# CMV reshapes lymphoid immunity in aging: a single-cell atlas with predictive modeling

**Authors:** Titas Grabauskas[1,2,]‡, Chris P. Verschoor[3,4] ‡, Luke Trinity[1]‡, Radu Marches[1], Asa Thibodeau[1], Djamel Nehar-Belaid[1], Giray Eryilmaz[1], Emilie Picard[3], Chia-Ling Kuo[2], Kenneth E. Schmader[5], Cathleen Colon-Emeric[5], Heather E. Whitson[5], Silke Paust[1], Adolfo García-Sastre[6-11], Jacques Banchereau[1,12]*, George A. Kuchel[2,13]*, and Duygu Ucar[1,14]*

**Affiliations:**
[1]The Jackson Laboratory for Genomic Medicine; Farmington, CT 06030, USA.
[2]University of Connecticut School of Medicine; Farmington, CT 06030, USA.
[3]Health Sciences North Research Institute; NOSM University; Sudbury, ON P3E 2H3, Canada.
[4]Department of Medicine, McMaster University; Hamilton, ON L8S 4L8 Canada
[5]Department of Medicine and Duke Center for the Study of Aging and Human Development (Aging Center), and GRECC, Durham VA Health Care System; Durham, NC 27705, USA.
[6]Department of Microbiology, Icahn School of Medicine at Mount Sinai; New York, NY 10029, USA.
[7]Global Health and Emerging Pathogens Institute, Icahn School of Medicine at Mount Sinai; New York, NY 10029, USA.
[8]Department of Medicine, Division of Infectious Diseases, Icahn School of Medicine at Mount Sinai; New York, NY 10029, USA.
[9]The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai; New York, NY 10029, USA.
[10]Department of Pathology, Molecular and Cell-Based Medicine, Icahn School of Medicine at Mount Sinai; New York, NY 10029, USA.
[11]The Icahn Genomics Institute, Icahn School of Medicine at Mount Sinai; New York, NY 10029, USA.
[12]Immunoledge LLC; Montclair, NJ 07042, USA.
[13]UConn Center on Aging, University of Connecticut School of Medicine; Farmington, CT 06032, USA.
[14]Institute for Systems Genomics, University of Connecticut Health Center; Farmington, CT 06032, USA.
‡ These authors contributed equally to this work.
* Corresponding authors: duygu.ucar@jax.org, jacques.banchereau@gmail.com, kuchel@uchc.edu

**Abstract:** Cytomegalovirus (CMV), a common herpesvirus, establishes lifelong latency and increases in prevalence with age; yet its systemic impact on the aging immune system remains incompletely understood. We profiled circulating immune cells from healthy older adults (median age: 73) who were CMV(+) or CMV(-) using single-cell RNA-sequencing and validated key findings by flow cytometry. CMV(+) individuals exhibited significant expansion of adaptive immune cells: $CD4^+$ and $CD8^+$ TEMRA T, $GZMK^+$ $CD8^+$ T, γδ T, and atypical B cells. Among innate immune cells, monocytes and dendritic cells remained largely unchanged while *KLRC2*$^+$ (adaptive) NK cells increased and $CD56^{dim}$ NK cells decreased. To facilitate CMV assessment in datasets with unknown CM serostatus, we developed *CMVerify*, a machine learning classifier that accurately predicts CMV serostatus from single-cell data across platforms and age groups (97%

accuracy). These findings reveal extensive CMV-associated immune remodeling in older adults and underscore the importance of incorporating CMV status in studies of immune aging.

**Main Text:**

The immune system undergoes significant changes with aging, characterized by a decline in naïve cell populations, expansion of memory and terminally differentiated T and B cells, and increased pro-inflammatory activity[1,2]. These changes contribute to weakened immune responses and increased susceptibility to disease[1,3,4]. However, immune aging varies significantly between individuals[5], contributing to differences in immune health, vaccine efficacy, and therapeutic responses among older adults of the same chronological age[6,7]. Factors contributing to this heterogeneity include sex, genetics, obesity, environmental exposures, and chronic latent infections such as Cytomegalovirus (CMV)[8–12].

CMV, a DNA herpesvirus transmitted through bodily fluids[13], establishes lifelong latency in hematopoietic stem cells and myeloid progenitor cell reservoirs[14]. It is one of the most prevalent chronic infections globally, with seropositivity increasing with age[15]. In the United States, 60% to 90% of adults aged 60 and older are CMV-positive (hereinafter referred to as CMV(+), with CMV(-) denoting CMV-negative)[15]. Although largely asymptomatic, in healthy adults[16], latent CMV has been linked to more severe outcomes of SARS-CoV-2[17] and tuberculosis[18] infections. Conversely, in melanoma patients, CMV seropositivity was associated with improved survival following anti-PD-1 checkpoint blockade[19]. Despite its high prevalence and clinical impact, CMV serostatus is often overlooked in immune aging studies, representing a significant but underappreciated confounder.

In addition to health outcomes, latent CMV infection has also been associated with alterations in T and NK cell subsets[1,20,21]. In T cells, CMV has been linked with the expansion of terminally differentiated $CD4^+$ and $CD8^+$ T cells —a phenomenon termed as "memory inflation."[22] CMV-

74 specific $CD8^+$ T cell clones can constitute up to 32% of total $CD8^+$ T cell pool[23], potentially
75 limiting responses to novel antigens. CMV is also associated with the accumulation of
76 unconventional γδ T lymphocytes[24–26], particularly circulating Vδ1 γδ T cells[24]. In NK cells, CMV
77 has been shown to correlate with expansion of adaptive NK cells[27,28] expressing NKG2C, which
78 recognizes CMV-derived peptides to restrict viral spread[29]. Despite these individual observations,
79 a systematic exploration of CMV's effects on the circulating immune system of older adults
80 remains lacking.
81
82 To address this gap, we profiled peripheral blood from healthy older adults (60–92 years old),
83 stratified by CMV serostatus. Our systems immunology approach uncovered extensive remodeling
84 of circulating immune cells, findings that were independently validated in a younger adult cohort.
85 This study presents the first single-cell transcriptional atlas of latent CMV infection in older adults,
86 defining a robust immune signature. Motivated by the strength of this signal—and to enable CMV
87 reporting in both newly generated and previously published single-cell datasets—we developed
88 *CMVerify*, a highly accurate machine learning model that predicts CMV serostatus. *CMVerify* is
89 easy to use and freely available as an open-source Python package (pypi.org/project/CMVerify/).
90
91 **Study design: older adult cohorts stratified by CMV serostatus**
92 We recruited two cohorts of healthy older adults during 2017-2018 from communities surrounding
93 Hartford, Connecticut, USA, and Sudbury, Ontario, Canada. The Discovery Cohort comprised 36
94 healthy older adults (n=17 CMV(+), n=19 CMV(-)), whose peripheral blood mononuclear cells
95 (PBMCs) were profiled using single-cell RNA sequencing (scRNAseq) and flow cytometry (Fig.
96 1a). The Validation Cohort included 63 older adults (n=40 CMV(+), n=23 CMV(-)) whose PBMCs
97 were profiled using flow cytometry alone (Fig. 1a). Participants were free from confounding

autoimmune diseases, were not undergoing treatments that might influence immune function and composition (e.g., immunotherapy) and were not recently vaccinated before blood collection. CMV serostatus was determined by detecting the CMV IgG antibodies in serum using ELISA.

The median age of the Discovery Cohort was 71 years old, including men (n=17) and women (n=19). CMV(+) and CMV(-) groups were comparable in age (p=0.90), body mass index (BMI; median 27.8, p=0.51), and frailty index[30] (median 0.085, p=0.50) (Fig. 1b, Extended Data Fig. 1a, and Supplementary Table 1). Other characteristics, including race, cancer history, and medication use, were also evenly distributed between the two groups (Supplementary Table 1).

The Validation Cohort had a median age of 75 years, including men (n=24) and women (n=39). Similar to the Discovery Cohort, the CMV(+) and CMV(-) groups were comparable in age (p=0.11), BMI (median 27.1, p=0.93), and frailty index[30] (median 0.09, p=0.88) (Fig. 1b and Supplementary Table 1). Race and cancer history were also evenly distributed between the groups (Supplementary Table 1).

**Expansion of memory $CD8^+$ T and γδ T cell subsets is associated with CMV positivity**

PBMCs from the Discovery Cohort were profiled using scRNAseq, capturing 144,324 cells (n=68,642 CMV(+), n=75,682 CMV(-)) after quality control (Fig. 1c,d, and Extended Data Fig. 1b). CMV(+) individuals had higher frequency of $CD8^+$ T cells compared to CMV(-) individuals (25% vs 11% of PBMCs; p=0.0012; Fig. 1e,f), confirmed by flow cytometry in the Validation Cohort (p<0.001; Fig. 1g). Absolute $CD8^+$ T cell counts were 1.78-fold higher in CMV(+) individuals (p=0.04; Fig. 1h), and the $CD4^+/CD8^+$ T cell ratio was lower in CMV(+) older adults in both cohorts (Discovery: p=0.006; Validation: p<0.001) (Fig. 1i).

122

123 Clustering of CD8$^+$ T cells (n=24,983) (Fig. 2a and Extended Data Fig. 2a) identified seven
124 subsets: (i) naïve (*CCR7*$^+$, *SELL*$^+$, *LEF1*+), (ii) central memory (CM; *CCR7*$^+$), (iii) mucosal-
125 associated invariant T (MAIT; *ZBTB16*$^+$, *KLRB1*$^+$), (iv) GZMK$^+$ CD8$^+$ (*CCR7*$^-$, *GZMK*$^+$), (v)
126 TEMRA (*GZMB*$^+$, *KLRD1*$^+$), (vi) *GZMK*$^+$ *γδ* (*TRDC*$^+$, *GZMK*$^+$), and (vii) *GZMB*$^+$ *γδ* (*TRDC*$^+$,
127 *GZMB*$^+$) cells. All subsets, except MAIT cells, were significantly altered with CMV positivity.
128 GZMK$^+$ CD8$^+$ ($p<0.001$), TEMRA ($p<0.001$), and GZMB$^+$ γδ ($p<0.001$) cells were the most
129 expanded subsets in CMV(+) older adults (Fig. 2b,c and Extended Data Fig. 2b).

130

131 CD8+ TEMRA cells exhibited a 9.5-fold expansion in CMV(+) compared CMV(-) adults (7.2%
132 vs 0.8% of PBMCs; $p<0.001$; Fig. 2c), which was confirmed by flow cytometry in the Validation
133 Cohort ($p<0.001$; Fig. 2d). Absolute CD8$^+$ TEMRA counts supported these results ($p=0.001$; Fig.
134 2e). γδ T cells, which recognize antigens independently of MHC presentation,[31] were also more
135 frequent in CMV(+) adults, particularly the GZMB$^+$ γδ subset which showed a 5.8-fold increase
136 (3.7% vs 0.6%; $p<0.001$; Fig. 2c and Extended Data Fig. 2b). Increase in γδ T cell frequency was
137 confirmed in the Validation Cohort by flow cytometry ($p=0.05$; Fig. 2f). GZMK$^+$ CD8$^+$ T cells
138 were also expanded in CMV(+) adults (6% vs 1.6% of PBMCs; $p<0.001$; Fig 2c).

139

140 To assess whether these CMV-associated alterations were specific to older adults, we re-analyzed
141 publicly available scRNAseq datasets from younger adults (Younger Adult Cohort; n=47, 25-35
142 years old; n=45, 55-65 years old; Extended Data Fig. 2c) with known CMV serostatus (CMV(+)
143 n=42; CMV(-) n=50)[32] (Extended Data Fig. 2b). CD8$^+$ TEMRA frequency, GZMB$^+$ γδ T cells,
144 GZMK$^+$ CD8$^+$ T cells were also higher in CMV(+) younger adults ($p<0.001$; Fig. 2g).

145

146 In-depth analyses of each $CD8^+$ subset revealed very few differentially expressed genes (DEGs;
147 FDR<0.05) between CMV(+) and CMV(-) older individuals (Supplementary Table 2). To further
148 explore cell-intrinsic differences, we calculated single-cell scores for cytotoxicity and lymphoid
149 innateness—a gradient that evaluates transcriptional similarity of T cells to innate lymphocytes
150 (Supplementary Table 3)[33,34]. These scores were largely comparable between CMV(+) and CMV(-
151 ) donors across all subsets, except for $GZMK^+$ $CD8^+$ T cells that displayed an increased
152 cytotoxicity score for CMV(+) individuals (Extended Data Fig. 2e, f).
153
154 Together, CMV seropositivity is associated with substantial compositional remodeling of the
155 $CD8^+$ T cell compartment with minor cell-intrinsic alterations in older adults.
156
157 **$CD4^+$ TEMRA T cells are expanded in CMV(+) older adults**
158 $CD4^+$ T cells (55,927) (Fig. 2h and Extended Data Fig. 3a) were clustered into four subsets: (i)
159 naïve (*CCR7*$^+$); (ii) T-regulatory cells (Tregs; *FOXP3*$^+$, *TIGIT*$^+$), (iii) effector memory (EM;
160 *S100A4*$^+$, *GZMB*$^-$), and (iv) TEMRA cells (*GZMB*$^+$, *GNLY*$^+$). CMV(+) individuals exhibited a
161 lower frequency of naïve $CD4^+$ T cells (p=0.04) and higher frequency of $CD4^+$ TEMRA cells
162 (p=<0.001; Fig. 2j and Extended Data Fig. 3b). $CD4^+$ TEMRA cells were 6.4-fold higher in
163 CMV(+) individuals (2.15% vs 0.34% of PBMCs), confirmed by flow cytometry (p=0.009; Figure
164 2k) in the Validation Cohort. Absolute counts of $CD4^+$ TEMRA cells were also higher in CMV(+)
165 (p=0.08; Fig. 2l). The Younger Adult Cohort displayed similar expansion of $CD4^+$ TEMRA cells
166 (p<0.001; Fig. 2m and Extended Data Fig. 3c). DEG analysis yielded few (n=3) significant genes
167 (Supplementary Table 2), with *CCL5* upregulated in Tregs from CMV(+) older adults.
168

169 Further clustering of the CD4$^+$ memory compartment (EM; n=19,691; Extended Data Fig. 3d)
170 revealed five memory subsets (i) T-helper 1-like (Th1-like; *IFNG-AS1*$^+$, *GZMK*$^+$), (ii) T-helper
171 17-like (Th17-like; *RORC*$^+$, *CCR6*$^+$, *CTSH*$^+$), (iii) T follicular helper-like (Tfh-like; *CXCR5*$^+$) (iv)
172 T-helper 2-like (Th2-like; *GATA3*$^+$ ), and (v) T-helper 10-like (Th10-like; *CCR10*$^+$)[35]. None of
173 these memory subsets remodeled with CMV seropositivity in our cohort (Extended Data Fig. 3e,f)
174 or in the Younger Adult Cohort (Extended Data Fig 3g,h).
175
176 Thus, CMV seropositivity in older adults is associated with higher frequency of CD4$^+$ TEMRA
177 cells without alterations of other memory CD4$^+$ T cell subsets.
178
179 **Atypical B cells (ABCs) are expanded in CMV(+) older adults**
180 B cells (n=12,573; Fig. 3a and Extended Data Fig. 4a) clustered into five subsets: (i) naïve (*IL4R*$^+$,
181 *IGHM*$^+$), (ii) transitional (*CD9*$^+$, *MME*$^+$), (iii) ABC (*CD27*$^-$, *ITGAX*/CD11c$^+$, *TBX21*/Tbet$^+$), (iv)
182 memory (*CD27*$^+$, *IGHG1*$^+$, *IGHA1*$^+$), and (v) plasmablasts (PB; *JCHAIN*$^+$, *MZB1*$^+$). CMV(+)
183 individuals had higher frequencies of ABCs, with a 3.4-fold increase (0.27% vs 0.9%; p=0.01; Fig.
184 3b,c and Extended Data Fig.4b). This expansion was also observed in 55–65-year-old CMV(+)
185 individuals from the Younger Adult Cohort (p=0.04; Fig. 3d and Extended Data Fig. 4c). No DEGs
186 were found between CMV(+) and CMV(-) older adults in any B cell subset (Supplementary Table
187 2).
188
189 Therefore, CMV seropositivity is linked with expansion of ABCs without affecting other B cell
190 subsets.
191
192 **CD56$^{dim}$NK cells are reduced and NKG2C$^+$ NK cells are expanded in CMV(+) older adults**

193 NK cells (n=17,222) (Fig. 3e,f) clustered into three subsets: (i) CD56$^{dim}$ NK cells (*SPON2*$^{+}$,
194 *FCGR3A*$^{+}$, *GZMB*$^{+}$), (ii) CD56$^{bright}$ NK cells (*NCAM1*/CD56 $^{+}$, *GZMK*$^{+}$, *XCL1*$^{+}$), and (iii)
195 KLRC2$^{+}$ NK cells (*KLRC2*$^{+}$/NKG2C), which closely resembled adaptive NK cells (Extended Data
196 Fig. 4d)[36,37].
197
198 CD56$^{dim}$ NK cell frequency was lower in CMV(+) individuals in both our older adult cohort (6.6%
199 vs 12.2% of PBMCs; p=0.02; Fig. 3g,h) and the Younger Adult Cohort (p<0.001; Fig. 3j and
200 Extended Data Fig. 4e). Conversely, KLRC2$^{+}$ NK cell frequency was higher by 2.3-fold in
201 CMV(+) older adults, (2% vs. 0.8%; p=0.011, Fig. 3h), which was confirmed by flow cytometry
202 in the Validation Cohort (p<0.001; Fig. 3i and Extended Data Fig. 4f). KLRC2$^{+}$ NK cells were
203 also more frequent in CMV(+) individuals within the Younger Adult Cohort (p<0.001; Fig. 3j and
204 Extended Data Fig. 4e). No DEGs were found between CMV(+) and CMV(-) older adults in NK
205 subsets (Supplementary Table 2).
206
207 **Monocytes and dendritic cells are comparable between CMV(+) and CMV(-) older adults**
208 Monocytes (n=30,132) (Fig. 4a and Extended Data Fig. 5a) clustered into four subsets: (i) CD14$^{+}$
209 monocytes (*CD14*$^{+}$, *LYZ*$^{+}$, *S100A8*$^{+}$, *S100A9*$^{+}$), (ii) ISG-high CD14$^{+}$ monocytes (*CD14*$^{+}$, *IFI44L*$^{+}$,
210 *MX1*$^{+}$), (iii) intermediate monocytes (*FCGR3A*$^{+}$, *CD14*$^{+}$, *HLA-DRB1*$^{+}$), and (iv) CD16$^{+}$ monocytes
211 (*FCGR3A*$^{+}$, *FCER1G*$^{+}$). No differences in monocyte subset frequencies or DEGs were observed
212 between CMV(+) and CMV(-) individuals (Fig. 4b, Extended Data Fig. 5b, and Supplementary
213 Table 2). Monocyte subset was also not altered with CMV in the Younger Adult Cohort (Extended
214 Data Fig. 5c).
215

216 Dendritic cells (DCs) (Fig. 4c) clustered into five subsets (i) monocyte-derived dendritic cells
217 (Mono-DC; *S100A8*$^{+}$, *CD14*$^{+}$, *CLEC10A*$^{+}$), (ii) type-2 classical dendritic cells (cDC2; *CLEC10A*$^{+}$,
218 *CD1C*$^{+}$), (iii) type-1 classical dendritic cells (cDC1; *CLEC9A*$^{+}$), (iv) AXL$^{+}$ dendritic cells (AXL$^{+}$
219 DC; *AXL*$^{+}$), and (v) plasmacytoid DCs (pDCs; *IRF7*$^{+}$, *JCHAIN*$^{+}$). Among these subsets, pDCs
220 showed a trend toward lower frequency in CMV(+) individuals (0.35% vs 0.6% of PBMCs)
221 (p=0.07; Fig. 4, d,e and Extended Data Fig. 5d). A modest but significant reduction in pDC
222 frequency was also observed in the Younger Adult Cohort (0.35% vs 0.41%; p=0.03; Fig. 4f and
223 Extended Data Fig. 5e). No DEGs were found between CMV(+) and CMV(-) older adults in DC
224 subsets (Supplementary Table 2).
225
226 Overall, our older adult cohorts revealed pronounced CMV-associated remodeling of lymphoid
227 immune cell subsets, with minimal alterations in myeloid cells (Fig. 4g and Supplementary Table
228 4). To estimate the contribution of clinical factors to lymphocyte frequency variance in the
229 Discovery Cohort, we conducted Principal Variance Component Analysis[38], which identified
230 CMV serostatus as the dominant contributor (43%), followed by BMI (11%) (Extended Data Fig.
231 5g). Finally, leveraging longitudinal samples from the Younger Adult Cohort, revealed that CMV-
232 associated immune signature remains stable over three years (Fig. 4h and Extended Data Fig. 5f),
233 underscoring the persistent nature of latent CMV-driven immune remodeling.
234
235 ***CMVerify*—a novel tool that accurately predicts CMV serostatus from scRNAseq data**
236 Given the increasing prevalence of latent CMV with age and its association with extensive immune
237 remodeling, we developed a machine learning model—*CMVerify*—to predict latent CMV
238 serostatus from existing single-cell transcriptomics datasets. *CMVerify* leverages a Random Forest

classification model that uses PBMC subtype frequencies as input (i.e., data features) to predict latent CMV serostatus (i.e., classification output) (Fig. 5a).

*CMVerify* was trained on the Younger Adult Cohort (n=92 total; n=42 CMV(+); n=50 CMV(-); Extended Data Fig. 2c) using leave-one-out cross-validation, achieving 96.7% training accuracy. Top predictive features included $CD4^+$ TEMRA, $CD8^+$ TEMRA, $KLRC2^+$ NK cells, and $GZMB^+$ γδ T cells (Extended Data Fig. 6a,b), consistent with our CMV-associated immune signature. We first validated *CMVerify* in our Discovery Cohort, where it correctly predicted CMV serostatus in 26 out of 27 individuals (Fig. 5b and Extended Data Fig. 6c). We then further tested *CMVerify* on two independent older adult cohorts with different chemistries using available CMV IgG and PBMC scRNAseq data: Cohort 1 (n=19, age 62-90) profiled using 10X 3' scRNAseq, Cohort 2: (n=20, age 64-88) profiled using 10X 5' scRNAseq. Across these datasets, *CMVerify* achieved 97% average accuracy, 98% positive predictive value (PPV), and 96% negative predictive value (NPV) (Fig. 5b, Extended Data Fig. 6c).

To evaluate longitudinal performance, *CMVerify* was applied to 77 individuals from the Younger Adult Cohort, each with PBMC scRNAseq data across 9 timepoints. The model maintained 97% accuracy, correctly classifying 670 of 693 longitudinal samples (Fig. 5c and Extended Data Fig 6d). Misclassifications primarily stemmed from two donors with unusually low $CD4^+$ TEMRA frequencies (Extended Data Fig 6e). Remarkably, in a donor who seroconverted over the study (35-year-old male, ID: BR1049), *CMVerify* successfully tracked the transition, shifting its prediction from CMV(-) to CMV(+) in parallel with rising CMV IgG titers (Fig. 5c,d). This transition was accompanied by rapid and dramatic changes in PBMC composition—including a 90.99-fold increase in $CD4^+$ TEMRAs, 8.74-fold increase in $CD8^+$ TEMRAs, and 3.37-fold

increase in KLRC2$^{+}$ NK cells (Fig. 5e and Extended Data Fig. 6f) — demonstrating the sensitivity of *CMVerify* to biological shifts following CMV infection.

Together, these results establish *CMVerify* as a robust, cross-platform tool for predicting CMV serostatus from human scRNAseq PBMC datasets. The publicly available Python package (pypi.org/project/CMVerify/) facilitates CMV predictions in both new and existing human studies that leverage scRNAseq profiling in PBMCs, enabling better separation of CMV- and age-driven immune changes.

## Discussion

Our systematic comparison of the circulating leukocytes from CMV(+) and CMV(-) older adults demonstrates that latent CMV infection is associated with extensive remodeling of the lymphoid cells, primarily linked to cytotoxicity, while having limited effects on myeloid cells. Key findings include the CMV-associated expansion of CD4$^{+}$ and CD8$^{+}$ TEMRA cells, GZMK$^{+}$ CD8$^{+}$ T cells, cytotoxic GZMB$^{+}$ γδ T cells, NKG2C$^{+}$ NK cells, and atypical B cells (ABCs), alongside a decline in NK CD56$^{dim}$.

The dramatic expansion of CD4$^{+}$ and CD8$^{+}$ TEMRA cells in CMV(+) individuals corroborates prior findings[22,32,39], but our study advances this by showing that GZMK$^{+}$ CD8$^{+}$ T cells also expand with CMV positivity in both older and younger adult cohorts. While these cells have previously shown to expand with age[2], our data highlight that expansion of these populations is not solely age driven.

286 Our study identified GZMB$^{+}$ γδ T cells as one of the subsets most strongly associated with CMV
287 seropositivity in both older and younger adults, advancing prior reports of Vδ1$^{+}$ γδ T cell
288 expansion[24]. CMV's ability to evade conventional T cell responses *via* MHC class I
289 downregulation[40] may promote MHC-independent recognition by γδ T cells to eliminate infected
290 cells[41]. Furthermore, NKG2C$^{+}$ NK cells, crucial for targeting CMV-infected cells through
291 activating receptors[29,42], were significantly expanded in CMV(+) older adults, extending
292 observations in younger cohorts[27,28,43]. In contrast, the reduction in CD56$^{dim}$ NK cells across both
293 younger and older CMV(+) cohorts suggests broader remodeling of the NK cell compartment with
294 latent CMV infection.
295
296 ABCs, associated with chronic viral infections in mice[44] and autoimmune conditions in humans[45],
297 were increased with CMV both in our older adult cohort and in the independent Younger Adult
298 Cohort. Although they expand with primary CMV infection during pregnancy[46], how they
299 accumulate and influence the immune system function with latent CMV is unknown. CMV-
300 induced chronic activation and differentiation through extra-follicular pathways[45,47] may drive
301 ABC expansion, potentially mediated by IFN-γ from expanded TEMRA and NK cells[48]. ABCs
302 might produce CMV-specific antibodies, potentially contributing to the control of latent infection.
303
304 Although CMV serostatus can be readily assessed from plasma, many genomic studies fail to
305 account for this important confounder. Tools that can predict latent CMV infection in the absence
306 of serum immunoglobulin data (typically, CMV IgG) or access to extra blood samples are therefore
307 highly valuable. Existing methods can leverage cytometry[49] or T cell receptor sequencing data[50]
308 to predict CMV serostatus with high accuracy. However, no such tool exists for scRNAseq
309 datasets, despite their increasing use in human immune aging studies. To address this gap, we

310 developed *CMVerify*, a machine learning tool that enables accurate CMV serostatus prediction
311 from single-cell transcriptomic data from PBMCs. Trained on younger adult data, *CMVerify*
312 successfully generalized to three independent older adult cohorts, underscoring the age-
313 independent nature of CMV-associated immune remodeling. By enabling CMV inference without
314 the need for plasma samples or additional assays (e.g., ELISA), *CMVerify* provides the field with
315 a practical and scalable solution to disentangle CMV-associated versus age-related immune
316 changes in existing and future single-cell datasets.
317
318 In summary, our findings highlight profound CMV-associated remodeling of lymphoid cells in
319 older adults, accompanied by limited changes in myeloid cells. These alterations should be taken
320 into consideration in immune aging and profiling studies.
321
322 **Limitations**
323
324 Our study has several limitations. First, the cohort predominantly consisted of Caucasian
325 individuals; race/ethnicity might influence the observed associations. Second, the cross-sectional
326 design of our study limits causal inferences regarding CMV's role in immune aging. Although we
327 demonstrated that CMV-associated changes remain stable over a two-year period, the longer-term
328 effects of CMV require further investigation. Additionally, the *CMVerify* model is intended for
329 use in healthy adult populations and is not recommended for other age groups, e.g., infants. Despite
330 these limitations, our study provides valuable and novel insights into the systemic effects of latent
331 CMV infection on immune aging.
332

333 **Materials and Methods**

334

335 **Cohort Description**

336

337 ***Discovery Cohort***. The Discovery Cohort comprised individuals recruited from two separate
338 studies:

339 1. **UConn Center on Aging Bacterial Pneumonia Vaccine Study**: Twenty healthy older
340 adults aged 60 years and older were recruited at the University of Connecticut Center on
341 Aging during the 2017 and 2018 seasons (IRB number: 16-071J; R01AG052608;
342 R01AI142086). Participants did not have confounding treatments or conditions that would
343 influence immune function and composition (e.g., immunotherapy, immunosuppressive
344 disorders, and prednisone in doses greater than 10 mg/day) [51].

345 2. **Immune Response to High-Dose vs. Standard Dose Influenza Vaccine at the**
346 **University of Connecticut Health Center (UCHC)**: Sixteen healthy older adults aged 65
347 years and older were recruited at UCHC and Health Sciences North Research Institute
348 (HSNRI) in 2017 to study resilience to influenza infections. Exclusion criteria for these
349 participants included known immunosuppressive disorders or medications, including
350 prednisone in doses greater than 10 mg/day.

351 ***Validation Cohort***. The Validation Cohort included 63 individuals recruited for the Immune
352 Response to High-Dose vs. Standard Dose Influenza Vaccine study at UCHC and HSNRI in 2018.
353 This cohort consisted of healthy older adults aged 65 years and older. Exclusion criteria were the
354 same as in the Discovery Cohort and included known immunosuppressive disorders or
355 medications, such as prednisone in doses greater than 10 mg/day.

356

357 **Participant Characteristics**. The CMV(+) and CMV(-) groups in both the Discovery and
358 Validation Cohorts had comparable distributions in age, sex, body mass index (BMI), and frailty
359 index (see Supplementary Table 1). All individuals in the Discovery Cohort (n=36) and the
360 Validation Cohort (n=63) were recruited from the communities of Sudbury, Ontario, Canada, and
361 Hartford, Connecticut, USA.

362

363 **Ethics Approval**. The study protocols were approved by the Institutional Review Board of the
364 University of Connecticut Health Center (UConn Health IRB #20-130-J-1 or JAX IRB #2019-
365 006-JGM) and the Health Sciences North Research Ethics Board (#985). Written informed consent
366 was obtained from all participants prior to inclusion in the study.

367

368 **Sample Collection**. Peripheral blood mononuclear cells (PBMCs) were isolated by Lymphoprep
369 (StemCell Technologies) gradient centrifugation within 1h after collection of blood samples by
370 Lymphoprep (StemCell Technologies) gradient centrifugation whereas serum was isolated from
371 the blood collected in red top collection tubes (BD Vacutainer). The isolated PBMCs were
372 resuspended in a cryoprotective solution containing 90% FBS and 10% DMSO, then transferred
373 to liquid nitrogen for long-term storage whereas serum was stored at -80°C. Further details on the
374 processing of previously published blood samples (Discovery Cohort, n=20; Validation Cohort,
375 n=63) can be found in the original publications [26,51,52].

376

377 **Cytomegalovirus Serostatus.** CMV serostatus was determined by ELISA using the CMV IgG
378 ELISA kits (Aviva Systems Biology and Genesis Diagnostics Ltd.). Serum samples were thawed

379 at room temperature, diluted, and analyzed according to manufacturer's recommendations.
380 Calculation of the results was done based on the appropriate controls included in the kit.
381
382 **Single-Cell RNA-Sequencing**. In the Discovery Cohort, we generated a total of 27 baseline
383 PBMC scRNAseq samples from older adults as part of two separate studies. All of these samples
384 were generated at the Jackson Laboratory Single Cell Biology Lab using the same pipeline.
385 Briefly, cells were washed and resuspended in PBS containing 0.04% BSA and immediately
386 processed as follows. Cell viability was assessed on a Countess FL II automated cell counter
387 (Thermofisher) using Trypan Blue staining, and up to 12,000 cells from each suspension were
388 loaded onto one lane of a 10x Genomics Chip G. Single cell capture, barcoding, and library
389 preparation were performed using the 10x Genomics Chromium Controller platform version 3.1
390 NEXTGEM chemistry and according to the manufacturer's protocol (CG000315, 10x Genomics).
391 cDNA and libraries were assessed for quality via Tapestation 4200 (Agilent) and Qubit
392 fluorometer (ThermoFisher), quantified by KAPA qPCR, and sequenced on a NovaSeq 6000
393 (Illumina) S4 v1.5 200 flow cell using a 28-10-10-90 asymmetric read configuration targeting a
394 read depth of 50,000 per cell.
395
396 **scRNAseq data analyses.** Illumina base call files for all libraries were converted to FASTQs using
397 bcl2fastq v2.20.0.422 (Illumina, CA, USA) and FASTQ files associated with baseline gene
398 expression libraries were aligned to the GRCh38 reference assembly with v32 annotations from
399 GENCODE (10x Genomics GRCh38 reference 2020-A) using the version 6.1.1 Cell Ranger count
400 pipeline (10x Genomics, CA, USA), as previously described[51]. The resulting gene-cell expression
401 matrices, were then corrected for ambient RNA contamination using SoupX[53]. The corrected gene-
402 cell expression matrices were then used for downstream analysis. We used the Scrublet[54] package

403 in Python to estimate the doublet score for each sample with an estimated doublet set to 0.1.
404 Doublets were manually confirmed to be doublets via expression of PBMC lineage markers before
405 being removed from downstream analyses. Cells with (i) genes that were not detected in at least 3
406 cells, (ii) mitochondrial reads greater than 25%, and (iii) fewer than 200 features were excluded.
407 Filtered gene expression matrices were merged and processed using the Scanpy toolkit (version
408 1.9.1)[55]. Read counts were log normalized to the total number of reads in each individual cell to
409 count-per-million (CPM) using the "normalized_total" and "log1p" functions. Genes with the
410 highest normalized dispersion were identified using the "highly_variable_genes" function. Gene
411 expression matrices were scaled using "scale" function and Principal Component Analysis (PCA)
412 was performed using the "pca" function. Harmony[56] was applied using the "harmony_integrate"
413 function, from Scanpy, to correct batch effects across samples. For the baseline samples, 100
414 Principal Components (PCs) were used to calculate nearest neighbors via the "neighbors" function
415 which were used to perform uniform manifold approximation and projection (UMAP) *via* the
416 "umap" function. Initial clusters were determined using the "leiden" function (resolution=0.8) for
417 all PBMCs. Multiple rounds of marker identification, cell type annotation, and manual inspection
418 and doublet removal, were performed to annotate PBMC subsets, resulting in 36 annotated PBMC
419 subsets from 144,324 cells across 27 samples.
420
421 **Publicly available scRNA data**. We obtained the publicly available dataset (Younger Adult
422 Cohort) from https://apps.allenimmunology.org/aifi/insights/dynamics-imm-health-
423 age/downloads/ which included samples from n=49 healthy young adults (25-35 years) and n=47
424 healthy older adults (55-65 years old)[32].
425

426 Our compositional analysis focused on the Flu Year 1 Day 0 baseline time point (Figures 2G, 2M,
427 2SC, 3D, 3J and 4F), which included n=1,539,411 cells. This dataset was filtered—by the original
428 authors—for doublets using Scrublet and further refined by removing cells with more than 5,000
429 genes or fewer than 200 genes, and mitochondrial threshold over 10%. The only additional
430 threshold we imposed for automated quality control was to remove genes detected in less than 3
431 cells, which left 32,069 genes detected. Cells were clustered and annotated using the same process
432 we applied to the Discovery Cohort. In Figure 2H we also include annotations following the above
433 procedure for Flu Year 2 Day 0 (Yr 2 V1) and Year 3 Stand-Alone (Yr 3 V1) timepoints. Finally,
434 in Figure 5E for the seroconverter we annotated cells from all available timepoints.
435
436 ***CMVerify* model training.** Baseline samples (Flu Year 1 Day 0) from the Younger Adult Cohort
437 (n=92)[32] were used for training *CMVerify*. To enable broad adoption, the input to our machine
438 learning pipeline was donor-level PBMC frequencies of the 72 cell types, labeled with the
439 standardized AIFI_L3 CellTypist model[32,57]. PBMC frequencies were scaled using
440 "StandardScaler" as provided by the scikit-learn package[58] to transform frequencies to have zero
441 mean and unit variance. A random forest classifier was used for model training. Optimal
442 parameters were selected using grid search and leave-one-out cross-validation over all possible
443 combinations of the following hyperparameters:
444

| *Parameter* | *Grid Search Options* ***(selected)*** |
|---|---|
| number of estimators | 50, **100**, 200 |
| max depth | **None**, 10, 20, 30 |
| minimum samples to split internal node | 2, **5**, 10 |
| minimum samples required at a leaf | 1, 2, **4** |
| number of features to consider when looking for best split | **square root (number of features)**, $\log_2$ (number of features) |
| sampling with replacement | True, **False** |

445 **Table 2**: Parameters tested and selected (in bold) for *CMVerify* training.

446
447 Optimal parameters (indicated in bold in Table 1) yielded the best cross validation score of 0.967.

448

449 Feature importances were calculated based on the mean decrease in impurity where impurity is a
450 measure of how mixed a node in a tree is with respect to target labels. Features (cell type
451 frequencies) that best reduce the impurity and more effectively split the CMV(+) and CMV(-)
452 samples will have the highest importance (Extended Data Fig. 6a, Table 2). Fourteen features were
453 selected by *CMVerify* based on feature importance scores (Table 2). To evaluate how the Cell
454 Typist labeling corresponds with our Discovery Cohort cell type labels (see Methods:scRNAseq
455 data analyses), we mapped the top five *CMVerify* selected Cell Typist features to our own cell type
456 annotations (Extended Data Fig. 6b).

457

| *Cell Type* | *Importance* |
|---|---|
| KLRF1- GZMB+ CD27- memory CD4 T cell (CD4$^+$ TEMRA) | 0.351594 |
| KLRF1- GZMB+ CD27- EM CD8 T cell (CD8$^+$ TEMRA) | 0.162693 |
| Adaptive NK cell (KLRC2$^+$ NK) | 0.137851 |
| KLRF1- effector Vd1 gdT (GZMB$^+$ γδ) | 0.092140 |
| KLRF1+ effector Vd1 gdT (GZMB$^+$ γδ) | 0.090094 |
| Proliferating NK cell | 0.033280 |
| GZMK- CD27+ EM CD8 T cell | 0.026493 |
| GZMK+ CD56dim NK cell | 0.025931 |
| GZMK- CD56dim NK cell | 0.024149 |
| KLRF1+ GZMB+ CD27- EM CD8 T cell | 0.018953 |
| ISG+ CD56dim NK cell | 0.010061 |
| GZMK+ memory CD4 Treg | 0.009613 |
| pDC | 0.009209 |
| CD8 MAIT | 0.007939 |

458 **Table 2**: Fourteen features were selected by *CMVerify* during training procedure, listed here in
459 descending order of importance (for predicting CMV serostatus).

460

461 **Test datasets for *CMVerify*.** Samples from the Independent Cohort 1: 3′ Older Adult Cohort have
462 been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession number

463 phs003259.v1.p1 and will be made publicly available upon acceptance of the study. We have
464 immediately made available cell type frequencies from this cohort in our Synapse repository along
465 with all code needed to reproduce our results in this cohort. Samples from Independent Cohort 2:
466 5′ Older Adult Cohort were from an ongoing in-house study, samples and cell type frequencies
467 from this cohort will be made available upon completion of the study. All PBMC samples were
468 processed using the same pipeline, as described above. For *CMVerify* testing, all cell type
469 annotations were performed using AIFI_L3 CellTypist model[32,57].

470 **Differential expression analyses.** Raw RNA counts were aggregated across cell types using the
471 function "ADPBulk" from adpbulk package (GitHub: noamteyssier, 2021, adpbulk). For each cell
472 type, samples containing fewer than 10 cells were excluded from the differential analysis. Raw
473 counts were first converted to a "DGEList" object, and transformed to normalized matrices using
474 the "calcNormFactors" and "cpm" functions in edgeR[59]. Genes with zero normalized counts across
475 all samples within a given cell type were removed. Differentially expressed gene (DEG) analysis
476 was performed using generalized linear models, employing the default normalization method
477 based on the trimmed mean of M-values (TMM) and controlling for batch, sex and site of sample
478 collection. P-values were adjusted using Benjamini–Hochberg correction for multiple comparisons
479 and genes with an adjusted P-value (FDR) less than 0.1 were selected as DEGs.

480

481 **Batch correction for gene expression.** Our baseline scRNAseq samples included two scRNAseq
482 libraries (previously published, n=11; and unpublished, n=16) that were sequenced separately.
483 Batch effects were corrected as follows: For each cell type of interest, raw RNA counts were
484 aggregated at the donor level (pseudobulk) as described above. Next, pseudobulk matrix values
485 were adjusted for batch using pyComBat[60]. Finally, the mean expression of genes of interest were
486 calculated at the donor level using batch-corrected pseudobulk matrix.

487

488 **Gene scoring.** Cytotoxicity score for each sample was calculated by computing the mean
489 expression of "Cytotoxic lymphocytes" (M9.1) gene set in $CD8^+$ T cells [33] (n=29 genes,
490 Supplementary Table 3). Innateness score was calculated by computing the mean expression of
491 innateness-related genes in $CD8^+$ T cells[34] (n=80 genes, Supplementary Table 3).

492

493 **Flow cytometry data generation and analyses.** The frequencies of the major cell populations in
494 PBMCs from CMV(+) and CMV(-) elderly donors were analyzed independently in the two
495 different cohorts recruited in Hartford, CT and Sudbury, Ontario, Canada. The samples collected
496 in Hartford, CT were stained as follows. Fluorescent-labeled antibody cocktails for cell-surface
497 staining panels were pre-mixed in BD Horizon Brilliant Stain Buffer (BD Biosciences) 10 minutes
498 before staining. Antibody cocktails were added over 100 μL aliquots of anticoagulated whole
499 blood in a 5 ml FACS tube within 60 minutes of blood collection. Samples were incubated for 15
500 minutes at room temperature then lysed and fixed with 2 ml of 1x FACS lysing solution (BD
501 Biosciences) for 8 minutes at room temperature. The lysed samples were washed twice to remove
502 the unbound antibodies, lysed RBCs, and platelets and finally resuspended in 250 μL of PBS to
503 which 50 μL of count beads suspension (Count Bright Absolute Counting Beads, Thermo Fisher)
504 were added for the detection of absolute cell counts. Cells were stained with fluorochrome-labeled
505 antibodies to the following surface markers: CD3 (clone UCHT1, BioLegend), CD4 (clone RPA-
506 T4, BD Biosciences), CD8 (clone SFC121Thy2D3, Beckman-Coulter), CD19 (clone HIB19, BD
507 Biosciences), CD14 (clone MSE2, BD Biosciences), CD197 (clone 150503, BD Biosciences), and
508 CD45RA (clone HI100, BD Biosciences), and CD127 (clone HIL-7R-M21, BD Biosciences). The
509 stained cells were acquired with LSR Fortessa X-20 (BD) and analyzed with FlowJo software
510 (BD).

Immunophenotyping for PBMCs in the Discovery Cohort (samples collected in Sudbury, Ontario, Canada) is detailed in the original publication [51]. For the Validation Cohort, PBMCs from participants were thawed and stained as previously described [26,52]. Following staining, PBMCs were analyzed on a CytoFLEX flow cytometer (Beckman Coulter, CA, USA). Analysis was performed using FlowJo v.10.1.5, following previously established protocols [26,52].

**Statistical tests.** In both the Discovery Cohort and the Validation Cohort, categorical participant characteristics including sex, race, previous cancer diagnosis, and medications used were compared using Fisher's exact test or the chi-square test, as applicable, using the SciPy package (GitHub: SciPy, 2011). Mann-Whitney U test was performed to compare CMV(+) and CMV(-) groups for continuous variables (e.g., age, BMI, and frailty), cell frequencies and gene set scores (e.g., Cytotoxicity score, Innateness score). Unless stated otherwise in figure legends, p-values in the main and supplementary figures are reported after multiple hypothesis testing correction (i.e. q) by Benjamini-Hochberg's procedure at FDR < 0.05.

## References and Notes


1. Nikolich-Žugich, J. The twilight of immunity: emerging concepts in aging of the immune system. *Nat Immunol* **19**, 10–19 (2018).
2. Terekhova, M. *et al.* Single-cell atlas of healthy human blood unveils age-related loss of NKG2C+GZMB−CD8+ memory T cells and accumulation of type 2 memory T cells. *Immunity* **56**, 2836-2854.e9 (2023).

535 3. Guo, J. *et al.* Aging and aging-related diseases: from molecular mechanisms to interventions and
536 treatments. *Sig Transduct Target Ther* **7**, 1–40 (2022).
537 4. Kennedy, B. K. *et al.* Geroscience: linking aging to chronic disease. *Cell* **159**, 709–713 (2014).
538 5. Mogilenko, D. A., Shchukina, I. & Artyomov, M. N. Immune ageing at single-cell resolution. *Nat Rev*
539 *Immunol* **22**, 484–498 (2022).
540 6. Hou, Y. *et al.* Insights into vaccines for elderly individuals: from the impacts of immunosenescence to
541 delivery strategies. *npj Vaccines* **9**, 1–13 (2024).
542 7. Bartleson, J. M. *et al.* SARS-CoV-2, COVID-19 and the aging immune system. *Nat Aging* **1**, 769–782
543 (2021).
544 8. Leng, S. X. & Margolick, J. B. Aging, sex, inflammation, frailty, and CMV and HIV infections. *Cellular*
545 *Immunology* **348**, 104024 (2020).
546 9. Brodin, P. *et al.* Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences.
547 *Cell* **160**, 37–47 (2015).
548 10. Khan, N. *et al.* Cytomegalovirus seropositivity drives the CD8 T cell repertoire toward greater clonality
549 in healthy elderly individuals. *J Immunol* **169**, 1984–1992 (2002).
550 11. Nehar-Belaid, D. *et al.* Baseline immune states (BIS) associated with vaccine responsiveness and
551 factors that shape the BIS. *Seminars in Immunology* **70**, 101842 (2023).
552 12. Márquez, E. J. *et al.* Sexual-dimorphism in human immune system aging. *Nat Commun* **11**, 751 (2020).
553 13. Cannon, M. J., Hyde, T. B. & Schmid, D. S. Review of cytomegalovirus shedding in bodily fluids and
554 relevance to congenital cytomegalovirus infection. *Reviews in Medical Virology* **21**, 240–255 (2011).
555 14. Schwartz, M. & Stern-Ginossar, N. Rethinking human cytomegalovirus latency reservoir. *Annals of the*
556 *New York Academy of Sciences* **1524**, 30–36 (2023).
557 15. Staras, S. A. S. *et al.* Seroprevalence of Cytomegalovirus Infection in the United States, 1988–1994.
558 *Clinical Infectious Diseases* **43**, 1143–1151 (2006).

559 16. Forte, E., Zhang, Z., Thorp, E. B. & Hummel, M. Cytomegalovirus Latency and Reactivation: An Intricate
560 Interplay With the Host Immune Response. *Front. Cell. Infect. Microbiol.* **10**, (2020).
561 17. Alanio, C. *et al.* Cytomegalovirus Latent Infection is Associated with an Increased Risk of COVID-19-
562 Related Hospitalization. *The Journal of Infectious Diseases* **226**, 463 (2022).
563 18. Kua, K. P., Chongmelaxme, B. & Lee, S. W. H. Association Between Cytomegalovirus Infection and
564 Tuberculosis Disease: A Systematic Review and Meta-Analysis of Epidemiological Studies. *The Journal*
565 *of Infectious Diseases* **227**, 471–482 (2023).
566 19. Milotay, G. *et al.* CMV serostatus is associated with improved survival and delayed toxicity onset
567 following anti-PD-1 checkpoint blockade. *Nat Med* 1–15 (2025) doi:10.1038/s41591-025-03647-1.
568 20. Pawelec, G., Derhovanessian, E., Larbi, A., Strindhall, J. & Wikby, A. Cytomegalovirus and human
569 immunosenescence. *Reviews in Medical Virology* **19**, 47–56 (2009).
570 21. Campos, C. *et al.* Effect of age and CMV on NK cell subpopulations. *Experimental Gerontology* **54**, 130–
571 137 (2014).
572 22. van der Heiden, M. *et al.* Differential effects of Cytomegalovirus carriage on the immune phenotype
573 of middle-aged males and females. *Sci Rep* **6**, 26892 (2016).
574 23. Wallace, D. L. *et al.* Human cytomegalovirus-specific CD8+ T-cell expansions contain long-lived cells
575 that retain functional capacity in both young and elderly subjects. *Immunology* **132**, 27–38 (2011).
576 24. Pitard, V. *et al.* Long-term expansion of effector/memory Vδ2− γδ T cells is a specific blood signature
577 of CMV infection. *Blood* **112**, 1317–1324 (2008).
578 25. Khairallah, C., Déchanet-Merville, J. & Capone, M. γδ T Cell-Mediated Immunity to Cytomegalovirus
579 Infection. *Front Immunol* **8**, 105 (2017).
580 26. Verschoor, C. P. *et al.* NK- and T-cell granzyme B and K expression correlates with age, CMV infection
581 and influenza vaccine-induced antibody titres in older adults. *Front Aging* **3**, 1098200 (2023).
582 27. Rückert, T., Lareau, C. A., Mashreghi, M.-F., Ludwig, L. S. & Romagnani, C. Clonal expansion and
583 epigenetic inheritance of long-lasting NK cell memory. *Nat Immunol* **23**, 1551–1563 (2022).

584 28. Schlums, H. *et al.* Cytomegalovirus Infection Drives Adaptive Epigenetic Diversification of NK Cells with
585 Altered Signaling and Effector Function. *Immunity* **42**, 443–456 (2015).
586 29. Watzl, C. Adaptive responses of innate lymphocytes. *Nat Immunol* **19**, 426–427 (2018).
587 30. Loeb, N. *et al.* Frailty Is Associated With Increased Hemagglutination-Inhibition Titers in a 4-Year
588 Randomized Trial Comparing Standard- and High-Dose Influenza Vaccination. *Open Forum Infectious*
589 *Diseases* **7**, ofaa148 (2020).
590 31. Deseke, M. & Prinz, I. Ligand recognition by the γδ TCR and discrimination between homeostasis and
591 stress conditions. *Cell Mol Immunol* **17**, 914–924 (2020).
592 32. Gong, Q. *et al.* Longitudinal Multi-omic Immune Profiling Reveals Age-Related Immune Cell Dynamics
593 in Healthy Adults. 2024.09.10.612119 Preprint at https://doi.org/10.1101/2024.09.10.612119 (2024).
594 33. Rinchai, D. *et al.* BloodGen3Module: blood transcriptional module repertoire analysis and visualization
595 using R. *Bioinformatics* **37**, 2382–2389 (2021).
596 34. Gutierrez-Arcelus, M. *et al.* Lymphocyte innateness defined by transcriptional states reflects a balance
597 between proliferation and effector functions. *Nat Commun* **10**, 687 (2019).
598 35. Caielli, S. *et al.* A CD4+ T cell population expanded in Lupus blood provides B cell help through IL10
599 and succinate. *Nat Med* **25**, 75–81 (2019).
600 36. Vivier, E., Tomasello, E., Baratin, M., Walzer, T. & Ugolini, S. Functions of natural killer cells. *Nat*
601 *Immunol* **9**, 503–510 (2008).
602 37. Rebuffet, L. *et al.* High-dimensional single-cell analysis of human natural killer cell heterogeneity. *Nat*
603 *Immunol* **25**, 1474–1488 (2024).
604 38. Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Computational Statistics* **2**, 433–459
605 (2010).
606 39. Zhang, W. *et al.* CMV Status Drives Distinct Trajectories of CD4+ T Cell Differentiation. *Front. Immunol.*
607 **12**, (2021).

608 40. Halenius, A., Gerke, C. & Hengel, H. Classical and non-classical MHC I molecule manipulation by human
609 cytomegalovirus: so many targets—but how many arrows in the quiver? *Cell Mol Immunol* **12**, 139–
610 153 (2015).
611 41. Hayday, A., Dechanet-Merville, J., Rossjohn, J. & Silva-Santos, B. Cancer immunotherapy by γδ T cells.
612 *Science* **386**, eabq7248 (2024).
613 42. Foley, B. *et al.* Cytomegalovirus reactivation after allogeneic transplantation promotes a lasting
614 increase in educated NKG2C+ natural killer cells with potent function. *Blood* **119**, 2665–2674 (2012).
615 43. Sun, J. C., Beilke, J. N. & Lanier, L. L. Adaptive immune features of natural killer cells. *Nature* **457**, 557–
616 561 (2009).
617 44. Barnett, B. E. *et al.* Cutting Edge: B Cell–Intrinsic T-bet Expression Is Required To Control Chronic Viral
618 Infection. *The Journal of Immunology* **197**, 1017–1022 (2016).
619 45. Jenks, S. A. *et al.* Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to
620 Pathogenic Responses in Systemic Lupus Erythematosus. *Immunity* **49**, 725-739.e6 (2018).
621 46. Dauby, N. *et al.* Primary Human Cytomegalovirus Infection Induces the Expansion of Virus-Specific
622 Activated and Atypical Memory B Cells. *The Journal of Infectious Diseases* **210**, 1275–1285 (2014).
623 47. Woodruff, M. C. *et al.* Dysregulated naive B cells and de novo autoreactivity in severe COVID-19.
624 *Nature* **611**, 139–147 (2022).
625 48. Zumaquero, E. *et al.* IFNγ induces epigenetic programming of human T-bethi B cells and promotes
626 TLR7/8 and IL-21 induced differentiation. *eLife* **8**, e41641 (2019).
627 49. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep
628 learning model for cytometry data. *Proceedings of the National Academy of Sciences* **117**, 21373–
629 21380 (2020).
630 50. Zhou, K. *et al.* Applying T-classifier, binary classifiers, upon high-throughput TCR sequencing output to
631 identify cytomegalovirus exposure history. *Sci Rep* **13**, 5024 (2023).

632 51. Ravichandran, S. *et al.* Distinct baseline immune characteristics associated with responses to
633 conjugated and unconjugated pneumococcal polysaccharide vaccines in older adults. *Nat Immunol*
634 **25**, 316–329 (2024).
635 52. Picard, E. *et al.* Markers of systemic inflammation are positively associated with influenza vaccine
636 antibody responses with a possible role for ILT2(+)CD57(+) NK-cells. *Immun Ageing* **19**, 26 (2022).
637 53. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell
638 RNA sequencing data. *GigaScience* **9**, giaa151 (2020).
639 54. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-
640 Cell Transcriptomic Data. *Cell Syst* **8**, 281-291.e9 (2019).
641 55. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis.
642 *Genome Biology* **19**, 15 (2018).
643 56. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat*
644 *Methods* **16**, 1289–1296 (2019).
645 57. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in
646 humans. *Science* **376**, eabl5197 (2022).
647 58. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*
648 **12**, 2825–2830 (2011).
649 59. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential
650 expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
651 60. Behdenna, A. *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular
652 data using empirical Bayes methods. *BMC Bioinformatics* **24**, 459 (2023).

653

654 **Acknowledgments:**

655 We gratefully acknowledge the contribution of the Single Cell Biology service, Genome
656 Technologies service, and Cyberinfrastructure high performance computing resources at

The Jackson Laboratory (JAX) for assistance with the work described herein. We thank Olivia Bart for help with dbGAP data upload. We thank members of the Ucar lab for critical feedback during the progress of the study.

**Funding:**

This study was made possible by generous financial support of the National Institutes of Health (NIH) grants under award number:

- UH2 AG056925 (to G.K, CCE, J.B., D.U.)
- U01 AI165452 (to D.U., G.K.)
- P30AG067988 Older Americans Independence Pepper Center (to G.A.K., R.M., D.U.)
- P30AG028716 Duke Older Americans Independence Center (to K.E.S, C.C.E, H.E.W).
- NIH grant U19 AI168631 (to A.G.-S.)
- CRIPT (Center for Research on Influenza Pathogenesis and Transmission), a as NIAID-funded Center of Excellence for Influenza Research and Response (CEIRR, contract # 75N93021C00014) (to A.G.S.)
- 1T32AG062409-01A1 (to L.T.)

Other partial sources of funding include:

- JAX Cancer Center (P30 CA034196)
- American Heart Association Predoctoral Fellowship (24PRE1186316) (to T.G.)

**Author contributions:**

Conceptualization: JB, GK, DU

Methodology: JB, GK, CPV, DU

Investigation: TG, CPV, LT, AT, DN-B, RM, EP, KES, CCE, HEW, SP, AGS, CK

Visualization: TG, LT, GE

Funding acquisition: JB, GK, CCE, DU

Project administration: TG, DU

Supervision: JB, GK, DU

Writing – original draft: TG, CPV, DU, LT

Writing – review & editing: TG, CPV, DU, JB, GK, LT, AT, DN-B, RM, CK, EP, KES, CCE, HEW, SP, AGS

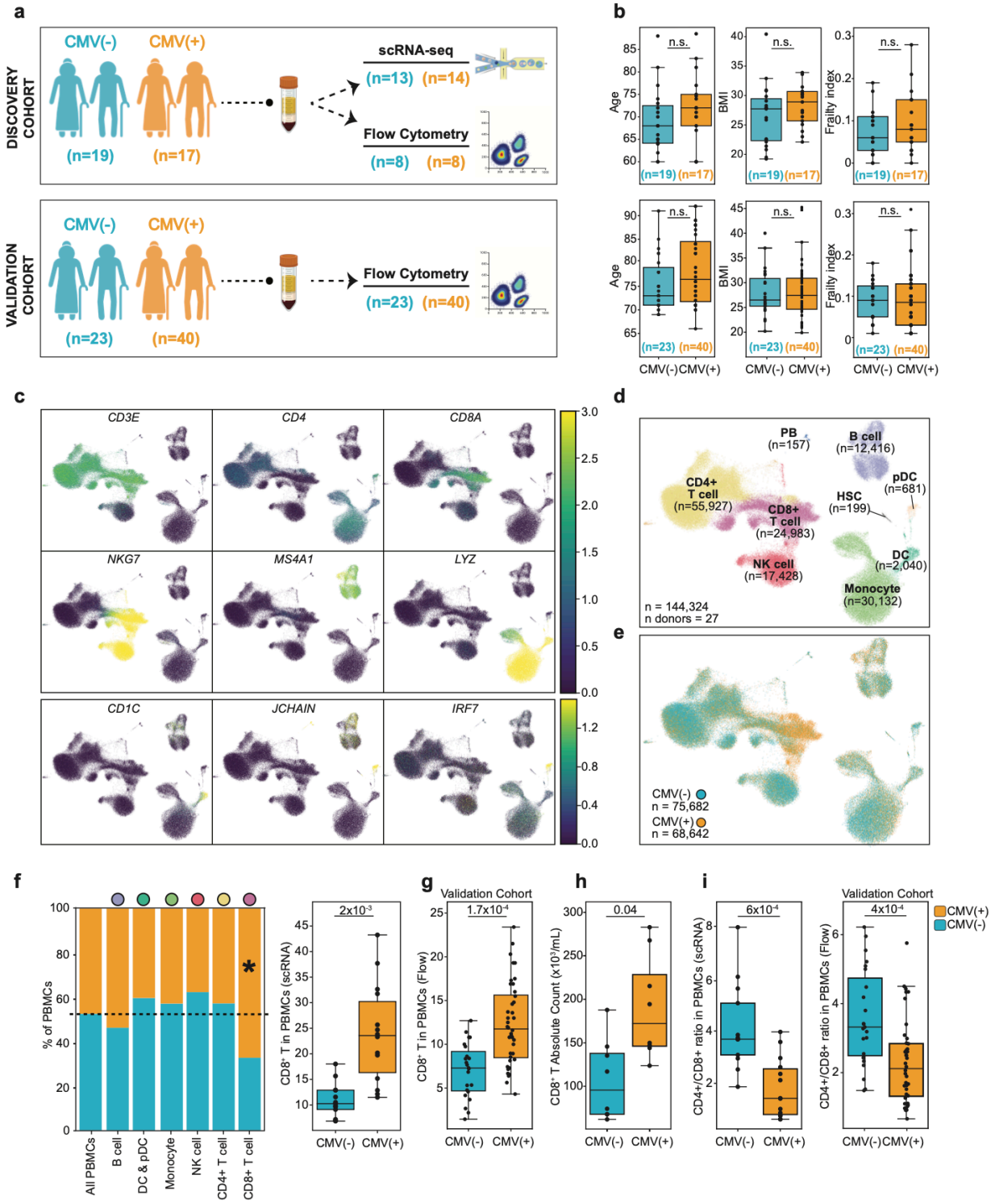
**Competing interests:**

The A.G.-S. laboratory has received research support from Avimex, Dynavax, Pharmamar, 7Hills Pharma, ImmunityBio and Accurius, outside of the reported work. A.G.-S. has consulting agreements for the following companies involving cash and/or stock: Castlevax, Amovir, Vivaldi Biosciences, Contrafect, 7Hills Pharma, Avimex, Pagoda, Accurius, Esperovax, Applied Biological Laboratories, Pharmamar, CureLab Oncology, CureLab Veterinary, Synairgen, Paratus, Pfizer, Virofend and Prosetta, outside of the reported work. A.G.-S. has been an invited speaker in meeting events organized by Seqirus, Janssen, Abbott, Astrazeneca and Novavax. A.G.-S. is inventor on patents and patent applications on the use of antivirals and vaccines for the treatment and prevention of virus infections and cancer, owned by the Icahn School of Medicine at Mount Sinai, New York, outside of the reported work.

**Data and materials availability:**

- Processed data and figure code is available on Synapse (under accession code https://doi.org/10.7303/syn68150215). Data are also browsable online. Raw read data will be made available on dbGAP upon acceptance of the manuscript (under accession code phs003598.v1.p1). *CMVerify* is available as a Python package: https://pypi.org/project/CMVerify/; open source code hosted at https://github.com/UcarLab/CMVerify.

714 **FIGURE 1**

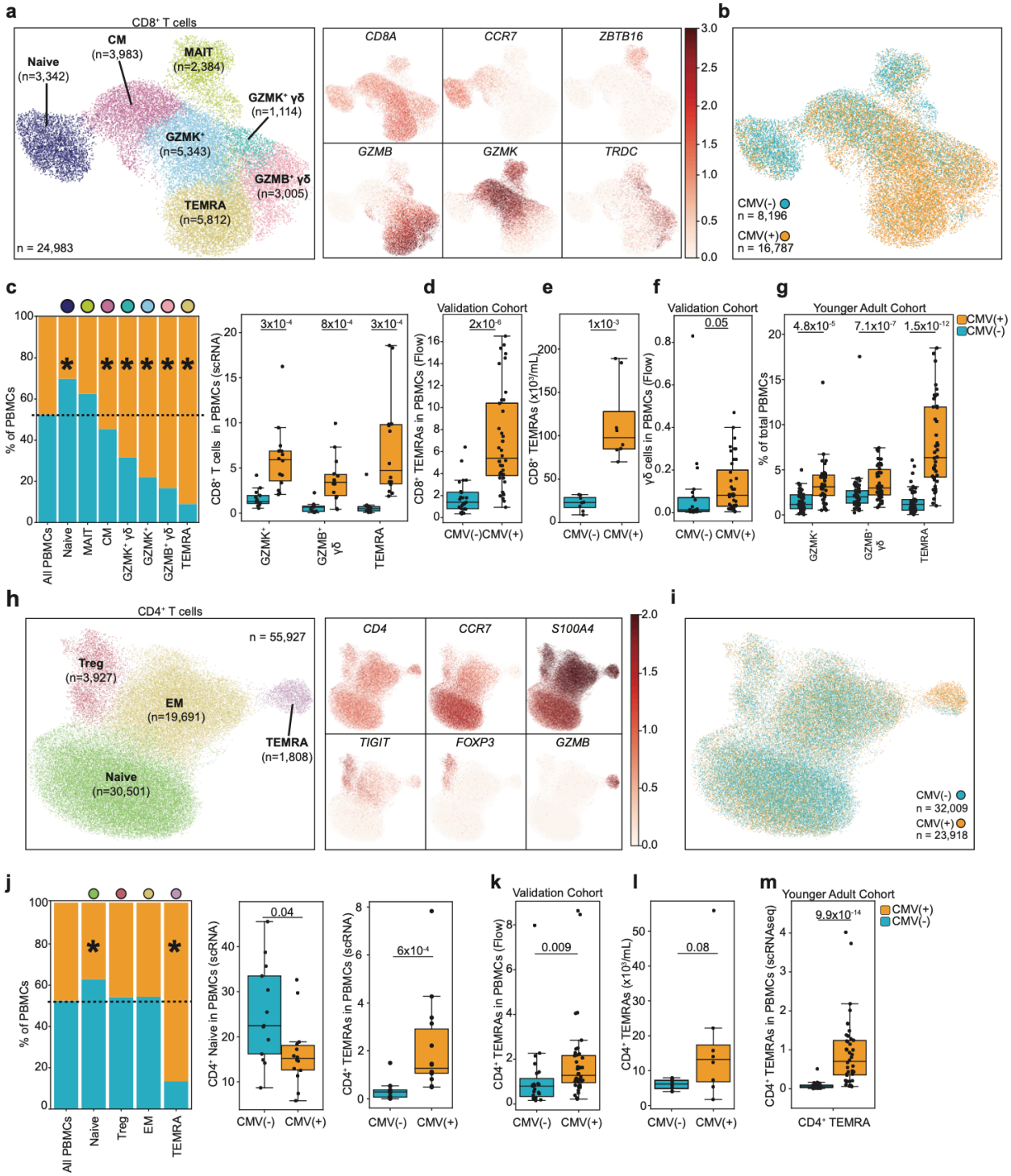



715
716 **Fig. 1| Study design and CD8⁺ T cell subset expansion associated with CMV. a,** Schematic
717 representation of the study design with two cohorts. *Top panel*: Discovery Cohort included n=19
718 CMV(-) and n=17 CMV(+) older adults. *Lower panel*: Validation Cohort included n=23 CMV(-)
719 and n=40 CMV(+) older adults. **b,** Distribution of age, BMI, and frailty index for CMV(-) and
720 CMV(+) groups in Discovery Cohort (*top panel*) and Validation Cohort (*bottom panel*). n.s.
721 indicates not significant. **c,** Uniform manifold approximation and projection (UMAP)

representing PBMCs from Discovery Cohort showing expression of subset-specific marker genes. **d,** 9 subsets (n=144,324) colored by immune cell annotations. NK, natural killer; pDC, plasmacytoid dendritic cell; PB, plasmablast; HSC, hematopoietic stem-cell. **e,** UMAP colored for CMV serostatus. **f,** *Left*: Immune cell frequencies in CMV(-) and CMV(+). Cell types with significant frequency differences between CMV(-) and CMV(+) groups are marked with an asterisk (*; P-adjusted [Benjamini-Hochberg] < 0.05). *Right*: Individual-level quantification of $CD8^+$ T cells within PBMCs in CMV(-) (n=13) and CMV(+) (n=14). **g,** $CD8^+$ T cell frequency within PBMCs quantified by flow cytometry in Validation Cohort: CMV(-) (n=23) and CMV(+) (n=40). **h,** Absolute count of $CD8^+$ T cells quantified by flow cytometry in Discovery Cohort: CMV(-) (n=8) and CMV(+) (n=8). **i,** *Left*: $CD4^+$ to $CD8^+$ T cell frequency ratios quantified *via* scRNAseq in Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). *Right:* $CD4^+$ to $CD8^+$ T cell frequency ratios quantified by flow cytometry in Validation Cohort: 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. Mann-Whitney-U test (two-sided) was used to compare cell counts between CMV(-) and CMV(+) (**b, f, g, h, i**). All *P*-values are shown after adjusting for multiple hypothesis testing using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

741 **FIGURE 2**

742




743

744 **Fig. 2| Expansion of memory CD8⁺ T, γδ T, and CD4⁺ TEMRA cell subsets associated with**
745 **CMV positivity. a,** *Left:* UMAP representing 7 subsets from CD8⁺ T cells (n= 24,983). CM,
746 central memory; MAIT, mucosal associated invariant T; TEMRA, terminally differentiated
747 effector memory. *Right:* UMAPs of CD8⁺ T cells showing expression of select marker genes. **b,**
748 UMAP color coded for CMV serostatus. **c,** *Left:* CD8⁺ T cell subset frequencies in CMV(-) and
749 CMV(+). Cell types with significant frequency differences between CMV(-) and CMV(+)

750 groups are marked with an asterisk (*; P-adjusted [Benjamini-Hochberg] < 0.05). *Right:* Select
751 $CD8^+$ T cell frequencies within PBMCs quantified *via* scRNAseq in Discovery Cohort: in
752 CMV(-) (n=13) and CMV(+) (n=14).**d,** $CD8^+$ TEMRA cell frequencies within PBMCs
753 quantified by flow cytometry in Validation Cohort: CMV(-) (n=23) and CMV(+) (n=40). **e,**
754 Absolute count of $CD8^+$ TEMRA cells quantified by flow cytometry in Discovery Cohort:
755 CMV(-) (n=8) and CMV(+) (n=8). **f,** γδ T cell frequencies within PBMCs quantified by flow
756 cytometry in Validation Cohort: CMV(-) (n=23) and CMV(+) (n=40). **g,** $GZMK^+$ $CD8^+$, $GZMB^+$
757 γδ, and $CD8^+$ TEMRA T cell frequencies within PBMCs quantified *via* scRNAseq in Younger
758 Adult Cohort: CMV(+) (n=42) and CMV(-) (n=50). **h,** *Left:* UMAP representing 4 subsets from
759 $CD4^+$ T cells (n= 55,927). Treg, $CD4^+$ T-regulatory; EM, effector memory; TEMRA, terminally
760 differentiated effector memory. *Right:* UMAPs of $CD4^+$ T cells showing expression of select
761 marker genes. **i,** UMAP color coded for CMV serostatus. **j,** *Left:* $CD4^+$ T cell subset frequencies
762 in CMV(-) and CMV(+). Cell types with significant frequency differences between CMV(-) and
763 CMV(+) groups are marked with an asterisk (*; P-adjusted [Benjamini-Hochberg] < 0.05).
764 *Middle:* Naïve $CD4^+$ T cell frequencies within PBMCs. *Right:* $CD4^+$ TEMRA cell frequencies
765 within PBMCs. All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and
766 CMV(+) (n=14). **k,** $CD4^+$ TEMRA frequencies within PBMCs quantified by flow cytometry in
767 Validation Cohort: CMV(-) (n=23) and CMV(+) (n=40). **l,** Absolute count of $CD4^+$ TEMRAs
768 quantified by flow cytometry in Discovery Cohort: CMV(-) (n=8) and CMV(+) (n=8). **m,** $CD4^+$
769 TEMRA T cell frequencies within PBMCs quantified *via* scRNAseq in Younger Adult Cohort:
770 CMV(+) (n=42) and CMV(-) (n=50). Box plots display the median and IQR (25–75%), with
771 whiskers representing the upper and lower quartiles ±1.5× IQR. Each dot represents an
772 individual sample. Mann-Whitney-U test (two-sided) was used to compare cell counts between
773 CMV(-) and CMV(+) (**c, d, e, f, g, j, k. l, m**). All *P*-values are shown after adjusting for multiple
774 hypothesis testing using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.
775
776

777 **FIGURE 3**

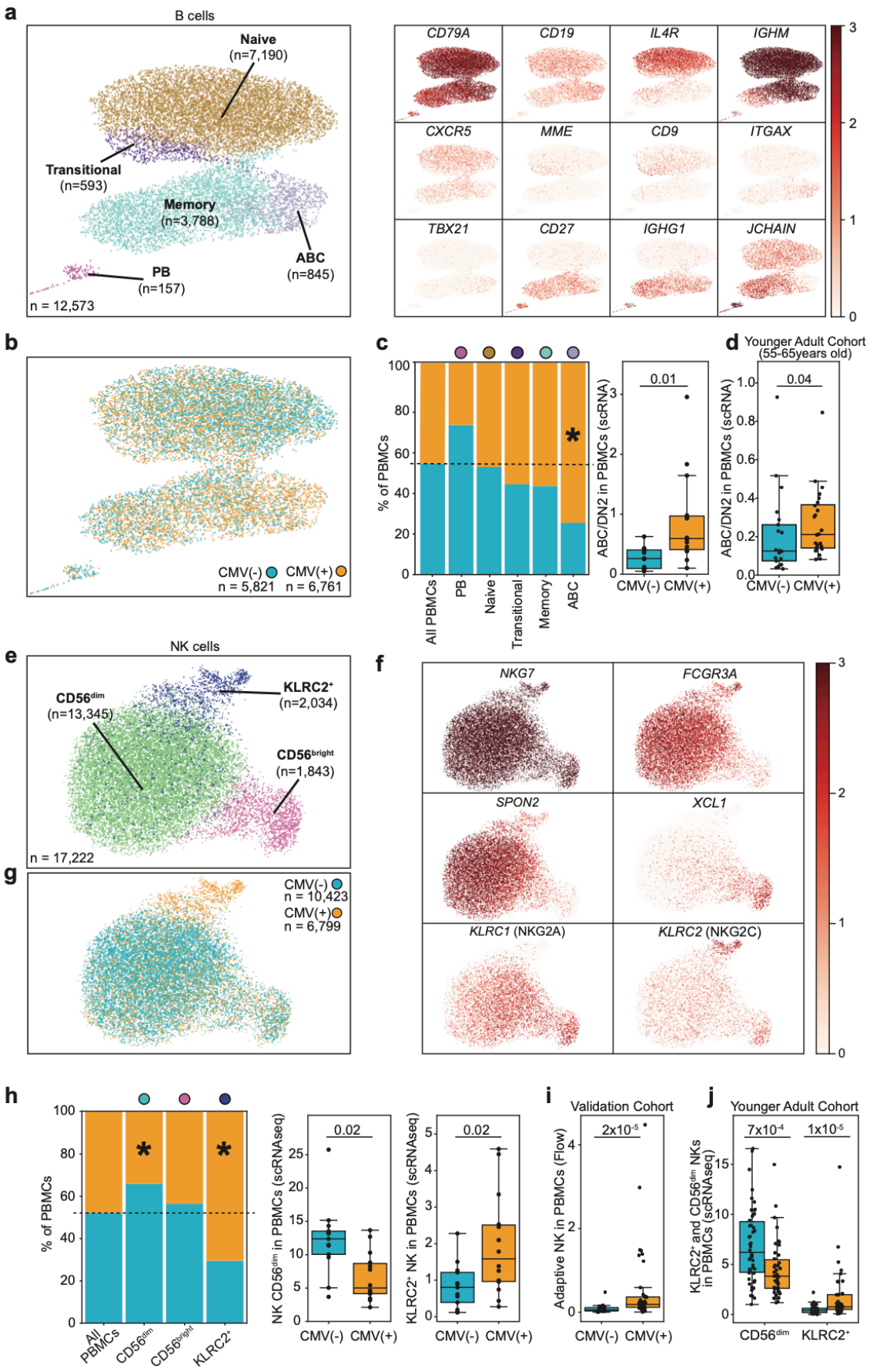
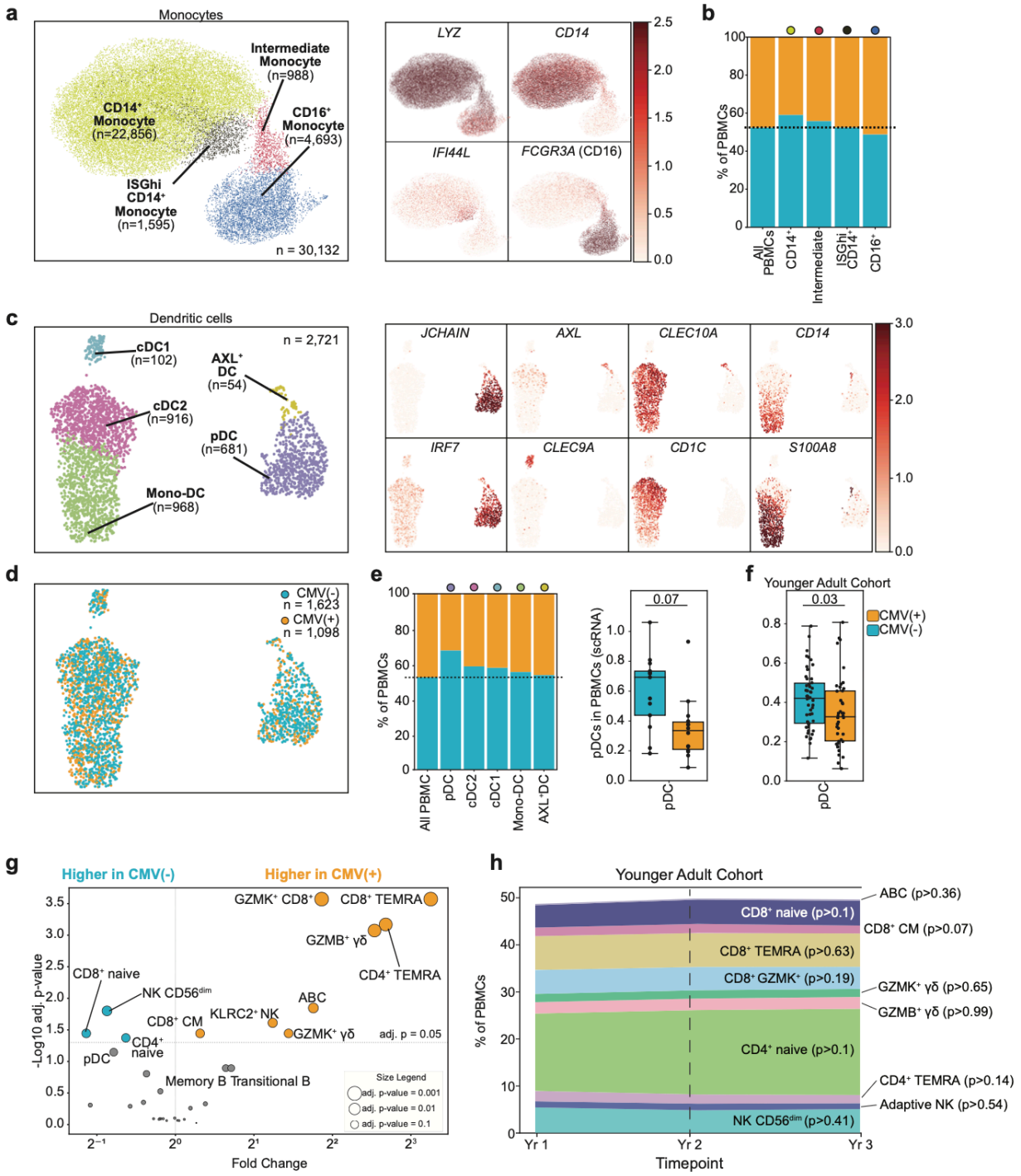

a
B cells
Naive
(n=7,190)
Transitional
(n=593)
Memory
(n=3,788)
ABC
(n=845)
PB
(n=157)
n = 12,573
CD79A
CD19
IL4R
IGHM
CXCR5
MME
CD9
ITGAX
TBX21
CD27
IGHG1
JCHAIN
b
CMV(-) n = 5,821
CMV(+) n = 6,761
c
% of PBMCs
All PBMCs
PB
Naive
Transitional
Memory
ABC
ABC/DN2 in PBMCs (scRNA)
0.01
CMV(-)
CMV(+)
d
Younger Adult Cohort
(55-65years old)
0.04
e
NK cells
CD56dim
(n=13,345)
KLRC2+
(n=2,034)
CD56bright
(n=1,843)
n = 17,222
f
NKG7
FCGR3A
SPON2
XCL1
KLRC1 (NKG2A)
KLRC2 (NKG2C)
g
CMV(-)
n = 10,423
CMV(+)
n = 6,799
h
% of PBMCs
All PBMCs
CD56dim
CD56bright
KLRC2+
NK CD56dim in PBMCs (scRNAseq)
0.02
KLRC2+ NK in PBMCs (scRNAseq)
0.02
i
Validation Cohort
Adaptive NK in PBMCs (Flow)
2x10-5
j
Younger Adult Cohort
KLRC2+ and CD56dim NKs in PBMCs (scRNAseq)
7x10-4
1x10-5
CD56dim
KLRC2+

**Fig. 3| ABC and adaptive NK cells were expanded while CD56$^{dim}$NK cells were reduced in CMV(+) older adults. a,** *Left:* UMAP representing 5 B cell subsets (n= 12,573). PB, Plasmablast; ABC, atypical B cell. *Right:* UMAPs of B cells showing expression of select marker genes. **b,** UMAP color coded for CMV serostatus. **c,** *Left:* B cell frequencies in CMV(-) and CMV(+). Cell types with significant frequency differences between CMV(-) and CMV(+) groups are marked with an asterisk (*; P-adjusted [Benjamini-Hochberg] < 0.05). *Right:* ABC frequencies within PBMCs All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **d,** ABC frequencies within PBMCs quantified *via* scRNAseq in Younger Adult Cohort (55-65 age group): CMV(+) (n=24) and CMV(-) (n=20). **e,** UMAP representing 3 NK cell subsets (n= 17,222). **f,** UMAPs of NK cells showing expression of select marker genes. **g,** UMAP color coded for CMV serostatus. **h,** *Left:* NK cell frequencies in CMV(-) and CMV(+). Cell types with significant frequency differences between CMV(-) and CMV(+) groups are marked with an asterisk (*; P-adjusted [Benjamini-Hochberg] < 0.05). *Middle:* NK CD56$^{dim}$ frequencies within PBMCs. *Right:* KLRC2$^{+}$ NK cell frequencies within PBMCs. All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **i,** Adaptive NK frequencies within PBMCs quantified by flow cytometry in Validation Cohort: CMV(-) (n=23) and CMV(+) (n=40). **j,** NK CD56$^{dim}$ and KLRC2$^{+}$ NK cell frequencies quantified *via* scRNAseq in Younger Adult Cohort: 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. For Mann-Whitney-U test (two-sided) was used to compare cell counts between CMV(-) and CMV(+) (**c**, **d, h, i, j**). All *P*-values are shown after adjusting for multiple hypothesis testing using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

804 **FIGURE 4**

805

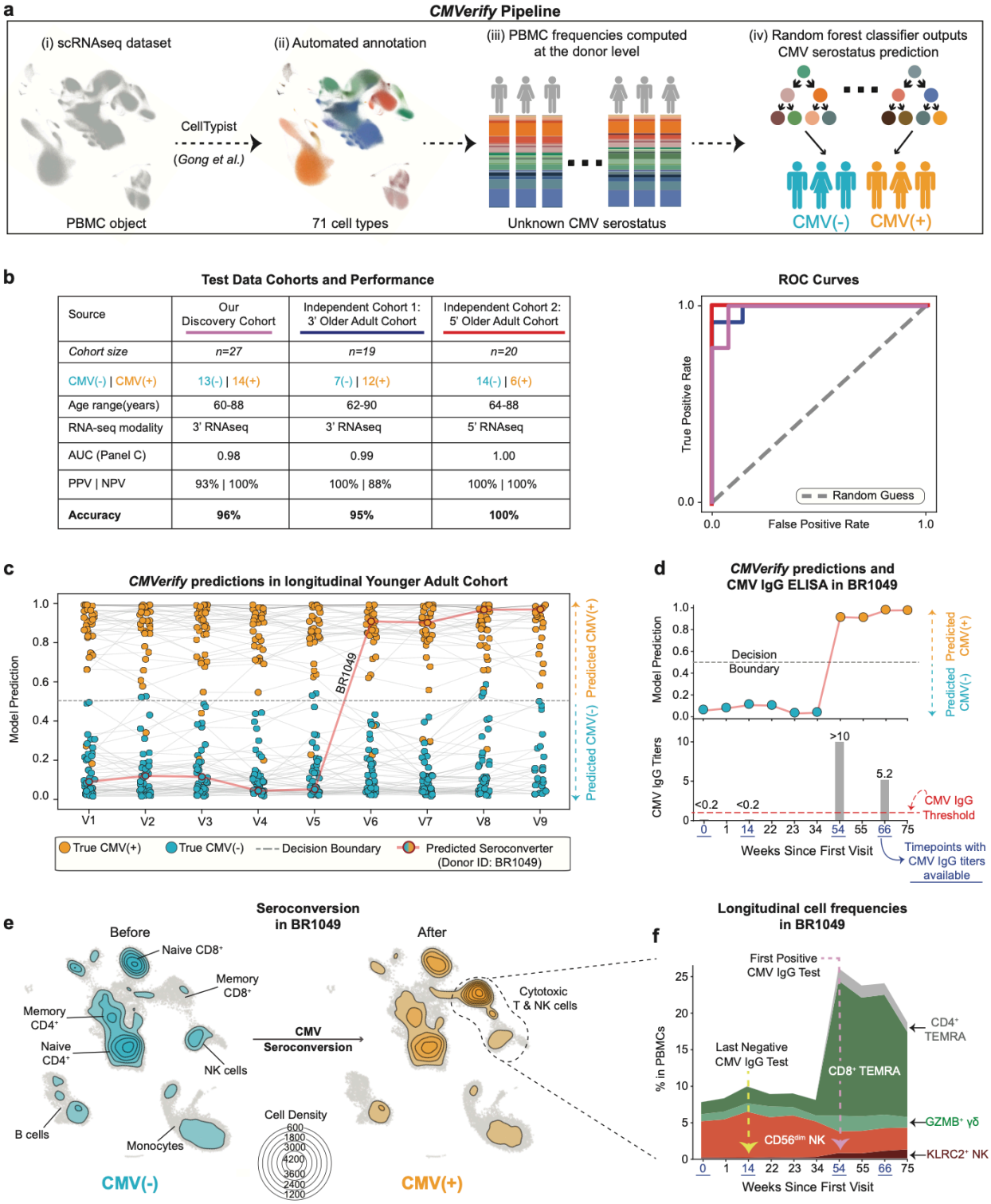



806

807 **Fig. 4| Monocytes and dendritic cells were comparable between CMV(+) and CMV(-) older**
808 **adults. a,** *Left:* UMAP representing 4 monocyte subsets (n= 30,132). ISGhi, interferon
809 stimulated gene-high expressing subset. *Right:* UMAPs of monocytes showing expression of
810 select marker genes. **b,** Monocyte frequencies in CMV(-) and CMV(+). **c,** *Left:* UMAP
811 representing 5 DC subsets (n= 2,721). DC, dendritic cell; Mono-DC, monocyte-derived dendritic
812 cell; cDC1, conventional type 1 dendritic cells; cDC2, Conventional type 2 dendritic cells; pDC,
813 plasmacytoid dendritic cell. *Right:* UMAPs of DCs showing expression of select marker genes.

**d,** UMAP color coded for CMV serostatus. **e,** *Left:* DC frequencies in CMV(-) and CMV(+). *Right:* pDC frequencies within PBMCs. All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **f,** pDC frequencies quantified *via* scRNAseq in Younger Adult Cohort: CMV(+) (n=42) and CMV(-) (n=50). **g,** Summary volcano plot showing PBMC compositional alterations associated with CMV serostatus in Discovery Cohort. Size of bubbles represent p-adjusted value (-Log scale). **h,** Average frequencies of CMV-associated cell types over a three-year period in CMV(+) individuals from Younger Adult Cohort (n=22). For **e,** and **f,** 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 (**e, f, h**). All *P*-values are shown after adjusting for multiple hypothesis testing using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

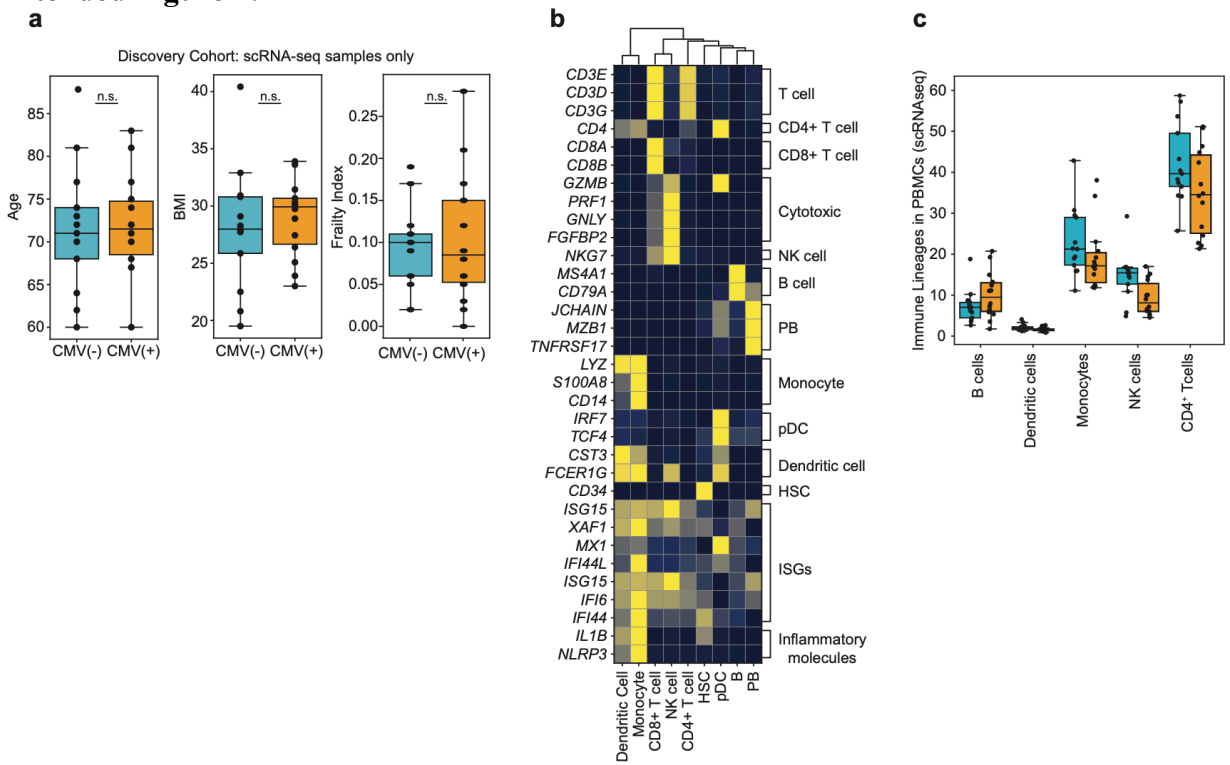
827 **FIGURE 5**




828
829
830 **Fig. 5| *CMVerify*—a novel tool that accurately predicts CMV serostatus from scRNAseq**
831 **data. a,** Schematic representation of *CMVerify* Pipeline: (i) *CMVerify* accepts a PBMC
832 scRNAseq dataset as input, (ii) applies CellTypist to automatically annotate the dataset, (iii)
833 computes PBMC frequencies at the donor level, and (iv) uses a random forest classifier to predict

CMV serostatus. **b,** *Left*: Table summarizing *CMVerify's* performance in three independent cohorts spanning different age groups and scRNAseq platforms. PPV, positive predictive value; NPV, negative predictive value; *Right*: Receiver Operating Characteristic (ROC) curves showing performance of *CMVerify* in the independent cohorts. **c,** *CMVerify* prediction results for the longitudinal Younger Adult Cohort. Y-axis represents model predicted probability, each bubble is a unique sample corresponding with a donor timepoint 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 true CMV serostatus (determined via CMV IgG ELISA). Red line connects model predicted probabilities for the seroconverter (donor ID: BR1049) across all timepoints. **d,** *Top*: *CMVerify* prediction probabilities across all timepoints in donor BR1049 identified in panel **c**. *Bottom*: CMV IgG titers for donor BR1049. Underlined timepoints (in blue) indicate presence of CMV IgG data at those timepoints. The red dotted line indicates CMV IgG titer threshold above which a sample is considered CMV(+). **e**, Contour UMAP plots show PBMC densities before and after CMV seroconversion in donor BR1049. The UMAP labeled 'Before' includes timepoints at weeks 14, 22, 23, and 34, while the UMAP labeled 'After' includes weeks 54, 55, 66, and 75. The Cell Density legend indicates the number of cells within each contour level shown in the UMAPs. **f**, CMV-driven changes in Cytotoxic T and NK cell frequencies over time in donor BR1049.

855 **Extended Figure 1:**



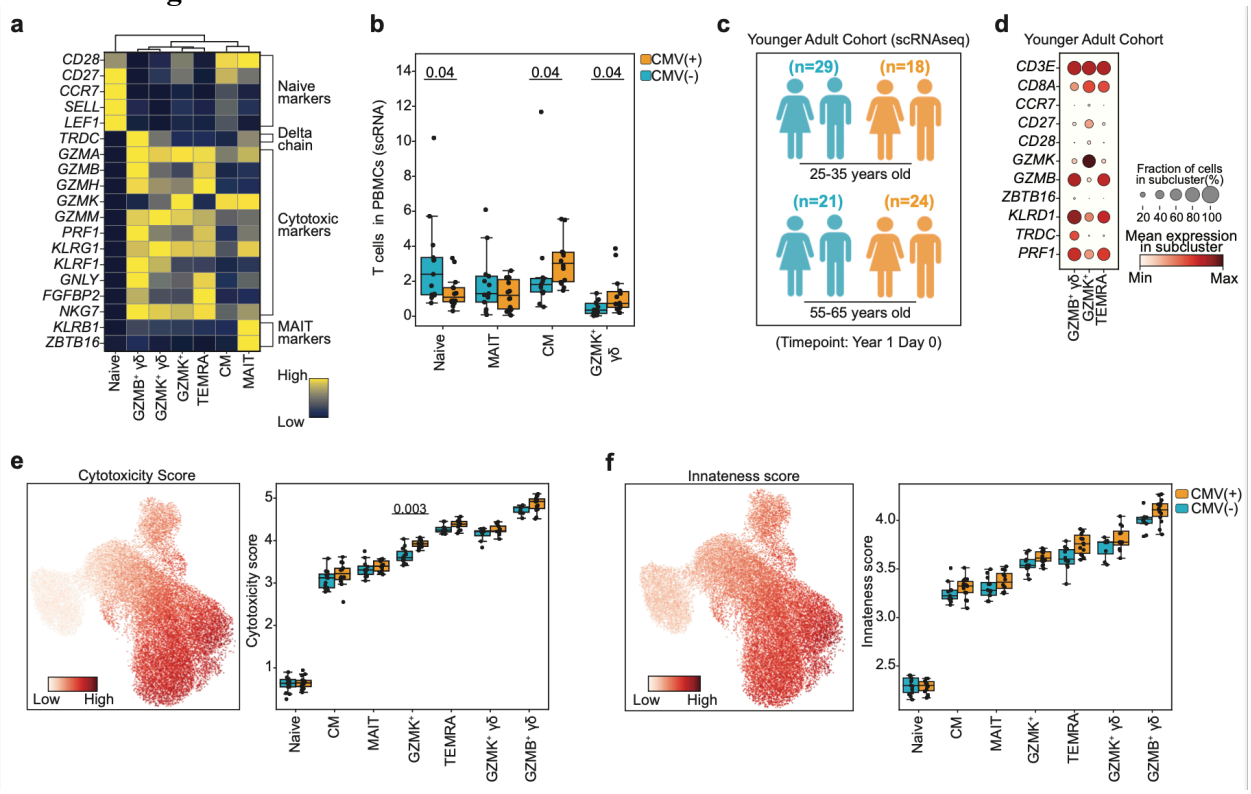

856

857 **Extended Data Fig. 1: PBMC-level annotations and analysis. a,** Distribution of age, BMI, and
858 frailty index in Discovery Cohort (scRNA-seq samples only n=27). n.s. indicates not significant.
859 **b,** Scaled expression of select marker genes. **c,** PBMC subset frequency ratios quantified *via*
860 scRNAseq in CMV(-) (n=13) and CMV(+) (n=14) in Discovery Cohort. Box plots display the
861 median and IQR (25–75%), with whiskers representing the upper and lower quartiles ±1.5× IQR.
862 Each dot represents an individual sample. For a and c, Mann-Whitney-U test (two-sided) was
863 used to calculate *P*-values between CMV(-) and CMV(+). **c**, none of the measurements were
864 significant after adjusting the *P*-values for multiple hypothesis testing, using Benjamini-
865 Hochberg at false discovery rate (FDR) < 0.05.

866

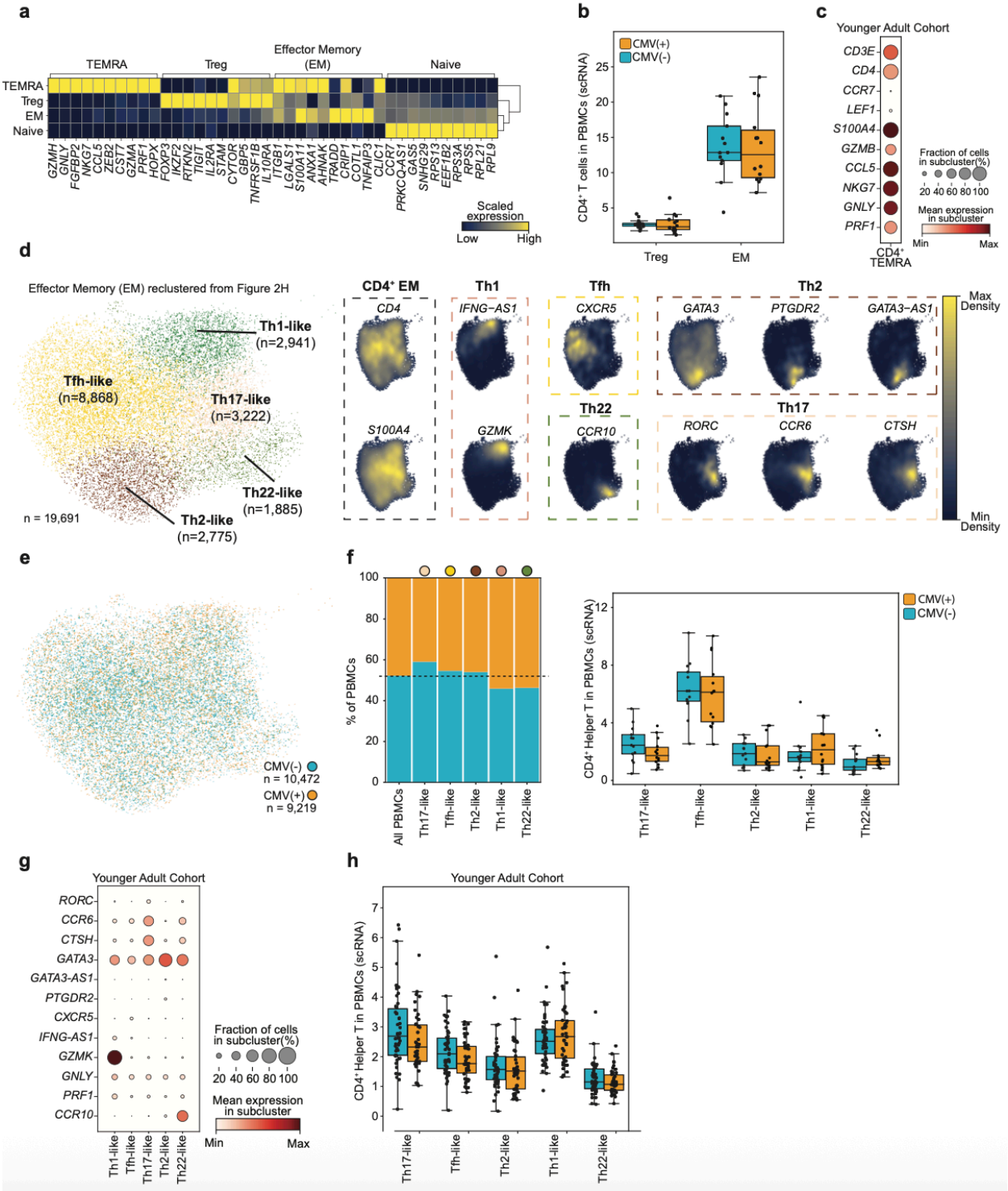
867

**Extended Figure 2:**




**Extended Data Fig. 2: 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* scRNAseq in Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **c,** Schematic representation of Younger Adult Cohort including two age groups (25-35 years old; 55-65 years old). **d,** Scaled expression of select marker for GZMB$^+$ γδ T, GZMK$^+$ T, and CD8$^+$ TEMRA T cells in Younger Adult Cohort. **e,** *Left:* UMAP showing Cytotoxicity score in CD8$^+$ T cell subsets. *Right:* individual-level cytotoxicity score in CD8$^+$ T cells subsets. **f,** *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, e, f**). All *P*-values are shown after adjusting for multiple hypothesis testing using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.

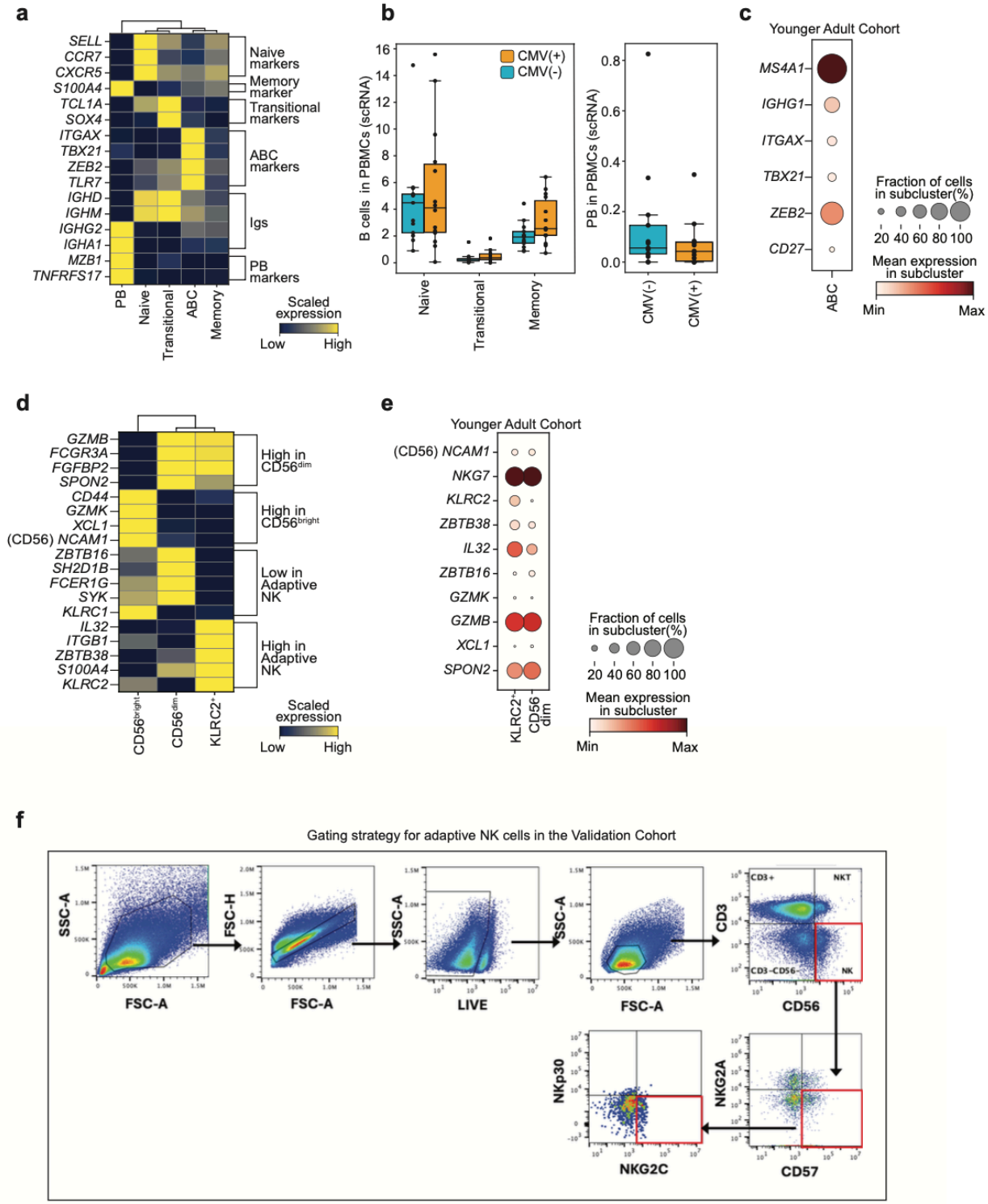
886
887 **Extended Figure 3:**




888
889 **Extended Data Fig. 3: CD4+ T helper subsets were comparable between CMV(-) and**
890 **CMV(+). a,** Scaled expression of top 10 marker genes in CD4+ naïve, CD4+ Treg, CD4+ EM,
891 and CD4+ TEMRA subsets. **b,** CD4+ Treg and CD4+ TEMRA T cell frequencies within PBMCs
892 quantified *via* scRNAseq in Discovery Cohort: CMV(-) (n = 13) and CMV(+) (n = 14). **c,** Scaled
893 expression of select marker genes for CD4+ TEMRA T cells in Younger Adult Cohort. **d,** *Left:*

894 UMAP representing re-clustered $CD4^+$ EM compartment and 5 resulting T helper subsets. Tfh-
895 like, T follicular helper-like; Th2-like, T helper 2-like; Th22-like, T helper 22-like; Th17-like, T
896 helper 17-like; Th1-like, T helper 1-like. *Right:* Gene density plot showing expression levels of
897 key marker genes, calculated using Nebulosa. Each gene density is grouped according to known
898 T helper marker genes. **e,** UMAP color coded for CMV serostatus. **f,** *Left:* $CD4^+$ T helper
899 frequencies in CMV(-) and CMV(+). *Right:* T helper cell frequencies within PBMCs quantified
900 *via* scRNAseq in Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **g,** Scaled expression
901 of select marker genes for T helper subsets in Younger Adult Cohort. **h,** T helper cell frequencies
902 within PBMCs quantified *via* scRNAseq in Younger Adult Cohort: CMV(+) (n=42) and CMV(-)
903 (n=50). Box plots display the median and IQR (25–75%), with whiskers representing the upper
904 and lower quartiles ±1.5× IQR. Each dot represents an individual sample. Mann-Whitney-U test
905 (two-sided) was used to compare cell counts between CMV(-) and CMV(+) (**b**, **e, h,**). None of
906 the measurements were significant after adjusting the *P*-values for multiple hypothesis testing,
907 using Benjamini-Hochberg at false discovery rate (FDR) < 0.05.
908
909

**Extended Figure 4:**

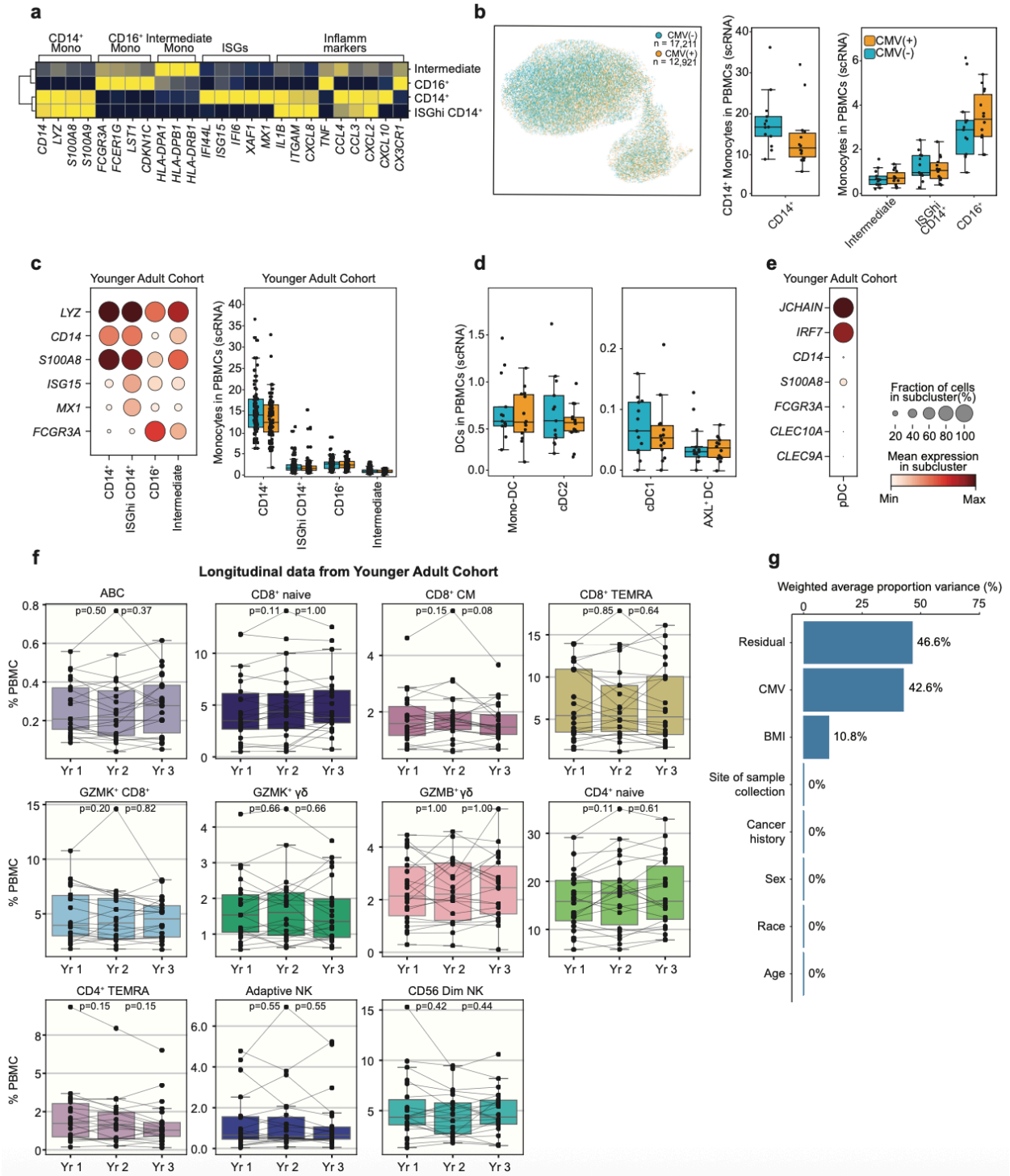



**Extended Data Fig. 4: B and NK cell annotations. a,** Scaled expression of select marker genes in B cell subsets. **b,** *Left:* naïve, transitional, and memory B cell frequencies within PBMCs. *Right:* PB cell frequencies within PBMCs. All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **c,** Scaled expression of select marker genes for ABCs in Younger Adult Cohort. **d,** Scaled expression of select marker genes in NK cell subsets.

**e,** Scaled expression of select marker genes for KLRC2$^{+}$ and CD56$^{dim}$ NK cells in Younger Adult Cohort. **f,** Gating strategy for adaptive NK cells in the Validation Cohort 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**). 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.

928 **Extended Figure 5:**

929

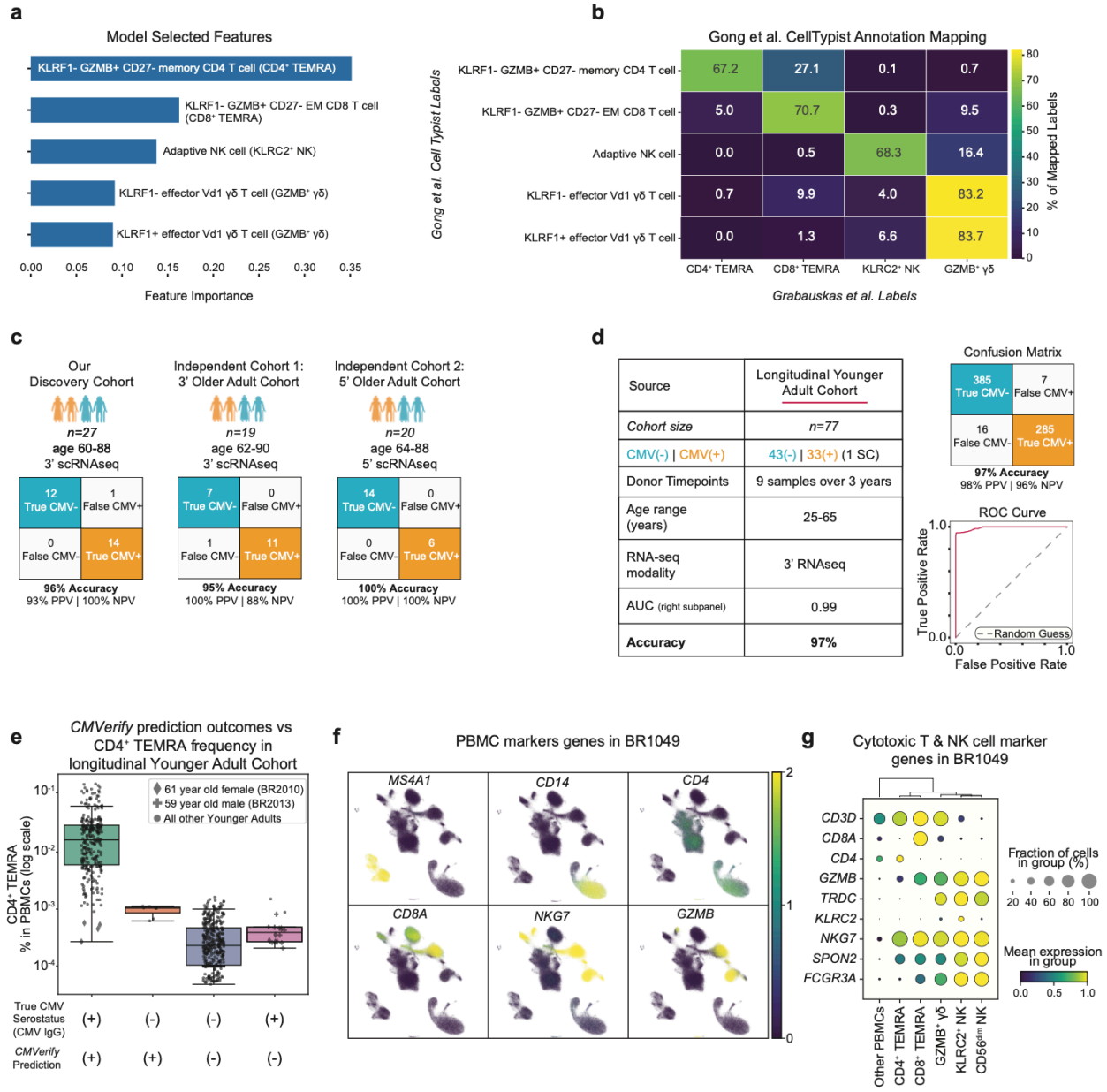



930

931 **Extended Data Fig. 5: Monocyte subset annotations and temporal stability of PBMC**
932 **frequencies associated with CMV. a,** Scaled expression of select marker genes in monocyte
933 subsets. **b,** *Left:* UMAP color coded for CMV serostatus. *Middle:* CD14$^+$ monocyte frequencies
934 within PBMCs. *Right:* Intermediate, ISGhi CD14$^+$, and CD16$^+$ monocyte frequencies within
935 PBMCs. All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and
936 CMV(+) (n=14) **c,** *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 scRNAseq data from Younger Adult Cohort: CMV(+) (n=42) and CMV(-) (n=50). **d,** *Left:* Mono-DC, and cDC2 frequencies within PBMCs. *Right:* cDC1 and AXL$^+$ DC frequencies within PBMCs. All panels represent scRNAseq data from Discovery Cohort: CMV(-) (n=13) and CMV(+) (n=14). **e,** Scaled expression of select marker genes for pDCs in Younger Adult Cohort. **f,** Individual-level frequencies of CMV-associated cell types over a three-year period in CMV(+) individuals from Younger Adult Cohort (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. **g**, Principal Variance Component Analysis (PVCA) of lymphocyte subset frequencies, estimating the proportion of variance explained by each clinical factor in the Discovery Cohort. 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 (**b**, **d, f**). 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.

**Extended Figure 6:**




**Extended Data Fig. 6: *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.[32] **b,** Heatmap showing percent of cell type labels mapped from AIFI_L3 CellTypist model to manual labels in our study. **c,** Confusion matrices for CMV serostatus prediction in each independent cohort from **Fig. 5b**. **d,** *Left*: table summarizing the longitudinal Younger Adult Cohort. '1 SC' stands for 'one seroconverter' (donor ID: BR1049). *Right top*: Confusion matrix for CMV serostatus predictions from the longitudinal Younger Adult Cohort in **Fig. 5d**. *Right lower*: Receiver Operating Characteristic (ROC) curve quantifying performance of *CMVerify* on the longitudinal Younger Adult Cohort in **Fig. 5d**. **e,** CMVerify prediction outcomes and CD4+ TEMRA

971 frequencies per sample in the longitudinal Younger Adult Cohort. Diamonds represent the 61-
972 year-old female (donor ID: BR2010), misclassified at five timepoints. Crosses represent the 59-
973 year-old male (donor ID: BR2013), misclassified at nine timepoints. Circles represent all other
974 samples. **f**, Expression of PBMC marker genes in donor BR1049 across all timepoints shown in
975 **Fig. 5e. g,** Expression of cytotoxic marker genes in select PBMC subsets from **Fig. 5f**. PPV,
976 positive predictive value; NPV, negative predictive value (**c,d**).
977
978