Supplementary material

TCRconv: Predicting recognition between T cell receptors and epitopes using contextualized motifs

S1. VDJdb confidence scores
All TCR-epitope pairs in VDJdb have been given confidence scores from 0-3 as follows:
9: low confidence or no information (a critical aspect of sequencing or specificity validation is missing)
1: moderate confidence (no verification or poor TCR sequence confidence)
2: high confidence (has some specificity verification, good TCR sequence confidence)
3: very high confidence (has extensive verification or structural data)

See more detailed description at https://github.com/antigenics/vdjdb-db

S2. Other embedding techniques for TCRs
We also attempted to make the ProBERT model more specialized to TCR sequences by fine-tuning it on 5 million TCRβ sequences from VDJdb (Bagaev et al., 2020), and studies of Emerson et al. (2017) and Dash et al. (2017) for eight epochs but this did not improve the prediction accuracies (mean AUROC 0.848 and AP 0.575 on VDJdb-small dataset). We also tested two ELMo (Embeddings from Language Models) architectures, classical ELMo (Peters et al., 2018) and masked ELMo (Senay and Salin, 2020), and trained them on a smaller dataset of 3 million TCRβ-sequences from the same sources as those used in the BERT fine-tuning. The main difference between these two models is that instead of unidirectional LSTMs, the masked ELMo uses a bidirectional two-layered LSTM and when trained in the token prediction task, the predicted token (amino acid) is masked to avoid leakage of information. We found that both ELMo models produced reasonable accuracies in the prediction task, and with masked ELMo we achieved almost as good accuracy as with the BERT embeddings (mean AUROC and AP 0.839 and 0.539 for ELMo and 0.847 and 0.571 for masked ELMo, on VDJdb-small dataset).

S3. Saliency maps
Gradient-based saliency maps can describe how much each position in a sequence influences the predicted epitope-specificity. We computed saliency maps for a TCRconv model with the protBERT model for computing embeddings for the CDR3 sequences using the full context (i.e., an embedding is first computed for the complete TCR determined by the CDR3, and V- and J-genes, and then the part corresponding to the CDR3 is extracted), trained with the VDJdb-large dataset. The saliency values were computed as the average over all absolute saliency values for all features at each position. To determine the importance of each residue individually, we compute the gradients of the true epitope binding w.r.t. the outputs of the non-contextualized layer (the input layer embedding's output) of protBERT. The values were scaled between 0 and 1 for each TCR separately. We chose to extract contextualized embeddings with protBERT without further fine-tuning its parameters (we did not find improvements when doing so, see Supplementary Section S2, and we wanted to avoid overfitting). Therefore, the gradients with respect to individual, uncontextualized residues need to propagate through the 30 untuned transformer layers in protBERT. Due to the complexity of the protBERT model, the high dimensionality of the embeddings, and the multiple convolutional layers in our predictor (four parallel convolutional layers and another consecutive convolutional layer, each with several filters), it is expected that identification of clear motifs can be challenging. This is illustrated by Supplementary Figure S15-S16 that show examples of these saliency maps for TCRs recognizing seven different epitopes, each from different epitope species. However, some more general observations can be made; On average the position-wise saliency values for the positions in CDR3 are higher than those outside the CDR3, and with the paired TCRβ, the average position-wise saliency was in general higher for the chain that had better predictive performance when used individually (see Supplementary Table S7 and Fig. 5). A few examples of saliency maps for paired TCRβ sequences are shown in Supplementary Fig. S17.

S4. Phenotypes of SARS-CoV-2 specific T cells in moderate and severe COVID-19
Count matrices, TCRβ-seq results, and metadata from Liao et al. (2020) were downloaded from GEO GSE145926. The data was analyzed mainly with Python package scv (Gayoso et al., 2022) (v 0.14.5) and R package Seurat (Hao et al., 2021) (v 4.0.4). Cells with > 10% mitochondrial gene counts, < 1000 UMI counts, < 200 or > 6000 detected genes, and cells with no detected TCR were filtered out. The highly variable genes were identified with "highly_variable_genes" function in scv tools with default parameters, which were then used to learn latent embeddings with "modelSCVI" function in scv tools with default parameters. The CD8+ T cells were then identified with SingleR (Aran et al., 2019) (v 1.6.1), and the process was repeated with scv tools. The obtained embeddings were then used for finding clusters with "FindNeighbors" and "FindClusters" functions and further visualized with UMAP dimensionality reduction with "RunUMAP" function using default parameters in Seurat. The optimal clustering threshold was chosen as 0.2 based on visual inspection of the clustering results in the UMAP reduced space. The markers used to test the clusters were found with Student's t-test using the "FindMarkers" function in Seurat with logfc.threshold = 0.25 from expression data that was scaled with “ScaleData” function with scaling factor of 10000. Patients C141, C142, and C144 have moderate COVID-19. Patients reported by Liao et al. (2020) to have severe (C143 and C145) or critical disease (C146, C148, C149, and C152) were considered to have severe COVID-19 in these analyses.

S5. Box plots
All boxplots presented in the paper have the same formatting, that uses the default settings used by Seaborn.

References
Aran, D. et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology, 20(2), 163–172.
Bagaev, D. V. et al. (2020). VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Research, 48(D1), D1057–D1062.
Dash, P. et al. (2017). Quantiﬁable predictive features deﬁne epitope-speciﬁc T cell receptor repertoires. Nature, 547(7661), 89–93.
Emerson, R. O. et al. (2017). Immunosequencing identiﬁes signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nature Genetics, 49(5), 659–665.
Gayoso, A. et al. (2022). A python library for probabilistic analysis of single-cell omics data. Nature Biotechnology, 40(2), 163–166.
Hao, Y. et al. (2021). Integrated analysis of multimodal single-cell data. Cell, 184(4), 3573–3587.
Liao, M. et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nature medicine, 26(6), 842–844.
Peters, M. E. et al. (2018). Deep contextualized word representations. arXiv.
Senay, G. and Salin, E. (2020). Masked ELMo: An evolution of ELMo towards fully contextualized RNN language models. arXiv.
Supplementary Fig. S1. TCR cross-reactivity in datasets a) VDJdbβ-small, b) VDJdbβ-large, and c) VDJdbβ-large. Each row of a heat map represents TCRs specific to the corresponding epitope and their fraction recognizing any of the epitopes present in the dataset. The bar plots on the right side of each heatmap show the average number of epitope specificities per TCR recognizing the epitope on the corresponding row. For example, TCRs specific to EBV epitope EBNA3A RLRAEAQVK recognize on average 2.2 different epitopes on (b) dataset VDJdbβ-large and 2.0 on (c) dataset VDJdbβ-large. TCRs recognizing certain epitopes have notable cross-reactivity. To highlight them we have marked DENV epitopes with pink, EBV epitopes with blue, and two HIV-1 epitopes (HIV-1 KRWIILGLNK and HIV-1 KRWIIMGLNK) with green.

Supplementary Fig. S2. Epitope-wise method comparison with respect to AUROC score on (a) VDJdbβ-small and (b) VDJdbβ-large datasets and with respect to average precision (AP) on (c) VDJdbβ-small and (d) VDJdbβ-large datasets. The results are sorted by increasing order of TCRconv predictions. To highlight the accuracies for epitopes with notably cross-reactive TCRs, we have highlighted such epitopes similarly to Supplementary Fig. S1: DENV epitopes with pink, EBV epitopes with blue, and two HIV-1 epitopes (HIV-1 KRWIILGLNK and HIV-1 KRWIIMGLNK) with green.
Supplementary Fig. S3. CDR3 edit distances on VDJdb-large from TCRs with chosen specificity to nearest TCR with same specificity (red) or other specificity (grey).
Supplementary Fig. S4. CDR3 edit distances on VDJdb\-large from TCRs with chosen specificity to all TCRs with same specificity (red) or to all TCRs with other specificity (grey). Y-axis has log-scale.
Supplementary Fig. S5. CDR3 edit distances on VDJdb\$small. (A) Edit distance from TCRs with chosen specificity to nearest TCR with same specificity (red) or other specificity (grey). (B) CDR3 edit distance from TCRs with chosen specificity to all TCRs with same specificity (red) or to all TCRs with other specificity (grey).
Supplementary Fig. S6: UMAP clustering of TCRs in VDJdbβ-small. Each dot corresponds to one TCR and is colored by its epitope specificity. TCRs specific to multiple epitopes are colored by only one of its specificities.
Supplementary Fig. S7. UMAP clustering of TCRs in VDJdbβ-small. Each dot corresponds to one TCR and is colored with red if it recognizes the epitope in the title and otherwise with grey.
Supplementary Fig. S8. Method comparisons. Mean AUROC and AP scores on (a) VDJdbβ-small and (b) VDJdbβ-large dataset.
Supplementary Fig. S9. TCRconv evaluation. All AUROC and AP scores are obtained over stratified 10-fold cross-validation.

(A) Pearson’s correlation between the diversity of epitope specific TCRs and the AUROC and AP scores. Panels (i) and (ii) show the mean AUROC scores for datasets VDJdbβ-small and VDJdbβ-large, respectively, and (iii) and (iv) mean AP scores for both datasets.

(B) Increasing embedding context size increases the predictive AUROC and AP scores. The schematics on the top show the approximate sections included in different context sizes. Complete TCR refers to using the complete TCR with the predictor, without extracting only the CDR3 part. Panels (i) and (ii) show the mean AUROC scores for datasets VDJdbβ-small and VDJdbβ-large, respectively, and (iii) and (iv) mean AP scores for both datasets.
Supplementary Fig. S10. HLA-types of the MHCs restricting the epitopes do not alone explain variance in results. All AUROC and AP scores are obtained over stratified 10-fold cross-validation.

(A) HLA-types of the MHCs restricting the epitopes in dataset VDJdb-β-large.

(B) TCRconv predictions for VDJdb-β-large dataset have some variation in terms of AUROC and AP scores when the predictions are divided into three groups (HLA-A, HLA-B, and HLA-DRA1) based on the HLA-gene.

(C) AUROC and AP scores for HLA-A*02 restricted epitopes are similar whether the TCRconv is trained only on TCRs specific to HLA-A*02 restricted epitopes or to TCRs specific to all epitopes in VDJdb-β-large dataset. For reference TCRconv model trained and tested with TCRs specific to all epitopes, corresponding to results shown in Fig. 1a ii, is shown on right.
Supplementary Fig. S11. TCRconv prediction performance for SARS-CoV-2 epitopes.
(A) TCRconv performance in terms of AUROC and AP scores when trained with 139099 TCRs specific to 188 peptide groups from SARS-CoV-2. Mean scores are shown above both boxplots. Each circle represents the score for one peptide group, colored by the genomic region and numbered according to Supplementary Table S3.
(B) TCRconv performance when trained with TCRs specific to 20 best performing peptides groups from SARS-CoV-2 combined with VDJdbβ-large dataset; above results for all 70 peptide (groups) and below for only the 20 SARS-CoV-2 peptides. For SARS-CoV-2 peptides coloring and numbering are the same as in panel (a), other epitopes are white, and the numbering corresponds to Supplementary Table S1.
(C) AUROC and AP scores from the model from (a) by the peptides’ genome location and the diversity of the TCRs specific to each peptide group by the peptides’ genome location.
Supplementary Fig. S12. Analysis with COVID-19 patient repertoires.

(A) Shannon clonality (i) for each dataset and (ii) by Days from diagnosis to sample.
(B) Subject age (i) by dataset and (ii) by Days from diagnosis to sample.
(C) Normalized frequency grouped by number of days from diagnosis to sample for (i) six EBV specific epitopes and (ii) for four HCV specific epitopes.
Supplementary Fig. S13. T cell phenotypes and specificity in COVID-19.

(A) Characteristics of scRNA+TCR UMAP representation of CD8+ T-cells based on their phenotypes, colored by patient. Patients C141, C142, and C144 have moderate COVID-19, while patients C143, C145, C146, C148, C149, and C152 have severe disease.

(B) Frequencies of T-cells predicted to be specific to the tested epitopes separately for each patient. Only T-cells with both TCR- and RNA-seq available are shown.
Supplementary Fig. S14. Predicted and experimentally validated specificity of TCRs for SARS-CoV-2 epitope SPIKE YLQPRTFLL.

Each TCR clone in the repertoire sample ADIRP0000273_20200527 is represented as a circle that is colored red if it has been validated in the MIRA experiment eQD123 and grey if not. Each circle is positioned by its productive frequency (y-axis) and TCRconv prediction score (x-axis). The two vertical black lines show prediction thresholds 0.643 and 0.944 that correspond to false positive rates of 0.001 and 0.0001 obtained from the 10-fold cross-validation with VDJdbβ-dataset. The TCRs with clone size one are shaded. The table below shows the true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR), and positive predictive value (PPV) for the two thresholds and for clones of size at least two or at least three.

| Threshold | Clone size at least 2 | Clone size at least 3 |
|-----------|-----------------------|-----------------------|
|           | TPR | FPR     | FDR     | PPV | TPR | FPR     | FDR     | PPV |
| 0.643     | 0.899 | 0.000488 | 0.707 | 0.293 | 0.823 | 0.000391 | 0.608 | 0.392 |
| 0.944     | 0.500 | 0.0000770 | 0.382 | 0.618 | 0.484 | 0.0000743 | 0.333 | 0.687 |
Supplementary Fig. S15. Saliency maps for CDR3β sequences, one example for each epitope species in VDJdbβ-large dataset (1/2).

Each plot consists of a sequence logo and a heatmap of CDR3 sequences with the most common length specific to an epitope. The height of a letter in a sequence logo corresponds to that amino acids frequency at that position, and the background color of the letter shows the average saliency for the amino acid at that position. The heatmap shows the saliency values for each CDR3 sequence individually. The sequences are clustered by the similarity of their saliency values, as illustrated by the dendogram on its left side.
Supplementary Fig. S16. Saliency maps for CDR3β sequences, one example for each epitope species in VDJdbβ-large dataset (2/2).

Each plot consists of a sequence logo and a heatmap of CDR3 sequences with the most common length specific to an epitope. The height of a letter in a sequence logo corresponds to that amino acid frequency at that position, and the background color of the letter shows the average saliency for the amino acid at that position. The heatmap shows the saliency values for each CDR3 sequence individually. The sequences are clustered by the similarity of their saliency values, as illustrated by the dendogram on its left side.
Supplementary Fig. S16. Saliency maps for paired TCRαβ sequences from VDJdb-large dataset, a few examples of TCRs specific to EBV epitope BMLF1GLCTLVAML, YFV epitope NS4BLLWNGPMAV, or SARS-CoV-2 epitope SpikeYLQPRTFLL. TCRs specific to BMLF1GLCTLVAML epitope have on average higher saliency values for the β-chain, TCRs specific to NS4BLLWNGPMAV for the α-chain and with SpikeYLQPRTFLL the saliency values are quite similar to both chains (see Supplementary Table S7).

(A) Paired CDR3αβ sequences with the two most common lengths specific to EBV epitope BMLF1GLCTLVAML, YFV epitope NS4BLLWNGPMAV, or SARS-CoV-2 epitope SpikeYLQPRTFLL. The height of a letter in a sequence logo corresponds to that amino acids frequency at that position, and the background color of the letter shows the average saliency for the amino acid at that position.

(B) Examples of paired TCRαβ sequences.
Supplementary Table S1. Three datasets of epitope-specific TCR-data collected from VDJdb. The datasets contain epitope-specific TCRs for Cytomegalovirus (CMV), Dengue virus types 1, 2 and 3 (DENV1, DENV2, DENV3-4), Epstein-Barr virus (EBV), Hepatitis C virus (HCV), Human immunodeficiency virus type 1 (HIV-1) and Influenza A virus (IAV), Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and Yellow Fever virus (YFV), as well as human stromal antigen 2 (BST2), insulin like growth factor 2 mRNA binding protein 2 (IGF2BP2), melanoma antigen (MLANA), and transketolase (TKT). VDJdb-large and VDJdb-small were collected in January 2021 and VDJdb-large in September 2021 which explains why some of the SARS-COV-2 epitopes are only present in VDJdb-large.

| # | Epitope Species | Epitope Gene | Epitope | MHC chain 1 | MHC chain 2 | VDJdb-large | VDJdb-small | VDJdb-large |
|---|----------------|--------------|---------|-------------|-------------|-------------|-------------|-------------|
| 1 | CMV | IE1 | KLGGAQLA0K | HLA-A*03 | B2M | 12693 | 13664 |
| 2 | CMV | IE2 | NEGYKAAW | HLA-B*44 | B2M | 118 | 62 |
| 3 | CMV | pp65 | VTEGRILLY | HLA-A*01 | B2M | 202 |
| 4 | CMV | pp65 | FSNYNHY | HLA-B*35 | B2M | 92 | 58 |
| 5 | CMV | pp65 | NLVPMHATV | HLA-A*02 | B2M | 4488 | 244 | 175 |
| 6 | CMV | pp65 | TPRVTGSGAM | HLA-B*07 | B2M | 197 | 122 |
| 7 | CMV | pp65 | LLGTDHRRVSGP3L | HLA-DRA1*01 | HLA-DRB1*15 | 304 |
| 8 | CMV | pp65 | MNPSNPV | HLA-A*02 | B2M | 73 |
| 9 | CMV | pp65 | YSHFPPFTSQY | HLA-A*01 | B2M | 52 |
| 10 | DENV1 | NS3 | GTGS pistN | HLA-A*11 | B2M | 185 | 59 |
| 11 | DENV2 | NS3 | GTGS pistDK | HLA-A*11 | B2M | 60 |
| 12 | DENV3-4 | NS3 | GTGS pistN | HLA-A*11 | B2M | 158 | 46 |
| 13 | EBV | BMLF1 | GLCLTVAIL | HLA-A*02 | B2M | 999 | 169 | 279 |
| 14 | EBV | BRLF1 | YVLDMHIVV | HLA-A*02 | B2M | 79 | 51 |
| 15 | EBV | BZLF1 | RAXFKGOL | HLA-B*08 | B2M | 842 | 151 | 1212 |
| 16 | EBV | EBNAA2A | ALRAEAGVNV | HLA-A*03 | B2M | 410 | 422 |
| 17 | EBV | EBNAA2A | AVFIRSDK | HLA-A*11 | B2M | 1643 | 1723 |
| 18 | EBV | EBNAA4 | IVDDFSIVIK | HLA-A*11 | B2M | 550 | 713 |
| 19 | HCV | NS3 | ATD doucheT0Y | HLA-A*01 | B2M | 169 | 139 |
| 20 | HCV | NS3 | CINSVCWTV | HLA-A*02 | B2M | 131 | 76 |
| 21 | HCV | NS3 | KVLAGDNHF | HLA-A*02 | B2M | 65 | 65 |
| 22 | HCV | NS8B | ARMILMTHF | HLA-B*27 | B2M | 66 |
| 23 | HV-1 | Gag | EIYKRVIII | HLA-B*08 | B2M | 148 | 60 |
| 24 | HV-1 | Gag | FRDVYDFQV/KLRAEFAQ/SQ | HLA-DRA1*01 | HLA-DRB1*01,07,11,15 | HLA-DRB1*01 | 367 | 95 |
| 25 | HV-1 | Gag | GPQHAKRVL | HLA-B*07 | B2M | 86 | 93 |
| 26 | HV-1 | Gag | KAFSPEVPMF | HLA-B*57 | B2M | 175 | 104 |
| 27 | HV-1 | Gag | KRRWILNK | HLA-B*27 | B2M | 320 | 141 |
| 28 | HV-1 | Gag | KRRWILNK | HLA-B*27 | B2M | 86 |
| 29 | HV-1 | Gag | SLYNTAVTT | HLA-A*02 | B2M | 57 |
| 30 | HV-1 | Gag | TPQDLNTMNL | HLA-B*42 | B2M | 101 | 40 |
| 31 | HV-1 | Nef | FLXKGGOL | HLA-B*08 | B2M | 144 | 78 |
| 32 | HV-1 | Nef | TPQGGR/VRP | HLA-B*07 | B2M | 63 |
| 33 | HV-1 | Nef | ISRPTLNAK | HLA-B*57 | B2M | 54 |
| 34 | HV-1 | Vif | MPHVI/S393VEVH | HLA-B*42 | B2M | 54 |
| 35 | HV-1 | Vpr | FPRWALKHLG | HLA-B*42 | B2M | 83 |
| 36 | HTLV-1 | Tax | SFHGLH/LD | HLA-A*01 | B2M | 132 | 45 |
| 37 | Human | BST2 | LULLGQLV | HLA-A*02 | B2M | 233 |
| 38 | Human | IGF2BP2 | NLSALGIFT | HLA-A*03 | B2M | 111 |
| 39 | Human | MLANA | ELAAGITLV | HLA-A*02 | B2M | 1305 | 388 |
| 40 | Human | SEC24A | FLYNWTR | HLA-A*02 | B2M | 61 |
| 41 | Human | TKT | AMFWPGYTV | HLA-A*02 | B2M | 82 |
| 42 | HIV | HA | PKTVKYNKLKAT | HLA-DRA1*01 | HLA-DRB1*01,04 | 388 | 89 | 59 |
| 43 | HIV | M1 | GLG/QVF | HLA-A*02 | B2M | 3430 | 160 | 1815 |
| 44 | HIV | M1 | GLYNRMDGAVTTEV | HLA-DRA1*01 | HLA-DRB1*01 | 121 |
| 45 | HIV | M1 | QARCMQVQAMRTGTPH | HLA-DRA1*01 | HLA-DRB1*01 | 124 |
| 46 | HIV | M1 | SOPLXAEADRLD | HLA-DRA1*01 | HLA-DRB1*01 | 64 | 76 |
| 47 | HIV | NP | DATYQRATLVR | HLA-A*68 | B2M | 102 | 92 |
| 48 | HIV | NP | DPFLRLLNSQVFS | HLA-DRA1*01 | HLA-DRB1*01 | 104 |
| 49 | HIV | NP | LPRRSSGAAGA | HLA-B*07 | B2M | 159 |
| 50 | HIV | NS4B | LUNWLPMAW | HLA-A*02 | B2M | 409 | 239 |
| 51 | SARS-CoV-2 | Spike | YLDPRFTFL | HLA-A*02 | B2M | 315 | 261 |
| 52 | SARS-CoV-2 | Spike | LDLTEIAQY | HLA-A*01 | B2M | 122 |
| 53 | SARS-CoV-2 | Spike | NQKLIQNF | HLA-B*15 | B2M | 71 |
| 54 | SARS-CoV-2 | Spike | TTDPSFLPORY | HLA-A*01 | B2M | 243 |
| 55 | SARS-CoV-2 | Spike | Nucleaseap | SPRRWYFYS | HLA-B*07 | B2M | 75 |

TOTAL epitope-TCR pairs: 32367
TOTAL unique TCRs: 30503
genomic location (Loc). The coloring of the genomic regions matches Supplementary Fig. 8.

Supplementary Table S2. Method comparison. Mean AUROC and AP scores for TCRconv, TCRGP, TCRdist, SETE, DeepTCR and ERGO-II from stratified 10-fold cross-validation. Mean AUROC and AP scores are macro averages over all epitopes. Standard deviation is given over all folds and over all epitopes (Eqn. ), showing that with all methods variation between folds is smaller than variation between different epitopes. With TCRconv we have used protBERT embeddings for CDR3 + full ORF1ab, ORF1ab, ORF1ab, ORF1ab, ORF1ab, ORF1ab

| Method | CDR3 + full context | Mean AUROC | Standard deviation folds Epit. | Mean AP | Standard deviation folds Epit. | VDJdb-small | VDJdb-large |
|--------|---------------------|------------|---------------------------------|--------|---------------------------------|-------------|-------------|
| TCRconv | 0.853 | 0.238 | 0.564 | 0.555 | 0.164 | 0.679 | 0.010 | 0.104 | 0.283 | 0.016 | 0.250 |
| TCRGP | 0.881 | 0.239 | 0.571 | 0.451 | 0.106 | 0.679 | 0.010 | 0.104 | 0.188 | 0.011 | 0.177 |
| all CDR3 | 0.880 | 0.238 | 0.562 | 0.544 | 0.173 | 0.728 | 0.014 | 0.106 | 0.222 | 0.014 | 0.194 |
| DeepTCR | 0.732 | 0.198 | 0.566 | 0.234 | 0.180 | 0.705 | 0.007 | 0.106 | 0.173 | 0.009 | 0.173 |
| CETE CDR3 | 0.679 | 0.225 | 0.569 | 0.207 | 0.175 | 0.769 | 0.005 | 0.104 | 0.213 | 0.012 | 0.194 |
| TCRdist | 0.702 | 0.072 | 0.568 | 0.334 | 0.153 | 0.741 | 0.008 | 0.125 | 0.183 | 0.009 | 0.169 |
| CETE CDR3 | 0.710 | 0.230 | 0.573 | 0.432 | 0.190 | 0.704 | 0.008 | 0.125 | 0.183 | 0.009 | 0.169 |
| ERGO-II CDR3 | 0.761 | 0.220 | 0.546 | 0.246 | 0.130 | 0.673 | 0.028 | 0.102 | 0.053 | 0.005 | 0.091 |

\[ \text{AUROC} = \frac{1}{N} \sum_{i=1}^{N} \text{AUROC}_i \]

\[ \text{AP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \]

\[ \text{deviation} = \frac{1}{N} \sum_{i=1}^{N} \left( \text{score}_i - \text{mean score} \right) \]

\[ S = \sum_{i=1}^{N} \frac{S_i}{N} \]

\[ \text{standard deviation} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \]

Supplementary Table S3. Overview of ImmuneCODE MIRA data used for training TCRconv classifiers for SARS-CoV-2 specific epitopes. Spike refers to surface glycoprotein, membrane to membrane glycoprotein, and nucleocap to nucleocapid phosphoprotein. The peptide groups are ordered by the start point of their genomic location (Loc). The coloring of the genomic regions matches Supplementary Fig. 8.
Supplementary Table S4. Healthy control and ImmuneCODE repertoire data used in the analysis for T-cell dynamics during COVID-19 (Fig. 2a). The controls consist of the first 72 TCR repertoires from healthy (CMV-) subjects in cohort 1 in the study of Emerson et al. that had over 250 000 TCRs, number of templates reported, and where the subject is known to be at least 18 years old (which is the age of the youngest subject in the ImmuneCODE data used here). From ImmuneCODE 493 repertoires with over 250 000 TCRs and “Days from diagnosis to sample” reported were selected from four separate datasets.

| Cohort type | Cohort name | Institution | Study description | Mean age and s.d. (years) | Number of samples | Samples with ≥ 250 000 TCRs and Days from diagnosis to sample reported |
|-------------|-------------|-------------|-------------------|--------------------------|------------------|------------------------------------------------------------------|
| Healthy control | Emerson | Fred Hutchinson Cancer Research Center | Human peripheral blood samples were obtained from the institution’s Research Cell Bank biorepository of healthy bone marrow donors. Donors underwent CMV serostatus testing at the time the samples were taken | 36.1 ± 8.9 | 72 | |
| COVID-19 | ImmuneRACE (ADI) | Adaptive Biotechnologies | Whole blood samples were collected from subjects from 24 geographic areas in the US with active infection, in convalescent phase, or exposed to SARS-CoV-2 | 42.6 ± 11.9 | 123 | 118 |
| COVID-19 | ISB | Institute for Systems Biology | Whole blood samples collected under the INCOVE project at Providence St. Joseph Health (Seattle, WA). Subjects were enrolled during the active phase and monitored through disease. | 66.1 ± 16.7 | 157 | 48 |
| COVID-19 | DLS | Discovery Life Sciences | Whole blood samples collected during routine care in acudes and convalescent phases procured through Discovery Life Sciences (Huntsville, AL.) | 64.1 ± 18.5 | 431 | 216 |
| COVID-19 | H12O | Hospital Universitario 12 de Octubre | Whole blood samples were collected at the Hospital Universitario 12 de Octubre (Madrid, Spain) during the active or convalescent phase. | 60.5 ± 19.1 | 612 | 111 |

Total: 110 × 493

Supplementary Table S5. Significance of case-control and age effects on frequency of virus specific T-cells. Linear regression analysis was performed to assess if COVID patients have significantly higher frequency of virus specific T-cells than healthy control subjects, and if frequencies are positively correlated with subjects’ age (see Methods). (A) The Benjamini-Hochberg adjusted p-values representing the significance of β_0 > 0. (B) The Benjamini-Hochberg adjusted p-values representing the significance of β_0 > 0. One-tailed t-test was used for computing the p-values and the multiple testing adjustments are done for each virus (column) separately. Adjusted p-values smaller than 0.1 are bolded.

Supplementary Table S6. Embedding comparison. Mean AUROC and AP scores from stratified 10-fold cross-validation with TCRCov on VDJdbβ-small and VDJdbβ-large datasets using different embeddings.

Supplementary Table S7. Average position-wise saliency values for TCRs specific to each epitope in VDJdbβ-large dataset. Values are given separately for α- and β-chains for the CDR3 region and the complete TCR, defined by the V- and J-genes and CDR3.