

# Supplementary materials

**Table S1:** Epitopes associated with TCR $\alpha$  clusters for Cohort I.

| antigen.epitope | antigen.species | count | pval     | cluster | all_clust_count | enrichment_score |
|-----------------|-----------------|-------|----------|---------|-----------------|------------------|
| PQPELPYPQPE     | Wheat           | 13    | 0.001307 | 0       | 13              | 1                |
| LMIERFVSL       | SARS-CoV-2      | 20    | 5.01E-21 | 0       | 20              | 1                |
| LTDEMIAQY       | SARS-CoV-2      | 22    | 8.44E-08 | 0       | 66              | 0.333333333      |
| AYAQKIFKI       | CMV             | 31    | 1.23E-07 | 0       | 32              | 0.96875          |
| MEVTPSGTWL      | SARS-CoV-2      | 33    | 0.000115 | 0       | 34              | 0.970588235      |
| IVTDFSVIK       | EBV             | 52    | 1.26E-05 | 0       | 211             | 0.246445498      |
| SPRWYFYLL       | SARS-CoV-2      | 52    | 3.12E-05 | 0       | 281             | 0.185053381      |
| LLWNGPMAV       | YFV             | 60    | 2.88E-17 | 0       | 138             | 0.434782609      |
| NLVPMVATV       | CMV             | 92    | 1.37E-17 | 0       | 642             | 0.143302181      |
| GILGFVFTL       | InfluenzaA      | 93    | 4.39E-11 | 0       | 504             | 0.18452381       |
| QPRAPIRPI       | EBV             | 6     | 4.06E-07 | 1       | 6               | 1                |
| MFLARGIVF       | SARS-CoV-2      | 29    | 1.61E-07 | 1       | 29              | 1                |
| QYIKWPWYI       | SARS-CoV-2      | 35    | 0.001283 | 1       | 68              | 0.514705882      |
| SGPLKAEIAQRLED  | InfluenzaA      | 36    | 2.03E-08 | 1       | 54              | 0.666666667      |
| KTFPPTPEPK      | SARS-CoV-2      | 43    | 0.000153 | 1       | 64              | 0.671875         |
| LLWNGPMAV       | YFV             | 53    | 1.89E-06 | 1       | 138             | 0.384057971      |
| RLRAEAQVK       | EBV             | 71    | 0.00087  | 1       | 198             | 0.358585859      |
| IVTDFSVIK       | EBV             | 75    | 7.13E-06 | 1       | 211             | 0.355450237      |
| RAKFKQLL        | EBV             | 106   | 0.000981 | 1       | 239             | 0.443514644      |
| PKYVKQNTLKLAT   | InfluenzaA      | 47    | 0.000552 | 2       | 78              | 0.602564103      |
| LPRWYFYLL       | HCoV-HKU1       | 32    | 9.06E-38 | 3       | 32              | 1                |
| FLRGRAYGL       | EBV             | 3     | 0.002733 | 4       | 7               | 0.428571429      |
| QVPLRPMTYK      | HIV-1           | 12    | 4.40E-08 | 4       | 15              | 0.8              |
| NEGVKAAW        | CMV             | 19    | 3.87E-09 | 4       | 51              | 0.37254902       |
| PKYVKQNTLKLAT   | InfluenzaA      | 19    | 2.12E-09 | 4       | 78              | 0.243589744      |
| TTDPSFLGRY      | SARS-CoV-2      | 22    | 1.57E-09 | 4       | 87              | 0.252873563      |
| RTLNAWVKV       | HIV-1           | 7     | 6.53E-26 | 5       | 9               | 0.777777778      |
| PKYVKQNTLKLAT   | InfluenzaA      | 8     | 6.57E-30 | 5       | 78              | 0.102564103      |
| IVTDFSVIK       | EBV             | 13    | 4.41E-07 | 5       | 211             | 0.061611374      |
| GILGFVFTL       | InfluenzaA      | 33    | 2.06E-11 | 5       | 504             | 0.06547619       |
| KLGGALQAK       | CMV             | 36    | 1.75E-08 | 5       | 2190            | 0.016438356      |
| RLITGRLQSL      | SARS-CoV-2      | 9     | 6.27E-06 | 6       | 9               | 1                |

|                       |                  |    |           |    |      |             |
|-----------------------|------------------|----|-----------|----|------|-------------|
| YINVFAFPF             | SARS-CoV-2       | 13 | 2.78E-13  | 6  | 13   | 1           |
| QVPLRPMTYK            | HIV-1            | 3  | 1.90E-17  | 9  | 15   | 0.2         |
| RLRAEAQVK             | EBV              | 19 | 3.91E-18  | 9  | 198  | 0.095959596 |
| KLGGALQAK             | CMV              | 21 | 7.24E-05  | 9  | 2190 | 0.009589041 |
| NLVPMVATV             | CMV              | 21 | 1.23E-23  | 9  | 642  | 0.03271028  |
| SPRWYFYLL             | SARS-CoV-2       | 21 | 1.78E-25  | 9  | 281  | 0.074733096 |
| DPFRLLQNSQVFS         | InfluenzaA       | 2  | 1.20E-08  | 10 | 2    | 1           |
| APFSESEQPVLG          | TriticumAestivum | 13 | 0.003376  | 10 | 13   | 1           |
| TTDPSFLGRY            | SARS-CoV-2       | 13 | 9.95E-12  | 10 | 87   | 0.149425287 |
| HWFVTQRNFYEPQII       | SARS-CoV-2       | 23 | 3.80E-15  | 10 | 23   | 1           |
| RISNCVADYSVLVNS       | SARS-CoV-2       | 24 | 1.77E-42  | 10 | 24   | 1           |
| KLGGALQAK             | CMV              | 89 | 5.85E-07  | 10 | 2190 | 0.040639269 |
| ALNTLVKQL             | SARS-CoV-2       | 2  | 6.21E-07  | 13 | 2    | 1           |
| RLRAEAQVK             | EBV              | 2  | 0.002061  | 13 | 198  | 0.01010101  |
| NTFSSTFNV             | SARS-CoV-2       | 5  | 1.84E-14  | 14 | 5    | 1           |
| PFPQPELPY             | Wheat            | 1  | 2.69E-05  | 22 | 1    | 1           |
| QARQMVMQAMRTIGTHP     | InfluenzaA       | 3  | 2.71E-19  | 22 | 4    | 0.75        |
| TTDPSFLGRY            | SARS-CoV-2       | 18 | 2.27E-07  | 22 | 87   | 0.206896552 |
| YVYSRVKNL             | SARS-CoV-2       | 2  | 1.99E-39  | 24 | 2    | 1           |
| NVILLNKHI             | SARS-CoV-2       | 2  | 9.16E-13  | 24 | 2    | 1           |
| FRDYVDRLFYKTLRAEQASQE | HIV-1            | 4  | 0.000569  | 24 | 27   | 0.148148148 |
| RLRAEAQVK             | EBV              | 18 | 1.35E-28  | 24 | 198  | 0.090909091 |
| NEGVKAAW              | CMV              | 18 | 1.37E-07  | 24 | 51   | 0.352941176 |
| MLDLQPETT             | HPV              | 18 | 1.26E-28  | 24 | 18   | 1           |
| GILGFVFTL             | InfluenzaA       | 19 | 0.001406  | 24 | 504  | 0.037698413 |
| SPRWYFYLL             | SARS-CoV-2       | 20 | 6.85E-09  | 24 | 281  | 0.071174377 |
| NLVPMVATV             | CMV              | 23 | 1.65E-05  | 24 | 642  | 0.035825545 |
| KLGGALQAK             | CMV              | 30 | 0.000218  | 24 | 2190 | 0.01369863  |
| NCTFEYVSQPFLMDL       | SARS-CoV-2       | 15 | 2.91E-23  | 29 | 81   | 0.185185185 |
| TFEYVSQPFLMDLE        | SARS-CoV-2       | 18 | 1.60E-30  | 29 | 88   | 0.204545455 |
| TSQWLTNIF             | SARS-CoV-2       | 2  | 1.99E-15  | 31 | 2    | 1           |
| YTVSCLPFT             | SARS-CoV-2       | 3  | 1.63E-108 | 31 | 3    | 1           |
| YLQPRTFLL             | SARS-CoV-2       | 42 | 3.24E-101 | 31 | 263  | 0.159695817 |
| KLGGALQAK             | CMV              | 65 | 0.001386  | 31 | 2190 | 0.029680365 |
| NCTFEYVSQPFLMDL       | SARS-CoV-2       | 66 | 1.73E-24  | 31 | 81   | 0.814814815 |
| TFEYVSQPFLMDLE        | SARS-CoV-2       | 69 | 1.16E-53  | 31 | 88   | 0.784090909 |
| RTLNAWVKV             | HIV-1            | 2  | 5.70E-06  | 37 | 9    | 0.222222222 |

|                 |            |    |              |     |      |             |
|-----------------|------------|----|--------------|-----|------|-------------|
| MHQKRTAMFQDPQER | HPV-16     | 1  | 9.61E-24     | 42  | 1    | 1           |
| SPRWYFYFL       | SARS-CoV-2 | 12 | 2.67E-13     | 42  | 281  | 0.042704626 |
| LLLEWLAMA       | SARS-CoV-2 | 13 | 6.70E-08     | 42  | 13   | 1           |
| KLGGALQAK       | CMV        | 21 | 3.00E-18     | 42  | 2190 | 0.009589041 |
| NLSDRVVFV       | SARS-CoV-2 | 1  | 0.00091      | 47  | 3    | 0.333333333 |
| NLSDRVVFV       | SARS-CoV-2 | 1  | 0.00091      | 49  | 3    | 0.333333333 |
| KLGGALQAK       | CMV        | 10 | 0.00096      | 52  | 2190 | 0.00456621  |
| NQKLIANQF       | SARS-CoV-2 | 9  | 2.52E-17     | 54  | 58   | 0.155172414 |
| KLGGALQAK       | CMV        | 18 | 1.69E-37     | 55  | 2190 | 0.008219178 |
| SLFNTVATLY      | HIV-1      | 1  | 0.00045<br>5 | 60  | 1    | 1           |
| KLGGALQAK       | CMV        | 9  | 0.00192<br>1 | 65  | 2190 | 0.004109589 |
| GPEPLPQGQLTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEGLPQGQLTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLGQGQLTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLPQAQLTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLPQGGLTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLPQGQGTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLPQGQLGAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLPQGQLTAY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| LPEPLPQGQLTGY   | EBV        | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| RLQSLQTYV       | SARS-CoV-2 | 1  | 0.00045<br>5 | 67  | 1    | 1           |
| TTDPSFLGRY      | SARS-CoV-2 | 2  | 0.00040<br>6 | 78  | 87   | 0.022988506 |
| YLQPRTFLL       | SARS-CoV-2 | 2  | 0.00362<br>3 | 85  | 263  | 0.007604563 |
| QLQFPQPPELPY    | Wheat      | 1  | 4.90E-15     | 87  | 1    | 1           |
| RAKFKQLL        | EBV        | 13 | 0.00048      | 87  | 239  | 0.054393305 |
| KLGGALQAK       | CMV        | 14 | 6.34E-07     | 87  | 2190 | 0.006392694 |
| NLVPMVATV       | CMV        | 4  | 0.00203<br>9 | 91  | 642  | 0.00623053  |
| RLNQLESKV       | SARS-CoV-2 | 1  | 0.00045<br>5 | 99  | 1    | 1           |
| NLSDRVVFV       | SARS-CoV-2 | 1  | 0.00091      | 108 | 3    | 0.333333333 |
| NLVPMVATV       | CMV        | 9  | 3.23E-08     | 110 | 642  | 0.014018692 |
| PTDNYITTY       | SARS-CoV-2 | 9  | 5.45E-23     | 110 | 10   | 0.9         |

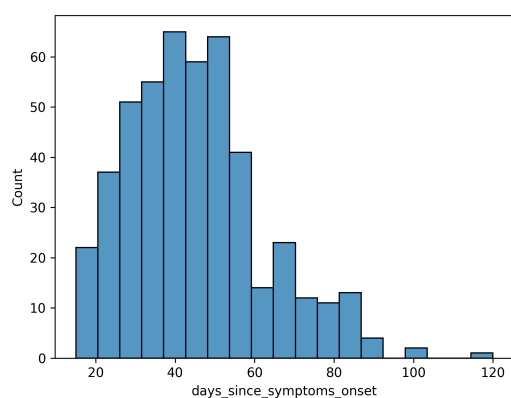


**Table S2:** Epitopes associated with TCR $\beta$  clusters for Cohort I.

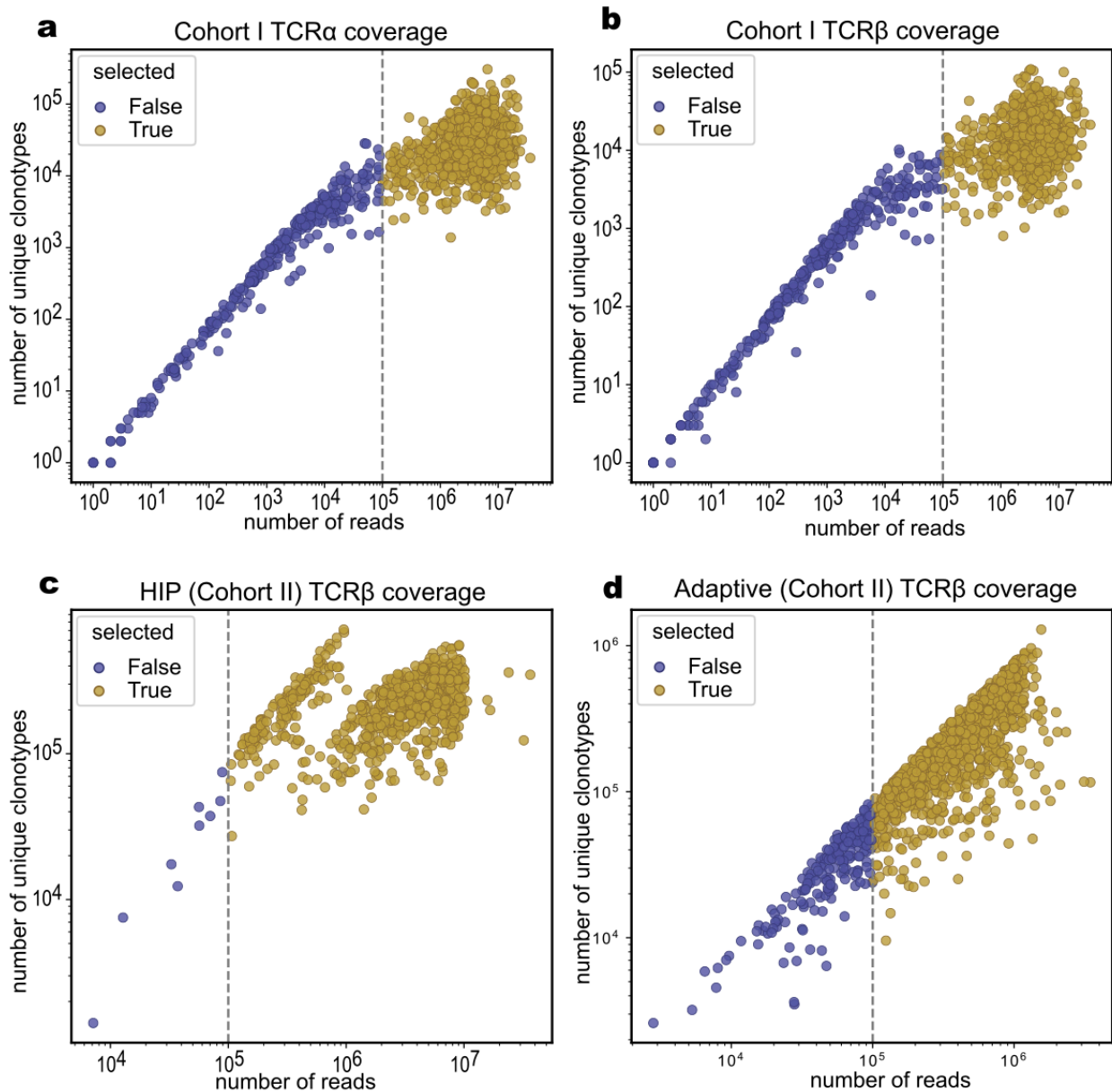
| antigen.epitope      | antigen.species | count | pval     | cluster | all_clust_count | enrichment_score |
|----------------------|-----------------|-------|----------|---------|-----------------|------------------|
| TPQDLNTML            | HIV-1           | 22    | 2.10E-05 | 0       | 25              | 0.88             |
| VTEHDTLLY            | CMV             | 33    | 5.01E-05 | 0       | 38              | 0.868421         |
| SIAYTMSL             | SARS-CoV-2      | 5     | 3.90E-05 | 3       | 5               | 1                |
| PKYVKQNTLKLAT        | InfluenzaA      | 17    | 2.86E-06 | 3       | 23              | 0.73913          |
| QARQMVQAMRTIGTH<br>P | InfluenzaA      | 18    | 0.000358 | 3       | 46              | 0.391304         |
| RLRAEAQVK            | EBV             | 19    | 3.63E-07 | 3       | 33              | 0.575758         |
| LTDEMIAQY            | SARS-CoV-2      | 20    | 5.65E-06 | 3       | 24              | 0.833333         |
| NLVPMVATV            | CMV             | 78    | 2.86E-06 | 3       | 299             | 0.26087          |
| NEGVKAAW             | CMV             | 20    | 4.47E-06 | 4       | 21              | 0.952381         |
| TMADLVYAL            | SARS-CoV-2      | 24    | 1.22E-10 | 4       | 24              | 1                |
| NLVPMVATV            | CMV             | 83    | 2.71E-06 | 4       | 299             | 0.277592         |
| RLRPGGKKK            | HIV-1           | 3     | 0.001204 | 6       | 3               | 1                |
| VTEHDTLLY            | CMV             | 3     | 0.000347 | 7       | 38              | 0.078947         |
| LLYDANYFL            | SARS-CoV-2      | 1     | 6.19E-05 | 11      | 3               | 0.333333         |
| RLNEVAKNL            | SARS-CoV-2      | 3     | 6.51E-07 | 16      | 3               | 1                |
| SSGDATTAY            | SARS-CoV-2      | 1     | 3.39E-05 | 54      | 1               | 1                |
| LLQTGIHVRVSQPSL      | CMV             | 4     | 2.21E-05 | 54      | 20              | 0.2              |

**Table S3:** Epitopes associated with both alpha and beta clusters for Cohort I. The odds and Pearson correlation scores for alpha-beta cluster co-occurrence is also shown.

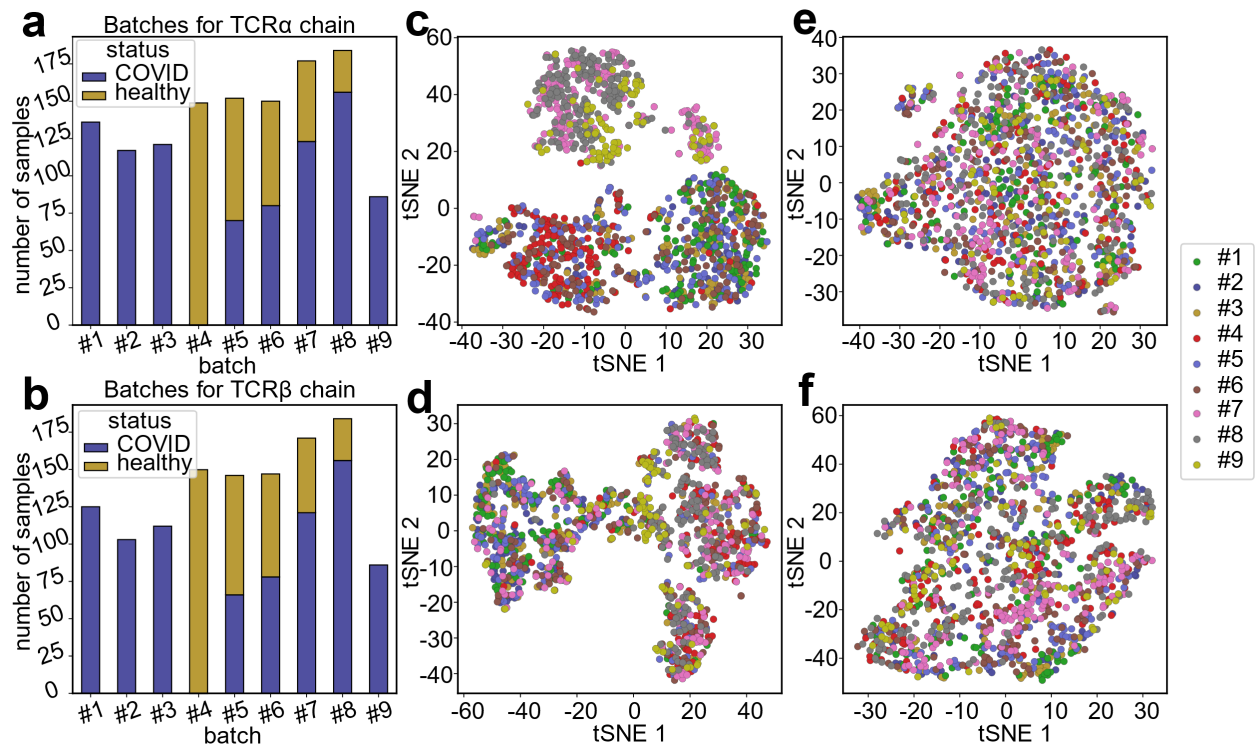
| antigen epitope      | antigen species | cluster alpha | cluster beta | odds     | corr     |
|----------------------|-----------------|---------------|--------------|----------|----------|
| RLRAEAQVK            | EBV             | 9             | 3            | 0.004513 | 0.49912  |
| RLRAEAQVK            | EBV             | 13            | 3            | 0.004578 | 0.361369 |
| RLRAEAQVK            | EBV             | 24            | 3            | 0.004357 | 0.585371 |
| RLRAEAQVK            | EBV             | 1             | 3            | 0.000902 | 0.338115 |
| NLVPMVATV            | CMV             | 9             | 3            | 0.004513 | 0.49912  |
| NLVPMVATV            | CMV             | 9             | 4            | 0.003094 | 0.473646 |
| NLVPMVATV            | CMV             | 24            | 3            | 0.004357 | 0.585371 |
| NLVPMVATV            | CMV             | 24            | 4            | 0.003049 | 0.550947 |
| NLVPMVATV            | CMV             | 110           | 3            | 0.004584 | 0.265179 |
| NLVPMVATV            | CMV             | 110           | 4            | 0.003117 | 0.280618 |
| NLVPMVATV            | CMV             | 0             | 3            | 0.001639 | 0.328225 |
| NLVPMVATV            | CMV             | 0             | 4            | 0.001598 | 0.314887 |
| NLVPMVATV            | CMV             | 91            | 3            | 0.004575 | 0.393508 |
| NLVPMVATV            | CMV             | 91            | 4            | 0.003117 | 0.360685 |
| SPRWYFYLL            | SARS-CoV-2      | 9             | 3            | 0.004513 | 0.49912  |
| SPRWYFYLL            | SARS-CoV-2      | 24            | 3            | 0.004357 | 0.585371 |
| SPRWYFYLL            | SARS-CoV-2      | 0             | 3            | 0.001639 | 0.328225 |
| SPRWYFYLL            | SARS-CoV-2      | 42            | 3            | 0.004567 | 0.552613 |
| NEGVKAAW             | CMV             | 24            | 4            | 0.003049 | 0.550947 |
| NEGVKAAW             | CMV             | 4             | 4            | 0.002123 | 0.572947 |
| TTDPSFLGRY           | SARS-CoV-2      | 78            | 0            | 0.003347 | 0.115375 |
| TTDPSFLGRY           | SARS-CoV-2      | 22            | 0            | 0.003316 | 0.529591 |
| TTDPSFLGRY           | SARS-CoV-2      | 10            | 0            | 0.002674 | 0.620498 |
| TTDPSFLGRY           | SARS-CoV-2      | 4             | 0            | 0.0022   | 0.613155 |
| LTDEMIAQY            | SARS-CoV-2      | 0             | 3            | 0.001639 | 0.328225 |
| QARQMVQ<br>AMRTIGTHP | InfluenzaA      | 22            | 3            | 0.00452  | 0.509571 |



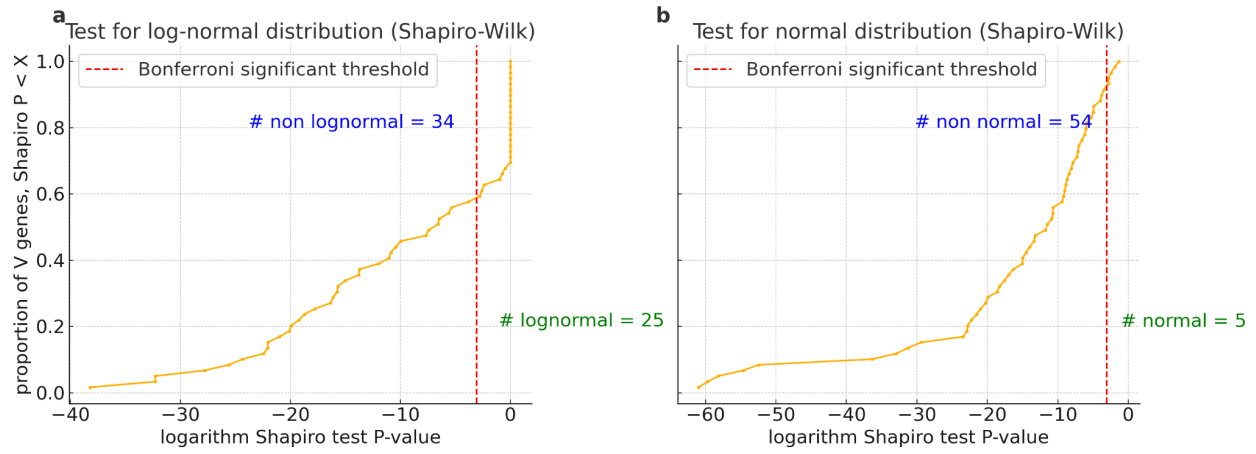
**Fig S1. Number of days since COVID-19 symptoms onset.** Time when blood was collected for repertoire sequencing since COVID-19 symptoms. Asymptomatic cases are not considered.



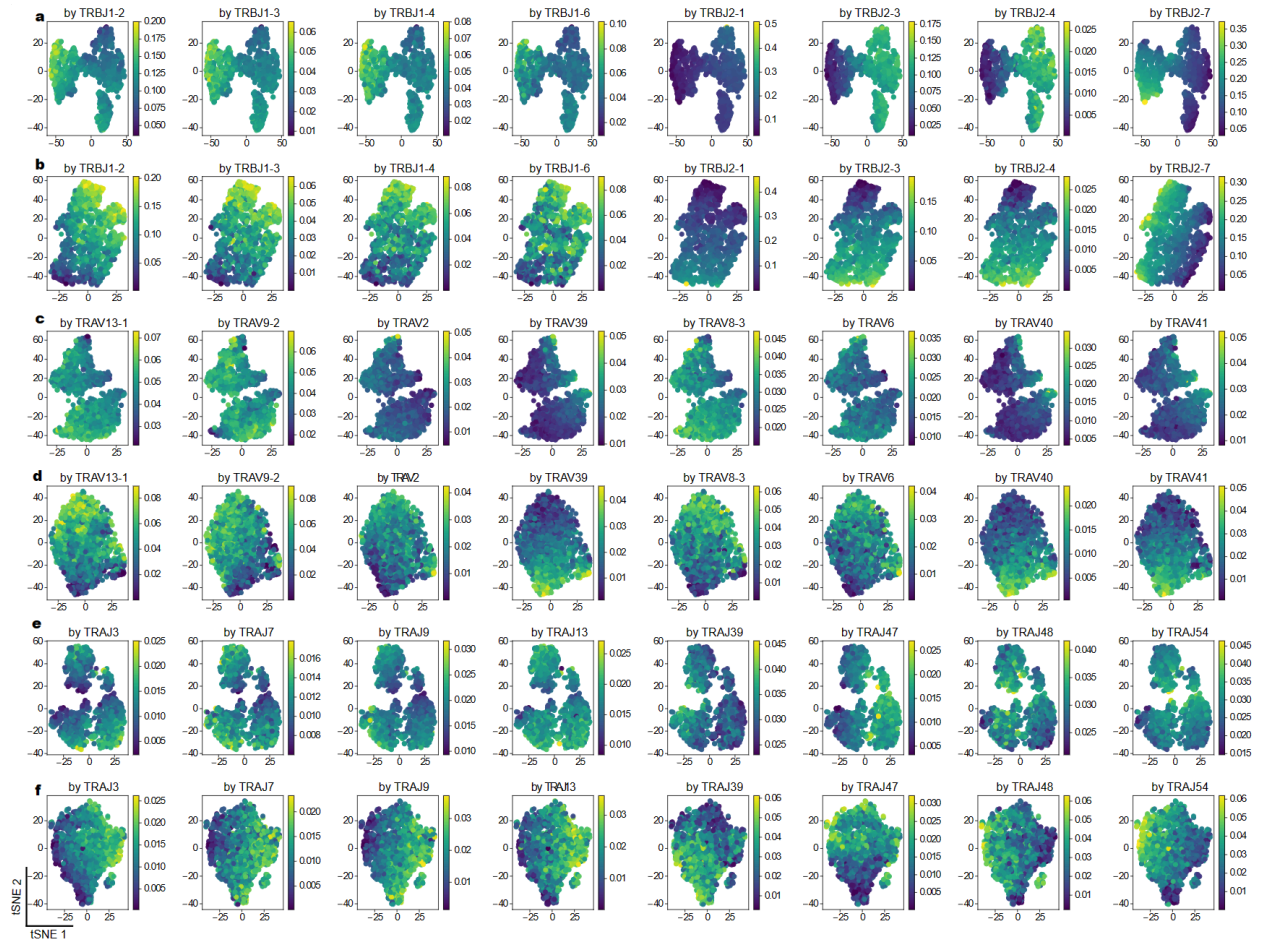
**Fig S2. Distribution of sample coverage across technologies.** **A.** Sample coverage for Cohort I TCR $\alpha$  clonotypes. There were 1637 donors, 1273 samples left after preprocessing. **B.** The same as **A** but for TCR $\beta$  clonotypes. There were 1637 donors, 1225 samples left after preprocessing. For both  $\beta$  and  $\alpha$  chains together, 1224 samples passed the read threshold. **C.** Distribution of coverage for TCR $\beta$  clonotypes in HIP cohort. There were 786 samples, but 9 were dropped during preprocessing. **D.** Distribution of coverage for TCR $\beta$  clonotypes in Adaptive cohort. There were 1414 samples, but 200 were dropped during preprocessing. HIP and Adaptive cohorts represent Cohort II.



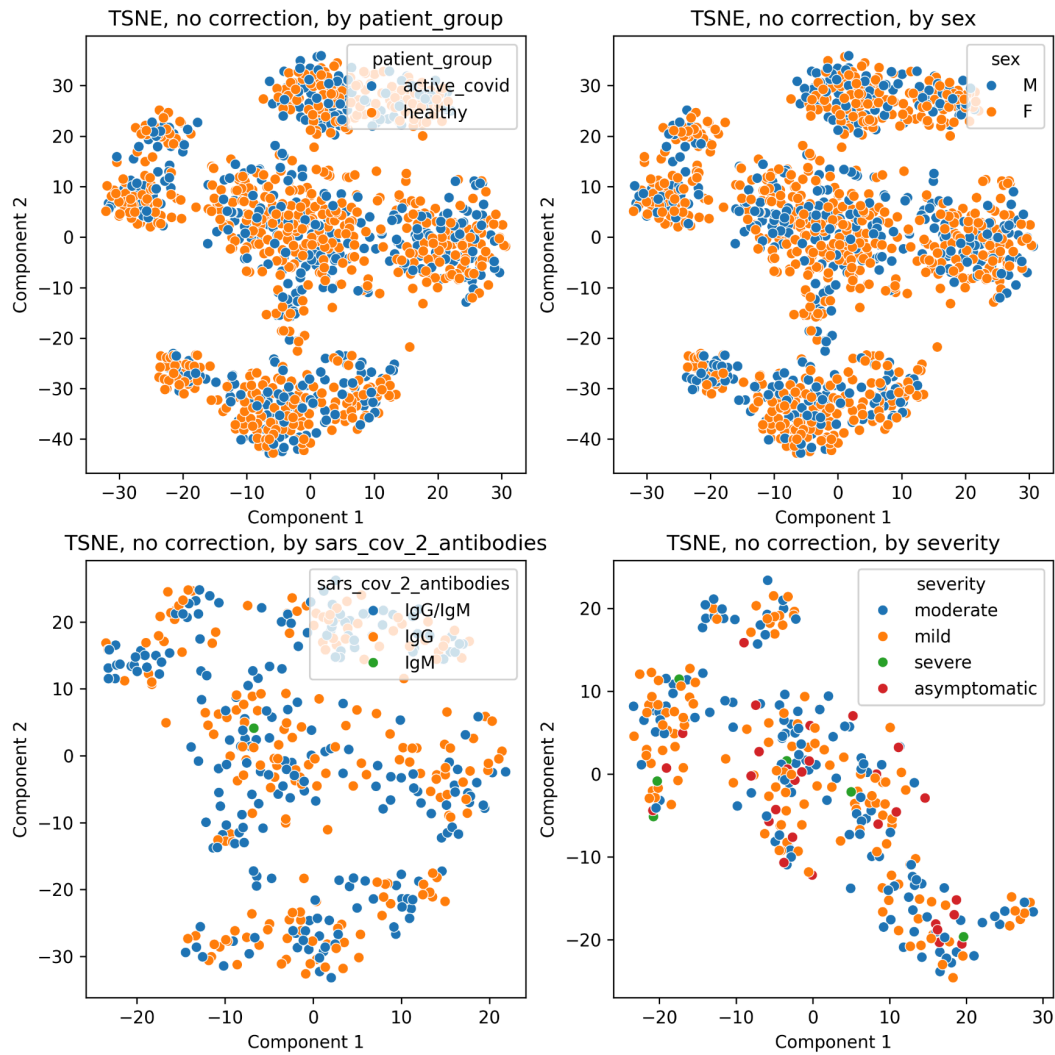
**Figure S3. Additional information on data.** **A, B.** A bar plot of total number of subjects, number of COVID-19-convalescent and healthy cases for 9 batches of Cohort I reported in this study. The distribution is shown for samples with TCR $\alpha$  (A) and TCR $\beta$  (B) high-quality data. **C, D.** Visualization of batch effects using t-SNE. Plots show TRAJ (C) and TRBJ (D) gene usage profiles between samples before batch effect correction procedure, samples are colored by batch. **E, F.** TRAJ (E) and TRBJ (F) gene usage profiles after batch effect correction procedure, samples are colored by batch.



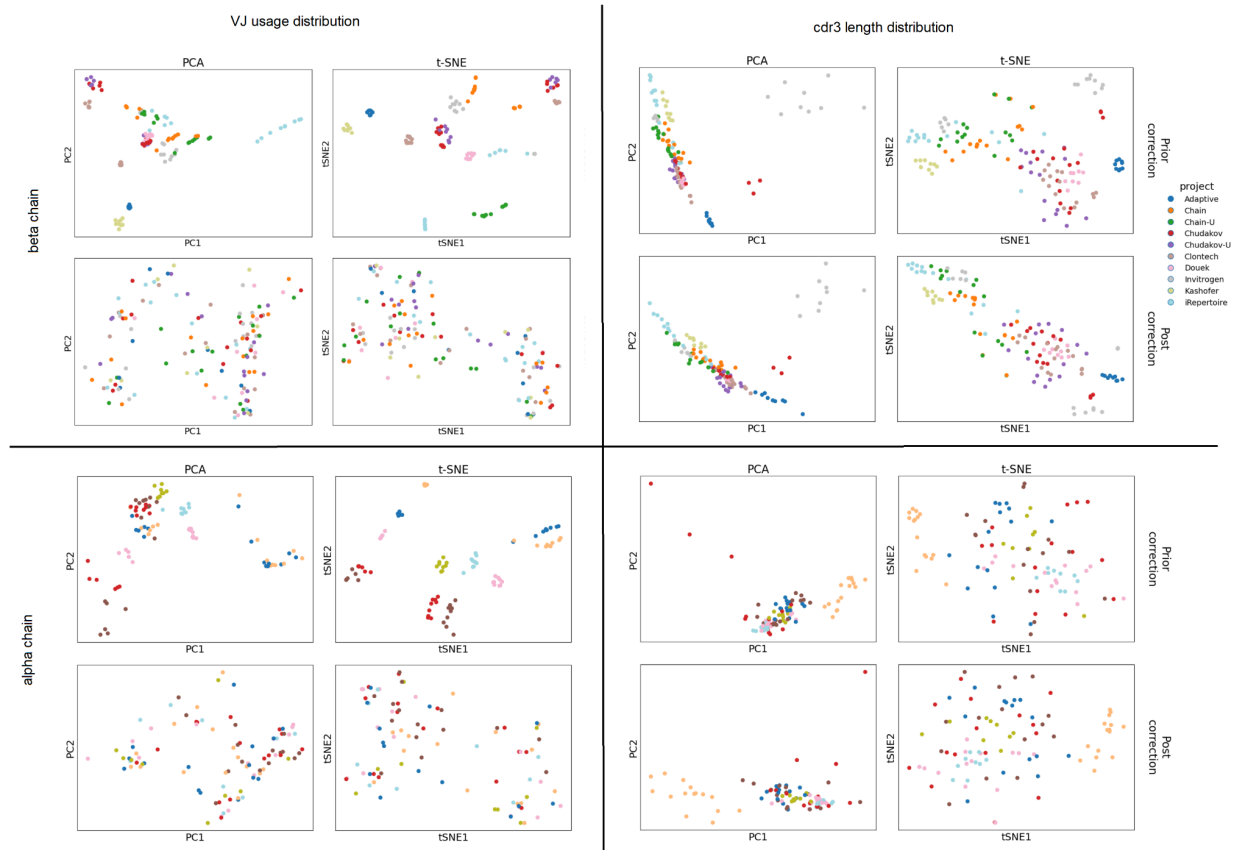
**Fig S4. Log-normality testing for V gene usage distribution. A.** Cumulative distribution of Shapiro-Wilk log p-values for lognormality test within each V gene. As can be seen from the plot that half of genes are rejected to be lognormally distributed with p-value adjustment applied, whereas another half of genes (25 genes) are not rejected to be lognormally distributed. **B.** The same as picture (A), but for normality checks in V gene distributions. In this case, only 5 genes are not rejected to be normally distributed, which is 5 times less than in lognormal tests.



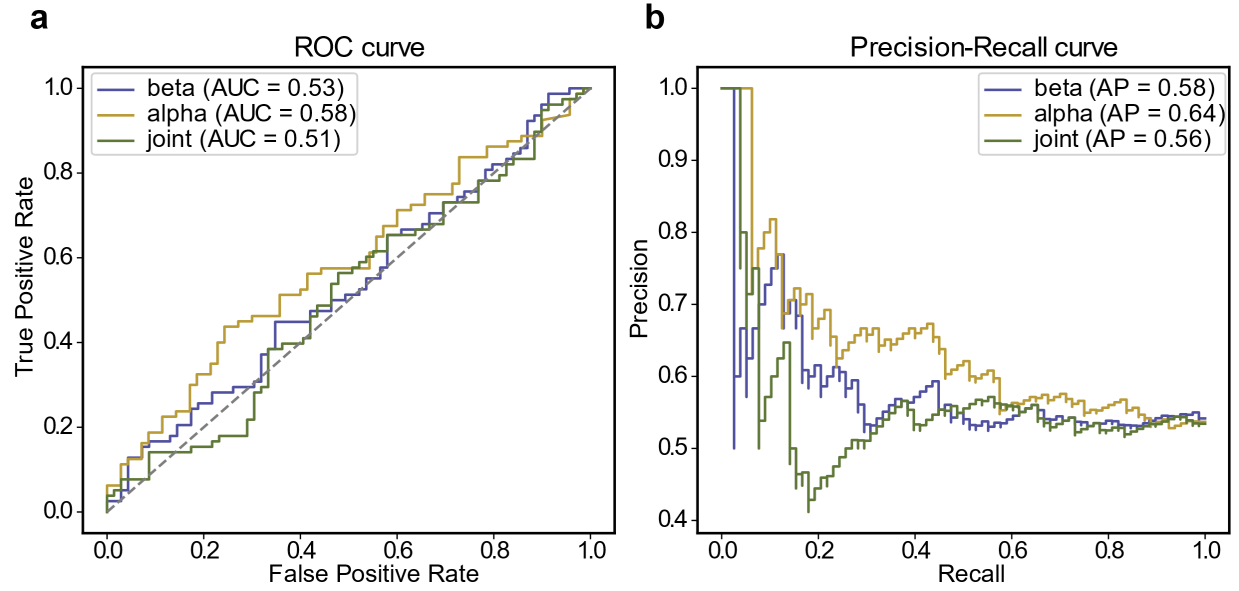
**Fig S5. Allele effects visualization for TCR $\beta$  J gene usages and TCR $\alpha$  V and J gene usages. A, B.** Visualization of biological effects pre (A) and post (B) batch effect correction procedure for TCR $\beta$  J gene usages. C, D. The same as (A, B), but for TCR $\alpha$  V gene usages. E, F. The same as (A, B), but for TCR $\alpha$  J gene usages.



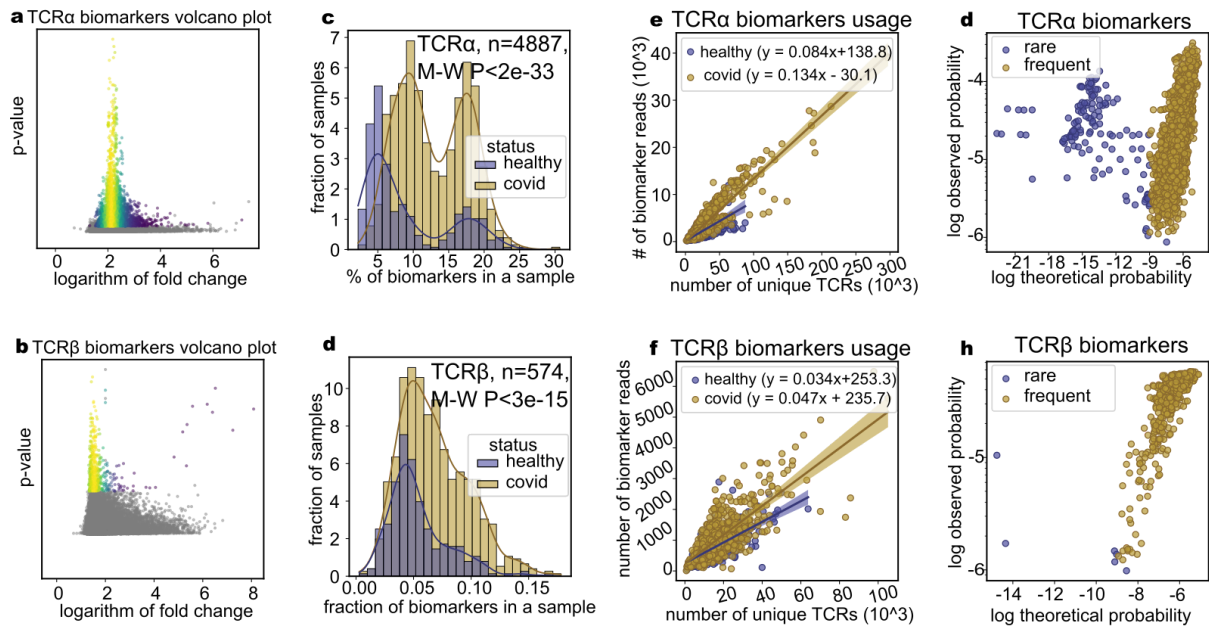
**Fig S6. V gene TSNE colored by categorical metadata.** We colored the TRBV gene usage TSNE by the categorical features (SARS-CoV-2 status, sex, presence of COVID-19 antibodies, severity). These plots do not show any correlations or patterns in TSNE coloring.



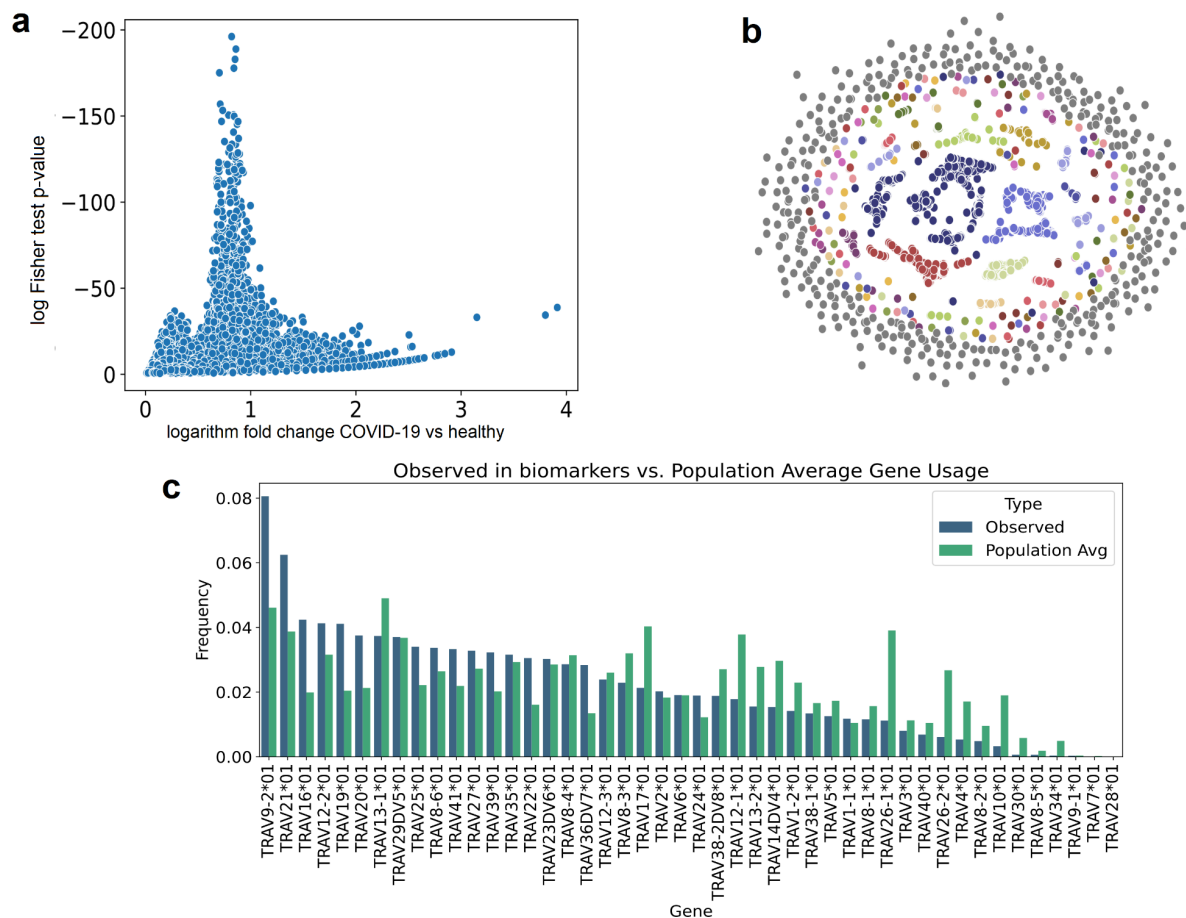
**Fig S7. Dimensionality reduction of VJ gene-usage profiles before and after batch-effect correction.** PCA (left) and t-SNE (right) projections of samples from multiple AIRR-seq projects (from Barennes et al., Nature Biotechnology 2020) colored by dataset of origin. Upper row contains raw, uncorrected data; lower row — data after applying the batch-effect correction procedure described in Methods. Before correction, samples cluster strongly by project/batch, reflecting substantial technical variation between sequencing runs and protocols. After correction, this batch-driven clustering is largely eliminated, and samples become more uniformly mixed across projects. Each section contains the distributions derived either from VJ usage patterns or from CDR3 length patterns for the clonotypes. The upper row contains data for TCR beta chain, whereas the lower row corresponds to the TCR alpha chain.



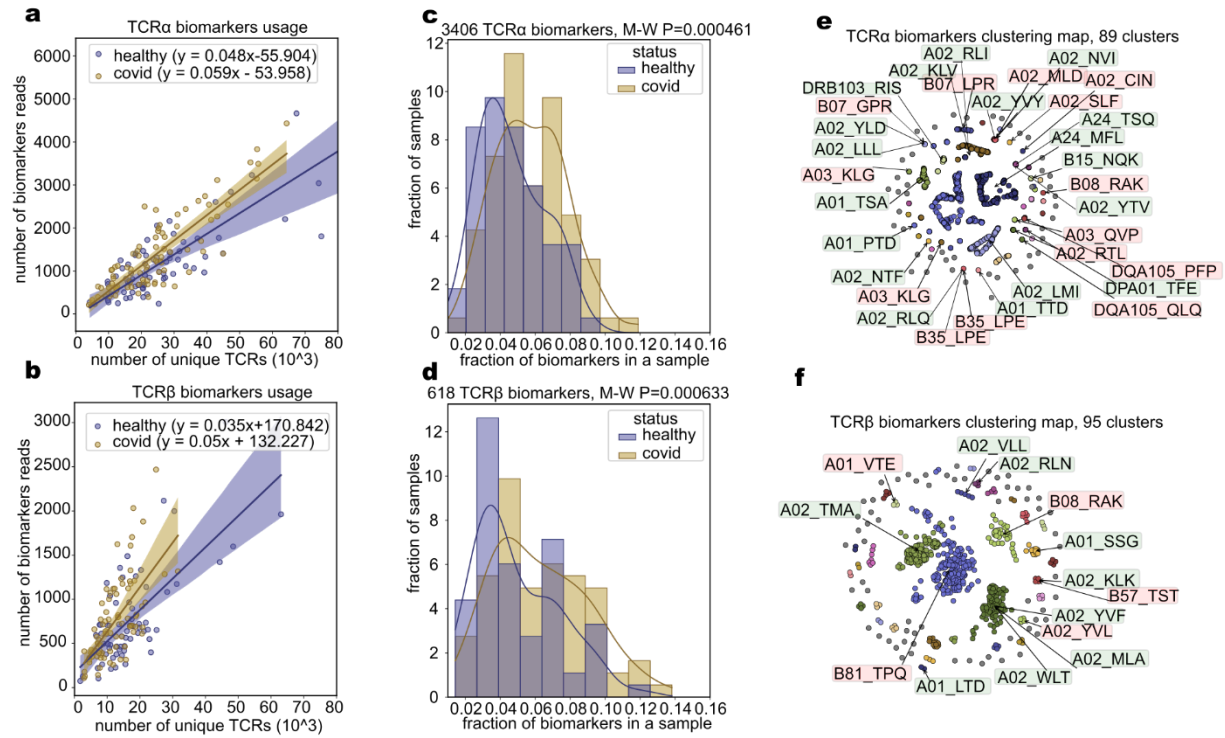
**Fig S8. COVID-19 status classification based on TCR  $\alpha$  and  $\beta$  VJ gene usage.** **A.** ROC Curve showing three classifier scores.  $\alpha$  corresponds to the classifier based on TRAV and TRAJ usages.  $\beta$  corresponds to TRBV and TRBJ usages. Joint is the classifier which uses both TCR  $\alpha$  and  $\beta$  genes. **B.** The same as (a), but for precision-recall curve.



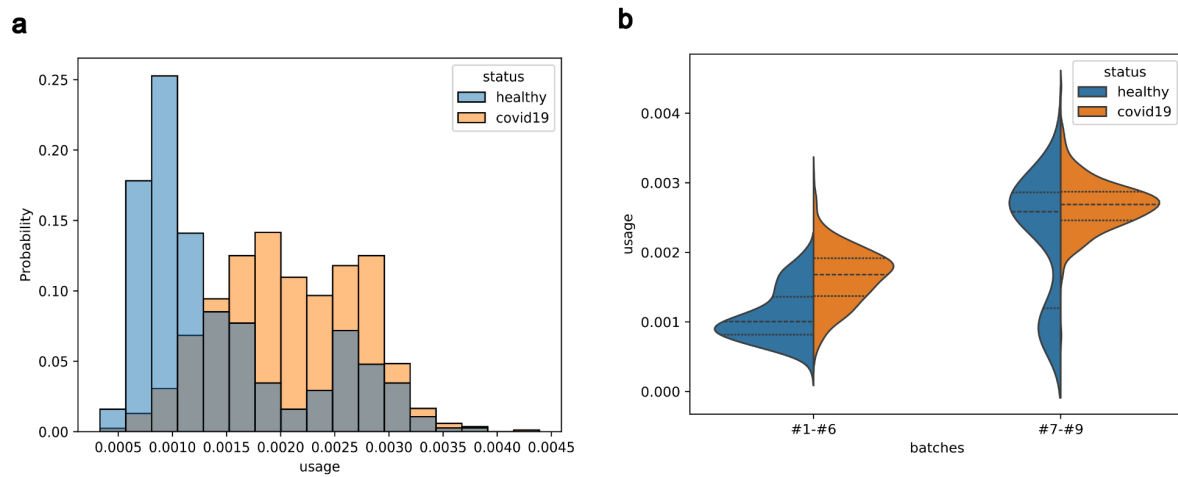
**Fig S9. Analysis of COVID-19 TCR  $\alpha$  and  $\beta$  biomarkers before cleanup procedure.** **A, B.** Volcano plot representing the position of selected clonotypes among all the clonotypes of interest for TCR $\alpha$  (a) and TCR $\beta$  (b) chain. **C, D.** Scatter plot showing the frequency of COVID-19-associated TCR $\alpha$  (c) and TCR $\beta$  (d) clonotypes plotted against the number of unique TCR $\beta$  clonotypes in Cohort I before the cleanup. **E, F.** Distribution of COVID-associated TCR $\alpha$  (e) and TCR $\beta$  (f) clonotypes across healthy and convalescent samples before the cleanup. **G, H.** Scatterplot showing the usage of TCR $\alpha$  (g) and TCR $\beta$  (h) clonotypes in samples against the expected frequency (calculated using probability based model software OLGA).



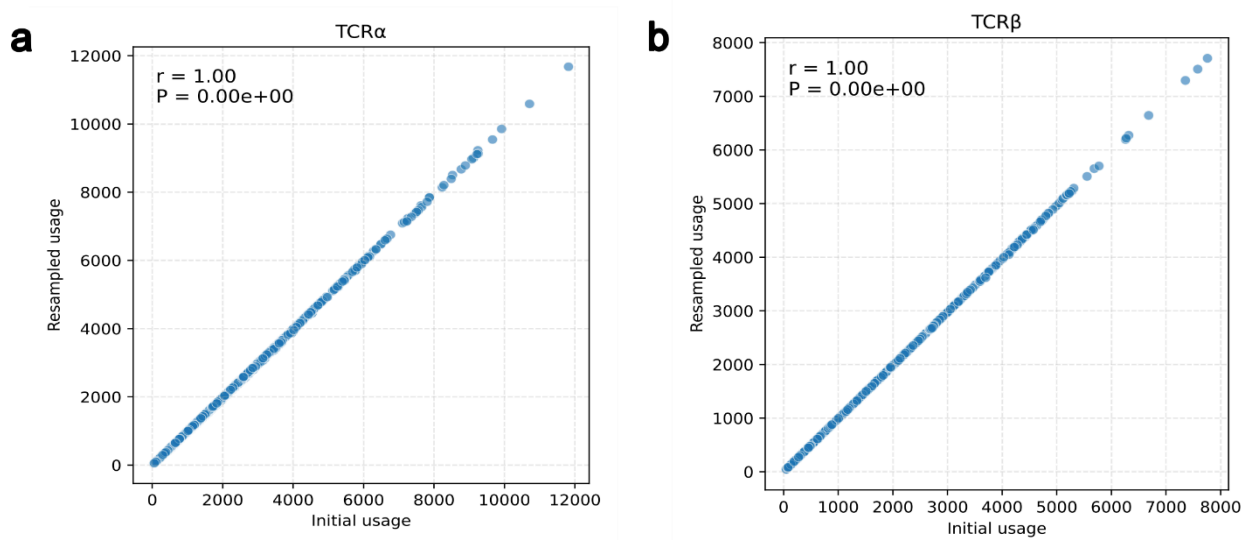
**Fig S10. TCR $\alpha$  biomarkers with equal V gene adjustment.** **A.** Fold change of biomarker usage in COVID-19 patients in comparison to healthy individuals. **B.** Clustering of found biomarkers (8901 biomarkers with P-adjusted < 0.05). The biomarkers are clustered in 545 clusters. **C.** The histogram of V gene usage differences within the biomarker set in comparison to the overall V gene usage in population. TRAV9-2\*01, TRAV21\*01, TRAV16\*01 V genes are significantly higher used in biomarkers set (Mann-Whitney P < 0.01).



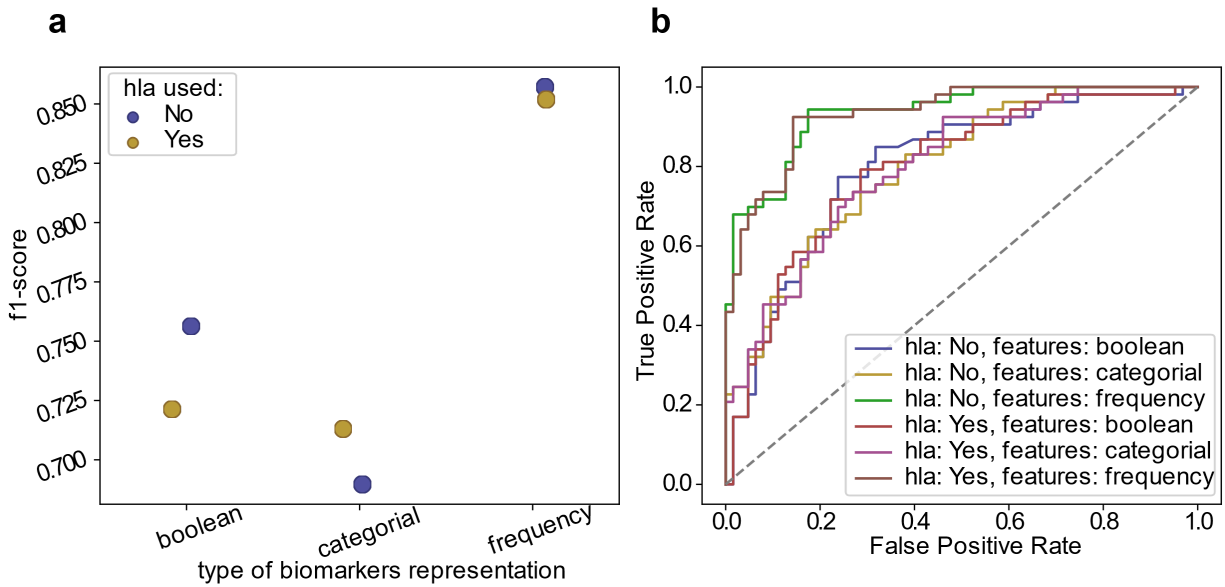
**Fig S11. Biomarkers analysis for statistical testing without a test batch (#6).** **A, B.** Scatter plot showing the frequency of COVID-19-associated TCRα (a) and TCRβ (b) clonotypes plotted against the number of unique clonotypes in the Cohort I for the dropped batch (#6). **C, D.** Distribution of COVID-associated TCRα (c) and TCRβ (d) clonotypes across healthy and convalescent samples for the dropped batch (#6). **E, F.** TCRα (e) and TCRβ (f) clonotype clustering. The epitopes marked with green are the SARS-CoV-2 epitopes, while the red ones are the epitopes belonging to other viruses.



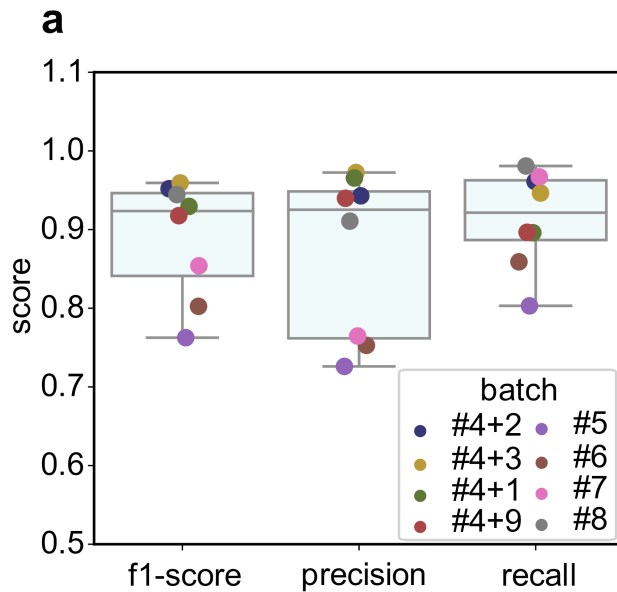
**Fig S12. TCR $\alpha$  biomarker usage in samples without batch effect correction.** **A.** Histogram of TCR $\alpha$  biomarker usage in healthy vs convalescent samples. Two peaks are visible for both healthy and convalescent distributions. **B.** The violin plots for batches #1-6 and #7-9 shows a significant difference in biomarker usage for both healthy and convalescent donors. The batches 1-6 contribute into the first peak of the histogram whereas the later batches form the higher peak.



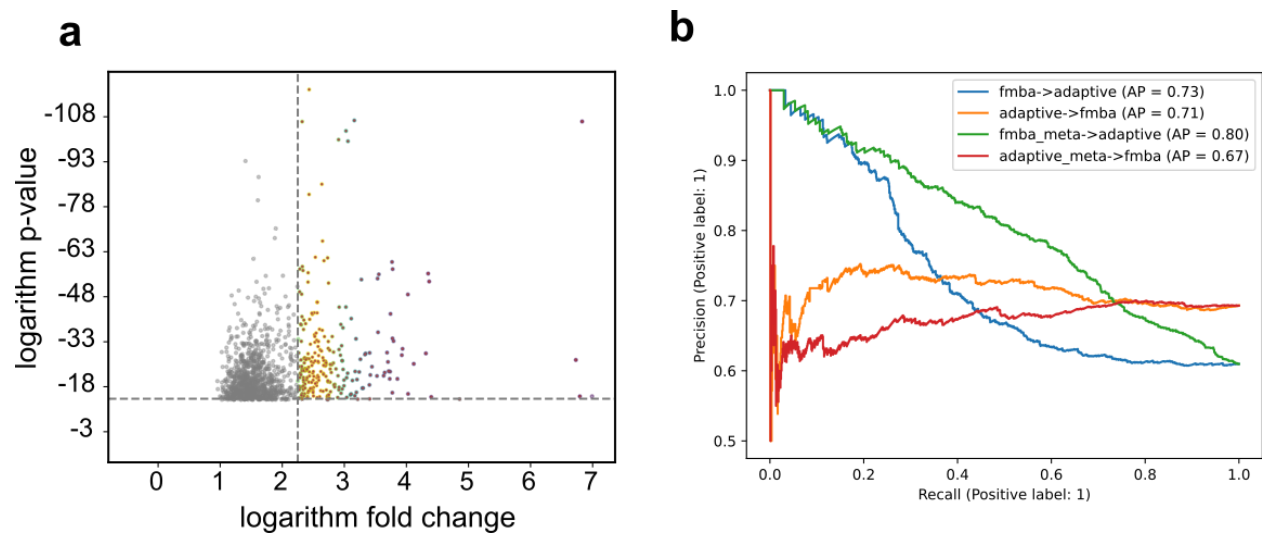
**Fig S13. Biomarker usage in pre- and post- corrected samples.** **A, B.** Biomarker usage correlation is shown for samples pre- and post- batch effect correction and resampling procedure.



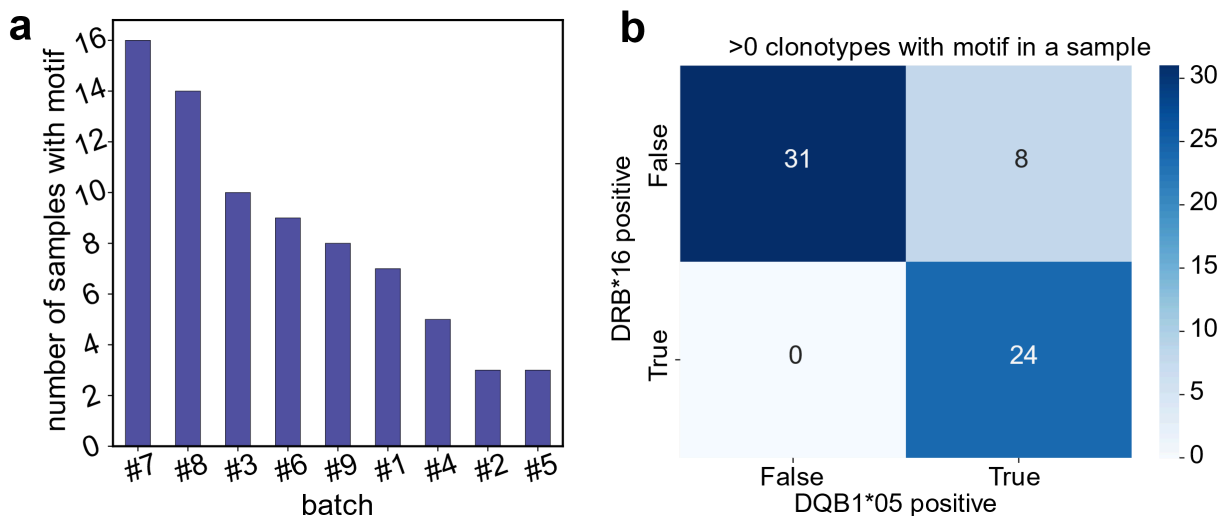
**Fig S14. Comparison of model settings for joint meta-classifiers.** **A.** F1 score of models trained on boolean/categorical/frequency features. The frequency features are gaining the highest f1-scores. **B.** Roc-curve, showing that frequency feature-based classifiers are getting the highest scores, whereas the HLA boolean features do not contribute to the model performance significantly.



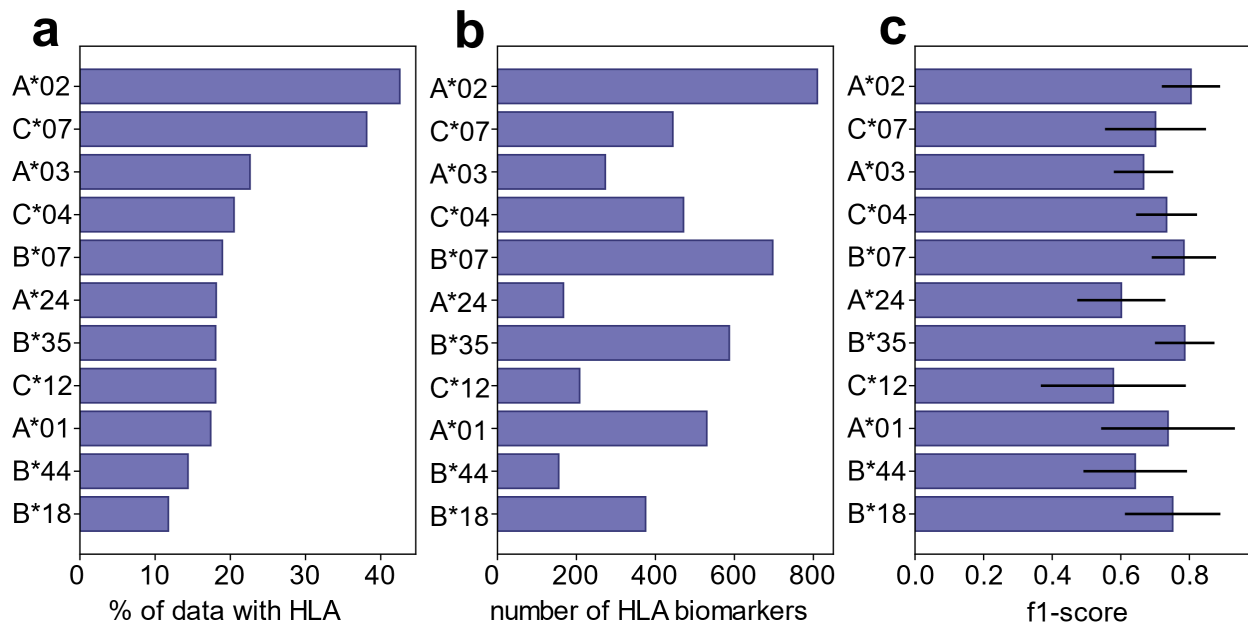
**Fig S15. TCR $\alpha$ + $\beta$  metaclonotype model cross validation between batches.** Evaluation of target metrics (f1-score, precision, recall) for one batch out cross validation. Each point represents the target metrics for one of the batches. Each batch is thrown out as a validation set and the entire model is being trained using the rest of the data. Joint metaclonotype model is used here. The error bars represent the range of metrics between thrown out batches. One can see that the summarized batches result in a better quality.



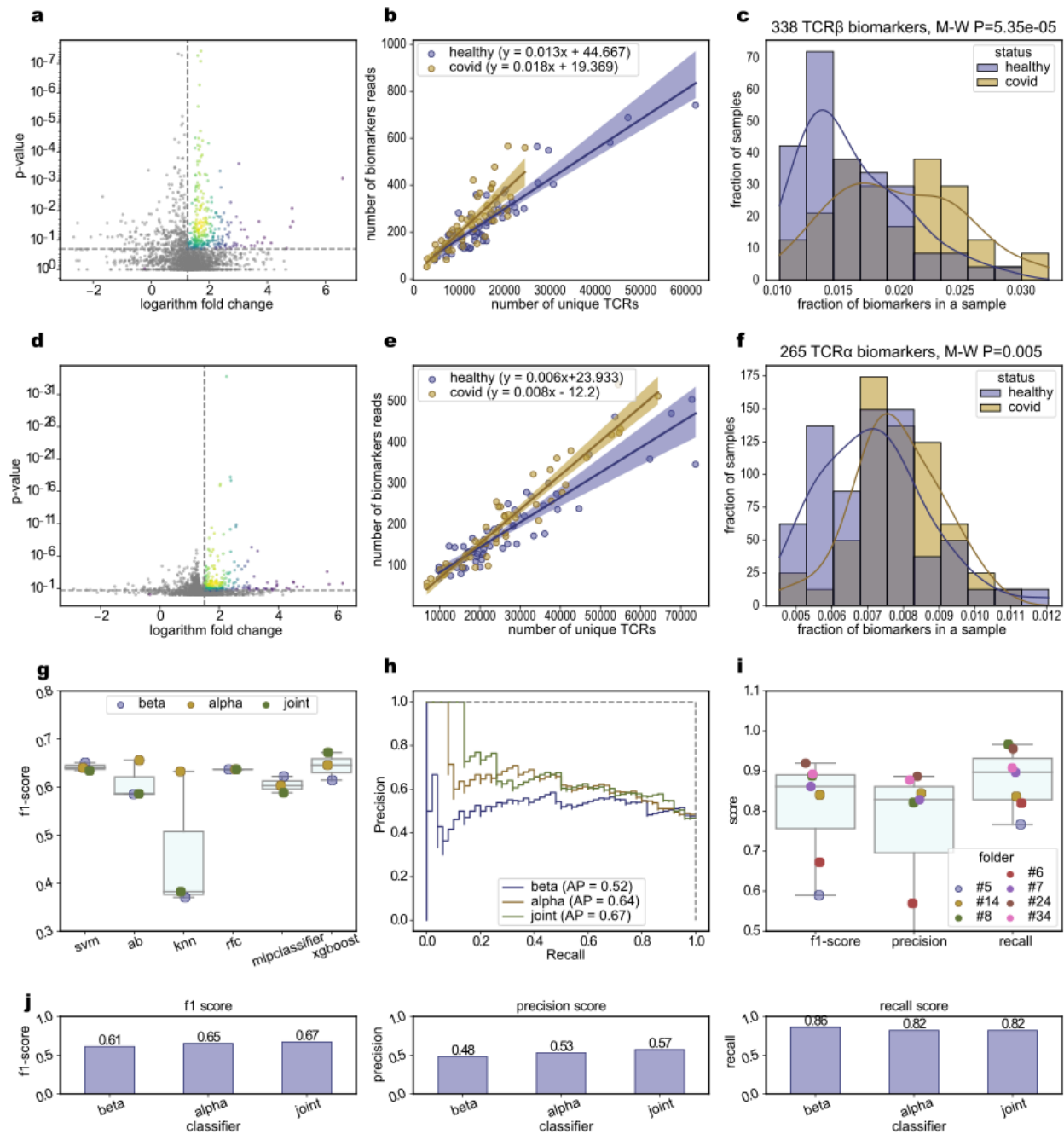
**Fig S16. Cohort II additional analysis. A.** A Fisher test-based approach revealed >20,000 statistically significant clonotypes. The volcano plot was used to select the clonotypes with the highest fold change, which is calculated as observed:expected incidence ratio in COVID-19 samples in comparison to healthy controls. **B.** Precision-Recall-curves for Cohort I- or Cohort II-based classifiers. Lines marked with “meta” sign correspond to classifiers built based on metaclonotypes. AP = average precision.



**Fig S17. Heatmaps of HLA-DRB1\*16 and HLA-DQB1\*05 presence in COVID-19 patients with clonotypes forming the CASSRTGXGSSYNSPLHF pattern. A.** Distribution of samples with motif between batches. The motif clonotypes are present in all the batches which demonstrates the absence of batch effect. **B.** Heatmap representing the linkage of HLAs of interest for all the COVID-19 patients with at least one read corresponding to pattern clonotypes of interest.



**Fig S18. Predicting donor HLA allele based on HLA-associated clonotypes.** **A.** Number of HLA positive donors in Cohort I. **B.** Number of clonotypes associated with presence of HLA for each allele. **C.** Average F1-scores for one-batch-left-out cross validation. Error bars show 25-75% quartiles for F1 score of each thrown out batch.



**Fig S19. Analysis of models built using VDJDdb TCR  $\alpha$  and  $\beta$  SARS-CoV-2 clonotypes.** **A.** Volcano plot for VDJDdb TCR $\beta$  SARS-CoV-2 clonotypes. The fold change of clonotype usage between healthy and convalescent cohorts in comparison to Fisher exact test p-value is shown. The dashed lines show the selected thresholds. Multiple hypothesis correction was not applied. **B.** Scatter plot showing the frequency of VDJDdb TCR $\beta$  clonotypes plotted against the number of unique TCR $\beta$  clonotypes. **C.** Distribution of COVID-associated TCR $\beta$  clonotypes across healthy and convalescent samples. **D,E,F.** The same as (a-c), but for VDJDdb TCR $\alpha$  clonotypes. **G.** Distribution of f1-scores for biomarker sets across different models. SVM shows most stable results. **H.** The precision-recall curve showing the difference in model performance. **I.** Evaluation of target metrics (f1-score, precision, recall) for one-batch-left-out cross validation. The metrics were evaluated for the TCR  $\alpha$ + $\beta$  biomarkers set. The boxplots represent the distribution of scores between different evaluation batches. **J.** Evaluation of target metrics for different biomarker sets.