the original study and published as supplemental materials. At the time, no epitope-specific TCR data was available for SARS-CoV-2 and thus it could not be analysed in this manner. If we apply all TCRex models (both SARS-CoV-2 and other viruses), we find strong enrichment for two SARS-CoV-2 epitopes in one donor (donor M), as can be seen in Figure 4. The TCRs associated with the most prominent epitope, namely YLQPRTFLL, clearly decrease over time in donor M as seen in the original TCR data set as can been seen in figure S3. Note that the YLQ epitope originates from the spike protein and is unique to SARS-CoV-2, which matches the previous findings. The other donor had no such enriched epitopes of any origin, indicating that these TCRs might still be resulting from a not-included set of epitopes.

Conclusions

In this paper, we have shown that there is sufficient SARS-CoV-2 epitope-TCR data to create a large number of epitope-specific TCR recognition models. These models can be used to screen TCR data from various individuals to track their T-cell immunity. In addition, using such models on longitudinal data reveals a potential difference in temporal dynamics between T cells predicted to react against epitopes that are unique to SARS-CoV-2 and those that are shared among other coronaviruses.

ACKNOWLEDGEMENTS
We wish to thank the scientists who have made their data available, and doing so have made these studies possible.

Bibliography

1. William J. Liu, Min Zhao, Kefang Liu, Kun Xu, Gary Wong, Wenjie Tan, and George F. Gao. T-cell immunity of SARS-CoV: Implications for vaccine development against MERS-CoV, jan 2017. ISSN 18729096.
2. Hsueh-Ling Janice Oh, Samuel Ken-En Gan, Antonio Bertoletti, and Yee-Joo Tan. Understanding the T cell immune response in SARS coronavirus infection. *Emerging Microbes & Infections*, 1(1):1–6, jul 2012. ISSN 2222-1751. doi: 10.1038/emi.2012.26.
3. Li Tao Yang, Hui Peng, Zhao Ling Zhu, Gang Li, Zi Tong Huang, Zhi Xin Zhao, Richard A. Koup, Robert T. Bailer, and Chang You Wu. Long-lived effector/central memory T-cell responses to severe acute respiratory syndrome coronavirus (SARS-CoV) S antigen in recovered SARS patients. *Clinical Immunology*, 120(2):171–178, aug 2006. ISSN 15216616. doi: 10.1016/j.clim.2006.05.002.
4. Nina Le Bert, Anthony T. Tan, Kamini Kunasegaran, Christine Y.L. Tham, Morteza Hafezi, Adeline Chia, Melissa Hui Yen Chng, Meiyin Lin, Nicole Tan, Martin Linster, Wan Ni Chia, Mark I.Cheng Chen, Lin Fa Wang, Eng Eong Ooi, Shirin Kalimuddin, Paul Anantharajah Tambyah, Jenny Guek Hong Low, Yee Joo Tan, and Antonio Bertoletti. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*, 584(7821):457–462, aug 2020. ISSN 14764687. doi: 10.1038/s41586-020-2550-z.
5. Zeyu Chen and E. John Wherry. T cell responses in patients with COVID-19. *Nature Reviews Immunology*, pages 1–8, jul 2020. ISSN 1474-1733. doi: 10.1038/s41577-020-0402-6.
6. Baoqing Sun, Ying Feng, Xiaoneng Mo, Peiyan Zheng, Qian Wang, Pingchao Li, Ping Peng, Xiaoqing Liu, Zhilong Chen, Huimin Huang, Fan Zhang, Wenting Luo, Xuefeng Niu, Peiyu Hu, Longyu Wang, Hui Peng, Zhifeng Huang, Liqiang Feng, Feng Li, Fuchun Zhang, Fang Li, Nanshan Zhong, and Ling Chen. Kinetics of SARS-CoV-2 specific IgM and IgG responses in COVID-19 patients. *Emerging Microbes and Infections*, 9(1):940–948, jan 2020. ISSN 22221751. doi: 10.1080/22221751.2020.1762515.
7. Davide F Robbiani, Christian Gaebler, Frauke Muecksch, Julio C C Lorenzi, Alice Cho, Marianna Agudelo, Christopher O Barnes, Anna Gazumyan, Thomas Hagglof, Thiago Y Oliveira, Charlotte Viant, Arlene Hurley, Heinrich Hoffmann, Katrina G Millard, Rhonda G Kost, Melissa Cipolla, Kristie Gordon, Filippo Bianchini, Spencer T Chen, Victor Ramos, Roshni Patel, Juan Dizon, Pilar Mendoza, Harald Hartweger, Lilian Nogueira, Maggi Pack, Jill Horowitz, Fabian Schmidt, Yiska Weisblum, Eleftherios Michailidis, Alison W Ashbrook, Eric Waltari, John E Pak, Kathryn E Huey-Tubman, Nicholas Koranda, Pauline R Hoffman, Charles M Rice, Theodora Hatziioannou, Pamela J Bjorkman, Paul D Bieniasz, Marina Caskey, and Michel C Nussenzweig. Convergent Antibody Responses to SARS-CoV-2 Infection in Convalescent Individuals 1 2. *bioRxiv*, page 2020.05.13.092619, may 2020. doi: 10.1101/2020.05.13.092619.
8. Fan Wu, Aojie Wang, Mei Liu, Qimin Wang, Jun Chen, Shuai Xia, Yun Ling, Yuling Zhang, Jingna Xun, Lu Lu, Shibo Jiang, Hongzhou Lu, Yumei Wen, and Jinghe Huang. Neutralizing Antibody Responses to SARS-CoV-2 in a COVID-19 Recovered Patient Cohort and Their Implications. *SSRN Electronic Journal*, page 2020.03.30.20047365, apr 2020. ISSN 1556-5068. doi: 10.1101/2020.03.30.20047365.
9. Alba Grifoni, John Sidney, Yun Zhang, Richard H. Scheuermann, Bjoern Peters, and Alessandro Sette. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host and Microbe*, 27(4):671–680.e2, apr 2020. ISSN 19346069. doi: 10.1016/j.chom.2020.03.002.
10. Lucie Rodriguez, Pirkka T. Pekkarinen, Tadepally Lakshmikanth, Ziyang Tan, Camila Rosat Consiglio, Christian Pou, Yang Chen, Constantin Habimana Mugabo, Ngoc Anh Nguyen, Kirsten Nowlan, Tomas Strandin, Lev Levanov, Jaromir Mikes, Jun Wang, Anu Kantele, Jussi Hepojoki, Olli Vapalahti, Santtu Heinonen, Eliisa Kekäläinen, and Petter Brodin. Systems-Level Immunomonitoring from Acute to Recovery Phase of Severe COVID-19. *Cell Reports Medicine*, 1(5):100078, aug 2020. ISSN 26663791. doi: 10.1016/j.xcrm.2020.100078.
11. Julian Braun, Lucie Loyal, Marco Frentsch, Daniel Wendisch, Philipp Georg, Florian Kurth, Stefan Hippenstiel, Manuela Dingeldey, Beate Kruse, Florent Fauchere, Emre Baysal, Maike Mangold, Larissa Henze, Roland Lauster, Marcus Mall, Kirsten Beyer, Jobst Roehmel, Juergen Schmitz, Stefan Miltenyi, Marcel Mueller, Martin Witzenrath, Norbert Suttorp, Florian Kern, Ulf Reimer, Holger Wenschuh, Christian Drosten, Victor Corman, Claudia Giesecke-Thiel, Leif-Erik Sander, and Andreas Thiel. Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. *medRxiv*, page 2020.04.17.20061440, apr 2020. doi: 10.1101/2020.04.17.20061440.
12. Jose Mateus, Alba Grifoni, Alison Tarke, John Sidney, Sydney I. Ramirez, Jennifer M. Dan, Zoe C. Burger, Stephen A. Rawlings, Davey M. Smith, Elizabeth Phillips, Simon Mallal, Marshall Lammers, Paul Rubiro, Lorenzo Quiambao, Aaron Sutherland, Esther Dawen Yu, Ricardo da Silva Antunes, Jason Greenbaum, April Frazier, Alena J. Markmann, Lakshmanane Premkumar, Aravinda de Silva, Bjoern Peters, Shane Crotty, Alessandro Sette, and Daniela Weiskopf. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science*, page eabd3871, aug 2020. ISSN 0036-8075. doi: 10.1126/science.abd3871.
13. Chloe H. Lee, Hashem Koohy, Katalin Wilkinson, Raffaele de Palma, Maria Bonsack, and Angelika Riemer. In silico identification of vaccine targets for 2019-nCoV [version 2; peer review: 3 approved]. *F1000Research*, 9:1–10, apr 2020. ISSN 1759796X. doi: 10.12688/f1000research.22507.2.
14. Kazuma Kiyotani, Yujiro Toyoshima, Kensaku Nemoto, and Yusuke Nakamura. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *Journal of Human Genetics*, 65(7): 569–575, jul 2020. ISSN 1435232X. doi: 10.1038/s10038-020-0771-5.
15. Syed Faraz Ahmed, Ahmed A. Quadeer, and Matthew R. McKay. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses*, 12(3):254, feb 2020. ISSN 1999-4915. doi: 10.3390/v12030254.
16. Vargab Baruah and Sujoy Bose. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. *Journal of Medical Virology*, 92(5): 495–500, may 2020. ISSN 0146-6615. doi: 10.1002/jmv.25698.
17. Ethan Fast, Russ Altman, and Binbin Chen. Potential T-cell and B-cell Epitopes of 2019-nCoV. *bioRxiv*, page 2020.02.19.955484, mar 2020. doi: 10.1101/2020.02.19.955484.
18. Alina S Shomuradova, Murad S Vagida, Savely A Sheetikov, Ksenia V Zornikova, Dmitriy Dmitry Kiryukhin, Aleksei Titov, Iuliia O Peshkova, Alexandra Khmelevskaya, Dmitry V Dianov, Maria Malasheva, Anton Shmelev, Yana Serdyuk, Dmitry V Bagaev, Anastasia Pivnyuk, Dmitrii S Shcherbinin, Alexandra V Maleeva, Naina T Shakirova, Artem Pilunov, Dmitry B Malko, Ekaterina G Khamaganova, Bella Biderman, Alexander V Ivanov,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Mikhail Shugay, and Grigory A Efimov. SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T-cell receptors. *medrxiv*, pages 1–27, may 2020. doi: 10.1101/2020.05.20.20107813.

19. Christoph Schultheiß, Lisa Paschold, Donjete Simnica, Malte Mohme, Edith Willscher, Lisa von Wenserski, Rebekka Scholz, Imke Wieters, Christine Dahlke, Eva Tolosa, Daniel G. Sedding, Sandra Ciesek, Marylyn Addo, and Mascha Binder. Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity*, jun 2020. ISSN 10974180. doi: 10.1016/j.immuni.2020.06.024.
20. Thomas M Snyder, Rachel M Gittelman, Mark Klinger, Damon H May, Edward J Osborne, Ruth Taniguchi, H Jabran Zahid, Ian M Kaplan, Jennifer N Dines, Matthew T Noakes, Ravi Pandya, Xiaoyu Chen, Summer Elasady, Emily Svejnoha, Peter Ebert, Mitchell W Pesesky, Patricia De Almeida, Hope O'donnell, Quinn Degottardi, Gladys Keitany, Jennifer Lu, Allen Vong, Rebecca Elyanow, Paul Fields, Julia Greissl, Lance Baldo, Simona Semprini, Claudio Cerchione, Massimiliano Mazza, Ottavia M Delmonte, Kerry Dobbs, Gonzalo Carreño-Tarragona, Santiago Barrio, Luisa Imberti, Alessandra Sottini, Eugenia Quiros-Roldan, Camillo Rossi, Andrea Biondi, Laura Rachele Bettini, Mariella D'angio, Paolo Bonfanti, Miranda F Tompkins, Camille Alba, Clifton Dalgard, Vittorio Sambri, Giovanni Martinelli, Jason D Goldman, James R Heath, Helen C Su, Luigi D Notarangelo, Joaquin Martinez-Lopez, Jonathan M Carlson, and Harlan S Robins. Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. *medRxiv*, page 2020.07.31.20165647, aug 2020. doi: 10.1101/2020.07.31.20165647.
21. Nicolas De Neuter, Wout Bittremieux, Charlie Beirnaert, Bart Cuypers, Aida Mrzic, Pieter Moris, Arvid Suls, Viggo Van Tendeloo, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics*, 70(3):159–168, mar 2018. ISSN 14321211. doi: 10.1007/s00251-017-1023-5.
22. Sofie Gielis, Pieter Moris, Wout Bittremieux, Nicolas De Neuter, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Frontiers in Immunology*, 10:2820, nov 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.02820.
23. P Meysman, N De Neuter, S Gielis, D Bui Thi . . . , and Undefined 2019. On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics*, 35 (9):1461–1468, 2019.
24. Brian D. Corrie, Nishanth Marthandan, Bojan Zimonja, Jerome Jaglale, Yang Zhou, Emily Barr, Nicole Knoetze, Frances M.W. Breden, Scott Christley, Jamie K. Scott, Lindsay G. Cowell, and Felix Breden. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories, jul 2018. ISSN 1600065X.
25. Anastasia A. Minervina, Ekaterina A. Komech, Aleksei Titov, Meriem Bensouda Koraichi, Elisa Rosati, Ilgar Z. Mamedov, Andre Franke, Grigory A. Efimov, Dmitriy M. Chudakov, Thierry Mora, Aleksandra M. Walczak, Yuri B. Lebedev, and Mikhail V. Pogorelyy. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19 infection. *arXiv*, may 2020.
26. Adrian M. Altenhoff, Natasha M. Glover, Clément Marie Train, Klara Kaleb, Alex Warwick Vesztrocy, David Dylus, Tarcisio M. De Farias, Karina Zile, Charles Stevenson, Jiao Long, Henning Redestig, Gaston H. Gonnet, and Christophe Dessimoz. The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, 46(D1):D477–D485, jan 2018. ISSN 13624962. doi: 10.1093/nar/gkx1019.
27. Javier A. Jaimes, Nicole M. André, Joshua S. Chappie, Jean K. Millet, and Gary R. Whittaker. Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. *Journal of Molecular Biology*, 432(10):3309–3325, may 2020. ISSN 10898638. doi: 10.1016/j.jmb.2020.04.009.
28. Tal Vider-Shalit, Ronit Sarid, Kobi Maman, Lea Tsaban, Ran Levi, and Yoram Louzoun. Viruses selectively mutate their CD8+ T-cell epitopes–a large-scale immunomic analysis. *Bioinformatics (Oxford, England)*, 25(12):i39–44, jun 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp221.
29. P. Meysman, D. Fedorov, V. Van Tendeloo, B. Ogunjimi, and K. Laukens. Immunological evasion of immediate-early varicella zoster virus proteins. *Immunogenetics*, 68(6-7), 2016. ISSN 14321211. doi: 10.1007/s00251-016-0911-4.
30. Fan Wang, Jiayan Nie, Haizhou Wang, Qiu Zhao, Yong Xiong, Liping Deng, Shihui Song, Zhiyong Ma, Pingzheng Mo, and Yongxi Zhang. Characteristics of Peripheral Lymphocyte Subset Alteration in COVID-19 Pneumonia. *The Journal of infectious diseases*, 221(11): 1762–1769, may 2020. ISSN 15376613. doi: 10.1093/infdis/jiaa150.
31. Jing Liu, Sumeng Li, Jia Liu, Boyun Liang, Xiaobei Wang, Hua Wang, Wei Li, Qiaoxia Tong, Jianhua Yi, Lei Zhao, Lijuan Xiong, Chunxia Guo, Jin Tian, Jinzhuo Luo, Jinghong Yao, Ran Pang, Hui Shen, Cheng Peng, Ting Liu, Qian Zhang, Jun Wu, Ling Xu, Sihong Lu, Baoju Wang, Zhihong Weng, Chunrong Han, Huabing Zhu, Ruxia Zhou, Helong Zhou, Xiliu Chen, Pian Ye, Bin Zhu, Lu Wang, Wenqing Zhou, Shengsong He, Yongwen He, Shenghua Jie, Ping Wei, Jianao Zhang, Yinping Lu, Weixian Wang, Li Zhang, Ling Li, Fengqin Zhou, Jun Wang, Ulf Dittmer, Mengji Lu, Yu Hu, Dongliang Yang, and Xin Zheng. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine*, 55, may 2020. ISSN 23523964. doi: 10.1016/j.ebiom.2020.102763.
32. Ulrik Stervbo, Sven Rahmann, Toralf Roch, Timm H Westhof, and Nina Babel. SARS-CoV-2 reactive T cells in uninfected individuals are likely expanded by beta-coronaviruses. *bioRxiv*, page 2020.07.01.182741, jul 2020. doi: 10.1101/2020.07.01.182741.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Note 1: SARS-CoV-2 epitope recognition models statistics

Epitope	Protein	No. training TCRs	Balanced accuracy	AUC ROC	AUC PR
ALSKGVHFV	ORF3a	129	0.53 ± 0.01	0.79 ± 0.03	0.4 ± 0.09
EILDITPCSF	S (ORF2)	68	0.63 ± 0.03	0.73 ± 0.06	0.46 ± 0.09
FADDLNQLTGY	ORF1ab	70	0.61 ± 0.04	0.7 ± 0.05	0.44 ± 0.11
FLNGSCGSV	ORF1ab	2332	0.61 ± 0.0	0.86 ± 0.01	0.56 ± 0.02
FLNRFTTTL	ORF1ab	104	0.54 ± 0.03	0.72 ± 0.07	0.38 ± 0.12
FLPRVFSAV	ORF1ab	773	0.59 ± 0.02	0.84 ± 0.01	0.51 ± 0.03
FPPTSFGPL	ORF1ab	621	0.69 ± 0.01	0.85 ± 0.01	0.61 ± 0.01
FTISVTTEIL	S (ORF2)	159	0.55 ± 0.03	0.73 ± 0.03	0.37 ± 0.04
FVDGVPFVV	ORF1ab	2420	0.54 ± 0.0	0.77 ± 0.01	0.37 ± 0.01
HLVDFQVTI	ORF6	69	0.61 ± 0.04	0.75 ± 0.06	0.44 ± 0.09
HTTDPSFLGRY	ORF1ab	5000	0.7 ± 0.01	0.9 ± 0.01	0.7 ± 0.02
ILGLPTQTV	ORF1ab	198	0.68 ± 0.02	0.83 ± 0.05	0.6 ± 0.08
ILHCANFNV	ORF1ab	185	0.61 ± 0.02	0.84 ± 0.03	0.52 ± 0.05
IPIQASLPF	ORF3a	104	0.51 ± 0.02	0.74 ± 0.09	0.35 ± 0.11
IQYIDIGNY	ORF8	149	0.56 ± 0.03	0.82 ± 0.04	0.47 ± 0.06
ITEEVGHTDLMAAY	ORF1ab	156	0.57 ± 0.02	0.78 ± 0.02	0.49 ± 0.04
IVDTVSALV	ORF1ab	36	0.51 ± 0.03	0.76 ± 0.12	0.36 ± 0.18
KAYNVTQAF	N (ORF9)	708	0.61 ± 0.02	0.83 ± 0.01	0.57 ± 0.03
KEIDRLNEV	S (ORF2)	57	0.58 ± 0.02	0.7 ± 0.08	0.43 ± 0.1
KLPDDFTGCV	S (ORF2)	1160	0.59 ± 0.01	0.84 ± 0.01	0.54 ± 0.02
KLSYGIATV	ORF1ab	2149	0.58 ± 0.0	0.84 ± 0.01	0.51 ± 0.01
KLWAQCVQL	ORF1ab	266	0.55 ± 0.01	0.8 ± 0.02	0.46 ± 0.05
KPLEFGATSAAL	ORF1ab	344	0.59 ± 0.02	0.8 ± 0.04	0.49 ± 0.07
KTSVDCTMYI	S (ORF2)	70	0.7 ± 0.05	0.89 ± 0.05	0.72 ± 0.08
LEPLVDLPI	S (ORF2)	367	0.54 ± 0.01	0.76 ± 0.02	0.35 ± 0.05
LLFNKVTLA	S (ORF2)	39	0.51 ± 0.03	0.76 ± 0.04	0.37 ± 0.1
LPPAYTNSF	S (ORF2)	127	0.53 ± 0.02	0.82 ± 0.03	0.45 ± 0.06
MPASWVMRI	ORF1ab	477	0.62 ± 0.02	0.82 ± 0.01	0.53 ± 0.01
NLDSKVGGNY	S (ORF2)	45	0.73 ± 0.06	0.96 ± 0.03	0.85 ± 0.05
NLNESLIDL	S (ORF2)	132	0.59 ± 0.03	0.74 ± 0.02	0.42 ± 0.05
NQKLIANQF	S (ORF2)	51	0.76 ± 0.02	0.95 ± 0.02	0.78 ± 0.07
QECVRGTTVL	ORF7a	147	0.71 ± 0.02	0.85 ± 0.02	0.66 ± 0.04
RQLLFVVEV	ORF1ab	841	0.57 ± 0.01	0.83 ± 0.01	0.46 ± 0.02
SEISMDNSPNL	ORF1ab	95	0.52 ± 0.01	0.79 ± 0.05	0.45 ± 0.13
SEPVLKGVKL	S (ORF2)	80	0.57 ± 0.02	0.79 ± 0.05	0.38 ± 0.1
SEVGPEHSLAEY	ORF1ab	250	0.54 ± 0.01	0.79 ± 0.01	0.42 ± 0.05
SSNVANYQK	ORF1ab	74	0.61 ± 0.04	0.77 ± 0.04	0.45 ± 0.08
TLDSKTQSL	S (ORF2)	107	0.83 ± 0.06	0.97 ± 0.02	0.88 ± 0.06
TLIGDCATV	ORF1ab	467	0.55 ± 0.02	0.73 ± 0.03	0.36 ± 0.04
TLVPQEHYV	ORF1ab	154	0.53 ± 0.03	0.72 ± 0.02	0.42 ± 0.07
TPINLVRDL	S (ORF2)	249	0.64 ± 0.02	0.81 ± 0.03	0.53 ± 0.04
VLWAHGFEL	ORF1ab	695	0.65 ± 0.02	0.85 ± 0.02	0.6 ± 0.04
WICLLQFAY	M (ORF5)	515	0.64 ± 0.02	0.81 ± 0.02	0.56 ± 0.06
YFPLQSYGF	S (ORF2)	357	0.53 ± 0.01	0.79 ± 0.03	0.4 ± 0.04
YLDAYNMMI	ORF1ab	197	0.62 ± 0.03	0.74 ± 0.03	0.42 ± 0.05
YLNTLTLAV	ORF1ab	390	0.58 ± 0.02	0.8 ± 0.01	0.46 ± 0.04
YLQPRTFLL	S (ORF2)	315	0.91 ± 0.01	0.97 ± 0.01	0.92 ± 0.01

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

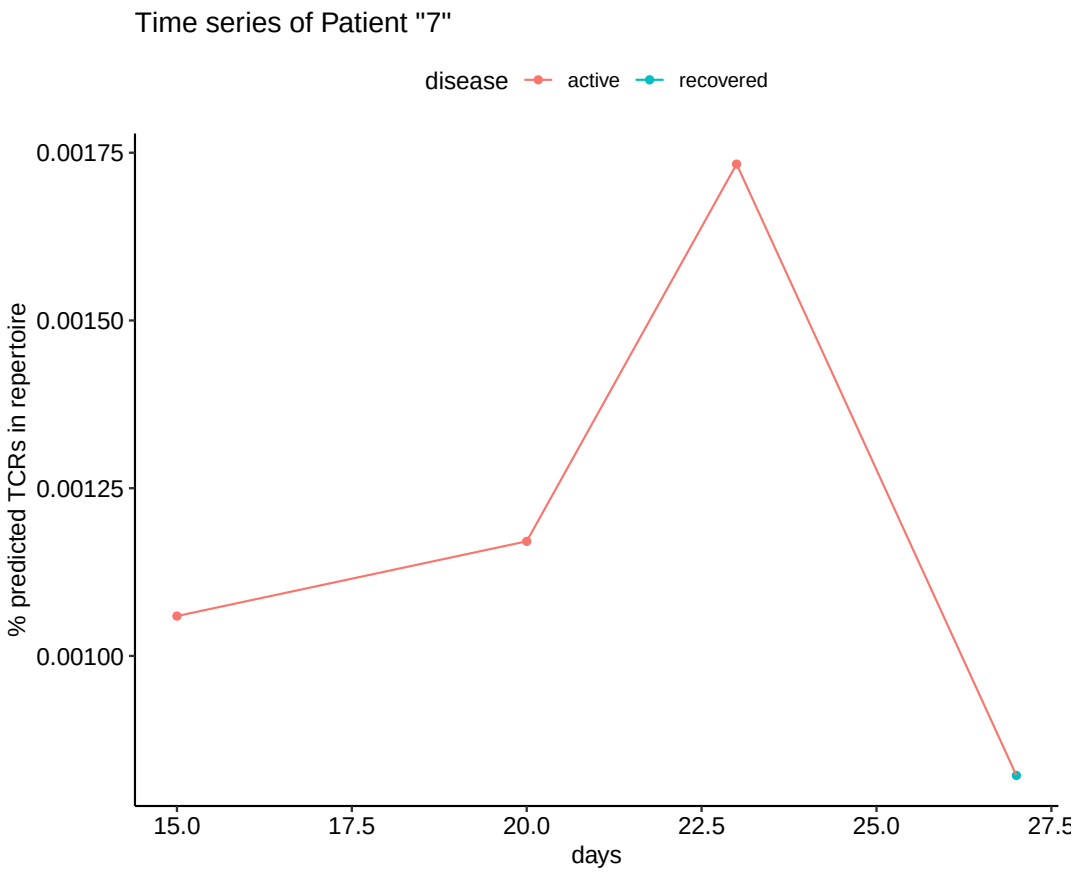

Fig. S1. SARS-CoV-2 predicted TCRs time profile of Patient "7".

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

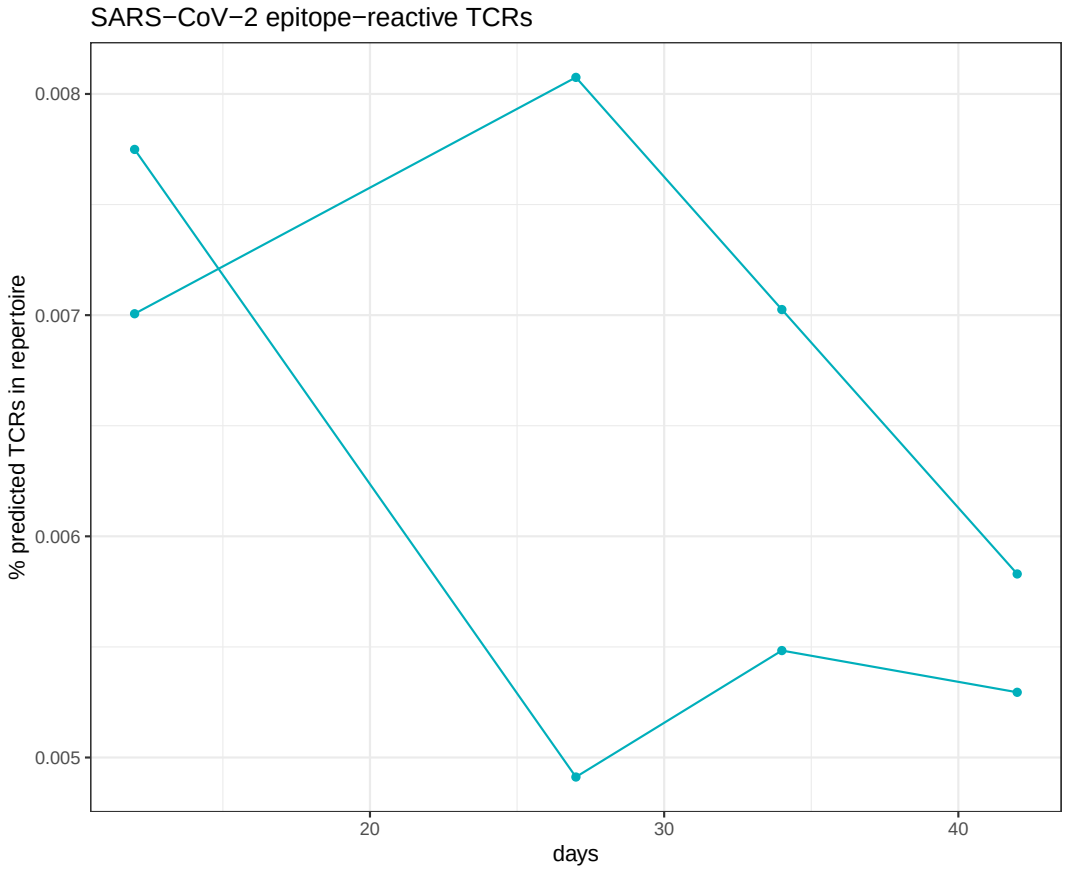

Fig. S2. SARS-CoV-2 predicted TCRs plotted against time on the data from Minervina et al.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.09.289355; this version posted September 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

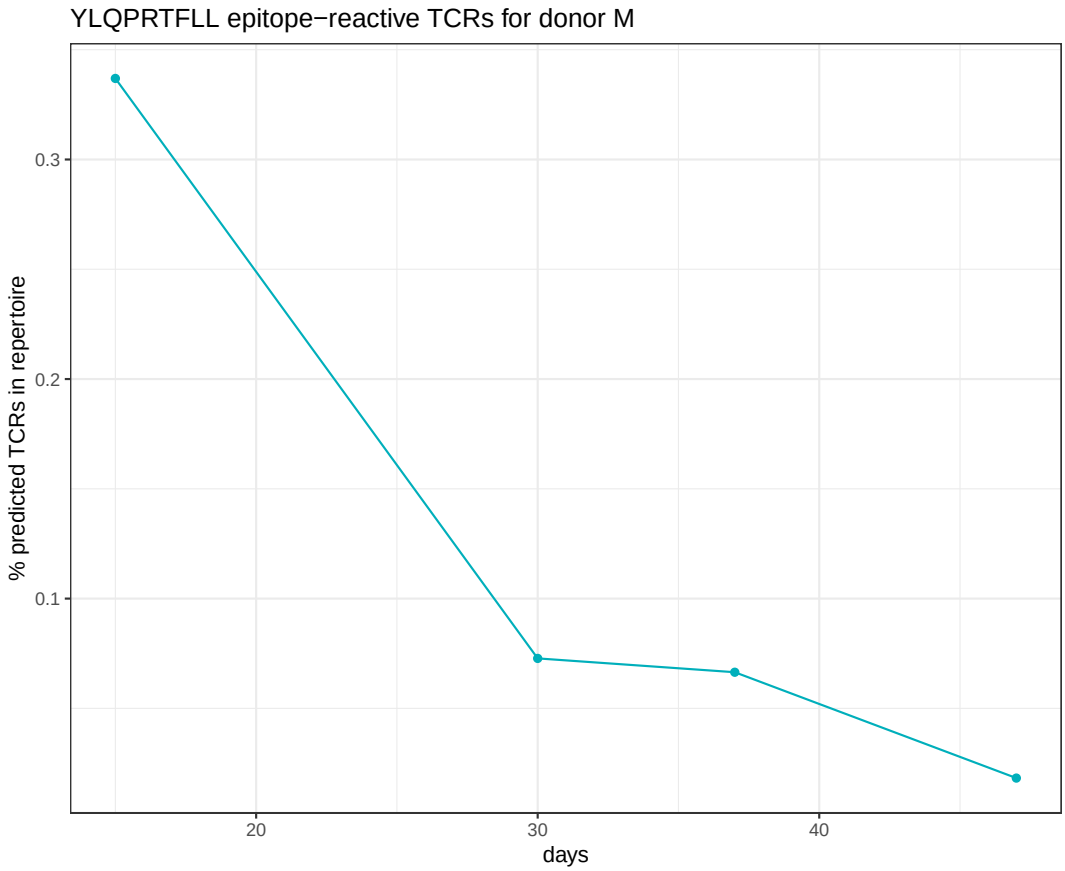

Fig. S3. YLQ epitope predicted TCRs from donor M over time on the data from Minervina et al. Samples at day 15, 30, 37, were all found to have a sufficient high level of YLQ reactive TCRs to be marked as significantly enriched by TCRex.