nature methods

Article

https://doi.org/10.1038/s41592-024-02184-y

TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses

Received: 24 April 2023

Accepted: 12 January 2024

Published online: 12 February 2024

Check for updates

Eric D. Sun [1], Rong Ma [2,3], Paloma Navarro Negredo[4], Anne Brunet [4,5,6] & James Zou [1]

Whole-transcriptome spatial profiling of genes at single-cell resolution remains a challenge. To address this limitation, spatial gene expression prediction methods have been developed to infer the spatial expression of unmeasured transcripts, but the quality of these predictions can vary greatly. Here we present Transcript Imputation with Spatial Single-cell Uncertainty Estimation (TISSUE) as a general framework for estimating uncertainty for spatial gene expression predictions and providing uncertainty-aware methods for downstream inference. Leveraging conformal inference, TISSUE provides well-calibrated prediction intervals for predicted expression values across 11 benchmark datasets. Moreover, it consistently reduces the false discovery rate for differential gene expression analysis, improves clustering and visualization of predicted spatial transcriptomics and improves the performance of supervised learning models trained on predicted gene expression profiles. Applying TISSUE to a MERFISH spatial transcriptomics dataset of the adult mouse subventricular zone, we identified subtypes within the neural stem cell lineage and developed subtype-specific regional classifiers.

Spatial transcriptomics technologies extend high-throughput characterization of gene expression to the spatial dimension and have been used to characterize the spatial distribution of cell types and transcripts across multiple tissues and organisms[1–6]. A major trade-off across all spatial transcriptomics technologies is between the number of genes profiled and the spatial resolution such that most imaging-based spatial transcriptomics technologies with single-cell resolution are limited to the measurement of a few hundred genes but typically not the whole transcriptome[7]. Given the resource-intensive nature of single-cell spatial transcriptomics data acquisition, computational methods for upscaling the number of genes by predicting the expression of additional genes of interest are highly desirable.

There exist several methods for imputing or predicting spatial gene expression using a paired single-cell RNA-sequencing (RNA-seq) dataset. Generally, these methods proceed by joint embedding of the spatial transcriptomics and RNA-seq datasets and then predicting expression of new spatial genes by aggregating the nearest neighboring cells in the RNA-seq data[8–11] or by joint probabilistic modeling, mapping or transport[6,12–16]. For example, SpaGE relies on joint embedding of spatial transcriptomics and RNA-seq data using PRECISE domain adaptation followed by *k*-nearest-neighbors regression[8,17]; a method referred to as 'Harmony' here relies on Harmony integration of the two data modalities and averaging of nearest cell expression profiles[10]; and Tangram uses an optimal transport framework with deep learning to devise a mapping between single-cell and spatial

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Department of Statistics, Stanford University, Stanford, CA, USA. [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]Department of Genetics, Stanford University, Stanford, CA, USA. [5]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA. [6]Glenn Center for the Biology of Aging, Stanford University, Stanford, CA, USA. e-mail: jamesz@stanford.edu

transcriptomics data[13]. Applications of these methods have been used in the characterization of spatial differences in aging of mouse neural and glial cell populations[10], recovery of immune signatures in primary tumor samples[14] and identification of spatial patterns in gene regulation[13].

Because the relative performance of these models varies substantially depending on the application area and underlying datasets, there is no best model across all use cases[7]. Moreover, variability in model performance may adversely affect downstream analysis, particularly in promoting false discoveries due to prediction errors. At the same time, few existing gene expression prediction methods provide uncertainty measures for the predicted expression profiles and there are no approaches for utilizing uncertainty in downstream analyses. As a result, it is often difficult to rely on predicted spatial gene expression profiles without extensive external validation or understanding of their context-specific uncertainties.

Here, we present TISSUE as a general wrapper framework around any spatial gene expression imputation or prediction model that produces well-calibrated uncertainty measures tailored to the context of each individual model and its use case. We show that TISSUE can be leveraged for improvements in various uncertainty-aware data analysis tasks including the calculation of prediction intervals, hypothesis testing, supervised learning (for example, cell-type classification and anatomic region classification), and clustering and visualization of spatial transcriptomics data. We further show that TISSUE can be used to identify new cell types and subtypes that have yet to be profiled using spatial transcriptomics.

Results

TISSUE: cell-centric variability and calibration scores

Spatial gene expression prediction generally relies on leveraging spatial transcriptomics and RNA-seq data from similar cell types. The RNA-seq data are used to impute the expression of genes not measured in the limited spatial transcriptomics panel and can recover up to whole-transcriptome coverage of genes (Fig. 1a). To motivate the need for uncertainty quantification, we benchmarked 3 popular spatial gene expression prediction methods (Harmony[10], SpaGE[8] and Tangram[13]) on 11 publicly available spatial transcriptomics datasets (spanning seqFISH[18], MERFISH[19], STARmap[20], ISS[21], FISH[22], osmFISH[23], ExSeq[24] and spatial enhanced resolution omics-sequencing (Stereo-seq)[25] technologies; spatial data are visualized in Extended Data Fig. 1a); paired with single-cell or single-nuclei RNA-seq datasets (spanning Smart-seq, Drop-seq and 10x Chromium technologies) from the same organism and tissue regions[7,20,23–39] (Supplementary Table 1). No method consistently outperformed other methods across all spatial transcriptomics datasets. Similarly, methods that performed the best under one metric (for example, gene-wise Spearman rank correlation between measured and predicted gene expression; Fig. 1b) did not necessarily perform the best under a different evaluation metric (for example, mean absolute error in predicted expression; Extended Data Fig. 1b). For a given method, there is also substantial heterogeneity in the relative performance of the model between genes and cells (Fig. 1b and Extended Data Fig. 1b,c), suggesting that accurate estimation of uncertainty in spatial gene expression predictions may improve confidence in interpretations and downstream analyses. We observed similar trends for gimVI[12], an independent spatial gene expression prediction method (Supplementary Fig. 1a,b).

Conformal inference is a statistically rigorous and distribution-free framework for quantifying uncertainty of black-box models[40–42]. Traditionally, conformal inference proceeds by fitting a machine learning model on labeled training data, evaluating the model predictions on a small amount of labeled calibration data to build calibrated uncertainties, and then deploying the model on unlabeled test data to obtain both the predicted labels and their uncertainty. Conformal inference has been used to quantify uncertainty of region segmentations in tissue image analysis[43], measure confidence of drug discovery predictions[44] and understand the robustness of clinical treatment effects[45]. To extend the traditional conformal inference framework to spatial gene expression prediction, we made several key modifications to build well-calibrated uncertainties in TISSUE (Methods). First, we established an initial measure of prediction uncertainty that is scalable to unseen observations and agnostic to the prediction error. To calibrate these uncertainties to the prediction error, we built distributions of calibration scores by linking these initial measures of uncertainty to the observed prediction errors on existing genes in the spatial transcriptomics data. Finally, these calibration score distributions were used for computing well-calibrated prediction intervals and improving downstream spatial transcriptomics data analysis.

To construct an initial measure of uncertainty that can be universally applied to all existing spatial gene expression prediction methods, we posited that, on average, spatially proximate cells with similar measured gene expression profiles will also have similar expression of genes that are not measured in the spatial transcriptomics gene panel (see Extended Data Fig. 2 for empirical observations supporting this assumption). As a result, large differences in predicted gene expression between neighboring cells of the same cell type would indicate low predictive performance, and highly similar predicted gene expressions between neighboring cells would signify high predictive performance for the spatial gene expression prediction method. To quantify this intuition, we introduce the cell-centric variability measure U_{ij} which, for given a gene, computes for each cell a weighted measure of deviation between the predicted expression of that cell and those of the cells within a spatial neighborhood of it (equations (1) and (2)).

$$U_{ij} = 1 + \sqrt{\frac{\sum_{k \in N_i} W_{ik} (\hat{X}_{kj} - \hat{X}_{ij})^2}{\sum_{k \in N_i} W_{ik}}} \tag{1}$$

$$W_{ik} = \exp\left(\frac{\hat{X}_{i,:} \cdot \hat{X}_{k,:}}{||\hat{X}_{i,:}|| ||\hat{X}_{k,:}||} \right) \tag{2}$$

Here, $\hat{X}_{ij}$ is the predicted gene expression of cell i and gene j. For a given cell i, its spatial neighborhood N_i corresponds to the K closest cells in the spatial transcriptomics data according to Euclidean distance. For all experiments, we use $K = 15$, but the cell-centric variability is generally robust to different choices of K (Extended Data Fig. 3a; see Methods for additional justifications). For each neighboring cell k, we compute a weight W_{ik} equal to the exponential cosine similarity in predicted gene expression profiles between the central cell i and its neighbor. These weights prioritize variability in predicted gene expression for similar cells (for example, cells of the same cell type) and downplay expected variability in gene expression from dissimilar cells without the need to explicitly define cell types or states.

The cell-centric variability is generally positively correlated with the absolute prediction error for spatial gene expression (Fig. 1c). However, the cell-centric variability does not provide an exact estimate of the magnitude of these errors and the relationship between these two quantities is highly context dependent (Extended Data Fig. 3b). To explicitly link the cell-centric variability to the prediction error, we adapt a conformal inference framework by computing the calibration score, which is defined as the ratio between the absolute prediction error and the cell-centric variability according to equation (3) (Fig. 1a):

$$s_{ij} = \frac{|X_{ij} - \hat{X}_{ij}|}{U_{ij}}, \tag{3}$$

where X_{ij} denotes the measured gene expression for cell i and gene j.

https://doi.org/10.1038/s41592-024-02184-y

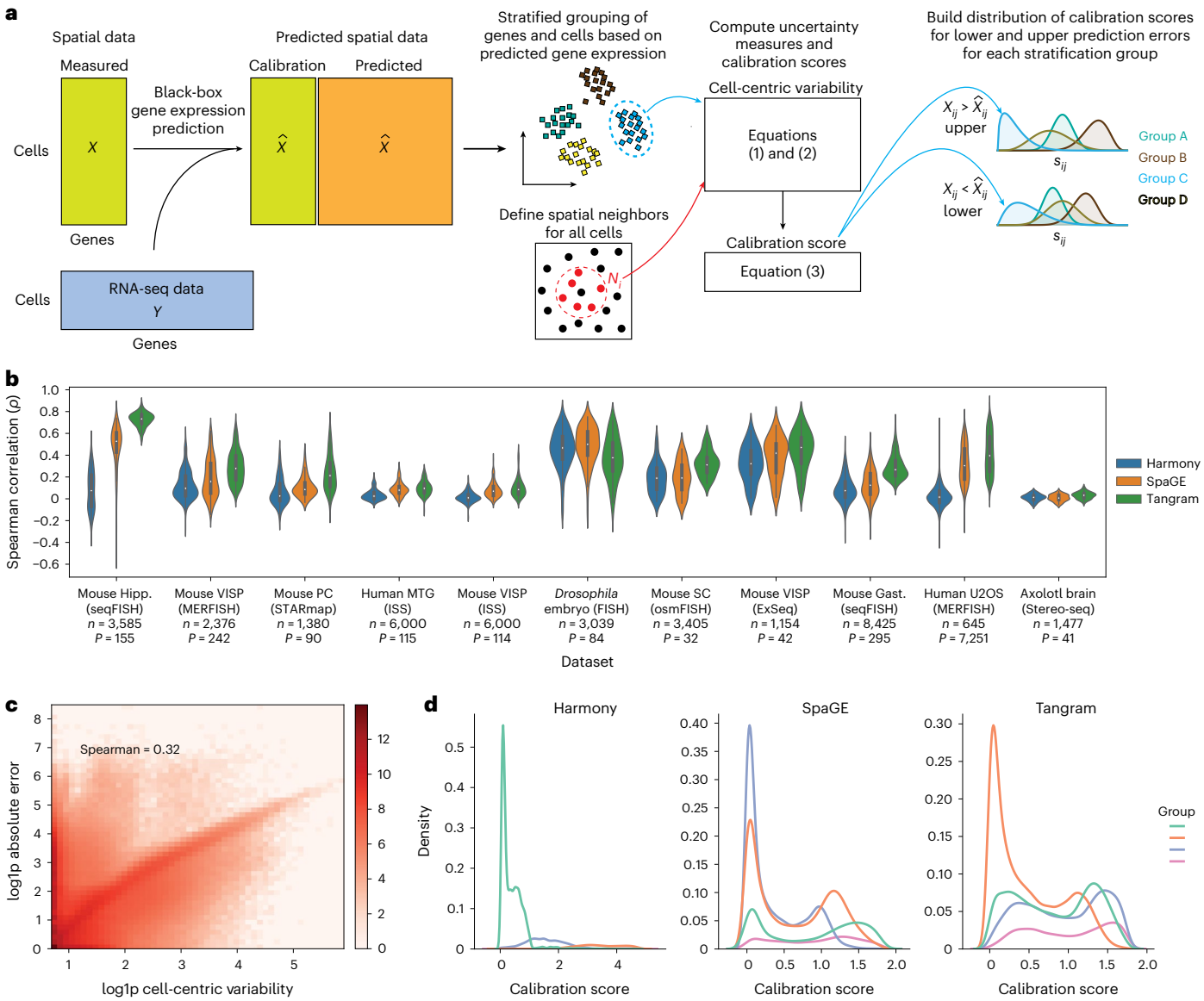


Fig. 1 | Cell-centric variability and calibration scores for conformal inference. **a**, Schematic of the TISSUE calibration score generation pipeline with black-box gene prediction (shown is an example method using paired spatial and RNA-seq datasets), stratified grouping of genes and cells, calculation of cell-centric variability measure, and computation and allocation of the calibration score to different stratified groups. **b**, Performance of three popular gene prediction methods (Harmony, SpaGE, Tangram) on 11 benchmark datasets as measured by the gene-wise Spearman correlation between predicted and actual gene expression over 10-fold cross-validation. Also shown are the number of cells (n) in the spatial transcriptomics datasets and the number of genes (P) shared between spatial and RNA-seq datasets. The inner box corresponds to quartiles of the correlation measures, and the whiskers span up to 1.5 times the interquartile range of the correlation measures. Hipp., hippocampus; VISP, primary visual cortex; PC, prefrontal cortex; MTG, middle temporal gyrus; SC, somatosensory cortex; Gast., gastrulation; U2OS, U-2 OS cell line. **c**, Correlation of TISSUE cell-centric variability and absolute prediction error across all dataset and prediction method combinations computed over 10-fold cross-validation. Log density with added pseudocount (log1p) is shown by color, with a maximum of 1,000 cells and 300 genes sampled from each dataset to provide more uniform representation. **d**, Distribution of TISSUE calibration scores on mouse hippocampus ISS dataset and all three prediction method combinations using $(k_g, k_c) = (4, 1)$. Details of each dataset and prediction method are available in Methods.

The distribution of s_{ij} can subsequently be used to calibrate uncertainties for new expression predictions by multiplying the cell-centric variability of those predictions by specific quantiles of the s_{ij} calibration score set, which returns values on the scale of prediction errors (see below for details). To permit flexible calibration schemes within the same spatial transcriptomics dataset, TISSUE allocates calibration scores to disjoint groups of genes and cells, referred to as the stratified calibration sets or groups, which are determined by k-means clustering of genes and then k-means clustering of cells by predicted gene expression (Fig. 1a and Methods). This stratified grouping scheme is motivated by the observation that there is generally positive correlation in pairwise similarities of predicted expression and of prediction error (Extended Data Fig. 3c). The number of gene and cell subsets (k_g, k_c) can be user-specified or determined using an automated method (Methods), but downstream results are generally robust to exact specifications of these stratified groupings. Empirically, the distribution of calibration scores can vary substantially across different identified subsets, suggesting the identification of heterogeneous calibration score sets (Fig. 1d and Extended Data Fig. 3d with $(k_g, k_c) = (4, 1)$). Due to the asymmetric distribution of transcript counts, the calibration scores are further separated by the sign of the prediction error into a lower set for $X_{ij} - \hat{X}_{ij} < 0$ and upper set for $X_{ij} - \hat{X}_{ij} > 0$ (Fig. 1a).

https://doi.org/10.1038/s41592-024-02184-y

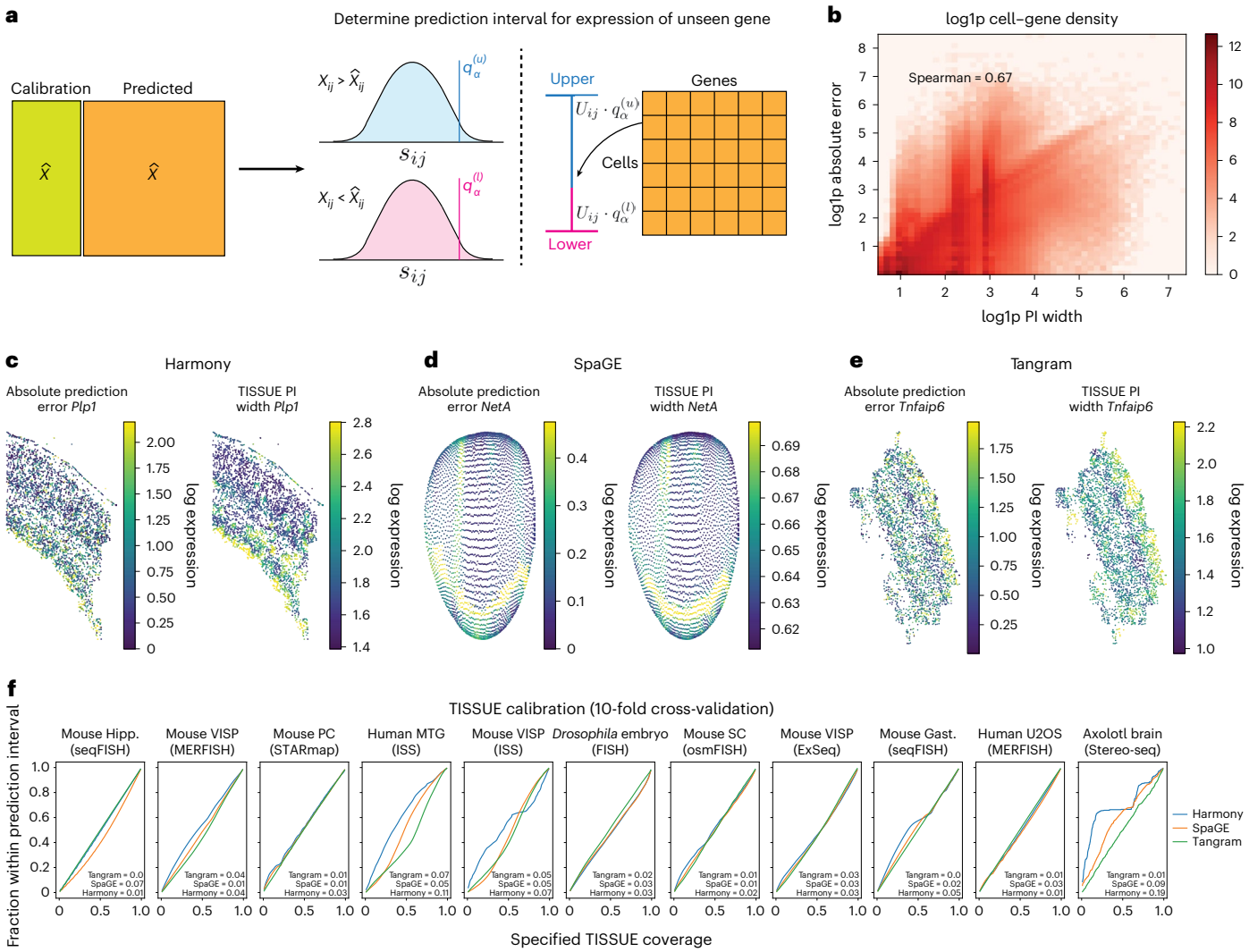


Fig. 2 | Prediction intervals for spatial gene expression. a, Schematic illustration of the TISSUE prediction interval retrieval process from the calibration scores for a given confidence level. **b**, Correlation of the 67% prediction interval (PI) width and the absolute prediction error across all dataset and prediction method combinations computed over 10-fold cross-validation. Log density with added pseudocount (log1p) is shown by color, with a maximum of 1,000 cells and 300 genes sampled from each dataset to provide more uniform representation. **c**–**e**, Comparison of absolute prediction error (left) and the 67% prediction interval width (right) for a representative gene in the mouse somatosensory cortex osmFISH dataset (**c**), in the virtual *Drosophila* embryo spatial transcriptomics dataset (**d**) and in the mouse primary visual cortex MERFISH dataset (**e**). **f**, Calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation. The calibration error is annotated for each prediction method (Methods). All prediction intervals were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping.

TISSUE prediction intervals for predicted gene expression

We leveraged a conformal inference framework to convert cell-centric variability of new spatial gene expression predictions into well-calibrated prediction intervals using the calibration scores derived from the measured gene panel. Given a gene expression prediction for cell a and gene b and confidence level α, we compute the cell-centric variability U_{ab} and multiply it by the $(\lceil (m+1)(1-\alpha) \rceil / m)$-th quantile of the upper and lower calibration sets of s_{ij} corresponding to all m predicted expression values from the same stratified group as the predicted expression of cell a and gene b, which yields the asymmetric TISSUE prediction interval with approximate $1-\alpha$ coverage (Fig. 2a). Using this approach, TISSUE prediction intervals can be obtained for every predicted gene expression and every cell in the spatial transcriptomics data (see Methods for details and mathematical guarantees). The TISSUE prediction interval width is positively correlated with the absolute prediction error of measured genes under cross-validation (Fig. 2b). This trend persists after normalizing by the magnitude of predicted expression (Extended Data Fig. 4a) and also exists for different choices of α for computing prediction interval widths (Extended Data Fig. 4b,c). For individual genes of interest, the TISSUE prediction interval width generally reflects the spatial distributions of absolute prediction errors such as for *Plp1* in osmFISH profiling of mouse somatosensory cortex (Fig. 2c), *NetA* in a virtual spatial transcriptomics profile of *Drosophila* embryo (Fig. 2d) and *Tnfaip6* in MERFISH profiling of mouse primary visual cortex (Fig. 2e). Averaged across all genes and cells, the TISSUE prediction interval provides well-calibrated coverage of prediction errors on unseen genes for a broad range of confidence levels and across all prediction methods and spatial transcriptomics datasets (Fig. 2f). For individual genes, there is a general tendency toward well-calibrated prediction intervals (Extended Data Fig. 4d). Similar calibration quality for TISSUE prediction intervals was observed under automated selection of k_g and k_c (Extended Data Fig. 4e), and when tested on gimVI, a deep generative model for spatial gene expression prediction[12] (Supplementary Fig. 1c,d). The calibration quality of

TISSUE was also highly reproducible across technical replicates within a spatial transcriptomics dataset (Extended Data Fig. 4f).

We used Sprod, a framework for denoising spatially resolved transcriptomics[46], and the mouse somatosensory cortex osmFISH dataset to investigate whether TISSUE calibration would be affected by spatial denoising or alternative formulations of the TISSUE neighborhood graph for computing cell-centric variability. Across different combinations of preprocessing (with or without Sprod denoising) and neighborhood graphs (TISSUE or Sprod cell similarity graph), TISSUE calibration quality was comparable to that from the default TISSUE settings (Extended Data Fig. 4g), suggesting that the TISSUE framework is likely to be robust to denoising and alternative definitions for cell neighborhoods. Similarly, the TISSUE prediction interval widths were highly correlated between the default TISSUE neighborhood graph and the alternative Sprod neighborhood graph (Extended Data Fig. 4h).

Uncertainty-aware hypothesis testing with TISSUE

Hypothesis testing of differences in gene expression between experimental conditions, cell types or other groupings is an important tool in understanding biological heterogeneity and perturbation effects using spatial transcriptomics. We extend TISSUE calibration scores for more robust hypothesis testing of differential predicted gene expression across conditions. Specifically, TISSUE hypothesis testing involves sampling multiple imputations for the predicted gene expression values by first sampling calibration scores $\hat{s}_{ij}$'s from corresponding calibration sets and then perturbing the original predicted expression values by $U_{ij} \times \hat{s}_{ij}$ with the direction of the perturbation dependent on whether the sampled score was in the upper or lower set (Fig. 3a). Repeating this process D times yields D possible imputations for each cell and gene. Using multiple imputation theory, TISSUE then derives corrected measures of statistical significance using a modified independent two-sample t-test (Fig. 3a and Methods). These corrected statistics account for the uncertainty in prediction as encoded by the sampling of scores for generating new imputations. This multiple imputation framework can be extended to other statistics of interest[47–50].

To compare TISSUE hypothesis testing to traditional hypothesis testing using only the predicted gene expression values, we generated synthetic data using SRTsim[51] in which there are two groups of cells with the same ground truth gene expression (see Methods for simulation settings). To simulate spatial gene expression prediction, we added biased Gaussian noise with mean μ to a portion of the genes in one of the cell groups but not the other, and standard Gaussian noise to all other gene expression values. Under this context, TISSUE hypothesis testing exhibited lower error rate in calling differentially expressed genes between the two cell groups (automated stratified grouping, Benjamini–Hochberg corrected P value cutoff for false discovery rate (FDR) = 5%) across different levels of prediction bias than traditional hypothesis testing (Fig. 3b).

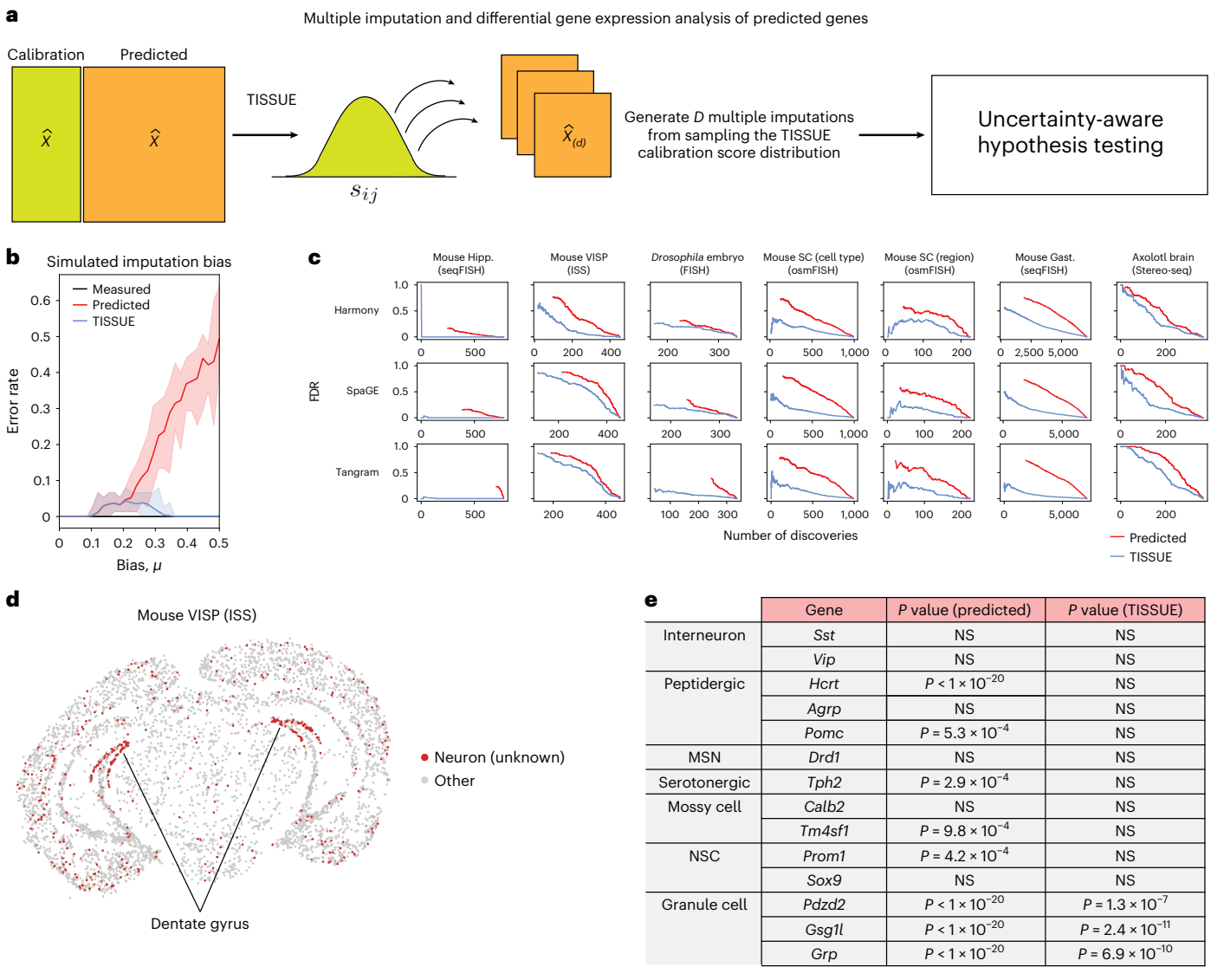
To further evaluate TISSUE hypothesis testing, we compared it to the traditional hypothesis testing approach on seven publicly available spatial transcriptomics datasets with associated cell type or anatomic region labels (see Methods for details on data labeling). For each label, we computed the statistical significance of gene expression differences within that label as compared to all cells with different labels (that is, one-versus-all approach). Statistical significance was assessed for all genes in the measured gene expression with traditional hypothesis testing and in the predicted gene expression with both TISSUE and traditional hypothesis testing. Using the differentially expressed genes detected using measured gene expression values as the ground truth, we observed a lower FDR of differentially expressed genes using TISSUE hypothesis testing as compared to traditional hypothesis testing across all prediction methods and datasets. The lower FDR was observed across different numbers of differentially expressed gene detections (Fig. 3c) and across different P value cutoffs (Extended Data Fig. 5a). This decrease in FDR was also observed for automated selection of k_g and k_c (Extended Data Fig. 5b). We also observed reduced FDR when using TISSUE with gimVI gene expression predictions (Supplementary Fig. 1e). The TISSUE multiple imputation framework can also be extended to non-parametric hypothesis testing and to spatially variable gene detection with SpatialDE[52], resulting in similar reductions in FDRs when using predicted gene expression profiles as input, especially when the number of intended discoveries is low (Extended Data Fig. 5c,d). For example, TISSUE selectively detected spatially variable expression of *Unc13c* in the mouse primary visual cortex ($P = 6.6 \times 10^{-5}$ for TISSUE SpatialDE on SpaGE predicted expression, $P = 0.08$ for SpatialDE on baseline SpaGE predicted expression), which encodes a protein that is linked to synaptic plasticity[53] and neuroprotection in Alzheimer's disease[54] and has not been previously identified as a spatially variable gene in this brain region. TISSUE hypothesis testing was also reproducible across different replicates within the same spatial transcriptomics dataset (Extended Data Fig. 5e). As such, application of TISSUE hypothesis testing robustly guards against false discoveries when performing differential gene expression analysis with predicted spatial gene expression profiles.

To illustrate a specific use case of TISSUE hypothesis testing, we applied the method to an in situ sequencing (ISS) mouse primary visual cortex dataset. Using unbiased Leiden clustering, we identified several broad neuronal cell-type clusters along with specific nonneuronal cell-type clusters. We were unable to further resolve the neuronal cell clusters and used spatial gene prediction with SpaGE to predict the expression of additional neuronal subtype markers that were not in the original ISS gene panel. For one neuronal cluster, which localized to the dentate gyrus (DG) of the hippocampus (Fig. 3d), differential expression was detected across most neuronal subtype markers, but under TISSUE hypothesis testing, we observed that this cluster had selective differential expression of predicted marker genes associated with granule cells (*Pdzd2*, *Gsg1l*, *Grp*), which are concentrated in the DG of the hippocampus[55], and no significant expression of other cell-type markers, including those for mossy cells (*Calb2*, *Tm4sf1*) and neural stem cells (*Prom1*, *Sox9*), which are other cell types found in the DG[55], and those for other neuronal subtypes (*Sst*, *Vip*, *Hcrt*, *Agrp*, *Pomc*, *Drd1*, *Tph2*; Fig. 3e). The identification of this neuronal cell cluster as a granule cell cluster was confirmed by spatial localization of these cells to the hippocampal DG and by further confirmation with measured gene marker *Lrrtm4*, which encodes a protein that has been previously implicated with granule cell processes[56]. Under traditional differential expression testing with the predicted spatial gene expression values, we were unable to recover the same specificity for granule cell markers. We observed similar but reduced trends for Tangram prediction and lack of significance for any markers under Harmony prediction, likely due to the low performance of that model on this dataset (Fig. 1b) and suggesting that comparison of TISSUE differential expression results from multiple prediction methods is advantageous.

Uncertainty-aware downstream analyses with TISSUE

Supervised learning is a common practice with single-cell and spatial transcriptomics data and can lead to useful models for predicting quantities of interest such as biological age[57], cell cycle state[58] and perturbational responses[59]. Similarly, cell clustering and visualization are commonly used to identify cell types in spatial transcriptomics data and intuitively understand high-dimensional differences between different groups. Substantial errors in spatial gene expression prediction may adversely affect the performance of these downstream tasks when relying on the predicted gene expression profiles as input. Here we introduce TISSUE cell filtering as an approach for retrieving a high-quality subset of predictions to be used as input for improved downstream training and evaluation of supervised classification models, clustering of cells and data visualization via dimensionality reduction. TISSUE cell filtering involves ranking of cells by the magnitude of

https://doi.org/10.1038/s41592-024-02184-y



e

	Gene	*P* value (predicted)	*P* value (TISSUE)
Interneuron	*Sst*	NS	NS
	Vip	NS	NS
Peptidergic	*Hcrt*	$P < 1 \times 10^{-20}$	NS
	Agrp	NS	NS
	Pomc	$P = 5.3 \times 10^{-4}$	NS
MSN	*Drd1*	NS	NS
Serotonergic	*Tph2*	$P = 2.9 \times 10^{-4}$	NS
Mossy cell	*Calb2*	NS	NS
	Tm4sf1	$P = 9.8 \times 10^{-4}$	NS
NSC	*Prom1*	$P = 4.2 \times 10^{-4}$	NS
	Sox9	NS	NS
Granule cell	*Pdzd2*	$P < 1 \times 10^{-20}$	$P = 1.3 \times 10^{-7}$
	Gsg1l	$P < 1 \times 10^{-20}$	$P = 2.4 \times 10^{-11}$
	Grp	$P < 1 \times 10^{-20}$	$P = 6.9 \times 10^{-10}$

Fig. 3 | Uncertainty-aware differential gene expression analysis with TISSUE. a, Schematic illustration of the TISSUE multiple imputation pipeline for hypothesis testing. Calibration scores are randomly sampled and used to compute new predicted gene expression profiles and statistics are compiled across all imputations using a modified two-sided *t*-test. **b**, Error rate for testing of significant gene expression differences between two homogeneous groups of cells ($n = 980$, $P = 1{,}000$) as a function of the selective bias in prediction error in approximately half of the genes for one group of cells using a Benjamini–Hochberg correction for 5% FDR. Shown are error rates for two-sided *t*-test on the measured gene expression profiles (black), two-sided *t*-test on predicted gene expression (red) and modified two-sided *t*-test using the TISSUE multiple imputation approach on predicted gene expression (blue). Results were obtained using automated stratified grouping (Methods). Bands represent the range, and the solid line denotes the mean error rate measured across 20 simulations. **c**, FDR of differentially expressed genes between cell type or anatomic region labels (one-versus-all approach) using the differentially expressed genes on the measured gene expression profiles as the ground truth across different numbers of discoveries. Discoveries are assessed across all genes for all class labels. Shown are results for all three prediction methods and all spatial transcriptomics datasets with cell type or region labels available. All calibration scores were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping. **d**, Mouse primary visual cortex ISS dataset with the unknown neuronal cluster colored in red and spatially localized to the DG of the hippocampus. **e**, Differential expression of neuronal marker genes using traditional hypothesis testing with two-sided *t*-test on the predicted gene expression (predicted) or using TISSUE hypothesis testing with modified two-sided *t*-test with multiple imputation (TISSUE). *P* values are shown for all predicted marker genes with significance threshold of Bonferroni-adjusted $P < 0.05$. All calibration scores were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping. MSN, medium spiny neuron; NSC, neural stem cell.

uncertainty (that is, prediction interval width) for each gene, followed by automated filtering of cells with the highest uncertainty ranking within each class label (for example, cell type; Fig. 4a and Methods).

To compare the performance of TISSUE cell filtering to traditional approaches for downstream tasks, we generated synthetic spatial transcriptomics data using SRTsim[51] with two distinct cell-type clusters where half of the profiled genes are higher in expression in one cell type (Methods). Predicted gene expression was simulated by selectively adding mix-in bias to a proportion of cells in one cell type such that the expression profiles of those cells resemble the ground truth of the other cell type, and zero-centered Gaussian prediction noise is added to all other cells (Methods and Fig. 4b). Under this simulation setting, supervised learning classification models trained and evaluated on TISSUE-filtered data generally outperformed classifiers trained and evaluated on unfiltered spatial transcriptomics data in separating the two cell groups across different levels of mix-in prediction bias and

https://doi.org/10.1038/s41592-024-02184-y

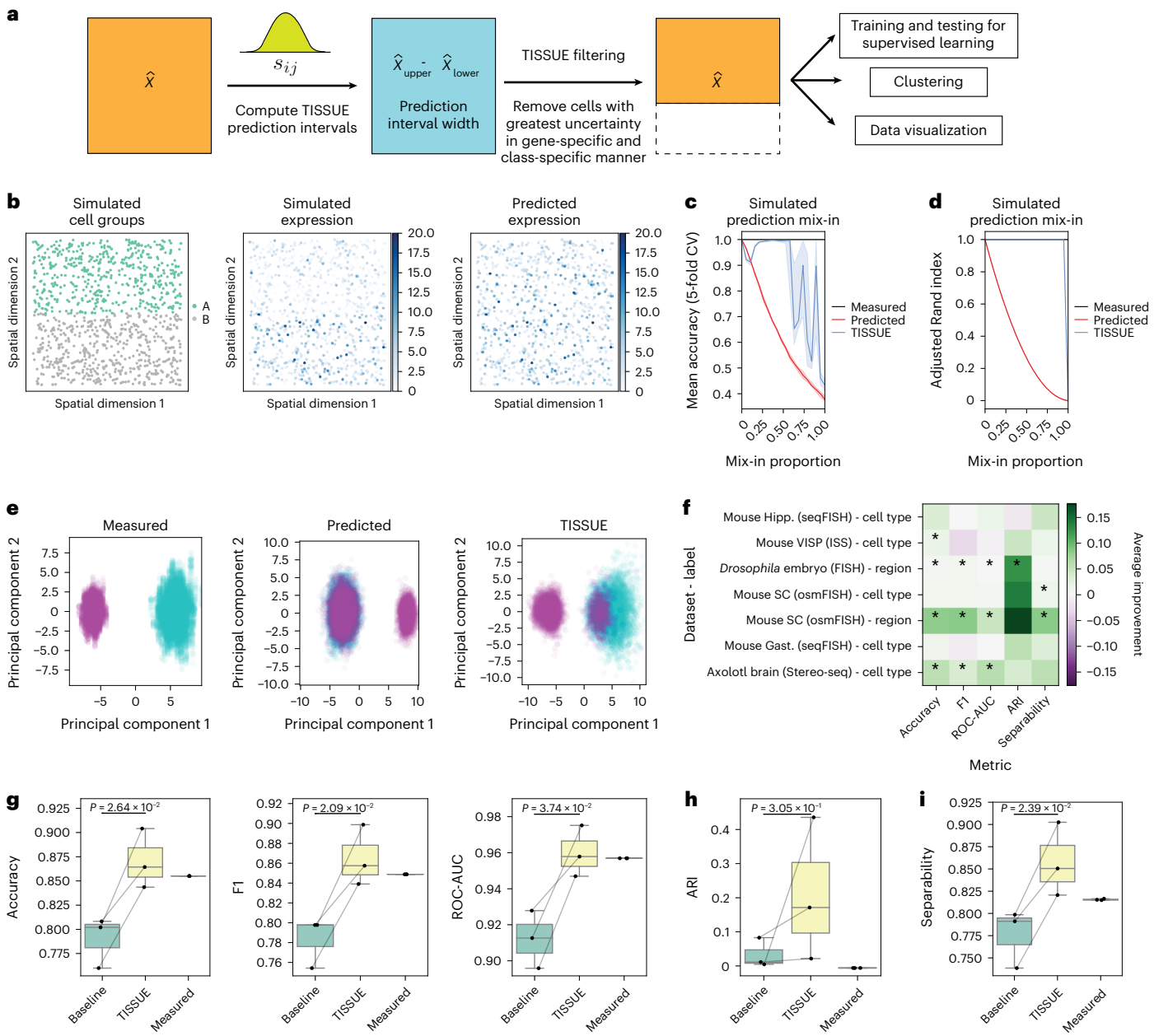


Fig. 4 | Uncertainty-aware supervised learning, clustering and visualization. **a**, Schematic illustration of the uncertainty-aware TISSUE cell filtering framework for downstream data analysis. **b**, Spatial visualization of the simulated dataset by two cell clusters (left), measured expression of a gene (middle) and predicted expression with mix-in bias of the same gene (right). **c**, Mean accuracy for logistic regression models trained to classify the two synthetic cell clusters as a function of mix-in bias. Accuracy was measured across 5-fold cross-validation (CV) for models trained on measured (black), predicted (red) and TISSUE-filtered predicted (blue) gene expression values. We used automated stratified grouping. Bands represent the interquartile range, and the solid line denotes median performance measured across 20 simulated datasets. **d**, ARI of k-means clustering on the top 15 principal components with $k = 2$ on the simulated dataset as a function of mix-in bias. Colors and bands are defined as in **c**. **e**, DynamicViz PCA plots (Methods) from measured gene expression profiles, predicted gene expression profiles and predicted gene expression profiles after TISSUE cell filtering with 90% mix-in of the two cell clusters (different colors). **f**, Average improvement of performance using TISSUE-filtered approach over unfiltered approach on predicted expression for supervised learning (accuracy, F1, receiver-operator characteristic area under the curve (ROC-AUC)), clustering (ARI) and visualization (linear separability) for the top three classes across all dataset and class label combinations. Asterisks denote a significant difference in performance metrics between TISSUE-filtered approach and unfiltered approach ($P < 0.05$) with P values computed using a paired two-sided t-test on $n = 3$ independent prediction methods. **g**–**i**, Downstream task performance metrics on the three most prominent anatomic region class labels in the mouse somatosensory osmFISH dataset. Shown are metrics for the predicted gene expression (baseline), TISSUE-filtered predicted gene expression or measured gene expression: accuracy, F1 score and ROC-AUC metrics for supervised learning (**g**); ARI for clustering (**h**); and linear separability for visualization (**i**). We set $(k_g, k_c) = (4, 1)$. P values were computed with a paired two-sided t-test on $n = 3$ independent prediction methods. The box outlines quartiles of the metrics, and whiskers span up to 1.5 times the interquartile range.

under cross-validation (Fig. 4c). Similarly, clustering of cells (k-means with $k = 2$ on the top 15 principal components) using the TISSUE-filtered gene expression predictions resulted in higher-quality clustering than clustering of cells using the unfiltered gene expression predictions as evidenced by higher adjusted Rand index (ARI) with respect to the true cell clusters across different levels of mix-in prediction bias (Fig. 4d).

https://doi.org/10.1038/s41592-024-02184-y

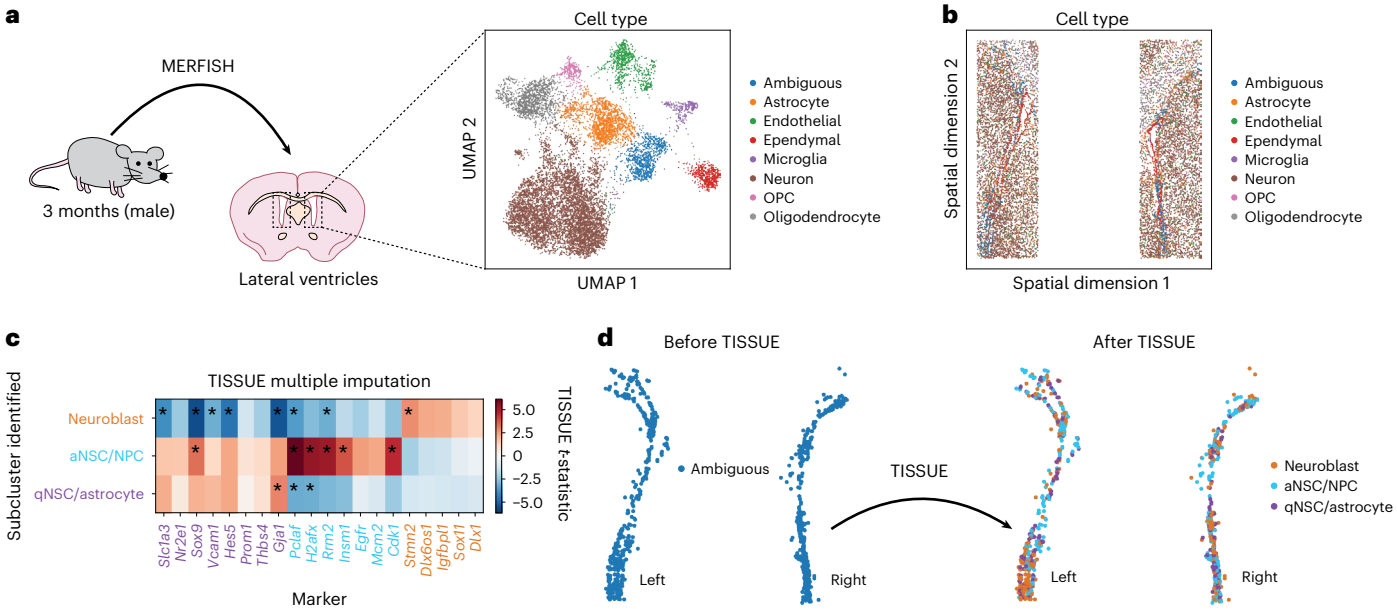


Fig. 5 | TISSUE discovers subtypes in neural stem cell lineage of the SVZ. **a**, Schematic of the 140-gene MERFISH dataset generated on adult mouse SVZ samples from both lateral ventricles (left) with uniform manifold approximation and projection (UMAP) visualization of cells colored by identified cell-type clusters including one ambiguous cell cluster (right). **b**, Spatial visualization of cells colored by identified cell-type clusters including one ambiguous cell cluster. **c**, Heat map of *t*-statistic values from the TISSUE multiple imputation two-sided *t*-test on the SpaGE predicted expression of new marker genes within cells of the three ambiguous cell subclusters compared to all other cells in the ambiguous cell cluster. Red boxes correspond to overexpression of that marker in that subcluster, while blue boxes correspond to underexpression of that marker in that subcluster. Boxes with asterisks denote expression differences with Bonferroni-adjusted $P < 0.05$ using the TISSUE multiple imputation two-sided *t*-test. Colored text corresponds to the identified cell subtypes and their markers, with some qNSC/astrocyte markers also known to be expressed in aNSCs/NPCs. **d**, Spatial visualization of the ambiguous cell-type cluster before the application of TISSUE (left) and then of the three TISSUE-identified subtypes of the ambiguous cell cluster including neuroblasts, aNSCs/NPCs and qNSCs/astrocytes (right). The left and right lateral ventricles are annotated.

To assess improvements in low-dimensional visualization of the data, we used DynamicViz[60] to rigidly align cells by their top two principal components across 20 independent simulations and observed that TISSUE-filtered visualizations were better able to separate the two cell groups while the unfiltered visualizations were unable to do so under 50% mix-in of the two cell groups (Fig. 4e). The DynamicViz variance score was also lower for the TISSUE-filtered visualization than for the unfiltered principal component analysis (PCA) visualization (median variance score of 0.198 compared to 0.381), indicating more stable visualization quality in the former, likely due to the improved representation of differences between the two cell-type clusters.

To assess improvements by TISSUE cell filtering on publicly available spatial transcriptomics datasets, we curated seven pairings of datasets and class labels (for example, cell type or anatomic region) and restricted our analyses to the three labels with greatest representation within each pairing. To evaluate supervised learning for classification, we compared the cross-validated performance of logistic regression models trained and evaluated on the TISSUE-filtered predicted spatial gene expression to the performance of logistic regression models trained and evaluated on the unfiltered predicted gene expression profiles. Across three different performance metrics (accuracy, area under the receiver-operator characteristic curve and F1 score), the TISSUE-filtered classifiers generally outperformed the unfiltered classifiers on prediction tasks (Fig. 4f), particularly for the osmFISH mouse somatosensory cortex dataset with region labels, where classification performance was comparable to that of models trained on the measured gene expression profiles (Fig. 4g). Similarly, clustering quality and visualization quality were generally improved by TISSUE cell filtering as evidenced by higher ARI with respect to the class labels and by higher linear separability of classes in low-dimensional PCA representations of the predicted gene expression, which was measured by fitting a support vector classifier with a linear kernel to the top 15 principal components (Fig. 4f,h,i).

Across all tasks, TISSUE cell filtering provided similar improvements when used with automated stratified grouping (Extended Data Fig. 6a–c), different prediction interval settings (Extended Data Fig. 6d,e) and gimVI predictions (Supplementary Fig. 1f). All benchmarks for evaluating TISSUE performance are summarized in Supplementary Table 2. For clustering and visualization tasks, we also considered an alternative framework to TISSUE cell filtering, where instead of filtering cells, we leveraged weighted principal component analysis (WPCA)[61] with weights related to the inverse TISSUE prediction interval width for each gene expression prediction (Extended Data Fig. 7a and Methods). Using the TISSUE-WPCA approach to obtain principal components improved linear separability between cell clusters on the synthetic datasets for a range of mix-in bias levels (Extended Data Fig. 7b,c) and improved clustering on several real spatial transcriptomics datasets (Extended Data Fig. 7d).

TISSUE resolves cell types of the neural stem cell lineage

The subventricular zone (SVZ) neurogenic niche is located in the lateral ventricles of the adult mammalian brain and is resident to neural stem cells that are important in homeostasis and for injury response and repair[62–64]. In addition to many other cell types, the SVZ contains cells of the neural stem cell lineage, which consists of neural stem cells (quiescent and activated subtypes), neural progenitor cells (NPCs) and neuroblasts, all of which have yet to be identified using spatial transcriptomics of the mammalian brain. To test the ability of TISSUE to identify and characterize these cell types, we used MERFISH to profile the spatial expression of 140 genes on two young adult mouse brain sections containing both lateral ventricles (Fig. 5a). We performed clustering on the data and using known marker genes, we identified several cell

types including astrocytes, endothelial cells, ependymal cells, microglia, neurons, oligodendrocyte progenitor cells and oligodendrocytes, along with an ambiguous cell cluster that localized to the lateral ventricles (Fig. 5a,b). Although the MERFISH gene panel contained several known transcriptomic markers for quiescent neural stem cells (qNSCs/astrocytes), activated neural stem cells (aNSCs)/NPCs and neuroblasts (Methods), they were insufficient for resolving the ambiguous cell cluster further into these subtypes (Extended Data Fig. 8a). As such, we used SpaGE and a single-cell RNA-seq dataset of the adult mouse SVZ[57] to predict additional genes that were not present in the original panel, including general NSC and qNSC/astrocyte markers (*Slc1a3*, *Nr2e1*, *Sox9*, *Vcam1*, *Hes5*, *Prom1*, *Thbs4*), aNSC/NPC markers (*Pclaf*, *H2ax*, *Rrm2*, *Insm1*, *Egfr*, *Prom1*, *Mcm2*, *Cdk1*) and neuroblast markers (*Stmn2*, *Dlx6os1*, *Igfbpl1*, *Sox11*, *Dlx1*). After subclustering the ambiguous cluster and then leveraging TISSUE multiple imputation and hypothesis testing, we observed differential expression of marker genes, which identified qNSC/astrocyte, aNSC/NPC and neuroblast cell clusters (Fig. 5c). In particular, TISSUE multiple imputation and hypothesis testing was necessary to resolve the identity of the aNSC/NPC subcluster, which could not be resolved from the SpaGE-imputed expression values alone (Extended Data Fig. 8b). Similar TISSUE-specific improvements in identifying subclusters were observed for other spatial gene expression prediction methods, and only TISSUE methods could identify the aNSC/NPC subcluster (Extended Data Fig. 8c). Consistent with known biology of the SVZ, all three cell subtypes were found throughout both lateral ventricles (Fig. 5d) and the relative proportions of each subtype were similar between the right and left ventricles (Extended Data Fig. 8d). Additionally, there were slightly more neuroblasts than aNSCs/NPCs in the MERFISH dataset, which is also reflected among several independent single-cell RNA-seq datasets of the adult mouse SVZ (Extended Data Fig. 8e)[57,65,66].

Recent efforts have uncovered biological heterogeneity in neural stem cell populations between the dorsal and ventral regions of the SVZ[67,68]. However, these existing transcriptomic characterizations of the SVZ have relied on dissociated single-cell transcriptomics data, thus precluding analyses involving the ground truth spatial location of the neural stem cells without resource-intensive regional micro-dissections. Using the spatial locations of cells determined via MERFISH imaging, we categorized cells into dorsal, ventral or other regional classes using horizontal boundaries (Extended Data Fig. 8f). Using TISSUE multiple imputation and hypothesis testing, we then performed whole-transcriptome differential gene expression on dorsal or ventral categories within each of the three subtypes. For each cell subtype, we selected the 20 most differentially expressed genes and trained penalized logistic regression models to predict dorsal or ventral regional origin from predicted spatial gene expression with or without TISSUE cell filtering. In all cell subtypes, the TISSUE-filtered models outperformed the baseline unfiltered models across several classification metrics including F1 score, accuracy, area under the receiver-operator curve and average precision (Extended Data Fig. 8g). Application of these TISSUE-filtered dorsal/ventral region classifiers to dissociated single-cell RNA-seq data may provide a useful first-step estimation of the regional origin of cells in the neural stem cell lineage without the need for laborious regional micro-dissections of the niche.

Discussion

We developed TISSUE to compute well-calibrated and context-specific measures of uncertainty for predicted spatial gene expression profiles. TISSUE provides general frameworks for leveraging uncertainty in downstream analysis such as differential gene expression analysis, clustering and visualization, and supervised learning. These frameworks are flexible and can be adapted into existing spatial transcriptomics data analysis workflows. For example, the differential gene expression analysis approach can be adapted to other hypothesis tests, which we show for non-parametric two-sample tests and spatially variable gene detection (Extended Data Fig. 5c,d). Likewise, the principal components obtained from TISSUE cell filtering can be used for any downstream algorithms that utilize the reduced dimensionality representation of PCA as input. Finally, the TISSUE-based filtering of training and evaluation data for supervised learning strictly modifies the data in a model-agnostic manner and therefore can be extended to both training and deployment of any supervised learning model across both regression and classification tasks. TISSUE motivates the future development and benchmarking of uncertainty-aware versions of single-cell and spatial transcriptomics data analysis methods, such as for spatial domain detection[69], embedding for label transfer[70] and cross-modality transformation[71].

As a case study, we applied TISSUE to predict the expression of additional cell-type markers in the adult mouse SVZ MERFISH dataset. TISSUE multiple imputation and hypothesis testing were then used to successfully annotate cell subtypes in the neural stem cell lineage, which, to our knowledge, constitute the first identification of these cell subtypes in spatial transcriptomics. In both the SVZ MERFISH dataset and in the primary visual cortex ISS dataset, TISSUE was necessary to uniquely identify ambiguous cell clusters, which was not possible using only the predicted gene expression. Together, these results indicate that TISSUE may serve as a promising framework for identifying new or previously unprofiled cell types in spatial transcriptomics data when the measured gene panel is insufficient.

TISSUE may have limited performance in contexts where spatial gene prediction patterns are not represented in the calibration set and for genes with extremely sparse expression patterns. Due to the assumptions underlying TISSUE, there may also be reduced performance on rare cell types that are not spatially colocalized. Since TISSUE performance is dependent on the size and diversity of the calibration sets, the method will generally scale better to spatial transcriptomics datasets with many cells, many genes or uniform representation of cell types. Additional benchmarking with spatial transcriptomics data collected on disease samples or other organs may further test the robustness of TISSUE. The main computational burden imposed by TISSUE is the cross-validated prediction of gene expression on the calibration set, which is necessary for building context-specific uncertainties. The burden for computing cell-centric variability, calibration scores and prediction intervals is comparatively less than that for generating the initial predictions (Extended Data Fig. 9).

Although TISSUE has thus far been tested in the spatial transcriptomics setting, the underlying assumptions can generalize to other spatial data modalities, such as spatial proteomics. As other spatial omics technologies mature, we anticipate that TISSUE will find additional use in the prediction and quantification of uncertainty for enhanced spatial data analysis across multiple modalities.

Online content

Any methods, additional references, Nature Portfolio 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/s41592-024-02184-y.

References

1. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
2. Asp, M. et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* **179**, 1647–1660 (2019).
3. 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).

4. Ji, A. L. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**, 497–514 (2020).
5. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* https://doi.org/10.1038/s41592-022-01409-2 (2022).
6. Wei, R. et al. Spatial charting of single-cell transcriptomes in tissues. *Nat. Biotechnol.* **40**, 1190–1199 (2022).
7. Li, B. et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* **19**, 662–670 (2022).
8. Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M. J. T. SpaGE: spatial gene enhancement using scRNA-seq. *Nucleic Acids Res.* **48**, e107 (2020).
9. Shengquan, C., Boheng, Z., Xiaoyang, C., Xuegong, Z. & Rui, J. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* **37**, i299–i307 (2021).
10. Allen, W. E., Blosser, T. R., Sullivan, Z. A., Dulac, C. & Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194–208(2023).
11. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
12. Lopez, R. et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *ICML Workshop on Computational Biology* (2019).
13. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
14. Vahid, M. R. et al. High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01697-9 (2023).
15. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).
16. Moriel, N. et al. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat. Protoc.* **16**, 4177–4200 (2021).
17. Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics* **35**, i510–i519 (2019).
18. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
19. 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).
20. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
21. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
22. Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on *Drosophila* polytene chromosomes. *Proc. Natl Acad. Sci. USA* **79**, 4381–4385 (1982).
23. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
24. Alon, S. et al. Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, eaax2656 (2021).
25. Wei, X. et al. Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science* **377**, eabp9444 (2022).
26. Long, B., Miller, J. & Consortium, T. S. SpaceTx: a roadmap for benchmarking spatial transcriptomics exploration of the brain. Preprint at http://arxiv.org/abs/2301.08436 (2023).
27. Joglekar, A. et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat. Commun.* **12**, 463 (2021).
28. Booeshaghi, A. S. et al. Isoform cell-type specificity in the mouse primary motor cortex. *Nature* **598**, 195–199 (2021).
29. Gyllborg, D. et al. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Res.* **48**, e112 (2020).
30. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
31. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).
32. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
33. Yao, Z. et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222–3241 (2021).
34. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
35. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
36. Lohoff, T. et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.* **40**, 74–85 (2022).
37. Lust, K. et al. Single-cell analyses of axolotl telencephalon organization, neurogenesis, and regeneration. *Science* **377**, eabp9262 (2022).
38. Zhou, Y. et al. Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat. Commun.* **11**, 6322 (2020).
39. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl Acad. Sci. USA* **116**, 19490–19499 (2019).
40. Angelopoulos, A. N. & Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Preprint at http://arxiv.org/abs/2107.07511 (2022).
41. Shafer, G. & Vovk, V. A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9**, 371–421 (2008).
42. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018).
43. Wieslander, H. et al. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE J. Biomed. Health Informatics* **25**, 371–380 (2021).
44. Alvarsson, J., Arvidsson McShane, S., Norinder, U. & Spjuth, O. Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* **110**, 42–49 (2021).
45. Jin, Y., Ren, Z. & Candès, E. J. Sensitivity analysis of individual treatment effects: a robust conformal inference approach. *Proc. Natl Acad. Sci. USA* **120**, e2214889120 (2023).
46. Wang, Y. et al. Sprod for de-noising spatially resolved transcriptomics data based on position and image information. *Nat. Methods* **19**, 950–958 (2022).
47. Palmer, C. & Pe'er, I. Bias characterization in probabilistic genotype data and improved signal detection with multiple imputation. *PLoS Genet.* **12**, e1006091 (2016).
48. Allison, P. D. *Missing Data* https://methods.sagepub.com/book/missing-data (SAGE Publications, 2002).
49. Little, R. J. A. & Rubin, D. B. Bayes and Multiple Imputation. In *Statistical Analysis with Missing Data* (eds Little, R. J. A. & Rubin, D. B.) 200–220 (John Wiley & Sons, Inc., 2002); https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119013563.ch10

50. Licht, C. New methods for generating significance levels from multiply-imputed data. Ph.D. thesis, Otto-Friedrich-Universität Bamberg, Fakultät Sozial- und Wirtschaftswissenschaften https://fis.uni-bamberg.de/handle/uniba/263 (2010).
51. Zhu, J., Shang, L. & Zhou, X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biol.* **24**, 39 (2023).
52. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
53. Yang, C. B., Kiser, P. J., Zheng, Y. T., Varoqueaux, F. & Mower, G. D. Bidirectional regulation of Munc13-3 protein expression by age and dark rearing during the critical period in mouse visual cortex. *Neuroscience* **150**, 603–608 (2007).
54. Miller, J. A., Woltjer, R. L., Goodenbour, J. M., Horvath, S. & Geschwind, D. H. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Med.* **5**, 48 (2013).
55. Artegiani, B. et al. A single-cell RNA sequencing study reveals cellular and molecular dynamics of the hippocampal neurogenic niche. *Cell Rep.* **21**, 3271–3284 (2017).
56. Siddiqui, T. J. et al. An LRRTM4-HSPG complex mediates excitatory synapse development on dentate gyrus granule cells. *Neuron* **79**, 680–695 (2013).
57. Buckley, M. T. et al. Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. *Nat. Aging* **3**, 121–137 (2023).
58. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
59. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
60. Sun, E. D., Ma, R. & Zou, J. Dynamic visualization of high-dimensional data. *Nat. Comput. Sci.* **3**, 86–100 (2023).
61. Delchambre, L. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Mon. Not. R. Astron. Soc.* **446**, 3545–3555 (2015).
62. Navarro Negredo, P., Yeo, R. W. & Brunet, A. Aging and rejuvenation of neural stem cells and their niches. *Cell Stem Cell* **27**, 202–223 (2020).
63. Doetsch, F. A niche for adult neural stem cells. *Curr. Opin. Genet. Dev.* **13**, 543–550 (2003).
64. Alvarez-Buylla, A. & Garcıia-Verdugo, J. M. Neurogenesis in adult subventricular zone. *J. Neurosci.* **22**, 629–634 (2002).
65. Dulken, B. W. et al. Single-cell analysis reveals T cell infiltration in old neurogenic niches. *Nature* **571**, 205–210 (2019).
66. Liu, L. et al. Exercise reprograms the inflammatory landscape of multiple stem cell compartments during mammalian aging. *Cell Stem Cell* **30**, 689–705 (2023).
67. Cebrian-Silla, A. et al. Single-cell analysis of the ventricular-subventricular zone reveals signatures of dorsal and ventral adult neurogenesis. *eLife* **10**, e67436 (2021).
68. Chaker, Z., Codega, P. & Doetsch, F. A mosaic world: puzzles revealed by adult neural stem cell heterogeneity. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 640–658 (2016).
69. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).
70. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
71. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).

Methods

Preprocessing of datasets

We followed data preprocessing approaches from prior benchmark comparisons of spatial gene prediction methods[7], which found highest predictive performance for these methods when non-normalized single-cell spatial transcriptomics data were paired with normalized single-cell RNA-seq data. The RNA-seq data were normalized using the Scanpy function pp.normalize_total() with its default settings followed by log transformation with an added pseudocount. We selected only genes expressed in at least 1% of cells in the RNA-seq data.

Prediction of spatial gene expression

General framework for spatial gene expression prediction. The spatial gene prediction problem involves paired data from spatial transcriptomics and RNA-seq that are approximately from the same tissue and organism. We denote the spatial transcriptomics data as $X \in \mathbb{R}^{n \times p}$ and the RNA-seq data as $Y \in \mathbb{R}^{m \times q}$, where rows are cells and columns are genes. Generally, spatial gene prediction considers the case where $q >> p$ and the genes present in X are a subset of those in Y. A spatial gene prediction method predicts the expression of a gene that is present in Y but not in X for each cell in X using information from both X and Y.

Harmony. Harmony (as referred to in this work) involves joint embedding of the spatial transcriptomics and RNA-seq data using the Harmony algorithm[72] followed by k-nearest-neighbor averaging to calculate predicted expression values for each spatial cell based on its nearest neighbors in the RNA-seq data. We implemented the Harmony algorithm following the description outlined in previous applications[10]. We used default Harmony settings in the Scanpy external.pp.harmony_integrate() implementation. We averaged across the ten nearest RNA-seq neighbors for each spatial cell using the first $\min\{30, p\}$ Harmony principal components.

SpaGE. SpaGE performs spatial gene prediction using a two-step approach consisting of alignment using the domain adaptation algorithm PRECISE[17] and then performing k-nearest-neighbor regression[8]. We used a local download of the SpaGE algorithm available at https://github.com/tabdelaal/SpaGE/ with a version corresponding to a download date of 19 July 2022. We set the number of principal vectors in SpaGE equal to 20 if $p < 40$ and to $\min\{n, p\}/2$ rounded to the nearest integer if $p \geq 40$ and otherwise used the default settings.

Tangram. Tangram uses a deep learning framework to create a mapping for projecting RNA-seq gene expression onto space[13]. We followed preprocessing details for Tangram (1.0.3) according to previous benchmarks[7], which consisted of Leiden clustering on the scaled highly variable genes in the spatial data using Scanpy methods with default settings unless otherwise specified: pp.highly_variable_genes(), pp.scale() with max_value = 10, tl.pca(), pp.neighbors() and tl.leiden() with resolution = 0.5. After preprocessing, the identified clusters were used by Tangram to project the RNA-seq cells onto space using map_cells_to_space() with mode = 'clusters' and density_prior = 'rna_count_based' and project_genes() with default settings.

gimVI. The gimVI algorithm uses deep generative modeling to impute spatial transcriptomics from paired single-cell RNA-seq data[12]. We used raw counts for both RNA-seq data and spatial transcriptomics data as input for the gimVI model and, unless otherwise specified, we used default settings for gimVI (0.20.3) training and prediction. Due to existing computational issues with gradients in the gimVI model, we were only able to extend TISSUE to gimVI for six dataset pairings, of which five are benchmark datasets. In some cases, several modifications had to be made to apply gimVI successfully. Instead of the default zero-inflated negative binomial generative distribution for the RNA-seq data, we used a Poisson distribution for the mouse prefrontal cortex (STARmap) dataset, *Drosophila* embryo (FISH) dataset and human U2OS (MERFISH) dataset. We used a negative binomial generative distribution for RNA-seq data for the mouse gastrulation (seqFISH) dataset, axolotl brain (Stereo-seq) dataset and adult mouse SVZ (MERFISH) dataset. Instead of the standard 10-fold cross-validation within TISSUE and for evaluating TISSUE, we used 3-fold cross-validation for the mouse gastrulation (seqFISH) dataset and 5-fold cross-validation for the adult mouse SVZ (MERFISH) dataset.

Calibration scores for spatial gene expression prediction

Modifications to standard conformal inference framework. Conformal inference provides a framework for developing rigorous prediction intervals for predictions made from machine learning models. We extend this framework to construct prediction intervals for predicted gene expression values in spatial transcriptomics. To adapt this framework, we made the following key modifications and additions to the traditional theory: (1) defining a scalar measure of uncertainty (cell-centric variability) that utilizes spatial context and can be measured in a single pass of any spatial gene expression method; (2) translating from the supervised learning setting to the unsupervised setting for spatial gene expression prediction, which includes using the entire spatial transcriptomics data for calibration; (3) calculating fine-resolution prediction intervals at the level of cell–gene combinations instead of general uncertainties for a given gene or a given cell; (4) calculating asymmetric prediction intervals that are more suitable to RNA count data; (5) building custom calibration of uncertainties with hierarchically stratified groupings consisting of combinations of genes and cells; (6) designing uncertainty-aware methods and algorithms for using TISSUE prediction intervals and uncertainties for a variety of downstream data analysis tasks.

Cross-validated spatial gene expression prediction. To compute calibration scores, we obtained estimated gene expression predictions on genes that are already measured in the spatial transcriptomics data. This is achieved through a cross-validation approach where a subset of the genes in the spatial transcriptomics data are left out and the gene prediction method is fitted to the remaining genes to make predictions on the left-out subset of genes (that is, calibration genes). In practice, we used 10-fold cross-validation to obtain predictions for all genes in the spatial transcriptomics data but the TISSUE implementation provides options to customize the cross-validation procedure according to user specifications.

Cell-centric variability. We outline three desiderata to guide the development of a scalar uncertainty measure for spatial gene prediction:

1. To ensure computational scalability, the measure can be calculated on a single set of predicted gene expression values.
2. To accurately measure heterogeneity in prediction performance, the measure provides specific values for any cell and gene.
3. The measure ideally leverages spatial and gene expression similarity information.

We introduce the cell-centric variability to satisfy these desiderata. Specifically, for a cell i and gene j, the cell-centric variability U_{ij} is computed according to equations (1) and (2) using cells in its local neighborhood N_i. We defined cell neighborhoods as the 15 nearest cells by Euclidean distance for each cell using squidpy.gr.spatial_neighbors()[73], and removed outliers by subsequently excluding neighbors with distance greater than Q3 + 1.5 × (Q3 − Q1), where Q3 and Q1 are the third and first quartiles of neighbor distances across all cell neighborhoods, respectively. We used this approach for defining cell neighborhoods for all experiments because it did not require gene expression information and thus could be generalized to unseen genes; was approximately

spatially scale-invariant such that in all 11 spatial transcriptomics datasets, 15 neighbors was between the average 1-hop and 2-hop neighborhood size for a Delaunay triangulation mesh graph of the cells; ensured a sufficiently high number of cells to compute the cell-centric variability reliably; and ensured a relatively fixed number of cells in each cell neighborhood such that cell-centric variability estimates could be comparable across different cells and different contexts.

The intercept term in equation (1) is included to ensure well-defined calibration scores and nonzero prediction interval widths for cells with no differences in gene expression across its neighborhood, which can result from the high sparsity of single-cell transcriptomics data. The weights W_{ik} in equation (2) are used to impose a soft weighting of the cell-centric variability for similar neighboring cells (that is, of the same cell type) over dissimilar neighboring cells (that is, of different cell types) without the need for user specification of discrete cell-type clusters. Current weights are based on the cosine similarity between gene expression of neighboring cells, but alternative definitions of weights based on distances in lower-dimensional embedding of gene expression may be advantageous when the dimensionality of the spatial transcriptomics dataset is high. The cell-centric variability can be computed for all cells and genes in both the calibration set (genes in the spatial transcriptomics data) and evaluation and test sets (genes to be predicted that are not in the spatial transcriptomics data).

Calculation of calibration scores from variability measure. To link the cell-centric variability to the prediction error, we compute the calibration score as shown in equation (3). We compute calibration scores for all pairs of cells and genes in the calibration set (that is, present in spatial transcriptomics data) and allocate them to their corresponding stratified groups (see below for details). Calibration scores are further separated by the sign of $X_{ij} - \hat{X}_{ij}$ to construct nonsymmetric uncertainty bounds around the predicted expression value with $X_{ij} > \hat{X}_{ij}$ designating inclusion in the calibration scores set for the upper interval and $X_{ij} < \hat{X}_{ij}$ designating inclusion in the calibration scores set for the lower interval.

Stratified cell and gene grouping for calibration scores. In addition to context-specific construction of calibration scores, TISSUE can also provide finer groupings of genes and cells, each with their own calibration score sets. These stratified groupings are specified by the number of gene groups k_g and the number of cell groups k_c for a total of $k_g \times k_c$ groups. Stratified grouping is performed for genes first through k-means clustering with $k = k_g$ of the genes on the first 15 principal components representing the transposed predicted gene expression matrix. Then, within each of the identified gene strata, we performed further k-means clustering with $k = k_c$ of the cells on the first 15 principal components representing the predicted gene expression matrix restricted to genes present in that stratum.

Since there is no guarantee that all stratified groups will contain genes in the calibration set or that all stratified groups will have an adequate number of scores for calibration, for stratified groups with less than 100 scores in either the upper interval or lower interval calibration score sets, we defaulted the calibration score set to the union of all calibration score sets across all stratified groups. To assess how representative the calibration set is for each stratified group, TISSUE includes options to measure the Wasserstein distance between the cell-centric variability of a group and that of the subset of genes in its calibration set.

TISSUE also includes an option ('auto') for automated selection of the stratified grouping parameters. The automated selection is achieved by first performing an increasing line search of integer $k_g > 1$ values and then performing k-means clustering of cells on the transposed predicted gene expression matrix. Finally, we compute the silhouette score on the identified clusters and increment k_g until the silhouette score decreases and then set k_g equal to the value at which maximum silhouette score was achieved. Then, we perform a similar incremental line search for $k_c > 1$, use k-means clustering of genes on the predicted gene expression matrix and return the k_c value for maximum silhouette score after the same early stopping as described previously.

Conformal prediction intervals

Retrieval of prediction intervals from calibration scores. For a given confidence level α, we construct the prediction interval with approximate probability coverage $(1-\alpha)$ by retrieving the $\frac{\lfloor (n+1)(1-\alpha) \rfloor}{n}$ –th quantiles of the upper interval calibration scores and lower interval calibration scores. Referring to these quantiles as $q_\alpha^{(u)}$ and $q_\alpha^{(l)}$, respectively, the nonsymmetric conformal prediction interval for the predicted gene expression of cell i and gene j can be computed according to equation (4):

$$I_{ij} = \left[\hat{X}_{ij} - U_{ij} \cdot q_\alpha^{(l)}, \hat{X}_{ij} + U_{ij} \cdot q_\alpha^{(u)}\right] \tag{4}$$

As this prediction interval does not explicitly depend on the measured prediction error, it can be calculated for all predicted gene expression values, even if the gene was not originally present in the calibration set.

Coverage guarantee for TISSUE prediction intervals. Under regularity conditions, the conformal inference framework provides consistent symmetric prediction intervals when applied to scalar uncertainty measures such as cell-centric variability[40]. Building on that result, we show that this consistency is still valid with the nonsymmetric prediction intervals that we compute using TISSUE.

Proposition 1. Let $\{X_{[:,j]}\}_{j\in[1,\ldots,p,\text{test}]}$ be independent and identically distributed from some distribution, then $P(X_{ij} \in [l_{ij}, u_{ij}]) \geq 1-\alpha$ for any confidence level $0 \leq \alpha \leq 1$, where l_{ij} and u_{ij} are the quantiles of the lower and upper calibration score sets corresponding to X_{ij}, respectively.

Here, 'test' refers to the index or set of indices for predicted genes that are not in the measured spatial transcriptomics data. Using the notation $l_{ij} = \hat{X}_{ij} - U_{ij} \cdot q_\alpha^{(l)}$ and $u_{ij} = \hat{X}_{ij} + U_{ij} \cdot q_\alpha^{(u)}$, the coverage of the TISSUE prediction interval for some confidence level α can be represented according to equation (5):

$$\begin{aligned} P(X_{ij} \in [l_{ij}, u_{ij}]) &= P(X_{ij} \in [l_{ij}, \hat{X}_{ij}]) + P(X_{ij} \in [\hat{X}_{ij}, u_{ij}]) \\ &= P(X_{ij} \in [l_{ij}, \hat{X}_{ij}] | X_{ij} < \hat{X}_{ij}) P(X_{ij} < \hat{X}_{ij}) \\ &+ P(X_{ij} \in [\hat{X}_{ij}, u_{ij}] | X_{ij} > \hat{X}_{ij}) P(X_{ij} > \hat{X}_{ij}) \\ &\geq (1-\alpha)(P(X_{ij} < \hat{X}_{ij}) + P(X_{ij} > \hat{X}_{ij})) \\ &\geq 1-\alpha, \end{aligned} \tag{5}$$

with the first inequality following from theorem 1 of ref. 40. And thus, given that symmetric intervals provide proper coverage[40], then we are also guaranteed proper coverage with the asymmetric prediction intervals produced by TISSUE, which is further evident through the empirical coverage assessment for TISSUE (Fig. 2f). This guarantee extends to the stratified group setting for $k_g > 1$ and/or $k_c > 1$ (see proposition 2 in ref. 40).

Evaluation of prediction intervals. We evaluate the empirical coverage of the prediction intervals using 10-fold cross-validation splits of the genes in the spatial transcriptomics data into a calibration set and an evaluation set. We leave out the evaluation set and use the calibration set to compute calibration scores. Then, using these calibration scores, for every value of α, we compute TISSUE prediction intervals for all cells and genes in the evaluation set. We then compute empirical coverage of the TISSUE prediction intervals, which is defined as the fraction of measured gene expression values in the evaluation set

that fall within their respective TISSUE prediction interval across all cells and genes (Fig. 2f and Extended Data Fig. 4e) or across all cells for each individual gene with nonzero predicted and measured expression (Extended Data Fig. 4d). Well-calibrated coverage of TISSUE prediction intervals are indicated by close equivalence of the empirical coverage ('fraction within prediction interval') and the theoretical coverage ('specified TISSUE coverage'). For datasets with a small number of cells, there will likely be worse calibration for choices of α that are very close to either 0 or 1 due to sparse calibration sets for those values.

We measured the 'calibration error' by measuring the area between the difference of the empirical calibration curve and the theoretically optimal calibration curve. Numerically, this involved computing the absolute difference between the empirical coverage and the theoretical coverage at each alpha value and then estimating the absolute area under this difference curve using the trapezoidal rule with default implementation of numpy.trapz().

Sprod denoising and alternative cell similarity graph. To investigate the effect of denoising on TISSUE calibration performance, we used Sprod to preprocess the mouse somatosensory cortex osmFISH dataset to yield a 'denoised' version and this was used in place of the original dataset in TISSUE calibration. We used the pseudo-image option in Sprod because the corresponding images for the dataset were not publicly available. All TISSUE settings were kept identical to the settings for the original analyses.

To investigate the effect of other cell similarity graphs on TISSUE calibration, we used the cell similarity graph constructed by Sprod in the output file 'sprod_Detected_graph.txt' in place of the default cosine similarity graph constructed by TISSUE. To ensure connectivity, we added the identity matrix to the Sprod cell similarity adjacency matrix, and otherwise kept all TISSUE settings at their default settings.

Uncertainty-aware hypothesis testing

Generating multiple imputations using calibration scores. We introduce a multiple imputation procedure for performing uncertainty-aware hypothesis testing for predicted spatial gene expression. Multiple imputations are generated by uniformly sampling calibration scores from the corresponding union of upper and lower interval calibration score sets for each predicted spatial gene expression value. Given a uniform random sample of such calibration scores, $S_{(d)} \in \mathbb{R}^{n\times p}$, we compute an alternative imputation given by equation (6):

$$\hat{X}_{(d)} = \hat{X} + \delta_{u/l,(d)} \times S_{(d)} \times U \tag{6}$$

where '×' denotes element-wise multiplication, $U \in \mathbb{R}^{n\times p}$ is equal to the cell-centric variability measures computed on $\hat{X}$, and $\delta_{u/l} = 1$ if the score was sampled from the upper interval set and $\delta_{u/l} = -1$ if the score was sampled from the lower interval set. This sampling is repeated $D-1$ times to generate D multiple imputations including the original imputation $\hat{X}$. We tempered the multiple imputations against outliers by restricting the sampling to the scores within the set corresponding to the 80% conformal prediction interval.

Modified two-sample *t*-test for multiple imputation. After generating D multiple imputations from the calibration scores, we performed hypothesis testing using a modified two-sided, two-sample t-test. Consider two groups with sets of sample/cell indices A and B. Then, the mean difference and variance under normal two-sample t-test for a single imputation are given by equations (7) and (8):

$$\mu_d = \frac{1}{n_A}\sum_{i\in A}\hat{X}^{(d)}_{i,:} - \frac{1}{n_B}\sum_{i\in B}\hat{X}^{(d)}_{i,:} \tag{7}$$

$$s_d^2 = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\left(\frac{(n_A-1)\,\mathrm{Var}\left\{\hat{X}^{(d)}_{i,:} : i\in A\right\} + (n_B-1)\,\mathrm{Var}\left\{\hat{X}^{(d)}_{i,:} : i\in B\right\}}{n_A+n_B-2}\right), \tag{8}$$

where $\hat{X}^{(d)}$ denotes the d-th imputation among the D multiple imputations. Extending these statistical measures to the multiple imputation case, we use the standard modification for multiple imputation[47–49,74], which results in the following mean and variance according to equations (9) and (10):

$$\mu_{MI} = \frac{1}{D}\sum_{d=1}^{D}\mu_d \tag{9}$$

$$s^2_{MI} = s^2_W + \left(1+\frac{1}{D}\right)s^2_B \tag{10}$$

where s^2_W is the within-imputation variance and s^2_B is the between-imputation variance, computed according to equations (11) and (12):

$$s^2_W = \frac{1}{D}\sum_{d=1}^{D}s^2_d \tag{11}$$

$$s^2_B = \frac{1}{D-1}\sum_{d=1}^{D}(\mu_d-\mu_{MI})^2. \tag{12}$$

Then, the modified test statistics for the two-sample t-test is given by equation (13):

$$\tilde{t}_{MI} = \frac{\mu_{MI}}{\sqrt{s^2_{MI}}}, \tag{13}$$

which is t-distributed with degrees of freedom $(D-1)\left(1+\frac{Ds^2_W}{(D+1)s^2_B}\right)^2$ and the resulting probability can be interpreted as the posterior probability of a significant difference in means between the two groups accounting for both evidence of this effect and the reliability of the imputations by inflating the standard error for this effect[47]. Under some regularity assumptions (approximate normality of imputed values and missing at random), the multiple imputation approach produces consistent estimates[48,49].

Empirical evaluation of TISSUE hypothesis testing. For each dataset, TISSUE hypothesis tests were computed using 10-fold cross-validation. In each cross-validation fold, we generated TISSUE spatial gene expression predictions and calibration scores on the calibration genes. Then, statistics for the TISSUE hypothesis test were computed for the held-out genes. This procedure is repeated across all folds to accrue statistics for all genes in the dataset. For TISSUE multiple imputation t-test and Wilcoxon test frameworks, we used 100 multiple imputations in each hypothesis test. For TISSUE multiple imputation SpatialDE framework, we used 10 multiple imputations in each hypothesis test due to the longer computational run time for SpatialDE. To evaluate the performance of TISSUE hypothesis testing, we applied the underlying tests (for example, t-test, Wilcoxon/Mann–Whitney, SpatialDE) to the measured spatial gene expression values of the held-out genes to obtain 'ground truth' statistics. Using these ground truth observations, we then benchmarked the discoveries made by TISSUE hypothesis tests against the discoveries made by the underlying tests directly applied on the predicted spatial gene expression values without TISSUE.

Alternative TISSUE hypothesis tests. To extend the multiple imputation testing framework to non-parametric two-sample tests, TISSUE can perform one-sided Wilcoxon/Mann–Whitney tests for 'greater than' or 'less than' comparisons between two multiply imputed samples.

For these tests, we use the scipy.stats.mannwhitneyu() implementation with either alternative = 'greater' or alternative = 'less' options. In a similar way, TISSUE can be extended to spatially variable gene detection methods such as SpatialDE[52], which provide *P* values as a measure of statistical significance of spatially variable expression. We use the standard implementation and transform the input to log-normalized counts before running SpatialDE.

The implementation of these frameworks is identical to the TISSUE multiple imputation *t*-test with the exception of using a different rule in combining inference across multiple imputations[50]. In this approach, we obtain *P* values of independent hypothesis tests (that is, one-sided Wilcoxon/Mann–Whitney test between two samples or SpatialDE across all cells with spatial coordinates) on each of *m* imputations, $\{p_1, \ldots, p_m\}$. Then, we transform the *P* values to approximate normality, $\{\tilde{p}_0, \ldots, \tilde{p}_m\}$, and combine these transformed values with $z_{MI} = \left(\frac{1}{m}\sum_{i=1}^{m}\tilde{p}_i\right) \Big/ \sqrt{1 + \mathrm{Var}\{\tilde{p}_0, \ldots, \tilde{p}_m\}}$. Finally, we transform the combined value back to the original scale to obtain a multiply imputed *P* value estimate, p_{MI}[50]. The transformation and inverse transformation are achieved in practice with scipy.stats.norm.ppf() and scipy.stats.norm.cdf(), respectively. Notably, this alternative approach involves running *m* independent hypothesis tests and is less computationally efficient than the multiple imputation *t*-test, which performs an aggregate test at the end of the procedure.

Due to computational constraints, we were only able to evaluate the Wilcoxon/Mann–Whitney framework on four of the seven labeled datasets, and we were only able to evaluate the SpatialDE framework on two small datasets, which were unlabeled since SpatialDE uses spatial coordinates of the cells for testing.

Simulated data for differential gene expression analysis

We generated synthetic data using SRTsim[51] for comparing the TISSUE uncertainty-aware hypothesis testing approach against traditional hypothesis testing. We used the reference-free SRTsim framework and generated a synthetic counts matrix using their native Shiny app. The data consist of two cell groups, referred to as 'A' and 'B', which were determined by manual drawing of a linear boundary between two spatial domains in the SRTsim Shiny application. In this setup, the separation is artificial with no simulated expression differences between the two groups. There are 465 'A' cells and 515 'B' cells for a total of 980 cells. To generate counts, we followed the default SRTsim recommendations and used the zero-inflated negative binomial model and set the zero proportion to 0.05, dispersion to 0.5 and mean to 2. We simulated 1,000 genes, where there was no systematic difference in expression of any gene between cell group 'A' and cell group 'B'. We used a random seed of 444 for SRTsim.

To simulate prediction bias for cells under condition 'B', we added shifted Gaussian noise (mean equal to $\mu \geq 0$, variance equal to one) to half of the genes for all cells in condition 'B'. Standard Gaussian noise was added to the other simulated expression values (that is, all cells and genes in group 'A' and the other half of genes for group 'B'). This simulation results in prediction errors that artificially produce a difference in predicted expression between the two groups for half of the genes despite the absence of any true expression differences in the original simulated data. For the main experiments, we varied the prediction bias parameter μ. We used random seeds of 444 in all sampling steps.

Metadata annotation for spatial transcriptomics datasets

For annotating cell types in the mouse hippocampus seqFISH dataset, we preprocessed the data using a standard Scanpy pipeline. Starting with the counts matrix, we normalized the data using pp.normalize_total() with default settings, log-transformed the data using pp.log1p() and scaled the data with pp.scale(). We computed principal components and a neighbors graph using tl.pca() followed by pp.neighbors() with 20 principal components. Finally, we performed Leiden clustering using tl.leiden() with resolution of 0.3, which yielded 5 cell clusters. We used tl.rank_genes_groups() with the Wilcoxon method to identify the top five marker genes for each cell cluster and manually identified the clusters using those markers. In total, we identified endothelial cells, oligodendrocytes, astrocytes and 2 neuron clusters.

For annotating cell types in the mouse primary visual cortex MERFISH dataset, we preprocessed the data using a standard Scanpy pipeline. Starting with the counts matrix, we normalized the data using pp.normalize_total() with default settings, log transformed the data using pp.log1p() and scaled the data with pp.scale(). We computed principal components and a neighbors graph using tl.pca() followed by pp.neighbors() with 20 principal components. Finally, we performed Leiden clustering using tl.leiden() with resolution of 0.1, which yielded 11 cell clusters. We used tl.rank_genes_groups() with the Wilcoxon method to identify the top five marker genes for each cell cluster and manually identified the clusters using those markers. In total, we identified endothelial cells, oligodendrocytes, astrocytes, and 8 neuron-like cell clusters.

For annotating anatomic regions in the *Drosophila* embryo dataset, we used the same preprocessing procedure as for the mouse primary visual cortex MERFISH dataset. We identified 7 Leiden clusters and grouped them into four region labels based on their spatial localization with 2 'posterior' clusters, 1 'anterior' cluster, 1 'bottom' cluster and 3 'middle' clusters.

We retrieved annotated class labels from publicly available metadata for the mouse somatosensory osmFISH dataset, mouse gastrulation seqFISH dataset and axolotl telencephalon Stereo-seq dataset. For the mouse somatosensory osmFISH dataset, we retrieved both anatomic region ('region') and cell-type ('ClusterName') labels from the metadata available at http://linnarssonlab.org/osmFISH/osmFISH_SScortex_mouse_all_cells.loom. For the mouse gastrulation seqFISH dataset, we retrieved cell-type ('celltype_mapped_refined') labels from the metadata available at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/ in the metadata.Rds file for 'embryo1' and 'z5'. For the axolotl telencephalon Stereo-seq dataset, we retrieved cell-type ('annotation') labels from the metadata available at https://db.cngb.org/stomics/artista/ for the Stage44.h5ad object file.

Replicate analysis for mouse gastrulation seqFISH dataset

To examine the reproducibility of TISSUE quantities across spatial transcriptomics replicates, we curated a replicate of the mouse gastrulation seqFISH dataset[36], which has not been previously included in benchmarking analyses for spatial gene expression prediction[7]. We mapped cell-type ('celltype_mapped_refined') labels from the metadata available at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/ in the metadata.Rds file for 'embryo1' and 'z2'. For replication experiments, we utilized identical settings for TISSUE calibration, prediction interval calculation and differential gene expression analysis as was used for the original dataset analysis.

Simulated data for downstream analyses

We generated synthetic data using SRTsim[51] for benchmarking TISSUE cell filtering and TISSUE-WPCA approaches for improved performance on downstream analysis tasks. We used the reference-free SRTsim framework and generated a synthetic counts matrix using their native Shiny application. The data consist of two cell groups, referred to as 'A' and 'B', which were determined by manual drawing of a linear boundary between two spatial domains in the SRTsim Shiny application. There are 476 'A' cells and 504 'B' cells for a total of 980 cells. To generate counts, we followed the default SRTsim recommendations and used the zero-inflated negative binomial model and set the zero proportion to 0.05, dispersion to 0.5 and mean to 2. We simulated 1,000 genes, consisting of 500 positive signal genes and 500 noise genes, where the positive signal genes had an average log fold change that was double in 'B' cells than in 'A' cells. We used a random seed of 444 for SRTsim.

To simulate prediction bias for cells in cell type 'A', we randomly sampled a proportion of cells in the 'A' group specified 'mix-in' proportion parameter, and then for each gene and selected cell, we updated their expression level with a random uniform sample of 'B' cell expression levels for that gene. For the sampled cells, this simulated prediction bias shifts their predicted gene expression profiles to be more similar to those of cell type 'B' rather than cell type 'A'. Finally, standard Gaussian noise was added to all other expression values for both cell types to simulate prediction noise. We used random seeds of 444 in all sampling steps.

Uncertainty-aware cell filtering for downstream tasks

Using the TISSUE prediction interval, we performed filtering of high-uncertainty cells (referred to as TISSUE cell filtering) to improve training/evaluation of supervised learning models, clustering and data visualization. We approximate the prediction uncertainty using the width of the 67% prediction interval (equivalent to the asymmetric standard error). Then, we convert all uncertainty values to z scores using the mean and standard deviation of expression for each gene in the data. For each cell, we assign a score equal to the average of its z scores across all genes. The cells with the highest scores are removed from the filtered data. The threshold for removal is automatically determined using Otsu's method, which finds a threshold that maximizes the variance between the filtered and unfiltered score sets. In the context of classification, we avoid inter-class differences in prediction uncertainty by performing this filtering procedure independently within each class.

Evaluation of TISSUE cell filtering for downstream tasks

We used several evaluation metrics to quantify the improvement of TISSUE cell filtering over using the predicted gene expression (baseline) for a variety of common downstream analysis tasks. To ensure relatively balanced representation of classes, we used dataset and class label pairs that were restricted to the three classes with greatest prevalence. To generate the initial predicted spatial gene expression, we iteratively made predictions on held-out folds of genes using one of the specified prediction methods and with 10-fold cross-validation (see 'Cross-validated spatial gene expression prediction' for further details). For supervised learning (classification), we performed 5-fold cross-validation where TISSUE cell filtering was applied independently on each train and test split. Within each fold, we fitted a logistic regression model on the train set using sklearn.linear_model.LogisticRegression() with penalty = 'l1' and solver = 'liblinear'. The model was evaluated on the test set and the classification accuracy, area under the receiver-operator curve and macro F1 score were computed. These performance metrics were then averaged across the five folds. For clustering and visualization, we applied TISSUE cell filtering to the predicted gene expression data and then performed standardization and PCA on the filtered data. For clustering, we then used k-means clustering with $k = 3$ on the top 15 principal components of the TISSUE-filtered data and measured clustering quality using the ARI with sklearn.metrics.adjusted_rand_score(). For visualization, we then fit a support vector classifier on the top 15 principal components of the TISSUE-filtered data using sklearn.svm.SVC() with kernel = 'linear' and random_state = 444, and measured the accuracy of separation of classes, which we refer to as linear separability. For comparison, we repeated each of these procedures for the unfiltered/baseline predicted gene expression. These assessment procedures were applied independently for each spatial gene expression prediction method (Harmony, SpaGE and Tangram) within each dataset.

Dynamic visualization of principal components

We generated dynamic visualizations of the first two principal components for visual comparison of PCA on the measured spatial gene expression, PCA on the predicted spatial gene expression and PCA on the TISSUE-filtered predicted spatial gene expression. We used DynamicViz (v.0.0.3) to center and rigidly align the cells across 20 two-dimensional PCA visualizations of the simulated datasets and visualized the resulting alignments using dynamicviz.viz.stacked(). Alignment was achieved on the subset of cells that overlapped between the reference and target visualizations for the TISSUE-filtered data. Robust visualizations can be consistently aligned across different replicates. We scored the variability of the resulting visualization by computing variance scores for each cell using dynamicviz.score.variance() with method = 'global'.

Weighted PCA for uncertainty-aware tasks

As an alternative to TISSUE cell filtering, we implemented a weighted version of PCA where each value in the gene expression matrix is assigned a scalar weight. We computed the weights according to the following steps. First, we compute the inverse of the TISSUE prediction interval width (that is, 67% prediction interval upper bound minus lower bound). Then, we normalize these values for each gene by the mean value across that gene to correct for expression level differences between genes. Finally, we binarize these normalized values so that the top 80% of normalized values will have 10-fold higher relative weight than the bottom 20% of normalized values. These binary values are used as weights for WPCA. Alternatively, we have also implemented a weighting scheme where we simply take the log transform of the normalized inverse prediction interval widths, which provides comparable performance to the previously described weighting scheme (Extended Data Fig. 7c). WPCA directly decomposes the weighted covariance matrix to obtain principal vectors, and then applies weighted least-squares optimization to retrieve the principal components[61]. We used the implementation of WPCA in the wpca (v.0.1) Python package with default settings and weights set according to our specification. The TISSUE implementation of WPCA is customizable with user options for specifying different weighting parameters.

Sample preparation and processing for MERFISH on SVZ

A healthy 3-month-old male C57BL/6 mouse was obtained from the National Institute on Aging Aged Rodent colony. The mouse was habituated for at least 2 weeks at Stanford University before use. It was housed at the ChEM-H/Neuroscience Vivarium at Stanford University and their care was monitored by the Veterinary Service Center at Stanford University under the Institutional Animal Care and Use Committee protocol 8661.

The mouse was euthanized in a carbon dioxide chamber in the morning. The whole brain was dissected and immediately embedded in ice-cold Tissue-Tek OCT compound in a cryomold and placed on dry ice. Once the sample was frozen, it was transferred and stored at −80 °C. The sample was shipped to Vizgen in dry ice for processing. At Vizgen, the brain was sectioned to obtain coronal sections of the SVZ followed by MERFISH laboratory service, transcript count detection and cell segmentation and allocation of counts to individual cells. The MERFISH dataset includes two consecutive coronal sections.

We curated a 140-gene panel for the MERFISH experiment. The panel included 2–5 known transcriptomic cell-type markers for each of aNSC/NPCs, qNSC/astrocytes, neuroblasts, microglia, endothelial cells, oligodendrocytes, T cells, mural cells, ependymal cells, neurons, macrophages and reactive astrocytes[57,65]. Markers for aNSC/NPCs were *Hmgb2*, *Hmgn2*, *Ccnd2* and *Sox2*; markers for qNSCs/astrocytes were *Aldoc*, *Clu*, *Mt3*, *Gfap* and *Id4*; markers for neuroblasts were *Tubb2b*, *Sox4*, *Tubb3* and *Dcx*. The remaining genes in the panel were related to neurogenesis, T cell activity, glycolysis, lipid metabolism and aging.

Data processing for MERFISH dataset of SVZ

The MERFISH dataset was cropped to around the left and right lateral ventricles using rectangular bounding boxes. The raw counts were normalized by the volume of the cell segmentation. To remove doublets and segmentation artifacts, we filtered out the top 2% and bottom

2% of cells by total normalized expression. For initial clustering and visualization of the data, we further normalized total expression of all cells to 250 (scanpy.pp.normalize_total() with target_sum = 250), log-transformed with an added pseudocount (scanpy.pp.log1p()) and scaled to *z* scores (scanpy.pp.scale() with max_value = 10). We performed PCA using scanpy.tl.pca(), built a neighbors graph with scanpy.pp.neighbors(), obtained UMAP visualization with scanpy.tl.umap(), and performed Leiden clustering with scanpy.tl.leiden() with a resolution of 0.5. Using visualizations of the cell-type markers in the MERFISH gene panel along with differential expression analysis, we manually identified nine cell-type clusters including two neuron clusters, astrocytes, oligodendrocytes, endothelial cells, ependymal cells, microglia, oligodendrocyte progenitor cells and an ambiguous cell-type cluster that localized to the lateral ventricles. We further subclustered the ambiguous cell cluster using the same Leiden clustering restricted to cells in this cluster and recovered three subclusters. For spatial region labels, we manually selected vertical coordinate cutoffs that corresponded to dorsal and ventral regions outlined in previous studies[67].

Subcluster identification for MERFISH dataset of SVZ

To assist in the identification of the ambiguous cell cluster, we used TISSUE to obtain uncertainty-aware SpaGE spatial gene expression predictions for additional cell-type marker genes that were not in the 140-gene MERFISH panel. These included general NSC and qNSC/astrocyte markers (*Slc1a3*, *Nr2e1*, *Sox9*, *Vcam1*, *Hes5*, *Prom1*, *Thbs4*), aNSC/NPC markers (*Pclaf*, *H2ax*, *Rrm2*, *Insm1*, *Egfr*, *Prom1*, *Mcm2*, *Cdk1*) and neuroblast markers (*Stmn2*, *Dlx6os1*, *Igfbpl1*, *Sox11*, *Dlx1*). For prediction, we used a single-cell RNA-seq dataset of the micro-dissected mouse SVZ[57]. To perform differential expression analysis, we used the TISSUE multiple imputation framework to perform two-sample *t*-tests to compare the expression of each predicted gene in one of the ambiguous subclusters to all other ambiguous subclusters. Final cell-type identifications were made by considering the markers with statistically significant overexpression within each of the subclusters after Bonferroni multiple-hypothesis correction.

Specifically, our procedure for annotating the cell subclusters for cell subtype is as follows. First, if there is significant overexpression of one or more marker genes for a cell subtype and not for the other two cell types, then we identify that subcluster as the marked cell type. Otherwise, if a subcluster has significant overexpression of markers from multiple cell subtypes, we identify the cluster as the cell subtype with the greatest proportion of markers that are significantly overexpressed. If the greatest proportion of significantly overexpressed markers is similar for two or more cell subtypes or if there are no significantly overexpressed markers for any cell subtypes, then we fail to identify a cell subtype label for that subcluster.

Spatial region classifiers for MERFISH dataset of SVZ

To train the regional SVZ classifiers, we performed TISSUE uncertainty-aware SpaGE spatial gene expression prediction of the whole-transcriptome (that is, all genes in the paired single-cell RNA-seq dataset) and obtained *P* values for differential expression in dorsal versus ventral regions for each cell type (qNSC/astrocyte, aNSC/NPC, neuroblast) using the TISSUE multiple imputation *t*-test. For each cell type, we selected the top 20 most differentially expressed genes with the lowest TISSUE multiple imputation *t*-test *P* values across the dorsal and ventral regions. Then, we trained cell-type-specific penalized logistic regression models (sklearn.linear_model.LogisticRegression() with penalty = 'l1' and solver = 'liblinear') to predict the regional origin of the cell from these 20 predicted gene expression features. The inputs were standardized before fitting the logistic regression model. We obtained class probabilities for each cell using 10-fold cross-validation, training and evaluating an independent model on each train and test split. For comparison, we used either the TISSUE-filtered input with the 67% prediction interval width or unfiltered input in fitting and evaluating the classifiers. For each cell type and approach (TISSUE-filtered or baseline unfiltered predicted expression), we measured the performance of the classifiers using the F1 score, accuracy, area under the receiver-operator curve and average precision score using the corresponding Scikit-learn implementations of these metrics.

Reporting summary

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

Data availability

All processed spatial transcriptomics and RNA-seq dataset pairings, including the final annotated adult mouse SVZ MERFISH dataset, have been deposited at https://doi.org/10.5281/zenodo.8259942. Other data files (raw images and large intermediate data files) can be provided upon reasonable request. Raw data were accessed from existing benchmark datasets[7] and are also available from the following studies:
Mouse hippocampus: Spatial transcriptomics (seqFISH) at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/; RNA-seq (10x Chromium) at GSE158450 in the Gene Expression Omnibus (GEO) for 'HIPP_sc_Rep1_10X sample'.
Mouse primary visual cortex: Spatial transcriptomics (MERFISH) at https://github.com/spacetx-spacejam/data; RNA-seq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex.
Mouse prefrontal cortex: Spatial transcriptomics (STARmap) at '20180419_BZ9_control' in https://www.starmapresources.com/data; RNA-seq (10x Chromium) at GSE158450 in the GEO for 'PFC_sc_Rep2_10X'.
Human middle temporal gyrus: Spatial transcriptomics (ISS) at https://github.com/spacetx-spacejam/data; RNA-seq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq.
Mouse primary visual cortex: Spatial transcriptomics (ISS) at https://github.com/spacetx-spacejam/data; RNA-seq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex.
Drosophila embryo: Spatial transcriptomics (FISH) at https://github.com/rajewsky-lab/distmap/; RNA-seq (Drop-seq) at GSE95025 in GEO.
Mouse somatosensory cortex: Spatial transcriptomics (osmFISH) at http://linnarssonlab.org/osmFISH/ for cortical region subset; RNA-seq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-smart-seq for mouse somatosensory cortex.
Mouse primary visual cortex: Spatial transcriptomics (ExSeq) at https://github.com/spacetx-spacejam/data; RNA-seq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex.
Mouse gastrulation: Spatial transcriptomics (seqFISH) at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/; RNA-seq (10x Chromium) 'Sample 21' in the MouseGastrulationData R package.
Human U2OS: Spatial transcriptomics (MERFISH) at https://www.pnas.org/doi/suppl/10.1073/pnas.1912459116/suppl_file/pnas.1912459116.sd12.csv; RNA-seq (10x Chromium) at 'BC22' in GSE152048 in the GEO database.
Axolotl brain: Spatial transcriptomics (Stereo-seq) at 'Stage44.h5ad' in https://db.cngb.org/stomics/artista/download/; RNA-seq (10x Chromium) at 'animal1' in 'all_nuclei_clustered_highlevel_anno.rds' at https://zenodo.org/records/6390083.

Code availability

The TISSUE Python package and associated code and documentation are available at https://github.com/sunericd/TISSUE/, and all code for generating figures and analyses is separately available at https://github.com/sunericd/tissue-figures-and-analyses/.

References

72. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
73. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
74. Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med. Res. Methodol.* **9**, 57 (2009).

Acknowledgements

Funding support was provided by Knight-Hennessy Scholars program (to E.D.S.), Paul and Daisy Soros Fellowship for New Americans (to E.D.S.), the National Science Foundation Graduate Research Fellowship Program (to E.D.S.), D. Donoho at Stanford University (to R.M.), National Institutes of Health P01AG036695 (to A.B.), NSF CAREER 1942926 (to J.Z.), National Institutes of Health P30AG059307 (to J.Z.), 5RM1HG010023 (to J.Z.) and grants from the Silicon Valley Foundation (to J.Z.) and the Chan Zuckerberg Initiative (to J.Z.). We thank L. Xu, O. Zhou and M. Yuksekgonul for helpful discussions.

Author contributions

E.D.S. and J.Z. conceived of the study. E.D.S. designed and implemented the method and ran all associated analyses with J.Z. and R.M providing input. P.N.N. and A.B. provided samples for the mouse SVZ MERFISH dataset and input on associated analyses. E.D.S. prepared a draft of the paper. R.M., P.N.N., A.B. and J.Z. edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

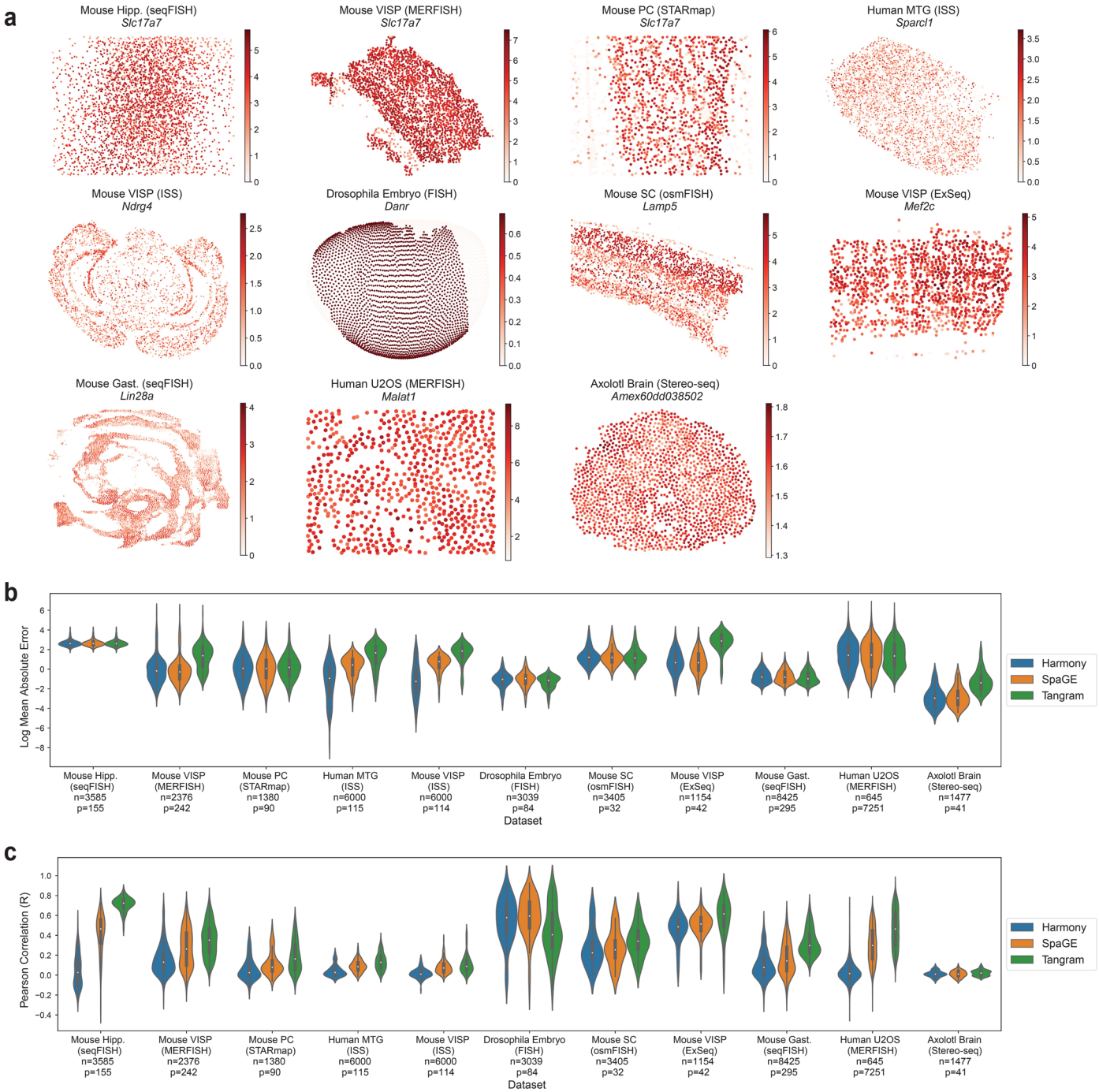
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02184-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02184-y.

Correspondence and requests for materials should be addressed to James Zou.

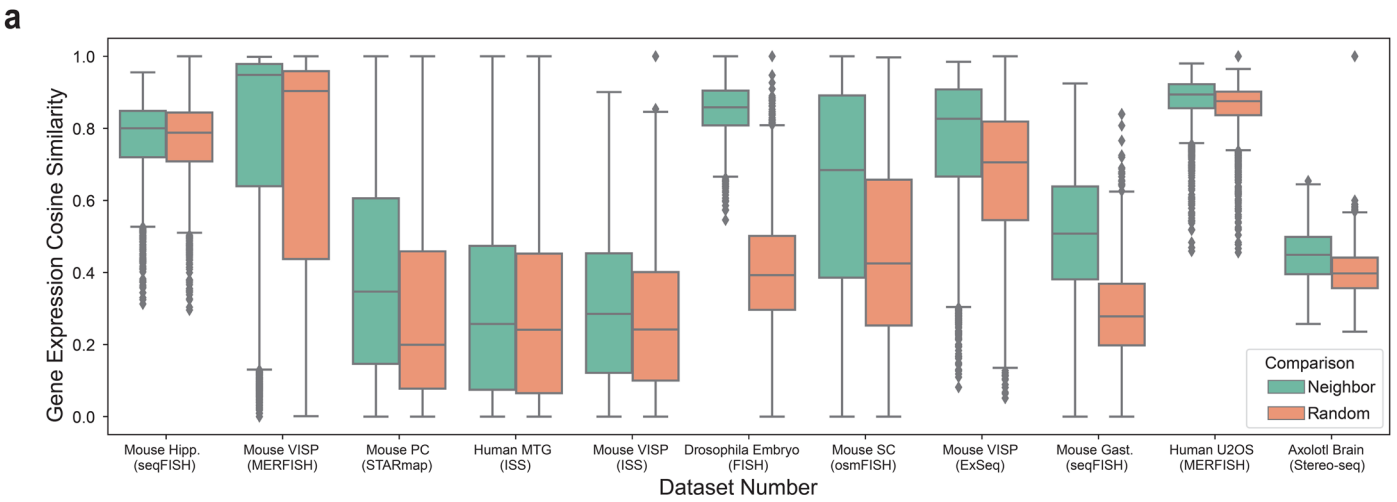
Peer review information *Nature Methods* thanks Nancy Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

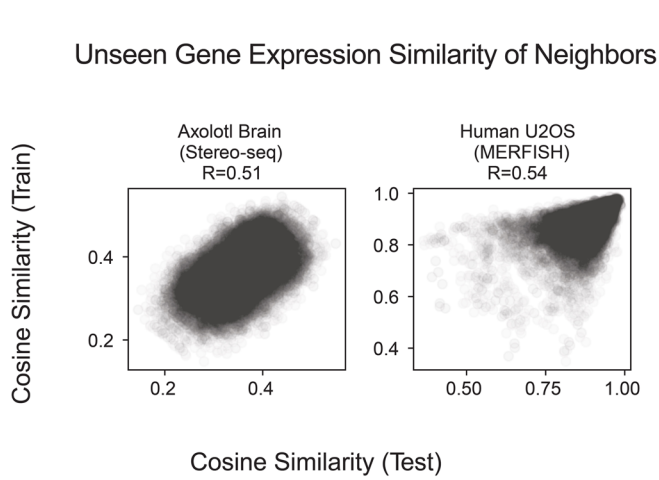
https://doi.org/10.1038/s41592-024-02184-y



Extended Data Fig. 1 | Overview of datasets and prediction performance. **a**, Visualization of cells in the eleven spatial transcriptomics datasets colored by the expression of the highest-expressed gene in each respective dataset. Abbreviations are as follows: hippocampus (Hipp.) primary visual cortex (VISP), prefrontal cortex (PC), middle temporal gyrus (MTG), somatosensory cortex (SC), gastrulation (Gast.), U-2 OS cell line (U2OS). **b**,**c**, Performance of all three gene prediction methods (Harmony, SpaGE, Tangram) on all datasets as measured by (**b**) gene-wise mean absolute error between predicted and actual gene expression over 10-fold cross-validation, and (**c**) gene-wise Pearson correlation between predicted and actual gene expression over 10-fold cross-validation. Shown also are the number of cells (n) in the spatial transcriptomics datasets and the number of genes (p) shared between spatial and RNAseq datasets. In panels **b**-**c**, the inner box corresponds to quartiles of the metrics and the whiskers span up to 1.5 times the interquartile range of the metrics.

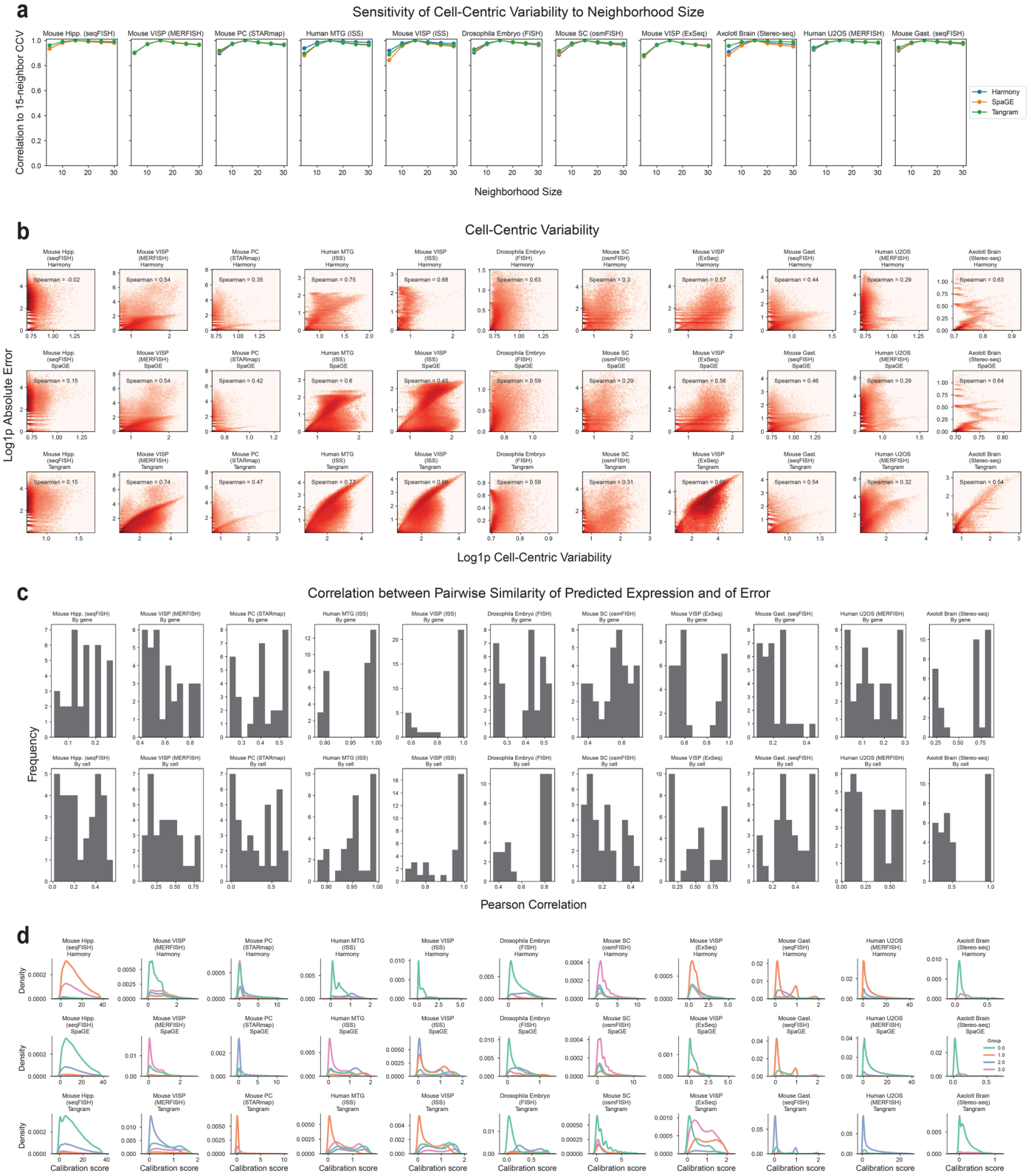
https://doi.org/10.1038/s41592-024-02184-y





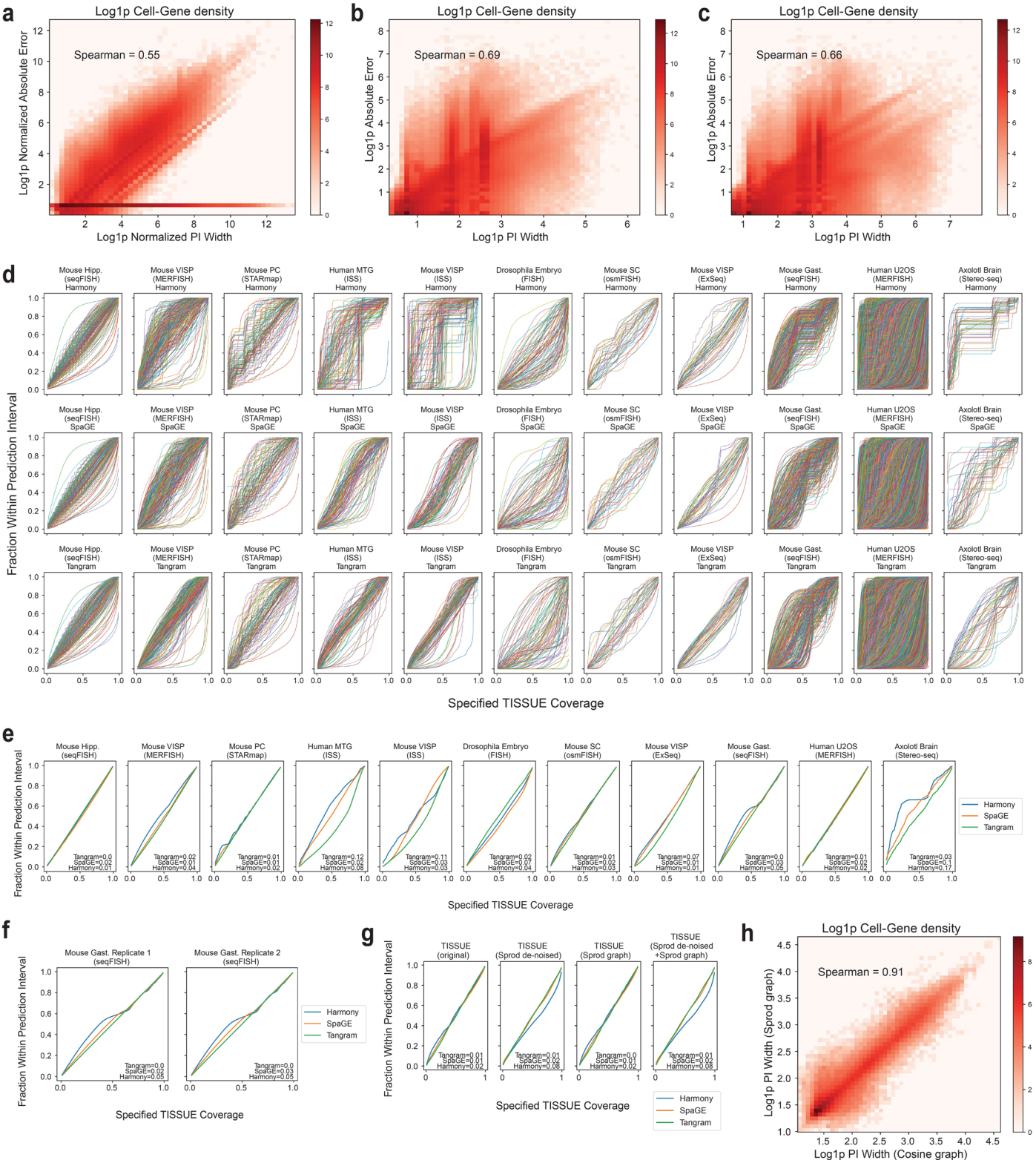
Extended Data Fig. 2 | Evidence of gene expression similarity between spatial neighbors. a, Cosine similarity of gene expression profiles for 250 cells paired with all their neighbors in the TISSUE spatial graph compared to pairings with randomly drawn cells across all eleven spatial transcriptomics datasets. The boxplot corresponds to the quartiles of the cosine similarity measurements. The center line corresponds to median cosine similarity, which was strictly higher in the neighbor-paired comparisons than the random-paired comparisons across all datasets. Whiskers span up to 1.5 times the interquartile range of the metrics and values outside this range are shown as dots. Abbreviations are as follows: hippocampus (Hipp.) primary visual cortex (VISP), prefrontal cortex (PC), middle temporal gyrus (MTG), somatosensory cortex (SC), gastrulation (Gast.), U-2 OS cell line (U2OS). **b**, Scatter plots of the cosine similarities of gene expression profiles for 250 cells paired with their neighbors for either the training gene set or the test gene set determined by random train-test split of all genes (50% train, 50% test). Shown are cosine similarity pairs for 10 train-test splits for the two benchmark spatial transcriptomics datasets with the most measured genes.

https://doi.org/10.1038/s41592-024-02184-y



Extended Data Fig. 3 | Cell-centric variability and calibration score distributions for individual datasets and prediction methods. a, Pearson correlation of all cell-centric variability measures obtained for different numbers of neighbors in building the TISSUE spatial graph compared to the default setting of 15 neighbors. **b**, Correlation of cell-centric variability and absolute prediction error shown individually for each dataset and prediction method combination computed over 10-fold cross-validation. Log density with added pseudocount (Log1p) is shown by color, with a maximum of 1000 cells and 300 genes sampled from each dataset to provide more uniform representation. **c**, Histograms showing the distribution of Pearson correlations between either gene-wise or cell-wise similarities of prediction errors and similarities of predicted expression values across all spatial transcriptomic datasets and across all prediction methods. **d**, Distribution of TISSUE calibration scores shown individually for each dataset and prediction method combination ($(k_g, k_c) = (4, 1)$). Details on each dataset and prediction method can be found in Methods. Abbreviations are as follows: hippocampus (Hipp.) primary visual cortex (VISP), prefrontal cortex (PC), middle temporal gyrus (MTG), somatosensory cortex (SC), gastrulation (Gast.), U-2 OS cell line (U2OS).

https://doi.org/10.1038/s41592-024-02184-y

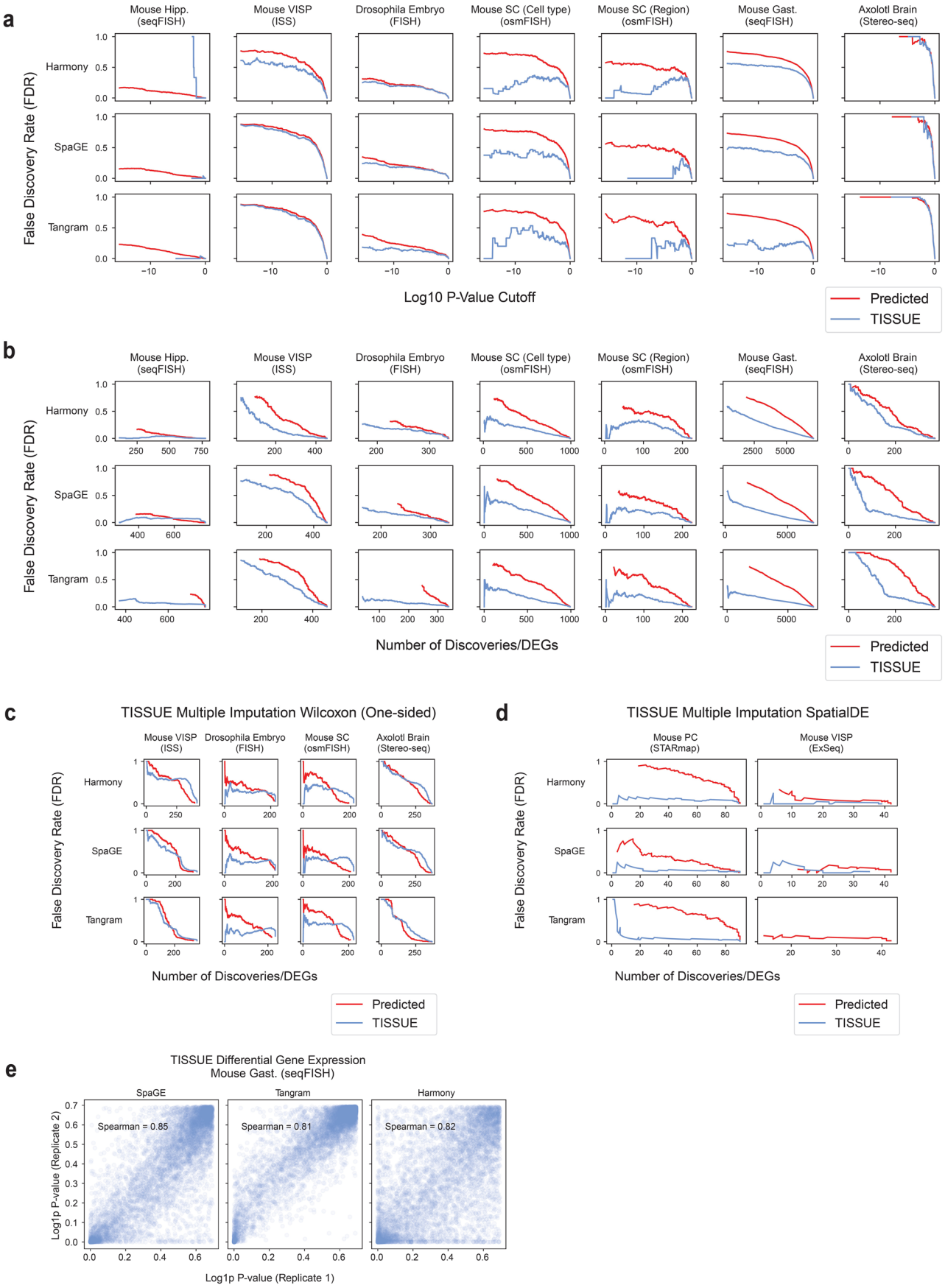


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41592-024-02184-y

Extended Data Fig. 4 | Further evaluation of TISSUE prediction intervals. **a**-**c**, Correlation plots across all dataset and prediction method combinations computed over 10-fold cross-validation for (**a**) the 67% prediction interval width and absolute prediction error, both normalized by the absolute value of the predicted expression; (**b**) 50% prediction interval width and absolute prediction error; (**c**) 80% prediction interval width and absolute prediction error. Log density with added pseudocount (Log1p) is shown by color, with a maximum of 1000 cells and 300 genes sampled from each dataset to provide more uniform representation. **d**, Gene-level calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation. Each line corresponds to an independent gene in the spatial transcriptomics dataset. Abbreviations are as follows: hippocampus (Hipp.) primary visual cortex (VISP), prefrontal cortex (PC), middle temporal gyrus (MTG), somatosensory cortex (SC), gastrulation (Gast.), U-2 OS cell line (U2OS). **e**,**f**, Calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation (**e**) under automated setting of (k_g, k_c) for stratified grouping; and (**f**) for two technical replicates of the mouse gastrulation seqFISH dataset with (k_g, k_c) = (4, 1). The calibration error is annotated for each prediction method (see Methods). **g**, Calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation for the mouse somatosensory cortex osmFISH dataset with different combinations of Sprod de-noising or Sprod-based spatial similarity graph instead of the TISSUE spatial neighbors graph. The calibration error is annotated for each prediction method (see Methods). **h**, Correlation plot of 67% prediction interval width with TISSUE spatial neighbors graph with cosine similarity weighting and 67% prediction interval width with Sprod similarity graph and weighting for the mouse somatosensory cortex osmFISH dataset and all prediction methods computed over 10-fold cross-validation.

https://doi.org/10.1038/s41592-024-02184-y

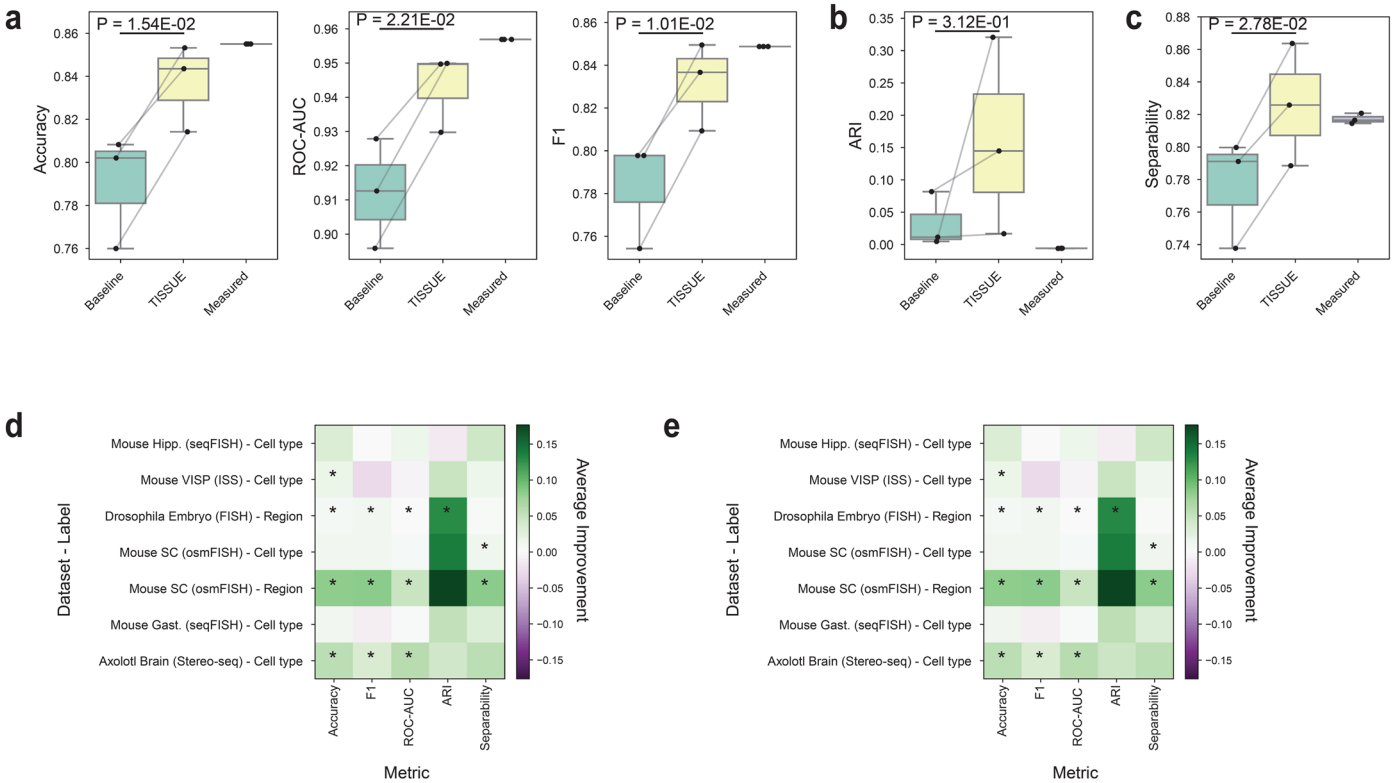


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41592-024-02184-y

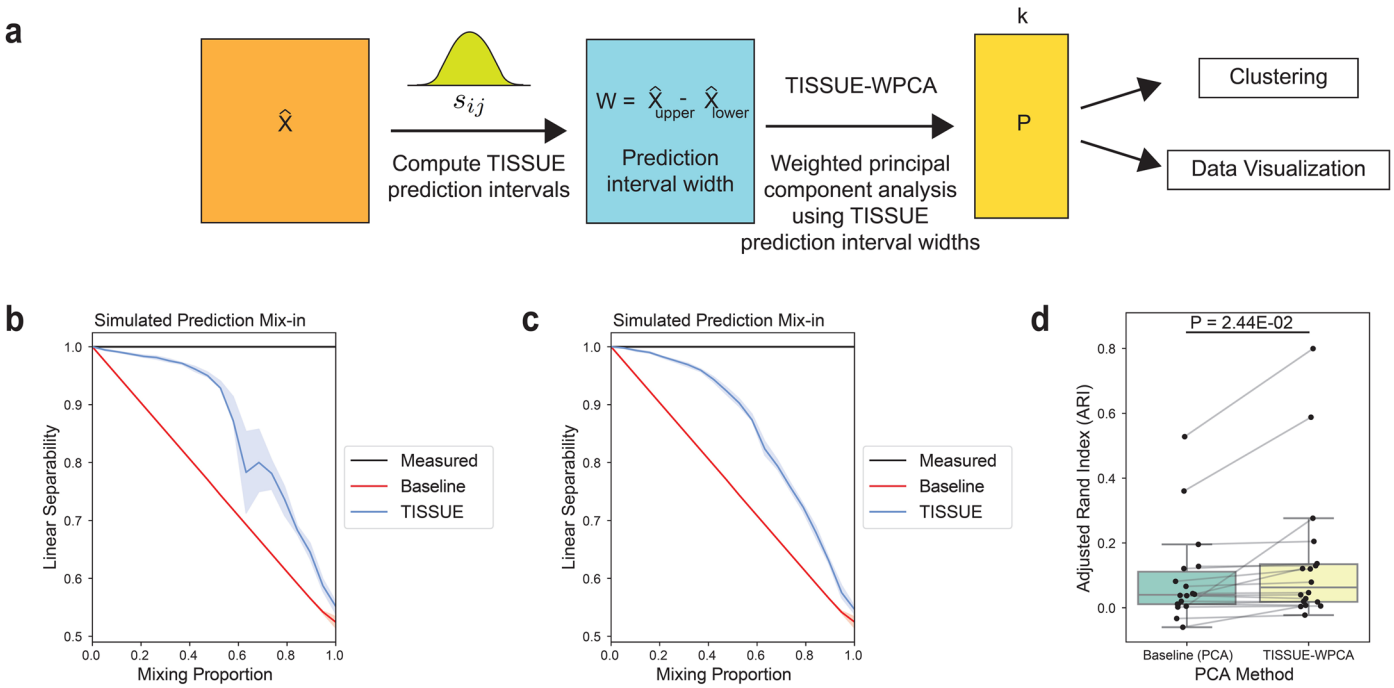
Extended Data Fig. 5 | Additional differential gene expression analysis with TISSUE. a, False discovery rate of differentially expressed genes between cell type or anatomic region labels (one versus all approach) using the differentially expressed genes on the measured gene expression profiles as the ground truth across different p-value cutoffs. P-values were computed using two-sided t-test. Discoveries are assessed across all genes for all class labels. Shown are results for all three prediction methods and all spatial transcriptomics datasets with cell type or region labels available. All calibration scores were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping. Abbreviations are as follows: hippocampus (Hipp.) primary visual cortex (VISP), middle temporal gyrus (MTG), somatosensory cortex (SC), gastrulation (Gast.). **b**, False discovery rate of differentially expressed genes between cell type or anatomic region labels (one versus all approach) as a function of the number of discoveries and with automated stratified grouping. **c**, False discovery rate of differentially expressed genes between cell type or anatomic region labels (one versus all approach) as a function of the number of discoveries and with $(k_g, k_c) = (4, 1)$ settings for stratified grouping for the alternative TISSUE multiple imputation framework using the 'greater than' one-sided Wilcoxon/Mann-Whitney test. **d**, False discovery rate of spatially variable genes as a function of the number of discoveries and with $(k_g, k_c) = (4, 1)$ settings for stratified grouping for the alternative TISSUE multiple imputation framework using the SpatialDE test. **e**, Correlation plot of the log p-values obtained from the TISSUE multiple imputation t-test framework between two technical replicates of the mouse gastrulation seqFISH dataset.

https://doi.org/10.1038/s41592-024-02184-y



Extended Data Fig. 6 | Additional experiments for uncertainty-aware supervised learning, clustering, and visualization. a-c, Downstream task performance metrics on the three most prominent anatomic region class labels for the mouse somatosensory osmFISH dataset. Shown are metrics for all three prediction methods with automated stratified grouping settings. P-value was computed using a paired two-sided t-test on n = 3 independent prediction methods. The box corresponds to quartiles of the metrics and the whiskers span up to 1.5 times the interquartile range of the metrics. (**a**) Accuracy, F1 score, and ROC-AUC (receiver-operator characteristic area under the curve) metrics for logistic regression models trained on the predicted gene expression, TISSUE-filtered predicted gene expression, or measured gene expression for classification. (**b**) Adjusted Rand index (ARI) for k-means clustering (k = 3) on the top 15 principal components obtained from the predicted gene expression, TISSUE-filtered predicted gene expression, or measured gene expression for classification. (**c**) Linear separability measured as classification accuracy of linear kernel support vector classifier fitted on the top 15 principal components obtained from the predicted gene expression, TISSUE-filtered predicted gene expression, or measured gene expression for classification. **d**, Average improvement of performance metrics using TISSUE-filtered approach in lieu of unfiltered approach on predicted expression for supervised learning (Accuracy, F1, ROC-AUC), clustering (adjusted Rand index (ARI)), and visualization (linear separability) for the top three classes across all dataset and class label combinations. Results were obtained using the 50% prediction interval width for filtering. Abbreviations are as follows: hippocampus (Hipp.) primary visual cortex (VISP), middle temporal gyrus (MTG), somatosensory cortex (SC), gastrulation (Gast.). Asterisks denote significant difference in performance metrics between TISSUE-filtered approach and unfiltered approach ($p<0.05$) with p-values computed using a paired two-sided t-test on n=3 independent prediction methods. **e**, Same as panel **d** except with the 80% prediction interval width for filtering.

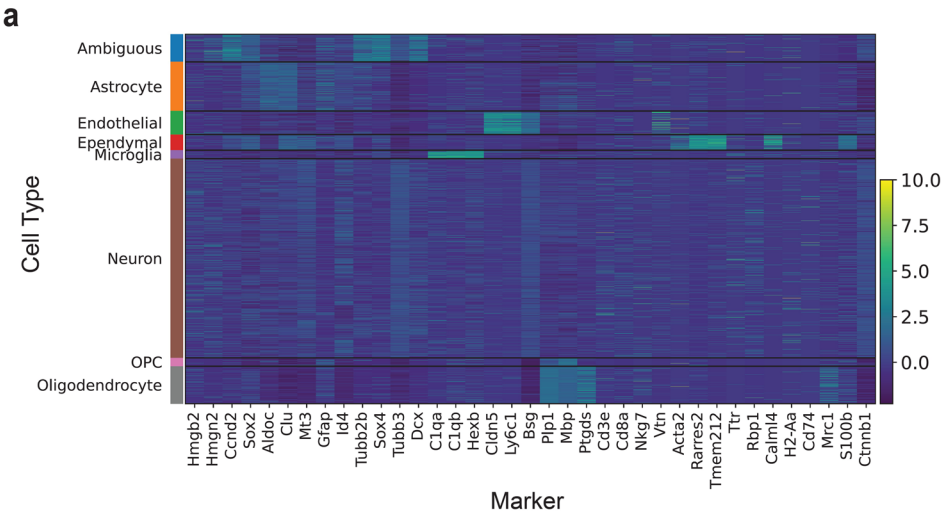
https://doi.org/10.1038/s41592-024-02184-y



Extended Data Fig. 7 | Uncertainty-aware clustering and label separation with TISSUE-WPCA. a, Schematic illustration of the weighted principal component analysis (WPCA) pipeline where the inverse TISSUE prediction interval width is used to obtain principal components from WPCA, which are then used for downstream tasks of clustering and label separation. **b**, Linear separability measured as the binary classification accuracy of a linear kernel support vector classifier fitted on the two cell clusters in the simulated spatial transcriptomics data as a function of the simulated mix-in proportion. The classifier was trained on the top 15 principal components obtained from the measured gene expression profiles with PCA, predicted gene expression profiles with PCA, and predicted gene expression profiles with TISSUE-WPCA. For TISSUE-WPCA, weights were determined by binarizing the inverse normalized 67% prediction interval width (see Methods). Results were obtained using automated stratified grouping. Bands represent the interquartile range and solid line denotes the median linear separability across 20 simulated datasets. **c**, Same as in panel **b** except with TISSUE-WPCA weighting using the log-transformed inverse normalized 67% prediction interval width. **d**, Adjusted Rand index (ARI) for k-means clustering (k = 3) on the top 15 principal components obtained from PCA on the predicted expression or TISSUE-WPCA on the predicted gene expression for six real spatial transcriptomics dataset and label pairings and all prediction methods. P-value was computed using a paired two-sided t-test on n=18 sets of predictions across 3 independent prediction methods and 6 independent dataset and class label combinations. The box corresponds to quartiles of the metrics and the whiskers span up to 1.5 times the interquartile range of the metrics.

https://doi.org/10.1038/s41592-024-02184-y

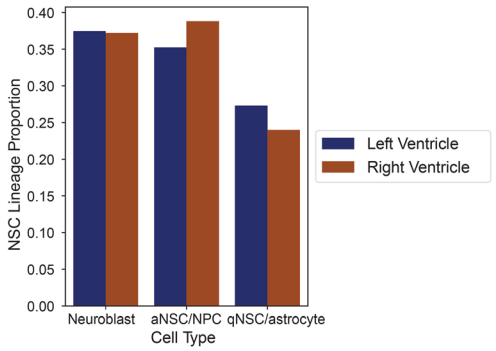
a



b

	Gene	P-value (Predicted)	P-value (TISSUE)
Quiescent neural stem cell/astrocyte (qNSC/astrocyte)	*Slc1a3*	**P = 1.4 E-6**	n.s.
	Nr2e1	**P = 8.4 E-15**	n.s.
	Sox9	**P = 5.8 E-10**	**P = 7.6 E-4**
	Vcam1	**P = 1.4 E-3**	n.s.
	Hes5	**P = 6.9 E-7**	n.s.
	Prom1	**P = 3.4 E-4**	n.s.
	Thbs4	**P = 2.3 E-5**	n.s.
	Gja1	**P = 1.1 E-5**	n.s.
Activated neural stem cell/neural progenitor cell (aNSC/NPC)	*Pclaf*	**P < 1 E-20**	**P = 1.2 E-9**
	H2afx	**P < 1 E-20**	**P = 9.2 E-8**
	Rrm2	**P < 1 E-20**	**P = 2.0 E-7**
	Insm1	**P = 1.8 E-15**	**P = 1.2 E-4**
	Egfr	**P < 1 E-20**	n.s.
	Mcm2	**P < 1 E-20**	n.s.
	Cdk1	**P < 1 E-20**	**P = 4.7 E-6**

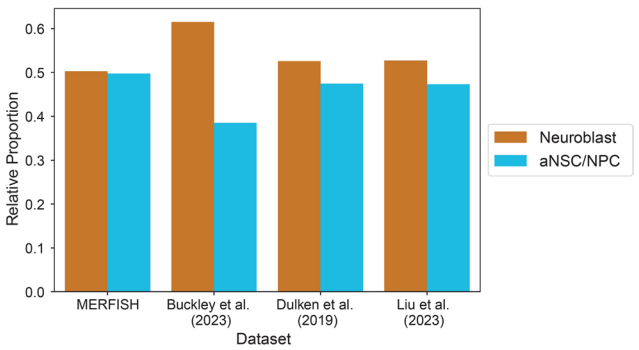
d



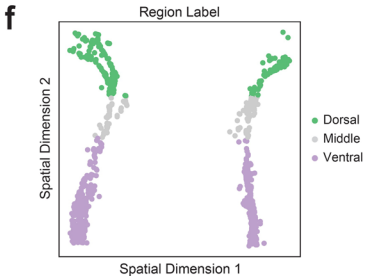
c

Cell Subtype	SpaGE + TISSUE	SpaGE	Tangram + TISSUE	Tangram	Harmony + TISSUE	Harmony	gimVI + TISSUE	gimVI
qNSC/astrocyte	✓	✓	✓	✗	✗	✗	✓	✓
aNSC/NPC	✓	✗	✗	✗	✓	✗	✓	✗
Neuroblast	✓	✓	✗	✗	✗	✗	✗	✓

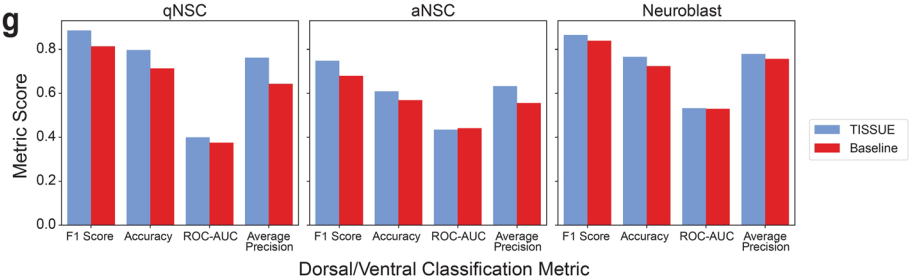
e



f



g

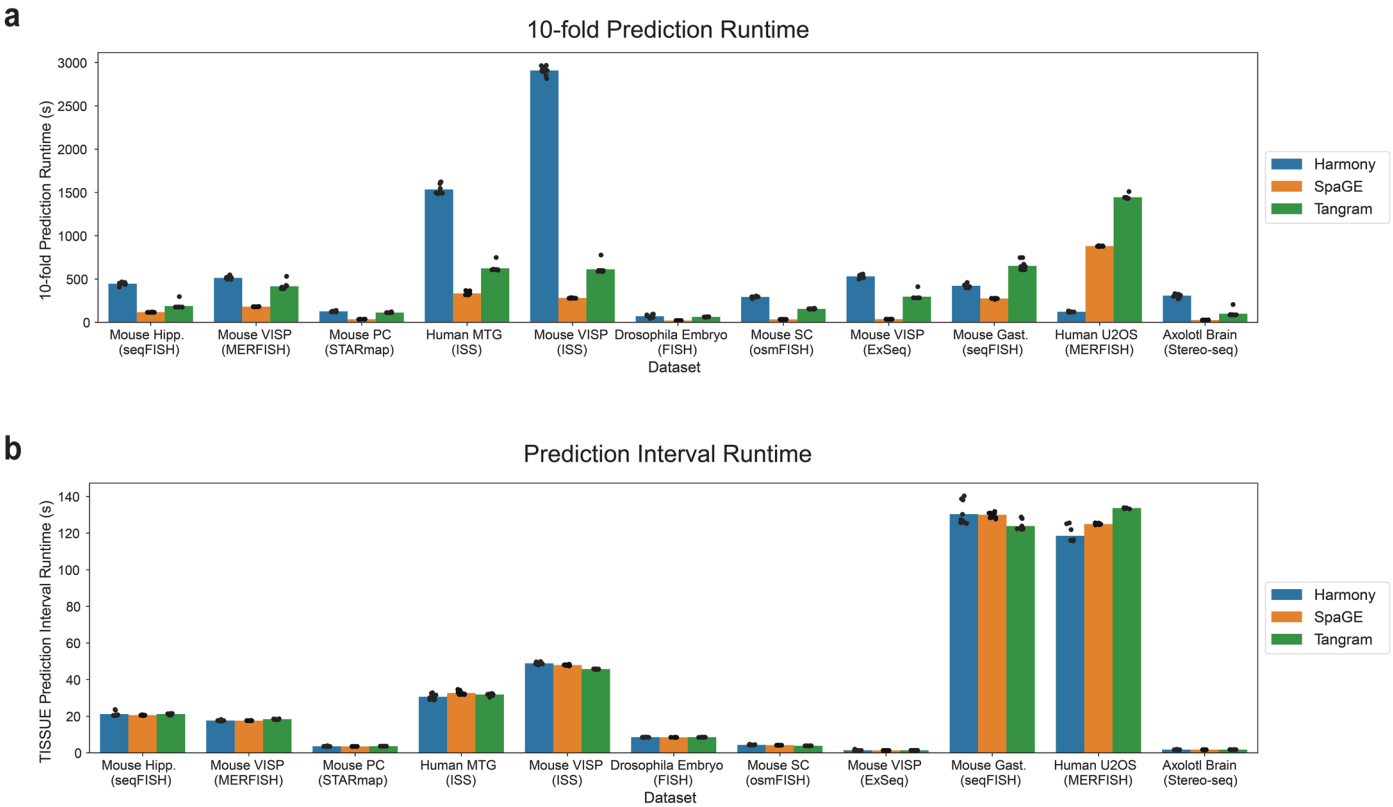


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41592-024-02184-y

Extended Data Fig. 8 | TISSUE is necessary to identify ambiguous NSC lineage subtype. a, Heatmap of the scaled log-normalized gene expression of original cell type markers in the adult mouse subventricular zone MERFISH dataset for each of the identified cell type clusters. The Ambiguous cell type cluster in the first row exhibits high expression of qNSC/astrocyte, aNSC/NPC, and neuroblast markers. **b**, Additional predicted marker genes for the second ambiguous subcluster are differentially expressed for all qNSC/astrocyte and aNSC/NPC markers under traditional hypothesis testing with two-sided t-test on the predicted gene expression (Predicted). With TISSUE multiple imputation two-sided t-test, there are substantially more aNSC/NPC markers that are differentially over-expressed in the ambiguous subcluster (TISSUE), permitting identification of this subcluster as an aNSC/NPC subtype cluster. P-values are shown for all predicted marker genes with significance threshold of Bonferroni-adjusted $p < 0.1$ for either two-sided t-test or TISSUE multiple imputation two-sided t-test. **c**, Table indicating whether each of the three cell subtypes of the NSC lineage could be resolved from predicted marker genes using baseline or TISSUE-based approaches. Green checks indicate successful identification of cell subtype and red crosses indicate unsuccessful identification of cell subtype. **d**, Relative proportion of each of the three TISSUE-identified subtypes in the neural stem cell lineage cluster for either the left or right lateral ventricle. **e**, Relative proportions of aNSC/NPC and neuroblast populations across the MERFISH dataset and three single-cell RNAseq datasets of the mouse subventricular zone. The qNSC/astrocyte proportions were not compared since they were aggregated with astrocytes of the striatum in the single-cell RNAseq datasets. **f**, Spatial visualization of the cells in the neural stem cell lineage cluster colored by dorsal or ventral spatial location labels. **g**, Dorsal versus ventral classification performance of TISSUE-filtered penalized logistic regression models and baseline unfiltered penalized logistic regression models evaluated using 10-fold cross-validation across F1 score, accuracy, area under the receiver-operator curve, and average precision.

https://doi.org/10.1038/s41592-024-02184-y



Extended Data Fig. 9 | Computational runtime for TISSUE. a, Bar plots of total runtimes for spatial gene expression prediction computations over 10 predictions to generate estimated predictions on all calibration genes. Bars denote the mean runtime across 10 instances of TISSUE prediction and each dot represents the runtime for one instance of generating TISSUE predictions using 10-fold cross-validation. **b**, Bar plots of total runtimes for TISSUE prediction interval calculation including computation of cell-centric variability and calibration score sets. Bars denote the mean runtime across 10 instances of TISSUE prediction interval calculation and each dot represents the runtime for one instance of TISSUE prediction interval calculation.

nature portfolio

Corresponding author(s): James Zou

Last updated by author(s): 1/4/2024

Reporting Summary

Nature Portfolio 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 Portfolio 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 specific software was used to collect data for this study.
Data analysis	python (3.8.13), tissue-sc (0.0.2), squidpy (1.2.3), scikit-learn (1.2.2), scanpy (1.9.3), scipy (1.9.1), wpca (0.1), tangram-sc (1.0.3), harmonypy (0.0.9), spage (version accessed July 19, 2022), dynamicviz (0.0.3) The TISSUE Python package and associated code and documentation are available at https://github.com/sunericd/TISSUE, and all code for generating figures and analyses are separately available at https://github.com/sunericd/tissue-figures-and-analyses.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All processed spatial transcriptomics and RNAseq dataset pairings, including the final annotated adult mouse subventricular zone MERFISH dataset, have been deposited at https://doi.org/10.5281/zenodo.8259942. Other data files (raw images and large intermediate data files) can be provided upon reasonable request. Raw data were accessed from existing benchmark datasets and are also available from the following studies:

Mouse Hippocampus: Spatial transcriptomics (seqFISH) at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/; RNAseq (10X Chromium) at GSE158450 in GEO for 'HIPP_sc_Rep1_10X sample'.

Mouse VISP: Spatial transcriptomics (MERFISH) at https://github.com/spacetx-spacejam/data/; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex (VISp).

Mouse Prefrontal Cortex (PC): Spatial transcriptomics (STARmap) at '20180419_BZ9_control' in https://www.starmapresources.com/data; RNAseq (10X Chromium) at GSE158450 in GEO for 'PFC_sc_Rep2_10X'.

Human Middle Temporal Gyrus (MTG): Spatial transcriptomics (ISS) at https://github.com/ spacetx-spacejam/data; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq.

Mouse VISP: Spatial transcriptomics (ISS) at https://github.com/spacetx-spacejam/data; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex (VISp).

{\it Drosophila} Embryo: Spatial transcriptomics (FISH) at https://github.com/rajewsky-lab/distmap; RNAseq (Drop-seq) at GSE95025 in GEO.

Mouse Somatosensory Cortex (SC): Spatial transcriptomics (osmFISH) at http://linnarssonlab.org/osmFISH/ for cortical region subset; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-smart-seq for mouse somatosensory cortex (SSp).

Mouse VISP: Spatial transcriptomics (ExSeq) at https://github.com/spacetx-spacejam/data; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex (VISp).

Mouse Gastrulation: Spatial transcriptomics (seqFISH) at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/; RNAseq (10X Chromium) 'Sample 21' in the MouseGastrulationData R package.

Human U2OS: Spatial transcriptomics (MERFISH) at https://www.pnas.org/doi/suppl/10.1073/pnas.1912459116/suppl_file/pnas.1912459116.sd12.csv; RNAseq (10X Chromium) at 'BC22' in GSE152048 in the GEO database.

Axolotl Brain: Spatial transcriptomics (Stereo-seq) at 'Stage44.h5ad' in https://db.cngb.org/stomics/artista/download/; RNAseq (10X Chromium) at 'animal1' in 'all_nuclei_clustered_highlevel_anno.rds' at https://zenodo.org/records/6390083.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	We used eleven publicly available datasets across different spatial transcriptomics and RNAseq modalities to evaluate the performance of TISSUE. The number of datasets was chosen to be comparable to the number of datasets used in existing benchmarking efforts for spatial transcriptomics prediction models based on literature review.
Data exclusions	No data was excluded.
Replication	We replicated our findings across eleven independent spatial transcriptomics and RNAseq dataset pairings and on synthetic datasets that we generated. We also considered technical and biological replicates for one dataset and showed that TISSUE performance was highly reproducible across replicates.
Randomization	Randomization was not necessary in our study since our evaluation was done on all datasets in a cross-validated manner.
Blinding	Blinding was not possible since there were no groups in our analysis.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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