Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

STEEL enables high-resolution delineation of spatiotemporal transcriptomic data

Yamao Chen
Fudan University
Shengyu Zhou
Fudan University
Ming Li
University of Chinese Academy of Sciences
Fangqing Zhao
Beijing Institutes of Life Science, Chinese Academy of Sciences https://orcid.org/0000-0002-6216-1235
Ji Qi (qij@fudan.edu.cn)
Fudan University https://orcid.org/0000-0001-7135-0524

Article

Keywords:

Posted Date: January 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1240258/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

STEEL enables high-resolution delineation of spatiotemporal transcriptomic data

Yamao Chen[1,6], Shengyu Zhou[1,6], Ming Li[2], Fangqing Zhao[3,4,5]* and Ji Qi[1]*

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China
[2]School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China
[4]University of Chinese Academy of Sciences, Beijing, China
[5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China
[6]These authors contributed equally: Yamao Chen and Shengyu Zhou

*Correspondence: Ji Qi (qij@fudan.edu.cn) and Fangqing Zhao (zhfq@biols.ac.cn)

Abstract

Advances in spatial transcriptomics enlarge the use of single cell technologies to unveil the expression landscape of the tissues with valuable spatial context. However, computational tools developed for single-cell transcriptomics have great limits in dealing with spatial transcriptomic data with high noise on detected transcript signals. Here we propose an unsupervised and manifold learning-based algorithm, STEEL, which identifies different cell types from spatial transcriptome by clustering cells/beads exhibiting both highly similar gene expression profiles and close spatial distance in the manner of graphs. Comprehensive evaluation of STEEL on various spatial transcriptomic datasets from 10X Visium platform demonstrates that it not only achieves a high resolution to characterize fine structures of mouse brain, but also enables the integration of multiple tissue slides individually analyzed into a larger one. STEEL outperforms previous methods to effectively distinguish different cell types of various tissues on Slide-seq datasets, featuring in higher bead density but lower transcript detection efficiency. Application of STEEL on spatial transcriptomes of early-stage mouse embryos (E9.5 to E12.5) successfully delineates a progressive development landscape of tissues from ectoderm, mesoderm and endoderm layers, and futher profiles dynamic changes on cell differentiation in heart and other organs. With the advancement of spatial transcriptome technologies, our method will have great applicability in high-resolution cell type identification and unbiased spatiotemporal data integration.

Introduction

Single cell transcriptome sequencing techniques (scRNA-seq), e.g. Drop-seq[1, 2], provide detailed profiles of gene transcription to differentiate cell types in a given tissue[3], as well as to monitor dynamic gene expression changes in various developmental stages with high resolution[4]. However, information of the locations of the cells in tissues is lost because of the procedure of tissue disruption[1]. This problem is solved by the advent of spatial transcriptome technologies like Visium platform of 10X Genomics and Slide-seq[5, 6], which barcode 10 or 50 μm beads on rubber-

coated glass to obtain the spatial context in tissues before high-throughput scRNA-seq[1]. It makes tremendous improvement on single cell sequencing technology by providing a map of gene expression for recognition of cell identities[7, 8]. Alternative strategies, including single molecule fluorescence in situ hybridization[9, 10, 11] and laser capture microdissection sequencing[12], are also capable of recognizing cell identities with spatiotemporal variation[13, 14].

Each bead of spatial transcriptomic data contains information of fixed coordinate on slide and gene expression values with considerable noises and amplification biases due to the deficiency of current techniques. Clustering single cells[15, 16, 17] based on transcription similarities usually requires substantial quality control on detected signals of gene expression, and dimension reduction by using PCA, tSNE[18] or UMAP[19] and etc. Cells/beads clustered are then identified as known cell types by well-defined marker genes before being positioned back onto the slide to illustrate their spatial patterns (methods of "clustering before projection" (CBP), Figure 1A). However, it brings great challenges for bioinformatics to correctly identify different cell types due to the disturbance of the high noise and amplification bias when dealing with high-resolution spatial transcriptomic data. Although many efforts have been made[20, 21, 22, 23, 24], these methods either provide limited spatial expression patterns or are very time consuming for analyzing data from recently developed platforms, e.g. 10X Visium (10^4 genes $\times$ 10^3 beads) and Slide-seq[5, 6] (10^4 genes $\times$ 10^4 beads). It leads us to seek alternative strategies for accurate cell type recognition by combining information of both gene expression and spatial context of beads into the clustering procedure (Figure 1A).

Here we present STEEL (Spatial Transcriptome based cEll typE cLustering, https://steel-st.sourceforge.io), an unsupervised method to effectively cluster beads from 10X Visium and Slide-seq[5, 6] data into different groups (cell type candidates). The STEEL method adopts spatial Gini coefficient score as measurement to discover spatially varying genes based on their expression patterns on slides, and then classifies beads/cells based on a simple assumption: beads/cells of the same cell type are more likely to have both similar expression profiles and relatively close spatial distance, to reduce the noises on detected gene expressions of individual beads especially for spatial transcriptome technologies with cellular resolution. Here we define the spatial distance in a "topological" manner: each bead on a slide denotes a vertice of a graph; a pair of beads has a distance of zero if they are within the same group (connected subgraph) or of one if they belong to different groups. Two conditional probability matrixes, P and Q, are calculated to represent similarities of any pairs of beads in high-dimensional expression space and two-dimensional slide space, respectively (Figure 1A). Cross-entropy $C(P, Q)$ is adopted to measure the faithfulness of modeling P with Q. In each clustering step, a pair of adjacent subgraphs (see Methods for details) with minimal increase of entropy is merged into one larger subgraph, and the clustering step is iteratively proceeded until all adjacent subgraphs are merged. The aim of STEEL is to assign beads into a number of clusters, which denotes as "clades" on the hierarchical tree of all beads, to infer cell type candidates in various resolution. Tested by sequencing data on either 10X Visium

or slide-seq[5, 6] platforms, STEEL succeeds in detecting dozens of cell types with higher accuracy compared with approaches based on gene expression only. Application of STEEL on mouse embryo of various stages (E9.5 to E12.5) sequenced in this study reveals clean and prominent identification of cell types highly consistent with classic Kaufman annotation of mouse embryo development[25, 26]. These results suggest that benefiting from integration of spatial information into the clustering procedure, the method is capable of dealing with spatial transcriptomic data from various technologies, regardless the bead size and number, spatial distribution and gene expression level among different tissues.

Results

Cell type recognition on mouse brain datasets (sagittal plane) of 10X Visium

Current spatial transcriptome technologies are now capable of covering a large tissue area, e.g., the entire sagittal or coronal plane of an adult mouse brain can be laid out in two slides on 10X Visium platform, and cell types can be distinguished according to the spatial patterns of gene expression of thousands of beads. To verify the performance of STEEL on analyzing 10X Visium data, we utilized gene expression and bead location profiles of four mouse brain samples, including an anterior and posterior section on sagittal planes and two coronal plane samples (Table S1). After the filtering procedure (see Methods for details), the data of the anterior section contains 2,690 valid beads (99.8% of the total) and 14,768 valid genes (45.7% of total), among which, a total of 508 genes with spatial Gini coefficient[27] ≥ 0.5, are adopted for further cell type identification. According to annotations of *the Mouse Brain in Stereotaxic Coordinates*[28] and the Allen Mouse Brain Atlas[29], STEEL clearly and efficiently clusters these beads into known cell types of olfactory bulb (8 groups), cortex (10 groups), olfactory area (11 groups), stratium (6 groups), white matter (4 groups), etc. (Figure 2A, Supplementary Figures S1 and S2), which are further identified by the expression of multiple marker genes (Table S2). For example, the cluster denoting mitral, nerve and granular layers of olfactory bulb (clusters A17-A18, A9 and A19 in Supplementary Figure S2) displays high consistency with beads expressing *Doc2g*, *S100a5* and *Gng4*, respectively (Table S2). Similarly, the expression of *Abi3bp* and *Cartpt* (Supplementary Figure S1) well supports the cluster denoting piriform cortex (cluster A31 in Supplementary Figure S2, deep blue beads in check mark shape in Figure 2A).

We further apply STEEL on the dataset of the posterior section covering 3,284 valid beads (99.8% of the total) with 14,567 valid genes after quality control. Armed with 1,066 spatially varying genes with high Gini coefficient score, the posterior section is clearly classified into 41 groups (clusters P1-P41): cortex (4 groups), hippocampus (6 groups), thalamus (2 groups), olfactory area (2 groups), cerebellum (5 groups), white matter (8 groups), brain ventricle (4 groups), midbrain (5 groups), and others (5 groups) (Figure 2A). These groups are well supported by the distribution patterns of cell type specific genes (Table S2). Taking cerebellum as an example, Purkinje cells, though with only one layer of beads (cluster P41, Supplementary Figure S3), are clearly recognized

by many marker genes including *Car8*, *Ppp1r17*, *Pcp2* etc. (Table S2). Other well-known cell types, e.g. granules, oligodendrocytes and molecular layers, are also clearly classified (clusters P37-P39, Supplementary Figure S3). The presence of tissue specific genes, as well as many more genes preferentially expressed in two or more types of cells (Table S2), greatly contributes to the successful classification of beads, e.g. *Tnnt1* is specifically expressed in thalamus, and *Tcf7l2* is preferentially expressed in both thalamus and middle brain.

To compare the performance between STEEL and commonly used methods, clustering results yielded from Seurat[15, 16], Mclust[17], BayesSapce[24] and STEEL are evaluated by considering number of outliers, which refers to beads classified into a different group from that of any of its neighbors on the same slide (see Methods for details). Seurat[15, 16], Mclust[17] and BayesSapce[24] exhibit 56-97 outliers in the 20-group classification upon the anterior dataset, and increases to 134-156 outliers in the 40-group classification (Figure 2B). In contrast, only 17 outliers are observed in the clustering results of STEEL in the 20-group classification and the number slightly increases to 22 in the 40-group classification. Similar observations are found on the analyses upon the posterior dataset, with 104-348 outliers for Seurat, Mclust and BayesSapce, and 31-39 for STEEL (Figure 2B and 2C). These results suggest that the noises on detected gene expressions of individual beads can be effectively reduced when the information of spatial context is involved in the clustering procedure. In addition, STEEL shows better performance on sharpening the edges of the adjacent groups (Figure 2B and 2C, see Methods for details) compared with the other three methods, especially when the datasets are divided into 30 or more groups (Supplementary Figures S4 and S5). For example, STEEL clearly recognizes stratium as two larger groups in the center denoting caudate putamen and accumbens nucleus, surrounded by four other single-bead layers (Figure 2A), while the other three methods classify stratium into subgroups with beads spreading in overlapping areas, lacking support from marker gene expression. A particular group in the stratium cluster (probably an olfactory tubercle layer, purple beads at the bottom), though spatially separated from the other groups by cells of olfactory area, exhibits higher gene expression similarity with other stratium groups, supported by *Gpr88*, *Adora2a*, *Lrrc10b*, *Rgs9* and many other marker genes (Table S2).

Integration of spatial context enables STEEL to discover fine structures of hippocampus[30] consistent with H&E stain (Figure 2D): field CA1 of hippocampus, composed of soma (cluster P28, Supplementary Figure S3) and neuropil (P27); CA3 in two groups (P33 and P36); dentate gyrus neurons (DG) in one group but in two spatially separated regions (P34). To further quantify the performance of different methods, we selected three genes preferentially expressed in CA1, CA3 or DG of hippocampus. *Spink8*, a known marker gene in CA1 (also in CA2), is expressed among 22 of 23 beads of Cluster P28 in the posterior section (3 reads per bead on average), exhibiting an AUC value (area under the curve of ROC) of 0.95 (Figure 2E). Similar observations are found in *Bok* (AUC=0.93, a marker gene of CA3, Figure 2F) and *C1ql2* (AUC=0.99, a marker gene of DG, Figure 2G). It is worth noting that the expressions of these genes are also detected on beads

of cortex, white matter or other areas, bringing noises for accurate cell type identification. Consistently, these genes are observed in dispersed area on the tSNE plot (left panels of Figure 2E-2G), suggesting it would be difficult to clearly group beads from CA1, CA3 and DG without taking into account the information of spatial context in the clustering procedure.

Performance evaluation using internal and external validation metrics

We performe comprehensive evaluations on the segmentation results of Seurat[15, 16], Mclust[17], BayesSapce[24] and STEEL by using six metrics, of which three are employed for internal validation based on gene expression and the other three for external validation based on annotated cell types of mouse neural system[31]. Baseline methods, e.g. random segmentation of beads, grouping according to amount of sequencing reads (UMI counts), and K-means algorithm are involved in the comparisons to provide calibrations.

To evaluate the classified structures of different methods, a cross-entropy test is performed to evaluate the similarity between beads and segmentation clusters by considering expression of all spatially varying genes (bootstrap resampling 1000 times). As for the anterior section, It turns out that STEEL outperforms the other three methods by giving the lowest average cross-entropy as 0.71, followed by Seurat (0.78), K-means (0.85), Mclust (0.92), BayesSpace (0.93), UMI thresholds (2.61) and random segmentation (3.61) (Figure S6A, upper panel). Here as for STEEL, an average cross-entropy of 0.71 indicates that the probability of assigning a bead to the labeled cluster is as high as 0.49, while the summed probabilities to all the other 39 clusters is only 0.51. Similarly, STEEL shows better performance by yielding an average cross-entropy of 0.86 on the posterior section (Figure S6A, lower panel), while those of the other methods range from 0.96 (Seurat) to 3.96 (random classification).

To further quantitatively assess the performance of different methods, we employ an atlas of central and peripheral neural system of mouse built by Zeisel et al.[31], which covers 265 cell types and includes most of the known cell types of mouse brain with manual annotation based on the Allen Mouse Brain Atlas[29]. Compared with annotated beads of the anterior section, STEEL exhibits the highest Adjusted Rand Index (ARI) scores of 0.63, 0.66 and 0.47 for segmentations with different cluster numbers (Figure S6B, upper panel), followed by Mclust (0.42, 0.33 and 0.24), Seurat (0.40, 0.32 and 0.24) and BayesSapce (0.37, 0.26 and 0.25). A similar observation is found in another external validation metric Fowlkes-Mallows Index (FMI) (Figure S6C). The evaluation based on Normalized Mutual Information (NMI) (Figure S6D) shows that all four methods perform well on the anterior section with STEEL displaying slightly higher scores (0.71 ~ 0.73 for different cluster numbers) than other methods (0.64 ~ 0.69). The same situation goes for the posterior section of which STEEL outperforms the other methods on 30 and 40 groups classification based on ARI and FMI, and all methods show a similar performance on the NMI metric.

Integration of multiple slides of mouse brain datasets (sagittal and coronal planes)

Dealing with large tissue specimens requires investigation and integration of cell types on multiple slides. Though the anterior and posterior datasets might have been neither collected from

the same plane of the same brain sample nor sequenced in the same run, not to mention that they are classified separately, many tissues flanking the border are identified as same types based on the similarity of gene expression profiles (Supplementary Figure S7). The paired/clustered groups include those of corpus callosum (clusters A2, A6, P4 and P5, Supplementary Figures S2, S3 and S7), field CA3 of hippocampus (clusters A33 and P36), white matter (clusters A1 and P1), cortex and olfactory area.

STEEL is further applied on two mouse brain samples of coronal plane: one has 2,895 valid beads (99.7% of the total) with 14,567 valid genes (44.3% of the total), the other has 2,803 valid beads (99.9% of the total) with 14,035 valid genes (43.5% of the total). The clustering procedures are implemented separately for both sections, resulting in a 35-group cluster for each sample. Although no information was shown that the two sections came from the same slide of the mouse brain, yet the section staining of the two samples displayed close if not the same anatomical position, and the clusters of the two slides also show highly consistent brain structures, e.g. groups of cortex (10 groups vs. 7 groups), hippocampus (7 vs. 7), thalamus (2 vs. 2), hypothalamus (1 vs. 1) and olfactory area (8 vs. 10) white matter (1 vs. 2), choroid plexus (2 vs. 2), stratium (2 vs. 1) are well displayed (Supplementary Figures S8, S9 and S10) according to previous annotations[28, 29]. As expected, integration of the two sections shows that 24 group pairs form a one-to-one correspondence based on gene expression similarity (Supplementary Figure S11) and are further identified by tissue-specific genes (Supplementary Figure S8B and Table S2). For example, the clusters denoting thalamus (CI6 vs. CII6, CI3 vs. CII7) are highly consistent with the expression areas of *Prkcd* and *Ctxn3*. Similar cases go to choroid plexus with *Ecrg4* and *Pltp*; stratium with *Gpr88*; white matter with the wide expression of *Mbp* and *Mobp*.

STEEL gives out a list of cell type specifically or preferentially expressed genes evaluated by spatial Gini coefficient. The similar expression patterns across samples/slides of individual genes enables the recognition of the anatomical positions of the anterior plane as 1.44mm according to annotated structure of mouse brain[28], the posterior plane as 1.68mm (upside) and 2.16mm (bottom side), and the coronal planes as 1.74mm. These results suggest that the clustering results of multiple individual slides detected by STELL can further combined to effectively reconstruct the overall structure of larger tissues. We compared the highly variable genes (HVGs) adopted by Seurat with the spatially varying genes by STEEL upon a 10X Visium dataset of the posterior section. Among the 1,066 genes with spatial Gini coefficient ≥ 0.5, 806 of them are included in the top 2,000 HVGs by Seurat (Supplementary Figure S12), suggesting that the selection of informative genes based on spatial Gini score is more stringent than that used by Seurat. For example, there are four genes shared by the top-10 Gini scores of STEEL and the top-10 HVGs of Seurat: *Ttr* and *Ecrg4* are specifically expressed in choroid plexus, *Car8* in Purkinje cells and *Trh* in olfactory area. Although there is no standard for threshold of spatial Gini coefficient, a score of 0.4~0.5 or higher is usually considered as distribution with high variance. To facilitate the

analysis on spatial transcriptomic data, STEEL provides an option "--gini=" to employ different Gini cut-off values.

STEEL works well on highly noisy and sparse Slide-seq datasets of high resolution

Spatial transcriptome technologies, e.g. Slide-seq with beads of 10 μm, though are making improvements in detecting the fine structures of tissues in cellular resolution, yet currently are only able to provide datasets with less gene numbers and coverage but higher noise (Figure 3A and 3B). It brings more difficulties in dealing with larger amount of data, particularly in accurate clustering and identification of cell types with sparse and fluctuate read counts. We downloaded the datasets of two samples, hippocampus (Puck_200115_08) and olfactory bulb (Puck_200127_15), to compare the performance of Seurat[15, 16], Mclust[17], BayesSapce[24] and STEEL on analyzing Slide-seq V2 data[6].

The high-resolution hippocampus sample contains detailed information of sequenced tissues in cellular base, waiting to be revealed by bioinformatics approaches. It is not difficult to recognize tissues like CA1, DG and choroid plexus, but it is most challenging to distinguish different CA tissues and thin layers like ependymal cells. It turns out that STEEL uncovers more cell types than the other methods, and is the only method that could successfully identify almost all major tissues of hippocampus[30], namely CA1, CA2, CA3, hilus, DG and fasciola cinereum (FC) (Figure 3C, Supplementary Figures S13 and S14). Moreover, single cell layer of ependymal and the central ventricle are also revealed by STEEL suggesting the method is capable to reveal very fine structures when dealing with data featuring in higher bead density but lower transcript detection efficiency. Many known gene markers for different fields of hippocampus, exhibiting relatively high spatial Gini coefficient value, provide evidence for the verification of detected cell types, e.g. *Pcp4* for CA2 and FC, *Ccdc153* for ependymal layer and *Rab3c* for the central ventricle (Figure 3C). Similarly, STEEL successfully divides the beads of the olfactory bulb sample into clear clusters of different cell types, easily recognized as external plexiform, granule, anterior commissure, leptomeninges and glomeruli (Figure 3D, Supplementary Figures S15 and S16). Likewise, these clusters can be verified by known markers genes, e.g. *Omp* for glomerular layer, *Igf2* for leptomeninges, *Nrsn1* for outer plexiform layer, *Doc2g* for mitral layer, *Gng4* for granular layer, *Sox11*, *Nrep* and *Macrod2* for core regions (Figure 3D).

Evaluations on the 667 spatially varying genes of the hippocampus dataset show that STEEL exhibits cross-entropy values lower than those of the other methods (Figure S17A, upper panel). Further evaluations based on three external validation metrics (Figure 4D) using annotated cell types as benchmark, demonstrate that STEEL exhibits more advantages on dealing with sparse and fluctuate Slide-seq V2 data of higher resolution. Another evaluation on the olfactory bulb dataset further shows that STEEL identifies cell types in a clean and prominent manner, highly consistent with previous annotations (Figure S17).

In a more challenging condition of Slide-seq V1[5] with high fluctuation in RNA detection efficiency, and most of genes exhibiting binary signals (0 or 1 read per bead), STEEL successfully

recognizes cell types of granule, Purkinje, cerebellar nucleus, cochlear nucleus, oligodendrocyte and ependymal in the cerebellum sample (Supplementary Figure S18A); CA1, DG, choroid plexus and others, in hippocampus sample (Supplementary Figure S18B); proximal convoluted tubule, distal convoluted tubule, etc. in kidney sample (Supplementary Figure S18C); pericentral and periportal cells in liver sample (Supplementary Figure S18D).

Spatiotemporal atlas of organogenesis in mouse embryo development

Although mouse organogenesis starts at earlier stage, the buds of most of the major organs form at stage E9.5 to E12.5, during which time a great many essential genes have been extensively studied[32, 33]. Therefore, it is of great interest to explore the tissue development and the spatiotemporal variation and distribution of identity determining genes spanning these stages using spatial transcriptome technology. We performed spatial transcriptome sequencing of C57BL/6 mouse embryos at the stages of E9.5, E10.5, E11.5 and E12.5 on 10X Visium platform (Figure 4A) to provide spatial cell atlas of mouse organogenesis and to discover the spatiotemporal expression patterns of cell type-specific genes.

As for the E9.5 sample, the discovery of 1,031 genes with spatial Gini coefficient score > 0.5 enables STEEL to classify 243 beads into 19 cell types, representing brain tissues, dorsal neural cells, heart, aorta and gut tissues (Figures 4B and Supplementary Figure S19). More genes (1,789-3,385) either preferentially expressed in specific tissues or with comprehensive expression patterns are detected in E10.5 (1,043 beads, yielding 25 groups, Figures 4B and Supplementary Figure S20), E11.5 (1,642 beads, 35 groups, Figures 4B and Supplementary Figure S21) and E12.5 (2,778 breads, 45 groups, Figures 4B and Supplementary Figure S22), suggesting a rapid cell type differentiation during embryonic development. The “pseudo-process” of tissue development is then illustrated in PCA analysis of the recognized groups of all four stages based on expression similarity of 4,564 spatially varying genes. Clusters of E9.5 are gathered at the corner, adjacent to those of E10.5 then those of E11.5 and E12.5 at the farthest away end. Altogether, 1,413 of the 4,564 genes (30.9%) exhibit high spatial coefficient scores in a single stage among the samples (Supplementary Figure S23). For example, *SOX10*, a transcription factor essential for neural crest development[34], shows high spatiotemporal variation across the four stages. It is worth noting that these samples are sequenced in the same slide to avoid batch effect influence in subsequent cross-stage comparisons. Furthermore, neural tissues of E9.5 including forebrain (group 13, Figure 4C and Supplementary Figure S19), hindbrain (group 14), dorsal neural cells (group 11) and two regions of brain vesicles (groups 4 and 19) appear at the top of PCA plot, together with similar tissues of E10.5, E11.5 and E12.5 (e.g. groups 37-45, Figure 5C and Supplementary Figures S19-S22). On the other hand, mesoderm originated tissues spread at the bottom, including heart and arteries (groups 1 and 15 of E9.5, groups 17, 18 and 19 of E11,5, groups 6, 8, 9 and 33 of E12,5). These results display a complex scenario of gene expression similarities among various cell types, revealing the expression patterns correlated with both different developmental stages and the orientation of different germ layers.

Expression heterogeneity among beads within clusters allows STEEL to further investigate sub-cell types or populations by involving more spatially varying genes for each stage. In the 19-group clustering results of the E9.5 sample, atria and ventricle have been recognized as two groups (Figure 4B) denoted by specifically expressed markers *Smpx* and *Myh6* (Figure 5A). Consistently, beads of heart tissue of E10.5, E11.5 and E12.5 are further divided into 2-3 subtypes (Figure 5A), with clearly clustered beads for atria and ventricle and an additional population, possible as outflow tracts of ventricles according to Kaufman annotation[25, 26], supported by specific expression of *Clu* (Figure 5A), *Kcne1* and many other markers. Trajectory analysis upon the three subtypes across different stages by Monocle[35] reveals that beads related to atria, ventricle and outflow tracts own a common start point at stage E9.5, while are differentiated via branched paths during development. Atrial beads begin to show intermediate trajectory at E10.5, and many of them reach to their relative maturation point at E11.5, and complete the transition at E12.5 (Figure 5A). Ventricular beads exhibit a similar transition pattern in the opposite branch. These observations provide information for more detailed exploration of heart tissue development.

Another case comes from descending aorta (group 6 of E12.5, Supplementary Figure S22) with three layers of beads, in which *Tagln* shows the variation of expression in radial direction with stronger signals in outer layers than in the center (Figure 5B). Further classification on the descending aorta beads yields two subgroups, one for vascular wall with continuous developmental trajectory and the other for blood cells with more homogeneity in global gene expression. It's worth noting that STEEL adopts a hierarchical clustering strategy, leading adjacent beads with similar gene expression profiles to be clustered prior to those with less-similar profiles. Thus, the beads of vascular wall own higher depth on the clustering tree than those of blood cells, and exhibit higher expression of *Tagln* (Figure 5C). The cluster denoting for vertebral column (group 27 of E12.5, Supplementary Figure S22) shows a more complex pattern: a list of 184 genes including *Sox9* shows an expression gradient in both radial direction with beads in the inner layer owning higher expression of *Sox9* and deeper depth on the clustering tree, and axial direction with higher expression in anterior end than in posterior end (Figure 5B and 5D). Consistently, beads of vertebral column are further classified into four groups by STEEL: one subtype occupying beads in the outer layer, associated with preferential expression of *Twist1* essential for normal development of mesoderm originated cells; the rest beads in the center of vertebral column are divided into three subtypes along with anteroposterior axis, consistent with the expression pattern of the 184 genes (Figure 5D).

Discussion

These results demonstrate that STEEL is effective and robust for analyzing both 10X Visium and Slide-seq data of various tissues, in which the number, spatial distribution and gene expression level of different cell types vary dramatically. We further examine the performance of STEEL on different hyper-parameter perplexity, which measures the number of neighbors of each bead in

high-dimensional expression space, with different values. STEEL displays robust classification on 10X Visium datasets when choosing the perplexity from 10 to 40 for the anterior and posterior sections of the sagittal plane, and the two sections of the coronal plane. Setting the perplexity with smaller values may yield more fragmented groups compared with those adopting larger perplexities, e.g. four cortex layers are yielded when perplexity = 10 while only two layers for perplexity = 40 in the 26-group classification. Another parameter which might affect the final clustering results is the quality control procedure in the first step of STEEL. By default, STEEL ignores those beads with 0.5% or lower of the total expressed genes. Under these thresholds, almost all beads (99.7% to 99.9%) of the four 10X Visium datasets pass the filtering procedure, while only 70.2% to 78.7% of the total beads of the two Slide-seq V2 datasets[6] are used for further clustering. When using a more stringent parameter, e.g. ignoring beads with 1% or less of the total expressed genes, there are only 55.7% to 61.4% of the total beads retained in the two Slide-seq V2 datasets[6], leading to the missed detection of some layers, e.g. oligodendrocytes, however, this parameter barely affects the analyses on the four 10X Visium datasets.

In general, STEEL presents effective and efficient performance on various spatial transcriptome technologies of various bead sizes. It takes STEEL only 1.2 minutes to analyze a 10X Visium dataset (the posterior section of the sagittal plane of mouse brain), similar to Seurat (2.1 minutes) and Mclust (1.4 minutes), must faster than BayesSpace (11.3 minutes) (Supplementary Figure S24A). Similar observations are found when applying the four methods on a Slide-seq V2 dataset (Puck_200127_15): 4.5 minutes for Mclust, followed by STEEL (6.5 minutes), Seurat (8.4 minutes) and BayesSpace (51.1 minutes). STEEL requires 176 Mb memory to perform the analysis of the 10X Visium dataset, much fewer than the other methods (3.2 to 4.8 Gb), while the analysis for the Slide-seq V2 data requires 2.0 Gb memory for STEEL, and 2.4 – 9.4 Gb for the other methods (Supplementary Figure S24B). STEEL can classify hundreds of thousands of beads/cells with affordable time, meeting the requirements of spatial transcriptome technologies with sharp increase of bead number and decrease of bead size, to recognize detailed structures of tissues with cellular resolution.

It is worth noting that cellular information contained in Hematoxylin and Eosin (H&E) staining is not utilized in this method. Integration of tissue structures recognized from H&E images through machine learning algorithms may enhance the accuracy of cell type identification by providing information independent of gene expressions, especially for those cells from narrow layers or on the border between different cell types. In addition, excessive pursuit of reducing outliers in the classification may lead to false identification of cell types in tissues with cell infiltration, and it may limit the applicable range of the method. Alternative strategies, e.g. using recognized cell types by STEEL as “a reference” to re-evaluate the identity of each bead according to similarities on gene expressions, may provide possible solutions to identify cells with dispersed pattern on tissues. Furthermore, alternative efforts have been made to detect spatial information by fluorescence in

situ hybridization, e.g. seqFISH[36] and MERFISH[37], and can be integrated into bioinformatics methods for spatial transcriptome analysis in future studies.

Methods

Mouse embryo dissection and spatial transcriptome sequencing on 10X Visium platform

C57BL/6 mice were obtained from the animal facility of School of Life Sciences, Fudan University. All procedures were approved by the Ethics Committee of School Life Sciences, Fudan University, and were carried out in accordance with guidelines for the Care and Use of Laboratory Animals of Shanghai Municipality, PR China. The protocol was approved by the Science and Technology Commission of Shanghai Municipality (Permit Number: SYXK(hu)2020-0011). The morning of the vaginal plug was defined as E0.5. Dissections were performed and embryos of E9.5-E12.5 were immediately frozen in liquid nitrogen-cooled isopentane and were embedded in pre-cooled OCT. Tissue preparation and optimization were performed on 10X Visium platform, including tissue sectioning, H&E staining, tissue permeabilization, fluorescent cDNA synthesis and imaging, according to the manufacturer's protocol. cDNAs of samples were sequenced on Illumina novaseq platform and were subjected to 150 cycles of paired-end (2 × 150 bp) sequencing.

The mouse reference genome (UCSC version mm10) and corresponding gene annotation models were adopted for the mapping of NGS short reads and measurement of gene expression. Sequenced reads were trimmed by SpaceRanger of 10X Genomics (https://support.10xgenomics.com/spatial-gene-expression/software/overview/welcome) and mapped to reference genome using STAR[38]. Unique molecular identifiers (UMIs) for each spot were counted to remove PCR duplicates. STEEL and Seurat[15, 16] were applied on the expression matrix for further analyses.

Identification of cell types from spatial transcriptomic datasets by STEEL

The nature of this algorithm is to cluster beads of spatial transcriptomics to their most likely cell types in an unsupervised way by combining information of gene expression and spatial context of tissues. We employ the following strategies to perform the clustering procedure: 1) spatial Gini coefficient is estimated for each gene, and only those with coefficient higher than or equal to 0.5 are adopted for further analyses; 2) expression similarity is calculated for any pairs of beads and is then represented as conditional probability ; 2) neighborhood of each bead on a slide is determined and only adjacent beads are possible to be clustered; 3) in the beginning of the clustering procedure, each bead represents a single sub-graph; in a certain step of clustering, a probability Q is calculated according to the partition of sub-graphs, and the cross-entropy $C(P,Q)$ is calculated to represent the fitness of Q to P; 4) if two adjacent sub-graphs are merged into one, a new entropy C' is obtained and only the sub-graph pairs with minimal increase of entropies are combined; 5) the clustering procedure proceeds until all adjacent sub-graphs are grouped and construct a "phylogenetic tree", in which the clades on the top levels are output for further inferring of cell type candidates; 6) genes preferentially presented in one or more sub-graphs are obtained.

1. Calculation of spatial Gini coefficient to discover spatially varying genes

STEEL first filters out genes with bead occupancy lower than 0.05%, and removes beads with gene occupancy lower than 0.5%. After the filtering procedures, all beads are divided into different cells on a grid by their positions on the slide. The size of each cell in the grid is calculated automatically according to the rectangular area covered by valid beads with square root of total beads as expected values for each cell. For each gene, a spatial Gini coefficient in forms of relative mean difference[27] is estimated by

$$Gini = \frac{\sum_i \sum_{j<i} |E_i - E_j|}{(n-1)\sum_i E_i}$$

where E_i is sum of expression values of all beads in the i_{th} cell of grid and n is total number of cells. Genes exhibiting spatial Gini coefficient < 0.5 or exhibiting maximum read count < 10 is ignored on the procedure of bead clustering.

2. Similarity among beads in the space of gene expression

STEEL then calculate Euclidean distance d_{ij} between each pair of beads i and j in the high-dimensional space of gene expression as described in the tSNE method[18]. Denote the similarity of bead j to bead i as the conditional probability $p_{j|i}$, and

$$p_{j|i} = \frac{exp\left(-\frac{d_{ij}^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} exp\left(-\frac{d_{ik}^2}{2\sigma_i^2}\right)}$$

Here we define $P_i = \{p_{j|i}\}$, representing the collection of conditional probability of other beads to bead i. The variance of distance between bead i and the others, denoted by σ_i, is determined by a hyperparameter perplexity, as defined in tSNE[18].

3. Similarity among beads in the space of spatial context of tissues

Euclidean distance between each pair of beads in the low-dimensional space of spatial context of tissues is obtained. For each bead, its distance to the $K-th$ closest beads is collected. A median value of $K-th$ distances across all beads is obtained to determine a common radius of neighborhood R. We then build a graph, in which each bead represents a vertex and only beads with distance shorter than R are potentially connected by an unweighted edge.

Suppose the graph is partitioned into a number of sub-graphs, among which δ_i is the sub-graph containing bead i, we then define the conditional probability of bead j to bead i in the graph as

$$q_{j|i} = \begin{cases} \frac{1}{(N - |\delta_i|)\alpha + |\delta_i|} & if\ j \in \delta_i \\ \frac{\alpha}{(N - |\delta_i|)\alpha + |\delta_i|} & if\ j \notin \delta_i \end{cases}$$

where $\alpha \ll 1$. It means $q_{j|i}$ is close to 1 if beads i and j belong to the same group, while $q_{j|i}$ is close to 0 if not.

4. Calculation of cross-entropy between P and Q

For bead i, the cross-entropy (or Kullback-Leibler divergence) between two probability distributions P_i and Q_i is

$$KL(P_i, Q_i) = \sum_j p_{j|i} ln\left(\frac{p_{j|i}}{q_{j|i}}\right) = \sum_j p_{j|i} ln(p_{j|i}) - \sum_j p_{j|i} ln(q_{j|i})$$

Since $p_{j|i}$ is a constant value in the clustering procedure, the cross-entropy is simplified as

$$KL(P_i, Q_i) \sim -\sum_j p_{j|i} ln(q_{j|i})$$

$$= -\sum_{j\in\delta_i} p_{j|i} ln\left(\frac{1}{(N-|\delta_i|)\alpha + |\delta_i|}\right) - \sum_{j\notin\delta_i} p_{j|i} ln\left(\frac{\alpha}{(N-|\delta_i|)\alpha + |\delta_i|}\right)$$

$$= ln[(N-|\delta_i|)\alpha + |\delta_i|] - \sum_{j\notin\delta_i} p_{j|i} ln(\alpha)$$

$$\sim ln(N\alpha + |\delta_i|) - \left(1 - \sum_{j\in\delta_i} p_{j|i}\right) ln(\alpha)$$

where N is the total number of beads. The sum of the cross-entropies of all beads is

$$C = \sum_i KL(P_i, Q_i) = \sum_i ln(N\alpha + |\delta_i|) + \sum_i \sum_{j\in\delta_i} p_{j|i}\, ln(\alpha) - N ln(\alpha)$$

5. Grouping the pair of sub-graphs with minimal increment of entropy

We define two sub-graphs, A and B, adjacent to each other only when there is at least one pair of beads, $i \in A$ and $j \in B$, which have Euclidean distance in the space of spatial context in tissues shorter than the common radius of neighborhood R.

Suppose two sub-graphs, A and B in step $n-1$, is clustered as $A \cup B$ in step n, we have increase of entropy as

$$C' - C = (|A \cup B|) ln(N\alpha + |A \cup B|) - |A| ln(N\alpha + |A|) - |B| ln(N\alpha + |B|)$$

$$+ \left(\sum_{i\in A}\sum_{j\in B} p_{j|i} + \sum_{i\in B}\sum_{j\in A} p_{j|i}\right) ln(\alpha)$$

then we have

$$C' - C = |A|\, ln\left(\frac{N\alpha + |A| + |B|}{N\alpha + |A|}\right) + |B|\, ln\left(\frac{N\alpha + |A| + |B|}{N\alpha + |B|}\right) + \left(\sum_{i\in A}\sum_{j\in B} p_{j|i} + \sum_{i\in B}\sum_{j\in A} p_{j|i}\right) ln(\alpha)$$

The first two items suggest the clustering "cost", which is proportional to the bead number of the two sub-graphs A and B. The last item suggests the clustering "benefit": the more similar between paired beads, $i \in A$ and $j \in B$, the more benefit is obtained (please note that $ln(\alpha) < 0$).

6. iterative clustering of beads

In the first step of clustering, each sub-graph includes a single bead. The pair of adjacent beads with distance smaller than R, is grouped into a larger sub-graph if exhibiting minimal increase of

cross-entropy, $C' - C$. The clustering continues until all the adjacent sub-graphs are grouped. In the late stage of clustering, two clusters, if both have bead numbers larger than $sqrt(N)$, are allowed to be merged even when they are not adjacent (no beads from the two groups have distance shorter than R). This step enables sub-graphs of the same cell type with distant spatial locations being clustered. In the late clustering step, e.g. when m sub-graphs remain (usually $2 \ll m \ll 40$), a profile is built for each of the m sub-graphs by summarizing expression values of all beads in the sub-graph. A $cosin$ similarity is calculated by using gene expression of each bead and each profile. A bead is re-labeled by the highest $cosin$ similarity and is output as the final clustering result in the current step (m sub-graphs). Gene preferentially presented in one or more groups is measured by using chi-square test using packages in *Numerical Recipes in C*[39].

Evaluation on public spatial transcriptomic datasets

1. 10X Genomics Visium datasets

Under the approval of 10X Genomics, we downloaded four spatial transcriptomic datasets of mouse brain and kidney from the official website of 10X Genomics: (https://www.10xgenomics.com/)

After analyzing each sample separately, we selected a slide of sagittal anterior section, one of posterior section and two coronal sections of mouse brain samples. The method is then applied on all four samples with default parameters: bead occupancy of genes higher than 0.05%, gene occupancy of bead higher than 0.5%, number of neighbors is 20; perplexity as 35. The evaluation of time and memory consuming were performed on a workstation with 32 cores and 512Gb memory by using linux command “time” and “top -n 1000 -d 10 | grep <process ID>”

A list of 37 mitochondrial genes were removed from analyses on slide-seq datasets by an option of “--excluded=” of STEEL. The names of these mitochondrial genes are available at Mouse Genome Informatics (http://www.informatics.jax.org/sequence/marker/MGI:102492).

2. Slide-seq datasets

We downloaded two datasets of Slide-seq V2[6]: hippocampus (Puck_200115_08) and olfactory bulb (Puck_200127_15); and four datasets of Slide-seq V1[5]: cerebellum (Puck_180430_1), hippocampus (Puck_180413_7), kidney (Puck_180528_23) and liver (Puck_180803_8) from the Broad institute’s single-cell repository: (https://singlecell.broadinstitute.org/single_cell/study/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution#study-summary) and (https://singlecell.broadinstitute.org/single_cell/study/SCP354/slide-seq-study)

For each sample, processed slide-seq data includes two files, “Puck*.digital_expression.txt” (or “MappedDGEForR.csv”) for gene expression of beads and “Puck*.bead_locations.csv” (or “BeadLocationsForR.csv”) for bead location on slides, and both files are adopted as input of STEEL. The method is then applied on all samples with default parameters.

3. Evaluation of clustering results

Clustering results of a spatial transcriptomic dataset are evaluated by number of outliers and edge strength between adjacent clusters.

Outlier number and edge strength: Briefly, the spatial context of beads in each cluster is denoted by a gray-scale image, on which beads within the cluster is labeled as 1 while the others as 0. An approximation of Laplace operator, with kernel as $\begin{vmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{vmatrix}$, is adopted on the image to detect outliers, whose Laplacian equals to 8. In other words, an outlier is defined as a bead classified into a group which is different from any of its neighbors on spatial context. The number of outliers are counted for each cluster and are summarized for performance comparison among different methods. In addition, beads on the edge of each cluster are then detected by applying the Laplacian on Gaussian-smoothed image. For each bead on the edge, the gradient magnitude is obtained by convolving the image with Sobel operator, $G_x = \begin{vmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{vmatrix}$ and $G_y = \begin{vmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{vmatrix}$. The gradient magnitude of the bead is given by $|G| = \sqrt{G_x^2 + G_y^2}$. An average of gradient magnitude of beads on the edge of all clusters is calculated for evaluation.

Cross-entropy validation: a cross-entropy test is performed to measure segmentations by considering expression of all spatially varying genes. Briefly, (1) Given a segmentation of N beads $\{b_1, b_2, \cdots b_N\}$ to M clusters $\{c_1, c_2, \cdots c_M\}$, let x_{ij} denote the expression value (log-transformed UMI) of gene g_j on bead b_i; (2) For a group c_k, calculate mean expression value μ_{kj} for each gene g_j; (3) Calculate distance between bead b_i and cluster c_k, $d_{ik} = \left(\sum_j (x_{ij} - \mu_{kj})^2\right)^{\frac{1}{2}}$, and then use the softmax function to convert distance to probability of bead b_i assigning to cluster c_k, as $p_{ik} = \frac{e^{-d_{ik}}}{\sum_k e^{-d_{ik}}}$; (4) Calculate cross-entropy of the probabilities of bead b_i to all M clusters and the segmentation of the bead (a one-hot encoded vector); (5) Calculate mean cross-entropy across all beads; (6) Bootstrap resampling of genes is repeated 1000 times to test the robustness of cross-entropy values.

External validation based on annotated cell types: Zeisel et al.[31] sequenced central and peripheral neural system of mouse, yielding gene expressions of 160,796 cells, and clustered 265 cell types with manual annotation based on the Allen Brain Atlas (http://mousebrain.org/). Expression values of genes per clusters are adopted as “references” to validate segmentations of baseline and ST methods. The ground truth is built by a K-NN like procedure as follows: (1) Collect expression values of all spatially varying genes for each annotated cell type; (2) Assign each bead to a cell type owning the highest Pearson's correlation coefficient; (3) Calculate Pearson's correlation coefficient for each pair of beads to determine their similarities; (4) Reassign each bead to a cell type according to the plurality vote of its K neighbors (here K=20) to build a ground truth for a given dataset; (5) Three metrics, ARI, FMI and NMI, are adopted to compare the ground truth

and segmentations by using function “compart” in the R package “flexclust” and functions “mutinformation” and “entropy” in package “infotheo”. Among the 265 cell types by Zeisel et al.[31], 22 types of spinal cord are ignored from analyses. Lists of 244 cell types are adopted for benchmarking the sagittal sections of mouse brain datasets, 60 types for hippocampus dataset and 15 types for olfactory bulb dataset. For Slide-seq datasets with sparse read counts, gene expression values of each bead are averaged according to its nearest 20 neighbors. All source codes and analyzed results are provided at SourceForge to facilitate validation by readers (https://sourceforge.net/projects/steel-st/files/Materials/Evaluation/).

Data availability

All spatial transcriptomic data of mouse embryo samples (E9.5, E10.5, E11.5 and E12.5) are available at NCBI Gene Expression Omnibus with accession number GSE178636 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178636).

Code availability

STEEL is a stand-alone C++ program, and its source code and binary tool files are available at SourceForge (https://steel-st.sourceforge.io), and are also deposited at CodeOcean platform (https://codeocean.com/capsule/2852687/tree). To facilitate the reproduction of analyses in this study, results of STEEL upon all datasets are available at SourceForge.

Author Contributions

J.Q., Y.C. and F.Z. designed the study and managed the project. J.Q. performed model construction and programming. Y.C. and F.Z. performed data analysis and validation. S.Z. performed source code refinement. J.Q. and M.L. performed theoretically derivation. J.Q., Y.C. and F.Z. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (grant number 32070247, 32025009) and funds from the State Key Laboratory of Genetic Engineering at Fudan University.

Decalration of Interests

The authors declare no competing interests.

References

1. Macosko EZ*, et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

2. Saunders A*, et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-+ (2018).

3. Zeisel A*, et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).

4. Morris SA. Gene-expression maps fall into place. *Nature* **569**, 197-199 (2019).

5. Rodriques SG*, et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).

6. Stickels RR*, et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol*, (2020).

7. Ebbing A*, et al.* Spatial Transcriptomics of C. elegans Males and Hermaphrodites Identifies Sex-Specific Differences in Gene Expression Patterns. *Dev Cell* **47**, 801-813 e806 (2018).

8. Giacomello S*, et al.* Spatially resolved transcriptome profiling in model plant species. *Nat Plants* **3**, 17061 (2017).

9. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5**, 877-879 (2008).

10. Choi HMT*, et al.* Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, (2018).

11. Codeluppi S*, et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* **15**, 932-935 (2018).

12. Nichterwitz S*, et al.* Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat Commun* **7**, 12139 (2016).

13. Eng CL*, et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239 (2019).

14. Peng G*, et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev Cell* **55**, 802-804 (2020).

15. Stuart T*, et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

16. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

17. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**, 289-317 (2016).

18. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *J Mach Learn Res* **9**, 2579-2605 (2008).

19. Becht E*, et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37**, 38-+ (2019).

20. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. *Nat Methods* **15**, 343-346 (2018).

21. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods* **17**, 193-200 (2020).

22. Miller BF, Bambah-Mukku D, Dulac C, Zhuang X, Fan J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities. *Genome Res*, (2021).

23. Dries R*, et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* **22**, 78 (2021).

24. Zhao E*, et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol*, (2021).

25. Kaufman MH. *The atlas of mouse development*.

26. Richardson L*, et al.* EMAGE mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Res* **42**, D835-844 (2014).

27. Rey SJ, Smith RJ. A spatial decomposition of the Gini coefficient. *Lett Spat Resour Sci* **6**, 55-70 (2013).

28. Paxinos G, Franklin KBJ. *The mouse brain in stereotaxic coordinates*, 3rd edn.

29. Lein ES*, et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168-176 (2007).

30. Schröder H, Moser N, Huggenberger S, SpringerLink (Online service). Neuroanatomy of the Mouse An Introduction.). 1st edn.

31. Zeisel A*, et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014 e1022 (2018).

32. Dickinson ME*, et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514 (2016).

33. Meehan TF*, et al.* Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat Genet* **49**, 1231-1238 (2017).

34. Deal KK, Cantrell VA, Chandler RL, Saunders TL, Mortlock DP, Southard-Smith EM. Distant regulatory elements in a Sox10-beta GEO BAC transgene are required for expression of Sox10 in the enteric nervous system and other neural crest-derived tissues. *Dev Dyn* **235**, 1413-1432 (2006).

35. Trapnell C*, et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).

36. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. *Nat Methods* **11**, 360-361 (2014).

37. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

38. Dobin A*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

39. Birchenhall C. Numerical Recipes in C - the Art of Scientific Computing. *Econ J* **104**, 725-726 (1994).

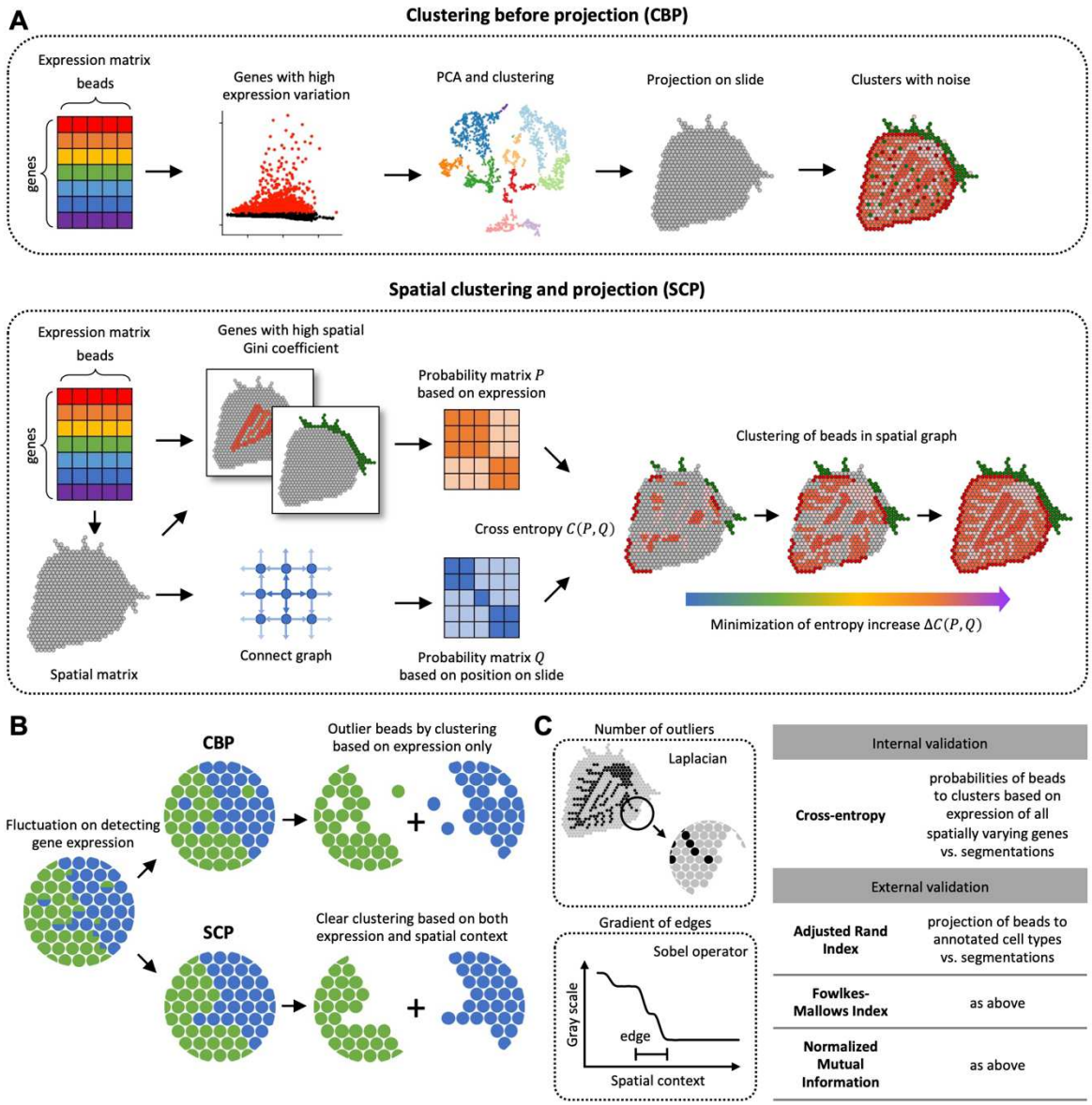

Figure 1. Schematic diagram of two strategies for analyzing spatial transcriptomic data. (A) The pipelines of the CBP strategy (clustering before projection) and SCP (spatial clustering and projection). (B) Noise in detected gene expression values might lead to different classification results by different strategies. (C) Validation metrics for quantitative evaluation.

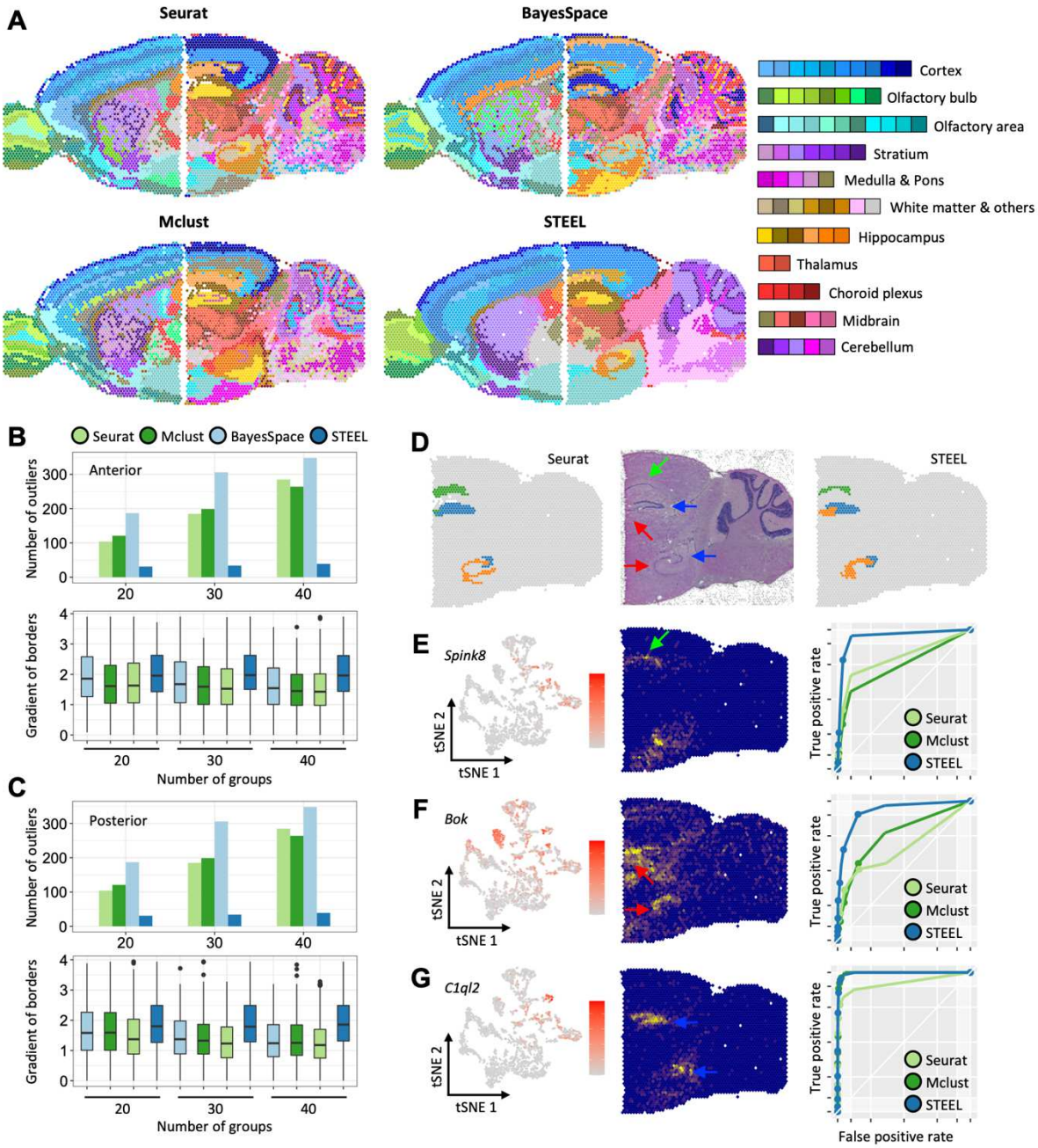

Figure 2. Performance comparisons of Seurat, Mclust, BayesSpace and STEEL on 10X Visium datasets of two sagittal sections (anterior and posterior) of mouse brain. (A) Illustration of 40 detected cell types in anterior and posterior sections by the four methods, respectively. (B) and (C) Lists of outlier numbers (upper panel) and average edge gradient (lower panel) for performance comparisons on the two datasets. (D) Illustration of the identified cell types of hippocampus on the posterior dataset by Seurat and STEEL. (E), (F) and (G) Beads with detected expression of three markers genes, *Spink8*, *Bok* and *C1ql2*, are shown on tSNE map (left) and spatial expression map (middle). AUC evaluation of classification accuracy of Seurat, Mclust and STEEL based on expression level of the three marker genes is shown on the right. BayesSpace is not involved in AUC evaluation as fields of hippocampus are not distinguished from those of cortex and olfactory area.

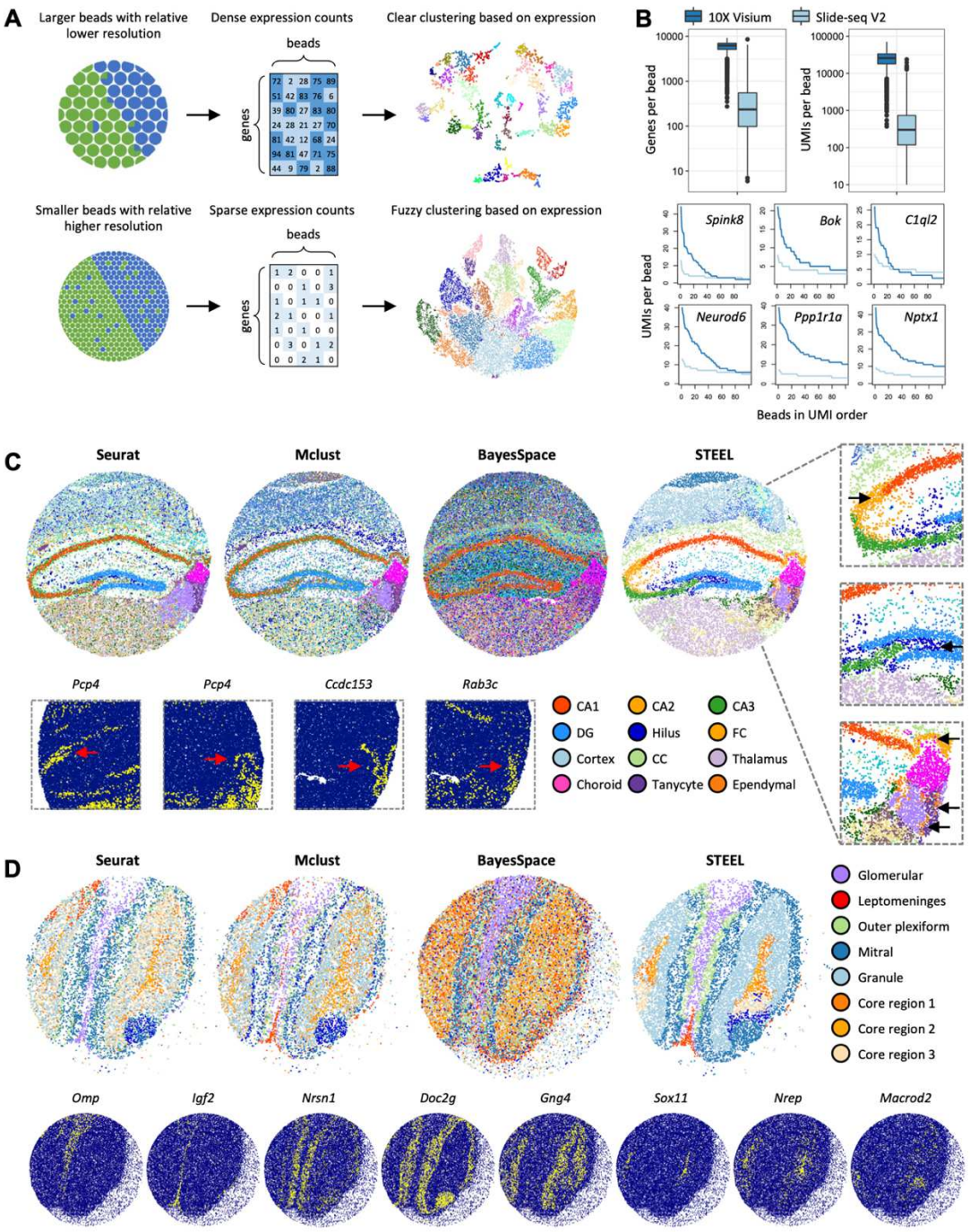

Figure 3. Application of STEEL for cell type detection on a cerebellum dataset by Slide-seq V2. (A) Cartoon representing the difference of bead densities and transcript detection efficiency between 10X Visium and slide-seq V2. (B) Comparison of gene number and UMI number per bead between 10X Visium (coronal dataset) and slide-seq V2 (Puck_200115_08). Detected UMIs of six marker genes of hippocampus by the two technologies are listed in the lower panels. (C) Comparison of clustering results (20 groups) by Seurat, Mclust, BayesSpace and STEEL on a hippocampus sample (Puck_200115_08) of Slide-seq V2. Fields of CA2, Hilus, FC, ependymal layer and the central ventricle are displayed in three insets. Heatmaps on the bottom denote for spatial expression of *Pcp4*, *Ccdc153* and *Rab3c*. (D) Comparison of clustering results (9 groups) by the four methods on an olfactory bulb sample (Puck_200127_15). Spatial expression of *Omp*, *Igf2*, *Nrsn1*, *Doc2g*, *Gng4 Sox11*, *Nrep* and *Macrod2* are displayed for evaluation of detected cell types of glomeruli, leptomeninges, outer plexiform, mitral cells, granule, and core regions, respectively.

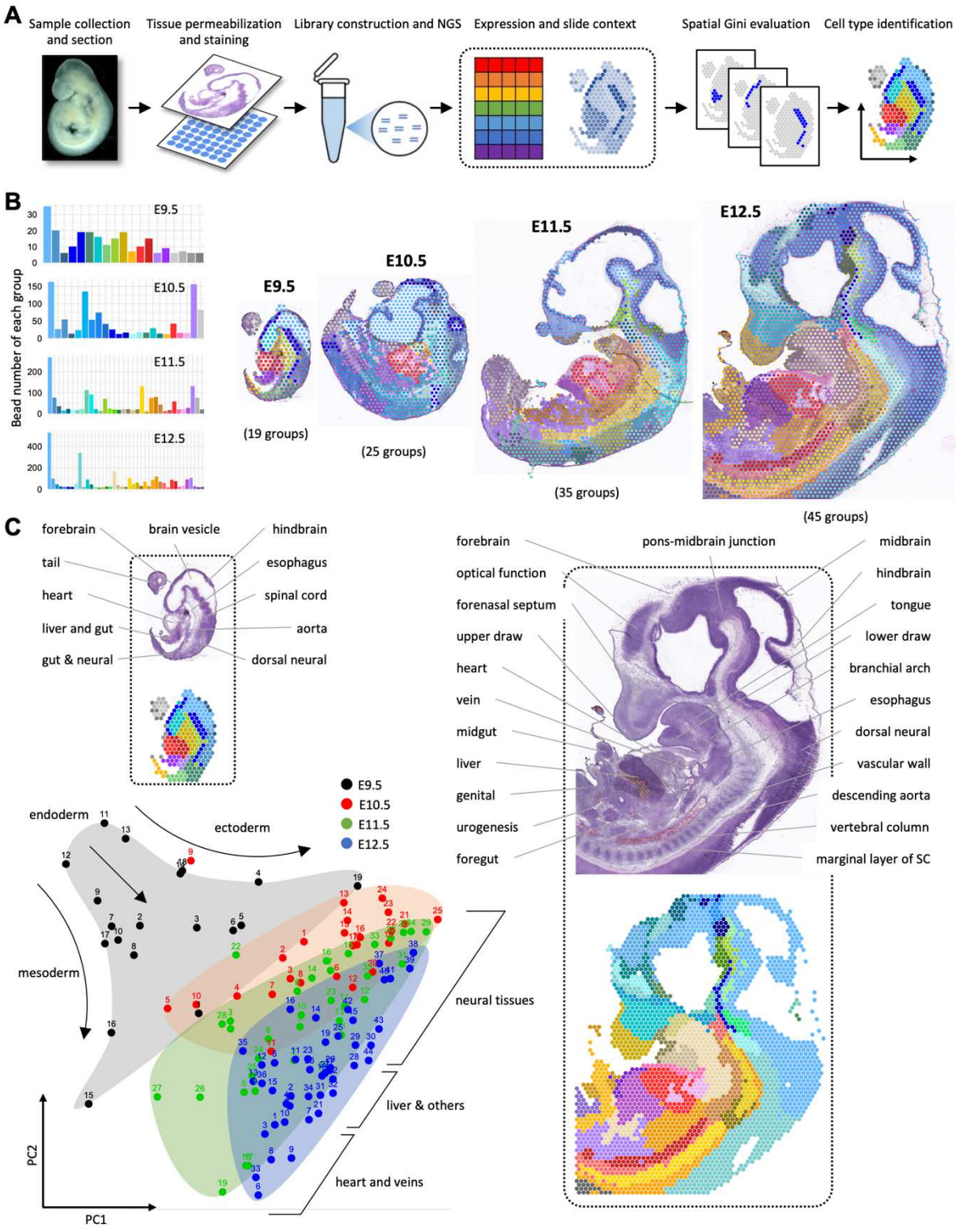

Figure 4. Reconstruction of organogenesis in early developmental stages of mouse embryo. (A) A workflow for sampling and sequencing of stages E9.5 to E12.5 of mouse embryo on 10X Visium platform. (B) Illustration of discovered cell types on the four stages (19 to 45 clusters) using STEEL, placing over corresponding H&E stain images. Bead numbers for each cluster are shown on the left. (C) Similarity of 118 detected cell types originated from three germ layers is displayed on PCA plot.

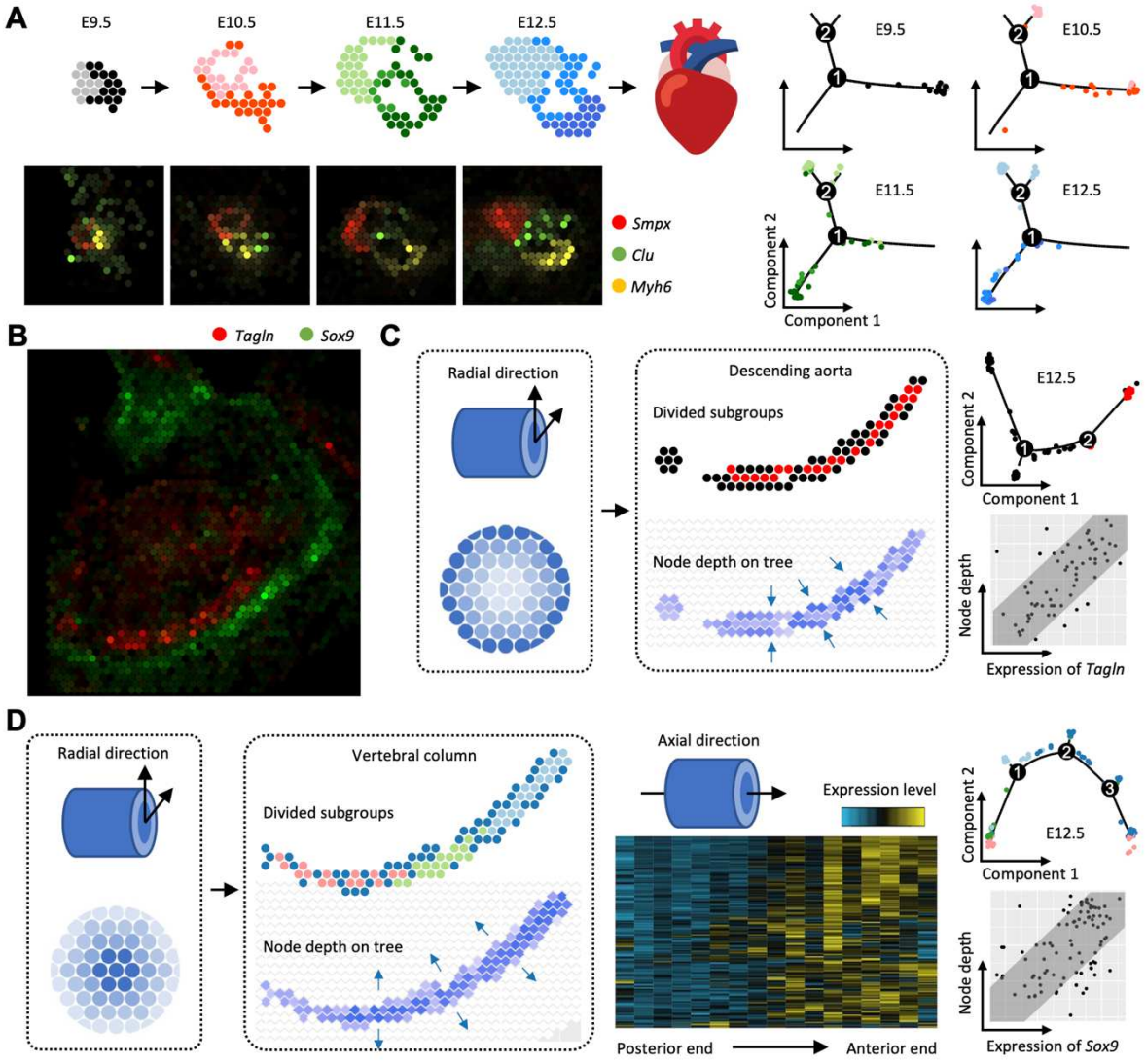

Figure 5. Discovery of subtypes on selected tissues of mouse embryo. (A) Heart tissues of E9.5-E12.5 are further classified into 2-3 groups by STEEL. Spatial expression patterns of three genes, *Smpx*, *Clu* and *Myh6*, are shown on lower left. Pesudo-time estimation of cell-cycle progression for each sub-type of E9.5-E12.5 is shown on the right. (B) Spatial expression patterns of *Tagln* and *Sox9* (E12.5), which are preferentially expressed in descending aorta and vertebral column. (C) Illustration of sub-types of descending aorta of E12.5 discovered by STEEL, and display of node depth for each bead on the tree of hierarchical clustering. (D) Sub-types and node depth for beads of vertebral column of E12.5. Expression patterns of 184 genes preferentially expressed in vertebral column are shown in the heatmap along with antero-posterior axis. Pseudo-time estimation and correlation between expression variation and node depth are shown on the right panel of (C) and (D).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingInformationcompressed.pdf