LIFE SCIENCES

Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

Accurate inference of genome-wide spatial expression with iSpatial

Chao Zhang[1,2,3]†, Renchao Chen[1,2,3]†, Yi Zhang[1,2,3,4,5]*

Spatially resolved transcriptomic analyses can reveal molecular insights underlying tissue structure and context-dependent cell-cell or cell-environment interaction. Because of the current technical limitation, obtaining genome-wide spatial transcriptome at single-cell resolution is challenging. Here, we developed a new algorithm named iSpatial to derive the spatial pattern of the entire transcriptome by integrating spatial transcriptomic and single-cell RNA-seq datasets. Compared to other existing methods, iSpatial has higher accuracy in predicting gene expression and spatial distribution. Furthermore, it reduces false-positive and false-negative signals in the original datasets. By testing iSpatial with multiple spatial transcriptomic datasets, we demonstrate its wide applicability to datasets from different tissues and by different techniques. Thus, we provide a computational approach to reveal spatial organization of the entire transcriptome at single-cell resolution. With numerous high-quality datasets available in the public domain, iSpatial provides a unique way to understand the structure and function of complex tissues and disease processes.

INTRODUCTION

In the past decade, single-cell RNA sequencing (scRNA-seq) has transformed our understanding of the cellular heterogeneity of various tissues/organs in multicellular organisms (*1*–*4*). With current scRNA-seq techniques, obtaining whole transcriptomic profiles of tens to hundreds of thousands of single cells has become routine. However, most high-throughput scRNA-seq methods use dissociated cells, and consequently, the spatial information of the analyzed cells is lost, which prevents directly connecting the molecular features of the analyzed cell types to their anatomic and functional features. On the other hand, the development of spatially resolved transcriptomic assays has enabled the transcript/cell location analysis in the tissue context, which has the potential to reveal how single-cell gene activity orchestrates the structure and function of complex tissues in health and disease (*5*).

In the past few years, different methods for spatial transcriptomic (ST) assays have been developed (*6*–*13*). Ideally, the spatial transcriptome data should provide genome-wide and spatially resolved expression measurements at single-cell resolution. However, because of technical limitations, either spatial resolution or gene coverage is compromised in most ST assays. For example, in situ capture and sequencing-based techniques are able to capture any mRNA molecules without preknowledge, but the spatial resolution is not at single-cell level (*6*, *14*). On the other hand, in situ sequencing and fluorescence in situ hybridization (FISH)–based mRNA measurement can achieve cellular or subcellular resolution, but most of these assays are limited with their throughput to genes that can be detected (usually 30 to 500) and require preknowledge for probe design (*7*–*9*).

With the rapid development of scRNA-seq and ST technologies, new bioinformatic tools have been developed to overcome the challenges in single-cell or ST data analysis (*15*–*20*). Several imputation methods for scRNA-seq data have emerged, including MAGIC (*21*), scImpute (*22*), DrImpute (*23*), and ALRA (*24*). However, methods developed for inferring ST data are still limited. Notably, by integrating scRNA-seq and spatially resolved profiling data, recent computational methods have leveraged the strength of different datasets and revealed information that otherwise cannot be obtained from a single experimental paradigm. For example, when the corresponding scRNA-seq is available, SpatialDWLS and RCTD could perform deconvolution on the "low-resolution" spatial dataset to estimate the cell type/proportion in each spatially resolved spot (*25*, *26*). On the other hand, Tangram and Cell2location could predict the spatial location of molecularly defined cell types (from scRNA-seq) on the basis of ST data (*27*, *28*). Seurat and Liger could impute transcriptome-wide spatial expression by integrating with corresponding scRNA-seq data and transferring the expression data to spatial transcriptome (*29*, *30*). Despite these existing tools, inferring expression patterns of all genes at high spatial resolution by integrating scRNA-seq and ST data are not straightforward, and different approaches have variable performance when applied to different datasets. Because the performance of this task is critical for downstream spatial analysis, a robust and convenient tool for spatial pattern prediction is highly desirable.

Here, we present iSpatial, an R-based bioinformatic tool that integrates scRNA-seq and ST profiling data to infer the expression pattern of each gene at high spatial resolution. We show that iSpatial outperforms existing approaches on its accuracy, and it can also reduce false-positive (FP) and false-negative (FN) signals in the original data. By applying iSpatial to datasets from different tissues (hippocampus, hemibrain, cortex, striatum, and liver) and generated with different techniques (Slide-seq, Stereo-seq, MERFISH, and STARmap), we have revealed both known and previously unknown spatial expression patterns in each dataset, indicating iSpatial is broadly applicable for analyzing different ST datasets. Collectively, our analyses demonstrate that iSpatial is a useful tool for resolving transcriptome-wide spatial expression patterns at single-cell resolution in complex tissues.

[1] Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA. [2] Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115, USA. [3] Division of Hematology/Oncology, Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115, USA. [4] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [5] Harvard Stem Cell Institute, WAB-149G, 200 Longwood Avenue, Boston, MA 02115, USA.
*Corresponding author. Email: yzhang@genetics.med.harvard.edu
†Co-first authors.

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

RESULTS

Overview of iSpatial

The FISH and in situ sequencing–based ST techniques, such as MERFISH (*7*), seqFISH (*8*), osmFISH (*12*), and STARmap (*9*), can simultaneously reveal gene expression and location at single-cell resolution, but with limited predefined gene targets (Fig. 1A, left). On the other hand, scRNA-seq can unbiasedly profile the whole transcriptome, but without providing spatial information (Fig. 1A, middle). We reasoned that by integrating the single-cell gene expression profiles (the gene by cell matrices) of the two methods, the missing information of nontargeted genes in each spatially profiled cells could be inferred on the basis of scRNA-seq data, resulting in genome-wide spatial expression information of the profiled cells (Fig. 1A, right).

To this end, we first performed dimension reduction on scRNA-seq and ST data separately, followed by expression stabilization, which

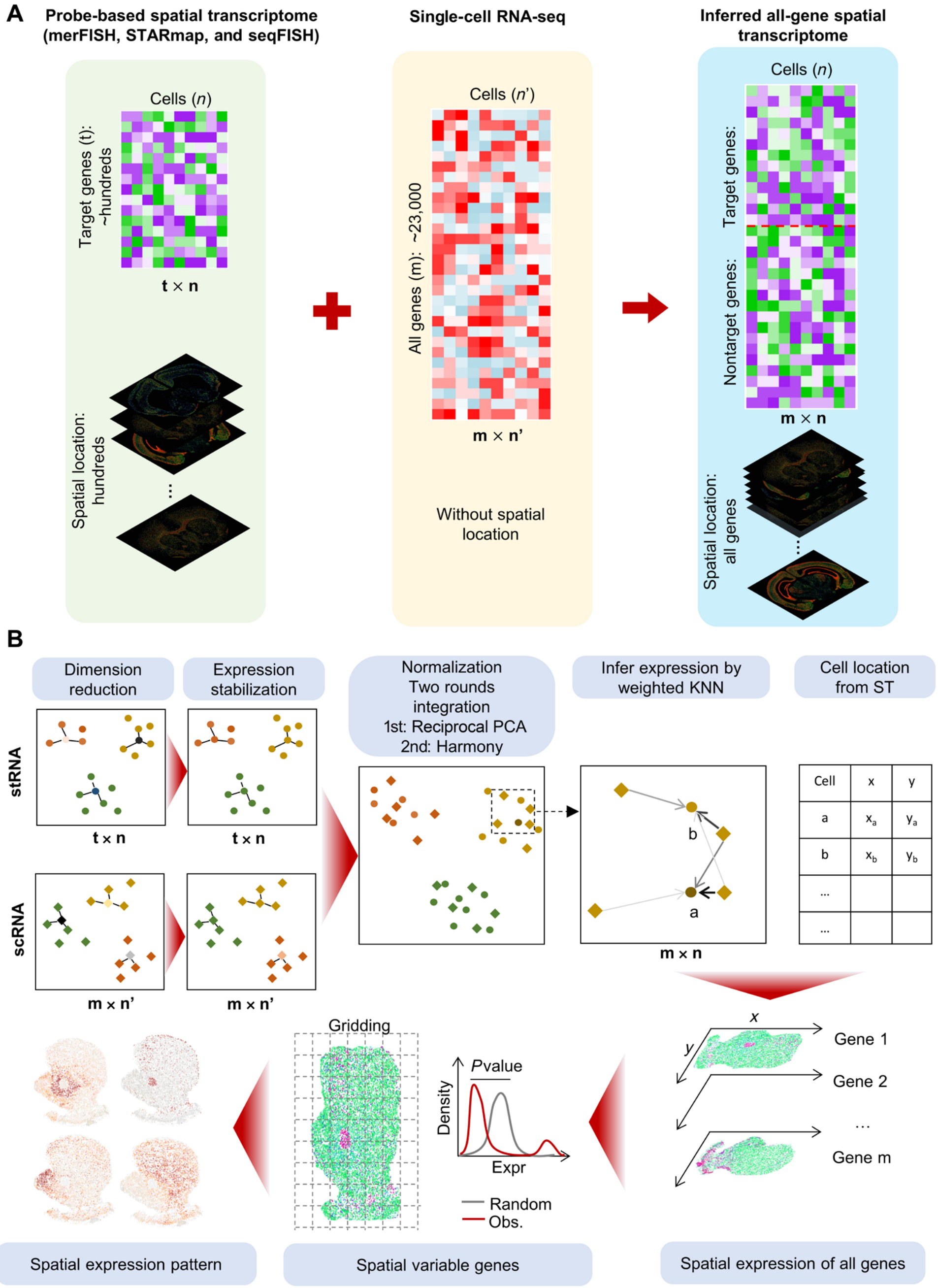

Fig. 1. Overview of iSpatial. (**A**) A diagram showing the rationale of iSpatial. The probe-based spatial transcriptome includes $t \times n$ (t, genes; n, cells) expression matrix and location of each cell. After integrating with $m \times n'$ scRNA-seq data, iSpatial infers the genome-wide transcriptional expression of all n cells. (**B**) The iSpatial pipeline consisted of (i) dimension reduction; (ii) expression stabilization (optional); (iii) expression normalization; (iv) inferring transcriptional expression; (v) spatial variable gene detection; and (vi) cluster spatial expression patterns.

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

removes potential noise/background expression based on the expression level of adjacent cells in principal components analysis (PCA) space. The two datasets were then normalized and embedded into a common space with two sequential rounds of integration: first by reciprocal PCA (RPCA) (*29*) and then through Harmony (*31*). On the basis of the common embedding, the expression value of each gene in each cell of the ST dataset is then inferred using a weighted k-nearest neighbors (KNN) model. Because the physical locations of these cells have been resolved in spatial profiling, the results represent a new single-cell gene expression profile with both genome-wide coverage and single-cell spatial resolution, which could be used for downstream analyses including identification of spatially variable genes (SVGs; Fig. 1B). Here, the genes that showed nonrandom distribution on spatial expression are defined as SVGs.

iSpatial outperforms existing tools in its accuracy on predicting spatial expression pattern

To evaluate the performance of iSpatial and compare it with existing tools, we used a mouse hippocampal dataset generated from Slide-seq V2 (*32*). Because this dataset includes the spatial expression of all genes, it can be used for evaluating the prediction performance (Fig. 2A). Specifically, we divided the dataset into the training and validation groups, which contain 3000 and ~20,000 genes, respectively. The training data group (mimic a ST dataset) was integrated with a scRNA-seq dataset covering the same brain region (hippocampus) (*1*, *33*) (but by a different method) to infer the expression level and spatial patterns of the genes in the validation data group. By comparing the inferred expression patterns with the "truth" determined by Slide-seq (validation data group), we found that iSpatial could predict the spatial expression pattern with high accuracy. For example, iSpatial inferred the expression of *Atp2b1*, *Prox1*, and *Fibcd1*, which were not included in the training data group, across the entire hippocampus, dentate gyrus, and CA1, respectively, consistent with the Slide-seq validation data and in situ hybridization (ISH) results from the Allen Brain Atlas (ABA; Fig. 2B) (*33*). We found that iSpatial could "enhance" the signals not well detected in the original data. For example, *Slit1*, *Tspan18*, *Efnb2*, *Car12*, and others were barely detectable in hippocampal cells by Slide-seq; thus, it was difficult to determine their spatial pattern. With iSpatial, the expression of these genes was clearly visible; thus, their spatial pattern could be clearly recognized. This is unlikely an artifact of imputation, as the spatial expression inferred by iSpatial is consistent with that of the ABA data (Fig. 2C and fig. S1A).

We further compared the performance of iSpatial with another two existing tools, Liger (*30*) and Seurat (*29*), on the same task using the Slide-seq dataset. Although these two methods could also infer the expression patterns of genes not included in the training data group, compared with iSpatial, the spatial patterns obtained from Liger and Seurat were more ambiguous with higher background in general (Fig. 2, B and C, and fig. S1A). To quantitatively benchmark these different methods, we calculated the expression and spatial correlation coefficient as well as the root mean square error (RMSE) between Slide-seq data (regarded as ground truth) and inferred results from iSpatial, Liger, or Seurat on each gene of the validation dataset. The results showed iSpatial exhibited significantly higher correlation coefficient and lower RMSE than the other methods across all gene groups with different expression levels, and the accuracy of prediction is positively correlated with the gene expression level (Fig. 2, D and E, and fig. S1B). In addition, cell type–specific expressed genes exhibit higher prediction accuracy (fig. S1C). This result suggests that iSpatial achieves higher prediction accuracy on functionally relevant genes. Furthermore, when comparing the SVGs identified from the original Slide-seq data with those identified from inferred data of different methods (fig. S1D), we found that iSpatial has the highest prediction power with area under the curve (AUC) greater than 0.84 on SVGs among the three methods (Fig. 2F).

We also used Stereo-seq data of an adult mouse coronal hemibrain section (*34*) to benchmark the performance of our method. Similar to Slide-seq V2, we randomly sampled 3000 genes as training dataset, and other genes as validation dataset (fig. S2A). After integrating with a single cell dataset of corresponding brain regions (*2*), we compared the performance of Liger, Seurat, and iSpatial on predicting the gene expression levels and patterns. The results showed that iSpatial achieved higher correlation than other methods on validation datasets (fig. S2, B to E). Collectively, these analyses indicate that iSpatial outperforms existing tools in terms of accuracy on predicting spatial expression pattern.

iSpatial is broadly applicable to different ST datasets

After validating the performance of iSpatial with Slide-seq and Stereo-seq data, we further tested whether iSpatial can be applied to other ST datasets generated from different tissues and with different techniques. To this end, we first used iSpatial to analyze a STARmap dataset that covered the primary visual cortex (V1) of mouse brain (Fig. 3A) (*9*). Although the original STARmap data only included 1020 gene targets, iSpatial successfully inferred the expression of over 20,000 genes by integrating a single-cell smart-seq dataset from ABA (Fig. 3B) (*35*). The spatial expression patterns of genes not included in the original STARmap data could be faithfully inferred by iSpatial. For example, the layer-specific expression of a number of genes was accurately predicted as evidenced by its similarity to that of the ABA ISH results (Fig. 3C). Notably, iSpatial not only correctly predicted the layer distribution of nontargeted genes but also detected the expression variation of certain genes based on their spatial locations. For example, it predicted (i) a high-to-low gradient of *Pvrl3* across upper cortical layers and (ii) strong expression of *Serinc2* and *Col5a1* in cortical layer VI but relatively weak expression in upper layer V, both of which were confirmed by ISH data (Fig. 3C).

In additional to the STARmap dataset, we analyzed a recently published MERFISH dataset of mouse striatum (*36*) with iSpatial (Fig. 3D). The original MERFISH dataset contained 253 target genes that allowed the identification of nine major cell types in the striatum (fig. S3A), with 175 target genes exhibiting significant enrichment in certain cell types (fig. S3B). By integrating this dataset with corresponding scRNA-seq data, iSpatial could infer the expression and location of ~9000 genes at single-cell resolution (Fig. 3E and fig. S3C), with over 2200 genes identified as cell type–specific expressed genes (fig. S3, D and E). The spatial patterns of inferred genes were largely consistent with those determined by ISH. For example, *Gpr37* was highly enriched in the anterior commissure, *Coch* formed a high-to-low gradient from the dorsolateral to the ventromedial striatum, and *Stard5* was specifically expressed in the medial nucleus accumbens (NAc); all these iSpatial inferred expression patterns are consistent with the results from ABA (Fig. 3F).

To globally evaluate the predication accuracy, we adopted a 10-fold cross validation approach and found that iSpatial showed higher correlation coefficient and lower RMSE than other methods in both datasets (fig. S4). Collectively, these results demonstrated the

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

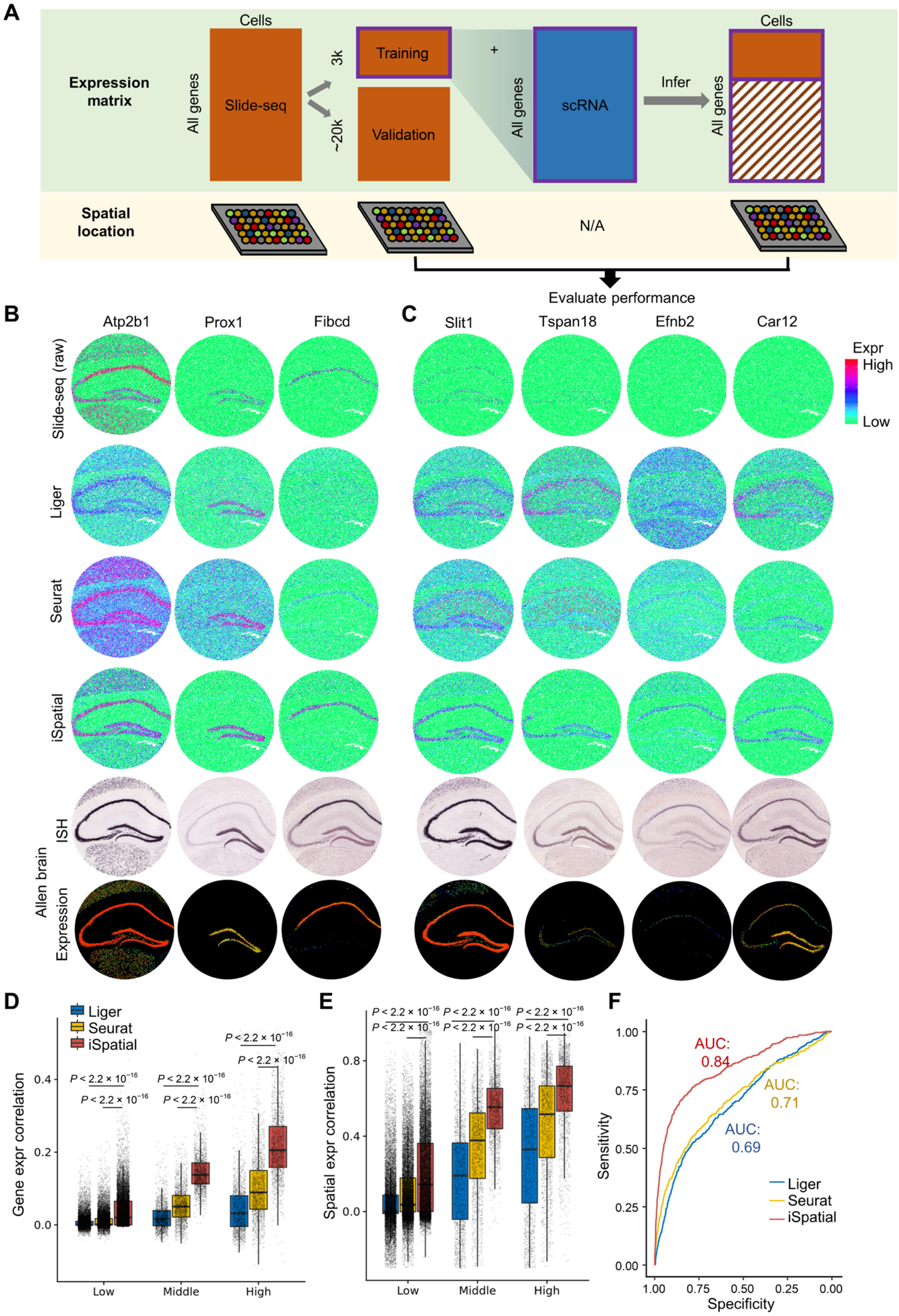

Fig. 2. Benchmarking the performance of iSpatial in inferring genome-wide spatial transcriptome. (**A**) A graphical illustration of the evaluation procedure. (**B**) Representative examples showing the performance of Liger, Seurat, and iSpatial in inferring spatial transcriptome. (**C**) Representative examples of inferred expression of genes barely detectable in raw Slide-seq data. (**D**) The gene expression correlation (Pearson's correlation) between Slide-seq raw data and Liger, Seurat, or iSpatial inferred data. The validation genes are divided into three groups on the basis of their expression levels. Two-sided Wilcoxon rank sum test was used. (**E**) Spatial expression correlation between inferred and raw Slide-seq data. Two-sided Wilcoxon rank sum test was used. (**F**) The receiver operating characteristic (ROC) curves comparing the prediction power of spatial variable genes by Liger, Seurat, and iSpatial. AUCs (area under the curves) are indicated.

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

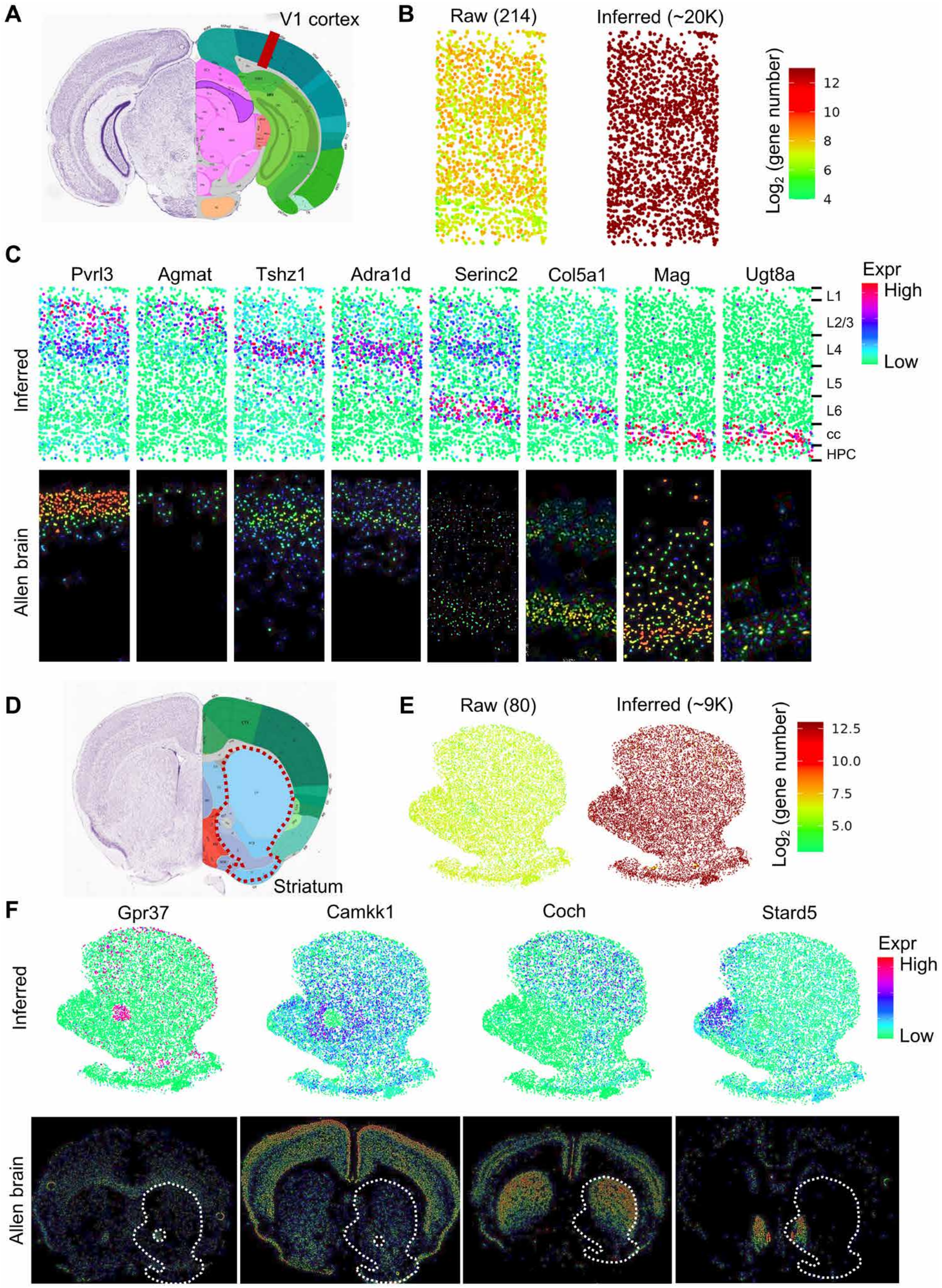

Fig. 3. iSpatial accurately infers the genome-wide spatial transcriptomes in mouse cortex and striatum. (**A**) Schematic of the anatomic region of mouse V1 cortex. (**B**) The numbers of detectable genes in each cell in raw STARmap (left) and after inferring by iSpatial (right). The mean numbers of detected genes in each cell are shown in brackets. (**C**) Inferred layer expression patterns of representative genes not targeted in the original STARmap library compared with the ISH data from the ABA. (**D**) Schematic of the anatomic region of mouse striatum. (**E**) The numbers of detectable genes in each cell in raw MERFISH (left) and after inferring by iSpatial (right). (**F**) Inferred spatial expression patterns of representative genes not targeted in the original MERFISH library compared with the ISH data from the ABA.

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

capacity of iSpatial in predicting the expression and location of genes using ST dataset generated from different tissues with different techniques.

iSpatial reduces FP and FN signals from spatial transcriptome

Although imaging-based transcriptomic assays have a higher detection efficiency when compared to that of sequencing-based methods, their performance is highly variable depending on the specific gene probes. For example, some transcripts are too short to be targeted by enough probes, which may lead to FN (dropout). On the other hand, some other genes may have close homologs that are difficult to distinguish with hybridization, leading to FP (background). We hypothesized that iSpatial could reduce these false signals by giving higher weights to cells of scRNA-seq when performing expression prediction, which were insensitive to gene length and could also unambiguously distinguish similar transcripts on the basis of sequence differences. To test this hypothesis, we first compared the expression pattern of some well-established cell type markers on the Uniform Manifold Approximation and Projection (UMAP) between the original STARmap data and iSpatial inferred data. We found that although these cell type–specific markers exhibited high enrichment in corresponding cell types, there were often FP signals in other cell types when analyzed by STARmap (Fig. 4A, top panels, *Slc17a7*, *Gad1*, *Plp1*, and *Cldn5*). In some cases, the expected expression patterns were not observed, likely due to FN (Fig. 4A, top panels, *Aqp4*). Consistent with our hypothesis, iSpatial could remove most FP signals from irrelevant cell types, without affecting the true-positive signals (Fig. 4A, bottom panels, *Slc17a7*, *Gad1*, *Plp1*, and *Cldn5*). Furthermore, iSpatial successfully inferred gene expression pattern that was missed in the original STARmap analysis (Fig. 4A, bottom panels, *Aqp4*), suggesting that iSpatial can also reduce FN results. On the basis of these findings, we further asked whether iSpatial could infer spatial patterns that were not well detected by STARmap. We found a number of genes whose expected layer-specific expression patterns were not detected in the original STARmap analysis. For example, *Nov*, *Rorb*, *Rspo1*, *Fezf2*, and *Foxp2* are established markers of different cortical layers. However, STARmap only detected sparse signals or even failed to detect real signals of these genes across all cortical layers (Fig. 4B). In contrast, iSpatial accurately captured the layer-specific expression patterns of these genes that were also detected by ISH (Fig. 4B).

In addition to the STARmap cortical dataset, a similar effect of iSpatial in correcting FP and FN expression on the MERFISH striatum data is also observed. Specifically, iSpatial removed most FP noise of known cell type–specific markers (fig. S5, A and B). It also accurately predicted the expression and spatial pattern of genes not well detected by MERFISH, such as *Tac2*, *Serpinb2*, and *Kctd4* (Fig. 4C and fig. S5C). Compared to the original MERFISH data, the spatial patterns inferred by iSpatial showed higher consistency with those determined by ISH (Fig. 4C and fig. S5C). For example, MERFISH indicates a broad distribution of *Serpinb2* across the striatum, but iSpatial suggested it was selectively expressed in a small group of cells located in the medial shell of NAc (Fig. 4C). ISH from ABA confirmed the accuracy of iSpatial's prediction (Fig. 4C), suggesting that iSpatial is capable of reducing noise. Collectively, these results showed that iSpatial can reduce FP and FN signals in the original ST data from different tissues generated by different techniques.

iSpatial enables whole transcriptome–level spatial analysis

One major goal of ST analysis is identifying SVGs, which are the molecular basis of structural/functional heterogeneity in different tissues. Because iSpatial could reliably infer genome-wide gene expression and their spatial locations, we sought to test whether iSpatial could augment the capability of a certain ST dataset in detecting SVGs and spatial gene expression patterns. To this end, we applied iSpatial to the STARmap cortex dataset to identify SVGs. We found that iSpatial inferred data markedly increased the number of detected SVGs (from 21 to 2122; fig. S6A). Clustering analysis of the SVGs revealed six major spatial patterns (fig. S6B), which resemble the known layer organization of mouse cortex. Notably, even when we restricted the analysis to the target genes of STARmap, iSpatial still identified more SVGs (162 in inferred data and 21 in original data), likely due to the correction of FP and FN signals in the original data (see above).

In addition to STARmap, we performed parallel analysis to evaluate iSpatial's effect on SVG identification using the MERFISH striatum dataset (Fig. 5A) and observed a similar increase in the SVGs number and statistic power of spatially variable test. Specifically, the SVG number increased by >20-fold (from 94 in the original data to 1968 in the inferred data; fig. S6C). Compared to the cerebral cortex, the anatomic organization of striatum is more ambiguous and less well understood, although recent studies have suggested distinct transcriptional features and cell types underlying its anatomic heterogeneity (*36*, *37*). By unbiased clustering analysis of the SVGs obtained from the iSpatial inferred data, we identified 12 distinct spatial patterns of SVGs (Fig. 5B). Many of these patterns closely resemble the known anatomic subregions in the striatum. For example, the C12 cluster is mainly expressed in the dorsal striatum, while C1 and C4 clusters are highly enriched in the NAc (Fig. 5, B and C). In addition, C7 and C11 clusters correspond to the core region of the NAc, while the C2, C9, and C10 clusters represent the shell region (Fig. 5, B and C) (*38*). The C8 cluster is specifically expressed in the medial shell (Fig. 5, B and C, and fig. S6D), a NAc subregion known to have distinct anatomic and functional features (*39*–*41*). These results indicate that iSpatial can facilitate identification of biologically relevant spatial gene expression patterns.

iSpatial improves analysis of ST data from liver

Having demonstrated the utility of iSpatial in the analysis of ST data from different brain regions, we next sought to test iSpatial's performance with ST data from other tissues. To this end, we analyzed a Vizgen MERFISH Mouse Liver Map dataset with 347 target genes included in the original data (https://vizgen.com/data-release-program/). By integrating the MERFISH data with a liver scRNA-seq dataset (*42*), iSpatial successfully inferred the expression of over 6000 genes on average in each single cell (Fig. 6, A and B), which increased by >20-fold from the original data. The inferred spatial patterns were largely consistent with established knowledge. For example, iSpatial predicted selective expression of *Slc1a2* and *Aldh1b1* in cells around the central vein (CV) and portal vein (PV), respectively (Fig. 6C). Similarly, *Cyp2e1* and *Cyp2f2* were predicted to be biased to CV and PV, but they have broader distribution than *Slc1a2* and *Aldh1b1* (Fig. 6C). All these spatial patterns were confirmed in previous studies (*43*, *44*). We further generated the UMAP on the basis of iSpatial inferred expression profile and found a close correlation between cells' positions on the UMAP and their in situ distribution along the CV-PV axis (Fig. 6, C and D), revealing a gradient expression profile along the CV-PV axis. Notably, although Liger and Seurat can also reveal a similar gradient expression pattern, a comparison among the three methods indicated that iSpatial achieved a higher specificity and accuracy, especially on genes with more spatially restricted expression patterns. For example, *Slc1a2* is selectively expressed in a monolayer surrounding the CV, which was accurately predicted by iSpatial (Fig. 6C), while Liger and Seurat revealed a more broad expression pattern (fig. S7, A and B). On the basis of the observed relationship between gene expression and spatial location of liver cells, we calculated a CV score for each cell (see Materials and Methods) to reflect its relative position to CV/PV, with a high/low CV score indicating close to CV/PV, respectively. As expected, the CV score showed gradual increase from PV to CV in both the liver tissue and the UMAP space (Fig. 6, E and F). Then, all the genes in the inferred dataset (from iSpatial) were ranked according to their correlation with the CV score, which enabled us to systematically

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

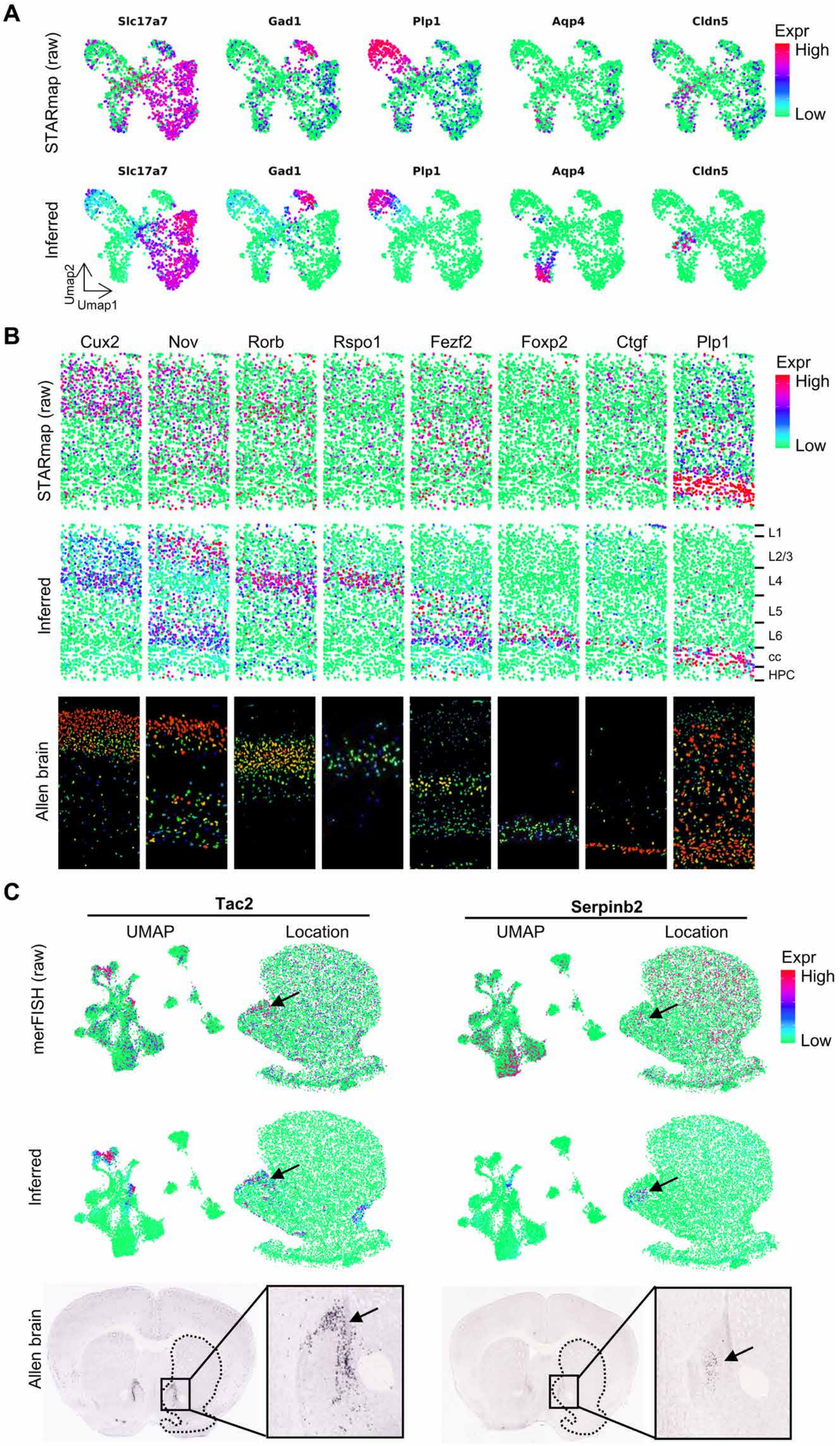

Fig. 4. iSpatial can reduce false-positive and false-negative signals in the original ST data. (**A**) UMAPs showing the expression levels of representative cell type markers in raw STARmap (top panels) and iSpatial inferred (bottom panels) data. Excitatory neuron (*Slc17a7*), inhibitory neuron (*Gad1*), oligodendrocyte (*Plp1*), astrocyte (*Aqp4*), and endothelial cell (*Cldn5*). (**B**) The spatial expression of cortex layer markers in the raw STARmap (top panels) and inferred by iSpatial (middle panels) compared with the ISH data from the ABA (bottom panels). Layer information: "L1 to L6," the six cortical layers; "cc," corpus callosum; "HPC," hippocampus. (**C**) The UMAP and spatial expression of *Tac2* and *Serpinb2* in the raw MERFISH (top panels) and inferred by iSpatial (middle panels) compared with the ISH data from the ABA (bottom panels).

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

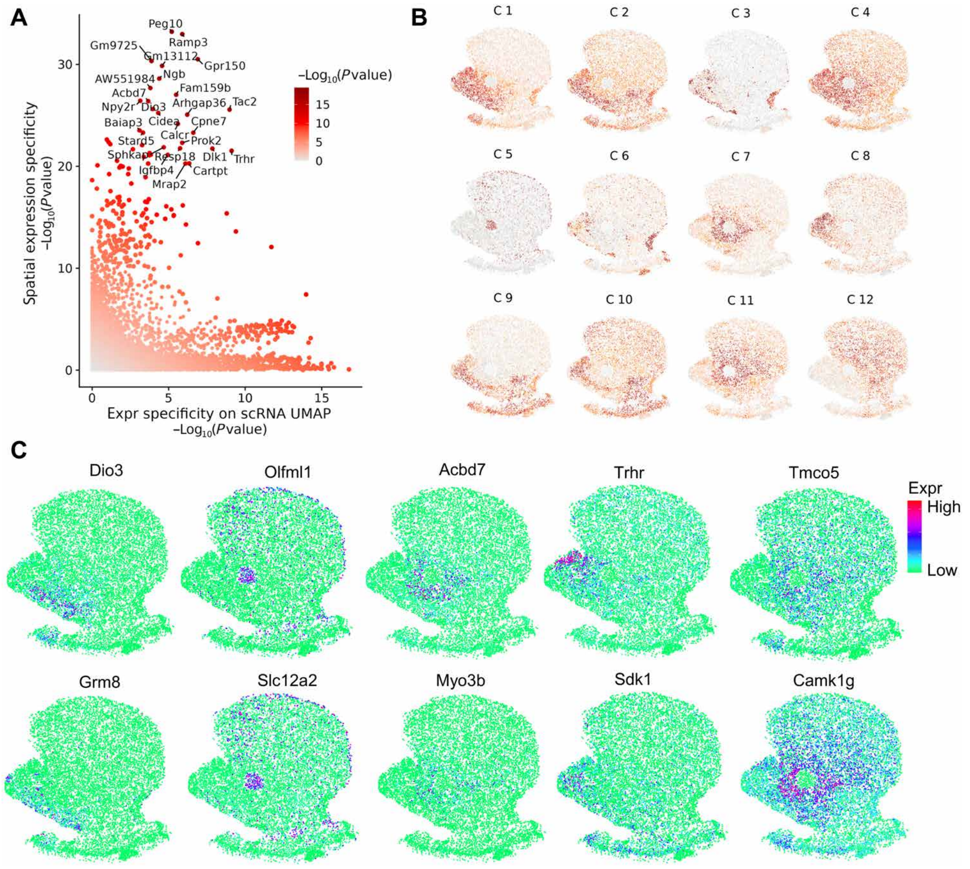

Fig. 5. iSpatial enables whole transcriptome–level spatial analysis. (**A**) Scatterplot showing genes with expression specificity in spatial location and UMAP projection of the corresponding scRNA-seq data. (**B**) The 12 clusters of the spatial variable genes in mouse striatum. The plot is color coded by the average gene expression in each cluster. (**C**) Inferred spatial expression signals of representative genes in different clusters. C1: *Dio3* and *Grm8*; C5: *Olfml1* and *Slc12a2*; C7: *Acbd7* and *Myo3b*; C8: *Sdk1* and *Trhr*; C11: *Camk1g* and *Tmco5*.

identify genes strongly related to the cell's spatial distribution (Fig. 6G). From this analysis, 141 and 692 genes with correlation to CV score >0.3 or <–0.3 were predicted to be strongly enriched in cells close to CV or PV. In contrast, only 3 and 17 of these SVGs were included in the original MERFISH dataset. As expected, many genes known to be biased to CV or PV were found, including *Gulo* and *Cyp2a5* enriched in cells adjacent to the CV, *Cdh1*, and *Etnppl* mainly expressed in cells around the PV (Fig. 6H). Furthermore, by applying Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis to CV and PV enriched genes, we found that they are involved in different biological functions (Fig. 6I). For example, the genes highly expressed around the CV were enriched in "drug metabolism" and "PPAR signaling pathway," while the genes highly expressed in the PV were enriched in "protein processing in endoplasmic reticulum (ER)" and "complement and coagulation cascades" (Fig. 6I and fig. S7C). These findings were consistent with previous reports (*43*, *44*). Together, the above analyses demonstrate that iSpatial can overcome the limited target gene numbers of various ST analyses to the whole-transcriptome level with high accuracy in different tissues.

DISCUSSION

ST assays simultaneously profile gene expression and their spatial location in tissue context, with the potential to unveil transcriptional features associated with tissue organization, cell-cell interaction, and region-specific physiological/pathological changes (*45*). Although sequencing and imaging-based ST techniques have been rapidly evolving (*5*), obtaining a genome-wide expression profile with single-cell spatial resolution is still challenging. To overcome this limitation, we developed a computational tool iSpatial to infer the genome-wide spatially resolved transcriptional information. iSpatial is especially useful for imaging-based ST analysis (such as MERFISH, seqFISH, and STARmap), which in general has high detection efficiency and single-cell/subcellular spatial resolution, but is usually limited by the predefined gene targets. By integrating such kind of ST data with corresponding scRNA-seq profiles, the expression levels of untargeted genes could be inferred from scRNA-seq data, while the spatial information is directly inherited from the ST data, enabling high-resolution spatial analysis at the whole-transcriptome scale.

To ensure accurate expression imputation, it is critical to account for intrinsic noise in different original datasets. Specifically, high-throughput scRNA-seq has low capture efficiency on mRNA molecules, leading to a large proportion of zero counts for expressed genes (dropout). On the other hand, because of the variable performance of different gene probes, imaging-based ST analysis may generate both FP and FN signals. iSpatial includes an expression stabilization step, which borrows the information from cells with similar global expression pattern to minimize random noise in the

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

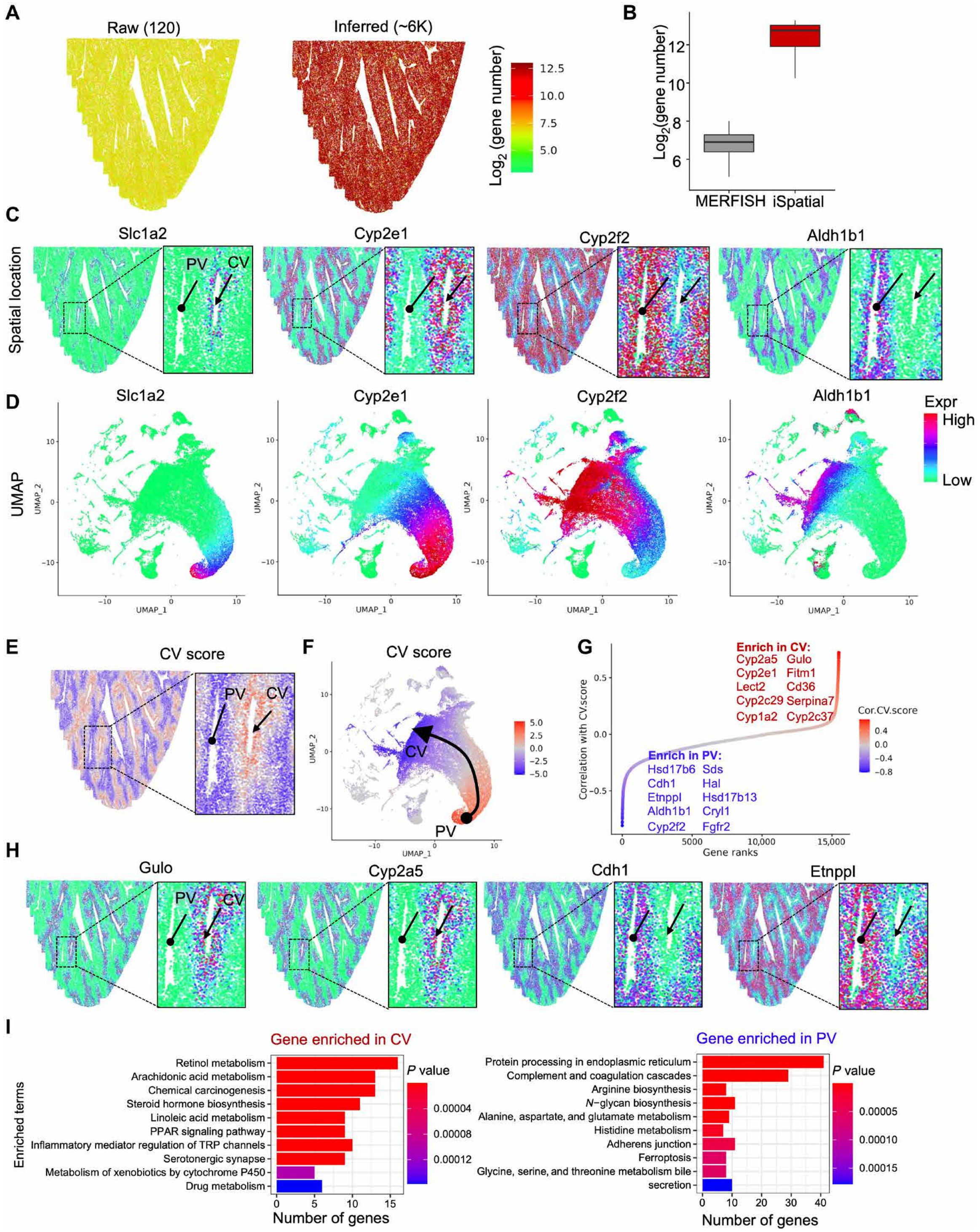

Fig. 6. iSpatial infers the spatial expression patterns in liver. (**A**) The numbers of detectable genes in each cell in raw MERFISH (left) and after inferring by iSpatial (right). (**B**) Boxplot showing the detectable gene numbers without or with interfering by iSpatial. (**C**) Representative examples of genes exhibiting spatial patterns enriched in central vein (CV; *Slc1a2* and *Cyp2e1*) or portal vein (PV; *Cyp2f2* and *Aldh1b1*). (**D**) UMAPs showing the expression levels of the genes shown in (C). The UMAPs were generated with iSpatial inferred expression. (**E**) The spatial location of each cell colored by CV score. (**F**) The UMAP of all cells colored by CV score. (**G**) The scatterplot showing the ranked correlations between gene expression level and CV score. The names of the top 10 enriched genes in PV or CV are listed. (**H**) Examples of genes selectively expressed nearby CV or PV. (**I**) The top 10 enriched Kyoto Encyclopedia of Genes and Genomes pathways of genes selectively expressed near CV or PV.

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

original data (fig. S8, A and B). The proper alignment of scRNA and ST is critical for this method. Harmony integration bias to keeping the global heterogeneity instead of local heterogeneity (fig. S8, C and D). iSpatial uses two-round integration to reduce potential technology bias and batch effect on PCA space, allowing accurate integration of ST and scRNA-seq datasets (fig. S8, C and D). A comparative analysis indicated that two-round integration resulted in more accurate prediction than one-round integration (fig. S9). As a result, iSpatial exhibits a significantly higher accuracy in both benchmark analysis with Slide-seq data and cross-validation with image-based ST data when compared with other existing imputation methods. Notably, iSpatial is not only able to faithfully predict the spatial expression of genes out of the original ST data but also shows robust performance on reducing the FP/FN signals in the raw data. In both cortical STARmap and striatal MERFISH datasets, iSpatial is able to remove random expression of cell type–specific markers in nonrelevant cell types and correct some nonspecific background in the raw data (probably caused by poor performance of certain probes) to generate spatial expression patterns that are consistent with established knowledge. Because it is difficult to directly validate the detected mRNA sequences that generate the signals in imaging-based ST techniques, iSpatial provides a useful method to evaluate the performance of different probes.

iSpatial is based on a KNN approach, where the setting of K value is important. Theoretically, large K value will dilute the signals for rare cell types, while small K value can not only increase specificity but also reduce coverage. In iSpatial, we used weighted KNN when performing expression inference: The neighbors close to the inquired cell will be assigned higher weights than neighbors far from the cell in expression imputation. This should reduce the oversmoothing effect for rare cell types when relatively large K is used, as the neighbors relatively far away from the rare cell types will have less impact on the inferred expression (fig. S10). Nevertheless, the optimal K value could vary for different datasets; thus, we implemented a function "recommend_k" in the iSpatial package to help the user determine the optimal K value.

In the real situation, the scRNA-seq and ST dataset would not always match exactly. To examine the performance of iSpatial under such situation, we compared Liger, Seurat, and iSpatial with unmatched scRNA-seq and ST datasets. Specifically, we used a MERFISH data covering a whole coronal mouse brain section (https://vizgen.com/data-release-program/) and an scRNA-seq dataset (*46*) generated from the mouse prefrontal cortex (fig. S11A). The results indicate that iSpatial can specifically infer the expression of cells in ST by inputting the corresponding scRNA-seq data (fig. S11B). *Cux1* and *Tle4* are established markers for cortical layer 2/3 and layer 6, respectively. iSpatial accurately inferred the expression patterns (fig. S11, C and D), demonstrating iSpatial can accurately impute the gene expression pattern in relevant cell types even when unmatched ST and scRNA-seq datasets were used.

We have tested iSpatial with multiple ST datasets generated from different tissues and techniques. In all the applications, iSpatial is able to expand spatial information from a predefined gene panel in the original ST data to the whole transcriptome, which renders several benefits for downstream analysis: First, it enables systematic identification of SVGs. In both the brain and liver datasets, we found that the number of SVGs increased from a few hundreds to several thousand after iSpatial imputation. Second, iSpatial enables the discovery of distinct spatial expression patterns across the tissue, which is achieved by spatial clustering analysis of SVGs. Many such unbiasedly identified expression patterns are biologically relevant. For example, we found that the SVG groups are organized into layer structure in the cortex, and they exhibit core/shell enrichment in the striatum, indicating a tight relationship between gene expression and tissue organization. Third, iSpatial enables bioinformatic analysis requiring sufficient gene number or high statistical power, such as KEGG analysis, to be performed on SVGs or SVG subgroups. For example, by inferring the transcriptome-wide spatial pattern in the liver, we found genes enriched in CV and PV are involved in distinct KEGG terms, suggesting a likely link between region-specific gene expression and function.

A potential limitation of iSpatial is that it requires corresponding ST and scRNA-seq data, which may not be always available. However, given the rapid development of ST and scRNA-seq techniques, and ongoing effort in large single-cell consortium and STs (*4*, *47*, *48*), we anticipate that iSpatial will be widely used to help understand the molecular basis of structural and functional heterogeneity in complex tissues of diverse organs in normal and disease states. To facilitate iSpatial implementation, we have made the iSpatial R package available at https://github.com/YiZhang-lab/iSpatial, where a tutorial on how to use iSpatial to integrate striatal scRNA-seq and MERFISH data to infer genome-wide spatial expression patterns can be found.

MATERIALS AND METHODS

iSpatial workflow

The workflow of iSpatial R-package contains the following steps: (i) expression stabilization, which removes possible FP and FN gene expression signals at single-cell level; (ii) integration, where two rounds of integration are performed to achieve accurate mapping of single-cell RNA-seq data and spatial transcriptome data; (iii) infer expression according to the weighted KNN; and (iv) downstream analysis, detection of spatial variable genes and patterns.

Expression stabilization

In the probe-based spatial transcriptome, we observed that some cell type markers are not fully detected in the corresponding cell cluster. However, these markers are indeed expressed across the cluster based on the scRNA-seq data. This indicates some probes have low binding affinities, which generates FN signals. We also found some markers could be detected in some cell types that should not be expressed (random distributed in whole slice). This can be caused by nonspecific binding of some probes, which cause FP signals. iSpatial tries to remove these FN/FP signals to correct the expression in each cell. We assume that these FN/FP signals are randomly distributed. Thus, it can be corrected using the cells showing similar global expression patterns but do not exhibit FN/FP on the same gene. To remove these FN and FP signals in each single cell *i*, we first find the KNNs ($KNN_{i,k}$) that most correlate with the cell *i* based on the global expression pattern.

E_i: the vector of gene expression in cell *i*

$\widehat{E}_i$: the vector of corrected gene expression in cell *i*

$KNN_{i,k}$: $\{KNN_{i,1}, KNN_{i,2}, \ldots, KNN_{i,k}\}$: the set of KNN of cell *i*

Then, we correct the gene expression $\widehat{E}_i$ by the expression of KNNs

$$\widehat{E}_i = \alpha E_i + (1-\alpha)\sum_k E_{\mathrm{KNN}_{i,k}}$$

where α controls the weight of expression from cell *i* and neighbors. By default, we set α = 0.5. Higher value strongly corrects the noise

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

by the KNNs, but loses cell specificity, which may cause lower detection power of some rarely expressed genes.

Integration of ST data and scRNA-seq data

To achieve accurate integration of ST data and scRNA-seq data, iSpatial uses a two-round integration approach. In the first round, we adopted an RPCA method from Seurat. At this step, the PCA spaces are calculated in both datasets. Then, one dataset is projected onto the other's PCA space and constrain the cell anchors with the same mutual neighbors. We use the "FindIntegrationAnchors" function with the parameter reduction = "rpca" in Seurat (version 4.0.5) to find the anchors and use "IntegrateData" to get the integrated data. For the second-round integration, we project the cells from both spatial transcriptome and scRNA-seq into a shared PCA embedding. Then, an iterative clustering method is used to remove the technology bias and batch effect on PCA space, which gets harmonious PCA embeddings for both datasets. Harmony (version 0.1.0) is then used to generate normalized PCA embeddings in this step.

Infer spatial expression

After integrating two datasets into one reduced dimension space, iSpatial uses a weighted KNN approach to infer the expression of nontargeted genes. For each cell t in spatial transcriptome data, iSpatial searches the KNNs ($KNN_{t,k}$).

t: cell in spatial transcriptome data

c: cell in scRNA-seq data

$KNN_{t,k}$: $\{KNN_{t,1}, KNN_{t,2}, \ldots, KNN_{t,k}\}$: the set of KNN of cell t

Then, $KNN_{t,k}$ are restricted to the cells from scRNA-seq data, because the cells from scRNA-seq data contain the expression information of the whole transcriptome

$$\mathrm{KNN}'_{t,k}\!:\!\mathrm{KNN}_{t,k} \in c$$

The final inferred expression $\widehat{E_t}$ of cell t is calculated by the expression of cell t itself and the expression of $KNN'_{t,k}$

$$\widehat{E_t} = (1-\beta)\,E_t + \beta\left(\textstyle\sum_k \omega_{t,k} E_{\mathrm{KNN}'_{t,k}}\right)$$

where ω is the weight of each neighbor k of cell t. For the genes targeted in cell t, β balances the expression from the spatial transcriptome and scRNA-seq. $\beta \in [0,1]$. For genes not measured in ST data, $\beta = 1$. The weights ω between cell t and its neighbor KNN'_t are defined by the normalized transcriptional distance $d_{t,k}$. Here, iSpatial uses 1 − Pearson's correlation coefficient to measure the distance

$$d_{t,k} = \mathrm{dist}(E_t, E_{\mathrm{KNN}'_{t,k}})$$

$$\omega_{t,k} = \frac{{d_{t,k}}^2}{\sum_k {d_{t,k}}^2}$$

and

$$\mathrm{dist}(E_t, E_{\mathrm{KNN}'_{t,k}}) = 1 - \mathrm{cor}(E_t, E_{\mathrm{KNN}'_{t,k}})$$
$$= \frac{\Sigma(E_t - \overline{E}_t)(E_{\mathrm{KNN}'_{t,k}} - \overline{E}_{\mathrm{KNN}'_{t,k}})}{\sqrt{\Sigma(E_t - \overline{E}_t)^2}\sqrt{\Sigma(E_{\mathrm{KNN}'_{t,k}} - \overline{E}_{\mathrm{KNN}'_{t,k}})^2}}$$

Identifying SVGs

To identify significant SVGs, the x and y axes of the spatial location are evenly divided into n bins, and then the spatial location is further grided into $n \times n$ grids. For each gene j, we calculate the average expression values E_j over the n^2 grids. We then randomly sample the spatial location of each cell and calculate the average expression E'_j over the randomly sampled n^2 grids. If a gene j has no specific spatial expression pattern, then the distribution of observed E_j should not be different from that of random E'_j. On the contrary, if a gene has strong spatial expression pattern, then the distribution E_j should be significantly different from E'_j. Thus, whether a gene exhibits spatial expression pattern depends on whether there is a difference between the distribution of E_j and E'_j. Here, we apply a nonparametric two-sided Mann-Whitney U test to determine the difference between E_j and E'_j. We also offer the Kolmogorov-Smirnov test to test the distribution differences.

Some studies found that a gene with spatial expression pattern always displays a specific expression bias on the scRNA-seq UMAP/t-distributed stochastic neighbor embedding (tSNE) projection. iSpatial also integrates scRNA-seq information into spatial variable gene detection. If a gene not only displays spatial expression pattern on spatial transcriptome data but also exhibits expression specificity on scRNA-seq UMAP/tSNE projection, then this gene has a higher confidence of spatial expression pattern. Similar to the detection of spatial expression gene on spatial location, iSpatial uses the same method to detect whether a gene displays a specific expression location on scRNA-seq UMAP/tSNE. To integrate these two lines of information, the final P value is equal to the P value from the ST data multiplied by the P value from the scRNA-seq data. Then, the adjusted P values are calculated to control the false discovery for multiple comparisons.

SVG clustering

The genes with spatial expression patterns could be grouped into clusters. For each gene, iSpatial captures the spatial expression features according to the average expression value over the n^2 grids, which was described before. On the basis of these features, Pearson's correlation coefficients are calculated for pairs of genes. Then, distances among genes are measured by 1 minus Pearson's correlation coefficients. Last, Hierarchical clustering is performed using "hclust" in R to spatial variable genes. "cutree" function from R is used to group these genes into desired number of groups.

Single-cell RNA-seq data processing

The initial gene × cell matrix for each study was downloaded according to the original papers. The expression matrix was then transferred into Seurat object for downstream analysis. The raw counts of gene expression profile of each cell were normalized to 10,000 counts and natural log transformed using the Seurat function "NormalizeData." To generate UMAP, we used standard pipeline from Seurat. In short, "FindVariableFeatures" was used to identify top variable genes, "ScaleData" was used to scale and center these genes in the data, and then PCA was performed by "RunPCA" on the basis of the selected features. Last, the top 30 principal components from PCA were used to generate the UMAP projection by "RunUMAP" with the parameters "dims = 1:30."

Mouse hippocampus data processing

The Slide-seq V2 of mouse hippocampus was downloaded from the Broad Institute single-cell portal website (https://singlecell.broadinstitute.org/single_cell/study/SCP815). Here, we only use the "Puck_200115_08" dataset (*32*). This dataset contains two files, the raw expression matrix and the barcode locations. The data processing of Slide-seq V2 followed the same procedure as that of

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

single-cell RNA-seq. The main difference is that Slide-seq contains spatial location of each cell. According to the vignettes of Seurat, the coordinate of each cell is stored as a "SlideSeq" class in Seurat object. For the scRNA-seq data of mouse hippocampus, we used a published dataset (*1*). To facilitate analysis, we used a preprocessed Seurat object (www.dropbox.com/s/cs6pii5my4p3ke3/mouse_hippocampus_reference.rds?dl=0) offered by the Satija Laboratory.

Mouse hemibrain data processing

The single-cell resolution Stereo-seq data of mouse hemibrains is downloaded from https://db.cngb.org/stomics/mosta/ (*34*). We first transfer single cell–level expression matrix to Seurat, and the spatial locations of each cell are inputted into Seurat object. Then, the cells with expressed genes over 500 are kept. The expression was normalized by NormalizeData. The corresponding scRNAseq data were downloaded from Linnarsson laboratory (http://mousebrain.org/adolescent/downloads.html) (*2*). We only used cells from the central nervous system and removed cells not in the brain regions profiled by Stereo-seq.

Mouse cortex data processing

The STARmap data of the mouse visual cortex were downloaded from the STARmap resource website (www.starmapresources.org/data). We chose the dataset "20180505_BY3_1kgenes" that profiles 1020 genes. The expression matrix data "cell_barcode_count.csv" were imported to Seurat object. The spatial location coordinate of each cell was extracted from "labels.npz" according to the method provided by the original paper (https://github.com/weallen/STARmap). The cell coordinates were integrated into Seurat object. The single-cell RNA-seq was downloaded from the ABA (https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq). The "mouse_VISp_2018-06-14_exon-matrix.csv" file was used to generate the expression profile. Low-quality cells were removed according to the meta data "mouse_VISp_2018-06-14_samples-columns.csv."

Mouse striatum data processing

The mouse striatum merFISH data were processed and normalized as described in the original paper (*36*). Briefly, the expression of each cell was normalized by cell size and total RNA counts. Then, the log-transformed data were applied to the expression matrix. Here, we used the data from a representative slice (slice 10). We used the preprocess Seurat object data in www.dropbox.com/s/ghkcovukgtctm76/NA_merFISH.RDS?dl=0. A down-sample version of these data is provided by the iSpatial package as a test dataset. After installing the R package, this command "data(NA_merFISH)" could load the merFISH data into the R environment.

The scRNA-seq data of mouse striatum were downloaded from GEO under accession GSE118020. To speed up the analysis, we down-sampled the full dataset to 10,000 cells using R function "sample." Then, these data were used to infer the spatial expression of all genes.

Mouse liver data processing

The mouse liver merFISH data are download from Vizgen MERFISH Mouse Liver Map (https://vizgen.com/data-release-program/), which targets 347 genes. This dataset contains multiply slices. We only used slice 3 from the first replicate in this manuscript. The merFISH data processing is the same as described above. The matched scRNA-seq data were download from GEO under the accession GSE166504 (*42*). The original dataset not only contains wild-type healthy livers but also livers with nonalcoholic fatty liver disease. Three healthy samples (Hepatocyte_Chow_Animal1_Capture1, Hepatocyte_Chow_Animal2_Capture1, and Hepatocyte_Chow_Animal3_Capture1) were used in this analysis.

CV score calculation

To measure the distance between each cell and CV, we calculated a CV score of each cell on the basis of well-known CV markers (*Slc1a2* and *Cyp2e1*) and PV markers (*Cyp2f2* and *Aldh1b1*). For each cell, the CV score was defined by the mean expression level of *Slc1a2* and *Cyp2e1* minus mean expression level of *Cyp2f2* and *Aldh1b1*. A positive value means more likely that the cell is located near the CV. On the contrary, a negative value indicates that the cell is located near the PV. After defining the CV score, we could then identify the genes specifically expressed in CV/PV. For each gene, we calculated the Pearson's correlation between the corresponding expression value and CV score across all cells. A positive correlation coefficient represents a gene preferentially expressed in CV, and a negative one represents a gene preferentially expressed in PV. According to the distribution of the correlation coefficient among all genes, we manually chose >0.3 or <−0.3 as cutoffs to define genes with most CV/PV bias.

Visualize the spatial expression

Visualization of gene spatial expression was achieved using the "SpatialFeaturePlot" in the Seurat package. iSpatial provides a function "spatial_signature_plot" for spatial visualization of the mean expression of a group of genes.

Benchmark methods

We chose two popular methods for comparison: Seurat (version 4.0.5) and rliger (version 1.0.0). The Seurat provides a "canonical correlation analysis" (CCA) method to integrate the spatial transcriptome and scRNA-seq. Here, we used the function "FindTransferAnchors" with parameter 'reduction = "cca"' to identify the integrated anchors. Then, genes expression values from scRNA-seq data were transferred to spatial transcriptome data using "TransferData" with the parameter 'weight.reduction = "cca".' Different from Seurat, Liger uses an integrative non-negative matrix factorization method to integrate the datasets, which is embedded in the function "optimizeALS." For imputing the nontargeted genes by rliger, we performed the following steps according to the rliger vignettes: "createLiger," "normalize," "scaleNotCenter," "optimizeALS," and "quantile_norm." Last, "imputeKNN" was used to search the nearest neighbors and impute the nontargeted genes.

Performance evaluation

The Slide-seq V2 data contain a total of 23,264 genes. We randomly sampled 3000 genes as the training dataset and used others as validation. After performing different methods to infer the spatial expression of all genes, we conducted the comparisons between the inferred expression values and the raw expression values. At cell level, we calculated the Pearson's correlation coefficient of each gene across all cells and then compared the overall differences of correlation coefficients among all genes of different methods. At spatial expression level, we grided the spatial location into 50 × 50 bricks. Then, we calculated the mean expression in each brick and compared the differences between raw expression and inferred one using 50 × 50

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

bricks as features. In addition, we compared the detection accuracies of spatial variable genes among three methods. The SVGs were also called using validation dataset as the ground truth. Receiver operating characteristic (ROC) curves and AUC were performed by R package pROC (version 1.18.0) (*49*).

Tenfold cross validation

For the datasets other than Slide-seq that we do not have the ground truth, we used 10-fold cross validation to evaluate the prediction performance. For the targeted genes in spatial transcriptome, we randomly separated it into 10 groups. Each time, we chose nine groups to infer the expression and used the left one to validate. After 10 rounds of inferring and validation, every gene was used to validate the prediction performance. Then, correlation of gene level and spatial expression level were calculated as described above.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abq0990

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. A. Saunders, E. Z. Macosko, A. Wysoker, M. Goldman, F. M. Krienen, H. de Rivera, E. Bien, M. Baum, L. Bortolin, S. Wang, A. Goeva, J. Nemesh, N. Kamitaki, S. Brumbaugh, D. Kulp, S. A. McCarroll, Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030.e16 (2018).
2. A. Zeisel, H. Hochgerner, P. Lonnerberg, A. Johnsson, F. Memic, J. van der Zwan, M. Haring, E. Braun, L. E. Borm, G. La Manno, S. Codeluppi, A. Furlan, K. Lee, N. Skene, K. D. Harris, J. Hjerling-Leffler, E. Arenas, P. Ernfors, U. Marklund, S. Linnarsson, Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
3. X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang, Y. Xu, W. Huang, M. Jiang, X. Jiang, J. Mao, Y. Chen, C. Lu, J. Xie, Q. Fang, Y. Wang, R. Yue, T. Li, H. Huang, S. H. Orkin, G. C. Yuan, M. Chen, G. Guo, Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107.e17 (2018).
4. O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, S. A. Teichmann, The Human Cell Atlas: From vision to reality. *Nature* **550**, 451–453 (2017).
5. V. Marx, Method of the year: Spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
6. P. L. Stahl, F. Salmen, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, A. Borg, F. Ponten, P. I. Costea, P. Sahlen, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisen, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
7. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
8. S. Shah, E. Lubeck, W. Zhou, L. Cai, seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron* **94**, 752–758.e1 (2017).
9. X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, G. P. Nolan, F. A. Bava, K. Deisseroth, Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
10. C. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G. C. Yuan, L. Cai, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
11. Y. Liu, M. Yang, Y. Deng, G. Su, A. Enninful, C. C. Guo, T. Tebaldi, D. Zhang, D. Kim, Z. Bai, E. Norris, A. Pan, J. Li, Y. Xiao, S. Halene, R. Fan, High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665–1681.e18 (2020).
12. S. Codeluppi, L. E. Borm, A. Zeisel, G. La Manno, J. A. van Lunteren, C. I. Svensson, S. Linnarsson, Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
13. L. E. Borm, A. M. Albiach, C. C. A. Mannens, J. Janusauskas, C. Özgün, D. Fernández-García, R. Hodge, E. S. Lein, S. Codeluppi, S. Linnarsson, Scalable in situ single-cell profiling by electrophoretic capture of mRNA. *bioRxiv* 2022.01.12.476082 (2022).
14. M. Asp, S. Giacomello, L. Larsson, C. Wu, D. Furth, X. Qian, E. Wardell, J. Custodio, J. Reimegard, F. Salmen, C. Osterholm, P. L. Stahl, E. Sundstrom, E. Akesson, O. Bergmann, M. Bienko, A. Mansson-Broberg, M. Nilsson, C. Sylven, J. Lundeberg, A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* **179**, 1647–1660.e19 (2019).
15. R. Dries, J. Chen, N. Del Rossi, M. M. Khan, A. Sistig, G. C. Yuan, Advances in spatial transcriptomic data analysis. *Genome Res.* **31**, 1706–1718 (2021).
16. B. Hie, J. Peters, S. K. Nyquist, A. K. Shalek, B. Berger, B. D. Bryson, Computational methods for single-cell RNA sequencing. *Annu. Rev. Biomed. Data Sci.* **3**, 339–364 (2020).
17. Z. Hu, A. A. Ahmed, C. Yau, CIDER: An interpretable meta-clustering framework for single-cell RNA-seq data integration and evaluation. *Genome Biol.* **22**, 337 (2021).
18. K. Polanski, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, J. E. Park, BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
19. B. Hie, B. Bryson, B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
20. L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
21. D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
22. W. V. Li, J. J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
23. W. Gong, I. Y. Kwak, P. Pota, N. Koyano-Nakagawa, D. J. Garry, DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 (2018).
24. G. C. Linderman, J. Zhao, M. Roulis, P. Bielecki, R. A. Flavell, B. Nadler, Y. Kluger, Zero-preserving imputation of single-cell RNA-seq data. *Nat. Commun.* **13**, 192 (2022).
25. R. Dong, G. C. Yuan, SpatialDWLS: Accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **22**, 145 (2021).
26. D. M. Cable, E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, R. A. Irizarry, Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2021).
27. T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, A. Segerstolpe, M. Zhang, I. Avraham-Davidi, S. Vickovic, M. Nitzan, S. Ma, A. Subramanian, M. Lipinski, J. Buenrostro, N. B. Brown, D. Fanelli, X. Zhuang, E. Z. Macosko, A. Regev, Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
28. V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, O. A. Bayraktar, Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
29. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
30. J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, E. Z. Macosko, Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887.e17 (2019).
31. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
32. R. R. Stickels, E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, F. Chen, Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
33. E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T. M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H. W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K. R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, A. R. Jones, Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
34. A. Chen, S. Liao, M. Cheng, K. Ma, L. Wu, Y. Lai, X. Qiu, J. Yang, J. Xu, S. Hao, X. Wang, H. Lu, X. Chen, X. Liu, X. Huang, Z. Li, Y. Hong, Y. Jiang, J. Peng, S. Liu, M. Shen, C. Liu, Q. Li, Y. Yuan, X. Wei, H. Zheng, W. Feng, Z. Wang, Y. Liu, Z. Wang, Y. Yang, H. Xiang, L. Han,

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

B. Qin, P. Guo, G. Lai, P. Munoz-Canoves, P. H. Maxwell, J. P. Thiery, Q. F. Wu, F. Zhao, B. Chen, M. Li, X. Dai, S. Wang, H. Kuang, J. Hui, L. Wang, J. F. Fei, O. Wang, X. Wei, H. Lu, B. Wang, S. Liu, Y. Gu, M. Ni, W. Zhang, F. Mu, Y. Yin, H. Yang, M. Lisby, R. J. Cornall, J. Mulder, M. Uhlen, M. A. Esteban, Y. Li, L. Liu, X. Xu, J. Wang, Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).

35. B. Tasic, Z. Yao, L. T. Graybuck, K. A. Smith, T. N. Nguyen, D. Bertagnolli, J. Goldy, E. Garren, M. N. Economo, S. Viswanathan, O. Penn, T. Bakken, V. Menon, J. Miller, O. Fong, K. E. Hirokawa, K. Lathia, C. Rimorin, M. Tieu, R. Larsen, T. Casper, E. Barkan, M. Kroll, S. Parry, N. V. Shapovalova, D. Hirschstein, J. Pendergraft, H. A. Sullivan, T. K. Kim, A. Szafer, N. Dee, P. Groblewski, I. Wickersham, A. Cetin, J. A. Harris, B. P. Levi, S. M. Sunkin, L. Madisen, T. L. Daigle, L. Looger, A. Bernard, J. Phillips, E. Lein, M. Hawrylycz, K. Svoboda, A. R. Jones, C. Koch, H. Zeng, Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

36. R. Chen, T. R. Blosser, M. N. Djekidel, J. Hao, A. Bhattacherjee, W. Chen, L. M. Tuesta, X. Zhuang, Y. Zhang, Decoding molecular and cellular heterogeneity of mouse nucleus accumbens. *Nat. Neurosci.* **24**, 1757–1771 (2021).

37. G. Stanley, O. Gokce, R. C. Malenka, T. C. Sudhof, S. R. Quake, Continuous and discrete neuron types of the adult murine striatum. *Neuron* **105**, 688–699.e8 (2020).

38. P. Voorn, L. J. Vanderschuren, H. J. Groenewegen, T. W. Robbins, C. M. Pennartz, Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.* **27**, 468–474 (2004).

39. D. C. Castro, K. C. Berridge, Opioid hedonic hotspot in nucleus accumbens shell: Mu, delta, and kappa maps for enhancement of sweetness "liking" and "wanting". *J. Neurosci.* **34**, 4239–4250 (2014).

40. R. Al-Hasani, J. G. McCall, G. Shin, A. M. Gomez, G. P. Schmitz, J. M. Bernardi, C. O. Pyo, S. I. Park, C. M. Marcinkiewcz, N. A. Crowley, M. J. Krashes, B. B. Lowell, T. L. Kash, J. A. Rogers, M. R. Bruchas, Distinct subpopulations of nucleus accumbens dynorphin neurons drive aversion and reward. *Neuron* **87**, 1063–1077 (2015).

41. J. W. de Jong, S. A. Afjei, I. Pollak Dorocic, J. R. Peck, C. Liu, C. K. Kim, L. Tian, K. Deisseroth, S. Lammel, A neural circuit mechanism for encoding aversive stimuli in the mesolimbic dopamine system. *Neuron* **101**, 133–151.e7 (2019).

42. Q. Su, S. Y. Kim, F. Adewale, Y. Zhou, C. Aldler, M. Ni, Y. Wei, M. E. Burczynski, G. S. Atwal, M. W. Sleeman, A. J. Murphy, Y. Xin, X. Cheng, Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).

43. K. B. Halpern, R. Shenhav, O. Matcovitch-Natan, B. Toth, D. Lemze, M. Golan, E. E. Massasa, S. Baydatch, S. Landen, A. E. Moor, A. Brandis, A. Giladi, A. S. Avihail, E. David, I. Amit, S. Itzkovitz, Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

44. F. Hildebrandt, A. Andersson, S. Saarenpaa, L. Larsson, N. Van Hul, S. Kanatani, J. Masek, E. Ellis, A. Barragan, A. Mollbrink, E. R. Andersson, J. Lundeberg, J. Ankarklev, Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nat. Commun.* **12**, 7046 (2021).

45. A. Rao, D. Barkley, G. S. Franca, I. Yanai, Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).

46. A. Bhattacherjee, M. N. Djekidel, R. Chen, W. Chen, L. M. Tuesta, Y. Zhang, Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nat. Commun.* **10**, 4169 (2019).

47. BRAIN Initiative Cell Census Network (BICCN), A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).

48. A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Gottgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef; Human Cell Atlas Meeting Participants, Science forum: The Human Cell Atlas. *eLife* **6**, e27041 (2017).

49. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, M. Muller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

Acknowledgments: We thank A. Bhattacherjee for some initial discussion and for the critical reading of the manuscript. **Funding:** This project was supported by 1R01DA042283, 1R01DA050589, and HHMI. R.C. is partially supported by the 2020 NARSAD Young Investigator Grant from the Brain and Behavior Research Foundation. Y.Z. is an investigator of the Howard Hughes Medical Institute. This article is subject to HHMI's Open Access to Publications policy. HHMI laboratory heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication. **Author contributions:** Y.Z. and C.Z. conceived the project. C.Z. developed the iSpatial package and performed the data analysis. C.Z., R.C., and Y.Z. designed the analyses, interpreted the results, and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Single-cell RNA-seq of mouse hippocampus was obtained from www.dropbox.com/s/cs6pii5my4p3ke3/mouse_hippocampus_reference.rds?dl=0:, and Slide-seq V2 of mouse hippocampus was downloaded from https://singlecell.broadinstitute.org/single_cell/study/SCP815. The Stereo-seq data of mouse hemibrains were from https://db.cngb.org/stomics/mosta/, and the corresponding scRNA-seq data were from Linnarsson laboratory (http://mousebrain.org/adolescent/downloads.html). The STARmap data of the mouse visual cortex were obtained from the STARmap resource website www.starmapresources.org/data, and corresponding scRNA-seq was downloaded from the Allen Brain Atlas (https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq). The mouse striatum MERFISH data were obtained from www.dropbox.com/s/ghkcovukgtctm76/NA_merFISH.RDS?dl=0, and the corresponding scRNA-seq was downloaded from GEO under accession GSE118020. The mouse liver MERFISH data were obtained from Vizgen MERFISH Mouse Liver Map (https://vizgen.com/data-release-program/), and the matched scRNA-seq data were download from GEO under the accession GSE166504. iSpatial package is available at Zenodo (https://doi.org/10.5281/zenodo.6828511) and GitHub (https://github.com/YiZhang-lab/iSpatial).

Submitted 17 March 2022
Accepted 13 July 2022
Published 26 August 2022
10.1126/sciadv.abq0990

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

Science Advances

Accurate inference of genome-wide spatial expression with iSpatial

Chao ZhangRenchao ChenYi Zhang

Sci. Adv., 8 (34), eabq0990. • DOI: 10.1126/sciadv.abq0990

View the article online
https://www.science.org/doi/10.1126/sciadv.abq0990
Permissions
https://www.science.org/help/reprints-and-permissions

Downloaded from https://www.science.org at Bank of Norway on August 26, 2022

Use of this article is subject to the Terms of service

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).