Figure S1:

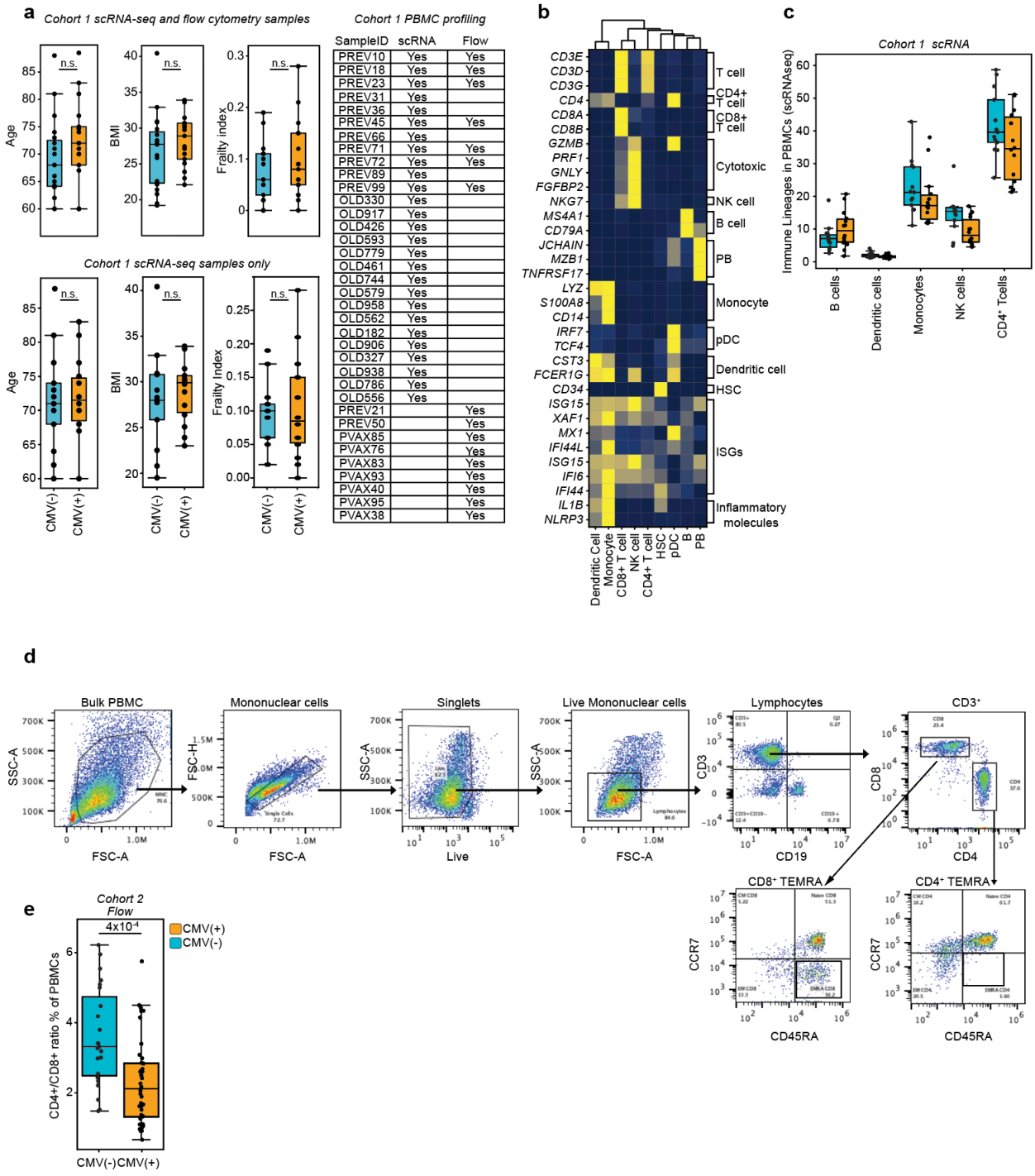

Figure S1: PBMC-level annotations and analysis. a, *Top Left:* Distribution of age, BMI, and frailty index in total Cohort 1 (scRNA-seq and flow cytometry profiled samples; n=36) as well as scRNA-seq profiled samples only (n=27). n.s. indicates not significant. *Bottom Left:* Donor-level distribution of samples profiled with scRNA-seq and/or flow cytometry in Cohort 1. *Right:* Donor-level distribution of modalities used to profile each sample via scRNA-seq and/or flow cytometry. **b,** Scaled expression of select marker genes. **c,** PBMC subset frequency ratios quantified *via* scRNA-seq in CMV(-) (n=13) and CMV(+) (n=14) in Cohort 1. **d**, Gating strategy for $CD8^+$ and

$CD4^+$ T cells (and TEMRA cells) in Cohort 2 using flow cytometry; FSC, forward scatter; SSC, side scatter. **e**, $CD4^+$ to $CD8^+$ T cell frequency ratios quantified by flow cytometry in Cohort 2: CMV(-) (n=23) and CMV(+) (n=40).

Box plots display the median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an individual sample. For a and c, Mann-Whitney-U test (two-sided) was used to calculate *P*-values between CMV(-) and CMV(+). **c**, none of the measurements were significant after adjusting the *P*-values for multiple hypothesis testing, using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

Figure S2:

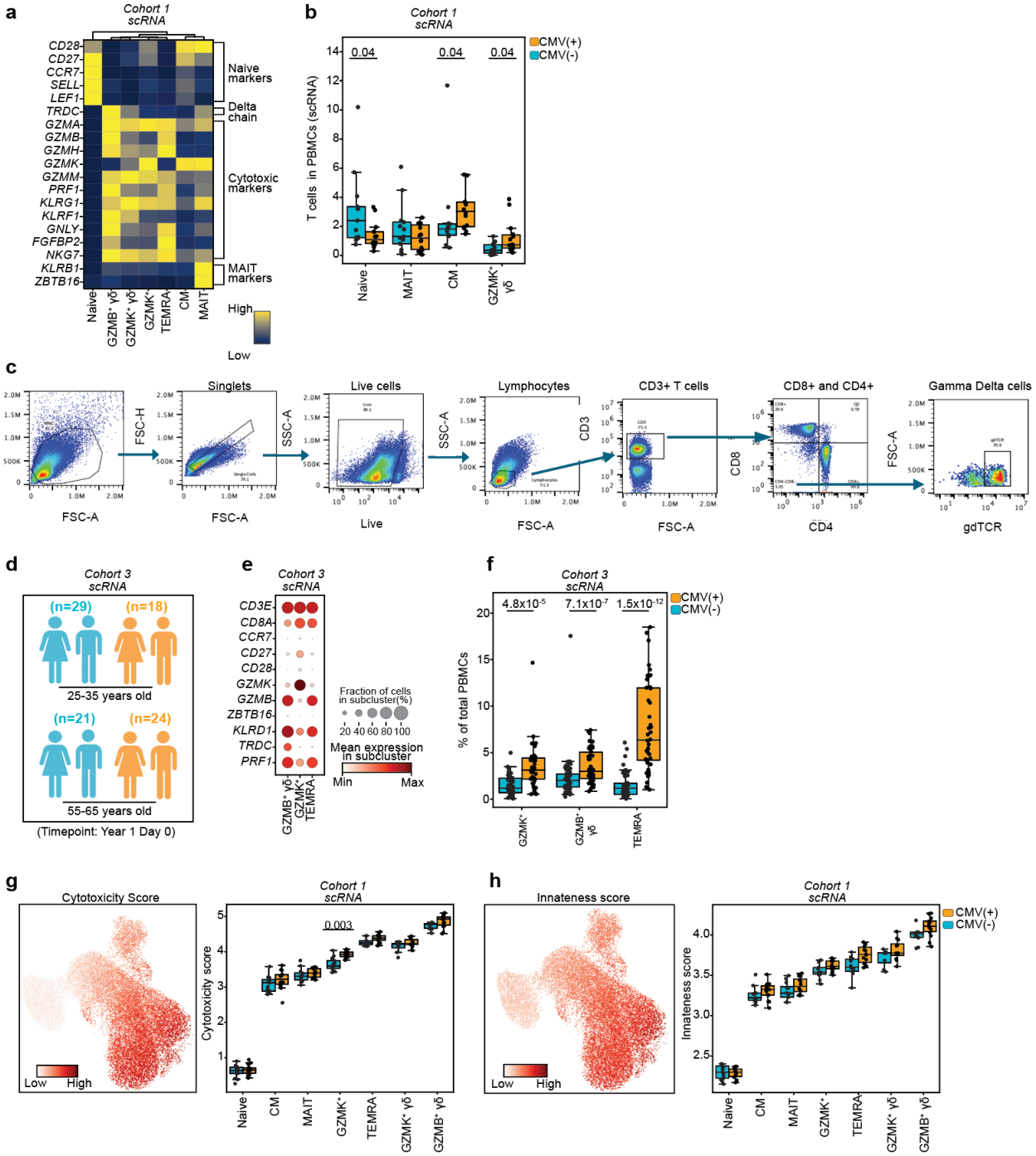

Figure S2: Cytotoxicity and innateness scores for CD8$^+$ T cell subsets. a, Scaled expression of select marker genes in CD8$^+$ T cell subsets. **b,** Select CD8$^+$ T cell frequencies within PBMCs quantified *via* scRNA-seq in Cohort 1: CMV(-) (n=13) and CMV(+) (n=14). **c**, Gating strategy for CD8$^+$/CD4$^+$ TEMRA and γδ T cells in Cohort 2 using flow cytometry; FSC, forward scatter; SSC, side scatter. **d,** Schematic representation of Cohort 3 including two age groups (25-35 years old; 55-65 years old). **e,** Scaled expression of select marker for GZMB$^+$ γδ T, GZMK$^+$ T, and CD8$^+$

TEMRA T cells in Cohort 3. **f,** Select $CD8^+$ T cell frequencies within PBMCs quantified *via* scRNA-seq in Cohort 3. **g,** *Left:* UMAP showing Cytotoxicity score in $CD8^+$ T cell subsets. *Right:* individual-level cytotoxicity score in $CD8^+$ T cells subsets. **h,** *Left:* UMAP showing Innateness score in $CD8^+$ T cell subsets. *Right:* individual-level Innateness score in $CD8^+$ T cells subsets. Box plots display the median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an individual sample. Mann-Whitney-U test (two-sided) was used to compare cell counts between CMV(-) and CMV(+) (**b, f, g, h**). All *P*-values are shown after adjusting for multiple hypothesis testing using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

Figure S3:

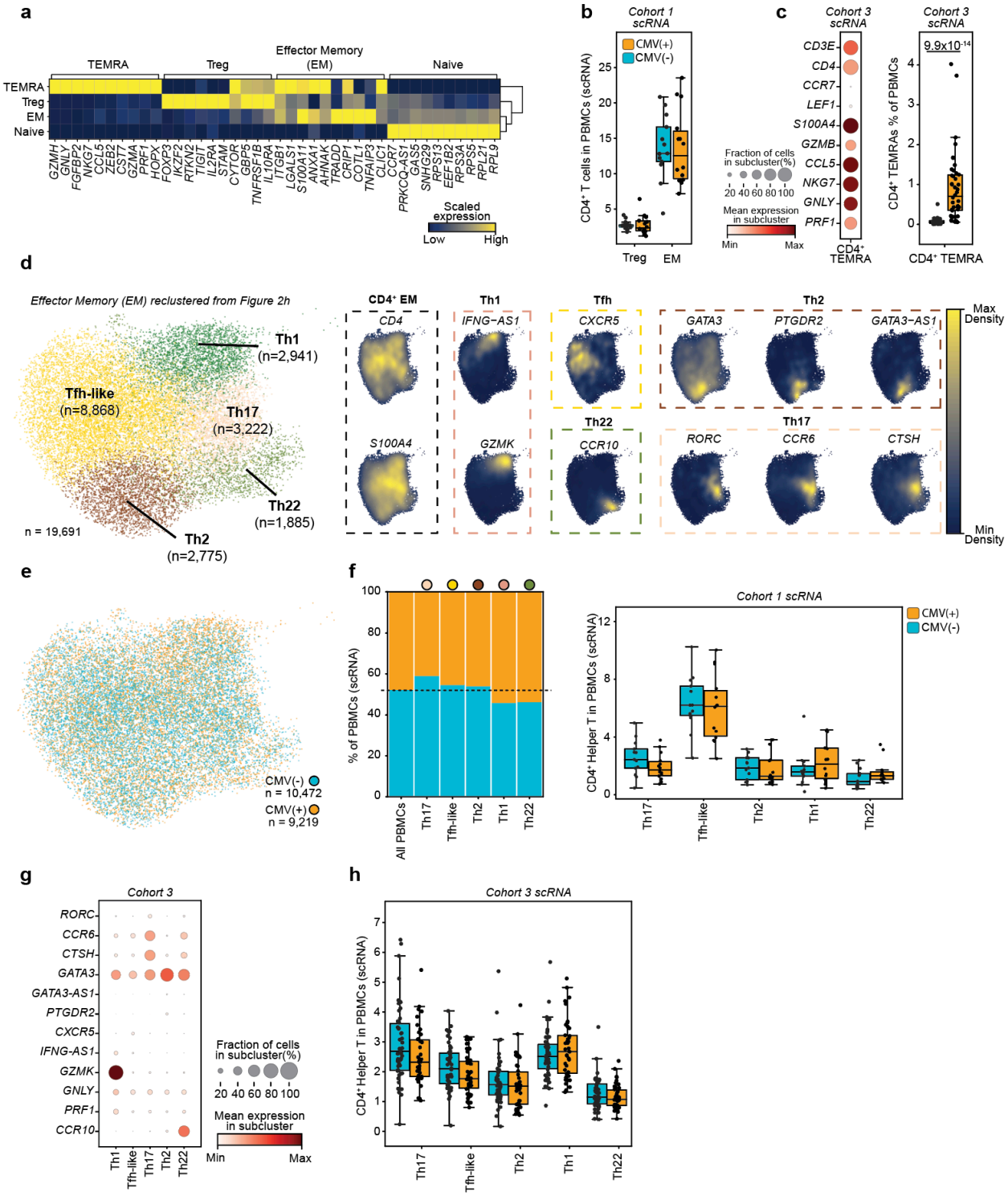

Figure S3: CD4$^+$ T helper subsets were comparable between CMV(-) and CMV(+). a, Scaled expression of top 10 marker genes in CD4$^+$ naïve, CD4$^+$ Treg, CD4$^+$ EM, and CD4$^+$ TEMRA subsets. **b,** CD4$^+$ Treg and CD4$^+$ EM T cell frequencies within PBMCs quantified *via* scRNA-seq in Cohort 1: CMV(-) (n = 13) and CMV(+) (n = 14). **c,** *Left:* Scaled expression of select marker genes for CD4$^+$ TEMRA T cells in Cohort 3. *Right:* CD4$^+$ TEMRA T cell frequencies within

PBMCs quantified *via* scRNA-seq in Cohort 3. **d,** *Left:* UMAP representing re-clustered $CD4^+$ EM compartment and 5 resulting T helper subsets. Tfh-like, T follicular helper-like; Th2, T helper 2; Th22, T helper 22; Th17, T helper 17; Th1. *Right:* Gene density plot showing expression levels of key marker genes, calculated using Nebulosa. Each gene density is grouped according to known T helper marker genes. **e,** UMAP color coded for CMV serostatus. **f,** *Left:* $CD4^+$ T helper frequencies in CMV(-) and CMV(+). *Right:* T helper cell frequencies within PBMCs quantified *via* scRNA-seq in Cohort 1: CMV(-) (n=13) and CMV(+) (n=14). **g,** Scaled expression of select marker genes for T helper subsets in Cohort 3. **h,** T helper cell frequencies within PBMCs quantified *via* scRNA-seq in Cohort 3: CMV(+) (n=42) and CMV(-) (n=50). Box plots display the median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an individual sample. Mann-Whitney-U test (two-sided) was used to compare cell counts between CMV(-) and CMV(+) (**b**, **c, f, h**). None of the T helper measurements were significant after adjusting the *P*-values for multiple hypothesis testing, using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

Figure S4:

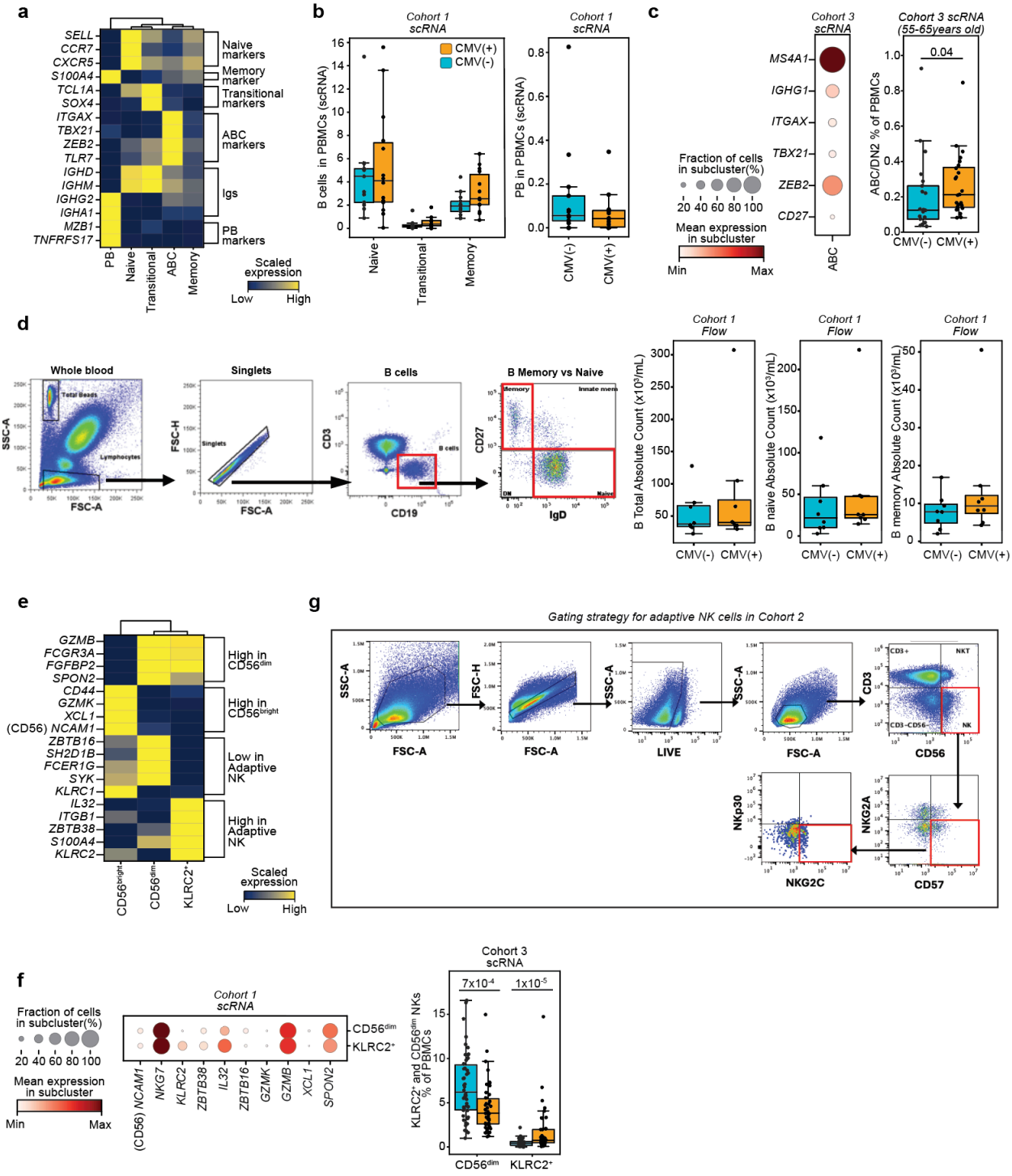

Figure S4: B and NK cell annotations. a, Scaled expression of select marker genes in B cell subsets. **b,** *Left:* naïve, transitional, and memory B cell frequencies within PBMCs. *Right:* PB cell frequencies within PBMCs. All panels represent scRNA-seq data from Cohort 1: CMV(-) (n=13) and CMV(+) (n=14). **c,** Scaled expression of select marker genes for ABCs in Cohort 3. *Right:* ABC frequencies within PBMCs quantified *via* scRNA-seq in Cohort 3 (55-65 age group): CMV(-) (n=20) and CMV(+) (n=24). **d**, *Left:* Gating strategy for memory and naïve B cells in Cohort 1

using flow cytometry; FSC, forward scatter; SSC, side scatter. *Right:* Total B, naïve B, and memory B total counts in Cohort 1. **e,** Scaled expression of select marker genes in NK cell subsets in Cohort 1. **f,** *Left:* Scaled expression of select marker genes for $KLRC2^{+}$ and $CD56^{dim}$ NK cells in Cohort 3. *Right:* NK $CD56^{dim}$ and $KLRC2^{+}$ NK cell frequencies quantified *via* scRNA-seq in Cohort 3: CMV(-) (n=50) and CMV(+) (n=42). **g**, Gating strategy for adaptive NK cells in Cohort 2 using flow cytometry; FSC, forward scatter; SSC, side scatter. Box plots display the median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an individual sample. Mann-Whitney-U test (two-sided) was used to compare cell counts between CMV(-) and CMV(+) (**b, c, d, f**).

Figure S5:

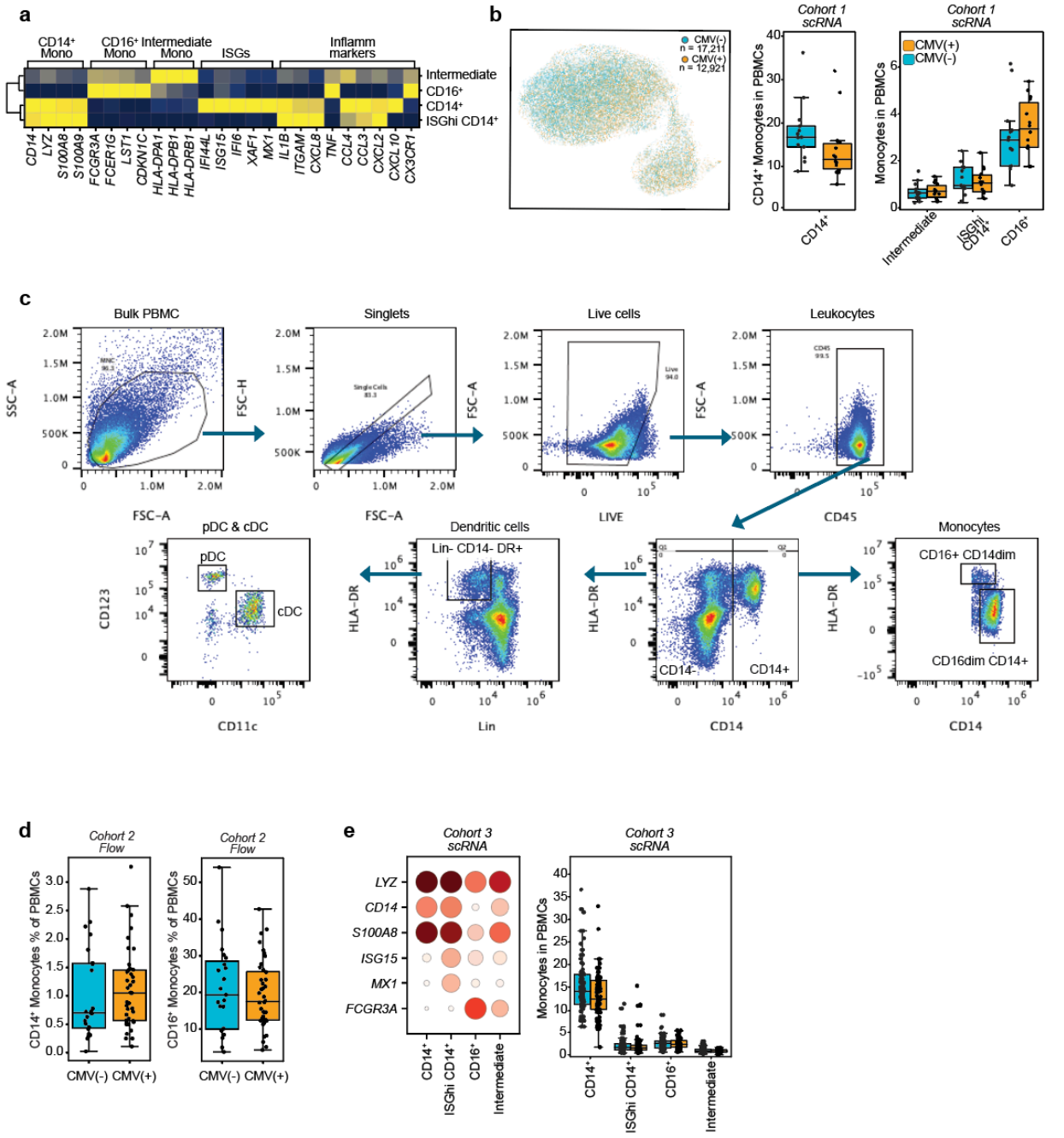

Figure S5: Monocyte subsets are not altered with CMV. a, Scaled expression of select marker genes in monocyte subsets. **b,** *Left:* UMAP color coded for CMV serostatus. *Middle:* CD14^{+} monocyte frequencies within PBMCs. *Right:* Intermediate, ISGhi CD14^{+}, and CD16^{+} monocyte frequencies within PBMCs. All panels represent scRNA-seq data from Cohort 1: CMV(-) (n=13) and CMV(+) (n=14) **c,** Gating strategy of dendritic cells and monocytes in Cohort 2. **d**, Donor-level quantification of CD14+ and CD16+ monocytes in Cohort 2. **e**, *Left:* scaled expression of select marker genes for CD14^{+}, ISGhi CD14^{+}, and CD16^{+}, and intermediate monocytes. *Right:*

monocyte subset frequencies within PBMCs. All panels represent scRNA-seq data from Cohort 3: CMV(-) (n=50) and CMV(+) (n=42).

Figure S6:

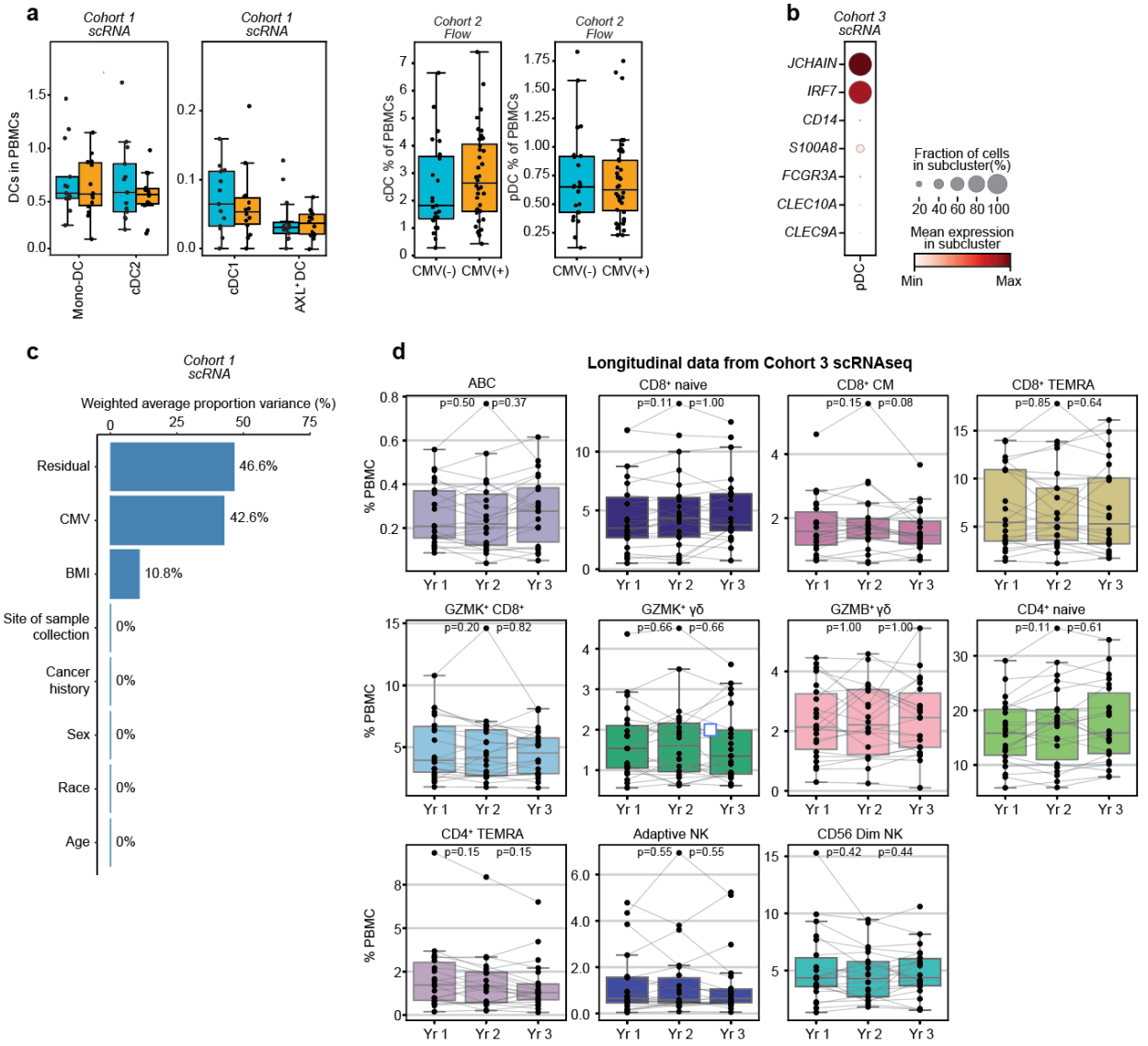

Figure S6: Dendritic cell subsets and temporal stability of PBMC frequencies with CMV. a, *Left:* Mono-DC, cDC2, cDC1 and AXL$^+$ DC frequencies within PBMCs in Cohort 1. *Right:* cDC and pDC frequencies within PBMCs in Cohort 2. **b,** Scaled expression of select marker genes for pDCs in Cohort 3. **c**, Principal Variance Component Analysis (PVCA) of lymphocyte subset frequencies, estimating the proportion of variance explained by each clinical factor in Cohort 1. **d,** Individual-level frequencies of CMV-associated cell types over a three-year period in CMV(+) individuals from Cohort 3 (n=22). Connecting lines represent individual-specific changes in cell frequencies between Year 1, Year 2, and Year 3 visits. *P*-values in each box plot are shown for (i) Year 1-Year 2 and (ii) Year 2-Year 3 comparisons. Box plots display the median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an individual sample. Mann-Whitney-U test (two-sided) was used to compare cell counts (**a**, **d**). None of the measurements were significant after adjusting the *P*-values for multiple hypothesis testing, using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

Figure S7:

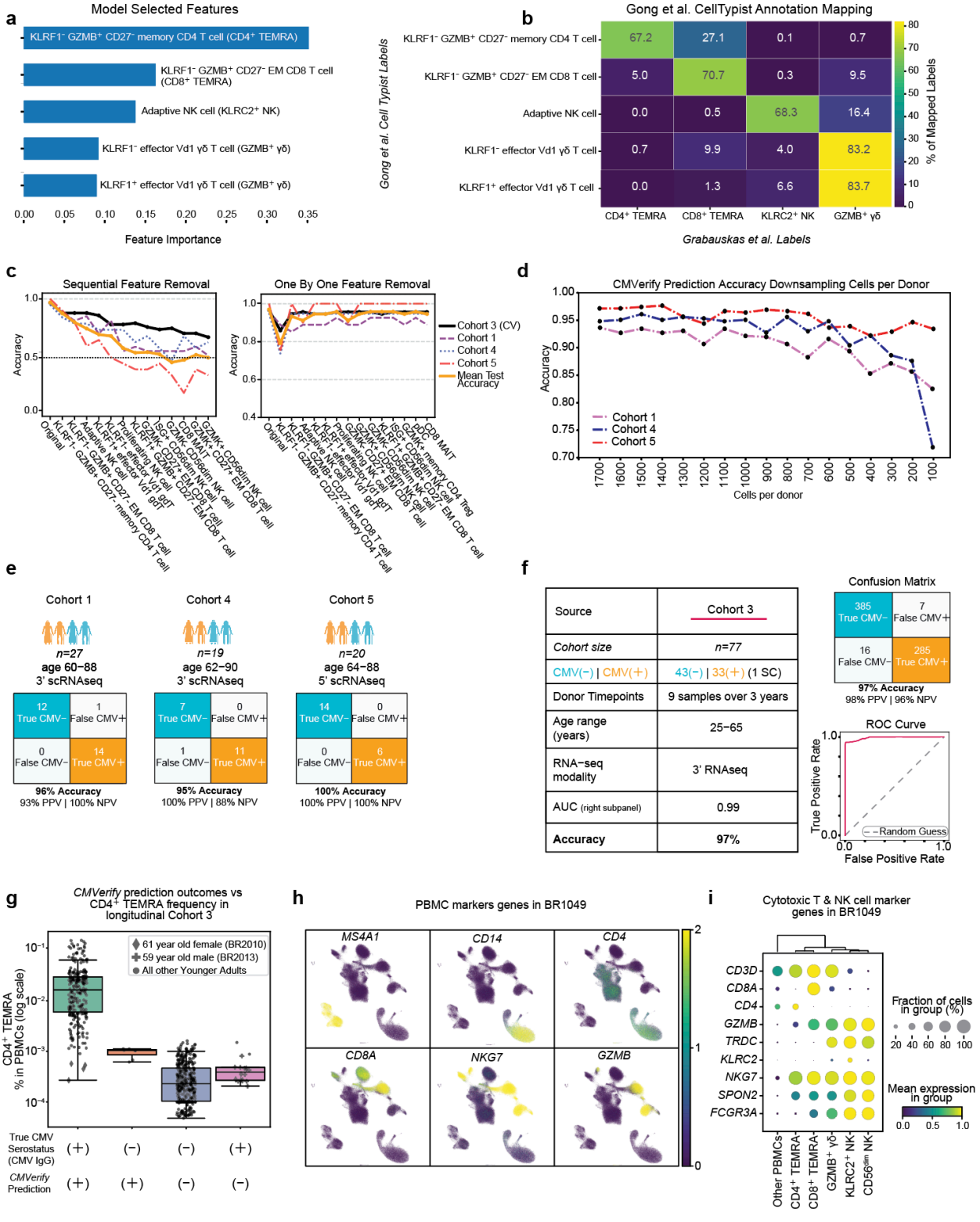

Figure S7: *CMVerify* selected features and extended prediction metrics. a, Top 5 selected features by *CMVerify* during the training procedure (see Methods) to predict CMV serostatus. Cell type labels were determined via the AIFI_L3 CellTypist model.[34] Top predictive feature: KLRF1-

GZMB+ CD27- memory CD4 T cell (i.e., $CD4^+$ TEMRA), and second-best predictive feature KLRF1- GZMB+ CD27- EM CD8 T cell (i.e., $CD8^+$ TEMRA). **b,** Heatmap showing percent of cell type labels mapped from AIFI_L3 CellTypist model to manual labels in our study. **c,** *Left:* Robustness of CMVerify predictions upon consecutive removal of predictive cell-type features. Predictive performance was iteratively evaluated after removing the top predictive feature at each iteration. After each removal step, the model was retrained and evaluated using cross-validation (CV) on the Cohort 3 training data as well as independent test cohorts (colored lines). *Right:* Robustness of CMVerify to removal of predictive cell-type features using a leave-one-feature-out ("knocking-out") analysis. In contrast to *Left* where features were removed sequentially according to importance, here each top predictive cell-type feature was removed individually while all remaining features were retained. The model was retrained after each feature removal and evaluated using independent test cohorts (colored lines). *Both:* The gold line indicates the mean test accuracy across cohorts, while the black line indicates the top cross-validation score during model training. **d,** Effect of reduced cell numbers on CMVerify performance. Cells were randomly downsampled (without replacement) to progressively lower cell numbers per donor across cohorts 1, 4, and 5. The procedure was repeated 20 times and prediction accuracy was averaged across runs. Accuracy remained high across cohorts despite substantial reductions in cell numbers. **e,** Confusion matrices for CMV serostatus prediction in each independent cohort from Figure 5b. **f,** *Left*: table summarizing the longitudinal Cohort 3. '1 SC' stands for 'one seroconverter' (donor ID: BR1049). *Right top*: Confusion matrix for CMV serostatus predictions from the longitudinal Cohort 3 in Figure 5c. *Right lower*: Receiver Operating Characteristic (ROC) curve quantifying performance of *CMVerify* on the longitudinal Cohort 3 in Figure 5c. **g,** CMVerify prediction outcomes and $CD4^+$ TEMRA frequencies per sample in the longitudinal Cohort 3. Diamonds represent the 61-year-old female (donor ID: BR2010), misclassified at five timepoints. Crosses represent the 59-year-old male (donor ID: BR2013), misclassified at nine timepoints. Circles represent all other samples. **h**, Expression of PBMC marker genes in donor BR1049 across all timepoints shown in Figure 5e**. i,** Expression of cytotoxic marker genes in select PBMC subsets from Figure 5f. PPV, positive predictive value; NPV, negative predictive value.

Figure S8:

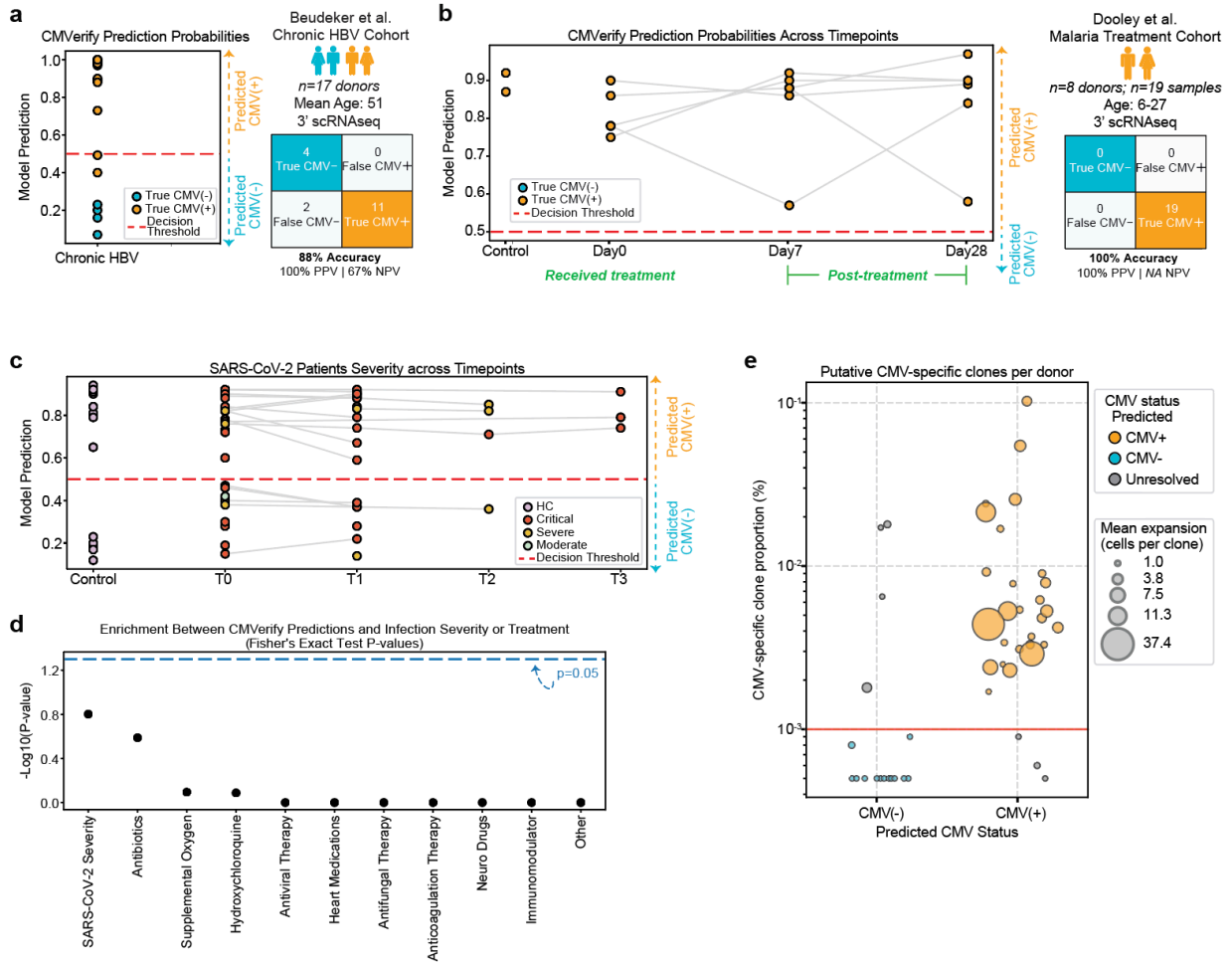

Figure S8: CMVerify performs accurately in non-healthy adult cohorts. a, CMVerify accurately predicted CMV serostatus in individuals with chronic HBV infections. *Left*: Y-axis represents model prediction probability, where each dot corresponds to a donor color coded for their CMV status using IgG ELISA. Samples above the decision Boundary (y = 0.5) are predicted to be CMV(+). *Right:* Confusion matrix for CMV serostatus predictions from this cohort. **b**, CMVerify accurately predicted CMV serostatus in individuals during and after malaria treatment. *Left*: Y-axis represents model prediction probability, each dot is a unique sample, color coded for known CMV serostatus (determined via CMV IgG ELISA). Samples above the Decision Boundary (y = 0.5) are predicted to be CMV(+). *Right:* Confusion matrix for CMV serostatus predictions from the malaria cohort. **c**, CMVerify predictions are not associated with SARS-CoV-2 infection severity and treatment. Dots are color-coded according to SARS-CoV-2 disease severity at study enrollment. Y-axis represents model prediction probability, each dot corresponds to a sample. Samples above the Decision Boundary (y = 0.5) are predicted to be CMV(+). **d,** Dot plot showing p-values (y-axis; log scale) from Fisher's exact test results evaluating enrichment of SARS-CoV-2 infection severity and treatments (x-axis) with respect to CMVerify predictions. Blue dotted line indicates the p=0.05 significance threshold, suggesting that no one of these metadata is significantly associated with CMV-predictions. **e,** Putative

CMV-specific clone proportion per donor. The Y-axis indicates donor-level CMV-specific clone proportion (defined as CMV-specific clones/ total clones). Bubbles are color coded for CMVerify predicted CMV status (blue for predicted CMV(-); orange for predicted CMV(+); gray for unresolved). Size of each bubble demonstrates mean CMV-specific clonal expansion (cells per clone).

Figure S9:

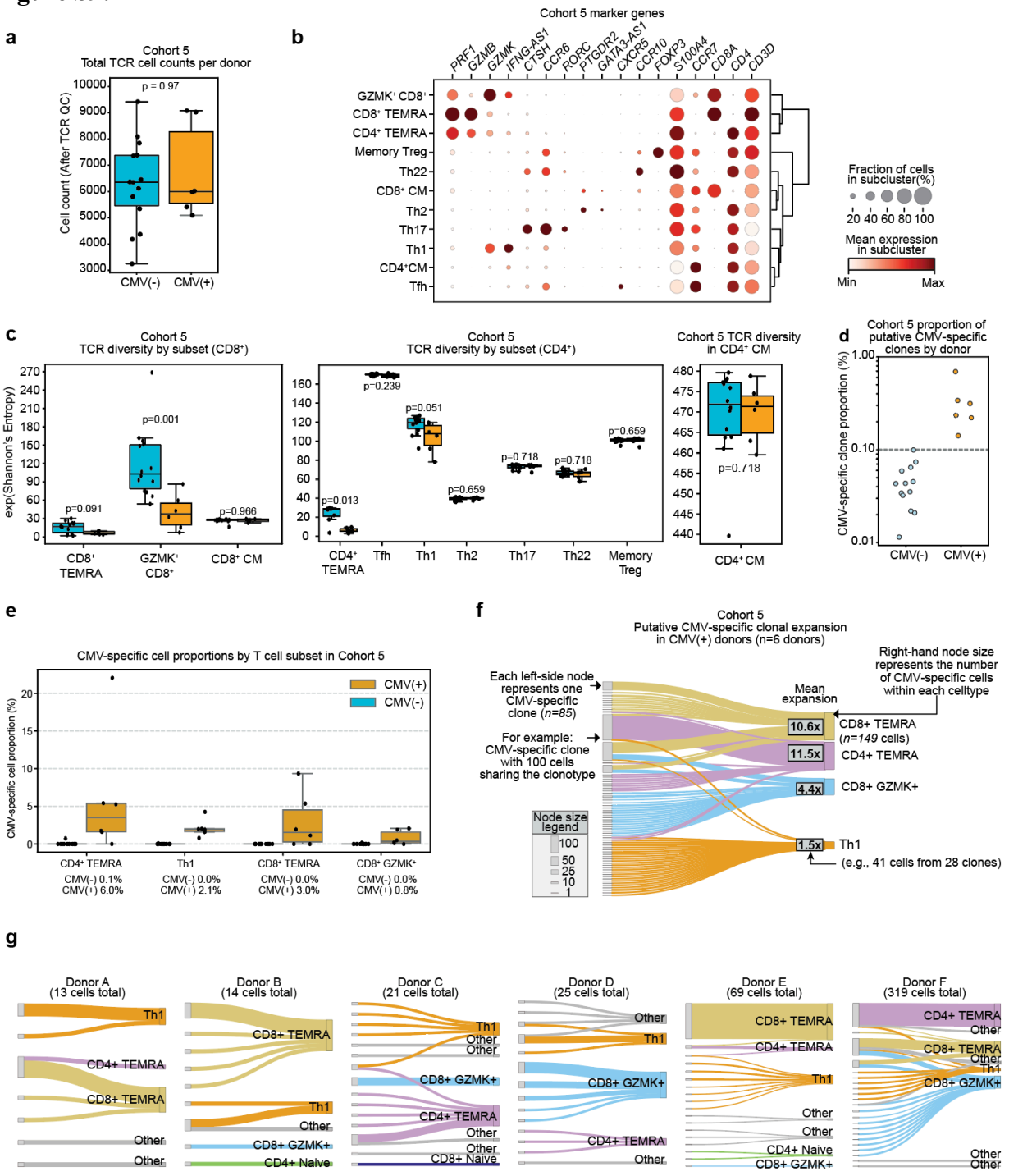

Figure S9: TCR diversity and CMV-specific clonal expansions in Cohorts 5 and 6. a, Total TCR cell counts after quality control in Cohort 5. **b,** Expression of marker genes in TCR-profiled cells in Cohort 5. **c,** TCR diversity in $CD8^+$ and $CD4^+$ T cells in Cohort 5. **d,** Putative CMV-specific clone proportion per donor in Cohort 5. **e,** CMV-specific T cells across expanded T cell subsets in Cohort 5. Boxplots show the percentage of T cells annotated as CMV-specific within each T cell subset for CMV(-) (blue) and CMV(+) (orange) donors. Mean values for each subset are shown below the x-axis, split by CMV status. **f,** Cohort-level Sankey plot of putative CMV-specific clones and their corresponding cells in CMV(+) individuals from Cohort 5. **g,** Donor-

level Sankey diagrams of putative CMV-specific clones and their corresponding cells in the six CMV(+) individuals from Cohort 5. For **f** and **g**, each node on the left represents one putative CMV-specific clone, with the node size proportional to the number of cells sharing the clonotype. Nodes on the right represent annotated cell types, with node size proportional to the number of putative CMV-specific cells detected within the cell type. For **a, c,** and **e,** box plots display the median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an individual sample. Mann-Whitney-U test (two-sided) was used to compare differences between groups. For **c**, TCR diversity was calculated using Shannon's entropy (e^H). Additionally, due to the small sample size in the CMV(+) group (6/20), diversity calculations within cell types were not FDR-corrected.

Figure S10:

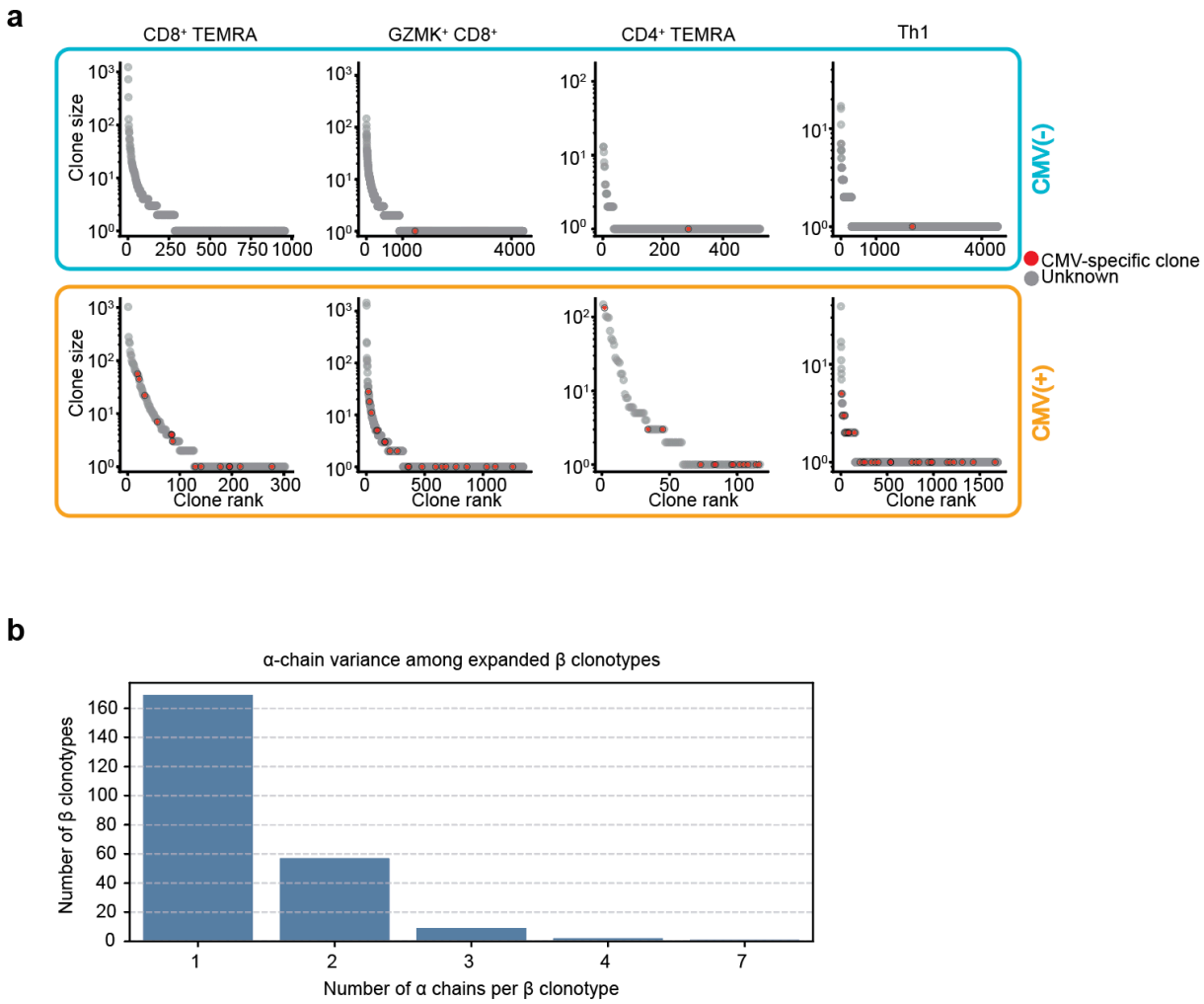

Figure S10: α-chain pairing among expanded β clonotypes and distribution of CMV-specific clonotype sizes in Cohort 5. a, Rank-abundance plots of TCR clonotypes within expanded T cell subsets in Cohort 5. Clonotypes are ordered by clone size (y-axis, log scale) and ranked by abundance (x-axis). Panels show clones for CMV- donors (top row) and CMV(+) donors (bottom row) across CD8+ TEMRA, GZMK+ CD8+, CD4+ TEMRA, and Th1 subsets. Each point represents a clonotype, with CMV-specific clones highlighted in red and non-CMV clones shown in gray. **b**, α-chain pairing among expanded β clonotypes. Distribution of the number of unique α chains associated with each expanded β clonotype (β chains detected in more than one cell). Each bar represents the count of β clonotypes paired with the indicated number of α chains (x-axis). Most expanded β clonotypes were paired with a single α chain, while a smaller number were associated with two α chains and only rare cases showed higher α variance.

Figure S11:

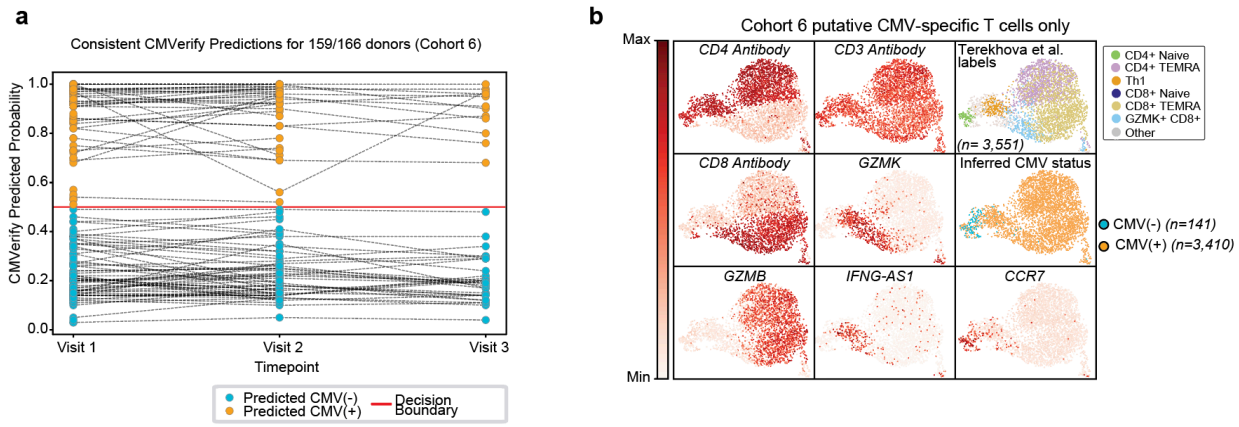

Figure S11: CMV status prediction and CMV-specific cells in Cohort 6. a, Longitudinal CMVerify predictions for each individual in Cohort 6. Y-axis represents model predicted probability, each bubble is a unique sample corresponding with a donor timepoint. Samples from the same donor are connected via gray lines. Based on the Decision Boundary (y = 0.5), all bubbles below the Decision Boundary (y < 0.5) represent prediction of CMV(-), vs. bubbles above the Decision boundary (y > 0.5) represent prediction of CMV(+). Bubbles are color coded for CMVerify prediction based on the prediction probability. **b**, Expression of marker genes and antibody staining in putative CMV-specific T cells only in Cohort 6.