Phenomics (2022) 2:323–335
https://doi.org/10.1007/s43657-022-00056-z

ARTICLE

A Hashing-Based Framework for Enhancing Cluster Delineation of High-Dimensional Single-Cell Profiles

Xiao Liu[1] · Ting Zhang[1] · Ziyang Tan[1] · Antony R. Warden[1] · Shanhe Li[1] · Edwin Cheung[2,3] · Xianting Ding[1]

Received: 9 February 2022 / Revised: 8 April 2022 / Accepted: 15 April 2022 / Published online: 19 May 2022

Abstract

Although many methods have been developed to explore the function of cells by clustering high-dimensional (HD) single-cell omics data, the inconspicuously differential expressions of biomarkers of proteins or genes across all cells disturb the cell cluster delineation and downstream analysis. Here, we introduce a hashing-based framework to improve the delineation of cell clusters, which is based on the hypothesis that one variable with no significant differences can be decomposed into more diversely latent variables to distinguish cells. By projecting the original data into a sparse HD space, fly and densefly hashing preprocessing retain the local structure of data, and improve the cluster delineation of existing clustering methods, such as PhenoGraph. Moreover, the analyses on mass cytometry dataset show that our hashing-based framework manages to unveil new hidden heterogeneities in cell clusters. The proposed framework promotes the utilization of cell biomarkers and enriches the biological findings by introducing more latent variables.

Keywords Cluster delineation · Hashing preprocessing · High dimensional · Single cell

Introduction

High-dimensional (HD) single-cell profile analyses, such as mass cytometry (CyTOF) (Good et al. 2018; Levine et al. 2021; Quintelier et al. 2021; Spitzer et al. 2017; Tu et al. 2019) and single-cell RNA sequencing (scRNA-seq) (Reid et al. 2018; Witt et al. 2019), provide extraordinary insights into the proteomics and genomics of single cells. Recent studies reveal new functional diversity and heterogeneity among cell populations through unbiased HD data analyses (Aghaeepour et al. 2017; Denis et al. 2018; Van Unen et al. 2016), with the aid of recently developed informatics techniques including spanning-tree progression analysis of density-normalized events (SPADE) (Anchang et al. 2016), hierarchical stochastic neighbor embedding (HSNE) (van Unen et al. 2017), independent component analysis (ICA) (Jin et al. 2019), and single-cell interpretation via multikernel learning (SIMLR) (Wang et al. 2017). Quality control (Kleinsteuber et al. 2016), gene selection (Tang et al. 2018), normalization (Finck et al. 2013), batch effects removal (Schuyler et al. 2019), and other preprocessing steps were frequently applied to control technical noise and improve the data quality. However, existing analysis methods commonly encounter with two inherent challenges. First, HD data often involves variables (markers of CyTOF data or genes of scRNA-seq data) whose expression values have no significant differences across all cells, which interferes with the process of cell clustering (Fig. 1a). Second, the large population of examined cells and their disturb variables lead to the vague cluster delineation, so that the cell clusters (i.e., cell populations) cannot be well visualized (Fig. 1a). These two inherent challenges appeal for new data processing methods to improve cell clustering accuracy.

Herein, we present a hashing-based framework to improve the delineation of cell clusters. Assuming that the expression value of a marker shows no significant difference across all cells that may be caused by the high correlation

Xiao Liu, Ting Zhang and Ziyang Tan contributed equally to this work.

✉ Xianting Ding
dingxianting@sjtu.edu.cn

[1] Institute of Personalized Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

[2] Cancer Centre, Faculty of Health Sciences, University of Macau, Taipa 999078, China

[3] Centre of Precision Medicine Research and Training, Faculty of Health Sciences, University of Macau, Taipa 999078, China

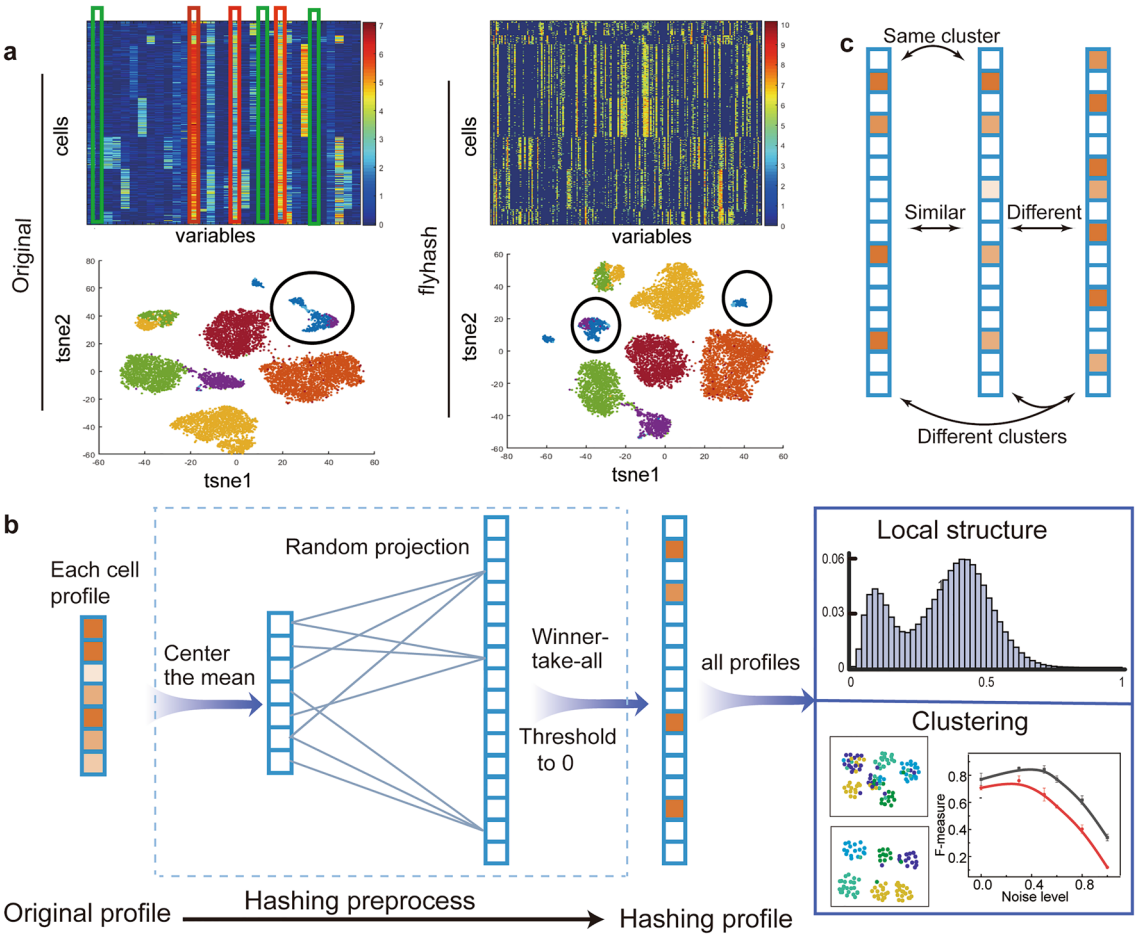

Fig. 1 The motivation and overview of our hashing-based framework. **a** The heatmap and tsne visualization of original and FHpre dataset. Variables with no significant differential expression across all cells in the original data disturb the cell cluster delineation (red and green box). After FHpre, the cells are more easily to cluster with clearer delineation (black circle). **b** The framework involves data projection and the down streaming analysis of local structure preservation and cell clustering. After normalization for each cell profile, our framework projects the normalized cell profile into a high dimension sparse tags by a sparse, binary random projection matrix. Then, a WTA strategy or 0-threshold-based binarization representation is used to transform the dense vectors of single-cell profiles into sparse tags. Finally, cell clustering is performed on all sparse cell profiles. The characteristic of local structure preservation of our framework is also verified on all sparse cell profiles. **c** Our framework retains the relationship between cell profiles, where profiles of the same cluster remain similar while the variations between different clusters are enhanced

between this marker and all cells, we hypothesize that such high correlation tends to connect cells together rather than separate them. Therefore, unlike the commonly used dimensionality reduction, we adopt dimensionality expansion strategy to extend markers into more latent variables in virtue of a theory that a highly correlated vector can readily be decomposed into the linear combinations of several independent vectors. These latent variables, which are merely highly associated with partial cells, promote the differentiation of cell populations and simultaneously ensure the full utilization of marker information. In other words, the highly correlated variables with all cells can be regarded as the irrelevant variables with subpopulation, and these variables are further decomposed into several subpopulation-relevant latent variables.

Locality-sensitive hashing-based (LSH) method (Dasgupta et al. 2017) can generate prolonged tags to decrease the influence of subpopulation-irrelevant variables. In addition, this method qualifies the property of preserving local data structures, that is to say, the neighborhood relationships of cells are basically unchanged via LSH preprocessing. By transferring the original input tags of single cells into prolonged sparse tags, it has been demonstrated that LSH-based method is compatible with many k-nearest neighbors (k-NN) based methods, such as SIMLR (Wang et al. 2017), locally linear embedding (LLE) (Roweis and Saul 2000), and t-SNE (Van der Maaten and Hinton 2008).

Herein, considering the superiority of LSH-based method and the computational cost, we introduce two variants of LSH into the framework for single-cell data preprocessing,

namely fly hashing (Dasgupta et al. 2017) preprocessing (FHpre) and densefly hashing preprocessing (dFHpre) (Chen et al. 2020). We validate the proposed framework through four aspects: first, we confirm that FHpre and dFHpre preserve local structures and elongate inter-cluster distances with local F1 and Spearman's correlation on four scRNA-seq and two CyTOF datasets (Supplementary Table S1, Supplementary Note). Second, applying the proposed framework to existing clustering methods, such as PhenoGraph (Levine et al. 2015) and ACCENSE (Shekhar et al. 2014), we benchmark its delineating ability by visualizing cell clusters, and verify its clustering improvement by comparing with the ground truth. Third, we show that in two separate public CyTOF data sets, the framework unveils new heterogeneities that are previously concealed. Lastly, we verify that the proposed framework is not only limited to single-cell data, but also can be scaled to computer vision data sets (MNIST and NORB, Supplementary Note). Our proposed framework offers a new platform revenue for the analysis of HD single-cell data. Of note, under the proposed framework, FHpre and dFHpre are both efficient, compatible, and comprehensible strategies that improve the clustering delineation of HD single-cell data.

Methods

Preliminaries

LSH

A hash function $\boldsymbol{h} : \mathcal{R}^d \rightarrow \mathcal{R}^m$ is locality-sensitive if for any two points $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{R}^d$, $Pr\left[h(\boldsymbol{p}) = h(\boldsymbol{q})\right] = sim(\boldsymbol{p}, \boldsymbol{q})$, where $sim(\boldsymbol{p}, \boldsymbol{q}) \in [0, 1]$ is a similarity function between the two input points (Dasgupta et al. 2017).

Simhashing

Simhashing was proposed to generate a binary hashing code for an original vector x (Charikar 2002). First, $20k$ (the hashing dimension) random projection vectors, $\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{20k}$, are generated, where each element is uniformly sampled from a Gaussian distribution $\mathcal{N}(0, 1)$. Then, the simhashing code for the i^{th} component of the original vector $\boldsymbol{x}$ was defined as

$$\boldsymbol{h}(x)_i = \begin{cases} 1, if \boldsymbol{r}_i \cdot x \geq 0, \\ 0, if \boldsymbol{r}_i \cdot x < 0. \end{cases} \tag{1}$$

According to the above definition, for the original cell-marker or cell-gene expression matrix $\boldsymbol{X} \in \mathcal{R}^{n \times m}$, the binary hashing by simhashing preprocessing (SHpre) was computed as

$$\boldsymbol{X}_{\boldsymbol{SHpre}} = \boldsymbol{X} \cdot \boldsymbol{R} \tag{2}$$

where $\boldsymbol{X}_{\boldsymbol{SHpre}}$ denotes the matrix preprocessed by simhashing, $\boldsymbol{R} = \left[\boldsymbol{r}_1 \boldsymbol{r}_2 \cdots \boldsymbol{r}_{20k}\right] \in \mathcal{R}^{m \times 20k}$ represents the random projection matrix.

Fly hashing

Similarly, fly hashing first projects the normalized input HD vectors (original tags) with a sparse, binary random projection matrix as follows (Dasgupta et al. 2017):

$$\boldsymbol{KC} = \boldsymbol{PN} \times \boldsymbol{P}. \tag{3}$$

Here, $\boldsymbol{KC}$ is the fly hashing tags ($n \times 20k$), $\boldsymbol{PN}$ represents the normalized input tags ($n \times m$) and $\boldsymbol{P}$ is the projection matrix ($m \times 20k$). $\boldsymbol{P}$ contains $n \times s\%$ ($s\%$ is the sampling ratio; $s = 10$ after optimization) entries of *1* randomly in each column while leaving other entries as *0*. n represents the event counts and m is the number of variables. k is called hash length as k entries are remained after the winner-take-all (WTA) (Yagnik et al. 2011) step. Longer hash length usually leads to better performance. Here, with consideration for trade-offs between performance and computational cost, we set k to 100 for MNIST and CyTOF data sets, 1000 for scRNA-seq data sets.

Afterwards, $\boldsymbol{KC}$ underwent a WTA step with Matlab code:

$$prc = prctile(\boldsymbol{KC}, 95, 2);$$

$$\boldsymbol{KC}(\boldsymbol{KC} < prc) = 0;$$

In this step, all entries with values smaller than the 95th percentile are set to *0* as a bionic process resembling the feedback inhibition from the anterior paired lateral (APL) neuron in the fruit fly olfactory circuit.

Densefly hashing

Densefly hashing first projects the normalized input HD vectors (original tags) as fly hashing in Eq. (3), but $\boldsymbol{KC}$ underwent a binarization transformation in dFHpre (Chen et al. 2020):

$$\boldsymbol{KC}(\boldsymbol{KC} < 0) = 0;$$

$$\boldsymbol{KC}(\boldsymbol{KC} \geq 0) = 1;$$

In this step, all entries with values smaller than *0* are set to *0* and the remaining to *1* resembling the procedures in Simhashing. For fair comparison, the hashing length of dFHpre and SHpre is set to the same as FHpre. We set the hashing length k for dFHpre and SHpre both to 100 for MNIST and CyTOF data sets, and 1000 for scRNA-seq data.

All parameters mentioned above are optimized after a trade-off or obtained directly from the previous research (Dasgupta et al. 2017).

Local F_1 score and distance of local structure

For each data set, a subset of 10,000 data points (or all data points if sample size is less than 10,000) was first randomly selected. Then, 1000 random query vectors were selected (100 random query vectors were selected for those data sets with less than 10,000 data points) from the subset and compared their true versus predicted top y% nearest neighbors. Taken the original data $\boldsymbol{X}_{n\times t}$ (i.e., true data) and FHpre data $\boldsymbol{Y}_{n\times p}$ (i.e., predicted) as examples, the randomly selected query vectors are denoted as $\boldsymbol{qx}_{1000\times t}$ and $\boldsymbol{qy}_{1000\times p}$, and the corresponding remaining data are, respectively, denoted as $\boldsymbol{x}_{9000\times t}$ and $\boldsymbol{y}_{9000\times p}$. For each row vector of $\boldsymbol{qx}_{1000\times t}$, we compute its true nearest neighbor based on cosine distance, while for each row vector of $\boldsymbol{qy}_{1000\times p}$, we compute its top y% predicted nearest neighbors based on cosine distance. Then, the mean average precision (MAP) (Yue et al. 2013) is calculated based on the overlap ratio of the true nearest neighbor and the top y% predicted nearest neighbors. Conversely, the mean average recall (MAR) can be obtained by the ratio of the overlap between the top y% true nearest neighbors and the predicted nearest neighbor. Here we traversed y from 2 to 10 with an interval 2. The Local F1 Score is defined as the harmonic average of MAP and MAR from 50 trials corresponding to different random projection matrices and queries, and the reciprocal of local F1 score is regarded as the local structure distance between input and output space. Local F1 score is a mutual metric, whose results from data set A to data set B would be the same vice versa. Therefore, the local structure distance is mutual as well, which is beneficial to the comparison of the local structure similarity. The larger distance represents bigger differences between the true and predicted local structure. It is noteworthy that the local structure distance defined here is not additive; however, this does not hamper the similarity comparison of local structures.

Evaluation Index

To evaluate the performance of our proposed framework, we adopt the following six evaluation indexes to measure the clustering accuracy.

F-measure

The weighted F-measure is defined as the weighted summary of F_i of each cluster over all clusters:

$$F = \sum_i \frac{n_i}{N} F_i,$$

where n_i denotes the number of cells of cluster i and N is the whole number of cells. F_i is the F-measure of cluster i, which is defined as the harmonic mean of precision and recall:

$$F_i = \frac{2P_i R_i}{P_i + R_i}$$

where P_i and R_i are the precision and recall of cluster i, respectively.

NMI

The normalized mutual information (NMI) is a widely used index to measure the similarity between the ground truth and the predicted result. The definition of NMI is formulated as

$$NMI(g,p) = \frac{-2\sum_{i=1}^{k1}\sum_{j=1}^{k2} \boldsymbol{M}_{ij} \log\left(\frac{\boldsymbol{M}_{ij}m}{\boldsymbol{M}_{i\cdot}\boldsymbol{M}_{\cdot j}}\right)}{\sum_{i=1}^{k1} \boldsymbol{M}_{i\cdot} \log\left(\frac{\boldsymbol{M}_{i\cdot}}{m}\right) + \sum_{j=1}^{k2} M_{\cdot j} \log\left(\frac{\boldsymbol{M}_{\cdot j}}{m}\right)},$$

where k_1 is the number of cell subpopulations of the ground truth, k_2 is the number of cell subpopulations of the algorithm, and m is the number of cells. The matrix $\boldsymbol{M}$ denotes the confusion matrix, where $\boldsymbol{M}_{i\cdot}$ and $\boldsymbol{M}_{\cdot j}$ are the sum over the i-th row and the j-th column of confusion matrix, respectively.

Accuracy

The accuracy is defined as the ratio of correctly clustered cells in the whole cells, which measures the proportion of how many cells are correctly classified. Assumed there are N cells, $\boldsymbol{g}$ is the cell labels of the ground truth, and $\boldsymbol{p}$ is the cell labels from the proposed algorithm. Then, the accuracy is defined as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \delta\left(\boldsymbol{g}_i, \text{map}\left(\boldsymbol{p}_i\right)\right),$$

where we adopt the Hungarian assignment algorithm as the map function.

Davies–Bouldin and Calinski–Harabasz

Davies–Bouldin and Calinski–Harabasz are two indexes evaluating the essential cluster structure of data and not relying on any ground truth. Davies–Bouldin measures the intra-cluster dispersion by averaging the distance between the data point and its corresponding centroid, while measures the inter-cluster dispersion by calculating the difference between

the two corresponding centroids. Unlike Davies–Bouldin, Calinski–Harabasz uses the averaged sum of squared distances within each cluster to measure the intra-cluster dispersion, and it uses the sum of squared distances between each cluster centroid and a fixed centroid of all data to measure the inter-cluster dispersion. These two indexes can be directly calculated by the MATLAB built-in function with the profile data and cluster label as input.

Silhouette

The silhouette for each point measures the similarity of that point with points in its own cluster, when compared to points in other clusters. The silhouette value for point i is defined as

$$S_i = \frac{b_i - a_i}{\max\left(a_i, b_i\right)},$$

where a_i is the average distance from the i-th point to the other points of cluster i, and b_i is the minimum average distance from the i-th point to points in all different clusters. The value of silhouette ranges from − 1 to 1, and a higher value indicates a better match.

Baseline Methods

Our proposed framework mainly involved three parts: data projection, cell clustering and visualization. For data projection, we compared three LSH-based methods: fly hashing (Dasgupta et al. 2017), densefly hashing (Chen et al. 2020) and simhashing (Charikar 2002), and discussed their capability of preserving local structure. Moreover, principal component analysis (PCA) and ICA were introduced as baseline methods for scRNA-seq data preprocessing. For cell clustering, we considered PhenoGraph (Levine et al. 2015) and ACCENSE (Shekhar et al. 2014), two commonly used single-cell classical clustering methods, as baseline clustering methods. Our framework improved the clustering accuracy, and additionally unveiled new heterogeneities in cell clusters. For the results visualization, we employed t-SNE and uniform manifold approximation and projection (UMAP) methods (Becht et al. 2019) as two baseline visualization methods, and we demonstrated that our framework clarify the visual delineation of cell clusters. Our hashing-based framework provides a new way to analyze single-cell data, which is compatible with many clustering methods.

Overview of the Proposed Hashing-Based Framework

Cells are usually clustered based on their variables, and cells of intra-cluster have similar variable expressions, while cells of inter-cluster have diverse variable expressions. Therefore, variables with similar expression values across all cells will disturb the process of cell clustering and make vague delineation of cell populations (Fig. 1a), such as variables CD133, CD14 and CD61 (Fig. 1a, green box), and they are highly correlated with all cells but not used to define any specific cell subpopulations. We considered such variables as subpopulation-irrelevant variables. In fact, these subpopulation-irrelevant variables usually contain informative features to unveil the new heterogeneous function of cell populations. To promote the utilization of such informative variables, we decompose such variables into multiple irrelevant latent variables based on a mathematical theory that a vector can be decomposed into a linear combination of unrelated vectors. In other words, the dimension of original data is further extended to even higher dimension. Simultaneously, taking computational cost into consideration, we propose a hashing-based framework to analyze single-cell data, including data preprocessing, local structure preservation and cell clustering (Fig. 1b).

The LSH family maps m-dimensional space into n-dimensional space while maintaining the local structures between the two spaces, specifically, the similarity within m-dimensional vectors (defined by distance metrics) remain identical in the n-dimensional space after mapping. Different from conventional hash projections that aim to avoid collisions for quick access of stored data, LSH aggregates similar vectors while avoids collisions between dissimilar vectors, which facilitates the cell clustering. Our framework adopts two LSH-based methods, namely fly hashing and densefly hashing, to map single-cell data.

Specifically, single-cell profiles are first normalized for comparability. Then, a random projection expands the input vector to $20{\cdot}k$-dimension (k is the hash length) by multiplying the input vector with a randomly generated sparse, binary projection matrix. This projection randomly sums up an optimized percentage of entries in the input vector as a new entry in the output vector. To decrease the computational cost, our framework adopts a strategy to further extract the sparse and expanded vector, such as a WTA process where entries below the 95th percentile are all set to zero while those above are retained in the final output vectors (fly hashing). Another strategy which binarizes the representation with a threshold of 0, where indices with values below 0 are set to 0, while the remaining are set to 1 (densefly hashing). The resulting vectors retain the similarity relationship between the same populations and enhance the heterogeneities between vectors of different populations (Fig. 1c). In this way, our framework generates prolonged tags that attenuate the effects of subpopulation-irrelevant variables and retains the most distinguishable variables. Taken the prolonged tags as input, our framework finally analyzes the local structure preservation and the clustering accuracy by FHpre and dFHpre.

Results

Local Structure Preservation and Inter-cluster Distance Elongation

To examine the preservation and elongation ability of our framework, we examined the pairwise distances and the similarity of neighborhood relationships between original and preprocessed data spaces, measured by Spearman's correlation coefficient (R) and local F_1 score, respectively. Here, we apply fly hashing and densefly hashing to preprocess the original data in our framework (Fig. 2 and Supplementary Fig. S1 and Fig. S2). For better comparison, simhashing algorithm is also incorporated and applied to preprocess the data (Supplementary Fig. S3).

Using Usoskin (Usoskin et al. 2015), Samusik (Samusik et al. 2016) and Levine (Levine et al. 2015) data sets as benchmarks, 10,000 data points (or all data points if sample size is less than 10,000) are randomly selected from each, leading to 49,995,000 sets of pairwise distances. These distances are clustered into 50 equal-width groups. For each group, it contains different number of distances, then we can calculate the median distance of each group and plot the boxplot for this group, these 50 boxplots are aligned based on the order of the median distance ranges from small to large, and the dashed line represents the line of equality (Fig. 2). The resulting median values of the boxplots form a crescent above the line of equality, which represents

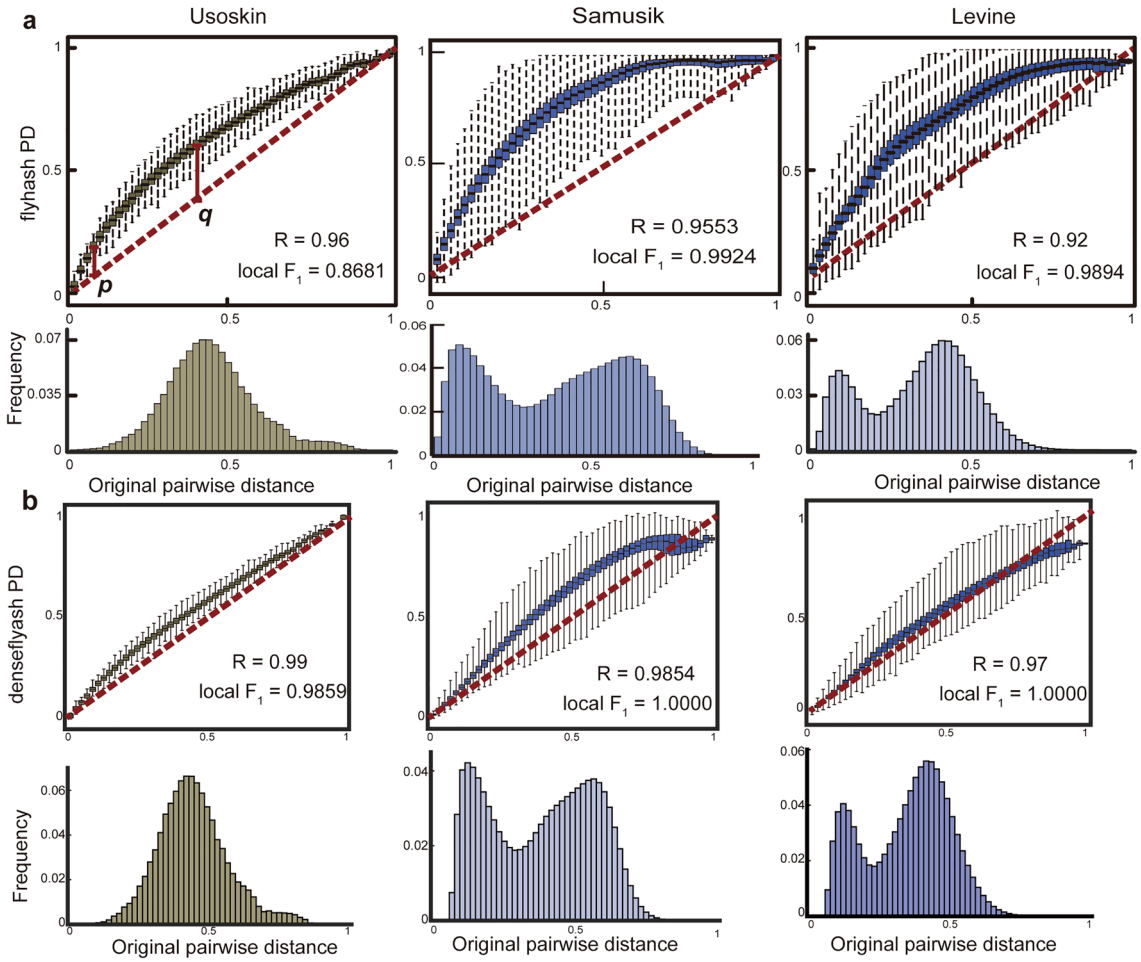

Fig. 2 Local structure preservation. **a** FHpre and **b** dFHpre preserve the local structure and elongates inter-cluster distances on three benchmark data sets. Box plots represent the changes of pairwise distances by FHpre (dFHpre). In each benchmark data set, 10,000 data points (or all data points if sample size is less than 10,000) are randomly selected, leading to 49,995,000 sets of pairwise distances, which are clustered into 50 equal-width groups. After applying FHpre (dFHpre), the distribution of corresponding distances within each group is summarized with a box plot. The dashed line is the identity line; p and q represent example offsets post-FHpre at short and medium distances. R of distances smaller than the 10th percentile and local F_1 score computed between the original and fly hashing (densefly hashing) pairwise distances are reported. Values closer to 1 represent better preservation of local structure. The lower row graphs of (**a** and **b**) show the histogram of pairwise distance distributions. PD: pairwise distance

larger changes in middle-range pairwise distances (q) than in short-range pairwise distances (p). Here, the middle-range pairwise distances (q) indicate the corresponding median distances distribute around 0.5, while the short-range pairwise distances (p) indicate the corresponding median distances distribute at the left bottom around 0. The data points of the same cluster have small pairwise distances, while the data points of different clusters have relatively large pairwise distances. Therefore, the short- and middle-range pairwise distances represent intra- and inter-cluster distances, respectively. In analyses, middle-range pairwise distances would interfere with clustering while short-range pairwise distances show less influence. Therefore, FHpre and dFHpre elongate inter-cluster distances while minimally increase intra-cluster distances.

R values and local F_1 scores are calculated to investigate whether the minor increase of intra-cluster distances disturb local structures (Fig. 2). R of pairwise distances measures the consistency of the local neighborhood pre- and post-FHpre (dFHpre). Examining different percentile of distances in the original and hashing pairwise space show the levels of the distance preservation on different distance scales (Supplementary Fig. S4a, b, d). The R values for Usoskin, Samusik, and Levine data sets based on the overall pairwise distances of dFHpre are 0.99, 0.98 and 0.97 (Fig. 2), respectively, higher than that of both FHpre (0.96, 0.95, and 0.92, respectively) and SHpre (0.86, 0.86, and 0.78, respectively) (Supplementary Figs. S2 and S3). The R values of other benchmark data sets are reported in Supplementary Figs. S1–S3. The R values confirm that the sequence of local pairwise distances of FHpre and dFHpre are undisturbed.

We then quantify the similarity of neighborhood relationships between data spaces, for both pre- and post-FHpre (dFHpre), using the local F_1 score to compare the true versus predicted nearest neighbors of randomly selected query vectors. Specifically, for each query vector, the top 2% of true nearest neighbors in the input space are identified, then their corresponding top 2% nearest neighbors in the output space post-FHpre (dFHpre) are calculated and termed as the predicted nearest neighbors. According to the true and predicted nearest neighbors of each query vector, the average precisions value from 100 queries are calculated to provide a more comprehensive and mutual index for neighborhood similarity. Similarly, the local F_1 scores of dFHpre calculated based on the Usoskin, Samusik, and Levine benchmark data sets are 0.98, 1.00, and 1.00 (Fig. 2), respectively, outperforming that of FHpre (0.86, 0.99, and 0.98, respectively) and SHpre (0.48, 0.76, and 0.70, respectively) (Supplementary Figs. S2 and S3). This indicates that the neighborhood relationships are mainly retained by post-FHpre and post-dFHpre. The local F_1 scores of other data sets are reported in Supplementary Figs. S1–S3 and the local F_1 score computed using varying percentages of the nearest neighbors are reported in Supplementary Fig. S4c. The above results indicate that dFHpre excels in local structure preservation and inter-cluster distances elongation when compared with FHpre and SHpre.

Retaining the Local Structure with Various Noise Levels

To properly simulate the HD single-cell profiles and demonstrate the restoration of local structure against noises, we employ the MNIST data set and augment it with increasing levels of noise (Supplementary Fig. S5a). These noises are randomly cut down from online landscape pictures and normalized for comparability. First, we compare the visual differences between standard t-SNE, FHpre t-SNE, dFHpre t_SNE and SHpre t-SNE layouts (Supplementary Figs. S6a and S7). As the noise level increases, the delineations between clusters become ambiguous in both layouts. However, FHpre and dFHpre decrease the ambiguity, especially at background levels 0.5 and 0.6, where FHpre and dFHpre t-SNE layouts provide much clearer delineations between clusters.

We regard ACCENSE as the base clustering method and perform it on the 2D space generated by t-SNE. The cluster delineations of FHpre are more distinguishable throughout all noise levels compared with the ACCENSE results on standard t-SNE maps (Supplementary Fig. S6b). We use F-measure to quantify the clustering accuracy between the true versus predicted labels. At all noise levels, especially at high noise level (Fig. 3a), FHpre obviously improves the clustering performance. The highest boost is 0.22 at noise level 1, almost tripling the F-measure value (0.34 vs 0.12). We also evaluate the clustering quality with additional three indexes: Calinski–Harabasz index (Caliński and Harabasz 1974) (the higher is better), Davies–Bouldin index (Davies and Bouldin 1979) (the lower is better), and silhouette coefficient (Kaufman and Rousseeuw 2009) (the higher is better) (Fig. 3a). These indexes quantify the clustering quality by maximizing the homogeneity of intra-clusters and minimizing heterogeneity of inter-clusters. The results of these three indexes further strengthen that FHpre improves the quality of ACCENSE clustering especially at high noise level.

Considering FHpre as a feature extraction method, we apply FHpre, dFHpre, SHpre, PCA and ICA to preprocess the MNIST data set with different noise levels. Then, PhenoGraph is performed on the processed data sets, and the corresponding F-measures are calculated to evaluate the clustering. As expected, FHpre shows higher F-measure at all noise levels (Fig. 3b), which is in accordance with visual comparison results. We then calculate the local structure distances between FHpre, PCA, and ICA processed data sets with various noise levels and the data set with no noise

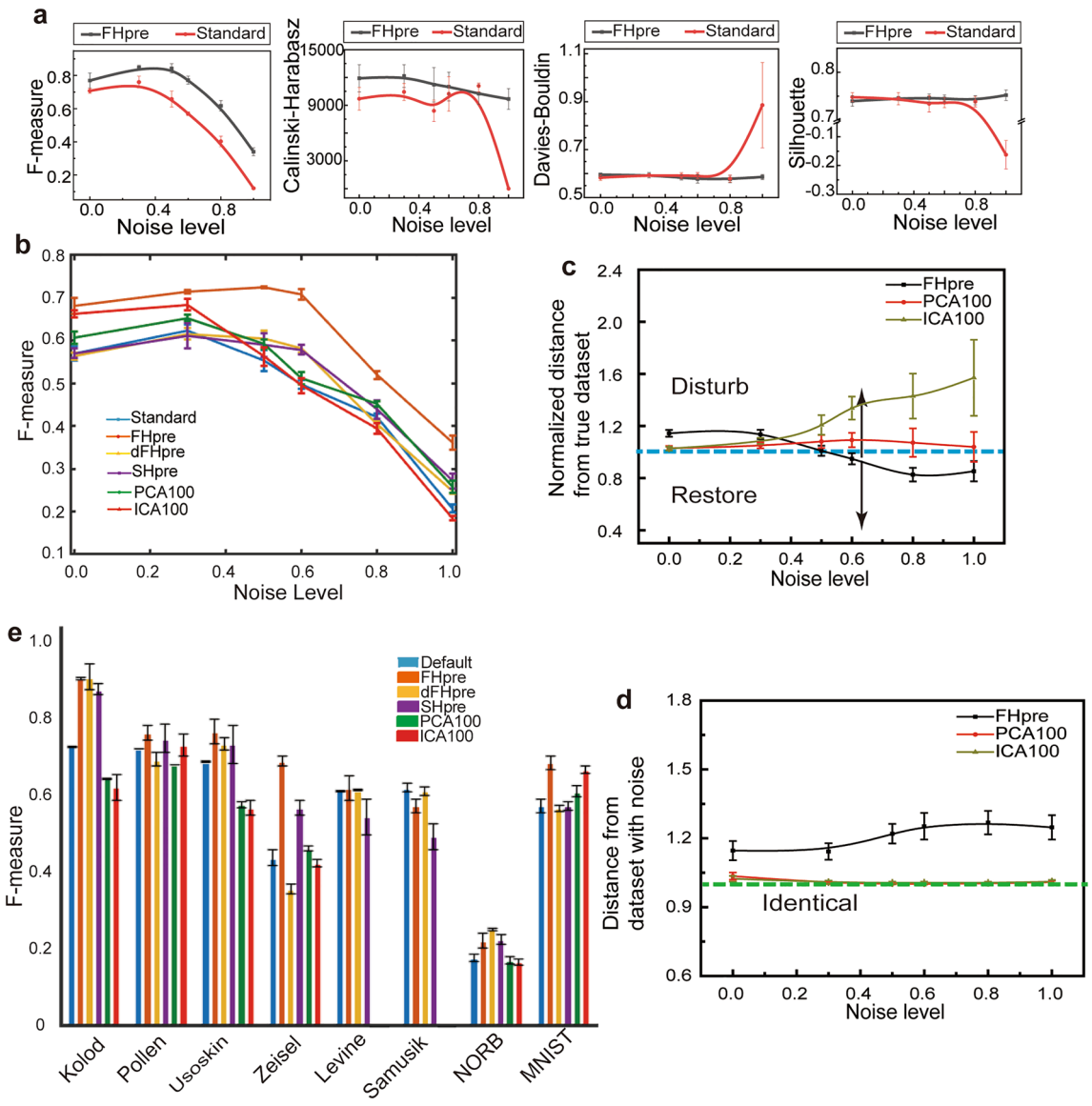

Fig. 3 The framework retains the local structure with various noise levels. **a** Evaluation of ACCENSE clustering results on original and FHpre MNIST dataset with noise level ranging from 0 to 1. The four evaluation indexes are *F*-measure, Calinski–Harabasz index, Davies–Bouldin index, and silhouette coefficient. **b** Comparisons of *F*-measure between different preprocessing methods with PhenoGraph as the clustering method. The component numbers of PCA and ICA take as 100. **c** Normalized local structure distances between true data space (with no noise) and FHpre, PCA, or ICA generated data spaces. Curves below the dashed line indicate the restoration of local structure against noise. **d** Local structure distances between data spaces with various level of noise and FHpre, PCA, or ICA generated data spaces. Curves further from the dashed line indicate more modification to data spaces with noise. **e** Comparisons of *F*-measure between different data sets with PhenoGraph as the clustering method. The component numbers of PCA and ICA are both 100

(denoted as the "true data set") (Fig. 3c and d). The distances are normalized by the distances between data sets with various noise levels and the true data set. The normalized distances demonstrate whether FHpre, PCA or ICA recover the local structure of true data set or further disturb it after the disturbance of artificial noise (Fig. 3c). Our result shows that after background level of 0.5, the FHpre curve descends below the identity line, which represents the tendency to restore local structure. In comparison, the curves of PCA and ICA are above the identity line, which represents the tendency to disturb local structure. Afterwards, we calculate the distances between data sets with various noise levels before and after FHpre, PCA, and ICA, respectively (Fig. 3d). The FHpre curve is above the identity line, representing that a large modification on the data sets disturbs artificial noise. In comparison, the modification of PCA and ICA on the disturbed data sets is minimal. Generally, FHpre restores the local structure against perturbations of various

noise levels (Supplementary Fig. S6b) and thus improves visual and quantitative analysis (Supplementary Fig. S5c).

In addition, we regard data sets with no artificial noise as true data set in the comparison as it is relatively closest to an ideal data set containing the least required variables for clustering. It is noteworthy that even at noise level of 0, FHpre, PCA, and ICA modified the true data set to some extent (Fig. 3c and d). The reason is that the true data set is free of artificial noise; however, it retains the intrinsic noise, which is the intrinsic subpopulation-irrelevant variables. One of the purposes of feature selection methods is to reduce the influence of these intrinsic noise, so all preprocessing improved the *F*-measure of clustering at noise level of 0 (Fig. 3b).

Visual and Quantitative Comparison of Benchmark Data Sets

To examine the clustering performance of elongated inter-cluster distances, we compare the dimension-reduction layouts and clustering accuracies of original with FHpre benchmark data sets. The t-SNE layouts (Fig. 4a–c) and UMAP layouts (Supplementary Fig. S8) of the benchmark data sets are presented for visual comparison. The color coding indicates the clustering label in accordance to the source data set. The short distances between clusters in t-SNE layout represent smaller distances between scatters associate with higher similarity among cells. FHpre does not significantly improve the layouts of MNIST and NORB data sets (Fig. 4a) because of their very vague cluster delineations in t-SNE visualization. Although the improvements of cluster delineations are comparatively small and invisible for Levine and Samusik data sets with a large number of cell clusters (Fig. 4b), FHpre makes scatters within the same subsets form clearer delineations in scRNA-seq data sets (Fig. 4c). Specifically, the circled scatters in the t-SNE layouts of Usoskin and Zeisel data sets, which contain mixed cells from several subsets in original data, show enhanced separation after applying FHpre. The similar result is also observed in the UMAP layouts (Supplementary Fig. S8). Unlike t-SNE layout, in which long-range distances are not associated with similarity (because t-SNE employs Gaussian distribution in the HD space and t-distribution in the low-dimension space), UMAP retains the association between distances in the layout and similarities of cells in both long- and short-range

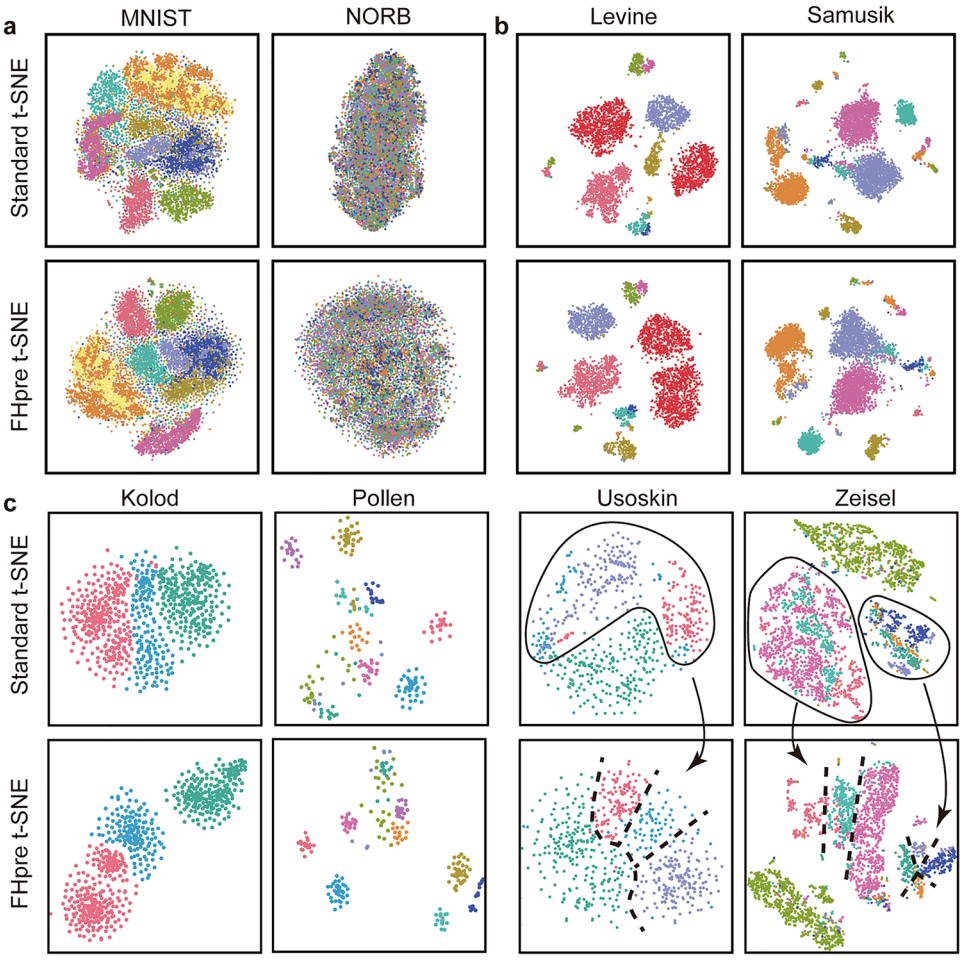

Fig. 4 Visual and quantitative comparison of benchmark data sets. The standard t-SNE and FHpre t-SNE layouts of **a** two image recognition data sets, **b** two CyTOF data sets and **c** four scRNA-seq data sets. The color coding indicates the true labels of subpopulation according to the source of each data set. Within a predefined distance, tighter scatters associate with higher similarity among cells. The circles and dashed lines denote better delineation between clusters post-FHpre

distances. As a result, the changes of inter- or intra- cluster distances would have more influence on UMAP than on t-SNE layouts. In the UMAP layouts of benchmark data sets, especially MNIST, Kolod and Zeisel, the clusters form tighter groups and the distances between clusters are increased after applying FHpre (Supplementary Fig. S8). These results demonstrate that FHpre increases the inter-cluster distances while its disturbance to intra-cluster distances is minimal.

In addition to visual comparison, we quantitatively compare the clustering accuracy for data set with various preprocessing. PhenoGraph is employed as the basic method for clustering because of its wide usage. Considering that PCA and ICA are commonly used to preprocess computer vision and scRNA-seq data sets, they are also introduced to compare with FHpre, dFHpre and SHpre. PCA and ICA are not usually applied in CyTOF data sets because they are considered as dimension-reduction methods, and the dimension of CyTOF data set is lower than that of computer vision and scRNA-seq data sets. First, we employ PhenoGraph on original data sets as well as data sets processed with FHpre, dFHpre, SHpre, PCA and ICA, generating six sets of predicted clustering labels. We then calculate *F*-measures between true labels and the six sets of predicted labels (Fig. 3e**,** representative visual comparisons displayed in Supplementary Fig. S9). In computer vision and scRNA-seq data sets, FHpre or dFHpre are of higher *F*-measure than others preprocessing, demonstrating improved clustering accuracy of our proposed framework. For a more comprehensive and in-depth comparison, *F*-measures of the four benchmark scRNA-seq data sets using PCA (left panel) and ICA (right panel) with principle component numbers ranging from 50 to 300 is displayed in Supplementary Fig. S10. The *F*-measures of PCA and ICA preprocessing with optimized principle component numbers remain inferior to FHpre. However, FHpre and dFHpre show slight or even no significant improvements on CyTOF data sets. That is, a bimodal curve appears between intra- and inter- cluster pairwise distances (Fig. 2), indicating more and smaller intra-cluster distances compared to other data sets. This distribution is more likely to appear in data sets with tighter clusters, which is beneficial to PhenoGraph clustering. The evaluation indexes of Accuracy and NMI on eight data sets (four scRNA-seq, two CyTOF and two computer vision data sets) with various preprocessing present similar performance with *F*-measure (Supplementary Fig. S11), which further verify the improvement of our framework for clustering.

Better Clustering Delineation of Our Proposed Framework

We employ two published CyTOF data sets (Horowitz et al. 2013; Mrdjen et al. 2018) and demonstrate the delineation ability of FHpre for clusters with functional diversities (Fig. 5 and Supplementary Fig. S12). The improvements of clustering efficacy from FHpre are examined using the CyTOF data set provided by Dunja Mrdjen and colleagues (Mrdjen et al. 2018). In this data set, cells are obtained from the central nervous system of 8-week-old C57BL/6 mice and characterized with a 43-parameter antibody panel. Processing and visualizing these cells with FHpre and t-SNE lead to four border-associated macrophage subsets and multiple dendritic cell subsets, which corroborates with published results. We find distinct distribution patterns of microglia subsets in standard t-SNE plot and FHpre t-SNE plot (Fig. 5a). We then obtain a detailed view of the phenotypic profile of microglia subsets (Fig. 5b). In the standard t-SNE layouts (Fig. 5b, upper panels), a group of subsets are crowded together with ambiguous delineation while a subset is distinctly separated from the group. The FHpre t-SNE layout reveals four microglia subsets based on differential expression of CD90 and CD172 (Fig. 5c), namely subset 1 ($CD90^{+}CD172^{+}$), subset 2 ($CD90^{+}CD172^{-}$), subset 3 ($CD90^{-}CD172^{+}$), and subset 4 ($CD90^{-}CD172^{-}$). As shown in Fig. 5b and c, the FHpre t-SNE layout reveals relatively clearer delineation of four microglia cell subsets than standard t-SNE layout because the elongated inter-microglia cell subset distance of FHpre t-SNE layout. However, for the standard t-SNE, the three subsets of microglia are mixed together tightly, and they are difficult to distinguish. The median expression values of all markers in different microglia subsets are summarized and compared to explore their detailed functional differences (Fig. 5d). FHpre can identify a heterogeneous subset 2, which exhibits distinct expression pattern (CCR2 + MerTK + Ter119 +). Together, these results show that microglia populations are heterogeneous and can be further divided into four subsets based on their distinct expression with the application of FHpre, which cannot be excavated in standard t-SNE visualization.

The performance of FHpre is also verified on another CyTOF data set comprising measurements of 36 markers for 20 PBMC samples with varying serology for cytomegalovirus (CMV). Both standard t-SNE and FHpre t-SNE reveal that cells from both CMV + and CMV − samples are well-mixed across most regions of the t-SNE plot (Supplementary Fig. S12a). Graphically, FHpre better delineates clinical associated clusters with functional diversity in the t-SNE plot. In summary, FHpre assists the visual analysis of CyTOF data sets by providing clear delineations of subpopulations with functional diversities.

Discussion

With increasing multi-parametricity of modern single-cell techniques, such as CyTOF and scRNA-seq, so does potentially uninformative and confounding variables overcrowd

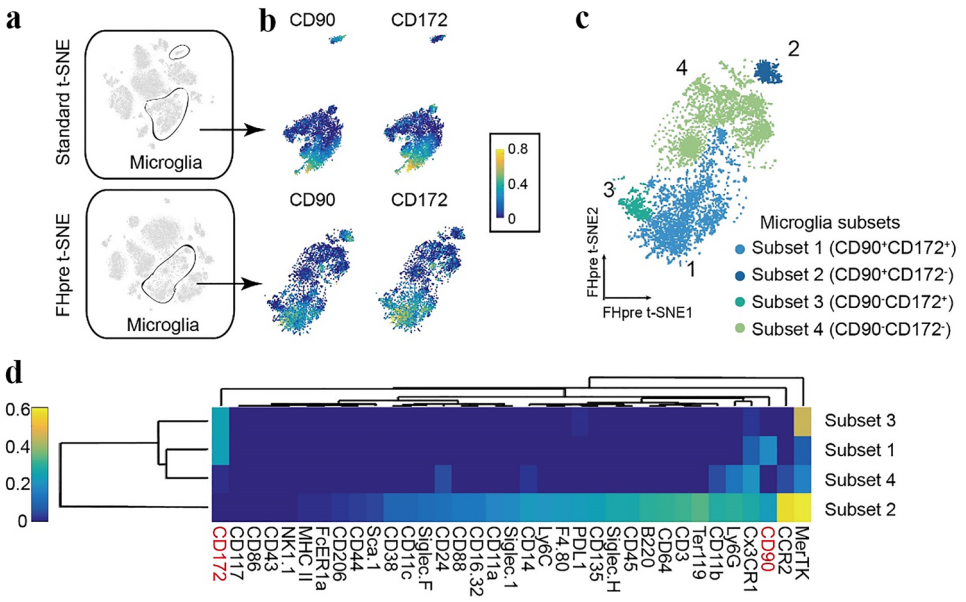

Fig. 5 FHpre better delineates CyTOF data subsets. **a** Standard t-SNE and FHpre t-SNE visualization of central nervous system data set. Microglia subsets (CD45loCD-11bloF4/80 + CD64 + MeTK + Cx3CR1hiSiglec-H + CD88lo) are manually annotated in the black circles on the t-SNE maps. **b** Magnified visualization of circled regions in (**a**). The color scale indicates the expression of selected markers. **c** The microglia subsets are well dispersed and clustered in FHpre t-SNE layout. Specifically, CD90 + CD172 + , CD90 + CD72 − , CD90 − CD172 + , and CD90 − CD172 − subsets are color annotated. **d** A clustergram summary of the median expression intensity of cell markers expressed by the four subsets identified in **c**. Color represents arcsin5-transformed marker expression. The subset-determining markers are colored with red. In contrast to other subsets, subset 2 exhibit distinct expression pattern (CCR2 + MerTK + Ter119 +), explaining the separation of subset 2 in the standard t-SNE layout in (**b**). Subsets 1, 3, and 4 are nearly identical with varying amounts of MerTK

neighborhoods, hindering distinct cluster delineation and impeding further clinical analysis (van Unen et al. 2017; Wang et al. 2017). Through FHpre and dFHpre, our proposed hashing-based framework provides new solutions to these limitations.

FHpre and dFHpre employ a newly developed LSH projection method that reduces the computational cost and further delineates dissimilar clusters. The merits of FHpre and dFHpre are examined with visual and quantitative results, which show its capability to further expand the horizon of computational tools for HD single-cell profile analysis. In addition to directly learning the structure of HD profiles, our pipeline manipulates the structure and aggrandizes the dissimilarity between clusters with proper mathematical lemmas (Abnousi et al. 2018).

FHpre and dFHpre efficiently improve the neighborhood clustering of input vectors with large dimensions, either around 50 (CyTOF) or more than 1000 (scRNA-seq). Our results reveal that dFHpre shows slightly better performance in preserving local structure and equivalent capability of anti-interference against high noise levels than FHpre. Yet, both FHpre and dFHpre are superior to SHpre and non-preprocessing in improving cluster delineation and enhancing PhenoGraph clustering accuracy. Compared FHpre with dFHpre, FHpre performs better than dFHpre on four scRNA-seq data sets, while dFHpre shows slightly higher F-measure than FHpre on two CyTOF data sets in improving the clustering accuracy on the basis of PhenoGraph. FHpre, dFHpre and SHpre are all compatible with extant k-NN-based clustering methods. Impressively, FHpre and dFHpre show unexpected performances in preprocessing the Samusik data, which did not show straightforward improvements for PhenoGraph. This discrepancy may be attributed to the argument about the former gold standard of manually gated labels. Many researchers have been re-considering the labels as 'biased' (Aghaeepour et al. 2017; Li et al. 2017), however, other researchers insist on the merits of labels manually gated by experienced experts (Mrdjen et al. 2018). Under this circumstance, many researchers rely on other judgments such as identifying rare subpopulation (van Unen et al. 2017), or achieving new subpopulations with high quality (Samusik et al. 2016). We agree on the performance of expert-gated labels; however, the lack of experienced experts also hinders the wider application of expert-gated labels. Here, we substantiate the ability of our framework by confirming published conclusions and unveiling cogent inferences that were previously concealed by insufficient preprocess. Especially when clinical information is accessible, our framework is validated by associations with the

granted ground truth, including the clinical labels such as CMV + .

Conclusion

In this study, we propose a hashing-based framework to improve the delineation of cell clusters and demonstrate the ability of our framework to achieve distinguishable clusters in a comprehensive collection of data sets, both simulated and experimental. Our framework can be applicable to a variety of highly multi-parametric data sources, such as medical imaging. Furthermore, our framework provides a general pattern for HD single-cell analysis, and other state-of-art preprocessing and clustering methods can be seamlessly integrated into our framework. With these corroborations, we suggest a wider application of the proposed framework as a routine procedure for analyzing HD single-cell profiles.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43657-022-00056-z.

Authors' Contributions XL, TZ, ZT, and XD conceived the study and experiments. XL and ZT performed data analysis and conceived the FHpre and dFHpre algorithm. XD, EC and TZ performed the biological analysis and interpretation. XL, TZ, ZT, SL and ARW wrote the manuscript with input from all the authors. All the authors discussed the results and commented on the manuscript.

Funding This work was supported by grants from the National Natural Science Foundation of China (Grant No. 81871448), Shanghai Municipal Science and Technology Project (Grant No. 2017SHZDZX01, 18430760500); Innovation Research Plan of the Shanghai Municipal Education Commission (Grant No. ZXWF082101), and National Key Research and Development Program of China (Grant No. 2017YFC0107603).

Material and Data Availability All used data are available from the corresponding original article, which we have detailed descripted in Supplementary Note.

Code Availability All the procedures are implemented with Matlab® 2020a, and are freely available at https://github.com/Lxc417/hashing_based_framework. Python version of the FHpre, dFHpre, and SHpre codes are also provided for the users' choice.

Declarations

Conflicts of Interest The authors declare that they have no competing interests.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

References

Abnousi A, Broschat SL, Kalyanaraman A (2018) Alignment-free clustering of large data sets of unannotated protein conserved regions using minhashing. BMC Bioinformatics 19(1):1–18. https://doi.org/10.1186/s12859-018-2080-y

Aghaeepour N, Ganio EA, Mcilwain D, Tsai AS, Tingle M, Van Gassen S, Gaudilliere DK, Baca Q, McNeil L, Okada R (2017) An immune clock of human pregnancy. Sci Immunol 2(15):n2946. https://www.science.org/doi/10.1126/sciimmunol.aan2946

Anchang B, Hart T, Bendall SC, Qiu P, Bjornson Z, Linderman M, Nolan GP, Plevritis SK (2016) Visualization and cellular hierarchy inference of single-cell data using SPADE. Nat Protoc 11(7):1264–1279. https://doi.org/10.1038/nprot.2016.066

Becht E, Mcinnes L, Healy J, Dutertre CA, Kwok I, Lai GN, Ginhoux F, Newell EW (2019) Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 37(1):38–44. https://doi.org/10.1038/nbt.4314

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat-Theor M 3(1):1–27. https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

Charikar MS (2002) Similarity estimation techniques from rounding algorithms. In: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing. pp 380–388.

Chen Y, Chen S, Zhang X (2020) Using DenseFly algorithm for cell searching on massive scRNA-seq datasets. BMC Genomics 21(5):1–9. https://doi.org/10.1186/s12864-020-6651-8

Dasgupta S, Stevens CF, Navlakha S (2017) A neural algorithm for a fundamental computing problem. Science 358(6364):793–796. https://www.science.org/doi/10.1126/science.aam9868

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE T Pattern Anal PAMI-1(2):224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Denis A, Sumithra S, Yim AKY, Ruteja B, Jeffrey M (2018) Single-Cell RNA-seq uncovers a robust transcriptional response to morphine by Glia. Cell Rep 24(13):3619–3629. https://doi.org/10.1016/j.celrep.2018.08.080

Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, Pe'Er D, Nolan GP, Bendall SC (2013) Normalization of mass cytometry data with bead standards. Cytom Part a 83(5):483–494. https://doi.org/10.1002/cyto.a.22271

Good Z, Sarno J, Jager A, Samusik N, Aghaeepour N, Simonds EF, White L, Lacayo NJ, Fantl WJ, Fazio G (2018) Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. Nat Med 24(4):474–483. https://doi.org/10.1038/nm.4505

Horowitz A, Strauss-Albee DM, Leipold M, Kubo J, Nemat-Gorgani N, Dogan OC, Dekker CL, Mackey S, Ma Ec Ker H, Swan GE (2013) Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. Sci Transl Med 5(208):145r–208r. https://www.science.org/doi/10.1126/scitranslmed.3006702

Jin C, Lagoudas GK, Zhao C, Bullman S, Bhutkar A, Hu B, Ameh S, Sandel D, Liang XS, Mazzilli S (2019) Commensal microbiota promote lung cancer development via γδ T cells. Cell 176(5):998–1013. https://doi.org/10.1016/j.cell.2018.12.040

Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York, 344. https://doi.org/10.1002/9780470316801

Kleinsteuber K, Corleis B, Rashidi N, Nchinda N, Walker BD (2016) Standardization and quality control for high-dimensional mass cytometry studies of human samples. Cytom Part a 89(10):903–913. https://doi.org/10.1002/cyto.a.22935

Levine J, Simonds E, Bendall S, Davis K, Amir EA, Tadmor M, Litvin O, Fienberg H, Jager A, Zunder E (2015) Data-driven phenotypic

dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 162(1):184–197. https://doi.org/10.1016/j.cell.2015.05.047

Levine LS, Hiam-Galvez KJ, Marquez DM, Tenvooren I, Madden MZ, Contreras DC, Dahunsi DO, Irish JM, Oluwole OO, Rathmell JC (2021) Single-cell analysis by mass cytometry reveals metabolic states of early-activated CD8+ T cells during the primary immune response. Immunity 54(4):829–844. https://doi.org/10.1016/j.immuni.2021.02.018

Li H, Uri S, Stanton KP, Yao Y, Montgomery RR, Yuval K (2017) Gating mass cytometry data by deep learning. Bioinformatics 33(21):3423–3430. https://doi.org/10.1093/bioinformatics/btx448

Mrdjen D, Pavlovic A, Hartmann FJ, Schreiner B, Utz SG, Leung BP, Lelios I, Heppner FL, Kipnis J, Merkler D (2018) High-dimensional single-cell mapping of central nervous system immune cells reveals distinct myeloid subsets in health, aging, and disease. Immunity 48(2):380–395. https://doi.org/10.1016/j.immuni.2018.01.011

Quintelier K, Couckuyt A, Emmaneel A, Aerts J, Saeys Y, Van Gassen S (2021) Analyzing high-dimensional cytometry data using FlowSOM. Nat Protoc 16(8):3775–3801. https://doi.org/10.1038/s41596-021-00550-0

Reid AJ, Talman AM, Bennett HM, Gomes AR, Lawniczak MK (2018) Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. Elife Sciences 7:e33105. https://doi.org/10.7554/eLife.33105.001

Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326. https://www.science.org/doi/10.1126/science.290.5500.2323

Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP (2016) Automated mapping of phenotype space with single-cell data. Nat Methods 13(6):493–496. https://doi.org/10.1038/nmeth.3863

Schuyler RP, Jackson C, Garcia-Perez JE, Baxter RM, Hsieh E (2019) Minimizing batch effects in mass cytometry data. Front Immunol 10:2367. https://doi.org/10.3389/fimmu.2019.02367

Shekhar K, Brodin P, Davis MM, Chakraborty AK (2014) Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). Proc Natl Acad Sci 111(1):202–207. https://doi.org/10.1073/pnas.1321405111

Spitzer MH, Carmi Y, Reticker-Flynn NE, Kwek SS, Engleman EG (2017) Systemic immunity is required for effective cancer immunotherapy. Cell 168(3):487–502. https://doi.org/10.1016/j.cell.2016.12.022

Tang C, Cao L, Zheng X, Wang M (2018) Gene selection for microarray data classification via subspace learning and manifold regularization. Med Biol Eng Comput 56(7):1271–1284. https://doi.org/10.1007/s11517-017-1751-6

Tu MM, Lee F, Jones RT, Kimball AK, Saravia E, Graziano RF, Coleman B, Menard K, Yan J, Michaud E (2019) Targeting DDR2 enhances tumor response to anti–PD-1 immunotherapy. Sci Adv 5(2):v2437. https://www.science.org/doi/10.1126/sciadv.aav2437

Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, Hjerling-Leffler J, Haeggström J, Kharchenko O, Kharchenko PV (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci 18(1):145–153. https://doi.org/10.1038/nn.3881

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(11): 2579–2605

Van Unen V, Li N, Molendijk I, Temurhan M, Hollt T, Der Meulen- V, De Jong AE, Verspaget HW, Mearin ML, Mulder CJ, Van Bergen J (2016) Mass cytometry of the human mucosal immune system identifies tissue- and disease-associated immune subsets. Immunity 44(5):1227–1239. https://doi.org/10.1016/j.immuni.2016.04.014

van Unen V, Höllt T, Pezzotti N, Li N, Reinders MJ, Eisemann E, Koning F, Vilanova A, Lelieveldt BP (2017) Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. Nat Commun 8(1):1–10. https://doi.org/10.1038/s41467-017-01689-9

Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 14(4):414–416. https://doi.org/10.1038/nmeth.4207

Witt E, Benjamin S, Svetec N, Zhao L (2019) Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. Elife 8:e47138. https://elifesciences.org/articles/47138

Yagnik J, Strelow D, Ross DA, Lin R (2011) The power of comparative reasoning. In: 2011 International Conference on Computer Vision. IEEE. pp 2431–2438.

Yue L, Rong J, Deng C, Yan S, Li X (2013) Compressed Hashing. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp 446–451.