Characterizing tissue composition through combined analysis of single-cell morphologies and transcriptional states

Feng Bao[1,*], Yue Deng[2,*], Sen Wan[3], Bo Wang[2], Qionghai Dai[3,#], Steven J. Altschuler[1,#], Lani F. Wu[1,#]

[1] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA.

[2] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China.

[3] Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing, China.

[*] These authors contributed equally to this work.

[#] Corresponding authors: L.F.W (Lani.Wu@ucsf.edu), S.J.A (Steven.Altschuler@ucsf.edu) and Q.D (qhdai@tsinghua.edu.cn)

Abstract

Advances in spatial transcriptomics technologies enable optical profiling of morphological and transcriptional modalities from the same cells within tissues. Here, we present multi-modal structured embedding (MUSE), an approach to deeply characterize tissue heterogeneity through analysis of combined image and transcriptional single-cell measurements. We demonstrate that MUSE can discover cellular subpopulations missed by either modality as well as compensate for modality-specific noise. MUSE identified biologically meaningful cellular subpopulations and stereotyped spatial patterning within heterogeneous mouse cortex brain tissues, profiled by seqFISH+ or STARmap technologies. MUSE provides a framework for combining multi-modal single-cell data to reveal deeper insights into the states, functions and organization of cells in complex biological tissues.

Introduction

Living tissues are built from ensembles of cells in different states. Microscopy provides a classical approach to identify and characterize cell types through similarities in morphology[1-3]. Developments in single-cell transcriptomics technologies provide complementary approaches to characterize cells types through similarities in transcriptional states[4-8]. Both microscopy and single-cell transcriptomics approaches have provided deep insights into cellular state, function and organization. Recent advances in spatial transcriptomics technologies, such as spatial transcriptomics (ST)[9, 10], sequential fluorescence *in situ* hybridization (seqFISH)[11, 12], multiplexed error-robust fluorescence *in situ* hybridization (MERFISH)[13, 14] and spatially-resolved transcript amplicon readout mapping (STARmap)[15], enable simultaneous morphological and transcriptional profiling from the same single cells. Here, we explore the exciting possibility that techniques in machine learning can be used to combine information from microscopy and single-cell transcriptomics to provide deeper insights into cell-type compositions that comprise tissues.

Methods that successfully combine multi-modal information hold the promise to identify biologically meaningful subpopulations that are missed by individual modalities and provide a more detailed description of tissue cell heterogeneity (**Fig. 1a**). However, effective approaches that combine multi-modal data need to overcome several challenges. Notable are requirements that: (**requirement 1**) discriminative information in each modality should be captured in the combined data structure; and (**requirement 2**) limited information from a lower-quality modality should improve—and not reduce—the resolution of the subpopulation structure that is learned from the higher-quality modality.

Here, we present a multi-modal structured embedding (MUSE) approach that addresses these requirements. MUSE uses a deep-learning architecture to extract and integrate information from each modality into a meaningful joint representation. A self-reconstruction loss ensures that information from each modality is not lost in the process of building the joint latent representation, and a self-supervised loss ensures that phenotypic similarity of cells in each modality is preserved in the joint representation. We demonstrate these ideas first using synthetic data with known ground truth. Then we apply MUSE to two single-cell spatial transcriptomics datasets of neural cortex[12, 15], which have the desirable properties for methods validation that many cell types have stereotyped morphologies, transcriptional biomarkers and positions within the tissue.

Results

MUSE architecture and training

MUSE is built on a standard multi-view autoencoder (AE) neural network architecture[16, 17] (**Fig. 1b**). Learning is conducted in three steps: 1) **modality-specific transformations:** the input features x and y are transformed into latent representations h_x and h_y; 2) **pseudo-label learning:** clustering on feature spaces h_x and h_y are performed independently to obtain pseudo-labels l_x and l_y for each modality; and 3) **joint feature learning**: the modality-specific features h_x and h_y are merged and transformed into a joint latent feature representation z. The learning process is guided by minimizing combined self-reconstruction and self-supervised loss functions. The self-reconstruction loss resembles the standard AE loss function, which encourages the learned joint feature representation (z) to faithfully retain information from the original individual input feature modalities (x and y). The self-supervised learning exploits triple-loss functions[18, 19] to encourage cells with the same cluster label (i.e. with the same pseudo label in either l_x or l_y) to remain close—and cells with different cluster labels to remain far apart—in the joint latent space. During model training, the transformation, pseudo-label learning and joint feature learning steps are iteratively performed. Model parameters in the whole neural network are jointly updated in each iteration (**Methods**). Finally, after model training, graph clustering is performed on the joint latent features (z) using PhenoGraph[20, 21] to identify latent subpopulations (**Methods**).

Combined analysis improves cell subpopulation identifications

To evaluate the performance of MUSE, we initially made use of simulated transcript and morphology data where ground truth subpopulation assignment for each cell and modality is known (**Methods** and **Supplementary Fig. 1**). As benchmarks, MUSE was compared with three existing approaches that enabled combining data: canonical correlation analysis (CCA)[22], multi-omics factor analysis v2 (MOFA+)[23, 24], and a multi-view autoencoder (AE). Results using a single modality were presented using principle component analysis (PCA) as reference. For each method, graph clustering was used to identify the underlying subpopulation structures, and the accuracy to correctly discover true cell subpopulations was quantified using the adjusted Rand index (ARI)[25].

We first used the simulated data to assess the ability of MUSE to capture discriminative information from each modality (**requirement 1 above**). How is performance affected as the ability to discriminate subpopulations in each modality decreases? We retained 10 ground truth subpopulations in the full multi-modal space and degraded the ability of both single modalities to resolve these subpopulations by randomly merging a different group of cell cluster assignments for each modality (**Methods**). Transcriptional data were simulated using a published single-cell RNA simulator[26, 27], and morphological features were simulated using a multi-layer neural network (**Methods**). As cluster numbers decreased, the factorization method MOFA+ maintained an accuracy level comparable to either single-modality approach while MUSE exceeded the single-modality benchmarks (**Figs. 1c**). Visualization of the latent space suggested the utility of the triplet-loss function: cells originating from the same subpopulation in either modality remained close and all true subpopulations remained well distinguished **(Supplementary Fig. 2**). How is performance affected as the number of ground-truth subpopulations increases? A potential advantage for multi-modal analysis is the ability discover more fine-grained population composition by combining heterogenous cellular properties; however, accuracies tend to decrease with increasing cellular heterogeneity. Here, we held the total number of cells constant but increased the number of subpopulations. As the number of clusters increased, CCA, AE and MOFA+ did not achieve higher ARIs than the single-modality benchmark methods; however, the guided combination of MUSE consistently outperformed single-modality methods (**Supplementary Fig. 3**).

We next assessed the performance of MUSE as data quality in one modality degrades (**requirement 2 above**). Two persistent problems in single-cell data are sequencing dropouts and noise in feature measurements[28, 29]. First, we varied dropout level for the transcript modality while leaving the simulation parameters for the morphology modality unchanged (**Methods**); as before, 10 ground-truth clusters were used. Morphology alone provided an average accuracy of ~0.6 ARI (**Fig. 1d**, horizontal dashed lines). As the dropout rate increased, the accuracy of all methods degraded, but only MUSE was able to retain the accuracy of morphology alone. Visualizing the results in latent space suggested that MUSE representations maintained a discernable subpopulation structure of 10 clusters (**Fig. 1e**). Second, we changed the noise level in both modalities, using additive Gaussian random noise with increasing variance (**Supplementary Fig. 4**). MUSE largely performed better or equal to the benchmark single-modality methods, though at extremely high noise levels the performance of MUSE and all other combined methods were strongly compromised.

Finally, we note that the choice of latent dimension had only a minimal effect on the accuracy of subpopulation identification for all compared methods (**Supplementary Fig. 5)**; further, the accuracy of MUSE was not affected by input dimension (**Supplementary Fig. 6**). While MUSE is reasonably fast for current experimental data sizes (e.g. 1000 samples in 1.5 minutes on a standard desktop, **Methods**), MUSE is slower than all other compared methods due to the inclusion of graph clustering during structured self-supervised training, (**Supplementary Fig. 7**). All simulation parameters used in experiments were summarized in **Supplementary Table 1**.

Together, these results indicated that direct combination of two feature modalities does not guarantee a better cell-subpopulation decomposition. The structured self-supervised used by MUSE enabled capturing and combining discriminative information that was not available from either modality alone. Further, MUSE was not unduly confounded by poor data quality in either one or both modalities. Thus, MUSE satisfied the two requirements we set *a priori* for a combined multi-modal method.

MUSE analysis of mouse cortex layers (seqFISH+)

A critical challenge in assessing single-cell analysis on real data is often a lack of ground truth. However, tissues with stereotyped spatial organization of cell types can provide independent evidence to evaluate the quality of learned representations and identified subpopulations[30, 31]. A particularly good example of this is the multi-layer pattern of spatial organization in the brain cortex[32, 33]. Thus, we applied MUSE to two experimental mouse cortex datasets.

The first cortex dataset was obtained using seqFISH+ technology[12]. This dataset includes expression profiles of 10,000 genes and cell images with DAPI and Nissl staining for 523 cells. For the transcript modality, we used a standard preprocessing pipeline for scRNA and selected highly variable genes as input features x (**Methods**). For the morphological modality, we input the DAPI and Nissl images for each cell (based on the provided cell masks) into a pre-trained deep neural network (Google Inception-v3[34]) to extract morphological properties as input features y (**Methods**). We extended subpopulation analyses to include four approaches in each of three classes—using only transcriptional features x (PCA, ZIFA, SIMILR and scScope; with detailed descriptions in **Methods**), only morphological features y (PCA, MDS, Isomap and tSNE), or the combination of both x and y (CCA, MOFA+, AE and MUSE). As before, cell clusters and cluster numbers were identified automatically by performing graph clustering on the latent cell representations z (**Methods**).

Many clusters identified from each method were spatially co-localized (**Methods**; **Supplementary Figs. 8** and **9**) and showed layer-like structure (**Figs. 2a** and **b**, **Supplementary Fig. 10**). Layer-specific markers were used to assign clusters to the cortex layers, and MUSE was able to identify all four layers (L2/3, L4, L5 and L6; **Figs. 2c, d** and **Supplementary Figs. 11, 12**). By examining MUSE clusters within the same cortex layers, we observed that the morphology modality refined distinctions provided by the transcript modality (e.g. in L2/3, distinct distributions of morphological features between the subpopulations were evident; **Fig. 2e** and **Supplementary Fig. 13**). Combing morphological and transcriptional profiles provided a refined dissection of cell diversity within the cortex.

MUSE analysis of mouse cortex layers (STARmap)

The second cortex dataset was obtained using STARmap technology. For the transcript modality, this dataset contained expression profiles of 1,020 genes; however, for the morphological modality, only cell shape masks were provided. The data were processed for the different comparison methods to obtain latent representations and subpopulations as in the previous cortex dataset (**Supplementary Fig. 14**).

We visualized the ability of different methods to "discover" cortical layer structure based on pseudo-colored cortex depth (**Fig. 3a**). SIMLR and MUSE identified the highest number of spatially co-localized clusters (**Fig. 3b**, **Supplementary Figs. 15** and **16**), and the clusters from MUSE were well separated in latent space (**Figs. 3a** and **c**). Based on anatomic annotations from the original paper, which labeled all seven layers in the cortex sample, MUSE successfully identified all neuron and non-neuron layers (**Fig. 3d**; **Methods**).

Analysis of multi-modal clusters identified by MUSE

As a case study, we analyzed STARmap clusters identified from individual (based on PCA) or combined (based on MUSE) modalities in the joint latent space provided by MUSE (**Fig. 4a**). We classified clusters based on whether MUSE 1) refined, 2) reproduced or 3) discovered new clusters compared to those obtained from single modal analyses (**Supplementary Figs. 17** and **18**).

The **"refined"** MUSE clusters were poorly separated based on transcript features, yet were reasonably well separated based on morphology features (**Fig. 4b**, and **Supplementary Fig. 19**). In the combined analysis, MUSE employed the morphological diversity to further dissect cells into subgroups. Cell masks provided in this dataset are shown sampled randomly from each cluster, and morphological differences (e.g. small cell sizes of Cluster 11) can be seen (**Fig. 4b**, bottom). The **"reproduced"** MUSE clusters were distinct based on transcript features alone (**Fig. 4c**, top panel). Differential expression analysis allowed us to annotate these clusters as astrocytes, hippocampus neurons, oligodendrocytes or smooth muscle cell (SMC) types (**Fig. 4c**, bottom panel), which have distinct and identifiable transcriptional expression patterns. The **"discovered"** MUSE clusters were missed from the single modalities, which individually provided only weak differences (**Fig. 4d**, top panel). Here, the combination of weak heterogeneities from both modalities enabled MUSE to identify distinct L2/3, L5 and L6 structures (**Fig. 4d**, right bottom panel).

Discussion

Characterizing cell heterogeneity is fundamental to understanding how tissues are organized and function. Two widely used and well-validated methods to study cell diversity are microscopy, to capture morphological differences, and scRNA sequencing, to capture transcriptional differences. Here, we developed a deep-learning framework, MUSE, to combine observations from both single-cell morphology and transcript modalities. MUSE makes use of a learning architecture that encourages synthesis of subpopulation structure observed in either modality. We demonstrated, for both synthetic and real biological data, that MUSE can reveal novel subpopulation structure and tissue organization missed by single-modalities or other methods.

It is evident that combined analysis across -omics modalities will increase our power to understand tissue heterogeneity[35, 36]. The machine learning approach of MUSE – with its parallelized autoencoder architecture and self-supervised learning approach – is designed to be extensible across modalities. MUSE is posed to leverage and coalesce advances as new measurement modalities, deeper profiling approaches, and modality-specific analysis are developed.

Methods

Multi-modal structured embedding

MUSE learns joint latent features by incorporating heterogeneity of morphological and transcriptional modalities. For a single-cell spatial transcriptomics dataset with n cells, transcriptional and morphological profiles are represented as $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, where the ith row of each matrix is the transcriptional (x_i) or morphological (y_i) feature from the same cell i.

Zero-inflated multi-modal autoencoder

The whole autoencoder structure is illustrated in **Supplementary Fig. 20**. Features from two modalities (x_i and y_i) are input into a multi-modal autoencoder, and a latent representation for each modality is learned by the encoder layer:

$$h_{x_i} = f_{\text{Encoder}_x}(x_i)$$

$$h_{y_i} = f_{\text{Encoder}_y}(y_i)$$

where $f_{\text{Encoder}_x}(\cdot)$, $f_{\text{Encoder}_y}(\cdot)$ are multi-layer neural networks for two modalities and $h_{x_i}, h_{y_i} \in \mathbb{R}^m$ are latent representations with the same low dimension extracted from high-dimensional original inputs. The activation function in the last layer of the two encoders is chosen as $\tanh(\cdot)$ to ensure the same scale for the two representations. Then the initial joint representation z_i is learned by combing h_{x_i} and h_{y_i}:

$$z_i = f_{\text{Encoder}_z}\left[\text{concat}\left(h_{x_i}, h_{y_i}\right)\right]$$

where $\text{concat}(\cdot)$ function concatenates two latent representations into one vector and the neural network encoder $f_{\text{Encoder}_z}(\cdot)$ further encodes the vector into a joint representation $z_i \in \mathbb{R}^k$. The joint representation z_i will be optimized by structured self-supervised loss.

Next, we use sparse weight matrices w_x and w_y to selectively activate entries in z_i for the reconstruction of the original features x_i and y_i:

$$z_i^{(x)} = z_i w_x$$

$$z_i^{(y)} = z_i w_y$$

where $w_x, w_y \in \mathbb{R}^{k\times k}$ are sparse matrices that only employ a subset of the entries in z_i to generate modality-specific features $z_i^{(x)}$ and $z_i^{(y)}$. Finally, features for each modality are reconstructed by decoders:

$$\hat{x}_i = f_{\mathrm{Decoder}_x}\left(z_i^{(x)}\right)$$

$$\hat{y}_i = f_{\mathrm{Decoder}_y}\left(z_i^{(y)}\right)$$

where $f_{\mathrm{Decoder}_x}(\cdot)$ and $f_{\mathrm{Decoder}_y}(\cdot)$ are multi-layer neural networks that expand latent representations into reconstructed features $\hat{x}_i$ and $\hat{y}_i$.

Self-reconstruction loss

For the transcriptional modality, dropout is a major limitation due to the challenges of tracking fluorescent spots across multiple imaging rounds. Therefore, transcript profiles from *in situ* sequencing usually include a large proportion of zeros. Here we use a zero-inflated reconstruction error for the transcript modality to remove the effects of zeros entries:

$$L_{\mathrm{reconstruct}_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{\|\mathrm{sign}(x_i)\circ(x_i-\hat{x}_i)\|_2}{\sum \mathrm{sign}(x_i)}$$

where $\mathrm{sign}(\cdot)$ is a sign function that returns either 0 or 1 based on whether x_i is zero or not (respectively); $\circ$ is the Hadamard product that conducts element-wise product of two vectors. The loss function calculates reconstruction errors for non-zero expression then averages them over non-zero entry numbers ($\sum \mathrm{sign}(x_i)$). *For the morphological modality,* we used the standard reconstruction loss:

$$L_{\mathrm{reconstruct}_y} = \frac{1}{n}\sum_{i=1}^{n}\|y_i-\hat{y}_i\|_2/q$$

The overall reconstruction loss is the combination of the two modality losses with the sparsity constraint:

$$L_{\mathrm{reconstruct}} = L_{\mathrm{reconstruct}_x} + L_{\mathrm{reconstruct}_y} + \lambda_{\mathrm{regularization}}\left(\|w_x\|_2 + \left\|w_y\right\|_2\right)$$

where $\lambda_{\mathrm{regularization}}$ is the regularization hyperparameter and is determined through analysis on simulation data (**Supplementary Fig. 21a**).

Structured self-supervised loss.

Feature properties, noise levels and distributions, dropout levels are intrinsically different between the image and transcript modalities. To extract useful information from each modality and increase the quality of the joint latent feature z_i, we further used structured self-supervised learning to encourage the structure of each modality to be maintained in the joint latent space.

To identify modality-specific population structure, clustering is performed for cell i using the latent feature from each modality (h_{x_i}, h_{y_i}) independently. Here, PhenoGraph[20] was used to identify sample structures. The optimal cluster number is determined automatically on sample graph structures, an approach that is widely used in single-cell analysis. Here, cluster labels for cell i with respect to modality features x_i and y_i are denoted by l_{x_i} and l_{y_i}, respectively. Cells with the same labels are similar to each other in (at least) one modality. Clusters from each modality are used as supervising labels to improve the learning of joint latent feature z_i via the triplet loss:

$$L_{\text{triplet}_x} = \frac{1}{n}\sum_{i=1}^{n} \max\left(\left\|z_i - z_{\text{pos}_x}\right\|_2 - \left\|z_i - z_{\text{neg}_x}\right\|_2 + \varepsilon, 0\right)$$

$$L_{\text{triplet}_y} = \frac{1}{n}\sum_{i=1}^{n} \max\left(\left\|z_i - z_{\text{pos}_y}\right\|_2 - \left\|z_i - z_{\text{neg}_y}\right\|_2 + \varepsilon, 0\right)$$

where in L_{triplet_x}, sample z_i is the anchor, z_{pos_x} is a positive sample from the same cluster as the anchor based on clusters from modality x, and z_{neg_x} is a negative sample from a different cluster; ε is the margin; L_{triplet_y} is defined in the same way using clusters from modality y. The triplet loss pushes the distance difference between anchor-positive and -negative samples to be greater than the margin so that the loss approaches its minimum (i.e. 0). As the choice of margin ε is hard to predetermine due to the uncertainty of feature distributions in two modalities, an adaptive method was used to automatically determine the margin value (refer to Optimization of MUSE).

Loss function

The overall loss function for training is the combination of the self-reconstruction and self-supervised losses:

$$L = L_{\text{reconstruct}} + \lambda_{\text{supervise}}\left(L_{\text{triplet}_x} + L_{\text{triplet}_y}\right)$$

where $\lambda_{\text{supervise}}$ is the hyperparameter to balance the contribution from triplet loss terms and was determined by simulation experiments (**Supplementary Fig. 21b**).

Optimization of MUSE

MUSE is trained on raw features and reference labels from two modality and optimizes joint latent features and cluster labels iteratively.

First, we obtain an estimate of the margin ε used in triplet loss. To accomplish this, we train the model without supervised terms by setting $\lambda_{\text{supervise}} = 0$, which is equivalent to a multi-modal autoencoder with zero-inflated loss in the transcript modality. We then estimate ε as the differences between medians in the top and bottom 20% values in the pairwise distance matrix from the initialized joint latent z_i.

Then, we optimize the whole MUSE model using iterative training (over the complete loss function):

1) Fixing the network parameters, update the cluster labels l_{x_i} and l_{y_i} by using clustering on h_{x_i}, h_{y_i} (see below).

2) Fixing cluster labels l_{x_i} and l_{y_i}, optimize the network parameters to obtain updated h_{x_i}, h_{y_i} and z_i.

Clustering

During training, clustering and labels for each independent modality were obtained using PhenoGraph[20] with the Louvain method, which determines optimal cluster number automatically. After optimization, the same procedure was used to obtain clusters and labels for the joint latent space. For all PhenoGraph analysis, we used the same default 30 nearest neighbors to construct graph. We note that the architecture of MUSE is flexible, and other (e.g. modality-specialized) clustering approaches can be used instead of PhenoGraph to provide cluster labels l_{x_i} and l_{y_i}.

Spatial transcriptomics data preprocessing

We made use of two cortex datasets from seqFISH+[12] and STARmap[15], respectively. The seqFISH+ data includes 523 cells from 5 fields of views in a mouse cortex. For each cell, 10,000 RNAs were profiled using *in situ* sequencing. DAPI and Nissl stains were used in imaging, and cells were segmented manually based on their morphology. The STARmap dataset mapped 973 cells for mouse visual cortex, and each cell has

1,020 gene measurements. Cells were identified using watershed segmentation and segmentation masks were provided.

Transcriptional analysis

For both datasets, we performed preprocessing based on gene count data. We selected the top 500 most variable genes (genes with zero counts for all cells were excluded by default). Gene counts were normalized by library size and transformed using log(1+x) before input into the tested models. Transcript analyses were performed using *scanpy* python package (version 1.4.4)[37].

Morphological analysis

For seqFISH+ data, we segmented single-cell images from tissue images for each imaging channel independently using the provided segmentation masks. Each cell-segmentation region was placed at the center of an empty image with 299×299 pixels. Then, DAPI and Nissl channels were independently input into the pretrained Inception v3 deep neural network. The output of the last network layer (with 1,024 dimensions) from each cell was concatenated into a long vector. PCA was applied to compress these feature vectors to 500-dimensional feature vectors. Vectors were scaled to have same mean value on all cells as in transcript features and then were used as input to the tested models. For STARmap data (which lacked cell markers), we directly placed the provided cell-segmentation masks over blank images then input them into the same pretrained neural network as for seqFISH+. Outputs from the last layers were also compressed using PCA and scaled to obtain single-cell morphological features. For both implementations, we used the Inception-v3 network[34] with pretrained parameters provided by TensorFlow Hub (https://tfhub.dev/).

Simulation experiment setup

We generated simulated ground-truth class labels $l \in \{1, \dots, L\}^n$ for n cells and L possible cluster (i.e., cell subpopulation) assignments (see **Supplementary Table 1** for values of all parameters below). We simulated the situation for which only a proportion of true cluster identities could be observed from each modality, but all clusters could be discriminated using both modalities (**Supplementary Fig. 1**). To accomplish this, we divided the true clusters into two non-overlapping groups that were each assigned to one of the two modalities. Then, in each group, clusters were merged with probability p providing observed cluster labels l_x, l_y for the two modalities.

For example, 10 ground-truth clusters, labeled $\{1, \ldots, 10\}$, could be divided into groups $G_1 = \{1, \ldots, 5\}, G_1 = \{6, \ldots, 10\}$ with modality 1 considering merges from G_1 but not G_1, and *vice versa* for modality 2; after merging, modality one might have seven clusters formed clusters $\{1 \cdot 2 \cdot 3, 4 \cdot 5, 6, 7, 8, 9, 10\}$ while modality two might have six clusters formed from $\{1, 2, 3, 4, 5, 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10\}$ (where " $\cdot$ " indicates merged clusters). While each modality can only distinguish a subset of the clusters, the combination has the potential to distinguish all of them.

For the transcriptional modality, we followed the same scRNA-seq simulation framework as used in SIMILR[26] and scScope[27]. In short, we generated latent codes $z_i^{(x)} \in \mathbb{R}^m$ for cell i using a multivariable normal (MVN) distribution:

$$z_i^{(x)} \sim \sum_{k=1}^{K} \pi_{k,i} \mathrm{MVN}(\mu_k, \Sigma_k),$$

where K is the total cluster number; $\pi_{k,i} = 1$ if cell i was assigned to cluster k in l_x and otherwise 0; $\mu_k \in \mathbb{R}^m$ was sampled from a uniform distribution with $\Sigma_k \in \mathbb{R}^{m \times m}$ the identity matrix. Raw transcriptional features were generated through a linear transformation by $x_i^{raw} = A^{(x)} z_i^{(x)}$, where entries in the random projection matrix $A^{(x)} \in \mathbb{R}^{p \times m}$ were randomly sampled from the uniform distribution between $[-0.5, 0.5]$. Gaussian noise was added to features $x_i^{noise} = x_i^{raw} + \varepsilon$, where ε was sampled from a Gaussian distribution $N(0, \sigma^2)$. Next, dropout in the count matrix with dropout rate proportional to expression level was simulated as:

$$x_i = x_i^{noise} \delta\left[\exp\left(-\alpha x_i^{noise}\right) < \eta\right],$$

where $\delta[\cdot]$ is an indicator function that outputs 1 if the argument is true and otherwise 0; α is the decay coefficient that controls dropout levels (set by default to 0.5); η is a random value sampled from the uniform distribution between $[0, 1]$. We input x_i to all methods for analysis.

For the morphological modality, we generated latent codes $z_i^{(y)} \in \mathbb{R}^m$ using the same mixture model procedure as above with modality labels l_y. To add complexity to these "image-based" features, we passed these latent codes through a two-layer, non-linear network:

$$y_i^{(1)} = \mathrm{sigmoid}\left(A_1^{(y)} z_i^{(y)}\right),$$

$$y_i^{(2)} = \mathrm{sigmoid}\left(A_2^{(y)} y_i^{(1)}\right),$$

where $A_1^{(y)} \in \mathbb{R}^{q\times m}$ and $A_2^{(y)} \in \mathbb{R}^{q\times q}$ were matrices randomly sampled from the uniform distribution $[-0.5, 0.5]$; $\mathrm{sigmoid}(\cdot)$ is the sigmoid function to non-linearly transform the data. Finally, as above we added random noise and dropouts to $y_i^{(2)}$ to obtain final morphological features y_i. As a heuristic, the number of dropouts in this modality was set to 0.1 in order to obtain reasonably similar ARI scores for clustering based on each modality alone.

Analysis of gene expression data

Differential analysis of expression data

With cluster labels, we identified differentially expressed genes using fold changes and p-values. For each cluster, we compared within *vs*. across cluster gene expression of cells. Log2 Fold changes were calculated based on mean gene expressions of these two groups to reveal the average expression differences. P-values were derived from one-sided ranksum test on expression profiles between the two groups to measure overall expression distribution differences.

Annotation of cortex layers

For the seqFISH+ data, marker genes that identify different cortex layers were obtained from the literature[32] (in our case, four different genes were used to identify four different layers; **Supplementary Fig. 11**). Next, clusters with layer-like structures were identified (see below for score). Finally, for each layer-like cluster, the maximally overexpressed marker gene was used to assign each cluster to a layer.

For the STARmap data, anatomic layer labels were provided[15]. This allowed clusters to be annotated based on their spatial positions in the tissue section. First, clusters with significant spatial colocalization patterns (based on spatial colocalization score) were identified for annotation. Next, a 1-dimensional kernel density estimation (KDE) with Gaussian kernels was performed along the provided x-coordinate of the image (corresponding to the cortex axis) for each cluster to model the spatial density of cells in the tissue. Finally, clusters were assigned to anatomic layers where peaks of cell spatial densities were located. In our implementation, kernel density estimation was performed using the KernelDensity function from sklearn python library with bandwidth determined by Scott's rule.

Evaluation of identified subpopulations

Spatial co-localization score and evaluation

To quantify the spatial enrichment in the tissue for cell clusters, we designed a spatial co-localization score based on the statistic used in gene-set enrichment analysis (GSEA)[38].

For all cells, we first calculated the cell-cell distance matrix $D = \{d_{ij}\} \in \mathbb{R}^{n\times n}$, where d_{ij} is the Euclidean distance between cells i and j on the image. The distance matrix was further converted into similarity $R = \{r_{ij}\} \in \mathbb{R}^{n\times n}$ by taking $r_{ij} = 1/d_{ij}, i \neq j$. As the similarity matrix is symmetric, only one similarity score is used for each cell pair $(i < j)$. All off-diagonal, upper-triangle entries $(r_{ij}, i < j)$ in R were ordered into a list and re-indexed by rank $L = \{r_k\}$, where r_k is the similarity score in position k of L. If n is the total number of cells, then the size of L is given by $N = (n-1)(n-2)/2$.

We define two scores that allow us to assess whether a cluster label, C, is consistent with distance similarities. First, let $S_C \subset L$ be the set of similarity scores r_k obtained from cells within C and define:

$$P_C(S_C, k) = \frac{1}{N_{s_C}} \sum_{\substack{r_l \in S_C,\\ l\leq k}} r_l,$$

where $N_{s_C} = \sum_{r_k \in S_C} r_k$. Second, for $r_k \notin S_C$ (i.e., at least one cell is not in C), define:

$$P_{\neg C}(S_C, k) = \frac{1}{N - N_H} \sum_{\substack{r_l \notin S_C,\\ l\leq k}} 1,$$

where $N_C = (n_c - 1)(n_c - 2)/2$ is the size of S_C (n_c is the number of cells in C). The spatial co-localization score (SCS) for C is defined as the maximal signed deviation between distributions $P_C(S_C, \cdot)$ and $P_{\neg C}(S_C, \cdot)$.

To derive a significance p-value, we constructed null SCS distribution by permuting cluster labels and calculating corresponding scores 1000 times. The p-value is defined as the proportion of scores greater than SCS on non-permuted labels. Source code for the SCS calculation is provide on GitHub.

Cluster accuracy evaluation with adjusted Rand index

For simulation studies, where ground truth subpopulation labels were given, we evaluated clustering performance using the adjusted Rand index (ARI)[25]. An ARI near 1 indicates a strong match to ground truth clustering, while values near 0 suggest random assignment. In the implementation, we used the adjusted_rand_score function from sklearn.metrics python package.

Feature quality evaluation with Silhouette Coefficient

The quality of latent features were evaluated by the compactness of the clusters in the latent space using the Silhouette coefficient[39]. A score of 1 indicates highest density in latent space. In our implementation, we employed the silhouette_score function from sklearn.metrics python package.

Comparing methods

All compared methods were run on the same input features (see above data processing section; single-modal methods took features from only one modality) to learn 100-dimensional latent representations. Subpopulations were identified based on latent representations using PhenGraph[20]. All methods were configured with default parameters (unless specifically noted) and were run on the same Linux desktop (Ubuntu 18.04.3 LTS operation system) with Xeon E5 CPU and Nvidia Titan X GPU (Driver Version: 418.87.00, CUDA Version: 10.1).

The software packages used for comparisons are as follows.

Transcriptional feature learning methods: Principle component analysis (PCA): sklearn 0.20.3 python package. Zero-Inflated Factor Analysis (ZIFA)[40]: ZIFA.fitModel() from ZIFA python package (v0.1). Single-cell Interpretation via Multi-kernel LeaRning (SIMLR)[26]: python implementation of SIMLR (v0.1.3). scScope[27]: python implementation scScope (v0.1.5).

Morphological feature learning methods: PCA: as above. Multi-dimensional scaling (MDS), Isometric mapping (Isomap), and t-distributed stochastic neighbor embedding (tSNE): sklearn.manifold python library (v0.20.3). We note that the tSNE method can only support maximal 3-dimension latent representations.

Multi-modal feature learning methods:

Correlation analysis (CCA) learns linear transformations of multiview data and maximizes their correlations in latent spaces, and we chose the transformation of the transcript data for clustering: sklearn python package (v0.20.3). Multi-omics factor analysis v2 (MOFA+)[24] was designed to combine multi-omics data using the multiview matrix factorization: mofapy2 package (v0.3) with factors = 100, iteration = 500, group number = 1 and view number = 2 to learn 100-dimension joint features. Autoencoder (AE) learns joint representations based on reconstruction loss from two modal features: in the implementation, we used the same neural network structure as in MUSE with standard reconstruction loss with all learning parameters (learning step, iteration numbers, etc.) the same as used in MUSE. Multi-modal structured embedding (MUSE): software is implemented in python 3.7.3 with NumPy, SciPy, PhenoGraph and TensorFlow packages (details were provided at https://github.com/AltschulerWu-Lab/MUSE); hyperparameter values were chosen through simulation study (**Supplementary Fig. 21**) and are provided in **Supplementary Table 2**.

Software used in the study

Software packages use in the study can be accessed *via* following links

PhenoGraph: https://github.com/jacoblevine/PhenoGraph

ZIFA: https://github.com/epierson9/ZIFA

SIMLR: https://github.com/bowang87/SIMLR_PY

scScope: https://github.com/AltschulerWu-Lab/scScope

MOFA+: https://github.com/bioFAM/MOFA2

seaborn: https://seaborn.pydata.org/

sklearn: https://scikit-learn.org/

Data and code availability

Single-cell spatial transcriptomics datasets:

seqFISH+: transcript data were downloaded from the GitHub page of seqFISH+ project (date: August 1, 2019; link: https://github.com/CaiGroup/seqFISH-PLUS). Nissl and DAPI stained images were provided by authors of seqFISH+ paper.

STARmap: raw data were downloaded from the project page (https://www.starmapresources.com/data at July 2, 2019). Transcript profiles and cell segmentation masks were extracted from data using the python pipeline provided by authors at https://github.com/weallen/STARmap.

Simulated tool for multi-modality data generation:

Simulation code is available from GitHub https://github.com/AltschulerWu-Lab/MUSE.

MUSE:

MUSE is provided as a python package under MIT license and can be installed through "`pip install muse_sc`". Source and demonstration code are available on GitHub https://github.com/AltschulerWu-Lab/MUSE.

Acknowledgements

We thank Chee-Huat Linus Eng at Caltech for providing seqFISH+ image data, Xiao Wang at the Broad Institute and MIT for providing information on STARmap data analysis, and Olivier Moindrot at Stanford for the open-source implementation of the triplet loss. We thank Heinz Hammerlindl, Lee Rao, Susan Shen, Xiaoxiao Sun and other members of the Altschuler and Wu labs for constructive feedback. S.J.A. and L.F.W gratefully acknowledge support from the UCSF Program for Breakthrough Biomedical Research, ProjectALS and the CZI NDNC Challenge Network.

Conflict of interest

The authors declare no conflict of interest.

Figure legends

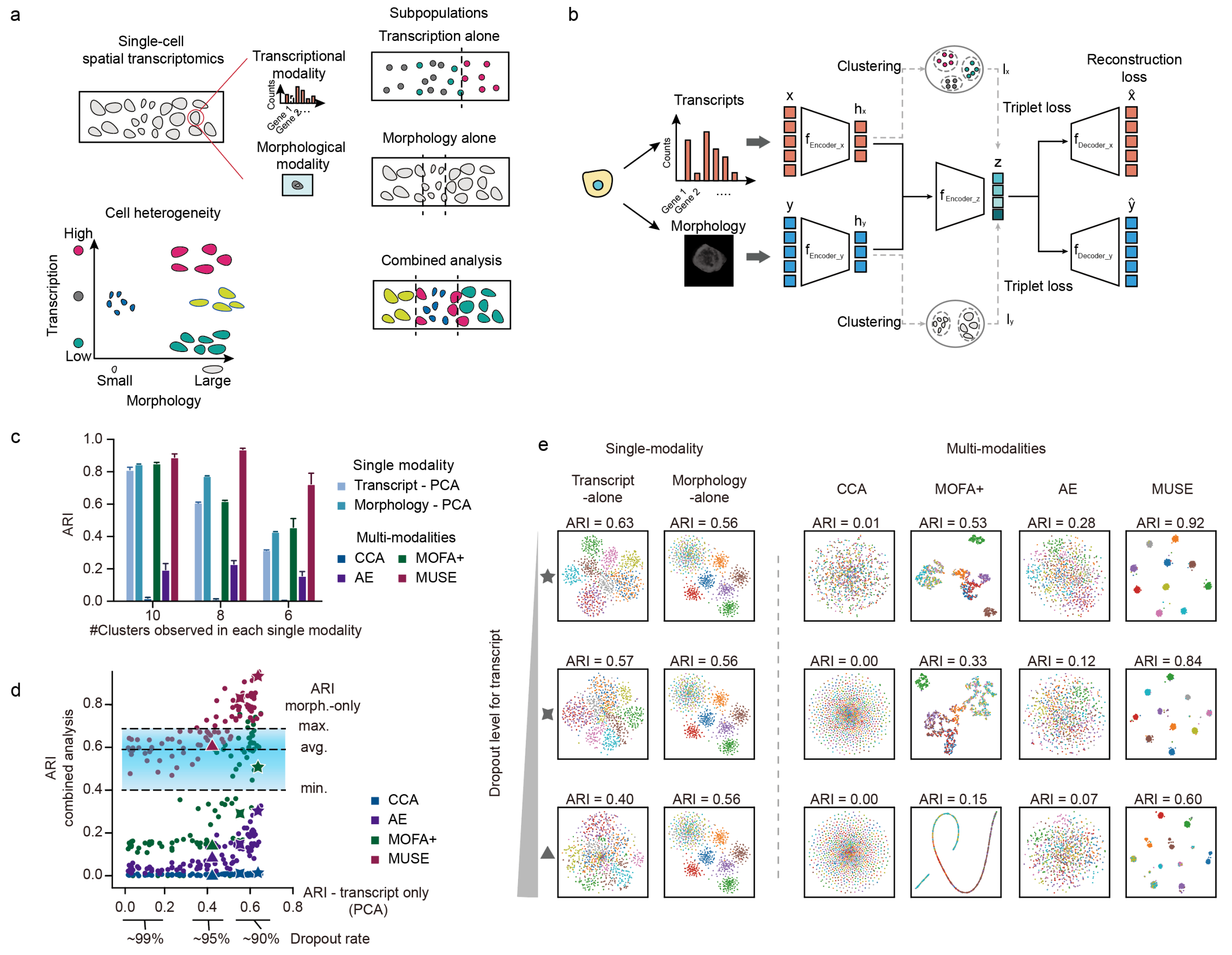

Figure 1 | Overview of MUSE and performance evaluation on simulated data. (**a**) Cartoon indicating how single-cell morphological and transcriptional data from a tissue (rectangular slide) can be combined to reveal high-resolution characterization of tissue heterogeneity. (**b**) Overview of MUSE architecture. MUSE combines features from transcripts (x) and morphology (y) into a joint latent representation z. The reconstruction and triplet losses encourage subpopulation structure from each modality to be faithfully maintained in z. (**c-e**) Performance evaluations using simulated data. (**c**) Accuracy of identifying ground-truth high-resolution subpopulations (k=10) from lower-resolution single-modality subpopulations (k=10, 8 or 6). 1,000 cells with transcriptional and morphological profiles were simulated. Cluster accuracy is quantified using the adjusted Rand index (ARI); mean+std is shown for 10 replicates. (**d**) Accuracy of identifying ground-truth clusters over a range of dropout levels from the transcriptional modality. Dashed lines: min, ave and max ARI of morphology-modality alone. x-axis: ARI of PCA analysis on transcript-modality alone. y-axis: ARI of combined-modality methods. 3-, 4-, 5-pointed shapes: comparison of results for randomly chosen datasets, also visualized in panel (**e**). (**e**) tSNE visualizations of latent representations from single- and combined-modality methods. Colors: ground truth subpopulation labels in simulation.

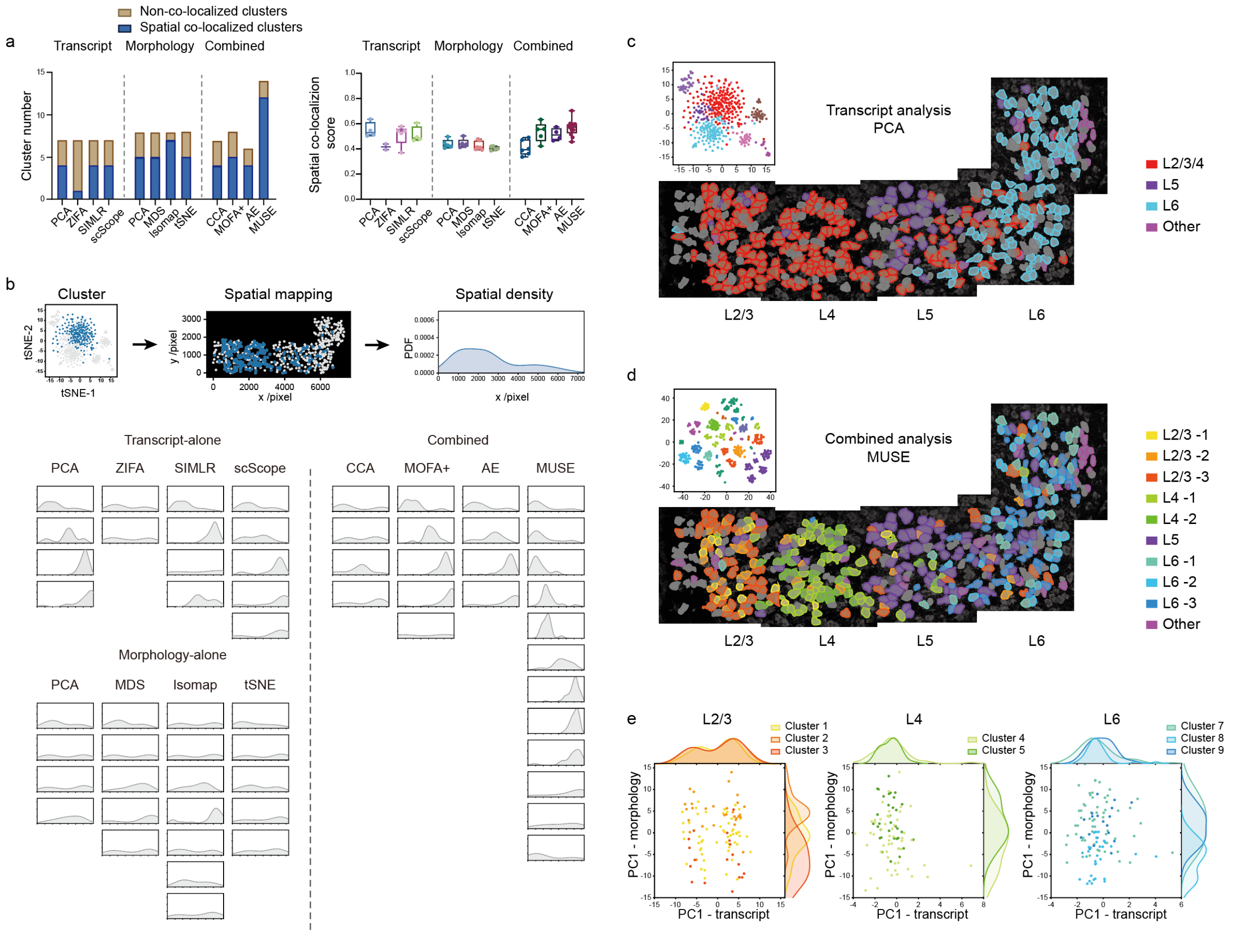

Figure 2 | Evaluation of MUSE on seqFISH+ mouse cortex data. Analysis of cells (n=523) based on transcript (top 500 variable genes) and/or morphology (DAPI and Nissl images) modalities. (**a**) Numbers (left) and scores (right) of clusters whose cells show spatial co-localization in the tissue (**Methods**). Right box plot: median (center line), interquartile range (box) and data range (whiskers). (**b**) Visualization of spatial density in the tissue section for clusters with co-localization patterns. Top: clusters were mapped to the tissue and spatial density was quantified by kernel density estimations. Bottom: spatial density plot for each cluster. Coordinates in each subfigure are the same as the density plot in the top. (**c-d**) Spatial mapping and tSNE plots of cell clusters with co-localization patterns based on (c) transcript-only analysis using PCA or (d) MUSE. Other transcript-only methods reveal similar spatial patterns (**Supplementary Fig. 10)**. Layers were annotated using marker genes identified from differential expression analysis (**Methods**). Cell colors: consistent with PCA or MUSE in (a) (respectively). (**e**) Visualization of clusters in the same layers identified by MUSE. Plots: first principle component (PC1) of raw features from each modality. Density graphs (top, right): Gaussian kernel density estimations.

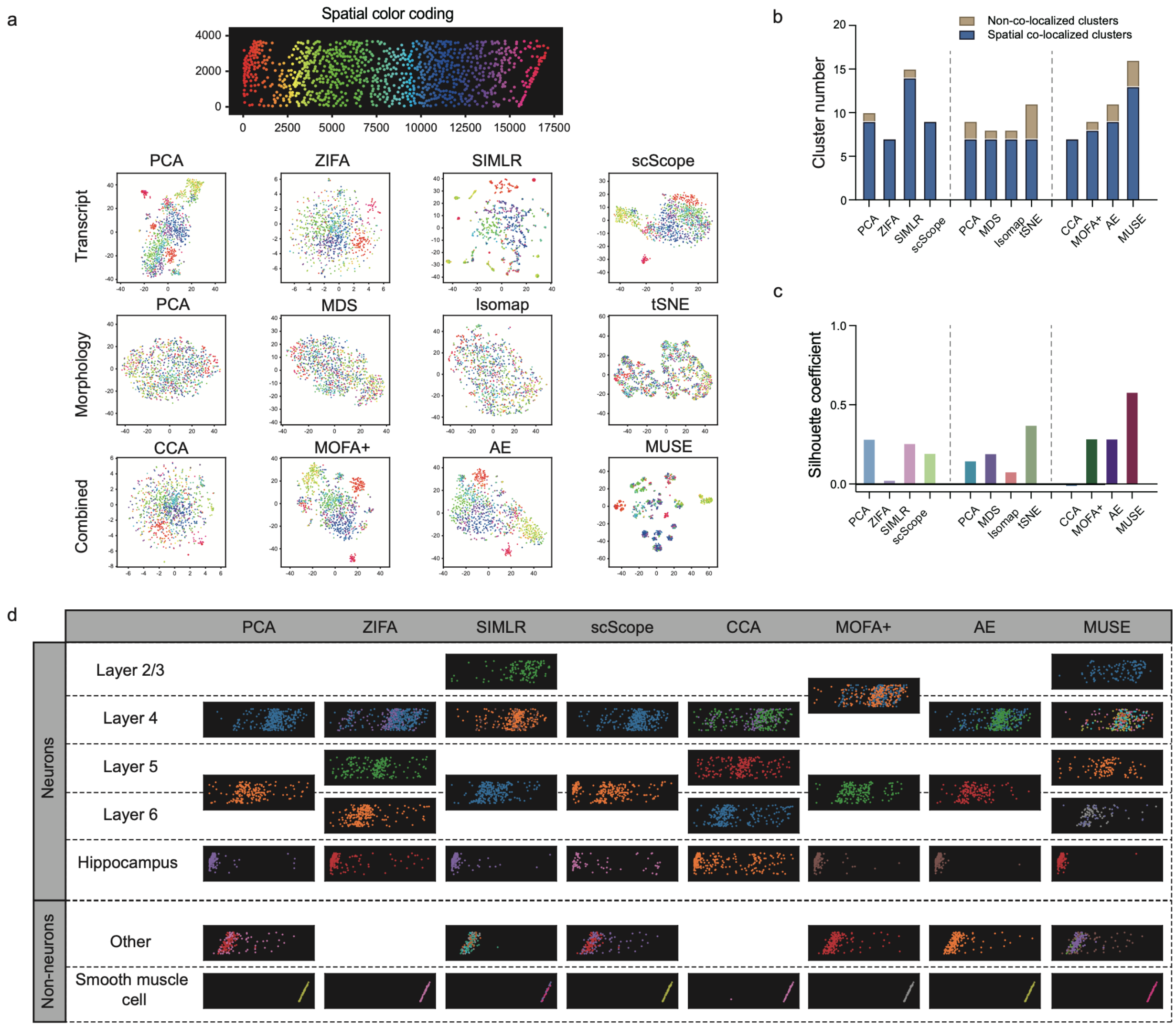

Figure 3 | Evaluation of MUSE on STARmap cortex data. Analysis of cells (n=972) based on transcript (top 500 variable genes) and/or morphology (segmentation mask) modalities. (**a**) tSNE visualization of latent representations by different methods with pseudo-colors labeling cortex depth along x-coordinate. (**b**) Numbers of identified clusters with or without significant spatial co-localization properties. (**c**) Feature quality evaluation by cluster compactness in latent space using Silhouette coefficient. (**d**) Spatial mapping and annotations of clusters with significant spatial co-localization patterns. Significantly co-localized clusters are identified using spatial co-localization score with permutation test. Clusters are assigned to one layer with respect to the anatomic annotations by original paper (**Methods**).

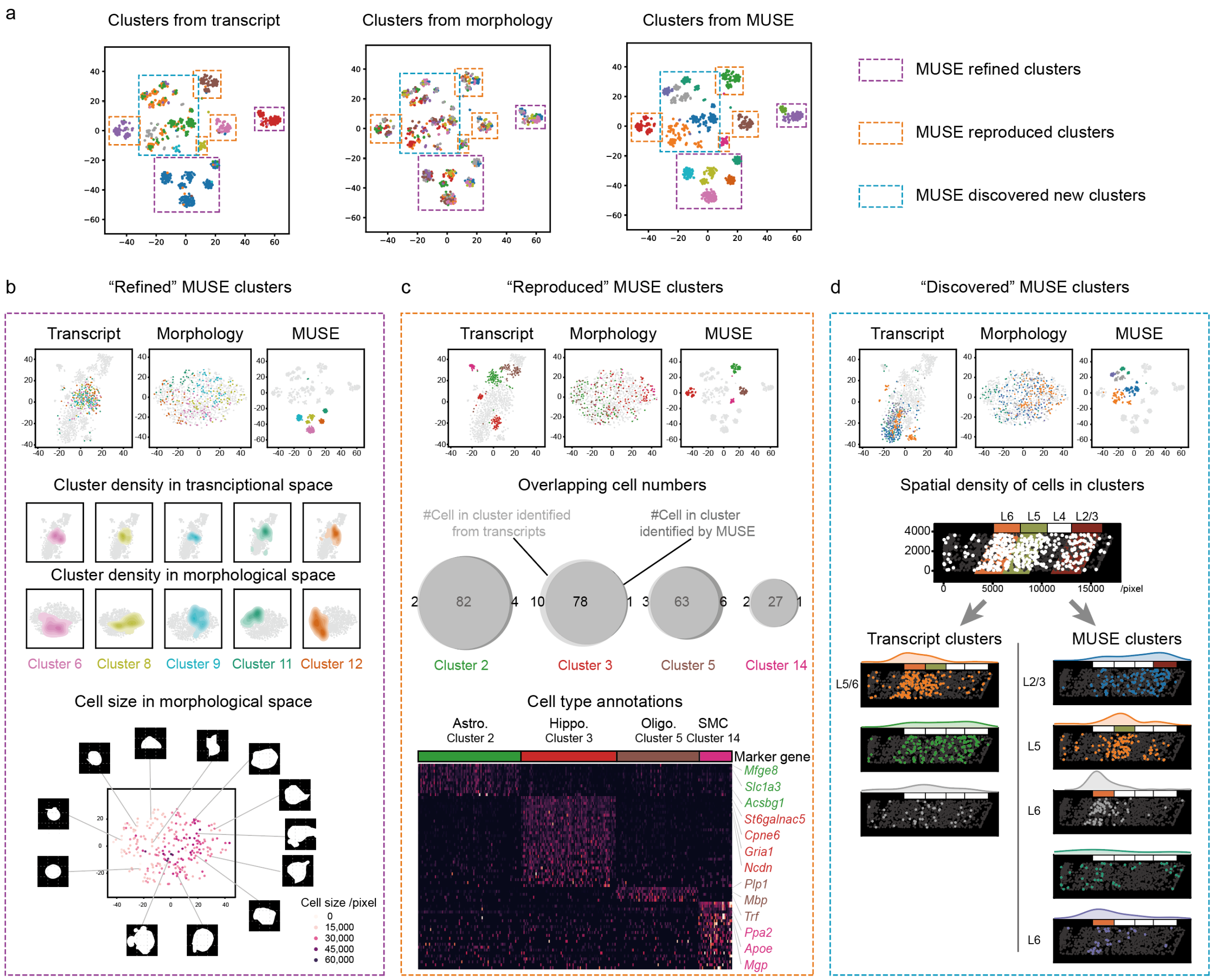

Figure 4 | Analysis of MUSE clusters. (**a**) MUSE-identified clusters are categorized based on whether they refined, reproduced or discovered clusters compared to the single-modality PCA-identified clusters. tSNE visualization using MUSE latent space: cluster labels from transcript-only (PCA, left), morphology-only (PCA, middle) or combined (MUSE, right) analyses. (**b**) "Refined" MUSE clusters. Top: tSNE visualization of latent spaces. Middle: density plots of 5 MUSE clusters in transcriptional and morphological spaces using 2D kernel density estimation. Bottom: topography of cell size in tSNE representation of morphological space. Color: cell sizes. Points: cells in the 5 MUSE clusters. Cell masks: randomly selected cells. (**c**) "Reproduced" MUSE clusters. Top: tSNE visualization of latent spaces. Middle: Venn diagrams: number of overlapping cells between PCA transcript-only (grey outline) and MUSE (black outline) clusters. Bottom: "Reproduced" MUSE clusters identify astrocyte (Astro.), hippocampus neurons (Hippo.), oligodendrocyte (Oligo.) and smooth muscle cell (SMC) types (**Methods**). (**d**) "Discovered" MUSE clusters.

Top: tSNE visualization of latent spaces. Bottom: “Discovered” MUSE clusters tend to be more layer specific than transcript-only clusters (density maps above tissue representations).

Reference

1. Perlman, Z.E. et al. Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194-1198 (2004).

2. Loo, L.-H., Wu, L.F. & Altschuler, S.J. Image-based multivariate profiling of drug responses from single cells. *Nature methods* **4**, 445-453 (2007).

3. Feldman, D. et al. Optical pooled screens in human cells. *Cell* **179**, 787-799. e717 (2019).

4. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).

5. Macosko, E.Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).

6. Shalek, A.K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369 (2014).

7. Rizvi, A.H. et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology* **35**, 551 (2017).

8. Gojo, J. et al. Single-cell RNA-seq reveals cellular hierarchies and impaired developmental trajectories in pediatric ependymoma. *Cancer cell* **38**, 44-59. e49 (2020).

9. Ståhl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).

10. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods* **16**, 987-990 (2019).

11. Shah, S. et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* **174**, 363-376. e316 (2018).

12. Eng, C.-H.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235-239 (2019).

13. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348** (2015).

14. Moffitt, J.R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* **113**, 11046-11051 (2016).

15. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** (2018).

16. Hinton, G.E. & Zemel, R.S. in Advances in neural information processing systems 3-10 (1994).

17. Baldi, P. in Proceedings of ICML workshop on unsupervised and transfer learning 37-49 (2012).

18. Chechik, G., Sharma, V., Shalit, U. & Bengio, S. Large Scale Online Learning of Image Similarity Through Ranking. *Journal of Machine Learning Research* **11** (2010).

19. Hoffer, E. & Ailon, N. in International Workshop on Similarity-Based Pattern Recognition 84-92 (Springer, 2015).

20. Levine, J.H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184-197 (2015).

21. De Meo, P., Ferrara, E., Fiumara, G. & Provetti, A. in 2011 11th international conference on intelligent systems design and applications 88-93 (IEEE, 2011).

22. Thompson, B. Canonical correlation analysis: Uses and interpretation. (Sage, 1984).

23. Argelaguet, R. et al. Multi‐Omics Factor Analysis—a framework for unsupervised integration of multi‐omics data sets. *Molecular systems biology* **14**, e8124 (2018).

24. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 1-17 (2020).

25. Hubert, L. & Arabie, P. Comparing partitions. *Journal of classification* **2**, 193-218 (1985).

26. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature methods* **14**, 414-416 (2017).

27. Deng, Y., Bao, F., Dai, Q., Wu, L.F. & Altschuler, S.J. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nature methods* **16**, 311-314 (2019).

28. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nature communications* **11**, 1-9 (2020).

29. Yuan, G.-C. et al. Challenges and emerging directions in single-cell analysis. *Genome biology* **18**, 84 (2017).

30. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).

31. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.-C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature biotechnology* **36**, 1183 (2018).

32. Belgard, T.G. et al. A transcriptomic atlas of mouse neocortical layers. *Neuron* **71**, 605-616 (2011).

33. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience* **19**, 335-346 (2016).

34. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in Proceedings of the IEEE conference on computer vision and pattern recognition 2818-2826 (2016).

35. Teichmann, S. & Efremova, M. Method of the Year 2019: single-cell multimodal omics. *Nat. Methods* **17** (2020).

36. Bock, C., Farlik, M. & Sheffield, N.C. Multi-omics of single cells: strategies and applications. *Trends in biotechnology* **34**, 605-608 (2016).

37. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

38. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).

39. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53-65 (1987).

40. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology* **16**, 1-10 (2015).