Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1073/pnas.1321405111">http://dx.doi.org/10.1073/pnas.1321405111</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>National Academy of Sciences (U.S.)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Wed Mar 16 08:51:01 EDT 2016</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/89083">http://hdl.handle.net/1721.1/89083</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)

Karthik Shekhar\textsuperscript{a,b,1}, Petter Brodin\textsuperscript{c,d,1}, Mark M. Davis\textsuperscript{e,d,2}, and Arup K. Chakraborty\textsuperscript{b,e,f,g,h,i,2}

\textsuperscript{a}Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; \textsuperscript{b}Ragon Institute of MGH, MIT and Harvard, Boston, MA 02129; \textsuperscript{c}Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94304; \textsuperscript{d}Institute of Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA 94304; and Departments of \textsuperscript{e}Chemical Engineering, \textsuperscript{f}Physics, \textsuperscript{g}Chemistry, and \textsuperscript{h}Biological Engineering, and \textsuperscript{i}Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

Contributed by Mark M. Davis, November 19, 2013 (sent for review October 5, 2013)

Mass cytometry enables an unprecedented number of parameters to be measured in individual cells at a high throughput, but the large dimensionality of the resulting data severely limits approaches relying on manual “gating.” Clustering cells based on phenotypic similarity comes at a loss of single-cell resolution and often the number of subpopulations is unknown a priori. Here we describe ACCENSE, a tool that combines nonlinear dimensionality reduction with density-based partitioning, and displays multivariate cellular phenotypes on a 2D plot. We apply ACCENSE to 35-parameter mass cytometry data from CD8\textsuperscript{T} cells derived from specific pathogen-free and germ-free mice, and stratify cells into phenotypic subpopulations. Our results show significant heterogeneity within the known CD8\textsuperscript{T} T-cell subpopulations, and of particular note is that we find a large novel subpopulation in both specific pathogen-free and germ-free mice that has not been described previously. This subpopulation possesses a phenotypic signature that is distinct from conventional naive and memory subpopulations when analyzed by ACCENSE, but is not distinguishable on a biaxial plot of standard markers. We are able to automatically identify cellular subpopulations based on all proteins analyzed, thus aiding the full utilization of powerful new single-cell technologies such as mass cytometry.

The immune system comprises many cell types that perform highly diverse functions and interact in complex ways during an immune response. The functional capabilities of individual cells are inextricably linked with their phenotypes, as defined by the expression levels of different proteins. These phenotypes are dynamic and alterations often occur, for example during the differentiation of lymphocytes from naive to memory cells upon encountering their specific antigens (1). Understanding which immune cell phenotypes exist is thus important for understanding the functional properties of the immune system as a whole. Flow cytometry, where cells are stained with fluorescently labeled antibodies and their protein targets quantified by light emission signals at single-cell resolution, has been the gold-standard technology for many years (2). Using this technique, hundreds of different immune cell populations have been defined based on differential protein expression. For example, T lymphocytes have been subdivided into helper T cells and killer T cells based on the expression of the coreceptors CD4 and CD8, respectively. In mice, these T-cell populations have also been further subdivided into antigen-naive cells (CD4\textsuperscript{+}CD62L\textsuperscript{+}) and multipotent subpopulations of antigen-naive cells [e.g., central memory (T\textsubscript{CM}, CD4\textsuperscript{+}CD62L\textsuperscript{−}), effector memory (T\textsubscript{EM}, CD4\textsuperscript{+}CD62L\textsuperscript{−}), and short-lived effector cells (T\textsubscript{SLEC}, CD4\textsuperscript{+}KLRG1\textsuperscript{+}CD122\textsuperscript{+})]. Corresponding populations also exist in humans, although the defining markers differ. In both species, these T-cell subpopulations also exhibit functional differences in their proliferative potential, killing capacity, and cytokine production (3).

Flow cytometry is currently constrained to 12–16 parameters per cell due to the limited light spectra and overlapping emission signals. In contrast, mass cytometry allows up to 42 parameters to be quantified on individual cells using metal-chelated probes without any significant signal overlap, thus resolving cellular phenotypes at an unprecedented level of detail (4). Using this technology, Newell et al. recently showed a continuous distribution of human CD8\textsuperscript{T} T-cell phenotypes and a previously unexpected level of functional diversity among these cells (5).

The high-dimensional data (Fig. 1A) generated by mass cytometry are challenging to interpret in biologically meaningful ways. Conventional flow cytometry analysis involves manual analysis through a laborious and highly subjective process known as “gating” (Fig. 1B) (6). As the number of biaxial plots to analyze increases combinatorially with the number of markers analyzed, this process becomes intractable beyond 10–12 parameters. Important advances have been made toward developing better analytic tools for multivariate cytometry data (7). Many of these tools cluster cells with similar protein expression, like the recently developed spanning-tree progression analysis of density normalized events (SPADE) algorithm, which has been applied to mass cytometry data (8, 9). SPADE uses multivariate information to define cellular clusters and displays the underlying phenotypic hierarchy in a tree-like structure. The main drawbacks of clustering approaches are the loss of single-cell resolution and the requirement for prespecification of the number of target clusters desired, introducing bias regarding a quantity that is rarely known.

As an alternative, dimensionality reduction approaches aim at finding low-dimensional representations of high-dimensional data. Important advances have been made toward developing better computational tools for multivariate cytometry data (7). Many of these tools cluster cells with similar protein expression, like the recently developed spanning-tree progression analysis of density normalized events (SPADE) algorithm, which has been applied to mass cytometry data (8, 9). SPADE uses multivariate information to define cellular clusters and displays the underlying phenotypic hierarchy in a tree-like structure. The main drawbacks of clustering approaches are the loss of single-cell resolution and the requirement for prespecification of the number of target clusters desired, introducing bias regarding a quantity that is rarely known.

As an alternative, dimensionality reduction approaches aim at finding low-dimensional representations of high-dimensional data. Important advances have been made toward developing better computational tools for multivariate cytometry data (7). Many of these tools cluster cells with similar protein expression, like the recently developed spanning-tree progression analysis of density normalized events (SPADE) algorithm, which has been applied to mass cytometry data (8, 9). SPADE uses multivariate information to define cellular clusters and displays the underlying phenotypic hierarchy in a tree-like structure. The main drawbacks of clustering approaches are the loss of single-cell resolution and the requirement for prespecification of the number of target clusters desired, introducing bias regarding a quantity that is rarely known.

As an alternative, dimensionality reduction approaches aim at finding low-dimensional representations of high-dimensional data. Important advances have been made toward developing better computational tools for multivariate cytometry data (7). Many of these tools cluster cells with similar protein expression, like the recently developed spanning-tree progression analysis of density normalized events (SPADE) algorithm, which has been applied to mass cytometry data (8, 9). SPADE uses multivariate information to define cellular clusters and displays the underlying phenotypic hierarchy in a tree-like structure. The main drawbacks of clustering approaches are the loss of single-cell resolution and the requirement for prespecification of the number of target clusters desired, introducing bias regarding a quantity that is rarely known.

As an alternative, dimensionality reduction approaches aim at finding low-dimensional representations of high-dimensional data. Important advances have been made toward developing better computational tools for multivariate cytometry data (7). Many of these tools cluster cells with similar protein expression, like the recently developed spanning-tree progression analysis of density normalized events (SPADE) algorithm, which has been applied to mass cytometry data (8, 9). SPADE uses multivariate information to define cellular clusters and displays the underlying phenotypic hierarchy in a tree-like structure. The main drawbacks of clustering approaches are the loss of single-cell resolution and the requirement for prespecification of the number of target clusters desired, introducing bias regarding a quantity that is rarely known.
data to allow easier visualization and interpretation, while retaining single-cell resolution. The spatial organization of datapoints in the low-dimensional space can be used to group cells into subpopulations with similar protein expression. Newell et al. applied principal component analysis (PCA) to 25-parameter mass cytometry data of human CD8⁺ T cells, and used the top three principal components (3D-PCA) to separate subpopulations (5). 3D-PCA represents the data in terms of three summary variables, each a linear combination of the original dimensions, defined so as to maximally capture the underlying variance in the data. That PCA finds the most optimal representation within the set of possible linear projections of the data is, however, also an important limitation—a linear projection may be too restrictive to yield accurate representations (10). To address this limitation, Amir et al. recently applied a nonlinear dimensionality reduction approach to visualize mass cytometry data (11). By using t-distributed stochastic neighbor embedding (or t-SNE) (12), multivariate cellular data could be represented on a 2D plot, similar to conventional biaxial flow plots. However, in contrast with these plots, wherein distance between cells reflects expression differences between only the two markers, distances on the t-SNE plot account for differences across all of the markers. Amir et al. demonstrated that t-SNE could effectively capture phenotypic relationships between cells, such as normal and leukemic bone marrow cells (11).

Here we combine t-SNE with density-based partitioning into a single tool, ACCENSE (Automatic Classification of Cellular Encoding by Nonlinear Stochastic Embedding), and use it to identify murine CD8⁺ T-cell subpopulations (SI Appendix, Tables S1 and S2) from high-dimensional mass cytometry data (SI Appendix 1) without having to predefine the number of expected populations. The work of Newell et al. (5) pertaining to human CD8⁺ T cells inspired us to ask what extent a similar scenario was applicable in laboratory mice, which have been extensively used to advance our understanding of basic immunology over the years. Our analysis not only recovers well-known naive and memory CD8⁺ T-cell populations, but also identifies phenotypically distinct subpopulations within and outside of these. We believe that ACCENSE will be important for exploratory analysis by automatically extracting and quantifying cell populations, based not on only a few, but on the combined expression of the many different proteins measured by mass cytometry.

Results

Computational Methods. Here, we provide a high-level overview of the embedding (using t-SNE) and clustering steps in ACCENSE (see also SI Appendix 2). Let x(0) represent the normalized N-dimensional protein expression vector encoding the phenotype of cell i (i = 1, 2, ..., M cells). We seek corresponding 2D vectors (y(1)) such that T cells with similar phenotype are embedded close to each other in the map, whereas phenotypically dissimilar cells are embedded far apart. t-SNE employs pairwise similarities (p_{ij}) between cells i and j such that p_{ij} is large if x(0) and x(0) are similar, and small otherwise. Let q_{ij} represent the corresponding quantity in the 2D map, encoding similarity between the embeddings y(1) and y(1). The embeddings (y(1)) that maximally conserve the information between the high-dimensional and low-dimensional representations are found by minimizing the Kullback–Leibler divergence (13) between (p_{ij}) and (q_{ij}),

\[ D_{KL}(\{p_{ij}\}, \{q_{ij}\}) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} \]

Thus, we compute a 2D distillation that faithfully preserves neighborhood relationships present in the high-dimensional data. Furthermore, (y(1)) can encode nonlinear relationships because they are not constrained to be linear combinations of (x(0)), as in standard PCA. The optimal embeddings are estimated by a numerical gradient descent procedure (SI Appendix 3), which, owing to the nonconvex objective function in Eq. 1, only guarantees a local minimum. Due to the O(M²) computational and memory complexity of t-SNE, we down-sampled the original dataset in a density-dependent manner (SI Appendix 1.5) to extract a smaller-size “training set,” which we explicitly embedded using the t-SNE algorithm.

Next, we used a kernel-based estimate of the 2D probability density K_{y}(y) (SI Appendix 4, Fig. S1) of cells in the embedding,
where the sum is over the locations of all cell locations in the embedding. Local maxima in $K_r(y)$ correspond to phenotypic subpopulations (Fig. 1 C and D) and were identified using a 2D peak-finding algorithm (14). Upon comparing the results produced by different choices of the kernel-bandwidth $\gamma$, we found a value that provided an accurate coarse-grained representation of the local and global features present in the phenotypic space (see SI Appendix 4 and Fig. S2). Although heuristic, this approach allows us to approximately identify clusters of CD8+ T cells in a data-driven manner without having to pre-specify their number. We also note that directly applying a 35-dimensional kernel to the original space of protein expression data to find cellular subpopulations without first performing dimensionality reduction is fraught with challenges, and is not practical (SI Appendix 2.2).

**Analyzing CD8+ T-Cell Populations in Specific Pathogen-Free Mice Using t-SNE.** CD8+ T cells derived from the blood of six specific-pathogen free (SPF) B6 mice (SI Appendix, Table S1) were assessed for expression of 35 markers, which included cell-surface and intracellular proteins (SI Appendix, Table S3). Each mouse was sampled twice, one of which (S) was stimulated for 5 h with phorbol-12-myristate-13-acetate (PMA) and Ionomycin (SI Appendix 1) while the other sample (U) was analyzed without any treatment. The complete dataset consisted of 36,309 cells, which we down-sampled in a density-dependent manner to obtain a training set of 18,304 cells (see SI Appendix 1.5). Fig. 1C shows the 2D embedding depicting the phenotypic space occupied by SPF mice T cells. The remaining cells were embedded onto this map based on their similarity to the training set (SI Appendix 5), which did not alter the global density profile of the original map (SI Appendix, Fig. S3). Different down-sampled datasets produced qualitatively similar maps.

The contiguous organization of mouse CD8+ T cells in Fig. 1B is consistent with human CD8+ T-cell data (5). The distribution of phenotypes exhibits a high degree of stereotypy, as is expected in these isogenic mice with similar environmental exposure (SI Appendix, Fig. S4). Because our samples were derived from mice of different ages, we also found age-related patterns (vide infra).

Despite the continuous organization of phenotypes, however, the nonuniform distribution of cells in Fig. 1C suggests that not all phenotypes are equally frequent among CD8+ T cells. Density-based partitioning of the t-SNE map identified 24 distinct subpopulations (Fig. 1D; Fig. S2) labeled $S_1$–$S_{24}$ (Fig. 2A). In contrast, projecting the same data along the top two principal components revealed only three distinct subpopulations, with >80% of the cells within one population (SI Appendix 6). Moreover, this representation captured only 21% of the underlying variance, and the spectrum of the covariance matrix indicated that the top 19 principal components altogether captured only 75% of the overall variance in the data (SI Appendix, Fig. S5). These observations underscore the limitations of linear dimensionality reduction and the need for an approach that can account for nonlinearities abundant in cytometry data.

![Fig. 2. ACCENSE identifies CD8+ T-cell subpopulations in SPF B6 mice. (A) Subpopulations $S_1$–$S_{24}$ identified as local maxima in the density map are marked. (B–D) Median expression values of CD44, CD62L, CD122, CD69, CD49d, and Ly6C within the most populous subpopulations that are part of the putative naive, $T_{CM}$, and $T_{EM}$ compartments within SPF CD8+ T cells, identified using ACCENSE. The complete phenotypic signatures of these subpopulations are described in SI Appendix, Figs. S6, S16, and S20, respectively. Bar heights indicate the median expression for each marker within the subpopulation, estimated from cells sampled close to the locations of the subpopulation peaks identified in Fig. 1D (1,500 cells were sampled for the large naive and $T_{CM}$ subpopulations $S_1$ and $S_{12}$, whereas 200 cells were sampled for the smaller $T_{EM}$ subpopulation $S_{10}$). For each marker $k$, the width of the error bar equals $\sigma_{mk}/\gamma$ (the median-absolute deviation (MAD) of the distribution of marker expression within the subpopulation (SI Appendix 7). The blue horizontal bar for each marker corresponds to its median expression $m_k$ across all the cells in the dataset. (E) CD44 vs. CD62L expression of cells in the dataset. (Left) Individual naive ($S_1$, $S_2$), $T_{CM}$ ($S_12$, $S_{13}$), and $T_{EM}$ ($S_{10}$, $S_{11}$) subpopulations are shown in magenta, blue, and green dots, respectively. Gray dots represent T cells from the remaining subpopulations. Dashed lines represent median protein expression. The location of the colored cells belonging to these subpopulations are consistent with their conventionally associated phenotype — e.g., naive, CD44+ CD62L−; $T_{CM}$, CD44+CD62L−; and $T_{EM}$, CD44−CD62L−. (Right) Cells from subpopulation $S_{24}$ shown as red circles, are not (as a group) clearly distinguishable as + or − for either marker. Gray circles represent T cells from all other subpopulations. (F) Distribution of CD44 expression within different subpopulations. "All" represents the CD44 expression distribution across all cells in the data ($M = 36,309$). For the naive (magenta), $T_{CM}$ (blue), and $T_{EM}$ (green) subpopulations, the corresponding distributions largely fail to one side of the population median (dashed black line), and therefore can be unequivocally classified as either a + or − phenotype. In contrast, the CD44 distribution of $S_{24}$ (red) peaks close to the population median and classifies as an "int" phenotype according to our convention.\n
204 | www.pnas.org/cgi/doi/10.1073/pnas.1321405111
Shekhar et al.
Phenotypic Coarse-Graining. We sought to extract the marker expression patterns of each of the CD8+ T-cell subpopulations depicted in Fig. 2A. For each subpopulation we compared the median expression for a marker within that subpopulation to the median expression of the same marker across all cells in the original dataset (see SI Appendix 7). Naively, one might be tempted to label a subpopulation as “+” for a particular marker if its median intrasubpopulation expression is higher than its median expression across all of the cells, and “−” if it is lower. However, such a rigid classification of phenotypes can be misleading for subpopulations identified here based on multivariate protein expression. This is because expression values of a particular marker k within a subpopulation Sk follow a distribution—therefore, labeling the subpopulation strictly according to the subpopulation median $m_k^{(i)}$ will not accurately capture the true phenotype if $m_k^{(i)}$ is close to the population median $m_k$, and if the underlying intrasubpopulation distribution of protein expression is wide (e.g., see the discussion on $S_{20}$ below). To alleviate this, we classified subpopulation phenotypes $S_k$ for which the width of the marker distributions $σ_k^2$ is incorporated in the simplest manner—namely, subpopulation $S_k$ is “+” for marker k if $m_k^{(i)} > m_k + σ_k^2/2$ and “−” for marker k if $m_k^{(i)} < m_k - σ_k^2/2$, else it is “int” (for intermediate). Using three ordinal categories in this manner, which incorporates the first two moments of the marker expression distributions, enables us to achieve a higher degree of precision in cell classification while avoiding the complexity of the entire distribution. The resulting coarse-grained “phenotypic signatures” of $S_{1}−S_{2}$ are shown in SI Appendix, Figs. S6–S29 and summarized in SI Appendix, Table S4. To test whether our procedure yielded subpopulations with distinct phenotypes, we analyzed each pair of subpopulations to compute the number of markers between them that were significantly different (see SI Appendix, Table S5).

Analysis of Identified CD8+ T-Cell Subpopulations in Mice. Phenotypic diversity among naive T cells. CD8+ T cells are conventionally divided into naive, central memory (TCM), and effector memory memory subpopulations (TEM) based on their expression of homing receptors and their propensity to traffic into secondary lymphoid organs (15). Naive T cells are characterized by their low and high expression of the cell-surface glycoprotein CD44 and the homing receptor CD62L, respectively (i.e., CD44−CD62L+). Studies have also associated a low expression of CD122 (subunit of a cytokine receptor), CD69 (an activation marker), CD49d (integrin subunit), and Ly6C (accessory glycoprotein) with naive cells (3). Traditionally, it was believed that naive cells differentiate to acquire a memory phenotype only upon antigen encounter, but this may also occur in the context of homeostatic proliferation (16), or T-cell receptor (TCR) cross-reactivity (17). Subpopulation 1 or $S_1$ on the t-SNE map (Fig. 2A and B) fulfills the traditional criteria for naive marker expression (SI Appendix, Fig. S6). Within the six mice, this subpopulation comprised 12–17% of the cells in the unstimulated samples and 6–8% of the stimulated samples (SI Appendix, Fig. S30A).

Adjacent to $S_1$ in the t-SNE map is $S_2$ (Fig. 2A), which occurred at a frequency of 12–17% in the unstimulated samples and 5–11% in the stimulated samples (SI Appendix, Fig. S30B). $S_3$ also had the “naive” phenotype CD44−CD62L+; nonetheless, it exhibited significant differences compared with $S_1$ across many markers. $S_3$ was CD45RB−CD45RC+, both isoforms of the phosphatase CD45, and CD8α−CD8β+ (SI Appendix, Fig. S7), whereas $S_1$ was either “−” or “int” for these markers. Additionally, the smaller subpopulations $S_4$, $S_5$, and $S_6$ also exhibit the naive phenotype CD44−CD62L+. But had significant differences across other markers compared with $S_1$ (SI Appendix, Figs. S8–S10). Two additional subpopulations $S_7$ and $S_8$, located proximal to the naive subpopulations 1–5 (Fig. 2A), had the intriguing phenotype CD44−CD62L+ (SI Appendix, Figs. S12 and S13). Together these results show a previously unappreciated heterogeneity within the naive CD8+ T-cell compartment.

Memory CD8+ T-cell subpopulations. $S_1$ exhibited the phenotype CD44−CD62L+, which is characteristic of CD8+ TCM cells in mice. Additionally this subpopulation was CD122+CD69+CD49d+Ly6C− (Fig. 2C), which is also consistent with TCM function (3). $S_1$ was present at a much higher frequency in the two 22-mo-old mice (24% and 17%) compared with the two 5-mo-old mice (7% and 9%) within the unstimulated samples (see SI Appendix, Fig. S31E), consistent with the increase of the memory pool with age in humans (16, 18). In the two 7.5-mo-old mice $S_1$ accounted for 22% and 14% of all cells. A neighboring subpopulation $S_2$ had a similar phenotype across all markers but had considerably lower median expression of CD44 and CD122, and higher CD49d compared with $S_1$ (see SI Appendix, Figs. S16 and S17). $S_2$ is also proximal to the “naive-like” subpopulation $S_2$. The organization of the naive subpopulations $S_{1}−S_{2}$ and the TEM subpopulations $S_{1}−S_{2}$ in close proximity to each other is suggestive of a phenotypic continuum of known subpopulations, in line with the human data (5).

Effector memory T-cells (TEM) circulate in the periphery and execute immediate effector functions, typically expressing CD44 but lacking the lymphoid homing receptor CD62L. $S_{15}$ and $S_{17}$ were CD44−CD62L−, consistent with the TEM phenotype (see Fig. 2D, SI Appendix, Figs. S20 and S22). However, 21 markers were significantly different between $S_{15}$ and $S_{17}$ (SI Appendix, Table S5). Interestingly, these subpopulations were embedded further away from the naive subpopulations other than the TEM subpopulations, suggesting that the TEM and naive phenotypes are more similar. Newell et al. also reported a continuous phenotypic progression from naive to TEM and finally to TCM in humans (5).

A large CD8+ T-cell population with CD44int phenotype. The t-SNE map (Fig. 2A) also showed a populous CD44int group $S_{20}$ which was distinct from the naive, TCM, and TEM subpopulations described above. $S_{20}$ was present at a frequency of 11.5–18.5% in all unstimulated samples, with a discernible decrease in older animals (SI Appendix, Fig. S33B). Compared with the naive subpopulation $S_1$ these T cells had, most notably, a significantly reduced expression of CD8β and (to a lesser extent) CD8α, and the CD45RB and -RC isoforms, while having a significantly up-regulated expression of TCRβ, CD3e, and CD5, as well as other markers (see SI Appendix, Figs. S6 and S25). The organization of subpopulations depicted in Fig. 2A clearly shows that $S_{20}$ is phenotypically distinct from the canonical naive, TCM, and TEM subpopulations. Interestingly, when we focused only on the expression of markers conventionally used to classify cells into naive and memory phenotypes, the distinctiveness of $S_{20}$ was less obvious. In particular, we observed that the median expression within $S_{20}$ of CD44, the typical naive–memory distinguishing marker, falls in an intermediate range (Fig. 2E, Right and Fig. 2F) in contrast with the conventional naive, TCM, and TEM subpopulations (Fig. 2E, Left). In a standard biaxial plot involving CD44 and CD62L, $S_{20}$ can be conflated with the conventional subpopulations despite being phenotypically distinct. This example illustrates the value of incorporating information across multiple markers while defining phenotypic subpopulations of cells.

Stimulation-associated phenotypes. Because we stimulated the T cells with PMA–Ionomycin, we were able to address phenotypic and functional changes resulting from this broad and unspecific stimulus. The $S_{10}$, $S_{14}$, and $S_{20}$ subpopulations, which were proximal in the t-SNE map to $S_1$, $S_{15}$, and $S_{20}$, respectively, were present in the stimulated samples in high proportions, but occurred with negligible frequency in the unstimulated controls across all mice (SI Appendix, Figs. S31D, S32B, and S33F). $S_{10}$ was also the most populous subpopulation in the stimulated samples occurring at a frequency of 20–50% in all mice. These responsive subpopulations were characterized by significantly
up-regulated expression of CD69, IL-2, CD107a (a marker of degranulation upon stimulation), CTLA-4, and loss of CD62L. That $S_{10}$, $S_{14}$, and $S_{24}$ are closer to $S_1$, $S_{11}$, and $S_{20}$ than to each other suggests that T cells might still retain their basal phenotypic characteristics across most markers, and only a few (e.g., CD69, IL-2, etc.) might change upon activation through such an unspecific stimulus. The degree of phenotypic change upon activation has not been addressed previously at this level of detail to the best of our knowledge.

Comparison of CD8 T-Cell Subpopulations Between SPF and Germ-Free Mice. Recent studies have found developmental defects in CD8$^+$ T cells in germ-free (GF) mice, which lack commensal bacterial microbiota (19). We wanted to obtain a more detailed view of the phenotypic profile of T cells in GF mice compared with the SPF mice described above. We collected protein expression data using the same staining panel (SI Appendix, Table S3) and focused on CD8$^+$ T cells ($M = 4,086$) derived from the blood of three GF mice (SI Appendix, Table S2). As in any automated cytometry analysis technique the presence of “batch effects” due to variations in staining intensity from day to day precluded direct merging of the SPF and GF datasets (SI Appendix 1.4). Instead, we elected to work around batch effects by computing a separate map for the GF data, which revealed eight subpopulations labeled $S_1-8$ (see Fig. 3A). Each subpopulation was phenotypically classified as “+”, “−” or “int” for each protein (see SI Appendix, Table S6). We then compared the presence or absence of specific CD8$^+$ T-cell subpopulations in GF mice with those identified in SPF mice. Importantly, we did not compare a GF subpopulation $S_i$ with an SPF subpopulation $S_j$ by the absolute marker expression values (which we expect to be corrupted by batch effects), but rather by their coarse-grained phenotypic signature across each marker (i.e., “+,” “−,” “int”).

For a GF subpopulation $S_i$ and SPF subpopulation $S_j$, we computed the phenotypic similarity $\text{Sim}(S_i, S_j) = \frac{1}{n} \sum_{k \in \text{markers}} \delta_{S_i,S_j}^k$, where $n = 35$ markers. Here $\delta_{S_i,S_j}^k = 1$ if $S_i$ and $S_j$ have the same coarse-grained phenotype for marker $k$, else $\delta_{S_i,S_j}^k = 0$. $\text{Sim}(S_i, S_j) = 1$ if the subpopulations have the same phenotype across all markers and 0 if they share no common phenotypes.

Fig. 3B, which depicts the matrix $\text{Sim}(S_i, S_j)$, suggests that the three most populous GF subpopulations—$S_1$, $S_4$, and $S_5$—have the greatest similarity (>75%) to the three most populous SPF subpopulations—$S_1$ (naive), $S_{11}$ ($T_{CM}$), and $S_{20}$, respectively. Notably, the corresponding $T_{CM}$ subpopulations $S_i$ and $S_{11}$ (both CD44$^+CD62L^-$) had >95% phenotypic similarity across all of the 35 markers analyzed. $S_2$ and $S_5$ shared >70% phenotypic similarity with $S_8$ and $S_{10}$, and likely represent phenotypic sub-divisions within the naive compartment in GF mice. $S_8$, the fourth most populous GF subpopulation, had a phenotype CD44$^+CD62L^+$ which was similar to $S_8$ in SPF mice. Our finding of a large $T_{CM}$ subpopulation is consistent with previous reports that T cells with a memory phenotype are found in GF mice (20). Intriguingly however, none of $S_1-8$ had significant overlap with the $T_{EM}$ subpopulations in SPF mice ($S_15$ or $S_17$), hinting at the possibility that such cells may be reduced or absent in the blood of GF mice.

Taken together, these results show a great degree of similarity in the phenotypic profile of blood CD8$^+$ T cells between SPF and GF mice. We have shown that the coarse-grained phenotypic signatures can be effectively used to compare cellular expression data collected across different experiments, where expression signals may be corrupted by batch effects. Whereas the number of subpopulations revealed by the t-SNE maps was fewer in GF mice, and an effector memory subpopulation was completely absent, further studies are required to fully establish whether or not these differences are due to the presence or absence of commensal microbiota.

Discussion

Conventional approaches that rely on the manual interpretation of a large number of biaxial plots are unscaleable in the context of multiparametric protein expression data across millions of cells. Their reliance on a handful of markers, one or two at a time, can potentially conflate populations that are phenotypically distinct based on multivariate protein expression patterns.

Recently, Amir et al. combined t-SNE with mass cytometry and demonstrated the potential of nonlinear dimensionality reduction in revealing important biological relationships in bone-marrow and leukemia datasets (11). The resulting output is a 2D map, where cells are organized according to their phenotype, taking into account the full protein expression vector in determining their relative positions. Cells with similar protein expression are embedded close to each other in the map and, unlike PCA, this representation effectively captures nonlinear relationships in the high-dimensional data.

Fig. 3. GF subpopulations and their phenotypic similarity with SPF subpopulations. (A) t-SNE map computed from CD8$^+$ T cells derived from the blood of GF mice ($M = 4,086$) indicating subpopulations $S_1-S_8$. (B) Phenotypic similarity $\text{Sim}(S_i, S_j)$ depicted as a heatmap. Rows are subpopulations $S_i$ in GF mice and columns are subpopulations $S_1-S_{24}$ in SPF mice. The color of the pixel at location ($i$, $j$) represents the fraction of marker phenotypes that are similar across $S_i$ and $S_j$. 

206 | www.pnas.org/cgi/doi/10.1073/pnas.1321405111

Shekhar et al.
Here, we have extended this and proposed a systematic framework for identifying phenotypic subpopulations from high-dimensional mass cytometry data and proposed a straightforward way to coarse-grain their phenotypes automatically, wherein the expression of a specific protein in a particular subpopulation is classified into one of three ordinal categories. Using this tool we termed ACCENSE, we identify phenotypic subpopulations in SPF and GF mice in a data-driven manner without directly prespecifying the number of clusters, and showed that phenotypic coarse-graining allows for comparison of data across batches, even when variations in instrument performance can preclude directly merging datasets for quantitative analysis.

Applying ACCENSE to cells derived from the blood of SPF mice enabled us to recover well-known CDSφ T-cell subpopulations—naive, central, and effector memory, and PMA–Ionomycin stimulation-associated. Phenotypic signatures within markers associated with these subpopulations are consistent with conventional wisdom. The presence of heterogeneous subpopulations and a continuous distribution of cells in the phenotypic space of CDSφ T cells in murine blood are consistent with the findings on human blood CDSφ T cells (5). Our map also elucidated further phenotypic subdivisions within these subsets based on the other markers analyzed. The most populous SPF subpopulations had clear analogs in GF mice, thereby demonstrating a great degree of phenotypic similarity at least among blood-derived CDSφ T cells in these mice. We were, however, unable to detect an effector memory population in the GF samples.

In both SPF and GF mice, we identified a large T-cell subpopulation (Sφ/Sφ) which was distinct from the naive and the memory subpopulations on the t-SNE map and was characterized by intermediate expression of CD44 and CD62L, but distinctly lower expression of the CD8β-chain and CD45 isoforms. These lower expression levels of the CD8β-chain and CD45 isoforms would have escaped notice as a distinct phenotype using standard CD44 vs. CD62L gating, as illustrated in Fig. 2E.

This population could either represent T cells wherein the coreceptor CD8 exists predominantly as an α- or β-homodimer or something completely different. Additional experiments are required to test these possibilities and assess their functional consequences. This result illustrates the potential of taking multivariate protein expression into account while assigning cells to phenotypic subpopulations.

The studies initiated here provide an important advance in the automatic classification of subpopulations from high-dimensional protein expression data, an important challenge in cytometry. Many algorithms exist for the automated analysis of flow cytometry, but generally use only four parameters at most and none of these has yet been applied to mass cytometry (7). The ACCENSE program described here provides a simple way to detect discrete populations in reduced dimensional space. We believe that this will help the full potential of mass cytometry data to be realized.

**Materials and Methods**

*SI Appendix* includes detailed descriptions of mouse experiments, mass cytometry, and data-preprocessing (*SI Appendix, Sec. 1*), mathematical details of t-SNE and density-based subpopulation identification (*SI Appendix, Secs. 2–5*), PCA on SPF data (*SI Appendix, Sec. 6*), and phenotypic signatures of SPF and GF subpopulations (*SI Appendix, Sec. 7*). Implementations of ACCENSE in MATLAB and R are freely available on the website, www.cellaccense.com.

**ACKNOWLEDGMENTS.** We are grateful to Dr. G. Berman and Prof. J. Shavevitz for sharing unpublished work, and to Prof. A. Ferguson for a critical reading of the manuscript. We thank Prof. J. Sonnenburg for providing the GF mice. This research was supported by a Poitras pre-doctoral fellowship (K.S.), the Ragon Institute of MGH, MIT and Harvard (K.S. and A.K.C.), the Wenner-Gren Foundation and the Swedish American Foundation (P.B.), U189 Al 090019 (to M.M.D.), and a National Institutes of Health PO1 AI091580 (to A.K.C.).