Temporal mixture modelling of single-cell RNA-seq data resolves a CD4$^+$ T cell fate bifurcation

Authors:

Tapio Lönnberg[1,2,#], Valentine Svensson[1,#], Kylie R. James[3,#], Daniel Fernandez-Ruiz[4], Ismail Sebina[3], Ruddy Montandon[2], Megan S. F. Soon[3], Lily G. Fogg[3], Michael J. T. Stubbington[1,2], Frederik Otzen Bagger[1,2,5,6], Max Zwiessele[7], Neil Lawrence[7], Fernando Souza-Fonseca-Guimaraes[3], William R. Heath[4,8], Oliver Billker[2], Oliver Stegle[1,*,†] Ashraful Haque[3,*,†], Sarah A. Teichmann[1,2,*,†]

Affiliations:

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK.

[2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

[3]QIMR Berghofer Medical Research Institute, Herston, Brisbane, Queensland, Australia.

[4]Department of Microbiology and Immunology, The Peter Doherty Institute, University of Melbourne, Parkville, Victoria, Australia.

[5]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

[6]National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge, UK

[7]Department of Computer Science, University of Sheffield, Sheffield, UK

[8]The Australian Research Council Centre of Excellence in Advanced Molecular Imaging, The University of Melbourne, Parkville, Victoria, Australia.

*Correspondence to: st9@sanger.ac.uk, Ashraful.Haque@qimrberghofer.edu.au or stegle@ebi.ac.uk

# denotes equal contribution

† denotes equal contribution

Abstract

Differentiation of naïve $CD4^+$ T cells into functionally distinct T helper subsets is crucial for the orchestration of immune responses. Due to multiple levels of heterogeneity and multiple overlapping transcriptional programs in differentiating T cell populations, this process has remained a challenge for systematic dissection *in vivo*. By using single-cell RNA transcriptomics and computational modelling of temporal mixtures, we reconstructed the developmental trajectories of Th1 and Tfh cell populations during *Plasmodium* infection in mice at single-cell resolution. These cell fates emerged from a common, highly proliferative and metabolically active precursor. Moreover, by tracking clonality from T cell receptor sequences, we infer that ancestors derived from the same naïve $CD4^+$ T cell can concurrently populate both Th1 and Tfh subsets. We further found that precursor T cells were coached towards a Th1 but not a Tfh fate by monocytes/macrophages. The integrated genomic and computational approach we describe is applicable for analysis of any cellular system characterized by differentiation towards multiple fates.

One Sentence Summary

Using single-cell RNA sequencing and a novel unsupervised computational approach, we resolve the developmental trajectories of two $CD4^+$ T cell fates *in vivo*, and show that uncommitted T cells are externally influenced towards one fate by inflammatory monocytes.

Introduction

T helper (Th) cells, also known as $CD4^+$ T cells, are key instructors of the immune system (*1*). They display extensive functional and phenotypic diversity in response to a spectrum of immune challenges, including viral, bacterial, fungal and parasitic infection, immunogenic cancers, and autoimmune and allergic stimuli. Th cell subsets are distinguished from each other most frequently by the cytokines they secrete. Th1 cells produce interferon-γ, leading to macrophage activation and enhanced killing of intracellular pathogens. Th2 cells produce IL-4, IL-5, and IL-13, prompting eosinophils to act against extracellular parasites and venom. Th17 cells produce IL-17 and IL-22, promoting neutrophilic responses against extracellular bacteria and fungi. Follicular T helper (Tfh) cells, a more recently defined Th subset, secrete IL-21, and drive somatic hypermutation of immunoglobulin genes in germinal centre B cells. This produces high affinity antibodies, upon which many licensed vaccines depend for efficacy. Since Th subsets can both control infections and drive immune-mediated diseases there remains tremendous interest in the molecular mechanisms that control their *in vivo* development.

In order for Th cells to develop, $CD4^+$ T cells must first be raised from an immunologically naïve state by antigenic stimulation of their highly diverse T cell receptors (TCR), which is followed by processes of clonal proliferation and differentiation. Recent *in vivo* data suggested that the unique TCR sequence of a single naïve $CD4^+$ T cell imparts a genetically programmed

preference towards a particular Th fate (*2*). However, co-stimulatory and cytokine signals can also profoundly influence both the magnitude of the response, and skewing towards particular Th fates. Several master transcription factors have been described in $CD4^+$ T cells that drive and stabilize Th fates, which supports a view of Th development as a choice between clearly distinct states. However, the relationship between Th subsets, particularly between Tfh and other Th fates remains unclear *in vivo*.

In many cases, immune challenges, such as infection or vaccination, induce concurrent differentiation into two or more Th fates within the same individual. Indeed, by performing a limiting dilution single-cell adoptive transfer of naïve $CD4^+$ T cells, it was suggested that daughter cells from a particular clone could bifurcate phenotypically to give rise to both Th1 and Tfh cells (*2*). However, it was not possible to visualize and pinpoint the bifurcation of Th1/Tfh cell fates *in vivo*.

Resolving Th cell fate decision-making *in vivo* using population-level approaches has been challenging, mainly due to extensive heterogeneity amongst differentiating cells. More specifically, $CD4^+$ T cells at any given time point display a distribution of intermediate and transitional states, which blurs the dynamics of Th cell developmental progression (*3*). Tfh differentiation, in particular, has been difficult to elucidate since it involves multiple stages with potential overlap with transcriptional programs of other Th subsets. Of particular note, computational tools for modelling bifurcations in cellular decision-making have not been available.

Th cell fate decisions are driven by both intrinsic factors and external signalling cues from other cells. Conventional dendritic cells (cDCs) are important cellular sources of antigenic stimulation, co-stimulation and cytokines for Th differentiation in secondary lymphoid tissues. Intra-vital imaging in lymph nodes has demonstrated that cDCs make long-lasting stable contacts with naïve $CD4^+$ T cells in order to initiate T cell priming (*4*). Once activated, $CD4^+$ T cells continue to require antigenic stimulation via their TCR to optimize their proliferation and Th differentiation (*5-7*). Continued signalling has been reported to be important for Th1 responses, although the cell types providing this signal remain unknown (*4*). A recent report suggested that CXCR3 expression by activated $CD4^+$ T cells facilitated continued interaction with adoptively-transferred CXCL9 and CXCL10-expressing cDCs (*8*), however, interactions with endogenous myeloid cell populations, including cDC subsets and monocytes have not been studied *in vivo*. While Tfh cells are sustained, once generated, via multiple molecular interactions with B cells in developing germinal centres (*9, 10*), possible roles for myeloid cells in providing early instruction towards a Tfh fate remain relatively unexplored. A recent study targeted antigens to two different cDC-subsets *in vivo,* and suggested that $CD8\alpha^-$ cDCs displayed the greater propensity for generating Tfh responses (*11*). Whether Th1/Tfh fate bifurcation can result from differential interactions with cDC subsets or activated monocytes currently remains unknown.

Herein we have used single-cell RNA sequencing (scRNA-seq) to study the various transcriptional states of individual $CD4^+$ T cells during blood-stage *Plasmodium chabaudi* infection in mice. This is an experimental model of malaria in which $CD4^+$ T cells are essential

for controlling parasite numbers, and which is characterized by concurrent development of Th1 and Tfh cells (*12*). We have used *Plasmodium*-specific TCR transgenic $CD4^+$ T (PbTII) cells to minimise the effects of TCR diversity on Th fate decisions.

Crucially, our approach builds on scRNA-seq profiling coupled with new computational strategies to reconstruct the differentiation trajectories of Th1 and Tfh cells at a single-cell resolution. Our data reveals, for the first time, the molecular detail of how a single antigen-specific $CD4^+$ T cell clone can undergo parallel development into Th1 and Tfh states *in vivo*, and reveals the hierarchical regulation of genes involved in this cell fate decision. Finally, we investigated intercellular interactions using scRNA-seq, and predicted roles for inflammatory monocytes, after cDC-dependent T cell activation, in coaching uncommitted $CD4^+$ T cells, specifically towards a Th1 fate.

Results

scRNA-seq resolves Th1 and Tfh cell fates during *Plasmodium* infection in mice

To study concurrent progression towards Th1 and Tfh fates, and to characterize the heterogeneity associated with this process during an *in vivo* $CD4^+$ T cell response, we performed scRNA-seq of PbTII cells during *Pc*AS infection (Figure 1A, Figure S1). We transferred naïve, proliferative dye-labeled PbTII cells into congenically marked wild-type mice, and recovered them at days 2, 3, 4, and 7 post-infection (p.i.) by fluorescence-activated cell sorting (FACS) of those expressing the early activation marker, CD69, or displaying dilution of the proliferative dye (Figure S2). Flow cytometric measurements of the canonical Th1 markers, T-bet (coded by *Tbx21*) and Interferon-γ, and Tfh markers, CXCR5 and Bcl6, indicated that these subsets emerged in parallel by day 7 p.i. (*13, 14*) (Figure 1B-D). Notably, markers of Th2, Th17 or Treg subsets were not upregulated on the PbTII cells (Figure S3).

We initially used Principal component analysis (PCA) to assess the overall heterogeneity of the PbTII cells (Figure 1E, Figure S4A). In all time points, the first principal component was strongly associated with the number of detected transcripts, which is reflective of changes in cellular RNA content and, in general, is linked to proliferative status (Figure S4B). As expected, the variability related to previously established Th1 and Tfh gene expression signatures became more prominent with the progression of time (*15*) (Figure S4C). Notably, at day 7 p.i., a PCA using these signature genes alone recapitulated the results of the genome-wide PCA (Spearman correlation -0.87) (Figure S5). Amongst the cells from day 7 p.i., two distinct subpopulations were apparent, separated along PC2 (Figure 1E). Notably, many of the genes associated with these subpopulations have been identified as associated with either Th1 or Tfh fates (Figure 1F, Table S1). Results from a global PCA of the entire dataset were largely in accordance with the time point information, with the Th1/Tfh signature genes showing separation along multiple PCs (Figure S6). Taken together, these results suggested a progressive commitment to Th1 and Tfh fates, and indicated that single-cell transcriptomes could be used for estimating both proliferative states and degrees of differentiation of individual cells.

Unbiased delineation of Th1 and Tfh trajectories using a Mixture of Gaussian Processes model

The results from the PCA suggest that gene expression variation in PbTII single-cell transcriptomes permit reconstruction of the transcriptional programs underlying Th1 and Tfh differentiation. To more explicitly model the temporal dynamics of the differentiation process, we developed and applied GPfates, a temporal mixture model that builds on the Gaussian Process Latent Variable Model (GPLVM) and Overlapping Mixtures of Gaussian Processes (OMGP) (*16*). This approach first reconstructs the differentiation trajectory from the observed data ("pseudotime", Figure 2A-B), thereby establishing an order for the cells. While our model uses the sample time as prior information, the inferred temporal orderings did not strictly adhere to these experimental time points (Figure S7). For example, cells from day 4 p.i. were mixed with some of the cells from day 3 and day 7 at either end of the day 4 pseudotime distribution. This was consistent with the idea that bulk assessments of cells at specific time points fail to take into account the heterogeneity and differential kinetics of responses made by single cells. We also repeated this analysis without supplying the experimental sampling times to the model, finding overall consistent results (Comp. Supp. Figure 8).

In a second step, GPfates uses a time series mixture model, which we adapted from a model that was initially developed to deconvolve temporal data into independent separate trends, and which is related to previous time series models for bulk gene expression time series (*16*). Using this approach, we identified two simultaneous trends (Figure 2C-D). These two alternative trajectories were in agreement with the Th1/Tfh signature genes identified by Hale et al. (*15*) (Figure 3A-D), indicating that the fitted mixture components correspond to cells with Th1 and Tfh phenotypes. Notably, these trends could not be identified by other published methods for reconstructing single-cell trajectories (*17, 18*) (Figure S8). Furthermore, the mixture modelling in GPfates could also successfully resolve bifurcation events in two other recently published scRNAseq datasets, which examined lung epithelial development in mice (Comp. Supp. Figure 11) (*19*) and primordial germ cell development in human embryos (Comp. Supp. Figure 12) (*20*). This suggests that pseudotime inference coupled with time series mixture modelling is applicable more generally for studying cellular differentiation in scRNAseq data.

Next, we sought to more clearly characterize the bifurcation time point. Using a change point model to annotate the inferred trajectories (see section 4.2 of the Computational Supplement), we could divide pseudotime into *before* and *after* bifurcation. We sought to characterize single cells that existed at the Th1/Tfh bifurcation point. Firstly, bifurcation initiated amongst cells from day 4 p.i. (see section 6.2 of computational supplement for a robustness analysis), specifically at a relatively early point in pseudotime compared with all day 4 p.i. cells (Figure 4A). Bifurcating PbTII cells also expressed the largest number of genes compared to those at all other points in pseudotime.

High transcriptional activity correlated with upregulation of *Mki67* and other known proliferation marker genes (*21*) (confirmed at the level of Ki-67, Figure 4B-C and S9A). It also correlated

with cell cycle activity, based on computational allocation of cells into cell cycle stages (*22*), and flow cytometric confirmation of DNA content and cell size (Figure 4D-E). Bifurcating PbTII cells also had increased expression of genes associated with aerobic glycolysis (data not shown), an indication of increased metabolic requirements being met by glucose metabolism and increased mTORC1 activity. Consistent with this was the observed elevated levels of ribosomal protein S6 phosphorylation by day 4 p.i. (Figure 4F).

Taken together, our data indicate that bifurcating PbTII cells exhibit a highly proliferative and metabolically active state, coupled with the upregulation of thousands of genes. Importantly, progression from the Th1/Tfh bifurcation point to either fate was marked by widespread silencing of gene expression across the genome. Although this decrease in gene expression can be partially explained by a deceleration in cell cycle speed, it is also consistent with other cellular differentiation processes characterized at a single-cell resolution (*19*).

Detectable expression of endogenous T cell receptor loci reveals breadth of clonotype fates

Since previous reports have suggested a role for TCR sequences in determining Th cell fate (*2*), our TCR transgenic approach was designed to minimize this potential source of variability. Importantly however, PbTII cells were generated in mice with functional *Rag1* and *Rag2* genes, and therefore, retained natural expression of highly diverse endogenous TCR chains in addition to the transgenic TCR. Sequence analysis of TCR transcripts in single PbTII cells confirmed universal expression of the PbTII Vα2 and Vβ12 chains in all cells (Supplementary Tables 2 & 3). Moreover, it confirmed highly diverse, though lower levels of expression of endogenous TCRα chains in many cells (Figure S10).

Given the vast combinatorial diversity of endogenous TCR sequences, we employed these as unique molecular barcodes to scan for PbTII cells that could be inferred with high confidence to have derived from a single common PbTII progenitor clone. Notably, we identified six clones comprising two or more sibling cells, while all other PbTII cells were individually unique. Of these six clones, two consisted of sibling cells that mapped close to the bifurcation point. For the remaining four clones, siblings exhibited highly diverging patterns of gene expression, with three sibling groups falling at the extremities of the Th1-Tfh phenotype spectrum (Figure 5A). These results demonstrate that during an *in vivo* infection, the progeny of a single $CD4^+$ T cell clone can differentiate into both Th1 and Tfh cells.

Transcriptional signatures associated with bifurcation of Th1 and Tfh fates

Next, we sought to identify genes whose expression followed the pattern of branching. We derived *bifurcation statistic* to estimate the concordance with the bifurcation for individual genes (see section 4.2 of the Computational Supplement text for details, Figure 5B). Among the highest-ranking bifurcating genes, the most common pattern was an increase in expression during progression to the Th1 fate. These genes were positively correlated with both pseudotime

and the Th1 trend assignment (Figure 5B). This suggests that Tfh cells are in fact developmentally closer to the highly proliferative progenitor state than Th1 cells as the Th1 fate involves up-regulation of numerous genes not expressed in either the progenitor or Tfh states.

The highest-ranking transcription factors were *Tcf7* for the Tfh fate, and *Mxd1, Bhlhe40, Hopx, Pgs1* and *Id2* for the Th1 fate (Figure 5C). In addition, the hallmark Tfh transcription factor *Bcl6* was strongly associated with the Tfh fate. *Tcf7* is required for T cell development, and has been recently shown to be instrumental for Tfh differentiation (*23, 24*). Notably, it represented one of the rare genes defined by a decrease in expression when moving towards the Th1 fate. Of the Th1-associated transcription factors, *Mxd1* is a negative regulator of the proliferation-associated, proto-oncogene, *Myc (25)* and *Bhlhe40* has been recently identified as a cofactor of *T-bet* (coded by *Tbx21*) *(26)*. *Id2* is known as an antagonist of *Tcf7 (27)* and as a regulator of effector $CD8^+$ T cell responses. Notably, while this manuscript was under revision, the role of Id2 as a key driver of Th1 responses was independently shown by another study (*28*).

Our results strongly support reciprocal regulation of *Id2* and *Tcf7* as a key feature of the Th1/Tfh bifurcation process. Expression of *Id2* and *Ifng* were highly correlated in the later stages of Th1 differentiation, and negatively correlated with *Tcf7,* both at a transcriptional and protein level (Figure 5D-F, Figure S11). Notably, the hallmark Th1 transcription factor *Tbx21* was induced before the bifurcation point, and showed only modest separation after bifurcation (Figure S12).

To validate the robustness of these gene signatures and the timing of the bifurcation, we repeated the infection, and at days 0, 4 and 7 sequenced additional single PbTII-cells using the Smart-seq2 protocol (*29*) (Figure 1A & S13A). Consistent with the original data, the cells from day 7 (but not day 4) segregated into two subpopulations correlating with Th1 and Tfh gene signatures (Figure S13A). Subset-characteristic co-expression patterns of the bifurcating genes identified by GPfates emerged by day 7 (Figure S13B). Notably, at this time, the cells from the different mice could be equally separated into distinct Th1- and Tfh- subpopulations using the top bifurcating genes (Figure S13C). Taken together, this indicated that the gene expression patterns associated with the cell fate bifurcation were reproducible across experiments and sequencing platforms.

In Th1 cells, a large fraction of the bifurcating genes were cytokine and chemokine receptors, including the top-ranked gene, *Cxcr6,* confirmed at protein level (Fig S14A and S14B), other established Th1 markers, *Ifngr1* and *Il18rap* (*30, 31*), and the chemokine receptors *Ccr2* and *Ccr5* (Figure 5C). These data were consistent with the idea that Th1 cells can migrate to peripheral tissues and remain receptive to external signals. In contrast, the only bifurcating chemokine receptor associated with a Tfh fate was *Cxcr5*, a gene established to mediate migration of Tfh cells into B cell follicles (*32, 33*).

Cxcr5 was among an early wave of chemokine receptor genes, including *Cxcr3* and *Ccr4* (Figure 5G) whose expression and translation into protein (Figure S14C) was initiated before the Th1/Tfh bifurcation point had been reached. We hypothesized that differences in the timing of expression of receptors reflected their roles in controlling differentiation or effector function. We reasoned, for instance, that while *Cxcr6*, *Ccr2* and *Ccr5* served to mediate trafficking and

effector function of Th1 cells, others such as *Cxcr3* and *Cxcr5* controlled Th cell fate via interactions with other immune cells (Figure 5H). Indeed, *Cxcr5* allows T cell trafficking towards B cells (*34, 35*), while *Cxcr3* has been associated with cDC-driven Th1 fates (*8*).

Myeloid cells support a Th1 but not Tfh fate

After activation and proliferation, PbTII cells reached an uncommitted state around the bifurcation and expressed chemokine receptors that indicated receptiveness to other chemokine-expressing cells. Given that B cells were essential, as expected, for supporting a Tfh fate in PbTII cells (Fig S15), we hypothesized that myeloid cells provided alternative, competing signals to promote a Th1 fate.

To study this, we performed scRNA-seq on splenic cDCs and inflammatory monocytes when activated PbTII cells were yet to bifurcate. We sorted $CD8\alpha^{+}$ and $CD11b^{+}$ cDCs and $Ly6C^{hi}$ monocytes from naïve and infected mice (Figure S16) and subjected these to single-cell analysis. PCA of cDCs firstly distinguished between the two naïve cell types, separating them along PC2 (Figure 6A & S17) with an efficiency consistent with recent data (*36*), and further highlighting a number of expected and previously unknown cDC subset-specific genes (Figure S18A-C). We next compared naïve cDCs with those from infection (Figure 6A & S16), and separated these along PC6 (Figure 6A). Analysis of differential gene expression between cDCs from naive and infected mice identified 30 genes, 29 upregulated (Figure 6B & S19), including interferon-associated transcription factors, *Stat1* and *Irf1*, and CXCR3-attractant chemokine genes, *Cxcl9* and *Cxcl10*. Notably, gene expression patterns amongst individual cDCs varied according to the gene. For example, *Stat1* and *Irf1* were heterogeneously expressed amongst individual naïve cDCs, and further upregulated during infection (Figure 6C). This was similar for *Cxcl9,* which was expressed by $CD8\alpha^{+}$ cDCs in naive mice, while *Cxcl10* was induced only upon infection (Figure 6C). These data revealed interferon-associated gene expression amongst individual cDCs, and also suggested interactions between cDCs and uncommitted $CXCR3^{+}$ PbTII cells, consistent with a recent study (*8*). Next, PCA of $Ly6C^{hi}$ monocytes from naïve and infected mice distinguished them from each other along PC2 (Figure 6D & S20). Differential gene expression analysis between naïve and infected groups uncovered ~100 genes, both up- and down-regulated during infection (Figure 6E & S21). This illustrated a fundamental difference in the directionality of transcriptional changes in individual monocytes compared to cDCs during *Plasmodium* infection, with only monocytes exhibiting down-regulation of gene expression (Figure 6B-C & E-F). Interestingly, a high proportion (~40%) of genes upregulated in cDCs were also induced in $Ly6C^{hi}$ monocytes, including transcription factors *Stat1* and *Irf1*, and the chemokine *Cxcl10* (Figure 6E & F), suggesting possible overlapping biological functions between these cell types. In addition, monocyte-specific chemokines were also observed, including *Cxcl2, Ccl2* and *Ccl3* (Figure 6E & F). Furthermore, specific examination of all immune cellular interaction genes (Figure S22) revealed emerging variable expression of *Tnf, Cd40, Pdl1*, *Ccl4, Ccl5, Cxcl16, Cxcl9, and Cxcl11* in monocytes, thus suggesting complex interactions and multiple roles for $Ly6C^{hi}$ monocytes during infection .

Given that *Cxcl9-11, Ccl2, Ccl3* and *Ccl5* signal through either *Cxcr3* or *Ccr4*, which were expressed by activated but uncommitted PbTII cells, we next hypothesized that Ly6C^{hi} monocytes, in addition to cDCs, might interact with PbTII cells, thereby influencing Th1/Tfh fate (*8*). To begin testing this, we first confirmed chemokine expression at protein level by Ly6C^{hi} monocytes, focussing on CXCL9 (Figure 6G). Kinetics of CXCL9 production was similar in cDCs and Ly6C^{hi} monocytes, consistent with a possible role in interacting with CXCR3^{+} PbTII cells. To test whether monocytes could influence Th1/Tfh bifurcation *in vivo*, we employed *LysMCre* x *iDTR* mice, in which Ly6C^{hi} monocytes could be depleted after PbTII cell activation, but before bifurcation (Figure 6H, Figure S22A). We also noted a modest reduction in CD68^{+} macrophages using this approach, with no evidence for depletion of cDCs or marginal zone macrophages (Figure S23). In this transgenic approach, Th1 fates, but not Tfh fates, were supported by monocytes/macrophages (Figure 6H). Together, these data supported a model in which progression of activated, uncommitted PbTII cells towards a Tfh fate was dependent upon B cells (Figure S14), and a Th1 fate was promoted by chemokine-expressing myeloid cells, including Ly6C^{hi} inflammatory monocytes.

Discussion

By capturing single CD4^{+} T cell transcriptomes over time, and using a novel analysis approach to reconstruct the continuous course of events, we have resolved the bifurcation of naive CD4^{+} T cells into Th1 and Tfh cells at an unprecedented level of molecular detail, and illustrated that external cellular signals influence Th fate around the point of bifurcation. Importantly, the GPfates modelling of scRNA-seq data is not limited to immune cells or single bifurcation events. The mixture of time series model we used can also be combined with existing computational workflows (*17, 37*) (see section 5.2 of the Computational Supplement). Therefore, it provides the means for high-resolution analysis of differentiation in any cellular system, mainly towards two fates, as shown by our examination of existing embryonic development and lung tissue regeneration data (Comp. Supp. Figure 11), and, in principle, also for differentiation into multiple cell types (Comp. Supp. Figure 12), for example, during haematopoiesis. The filtered expression data and gaussian process models presented in this study can be found on our interactive web application at data.teichlab.org, where users can visualise their own genes of interest.

Our data reveals the developmental relationship between Th1 and Tfh cells on a genomic scale, and shows that the same naïve precursor can give rise to both fates simultaneously. It provides insights for the early stages of differentiation, and describes the order of transcriptional events before and after the bifurcation of Th1 and Tfh fates. To date, this process has remained incompletely characterised. Here, we use pseudo-temporal ordering of cells to reveal the hierarchy of transcriptional regulation of these events at an unprecedented resolution. Our data highlight the importance of stochastic expression of transcription factors as well as chemokine receptors, suggesting a role for noisy gene expression in Th development.

Transcriptomic profiling previously suggested developmental similarities between Tfh and Th1 cells (*38*), with *in vitro* studies suggesting relatively late bifurcation of Tfh and Th1 cells (*39*). However, highly immunogenic viral or bacterial infections induced $CD4^+$ T cells to segregate into $Bcl6^+$ (Tfh) or $Blimp\text{-}1^+$ (Th1) subpopulations within two days, and by three days, fate-committed Tfh cells had developed (*40-42*). In our parasitic model, single $CD4^+$ T cell transcriptomes remained remarkably similar until four days of infection. Although it is difficult to directly compare viral or bacterial systems with our parasitic model, we speculate that due to infection-related differences in antigen-presenting cell function, antigen load and availability, *Plasmodium* infection in mice does not drive Th bifurcation as early as observed with highly immunogenic viruses or bacteria. Evidence of sub-optimal MHCII antigen-presenting cell function early during *Plasmodium* infection (*43, 44*) raises the hypothesis that Th bifurcation is sensitive to immune-suppression. Our data indicate that uncommitted, activated $CD4^+$ T cells are heterogeneous, but nevertheless closely related at a transcriptional level, suggesting considerable flexibility throughout the proliferative phase of their response. Such plasticity during Th differentiation has been proposed to be beneficial as a means of countering evolution of immune-evasion strategies by pathogens (*3*).

As $CD4^+$ T cells progress from immunological naivety towards a Th fate, they may experience different cellular microenvironments, even within the confines of secondary lymphoid tissue. The observation that bifurcation towards Th1 and Tfh cells was preceded by upregulation of chemokine receptors prompted us to investigate possible interactions with chemokine-expressing myeloid cells. Previous studies have highlighted the potential for cDCs in lymph nodes to produce Th1-associated chemokines (*8*). Our study, which focused on the spleen, was consistent with this concept, and, furthermore, implicated inflammatory monocytes in Th1 support. However, since our transgenic approach for depleting monocytes also removed a portion of splenic red pulp macrophages, we cannot discount the possibility that red pulp macrophages may partly contribute to a Th1 fate. Nevertheless, our data support a model in which myeloid cells in the spleen influence bifurcation, and support a Th1 fate during *Plasmodium* infection. Moreover, our studies emphasise that although cDCs are the predominant professional antigen-presenting cell for initiating $CD4^+$ T cell activation in the spleen, other myeloid cells also exhibit a capacity to influence towards a Th1 fate. In contrast, Th bifurcation towards a Tfh fate was not supported by monocytes/macrophages. Instead, given that CXCR5 was the only chemokine receptor significantly associated with bifurcation towards a Tfh fate, cellular interaction with B cell follicles may be the primary mechanism for supporting activated $CD4^+$ T cells towards a Tfh fate. Our model suggests that activated, uncommitted $CD4^+$ T cells become receptive to competing chemoattractant signals from multiple cell types in different zones of the spleen. This model focuses on intercellular communication as the main driver of bifurcation. However, upstream of these processes, internal stochasticity in uncommitted $CD4^+$ T cells may control the balance of chemokine receptor expression (*45*), thus mediating differential trafficking and variation in intercellular interactions. Future experiments combining our integrated single-cell genomics and computational approach with *in vivo* positional and trafficking data may reveal molecular relationships between internal stochasticity, migratory behaviour, Th fate and perhaps immunological memory.

Figure 1

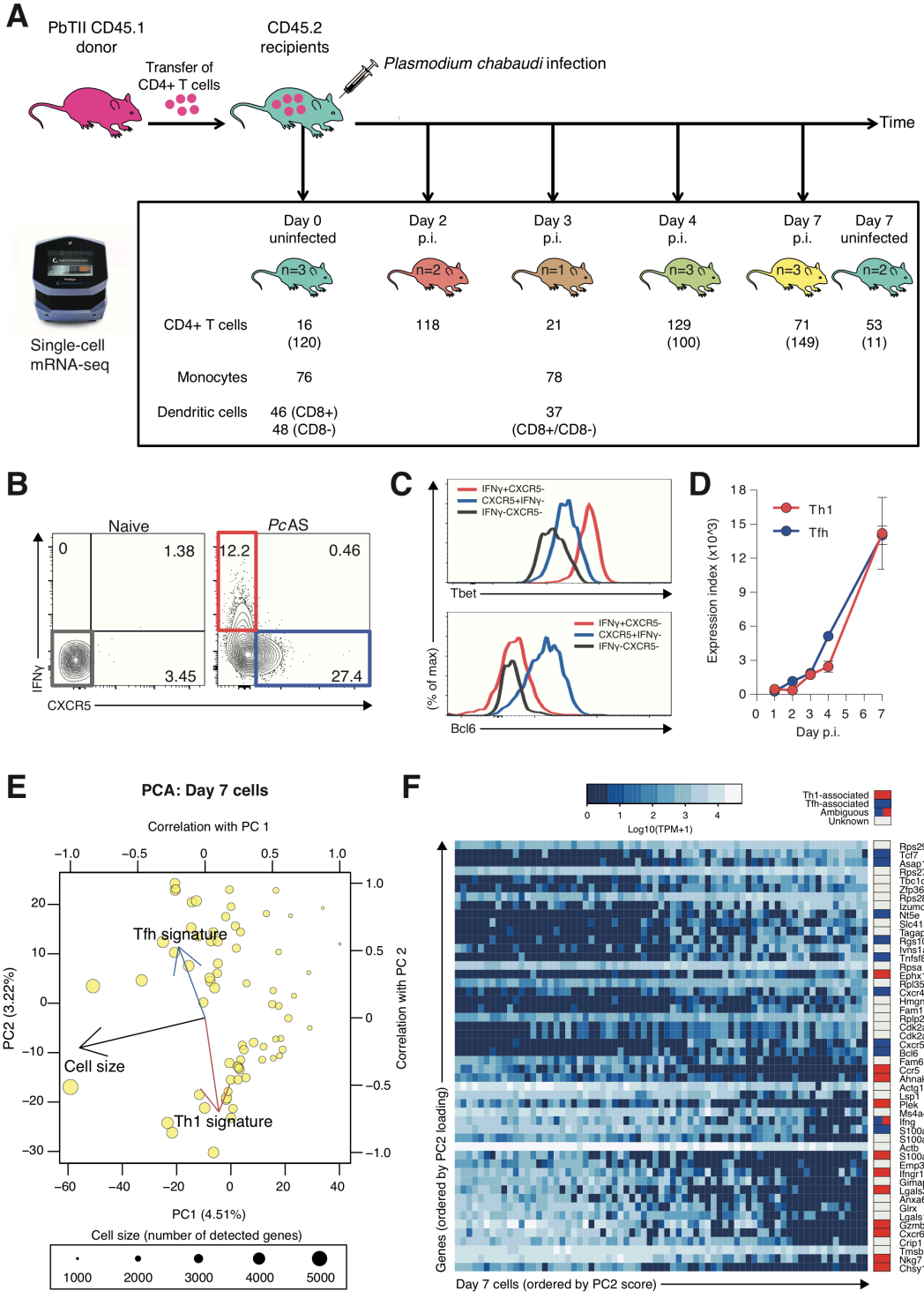


Figure 1. Single-cell mRNA-sequencing of activated antigen-specific PbTII CD4$^+$ T cells.

(**A**) Experimental setup. PbTII cells were transferred from a single donor to multiple recipients. "n" refers to the number of recipient mice per time point. Also shown are the numbers of single cells from which high-quality mRNA-seq data was successfully recorded. The numbers in parentheses refer to the experiment presented in Fig. S12. p.i., post-infection.

(**B-C**) Representative FACS plots showing bifurcation of splenic Th1 (IFNγ$^+$T-bet$^+$) and Tfh (CXCR5$^+$Bcl6$^+$) PbTII CD4$^+$ T cells at day 7 post-infection with *Pc*AS.

(**D**) Flow cytometry data indicate concurrent differentiation of Th1 (IFNγ$^+$) and Tfh (CXCR5$^+$) PbTII CD4$^+$ T cells within the spleen of *Pc*AS-infected mice (n=4). Index expression is the product of MFI and proportion IFNγ$^+$ or CXCR5$^+$. These data are representative of two independent experiments. MFI, mean fluorescence intensity.

(**E**) PCA of single PbTII cells at 7 days post-infection with *Pc*AS. The arrows represent the Pearson correlation with PC1 and PC2. Cell size refers to the number of detected genes. The size of the data points also represents cell size. "Th1 signature" and "Tfh signature" refer to cumulative expression of genes associated with Th1 or Tfh phenotypes (*15*). PC, Principal Component.

(F) Expression of top 50 genes with largest PC2 loadings of day 7 cells (D). The genes were annotated as Th1- or Tfh-associated based on public datasets (*15, 38, 46, 47*). **Cdk2ap2* appears twice because two alternative genomic annotations exist. PC, Principal Component

Figure 2

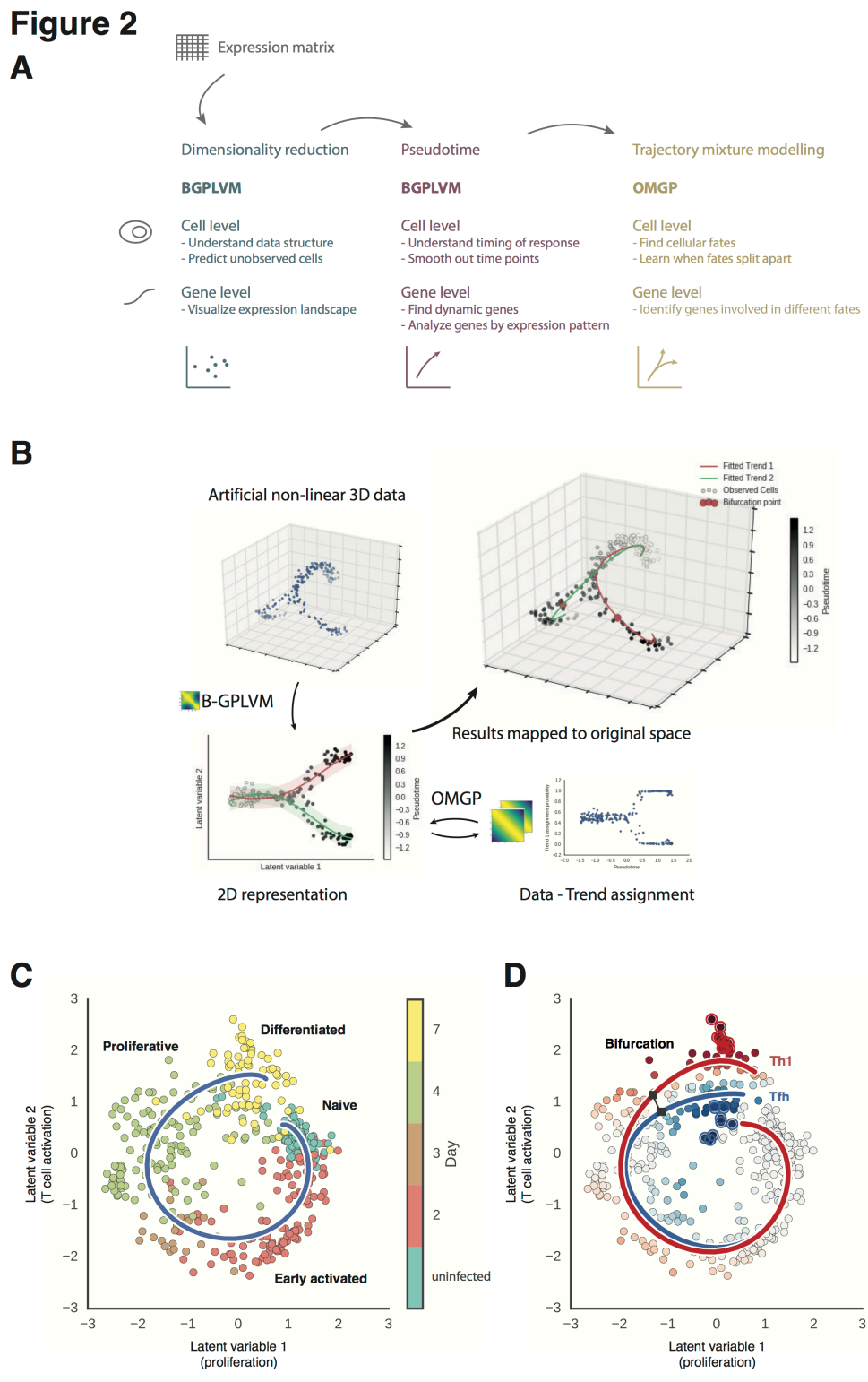


Figure 2. GPfates modelling of bifurcation processes using scRNA-seq data.

(A) Overview of analysis abilities from the framework of Gaussian Processes. Data is modelled and interpreted on the cellular level using the global genomic level data. Through downstream analysis from these models, it is possible to investigate individual genes to explain the drivers of the different models.

(B) Sketch of the analysis workflow. A low-dimensional model of the non-linear high-dimensional data is inferred by Bayesian GPLVM. The low-dimensional representation is then modelled as an Overlapping Mixture of Gaussian Processes. This gives us a data-trend assignment per cell which can be used for interpretation. Since the models are all predictive, the low-dimensional model can be interpreted in the original high-dimensional space.

(C) The low-dimensional representation of our data. The blue line depicts the progression of pseudotime. The text labels illustrate features of typical cells on that region of the pseudotime, and are provided purely as a visual aid.

Figure 3

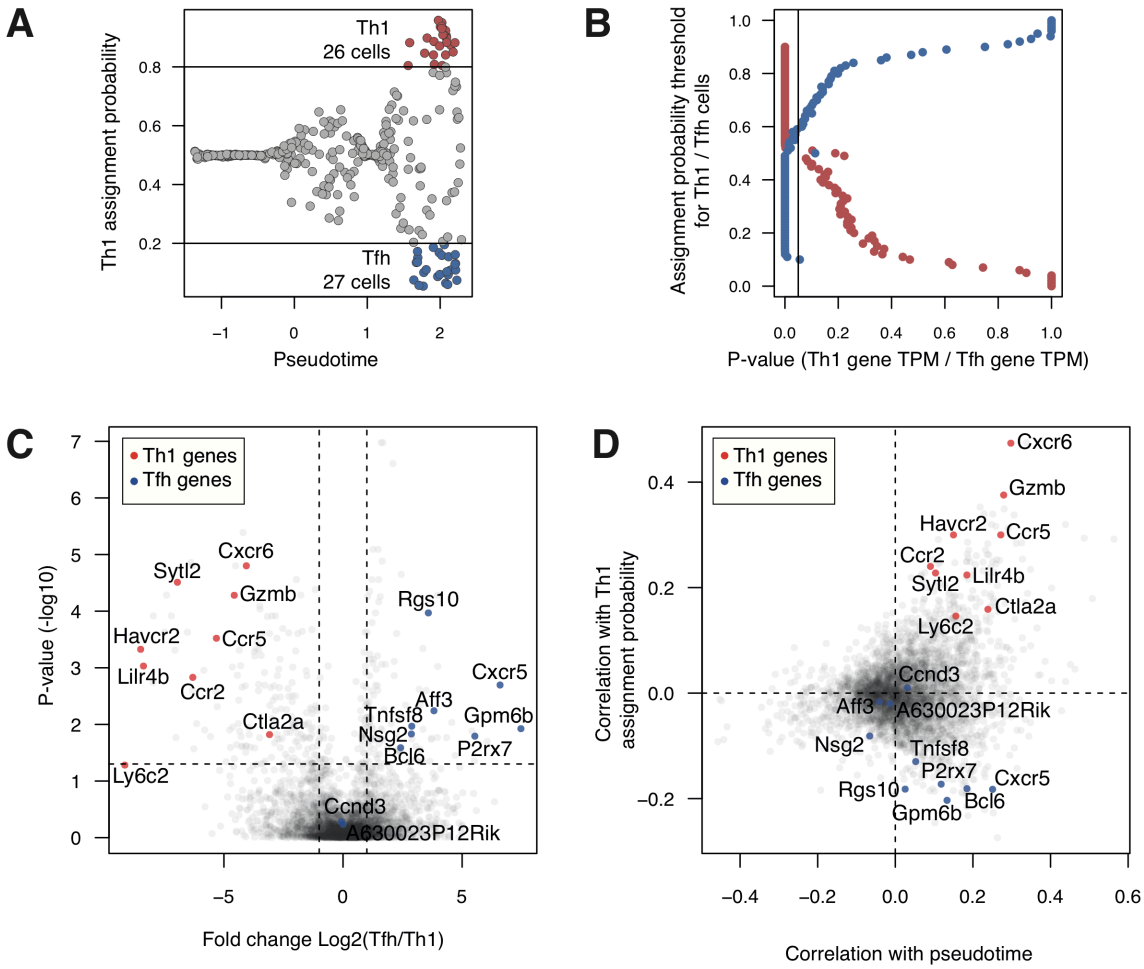


Figure 3. The relationship of known Th1- and Tfh-transcriptomics signatures and the trajectories determined using the GPfates analysis.

(**A**) Th1 and Tfh states were defined as cells with assignment probability of ≥0.8 for the respective trend. For each single cell, cumulative expression of Th1 and Tfh signature genes (*15*) was calculated as in Figure 1E.

(**B**) The effect of the probability threshold on the cumulative expression of Th1 and Tfh signature genes. The p-values were calculated using Wilcoxon rank sum test.

(C) The expression of Th1 and Tfh signature genes in Th1 and Tfh cells defined using the GPfates model (A). For all genes expressed by at least 20% of the single cells, fold changes were

calculated. The p-values were calculated using Wilcoxon rank sum test and adjusted for multiple testing using Benjamini & Hochberg correction.

(**D**) Correlation of expression of Th1 and Tfh signature genes with pseudotime and with Th1 assignment probability.

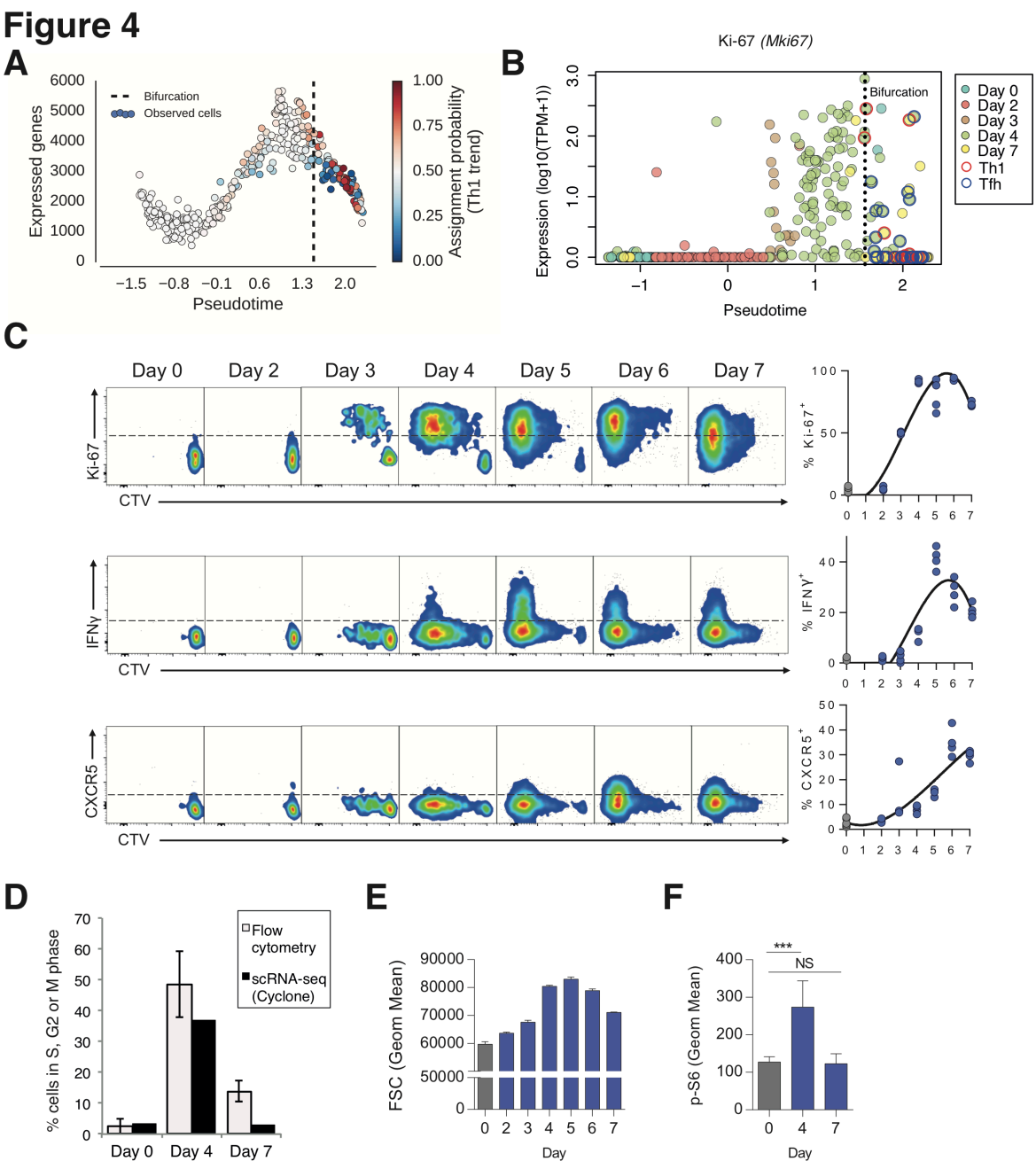


Figure 4. The bifurcation of T cell fates is accompanied by changes in transcription, proliferation and metabolism.

(**A**) The relationship of Th1-Tfh bifurcation point and the number of detected genes per cell.

(**B**) The expression level of the proliferative marker Ki-67 (encoded by the *Mki67* gene) across pseudotime.

(**C**) Representative FACS plots showing kinetics of CellTrace™ Violet (CTV) dilution and Ki67, IFNγ or CXCR5 expression, with summary graphs showing % of PbTII cells expressing these (after 10^6 PbTII cells transferred) in uninfected (Day 0) and *PcAS*-infected mice at indicated days post-infection (n=4 mice/timepoint, with individual mouse data shown in summary graphs; solid line in summary graphs indicates results from third order polynominal regression analysis.) Data are representative of two independent experiments.

(**D**) Experimental and computational analysis of cell cycle speed of PbTII $CD4^+$ T cells activated in response to *Pc*AS. The allocation of cells to cell cycle phases was performed by flow cytometry using Hoechst staining (Figure S9B) and computationally using the Cyclone algorithm (*22*). The relative cell-cycle speed was determined by measuring the fraction of cells in S, G2, or M phases.

(**E**) Cell size estimation using FSC (Forward Scatter) measurements of PbTII cells.

(**F**) Cellular metabolic activity of PbTII cells in naive mice (n=3) and at days 4 and 7 post-infection (n=6) as determined by flow cytometric assessment of ribosomal protein S6 phosphorylation (p-S6). Histogram and proportions are representative of two independent experiments. Statistics are one-way ANOVA and Tukey's multiple comparisons tests ***$p<0.001$.

Figure 5

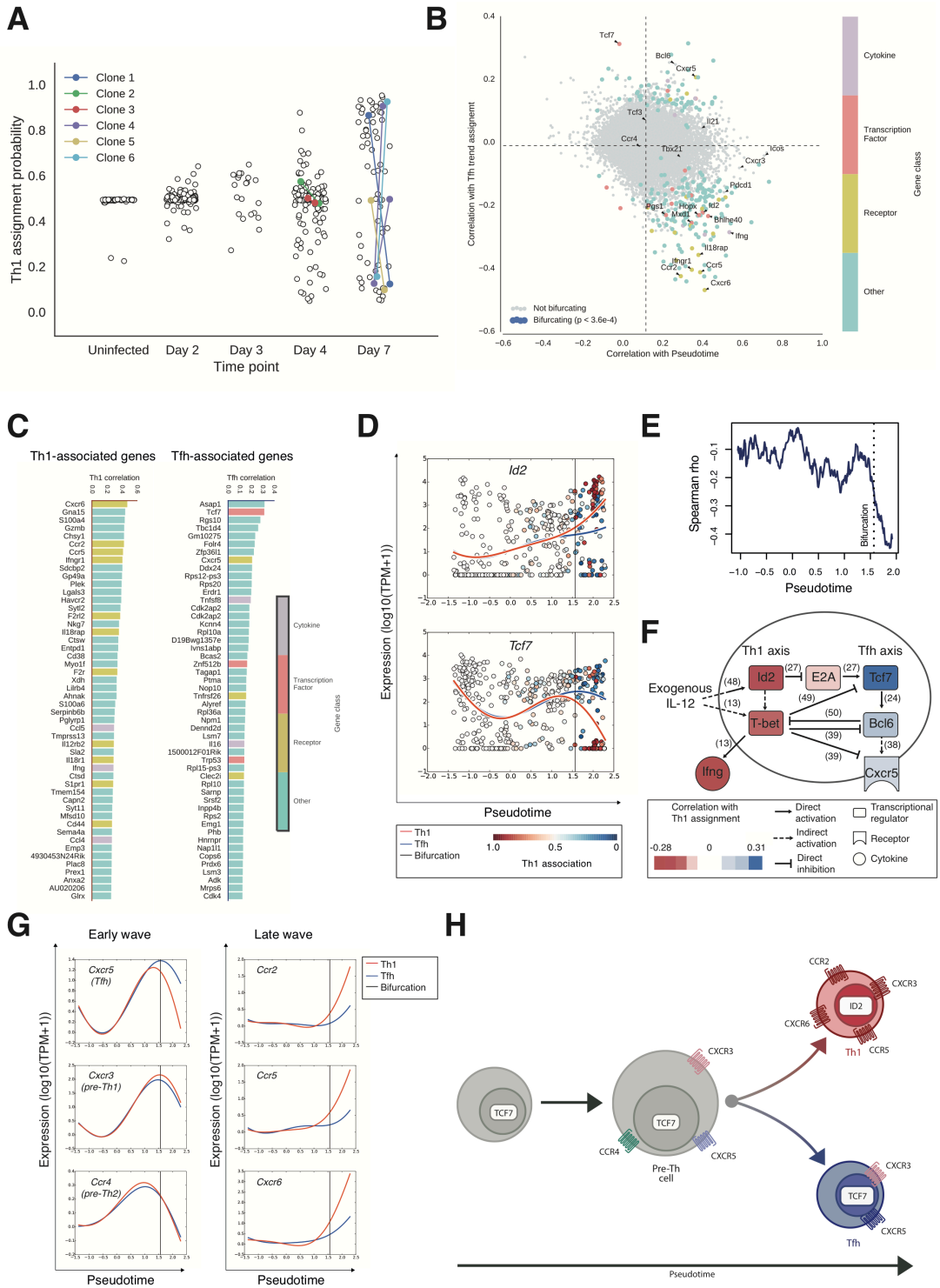


Figure 5. Mechanisms underlying the differentiation of Th1 and Tfh cells.

(**A**) Parallel Th1 and Tfh differentiation within cells of a single $CD4^+$ T cell clone. The colours represent clones determined by sequence analysis of secondary T cell receptor genes (Supplementary Tables 2 and 3).

(**B**) Identification of genes associated with the differentiation of Th1 or Tfh cells. For every gene, the correlation of its expression with pseudotime (x-axis) and Tfh trend assignment (y-axis) are shown. Statistical significance was determined using the bifurcating score (methods). Genes satisfying the significance threshold of FDR<0.002 are represented in colours according to the functional classification of the genes (methods and Supplementary Table 4). FDR, False Discovery Rate, estimated by performing the same analysis with permuted data.

(**C**) The genes with strongest association with Th1 (left) or Tfh differentiation (right). The genes were filtered using the bifurcation score as in (B). The genes were then ranked in descending order of association with either Th1 or Tfh trend. *Cdk2ap2* appears twice because two alternative genomic annotations exist.

(**D**) The expression of *Id2* (upper panel) and *Tcf7* (lower panel) across the pseudotime. The curves represent the Th1 (red) and Tfh (blue) trends when weighing the information from data points according to trend assignment. The colour of the data points represents the strength of the relationship with the two alternative trends.

(**E**) The correlation of *Id2* and *Tcf7* expression at single-cell level. Using a rolling window method, Spearman rho was calculated in windows of 100 cells. The pseudotime values are mean values within each window.

(**F**) A model depicting known interactions of *Id2* and *Tcf7*. The colours represent the Pearson correlation of gene expression and the Th1 trend assignment in single cells. The numbers in parentheses refer to the original publications (*13, 24, 27, 38, 39, 48-50*).

(**G**) The expression kinetics of the chemokine receptor genes *Cxcr5, Cxcr3, Ccr4, Ccr2, Ccr5,* and *Cxcr6* across pseudotime. The curves represent the expression patterns associated with the Th1 (red) and the Tfh (blue) trends.

(**H**) A model summarizing the expression patterns of *Id2*, *Tcf7*, and the chemokine receptors during Th1-Tfh cell fate determination. The size of the cell represents proliferative capacity (Fig. 4 (B-E). The colour of receptors and transcription factors represent differences in expression level.

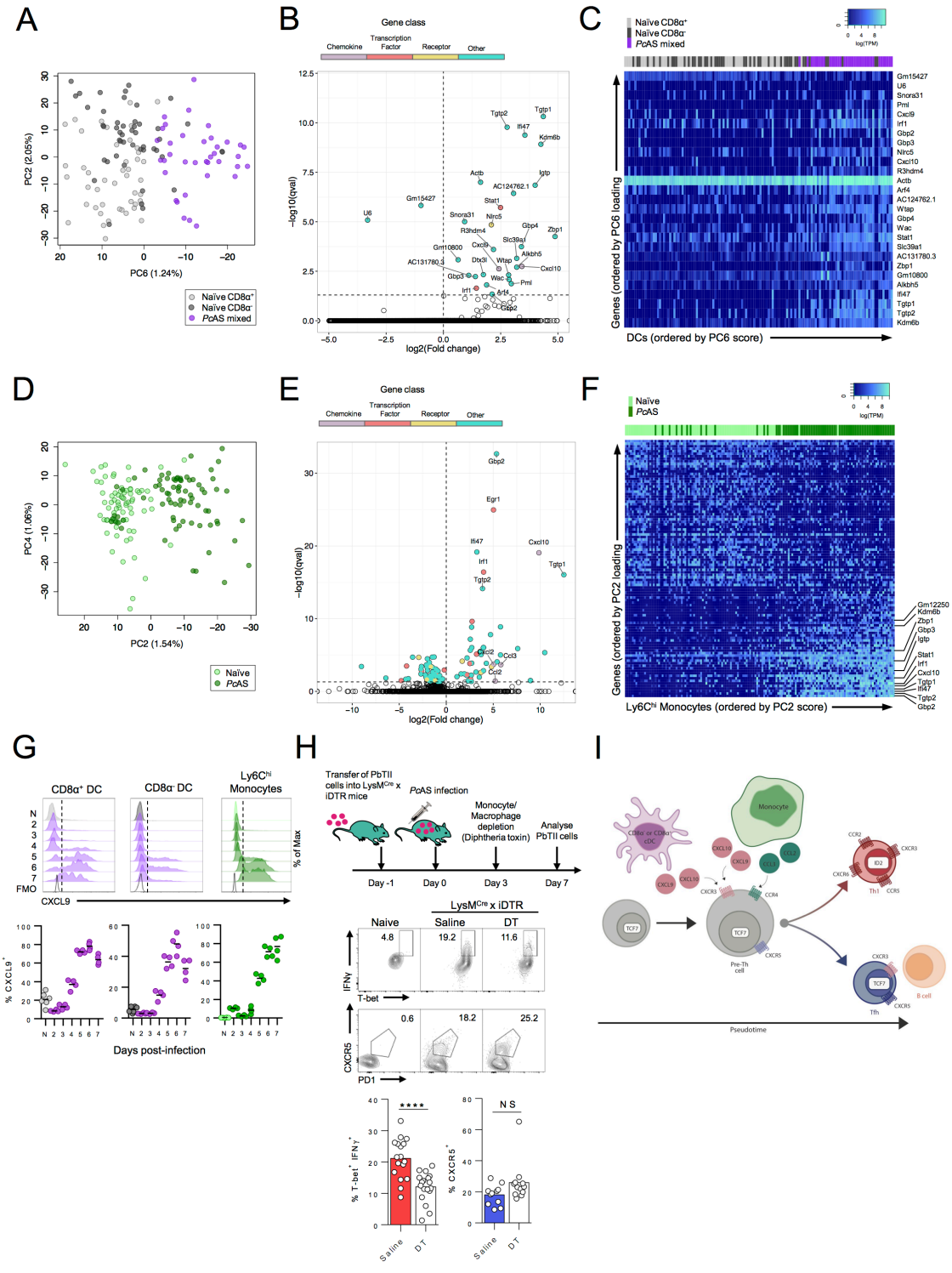


Figure 6. Myeloid cells influence Th bifurcation in uncommitted PbTII cells.

(**A-C**) 131 single splenic CD8α^{+} and CD11b^{+} CD8α^{-} cDCs from a naïve mouse, mixed cDCs from a day 3-infected mouse, and **(D-F)** 154 single Ly6C^{hi} monocytes from naïve and infected mice were analysed by scRNAseq, with mRNA reads filtered by minimum expression of 100 TPM in at least 2 cells. (**A** & **D)** Principal Component Analyses, showing PC combinations best separating populations of (A) cDCs, and (D) Ly6C^{hi} monocytes from naïve and infected mice. (**B** & **E**) Volcano plots showing fold-change and confidence for differentially expressed genes (*17*) between (B) cDCs or (E) monocytes in infected versus naïve mice - genes filtered on expression in >10 cells; genes satisfying qval < 0.05 are represented in colours according to functional classification displayed. Full gene lists are provided in Figure S19 and S21, respectively. **(C** & **F)** Expression heatmaps for significantly (qval<0.05) differentially expressed genes in (C) cDCs and (F) Ly6C^{hi} monocytes, between naive and infected mice: cells and genes are ordered according to PC score and loading respectively, using PC6 for cDCs, and PC2 for Ly6C^{hi} monocytes. The 12 common genes between cDCs and monocyte heatmaps are annotated in (F). (**G**) Representative FACS histograms and proportions of splenic CD8α^{+} cDCs, CD8α^{-} cDCs and Ly6C^{hi} monocytes expressing CXCL9 in naive and infected mice between 2-7 days post-infection - individual mouse data plotted with line at mean; data representative of two independent experiments (n=4 mice/time point/experiment). (**H**) Scheme depicting experimental design: PbTII cells were transferred into *LysM*^{Cre} x *iDTR* mice 1 day prior to infection. At 3 days p.i., mice were treated with diphtheria toxin (DT) or control saline, with PbTII Th1/Tfh responses assessed at 7 days p.i.. Representative FACS plots (gated on splenic PbTII cells) showing Th1 proportions (T-bet^{hi} IFNγ^{+}) and Tfh proportions (CXCR5^{+}) in DT or saline-treated *LysM*^{Cre} x *iDTR* mice; data pooled from two independent experiments. Numbers depict proportions within respective gates. Statistics: Mann-Whitney U Test. ****p<0.0001; NS, not significant. (**I**) Summary model proposing chemokine interactions between non-bifurcated PbTII cells and myeloid cells support a Th1 fate, while Tfh fates are sustained by B cells.

References and Notes:

1. J. Zhu, H. Yamane, W. E. Paul, Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* **28**, 445 (2010).
2. N. J. Tubo *et al.*, Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. *Cell* **153**, 785 (May 9, 2013).
3. S. Crotty, T follicular helper cell differentiation, function, and roles in disease. *Immunity* **41**, 529 (Oct 16, 2014).
4. S. Celli, F. Lemaitre, P. Bousso, Real-time manipulation of T cell-dendritic cell interactions in vivo reveals the importance of prolonged contacts for CD4+ T cell activation. *Immunity* **27**, 625 (Oct, 2007).
5. M. Bajenoff, O. Wurtz, S. Guerder, Repeated antigen exposure is necessary for the differentiation, but not the initial proliferation, of naive CD4(+) T cells. *Journal of immunology* **168**, 1723 (Feb 15, 2002).
6. R. Obst, H. M. van Santen, D. Mathis, C. Benoist, Antigen persistence is required throughout the expansion phase of a CD4(+) T cell response. *The Journal of experimental medicine* **201**, 1555 (May 16, 2005).
7. C. Kim, T. Wilson, K. F. Fischer, M. A. Williams, Sustained interactions between T cell receptors and antigens promote the differentiation of CD4(+) memory T cells. *Immunity* **39**, 508 (Sep 19, 2013).
8. J. R. Groom *et al.*, CXCR3 chemokine receptor-ligand interactions in the lymph node optimize CD4+ T helper 1 cell differentiation. *Immunity* **37**, 1091 (Dec 14, 2012).
9. S. M. Kerfoot *et al.*, Germinal center B cell and T follicular helper cell development initiates in the interfollicular zone. *Immunity* **34**, 947 (Jun 24, 2011).
10. D. Baumjohann *et al.*, Persistent antigen and germinal center B cells sustain T follicular helper cell responses and phenotype. *Immunity* **38**, 596 (Mar 21, 2013).
11. C. Shin *et al.*, CD8alpha(-) Dendritic Cells Induce Antigen-Specific T Follicular Helper Cells Generating Efficient Humoral Immune Responses. *Cell reports* **11**, 1929 (Jun 30, 2015).
12. D. Perez-Mazliah, J. Langhorne, CD4 T-cell subsets in malaria: TH1/TH2 revisited. *Frontiers in immunology* **5**, 671 (2014).
13. S. J. Szabo *et al.*, A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* **100**, 655 (Mar 17, 2000).
14. R. J. Johnston *et al.*, Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation. *Science* **325**, 1006 (Aug 21, 2009).
15. J. S. Hale *et al.*, Distinct memory CD4+ T cells with commitment to T follicular helper- and T helper 1-cell lineages are generated after acute viral infection. *Immunity* **38**, 805 (Apr 18, 2013).
16. N. D. Lawrence, M. K. Titsias, Bayesian gaussian process latent variable model. *Proceedings of the 13th international conference on artificial intelligence and statistics*, 844 (2010).
17. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381 (Apr, 2014).
18. E. Marco *et al.*, Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E5643 (Dec 30, 2014).
19. B. Treutlein *et al.*, Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371 (May 15, 2014).
20. F. Guo *et al.*, The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161**, 1437 (Jun 4, 2015).
21. M. L. Whitfield, L. K. George, G. D. Grant, C. M. Perou, Common markers of proliferation. *Nature reviews. Cancer* **6**, 99 (Feb, 2006).
22. A. Scialdone *et al.*, Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54 (Sep 1, 2015).
23. Y. S. Choi *et al.*, LEF-1 and TCF-1 orchestrate T(FH) differentiation by regulating differentiation circuits upstream of the transcriptional repressor Bcl6. *Nature immunology* **16**, 980 (Sep, 2015).

24. L. Xu *et al.*, The transcription factor TCF-1 initiates the differentiation of T(FH) cells during acute viral infection. *Nature immunology* **16**, 991 (Sep, 2015).
25. B. Luscher, MAD1 and its life as a MYC antagonist: An update. *Eur J Cell Biol* **91**, 506 (Jun-Jul, 2012).
26. M. Kanda *et al.*, Transcriptional regulator Bhlhe40 works as a cofactor of T-bet in the regulation of IFN-gamma production in iNKT cells. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3394 (Jun 14, 2016).
27. F. Masson *et al.*, Id2-mediated inhibition of E2A represses memory CD8+ T cell differentiation. *Journal of immunology* **190**, 4585 (May 1, 2013).
28. L. A. Shaw *et al.*, Id2 reinforces TH1 differentiation and inhibits E2A to repress TFH differentiation. *Nature immunology* **17**, 834 (Jul, 2016).
29. S. Picelli *et al.*, Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171 (Jan, 2014).
30. R. J. Lund, E. K. Ylikoski, T. Aittokallio, O. Nevalainen, R. Lahesmaa, Kinetics and STAT4- or STAT6-mediated regulation of genes involved in lymphocyte polarization to Th1 and Th2 cells. *European journal of immunology* **33**, 1105 (Apr, 2003).
31. C. H. Kim *et al.*, Bonzo/CXCR6 expression defines type 1-polarized T-cell subsets with extralymphoid tissue homing potential. *The Journal of clinical investigation* **107**, 595 (Mar, 2001).
32. S. Hardtke, L. Ohl, R. Forster, Balanced expression of CXCR5 and CCR7 on follicular T helper cells determines their transient positioning to lymph node follicles and is essential for efficient B-cell help. *Blood* **106**, 1924 (Sep 15, 2005).
33. T. Junt *et al.*, CXCR5-dependent seeding of follicular niches by B and Th cells augments antiviral B cell responses. *Journal of immunology* **175**, 7109 (Dec 1, 2005).
34. D. Breitfeld *et al.*, Follicular B helper T cells express CXC chemokine receptor 5, localize to B cell follicles, and support immunoglobulin production. *The Journal of experimental medicine* **192**, 1545 (Dec 4, 2000).
35. P. Schaerli *et al.*, CXC chemokine receptor 5 expression defines follicular homing T cells with B cell helper function. *The Journal of experimental medicine* **192**, 1553 (Dec 4, 2000).
36. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776 (Feb 14, 2014).
37. S. C. Bendall *et al.*, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714 (Apr 24, 2014).
38. X. Liu *et al.*, Bcl6 expression specifies the T follicular helper cell program in vivo. *The Journal of experimental medicine* **209**, 1841 (Sep 24, 2012).
39. S. Nakayamada *et al.*, Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* **35**, 919 (Dec 23, 2011).
40. Y. S. Choi *et al.*, Bcl6 expressing follicular helper CD4 T cells are fate committed early and have the capacity to form memory. *Journal of immunology* **190**, 4014 (Apr 15, 2013).
41. M. Pepper, A. J. Pagan, B. Z. Igyarto, J. J. Taylor, M. K. Jenkins, Opposing signals from the Bcl6 transcription factor and the interleukin-2 receptor generate T helper 1 central and effector memory cells. *Immunity* **35**, 583 (Oct 28, 2011).
42. Y. S. Choi *et al.*, ICOS receptor instructs T follicular helper cell versus effector cell differentiation via induction of the transcriptional repressor Bcl6. *Immunity* **34**, 932 (Jun 24, 2011).
43. R. J. Lundie *et al.*, Blood-stage Plasmodium berghei infection leads to short-lived parasite-associated antigen presentation by dendritic cells. *European journal of immunology* **40**, 1674 (Jun, 2010).
44. A. Haque *et al.*, Type I IFN signaling in CD8- DCs impairs Th1-dependent malaria immunity. *The Journal of clinical investigation* **124**, 2483 (Jun, 2014).
45. J. M. Marchingo *et al.*, T cell signaling. Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. *Science* **346**, 1123 (Nov 28, 2014).
46. H. D. Marshall *et al.*, Differential expression of Ly6C and T-bet distinguish effector and memory Th1 CD4(+) cell properties during viral infection. *Immunity* **35**, 633 (Oct 28, 2011).
47. M. J. Stubbington *et al.*, An atlas of mouse CD4(+) T cell transcriptomes. *Biology direct* **10**, 14 (2015).

48. C. Y. Yang *et al.*, The transcriptional regulators Id2 and Id3 control the formation of distinct memory CD8+ T cell subsets. *Nature immunology* **12**, 1221 (Dec, 2011).
49. K. J. Oestreich, A. C. Huang, A. S. Weinmann, The lineage-defining factors T-bet and Bcl-6 collaborate to regulate Th1 gene expression patterns. *The Journal of experimental medicine* **208**, 1001 (May 9, 2011).
50. K. Hatzi *et al.*, BCL6 orchestrates Tfh cell differentiation via multiple distinct mechanisms. *The Journal of experimental medicine* **212**, 539 (Apr 6, 2015).
51. A. Crawford *et al.*, Molecular and transcriptional basis of CD4(+) T cell dysfunction during chronic infection. *Immunity* **40**, 289 (Feb 20, 2014).
52. B. Mahata *et al.*, Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell reports* **7**, 1130 (May 22, 2014).

Acknowledgments:

We thank the Wellcome Trust Sanger Institute Sequencing Facility for performing Illumina sequencing, Wellcome Trust Sanger Institute Single-cell Genomics Core Facility for single-cell sample processing and the Wellcome Trust Sanger Institute Research Support Facility for care of the mice used in these studies. We thank QIMR Berghofer Flow Cytometry and Animal Facilities for expert advice and care of wild-type and transgenic mice. We wish to acknowledge Stephan Lorenz, Joanna Cartwright and Tom Metcalf for expert technical assistance. We thank Guy Emerton for constructing the database and the interface for accessing the data. We thank Michel Raymond for his work in defining cytokines and cell-surface receptors. Susanna Ng is acknowledged for assistance with graphic design.

This work was supported by European Research Council grant ThSWITCH (number 260507), Australian National Health and Medical Research Council Project grant (number 1028641) and Career Development Fellowship (no. 1028643), University of Queensland, Australian Infectious Disease Research Centre grants and the Lister Institute for Preventative Medicine. KRJ was supported by grants from EMBL Australia and OzEMalaR. FOB was supported by the Lundbeck Foundation. MZ was supported by the Marie Curie ITN grant "Machine Learning for Personalized Medicine" (EU FP7-PEOPLE Project Ref 316861, MLPM2012).

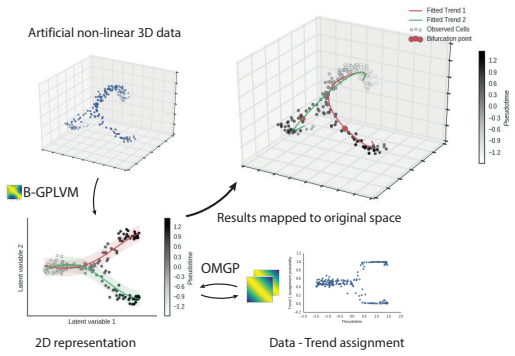
The data presented in this paper is publically available in the ArrayExpress database.

Supplementary Computational Methods - The GPfates model

July 27, 2016

1 Introduction

GPfates is based on a three-stage approach that first i) infers a low-dimensional representation of single-cell RNA-seq data, then ii) infers pseudotime to iii) model the temporal dynamics of gene expression profiles with a mixture model. These steps build on existing modeling components: The Gaussian Process Latent Variable Model [Lawrence, 2006], and the Overlapping Mixture of Gaussian Processes [Lázaro-Gredilla et al., 2012]. For a graphical illustration of the major steps involved in this analysis, see Figure 2D of the main text (as well as Supp. Comp. Fig 1). In Sections 2 and 3 we describe the statistical models that underlie the components of GPfates. In Section 4 we describe downstream analysis methods for interpreting the fitted model. Finally, in Section 5, we present additional validation experiments using simulations, robustness analyses and by analyzing multiple existing data sets.



Sup. Comp. Fig. 1: Illustration of the analysis workflow. A low dimensional parametrization of the data is found using Bayesian GPLVM. The low-dimensional representation is viewed as a mixture problem, and solved by an Overlapping Mixture of Gaussian Processes. This allows us to represent our cells as members of different smooth processes. But also interpret in terms of the high-dimensional space parametrized by the GPLVM.

2 Pseudotime inference

2.1 Gaussian Process Regression

A main component of GPfates is to model temporal transitions. We use the Gaussian process (GP) framework, thereby casting this problem as non-parametric regression. Let us begin by assuming that the developmental time t for each cell we observe is known. Then, the output y_g (i.e. expression of gene g) is modelled as a continuous function of the input t (i.e. developmental progression)

$$y_g = f(t) + \varepsilon, \tag{1}$$

where

$$p(\varepsilon) = \mathcal{N}(0, \sigma^2)$$

is Gaussian distributed residual noise and $f(t)$ denotes the unknown regression function. In this work y_g is considered to be an N-dimensional vector of N cells with observed expression of the gene g. We denote the expression of g in an individual cell n as $[y_g]_n$.

A GP can be interpreted as a function-valued prior on the elements of f, which is defined by a covariance function that in turn is parametrized by the input (developmental time) t:

$$\mathrm{cov}(f(t_{n_1}), f(t_{n_2})) = k(t_n, t_{n_2}).$$

The *covariance function* $k(t_{n_1}, t_{n_2})$ encodes prior assumptions on the smoothness and lengthscales of the function $f(t)$. The most widely used covariance function is the Squared Exponential (SE) covariance function,

$$k(t_{n_1}, t_{n_2}) = \sigma^2_{\mathrm{SE}} \exp\left(-\frac{|t_{n_1} - t_{n_2}|^2}{2l^2_{\mathrm{SE}}}\right), \tag{2}$$

and this is the covariance function we will generally be used in this work. This covariance has the hyperparameters $\theta = (\sigma^2_{\mathrm{SE}}, l^2_{\mathrm{SE}})$, which parametrize the amplitude (σ^2_{SE}) and the lengthscale (l^2_{SE}) of functions under the prior. Throughout the remainder of the text we will omit the hyperparameters from equations for the sake of brevity. Note that there is a whole compendium of valid covariance functions, which can also be combined using sum or multiplication; see [cite: Rasmussen, GP 2006] for an overview.

We write that a function f is *Gaussian Process distributed* by

$$f(t) \sim \mathcal{GP}(0, k(t_{n_1}, t_{n_2})).$$

This prior on the function f can be linked to the finite

observed data using a Gaussian likelihood:

$$p(y_g|f) = \prod_{n=1}^{N} \mathcal{N}([y_g]_n|f_n, \sigma^2).$$

Together with the prior on the corresponding (finite) elements of f,

$$p(f) = \mathcal{N}(f|0, \boldsymbol{K}_t),$$

this results in the marginal likelihood

$$p(y_g|t) = \mathcal{N}(y_g|0, \boldsymbol{K}_t + \sigma^2 \cdot I).$$

Here $\boldsymbol{K}_t$ is an $N \times N$ matrix of pairwise evaluations of the covariance functions at the observed times t. I.e.

$$[\boldsymbol{K}_t]_{n,m} = k(t_n, t_m). \tag{3}$$

By considering the joint distribution of the observed data y_g and an unseen function value $f(t_\star)$, it is possible to derive the predictive distribution for $f(t_\star)$:

$$p(f(t_\star)|t, y_g, t_\star) = \mathcal{N}\left(\overline{f(t_\star)}, k(t_\star)\right),$$

where

$$\overline{f(t_\star)} = k(t_\star, t)[\boldsymbol{K}_t + \sigma^2 \cdot I]^{-1} y_g, \text{ and}$$
$$k(t_\star) = k(t_\star, t_\star) - k(t_\star, t)[\boldsymbol{K}_t + \sigma^2 \cdot I]^{-1} k(t, t_\star).$$

For a full review on Gaussian Processes, see Williams and Rasmussen [2006].

So far, we have only described Gaussian Process Regression for expression y_g of a single gene g. If we consider a collection of G genes $\{1, \ldots, G\}$, their expression can be modelled together by

$$(y_1, \ldots, y_G) = (f^1(t), \ldots, f^G(t)) + (\varepsilon, \ldots, \varepsilon). \tag{4}$$

We use Y to compactly denote the $N \times G$ expression matrix of cells $\times$ genes, where

$$Y_{n,g} = [y_g]_n.$$

The assumption that all genes are governed by similar functional relationships with t means we place the same GP prior (with shared covariance function):

$$p(Y|t) = \prod_{g=1}^{G} p(y_g|t) = \prod_{g=1}^{G} \mathcal{N}(y_g|0, \boldsymbol{K}_t + \sigma^2 \cdot I). \tag{5}$$

In the next section we will see the usefulness of considering multiple genes at once.

2.2 Pseudotime inference by Bayesian GPLVM with *per*-cell prior

The Gaussian Process regression framework described above assumes we know the time t of each cell. While in many single-cell RNA-seq experiments record a collection times over some time-course, these are rather sparse, and it has been pointed out [Trapnell et al., 2014] that cells are sampled from a population where responses are *unsynchronized*. Each cell has reached a certain stage in the differentiation process under investigation, which we do not observe directly. The progress in to this process is referred to as *pseudotime*. We can however infer this from the data. In the Gaussian Process Latent Variable Model (GPLVM) [Lawrence, 2006], we use the multiple output case of Gaussian Process regression (equation 4), but consider the values of t to be parameters which we wish to infer.

The joint probability of the GPLVM is

$$p(Y, t) = p(Y|t)p(t),$$

where $p(Y|t)$ is defined in equation 5, and the prior $p(t)$ is such that for cell n,

$$p(t_n) = \mathcal{N}(0, 1).$$

Following Reid and Wernisch [2016], we can also consider the prior $p(t)$ to be informed about the experimental ordering of collection times of the cells, putting the mean of t_n to correspond to the time point of cell n. If we use our Malaria time course as and example, we can put the prior on t so that

$$p(t_n) = \mathcal{N}(\text{day}_n, \sigma^2_{\text{prior}}),$$

where $\text{day}_n \in \{1, 2, 3, 4, 5\}$ correspond to the collection order of those cells. The parameter σ^2_{prior} alters the strength of the prior.

The objective of Bayesian GPLVM [Titsias and Lawrence, 2010], is to find the posterior probability distribution $p(t|Y) \propto p(Y|t)p(t)$. This is intractable though, due to the t values appearing non-linearly in the matrix inverse $[\boldsymbol{K}_t + \sigma^2 \cdot I]^{-1}$.

In Titsias and Lawrence [2010], a lower bound to the marginal likelihood is calculated by estimating the posterior $p(t|Y)$ by a variational distribution $q(t)$. The distribution

$$q(t) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mu_n, S_n)$$

is described in that paper, and Bayesian training of the model to maximize this lower bound. This is the method we use.

Because the scale of t is ambiguous, in particular if no priors are specified, we prefer to at times scale the inferred t to the range $[0, 1]$ when reporting the pseudotime, to avoid confusion about negative "time". In these cases we refer to pseudotime as *scaled pseudotime* in the legends.

2.3 Dimensionality reduction

In many cases it is useful to work on a reduced representation of cellular expression profiles. For example, when modelling transcriptomic data, fitting a model to a low-dimensional representation can be preferable to

fitting it to expression profiles of thousands of genes. Formally, the objective of *dimensionality reductions* is to find some M-dimensional representation of the G-dimensional expression measurements, where $M << G$. Typically M is 2 or 3, which aides visual interpretation. Analogous to the pseudotime inference, these latent cell states can also be inferred using the GPLVM. Say X is an $M \times N$ matrix so that each cell n correspond to an M-dimensional vector,

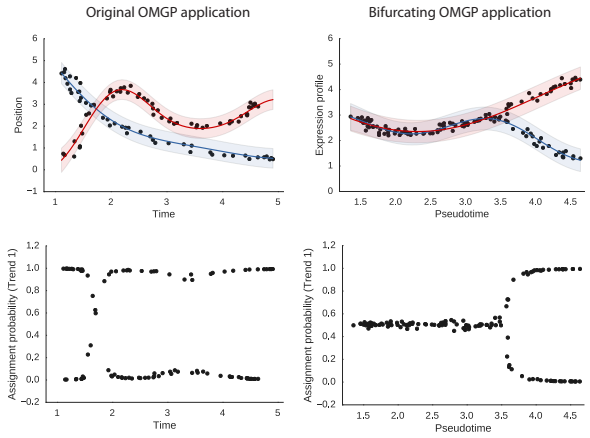
$$X_n = (x_n^1, \ldots, x_n^M).$$

We want to model the expression matrix Y so that

$$[y_g]_n = f^g(x_n^1, \ldots, x_n^M) + \varepsilon = f^g(X_n) + \varepsilon.$$

Note that now the covariance function is evaluated as $k(X_{n_1}, X_{n_2})$, where, in the Squared Exponential covariance function in equation 2, the operator $|\cdot|$ is evaluated as the Euclidean norm for vectors, rather than absolute value.

Just as the t_n values are inferred from data above, so can the X_n vectors be inferred from the data.



Sup. Comp. Fig. 2: Comparison of the original OMGP use case (left) and our use case (right), in both cases where the number of trends $K = 2$. In the original use case trends are expected to be independent throughout time, albeit with some ambiguity in some locations. In our application, we interpret ambiguous cell assignment to be in a common precursor state.

3 Bifurcation inference using overlapping mixtures of Gaussian processes

In a continuous setting, a bifurcating process can be seen as one function, splitting apart into two functions over time. One approach to model this could be to consider two functions throughout time, but before the bifurcation happens, the two functions are identical. With this in mind, we can use a mixture model to tease apart the shared and bifurcated functions.

3.1 Mixture model

Mixture models are hierarchical models where an observation is assumed to be generated from one of C *components*, each of which is described by its own model. The goal of mixture models is to infer which component an observation stems from, and at the same time model that component.

The Overlapping Mixture of Gaussian Processes (OMGP) model [Lázaro-Gredilla et al., 2012] assumes there are C different underlying latent functions producing the N observed cells. This model was originally developed for the application of missile tracking, and in that setting an observation is e.g. a radar based location at a given real time point. As such, the main focus of the definition of the model is for the case of C completely independent components. The approach presented here is based on the realisation that the model would also be able to handle the case of *branching* trajectories. There would simply be a time interval where it does not matter which mixture trajectory data is sampled from. In our setting, an observation is a single cell, and the analog to real time is pseudotime (Supp. Comp. Fig. 2). As an additional extension, we phrase a version of the OMGP model which is non-parametric in the number of trajectories.

In the original regression case described in equation 1, data is assumed to be generated by a single smooth unknown function. When modeling our gene expression data with the Overlapping Mixture of Gaussian Processes, data is considered to be generated by

$$X = f_c(t) + \varepsilon.$$

However, we are lacking information about which latent function f_c generated any given observation (t_n, X_n) of pseudotime and gene expression for the N observed cells. Here X correspond to some representation of the transcriptional state of the cells. It could be the expression of all genes ($X = Y$), a single gene ($X = y_g$), or an M-dimensional inferred representation as discussed above.

This is viewed as a mixture modelling problem, where each cell has a latent variable z_i specifying to which component f_c the cell should be allocated to. Write F for the collection of all latent functions. The covariance functions k_c for each f_c can be different from each other, though for the applications we discuss here, we take them as Squared Exponential covariance functions with different hyperparameter values.

In the OMGP formulation, the likelihood is

$$p(X|F, T, Z) = \prod_{n=1}^{N} \prod_{c=1}^{C} \mathcal{N}(x_i|f_c(t_n), \sigma^2)^{z_{nc}}.$$

We specify a multinomial prior on the latent variables

Z, namely

$$p(Z) = \prod_{n=1}^{N}\prod_{c=1}^{C} \Pi_{n,c}^{z_{n,c}},$$

$$\sum_{c=1}^{C} \Pi_{n,c} = 1.$$

Additionally, each of the latent functions f_c has an independent Gaussian process prior:

$$p(F|T) = \prod_{c=1}^{C} \mathcal{N}(f_c|0, \boldsymbol{K}_t^c).$$

The covariance matrices $\boldsymbol{K}_t^1, \ldots, \boldsymbol{K}_t^C$ for the latent functions $f_1, \ldots, f_C$ are generated from a covariance functions $k_1(t_{n_1}, t_{n_2}), \ldots, k_C(t_{n_1}, t_{n_2})$ like in equation 3.

Now we rephrase this as a Dirichlet Process Gaussian Process mixture model [Hensman et al., 2015]. Let every latent function f_c have an associated "stick-breaking length" v_c, based on the "stick-breaking" formulation of the Dirichlet Process. Here $V = [v_1, \cdots, v_\infty]$ is the collection of stick-breaking lengths for constructing the Dirichlet process for the assignment. The joint distribution of the OMGP model is

$$p(X, Z, V, F) = p(F|T)p(X|F, Z)p(Z|V)p(V|\alpha).$$

The value α is a parameter of the model which controls the expected concentrations of mixtures (which we in practice take as $\alpha = 1$, a common default), and

$$p(V|\alpha) = \prod_{c=1}^{\infty} Beta(v_c|1, \alpha),$$

where $Beta(\cdot, \cdot)$ is the beta distribution. The prior distribution over the collection of Gaussian Processes is

$$p(F|T) = \prod_{c=1}^{\infty} \mathcal{N}(f_c|0, \boldsymbol{K}_c).$$

Following the stick-breaking formulation,

$$p(Z|V) = \prod_{i=1}^{N}\prod_{c=1}^{\infty} \pi_c(V)^{z_{i,c}},$$

where $\pi_c(V) = v_c \prod_{j=1}^{c-1}(1 - v_j)$.

The assignments between observations X and the latent functions F is given by a binary $N \times C$ matrix Z. The assignments to latent functions are considered as additional variational parameters. Let $\boldsymbol{\phi}$ be an $N \times C$ matrix where ϕ_{nc} is the approximate posterior probability of assigning the nth observation to the cth latent function. The $\boldsymbol{\phi}$ parameters are inferred by collapsed variational inference as described in Hensman et al. [2012]. Overall, the likelihood of the model is

$$p(X|F, Z) = \prod_{n=1}^{N}\prod_{c=1}^{\infty} \mathcal{N}(x_n|f_c, \boldsymbol{K}_c)^{z_{n,c}}.$$

(It should be noted that everything described generalizes to the case where the latent functions f_c are vector valued, as long as all output dimensions of such a function share the same covariance function. In this case, probabilities factorize over output dimensions, but beyond that all calculations are the same.)

3.2 Parameter inference

In Lázaro-Gredilla et al. [2012] the latent variables $\mathbf{Z}$ in the parametric version of OMGP were inferred using an expectation-maximization scheme. Here we describe how we perform variational inference for the $\boldsymbol{\phi}$-parameters in the non-parametric version of the model.

To make the inference problem tractable, the variational distribution $q(Z)$ is introduced with variational parameters ϕ, at a given truncation level C such that

$$q(Z) = \prod_{n=1}^{N}\prod_{c=1}^{C} \phi_{n,c}^{z_{n,c}}.$$

with the objective of approximating $p(Z|F, X, T)$.

The lower bound of the log-likelihood of the OMGP model, which we write as $\mathcal{L}_{\mathrm{KL}}$, when approximating $p(Z)$ by $q(Z)$ can be split up in three terms as

$$\mathcal{L}_{\mathrm{KL}} = \mathcal{L}^{\mathrm{M}} + \mathcal{L}^{\mathrm{MP}} + \mathcal{L}^{\mathrm{H}}.$$

Here $\mathcal{L}^{\mathrm{M}} = \sum_{c=1}^{C} \mathcal{L}_c^{\mathrm{M}}$ is the log-likelihood of the latent functions as represented by Gaussian processes. For the cth latent function, the variational distribution of f_c which maximizes the lower bound was derived in Lázaro-Gredilla et al. [2012] to be

$$q(f_c) = \mathcal{N}(f_c|\mu_c, \boldsymbol{\Sigma}_c)$$

where $\boldsymbol{\Sigma}_c = (\boldsymbol{K}_c^{-1} + \boldsymbol{B}_c)^{-1}$, and $\mu_c = \boldsymbol{\Sigma}_c \boldsymbol{B}_c y_c$. Here $\boldsymbol{B}_c$ is a diagonal matrix with entries $[\boldsymbol{B}_c]_{i,i} = \frac{\phi_{i,c}}{\sigma^2}$. Thus the log-likelihood for a particular latent function f_c, assuming we have optimal assignments ϕ, is

$$\mathcal{L}_c^{\mathrm{M}} = -\frac{1}{2} y^T \boldsymbol{\Sigma}_c^{-1} y - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| - \frac{N}{2} \ln 2\pi.$$

The second and third parts of $\mathcal{L}_{KL}$ were derived in [Hensman2014] as

$$\begin{aligned}\mathcal{L}^{\mathrm{MP}} &= \ln \int \exp\{\mathbb{E}_{q(Z)}\left[\ln p(Z|V)\right]\} p(V) dV \\ &= \ln \prod_{c=1}^{C} \left(\frac{\Gamma(\hat{\phi}_c + 1)\Gamma(\tilde{\phi}_c + \alpha)\alpha}{\Gamma(\hat{\phi}_c + \tilde{\phi}_c + \alpha + 1)} \right)\end{aligned}$$

and

$$\mathcal{L}^{H} = -\mathbb{E}_{q(Z)}\left[\ln q(Z)\right].$$

For optimizing variational mixture assignment parameters we follow Hensman et al. [2012], and use *natural gradient descent*. For hyperparameters of the kernels, as well as the variance parameter σ^2 of the model, we perform gradient descent.

If we know $\frac{\partial \mathcal{L}_{KL}}{\partial \phi}$ we can calculate the natural gradient by equation (22) in Hensman et al. [2015]. The gradients $\frac{\partial \mathcal{L}^{MP}}{\partial \phi}$ and $\frac{\partial \mathcal{L}^{H}}{\partial \phi}$ were derived in Hensman et al. [2015], the only unknown part is $\frac{\partial \mathcal{L}^{M}}{\partial \phi}$.

We then use the identity $\frac{\partial \mathcal{L}_k^M}{\partial \phi_{n,i}} = \frac{1}{2}\,\mathrm{Tr}\left(\frac{\partial \mathcal{L}_c^M}{\partial \boldsymbol{B}_c^{-1}} \cdot \frac{\partial \boldsymbol{B}_c^{-1}}{\partial \phi_{n,i}}\right)$. Here $\frac{\partial \mathcal{L}_c^M}{\partial \boldsymbol{B}_c^{-1}} = \alpha\alpha^T - (\boldsymbol{K}_c + \boldsymbol{B}_c^{-1})^{-1}$, and the matrix $\frac{\partial \boldsymbol{B}_c^{-1}}{\partial \phi_{n,i}}$ will be zero everywhere, except in the diagonal element (n,n) where it will be $\frac{-\sigma^2}{\phi_{n,i}^2}$.

Using the chain rule, we can calculate log-likelihood gradients of the model hyperparameters for any covariance function, since we know $\frac{\partial \boldsymbol{K}_c}{\partial \theta}$, resulting in a very general and modular framework. We only need $\frac{\partial \mathcal{L}_{KL}}{\partial K_c} = \frac{\partial \mathcal{L}_c^M}{\partial B_c^{-1}} = \alpha_c \alpha_c^T - (K_c + B_c^{-1})^{-1}$. In the case of the model variance σ^2 we have $\frac{\partial \mathcal{L}^M}{\partial \sigma^2} = \sum_k \frac{1}{2}\,\mathrm{Tr}\left(\frac{\partial \mathcal{L}_c^M}{\partial B_c^{-1}} \cdot \frac{\partial B_c^{-1}}{\partial \sigma^2}\right)$ where $\frac{\partial B_c^{-1}}{\partial \sigma^2}$ will be a diagonal matrix with $\frac{1}{\phi_{i,c}}$ on element (i,i) for all i.

4 Downstream analysis

4.1 Ranking genes by bifurcation

Once the OMGP model has been fit, it can be used to investigate individual genes in terms of their bifurcating trajectory.

The log-likelihood of the OMGP model depends on the covariance matrices $\mathbf{K}_t = \{\boldsymbol{K}_t^c, c = 1, \ldots, C\}$, the variational mixture parameter matrix ϕ, and the N observations (t, X). Let us assume that we have mixture parameters ϕ_b which have been found to distinguish a bifurcating trend based on some X response variables. We can now keep the fitted parameters and evaluate the marginal likelihood of a model where the response variables X are replaced by gene expression values y_g. We call this new model $\mathcal{H}_{\text{bifurcating}}$. We wish to find genes which fit this bifurcating model better than a model where this is no bifurcation. To this end, we make a third model $\mathcal{H}_{\text{not bifurcating}}$ identical to the precious one, except we replace ϕ_b with ambiguous assignments ϕ_a. To asses whether a given gene g is better described by the *bifurcating* or the *not bifurcating* model, we evaluate the Bayes factor:

$$BF_g = \log p(y_g|\mathcal{H}_{\text{bifurcating}}) - \log p(y_g|\mathcal{H}_{\text{not bifurcating}}).$$

We refer to this ratio as the *bifurcation statistic*.

To estimate p-values, we used a permutation approach where we perform the same analysis for every gene g, except with permuted t values to estimate a null distribution.

As a proxy for *effect size* of bifurcation, we consider how well the expression values of a gene correlate with the trend assignments to a latent function. Strong positive correlation will mean the gene is particularly up-regulated in the cells unambiguously belonging to the trend. Conversely, a strong negative correlation indicates the gene is down-regulated in the strongly assigned cells compared to all cells.

4.2 Inferring the bifurcation time point

It is possible to qualitatively appreciate from the GP assignment probability (ϕ_c) for each trajectory ($f_c(t)$) of the OMGP model, which cells are *ambiguous* and which cells are *exclusive* to individual GP's. In the case of two trends, ambiguous cells have assignment probability (ϕ) close to 0.5. A model where the data can be described by two trends, but not by one, will have a higher likelihood. Similarly, if only a *region* of the ϕ parameters over time are replaced by ambiguous cell assignment values, the new model will have a lower likelihood.

For the sake of clarity, we make the assumption that the OMGP will begin as ambiguous, and then become less ambiguous over time, splitting into two trends, in this special case. To investigate these cases, we pick a time-point t_b in an OMGP, then replace all ϕ values prior to t_b with 0.5. We define this new ϕ as $\phi_{>t_b}$:

$$\begin{aligned} [\phi_{>t_b}]_{i,c} &= 0.5 && t_i < t_b \\ [\phi_{>t_b}]_{i,c} &= \phi_{i,k} && t_i \geq t_b. \end{aligned}$$

Now we can evaluate the model likelihood for this particular t_b and define

$$\mathcal{L}_{t_b} = \mathcal{L}_{KL}(\phi_{>t_b}, \mathbf{K}_t, \sigma^2 | X, T).$$

This procedure is repeated for multiple ts over the predictor variable of the OMGP model. In our implementation, we consider 30 evenly spaced bins by default, which has given enough resolution for the data investigated (though the number of bins can easily be changed).

The likelihood has to decrease by definition. However, after the bifurcation the decrease is much more pronounced. We use a break-point heuristic to detect this elbow, which is indicative of the bifurcation time.

To identify the region at which the likelihood decreases more rapidly, we fit a piece-wise linear curve to the log-likelihoods, defined by

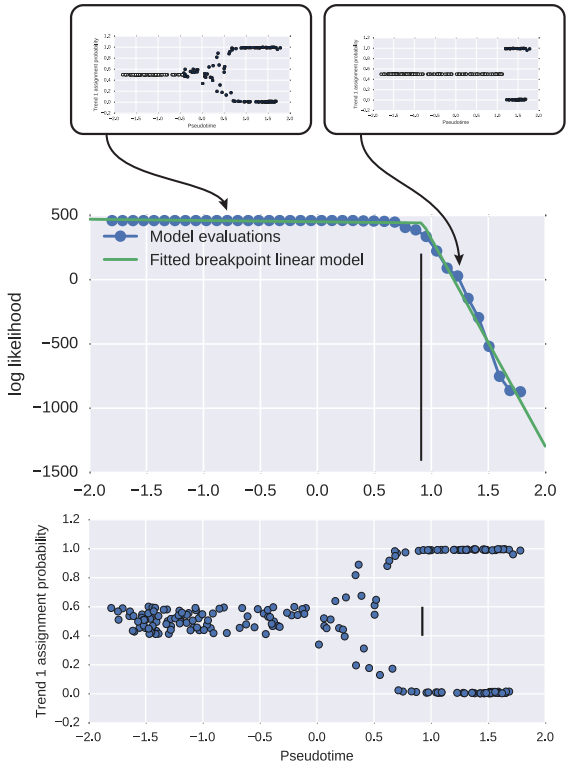
$$\begin{aligned} \mathcal{L}_{t_b} &= k_1 \cdot t + c_1 && t < t_b \\ \mathcal{L}_{t_b} &= k_2 \cdot t + (k_1 - k_2) \cdot p + c_1 && t \geq t_b \end{aligned}$$

This curve consists of two linear pieces, broken up at the point p. When the curve is fitted, we consider the break-point p to be the point after which we can be confident that a bifurcation has occurred, see Supp. Comp. Fig 3.

5 Implementation and combination with existing workflows

5.1 Practical use of GPfates

The basis principle of GPfates is the combination of pseudotime and mixture modelling.



Sup. Comp. Fig. 3: Inferring bifurcation point. The plot illustrates how different points along the pseudotime are sampled. Ambigous assignment probabilities replaces trained assignment probabilities in the observations earler than the sampled points. The breakpoint model identifies the points where a decrease in likelihood differences becomes more extreme.

Input to the GPLVM is an expression table consisting of log scaled relative abundance values Transcripts Per Million, TPM, with a value of 1 added to handle cases where expression is 0. As relative abundance follow a log-normal distribution, the Gaussian likelihood used for Gaussian Process regression should be appropriate.

In practice, the pseudotime should represent the biological process of interest. If this process is clear, the expression data should be usable without pre-processing. In single cell time course experiments where the process of interest is less immediate, a strategy highlighted in Trapnell et al. [2014] is to select the gene set used could be to rank the genes by an ANOVA test over the time points, and select a larger number of significant genes.

Similarly, the low-dimensional representation of the transcriptomic cell state should represent the biological response of interest. It can be beneficial to select the parts of the representation which correspond to this. For example, in the analysis of CD4+ T cell time course, we use the second GPLVM latent variable as a representation of T cell response, and model this factor by the OMGP.

While the pseudotime can be inferred directly from the expression matrix Y, in many cases it helps interpretation to perform an intermediate step of dimensionality reduction. This process could also be beneficial if the data has a very complex structure.

Another practical consideration we must consider is that single cell expression vales can be quite noisy. This limits the time-scale at which we can expect to measure proper functional differences in expression levels. Due to this, we tend to put lower limits on the lengtscale l_{SE} of the squares exponential covariance function.

5.2 Integration of existing methods

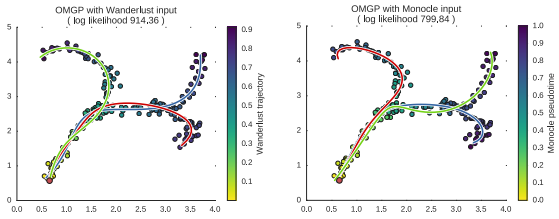
We have presented use of the GPfates method when pseudotime or low-dimensional representations have been based on the GPLVM. This is because the OMGP follows from this framework, and the statistical assumptions are consistent between the models.

In practice, other methods for inferring pseudotime or low-dimensional representations could also be considered. Here we briefly outline possible strategies for applications of GPfates downstream of popular single-cell analysis methods.

Recall that to perform the GPfates inference, we need pseudotime t and some representation of transcriptomic state X. These variables can be set as the output from other methods.

In Monocle [Trapnell et al., 2014], the low-dimensional representation X is found by independent component analysis, and the pseudotime t for each cell is defined by the path distance to a starting cell through a minimum spanning tree in the coordinates of X.

In Wanderlust, a heuristic is used to build a stable k Nearest Neighbor (kNN) graph of the data in the high-dimensional space of protein measurements. The pseu-



Sup. Comp. Fig. 4: OMGP is compatible with e.g. Wanderlust and Monocle, as demonstrated with a toy data set

dotime t for a cell is then defined as the average shortest path from a known starting cell through the kNN graph. Note that for CyTOF data, which Wanderlust is designed for, only up to 40 analytes can be measured at once, so it could be feasible to take X to be the original expression matrix (Y in our notation).

Another dimensionality reduction technique which have been used for single cell RNA seq data is Diffusion Maps [Haghverdi et al., 2015]. Here X is a spectral embedding of the data manifold, based on the Laplace operator. It has been pointed out that these embeddings preserve branching structure in the data. Taking the pseudotime t as the Diffusion Pseudotime [Haghverdi et al., 2016], an approximation of geodesic distance over the data manifold (from a known starting cell), based on the diffusion map, GPfates modelling could be used downstream to quantify the branching structure of the data.

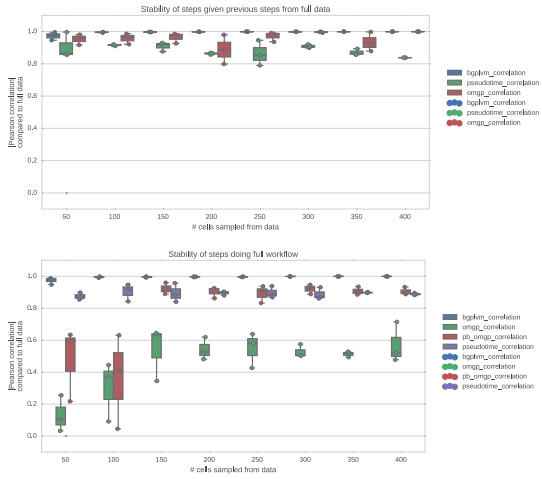
We list alternative compatible pseudotime methods in table 1.

As a demonstration, we generated a toy data set with three branches, and extracted the pseudotime using both the Monocle method and the Wanderlust method. Then fitted and OMGP with $C = 3$ on the output. The results can be seen in Comp. Supp. Fig 4, which illustrates the correct identification of the branching processes for either input.

5.3 Software availability

We have made a software package for using the GPfates method, which is available at `https://github.com/Teichlab/GPfates`. It provides guidance and sensible defaults for the kind of analysis we have described here. It makes extensive use of the `GPy`[1] package, and the `GPclust`[2] package, where we implemented the non-parametric OMGP model.

[1] `https://github.com/SheffieldML/GPy`
[2] `https://github.com/SheffieldML/GPclust`



Sup. Comp. Fig. 5: Robustness of analysis steps by subsampling. Parameters inferred from a subsample of the data are compared to parameters inferred using the full data. The top panel indicates this analysis for independent steps assuming the previous step is known. The lower panel shows the result when running the workflow from start to end.

6 Assessment of GPfates on simulated and real data

6.1 Sample-size robustness analysis

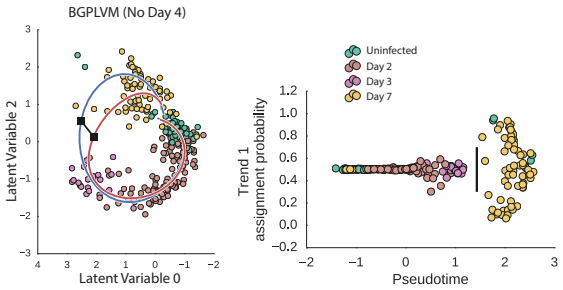
Our full analysis consists of several independent consecutive steps: first the GPfates method where we are i) finding a low-dimensional representation, ii) smoothing the data over a pseudotime, and iii) finding a trend mixture model. After this we perform downstream analysis where we are iv) dentifying the end states and bifurcation.

How much data do we need to accurately reconstruct trends from all four of the above steps, and how much data is needed for individual steps? We investigated both how stable the full procedure is, as well as the individual steps, by re-running it on sub-sampled datasets with fewer cells than the entire dataset.

To measure the stability of the methods, we consider absolute Pearson correlation of the parameters inferred for sub-sampled data relative to the results obtained from performing the analysis on the full data set.

We found that recovering a low-dimensional representation is extremely stable with respect to the number of cells (Supp. Comp. Fig. 5), with almost perfect correlations between analysis of the sub-sampled data and the original GPLVM values. (For example, the lowest absolute Pearson correlation for a run with 50 cells was 0.96). Similarly, the pseudotime inference is also very stable to sub-sampling.

Finding the entire OMGP mixture model over pseudotime requires a larger number of cells. We don't see



Sup. Comp. Fig. 6: Complete reanalysis of our T-cell data excluding cells cellected at day 4. The bifurcation point is identified as being between Day 3 and Day 7, and is not forced in to either of the days.

any higher degrees of consistency until we reach 150 sub-sampled cells, with correlations around 0.5. It is rare to see single cell studies with so few cells, and in the study accompanying this text we had many more cells (408). Identifying only the end states is rather robust (but in many cases might be best analyzed as a cluster problem rather than a continuous value problem), where we start seeing a correlation of 0.9 at 150 cells.

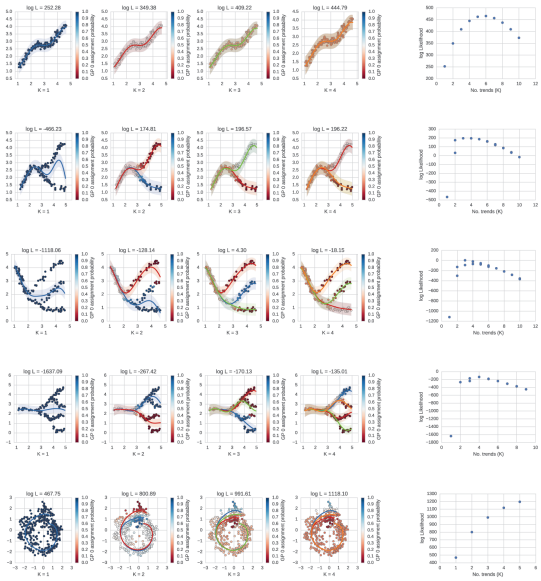
The individual steps were in general very stable to sub-sampling, relative to the "gold standard" of using the full data set. When running the entire procedure, we see that smaller errors early on in the analysis will propagate and affect later steps.

6.2 Predicted bifurcation time is not biased by collection times

We consider the risk that the identified bifurcation point in the CD4+ T cell data potentially just reflects the time points at which we have collected data. We test the robustness of the prediction of the bifurcation as having happened at Day 4 by re-running the analysis after removing cells collected at Day 4. In this analysis we find that the bifurcation happens at some point between Day 3 and Day 7 where we don't have any observed cells. The alternate hypothesis would have been that the bifurcation would be found in either Day 3 or Day 7. This provides confidence both in the bifurcation point identification, and more generally in the meaningfulness of the low-dimensional GPLVM representation of the data (Supp. Comp. Fig. 6).

6.3 Assessment of the ability to select the number of trends in OMGP

In principle, the marginal log likelihood of the OMGP model should let us select the C number of trends which optimally explain the data. We investigated this by generating four synthetic data sets with between 1 and 4 underlying trends. For each of the data sets, we optimized OMGP models with the number of trends C varying

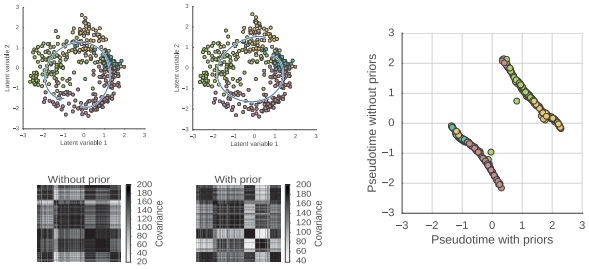


Sup. Comp. Fig. 7: Attempts detecting number of trends with OMGP. Simulated data with expected numbers of trends where fitted with OMGP, where the C cutoff was set to a range of values.

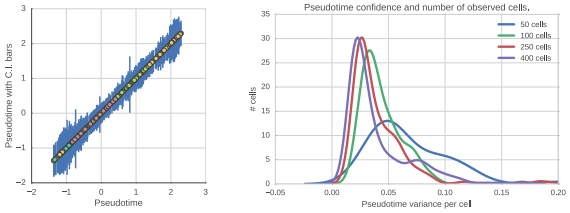
from 1 to 9 (three times per C value). We found that the marginal likelihood of the models corresponded to the correct number of trends in the cases of 3 and 4 ground truth trends, but not for the 1-trend and 2-trend synthetic data. For 1 trend, the likelihood was lowest for a larger number of trends, and for 2 trends, the likelihood was very similar for 2 and 3 trends. This suggests that the OMGP may have a tendency to overestimate the number of trends if there is a single progression. Supp. Comp. Fig. 7.

For our CD4+ T cell data, we found that the marginal likelihood continuously increased with the number C. We elected to keep the model simple and made the assumption that we could sufficiently explain the data with $C = 2$.

It is possible that the optimal likelihood for K is not well defined when we have trends branching off from a common trend. In the original application of the OMGP model, the assumption is that the trends will be completely independent of each other. As we are already to some extent failing to fit two models in the ambiguous case, this might cause the likelihood to reflect a poor fit. For quantitatively determining the number of trends in the data, further work is needed, probably with a model which explicitly considers branching from a common original trend. The marginal likelihood of the model is an indication, but the choice of C should also reflect the biological system under consideration.



Sup. Comp. Fig. 8: Comparison of pseudotime with and without per-cell priors. The upper left shows the fit of the pseudotime predicted in to the 2D GPLVM with and without priors. Below are the corresponding inferred covariance matrices. The right plot shows the relations between the two versions of pseudotime, clearly indicating that they have an approximate one-to-one mapping.



Sup. Comp. Fig. 9: Investigation of uncertainty of inferred pseudotimes. Left panel, since the Bayesian GPLVM fits the variance of the pseudotime for each cell, we can compare the assignments with each other. The bars correspond to 95% confidence intervals. On the right panel we see how the lengths of the confidence intervals globally decrease as the number of cells used increases.

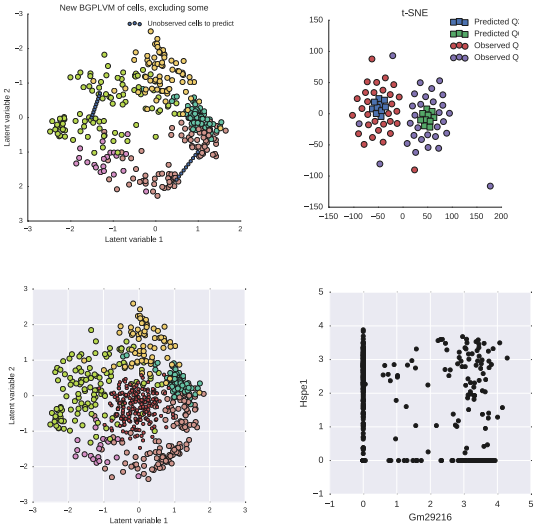
6.4 Comparison of pseudotime inference with and without priors

For the 1-dimensional Bayesian GPLVM which we use to find the pseudotime of the data, we put priors on the cells based on their known time points. This is not strictly necessary, but helps to enhance interpretability as there is intrinsic invariance to the inferred values. If we do not use priors, qualitatively, the same trajectory is identified. Additionally, comparing the two versions of pseudotime against each other, we see that they correspond to a circular shift relative to each other. The covariance matrices inferred using either strategy have a very similar block structure (Frobenius norm ... of difference) indicating that neighbor relations are consistent. Supp. Comp. Fig. 8.

6.5 Pseudotime uncertainty

As pointed out in Campbell and Yau [2015], we can use the posterior distribution of pseudotime from the Bayesian GPLVM to assess how meaningful the ordering is. By investigating the confidence intervals of the pseudotime for each cell compared to neighboring cells, we see that the ordering is quite meaningful (few cells overlap in confidence interval). (Supp. Comp. Fig. 9)

We also investigated how the confidence of the pseudotime depends on the number of cells observed. As the number of observed cells increases, the distribution of variance per cell decreases towards zero. (Supp. Comp. Fig. 9)



Sup. Comp. Fig. 10: Stability of GPLVM representation, and prediction through GPLVM. Top row: Predicting cells from regions of show higher similarity with left out real cells from corresponding regions than non-corresponding regions. Bottom row: Predicting cells from unobserved regions potentially identifies antagonizing gene combinations.

6.6 Stability of the circular shape of the GPLVM representation

We wanted to rule out the possibility that the latent variable representations of data which appear circular might be artifacts due to random noise, as suggested by Diaconis et al. [2008]. To make sure this was not the case for our CD4+ T cell data, we removed two 'slices' of cells from the circular 2D GPLVM pattern. Following this, we fitted a new GPLVM with this reduced data set. After optimizing the GPLVM, a representation was found which was again missing the same slices, Supp. Comp. Fig. 10A. (The correlation between the two representation for the common cells is very high as well, XX). This control experiment strongly suggests that the GPLVM learns the actual topology of the data.

6.7 Assessing the accuracy of imputing virtual cells

Unlike many other dimensionality reduction techniques, the GPLVM creates a model which maps into the high dimensional observed space. It is, however, not clear how meaningful this representation is. We assessed this by taking the "slice-less" model described above, and in the empty areas corresponding to the removed cells, predicting "virtual cells" (Supp. Comp. Fig 10A). Using an independent clustering technique, t-SNE [Van der Maaten and Hinton, 2008], on both the left out slices of cells and the predicted virtual cells, we find that single cell transcriptomes predicted from a given slice coincide with the real cells from the corresponding slice (Supp. Comp. Fig. 10). This indicates that GPLVM prediction in to high-dimensional spaces is not simply producing overfitted results.

Following from this, we investigated the "hole" in our CD4+ T cell data. We create number of virtual cells from the hole region and compare which genes would be expressed in these cells compared to genes expressed in all cells (Supp. Comp. Fig. 10C). The underlying reasons for data being non-linear is that particular combinations of gene expression patterns do not occur together. If we find genes which are high in the virtual cells but are not observed at the same time in actual cells, this might indicate that they are incompatible with each other. This might be a good complementary tool for generating hypotheses about regulation. For instance, we identified the genes Hspe1 and Gm29216 which would be co-expressed in the hole, but are generally not co-expressed in observed cells (Supp. Comp. Fig. 10D).

7 Validating the BGPLVM and OMGP approach by application to other data sets

In order to further corroborate our analysis approach, we considered two recently published single cell data sets produced to investigate progression of single cells in two developmental contexts: mouse fetal lung and human fetal primordial germ cells.

7.1 Analysis of lung development data

We downloaded the data from Treutlein et al. [2014] and quantified the expression using Salmon. To smooth the data over pseudotime, we found genes that vary over the *a priori* known time points by a likelihood ratio test of an ANOVA model of the time points. The expression values for the top varying genes were run on a GPLVM. One of the factors of the optimized GPLVM was used as pseudotime, and the top two factors of the GPLVM were used to represent the entire data set. An OMGP was then optimized on this low-dimensional representation to identify the two trends corresponding to the AT1 and AT2 lung cell lineages without prior annotation. The bifurcation statistic for all expressed genes in these cells reconstituted many of the genes identified in a largely manual manner by Treutlein et al. [2014].

7.2 Analysis of human primordial germ cell data

The data from Guo et al. [2015] was downloaded and quantified with Salmon as with the other data, but with an index based on the human transcriptome: Ensembl 78 annotationa of GRCh38, together with ERCC sequences and human specific repeats from RepBase. To smooth the time course data, we used a likelihood ratio test to find the top genes which were described linearly along the time points in the data. The expression of these genes were then used to fit a GPLVM. This low-dimensional representation of the data was then used to fit an OMGP, taking one of the latent factors as pseudotime.

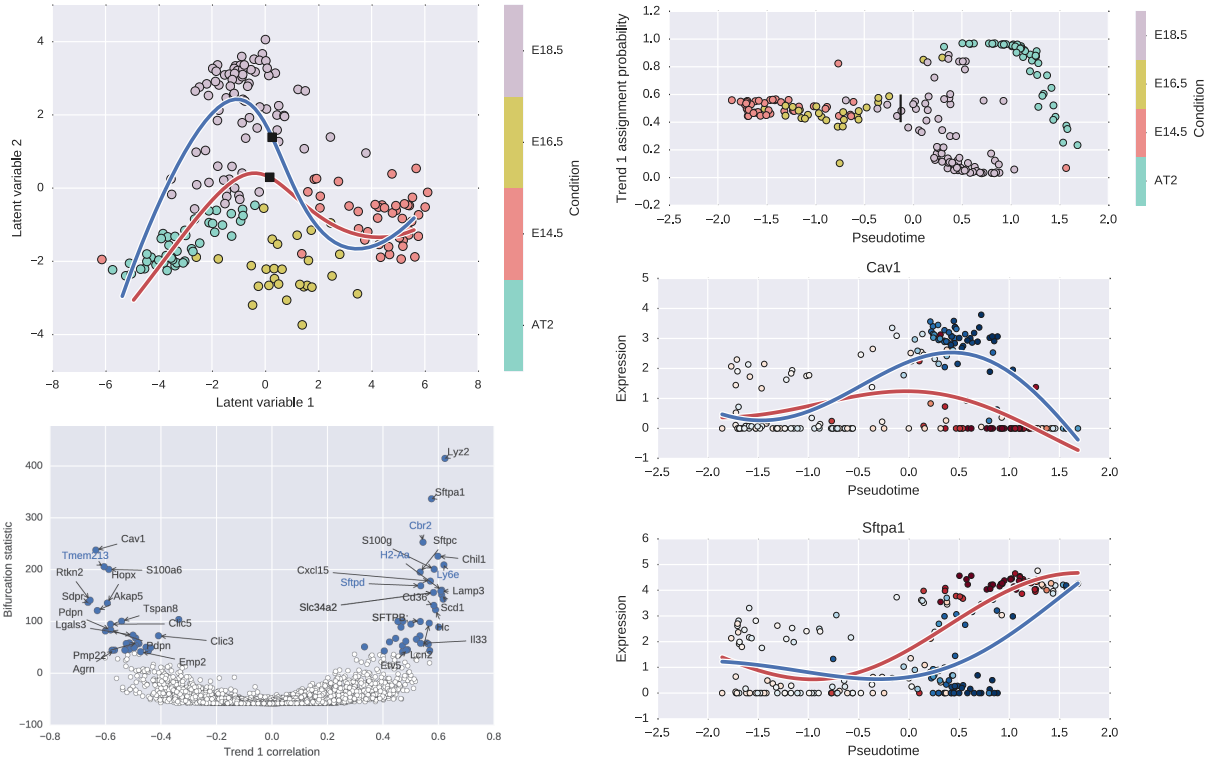
In this data set, the ground truth about the sex of the cells is known, and thus we could have use a supervised approach such as `GPTwoSample` [Stegle et al., 2010] or `DETime` [Yang et al., 2016]. Interestingly, the OMGP model identifies the split between male and female cells in an unsupervised fashion.

We applied the bifurcation statistic test to identify genes which follow the male and female development differently.

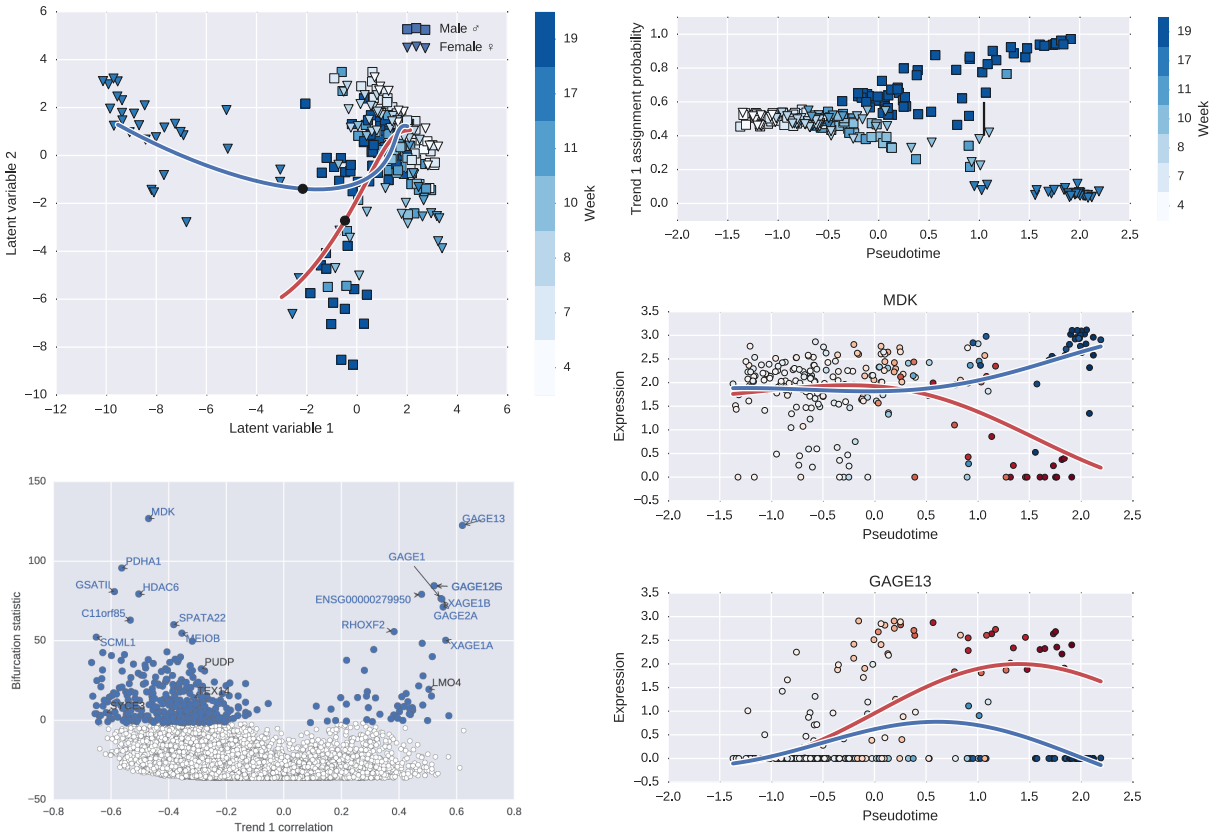
Unlike in the case of the lung development data, the majority of the genes we identify are not discussed in the original study. In the original study, the authors focused on genes specific to given conditions (e.g. Male PGC's from week 11 compared to all other cells). In our analysis, we consider the dynamics of gene expression over development. We find that in the male branch, the GAGE family is highly upregulated over development. Additionally we find the Y-linked gene ENSG00000279950. Also among the top male hits is RHOXF2, a gene linked to male reproduction [Niu et al., 2011]. Further down the list we also interestingly find PIWIL4, a gene with function in development and maintenance of germline stem cells [Sasaki et al., 2003]. On the female side, the top hit is MDK, a gene involved with fetal adrenal gland development (by similariry: UniProtKB P21741). Other top hits include MEIOB, a meiosis related gene, and the satellite repeat GSATII. Surprisingly, we also see upregulation of SPATA22, a gene associated with spermatogenesis.

8 Discussion

We have demonstrated our GPfates method, where we use latent variable modelling to infer temporal expression dynamics, and Gaussian process mixture modelling to identify diverging global trends. The method has been investigated in terms of robustness, and applied



Sup. Comp. Fig. 11: Summary of GPfates result of Treutlein et al developing lung data.



Sup. Comp. Fig. 12: Summary of GPfates result of Guo et al developing primordial germ cell data.

on several simulated and real data sets showing good results.

Of course there is no silver bullet for these sorts of problems, and it would not be surprising if other methods than the ones we have used work better for some biological systems. We have illustrated that the main component, the Gaussian process mixture modelling, is compatible with other methods in these cases.

A benefit from the methods we use is that diagnostics such as marginal likelihood can be used to aide the user with regards to the models to use. Still, the user will need to keep the biological system in mind, and be critical of results.

References

Sean C Bendall, Kara L Davis, El-Ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–725, 24 April 2014. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2014.04.005. URL http://dx.doi.org/10.1016/j.cell.2014.04.005.

Kieran Campbell and Christopher Yau. Bayesian gaussian process latent variable models for pseudotime inference in single-cell RNA-seq data. Technical report, 15 September 2015. URL http://biorxiv.org/lookup/doi/10.1101/026872.

Persi Diaconis, Sharad Goel, and Susan Holmes. Horseshoes in multidimensional scaling and local kernel methods, September 2008. URL http://projecteuclid.org/euclid.aoas/1223908041.

Fan Guo, Liying Yan, Hongshan Guo, Lin Li, Boqiang Hu, Yangyu Zhao, Jun Yong, Yuqiong Hu, Xiaoye Wang, Yuan Wei, Wei Wang, Rong Li, Jie Yan, Xu Zhi, Yan Zhang, Hongyan Jin, Wenxin Zhang, Yu Hou, Ping Zhu, Jingyun Li, Ling Zhang, Sirui Liu, Yixin Ren, Xiaohui Zhu, Lu Wen, Yi Qin Gao, Fuchou Tang, and Jie Qiao. The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell*, 161(6):1437–1452, 4 June 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.05.015. URL http://dx.doi.org/10.1016/j.cell.2015.05.015.

Laleh Haghverdi, Florian Buettner, and Fabian J Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 15 September 2015. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btv325. URL http://dx.doi.org/10.1093/bioinformatics/btv325.

Laleh Haghverdi, Maren Buettner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. Technical report, 29 February 2016. URL http://biorxiv.org/lookup/doi/10.1101/041384.

J Hensman, M Rattray, and N D Lawrence. Fast nonparametric clustering of structured Time-Series. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):383–393, February 2015. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2318711. URL http://dx.doi.org/10.1109/TPAMI.2014.2318711.

James Hensman, Magnus Rattray, and Neil D Lawrence. Fast variational inference in the conjugate exponential family. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2888–2896. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4766-fast-variational-inference-in-the-conjugate-exponential-family.pdf.

Neil D Lawrence. The gaussian process latent variable model. *Technique Report*, 2006.

Miguel Lázaro-Gredilla, Steven Van Vaerenbergh, and Neil D Lawrence. Overlapping mixtures of gaussian processes for the data association problem. *Pattern recognition*, 45(4):1386–1395, April 2012. ISSN 0031-3203. doi: 10.1016/j.patcog.2011.10.004. URL http://www.sciencedirect.com/science/article/pii/S0031320311004109.

Eugenio Marco, Robert L Karp, Guoji Guo, Paul Robson, Adam H Hart, Lorenzo Trippa, and Guo-Cheng Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 111(52):E5643–50, 30 December 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1408993111. URL http://dx.doi.org/10.1073/pnas.1408993111.

Ao-Lei Niu, Yin-Qiu Wang, Hui Zhang, Cheng-Hong Liao, Jin-Kai Wang, Rui Zhang, Jun Che, and Bing Su. Rapid evolution and copy number variation of primate RHOXF2, an x-linked homeobox gene involved in male reproduction and possibly brain function. *BMC evolutionary biology*, 11:298, 12 October 2011. ISSN 1471-2148. doi: 10.1186/1471-2148-11-298. URL http://dx.doi.org/10.1186/1471-2148-11-298.

John E Reid and Lorenz Wernisch. Pseudotime estimation: Deconfounding single cell time series. *Bioinformatics*, 17 June 2016. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btw372. URL http://dx.doi.org/10.1093/bioinformatics/btw372.

Takashi Sasaki, Aiko Shiohama, Shinsei Minoshima, and Nobuyoshi Shimizu. Identification of eight members

Pseudotime Method	Strategy	OMGP Compatibility
Monocle Pseudotime [Trapnell et al., 2014]	Minimum Spanning Tree path length in 2D ICA space	Yes
Diffusion maps [Haghverdi et al., 2015]	Spectral embedding of data manifold	With postprocessing, e.g. DPT [Haghverdi et al., 2016]
Wanderlust [Bendall et al., 2014]	Heuristic k-NN graph geodesic distance	Yes
GPLVM	Latent data parametrization	Yes
SCUBA Pseudotime [Marco et al., 2014]	Principal curve in 3D t-SNE embedding	Yes

Bifurcation Method	Strategy
SCUBA [Marco et al., 2014]	Transitions between clusters in pseudotime bins
Monocle states [Trapnell et al., 2014]	Create k PQ trees from a Minimum Spanning Tree
OMGP	Model data as mixture of continuous processes

Table 1: Examples of other pseudotime methods and bifurcation methods.

of the argonaute family in the human genome. *Genomics*, 82(3):323–330, September 2003. ISSN 0888-7543. URL http://www.ncbi.nlm.nih.gov/pubmed/12906857.

Oliver Stegle, Katherine J Denby, Emma J Cooke, David L Wild, Zoubin Ghahramani, and Karsten M Borgwardt. A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of computational biology: a journal of computational molecular cell biology*, 17(3):355–367, March 2010. ISSN 1066-5277, 1557-8666. doi: 10.1089/cmb.2009.0175. URL http://dx.doi.org/10.1089/cmb.2009.0175.

Michalis K Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010. URL http://jmlr.csail.mit.edu/proceedings/papers/v9/titsias10a/titsias10a.pdf.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4): 381–386, April 2014. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2859. URL http://dx.doi.org/10.1038/nbt.2859.

Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375, 15 May 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13173. URL http://dx.doi.org/10.1038/nature13173.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research: JMLR*, 9(2579-2605):85, 2008. ISSN 1532-4435. URL http://siplab.tudelft.nl/sites/default/files/vandermaaten08a.pdf.

Christopher K I Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006. URL http://www-old.newton.ac.uk/programmes/BNR/seminars/2007080914001.pdf.

Jing Yang, Christopher A Penfold, Murray R Grant, and Magnus Rattray. Inferring the perturbation time from biological time course data. 4 February 2016. URL http://arxiv.org/abs/1602.01743.

Supplementary Materials:

Materials and Methods

Figures **S1-S23**

Fig. S1. Enrichment of PbTII cells for adoptive transfer.

Fig. S2. Sorting strategy for PbTII cells.

Fig. S3. Expression of subset-specific marker genes in PbTII cells.

Fig. S4. Heterogeneity of activated PbTII cells.

Fig. S5. Heterogeneity of Th1/Tfh signature gene expression in activated PbTII cells.

Fig. S6. The contribution of Th1 and Tfh signature genes to the overall heterogeneity of the PbTII time series.

Fig. S7. The relationship of pseudotime with time points and with the Th1 assignment probability.

Fig. S8. Modelling the data using Monocle and SCUBA.

Fig. S9. Proliferative burst of activated PbTII cells.

Fig. S10. Expression of transgenic and endogenic TCRs.

Fig. S11. Correlation of expression of *Ifng* with *Tcf7 and Id2* across pseudotime.

Fig. S12. The expression of *Tbx21* (left) and *Bcl6* (right) across pseudotime

Fig. S13. Robustness of top bifurcating genes across experiments.

Fig. S14. Flow cytometric validation of select marker genes in PbTII cells prior to and after bifurcation.

Fig. S15. B cells are essential for Tfh responses in PbTII cells during *Pc*AS infection.

Fig. S16. Sorting strategy for myeloid cells.

Fig. S17. Principal Component Analysis of cDCs from naïve and infected mice.

Figure S18. Differential gene expression between single splenic CD8α+ and CD8α- cDCs.

Figure S19. Differentially expressed genes between single naïve and day 3 *Pc*AS-infected cDCs.

Figure S20. Principal Component Analysis of Ly6Chi monocytes from naïve and infected mice.

Figure S21. Differentially expressed genes between single Ly6Chi monocytes from naïve and day 3 *Pc*AS-infected mice.

Fig. S22. Expression of immune signalling genes by cDCs and monocytes.

Figure S23. Myeloid cell depletion in LysMCre x iDTR mice.

Tables **S1-S4**

Table S1 The expression data for all genes on day 7, the PCA loadings for PC1-PC10, and functional annotations for the genes (external file). Th1 annotations are based on studies by Hale *et al.* (SMARTA transgenic, day 6 of LCMV infection, $CXCR5^{-}Ly6c^{hi}$), Marshall *et al.* (SMARTA transgenic, day 8 of LCMV infection, $PSGL1^{hi}Ly6c^{hi}$), Stubbington *et al.* (*In vitro*, day 4) (*15, 46, 47*). Tfh annotations are based by studies by Hale *et al.* ($CXCR5+Ly6c^{lo}$, Marshall *et al.* ($PSGL1^{lo}Ly6c^{lo}$) and Liu *et al.* (Bcl6-RFP reporter, KLH immunization, $CXCR5^{+}Bcl6^{hi}$) (*38*). Th2 and Th17 annotations are based on Stubbington *et al*.. Annotations for genes associated with exhausted $CD4^{+}$ T cell phenotype are based on Crawford *et al*. (Day 30 of LCMV infection, genes upregulated in exhausted cells but not in memory cells) (*51*).

Table S2 TraCeR detection statistics for T cell receptor sequences in single-cell RNA-seq data from the first set of experiments, performed using the C1 platform (external file).

Table S3 TraCeR detection statistics for T cell receptor sequences in single-cell RNA-seq data from the second set of experiments, performed using the Smart-seq2 platform (external file).

Table S4 Annotation of receptors, cytokines and transcription factors.

Supplementary Computational Methods - The GPfates model (external file)

Materials and Methods:

Ethics and approval

All animal procedures were in accordance with the Animals (Scientific Procedures) Act 1986 and approved by the Animal Welfare and Ethical Review Body of the Wellcome Trust Genome Campus, or in accordance with Australian National Health and Medical Research Council guidelines and approved by the QIMR Berghofer Medical Research Institute Animal Ethics Committee (approval no. A02-633M).

Mice

C57BL/6 mice were purchased from Australian Resource Center (Canning Vale) or bred in-house. C57BL/6, PbTIIxCD45.1 and *LysMCre* x *iDTR* mice were maintained under specific pathogen-free conditions within animal facilities at the Wellcome Trust Genome Campus Research Support Facility (Cambridge, UK), registered with the UK Home Office, or at QIMR Berghofer Medical Research Institute (Brisbane, Australia). All mice were female and used at 8-12 weeks of age.

Adoptive transfer

Spleens from PbTIIxCD45.1 mice were aseptically removed and homogenised through a 100µm strainer before lysis of erythrocytes with RBC lysis buffer (eBioscience). $CD4^+$ T cells were enriched (purity >80%) using CD4 microbeads according to the manufacturer's instructions (Miltenyi Biotech) and stained with CellTrace™ Violet (Invitrogen) at 1µM in PBS for 15 minutes at 37°C in the dark. Violet CellTrace-labelled cells were resuspended in PBS and injected (10^6/200µl RPMI) via a lateral tail vein.

Infections

Plasmodium chabaudi chabaudi AS parasites were used after one *in vivo* passage in WT C57BL/6 mice. Mice were infected with 10^5 pRBCs i.v. and blood parasitemia was monitored by Giemsa-stained thin blood smears obtained from tail bleeds.

Flow cytometry and cell isolation

Single-cell suspensions were prepared by homogenising spleens through 100 µm strainers and lysing erythrocytes using RBC lysis buffer (eBioscience). Fc receptors were blocked using anti-CD16/32 antibody (BD Pharmigen or in-house). T cells were stained with the following antibodies (Biolegend): CD4-APC (GK1.5), TCRβ-APC-Cy7 (H57-597), CD45.1-FITC (A20), Vα2-FITC (B20.1), Vβ12-eFluor710 (MR11-1) (eBioscience), CD69-PE (H1.2F3), PD-1-APCCy7 (29F.1A12), CXCR5-Biotinylated (2G8) (BD Pharmigen), Streptavidin-PeCy7, CD183 (CXCR3)-PE (CXCR3-173), IFNγ-BV421 (XMG1.2), Bcl6-PercpCy5.5 (K112-91) (BD Pharmigen), Ki-67-PE (16A8), T-bet-eFluor660 (4-B10) (eBioscience) and TCF-1-PE (CD63D9) (Cell Signaling Technology). Dendritic cells and monocytes were stained with the

following antibodies (Biolegend): CD11c-Percp-Cy5.5 (N418), MHCII (1-A/1/E)-APC (M5/114-15.3), B220-AlexaFluor700 (RA3-6B2), TCRβ-APC-Cy7 (H57-597), Ly6C-FITC (HKJ.4), CD11b-BV421 (M1/70), Ly6G-PE (IA8), CD8-PE-Cy7 (53-6.7) and CXCL9-PE (MIG-2F5.5). Intracellular staining for IFNγ, T-bet, Bcl6, Ki-67, TCF-1 or CXCL9 was performed with the eBioscience FoxP3 intracellular kit. Intracellular staining for p-S6 was performing using a monoclonal antibody (D57.2.2E), or respective isotype control (Cell Signaling Technology) with Cell Signaling Buffer Set A (Miltenyi Biotech) according to manufacturer's protocol.
For DNA/RNA staining, Hoechst33342 (10µg/ml; Sigma) was added at 1/500 v/v to cell preparation 15 minutes prior to acquisition using a BD LSRFortessa IV (BD Bioscience).
Cells were sorted using a MoFlo XDP (Beckman Coulter), a FACSAria II (Becton Dickinson) or an Influx (Becton Dickinson) instrument. Activated PbTII T cells were sorted as $CD4^{+}TCR\beta^{+}$ and $CD69^{+}$ and/or divided at least once as measured using the CellTrace Violet proliferation dye. Dendritic cells were sorted as $CD11c^{hi}MHCII^{hi}TCR\beta^{-}B220^{-}$. Naive dendritic cells were further sorted as $CD8\alpha^{+}CD11b^{-}$ or $CD8\alpha^{-}CD11b^{+}$. Inflammatory monocytes were identified as $CD11b^{hi}Ly6C^{hi}Ly6G^{lo}TCR\beta^{-}B220^{-}$.

Single-cell mRNA sequencing

Single cell capture and processing with the Fluidigm C1 system was performed as described in (*52*). The cell suspension obtained from sorting was loaded onto the Fluidigm C1 platform using small–sized capture chips (5-10µm cells). 1 µl of a 1:4000 dilution of External RNA Control Consortium (ERCC) spike-ins (Ambion, Life Technologies) was included in the lysis buffer. Reverse transcription and pre-amplification of cDNA were performed using the SMARTer Ultra Low RNA kit (Clontech).
For processing with the Smart-seq2 protocol (*29*), the cells were sorted into 96-well plates containing lysis buffer using either a MoFlo XDP (Beckman Coulter) or an Influx (Becton Dickinson) instrument. The Smart-seq2 amplification was performed as described in (*29*), with the lysis buffer containing Triton-X, RNase inhibitor, dNTPs, dT30 primer and ERCC spike-ins (Ambion, Life Technologies, final dilution 1:100 million). The cDNA amplification step was performed with 24 cycles.
The sequencing libraries were prepared using Nextera XT DNA Sample Preparation Kit (Illumina), according to the protocol supplied by Fluidigm (PN 100-5950 B1). Libraries from up to 96 single cells were pooled and purified using AMPure XP beads (Beckman Coulter). Pooled samples were sequenced on an Illumina HiSeq 2500 instrument, using paired-end 100 or 125-base pair reads.

Processing and QC of scRNA-Seq data

Gene expressions were quantified from the paired end reads of the samples using Salmon (*41*), version 0.4.0. An example command for a one sample would be "salmon quant -i mouse_cdna_38.p3.78_repbase_ercc_index -l IU -p 4 -1 1771-026-195-H4_1.fastq -21771-026-195-H4_2.fastq -o1771-026-195-H4_salmon_out -g mouse_cdna38.78_repbase_ercc_index_gene_map.txt". The parameter libType=IU, and a transcriptome index built on Ensembl version 78 mouse cDNA sequences. We also had

sequences from the ERCC RNA spike-ins in the index, as well as 313 mouse specific repeat sequences from RepBase to potentially capture transcribed repeats.
For quality control of the single-cell data we assessed the number of input read pairs, and the amount of mitochondrial gene content. For all cells, we considered samples with less than 100,000 reads or more than 35% mitochondrial gene content as failed. For T cells, we additionally considered cells where number of genes was less than 100 + 0.008 * (mitochondrial gene content) as failed. For the data generated using a 96-well plate-based Smart-seq2 protocol, which does not permit visual inspection of the captured cells, we additionally excluded low-quality cells from which fewer than 2000 genes were detected, motivated by negative control wells. To verify that that the cells sorted in the wells were PbTII cells, we only selected cells from which both the transgenic TCR alpha and beta chains were detected (Supplementary Tables 2 and 3). Excluded cells were removed from all further analyses, and the remainder of the samples were taken as individual single cells.
For expression values, the Transcripts Per Millions (TPM's) estimated by Salmon included ERCC spike-ins. Thus, for analysis of the cells, we removed ERCC's from the expression table and scaled the TPM's so they again summed to a million. This way we get *endogenous* TPM values, representing the relative abundance of a given gene *within a cell*. We also globally removed genes from analysis where less than three cells expressed the gene at minimum 1 TPM, unless stated otherwise.

Latent Variable Modelling of data

We modelled the data using an unsupervised Bayesian Gaussian Process Latent Variable model (BGPLVM) (*14*) on log10 transformed TPM values (with a scaling factor 1 added). The BGPLVM was run with 5 latent variables. As we used an ARD (Automatic Relevance Determination) squared exponential covariance function, we could infer that two latent factors explained the data. All other parameters to the BayesianGPLVM model in GPy (version 1.0.9) were left as default. Upon inspection we noticed a circular pattern. This corresponds to a 1-dimensional topology, which requires two dimensions for a faithful representation. Thus we inferred a new latent variable by 1-dimensional BGPLVM, with priors on the latent variable based on the cell collection times (see the Computational Supplement), where we used the 2D latent variables as input. This way we inferred smoothed "pseudotime" values for the every data point representing the progressive response to the malaria infection.
In the 2-dimensional model of the data, we searched for genes that highly correlated with either of the two explanatory latent factors. Performing functional enrichment analysis using gProfiler (*42*) on the top genes revealed that factor 1 (which explained most of the variation) was related to cell proliferation. The second factor was largely explained by genes involved with immune response. Upon inspection, it seemed as the second latent factor terminated in two groups of cells. We investigated this in terms of a bifurcating time series.

Bifurcating time series analysis

To study the cells in terms of a bifurcating time series, we implemented an Overlapping Mixture of Gaussian Processes (OMGP) model (*16*), see the Computational Supplement for details. The

model uses an optimization procedure to associate observed data with a given number of individual independent trends over a time variable. The model was run with pseudotime as input, and the immune response related latent variable as output. For the mixture model, we assumed two trends. The two trends were given squared exponential covariance functions, where we fixed the length scale to 1 based on our prior assumptions on smoothness over pseudotime. We also constrained the model variance to 0.05, which allows trends to share observed data points. Remaining hyperparameters were optimized by gradient descent. (See the Computational Supplement for details)

Testing genes for bifurcated expression

The output of the OMGP model is a soft assignment to each of two trends for every observed cell. The original model was fitted with the 2nd latent variable from the latent variable analysis. To find genes that significantly drive this bifurcation, we keep all parameters fixed but change the data to be individual genes expression levels and calculate the data likelihood. In order to get a null distribution to assess significance, we performed the same analysis but with randomly permuted pseudotime-values. This is described in detail in the Computational Supplement.

To measure in which *direction* a gene is involved with the bifurcation, we used correlation between expression and trend assignment. For example, a gene's expression being strongly positively correlated with a trends assignment means it is being upregulated on that bifurcated branch.

Monocle

The Monocle analysis was performed with version 1.2.0 of the Monocle package, following the steps outlined in the original vignette (*17*). In brief, the analysis was performed using the size-normalized data (TPM) including all genes expressed in ≥10 cells (11439 genes) with default parameters. The genes used for the ordering of cells were defined by carrying out a differential expression analysis of the time points using the differentialGeneTest embedded in the package. Following the original vignette, genes with q-value <0.01 were selected (7738 genes). The num_paths option was set as 2.

SCUBA

(https://github.com/gcyuan/SCUBA/tree/2ffa4fe5842dfe88db0207c82088bce0e5b97be7) was run using 3003 genes and provided information about time point. RNAseq_preprocess.m and SCUBA scripts were run according to instructions. SCUBA did not find any bifurcation points. Similarly, using 1000 most informative genes (SCUBA default), or scaling of the data (to aid the sensitivity), did not result in any bifurcation either. Changing the number of was done by the variable ngene_select in RNAseq_preprocess.m. All other variables were kept at default.

Annotation of cell-surface receptors, cytokines and transcription factors

Genes likely to encode transcription factors, cell-surface receptors or cytokines were found by combining information from KEGG (http://www.genome.jp/kegg/), the Gene Ontology Consortium (http://geneontology.org/, PANTHER (http://www.pantherdb.org/) along with the more specific databases detailed below.

Transcription factors were found by searching the Gene Ontology Consortium database using the following ontology term: *GO:0003700 (sequence-specific DNA binding transcription factor activity)*; KEGG for *ko03000 (Transcription Factors)*; PANTHER for *PC00009 (DNA binding) AND PC00218 (Transcription Factors)*. The presence of genes in the following databases was also used as evidence for transcription factor activity: AnimalTFDB (http://www.bioguo.org/AnimalTFDB/index.php), DBD (http://www.transcriptionfactor.org), TFCat (http://www.tfcat.ca), TFClass (http://tfclass.bioinf.med.uni-goettingen.de/tfclass), UniProbe (http://the_brain.bwh.harvard.edu/uniprobe) and TFcheckpoint (http://www.tfcheckpoint.org).

Cell-surface receptors were found by searching the Gene Ontolotgy Consortium database using the following ontology terms *GO:0004888 (transmembrane signaling receptor activity) OR GO:0008305 (integrin complex)) AND NOT (GO:0004984 (olfactory recep- tor activity) OR GO:0008527 (taste receptor activity)*; KEGG for *ko04030 (G-Protein Coupled Receptors) OR ko04050 (Cytokine Receptors) OR ko01020 (Enzyme-linked Receptors)*; PANTHER for *PC00021 (G-Protein Coupled Receptors) OR PC00084 (Cy- tokine Receptors) OR PC00194 (Enzyme-linked Receptors)*. Annotation of genes as receptors in the ImmPort (https://immport.niaid.nih.gov/), GPCRDB (http://gpcrdb.org/) or IUPHAR (http://www.guidetopharmacology.org/) databases was also used as evidence for receptor functionality.

Cytokines were found by searching the Gene Ontolotgy Consortium database using the following ontology terms *GO:0005125 (cytokine activity)*; KEGG for *ko04052 (Cytokines)*; PANTHER for *PC00083 (Cytokines)*. Annotation of genes as cytokines in ImmPort was also used in this case.

Genes were scored according to the number of databases and search results in which they occurred. Scores were weighted according to the strength of evidence provided by each database such that functional annotations supported by manually reviewed experimental evidence were given a higher score than those that were solely computationally generated (Table)

Annotation source	Score
KEGG	3
Gene Ontology Consortium evidence codes IDA, IPI, IMP, IGI, IEP, ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA, TAS, IC	5
Gene Ontology Consortium evidence codes IEA, NAS, ND	1
PANTHER	2
AnimalTFDB	4
DBD	1
TFCat, classed as ‘transcription factor’	7

TFCat, classed as 'candidate'	5
TFClass	4
UniProbe	7
TFcheckpoint (if manually reviewed)	6
ImmPort	4
GPRCDB	2
IUPHAR	7

Genes were assigned as likely cell-surface receptors or cytokines if they had a cumulative score greater than or equal to 5 in that category. Genes were assigned as likely transcription factors if they had a cumulative score greater than or equal to 6 in that category.

In vivo cell depletion

Cellular depletion in *LysMCre* x *iDTR* mice was performed by intraperitoneal injection of 10ng/g DT (Sigma-Aldrich) in 200µl 0.9% saline (Baxter) at day 3 post-infection. Control mice were given 0.9% saline only.

For B cell depletion, anti-CD20 (Genentech) or isotype control antibody was administered in a single 0.25mg dose via i.p. injection in 200ul 0.9% NaCl (Baxter), 7 days prior to infection.

Confocal microscopy

Confocal microscopy was performed on 10–20 µm frozen spleen sections. . Briefly, splenic tissues were snap frozen in embedding optimal cutting temperature (OCT) medium (Sakura) and stored at -80^{o}C until use. Sections were fixed in ice-cold acetone for 10 minutes prior to labelling with antibodies against B220-PE (clone-RA3-6B2) as well as CD68-Alexa Fluor 594 (clone-FA-11) or SIGN-R1-Alexa Fluor 647 (clone ER-TR9). Antibodies against CD68 were obtained from Biolegend (San Diego, CA), and against SIGN-R1 from BIO-RAD (USA). DAPI was used to aid visualization of white pulp areas. Samples were imaged on a Zeiss 780-NLO laser-scanning confocal microscope (Carl Zeiss Microimaging) and data analysed using Imaris image analysis software, version 8.1.2 (Bitplane). Cells were identified using the spots function in Imaris, with thresholds <10mM and intensities <150. All objects were manually inspected for accuracy before data were plotted and analyzed in GraphPad prism (version 6).

Figures S1-S23

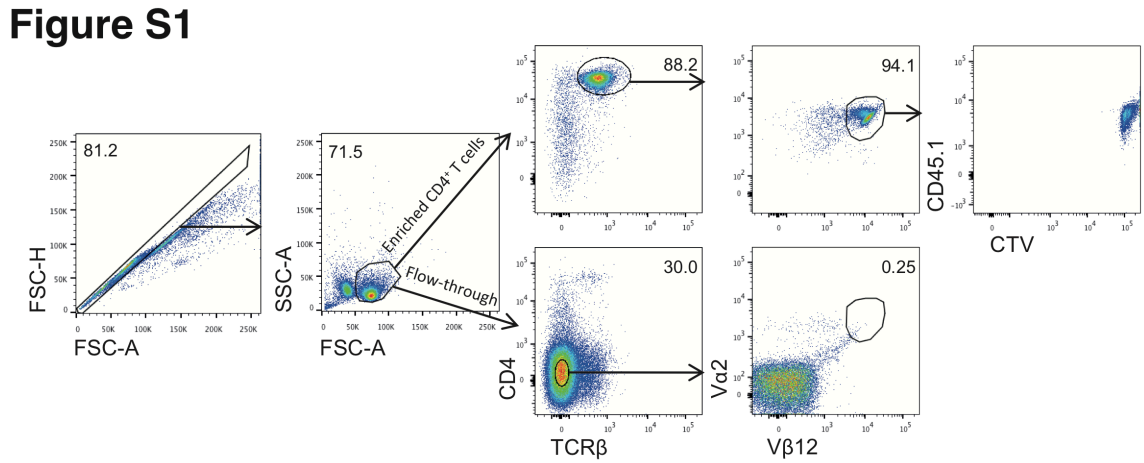


Fig. S1. Enrichment of PbTII cells for adoptive transfer.

(A) $CD4^+$ T cells were enriched using positive selection (MACS microbeads) from the spleen of a naive, PbTII x CD45.1 mouse. FACS plots show purity, expression of Vα2 and Vβ12 transgenes, and CellTrace™ Violet (CTV) staining of enriched PbTII compared to corresponding flow-through from the enrichment process.

A

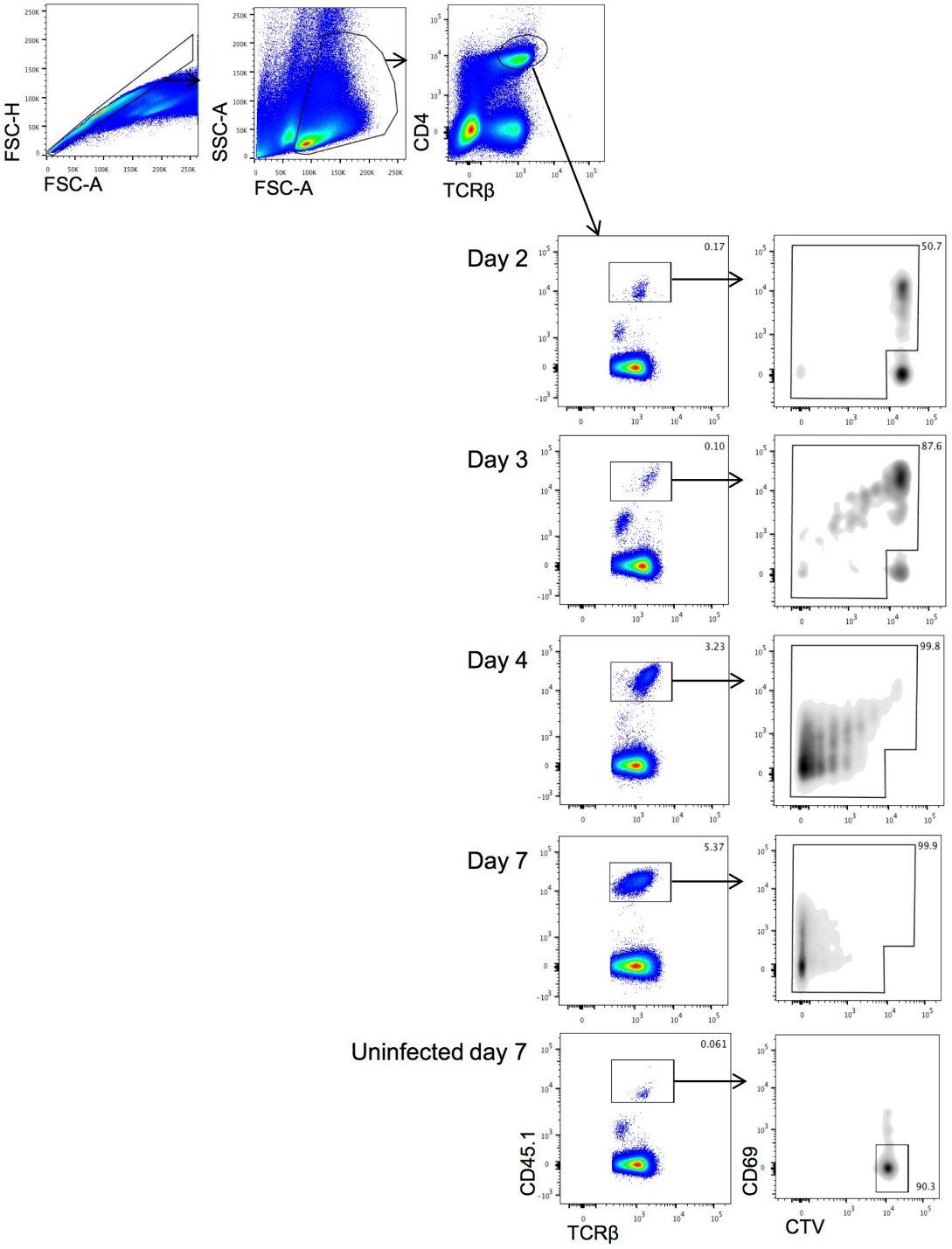


Fig S2

Fig. S2. Sorting strategy for PbTII cells.

(A) PbTII cells ($CD4^+$ $TCR\beta^+$ $CD45.1^+$) were adoptively transferred into WT congenic ($CD45.2^+$) recipient mice At indicated days, early activated ($CD69^+$) and/or proliferated (CTV^{lo}) PbTII cells were cell-sorted from spleens of *Plasmodium chabaudi chabaudi* AS infected mice, and naïve PbTII cells ($CD69^{lo}CTV^{hi}$) were cell-sorted from the spleens of naïve mice at day 7 post-transfer.

Figure S3

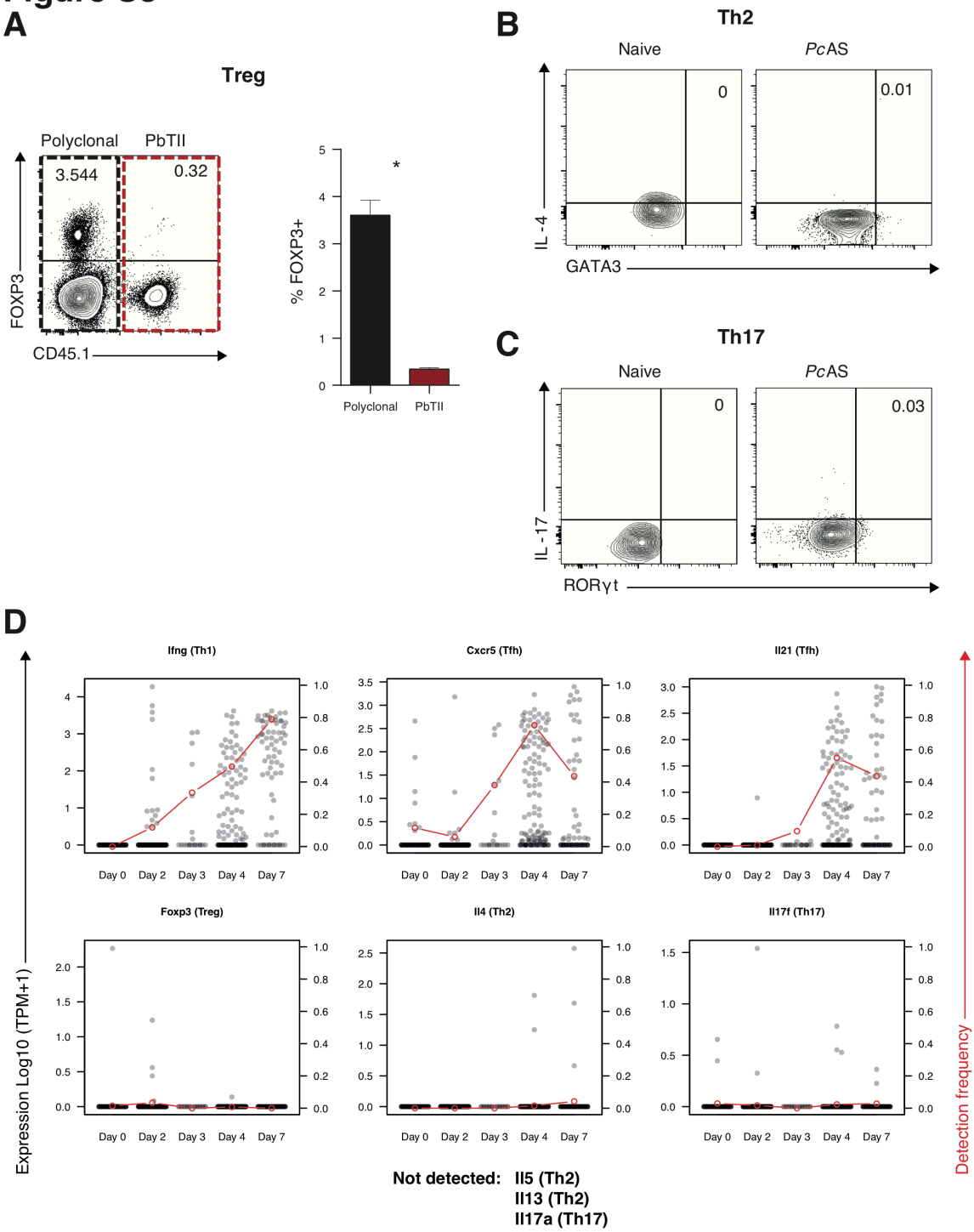


Fig. S3. Expression of subset-specific marker genes in PbTII cells.

(A) Representative FACS plot (gated on $CD4^+$ $TCR\beta^+$ live singlets) and proportion of $FOXP3^+$ (Treg) splenic PbTII (10^4 transferred) ($CD45.1^+$; red dashed box) or polyclonal $CD4^+$ T ($CD45.1^-$; black dashed box) cells from mice (n=6) at day 7 post-infection.

(B-C) Representative FACS plots (gated on $CD45.1^+$ $CD4^+$ $TCR\beta^+$ live singlets) of (B) IL-4^+GATA3^+ (Th2) and (C) IL-$17^+ROR\gamma t^+$ (Th17) splenic PbTII cells in naive (receiving 10^6 cells, n=3) or *Pc*AS-infected mice (receiving 10^4 cells, n=6) at day 7 post-infection. (A-C) Data are representative of two independent experiments. Statistics: Mann-Whitney U test; *$p<0.05$.

(D) The mRNA expression of selected subset-specific cytokines and the Treg hallmark transcription factor *Foxp3* in PbTII cells. The red dots and line indicate the fraction of cells in each time point where the particular mRNA was detected.

Figure S4

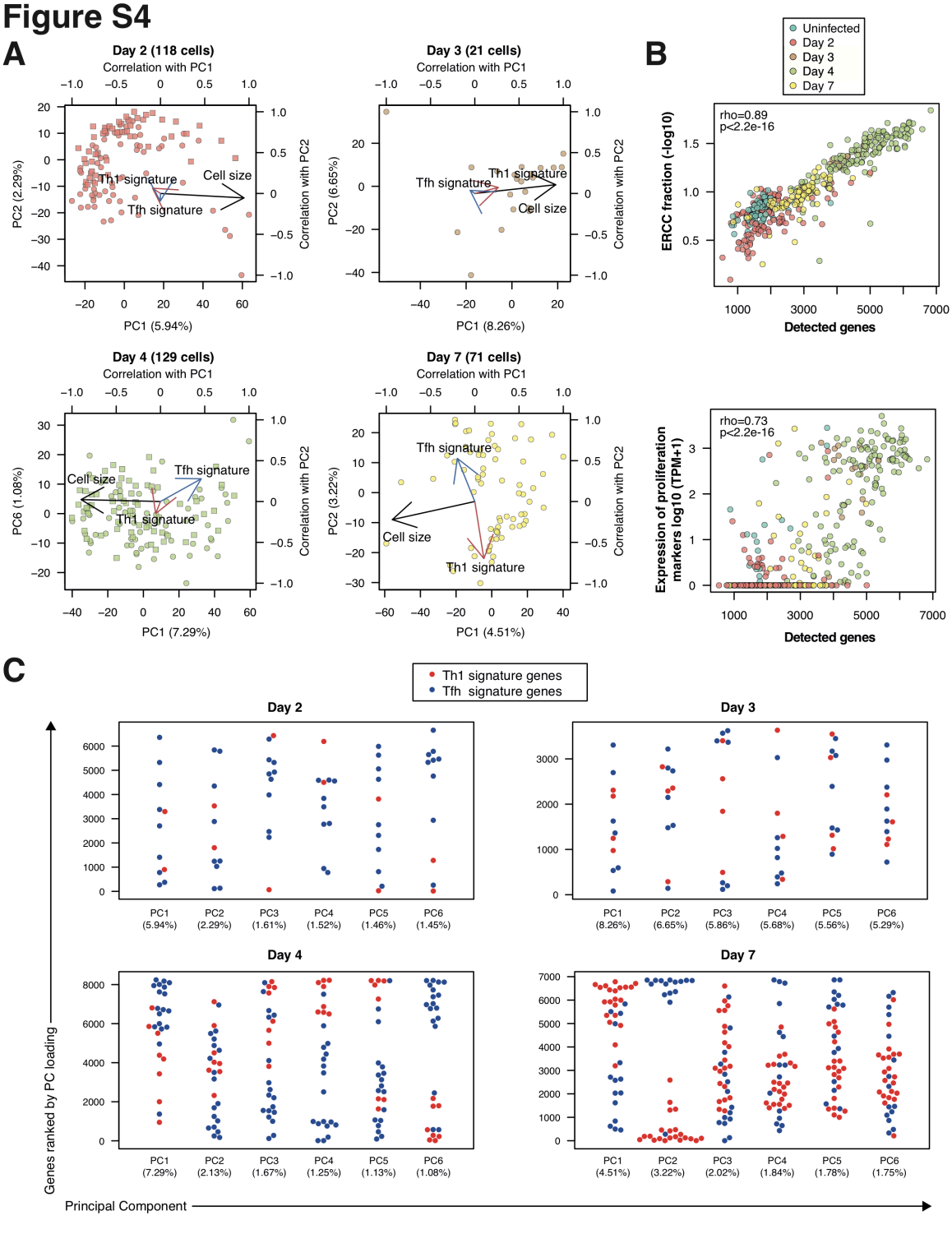


Fig. S4. Heterogeneity of activated PbTII cells.

(A) PCA of single PbTII cells at 2, 3, 4 and 7 days post-infection with *Pc*AS. The PCA was based on all genes expressed at ≥100 TPM in at least 2 cells. The arrows represent the Pearson correlation with PC1 and PC2. Cell size refers to the number of detected genes. "Th1 signature" and "Tfh signature" refer to cumulative expression of top 30 signature genes associated with Th1 and Tfh phenotypes (*15*). The numbers in parenthesis show proportional contribution of respective PC.

(B) The relationship of detected cell number with the fraction or reads mapping to ERCC spike-in RNA (top) and with cumulative expression of proliferation markers *Mki67, Mybl2, Bub1, Plk1, Ccne1, Ccnd1* and *Ccnb1* (*21*) (Figure 4B and S9).

(C) Ranked loading scores for PC1-PC6 of the Th1 and Tfh signature genes in the PCA shown in (A). The numbers in parenthesis show proportional contribution of respective PC.

Figure S5

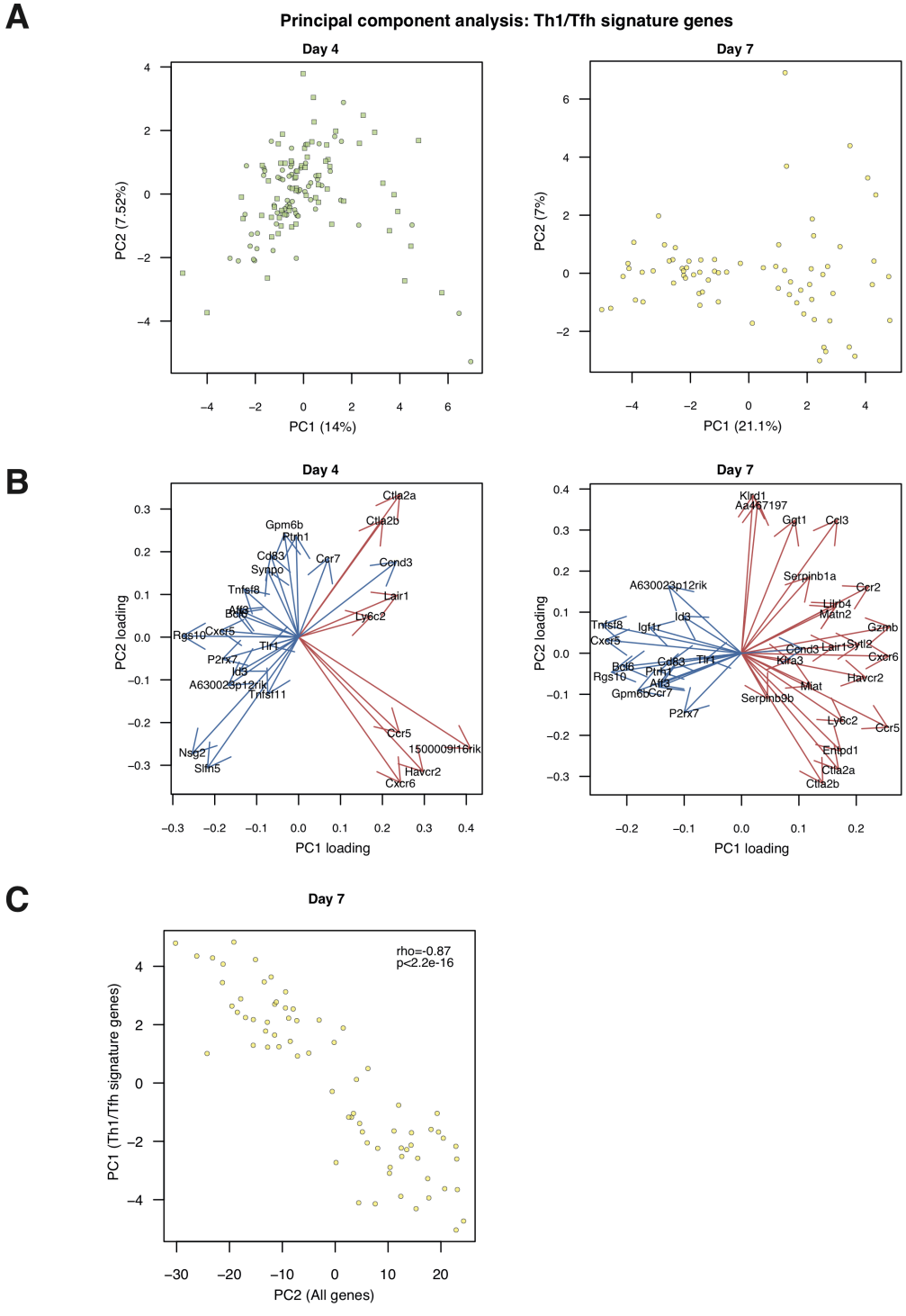


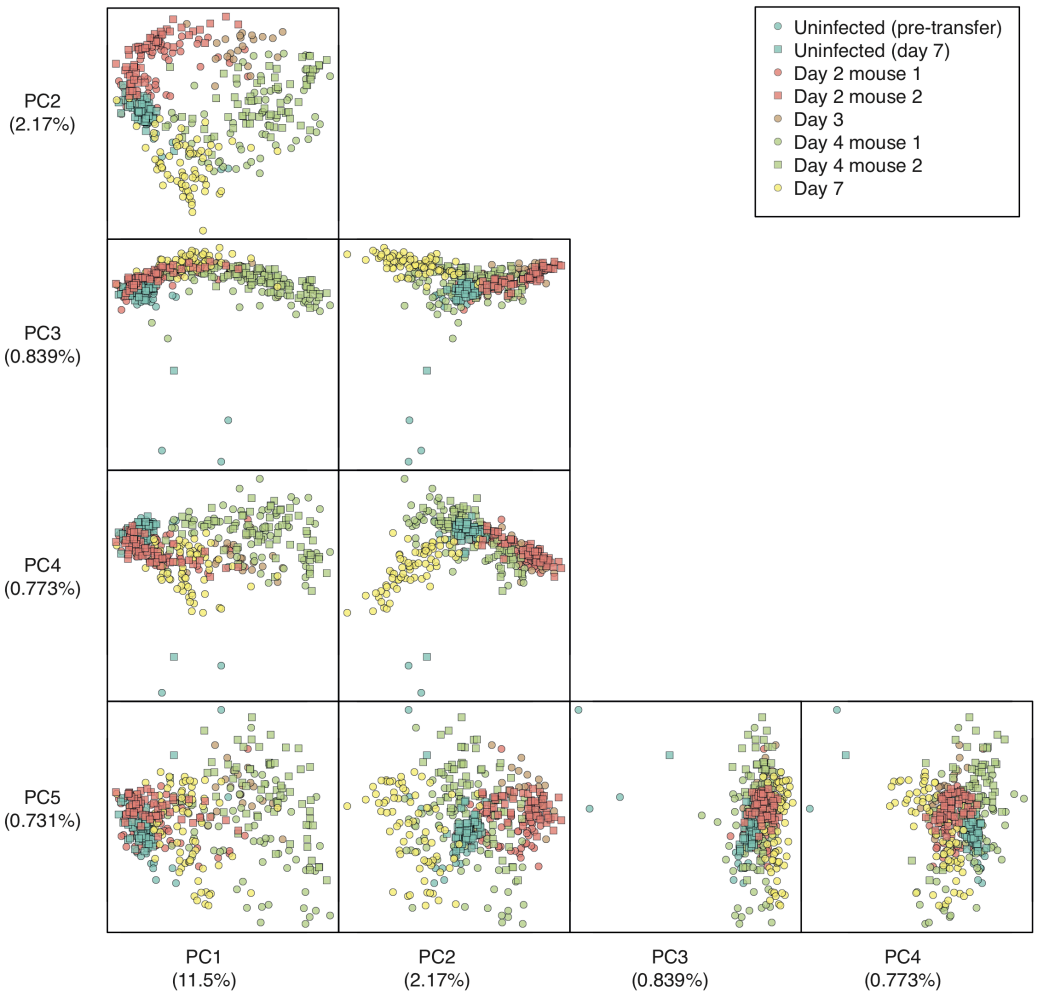
Fig. S5. Heterogeneity of Th1/Tfh signature gene expression in activated PbTII cells.

(**A**) Principal component analysis of day 4 (left) and day 7 (right) PbTII cells using Th1/Tfh signature genes (*15*) detected at the level ≥ 100 TPM in at least 2 cells. The numbers in parenthesis show proportional contribution of respective PC.

(**B**) The PC1 and PC2 loadings of individual Th1 (red) and Tfh (blue) signature genes in PCA of day 4 and day 7 PbTII cells (A). PC, Principal Component

(**C**) The correlation of PC1 from the analysis with the signature genes alone and PC2 of the genome-wide analysis.

Figure S6



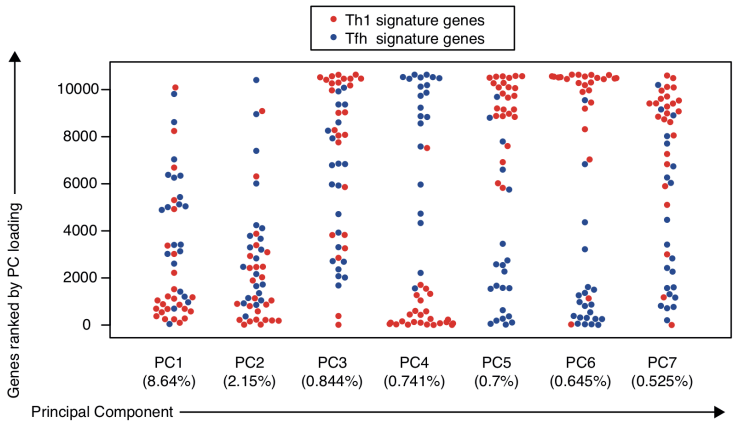


Fig. S6. The contribution of Th1 and Tfh signature genes to the overall heterogeneity of the PbTII time series.

(A) The first five components of the Principal Component Analysis of the entire time series. The numbers in parenthesis show proportional contribution of respective PC.

(B) The rankings of the Th1 and Tfh signature genes among the loadings of Principal Components 1-7.

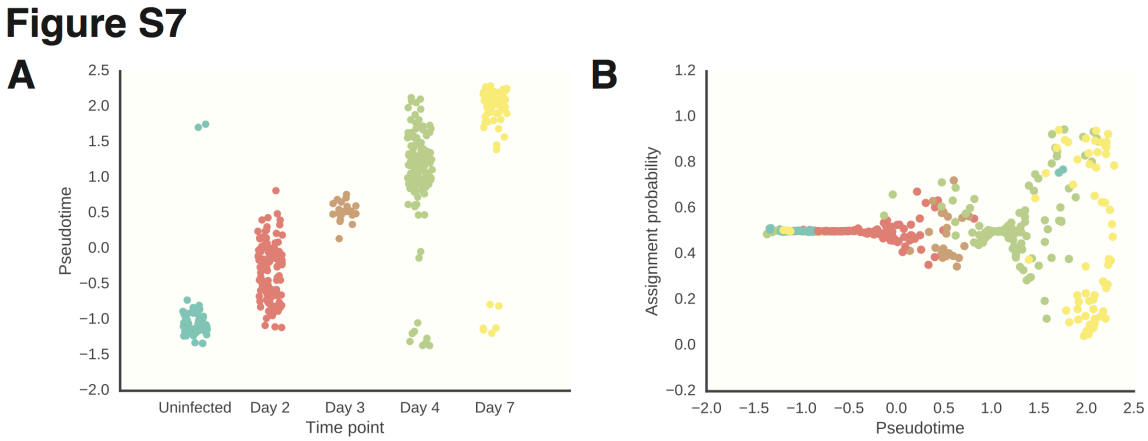


Fig. S7. The relationship of pseudotime with time points **(A)** and with the Th1 assignment probability **(B)**.

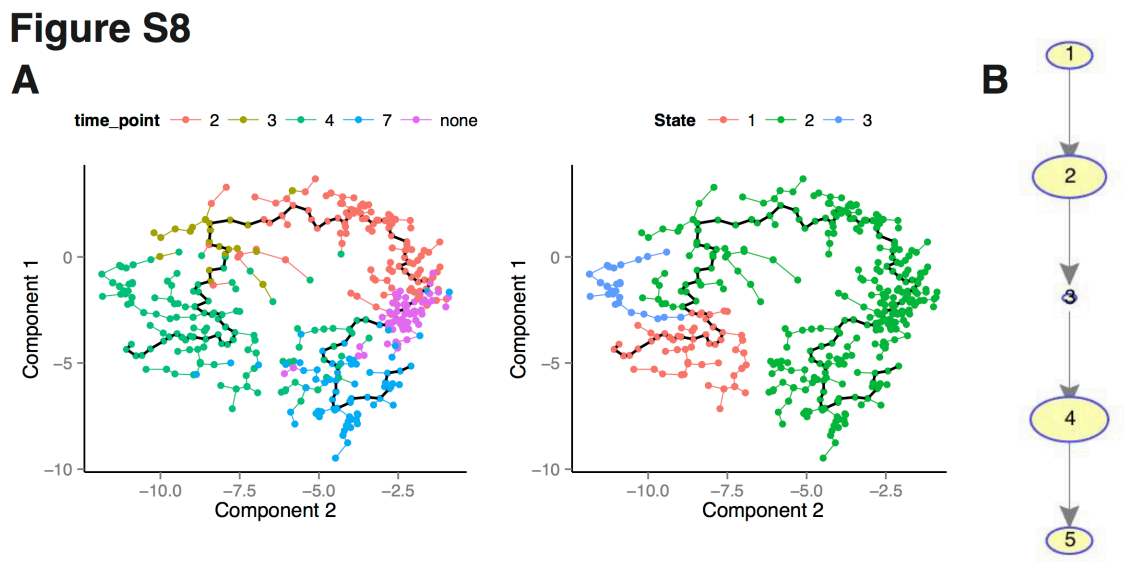
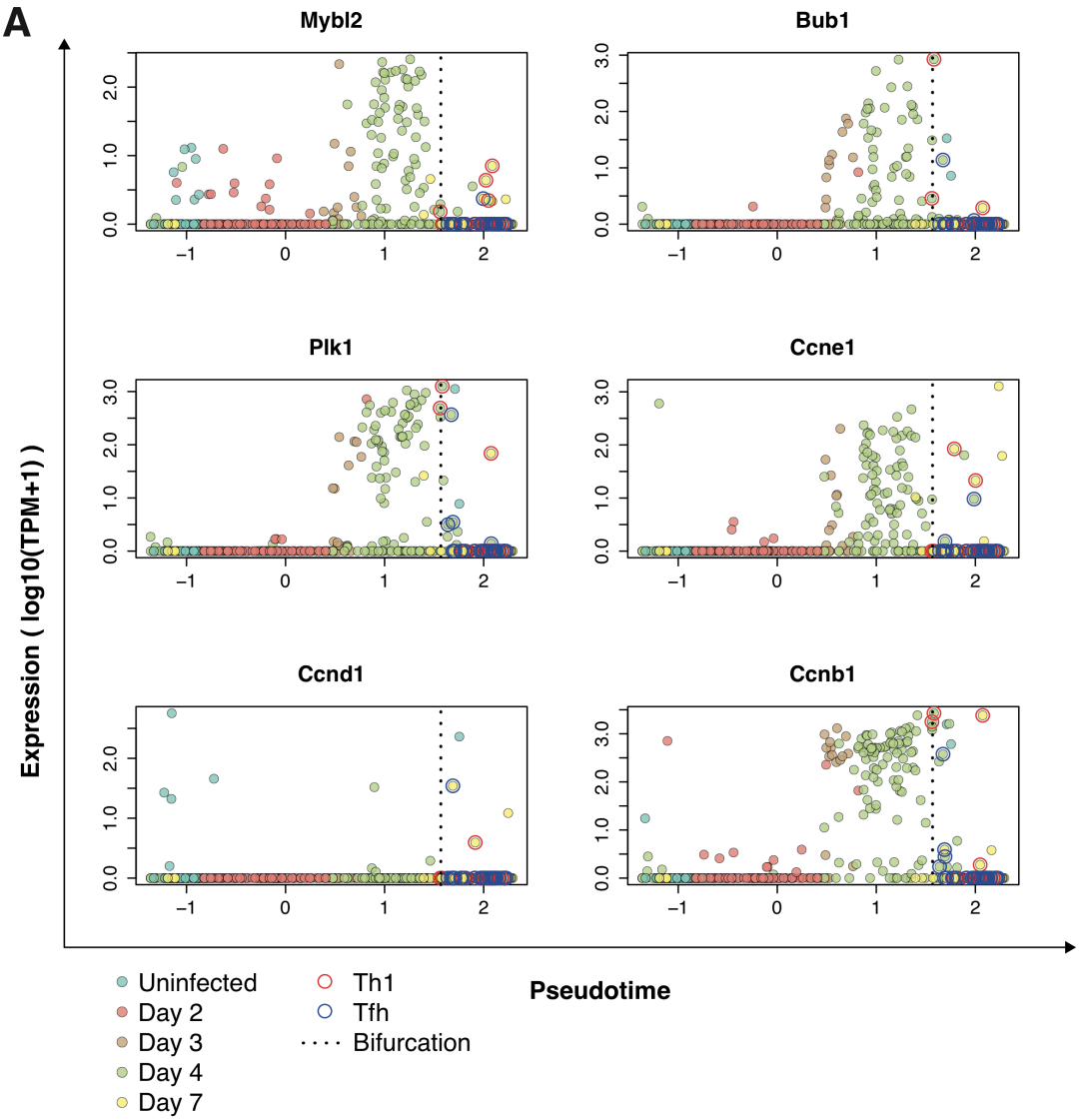


Fig. S8. Modelling the data using Monocle and SCUBA.

(A) Monocle model of the data, coloured by time points (left) and cell states identified by Monocle (right).

(B) SCUBA bifurcation analysis failed to yield any bifurcating points. Sizes of bubbles are according to number of cells.

Figure S9



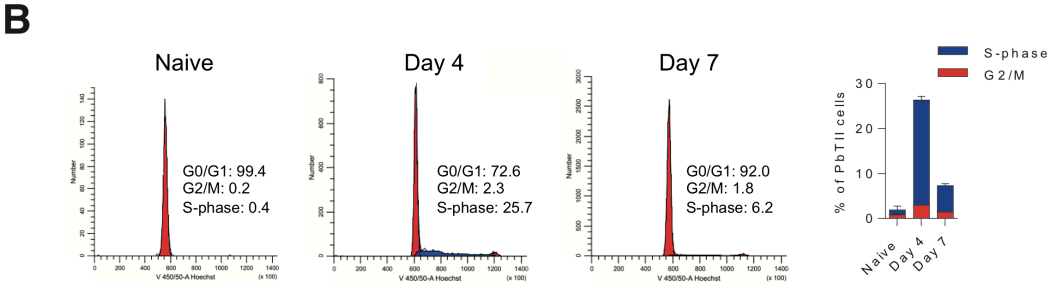


Fig.

Fig. S9. Proliferative burst of activated PbTII cells.

(A) The expression of established proliferation genes (*21*) along pseudotime.

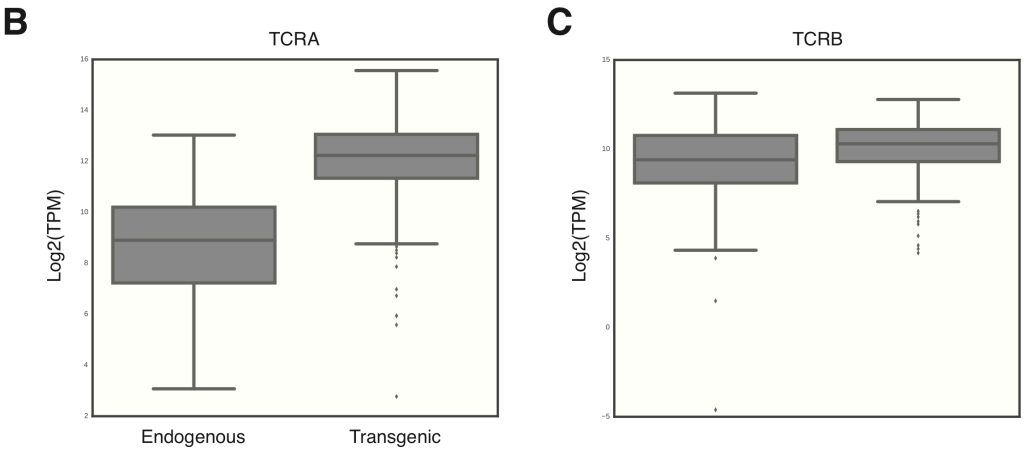
(B) ModFit plots and proportions of PbTII cells in G0/G1, G2/M and S-phase of cell cycle as determined by Hoechst staining.

Figure S10

A

TCRB

TCRA	Transgenic	Endogenous	Not detected
Transgenic	287	15	76
Endogenous	0	2	1
Not detected	4	5	18



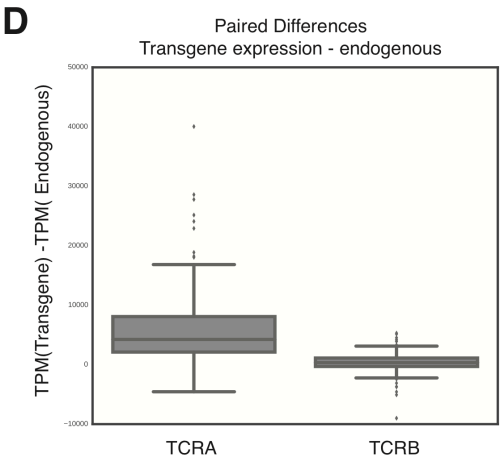


Fig. S10. Expression of transgenic and endogenic TCRs.

(**A**) Statistics of TCR sequence detection. Numbers correspond to single cells in which the corresponding transcript was detected.
(**B**) Expression levels (log2(TPM)) of for the endogenous or transgenic TCRA chains across the entire dataset.

Figure S11

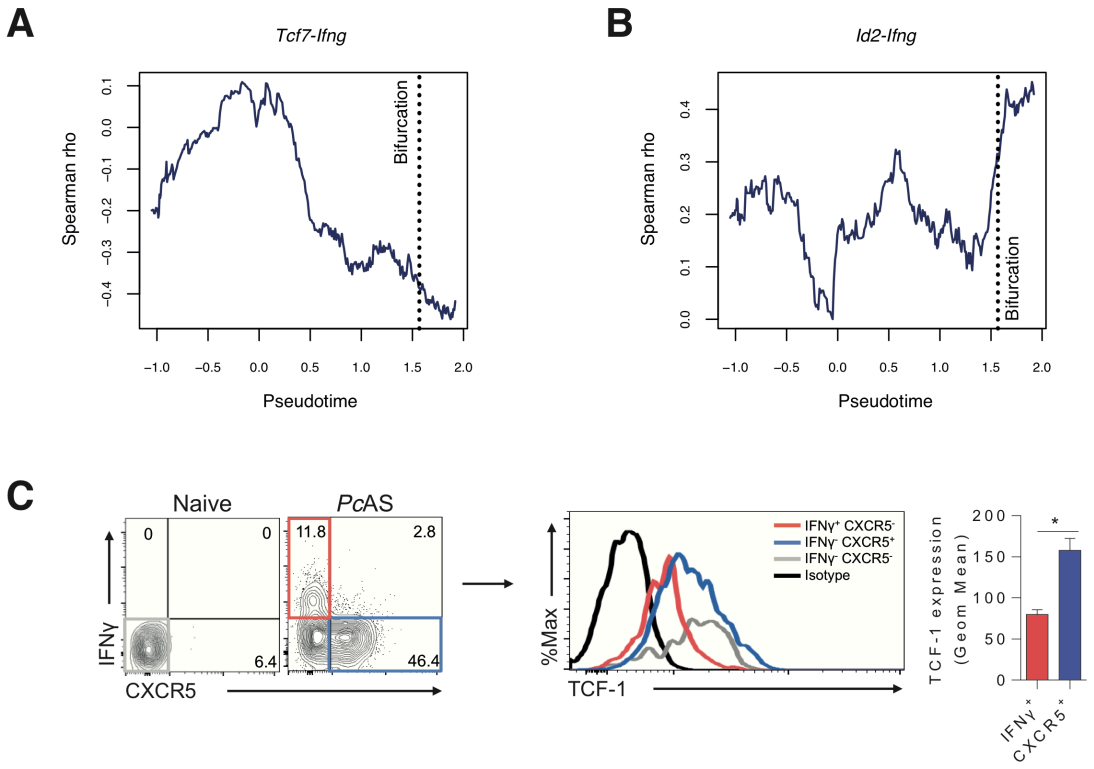


Fig. S11. Correlation of expression of *Ifng* with *Tcf7 and Id2* across pseudotime.

(A-B) The correlation of the expression *Ifng* with *Tcf7* (A) and with *Id2* (B) at single-cell level. Using a rolling window method, Spearman rho was calculated in windows of 100 cells. The pseudotime values are mean values within each window.

(C) Representative FACS plots showing TCF-1 (gene product of *Tcf7*) expression in CXCR5+ (blue gate) and IFNγ+ (red gate) PbTIIs, compared to naïve PbTIIs (gray) (isotype control shown in black in FACS histogram) at 7 days post-infection. Summary graph shows mean & standard deviations for geometric mean fluorescence intensity of TCF-1 expression in gated PbTII populations (n=4 mice) Statistics: Mann-Whitney U test *p<0.05.

Figure S12

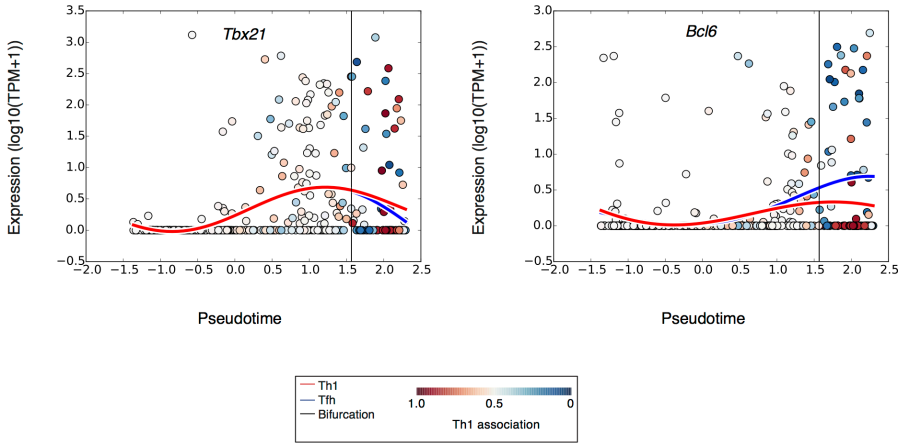


Fig. S12. The expression of *Tbx21* (left) and *Bcl6* (right) across pseudotime.

The curves represent the Th1 (red) and Tfh (blue) trends when weighing the information from data points according to trend assignment. The color of the data points represents the strength of the relationship with the Th1 trend.

Figure S13

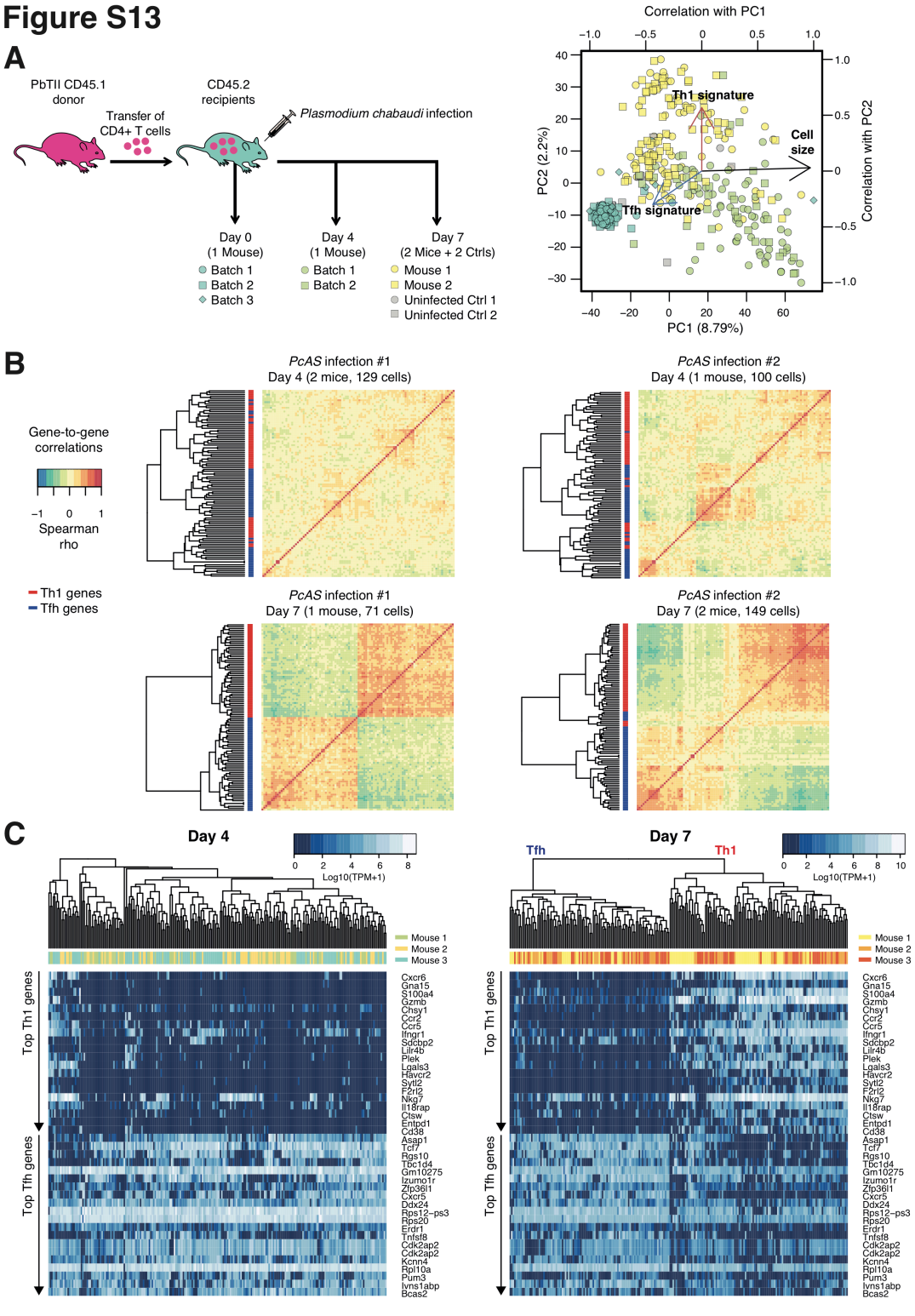


Fig. S13. Robustness of top bifurcating genes across experiments.

(A) Principal Component Analysis of the single cells from the replicate *Pc*AS infection. The single cells were sorted on 96-well plates and cDNA was amplified using the Smart-seq2 protocol (*29*). The arrows represent the Pearson correlation with PC1 and PC2. Cell size refers to the number of detected genes. "Th1 signature" and "Tfh signature" refer to cumulative expression of genes associated with Th1 or Tfh phenotypes (*15*). PC, Principal Component.

(B) The emergence of subset-specific gene patterns at day 7 of infection. For the top bifurcating genes (Fig S5C) pairwise gene-to-gene Spearman correlations were calculated. The rowside colours represent the association of the gene with either Th1 fate (red) or Tfh fate (blue).

(C) The expression of top 20 Th1 and Tfh associated genes identified using GPfates in single PbTII cells at days 4 and 7. *Cdk2ap2* appears twice because two alternative genomic annotations exist.

Figure S14

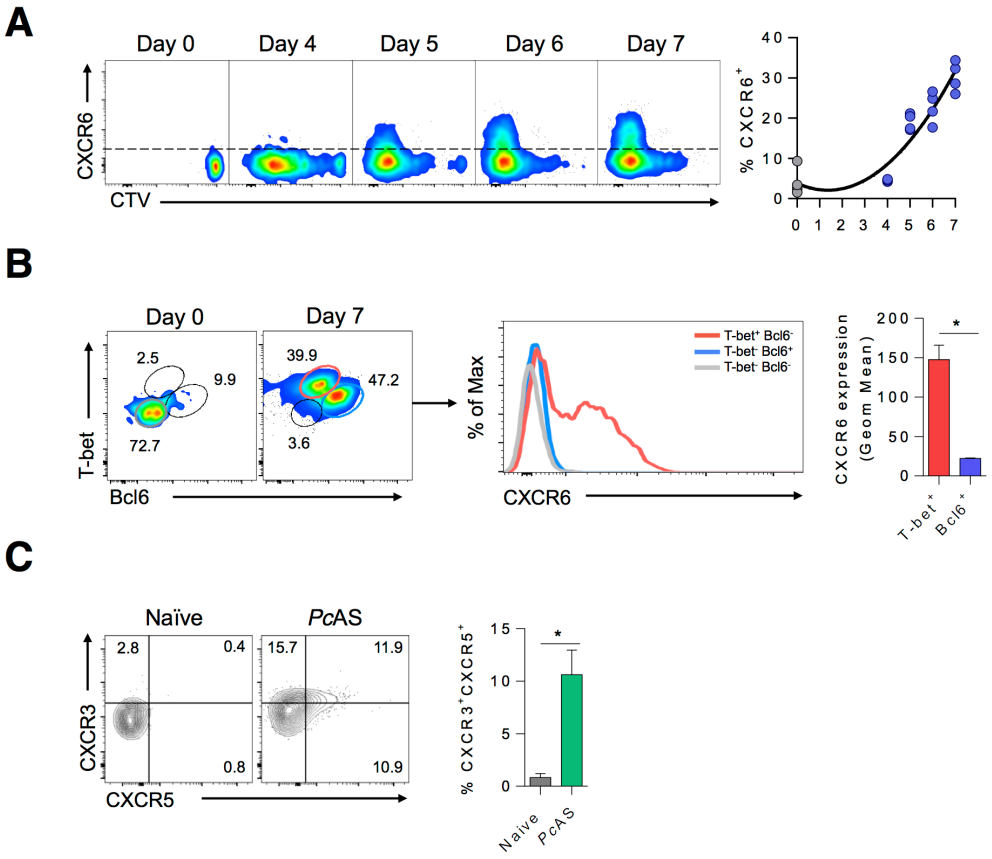


Fig. S14. Flow cytometric validation of select marker genes in PbTII cells prior to and after bifurcation.

(A) Representative FACS plots showing kinetics of CellTraceTM Violet (CTV) dilution and CXCR6 expression, with summary graphs showing % of PbTII cells expressing this (after 106 PbTII cells transferred) in un-infected (Day 0) and *PcAS*-infected mice at indicated days post-infection (n=4 mice/timepoint, with individual mouse data shown in summary graphs; solid line in summary graphs indicates results from third order polynominal regression analysis.) Data are representative of two independent experiments.

(B) Representative FACS plots showing CXCR6 expression in Tbethi (red gate) and Bcl6hi (blue gate) PbTII cells, compared to naïve PbTIIs (grey) at 7 days post-infection. Summary graph shows mean & standard deviations for geometric mean fluorescence intensity of CXCR6 expression in gated PbTII populations (n=4 mice) Statistics: Mann-Whitney U test *p<0.05.

(C) Representative FACS plots and proportions of splenic PbTII cells co-expressing CXCR5 and CXCR3 in naive (gray; n=3) or infected mice (green; n=6) at 4 days post-infection with *P.chabaudi chabaudi* AS (*Pc*AS). Results are representative of two independent experiments. Statistics: Mann-Whitney U test *p<0.05.

Figure S15

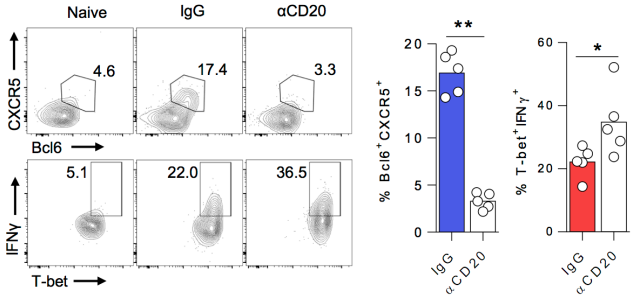


Fig. S15. B cells are essential for Tfh responses in PbTII cells during *Pc*AS infection.

Representative FACS plots (gated on CD4+ TCRβ+ CD45.1+ live singlets) of splenic PbTII cells, showing proportions exhibiting Tfh (Bcl6+ CXCR5+) and Th1 (Tbet+ IFNγ+) phenotypes in WT mice (receiving 104 PbTII cells), treated with anti-CD20 monoclonal antibodies (0.25mg) to deplete B-cells, or control IgG, and infected for 7 days with *Pc*AS. Individual mice data (n=5) shown in summary graph. Mann-Whitney U test *p<0.05; **p<0.01. Results are representative of two independent experiments.

Figure S16

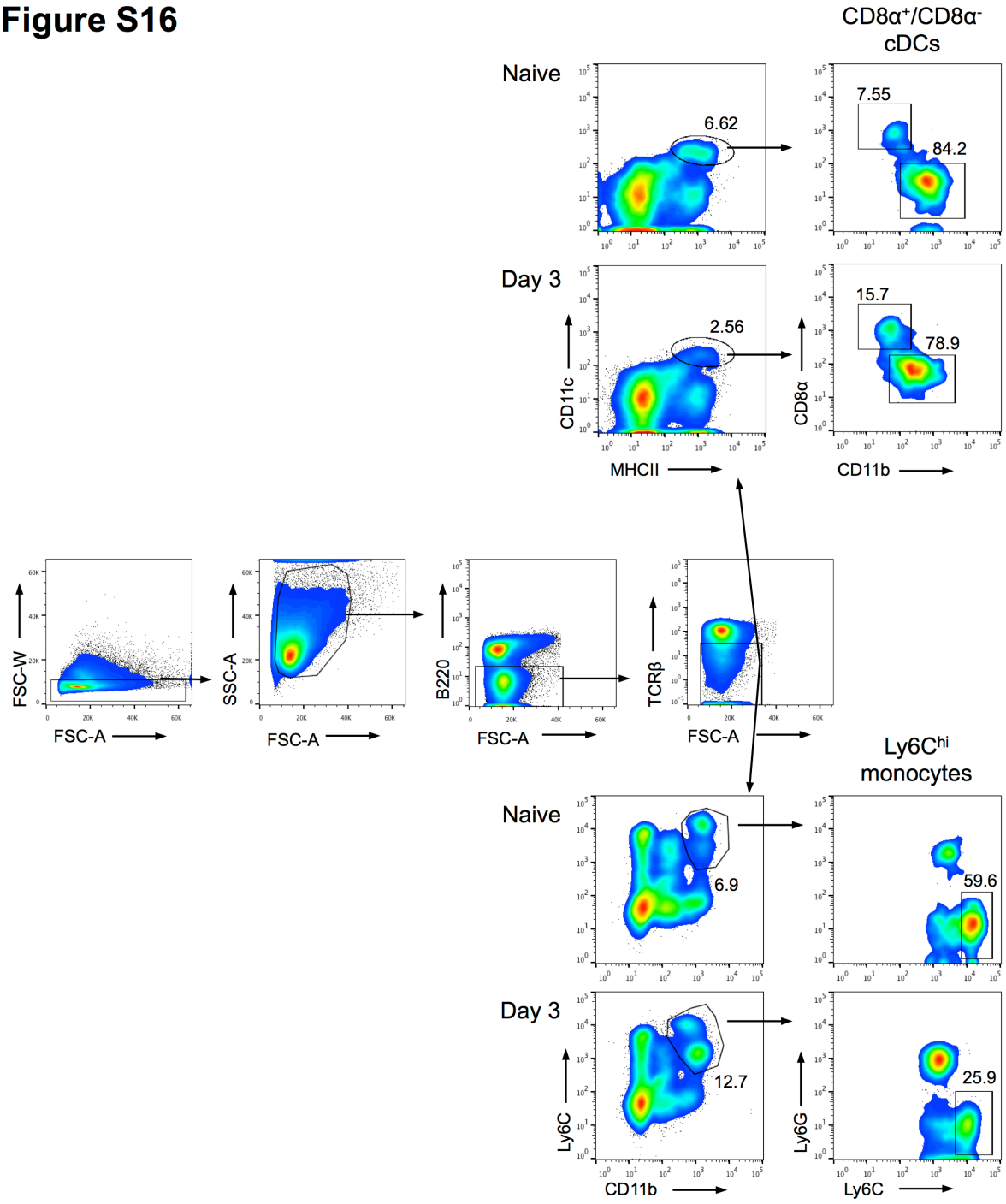


Fig. S16. Sorting strategy for myeloid cells.

Representative FACS plots showing sorting strategy for CD8α+ and CD11b+ cDC, and Ly6Chi inflammatory monocytes from the spleens of naive and 3-day *Pc*AS-infected mice.

Figure S17

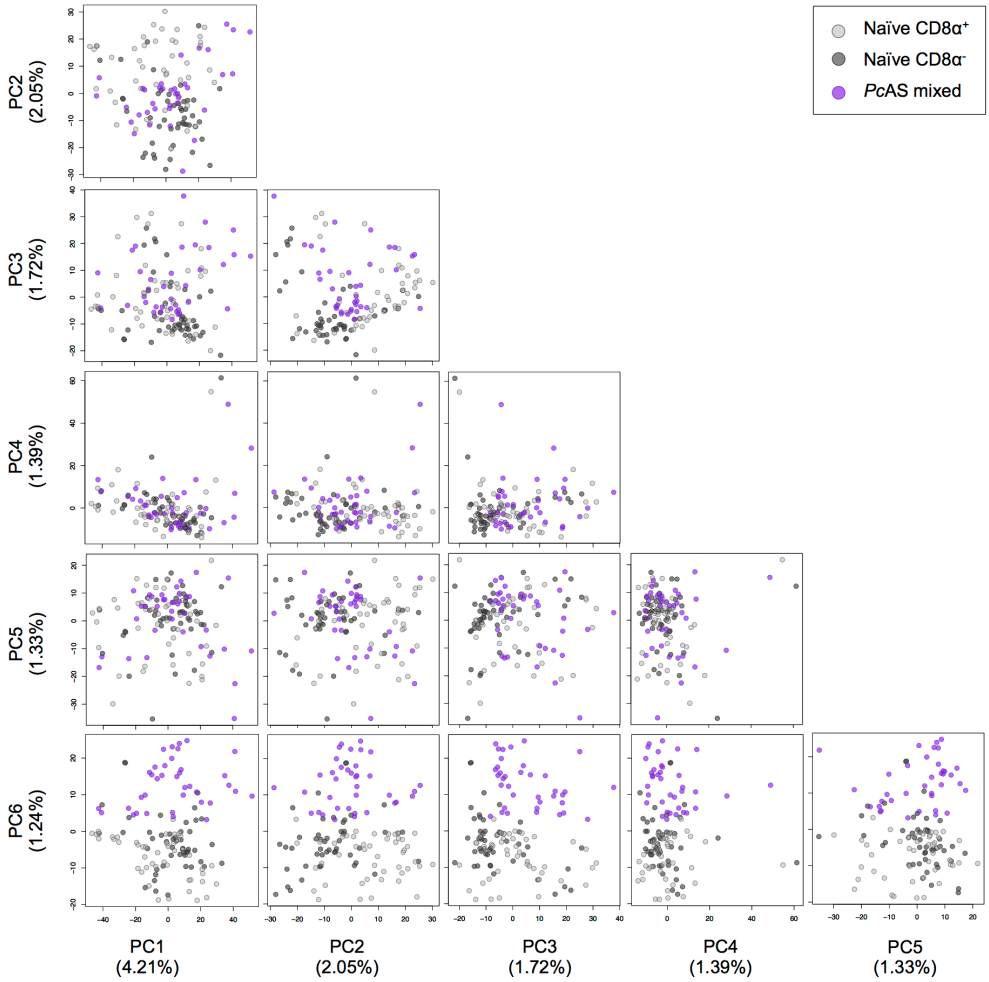
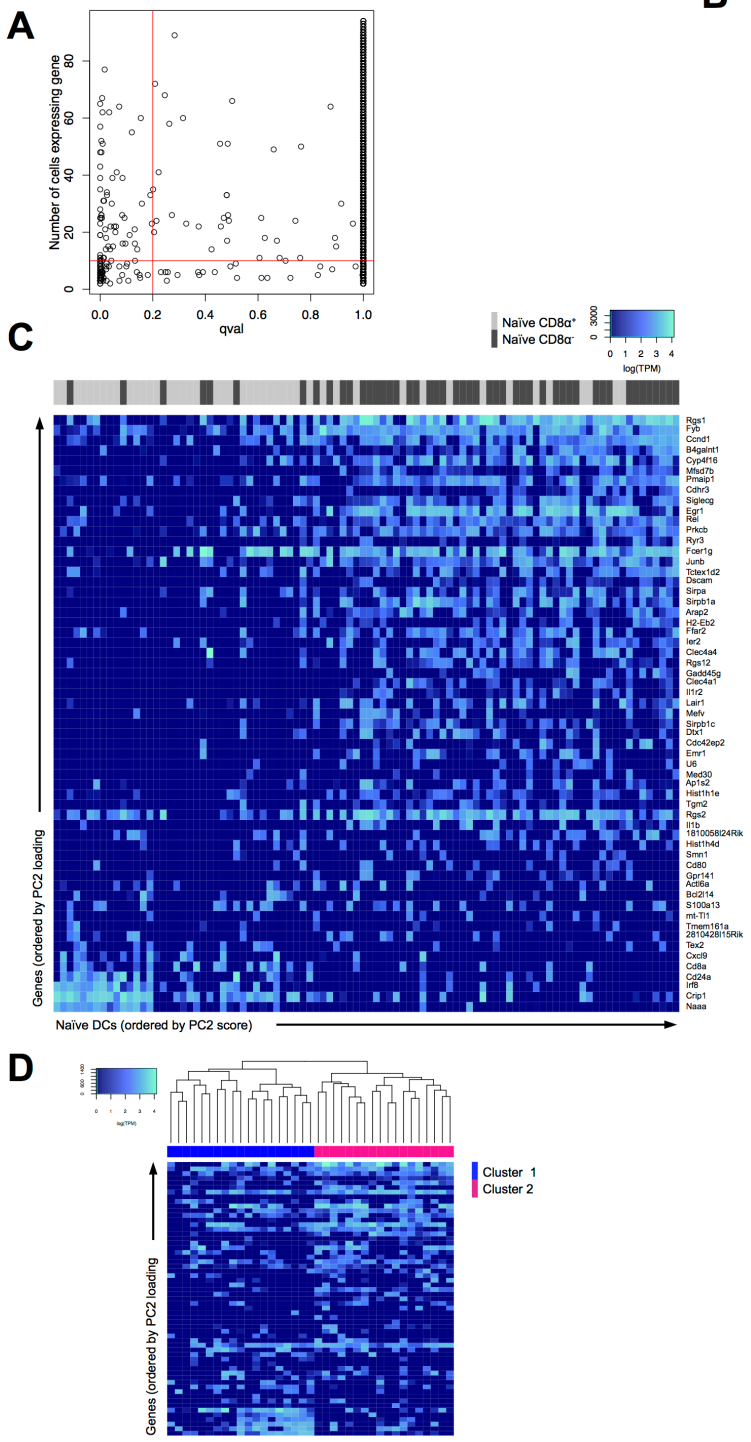


Fig. S17. Principal Component Analysis of cDCs from naïve and infected mice.

Results of Principal Component (PC) Analysis on scRNAseq mRNA reads (filtered by minimum expression of 100 TPM in at least 2 cells) from 131 single splenic naïve CD8α+ and CD8α- and mixed day 3 *Pc*AS-infected cDC. PC1-PC6 shown. Axis labels show proportional contribution of respective PC.

Figure S18



B

Gene	qval	Fold change Log2(CD8α-/CD8α+)
Sirpa	6.48E-08	2.69E+00
U6	1.33E-07	1.44E+00
Rgs1	2.07E-07	1.93E+00
Cd8a	1.57E-06	-1.39E+01
Hist1h4d	1.85E-06	2.59E+00
Tmem161a	2.45E-06	-3.40E+00
Med30	6.52E-06	5.14E+00
S100a13	2.67E-05	8.50E-02
Sirpb1a	2.83E-05	2.52E+00
Arap2	1.26E-04	1.40E+00
Irf8	1.33E-04	-3.82E+00
Siglecg	1.35E-04	2.69E+00
Actl6a	2.22E-04	1.77E+00
Junb	3.45E-04	3.92E+00
Dscam	2.10E-03	4.77E+00
Cyp4f16	4.32E-03	1.89E+00
Ap1s2	4.37E-03	2.83E+00
Crip1	4.61E-03	-1.89E+00
1810058I24Rik	5.19E-03	1.02E+00
Dtx1	6.65E-03	4.77E+00
Rgs2	7.53E-03	2.39E+00
Ccnd1	8.80E-03	2.21E+00
Egr1	9.33E-03	1.50E+00
mt-Tl1	1.19E-02	2.48E+00
Mfsd7b	1.23E-02	3.30E+00
Smn1	1.40E-02	1.12E+00
Lair1	1.46E-02	1.73E+00
Fyb	1.69E-02	8.13E-01
Cdhr3	1.92E-02	4.25E+00
Cd80	2.14E-02	1.27E+00
Ryr3	2.25E-02	1.02E+01
Hist1h1e	2.54E-02	2.22E+00
Rgs12	2.59E-02	1.36E+00
Tex2	3.08E-02	-2.52E+00
Prkcb	3.43E-02	8.13E-01
Gadd45g	3.81E-02	6.36E+00
Emr1	3.93E-02	2.72E+00
Clec4a4	4.40E-02	2.68E+00
B4galnt1	4.65E-02	2.38E+00
Cxcl9	4.87E-02	-7.54E+00
Il1r2	5.30E-02	3.38E+00
Mefv	5.77E-02	4.60E+00
H2-Eb2	6.02E-02	5.62E+00
Ffar2	6.27E-02	1.90E+00
Fcer1g	7.25E-02	9.84E-01
Cd24a	8.37E-02	7.83E-02
Tctex1d2	8.51E-02	2.11E+00
2810428I15Rik	8.54E-02	2.85E-01
Naaa	9.25E-02	-1.50E+00
Gpr141	9.61E-02	6.36E+00
Il1b	1.12E-01	6.61E+00
Rel	1.20E-01	1.09E+00
Tgm2	1.31E-01	2.26E+00
Cdc42ep2	1.32E-01	5.82E+00
Bcl2l14	1.41E-01	-1.34E+01
Pmaip1	1.54E-01	1.64E+00
Sirpb1c	1.59E-01	4.78E-01
Ier2	1.91E-01	3.26E+00
Clec4a1	1.96E-01	3.66E+00

Figure S18. Differential gene expression between single splenic CD8α+ and CD8α- cDCs.

(A) Results of differential gene expression analysis between naïve splenic CD8α+ and CD8α- cDCs, for all genes expressed in greater than 2 cells.

(B) Complete list of differentially-expressed genes between naïve CD8α+ and CD8α- cDCs, which were expressed in >10 cells of either subset with a qval <0.2 as determined in (A).

(C) Heatmap of naïve cDCs ordered by PC2 (Fig. 6A) and expression of genes from (B) ordered by PC2 loading in (Fig 6A).

(D) Heatmap examining hierarchical clustering of mixed CD8α+ and CD8α- CD11b+ day 3-infected cDCs (cell-sorted and mixed at a ratio of 50:50 prior to scRNAseq) using differentially expressed genes from (B) ordered by PC2 loading shown in (Fig 6A).

Figure S19

Gene	Log2 (Fold change)	qval
Tgtp1	4.36E+00	4.84E-11
Tgtp2	2.78E+00	1.68E-10
Ifi47	3.56E+00	4.23E-10
Kdm6b	4.25E+00	1.22E-09
Actb	1.62E+00	1.01E-07
Igtp	4.00E+00	1.43E-07
AC124762.1	3.06E+00	3.69E-07
Gm15427	-9.78E-01	1.51E-06
Stat1	2.49E+00	1.92E-06
U6	-3.31E+00	8.27E-06
Snora31	9.27E-01	9.92E-06
Nlrc5	2.09E+00	1.43E-05
Gm12250	5.43E+00	1.61E-05
Zbp1	4.87E+00	5.59E-05
Gbp4	3.41E+00	1.82E-04
R3hdm4	2.19E+00	2.52E-04
Slc39a1	3.20E+00	7.09E-04
Gm10800	6.42E-01	8.35E-04
Cxcl10	3.43E+00	1.81E-03
Alkbh5	3.20E+00	1.98E-03
Cxcl9	2.42E+00	2.38E-03
Dtx3l	1.74E+00	4.69E-03
Wtap	2.85E+00	4.87E-03
AC131780.3	1.11E+00	5.07E-03
Gbp3	1.40E+00	5.89E-03
Wac	2.87E+00	8.31E-03
Pml	2.96E+00	1.35E-02
Arf4	1.89E+00	1.56E-02
Irf1	1.44E+00	2.27E-02
Gbp2	2.12E+00	4.55E-02

Figure S19. Differentially expressed genes between single naïve and day 3 *Pc*AS-infected cDCs. List of differentially expressed genes, expressed in >10 cells (qval<0.05) between naïve and day 3-infected cDCs. Mean TPM fold-change in gene expression relative to naïve levels.

Figure S20

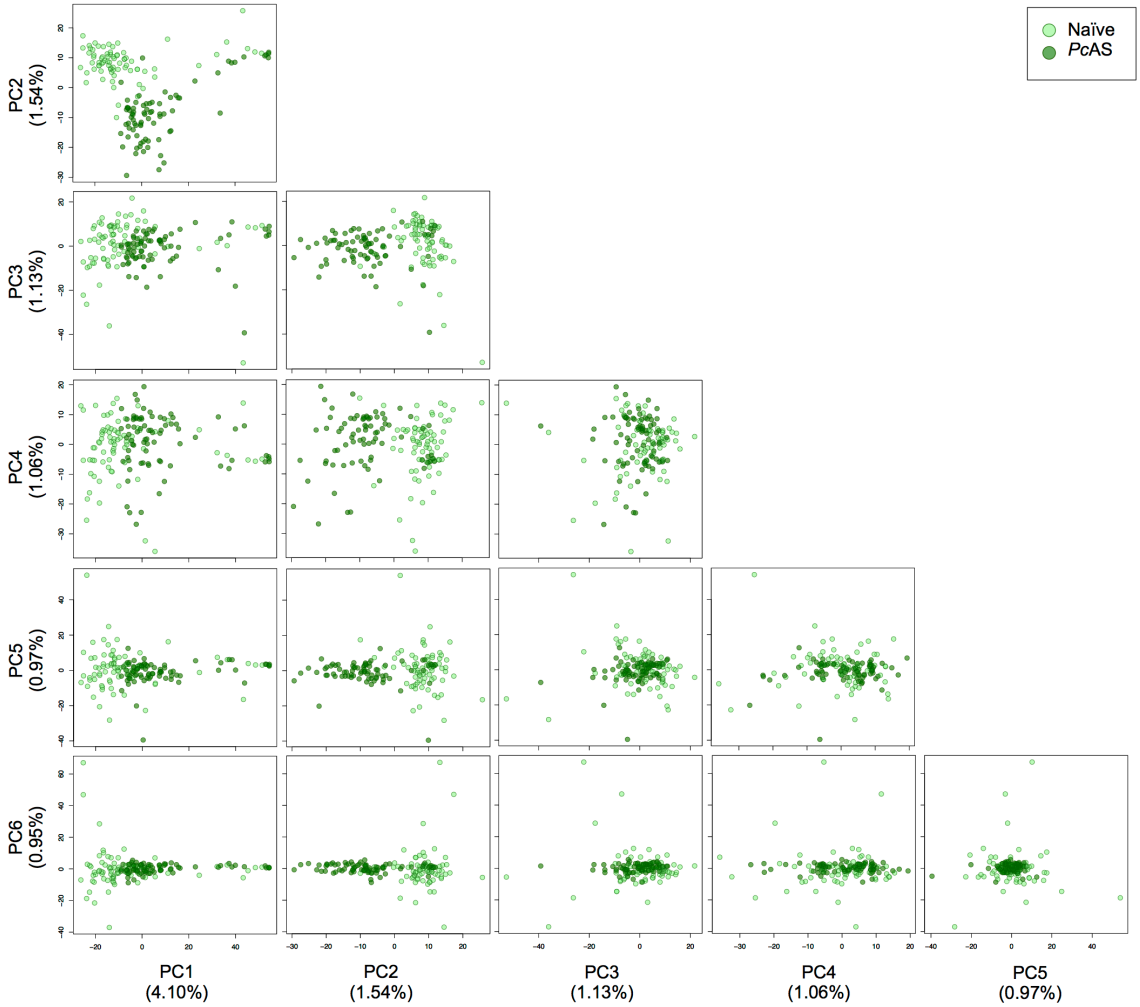


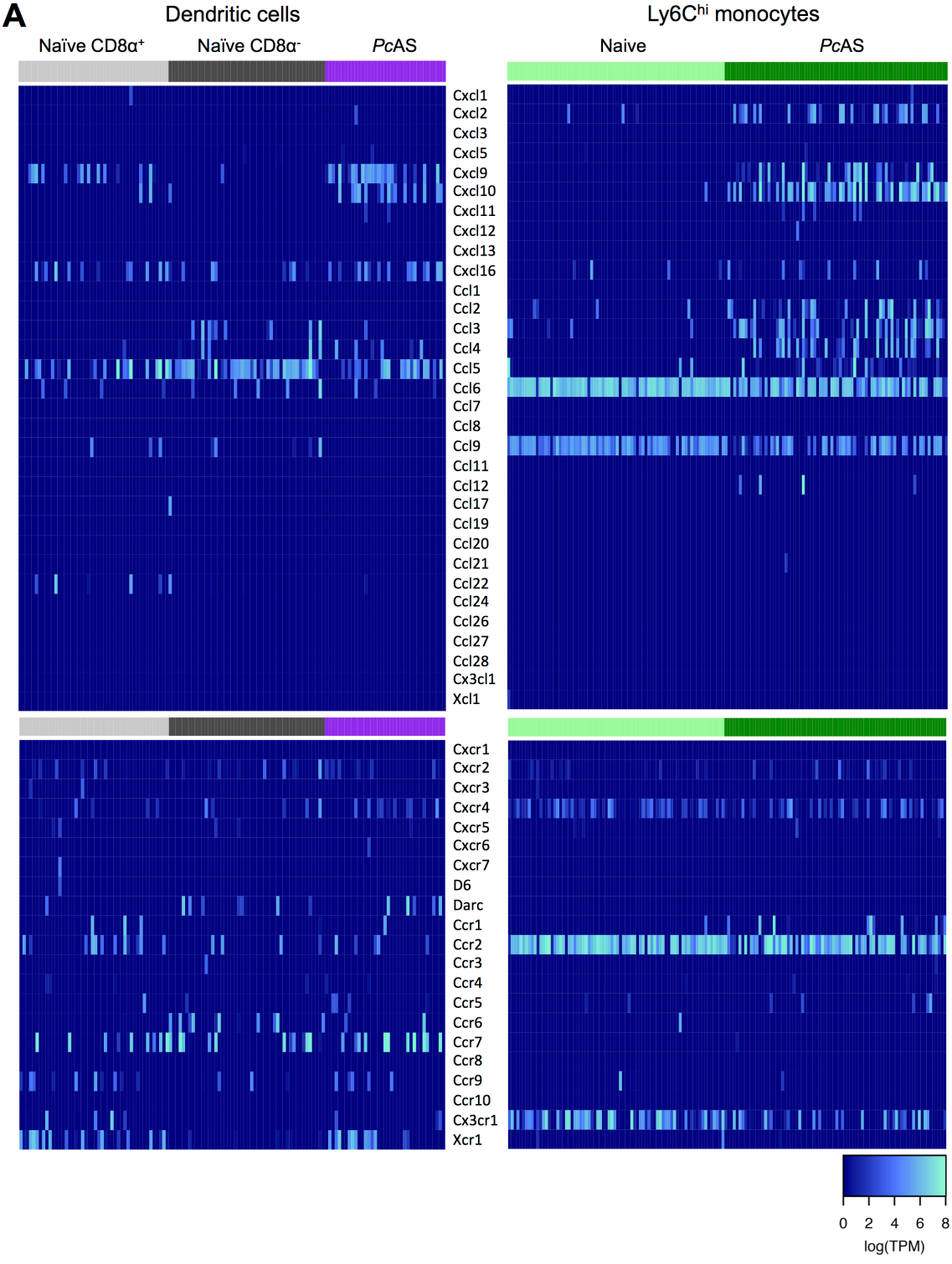
Figure S20. Principal Component Analysis of Ly6Chi monocytes from naïve and infected mice. Results of Principal Component (PC) Analysis using scRNAseq mRNA reads (filtered by minimum expression of 100 TPM in at least 2 cells) of 154 single splenic Ly6Chi monocytes from naïve and infected mice. PC1-PC6 shown. Axis labels show proportional contribution of respective PC.

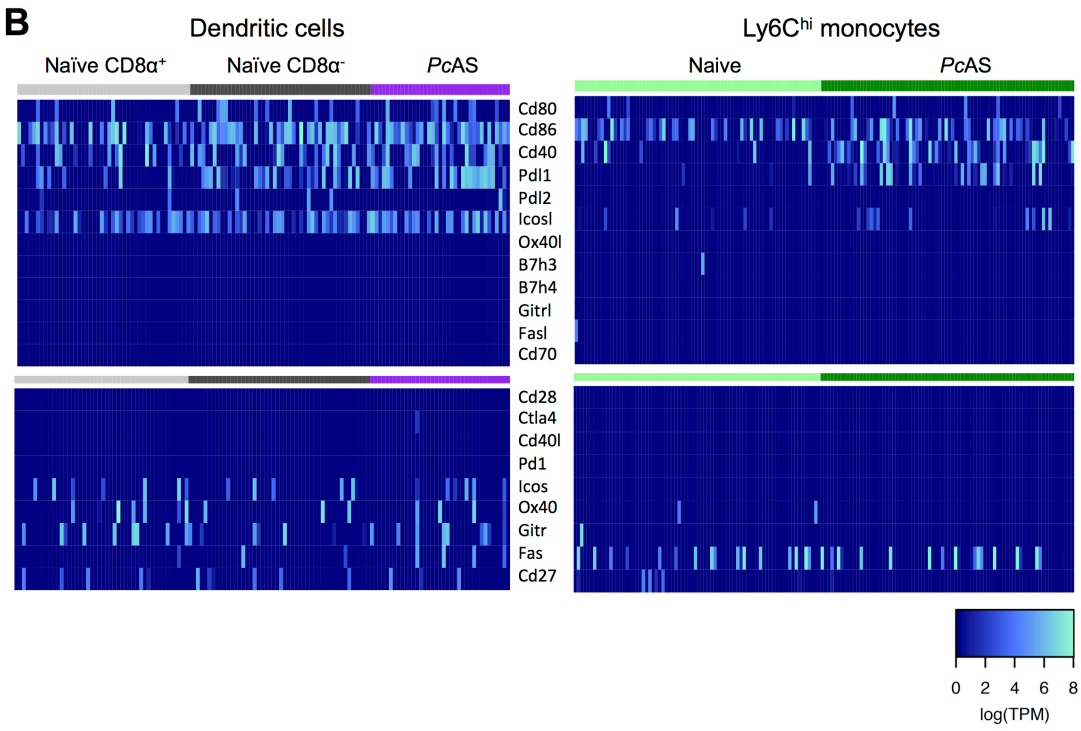
Figure S21

Gene	Log2 (Fold change)	qval
Gbp2	5.34E+00	2.14E-33
Egr1	5.03E+00	1.11E-25
Ifi47	3.25E+00	6.85E-20
Cxcl10	9.85E+00	8.78E-20
Irf1	3.97E+00	3.93E-17
Tgtp1	1.25E+01	9.12E-17
Tgtp2	3.86E+00	7.32E-15
Stat1	2.75E+00	2.36E-10
Gbp4	5.77E+00	1.34E-09
Igtp	2.70E+00	1.46E-09
Fam26f	4.71E+00	1.56E-08
Ifi205	2.56E+00	6.71E-08
Gbp7	2.72E+00	9.43E-07
U2	3.96E+00	9.47E-07
Gbp6	7.60E+00	1.30E-06
Irgm1	2.55E+00	2.23E-06
Gbp3	3.65E+00	3.14E-06
Serpina3g	1.05E+01	4.62E-06
Nfkbiz	3.22E+00	7.52E-06
Gm17334	3.71E+00	8.59E-06
Gbp5	5.77E+00	9.75E-06
D4Wsu53e	-1.29E+00	1.86E-05
Ceacam1	-1.49E+00	1.94E-05
Gpcpd1	-3.58E+00	1.98E-05
Hpgd	-2.92E+00	2.11E-05
Pim1	2.21E+00	2.51E-05
Cd244	-2.01E+00	3.26E-05
Cd40	1.70E+00	7.05E-05
Cd274	5.40E+00	8.55E-05
Cd300lb	-1.80E+00	1.03E-04
Kdm6b	2.71E+00	1.30E-04
CT572998.1	4.26E+00	1.39E-04
Nfkbia	2.51E+00	1.47E-04
Cxcl2	5.03E+00	1.88E-04
Ccl3	5.80E+00	2.16E-04
Pira2	-2.11E+00	2.24E-04
Susd3	-1.73E+00	2.88E-04
Nfic	-4.22E+00	3.26E-04
Ogt	-1.84E+00	3.37E-04
Mll3	-2.64E+00	3.72E-04
Mast3	-9.02E+00	3.78E-04
Ptpre	-1.63E+00	4.04E-04
Ccrl2	4.81E+00	4.17E-04
Lilra6	-2.05E+00	4.38E-04
Atp2b1	-1.49E+00	4.72E-04
Ptplad2	-1.12E+00	8.36E-04
Dnm1l	-2.68E+00	8.92E-04
Stk4	-1.04E+00	1.07E-03
Hmgb2	-2.08E-01	1.16E-03
Mar-01	-1.99E+00	1.32E-03
Ifit2	3.96E+00	1.44E-03
Klf4	-2.63E+00	1.59E-03
Calm2	-8.75E-01	1.61E-03
Itfg3	-1.45E+00	2.17E-03
Dnase1l3	3.16E+00	2.21E-03
Nrp2	3.68E+00	2.52E-03
Zbp1	1.92E+00	3.89E-03
Gbp9	1.57E+00	3.91E-03
Tsc22d3	-1.15E+00	3.97E-03
Rell1	-2.69E+00	4.03E-03
Srsf2	-1.23E+00	4.08E-03
Jun	2.22E+00	5.32E-03
Hnrnph1	-2.23E+00	5.49E-03
Itgb2	-1.02E+00	5.81E-03
Ifit3	2.67E+00	5.90E-03
Eif4a2	-1.39E+00	6.05E-03
Rps6ka1	-1.72E+00	6.59E-03
Rtp4	2.19E+00	6.62E-03
Sfr1	-7.00E-01	6.70E-03
Fam107b	-1.63E+00	7.79E-03
Tmem164	-1.11E+00	7.98E-03
Mnda	1.14E+00	9.66E-03
Glg1	-7.70E-01	1.06E-02
Irgm2	3.51E+00	1.10E-02
Emilin2	-1.78E+00	1.16E-02
Tlk1	-2.16E+00	1.19E-02
Tmem126b	-6.42E+00	1.47E-02
Lrp1	-1.51E+00	1.68E-02
Zzef1	-4.02E+00	1.78E-02
Pik3cg	-1.62E+00	1.89E-02
Camk1d	-3.82E-01	1.97E-02
Cd84	-1.75E+00	1.98E-02
Vps13c	-1.66E+00	2.08E-02
Cd300ld	-4.31E+00	2.15E-02
Atf3	2.59E+00	2.17E-02
Emr4	-1.48E+00	2.27E-02
Cx3cr1	-2.06E+00	2.34E-02
Nedd8	-4.54E-01	2.45E-02
Tubb6	-1.63E+00	2.51E-02
Ndrg1	-1.90E+00	2.83E-02
Mknk2	-1.74E+00	2.87E-02
Pqlc3	-1.40E+00	2.92E-02
Nfam1	-1.19E+00	3.09E-02
Fam89b	-1.85E+00	3.20E-02
Foxj2	-4.79E+00	3.36E-02
H2-Q10	2.32E+00	3.68E-02
Ccl2	5.21E+00	3.84E-02
Idh1	-1.69E+00	4.07E-02
Gm12250	2.68E+00	4.36E-02
Fam46a	-6.38E-01	4.39E-02
Ankrd44	-7.78E-01	4.57E-02
Pdcd4	-8.95E-01	4.74E-02

Figure S21. Differentially expressed genes between single Ly6Chi monocytes from naïve and day 3 *Pc*AS-infected mice. List of differentially expressed genes, expressed in >10 cells (qval<0.05) between Ly6Chi monocytes from naïve and day 3-infected mice. Mean TPM fold-change in gene expression relative to naïve levels.

Figure S22





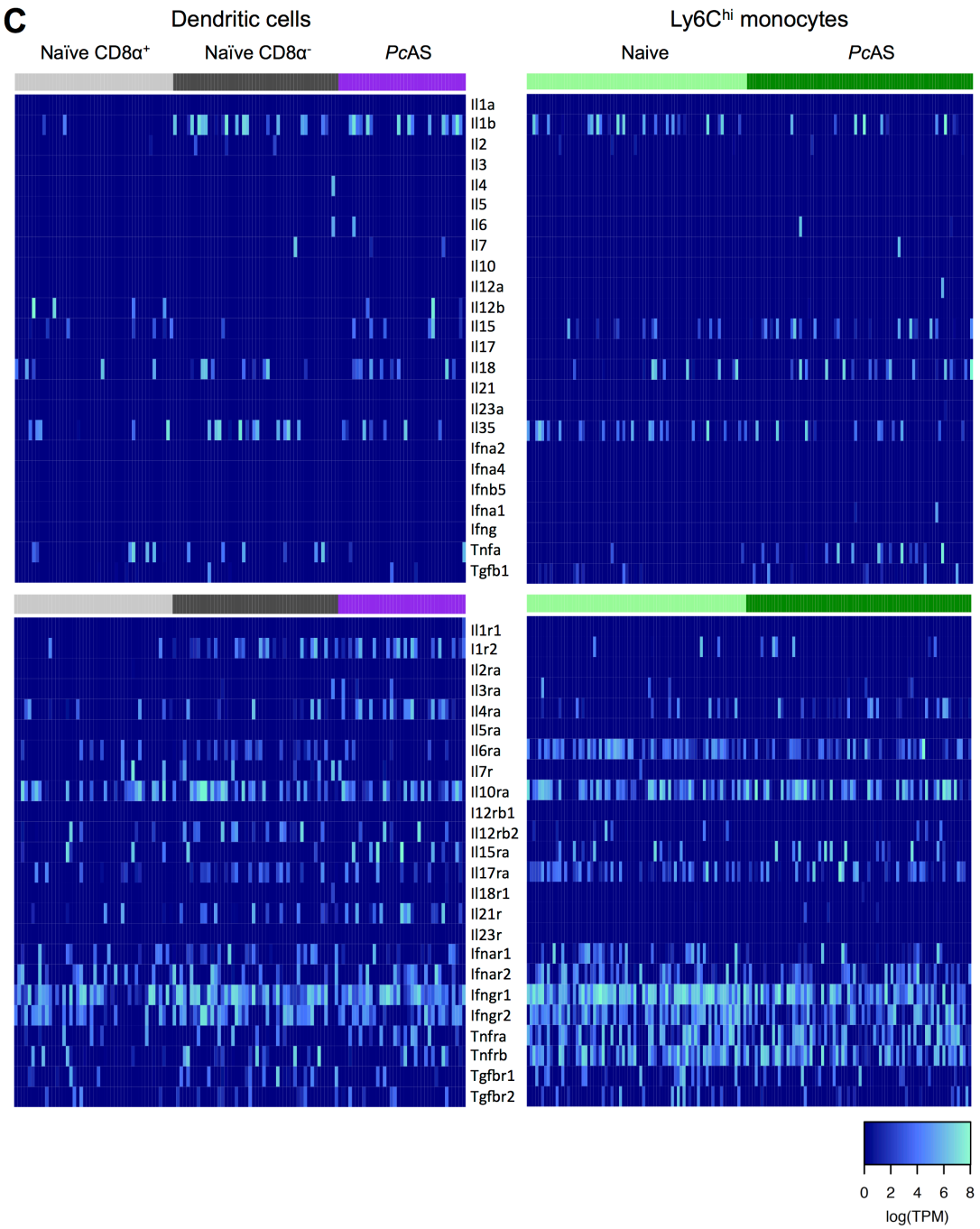


Fig. S22. Expression of immune signalling genes by cDCs and monocytes.

(A-C) Heatmaps showing normalised mRNA expression of select **(A)** chemokines, **(B)** costimulatory molecules and **(C)** cytokines and respective receptors (rows) by single splenic cDCs and Ly6Chi monocytes (columns) from naïve or 3-day *Plasmodium chabaudi chabaudi* AS-infected mice.

Figure S23

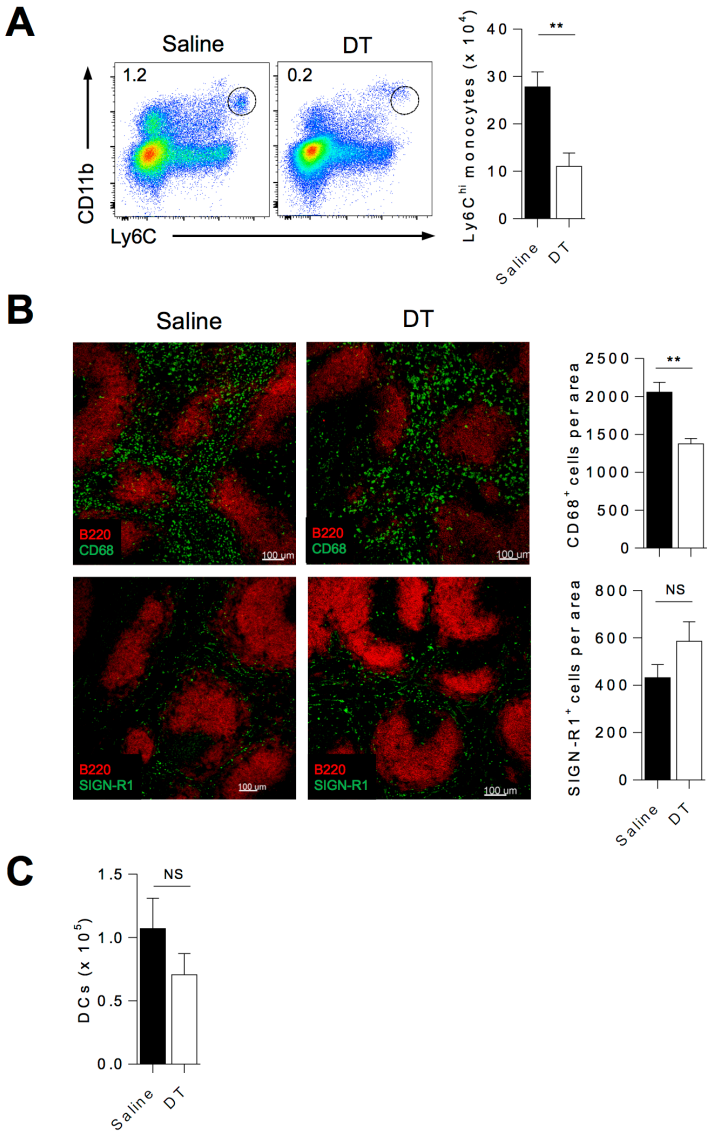


Figure S23. Myeloid cell depletion in LysMCre x iDTR mice.

LysMCre x iDTR mice were infected with *Pc*AS, and treated 3 days later with DT (10ng/g intraperitoneal injection) or control saline (n=6 per group). 24 hours later spleens were harvested for cellular compositional analysis:

(A) Representative FACS plots enumerating splenic inflammatory monocytes (Ly6Chi CD11bhi Ly6G- B220- TCRβ-).

(**B)** Representative fluorescence micrographs showing spleen tissue sections co-stained for B cells (B220 in red) and macrophages (CD68 (top panel) or SIGN-R1 (bottom panel) in green) and summary graphs of average cell number in three fields of view covering the total cross section of a spleen.

(C) Flow cytometric enumeration of splenic cDC (CD11chi MHCIIhi B220- TCRβ-).