nature immunology

Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses

Janilyn Arsenio[1,4], Boyko Kakaradov[2,4], Patrick J Metz[1], Stephanie H Kim[1], Gene W Yeo[2,3] & John T Chang[1]

© 2014 Nature America, Inc. All rights reserved.

T lymphocytes responding to microbial infection give rise to effector cells that mediate acute host defense and memory cells that provide long-lived immunity, but the fundamental question of when and how these cells arise remains unresolved. Here we combined single-cell gene-expression analyses with 'machine-learning' approaches to trace the transcriptional 'roadmap' of individual CD8+ T lymphocytes throughout the course of an immune response *in vivo*. Gene-expression signatures predictive of eventual fates could be discerned as early as the first T lymphocyte division and may have been influenced by asymmetric partitioning of the receptor for interleukin 2 (IL-2Rα) during mitosis. Our findings emphasize the importance of single-cell analyses in understanding fate determination and provide new insights into the specification of divergent lymphocyte fates early during an immune response to microbial infection.

During microbial infection, responding T lymphocytes give rise to two distinct classes of cellular progeny: effector cells that provide acute host defense, and long-lived memory cells that provide durable immunity[1]. Terminally differentiated, short-lived effector T cells (T_{SLE} cells) can be identified phenotypically by their high expression of the lectin-like receptor KLRG1 and low expression of the receptor for interleukin 7 (IL-7R)[2]. At least two distinct subsets of long-lived memory cells, central memory T cells (T_{CM} cells) and effector memory T cells (T_{EM} cells), have been described; these can be distinguished on the basis of their proliferative capacity, cytotoxicity, anatomic localization and expression of certain homing and chemokine receptors, including L-selectin (CD62L) and CCR7 (refs. 3,4).

Published studies using single-cell adoptive transfer and 'genetic barcoding' approaches[5,6] have demonstrated that a single naive CD8+ T lymphocyte can give rise to cells with more than one fate and, notably, is able to generate all of the diverse cellular fates necessary for an immune response. The process by which a single activated T lymphocyte yields effector- and memory-fated progeny and the time at which those differentiation pathways begin to diverge, however, remain unresolved. One possibility is that the progeny of an activated naive CD8+ T lymphocyte progress along a linear differentiation path, initially becoming effector cells, with a subset of those cells later acquiring the memory fate[1,7,8]. An alternative possibility is that the first CD8+ T cell division *in vivo* is asymmetric[9,10] and thus enables lymphocyte fates to diverge early during an immune response owing to unequal inheritance of certain determinants, such as the receptor for interferon-γ (IFN-γ) and the transcription factor T-bet.

Tracing individual lymphocytes sequentially as they differentiate *in vivo* might distinguish whether lymphocytes progress along a linear differentiation pathway[1,7,8] or diverge early during an immune response. While genomic profiling studies have begun to elucidate the transcriptional networks that control lymphocyte fate specification[11–13], such studies have been based on analyses of bulk cellular populations, which makes it impossible to discern cell-fate 'decisions' made by individual T cells. Technological advances that have coupled microfluidics technologies with high-throughput quantitative RT-PCR analyses have enabled detailed analyses of cell-fate 'decisions' during the development of *Caenorhabditis elegans*, induced stem-cell reprogramming and cancer biology[14–17]. Here we used single-cell gene-expression profiling to investigate the ontogeny of effector and memory CD8+ T lymphocytes during microbial infection *in vivo* and found evidence of heterogeneity in gene expression within individual lymphocytes early after the initiation of an adaptive immune response.

RESULTS

Single-cell gene-expression analyses of CD8+ T cells

To delineate the hierarchy and mechanism of CD8+ T cell differentiation during an adaptive immune response at the single-cell level, we used an experimental system that allowed us to investigate the gene expression of individual CD8+ T lymphocytes throughout the course of a microbial infection *in vivo*. We adoptively transferred OT-I CD8+ T cells, which have transgenic expression of T cell antigen receptor that recognizes a specific ovalbumin (OVA) epitope, into wild-type recipient mice. We then infected the recipient mice

[1]Department of Medicine, University of California San Diego, La Jolla, California, USA. [2]Department of Cellular and Molecular Medicine, UCSD Stem Cell and Bioinformatics Programs, and Institute for Genomic Medicine, University of California San Diego, La Jolla, California, USA. [3]Department of Physiology, National University of Singapore and Genome Institute of Singapore and Molecular Engineering Laboratory, A*STAR, Singapore. [4]These authors contributed equally to this work. Correspondence should be addressed to G.W.Y. (geneyeo@ucsd.edu) or J.T.C. (changj@ucsd.edu).

Received 12 December 2013; accepted 30 January 2014; published online 2 March 2014; doi:10.1038/ni.2842

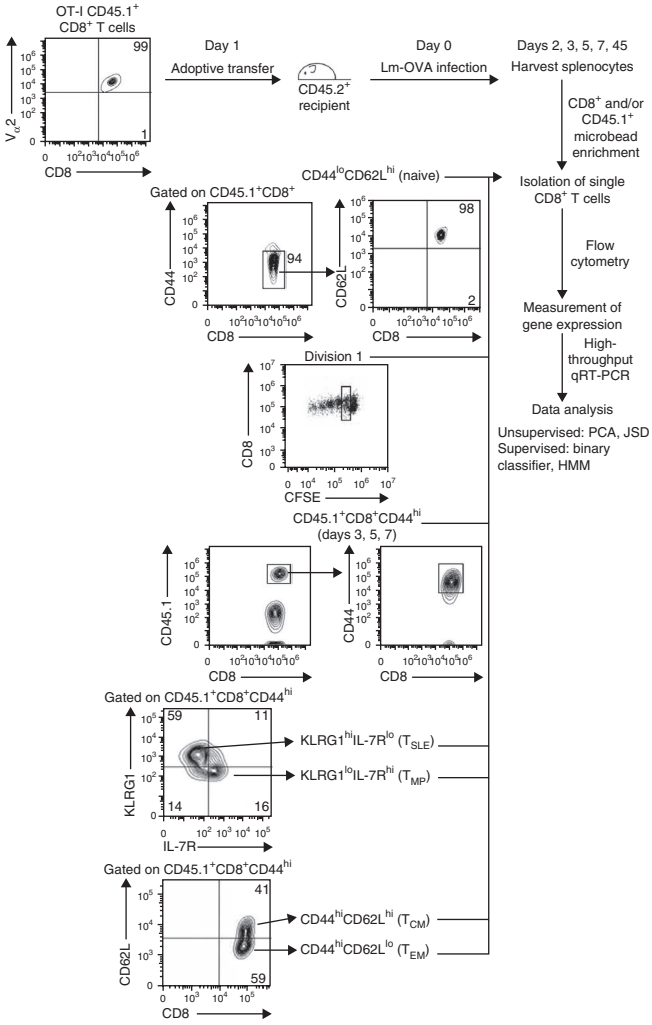

Figure 1 Gating strategy and experimental approach. Unlabeled or CFSE-labeled CD45.1$^+$ OT-I CD8$^+$ T cells were adoptively transferred intravenously into CD45.2$^+$ recipient mice, which were then left uninfected (naive; CD8$^+$CD45.1$^+$CD44^{lo}CD62L^{hi}) or infected with Lm-OVA at 24 h after cell transfer. CD8$^+$ T cell subsets were isolated at various times after infection (division 1 (CD8$^+$CD45.1$^+$CD44^{hi} cells within the second brightest CFSE peak); days 3, 5, and 7 after infection; T_{SLE} cells at day 7 (CD8$^+$CD45.1$^+$CD44^{hi}KLRG1^{hi}IL-7R^{lo}); putative T_{MP} cells at day 7 (CD8$^+$CD45.1$^+$CD44^{hi}KLRG1^{lo}IL-7R^{hi}); T_{CM} cells at day 45 (CD8$^+$CD45.1$^+$CD44^{hi}CD62L^{hi}); and T_{EM} cells at day 45 (CD8$^+$CD45.1$^+$CD44^{hi}CD62L^{lo})), followed by single-cell gene-expression analysis and data analysis with PCA and JSD (unsupervised) and binary classifier and HMM (supervised). Numbers adjacent to outlined areas or in quadrants indicate percent cells in each. Data are representative of three experiments.

intravenously 24 h later with recombinant *Listeria monocytogenes* bacteria expressing ovalbumin (Lm-OVA) and isolated CD8$^+$ T cells throughout the course of infection for single-cell analysis (**Fig. 1**). In addition, we selected for analysis terminally differentiated T_{SLE} cells (KLRG1^{hi}IL-7R^{lo})[2], putative memory precursor T cells (T_{MP} cells; KLRG1^{lo}IL-7R^{hi})[2], T_{CM} cells (CD44^{hi}CD62L^{hi}) and T_{EM} cells (CD44^{hi}CD62L^{lo})[3,4] (**Fig. 1**).

We used Fluidigm 96.96 Dynamic Arrays for quantitative real-time PCR analysis, which allowed simultaneous measurement of the expression of 96 genes in 96 individual cells (**Supplementary Fig. 1a**). Among the 94 genes (**Table 1** and **Supplementary Table 1**) we selected for analysis were those encoding transcriptional regulators reported to influence CD8$^+$ T lymphocyte differentiation[18–25]; cytokines, chemokines, and their receptors[19]; and molecules associated with tissue homing and survival[19]. After excluding failed reactions, we retained expression data from 1,300 single cells for in-depth analyses (**Supplementary Fig. 1b**). Because expression of 'housekeeping' genes has been shown to vary substantially across cell types and states of differentiation[26], we used the expression of each gene of interest without normalization for all our analyses here.

We used principal-component analysis (PCA) to visualize the expression data globally. PCA is an unsupervised dimensionality-reduction method that we used to project the data into two dimensions by its coordinates in the first two principal components (PC1 and PC2) that account for the largest variations in the data. Those principal components were linear combinations of the expression data for the 94 original genes. PCA revealed that naive cells, T_{SLE} cells, T_{EM} cells and T_{CM} cells clustered distinctly (**Fig. 2a**). Expression of *Sell* and *Tcf7*, which encode the trafficking molecule CD62L and the transcription factor TCF-1, respectively, distinguished naive cells from T_{SLE} cells (**Fig. 2a**), consistent with published findings[2,4]. Although T_{SLE} cells formed a distinct cluster, these cells were projected closest to T_{EM} cells (**Fig. 2a**), which suggested that related gene-expression profiles may underlie some of their functional similarities, such as cytotoxicity and the secretion of proinflammatory cytokines[27]. That clustering was driven by expression of *Zeb2*, which encodes a transcription factor expressed in T_{SLE} cells[12]. In addition, T_{EM} cells and T_{CM} cells occupied distinct clusters, with higher expression of *Tcf7*, *Il2rb*, *Il7r*, *Cxcr3* and *Sell* mRNA in T_{CM} cells and higher expression of *Zeb2* mRNA in T_{EM} cells accounting for the variance between these memory cell populations. We confirmed at the protein level some of the disparities observed at the transcriptional level (**Fig. 2b**), which supported our finding that T_{CM} cells and T_{EM} cells were molecularly distinct. The higher expression of *Il7r* and *Tcf7*, which encode regulators of T lymphocyte survival and longevity[25,28], that we observed in T_{CM} cells may underlie the superior ability of these cells to persist *in vivo*[29]. Putative T_{MP} cells did not form a distinct cluster but overlapped T_{SLE} cells, T_{EM} cells and T_{CM} cells (**Fig. 2c**). These results suggested that putative T_{MP} cells are molecularly heterogeneous, which raised the possibility that this population may not represent memory precursor cells but instead may consist of 'mature' memory cells and terminally differentiated effector cells. Together these findings suggested that T_{SLE} cells, T_{CM} cells and T_{EM} cells exhibited similar gene-expression profiles at the single-cell level, but putative T_{MP} cells did not.

Molecular heterogeneity at the single-cell level

To assess whether single responding CD8$^+$ T cells were molecularly heterogenous early after infection, we analyzed the gene-expression profiles of individual CD8$^+$ T cells (**Fig. 3a**) isolated throughout the course of the infection described above (**Fig. 1**). PCA revealed greater heterogeneity among cells isolated early after infection (division 1 and day 3) than among cells isolated at later time points (day 5 and day 7; T_{SLE}, T_{CM}, and T_{EM} cells). The first two principal components captured 17% of the variance in our data set (**Fig. 3**), slightly lower than that in published results[15]; this was probably a reflection of a higher degree of heterogeneity in lymphocytes during differentiation and the greater number of genes analyzed in our study. In agreement with our findings obtained by PCA, an alternative 'unsupervised' method (t-distributed stochastic neighbor-embedding analysis[30]) produced similar results (**Supplementary Fig. 2**). To determine whether the heterogeneity observed with data from single cells could be recapitulated with data from bulk cells, we formally compared the analyses with data derived from single populations versus those derived from bulk

© 2014 Nature America, Inc. All rights reserved.

Table 1 Functional classification of gene targets

Class	Gene
Apoptosis	*Bnip2*, *Bnip3l*, *Casp3*, *Casp9*, *Cflar*, *Pdcd1*
Cytokine or chemokine receptor	*Ccr5*, *Ccr6*, *Ccr7*, *Cxcr3*, *Ifngr1*, *Ifnar1*, *Il2ra*, *Il2rb*, *Il7r*, *Klrc1*, *Klrg1*, *Tnfrsf1*, *Tnfrsf9*
Cytokine, chemokine, granzyme	*Ccl3*, *Ccl4*, *Ccl5*, *Cxcl10*, *Gzmb*, *Ifng*, *Il2*, *Il3*, *Lif*, *Xcl1*
Housekeeping	*Actb*, *Gapdh*, *Rn18s*, *Rpl35*
Polarity-proteasome	*Prkcz*, *Psmb1*, *Psmb7*
Signaling, proliferation, self-renewal	*Bag2*, *Bmi1*, *Bmp2*, *Cd28*, *Cd4*, *Cd44*, *Cd69*, *Cd8a*, *Grap2*, *Hk2*, *Lag3*, *Lgals1*, *Mapk3*, *Mapk8*, *Mapk14*, *Mela*, *Mtor*, *Myc*, *Ptprc*, *RelA*, *Sema7a*, *Serpinb6b*, *Serpinb9*, *Setd7*, *Sell*, *Thy1*
Transcription factor	*Atf1*, *Bcl11b*, *Bcl6*, *Bhlhe40*, *Eomes*, *Foxo1*, *Foxo3*, *Foxp1*, *Foxp3*, *Gata3*, *Hopx*, *Id2*, *Id3*, *Irf4*, *Irf8*, *Klf2*, *Lef1*, *Nfatc1*, *Nfatc2*, *Prdm1*, *Rel*, *Runx1*, *Runx2*, *Stat1*, *Stat4*, *Tbx21*, *Tcf3*, *Tcf7*, *Tcf12*, *Tox*, *Zeb2*, *Zfp281*

Functional classification of the products of 94 selected gene targets.

populations (**Fig. 3a**). We found that the heterogeneity we observed at the single-cell level in putative T_{MP} cells and cells isolated early after infection was not apparent in the bulk analysis (**Fig. 3a**); this illustrated the power and necessity of using a single-cell approach.

To further evaluate the degree of heterogeneity within and between cell populations at each time point, which was not previously possible with bulk analysis, we applied the Jensen-Shannon divergence (JSD) metric, a nonparametric, model-free measure of similarity between two empirical probability distributions. In general, the intrapopulation JSD metric was lowest for naive cells and highest for cells isolated early after infection (**Fig. 3b** and **Supplementary Fig. 3a**). We observed that the intrapopulation JSD metric decreased as a function of time after infection, with the notable exception of putative T_{MP} cells (**Fig. 3b**). Those cells exhibited a high degree of intrapopulation divergence, consistent with the apparent heterogeneity of those cells by PCA (**Fig. 3a**). Pairwise comparison of JSD metrics for all cell populations (naive cells, cells at division 1, cells at days 3, 5 and 7 after infection, and T_{MP}, T_{SLE}, T_{CM} and T_{EM} cells) yielded similar results, with the greatest divergence between cells isolated early after infection versus those isolated late after infection (**Fig. 3b** and **Supplementary Fig. 3a**). Notably, the interpopulation JSD metric was not affected by group size (**Supplementary Fig. 3b**). Together these results demonstrated that $CD8^+$ T lymphocytes responding to a microbe exhibited substantial molecular heterogeneity at the single-cell level early after infection and that this heterogeneity diminished with time.

Distinct transcriptional signatures early after infection

We hypothesized that the heterogeneity observed for lymphocytes early after infection might have reflected distinct gene-expression patterns that are predictive of more differentiated cells. We reasoned that 'supervised' classifiers 'trained' on relatively well-defined, differentiated cellular fates, such as sorted T_{CM} cells and T_{SLE} cells, could be used to assess whether cells isolated early after infection might be fated toward specific $CD8^+$ T lymphocyte subsets. We chose to use 'boosted decision trees'[31] rather than other classification frameworks with similar performance characteristics because the 'learned trees' are easily interpretable. A 'decision tree' built from the data consisted of several predictive rules that compared the expression of *Ptprc*, *Sell* and *Ccl5* to thresholds 'learned' from that data to 'decide' whether a cell was more like a T_{CM} cell or a T_{SLE} cell (**Supplementary Fig. 4a**). Ensembles of 'decision trees' were 'trained' with the RobustBoost algorithm[32] to generate a binary classifier that achieved a misclassification error of approximately 4% in 'leave-one-out' cross-validation, which was split evenly when the classifier was distinguishing between T_{CM} cells and T_{SLE} cells (**Fig. 4a** and **Supplementary Fig. 4b**). The classifier revealed that *Sell* and *Il7r* were among the most predictive genes whose high expression accurately described T_{CM} cells, whereas the lack of expression of *Sell* and *Il7r*, along with high expression of *Zeb2*, defined T_{SLE} cells (**Fig. 4b**). Application of the classifier to cells isolated at days 5 and 7 after infection revealed that 49% and 57%, respectively, of total $CD8^+$ T cells at these time points were more like T_{SLE} cells than T_{CM} cells (**Fig. 4c**), consistent with the expected frequency of T_{SLE} cells at days 5 and 7 after infection[2].

We next sought to determine whether the classifier could discern the fates of responding lymphocytes isolated early during an immune response. It has been suggested that asymmetric division of $CD8^+$ T lymphocytes yields daughter cells proximal to the immunological synapse ('proximal daughter cells') and distal to the immunological synapse ('distal daughter cells') that adopt different fates[6], which raises

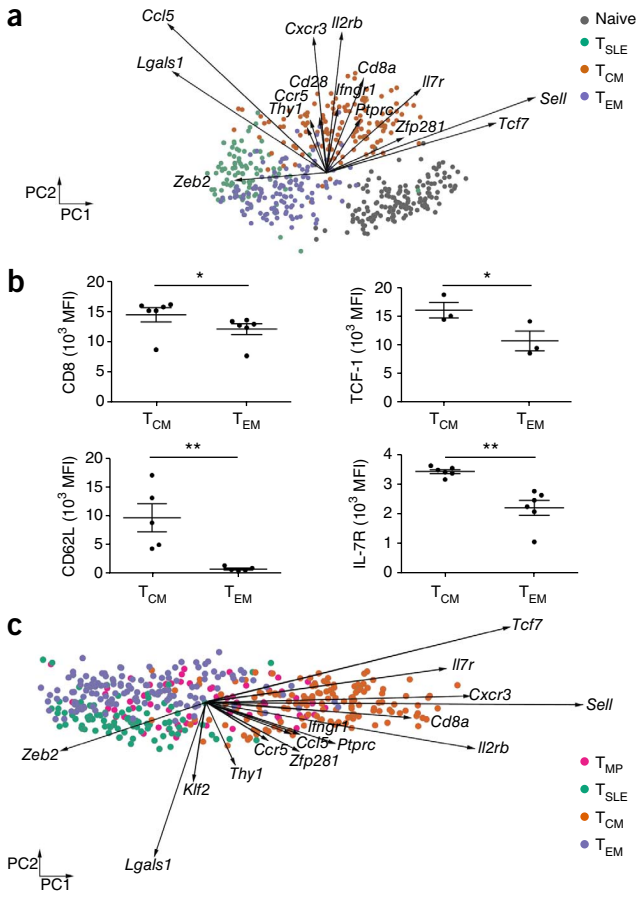

Figure 2 Effector and memory $CD8^+$ T lymphocyte subsets are molecularly distinct on a single-cell level. (**a**) Projections of PC1 and PC2 (bottom left) for single-cell gene-expression data derived from individual lymphocytes from populations of naive, T_{SLE}, T_{CM} and T_{EM} cells (key). Each symbol represents an individual cell; each vector emanating from the origin represents an individual gene. PC1 and PC2 account for 11% and 9% of the variance, respectively. (**b**) Expression of CD8 (encoded by *Cd8a*), TCF-1 (encoded by *Tcf7*), CD62L (encoded by *Sell*) and IL-7R (encoded by *Il7r*) in T_{CM} and T_{EM} cells, assessed by flow cytometry and presented as mean fluorescence intensity (MFI). Each symbol represents an individual mouse; small horizontal lines indicate the mean (± s.e.m.). $^*P < 0.05$ and $^{**}P < 0.01$ (Kolmogorov-Smirnov test). (**c**) PC projections as in **a** of T_{MP}, T_{SLE}, T_{CM} and T_{EM} cells (key). PC1 and PC2 account for 11% and 6% of the variance, respectively. Data are representative of at least two experiments (**a**,**c**) or two experiments with at least three mice in each (**b**).

© 2014 Nature America, Inc. All rights reserved.

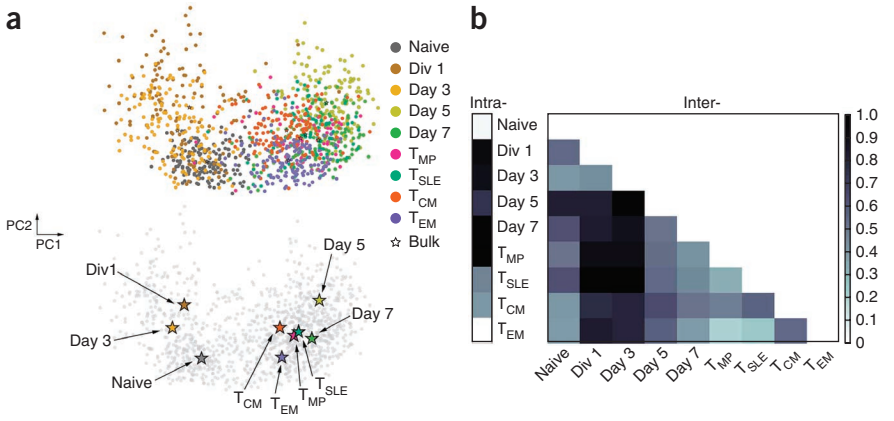

Figure 3 Early heterogeneity of gene expression in individual CD8$^+$ T lymphocytes during an immune response. (**a**) PC projections (as in **Fig. 2a**) of single-cell gene expression in mice infected as described above (**Fig. 1**), for data derived from individual lymphocytes among populations of naive cells, cells at division 1 (Div 1) or at days 3, 5 and 7 after infection, and T_{MP}, T_{SLE}, T_{CM} and T_{EM} cells (key). PC1 and PC2 account for 10% and 7% of the variance, respectively. Below, analysis of pooled 'bulk' samples in each experimental condition (star colors match key colors above); single-cell data points are in gray (background) for clarity. (**b**) Intrapopulation (Intra-) and interpopulation (Inter-) JSD metrics of mean gene expression within and between CD8$^+$ T cell populations (defined as in **a**). Data are representative of at least two experiments.

the possibility that such cells might already exhibit distinct gene-expression patterns that are predictive of their eventual fates as early as the first cell division. To assess this possibility, we sorted putative proximal and distal daughter cells, which can be distinguished by their relative abundance of CD8 and CD11a[9], and analyzed the sorted cells. The classifier revealed that most proximal daughter cells more closely resembled T_{SLE} cells, while most distal daughter cells more closely resembled T_{CM} cells (**Fig. 4c**), which suggested that these cells may indeed adopt different fates.

As further evidence that proximal and distal daughter cells display unique molecular patterns that might drive their distinct fates, we observed that these cells exhibited a pronounced disparity in the expression of genes associated with the effector fate or memory fate (**Fig. 4d**). Certain genes associated with the memory fate in CD8$^+$ T cells, including *Eomes*, *Sell*, *Il7r*, *Il2rb*, *Tcf7*, *Id3* and *Bcl6* (refs. 18,19,21,24,25), had higher expression in distal daughter cells. Conversely, we detected certain genes associated with terminally differentiated effector cells, such as *Tbx21*, *Prdm1* and *Grzmb*[19,20,22], only in proximal, not distal, daughter cells. While it remains possible that the gene-expression patterns of early lymphocytes might change as the cells continue to differentiate, together these results were indicative of distinct molecular patterns, suggestive of a possible predisposition toward different fates, in cells that may have undergone an asymmetric division *in vivo*.

Temporal expression of determinants of CD8$^+$ T cell fates

Having determined that the gene-expression patterns of less-differentiated cells could be used to predict their eventual fates, we next sought to develop a simple generative model for the fate specification of CD8$^+$ T lymphocytes that would capture key genes encoding molecules involved in each step of the differentiation pathway of an individual naive cell. In contrast to the classifiers we 'trained' on cells purified by sorting, to discriminate between different cellular fates (T_{CM} versus T_{SLE}), we used a hidden Markov model (HMM) 'trained' on lymphocytes representative of intermediate states of differentiation (division 1, day 3, day 5) between the naive state and the differentiated fates (**Fig. 5a**). HMMs have been applied to sequential and time-series analyses in diverse fields and have been particularly useful for modeling hidden, unobserved states during biological processes[33,34]. HMMs not only capture the static expression profiles of subpopulations at a particular stage but also can detect dynamic expression changes responsible for the transitions between them. To construct a temporal paradigm of T lymphocyte fate specification *in vivo*, we first defined six linear HMMs and twelve divergent HMMs (**Supplementary Fig. 4c**) that represented possible hypothetical states (such as pre-T_{SLE} or pre-memory) through which an individual naive T lymphocyte could transition before differentiating into any of three observed fates (T_{SLE}, T_{CM} and T_{EM}). To evaluate each HMM, we analyzed all possible paths for each individual cell (**Supplementary Fig. 4c**). Incorporating the single-cell measurements obtained serially for CD8$^+$ T lymphocytes differentiating *in vivo*, we calculated the likelihood of each of the possible differentiation paths for each defined linear or divergent HMM (**Supplementary Fig. 4c**). To determine both the significance and robustness of each HMM, we randomly varied the initial values of the transition matrices by 10% and computed the 'log likelihood' for each iteration. Our results showed that the divergent models generally outperformed the linear models, and we identified an early divergent model as the most likely pathway (**Fig. 5b** and **Supplementary Fig. 4d**). We further evaluated the performance of that final model by randomly ordering the population labels of the cells as well as the associated expression values. Notably, the likelihood of the best model was significantly higher than the likelihood for shuffled data ($P = 0.00034$), which showed that the model robustly indicated that an activated CD8$^+$ T lymphocyte gave rise to cells that transitioned through either a hypothetical pre-T_{SLE} state or pre-memory state. Pre-T_{SLE} cells could undergo further differentiation to acquire the T_{SLE} fate, whereas pre-memory cells could further diverge to give rise to T_{CM} cells or T_{EM} cells. Together these findings suggested that an early divergent model may be the most likely pathway that underlies lymphocyte fate specification *in vivo*.

We analyzed the changes in expression of all 94 genes noted above (**Table 1**) during each of those five unique transitions: naive to pre-T_{SLE}, naive to pre-memory, pre-T_{SLE} to T_{SLE}, pre-memory to T_{CM}, and pre-memory to T_{EM} (**Fig. 5b**,**c** and **Supplementary Fig. 5**). This analysis revealed both shared and unique molecular features of each transition. The naive–to–pre-T_{SLE} and naive–to–pre-memory transitions, for example, were both associated with increased expression of *Lgals1*. Notably, however, the naive–to–pre-T_{SLE} transition was associated with higher expression of *Il2ra* and lower expression of *Cxcr3*, *Sell* and *Tcf7* than that of the naive–to–pre-memory transition, which raised the possibility that these genes might encode molecules that influence whether a cell proceeds along the pathway toward terminal differentiation or self-renewal. The pre-memory–to–T_{CM} and pre-memory–to–T_{EM} transitions were both characterized by increased expression of *Ccl5* and decreased expression of *Foxo1* and *Cxcr3*. However, the pre-memory–to–T_{CM} transition was uniquely associated with increased expression of *Tcf7*, *Il7r* and *Sell*. In contrast, the pre-T_{SLE}–to–T_{SLE} transition was associated with increased expression of *Ccl5* and decreased expression of *Il2ra*, *Il2rb*, and *Foxo1*. Together

© 2014 Nature America, Inc. All rights reserved.

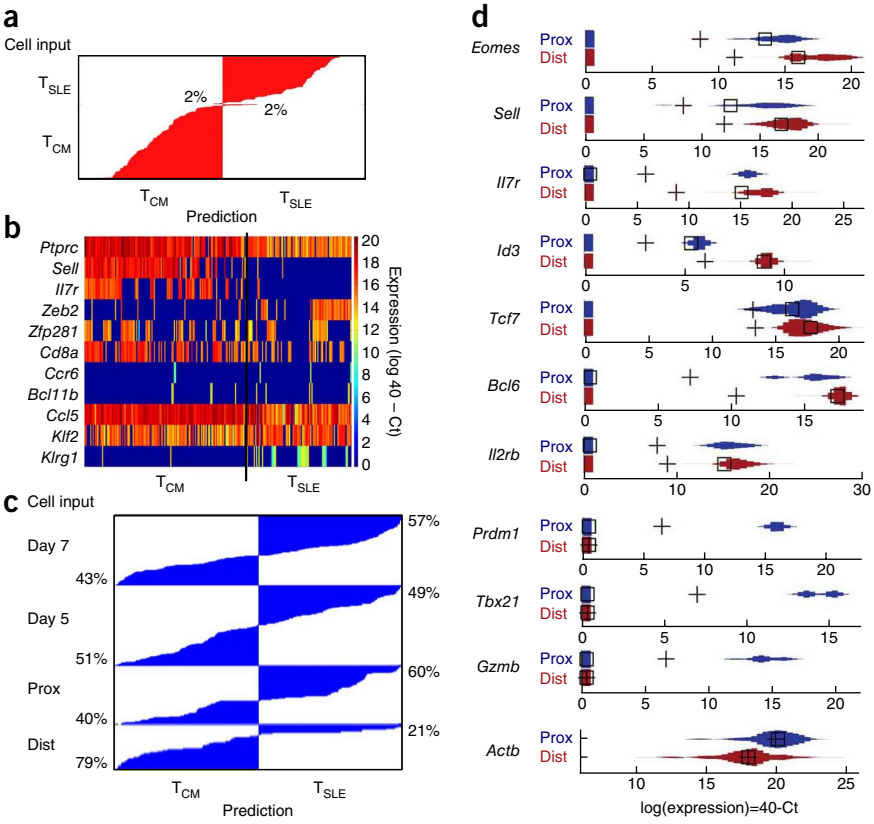

Figure 4 Classifier analysis allows prediction of eventual fates of individual CD8⁺ T lymphocytes. (**a**) Predictions made by the classifier for T_{CM} and T_{SLE} cells purified by sorting, cross-validated during 'training'. Horizontal red bars indicate 'voting margin' for each individual cell, with internal confidence of the classifier's prediction for that cell; numbers indicate rate of misclassification of a T_{SLE} cell as a T_{CM} cell or of a T_{CM} cell as a T_{SLE} cell. (**b**) Binary classifier 'trained' to distinguish between a pair of differentiated cell fates (T_{CM} versus T_{SLE}): each (narrow) column represents an individual T_{CM} or T_{SLE} cell purified by sorting; colors indicate its expression of each gene along left margin (key, right margin: logarithm of the expression measured as '40 – cycling threshold (Ct)'). (**c**) Classifier analysis of individual CD8⁺ T cells (thin stacked horizontal blue bars) from various populations (cells isolated at day 5 or 7 after infection (as in **Fig. 1**) and proximal (Prox) or distal (Dist) daughter cells at the first division), with predictions sorted by confidence (length of blue bars) from the most T_{CM}-like cells to most T_{SLE}-like cells. Numbers along margins indicate proportion of cells predicted to be more T_{CM}-like (left) or T_{SLE}-like (right) in each cell population. (**d**) Expression (as measured in **b**) of various genes (left margin) by proximal (blue) and distal (red) daughter cells at the first division, presented as 'violin plots'; black crosses and squares indicate mean and mode, respectively. Data are representative of at least two experiments.

these results revealed the temporal expression patterns of key genes encoding molecules that influence the fates of CD8⁺ T lymphocytes responding to microbial infection *in vivo*.

IL-2Rα asymmetry is associated with distinct cellular fates

We found the prediction, raised by our temporal model, that IL-2Rα might represent an early molecular switch that promotes the pathway toward terminal differentiation intriguing, given published work suggesting a role for IL-2 signaling in the differentiation of CD8⁺ T lymphocytes[35–39]. To determine how early we could detect a possible disparity in *Il2ra* expression in lineages destined for an effector fate versus those destined for a memory fate, we used flow cytometry to examine the expression of IL-2Rα in CD8⁺ T cells that had undergone their first division *in vivo* in response to microbial infection. We observed that differences in the abundance of IL-2Rα on the cell surface distinguished two populations of first daughter cells (**Fig. 6a**) and that IL-2Rα abundance was inversely correlated with CD62L expression (**Fig. 6a**). Furthermore, cells with higher expression of IL-2Rα also exhibited increased production of IFN-γ and granzyme B, characteristic of effector cells (**Fig. 6b**).

To test the hypothesis that the amount of IL-2Rα expression conferred a distinct predisposition toward the effector lineage or

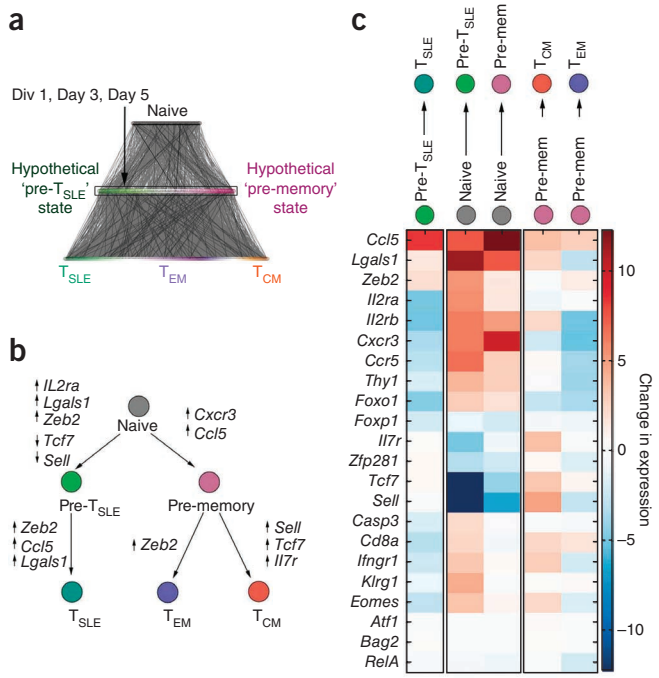

Figure 5 Temporal model for predicting the differentiation paths of individual CD8⁺ T lymphocytes. (**a**) Ranking of cells in early states of differentiation (division (Div) 1, day 3, day 5) by their T_{SLE}- or memory-like expression profile (green-to-purple gradient; middle), followed by random linkage to sorted naive cells (top) and sorted T_{SLE}, T_{EM} and T_{CM} cells (green-to-purple-to-orange gradient; bottom), for the formation of hypothetical differentiation paths (black lines) analyzed through the use of an HMM. (**b**) Most likely model for the differentiation of a CD8⁺ T lymphocyte, with key changes in gene expression (upward arrows, upregulation; downward arrows, downregulation) associated with each of the following five unique transitions: naive to pre-T_{SLE}; naive to pre-memory; pre-T_{SLE} to T_{SLE}; pre-memory to T_{CM}; and pre-memory to T_{EM} (filled symbols). (**c**) Genes with consistent changes in expression during each transition phase, predicted by temporal model for the differentiation of CD8⁺ T lymphocytes (key; change in expression during one transition compared with change in expression during other transitions). Data are representative of at least two experiments.

© 2014 Nature America, Inc. All rights reserved.

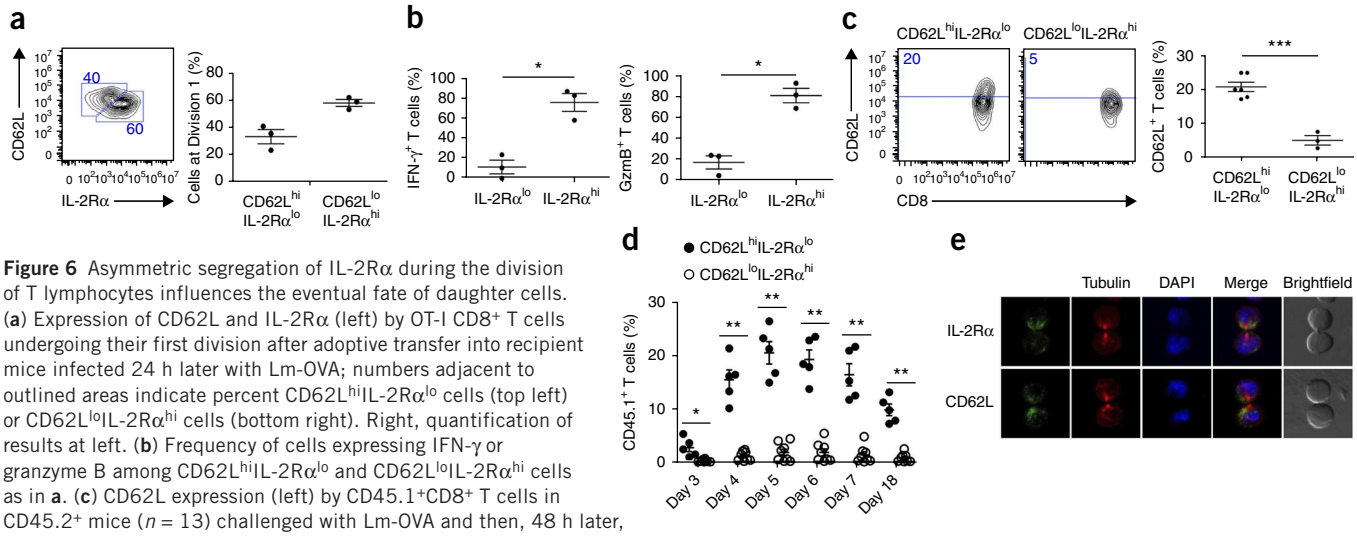

Figure 6 Asymmetric segregation of IL-2Rα during the division of T lymphocytes influences the eventual fate of daughter cells. (**a**) Expression of CD62L and IL-2Rα (left) by OT-I CD8$^+$ T cells undergoing their first division after adoptive transfer into recipient mice infected 24 h later with Lm-OVA; numbers adjacent to outlined areas indicate percent CD62L^{hi}IL-2Rα^{lo} cells (top left) or CD62L^{lo}IL-2Rα^{hi} cells (bottom right). Right, quantification of results at left. (**b**) Frequency of cells expressing IFN-γ or granzyme B among CD62L^{hi}IL-2Rα^{lo} and CD62L^{lo}IL-2Rα^{hi} cells as in **a**. (**c**) CD62L expression (left) by CD45.1$^+$CD8$^+$ T cells in CD45.2$^+$ mice (n = 13) challenged with Lm-OVA and then, 48 h later, given IL-2Rα^{lo}CD62L^{hi} or IL-2Rα^{hi}CD62L^{lo} cells at division 1 (purified by sorting), assessed on day 49 after infection of recipients; numbers in top left corners indicate percent CD62L$^+$ cells. Right, quantification of CD62L$^+$ T cells among CD45.1$^+$ T cells (as at left). (**d**) Population expansion of CD45.1$^+$CD8$^+$ T cells in blood obtained (time, horizontal axis) from mice in **c** that were subsequently rechallenged with Lm-OVA at day 50 after the primary infection. (**e**) Confocal microscopy of IL-2Rα (green; top row) or CD62L (green; bottom row), β-tubulin (red) and DNA (blue; stained with the DNA-intercalating dye DAPI) in sorted OT-I CD8$^+$ T cells undergoing their first division after adoptive transfer into Lm-OVA-infected recipient mice. Each symbol (**a**–**c** (left), **d**) represents an individual mouse; small horizontal lines indicate the mean (± s.e.m.). $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P = 0.0002$ (Kolmogorov-Smirnov test). Data are representative of three (**a**,**b**,**e**) or two (**c**,**d**) experiments.

memory lineage, we sorted CD62L^{lo}IL-2Rα^{hi} or CD62L^{hi}IL-2Rα^{lo} cells that had undergone their first division *in vivo*. We then adoptively transferred those cells into recipient wild-type mice that we had infected 48 h before with Lm-OVA. We tracked the progeny of adoptively transferred cells at multiple time points throughout the course of the primary response and found that the progeny of both CD62L^{lo}IL-2Rα^{hi} cells and CD62L^{hi}IL-2Rα^{lo} cells were detectable after infection (**Supplementary Fig. 6**). Notably, however, the progeny of the transferred IL-2Rα^{lo} cells exhibited a fourfold greater capacity to give rise to T_{CM} cells than did the progeny of transferred IL-2Rα^{hi} cells (**Fig. 6c**). To confirm functionally that those cells were indeed memory lymphocytes, we assessed their ability to respond to microbial rechallenge. We rechallenged the recipient mice with Lm-OVA at day 50 after primary infection and observed tenfold more population expansion by the progeny of transferred CD8$^+$ T cells in recipient mice that had received CD62L^{hi}IL-2Rα^{lo} cells than in mice that had received CD62L^{lo}IL-2Rα^{hi} cells (**Fig. 6d**), which suggested that these cells had different abilities to give rise to memory lymphocytes.

Because certain cytokine and immunological receptors can undergo unequal partitioning during cell division[9], we hypothesized that asymmetric segregation of IL-2Rα and CD62L during mitosis might provide a mechanism underlying their difference in abundance on daughter cells that had undergone their first division *in vivo*. We used an experimental system that has allowed analysis of T cells preparing for their first division in response to a microbe[9]. We labeled OT-I CD8$^+$ T cells with the division-tracking fluorescent dye CFSE and adoptively transferred those cells into recipient mice that we had infected 24 h before with Lm-OVA. We isolated undivided donor CD8$^+$ T cells by flow cytometry at 36 h after transfer and examined them by confocal microscopy. We observed that IL-2Rα and CD62L exhibited pronounced asymmetric distribution in cells that were preparing for division (asymmetric segregation of IL-2Rα in 60% of cells (n = 96) and of CD62L in 62% of cells (n = 74), respectively. **Fig. 6e**). Together these results suggested that the asymmetric segregation of IL-2Rα and CD62L during the first division of a CD8$^+$ T lymphocyte *in vivo* may influence the transcriptional profiles of the nascent daughter cells and their eventual fates.

DISCUSSION

Advances in high-throughput single-cell gene-expression profiling have enabled the use of this technique in such diverse fields as embryonic development, hematopoiesis, stem cell reprogramming and cancer biology[14–17]. Those advances, coupled with computational modeling approaches, enabled us to investigate, on a level of molecular detail not previously possible, the ontogeny of effector and memory lymphocytes during microbial infection *in vivo*. We found evidence of considerable heterogeneity in gene expression among individual CD8$^+$ T lymphocytes early after the initiation of microbial infection. Notably, we demonstrated that such heterogeneity could not be identified by traditional bulk-population analyses and that many of our computational analyses here, including JSD, classifier and HMM, were possible only with data derived from single cells. Our observations provide a compelling argument for the integration of single-cell approaches into future studies of the fate specification of cells of the immune system.

Through the use of sequential single-cell gene-expression measurements in activated lymphocytes during the course of a microbial infection *in vivo*, we constructed a temporal model that allowed us to predict the timing and changes in the expression of key genes in individual lymphocytes as they transitioned from the naive state toward each of several cellular fates. We provided experimental evidence in support of a prediction of this temporal model: that differences in the expression of IL-2Rα may reflect one of the earliest molecular determinants that influence the memory-versus-effector fate 'decision'. Moreover, we demonstrated that unequal partitioning of IL-2Rα during the first asymmetric division *in vivo* may result in its disparate abundance in daughter lymphocytes and potentially contribute to their acquisition of distinct gene-expression profiles and cellular fates.

Along with published evidence that other critical signaling molecules, such as the receptor for IFN-γ, can be unequally partitioned[9],

© 2014 Nature America, Inc. All rights reserved.

our results suggest that asymmetric segregation of cytokine receptors during lymphocyte division may result in proximal daughter cells' encountering more signaling by IL-2 and IFN-γ than that encountered by distal daughters. As IL-2 has been shown to induce *Prdm1* and repress *Bcl6* and *Il7ra*[37,38], while IFN-γ is known to induce *Tbx21* (refs. 40,41), differences in cytokine signaling encountered by proximal and distal daughter cells may initiate a pre-effector gene-expression program or pre-memory gene-expression program, respectively, consistent with our experimental observations and with published work showing that cells that receive prolonged IL-2 signals acquire characteristics of terminally differentiated effector cells[37]. Continued changes in gene-expression patterns, influenced by environmental signals, may allow lymphocytes to continue along distinct pathways toward terminal differentiation or self-renewal.

Published reports describing additional subsets of memory T lymphocytes, however, raise the possibility that the effector lineage or central memory lineage may not be the exclusive fate adopted by the progeny of $CD62L^{lo}IL\text{-}2R\alpha^{hi}$ cells and $CD62L^{hi}IL\text{-}2R\alpha^{lo}$ cells. Tissue-resident memory T cells[42,43] do not circulate and instead remain in the peripheral tissues after pathogen clearance, while so-called 'effector-phenotype' memory T cells share certain phenotypic characteristics with terminally differentiated effector cells and mediate robust immunological protection in certain infectious settings despite exhibiting poor proliferative recall responses[44]. Indeed, some of the progeny of $IL\text{-}2R\alpha^{hi}CD62L^{lo}$ cells seemed to give rise to a population of lymphocytes that, although poorly proliferative in response to microbial rechallenge, persisted *in vivo*, reminiscent of effector-phenotype memory cells. Thus, it remains possible that the first cellular division, in addition to mediating a divergence of the effector and memory fates, may also facilitate the specification of distinct memory cell subset fates.

Although the generation of long-lived memory lymphocytes is an essential feature of an adaptive immune response, the fundamental question of when and how these cells arise has remained controversial. Resolving whether lymphocytes progress along a linear differentiation pathway or diverge early during an immune response, owing to asymmetric cell division, necessitated our tracing individual lymphocytes as they underwent differentiation *in vivo*. By investigating the gene-expression patterns of individual lymphocytes during an immune response to a microbial pathogen, we reconstructed the lineage paths of single lymphocytes as they differentiated *in vivo*. This approach has yielded new insights into lymphocyte fate specification and has provided new evidence in support of the proposal of an early divergence of lymphocyte fates, via asymmetric division, during an adaptive immune response to microbial infection. More broadly, we anticipate that single-cell gene-expression approaches undertaken by investigators across scientific disciplines, along with ever-improving advances in such technologies as single-cell RNA sequencing[45,46] and single-cell mass cytometry[47], will continue to provide unprecedented molecular insights into cell-fate specification in diverse biological settings, including immunity, development and cancer.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GEO: gene-expression array data, GSE54321.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Hedrick, J. Bui, A. Goldrath, S. Schoenberger and members of the Chang and Yeo laboratories for discussions and critical reading of the manuscript. Supported by the US National Institutes of Health (DK080949, OD008469 and AI095277 to J.T.C., and HG004659 and NS075449 to G.W.Y.), the UCSD Digestive Diseases Research Development Center (DK80506), the California Institute for Regenerative Medicine (RB1-01413 and RB3-05009 to G.W.Y.), the National Science Foundation (B.K.), the Alfred P. Sloan Foundation (G.W.Y.) and the Howard Hughes Medical Institute (J.T.C.).

AUTHOR CONTRIBUTIONS

J.A. and J.T.C. designed experiments; J.A., P.J.M. and S.H.K. did experiments. B.K. and G.W.Y. analyzed data; and J.A., B.K., G.W.Y. and J.T.C. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ahmed, R. & Gray, D. Immunological memory and protective immunity: understanding their relation. *Science* **272**, 54–60 (1996).
2. Joshi, N.S. *et al.* Inflammation directs memory precursor and short-lived effector $CD8^+$ T cell fates via the graded expression of T-bet transcription factor. *Immunity* **27**, 281–295 (2007).
3. Masopust, D., Kaech, S.M., Wherry, E.J. & Ahmed, R. The role of programming in memory T-cell development. *Curr. Opin. Immunol.* **16**, 217–225 (2004).
4. Sallusto, F., Lenig, D., Forster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401**, 708–712 (1999).
5. Stemberger, C. *et al.* A single naive $CD8^+$ T cell precursor can develop into diverse effector and memory subsets. *Immunity* **27**, 985–997 (2007).
6. Gerlach, C. *et al.* One naive T cell, multiple fates in $CD8^+$ T cell differentiation. *J. Exp. Med.* **207**, 1235–1246 (2010).
7. Buchholz, V.R. *et al.* Disparate individual fates compose robust $CD8^+$ T cell immunity. *Science* **340**, 630–635 (2013).
8. Gerlach, C. *et al.* Heterogeneous differentiation patterns of individual $CD8^+$ T cells. *Science* **340**, 635–639 (2013).
9. Chang, J.T. *et al.* Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science* **315**, 1687–1691 (2007).
10. Chang, J.T. *et al.* Asymmetric proteasome segregation as a mechanism for unequal partitioning of the transcription factor T-bet during T lymphocyte division. *Immunity* **34**, 492–504 (2011).
11. Kaech, S.M., Hemby, S., Kersh, E. & Ahmed, R. Molecular and functional profiling of memory CD8 T cell differentiation. *Cell* **111**, 837–851 (2002).
12. Best, J.A. *et al.* Transcriptional insights into the $CD8^+$ T cell response to infection and memory T cell formation. *Nat. Immunol.* **14**, 404–412 (2013).
13. Sarkar, S. *et al.* Functional and genomic profiling of effector CD8 T cell subsets with distinct memory fates. *J. Exp. Med.* **205**, 625–640 (2008).
14. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
15. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
16. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011).
17. Lu, R., Neff, N.F., Quake, S.R. & Weissman, I.L. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29**, 928–933 (2011).
18. Ichii, H. *et al.* Role for Bcl-6 in the generation and maintenance of memory $CD8^+$ T cells. *Nat. Immunol.* **3**, 558–563 (2002).
19. Kaech, S.M. & Cui, W. Transcriptional control of effector and memory $CD8^+$ T cell differentiation. *Nat. Rev. Immunol.* **12**, 749–761 (2012).
20. Kallies, A., Xin, A., Belz, G.T. & Nutt, S.L. Blimp-1 transcription factor is required for the differentiation of effector $CD8^+$ T cells and memory responses. *Immunity* **31**, 283–295 (2009).
21. Pearce, E.L. *et al.* Control of effector $CD8^+$ T cell function by the transcription factor Eomesodermin. *Science* **302**, 1041–1043 (2003).
22. Rutishauser, R.L. *et al.* Transcriptional repressor Blimp-1 promotes $CD8^+$ T cell terminal differentiation and represses the acquisition of central memory T cell properties. *Immunity* **31**, 296–308 (2009).
23. Szabo, S.J. *et al.* Distinct effects of T-bet in TH1 lineage commitment and IFN-γ production in CD4 and CD8 T cells. *Science* **295**, 338–342 (2002).
24. Yang, C.Y. *et al.* The transcriptional regulators Id2 and Id3 control the formation of distinct memory $CD8^+$ T cell subsets. *Nat. Immunol.* **12**, 1221–1229 (2011).
25. Zhou, X. *et al.* Differentiation and persistence of memory $CD8^+$ T cells depend on T cell factor 1. *Immunity* **33**, 229–240 (2010).
26. Warren, L., Bryder, D., Weissman, I.L. & Quake, S.R. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc. Natl. Acad. Sci. USA* **103**, 17807–17812 (2006).

© 2014 Nature America, Inc. All rights reserved.

27. Masopust, D., Vezys, V., Marzo, A.L. & Lefrancois, L. Preferential localization of effector memory cells in nonlymphoid tissue. *Science* **291**, 2413–2417 (2001).
28. Schluns, K.S., Kieper, W.C., Jameson, S.C. & Lefrancois, L. Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells *in vivo*. *Nat. Immunol.* **1**, 426–432 (2000).
29. Wherry, E.J. *et al.* Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nat. Immunol.* **4**, 225–234 (2003).
30. van der Maaten, L.J.P. & Hinton, G.E. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. Freund, Y. & Mason, L. The Alternating Decision Tree Learning Algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning* (eds. Bratko, I. & Dzeroski, S.) 124–133 (Morgan Kaufmann Publishers, 1999).
32. Freund, Y. Invited talk: Drifting games, boosting and online learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* 162, (ACM, 2009).
33. Beerenwinkel, N. & Drton, M. A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. *Biostatistics* **8**, 53–71 (2007).
34. Bulla, J. & Bulla, I. Stylized facts of financial time series and hidden semi-Markov models. *Comput. Stat. Data Anal.* **51**, 2192–2209 (2006).
35. Feau, S., Arens, R., Togher, S. & Schoenberger, S.P. Autocrine IL-2 is required for secondary population expansion of CD8$^+$ memory T cells. *Nat. Immunol.* **12**, 908–913 (2011).
36. Williams, M.A., Tyznik, A.J. & Bevan, M.J. Interleukin-2 signals during priming are required for secondary expansion of CD8$^+$ memory T cells. *Nature* **441**, 890–893 (2006).
37. Kalia, V. *et al.* Prolonged interleukin-2Ralpha expression on virus-specific CD8$^+$ T cells favors terminal-effector differentiation in vivo. *Immunity* **32**, 91–103 (2010).
38. Pipkin, M.E. *et al.* Interleukin-2 and inflammation induce distinct transcriptional programs that promote the differentiation of effector cytolytic T cells. *Immunity* **32**, 79–90 (2010).
39. Obar, J.J. & Lefrancois, L. Early signals during CD8 T cell priming regulate the generation of central memory cells. *J. Immunol.* **185**, 263–272 (2010).
40. Afkarian, M. *et al.* T-bet is a STAT1-induced regulator of IL-12R expression in naive CD4$^+$ T cells. *Nat. Immunol.* **3**, 549–557 (2002).
41. Lighvani, A.A. *et al.* T-bet is rapidly induced by interferon-γ in lymphoid and myeloid cells. *Proc. Natl. Acad. Sci. USA* **98**, 15137–15142 (2001).
42. Gebhardt, T. *et al.* Memory T cells in nonlymphoid tissue that provide enhanced local immunity during infection with herpes simplex virus. *Nat. Immunol.* **10**, 524–530 (2009).
43. Masopust, D. *et al.* Dynamic T cell migration program provides resident memory within intestinal epithelium. *J. Exp. Med.* **207**, 553–564 (2010).
44. Olson, J.A., McDonald-Hyman, C., Jameson, S.C. & Hamilton, S.E. Effector-like CD8$^+$ T cells in the memory population mediate potent protective immunity. *Immunity* **38**, 1250–1260 (2013).
45. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
46. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
47. Bendall, S.C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Mice. All animal work was done in accordance with Institutional Animal Care and Use Guidelines of the University of California, San Diego. All mice were housed in specific pathogen–free conditions before use. Wild-type C57/BL6J mice were from the Jackson Laboratory, and OT-I mice that recognize the OVA peptide SIINFEKL (residues 257–264) presented by H-2K^b were used.

Antibodies and flow cytometry. The following antibodies were used: anti-CD8α (53-6.7), anti-CD45.1 (A20), anti-CD62L (MEL-14), anti-KLRG1 (2F1), anti-IFN-γ (XMG1.2), anti-CD44 (IM7), anti-IL-2Rα (PC61), anti-V_α2 (B20.1), anti-CD4 (RM4-5), anti-B220 (RA3-6B2), anti-CD11b (M1/70), anti-CD11c (N418), anti-F4/80 (BM8), and anti-IL-7R (A7R34; all from Biolegend); F(ab′)2 antibody to rabbit IgG (polyclonal; eBioscience); rabbit anti-TCF-1 (C63D9; Cell Signaling Technology; and antibody to human granzyme B (GB11; Life Technologies). For intracellular detection of IFN-γ, CD8$^+$ T cells were stimulated for 4 h at 37 °C *ex vivo* with 0.25 ng/ml SIINFEKL in the presence of brefeldin A (Sigma); cells were fixed in 4% paraformaldehyde (Electron Microscopy Services) and were permeabilized before staining. All samples were analyzed on an Accuri C6 or FACSCanto (BD Biosciences).

Adoptive cell transfer and infection. 5×10^3 OT-I CD45.1$^+$ CD8$^+$ T cells were adoptively transferred into congenic wild-type CD45.2$^+$ recipient mice, followed by intravenous infection 1 d later with 5×10^3 colony-forming units of *L. monocytogenes* expressing full-length chicken OVA (Lm-OVA). Splenocytes were isolated from recipient mice at 5, 7 or 45 d after infection. For the isolation of cells at 3 d after infection, 2×10^4 OT-I CD8$^+$ T cells were adoptively transferred. For the isolation of cells that had undergone their first division, 2×10^6 OT-I CD8$^+$ T cells were first labeled with CFSE (carboxyfluorescein diacetate succinimidyl ester) before adoptive transfer, and recipient mice were killed at 48 h after infection. Cells were stained with fluorochrome-labeled anti-CD8, anti-CD44, anti-CD4, anti-CD11b, anti-CD11c and anti-F4/80 (identified above) and were sorted on a MoFlo (Beckman Coulter) or FACSAria II (BD Biosciences).

Microbead-based enrichment. Magnetic bead–based enrichment was done as described[48]. Single-cell suspensions prepared from infected mice that had received OT-I CD8$^+$ T cells were stained with phycoerythrin-conjugated anti-CD45.1 (identified above), then were washed, stained with anti-PE magnetic microbeads (Miltenyi Biotec) and subjected to enrichment via a magnetic column. Cells were then stained and sorted as described above.

Lymphocyte fate–tracking experiments. Splenocytes from infected recipient mice that had received CFSE-labeled OT-I CD8$^+$ T cells were stained with fluorochrome-conjugated anti-CD8, anti-CD62L and anti-IL-2Rα (identified above). Cells that had undergone their first division (the second-brightest CFSE peak) were electronically gated, and IL-2Rα^{hi}CD62L^{lo} or IL-2Rα^{lo}CD62L^{hi} cells were sorted. 350 cells of each phenotype were adoptively transferred into separate infection-matched CD45.2$^+$ wild-type recipient mice. The progeny of transferred CD45.1$^+$ T cells were monitored throughout the primary response in blood obtained by serial submandibular venipuncture. At 50 d after infection, recipient mice were rechallenged with 5×10^5 colony-forming units of Lm-OVA, and population expansion of the progeny of donor CD45.1$^+$ T cells was tracked in the peripheral blood.

Single-cell gene-expression assays. TaqMan gene-expression assays (20×; Life Technologies) were pooled together in DNA Suspension Buffer (Teknova) to a final concentration of 0.2× for each of the 94 gene-expression assays. Single CD8$^+$ T cells were sorted directly into RT-PreAmp Master Mix (Life Technologies) containing the pooled assays. Cell lysis, sequence-specific reverse-transcription and sequence-specific amplification of cDNA were done as described[14] and high-throughput quantitative PCR was done on 96.96 Dynamic Arrays with a BioMark system (Fluidigm). Cycling threshold values were calculated with BioMark system software. Cells in which the expression of both *Actb* mRNA and *Rn18s* mRNA was detected were retained for further analysis.

Statistical analysis. For statistical analysis, the Kolmogorov-Smirnov test was used for model-free comparisons involving two groups (**Figs. 2b** and **6b–d**). Differences with a P value of <0.05 were considered significant.

Confocal microscopy of T lymphocytes. Immunofluorescence of T cells was analyzed as described[9] with the following antibodies: anti-β-tubulin (AA2; Sigma), anti-IL-2Rα (PC61; eBioscience), anti-CD62L (MEL14; eBioscience), Alexa Fluor 488–conjugated anti-mouse IgG (A11029; Invitrogen) and Alexa Fluor 633–conjugated anti-rat IgG (A21094; Invitrogen). DAPI (4,6-diamidino-2-phenylindole; Life Technologies) was used for the detection of DNA. Cells undergoing cytokinesis were identified by brightfield as those having dual nuclei and a pronounced cytoplasmic cleft. Acquisition of image stacks was done as described[9] with an FV1000 laser-scanning confocal microscope (Olympus). The volume of three-dimensional pixels ('voxels') containing the designated receptor fluorescence was quantified in each nascent daughter in cytokinetic cells as described[9] with ImageJ software.

Data and preprocessing. The log expression of each gene 'g' was computed as follows: $\log E_{g,c} = 40 - Ct_{g,c}$ where c is the cell and $Ct_{g,c}$ is the cycling threshold value obtained from the BioMark (Fluidigm). Cells c' with undefined cycling threshold values ($Ct_{g,c}{}' = 999$) for both $g = Rn18s$ and $g = Actb$, or cells c'' with at least $60 \leq \Sigma_{g=1}^{94} 1\{E_{g,c}{}'' \leq 0\}$ unexpressed genes were also removed from our analyses. The remaining 'good' cells in each population were deemed sufficient for all subsequent analyses, since they exceeded the number of free parameters for any supervised model by a factor of at least 5.

PCA. We used PCA to diminish dimensionality of the data with a linear transformation and projected data X from their original 94 dimensions to the first two principal components. MATLAB software was used for PCA, with the function 'pca'. To visualize the clustering of populations, we projected the cells from their original 94-gene space to the first two principal components of X. Each principal component, also known as 'eigen-gene', captures some percentage of the total variance in X proportional to its corresponding eigen value in the singular value decomposition of X. The first two eigen-genes have the largest eigen values. To visualize the contribution of each original dimension to those eigen-genes, we projected the 94 unit vectors on to the two-dimensional space spanned by the principal components. We combined these projections into the scatter and spike plots presented here (**Figs. 2a**,**c** and **3a**).

***t*-distributed stochastic neighborhood embedding.** To confirm our unsupervised clustering results, we used *t*-distributed stochastic neighborhood embedding (tSNE)[29], one of the most powerful dimensionality reduction methods, with our data set. This is specifically designed for the visualization of high-dimensional data and has been shown to capture more useful variance and more complex clustering patterns in data by attempting to preserve the distances between data points from high dimension to low dimension without any prior assumptions about the distribution of the data. In contrast, PCA captures only linear relationships between genes and principal components and assumes a single homoscedastic (spherical) Gaussian distribution for the entire data set.

JSD. To quantify differences between populations and heterogeneity in each population, we used JSD, a symmetric version of the Küllback-Liebler (KL) divergence, a parameter- and model-free metric of the distance between empirical distributions. Given two sets of experimental measurements, $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$, such as expression profiles for individual cells from the T_{CM} and T_{EM} populations (in this case $x_i \in \mathbb{R}^{94}$), we used JSD to characterize the distance between the two empirical distributions P_x and P_y indicated by the expression by T_{CM} and T_{EM} cells, respectively.

$$JS(P_x, P_y) = \frac{1}{2} KL(P_x \,||\, M) + \frac{1}{2} KL(P_y \,||\, M) \quad (1)$$

$$KL(P \,||\, M) = \int_z P(z) \ln \frac{P(z)}{M(z)} \quad (2)$$

© 2014 Nature America, Inc. All rights reserved.

where $M = (P_x + P_y) / 2$ is an equal mixture of the two distributions, and the KL divergence can be approximated over 'discretized' histograms of its two input distributions:

$$\hat{a}P_x \approx \hat{P}_x(i) = \int_i^{i+1} \quad \text{and} \quad P_y \approx \hat{P}_y(i) = \int_i^{i+1} P_y dy$$

This is the common form of JSD, which does not take into account the group sizes m and n. Instead of using the more general form, which allows arbitrary reweighting of the contribution from each distribution, we randomly subsampled the larger group and concluded that the common form we used was not sensitive to differences in group size when those sizes were within a factor of 2; i.e., $\min(m,n) \geq \max(m,n)/2$.

We interpreted each cell's expression profile as a sample from the 94-dimensional empirical distribution of its population. Expression values for each of the 94 genes is 'discretized' in the same bins, so we simply added the single-dimensional JSD of the two populations for each gene. Moreover, we identified the genes with the greatest and least difference in expression in the two populations, which did not need to match the PCA results exactly, since the JSD analysis does not make the simple linear modeling assumption that PCA does. Finally, to quantify the heterogeneity within a single population, we randomly partitioned the population in half and measured the JSD of the two halves. Averaging that intrapopulation JSD for multiple random partitions gave an estimate of the true variation in each population.

That approach is more principled than a published application of JSD for measuring single-cell diversity, which arbitrarily converted each cell's expression profile into a separate probability distribution over RNA molecules[15]. That was a misrepresentation of the BioMark output, which does not distribute a fixed 'budget' of expression units over the 94 genes of interest but instead measures the doubling times for each PCR primer and can be justified only for single-cell RNA-Seq experiments in which similar numbers of 'reads' are produced for each cell.

Rationale for approach to supervised analysis of gene-expression data. PCA and other unsupervised dimensionality-reduction methods aid in understanding the structure of a cell population. However, such determinations are made by visual inspection. Given a heterogeneous (unsorted) population of cells X', to classify a new cell (i.e., to identify which subpopulation it belongs to), we could cluster together the new samples with existing labeled data in X. Such an approach is suboptimal for two reasons: efficiency and accuracy. That approach is not efficient because to classify even one new cell x' in X', we must redo PCA of the original data X extended by a single row x'. More notably, the accuracy of that approach depends not only on the quality of X but also on that of x', which we are trying to assess. If some of the new samples in X' contain bad or 'noisy' readings that are not filtered by our criteria for X, the variance inherent in X' will eclipse the useful structure observed in X and the coclustering result will be unrelated to or, even worse, counter to the original clustering of X. To resolve both of these problems, we decided to use supervised analysis, which 'learns' to distinguish between subpopulations of the labeled data X in the form of 'boosted' classifiers and applies the classifiers to the remainder of the cells in X.

Robust boosting. We used the RobustBoost algorithm[32] to 'train' an ensemble of 'decision trees' at depths of ≤20. We chose boosting rather than other classification frameworks because the models that are 'learned' are easily interpretable. For example, the alternating decision tree (ADTree) method[31] for the T_{CM}-versus-T_{SLE} classifier (**Supplementary Fig. 4a**) consisted of simple rules by which the expression of *Ptprc*, *Ccl5* and *Sell* was compared with thresholds 'learned' from the data. The classifier's confidence was measured by the margin of each prediction (red bars, **Fig. 4a**). We evaluated the performance of the classifier by its prediction accuracy in 'leave-one-out' cross-validation, where the m classifiers $b_1, b_2, \ldots, b_m$ are each trained on a different subset of $m - 1$ cells. Each classifier b_i was tested on cell x_c, which corresponded to the c^{th} row of the data matrix X, after being trained on the remaining cells $X_{-c} = \{x_1, x_2, \ldots, x_{c-1}, x_{c+1}, \ldots, x_m\}$. Such cross-validation produced a group of m classifiers that provided an estimate of the generalization error $\varepsilon = \Sigma_{c=1}^{m} b_c(x_c) / m$ on the validation examples. This also generated an overall margin $\gamma = \Sigma_{c=1}^{m} \gamma_c$ on the training examples by tallying the predictions of $m - 1$ informed and 1 uninformed classifier for each of the m cells, where $\gamma_c = ||b_{1\ldots m}(x_c) - l_c||$ and l_c is the label of cell c (in this case $l_c = -1$ means T_{CM} and $l_c = 1$ means T_{SLE}).

Temporal model of CD8$^+$ T cell differentiation. Akin to the Heisenberg uncertainty principle, the problem of observing a cell's gene expression is that a cell must be modified (i.e., destroyed) to allow observation of its gene expression. Although it is not a concern for single-cell analysis of static populations, it is a limitation in capturing the dynamics of tracing the lineage of a cell. We propose a statistical modeling approach to overcome that limitation with approximate single-cell histories sampled from the available time-series gene-expression data[33]. We constructed hypothetical differentiation paths and trained an HMM on the resulting expression time courses. Starting from each naive cell, we sampled cells in successively more mature stages whose expression profiles satisfied an ensemble of predictors for one of the terminal fates, matched those samples in the early differentiation stages (division 1 and day 3), connected both ends of each path and finally estimated the transition and observation parameters of a six-state HMM to learn the state-to-state transition probabilities and in-state mixture components that capture the dynamics of gene expression in the hypothetical histories.

Input data. To capture the temporal structure of T cell differentiation in our time-course gene-expression data from single cells, we developed a semi-supervised method based on the fate classifier predictions in early heterogeneous populations (**Fig. 4d**) and on the expression profiles of putative pre-memory and pre-effector cells purified from T_{MP} and T_{SLE} populations by sorting on day 5 after infection. Then, we constructed hypothetical differentiation histories of single cells starting from the naive population, going through an intermediate stage and ending in one of the three terminal fates: T_{CM}, T_{EM} or T_{SLE}. To approximate the real distribution of proliferation transitions between these stages, we used 1,000 bootstrap samples from each subpopulation 'stringed' along one of the three main paths according to their classifier scores. This resulted in an empirical distribution over early transitions (naive to pre-T_{SLE}, and naive to pre-memory) and another distribution over late transitions (pre-memory to T_{CM}, pre-memory to T_{EM}, and pre-T_{SLE} to T_{SLE}). The early transitions were then connected to the late transitions by cells at the intermediate states.

Model structure. Since the differentiation dynamics of individual proliferating T cells are not yet well described, we used an HMM to model the data because of its simple yet powerful structure, which decouples uncertainty in the lineage reconstruction (state transitions) from measurement noise (observations-emissions). We constructed an HMM with six states (naive, pre-memory, T_{CM}, T_{EM}, pre-T_{SLE} and T_{SLE}) to capture the signal in each empirical distribution from our temporal approximation input. Each state 'emits' gene expression from a mixture of two 94-dimensional Gaussian distributions with full covariance matrices.

Because of concerns about our model's sensitivity to initialization, we constructed 18 biologically plausible differentiation pathways (6 sequential and 12 bifurcating; structures, **Supplementary Fig. 4c**) and fixed the transition parameters to the corresponding adjacency matrix of each structure in turn. Using the learning algorithm described below, we calculated the posterior log-likelihood of each pathway. To address any further concerns about the robustness of these results, we reinitialized each structure twice more with 10% random noise drawn from the standard uniform distribution over the [0,1] interval, which also ensured that there were no zero-probability transitions between any two states.

Transition parameters. For a cell c in state f, the probability of transitioning to state t is $T^c_{f,t}$. We assumed that other cells whose expression profile in state f was similar to that of cell c would have similar differentiation potential and, in particular, have a similar probability of transitioning to state t. That assumption allowed us to share the parameters $T_{f,t} = P(f \rightarrow t)$ which gives the probability of any cell in state f to proliferate to state t.

Observation parameters. Because of the bimodal nature of our 'violin plots' (**Fig. 4d**), we modeled the observed expression x of cell c in state i as a mixture model of two Gaussian distributions with 94-dimensional means μ_i^c and η_i^c,

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/ni.2842

94 × 94 full covariance matrices Σ_i^c and Ξ_i^c. Like the transitions, parameter sharing between cells allowed us to simplify the observation parameters, which resulted in the following observation model:

$$P(x \mid s = i) \propto a_i \mathcal{N}(x; \mu_i, \Sigma_i) + b_i \mathcal{N}(x; \eta_i, \Xi_i)$$

Learning algorithm. First, we initialized the model parameters to their prior distributions. Specifically, the transitions $P(f \rightarrow t)$ were initialized to the matrix $T^0_{f,t}$ (**Fig. 5b**). The emission parameters for the naive, T_{SLE}, T_{CM} and T_{EM} states were initialized to the maximum-likelihood fit for a mixture of two Gaussian distributions to the empirical histograms of gene expression for the respective population. The emission parameters in the intermediate states, pre-memory and pre-T_{SLE}, were fit to the empirical histograms accumulated over all intermediate states. The transition parameters were fixed throughout the duration of each learning run but were randomized with up to 10% noise as described above.

Finally, we optimized the parameters of the HMM with the expectation maximization algorithm implemented in pmtk3, the probabilistic modeling toolkit for MATLAB or Octave software[34]. The learned emission parameters were used to identify the genes whose relative expression changed the most during each transition (**Fig. 5b–c**). While we did not learn the transition probabilities, we did resample them from 18 plausible structures and picked the most likely structure (transition matrix, **Supplementary Fig. 4d**; adjacency graph, far bottom right, **Supplementary Fig. 4c**). To determine the most likely structure, we calculated the posterior likelihoods of each HMM (starting from ten random re-initializations) and compared their cumulative distribution functions (**Supplementary Fig. 4c**). To further gauge the statistical significance of the best model, we randomly shuffled the input data 100 times for each model and built a background distribution of the resulting log-likelihoods. To determine both the significance and robustness of each HMM compared with its own shuffled background consisting of 100 random shuffles of the data, we used the nonparametric Kolmogorov-Smirnov method to determine if the log-likelihoods of each model were significantly higher for the real data than for the shuffled data and to provide a *P* value for the significance of each result. To test the reproducibility of our results relative to that of the bootstrap sampling method, we resampled the data and did another 10 randomly initialized training runs on the best model (**Supplementary Fig. 4c**, bottom right). Finally, we did Kolmogorov-Smirnov tests of each proposed structure to ensure that the best model's cumulative distribution function represented significantly higher log-likelihoods than those of the other models or its own background.

48. Moon, J.J. *et al.* Tracking epitope-specific T cells. *Nat. Protoc.* **4**, 565–581 (2009).

© 2014 Nature America, Inc. All rights reserved.