



OPEN

# Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection

Samuel W. Kazer<sup>1,2,3,4</sup>, Toby P. Aicher<sup>1,2,4</sup>, Daniel M. Muema<sup>5,6</sup>, Shaina L. Carroll<sup>7</sup>, Jose Ordovas-Montanes<sup>1,2,3,4,8,9</sup>, Vincent N. Miao<sup>1,2,4,10</sup>, Ang A. Tu<sup>4,11,12</sup>, Carly G. K. Ziegler<sup>1,2,4,10</sup>, Sarah K. Nyquist<sup>1,2,4,13,14</sup>, Emily B. Wong<sup>5,15,16,17</sup>, Nasreen Ismail<sup>6</sup>, Mary Dong<sup>1</sup>, Amber Moodley<sup>6</sup>, Bonnie Berger<sup>14,18</sup>, J. Christopher Love<sup>1,11,12</sup>, Krista L. Dong<sup>1</sup>, Alasdair Leslie<sup>5,16</sup>, Zaza M. Ndhlovu<sup>1,5,6,19</sup>, Thumbi Ndung'u<sup>1,5,6,16,20</sup>, Bruce D. Walker<sup>1,6,19,21</sup> and Alex K. Shalek<sup>1,2,3,4,10,11,13,21</sup>

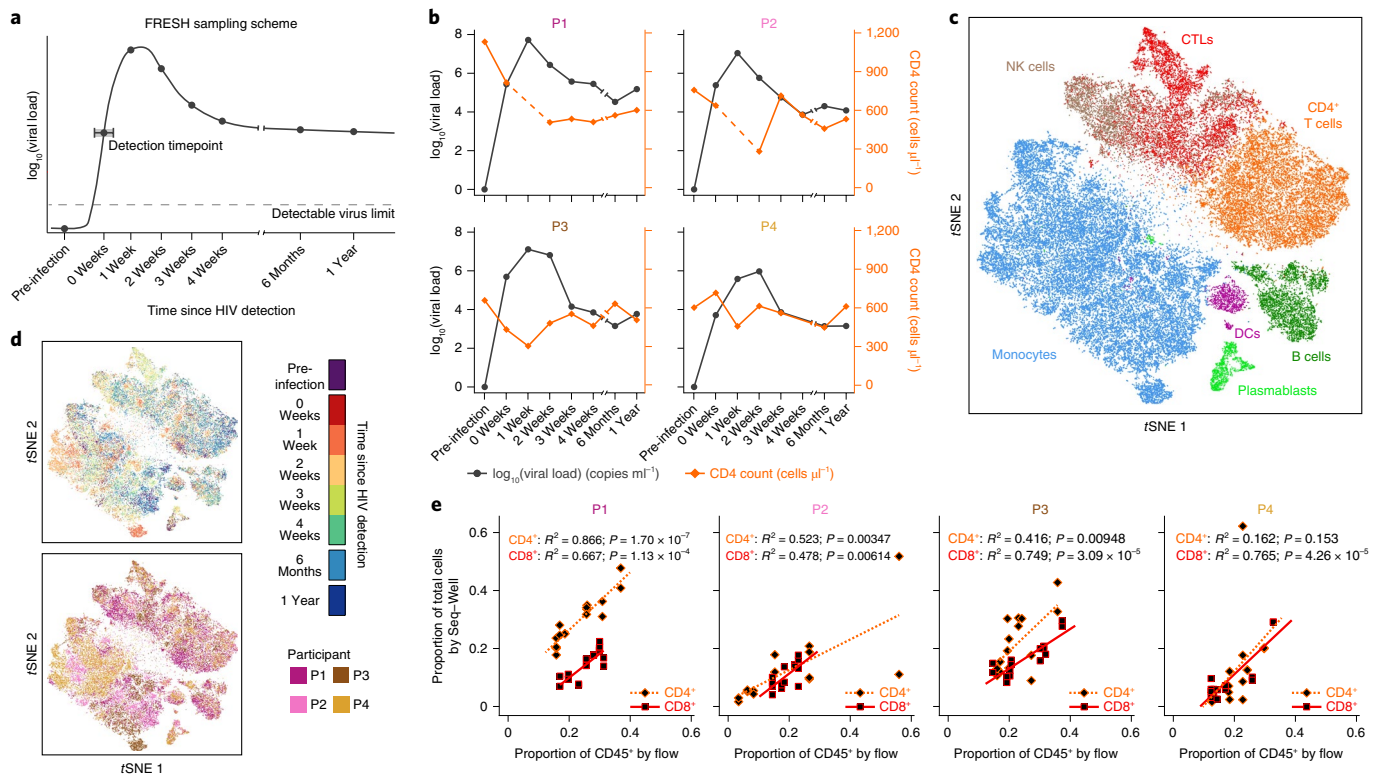
**Cellular immunity is critical for controlling intracellular pathogens, but individual cellular dynamics and cell-cell cooperativity in evolving human immune responses remain poorly understood. Single-cell RNA-sequencing (scRNA-seq) represents a powerful tool for dissecting complex multicellular behaviors in health and disease<sup>1,2</sup> and nominating testable therapeutic targets<sup>3</sup>. Its application to longitudinal samples could afford an opportunity to uncover cellular factors associated with the evolution of disease progression without potentially confounding inter-individual variability<sup>4</sup>. Here, we present an experimental and computational methodology that uses scRNA-seq to characterize dynamic cellular programs and their molecular drivers, and apply it to HIV infection. By performing scRNA-seq on peripheral blood mononuclear cells from four untreated individuals before and longitudinally during acute infection<sup>5</sup>, we were powered within each to discover gene response modules that vary by time and cell subset. Beyond previously unappreciated individual- and cell-type-specific interferon-stimulated gene upregulation, we describe temporally aligned gene expression responses obscured in bulk analyses, including those involved in proinflammatory T cell differentiation, prolonged monocyte major histocompatibility complex II upregulation and persistent natural killer (NK) cell cytolytic killing. We further identify response features arising in the first weeks of infection, for example proliferating natural killer cells, which potentially may associate with future viral control. Overall, our approach provides a unified framework for characterizing multiple dynamic cellular responses and their coordination.**

Despite advances in pre-exposure prophylaxis, there were 1.7 million new cases of HIV infection in 2018 (ref. <sup>6</sup>), highlighting the need for effective HIV vaccines. A better understanding of key immune responses during the earliest stages of infection, especially Fiebig stage I and II, before and at peak viral load, respectively, could help identify future prophylactic and therapeutic targets<sup>7</sup>. Using historical samples, collected before standard-of-care included treatment during acute infection, from the Females Rising through Education, Support and Health (FRESH) study<sup>5</sup>, we assayed evolving immune responses during hyperacute (1–2 weeks post-detection) and acute (3 weeks to 6 months) HIV infection.

We performed Seq-Well-based massively parallel scRNA-seq on peripheral blood mononuclear cells (PBMCs) from four FRESH participants who became infected with HIV during study. We analyzed multiple timepoints from pre-infection through 1 year following viral detection (Fig. 1a, Supplementary Table 1 and Methods) over which all four demonstrated a rapid rise in plasma viremia and a drop in CD4<sup>+</sup> T cell counts<sup>8</sup> (Fig. 1b and Extended Data Fig. 1a). Altogether, we captured 59,162 cells after performing quality controls, with an average of 1,976 cells per participant per timepoint (Extended Data Fig. 1b and Supplementary Table 2).

To assign cellular identity, we analyzed the combined data from all participants and timepoints (Methods). These analyses yielded few participant-specific features, suggesting that disease biology, rather than technical artifact, is the main driver of variation (Fig. 1d and Extended Data Fig. 1c,d). We annotated clusters by comparing differentially expressed genes defining each to known lineage markers and previously published datasets (Extended Data Fig. 1e,f and Supplementary Table 3). These clusters recapitulate several

<sup>1</sup>Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. <sup>2</sup>Institute for Medical Engineering and Science (IMES), Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>3</sup>Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>4</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>5</sup>African Health Research Institute, Durban, South Africa. <sup>6</sup>HIV Pathogenesis Programme, Nelson R. Mandela School of Medicine, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa. <sup>7</sup>Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA. <sup>8</sup>Division of Gastroenterology, Boston Children's Hospital, Boston, MA, USA. <sup>9</sup>Harvard Stem Cell Institute, Cambridge, MA, USA. <sup>10</sup>Program in Health Sciences and Technology, Harvard Medical School & Massachusetts Institute of Technology, Boston, MA, USA. <sup>11</sup>Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>12</sup>Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>13</sup>Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>14</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>15</sup>Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA. <sup>16</sup>Division of Infection and Immunity, University College London, London, UK. <sup>17</sup>Harvard Medical School, Boston, MA, USA. <sup>18</sup>Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>19</sup>Howard Hughes Medical Institute, Chevy Chase, MD, USA. <sup>20</sup>Max Planck Institute for Infection Biology, Berlin, Germany. <sup>21</sup>These authors contributed equally: Bruce D. Walker, Alex K. Shalek. ✉e-mail: [BWALKER@mgh.harvard.edu](mailto:BWALKER@mgh.harvard.edu); [shalek@mit.edu](mailto:shalek@mit.edu)



**Fig. 1 | Longitudinal profiling of peripheral immune cells in hyperacute and acute HIV infection by scRNA-seq.** **a**, Depiction of the typical trajectory of HIV viral load in the plasma during hyperacute and acute HIV infection adapted from Fiebig et al.<sup>8</sup> and the timepoints sampled in this study. Since participants were tested twice weekly, there was an uncertainty of up to 3 d in where on the viral load curve the first detectable viremia occurred (error bar is representative). The exact days sampled are available in Supplementary Table 1. **b**, Viral load and CD4<sup>+</sup> T cell count for four participants assayed in this study. Dotted lines indicate a missing data point for the metric. **c**, *t*-distributed stochastic neighbor embedding (tSNE) analysis of PBMCs from all participants and timepoints sampled ( $n = 59,162$ ). Cells are annotated based on differential expression analysis on orthogonally discovered clusters. **d**, tSNE in **c** annotated by timepoint (top) and participant (bottom). **e**, Scatter-plot depicting the correlation between cell frequencies of CD4<sup>+</sup> and CD8<sup>+</sup> T cells measured by Seq-Well ( $n = 2$  array replicates) and FACS ( $n = 1$  flow replicate).  $R^2$  values reflect variance described by an F-test for linear regression.

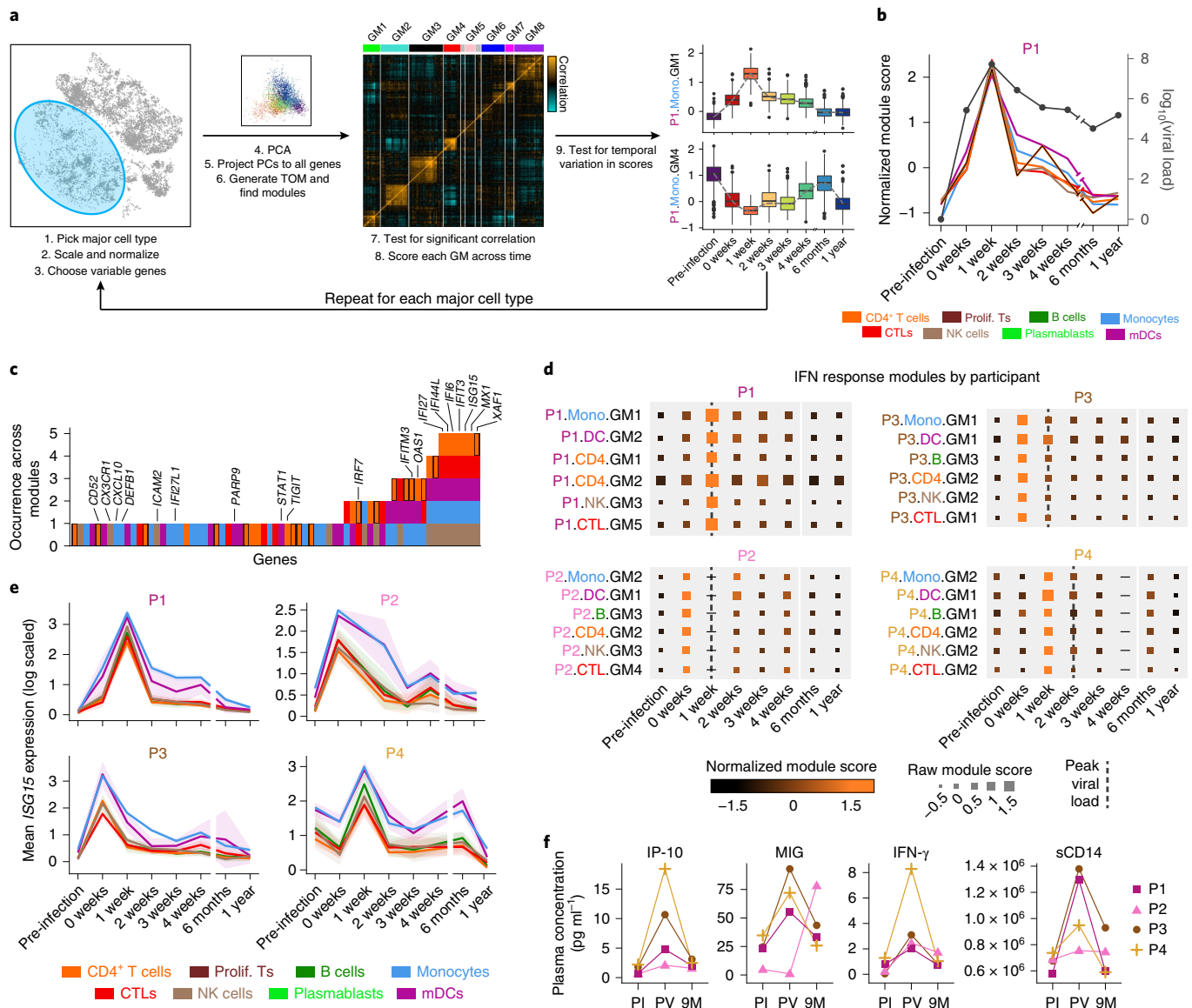
well-established PBMC subsets (Fig. 1c), revealed phenotypic subgroupings of both monocytes (antiviral, inflammatory and nonclassical) and cytotoxic T cells (CTLs) (CD8<sup>+</sup> CTL, proliferating; Extended Data Fig. 1g) and highlighted subset frequency dynamics such as natural killer (NK) cell expansion after 2–3 weeks. Flow cytometry measurements of CD45<sup>+</sup>CD3<sup>+</sup>CD4<sup>+</sup> and CD45<sup>+</sup>CD3<sup>+</sup>CD8<sup>+</sup> frequencies over the course of infection correlated with those measured by Seq-Well (Extended Data Fig. 2a,b and Fig. 1e). Whole blood monocyte counts, meanwhile, confirmed monocyte expansion following infection (Extended Data Fig. 2c).

Having mapped cell type frequency dynamics during acute HIV-1 infection, we next examined how different cellular phenotypes shifted over time. Previous applications of scRNA-seq to evolving cellular responses have either emphasized pseudotemporal ordering in development<sup>9</sup> to delineate well-ordered progressions through cell fate<sup>10</sup> or identified transcriptional differences<sup>11</sup> associated with disease treatment<sup>12</sup>. As our dataset includes multiple, noncontiguous timepoints and complex nonlinear dynamics spaced over days to weeks, it required distinct treatment. Therefore, we developed a framework to examine how each cell type varied in phenotype over the course of infection by adapting weighted gene correlation network analysis (WGCNA) to discover, in an unbiased manner at single-cell resolution, gene modules (GMs) whose expression varied significantly over time (Fig. 2a, Methods and Supplementary Tables 4 and 5). Given the small number of participants and heterogeneity in disease response, we opted to characterize each participant and cell type independently to: 1) identify cellular responses

associated with plasma viremia; 2) group modules within individuals over time; and 3) nominate molecular drivers and potential cell-cell signaling.

Within each individual, the discovered GMs demonstrated common transient patterns over the course of infection, indicating the utility of our approach in uncovering responses with shared dynamics across multiple cell types. Looking at GMs associated with changes in plasma viral load in participant 1 (P1), we identified a set of six spanning multiple cell types all sharing their highest relative module score at peak viremia (Fig. 2b). Despite being generated in distinct cell types, each GM included *IFI27*, *IFI44L*, *IFI6*, *IFIT3*, *ISG15* and *XAF1* (Fig. 2c and Extended Data Fig. 3a), in addition to other interferon (IFN)-stimulated genes (ISGs)<sup>13</sup>. Collectively, these expression patterns reveal cell-type-specific genes and functions correlated with a core ISG signature in P1, including monocyte antiviral activity (*CXCL10*, *DEFB1*)<sup>14,15</sup>, dendritic cell (DC) activation (*PARP9*, *STAT1*)<sup>16,17</sup>, naive CD4<sup>+</sup> T cell differentiation (*CD52*, *TIGIT*)<sup>18,19</sup> and NK cell trafficking (*CX3CR1*, *ICAM2*)<sup>20</sup>. Moreover, in P1, monocytes and DCs uniquely expressed genes (*CXCL10*, *LGALS3BP*) measured in bulk responses in acute simian immunodeficiency virus (SIV) infection in rhesus macaques<sup>21</sup>, which may shed light on the cellular sources of these antiviral molecules (Extended Data Fig. 3b and Methods).

Because we did not directly detect expression of IFN-I genes, potentially due to the sites or timepoints analyzed, we characterized the expression of their upstream regulator *IRF7* to infer which cell type(s) may be responsible for their production<sup>22</sup>. In P1, six of eight



**Fig. 2 | GM discovery reveals ubiquitous response to IFN with cell-type-specific features.** **a**, Schema depicting temporal GM discovery (see Methods). This procedure was repeated for each major cell type (monocytes, CD4<sup>+</sup> T cells, CTLs, proliferating T cells, NK cells, B cells, plasmablasts and mDCs) on a participant-by-participant basis, generating 0–8 GMs per cell type. Modules were arbitrarily numbered within a given cell type and participant. mDC, myeloid dendritic cell; PCA, principal-component analysis; PC, principal component; TOM, topological overlap matrix. **b**, In P1, six GMs across multiple cell types exhibited similar temporal expression profiles; each GM’s score (colored lines) peaked at the same timepoint as for peak viremia (line-and-dot plot). **c**, Number of occurrences of gene membership for all genes present across the six GMs in **b**. **d**, GM expression scores for IFN response modules in each participant. Normalized GM score is depicted by heat (black to orange), whereas raw module score is depicted by box size. The timepoint closest to peak viral load is indicated by a dotted line. **e**, Mean expression of *ISG15* separated by timepoint and individual. Shaded area denotes 95% CI of the mean. See Supplementary Table 2 for the number of cells per timepoint per cell type. **f**, Luminex measurements of IP-10, MIG and IFN- $\gamma$  and ELISA of soluble CD14 (sCD14) in matching plasma samples. Points are averages of duplicate measurements. PI, pre-infection; PV, peak viremia; 9M, 9 months post-detection.

cell types studied demonstrated higher expression of *IRF7* at peak viremia compared to pre-infection and 1-year timepoints (Extended Data Fig. 3c). We also assayed plasmacytoid DCs (pDCs), which produce IFN- $\alpha$  and IFN- $\beta$  in response to HIV<sup>23</sup>, at peak viremia and 1-year post-infection (Extended Data Fig. 3d,e and Methods) but did not find IFN-I gene expression or a significant change in *IRF7* expression (two-sided Wilcoxon rank-sum test, false discovery rate (FDR) corrected  $q < 1$ ).

The three other participants studied (P2–P4) each had pDC responses and sets of ISG GMs similar to P1 at, or the week before, peak viremia, which we corroborated at the individual gene level (Fig. 2d,e and Extended Data Fig. 3f–h). Comparing GMs across

individuals, we noted common ISGs (present in three or more cell types) that were shared in two or more participants (*ISG15*, *IFIT3*, *XAF1*) as well as some specific to a single participant (*APOBEC3A*, *IFI27*, *STAT1*; Extended Data Fig. 3i). To independently confirm the presence of IFNs and downstream cytokines, we measured IFN- $\gamma$ , MIG (CXCL9) and IP-10 (CXCL10; previously associated with disease progression and infection outcome<sup>24</sup>; Fig. 2f and Methods). All participants demonstrated higher levels of IFN- $\gamma$  and IP-10 at peak viremia and three demonstrated elevated MIG. We also observed increased soluble CD14, known to be associated with monocyte activation<sup>25</sup>.

Given concerted and cell-type-specific IFN responses during hyperacute HIV infection, we next explored whether other modules



exhibited shared expression dynamics. We applied fuzzy *c*-means clustering to the median module scores at each timepoint across all cell types on a participant-by-participant basis, generating clusters of modules which we refer to as meta-modules (MMs) (Methods). MMs represent gene programming across distinct cell types with coordinated temporal dynamics—here, synchronized responses to infection—enabling us to link cellularly discrete but contemporaneous behaviors to both common and unique propagators.

We next identified MMs from every participant and grouped them by their expression dynamics (Extended Data Fig. 4). We labeled four of these on the basis of their transient peak expression score patterns: sharp positive (MMsp), sharp negative (MMsn), gradual positive (MMgp) and gradual negative (MMgn); three additional MMs, labeled a–c, demonstrated more complex patterns. Besides MMsp, which contained the majority of the ISG modules, only MMs<sub>sn</sub>, enriched for ribosomal protein-coding genes (Supplementary Table 4) previously shown to indicate cellular quiescence<sup>26</sup>, spanned five or more cell types. In parallel, we attempted to discover conserved modules across individuals, using cells from all four participants binning timepoints by viral load (Methods, Extended Data Fig. 5a and Supplementary Table 6). All but four of these cross-participant modules recapitulated those found in our participant-specific approach (Extended Data Fig. 5b–d). However, this pan-participant analysis did not reveal any GMs with consistent expression trends (in at least three out of four participants) besides the already identified ISG (MMsp) and ribosomal protein (MMsn) modules and failed to discover several participant-specific modules within MMgp.

Notably, MMgp consisted of responses sustained throughout acute infection, but implicated different cell types in each participant. For example, in P2, MMgp consisted of monocyte, B cell, plasmablast, CTL and proliferating T cell GMs (Fig. 3a). Unlike MMsp (ISGs), these GMs spanned several distinct gene expression programs, such as antigen presentation (monocytes and B cells), interleukin (IL)-6 and IL-8 production (plasmablasts) and granzyme B production (CTLs; Fig. 3b,c and Supplementary Table 7). As these overlap in time, they may represent cell subsets responding to common stimuli and/or one another. Looking for known relationships between genes within and across cell types, we generated a network model describing potential axes of cell–cell signaling, both direct (via receptor–ligand) and indirect (signaling via chemokines and cytokines), in P2 (Fig. 3d and Methods). Expression of IL-8 and IL-6 in B cells and plasmablasts<sup>27</sup> may attract monocytes presenting antigen to prime CD4<sup>+</sup> T cells, potentially leading to IL-17 production<sup>28</sup> and *BCL2* upregulation, known to restrict CTL-mediated killing of infected cells<sup>29</sup>. Together, this suggests the IL-6–IL-8–IL-17 signaling axis as a potential target for future HIV treatment.

Given the diverse, participant-specific GMs in MMgp (Extended Data Fig. 4), we next looked whether any acute infection responses were present in multiple participants. In CD4<sup>+</sup> T cells, monocytes, NK cells, CTLs and proliferating T cells, we found GMs in MMgp that shared genes in two or more participants (Fig. 3e–i; see Supplementary Table 8 for overlapping genes). While DCs and B cells also expressed multiple GMs within MMgp, some did not share any genes across participants (Extended Data Fig. 6a) or had low membership scores and were thus excluded (membership <0.25, labeled with † in Extended Data Fig. 4; Methods).

We next qualitatively compared GM functional annotations within MMgp for each cell type across participants (Fig. 3e–i and Methods). Despite variable temporal dynamics and unique gene memberships, we observed significant enrichment for  $\geq 15$  of the same underlying pathways and functions in at least two participants ( $P < 0.01$ , Supplementary Table 9), suggesting the existence of common features across individuals despite heterogeneity in infection response. For example: 1) CD4<sup>+</sup> T cells (P3 + P4) expressed genes associated with nonclassical viral entry by endocytosis<sup>30</sup> and

adhesion, suggesting migration and viral dissemination throughout the body; 2) monocytes (P2 + P3 + P4) expressed genes associated with antigen presentation, potentially indicating generalized IFN responses or the potential to promote active T helper and CTL responses<sup>31</sup>; and 3) NK cells (P1 + P3), CTLs (P1 + P2) and proliferating T cells (P2 + P3 + P4) upregulated genes associated with killing of target cells by perforin and granzyme release, highlighting the joint role of innate and adaptive lymphocytes in combating viremia<sup>32,33</sup> (see Supplementary Table 7 and Extended Data Fig. 6b for shared genes). Gene expression data corroborate these GM expression trends (Extended Data Fig. 6c).

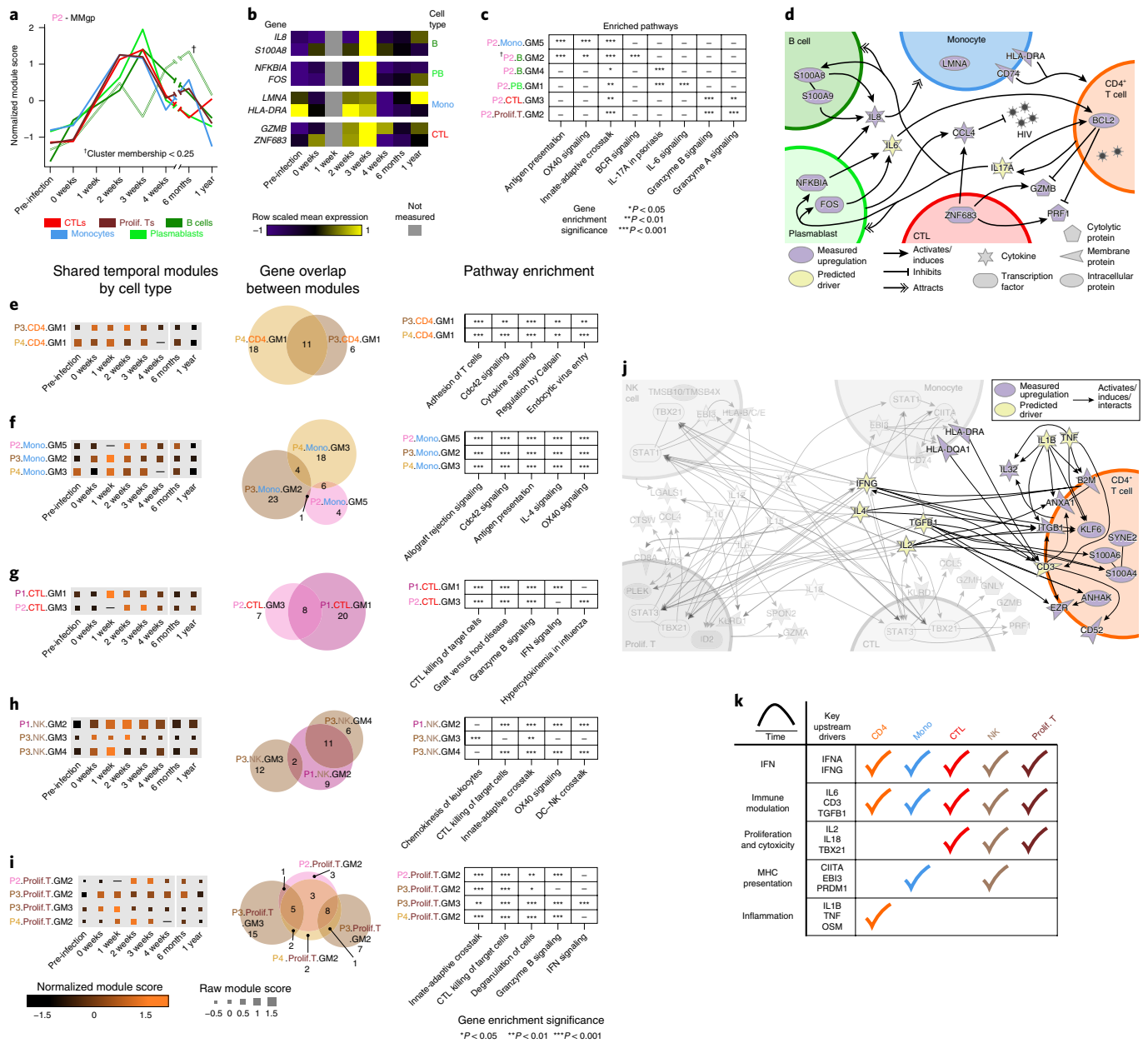
We hypothesized that there may be a common set of immune drivers coordinating these gene responses during infection. To identify potential inducers of the GMs in MMgp, we generated a list of predicted upstream drivers for each. Using hits that were significant for two or more GMs, we constructed a network detailing putative upstream signaling (Supplementary Table 9, Extended Data Fig. 6d,e and Methods), highlighting potential roles for: 1) IFN- $\alpha$  and IFN- $\gamma$  across all five cell types; 2) IL-15, IL-12 and IL-21 in CTLs, NK and proliferating T cells; and 3) IL-1 $\beta$  and tumor necrosis factor (TNF) restricted to CD4<sup>+</sup> T cells. Parallel Luminex measurements confirmed increased IP-10, MIG and IL-12, but not IFN- $\gamma$ , in plasma at 4 weeks, near when MMgp peaked in each individual (Extended Data Fig. 6f).

Re-scoring cell types against enriched genes for each driver revealed variable kinetics in the onset, intensity and length of immune responses across different cell types (Extended Data Fig. 7a). We note the following gene-programming upregulation trends in all participants: 1) CD4<sup>+</sup> T cells activity from before peak viremia throughout acute infection; 2) CTL and proliferating T cell programs are induced during hyperacute infection; and 3) NK cell and monocyte activity persists throughout the first month of infection, highlighting a persistent role for innate immunity throughout acute infection. Based on cell type, gene and functional enrichments, we summarize the shared ( $\geq 2$  participants) immune responses with sustained gene expression over the course of the first month of HIV infection, their potential drivers and putative cell–cell signaling, emphasizing CD4<sup>+</sup> T cells, which was the only cell type expressing genes downstream of proinflammatory cytokines (Fig. 3j,k and Extended Data Fig. 7b). Thus, our module discovery approach readily reveals immune responses and potential interactions among several cell types during acute HIV infection.

In our analyses, we observed GMs that demonstrated similar temporal response patterns within the same cell type but distinct pathway enrichments, implying orthogonal biological functionality: for example, the NK cytokine signaling GM3 module (*CCL3*, *CCL4*) and the cytotoxic GM4 module (*PRF1*, *GZMB*) in P3 (Fig. 3h). To understand how these GMs might be linked, we looked across single cells for module coexpression (observed in bulk approaches). Surprisingly, the strength of the correlation between expression of these modules across single NK cells changes with time, decreasing later in infection (Extended Data Fig. 8a,b). K-means clustering separated cells by variable expression of GM3 and GM4 (Extended Data Fig. 8c). Variation in the correlation of GM3 and GM4 may reflect NK cell plasticity with dual cytotoxic and signaling programming near peak viremia.

Examining MMsp (ISG GMs), we also observed that P3 exhibited temporally similar modules in monocytes (GM1 and GM3); however, these did not variably correlate over time. Instead, they were highly coexpressed, but only at HIV-detection (Extended Data Fig. 8d,e). Gene-set analysis demonstrated that monocyte GM1 consisted of antiviral response genes, while GM3 was enriched for genes associated with inflammation (Extended Data Fig. 8f). Thus, monocytes in P3 at the time of HIV detection are simultaneously expressing both antiviral and inflammatory gene programs,

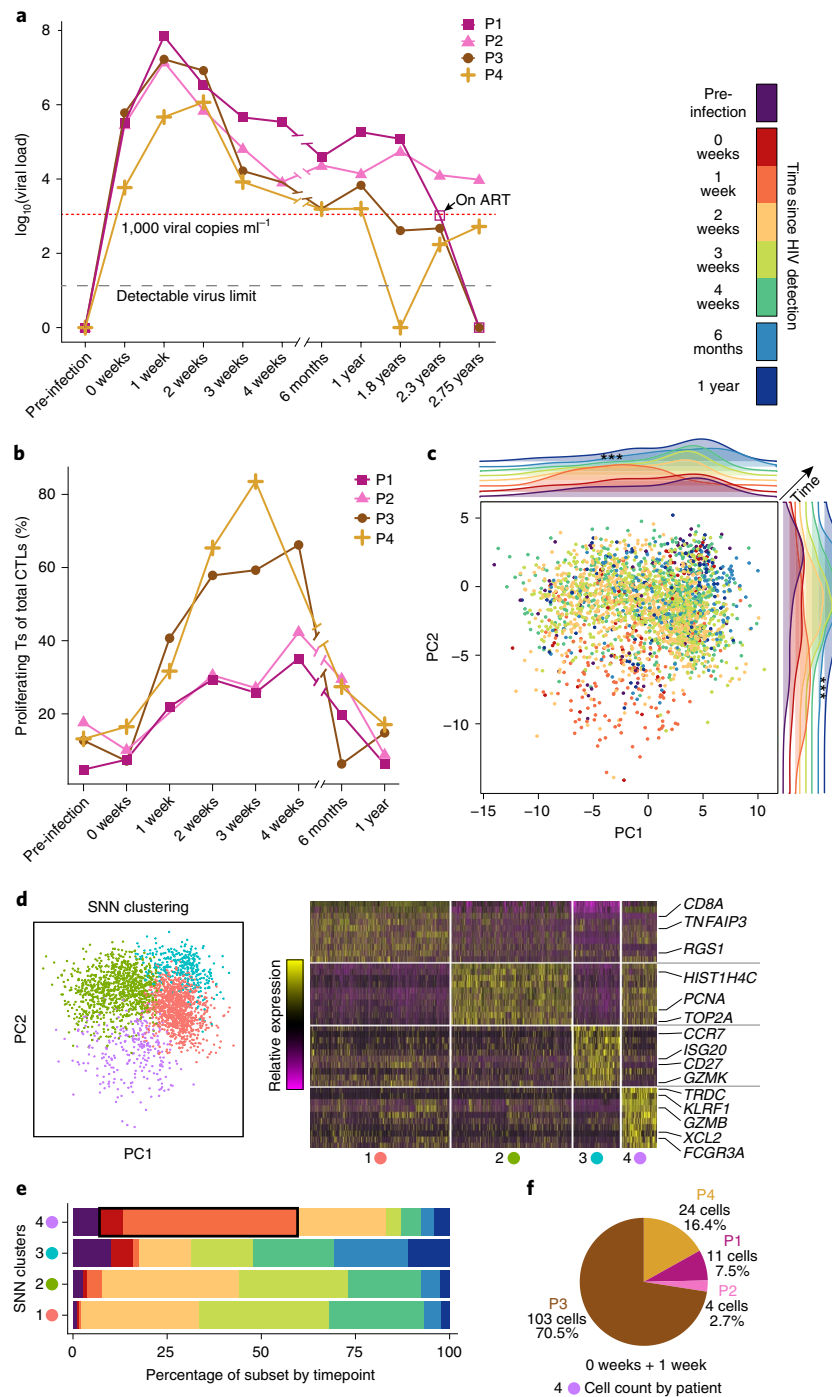




**Fig. 3 | Distinct modules across different cell types share temporal expression patterns in acute HIV infection and suggest shared and cell-type-specific drivers of immune response.** **a**, Normalized module expression scores for the six GMs clustered into meta modules: gradual positive (MMgp) in P2. † indicates GMs with MM membership score < 0.25 **b**, Mean gene expression of representative genes from modules in **a**; see Supplementary Table 4 for full gene lists. **c**, Select enriched pathways for the genes in each module from **a**; gene-set enrichment performed in ingenuity pathway analysis (IPA); see Supplementary Table 6 for the full list, using a right-tailed Fisher’s exact test. See Supplementary Table 4 for gene list sizes. **d**, Putative cell-cell signaling network. Nodes represent gene products with either measured gene upregulation in the modules in **a** or predicted drivers from IPA. Edges were drawn from connections nominated by IPA and curated from the literature. **e–i**, Module scores (left), gene overlaps between modules (middle) and enriched pathways (right; IPA) for modules grouped in MMgp and shared across participants in CD4+ T cells (**e**), monocytes (**f**), CTLs (**g**), NK cells (**h**) and proliferating T cells (**i**). Enriched pathways were determined using a right-tailed Fisher’s exact test. See Supplementary Table 4 for gene lists and their sizes. **j**, Putative cell-cell signaling network derived from genes shared across ≥2 participants from modules in **e–i**; see Supplementary Table 7. Nodes and edges are drawn as in **d**. Here we highlight those molecules interacting with or measured in CD4+ T cells; the full network is presented in Extended Data Fig. 7b. To reduce complexity, we omitted nodes depicting expression of GZMB, PRF1 and GNL1 by both NK cells and proliferating T cells and CCL5 by proliferating T cells. **k**, Summary table of immune responses to related and distinct stimuli with similar temporal dynamics, defined by transient increased module expression for several weeks after peak viremia.

a previously unappreciated phenotype. While both gene programs strongly contributed to the major axes of monocyte variation in all individuals, we were unable to identify polyfunctional monocytes in the other participants (Extended Data Fig. 9a–c). Meanwhile, non-classical monocytes displayed disparate temporal dynamics across

participants (Extended Data Figs. 1d and 9d). Comparing differentially expressed genes at peak response timepoints (1–2 weeks) further highlighted other participant-specific differences: monocytes in all participants produced antiviral factors (Extended Data Fig. 9e,f), but only P2 and P3 were enriched for inflammatory



**Fig. 4 | Future controllers exhibit higher frequencies of proliferating CTLs and a precocious subset of NK cells 1 week after detection of HIV viremia.**

**a**, Viral load by PCR with reverse transcription of the plasma of four participants assayed out to 2.75 years. Controllers of HIV maintain levels of plasma viremia  $<1,000$  viral copies  $\text{ml}^{-1}$ . P1 initiated ART before the 2.3-year timepoint. **b**, Proportion of proliferating T cells of total CTLs as a function of time and individual measured by Seq-Well. See Supplementary Table 2 for the number of cells per timepoint per cell type. **c**, PCA of proliferating T cells from all four individuals. Cells assayed from the 1-week timepoint strongly separated along PC1 and PC2; two-sample Mann-Whitney  $U$ -test; 174 cells (1 week) versus 2,465 cells (all other timepoints);  $***P < 2.2 \times 10^{-16}$ . **d**, Shared-nearest neighbor (SNN) clustering over the top six PCs reveals four subclusters (left) with distinct gene programs (right). Two-sample Wilcoxon rank-sum test was used for analysis; numbers of cells per cluster: 1–1,081; 2–929; 3–359; 4–270. **e**, Percentage of cells in each subcluster by timepoint. **f**, Number of cells from each individual within the cells sampled at 0 weeks and 1 week in the NK cell cluster (4, lilac; black box in **e**).

responses and only P3 for TNF signaling via NF- $\kappa$ B ( $q < 0.001$ ). Chronic inflammation has been associated with susceptibility to infection<sup>34</sup> and our data show variable inflammatory gene expression before infection with subsequent mixed expression changes in hyperacute infection across participants (Extended Data Fig. 9g).

As we have previously demonstrated that natural control of HIV is associated with diverse cellular phenotypes in CTLs<sup>35</sup> and DCs<sup>4</sup>, we looked to see whether the presence of polyfunctional monocytes in P3 might link to disease progression in chronic infection. We observed that both P3 and P4 maintained low levels of viremia

(<1,000 viral copies ml<sup>-1</sup>) at 2.75 years after infection in the absence of antiretroviral therapy (ART) (Fig. 4a). HIV infected persons who naturally maintain low levels of viremia in chronic infection (HIV controllers) demonstrate enhanced immune responses systemically<sup>4,36</sup>. However, whether early events in acute HIV infection reflect or contribute to control of chronic viremia is unknown.

As CD8<sup>+</sup> T cells are known to contribute to controlling chronic HIV infection<sup>35,36</sup>, we also analyzed CTLs from all participants, noting increasing levels of *PRF1* and *GZMB* during acute infection (Fig. 3g). Further unsupervised and directed approaches did not demonstrate significant differences in CTL responses across participants (Extended Data Fig. 10a,b and Supplementary Table 10). In FRESH, we demonstrated that the majority of proliferating CTLs in acute infection are HIV specific<sup>37</sup>. Therefore, we looked for differences in proliferating T cell responses by participant. On average, proliferating T cells expressed similar levels of cytotoxic genes as non-proliferating CTLs. Differential expression analysis highlighted genes associated with cell cycle and memory for proliferating and nonproliferating CTLs, respectively (Extended Data Fig. 10c,d and Supplementary Table 10). T cell receptor (TCR) pull-down and enrichment (TCR-β CDR3) revealed few expanded clones (Extended Data Fig. 10e,f, Supplementary Table 10 and Methods); this, however, may be affected by sample size (CDR3s were detected in 982 proliferating T cells). Relative to P1 and P2, both controllers (P3 and P4) displayed higher frequencies of proliferating cytotoxic cells within the first month of infection compared to pre-infection (Fig. 4b).

We next used unsupervised analyses to examine differences in proliferating T cell responses over time among participants (Fig. 4c and Extended Data Fig. 10g). Clustering over all proliferating T cells, we identified four subsets of cells with distinct gene programs (Fig. 4d and Supplementary Table 10): traditional CD8<sup>+</sup> T cells, hyperproliferative CD8<sup>+</sup> T cells, naive CD4<sup>+</sup> T cells and a subset of cells that were CD8A<sup>-</sup> but TRDC<sup>+</sup> and FCGR3A<sup>+</sup> (CD16). Using signatures from a single-cell study of cytotoxic cells, we determined that the FCGR3A<sup>+</sup> cells were NK cells (Extended Data Fig. 10h and Methods). Looking at the distribution of cells within each of these clusters, the NK cluster contained the highest proportion of proliferating cells at HIV detection and 1 week thereafter (Fig. 4e,f). The majority of these were from P3 and P4. Thus, our data show that the two participants who maintain viral loads <1,000 viral copies ml<sup>-1</sup> at 2.75 years after infection without ART exhibit a subset of proliferative, cytotoxic NK cells during the earliest stages of acute infection before the majority of HIV-specific CD8<sup>+</sup> T cells arise.

Here, we present and apply a novel scRNA-seq-based framework to a unique longitudinal study of human infection in order to characterize conserved immune response dynamics, as well as cell subsets and gene programs with potential therapeutic and preventative applications. By analyzing hundreds of cells per timepoint and cell type, we were powered to identify significant changes in abundant cellular phenotypes over time in each participant; however, we could not account for all potential sources of cellular heterogeneity or their impact on infection outcome. Nevertheless, we discovered interrelated temporal GM expression patterns in distinct cell types and nominated mechanisms by which multiple components of the immune system may respond collectively—sometimes with different gene programs—to HIV infection. By identifying upstream drivers that may induce our MMs, we hypothesized when and how various cytokines, chemokines and transcription factors might orchestrate immune responses during infection. Together, our work affords a unique reference dataset for studying the earliest moments of HIV infection after detection and suggests potentially new roles for monocytes, NK cells and CD4<sup>+</sup> T cells in acute infection.

Our single-cell approach also enabled us to identify cellular subsets present during hyperacute HIV infection in two individuals (P3 and P4) who maintained low viremia in chronic infection. In addition to polyfunctional monocytes identified in P3, we found a

subset of cytotoxic, proliferating NK cells in P3 and P4. In other infection settings<sup>38,39</sup>, NK cells have demonstrated antigenic memory, suggesting that these cells could be responding to some previously encountered antigen; however, while all participants tested negative for sexually transmitted infections pre-HIV-infection, we did not screen for other chronic infections or exposures. These proliferating NK cells may function alongside CTLs early in infection, mitigating CTL antigenic load and subsequent exhaustion<sup>40</sup>. Although there are ethical and practical difficulties associated with collecting additional samples from untreated HIV infected persons, follow-up studies in nonhuman primates could determine the functional, actionable importance of early immune responses for long-term viral control. Future work in FRESH will seek to test the effects of early administered ART on longitudinal HIV response dynamics. Parallel efforts in human cohorts with other viral and bacterial infections, as well as inflammatory diseases and cancers, will enable assessment of the broad utility of the framework described herein, and begin to reveal the common and unique immune response motifs that inform human immunity for future modulation with drugs or treatments.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-020-0799-2>.

Received: 8 July 2019; Accepted: 12 February 2020;

Published online: 23 March 2020

### References

- Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
- Gomes, T., Teichmann, S. A. & Talavera-López, C. Immunology driven by large-scale single-cell sequencing. *Trends Immunol.* **40**, 1011–1021 (2019).
- Shalek, A. K. & Benson, M. Single-cell analyses to tailor treatments. *Sci. Transl. Med.* **9**, eaan4730 (2017).
- Martin-Gayo, E. et al. A reproducibility-based computational framework identifies an inducible, enhanced antiviral state in dendritic cells from HIV-1 elite controllers. *Genome Biol.* **19**, 10 (2018).
- Ndung'u, T., Dong, K. L., Kwon, D. S. & Walker, B. D. A FRESH approach: combining basic science and social good. *Sci. Immunol.* **3**, eaau2798 (2018).
- Joint United Nations Programme on HIV/AIDS. *UNAIDS Data 2019*. (UNAIDS, 2019).
- Robb, M. & Ananworanich, J. Lessons from acute HIV infection. *Curr. Opin. HIV AIDS* **11**, 555–560 (2016).
- Fiebig, E. et al. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS* **17**, 1871–1879 (2003).
- Pijuan-Sala, B., Guibentif, C. & Göttgens, B. Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nat. Rev. Mol. Cell Biol.* **19**, 399 (2018).
- Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
- Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA-sequencing data. *BMC Bioinformatics* **20**, 40 (2019).
- Tirosh, I. & Suvà, M. L. Deciphering human tumor biology by single-cell expression profiling. *Annu. Rev. Cancer Biol.* **3**, 151–166 (2019).
- Schoggins, J. W. & Rice, C. M. Interferon-stimulated genes and their antiviral effector functions. *Curr. Opin. Virol.* **1**, 519–525 (2011).
- Corleis, B. et al. Early type 1 Interferon response induces upregulation of human β-defensin 1 during acute HIV-1 infection. *PLoS ONE* **12**, e0173161 (2017).
- Vargas-Inchaustegui, D. A. et al. CXCL10 production by human monocytes in response to *Leishmania braziliensis* infection. *Infect. Immun.* **78**, 301–308 (2010).
- Luban, J. Innate immune sensing of HIV-1 by dendritic cells. *Cell Host Microbe* **12**, 408–418 (2012).
- Ng, D. & Gommerman, J. L. The regulation of immune responses by DC derived type 1 IFN. *Front. Immunol.* **4**, 94 (2013).
- Kurtulus, S. et al. TIGIT predominantly regulates the immune response via regulatory T cells. *J. Clin. Invest.* **125**, 4053–4062 (2015).



19. Samten, B. CD52 as both a marker and an effector molecule of T cells with regulatory action: identification of novel regulatory T cells. *Cell Mol. Immunol.* **10**, 456–458 (2013).
20. Lugli, E., Marcenaro, E. & Mavilio, D. NK cell subset redistribution during the course of viral infections. *Front. Immunol.* **5**, 390 (2014).
21. Bosinger, S. E. et al. Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys. *J. Clin. Invest.* **119**, 3556–3572 (2009).
22. Bosinger, S. E. et al. Intact type I interferon production and IRF7 function in sooty mangabeys. *PLoS Pathogens* **9**, e1003597 (2013).
23. O'Brien, M., Manches, O. & Bhardwaj, N. Plasmacytoid dendritic cells in HIV infection. *Adv. Exp. Med. Biol.* **762**, 71–107 (2013).
24. Jiao, Y. et al. Plasma IP-10 is associated with rapid disease progression in early HIV-1 infection. *Viral Immunol.* **25**, 333–337 (2012).
25. Shive, C. L., Jiang, W., Anthony, D. D. & Lederman, M. M. Soluble CD14 is a nonspecific marker of monocyte activation. *AIDS* **29**, 1263–1265 (2015).
26. Athanasiadis, E. I. et al. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat. Commun.* **8**, 2045 (2017).
27. Matsusaka, T. et al. Transcription factors NF-IL6 and NF-κB synergistically activate transcription of the inflammatory cytokines, interleukin 6 and interleukin 8. *Proc. Natl Acad. Sci. USA* **90**, 10193–10197 (1993).
28. Yue, F. Y. et al. Virus-specific interleukin-17-producing CD4<sup>+</sup> T cells are detectable in early human immunodeficiency virus type 1 infection. *J. Virol.* **82**, 6767–6771 (2008).
29. Hou, W., Jin, Y.-H., Kang, H. S. & Kim, B. S. Interleukin-6 (IL-6) and IL-17 synergistically promote viral persistence by inhibiting cellular apoptosis and cytotoxic T cell function. *J. Virol.* **88**, 8479–8489 (2014).
30. Sloan, R. D. et al. Productive entry of HIV-1 during cell-to-cell transmission via dynamin-dependent endocytosis. *J. Virol.* **87**, 8110–8123 (2013).
31. Jakubzick, C. V., Randolph, G. J. & Henson, P. M. Monocyte differentiation and antigen-presenting functions. *Nat. Rev. Immunol.* **17**, 349–362 (2017).
32. Gulzar, N. & Copeland, K. F. T. CD8<sup>+</sup> T-cells: function and response to HIV infection. *Curr. HIV Res.* **2**, 23–37 (2004).
33. Scully, E. & Alter, G. NK cells in HIV disease. *Curr. HIV AIDS Rep.* **13**, 85–94 (2016).
34. Kaspersen, K. A. et al. Low-grade inflammation is associated with susceptibility to infection in healthy men: results from the Danish blood donor study (DBDS). *PLoS ONE* **11**, e0164220 (2016).
35. Ranasinghe, S. et al. Antiviral CD8<sup>+</sup> T cells restricted by human leukocyte antigen class II exist during natural HIV infection and exhibit clonal expansion. *Immunity* **45**, 917–930 (2016).
36. Walker, B. D. & Yu, X. G. Unravelling the mechanisms of durable control of HIV-1. *Nat. Rev. Immunol.* **13**, 487–498 (2013).
37. Ndhlovu, Z. M. et al. Magnitude and kinetics of CD8<sup>+</sup> T cell activation during hyperacute HIV infection impact viral set point. *Immunity* **43**, 591–604 (2015).
38. Reeves, R. K. et al. Antigen-specific NK cell memory in rhesus macaques. *Nat. Immunol.* **16**, 927–932 (2015).
39. Cerwenka, A. & Lanier, L. L. Natural killer cell memory in infection, inflammation and cancer. *Nat. Rev. Immunol.* **16**, 112–123 (2016).
40. Hoffmann, M. et al. Exhaustion of activated CD8 T cells predicts disease progression in primary HIV-1 infection. *PLoS Pathogens* **12**, e1005661 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

## Methods

**Study participants.** All participants in this study were enrolled in the FRESH cohort<sup>541</sup>. This prospective study recruits women who are HIV negative, aged 18–24 years and are tested for HIV-1 RNA in plasma twice weekly for 1 year. Each time the women come to the study center, they participate in peer-support groups and receive a stipend. In addition to semi-weekly virus testing by PCR with reverse transcription, whole blood is collected four times (including during enrollment) throughout the year from participants. If a plasma test comes back positive, the participant is asked to come back to the clinic that day to collect a blood sample. Samples are then collected weekly through the first 6 weeks of infection and regularly afterward as long as the participant continues to return to the study center. In the arm of the study described herein, participants were initiated on ART when their CD4 count fell below 350 cells  $\mu\text{l}^{-1}$ , per standard treatment guidelines at the time of enrollment. A second arm of the study was initiated in 2014 and is currently still in place; in that arm, participants who test positive for viral RNA are initiated on ART when they are called back into the study center for their first post-infection sample collection. To the best of our knowledge, all participants in this study had not yet started ART for the timepoints processed here. FRESH was performed in accordance with protocols approved by the Institutional Review Board at Partners (Massachusetts General Hospital), MIT and the Biomedical Research Ethics Committee of the University of KwaZulu-Natal. All FRESH participants consented for genetic and genomic data collection and analysis.

**Cell preparation, flow cytometry and cell sorting.** The Life Sciences Reporting Summary contains information on the sample preparation, antibodies, gating strategy and sort strategy used in this study.

**Single-cell RNA-seq with Seq-Well.** The Seq-Well platform was utilized as previously described<sup>42</sup> to capture the transcriptomes of single cells on barcoded mRNA capture beads. In brief, 10  $\mu\text{l}$  of sorted CD45<sup>+</sup>Calcein Blue<sup>+</sup> PBMCs were mixed at 1:1 dilution with Trypan blue and counted using a hemocytometer. The cells were resuspended in RPMI + 10% FBS at a final concentration of ~100,000 live cells  $\text{ml}^{-1}$  and 20,000–25,000 cells in 200  $\mu\text{l}$  were added to each Seq-Well array preloaded with barcoded mRNA capture beads (ChemGenes). Two arrays were used for each sample to increase cell numbers. The arrays were then sealed with a polycarbonate membrane (pore size of 0.01  $\mu\text{m}$ ), cells were lysed, transcripts were hybridized to the beads and the barcoded mRNA capture beads were recovered and pooled for reverse transcription using Maxima H-RT (Thermo Fisher EPO0753) and all subsequent steps. After an Exonuclease I treatment (NEB M0293L) to remove excess primers, whole transcriptome amplification (WTA) was carried out using KAPA HiFi PCR Mastermix (Kapa Biosystems KK2602) with 2,000 beads per 50  $\mu\text{l}$  of reaction volume. Libraries were then pooled in sets of eight (totaling 16,000 beads) and purified using Agencourt AMPure XP beads (Beckman Coulter, A63881) by a 0.6 $\times$  volume wash followed by a 0.8 $\times$  volume wash and quantified using Qubit dsDNA Assay (Thermo Fisher Q32854). Quality of the WTA product was assessed using the Agilent hsD5000 Screen Tape System (Agilent Genomics) with an expected peak >800 bp tailing off to beyond 3,000 bp, and a small or non-existent primer peak, indicating a successful preparation. Libraries were then constructed using a Nextera XT DNA library preparation kit (Illumina FC-131-1096) on a total of 750 pg of pooled cDNA library from 16,000 recovered beads using index primers as previously described<sup>43</sup>. Tagmented and amplified sequences were purified using a 0.8 $\times$  volume AMPure XP bead wash yielding library sizes with an average distribution of 500–750 bp in length as determined using an Agilent hsD1000 Screen Tape System (Agilent Genomics). Two Seq-Well arrays were sequenced per NextSeq500 sequencing run with an Illumina 75 Cycle NextSeq500/550 v2 kit (Illumina FC-404-2005) at a final concentration of 2.4 pM. The read structure was paired end with Read 1, starting from a custom read 1 primer, covering 20 bases inclusive of a 12-bp cell barcode and 8-bp unique molecular identifier (UMI), then an 8-bp index read and finally Read 2 containing 50 bases of transcript sequence.

**Seq-Well alignment, cell identification and cell type separation.** Read alignment, cell barcode discrimination and UMI per transcript collation were performed as by Ordovas-Montanes et al.<sup>43</sup> using a hg19 reference. Initially, we aligned the sequences from P1 to a combined HIV + hg19 genome using the consensus sequence of HIV clade C viruses from the HIV Sequence Database (<https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html>). After alignment, however, we measured 0–2 cells with HIV transcript alignments per array; therefore, we used the standard hg19 reference for our analysis. UMI-collapsed data were used as input into Seurat<sup>44</sup> (v2.3.4) for cell and gene trimming and downstream analysis. The following steps were performed on all of the arrays processed from a single participant, on a participant-by-participant basis. Any cell with <750 UMIs or >6,000 UMIs (0–5 cells per array) and any gene expressed in fewer than five cells were discarded from downstream analysis. This cells-by-genes matrix was then used to create a Seurat object for each participant. Cells with >20% of UMIs mapping to mitochondrial genes were then removed (50–100 cells per array). These objects (one per

participant) were then merged into one object for pre-processing and cell type identification

The combined Seurat object was log-normalized with a size factor of 10,000 and scaled without centering. Additionally, linear regression was performed to remove unwanted variation due to cellular complexity (nUMI) and low-quality cells (percent.mito). Subsequently, 3,251 variable genes were identified using the 'LogVMR' function and the following cutoffs: x.low.cutoff=0.01, x.high.cutoff=10 and y.cutoff=0.25. PCA was performed over these genes and the top 17 PCs were chosen for clustering and embedding on the basis of the curve of variance described by each PC and the genes most contributing to each PC. Next, FindClusters (SNN graph + modularity optimization) with a resolution of 0.5 was used to generate 13 clusters and the Fourier transform *t*-distributed stochastic neighbor embedding (tSNE) implementation<sup>45</sup> with 2,000 iterations to embed the data into two-dimensional space.

Cluster identity was assigned by finding differentially expressed genes using Seurat's implemented Wilcoxon rank-sum test and then comparing those cluster-specific genes to previously published datasets<sup>46–48</sup>. One cluster exhibited no cluster-specific genes; the cells from this cluster were embedded centrally in the tSNE, and on further investigation expressed both myeloid and lymphocyte markers. Therefore, these cells were removed as multiplets (when multiple cells enter the same well in the Seq-Well array). After multiplet removal, 59,162 cells were captured across all samples processed. The remaining 12 clusters included subsets of major circulating immune cells (see Supplementary Table 3 for marker genes). These clusters were merged by parent cell type (T cell, cytotoxic T cell, B cell, plasmablast, DC and monocyte) for downstream analysis, as variation in the SNN graph parameters weakly affected cluster assignment to the subsets.

As NK cells share many markers transcriptionally with cytotoxic T cells<sup>46</sup>, clustering in our dataset did not separate these two cytotoxic cell types. NK cells were annotated based on lacking expression of CD3 (*CD3D*, *CD3E*, *CD3G*) and nonzero expression of CD16 (*FCRG3A*) and *KLRF1*. CD56 (*NCAM*) was not highly expressed in our data and therefore was not used to separate NK cells. Any cell with a cluster identity belonging to the cytotoxic T cell cluster that lacked CD3 expression or expressed *CD16/KLRF1* was annotated as an NK cell. With this annotation, we noted distinct transcriptional responses between NK cells and CTLs both as a function of time and gene membership (Fig. 2c and Fig. 3g,h).

For downstream analysis of temporal variation in expression, the dataset was separated by participant and cell type: CD4<sup>+</sup> T cells, NK cells, CTLs, proliferating T cells, B cells, plasmablasts, mDCs and monocytes. The expression matrix and associated metadata can be accessed online through the Single Cell Portal hosted by Broad Institute of MIT and Harvard (see Data Availability; [https://singlecell.broadinstitute.org/single\\_cell/study/SCP256](https://singlecell.broadinstitute.org/single_cell/study/SCP256)).

**Cell type normalization.** Once separated by cell type and participant, the single-cell transcriptomes were processed on a cell-type-by-cell-type basis across all timepoints. For each cell type, the presence of residual contaminant RNA or doublets was assayed by scoring every cell against a set of contaminant genes from other cell types built from our marker list used to discern cluster identity (see Supplementary Table 11 for cell-type-specific contaminant gene lists and cutoffs). Cells with high contamination scores (0–10% of cells) were subsequently removed from further analysis to avoid unwanted variation in the subsequent unsupervised module discovery. Following contamination filtering, data underwent scaling and normalization, followed by variable gene discovery (~400–1,000 genes, dependent on cell type and cell number). PCA was then applied on the limited set of genes, followed by projection to the rest of the genes in the dataset.

**Module discovery.** For the module analysis, we subset our data on the top and bottom 50 genes, after projection, for the first 3–9 PCs (dependent on the variance described by each PC and genes contributing to each PC) as input for WGCNA<sup>49,50</sup> functions. Following the WGCNA tutorial (<https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/>), an appropriate soft-power threshold was chosen to calculate the adjacency matrix. As scRNA-seq data is impacted by transcript dropout (failed capture events), adjacency matrices with high power further inflate the impact of this technical limitation and yield few correlated modules. Therefore, when possible, a power was chosen as suggested by the authors of WGCNA (that is, the first power with a scale-free topology >0.8); however, in instances where this power yielded few modules (fewer than three), we decreased our power. As a rule, smaller soft powers lead to fewer large-sized modules, whereas larger soft powers lead to more small-sized modules. In our analysis, there was frequently a distinct tipping point where, as we increased soft power, modules would fail to be identified by WGCNA (due to low connectivity). We ran our analyses with several soft powers to find an appropriate balance to generate a maximal number of modules without losing GM membership. Next, an adjacency matrix was generated using the selected soft power and it was transformed into a TOM. Subsequently, this TOM was hierarchically clustered and the *cutreeDynamic* function with method 'tree' was used to generate modules of correlated genes (minimum module size of ten). Similar modules were then merged using a dissimilarity threshold of 0.5 (that is, a correlation of 0.5); WGCNA typically suggests dissimilarity thresholds of 0.8–0.95, but we sought to avoid any spurious cluster separation potentially associated with the chosen soft power.

To test the significance of the correlation structure of a given module, a permutation test was implemented. Binning genes in the true module by average gene expression (number of bins was ten), genes with the same distribution of average expression from the total list of genes used for module discovery were randomly picked 10,000 times. For each of these random modules, a one-sided Mann–Whitney *U*-test was performed to compare the distribution of dissimilarity values between the genes in the true module and the distribution of dissimilarity values between the genes in the random module. Correcting the resulting *P* values for multiple hypothesis testing by Benjamini–Hochberg FDR correction, a module was considered significant if <500 tests ( $P < 0.05$ ) had  $FDR > 0.05$ . We note that if we chose a smaller soft power for TOM generation, which in turn resulted in larger modules with fewer excluded genes, fewer modules passed this permutation test, likely due to noisier genes that maintained weak correlations with all other genes in the analysis.

Since we were interested in identifying modules of genes that changed in expression as a function of time, another permutation test was implemented to identify modules that significantly vary from pre-infection. First, every cell was scored for the genes within the module, using the AddModuleScore function in Seurat. This function calculates an average module score by calculating the mean expression of the genes within the module corrected for expression of other genes with similar means across the dataset. Thus, this score functions as an expression estimate of the genes within a module in any given single cell. As testing for differences in distribution is sensitive to sample number, a sample size (*s*) was selected based on the number of cells present at any given timepoint within a cell type. The smallest *s* used was ten; this cutoff was chosen based on the least frequent cell types having ~100 cells total across all timepoints within a participant. If a timepoint had fewer than ten cells, that point was not used in the testing. In the case of plasmablasts and mDCs in multiple participants, more than three timepoints had fewer than ten cells and therefore no modules were considered significantly variant in time. To determine whether module expression varied over time, 1,000 two-sided Mann–Whitney *U*-tests between the distribution of scores from *s* random cells at pre-infection and *s* random cells from each other timepoint were performed. For each timepoint, the *P* values from the 1,000 tests were averaged. After FDR correction, if  $q < 0.05$  for any timepoint, the module was considered to significantly vary in expression in time. Our approach and tests have been written as functions in R and have been included as Supplementary Software.

While a similar approach is possible using bulk RNA-seq data, here, we are powered to identify temporally similar modules active in distinct subsets of cells both within and across time and we can use each cell of a specific type as a well-controlled, independent biological replicate to identify, from a single sample, essential response features and their putative upstream drivers. Compared to a directed approach, this discovery-based identification of temporally variant modules enables unbiased selection of coordinated genes and pathways and immediately reveals differences in response dynamics among cell types, states and participants.

For the cross-participant module discovery analysis (Extended Data Fig. 5), we applied the WGCNA framework to all cells of a given cell type across all four participants at all timepoints sampled. Here, the number of genes input into the framework varied between ~350–850 genes by choosing the top and bottom 100 genes from the most significant PCs, determined by finding the asymptote in the PC elbow plot (ranked s.d. of each PC). These modules were then tested for significant correlation against random sets of genes using the same permutation test outlined above. To test for temporal variability in module expression across all four participants, we binned timepoints into pre-infection, peak, post-peak and 1-year and implemented an analysis of variance (ANOVA) across binned timepoints accounting for participant heterogeneity (see Extended Data Fig. 5a). Specifically, we fitted a linear regression to the data across binned timepoints using two models: (1) null hypothesis ~1 + participant; and (2) alternative hypothesis ~1 + participant + time.bin. We then calculated the *F* statistic for the ANOVA between these two models. Peak and post-peak timepoints were chosen based on the score maxima for the modules discovered in each participant in MMsp and MMgp (see Extended Data Fig. 4).

**Module grouping and gene-set analysis.** To more easily compare modules by temporal pattern within and between participants, fuzzy *c*-means clustering was applied to all of the modules in a given participant using the Mfuzz package<sup>51</sup> (v.2.38.0). We chose to use fuzzy *c*-means clustering to allow us to understand the extent of membership of a given module to its assigned cluster. For each participant, *c* was chosen to be 5–7, such that diverse temporal patterns were separated, minimizing the number of clusters containing fewer than three modules. These groupings of modules were then annotated by similar scoring patterns across participants, taking into consideration that infection time is not the same for every participant (Extended Data Fig. 4). We named four of these MMs on the basis of the transient module score dynamics of each: MMsp, MMsn, MMgp and MMgn. The remaining three MMs were named a–c given their more complex score dynamics.

Gene-set analysis on modules was performed using IPA (Qiagen) given its better performance with low gene numbers; our modules were sized between 10–66 genes. Only gene names were supplied for analysis and submitted for core

analysis with the experimentally observed confidence setting. In Fig. 3d–i, the pathways annotated were taken from either the canonical pathways or diseases and functions results. For the upstream driver analysis in Extended Data Fig. 6d,e, upstream drivers were selected by the following criteria: significant ( $P < 0.001$ ) in at least two modules of any given cell type, with at least five genes in the gene set. As the gene sets annotated in IPA are quite large and share many genes, the edges in our network were restricted to only those upstream drivers that shared three or more genes. To achieve finer grain temporal resolution on putative inducers of immune response, the union of enriched genes for each upstream driver from modules within a given cell type was used to generate scores against the single-cell expression data. Only upstream driver scores that demonstrated temporal variability (as described above) were included. We report the median scores at each timepoint for each upstream driver.

We chose to use parts of MSigDB v.6.2 (<http://software.broadinstitute.org/gsea/msigdb>) for the gene-set enrichment analysis in Extended Data Figs. 8 and 9, given higher gene numbers (>100), allowing for more conservative *P* values. Multiple-hypothesis testing was corrected by the Benjamini–Hochberg FDR procedure. The specific collections of gene sets used are reported in the figure legends.

**Cell–cell signaling network curation.** The cell–cell signaling networks in Fig. 3d,j and Extended Data Fig. 7b were generated using connections annotated in IPA<sup>52</sup>. Molecules of interest were chosen from genes in the modules belonging to MMgp in P2 (Fig. 3d) or shared among at least two participants (Fig. 3j, Extended Data Fig. 7b and Supplementary Table 7), respectively. We also included select upstream drivers found to be significant by IPA given enrichment of downstream genes within the modules. Edges were drawn between all nodes (genes or predicted upstream drivers) with the ‘Connect’ tool in ‘My Pathways’ using both ‘Direct’ and ‘Indirect’ interactions. Subsequently, edges were manually trimmed by looking at the provided support for the connections and discarding any connections not supported by demonstrations of expression or activation in the literature. For Fig. 3j, any predicted upstream driver–gene edge that connected to a cell for which that upstream driver was not significantly enriched was also trimmed (for example, only edges between IL-1 $\beta$  and nodes for genes in CD4<sup>+</sup> T cells were kept). Contextualization of these cell–cell signaling networks is further explored online: <http://shaleklab.com/resource/immune-dynamics-of-acute-hiv-infection>.

**Comparison to microarray data generated in acute SIV infection.** To contextualize our single-cell results against previously published datasets in SIV infection, we compared genes differentially expressed at peak viremia (compared to pre-infection) in all eight cell types studied in P1 to genes upregulated in rhesus macaques 0–180 d after SIV infection<sup>21</sup>. Any genes found to be differentially expressed (FDR-corrected  $q < 0.05$ ) in our data are depicted alongside any gene that demonstrated log<sub>2</sub>(fold change) upregulation at any timepoint in the Bosinger et al. microarray experiments<sup>21</sup> (Extended Data Fig. 3b). We note that compared to many of the ISGs upregulated in acute SIV infection, where upregulation persists for >2 weeks, most ISGs in P1 were only differentially expressed at the 1-week timepoint, indicating potential differences in the evolution of immune responses between humans and macaques.

**scRNA-seq of pDCs with SMART-seq2 and analysis.** Reverse transcription, WTA and library preparation of single pDCs in 96-well plates was performed as previously described<sup>33</sup>. Samples were sequenced on an Illumina NextSeq500/550 instrument with an Illumina 75 Cycle NextSeq500/550 v2 kit (Illumina FC-404-2005) using 30-bp paired-end reads. Given difficulties acquiring pDCs from pre-infection samples due to limited cell numbers, we sequenced pDCs from the peak IFN response and the 1-year timepoints in each participant. Reads ( $5 \times 10^5$  to  $3 \times 10^6$  per cell) were aligned to the hg38 (GENCODE v.21) transcriptome and genome using RSEM<sup>54</sup> and Tophat<sup>55</sup>, respectively. After trimming low-quality cells (cells with <25,000 mapped reads or <1,000 genes), the remaining cells had a median of 122,000 mapped reads and 2,866 genes. Pre-processing and differential expression analysis were conducted in Seurat<sup>56</sup> using the Wilcoxon rank-sum test. To test for differences in IFN responsiveness, participant-specific IFN response gene lists were used to generate scores in the pDCs using the AddModuleScore function in Seurat. The gene list used to score in each participant was chosen by including any gene that appeared at least twice in the modules that belonged to MM3 for that participant (see Extended Data Fig. 3i).

**Luminex and ELISA cytokine measurements.** Matching plasma cytokine levels were determined in duplicate using a multiplexed magnetic bead assay (catalog no. LHC6003M, Life Technologies) in accordance with the manufacturer's instructions. Briefly, a mixture of beads that were coated with anticytokine antibodies were prewashed and then incubated with the plasma samples. They were then co-incubated with a mixture of biotinylated detector antibodies followed by R-phycoerythrin (R-PE)-conjugated streptavidin. A magnetic separator was used to wash the beads between incubations. Fluorescence intensity was determined on a Bio-Plex 200 system. Concentrations of the cytokines in the samples were determined by interpolating on sigmoid four-parameter logistic regression standard curves.



Matching plasma soluble CD14 levels were measured using human CD14 DuoSet ELISA Kit (catalog no. DY383, R&D Systems) in accordance with the manufacturer's instructions. Briefly, a 96-well microplate was coated with anti-CD14 capture antibody overnight. The plates were blocked with reagent diluent for 1 h and then incubated with recombinant standards or plasma samples (diluted 1:600 in reagent diluent) for 1.5 h. They were further incubated with detection antibody for 1.5 h, followed by streptavidin–HRP for 20 min. The substrate was added for 20 min for color development. The reaction was stopped by adding stop solution. Optical density (OD) of each well was determined at 450 nm (corresponding ODs at 530 nm were subtracted for wavelength correction). The concentrations of soluble CD14 in the samples were determined by interpolating on a sigmoid four-parameter logistic regression standard curve. All incubations were done at room temperature.

**T cell receptor CDR3 pulldown and analysis.** To directly sequence the CDR3s from proliferating T cells assayed by Seq-Well, we applied a recently published TCR pulldown method<sup>56</sup> to WTA products from the 2-week, 3-week and 4-week timepoint samples from all four participants. Briefly, biotinylated capture probes from the TRBC region were annealed to melted WTA cDNA. Magnetic streptavidin beads were then used to pull down cDNA enriched for TRBC; this cDNA was subsequently amplified using KAPA HiFi Mastermix (Kapa Biosystems) and purified using 0.75× SeraPure beads to select for 0.8–1-kb sized DNA fragments. To select for sequences with full CDR3 regions, a pool of V-region primers was used to further amplify sequences of interest. Step-out PCR was used to add sequencing handles and the resulting libraries were sequenced on a NextSeq 550 using a 150-cycle NextSeq kit with 148 cycles for Read 1 (CDR3) and 20 cycles for Index 1 (BC + UMI). Sequences of the primers used are available in Tu et al.<sup>56</sup>. CDR3 consensus sequences were aligned and determined as outlined previously. Across the entire dataset, we detected ~50% of TCR-β chain CDR3s.

**Determining the cellular identity of early proliferating cytotoxic cells.** A recent scRNA-seq study on cytotoxic innateness characterized cytotoxic γδT and NK cells in healthy humans, noting basal levels of *TRDC* in both cell types<sup>46</sup>. To determine whether *TRDC*<sup>+</sup>*FCGR3A*<sup>+</sup> cells were γδT or NK cells, we scored them, as well as nonproliferating CTLs and NK cells, against gene signatures described in that study (Extended Data Fig. 10g). Based on score similarity to NK cells, and the relative downregulation of CD3 compared to the other proliferating T cell subsets (FDR-corrected Wilcoxon rank-sum test; *CD3D*: log(FC) = −0.895, *q* = 2.7 × 10<sup>−42</sup>; *CD3G*: log(FC) = −0.923, *q* = 8.9 × 10<sup>−37</sup>), we determined cluster 4 (lilac) to be proliferating NK cells.

**Statistics.** WGCNA module significance was tested using a permutation test (*n* = 10,000) on dissimilarity values and compared to the distribution of true values using a one-sided Mann–Whitney *U*-test. After FDR correction, modules that failed <500 tests (*P* < 0.05) were considered significant. Participant-specific modules were tested for temporal variation in score using a two-sided Mann–Whitney *U*-test with 1,000 subsamplings. Modules discovered across all participants were tested for temporal variation with binned samples using an ANOVA with two models: (1) null ~1 + participant; and (2) alternative ~1 + participant + time.bin (*P* < 0.05). Gene-set analysis was performed on GMs using IPA (Qiagen). For the differential expression analysis in monocytes, we performed hypergeometric tests using gene lists from MSigDB v.6.2 and corrected for multiple-hypothesis testing using the FDR procedure. Differential expression analysis between groups of single cells was performed using either a two-sided Wilcoxon rank-sum test or the 'bimod' test as implemented in Seurat.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The expression matrix and associated metadata can be visualized and downloaded from the Alexandria Project, a Bill and Melinda Gates Foundation-funded portal (part of the Single-Cell Portal hosted by The Broad Institute of MIT and Harvard): [https://singlecell.broadinstitute.org/single\\_cell/study/SCP256](https://singlecell.broadinstitute.org/single_cell/study/SCP256). It is also available for download, alongside accompanying supplementary material, at our website: <http://shaleklab.com/resource/immune-dynamics-of-acute-hiv-infection>. De-identified raw data are available on request through the corresponding authors, given the at-risk nature of people with HIV. The raw data will also be submitted to dbGaP, pending IRB approval.

## Code availability

The custom code used for GM discovery analysis is in the Supplementary Software.

## References

- Dong, K. L. et al. Detection and treatment of Fiebig stage I HIV-1 infection in young at-risk women in South Africa: a prospective cohort study. *Lancet HIV* **5**, e35–e44 (2018).

- Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
- Ordovas-Montanes, J. et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649 (2018).
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
- Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast interpolation-based *t*-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **16**, 243 (2019).
- Gutierrez-Arcelus, M. et al. Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nat. Commun.* **10**, 687 (2019).
- Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
- Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
- Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
- Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
- Kumar, L. & E. Futschik, M. Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics* **2**, 5–7 (2007).
- Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).
- Trombetta, J. J. et al. Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 1–17 (2014).
- Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
- Tu, A. A. et al. TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures. *Nat. Immunol.* **20**, 1692–1699 (2019).

## Acknowledgements

We thank the individuals who participated in the FRESH study, as well as FRESH and HIV Pathogenesis Programme staff; M. Waring, N. Bonheur and E. Koscher for flow cytometry and sorting services through the Ragon Institute Imaging Core-Flow Cytometry Facility; A. Piechocka-Trocha for sample handling and reagent preparation; M. Cole and N. Yosef for advice on computational methodology; D. Kwon, S. Bloom and M. Hayward for participant intake and sexually transmitted infection data; M. Carrington and A. Bashirova for HLA genotyping; O. Nduumego, S. Rasehlo and N. A. Akilimali for support with Luminex and ELISA; H. Kloverspris, D. Lingwood and S. Pillai for insightful discussions; and R. Leeson for copy editing. This work was supported, in part, by the Searle Scholars Program, the Beckman Young Investigator Program, the Pew-Stewart Scholars Program for Cancer Research, a Sloan Fellowship in Chemistry, the National Institutes for Health (2U19AI089992, 2R01HL095791, 1U54CA217377, 2P01AI039671, 5U24AI118672, 2R01HG006193, 1R01AI138546, 1R37AI067073, 1R01HL134539, 1R01DA046277, 1K08AI118538, U01AI100663, U01AI144462 and 1R01AI145305), the Bill and Melinda Gates Foundation (OPP1066973 and INV-002703), an NSF Graduate Student Fellowship Award, the Hugh Hampton Young Memorial Fund Fellowship, the Damon Runyon Cancer Research Foundation (HHMI Fellow DRG-2274-16), Gilead Sciences (grant ID 00406), the International AIDS Vaccine Initiative (UKZNRSA1001), the South African National Research Foundation (grant no. 64809), the Witten Family Foundation, Dan and Marjorie Sullivan Foundation, the Mark and Lisa Schwartz Foundation, Ursula Brunner, the AIDS Healthcare Foundation and the Harvard University Center for AIDS Research (CFAR, P30 AI060354, which is supported by the following institutes and centers co-funded by and participating with the United States National Institutes of Health: NIAID, NCI, NICHD, NHLBI, NIDA, NIMH, NIA, FIC and OAR), the Sub-Saharan African Network for TB/HIV Research Excellence and DELTAS Africa Initiative (grant no. DEL-15-006). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (grant nos. 107752/Z/15/Z and 210662/Z/18/Z) and the United Kingdom Government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the United Kingdom Government.

## Author contributions

S.W.K., J.O.-M., E.B.W., A.L., Z.M.N., T.N., B.D.W. and A.K.S. conceived and designed experiments. S.W.K., T.P.A., S.K.N., B.B. and A.K.S. developed the computational framework. S.W.K., T.P.A., D.M.M., S.L.C. and V.N.M. performed scRNA-seq, flow cytometry, Luminex, ELISA and TCR pulldown assays. S.W.K. and T.P.A. analyzed the

single-cell RNA-seq data. A.A.T., J.C.L. and A.K.S. designed the TCR pulldown method. S.W.K. and A.A.T. aligned and analyzed TCR pulldown data. S.W.K., T.P.A., D.M.M., J.O.-M., C.G.K.Z., S.K.N., A.L., Z.M.N., T.N., B.D.W. and A.K.S. interpreted data. N.I., M.D., A.M., K.L.D., Z.M.N., T.N. and B.D.W. established and/or currently maintain the FRESH study. S.W.K., T.P.A., D.M.M., J.O.-M., C.G.K.Z., Z.M.N., T.N., B.D.W. and A.K.S. wrote the manuscript with input from all authors.

### Competing interests

The authors declare no competing interests.

### Additional information

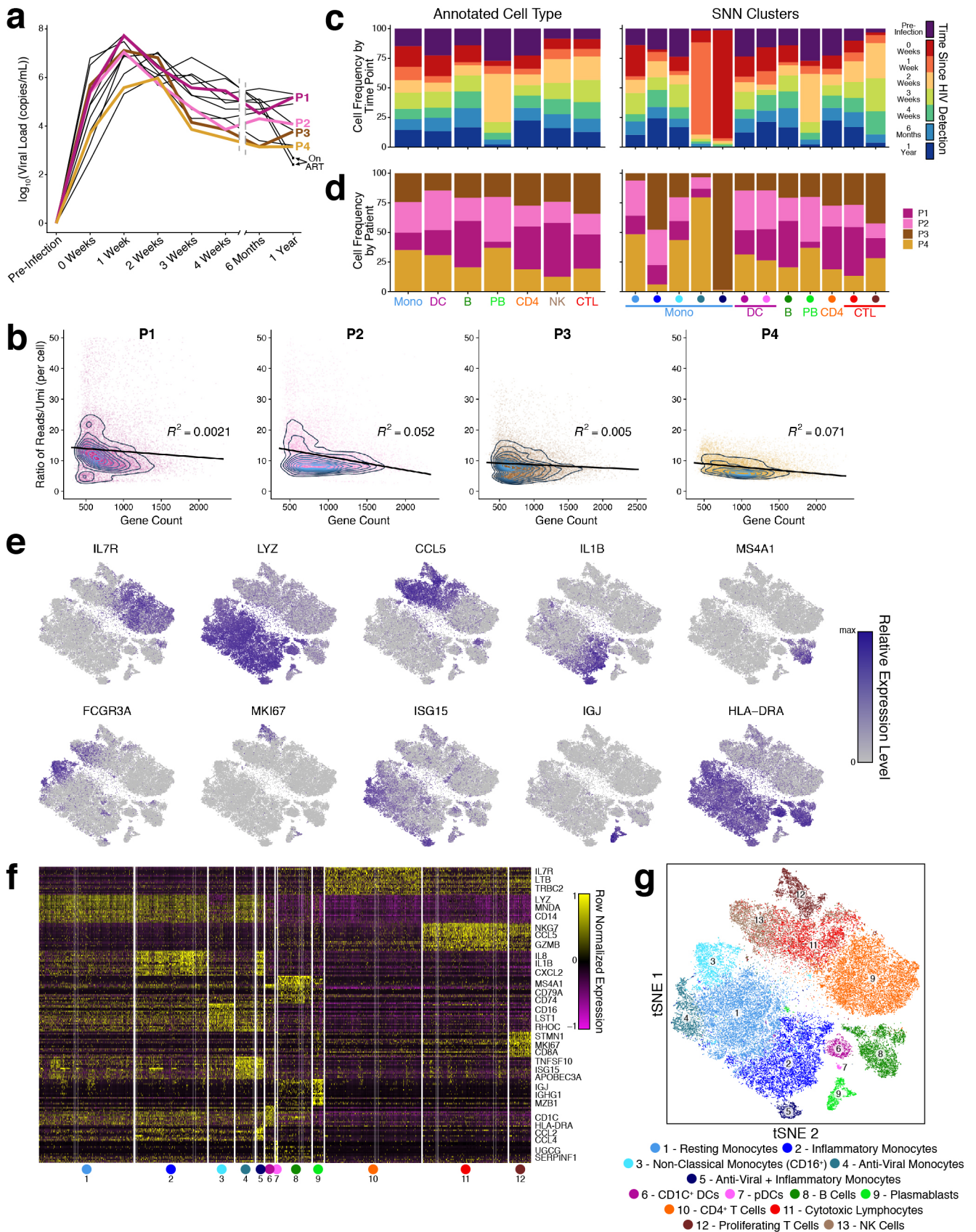
**Extended data** is available for this paper at <https://doi.org/10.1038/s41591-020-0799-2>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41591-020-0799-2>.

**Correspondence and requests for materials** should be addressed to B.D.W. or A.K.S.

**Peer review information** Alison Farrell is the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

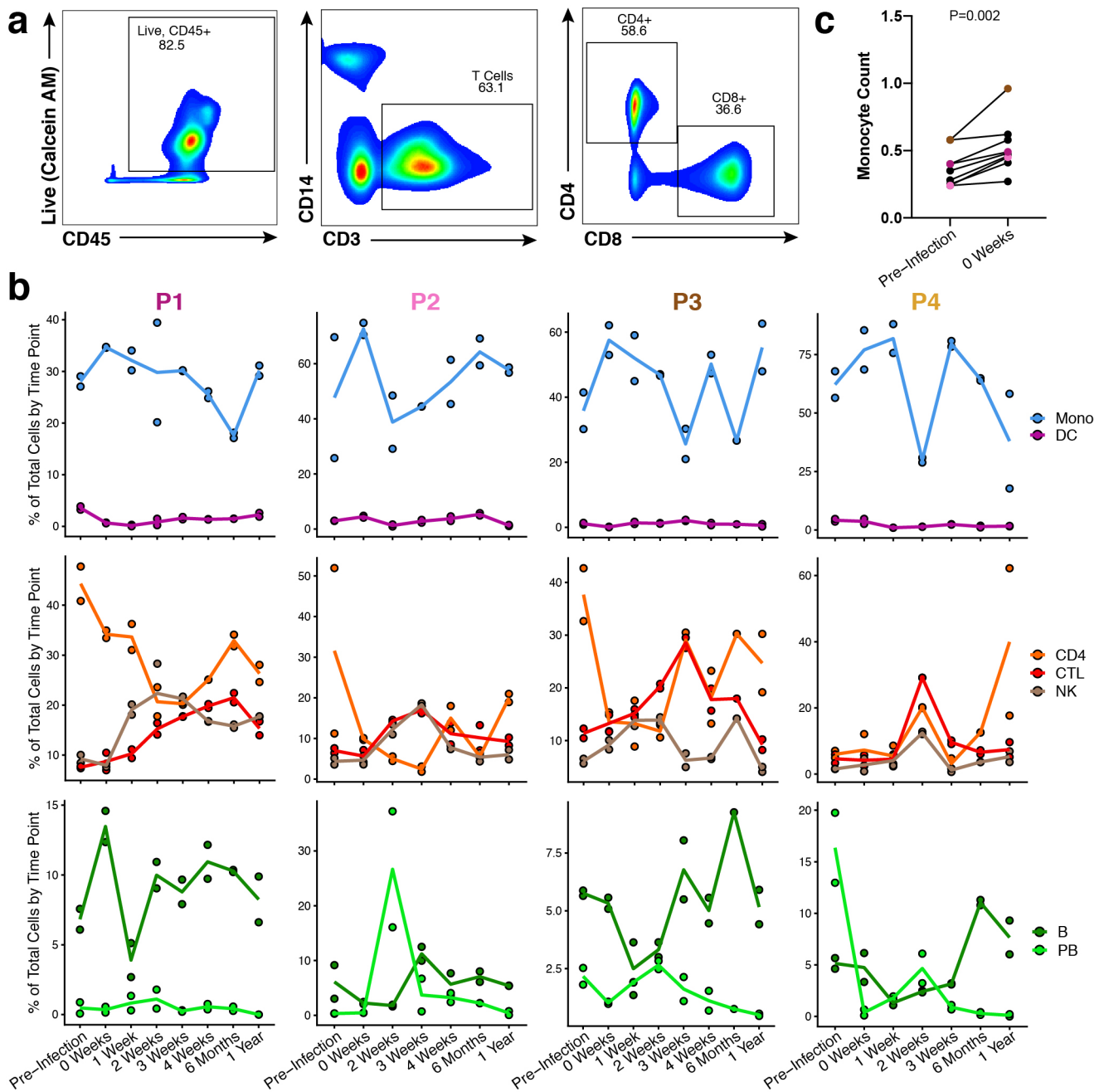
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



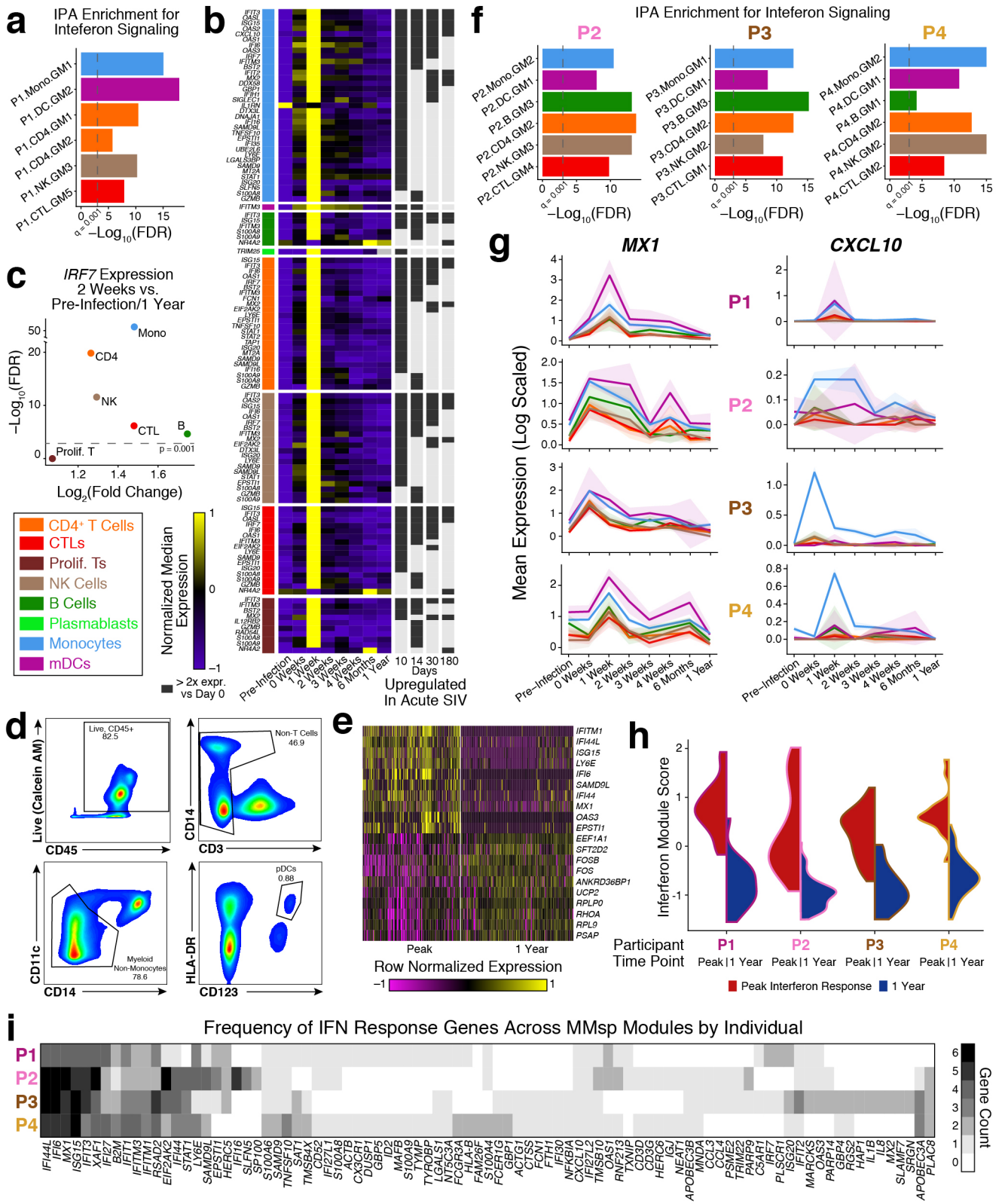
Extended Data Fig. 1 | See next page for caption.



**Extended Data Fig. 1 | Participant and time point breakdown by cluster and cluster annotation.** **a**, Plasma viral load for the 12 participants studied in Ndhlovu et al.<sup>37</sup> with the four individuals characterized here annotated. **b**, Average ratio of number of reads per number of UMIs measured per single cell. R-squared values reflect variance described by a linear model. Number of cells per participant: P1 – 15,259; P2 – 13,128; P3 – 15,927; P4 – 15,425. **c**, Time point and **d**, participant cell frequency by annotated cell type (left) and shared-nearest neighbors (SNN) clusters (right). **e**, Relative expression levels of exemplary marker genes used for cell type identification projected on the FitSNE. **f**, Heatmap of the top 10 genes differentiating each SNN cluster. Up to 500 random cells are depicted for each cluster. Two-sample Wilcoxon rank sum test; see Supplementary Table 2 for cell numbers. **g**, tSNE embedding of dataset colored by SNN cluster and annotated based on genes in **e**, **f**, and Supplementary Table 3.



**Extended Data Fig. 2 | Cell frequency by participant and cell type.** **a**, Representative gating scheme for CD4<sup>+</sup> and CD8<sup>+</sup> T cells.  $n = 30$  samples. **b**, Cell type frequency calculated from total cells measured within an array. Lines represent average between duplicate arrays. Columns are separated by participant. **c**, Monocyte count from whole blood for 12 participants studied in Ndhlovu et al.<sup>37</sup>. Two-sided paired t test,  $n = 12$ .

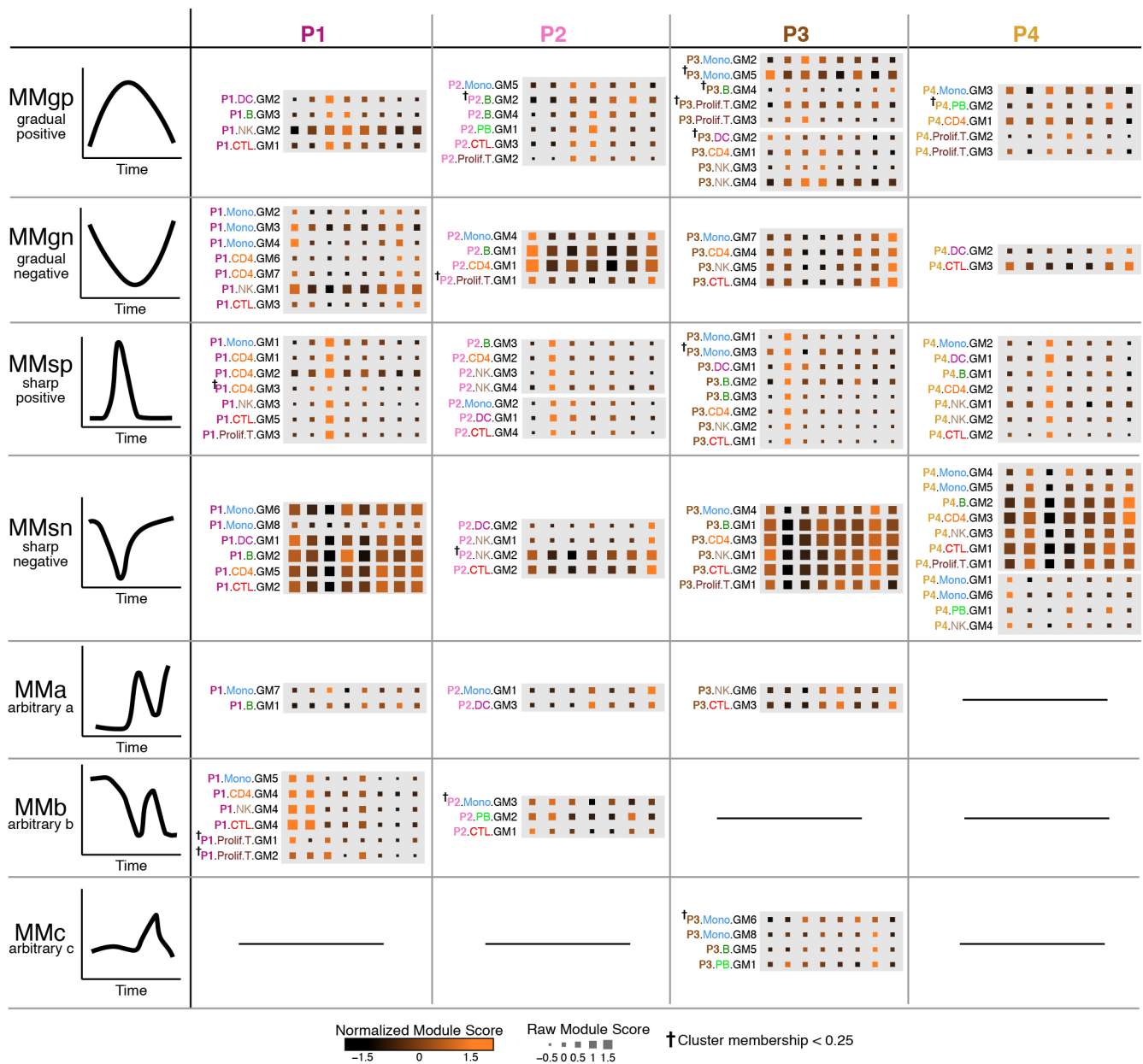


Extended Data Fig. 3 | See next page for caption.

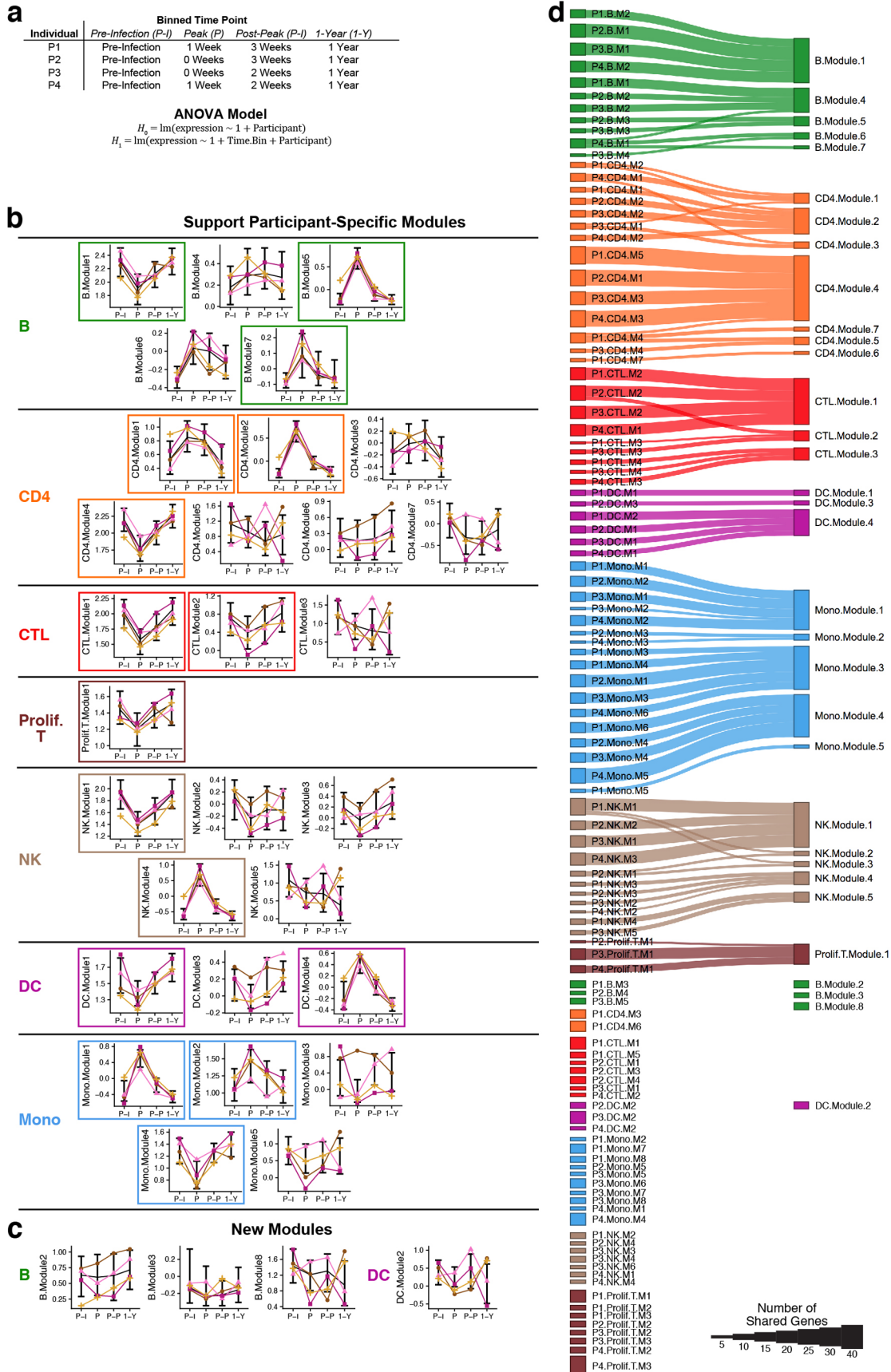


**Extended Data Fig. 3 | Gene modules that align near peak viremia and are differentially expressed in pDCs are enriched for response to interferon.**

**a**, Enrichment of modules from P1 in Fig. 2b against the IPA Interferon Signaling canonical pathway; FDR corrected Right-Tailed Fisher's Exact Test. See Supplementary Table 4 for gene lists and their sizes. **b**, Heatmap of median expression of genes upregulated in PBMCs from SIV infection of rhesus macaques (n=8) compared to day 0 (fold change > 2) in Bosinger et al.<sup>21</sup> depicted independently for each cell type in P1. Only genes differentially expressed across all time points by ANOVA in a given cell type are shown. **c**, Differential expression results for *IRF7* in each cell type (except plasmablasts and mDCs which do not have enough cells to test, n < 4) between cells from 2 weeks and pre-infection + 1 year; implemented using the "bimod" likelihood ratio test in Seurat<sup>44</sup>. See Supplementary Table 2 for cell numbers per cell type per timepoint. **d**, Representative gating scheme for single-cell pDC sorts. n = 8 samples. **e**, Heatmap of genes differentially expressed between pDCs captured at the same timepoints as peak interferon responses and 1-year post HIV infection. Two-sided Wilcoxon rank sum test; number of cells per timepoint: Peak - 159; 1-Year - 184. **f**, Same as **a** for the modules in Fig. 2d for P2, P3, and P4. See Supplementary Table 2 for cell numbers per cell type per timepoint. **g**, Mean gene expression (log scaled) of *MX1* and *CXCL10* over time in each individual separated by cell type. Shaded area denotes 95% confidence interval of the mean. **h**, Scoring of pDCs in each participant using a core interferon signature specific to that participant. Number of cells per participant and time point: P1 Peak - 31; P1 1-Year - 79; P2 Peak - 48; P2 1-Year - 20; P3 Peak - 42; P3 1-Year - 28; P4 Peak - 16; P4 1-Year - 57. **i**, Heatmap of gene frequency across interferon response GMs in each participant.



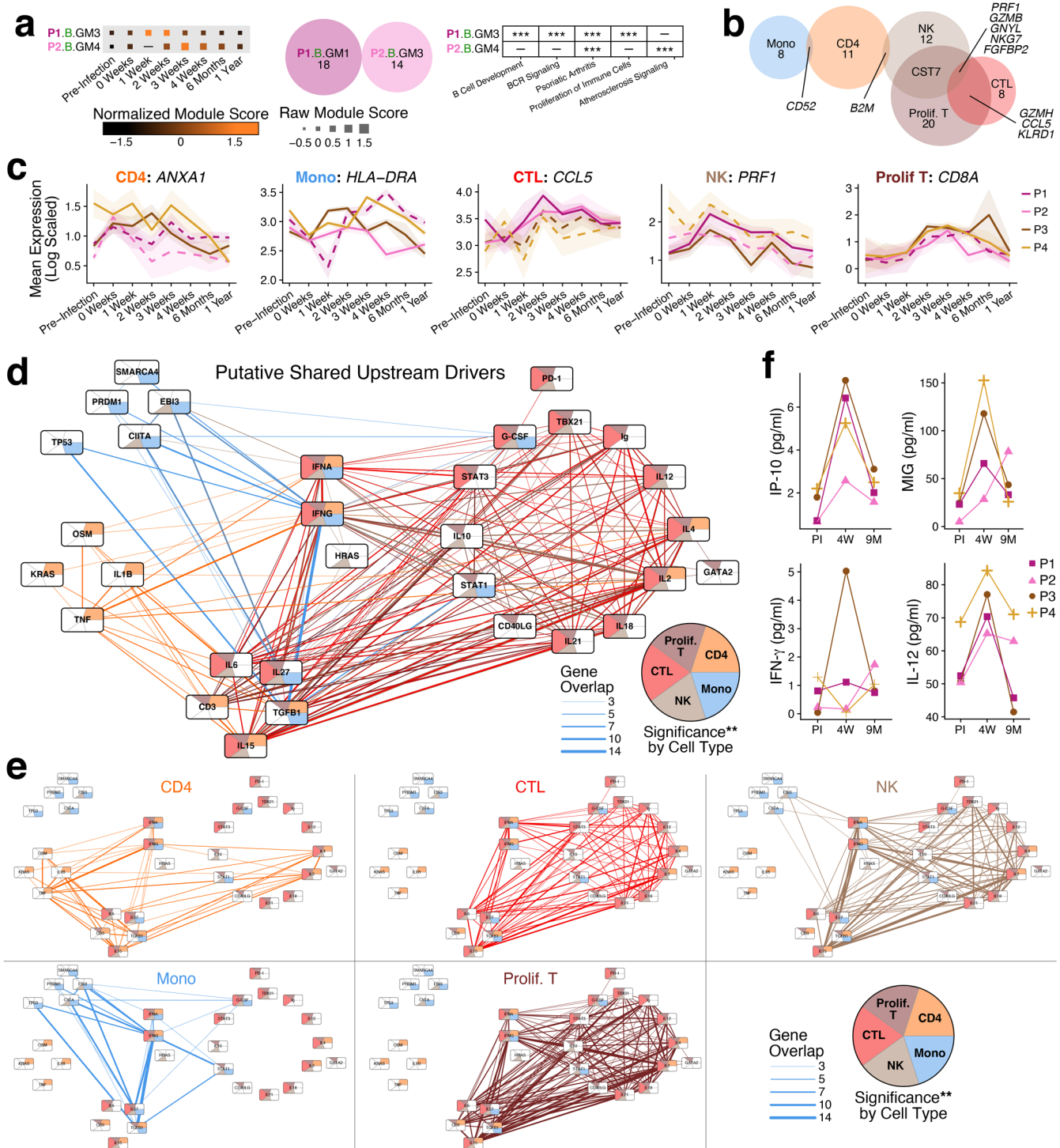
**Extended Data Fig. 4 | All significant temporally variant modules in all participants grouped by fuzzy c-means clustering.** Modules grouped by fuzzy c-means clustering (see Methods for choice of  $c$ ) reside in the same gray box. Each group of modules, or meta module (MM), were then aligned across participants based on overall temporal trend (left column). Some participants had multiple MM with similar temporal dynamics and were grouped within the same MM. Since fuzzy c-means clustering assigns membership values to each member of a cluster, we report any modules that demonstrated low cluster membership with †.



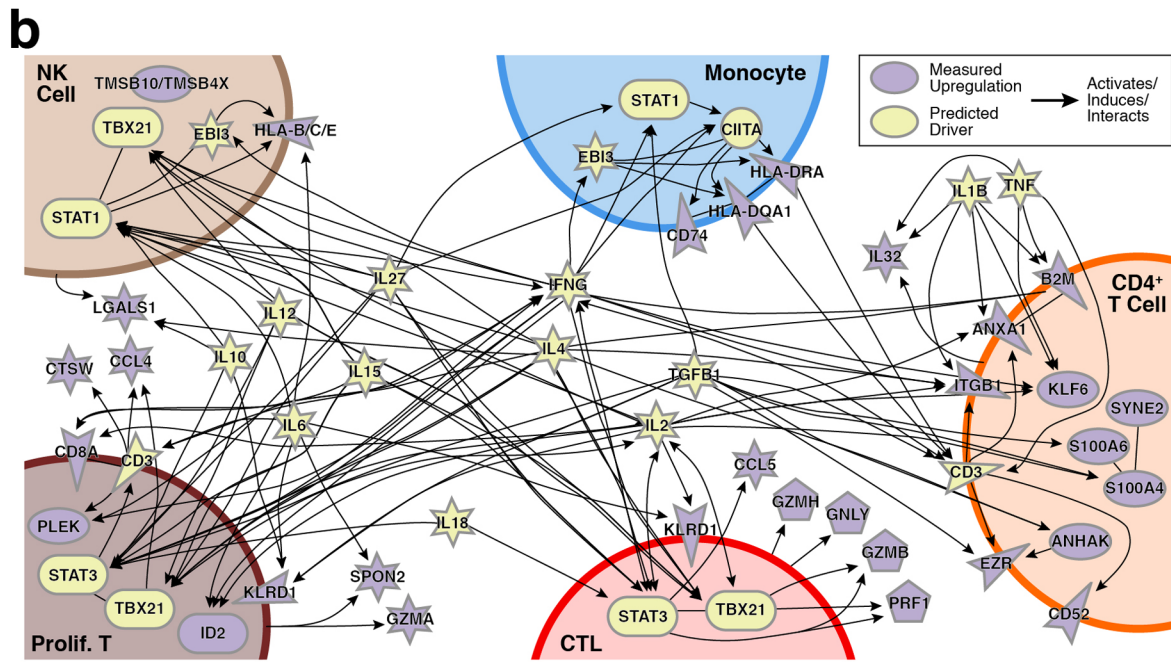
Extended Data Fig. 5 | See next page for caption.



**Extended Data Fig. 5 | Cross-participant module discovery recapitulates aspects determined by calculating participant-specific modules. a**, Timepoints chosen for testing significant changes in module expression over time across all participants. Timepoints were chosen based on peak expression of modules in MM1 and MM3. We used an ANOVA model to account for participant-specific features. **b**, Significant cross-participant modules that map to participant-specific modules (share at least 5 genes), separated by cell type. Median module expression is plotted for each module split by participant. Error bars depict the upper and lower quartiles for all cells across all four individuals at each time point. Boxed modules demonstrate consistent directional trends in expression between each pair of timepoints in at least 3/4 participants. P-I = pre-infection; P = peak; P-P = post-peak; 1-Y = 1 year. See **a** and Supplementary Table 2 for cell numbers per cell type per participant. **c**, Significant cross-participant modules that do not correspond to participant-specific modules. **d**, Sankey diagram demonstrating the gene overlap between participant specific modules (left) and cross-participant modules (right). Node size correlates with the number of genes within the module and edge width correlates with the number of shared genes between modules. Only overlaps consisting of  $\geq 5$  genes have edges depicted.



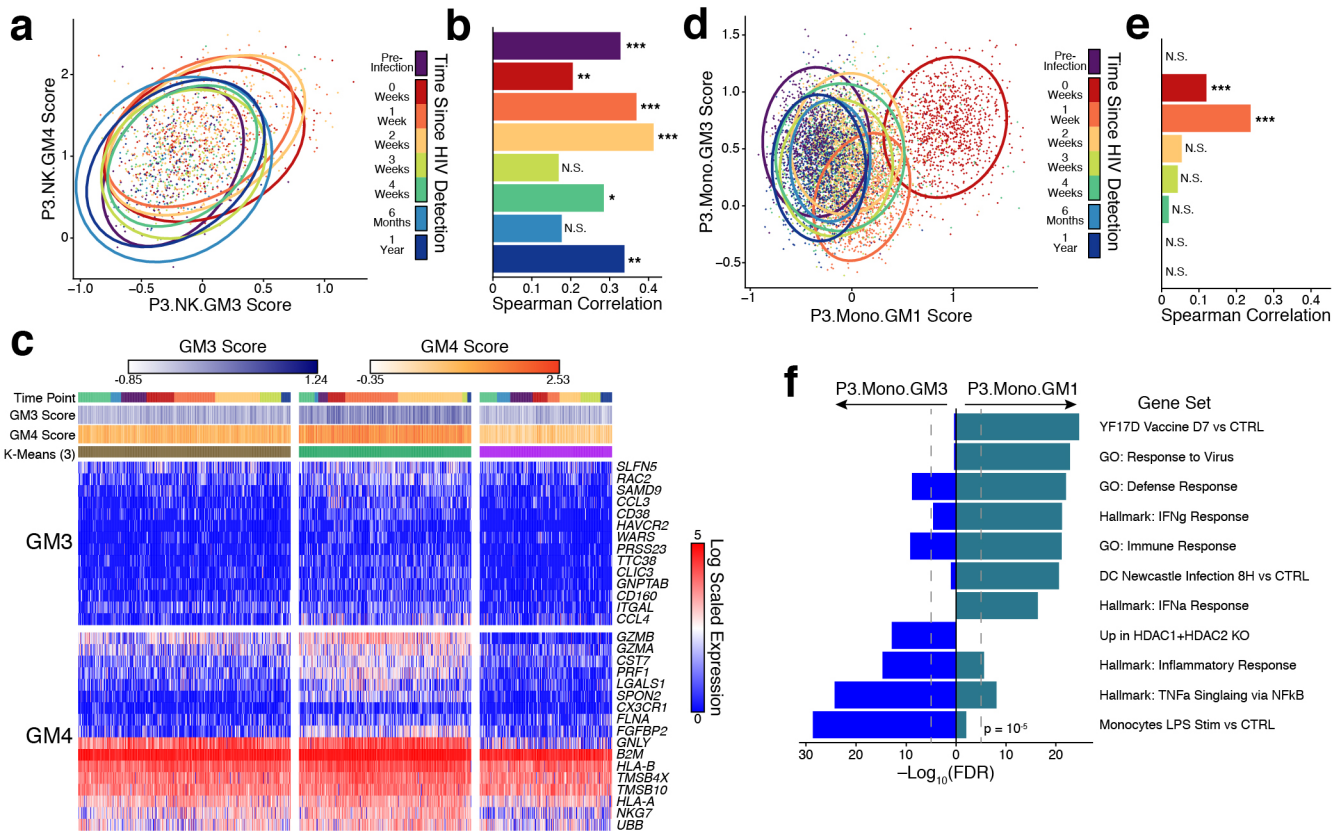
**Extended Data Fig. 6 | Sustained B cell modules and shared genes and upstream drivers between participants.** **a**, B cell modules in MM1 with high cluster membership. **b**, Euler diagram of conserved overlapping genes between cell types from Fig. 3e-i, see Supplementary Table 8. **c**, Mean expression of *ANXA1*, *HLA-DRA*, *CCL5*, *PRF1*, and *CD8A* in CD4<sup>+</sup> T cells, monocytes, CTLs, NK cells, and proliferating T cells, respectively. Shaded area denotes 95% confidence interval of the mean. Participants who did not have modules with shared temporal expression pattern as outlined in Fig. 3e-i are shown as dashed lines. See Supplementary Table 2 for cell numbers per timepoint per participant. **d**, Network of predicted upstream drivers of modules in Fig. 3e-i. Nodes are colored by significance in each cell-type. Edge width and color reflect the number of shared genes (width) in the gene sets of the upstream drivers for a given cell-type (color; see Methods). **e**, **d** displayed with only the edges from a given cell type. **f**, Luminex measurements of IP-10, MIG, IFN- $\gamma$ , IL-12 in matching plasma samples. Points are averages of duplicate measurements. PI = pre-infection; 4W = 4 weeks post-detection; 9M = 9 months post-detection.



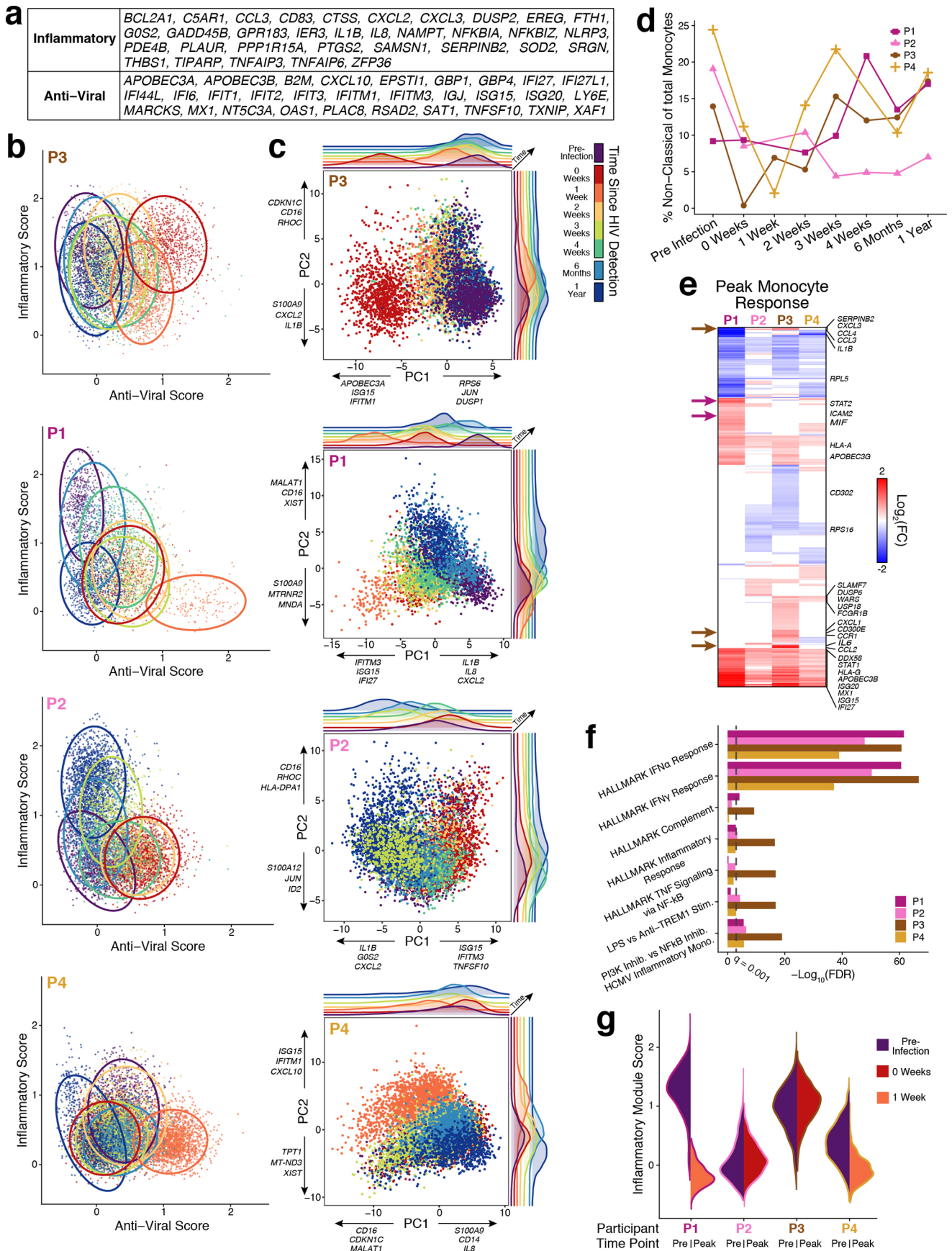
Extended Data Fig. 7 | See next page for caption.

**Extended Data Fig. 7 | Putative upstream drivers highlight variable response dynamics and cell-cell signaling.** **a**, Median gene set scores for significantly temporally variant ( $p < 0.05$ ) upstream drivers in all participants. Gray boxes indicate that the upstream driver was not significantly variant in that cell type and participant. Right-Tailed Fisher Exact Test; see Supplementary Table 2 for cell numbers per timepoint per cell type. **b**, Putative cell-cell network described in Fig. 3j, but all nodes and connections depicted. Nodes represent genes with either measured upregulation in the modules in Fig. 3e–i or predicted drivers from IPA. Edges were drawn from connections nominated by IPA and curated from the literature.



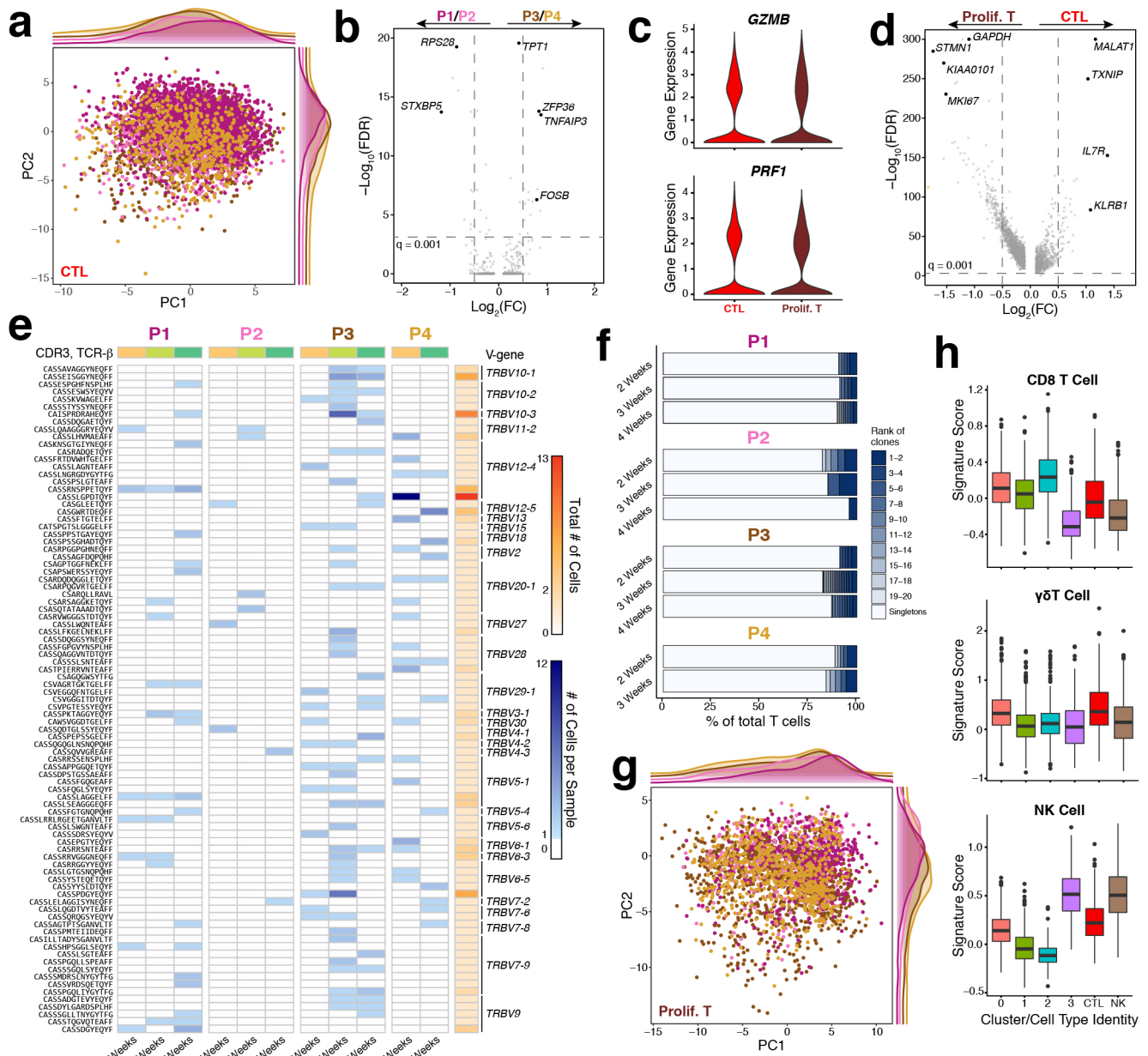


**Extended Data Fig. 8 | Two cases of similar temporal modules in NK cells and monocytes: variable correlation and variable co-expression.** **a**, Module scores in NK cells for NK GM3 and NK GM4 in P3. Ellipses drawn at 95% confidence interval for cells from each timepoint. See Supplementary Table 2 for cell numbers per timepoint. **b**, Correlation (spearman's rho) between the scores for NK GM3 and NK GM4 at each timepoint. Two-sided asymptotic t approximation; FDR corrected q-value: N.S = not significant; \* q < 0.05; \*\* q < 0.01, \*\*\* q < 0.001. **c**, Gene expression heatmap of all NK cells in P3 for those genes in NK GM3 and NK GM4. Cells are separated based on k-means (k=3) over the depicted genes. **d** & **e**, Same as in **a** & **b** but for monocyte modules Mono GM1 and Mono GM3 in P3. **f**, Gene set enrichment analysis of the genes in Mono GM1 and Mono GM3 against the following MSigDB collections: Hallmark, C2, C3, C5, and C7. FDR corrected hypergeometric test; number of genes: GM1 - 33, GM3 - 52.



Extended Data Fig. 9 | See next page for caption.

**Extended Data Fig. 9 | Participants demonstrate diverse monocyte responses prior to and immediately after HIV detection, wherein one participant who goes on to control infection presents a poly-functional subset of monocytes.** **a**, Inflammatory and anti-viral genes shared between participants (present in at least two modules). **b**, Inflammatory and anti-viral scores of monocytes in each participant using gene lists in **a**. Ellipses drawn at 95% confidence interval for cells from each timepoint. See Supplementary Table 2 for cell numbers per timepoint per cell type. **c**, Principal component analysis (PCA) of all monocytes from each participant. Density of cells in PC1 vs PC2 space annotated by timepoint are depicted, and 3/5 of the top loading genes for PC1 and PC2 are also annotated. **d**, Percent of Non-Classical (CD16<sup>+</sup>/FCGR3A) monocytes of total monocytes as a function of time in each participant. Percentage calculated from cluster assignment (see Extended Data Fig. 1d). **e**, Heatmap of differentially expressed genes (FDR corrected  $q < 0.05$  in at least one participant) between monocytes at the peak response timepoint (0 weeks/1 week) vs. pre-infection. Arrows indicate genes specific to P3 (dark-brown) and P1 (violet). Two-sided Wilcoxon rank sum test; see Supplementary Table 2 for cell numbers per timepoint per participant. **f**, Enriched pathways for the differentially expressed genes in **e**, using the MSigDB Hallmark Gene Sets. Hypergeometric test; FDR corrected  $q$  values; number of differentially expressed genes: P1 - 1,350; P2 - 436; P3 - 857; P4 - 514. **g**, Violin plot of the Inflammatory Module Score (see **a** for genes in the module) for monocytes at pre-infection (Pre) and peak transcriptional response time points (Peak) in each participant. See Supplementary Table 2 for cell numbers per timepoint per participant.



**Extended Data Fig. 10 | Non-proliferating and proliferating cytotoxic T cells.** **a**, Principal component analysis of non-proliferating CTLs with participant density annotated along PC1 and PC2. Number of cells: P1 - 1828; P2 - 968; P3 - 1503; P4 - 670. **b**, Volcano plot of differentially expressed genes between the participants who control (P3/P4) and those who do not (P1/P2); implemented using a two-sided Wilcoxon rank sum test. See **a** for cell numbers. **c**, Expression of *GZMB* and *PRF1* in all CTLs and proliferating T cells. Number of cells: CTL - 4,969; Proliferating T cells - 2,639. **d**, Volcano plot of differentially expressed genes between non-proliferating CTLs and proliferating T cells; implemented using a two-sided Wilcoxon rank sum test; see **c** for cell numbers. **e**, Heatmap of detected TCR- $\beta$  CDR3s in proliferating T cell clusters 0 & 1 at 2 weeks, 3 weeks, and 4 weeks post-HIV detection. See Supplementary Table 10 for full list of CDR3s. **f**, Distribution of ranked TCR- $\beta$  CDR3 clones (by total cell number) and singletons measured from all T cells (i.e. CD4<sup>+</sup> T cells, CTLs, and proliferating T cells) detected at 2 weeks, 3 weeks, and 4 weeks post-HIV detection in at least two single cells in each participant. Here, except for the singletons, each sliver represents the percentage of CDR3s ascribed to the top  $n-n+1$  clones for that timepoint and participant. **g**, Same as in **a** but over proliferating T cells. Number of cells: P1 - 483; P2 - 273; P3 - 1193; P4 - 690. **h**, CD8 T cell (top),  $\gamma\delta$ T cell (middle), and NK cell (bottom) scores for each proliferating T cell cluster (see **g** for cell numbers) and 500 randomly sampled CTLs, and 500 randomly sampled NK cells. Signatures were established from differential expression over the single-cell dataset published by Gutierrez-Arcelus et al.<sup>46</sup>. See Supplementary Table 10 for all differentially expressed genes and signature score gene lists. Box plots features depict: minimum = 25<sup>th</sup> percentile - 1.5 \* inter-quartile range (IQR; smallest value within); lower = 25<sup>th</sup> percentile; center = 50<sup>th</sup> percentile; upper = 75<sup>th</sup> percentile; maximum = 75<sup>th</sup> percentile + 1.5 \* IQR (largest value within).



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Sequencing data was collected using a NextSeq500. Flow cytometry data was collected on a BD SORP FACSAria II cell sorter using BD FACSDiva software (version 8.0.2).

Data analysis

Reads were aligned against hg19 (Ensembl: Homo\_sapiens.GRCh37.74) using Drop-seq Tools (v.1.13). After alignment, data were processed using R (v3.4.4), Seurat (v.2.3.4), WGCNA (v1.66), Mfuzz (v.2.38.0), and custom scripts provided in the Supplemental Software. Gene Set Analysis was performed using Ingenuity Pathway Analysis (Qiagen, Summer 2019 Release) and MSigDB (v6.2). Flow cytometry data was analyzed using Flowjo (v10.4).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The expression matrix and associated meta data can be visualized and downloaded from The Alexandria Project, a Bill and Melinda Gates Foundation funded portal (part of the Single Cell Portal hosted by The Broad Institute of MIT and Harvard): [https://singlecell.broadinstitute.org/single\\_cell/study/SCP256](https://singlecell.broadinstitute.org/single_cell/study/SCP256). De-identified raw data is available upon request through the corresponding authors given the at-risk nature of HIV infected persons. The raw data will also be submitted to dbGaP pending IRB approval.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was selected based on limited sample availability from the FRESH study (n=4). We chose participants for which we had at least 7/8 of the following time points with samples remaining: pre-infection, HIV detection (0 weeks), 1 week, 2 weeks, 3 weeks, 4 weeks, 6 months, and 1 year post HIV detection. With longitudinal sampling, we have 30 independent samples. Within an individual, we have 7-8 samples, which provided robust statistics based on cell number replicates (~14k-16k cells per participant).
Data exclusions	Cells with fewer than 750 or more than 6,000 unique transcript reads were removed from analysis as low quality cells or potential doublets; these cells would add unwanted noise to downstream analysis. Any cells whose measured transcriptome was greater than 20% mapping to mitochondrial genes were also removed, as these have been shown to be low quality cells. Both exclusion criteria were pre-established for this data analysis.
Replication	Every sample was processed by Seq-Well in duplicate (except for one: P1 - 6 months, due to experiment failure). In the majority of cases, cell frequencies were confirmed by duplicate measurements. Due to sample limitations and current standard of care for HIV infection, we were unable to assay additional individuals to further validate our findings above the participants described here (n=4).
Randomization	All participants contracted HIV-infection in this study. Pre-infection samples were determined by negative RT-PCR of viral RNA and other HIV tests. All samples classified as post-HIV detection were characterized by positive RT-PCR of viral HIV RNA in the plasma. Controller phenotype was determined by a HIV viral load of fewer than 1,000 viral copies/mL for 3 consecutive tests, spaced 6 months apart, starting at 1.8 years post initial infection.
Blinding	Blinding was not relevant as there was no placebo group. All HIV- samples were collected prior to infection as participant-internal controls.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

### Antibodies used

Alexa Fluor 700 - CD45; Biolegend; clone 2D1; Cat# 368514; Lot B 248834; 1:200 Dilution  
 BUV737 - CD3; BD Biosciences; clone UCHT1; Cat# 564307; Lot 7335654; 1:200 Dilution  
 BV711 - CD4; Biolegend; clone OKT4; Cat# 317440; Lot B249480; 1:200 Dilution  
 BUV395 - CD8; BD Biosciences; clone RPA-T8; Cat# 563795; Lot 7069910; 1:200 Dilution  
 BV605 - CD14; Biolegend; clone M5E2; Cat# 301834; Lot B247581; 1:200 Dilution  
 BV510 - HLA-DR; BD Biosciences; clone G46-6; Cat# 563083; Lot B249262; 1:200 Dilution  
 BV650 - CD123; Biolegend; clone 6H6; Cat# 306020; Lot B241226; 1:200 Dilution  
 APC-Cy7 - CD11c; Biolegend; clone 2D1; Cat# 337218; Lot B254813; 1:200 Dilution

### Validation

Alexa Fluor 700 - CD45. Validated by Biolegend on human peripheral blood lymphocytes (mouse IgG1 K control).  
 BUV737 - CD3. Validated by BD Biosciences on human peripheral blood lymphocytes (mouse IgG1 K control).  
 BV711 - CD4. Validated by Biolegend on human peripheral blood lymphocytes (mouse IgG1 K control).  
 BUV395 - CD8. Validated by BD Biosciences on human peripheral blood lymphocytes (mouse IgG1 K control).  
 BV605 - CD14. Validated by Biolegend on human peripheral blood lymphocytes (mouse IgG1 K control).  
 BV510 - HLA-DR. Validated by BD Biosciences on human peripheral blood lymphocytes and monocytes (mouse IgG1 K control).

BV650 – CD123. Validated by Biologend on human peripheral blood leukocytes (mouse IgG1 K control).  
 APC-Cy7 – CD11c. Validated by Biologend on human peripheral blood granulocytes (mouse IgG1 K control).  
 All validation information is available on the Biologend and BD Biosciences websites:  
<https://www.biologend.com/>  
<https://www.bdbiosciences.com/en-us>

## Human research participants

Policy information about [studies involving human research participants](#)

### Population characteristics

For P1, P2, P3, P4 respectively:  
 Age at HIV Detection: 24, 21, 24, 21.  
 Gender: F, F, F, F.  
 Feibig Stage at HIV Detection: I, I, I, I.  
 HIV Status at enrollment: HIV-, HIV-, HIV-, HIV-.  
 Controller Phenotype: No, No, Yes, Yes.  
 HLA-A: 24:02/29:02, 68:01/68:02, 02:05/66:01, 01:01/66:01.  
 HLA-B: 07:02/44:03, 57:02/58:02, 14:01/39:10, 39:10/81:01.  
 HLA-C: 07:01/07:02, 06:02/18, 8:04/12:03, 12:03/18.

### Recruitment

Eligible women were HIV uninfected, aged 18–23 years, sexually active, not pregnant, non-anemic (hemoglobin  $\geq 10$  g/L), without other barriers to participation (serious chronic illness, enrollment in another study, or family responsibilities), and gave written consent to enrollment. We targeted disadvantaged, at-risk women, favoring the enrollment of those who were unemployed and not attending school. Participants were recruited at local sites frequented by young people, including cafes, nightclubs, and shopping malls. (see Dong et al., The Lancet, 2018). Selection bias in participant recruitment is possible given the monetary incentive to join the study. This may have led to increased numbers of women seeking socio-economic gain. Nevertheless, in our study of  $n=4$ , we cannot comment on how any biases may impact the results herein, especially given our pre-infection control samples.

### Ethics oversight

Biomedical research ethics committee of the University of KwaZulu-Natal and the IRB of Massachusetts General Hospital.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

#### Sample preparation

PBMCs were purified from whole blood by ficoll phase separation, and subsequently frozen for long term LN2 storage. Frozen peripheral blood mononuclear cells (PBMCs) were thawed and washed twice with warm RPMI supplemented with 10% fetal bovine serum. Next, the cells were resuspended in FACS buffer (PBS supplemented with 1% FBS) and stained with antibodies on ice for 30 minutes in FACS buffer. Afterward, the cells were washed and stained with the viability stain Calcein Blue, AM (Invitrogen, C34853) for 15 minutes on ice. Finally, the stained cells were washed twice with FACS buffer.

#### Instrument

BD SORP FACSAria II

#### Software

BD FACSDiva (v8.0.2) was used to collect flow cytometry data. Subsequent analysis was performed using FlowJo v10.4.

#### Cell population abundance

Up to 250,000 viable immune cells (CD45+Calcein Blue+) were sorted into 1 ml of RPMI + 10% FBS for Seq-Well. For Smart-Seq2 of pDCs, cells were directly sorted into 10  $\mu$ l of RLT (Qiagen) + 1% BME in 96 well plates (24-96 per sample).

#### Gating strategy

FSC-A and SSC-A were gated to liberally include all lymphocytes and myeloid cells. Subsequent gating on intact cells was performed on FSC-W, and on singlets by SSC-W.  
 Live, Immune cells: Calcein+CD45+  
 CD4+ T cells: Calcein+CD45+CD14–CD3+CD4+.  
 CD8+ T cells: Calcein+CD45+CD14–CD3+CD8+.  
 pDCs: Calcein+CD45+CD14–CD3–CD11c–HLA-DR+CD123++.

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.