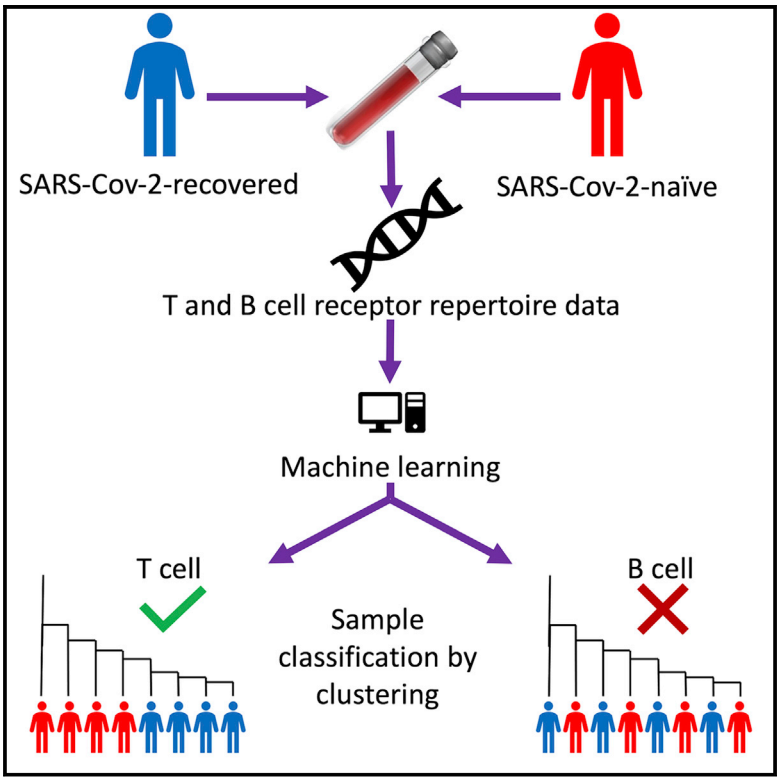
Cell Reports
Medicine

Report

Use of machine learning to identify a T cell response to SARS-CoV-2

Graphical Abstract



Authors

M. Saad Shoukat, Andrew D. Foers, Stephen Woodmansey, Shelley C. Evans, Anna Fowler, Elizabeth J. Soilleux

Correspondence

ejs17@cam.ac.uk

In Brief

To understand T cell responses to SARS-CoV-2, Shoukat et al. analyze TCR beta repertoire data from recovered COVID-19 patients and SARS-CoV-2 infection-naïve controls. Their machine learning approach can classify samples with up to 96.4% training accuracy and 92.9% testing accuracy. This method may detect T-cell responses acquired through natural infection or vaccination.

Highlights

- Machine learning can classify patient samples using their T cell receptor sequences
- T cell receptor sequence analysis accurately identifies recovered COVID-19 patients
- B cell receptor sequence analysis cannot identify recovered COVID-19 patients
- This method may detect T cell responses acquired through infection or vaccination

Shoukat et al., 2021, Cell Reports Medicine 2, 100192
February 16, 2021
https://doi.org/10.1016/j.xcrm.2021.100192

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Report

Use of machine learning to identify a T cell response to SARS-CoV-2

M. Saad Shoukat,[1,3] Andrew D. Foers,[1,3] Stephen Woodmansey,[1] Shelley C. Evans,[1] Anna Fowler,[2,4] and Elizabeth J. Soilleux[1,4,5,*]

[1]Department of Pathology, University of Cambridge, Cambridge, UK
[2]Department of Health Data Science, Institute of Population Health, University of Liverpool, Liverpool, UK
[3]These authors contributed equally
[4]These authors contributed equally
[5]Lead contact
*Correspondence: ejs17@cam.ac.uk
https://doi.org/10.1016/j.xcrm.2021.100192

SUMMARY

The identification of SARS-CoV-2-specific T cell receptor (TCR) sequences is critical for understanding T cell responses to SARS-CoV-2. Accordingly, we reanalyze publicly available data from SARS-CoV-2-recovered patients who had low-severity disease (n = 17) and SARS-CoV-2 infection-naive (control) individuals (n = 39). Applying a machine learning approach to TCR beta (TRB) repertoire data, we can classify patient/control samples with a training sensitivity, specificity, and accuracy of 88.2%, 100%, and 96.4% and a testing sensitivity, specificity, and accuracy of 82.4%, 97.4%, and 92.9%, respectively. Interestingly, the same machine learning approach cannot separate SARS-CoV-2 recovered from SARS-CoV-2 infection-naive individual samples on the basis of B cell receptor (immunoglobulin heavy chain; IGH) repertoire data, suggesting that the T cell response to SARS-CoV-2 may be more stereotyped and longer lived. Following validation in larger cohorts, our method may be useful in detecting protective immunity acquired through natural infection or in determining the longevity of vaccine-induced immunity.

INTRODUCTION

The identification of public SARS-CoV-2-specific T cell receptor (TCR) sequences, that is those TCR sequences shared between individuals who have recovered from the infection, is critical for understanding T cell responses to SARS-CoV-2. A recent study of T and B cell receptor (BCR) repertoires from coronavirus disease 2019 (COVID-19) patients demonstrated an association between TCR and BCR repertoire data and severity of disease.[1] Datasets from this study provide opportunities to look for a TCR/BCR signature of a SARS-CoV-2 adaptive immune response in patients who have recovered, comparing their samples with those from SARS-CoV-2 infection-naive (control) individuals as a first step toward identifying a signal that might indicate that an individual has protective immunity to SARS-CoV-2. Detection of such a signal could be useful in indicating that an individual has developed protective immunity through natural infection or in post-vaccination follow-up when considering the longevity of vaccine-induced immunity.

Until a safe and effective vaccine becomes widely available, determining likely protective immunity to SARS-CoV-2 at an individual level is paramount[2] both for healthcare personnel and to enable wider society to return to a level of normality, with attendant economic recovery. Immune status may be assessed by testing for SARS-CoV-2-specific antibodies, although it remains unclear whether the presence of antibodies confers robust SARS-CoV-2 protective immunity,[3] and studies of other human coronaviruses have demonstrated re-infection despite the presence of virus-specific antibodies.[4] Antibody decay, a recognized phenomenon in response to other human coronaviruses,[5] occurs in post-COVID-19 patients.[2,6] This is more common in individuals who experienced mild/asymptomatic infection and had low antibody titers in the early convalescent period.[7] Complete absence of a detectable antibody response is also described in some individuals following mild/asymptomatic infection.[8] Transient or absent humoral responses to SARS-CoV-2 may be due to dysregulated induction of B cell responses, likely correlating with observations of ineffective differentiation of T follicular helper cells and a reduction in Bcl6+ germinal center B cell levels.[9] Despite these poor humoral responses, recent studies reveal that seronegative individuals with mild or no symptoms and seronegative but exposed family members can produce a SARS-CoV-2-specific T cell response.[10,11] Longitudinal studies in other human coronaviruses suggest virus-specific T cell responses are more enduring than antibodies, persisting for at least 11 years.[5,12] T cell analysis may therefore represent a longer lasting and more sensitive means of evaluating immunity and might be particularly important in individuals experiencing mild/asymptomatic infection, who are less likely to have undergone RNA-based testing for active viral infection[2] and may have no detectable antibodies to SARS-CoV-2.[7,8]

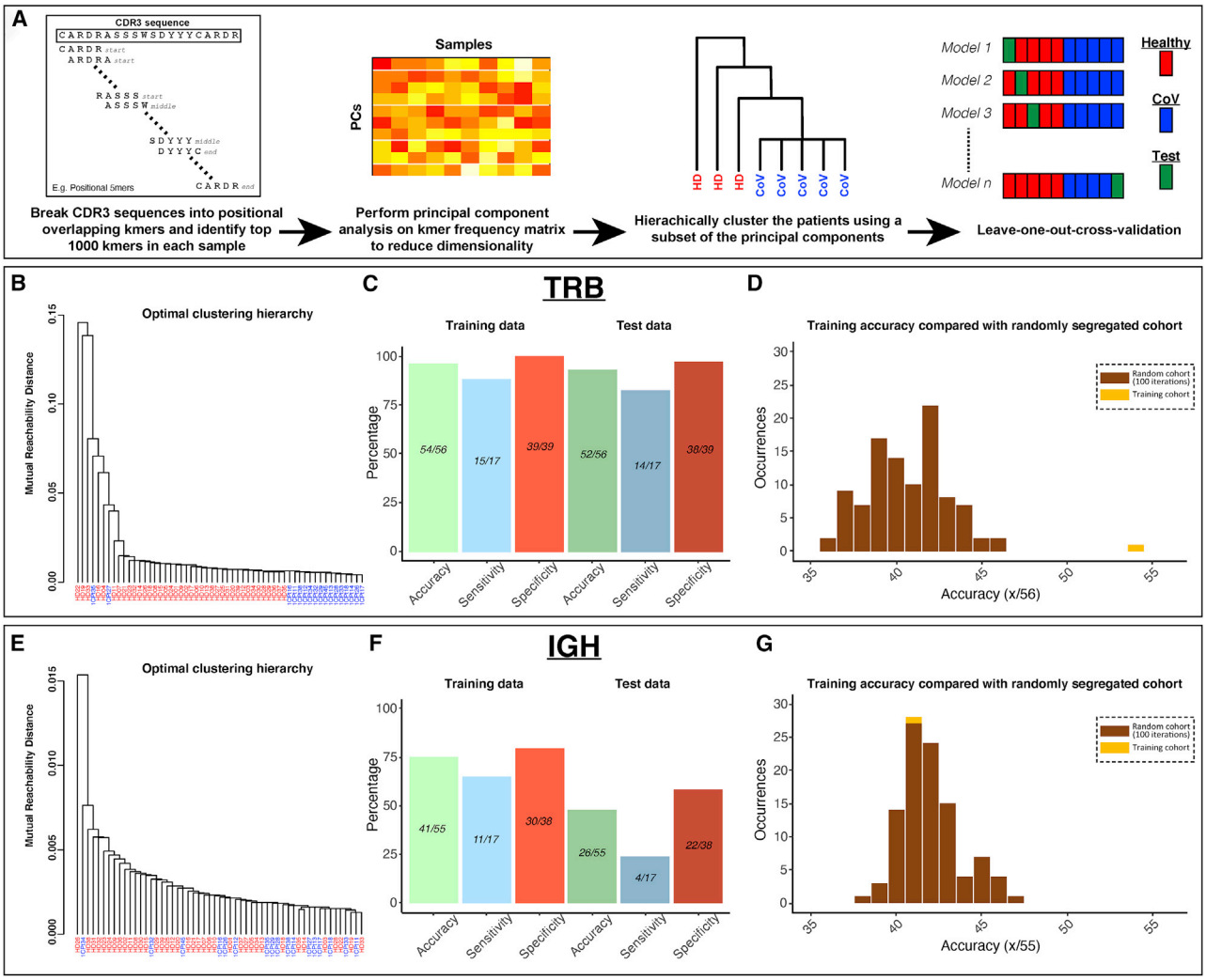


Figure 1. Clustering method for sample classification on the basis of T or B cell receptor repertoires

(A) The nucleic acid sequence is first translated to amino acid sequences, and functional CDR3 regions are defined with MiXCR. To take account of similar, but not clonotypically identical, TCR sequences, the entire CDR3 sequence of each TCR is split into short overlapping segments of length k, designated kmers, where k = 3–9 amino acids. Because the same kmer occurring at substantially different positions within the CDR3 is likely to differ in its effect on antigen binding, we positionally annotated kmers (start/middle/end) in the functional CDR3s. For example, the kmer sequence CARDR occurring toward the N terminus of a CDR3 sequence is regarded as distinct from the kmer sequence CARDR located toward the C terminus. In each patient sample, kmer frequencies are normalized and the 1,000 most frequent kmers are selected. Individual patient data are then merged into a single frequency matrix. Principal-component analysis (PCA) is used to reduce dimensionality while retaining major sources of variation, simplifying downstream computational steps. To classify samples, hierarchical clustering is applied to a subset of the principal components (PCs); this iteratively groups together samples forming a dendrogram. Samples for which the true underlying disease status is known are used to select optimal parameter sets, consisting of the value of k and the subsets of PCs, by means of a machine learning approach (i.e., to train the model). Selection of the value of k and the subsets of PCs essentially constitutes the machine learning component. The optimal parameters generate separate clusters for each immune state (i.e., SARS-CoV-2 recovered versus SARS-CoV-2 naive). For each kmer length, the first 10 PCs are considered in all possible combinations. To test the model, a leave-one-out-cross-validation approach is used, where each sample is iteratively removed and reintroduced. Upon reintroduction, the sample is blindly assessed as either a healthy or post-SARS-CoV-2 sample, based on which cluster it falls into.

(B) Analyzing the TRB data, a kmer length of 5 achieved greatest accuracy, with a top accuracy score of 96.4% (54/56 samples classified correctly). Here, 2 PC combinations gave 96.4% accuracy, with PC subset 2, 3, 4, 7, and 8 giving the greatest separation between diagnostic groups, with the greatest vertical distance (mutual reachability distance) between branches on the cluster plot. A range of other PC combinations also gave training classification accuracies >90% (not shown), indicating the robustness of our approach.

(C) Bar chart for TRB data, demonstrating training accuracy, sensitivity, and specificity compared with testing accuracy, sensitivity, and specificity, produced using the leave-one-out-cross-validation approach.

(D) Accuracy of best performing cluster analyses for TRB kmers of length 5 for 100 randomly labeled permutations of the patient data, giving a mean best score of 40.6/56 or 72.5(±4.11 SD)% training accuracy across all permutations, compared to 54/56 (96.4%) for the training dataset ($p = 0.01$ using a permutation test).

(E) Analyzing the BCR (IGH) data, a kmer length of 3 achieved a maximum training accuracy of 41/55 (74.5%). Here, 2 PC combinations gave 74.5% accuracy, with PC subset 7, 8, and 10 showing the greatest mutual reachability distance between diagnostic groups.

(legend continued on next page)

A small number of studies have analyzed relatedness of TCR/BCR repertoires to classify samples into groups correlating with diagnosis,[13] with machine learning approaches showing promise in this area.[14] For example, Beshnova et al.[15] demonstrated that a deep-learning model applied to peripheral TCR repertoire data can identify multiple cancer types with accuracies ≥95%. Similarly, we recently demonstrated that machine-learning-based analysis of TCR repertoires from duodenal gamma/delta T cells can separate patients with celiac disease from controls with ≥91% accuracy.[16] We chose to investigate whether a similar approach could identify individuals who had recovered from SARS-CoV-2 infection.

RESULTS

To investigate whether TCR/BCR analysis can identify patients with evidence of adaptive responses to SARS-CoV-2, we passed repertoire data from Schultheiß et al.[1] through our TCR/BCR classification algorithm and compared samples from SARS-CoV-2-recovered patients who had had low-severity disease (n = 17) with the SARS-CoV-2 infection-naive (control) cohort (n = 39). Our classification approach is based on the hypothesis that there are multiple related TCR/BCR sequences with similar specificities capable of binding SARS-CoV-2 antigens, permitting training of a machine learning algorithm to cluster these samples together by analyzing the amino acid sequence of the hypervariable part of the TCR/BCR complementarity determining region 3 (CDR3) (Figure 1A).[17] To take account of closely related sequences, we break the CDR3 sequences into overlapping kmers (amino acid sequences of length k, range 3–9), label the kmers with their position within the CDR3 sequence (start/middle/end), and compile a matrix detailing frequency of each kmer in each sample (Figure 1A). To focus on kmers most likely involved in immune responses, only the 1,000 most frequent kmers in each sample were analyzed. To decrease dimensionality of the data, we take principal components (PCs) and perform hierarchical clustering on the basis of these PCs, selecting the kmer length and PC combination that gives the greatest clustering accuracy.

Applying our bioinformatic algorithm[16] to Schultheiß's TCR beta (TRB) repertoire dataset, samples were classified with a training sensitivity, specificity, and overall accuracy of 88.2%, 100%, and 96.4%, respectively (Figures 1B and 1C). Due to the relatively small sizes of the cohorts (low-severity infection: n = 17; uninfected: n = 39), we were unable to use a fully independent test cohort. We therefore undertook a leave-one-out-cross-validation (LOOCV) approach, achieving a testing sensitivity, specificity, and overall accuracy of 82.4%, 97.4%, and 92.9%, respectively (Figure 1C).

We noted from the publication by Schultheiß et al.[1] that the samples could not be reliably separated by TRB clonality, Shannon diversity, or richness score. To further assess the robustness of our result, we performed a permutation test in which we randomly permuted the sample labels and attempted to separate the random groups using our methodology, testing the same range of parameters as we did for the true labels. In each of the 100 random groups, we considered all combinations of the first 10 PCs and recorded the maximum accuracy achieved for each random group. This gave a mean maximum accuracy score across all 100 permutations of 40.6/56 (72.5%; range = 63.3%–82.1%), compared with 54/56 (96.4%) for the training dataset ($p < 0.01$; permutation test; Figure 1D). We applied the same methodology to the B cell repertoire (IGH) datasets (Figures 1E–1G) and were not able to separate the two groups by this means.

DISCUSSION

Here, we demonstrate successful application of a machine learning method to the analysis of peripheral blood TCR sequence data in order to separate a cohort on the basis of whether or not individuals have previously had low-severity SARS-CoV-2 infection. Our identification of a TCR/BCR signature of a SARS-CoV-2 adaptive immune response in patients who have recovered, comparing their samples with those from SARS-CoV-2 infection-naive (control) individuals, represents a first step toward identifying a signal that might indicate that an individual has protective immunity to SARS-CoV-2. We particularly focused on low severity, because it is individuals who have recovered from asymptomatic or mildly symptomatic SARS-CoV-2 infection who are least likely to undergo testing for viral infection and thus least likely to know their immune status.

Although our methodology was successful for the analysis of TRB repertoire data (Figures 1A–1D), it could not separate the cohort into SARS-CoV-2 recovered and SARS-CoV-2 infection-naive on the basis of BCR (IGH) repertoire data (Figures 1E–1G). This suggests a more stereotyped and possibly longer-lived T cell response to SARS-CoV-2. This result is also consistent with Schultheiß et al.'s observation of enriched shared TCR compared with BCR motifs between SARS-CoV-2 recovered patients.[1] Considering longevity of anti-coronavirus T cell responses[12] and anti-SARS-CoV-2 T cell, but not antibody responses, in individuals with previous mild COVID-19,[10] our data indicate that analysis of TCR sequences, rather than serological assays, shows greater promise for identifying long-lived SARS-CoV-2 adaptive immune responses. Although, in some studies, T cell analysis in COVID-19 has been complicated by cross-reactivity with T cell responses generated through exposure to "common cold" coronaviruses,[18,19] the ability of our method to correctly classify SARS-CoV-2 infection-naive individuals suggests that any pre-existing T cell immunity to endemic coronaviruses does not confound the SARS-CoV-2 specificity of our approach. Importantly, further studies of larger cohorts are required to validate our findings and to investigate the longevity of T cell immunity to SARS-CoV-2, as well as

(F) Bar chart for BCR data, demonstrating training accuracy, sensitivity, and specificity compared with test accuracy, sensitivity, and specificity, produced using the leave-one-out-cross-validation approach.

(G) Accuracy of best performing cluster analyses for BCR kmers of length 3 for 100 randomly labeled permutations of the patient data, giving a mean best score of 42.0/55 (76.3 ± 3.14 SD%) training accuracy across all permutations, compared to 41/55 (74.5%) for the training dataset (not significant.; permutation test).

providing further insight into the relative importance of TCR and antibody-mediated immunity to SARS-CoV-2.

In summary, we describe a machine learning approach to TCR repertoire analysis that, when applied to a TRB dataset,[1] can accurately identify prior SARS-CoV-2 infection from long-lasting TRB profiles. The Biomed-2 primer sets used for TCR amplification in this study[1] are already in clinical use for the diagnosis of lymphoma and leukemia. Therefore, this method is amenable to existing diagnostic pathways in fully accredited clinical laboratories and could be rapidly scaled up, permitting the introduction of a novel test for immunity to SARS-CoV-2.

Limitations of study

This is a small and preliminary study, utilizing an analytical approach that we have previously successfully applied to the diagnosis of celiac disease using duodenal biopsy samples.[16] A larger dataset with separate training and test sets is required to corroborate our findings. Such a dataset will ideally need to have been generated using the same TCR/BCR repertoire sequencing methodology, preferably in one or more separate laboratories, to investigate the effect of laboratory-to-laboratory variation in sequencing methodology. A subsequent study to corroborate our findings using datasets produced with different TCR/BCR repertoire sequencing methods is ideally required to provide broader corroboration of the bioinformatic method. The methodology also requires validation on additional datasets drawn from different racial groups and/or individuals with different HLA types, as these may confound the analysis. Finally, the current study does not provide proof that the signal we identify is indicative of protective immunity, and individuals with a TCR signal indicative of a T cell immune response to SARS-CoV-2, as identified by our method, would need to be followed up to determine whether they can still become infected with SARS-CoV-2.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

ACKNOWLEDGMENTS

We are grateful to Schultheiß et al. for the provision of the publicly available datasets, which we have analyzed. This work was funded by a charitable donation from the Snudden Family Trust, UK.

AUTHOR CONTRIBUTIONS

M.S.S., A.D.F., A.F., and E.J.S. designed the study. M.S.S. and A.D.F. undertook data analysis. S.W., S.C.E., M.S.S., A.D.F., A.F., and E.J.S. contributed to the writing of the manuscript, and all authors have approved the final version.

DECLARATION OF INTERESTS

E.J.S. and A.F. are inventors of the following: GB patent application no. 1718238.7, for Oxford University Innovation, dated 3 November 2017; international patent application no. PCT/GB2018/053198 for Cambridge Enterprise, based on GB application no. 1718238.7, dated 5 November 2018.

Received: October 2, 2020
Revised: December 8, 2020
Accepted: January 12, 2021
Published: January 16, 2021

REFERENCES

1. 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.
2. Ibarrondo, F.J., Fulcher, J.A., Goodman-Meza, D., Elliott, J., Hofmann, C., Hausner, M.A., Ferbas, K.G., Tobin, N.H., Aldrovandi, G.M., and Yang, O.O. (2020). Rapid decay of anti–SARS-CoV-2 antibodies in persons with mild Covid-19. N. Engl. J. Med. *383*, 1085–1087.
3. To, K.K.-W., Hung, I.F.-N., Ip, J.D., Chu, A.W.-H., Chan, W.-M., Tam, A.R., Fong, C.H.-Y., Yuan, S., Tsoi, H.-W., Ng, A.C.-K., et al. (2020). Coronavirus disease 2019 (COVID-19) re-infection by a phylogenetically distinct severe acute respiratory syndrome coronavirus 2 strain confirmed by whole genome sequencing. Clin. Infect. Dis. Published online August 25, 2020. https://doi.org/10.1093/cid/ciaa1275.
4. Edridge, A.W.D., Kaczorowska, J., Hoste, A.C.R., Bakker, M., Klein, M., Loens, K., Jebbink, M.F., Matser, A., Kinsella, C.M., Rueda, P., et al. (2020). Seasonal coronavirus protective immunity is short-lasting. Nat. Med. *26*, 1691–1693.
5. Tang, F., Quan, Y., Xin, Z.-T., Wrammert, J., Ma, M.-J., Lv, H., Wang, T.-B., Yang, H., Richardus, J.H., Liu, W., and Cao, W.C. (2011). Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: a six-year follow-up study. J. Immunol. *186*, 7264–7268.
6. Seow, J., Graham, C., Merrick, B., Acors, S., Steel, K.J.A., Hemmings, O., Bryne, A., Kouphou, N., Pickering, S., Galao, R., et al. (2020). Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection. medRxiv, 2020.07.09.20148429.
7. Long, Q.-X., Tang, X.-J., Shi, Q.-L., Li, Q., Deng, H.-J., Yuan, J., Hu, J.-L., Xu, W., Zhang, Y., Lv, F.-J., et al. (2020). Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. Nat. Med. *26*, 1200–1204.
8. Liu, Z.-L., Liu, Y., Wan, L.-G., Xiang, T.-X., Le, A.-P., Liu, P., Peiris, M., Poon, L.L.M., and Zhang, W. (2020). Antibody profiles in mild and severe cases of COVID-19. Clin. Chem. *66*, 1102–1104.
9. Kaneko, N., Kuo, H.-H., Boucau, J., Farmer, J.R., Allard-Chamard, H., Mahajan, V.S., Piechocka-Trocha, A., Lefteri, K., Osborn, M., Bals, J., et al.; Massachusetts Consortium on Pathogen Readiness Specimen Working Group (2020). Loss of Bcl-6-expressing T follicular helper cells and germinal centers in COVID-19. Cell *183*, 143–157.e13.
10. Gallais, F., Velay, A., Wendling, M.-J., Nazon, C., Partisani, M., Sibilia, J., Candon, S., and Fafi-Kremer, S. (2020). Intrafamilial exposure to SARS-CoV-2 induces cellular immune response without seroconversion. medRxiv, 2020.06.21.20132449.
11. 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.; Karolinska COVID-19 Study Group (2020). Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. Cell *183*, 158–168.e14.

12. Ng, O.-W., Chia, A., Tan, A.T., Jadi, R.S., Leong, H.N., Bertoletti, A., and Tan, Y.-J. (2016). Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. Vaccine *34*, 2008–2014.

13. Greiff, V., Weber, C.R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., and Reddy, S.T. (2017). Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. J. Immunol. *199*, 2985–2997.

14. Ostmeyer, J., Christley, S., Rounds, W.H., Toby, I., Greenberg, B.M., Monson, N.L., and Cowell, L.G. (2017). Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. BMC Bioinformatics *18*, 401.

15. Beshnova, D., Ye, J., Onabolu, O., Moon, B., Zheng, W., Fu, Y.-X., Brugarolas, J., Lea, J., and Li, B. (2020). De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. Sci. Transl. Med. *12*, eaaz3738.

16. Foers, A.D., Shoukat, M.S., Welsh, O.E., Donovan, K., Petry, R., Evans, S.C., FitzPatrick, M.E.B., Collins, N., Klenerman, P., Fowler, A., and Soilleux, E.J. (2020). Classification of intestinal T-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. J. Pathol. Published online November 22, 2020. https://doi.org/10.1002/path.5592.

17. Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381.

18. 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.

19. 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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw sequence data	Schultheiß et al.[1]	https://www.ebi.ac.uk/ena/browser/view/PRJEB38339
Software and algorithms		
MiXCR	Bolotin et al.[17]	https://github.com/milaboratory/mixcr
In-house software written in R Statistical Software (4.0.2)	Foers et al.[16]	https://doi.org/10.5281/zenodo.3964131; 27th July 2020

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Elizabeth Soilleux (ejs17@cam.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate any unique datasets. Raw FASTQ sequencing files, corresponding to the TRB and IGH genetic loci, amplified by multiplexed PCR reactions (using BIOMED2-FR1 (IGH) or –TRB-A/ B primer pools), were obtained from the European Nucleotide Archive (accession number: PRJEB38339; https://www.ebi.ac.uk/ena/browser/view/PRJEB38339) from a previous study[1]. The source code, written in R Statistical Software (4.0.2), supporting the current study is available for research purposes, upon request, from Zenodo (https://doi.org/10.5281/zenodo.3964131; 27th July 2020) without restriction.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We used publicly available T/ B cell receptor repertoire data, acquired from next-generation sequencing of blood sample-derived DNA, from 17 recovered COVID-19 patients, who were described as having had mild disease courses, not requiring hospitalisation, with 1 patient having been asymptomatic, (Schultheiß's cohort 1) and 39 non-infected controls from this study[1]. The recovered patient cohort comprised 10 males and 7 females and had a median age of 34, but individual ages are not available to us. 1 male patient, in the 20-29 year age range had hypertension, but no other patients had health conditions known to increase individual risk from SARS-CoV-2 infection. All available details regarding these cohorts can be found in the supplemental material associated with the original publication of the datasets[1].

METHOD DETAILS

Productive CDR3 sequences, corresponding to the most variable part of each TCR/BCR sequence, were obtained from FASTQ files with MiXCR software[17] using the amplicon command and default parameters. Within each sample, normalized frequencies for each unique CDR3 amino acid sequence were calculated by dividing the CDR3 count by the sum of all productive CDR3 counts for the sample.

Each CDR3 amino acid sequence was broken into its constituent overlapping kmers (amino acid strings of length k), for a range of k of 3 – 9 amino acids, with kmers being positionally annotated as start, middle or end, to indicate which third of the parent CDR3 sequence contained the largest component of the kmer. Kmers with identical sequences, but occurring in different thirds of the CDR3 region, were treated as non-identical, as described previously[16]. Individual kmer frequencies were assigned as per the frequency of the parent CDR3 sequence and a matrix of kmer sequences and their total frequencies in each sample was generated for each sample. The 1000 most frequent kmers in each patient sample were then selected for sample classification.

Sample classification was performed as previously described[16]. Briefly, kmer matrix dimensionality was reduced by principal component analysis (PCA). For each value of k (3-9), combinations of the first 10 principal components (PCs) were assessed by hierarchical clustering, permitting determination of the value of k and the combination of PCs that could separate SARS-CoV-2-recovered patients from uninfected controls with greatest accuracy.

QUANTIFICATION AND STATISTICAL ANALYSIS

Using the optimal value of k and optimal combination of PCs, as determined above, a leave-one-out-cross-validation approach was implemented to calculate test accuracy, sensitivity and specificity. For leave-one-out-cross-validation, each sample was iteratively removed and reintroduced. Upon reintroduction, the sample was blindly assessed as being either from a SARS-CoV-2-recovered patient or from a SARS-CoV-2 infection-naive individual, based on which cluster it fell into, permitting calculation of test accuracy, sensitivity and specificity. Since classifications were decided according to the majority in each cluster, when there were 2 clusters, the minimum achievable accuracy in the permutation test was 50%, although in a small number of cases, in which samples segregated into more than 2 clusters, the minimum accuracy could potentially be slightly lower.