nature methods

Resource

https://doi.org/10.1038/s41592-023-01773-7

SODB facilitates comprehensive exploration of spatial omics data

Received: 10 August 2022

Accepted: 6 January 2023

Published online: 16 February 2023

Check for updates

Zhiyuan Yuan[1,2,7] ✉, Wentao Pan[2,3,7], Xuan Zhao[2], Fangyuan Zhao[4,5], Zhimeng Xu[2], Xiu Li[3], Yi Zhao[4,5], Michael Q. Zhang[6] ✉ & Jianhua Yao[2] ✉

Spatial omics technologies generate wealthy but highly complex datasets. Here we present Spatial Omics DataBase (SODB), a web-based platform providing both rich data resources and a suite of interactive data analytical modules. SODB currently maintains >2,400 experiments from >25 spatial omics technologies, which are freely accessible as a unified data format compatible with various computational packages. SODB also provides multiple interactive data analytical modules, especially a unique module, Spatial Omics View (SOView). We conduct comprehensive statistical analyses and illustrate the utility of both basic and advanced analytical modules using multiple spatial omics datasets. We demonstrate SOView utility with brain spatial transcriptomics data and recover known anatomical structures. We further delineate functional tissue domains with associated marker genes that were obscured when analyzed using previous methods. We finally show how SODB may efficiently facilitate computational method development. The SODB website is https://gene.ai.tencent.com/SpatialOmics/. The command-line package is available at https://pysodb.readthedocs.io/en/latest/.

Quantifying molecular profiles within the endogenous spatial context can enable the systematic understanding of tissue organization. In recent years, people have witnessed great advances in diverse spatial technologies, and the molecules that can be spatially resolved include, so far, messenger RNAs (mRNAs)[1–3], proteins[4–6], metabolites[7–9] and DNAs[10]. Spatial transcriptomics (technologies for spatially quantifying mRNA expressions)[11], which is the most developed and widely used class of technologies[12,13], can be divided into imaging-based[14–17] and next-generation sequencing (NGS)-based[1,2,18,19] categories. Although enabling whole-transcriptome profiling, classical NGS-based spatial transcriptomics (for example, spatial transcriptomics (ST)[1], 10X Visium[21] and Slide-seq[2]) inherently suffer from limited spatial resolution and low mRNA capture rate[20,21]. Imaging-based spatial transcriptomics (for example, MERFISH[14] and seqFISH[16]) have some complementary advantages to the NGS-based methods, such as high spatial resolution and capturing rate, but they are limited both in the number of targeted genes to profile and in the sizes of the fields of view. In this respect, several recent works, such as HDST[22], Slide-seqV2 (ref. [18]) and Stereo-seq[23], have been introduced as potential alternatives by improving on various limitations. Spatial proteomics (strictly speaking, spatially resolved high-plex protein profiling, SRHP)[21,24,25], another important class of spatial technologies, is mainly fulfilled by multiplexed antibody-based imaging. On the basis of different antibody-labeling strategies, mainstream SRHP technologies include fluorophore-labeled (for example, t-CyCIF[26] and 4i[27]), DNA-labeled (for example, CODEX[6]) and metal-labeled (for example, MIBI-TOF[28] and IMC[29]) technologies. In addition, spatial metabolomics[30], spatial genomics[10] and spatial multi-omics (simultaneously quantifying two or more types of molecules)[31–37] are emerging spatial technologies that are increasingly gaining more attention. In the literature, all the

[1]Institute of Science and Technology for Brain-Inspired Intelligence; MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence; MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China. [2]Tencent AI Lab, Shenzhen, China. [3]Shenzhen International Graduate School, Tsinghua University, Shenzen, China. [4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. [5]University of Chinese Academy of Sciences, Beijing, China. [6]Department of Biological Sciences, Center for Systems Biology, The University of Texas, Richardson, TX, USA. [7]These authors contributed equally: Zhiyuan Yuan, Wentao Pan. ✉e-mail: zhiyuan@fudan.edu.cn; michael.zhang@utdallas.edu; jianhuayao@tencent.com

above spatial technologies, which enable multiplexed profiling, are termed 'spatial omics'.

Thanks to these technical advances, a tremendous amount of data is generated to fuel global investigations of complex organism spatial biology, especially in disease[38–41], tumor microenvironment[42–52], normal tissue homeostasis[2,32,41,53–56] and development[23,57–62]. Due to differences in research purposes and the origins in diverse laboratories, these data are originally deposited in a variety of repository platforms (Supplementary Fig. 1). Some data are maintained in general-purpose data repositories, such as Gene Expression Omnibus, Zenodo and figshare (Supplementary Fig. 1). Some data are maintained by dedicated scientific institutions, for example, Single Cell Portal and Spatial Research (Supplementary Fig. 1). Other data generated by large consortium projects are stored on their webservers, for example, Human Tumor Atlas Network (HTAN) for three-dimensional (3D) atlases of different human cancers, ImmunoAtlas for immune atlas construction and Brain Initiative Cell Census Network (BICCN) for multi-omics atlas of brain cell types (Supplementary Fig. 1). Data generated by commercial companies are sometimes listed as sample datasets on their own websites, for example, IONpath for MIBI data and 10X Genomics for 10X Visium data (Supplementary Fig. 1). Such heterogeneous data resources and representations demand great efforts and tedious operations for ordinary researchers to process and utilize them.

Several databases have been presented for spatial data deployment and to provide basic analytical modules online[63–65]. However, they separately suffer from drawbacks such as limitations in user interaction, lack of cell-type/tissue region annotations or a lack of newly released data types/technologies. Importantly, they are only focused on depositing spatial transcriptomics datasets, meaning that other classes of spatial omics technologies, such as spatial proteomics, metabolomics, genomics and multi-omics data, are ignored.

In this manuscript, we propose SODB, an online platform that combines a large-scale data deployment for general spatial omics datasets and a suite of interactive analytical modules. We first take an overview of SODB's datasets and functions and compare them with other existing platforms. Next, we systematically evaluate SODB's data characteristics and statistics. Then, we introduce SODB's interactive modules using an annotated spatial transcriptomics dataset, and demonstrate SODB's scalability on large-scale datasets using two SRHP datasets. Furthermore, we explain SOView, a unique interactive visualization module, and its algorithmic and application principles on various spatial omics datasets. We additionally demonstrate how SOView can be used to delineate known tissue structures and identify tissue structures that cannot be seen by other methods. By combining SOView's ability to locate regions of interest within tissues and the interactive region selection function, we demonstrate how SODB is used to identify unexpected tissue regions and companion marker genes. Finally, we show how SODB could fuel the development of computational methods.

Results

Overview

Currently, there are three popular databases (namely, SpatialDB[63], STomicsDB[65] and SOAR[64]) designed for spatial data visualization and deposition. In this section, we summarize their features and introduce SODB's unique advantages compared with them.

As a pioneer database for spatial transcriptomics, SpatialDB[63] provides datasets from eight different biotechnologies, and implements basic functions, such as data searching, downloading, gene comparison and spatial expression visualization[63]. However, it only provides raw text data format, which needs to be further transformed to be processed with downstream computational software[63]. The weakness in interactive functions and limited data types also hinder its wider usability. In comparison, STomicsDB[65] is an improvement with regard to the range of spatial data types and user interaction interface[65]. One can effectively inspect the gene expression values or meta information by simply mouse-hovering over the interested cell on the spatial map[65]. STomicsDB also provides analytical results, such as gene distributions and spatial marker genes[65]. SOAR[64] was released shortly after STomicsDB, and covers a similar range of data types as STomicsDB. The main strength of SOAR is its online spatial analytical modules, such as spatially variable gene analysis and cell-type interaction analysis[64]. SOAR's weaker points, compared with STomicsDB, are that it provides gene expression maps and cell-type annotation maps as constant images, preventing users from browsing cells/regions of interest.

Like the databases described above, SODB also provides both spatial data deployment and interactive data exploration (Fig. 1). Specifically, datasets can be efficiently accessed either by browsing according to a tree structure (Supplementary Figs. 2 and 3) or by searching according to the dataset properties (Supplementary Fig. 4). SODB provides interactive data exploration, including easy inspection (for example, mouse-hovering and selection on cell or tissue of interest), automatic statistics (for example, cell-type composition and expression-value distribution within user-selected regions) and basic spatial analysis (for example, gene comparison and spatially variable gene analysis). In addition, SODB presents data using a unified data format for convenient interaction with downstream analytical pipelines, for example Scanpy[66] and Squidpy[67]. With this data format, cell-wise and feature-wise annotation are easily incorporated.

In addition to these features, there are four additional distinguishing strengths of SODB. The first is the wide range of spatial data types and the large volume of datasets (Supplementary Tables 1–3). SODB covers multiple classes of spatial technologies (for example, spatial transcriptomics, proteomics, metabolomics, genomics and multi-omics) compared with other existing databases, which only provide spatial transcriptomics datasets. The volume of included datasets is larger than any existing database (see Methods and Fig. 1g). The second strength is that SODB provides an interactive visualization module named SOView, which can be used to quickly preview the global structure of tissue, and also to identify subtle but important tissue structures that are obscured in previous analyses. The third strength is that SODB provides an interactive display panel, which can be combined with SOView to automatically produce molecular markers for the user-defined regions. The fourth strength is that we have provided a command-line package for more efficient data fetching for computational groups (Methods).

Data organization and export

In SODB, data are organized using a hierarchical tree containing five levels, that is, root, Biotech category, Biotechnology, Dataset and Experiment (Supplementary Fig. 2). In this tree, the children of a node are subordinate to the node itself. The Biotech category level (Supplementary Fig. 2b) consists of different classes of spatial technologies, including spatial transcriptomics, SRHP, spatial metabolomics, spatial genomics and spatial multi-omics. Biotechnology (Supplementary Fig. 2c) is the next level of the Biotech category, which contains specific spatial technologies. For example, the children of 'spatial transcriptomics' of the Biotech category level contains 10X Visium, ST, Slide-seq and 11 other spatial transcriptomics technologies (Supplementary Fig. 2c). Each Biotechnology node has one or more datasets (Dataset level), and each dataset is generally attached to a project or publication (Supplementary Fig. 2d). One dataset may consist of multiple replicates or control slices, and we term each slice as an "experiment", which is the leaf node of the tree (Supplementary Fig. 2e).

Each experiment contains the spatially resolved molecular profiles of a set of spots. (For the sake of simplicity, we denote the observation unit of spatial omics as a 'spot', which can be understood as cell/pixel/bead/other according to the spatial technology used.) In SODB, the Experiment is the smallest object available for analysis and downloading. As typical spatial omics data are presented, the data of each

https://doi.org/10.1038/s41592-023-01773-7

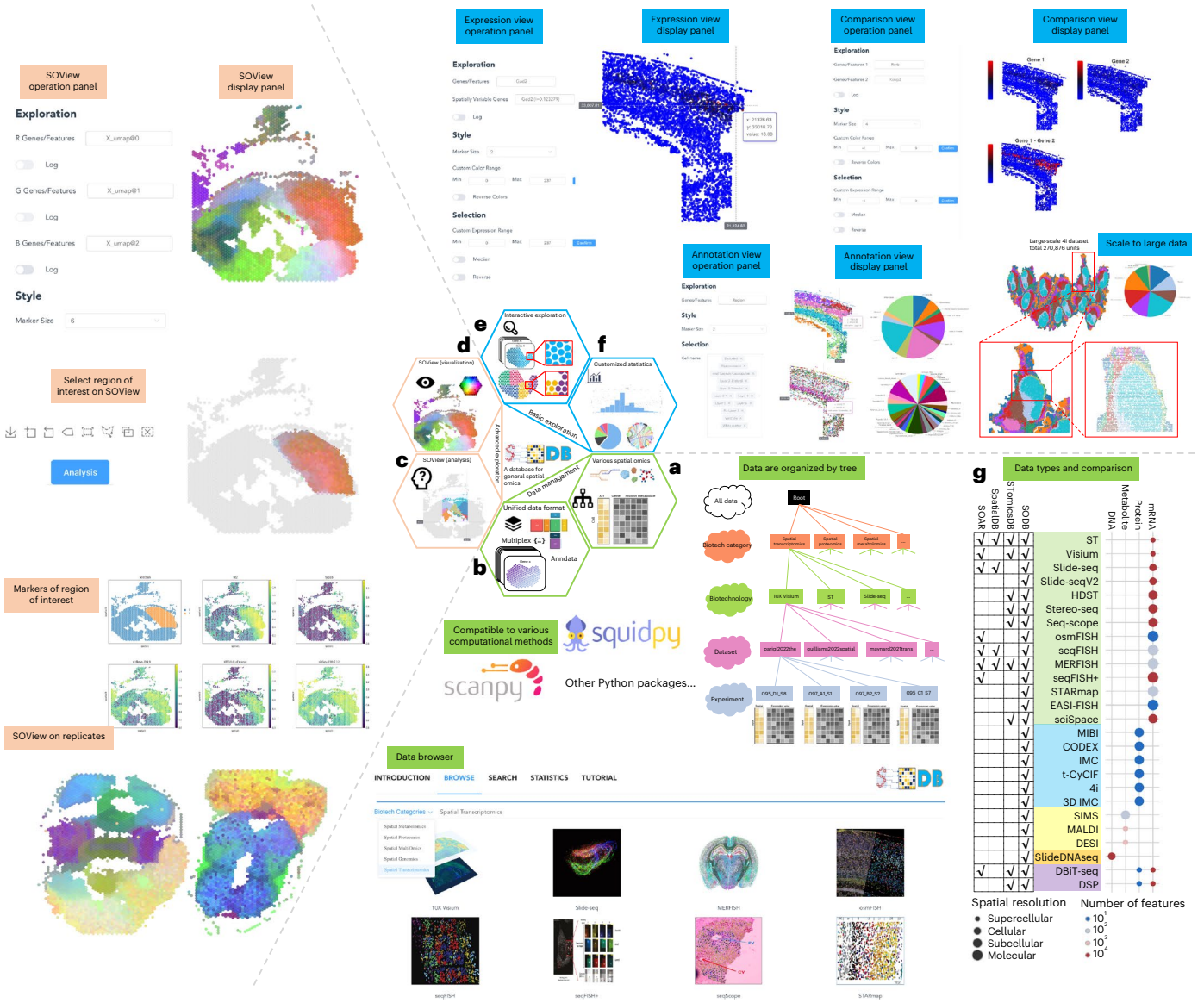


Fig. 1 | Overview. a–f, The overall design of SODB. The six hexagons summarize the six features of SODB. SODB contains various types of spatial omics data (**a**), and these data are processed and presented as a unified Anndata format (**b**). SODB supports interactive exploration (**e**) and customized data statistics modules (**f**). SODB further provides a unique data exploration module named Spatial Omics View (SOView). SOView supports both efficient data visualization (**d**) and interactive data analysis (**c**) functions. The six hexagon features are categorized into three groups: data management (**a**,**b**), basic exploration (**e**,**f**) and advanced exploration (**c**,**d**). For each group, more detailed information is shown with different colored titles. **g**, Information of different spatial technologies, including the types of profiled molecules, the number of molecular features and the spatial resolution. And the comparison of included spatial technologies among different platforms: SODB, STomicsDB, SpatialDB and SOAR.

experiment consist of two matrices (that is, a spatial coordination matrix and a molecular expression matrix) with the same number of rows standing for spots, and the columns of the two matrices stand for the x–y coordinates and molecular features, respectively (Supplementary Fig. 2e).

SODB provides both graphical user interface (GUI) and non-GUI ways for exporting data. Using the GUI, SODB provides the 'download' button in every experiment page, so that users can directly obtain the data of interest by clicking the button. For non-GUI users, SODB provides a command-line package, pysodb, for efficient data downloading (Methods: 'Command-line package'). By benchmarking against conventional data loading practice, we show that pysodb can substantially save time and memory usage for biologists and bioinformaticians (Methods: 'Method comparisons on data loading'). When loading Slide-seqV2 datasets (Supplementary Table 10), the conventional data loading approach costs 19.04 minutes and 21.97 gigabytes on average, which is hardly possible with personal computers, while SODB reduced the time and peak memory usage to 7.16 seconds and 0.04 gigabytes on average (Supplementary Fig. 24). If the same dataset has been previously loaded in the same machine, the time cost could be further reduced to 0.20 seconds.

Data characteristics and statistics

We collected spatial omics data according to the data availability information provided by the original publications (Supplementary Table 2). These data were generally deposited in various platforms (Supplementary Fig. 1) in raw formats. We manually curated these data based on original publications as well as well-established data-processing pipelines, which resulted in more than 2,000 experiments (Supplementary Table 1) of tissue samples from seven different species (Fig. 2a).

https://doi.org/10.1038/s41592-023-01773-7

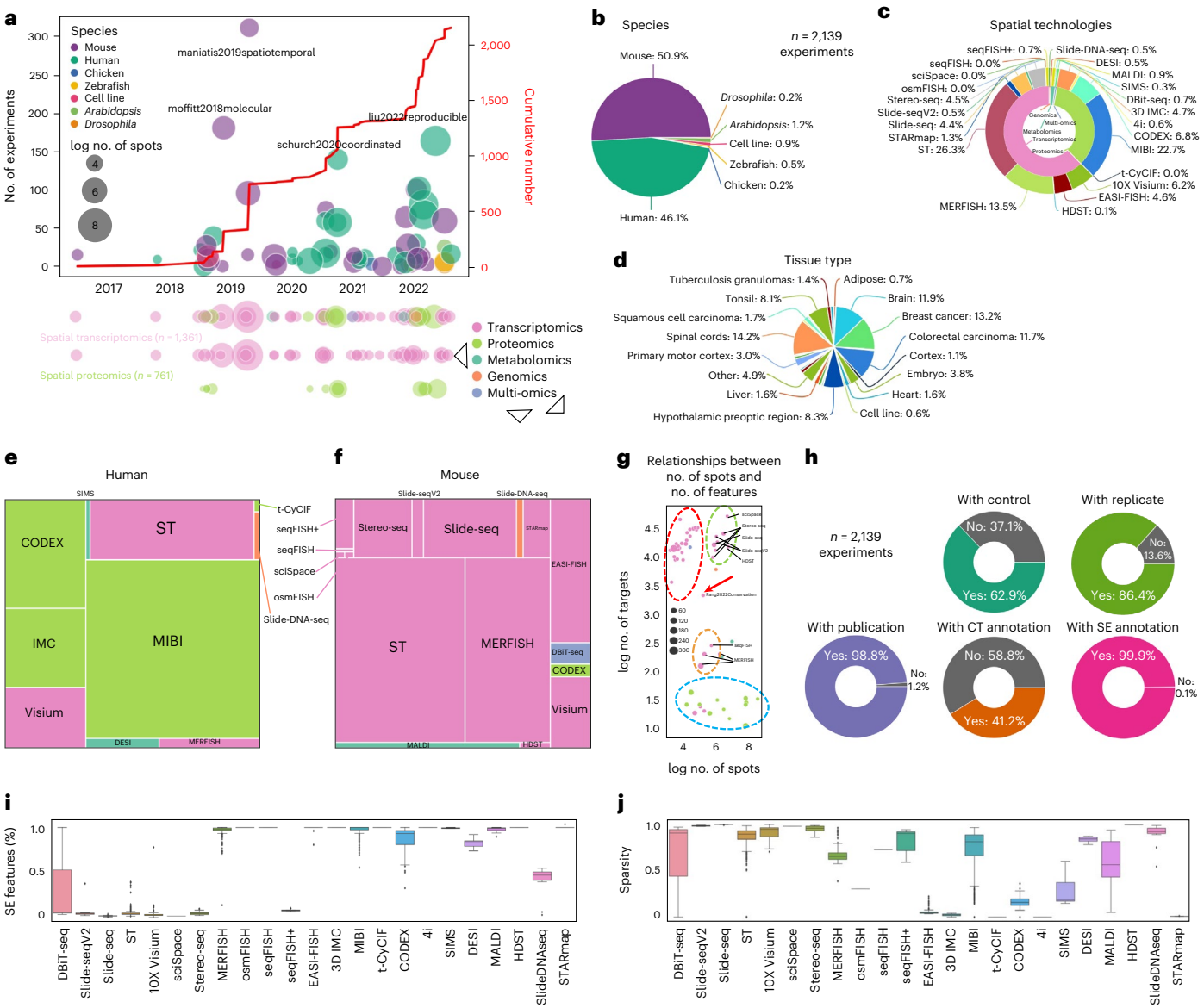


Fig. 2 | Data characteristics and statistics. a, The scatter plot shows the generated date of datasets, each dot is a dataset and is colored by the species and sized by the number of spots, the *y* axis is the number of experiments in the dataset and the *x* axis is the release time of the dataset. The line plot shows the cumulative number of experiments across time. The timeline plot shows the number of experiments of different categories of spatial technologies over time. **b**–**d**, Pie charts showing the distribution of experiments by species (**b**), technologies (**c**) and tissue type (**d**). **e**,**f**, Tree map showing the distribution of technologies used in human (**e**) and mouse (**f**) studies. **g**, Scatter plot showing the relationships between the number of spots and the number of molecular features. Each dot is a dataset and is colored by the categories of spatial transcriptomics technologies and sized by the number of experiments. Dashed circles of different colors are different clusters. Black solid lines indicate the spatial technologies of datasets. The red arrow indicates a representative MERFISH dataset. **h**, Pie charts showing the distribution of experiments by different quality controls. **i**, Box plots showing the percentage of SE features grouped by different technologies. Each point in the box plot is the percentage of SE features (molecular features with Moran's $I > 0$) of one experiment. Total $N = 2,139$ independent experiments. **j**, Box plots showing the sparsity of expression matrix grouped by different technologies. Each point in the box plot is the sparsity (the percentage of zeros in expression matrix) of one experiment. Total $N = 2,139$ independent experiments.

Mouse and human were the two most studied species, and consisted of 50.9% and 46.1% of all experiments, respectively (Fig. 2b). Mouse data occupied the majority of studies before late 2019, while human studies increased substantially to an amount comparable to the number of mouse studies after 2020 (Fig. 2a).

With regard to tissue types (Fig. 2d), different brain regions were among the most studied, including cortex regions, which were often used to benchmark new spatial technologies[17,18,68], and also the focused regions of the recent BICCN products[69,70]. Other brain regions, such as hypothalamic preoptic[71], nucleus accumbens[72] and olfactory bulb[1,22] were also studied. In addition, whole brain data were available but in low spatial resolution[19,73]. With the development of large field-of-view (FOV) and high-resolution technologies (for example, Stereo-seq[23]), a single-cell 3D atlas of the whole brain will hopefully be available in the near future. Apart from neuroscience studies, other organs, such as liver[53,54] and heart[57,58], are also preferred targets (Fig. 2d), in which liver has well-studied zonation patterns[74] and heart development stages were investigated in both human[58] and chicken[57]. In cancer research, breast cancer[1,22,45,50,75] and colorectal carcinoma[47,76] are two prominent targets (Fig. 2d).

We also investigated the various categories of spatial technologies involved in SODB (Fig. 2c). As expected, spatial transcriptomics and proteomics are the two major classes, accounting for 62.6% and 35.3% in all experiments, respectively (Fig. 2c). Data for spatial genomics and multi-omics were limited since the available technologies were rare, but we reserved space and interfaces for more such data in the future (Fig. 2c). In spatial transcriptomics data, ST, as the earliest spatial transcriptomics technology, accounted for the largest proportion (26.3%) of all experiments, followed by MERFISH (13.5%), which is the most widely used imaging-based spatial transcriptomics (Fig. 2c). In SRHP data, MIBI, a successful commercialized technology, explained the largest proportion (22.7%) of all experiments, while CODEX took the second position (6.8%) (Fig. 2c).

Of note, human and mouse studies substantially differed in the spatial technologies used (Fig. 2e,f). In human, more than half of the experiments were generated by SRHP technologies (for example, MIBI, IMC and CODEX), while a few relatively mature spatial transcriptomics technologies (for example, ST and Visium) were also used for human study (Fig. 2e). In mouse, on the contrary, almost all experiments were produced by spatial transcriptomics technologies (for example, ST, MERFISH, Visium and Slide-seq) (Fig. 2f).

Among most spatial technologies there existed a trade-off between the number of spots and molecular features (Fig. 2g). In the scatter plot (Fig. 2g), all SRHP datasets clustered together at the bottom right (Fig. 2g, blue dashed circle), which is expected since SRHP technologies generally enjoy strength in finer spot resolution while suffering from limited (<100) protein multiplexing[21]. A similar statement could be made for spatial metabolomics, but these had limited interpretation and annotation for their features (that is, mass-to-charge ratio, m/z). Specifically, time of flight-secondary ion mass spectroscopy data relied on chemical standards or isotope tracing to annotate metabolites[9,77], while matrix-assisted laser desorption/ionization (MALDI) and desorption electrospray ionization enjoyed some computational tools for annotation inference[78,79]. Spatial transcriptomics exhibited a rather diverse distribution in the plot (Fig. 2g). The major proportion of spatial transcriptomics datasets was located at the top left (high gene throughput and low number of spots; Fig. 2g, red dashed circle), and these datasets were mainly generated by classical spatial transcriptomics technologies, for example, 10X Visium[19] and ST[1]. New technologies, such as sciSpace[55], Slide-seqV2 (ref. [18]), HDST[22] and Stereo-seq[23] were improved with regard to spatial resolutions as well as spot throughput (Fig. 2g, green dashed circle). Another cluster of spatial transcriptomics datasets (mainly imaging-based technologies; Fig. 2g, yellow dashed circles) had a smaller number of targeted genes compared with traditional ones, while they contained larger numbers of spots (cells), which was competitive with those in green dashed circles. One dataset, which was distal from any clusters (Fig. 2g, red arrow), was from a very recent MERFISH paper[80], which improved the number of targeted genes to several thousands.

We then evaluated the data quality of the experiments (Fig. 2h, $n = 2{,}139$). More than 98% of experiments were published in peer-reviewed journals (Fig. 2h), and the remaining experiments were either from sample data from commercial websites or from well-established databases. There were 62.9% experiments with a control, and 86.4% experiments with replicates (Fig. 2h). In general, spatial transcriptomics had better replicates and worse controls than SRHP (Supplementary Fig. 7). We curated the cell-type annotation according to the original manuscript, resulting in 41.2% well-annotated experiments (Fig. 2h). We also provided spatially variable (SE)[67,81] annotations for molecular features in almost all datasets (>99.9%), which could be conveniently accessed on the website (Fig. 2h and Supplementary Fig. 9b).

We further assessed the percentage of molecular features that exhibited spatially variable patterns (Methods) across all spatial technologies (Fig. 2i). Not surprisingly, sequencing-based spatial transcriptomics technologies had low SE percentages (~0), compared with most imaging-based spatial transcriptomics technologies, which had high SE percentages (~1). One exception was seqFISH+, which was the extension of seqFISH and could cover transcriptome-scale gene expressions (~10,000)[68]. All SRHP technologies had SE percentages of ~1 because the targeted proteins tended to be spatially informative markers. Other spatial technologies, for example, spatial genomics and metabolomics, showed modest SE percentages (Fig. 2i).

We finally quantified the data sparsity (Methods) for each spatial technology (Fig. 2j), which showed diverse distributions. As expected, all sequencing-based spatial transcriptomics technologies showed high data sparsity (~1), especially for high spatial resolution technologies, such as HDST[22] and Slide-seq[2]. Imaging-based spatial transcriptomics technologies (except seqFISH+) had lower sparsity because the targeted genes were typically densely distributed and the technologies, per se, had higher mRNA capturing rates[20] (Fig. 2j). Similarly, SRHP technologies also tended to have lower sparsity (generally even lower than imaging-based spatial transcriptomics technologies), since their number of targeted proteins was below 100 (Fig. 2g,j). Spatial metabolomics exhibited larger sparsity compared with proteomics since they share similar data representation (meshed pixels), while the former had much more molecular features. The only spatial genomics technology, that is, slide-DNA-seq[10], whose experimental protocol was inherited from Slide-seq[2], also had data sparsity of ~1 (Fig. 2j).

Data exploration

SODB provides convenient ways to interactively explore data. For each experiment, the data consist of the molecular expressions of spots, the associated spatial coordinates, as well as some attributes of each spot (such as cell-type annotation, tissue-domain annotation, etc.). SODB provides four data exploration views, namely Expression view (Fig. 3a), Annotation view (Fig. 3f), Comparison view (Supplementary Fig. 10) and SOView (Fig. 4). The first three basic views are introduced in this section, and SOView will be introduced in the following sections. We use Slide-seqV2 (ref. [18]) data in the mouse hippocampus region to demonstrate the three interactive data exploration views. This demonstration can be accessed at https://gene.ai.tencent.com/SpatialOmics/dataset?datasetID=1. The web page contains two panels, the top panel is for dataset detail (Supplementary Fig. 8a), which is used to display the necessary information from the dataset, and the bottom panel is for data exploration (Supplementary Fig. 8b), which is used to provide users with interactive data exploration.

Expression view. This view can be used to explore the spatial expression values of selected genes (or other molecules), one gene at a time. The gene to be displayed is selected by the user from the drop-down menu in the 'operation panel' (Supplementary Fig. 9a). The spatial expression of the selected gene is shown on the 'display panel' (Supplementary Fig. 9f). The user can freely choose whether to perform a logarithm operation on the raw count, through the log switch (Supplementary Fig. 9c). Note that SODB additionally provides a spatially varying (SE) gene drop-down menu (Supplementary Fig. 9b), in parallel to the gene selection menu (Supplementary Fig. 9a), to help users select those genes exhibiting spatially varying patterns (Supplementary Fig. 11). Users can also customize the marker and color style (Supplementary Fig. 9d), which will respond instantly on the display panel. The display panel of the Expression view also supports interactive operations, including region zoom-in or zoom-out (Fig. 3a,b) by clicking the button designated by the red arrow in Supplementary Fig. 9f, and inspecting the expression value of a spot of interest by mouse-hovering (Fig. 3c). Additionally, users can also select a region of interest (ROI) using a rectangular or polygon selector by clicking the button indicated by the green arrow in Supplementary Fig. 9f, then the expression-value distribution in the selected region will be displayed synchronously in the bottom part of the display panel (Fig. 3d). Another way for the user

https://doi.org/10.1038/s41592-023-01773-7

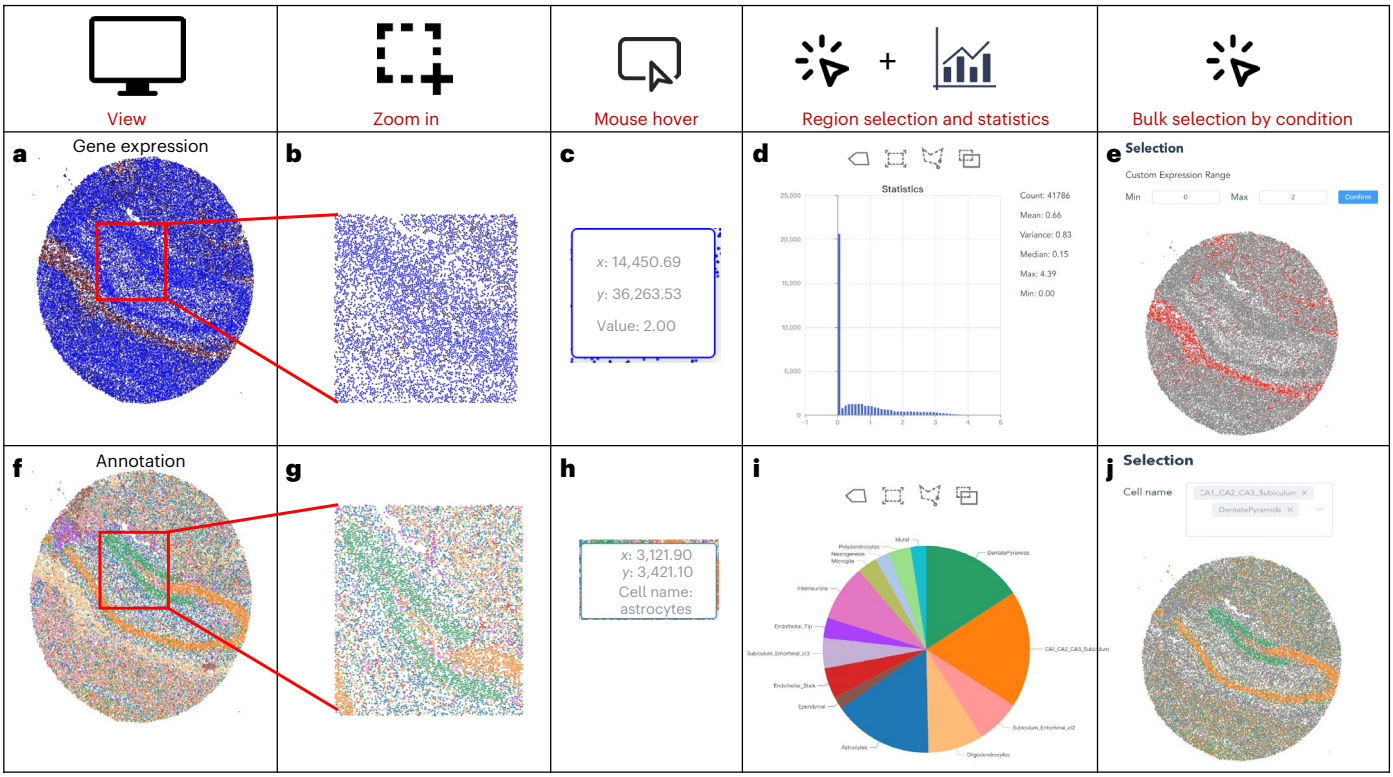


Fig. 3 | Interactive views of SODB. a,**f**, There are four views for interactive data exploration: Expression view (**a**), Annotation view (**f**), Comparison view (Supplementary Fig. 10) and SOView (Fig. 4 and Supplementary Fig. 19). **a**–**e**, In Expression View (**a**), one can zoom in to see the detailed spatial expression for ROI (**b**), mouse-hover to see the spatial and expression information of spots of interest (**c**), obtain distribution information of gene expression within selected ROIs (**d**) and inspect the spatial distribution of spots selected by condition (**e**). **f**–**j**, In Annotation view (**f**), one can zoom in to see the detailed spatial cell-type distribution within the ROI (**g**), mouse-hover to see the spatial and cell-type information of spots of interest (**h**), obtain distribution information of cell types within selected ROIs (**i**) and view the spatial distribution of selected cell types (**j**).

to select regions (Fig. 3e) is by setting the condition in the operation panel (Supplementary Fig. 9e).

Annotation view. The purpose of this view is to explore the spatial distribution of a selected property of a spot. The property could be cell-type annotation, tissue-domain annotation or other category of property of the spots. Similar to the Expression view, the user can select the annotation to be displayed through the drop-down menu on the operation panel (Supplementary Fig. 12a). The marker size can be customized as in Supplementary Fig. 12b. The Annotation view also provides functions such as zoom-in and zoom-out of selected regions (Fig. 3f,g), hovering the mouse to display annotation information (Fig. 3h) and selecting ROIs by rectangle or polygon selector to obtain the included cell-type ratios (Fig. 3i). Annotation view also provides the option to highlight the interested cell type(s) or other annotations (Fig. 3j and Supplementary Fig. 12c), which is especially useful when viewing large numbers of cell types when only a small number of them are of interest.

Comparison view. This view is for efficiently comparing the relative expression levels of two selected genes, and showing their differences in space. The user can select two genes of interest through the operation panel (Supplementary Fig. 10a). The spot-wise difference of the raw count values of the selected genes is then displayed instantly in the display panel (Supplementary Fig. 10c). Other options in the operation panel (for example, style and selection) are similar to those in Expression view (Supplementary Fig. 10b).

Delineating known tissue structures with SOView

An important feature of SODB is the provision of a unique interactive visualization tool, Spatial Omics View (SOView). With SOView, users can easily get an overview of the tissue structure characterized by the rich molecular features from a single map (Fig. 4). We first use an SRHP experiment to demonstrate how SOView processes the raw multiplexed data to generate the colorful SOView map (Fig. 4a). This experiment used 4i[27] technology to measure the subcellular resolved ~50 protein measurements of 13 HeLa cells (Supplementary Fig. 14a). By comparing with the annotation (Supplementary Fig. 14c) provided by the original publication[27,67], one can see that the SOView map could clearly differentiate the nucleus and cytoplasm with distinct colors (Supplementary Fig. 14b,c), and 13 nuclei share similar pink colors (Supplementary Fig. 14b). It is worth noting that although these 13 cells are all HeLa cells, the color of their cytoplasm in the SOView map is slightly different; for example, the cytoplasm of some cells is blue (Supplementary Fig. 14b, blue arrow) and of some cells is green (Supplementary Fig. 14b, green arrow). We reasoned that there might be some proteins that were differentially expressed in these two groups of cells. After exploring the data using the Expression view of SODB, we found that CTNNB1 protein was at a higher level in cells 141, 143, 120, 122, 142 and 136 (with green cytoplasm) than in other cells (with blue cytoplasm), both visually (Supplementary Fig. 14d) and quantitatively (Supplementary Fig. 14e). The elevated expression of CTNNB1 might be related to cell crowding according to the original report[27].

To test SOView on spatial metabolomics data, and to use a simple tissue with just three dominate structural factors to illustrate SOView's advantages for visualization, we use a MALDI dataset for wheat seed (Fig. 4b). We use dictionary learning[82] to extract three dominant classes of metabolic features from ~1,000 ion images and explore the metabolic expression using SODB's Expression view (Supplementary Fig. 15), and the merged image of the three representative ion images could reflect the basic structure of the data (Fig. 4b, left). As expected,

https://doi.org/10.1038/s41592-023-01773-7

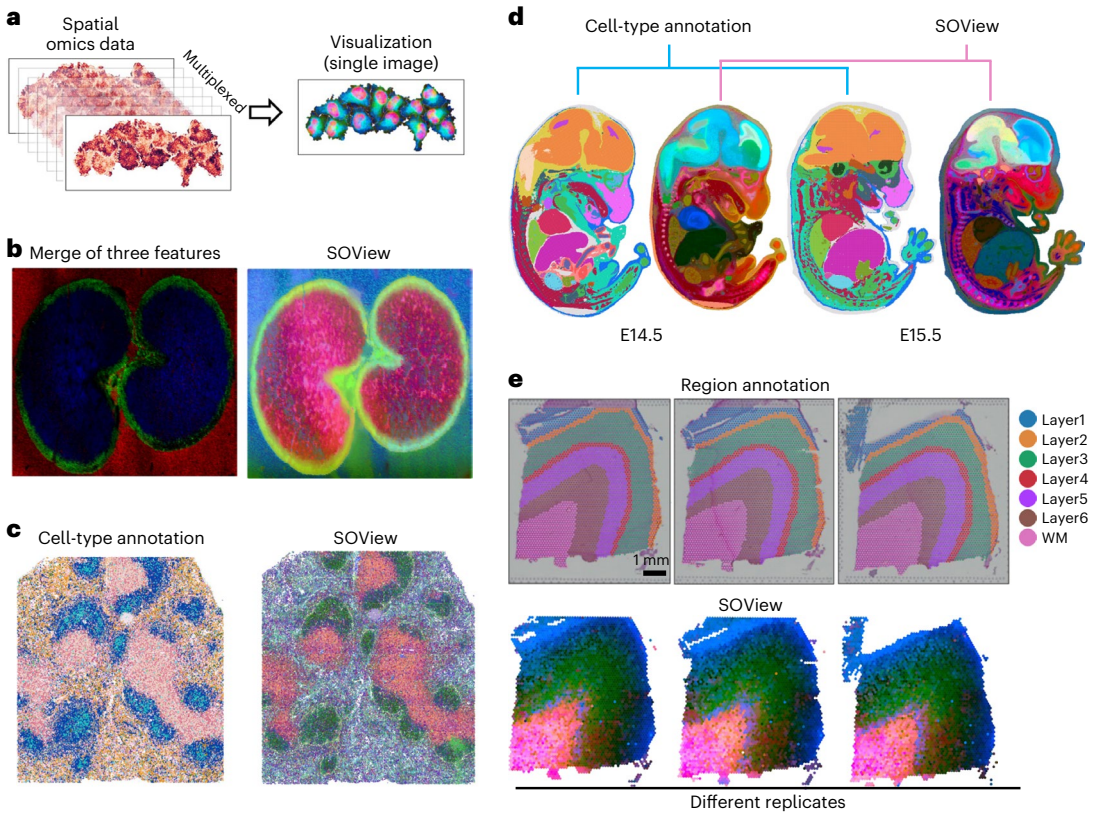


Fig. 4 | SOView demonstration of various spatial omics datasets. a, SOView inputs spatial omics data and outputs a single colorful image for interactive visualization. **b**, Visualization comparison between a merge of three individual ion images (left) and SOView (right) on a spatial metabolomics wheat seed dataset. **c**, Visualization comparison between cell-type mapping (left) and SOView (right) on a spatial proteomics mouse spleen dataset. The cell-type mapping is generated by SODB, the cell-type color legend is in Supplementary Fig. 16d. **d**, Visualization comparison between cell-type mapping (left) and SOView (right) on a spatial transcriptomics mouse embryo dataset. The cell-type mapping is generated by SODB, the cell-type color legend is in Supplementary Fig. 17. **e**, Visualization comparison between region annotation (top) and SOView (bottom) on a spatial transcriptomics human cortex dataset.

we find that the SOView map (Fig. 4b, right) shows the symmetrical structure of the seeds very clearly and is consistent with the merged image in Fig. 4b (left). Although a merged map of three single molecular maps can obtain similar visualization results as SOView, the advantage of SOView is that one does not need any knowledge of the tissue in advance, nor is it necessary to manually or computationally select those important molecular features to visualize the global view of the tissue structure. This strength of SOView is more obvious in more complex tissues with >3 dominant factors.

The spatial cell-type map, which is typically obtained by clustering followed by cell-type labeling, is a standard practice to display tissue structure and study the spatial distribution of cell types. However, there are some inconveniences when using this practice for the purposes of data visualization, especially with a large number of cell types. To demonstrate visualization with SOView in such cases, we take the SRHP data of a mouse spleen[6] with 28 cell types (Fig. 4c and Supplementary Fig. 16c,d) as an example to compare the visualization performance between cell-type map and SOView. In the SOView result, the four major classic splenic compartments, that is, red pulp, B cell follicle, PALS (periarteriolar lymphoid sheath) and marginal zone (Supplementary Fig. 16a,b) can be clearly differentiated by distinguishing colors. In particular, SOView outlines PALS, red pulp and B-follicle in very different colors, reflecting their substantial differences in protein content. On the contrary, the cell-type map (Fig. 4c, left, and Supplementary Fig. 16c,d) is slightly cluttered due to its random coloring, so that the color difference cannot represent the difference between cell types, which hinders the user in distinguishing the main structure of spatial data for visualization purposes. In addition, clustering and cell-type labeling additionally require complex clustering parameter tuning and time-consuming manual labeling. It can be concluded that when users browse spatial omics data, we believe that SOView is more suitable as a quick visualization tool, in which the color is more meaningful than the color of the cell-type map. But, when the user intends to dive into cell-type interaction analysis in detail, clustering and cell-type labeling steps are necessary.

For samples with more molecular feature dimensions (for example, spatial transcriptomics) and more complex tissue structures, global visualization of the tissue landmarks of the whole sample is even more important. We next used the spatiotemporal transcriptomics data (Stereo-seq technology) for mouse embryonic development[23] (Fig. 4d) at E14.5 and E15.5. By comparing SOView with the spots annotation (Fig. 4d and Supplementary Fig. 17), one can observe that SOView can not only differentiate different organs with discriminative colors, but also find subcompartments inside individual organs, such as brain, heart, liver, lung and pancreas (Supplementary Fig. 17). This strength of SOView allows users to quickly understand the global structure of a whole sample and reveals some local heterogeneity within specific subregions.

Since SOView coloring is based on the similarity of the molecular expression profiles of spots, we want to test SOView using tissues with known spatial continuity to verify whether SOView could generate gradient color patterns. For this, we adopt the dorsolateral prefrontal cortex (DLPFC) dataset[83] of 10X Visium spatial transcriptomics, which covers cortex layers from one to six and white matter (WM). By comparing the SOView results and the region annotation results (Fig. 4e), one can observe both the gradient of the colors from Layer1 to Layer6 in

the three replicates, and the color gap between Layer6 and WM (Fig. 4e, bottom). This is also consistent with the recent conclusion published with the BICCN project that the cell types and gene expression exhibit a gradient distribution along the cortex axis[69]. From a practical view, if one has no previous knowledge of the gradient nature of the cerebral cortex, it is difficult to observe the continuity of gene expression in this tissue, either by means of region annotation (Fig. 4e, top) or by exploring individual gene expression. In contrast, SOView can both easily achieve this goal, to reveal the continuity and gradient, and interactively inspect the pattern by mouse-hovering and region selection.

In summary, we conclude that SOView, as a spatial omics visualization tool: (1) can support a global overview of the tissue structure and reveal heterogeneity within substructures; (2) can reflect the difference in the molecular expression profile with the color differences of its automatic color assignment, thus revealing the underlying continuity of the tissue; (3) can compare well with cell-type maps, where SOView is more suitable for visualization purposes, which can avoid clustering parameter adjustment, laborious cell-type labeling and avoid the problem of color crowding in cases with large numbers of cell types. In the next section, we demonstrate the advantages of SOView over other methods in discovering unexpected tissue structures, and its ability to discover region-specific markers by combining with SODB interactive features.

Characterizing tissue structures obscured in other analyses

The previous section introduced the strengths of SOView for visualization purposes compared with other methods. This section describes the even more powerful ability of SOView by combining with SODB interactive functions, which is one of the key features of our work. To this end, we used a more complex dataset to show that SOView can identify some important tissue structures that other methods cannot, and to interactively find the companion markers of these structures.

We used a sagittal mouse brain posterior dataset generated by 10X Visium spatial transcriptomics technology[19]. One can freely access the data by visiting SODB (https://gene.ai.tencent.com/SpatialOmics/dataset?datasetID=78), selecting 'V1_Mouse_Brain_Sagittal_Posterior_filtered_feature_bc_matrix' in the data selector drop-down menu (Supplementary Fig. 18), then clicking 'SOView' (Supplementary Fig. 19a) to explore the global structure of the tissue (Supplementary Fig. 19a, black arrow). Current methods for identifying tissue structures for spatial transcriptomics data include three categories: (1) clustering using gene expression profiles alone, the representative method is Louvain[66]; (2) clustering using both gene expression profiles and spatial location, the representative method is BayesSpace[84] and (3) clustering using gene expression profiles, spatial location and histological image, the representative method is SpaGCN[85]. By referring to the Allen brain map[86] (Fig. 5a) and the paired histological images (Fig. 5b,c), we next compared these methods, that is, Louvain (Fig. 5d), BayesSpace (Fig. 5e), SpaGCN (Fig. 5f) and SOView (Fig. 5g), for their ability to identify different tissue structures.

First, we focus on the cerebellum region (Fig. 5c, red dashed line, and Supplementary Fig. 20a), and we can see that all methods show similar results (Fig. 5d–g), clearly distinguishing the molecular layer and granular layer of the hemispheric region, and the fiber tracts. For the brain stem region (Fig. 5c, orange dashed line, and Supplementary Fig. 20b), the Louvain results seemed more scattered (Fig. 5d), since the spatial location information was not considered when clustering. In contrast, the methods that utilized spatial information, BayesSpace (Fig. 5e) and SpaGCN (Fig. 5f) could better recover the spatially coherent tissue domain, and SOView (Fig. 5g) visualization was comparable with BayesSpace and SpaGCN, even without using either spatial or histological information. For the isocortex (Fig. 5c, dark blue dashed line, and Supplementary Fig. 20c), BayesSpace (Fig. 5e) distinguished the main region from other regions but failed to identify subregions (for example, cortex layers). Louvain (Fig. 5d) differentiated cortex Layer1 from the other cortex layers but failed to distinguish Layer1 from a hippocampal formation (HPF) region. SpaGCN (Fig. 5f) mistakenly mixed Layer2–Layer6 of the isocortex with CA1 of the HPF region. In contrast, SOView (Fig. 5g) not only differentiated the isocortex from the other regions using different colors, but also revealed a gradient trend from Layer1 to Layer6, consistent with all the three cortex replicates in Fig. 4e, as well as with the previous report[69,83].

Structure characterization is even more challenging when considering the hippocampal formation (HPF) region (Fig. 5c, green dashed line, and Supplementary Fig. 20d). For HPF1 (Fig. 5c), Louvain (Fig. 5d) could distinguish the '(' shaped Ammon's horn (CA) region from the ')' shaped dentate gyrus (DG) region, but failed to differentiate CA from an isocortex region. BayesSpace (Fig. 5e) mixed CA and DG together, possibly due to over weighting of the spatial information. Although SpaGCN (Fig. 5f) successfully distinguished CA and DG, it mixed CA with a part of the isocortex, and also failed to identify subregions of CA and DG. In contrast, SOView not only distinguished CA and DG by different colors, but also identified the heterogeneity inside CA and DG with slight color variations.

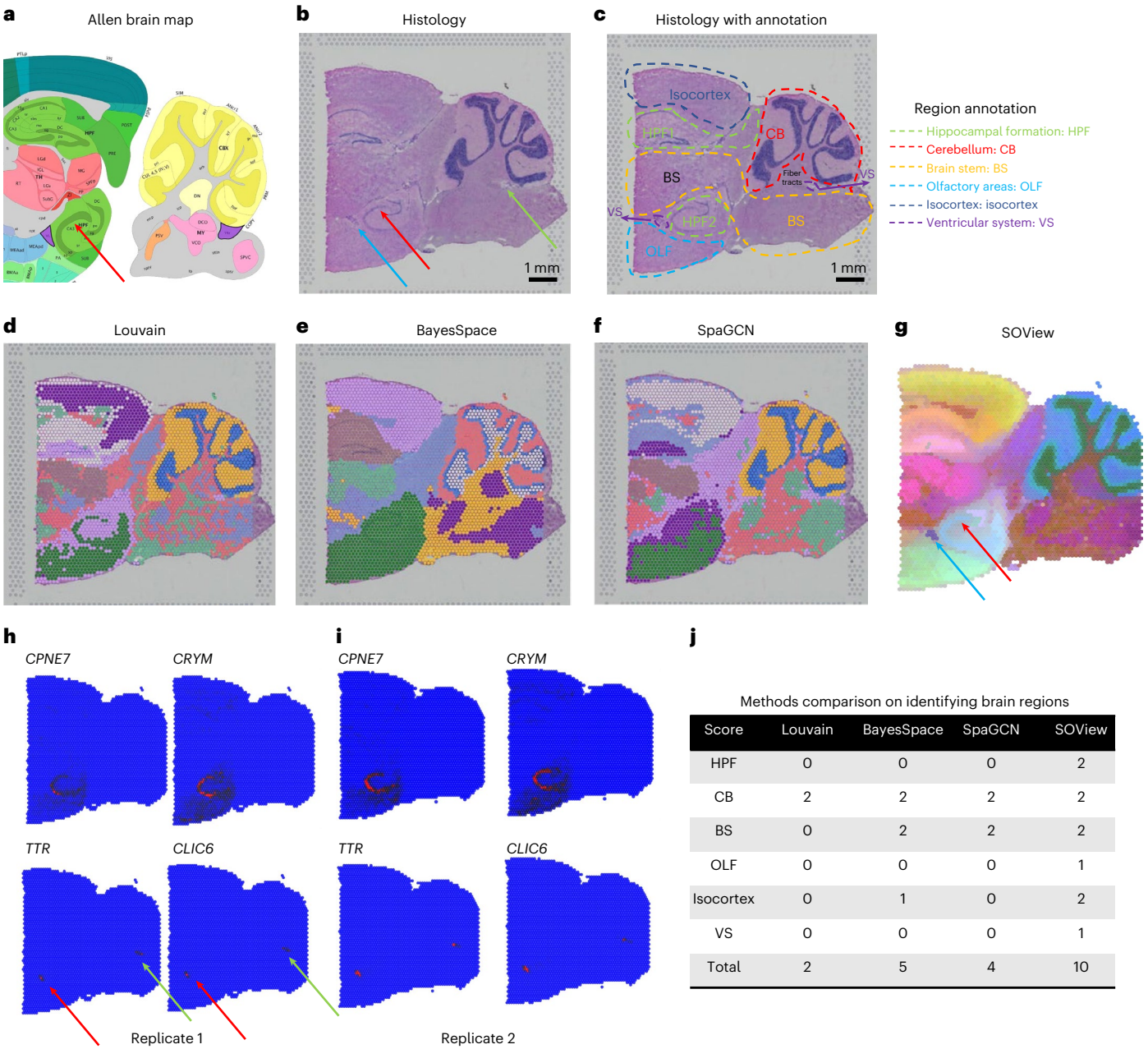
For HPF2 (Fig. 5c), Louvain (Fig. 5d) was able to outline the '(' shaped pyramidal layer of CA3 (CA3sp; Fig. 5b, red arrow), but mistakenly mixed it with olfactory areas (OLF) (Fig. 5c, blue dashed line). Neither BayesSpace (Fig. 5e) nor SpaGCN (Fig. 5f) could distinguish between OLF and HPF2, or among the CA3sp subregions in HPF2 (Fig. 5b, red arrow). In contrast, SOView was the only method that successfully delineated CA3sp (Fig. 5b,g, red arrows), which could be verified in both the paired histological image (Fig. 5b) and the Allen brain map (Fig. 5a, red arrow). We next asked whether this region is characterized by corresponding marker genes. Since the other methods could not identify this region (CA3sp; Fig. 5b, red arrow), naturally the marker genes of this region might not be identifiable either. However, by combining the interactive function of SODB, SOView enabled the identification of such subtle regions and the corresponding marker genes. Users could just use either the box selector (Supplementary Fig. 19a, red arrow) or the polygon selector (Supplementary Fig. 19a, yellow arrow) to select the ROI in the SOView display panel, then click the 'Analysis' button (Supplementary Fig. 19a, blue arrow). The marker genes of the selected ROI will be automatically presented in the web page (Supplementary Fig. 19b). We used this method to analyze the marker genes of CA3sp in HPF2 (Supplementary Fig. 19b), and found that they were specifically highly enriched in the CA3sp region (Fig. 5h, top row). We also verified the expression of these marker genes in a replicate experiment (Fig. 5i, top row).

We also found a small distinctive region in the SOView map (Fig. 5g, blue arrow). To dissect this region, we circled it on the SOView map (Supplementary Fig. 21a, red circle) and obtained its differentially expressed genes (Supplementary Fig. 21b and Fig. 5h, bottom row). We found that these marker genes were highly expressed not only in the ROI that we circled (Fig. 5h, bottom row, red arrow), but also in another small region (Fig. 5h, bottom row, green arrow). The marker genes showed consistent patterns in another replicate (Fig. 5i, bottom row). By referring to the histological image (Fig. 5b,c), one could find that these two small regions did have distinctive morphological features from the surrounding regions (Fig. 5b, blue and green arrows), and their relative positions on the Allen brain map corresponded to the ventricular system (VS) (Fig. 5a,c, purple region and purple dashed line). Note that both this region and the marker genes cannot be discovered by any other methods (Fig. 5d–f).

Furthermore, we scored different methods according to their abilities to characterize different regions of the brain (Methods). The results showed that SOView could exclusively identify three brain regions that could not be identified by other methods, and was better than the other methods in the overall score (Fig. 5j).

To comprehensively compare more methods on a wider range of parameter settings, we additionally compared a list of

https://doi.org/10.1038/s41592-023-01773-7



j

Methods comparison on identifying brain regions

Score	Louvain	BayesSpace	SpaGCN	SOView
HPF	0	0	0	2
CB	2	2	2	2
BS	0	2	2	2
OLF	0	0	0	1
Isocortex	0	1	0	2
VS	0	0	0	1
Total	2	5	4	10

Fig. 5 | SOView discovers both known and unknown functional tissue domains. a,**b**, Allen brain map reference (**a**) and histological image (**b**) of a mouse brain posterior part. **c**, Major brain regions are annotated by reference to **a** and Supplementary Fig 20. **d**–**g**, Result of different methods in delineating different brain structures: Louvain (**d**), BayesSpace (**e**), SpaGCN (**f**) and SOView (**g**). **h**, Top, top marker genes of CA3sp region, identified in SOView (**g**) red arrow. The top marker genes are obtained by SODB (Supplementary Fig. 19). **h**, Bottom, top marker genes of VS region, identified in SOView (**g**) blue arrow. The top marker genes are obtained by SODB (Supplementary Fig. 21). **i**, The same genes as in **h** show consistent patterns in another replicate. **j**, Comparison of different methods in the ability to distinguish different brain regions.

methods for identifying spatial domains (Methods: 'Method comparisons on identifying tissue domains' and Supplementary Table 7) under ten different parameter settings (Supplementary Table 8) for each method on two replicate samples. We show all the results in Supplementary Figs. 25–42. In these results, we not only showed each method output in its whole, but also showed the zoomed in regions that were specifically well recognized by SOView (that is, VS and HPF2, as shown in Fig. 5c,g). All these results suggest that no matter how the parameters were tuned in these methods, they could not find the functional tissue domains that were easily identified by SOView.

Advancing computational methods development

To show that SODB could facilitate development of computational methods (for example, provide datasets for reproducing and benchmarking existing methods and provide potential new datasets for novel methods development), we took the field of spatial transcriptomics as an example. For this, we collected mainstream computational methods covered in six recent review articles[12,20,87–90] (Supplementary Table 4), including a total of 68 methods in 11 types (Fig. 6a and Supplementary Table 5). We also summarized the datasets used by these methods, and matched them with datasets in SODB, to see how SODB might support these existing methods (Supplementary Table 6).

https://doi.org/10.1038/s41592-023-01773-7

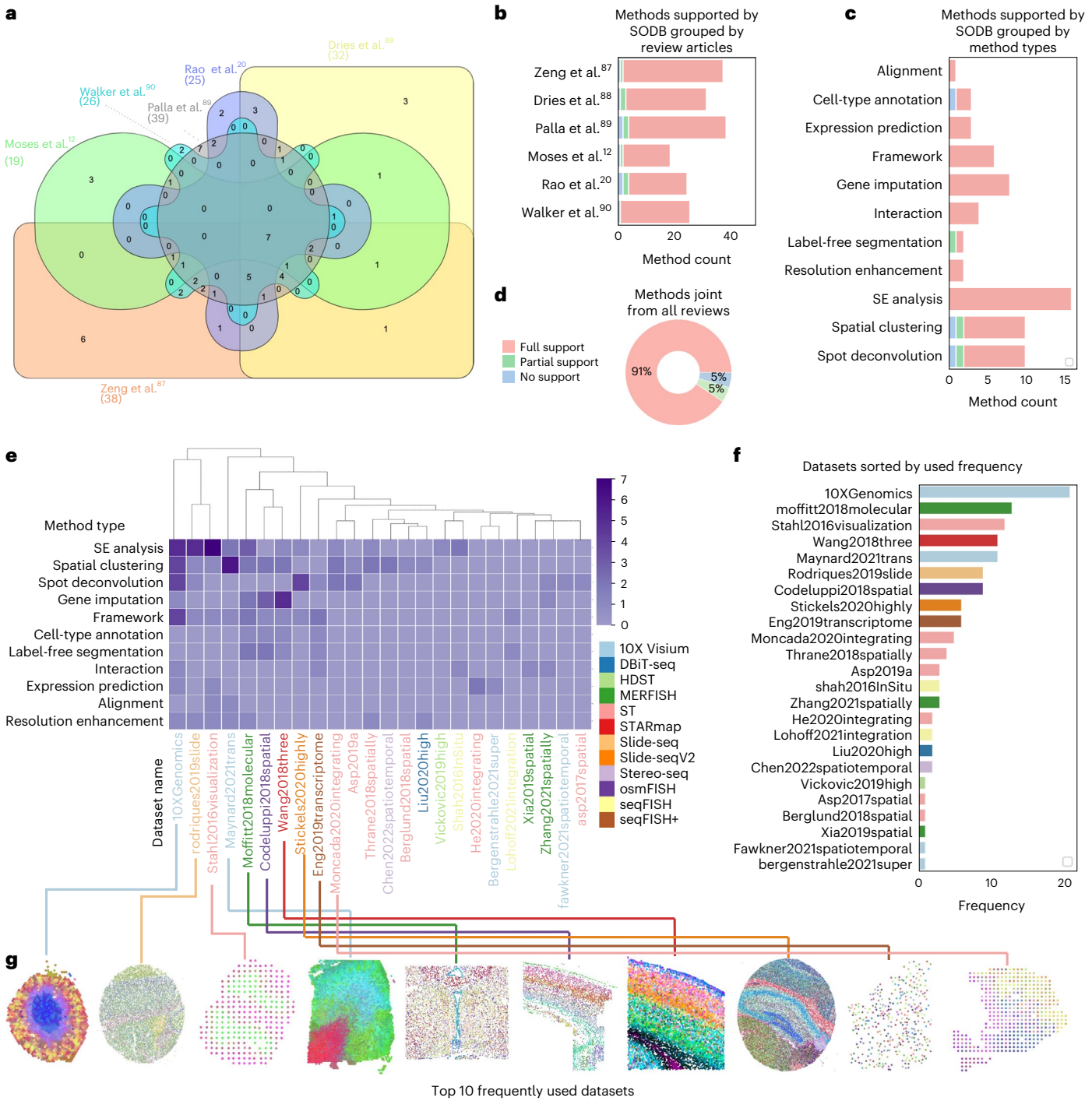


Fig. 6 | SODB advances computational methods development. a, Venn plot showing the computational methods mentioned in six high-impact review articles. **b**,**c**, Bar plot showing how many computational methods (grouped by review articles (**b**) or method types (**c**)) are fully/partially/not supported by SODB. **d**, Pie plot showing how many methods (joint from **a**) are fully/partially/not supported by SODB. **e**, Heat map showing the frequencies of each dataset used by each method type. The dataset name is colored by spatial technology. The first entry of the heat map (SE analysis, 10X Genomics) is 6, this means that there are six methods of SE analysis that used 10X Genomics dataset. **f**, Bar plot showing the top frequently used datasets. **g**, SOView visualization of top ten frequently used datasets. In each dataset, one experiment is used for visualization.

We find that SODB could fully support most (>90%) of the computational methods of the six review articles (Fig. 6b). By joining all covered methods together, SODB could fully support 91% of them (Fig. 6d). After grouping these methods by type, we found that SODB could fully support seven out of eleven types of methods, including alignment, expression prediction, framework, gene imputation, interaction, resolution enhancement and SE analysis (Fig. 6c).

We next summarized the frequencies of each dataset used by each type of method (Fig. 6e). These datasets were selected because they were used at least once by one of the 68 most popular methods, and the names of the datasets (colored by spatial biotechnologies) were its dataset ID in SODB (Methods). In the statistics, the Visium sample data provided by the 10X Genomics website was the most widely used dataset in four method types (Fig. 6e). Datasets

published with original ST (Stahl2016visualization)[1] and Slide-seq (rodriques2019slide)[2] papers were mainly used in SE analysis algorithms and used less in other method types (Fig. 6e). The DLPFC dataset (Maynard2021trans)[83], which contained 12 replicates of human brain cortex with well-annotated region labels was a widely used standard dataset in spatial clustering methods (Fig. 6e). The gene imputation methods mainly used imaging-based spatial transcriptomics datasets, such as MERFISH profiling on mouse hypothalamic preoptic region (Moffitt2018molecular)[71], osmFISH profiling on mouse somatosensory cortex (Codeluppi2018spatial)[15] and STARmap profiling on mouse visual cortex (Wang2018three)[17] (Fig. 6e). Novel methods developers could easily find the necessary datasets to be used for benchmarking according to the method types, and also explore new datasets for novel applications.

We also compiled the statistics (Fig. 6f) on the overall usage frequencies of SODB datasets (the datasets that were not used by any methods were excluded), and found that early datasets and well-organized datasets were more likely to be frequently used (see the top ten frequently used datasets in Fig. 6g). Some newly generated datasets, such as MERFISH profiling on mouse primary motor cortex (Zhang2021spatially)[69], seqFISH profiling on embryo development (Lohoff2021integration)[60] and Stereo-seq profiling on embryo development (Chen2022spatiotemporal)[23], were not widely used, but their high data quality (for example, large field-of-view, high mRNA capture ratio and high throughput) could make them potentially more popular in the future. The generation of these new datasets would also stimulate new algorithm developments. For example, more MERFISH datasets in diverse brain tissues (for example, Fang2022conservation[80] and chen-2021decoding[72]) would suggest integrating cell morphological features with gene expression profiles to achieve more comprehensive cell identity learning. More 3D spatial datasets (for example, Wang2021easi[91] and kuett2021three[52]) would call for new 3D analytical methods. Spatial datasets with large FOV (for example, Srivatsan2021embryo[55] and Chen2022spatiotemporal[23]) could facilitate high-quality data registration and cell segmentation method development.

Discussion

We present SODB, a web-based platform combining large-scale data deposition and interactive data exploration for general spatial omics data. SODB presents various types of spatial omics datasets (for example, spatial transcriptomics, proteomics, metabolomics, genomics and multi-omics) with a downloadable and unified data format, which could be directly fed into many mainstream analytical packages. Apart from data, SODB also provides a suite of interactive data exploration modules. Among these, SOView is a key feature of SODB, which can be used to visualize the global tissue structure, and identify some subtle but important local or subtissue structures. Combing SOView and the interactive interface of SODB, one can characterize user-defined ROIs with automatically generated marker genes. SODB could also fuel the future development of various spatial omics computational methods.

We anticipate some potential future improvements, and we welcome user feedback. For example, due to the heterogeneous pipelines for processing different data formats, users are currently not allowed to upload their own datasets to SODB, and data submission should be accomplished by contacting the corresponding authors via email. Our group will process and update the database biweekly. At present, SODB can explore data consisting of up to 10^6 spots (Supplementary Fig. 22), the scalability to larger scale data needs to be further optimized in the future.

As one of the most watched technologies in recent years[31], the spatial omics community will contribute more novel biotechnologies and new datasets in the future. Integrating comprehensive spatial omics data with interactive analytical modules, SODB will greatly assist its users in gaining more functional insights by providing a multifaceted view of tissue-level molecular profiles and biological pathways.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01773-7.

References

1. Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
2. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463 (2019).
3. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. Y. & Zhuang, X. W. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* https://doi.org/10.1126/science.aaa6090 (2015).
4. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).
5. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
6. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981.e15 (2018).
7. Sun, C. et al. Spatially resolved metabolomics to discover tumor-associated metabolic alterations. *Proc. Natl Acad. Sci. USA* **116**, 52–57 (2019).
8. Rappez, L. et al. SpaceM reveals metabolic states of single cells. *Nat. Methods* **18**, 799–805 (2021).
9. Passarelli, M. K. et al. The 3D OrbiSIMS-label-free metabolic imaging with subcellular lateral resolution and high mass-resolving power. *Nat. Methods* **14**, 1175–1183 (2017).
10. Zhao, T. et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).
11. Marx, V. Method of the year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
12. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
13. Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nat. Rev. Genet.* **23**, 741–759 (2022).
14. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
15. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
16. Shah, S. et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* **174**, 363–376.e16 (2018).
17. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* https://doi.org/10.1126/science.aat5691 (2018).
18. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2020).
19. Gracia Villacampa, E. et al. Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genomics* https://doi.org/10.1016/j.xgen.2021.100065 (2021).
20. Rao, A., Barkley, D., Franca, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
21. Lewis, S. M. et al. Spatial omics and multiplexed imaging to explore cancer biology. *Nat. Methods* **18**, 997–1012 (2021).
22. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**, 987–990 (2019).
23. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).

24. Hickey, J. W. et al. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nat. Methods* **19**, 284–295 (2021).
25. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).
26. Lin, J.-R. et al. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife* **7**, e31657 (2018).
27. Gut, G., Herrmann, M. D. & Pelkmans, L. Multiplexed protein maps link subcellular organization to cellular states. *Science* https://doi.org/10.1126/science.aar7042 (2018).
28. Keren, L. et al. MIBI-TOF: a multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci. Adv.* https://doi.org/10.1126/sciadv.aax5851 (2019).
29. Damond, N. et al. A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metab.* **29**, 755–768.e55 (2019).
30. Yuan, Z. et al. SEAM is a spatial single nuclear metabolomics method for dissecting tissue microenvironment. *Nat. Methods* **18**, 1223–1232 (2021).
31. Eisenstein, M. Seven technologies to watch in 2022. *Nature* **601**, 658–661 (2022).
32. Liu, Y. et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665–1681 (2020).
33. Fan, R. et al. Spatial-CITE-seq: spatially resolved high-plex protein and whole transcriptome co-mapping. Preprint at *Res. Sq.* https://doi.org/10.21203/rs.3.rs-1499315/v1 (2022).
34. Merritt, C. R. et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat. Biotechnol.* **38**, 586–599 (2020).
35. Vickovic, S. et al. SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nat. Commun.* https://doi.org/10.1038/s41467-022-28445-y (2022).
36. Fan, R. et al. Spatially resolved epigenome-transcriptome co-profiling of mammalian tissues at the cellular level. Prerpint at *Res. Sq*. https://doi.org/10.21203/rs.3.rs-1728747/v1 (2022).
37. Chung, H. et al. Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* **18**, 1204–1212 (2021).
38. Chen, W. T. et al. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991.e19 (2020).
39. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89–93 (2019).
40. Marshall, J. L. et al. High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *iScience* **25**, 104097 (2022).
41. Chen, H. et al. Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell Rep.* **37**, 109915 (2021).
42. Ji, A. L. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**, 497–514.e22 (2020).
43. Berglund, E. et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 2419 (2018).
44. Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I. & White, R. M. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat. Commun.* **12**, 6278 (2021).
45. Keren, L. et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**, 1373–1387 (2018).
46. Wu, R. et al. Comprehensive analysis of spatial architecture in primary liver cancer. *Sci. Adv.* **7**, eabg3750 (2021).
47. Hartmann, F. J. et al. Single-cell metabolic profiling of human cytotoxic T cells. *Nat. Biotechnol.* **39**, 186–197 (2020).
48. Risom, T. et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* **185**, 299–310.e18 (2022).
49. Danenberg, E. et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nat. Genet.* **54**, 660–669 (2022).
50. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
51. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
52. Kuett, L. et al. Three-dimensional imaging mass cytometry for highly multiplexed molecular and cellular mapping of tissues and the tumor microenvironment. *Nat. Cancer* **3**, 122–133 (2021).
53. Hildebrandt, F. et al. Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nat. Commun.* **12**, 7046 (2021).
54. Cho, C. S. et al. Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* **184**, 3559–3572.e22 (2021).
55. Srivatsan, S. R. et al. Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111–117 (2021).
56. Goh, J. J. L. et al. Highly specific multiplexed RNA imaging in tissues with split-FISH. *Nat. Methods* **17**, 689–693 (2020).
57. Mantri, M. et al. Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. *Nat. Commun.* **12**, 1771 (2021).
58. Asp, M. et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* **179**, 1647–1660.e19 (2019).
59. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–826.e23 (2021).
60. Lohoff, T. et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.* **40**, 74–85 (2021).
61. Wang, M. et al. High-resolution 3D spatiotemporal transcriptomic maps of developing *Drosophila* embryos and larvae. *Dev. Cell* **57**, 1271–1283 (2022).
62. Liu, C. et al. Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Dev. Cell* **57**, 1284–1298 (2022).
63. Fan, Z., Chen, R. & Chen, X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Res.* **48**, D233–D237 (2019).
64. Li, Y. et al. SOAR: a spatial transcriptomics analysis resource to model spatial variability and cell type interactions. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.17.488596 (2022).
65. Xu, Z. et al. STOmicsDB: a database of spatial transcriptomic data. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.11.481421 (2022).
66. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
67. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
68. Eng, C. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
69. Zhang, M. et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).
70. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
71. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* https://doi.org/10.1126/science.aau5324 (2018).
72. Chen, R. et al. Decoding molecular and cellular heterogeneity of mouse nucleus accumbens. *Nat. Neurosci.* **24**, 1757–1771 (2021).

73. Ortiz, C. et al. Molecular atlas of the adult mouse brain. *Sci. Adv.* **6**, eabb3446 (2020).
74. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
75. Risom, T. et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* **185**, 299–310 (2022).
76. Zhang, R. et al. Spatial transcriptome unveils a discontinuous inflammatory pattern in proficient mismatch repair colorectal adenocarcinoma. *Fundamental Res*. https://doi.org/10.1016/j.fmre.2022.01.036 (2022).
77. Taylor, M. J., Lukowski, J. K. & Anderton, C. R. Spatially resolved mass spectrometry at the single cell: recent innovations in proteomics and metabolomics. *J. Am. Soc. Mass Spectrom.* **32**, 872–894 (2021).
78. Palmer, A. et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Methods* **14**, 57–60 (2017).
79. Abdelmoula, W. M. et al. Peak learning of mass spectrometry imaging data using artificial neural networks. *Nat. Commun.* https://doi.org/10.1038/s41467-021-25744-8 (2021).
80. Fang, R. et al. Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. *Science* **377**, 56–62 (2022).
81. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
82. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
83. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
84. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).
85. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
86. Wang, Q. et al. The Allen mouse brain common coordinate framework: a 3D reference atlas. *Cell* **181**, 936–953.e20 (2020).
87. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol.* **23**, 83 (2022).
88. Dries, R. et al. Advances in spatial transcriptomic data analysis. *Genome Res.* **31**, 1706–1718 (2021).
89. Palla, G., Fischer, D. S., Regev, A. & Theis, F. J. Spatial components of molecular tissue biology. *Nat. Biotechnol.* **40**, 308–318 (2022).
90. Walker, B. L., Cang, Z., Ren, H., Bourgain-Chang, E. & Nie, Q. Deciphering tissue structure and function using spatial transcriptomics. *Commun. Biol.* **5**, 220 (2022).
91. Wang, Y. et al. EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell* **184**, 6361–6377.e24 (2021).

Methods

Data collection and processing

We collected datasets according to the 'data availability' statement of each original manuscript. The provided data links (Supplementary Table 2, 'access' column) were typically referred to public data deployment platforms (Supplementary Fig. 1). For the sake of different technologies and different laboratories, the raw format of generated data is highly diverse. In general, for meshed data, such as 10X Visium[83] and ST[1], the spatial information was obtained using the coordinate of each spot; for subcellular resolution data, the spatial information was obtained using the center position of each segmentation cell; and for imaging data whose raw data is presented by multi-channel images, the spatial information was obtained using the relative positions of pixels. Due to the heterogeneous data format, which cannot be processed by a unified pipeline, we designed different scripts to manually process them into Anndata[66] format. The obtained spatial information was written into Anndata.obsm['spatial'], and other properties associated with spots (for example, spot annotations, region labels or cell sizes) were written into Anndata.obs.

For every dataset we provided a clustering result (Anndata.obs['leiden']) for users' convenience. We used the standard pipeline provided by Scanpy[66]. Specifically, we used the total counts to normalize the raw count values, followed by log transformation. For data with feature dimensions >2,000, we selected the top 2,000 highly variable genes using highly_variable_genes(adata, flavor = 'seurat', n_top_genes = 2000). Then principal component analysis (PCA), neighbors and Leiden algorithms were run with default parameters. We admit that different applications and data need different clustering algorithms with different parameters to be tuned. The clustering results with this unified procedure were aimed at providing a reference for users. One can download the Anndata-formatted data and process it using the customized pipelines. It is worth noting that the downloaded data from SODB remain as the full gene set, rather than just highly variable genes, which is ready to interact with downstream pipelines.

SOView

The visualization and analysis of high-plex spatial omics data is challenging. For a typical spatial transcriptomics data (such as 10X Visium), there are more than 20,000 genes profiled on several thousands of spots. One cannot have a comprehensive understanding of the assayed tissue by visually inspecting each gene's spatial expression. To alleviate this problem, we propose SOView, an interactive visualization method for efficient spatial omics data exploration. SOView does not need any parameter-tuning steps, and all the datasets were run with consistent settings. This is beneficial for batch processing of a large number of datasets. The main idea of SOView is to utilize the interactive capabilities of the SODB website to visualize and explore the spatially resolved molecular landscape of the target tissue by combining manifold learning and visually understandable color coding. In the following, we explain the internal happenings of SOView at three time points: (1) before data is updated to SODB, (2) when the user visualizes data with SOView and (3) when the user interacts with SOView.

Before data is updated to SODB. This step happens after the data is processed into Anndata format and before input into SODB. Without loss of generality, suppose this data consists of one gene expression matrix of n spots × m genes (stored in Anndata.X), and one spatial coordinate matrix (SCM) of n spots × 2 (stored in Anndata.obsm['spatial']). First, the gene expression matrix is reduced to $p = \min(50, m-1)$ dimensions by PCA to form one n spots × p principal components matrix. Next, the connectivity of each pair of spots is estimated by an efficient neighborhood search[92], to generate a sparse neighborhood graph of spots with size $n \times n$. Then, this neighborhood graph is input into a uniform manifold approximation and projection (UMAP) algorithm[93] to obtain a dimensionality reduced matrix with $n \times 3$ (stored in Anndata.obsm['X_umap']).

When the user visualizes data with SOView. For the display panel of SOView (Supplementary Fig. 19a, black arrow), we exploited the open-source JavaScript visualization library Echarts in the front-end of the SODB website. When the user needs to visualize data with a specified 'data id' and properties with SOView, the front-end code first sends a request using HTTP-POST function to the back-end Python code, with the 'data id' and three property names of spots (by default, three UMAP components, that is, X_umap@0, X_umap@1 and X_umap@2; Supplementary Fig. 13a) as parameters. After receiving the POST request, the back-end Python code accesses the location where the specified Anndata data are stored on the back-end server according to this 'data id', followed by reading the Anndata file (stored as h5ad file) into memory. The back-end code reads the three-dimensional dimensionality reduced matrix via Anndata.obsm['X_umap'] and rescales it to 0–255 by each column to form an RGB color matrix (RCM). Then RCM and SCM (stored in Anndata.obsm['spatial']) are transmitted to the front end through the HTTP-POST function. The front-end JavaScript code passes the two matrices to Echarts, and Echarts plots an interactive color map based on the spatial position and color of each spot. In this way, the similarities in color between spots reflect the similarities in gene expression between spots, and users can visualize the tissue heterogeneity through a single plot.

Note that the three properties to be encoded in colors could be any other spot-level features, such as PCA components, t-SNE components and gene expression values. SOView provides an operation panel (Supplementary Fig. 13a,b) for users to customize the features to be encoded in RGB colors. The default option is the first three components of UMAP (that is, X_umap@0, X_umap@1 and X_umap@2; Supplementary Fig. 13a). One can freely change these options using the drop-down menus (Supplementary Fig. 13a).

When the user interacts with SOView. Some basic user interactions include mouse-hovering, zoom-in, zoom-out and region selection on SOView plot (Supplementary Fig. 19a, black arrow). Another important interaction function is to detect the marker genes of user-selected ROIs. This function is warranted because clustering-based methods cannot identify some important but subtle tissue domains (examples can be found in Fig. 5), and the markers of these domains would be missed. The powerful functions of SOView can help users to locate these tissue domains, and then the marker gene in this domain will be detected. Specifically, to detect the marker genes (or other molecular features) of user-selected ROIs, there were three steps: (1) the front-end code passed the index list of the user-selected spots to the back-end code; (2) the spots in the ROI were compared with other spots in the tissue using t-test comparison between the two groups and (3) the genes (or other molecular features) were ranked by the score output by scanpy.tl.rank_genes_groups (https://scanpy.readthedocs.io/en/stable/generated/scanpy.tl.rank_genes_groups.html#scanpy.tl.rank_genes_groups) and displayed on the display panel of SOView (Supplementary Fig. 19b).

SOView is designed to maximally serve the SODB database by lifting the heavy computational burdens from users in the data preparation stage, so that users do not feel the computational pressure when exploring and interacting with data, even for large-scale data up to more than 10^6 spots.

Database comparison

According to the STomicsDB[65] website (https://db.cngb.org/stomics/; Supplementary Fig. 5a), the number of spots is 754,063 (*Homo sapiens*), 3,222,307 (*Mus musculus*), 95,749 (*Macaca fascicularis*),

160,258 (*Danio rerio*) and 155,684 (*Drosophila melanogaster*), the listed biotechnologies with spot-wise spatial coordinates are 10X Visium, ST, Stereo-seq, DBiT-seq, HDST, MERFISH, seqScope, DSP and sciSpace (Supplementary Fig. 5c).

According to the SODB website, the numbers of biotechnologies are 14 (spatial transcriptomics), six (SRHP), three (spatial metabolomics), one (spatial genomics) and two (spatial multi-omics), respectively. So the total number of biotechnologies in SODB is 26 (Supplementary Fig. 5b).

According to SOAR's manuscript[64] (https://doi.org/10.1101/2022.04.17.488596), there are a total of 1,633 samples (experiments). The SOAR website (Supplementary Fig. 5d) shows that SOAR covers eight different spatial technologies (10X, DBiT-seq, MERFISH, osmFISH, seqFISH, seqFISH+, Slide-seq and ST).

Command-line package (pysodb)

Besides the GUI-based data access in the SODB website, SODB provides another command-line package (pysodb) to access the data for computational groups. Specifically, pysodb contains the following functions:

- Function 1: Pysodb.list_dataset()
  This function returns a list containing the names of all datasets.
- Function 2: Pysodb.list_dataset_by_biotech(biotech_name)
  This function takes 'biotech_name' as the input (for example, 10X Visium, slide-seq, MIBI or CODEX), and outputs a list that contains the names of datasets in that biotech.
- Function 3: Pysodb.list_biotech(biotech_category)
  This function takes 'biotech_category' as input (for example, spatial transcriptomics/spatial proteomics/spatial metabolomics/spatial genomics/spatial multi-omics), and outputs a list that contains the names of possible biotechnologies belonging to the category. For example, if the input is 'spatial transcriptomics', the expected output is 10X Visium, Slide-seq, MERFISH, osmFISH, seqFISH, seqFISH+, seqScope, STARmap, EASI-FISH, Slide-seqV2, HDST, ST, Stereo-seq and sciSpace.
- Function 4: Pysodb.load_dataset(dataset_name)
  This function takes 'dataset_name' as the input (such as 'wang2021easi', please refer to Methods: 'Names of datasets', and 'Short name' in the dataset detail page, for example, https://gene.ai.tencent.com/SpatialOmics/dataset?datasetID=41), and returns Python dict object. The keys of the dict are the names of the experiments within the dataset, and the corresponding values are the Anndata objects.
- Function 5: Pysodb.
  load_experiment(dataset_name,experiment_name)

This function takes two parameters as input: 'dataset_name' and 'experiment_name'. Since function 4 might take a long time if input a dataset_name that contains a large number of experiments. Function 5 is designed to eliminate this problem by only downloading and loading one experiment of a dataset. This function returns a single Anndata object.

Note that if one dataset was previously loaded by pysodb, it would be cached locally, so that loading that data is much more time efficient than loading it for the first time.

The code of pysodb is publicly available at https://github.com/TencentAILabHealthcare/pysodb. The document for the package is available at https://pysodb.readthedocs.io/en/latest/. The demonstration code to show how pysodb can interact with downstream pipelines is available at https://github.com/yuanzhiyuan/SODB_analysis/tree/master/Demonstration.

Data sparsity and SE analysis

The data sparsity of each experiment was measured by computing the percentage of zeroes entries in the expression matrix[94]. The spatially variable (SE) analysis was performed by Moran's *I*[67].

Names of datasets

In SODB, the dataset identities were named as 'ABC', where A is the last name of the first author, B is the published year and C is the first word of the paper/project title.

Platform implementation

SODB provides data and analytical tools online through the website. The front end is developed based on Vue.js (v.3.2.13) and Element Plus (v.2.0.5), where Vue.js is a popular progressive JavaScript framework for single page applications and Element Plus is a Vue3-based component library. We construct data visualization and analytical tools with the open-source JavaScript visualization library Echarts (v.5.3.2). The back end is built with Python (v.3.8.13) and Flask (v.2.1.2). Flask is a lightweight WSGI web application framework. We utilize SQLite3 (v.2.6.0) to store metadata. Nginx (v.1.20.1) is used to reverse proxy.

Method comparisons on identifying tissue domains

To benchmark the performance of identifying tissue domains, we compared ten computational methods with SOView. The ten methods include traditional nonspatial clustering methods (Louvain[66]), spatial clustering methods with (SpaGCN[85]) or without (BayesSpace[84], CCST_leiden[95], CCST_louvain[95], conST[96], SCAN-IT[97], SEDR[98], SpaceFlow[99] and STAGATE[100]) the integration of hematoxylin and eosin images. All the methods information is summarized in Supplementary Table 7. We used the replicate samples provided in https://support.10xgenomics.com/spatial-gene-expression/datasets. These samples can be downloaded by the visium_sge function provided in SCANPY, with accession code 'V1_Mouse_Brain_Sagittal_Posterior' and 'V1_Mouse_Brain_Sagittal_Posterior_Section_2'. These datasets can also be explored and downloaded by SODB, via the GUI link https://gene.ai.tencent.com/SpatialOmics/dataset?datasetID=78, or by the command-line pysodb.

We ran all the compared methods according to the tutorials provided on their websites (Supplementary Table 7). We tested each method on each dataset, with a range of parameters to fully test their abilities to identify spatial domains at different cluster granularity. We have provided the parameter settings, figures and reproducible codes in Supplementary Table 8.

Methods comparison on identifying brain regions

The comparison shown in Fig. 5j proceeds as follows: method A on region I, if A cannot exclusively distinguish region I from other regions, then the score for method A in region I is 0. If A can exclusively distinguish region I from other regions, but A cannot identify at least one subregion of region I, then the score for method A in region I is 1. If A can exclusively distinguish region I from other regions, and also can identify at least one subregion of region I, then the score for method A in region I is 2.

For example, the gray spots in the Louvain result (Fig. 5d) are in both HPF1 and isocortex regions (Fig. 5c), which means Louvain could not exclusively distinguish HPF1 and isocortex, so the scores of Louvain in HPF1 and isocortex are 0. BayesSpace can distinguish isocortex region from other regions (Fig. 5c,e), but it cannot identify subregions within isocortex, so the score of BayesSpace in isocortex is 1. SOView can not only distinguish HPF region from other regions in color, but also identify subregions of HPF (for example, CA and DG), so the score for SOView in HPF is 2.

SODB can help with the development of spatial clustering algorithms

One necessary part of developing a spatial clustering algorithm is to quantitatively evaluate the algorithm performance on the data with spatial domain ground truth and compare with existing algorithms. SODB provides various such data, such as maynard2021trans[83], codelupi2018spatial[15], Wang2018Three_1k[17], to name a few. Many existing algorithms for spatial clustering have already used some datasets (they are also provided in SODB); please refer to Supplementary Table 5 for details.

Method comparisons on data loading

We highlighted the importance of pysodb in improving data access for biologists and bioinformaticians. In detail, we recorded the peak memory and time cost of loading spatial omics datasets. We loaded data in three ways: (1) load from raw data provided by the original paper, in which data are typically presented as multiple CSV formatted files; (2) load using pysodb's load_experiment function and (3) load using pysodb's load experiment function (preload). The difference between methods (2) and (3) is that the former loads the data the first time at the machine, and the latter loads data that was previously loaded at the machine (which means that the data was cached locally). The datasets we compare are Slide-seqV2 (Supplementary Table 10). We use the Python 'time' function to record the time cost, and the Python "tracemalloc' function to record the peak memory. For code availability, see Supplementary Table 9.

Updating and maintenance

We are a multi-institute research team from Fudan University, Tsinghua University and Tencent AI Laboratory, specializing in spatial omics research. The server clusters are stably maintained by the Tencent Cloud, which is one of the most robust cloud services in China. We have also backed up the docker image in different sites (Fudan and Tsinghua) for data redundancy, so that the database could be resumed when necessary.

The regular data update will be biweekly, the data collection people will search a set of keywords (spatial transcriptomics, spatially resolved transcriptomics, Visium, Slide-seq, Slide-seqV2, spatial proteomics, MIBI and other technology names involved in SODB) to find related publications over the preceding two weeks. We also subscribe to Google alerts associated with these keywords. Then, the data will be downloaded and preprocessed by the customized code for specific technologies, followed by updating to the database.

We have updated the data curation code to Github for the SODB project. Interested researchers will be able to propose new data/function suggestions via the Github Issue or email to the corresponding author (we have provided the emails for our team on the SODB website).

Statistics

All box plots in the manuscript share the same settings: the lower and upper hinges show the first and third quartiles (the 25th and 75th percentiles); the center lines correspond to the median; the upper whisker extends from the upper hinge to the largest value, which should be less than 1.5× the interquartile range (or distance between the first and third quartiles); and the lower whisker extends from the lower hinge to the smallest value, which is at most 1.5× the interquartile range. Data beyond the end of the whiskers are 'outlying' points and are plotted individually.

Visualization was performed by matplotlib (https://matplotlib.org/) and seaborn (https://seaborn.pydata.org/), statistical analysis was performed by scipy, numpy and scikit-learn[82]. Omics data analysis was performed by SCANPY[66] and Squidpy[67].

Reporting summary

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

Data availability

All the primary links of raw data are provided on the web page of datasets. All processed data can be downloaded via the SODB website (https://gene.ai.tencent.com/SpatialOmics/) or pysodb package (https://pysodb.readthedocs.io/en/latest/).

Code availability

The SODB website is available at https://gene.ai.tencent.com/SpatialOmics/. Code for the SODB project is available at https://github.com/yuanzhiyuan/SODB_analysis/. Code for pysodb is available at https://github.com/TencentAILabHealthcare/pysodb. Please refer to Supplementary Table 9 for detailed information on code and resources.

References

92. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1802.03426 (2018).
93. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
94. Li, B. et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).
95. Li, J., Chen, S., Pan, X., Yuan, Y. & Shen, H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Nat. Comput. Sci.* **2**, 399–408 (2022).
96. Zong, Y. et al. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.01.14.476408 (2022).
97. Cang, Z., Ning, X., Nie, A., Xu, M. & Zhang, J. SCAN-IT: domain segmentation of spatial transcriptomics images by graph neural network. In *Proc. 32nd British Machine Vision Conference* 22–25 November (2021).
98. Fu, H. et al. Unsupervised spatial embedded deep representation of spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.15.448542 (2021)..
99. Ren, H., Walker, B. L., Cang, Z. & Nie, Q. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nat. Commun.* **13**, 4076 (2022).
100. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).

Acknowledgements

Z.Y. acknowledges the support from the Shanghai Municipal Science and Technology Major Project (no. 2018SHZDZX01), ZJ Laboratory, Shanghai Center for Brain Science and Brain-Inspired Technology and 111 Project (no. B18015). M.Q.Z. acknowledges support by the Cecil H. and Ida Green Distinguished Chair. We thank L. Wang of Tencent for technical support.

Author contributions

J.Y., Z.Y. and M.Q.Z. designed the project. Z.Y. performed data collection. Website design was by Z.Y. and X.Z. J.Y., X.L. and Y.Z. provided technical support. Biological interpretation was by M.Q.Z. and Y.Z. Data statistics were performed by Z.Y. Website implementation was by X.Z. and W.P. Figure generation was by Z.Y. and F.Z. Z.Y. and W.P. wrote the manuscript. Z.X. maintains the website. J.Y. and M.Q.Z. reviewed the manuscript. All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01773-7.

Correspondence and requests for materials should be addressed to Zhiyuan Yuan, Michael Q. Zhang or Jianhua Yao.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contributions to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

nature research

Corresponding author(s): Michael Q. Zhang

Last updated by author(s): Aug 13, 2022

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	scanpy=1.9.1, squidpy=1.1.2, numpy=1.23.3, scipy=1.9.1, scikit-learn=1.1.2
Data analysis	These raw data were processed using Python. The front-end is developed based on Vue.js (version 3.2.13) and Element Plus (version 2.0.5), where Vue.js is a popular progressive JavaScript framework for single page applications and Element Plus is a Vue3 based component library. We construct data visualization and analytical tools with the open source JavaScript visualization library Echarts (version 5.3.2). The backend is built with Python (version 3.8.13) and Flask (version 2.1.2). Flask is a lightweight WSGI web application framework. We utilize SQLite3 (version 2.6.0) to store metadata. Nginx (version 1.20.1) is used to reverse proxy.

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

Data

Policy information about availability of data

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

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

This is a database paper, containing more than 2000 datasets. It is impossible to list all the links here. Please find the full link list in Supplementary Table 1~2. Also, all the primary links of raw data are also provided on the web page of datasets (e.g., https://gene.ai.tencent.com/SpatialOmics/dataset?datasetID=3). All processed data can be downloaded via SODB website (https://gene.ai.tencent.com/SpatialOmics/) or pysodb package (https://pysodb.readthedocs.io/en/latest/).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	All the data used in this study are from publicly available sources, and no original data was collected in this study.
Data exclusions	None
Replication	In all computational analysis, all replicated experiments (at least 5 times) were successful.
Randomization	The order of samples and cells in our method is inherently arbitrary and we have therefore not incorporated randomization in our method or results.
Blinding	We collect all the public data and know their tissue origin in order to annotate these information in our website.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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