AUTOMATED OPTIMIZED 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]BD Life Sciences – FlowJo, 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 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 superior data resolution in t-SNE space and thereby more accurate 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, 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 PCA, are mostly unsuitable for cytometry data visualization as such techniques cannot faithfully present the non-linear relationships. 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. 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]. Visualizations of cytometry data produced with other non-linear embedding algorithms, such as LargeVis [8], UMAP [9], and EmbedSOM [10], can be interpreted and interrogated in a similar manner. Clustering algorithms that directly interrogate high-dimensional data, such as FlowSOM [11] and PhenoGraph [12], are often used in conjunction with 2D maps to present annotated cell 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 [7,8]. 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 [13]. 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. Consequently, researchers often resort to either subsampling their data to the very limit of detection of rare populations [14] or to exporting specific populations from their dataset, thus compromising the 'unbiased' data analysis approach[15,16].

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 Barnes-Hut (BH) implementation of t-SNE (or viSNE, as it was renamed [3]) on mass cytometry data; since then, BH-tSNE has been integrated into the majority of commercial and open-source platforms for cytometry analysis. 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 and hence are still widely used. Recently published modifications of t-SNE, such as HSNE[17] and FIt-SNE[18], do not fully address these limitations. Therefore, development of rigorous methodology to release the full potential of t-SNE for single cell data comprehension, and provide clues for optimization of second generation t-SNE-like algorithms 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 commonly prescribed best practices, and then iteratively modified parameters of embedding to identify conditions that ensure improved visualization. As a result, we propose a new method to automatically find dataset-tailored t-SNE parameters via fine-tuning of the early exaggeration phase of t-SNE embedding in real time. We call our approach opt-SNE, for *opt*imized t-SNE. opt-SNE adjustments can greatly 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 most favorable selection of t-SNE

parameters, potentially saving many hours of computation time per research project. Finally, we implemented opt-SNE as an open source fast multicore t-SNE C++ package which allows much faster computation of t-SNE embeddings than original single-core Barnes-Hut t-SNE implementation.

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 lineage parameters used for embedding) of 1 million datapoints concatenated from 5 samples of human bone marrow cells	[19]
Flow18parameter	Flow cytometry	18 parameter dataset (11 lineage parameters used for embedding) of 1 million datapoints concatenated from 2 samples of human PBMC	[20]
Flow20M	Flow cytometry	18 parameter dataset (15 lineage parameters used for embedding) of 20 million datapoints concatenated from 27 samples of human PBMC	[20]
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 [21]
van Unen et al	Mass cytometry	32 parameter dataset (26 lineage 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 [22] 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 [19] or flow cytometry [20] 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 (with cofactor 5 for mass cytometry data and with cofactor 150 for flow14parameter data) or biexponential (flow20M) transformation. Light scatter parameters were log-transformed.

The standard t-SNE configuration for cytometry applications. As described in detail previously[2], 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[1] 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 EE phase (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 the BH-tSNE 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[2] 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 in Omiq and integrated into FlowJo and SeqGeq programs.

Data analysis. A desktop C++ Barnes-Hut implementation of t-SNE for Mac OS was used for t-SNE analyses [2] with original implementation adapted to allow user input for the early exaggeration stop iteration, perplexity, total number of iterations, early exaggeration factor value, and learning rate value. A multicore modification of Barnes-Hut t-SNE[23] was used as a source for building the opt-SNE package with user-accessible parameters similar to listed above. The Kullback-Liebler Divergence (KLD) value and t-SNE coordinates were reported during each iteration or as frequently as requested. All datasets were embedded in 2D space. Visually comparable t-SNE maps were generated with the same random seed values used when permutations for specific parameters were tested and compared; all experiments were repeated with several values of random seed. Benchmarking data were generated on an iMac personal station with Intel Core i7 quad-core i7 processor and 16GB of RAM. For cross-validation, we utilized cloud-based Cytobank[24], cloud-based Omiq, FlowJo V10.3-10.5 and FlowJo V9.9.6. R flowcore package, Cytobank and FlowJo

were used to generate FCS files and graphical outputs from tabular data. Omiq platform was used for FlowSOM[11], FIt-SNE (Fast Fourier Transform-accelerated Interpolation-based t-SNE[18]) and UMAP (Uniform Manifold Approximation and Projection) [9]analyses. Also, the FIt-SNE plugin for FlowJo and FIt-SNE prototype in FlowJo were used to ensure FIt-SNE compatibility with opt-SNE. Logs of t-SNE runs were batch-processed with VBA scripts and analyzed with GraphPad Prism. Expert-guided (manual) analysis of cell populations was performed in FlowJo 10.3-10.5 as described previously for specific datasets [19,22] or as explained below, and used for map annotations as cluster classes.

We used SeqGeq 1.3 for scRNAseq analysis and leveraged PCA projections provided in the 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 [21].

The quality of embedding was assessed by a 1-nearest neighbor (1NN) classifier based on previous reports on t-SNE accuracy evaluation [1,2,8] . The classifier was written in R (version 3.5.0). For each invocation, the classifier was provided a complete embedding and an accompanying array of assigned class values for each observation in the embedding. Ten thousand cells were sampled from the embedding as a training set using uniform pseudorandom sampling with the "sample" function from base R. Next, a set of 50,000 cells disjoint to the training set was sampled from the embedding as the test set. Each cell in the test set was then mapped to its single nearest neighbor within the test set and assigned the class of that neighbor using the "knn" function of the "FNN" R package (version 1.1.3, available on CRAN). The number of correct class assignments in the test set was calculated on a total basis and also on a per-class basis using the supplied class values. This process was repeated five times for the provided embedding using different seeds for the sampling step. Replacement was allowed for pseudorandom sampling between the five internal repeat runs. The results were then presented as the average of these five repetitions. The seeds used for the five repetitions were kept consistent such that classification results between different embeddings of the same dataset tested the same cells.

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 [25]. 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 typical ($< 5x10^5$) datasets previously reported in benchmark studies of cytometry algorithmic tools [26]. 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.

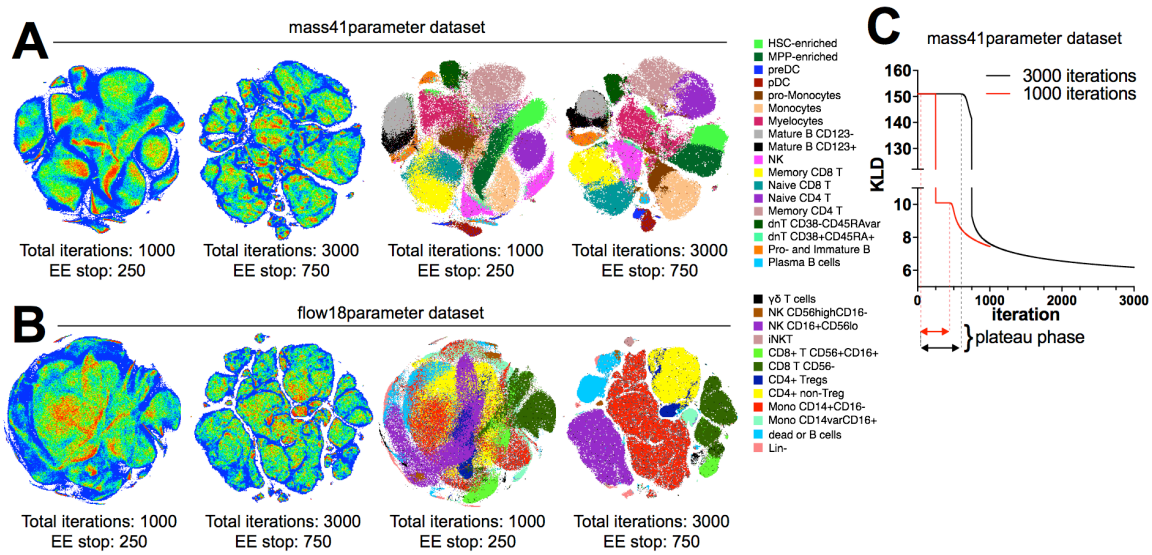


Figure 1. Performance of Barnes-Hut t-SNE implementation for cytometry data visualization. Standard (1000 iterations) and extended (3000 iterations) embeddings of mass cytometry (A) or flow cytometry (B) data are presented as heatmap density plots (left) or color-coded population overlays based on ground-truth 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.

Empirically, analysts have observed that increasing the number of iterations of t-SNE computation results in better quality maps[9,27]. 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 standard 1000 iterations per run and with an extended 3000 iteration computation (Fig. 1). As expected, 1000-iteration runs produced maps with poor visualization compared to 3000-iteration runs. 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 (Fig.1A, B) The 1NN accuracy of embedding was also higher in 3000-iteration embeddings for both datasets compared to the "standard" 1000-iteration t-SNE (Suppl. Fig. 1). 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 "standard" 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 in BH-tSNE [1]. We applied EE over 250 iterations in the standard t-SNE configuration and over 750 iterations in the extended run with 3000 iterations since most platforms have EE scaled to 25% of total iteration number. Notably, the KLD did not immediately start to minimize at the beginning of the EE; instead, the graph of KLD over time is a plateau that is

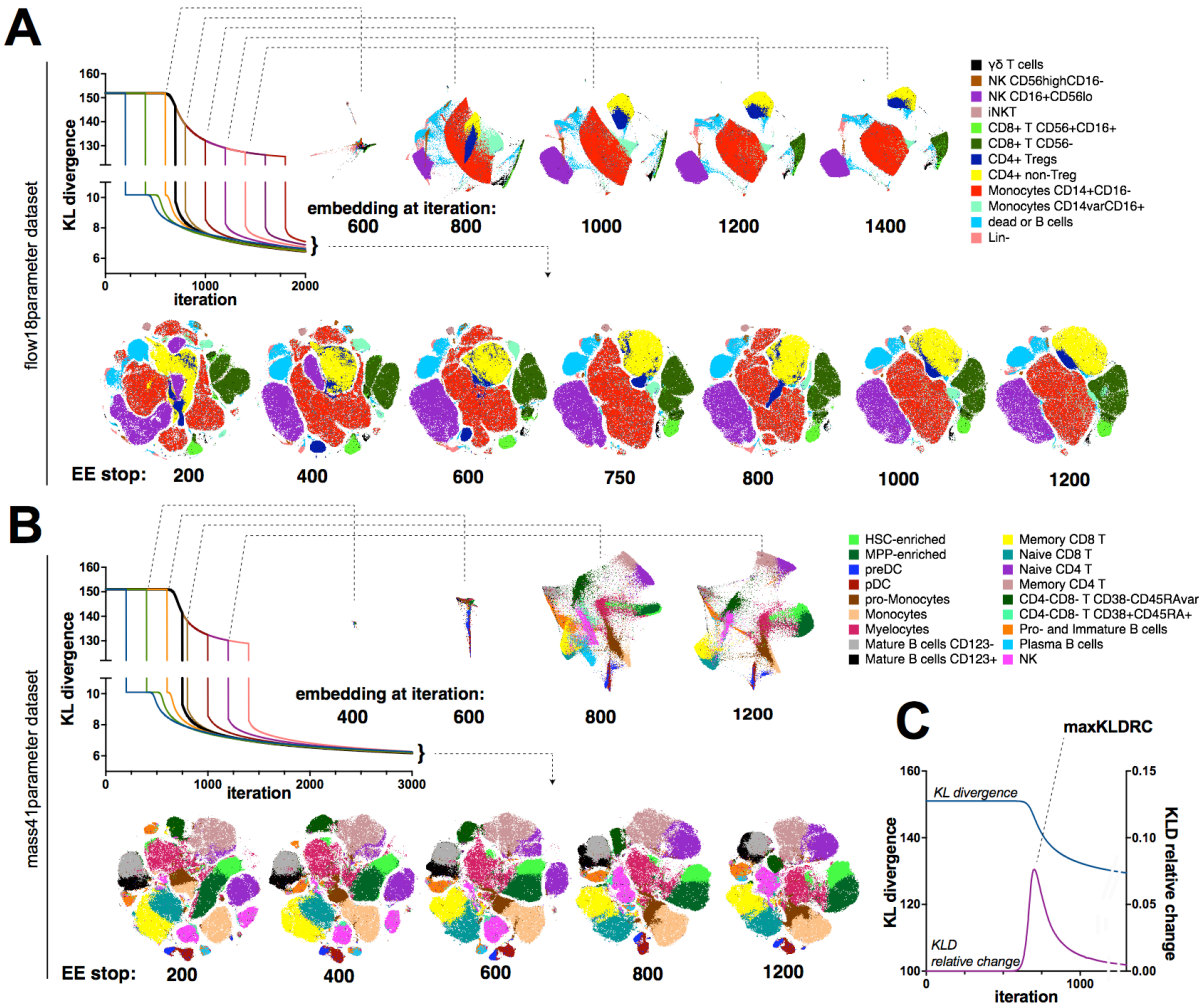


Figure 2. Effect of EE plateau phase on t-SNE visualization. EE was stopped after varying number of iterations and embedding visualization was examined at several intermediate timepoints and in the end of embedding for flow cytometry (total of 2000 iterations, A) and mass cytometry (total of 3000 iterations, B). 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. C. KLD and KLD relative change plotted against iteration time for the mass41parameter embedding. All embeddings were generated with standard BH-tSNE implementation and representative examples of multiple runs with varying seed values are shown.

followed by a curve that captures the incremental decrease of KLD, indicating the gradient descent. In the standard t-SNE run, the plateau was interrupted when the EE was stopped, then continued with a non-exaggerated value of KLD.

Van der Maaten and Hinton called EE a "trick" that improved resolution of the global structure of the data visualization that would not otherwise converge to separated clusters[1]. As the suboptimal quality of the 1000-iteration t-SNE maps shown in Fig. 1A and 1B demonstrated poor global structure resolution, we hypothesized that increasing the total iteration number in conventional cytometry analysis platforms would inadvertently increase the number of EE iterations resulting in improvement in map visualization. To test this hypothesis, we compared runs that differed in timing of the EE stop by plotting the embedding at the iteration when the EE stops and also at later

224 iterations, thus assessing the effects of EE and our perturbations on both flow cytometry (Fig. 2A)
225 and mass cytometry (Fig. 2B) data visualization.
226 We found notable differences in map quality between the runs with shorter and longer EE.
227 Although the map after EE200/total3000 iterations may appear visually more appealing than
228 previously shown EE250/total1000 map, ground-truth labeling indicated that phenotype-defined
229 clusters are fragmented in both flow and mass cytometry maps. When cluster fragments were
230 plotted on a biaxial plot against parameters that were used in the t-SNE dimension reduction, we
231 were not able to identify protein marker expression patterns that immediately contributed to their
232 fragmentation (Suppl. Fig. 2). Conversely, tight clusters that form at the end of the plateau remain
233 mostly unchanged as long as the EE is being applied to the computation (Fig. 2 A, B). The KLD
234 minimization in that case could be explained by the gradual shrinking of the 2D space (data not
235 shown). Once EE is removed, the attractive forces within each cluster are weakened and the local
236 structure of the data is fully resolved within each cluster.
237 Overall, our observations suggest that the EE phase of the gradient descent is essential for data
238 clustering while the non-exaggerated descent results in the resolution of local structures. When the
239 EE is too short, the cell clusters continue to be resolved simultaneously with the local structure of
240 each cluster being unfolded, leading to fragmented, overlapped or deformed "islands" in the
241 resulting map (Fig. 2 and Suppl.Fig. 2). To locate the timepoint when to end EE once the plateau
242 phase is completed, we tracked the relative rate of KLD change ($KLDRC_N = 100\% *((KLD_{N-1} -$
243 $KLD_N)/ KLD_{N-1})$ where N is the iteration number) and identified the local maximum of KLDRC
244 (maxKLDRC) (Fig. 2C). Since KLD is computed at each iteration, the maxKLDRC 'sensor' can
245 be added to the algorithm programmatically and would stop EE at the next iteration past
246 maxKLDRC. For both datasets the maxKLDRC was detected at about 700-720 iterations (Fig. 2C),
247 and, when the EE phase was followed by 300-1300 iterations without EE, the resulting visualization
248 was similar to the EE750/3000 map at Fig. 1A and no cluster fragmentations were observed (Suppl.
249 Fig. 3), demonstrating more relevant visualization of cytometry data compared to standard t-SNE
250 settings.
251 **Moderate adjustments of EE factor and perplexity do not impact visualization.** Once
252 we found EE to be crucial for map optimization, we next examined if the value of the EE factor α
253 can also be tuned to improve the results of t-SNE. We made α user-accessible in our C++ t-SNE
254 code since it is hard-coded in the original Barnes-Hut C++ t-SNE implementation and all
255 aforementioned results were obtained with default value of $\alpha = 12$. First, we tested how the
256 optimization would proceed without EE ($\alpha = 1$). We expected the run to fail or produce extremely
257 crowded results as explained in the original t-SNE report [1]; however, we did not see much overlap
258 in cluster positioning, likely due to the substantial number of map iterations run (Fig. 3A).
259 Nevertheless, the resulting map showed considerable fragmentation, reflecting an extreme case of
260 an interrupted plateau phase. Even when run for as many as 3000 iterations, the fragmentation could
261 not be remedied, again demonstrating the necessity of EE. As expected, higher values of α lead to

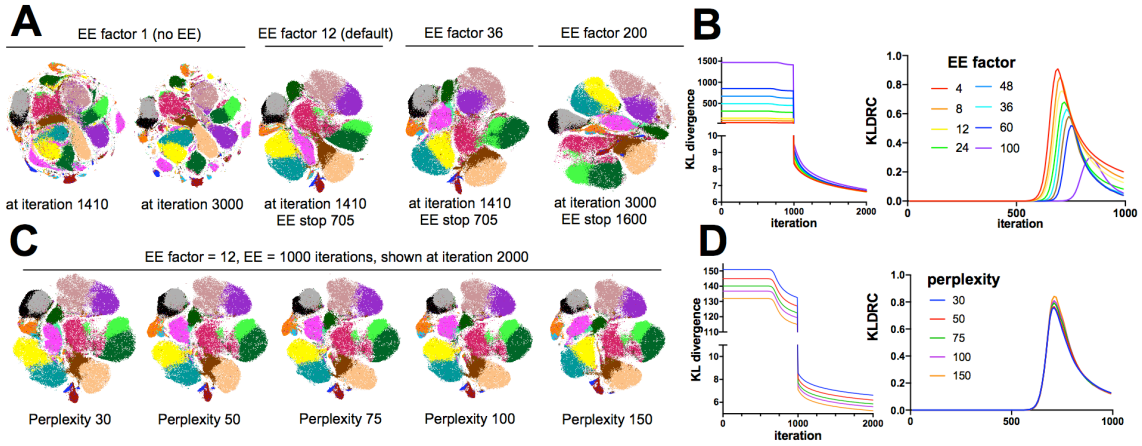


Figure 3. Effects of perplexity and EE factor adjustments on t-SNE visualization of cytometry data. A, B. KLD, KLDRC and t-SNE biaxial plots generated with varying EE factor values. C, D. KLD, KLDRC and t-SNE biaxial plots generated with varying perplexity. Graphs showing KLD and KLDRC change over iteration time are color-labeled to distinguish curves corresponding to experiment perturbations. 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.

much higher KLD during EE, however, the KLD values were similar at 2000 iterations when α varied between 4 and 60 (Fig. 3B) and EE was stopped at the max KLDRC iteration. Larger α prolonged the plateau phase and became detrimental for KLD values when over 100. Visually α = 200 resulted 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 hard-coded in BH-tSNE [2], or reverted to α = 4, as originally proposed by van der Maaten and Hinton[1] since per our results any value between 4 and 20 leads to similar 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 [25]. 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 perplexity results in visibly improved data visualization (Fig. 3C) or faster resolution of clusters (Fig. 3D). However, while changing α does not affect t-SNE computation time, perplexity is linearly related to the time required to create the embedding in both implementations of BH-tSNE that we tested (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.

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 [28]. 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. 4A) and *h* = 200 finished the plateau in ~700 iterations as previously shown. With

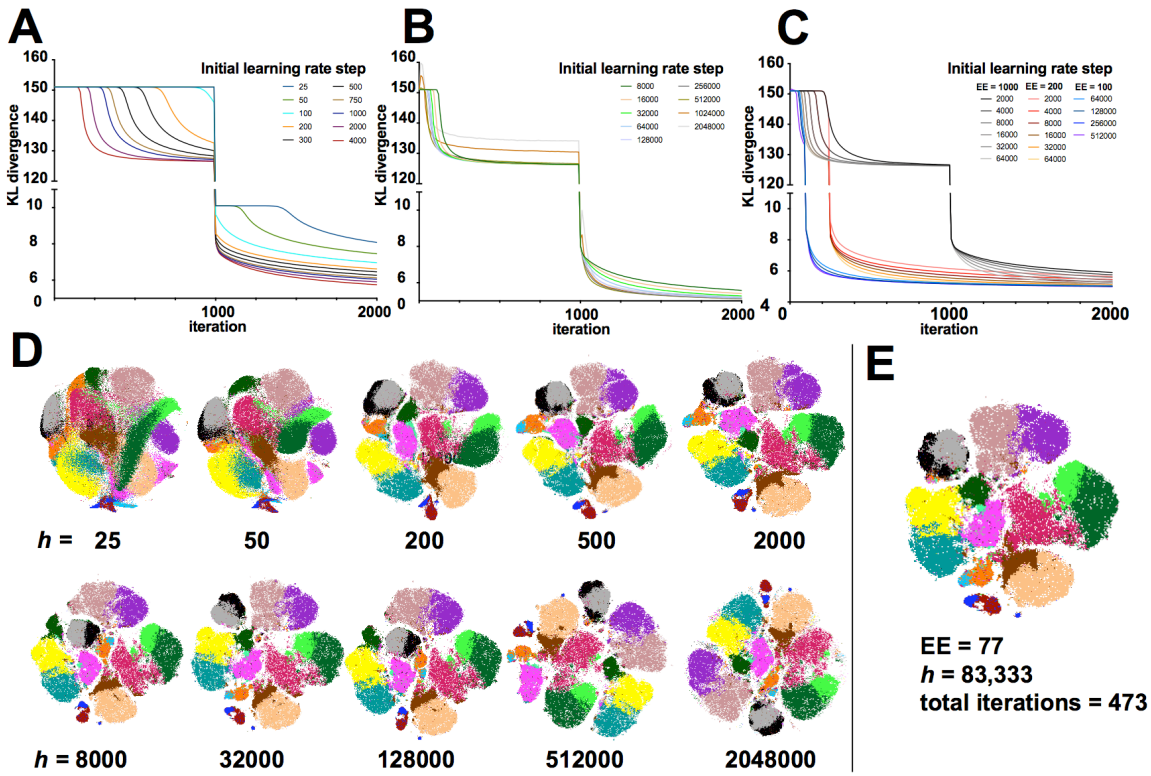


Figure 4. Learning step size optimization for t-SNE visualization of large datasets. A, B, C KLD change over iterations for embeddings with varying values of initial learning rate step size, color coded as indicated. A, EE = 1000 iterations, learning rate step = 25-4,000; B, EE = 1000 iterations, learning rate step = 8,000-2,048,000; C, EE = 100-1000 iterations, learning rate step = 2,000-512,000. D, representative t-SNE plots of embeddings graphed on A. E, t-SNE plot of an optimized 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.

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 (in most platforms, h is restricted to ranges below 3,000) (Fig.4B). At $h \sim 256{,}000$ we observed irregular peaks in the KLD graph indicating that the prescribed step size rendered gradient descent ineffective. However, using lower values of h we were able to create 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. 4C). Visual inspection of the embedded maps over the range of h values agrees with the KLD values (Fig. 4D).

In a recent publication, Linderman and Steinerberger [29] 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, we found the most efficient 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 t-SNE visualization (Fig. 4E).

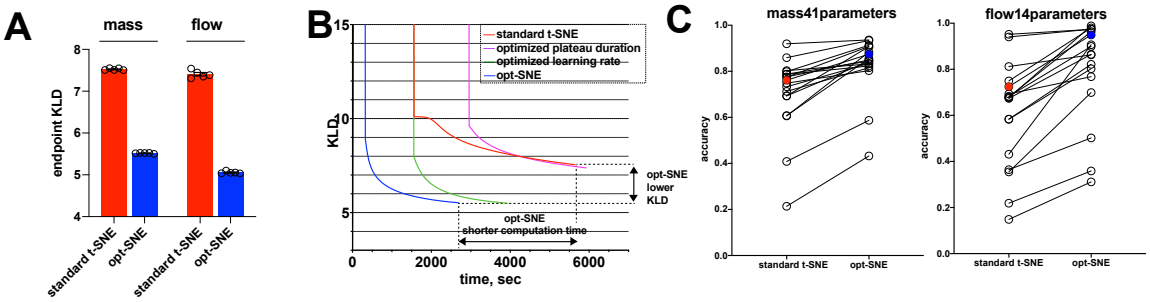


Figure 5. Evaluation of opt-SNE embeddings. A, endpoint KLD values for standard t-SNE (initial learning rate step = 200, EE stop = 250 iterations) and opt-SNE (initial learning rate = n / α, EE stop at maxKLDRC iteration). N = 5 seeds used for random initialization; error bars denote SEM. B. Post-EE graph of KLD minimization over physical time for standard t-SNE, adjusted parameter (as indicated) t-SNE and opt-SNE (representative examples of mass cytometry data embeddings are shown). C. 1NN accuracy scores for standard t-SNE and opt-SNE embeddings of of mass cytometry (left) and flow cytometry (right) data per assigned class values (cell subsets, open circles; overal scores, filled circles). Representative examples of multiple runs initiated with varying seed values are shown.

opt-SNE provides faster, higher quality embedding of cytometry datasets. The standard C++ BH-tSNE implementation that we used to optimize t-SNE parameters only utilizes a single processor core and requires considerable computation time. We thus adopted a recently developedmulticore modification of Barnes-Hut t-SNE[23] to implement our proposed optimization techniques into a single workflow labeled 'opt-SNE', that includes KLD plateau monitoring and dataset size-specific learning rate step scaling (Table 2). To compare opt-SNE to standard t-SNE and individual t-SNE parameter tuning results, we generated embeddings that were automatically finalized when $(KLD_{N-1} - KLD_N) < KLD_N/5{,}000$ (i.e. when each new iteration only improved KLD by less than 0.05%). When we moved from original C++ BH-tSNE to multi-core BH-tSNE[23], we observed a 2-3X boost in computation speed with no penalty in embedding quality, endpoint KLD values, or total number of iterations to reach endpoint KLD (Suppl. Fig. 4A and data not shown). When compared with standard t-SNE run in multicore enviroment, opt-SNE visualizations of both flow and mass cytometry data exhibited an additional 2X or greater speedup due to the smaller number of iterations needed to complete the data embedding (Fig. 5B and Suppl. Fig. 4B). This improvement in computation time scaled with dataset size, and almost no variation was observed when runs initiated from various random seeds were compared (Suppl. Fig. 4B-E). As discussed above, opt-SNE maps demonstrated superior quality compared to standard t-SNE embeddings when visually inspected. We also assessed the quality and accuracy of the embeddings quantitatively by comparing their endpoint KLD values and 1NN accuracy scores. opt-SNE embeddings achieved values of endpoint KLD that were significantly lower than standard t-SNE (Fig. 5A, B) and such values were not reached by the standard t-SNE calculation even after 20,000 iterations (data not shown), indicating better retention of multidimensional data structure. When compared on 1NN classifier accuracy, opt-SNE embeddings achieved better scores than embeddings produced with standard t-SNE methodology (Fig. 5C). Also, by the same metric, the $KLD_N/5{,}000$ endpoint criterion has proven to mark sufficiently long opt-SNE runs since no improvement of accuracy was observed when the embedding calculation continued beyond that point (Suppl. Fig. 4E). For further algorithm tuning at the discretion of the user, we incorporated automated opt-SNE termination as a tunable parameter that sets a fraction of KLD minimization per iteration.

Table 2. Summary of t-SNE optimizations proposed in opt-SNE workflow

Parameter	opt-SNE setup	Suggested use with cytometry data
Gradient descent learning rate	Adaptive learning rate with initial value $h = n / \alpha$, where n is the number of datapoints and α is the early exaggeration factor	Automated per dataset
Early exaggeration factor	Standard t-SNE setup considerations apply	4-12
Perplexity	Standard t-SNE setup considerations apply	30-50
Early exaggeration termination	KLD value (cost function) is monitored in real time, and early exaggeration is removed at maxKLDRC	Automated per dataset
t-SNE termination	KLD value (cost function) is monitored in real time and the embedding is finalized when $(KLD_{N-1} - KLD_N) < KLD_N/X$	X = 5000

340

341 **opt-SNE allows visualization of mega-scale, real life single cell data.** To test the performance
342 of opt-SNE on even larger datasets, we used a 20.1 x 10^6 event fluorescent cytometry dataset
343 concatenated from two independent batches of PBMC samples (N=27) stained with a variation of
344 the OMIP-037 fluorescent cytometry panel [22] that allows detailed assessment multiple immune
345 subsets including naïve and memory CD4+ and CD8+ T cells, NK cells and γδ T cells (Fig. 6A-B).
346 The embedding completed in 770 iterations with only 73 iterations required to pass the EE step
347 (Suppl. Fig. 5A) and resulted in clear separation of cell clusters as evaluated by cell type annotation
348 (Fig. 6C). The majority of clusters appear to be populated by cells from all subjects with the
349 exceptions of several populations that contained sample-unique debris features (Fig. 6B, dashed
350 arrows), confirming an absence of batch effects. A detailed breakdown of identified populations is
351 presented in Fig. 6C that shows subsets of CD4+ and CD8+ T cells, NK cells, γδ T cells, B cells,
352 and monocytes. Importantly, B cell and monocyte lineage markers were detected together with a
353 viability dye in a dump channel in this panel and therefore could not be gated accurately via
354 traditional biaxial plot analysis. However, opt-SNE helped to visualize these populations as clear
355 cell groupings in the embedding analogously to their presence in the original high-dimensional
356 space defined by the combination of several surface labels and light scatter characteristics, and the
357 resulting opt-SNE islands were minimally mixed with dead cells. Use of the standard t-SNE
358 algorithm on this dataset completely failed to reveal its structure, even with several thousands of
359 iterations (Fig. 6D).
360 To test the suitability of opt-SNE for applications beyond flow and mass cytometry, we analyzed
361 a 1.3x10^6 cell single-cell RNA-seq dataset of mouse embryonic brain cells published by 10X
362 Genomics. We used pre-calculated PCA projections included in the dataset to generate opt-SNE
363 maps that we compared with 10X standard t-SNE embedding (Fig. 6E). 10X used EE = 1000/total

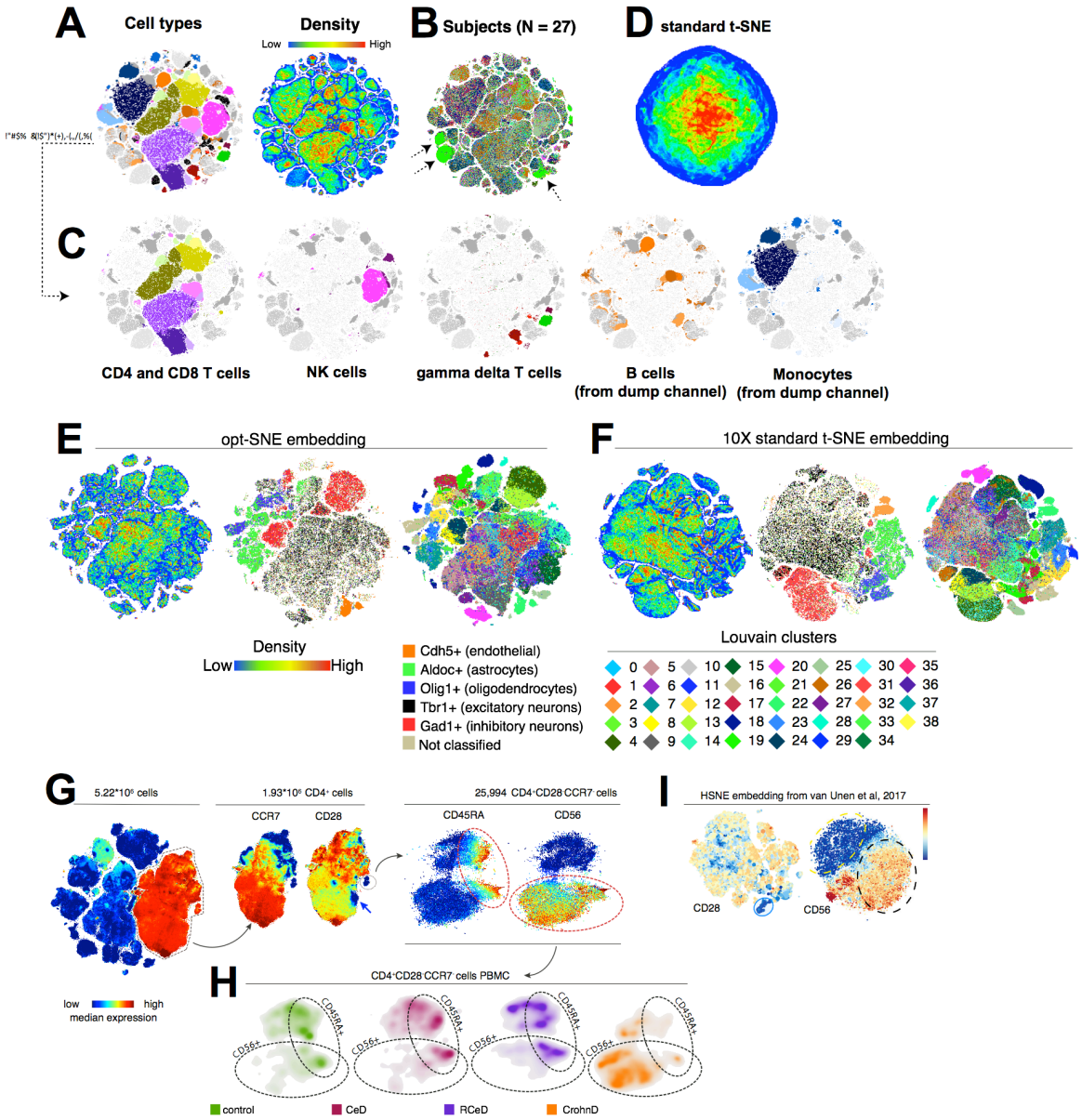


Figure 6. opt-SNE allows high-quality visualization of large cytometry and transcriptomics datasets. **A-D**: 20 million datapoints from fluorescent cytometry dataset concatenated from 27 subjects vizualized in 2D space. **A** and **C**, cell type classes and density overlaid on 2D opt-SNE embedding. **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.

4000 iterations of standard t-SNE while we used opt-SNE settings with h = 97,959, EE = 66/total 885 iterations (Suppl. Fig. 5B). Non-immune single cell transcriptomics data are more 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 clustering algorithm using the Scanpy Python package [21] 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 [7,17]. To directly compare HSNE and opt-SNE, we applied opt-SNE to the 5.2 million point dataset that was reported in HSNE analysis of mass cytometry data[7], and compared opt-SNE visualization to the full resolution level of HSNE embedding (Fig. 6G, Suppl. Fig. 6). In the CD4+ subset, opt-SNE visualization revealed two groups of CD4+CD28- cells likely representing terminally differentiated memory CD4+ T cells [30] 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). However, 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 maps revealed 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 [11,31]. 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 [32]. 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 [33]. 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[34-36], yet the proper data assessment tools are lacking for general use. DNA-barcoded antibodies have been recently utilized to allow

410 simultaneous protein-epitope and transcriptome measurements in single cells [37] thus expanding the
411 repertoire of traditional cytometry methods that could employ t-SNE as a staple method of data
412 visualization and presentation.

413 One approach to large dataset t-SNE embedding is to model visualization with a subset of
414 datapoints curated [7,17] or randomly selected (as implemented in cytometry data analysis platforms
415 such as FCS Express and Gemstone) from the nearest neighbor graph. Such techniques may fail to
416 project extremely rare datapoints, as demonstrated by CD4+CD28-CCR7+ cells that were
417 identified via use of opt-SNE, but not HSNE in the van Unen et al dataset[7] (Fig. 6). This T cell
418 subset was significantly less abundant in subjects with severe inflammatory conditions including
419 Crohn's disease as compared to controls (Suppl. Fig. 6). However, the cells expressing
420 CD4+CD28-CCR7- phenotype, while also low in frequency in most PBMC samples (<1% of all
421 CD4+ cells), was likely successfully clustered by HSNE because two of the 11 Crohn disease
422 subjects in the dataset showed an unusually high frequency of this population (40.7% and 31.0%
423 of all CD4+ cells), allowing it to be well represented in the kNN graph of the full dataset.

424 Several attempts to successfully apply t-SNE-like methods to massive datasets have been
425 recently reported including aforenoted HSNE [7,17,38], LargeVis [8] and net-SNE [39]. However, when
426 applied to large datasets, these methods often require/benefit from considerable computational
427 resources; for instance, the LargeVis study was performed on a 512GB RAM, 32 core station.
428 However, it is ideal for such data analysis to be performed successfully with the computational
429 resources found in most laboratories. We benchmarked our opt-SNE modification of multicore t-
430 SNE C++ implementation on a quad-core, 16GB RAM workstation although even further
431 improvement in computation speed would occur if more cores were available. FIt-SNE, a recently
432 published alternative to Barnes-Hut t-SNE that uses fast Fourier transform for much faster
433 computation of repulsive forces approximation[18], renders opt-SNE even more feasible on personal
434 computers. Notably, we have not observed differences in quality between embeddings generated
435 with fast Fourier transform versus Barnes-Hut approximations when opt-SNE parameter
436 adjustments were used to compute the visualization (Suppl.Fig.7 and data not shown). Therefore,
437 we expect opt-SNE to be applicable for existing or future adaptations of t-SNE even if alternative
438 methods of computation are utilized [40,41] provided that they retain the core principles of t-SNE
439 embedding. A promising approach that may be integrated with opt-SNE is the smart EE adjustment
440 implemented in A-tSNE (approximated t-SNE)[38] algorithm where EE is removed gradually and on
441 a per-point basis. Cytosplore, a novel software platform that includes HSNE and A-tSNE, allows
442 the analyst to interactively initiate the local refinement of the map, resulting a significant
443 improvement in computation time[7].

444 Often compared to t-SNE, Uniform Manifold Approximation and Projection (UMAP) is a
445 recently developed novel technique for dimension reduction that can be successfully applied to
446 cytometry datasets [9]. As has been the case for other groups [9,27], our comparison of t-SNE and opt-
447 SNE to UMAP generated mixed results. Specifically, we found that UMAP parameter selection
448 had dramatic effects on UMAP computation performance and embedding quality assessed by both
449 human perception and 1NN classifier accuracy (Supp. Fig. 8A-E). When benchmarked for resource
450 use in the Omiq cloud, UMAP consumed up to 128GB of RAM when embedding 5.2 million cell
451 van Unen dataset, while opt-SNE and FIt-SNE peaked at 16GB RAM consumption; the UMAP
452 embedding took a comparable amount of time to complete as opt-SNE. With careful parameter

453 selection, UMAP accuracy was approaching, but not surpassing, opt-SNE accuracy (Supp. Fig.
454 8A,B).
455 Similar to other types of biological data, the structure of cytometry data is difficult to project
456 due to its mixed nature, often comprised of cluster-like, manifold-like and/or hierarchical
457 components [33,42]. In this paper we propose multiple techniques that are essential for superior quality
458 t-SNE data projection and are all germane to the fine-tuning of the early exaggeration phase of t-
459 SNE embedding. EE facilitates cluster formation on a 2D plane[1,43] and serves as a necessary
460 compromise that allows clusters to escape the crowding effect. We designed an efficient measure
461 to ensure that the cluster-like global structure of the data is fully revealed during the EE phase by
462 monitoring the KLD output of the embedding in real time. As indicated by the lower KLD values
463 of opt-SNE embedding where EE was limited to fewer iterations (Fig. 4E), prolonged amplification
464 of the attractive forces that drive tight cluster formation in EE may be detrimental for the manifold-
465 like local data structure represented by signal distributions that are continuous with background
466 signal. These 'continuum expression' molecules include proteins used to define classic immune
467 cell subsets as well as markers indicating disease phenotypes, cell activation, and/or exhaustion
468 state[44]. Conversely, the non-exaggerated phase of t-SNE allows local data structures to be revealed.
469 [1]. Therefore, cytometry data analysis would be missing valuable information if we reduced t-SNE
470 applicability to finding only well separated clusters, especially since other techniques would
471 perform that task better and faster. However, some workflows call for t-SNE pre-processing to
472 facilitate extraction of cluster features from multidimensional data [5,45]. In those cases, it may be
473 helpful to adapt the opt-SNE toolkit to terminate the embedding calculation immediately at the EE
474 stop iteration and re-assess the high-dimensional structure within each cluster. Alternatively, a 'late
475 exaggeration approach'[27,41] can be cautiously applied to create very tight clusters, although in our
476 experience this approach was only marginally beneficial for global structure representation while
477 detrimental for local structure (data not shown). Interestingly, UMAP embeddings also presented
478 a challenge to perceive both cluster data structure and continuum marker expression data structure
479 within the same embedding, since increasing the number of neighbors and decreasing the minimum
480 distance parameters seemed to benefit the former but penalize the latter task (Suppl.Fig. 8C-E and
481 [27]). These results suggest that UMAP is highly promising but not readily available yet to completely
482 replace t-SNE as a state-of-the-art method for cytometry data visualization.
483 It is advisable to note that certain data structures, such as cluster hierarchy, cannot be revealed
484 with t-SNE [46,47]. t-SNE accessibility in cytometry analysis software lead to its not infrequent misuse
485 with cytometry data, evident when the cluster-like structure is not prominent in the map. Therefore,
486 the features identified from t-SNE embedding in its current form should be verified with alternative
487 methods when possible for confirmational purposes. Im et al also suggest that if a continuous
488 manifold structure exists in the data, large perplexity values may cause artificial breaks
489 (overclustering) [47]. The perplexity values commonly used in cytometry analysis are on the lower
490 end of the suggested range for efficient clustering, as it is often advised to scale the number of
491 nearest neighbors to the average cluster size [48]; however, if computationally feasible, higher
492 perplexity values might facilitate feature preservation for markers whose expression is not
493 bimodally distributed. To preserve global geometry, Kobak et al[27] propose to perform the
494 embedding in two steps, and use the kNN-based extrapolation of a high-perplexity t-SNE of a
495 subset of the large dataset as initialization to run the t-SNE of the whole dataset.

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

Acknowledgements. The authors would like to thank Yvan Saeys, Gary Kazantsev, Jonathan Irish, Dmitry Kobak, 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 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 and https://github.com/omiq-ai/Multicore-opt-SNE. opt-SNE is also available in the FlowJo data analysis software version 10.6 and later, and the SeqGeq version 1.5 or later, both downloadable from http://www.flowjo.com. The code of modified van der Maaten BH-tSNE C++ implementation with user-accessible parameters is available upon request. Van Unen et al dataset[7] 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 .

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 Beckton Dickinson (BD); FlowJo is a subsidiary of BD.

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).

539 4 Wong, M. T. *et al.* Mapping the Diversity of Follicular Helper T Cells in Human
540 Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell*
541 *reports* **11**, 1822-1833, doi:10.1016/j.celrep.2015.05.022 (2015).
542 5 Becher, B. *et al.* High-dimensional analysis of the murine myeloid cell system.
543 *Nat Immunol* **15**, 1181-1189, doi:10.1038/ni.3006
544 http://www.nature.com/ni/journal/v15/n12/abs/ni.3006.html#supplementary-
545 information (2014).
546 6 Chen, H. *et al.* Cytofkit: A Bioconductor Package for an Integrated Mass
547 Cytometry Data Analysis Pipeline. *PLoS Comput Biol* **12**, e1005112,
548 doi:10.1371/journal.pcbi.1005112 (2016).
549 7 van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical
550 stochastic neighbour embedding reveals rare cell types. *Nat Commun* **8**, 1740,
551 doi:10.1038/s41467-017-01689-9 (2017).
552 8 Tang, J., Liu, J., Zhang, M. & Mei, Q. in *Proceedings of the 25th International*
553 *Conference on World Wide Web %@ 978-1-4503-4143-1* 287-297
554 (International World Wide Web Conferences Steering Committee,
555 Montréal, Québec, Canada, 2016).
556 9 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
557 Projection for Dimension Reduction. *eprint arXiv:1802.03426*,
558 arXiv:1802.03426 (2018).
559 10 Kratochvíl, M. *et al.* SOM-based embedding improves efficiency of high-
560 dimensional cytometry data analysis. *bioRxiv*, 496869, doi:10.1101/496869
561 (2019).
562 11 Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and
563 interpretation of cytometry data. *Cytometry A* **87**, 636-645,
564 doi:10.1002/cyto.a.22625 (2015).
565 12 Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals
566 Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197,
567 doi:10.1016/j.cell.2015.05.047 (2015).
568 13 Donnenberg, A. D. & Donnenberg, V. S. Rare-event analysis in flow cytometry.
569 *Clin Lab Med* **27**, 627-652, viii, doi:10.1016/j.cll.2007.05.013 (2007).
570 14 DiGiuseppe, J. A., Tadmor, M. D. & Pe'er, D. Detection of minimal residual
571 disease in B lymphoblastic leukemia using viSNE. *Cytometry. Part B, Clinical*
572 *cytometry* **88**, 294-304, doi:10.1002/cyto.b.21252 (2015).
573 15 Lin, L. *et al.* Identification and visualization of multidimensional antigen-
574 specific T-cell populations in polychromatic cytometry data. *Cytometry A* **87**,
575 675-682, doi:10.1002/cyto.a.22623 (2015).
576 16 Hirakawa, M. *et al.* Low-dose IL-2 selectively activates subsets of CD4(+) Tregs
577 and NK cells. *JCI insight* **1**, e89278, doi:10.1172/jci.insight.89278 (2016).
578 17 Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical
579 Stochastic Neighbor Embedding. *Computer Graphics Forum* **35**, 21-30,
580 doi:10.1111/cgf.12878 (2016).
581 18 Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast
582 interpolation-based t-SNE for improved visualization of single-cell RNA-seq
583 data. *Nature Methods* **16**, 243-245, doi:10.1038/s41592-018-0308-4 (2019).

19 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).

20 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).

21 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).

22 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).

23 Ulyanov, D. (Github, https://github.com/DmitryUlyanov/, 2016).

24 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).

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

26 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).

27 Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *bioRxiv*, 453449, doi:10.1101/453449 (2018).

28 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).

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

30 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).

31 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).

32 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).

33 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).

34 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).

35 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).

36 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).

37 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).

38 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).

39 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).

40 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).

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

42 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).

43 Arora, S., Hu, W. & Kothari, P. K. in *Proceedings of the 31st Conference On Learning Theory* Vol. 75 (eds Bubeck Sébastien, Perchet Vianney, & Rigollet Philippe) 1455--1462 (PMLR, Proceedings of Machine Learning Research, 2018).

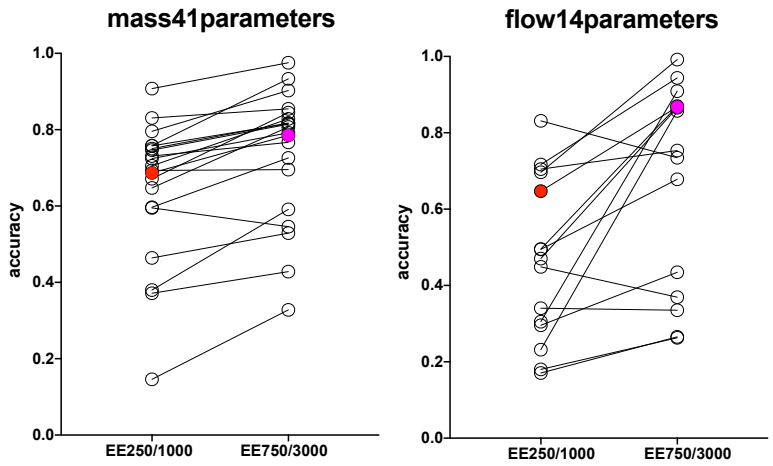
44 Chattopadhyay, P. K. & Roederer, M. Cytometry: today's technology and tomorrow's horizons. *Methods* **57**, 251-258, doi:10.1016/j.ymeth.2012.02.009 (2012).

45 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).

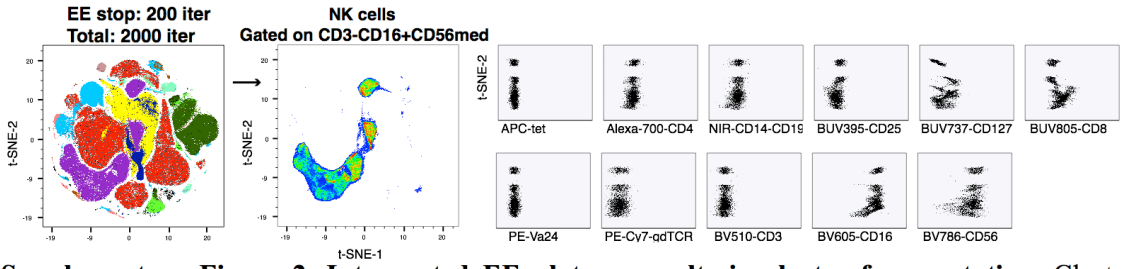
46 Amid, E. & Warmuth, M. K. A more globally accurate dimensionality reduction method using triplets. *arXiv:1803.00854* (2018).

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

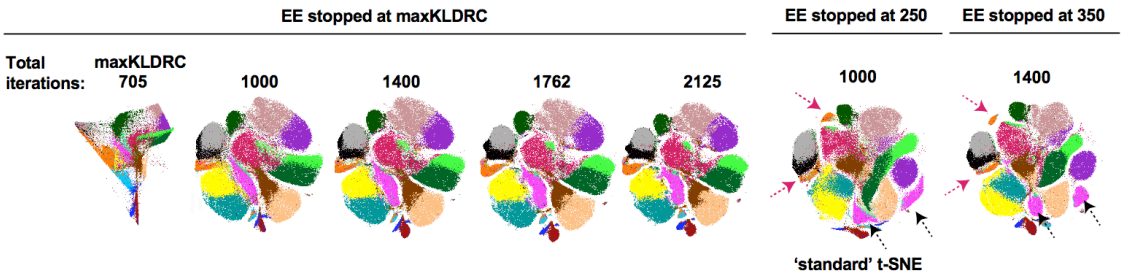
48 Cao, Y. & Wang, L. Automatic Selection of t-SNE Perplexity. *arXiv:1708.03229* (2017).



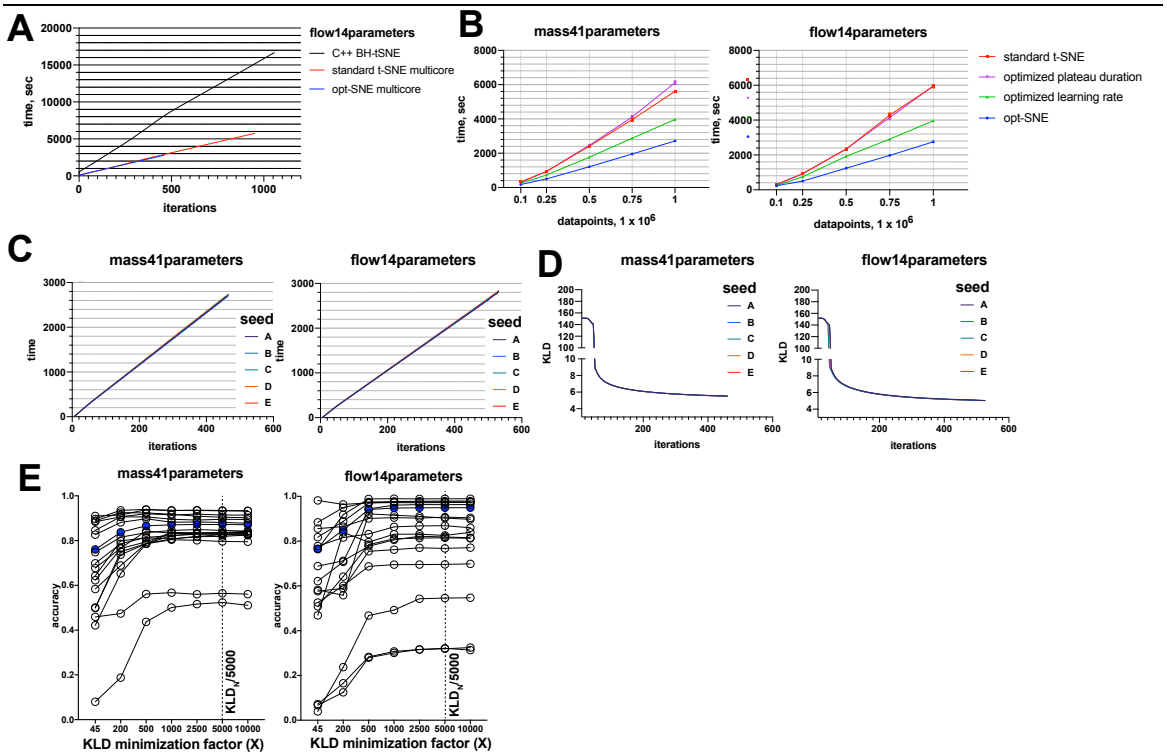
Supplementary Figure 1. 1NN accuracy of 'default' t-SNE embedding compared to 'extended' setup with longer EE. Accuracy scores for standard EE250/1000 and 'extended' EE750/3000 computation of mass cytometry (left) and flow cytometry (right) data per assigned class values (cell subsets as listed on Fig.1-2, open circles; overal scores, filled circles). Representative examples of multiple runs initiated with varying seed values are shown.



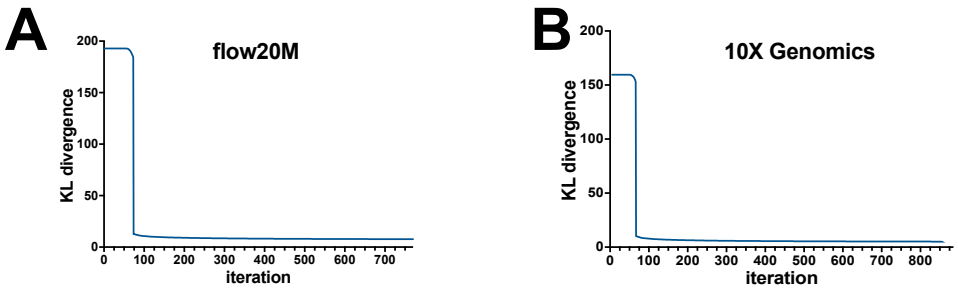
Supplementary Figure 2. Interrupted EE plateau results in cluster fragmentation. Clusters corresponding to CD3-CD16+CD56med NK cells were subsetted from the dataset and assessed as biaxial plots of different parameters plotted versus t-SNE-2 axis. Color overlays correspond to cell type classes labeled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



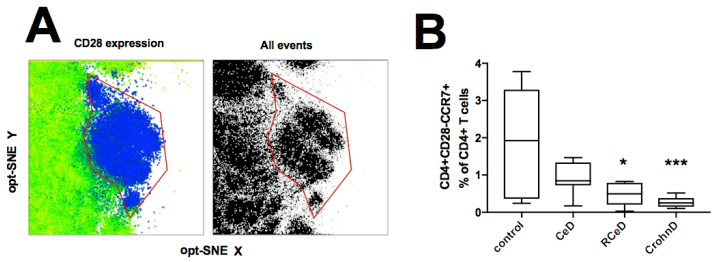
Supplementary Figure 3. KLD plateau effect on subset visualization in t-SNE maps. Mass cytometry data visualizations are shown as generated with varying duration of post-EE iteration time. Arrows mark cluster fragmentation. All color overlays correspond to cell type classes labeled as in Fig. 1-2. Representative examples of multiple runs with varying seed values are shown.



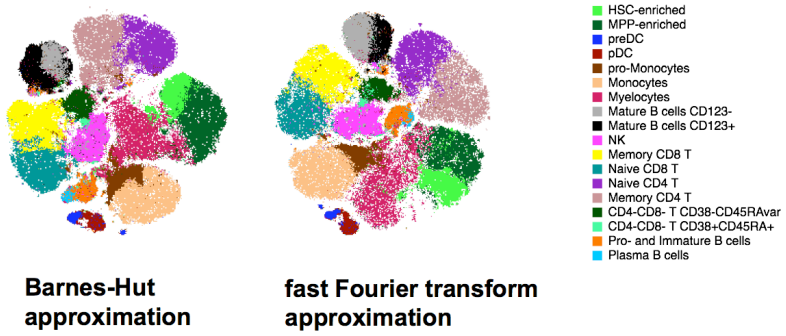
Supplementary Figure 4. Evaluation of opt-SNE embeddings. A, Speed benchmarks of original BH-tSNE (black), multicore BH-tSNE with standard settings (red), and multicore opt-SNE (blue). Representative examples of multiple runs with varying seed values are shown. B. Speed of t-SNE embedding (multicore BH-tSNE) with standard settings (red), optimized plateau duration/standard h = 200 (purple), optimized h/standard EE=250 iterations (green), and opt-SNE (blue). Average of 5 random seed runs for 5 dataset sizes (0.1-1 x 10^6 cells) of mass and flow cytometry data are plotted; error bars denote SEM. C, speed, and D, KLD graph of opt-SNE embeddings (multicore BH-tSNE) initiated from 5 random seed runs for 1 x 10^6 cells of mass (left) and flow (right) cytometry data are plotted; color indicates various seeds. E, 1NN accuracy scores for opt-SNE embeddings of of mass cytometry (left) and flow cytometry (right) data per assigned class values (cell subsets, open circles; overal scores, filled circles) for embeddings terminated based on various KLD minimization criteria X in $(KLD_{N-1} - KLD_N) < KLD_N/X$, with X plotted on horizontal axis. Representative examples of multiple runs initiated with varying seed values are shown.



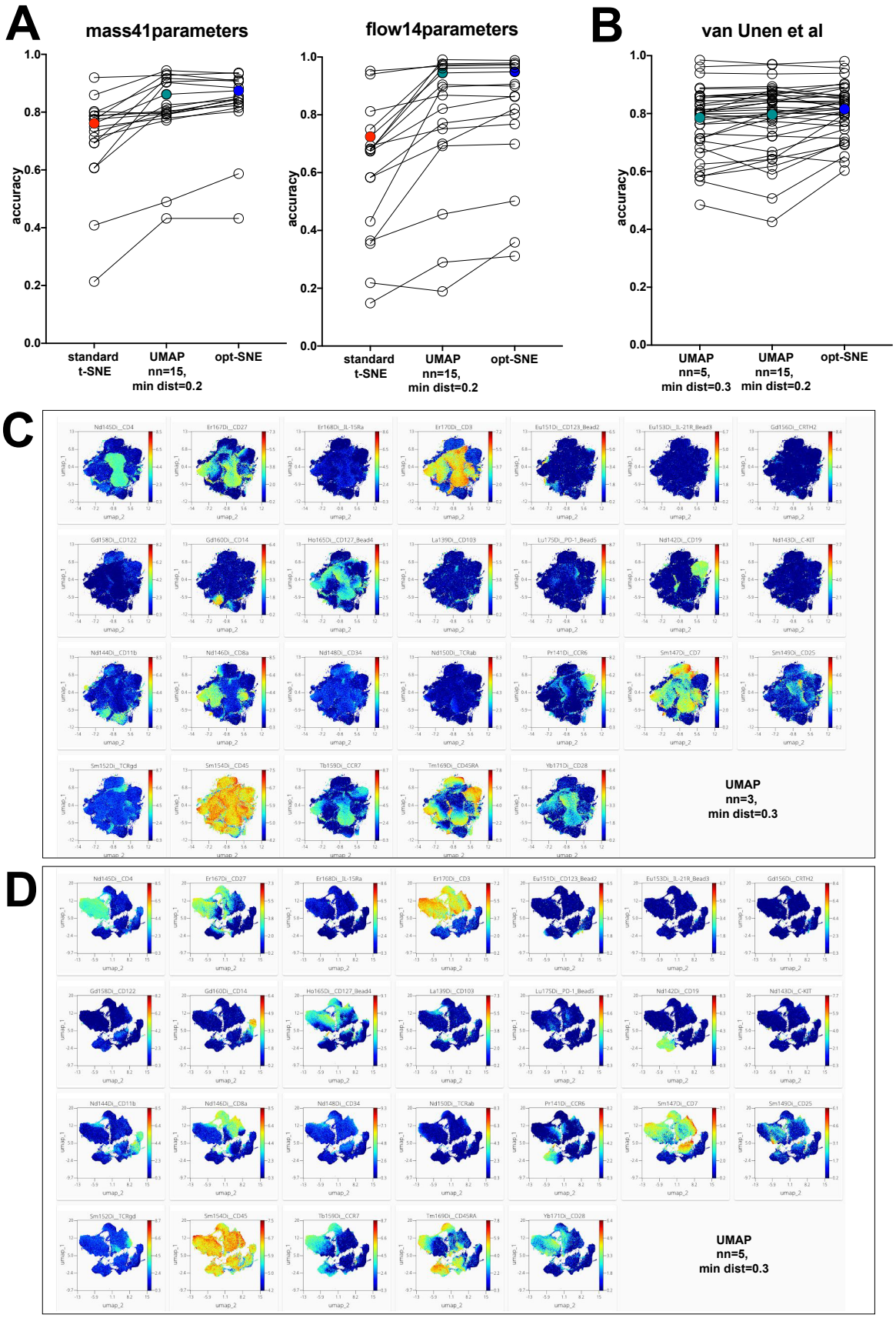
Supplementary Figure 5. KLD graphs for opt-SNE embedding of flow20M dataset (A) and 10X Genomics 1.3 million datapoints scRNA-seq dataset (B).



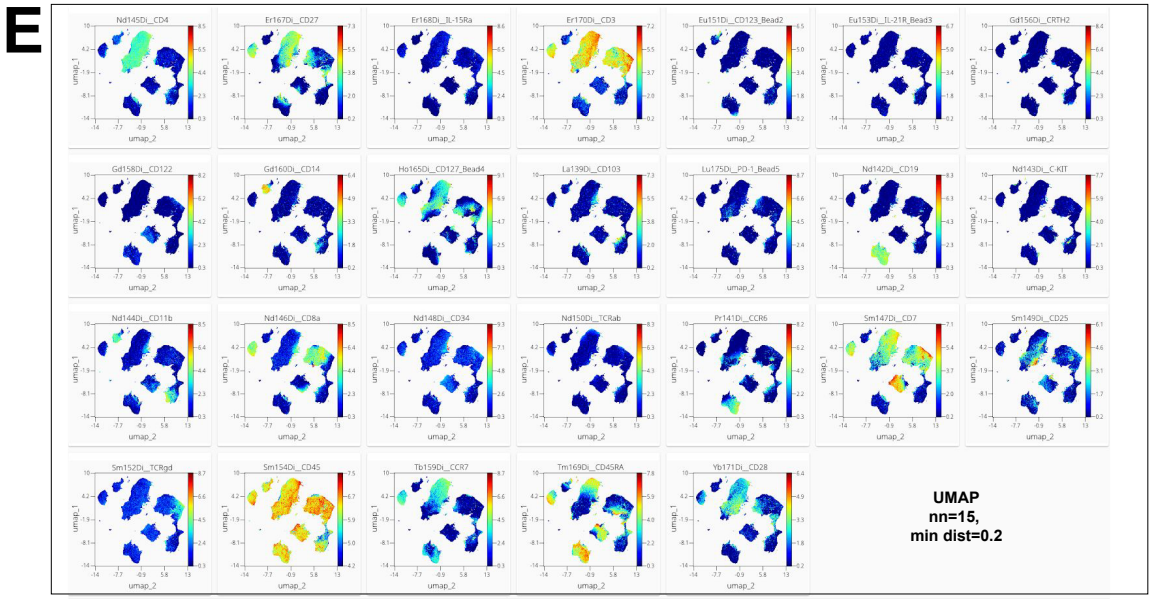
Supplementary Figure 6. A. Enlarged fragment of opt-SNE embedding showing CD4+CD28-CCR7+ cell cluster (left: color indicates CD28 expression intensity; right: dot plot). B. Frequencies of CD4+CD28-CCR7+ cells in CD4+ PBMC compared between different cohorts of subjects. CeD, Celiac disease; RCeD, refractory celiac disease; CrohnD, Crohn's disease. Two-tailed T tests were performed for statistical analysis. * p = 0.0258; *** p = 0.0002.



Supplementary Fig. 7. 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.


A
mass41parameters
flow14parameters
B
van Unen et al
accuracy
1.0
0.8
0.6
0.4
0.2
0.0
standard t-SNE
UMAP nn=15, min dist=0.2
opt-SNE
UMAP nn=5, min dist=0.3
C
UMAP nn=3, min dist=0.3
D
UMAP nn=5, min dist=0.3

691



Supplementary Figure 8. NN accuracy comparison of standard t-SNE, UMAP and opt-SNE. A, Accuracy scores for gated cytometry data (biologically relevant subsets as listed in Fig.1-2, open circles; overal scores, filled circles). UMAP parameters were applied as indicated. A representative sample of 5 random seed runs is shown. B, van Unen et al dataset was clustered with FlowSOM and accuracy scores for visualizations generated with indicated UMAP parameters as well as opt-SNE embedding are presented. Representative examples of multiple runs initiated with varying seed values are shown. C-E, UMAP embeddings visualized with mass cytometry labeling median intensities overlaid as color gradients.

692