VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2

To the Editor — Here, we report the VDJdb database (https://vdjdb.cdr3.net) update prepared between 2019 and 2022, marked by the emergence of SARS-CoV-2, the causative agent of COVID-19.

In 2016, we started a community effort to gather and curate publicly available sequence data acquired from T cell receptor (TCRs) with defined antigen specificities, as well as communicated datasets from our colleagues, by developing the VDJdb database, which has since been extended with a web interface that allows batch querying of adaptive immune receptor repertoire sequencing (AIRR-seq) datasets and the identification of TCR sequence motifs linked with specific epitopes[1].

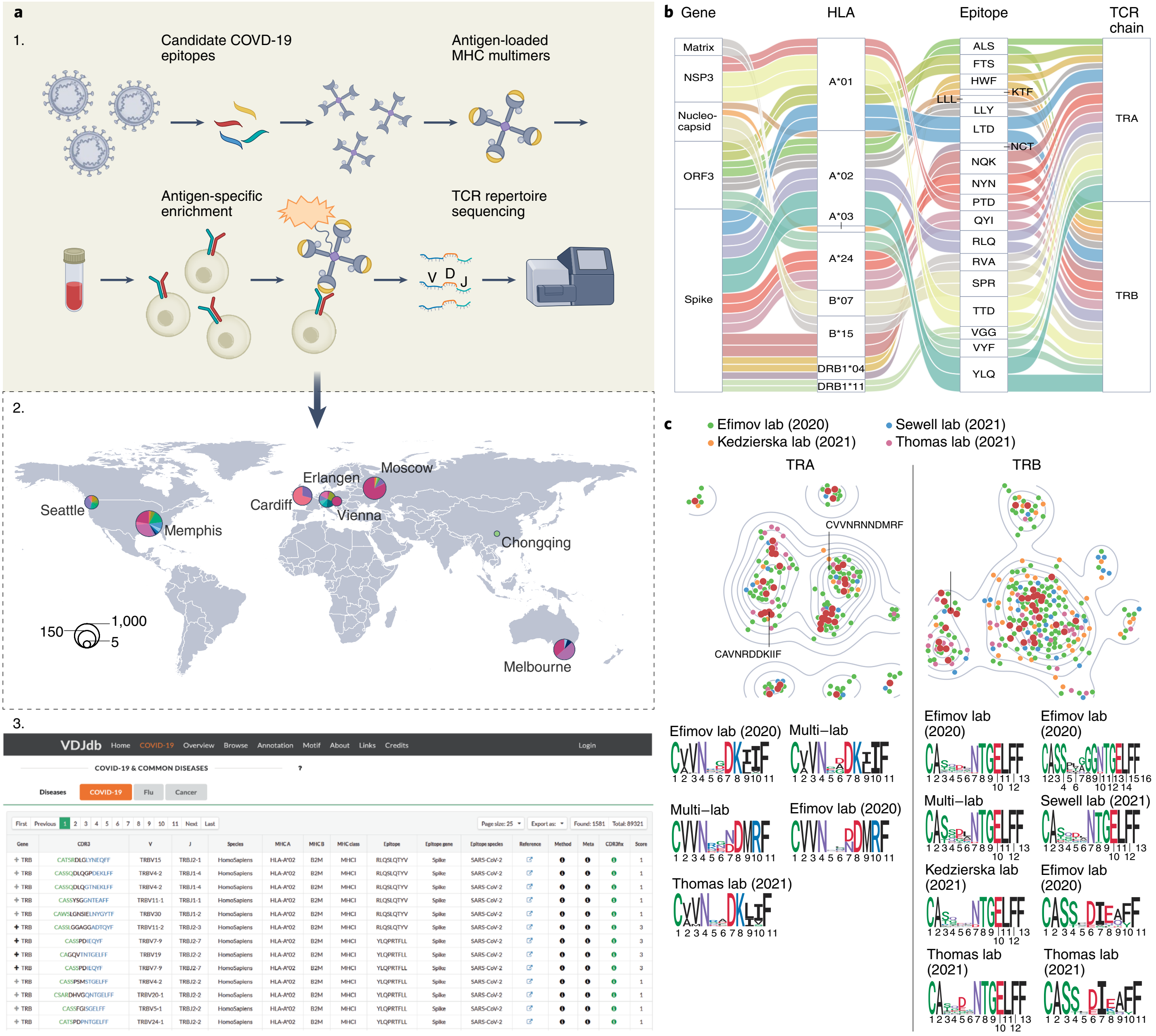
In the current pandemic era, a large majority of recent T cell repertoire profiling and antigen-specificity studies have focused on TCR variants that target the SARS-CoV-2 coronavirus[2–4]. As a consequence, millions of TCR sequences have now been isolated from donors with COVID-19. To complement these efforts, in the latest release of VDJdb, we incorporated TCR specificity data from various studies of COVID-19. We collected data from an international network of laboratories focused on assaying antigen-specific T cell responses in COVID-19 (Fig. 1a). Data acquired from multiple laboratories across the world feature over 3,000 TCR α and β chain sequences recognizing dozens of SARS-CoV-2 epitopes. These analyses revealed a set of reproducible TCR motifs that could find utility in large-scale clinical and experimental studies focused on COVID-19. We showed consistency and reproducibility of TCR specificity data across laboratories. Inferred TCR motifs will facilitate the tracking SARS-CoV-2-specific T cells and the discovery of immune signatures associated with protection against COVID-19. T cell antigen specificity is encoded by somatically rearranged TCRs. Current techniques allow the comprehensive profiling of TCR repertoires via high-throughput sequencing, which is compatible with various methods for elucidating the antigen specificity of T cell populations[5].

The first set of TCR repertoires with known specificity for SARS-CoV-2 epitopes was acquired from the Efimov laboratory[4]. This work prioritized the HLA-A*02-restricted YLQ and RLQ epitopes, producing 573 VDJdb records (unpaired TCR α and β chains), which were subsequently detected in other studies and served as a template for the first SARS-CoV-2-specific TCR–peptide–MHC crystal structures[6]. This submission was followed by a number of studies from different laboratories performed in 2021. One dataset reported multiple TCR sequences specific for SARS-CoV-2 epitopes restricted by HLA-A*24[7], a prominent HLA class I allotype among indigenous Asian populations. A report from the Kedzierska laboratory complemented these data with the addition of TCR sequences specific for SARS-CoV-2 epitopes restricted by HLA-A*02, HLA-A*24 and HLA-B*07[3]. A large set of paired TCRαβ sequences specific for a range of SARS-CoV-2 epitopes was acquired from the Thomas laboratory[8]. Smaller datasets were also imported from other published works and private communications (all listed in the issue section of the VDJdb github repository), including one notable study that reported TCR sequences specific for SARS-CoV-2 epitopes restricted by HLA class II allotypes[9]. In total, the current VDJdb release features 3,187 unique TCR specificity records spanning 46 distinct SARS-CoV-2 epitopes (Fig. 1b and Supplementary Table 1).

An important test of consistency for any biological dataset is independent reproducibility, and TCR repertoire sequencing in particular is prone to methodological and operator-dependent biases. To explore potential biases in the SARS-CoV-2-related VDJdb dataset, we performed a comparative analysis of TCR α and β chain specificity records for the most widely studied epitope, YLQ-HLA-A*02. No preferential clustering of these specificity records was observed across laboratories (Fig. 1c, top), while the overall structure of the TCR similarity map was preserved, suggesting that different laboratories sampled uniformly from the same space of epitope-specific TCR sequences.

Conversely, the independently generated data validated a set of TCR complementarity-determining region 3 (CDR3) sequences, which clustered as clearly defined motifs across different laboratories (Fig. 1c). Of note, the most commonly obtained CDR3 sequences were used successfully in crystallographic studies

Fig. 1 | Overview of COVID-19 data compendium stored in VDJdb. a, General pipeline used to acquire and store COVID-19 TCR specificity data. SARS-CoV-2 epitopes of interest are selected and used to construct MHC multimers, which are in turn used to enrich T cells and select T cells specific to a given epitope; those T cells are then subjected to a conventional TCR repertoire sequencing procedure (part 1). The data on TCR receptor sequences and their cognate epitopes is acquired independently by proficient laboratories around the globe; pie chart sizes reflect the number of TCR specificity records, with chart colors representing distinct epitopes (part 2). Data is processed, curated and stored in the VDJdb, which provides means to browse the COVID-19 compendium and annotate novel TCR sequences of unknown specificity (part 3). Maps are adapted (see https://github.com/antigenomics/vdjdb-db/blob/master/summary/vdjdb_summary.Rmd for code) from open-source R package "maps" released under GPL-2 license (https://CRAN.R-project.org/package=maps), copyright 2015–2022 VDJdb Developers and reproduced with permission of VDJdb Developers. **b**, Numbers of TCR specificity records for SARS-CoV-2 epitopes presented by various HLAs. Correspondence is shown using an alluvial plot with bands colored by epitopes. First three letters are used to code epitopes; only epitopes with ≥10 records are shown; band widths represent log-scaled number of records. **c**, Comparing TCR repertoires specific for the HLA-A*02-restricted YLQ epitope from SARS-CoV-2 obtained by different laboratories using sequence similarity map, with each dot representing a unique CDR3 sequence (top). Dot locations are based on CDR3 sequence similarity graphs generated using the TCRNET algorithm (see Supplementary Methods). Each dot is colored according to the parental dataset (key). Large red dots represent CDR3 sequences that were identified in multiple datasets. Left, TCR α chains; right, TCR β chains. Labels highlight TCRs that were successfully used to refold TCR–peptide–MHC complexes[6]. Sequence motif logos for clusters from the similarity map are shown below. Two recurring motifs each, CVVNXXDKIIF and CVVNXXDDMRF for TCRα and CAS-NTGELFF and CASSXDIEAFF for TCRβ, were shared among datasets ("Multi-lab" means shared across all laboratories).



VDJdb Home COVID-19 Overview Browse Annotation Motif About Links Credits Login

COVID-19 & COMMON DISEASES ?

Diseases COVID-19 Flu Cancer

First Previous 1 2 3 4 5 6 7 8 9 10 11 Next Last

Page size: 25 ▾ Export as: ▾ Found: 1581 Total: 89321

Gene	CDR3	V	J	Species	MHC A	MHC B	MHC class	Epitope	Epitope gene	Epitope species	Reference	Method	Meta	CDR3fix	Score
+ TRB	CATSRDLGLYNEQFF	TRBV15	TRBJ2-1	HomoSapiens	HLA-A*02	B2M	MHCI	RLQSLQTYV	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CASSQDLQGPDEKLFF	TRBV4-2	TRBJ1-4	HomoSapiens	HLA-A*02	B2M	MHCI	RLQSLQTYV	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CASSQDLQGTNEKLFF	TRBV4-2	TRBJ1-4	HomoSapiens	HLA-A*02	B2M	MHCI	RLQSLQTYV	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CASSYSGGNTEAFF	TRBV11-1	TRBJ1-1	HomoSapiens	HLA-A*02	B2M	MHCI	RLQSLQTYV	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CAWSLGNSIELNYGYTF	TRBV30	TRBJ1-2	HomoSapiens	HLA-A*02	B2M	MHCI	RLQSLQTYV	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CASSLGGAGGADTQYF	TRBV11-2	TRBJ2-3	HomoSapiens	HLA-A*02	B2M	MHCI	RLQSLQTYV	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	3
+ TRB	CASSPDIEQYF	TRBV7-9	TRBJ2-7	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	3
+ TRB	CAGQVTNTGELFF	TRBV19	TRBJ2-2	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	3
+ TRB	CASSPDIEQYF	TRBV7-9	TRBJ2-7	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	3
+ TRB	CASSPSMSTGELFF	TRBV4-2	TRBJ2-2	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CSARDHVGQNTGELFF	TRBV20-1	TRBJ2-2	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CASSFGISGELFF	TRBV5-1	TRBJ2-2	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1
+ TRB	CATSPDPNTGELFF	TRBV24-1	TRBJ2-2	HomoSapiens	HLA-A*02	B2M	MHCI	YLQPRTFLL	Spike	SARS-CoV-2	↗	ⓘ	ⓘ	ⓘ	1

to generate ternary structures[6], providing new insights into the molecular mechanisms that underpin TCR recognition of the YLQ epitope in complex with HLA-A*02.

Imprints of common infections can be detected in TCR repertoire sequencing datasets[10], which in turn can be used to predict immune responses and stratify patients with COVID-19[5]. VDJdb has been used successfully in the past for similar purposes and currently serves as a benchmark standard for testing TCR-specificity prediction algorithms[2]. In this work we demonstrated that the COVID-19 TCR-specificity compendium is unaffected by inter-laboratory biases and thus can be employed as a reference in TCR repertoire annotation. These precedents suggest that VDJdb can be used in the future to build classifiers trained to identify biologically relevant T cell responses in patients with COVID-19. Overall, we anticipate that the present release will enhance the versatility of VDJdb in the pandemic era, supporting the development of more effective vaccines and addressing future challenges associated with viral evolution and the emergence of new pathogens beyond SARS-CoV-2.

Data availability

All code and data are available at https://github.com/antigenomics/vdjdb-db, https://github.com/antigenomics/vdjdb-motifs and https://github.com/antigenomics/vdjdb-web, released under open-source Apache 2.0 and CC BY-ND 4.0 licenses. ❐

Mikhail Goncharov[1,2,13], Dmitry Bagaev[1,3,13], Dmitrii Shcherbinin[4], Ivan Zvyagin[1,4], Dmitry Bolotin[1,4], Paul G. Thomas[5], Anastasia A. Minervina[5], Mikhail V. Pogorelyy[5], Kristin Ladell[6], James E. McLaren[6], David A. Price[6,7], Thi H. O. Nguyen[8], Louise C. Rowntree[8], E. Bridie Clemens[8], Katherine Kedzierska[8], Garry Dolton[9], Cristina Rafael Rius[9], Andrew Sewell[7,9], Jerome Samir[10], Fabio Luciani[10], Ksenia V. Zornikova[11,12], Alexandra A. Khmelevskaya[11], Saveliy A. Sheetikov[11,12], Grigory A. Efimov[11], Dmitry Chudakov[1,2,4] and Mikhail Shugay[1,4] ✉

[1]Institute of Bioorganic Chemistry of Russian Academy of Sciences, Moscow, Russia. [2]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia. [3]Signal Processing Group, Eindhoven University of Technology, Eindhoven, the Netherlands. [4]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Translational Medicine, Pirogov Russian National Research Medical University, Moscow, Russia. [5]Department of Immunology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [6]Division of Infection and Immunity, Cardiff University School of Medicine, Cardiff, UK. [7]Systems Immunity Research Institute, Cardiff University School of Medicine, Cardiff, UK. [8]Department of Microbiology and Immunology, University of Melbourne, Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia. [9]T-Cell Modulation Group, Division of Infection and Immunity, Cardiff University School of Medicine, Cardiff, UK. [10]Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia. [11]National Research Center for Hematology, Moscow, Russia. [12]Biological Faculty, Lomonosov Moscow State University, Moscow, Russia. [13]These authors contributed equally: Mikhail Goncharov, Dmitry Bagaev.

✉*e-mail: mikhail.shugay@gmail.com*

Published online: 15 August 2022
https://doi.org/10.1038/s41592-022-01578-0

References

1. Dolton, G. et al. *Front. Immunol.* **9**, 1378 (2018).
2. Nguyen, T. H. O. et al. *Immunity* **54**, 1066–1082.e5 (2021).
3. Shomuradova, A. S. et al. *Immunity* **53**, 1245–1257.e5 (2020).
4. Shoukat, M. S. et al. *Cell Rep. Med.* **2**, 100192 (2021).
5. Bagaev, D. V. et al. *Nucleic Acids Res.* **48**, D1057–D1062 (2020).
6. Chaurasia, P. et al. *J. Biol. Chem.* **297**, 101065 (2021).
7. Rowntree, L. C. et al. *Immunol. Cell Biol.* https://doi.org/10.1111/imcb.12482 (2021).
8. Minervina, A. A. et al. *Nat. Immunol.* **23**, 781–790 (2022).
9. Verhagen, J. et al. *Clin. Exp. Immunol.* **205**, 363–378 (2021).
10. Pogorelyy, M. V. et al. *Genome Med.* **10**, 68 (2018).

Acknowledgements

This work was supported by a grant from the Ministry of Science and Higher Education of the Russian Federation (075-15-2019-1789). Additional funds were provided by the National Health and Medical Research Council (NHMRC; Australia) via a Leadership Investigator Grant (no. 1173871 to K.K.), the Research Grants Council of the Hong Kong Special Administrative Region, China (no. T11-712/19-N to K.K.) and the Medical Research Future Fund (Australia; no. 2005544 to K.K.). T.H.O.N. was supported by an NHMRC Emerging Leadership Level 1 Investigator Grant (no. 1194036). E.B.C. was supported by an NHMRC Peter Doherty Fellowship (no. 1091516). D.A.P. was supported by a Wellcome Trust Senior Investigator Award (UK; 100326/Z/12/Z). G.A.E. was supported by Russian Science Foundation Grant (20-15-00395).

Author contributions

M.G., M.S., D.S. and I.Z. proofread and incorporated sequencing data into the database and performed statistical analysis. D. Bagaev and D. Bolotin implemented, hosted and supported the web interface for the database. P.G.T., A.A.M., M.V.P., K.L., J.E.M., D.A.P., T.H.O.N., L.C.R., E.B.C., K.K., G.D., C.R.R., A.S., J.S., F.L., K.V.Z., A.A.K., S.A.S. and G.A.E. gathered, formatted and submitted sequencing data to the database. M.S., I.Z. and D.C. designed and curated the study. M.S., D.C., D.A.P., P.G.T., K.K., F.L., G.A.E. and A.S. wrote and edited the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01578-0.

Peer review information *Nature Methods* thanks Sam Darko, Baojun Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.