AUTOMATED OPTIMAL PARAMETERS FOR T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING IMPROVE VISUALIZATION AND ALLOW ANALYSIS OF LARGE DATASETS

ANNA C. BELKINA[1,2,*], CHRISTOPHER O. CICCOLELLA[4], RINA ANNO[5], RICHARD HALPERT[6], JOSEF SPIDLEN[6], JENNIFER E. SNYDER-CAPPIONE[2,3]

[1]Department of Pathology, [2]Flow Cytometry Core Facility and [3]Department of Microbiology, Boston University School of Medicine, Boston, MA; [4]Omiq, Inc, Santa Clara, CA, [5]Department of Mathematics, Kansas State University, Manhattan, KS, [6]FlowJo, LLC, Ashland, OR

*CORRESPONDING AUTHOR: ANNA C. BELKINA, M.D., PH.D. BELKINA@BU.EDU

Accurate and comprehensive extraction of information from high-dimensional single cell datasets necessitates faithful visualizations to assess biological populations. A state-of-the-art algorithm for non-linear dimension reduction, t-SNE, requires multiple heuristics and fails to produce clear representations of datasets when millions of cells are projected. We developed opt-SNE, an automated toolkit for optimal t-SNE parameter selection that utilizes Kullback-Liebler divergence evaluation in real time to tailor the early exaggeration and overall number of gradient descent iterations in a dataset-specific manner. The precise calibration of early exaggeration together with opt-SNE adjustment of gradient descent learning rate dramatically improves computation time and enables high-quality visualization of large cytometry and transcriptomics datasets, overcoming limitations of analysis tools with hard-coded parameters that often produce poorly resolved or misleading maps of fluorescent and mass cytometry data. In summary, opt-SNE enables optimal data resolution in t-SNE space and more precise data interpretation.

Keywords: dimensionality reduction, data visualization, t-SNE, mass cytometry, flow cytometry, scRNA-seq, cytometry analysis, machine learning, viSNE

Visual exploration of high-dimensional data is imperative for the comprehensive analysis of single cell datasets. Fluorescence, mass and sequencing-based cytometric data analysis requires tools that are able to reveal the combinations of proteomic and/or transcriptomic markers that define complex and diverse cell phenotypes in a mixed population. While traditional biaxial data presentation via expert-driven gating is still the standard analysis method for cytometry data to date, with the advent of the modern multi-parameter era an analysis tool that can accurately and comprehensively visualize multi-dimensional data is direly needed to relieve the current cytometry data-processing bottleneck.

To date, multiple dimensionality reduction techniques have been applied to cytometry data with variable success. Linear methods, such as principal component analysis (PCA) which generates a low-dimensional representation of data with a linear mapping matrix, are mostly unsuitable for cytometry data visualization as such techniques cannot faithfully present the non-linear relationships in the data. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a state-of-the-art dimensionality reduction algorithm for non-linear data representation that creates a low-dimensional distribution, or a 'map', of high-dimensional data [1,2]. Conspicuous groupings of datapoints, or 'islands', correspond to observations that are similar in the original high-dimensional

space and help to visualize the general structure and heterogeneity of a dataset. t-SNE was developed as a machine learning technique for a broad range of data types and has been adopted [2] for single-cell applications. When t-SNE embeds single cell data, the islands represent cells with similar phenotypes, as defined by a cytometric or genomic signature, thereby allowing to reveal biological data structure and to surface important differences between samples and/or subject groups [3].

In addition, t-SNE maps are used to categorize single cell data into relevant biological populations for downstream quantification, achievable through expert-guided filtering ('gating')[4] or unsupervised clustering of the map [5-7]. Cytometry clustering algorithms that directly interrogate high-dimensional data, such as FlowSOM [8] and PhenoGraph [9], are often used in conjunction with t-SNE maps to present annotated clusters to the viewer.

A limitation of t-SNE in its current form is its inability to scale to datasets with large numbers of observations [7,10]. This restrains t-SNE's utility for cytometry datasets that often include millions of observations ('events') routinely collected for phenotypic analysis. Unlike PCA, t-SNE learns the embedding non-parametrically, and hence new pieces of data cannot validly be added to an existing analysis, necessitating the whole dataset to be analyzed within one computation. When the full dataset is comprised of multiple samples, each representing a subject in a large cohort or an independent experimental condition, retaining statistically significant representation of small subpopulations in each sample requires inflating the dataset size [11]. However, even within a single measurement, subsampling the data risks preventing rare subsets from being identified. These limitations cannot be overcome via application of the currently understood best practices for t-SNE use. Not only are large datasets computationally expensive to analyze, but also the resulting t-SNE maps provide poor visualization and incomprehensive representation of high-dimensional data. As a result, upon finding initial poor t-SNE visualizations from large datasets, researchers often resort to either subsampling their data to the very limit of detection of rare populations [12] or to exporting specific populations from their dataset, thus compromising the 'unbiased' data analysis approach[13,14].

Although t-SNE has been widely adopted by the scientific community, to our knowledge no rigorous theoretical or empirical testing of t-SNE for cytometry applications has been performed. In 2013, Amir et al first reported the use of t-SNE (or viSNE, as it was renamed [3]) on mass cytometry data; since then, t-SNE has been implemented in the majority of commercial and open-source platforms for cytometry analysis (FlowJo, Cytobank, FCSExpress, cytofkit, etc). In most implementations, few or no adjustments were made to the Barnes-Hut t-SNE algorithm for the requirements of cytometry datasets; the default and hard-coded parameter settings that were originally tested and optimized with non-cytometry datasets like CIFAR (image dataset) or MNIST (handwritten digits) are retained in these cytometry programs. Developing rigorous methodology to release the full potential of t-SNE for single cell data comprehension is the primary motivation for this work.

In this study, we first assessed the behavior of t-SNE computation with routinely used settings that match common best practices, and then iteratively modified parameters of embedding to identify conditions that ensure optimal visualization. As a result of this work, we propose a new method to automatically find optimal t-SNE parameters via fine-tuning of the early exaggeration stage of t-SNE embedding in real time. We call our approach opt-SNE, for *opt*imal t-SNE. We find that our adjustments can tremendously shorten the number of iterations required to obtain

visualizations of large cytometry datasets with superior quality. Our approach also eliminates the need for trial-and-error runs intended to empirically find the optimal selection of t-SNE parameters, potentially saving hundreds of hours of computation time per research project.

Materials and methods

Datasets. All datasets used in the study are summarized in Table 1.

Table 1. Datasets used in this manuscript

Dataset	Data type	Details	References
Mass41parameter	Mass cytometry	41 parameter dataset (14 parameters used for embedding) of 1 million datapoints concatenated from 5 samples of human bone marrow cells	[15]
Flow18parameter	Flow cytometry	18 parameter dataset (11 parameters used for embedding) of 1 million datapoints concatenated from 2 samples of human PBMC	[16]
Flow20M	Flow cytometry	18 parameter dataset (15 parameters used for embedding) of 20 million datapoints concatenated from 27 samples of human PBMC	Panel based on [16]
10X Genomics	scRNA-seq	Single cell gene expression data of E18 mouse brain pre-processed into 20 PCA projections used for t-SNE embedding	https://support.10xgenomics.com/single-cell-gene-expression/datasets and [17]
van Unen et al	Mass cytometry	18 parameter dataset (14 parameters used for embedding) of 5.22 million datapoints concatenated from 104 samples of human peripheral blood mononuclear cells (PBMC) and gut biopsy cells	[7]

Primary samples. Flow18parameter and Flow20M data were collected as described [18] with minor modifications of the flow cytometry reagent list. Study protocols were approved by the institutional review boards and all subjects provided written informed consent. The committee that approved the research protocol was the Boston University Institutional Review Board, IRB# H-33095.

Data pre-processing. Singlet events from several data recordings were digitally concatenated and a randomly subsampled file of 1,000,000 mass [15] or flow cytometry [16] events was created and used for analyses of the mass41parameter and flow18parameter datasets. All observations from 27 recordings of flow cytometry data were concatenated to generate the flow20M dataset. All observations from 104 recordings of mass cytometry data [7] were concatenated to generate the van Unen et al dataset.

All flow cytometry data were compensated with acquisition-defined compensation matrices. Prior to t-SNE analysis, all cytometry data were transformed using asinh (all mass cytometry data and flow14parameter data) or biexponential (flow20M) transformation. Light scatter parameters were log-transformed.

Data analysis. A desktop C++ Barnes-Hut implementation of t-SNE for Mac OS was used for most t-SNE analyses [2]. All datasets were embedded in 2D space. Original code was edited to allow user input for early exaggeration stop iteration, perplexity, total number of iterations, early exaggeration factor value, and learning rate value. KLD value and t-SNE coordinates were reported during each iteration or as frequently as requested. To allow generation of visually comparable t-SNE maps, the same random seed value was used and all experiments were repeated with several values of random seed. We did not observe noticeable differences in reported results between runs initiated with different seed values. For cross-validation and to benchmark against standard platforms, we utilized cloud-based Cytobank [19], FlowJo V10.3-10.5 and FlowJo V9.9.6. Cytobank and FlowJo platforms were used to generate FCS files and graphical outputs from tabular data. FIt-SNE (Fast Fourier Transform-accelerated Interpolation-based t-SNE [20] plugin for FlowJo was used as indicated. Logs of t-SNE runs were batch-processed with VBA scripts and analyzed with GraphPad Prism 7. Expert-guided (manual) analysis of cell populations was performed in FlowJo 10.3-10.5 as described previously for specific datasets [15,18] or as explained below and used for map annotations as cluster classifiers.

For scRNAseq analysis, we used SeqGeq 1.3 package. We used PCA projections provided in 10X Genomics dataset to calculate the t-SNE embedding and annotated it using marker genes for major cell types. Louvain cluster classification was adopted from the SCANPY data analysis study [17]

The quality of the embeddings was assessed by a nearest neighbor (NN) classifier strategy as described in previous reports on t-SNE accuracy evaluation [1,2,10]. Briefly, for each observation, the k nearest neighbors (by Euclidean distance) were calculated using the 2D coordinates of the t-SNE map and the class assigned by expert gating was compared to the most common class of its k neighborhood. The rate of correct matches was tallied and represented as the overall nearest neighbor accuracy. The accuracy was also calculated on a per-class basis; different values for k (1,10,20,30,40,50) were reported.

The standard t-SNE configuration for cytometry applications. As described in detail in van der Maaten 2014, t-SNE computes low-dimensional coordinates of high-dimensional data resulting in similar and dissimilar data points in the raw data space placed proximally and at a distance, respectively, in the dimensionally reduced map. This map placement is achieved via t-SNE modeling the probabilities as a Gaussian distribution around each data point in the high-dimensional space and modeling the target distribution of pairwise similarities in the lower-dimensional space using Cauchy distribution (Student t-distribution with 1 degree of freedom). Then, the Kullback-Liebler Divergence (KLD) between the distributions is iteratively minimized via gradient descent. The gradient computation is essentially an N-body simulation problem with attractive forces (approximated to nearest neighbors using vantage-point trees) pulling similar points together and repulsive forces (approximated at each iteration using the Barnes-Hut algorithm) pushing dissimilar points apart.

An important part of t-SNE gradient descent computation is the "early exaggeration" (EE) that was proposed by van der Maaten and Hinton (2008) to battle the "overcrowding" artifact of embedding. With EE, all probabilities modeling distances in high-dimensional space are multiplied by a factor (early exaggeration factor, EEF, or α) for the duration of the first (typically 250, or 25% of the total number of) iterations. EE coerces data to form tight and widely separated clusters in the map and is considered to enable the map to find a better global structure.

Multiple software platforms incorporate t-SNE algorithm specifically for analysis of cytometry data, including commonly used cytometry analysis desktop packages (FlowJo, FCS Express), and the cloud-based analysis platform Cytobank. Also, implementations of t-SNE are available as open-source packages in popular programming languages such as R (rtsne) and Python (sci-kit learn). Most of implementations wrap or re-write original C++ code of Barnes-Hut t-SNE (van der Maaten, 2014) and produce comparable analysis results upon direct comparison. Here, we customized the standard t-SNE C++ code to implement the parameter adjustments described in this work and published this customization as an open source solution to enable the research community to use this optimized t-SNE algorithm. Also, the equivalent adjustments of t-SNE have been made available as a cloud application from Omiq and integrated into FlowJo and SeqGeq programs.

Results.

The standard t-SNE configuration fails to visualize large datasets. The t-SNE algorithm can be guided by a set of parameters that finely adjust multiple aspects of the t-SNE run [21]. However, cytometry data analysis software often locks or severely restrains the tunability of those parameters, likely to provide a simplified, 'one-size-fits-all' solution for t-SNE use in the software packages. Although each software platform has a unique combination of possible adjustments, most allow changes to both the number of iterations and to the perplexity (a soft measure for the number of nearest neighbors considered for each data point).

The datasets used throughout this study include at least 1 million datapoints of fluorescent or mass cytometry data and are therefore considerably larger than the smaller ($< 5x10^5$) datasets previously reported in benchmark studies of cytometry algorithmic tools [22]. Cytometry datasets larger than approximately $5x10^5$ events are generally observed to produce poor quality t-SNE maps and are therefore usually subsampled prior to analysis.

Empirically, cytometrists have observed that increasing the number of iterations of t-SNE computation results in better quality maps. We hypothesized that the resolution of t-SNE maps created from higher event counts could be dramatically improved via fine-tuning of t-SNE parameters. We first directly tested the relationship between iteration number and map quality by running two datasets (mass41parameter and flow18parameter, as described in Table 1) at the default 1000 iterations per run and with an "extended" 3000 iteration computation (Fig. 1A, B). To

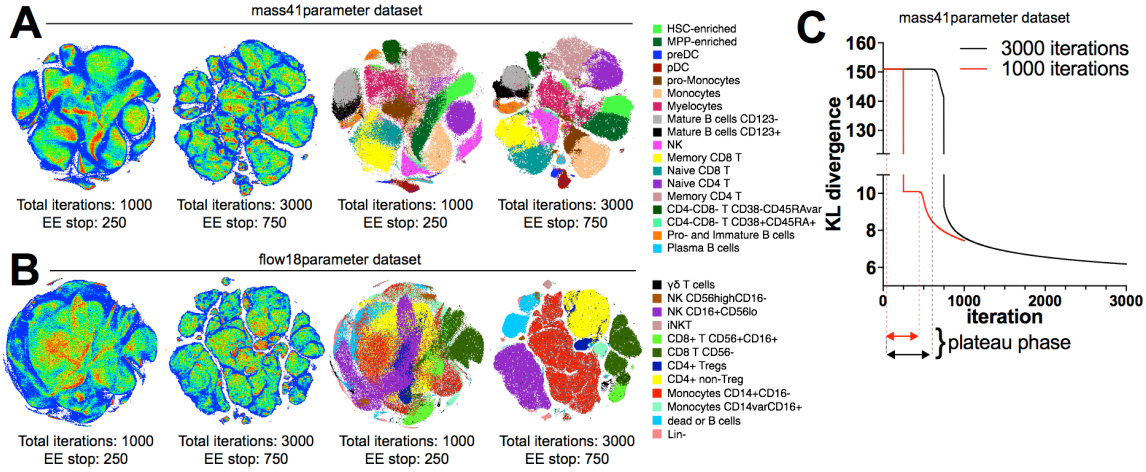


Figure 1. Performance of Barnes-Hut t-SNE implementation for cytometry data visualization showing comparison of standard (1000 iterations) and extended (3000 iterations) embeddings of mass cytometry (A) or flow cytometry (B) data presented as heatmap density plots (left) or color-coded population overlays based on ground-truth classification of single cell in the datasets. C.KLD change over iteration time of gradient descent for standard 1000 iterations (red line) or extended 3000 iterations (black line) embeddings of mass41parmeter dataset. Representative examples of multiple runs with varying seed values are shown.

aid visualization in Fig 1 and the following figures maps are overlaid with color-coded populations derived from expert-driven ('manual') gating of the same dataset which serves as a ground-truth basis for data classification. As expected, 1000-iteration runs produced maps with poor visualization (overall 1-NN accuracy of embedding was 65% and as low as 18% for certain populations). Specifically, massive overlaps and random fragmentation of populations were observed. In contrast, the 3000-iteration runs resulted in maps with defined "islands" comprised of clearly isolated populations and no random fragmentations (overall 1-NN accuracy of embedding 96%; see Suppl. Table 1 for the detailed results of accuracy evaluations). Therefore, these findings are in agreement with the concept of a higher number of iterations resulting in higher quality t-SNE maps.

KLD plateau phase resolves global cluster structure in t-SNE visualization. In order to determine the cause of the difference in cluster resolution between the "default" and "extended" t-SNE runs, we examined the behavior of KLD over the duration of t-SNE embeddings (Fig 1C). As expected, the KLD value was inflated during the EE since EEF is factored into gradient and KLD value calculation [1]. The EE is applied over 250 iterations in the "default" (red line, Fig. 1C, 1000 iterations) t-SNE configuration and 750 iterations in the "extended" run with 3000 iterations (black line Fig. 1C, 3000 iterations) since most platforms have EE scaled to 25% of total iteration number.

Notably, the KLD did not immediately minimize at the start of the EE in both the "default" and "extended" t-SNE runs; instead, the graph of KLD over time is a plateau that is followed by a curve that captures the incremental decrease of KLD, indicating the gradient descent. In the "default" run, the plateau was interrupted when the EE was stopped and KLD dropped, then continued with a non-exaggerated value of KLD.

According to van der Maaten and Hinton, EE was introduced as a "trick" to improve resolution of the global structure of the data visualization that would not otherwise converge to separated clusters. As the suboptimal quality of the 1000-iteration t-SNE maps shown in Fig. 1A, B demonstrates poor global structure resolution, we hypothesized that by increasing the total

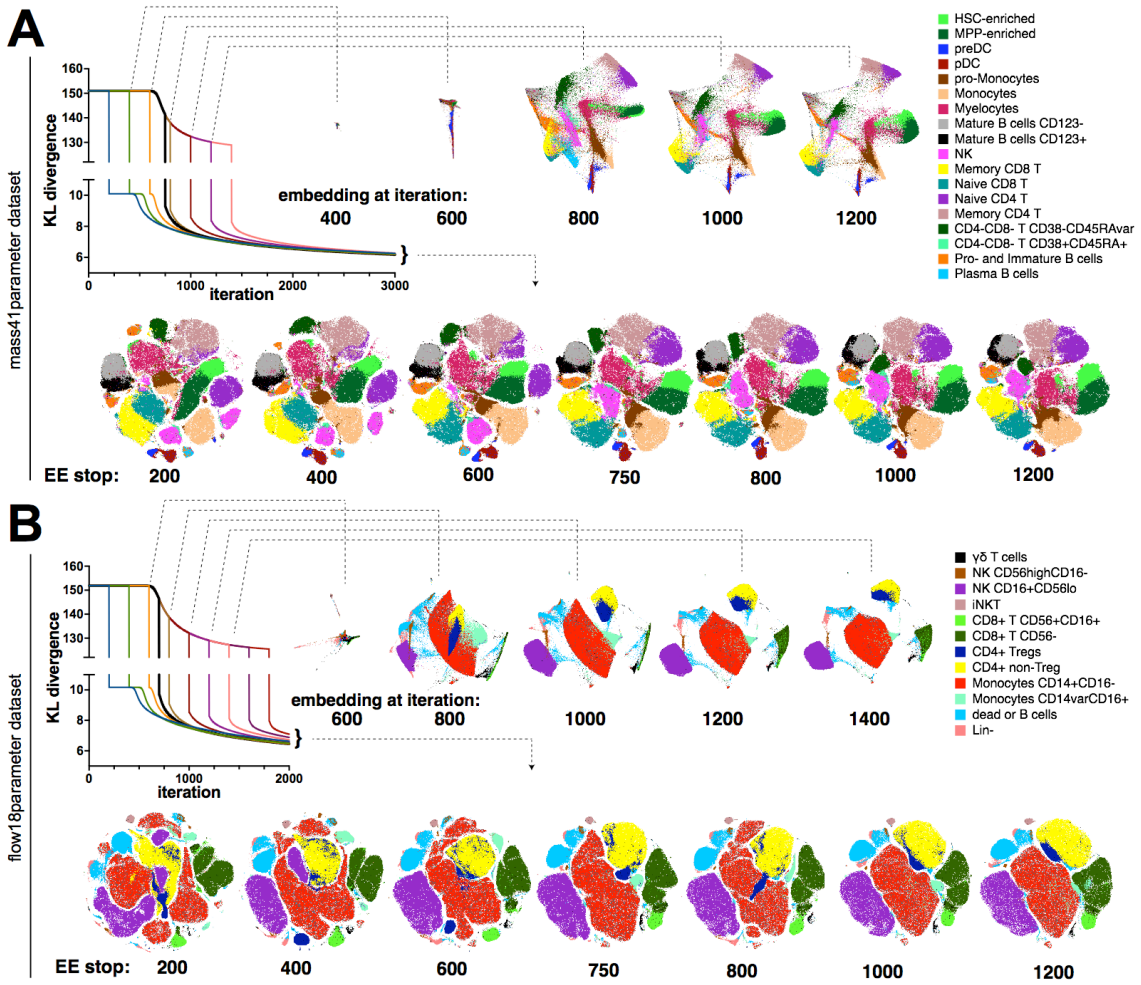


Figure 2. Effect of EE plateau phase on t-SNE visualization. EE was stopped after varying number of iterations and embedding output was examined at several intermediate timepoints and in the end of embedding for mass cytometry (total of 3000 iterations, A) and fluorescent cytometry (total of 2000 iterations, B) data visualization. Graphs showing KLD change over iteration time are color-labeled to distinguish curves corresponding to experiment perturbations, with black line indicating the run with the shortest EE but uninterrupted plateau. t-SNE maps are annotated with color-coded population overlays based on ground-truth classification of single cell in the datasets. Representative examples of multiple runs with varying seed values are shown.

iterations an analyst who uses t-SNE via conventional cytometry analysis platforms may inadvertently increase the number of EE iterations and this specific alteration may be the cause of the improvement in map visualization. To test this hypothesis, we compared multiple 2000- or 3000-iteration runs that differed in timing of the EE stop (Fig. 2A) by plotting the embedding at the iteration both when the EE stops and at later iterations, thus assessing the effects of EE and our perturbations on both mass cytometry (Fig. 2A) and flow cytometry (Fig. 2B) data visualization.

We found notable differences in map quality between the shorter and longer EE runs. Although the map after EE200/total3000 iterations appears visually superior than EE250/total1000 (Fig 1A, B) and could be considered a successful visualization, ground-truth labeling indicated that cluster fragmentation is present in both maps. When cluster fragments were plotted on a biaxial plot against parameters that were used in the t-SNE dimension reduction, we were not able to identify parameters that immediately contributed to their fragmentation (Suppl. Fig. 1).

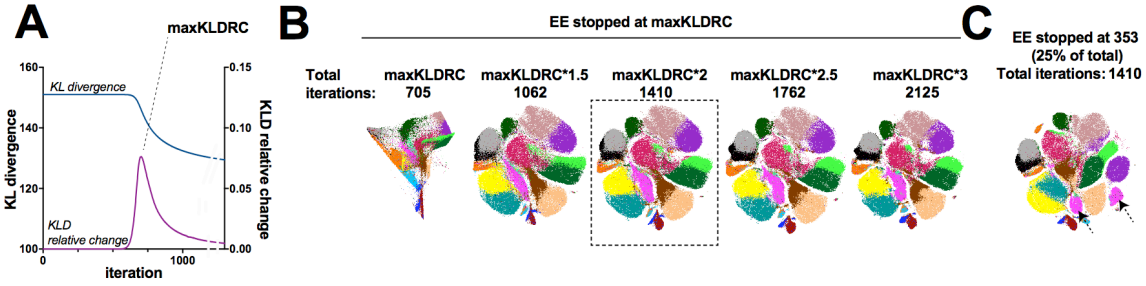


Figure 3. Early exaggeration plateau ensures optimal quality of visualization. A, KLD and KLD relative change plotted against iteration time. B. Mass cytometry data visualiztions generated with varying duration of post-EE iteration time. Dashed box indicates visualizations used for comparison with default setup. C. Mass cytometry data visualization generated with default (25% of total) duration of EE. Arrows mark cluster fragmentation. All color overlays correspond to cell type classes labeled as in Fig. 1-2. Representative examples of multiple runs with varying seed values are shown.

Conversely, tight clusters that form at the end of the plateau remain mostly unchanged as long as the EE is being applied to the computation (Fig 2 A, B). The KLD minimization in that case could be explained by the gradual shrinking of the 2D space (data not shown). Once EE is removed, the attractive forces within each cluster are weakened and the local structure of the data is fully resolved within each cluster. Overall, these observations suggest that the EE stage of the gradient descent is essential for data clustering while the non-exaggerated descent results in resolution of local structures.

Real-time monitoring of the KLD plateau results in optimal quality of t-SNE maps. We have demonstrated that when the EE is too short, cell clusters continue to be resolved simultaneously with the local structure of each cluster being unfolded, leading to fragmented, overlapped or deformed "islands" in the resulting map (Fig. 2). Due to these results, we constructed an equation to find optimal EE timing. Specifically, we tracked the relative rate of KLD change ($KLDRC_N = 100\% *((KLD_{N-1} – KLD_N)/ KLD_{N-1})$ where N is the iteration number) and identified the local maximum of KLDRC (maxKLDRC) (Fig. 3A). Since KLD is computed at each iteration, the maxKLDRC 'sensor' can be easily added to the algorithm programmatically and would stop EE at the next iteration past maxKLDRC. For the mass41parameter dataset of 1M datapoints, the maxKLDRC was detected at iteration 705 (Fig. 3A). Next, we ran t-SNE with an EE stop at iteration 706 and sampled map development at 706, 1.5x706, 2x706, 2.5x706 and 3x706 iterations (Fig 3B). As expected, at the maxKLDRC iteration the map contained the primordial clusters only; it was well shaped at 2 x max KLDRC and there was no visible improvement in map quality past that step and the visualization was very similar to the EE750/3000 map at Fig. 1A. When compared to the map created with 'default' settings of EE taking 25% of the run yet computed with the same number of iterations (1410), the maxKLDRC-triggered t-SNE produced visually and KLD-wise superior results within the similar computation time, and it also eliminated extensive trial-and-error calibration of t-SNE parameters (Fig. 3C). We propose a conservative approach to finalize the embedding automatically when $(KLD_{N-1} – KLD_N) < KLD_N/10{,}000$. Alternatively, t-SNE projection output can be evaluated in real time to justify the termination of embedding.

Moderate adjustments of EE factor and perplexity do not impact visualization. Once we found EE to be crucial for map optimization, we next examined if the value of the EE factor α can also be tuned to improve the results of t-SNE. We made α user-accessible in our C++ t-SNE

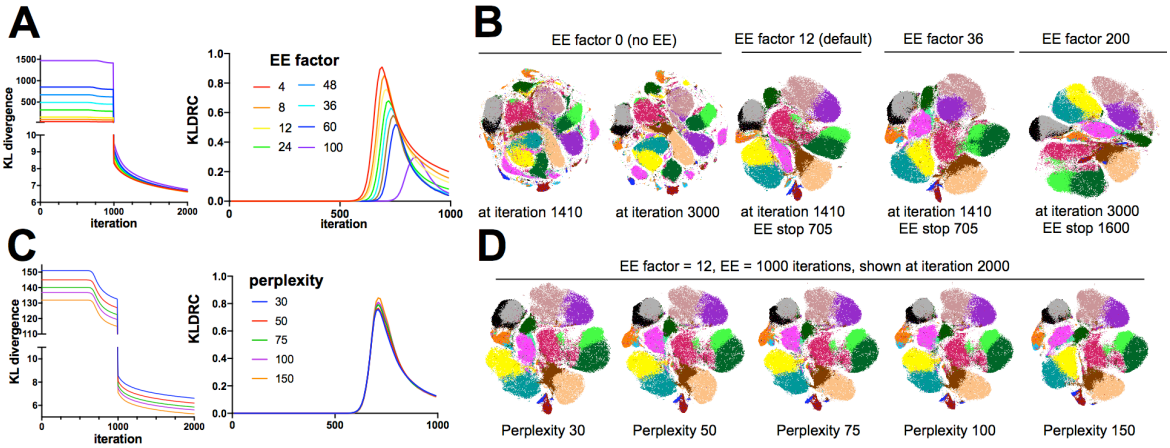


Figure 4. Effects of perplexity and EE factor adjustments on t-SNE visualization of cytometry data. A, B. KLD, KLDRC and t-SNE biaxial plots generated with varying EE factor values. C, D. KLD, KLDRC and t-SNE biaxial plots generated with varying perplexity. Graphs showing KLD and KLDRC change over iteration time are color-labeled to distinguish curves corresponding to experiment perturbations. All color overlays on t-SNE plots correspond to cell type classes labeled as in Fig. 1-2. Representative examples of multiple runs with varying seed values are shown.

code since it is hard-coded in the original Barnes-Hut C++ t-SNE implementation and all aforementioned results were obtained with default value of $\alpha = 12$. We chose the parameters defined above ($\alpha = 12$, EE = 706 iterations, 1410 iterations total) as our baseline for comparison since they provided optimal balance of map quality versus computation time. First, we tested how the optimization would proceed without EE ($\alpha = 1$). We expected the run to fail or produce extremely crowded results as explained in the original t-SNE report [1]; however, we did not see much overlap in cluster positioning, probably due to the substantial number of map iterations run (Fig 4B). Nevertheless, the resulting map showed a lot of fragmentation proving to be an extreme case of an interrupted plateau phase. Even when run for as many as 3000 iterations, the fragmentation could not be remedied, again demonstrating the necessity of EE.

As expected, higher values of α lead to much higher KLD during EE, however, the KLD values were similar at 2000 iterations when α varied between 4 and 60 (Fig. 4A). Larger α prolonged the plateau phase and became detrimental for KLD values when over 100. Visually $\alpha = 200$ results in a distorted map with smaller populations lost. Therefore, we suggest that for cytometry applications the α parameter may remain unchanged and set to 12, as suggested in van der Maaten 2014, or reverted to $\alpha = 4$, as originally proposed in van der Maaten and Hinton (2008) since per our results any value between 4 and 20 leads to comparable outcomes.

Increased perplexity has been proposed to be an intuitively beneficial method for visualization improvement since it translates to a larger number of considered nearest neighbors and hence a more accurate approximation of attractive forces, while decreased perplexity can completely fail the visualization [21]. KLD values for runs with varying perplexity cannot be directly compared since the KLD value is related to perplexity; however, KLD records over time do not show that increased

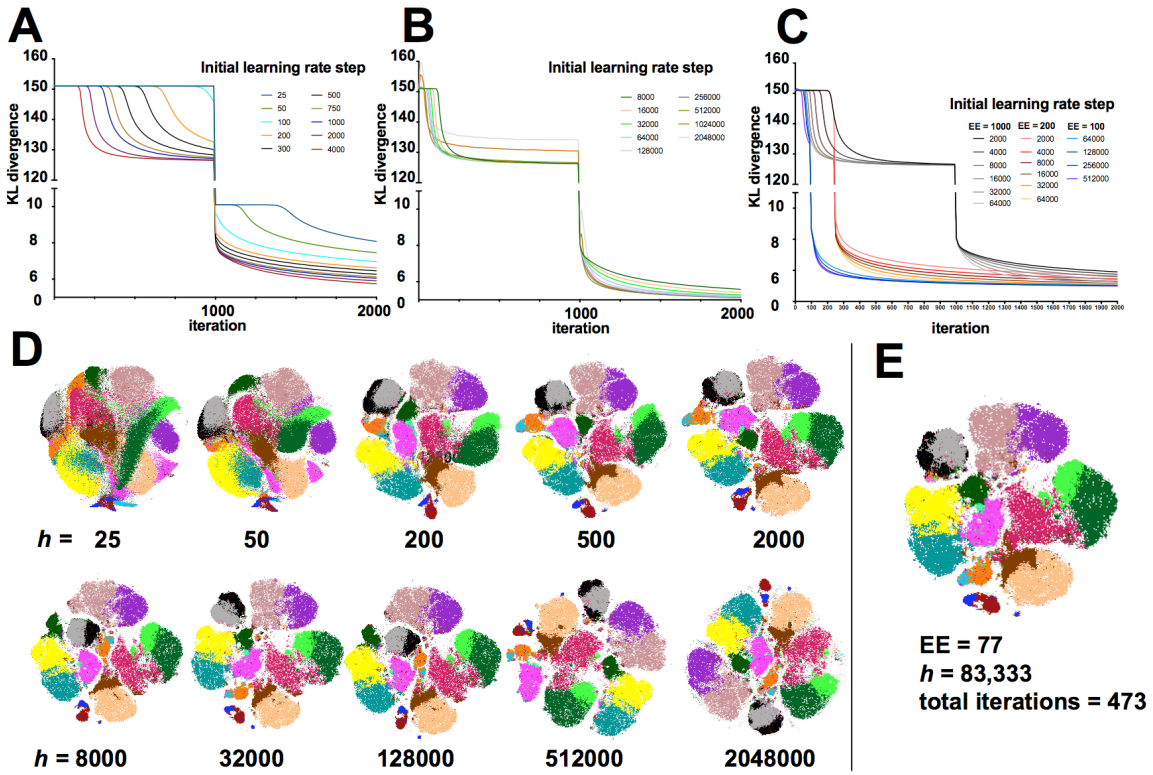


Figure 5. Learning step size optimization for t-SNE visualization of large datasets. A, B, C KLD change over iterations for embeddings with varying values of initial learning rate step size, color coded as indicated. A, EE = 1000 iterations, learning rate step = 25-4000; B, EE = 1000 iterations, learning rate step = 8K-2048K; C, EE = 100-1000 iterations, learning rate step = 2K - 512K. D, representative t-SNE plots of embeddings graphed on A. E, t-SNE plot of an optimal embedding. All color overlays on t-SNE plots correspond to cell type classes labeled as in Fig. 1-2. Representative examples of multiple runs with varying seed values are shown.

perplexity results in faster resolution of clusters (Fig 4C) or cleaner data visualization (Fig 4D). However, while changing α does not affect t-SNE computation time, perplexity is linearly related to the time and memory required to create the embedding (data not shown). Although we and others have found some benefits of perplexity increases to map quality in otherwise suboptimal t-SNE runs, optimizing the EE step as described above and further in this work does not leave much space for improvement with perplexity tuning (Fig 4E).

Learning step size is a key parameter to ensure t-SNE visualization of large datasets. The step size in t-SNE gradient descent is updated at each iteration per Jacobs adaptive learning rate scheme [23]. This method increases the learning rate in directions in which the gradient is stable. A conservative initial value of 200 is hard-coded into most platforms. We hypothesized that larger datasets may stay longer on KLD plateau due to the number of iterations it takes to build up a sufficient learning rate step size. To evaluate this possibility, we titrated the step size *h* while observing the KLD with fixed EE=1000 iterations in Mass41parameter dataset. In agreement with our hypothesis, *h* = 25 and *h* = 50 runs failed to resolve from KLD plateau within 1000 iterations of EE (Fig. 5A) and *h* = 200 finished the plateau in ~700 iterations as previously shown. With further increases in *h,* we found that not only are progressively fewer iterations required to complete the plateau, but also that the final KLD of the maps scored at lower values. KLD is directly related

to the quality of visualization since it reflects the faithfulness of representation of high-dimensional data in t-SNE space; therefore, lower KLD values indicate superior visualization quality.

We continued to see improvement in plateau duration and KLD values with higher h values up to $h \sim 64{,}000$, a value that is drastically far from the "default" $h = 200$ setting (note that in most platforms, h is restricted to ranges below 3,000) (Fig. 5B). At $h \sim 256{,}000$ we observed irregular peaks in KLD graph indicating that the prescribed step size rendered gradient descent ineffective. However, using lower values of h we were able to converge the map with lowest KLD values at a fraction of time when limiting the EE step to 200 and even 100 iterations despite the 10^6 size of the dataset (Fig. 5C). Visual inspection of the embedded maps over the range of h values agrees with KLD values (Fig 5D).

In a recent publication, Linderman and Steinerberger [24] prove that in general t-SNE embedding will not converge if a product of EE factor α and of fixed learning rate step size h is larger than the number of datapoints n (i.e. if $\alpha h > n$). Since we employ an adaptive learning rate, our selection of initial h value is more forgiving; however, in our experiments we found the optimal settings of h to be close to $h = n / 12$ for computations where $\alpha = 12$. Therefore, we propose to initiate the gradient descent with $h = n / \alpha$ to create optimal t-SNE visualization (Fig 5E).

Opt-SNE allows optimal embedding of massive datasets. We implemented all proposed techniques including: (1) a dataset-specific automated early exaggeration step controlled by the KLDRC sensor, (2) an optimal learning rate step size, and (3) a KLDRC-driven embedding termination into a single workflow labeled 'opt-SNE', for 'optimized t-Stochastic neighbor embedding'. To test the performance of opt-SNE, we used a 20.1 x 10^6 event fluorescent cytometry dataset concatenated from two independent batches of PBMC samples (N=27) stained with a variation of the OMIP-037 fluorescent cytometry panel [18] that allows detailed assessment of naïve and memory CD4+ and CD8+ T cells, NK cells and γδ T cells (Fig. 6A). The embedding completed in 770 iterations with only 73 iterations required to pass the EE step (Suppl. Fig. 3) and resulted in clear separation of cell clusters as evaluated by cell type annotation (Fig 6C). The majority of clusters appear to be populated by cells from all subjects with the exceptions of several populations that contained sample-unique debris features (Fig 6B, dashed arrows), confirming an absence of batch effects. A detailed breakdown of identified populations is presented in Fig 6C that shows subsets of CD4+ and CD8+ T cells, NK cells, γδ T cells, B cells, and monocytes. Importantly, B cell and monocyte lineage markers were detected together with a viability dye in a dump channel in this panel and therefore cannot be gated accurately via traditional biaxial plot analysis. However, opt-SNE identified them in high-dimensional space through the combination of several surface labels and light scatter characteristics, and each was successfully clustered into populations that were minimally mixed with dead cells. Use of the standard t-SNE algorithm completely failed to reveal the structure of the multi-million event flow cytometry dataset, even with several thousands of iterations (Fig 6D).

In order to test the suitability of opt-SNE for applications beyond flow and mass cytometry, we analyzed a 1.3x10^6 cell single-cell RNA-seq dataset of mouse embryonic brain cells published by 10X Genomics. We used pre-calculated PCA projections included in the dataset to generate opt-SNE maps that we compared with 10X standard t-SNE embedding (Fig 6E). 10X used EE = 1000/total 4000 iterations of standard t-SNE while we used opt-SNE settings with $h = 97{,}959$, EE = 66/total 885 iterations (Suppl. Fig. 2). Non-immune single cell transcriptomics data are more

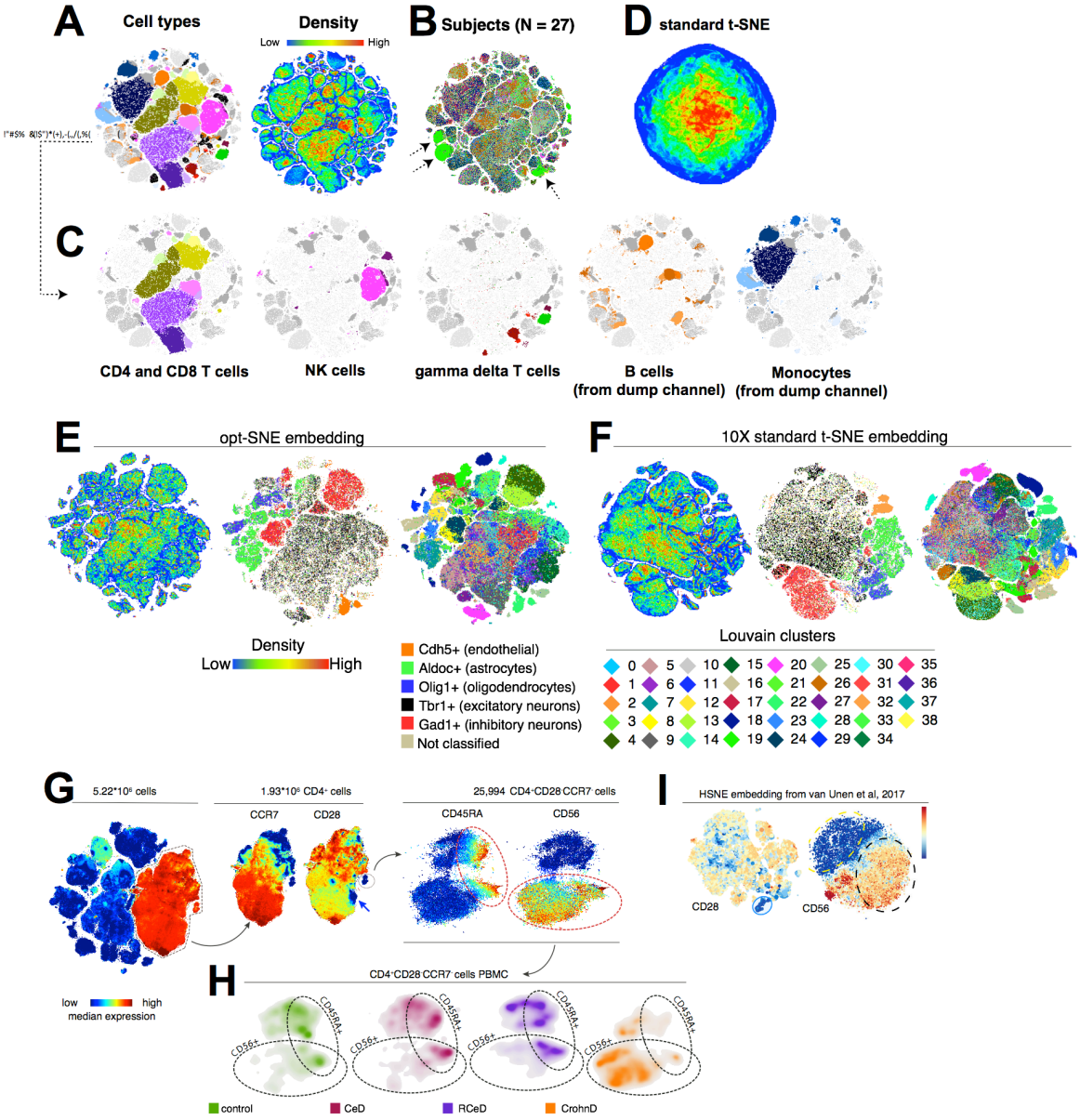


Figure 6. opt-SNE allows high-quality visualization of large cytometry and transcriptomics datasets. **A-D**: 20 million datapoints from fluorescent cytometry dataset concatenated from 27 subjects vizualized in 2D space. **A** and **C**, cell type classes and density overlaid on 2D opt-SNE embedding. **B**, subject identifier overlaid on 2D opt-SNE embedding. Dashed arrows indicate clusters represented by datapoints from a single subject; **D**, standard t-SNE visualization (4000 iterations). **E-F**, 10X Genomics mouse brain scRNA-seq dataset (1.3 million datapoints) visualized in 2D space with opt-SNE (E) or standard t-SNE (F). From left to right: density features, single gene classes, and Louvain clusters (0-38) overlays. **G,** 5.22 million datapoints from mass cytometry dataset used in van Unen et al (2017) visualized in 2D space with opt-SNE. From left to right: CD4 expression overlaid on opt-SNE embedding; CCR7 and CD28 expression overlaid on CD4+ opt-SNE cluster; CD45RA and CD56 expression intensity overlaid on CD4+CD28-CCR7- cluster. **H,** CD4+CD28-CCR7- cells from control, celiac disease (CeD), refractory celiac disease (RCeD) and Crohn's disease (CrohnD) subjects presented on density plots. Dashed encirclements indicate CD45RA+ and CD56+ areas of the cluster as defined in G. **I,** hierarchical t-SNE (HSNE) embedding of the CD4 (left) and CD4+CD28- cluster (right) from van Unen et al (2017). Color indicates marker expression intensity.

difficult to interpret with ground-truth classes since much fewer scRNA-seq markers can be easily interpreted for population identification. Therefore, we utilized both single gene classification and classification through the Louvain algorithm clustering using the Scanpy Python package [17] to

annotate the data. While opt-SNE and t-SNE both capture the macro-structure illustrated by gene overlays, several Louvain clusters (3, 18, 19, 25, 28) appear partially or completely overlapped by other clusters in standard t-SNE but not in the opt-SNE embedding. Therefore, opt-SNE allowed equivalent or superior resolution of single cell transcriptomics data as with standard t-SNE but with ~5x smaller iteration time (885 iterations of opt-SNE vs 4000 iterations of standard t-SNE).

HSNE (hierarchical SNE) is a t-SNE adaptation that was recently reported to facilitate analysis of large datasets by constructing a hierarchy of embeddings that can be explored from the overview of 'landmark' populations up to a single-cell level or resolution [7]. We applied opt-SNE to the 5.2 million point dataset that was reported in HSNE analysis of mass cytometry data, and compared opt-SNE visualization to the full resolution level of HSNE embedding (Fig. 6G, Suppl. Fig. 3). In the CD4+ subset, opt-SNE visualization revealed two groups of CD4+CD28- cells likely representing terminally differentiated memory CD4+ T cells [25] with different levels of CCR7 expression that may reflect distinct differentiation states of the two populations. While HSNE allowed identification of CD4+CD28-CCR7- cells, it was unable to visualize CD4+CD28-CCR7+ cells as a single cluster and projected these cells sparsely in the CD4+ island (Fig.6I, left). Also, both algorithms projected heterogenous expression of CD56 in CD4+CD28-CCR7- cells, but HSNE embedding did not resolve separate CD56+ and CD56- clusters within CD4+CD28-CCR7- cells, loosely mapping them to the poles of the single round cluster (Fig. 6I, right). On the contrary, opt-SNE embedding visualized both the CD56+ and CD56- clusters and disparate CD45RA expression within each cluster, revealing distinct phenotypes for the control and diseased subject groups (Fig. 6H). Also, opt-SNE demonstrated that the CD56+CD45RA- cells in the cluster originate from several subjects with Crohn's disease (Fig. 6H). In summary, these results confirm that opt-SNE embedding provides superior visualization quality for complex cytometry data.

Discussion

Visual exploration of data drives hypothesis formation and/or serendipitous discoveries; therefore, t-SNE is an extremely valuable tool for data comprehension. It is often used to facilitate data perception when hypothesis generation is automated by robust computational methods [8,26]. Comparison of t-SNE embeddings from multiple experimental conditions, timepoints, or subjects is invaluable to visualize sample-to-sample differences including disease hallmarks and longitudinal observations [27]. t-SNE is also valuable for quality assessment of data, when abnormal clustering could be traced back to sample preparation, data acquisition and preprocessing artifacts [28]. Therefore, batch embedding of multiple experimental points is essential for sample comparison and can only be enabled when t-SNE accommodates large datasets.

t-SNE was first introduced in cytometry research as a tool to visualize CyTOF data, since fluorescence-based high-parameter datasets were less common at that time. With recent advances in instrumentation and reagent availability, flow cytometry datasets with >20 parameters are quickly becoming prevalent and even standard in the field[29-31], yet the proper data assessment tools are lacking for general use. DNA-barcoded antibodies have been recently utilized to allow simultaneous protein-epitope and transcriptome measurements in single cells [32] thus expanding the repertoire of traditional cytometry methods that could employ t-SNE as a staple method of data visualization and presentation.

One approach to large dataset t-SNE embedding is to model visualization with a subset of datapoints curated [7,33] or randomly selected (as implemented in cytometry data analysis platforms such as FCS Express and Gemstone) from the nearest neighbor graph. Such techniques may fail to project extremely rare datapoints, as demonstrated by CD4+CD28-CCR7+ cells that were identified with opt-SNE, but not HSNE in van Unen et al dataset (Fig. 6). This T cell subset was significantly less abundant in subjects with severe inflammatory conditions including Crohn's disease as compared to controls (Suppl. Fig. 3), marking this population as extremely rare. On the other hand, the CD4+CD28-CCR7-, while also low in frequency in most PBMC samples (<1% of all CD4+ cells), was likely successfully clustered by HSNE because two of the 11 Crohn disease subjects in the dataset showed an unusually high frequency of this population (40.7% and 31.0% of all CD4+ cells), allowing it to be well represented in the kNN graph of the full dataset.

Several attempts to successfully apply t-SNE-like methods to massive datasets have been recently reported including aforenoted HSNE [7,33,34], LargeVis [10] and net-SNE [35]. However, these improved methods, when applied to large datasets, often require/benefit from considerable computational resources; for instance, the LargeVis study was performed on a 512Gb RAM, 32 core station. However, routine data analysis in a typical research laboratory should be possible with less available resources. We have not explicitly focused on computational efficiency in this work since we benchmarked the algorithm against itself with no specific emphasis on shortening computation time, which occurred due to the fewer number of iterations required to complete the data embedding with our adjustments. However, we have addressed several aspects of computation efficiency. For public use, we released an opt-SNE modification of multicore t-SNE C++ implementation (See Data and Software Availability) since the original Barnes-Hut t-SNE does not employ multi-threading. FIt-SNE, a recently published alternative to Barnes-Hut t-SNE that uses fast Fourier transform for much faster computation of repulsive forces approximation [20], renders opt-SNE even more feasible on personal computers; when we combined our automated opt-SNE setup with FIt-SNE approximation in the van Unen dataset analysis (Fig. 6G-H), it was completed in about two hours on a 16Gb RAM personal notebook with 2 cores. Notably, we have not observed differences in embedding quality between embeddings generated with fast Fourier transform versus Barnes-Hut approximations when opt-SNE was used to control the visualization (Suppl.Fig.4 and data not shown). All other analyses performed in this manuscript were performed using Barnes-Hut approximation on personal computers with the exception of the 20M embedding that required ~60Gb RAM at its peak and was run for several days on a multicore workstation using SeqGeq implementation of opt-SNE. Therefore, we expect opt-SNE to be applicable for existing or future adaptations of t-SNE even if alternative methods of computation are utilized [20,36] provided that they retain the core principles of t-SNE embedding. A promising approach that may be integrated with opt-SNE is the smart EE adjustment implemented in A-tSNE (approximated t-SNE)[34] algorithm where EE is removed gradually and on a per-point basis. Cytosplore, a novel software platform that includes HSNE and A-tSNE, allows the analyst to interactively initiate the local refinement of the map, resulting a significant improvement in computation time[7].

Similar to other types of biological data, the structure of cytometry data is difficult to project due to its mixed nature, often comprised of cluster-like, manifold-like and/or hierarchical components [28,37]. In this paper we propose multiple techniques that are essential for optimal t-SNE data projection and are all germane to the fine-tuning of the early exaggeration stage of t-SNE embedding. EE facilitates cluster formation on a 2D plane [1] and serves as a necessary compromise

that allows clusters to escape the crowding effect. We designed an efficient measure to ensure that the cluster-like global structure of the data is fully revealed during the EE stage by monitoring the KLD output of the embedding in real time. As indicated by the lower KLD values of opt-SNE embedding where EE was limited to fewer iterations (Fig. 5E), prolonged amplification of the attractive forces that drive tight cluster formation in EE may be detrimental for the manifold-like local data structure represented by signal distributions that are continuous with background signal. These 'continuum expression' molecules include the proteins used to define classic immune cell subsets, as well as those linked with activation and/or exhaustion and markers indicating disease phenotypes[38]. Conversely, the non-exaggerated stage of t-SNE allows local data structures to be revealed. [1]. Therefore, cytometry data analysis would be missing valuable information if we limited t-SNE applicability to finding only well separated clusters, especially since other techniques would perform that task better and faster. However, some workflows call for t-SNE pre-processing to facilitate extraction of cluster features from multidimensional data [5,39]. In those cases, it may be helpful to adapt the opt-SNE toolkit to terminate the embedding calculation immediately at the EE stop iteration and re-assess the high-dimensional structure within each cluster. Alternatively, a 'late exaggeration approach' [20] can be cautiously applied to create very tight clusters, although in our experience this approach was only marginally beneficial for global structure representation but detrimental for local structure (data not shown).

It is advisable to note that certain data structures, such as cluster hierarchy, cannot be revealed with t-SNE [40,41]. t-SNE accessibility in cytometry analysis software lead to its not infrequent misuse with cytometry data, evident when the cluster-like structure is not prominent in the map. Therefore, the features identified from t-SNE embedding in its current form should be verified with alternative methods when possible for confirmational purposes. Im et al also suggest that if a continuous manifold structure exists in the data, large perplexity values may cause artificial breaks (overclustering) [41]. The perplexity values commonly used in cytometry analysis are on the lower end of the suggested range for efficient clustering, as it is often advised to scale the number of nearest neighbors to the average cluster size [42]; however, if computationally feasible, higher perplexity values might facilitate feature preservation for markers whose expression is not bimodally distributed.

In summary, we believe that opt-SNE is a powerful optimization toolkit that removes major limitations of t-SNE use for cytometric datasets and could thus potentiate novel data-driven findings in single cell research.

Acknowledgements. The authors would like to thank Yvan Saeys, Gary Kazantsev, Jonathan Irish, Allison Irvine and Katherine Drake for helpful discussions, and Geoff Kraker for technical assistance. We thank Sean Bendall from Stanford University and Vincent van Unen from Leiden University Medical Center for sharing their data, and Brian Tilton and Riley Pihl from BUSM Flow Cytometry Core Facility for assistance with data collection at BUSM. Anna C. Belkina is an ISAC (International Society for Advancement of Cytometry) SRL Emerging Leader 2015-2019 and thanks the ISAC organization and members for continuous support and encouragement. Josef Spidlen is an ISAC Marylou Ingram Scholar.

Data and software Availability

All code including the open source multicore t-SNE C++ implementation, usage instructions, flow18parameters and mass41parameters datasets, and the cloud version of opt-SNE are available at http://www.omiq.ai/opt-SNE (the C++ code with a Python wrapper is available from https://github.com/omiq-ai/Multicore-opt-SNE).

Van Unen et al dataset is available at http://flowrepository.org/id/FR-FCM-ZYRM.

10X Genomics 1.3M scRNA-seq dataset is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets

To facilitate availability to flow cytometry and scRNA-seq data analysts, opt-SNE has been incorporated into FlowJo version ≥ 10.5.2 and SeqGeq version ≥ 1.4. At the time of this publication, this option was considered experimental and therefore hidden, but users can enable it by adding `<DRPlatform showAutoLearning="1" />` to the FlowJo10.prefs (or SeqGeq.prefs) XML file. It will be fully integrated and available by default in future releases of FlowJo and SeqGeq.

Author contributions

ACB has conceived the study. ACB and COC developed the opt-SNE method. COC developed implementation in C++ and RH and JS developed implementation in FlowJo and SeqGeq. ACB has analyzed flow and mass cytometry datasets. RH and JS have analyzed the scRNA-seq dataset. RA and JSC provided conceptual input. ACB wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests:

COC is a founder of Omiq, Inc. RH and JS are employees of FlowJo, LLC.

References

1 van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *The Journal of Machine Learning Research* **9**, 85 (2008).

2 Van Der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *The Journal of Machine Learning Research* **15**, 3221-3245 (2014).

3 Amir el, A. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* **31**, 545-552, doi:10.1038/nbt.2594 (2013).

4 Wong, M. T. *et al.* Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell reports* **11**, 1822-1833, doi:10.1016/j.celrep.2015.05.022 (2015).

5 Becher, B. *et al.* High-dimensional analysis of the murine myeloid cell system. *Nat Immunol* **15**, 1181-1189, doi:10.1038/ni.3006 http://www.nature.com/ni/journal/v15/n12/abs/ni.3006.html#supplementary-information (2014).

6 Chen, H. *et al.* Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. *PLoS Comput Biol* **12**, e1005112, doi:10.1371/journal.pcbi.1005112 (2016).

7 van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat Commun* **8**, 1740, doi:10.1038/s41467-017-01689-9 (2017).

8 Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636-645, doi:10.1002/cyto.a.22625 (2015).

9 Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197, doi:10.1016/j.cell.2015.05.047 (2015).

10 Tang, J., Liu, J., Zhang, M. & Mei, Q. in *Proceedings of the 25th International Conference on World Wide Web %@ 978-1-4503-4143-1* 287-297 (International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, 2016).

11 Donnenberg, A. D. & Donnenberg, V. S. Rare-event analysis in flow cytometry. *Clin Lab Med* **27**, 627-652, viii, doi:10.1016/j.cll.2007.05.013 (2007).

12 DiGiuseppe, J. A., Tadmor, M. D. & Pe'er, D. Detection of minimal residual disease in B lymphoblastic leukemia using viSNE. *Cytometry. Part B, Clinical cytometry* **88**, 294-304, doi:10.1002/cyto.b.21252 (2015).

13 Lin, L. *et al.* Identification and visualization of multidimensional antigen-specific T-cell populations in polychromatic cytometry data. *Cytometry A* **87**, 675-682, doi:10.1002/cyto.a.22623 (2015).

14 Hirakawa, M. *et al.* Low-dose IL-2 selectively activates subsets of CD4(+) Tregs and NK cells. *JCI insight* **1**, e89278, doi:10.1172/jci.insight.89278 (2016).

15 Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science (New York, N.Y.)* **332**, 687-696, doi:10.1126/science.1198704 (2011).

16 Belkina, A. *et al.* Multivariate Computational Analysis of Gamma Delta T cell Inhibitory Receptor Signatures Reveals the Divergence of Healthy and ART-Suppressed HIV+ Aging. *bioRxiv* (2018).

17 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).

18 Belkina, A. C. & Snyder-Cappione, J. E. OMIP-037: 16-color panel to measure inhibitory receptor signatures from multiple human immune cell subsets. *Cytometry A* **91**, 175-179, doi:10.1002/cyto.a.22983 (2017).

19 Chen, T. J. & Kotecha, N. Cytobank: providing an analytics platform for community cytometry data analysis and collaboration. *Curr Top Microbiol Immunol* **377**, 127-157, doi:10.1007/82_2014_364 (2014).

20 Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. *arXiv:1712.09005* (2017).

21 Wattenberg, M. V., Fernanda; Johnson, Ian. How to Use t-SNE Effectively. *Distill*, doi:10.23915/distill.00002 (2016).

22 Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A* **89**, 1084-1096, doi:10.1002/cyto.a.23030 (2016).

23 Jacobs, R. A. Increased rates of convergence through learning rate adaptation. *Neural Networks* **1**, 295-307, doi:https://doi.org/10.1016/0893-6080(88)90003-2 (1988).

24 Linderman, G. C. & Steinerberger, S. Clustering with t-SNE, provably. *arXiv:1706.02582* (2017).

25 Mou, D., Espinosa, J., Lo, D. J. & Kirk, A. D. CD28 negative T cells: is their loss our gain? *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* **14**, 2460-2466, doi:10.1111/ajt.12937 (2014).

26 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411, doi:10.1038/nbt.4096
https://www.nature.com/articles/nbt.4096#supplementary-information (2018).

27 Wogsland, C. E. *et al.* Mass Cytometry of Follicular Lymphoma Tumors Reveals Intrinsic Heterogeneity in Proteins Including HLA-DR and a Deficit in Nonmalignant Plasmablast and Germinal Center B-Cell Populations. *Cytometry. Part B, Clinical cytometry* **92**, 79-87, doi:10.1002/cyto.b.21498 (2017).

28 Mazza, E. M. C. *et al.* Background fluorescence and spreading error are major contributors of variability in high-dimensional flow cytometry data visualization by t-distributed stochastic neighboring embedding. *Cytometry Part A* **93**, 785-792, doi:10.1002/cyto.a.23566 (2018).

29 Staser, K. W., Eades, W., Choi, J., Karpova, D. & DiPersio, J. F. OMIP-042: 21-color flow cytometry to comprehensively immunophenotype major lymphocyte and myeloid subsets in human peripheral blood. *Cytometry A* **93**, 186-189, doi:10.1002/cyto.a.23303 (2018).

30 Mair, F. & Prlic, M. OMIP-044: 28-color immunophenotyping of the human dendritic cell compartment. *Cytometry A* **93**, 402-405, doi:10.1002/cyto.a.23331 (2018).

31 Nettey, L., Giles, A. J. & Chattopadhyay, P. K. OMIP-050: A 28-color/30-parameter Fluorescence Flow Cytometry Panel to Enumerate and Characterize Cells Expressing a Wide Array of Immune Checkpoint Molecules. *Cytometry A*, doi:10.1002/cyto.a.23608 (2018).

32 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865, doi:10.1038/nmeth.4380
https://www.nature.com/articles/nmeth.4380#supplementary-information (2017).

33 Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. *Computer Graphics Forum* **35**, 21-30, doi:10.1111/cgf.12878 (2016).

34 Pezzotti, N. *et al.* Approximated and User Steerable tSNE for Progressive Visual Analytics. *IEEE Trans Vis Comput Graph* **23**, 1739-1752, doi:10.1109/TVCG.2016.2570755 (2017).

35 Cho, H., Berger, B. & Peng, J. Generalizable and Scalable Visualization of Single-Cell Data Using Neural Networks. *Cell Syst* **7**, 185-191 e184, doi:10.1016/j.cels.2018.05.017 (2018).

36 Chan, D. M., Rao, R., Huang, F. & Canny, J. F. t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data. *arXiv:1807.11824* (2018).

37 Finn, W. G., Carter, K. M., Raich, R., Stoolman, L. M. & Hero, A. O. Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: Treating flow cytometry data as high-dimensional objects. *Cytometry Part B: Clinical Cytometry* **76B**, 1-7, doi:10.1002/cyto.b.20435 (2008).

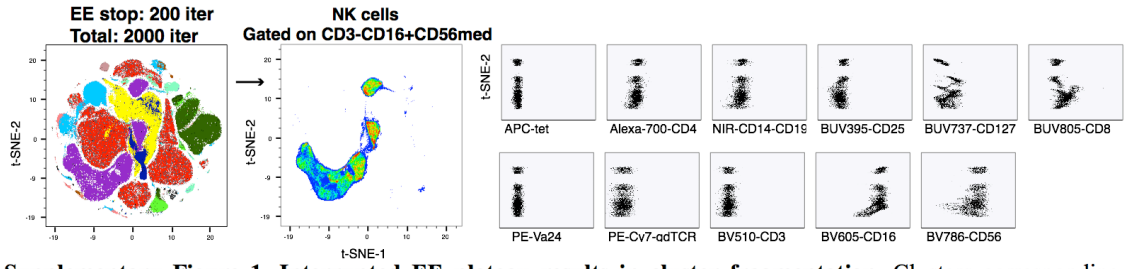
38 Chattopadhyay, P. K. & Roederer, M. Cytometry: today's technology and tomorrow's horizons. *Methods* **57**, 251-258, doi:10.1016/j.ymeth.2012.02.009 (2012).

39 Shekhar, K., Brodin, P., Davis, M. M. & Chakraborty, A. K. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proceedings of the National Academy of Sciences* **111**, 202 (2014).

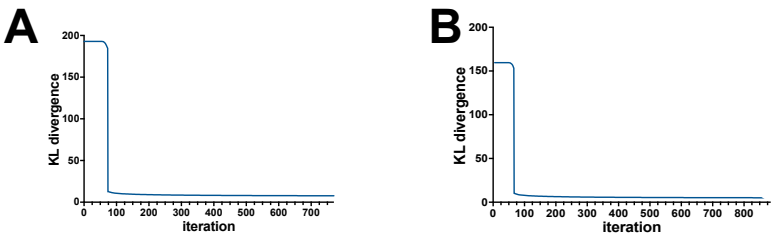
40 Amid, E. & Warmuth, M. K. A more globally accurate dimensionality reduction method using triplets. *arXiv:1803.00854* (2018).

41 Im, D. J., Verma, N. & Branson, K. Stochastic Neighbor Embedding under f-divergences. *arXiv:1811.01247* (2018).

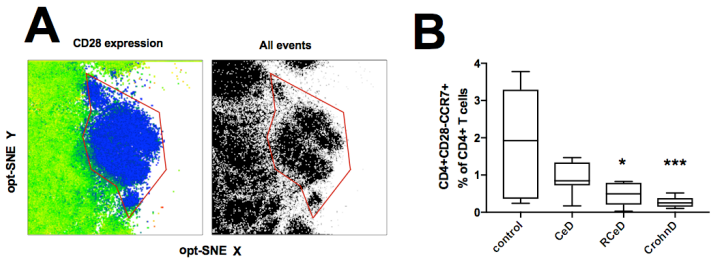
42 Cao, Y. & Wang, L. Automatic Selection of t-SNE Perplexity. *arXiv:1708.03229* (2017).



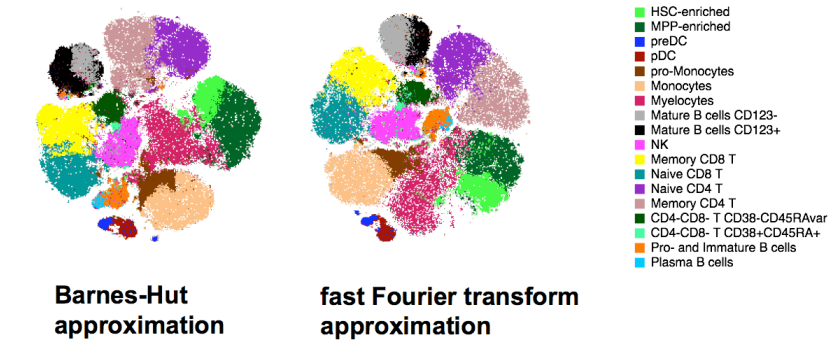
Supplementary Figure 1. Interrupted EE plateau results in cluster fragmentation. Clusters corresponding to CD3-CD16+CD56med NK cells were subsetted from the dataset and assessed as biaxial plots of different parameters plotted versus t-SNE-2 axis. Color overlays correspond to cell type classes labeled as in Fig. 2B (left) or density heatmap (middle).



Supplementary Figure 2. KLD and KLDRC graphs for opt-SNE embedding of flow20M dataset (A) and 10X Genomics 1.3 million datapoints scRNA-seq dataset (B).



Supplementary Figure 3. A. Enlarged fragment of opt-SNE embedding showing CD4+CD28-CCR7+ cell cluster (left: color indicates CD28 expression intensity; right: dot plot). B. Frequencies of CD4+CD28-CCR7+ cells in CD4+ PBMC compared between different cohorts of subjects. CeD, Celiac disease; RCeD, refractory celiac disease; CrohnD, Crohn's disease. Two-tailed T tests were performed for statistical analysis. * p = 0.0258; *** p = 0.0002.



Supplementary Fig. 4. Opt-SNE embeddings of mass41parameter datasets with Barnes-Hut (left) and fast Fourier transform (right) used to compute approximation. Color overlays correspond to cell type classes.

Fluorescent cytometry data																															
	NN		overall	CD8+_T_CD56-CD16-	B_cells	CD14+_CD16- Monocytes	all_monocytes_CD14medCD16-	CD4+_non-Tregs	NK_CD16+CD56-	INKT	all_monocytes_CD14varCD16+	dead_cells	Un-	dn_T_CD4-CD8-	CD4+_Tregs	CD8+_T_CD56+CD16+	dp_T_CD4+CD8+	NK_CD56+highCD16-	gd_T_cells											
250/1000	1	0.655265	0.8351276	0.4740508	0.7192861	0.4697426	0.5034426	0.7038255	0.9665833	0.364868	0.2157275	0.7221916	0.5170367	0.3254001	0.5031605	0.5047785	0.2726858	0.2456989													
250/1000	10	0.7351738	0.8742937	0.5647114	0.8316071	0.4897221	0.5458586	0.8144181	0.9959082	0.3212974	0.1970956	0.7330941	0.4804052	0.3734168	0.5974068	0.5447437	0.2366621	0.2284946													
250/1000	20	0.7432323	0.8777411	0.5741742	0.8442133	0.4909388	0.5451846	0.8296501	0.9963628	0.3043628	0.1917983	0.7296439	0.4590197	0.3842728	0.6133712	0.5377932	0.2334701	0.2139785													
250/1000	30	0.7458495	0.8791322	0.5774068	0.8490957	0.489658	0.5450996	0.8341927	0.9963628	0.2961538	0.1882364	0.7282639	0.4453085	0.3819763	0.6185575	0.5482189	0.2193342	0.2026882													
250/1000	40	0.7470276	0.880601	0.5811097	0.8511621	0.4900423	0.5446321	0.8357519	0.9963628	0.2911596	0.1852224	0.7284019	0.4393738	0.3824635	0.6214749	0.5499566	0.2134063	0.2021505													
250/1000	50	0.7479526	0.881154	0.5809921	0.852917	0.4893699	0.5447961	0.8375701	0.9963628	0.2860505	0.1831217	0.7288159	0.438918	0.3823939	0.6201783	0.5508254	0.2102143	0.2016129													
750/3000	1	0.950853	0.9681263	0.9032855	0.9799056	0.6929752	0.976928	0.9901099	0.966356	0.8481056	0.3066947	0.7667679	0.7823596	0.9191371	0.8658023	0.5499566	0.8435933	0.3537634													
750/3000	10	0.9697462	0.9912734	0.9525979	0.9973461	0.7314613	0.9931634	0.9971799	0.9988634	0.88938	0.2933601	0.7808446	0.807531	0.9522617	0.9267423	0.6133797	0.8969448	0.372043													
750/3000	20	0.9700532	0.9922584	0.954126	0.9974119	0.7365523	0.9932484	0.9972137	0.9997727	0.8880597	0.2875148	0.7801546	0.8080426	0.9530967	0.9273906	0.6324935	0.8937528	0.3634409													
750/3000	30	0.9699792	0.9923707	0.9545668	0.9973461	0.7358799	0.9932727	0.9972137	0.9997727	0.8871412	0.2870582	0.7742203	0.8093728	0.9540014	0.9271475	0.6377063	0.8910169	0.3537634													
750/3000	40	0.9699482	0.9924917	0.9555954	0.9973359	0.7358158	0.9931938	0.9971855	0.9997727	0.8869116	0.2830395	0.7731162	0.8074286	0.954071	0.9284441	0.6455256	0.8878249	0.3537634													
750/3000	50	0.9699402	0.9924917	0.9567121	0.9973258	0.736136	0.9931573	0.9971236	0.999318	0.8859357	0.2826742	0.7714601	0.8068147	0.9542102	0.9287682	0.6533449	0.8878249	0.3489247													

Mass cytometry data																															
	NN		overall	MPP-enriched	Naive_CD4_T	Monocytes	memory_CD8_T	Pro_Imm_B_cells	NK	Mature_B_cells_CD123+	memory_CD4_T	Myelocytes	dnT_CD38-	Mature_B_cells_CD123-	Naive_CD8_T	pro-Mono	negative_crap	dnT_CD38+	preDC	HSC-enriched	Plasma_B_cells	pDC									
250/1000	1	0.7621149	0.782215	0.863678	0.7726368	0.7598064	0.7360379	0.6640332	0.755357	0.8931913	0.7213852	0.7748872	0.7623388	0.7881323	0.3764868	0.6020365	0.1796563	0.6167708	0.6961161	0.8391708	0.7356503									
250/1000	10	0.8239847	0.879642	0.8724507	0.8274643	0.8333474	0.8180381	0.6873326	0.830432	0.925903	0.8287211	0.8064647	0.8474304	0.8509405	0.489584	0.6214517	0.1509666	0.6902931	0.761034	0.874732	0.8836264									
250/1000	20	0.8295168	0.8868671	0.8717855	0.8355174	0.8396505	0.8221094	0.6868119	0.8367088	0.9296435	0.8371309	0.8106706	0.8555267	0.8577964	0.5196287	0.6213535	0.1422215	0.694618	0.7664077	0.8749107	0.8929993									
250/1000	30	0.8309218	0.8896337	0.872024	0.8383637	0.8411419	0.8226842	0.6852794	0.8390897	0.9301004	0.8387194	0.8123325	0.8557411	0.8599232	0.5350292	0.6214845	0.1337834	0.6972609	0.7667898	0.8733024	0.8959025									
250/1000	40	0.8318248	0.8912831	0.8718232	0.83951	0.8422393	0.8228279	0.6846843	0.8408522	0.9306425	0.8404014	0.8132822	0.8559556	0.8608976	0.543142	0.6219101	0.1319423	0.6972609	0.7665606	0.8727663	0.8974784									
250/1000	50	0.8321128	0.8919109	0.8719612	0.8396946	0.8426332	0.8230194	0.6838957	0.8427074	0.9306037	0.8407464	0.8142998	0.8558216	0.8615511	0.5474734	0.6222048	0.1276465	0.7006247	0.7662549	0.8727663	0.8967319									
750/3000	1	0.8607732	0.9044564	0.9035756	0.9380519	0.8925798	0.8012741	0.818455	0.8286076	0.9342823	0.7986688	0.7832649	0.81282	0.9062392	0.5717429	0.7443932	0.4679349	0.6309467	0.8394499	0.8561472	0.7719808									
750/3000	10	0.9008077	0.9216412	0.9166405	0.9588308	0.9250098	0.8889261	0.848658	0.8854395	0.9591413	0.8785041	0.8336329	0.8887429	0.9352313	0.6842901	0.773205	0.4868058	0.7090341	0.8809882	0.887777	0.8871931									
750/3000	20	0.9037007	0.9224287	0.9167033	0.9597245	0.9291744	0.8916563	0.8486133	0.8874184	0.9626262	0.8850666	0.8349218	0.8975095	0.9371443	0.6972155	0.774842	0.4815894	0.7172033	0.8830001	0.8874196	0.8958195									
750/3000	30	0.9043998	0.9230458	0.9169041	0.9597828	0.9299623	0.8928537	0.8483455	0.8877277	0.9634704	0.8869713	0.8342774	0.8990644	0.9381543	0.6971468	0.7746783	0.475606	0.7234503	0.8841717	0.8858113	0.8983079									
750/3000	40	0.9045688	0.9227904	0.9168539	0.9596954	0.9306517	0.8933806	0.8475867	0.8878204	0.9641286	0.887381	0.8344131	0.8998686	0.9385939	0.6999656	0.7739253	0.47085	0.7239308	0.8837387	0.884203	0.8998009									
750/3000	50	0.9045178	0.922833	0.9168539	0.9593359	0.9306939	0.8929016	0.846947	0.8875112	0.9641054	0.887992	0.8335651	0.9008338	0.9386889	0.7011344	0.7725502	0.4690089	0.7234503	0.8835095	0.8825947	0.899635									

Supplementary Table 1. Nearest Neighbor (NN) accuracy evaluation of standard (EE250/1000 iterations) and extended (EE750/3000 iterations) t-SNE embeddings shown on Fig. 1.