Accurate inference of genome-wide spatial expression with iSpatial

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

[1]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, Massachusetts 02115, USA; [2]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, Massachusetts 02115, USA; [3]Division of Hematology/Oncology, Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts 02115, USA; [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [5]Harvard Stem Cell Institute, WAB-149G, 200 Longwood Avenue, Boston, Massachusetts 02115, USA; †Co-first authors.

‡To whom correspondence should be addressed

E-mail: yzhang@genetics.med.harvard.edu

Keywords: Spatial transcriptomics, scRNA-seq, MERFISH, Slide-seq, STARmap, iSpatial

Manuscript information: 52 pages, 6 Figures, 11 Supplementary figures

HIGHLIGHTS

- iSpatial infers genome-wide spatial gene expression pattern by integrating spatial transcriptomic and scRNA-seq data
- iSpatial outperforms existing approaches in inferring spatial gene expression patterns
- iSpatial reduces false-positive/negative signals of the original spatial transcriptome
- iSpatial is applicable to spatial transcriptomic datasets from different tissues and techniques

ABSTRACT

Spatially resolved transcriptomic analyses can reveal molecular insights underlying tissue structure and context-dependent cell-cell or cell-environment interaction. Due to 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 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 their 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 for understanding the structure, function of complex tissues and disease processes.

53 **INTRODUCTION**

54

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

65

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

75

76 With the rapid development of scRNA-seq and ST technologies, new bioinformatic tools have
77 been developed to overcome the challenges in single-cell or spatial transcriptomic data analysis
78 (*15-20*). Several imputation methods for scRNA-seq data have emerged, including MAGIC (*21*),
79 scImpute (*22*), DrImpute (*23*), ALRA (*24*). However, methods developed for inferring ST data
80 are still limit. Notably, by integrating scRNA-seq and spatial resolved profiling data, recent
81 computational methods have leveraged the strength of different datasets and revealed
82 information that otherwise cannot be obtained from a single experimental paradigm. For
83 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) based on 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 is not straightforward, and different approaches have variable performance when applied to different datasets. Since 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 and false-negative signals in the original data. By applying iSpatial to datasets from different tissues (hippocampus, hemibrain, cortex, striatum, liver) and generated with different techniques (Slide-seq, Stereo-seq, MERFISH, STARmap), we have revealed both known and novel 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.

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 pre-defined 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 non-targeted genes in each spatially profiled cells could be inferred based on 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 removes potential noise/background expression based on the expression level of adjacent cells in PCA space. The two datasets were then normalized and embedded into a common space with two sequential rounds of integration: first by reciprocal PCA (*29*) and then through Harmony (*31*). Based on 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. Since 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**).

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*). Since 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 3,000 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

146 data group. By comparing the inferred expression patterns with the "truth" determined by Slide-
147 seq (validation data group), we found that iSpatial could predict the spatial expression pattern
148 with high accuracy. For example, iSpatial inferred the expression of *Atp2b1*, *Prox1* and *Fibcd1*,
149 which were not included in the training data group, across the entire hippocampus, dentate gyrus
150 (DG), and CA1, respectively, consistent with the Slide-seq validation data and ISH results from
151 Allen Brain Atlas (ABA) (*33*) (**Fig. 2B**). Interestingly, we found that iSpatial could "enhance"
152 the signals not well detected in the original data. For example, *Slit1*, *Tspan18*, *Efnb2*, *Car12* and
153 others were barely detectable in hippocampal cells by Slide-seq, thus difficult to determine their
154 spatial pattern. With iSpatial, the expression of these genes was clearly visible, thus their spatial
155 pattern could be clearly recognized. This is unlikely an artifact of imputation, as the spatial
156 expression inferred by iSpatial is consistent with that of the ABA data (**Fig. 2C and fig. S1A**).
157
158 We further compared the performance of iSpatial with another two existing tools, Liger (*30*) and
159 Seurat (*29*), on the same task using the Slide-seq dataset. Although these two methods could also
160 infer the expression patterns of genes not included in the training data group, compared with
161 iSpatial, the spatial patterns obtained from Liger and Seurat were more ambiguous with higher
162 background in general (**Fig. 2B, 2C and fig. S1A**). To quantitatively benchmark these different
163 methods, we calculated the expression and spatial correlation between Slide-seq data (regarded
164 as ground truth) and inferred results from iSpatial, Liger or Seurat on each gene of the validation
165 dataset. The results showed iSpatial exhibited significantly higher accuracy than the other
166 methods across all gene groups with different expression levels, and the accuracy of prediction is
167 positively correlated with the gene expression level (**Fig. 2D and 2E**). In addition, cell type
168 specifically expressed genes exhibit higher prediction accuracy (**fig. S1B**). This result suggests
169 that iSpatial achieves higher prediction accuracy on functionally relevant genes. Furthermore,
170 when compared the SVGs identified from the original Slide-seq data with those identified from
171 inferred data of different methods (**fig. S1C**), we found that iSpatial has the highest prediction
172 power with area under the curve (AUC) over 0.84 on SVGs among the three methods (**Fig. 2F**).
173
174 We also used Stereo-seq data of an adult mouse coronal hemibrain section (*34*) to benchmark the
175 performance of our method. Similar to Slide-seq V2, we randomly sampled 3,000 genes as
176 training dataset, and other genes as validation dataset (**fig. S2A**). After integrating with a single

177 cell dataset of corresponding brain regions (*2*), we compared the performance of Liger, Seurat
178 and iSpatial on predicting the gene expression levels and patterns. The results showed that
179 iSpatial achieved higher correlation than other methods on validation datasets (**fig. S2B-E**).
180 Collectively, these analyses indicate that iSpatial outperforms existing tools in terms of accuracy
181 on predicting spatial expression pattern.
182
183 **iSpatial is broadly applicable to different spatial transcriptomic datasets**
184 After validating the performance of iSpatial with Slide-seq and Stereo-seq data, we further tested
185 whether iSpatial can be applied to other ST datasets generated from different tissues and with
186 different techniques. To this end, we first used iSpatial to analyze a STARmap dataset which
187 covered the primary visual cortex (V1) of mouse brain (*9*) (**Fig. 3A**). Although the original
188 STARmap data only included 1,020 gene targets, iSpatial successfully inferred the expression of
189 over 20,000 genes by integrating a single cell smart-seq dataset from ABA (*35*) (**Fig. 3B**).
190 Importantly, the spatial expression patterns of genes not included in the original STARmap data
191 could be faithfully inferred by iSpatial. For example, the layer-specific expression of a number of
192 genes were accurately predicted as evidenced by its similarity to that of the ABA ISH results
193 (**Fig. 3C**). Notably, iSpatial not only correctly predicted the layer distribution of non-targeted
194 genes, but also detected the expression variation of certain genes based on their spatial locations.
195 For example, it predicted: 1) a high-to-low gradient of *Pvrl3* across upper cortical layers, and 2)
196 strong expression of *Serinc2* and *Col5a1* in cortical layer VI but relatively weak expression in
197 upper layer V, both of which were confirmed by ISH data (**Fig. 3C**).
198
199 In additional to the STARmap dataset, we analyzed a recently published MERFISH dataset of
200 mouse striatum (*36*) with iSpatial (**Fig. 3D**). The original MERFISH dataset contained 253 target
201 genes which allowed the identification of 9 major cell types in striatum (**fig. S3A**), with 175
202 target genes exhibit significant enrichment in certain cell types (**fig. S3B**). By integrating this
203 dataset with corresponding scRNA-seq data, iSpatial could infer the expression and location of
204 ~9,000 genes at single-cell resolution (**Fig. 3E and fig. S3C**), with over 2,200 genes identified as
205 cell type specific expressed genes (**fig. S3D and S3E**). Importantly, the spatial patterns of
206 inferred genes were largely consistent with those determined by ISH. For example, *Gpr37* was
207 highly enriched in the anterior commissure (AC), *Coch* formed a high-to-low gradient from the

208 dorsolateral to the ventromedial striatum, and *Stard5* was specifically expressed in the medial
209 nucleus accumbens (NAc), all these iSpatial inferred expression patterns are consistent with the
210 results from ABA (**Fig. 3F**).
211
212 To globally evaluate the predication accuracy, we adopted a 10-fold cross validation approach
213 and found that iSpatial showed significantly higher accuracy than Liger and Seurat in predicting
214 the expression and spatial location of genes in both datasets (**fig. S4**). Collectively, these results
215 demonstrated the capacity of iSpatial in predicting the expression and location of genes using ST
216 dataset generated from different tissues with different techniques.
217
218 **iSpatial reduces false-positive and false-negative signals from spatial transcriptome**
219 Although imaging-based transcriptomic assays have a higher detection efficiency when
220 compared to that of sequencing-based methods, their performance is highly variable depending
221 on the specific gene probes. For example, some transcripts are too short to be targeted by enough
222 probes, which may lead to false-negative (dropout). On the other hand, some other genes may
223 have close homologs which are difficult to distinguish with hybridization, leading to false-
224 positive (background). We hypothesized that iSpatial could reduce these false signals by giving
225 higher weights to cells of scRNA-seq when performing expression prediction, which were
226 insensitive to gene length and could also unambiguously distinguish similar transcripts based on
227 sequence differences. To test this hypothesis, we first compared the expression pattern of some
228 well-establish cell type markers on the UMAP between the original STARmap data and iSpatial
229 inferred data. We found that although these cell type-specific markers exhibited high enrichment
230 in corresponding cell types, there were often false-positive signal in other cell types when
231 analyzed by STARmap (**Fig. 4A**, upper panels, *Slc17a7*, *Gad1*, *Plp1* and *Cldn5*). In some cases,
232 the expected expression patterns were not observed, likely due to false-negative (**Fig. 4A**, upper
233 panels, *Aqp4*). Consistent with our hypothesis, iSpatial could remove most false-positive signals
234 from irrelevant cell types, without affecting the true positive signals (**Fig. 4A**, lower panels,
235 *Slc17a7*, *Gad1*, *Plp1* and *Cldn5*). Furthermore, iSpatial successfully inferred gene expression
236 pattern that was missed in the original STARmap analysis (**Fig. 4A**, lower panels, *Aqp4*),
237 suggesting that iSpatial can also reduce false-negative results. Based on these findings, we
238 further asked whether iSpatial could infer spatial patterns which were not well detected by

239 STARmap. Indeed, we found a number of genes, whose expected layer-specific expression
240 patterns were not detected in the original STARmap analysis. For example, *Nov*, *Rorb*, *Rspo1*,
241 *Fezf2*, *Foxp2* are established markers of different cortical layers. However, STARmap only
242 detected sparse signals or even failed to detect real signals of these genes across all cortical
243 layers (**Fig. 4B**). In contrast, iSpatial accurately captured the layer-specific expression patterns of
244 these genes that were also detected by ISH (**Fig. 4B**).
245
246 In addition to the STARmap cortical dataset, a similar effect of iSpatial in correcting false-
247 positive and false-negative expression on the MERFISH striatum data is also observed.
248 Specifically, iSpatial removed most false-positive noise of known cell type-specific markers (**fig.**
249 **S5A and B**). It also accurately predicted the expression and spatial pattern of genes not well
250 detected by MERFISH, such as *Tac2*, *Serpinb2* and *Kctd4* (**Fig. 4C and fig. S5C**). Compared to
251 the original MERFISH data, the spatial patterns inferred by iSpatial showed higher consistency
252 with those determined by ISH (**Fig. 4C and fig. S5C**). For example, MERFISH indicates a broad
253 distribution of *Serpinb2* across the striatum, but iSpatial suggested it was selectively expressed in
254 a small group of cells located in the medial shell of NAc (**Fig. 4C**). ISH from ABA confirmed
255 the accuracy of iSpatial's prediction (**Fig. 4C**), suggesting that iSpatial is capable of reducing
256 noise. Collectively, these results showed that iSpatial can reduce false-positive and false-
257 negative signals in the original ST data from different tissues generated by different techniques.
258
259 **iSpatial enables whole-transcriptome level spatial analysis**
260 One major goal of ST analysis is identifying SVGs, which are the molecular basis of
261 structural/functional heterogeneity in different tissues. Since iSpatial could reliably infer
262 genome-wide gene expression and their spatial locations, we sought to test whether iSpatial
263 could augment the capability of certain ST dataset in detecting SVGs and spatial gene expression
264 patterns. To this end, we applied iSpatial to the STARmap cortex dataset to identify SVGs. We
265 found that iSpatial inferred data dramatically increased the number of detected SVGs (from 21 to
266 2,122, **fig. S6A**). Clustering analysis of the SVGs revealed six major spatial patterns (**fig. S6B**),
267 which resemble the known layer organization of mouse cortex. Notably, even when we restricted
268 the analysis to the target genes of STARmap, iSpatial still identified more SVGs (162 in inferred

269 data and 21 in original data), likely due to the correction of false-positive and false-negative
270 signals in the original data (see above).
271
272 In addition to STARmap, we performed parallel analysis to evaluate iSpatial's effect on SVG
273 identification using the MERFISH striatum dataset (**Fig. 5A**) and observed a similar increase in
274 the SVGs number and statistic power of spatially variable test. Specifically, the SVG number
275 increased by > 20 folds (from 94 in the original data to 1,968 in the inferred data) (**fig. S6C**).
276 Compared to cerebral cortex, the anatomic organization of striatum is more ambiguous and less
277 well understood, although recent studies have suggested that distinct transcriptional features and
278 cell types underlying its anatomic heterogeneity (*36, 37*). By unbiased clustering analysis of the
279 SVGs obtained from the iSpatial inferred data, we identified 12 distinct spatial patterns of SVGs
280 (**Fig. 5B**). Importantly, many of these patterns closely resemble the known anatomic subregions
281 in the striatum. For example, the C12 cluster is mainly expressed in the dorsal striatum, while C1
282 and C4 clusters are highly enriched in the NAc (**Fig. 5B and 5C**). Additionally, C7 and C11
283 clusters correspond to the core region of the NAc, while the C2, C9 and C10 clusters represent
284 the shell region (*38*) (**Fig. 5B and 5C**). Interestingly, the C8 cluster is specifically expressed in
285 the medial shell (**Fig. 5B, 5C and fig. S6D**), a NAc subregion known to possess distinct
286 anatomic and functional features (*39-41*). These results indicate that iSpatial can facilitate
287 identification of biologically relevant spatial gene expression patterns.
288
289 **iSpatial improves analysis of spatial transcriptomic data from liver**
290 Having demonstrated the utility of iSpatial in the analysis of ST data from different brain
291 regions, we next sought to test iSpatial's performance with ST data from other tissues. To this
292 end, we analyzed a Vizgen MERFISH Mouse Liver Map dataset with 347 target genes included
293 in the original data (https://vizgen.com/data-release-program/). By integrating the MERFISH
294 data with a liver scRNA-seq dataset (*42*), iSpatial successfully inferred the expression of over
295 6,000 genes on average in each single cell (**Fig. 6A and 6B**), which increased by > 20 folds from
296 the original data. Importantly, the inferred spatial patterns were largely consistent with
297 established knowledge. For example, iSpatial predicted selective expression of *Slc1a2* and
298 *Aldh1b1* in cells around the central vein (CV) and portal vein (PV), respectively (**Fig. 6C**).
299 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 UMAP based on iSpatial inferred expression profile and found a close correlation between cells' positions on UMAP and their *in-situ* distribution along the CV-PV axis (**Fig. 6C and 6D**), revealing a gradient expression profile along the CV-PV axis. Notably, although Liger and Seurat can also reveal a similar gradient expression patten, 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 broader expression pattern (**fig. S7A and S7B**). Based on the observed relationship between gene expression and spatial location of liver cells, we calculated a CV score for each cell (see 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. 6E and 6F**). 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 identify genes strongly related to 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 spatially variable genes 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 CV, *Cdh1* and *Etnppl* mainly expressed in cells around PV (**Fig. 6H**). Furthermore, by applying 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 CV were enriched in "drug metabolism" and "PPAR signaling pathway", while the genes highly expressed in PV were enriched in "protein processing in ER", and "complement and coagulation cascades" (**Fig. 6I and fig. S7C**). These findings were consistent with previous reports (*43, 44*). Taken 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.

330 **DISCUSSION**

331

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

345

346 To ensure accurate expression imputation, it is critical to account for intrinsic noise in different
347 original datasets. Specifically, high-throughput scRNA-seq has low capture efficiency on mRNA
348 molecules, leading to a large proportion of zero counts for expressed genes (dropout). On the
349 other hand, due to the variable performance of different gene probes, imaging-based ST analysis
350 may generate both false-positive and false-negative signals. iSpatial includes an expression
351 stabilization step, which borrows the information from cells with similar global expression
352 pattern to minimize random noise in the original data (**fig. S8A and B**). The proper alignment of
353 scRNA and ST is critical for this method. Harmony integration bias to keeping the global
354 heterogeneity instead of local heterogeneity (**fig. S8C and D**). iSpatial uses two-rounds
355 integration to reduce potential technology bias and batch effect on PCA space, allowing accurate
356 integration of ST and scRNA-seq datasets (**fig. S8C and D**). A comparative analysis indicated
357 that two-round integration resulted in more accurate prediction than one-round integration (**fig.
358 S9**). As a result, iSpatial exhibits a significantly higher accuracy in both benchmark analysis with
359 Slide-seq data and cross-validation with image-based ST data when compared with other existing
360 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 false-positive/negative 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 non-relevant cell types and correct some non-specific background in the raw data (probably caused by poor performance of certain probes) to generate spatial expression patterns that are consistent with established knowledge. Since 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 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 over-smoothing 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 exam 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 a scRNA-seq dataset (*46*) generated from 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. S11C 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 pre-defined 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. Importantly, 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 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. But given the rapid development of ST and scRNA-seq techniques, and ongoing effort in large single-cell consortium and spatial transcriptomics (*4, 47, 48*), we anticipate that iSpatial will be widely used to help understanding 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: 1) Expression stabilization, which removes possible false positive and false negative gene expression signals at single cell level; 2) Integration. Two rounds of integration are performed to achieve accurate mapping of single cell RNA-seq data and spatial transcriptome data; 3) Infer expression according to the weighted k-nearest neighbors (KNN); 4) Downstream analysis. Detection of spatial variable genes and patterns.

Expression stabilization

In the probe-based spatial transcriptome, we observed some cell type markers are not fully detected in 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 false negative (FN) signals. We also found some markers could be detected in some cell types which should not be expressed (random distributed in whole slice). This can be caused by nonspecific binding of some probes, which cause false positive (FP) signals. iSpatial try 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 show similar global expression patterns, but don't exhibit FN/FP on the same gene. To remove these FN and FP signals in each single cell i, we first find the K-nearest neighbors ($KNN_{i,k}$) which most correlate with the cell i based on the global expression pattern.

E_i: the vector of gene expression in cell i

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

$KNN_{i,k}$: $\{KNN_{i,1}, KNN_{i,2}, \dots, KNN_{i,k}\}$: the set of k-nearest neighbor of cell i

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

$$\hat{E}_i = \alpha E_i + (1-\alpha) \sum_k E_{KNN_{i,k}}$$

where α controls the weight of expression from cell i and neighbors. By default, we set $\alpha = 0.5$. Higher value strongly corrects the noise by the KNNs, but loss cell specificity, which may cause lower detection power of some rarely expressed genes.

449

450 ***Integration of spatial transcriptomic data and scRNA-seq data***

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

462

463 ***Infer spatial expression***

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

467 t: cell in spatial transcriptome data

468 c: cell in scRNA-seq data

469 $KNN_{t,k}$: $\{KNN_{t,1}, KNN_{t,2}, \dots, KNN_{t,k}\}$: the set of k-nearest neighbor of cell t

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

472 $KNN'_{t,k}$: $KNN_{t,k} \in c$

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

475 $$\widehat{E_t} = (1-\beta)E_t + \beta(\sum_k \omega_{t,k} E_{KNN'_{t,k}})$$

476 where ω is the weight of each neighbor k of cell t. For the genes targeted in cell t, β balance the
477 expression from spatial transcriptome and scRNA-seq. $\beta \in [0,1]$. For genes not measured in ST
478 data, the $\beta = 1$. The weights ω between cell t and its neighbor KNN'_t are defined by the

normalized transcriptional distance $d_{t,k}$. Here, iSpatial use 1 - Pearson's correlation coefficient to measure the distance:

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

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

And,

$$dist\left(E_t, E_{KNN'_{t,k}}\right) = 1 - cor\left(E_t, E_{KNN'_{t,k}}\right)$$

$$= \frac{\sum(E_t - \bar{E}_t)(E_{KNN'_{t,k}} - \bar{E}_{KNN'_{t,k}})}{\sqrt{\sum(E_t - \bar{E}_t)^2}\sqrt{\sum(E_{KNN'_{t,k}} - \bar{E}_{KNN'_{t,k}})^2}}$$

Identifying spatially variable genes

To identify significant spatially variable genes, the x and y axes of the spatial location are evenly divided into n bins, 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, 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, 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 non-parametric two-sides 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 display a specific expression bias on scRNA-seq UMAP/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

506 UMAP/tSNE projection, this gene has a higher confidence of spatial expression pattern. Similar
507 to the detection of spatial expression gene on spatial location, iSpatial uses the same method to
508 detect whether a gene displays a specific expression location on scRNA-seq UMAP/tSNE. To
509 integrate these two lines of information, the final p value is equal to the p value from the spatial
510 transcriptomic data multiplied by the p value from the scRNA-seq data. Then the adjusted p
511 values are calculated to control the false discovery for multiple comparisons.
512
513 ***Spatially variable gene clustering***
514 The genes with spatial expression patterns could be grouped into clusters. For each gene, iSpatial
515 captures the spatial expression features according to the average expression value over the n^2
516 grids, which was described before. Based on these features, Pearson's correlation coefficients are
517 calculated for pairs of genes. Then, distances among genes are measured by 1 minus Pearson's
518 correlation coefficients. Finally, Hierarchical clustering is performed using "hclust" in R to
519 spatial variable genes. "cutree" function from R is used to group these genes into desired number
520 of groups.
521
522 ***Single-cell RNA-seq data processing***
523 The initial gene×cell matrix for each study was downloaded according to the original papers. The
524 expression matrix was then transferred into Seurat object for downstream analysis. The raw
525 counts of gene expression profile of each cell were normalized to 10,000 counts and natural log
526 transformed using the Seurat function "NormalizeData". To generate UMAP, we used standard
527 pipeline from Seuart. In short, "FindVariableFeatures" was used to identify top variable genes,
528 "ScaleData" was used to scale and center these genes in the data, then PCA was performed by
529 "RunPCA" based on the selected features. Finally, the top 30 principal components from PCA
530 were used to generate the UMAP projection by "RunUMAP" with the parameters "dims=1:30".
531
532 ***Mouse hippocampus data processing***
533 The Slide-seq V2 of mouse hippocampus was download from the Broad Institute single cell
534 portal website (https://singlecell.broadinstitute.org/single_cell/study/SCP815). Here, we only use
535 the "Puck_200115_08" dataset (*32*). This dataset contains two files, the raw expression matrix
536 and the barcode locations. The data processing of Slide-seq V2 followed the same procedure as

537 that of single cell RNA-seq. The main difference is that Slide-seq contains spatial location of
538 each cell. According to the vignettes of Seurat, the coordinate of each cell is stored as a
539 "SlideSeq" class in Seurat object. For the scRNA-seq data of mouse hippocampus, we used a
540 published dataset (*1*). To facilitate analysis, we used a pre-processed Seurat object
541 (https://www.dropbox.com/s/cs6pii5my4p3ke3/mouse_hippocampus_reference.rds?dl=0)
542 offered by the Satija Lab.
543
544 ***Mouse hemibrain data processing***
545 The single cell resolution Stereo-seq data of mouse hemibrains is download from
546 (https://db.cngb.org/stomics/mosta/) (*34*). We first transfer single cell level expression matrix to
547 Seurat. And the spatial locations of each cell are inputted into Seurat object. Then, the cells with
548 expressed genes over 500 are kept. The expression was normalized by NormalizeData. The
549 corresponding scRNAseq data was download from Linnarsson lab
550 (http://mousebrain.org/adolescent/downloads.html) (*2*). We only used cells from central nervous
551 system, and removed cells not in the brain regions profiled by Stereo-seq.
552
553 ***Mouse cortex data processing***
554 The STARmap data of mouse visual cortex was download from the STARmap resource website
555 (https://www.starmapresources.org/data). We chose the dataset "20180505_BY3_1kgenes"
556 which profiles 1020 genes. The expression matrix data "cell_barcode_count.csv" was imported
557 to Seurat object. The spatial location coordinate of each cell was extracted from "labels.npz"
558 according to the method provide by the original paper (https://github.com/weallen/STARmap).
559 The cell coordinates were integrated into Seurat object. The single cell RNA-seq was download
560 from the Allen Brain Atlas (https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-
561 alm-smart-seq). The "mouse_VISp_2018-06-14_exon-matrix.csv" file was used to generate the
562 expression profile. Low-quality cells were removed according to the meta data
563 "mouse_VISp_2018-06-14_samples-columns.csv".
564
565 ***Mouse striatum data processing***
566 The mouse striatum merFISH data was processed and normalized as described in the original
567 paper (*36*). Briefly, the expression of each cell was normalized by cell size and total RNA

568 counts. Then, the log-transformed data was applied to the expression matrix. In this manuscript,
569 we used the data from a representative slice (slice 10). We used the pre-process Seurat object
570 data in https://www.dropbox.com/s/ghkcovukgtctm76/NA_merFISH.RDS?dl=0. A down-sample
571 version of this data are provided by iSpatial package as a test dataset. After installing the R
572 package, this command "data(NA_merFISH)" could load the merFISH data into R environment.
573 The scRNA-seq data of mouse striatum was downloaded from GEO under accession
574 GSE118020. To speeding up the analysis, we down-sampled the full dataset to 10,000 cells using
575 R function "sample". Then this data was used to infer the spatial expression of all genes.
576

577 ***Mouse liver data processing***

578 The mouse liver merFISH data is download from Vizgen MERFISH Mouse Liver Map
579 (https://vizgen.com/data-release-program/), which targets 347 genes. This dataset contains
580 multiply slices. We only used slice 3 from the first replicate in this manuscript. The merFISH
581 data processing is the same as described above. The matched scRNA-seq data was download
582 from GEO under the accession GSE166504 (*42*). The original data set not only contains wild-
583 type healthy livers but also livers with nonalcoholic fatty liver disease (NAFLD). Three healthy
584 samples (Hepatocyte_Chow_Animal1_Capture1, Hepatocyte_Chow_Animal2_Capture1,
585 Hepatocyte_Chow_Animal3_Capture1) were used in this analysis.
586

587 ***CV score calculation***

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

599

600 ***Visualize the spatial expression***

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

604

605 ***Benchmark methods***

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

617

618 ***Performance evaluation***

619 The Slide-seq V2 data contains a total of 23,264 genes. We randomly sampled 3,000 genes as the
620 training dataset and used others as validation. After performing different methods to infer the
621 spatial expression of all genes, we conducted the comparisons between the inferred expression
622 values and the raw expression values. At cell level, we calculated the Pearson's correlation
623 coefficient of each gene across all cells, then compared the overall differences of correlation
624 coefficients among all genes of different methods. At spatial expression level, we grided the
625 spatial location into 50×50 bricks. Then, we calculated the mean expression in each brick and
626 compared the differences between raw expression and inferred one using 50×50 bricks as
627 features. In addition, we compared the detection accuracies of spatial variable genes among three
628 methods. The spatially variable genes were also called using validation dataset as the ground

truth. Receiver operating characteristic (ROC) curves and area under the curve (AUC) were performed by R package pROC (version 1.18.0) (*49*).

Ten-fold cross validation

For the datasets other than Slide-seq which we don't 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 ten groups. Each time we chose nine groups to infer the expression and used the left one to validate. After ten 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.

Data and Software Availability

iSpatial is available as an R package on github (https://github.com/YiZhang-lab/iSpatial).

Acknowledgements

We thank Dr. Aritra Bhattacherjee for some initial discussion and for critical reading of the manuscript. 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.

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 no competing interests.

658 **References**

659 1. A. Saunders, E. Z. Macosko, A. Wysoker, M. Goldman, F. M. Krienen, H. de Rivera, E. Bien, M.
660 Baum, L. Bortolin, S. Wang, A. Goeva, J. Nemesh, N. Kamitaki, S. Brumbaugh, D. Kulp, S. A.
661 McCarroll, Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain.
662 *Cell* **174**, 1015-1030 e1016 (2018).
663 2. A. Zeisel, H. Hochgerner, P. Lonnerberg, A. Johnsson, F. Memic, J. van der Zwan, M. Haring, E.
664 Braun, L. E. Borm, G. La Manno, S. Codeluppi, A. Furlan, K. Lee, N. Skene, K. D. Harris, J. Hjerling-
665 Leffler, E. Arenas, P. Ernfors, U. Marklund, S. Linnarsson, Molecular Architecture of the Mouse
666 Nervous System. *Cell* **174**, 999-1014 e1022 (2018).
667 3. X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang,
668 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.
669 Huang, S. H. Orkin, G. C. Yuan, M. Chen, G. Guo, Mapping the Mouse Cell Atlas by Microwell-
670 Seq. *Cell* **173**, 1307 (2018).
671 4. O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, S. A. Teichmann, The Human Cell Atlas:
672 from vision to reality. *Nature* **550**, 451-453 (2017).
673 5. V. Marx, Method of the Year: spatially resolved transcriptomics. *Nat Methods* **18**, 9-14 (2021).
674 6. P. L. Stahl, F. Salmen, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M.
675 Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, A. Borg, F. Ponten, P. I.
676 Costea, P. Sahlen, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisen, Visualization and analysis of
677 gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).
678 7. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, RNA imaging. Spatially resolved,
679 highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
680 8. S. Shah, E. Lubeck, W. Zhou, L. Cai, seqFISH Accurately Detects Transcripts in Single Cells and
681 Reveals Robust Spatial Organization in the Hippocampus. *Neuron* **94**, 752-758 e751 (2017).
682 9. X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C.
683 Ramakrishnan, J. Liu, G. P. Nolan, F. A. Bava, K. Deisseroth, Three-dimensional intact-tissue
684 sequencing of single-cell transcriptional states. *Science* **361**, (2018).
685 10. C. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G. C. Yuan,
686 L. Cai, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-
687 239 (2019).
688 11. Y. Liu, M. Yang, Y. Deng, G. Su, A. Enninful, C. C. Guo, T. Tebaldi, D. Zhang, D. Kim, Z. Bai, E.
689 Norris, A. Pan, J. Li, Y. Xiao, S. Halene, R. Fan, High-Spatial-Resolution Multi-Omics Sequencing
690 via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681 e1618 (2020).
691 12. S. Codeluppi, L. E. Borm, A. Zeisel, G. La Manno, J. A. van Lunteren, C. I. Svensson, S. Linnarsson,
692 Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* **15**, 932-
693 935 (2018).
694 13. L. E. Borm, A. M. Albiach, C. C. A. Mannens, J. Janusauskas, C. Özgün, D. Fernández-García, R.
695 Hodge, E. S. Lein, S. Codeluppi, S. Linnarsson, Scalable in situ single-cell profiling by
696 electrophoretic capture of mRNA. *bioRxiv*, 2022.2001.2012.476082 (2022).
697 14. M. Asp, S. Giacomello, L. Larsson, C. Wu, D. Furth, X. Qian, E. Wardell, J. Custodio, J. Reimegard,
698 F. Salmen, C. Osterholm, P. L. Stahl, E. Sundstrom, E. Akesson, O. Bergmann, M. Bienko, A.
699 Mansson-Broberg, M. Nilsson, C. Sylven, J. Lundeberg, A Spatiotemporal Organ-Wide Gene
700 Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660 e1619 (2019).
701 15. R. Dries, J. Chen, N. Del Rossi, M. M. Khan, A. Sistig, G. C. Yuan, Advances in spatial
702 transcriptomic data analysis. *Genome Res* **31**, 1706-1718 (2021).
703 16. B. Hie, J. Peters, S. K. Nyquist, A. K. Shalek, B. Berger, B. D. Bryson, Computational Methods for
704 Single-Cell RNA Sequencing. *Annual Review of Biomedical Data Science* **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 e727 (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*, (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*, (2022).
29. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (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 e1817 (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, 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 e1721 (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 e688 (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 e137 (2019).

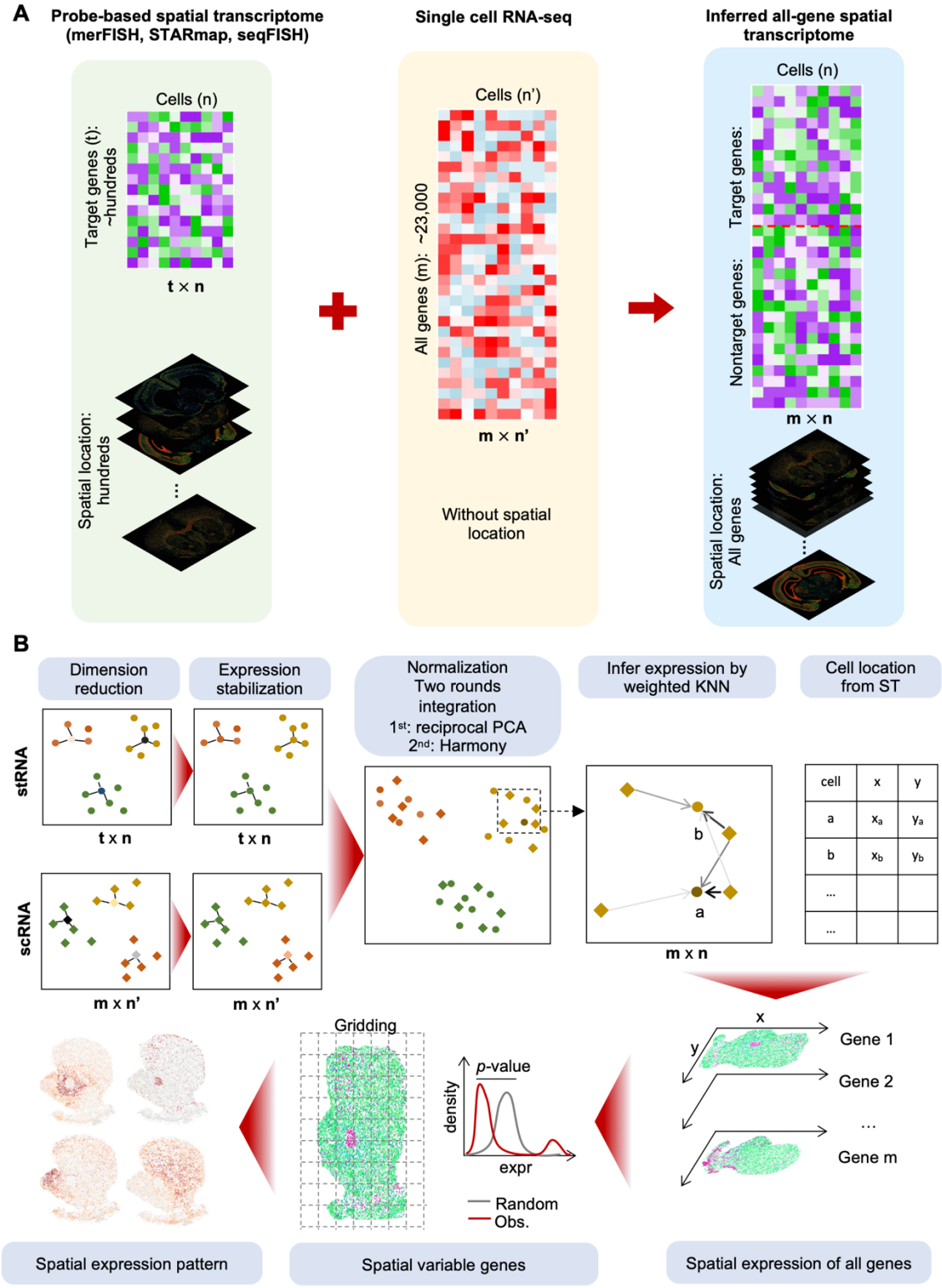
42. Q. Su, S. Y. Kim, F. Adewale, Y. Zhou, C. Aldler, M. Ni, Y. Wei, M. E. Burczynski, G. S. Atwal, M. W. Sleeman, A. J. Murphy, Y. Xin, X. Cheng, Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).

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).

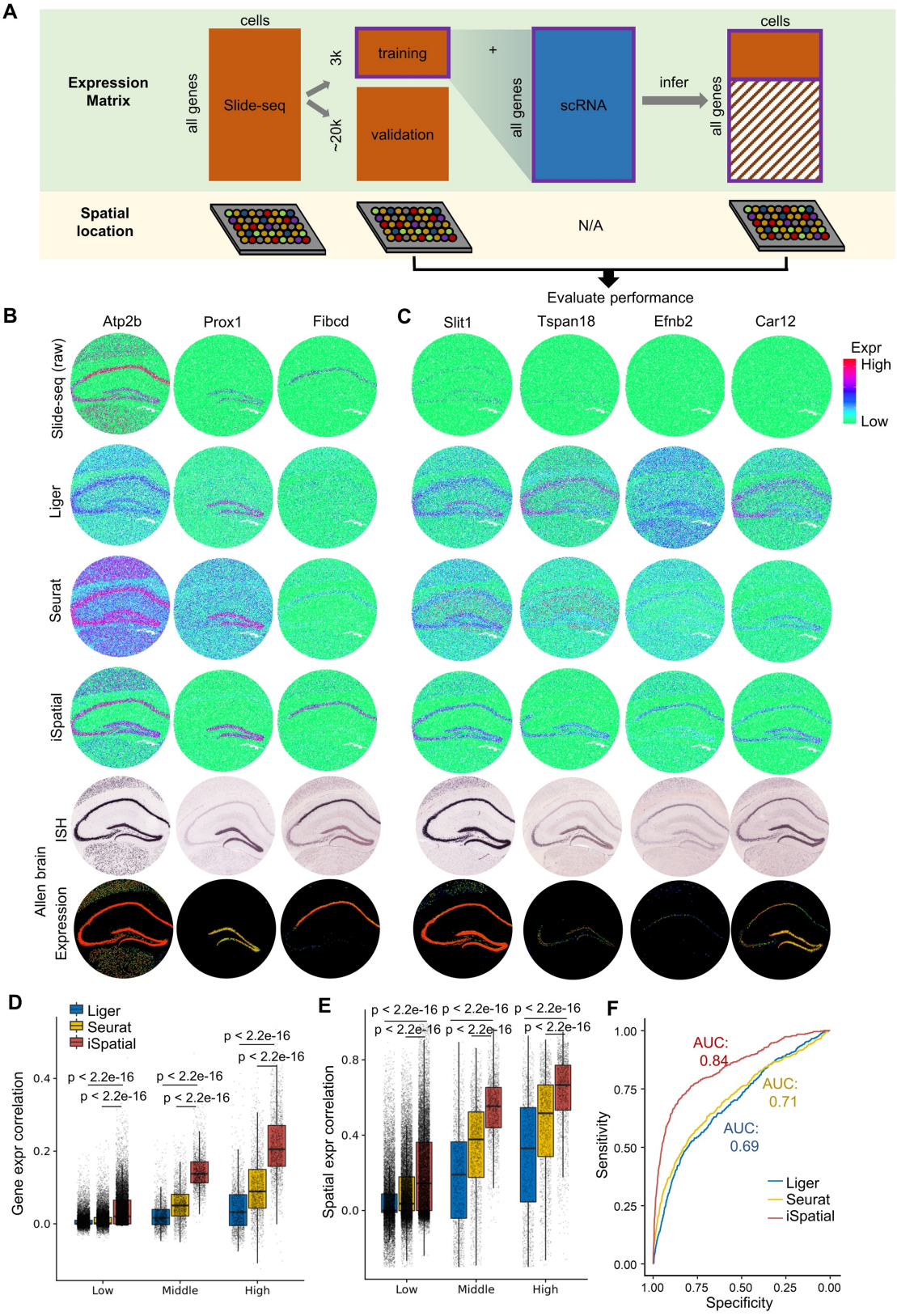
44. F. Hildebrandt, A. Andersson, S. Saarenpaa, L. Larsson, N. Van Hul, S. Kanatani, J. Masek, E. Ellis, A. Barragan, A. Mollbrink, E. R. Andersson, J. Lundeberg, J. Ankarklev, Spatial Transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nat Commun* **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. B. I. C. C. Network, 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, P. Human Cell Atlas Meeting, The Human Cell Atlas. *Elife* **6**, (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).

827 **Figure**



828

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 is consisted of: 1) Dimension reduction and expression stabilization; 2) Expression normalization; 3) Inferring transcriptional expression; and 4) Downstream analysis such as spatial variable gene detection.


A
Expression Matrix
cells
all genes
Slide-seq
3k
~20k
training
validation
+
all genes
scRNA
infer
cells
all genes
Spatial location
N/A
Evaluate performance
B
Atp2b
Prox1
Fibcd
C
Slit1
Tspan18
Efnb2
Car12
Expr
High
Low
Slide-seq (raw)
Liger
Seurat
iSpatial
Allen brain
ISH
Expression
D
Liger
Seurat
iSpatial
Gene expr correlation
p < 2.2e-16
Low
Middle
High
E
Spatial expr correlation
p < 2.2e-16
Low
Middle
High
F
Sensitivity
Specificity
AUC: 0.84
AUC: 0.71
AUC: 0.69
Liger
Seurat
iSpatial

836

837 **Fig. 2: Benchmarking the performance of iSpatial in inferring genome-wide spatial**
838 **transcriptome**. **A**, A graphical illustration of the evaluation procedure. **B**, Representative
839 examples showing the performance of Liger, Seurat and iSpatial in inferring spatial
840 transcriptome. **C**, Representative examples of inferred expression of genes barely detectable in
841 raw Slide-seq data. **D**, The gene expression correlation between Slide-seq raw data and Liger,
842 Seurat, or iSpatial inferred data. The validation genes are divided into 3 groups based on their
843 expression levels. Two-side Wilcoxon Rank Sum test was used. **E**, The spatial expression
844 correlation between inferred and raw Slide-seq data. Two-side Wilcoxon Rank Sum test was
845 used. **F**, The ROC curves comparing the prediction power of spatial variable genes by Liger,
846 Seurat, and iSpatial. AUCs (Area under the curves) are indicated.
847

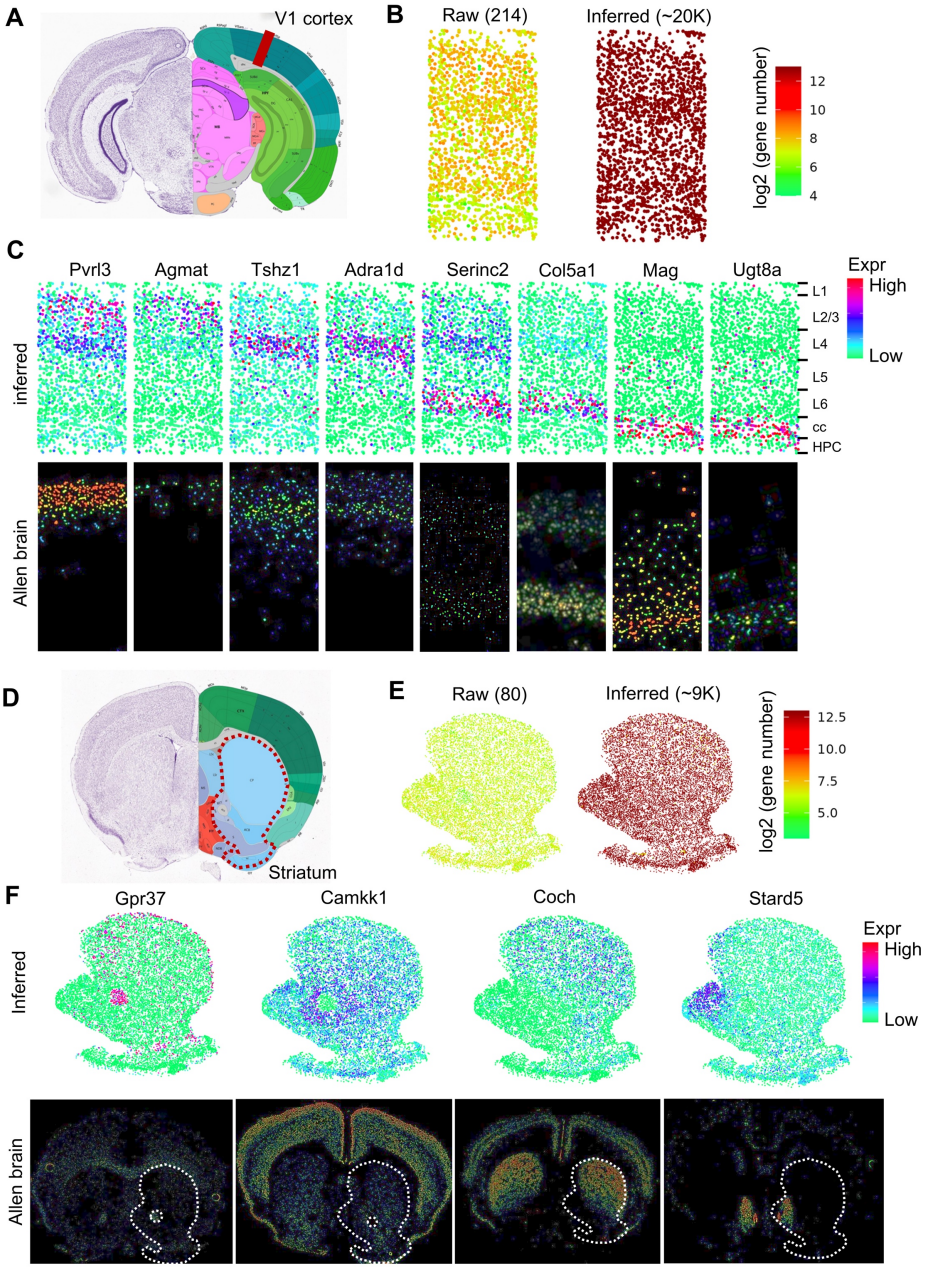
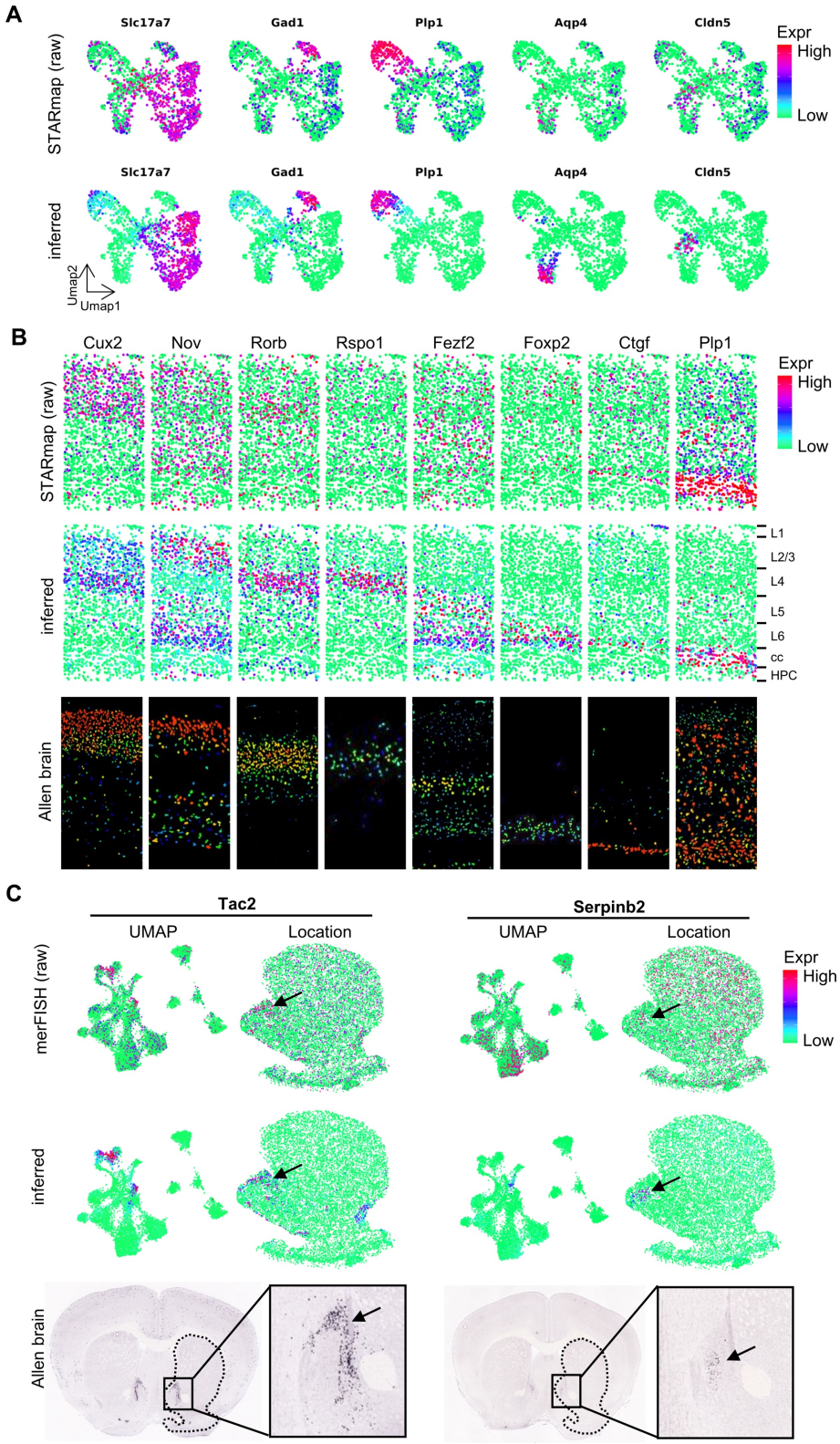


Fig. 3: iSpatial accurately infers the genome-wide spatial transcriptomes in mouse cortex and striatum. **A**, Schematic of the anatomic region of mouse V1 cortex. **B**, The numbers of detectable genes in each cell in raw STARmap (left) and after inferring by iSpatial (right). The mean number of detected 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.


A
Slc17a7
Gad1
Plp1
Aqp4
Cldn5
Expr
High
Low
STARmap (raw)
inferred
Umap2
Umap1
B
Cux2
Nov
Rorb
Rspo1
Fezf2
Foxp2
Ctgf
Plp1
STARmap (raw)
inferred
Allen brain
L1
L2/3
L4
L5
L6
cc
HPC
C
Tac2
Serpinb2
UMAP
Location
merFISH (raw)
inferred
Allen brain

859

860 **Fig. 4: iSpatial can reduce false-positive and false-negative signals in the original spatial**
861 **transcriptomic data**. **A**, UMAPs showing the expression levels of representative cell type
862 markers in raw STARmap (top panels) and iSpatial inferred (bottom panels) data. Excitatory
863 neuron (*Slc17a7*), Inhibitory neuron (*Gad1*), Oligodendrocyte (*Plp1*), Astrocyte (*Aqp4*), and
864 Endothelial cell (*Cldn5*). **B**, The spatial expression of cortex layer markers in the raw STARmap
865 (top panels) and inferred by iSpatial (middle panels) compared with the ISH data from the ABA
866 (bottom panels). Layer information: "L1 to L6", the six cortical layers; "cc", corpus callosum;
867 "HPC", hippocampus. **C**, The UMAP and spatial expression of *Tac2* and *Serpinb2* in the raw
868 MERFISH (top panels) and inferred by iSpatial (middle panels) compared with the ISH data
869 from the ABA (bottom panels).
870

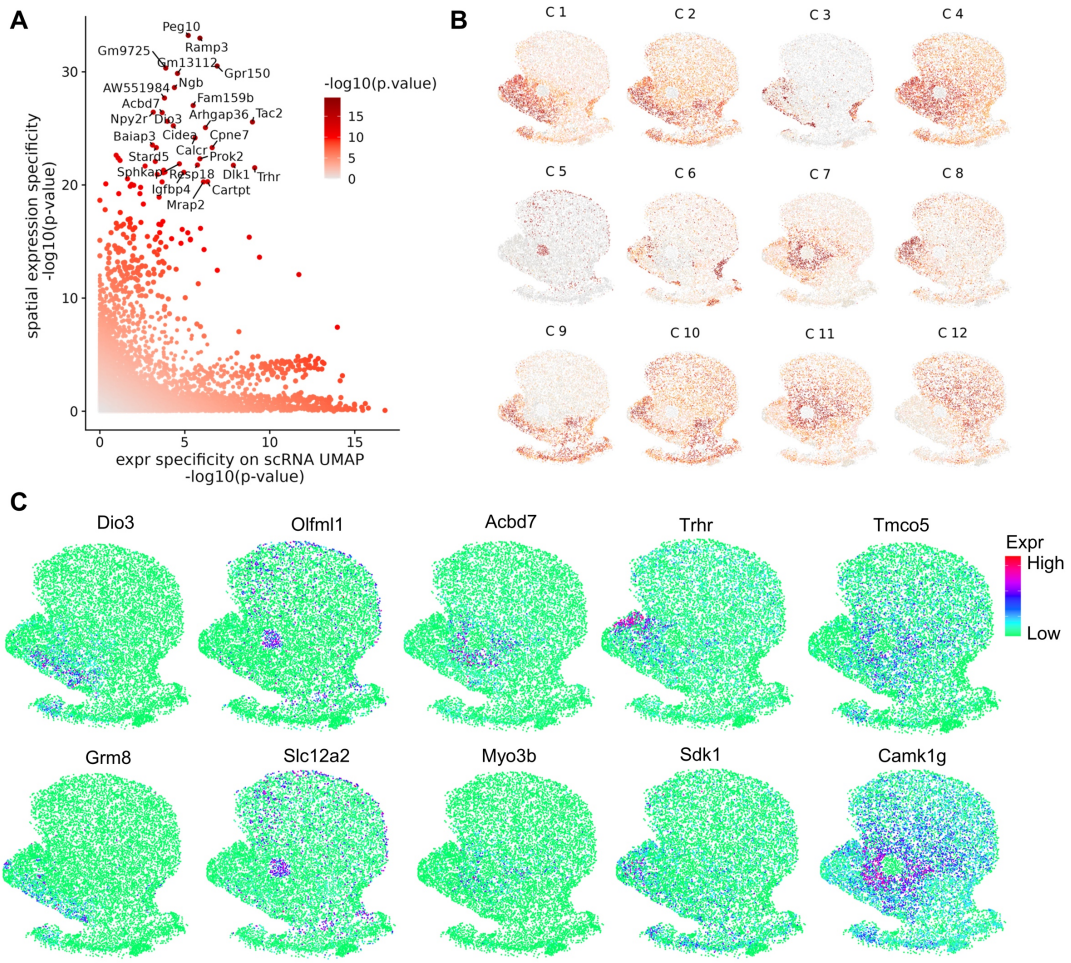


Fig. 5: iSpatial enables whole-transcriptome level spatial analysis. **A**, Scatterplot showing genes with expression specificity in spatial location and UMAP projection of the corresponding scRNA-seq data. **B**, The 12 clusters of the spatial variable genes in mouse striatum. The plot is colored coded 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*.

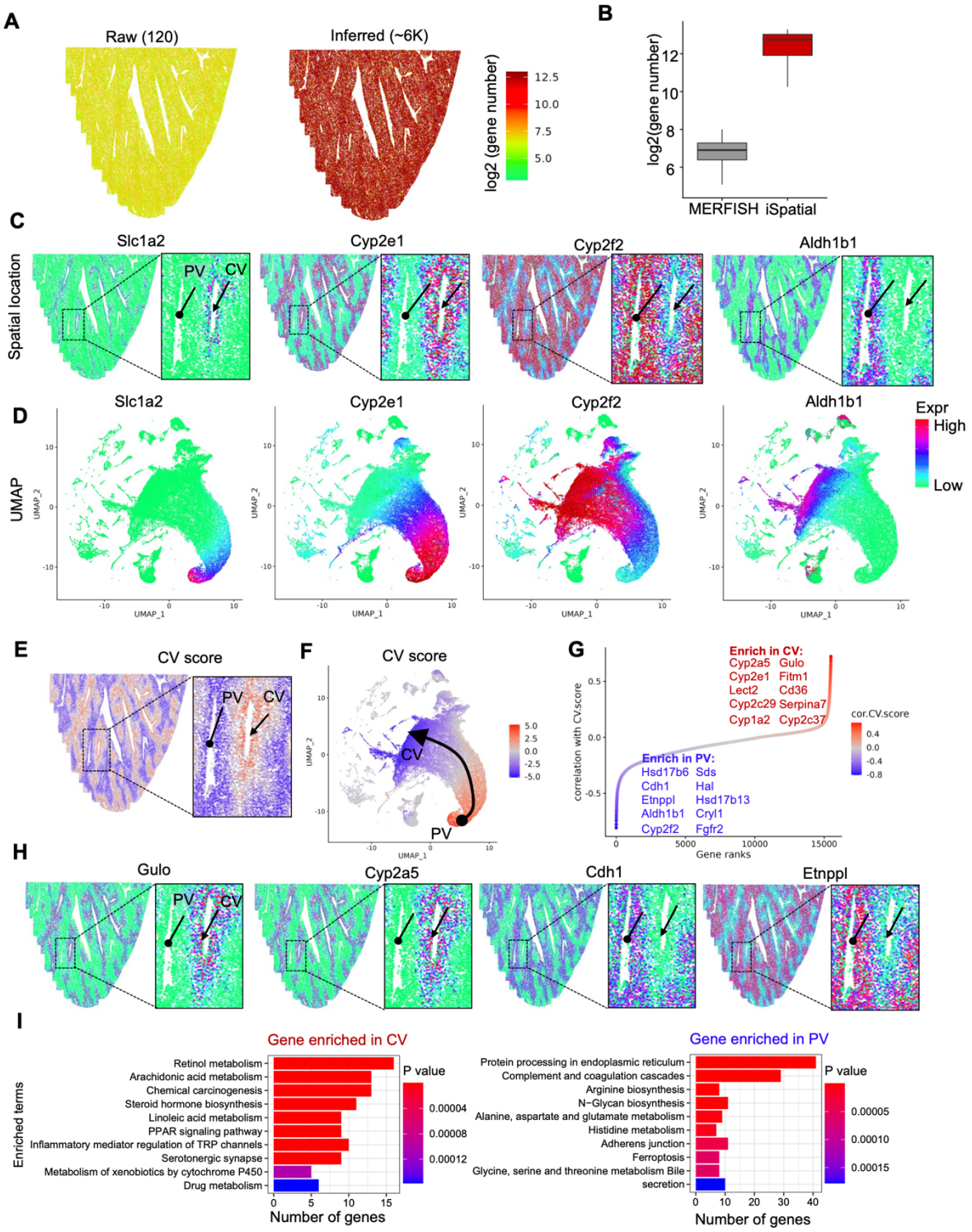


Fig. 6: iSpatial infers the spatial expression patterns in liver. **A**, The numbers of detectable genes in each cell in raw MERFISH (left) and after inferring by iSpatial (right). **B**, Boxplot showing the detectable gene numbers without or with interfering by iSpatial. **C**, Representative

examples of genes exhibiting spatial patterns enriched in CV (*Slc1a2*, *Cyp2e1*) or PV (*Cyp2f2*, *Aldh1b1*). **D**, UMAPs showing the expression levels of the genes showed 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 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 KEGG pathways of genes selectively expressed near CV or PV.

Supplementary Figures

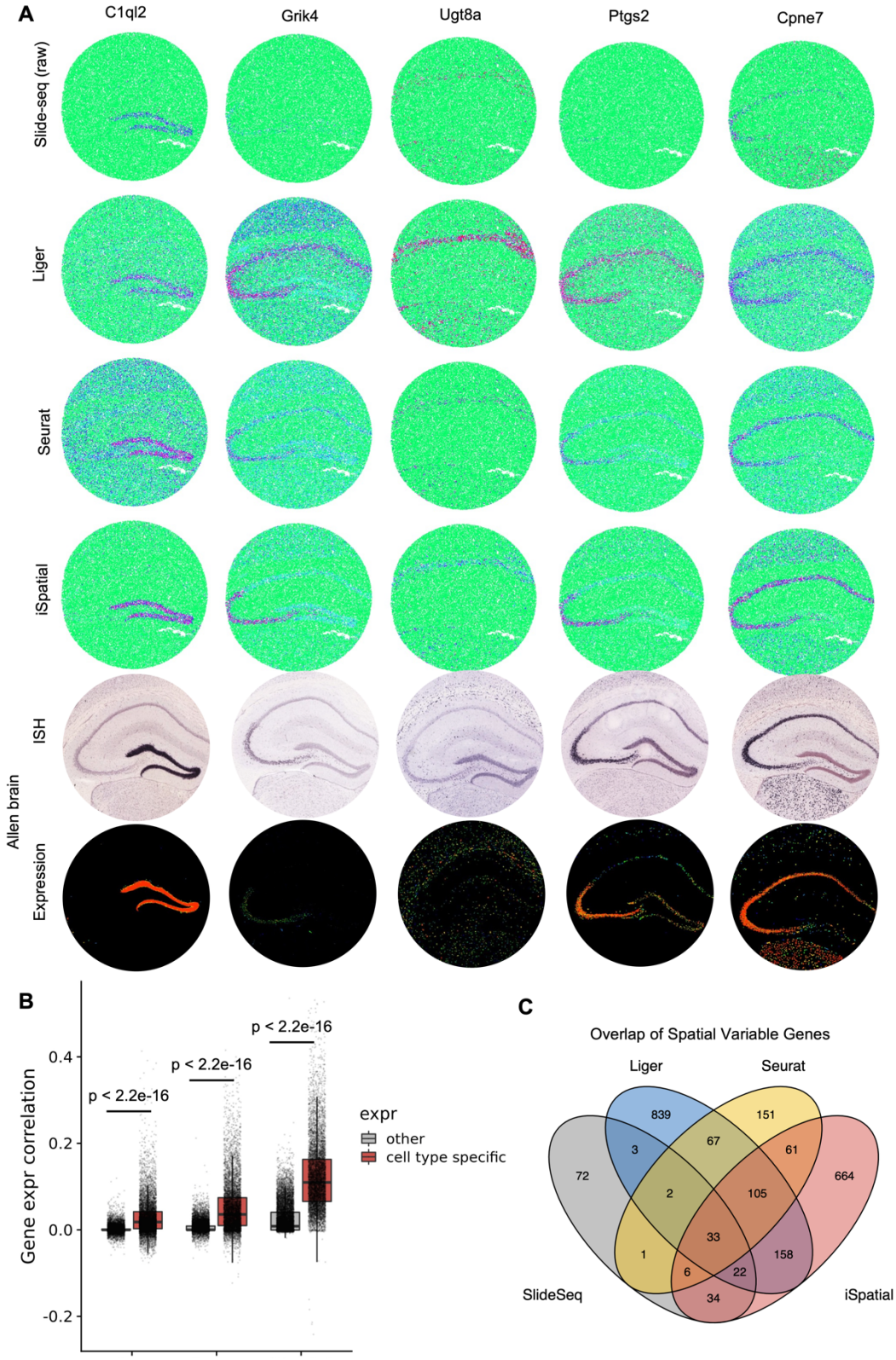
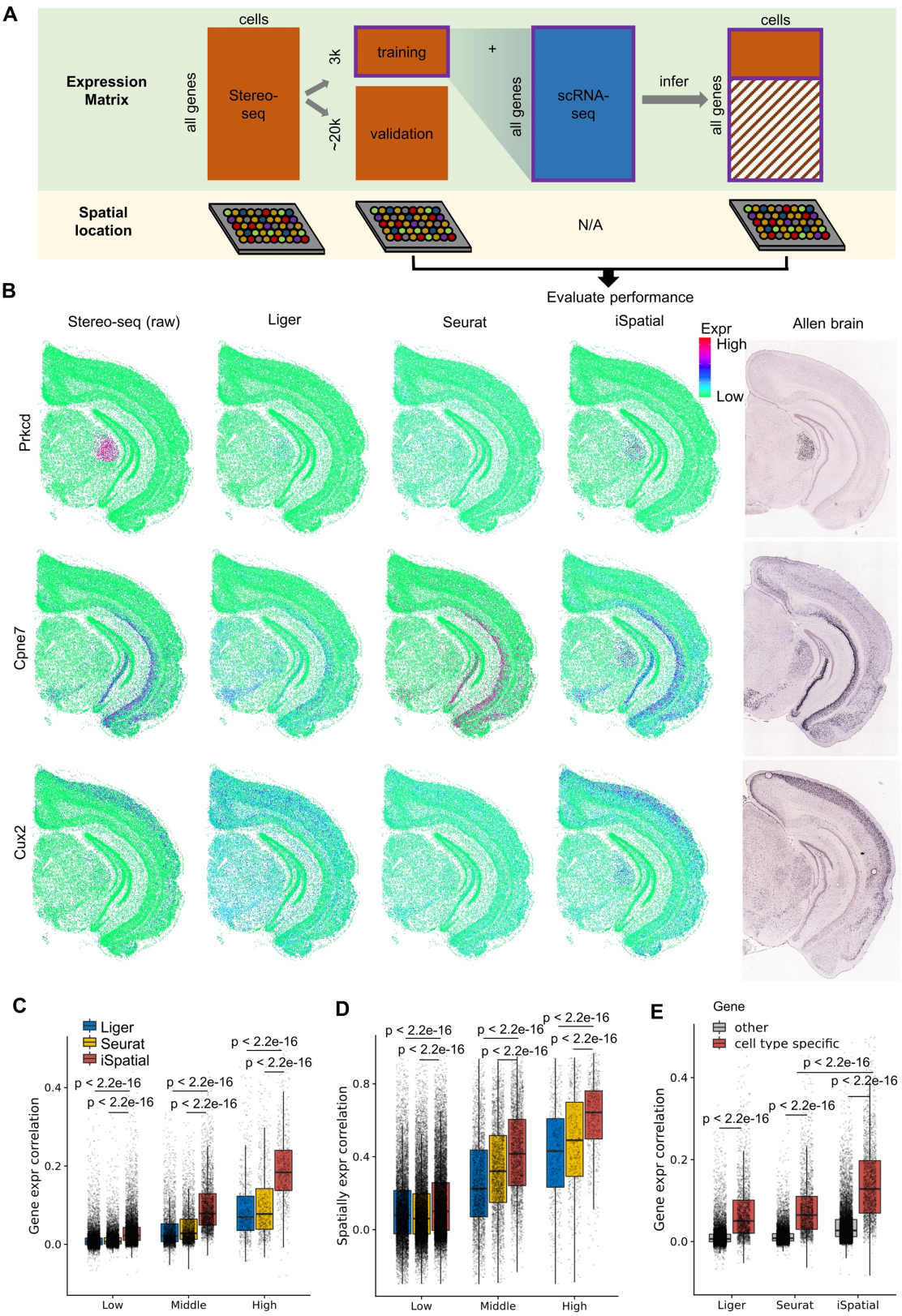


fig. S1: Benchmarking the performance of iSpatial in inferring genome-wide spatial transriptome. **A**, Representative examples showing the performance of Liger, Seurat and iSpatial in inferring spatial transcriptome. **B**, The gene expression correlation between raw data and Liger, Seurat, or iSpatial inferred data of cell type specific genes or other genes. **C**, The overlaps of the spatial variable genes detected by different methods.


A
Expression Matrix
cells
all genes
Stereo-seq
3k
~20k
training
validation
+
all genes
scRNA-seq
infer
cells
all genes
Spatial location
N/A
Evaluate performance
B
Stereo-seq (raw)
Liger
Seurat
iSpatial
Expr
High
Low
Allen brain
Prkcd
Cpne7
Cux2
C
Liger
Seurat
iSpatial
Gene expr correlation
p < 2.2e-16
Low
Middle
High
D
Spatially expr correlation
E
Gene
other
cell type specific
Gene expr correlation
Liger
Seurat
iSpatial

899

900 **fig. S2: Benchmarking the performance on Stereo-seq data. A**, A graphical illustration of the
901 evaluation procedure. **B**, Representative examples showing the raw or inferred spatial expression
902 by Liger, Seurat and iSpatial. Allen brain ISH data are also shown. **C**, The gene expression
903 correlation between Stereo-seq raw data and Liger, Seurat, or iSpatial inferred data. The
904 validation genes are divided into 3 groups based on their expression levels. **D**, The spatial
905 expression correlation between inferred and raw Stereo-seq data. **E**, The gene expression
906 correlation between Stereo-seq raw data and Liger, Seurat, or iSpatial inferred data of cell type
907 specific genes or other genes. Two-side Wilcoxon Rank Sum test was used.
908

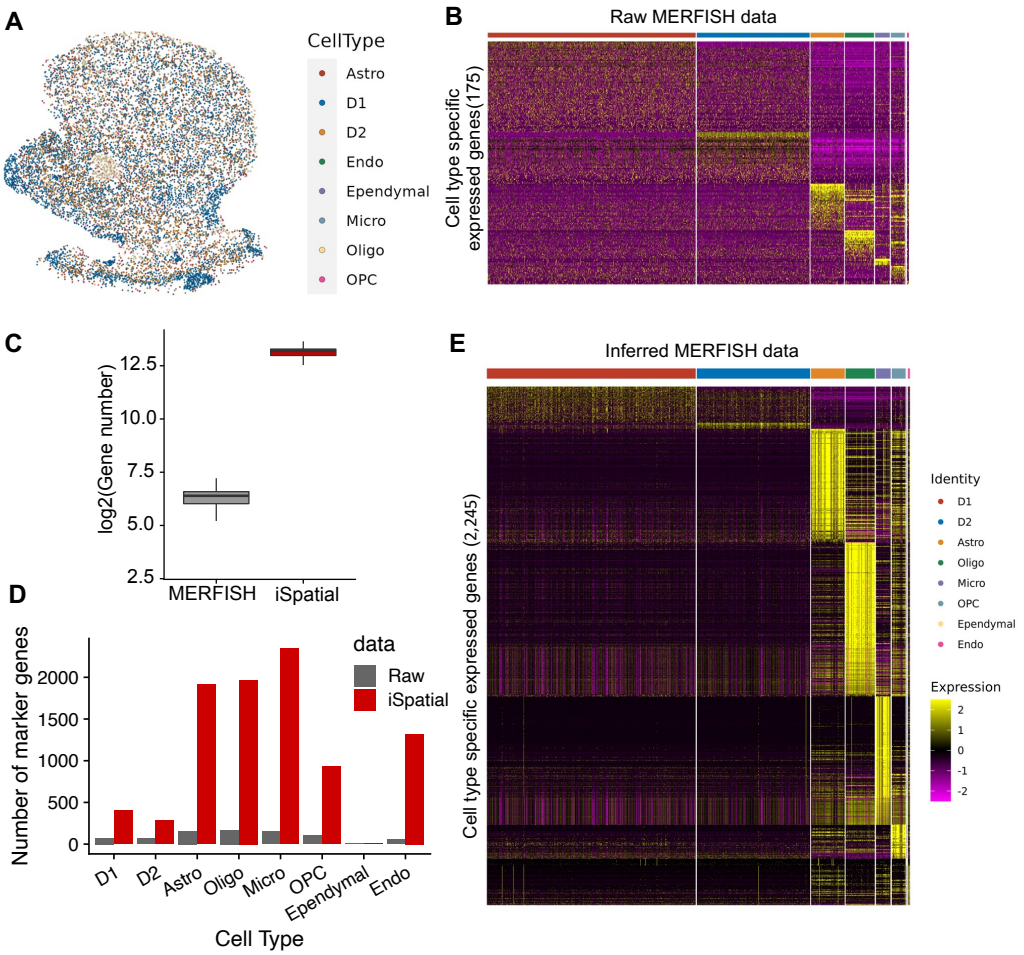


fig. S3: iSpatial is applicable to MERFISH dataset. **A**, Spatial distribution of the cell types detected by MERFISH. **B**, The heatmap of the 175 cell type markers detected in the raw MERFISH data. **C**, The numbers of detected genes in raw MERFISH data and after interfering by iSpatial. **D**, The number of cell type specific expressed genes in each cell type detected in raw and iSpatial inferred data. **E**, The heatmap of all maker genes detected after inferring by iSpatial.

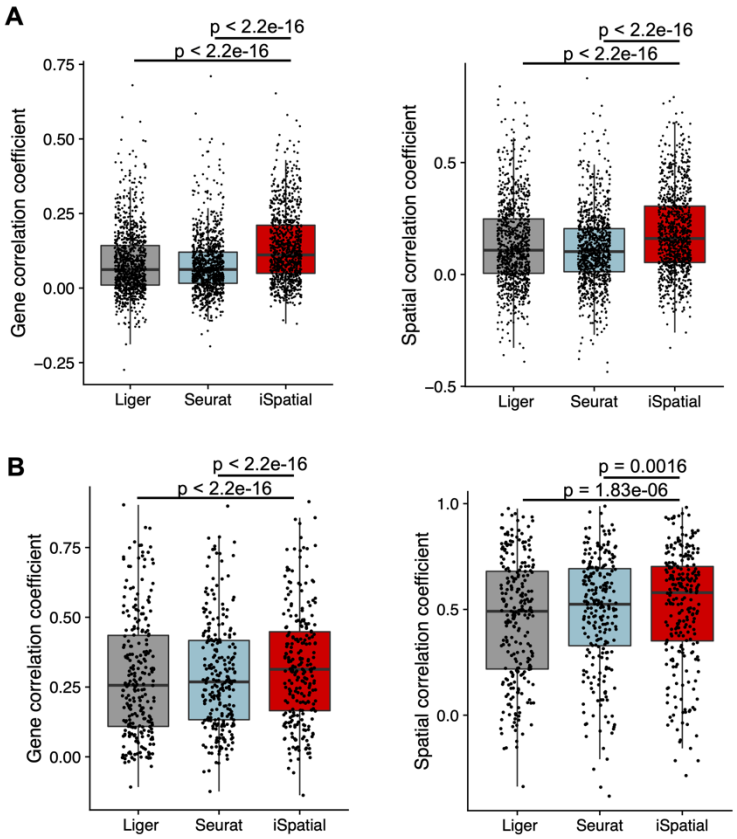


fig. S4: Ten-fold cross validation on the performance of the three methods in mouse cortex (A) and striatum (B). Left panels: gene expression correlation between raw expression data and inferred data. Right panels: spatial expression correlation between raw expression data and inferred data.

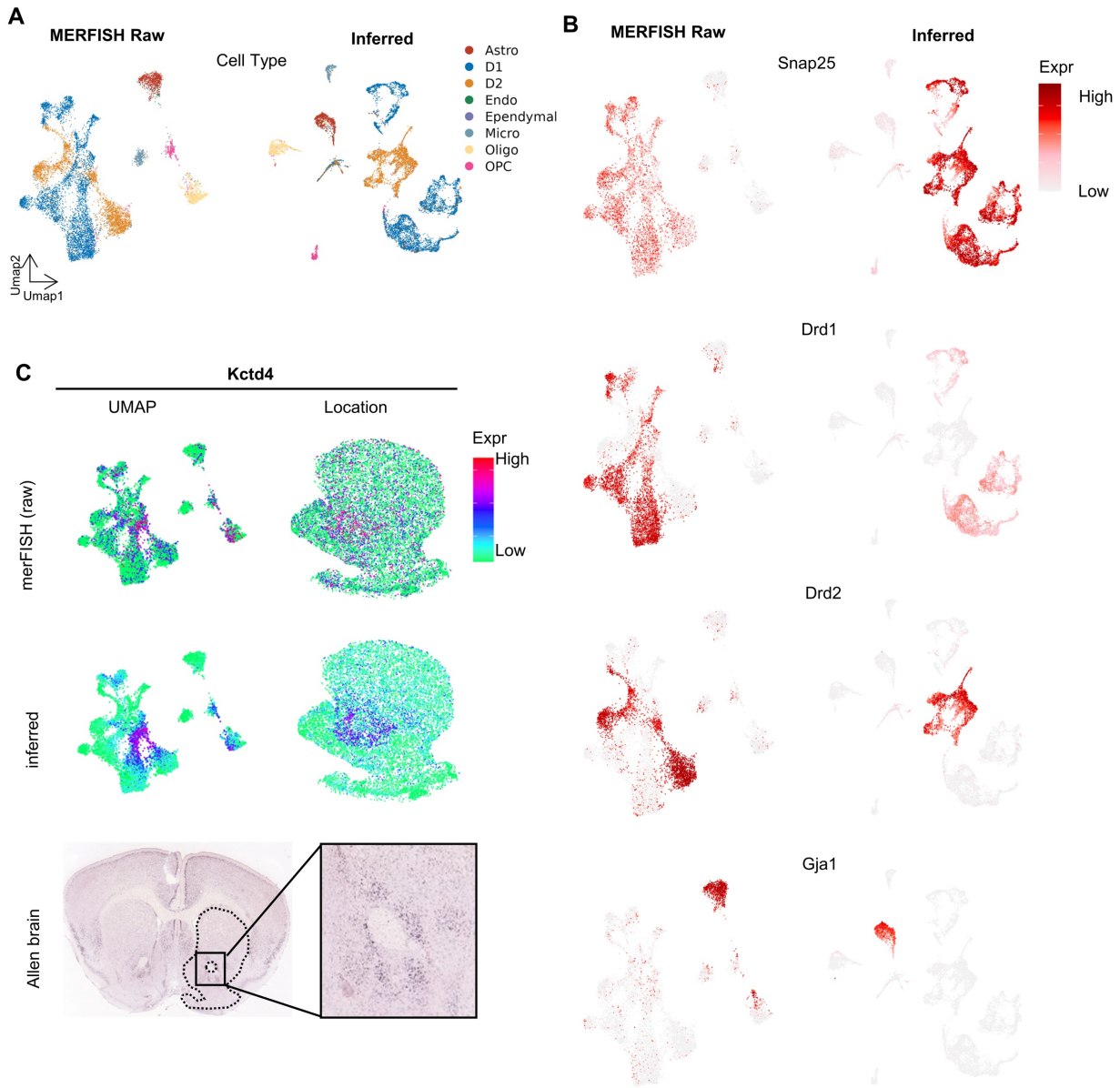
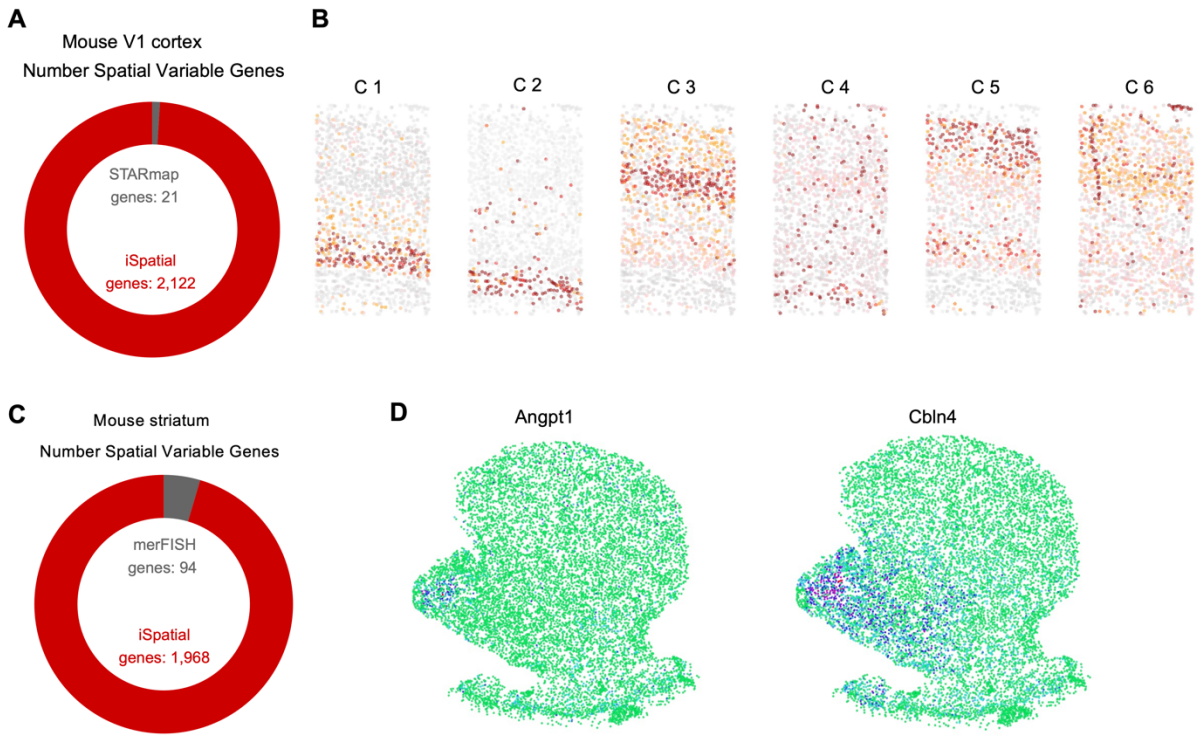


fig. S5: iSpatial can reduce false-positive and false-negative signals in the original MERFISH data. A, Cell type clusters on the UMAPs derived from raw or inferred data. **B**, UMAPs showing the expression of representative cell type marker genes. The UMAPs are generated from raw or iSpatial inferred expression profiles, respectively. Marker genes for neuron (*Snap25*), D1 medium spiny neuron (*Drd1*), D2 medium spiny neuron (*Drd2*), and astrocytes (*Gja1*) are presented. **C**, The UMAP and spatial expression of *Kctd4* in the raw

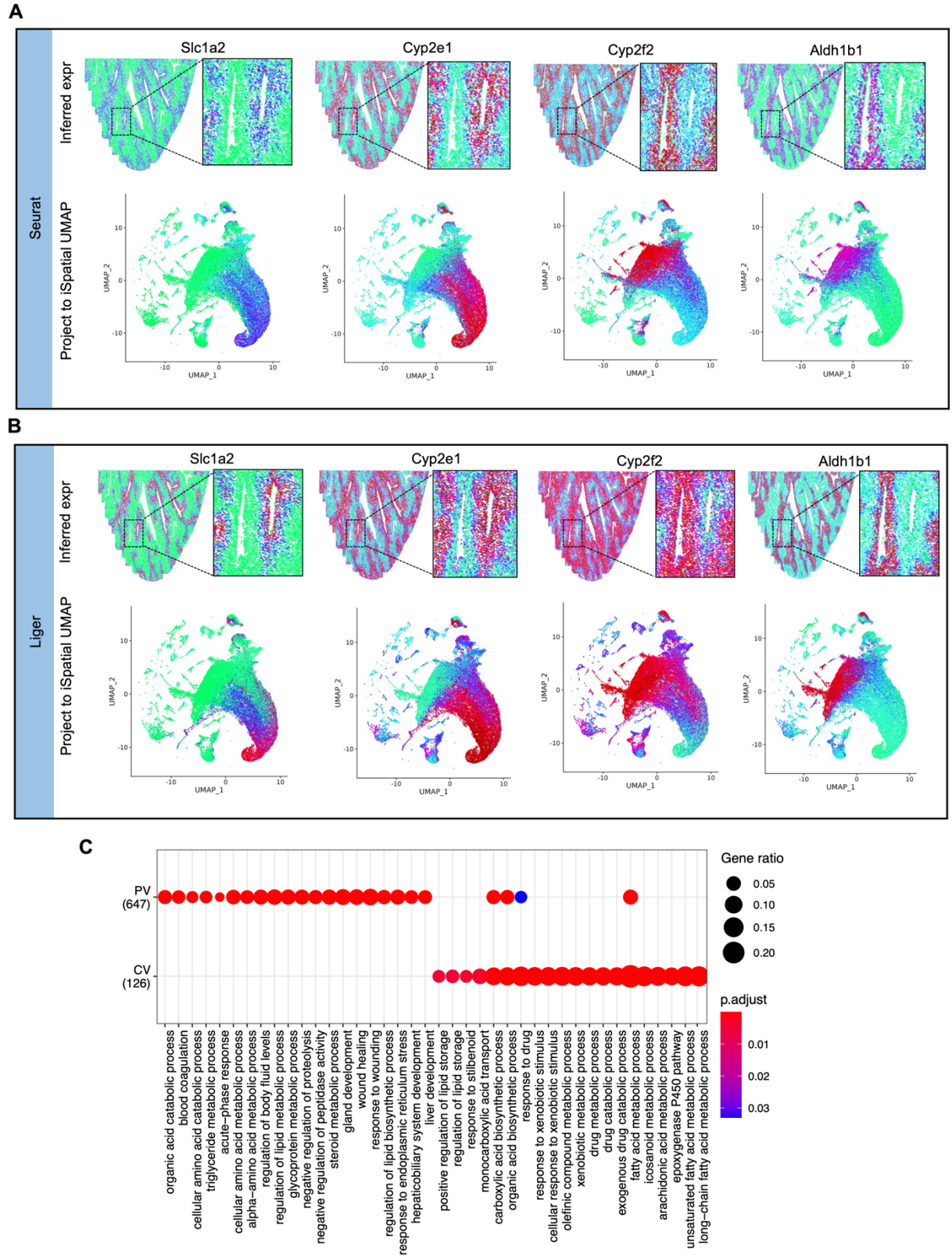
929 MERFISH (top panels) and inferred by iSpatial (middle panels) compared with the ISH data
930 from the ABA (bottom panel).
931



932

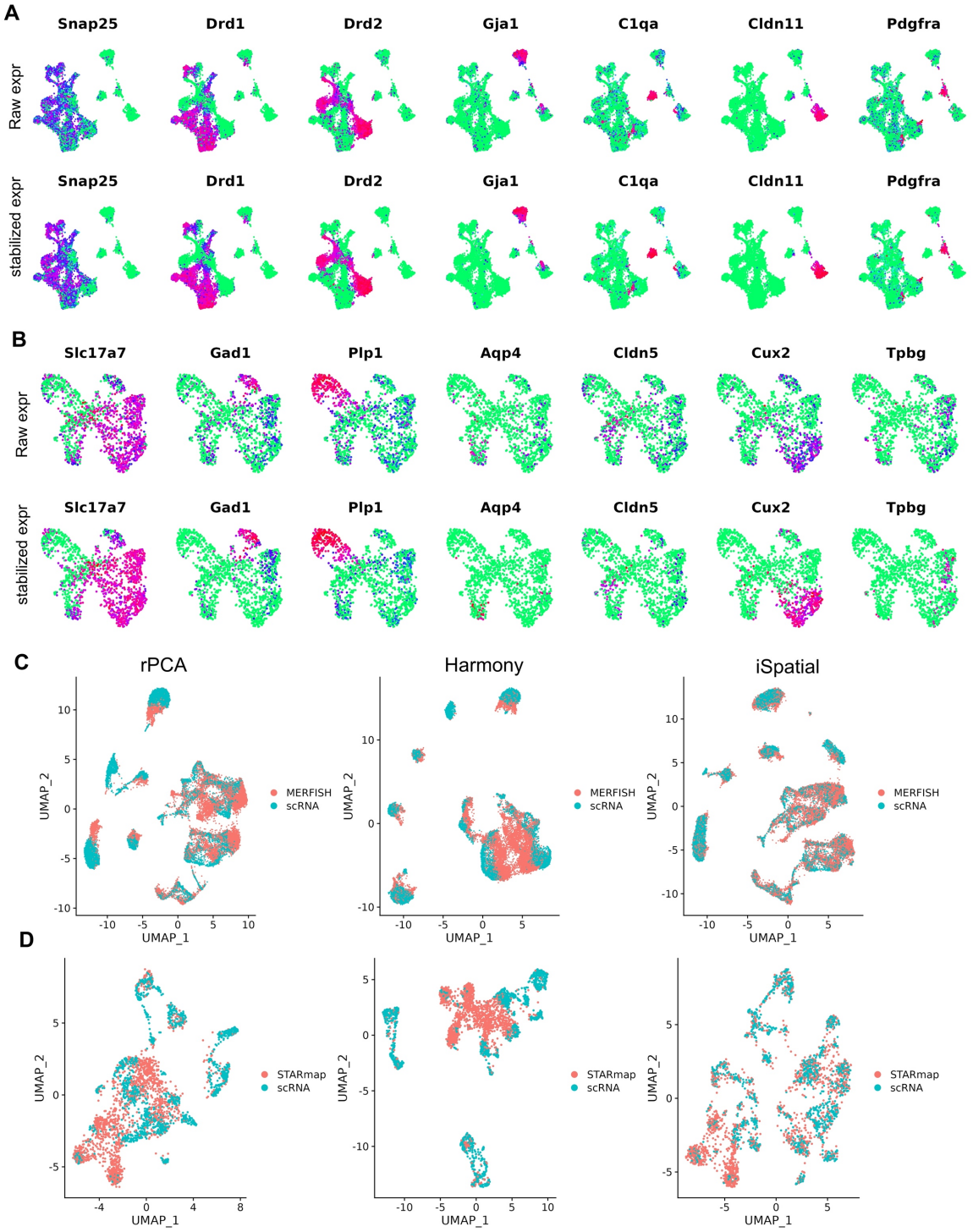
933 **fig. S6: iSpatial detects spatial variable genes in mouse cortex (A, B) and striatum (C, D)**.
934 **A**, The number of spatial variable genes in mouse V1 cortex detected by STARmap and after
935 iSpatial interferring. **B**, The layer distribution of the 6 clusters spatial variable genes. The plot is
936 colored by the average expression level of the genes in each cluster. **C**, The number of spatial
937 variable genes in mouse striatum detected by MERFISH and after iSpatial interfering. **D**,
938 Inferred spatial expression signals of representative genes in the C8 cluster.

939


A
Seurat
Inferred expr
Project to iSpatial UMAP
Slc1a2
Cyp2e1
Cyp2f2
Aldh1b1
UMAP_1
UMAP_2
B
Liger
Inferred expr
Project to iSpatial UMAP
Slc1a2
Cyp2e1
Cyp2f2
Aldh1b1
UMAP_1
UMAP_2
C
PV (647)
CV (126)
organic acid catabolic process
blood coagulation
cellular amino acid catabolic process
triglyceride metabolic process
acute−phase response
cellular amino acid metabolic process
alpha−amino acid metabolic process
regulation of body fluid levels
regulation of lipid metabolic process
glycoprotein metabolic process
negative regulation of proteolysis
negative regulation of peptidase activity
steroid metabolic process
gland development
wound healing
response to wounding
regulation of lipid biosynthetic process
response to endoplasmic reticulum stress
hepaticobiliary system development
liver development
positive regulation of lipid storage
regulation of lipid storage
response to stilbenoid
monocarboxylic acid transport
carboxylic acid biosynthetic process
organic acid biosynthetic process
response to drug
response to xenobiotic stimulus
cellular response to xenobiotic stimulus
olefinic compound metabolic process
xenobiotic metabolic process
drug metabolic process
drug catabolic process
exogenous drug catabolic process
fatty acid metabolic process
icosanoid metabolic process
arachidonic acid metabolic process
epoxygenase P450 pathway
unsaturated fatty acid metabolic process
long−chain fatty acid metabolic process
Gene ratio
0.05
0.10
0.15
0.20
p.adjust
0.01
0.02
0.03

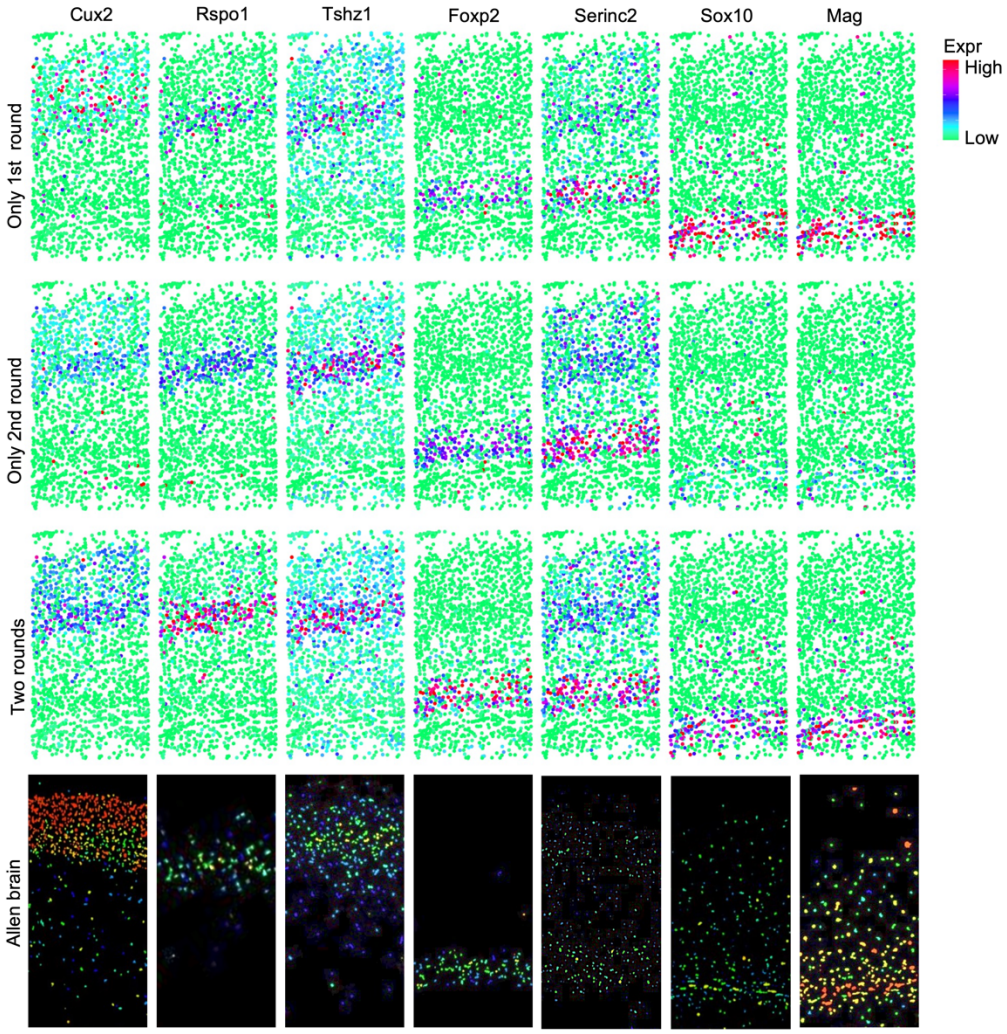
940

941 **fig. S7: iSpatial infers the spatial expression patterns in liver**. **A, B**, Examples of spatial
942 expressed genes enriched in CV (*Slc1a2*, *Cyp2e1*) or PV (*Cyp2f2*, *Aldh1b1*), inferred by Seurat
943 (A) and Liger (B). The upper panels show the spatial expression patterns. The bottom panels
944 show the inferred genes' expression by the indicated method projects on the iSpatial UMAP. **C**,
945 The top 20 enriched Gene Ontology terms of genes selectively expressed near CV or PV.
946


A
Snap25
Drd1
Drd2
Gja1
C1qa
Cldn11
Pdgfra
Raw expr
stabilized expr
B
Slc17a7
Gad1
Plp1
Aqp4
Cldn5
Cux2
Tpbg
C
rPCA
Harmony
iSpatial
MERFISH
scRNA
UMAP_1
UMAP_2
D
STARmap
scRNA

947

fig. S8: Stabilization and two-rounds integration steps contribute to the improved performance of iSpatial. A, B, UMAPs showing the expression signals of representative markers before (top panels) and after (bottom panels) expression stabilization in mouse striatum (**A**) or cortex (**B**) data. **C, D**, The co-embed UMAPs showing the integration of ST and scRNA-seq data in mouse striatum (**C**) or cortex (**D**) data.



954

955 **fig S9: Comparing the two-rounds integration with one round.** The spatial expression of the
956 representative genes of the mouse cortex raw STARmap data inferred by one round (top two
957 rows) or two rounds (third row) of integration by iSpatial compared with the ISH data from the
958 ABA (bottom row). Reciprocal PCA is used for the first-round integration, and Harmony is used
959 for the second-round integration.

960
961
962
963
964
965
966
967
968

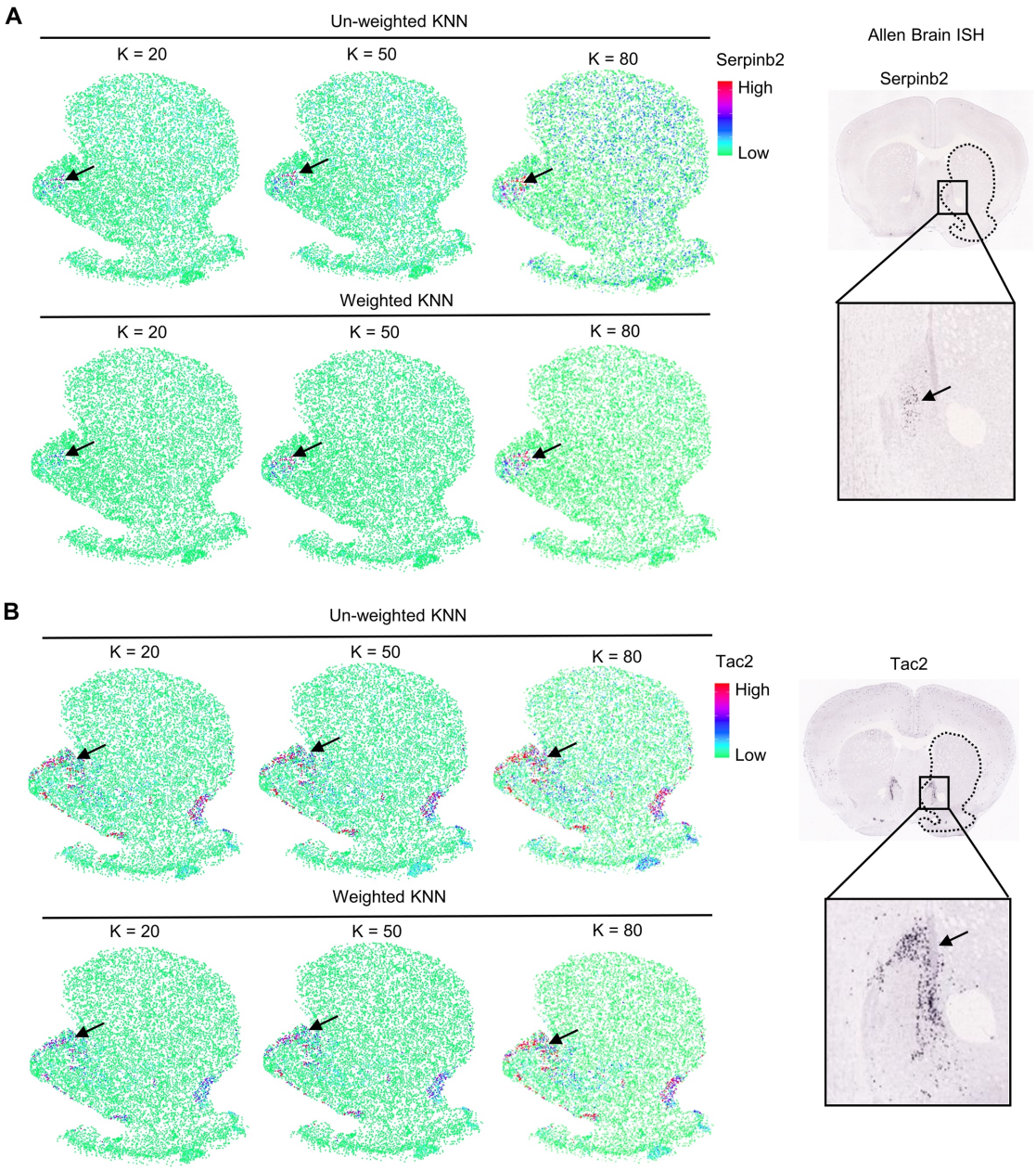


fig S10: Comparing the performance of weighted or un-weighted KNN. A, B, The inferred spatial expression of *Serpinb2* (**A**) and *Tac2* (**B**) in rare cell population using un-weighted KNN or weighted KNN under different K value. Allen brain ISH data were showed on right. Arrows indicate the location of the rare cell population.

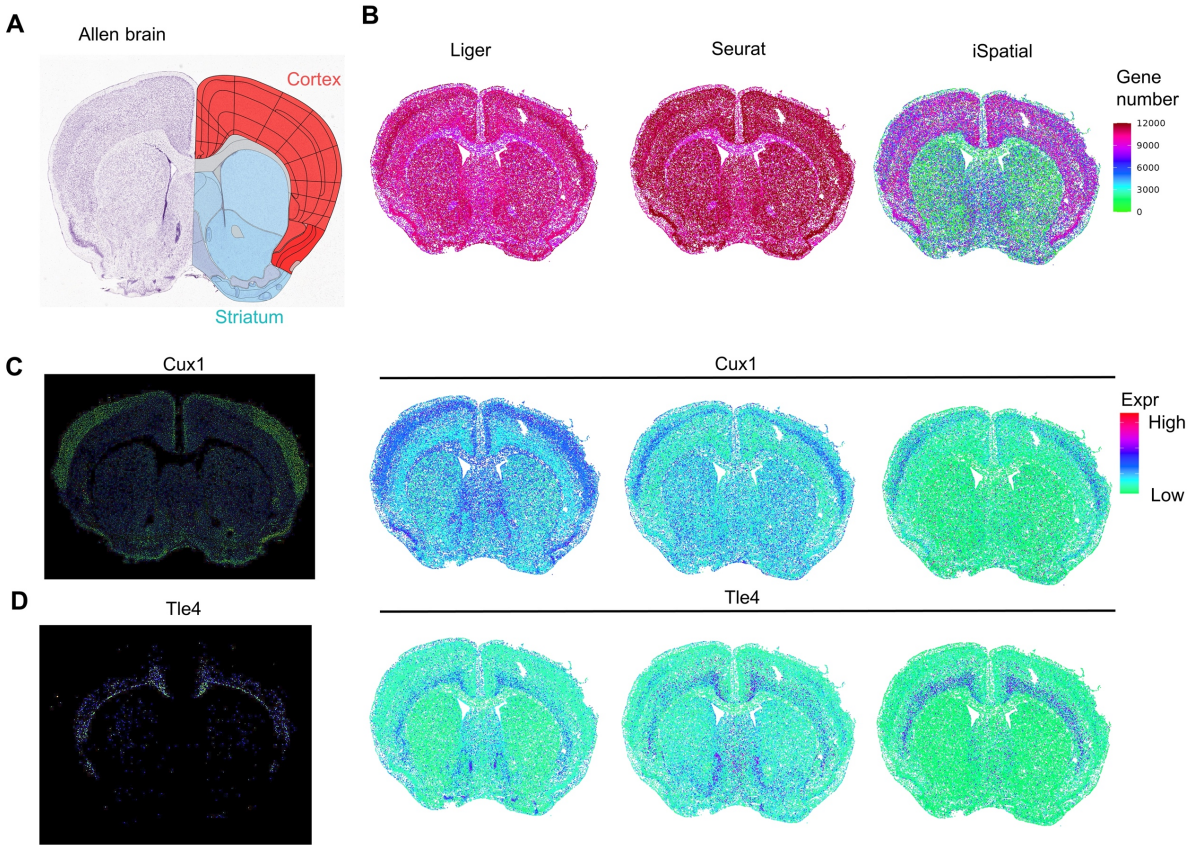


fig. S11: Infer spatial expression by unmatched ST and scRNAseq data. A, Schematic of the anatomic region of mouse cortex and striatum. **B**, Varied number of measured genes in each cell of a whole brain slice inferred by Liger, Seurat and iSpatial. Unmatched MERFISH and scRNA-seq data are used. MERFISH data covers the entire slice. scRNA-seq data only covers cortex region. **C and D**, Representative examples showing the performance of Liger, Seurat and iSpatial in inferring spatial transcriptome with unmatched scRNA-seq data. Allen brain ISH data were showed in left.