nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-022-01251-z

Integrative spatial analysis of cell morphologies and transcriptional states with MUSE

Feng Bao[1,8], Yue Deng[2,3,8], Sen Wan[4], Susan Q. Shen[1,5], Bo Wang[2], Qionghai Dai[4,6,7] ✉, Steven J. Altschuler[1] ✉ and Lani F. Wu[1] ✉

Spatial transcriptomics enables the simultaneous measurement of morphological features and transcriptional profiles of the same cells or regions in tissues. Here we present multi-modal structured embedding (MUSE), an approach to characterize cells and tissue regions by integrating morphological and spatially resolved transcriptional data. We demonstrate that MUSE can discover tissue subpopulations missed by either modality as well as compensate for modality-specific noise. We apply MUSE to diverse datasets containing spatial transcriptomics (seqFISH+, STARmap or Visium) and imaging (hematoxylin and eosin or fluorescence microscopy) modalities. MUSE identified biologically meaningful tissue subpopulations and stereotyped spatial patterning in healthy brain cortex and intestinal tissues. In diseased tissues, MUSE revealed gene biomarkers for proximity to tumor region and heterogeneity of amyloid precursor protein processing across Alzheimer brain regions. MUSE enables the integration of multi-modal data to provide insights into the states, functions and organization of cells in complex biological tissues.

Tissues are built from ensembles of cells in different states. Microscopy enables the identification and characterization of cell types through similarities in morphology[1–3]. Single-cell transcriptomics provides complementary approaches to characterize cell types through similarities in transcriptional states[4–8]. Both microscopy and single-cell transcriptomics approaches can elucidate cellular state, function and organization. Recent advances in spatial transcriptomics have combined the two approaches, allowing simultaneous morphological and transcriptional profiling from the same single cells or tissue regions, in methods such as Spatial Transcriptomics (ST or Visium)[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]. However, few generalizable methods for the integrated analysis of both morphological and transcriptomics data from the same cell exist. Here we show that deep learning techniques can be used to combine information from state-of-the-art spatial transcriptomics and microscopy technologies to provide insights into the organization, function and disease progression of heterogeneous tissues.

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

Here we present a MUSE approach that addresses these requirements (Fig. 1b). 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-supervision loss ensures that phenotypic similarity of samples in each modality is preserved in the joint representation. We first benchmark this approach using synthetic data with known ground truth and compare with known multi-modal approaches (Methods), including correlation-based method canonical correlation analysis (CCA)[16], matrix factorization-based method multi-omics factor analysis v2 (MOFA+)[17,18] and a multi-view autoencoder (AE). We then apply MUSE to a variety of datasets obtained using different spatial transcriptomics and imaging technologies. We use examples of profiled brain cortex, tumors, intestine and neurodegenerative disease progression studies to demonstrate how combined multi-modal analysis from MUSE can improve the dissection and interpretation of functional spatial heterogeneity (Extended Data Fig. 1a).

Results

MUSE architecture and training. MUSE is built on a standard multi-view AE neural network architecture[19,20] with the addition of a self-supervision loss function (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 is 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

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA. [2]School of Astronautics, Beihang University, Beijing, China. [3]Institute of Artificial Intelligence, Beihang University, Beijing, China. [4]Department of Automation and Institute for Brain and Cognitive Science, Tsinghua University, Beijing, China. [5]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA. [6]Beijing National Research Center for Information Science and Technology, Beijing, China. [7]Beijing Laboratory of Brain and Cognitive Intelligence and Beijing Key Laboratory of Multi-dimension & Multi-scale Computational Photography, Beijing, China. [8]These authors contributed equally: Feng Bao, Yue Deng. ✉e-mail: qhdai@tsinghua.edu.cn; steven.altschuler@ucsf.edu; lani.wu@ucsf.edu

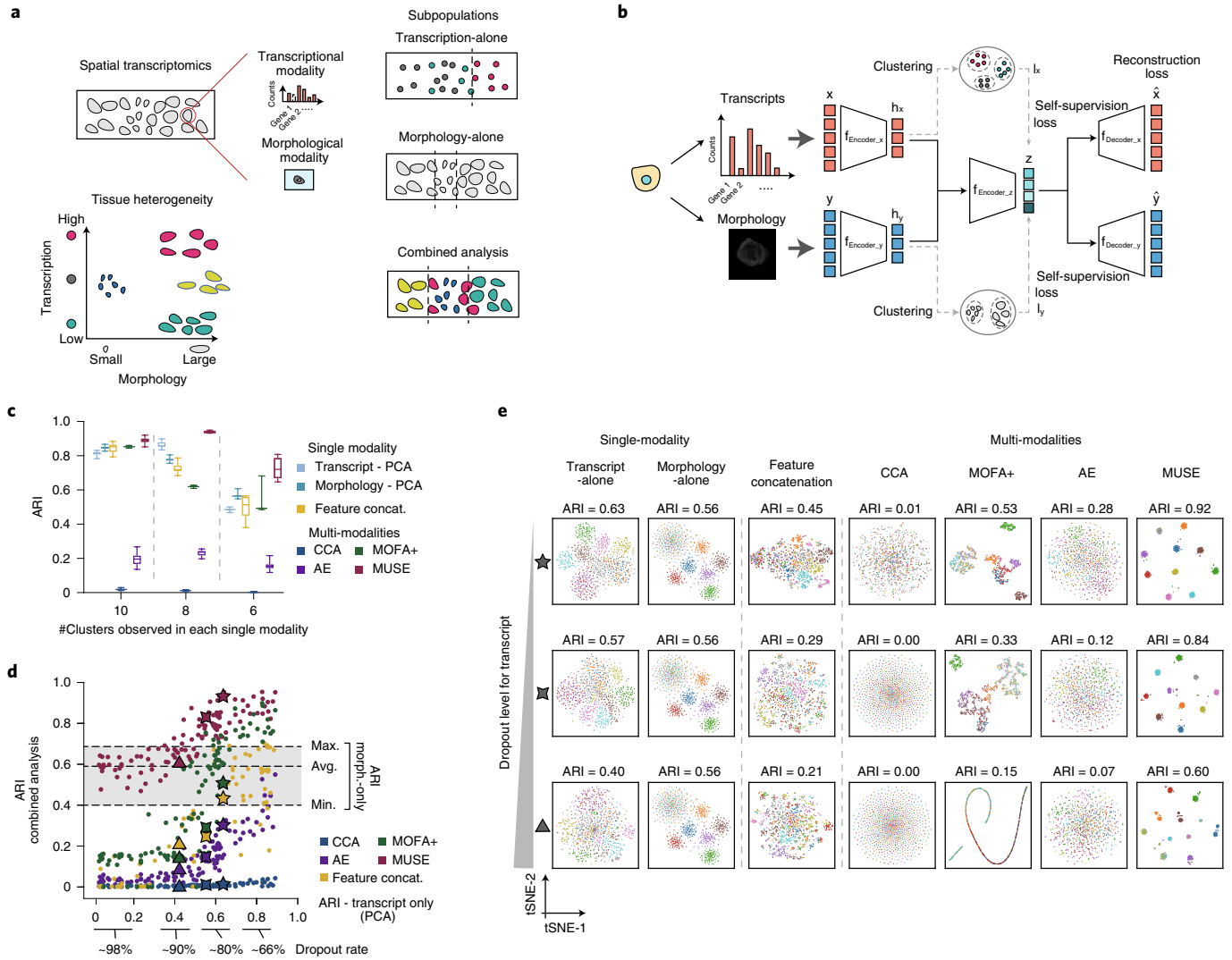


Fig. 1 | Overview of MUSE and performance evaluation on simulated data. a, Cartoon indicating how single-cell morphological and transcriptional data from a tissue (rectangular slide) can be combined to reveal high-resolution characterization of tissue heterogeneity. **b**, Overview of MUSE architecture. 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). In total, 1,000 cells with transcriptional and morphological profiles are simulated. Cluster accuracy is quantified using the ARI; box plot is based on $n=10$ replicates: median (center line), interquartile range (box) and data range (whiskers). **d**, Accuracy of identifying ground truth clusters over a range of dropout levels from the transcriptional modality. Dashed lines: minimum, average and maximum ARI of morphology modality alone. x axis: ARI of PCA on transcript modality alone. y axis: ARI of combined-modality methods. 3-, 4- and 5-pointed shapes: comparison of results for randomly chosen datasets, also visualized in **e**. **e**, tSNE visualizations of latent representations from single- and combined-modality methods. Colors: ground truth subpopulation labels in simulation.

combined self-reconstruction and self-supervision loss functions. The self-reconstruction loss (motivated by the standard AE loss function) encourages the learned joint feature representation (z) to faithfully retain information from the original individual input feature modalities (x and y). The self-supervision loss (using a triple-loss function[21,22]) encourages cells and tissue spots with the same cluster label in a single modality (that is, with the same pseudo-label in either l_x or l_y) to remain close—and those 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 (Extended Data Fig. 1b). Model parameters in the whole neural network are jointly updated in each iteration (Methods). Finally, after model training, the joint latent features (z) can be used in various tasks, such as clustering and trajectory inference. In this work, we focused on identifying latent subpopulations (Methods).

Combined analysis improves subpopulation identifications. To evaluate the performance of MUSE, we initially made use of simulated transcript and morphology data, in which the ground truth subpopulation assignment for each sample (cell or tissue spot) is known (Methods and Extended Data Fig. 1c). As benchmarks, MUSE was compared to several existing approaches that combine data (CCA, MOFA+ and AE) as well as simple concatenation of features from the two modalities as a baseline (Methods). CCA attempts to learn representations with maximal cross-modality correlation[16]. MOFA+ employs matrix factorization to decompose multi-view features into shared latent factors[17,18]. Finally, the AE

combines multi-modal data through a bottleneck layer that can be used to reconstruct original features. MUSE uses the same network architecture as a standard AE with the addition of a self-supervision loss. For benchmarking, single-modality feature spaces were reduced via principal component analysis (PCA) to match the latent space dimensions of the compared multi-modal approaches. Unless otherwise noted, graph clustering (PhenoGraph[23]) with default parameters was used to automatically identify subpopulation numbers, and the adjusted Rand index (ARI)[24] was used to assess accuracy of discovering true subpopulations.

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 ten 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 sample cluster assignments for each modality (Methods). Transcriptional data were simulated using a published single-cell RNA simulator[25,26], and morphological features were simulated using a multi-layer neural network (Extended Data Fig. 1c). As cluster numbers decreased, the factorization method MOFA+ and feature concatenation maintained an accuracy level similar to either single-modality approach, whereas MUSE exceeded the single-modality benchmarks (Fig. 1c). Visualization of the MUSE 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 (Extended Data Fig. 1d). How is performance affected as the number of ground truth subpopulations increases? A potential advantage for multi-modal analysis is the ability to discover more fine-grained population composition by combining heterogenous cellular and tissue properties. Here, we held the total number of cells constant and found that, with increased numbers of subpopulations, MUSE outperformed single-modality methods (Extended Data Fig. 1e).

We next assessed the performance of MUSE when data quality in one modality degrades (requirement 2 above). Two persistent problems in single-cell data are sequencing dropouts and noise in feature measurements[27,28]. First, we varied the level of transcript dropout while leaving the morphology modality unchanged (Methods); as before, ten ground truth clusters were used. Morphology-alone analyses provided an average accuracy of ~0.6 ARI (Fig. 1d, horizontal dashed lines). As expected, as the transcriptional signal degraded (Fig. 1d, from right to left on the x axis), the accuracy of all multi-modal methods dropped (Fig. 1d, y axis). However, the accuracy of MUSE only decreased to the range of accuracy estimated using morphology features alone (Fig. 1d, shaded range between 'min' and 'max'), suggesting that a lower-quality modality did not unduly harm MUSE's ability to use a higher-quality modality. Visualizing the results in latent space suggested that MUSE representations maintained a discernable subpopulation structure of ten clusters (Fig. 1e). Multi-modal methods that relied on maximizing multi-modal correlation (CCA), reconstruction accuracy without self-supervision (AE and MOFA+) or feature concatenation were strongly affected by data degradation in any one modality.

Second, we simultaneously changed the noise level in both the transcript and morphology modalities, using additive Gaussian random noise with increasing variance (Extended Data Fig. 1f). MUSE performed well for lower variance noise levels. However, for higher variance noise levels, the performance of MUSE and all compared multi-modal methods was strongly compromised. MUSE performed as expected in 'control' benchmark settings, including: not identifying subpopulations when single-cell transcript and image data were uncoupled across cells (Extended Data Fig. 1g); performing better than simple feature concatenation controls (Extended Data Fig. 1g); not over-clustering in comparison to simple superimposition of clusters obtained from single-modality analysis (Extended Data Fig. 1h); remaining robust to 'semi-simulated' data with real image features (Methods and Extended Data Fig. 1i); and not degrading when one modality was completely homogeneous (Extended Data Fig. 1j).

Finally, we surveyed the sensitivity of results to MUSE default settings. We found that the accuracy (ARI score) of MUSE was robust with respect to varying latent dimension (Extended Data Fig. 1k), input feature dimension (Extended Data Fig. 1l), latent dimension of single modality (Extended Data Fig. 1m) and the choice of post-learning clustering approaches (Extended Data Fig. 1n). In terms of runtime performance, MUSE is reasonably fast for current experimental data sizes (for example, 1,000 samples in 1.5 minutes on a benchmark desktop; Methods and Extended Data Fig. 1o). During training, graph clustering and iterative updating of cluster assignment increase MUSE runtime; as such, runtime can be accelerated by using other clustering methods (Extended Data Fig. 1p) or fixing cluster labels before network training, although at a loss of accuracy (Extended Data Fig. 1q). All simulation parameters used in experiments are summarized in Supplementary Table 1.

In summary, the synthetic data demonstrated that MUSE satisfied our two a priori requirements for a successful multi-modal method. Namely, the structured self-supervision approach used by MUSE facilitated capturing and combining discriminative information that was not available from either modality alone. Furthermore, MUSE was not unduly confounded by poor data quality in either one or both modalities.

MUSE analysis of mouse cortex layers from seqFISH+ data. Assessing inferred subpopulation structure on real data can be challenging due to 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[29,30]. A particularly good example of this is the brain cortex[31,32], whose multi-layer patterning provides orthogonal information to evaluate discovered subpopulations. 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 pre-processing pipeline for single-cell RNA 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 (ref. [33])) to extract morphological properties as input features y (Methods and Extended Data Fig. 2a). We extended subpopulation analyses to include four approaches in each of three classes, based on: (1) only transcriptional features x (PCA, ZIFA[34], SIMILR[25] and scScope[26], with detailed descriptions in Methods); (2) only morphological features y (PCA, multi-dimensional scaling (MDS), isometric mapping (Isomap) and t-distributed stochastic neighbor embedding (tSNE)); or (3) the combination of both x and y (CCA, MOFA+[18], AE and MUSE). As before, cell clusters and cluster numbers were identified automatically by performing Louvain clustering on the latent cell representations z (Methods).

MUSE identified a relatively large number of clusters that were spatially co-localized (Methods and Fig. 2a,b). Clusters were annotated using layer-specific markers (Extended Data Fig. 2b). Only MUSE identified clusters specific to each of the four layers (L2/3, L4, L5 and L6; Fig. 2c,d and Extended Data Fig. 2b,c). Neither subdividing transcript-alone clusters (Extended Data Fig. 2d) nor increasing cluster numbers from multi-modal embeddings (Extended Data Fig. 2e) provided better layer separation. In some cases, MUSE clusters within the same cortex layer could be seen to have different distributions of morphology (Fig. 2e) and/or matched to

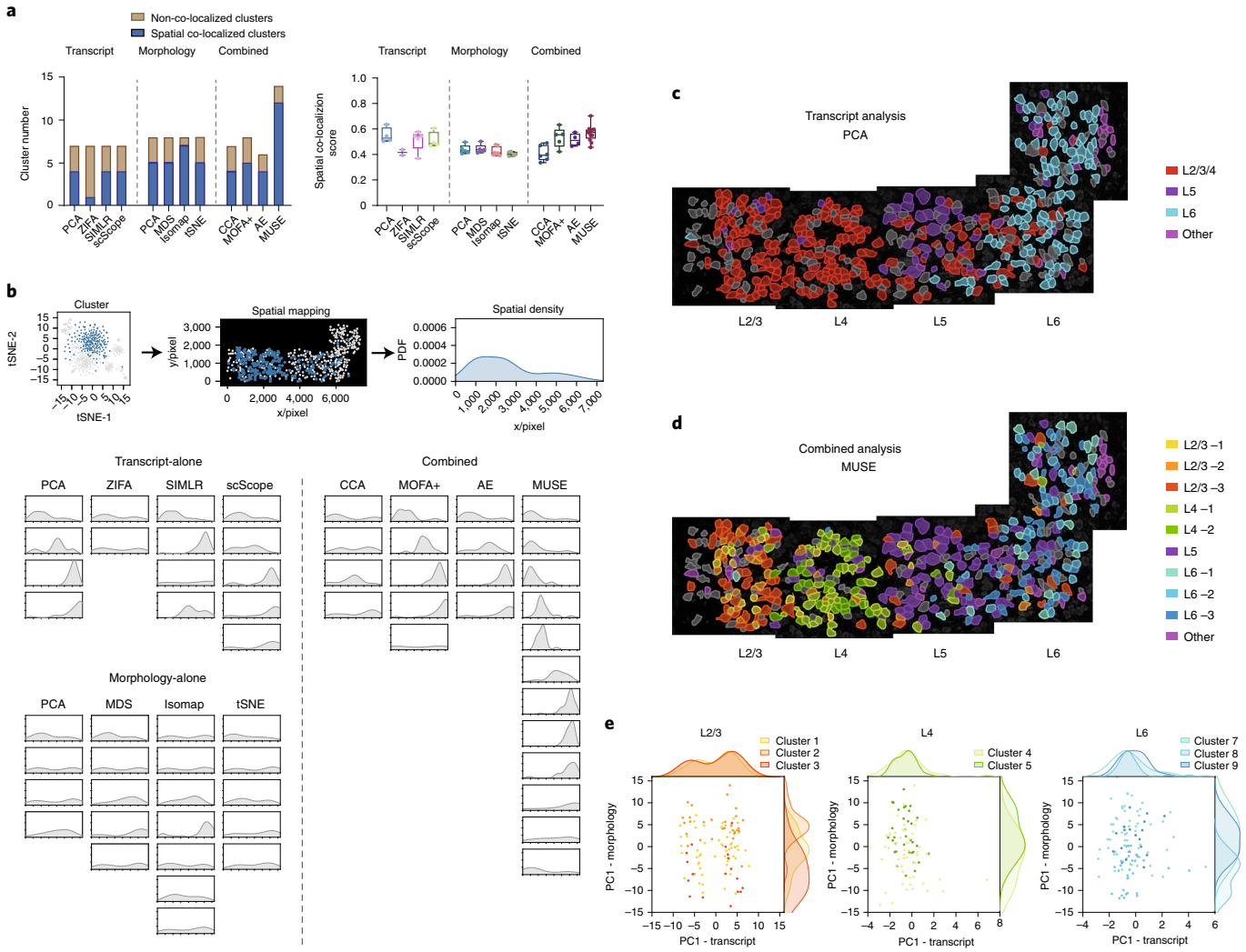


Fig. 2 | Evaluation of MUSE on seqFISH+ mouse cortex data. Analysis of cells ($n = 523$) based on transcript (top 500 variable genes) and/or morphology (DAPI and Nissl images) modalities. **a**, Numbers (left) and scores (right) of clusters whose cells show spatial co-localization in the tissue (Methods). 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 KDEs. 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 of cell clusters with co-localization patterns based on transcript-only analysis using PCA (**c**) or combined analysis using MUSE (**d**). Layers were annotated using marker genes identified from differential expression analysis (Methods). **e**, Visualization of clusters in the same layers identified by MUSE. Plots: PC1 of raw features from each modality. Density graphs (top, right): Gaussian KDEs.

different glutamatergic cell types reported in a recent single-cell RNA sequencing (scRNA-seq) study[35] (Extended Data Fig. 2f).

MUSE analysis of mouse cortex layers from STARmap data. The second cortex dataset was obtained using STARmap technology. For the transcript modality, this dataset contained 973 single-cell expression profiles of 1,020 genes; however, for the morphological modality, only cell shape masks were provided. The data were processed in the same manner as for the previous cortex dataset to obtain latent representations and subpopulations.

We visualized the ability of different methods to 'discover' cortical layer structure based on pseudo-colored cortex depth (Extended Data Fig. 3a). SIMLR and MUSE identified the highest number of spatially co-localized clusters (Extended Data Fig. 3b, top), with the clusters from MUSE well separated in latent space (Extended Data Fig. 3b, bottom). We note that the MUSE clusters were also robust to cluster parameter changes (Extended Data Fig. 3c).

Based on anatomic annotations from the original paper, which labeled all seven layers in the cortex sample, MUSE successfully identified all neuronal and non-neuronal layers (Extended Data Fig. 3d and Methods). 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. 3a). We classified clusters based on whether MUSE (1) refined, (2) reproduced or (3) discovered new clusters compared to those obtained from single-modal analyses (Extended Data Fig. 3e,f).

The 'refined' MUSE clusters were poorly separated based on transcript features yet were reasonably well separated based on morphology features (Fig. 3b). In the combined analysis, MUSE employed morphological diversity to further dissect cells into subgroups. The 'reproduced' MUSE clusters were distinct based on transcript features alone (Fig. 3c). Differential expression analysis allowed us to annotate these clusters as astrocyte (Astro.), hippocampal neuron (Hippo.), oligodendrocyte (Oligo.) or smooth

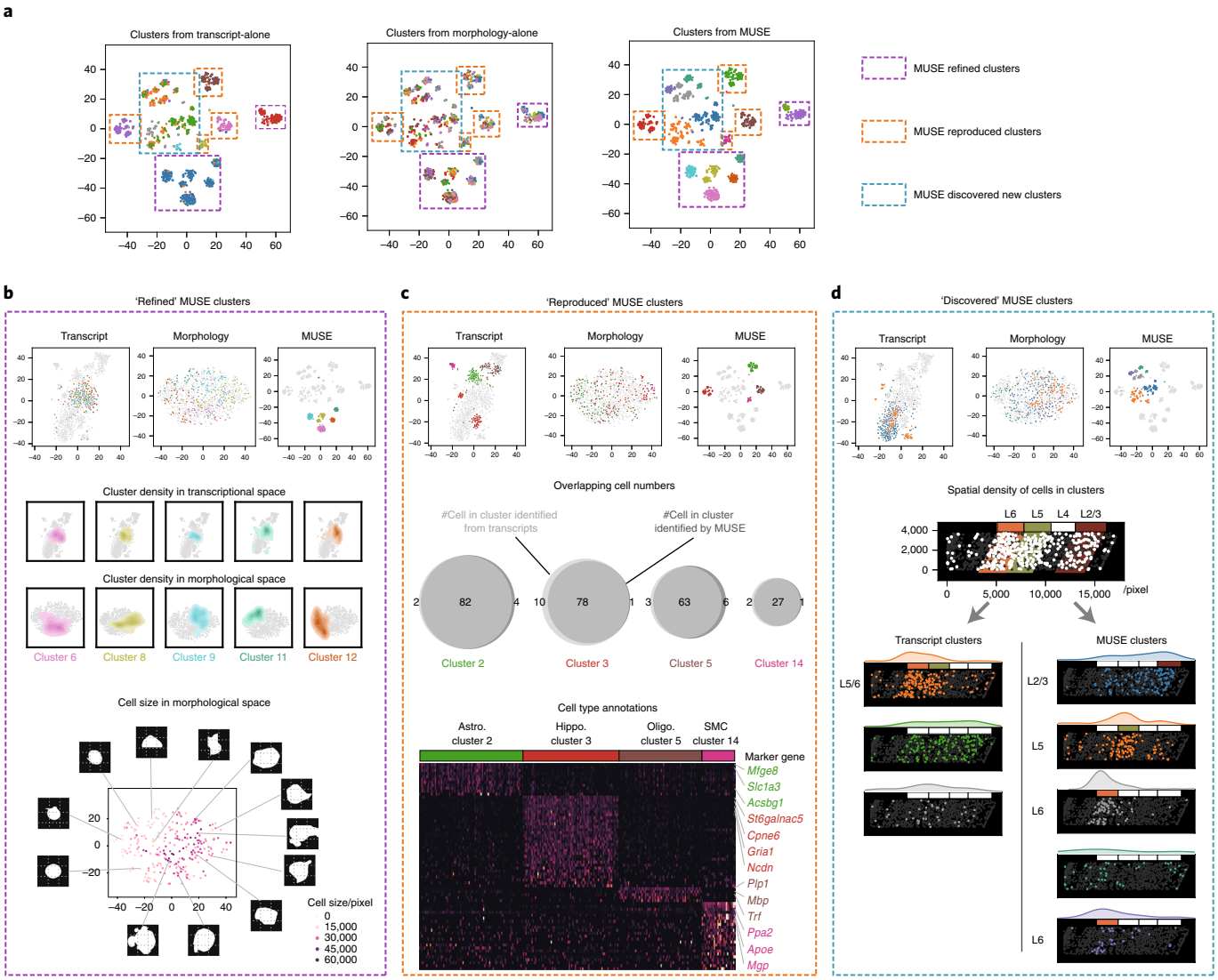


Fig. 3 | Analysis of MUSE clusters on STARmap cortex dataset. a, MUSE-identified clusters are categorized based on whether they refined, reproduced or discovered clusters compared to the single-modality PCA-identified clusters. tSNE visualization using MUSE latent space: cluster labels from transcript-alone (PCA, left), morphology-alone (PCA, middle) or combined (MUSE, right) analyses. **b**, 'Refined' MUSE clusters. Top: tSNE visualization of latent spaces. Middle: density plots of five MUSE clusters in transcriptional and morphological spaces using two-dimensional KDE. Bottom: topography of cell size in tSNE representation of morphological space. Color: cell sizes. Points: cells in the five 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 (gray outline) and MUSE (black outline) clusters. Bottom: 'Reproduced' MUSE clusters identify astrocyte (Astro.), hippocampus neurons (Hippo.), oligodendrocyte (Oligo.) and 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).

muscle (SMC) cells. The 'discovered' MUSE clusters were missed from either single modality, which individually provided only weak differences (Fig. 3d). Here, the combination of weak heterogeneities from both modalities enabled MUSE to identify distinct L2/3, L5 and L6 structures.

MUSE analysis of pancreatic ductal adenocarcinoma tissue from ST data. Genetic diversity and tumor microenvironment variation can greatly affect cancer progression, diagnosis and treatment[36–38]. Here, we made use of a pancreatic ductal adenocarcinoma (PDAC) dataset[39] collected by ST, which provides tissue-spot-based (rather than single-cell-based) reporting of transcriptional states. In this dataset, each ST spot contains multiple cells, which limits the resolution at which subpopulations can be discovered. For transcripts, 428 tissue spots were sequenced, and, for imaging, corresponding tissue sections were stained with hematoxylin and eosin (H&E). We took the top 500 variable genes as the input for the transcript modality, and we segmented the corresponding H&E image regions for ST spots to learn deep image embeddings (Methods and Extended Data Fig. 4a,b). To provide high-resolution references, the provided scRNA-seq data from the same tumor tissues were used (Extended Data Fig. 4c,d).

In the original study[39], transcript analysis of the regionally averaged ST data identified four regions, including one identified as cancer; analysis of high-resolution scRNA-seq data further revealed the presence of two cancer clones with the signature

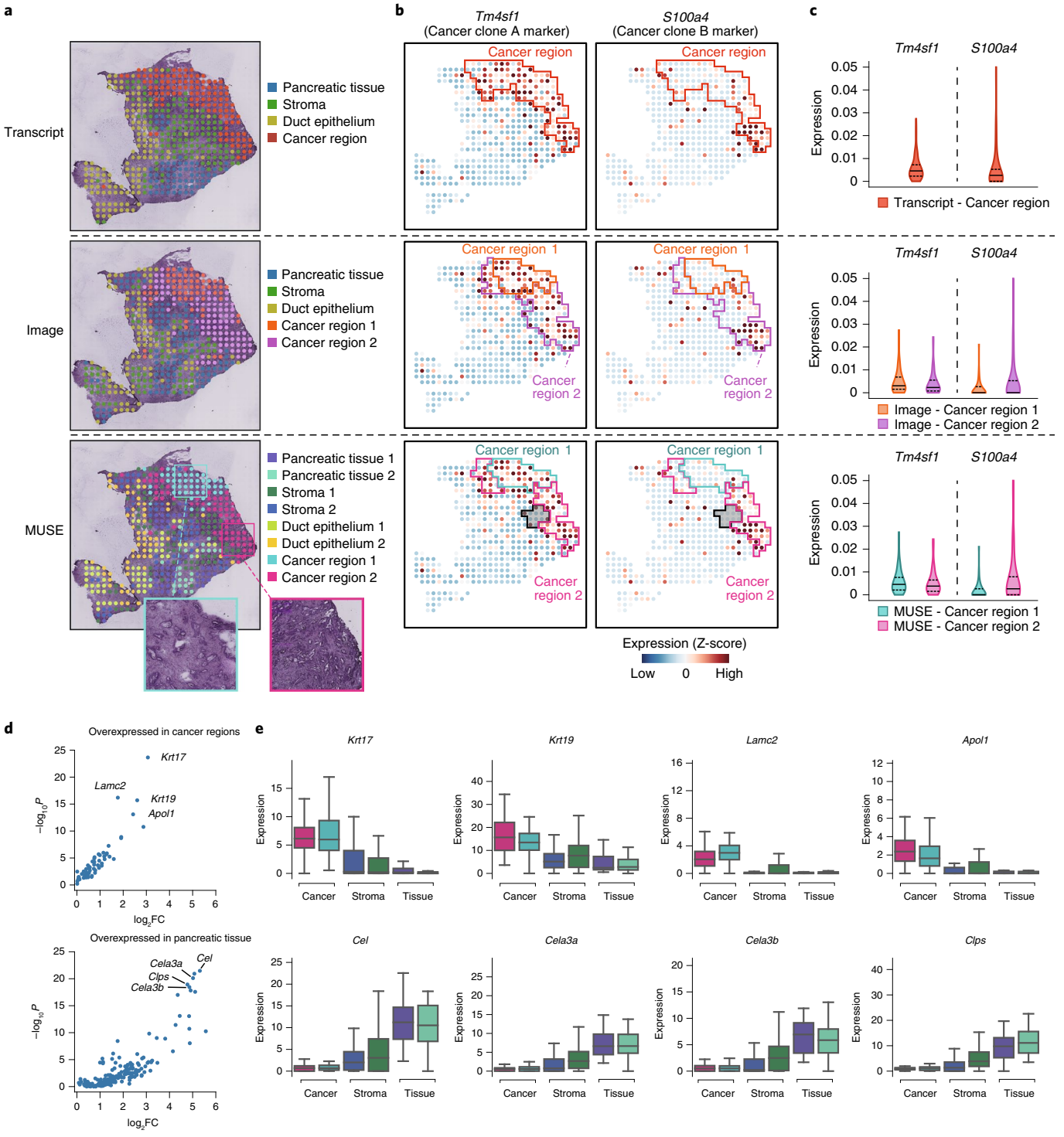


Fig. 4 | Application of MUSE to an ST PDAC dataset. a, Spatial region clusters from transcript-alone analysis (top), H&E image-alone analysis (middle) or combined analysis from MUSE (bottom). **b**, Expression levels of cancer clone A marker (*Tm4sf1*) and clone B marker (*S100a4*) measured by ST. Colored lines: identified cancer regions. Gray-shaded region with black outline: domain annotated as cancer from image analysis but excluded from cancer regions in MUSE. **c**, Expressions (normalized by total gene counts) of two clone makers in cancer regions identified by three analyses (from top to bottom: transcript, image and MUSE). Dashed line in violin plot: ¼ and ¾ quantiles. Concrete line: median. **d**, Overexpressed genes in cancer regions (top) or pancreatic tissues (bottom) through differential expression analysis between pancreatic tissue regions and cancer regions characterized by MUSE. *Krt17*, *Krt19*, *Lmac2* and *Apol1*: PDAC biomarkers; *Cel*, *Cela3a*, *Cela3b* and *Clps*: genes overexpressed in tissues with worse prognosis[44]. **e**, Expression changes (counts after median normalization) of top overexpressed genes (annotated in **d**) from cancer, stroma or pancreatic tissue regions based on MUSE clusters. Box plot: center line, median; box, interquartile range; and whiskers, minimum–maximum range. FC, fold change.

marker genes *Tmfsf1* (high in both clones) and *S100a4* (high in one clone) (Extended Data Fig. 4d). We applied MUSE to the available ST and image multi-modal data and identified two morphologically distinct cancer regions, each of which captured one of the two distinct clones based on the signature marker genes (Fig. 4a,c). These two distinct cancer clones were not well identified by either transcript-only clustering or subclustering (Extended Data Fig. 4e) or by image-only (Fig. 4b,c) analysis.

Outside of the cancer regions, MUSE also dissected non-tumor tissue into spatially distinct subregions. We performed differential expression analysis on these clusters to investigate changes across tissue regions (Fig. 4d and Methods). The top overexpressed cancer-region genes were previously identified PDAC biomarkers (*Krt19* (ref. [40]), *Apol1* (ref. [41]), *Krt17* (ref. [42]) and *Lamc2* (ref. [43])); these genes showed a decreasing trend with increasing distance from the cancer regions (Fig. 4e, top, and Extended Data Fig. 4f, top). In comparison, the top overexpressed non-cancer-region genes showed the reverse trend for distance (Fig. 4e, bottom, and Extended Data Fig. 4f, bottom). Interestingly, eight out of the top ten of these overexpressed non-cancer-region genes (Supplementary Table 2) were also previously reported to be overexpressed in bulk tissue samples from patients with PDAC with worse prognosis[44].

MUSE analysis of human intestine tissue from Visium data. Next, we applied MUSE to a recent dataset generated from the intestine of a male adult colon using the commercially available 10x Visium spatial platform[45]. After removing spots outside of the tissue regions, we analyzed 2,807 sequenced tissue spots (Methods). For the transcript modality, we used the same pipeline as before to select the top 500 variable genes. For the image modality, we extracted deep image features from regions in the H&E images corresponding to the tissue sequencing spots.

For this dataset, the image modality showed clear clustering structure (Extended Data Fig. 5a). MUSE, as well as single-modality analysis, revealed layered intestinal structure (Extended Data Fig. 5b,c). Inspired by a recent study[46], we used known markers from major cell types in this tissue to evaluate the coherence of identified clusters. We focused on four major cell types, which were labeled as epithelium, muscle cells, immune cells and endothelium. Epithelium and muscle layers were clearly identified within the clusters (Fig. 5a,b). MUSE and transcript clusters generally showed higher enrichment of layer-specific genes compared to the image modality alone (Fig. 5c,d). Immune and endothelial cells appear spatially grouped (rather than layered) in H&E images. We observed one major region for each of these two cell types, which were annotated in the previous publication (Fig. 5e, left). We used the identified clusters covering each region (Fig. 5e, right) and visualized cell marker expression (Fig. 5f,g). For the immune region, the image-alone and MUSE subpopulations showed higher marker enrichment, whereas, for the endothelial region, the transcript-alone and MUSE showed higher marker enrichment. An interesting possibility is to use the computational resolution-enhancement method BayesSpace[47] to infer sub-spot transcript profiles and enhance the spatial resolution for ST data before inputting into MUSE. With this additional pre-processing, we observed instances of improved spatial and subpopulation resolution (Extended Data Fig. 5d,e), suggesting a fruitful direction of future integration.

MUSE analysis of Alzheimer's disease from ST data. Finally, we investigated how MUSE could leverage combined ST and pathology biomarker image data. For this, we made use of a recent study of the deposition of amyloid-beta (Aβ) peptide in brain, which is a key pathophysiological hallmark of Alzheimer's disease (AD)[48,49]. Here, we evaluated MUSE on a multi-modal AD dataset with regional information on transcripts (ST) and Aβ distributions (immunofluorescent imaging)[50]. The data consisted of brain samples from ten $App^{NL\text{-}G\text{-}F}$ knock-in mice of four different ages (3, 6, 12 and 18 months), using Aβ accumulation as a proxy for disease progression (Extended Data Fig. 6a). For each sample, we analyzed a tissue section that was spatially barcoded for ST sequencing as well as an adjacent section immunostained for Aβ (Fig. 6a). In total, 5,009 ST spots were sequenced, and we selected the top 500 variable genes as before. For the morphological modality, we segmented the corresponding region in the adjacent fluorescent image (Aβ-channel only) for each ST spot and learned the deep embeddings using the Inception v3 model.

Image-alone analysis highlighted the progression of AD as viewed by age or by the previously defined Aβ index[50] (Fig. 6b and Extended Data Fig. 6b,c). Transcript-alone analysis highlighted expression differences across brain regions (Fig. 6c). Satisfyingly, in combined multi-modal analysis, both the disease progression trajectory (from fluorescent images) and brain region differences (from transcripts) were captured in the MUSE latent space (Fig. 6d). (We note that the 6-month transcript data are distinct from the other data; Fig. 6c,d, dashed ellipses.) We further observed that MUSE clusters captured regional and temporal specificity (Fig. 6e).

We made use of the MUSE clusters to look for Aβ-related genes. We identified four sets of MUSE clusters that had similar regional compositions (enriched in thalamus, hypothalamus, hippocampus and cortex) but different age compositions (Fig. 6f). First, we investigated differentially expressed (DE) (in age) genes within each cluster set. The top DE genes shared across all four regions (Fig. 6f) included AD risk genes identified from previous studies, such as *Ctsd* (top-ranked DE genes for all four cluster sets), *C4b*, *Apoe* and *Trem2* (Supplementary Tables 3–5). Second, we performed pathway enrichment analysis using the DE genes from each of the four sets of clusters (Fig. 6g, top). This allowed us to identify regional differences in aspects of amyloid precursor protein (APP) processing (Fig. 6g, bottom). For example, the hypothalamus is enriched for DE genes related to 'APP catabolic processes', whereas the cortex is enriched for DE genes related to 'cellular response to Aβ formation'. Specific examples of changes in older $App^{NL\text{-}G\text{-}F}$ mice include known AD-related genes *Ranbp9* (downregulated in hypothalamus), *Igf1* (upregulated in cortex) and *Sorl1* (upregulated in hypothalamus but downregulated in cortex). In summary, analysis of MUSE clusters revealed regional, temporal and biological differences reflecting AD progression and raised the hypothesis that APP processing is not uniform across brain regions.

Discussion

The characterization of cellular heterogeneity is fundamental to understanding the organization and function of tissues in health and disease. Two widely used and well-validated methods to study tissue diversity are microscopy (to capture morphological differences) and scRNA-seq (to capture transcriptional differences). MUSE leverages a learning architecture containing self-supervision and self-reconstruction losses that encourage the synthesis of subpopulation structure observed in these distinct modalities. We demonstrated, for both synthetic data and a diverse collection of biological data from different platforms, that MUSE can reveal novel subpopulation structures and tissue organization missed by single modalities or other methods.

Identified cellular subpopulations can, in principle, always be further subdivided to reflect a finer characterization of observed heterogeneity. Subdivision based on multiple modalities provides an opportunity to identify more meaningful biological distinctions. From a technical perspective, multi-modal data may also help to enhance signal from a low-quality or less informative modality (for example, spatially averaged cell measurements, limited numbers of measured cells or low transcript depth). In our studies, subpopulation refinements obtained by combining image and transcriptomics data, rather than transcriptomics alone, appeared to reflect meaningful

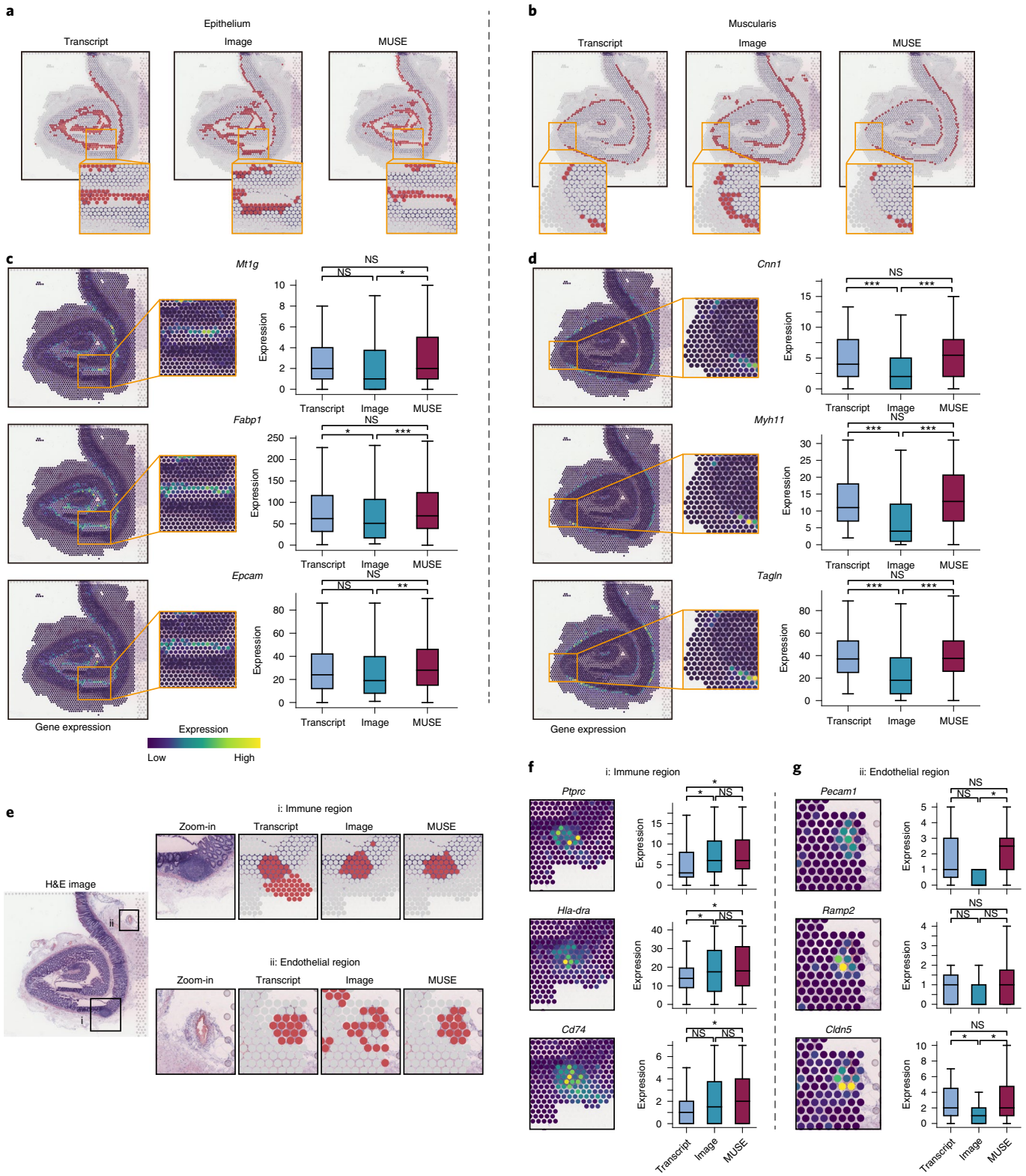


Fig. 5 | Application of MUSE to a 10x Visium human intestine dataset. a, **b**, Epithelium (**a**) and muscularis (**b**) clusters (red spots) identified by transcript-alone (left), image-alone (middle) or combined analysis from MUSE (right). **c**, **d**, Spatial expression heat maps of marker genes for epithelium (**c**) and muscularis (**d**) and their abundances in corresponding clusters from three analyses. Two-sided *t*-tests were used. NS, not significant; **P* value between 0.05 and 0.01; ***P* value between 0.01 and 0.005; ****P* value below 0.005. Box plot: center line, median; box, interquartile range; and whiskers, minimum–maximum range. Same statistical tests and figure annotations apply to **f** and **g**. **e**, Regions enriched in immune cells (i) and endothelium (ii) and subpopulation clusters covering these two regions by each method. **f**, **g**, Spatial expression heat maps of marker genes for immune cells (**f**) and endothelium (**g**) and their abundances in clusters in these regions.

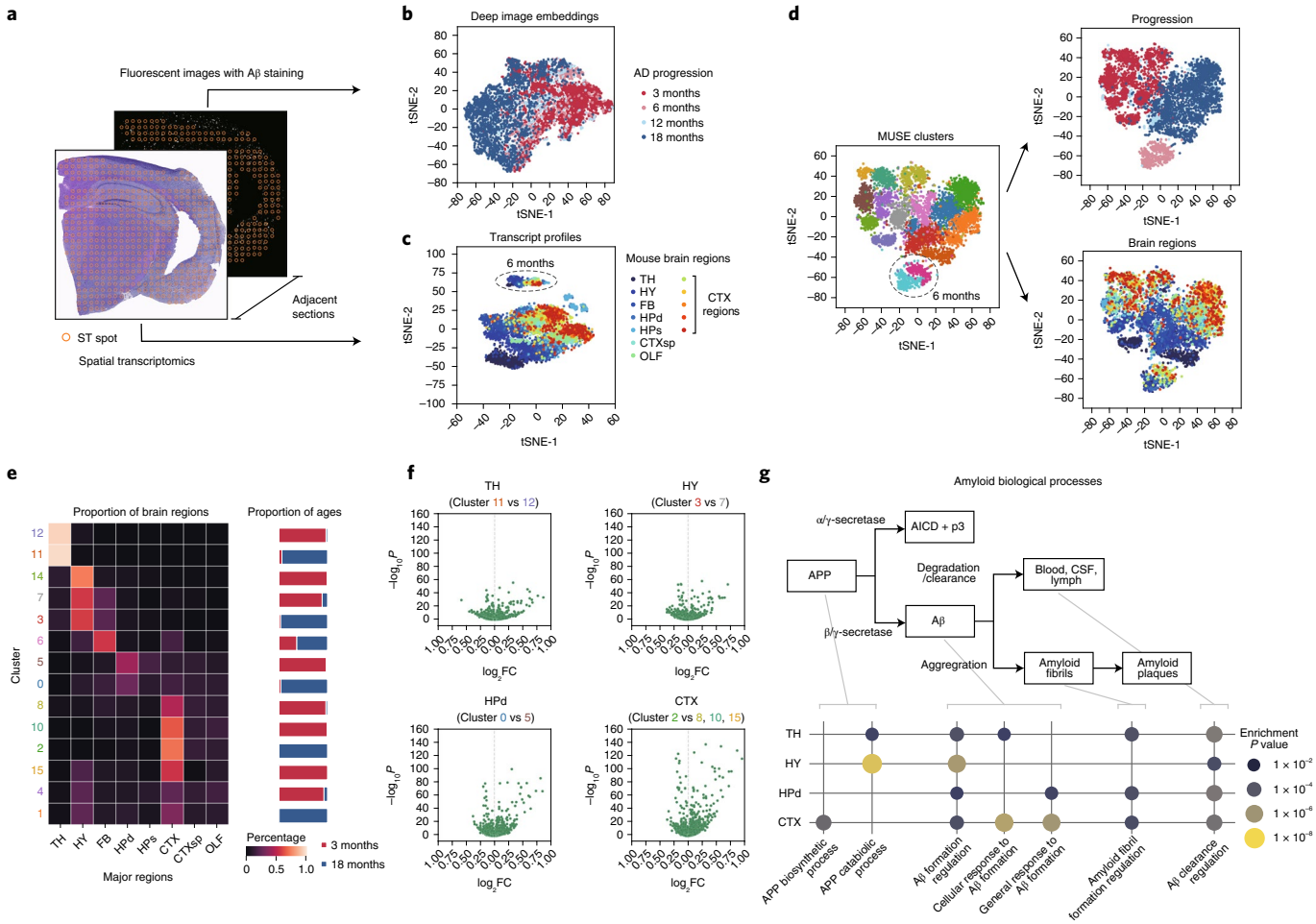


Fig. 6 | Application of MUSE to a multimodal AD dataset. a, Adjacent brain sections were profiled by ST and immunofluorescent imaging with Aβ staining. Transcripts and corresponding image regions for ST spots were used in analyses. **b**–**d**, tSNE visualizations of deep embeddings from Aβ images alone with age annotations (**b**), transcript profiles alone with brain region annotations (**c**) and MUSE joint latent representations from both modalities (**d**). TH, thalamus; HY, hypothalamus; FB, fiber tract; HPd, dendritic hippocampus; HPs, somatic hippocampus; CTXsp, cortical subplate; OLF, olfactory area; CTX, isocortex. **e**, Brain region (left) and age (right) compositions of spots in MUSE clusters. 6-month and 12-month proportions are not reported due to small sample sizes (provided in Extended Data Fig. 6d). Cluster ID colors correspond to MUSE clusters in **d**. **f**, Volcano plots of genes in four cluster sets in **e** with similar brain region compositions (predominant brain region annotated at top). *P* values were derived from one-sided rank-sum test. **g**, Amyloid-related pathways (top) and significantly enriched (enrichment $P < 1 \times 10^{-3}$) processes (bottom) based on **f**. AICD, APP intracellular domain; CSF, cerebrospinal fluid; FC, fold change.

biological differences, including spatial coherence within a layered or heterogeneous tissue, subclonal resolution of growth within a tumor and stages of neurodegenerative disease progression.

We anticipate that MUSE can be extended to integrate advances in spatial resolution[47,51], new measurement technologies[52] or increased numbers of modalities. Computational analysis across -omics modalities, even beyond spatial transcriptomics, holds the potential to increase our power to meaningfully dissect and interpret tissue heterogeneity[47,53,54]. The deep learning approach of MUSE—with its parallelized AE architecture and self-supervision learning approach—is extensible and designed to leverage and combine future advances in measurement modalities.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01251-z.

Received: 5 September 2020; Accepted: 8 February 2022;
Published online: 28 March 2022

References

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. *Nat. Methods* **4**, 445–453 (2007).
3. Feldman, D. et al. Optical pooled screens in human cells. *Cell* **179**, 787–799 (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. *Nat. Biotechnol.* **35**, 551–560 (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 (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. *Nat. 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 (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**, aaa6090 (2015).
14. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
15. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
16. Thompson, B. *Canonical Correlation Analysis: Uses and Interpretation* (Sage, 1984).
17. Argelaguet, R. et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
18. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
19. Hinton, G.E. & Zemel, R.S. In: *Advances in Neural Information Processing Systems* 3–10 (MIT Press, 1994).
20. Baldi, P. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* 37–49 (MLR Press, 2012).
21. Chechik, G., Sharma, V., Shalit, U. & Bengio, S. Large scale online learning of image similarity through ranking. *J.Mach. Learn. Res.* **11**, 1109–1135 (2010).
22. Hoffer, E. & Ailon, N. In: *International Workshop on Similarity-Based Pattern Recognition* 84–92 (Springer, 2015).
23. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
24. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
25. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
26. 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. *Nat. Methods* **16**, 311–314 (2019).
27. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1169 (2020).
28. Yuan, G.-C. et al. Challenges and emerging directions in single-cell analysis. *Genome Biol.* **18**, 84 (2017).
29. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
30. 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. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4260 (2018).
31. Belgard, T. G. et al. A transcriptomic atlas of mouse neocortical layers. *Neuron* **71**, 605–616 (2011).
32. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
34. Pierson, E. & Yau, C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 (2015).
35. Yao, Z. et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222–3241 (2021).
36. Mbeunkui, F. & Johann, D. J. Cancer and the tumor microenvironment: a review of an essential relationship. *Cancer Chemother. Pharmacol.* **63**, 571–582 (2009).
37. Sun, Y. et al. Treatment-induced damage to the tumor microenvironment promotes prostate cancer therapy resistance through WNT16B. *Nat. Med.* **18**, 1359–1368 (2012).
38. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
39. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).
40. Yao, H. et al. Glypican-3 and KRT19 are markers associating with metastasis and poor prognosis of pancreatic ductal adenocarcinoma. *Cancer Biomark.* **17**, 397–404 (2016).
41. Liu, X. et al. A new panel of pancreatic cancer biomarkers discovered using a mass spectrometry-based pipeline. *Br. J. Cancer* **117**, 1846–1854 (2017).
42. Roa-Peña, L. et al. Keratin 17 identifies the most lethal molecular subtype of pancreatic cancer. *Sci. Rep.* **9**, 11239 (2019).
43. Yang, C. et al. Evaluation of the diagnostic ability of laminin gene family for pancreatic ductal adenocarcinoma. *Aging (Albany NY)* **11**, 3679–3703 (2019).
44. Van den Broeck, A., Vankelecom, H., Van Eijsden, R., Govaere, O. & Topal, B. Molecular markers associated with outcome and metastasis in human pancreatic cancer. *J. Exp. Clin. Cancer Res.* **31**, 68 (2012).
45. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–826 (2021).
46. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
47. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).
48. Hardy, J. A. & Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256**, 184–186 (1992).
49. Murphy, M. & Levine, H. III Alzheimer's disease and the amyloid-β peptide. *J. Alzheimers Dis.* **19**, 311–323 (2010).
50. Chen, W.-T. et al. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991(2020).
51. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
52. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
53. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
54. Pham, D. et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2020.05.31.125658v1 (2020).

Methods

Overview. The ability of MUSE to combine multi-modal data is built around the following two principles. First, MUSE makes use of an AE architecture with self-reconstruction loss for each modality. The AE encodes data in the original space (capturing high-dimensional, multi-modal measures of the cells) into points in a joint 'latent' space (a low-dimensional space that is allowed to mix the modalities) that, although low dimensional, still contains enough information to be decoded faithfully back to the original space. A self-reconstruction loss measures 'faithfulness' as the difference between the original and decoded data and is minimized during training of MUSE. Second, MUSE makes use of a self-supervision loss, which is simultaneously minimized with the reconstruction loss during training of MUSE. Subpopulation structure in each modality is calculated in each training iteration. During training, the self-supervision loss encourages points that are near (or far from) each other in each modality to remain near (or far from) each other in the joint latent space. Taken together, MUSE provides a way to combine features across modalities that respects heterogeneity within each modality. Although MUSE is demonstrated using image and transcriptomics data, the approach is general.

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

Zero-inflated multi-modal AE. The whole AE structure is illustrated in Extended Data Fig. 1r. Features from two modalities (x_i and y_i) are input into a multi-modal AE, 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 $f_{\text{Encoder}_x}(\cdot)$ and $f_{\text{Encoder}_y}(\cdot)$ 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 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-supervision loss.

Next, we use weight matrices w_x and w_y to transform information in z_i for the reconstruction of the original features from i^{th} sample 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 matrices that transform information in z_i to generate modality-specific features $z_i^{(x)}$ and $z_i^{(y)}$. The Frobenius-norm (F-norm) is used on w_x and w_y to alleviate overfitting. Finally, features for each modality are reconstructed by decoders:

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

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

where $f_{\text{Decoder}_x}(\cdot)$ and $f_{\text{Decoder}_y}(\cdot)$ 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 or sequencing on a small number of cells. Therefore, transcript profiles 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:

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

where the function sign $(\cdot)$ returns a vector of 0s or 1s for each entry of the vector x_i depending on whether the entry is zero or not (respectively); $\sum\text{sign}(x_i)$ returns the total number of non-zero entries in the vector x_i; and $\circ$ is the element-wise product of two vectors. Thus, the numerator returns the total reconstruction error for x_i using genes with non-zero expressions, and the ratio calculates a normalized reconstruction error for expressed genes in the i^{th} sample.

For the morphological modality, we used the standard reconstruction loss:

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

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

$$L_{\text{reconstruct}} = L_{\text{reconstruct}_x} + L_{\text{reconstruct}_y} + \lambda_{\text{regularization}}\left(\|w_x\|_{\text{F}} + \|w_y\|_{\text{F}}\right)$$

where $\lambda_{\text{regularization}}$ is the regularization hyperparameter and is determined through analysis on simulation data (Extended Data Fig. 1s). $\|\cdot\|_{\text{F}}$ represents the F-norm of matrix. Effects from $\lambda_{\text{regularization}}$ and cluster number to regularization value were quantified in Extended Data Fig. 1t.

Self-supervision loss. To extract useful information from each modality and increase the quality of the joint latent feature z_i, we further used structured self-supervision 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[23] 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, and z_{pos_x} is a positive sample from the same cluster as the anchor based on clusters from modality x. z_{neg_x} is a negative sample from a different cluster; ε is the margin; and 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 anchor-negative samples to be greater than the margin so that the loss approaches its minimum (that is, 0). Positive and negative samples were randomly selected during training. 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 the 'Optimization of MUSE' section below).

Loss function. The overall loss function for training is the combination of the self-reconstruction and self-supervision 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 (Extended Data Fig. 1s-(2)).

Optimization of MUSE. MUSE is trained on raw features and reference labels from two modalities 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 AE 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[23] with the Louvain method, which determines

the 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 the graph. We also tested the effects of nearest neighbor numbers on the cluster numbers and MUSE accuracies (Extended Data Fig. 1u). We note that the architecture of MUSE is flexible, and other (for example, modality-specialized) clustering approaches can be used instead of PhenoGraph to provide cluster labels l_{x_i} and l_{y_i}.

Spatial transcriptomics data pre-processing. In this work, we made use of five datasets using seqFISH+, STARmap, ST, Visium and ST with fluorescent imaging, respectively.

seqFISH+ cortex dataset. The seqFISH+ data includes 523 cells from five fields of views in a mouse cortex. For transcriptional modality, 10,000 RNAs were profiled using in situ sequencing for each cell. We performed pre-processing based on gene count data and 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 the scanpy Python package (version 1.4.4)[55]. For morphology modality, DAPI and Nissl stains were used in imaging, and cells were segmented manually based on their morphology by seqFISH+ authors. Each cell segmentation region was placed at the center of an empty image with 299×299 pixels. Then, DAPI and Nissl channels were input into the pre-trained Inception v3 deep neural network independently. The output of the last network layer (with 2,048 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 the same mean value for transcript features across cells and then were used as input to the tested models. We used the Inception v3 network[33] with pre-trained parameters provided by TensorFlow Hub (https://tfhub.dev/).

STARmap cortex dataset. 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. We used the same analysis pipeline modality as in seqFISH+ for transcript. For morphology modality where cell markers were absent, we directly placed the provided cell segmentation masks over blank images and input them into the same Inception v3 neural network as before. Outputs from the last layers were also compressed using PCA and scaled to obtain single-cell morphological features.

Spatial transcriptomics dataset of PDAC. The PDAC dataset employed the ST technology and sequenced 428 spots in the tumor tissue. For transcript data, we followed the pre-processing in the original publication[39], normalized each spot by total counts and then scaled using median transcript counts. We also performed log(1+x) transformation and selected the top 500 most variable genes. For image, we identified corresponding H&E regions for each spot using ST positions, resized image tiles to 299×299 pixels and learned deep embeddings using Inception v3. To test the region size that provided the best description of microenvironment, we considered several different image sizes and used the one that gave the best separation (Extended Data Fig. 4g).

10x Visium dataset of human intestine. The sample was from the colon of a male patient aged 66 years and was labeled as 'A1' in the original publication[45]. The 10x Visium array covered 4,992 spots. After removing spots that did not map to the tissue region, 2,807 tissue spots were left for the subsequent analysis. For the transcript modality, we again used median normalization and the log1p transform and selected the top 500 most variable genes. For H&E images, tiles corresponding to tissue spots from the downsampled image were segmented, resized and input to Inception v3 to obtain deep features.

Spatial transcriptomics and fluorescent imaging dataset of AD. The AD dataset contains six $App^{NL\text{-}G\text{-}F}$ knock-in mice in the age of 3 (two mice), 6, 12 and 18 (two mice) months. In total, ten brain samples were collected, with eight of them from 3- or 18-month-old mice (Extended Data Fig. 6a). For each sample, three adjacent sections were obtained, where the middle layer was used for ST sequencing and the other two were immunostained for imaging (except for the 6-month sample where only one section was available). Each section included ~500 ST spots, and, together, 5,009 spots were obtained. For transcript modality, we employed the filtered logCPM data provided in the original publication[13] and selected the top 500 variable genes. For morphological modality, we selected the Aβ channel of fluorescent images and segmented regions using the coordinates from adjacent ST spots. Then, Aβ images were resized and transformed to deep embeddings using the Inception network and compressed into 500-dimension morphological input.

Rationale for using deep image features. For image data contained within the spatial transcriptomics dataset, data scale is often small, and subpopulation annotation is less available. Here, we posited that deep features could provide an efficient method for constructing rich, informative image profiles. To this end, for all datasets in this work, we used Google Inception v3, trained on ImageNet (ILSVRC-2012-CLS dataset), to extract deep features from cell or tissue spot images. The deep neural network architecture of Inception v3 (with ~25 million parameters) and large training set (1.2 million images from 1,000 categories) enabled the model to capture and encode general image features at different scales into highly compact representations. Although the interpretation of deep image features is not always immediate, in some cases it is possible to find intuition by determining where more traditional cell and tissue properties (for example, cell shape) map into the learned deep image feature space (Extended Data Fig. 2a).

Simulation experiment setup. We generated simulated ground truth class labels $l \in \{1, \ldots, L\}^n$ for n cells and L possible cluster (that is, 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 separately, but all clusters could be discriminated using the combination of both modalities (Extended Data Fig. 1c). 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 and l_y for the two modalities.

For example, ten ground truth clusters, labeled $\{1, \ldots, 10\}$, could be divided into groups $G_1 = \{1, \ldots, 5\}$, $G_2 = \{6, \ldots, 10\}$, with modality 1 considering merges from G_1 but not G_2 and vice versa for modality 2; after merging, modality 1 might have seven clusters $\{1 \cdot 2 \cdot 3, 4 \cdot 5, 6, 7, 8, 9, 10\}$, whereas modality 2 might have six clusters formed from $\{1, 2, 3, 4, 5, 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10\}$ (where '·' indicates merged clusters). Although each modality can distinguish only 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[25] and scScope[26]. In short, we generated latent codes $z_i^{(x)} \in \mathbb{R}^m$ for cell i using a multivariable normal distribution:

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

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 $x_i^{\mathrm{raw}} = A^{(x)} z_i^{(x)}$, 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 $x_i^{\mathrm{noise}} = x_i^{\mathrm{raw}} + \varepsilon$, 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:

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

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); and 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 $[-0.5, 0.5]$; 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 to obtain reasonably similar ARI scores for clustering based on each modality alone.

We generated 'semi-simulated' data using real images (Extended Data Fig. 1i). To create the image modality data, we used the image features from the STARmap data and PhenoGraph to obtain image clusters. Next, we randomly subdivided the image clusters to create ground truth labels for all cells and then simulated the associated transcript modality data as described in the previous section.

Analysis of gene expression data. *Differential analysis of expression data.* With cluster labels, we identified DE genes using fold changes and P values. For each cluster, we compared within versus across cluster gene expression of cells. $\log_2$ 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 rank-sum test on expression profiles between the two groups to measure overall expression distribution differences. For the STARmap dataset, the analysis was conducted among 'reproduced' MUSE clusters (Fig. 3c). For the PDAC dataset, the

differential analysis was conducted between two cancer regions and two pancreatic tissue regions (Fig. 4d). For the AD dataset, clusters with similar brain region compositions were analyzed independently (highlighted in bracket in Fig. 6e; late stage versus early stage).

Annotation of cortex layers. For the seqFISH+ data, marker genes that identify different cortex layers were obtained from the literature[31] (in our case, four different genes were used to identify four different layers; Extended Data Fig. 2b). 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 co-localization patterns (based on spatial co-localization score) were identified for annotation. Next, a one-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, KDE was performed using the KernelDensity function from the sklearn Python library with bandwidth determined by Scott's rule.

Annotation of tissue regions for PDAC data. For transcript-only analysis results, we directly used spot clusters and annotations from the original paper[39]. For image and MUSE clusters, we used the histological annotations based on H&E image (also provided in original publications; Extended Data Fig. 4b) as a reference.

Gene Ontology enrichment analysis. Based on differential analysis, we sorted genes by P values. Then, we made use of the online Gene Ontology analysis tool GOrilla[56] and uploaded the ordered gene list to quantify the significance of biological processes.

Subpopulation identification and evaluation. *Subpopulation discovery.* Unless otherwise noted, PhenoGraph was used on the latent representation learned from different methods to identify clusters. PhenoGraph automatically determines optimal cluster number, which is valuable for analyzing new datasets where the true subpopulation size is unknown. We used default PhenoGraph parameters in the analysis. The effect of varying the number of neighbors used for graph construction on MUSE performance was analyzed in Extended Data Fig. 1u.

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[57]. We note that, throughout this work, spatial coordinates were used only in evaluation and never used as input to a method.

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:

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

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

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

where $N_H = (n_c - 1)(n_c - 2)/2$ 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 1,000 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 ARI. For simulation studies, where ground truth subpopulation labels were given, we evaluated clustering performance using the ARI[24]. An ARI near 1 indicates a strong match to ground truth clustering, whereas values near 0 suggest random assignment. In the implementation, we used the adjusted_rand_score function from the sklearn.metrics Python package.

Feature quality evaluation with silhouette coefficient. The quality of latent features was evaluated by the compactness of the clusters in the latent space using the silhouette coefficient[58]. A score of 1 indicates highest density in latent space. In our implementation, we employed the silhouette_score function from the sklearn.metrics Python package.

Comparing methods. All compared methods were run on the same input features (see the 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[23]. 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. PCA: sklearn 0.20.3 Python package. ZIFA[34] is a Bayesian approach that uses a statistical graph model to simulate the generation of gene count data. We used ZIFA.fitModel() from the ZIFA Python package (version 0.1) with latent code $k = 100$. SIMLR[25] uses multiple kernels to construct the sample similarity matrices at multiple metrics and then decomposes similarity into low-dimensional representations. Here, we used the Python implementation of SIMLR (version 0.1.3) with 30 neighbors and maximal five iterations to construct the graph. scScope[26] tackles the gene dropout using a recurrent AE and takes the bottleneck layer as latent representations. We used the Python implementation scScope (version 0.1.5) with two recurrent layers, 64 batch size and 100 epochs.

Morphological feature learning methods. PCA: as above. MDS, Isomap and tSNE: sklearn.manifold Python library (version 0.20.3). We note that the tSNE method can only support maximal three-dimension latent representations.

Multi-modal feature learning methods. CCA (from the sklearn Python package version 0.20.3) learns linear transformations of multi-view data and maximizes their correlations in latent spaces. We chose the CCA transformation of the transcript data for clustering. This widely used multi-modal analysis approach is best applied when the two modalities are equally informative. MOFA+[18] is designed to combine multi-omics data using multi-view matrix factorization: mofapy2 package (version 0.3) with factors = 100, iteration = 500, group number = 1 and view number = 2 to learn 100-dimension joint features. 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. For the concatenation analyses of two modalities (Extended Data Fig. 1g), we used three approaches, including standard scaling, min–max scaling or quantile transformation (implemented in the sklearn Python package), to normalize features and then concatenated them into joint features and extracted the top 100 principal components (PCs) for downstream analysis. MUSE: software is implemented in Python 3.7.3 with NumPy (version ≥1.16.2), SciPy (version ≥1.4.1), PhenoGraph (version ≥1.5.4) and TensorFlow (version ≥1.14.0) packages (details provided at https://github.com/AltschulerWu-Lab/MUSE); hyperparameter values were chosen through simulation study (Extended Data Fig. 1s) and are provided in Supplementary Table 6. The same parameters were used in all studied datasets.

Combined enhancement analysis using BayesSpace and MUSE. We followed the BayesSpace analysis pipeline and used the top 15 PCs of the transcripts as the input. BayesSpace (R package version 1.2.0) was run with default parameters (nrep = 1×10^4; nrep = 2×10^5; gamma = 3; jitter scale = 5.5; jitter prior = 0.3) to enhance each tissue spot into six subspots. The expression profiles, as well as spatial coordinates (also in 15 dimensions), were inferred (Extended Data Fig. 5d, left). Using the subspots' positions, we identified corresponding image regions and extracted deep image features in the same procedure as before. Then, MUSE was performed on enhanced transcript profiles and image features to learn joint representations. As the transcript profiles from BayesSpace were the top PCs instead of counts, we modified the transcriptional reconstruction loss in MUSE to include all entries.

Software used in the study. Software packages use in the study can be accessed via the 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
BayesSpace: https://github.com/edward130603/BayesSpace
GOrilla: http://cbl-gorilla.cs.technion.ac.il/
TensorFlow Hub: https://www.tensorflow.org/hub

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

Data availability

seqFISH+ mouse cortex dataset: Transcript data were downloaded from the GitHub page of the seqFISH+ project (https://github.com/CaiGroup/seqFISH-PLUS) on 1 August 2019. Nissl and DAPI stained images were provided by the authors of the seqFISH+ paper.
STARmap mouse cortex dataset: Raw data were downloaded from the project page (http://clarityresourcecenter.org/) on 2 July 2019. Transcript profiles and cell segmentation masks were extracted from data using the Python pipeline provided by the authors at https://github.com/weallen/STARmap.
PDAC dataset: Both spatial transcriptomics (including gene expressions and H&E images) and scRNA-seq datasets were downloaded from the Gene Expression Omnibus (GEO) database with accession number GSE111672.
Intestine dataset: 10x Visium spatial transcriptomics were downloaded from the GEO database with accession number GSE158328.
AD dataset: Raw and normalized count matrix of the spatial transcriptomics were downloaded from the GEO database of the project (accession number GSE152506). Immunofluorescence images (Abeta, GFAP, NeuN and DAPI staining) that correspond to spatial transcriptomics data were downloaded from the 'synapse.org' page of the project (https://www.synapse.org/#!Synapse:syn22153884/wiki/603937) on 31 October 2020.

Code availability

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 code and demonstration code are available on GitHub (https://github.com/AltschulerWu-Lab/MUSE).

References

55. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
56. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
57. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
58. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

Acknowledgements

We thank C.-H. L. Eng at Caltech for providing seqFISH+ image data; X. Wang at the Broad Institute and MIT for providing information on STARmap data analysis; R. Moncada at NYU for advice on PDAC data analysis; H. Koohy and A. Antanavicute from Oxford for providing full-resolution human intestine images; and O. Moindrot at Stanford for the open-source implementation of the triplet loss. We thank J. Bieber, H. Hammerlindl, L. Rao, X. Sun and other members of the Altschuler and Wu laboratories 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. Q.D. was supported by the projects of NSFC (no. 62088102) and the MOST (no. 2020AA0105500). Y.D. was supported by the projects of NSFC (no. 61971020 and 62031001) and the MOST (no. 2020AAA0105502).

Author contributions

F.B., Y.D. and Q.D. developed the approach and conducted simulation experiments. F.B., Y.D., S.W., B.W., S.Q.S., S.J.A. and L.F.W. conducted experimental analyses on biological datasets. The manuscript was written by F.B., Y.D., S.Q.S., S.J.A. and L.F.W. All authors read and approved the manuscript.

Competing interests

S.J.A. and L.F.W. have consulting agreements with Nine Square Therapeutics and BAKX Therapeutics involving cash and/or equity compensation. All other authors declare no competing interests.

Additional information

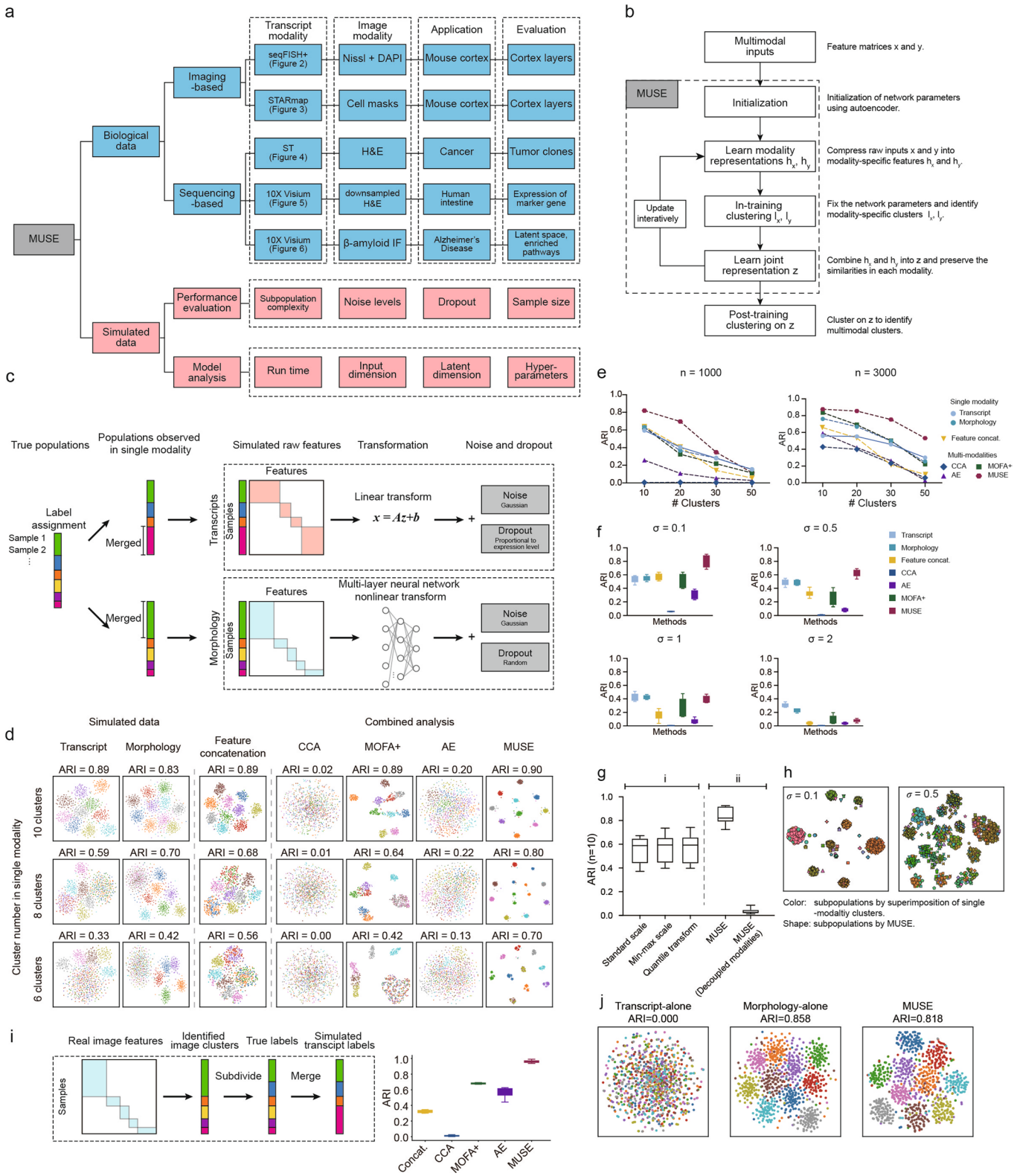
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01251-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01251-z.

Correspondence and requests for materials should be addressed to Qionghai Dai, Steven J. Altschuler or Lani F. Wu.

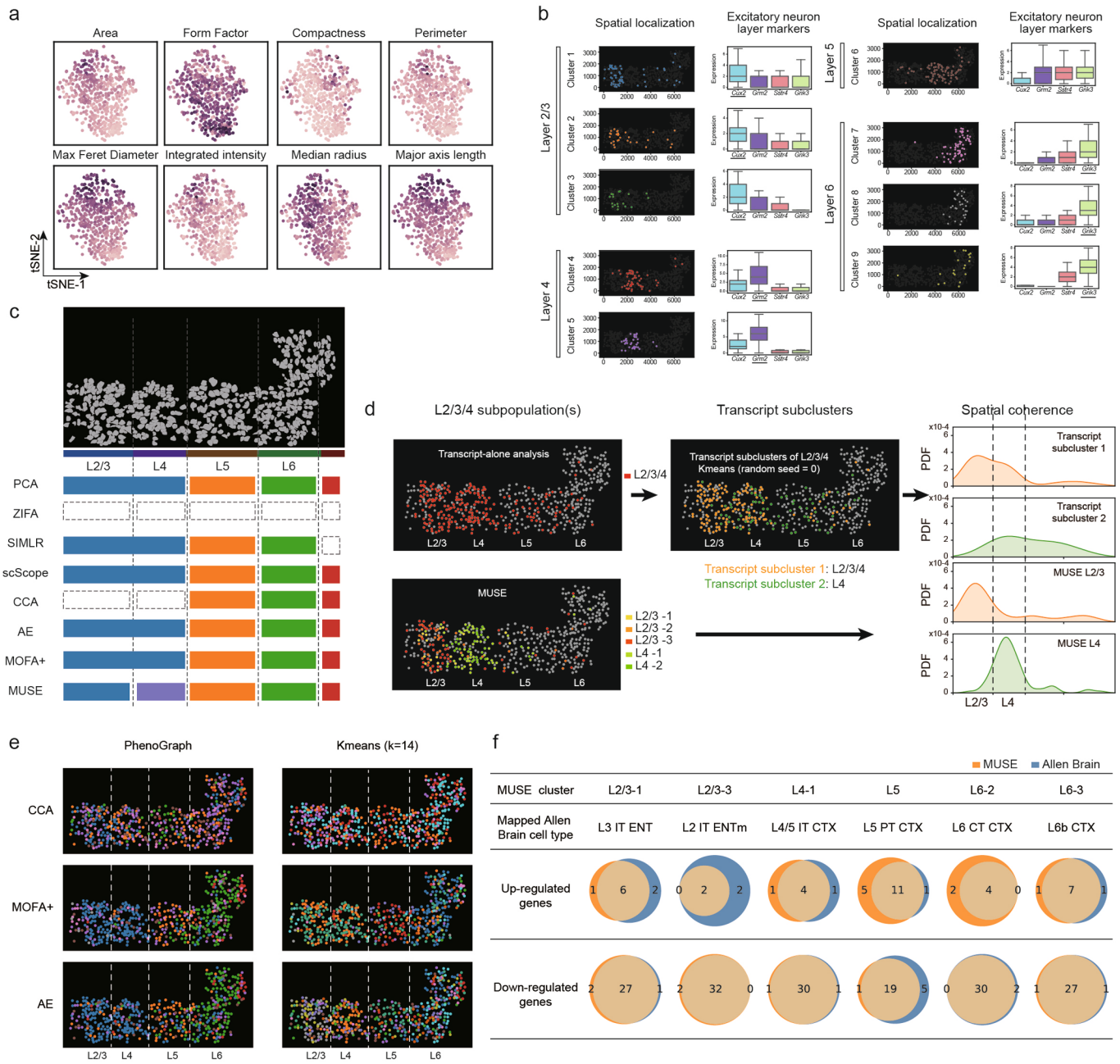
Peer review information *Nature Biotechnology* thanks Itai Yanai, Raphael Gottardo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

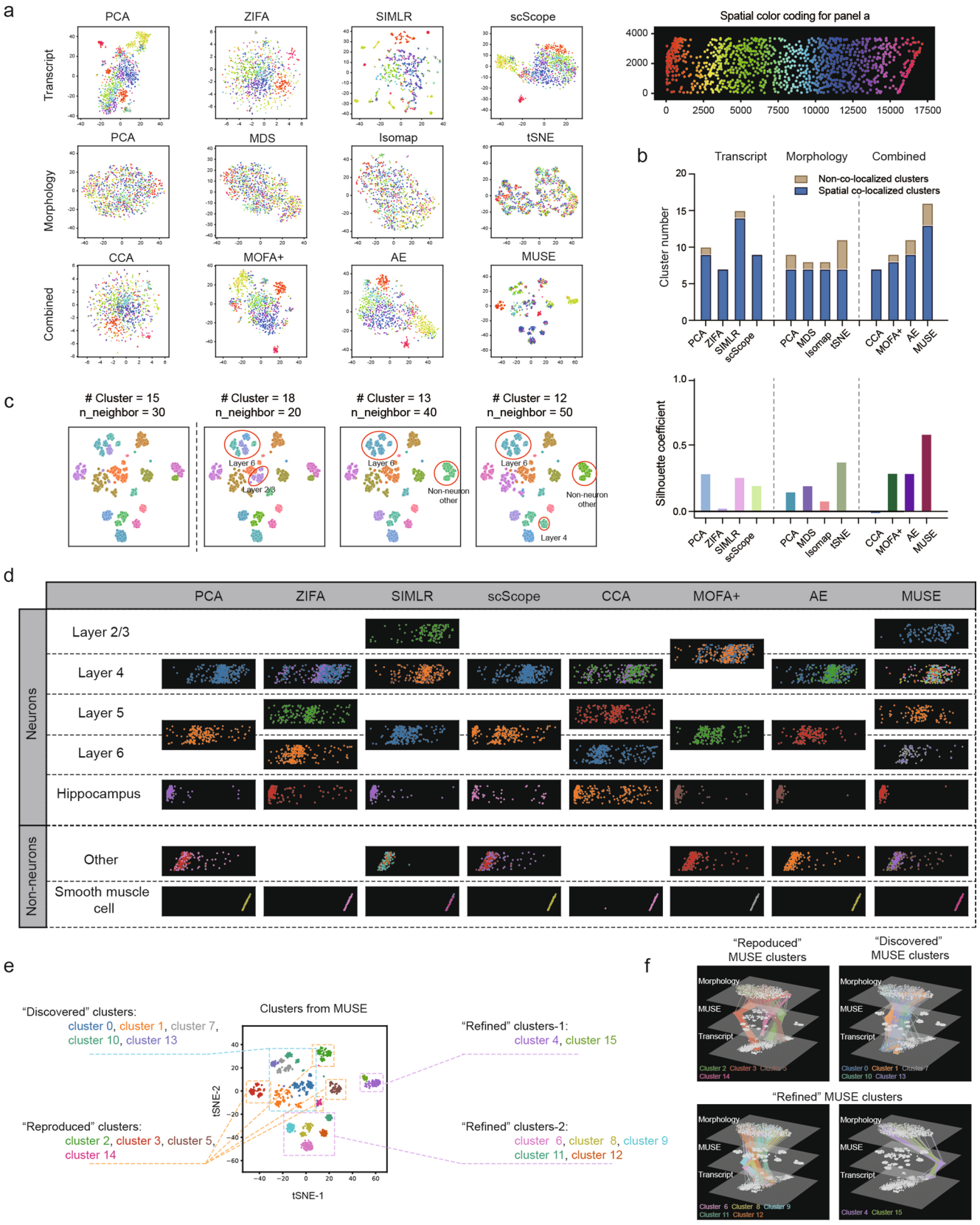


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Overview and simulation studies of MUSE, related to Fig. 1. Parameters used in simulation were listed in Supplementary Table 1. (**a**) Summary of data and analysis used in this work. (**b**) A flowchart of MUSE analysis pipeline. (**c**) Simulation design (Methods) to generate sample profiles with two modalities used for (d-s) below. (**d**) tSNE visualizations of latent representations from single- and combined-modality methods for randomly selected simulation experiments in Fig. 1c. Colors: ground-truth subpopulation labels in simulation. (**e**) Evaluation of combined methods in simulated data with different ground-truth cluster numbers. n = 1,000 (top) and 3,000 (bottom) samples were considered in simulations. (Note: for n = 1,000 and cluster number ≥30, each cluster may only contain a small number of samples.) (**f**) Evaluation of multi-modal methods in simulated data with Gaussian noise for increasing variance (σ). (**g**) Clustering accuracies for (i) analyses of concatenated modality features using various normalization approaches (Methods), and (ii) MUSE multi-modal analysis on matched or unmatched (randomly permuted sample order on one modality) data. ARIs were calculated based on n = 10 repeats. Boxplot: center line, median; box, interquartile range; whiskers, minimum–maximum range; same annotation also applies to other boxplots in this figure. (**h**) Example t-SNE visualization of MUSE subpopulations (indicated by shapes) and simple superimposition of single-modality clusters (indicated by colors) with simulation parameters chosen as in (f). (**i**) Simulation design using real morphological features from STARmap (Methods; dataset details were described in Fig. 3) and performances of multimodal methods (right n = 10). (**j**) Multimodal analysis on data with homogeneous features in one modality. Transcript profiles (left) were generated from a normal distribution while morphological features (middle) were simulated from known subpopulations as before. (**k**) Evaluation of clustering accuracy under different dimensions of joint latent representations (n = 10). (**l**) Clustering accuracy of MUSE while changing dimension of morphological features between 100 to 1,000 (n = 10). (**m**) Clustering accuracy of MUSE when fixing the latent representation of single modality (h_x, h_y) to different dimensions. ARIs were averaged on 10 repeats. Red underlines: parameters selected as default. (**n**) Effects of clustering methods on accuracies (n = 10). Cluster numbers for hierarchical and Kmeans methods were chosen using the elbow method with distortion score. (**o**) Run times for compared methods on simulated data; n = 1,000 cells. Note: for fair comparison, all methods were run under CPU mode. (**p**) Run times of MUSE on datasets with larger sample sizes using different clustering methods in label updating during training. (**q**) Accuracies and run times when fixing single modality labels (denoted as l_x and l_y in Methods) to the initial labels in training. Each dot represented one independent experiment. (**r**) Model structure of multi-modal autoencoder used in MUSE. (**s**) Performance evaluation of MUSE with different hyperparameter settings (n = 10): 1) weight of regularization term; 2) weight of supervision term; 3) learning rate; and 4) iteration intervals between cluster updating in training. Red underlines: parameters selected as default in MUSE package. (**t**) F-norms of selective matrices w_x and w_y to different true cluster numbers (left) in data and choices of regularization hyperparameter $\lambda_{regularization}$ (right); n = 10. (**u**) Clustering accuracies (left) and number of clusters (right) from PhenoGraph when change the hyperparameter of *n_neighbor* (n = 10).

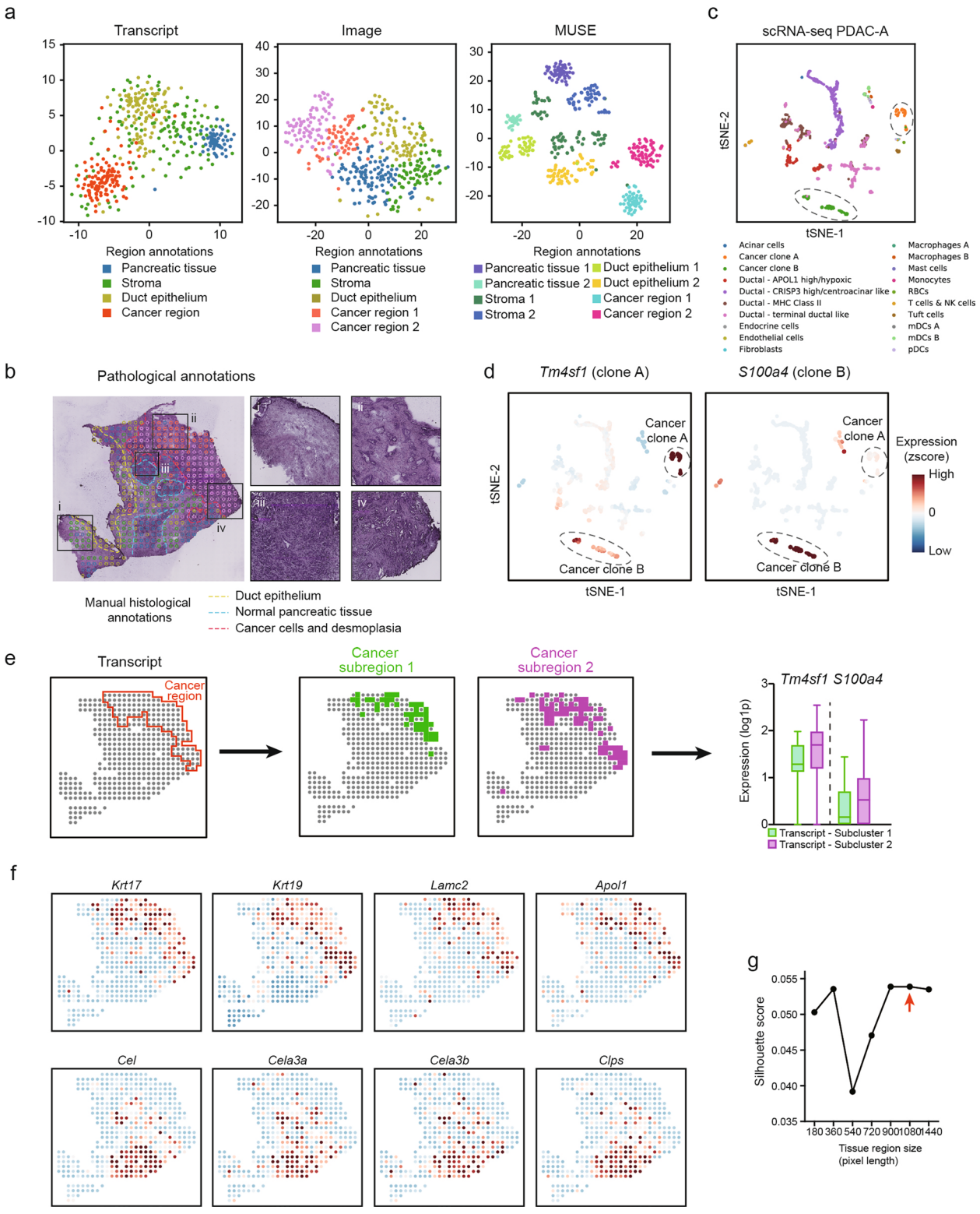


Extended Data Fig. 2 | Analysis of mouse cortex dataset from seqFISH+, related to Fig. 2. (**a**) tSNE visualization of latent space from deep image features, overlaid with various cellular properties from CellProfiler. (**b**) Layer annotations of MUSE clusters based on layer gene markers. Spatial localization of cell clusters (first column) and marker expression abundances (second column) were shown. For each cluster, gene names with maximal overexpression levels were underlined. Boxplot: center line, median; box, interquartile range; whiskers, minimum–maximum range. (**c**) Comparison of discovered cortical layers by transcriptional or combined methods. 5 layers are shown. Squares with the same color and across multiple layers indicate the method discovered merged layers. Squares with no color indicate the method failed to discover the corresponding layer. (**d**) Subclustering analysis on transcript L2/3/4 cluster from Fig. 2c. Kmeans clustering were performed to divide L2/3/4 into two subclusters (middle). Spatial coherences with cortex layers were shown using cell density plots (right). (**e**) Comparisons of subpopulations identified by different clustering methods from multimodal features. In Kmeans, target cluster number (k) was set to the subpopulation size from MUSE analysis. (**f**) Shared up- and down-regulated glutamatergic marker genes between MUSE clusters and cell types from Allen Brain Atlas. Marker genes were obtained from recent Allen Brain Atlas publication; 36 markers were measured in both the seqFISH+ and Allen Brain datasets.



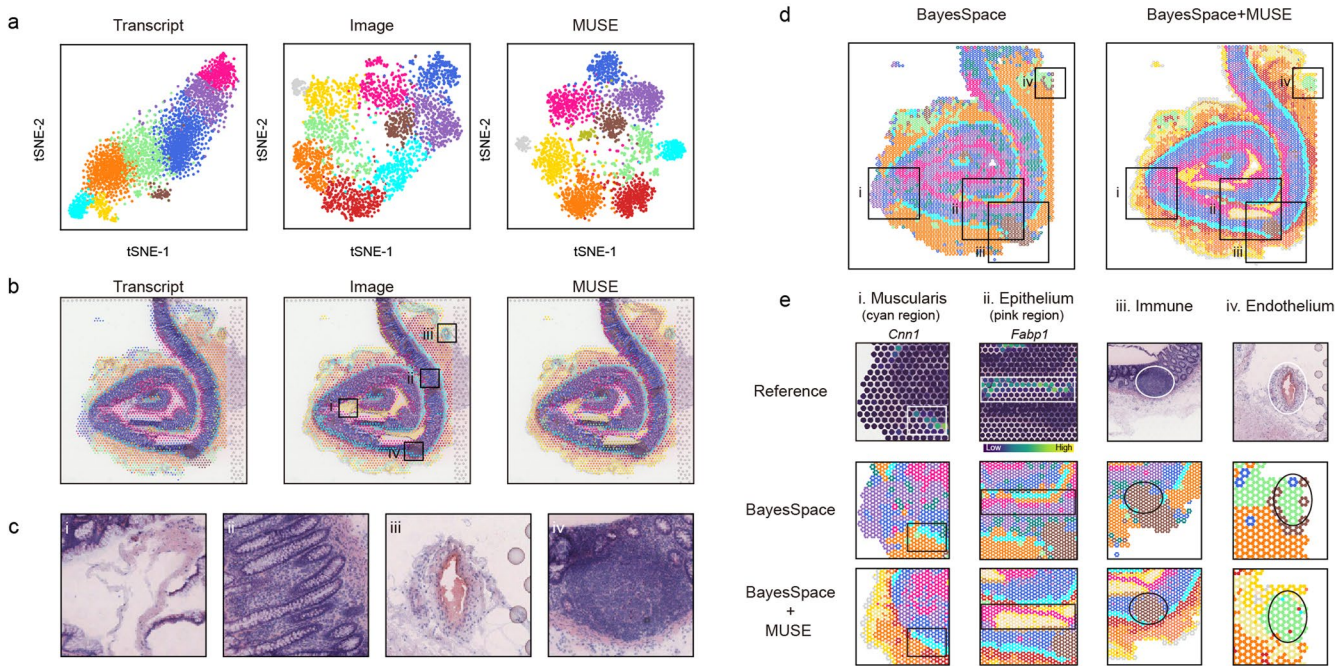
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Comparison of methods on mouse cortex dataset from STARmap, related to Fig. 3. (**a**) tSNE visualization of latent representations by different methods with pseudo-colors labeling cortex depth along x-coordinate (on right side). (**b**) Comparison of cell clusters on (top) numbers of identified clusters with or without significant spatial co-localization properties and (bottom) feature quality evaluation by cluster compactness in latent space using Silhouette coefficient. (**c**) Stability analysis of identified clusters to the choice of hyperparameter *n_neighbor* in PhenoGraph. Red circles: major differences in subpopulations compared with the result using default parameters (left panel) annotated with affected cortex layers. (**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). (**e**) tSNE visualization of MUSE clusters in MUSE latent space. All clusters were classified into 'Refined', 'Reproduced' or 'Discovered' types based on comparison with clusters identified from transcript-alone or morphological-alone analysis (corresponding to Fig. 3a). (**f**) 3D mapping of three types of MUSE clusters in the latent space of morphological features (top layer of each 3D plot), MUSE latent features (middle layer) or transcriptional features (bottom layer). Lines connect the same cells across the three spaces.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Application of MUSE to a multimodal pancreatic ductal adenocarcinoma (PDAC) dataset, related to Fig. 4. (**a**) tSNE visualizations of latent representations and identified clusters by transcripts-alone (left), H&E image-alone (middle) and MUSE (right) analyses (corresponding to Fig. 4a). (**b**) Manual histological annotations (colored lines) provided in original publications overlaid with regional clusters (colored circles) from image analysis. Highlighted regions show the morphological differences. (**c-d**) Analysis of single-cell RNA-seq data from the same PDAC tissue. tSNE visualization with cell type annotations (**c**) and signature gene expressions of two cancer clones (**d**). Cell type annotations from original publication. (**e**) Subclustering analysis of transcript cancer region using Kmeans method and comparisons of clone signature expressions between transcript subclusters and MUSE cancer regions. Boxplot: center line, median; box, interquartile range; whiskers, minimum–maximum range. n = 44 for subcluster 1 and n = 71 for subcluster 2. (**f**) Spatial expression maps of overexpressed genes in cancer regions (top) or pancreatic tissues (bottom) through differential expression analysis between pancreatic tissue regions and cancer regions characterized by MUSE (Methods) (**g**) Cluster separateness of tissue image spots with different size. We segmented image tiles with different pixel sizes and input them into Inception-v3 to learn deep features. Then we performed clustering on features and used Silhouette score to quantify the separateness of clusters. Red arrow indicates the chosen region size.

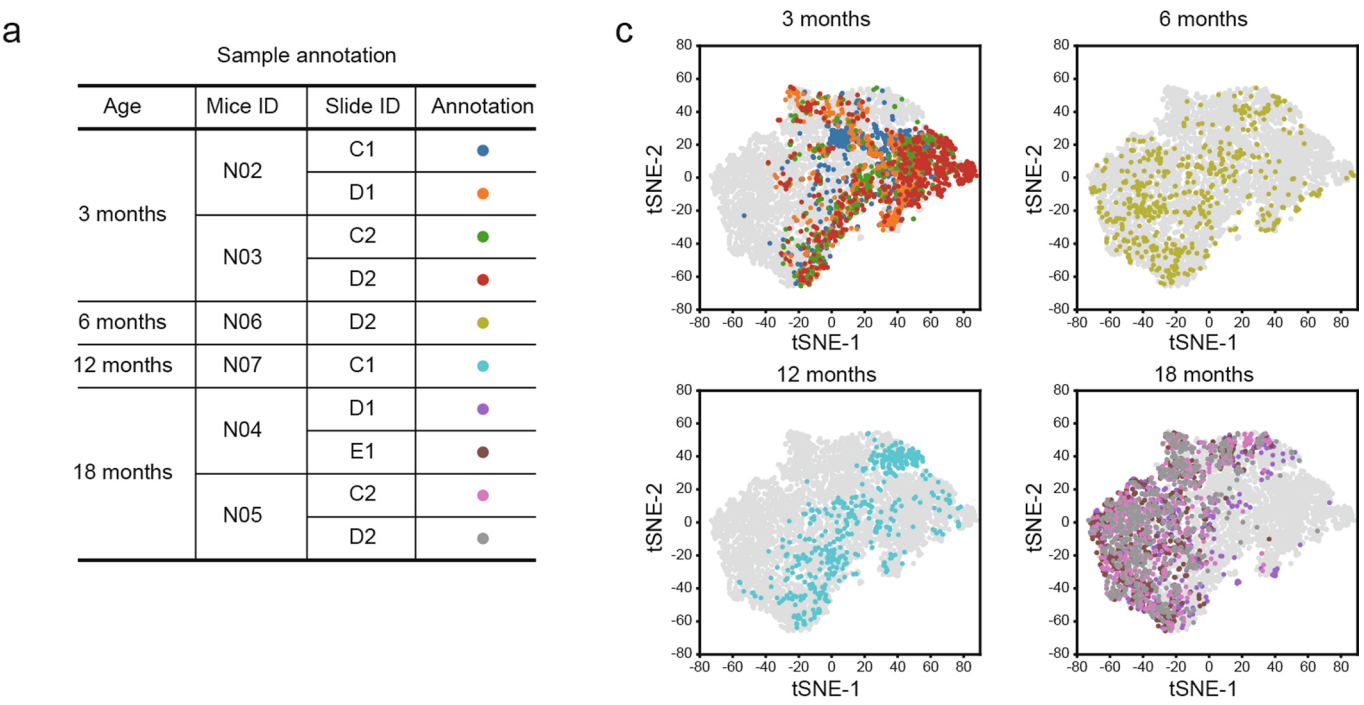


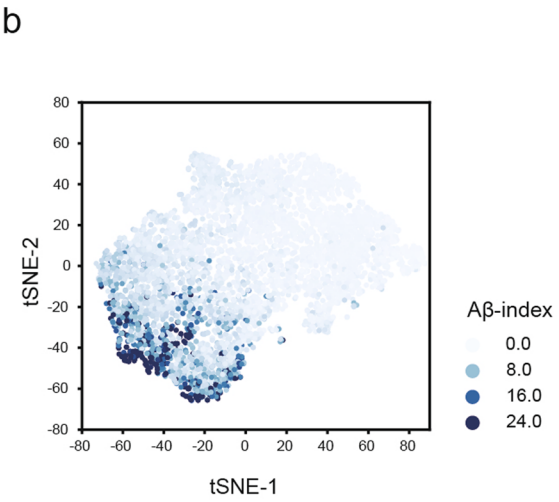
Extended Data Fig. 5 | Application of MUSE to a Visium human intestine dataset, related to Fig. 5. (**a**-**b**) tSNE visualizations of latent representations (**a**) and spatial plots (**b**) of identified clusters by transcripts-alone (left), H&E image-alone (middle) and MUSE (right) analyses. (**c**) Selected regions with various morphological patterns in the tissue. (**d**) Enhanced spatial maps of subpopulations from BayesSpace (left) or BayesSpace + MUSE (right). Details of the analysis were provided in Methods. (**e**) Selected zoom-in region examples with marker gene expressions or morphological patterns (top) and subpopulations defined from BayesSpace (middle) and MUSE (bottom) for four analyzed cell types in Fig. 5.

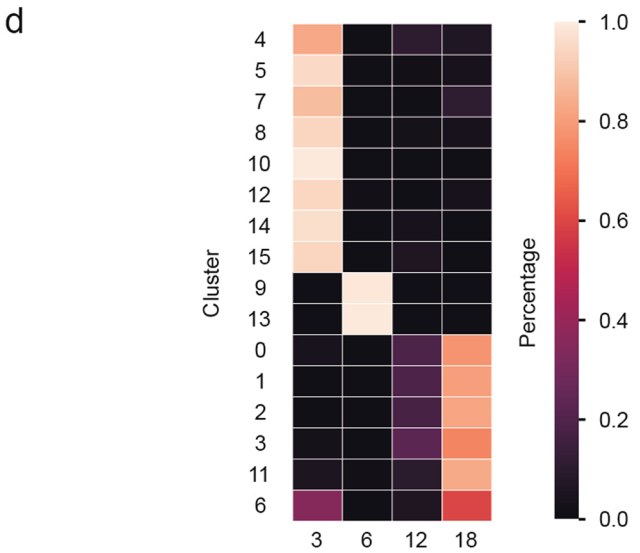
a

Sample annotation

Age	Mice ID	Slide ID	Annotation
3 months	N02	C1	
		D1	
	N03	C2	
		D2	
6 months	N06	D2	
12 months	N07	C1	
18 months	N04	D1	
		E1	
	N05	C2	
		D2	







Extended Data Fig. 6 | Application of MUSE to a multimodal Alzheimer's disease dataset, related to Fig. 6. (**a**) A summary of samples collected in the Alzheimer's disease dataset. (**b**) tSNE was fitted on MUSE deep embeddings and each spot was colored by the Aβ index (defined by standard deviation of intensity in the previous study). (**c**) Visualization of deep embeddings of Aβ spots in the same ages. Color annotations as in (**a**). (**d**) Proportion of samples from all 4 timepoints in each MUSE cluster, related to Fig. 6e.

nature research

Corresponding author(s): Steven Altschuler

Last updated by author(s): Jan 31, 2022

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Principle component analysis (PCA),Multi-dimensional scaling (MDS), Isometric mapping (Isomap), Canonical correlation analysis (CCA), and t-distributed stochastic neighbor embedding (tSNE): sklearn 0.20.3 python package. Zero-Inflated Factor Analysis (ZIFA): ZIFA python package (v0.1). Single-cell Interpretation via Multi-kernel LeaRning (SIMLR): SIMLR python package (v0.1.3). scScope: python implementation scScope (v0.1.5). Multi-omics factor analysis v2 (MOFA+): mofapy2 package (v0.3). Multi-modal structured embedding (MUSE): software is implemented in python 3.7.3 with NumPy (version ≥1.16.2), SciPy (version ≥ 1.4.1), PhenoGraph (version ≥ 1.5.4) and TensorFlow (version ≥ 1.14.0) packages (details were provided at https://github.com/AltschulerWu-Lab/MUSE). BayesSpace: R package v1.2.0. Inception v3 pretrained model was obtained from TensorFlow Hub library (version 0.5.0; https://tfhub.dev/). All methods were configured with default parameters (unless specifically noted in the manuscript) 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).

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

Data

Policy information about availability of data

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

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

seqFISH+ mouse cortex dataset (related to Fig. 2): 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 mouse cortex dataset (related to Fig. 3): raw data were downloaded from the project page (http://clarityresourcecenter.org/ 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. PDAC datasets (related to Fig. 4): Both spatial transcriptomics (including gene expressions and H&E images) and scRNA-seq datasets were downloaded from Gene Expression Omnibus (GEO) data base with accession number GSE111672. Intestine dataset (related to Fig. 5): 10X Visium spatial transcriptomics were downloaded from the GEO database with the accession number GSE158328. AD dataset (related to Fig. 6): Raw and normalized count matrix of the spatial transcriptomics were downloaded from the GEO database of the project (accession number: GSE152506). Immunofluorescence images (Abeta, GFAP, NeuN and DAPI staining) that correspond to spatial transcriptomics data were downloaded from the "synapse.org" page of the project (date: October 31, 2020; link: https://www.synapse.org/#!Synapse:syn22153884/wiki/603937).

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Sample size were determined in original publications. seqFISH+ cortex data: n=523 cells; STARmap cortex data: n=973 cells; PDAC ST data: n=428 tissue spots; Intestine Visium data: 2,807 tissue spots; AD ST data: n=5,009 tissue spots.
Data exclusions	We did not exclude any samples in the analysis.
Replication	In the simulation study, we took 10 replicates for each experiment, estimated the variances of outcomes and visualized them in results (related to Fig. 1 and Supplementary Fig. 1). In AD data, 4 replicated brain tissue samples were collected from two mice (2 samples from each mouse) for 3- and 18-month time-points (Supplementary Fig. 6a). Tissue spots from the same conditions were well mixed (Supplementary Fig. 6c). The results from replications were illustrated with boxplots or violinplots.
Randomization	Randomization is not relevant to this study. No new datasets were collected in this study.
Blinding	For all datasets, only gene expressions and morphological features were input to the analyses; spatial positions/sample condition/disease status information was blind to analysis and was used as orthogonal evidences for result evaluations. Blinding is not relevant to other experiments as we did not design and generate new datasets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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