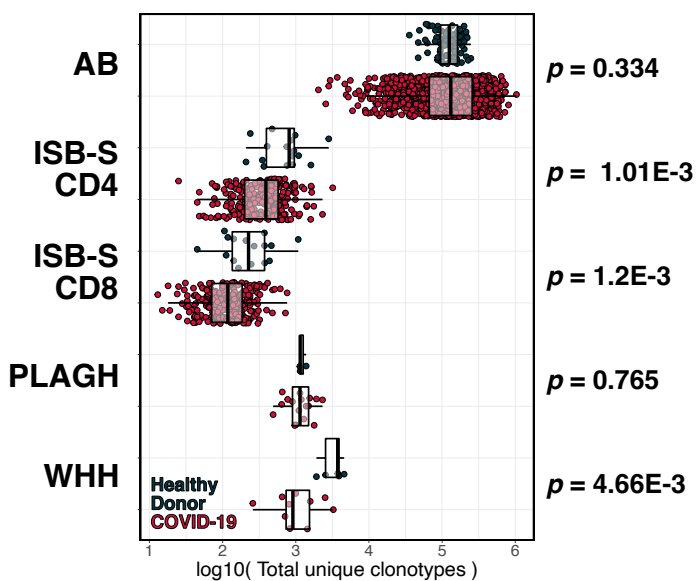
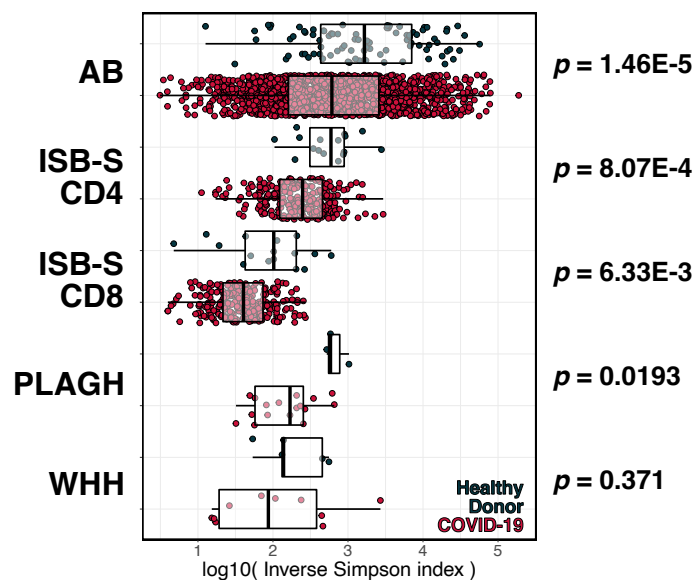


# Figure S1

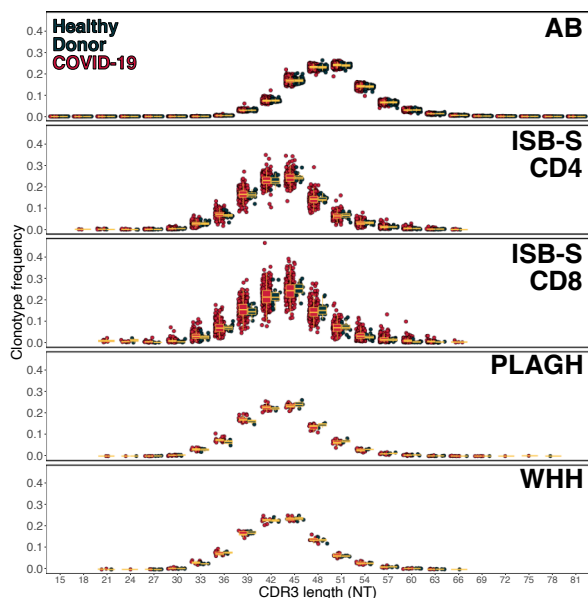
**a**



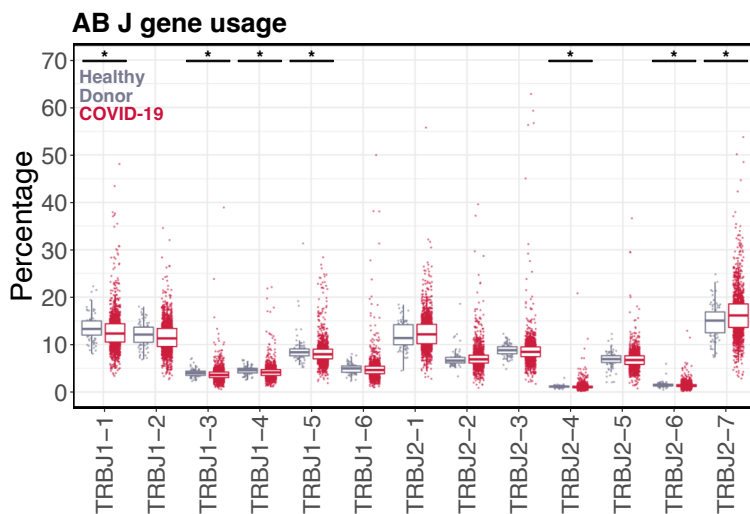
**b**



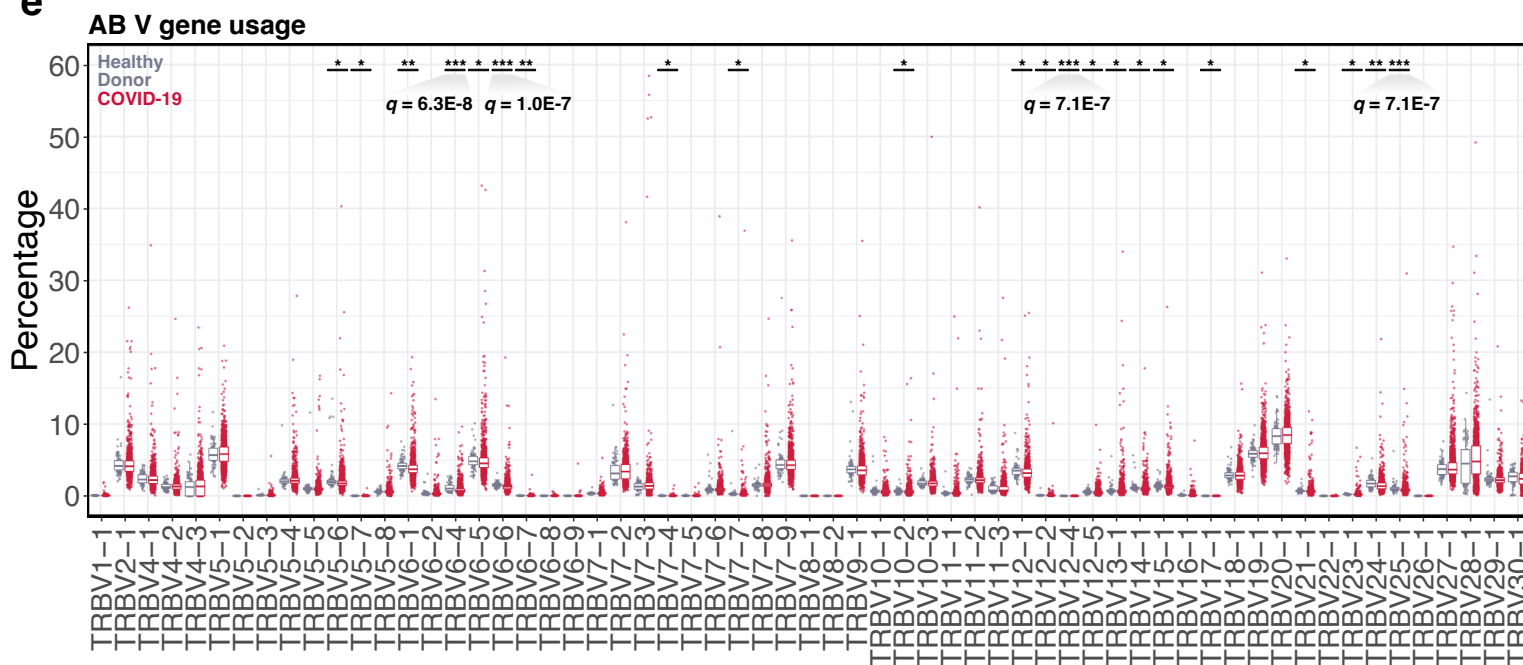
**c**



**d**



**e**



## Supplemental Figure Legends

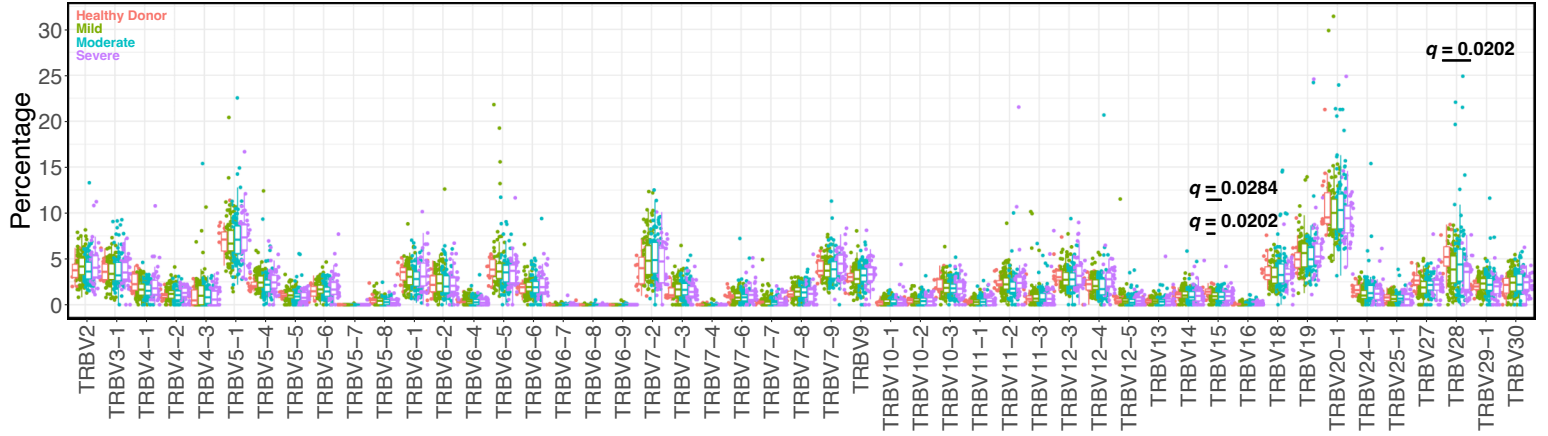
### Figure S1. Diversity metrics and gene usages for TCR repertoire datasets.

- (a) Boxplot of total unique clonotypes for COVID-19 patients and healthy donors for each repertoire dataset. P-values were obtained using the two-sided Wilcoxon rank-sum test.
- (b) Boxplot of inverse Simpson indices for COVID-19 patients and healthy donors for each repertoire dataset. P-values were obtained using the two-sided Wilcoxon rank-sum test.
- (c) Boxplot of clonotype frequencies by CDR3 length for COVID-19 patients and healthy donors for each repertoire dataset.
- (d) Boxplot of J gene usages for samples from the AB dataset. Gray dots represent healthy donor samples; red dots represent COVID-19 samples. Statistical significance determined using the two-sided Wilcoxon rank-sum test and adjusted using the Benjamini & Hochberg method. \* adj.  $P < 0.05$ , \*\* adj.  $P < 1e-4$ , \*\*\* adj.  $P < 1e-6$ .
- (e) Boxplot of V gene usages for samples from the AB dataset. Gray dots represent healthy donor samples; red dots represent COVID-19 samples. Statistical significance determined using the two-sided Wilcoxon rank-sum test and adjusted using the Benjamini & Hochberg method. \* adj.  $P < 0.05$ , \*\* adj.  $P < 1e-4$ , \*\*\* adj.  $P < 1e-6$ .

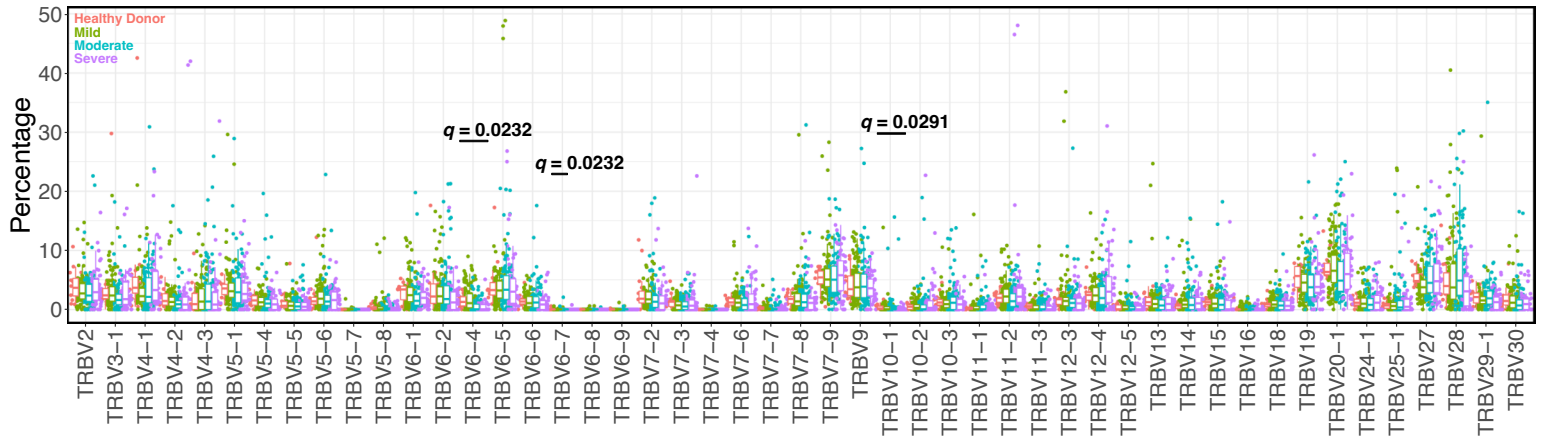
# Figure S2

**a**

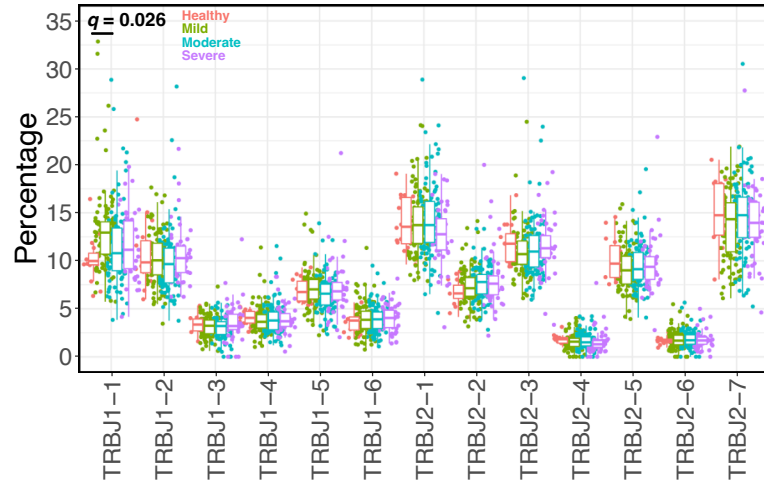
## ISB-S CD4 V gene usage

**b**

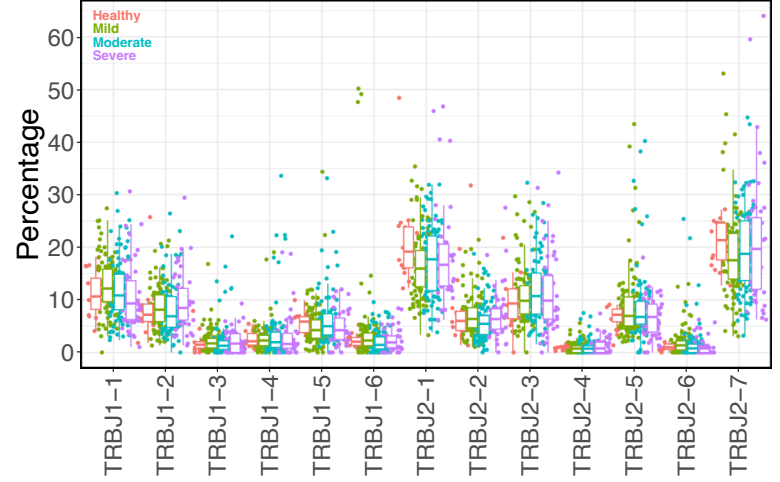
## ISB-S CD8 V gene usage

**c**

## ISB-S CD4 J gene usage

**d**

## ISB-S CD8 J gene usage



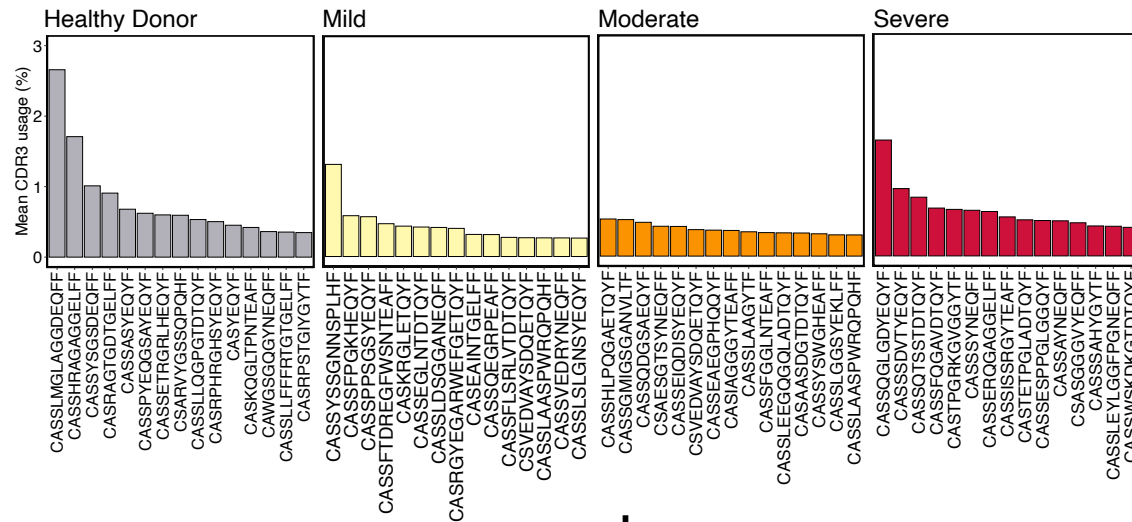
**Figure S2. V and J gene usages by disease severity for ISB-S datasets.**

- (a) Boxplot of V gene usages for CD4 samples from the ISB-S dataset. Red dots represent healthy donor samples; green dots represent mild; blue dots represent moderate; purple dots represent severe. Statistical significance determined using the two-sided Wilcoxon rank-sum test and adjusted using the Benjamini & Hochberg method. Adj.  $P < 0.05$  labelled on plot.
- (b) Boxplot of V gene usages for CD8 samples from the ISB-S dataset.
- (c) Boxplot of J gene usages for CD4 samples from the ISB-S dataset.
- (d) Boxplot of J gene usages for CD8 samples from the ISB-S dataset.

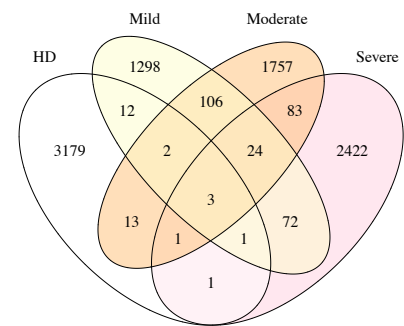


# Figure S3

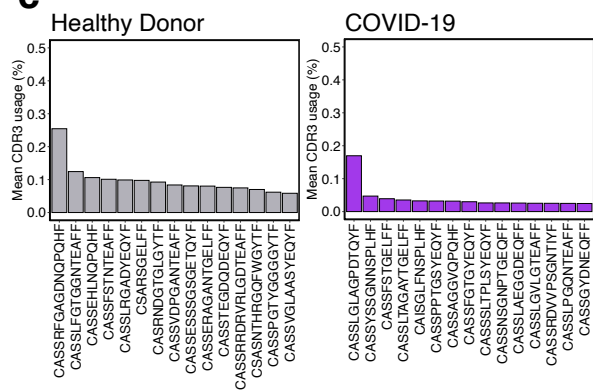
**a**



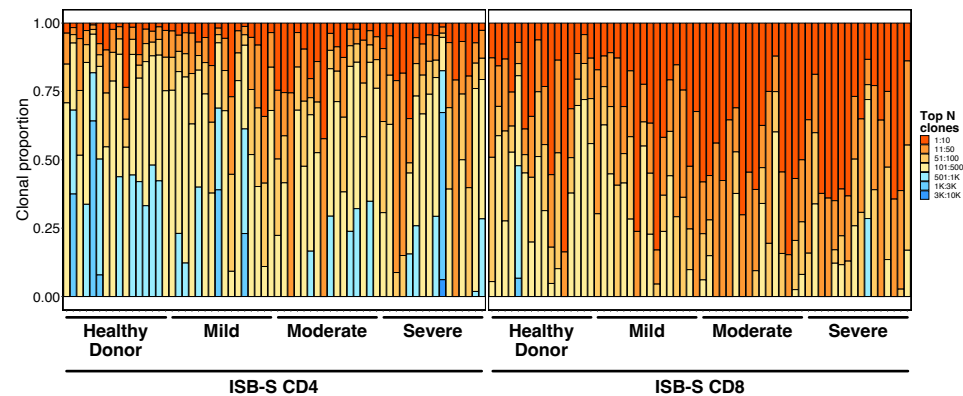
**b**



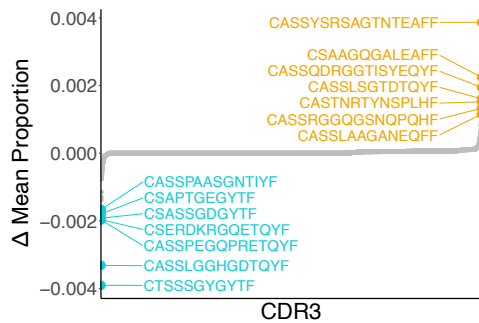
**c**



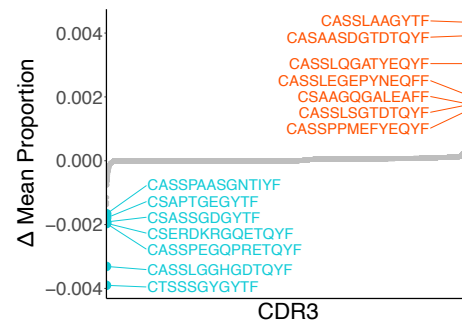
**d**



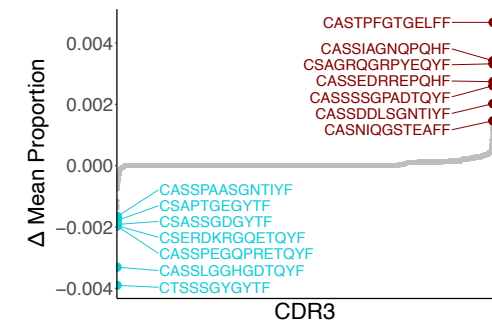
**e**



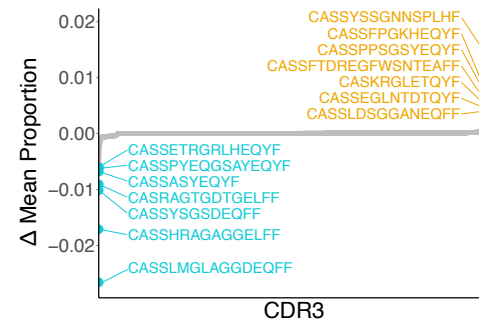
**f**



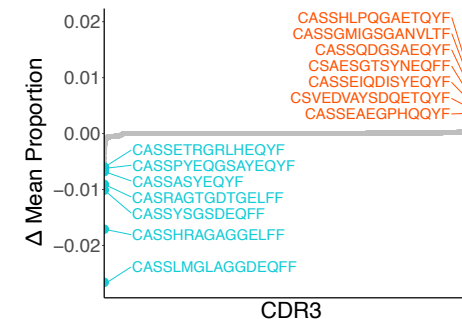
**g**



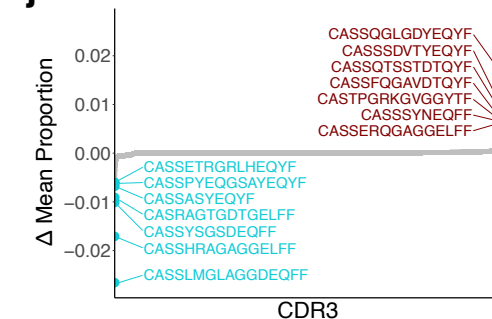
**h**



**i**



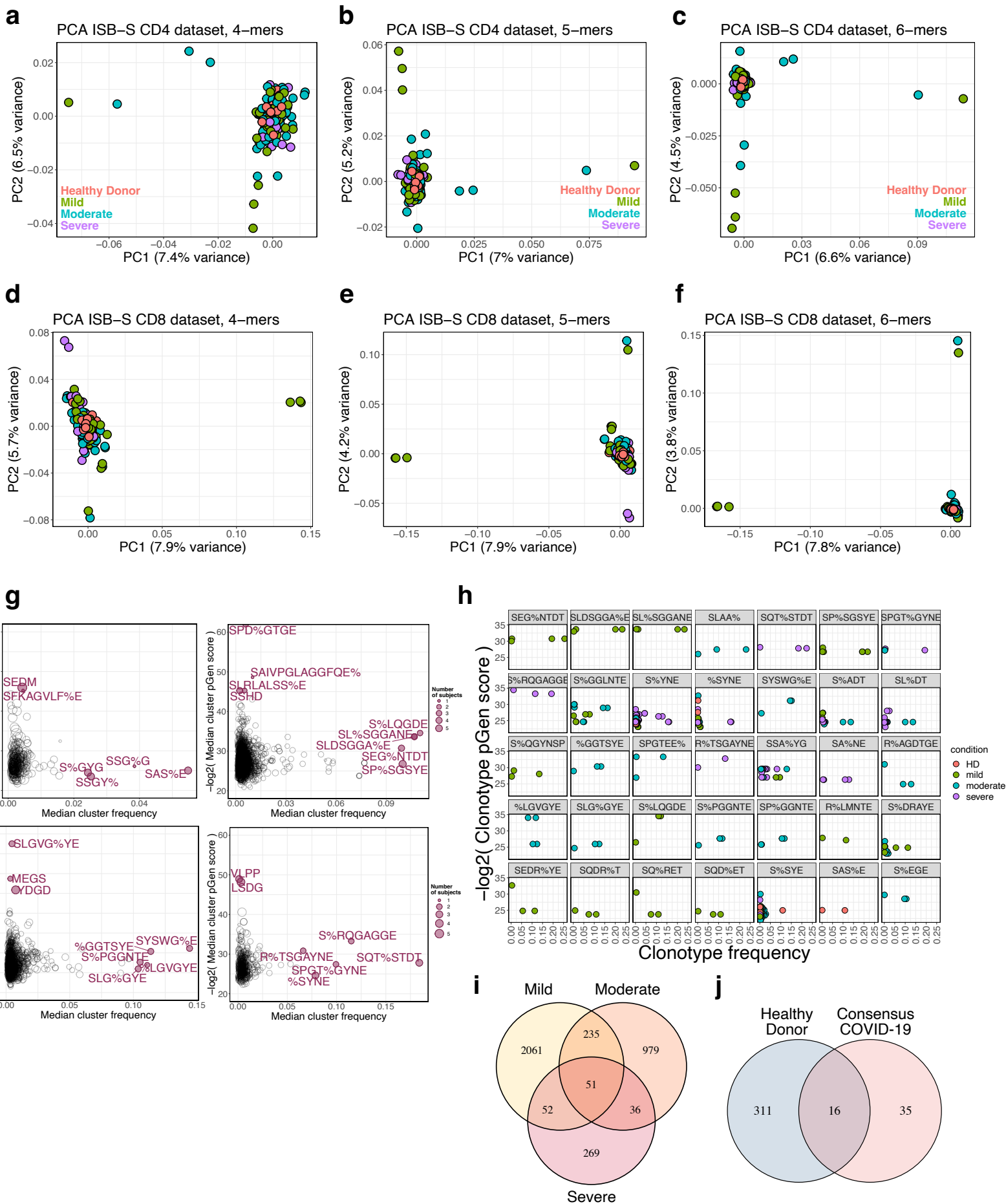
**j**



**Figure S3. Additional CDR3 gene usage statistics.**

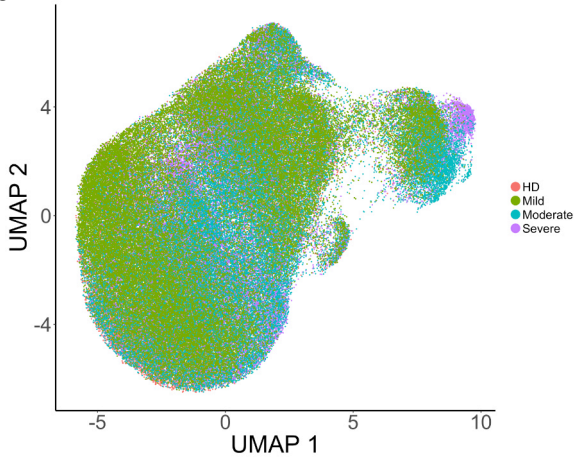
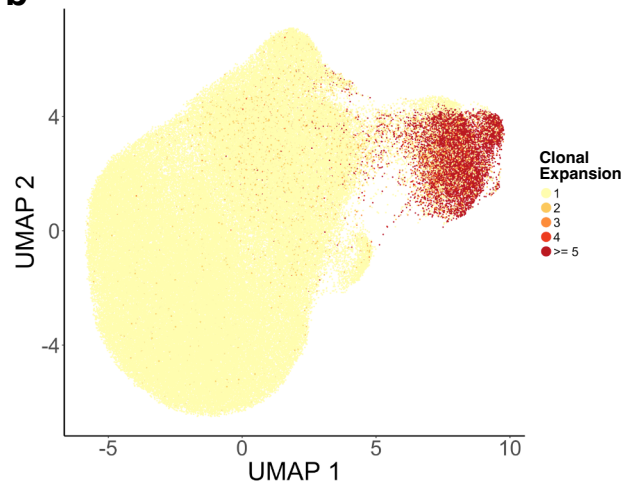
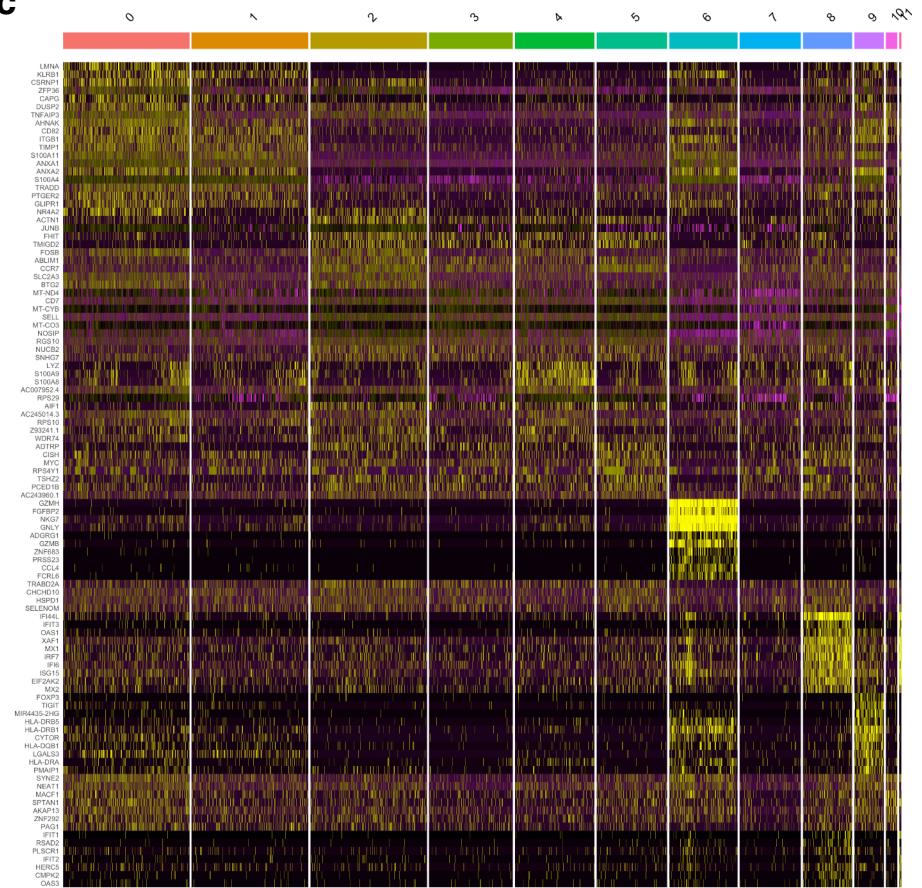
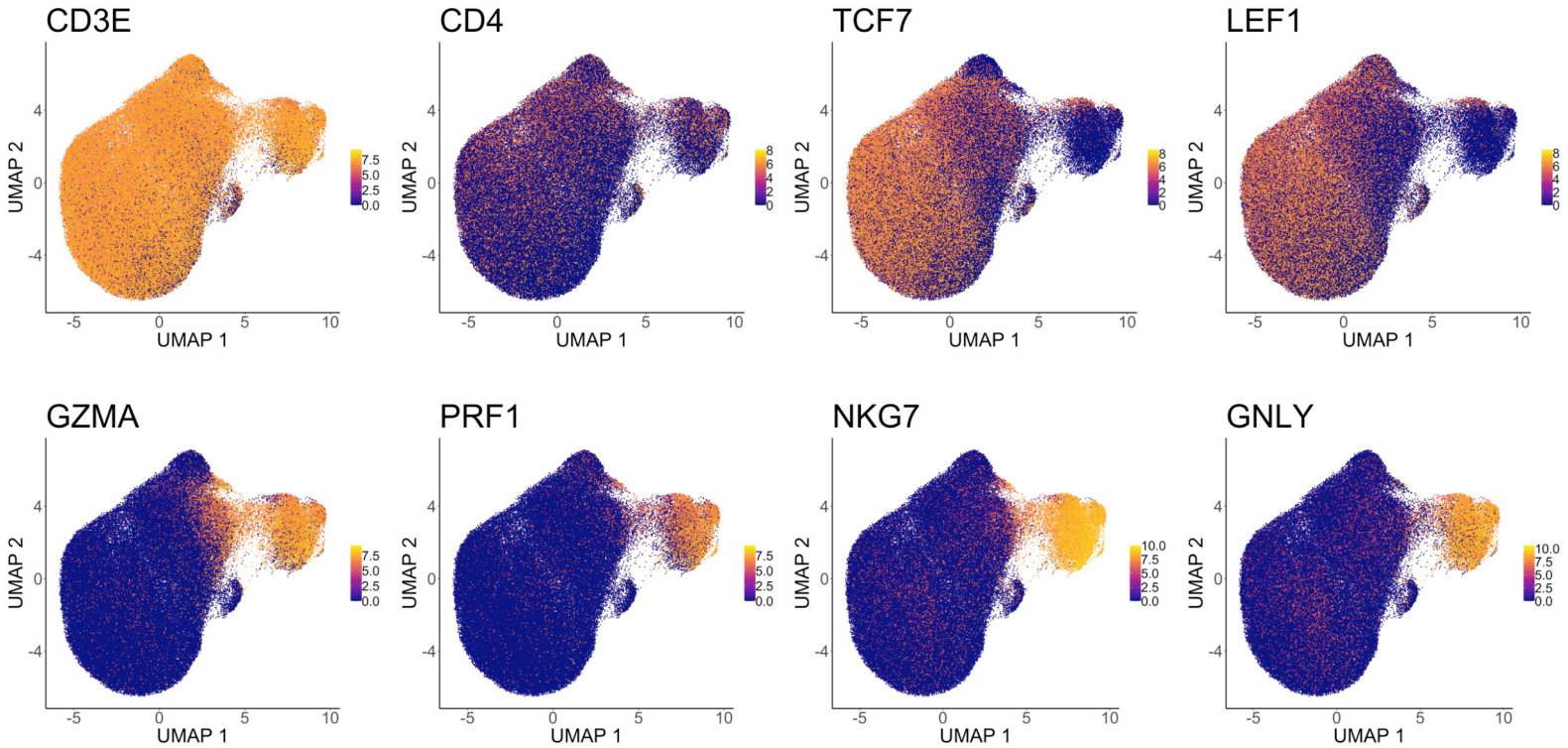
- (a) Bar plots showing the top 15 mean CDR3 usages for patients in the ISB-S CD8 dataset grouped by disease severity (healthy donor = 16, mild = 108, moderate = 93, severe = 49).
- (b) Venn diagram showing overlap of top mean CDR3 usages (proportion threshold = 0.0001) for patients in the ISB-S CD8 dataset grouped by disease severity.
- (c) Bar plots showing the top 15 mean CDR3 usages for patients in the AB dataset grouped by disease status (healthy donor = 88, COVID-19 = 1,475).
- (d) Bar plot depicting relative abundance for groups of top clonotypes by disease condition for sampled repertoires (n = 16 per condition) from ISB-S datasets.
- (e) Dotted waterfall plot of CDR3 gene usage differentials between mild disease COVID-19 patients and healthy donors (delta mean proportion) in the ISB-S CD4 dataset. Yellow dots are CDR3 sequences enriched in moderate disease repertoires; light blue dots are CDR3 sequences enriched in healthy donors; grey dots are all other CDR3 sequences.
- (f) Dotted waterfall plot of CDR3 gene usage differentials between moderate disease COVID-19 patients and healthy donors in the ISB-S CD4 dataset. Orange dots are CDR3 sequences enriched in moderate disease repertoires.
- (g) Dotted waterfall plot of CDR3 gene usage differentials between severe disease COVID-19 patients and healthy donors in the ISB-S CD4 dataset. Red dots are CDR3 sequences enriched in severe disease repertoires.
- (h) Dotted waterfall plot of CDR3 gene usage differentials between mild disease COVID-19 patients and healthy donors in the ISB-S CD8 dataset. Yellow dots are CDR3 sequences enriched in mild disease repertoires.
- (i) Dotted waterfall plot of CDR3 gene usage differentials between moderate disease COVID-19 patients and healthy donors in the ISB-S CD8 dataset. Orange dots are CDR3 sequences enriched in moderate disease repertoires.
- (j) Dotted waterfall plot of CDR3 gene usage differentials between severe disease COVID-19 patients and healthy donors in the ISB-S CD8 dataset. Red dots are CDR3 sequences enriched in severe disease repertoires.

# Figure S4



**Figure S4. Additional k-mer and motif analyses.**

- (a) PCA of 4-mer representations of TCR repertoires from the ISB-S CD4 dataset.
- (b) PCA of 5-mer representations of TCR repertoires from the ISB-S CD4 dataset.
- (c) PCA of 6-mer representations of TCR repertoires from the ISB-S CD4 dataset.
- (d) PCA of 4-mer representations of TCR repertoires from the ISB-S CD8 dataset.
- (e) PCA of 5-mer representations of TCR repertoires from the ISB-S CD8 dataset.
- (f) PCA of 6-mer representations of TCR repertoires from the ISB-S CD8 dataset.
- (g) Median frequency and pGen scores of COVID-19 and healthy donor associated T cell clusters from GLIPH2 analysis of the ISB-S CD8 dataset, grouped by disease condition.
- (h) Detailed view of frequencies and pGen scores of specific clonotypes associated with high frequency T cell clusters from CD8 dataset. Clonotypes are colored by patient disease condition.
- (i) Venn diagram showing overlap of COVID-19-associated T cell clusters for patients in the ISB-S CD8 dataset grouped by disease condition.
- (j) Venn diagram showing overlap between consensus COVID-19-associated T cell clusters (taken from intersection of disease conditions) and healthy donors for repertoires in the ISB-S CD8 dataset.

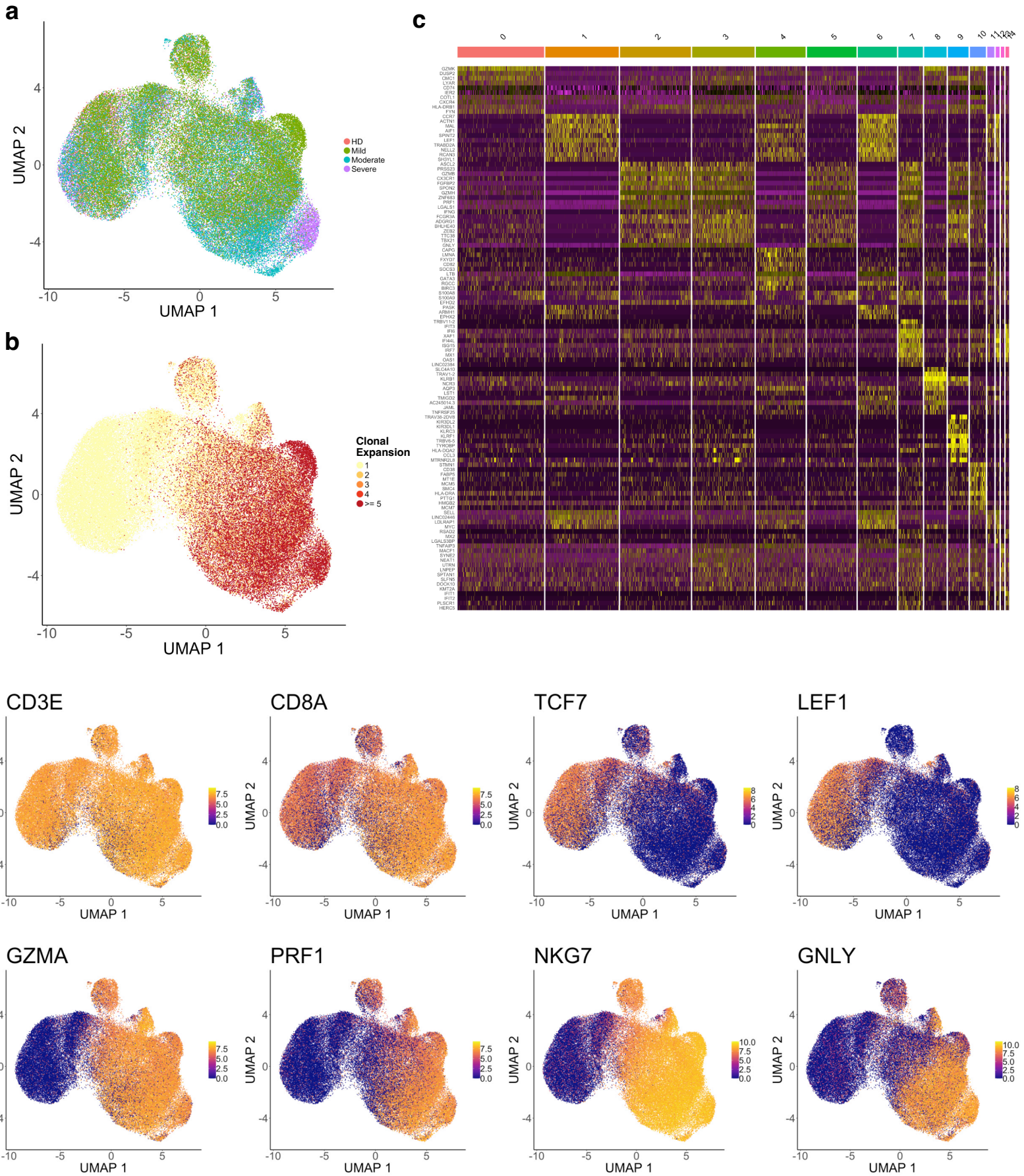
**Figure S5****a****b****c****d**

**Figure S5. Additional single-cell transcriptional analyses for CD4 T cells.**

- (a) UMAP visualization of 137,075 CD4 T cell single-cell transcriptomes from the ISB-S CD4 dataset labelled by disease condition.
- (b) UMAP visualization of CD4 T cell single-cell transcriptomes labelled by clonal expansion.
- (c) Heatmap of differentially expressed markers for all identified clusters (n = 12).
- (d) UMAP visualizations highlighting expression levels of individual genes for cell phenotyping.



**Figure S6**



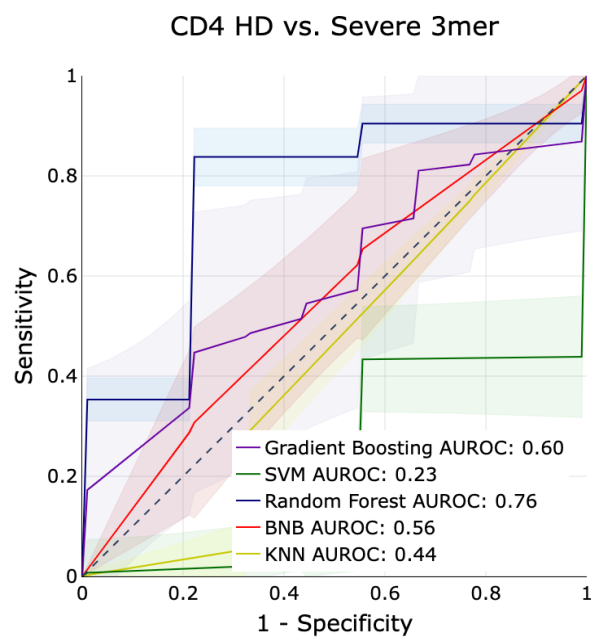
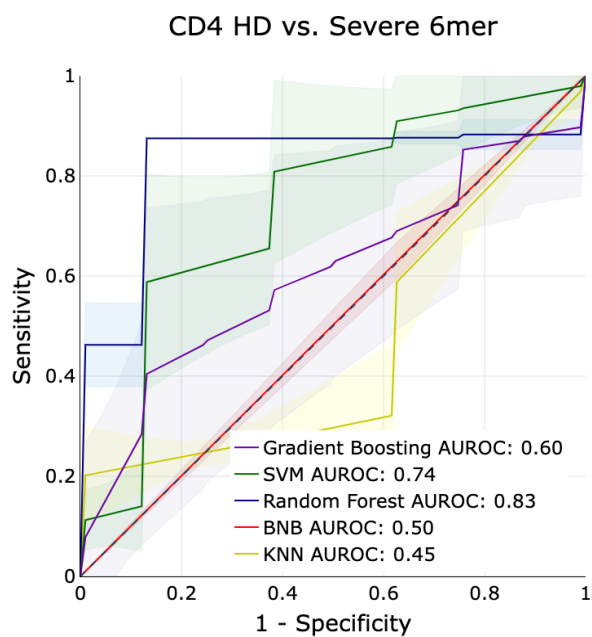
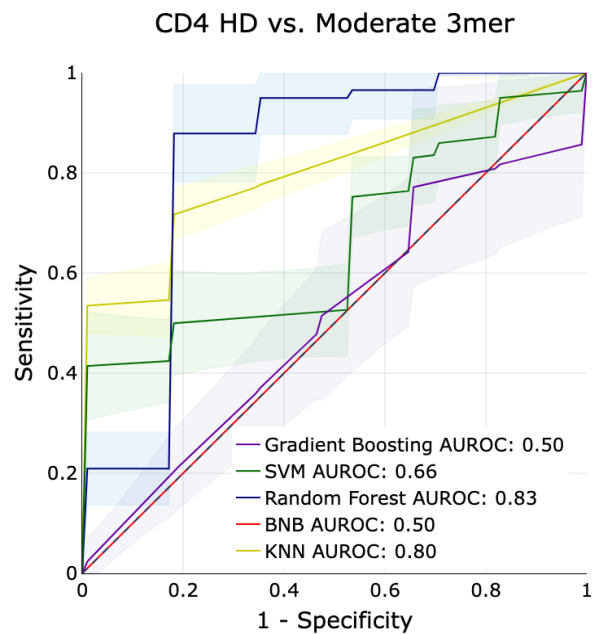
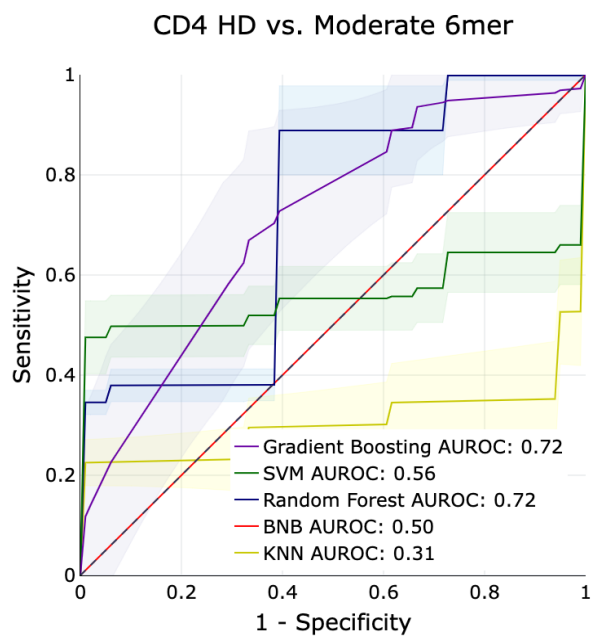
**Figure S6. Additional single-cell transcriptional analyses for CD8 T cells.**

- (a)** UMAP visualization of 70,237 CD8 T cell single-cell transcriptomes from the ISB-S CD8 dataset labelled by disease condition.
- (b)** UMAP visualization of CD8 T cell single-cell transcriptomes labelled by clonal expansion.
- (c)** Heatmap of differentially expressed markers for all identified clusters (n = 15).
- (d)** UMAP visualizations highlighting expression levels of individual genes for cell phenotyping.



# Figure S7

a



**Figure S7. Additional machine learning analyses for CD4 T cell subset.**

**(a)** AUROC curves for five machine learning models using 3-mer and 6-mer representations of TCR repertoire data from the ISB-S CD4 dataset. Models were trained to predict disease severity (moderate, severe) vs healthy donors for CD4 samples. Training and evaluation was performed using 100 repetitions of 5-fold cross-validations per model, average performance +/- 1 standard deviation shown on individual plots.

## **Supplementary Datasets**

### **Supplementary Dataset**

Dataset S1. Metadata for TCR repertoire samples obtained for all datasets used in study.

Dataset S2. Number of clones and unique clonotypes for each sample across datasets.

Dataset S3. CDR3 length statistics for each sample across datasets.

Dataset S4. Diversity statistics including Chao1 estimators, Gini-Simpson indices, and inverse Simpson indices for each sample across datasets.

Dataset S5. V and J gene usage statistics for each sample in Adaptive Biotechnologies datasets.

Dataset S6. V and J gene usage statistics for each sample in ISB-Swedish datasets.

Dataset S7. Principal components analysis results for 3-mer, 4-mer, 5-mer, and 6-mer representations of each sample in ISB-Swedish datasets.

Dataset S8. GLIPH clustering analysis patterns, scores, and statistics for ISB-Swedish datasets by T cell type and disease condition.

Dataset S9. OLGA analysis inputs of structured ISB-Swedish datasets by T cell type and disease condition.

Dataset S10. OLGA analysis output pGen scores of ISB-Swedish datasets by T cell type and disease condition.

Dataset S11. COVID-19-associated clusters in ISB-Swedish datasets by T cell type.

Dataset S12. UMAP coordinates for CD4 and CD8 T cell single-cell transcriptome analyses.

Dataset S13. Cell proportions and counts for clonally expanded groups from CD4 and CD8 T cell single-cell transcriptome analyses.

Dataset S14. Differential gene expression for Cluster 6 vs all other cells in CD4 T cell transcriptome analysis and Expanded group cells vs all other cells in CD8 T cell transcriptome analysis.

Dataset S15. Upregulated and downregulated genes for CD4 and CD8 T cell clonal expansion differential gene expression analysis using threshold q-value  $< 1e-4$ .

Dataset S16. DAVID gene ontology biological process annotations for CD4 and CD8 T cell clonal expansion differential gene expression analysis using threshold q-value  $< 1e-4$ .

Dataset S17. Average AUROC scores for machine learning models trained to predict disease severity from healthy donors using different k-mer and GLIPH2 representations of TCR repertoires.