Science Immunology
AAAS

immunology.sciencemag.org/cgi/content/full/5/51/eabb8786/DC1

Supplementary Materials for

Transcriptomic and clonal characterization of T cells in the human central nervous system

Jenna L. Pappalardo, Le Zhang, Maggie K. Pecsok, Kelly Perlman, Chrysoula Zografou, Khadir Raddassi, Ahmad Abulaban, Smita Krishnaswamy, Jack Antel, David van Dijk*, David A. Hafler*

*Corresponding author. Email: david.vandijk@yale.edu (D.v.D.); david.hafler@yale.edu (D.A.H.)

Published 18 September 2020, *Sci. Immunol.* **5**, eabb8786 (2020)
DOI: 10.1126/sciimmunol.abb8786

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at immunology.sciencemag.org/cgi/content/full/5/51/eabb8786/DC1)

Table S10 (Microsoft Excel format). Raw data.

Supplementary Materials

Supplementary Materials and Methods

Single-cell RNA sequencing pre-processing and cell type identification

Following preliminary analysis of each sample individually, cell by gene matrices from the blood and CSF of healthy individuals (n=6) and MS patients (n=5) were concatenated into a single file. Cells were retained that had library sizes between 800 and 15,000 unique molecular identifiers (UMI) and expressed at least 300 genes. Genes were retained if they were expressed in at least 50 cells. Cells were then library size normalized using default parameters in scprep (v0.1.1, https://github.com/KrishnaswamyLab/scprep) and the top 5% of cells were removed based on mitochondrial content. Normalized data were log transformed ($\log_2 + 1$). Genes were filtered to include only those expressed by all individuals, either all healthy or all MS patients, within the blood and CSF separately. Mitochondrial and ribosomal genes were removed from the data(*80*). Blood and CSF were batch-corrected separately with all default parameters in a python implementation of ComBat (combat.py, https://github.com/brentp/combat.py) and the donor identifier as the batch. This approach performed best for batch correcting between individuals while retaining biological differences between tissues that were seen when examining the data of each individual separately. Batch corrected data were used for downstream analysis relying on principal component analysis (PCA) (visualization, clustering, PHATE(*26*), Markov affinity-based graph imputation of cells (MAGIC)(*81*), MELD(*49*), Euclidean distances between PCs). Importantly, uncorrected data was used for all differential expression analysis. Data were initially clustered at high granularity using PhenoGraph(*82*) on 50 PCs with k=30 and visualized via tSNE to isolate small doublet populations, which were identified as groups of cells expressing lineage genes from different cell subsets. After doublets were removed, data was re-clustered using Phenograph on 50 PCs with k=200 to identify cell populations. Known lineage genes were used to identify cell types.

Differential expression analysis

Differential expression analysis was performed on $\log_2$(normalized + 1) data using model-based analysis of single-cell transcriptomics (MAST)(*83*) (R, v1.8.2) with the number of genes expressed by each cell used as a covariate. Differentially expressed genes were then filtered to contain those expressed by at least 1% of cells in the cell group of interest, specified in supplementary tables.

Single-cell T cell and TCR analysis of blood and CSF

T cell clusters were extracted from the data based on high expression of T cell genes (*CD3E, CD3G, TRAC*) and low expression of lineage markers from other cell types. T cells were then filtered by genes defining other lineages (*CD14, CD19, FCGR3A, C1QA, CSF1R, MNDA, IGHG1, IGHA1, TRDC*) to minimize contamination of other cell types. Genes were filtered by retaining ones that were expressed in at least 50 cells and genes were retained that were expressed by all individuals, either all healthy or all MS patients, either within the blood and CSF. After initial clustering of T cells (Phenograph, 50 PCs, k=200), an additional cluster containing doublets and/or contaminating non-T cells was identified and this cluster as well as a cluster of Tregs was removed, leaving conventional CD4 and CD8 T cells that were clustered (Phenograph, 20 PCs, k=100) resulting in the 8 clusters characterized in the main figures. To

define the identity of these clusters, differential expression analysis was performed on the cluster of interest compared to all other cells within healthy individuals (Supplementary Table 2) and genes were included in differential expression analysis if they were expressed by all healthy donors in the blood and/or CSF. T cell clusters in figure 1 were annotated using genes with a MAST-estimated $\log_2$ fold change of at least log2(1.15) (~0.2, 15% difference). To summarize the function of each cluster in healthy individuals, violin plots of the mean $\log_2$(normalized + 1) gene expression for each cell were made for the following gene lists: Th1 function (*TBX21, CXCR3, IFNG, CCR5, HOPX, RUNX3, STAT4, IL12RB2, IFNG-AS1*)(*84, 85*); Tissue residence (*PDCD1, ITGAE, CXCR6, LAG3, PTGER4, LGALS3, CD69, PRDM1, LGALS1*)(*24, 36, 42, 43, 86, 87*); cytotoxicity (*GNLY, GZMB, GZMM, GZMK, NKG7, GZMA, GZMH*)(*80*).

PHATE, Markov affinity-based graph imputation of cells (MAGIC) and MELD, which are tools for visualization, gene imputation, and differential experimental condition analysis (here, referred to as the tissue score), respectively, rely on the same underlying graph that was constructed in graphtools (1.1.0, https://github.com/KrishnaswamyLab/graphtools)(n_pca=100, knn=5, decay=40) on all batch-corrected data from healthy donors and MS patients together. This graph was then used to run PHATE (gamma=1, n_components=3), MAGIC (all default parameters, automatically selected t), and MELD (beta=2.75), which were then used for downstream analysis. For MELD, the raw experimental signal was the original tissue of the cell (with 0 representing that the cell came from blood and 1 representing that the cell came from CSF), which resulted in a continuous score reflecting how blood-like or CSF-like each cell is. This score was then centered by subtracting the mean of the score.

To characterize gene expression trends across the blood-CSF continuum, MAGIC-imputed data was used to group genes into shape-based clusters. Imputed data was only used for the characterization of gene trends. The top 1,500 consensus variable genes across T cells from healthy individuals were found(*80*) (https://github.com/cssmillie/ulcerative_colitis) using 20 bins and a minimum of 50 cells on un-imputed, non-corrected data and then narrowed down to include genes expressed in at least 2% of cells in the tissue where the gene was more highly expressed based on the mean of $\log_2$(normalized+1) expression and with a conditional-Density Resampled Estimate of Mutual Information (DREMI)(*88*) score of at least 0.5 suggesting that the gene has a dynamic relationship with the tissue score, leaving 1,300 genes. This gene list was only used for gene clustering analysis. A normalized 2D histogram (numpy, bins = 10) was used to determine the density of the tissue score by gene expression plots imputed with MAGIC. The array for each gene was flattened and clustered (Phenograph, k=100). Genes with higher mean expression in the blood or CSF were clustered separately with the same parameters. Genes lists for each cluster were entered into Ingenuity Pathway Analysis (IPA) (Qiagen) for canonical pathway analysis and the top 5 pathways by p value are shown. Genes representative of different shape-based clusters shown in the figure were selected based on their role in T cell function and/or being representative of associated canonical pathways.

For TCR analysis, TCRs were included if they had both a CDR3α and CDR3β sequence and cells were considered part of the same clonotype if they had the same amino acid sequence of the CDR3α and CDR3β chains. TCRs were merged with transcriptional information based on cellular barcode based on the filtered contigs output from CellRanger after all filtering and clustering, so TCRs were only retained in analysis if the cell they were associated with a T cell that passed all other filters and clonotype size was determined as the number of T cells sharing CDR3α/CDR3β amino acid sequences at this step. Clonal groups shared between tissues were identified by cells in the blood and CSF sharing their CDR3α/CDR3β amino acid sequences

within an individual. Cells were considered unexpanded if a CDR3α/CDR3β was recovered from that cell but did not match other TCRs at this step. For healthy donors, donors HD1, HD3, HD4, HD5, and HD6 had expanded cells in the CSF; no expanded groups with both CDR3α/CDR3β were found in HD2, possibly due to low cell number. All pairwise Euclidean distances (scikit-learn) were calculated between cells from the blood and CSF in shared clonal groups and for cells in clonal groups unique to the blood or CSF on the first 20 PCs of healthy T cells for all 6 healthy donors. The mean distance for each clonal group is plotted. Differential expression was performed between cells from the CSF compared to the blood in shared clonal groups using genes expressed by all individuals in blood and/or CSF cells. Differential expression analysis was performed using donors HD1, HD3, HD4, HD5, and HD6, as only one pair of cells with shared TCRs was found in HD2. The expansion score was calculated using the $\log_2$(number of clones) in expanded groups and was assigned to each cell belonging to that clonal group. Differential expression was performed between highly expanded (>2 cells/clonal group) and unexpanded CD4 and CD8 T cells separately. Cells were first determined to be CD4 or CD8 based on their cluster assignment and then CD4 T cells were filtered to exclude cells expressing *CD8A* and CD8 T cells were filtered to exclude cells expressing *CD4*. Donors were included in differential expression if they contained at least 10 highly expanded CD4 or CD8 cells after this filtering (CD4 n=3: HD3, HD5, HD6; CD8 n=4: HD3, HD4, HD5, HD6). Genes input for differential expression analysis were those that were shared among all donors in highly expanded cells. Heatmaps and line plots showing differentially expressed genes between expanded and unexpanded CD4 and CD8 cells include the donors used for differential expression and were filtered for *CD8A*/*CD4* expression respectively and also include duplicated cells (2 cells/clonal group) that were not directly used in differential expression analysis.

For each cluster, differential expression analysis was performed on all cells from the total cluster between MS patients and healthy controls as well as between expanded (2 or more cells/clonal group) from MS patients and healthy controls. Genes used in both comparisons were those that were expressed by all MS patients and/or healthy donors in each cluster. Genes discussed in the results section that were differentially expressed between MS patients and healthy donors in the total cluster all had MAST-estimated log2 fold changes over log2(1.1), or a 10% difference, in the clusters where they were mentioned. Donors were included in the analysis of expanded cells if they had at least 10 expanded cells/cluster, which included all MS patients in all three clusters and HD3, HD4, HD5, and HD6 in all three clusters. Genes included in the MS gene score were the top 20 genes for each cluster that were up-regulated (fdr < 0.05) in MS in the analysis of MS patients vs. healthy donors both in the total cluster and expanded cells and had the largest increase in MAST-estimated $\log_2$FC, or (log fold change expanded – log fold change total cluster), excluding specific TCR and HLA genes. TCR and HLA genes were excluded as they may be influenced by expression in specific clonal groups or by individual donors. Violin plots in Figure 5E show the mean $\log_2$(normalized + 1) expression for the 20 MS gene signatures for each cell in each cluster shown in Figure 5D.

Single-cell and single-nucleus RNA analysis of brain parenchyma

For snRNA-seq, cell by genes matrices for total and non-neuronal nuclei were concatenated into a single data matrix. For snRNA-seq, cells were retained that had library sizes between 500 and 15,000 UMI and for scRNA-seq, cells were retained that had library sizes between 800 and 15,000. For both datasets, cells were retained if they expressed at least 200 genes and genes were retained that were expressed in at least 5 cells. Cells were then library size

normalized (scprep) and log-transformed $\log_2$(normalized + 1). For the scRNAseq dataset, the top 1% of cells with the highest mitochondrial gene expression were removed. Mitochondrial and ribosomal genes were removed from both datasets. Datasets were separately batch-corrected using ComBat with the patient identifier input as the batch using genes that were shared between all donors. Data were then clustered (snRNA-seq: Phenograph, 50 PCs, k=300 scRNA-seq: Phenograph, 50 PCs, k=200) and identities were assigned based on known lineage genes. One cluster of doublets, identified by their expression of genes associated with multiple lineages, was removed from each dataset. Cell types that were split into several clusters were merged into one cluster identity (Supplementary Figure 9). For analysis involving T cells, the T cell cluster from each dataset was filtered to exclude cells expressing non-T cell lineage genes (snRNAseq: *FCGR3A, CD14, IGHM,* scRNAseq: *FCGR3A, CD14*). Differential expression was performed between filtered T cells and all other cells on genes expressed in T cells by all donors.

Cell-cell Interaction Analysis

Using merged clusters and filtered T cells described above from snRNA-seq data, library size normalized counts matrices were input into the CellPhoneDB(*89*) (https://github.com/Teichlab/cellphonedb, v2.0.0) python package along with a metadata file containing cluster names. Genes were included that were expressed by all donors in the T cell or non-T cell clusters. The fraction of cells in a cluster expressing a gene was set to 0.02 and statistical analysis with a p value cutoff of 0.01 was performed using 1000 iterations. A lenient cutoff was set for the fraction of cells expressing a gene to allow for the identification of potential interactions involving lowly-expressed genes and due to the fact that the proportion of cells expressing a gene in nuclei may be different than in the total cell. All other parameters were left as default. Full resulting analysis tables with all cell-cell interactions are present in Supplementary Table 9.

Supplementary Figure 1. Cell quality and lineage identification of healthy and MS PBMCs and CSF.

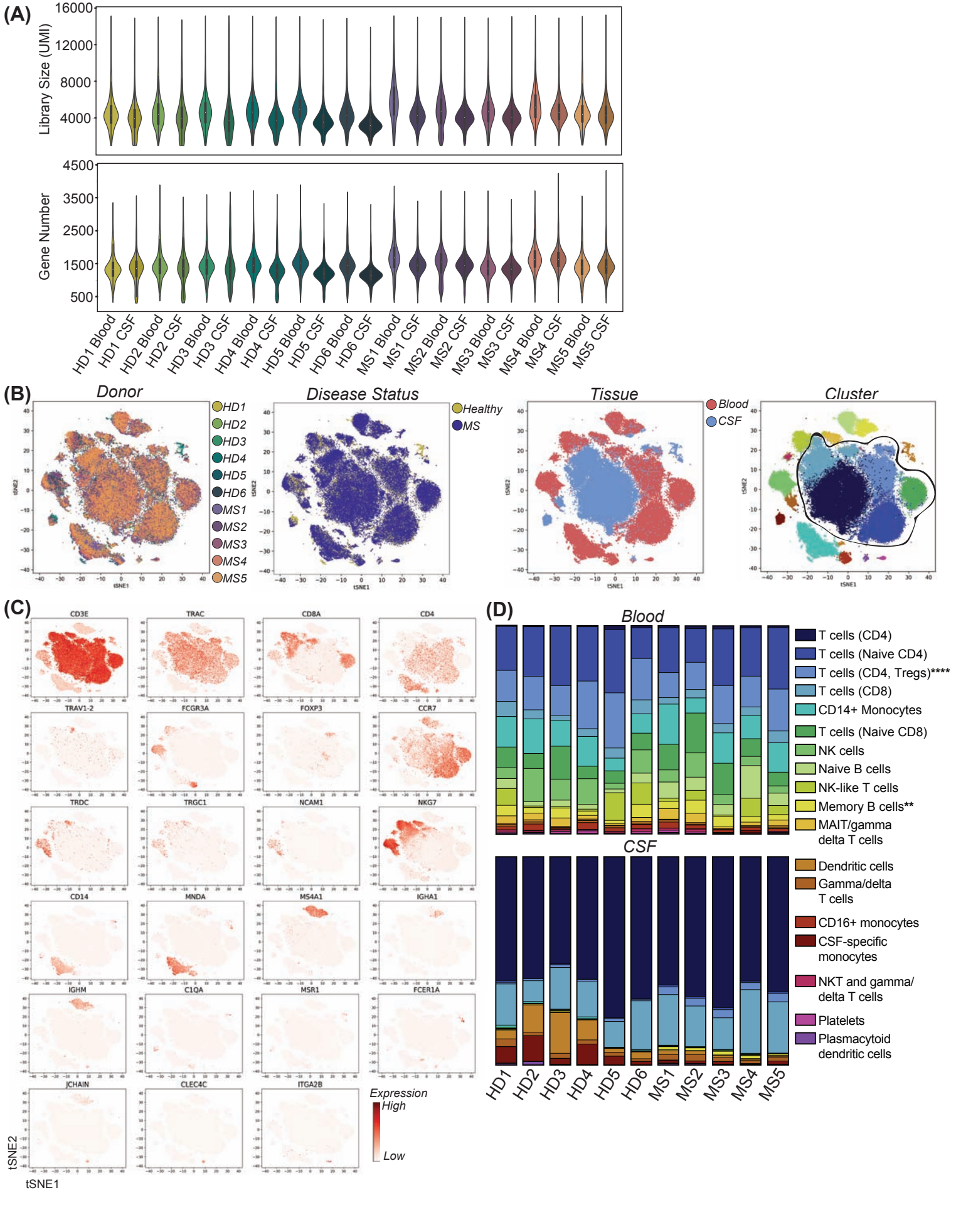
(A) Violin plots showing the library size (total number of UMIs) and number of genes detected for healthy blood and CSF using 10X Genomics following library size filtering and rare gene removal. HD = healthy donor, MS = multiple sclerosis. Violin plots include miniature internal box plots.

(B) tSNE of batch corrected total PBMCs and CSF (n=6 healthy donors, n=5 MS patients, 105,159 cells) after doublet removal. Colored by donor (left), disease status (left-center), tissue (right-center), and cluster (right). Line in right panel shows clusters that were selected and re-clustered for further T cell analysis. Color code for clusters present in **(d)**.

(C) tSNE colored by batch corrected gene expression of known lineage genes used to identify cell types.

(D) Parts of whole plots showing the distribution of clusters for each donor. Cluster composition for healthy donors and MS patients were compared using multiple two-tailed t-tests with FDR correction (desired FDR = 1%). Cluster composition in healthy blood compared to MS blood: No differences in cluster composition were found in the blood between healthy donors and MS patients. Cluster composition in healthy CSF compared to MS CSF: T cells (CD4, Treg) (MS > healthy): ****p=0.0000013, Memory B cells: **p=0.0013 (Supplementary Table 5).

Supplementary Figure 1



Supplementary Figure 2. PHATE representation of T cells.

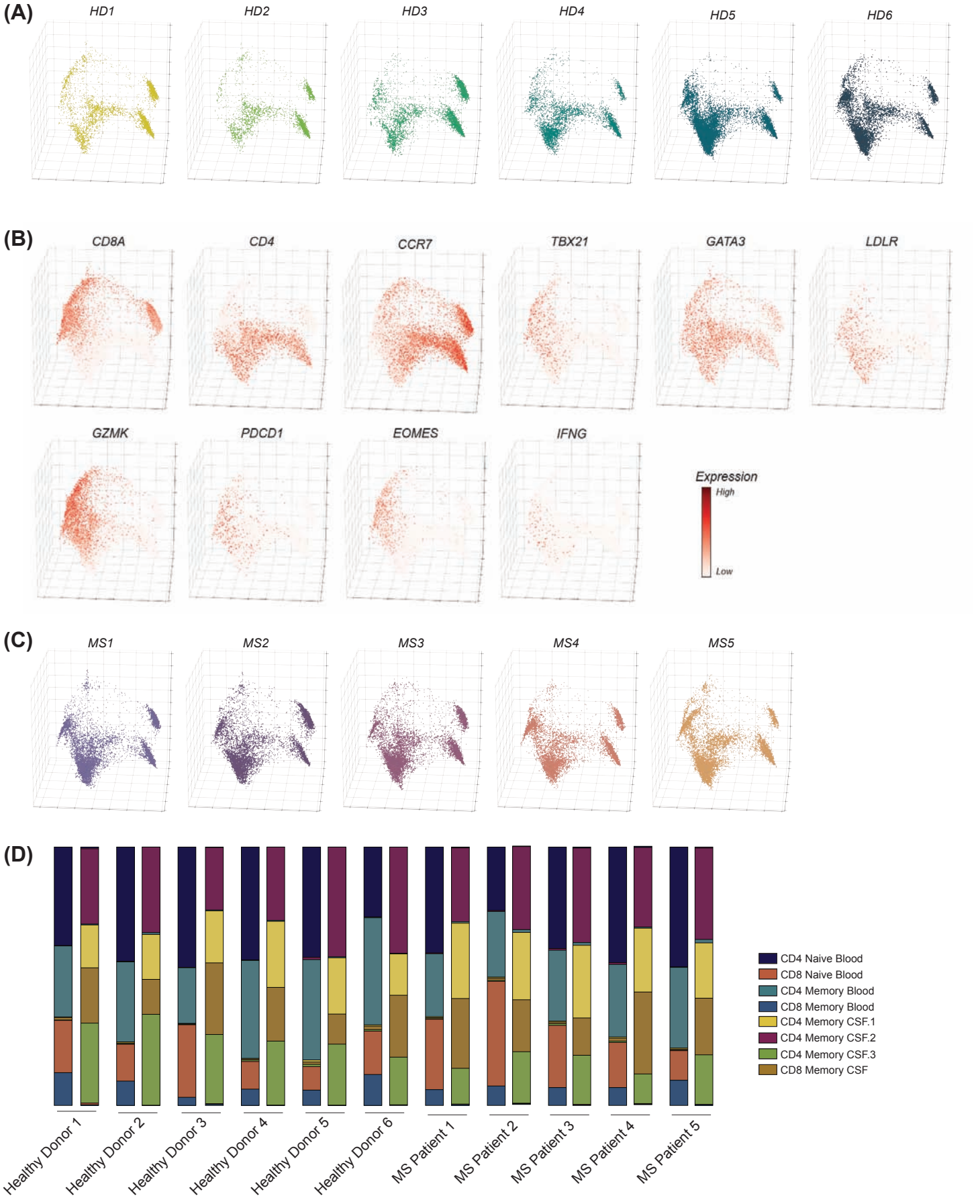
(A) PHATE embedding for each healthy donor.

(B) PHATE embedding of healthy donors colored by batch corrected gene expression of T cell lineage genes as well as select genes that were found to characterize clusters or the blood to CSF continuum.

(C) PHATE embedding for each patient with MS.

(D) Conventional T cell cluster composition for each healthy donor and patient with MS.

Supplementary Figure 2



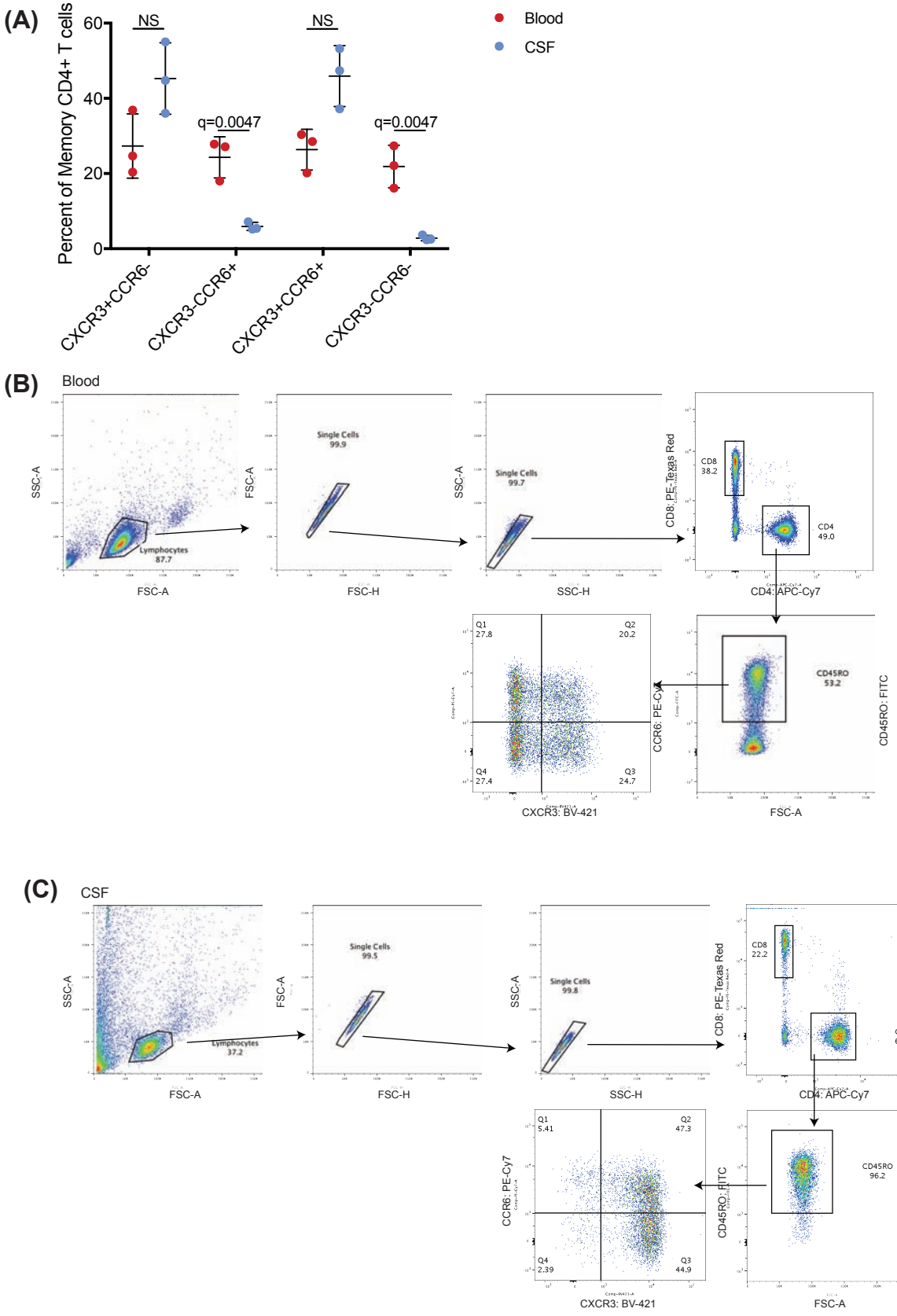
Supplementary Figure 3. Chemokine receptor flow cytometry of healthy PBMCs and CSF.

(A) Plots showing the percent of CD4+CD45RO+ cells expressing combinations of CXCR3 and CCR6 in the blood and CSF of healthy individuals (n=3 donors, Supplementary Table 1) with the mean and standard deviation. Frequencies between the blood and CSF were compared using multiple two-tailed t-tests with FDR correction (desired FDR = 1%): CXCR3-CCR6+ (blood > CSF): q = 0.0047, CXCR3-CCR6- (blood > CSF): q = 0.0047 (Supplementary Table 5).

(B) Gating strategy for the blood shown in **(A)** from donor 3.

(C) Gating strategy for the CSF shown in **(A)** from donor 3.

Supplementary Figure 3



Supplementary Figure 4. Gating strategy for total T cell selection for surface and intracellular staining.

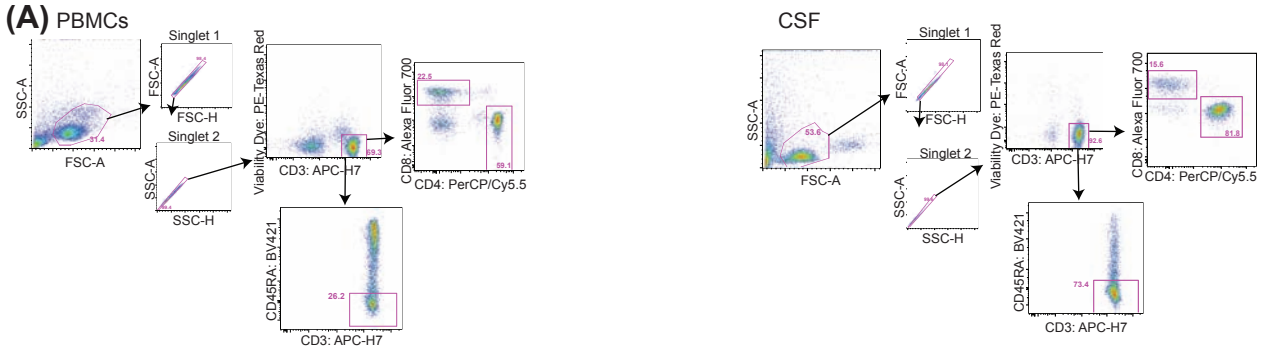
(A) Gating strategy to isolate total T cells (CD3+) from PBMCs and CSF for Figure 3 and CD4+, CD8+ or memory (CD45RA-) T cells for Supplementary Figure 4 B-E.

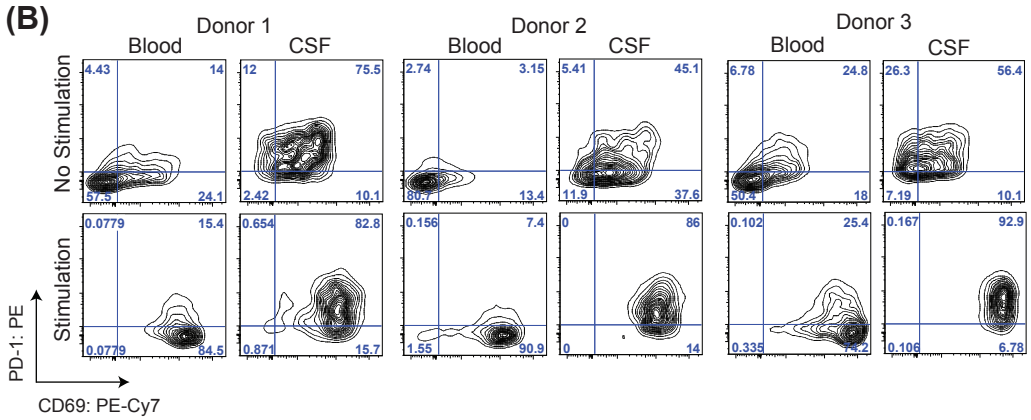
(B) Contour plots showing CD69 and PD-1 or IFNγ and PD-1 for total CD3+ T cells in the blood and CSF for each healthy donor shown in Figure 3 (n=3).

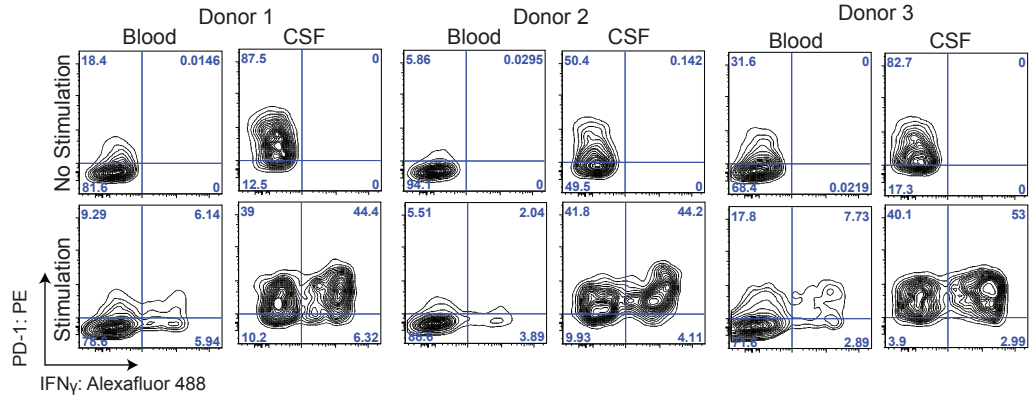
(C) Contour plots showing CD69 and PD-1 or IFNγ and PD-1 for CD3+CD4+ T cells in the blood and CSF for each healthy donor.

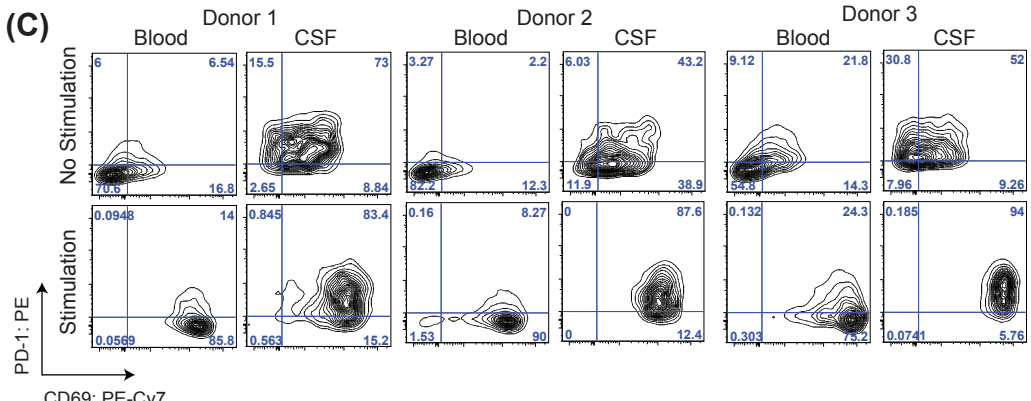
(D) Contour plots showing CD69 and PD-1 or IFNγ and PD-1 for CD3+CD8+ T cells in the blood and CSF for each healthy donor.

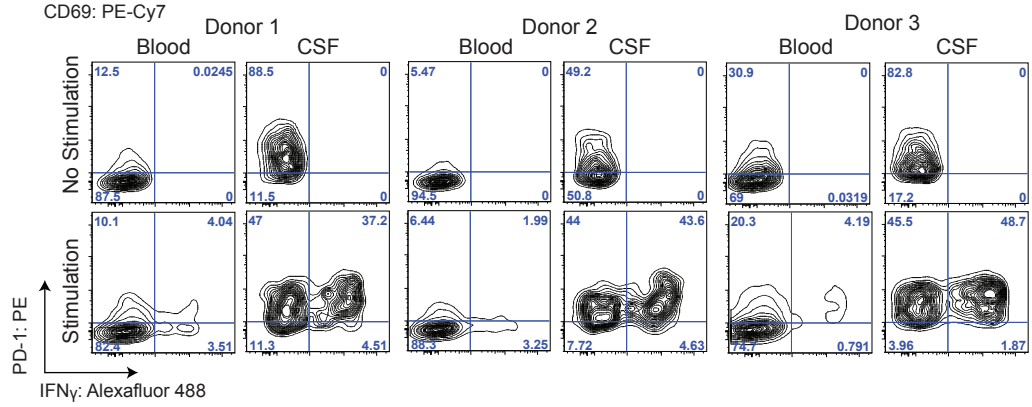
(E) Contour plots showing CD69 and PD-1 or IFNγ and PD-1 for CD3+CD45RA- T cells in the blood and CSF for each healthy donor.

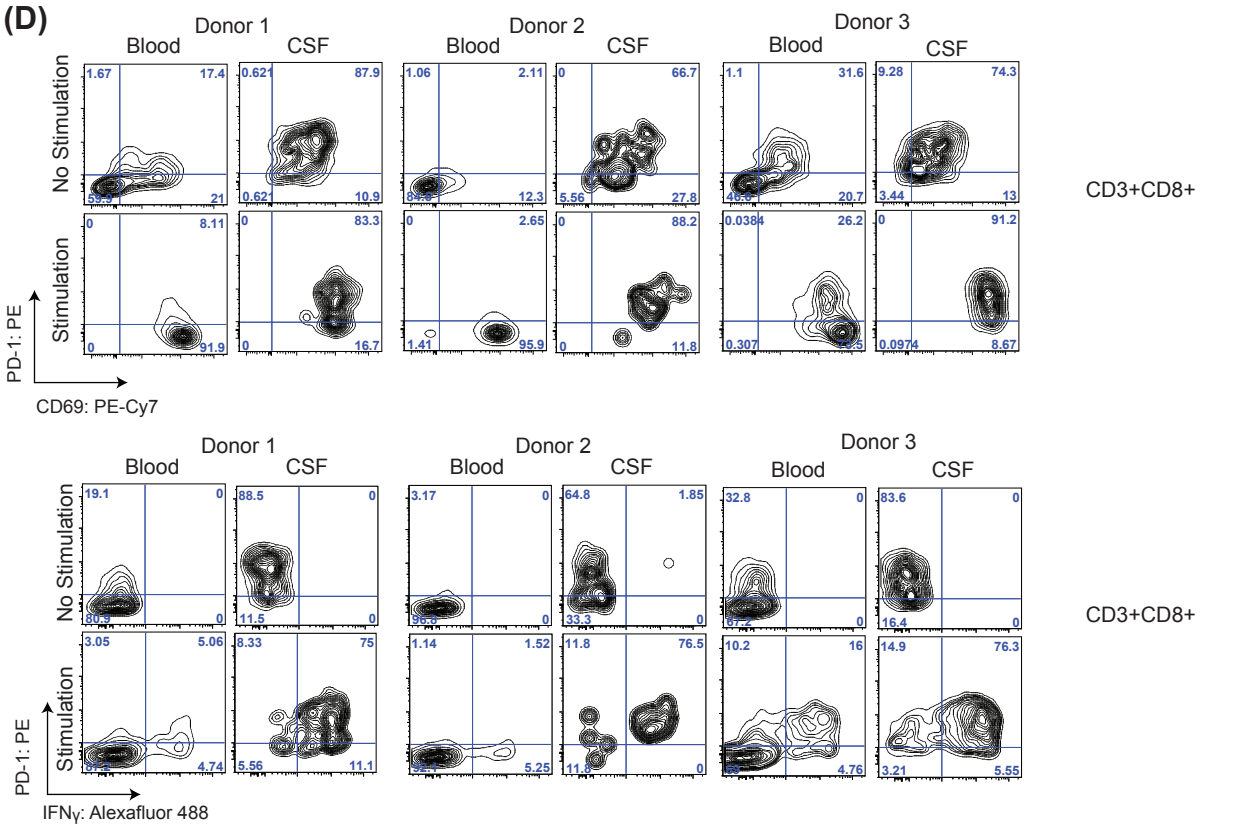

(A) PBMCs
Singlet 1
Singlet 2
SSC-A
FSC-A
FSC-H
SSC-H
Viability Dye: PE-Texas Red
CD3: APC-H7
CD8: Alexa Fluor 700
CD4: PerCP/Cy5.5
CD45RA: BV421
CSF

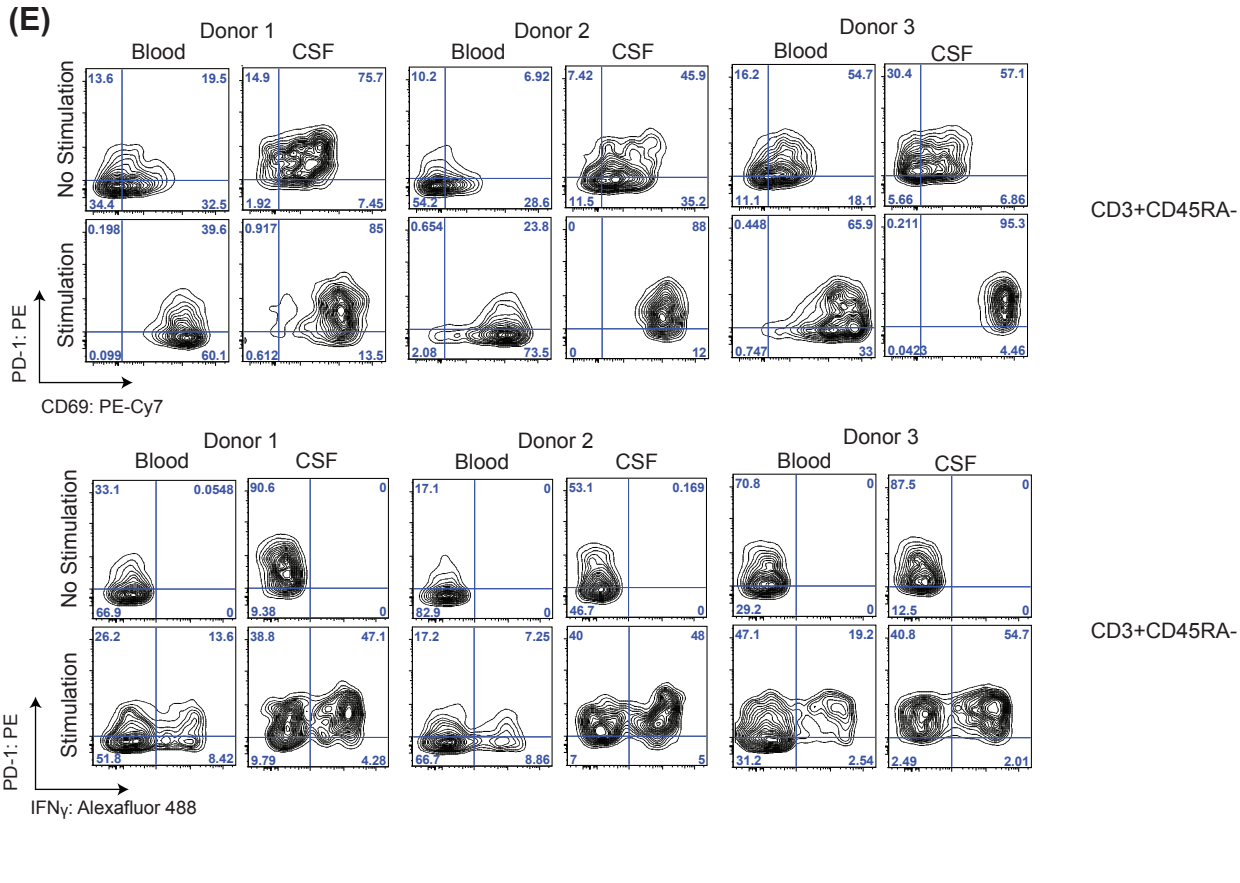

(B)
Donor 1
Donor 2
Donor 3
Blood
CSF
No Stimulation
Stimulation
PD-1: PE
CD69: PE-Cy7
Total CD3+


Donor 1
Donor 2
Donor 3
Blood
CSF
No Stimulation
Stimulation
PD-1: PE
IFNγ: Alexafluor 488
Total CD3+


(C)
Donor 1
Donor 2
Donor 3
Blood
CSF
No Stimulation
Stimulation
PD-1: PE
CD69: PE-Cy7
CD3+CD4+


Donor 1
Donor 2
Donor 3
Blood
CSF
No Stimulation
Stimulation
PD-1: PE
IFNγ: Alexafluor 488
CD3+CD4+


(D)
Donor 1
Donor 2
Donor 3
Blood
CSF
No Stimulation
Stimulation
PD-1: PE
CD69: PE-Cy7
IFNγ: Alexafluor 488
CD3+CD8+
CD3+CD8+


(E)
Donor 1
Donor 2
Donor 3
Blood
CSF
No Stimulation
Stimulation
PD-1: PE
CD69: PE-Cy7
IFNγ: Alexafluor 488
CD3+CD45RA-
CD3+CD45RA-

Supplementary Figure 5. Analysis of healthy TCRs.

(A) Bar plot showing the fraction of expanded CSF T cells that were shared with the blood or only found in the CSF for the 5 healthy donors where expanded T cells were identified in the CSF.

(B) Stacked bar plot showing all 284 clonal TCR groups that had cells in both the blood and CSF (n=6 healthy donors). Total height of the bar reflects the total number of cells in the clonal group with the fraction of cells from the blood or CSF shown.

(C) All clonal groups that contain cells in both the blood and CSF plotted on the 3D PHATE representation and colored by the original tissue of each cell.

(D) Parts of whole plots showing the cluster distribution for clonal groups that are unique to the blood or CSF and shared between the blood and CSF in healthy donors.

(E) Violin plots for select genes that were found to be differentially-expressed between CSF and blood cells that were part of the same clonal groups but present in different tissues (n=5 donors, 1,008 cells) (Supplementary Table 4).

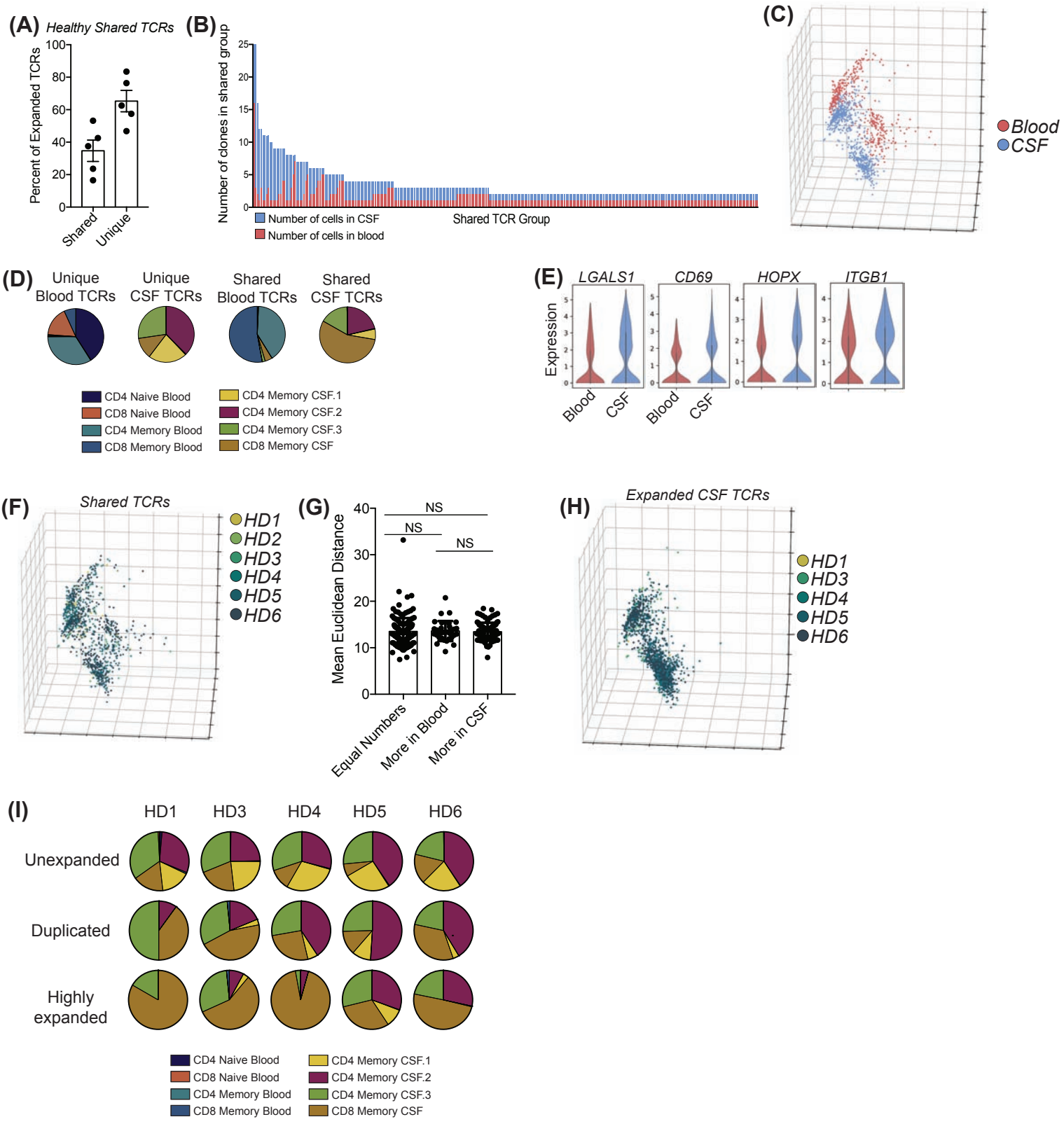
(F) All clonal groups that contain cells in the blood and CSF plotted on the 3D PHATE representation and colored by the donor identity of each cell.

(G) Bar plot showing the mean Euclidean distance of clonal groups present in both tissues based on the ratio of cells in the blood or CSF. Kruskal-Wallis test with Dunn's multiple comparisons test: NS: not significant, all p-values >0.9999.

(H) Expanded T cell groups present in the CSF colored by the donor identity.

(I) Parts of whole plots showing the cluster distribution for unexpanded, duplicated, and highly-expanded T cells by healthy donor.

Supplementary Figure 5



Supplementary Figure 6. Analysis of MS TCRs.

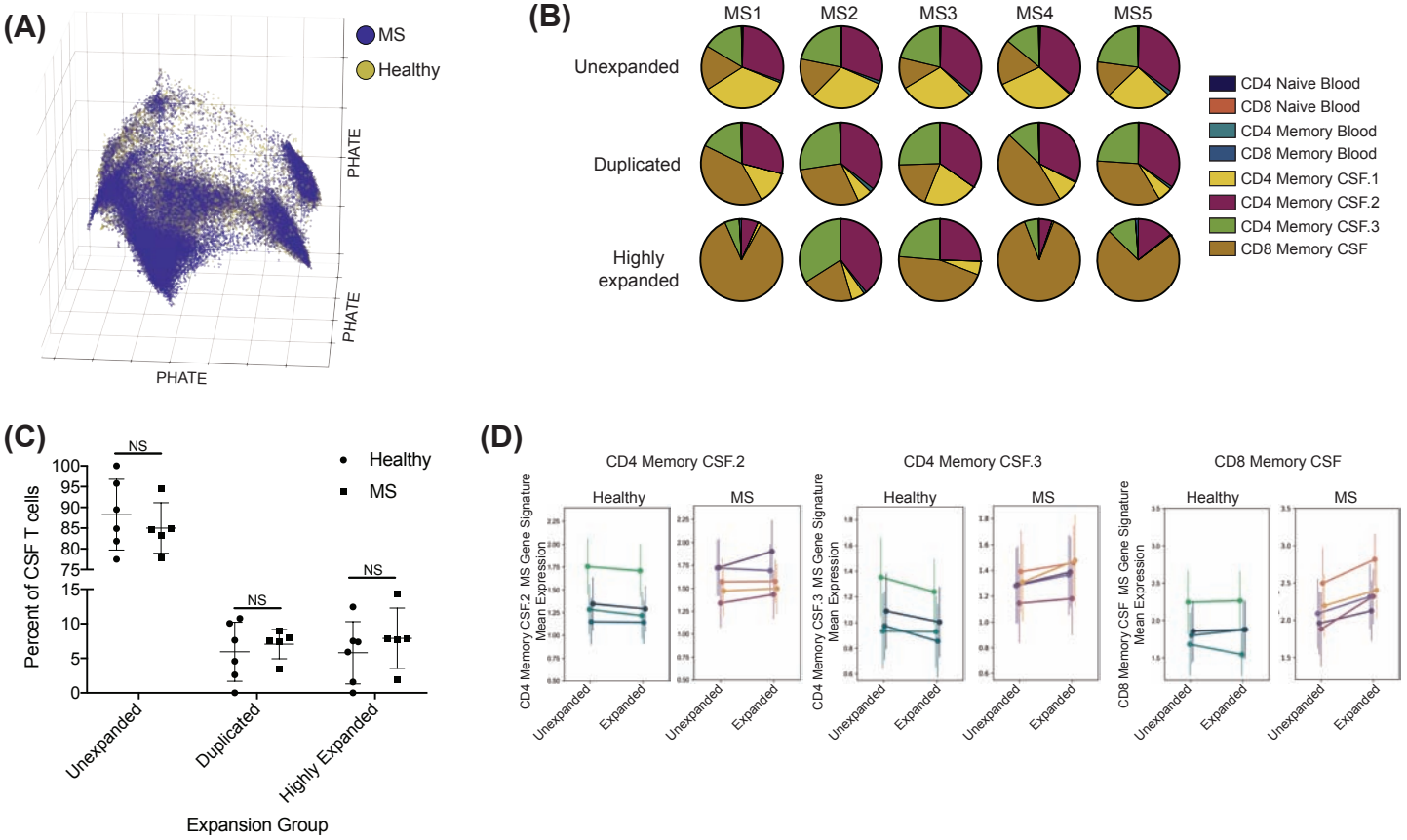
(A) PHATE representation of all conventional T cells colored by disease.

(B) Parts of whole plots showing the cluster distribution for unexpanded, duplicated, and highly-expanded T cells by MS donor.

(C) Frequency of unexpanded, duplicated, and highly-expanded T cells in healthy controls (n=6) and patients with MS (n=5) shown with mean and standard deviation. Clonal group composition for healthy donors and MS patients were compared using multiple two-tailed t-tests with FDR correction (desired FDR = 1%). NS: not significant (MS vs. Healthy unexpanded: $p = 0.4998$ ($q = 0.6166$), duplicated: $p = 0.6105$ ($q = 0.6166$), highly expanded: $p = 0.4547$ ($q = 0.6166$).

(D) Plots of the mean and standard deviation MS gene scores by individual in unexpanded and expanded cells for the 3 clusters shown in Figure 5E (n=4 healthy donors, n=5 MS patients).

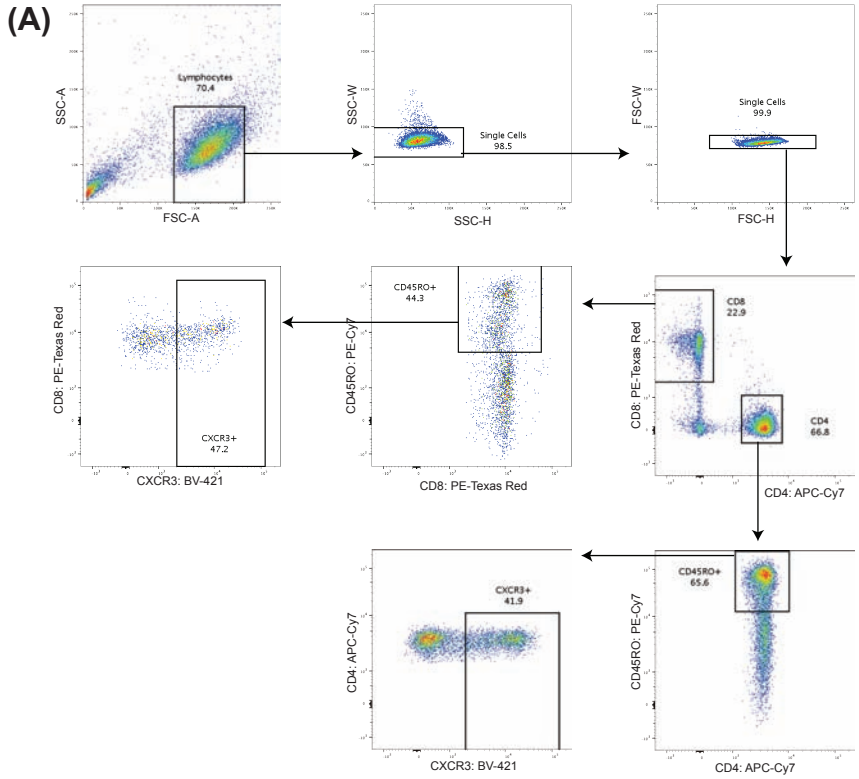
Supplementary Figure 6



Supplementary Figure 7. Sample sort strategy for *in vitro* experiments.

(A) Representative sort strategy to isolate CD4+CD45RO+CXCR3+ and CD8+CD45RO+CXCR3+ T cells from healthy blood after PBMC isolation and T cell enrichment for *in vitro* culture for Figure 6A.

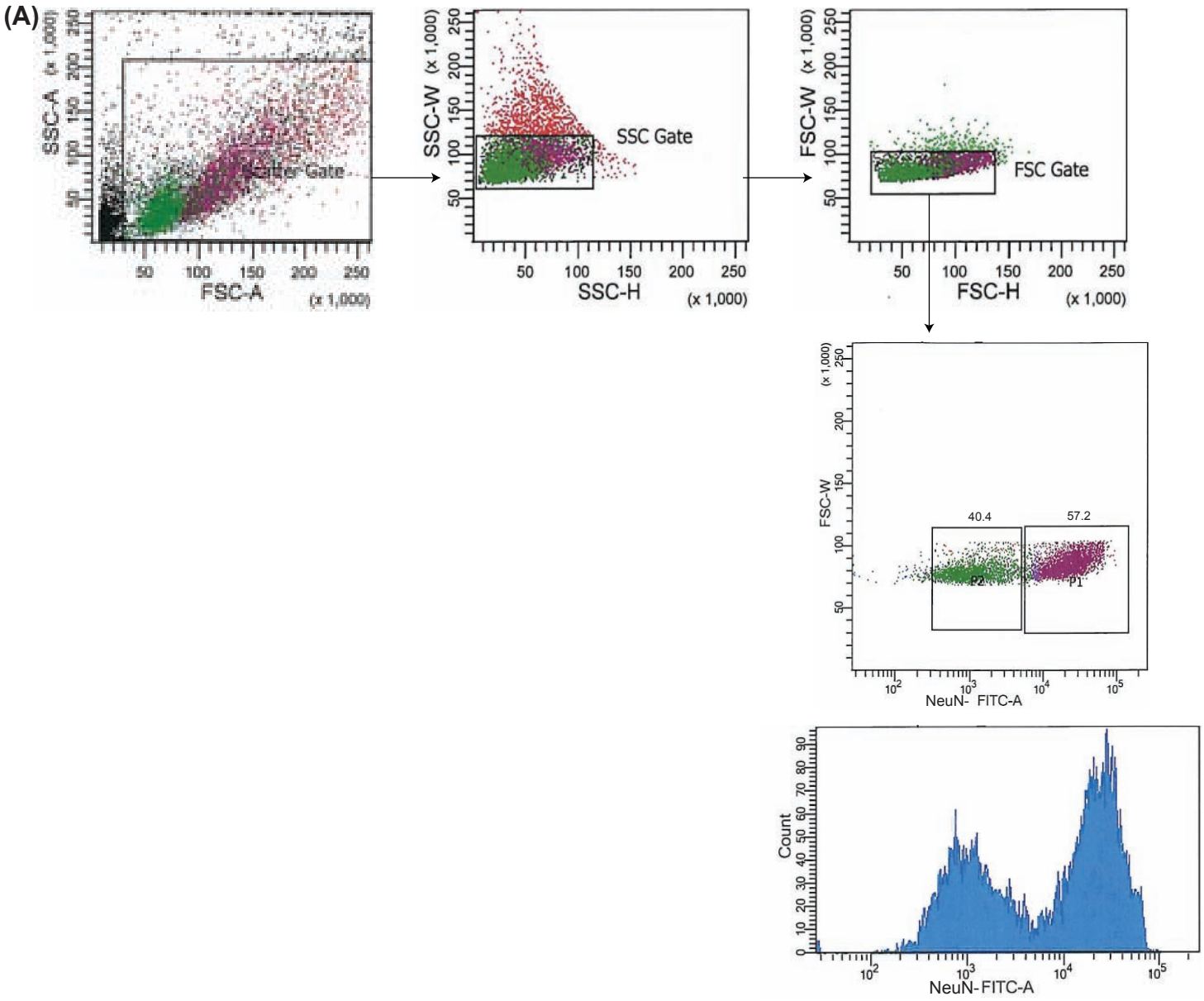
Supplementary Figure 7



Supplementary Figure 8. Sample sort strategy for non-neuronal nuclei enrichment.

(A) Example sort strategy for nuclei stained with NeuN to exclude neuronal nuclei prior to single-cell RNA sequencing including a histogram of NeuN staining.

Supplementary Figure 8



Supplementary Figure 9. Cell quality and lineage identification in the brain parenchyma in snRNA-seq and scRNA-seq datasets.

(A) Violin plots showing the library size (total number of UMIs) and number of genes detected for single-nucleus RNA sequencing following library size filtering and rare gene removal in total nuclei or non-neuronal (NeuN-negative) nuclei (n=3 donors).

(B) Violin plots showing the mean $\log_2$(normalized) expression of mitochondrial (MT-) and ribosomal (RPL, RPS) genes in all nuclei (total and non-neuronal combined) for the three donors shown in **(A)** present in the data before mitochondrial and ribosomal genes were removed.

(C) tSNE of batch corrected total and non-neuronal nuclei for 3 donors. tSNE plots are colored by donor and type of processing (left), Phenograph cluster (middle), and clusters merged based on overall cell type (right).

(D) tSNE colored by gene expression of known lineage genes used to identify cell types.

(E) Violin plots showing the library size (total number of UMIs) and number of genes detected for single-cell RNA sequencing (n=4 donors).

(F) tSNE of batch corrected single-cell RNA sequencing for 4 donors. tSNE plots are colored by donor (left), Phenograph cluster (middle), and clusters merged based on overall cell type (right).

(G) tSNE colored by gene expression of known lineage genes used to identify cell types.

(H) Parts of whole plots showing the distribution of cells for each donor among merged clusters for both the single-nucleus RNA sequencing (left) and single-cell RNA sequencing (right) datasets.

Supplementary Figure 9