

## Methods

### Settings

The study was performed on deidentified, cryopreserved PBMC samples of 10 COVID-19 patients and 13 matched controls, obtained with informed consent on a protocol approved by Yale Human Research Protection Program Institutional Review Boards (FWA00002571, Protocol ID. 2000027690).

### Patients and samples

Ten COVID-19 patients hospitalized at Yale-New Haven Hospital (YNHH) were recruited for this study. All were confirmed to have COVID-19 by RT-PCR testing of nasopharyngeal samples. Four of the patients were "Progressive" (TP6, TP7, TP8, TP9), defined as patients who required admission to the intensive care unit (ICU) and eventually succumbed to the disease. At the same time, the other 6 were "Stable" (NS0, NS1, TS2, TS3, TS4, TS5), defined as patients hospitalized in non-ICU internal medicine wards who were eventually discharged. Eight patients (80%) were treated with Tocilizumab, a humanized anti-IL6 receptor antibody. Only patients NS0 and NS1 did not receive this drug (designated with "N"). Tocilizumab was given once at a dose of 8 mg/kg (up to a maximal dose of 800 mg). All patients were treated with antivirals (Atazanavir, except for patient NS1 which was treated with Remdesivir) and with Hydroxychloroquine (except for NS1). Two progressive subjects were treated with corticosteroids: patient TP7 was treated with Prednisone 40 mg daily for 2-3 days just before blood draw A, and patient TP9 was treated with Methylprednisolone 120 mg daily for 1-2 days prior to blood draw B. No other immunosuppressive, immunomodulatory, or antiviral agents were used.

Eighteen blood samples were collected from these ten patients, at different time-points as described in the results section and Fig 1A-B. Thirteen control subjects were recruited prior to the COVID-19 pandemic. Baseline characteristics of COVID-19 patients and controls are presented in Supp Table ST1. The timing of symptom onset, hospitalization, tocilizumab treatment, and blood draws for each patient is shown in Fig 1D.

### Isolation of PBMC and cryopreservation

PBMCs were isolated from whole blood using density gradient centrifugation, according to the following protocol: Histopaque 20 ml was added to a 50 ml SepMate tube, then overlaid with fresh blood 1:1 diluted in PBS 2% fetal bovine serum (FBS)

and centrifuged at 1200 x g for 10 minutes. The PBMC layer was collected by quickly pouring the remaining contents above the SepMate insert into a fresh tube, and washed once with PBS at 650 x g for 10 min. The supernatant was decanted and ACK red blood cell lysis buffer (2ml/sample) was added for 2 minutes; another wash with PBS 2% FBS was done, followed by centrifugation at 290 x g to remove platelets and supernatant aspirated. Following resuspension of the pellet, PBMCs were cryopreserved in aliquots of  $5 \times 10^6$  cells using 10% DMSO in heat inactivated-FBS as the cryopreservation solution. Cryovials were placed in a freezing container (Mr. Frosty) and transferred immediately to a -80 °C freezer for >24 hours before being transferred to long-term liquid nitrogen storage.

#### Sample preparation and 10x barcoding

All sample processing steps were done in a biosafety level 2+ laboratory. Samples were thawed in a water bath at 37°C for ~2 min without agitation, and removed from the water bath when a tiny ice crystal still remains. After thawing, cells were gently transferred to a 50 mL conical tube using a wide-bore pipette tip, the cryovial was rinsed with cold growth medium (10% FBS in DMEM) to recover leftover cells, and the rinse medium was added dropwise (1 drop per 5 sec) to the 50 mL conical tube while gently shaking the tube. Next, we conducted serial dilutions with cold growth medium a total of 5 times by 1:1 volume addition with ~1 min wait between additions. Cold growth medium was added at a speed of 3-5 ml/sec, achieving a final volume of 32 mL. The cells were then centrifuged at 300 x g for 5 minutes at 4°C, and the supernatant was removed without disrupting the cell pellet. The pellet was resuspended in 1X PBS with 0.04% BSA, and the sample was filtered with a 40 µM strainer. Cell concentration was determined using Trypan blue staining with a Countess automated cell counter (ThermoFisher). Following this cell count, each sample was split into two parts (Fig 1C): one was immediately loaded onto the 10x Chromium Next GEM Chip G, according to the manufacturer's user guide (document number CG000208, revision E, February 2020), and the other was further processed for CITE-seq as described in the next section, and then loaded to the 10x Chromium Chip G. In total, we loaded 18 'conventional' samples into 18 Chip G lanes (aiming for recovery of 10,000 cells per lane), and 17 out of 18 'CITE-seq' samples into 6 Chip G lanes (each lane containing 5-6 pooled hashed samples, as portrayed in Supp Table ST7). One out of 18 CITE-seq samples (TP8B) was not pooled because of very low cell concentration.

#### CITE-seq and cell hashing

The lyophilized Total-seq C human panel (BioLegend) was resuspended with 35  $\mu$ L of wash buffer, vortexed for 10 sec and incubated for 5 min at RT. Total-seq C human Hashtag antibodies (Biolegend) were centrifuged at 20000g for 10 min and 6-fold diluted with wash buffer (2% FBS and 1mM EDTA in PBS). To maximize performance, both were centrifuged at 20000g for 10 min just before adding to the cells. See Supp Table ST6 for a list of antibodies, clones, and barcodes used for CITE-seq and hashing samples.

PBMCs from each sample were reconstituted with wash buffer at the concentration of 10-20  $\times 10^6$  cells/ml and incubated on ice for 10 mins with 5ul of Human Fc block (BD Biosciences) and 5  $\mu$ L of TrueStain Monocyte Blocker (Biolegend). 10-20  $\mu$ L (0.1-0.2  $\times 10^6$  cells) were transferred into a new tube and incubated on ice for 30 mins with 5  $\mu$ L of CITE-seq panels and 5  $\mu$ L of Hashtag antibodies prepared as above. Cells were washed twice with wash buffer and with 2% FBS in PBS for the third wash. Samples were pooled into one tube based on cell counts, and super-loaded onto the 10x Chromium Chip G, aiming for recovery of  $\sim$ 20,000 cells per sample. See Supp Table ST6 for the details of 6 pooled samples (CITE#1-CITE#6).

#### cDNA libraries preparation and sequencing

The loaded Chip G was placed in the 10x Chromium controller to create Gel Beads-in-emulsion (GEMs). The next steps were carried out according to the manufacturer's user guide, including GEM-RT incubation, post GEM-RT Dynabead cleanup, and cDNA amplification. The cDNA samples were used to construct 4 types of cDNA libraries, according to the steps outlined in the user guide: gene expression libraries, T-cell receptor libraries, B-cell receptor libraries, and cell surface protein libraries (the latter only for samples processed with CITE-seq). cDNA libraries were then sequenced on an Illumina Novaseq 6000 platform.

#### Flow cytometry

Freshly isolated PBMCs were incubated with Fc block reagent (Biolegend) for 10min and stained with LIVE/DEAD Fixable Aqua Dead Cell Stain kit (Thermo Fisher) for 20 minutes at 4°C. Following a wash, cells were then blocked with Human TruStan FcX (BioLegend) for 10 minutes at RT. Cocktails of following antibodies were directly added to this mixture for 30 minutes at RT. BB515 anti-HLA-DR (G46-6), BV605 anti-CD3 (UCHT1), BV785 anti-CD4 (SK3), APCFire750 anti-CD8 (SK1), BV421 anti-CCR7 (G043H7), AlexaFluor 700 anti-CD45RA (HI100), PE anti-PD1 (EH12.2H7), APC anti-TIM3 (F38-2E2), BV711 anti-CD38 (HIT2), BB700 anti-CXCR5 (RF8B2), PE-Cy7 anti-CD127 (HIL-7R-M21), PE-CF594 anti-CD25 (BC96), BV711 anti-CD127 (HIL-7R-

M21). Cells were washed two times with staining buffer and acquired on a BD Fortessa flow cytometer. FlowJo software (Treestar) was used for analysis.

#### SARS-CoV-2 viral load measurements

Nasopharyngeal swabs and saliva samples were collected from COVID-19 diagnosed inpatients at -Yale-New Haven Hospital, as described elsewhere<sup>1</sup>. We extracted total nucleic acid using the MagMax Viral/Pathogen Nucleic Acid Isolation kit (ThermoFisher Scientific, Waltham, MA, USA) with 300  $\mu$ L of input sample eluted into 75  $\mu$ L, using a slightly modified protocol ([dx.doi.org/10.17504/protocols.io.bg3pjymn](https://doi.org/10.17504/protocols.io.bg3pjymn)). A total of 5  $\mu$ L of extracted nucleic acid was used as input in the RT-qPCR assay for SARS-CoV-2 detection, as described elsewhere<sup>2</sup>. Briefly, we used the Luna Universal Probe One-Step RT-qPCR kit (New England Biolabs, Ipswich, MA, USA) with the CDC 2019-nCoV\_N1, 2019-nCoV\_N2, and human RNase P (RP) primer-probe sets (Integrated DNA Technologies, Coralville, IA, USA). Viral RNA copy numbers were calculated based on 10-fold dilution standard curves of the previously generated nucleocapsid (N) transcript standard<sup>2</sup>.

#### Data processing of raw sequencing reads

Raw sequencing reads were demultiplexed using Cell Ranger mkfastq pipeline to create FASTQ files. Next, Cell Ranger count pipeline (v3.1) was employed in order to perform alignment (using STAR), filtering, barcode counting, and UMI counting. We have used GRCh38 (Ensembl 93) as the genome reference (corresponding to Cell Ranger reference GRCh38-3.0.0).

#### ScRNA-seq sample aggregation

10x cellranger count filtered output data of PBMCs from thirteen healthy controls were added to that of the eighteen COVID-19 samples. Seurat package<sup>3, 4</sup> (v3.1) was used for all downstream analyses. 10x gene expression matrices for each sample were converted and combined into one Seurat object. Cells with mitochondrial gene percentages higher than 12% and cells with less than 200 genes were excluded from the study to filter out dead and dying cells. For CITE-seq samples, following de-hashing, cell barcodes of multiplets (i.e. with 2 or more hashing antibody signals) or uncertain origin (i.e with no clear hashing signal) were also removed. After these filtering steps, the gene-barcode matrix contained 35,538 genes and 163,452 barcoded cells.

#### Integration, principal components analysis and clustering

In accordance with the standard Seurat pre-processing workflow, sample gene expressions were normalized using Seurat's "LogNormalize" method<sup>3, 4</sup>. The "FindVariableFeatures" function selected the 3,000 genes with the highest variance to mean ratio using the "vst" method. To remove single-subject effects, samples were integrated on a subject level using 2000 anchors with a dimensionality of 30<sup>4</sup>. The integrated data was then scaled with the "ScaleData" function.

Principal Component Analysis (PCA) was performed on the integrated data, and the first 30 Principal Components (PCs) were used in the "FindNeighbors" algorithm. The Louvain modularity optimization algorithm in "FindClusters" generated the clusters while the resolution was set to 0.75. Thirty PCs were used in the "RunUMAP" function to create the final UMAP, and thirty clusters were generated from the aforementioned pipeline (Supp Fig S2).

These thirty clusters were first annotated with the SingleR software (Supp Fig S3) and then annotated manually (Fig 1F) by using cell-specific markers (Supp Fig S1) plotted on UMAP space, and by examining the output of "FindAllMarkers" per cluster. Five clusters out of thirty were removed; namely: a nonspecific cluster of low UMI cells (cluster #8), monocyte-platelet multiplets (#22), B and T/NK multiplets (#24), erythroid cell contamination from a single subject (#25), and B cell-platelet multiplets (#29). Following the removal of these clusters, the final Seurat object contained 153,554 cells.

#### Cell Type Proportions Analysis

For each subject, the number of cells within a given cell type was normalized by the subject's total number of cells. For each cell type, cell proportions were plotted in a boxplot by disease group, namely by (1) controls, (2) stable patients at time point A, (3) stable patients at time point B, (4) progressive patients at time point A, and (5) progressive patients at time point B.

#### Differential Gene Expression Analysis

The "FindMarkers" function was used to identify differentially expressed genes (DEGs) across the following conditions: per cell type, (1) time point A versus time point B; (2) controls versus COVID-19 patients at time point A; and (3) progressive vs stable. For the time point comparison (comparison 1), the logistic regression test for differential expression with subjects set as latent variables was used to account for paired samples. For comparisons (2) and (3), the default Wilcoxon rank test was used. Genes were ranked by absolute log<sub>2</sub> fold-change (logFC), and those with p-values > 0.05 (adjusted for multiple comparisons) were removed.

### Heatmap visualization of DEGs

DEGs were visualized as heatmaps which were generated by using the ComplexHeatmap package<sup>5</sup>. Cell types were binned into monocytes, CD4 T cells, CD8 T cells, and B cells and “FindMarkers” distinguished DEGs for each cell type bin for (1) time point A versus time point B and (2) progressive versus stable. Genes with greater than 0.5 absolute logFC were included in visualization and EnrichR pathway analysis. Samples for the progressive versus stable time-point were hierarchically clustered.

### Gene Pathway Annotation

Gene list outputs from the “FindMarkers” function were fed into EnrichR for pathway and ontology analysis<sup>6, 7</sup>. Gene set enrichment analysis<sup>8</sup> was also performed on “Dividing T cells” cluster using KEGG<sup>9, 10, 11</sup> and MSigDB Hallmark gene sets<sup>12</sup>, and custom gene sets (Supp Table ST9).

### Gene List Score Analysis

Seurat Function “AddModuleScore” was used to combine the expression of genes from IFN Score A<sub>13</sub> (*ISG15, IFI44, IFI27, CXCL10, RSAD2, IFIT1, IFI44L, CCL8, XAF1, GBP1, IRF7, CEACAM1*). This function was also used to combine other gene list scores as well, including scores for HLA type II (*HLA-DRA, HLA-DQA1, HLA-DQA2, HLA-DPA1, HLA-DRB1, HLA-DPB1, HLA-DQB2, HLA-DRB5, HLA-DQB1, HLA-DMA, HLA-DMB*) and the IL6 pathway (*ARID5A, SOCS3, PIM1, BCL3, BATF, MYC*)<sup>14</sup>. Similarly, a platelet score was calculated with 6 genes that have high specificity for platelets within the PBMC compartment (*PPBP, TUBB1, PF4, CAVIN2, SPARC, CLU*). The differences in gene list scores were compared between (1) control versus COVID-19 patients, (2) time point A and time point B, and (3) progressive versus stable patients.

### Demultiplexing (de-hashing) of CITE-seq samples

In order to demultiplex cells in the CITE-seq samples and attribute them a biological sample, hashing antibody-derived tag (ADT) counts were normalized by library size, square-root transformed, and normalized for every row in the data matrix of each CITE-seq sample. To account for the inherent background noise of ADT and accurately identify a cell as tagged by a hashing ADT, histograms of each hashing ADT counts in each CITE-seq sample were used to determine the optimal threshold of significance for hashing ADTs. As distributions appeared bimodal for the majority of hashing, we manually set the threshold between the two modes.

Based on the previous threshold, data matrix rows with two or more significant ADT were flagged as doublets, and rows with zero significant ADT flagged as unidentified, thus removed for downstream analysis.

#### CITE-seq ADT preprocessing and downstream analysis

Once the cells were demultiplexed and hashing ADT counts were removed, the remaining ADT counts (192) for each CITE-seq sample were combined into one single matrix. The counts for the remaining 43,349 cells were normalized by library size and square root transformed. We visualized the dataset using Uniform Manifold Approximation and Projection (UMAP). Cells were clustered using the Louvain community detection on a 15-nearest neighborhood graph and were manually annotated using a panel of ADT markers for each cell type (cell types include: CD4 T cells, CD8 T cells, B cells, NK cells, monocytes, macrophages, DCs, plasma cells, neutrophils, eosinophils, platelets and red blood cells). We also manually annotated the clusters based on surface marker lists proposed by the Human Immunology Project Consortium (HIPC) (<https://www.immuneprofiling.org>) (Supp Table ST8). As gene expression (GEX) data from CITE-seq was incorporated in the standard scRNA-seq analysis, there were two different annotations: one based on GEX, and one based on ADT. To measure the concordance between the two annotations, the percentage of shared cells between each annotated cluster was computed (Fig 4C).

Differential expression analysis was performed using the Wilcoxon rank sum test, and p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg correction. Data preprocessing and analysis (for ADT analysis only) was performed in Python (version 3.8.0) using Scanpy (version 1.4.6)<sup>15</sup>.

#### Differential connectivity (connectomic) analysis

For connectomic analysis, the cell parcellation shown in Figure 1 was used except for IFN-activated CD8 T cells, which were lumped into the Effector T cell cluster. These data were then mapped against a version of the FANTOM5 database of ligand-receptor interactions, modified to include additional immunomodulatory cues of interest to the authors (Supp Table ST10). Each parcellation, in a given experimental condition, was then treated as a single signaling node for network analysis. Average expression values were calculated for all ligand and receptor genes on a per-cell-type basis. Then an unfiltered edgelist ('connectome') was created linking all producers of a ligand to all producers of a receptor, with associated quantitative edge attributes, as previously described.

To compare experimental conditions, the connectomes from two experimental conditions were directly compared to yield log-fold changes for the sending (ligand) side and receiving (receptor) side of all edges. In addition, a ‘perturbation score’ was calculated, which allows plotting of differential edges proportional to the degree of change, allowing both negative and positive log fold changes and incorporating information from both sides of a given edge. The perturbation score that we used was defined, for every cell vector from Cell<sub>i</sub> to Cell<sub>j</sub> for ligand-receptor mechanism (k), as:

$$score^{ijk} = \left| \log \left( \frac{Cell_i^{L_k^{test}} - Cell_i^{L_k^{control}}}{Cell_i^{L_k^{control}}} \right) \right| \times \left| \log \left( \frac{Cell_j^{R_k^{test}} - Cell_j^{R_k^{control}}}{Cell_j^{R_k^{control}}} \right) \right|$$

Edges were then plotted which 1) had >10% of the sending and receiving cluster expressing the given ligand and receptor, respectively, and 2) which displayed an adjusted p-value < 0.05 via a Wilcoxon Rank Sum test comparing identical cell types to each other across experimental conditions.

Chord diagram plotting was performed using a custom implementation of the *circlize* package with directed edge thickness between cell type nodes proportional to the above described perturbation score, scaled per-plot.

#### Tocilizumab treatment effect analysis

To investigate the treatment effects of tocilizumab on transcription levels for different cell types, we conducted differential expression analysis between the two sampling time points for patients in the tocilizumab treatment group and those in the non-tocilizumab group separately. The logFC from these two separate analyses, i.e. for the tocilizumab group and non-tocilizumab group, was scatter plotted for each cell type in order to identify genes in which the differential expression pattern observed between the two-time points is due to a treatment effect rather than the natural course of the disease progression (Fig 3D). In addition, we investigated the correlations across cell types and compared results between the tocilizumab and non-tocilizumab groups (Supp Fig S13B, C). Six IL-6 pathway related genes, which are known to be associated with tocilizumab treatment<sup>14</sup>, are highlighted in red (Fig 3D, Supp Fig S13A-C).

#### T cell receptor V(D)J data processing

The raw sequencing reads of the T cell receptor (TCR) libraries were processed using the Cell Ranger V(D)J pipeline by 10XGenomics™, which assembled read-pairs into V(D)J contigs for each cell, identified cell barcodes from targeted cells, annotated the assembled contigs with V(D)J segment labels and located the CDR3 regions. We only



considered V(D)J contigs with high confidence defined by cell ranger under the default settings for downstream analysis. Contigs that were not recognized as either alpha chain or beta chain and cells with no beta chains were removed. Only the alpha and beta chains with the largest UMI count were kept for cells with more than one alpha and/or beta chains. After the filtering, each cell has only one beta chain contig and zero or one alpha chain contig.

The data were further examined and processed for sample to sample contamination and potential cell doublets. First, we removed cells with cell barcodes found in more than 2 samples. Second, cells barcodes overlapped between TCR and BCR data were extracted and checked for their cell types determined based on the scRNA-seq gene expression data. Only cell barcodes from T cells were kept. Finally, we checked the gene expression-based cell types of all cells and cells without an assigned cell type or not belonging to the T cell category were removed. The T cell category includes 13 cell types: Naive CD4 T, Tregs, Naive CD8 T, Effector T, NK CD56dim, Memory CD4 T, NK CD56bright, Dividing T & NK, Memory CD8 T, Dying T & NK, Memory CD4 & MAIT, Gamma delta T, and IFN-activated CD8 T.

#### TCR Clone Identification

Before defining clones, we re-annotated the contigs using Change-O<sub>16</sub>. A TCR clone was defined as a group of cells sharing an identical nucleic acid sequence of TCR alpha chain and beta chain in the repertoire.

#### Specificity group identification by GLIPH2

It was observed that antigen-specific pools of TCRs were enriched for similar CDR3 sequences<sup>17</sup>. To identify clone clusters of TCRs with a high probability of sharing antigen specificity (specificity groups), we applied GLIPH2<sup>18</sup> to cluster CD4 and CD8 TCR clones from all samples. Clones from the same cluster are predicted to bind the same antigen. Significant clonal groups reported by GLIPH2 were identified based on either local motif-based similarity (shared CDR3 amino acid motifs are comparatively rare in a reference population of naive T-cells) or global similarity (CDR3 differing by up to one amino acid). GLIPH2 assesses the quality of clusters by their global/local similarities, cluster size, and enrichment of common V-genes, a limited CDR3 length distribution and clonally expanded clones. The confidence of identified clusters was examined by Fisher's exact test which tests for the enrichment of unique CDR3s in each cluster compared to the reference naïve CD4 and CD8 T cell repertoire provided in GLIPH2. The V and J gene usage was calculated as the frequency of clones with the corresponding genes in a given clone cluster.

### B cell receptor V(D)J data processing

B cell receptor (BCR) V(D)J repertoire data processing and analysis were carried out using tools in the Immcantation framework ([www.immcantation.org](http://www.immcantation.org)). V(D)J genes were re-assigned from CellRanger output using IgBLAST v.1.15.0. Cells with multiple IGH V(D)J sequences were assigned to the most abundant IGH V(D)J sequence by UMI count. Following V(D)J gene annotation, non-functional sequences were removed from further analysis and functional V(D)J sequences were assigned into clonal groups using Change-O v.1.0.0. Sequences were first partitioned based on common IGHV gene annotations, IGHJ gene annotations, and junction lengths (the junction region is defined as the complementarity-determining region-3 plus the conserved flanking amino acid residues). Within these groups, sequences differing from one another by a length normalized Hamming distance of 0.15 within the junction region were defined as clones by single-linkage clustering<sup>19</sup> using the DefineClones function from Change-O v.1.0.0 package. This distance threshold was determined at equal distance between the two modes of the within-sample bimodal distance-to-nearest histogram across all patients. The distance-to-nearest distribution was calculated using distToNearest function from SHazaM v.1.0.0 in R v.3.6.3. Germline sequences were then reconstructed for each sequence with D segment and N/P regions masked (replaced with "N" nucleotides) using the CreateGermlines.py function within Change-O v.1.0.0. The IMGT/GENE-DB v3.1.26 reference database was used to assign B cell gene segments.

### Expanded B cell clonal lineages identification

We identified expanded clonal lineages based on the fractional abundance of each lineage. The fractional abundance of a lineage is defined as the number of cells within that lineage divided by the total number of cells observed in the repertoire at a given time point. Expanded lineages were identified among lineages with fractional abundance above 1% of the repertoire at either time point. To account for the low sequencing depth, we further required expanded clones to contain at least 5 cells.

### Analysis of somatic hypermutation (SHM) from single-cell V(D)J library

Mutations in IGHV and IGHJ relative to germline sequences were quantified using SHazaM v.1.0.0 in R v.3.6.3.

### CDR3 amino acids information content

For a given patient we computed the frequency of observed amino acids in the CDR-H3 segment for each time point A and B. Then, the fold changes were calculated as

the log<sub>2</sub> ratio of each amino acid frequency at B divided by the corresponding amino acid frequency at A. Finally, each full change was multiplied by the frequency of amino acid at B to calculate the conditional information content of the given amino acid.

#### Convergent antibody identification

To identify putative SARS-CoV-2-specific antibody signatures, we first grouped together heavy chain sequences that utilized the same IGHV and IGJ gene, and had CDR-H3 regions with the same length. We then grouped these sequences using single-linkage clustering with a threshold of 85% amino acid identity in the CDR-H3 sequence. Within these clusters, we identified sequences that were found in at least two COVID-19 patients.

#### Identification of unmutated IGHG clones

As specified in recent study<sup>20</sup>, B cell clones consisting of any cellular subtype (naïve, memory, plasma) were separated by isotype. These isotype-specific clonal clusters were considered “unmutated” if the median SHM frequency of their constituent sequences was < 1%.

#### Lineage tree analysis

B cell lineage trees were built for all clones found at both time points using the procedure detailed in recent study<sup>21</sup> using IgPhyML v1.1.3 and Change-O v1.0.0<sup>16</sup>. Within each time-point, identical sequences and those differing only by ambiguous characters (e.g. “N”) were collapsed. Only clones containing at least three distinct sequences (i.e. sequences that were either unique or sampled at different time points) were included. We estimated maximum likelihood tree topologies and branch lengths for each clone, as well as repertoire-wide model parameters shared among all clones, using the GY94 model<sup>22</sup>. Using these tree topologies, we then estimated maximum likelihood branch lengths for each clone and repertoire-wide substitution model parameters using the HLP19 model with separate  $\omega$  parameters for FWR and CDR partitions and separate  $h$  parameters for all six canonical somatic hypermutation (SHM) hot- and cold-spot motifs<sup>21</sup>. Branches with lengths < 0.001 were collapsed to zero. Trees were visualized using ggtree v2.0.2<sup>23</sup>. We used a root-to-tip correlation test<sup>24</sup> to test for evidence of continued SHM between time points within these B cell lineage trees. For each tip we calculated the divergence, which is the sum of branch lengths leading to the most recent common ancestor (MRCA) of all observed sequences. Predicted germline sequences were excluded because their sampling time is unknown. Clones in which all sequences were equally diverged from the MRCA

were discarded. We then calculated the Pearson correlation coefficient between divergence and time point ( $A = 0$ ,  $B = 1$ ). If B cell clones continued to accumulate SHM between time points, we would expect a positive correlation between divergence and time. We tested the significance of this correlation by randomizing time point labels within each tree, re-calculating the correlation between divergence and time, and repeating for 10,000 repetitions. We calculated the p-value that the correlation was positive as the proportion of repetitions in which the observed correlation was less than or equal to the correlation in randomized trees.

Yale IMPACT Research Team Authors

Abeer Obaid, Adam J. Moore, Alice Lu-Culligan, Allison Nelson, Aryn A. Malik, Anderson Brito, Angela Nunez, Anjelica Martin, Annie Watkins, Annsea Park, Arvind Venkataraman, Benjamin Goldman-Israelow, Bertie Geng, Camila Odio, Carolina Lucas, Chaney Kalinich, Christina Harden, Codruta Todeasa, Cole Jensen, Daniel Kim, David McDonald, Denise Shepard, Edward Courchaine, Elizabeth B. White, Erin Silva, Eric Song, Eric Y Wang, Eriko Kudo, Erin Silva, Feimei Liu, Harold Rahming, Hong-Jai Park, Irene Matos, Isabel M. Ott, Jessica Nouws, Ji Eun Oh, Jon Klein, Jordan Valdez, Joseph Fauver, Joseph Lim, Julio Silva, Kadi-Ann Rose, Kelly Anastasio, Kristina Brower, Laura Glick, Lina Vadlamani, Lorenzo Sewanan, Lynda Knaggs, M. Catherine Muenker, Marcella Nunez-Smith, Maria Batsu, Maria Tokuyama, Mary Petrone, Maura Nakahata, Maxine Kuang, Melissa Campbell, Melissa Linehan, Michael H. Askenase, Mikhail Smolgovsky, Molly L. Bucklin, Nicole Sonnert, Nida Naushad, Pavithra Vijayakumar, Peiwen Lu, Rebecca Earnest, Rupak Datta, Ryan Handoko, Saad Omer, Sarah Lapidus, Sarah Prophet, Sean Bickerton, Sofia Velazquez, Staci Cahill, Tara Alpert, Tianyang Mao, Tyler Rice, William Khoury-Hanold, Xiaohua Peng, Yexin Yang, Yiyun Cao, Yvette Strong.

## References

1. Wylie, A.L. *et al.* Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs. *medRxiv* (2020).
2. Vogels, C.B.F. *et al.* Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR primer-probe sets. *medRxiv* (2020).
3. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
4. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
5. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).
6. Chen, E.Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
7. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97 (2016).
8. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
9. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

10. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res* **47**, D590-D595 (2019).
11. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* **28**, 1947-1951 (2019).
12. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
13. El-Sherbiny, Y.M. *et al.* A novel two-score system for interferon status segregates autoimmune diseases and correlates with clinical features. *Sci Rep* **8**, 5793 (2018).
14. Saito, Y. *et al.* AT-rich-interactive domain-containing protein 5A functions as a negative regulator of retinoic acid receptor-related orphan nuclear receptor gamma-induced Th17 cell differentiation. *Arthritis Rheumatol* **66**, 1185-1194 (2014).
15. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
16. Gupta, N.T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356-3358 (2015).
17. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94-98 (2017).
18. Huang, H., Wang, C., Rubelt, F., Scriba, T.J. & Davis, M.M. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol* (2020).

19. Gupta, N.T. *et al.* Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *J Immunol* **198**, 2489-2499 (2017).
20. Nielsen, S.C.A. *et al.* Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. *bioRxiv*, 2020.2007.2008.194456 (2020).
21. Hoehn, K.B. *et al.* Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proc Natl Acad Sci U S A* **116**, 22664-22672 (2019).
22. Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936 (1998).
23. Yu, G., Lam, T.T., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* **35**, 3041-3043 (2018).
24. Murray, G.G. *et al.* The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol Evol* **7**, 80-89 (2016).