The Journal of Immunology

RESEARCH ARTICLE | AUGUST 01 2020

A Comprehensive Workflow for Applying Single-Cell Clustering and Pseudotime Analysis to Flow Cytometry Data ✓

Janine E. Melsen; ... et. al

J Immunol (2020) 205 (3): 864–871.

https://doi.org/10.4049/jimmunol.1901530

Related Content

An Integrated Workflow To Assess Technical and Biological Variability of Cell Population Frequencies in Human Peripheral Blood by Flow Cytometry

J Immunol (February,2017)

A Comprehensive Workflow for Applying Single-Cell Clustering and Pseudotime Analysis to Flow Cytometry Data

Janine E. Melsen,* Monique M. van Ostaijen-ten Dam,* Arjan C. Lankester,*
Marco W. Schilham,*,[1] and Erik B. van den Akker†,‡,[1]

The introduction of single-cell platforms inspired the development of high-dimensional single-cell analysis tools to comprehensively characterize the underlying cellular heterogeneity. Flow cytometry data are traditionally analyzed by (subjective) gating of subpopulations on two-dimensional plots. However, the increasing number of parameters measured by conventional and spectral flow cytometry reinforces the need to apply many of the recently developed tools for single-cell analysis on flow cytometry data, as well. However, the myriads of analysis options offered by the continuously released novel packages can be overwhelming to the immunologist with limited computational background. In this article, we explain the main concepts of such analyses and provide a detailed workflow to illustrate their implications and additional prerequisites when applied on flow cytometry data. Moreover, we provide readily applicable R code covering transformation, normalization, dimensionality reduction, clustering, and pseudotime analysis that can serve as a template for future analyses. We demonstrate the merit of our workflow by reanalyzing a public human dataset. Compared with standard gating, the results of our workflow provide new insights in cellular subsets, alternative classifications, and hypothetical trajectories. Taken together, we present a well-documented workflow, which utilizes existing high-dimensional single-cell analysis tools to reveal cellular heterogeneity and intercellular relationships in flow cytometry data. ***The Journal of Immunology*****, 2020, 205: 864–871.**

Flow cytometry data are traditionally analyzed by manual and sequential gating of subpopulations on two-dimensional plots. This approach is highly dependent on the user's interpretation and knowledge and is time-consuming. Moreover, it critically underappreciates the full spectrum of naturally occurring variation in coexpression and intensity expression of markers and thus seems insufficient to capture the full underlying cellular complexity.

The introduction of new experimental platforms such as single-cell RNA sequencing and mass cytometry (1), which can acquire >20,000 and 40 parameters per cell, respectively, has advanced the development of high-dimensional data analysis tools (2). In flow cytometry, technological advances in equipment and development of new fluorescent dyes have led to a major increase in dimensionality of the acquired datasets. With the introduction of new conventional and spectral flow cytometers, >30 parameters can be simultaneously measured (3, 4). Hence, the need to apply the existing single-cell analysis tools on flow cytometry datasets increases.

In general, two analytical approaches have been developed to capture the underlying cellular heterogeneity. The first approach aims to define phenotypically similar cells by clustering. Clustering can be achieved by applying general clustering methods such as hierarchical clustering or K-means or methods developed for cytometry data (5) [e.g., Gaussian mean shift (GMS) clustering (6) as implemented in Cytosplore, FlowSoM (7), PhenoGraph (8)] or single-cell RNA-sequencing data (9) [e.g., *k*-nearest neighbor–based Louvain clustering (10), as implemented in Seurat (11)]. In contrast, the second analytical approach assumes a continuum of cellular states and aims to reveal cellular progression by inferring cellular trajectories, called pseudotime analysis. This type of analysis is particularly useful for studying cellular differentiation or disease progress. The method has been developed for single-cell RNA-sequencing data by Monocle (12). Nowadays, >70 methods are available and reviewed (13), including Slingshot (14), Wishbone (15), and partition-based graph abstraction (16). Hence, depending on the assumptions and purposes for the data, current analytical approaches are either aimed at discovering distinct subsets of cells or model cells as a differentiating continuum. In practice, analytical methods are often applied consecutively to progressively unravel the structure in single-cell datasets.

Visualization of high-dimensional data requires dedicated methods, so-called dimensionality reduction methods, to comprehensively represent the cellular heterogeneity assessed by many parameters into a two-dimensional scatterplot. To date, a wide array of dimensionality reduction methods are available and already extensively reviewed elsewhere (5, 17, 18). For instance, *t*–stochastic neighbor embedding (19), hierarchical stochastic neighbor embedding (HSNE; as implemented in Cytosplore) (20, 21), and uniform manifold approximation and projection (UMAP) (22, 23) are commonly applied to visualize results of cluster analysis. The result of pseudotime analysis is preferably visualized in a reduced

*Department of Pediatrics, Leiden University Medical Center, 2333 ZA Leiden, the Netherlands; †Department of Biomedical Data Sciences, Leiden University Medical Center, 2333 ZC Leiden, the Netherlands; and ‡Pattern Recognition and Bioinformatics Group, Delft University of Technology, 2628 XE Delft, the Netherlands

[1]M.W.S. and E.B.v.d.A. contributed equally to this work.

ORCIDs: 0000-0001-5322-7194 (J.E.M.); 0000-0003-4391-6003 (M.W.S.).

Received for publication December 30, 2019. Accepted for publication June 1, 2020.

J.E.M. was supported by fellowships from the Leiden University Medical Center, the graduate program of Nederlandse Organisatie voor Wetenschappelijk Onderzoek, and a grant from Stichting Zeldzame Ziekten Fonds (31545).

Address correspondence and reprint requests to Janine E. Melsen, Department of Pediatrics, Leiden University Medical Center, PO Box 9600, Albinusdreef 2, 2333 ZA Leiden, the Netherlands. E-mail address: j.e.melsen@lumc.nl

The online version of this article contains supplemental material.

Abbreviations used in this article: CM, central memory; EM, effector memory; EMRA, EM reexpressing CD45RA; GMS, Gaussian mean shift; HSNE, hierarchical stochastic neighbor embedding; UMAP, uniform manifold approximation and projection.

www.jimmunol.org/cgi/doi/10.4049/jimmunol.1901530

dimensional space, which orders cells along a trajectory, such as diffusion map (24, 25). Overall, visualization by dimensionality reduction forms the key element of each analysis of single-cell data.

To reveal in detail the immunological landscape at the single-cell resolution in large flow cytometry datasets, we have compiled a comprehensive workflow illustrated with readily applicable R code, accessible to the immunologist with limited computational background. Specifically, we offer a workflow for high-dimensional single-cell analysis from data preprocessing to visualization, clustering, and pseudotime analysis and include a description of limitations and potential pitfalls.

Materials and Methods

Data

The flow cytometry dataset FR-FCM-ZYQ9 was downloaded from the FlowRepository Web site (26, 27). Eight healthy bone marrow donors were selected for further analysis of the T cells. The panel included the following markers: a live/dead marker (Aquablue), CD14+CD19 (PE-Dazzle), CD3 (BV785), CD4 (BV605), CD8 (APC-Fire), CCR7 (BV421), CD45RA (BV650), CD27 (FITC), CD95 (APC), CD49b (PE), CD69 (PE-Cy5), CD103 (PE-Cy7), and CXCR4 (PerCP-Cy5.5). CXCR4 was removed from the analysis because we observed no reliable CXCR4 staining. Donor B (47 y), L (57 y), and M (60 y) were measured on a different day than donor F (41 y), K (84 y), N (67 y), R (31 y), and W (28 y). The complete workflow as discussed below is depicted in Fig. 1.

Preprocessing

The fcs files were manually compensated in the conventional analysis software Kaluza (v2.1; Beckman Coulter, Brea, CA). Next, the following gating strategy was applied. Lymphocytes were gated based on forward and side scatter, dead cells were excluded by the live/dead staining, doublets were excluded by plotting the width and height of the forward and side scatter, and finally T cells were gated as CD14$^-$CD19$^-$CD3$^+$ living single lymphocytes (Supplemental Fig. 1). The T cell data were exported as csv files and imported into R (v3.6; R Foundation for Statistical Computing, Vienna, Austria). Transformation was applied on the eight samples (Supplemental Fig. 2). We compared our manual hyperbolic arcsine (arcsinh) transformation to arcsinh transformation with cofactor 150 and to automated parameter optimized transformations, as implemented in FlowCore (logicle) (28), FlowVS (arcsinh) (29), and FlowTrans (arcsinh) (30) (Supplemental Fig. 2). After transformation, normalization was applied by either gaussNorm or fdaNorm as implemented in the FlowStats package (31) to correct for technical intersample variation (Supplemental Fig. 3). The transformed and normalized expression values were exported to fcs files by use of the FlowCore package. The R scripts and fcs files are available on GitHub (https://github.com/janinemelsen/Single-cell-analysis-flow-cytometry).

Dimensionality reduction

Three dimensionality reduction methods were used: HSNE, diffusion map [as implemented in the destiny package (25)], and UMAP [as implemented in the uwot package (32)]. For HSNE, we imported the fcs files containing the T cells from the eight donors (in total 618,288 cells) in Cytosplore (21) and performed a four-level HSNE with default parameters based on the expression levels of CD3, CD4, CD8, CCR7, CD45RA, CD27, CD95, CD49b, CD69, and CD103.

Clustering

As clustering methods, we applied GMS clustering (in Cytosplore), FlowSOM, and PhenoGraph. The HSNE-based GMS clustering with a σ of 20 was performed to identify cell clusters. Phenotypically similar clusters were manually merged, to avoid overclustering. The 14 clusters identified in the CD4$^+$ T cell compartment in the second level were exported as individual fcs files for further analyses.

Pseudotime

The 14 fcs files from the CD4$^+$ T cell clusters (in total 275,856 cells) were exported from Cytosplore and imported in R. Next, we ran Slingshot (14) on the transformed CD4$^+$ T cell dataset. Based on the HSNE-based GMS clusters, the minimum spanning tree was calculated to identify lineages. We specified the naive CD4$^+$ T cells as the initial cluster. The pseudotime variable was inferred by fitting simultaneous principal curves. Visualization was achieved by plotting the pseudotime variable as a color scale on a diffusion map. The R code is available on GitHub (https://github.com/janinemelsen/Single-cell-analysis-flow-cytometry).

Results

To apply high-dimensional single-cell analysis on flow cytometry data, we developed a workflow that consists of four sections: data preprocessing (compensation, export cells of interest, transformation, normalization), dimensionality reduction, clustering, and pseudotime analysis (Fig. 1). To enable readers to perform the analyses themselves, we demonstrated our workflow by reanalyzing publicly available flow cytometry data downloaded from FlowRepository (FR-FCM-ZYQ9) and comparing the results with the results obtained after standard gating. Bone marrow mononuclear cells from eight healthy donors, which were measured on two different days, were included for further analysis. We selected the T cell panel, which includes a live/dead marker, CD14, CD19, CD3, CD4, CD8, CCR7, CD45RA, CD27, CD95, CD49b, CD69, and CD103.

Preprocessing

The initial preprocessing steps (compensation and exporting cells of interest) of a data-driven analysis of flow cytometry data are not different from the traditional approach and can be performed in conventional gating software.

Compensation. We corrected for spillover of fluorochromes by applying compensation. It should be noted that the compensation can differ between days but also between samples that were measured on the same day.

Export cells of interest. After compensation, dead cells and doublets need to be excluded. Lymphocytes were gated based on forward and side scatter, dead cells were excluded by the live/dead staining, doublets were excluded by plotting the width and height of the scatters, and for illustrative purposes, T cells were gated as CD14$^-$CD19$^-$CD3$^+$ (Supplemental Fig. 1). In total, we exported 618,288 T cells as csv from eight samples.

Transformation. Conventional gating software often visualizes data on biexponential axes, meaning that lower (negative) values are plotted on a linear scale, whereas higher (positive) values are plotted on a logarithmic scale. This standard transformation, however, is generally not applied on the exported data. Hence, our workflow starts with options for data transformation.

Automated transformations as implemented in the FlowVS (29), FlowCore (28), or FlowTrans package (30) were applied on the data. In contrast to FlowTrans, FlowVS and FlowCore calculate the optimal transformation per parameter and work satisfactorily in our hands, except for markers that have low-intensity expression or are expressed at low frequency. For instance, the parameters CD103, CD49b, and CCR7 exhibit artificial positive peaks after automated transformation (Supplemental Fig. 2A–C). Hence, in our opinion, a manual inspection (and if necessary, adjustment) of the applied transformations remains essential.

In parallel, we conducted a manual transformation of the data for comparison with the automated transformation methods (Supplemental Fig. 2D). We choose to apply a manual arcsinh transformation on the data, which serves a similar purpose as the biexponential transformation and allows tuning of the linear region around zero by adjusting the cofactor. The cofactor of the arcsinh transformation equals the size of the linear region on the positive or negative side of the zero (Fig. 2A). Whereas for mass cytometry–derived datasets, the arcsinh transformation is commonly applied with a cofactor of 5; for flow cytometry data, a cofactor of 150 is often applied (5, 33).

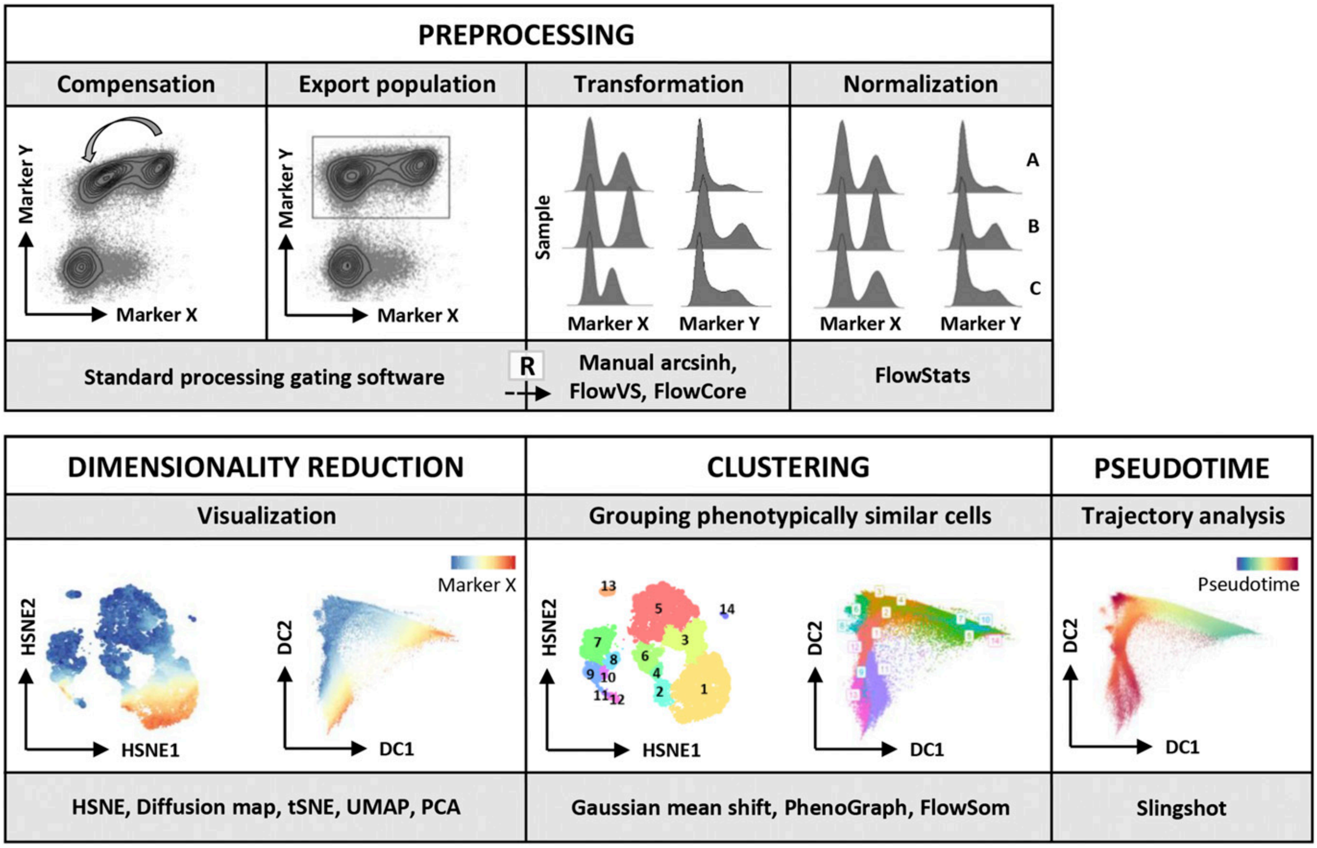

FIGURE 1. Overview workflow. The preprocessing of the data includes compensation, exporting the population of interest, transformation (either with a manual determined cofactor or automated), and normalization to correct for technical intersample variation. Compensation and gating is standard procedure in analysis of flow cytometry data. Hereafter, the data are imported in R. The transformed and normalized data are used as input to define phenotypically similar cells by clustering, visualized in a reduced dimensional space. For both dimensionality reduction and clustering, multiple methods are available. In this figure, we demonstrate HSNE with GMS clustering as implemented in Cytosplore and clusters calculated by FlowSOM, projected on a diffusion map. As an alternative to clustering, cellular trajectories can be studied by pseudotime analysis as implemented by Slingshot.

However, transforming a parameter with a fixed (yet in our hands, generally too low) cofactor of 150 can result in a false-positive peak (Fig. 2A third panel, Supplemental Fig. 2E).

Because each fluorochrome has a distinct staining pattern, each parameter requires an individualized transformation. By adjusting the linear size of the axis until an optimal distribution of the peaks is observed (which can be done in conventional gating software), the optimal cofactor can be determined. The negative peak of a histogram should be positioned in the linear region of the biexponential scale. If the cofactor is set too high, the positive peak will be artificially positioned in the linear region, resulting in a false-negative peak. If the cofactor is set too low, the negative peak will be artificially positioned in the logarithmic region, resulting in a false-positive peak (Fig. 2A).

Optimal cofactors could differ slightly between samples that were measured on different days. For instance, the distribution of the data of donor F and M, which were measured on two distinct days, varied for CCR7 (Fig. 2B). In this case, the highest cofactor can be applied on both samples as long as no false-negative peak is observed (Fig. 2B, Supplemental Fig. 2D). In this manner, the samples are equally transformed.

Normalization. As a result of interexperiment variability, variation in staining intensity can be observed. For instance, use of a new Ab lot, different temperature staining conditions, or flow cytometer variability can lead to technical variation in the data. Therefore, interexperiment variability does require normalization, but biological variation should be conserved. We observed for CD49b, CD4, CD27, CCR7, CD8, and CD3 subtle differences in signal intensities between the donors and therefore applied normalization by either fdaNorm or gaussNorm as implemented in the FlowStats package (31) (Fig. 2C, Supplemental Fig. 3). Because fdaNorm automatically detects the number of peaks present in the datasets, we prefer the use of fdaNorm. After normalization of the transformed data, the FlowCore package was used to export the data as fcs again (28).

Dimensionality reduction

For visualization purposes, the total number of dimensions (which is equal to the number of parameters) needs to be reduced to two. A wide range of methods is available, with each having its own limitations and strengths. Datasets acquired by flow cytometry and mass cytometry often include millions of cells. Because conventional *t*–stochastic neighbor embedding can only handle 150,000 cells, the HSNE has recently been introduced to eliminate the need for downsampling (20). The strength of HSNE is the presentation of millions of cells in multiple levels of clusters in limited computational time. At the overview level, major lineages can be identified, whereas at the deeper levels, phenotypical details of subpopulations can be revealed. HSNE has been implemented in Cytosplore, which allows interactive exploration of the data (21) (Fig. 3A, 3B). In addition, we will demonstrate the use of diffusion map and UMAP (Fig. 4A, 4B, Supplemental Fig. 4). In contrast to HSNE, both diffusion map and UMAP better preserve the global structure of the data and are therefore more suitable for

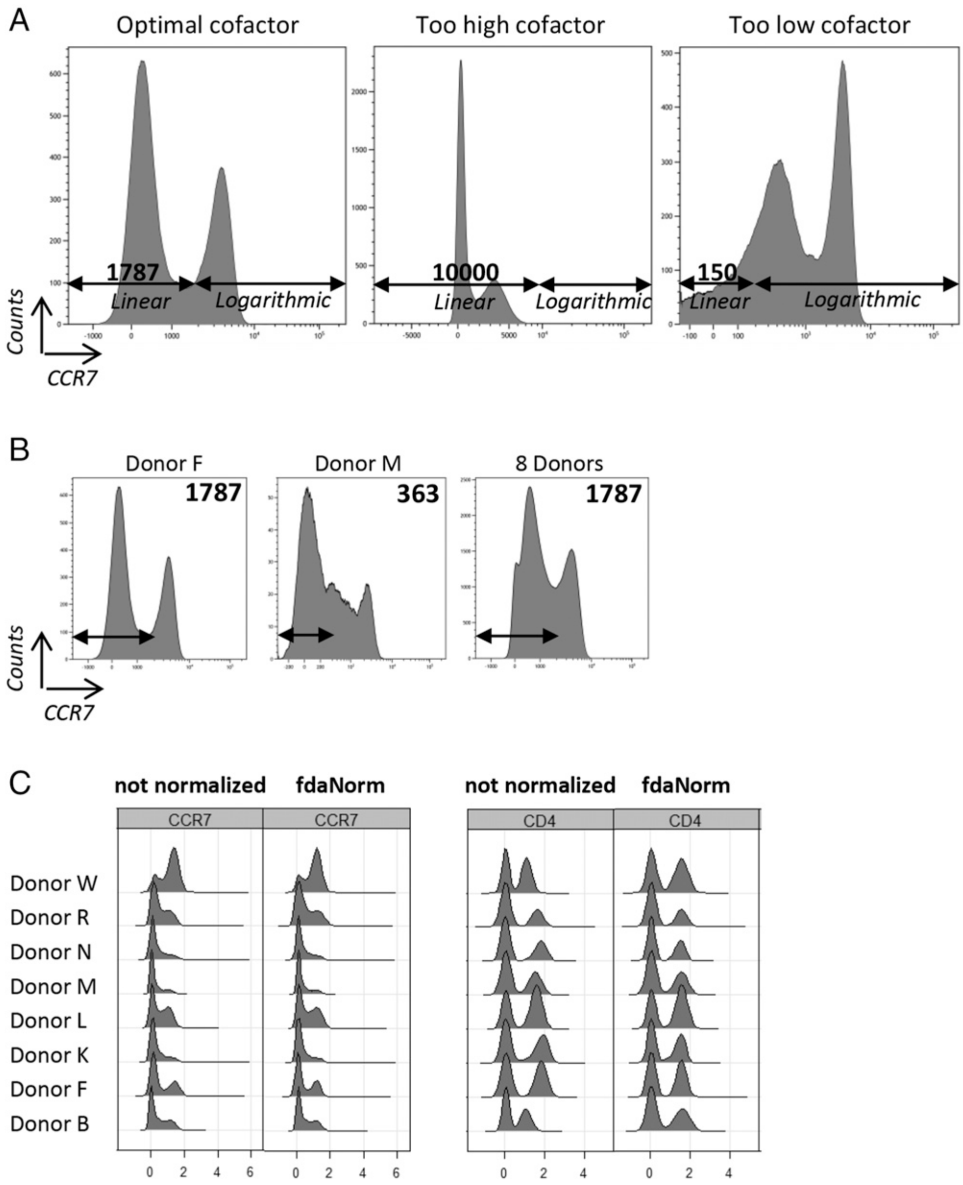

FIGURE 2. Transformation and normalization. (**A**) The CCR7 expression on the T cells of donor F are shown on three distinct biexponential scales. By tuning the linear region around zero on the biexponential scale, the optimal cofactor of the arcsinh transformation can be determined. The size of the linear region on the positive or negative side of the zero equals the cofactor (1787, 10,000, or 150 are shown). Setting the cofactor too high results in a false-negative peak, whereas setting the cofactor too low results in a false-positive peak. (**B**) The cofactor can differ between samples measured on different days. The expression of CCR7 is shown with an optimal cofactor of 1787 for donor F and 363 for donor M. To equally transform the samples and reduce intersample variation, the highest cofactor can be applied on all samples. (**C**) To correct for signal intensity differences between the donors, we applied normalization by using fdaNorm, as implemented in FlowStats. As an example, CCR7 and CD4 are shown pre- and postnormalization.

visualization of cellular progresses. We will further discuss these visualization tools in the clustering and pseudotime section.

Clustering

HSNE-based GMS clustering. We imported the compensated, transformed, and normalized fcs files of the eight donors into Cytosplore and performed an HSNE analysis on a total of 618,288 cells based on all parameters, except the forward scatter, side scatter, live/dead marker, CXCR4, CD14, and CD19. In the overview level, we identified the CD4$^+$ T cells and CD8$^+$ T cells (Fig. 3A). We zoomed in on the CD4$^+$ T cells and generated 14 clusters at the second level, guided by the GMS clustering based on the density representation of the embedding (Fig. 3B, 3D).

We identified CD49b$^+$ and CD49b$^-$ naive T cells (CCR7$^+$CD45RA$^+$CD95$^-$, cluster 1, 2) and 12 subsets of memory T cells, including the CD27$^+$ central memory (CM) T cells (cluster 3, 4), CD27$^+$ effector memory (EM) T cells (cluster 5, 6), CD27$^-$ EM T cells (cluster 7, 8, 12), CD27$^-$ terminally differentiated EM reexpressing CD45RA (EMRA) T cells (cluster 9, 10, 11), CD69$^+$CD103$^-$ EM T cells (cluster 13), and CD103$^+$ EM T cells (cluster 14). The frequency of each cluster as percentage of total CD4 per individual donor is depicted in Fig. 3E. Although not present in all donors, clusters 11 and 12 represented CD27$^-$ EM and EMRA CD4$^+$ T cells, which expressed low levels of CD8 (Fig. 3B, 3D). When transformation with a suboptimal cofactor of 150 is applied, these CD8^{dim}CD4$^+$ T cells can be falsely interpreted as CD8$^+$CD4$^+$ T cells (Fig. 3C compared with Fig. 3B). Therefore, correct transformation of the data is crucial for the visualization and interpretation of the data.

When zooming in on each cluster, multiple subclusters can be identified. For instance, zooming in up to the data level for CD103$^+$ EM T cells revealed the heterogenous expression of CD49b and a subpopulation that coexpressed CD69 and CD103 (Supplemental Fig. 5). Moreover, by plotting the sample color, which could be relevant in clinical settings, sample specific clusters can be identified (Supplemental Fig. 5).

To further evaluate the data-driven single-cell clustering, we compared our approach to the standard gating analyses performed by Oetjen et al. (27). First, we revealed a distinct hierarchy of cells, namely the CD8^{dim}CD4$^+$ T cells that were classified as bona fide CD4 T cells, rather than as a separate double-positive T cell population. Second, we included more markers in our analysis

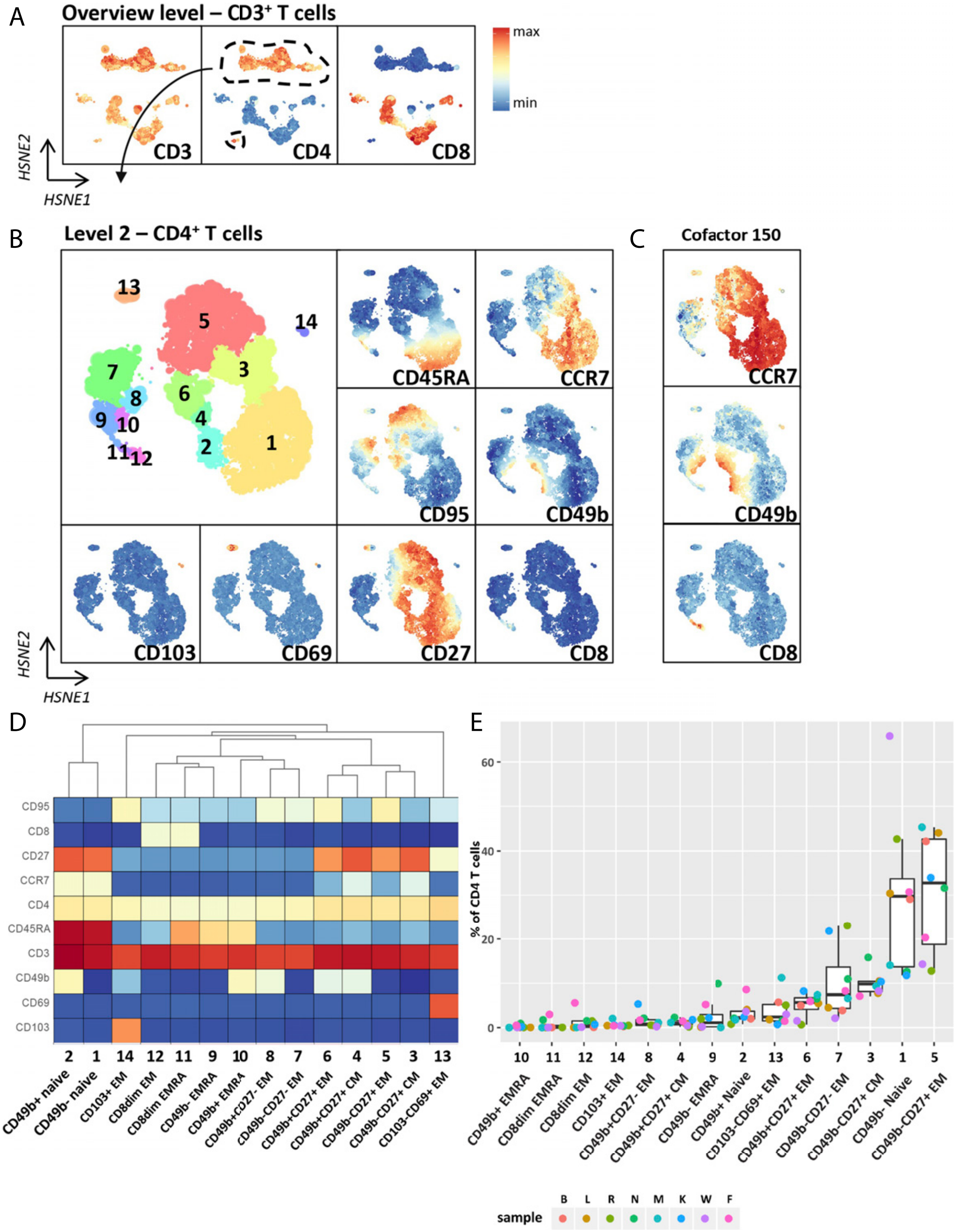

FIGURE 3. Dimensionality reduction and clustering by HSNE and GMS. The exported T cells from eight donors (in total 618,288 cells) were imported in Cytosplore, and an HSNE analysis with default parameters was performed. (**A**) At the overview level, CD4$^+$ T cells were identified and selected for further analysis. (**B**) At the second level of HSNE, multiple CD4$^+$ T cells subsets were identified. The clustering was guided by the GMS clustering, which is based on the density representation of the embedding. (**C**) To demonstrate the effect of changing the cofactor on the color scale, we applied an arcsinh transformation with a suboptimal cofactor of 150 on CCR7, CD49b, and CD8 and plotted the results on the HSNE embedding. Applying a wrong cofactor could lead to misinterpretation of the expression values. (**D**) The heatmap, as generated in Cytosplore, illustrates the median expression values of the T cell markers for each cluster. (**E**) The frequencies of the CD4 T cell clusters in each individual donor are shown. The boxplots indicate the median, interquartile range, and outliers.

(CD49b and CD27), which allowed us to identify additional subpopulations of cells. Third, by considering all markers on each single cell to define populations, instead of only two, we revealed alternate definitions. For instance, in the original gating strategy, only CCR7 and CD45RA were considered for the definition of the naive, CM, EM, and EMRA CD4 T cells, whereas with our approach, all markers were considered (Supplemental Fig. 6A). As a consequence, the median frequency of the CM and EM CD4 T cells differed between the two approaches (Supplemental Fig. 6B). Altogether, these findings indicate the relevance of data-driven

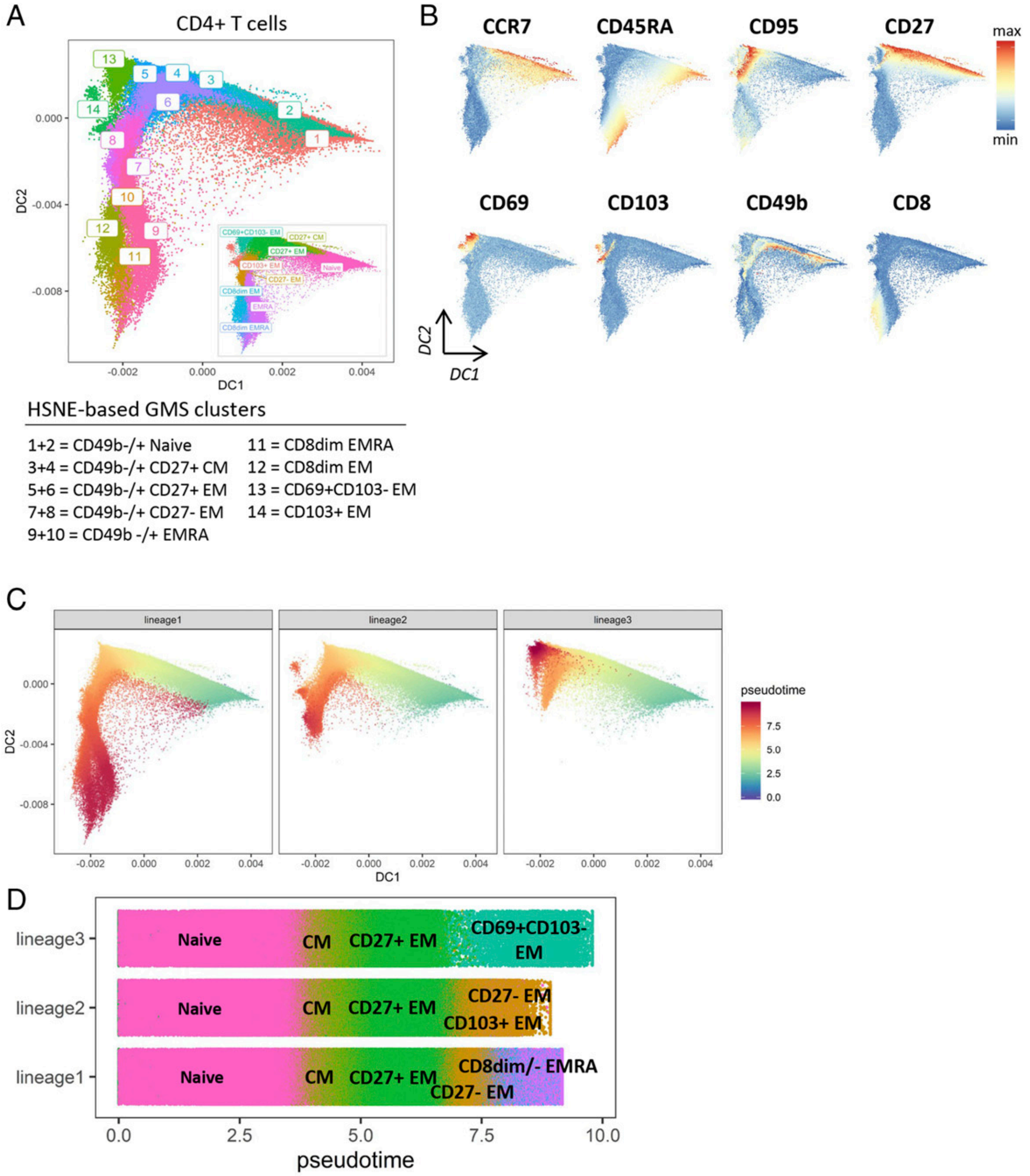

FIGURE 4. Pseudotime analysis by Slingshot. To perform pseudotime analysis on the CD4$^+$ T cells, Slingshot was applied. Slingshot requires clusters and transformed data as input for lineage identification and pseudotime calculation. To visualize the pseudotime values, the dimensionality reduction method diffusion map was used. (**A**) The diffusion map was calculated based on the transformed marker expression values of CD3, CD4, CCR7, CD45RA, CD95, CD27, CD69, CD103, CD49b, and CD8. For illustrative purposes, the HSNE-based cluster assignments were superimposed on the diffusion map. The diffusion map captured the same cellular organization as identified by the clustering, except for the CD49b$^-$ and CD49b$^+$ subsets, which were plotted on top of each other. Those subsets where pooled for pseudotime calculation (as indicated in the legend and the lower-right corner). (**B**) The expression of the individual parameters is shown on the diffusion map. (**C**) Three lineages were identified by Slingshot, which connected the different HSNE-based clusters. The cells are colored according to their pseudotime value. (**D**) All three lineages are characterized by loss of CD45RA, followed by either loss of CD27, gain of CD45RA and possible gain of CD8 (lineage1), loss of CD27 and possible gain of CD103 (lineage2), or gain of CD69 (lineage 3). The color of the cells corresponds to the color of the merged clusters in (A).

single-cell analysis because it could lead to new insights in subset identification and classification.

Alternative clustering methods. In addition to the HSNE-based GMS clustering, we applied the clustering methods PhenoGraph (8) and FlowSOM (7) on the CD4 T cells. To demonstrate the different visualization possibilities, we visualized the clustering results on a diffusion map and UMAP (Supplemental Fig. 4). All three clustering methods allow tuning of the number of clusters. It can be noted that each clustering tool provides slightly different results.

Pseudotime analysis

By clustering, we assume that cells can be categorized in discrete well-defined subpopulations; however, from a biological point of view, it is equally plausible that cells are related to each other in continuous paths of differentiation or maturation. Therefore, we demonstrated how to infer cell trajectories from flow cytometry data and reveal transitional cellular states. To this end, we exported the CD4 clusters (in total 275,856 cells) as assigned in Fig. 3B and imported them in R.

Recently, Slingshot (14) has been introduced, which uses cell clusters to build a minimum spanning tree and reveal the cell lineages. Hereafter, smooth curves are constructed, and the pseudotime variable (i.e., a numeric value representing each cell's progression along a trajectory) is assigned to each cell. In case of single-cell RNA sequencing, the input for clustering and lineage identification is a certain number of (reduced) dimensions. Because flow cytometry already has a limited number of dimensions (equal to the number of parameters), all parameters can be used to calculate the clusters and the pseudotime variable. To visualize the lineages, we plotted the transformed data in a reduced dimensional space.

Slingshot allows the user to choose the dimensionality reduction and clustering method. We used diffusion map because it reorders the cells along a potential differentiation trajectory (Fig. 4A). For illustrative purposes, we chose the HSNE-based GMS clustering as a clustering method and superimposed the cluster assignments on the diffusion map (Fig. 4A). The diffusion map captured the same cellular organization as identified by the clustering, except for the CD49b^{-} and CD49b^{+} subsets, which were plotted on top of each other (Fig. 4A, 4B). For the pseudotime calculation by Slingshot, we pooled the CD49b^{-} and CD49b^{+} subsets because too many clusters will result in artificial lineage calculation. Lineages and pseudotimes were further calculated based on those merged clusters and on the expression levels of the T cell markers, which we visualized on the diffusion map (Fig. 4B). The naive T cells were designated as starting cluster.

In total, three lineages were identified that connected the different clusters (Fig. 4C). By plotting the pseudotime value on the diffusion map, the potential order of differentiation was identified (Fig. 4C). In line with the current T cell development knowledge, the naive CD4^{+} T cells were followed by the CCR7^{+}CD45RA^{-} CM CD4^{+} T cells and CD27^{+}CCR7^{-}CD45RA^{-} EM T cells (Fig. 4B–D). Lineage 1 was further characterized by loss of CD27 and re-expression of CD45RA in the presence or absence of CD8^{dim} expression. In literature, the EMRA CD4 T cells are indeed defined as terminally differentiated memory T cells (34). Although the unconventional CD4^{+}CD8^{dim} T cells are poorly described, they are considered to represent mature memory T cells (35). In lineage 2, cellular differentiation was also defined by loss of CD27 but included both the CD103^{-} and CD69$^{\pm}$CD103^{+} memory T cells. Both CD69 and CD103 are markers associated with tissue residency. Lineage 3 assigned the CD69^{+}CD103^{-} memory CD4^{+} T cells, which are likely to represent the CD4 analogue of the earlier-described bone marrow–resident CD69^{+}CD103^{-}CD8^{+} memory T cells, as end stage (36). Multiple lines of evidence suggest that tissue-resident memory T cells indeed represent a separate lineage, but these studies were mainly focused on CD8 T cells (37). Altogether, Slingshot can be used as hypothesis-generating tool to study cellular trajectories.

Discussion

Flow cytometry is a high-throughput technique with an increasing number of parameters being measured, reinforcing the need for novel analyses to explore cellular heterogeneity. Therefore, we demonstrated the application of tools to facilitate the transition from two-dimensional to high-dimensional single-cell analysis. After the initial steps of compensation and selecting the cells of interest, subsequent steps to continue at the single-cell resolution require a more-advanced approach than for conventional analysis.

The quality of the transformation of the data has a major impact on the downstream analyses (30). We highlighted the importance of an individualized cofactor for each parameter to achieve correct visualization. This is probably caused by the distinct spectral properties of fluorochromes and relative expression of the molecules being recognized. Another aspect of data preprocessing is data normalization. To minimize technical variation, experimental conditions should be kept as consistent as possible. In case some variation is present in a dataset, we showed that normalization can be achieved by the methods implemented in the FlowStats package (31).

Depending on the computational knowledge, any software with a graphical interface (e.g., Cytosplore) or open source packages in R or other programming languages can be explored for further analysis. In this article, we propose two approaches for single-cell analysis of flow cytometry data. First, we used clustering, aiming to define groups of cells with the same phenotype. The differences in subset definition and classification between the conventional gating strategy in the original article (27) and our clustering approach highlight the importance of single-cell analysis. By considering all the markers at once on each single cell, the cellular heterogeneity can more easily be uncovered and visualized. Moreover, the increasing number of parameters detected with flow cytometry makes sequential gating virtually unfeasible.

As a second approach, we demonstrated pseudotime analysis by Slingshot, which enables visualization of transitional phenotypes to study, for instance, development or activation processes. In this analysis, we demonstrated that the differentiation and maturation of the CD4 naive to the EMRA T cells hypothetically occurs in the presence or absence of CD8^{dim} expression. This observation raises the questions of which and why CD4 T cells acquire CD8 expression and would have been missed by the conventional gating analysis if the CD8^{dim} cells were not included in the CD4 gate.

Despite the capacity of clustering and pseudotime tools to identify new subsets and cellular trajectories, respectively, it is important to be aware that the results are tool and setting dependent. Hence, clustering and pseudotime algorithms are useful to explore the presence of potentially interesting new immune subsets and developmental pathways yet always warrant a rigorous validation by additional experiments.

In conclusion, we compared and illustrated the use of multiple methods for dimensionality reduction, clustering, and pseudotime analysis after preprocessing of the data. To support high-dimensional single-cell analysis of flow cytometry data, we provide a well-documented workflow, which could contribute to a deeper understanding of the cellular heterogeneity and intercellular relationships.

Acknowledgments

We thank Dr. Meghali Goswami and Dr. Christopher Hourigan (Laboratory of Myeloid Malignancies, National Heart Lung and Blood Institute, Bethesda, MD) for providing the gating strategy and CD4 subset frequencies of the eight donors.

Disclosures

The authors have no financial conflicts of interest.

References

1. Bandura, D. R., V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner. 2009. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* 81: 6813–6822.
2. Chattopadhyay, P. K., A. F. Winters, W. E. Lomas, III, A. S. Laino, and D. M. Woods. 2019. High-parameter single-cell analysis. *Annu. Rev. Anal. Chem. (Palo Alto, Calif.)* 12: 411–430.
3. Mair, F., and M. Prlic. 2018. OMIP-044: 28-color immunophenotyping of the human dendritic cell compartment. [Published erratum appears in 2019 Cytometry A 95: 925–926.] *Cytometry A* 93: 402–405.
4. Robinson, J. P. 2019. Spectral flow cytometry-Quo vadimus? *Cytometry A* 95: 823–824.

5. Weber, L. M., and M. D. Robinson. 2016. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* 89: 1084–1096.
6. Comaniciu, D., and P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 24: 603–619.
7. Van Gassen, S., B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys. 2015. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* 87: 636–645.
8. Levine, J. H., E. F. Simonds, S. C. Bendall, K. L. Davis, el.-A. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, et al. 2015. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162: 184–197.
9. Duò, A., M. D. Robinson, and C. Soneson. 2018. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000 Res.* 7: 1141.
10. Waltman, L., and N. J. van Eck. 2013. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* 86: 471.
11. Satija, R., J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. 2015. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33: 495–502.
12. Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32: 381–386.
13. Saelens, W., R. Cannoodt, H. Todorov, and Y. Saeys. 2019. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37: 547–554.
14. Street, K., D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19: 477.
15. Setty, M., M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34: 637–645.
16. Wolf, F. A., F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* 20: 59.
17. Saeys, Y., S. Van Gassen, and B. N. Lambrecht. 2016. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* 16: 449–462.
18. Palit, S., C. Heuser, G. P. de Almeida, F. J. Theis, and C. E. Zielinski. 2019. Meeting the challenges of high-dimensional single-cell data analysis in immunology. *Front. Immunol.* 10: 1515.
19. Van Der Maaten, L., and G. Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9: 2579–2605.
20. van Unen, V., T. Höllt, N. Pezzotti, N. Li, M. J. T. Reinders, E. Eisemann, F. Koning, A. Vilanova, and B. P. F. Lelieveldt. 2017. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* 8: 1740.
21. Höllt, T., N. Pezzotti, V. van Unen, F. Koning, E. Eisemann, B. Lelieveldt, and A. Vilanova. 2016. Cytosplore: interactive immune cell phenotyping for large single-cell datasets. *Comput. Graph. Forum* 35: 171–180.
22. McInnes, L., J. Healy, N. Saul, and L. Großberger. 2018. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3: 861.
23. Becht, E., L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. 2018. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37: 38–44.
24. Haghverdi, L., F. Buettner, and F. J. Theis. 2015. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31: 2989–2998.
25. Angerer, P., L. Haghverdi, M. Büttner, F. J. Theis, C. Marr, and F. Buettner. 2016. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 32: 1241–1243.
26. Spidlen, J., K. Breuer, C. Rosenberg, N. Kotecha, and R. R. Brinkman. 2012. FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry A* 81: 727–731.
27. Oetjen, K. A., K. E. Lindblad, M. Goswami, G. Gui, P. K. Dagur, C. Lai, L. W. Dillon, J. P. McCoy, and C. S. Hourigan. 2018. Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight* 3: e124928.
28. Hahne, F., N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. 2009. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10: 106.
29. Azad, A., B. Rajwa, and A. Pothen. 2016. flowVS: channel-specific variance stabilization in flow cytometry. *BMC Bioinformatics* 17: 291.
30. Finak, G., J.-M. Perez, A. Weng, and R. Gottardo. 2010. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinformatics* 11: 546.
31. Hahne, F., A. H. Khodabakhshi, A. Bashashati, C.-J. Wong, R. D. Gascoyne, A. P. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, et al. 2010. Per-channel basis normalization methods for flow cytometry data. *Cytometry A* 77: 121–131.
32. Melville, J. 2020. uwot: the Uniform Manifold Approximation and Projection (UMAP) method for dimensionality reduction. R package version 0.1.8. Available at: https://CRAN.R-project.org/package=uwot. Date accessed: December 2, 2020.
33. Bendall, S. C., E. F. Simonds, P. Qiu, el.-A. D. Amir, P. O. Krutzik, R. Finck, R. V Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, et al. 2011. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332: 687–696.
34. Moro-García, M. A., R. Alonso-Arias, and C. López-Larrea. 2013. When aging reaches CD4+ T-cells: phenotypic and functional changes. *Front. Immunol.* 4: 107.
35. Nascimbeni, M., E. C. Shin, L. Chiriboga, D. E. Kleiner, and B. Rehermann. 2004. Peripheral CD4(+)CD8(+) T cells are differentiated effector memory cells with antiviral functions. *Blood* 104: 478–486.
36. Melsen, J. E., G. Lugthart, C. Vervat, S. M. Kielbasa, S. A. J. van der Zeeuw, H. P. J. Buermans, M. M. van Ostaijen-Ten Dam, A. C. Lankester, and M. W. Schilham. 2018. Human bone marrow-resident natural killer cells have a unique transcriptional profile and resemble resident memory $CD8^+$ T cells. *Front. Immunol.* 9: 1829.
37. Amsen, D., K. P. J. M. van Gisbergen, P. Hombrink, and R. A. W. van Lier. 2018. Tissue-resident memory T cells at the center of immunity to solid tumors. *Nat. Immunol.* 19: 538–546.