ARTICLES
https://doi.org/10.1038/s41590-020-00836-7

nature immunology

Check for updates

Gut CD4$^+$ T cell phenotypes are a continuum molded by microbes, not by T_H archetypes

Evgeny Kiner[1,2,26], Elijah Willie[3], Brinda Vijaykumar[1,2], Kaitavjeet Chowdhary[1,2], Hugo Schmutz[1,2], Jodie Chandler[4], Alexandra Schnell[2], Pratiksha I. Thakore[5], Graham LeGros[4], Sara Mostafavi[6,7,8], Diane Mathis[1,2] ✉, Christophe Benoist[1,2] ✉ and The Immunological Genome Project Consortium*

CD4$^+$ effector lymphocytes (T_{eff}) are traditionally classified by the cytokines they produce. To determine the states that T_{eff} cells actually adopt in frontline tissues in vivo, we applied single-cell transcriptome and chromatin analyses to colonic T_{eff} cells in germ-free or conventional mice or in mice after challenge with a range of phenotypically biasing microbes. Unexpected subsets were marked by the expression of the interferon (IFN) signature or myeloid-specific transcripts, but transcriptome or chromatin structure could not resolve discrete clusters fitting classic helper T cell (T_H) subsets. At baseline or at different times of infection, transcripts encoding cytokines or proteins commonly used as T_H markers were distributed in a polarized continuum, which was functionally validated. Clones derived from single progenitors gave rise to both IFN-γ- and interleukin (IL)-17-producing cells. Most of the transcriptional variance was tied to the infecting agent, independent of the cytokines produced, and chromatin variance primarily reflected activities of activator protein (AP)-1 and IFN-regulatory factor (IRF) transcription factor (TF) families, not the canonical subset master regulators T-bet, GATA3 or RORγ.

T_{eff} cells are key drivers of both humoral and cellular immune responses, orchestrating adaptive (antibodies, cytotoxic cells) and innate (macrophages, granulocytes) immune responses. This range of abilities has long raised the issue of functional diversity, which was documented by functional assays even before the molecular identification of major histocompatibility complex (MHC) and T cell antigen receptor (TCR) molecules, the central axis of T cell activation and differentiation[1,2]. A key advance was the demonstration that functional phenotypes of different T cell clones were keyed to the cytokines they produced[3,4], coining the T_H1/T_H2 nomenclature of T_H subsets. T_H1 cells secrete IFN-γ and mainly support inflammatory and cytotoxic responses; T_H2 cells produce IL-4, IL-5 or IL-13 and principally help B cells produce antibodies. This division has since been revised several times to add more subsets (IL-17-secreting T_H17 cells, IL-9-secreting T_H9 cells, follicular helpers (T_{FH})[5–7]), but the core notion that T_{eff} cells belong to discrete and largely stable states defined by the cytokines they produce has endured[8,9]. Different types of infectious or allergic challenges elicit different T_{eff} 'flavors' (T_H1 cells are generally associated with intracellular pathogens, T_H2 cells with helminth parasites, T_H17 cells with bacterial and fungal infections), and these T_H distinctions also have implications for immune-mediated diseases[10]. Indeed, the T_H paradigm has elicited parallel cosmologies in macrophages, γδ T cells or innate lymphoid cells (ILCs)[11].

However, this model was questioned almost since its inception[12,13]. First, because its attractive simplicity could lead to shoehorning of immune functions (for example, publications in the 1990s erroneously tagged immune diseases as either T_H1 or T_H2). Second, many reports documented that the secretion of IFN-γ, IL-4 or IL-17 is not always mutually exclusive[14]. Plasticity between T_H subtypes was demonstrated, suggesting that these cell states are not as stable and terminally differentiated as originally inferred from T_H lines grown in supraphysiological cytokine concentrations[9,13]. Further, while some cell surface markers were proposed as indicators of differentiated T_H types, they often proved non-exclusive. Thus, while T_H subsets were most precisely defined in vitro, their in vivo counterparts remained elusive.

Here, we aimed to assess the spectrum of phenotypic states that T_{eff} cells can adopt in vivo, leveraging the unbiased potential of single-cell genomics[15]. In essence, we returned to the clonal analysis that founded the T_H paradigm[3,4], but now with the ability to evaluate the entirety of a cell's transcriptome and chromatin structure, rather than only a few preselected cytokines or markers. We analyzed T cells in the colonic lamina propria (LP), a frontline tissue under continuous and diverse challenge, by comparing CD4$^+$ T cells in mice under germ-free conditions, carrying normal commensal microbiota or infected with agents that elicit diversely biased T_{eff} responses. The results indicate that T_{eff} cells form a continuum in transcriptional space, but highlight some novel phenotypes. The production of key cytokines did show skewed distributions, but these did not identify the discrete cell clusters that might have been expected from the T_H paradigm.

Results

A continuum of effector phenotypes in colonic CD4$^+$ T_{eff} cells. To probe the transcriptional landscape of CD4$^+$ T_{eff} cells in an unbiased manner, we performed single-cell RNA sequencing (scRNA-seq) on total CD4$^+$ T cells from the colonic LP, starting with conventionally

[1]Department of Immunology, Harvard Medical School, Boston, MA, USA. [2]Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA. [3]Bioinformatics Program, University of British Columbia, Vancouver, British Columbia, Canada. [4]Malaghan Institute of Medical Research, Wellington, New Zealand. [5]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Departments of Statistics and Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [7]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [8]Vector Institute, Toronto, Ontario, Canada. [26]Present address: Immunai, New York, NY, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: cbdm@hms.harvard.edu; cbdm@hms.harvard.edu

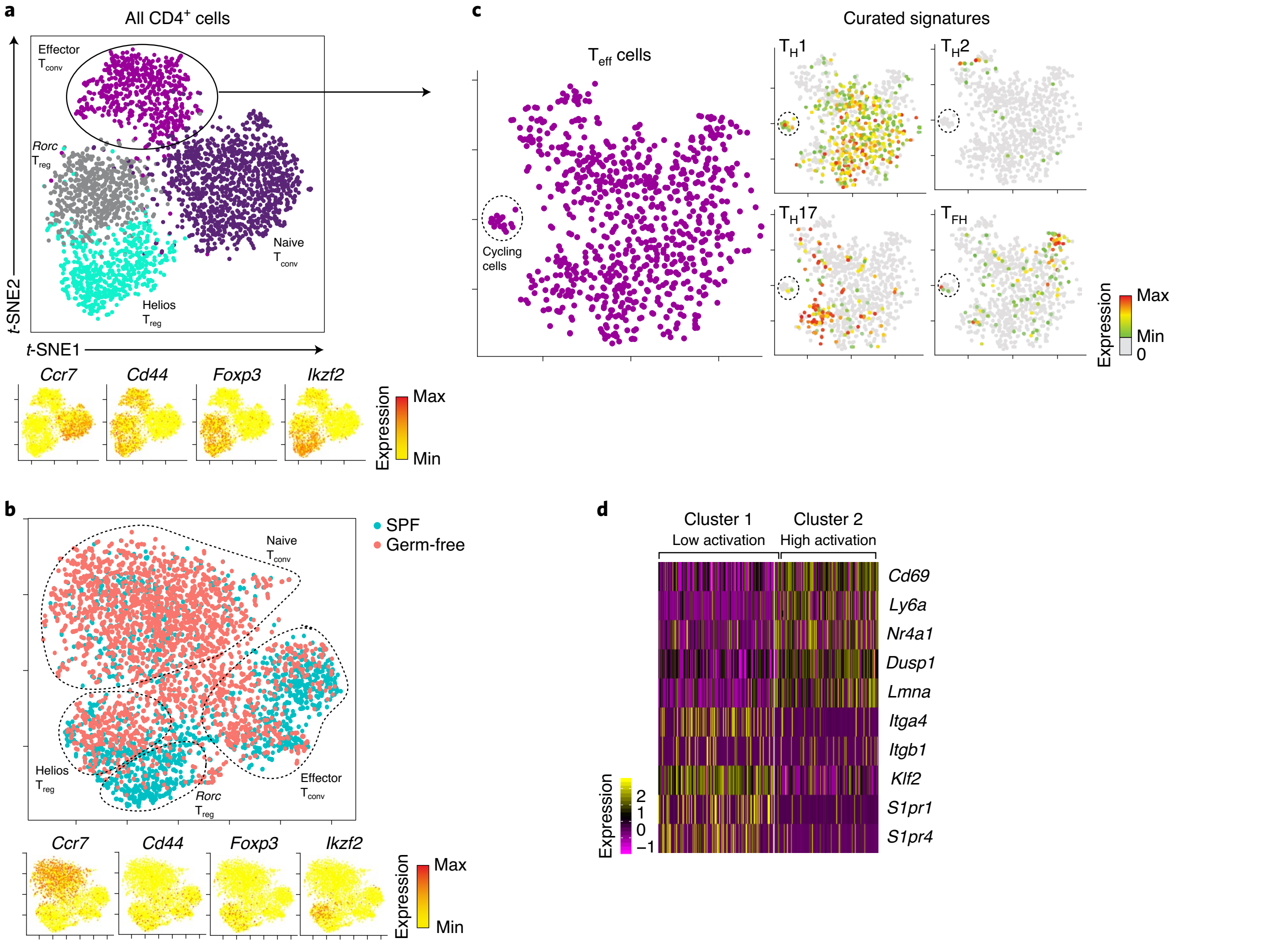


Fig. 1 | The transcriptional landscape of CD4$^+$ T cells in the colon. a, scRNA-seq analysis of total colonic LP CD4$^+$ T cells from SPF mice (computed from the 658 most variable genes). *t*-SNE representation, color coded by *k*-nearest neighbor (KNN) cell clusters (top), identified based on the expression of prototypic transcripts (bottom). **b**, scRNA-seq analysis of total colonic LP CD4$^+$ T cells from germ-free and SPF mice. Top, *t*-SNE representation, color coded by cell of origin. Bottom, marked clusters are identified based on the expression of prototypic transcripts. **c**, *t*-SNE representation, restricted to the CD4$^+$ T$_{\text{eff}}$ cells selected in **a** (*t*-SNE computed from the 584 most variable genes). Right, overlay of combined expression of prototypic T$_{\text{H}}$ gene sets (Supplementary Table 2). **d**, Heatmap of T$_{\text{eff}}$ cells divided into two clusters by KNN clustering. Representative genes overexpressed in each cluster are shown.

housed (SPF) C57BL/6 mice. Two experiments were performed with droplet-based scRNA-seq (Extended Data Fig. 1a,b and Supplementary Table 1; datasets were analyzed individually; replicates served to confirm conclusions). It was straightforward to parse, with standard clustering, CD4$^+$ T cells into the four main groups expected from flow cytometry (Fig. 1a): regulatory T cells (T$_{\text{reg}}$; *Foxp3*$^+$) and their *Rorc*$^+$ and *Ikzf2*$^+$ (Helios) subsets, naive T conventional cells (naive T$_{\text{conv}}$; *Cd44*$^-$*Ccr7*$^+$) and T$_{\text{eff}}$ cells (*Cd44*$^+$*Ccr7*$^-$). To assess the influence of the commensal microbiota on this distribution, we generated scRNA-seq datasets of colonic CD4$^+$ T cells from SPF and germ-free mice (Fig. 1b), revealing similar clusters, with fewer RORγ$^+$ T$_{\text{reg}}$ and T$_{\text{eff}}$ cells in germ-free mice, as expected.

To assess which phenotypic states gut T$_{\text{eff}}$ cells can adopt, we reclustered the T$_{\text{eff}}$ population from SPF mice. Here, with the exception of cycling cells, we could not observe any clear partitioning of cells, but rather a quasi-continuous cloud (Fig. 1c, left). To search for distinctions corresponding to the major recognized T$_{\text{eff}}$ types, we manually curated from published signatures short but robust and highly specific gene sets, which included the defining cytokines, driving TFs and a few correlated transcripts but left out generic activation-associated transcripts or transcripts with poor specificity (Supplementary Table 2). The T$_{\text{H}}$2 signature showed polarized expression, while cells expressing the T$_{\text{H}}$17- and especially T$_{\text{H}}$1-associated signatures were dispersed more widely across the continuum (Fig. 1c, right). To ensure that this continuum was not due to the high dropout rate of scRNA-seq, we reanalyzed a published dataset from colonic T$_{\text{eff}}$ cells that included fewer cells but was sequenced to greater depths[16]. These data also showed a continuous distribution and dispersion of the T$_{\text{H}}$ signatures (Extended Data Fig. 1c). If cytokines do not represent the main axes of variance in colonic T cells, what does? To this end, we used a simple clustering strategy, which showed that the driving variance lay in the degree of activation of T$_{\text{eff}}$ cells, represented by typical activation transcripts such as *Cd69* or *Nr4a1* (Fig. 1d). T$_{\text{eff}}$ cells with a lower degree of activation overexpressed *Klf2* and *S1pr1*, a combination shown to restrain CD4$^+$ T cell differentiation[17]. Thus, the main heterogeneity of T$_{\text{eff}}$ cells in the colonic LP corresponds to a gradient of activation in response to commensal microbiota but not predominantly to a commitment to produce one cytokine or the other.

Different intestinal infections elicit divergent T$_{\text{eff}}$ phenotypes. It thus seemed difficult to identify discrete T$_{\text{H}}$1 or T$_{\text{H}}$17 cell populations in normal mice. We hypothesized that under baseline conditions, T$_{\text{eff}}$ cells were only partially polarized because they were incompletely activated by commensals, with only 'stubs' of more differentiated states that the cells could potentially reach. We thus further polarized the T cell pools by infecting mice with pathogens known to elicit biased immune responses: (1) Δ*aroA* *Salmonella enterica* (serovar Typhimurium), a non-invasive mutant that elicits IFN-γ-dominated responses, (2) *Citrobacter rodentium*, a strong inducer of IL-17 and (3) *Heligmosomoides polygyrus* and *Nippostrongylus brasiliensis*, two helminths that provoke prototypic type 2 responses (Fig. 2a). Infection times (11–13 d) allowed responses to develop and achieve full bias. Flow cytometry confirmed the production of the expected cytokines, including some IFN-γ and IL-17A double-producer cells, as expected (Fig. 2b). In a first experiment, T_{eff} cells from control or infected mice were tagged with DNA-coded antibodies ('hashtagged'; ref. [18]) and comingled for sorting, microfluidic bead capture and library construction, making for a robust intra-batch comparison (Fig. 2a). As in uninfected mice, $CD4^+$ T cells clustered into T_{reg} cells, naive T cells and T_{eff} cells (Extended Data Fig. 2a).

T_{eff} cells were then considered on their own, with dimensionality reduction on *t*-distributed stochastic neighbor embedding (*t*-SNE) (Fig. 2c) or uniform manifold approximation and projection (UMAP) (Extended Data Fig. 2b) plots, which revealed a dominant partitioning according to the infectious agent used. Outside the main 'blob', some T_{eff} cells did break out into discrete populations, but we could not detect well-demarcated cell clusters that expressed characteristic T_H gene sets. These mapped to skewed but broad swaths of cells (cells with high levels of the T_H2 gene set were best demarcated, those with the T_H17 gene set were biased but dispersed, and cells with high levels of the T_H1 gene set were found almost throughout; Fig. 2c and Extended Data Fig. 2b). This lack of segregation was robust across gene sets (if anything, it was more diffuse using another curated signature set based on ref. [19]) (Extended Data Fig. 2c). The expression of *Ifng* and *Il17a* transcripts also overlapped, consistent with the double-producer cells detected by flow cytometry (Extended Data Fig. 2d). These conclusions were also true for a replicate set of colonic $CD4^+$ T_{eff} cells from mice infected with the same pathogens (Extended Data Fig. 2e). The dominant influence of the infectious microbe over the T_H phenotype marked by cytokine production was objectivized by comparing the overall Euclidean distance between all cells expressing *Il17a* and *Ifng* transcripts from the different conditions; T_{eff} cells expressing *Ifng* or *Il17a* transcripts from each infection type were much closer than their cytokine-sharing counterparts in mice with other infections (Fig. 2d).

We applied a panel of clustering and biclustering algorithms in an attempt to break the cell cloud into clusters that coincided with the expression of T_H signature sets, but none of the clusters thus generated were uniquely enriched for any one T_H signature or cytokine (Extended Data Fig. 3a–c). To objectively verify the continuity in the distribution of transcriptomes of T_{eff} cells, we used Hartigan's dip test of multimodality[20] after applying a projection defined by the minimum separation hyperplane[21] to the expression of the most variable genes. The results showed that T_{eff}, T_{reg} and naive T_{conv} cells significantly segregated by Hartigan's test (Fig. 2e, top), while there was no significant break in the distances within T_{eff} pools (Fig. 2e, middle and bottom). These results confirmed that T_{eff} cells occupy a continuum point cloud that is not easily separable into distinct clusters.

One explanation for this continuous T_{eff} distribution is that they included different subsets of the canonical T_H1, T_H2 and T_H17 archetypes. However, projection of differentiating genes reported for the 'pathogenesis subsets' within T_H17 cells[22,23] did not demarcate distinct subsets of IL-17-producing cells, although it showed a skewed distribution more generally (Extended Data Fig. 3d). Similarly, a reported distinction between 'homeostatic' and 'inflammatory' T_H17 cells[24], the latter being elicited by *C. rodentium* infection, may have mostly resulted from infection rather than from distinct T_H17 subsets, as the corresponding signature did not specifically demarcate IL-17-producing cells (Extended Data Fig. 3e).

Deep machine learning tools can efficiently discover combinatorial and non-linear patterns that are difficult to discern conventionally. In another attempt to identify patterns that would uniquely identify IL-17- or IFN-γ-producing cells, we optimized and trained a deep neural network (DNN) to classify cells into IL-17- and IFN-γ-producing groups based on their single-cell transcriptomes. As a positive control, this architecture could be trained to recognize T_{eff} and T_{reg} cells from the held-back test set (Methods). The DNN did partially identify *Ifng*- and *Il17a*-positive cells in the test set (Extended Data Fig. 4a,b; 90.2% and 60.7% accuracy for *Ifng*- and *Il17a*-positive cells, respectively). However, using the integrated gradients method to measure the importance of the transcripts used by the model to support this identification showed little reproducibility in independent training runs (Extended Data Fig. 4c). Beyond a few transcripts known to correlate with *Il17a* (*Tmem176a*, *Capg*), only *Il22* had a strong and reproducible influence, which is an internal control given its known coregulation with *Il17*. Indeed, when *Il22* was left out, prediction efficacy dropped to 28.7%. Hence, even with a pliable artificial intelligence tool, it seemed difficult to identify robust T_H1 or T_H17 transcriptome patterns.

Finally, we assessed the distribution of surface markers that are associated with T_H subsets and are commonly used for cell sorting (*Ccr5* and *Cxcr3* for T_H1; *Ccr6* and *Il1r2* for T_H17). *Ccr6* and *Il1r2* proved to be mutually exclusive, with only partial overlap with cells transcribing *Il17a* (Fig. 2f). *Ccr5* and *Cxcr3* transcripts were widely distributed across the cloud and only partially overlapped with the T_H1 signature. Flow cytometric analysis of LP cells after *Salmonella* infection confirmed these results (Fig. 2f). Thus, not only were classic T_H subsets not clearly identifiable in the transcriptional data, but the flow cytometry markers used to identify them had limited congruence in this context.

T_{eff} phenotypes are distinguished by infecting agents, not by T_H type. Colonic T_{eff} cells clustered according to the type of infection, rather than by the cytokine they expressed. Accordingly, analysis of variable transcripts present in *Il17a*- and *Ifng*-expressing cells revealed divergent patterns, with blocks of coexpressed transcripts that largely aligned with the infection (Fig. 3a). To validate this result and exclude technical pitfalls of scRNA-seq, we used an *Il17a*-GFP reporter mouse line and performed population RNA-seq on colonic GFP-positive cells at baseline or after infection with *Salmonella* or *Citrobacter* (Fig. 3b). Echoing the single-cell data, principal component analysis (PCA) showed that IL-17A$^+$ cells from each condition clustered separately from each other (Fig. 3c). The direct comparison of IL-17A$^+$ cells from *Salmonella*- or *Citrobacter*-infected mice yielded 277 differential transcripts (at fold change (FC) > 2, false discovery rate < 0.05; Fig. 3d). Among this set, transcripts with differential representation in the single-cell data showed similar biases. Thus, the majority of changes imparted by infection were unrelated to *Il17* or *Ifng* expression or membership in a T_H class.

The primary determinants of T_{eff} variability. Their expression patterns within the projection plots of Figs. 1c and 2c indicated that prototypic T_H1 or T_H17 signature sets did not mark discrete sets of cells. To turn the question to a gene-centric perspective, we asked which coregulated modules of transcripts existed among these $CD4^+$ T_{eff} cells, and whether these might track with cytokine production. First, a PCA showed that the gene sets in the principal components (PCs) with the most variance contained few T_H-associated signature genes (Extended Data Fig. 5a). Next we analyzed gene–gene correlation, leveraging coexpression across thousands of individual cells[25]. The transcripts for some cytokines did show significant positive coexpression (*Il4* or *Il5* and *Il13*; *Il17a* and *Il17f*; Extended Data Fig. 5b). We separated coregulated gene modules (affinity propagation) that

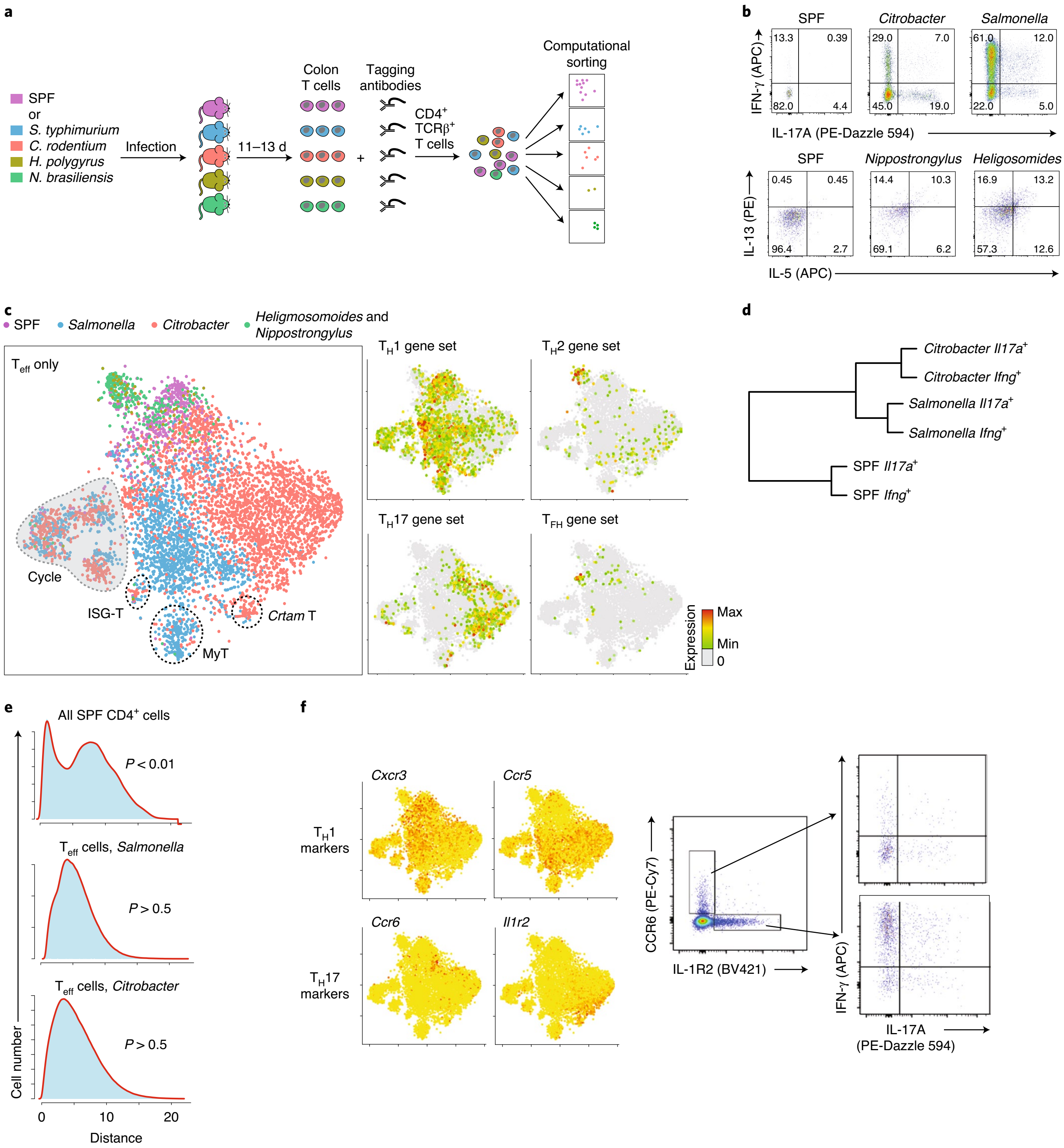


Fig. 2 | Variation in T_{eff} transcriptomes shows continuous distribution that is not dictated by 'T_H subsets'. a, Schematic of the hashtagging experiment. Mice were infected with different pathogens, and their colonic LP cells were extracted, labeled with hashtagging antibodies, sorted as $CD4^+$ T cells and processed as a single batch on the 10x Chromium Controller. Sample demultiplexing was performed computationally. **b**, Flow cytometric confirmation of intestinal infections after intracellular staining for the cytokines shown (gated on $CD4^+TCR\beta^+FOXP3^-CD44^{hi}$ cells). **c**, *t*-SNE representation of T_{eff} scRNA-seq data from mice under different infection conditions (computed from 930 variable genes). Left, data are color-coded by condition or infection. Right, overlay of combined expression of prototypic T_H gene sets. **d**, Dendrogram of Euclidean distances between cells in the scRNA-seq dataset in **c** that splits cells that express *Ifng* or *Il17a* in each of the infection conditions. **e**, Hartigan's dip test was applied to whole colonic $CD4^+$ T cells from SPF mice (top) or only to T_{eff} cells from *Salmonella*-infected (middle) or *Citrobacter*-infected mice (bottom). MyT and cycling cells were not included in this analysis. **f**, Expression of commonly used markers of T_H subsets. Left, RNA expression in the scRNA-seq data (overlaid on the *t*-SNE plot from **c**). Right, protein expression by flow cytometry in $CD4^+$ T_{eff} cells (gated on $CD4^+TCR\beta^+FOXP3^-CD44^{hi}$ cells) from *Salmonella*-infected mice.

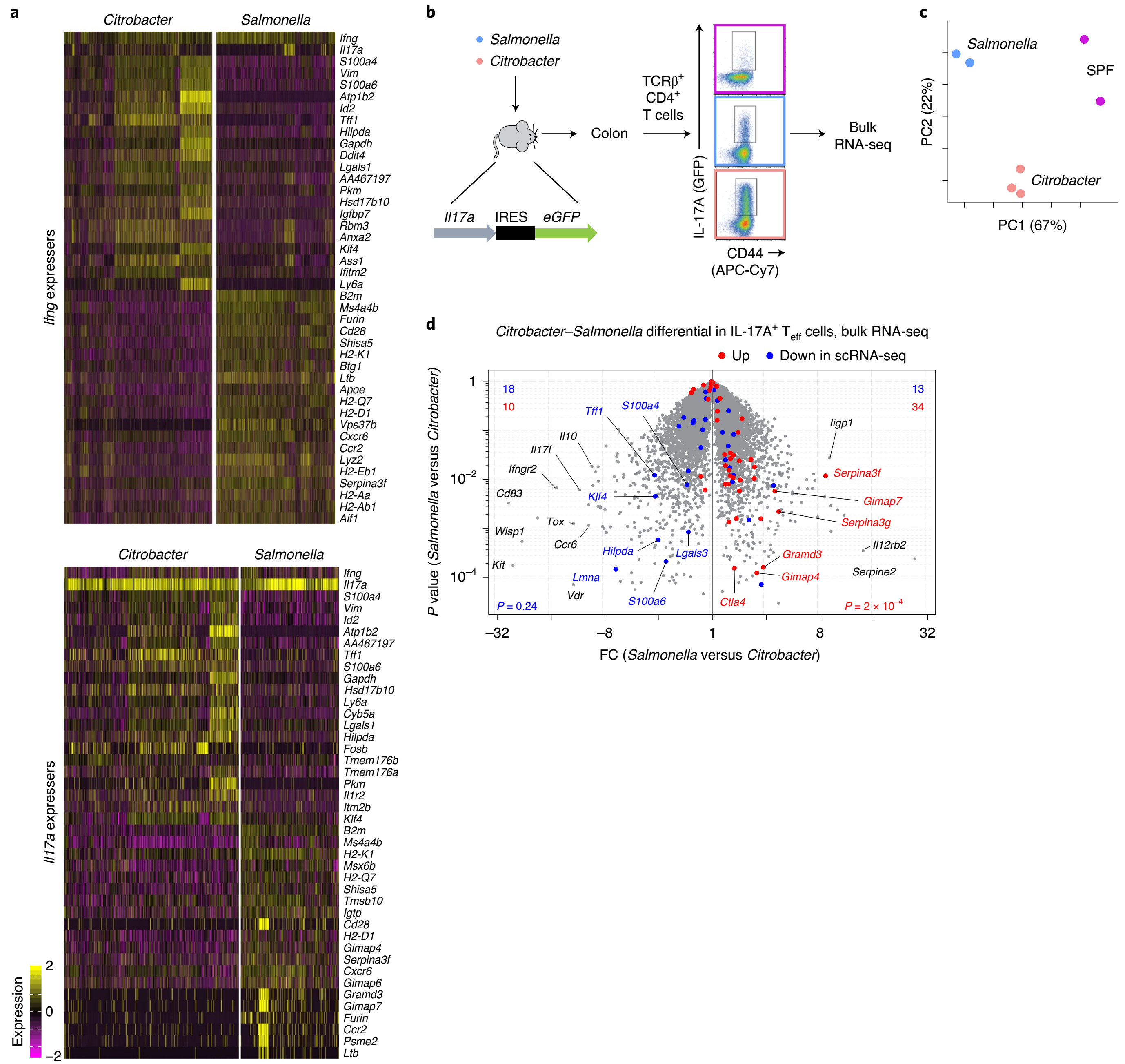


Fig. 3 | T_{eff} phenotypes are distinguishable by infection rather than by T_H type. a, Heatmap of most differentially expressed transcripts in *Ifng*$^+$ (top) and *Il17a*$^+$ (bottom) T_{eff} cells from *Citrobacter*- or *Salmonella*-infected mice. **b**, Sorting of IL-17a-expressing CD4$^+$CD44$^+$ T_{eff} cells from *Il17a*-IRES-GFP reporter mice (infected with *Salmonella* or *Citrobacter* or uninfected) for expression profiling by ultra-low-input RNA-seq. **c**, PCA analysis of datasets from **b**. **d**, Volcano plot of bulk RNA-seq data from **c**, comparing IL-17A$^+$ T_{eff} cells from *Salmonella*- or *Citrobacter*-infected mice. Red and blue highlights, transcripts that were differentially expressed by scRNA-seq.

defined independent transcriptional programs (Extended Data Fig. 5c). Gene ontology analysis showed that while most modules were related to generic functions (Supplementary Table 3), a few small modules (M7, M11, M13) included some elements of prototypical T_H signatures, for example, cytokines and TFs (*Ifng*, *Il13*, *Tbx21*, *Gata3*). But when projected across the cell space, most modules showed broadly differential representation as a gradient across all cells, cutting across cells expressing *Ifng* or *Il17a* transcripts (Extended Data Fig. 5d, with the exception of cell cycle genes in M1 and M2, the MHC-II module in M9 and the T_H2-like cluster in M7). Thus, the major components of variability among T_{eff} cells highlighted a continuous cloud of phenotypic variance, rather than discrete cell sets.

T_{eff} phenotypes over time. A possible explanation for the lack of discrete T_H1 and T_H17 identities was that the 13-d time point chosen for analysis might be not be ideal and that at 13 d, polarized cells might have faded or have yet to appear. To test this possibility, we analyzed LP CD4$^+$ cells at different times after *Salmonella* infection, again hashtagged in a single batch. The CD4 response, denoted by total CD4$^+$ T proportions and the effector:naive cell ratio, was highest in the day 10–17 window (Extended Data Fig. 6a,b).

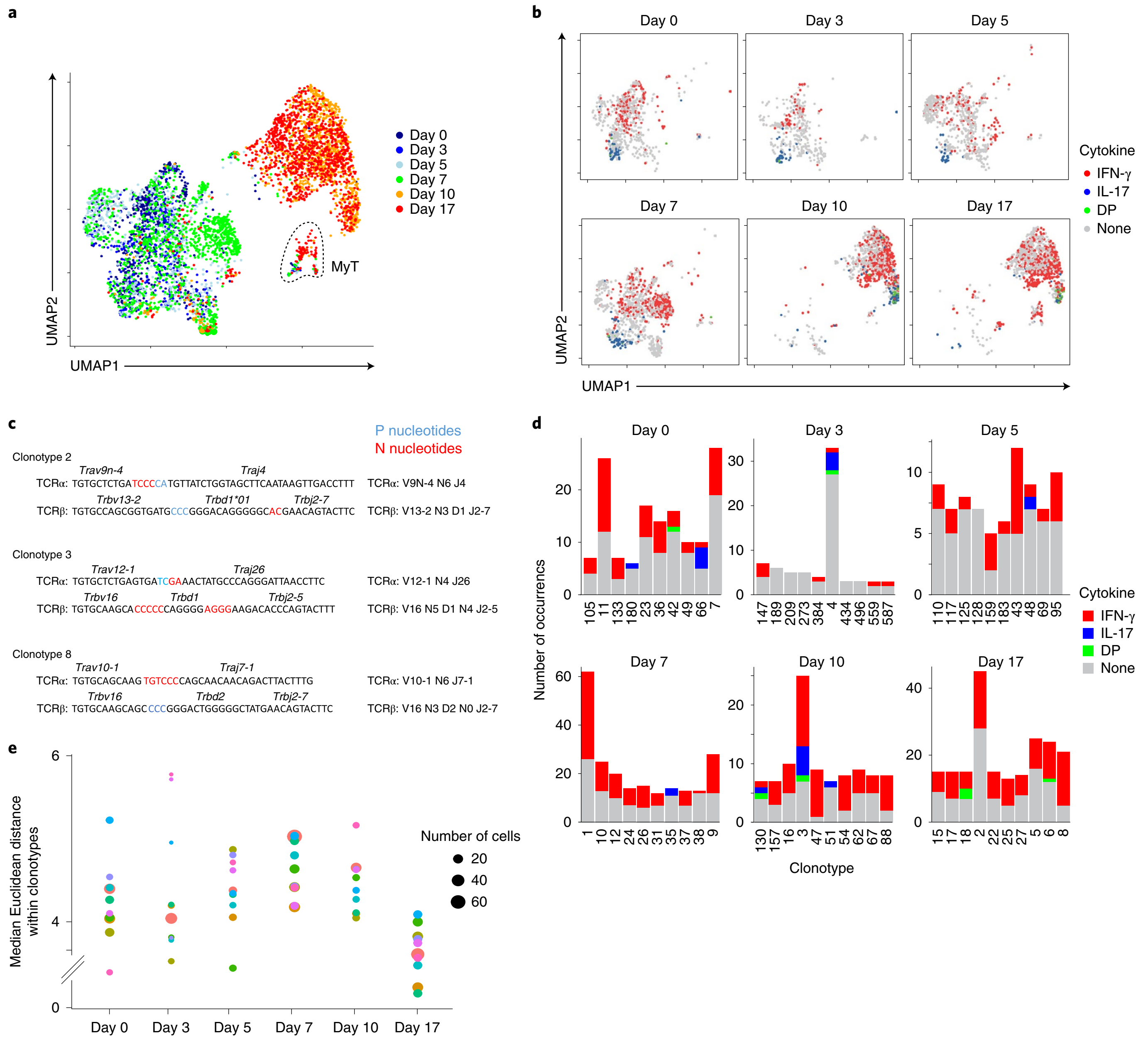


Fig. 4 | Repeated clonotypes can adopt different phenotypes and do not diverge over time. a, UMAP representation of T_{eff} cells from murine LP at different time points post-infection with *Salmonella*. MyT cells are circled. **b**, Cells from different time points. Cytokine-producing cells are highlighted as shown. **c**, Representative examples of clonotypes with unique complementarity-determining region (CDR)3 identified by scTCR sequencing (non-germline non-templated (N) and palindromic (P) nucleotides shown). **d**, Numbers of *Il17a*-, *Ifng*- or *Il17a*- and *Ifng*- (double-positive, DP) expressing cells in the ten most frequent clonotypes identified in each individual time point. **e**, Median Euclidean distances between cells within the same clonotype across the top ten clonotypes for each time point. Euclidean distance was calculated based on the T_H genes from Supplementary Table 2. Clonotypes are color-coded, and the size denotes the number of cells that express each clonotype.

A marked shift in the overall T_{eff} transcriptomes occurred from day 10 onwards (Fig. 4a). Transcripts that distinguished these two superclusters included many of the *Salmonella*-specific transcripts identified above but no prototypical T_H signature transcripts (Extended Data Fig. 6c). In these samples, IL-17^+ cells were better demarcated than in earlier experiments, and IFN-γ^+ cells were again broadly spread out, with no indication of a time-dependent convergence (Fig. 4b). Both types of cytokine-producing cells were shifted during the 'day 10 transition', again implicating the infectious agent as the dominant driver of T_{eff} phenotypes at the height of infection.

Next, we asked whether one could identify distinct lineages of *Il17*- and *Ifng*-expressing cells within $CD4^+$ T_{eff} cells at different infection times, using the sequences of rearranged *Tcra* and *Tcra* genes to lineage-trace cells originating from the same progenitor. A total of 579 repeated clonotypes were observed (defined by shared nucleotide sequences for both chains and P or N nucleotide addition that ensured true clonal amplification; examples in Fig. 4c). These repeated clonotypes expanded with time in T_{eff} cells but not in naive T_{conv} or in T_{reg} cells, consistent with infection-driven expansion (Extended Data Fig. 6d). Importantly, expanded T_{eff} clones were not restricted to the expression of one cytokine; most *Il17*-expressing cells within a clonotype had cousins that expressed *Ifng* or both *Ifng* and *Il17* (Fig. 4d). That expanded clonotypes did not appear committed to produce a single cytokine could be explained by parallel differentiation of the initial precursor. However, the median Euclidean distance between members of a clonotype did not increase with time, if anything, it contracted beyond day 10, whether computed from the T_H signature gene sets (Fig. 4e) or the most variable genes (Extended Data Fig. 6e), indicating that cells were not diversifying. Thus, this lineage tracing revealed no parallel tracks of differentiation for *Ifng* and *Il17* expression; the *Salmonella*-driven dominance of IFN-γ production extended across all amplified clonotypes.

T_{eff} phenotypes at the chromatin level. Accessibility of enhancer elements in chromatin is a more proximal readout of a cell's differentiated state than mRNA levels, which are affected by post-transcriptional events. To explore the relationship between *Il17*- and *Ifng*-expressing cells at the chromatin level, we performed single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq)[26] on 4,671 LP CD4$^+$ T cells from colons of *Salmonella*-infected mice. As with scRNA-seq, three distinct clusters could be distinguished by clustering and identification based on the accessibility of typical indicator genes (T_{reg}, naive T_{conv} cells and a cloud of T_{eff} cells; Fig. 5a). We leveraged a framework of pan-immune open chromatin regions (OCRs) and charts of those most likely to be associated with activity of a given gene[27], extracting OCRs that best predicted the expression of *Tbx21* and *Rorc* and averaging their accessibility to calculate *Tbx21* and *Rorc* 'chromatin scores' per cell. We validated these scores by showing that they clearly distinguished ATAC-seq profiles from in vitro-derived T_H1 and T_H17 cells (Fig. 5b, top). However, when projected on the UMAP plot of ex vivo T_{eff} cells, the *Tbx21* and *Rorc* scores were broadly distributed, with diffuse local maxima, but no cell cluster displayed either exclusively (Fig. 5b, bottom). We also examined chromatin profiles across the *Rorc* and *Tbx21* loci themselves, by collapsing the reads from cells selected as having high or low signals at *Rorc*- or *Tbx21*-controlling OCRs and asking whether one would be anticorrelated with the other. Clearly, chromatin openness at one locus was independent of the state at the other locus (Fig. 5c). Thus, chromatin opening at master regulator loci did not split identifiable T_H1 and T_H17 subsets.

As an alternative to analyzing the *Rorc* and *Tbx21* loci, we computationally mapped the differential activity of OCRs enriched in DNA motifs recognized by these TFs relative to background OCRs[28]. T-bet and GATA3 motif scores were broadly distributed (Fig. 5d), with a more concentrated over-representation of RORγ motif scores (acknowledging the caveat that these motifs may be recognized by the related TFs EOMES and RORα, respectively).

If RORγ and T-bet are not the main discriminators of chromatin accessibility of T_{eff} cells, then what is? We broadened the analysis to all TF motifs in the JASPAR database, ranking them by their overall variability (null distribution from randomized data; Fig. 5e). This ranking was dominated by motifs for several factors, foremost those for the AP-1 (FOS, JUN, etc.) and the IRF (IRF4, IRF2, IRF9) families or for other factors related to T cell activation (BACH2), while the T-box and nuclear receptor families (T-bet, EOMES and RORγ, RORα) figured less prominently. Correspondingly, scores for *Fos* and *Irf4* motifs segregated most distinctly (Fig. 5f). Thus, in line with mRNA data, which showed that generic activation was the main driver of T_{eff} diversity, activation drivers (AP-1, IRF4, BACH2) seemed to have a more important contribution in parsing T_{eff} cells than classic master regulators.

A functional continuum of CD4$^+$ T_{eff} cells. A continuum in which different functions are distributed along poles and gradients is more challenging to address experimentally than demarcated groups of cells. To validate the notion of a continuum of T_{eff} phenotypic states, we followed a strategy similar to one described recently[29–31] in which cell sorting was not steered to well-defined cell populations but performed by integrating information in a multidimensional marker space (Fig. 6a). We first identified transcripts in the scRNA-seq data that showed different gradients of expression through the T_{eff} continuum and encoded cell surface molecules detectable by flow cytometry (*Klrg1*, *Cxcr6*, *Icos*, *Cd69*, *Ly6a* (encodes SCA-1); Fig. 6b). Colon LP cells were resolved by flow cytometry with antibodies against these markers, combining results in a multiparameter *t*-SNE projection (Fig. 6c). In this proteomic space, no specific clusters of cells were identified by any one marker (perhaps with the exception of the receptor KLRG1); all were distributed as quantitative gradients as for the mRNA data. We then empirically determined gates to pilot a cell sorter to purify cells belonging to specific areas of the cell cloud (Fig. 6d), yielding three distinct cell populations. Such cells were sorted from colon LP of *Salmonella*-infected mice for phenotypic and functional testing.

Conventional RNA-seq on these sorted populations showed a differential transcript representation, with enrichments that corresponded well to signatures predicted from the scRNA-seq data (Fig. 6e). Differentially expressed genes included, in population B, transcripts associated with a more resting state (*Ccr7*, *Sell* and *Tcf7*), while IL-17-associated transcripts (*Rorc*, *Il23r*, *Il17re*) were over-represented in population A. For a test of function, we stimulated these sorted cells and measured cytokine secretion by ELISA (Fig. 6f). Distinctive patterns were observed, although, as expected, no single pool was associated with the exclusive secretion of any one cytokine. Populations A and C secreted significantly more IFN-γ than did population B, whereas population A secreted more lL-17A and IL-22. But both populations encompassed all potentialities, only in quantitatively different amounts, confirming that the scRNA-seq data captured true continuous T_{eff} heterogeneity.

New T_{eff} populations. As presented above, the T_{eff} pool in SPF or infected mice included, beyond the main 'cloud', a few well-distinguished populations (Fig. 2c).

1. A small T_{eff} population (ISG-T) was peculiar because it expressed high levels of IFN-induced signature transcripts (ISGs) (Fig. 7a, left) and was over-represented after infection with *Salmonella* or *Citrobacter*. Comparison with profiles induced in T cells by type 1 or type 2 IFNs indicated that ISG-T cells likely respond to type 1 IFNs (Fig. 7a, right). Their existence suggested either a small subset uniquely responsive to IFN or normal T_{eff} cells that happened to reside in a small anatomical compartment where IFN was particularly abundant. Similar subsets have been described in CD4$^+$ T cells from house dust mite-infected lungs and kidneys from patients with lupus nephritis[32,33].

2. Another population expressed high levels of the surface markers *Cd160*, *Crtam* and *Lag3*, the neural gene *Nrgn* and several chemokines (Fig. 7b). We sorted this CD4$^+$CRTAM$^+$ population for bulk RNA-seq, confirming the particular signature (Fig. 7c). Pathway analysis showed enrichment of signal transducer and activator of transcription (STAT)3, prolactin and neuregulin signaling pathways, hinting at a possible origin.

3. The most intriguing population was myeloid-like T (MyT) cells, which unexpectedly showed many myeloid cell transcripts, such as *Apoe*, *Lyz2* or *C1qa*, and several transcripts for MHC-II (Fig. 7d). This expression of myeloid transcripts was not wholesale; only a fraction of genes with strong T versus myeloid differential expression was represented in MyT cells (Fig. 7e), several of which corresponded to innate antimicrobial receptors or defense mechanisms (*Lyz2*, *C1q*, *Cfp*, *Tyrobp*). Correspondingly, a small MHC-II$^+$ subset of TCRβ$^+$CD4$^+$CD44^{hi} T_{eff} cells was detected by cytometry (Extended Data Fig. 7a), for which the RNA-seq transcriptome confirmed the single-cell data (Extended Data Fig. 7b). We applied 'CITE-seq' for protein detection with DNA-barcoded antibodies[34], revealing a good correspondence between mRNA and surface

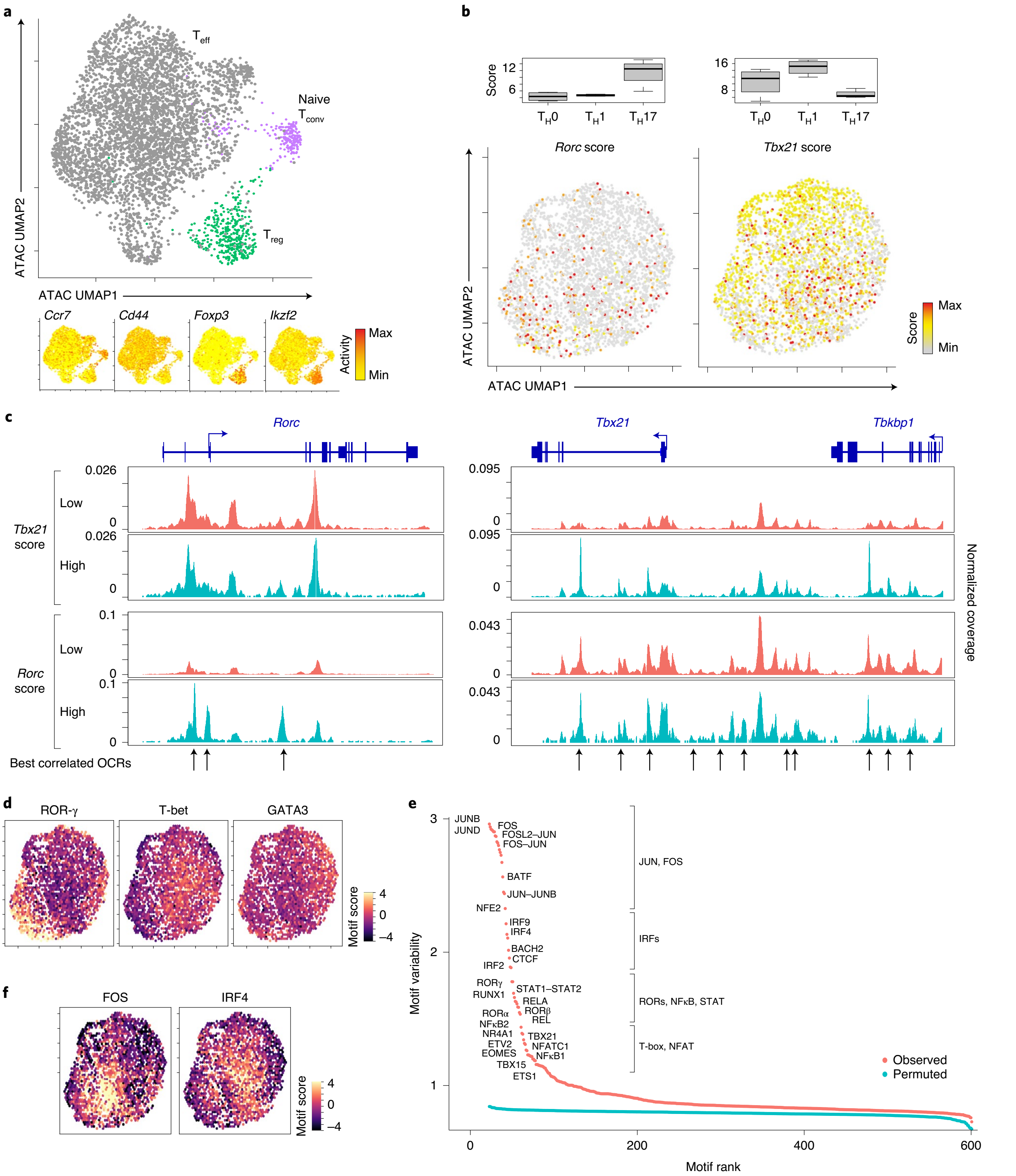


Fig. 5 | The chromatin states of T_{eff} cells are found on a continuum. a, scATAC-seq of total LP CD4+ T cells from *Salmonella*-infected mice. UMAP representation with T_{reg} cells and naive T_{conv} cells identified (top) based on gene activity at prototypic loci (bottom). **b**, Cell chromatin scores for *Rorc* and *Tbx21* loci, computed from the accessibility of expression-correlated OCRs. Top, scores in T_H0, T_H1 and T_H17 cells differentiated in vitro. $n = 4$ biological replicates for each condition. Center, median; box limits, first and third percentiles; whiskers, 1.5× interquartile range. Bottom, scores for each cell in the scATAC-seq data from **a** (T_{eff} only) in a UMAP plot. **c**, Aggregated coverage maps around *Rorc* and *Tbx21* loci in T_{eff} cells, split based according to their chromatin score at each locus (shown at left); arrows, locations of the best expression-correlated OCRs used to compute the scores. **d**, Over-representation in each T_{eff} cell (data from **a**) of TF-binding motifs in accessible chromatin (TF motif deviation scores that were bias-corrected by chromVAR[28]) for classic master regulators (UMAP framework from **b**). **e**, Combined variability across the T_{eff} scATAC-seq data for OCRs that contain motifs for different TFs (blue, null distribution for permuted dataset). TF families across the ranking shown at right. **f**, TF motif deviation scores per T_{eff} cell (as in **d**) for FOS and IRF4 motifs.

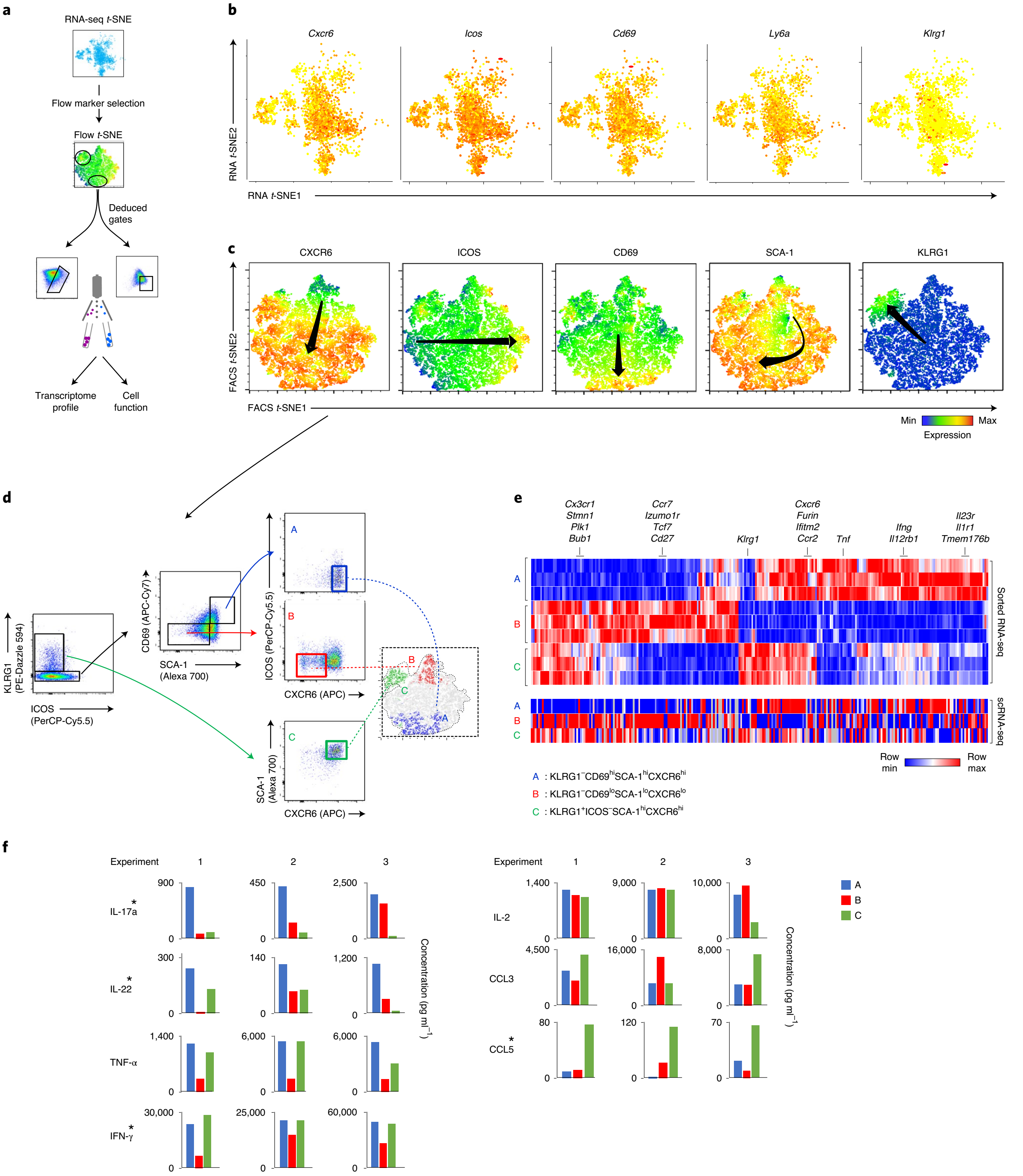
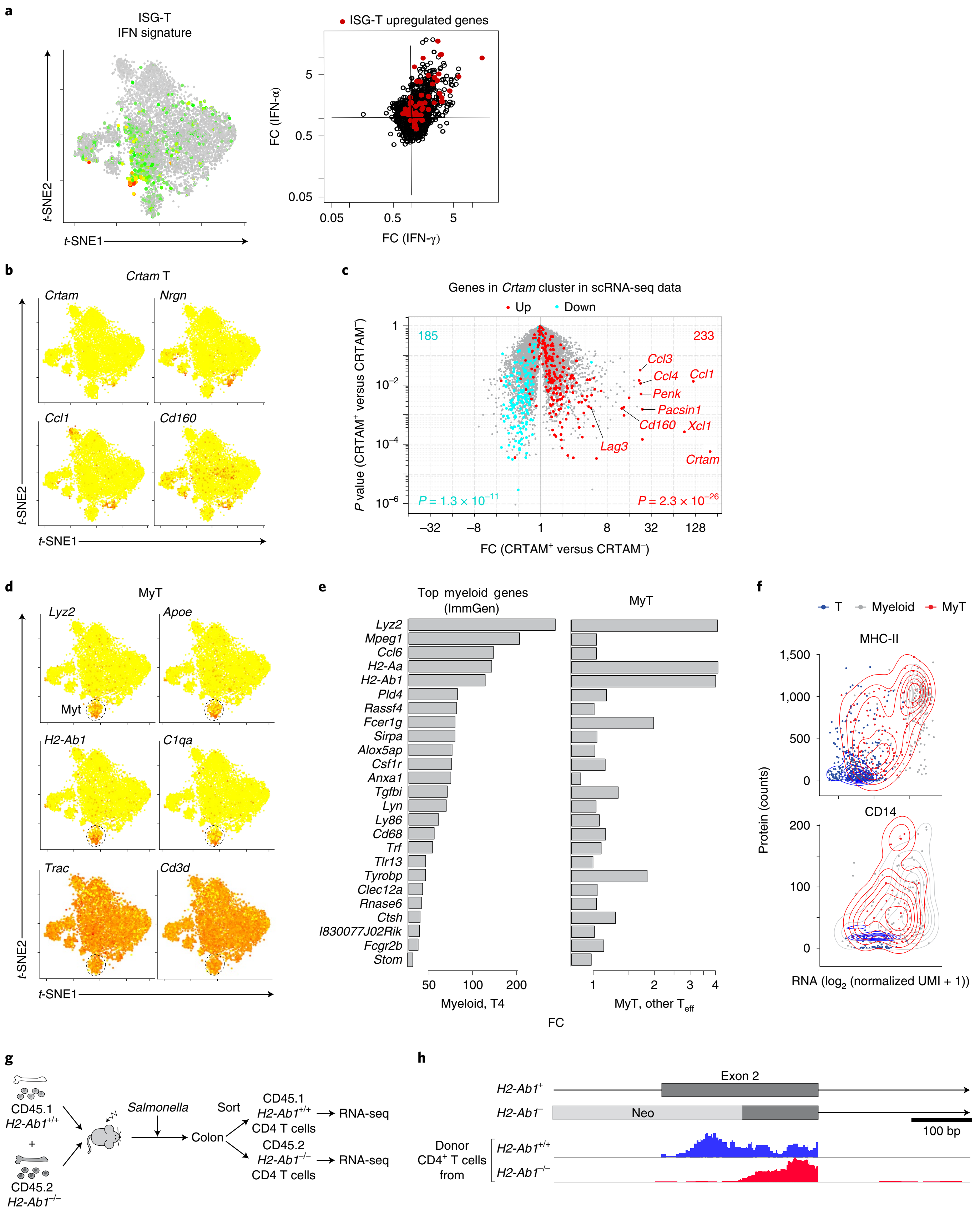


Fig. 6 | Transcriptional and functional validation of T_{eff} continuity. a, Experimental schematic. Surface markers with continuous distribution in the scRNA-seq were selected, and cells were stained with the corresponding antibodies for flow cytometry. *t*-SNE plots were computed from the cytometry data, from which sorting gates were set to prepare cells for transcriptional and functional analyses. **b**, Gene expression of selected surface markers in colonic T_{eff} cells from *Salmonella*-infected mice (scRNA-seq plot from Fig. 2c). **c**, Flow cytometry *t*-SNE generated from fluorescence intensities of CD4$^+$ T_{eff} cells stained for these markers. **d**, Sorting strategy, corresponding to the poles of the flow *t*-SNE data from **c**. The *t*-SNE positions of the sorted cells are shown at the right. **e**, Heatmap comparing differentially expressed genes in the bulk RNA-seq profiling of populations A, B and C, sorted in **d**. Data were hierarchically clustered and row-mean normalized. **f**, Multiplex ELISA comparing the secretion of cytokines and chemokines from populations A, B and C, where each bar is an independent biological replicate. *, cytokines with significant differences (at $P<0.05$) between any two populations as determined by paired Student's *t*-test (IL-17A, A versus B, $P=4.0\times10^{-4}$; IL-22, A versus B, $P=1.3\times10^{-2}$ and A versus C, $P=3.6\times10^{-2}$; IFN-γ, A versus B, $P=1.5\times10^{-2}$; CCL-5, A versus C, $P=2.7\times10^{-2}$).



H2-A^b and CD14 proteins (Fig. 7f) in MyT cells, at protein levels that were only somewhat lower than those seen in true myeloid cells. In the experiments in Fig. 4, αβTCRs detected in MyT cells were shared with other T$_{eff}$ cells from the same mice, suggesting that the MyT phenotype was not acquired during thymic differentiation, but late in the periphery after antigen encounter. MyT cells may

Fig. 7 | New T_{eff} populations. a, The ISG-T subset. Left, IFN type I signature[50] overlaid on the T_{eff} *t*-SNE plot. Right, genes overexpressed in the ISG-T cluster overlaid on top of genes upregulated in CD4+ T cells upon administration of IFN-α or IFN-γ[50]. **b**, scRNA-seq expression data of genes in the *Crtam*+ cluster. **c**, Volcano plot from RNA-seq data of sorted CRTAM+ versus CRTAM− colon T_{eff} cells; over- and underexpressed genes in the *Crtam*+ T cluster in scRNA-seq data are shown in red and blue, respectively, with significance of overlap. **d**, Expression in MyT cells of genes overlaid on the general *t*-SNE plot of Fig. 2c. Typical myeloid cell transcripts (top) and typical T cell transcripts (bottom). **e**, FC histograms of myeloid-specific genes. In myeloid versus CD4+ T cells (ImmGen RNA-seq data) (left) and in MyT versus other colon T_{eff} cells (*Salmonella*-infected, data from Fig. 2c) (right). The *x* axis is on a logarithmic scale. **f**, Contour plot representing RNA and protein expression in the single-cell data from Fig. 5 (*x* axis, normalized scRNA-seq; *y* axis, raw cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) counts) for MHC-II (top) or CD14 (bottom). Individual cells are represented by dots and are colored by their classification based on unsupervised clustering. **g**, Experimental schematic. Bone marrow from WT CD45.1 and CD45.2 *H2-Ab1*−/− mice was mixed and transferred to irradiated *Cd45.1/2* hosts. After 8 weeks, mice were infected with *Salmonella*, and 13 d later, the WT or KO LP CD4+ T cells were sorted for RNA-seq. **h**, Top, schematic representation of the WT or KO *H2-Ab1* loci (a neomycin resistance gene (neo) was inserted into the second exon[36]). Bottom, the position of RNA-seq reads in colonic CD4+ T_{eff} cells stemming from WT or *H2-Ab1* KO stem cells in mixed bone marrow chimeras infected with *Salmonella*.

correspond to the unusual CD3+CD14+ cells in human blood recently reported to also increase upon infection[35], which was attributed to stable doublets. However, several arguments suggested that MyT cells are not doublets (there were very few myeloid cells in the sorted CD4+TCRβ+ datasets, normal unique molecular identifiers (UMI) or cell counts; partial myeloid gene representation). Other than doublet formation or cell fusions, explanations for MyT cells include exosomal transfer of transcripts from myeloid to T_{eff} cells or the activation of unusual transcriptional modules. To formally resolve this issue, we created bone marrow chimeras with a 50:50 mix of congenically marked stem cells from wild-type (WT) and MHC-II-deficient donors with an inactivating neomycin insertion in *H2-Ab1*, a mutation that results in altered *Ab1* transcripts[36] (Fig. 7g). After reconstitution for 10 weeks, we sorted TCRβ+CD4+ cells of both donor origins for RNA-seq, analyzing the sequence reads at *H2-Ab1*. Should MyT cells result from doublets or mRNA transfer, TCRβ+CD4+ cells of knockout (KO) origin would have acquired WT *H2-Ab1* transcripts. This was not the case (Fig. 7h), as these T cells expressed transcripts from their own *H2-Ab1* gene. Thus, MyT cells are bona fide αβTCR+ T cells that activate a segment of the myeloid transcriptome. Their origin and significance remain to be established. However, there may be a precedent in the myeloid-like T cells that constitute the high-risk 'mixed phenotype acute leukemia' (ref. [37]).

Discussion

Our study set out to map the landscape of phenotypes that T_{eff} cells in the gut can adopt when stressed by microbial infection, which is related to the long-running question of T_{eff} cell heterogeneity. Whether evaluated at the transcriptome or the chromatin level, our results show that T_{eff} cells are molded by infections in a profound and specific manner, one that does not readily conform to T_H stereotypes and also gives rise to other intriguing new cell states.

From the realization over 40 years ago that distinct functions of T_H cells reside in different cells[1], the field has striven to subdivide T_{eff} cells into discrete subsets. Since the seminal discoveries of Mossman and Coffman[3] and the coining of the T_H1 and T_H2 semantic, these distinctions have been anchored by cytokine production, an anchor which has persisted despite repeated demonstrations of dual-expressing cells, T_H sub-subsets[22,23,38] and plasticity between T_H states[9,13,14]. Our results suggest that T_{eff} transcriptional identities form a 'polarized continuity' and cannot be parsed out into discrete T_H cell types, even in the context of infections expected to drive focused differentiation. Nor does progressing infection result in phenotypic divergence between clearly distinct states. This model does not imply homogeneity, however, as the different poles of the phenotypic cloud do show a strong preference for producing one cytokine over another (most marked for IL-4 or IL-5).

This view of T_{eff} cell heterogeneity differs from previously proposed concepts of cell plasticity, in which cells of defined phenotypes can switch between states that are otherwise coherent and reproducible[9,13,14]. The plasticity concept implies that discrete states do exist, but are not irrevocable. We find that there are no defined states to interconvert between. This view also diverges from the notion of sub-subsets (for example, pathogenic T_H17 cells (refs. [22,23,39,40])), which also implied discrete cell sets that could be further subdivided. Such sub-subsets also seemed absent and, in hindsight, may represent the spread of IL-17-producing cells across different regions of the phenotypic cloud.

One might argue that the polarized continuity represents transient intermediates between cell states. But, then, most cells would be intermediates. Velocity testing of differentiation within the T_{eff} continuum[41] gave no indication of directional progression, and the time course study showed no particular convergence toward more distinct T_{eff} phenotypes, overall or for amplified progeny of the same precursor. Importantly, chromatin analysis revealed that key controlling loci, *Rorc* and *Tbx21*, opened largely independently of each other.

Several studies are also consistent with this view of 'polarized continuity' within T_{eff} cells that is dominantly molded by microbes. Cloned human memory CD4+ T cells showed phenotypic divergence related to the initiating microbe[42]. Proteomic analysis by mass cytometry revealed a wide phenotypic range in CD4+ T_{eff} cells unleashed by *Ctla4* deficiency[43]. In tumor-infiltrating cells, scRNA-seq studies also found gradients of transcriptional phenotypes[44,45], as in other broad 'landscape' studies in which T cells were notoriously difficult to parse finely[46,47]. A recent analysis of airway-resident T cells also reported a continuous disposition of T_{eff} cells in house dust mite infection[32], showing that our results are not gut-specific. A continuous phenotypic spectrum was described for ILCs[48], contrasting with commonly used categorization[11]. Rather, ILC phenotypes can be described by a series of 'topics'[48] that are conceptually similar to and partially overlapping with the modules reported here. While this work was under review, Cano-Gamez et al. also proposed a model of human T cell activation in vitro dominated by 'continuous effectorness' (ref. [49]).

In conclusion, this study sheds light on the T cell response to infectious challenges: broad responses that adapt to each microbe, dominant coregulated gene modules that are not anchored by cytokines, different leading transcriptional drivers and intriguing new cell subsets.

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/s41590-020-00836-7.

Received: 9 September 2019; Accepted: 6 November 2020;
Published online: 18 January 2021

References

1. Araneo, B. A., Marrack, P. C. & Kappler, J. W. Functional heterogeneity among the T-derived lymphocytes of the mouse. II. Sensitivity of subpopulations to anti-thymocyte serum. *J. Immunol.* **114**, 747–751 (1975).
2. Bottomly, K. A functional dichotomy in CD4$^+$ T lymphocytes. *Immunol. Today* **9**, 268–274 (1988).
3. Mosmann, T. R. et al. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol.* **136**, 2348–2357 (1986).
4. Killar, L. et al. Cloned, Ia-restricted T cells that do not produce interleukin 4 (IL-4)/B cell stimulatory factor 1 (BSF-1) fail to help antigen-specific B cells. *J. Immunol.* **138**, 1674–1679 (1987).
5. Korn, T., Bettelli, E., Oukka, M. & Kuchroo, V. K. IL-17 and T$_H$17 cells. *Annu. Rev. Immunol.* **27**, 485–517 (2009).
6. Jabeen, R. & Kaplan, M. H. The symphony of the ninth: the development and function of T$_H$9 cells. *Curr. Opin. Immunol.* **24**, 303–307 (2012).
7. Crotty, S. Follicular helper CD4 T cells (T$_{FH}$). *Annu. Rev. Immunol.* **29**, 621–663 (2011).
8. Murphy, K. M. et al. Signaling and transcription in T helper development. *Annu. Rev. Immunol.* **18**, 451–494 (2000).
9. O'Shea, J. J. & Paul, W. E. Mechanisms underlying lineage commitment and plasticity of helper CD4$^+$ T cells. *Science* **327**, 1098–1102 (2010).
10. Zhu, J. & Paul, W. E. CD4 T cells: fates, functions, and faults. *Blood* **112**, 1557–1569 (2008).
11. Spits, H. et al. Innate lymphoid cells—a proposal for uniform nomenclature. *Nat. Rev. Immunol.* **13**, 145–149 (2013).
12. Kelso, A. T$_H$1 and T$_H$2 subsets: paradigms lost? *Immunol. Today* **16**, 374–379 (1995).
13. Murphy, K. M. & Stockinger, B. Effector T cell plasticity: flexibility in the face of changing circumstances. *Nat. Immunol.* **11**, 674–680 (2010).
14. Geginat, J. et al. Plasticity of human CD4 T cell subsets. *Front. Immunol.* **5**, 630 (2014).
15. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).
16. Miragaia, R. J. et al. Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation. *Immunity* **50**, 493–504 (2019).
17. Lee, J. Y. et al. The transcription factor KLF2 restrains CD4$^+$ T follicular helper cell differentiation. *Immunity* **42**, 252–264 (2015).
18. Stoeckius, M. et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
19. Yosef, N. et al. Dynamic regulatory network controlling T$_H$17 cell differentiation. *Nature* **496**, 461–468 (2013).
20. Hartigan, J. A. & Hartigan, P. M. The dip test of unimodality. *Ann. Stat.* **13**, 70–84 (1985).
21. Pavlidis, N. G., Hofmeyr, D. P. & Tasoulis, S. K. Minimum density hyperplanes. *J. Mach. Learn. Res.* **17**, 5414–5446 (2016).
22. Lee, Y. et al. Induction and molecular signature of pathogenic T$_H$17 cells. *Nat. Immunol.* **13**, 991–999 (2012).
23. Ghoreschi, K. et al. Generation of pathogenic T$_H$17 cells in the absence of TGF-β signalling. *Nature* **467**, 967–971 (2010).
24. Omenetti, S. et al. The intestine harbors functionally distinct homeostatic tissue-resident and inflammatory T$_H$17 cells. *Immunity* **51**, 77–89 (2019).
25. Meredith, M., Zemmour, D., Mathis, D. & Benoist, C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. *Nat. Immunol.* **16**, 942–949 (2015).
26. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
27. Yoshida, H. et al. The *cis*-regulatory atlas of the mouse immune system. *Cell* **176**, 897–912 (2019).
28. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
29. Aghaeepour, N. et al. GateFinder: projection-based gating strategy optimization for flow and mass cytometry. *Bioinformatics* **34**, 4131–4133 (2018).
30. Becht, E. et al. Reverse-engineering flow-cytometry gating strategies for phenotypic labelling and high-performance cell sorting. *Bioinformatics* **35**, 301–308 (2019).
31. Baron, C. S. et al. Cell type purification by single-cell transcriptome-trained sorting. *Cell* **179**, 527–542 (2019).
32. Tibbitt, C. A. et al. Single-cell RNA sequencing of the T helper cell response to house dust mites defines a distinct gene expression signature in airway T$_H$2 cells. *Immunity* **51**, 169–184 (2019).
33. Arazi, A. et al. The immune cell landscape in kidneys of patients with lupus nephritis. *Nat. Immunol.* **20**, 902–914 (2019).
34. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
35. Burel, J. G. et al. The challenge of distinguishing cell–cell complexes from singlet cells in non-imaging flow cytometry and single-cell sorting. *Cytometry* **97**, 1127–1135 (2020).
36. Cosgrove, D. et al. Mice lacking MHC class II molecules. *Cell* **66**, 1051–1066 (1991).
37. Alexander, T. B. et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373–379 (2018).
38. Priceman, S. J. et al. Regulation of adipose tissue T cell subsets by Stat3 is crucial for diet-induced obesity and insulin resistance. *Proc. Natl Acad. Sci. USA* **110**, 13079–13084 (2013).
39. McGeachy, M. J. et al. TGF-β and IL-6 drive the production of IL-17 and IL-10 by T cells and restrain T$_H$-17 cell-mediated pathology. *Nat. Immunol.* **8**, 1390–1397 (2007).
40. Krausgruber, T. et al. T-bet is a key modulator of IL-23-driven pathogenic CD4$^+$ T cell responses in the intestine. *Nat. Commun.* **7**, 11627 (2016).
41. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
42. Becattini, S. et al. Functional heterogeneity of human memory CD4$^+$ T cell clones primed by pathogens or vaccines. *Science* **347**, 400–406 (2015).
43. Wei, S. C. et al. Negative co-stimulation constrains T cell differentiation by imposing boundaries on possible cell states. *Immunity* **50**, 1084–1098 (2019).
44. Li, H. et al. Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell* **176**, 775–789 (2019).
45. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
46. Wu, T. D. et al. Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature* **579**, 274–278 (2020).
47. Zilionis, R. et al. Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity* **50**, 1317–1334 (2019).
48. Bielecki, P. et al. Skin inflammation driven by differentiation of quiescent tissue-resident ILCs into a spectrum of pathogenic effectors. Preprint at *bioRxiv* https://doi.org/10.1101/461228 (2018).
49. Cano-Gamez, E. et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4$^+$ T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).
50. Mostafavi, S. et al. Parsing the interferon transcriptional network and its disease associations. *Cell* **164**, 564–578 (2016).

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

The Immunological Genome Project Consortium

Oscar Aguilar[9], Rhys Allan[10], Jilian Astarita[11], K. Frank Austen[12], Nora Barrett[12], Alev Baysoy[13], Christophe Benoist[13], Brian D. Brown[14], Matthew Buechler[11], Jason Buenrostro[15], Maria Acebes Casanova[16], Kyunghee Choi[17], Kaitavjeet Chowdhary[13], Marco Colonna[17], Ty Crowl[18], Tianda Deng[18], Jigar V. Desai[19], Fiona Desland[16], Maxime Dhainaut[14], Jiarui Ding[20], Claudia Dominguez[11], Daniel Dwyer[12], Michela Frascoli[21], Shani Gal-Oz[22], Ananda Goldrath[18], Ricardo Grieshaber-Bouyer[12],

Baosen Jia[16], Tim Johanson[10], Stefan Jordan[16], Joonsoo Kang[21], Varun Kapoor[11], Ephraim Kenigsberg[16], Joel Kim[16], Ki wook Kim[17], Evgeny Kiner[13], Mitchell Kronenberg[23], Lewis Lanier[13], Catherine Laplace[13], Caleb Lareau[15], Andrew Leader[16], Jisu Lee[21], Assaf Magen[16], Barbara Maier[16], Alexandra Maslova[24], Diane Mathis[13], Adelle McFarland[17], Miriam Merad[16], Etienne Meunier[24], Paul Monach[12], Sara Mostafavi[24], Soren Muller[11], Christoph Muus[20], Hadas Ner-Gaon[22], Quyhn Nguyen[18], Peter A. Nigrovic[12], German Novakovsky[24], Stephen Nutt[10], Kayla Omilusik[18], Adriana Ortiz-Lopez[13], Mallory Paynich[23], Vincent Peng[17], Marc Potempa[13], Rachana Pradhan[11], Sara Quon[18], Ricardo Ramirez[13], Deepshika Ramanan[13], Gwendalyn Randolph[17], Aviv Regev[20], Samuel A. Rose[14], Kumba Seddu[13], Tal Shay[22], Avishai Shemesh[13], Justin Shyer[11], Christopher Smilie[20], Nick Spidale[21], Ayshwarya Subramanian[20], Katelyn Sylvia[21], Julie Tellier[10], Shannon Turley[11], Brinda Vijaykumar[13], Amy Wagers[15], Chendi Wang[24], Peter L. Wang[17], Aleksandra Wroblewska[14], Liang Yang[13], Aldrin Yim[17] and Hideyuki Yoshida[25]

[9]Department of Microbiology & Immunology, University of California, San Francisco, San Francisco, CA, USA. [10]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [11]Department of Cancer Immunology, Genentech, South San Francisco, CA, USA. [12]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, USA. [13]Department of Immunology, Harvard Medical School, Boston, MA, USA. [14]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [15]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [16]Immunology Institute and Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [17]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. [18]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [19]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. [20]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA. [21]Department of Pathology, University of Massachusetts Medical School, Worcester, MA, USA. [22]Department of Life Sciences, Ben-Gurion University of the Negev, Be'er Sheva, Israel. [23]La Jolla Institute for Immunology, La Jolla, CA, USA. [24]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. [25]YCI Laboratory for Immunological Transcriptomics, RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan.

Methods

Mice. Male C57BL/6 mice were purchased from Jackson Laboratory. *Il17A*$^{GFP/+}$ mice (JAX, C57BL/6-*Il17a*^{tm1Bcgen}/J) were a gift from J. Huh (Harvard University). OT-II TCR-transgenic mice were obtained from Jackson Laboratory (B6. Cg-Tg(*TcraTcrb*)425Cbn/J). *H2-Ab1*-deficient mice were previously described[36]. To construct bone marrow chimeras, bone marrow cells were harvested from both femurs and tibias and treated with ACK buffer (Lonza) to remove red blood cells. *Cd45.1*$^{-/-}$;*Cd45.2*$^{-/-}$ mice were irradiated (10 Gy) and reconstituted with equal proportions (~5 million cells each) of *Cd45.1* and *Cd45.2* (*Ab* KO) bone marrow cells. All mice were bred and maintained in our specific pathogen-free facilities at Harvard Medical School (IACUC protocols IS1257, IS187-3, IS2221).

Flow cytometry. Cells from colon LP were prepared as previously described[51]. Briefly, intestinal tissues were treated with RPMI containing 1 mM dithiothreitol, 20 mM EDTA and 2% FBS at 37 °C for 15 min to remove epithelial cells, and then they were minced and dissociated in collagenase solution (1.5 mg ml^{-1} collagenase II (Gibco), 0.5 mg ml^{-1} dispase (Gibco) and 1% FBS in RPMI) with constant stirring at 37 °C for 45 min. Single-cell suspensions were then filtered and washed with a 4% RPMI solution. For cytokine analyses, cells were treated with 10 ng ml^{-1} phorbol 12-myristate 13-acetate (Sigma), 1 μM ionomycin (Sigma) and 1× protein transport inhibitor cocktail (eBioscience, 00-4980-03) for 3.5 h in 10% FBS, RPMI. For intracellular staining of cytokines and TFs, cells were stained for surface markers and fixed in eBioscience Foxp3 buffer overnight, followed by permeabilization in eBioscience (both 00-5523-00) buffer for 45 min in the presence of antibodies. Fluorescence profiles were acquired on a BD Symphony instrument, and analyses were performed with FlowJo (Tree Star) software.

Antibodies used in the study included anti-mouse (m)CD45 (30-F11), anti-mCD19 (6D5), anti-mCD4 (RM4-5), anti-mTCRβ (H57-597), anti-mCD44 (IM7), anti-mCD25 (PC61), anti-mFOXP3 (FJK-16s), anti-mIFN-γ (XMG1.2), anti-mIL-17A (TC-11-18H10.1), anti-mIL-5 (TRFK5), anti-mIL-13 (W17010B), anti-mCCR6 (29-2L17), anti-mIL-1R2 (4E2), anti-mKLRG1 (2F1/KLRG1), anti-mICOS (C398.4A), anti-mCXCR6 (SA051D1), anti-mCD69 (H1.2F3), anti-SCA-1 (D7), anti-mCRTAM (11-5/CRTAM), anti-I/A I-E(M5/114.15.2), anti-mCD45.1 (A20), and anti-mCD45.2 (104). All antibodies were diluted 1:100, with the exception of CD25 (1:50).

For the '*t*-SNE sort', the goal was to sort cells defined combinatorially by a panel of markers, even if they were not readily identifiable as well-demarcated populations on conventional two-parameter flow cytometry profiles. Flow cytometry *t*-SNE plots were generated in FlowJo version 10 from gated CD4$^+$TCRβ$^+$CD44$^+$CD25$^-$ cells stained for markers found by manual inspection to have non-discrete and non-correlated expression in the scRNA-seq data and thus were most appropriate to represent the phenotypic continuity within the T$_{eff}$ phenotypic cloud (KLRG1, ICOS, CD69, SCA-1 and CXCR6). Selected regions that demarcated clusters on the *t*-SNE continuum were then backgated onto normal two-parameter plots, from which gating instructions interpretable by a cell sorter were drawn (by a manual and iterative process). The gates were drawn such that more than 90% of the events in the selected region would be within the sort gates. These combinatorial gates were then applied to sort colonic T$_{eff}$ cells.

Multiplex ELISA. Cells (1,000 to 10,000) were sorted (single sort) into 100 μl T cell medium (RPMI 1640, 10% FBS, 20 mM HEPES, 1 mM sodium pyruvate, 0.05 mM 2-mercaptoethanol, 2 mM L-glutamine, 100 mg ml^{-1} streptomycin and 100 mg ml^{-1} penicillin). Cells were plated in round-bottom 96-well plates with a 1:1 ratio of anti-CD3/CD28 beads (Miltenyi) and were incubated at 37 °C for 24 h. Supernatants were collected and analyzed with the LEGENDplex T Helper Cytokine version 2 and the Proinflammatory Chemokine Panel kits (BioLegend) according to the manufacturer's instructions. Samples were acquired with the BD Symphony instrument and analyzed with LEGENDplex software. Paired Student's *t*-test was used for significance assessment.

Infections. For infection with *Salmonella*, mice were gavaged with 100 μl of 200 mg ml^{-1} streptomycin in water and, 24 h later, gavaged with 10^9 *S. enterica* (serovar Typhimurium) Δ*aroA*[52] (a gift from D. Littman, New York University). For infection with *Citrobacter*, mice were gavaged with 5×10^8 *C. rodentium*[53]. Unless noted otherwise, mice were sacrificed at day 13 after infection. For helminth infections[54], mice were gavaged with 200 *H. polygyrus* L3 larvae in 200 μl H_2O or subcutaneously injected with 500 L3 larvae of *N. brasiliensis* in 100 μl PBS and sacrificed 11 d later.

Low input RNA-seq. All cells were double-sorted. For the final sort, 1,000 cells were collected directly into 5 μl lysis buffer (TCL buffer (Qiagen) with 1% 2-mercaptoethanol), and the lysates were frozen after 5 min. Smart-seq2 libraries were prepared as previously described[12]. Reads were aligned to the mouse genome (GENCODE GRCm38/mm10 primary assembly and gene annotation version M16; https://www.gencodegenes.org/mouse/release_M16) or to the human genome (GENCODE human release 27; reference genome sequence, GRCh38/hg38; annotation, GENCODE version 27) with STAR 2.5.4a. The ribosomal RNA gene annotations were removed from the general transfer format (GTF) file. The gene-level quantification was calculated by featureCounts (http://subread.sourceforge.net/). Raw read count tables were normalized by the median of ratios method with the DESeq2 package from Bioconductor and then converted to GenePattern GCT and CLS format. Samples with less than 3 million uniquely mapped reads were automatically excluded from normalization to mitigate the effect of samples with poor quality on normalized counts. Normalized read counts were filtered for robust expression (>10) to avoid confounders from low-level noise and processed in the Multiplot suite and Morpheus (https://software.broadinstitute.org/morpheus/). PCA was done using the prcomp function in R on all genes with expression higher than 0 in any sample.

Single-cell RNA-seq. Intestinal tissues were treated with RPMI containing 1 mM dithiothreitol, 20 mM EDTA and 2% FBS at 37 °C for 15 min to remove epithelial cells, and then they were minced and dissociated in 1 mg ml^{-1} collagenase VIII (Sigma), 1 μg ml^{-1} DNase and 1% FCS in RPMI with constant stirring at 37 °C for 20 min. Single-cell suspensions were then filtered and washed with 4% FCS in RPMI medium. Single-cell suspensions were stained on ice for 30 min with antibodies to CD4, TCRβ, CD19 and CD45 (BioLegend) and 20 ng ml^{-1} DAPI (BioLegend) as a viability dye. T cells were then sorted on an Astrios MoFlo instrument (Beckman Coulter) as DAPI$^-$CD45$^+$CD4$^+$TCRβ$^+$CD19$^-$ cells. For single-sample processing, cells were sorted directly into PBS with BSA for a final concentration of 0.04% BSA. For cell hashtagging, TotalSeq-A hashtag antibodies (SPF, hashtag 1; *C. rodentium*, hashtag 2; *S. enterica*, hashtag 3; *N. brasiliensis*, hashtag 4; *H. polygyrus*, hashtag 5) were added to each sample individually at the same time as other antibodies. All samples were sorted together directly into RPMI with 2% FCS and subsequently spun down and reconstituted in 33 μl PBS with 0.04% BSA. All samples were loaded on the 10x Chromium Controller (10x Genomics) within 30 min of sorting. Libraries were prepared using Chromium Single Cell 3′ Reagent Kits version 2 according to the manufacturer's protocol. Hashtag oligonucleotide (HTO) libraries were prepared as described in ref. [18]. Libraries were sequenced together on the Illumina HiSeq 4000.

Single-cell RNA-seq data analysis. Gene counts were obtained by aligning reads to the mm10 genome using Cell Ranger software (version 1.3) (10x Genomics). HTO counts were obtained by using the CITE-seq-Count package[34]. Single-cell data were initially analyzed using the Seurat package[55]. HTOs were assigned to cells using the HTODemux function, and doublets were eliminated from analysis. Cells with less than 1,000 UMIs or 400 genes and more than 4,000 UMIs or 0.05% of reads mapped to mitochondrial genes were also excluded from the analysis. T$_{reg}$ cells and naive CD4$^+$ cells were removed from analysis by using the SubsetData function. Data were normalized using the NormalizeData function and scaled using the ScaleData function, regressing out number of UMIs and percentage of expressed mitochondrial genes. Variable genes were found by the FindVariableGenes function, using genes with mean expression over 0.0125 and four UMIs per cell. Dispersion cutoff was calculated based on the Fano factor distribution per gene. By these means, 550–950 variable genes were selected in different T$_{eff}$ datasets. PCs were calculated using the RunPCA function, and significant PCs were selected using the JackStraw function. *t*-SNE and KNN clusters were computed on significant PCs using the RunTSNE and FindClusters functions, respectively. UMAP dimensionality reduction was calculated on significant PCs using the RunUMAP function. T$_H$ signatures scores were computed as the mean expression of signature genes per cell.

Diffusion maps are useful for identifying differentiation trajectories, as they allow for pseudotemporal ordering of single cells in a high-dimensional gene expression space[56]. Diffusion maps were generated using the Seurat package RunDiffusion function with default settings.

Imputation can denoise the cell count matrix and fill in missing transcripts by data diffusion[57]. Imputation was performed using the built-in Seurat AddImputedScore function with default parameters on all variable genes. PCs and *t*-SNE data were then *r*-recomputed based on the imputed values.

PCs were identified and plotted using the Seurat PCHeatmap function with default parameters.

Correlation coefficient analysis (CCA)[55] was performed by running the RunMultiCCA function on 500 variable genes between the four samples. Twenty significant correlation coefficients (CC) were selected for alignment using the AlignSubspace function. *t*-SNE and KNN clustering were run as previously, based on 20 CCs.

To compute Euclidean distances within groups of cytokine-expressing cells, cytokine-positive cells were identified as expressing one or more normalized UMIs. Distances between each selected cell to other cells were calculated for the 1,000 top variable genes using the dist function in R. *P* values were computed using the Mann–Whitney test. For dendrogram analysis, cytokine-expressing cells were identified as above, and distances between different samples expressing different cytokines were computed by the dist function in R with default settings on the top 1,000 variable genes. Hierarchical clustering (hclust function in R) was then employed to generate the dendrogram.

Highly T$_H$-specific gene sets (Supplementary Table 2) were generated by manual curation, starting mainly from published signatures as well as other scRNA-seq datasets[19,32,58–61] and selecting genes that were reproducibly present in these signatures. We removed transcripts that overlapped between resulting T$_H$

gene sets, often simple markers of cell activation frequent in such signatures, as well as some non-T transcripts that frequently contaminate published signatures (for example, *Cd19*, *Cd79a*, *Cd8a*). We also added several transcripts known to correlate with *Ifng*, *Il17a* or *Il4* and *Il13* (*Cxcr3*, *Tmem176a*, *Areg*). The gene signature average for these genes was then calculated with the AddModuleScore function in Seurat version 3. Expression of cell cycle genes was calculated based on the CellCycleScoring function in Seurat version 3 (cc.genes based on ref. [62])

Gene module generation. After filtering transcripts for robust expression (those that appeared in more than ten cells in any one of the infected or SPF samples), gene–gene correlations (Pearson, cor function in R) were calculated within each dataset. The ten matrices (one for each replicate and condition) of pairwise gene–gene correlations were then averaged for Extended Data Fig. 5c.

To select the genes with the highest correlations, a threshold correlation score in the 98th percentile was calculated for each gene, and 588 genes with correlation scores higher than 0.05 were selected for further analysis. Gene modules were then identified by affinity propagation[63] using the APCluster R package with a negative distance similarity function, and the number of input similarities (q) was set to 0. Gene modules were overlaid on the t-SNE plot by computing the mean expression of module genes for each cell.

Clustering approaches. *BackSPIN.* The data were normalized with Seurat parameters and then subset to the top 588 most variable genes according to the Seurat pipeline. To determine whether significant clusters would emerge from more elaborate clustering methods, we used BackSPIN, an unsupervised biclustering method that sorts both genes and cells into clusters[64]. The motivation behind BackSPIN was that by iterative partitioning, the algorithm would be able to cluster true cell subsets and gene subsets together. One important parameter for BackSPIN involves defining the partitioning 'rate' (that is, how much to subset the groups at each iterative process). This was set at the default of 0.1. Other parameters specified were the number of levels (numLevels) to partition by (set at 2), the number of top variable genes to cluster (set at 596), the initial number of iterations (first_run_iters; set to 10) and subsequent number of iterations (runs_iters; set to 8). The default initial decrease rate of 0.1 (first_run_iters) and the default subsequent decrease rate (runs_step) of 0.3 were used. The decrease rates helped to determine the precision of clusters. Finally, threshold values were set at the default value of 2 for both minimum numbers of cells (split_limit_c) and genes (split_limit_g). A threshold score of 1.15 was used to determine when to stop partitioning the data (stop_const), and the default threshold for determining which group a gene would be assigned to was kept at 0.015.

BISCUIT. BISCUIT iteratively learns to identify features in each cluster and create clusters with these specific features by imputing and normalizing the data[45]. The motivation behind BISCUIT is that by imputing the data, variation provided by genes that may have dropped out is captured. The major parameter for BISCUIT is the dispersion parameter (α) that allows the algorithm to sort cells into more clusters or less clusters, which was set to 1. The following parameters were used to run BISCUIT: the default setting of 20 genes per batch, the default number of 20 iterations and 100 as the number of cells in each batch. Once complete, the final clusters were projected onto the t-SNE plot of Fig. 2c computed by Seurat. Cell cluster outputs from BISCUIT were projected onto the t-SNE data computed by Seurat.

Dip test. Data were normalized with standard Seurat parameters as described previously[55]. The same number of variable genes, defined by Seurat, was used in the continuity analysis. To test for 'discontinuity' in transcriptomic-based representation of a set of cells, the Hartigan's dip test of multimodality was used[20]. The dip test asks whether the pairwise distances between all pairs of cells can be best supported by a unimodal or a multimodal distribution. The intuition behind this test comes from the fact that if there are two or more clear subpopulations of cells that cluster together with clear boundaries, then, given a high-dimensional representation of these cells (that is, vectors of length g consisting of gene expression levels for g genes), there would be one or more region(s) of low density in between highly dense regions in this space. These low-density regions would thus create a 'dip' in the distribution of pairwise distances between all cells in this space. One important parameter here is the representation of gene expression data used in computing the pairwise distances between cells. To support the ability of the dip test to identify regions of low density, we first applied a projection defined by minimum separation hyperplane[21] to gene expression data from variable genes (defined by Seurat) and then applied the dip test to the distances computed on the projected data.

Binary classification of *Il17*- or *Ifng*-expressing cells. We trained a DNN run on the Keras platform (https://keras.io/). The input gene set was the 500 most variable genes across the entire scRNA-seq dataset of Fig. 2 (naturally leaving out *Il17* and *Ifng* transcripts), and the network was trained to classify *Ifng*- or *Il17a*-expressing cells (randomly assigned to 80% training set, 20% test set). The data matrix was normalized by the mean of the expression of each gene across the 2,885 cells (otherwise the transcripts with highest expression levels dominate the output). The DNN was composed of three hidden layers with the following features: size of the hidden layers, 512, 128 or 64 with random weights initialization; activation function, sigmoid; optimizer for backward propagation, ADAM; number of epochs, 50; training and testing on CPU; batch size, 100. We added a decision function downstream with the possibility of NoCall (for non-producing cells); the classification as *Il17*-expressing was accepted if the output softmax score of the cell was above 0.95 (and below 0.05 for *Ifng*-expressing cells), otherwise the NoCall decision was made. We voluntarily overfitted the model to fit the distribution of the output softmax score with the decision function constraints. We used a Keras-based (version 2.2.4) neural network (https://keras.io/) on Python 2. The integrated gradients library was used to compute the overall contribution score of each gene as the mean of its contribution scores across the whole dataset. To test the reproducibility of the integrated gradients, we randomly split the dataset into two subdatasets on which we independently trained models, repeating the operation 100 times on each dataset and taking the mean of these 100 scores. As a positive control, the same architecture was used to distinguish T_{eff} from T_{reg} cells, which could be done with 98.8% (T_{eff}) and 89.7% (T_{reg}) accuracy on average, as shown below.

Accuracy, 97.58–98.05%	Actual T_{reg}	Actual T_{eff}
Prediction T_{reg} (in ten independent runs)	1,120–1,170	42–64
Prediction T_{eff} (in ten independent runs)	78–128	5,700–5,722
Total	1,248	5,764

Total number of cells, 7,012.

Clonotype analysis and CITE-seq. Mice were infected with *Salmonella* as above, and colon single-cell suspensions were prepared as above. Antibody staining (cell hashing and CITE-seq) was performed simultaneously by adding TotalSeq-C hashtags 1–7 (day 0, hashtag7; day 3, hashtag6; day 5, hashtag 5; day 7, hashtag 4; day 10, hashtag 3; day 17, hashtag 1), anti-CD14 (C0424) and anti-I-A/I-E (C0117) (BioLegend) to the cells at a ratio of 1:100 in RPMI with 2% FCS and incubating the mixture on ice for 15 min. Cells were then washed twice with RPMI, 2% FCS and sorted as described above before encapsulation (10x Genomics). Gene expression, feature and TCR V(D)J libraries were prepared using the 5′ V(D)J version 1 kit (10x Genomics). Rearranged TCRs were identified by running Cell Ranger vdj 3.0, and TCR chains and N and P nucleotides per clonotype were determined with the help of the IMGT database (http://www.imgt.org/IMGT_vquest/input). Repeated clonotypes were defined by shared TCRα and -β receptors with identical *Cdr3* sequences at the nucleotide level. Cells in cycle were excluded from UMAP and clonotype analyses. *Ifng*- or *Il17a*-expressing cells were defined as cells that had reads for either transcript. Euclidean distances between cells expressing the same repeated TCR clonotype were measured using the dist() function on either the T_H gene set (Supplementary Table 2) or the 1,000 most variable genes.

Single-cell ATAC-seq. Total CD4$^+$ T cells were isolated from the colons of *Salmonella*-infected mice as described for scRNA-seq, except collagenase II and dispase was used instead of collagenase VIII. Cells (25×10^3) were sorted directly into 2% FCS, RPMI and subsequently spun down and reconstituted in 0.04% PBS. Nuclei isolation, GEM generation and library preparation were performed as described in the Chromium Single Cell ATAC (10x Genomics) manual (https://support.10xgenomics.com/single-cell-atac). Libraries were sequenced on the Illumina NextSeq system. OCR counts were obtained by aligning reads to the mm10 genome using Cell Ranger ATAC software (version 1.1) (10x Genomics). scATAC-seq was analyzed using the Seurat–Signac pipeline (https://satijalab.org/signac/index.html). For QC, cells with less than 5,000 peak calls and less than 20% of reads mapped to peaks were filtered out. For the normalization of peak counts used to drive the UMAP representation, the RunTFIDF function was used to calculate the term frequency–inverse document frequency (TF–IDF). For dimensionality reduction, data structure was learned via latent semantic indexing (RunLSI function) and single value decomposition (RunSVD function). Contaminating non-T cells were taken out, and UMAP and cell clusters were then recalculated. Naive T, T_{reg} and T_{eff} cell clusters were identified and attributed based on the gene activity matrix, constructed using the FeatureMatrix function and the Gencode version 18 annotation; peaks that were found within the gene body and up to 2 kb upstream of transcription start sites (TSSs) were assigned to the corresponding genes.

To calculate the *Tbx21* and *Rorc* scores shown in Fig. 5b, we counted the raw reads falling into 300-bp intervals centered on OCRs that were highly correlated with the expression of corresponding genes in the ImmGen compendium (according to Supplementary Tables 3f and ref. [27]). For the *Rorc* locus, signals at three OCRs with TSS gene–OCR correlation scores >10 were used, and for the *Tbx21* locus, 11 OCRs with TSS gene–OCR scores >15 were used (Fig. 5c). Read counts were then summed and averaged into a score per cell using the AddModuleScore function. Cells were assigned as *Rorc*$^+$ or *Tbx21*$^+$ if the average OCR score for these loci was greater than 0. Coverage maps were then generated using the CoveragePlot function, applied only to T_{eff} cells.

Raw bulk ATAC-seq data from T_H0, T_H1 and T_H17 cells differentiated in vitro were generously provided by P. Thakore and A. Schnell (Harvard University)[65]. The *Tbx21* and *Rorc* chromatin scores were computed as above from read counts (normalized to the total read number for each biological replicate).

TF deviation and variability scores were calculated using the chromVAR package (version 1.8)[28] with motifs from the JASPAR 2018 database. The filtered T_{eff}-only scATAC-seq count matrix was used as input, with peaks overlapping motifs determined using the motifmatchr matchMotifs function. The chromVAR computeDeviations function was used to calculate the bias-corrected deviation scores for each TF motif. Briefly, this method computes the difference between observed fragments within peaks containing a given motif and the total expected number of fragments using the average of all cells. These 'raw deviation' scores are then normalized for technical biases using a set of background peaks matched for GC content and accessibility to yield the 'bias-corrected deviation scores'. Variability of TF motifs across the T_{eff} data was calculated using the chromVAR computeVariability function.

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

Data availability

The data reported in this paper were deposited in the Gene Expression Omnibus (GEO) database under accession no. GSE160055).

References

51. Sefik, E. et al. Individual intestinal symbionts induce a distinct population of RORγ$^+$ regulatory T cells. *Science* **349**, 993–997 (2015).
52. Hess, J., Ladel, C., Miko, D. & Kaufmann, S. H. *Salmonella typhimurium aroA*$^-$ infection in gene-targeted immunodeficient mice: major role of CD4$^+$ TCR-αβ cells and IFN-γ in bacterial clearance independent of intracellular location. *J. Immunol.* **156**, 3321–3326 (1996).
53. Collins, J. W. et al. *Citrobacter rodentium*: infection, inflammation and the microbiota. *Nat. Rev. Microbiol.* **12**, 612–623 (2014).
54. Camberis, M., Le, G. G. & Urban, J. Jr. Animal model of *Nippostrongylus brasiliensis* and *Heligmosomoides polygyrus*. *Curr. Protoc. Immunol.* **55**, 19.12.1–19.12.27 (2003).
55. Butler, A. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
56. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
57. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
58. Ciofani, M. et al. A validated regulatory network for T_H17 cell specification. *Cell* **151**, 289–303 (2012).
59. Muranski, P. et al. T_H17 cells are long lived and retain a stem cell-like molecular signature. *Immunity* **35**, 972–985 (2011).
60. Nurieva, R. I. et al. Generation of T follicular helper cells is mediated by interleukin-21 but independent of T helper 1, 2, or 17 cell lineages. *Immunity* **29**, 138–149 (2008).
61. Yusuf, I. et al. Germinal center T follicular helper cell IL-4 production is dependent on signaling lymphocytic activation molecule receptor (CD150). *J. Immunol.* **185**, 190–202 (2010).
62. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
63. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**, 2463–2464 (2011).
64. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
65. Wang, C. et al. Metabolic and epigenomic regulation of T_H17/T_{reg} balance by the polyamine pathway. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.23.911966 (2020).

Acknowledgements

We thank K. Murphy, N. Yosef, V. Kuchroo, R. Ramirez, D. Ramanan, M. Sassone-Corsi, E. Pamer, A. Anderson and J. Huh for insightful discussions, data and mouse lines and K. Hattori, C. Araneo, K. Seddu, D. Dionne and the Klarman Cell Observatory team and N. El-Ali and the Bauer Core Facility for help with mice, cell sorting and single-cell profiling. This work was supported by grants from the NIH to C.B. and D.M. (AI125603) and to the ImmGen consortium (AI072073). E.K. was supported by a PhD fellowship from Boehringer Ingelheim Fonds. K.C. was supported by NIGMS-T32GM007753 and a Harvard Stem Cell Institute MD/PhD Training Fellowship.

Author contributions

E.K. and E.W. performed experiments. E.K., B.V., K.C., H.S., S.M. and C.B. analyzed and interpreted data. A.S., P.I.T., J.C. and G.L. provided data or reagents. E.K., S.M., D.M. and C.B. designed the study and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

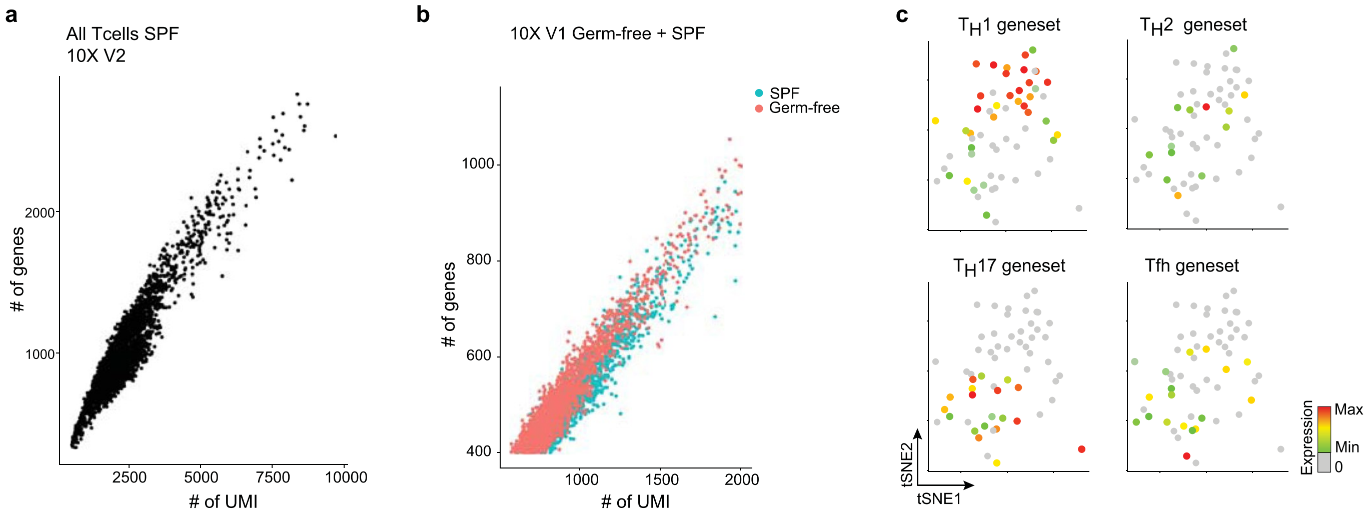
Extended data is available for this paper at https://doi.org/10.1038/s41590-020-00836-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41590-020-00836-7.

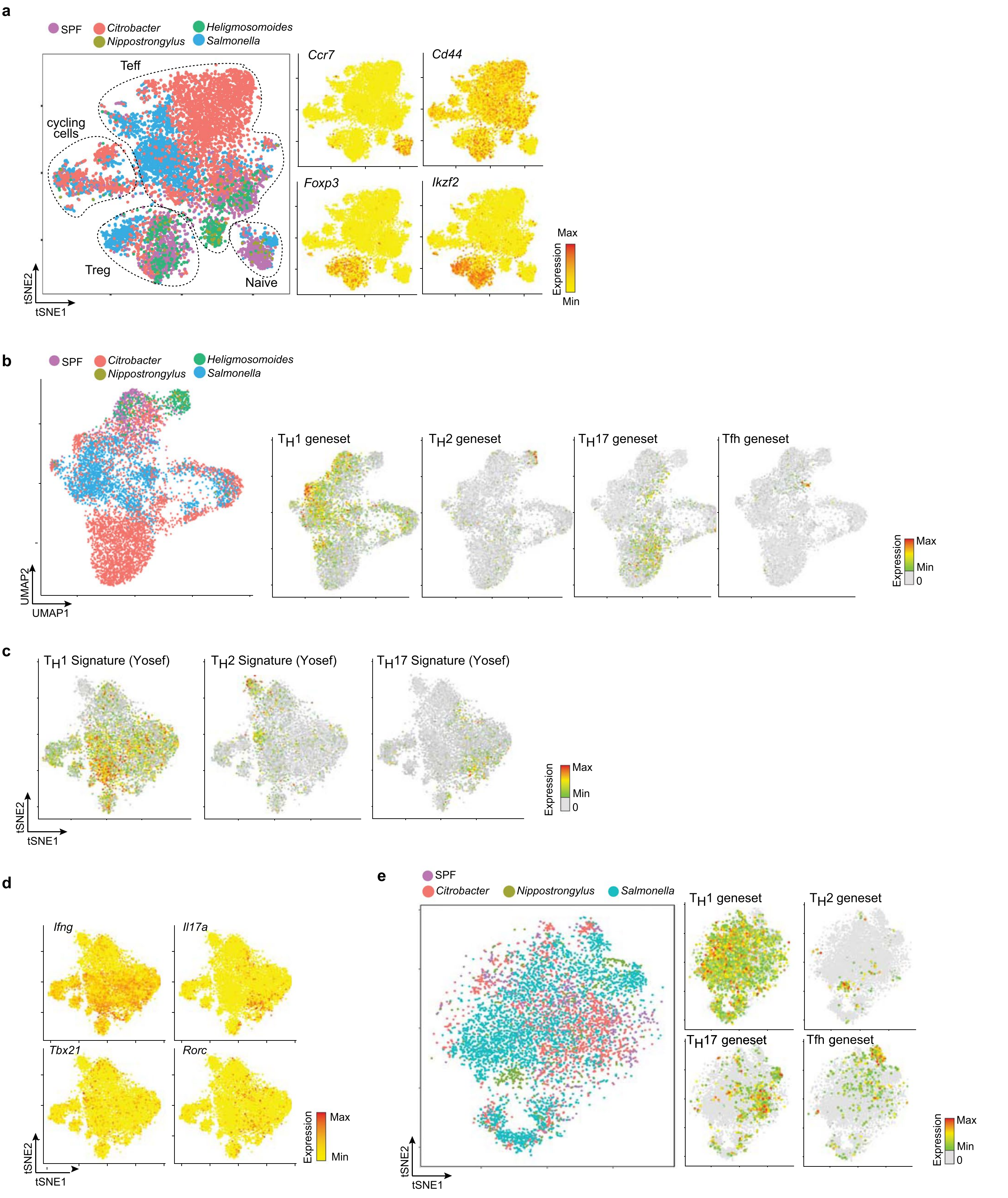
Correspondence and requests for materials should be addressed to D.M. or C.B.

Peer review information *Nature Immunology* thanks Thomas Korn, Evan Newell and Masahiro Ono for their contribution to the peer review of this work. Zoltan Fehervari was 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.

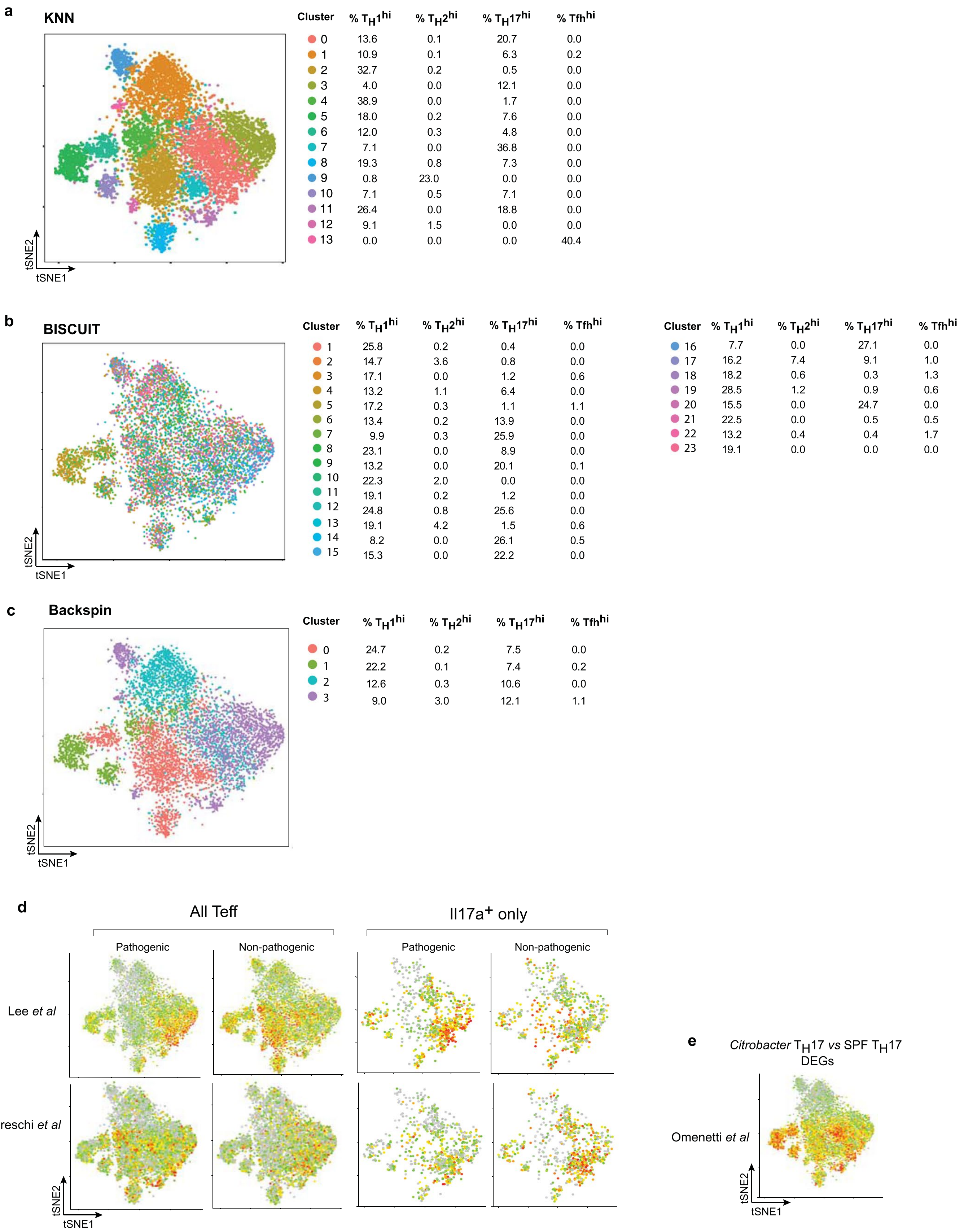


Extended Data Fig. 1 | scRNAseq of Teff under normal conditions. a, Quality control plots (per-cell number of unique reads vs number of transcripts detected) for the scRNAseq data from total colonic CD4$^+$ T cells (data from Fig. 1a). **b**, Same plots as (**a**), for CD4$^+$ QC of scRNAseq data from total colonic CD4$^+$ T cells of germ-free and SPF mice. **c**, SMART-SEQ2 single-cell data from colon T memory cells (from ref. [16]). Aggregate expression of Th-specific genesets (defined as for Fig. 1) are overlayed on the tSNE.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | scRNAseq of Teff under infectious conditions. a, tSNE representation of all CD4$^+$ T cells in the scRNAseq data from the parallel infection experiment of Fig. 2. Left panel: each color represents cells from a different infection condition. Tregs, naive Tconvs, cycling cells and Teffs are circled; right panel: expression of key genes. **b**, UMAP representation of Teff cells from the same experiment, colored by condition; Right panels: Overlay of T_H genesets (per Fig. 2). **c**, Data from the same parallel-infection experiment as Fig. 2c and displayed using the same tSNE coordinates, highlighted with aggregate expression of T_H signature genes from ref. [19]. **d**, Expression of key cytokines and transcription factors in the same scRNAseq data as Fig. 2c. **e**, Independent parallel infection experiment. Samples were not hash-tagged, and processed in parallel encapsulations, and cell data were aligned by canonical correlation analysis (CCA) for tSNE representation, color-coded by sample. Right: expression of Th-specific genesets, defined as for Fig. 2c.



a KNN

Cluster	% T_H1^{hi}	% T_H2^{hi}	% T_H17^{hi}	% Tfh^{hi}
0	13.6	0.1	20.7	0.0
1	10.9	0.1	6.3	0.2
2	32.7	0.2	0.5	0.0
3	4.0	0.0	12.1	0.0
4	38.9	0.0	1.7	0.0
5	18.0	0.2	7.6	0.0
6	12.0	0.3	4.8	0.0
7	7.1	0.0	36.8	0.0
8	19.3	0.8	7.3	0.0
9	0.8	23.0	0.0	0.0
10	7.1	0.5	7.1	0.0
11	26.4	0.0	18.8	0.0
12	9.1	1.5	0.0	0.0
13	0.0	0.0	0.0	40.4

b BISCUIT

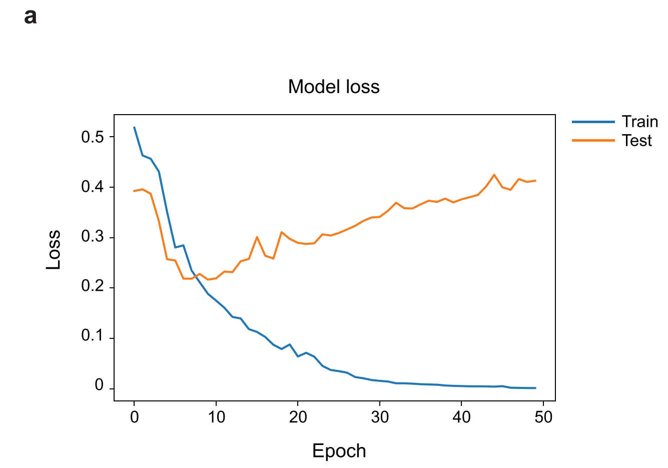
Cluster	% T_H1^{hi}	% T_H2^{hi}	% T_H17^{hi}	% Tfh^{hi}
1	25.8	0.2	0.4	0.0
2	14.7	3.6	0.8	0.0
3	17.1	0.0	1.2	0.6
4	13.2	1.1	6.4	0.0
5	17.2	0.3	1.1	1.1
6	13.4	0.2	13.9	0.0
7	9.9	0.3	25.9	0.0
8	23.1	0.0	8.9	0.0
9	13.2	0.0	20.1	0.1
10	22.3	2.0	0.0	0.0
11	19.1	0.2	1.2	0.0
12	24.8	0.8	25.6	0.0
13	19.1	4.2	1.5	0.6
14	8.2	0.0	26.1	0.5
15	15.3	0.0	22.2	0.0
16	7.7	0.0	27.1	0.0
17	16.2	7.4	9.1	1.0
18	18.2	0.6	0.3	1.3
19	28.5	1.2	0.9	0.6
20	15.5	0.0	24.7	0.0
21	22.5	0.0	0.5	0.5
22	13.2	0.4	0.4	1.7
23	19.1	0.0	0.0	0.0

c Backspin

Cluster	% T_H1^{hi}	% T_H2^{hi}	% T_H17^{hi}	% Tfh^{hi}
0	24.7	0.2	7.5	0.0
1	22.2	0.1	7.4	0.2
2	12.6	0.3	10.6	0.0
3	9.0	3.0	12.1	1.1

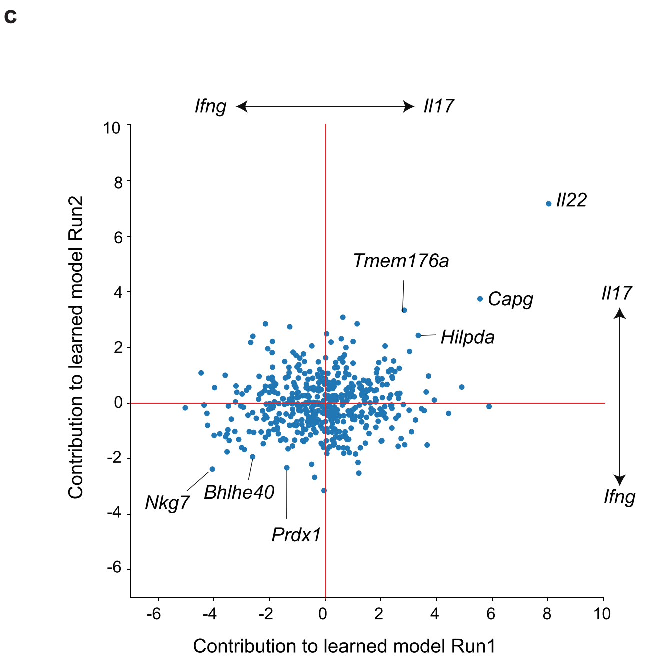
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Different clustering approaches and signatures do not parse out the data into T$_H$ subsets. a, KNN clusters shown on hash-tagged tSNE. Percentages of cells corresponding to each signature in each KNN cluster are shown in the table. **b**, Biscuit clusters shown on hash-tagged tSNE. Percentages of cells corresponding to each signature in each Biscuit cluster are shown in the table. **c**, Backspin clusters shown on hash-tagged tSNE. Percentages of cells corresponding to each signature in each Backspin cluster are shown in the table. **d**, Overlay of pathogenic T$_H$17 signatures from refs. [22,23]. Left panel: all Teff; right panel: only Il17a$^+$ Teff. **e**, Overlay of *Citrobacter* T$_H$17 signature from ref. [24] on the tSNE plot.

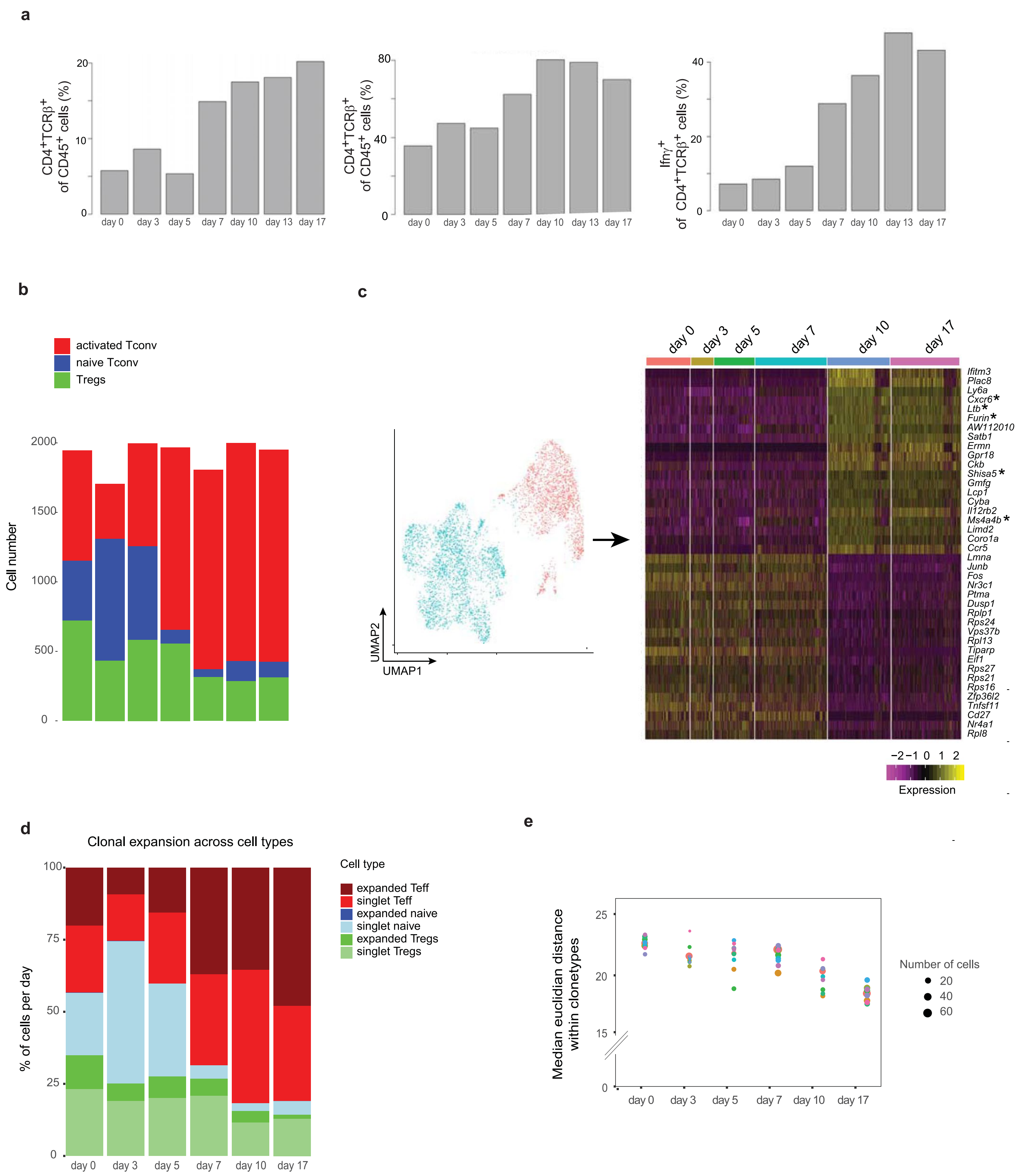


b

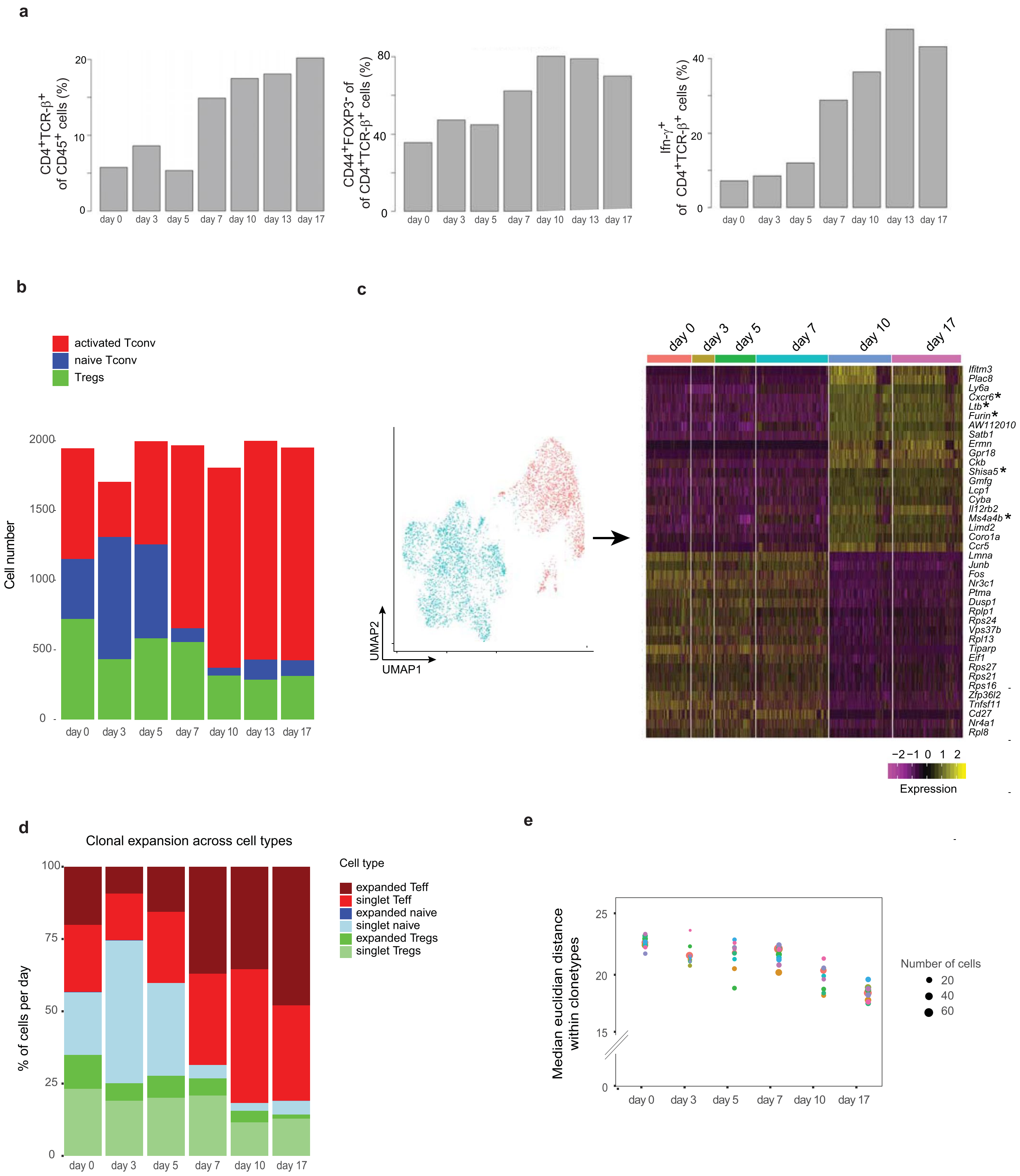
nCells=577 (test set)	Actual Ifng$^+$ (478)	Actual Il17$^+$ (99)
Predicted Ifng$^+$	409-464	13-25
Predicted Il17$^+$	10-35	45-51
No Call	25-56	5-14



Extended Data Fig. 4 | Neural Network prediction of IFN-γ and Il17-producing phenotypes. a, A Keras neural network was trained to use as input the expression of 500 most variable genes in Teff single-cell RNAseq data to predict *Ifng* or *Il17a* expression in each cell. Loss as a function of training epochs plotted here. Note the overfitting beyond 10 epochs (representative of >50 independent training runs with random 80/20 training/test)**. b**, Accuracy of DNN-predicted cytokine expression by individual Teff cells, relative to their actual expression in the test scRNAseq data (non-expressing cells were not included as input, since there is uncertainty as to their real nature given drop-out frequencies in scRNAseq data). Numbers shown represent the range observed in 10 independent training runs (with different training/test sets). **c**, Contribution of each transcript to the prediction of *Il17a* or *Ifng* expression, as score in the Integrated Gradients, comparing the model learned in two independent runs. A positive score indicates influence on predicting *Il17a* expression, a negative score influence in predicting *Ifng* expression.

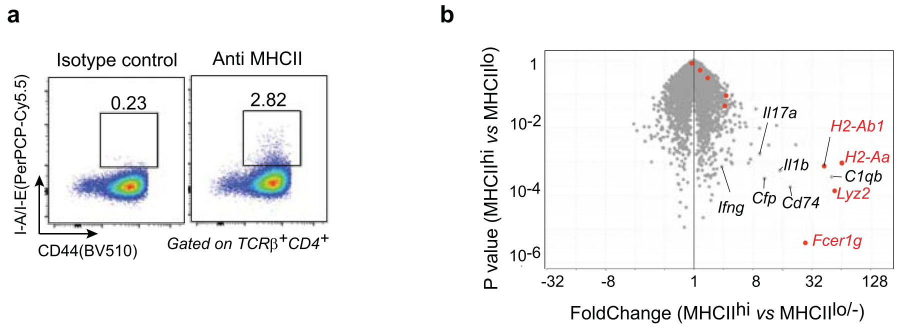


Extended Data Fig. 5 | Th-associated genes are not the main drivers of Teff heterogeneity. a, Distribution of Top 6 PCs of Teffs from all hash-tagged samples, with cell cycle genes regressed out. Genes that are Th-associated are highlighted. **b**, Co-expression of key cytokines across all samples. Mean Pearson gene:gene correlation of cytokine genes across all samples. Only significantly correlated cytokines are colored ($p < 0.05$, χ^2 test). Significant P values: *Il4/Il13* 6.3×10^{-3}, *Il4/Il5* 1.8×10^{-98}, *Il5/Il13* 5.5×10^{-129}, *Il17a/Il17f* 1.3×10^{-4}. **c**, Coregulated gene modules in Teff single-cells. Gene:gene correlation between 588 most variable genes was calculated independently within each condition/infection of the single-cell datasets, then averaged between conditions. 16 gene modules were determined by Affinity Propagation within this matrix, annotated at right. **d**, Overlay of average expression of these gene modules on Teff tSNE (per 2c) with barplots showing genes with highest mean correlation (full list in Supplementary Table 3).



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Unique clonotypes are not restricted to a T_H type and do not diversify over time. a, Quantification of flow cytometry data on cells from mouse LP at different timepoints of infection; Left: Proportion of CD4$^+$ T cells within total CD45$^+$; Middle: Proportion of Teff (CD44^{hi} Foxp3$^-$) within total CD4$^+$ T; Right: Proportion of IFN-γ$^+$ cells within total CD4 T. **b**, Cell numbers per scRNAseq clustering by day post infection. Treg clusters were identified as Foxp3$^+$, naive cluster as Foxp3$^-$ Ccr7$^+$ and Teff clusters as Foxp3$^-$ Cd44$^+$. **c**, Left: UMAP as in 5a, showing two groups of cell clusters: cells taken from mice after day 10 are colored in red, and cells taken prior to day 7 are colored in blue. Right: DEG analysis on top 20 differentially expressed genes between the two cluster groups. Asterisks represent genes that overlap with genes that are higher in Teff after *Salmonella* infection in Fig. 3a. **d**, Bar graph representing proportions of cells belonging to singlet clones (clones that appear only once) or expanded clones (clones that appear more than once) in each of the clusters defined in S6b, grouped by day post infection. **e**, Median Euclidean distances between cells within the same clonotype across the top 10 clonotypes for each timepoint. Euclidean distance was calculated based on the top 1000 variable genes. Each color dot represents a unique clonotype, and the size of the dot signifies the number of cells within each clonotype.



Extended Data Fig. 7 | The unexpected MyT subset. **a**, Flow cytometric analysis (gated $CD4^+TCR\beta^+FOXP3^-$ Teff) cells from colonic LP of Salmonella infected mice. **b**, Volcano plot of bulk RNAseq from colonic Teff sorted as in C (LP of *Salmonella* infected mice). Genes highlighted in red belong to the myeloid genes listed in B.

nature research

Corresponding author(s): Christophe Benoist

Last updated by author(s): Oct 22, 2020

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 our Editorial Policies 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	Cellranger software (V1.1.0 for 10x 3' V1, V2.1.0 for 10x 3' V2 and V1.3.0 for 10x 5')
Data analysis	All software and analysis steps are detailed in Methods. Initial processing of the scRNAseq data with Seurat.v2, later clustering with BackSPIN v1.0 , BISCUIT v1.0 ; Deep Neural Network training with Keras v2.2.4.

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 and 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

All data (raw fastq as well as gene tables) have been deposited at NCBI/GEO, and are served on the ImmGen single-cell browser. ATACseq and bulk RNAseq data also at GEO.
No restrictions on data availability.

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

Life sciences study design

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

Sample size	N/A
Data exclusions	Uninformative cells with low UMI/gene counts were excluded from all single-cell datasets, per usual practice, as described in Methods.
Replication	All single-cell RNAseq profiling experiments were performed at least in duplicate.
Randomization	N/A
Blinding	N/A

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
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All mAbs were established standards, obtained were from commercial suppliers, clones and dilutions listed in Methods.
Validation	Validation was based on supplier's catalog, as well as matching expected frequencies in the immunocyte populations analyzed.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	7-9 week old C57BL/6 males were used for all single cell experiments unless otherwise specified.
Wild animals	No wild animals were used in the study.
Field-collected samples	No field-collected samples were used in the study.
Ethics oversight	Harvard Medical School IACUC protocols IS1257, IS187-3, IS2221

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	Cells from colon LP were prepared as previously described {10155}. Briefly, Intestinal tissues were treated with RPMI containing 1 mM DTT, 20 mM EDTA and 2% FBS at 37ºC for 15 min to remove epithelial cells, minced and dissociated in collagenase solution (1.5mg/ml collagenase II (Gibco), 0.5mg/ml Dispase (Gibco) and 1%FBS in RPMI) with constant stirring at 37ºC for 45min. Single cell suspensions were then filtered and washed with 4% RPMI solution.
Instrument	FACS Aria, Symphony, MoFlo Astrios
Software	FACS Diva, FlowJo
Cell population abundance	Variable, mostly relatively abundant (15-40%) CD4+ T cells
Gating strategy	As indicated in Figures where relevant

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