nature biotechnology

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht[1], Leland McInnes[2], John Healy[2], Charles-Antoine Dutertre[1], Immanuel W H Kwok[1], Lai Guan Ng[1], Florent Ginhoux[1] & Evan W Newell[1,3]

Advances in single-cell technologies have enabled high-resolution dissection of tissue composition. Several tools for dimensionality reduction are available to analyze the large number of parameters generated in single-cell studies. Recently, a nonlinear dimensionality-reduction technique, uniform manifold approximation and projection (UMAP), was developed for the analysis of any type of high-dimensional data. Here we apply it to biological data, using three well-characterized mass cytometry and single-cell RNA sequencing datasets. Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters. The work highlights the use of UMAP for improved visualization and interpretation of single-cell data.

The past decades have witnessed a large increment in the number of parameters analyzed in single-cell cytometry and transcriptome studies. Parameter numbers currently reach ~20 for flow cytometry, ~40 for mass cytometry and >20,000 in single-cell RNA sequencing (scRNAseq). Dimensionality reduction techniques have been pivotal in enabling researchers to visualize high-dimensional data. Although principal component analysis (PCA) has historically been the most commonly used method for dimensionality reduction, the importance of nonlinear dimensionality reduction techniques has recently been recognized. Nonlinear dimensionality reduction techniques are, notably, able to avoid overcrowding of the representation, wherein distinct clusters are represented on an overlapping area. Nonlinear dimensionality reduction methods[1] include Isomap[2], Diffusion Map[3] and *t*-distributed stochastic neighborhood embedding (t-SNE[4], renamed viSNE[5]). t-SNE is currently the most commonly used technique in single-cell analysis. It has been used to efficiently reveal local data structure and is widely used to identify distinct cell populations in cytometry and transcriptomic data. However, t-SNE suffers from limitations such as loss of large-scale information (the intercluster relationships), slow computation time and inability to meaningfully represent very large datasets[6]. A new algorithm, called uniform manifold approximation and projection (UMAP) has been recently published[7,8] and is claimed to preserve as much of the local and more of the global data structure than t-SNE, with a shorter run time. Given the wide use of t-SNE in the analysis of flow and mass cytometry data, as well as scRNAseq data, here we test these claims on three well-characterized single-cell datasets[9–11]. We also visually and quantitatively compare the performance of UMAP with the widely used Barnes–Hut implementation of t-SNE[12]; the heavily optimized Fourier-interpolated t-SNE, with or without late exaggeration (FIt-SNE l.e. or FIt-SNE, respectively)[13]; and the autoencoder neural network scvis[14].

RESULTS

Qualitative comparison of UMAP with t-SNE

We ran UMAP and t-SNE simultaneously on a dataset covering 35 samples originating from 8 distinct human tissues enriched for T and natural killer (NK) cells, of more than >300,000 events with 39 protein targets[11] (the Wong dataset; **Supplementary Table 1**). Using the Louvain clustering-based Phenograph[15] algorithm and manual cluster labeling, we classified events into six broad cell populations (**Supplementary Fig. 1a**). UMAP and t-SNE were both successful at pulling together only clusters corresponding to similar cell populations with generally very good correspondence with Phenograph clustering (**Fig. 1a** and **Supplementary Fig. 1b**). However, t-SNE separated cell populations into distinct clusters more commonly than UMAP, notably splitting CD8 T cells, γδ T cells and contaminating cells (likely including B cells) into two distinct clusters each. Nonetheless, while these cells were not always segregated into completely distinct clusters by UMAP, these cell populations remained similarly identifiable in UMAP as compared to t-SNE, both techniques surpassing PCA (**Supplementary Fig. 1b**).

By color-coding the tissues of origin on the UMAP and t-SNE maps, we observed that t-SNE separated cell populations according to their tissue of origin more often than UMAP (**Fig. 1b** and **Supplementary Fig. 2**). UMAP instead ordered events according to their origin within each major cluster, roughly from cord blood and peripheral blood mononuclear cells, to liver and spleen, and to tonsils on the one end to skin, gut and lung on the other. The sample type was not given as an input of either of the two algorithms. We observed that UMAP was able to recapitulate the differentiation stage of T cells within

[1]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [2]Tutte Institute for Mathematics and Computing, Ottawa, Ontario, Canada. [3]Fred Hutchinson Cancer Research Center, Vaccine and Infectious Disease Division, Seattle, Washington, USA. Correspondence should be addressed to E.W.N. (enewell@fredhutch.org).

Received 11 April; accepted 5 November; published online 3 December 2018; doi:10.1038/nbt.4314

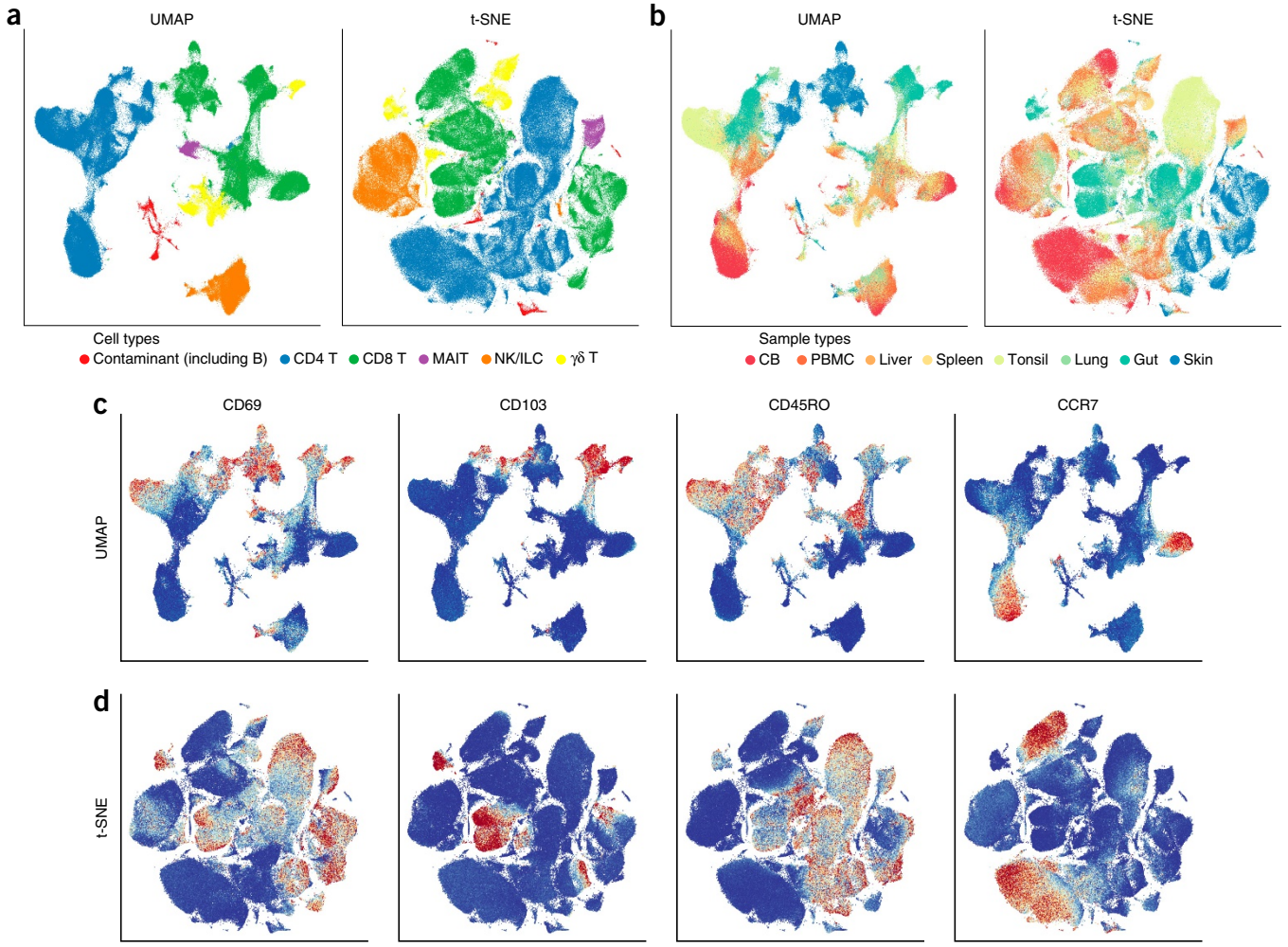

Figure 1 UMAP embeds local and large-scale structure of the data. UMAP and t-SNE projections of the Wong *et al.* dataset colored according to (**a**) broad cell lineages, (**b**) tissue of origin, and for (**c**) UMAP and (**d**) t-SNE, the expression of CD69, CD103, CD45RO and CCR7. For **c** and **d**, blue denotes minimal expression, beige intermediate and red high. MAIT, mucosal-associated invariant T cell; ILC, innate lymphoid cell; CB, cord blood; PBMC, peripheral blood mononuclear cell.

each major cluster, as seen by the expression levels of events for the resident-memory T cell markers CD69 and CD103, the memory T cell marker CD45RO and the naive T cell marker CCR7 on the UMAP projection (**Fig. 1c**). By contrast, while t-SNE identified similar continua within clusters, they had no apparent structure along a common axis that made them easily identifiable (**Fig. 1d**).

UMAP better represents the multi-branched continuous trajectory of hematopoietic development

To investigate how UMAP handles continuity of cell phenotypes, we applied it alongside t-SNE on the well-documented subject of bone marrow hematopoiesis using both a mass cytometry (the Samusik_01 dataset[10]: >86,000 events, 38 parameters and 24 cell populations annotated by its authors; **Supplementary Table 1**) and a scRNAseq dataset (the Han dataset[9]: three sample types, 51,252 cells and 25,912 dimensions reduced to 100 approximate principal components; **Supplementary Table 1**). On the mass cytometry dataset, UMAP visually revealed eight main cell clusters (**Fig. 2a**). One was composed of all B cell subsets (and close to a small cluster of plasma cells) and one of all T cell subsets. Four small, homogeneous clusters corresponded respectively to macrophages, NK cells, eosinophils and nonclassical monocytes. The last cluster contained 11 out of the 24 manually gated populations and appeared most interesting with respect to hematopoiesis. Indeed, these populations were ordered according to a five-branched structure that was consistent with hematopoietic differentiation. Hematopoietic stem cells (HSCs) overlapped with multipotent progenitors (MPPs). These cells neighbored common lymphoid progenitors (CLPs) on one side and common myeloid progenitors (CMPs) on the other. CMPs led to myeloid-erythroid progenitors (MEPs), which led to unlabeled erythrocytes (**Supplementary Fig. 3**) and to granulocyte-myeloid progenitors (GMPs). GMPs then led to classical monocytes that further led to myeloid dendritic cells on one branch and to cells labeled as intermediate monocytes on another branch. UMAP linked basophils to a population of $Lin^{-}cKit^{+}Sca1^{-}CD34^{+}Fc\gamma RII/III^{+}Fc\varepsilon RI\alpha^{+}$ cells, consistent with a previously described phenotype for committed basophil progenitors[16]. These putative progenitors appeared closer to CMPs than to GMPs, a topic that is still intensely debated. Neutrophils were gated out from the dataset by its authors and are thus absent from this representation[10].

t-SNE identified relatively similar clusters, with a few differences (**Fig. 2b**), notably singling out more clusters than UMAP. $CD4^{+}$ T cells

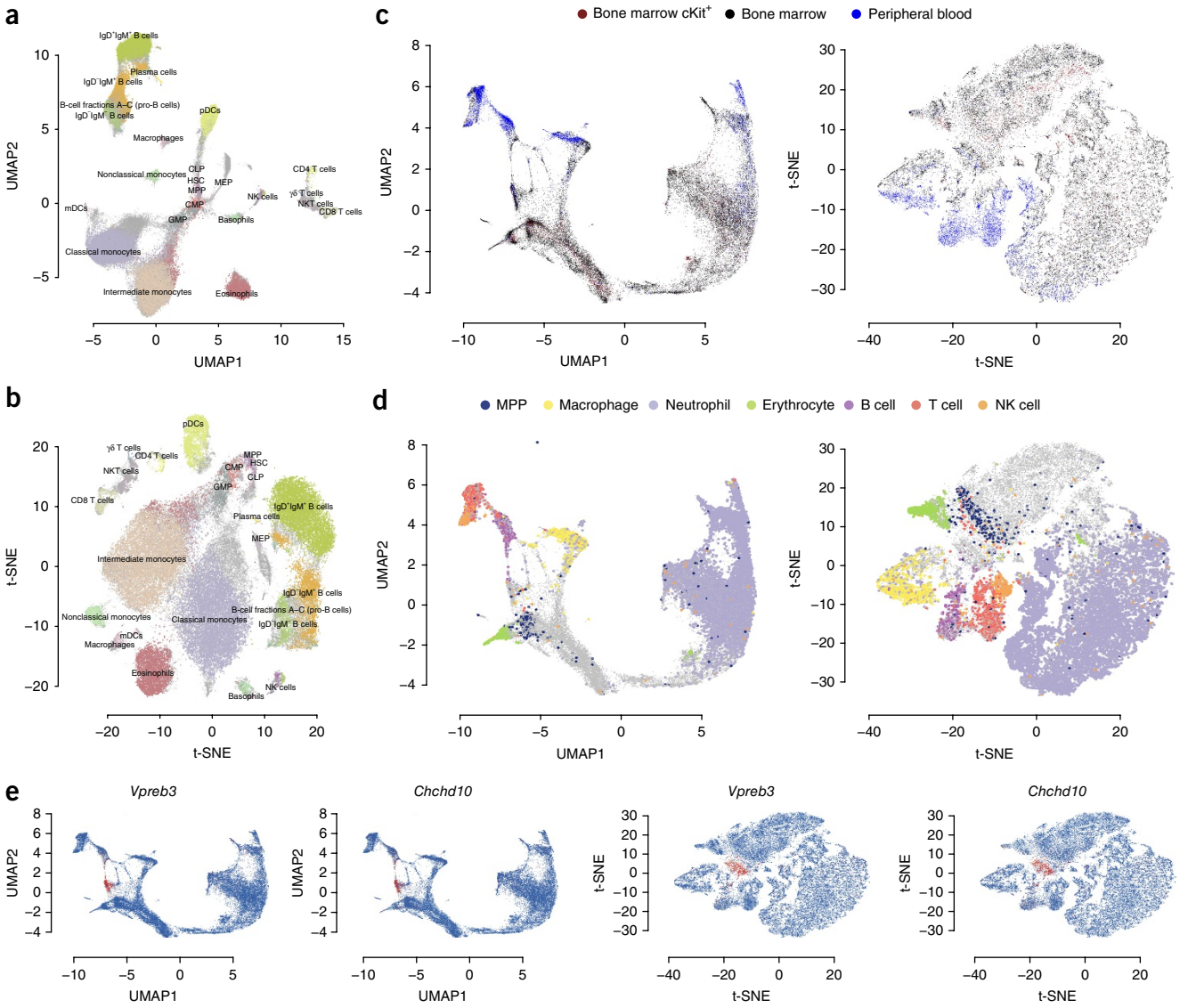

Figure 2 UMAP embeddings of bone marrow and blood samples recapitulate hematopoiesis. (**a**) UMAP and (**b**) t-SNE projection of the Samusik_01 dataset. Events are color-coded according to manual gates provided by the authors of the dataset. (**c**,**d**) UMAP and t-SNE projections of the Han dataset, color-coded by (**c**) tissue of origin or (**d**) cell populations. (**e**) Expression of the V-set pre-B cell surrogate light chain 3 (*Vpreb3*) and *Chchd10* genes on the UMAP and t-SNE projections of the Han dataset. Blue denotes minimal expression, beige intermediate and red high. pDC, plasmacytoid dendritic cell; mDC, myeloid dendritic cell; NKT, natural killer T.

were separated from other T cell subsets. As noted by others[17], t-SNE expands low density areas and tends to ignore global relationships. Thus, while some paths from HSCs and MPPs to differentiated populations were still apparent—notably, from HSCs to monocytes—the overall structure was less clear, as no narrow 'neck' led to larger terminal clusters. t-SNE also separated basophils from their putative precursors close to CMPs and GMPs and separated plasmacytoid dendritic cells from CLPs. The density of events in the dimensionally reduced space also appeared less uniform in t-SNE, with large clusters in the t-SNE space being less dense than the smaller ones. In contrast, the density of UMAP clusters appeared more uniform, which could help avoid biases in interpreting phenotypic heterogeneity in large versus small clusters (**Supplementary Fig. 4**).

From the scRNAseq dataset we analyzed the transcriptomes of cells isolated from bone marrow, cKit+ bone marrow, and peripheral blood to facilitate identifying mature versus progenitor cell populations (**Fig. 2c**). We first removed low-abundance cell types such as basophils and eosinophils, contaminants such as mature erythrocytes, and outlier cells originating from unique samples and highly expressing mitochondrial transcripts (**Supplementary Fig. 5**). Using published cell signatures specific for mouse bone marrow cell populations[18], we were able to identify cell clusters that corresponded to MPPs, MEPs, macrophages, B cells, T cells and NK cells (**Fig. 2d**). Consistently with the UMAP projection of the mass cytometry dataset, MPPs were found amid a larger group of clusters that led to differentiated cells originating from peripheral blood samples (**Fig. 2c**). Peripheral blood events consisted of distinct clusters of lymphocytes (T, NK and B cells), macrophages, MEPs and neutrophils (**Fig. 2d**). Although this does not prove that cells lying between MPPs and differentiated cells are committed progenitors, these results suggest

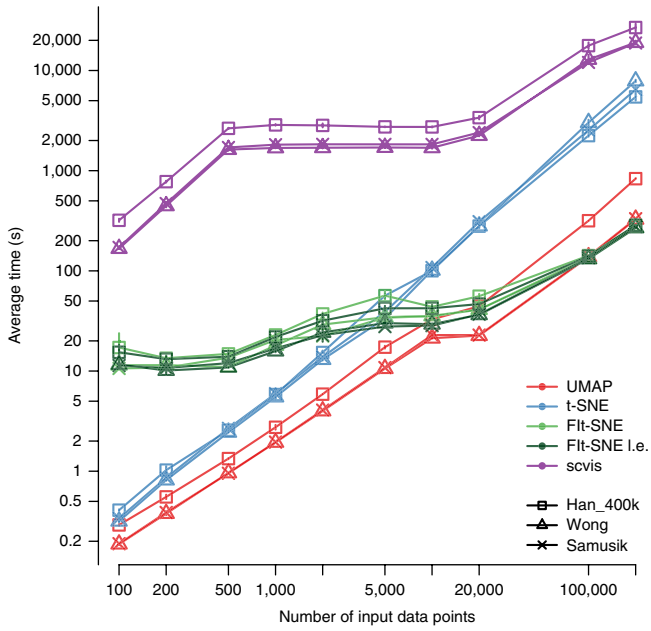

Figure 3 Run times of five dimensionality reduction methods for inputs of varying sizes. The average run time of three random subsamples is represented, with vertical bars representing s.d. after log-transforming the run times.

that UMAP could be used as a hypothesis-generating tool to identify putative markers for such cells. By investigating a small cluster of cells lying between MPPs and mature B cells in the UMAP projection, we were indeed able to identify the pre-B cell marker[19] Vpreb3 and to hypothesize that Chchd10 could be another gene marker for pre-B cells in mouse bone marrow (**Fig. 2e**). These conclusions and hypotheses would have been more difficult to draw using t-SNE, which blurred the relationship of terminal clusters to MPPs (**Fig. 2d**,**e**).

Quantitative analysis of UMAP, t-SNE and scvis outputs

Linderman *et al.* recently optimized t-SNE to decrease its run time by using Fourier interpolation to speed up the convolution step (the FIt-SNE algorithm)[13]. In addition, these authors proposed a late-exaggeration parameter that magnifies gaps between distinct clusters (FIt-SNE l.e.). Ding *et al.* recently published the autoencoder neural network scvis algorithm[14]. To formalize the qualitative observation we reported above, we quantitatively benchmarked computational aspects of UMAP along with these modern dimensionality reduction methods on large (300,000–800,000 cells) and high-dimensional (38–100) datasets, including two mass cytometry datasets and one scRNAseq dataset. We report on the duration each algorithm takes to complete on data subsamples of various sizes and their abilities to represent distinct cell clusters in a non-overlapping fashion, to produce reproducible clusters and to preserve distances.

We measured the execution time of each algorithm across data subsamples of sizes ranging from 20 to 200,000 cells (**Fig. 3**). Across all ranges of subsample sizes, timings showed little variability across replicates. scvis appeared consistently slow across all dataset range, taking 326 s for subsamples of size 100. Its timing plateaued in the intermediate range around 2,000 s for 500 to 20,000 data points and rose again to up to 10 h for 200,000 data points. Among the remaining algorithms, FIt-SNE appeared slightly slower at low data subsample sizes, likely owing to input/output operations (as the author-provided R script we used writes data to disk before running FIt-SNE on it). However, it was faster than Barnes–Hut t-SNE at data sizes over 5,000 data points and faster than UMAP at data sizes above 100,000. UMAP appeared the fastest algorithm for small datasets and appeared competitive with FIt-SNE for large datasets (100,000 events and more). UMAP was also 2.5 times slower on the un-subsetted, entire Han dataset (Han_400k, **Supplementary Table 1**) compared to the other two datasets, likely owing to its increased dimensionality (100 versus 38 or 39 dimensions for the other datasets). Late exaggeration did not affect the run time of FIt-SNE. In practice, both UMAP and FIt-SNE appear to be much faster than current standards for dimensionality reduction, while the time required for scvis is currently a limitation for the sizes of datasets typically analyzed for both mass cytometry and many scRNAseq experiments.

To investigate the ability of each method to separate cell populations in two-dimensional embeddings, we trained random forests to predict Phenograph clusters' identities from the position of the data points in embeddings and measured the accuracy of the predictions on held-out data (**Fig. 4a**). Nonlinear dimensionality reduction methods all led to higher accuracy than PCA, even when including up to five principal components. UMAP, t-SNE and FIt-SNE with or

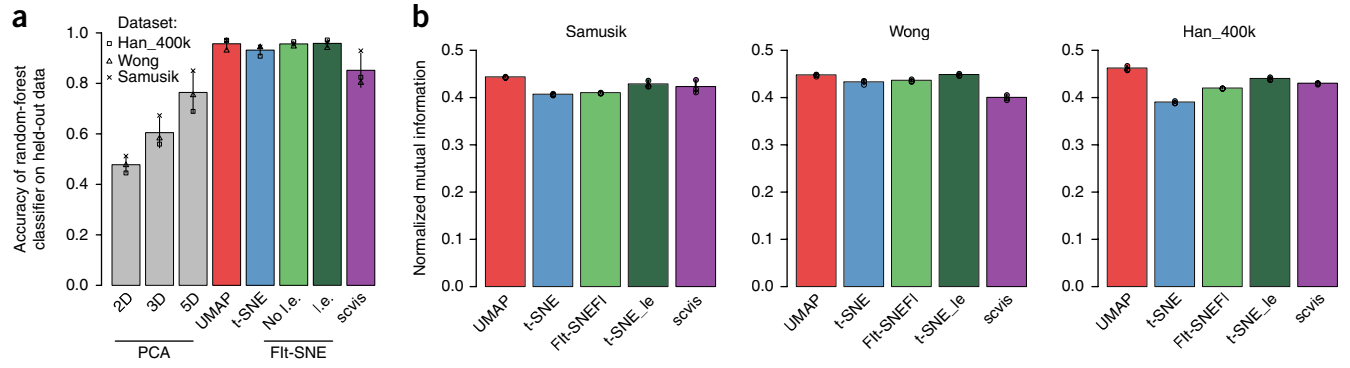

Figure 4 Analysis of local data structure in embeddings produced by each algorithm. (**a**) Accurate classification rate on held-out data of random-forest classifiers predicting Phenograph cluster labels using embedded coordinates as input. The average across the three datasets is shown, with vertical bars representing s.d. (**b**) Average normalized mutual information of *k*-means clustering ($k = 100$) performed on the embeddings of data subsamples and *k*-means clustering ($k = 100$) performed on total datasets. The average across the three random subsamples of size 200,000 is shown, with vertical bars representing s.d.

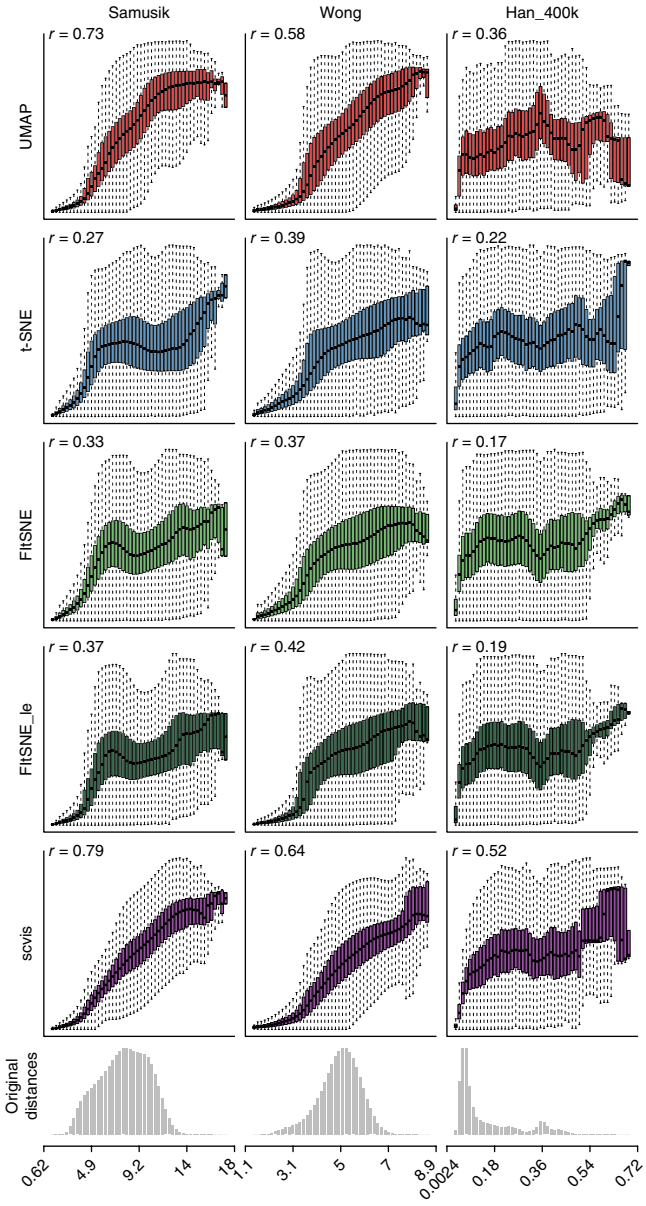

Figure 5 Preservation of pairwise distances in embeddings. Box plots represent distances across pairs of points in the embeddings, binned using 50 equal-width bins over the pairwise distances in the original space using 10,000 randomly selected points, leading to 49,995,000 pairs of pairwise distances. The last row of graphs represents counts of pairwise distances in each bin of distances from the original space as histograms. The value of the Pearson correlation coefficient computed over the pairs of pairwise distances is reported. For the box plots, the central bar represents the median, and the top and bottom boundary of the boxes represent the 75th and 25th percentiles, respectively. The whiskers represent 1.5 times the interquartile range above (or, respectively, below) the top (or, respectively, bottom) box boundary, truncated to the data range if applicable.

without late exaggeration all appeared to perform well in this setting, leading to random-forest accuracies around 95%. scvis appeared less efficient at separating cell populations, leading to an accuracy of 85%. These conclusions were visually consistent with the annotation of embeddings using manually labeled cell populations (either on manual gating for the Samusik dataset, annotated clustering for the Wong dataset or sample of origin for the Han_400k dataset), where t-SNE-based methods and UMAP appeared best at cleanly separating cell populations, followed by scvis and finally two-dimensional PCA (**Supplementary Fig. 6**).

We also benchmarked how reproducible the embeddings were on a local scale. To do so, for each dataset, we used *k*-means clustering (with $k = 100$ to obtain small clusters) on the embeddings of three data subsamples of size 200,000, as well as on the embedding of the full dataset. We then measured how informative the knowledge of cluster identities on the embeddings of the subsamples were to predict the cluster identities obtained from the embeddings of the full datasets (**Fig. 4b**). In this benchmark, each algorithm appeared to produce embeddings of subsamples consistent with embeddings of the full dataset, with normalized mutual information coefficients between 0.4 and 0.5. UMAP nonetheless ranked first on two datasets and second on the third one, with FIt-SNE with late exaggeration appearing second best at consistently embedding data points on a local scale.

Finally, we analyzed how each algorithm dealt with global data structure by comparing distances between random pairs of points in the original (high-dimensional) or embedded (two-dimensional) data, as well as by quantifying the reproducibility of embeddings. Examining how distances in the original datasets were related to distances in the embeddings showed that each algorithm consistently preserved distances on a small scale. However, t-SNE-based algorithms appeared to ignore distances on a moderate scale (**Fig. 5**). scvis and UMAP both appeared to better preserve large-scale distances. scvis appeared marginally better than UMAP in that regard. To quantify the reproducibility of embeddings on a large scale, we measured the correlation of coordinates on random subsamples of varying sizes versus embeddings of the full dataset, up to symmetries across the axes (**Fig. 6** and **Supplementary Fig. 7**). t-SNE-based methods appeared poorly reproducible across datasets and subsample sizes. scvis and UMAP both appeared more reproducible, with UMAP ranking first for most sample sizes across all datasets. These finding were consistent with the idea that both UMAP and scvis perform optimizations that are sensitive to global features of the data, thus reaching similar arrangements more consistently.

Taken together, these analyses suggest that UMAP achieves delineation of cell subsets that is comparable to that of t-SNE, in addition to preserving global distances similarly to the autoencoder-based scvis. In addition, UMAP is relatively fast, with run times only slightly above FIt-SNE. It thus appears to be a robust method, able to preserve both local and global structure in the data without compromising much on any of these aspects, while being very fast for current standards.

DISCUSSION

Our work shows that UMAP produces equally meaningful representations compared with t-SNE, particularly in its ability to resolve subtly differing cell populations. It also provides the useful and intuitively pleasing feature that it preserves more of the global structure and, notably, the continuity of the cell subsets. In addition to making plots easier to interpret, we note that this also improves its utility for generating hypotheses related to cellular development. UMAP outputs are faster to compute compared with Barnes–Hut t-SNE, much faster than scvis, and comparable to FIt-SNE. UMAP embeddings are more reproducible than other methods, notably more so than those from t-SNE implementations. We systematically and quantitatively benchmarked all these qualitative aspects. UMAP was found to be the best or close to best method in every aspect investigated, and thus appears as a robust all-around method for dimensionality reduction for single-cell methods. Although this will also be possible with UMAP, the

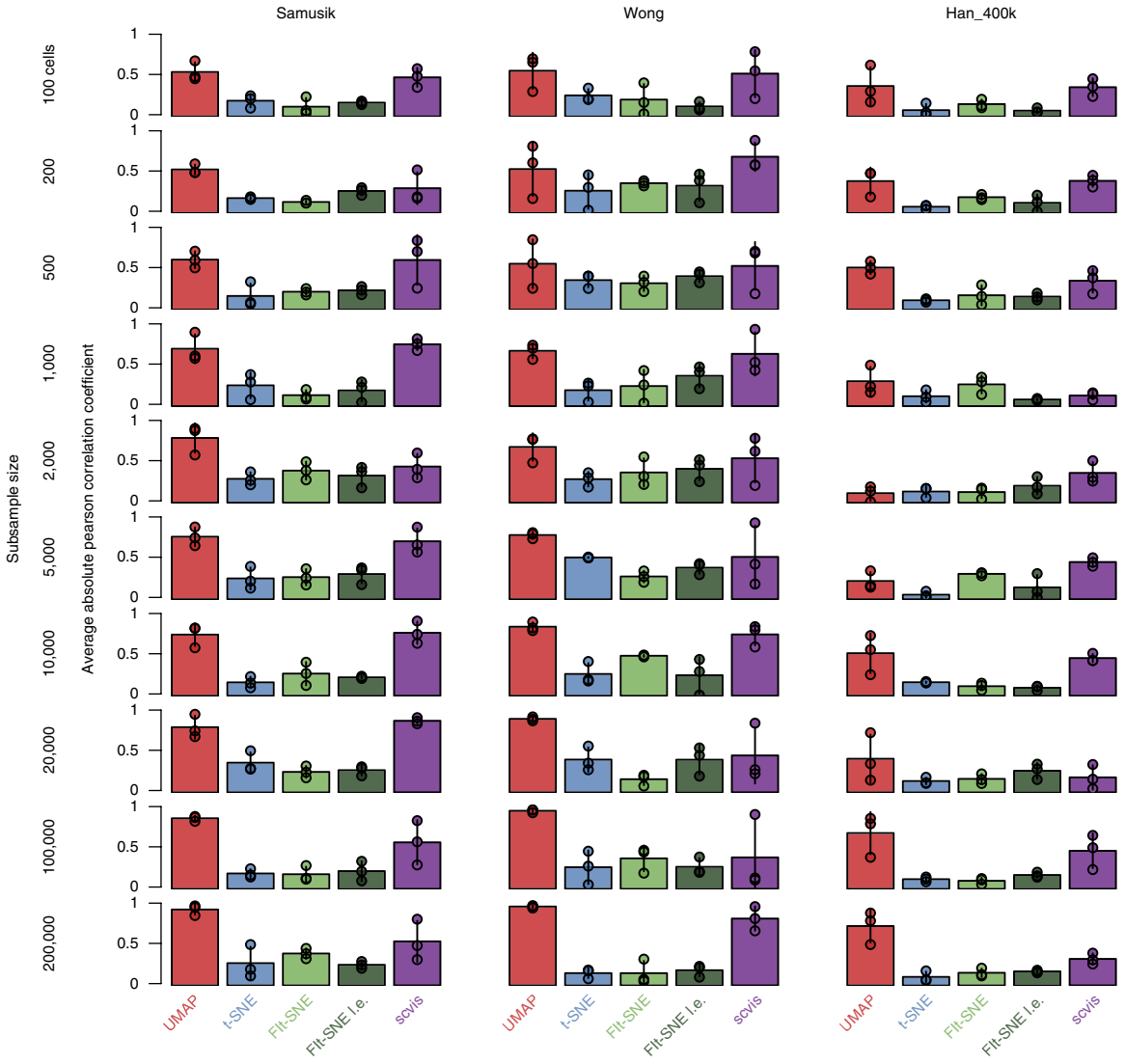

Figure 6 Reproducibility of large-scale structures in embeddings. Bar plots represent the average unsigned Pearson correlation coefficient of the points' coordinates in the embedding of subsamples versus in the embedding of the full dataset, thus measuring the correlation of coordinates in subsamples versus in the embedding of the full dataset, up to symmetries along the graph axes. Bar heights represent the average across three replicates and vertical bars the corresponding s.d.

autoencoder scvis currently offers the advantage of being able to append new data points to an existing embedding. The scRNAseq analysis toolkits scanpy[20] and Seurat[21] recently implemented UMAP as a possible tool for dimensionality reduction, and the popular commercial software platform for flow cytometry analysis FlowJo recently released a plug-in to run UMAP. Altogether, on the basis of its ease of use and results of our benchmarking analyses, we anticipate that UMAP will be a valuable tool that can be rapidly adopted by the single-cell analysis community.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

We thank members of the Singapore Immunology Network and notably members of the E.W.N. laboratory. We thank S. Li, Y. Simoni, M. Chng, Y. Cheng, J.W. Lim and M. Fehlings for their insightful feedback. This study was funded by A-STAR/SIgN core funding and A-STAR/SIgN immunomonitoring platform funding.

AUTHORS CONTRIBUTIONS

E.B., L.M., J.H., C.-A.D., I.W.H.K. and E.W.N. analyzed data. L.G.N., F.G. and E.W.N. helped supervise the project. L.M. and J.H. developed UMAP. All authors participated in writing and revising the manuscript.

COMPETING INTERESTS

E.W.N. is a board director and shareholder of immunoSCAPE Pte. Ltd., which is an immune profiling service provider.

1. Saeys, Y., Van Gassen, S. & Lambrecht, B.N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **16**, 449–462 (2016).
2. Tenenbaum, J.B., De Silva, V. & Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
3. Coifman, R.R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. USA* **102**, 7426–7431 (2005).
4. Van Der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-SNE. journal of machine learning research. *J. Mach. Learn. Res.* **9**, 26 (2008).
5. Amir, A.D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
6. van Unen, V. *et al.* Mass cytometry of the human mucosal immune system identifies tissue- and disease-associated immune subsets. *Immunity* **44**, 1227–1239 (2016).
7. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
8. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
9. Han, X. *et al.* Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107. e17 (2018).
10. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L. & Nolan, G.P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).
11. Wong, M.T. *et al.* A high-dimensional atlas of human T cell diversity reveals tissue-specific trafficking and cytokine signatures. *Immunity* **45**, 442–456 (2016).
12. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
13. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. & Kluger, Y. Efficient algorithms for *t*-distributed stochastic neighborhood embedding. Preprint at https://arxiv.org/abs/1712.09005 (2017).
14. Ding, J., Condon, A. & Shah, S.P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).
15. Levine, J.H. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
16. Huang, H., Li, Y. & Liu, B. Transcriptional regulation of mast cell and basophil lineage commitment. *Semin. Immunopathol.* **38**, 539–548 (2016).
17. Wattenberg, M., Viégas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1**, e2 (2016).
18. de Graaf, C.A. *et al.* Haemopedia: an expression atlas of murine hematopoietic cells. *Stem Cell Rep.* **7**, 571–582 (2016).
19. Mårtensson, I.-L., Keenan, R.A. & Licence, S. The pre-B-cell receptor. *Curr. Opin. Immunol.* **19**, 137–142 (2007).
20. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
21. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

ONLINE METHODS

Datasets. The main characteristics of the datasets we analyzed and their accession identifiers are presented in **Supplementary Table 1**. For the Wong *et al.* dataset, live CD45$^+$ non-B (CD19$^-$), non-monocyte (CD14$^-$) events were selected using FlowJo software. To partially equalize weighting of each human tissue, a maximum of 10,000 events were randomly sampled from each of the 39 samples before analysis. Other datasets were used as described in the table. For the Samusik dataset, we used either its restriction to the first bone marrow sample analyzed (Samusik_01) or the full dataset (Samusik). For the Han dataset, we used either its restriction to hematopoietic samples (Han) or the full dataset encompassing 50 distinct tissues (Han_400k).

Transformations and preprocessing. For the Samusik CyTOF datasets we used an arcsinh transformation with a cofactor of 1, and for the Wong CyTOF dataset we used a logicle transform (parameters $w = 0.25$, $t = 16{,}409$, $m = 4.5$, $a = 0$). For the scRNAseq Han dataset we transformed count into reads per million (thus normalizing the number of reads per cell to 1) and reduced to 100 approximate principal components using the IRLBA R package.

Running dimensionality reduction algorithms. We used a total of five nonlinear algorithms (UMAP, Barnes-Hut t-SNE, FIt-SNE, FIt-SNE l.e. and scvis) and one linear dimensionality reduction algorithms (PCA). Software versions, accession links and parameters used are listed in **Supplementary Table 2**.

Cell annotations. For the Samusik datasets, we used cell annotations provided by the authors and available from the public repository. For all datasets we used Phenograph clustering (with default parameters $k = 30$), and for the Wong dataset we manually labeled the clusters into broad cell populations. For the Han dataset we used the AUCell R package[22], which computes the AUC of gene sets within each single cell, using gene sets from the Haemopedia[18] resource to annotate cell lineages. We then manually thresholded these AUC scores to obtain categorical labels. Cells that were assigned to multiple lineages were set to unlabeled.

Quantitative benchmarks. Each algorithm was run on each total dataset. Then subsamples of sizes 100, 200, 500, 1,000, 2,000, 5,000, 10,000, 20,000, 100,000 and 200,000 were uniformly drawn three times, generating 30 data subsamples of varying sizes. We then ran each algorithm on each data subsample, saving run times and embeddings.

Run time. For UMAP, Barnes–Hut t-SNE, FIt-SNE and FIt-SNE l.e., timing was determined in R using the "elapsed" (wall clock) time measurements, to allow for consistent timing across methods that made different uses of multithreading or input/output methods. For scvis we used the self-reported run time output by the software upon completion.

Separability of cell populations. To investigate the ability of each dimensionality reduction method to preserve the cohesiveness of cell populations in the embeddings, we first generated categorical labels for the complete Wong and the Han_400k datasets using Phenograph clustering with hyperparameter $k = 30$ (default value). For each dataset, we sampled 50,000 cells as a training set. For each algorithm, we trained on the downsampled embedding a random-forest classifier using the cluster labels as target variable and the embeddings' coordinates as training variables. We then used these classifiers to predict cluster identities for a non-overlapping random test set of 50,000 cells and computed the accuracy of these predictions, thus assessing the ability of each method to separate cell clusters. For a more qualitative analysis of the separability of cell populations, we used categorical labels from either manual annotation (Wong and Samusik datasets) or tissue and sample of origin (Han dataset), which we used to annotate the dimensionality reduction plots.

Reproducibility of local structure across replicates. From the complete datasets and the three replicates of their largest data subsamples (200,000), we used k-means clustering on each embedding using $k = 100$ clusters. We then computed the mutual information between clusters obtained on the embedding of the total dataset and each of its largest subsamples, normalized for the entropy in the total datasets' clustering. This metric measures the proportion of the information of the clustering on the total dataset's embedding that can be restored from the knowledge of the clusters on the embeddings of subsamples, allowing a comparison across dimensionality reduction methods and replicates.

Correlation of pairwise distances. From each dataset we sampled 10,000 data points. We then computed the corresponding 49,995,000 pairwise distances on the original dataset and each embedding. To quantify the preservation of distances, we computed the Pearson correlation coefficient between pairwise distances in the embedding and the original space. To facilitate visualization of this relationship, we binned the distances in the original space into 50 equal-width bins and used box plots to summarize pairwise distances on the embeddings conditional on the distance bin on the original space.

Reproducibility of large-scale structure. For each dataset and algorithm, we obtain two vectors of embedded coordinates (x,y). For each subsample, we also obtain such a pair of embedded coordinates (x',y'). From this we computed $(|\mathrm{cor}(x,x')| + |\mathrm{cor}(y,y')|)/2$, where $|.|$ denotes the absolute value and cor the computation of the Pearson correlation coefficient. This quantity thus measures the average correlation of coordinates between the full embedding and subsamples from various sizes, maximized across axial symmetries across the x axis and/or y axis.

Code availability. The source code used to generate main and supplementary figures is published at https://github.com/ebecht/DR_benchmark and bundled with input data at https://figshare.com/s/9c3a0136f12b97f1dadd. The UMAP source code used herein is available at https://github.com/lmcinnes/umap/archive/0.2.4.tar.gz.

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

Data availability. The results presented herein are based only on publicly available datasets. They are available using identifiers listed in **Supplementary Table 2**.

22. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

nature research

Corresponding author(s): Evan Newell

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used public data as described in the methods section. We notably used FlowRepository datasets (FR-FCM-ZZTM and FR-FCM-ZZPH) as well as a subset of the the scRNAseq dataset available at Figshare (865e694ad06d5857db4b). No specific software was used to facilitate the retrieval of these data
Data analysis	UMAP: umap-learn Python package version 2.4.0 (https://github.com/lmcinnes/umap/archive/0.2.4.tar.gz), Barnes Hut t-SNE: Rtsne R package version 0.13) (https://cran.r-project.org/web/packages/Rtsne/index.html), Fit-SNE (https://github.com/KlugerLab/Fit-SNE), scvis version 0.1.0 (https://bitbucket.org/jerry00/scvis-dev), PCA: stats R package version 3.5.0 (base R package), Phenograph : R package Rphenograph version 0.99.1 (https://github.com/JinmiaoChenLab/Rphenograph), AUCell: R package version 1.2.4 (https://bioconductor.org/packages/release/bioc/html/AUCell.html), FlowJo: Commercialized by Treestar Inc., version 10.5.0. Custom code written in R and available at https://github.com/ebecht/DR_benchmark as well as at https://figshare.com/s/9c3a0136f12b97f1dadd

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Nature Biotechnology: doi:10.1038/nbt.4314

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analysed during the current study are available in the figshare repository, https://figshare.com/s/9c3a0136f12b97f1dadd.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	The only statistical test we perform is for the algorithm runtime where we used 3 replicates. Variance in these replicates is very low compared to the runtime, and 3 replicates were thus enough to accurately show the difference in runtime across conditions.
Data exclusions	For the single cell-RNAseq dataset, we excluded small populations as well as two relatively big cell clusters that appeared to be artifacts. These two big clusters were presented in one sample each and expressed a very large number of unique genes at a low level. Exclusion criteria were not decided beforehand but we made the exclusion process clear in the text and in a dedicated supplementary figure.
Replication	We assessed our findings on three datasets and using different random subsamples in triplicates. All replication attempts were successful.
Randomization	There is no grouping in our study as we do not perform statistical tests based on subgroups.
Blinding	There is no blinding in our study as we do not perform statistical tests or decision based on subgroups.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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

Nature Biotechnology: doi:10.1038/nbt.4314