Robust detection of SARS-CoV-2 exposure in population using T-cell repertoire profiling

Elizaveta Vlasova[1,2], Alexandra I. Nekrasova[3], Alexander Y Komkov[4,5,6], Mark Izraelson[4], Ekaterina A. Snigir[3], Sergey I. Mitrofanov[3], Vladimir S. Yudin[3], Valentin V. Makarov[3], Anton A. Keskinov[3], Darya Kirienko[1], Anastasia Pivnyuk[7], Pavel V Shelyakin[4,6], Ilgar Z Mamedov[4], Denis V Rebrikov[1], Dmitry M Chudakov[1,4,6], Sergey M. Yudin[3], Veronika I. Skvortsova[8], Olga V Britanova[1,4], Mikhail Shugay[1,4]

[1]Institute of Translational Medicine, Pirogov Russian National Research Medical University, Moscow, Russia

[2]ITMO University, Saint-Petersburg, Russia

[3]Federal State Budgetary Institution "Centre for Strategic Planning and Management of Biomedical Health Risks", Federal Medical Biological Agency, Moscow, Russia

[4]Department of Genomics of Adaptive Immunity, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia

[5]Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia

[6]Abu Dhabi Stem Cells Center, Abu Dhabi, United Arab Emirates

[7]Skolkovo Institute of Science and Technology, Moscow, Russia

[8]The Federal Medical Biological Agency (FMBA), Moscow, Russia

Abstract

The COVID-19 pandemic clearly demonstrates the need to monitor the spread of infectious diseases and population immunity. Probing adaptive immunity by sequencing the repertoire of antigen receptors (Rep-Seq) encoding specificity and immunological memory has become a method of choice for immunology studies. Rep-Seq can detect the imprint of past and ongoing infections and study individual responses to SARS-CoV-2 as shown in a number of recent studies. Here we apply a machine learning approach to two large datasets with more than 1200 high-quality repertoires from healthy and COVID-19-convalescent donor repertoires to infer T-cell receptor (TCR) repertoire features that were induced by SARS-CoV-2 exposure. Proper standardization of Rep-Seq batches, access to human leukocyte antigen (HLA) typing and both α- and β-chain sequences of TCRs allowed us to generate a high-quality biomarker database and build a robust and highly accurate classifier for COVID-19 exposure applicable to individual TCR repertoires

obtained using different protocols, paving a way to Rep-Seq-based immune status assessment in large cohorts of donors.

Introduction

Pathogen exposure leaves an imprint on the T-cell immunity, with pathogen recognition being guided by antigens presented by HLA molecules and resulting in T-cell expansion and formation of immunological memory[1]. This imprint can be explored by sequencing the T-cell receptor (TCR) repertoire[2] and a number of recent studies have revealed that the structure of TCR repertoire from peripheral blood mononuclear cells (PBMC) can be explained by a relatively simple V(D)J rearrangement model[3] together with the signs of past infections[4]. A large number of studies utilize AIRR-Seq data to infer biomarkers associated with a certain donor phenotype and pathology[5,6], including efforts directed at predicting common pathogen exposure from repertoire, such as CMV or SARS-CoV-2 exposure[7–9].

The recent COVID-19 pandemic has increased efforts to analyze TCR repertoires[10,11]. However, most of the studies have been conducted on small cohorts of individuals (<500 donors), which may not be sufficient for large-scale statistical analysis[12]. Other existing studies for large cohorts show the high quality classification of COVID-19 severity based on VJ usage or TCR sequence patterns[13,14], but the authors do not provide the batch effect correction methods, which could affect the quality of the results. Interestingly, the COVID-19 immune response traces are identified in both healthy donors and patients[15], which may indicate cross-reactivity of the T cell immune response to SARS-CoV-2. In addition, the virus is rapidly evolving[16–18] and it has recently been shown that SARS-CoV-2 specific immunity is mediated by the diversity of the antigen-specific T cells[19], highlighting the importance of further research . The immune response is strongly differentiated by the patient's HLA alleles, which is important to account for when building the AIRR-based classifiers.

In the present study we report a classifier that can distinguish SARS-CoV-2-exposed patients from uninfected donors sampled before the COVID-19 pandemic. The main idea of this approach is to use complementarity determining region 3 (CDR3) sequences that are highly diverse and encode most information on TCR antigen specificity as biomarkers for training the classifier, extending far beyond the set of previously reported TCRs recognizing cognate SARS-CoV-2 epitopes[8,20,21].

The cohort reported in the present study features individuals with known COVID-19 status and HLA class I and II haplotype that were subject to deep profiling of both TCR α and β chain repertoires. We extract CDR3 sequences that can serve as biomarkers of COVID-19 status in our cohort and assess them by studying HLA-association, TCRα-β chain pairing and discover a large number of homologous groups of TCR sequences (‘metaclonotypes’) similar to known SARS-CoV-2-related TCRs. We aim at minimizing batch effects that are prominent in Rep-Seq studies which allows us to generalize our classifier across data coming from different batches and remove potential biases[22–24], resulting in a robust classifier that shows good performance on unseen data batches. We use previously reported cohort of convalescent and healthy individuals by Adaptive Biotechnologies[8,25] to further validate the performance of our classifier.

Results

Study design overview

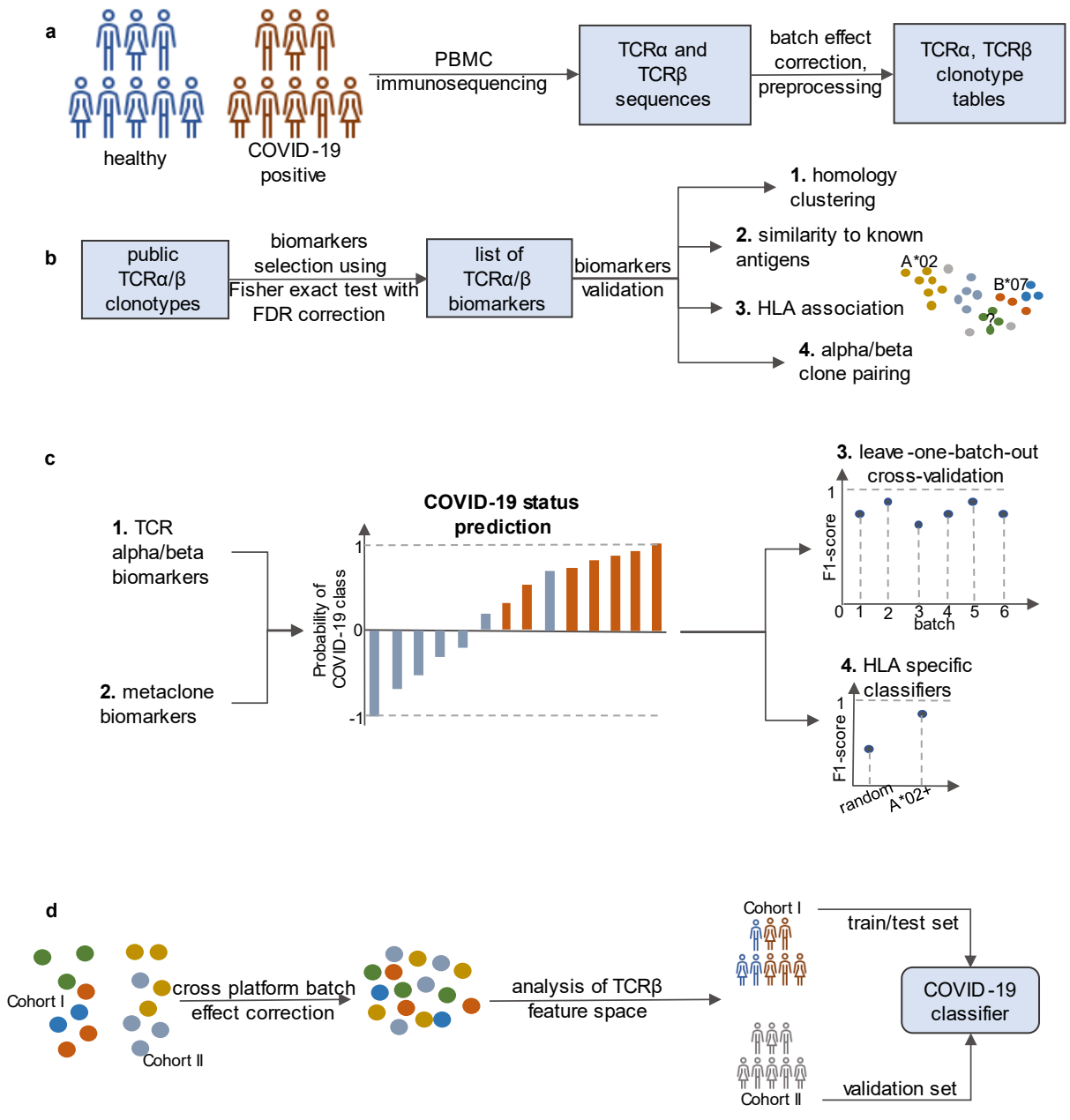

Figure 1. Experimental and analytical procedure overview. A. Peripheral blood samples were obtained for donors from Cohort-I reported here with known COVID-19 status and HLA context. TCR α and β loci were sequenced using conventional DNA-based protocol involving multiplex-PCR (see **Methods**). Samples were preprocessed and mapped to Variable, Diversity and Joining gene library to identify V/D/J alleles and extract CDR3 regions. Batch effect correction procedure was applied to normalize V(D)J rearrangement frequencies. **B.** TCRα and TCRβ biomarkers were selected for both T-cell chains separately using Fisher exact test. These biomarkers were aggregated by clustering TCR sequences into 'metaclonotypes'. Annotation of these clusters was performed using a database of TCR sequences with known antigen specificity, association with HLA metadata, and the results of α-β chain pairing analysis. **C.** Classifiers including various sets of features were constructed based on metaclope biomarkers described above. These classifiers were trained and evaluated using leave-one-batch-out cross validation technique. **D.** We used another cohort (Cohort-II published previously) to study the robustness of our classifier. Batch effect correction and sample preprocessing techniques were applied to both cohorts and the classifier was trained entirely on one of the cohorts and validated on another cohort.

The main aim of our study is to find TCR sequence biomarkers and develop a bioinformatic pipeline that allows building an accurate and robust classifier that distinguishes COVID-19-convalescent donors from unexposed individuals, as summarized in **Figure 1**. We performed immunosequencing of the rearranged TCR α and β regions for PBMCs. For the cohort described in this study (Cohort-I) we sequence both chains of the TCR heterodimer as both of these chains are required to properly predict antigen recognition[26]. We ran conventional T-cell repertoire data analysis (see **Methods**) and pre-processed data to remove low-coverage samples (**Supplementary Figure 1A,B**).

After sequencing, we processed the data to extract clonotype tables from each sample, that is, TCR sequences rearranged from a certain Variable and Joining genes and having a specific CDR3 region sequence. Hereafter we will focus on CDR3 amino acid sequences as a proxy for a specific T-cell clonotype and the number of sequencing reads corresponding to it as the clonotype size.

Of samples in Cohort-I which passed read count threshold, 383/377 TCR α/β samples were from healthy donors (SARS-CoV-2 PCR test negative or obtained prior to pandemic) and 890/848 were from COVID-19-positive patients. The majority of samples were accompanied by information on HLA class I and II alleles. Samples were prepared and sequenced in nine batches.

We proceed with selecting a set of CDR3 sequences that can serve as biomarkers and form a feature list for COVID-19 status classifier. Selecting an unbiased set is problematic due to the batch effect that stems from cross-sample contamination[27] and method-specific biases[23]. Note that such biases can also affect the classifier training procedure as it can learn from them when batches are unbalanced in terms of cases. In order to tackle batch effects we perform a standardization procedure that allows us to correct for differences in Variable and Joining gene usage that are the major source of differences across batches.

We also validate the resulting set of clonotypes in several ways. Co-occurence of specific TCR α and β clonotypes can serve as an independent validation for biomarkers and their co-association with some specific pathogen. Additional information on donor HLAs is provided to filter the set of biomarkers based on HLA restriction: association with donor HLA serves as an additional evidence for TCR specificity to a specific set of antigens presented in a given donor and allows detecting the fingerprint of past and present infection[28]. Furthermore, clonotypes with similar sequences can be aggregated into 'metaclonotype' biomarkers based on clonotype graph analysis.

Finally, we train various COVID-19 status classifiers on selected batches from Cohort-I data using different algorithms and incorporating different feature sets. Verification of the robustness of our results was performed using independent batches of the Cohort-I and data from Cohort-II published previously[8] that

contains TCRβ chain sequencing of COVID-19-convalescent donors merged with control samples obtained prior to the pandemic published in Emerson *et al.*[25](**Supplementary Figure 1C,D**).

Batch effects and their correction in TCR repertoires

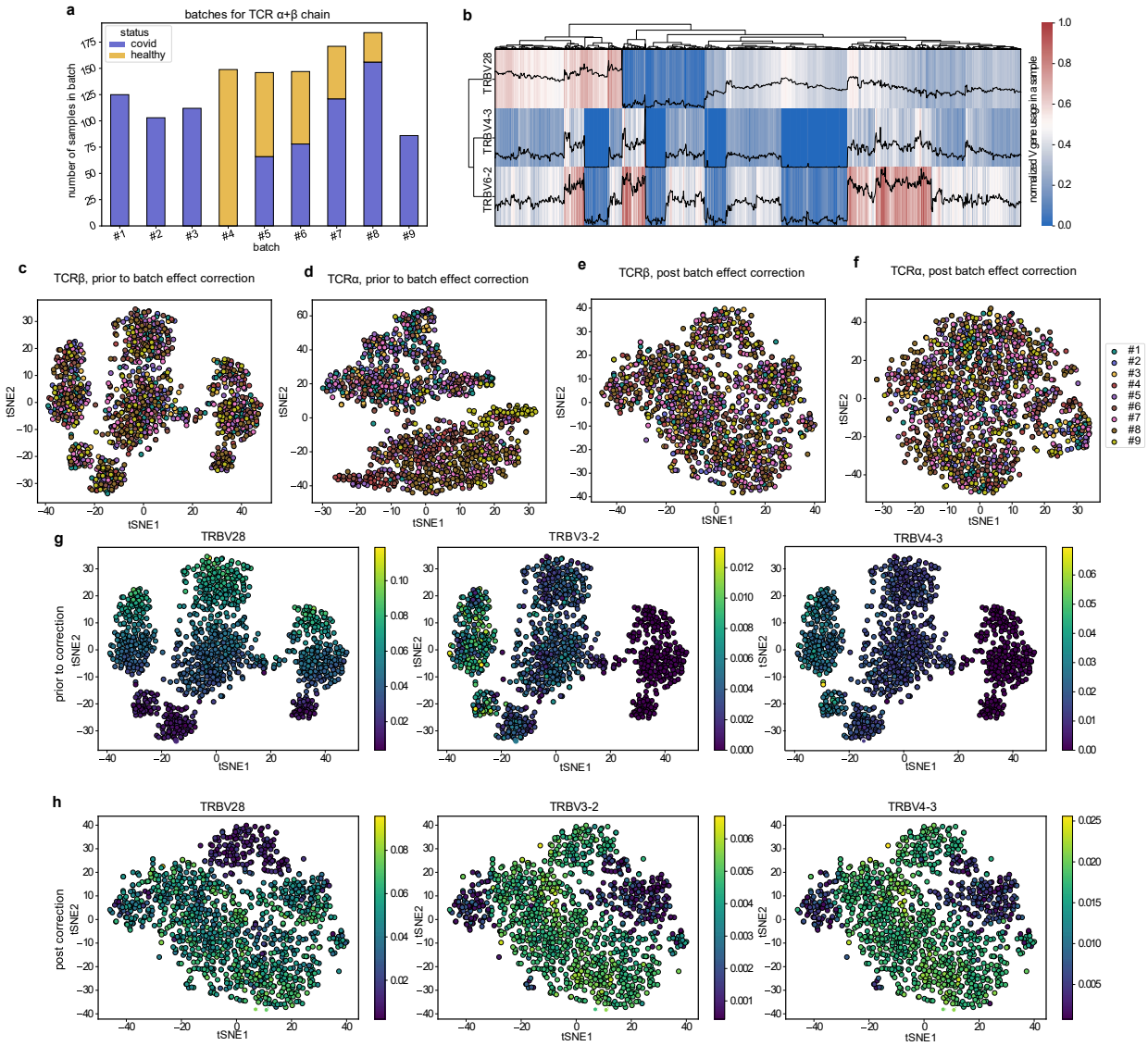

Figure 2. Batch effect and genotype related differences in Variable gene usage. A. A bar plot of total number of subjects, number of COVID-19-convalescent and healthy cases for 9 batches of Cohort I reported in this study. The distribution is shown for samples that pass the sequencing depth cutoff for both TCRα and β chains. **B.** Heatmap of gene usage showing the imprint of TRBV28/3-2/4-3 haplotype in the dataset. Black lines show gene usage Z scores for each of three genes. **C, D.** Visualization of batch effects using t-SNE. Plots show similarity of TRBV (**C**) and TRAV (**D**) gene usage profiles between samples, samples are colored by batch. **E, F.** Same as **C, D.**, but for TRBV (**E**) and TRAV (**F**) gene usage profiles post batch effect correction. **G, H.** Coloring of TRBV gene usage t-SNE plots **C** and **E** according to the frequency of TRBV28, TRBV3-2 and TRBV4-3genes.

Similar to many large-scale studies involving hundreds of subjects, our dataset was prepared in several batches. These batches were both uneven in the number of subjects and imbalanced in the ratio of COVID-19-convalescent patients to healthy donors (**Figure 2A, Supplementary Figure 2AB**). Thus, as TCR repertoire sequencing data is extremely high-dimensional and complex, even minor differences in library preparation and sequencing between batches can lead to biases in comparative analysis of repertoires and obfuscate COVID-19-related features. We utilize Variable and Joining gene frequency profiles (termed 'gene usage') of repertoires to monitor for these biases and biological effects.

We start with Variable gene usage clustering analysis and use standard multidimensional reduction approaches such as PCA and t-SNE for visualization. From the gene usage heatmap analysis one can spot a prominent and peculiar pattern involving TRBV28, TRBV6-2 and TRBV4-3 co-usage (**Figure 2B**). This pattern can be attributed to germline differences in TCRβ locus[29]: there is a well-known 21-kb deletion polymorphism involving both TRBV6-2 and TRBV4-3, while TRBV28 is involved in a single-gene deletion polymorphism. The same phenomenon was observed for TRBJ genes (**Supplementary Figure 3A**). As can be seen from **Figure 2C**, there is a three cluster pattern in TRBV usage t-SNE that seems to be unrelated to differences between batches. Further analysis shows that it is related to TRBV28 deletion in component 1 and TRBV6-2/TRBV4-3 deletion in component 2 (**Figure 2G**). Ideally, when correcting for batch effect, one should still be able to recover biological differences between samples such as aforementioned germline polymorphisms. On the other hand, TRAV usage demonstrates clear batch-related clustering, with samples from batches 1 and 8 are predominantly located on the positive and negative sides of component 1 in the t-SNE plot (**Figure 2D**). Batch effects were also detected for TRBJ and TRAJ genes (**Supplementary Figure 3CE**); for both chains, the batches sequenced later in time are clustered together (e.g. batches #8 and #9) (**Supplementary Figure 2BE**). The allelic effects were also visible for the TRAV and TRAJ genes, but we didn't find any previous reports of them (**Supplementary Figure 3CF**), which we attribute to the fact that most AIRR-seq studies focus on the β chain.

We proceed to the batch effect correction procedure described in detail in the **Methods** section. Briefly, we aim to normalize gene usage profiles across samples by using Z-scores calculated across each batch for log-transformed gene frequencies, these values are then transformed and re-scaled to reflect global gene usage hierarchy of the entire dataset. As one can see from **Figure 2EF, S2CF**, such normalization negates batch effects and samples are uniformly distributed on the t-SNE plot. On the other hand, our correction procedure leaves biological variability intact, as can be seen for the TRBV28, TRBV6-2 and TRBV4-3 case in **Figure 2H**. The biological variability is also preserved for TRAV and TRAJ genes as well as for TRBJ gene usages (**Supplementary Figure 3BDE**).

After correcting gene usage profiles one can re-sample clonotypes in the repertoires as described in **Methods** and proceed with comparative analysis of donor repertoires to search for COVID-19-associated clonotypes. We trained COVID-19 status classifiers based on V-J gene usage profiles alone, but did not succeed (**Supplementary Figure 4**).

Identification of COVID-19-associated TCR α and β clonotypes

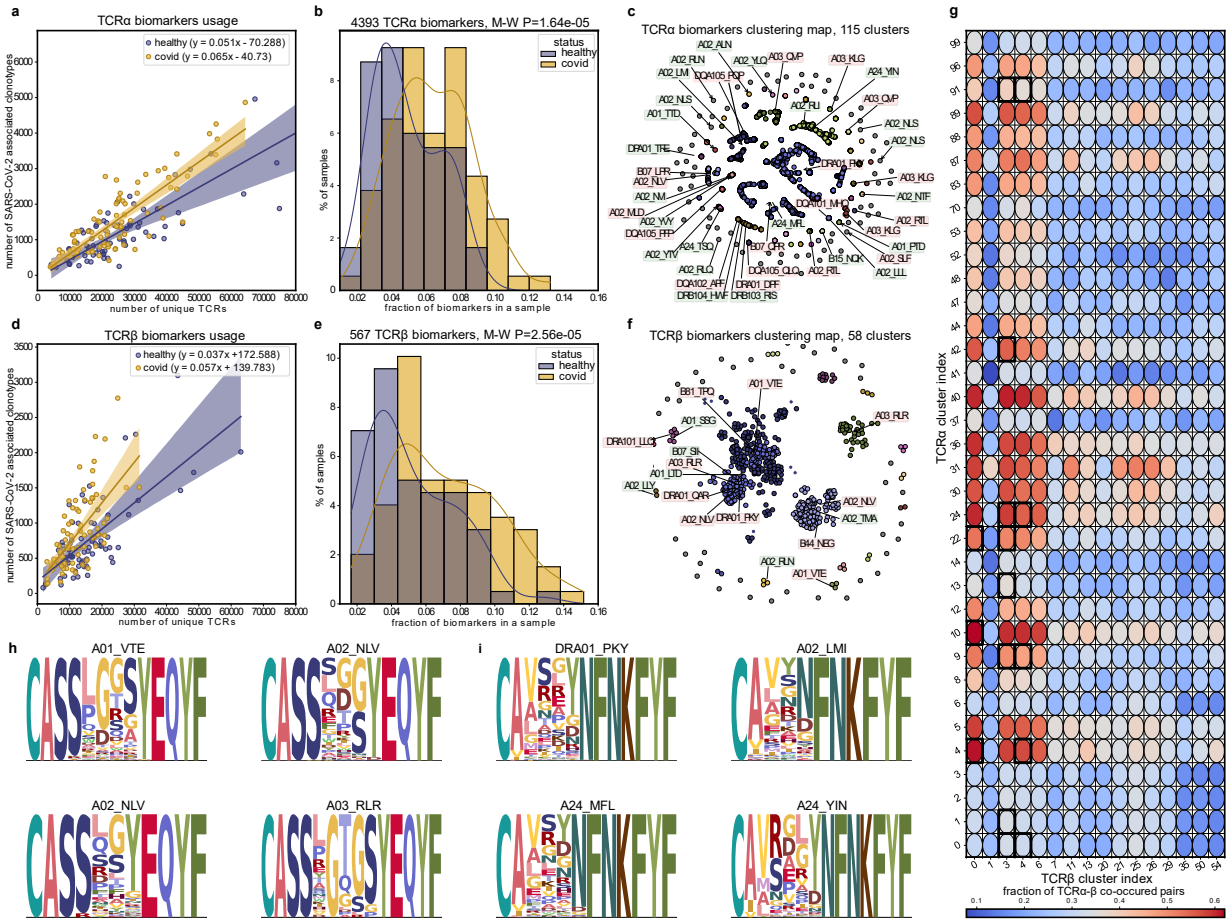

Figure 3. Assessment of COVID-19 TCR α and β biomarkers. A. Scatter plot showing the number of COVID-19-associated TCRα clonotypes (Y axis) plotted against the total number of clonotypes in each sample (X axis) from the validation batch. All numbers are given in terms of unique rearrangements, matching to associated clonotypes allows up to one amino acid substitution in CDR3. **B.** Distribution of COVID-19-associated TCRα clonotype fraction across healthy and convalescent samples. **C.** Layout of the CDR3 sequence similarity graph of COVID-19-associated TCRα clonotypes, edges (not shown) connect sequences with up to one amino acid substitution. Each connected component (cluster) is highlighted with its own color. Predicted antigen specificity according to VDJdb is shown with arrows and labels. The green labels correspond to SARS-CoV-2 epitopes, the red ones to the other viruses. The detailed list of associations is available in **Supplementary Table 1**. **D, E, F.** Same as **A, B, C** but for TCRβ chain clonotypes. The detailed list of associations is available in **Supplementary Table 2**. **G.** Co-occurrence of CDR3s in α and β chain clonotype clusters. Co-ouccurence was calculated as the Tanimoto index for the corresponding TCR α and β clusters. The color corresponds to the correlation coefficient between the usage of α and β cluster clonotypes in the specified clusters. Bold squares show α-β pairings that demonstrate significant association with the same antigen according to VDJdb. **H.,I.** CDR3 sequence logos of top four largest clusters in **C.,F.** (TCR α and β respectively).

We proceed with statistical inference of COVID-19-associated TCR clonotypes from batch-corrected repertoires. Inspired by previous work identifying Cytomegalovirus-associated TCR sequences[25] we perform Fisher exact test for the contingency table of presence/absence of a given TCR and positive/negative COVID-19 status with one critical modification described below. Note that at this step we consider all batches in our analysis: due to extremely high diversity of TCR repertoires and sparse overlap each batch will likely bring many novel biomarkers.

First, we made a short-list of candidate TCR α and β clonotypes that were rearranged at least twice, i.e. supported by at least two unique nucleotide variants or found in at least two samples in our entire dataset. After computing P-values for the association we ran false-discovery rate (FDR) correction for multiple

testing to select significant hits at q = 0.05. Surprisingly, no significant hits were found for TCRβ clonotypes. We suggest this is due to limited sampling depth of individual repertoires and low cell counts of individual TCR variants in a donor. We then modified our method to allow a single mismatch in the CDR3 amino acid sequence when searching for a clonotype of interest in the repertoire. This adjustment is justified by the fact that a single mismatch in CDR3 doesn't change antigen specificity most of the time[30]. The other rationale is that the estimate of clonotype population frequency computed this way is in perfect agreement with theoretical baseline V(D)J rearrangement probability[3]. After the adjustment many significant hits were obtained at q = 0.05.

For the final list of COVID-19-associated TCRs we use the q = 0.01 threshold arriving at 4887 and 574 significant TCRα and β biomarkers (**Supplementary Figure 6AE**). This almost 10-times difference in the number of variants for TCRα and β can be explained by the fact that TCRβ clonotypes are less public and more dissimilar due to the presence of Diversity gene, so finding a significant association for them is harder and requires more samples than for TCRα.

Basic analysis of the incidence rate of COVID-19-associated clonotypes shows that they constitute a detectable fraction in COVID-19-convalescent donor samples (24.1% and 6.8% on average of all clonotypes in TCRα and TCRβ repertoires respectively, **Supplementary Figure 5BF**) that is significantly higher compared to healthy donors ($P = 1.2x10^{-33}$ and $2.6x10^{-15}$ for TCRα and TCRβ respectively, Mann-Whitney test, **Supplementary Figure 5CG).**

Manual inspection of clonotype lists revealed that some of the found clonotypes had spurious CDR3 sequences that are either extremely short or too long, or non-canonical (CDR3 not starting with Cys and ending with Phe/Trp). We used the V(D)J rearrangement model[31] to compute the generation probability for each of the clonotypes in our list. We found outliers both for TCR α and β sequences (**Supplementary Figure 4DH**) that match spurious sequences detected manually and removed them, leaving 4393 and 567 TCRα and β biomarkers for further analysis. The results remain unchanged after spurious clonotype removal (**Figure 3AD** and **Figure 3BE**).

In order to check the robustness of our results to batch removal, we repeated the analysis by leaving out a validation batch. We were able to reproduce the results above: there is a detectable number of rearrangements matching COVID-19-associated TCRs in convalescent donors (**Supplementary Figure 5AE**) and those are significantly more frequent than those in healthy ones ($P < 10^{-4}$ for both α and β, Mann-Whitney test, **Supplementary Figure 5BF**).

Assessment of COVID-19-associated TCR biomarkers

We explored our list of COVID-19-associated TCR clonotypes to see if there are some characteristic motifs that may be linked to antigen specificity. Firstly, we calculated the CDR3 amino acid sequence similarity graph with a Hamming distance of 1 threshold for edges shown in **Figure 3CF**. The presence of prominent hubs in the graph suggests that the clustering is non-random and there may be large groups among identified biomarkers that share the same CDR3 motifs and have similar antigen specificity. After computing connected components we get 115 and 58 TCR α and β clonotype clusters (metaclonotypes), an order of magnitude reduction in the number of biomarkers compared to the original clonotype list.

We performed a search against the VDJdb database[32] allowing for a single amino acid substitution in the CDR3 region to query for potential antigen specificities of metaclonotypes. In order to find reliable specificity predictions we tested if the fraction of clonotypes associated with a given antigen in our cluster is higher than the fraction observed for the entire clonotype list (see **Methods**). In cases when multiple antigens are associated with the same cluster, we selected the most enriched one. By setting $P < 0.05$ threshold for Fisher exact test and choosing the epitopes with maximum enrichment score (see **Methods**) we found that 23 and 5 of TCR α and β clusters are reliably associated with SARS-CoV-2 epitopes (**Supplementary Data 1, 2**). Three matches to epitopes other than SARS-CoV-2 were found for β clusters and 27 matches to distinct epitopes for α clusters. The latter might be either false-positive hits that can arise

as we consider only one chain while specificity is defined by both, or bystander clonal expansions in COVID-19 such as the association of α cluster#3 with an epitope from HCoV-HKU1.

In order to bypass the limitation on the absence of TCR α and β pairing information in bulk T-cell sequencing we performed an *in-silico* chain pairing procedure by studying the correlation of α and β clonotype presence across the samples (see **Methods**). This approach has its own limitations, as it will not be able to detect pairing for extremely rare or extremely frequent metaclonotypes. The resulting heatmap of α and β pairing for our clusters is shown in **Figure 3G**. We next performed VDJdb-based antigen specificity prediction for pairs of clusters using only those VDJdb records that contain both TCR chains (see **Methods**). Interestingly, only strongly correlated TCRα and β clusters matched to the same epitope according to our annotation (**Figure 3G**). The majority of these epitopes comes from SARS-CoV-2 (**Supplementary Table 3**), yet there are some matches to Epstein-Barr virus epitopes and even neoantigens that may be attributed to cross-reactivity. Characteristic patterns of CDR3 amino acid sequences for four largest α and β clusters are shown in **Figure 3HI**.

The database of biomarkers is available as a Supplementary and can be accessed via https://covidbiomarkers.cdr3.net/.

Building COVID-19 status classifier using TCRα and TCRβ biomarkers

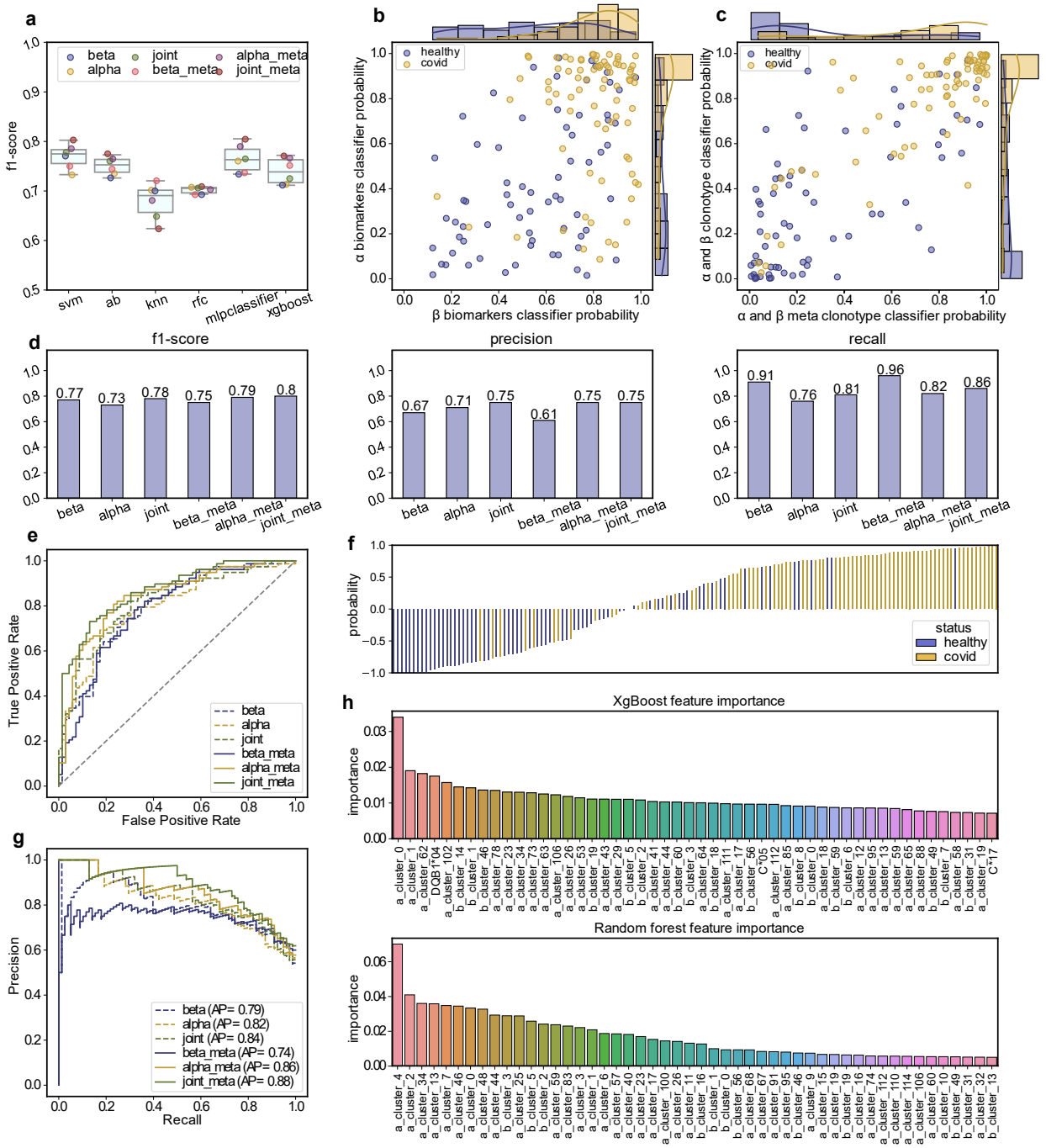

Figure 4. Analysis of machine learning approaches applied to Cohort I TCRβ and TCRα biomarkers. Here batch#6 was left out for validation purposes in all the figures, while the classifiers were trained using all the other batches **A.** Comparison of ML models f1-scores produced for each feature set. **B.** The scatterplot describing the relation of the probabilities to label the sample as the COVID-19 positive for TRCα and TCRβ based classifiers. **C.** The scatterplot describing the relation of the probabilities to label the sample as the COVID-19 positive for classifier based on both TRCα,TCRβ biomarkers and TRCα,TCRβ metaclonotype cluster features. **D.** Distribution of target metrics (f1-score, precision, recall) for all the evaluated models. **E.** The Roc-curve for SVM based classifiers for all sets of features. **F.** The waterfall plot represents the probability of each sample to be labeled as COVID-19 positive (> 0) or healthy (< 0). Samples coming from healthy donors are colored with

blue, COVID-19 samples are colored with orange. **G.** Precision-recall curve for SVM based classifiers. Different biomarker sets are assessed. **H.** Feature importance plot for the XGBoost/RandomForestClassifier models based on TRCα and TCRβ based meta-biomarkers and HLA presence features.

We proceed with TCRα and β chain biomarkers obtained in the previous section to build an accurate COVID-19 status classifier. We consider both clustered biomarkers (metaclonotypes) and unclustered ones (clonotypes) as features, training for either one of the chains or both. Note that here we consider the biomarker list obtained for all batches in order to provide uniform coverage of repertoires: leaving some batches out will decrease classifier recall (not shown). Among possible parameters of a classifier is the choice of biomarker frequency encoding and the addition of binary-encoded HLA features. Biomarker frequency used in the classifier can be encoded as boolean (present or absent), categorical (supported by 0, 1, 2 or 3+ reads) and real (fraction of reads) value. One can also add binary-encoded HLA information expecting that biomarkers have different feature importance depending on the HLA context. As can be seen from training an example SVM classifier (**Supplementary Figure 7**), real biomarker frequency encoding gives the highest accuracy, while there is almost no effect from the addition of HLA information. The non-informativeness of HLAs for our list of biomarkers can be explained by the fact that they were obtained from the entire dataset with a diverse mixture of HLAs and thus had to be robust to donor HLA haplotype (this is supported by observations for classifier training on allele-restricted set of samples described later in the text).

As some batches contained exclusively healthy or COVID-19-convalescent individuals, we combined them to balance the number of cases as described in **Methods**. We also held-out batch#6 that is relatively balanced to monitor for batch effects in classifier training. We considered 6 commonly used classifier models and measured the F1 score that is robust to case imbalance (see **Methods**) to select the optimal classifier based on training with various sets of features (using a single or both chains, considering metaclonotype biomarkers). As can be seen from **Figure 4A**, support vector machine (SVM) shows optimal performance, tied with multi-linear perceptron (MLP) having slightly lower median score. The former was chosen for in-depth evaluation of classifier performance. As one can observe from **Figure 4A**, using both α and β chains simultaneously results in a highly accurate classifier, suggesting that biomarkers from each chain convey additional information that benefits overall classification. This is supported by the fact that classification results obtained using α and β chain biomarkers are independent (**Figure 4B,** Pearson R=0.44). Interestingly, using metaclonotypes in αβ classifier provides far better separation between COVID-19-convalescent and healthy cases (**Figure 4C**), perhaps owing to less sparsity in feature values. All these factors result in optimal classifier metrics and ROC curve (**Figure 4DE**), leading to near-perfect sample classification shown in the waterfall plot (**Figure 4F**).

We additionally assessed the impact of batch effect on classifier performance. We performed a leave-one-batch-out cross validation procedure to compute classifier accuracy and focused on the difference between combined batches (batch#4+1, batch#4+2, batch#4+3, batch#4+9), batches #5, #7 and #8, and held-out batch#6 (**Figure S8**). As expected, combined batches demonstrate higher score due to batch effect, however, batch#6 that was not used at the biomarker selection step was classified with accuracy similar to batch #5, #7 and #8, suggesting that our set of biomarkers is robust.

We analyzed the feature importance of the TCRαβ metaclonotypes-based classification by training XgBoost and RandomForest classifiers (**Figure 4H**). RandomForest almost exclusively ranks α chain features as being most important, with a couple of β chain features showing medium to low importance. XgBoost model, on the other hand, ranks many β chain features as being important (17 out of 58 β metaclonotypes among top-50 most important features), and additionally includes 3 HLA-related features. One can suggest that XgBoost ranking of β and HLA features is due to its ability (as a boosting model) to highlight features that are important for small subsets of data.

Applying COVID-19 status classifier to an independent dataset

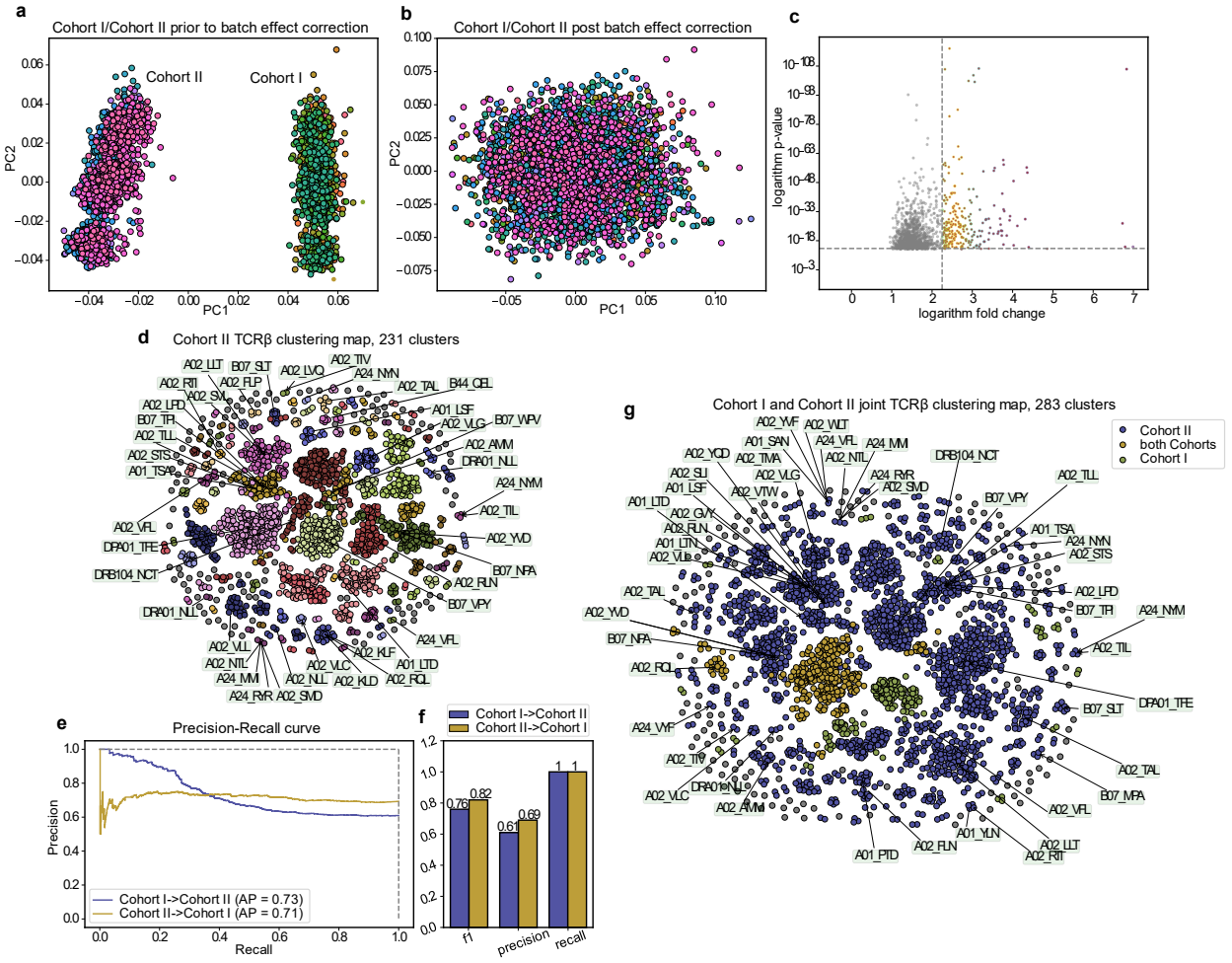

Figure 5. Comparative analysis of classifiers between cohorts. A, B. Visualization of batch effect prior (A) and post (B) correction procedure for TRBV genes. Color shows sample batch. **C.** The Fisher test based approach resulted in >20k statistically significant clonotypes found. The volcano plot was used to select the clonotypes with the highest fold change. **D.** The layout of homology graph of CDR3 sequences of COVID-19-associated TCRβ clonotypes for Cohort II cohort. For each cluster which is associated with SARS-CoV-2 the most probable epitope is marked. **E.** Comparison precision-recall curves for Cohort I/Cohort II based classifiers. The Cohort I model was trained using Cohort I data and tested using Cohort II data. The Cohort II based model was trained using Cohort II dataset and tested for Cohort I. AP corresponds to the average precision score. The detailed list of associations is available in **Supplementary Data 3**. **F.** Comparison of metrics for the models described in **Figure 6E**. **G.** The homology graph of CDR3 sequences of COVID-19-associated TRB clonotypes for both Cohort I and Cohort II TCRβ biomarkers. Cohort I and Cohort II clusters overlap forming 5 joint clusters which means that the selected clonotypes are similar to each other between platforms. The clusters which contain clonotypes from both platforms are colored with orange. The detailed list of associations is available in **Supplementary Data 4**.

We use a previously published data compendium by Adaptive Biotechnologies[8] (Cohort II) to validate our classifier and see if we can train an efficient classifier on Cohort II and apply it to Cohort I. The dataset contains 777 healthy and 1214 COVID-19 samples prior after coverage filtering **(Supplementary Figure 1CD)** and is limited to TCRβ chain filtering. The healthy donors were taken from HIP and KECK cohorts which were initially used for CMV status research[25]. The main problem with this dataset is that healthy and COVID-19 convalescent samples come from different batches sequenced several years apart, which can

result in a strong batch effect. Moreover, this dataset was generated using a proprietary protocol and primers that are different to the ones used in present study, so a strong batch effect is also expected to take place when comparing Cohorts I and II. We applied the batch effect correction technique discussed previously and resampled the clonotypes afterwards (**Figure 5AB**). As the average number of reads (sequencing depth) in Cohort II is far larger than in Cohort I, each sample in this cohort was down-sampled to 50,000 reads.

We repeated the feature selection pipeline for the resampled Cohort II data and found 17185 TCRβ biomarkers significantly associated with COVID-19 status (compared to ~500 in Cohort I). However, due to the inability to decouple COVID-19 status and batch we suspect the majority of these features reflect batch effect. To narrow down the biomarker list we used a log enrichment (observed to expected incidence ratio in COVID-19-convalescent samples) threshold of 2.5 resulting in 2066 biomarkers (**Figure 5C**). Interestingly, there were no spurious clonotypes and clonotype incidence rates were in good agreement with those predicted by the theoretical model of V(D)J rearrangement (**Figure 5D**).

We proceed with the similarity clustering of TCRβ clonotypes and arrive at 231 metaclonotypes, 33 of which were specific to SARS-CoV-2 epitopes according to VDJdb (**Figure 5E**). The fraction of SARS-CoV-2 epitope-specific metaclonotypes is 14% which is higher than 8.6% observed for TCRβ in Cohort I. We additionally checked whether TCRβ clonotype biomarkers identified separately for Cohort I and II have similar sequences and observed that 5 clusters of clonotypes are mixed which is a good proof that the found clonotypes are similar between the platforms. (**Figure 5H**).

We next test the performance of a classifier trained for Cohort I on Cohort II data and vice versa. First, we took an existing TCRβ classifier trained for 5670 biomarkers of Cohort I data and applied it to Cohort II with prior batch effect correction between cohorts. In this case we did not downsample Cohort II samples to lower the number of reads as we expect that the Cohort I biomarkers will be quite rare and require additional depth to be detected in Cohort II. Unfortunately, 15 of the described biomarkers were still absent in all the samples of Cohort II. This approach resulted in the validation f1-score equal to 0.76, precision score of 0.61 and recall 1 for the preselected model and an optimal precision-recall (PR) curve (**Figure 5FG**). Conversely, we trained a SVM classifier based on Cohort II biomarkers and downsampled data and applied it to Cohort I. In this case we got almost similar performance results, yet the PC curve showed suboptimal average precision (**Figure 5FG**). Again, one can suggest that the design of Cohort II where COVID-19 status cannot be decoupled from batch does not allow training an optimal classifier.

HLA association analysis reveals a unique COVID-19 biomarker motif

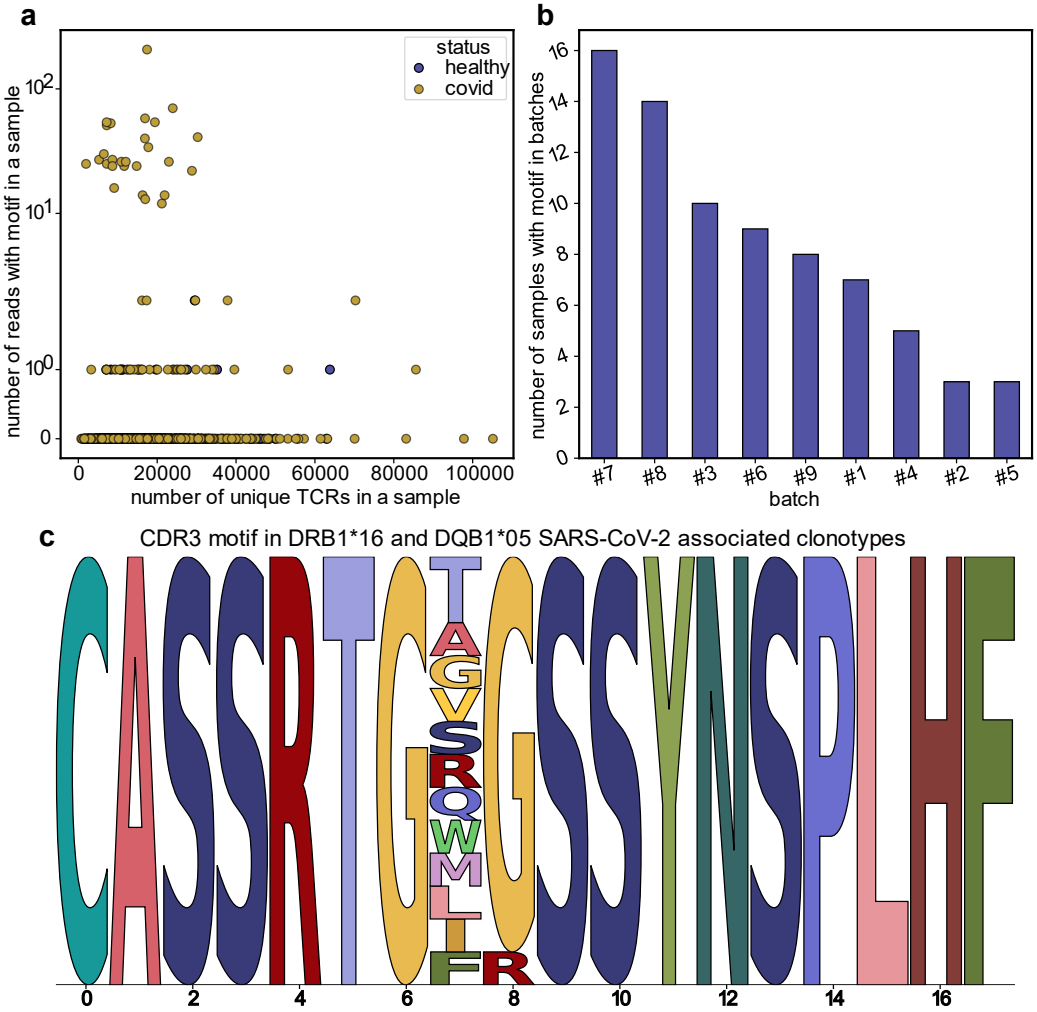

Figure 6. Analysis of DRB1*16, DQB1*05 and COVID-19 associated TCRβ biomarkers. A. The distribution of number of motif biomarkers compared between healthy and COVID-19 cohorts. **B.** The number of samples containing biomarkers with motif in each batch of Cohort I data. **C.** The weblogo describing the pattern of DRB1**16, DQB1**05 and COVID-19 associated TCRβ biomarkers.

Imprint of past and present infections on TCR repertoires is associated with individual HLA haplotypes as the set of antigens available to T-cells is shaped by HLA restriction[33]. The same should be the case for COVID-19 so we decided to search for TCRs that were both donor status and HLA-associated as was suggested previously[28]. HLA-associated clonotypes were found using the Fisher exact test comparing incidence in individuals having a specific allele and those without it. COVID-19 association was then calculated for the subset of samples having HLA of interest. After obtaining lists of HLA-associated clonotypes and COVID-19-associated clonotypes for every HLA we intersected them to select the list of HLA-restricted biomarkers. This allowed us to explore rare clonotypes present in COVID-19 patients depending on HLA context.

This approach led to identification of 13 TCRβ biomarkers that are exclusive to COVID-19 patients among donors with HLA-DQB1*05 or HLA-DRB1*16 alleles (**Supplementary Figure 9A**), this pattern of clonotypes were found in 75 donors. The presence of these biomarkers is almost always associated with positive COVID-19 status (63 out of 75 donors were infected with SARS-CoV-2) in donors (**Figure 6A**) allowing straightforward classification of a fraction of convalescent donors with common HLA allele (**Figure 6B**). Moreover, 11 of the 12 healthy donors with clonotypes matching the found motif had only

one read with the clonotypes of interest. It's also interesting to note that having dropped all the samples having only one read with corresponding pattern the HLA linkage becomes even more evident (**Supplementary Figure 9B**).

Interestingly, while the search was performed using the CDR3 region that contains only few amino acids (CASS subsequence that is common across many TRBV alleles), all of these clonotypes are rearranged using TRBV12-3. The amino acid motif corresponding to the found clonotypes has only one variable position (position 8), whereas all the other positions remain fixed (**Figure 6C**). It is also interesting to note that the presence of both HLA-DQB1*05 and HLA-DRB1*16 (which have high linkage disequilibrium) is associated with various autoimmune diseases[34,35].

Separate HLA classifiers provide a better classification quality

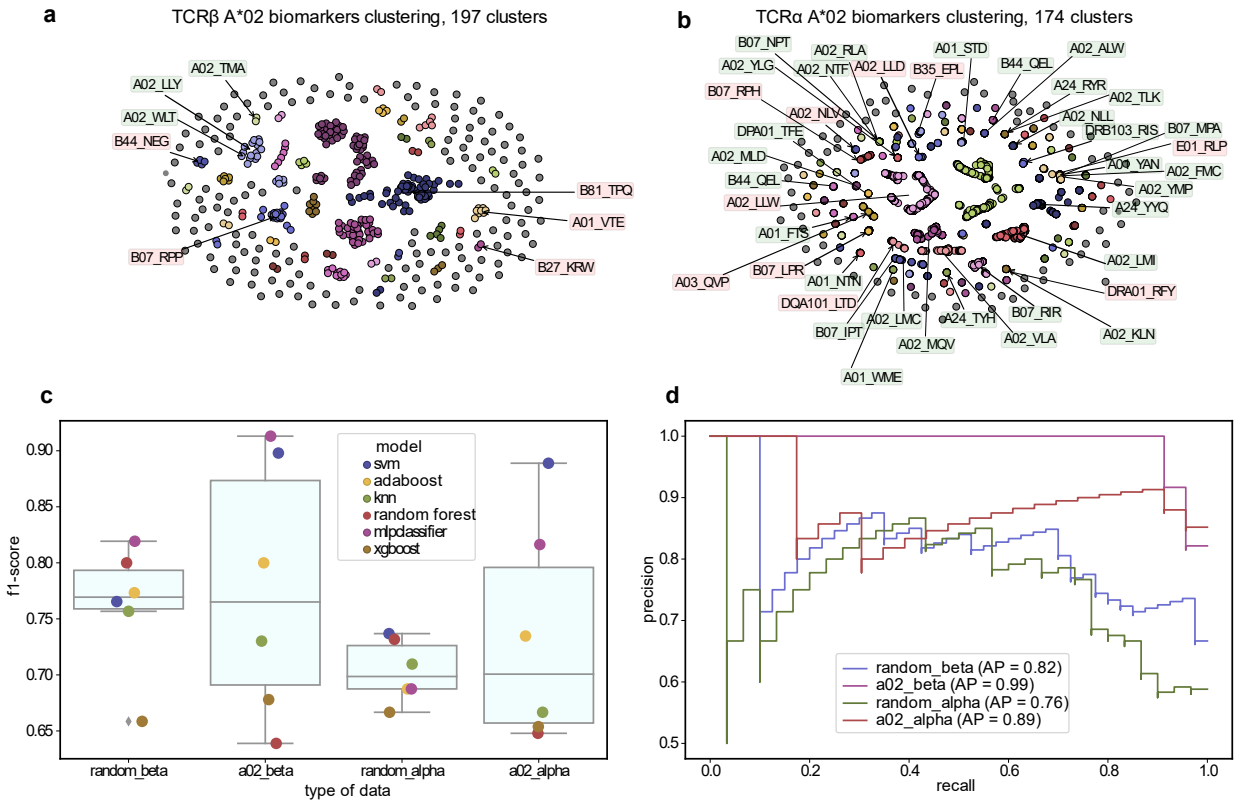

Figure 7. Comparison of classifiers built using biomarkers derived from random samples and A*02+ samples. A. The layout of homology graph of CDR3 sequences of COVID-19-associated TCRβ clonotypes derived using only A*02 positive samples. For each cluster which is associated with SARS-CoV-2 the most probable epitope is marked. The epitopes marked with green are the SARS-CoV-2 epitopes, while the red ones are the epitopes belonging to other viruses. The detailed information on cluster associations is available in **Supplementary Data 5**. **B.** Same as **A**, but for TCRα biomarkers. Only matches for clusters of size >9 are shown. For more information see **Supplementary Data 5**. **C.** Comparison of model performance for different data sets. Random marker corresponds to classifiers built using biomarkers derived from random 545 and 521 samples for TCR α and β chain respectively in the cohort, whereas a02 corresponds to classifiers built for A*02+ derived biomarkers and samples. **D.** Precision-recall curve for SVM classifiers built for the datasets described in **C.** The plot shows that the classifiers made using single HLA allele samples and single HLA allele SARS-CoV-2 associated clonotypes provide a better classification.

While our previous attempts to incorporate HLA information in COVID-19 classifier didn't result in any gains in accuracy, we suggest that it might be due to limited cohort size and the presence of a large number of rare HLA alleles. Thus we decided to directly check if using HLA information can improve classifier performance by selecting samples carrying a certain HLA allele for our pipeline. We chose HLA-A*02+ as it is among most frequent HLA alleles and there were 545 positive donors in Cohort I for TCRα chain and 521 positive donors for TCRβ chain. We repeat our COVID-19 association analysis resulting in 3277 and 358 TCR α and β clonotype biomarkers that display the presence of large hubs in TCR similarity graph that can be clustered into 174 and 197 TCR α and β meta-clonotypes (**Figure 7AB)**, the majority of which are found to be specific to HLA-A*02-restricted epitopes according to VDJdb annotation.
Next, we used HLA-A*02-restricted biomarker set to train classifiers for HLA-A*02+ subset of the cohort resulting in highly accurate COVID-19 status predictions. To perform a proper comparison with the HLA-unrestricted classifier we repeated the whole procedure described in this section for a randomly chosen set of 545 and 521 donors for TCR α and β chains respectively from Cohort I selected regardless of their HLA. The resulting classifier performs worse than the classifiers for our entire cohort as we decreased the dataset size almost twice. Comparing both classifiers shows superior performance of the HLA-restricted classifier (**Figure 7CD**) suggesting that HLA information can greatly boost classifier performance when one is provided with sufficient number of samples with a given HLA and is correctly handling them. Moreover, our approach can be utilized for accurate HLA allele prediction for frequent alleles (**Supplementary Figure 10**).

Classification based on published SARS-CoV-2 epitope-specific TCR sequences

Our previous results indicate that many of inferred biomarkers are similar to TCR sequence variants known to recognize SARS-CoV-2 epitopes according to previously published studies. Thus, we decided to test whether we can create an accurate COVID-19 status classifier by solely relying on publicly available TCR sequences as biomarkers.
We start with 2909 and 3816 SARS-CoV-2 epitope-specific TCR α and β sequences from the VDJdb database[32] as a biomarker set. Overall, there was no significant difference in incidence of clonotypes matching with up to one amino acid substitution to VDJdb biomarkers in convalescent donors compared to healthy ones. We decided to narrow the set and proceed with Fisher test for COVID-19 status association (**Supplementary Figure 11AD**) and observed that there was no significant difference for the majority of clonotypes as most of them are found in only a handful of samples in the dataset. Nevertheless, we were able to select 261 TCRα (**Supplementary Figure 11EF**) and 348 TCRβ (**Supplementary Figure 11BC**) biomarkers for further analysis.
We evaluated the classifiers based on this set of biomarkers and found that it yields classification accuracy of F1 between 0.6 and 0.7 that is substantially lower compared to the classifier built using Cohort I data displaying F1 between 0.7 and 0.9 (**Supplementary Figure 11G-J**). However, one can still see the improvement of classification metrics in the paired-chain classifier model compared to single-chain one. These results suggest that the number of publicly available TCR specificity records is still too small to provide means for accurate COVID-19 status prediction in a large and diverse population of donors.

Materials and Methods

Cohort collection

This study is a multicenter cross-sectional stratified epidemiological research. The study was conducted in FMBA medical organizations in 54 regions of the Russian Federation. The design of the study assumed the inclusion of volunteers in two cohorts. The first cohort (control group) included conditionally healthy people, without signs of respiratory diseases at the time of inclusion in the study, with a SARS-CoV-2 PCR negative test and samples of healthy individuals collected before the COVID-19 pandemic ("precovid"). The second cohort (core group) included people with COVID-19, with a positive test for SARS-CoV-2, lung CT findings, as well as with a history of diagnosis. The core group included participants with different severity of COVID-19: asymptomatic, mild, moderate, severe, extremely severe patients, as provided in the recommendations of the Ministry of Health of the Russian Federation. The core group included patients that have had no more than 28 days since the diagnosis of COVID-19. A total of 1,253 participants were included in this study 392 healthy volunteers were included in the control group and 861 patients with COVID-19 in the core group in the period from June to October 2020.

All participants underwent a medical inspection, physical examination, collection of blood samples, blood serum and oropharyngeal swab. The samples were stored at a temperature of +4 -+6°C no more than 2 hours after sampling and transfer to the laboratory, then in a low-temperature refrigerator at a temperature of -80°C before the analysis. The samples were transported at a temperature no higher than -80°C in thermobox with dry ice.

Ethical considerations

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Local Ethics Committee of the Federal State Budgetary Institution "Centre for Strategic Planning and Management of Biomedical Health Risks" of the Federal Medical Biological Agency (Protocol No. 2 of May 28, 2020). All participants provided informed consent.

DNA isolation

For HLA typing isolation of genomic DNA (gDNA) from whole blood samples was performed manually using DNA Blood Mini Kit (Qiagen, Germany) in accordance to the manufacturer's protocol. The yield and purity of the isolated gDNA were manually determined using Qubit 4.0 fluorimeter (Thermo Fisher Scientific, USA) and NanoDropTM One C Microvolume UV-Vis (Thermo Fisher Scientific, USA), respectively. Only gDNA samples with absorbance ratios A260/280 of 1.7-1.9 and A230/260 of 1.8-2.2, in accordance, were selected for further analysis.

For TCR library preparation 200 ul of frozen whole blood was used for genomic DNA isolation using Amplitech E1 automatic sample preparation station (LLC "Amplitech", Russia). The first step was lysis of blood samples, followed by washing. The gDNA was eluted with 50 ul of TE buffer. The DNA concentration was manually determined using Qubit 2.0 fluorimeter (Thermo Fisher Scientific, USA).

HLA-typing

A total of 150-500 ng of gDNA was used to prepare Next-Generation Sequencing (NGS) libraries. Libraries were prepared using Illumina DNA Prep kit (Illumina, Inc. USA) according to the manufacturer's recommendations using Tecan Freedom EVO robotic station (Tecan, Switzerland). The size of the resulting libraries was determined with Agilent D1000 reagent kit using Agilent 4200 TapeStation (Agilent Technologies, Inc., USA). Pooling was performed automatically using a Tecan Freedom EVO robotic station (Tecan, Switzerland). Pool quality control was performed with Agilent High Sensitivity D1000 Screen Tape reagent kit using Agilent 4200 TapeStation (Agilent Technologies, Inc., USA). Whole-genome sequencing (WGS) was performed using the Illumina NovaSeq 6000 with the reagent kit S4 (300 cycles) (Illumina, USA) with 2 × 150 bp paired-end reads. The samples were typed at 6 main loci of HLA classes I and II: HLA-A, -B, -C, -DRB1, -DPB1, -DQB1 with a resolution of 2 fields with the xHLA bioinformatic software[36].

TCR library preparation

The sequencing libraries for TCR α and β chains profiling were prepared using Human TCR DNA Multiplex kit (MiLaboratories, USA) according to the manufacturer's protocol. Briefly, 600 ng DNA (equivalent to 92,000 cells genomes with expected ~50% related to T cells) total extracted from whole blood was used as input in two separate 50 µl multiplex PCR replicates containing 1X Turbo buffer, 10 units of HS Taq polymerase, 200 µM of each dNTP (all Evrogen, Russia) and 1X TR hum DNA PCR1 primer mix from Human TCR DNA Multiplex kit. The amplification profile was 94°C for 3 min, followed by 8 cycles of 94°C for 20 s, 57°C for 90 s, and 72°C for 40 s, followed by an additional 17 cycles of 94°C for 20 s and 72°C for 80 s. Denaturation steps (94°C) were performed with Ramp 4°C/s, and other steps were performed with Ramp 0.5°C/s. 10 µl of each replicate amplicon were pooled and purified using 1V AmPure XP beads (Beckman Coulter) and used as a template for the second PCR. The second (indexing) 25 µl PCR contained 1X Turbo buffer, 2.5 units of HS Taq polymerase, 200 µM of each dNTP, and 1 µl of Unique dual index primers (Illumina, USA). The amplification profile was as follows: 94°C for 3 min, followed by 15 cycles of 94°C for 20 s, 55°C for 20 s, and 72°C for 40 s. Obtained amplicons were purified using 0.8V AmPure XP beads (Beckman Coulter, USA), pooled, and sequenced on the Illumina NovaSeq 6000 system in 150PE mode.

External datasets

Data for the Adaptive Biotechnologies cohort was obtained from *Emerson et al.*[25] (HIP and KECK batches, pre-COVID-19 control) and *Nolan et al.*[8] (convalescent donor data).

Data processing and analysis

All datasets were processed with MIGEC and MiXCR software as reported previously[27,37]. Clonotype sequences for Cohort-II datasets were re-mapped using MiXCR to keep consistent V(D)J naming across datasets. Clonotype tables were formatted and standardized using VDJtools software tool[38]. Formatted datasets can be downloaded from Zenodo[39], the list of TCR biomarkers can be downloaded at https://covidbiomarkers.cdr3.net/.

All data analysis was performed using in-house R and python scripts available at GitHub (https://github.com/antigenomics/tcr-covid-classifier), the analysis can be reproduced by running a Snakemake workflow stored in the repository.

Batch effect correction and data normalization

We focus on sample preparation and sequencing biases manifested in the difference in frequency of Variable and Joining gene rearrangements across batches and normalize samples as follows. We assume that gene usage (P, fraction of reads mapping to a given gene in a given sample) is log-normally distributed for each gene across samples in a given batch with a specific mean and standard deviation $P(gene, sample) \sim LogNormal(\mu, \sigma \mid batch)$. We therefore log-transform gene usage and compute Z-scores $Z = (\log P - \mu) / \sigma$ for each gene so that the distribution of gene usage becomes a standard normal distribution $N(0, 1)$ across all samples.

As we perform the analysis on clonotype level, we need means to re-define and correct raw clonotype frequencies f_i, therefore put Z-scores back into the $[0, 1]$ interval so that we can use them to scale clonotype frequencies. We use sigmoid transform for this purpose, adjusted to preserve the original average gene frequency across all batches $P_{avg}(gene)$ to preserve gene usage hierarchy. We can compute target frequencies by using the following formula $P_{final}(gene, sample) = 2 \times P_{avg}(gene) / (1 + \exp[-Z(gene, sample)])$.

We can then correct frequencies of clonotypes f_i from sample S rearranged from gene G as $f_i' = f_i \times P_{final}(G, S) / P(G, S)$. We next adjust the sample clonotype composition according to new frequencies. As our frequency adjustment can lead to non-integer clonotype counts, straightforward sampling from a multinomial distribution is not feasible. We therefore randomly sample numbers from a uniform distribution $U[0, 1]$ equal to the read count in the sample and perform a roulette-wheel selection for clonotypes according to their adjusted frequencies f_i'' normalized to a sum of 1.

In-silico α-β pairing of clonotype groups

This approach is inspired by the work of Howie *et al.*[40] and extended by addition of fuzzy CDR3 matching and working with clonotype groups instead of clonotypes for robustness. We estimate the pairing between selected TCR α and β chain clonotype groups as follows. For a given group of clonotypes g we estimate the number of matching clonotypes in a given sample s as X_{gs} that is equal to the number of CDR3 amino acid sequences in the sample that match at least one sequence in g with up to one amino acid mismatch allowed (*i.e.* Hamming distance less or equal to one).

Let us consider i^{th} group of α chain clonotypes and j^{th} group of β chain clonotypes. We can compute the Pearson correlation ρ_{ij} of X_{is} and X_{js} across all samples and use it as a score for pairing between these two groups. Alternatively, one can compute $n_i = \sum_s X_{is} > 0$ and $n_{ij} = \sum_s X_{is} > 0 \wedge X_{js} > 0$ and define the odds ratio of detecting a given pair as $\theta_{ij} = (n_{ij} \times n) / (n_i \times n_j)$ where n is the total number of samples. Note that the potential of in-silico pairing is limited: in case of a too frequent group having high incidence across samples it will be paired with almost all groups of another chain, while for a rare group there may not be enough statistical power to detect a matching group.

Clonotype group antigen specificity annotation

We annotated clonotype groups with putative antigen they recognize by matching TCR α and β sequences to antigen specificity records from the VDJdb database[32]. In general, we consider a match between CDR3 sequences up to a single amino acid substitution to a database record to indicate that a given TCR sequence recognizes a corresponding antigen. This way we arrive at the putative set of antigens $\{A\}_G$ for a given clonotype group G.

The overall matching between a given clonotype group and antigen A can be assessed as follows: let n_A be the number of matching clonotypes in a group of interest, m be the size of group of interest, N_A be the number of matches in all groups and M be the total number of clonotypes in groups, then we can use Binomial distribution with parameters n_A, m and $p = N_A / M$ to test the significance of overlap n_A / m as described in *Pogorelyy et al.* [28].

For a given α and β chain clonotype group pair $(G\alpha, G\beta)$ we can narrow down the search to the intersection of their putative antigen sets. We can exploit the true α and β pairing information present in VDJdb to find high-confidence antigen matches for pairs of clonotype groups as follows. Let us consider only paired records in VDJdb for an antigen A, we next iterate over all pairs of TCR α and β sequences in our clonotype group pair and see if it matches to any paired database record with a single amino acid substitution in both CDR3s of the chains allowed. Counting the extent of pairing x_A, i.e. the total number of TCR α and β sequences matched to A, we can compare it to the expected value by using Hypergeometric test with the maximum potential number of matches $X_A = min(|G\alpha|, |G\beta|)$, the number of paired records belonging to antigen A in VDJdb and the total number of paired records in VDJdb.

Selecting COVID-19 classifier model

A set of commonly used machine learning approaches was compared in their ability to classify COVID-19 status using TCR repertoire data, a set of pre-selected TCR features and HLA haplotype: SVM, AdaBoost (referred to as *ab*), KNN, Random Forest Classifier (referred to as *rfc*), Multi Linear Perceptron Classifier (referred to as *mlpclassifier*), XgBoost. These classifier models were compared according to their performance and *ab* and XgBoost were additionally used to analyze feature importance.
Grid search based on 5-fold cross validation was used to optimize the parameters of each classifier model for the training data that includes all batches except 6. Batch#6 was used as a verification set, it was held out and not included in classifier model selection and training.
As the data was still imbalanced, we focused on the F1-score (harmonic mean of the precision and recall metrics) and precision-recall (PR)-curves to assess classifier performance. The best model was chosen by the maximum F1-score on the test sets.
We obtained final metrics for best models using leave-one-batch cross validation. Here, as batches 1-3 contain no negative cases and almost all cases in batch 4 were positive, we constructed merged batches batch14, batch24 and batch34 by evenly distributing batch 4 to the first three batches and used these batches instead of original batches 1-4.

Discussion

Since the advent of AIRR-seq technology there was always an expectation that it will yield a wealth of novel diagnostic biomarkers and will complement commonly used immunological assays in infectious disease and vaccine studies. However, both AIRR-seq protocol complexity, extreme diversity of antigen receptors, variability in the strength of immune response in different donors and the phenomenon of public (shared across multiple donors) and private immune responses make it extremely hard to employ this technique in a clinical setting. In the present study we used a large cohort of COVID-19-convalescent and healthy donors to design a proof-of-concept wet lab and bioinformatic pipeline based on AIRR data that can be employed to accurately stratify patients based on the history of this infection.

The main difficulty in the way of designing a universal pipeline based on a robust set of COVID-19 TCR biomarkers is the presence of strong batch effects in AIRR-seq data[23] that can lead to data overfitting and misclassification of donor samples. As was observed previously, most batch effects manifest in the form of skewed Variable and Joining gene usage profiles and presence of ultra-public TCR variants due to cross-sample contamination. Here we designed and applied a gene usage normalization strategy that was able to remove biases between batches and technologies making our data consistent and comparable with previously published datasets that is useful in cases when each donor group is sequenced in a separate batch[8]. Notably, this critical step can benefit research comparing AIRR data from different groups of donors and performing meta-analysis of different AIRR studies[41]. We also propose to use V(D)J rearrangement model software such as OLGA[31] to control for potential spurious clonotypes that can be either artifacts or contaminants. Importantly, we explicitly show that with our approach one can derive a set of biomarkers that can be transferred and applied to an independent COVID-19 dataset obtained for donors from a population with distinct demographics using a different AIRR-seq protocol.

The ability to sequence only one TCR chain is one of the major limitations of bulk AIRR-seq methods. Most studies limit their analysis to TCRβ sequencing data[42], while interaction of both chains is required for proper antigen binding[43] and there are multiple cases of prominent TCRα motifs recognizing immunogenic epitopes[41,44]. In the present study we clearly demonstrate TCRα and β chain biomarkers contain complementary information and joint analysis of both chains can greatly boost COVID-19 status classifier performance even in the absence of pairing data obtained at single-cell level[40]. The latter can be partially recovered by performing *in silico* pairing based on biomarker co-occurrence.

Given the fact that T-cells observe pathogens through a prism of HLA alleles presenting distinct peptides, we combined AIRR data reported in this paper with donor HLA typing. Our results show that in general, unsupervised feature selection leads to an HLA-agnostic set of biomarkers as classifiers built using them cannot be improved by including donor HLA information. On the other hand, dedicated classifiers trained using donors with shared HLA alleles (e.g. HLA-A*02 allele that is relatively common in population) show higher accuracy than those trained for sets of donors of similar size with mixed HLA alleles or for the whole dataset. We were also able to find relatively rare (in terms of incidence in donors) subset of biomarkers that is nevertheless tightly linked to donor HLA class II haplotype DRB1*16/DQB1*05 and COVID-19 positive status. Overall our findings suggest that HLA information can greatly benefit AIRR-based donor classifiers if handled properly[45,46].

The pipeline described in present study can be easily adapted to AIRR studies of various infectious diseases and pathologies. The set of biomarkers produced in this work can be applied to monitor T-cell memory for COVID-19 in donor samples and test if similar TCR sequences are induced by vaccination with SARS-CoV-2 antigens. We hope that further AIRR studies containing large and diverse cohorts of donors will

extend the TCR biomarker knowledgebase to other conditions and make AIRR-based diagnostics and monitoring more common in a clinical setting in the near future.

Acknowledgements

This work was supported by a grant from the Ministry of Science and Higher Education of the Russian Federation (075-15-2019-1789).

References

1. Immunobiology - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK10757/.
2. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135, 183–191 (2012).
3. Pogorelyy, M. V. et al. Exploring the pre-immune landscape of antigen-specific T cells. Genome Med. 10, 68 (2018).
4. DeWitt, W. S., III et al. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. eLife 7, e38358 (2018).
5. Rosati, E. et al. Identification of Disease-associated Traits and Clonotypes in the T Cell Receptor Repertoire of Monozygotic Twins Affected by Inflammatory Bowel Diseases. J. Crohns Colitis 14, 778–790 (2020).
6. Nakayama, M. & Michels, A. W. Using the T Cell Receptor as a Biomarker in Type 1 Diabetes. Front. Immunol. 12, 777788 (2021).
7. Dalai, S. C. et al. Clinical Validation of a Novel T-Cell Receptor Sequencing Assay for Identification of Recent or Prior Severe Acute Respiratory Syndrome Coronavirus 2 Infection. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. 75, 2079–2087 (2022).
8. Nolan, S. et al. A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. Res. Sq. doi:10.21203/rs.3.rs-51964/v1.
9. Schultheiß, C. et al. 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 (2020).
10. Wang, Y. et al. Analysis of TCR Repertoire by High-Throughput Sequencing Indicates the Feature of T Cell Immune Response after SARS-CoV-2 Infection. Cells 11, 68 (2021).
11. Kockelbergh, H. et al. Utility of Bulk T-Cell Receptor Repertoire Sequencing Analysis in Understanding Immune Responses to COVID-19. Diagnostics 12, 1222 (2022).
12. Peng, K. et al. Diversity in immunogenomics: the value and the challenge. Nat. Methods 18, 588–591 (2021).
13. Park, J. J. et al. Machine learning identifies T cell receptor repertoire signatures associated with COVID-19 severity. Commun. Biol. 6, 1–13 (2023).
14. Xu, J. et al. T cell receptor β repertoires in patients with COVID-19 reveal disease severity signatures. Front. Immunol. 14, 1190844 (2023).
15. Braun, J. et al. SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. Nature 587, 270–274 (2020).
16. Nersisyan, S., Zhiyanov, A., Shkurnikov, M. & Tonevitsky, A. T-CoV: a comprehensive portal of HLA-peptide interactions affected by SARS-CoV-2 mutations. Nucleic Acids Res. 50, D883–D887 (2022).
17. Lu, X. & Yamasaki, S. Current understanding of T cell immunity against SARS-CoV-2. Inflamm. Regen. 42, 51 (2022).
18. Zornikova, K. V., Sheetikov, S. A., Rusinov, A. Y., Iskhakov, R. N. & Bogolyubova, A. V. Architecture of the SARS-CoV-2-specific T cell repertoire. Front. Immunol. 14, (2023).
19. Zornikova, K. V. et al. Clonal diversity predicts persistence of SARS-CoV-2 epitope-specific T-cell response. Commun. Biol. 5, 1–11 (2022).

20. Khoo, W. H. et al. Tracking the clonal dynamics of SARS-CoV-2-specific T cells in children and adults with mild/asymptomatic COVID-19. Clin. Immunol. Orlando Fla 246, 109209 (2023).
21. Shomuradova, A. S. et al. SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. Immunity 53, 1245-1257.e5 (2020).
22. Trück, J. et al. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. eLife 10, e66274 (2021).
23. Barennes, P. et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nat. Biotechnol. 39, 236–245 (2021).
24. Breden, F. et al. Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. Front. Immunol. 8, (2017).
25. Emerson, R. O. et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat. Genet. 49, 659–665 (2017).
26. Springer, I., Tickotsky, N. & Louzoun, Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. Front. Immunol. 12, (2021).
27. Shugay, M. et al. Towards error-free profiling of immune repertoires. Nat. Methods 11, 653–655 (2014).
28. Pogorelyy, M. V. & Shugay, M. A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies. Front. Immunol. 10, (2019).
29. Omer, A. et al. T cell receptor beta germline variability is revealed by inference from repertoire data. Genome Med. 14, 2 (2022).
30. Shugay, M. et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res. 46, D419–D427 (2018).
31. Sethna, Z., Elhanati, Y., Callan, C. G., Jr, Walczak, A. M. & Mora, T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics 35, 2974–2981 (2019).
32. Goncharov, M. et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. Nat. Methods 19, 1017–1019 (2022).
33. Karnaukhov, V. et al. HLA binding of self-peptides is biased towards proteins with specific molecular functions. BioRxiv Prepr. Serv. Biol. 2021.02.16.431395 (2021) doi:10.1101/2021.02.16.431395.
34. Goni, E. et al. HLA-DRB1∗16 and -DQB1∗05 alleles are strongly associated with autoimmune pancreatitis in a cohort of hundred patients. Pancreatol. Off. J. Int. Assoc. Pancreatol. IAP Al 22, 466–471 (2022).
35. Testi, M. et al. Association of HLA-DQB1∗05:02 and DRB1∗16 Alleles with Late-Onset, Nonthymomatous, AChR-Ab-Positive Myasthenia Gravis. Autoimmune Dis. 2012, 541760 (2012).
36. Xie, C. et al. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc. Natl. Acad. Sci. 114, 8059–8064 (2017).
37. Bolotin, D. A. et al. Antigen receptor repertoire profiling from RNA-seq data. Nat. Biotechnol. 35, 908–911 (2017).
38. Shugay, M. et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLOS Comput. Biol. 11, e1004503 (2015).
39. Vlasova, E. & Shugay, M. Predicting SARS-CoV-2 exposure using T-cell repertoire sequencing and machine learning. (2023) doi:10.5281/zenodo.8362803.

40. Howie, B. et al. High-throughput pairing of T cell receptor α and β sequences. Sci. Transl. Med. 7, 301ra131 (2015).
41. Bovay, A. et al. T cell receptor alpha variable 12-2 bias in the immunodominant response to Yellow fever virus. Eur. J. Immunol. 48, 258–272 (2018).
42. Rosati, E. et al. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnol. 17, 61 (2017).
43. De Simone, M., Rossetti, G. & Pagani, M. Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges. Front. Immunol. 9, 1638 (2018).
44. Cole, D. K. et al. Germ Line-governed Recognition of a Cancer Epitope by an Immunodominant Human T-cell Receptor. J. Biol. Chem. 284, 27281–27289 (2009).
45. Augusto, D. G. et al. A common allele of HLA is associated with asymptomatic SARS-CoV-2 infection. Nature 620, 128–136 (2023).
46. Karnaukhov, V. et al. HLA variants have different preferences to present proteins with specific molecular functions which are complemented in frequent haplotypes. Front. Immunol. 13, (2022).