Spatial transcriptomics resources to model spatial variability and cell-cell interactions

Yiming Li[1,#], Saya Dennis[1,#], Meghan R. Hutch[1,#], Yawei Li[1], Mia S. Broad[1], Zexian Zeng[2,3,]* and Yuan Luo[1,]*

[1]Department of Preventive Medicine, Northwestern University, USA
[2]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
[3]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

*Corresponding authors

[#]Co-first authors

Email addresses:

zexianzeng@pku.edu.cn (Zexian Zeng)

yuan.luo@northwestern.edu (Yuan Luo)

Abstract

Spatial transcriptomics allows researchers to investigate transcriptomes in the tissue context. Although large volumes of spatial transcriptomics data have been generated, a lack of systematic curation and analysis makes data reuse challenging. Herein, we present Spatial transcriptOmics Analysis Resource (SOAR), a resource with an extensive systematically curated collection of spatial transcriptomics data and multi-pronged analysis capability.

Main

Spatially resolved transcriptomics preserves the spatial organization of capture locations, which allows researchers to study tissue functions and disease pathology in the morphological context[1]. In recent years, different spatial transcriptomics technologies have been widely applied, leading to discoveries in cancer research, neuroscience, and developmental biology. To facilitate data exploration and meta-analysis in spatial transcriptomics research, we present Spatial transcriptOmics Analysis Resource (SOAR; https://soar.fsm.northwestern.edu/), a public database with an extensive collection of spatial transcriptomics data, analysis functions, and interactive visualization features. SOAR hosts an interactive web interface for users to analyze and visualize spatial gene expressions in different tissue samples. SOAR also provides functions for users to investigate spatial variability and cell-cell interaction across datasets, allowing systematic comparison across samples, cell types, and disease types.

Multiple existing data resources are available for spatial transcriptomics, such as SpatialDB[2], STOmicsDB[3], SPASCER[4], Museum of Spatial Transcriptomics[5], spatialLIBD[6], STAR-FINDer[7], and Amyotrophic Lateral Sclerosis Spinal Cord Atlas[8]. SpatialDB[2] is a manually curated spatial transcriptomics database with 305 samples. It provides spatial expression visualization and pre-computed results for sample-wise spatially variable gene identification.

STOmicsDB[3] hosts 1,033 spatial transcriptomics samples, of which 463 samples are processed and have spatial coordinates. Users can perform sample-wise clustering and spatial variability analysis on the processed datasets using STOmicsDB. SPASCER[4] is a spatial transcriptomics database providing clustering, spatial variability, pathway, gene regulatory network, and neighborhood-based cell-cell interaction analyses. Out of all the 1,082 samples on SPASCER, 1,030 samples have spatial transcriptomics data analysis results. Museum of Spatial Transcriptomics[5] is an annotated list of spatial transcriptomics literature with links to publications and their datasets. Multiple data atlases are available, including spatialLIBD[6], STAR-FINDer[7], and Amyotrophic Lateral Sclerosis Spinal Cord Atlas[8]. These atlases provide interfaces for visualizing spatial expressions, but their scopes are limited to one particular tissue type (brain, intestine, and spinal cord, respectively). To the best of our knowledge, SOAR is the largest spatial transcriptomics database that provides unified data access, exploration, and analysis functions across different tissue types. In addition to hosting a large number of systematically curated and uniformly processed datasets, SOAR provides easy-to-use functions for users to interactively visualize spatial gene expression, perform clustering, explore spatial variability, analyze neighborhood-based and distance-based cell-cell interactions, and conduct cross-dataset meta-analyses (Supplementary Table 1). In summary, SOAR can be a great utility to study cell functions and disease pathologies in the tissue context.

We systematically curated 1,735 samples across six species and 23 tissue types from 145 datasets (Supplementary Table 2). The curated datasets were generated by nine different spatial transcriptomics technologies (Fig. 1a-h). The datasets were reviewed, annotated, and pre-processed by a standardized workflow that we developed (Methods). In SOAR's data browser, users could browse the curated datasets to pinpoint samples of interest using

metadata-based filters and visualize spatial expressions (Fig. 2b). The processed data, including normalized expression, images, coordinates, and phenotypic data of all samples are available for download. In addition to providing data access and visualization, SOAR could help users evaluate gene spatial variability and investigate cell-cell interactions across multiple datasets.

The spatial patterns of gene expression inform us about the migration of cells, their responses to different tissue environments, and cell-cell interactions[1, 9]. SOAR allows users to evaluate the statistical significance of a gene's spatial variability in different tissue samples and cell types. As a case study, we visualized the spatial variability of CXCL9 and CXCL13 in our curated breast cancer samples (Fig. 2a). Cytokine genes CXCL9 and CXCL13 are known to be associated with the degree and prognosis of cancer[10, 11], and their expressions were found to have significant spatial pattern variation in cancer samples[9, 12]. Our results showed significant spatial expression variation of these genes in malignant capture locations. In addition, we found that CXCL9 and CXCL13 were also expressed in a visually distinct region along with the interleukin 2 receptor subunit (IL2RB) (Fig. 2b). Interleukin is an inflammatory cytokine that activates and regulates immune cells[13]. As an interleukin signalling gene, IL2RB is involved in T cell-mediated immune responses[13]. Our results indicate that the spatial co-expression of CXCL9, CXCL13, and IL2RB might be associated with a pro-inflammatory anti-tumor response, which aligns with previous findings[9, 10]. This illustration demonstrates that even without histological annotation, SOAR could identify biologically relevant genes with localized expression patterns (Supplementary Material).

Gene expression variation at the cell borders, especially between different cell types, reflects cell-cell interaction levels [14]. With SOAR, users could explore whether the expression of a

given gene in a query cell type is associated with other cell types in its neighborhood (Fig. 2c, Supplementary Material). As a use case, we studied how neighbouring cell types may influence complement component 3 (C3) expression in cancer tissue fibroblasts. C3 plays a central role in complement activation, which may aid immune surveillance against tumor cells[15, 16]. In tumor microenvironments, cancer-associated fibroblasts are often the most prominent cell type[17, 18] and known to express C3[19]. Existing studies proposed that manipulating the function of cancer-associated fibroblasts is a strategy to control tumor progression[16, 20]. Therefore, it is important to study how cell-cell interactions influence the expression of C3 in cancer-associated fibroblasts. Our results showed that fibroblasts neighboring with CD4-infiltrated capture locations had higher C3 expression levels compared with those farther away from CD4-infiltrated regions (Fig. 2c). Our results are in line with existing knowledge of the positive relationship between C3 activation and T cell infiltration in tumor tissues[21]. This example demonstrates SOAR's capacity to identify neighborhood-based cell-cell interaction signals.

Interactions between cells can occur beyond simple adjacency through the secretion of cytokines[14]. To characterize the distance-based interactions between cells, we developed a Gaussian process model based on the Spatial Variance Component Analysis framework[22]. Our method models gene expressions in a query cell type using its distance from other capture locations and the expression patterns in other cell types (Methods). Using this function, users could visualize and investigate potential far-reaching cell-cell interactions (Fig. 2d, Supplementary Material). As an example, we studied how the expression of the major histocompatibility complex (MHC) genes in cancer cells is affected by their distances from CD4 and CD8 infiltrated regions. MHC is a family of genes encoding cell surface proteins important in the adaptive immune system[23], and the expression of MHC genes in

cancer cells is known to be induced by cytokines secreted by T cells[24]. Our analysis revealed that malignant capture locations have higher expression levels of MHC genes when they are closer to CD4 (Supplementary Fig. 2) and CD8 (Fig. 1i) infiltrated regions, which is corroborated by the existing knowledge[24]. Compared with neighborhood-based analysis, our distance-based model produces consistent results at the center-to-center distance between adjacent capture locations. Moreover, it provides better power in detecting cell-cell interaction signals and enables the evaluation of long-distance cell-cell interactions (Supplementary Fig. 3-4).

In summary, SOAR hosts a large number of systematically processed and annotated spatial transcriptomics datasets as a freely available resource. Besides data access and download, SOAR provides interactive analysis functions for visualizing spatial gene expression, evaluating gene spatial variability, and studying both neighborhood-based and distance-based cell-cell interactions. SOAR will be continuously maintained to provide regularly updated and scalable utility to the biological, biomedical, and clinical research communities for harnessing the power of spatial transcriptomics data.

References

1. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome biology* **23**, 1-23 (2022).
2. Fan, Z., Chen, R. & Chen, X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Research* **48**, D233-D237 (2020).
3. Xu, Z. et al. STOmicsDB: a database of Spatial Transcriptomic data. *bioRxiv* (2022).
4. Fan, Z. et al. SPASCER: spatial transcriptomics annotation at single-cell resolution. *Nucleic Acids Research* (2022).
5. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nature Methods*, 1-13 (2022).
6. Maynard, K.R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* **24**, 425-436 (2021).
7. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810-826. e823 (2021).
8. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89-93 (2019).
9. Svensson, V., Teichmann, S.A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nature methods* **15**, 343-346 (2018).

10. Bronger, H. et al. CXCL9 and CXCL10 predict survival and are regulated by cyclooxygenase inhibition in advanced serous ovarian cancer. *British journal of cancer* **115**, 553-563 (2016).
11. Fukuda, Y. et al. Endogenous CXCL9 affects prognosis by regulating tumor-infiltrating natural killer cells in intrahepatic cholangiocarcinoma. *Cancer science* **111**, 323-333 (2020).
12. Pelka, K. et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734-4752. e4720 (2021).
13. Jia, Z. et al. Effect of IL2RA and IL2RB gene polymorphisms on lung cancer risk. *International immunopharmacology* **74**, 105716 (2019).
14. Armingol, E., Officer, A., Harismendy, O. & Lewis, N.E. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics* **22**, 71-88 (2021).
15. Afshar-Kharghan, V. The role of the complement system in cancer. *The Journal of clinical investigation* **127**, 780-789 (2017).
16. Shu, C. et al. C3a-C3aR signaling promotes breast cancer lung metastasis via modulating carcinoma associated fibroblasts. *Journal of Experimental & Clinical Cancer Research* **39**, 1-14 (2020).
17. Quail, D.F. & Joyce, J.A. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine* **19**, 1423-1437 (2013).
18. Sahai, E. et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nature Reviews Cancer* **20**, 174-186 (2020).
19. Sun, D. et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic acids research* **49**, D1420-D1430 (2021).
20. Houthuijzen, J. & Jonkers, J. Cancer-associated fibroblasts as key regulators of the breast cancer tumor microenvironment. *Cancer and Metastasis Reviews* **37**, 577-597 (2018).
21. Lin, K. et al. Complement component 3 is a prognostic factor of non□small cell lung cancer. *Molecular medicine reports* **10**, 811-817 (2014).
22. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell reports* **29**, 202-211. e206 (2019).
23. Haworth, K.B. et al. Going back to class I: MHC and immunotherapies for childhood cancer. *Pediatric blood & cancer* **62**, 571-576 (2015).
24. Chen, L. & Flies, D.B. Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nature reviews immunology* **13**, 227-242 (2013).

Methods

Data Collection

We queried the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) for human and mouse spatial transcriptomics datasets using the keywords "spatial+transcriptomics", "spatial+transcriptome", "spatial+RNA-seq", and "spatial+RNA+sequencing", and downloaded 361 datasets from unique GEO series (GSE) accessions. Additionally, we manually reviewed the papers in the Museum of Spatial Transcriptomics[1] and collected 75 publicly available datasets. We also collected 114 datasets from other resources including Single Cell Portal (https://singlecell.broadinstitute.org/single_cell), 10x Genomics spatial gene expression demonstration datasets (https://support.10xgenomics.com/spatial-gene-expression/datasets), Spatial Research Lab (https://www.spatialresearch.org/resources-published-datasets/), 10x Genomics spatial publication list (https://www.10xgenomics.com/resources/publications), Amyotrophic Lateral Sclerosis Spinal Cord Atlas[2], spatialLIBD[3], STAR-FINDer[4], and Brain Research through Advancing Innovative Neurotechnologies Initiative – Cell Census Network (https://biccn.org/data). Next, we removed the duplicative datasets, validated that the downloaded data used fluorescence in situ hybridization (FISH) or next-generation sequencing (NGS) based spatial transcriptomics technology, and excluded the datasets missing spatial coordinates information.

In total, we have collected 145 datasets containing 1,735 spatial transcriptomics samples from nine different technologies and six species (human, mouse, chicken, pig, thale cress, and aspen). The human and mouse samples come from different organs (brain, heart, intestine, joints, kidney, liver, lung, lymph node, muscle, prostate, skin, spinal cord, testis, uterus) and other specific tissues including body fat, cancer, and embryonic tissues.

Data Processing

We downloaded the count matrices and coordinate information for each dataset and applied a systematic data processing workflow to all the collected datasets. To account for the resolution and sequencing depth difference among spatial transcriptomics techniques, samples measured using different technologies were processed with corresponding quality control (QC) protocols. For 10x Visium, ST[5], DBiT-seq[6], and sci-Space[7] datasets, we removed the capture locations with fewer than 500 unique molecular identifiers (UMIs), fewer than 500 genes, or ≥ 25% mitochondrial reads[8]. We further excluded the capture locations with a total UMI count (or a total number of genes) three standard deviations below the median[9]. Finally, we filtered out the genes that are expressed in less than five capture locations. FISH technologies like MERFISH[10, 11], osmFISH[12], seqFISH[10, 13, 14], and seqFISH+[15] typically measure a smaller number of genes at lower sequencing depth. Therefore, we only performed capture-location-wise QC on these datasets. In the QC pipeline for FISH datasets, genes expressed in fewer than five capture locations were excluded[9], and capture locations with fewer than 500 UMIs or ≥ 25% mitochondrial reads were removed[8]. We performed QC on Slide-seq[16, 17] samples so that the genes with total UMI counts less than 300 were excluded[18]. Additionally, only the capture locations with total UMI counts greater than 100 and less than 25% mitochondrial reads were included[18]. Supplementary Fig. 5 and Fig. 1d-g demonstrates the average number of capture locations and genes per sample, as well as the average number of UMI counts and genes per capture location of samples generated by different technologies before and after QC, respectively. After QC, we normalized and transformed the raw datasets using SCTransform, a framework for the normalization and variance stabilization of molecular count data[19]. We next performed principal component analysis on normalized data and clustered the capture locations through a shared nearest neighbor approach[9]. All data processing was conducted using R v4.0.5 and Seurat V3[9]. The

processed samples were stored in a standardized format including an expression table (gene by capture location) and a coordinates table (capture location by Cartesian coordinates).

Cell Type Annotation

We performed cell type annotation on SOAR's spatial transcriptomics datasets using SingleR[20], a method capable of annotating the cells in test datasets based on their similarities to reference single-cell RNA sequencing (scRNA-seq) datasets with known cell types. To identify such reference datasets, we queried the GEO and curated an average of two annotated scRNA-seq datasets for each tissue type featured in SOAR. Next, we processed these datasets using an approach similar to that of the spatial transcriptomics dataset, including QC, normalization, and transformation. The scRNA-seq datasets were then used as references for annotating the cell types of spatial transcriptomic capture locations of the corresponding tissue type. In particular, for complex tissues such as tumor and brain, we adopted the following heuristic-guided approaches to improve the performance of cell type annotation.

For cancer datasets (197 samples in total), we first identified possible lymphocyte capture locations using known biomarkers (CD3D, CD3E, and CD3G) through differential expression analysis. The clusters with greater than 1.2 average log fold-changes for one or more lymphocyte genes and over 20% capture locations expressing at least one of the biomarkers were noted as potential lymphocyte clusters. Next, we identified capture location clusters that differentially expressed CD4, CD8A, or CD8B genes in a similar manner. Among these clusters, we annotated those that were also found to be possible lymphocyte clusters in the first step as CD4-infiltrated or CD8-infiltrated based on the percentage of capture locations in the cluster expressing CD4 versus CD8A and CD8B genes. Next, we performed cell type annotation on all the remaining clusters using SingleR[20], with scRNA-seq

datasets of the same cancer type curated in a previous study[21] as the reference. The cell type of each cluster was then determined by the annotation results from different references, weighted by the number of cells in the corresponding reference dataset.

For non-cancer brain datasets (911 samples in total), we followed the Common Cell Type Nomenclature (CCN)[22] and annotated their capture locations as glutamatergic, GABAergic, or non-neuronal. Two scRNA-seq datasets from the Allen Brain Map (https://portal.brain-map.org/atlases-and-data/rnaseq) were used as the references – the Human Multiple Cortical Areas SMART-seq dataset (for annotating human samples) and the Mouse Whole Cortex and Hippocampus dataset (for annotating mouse samples). Firstly, we identified marker gene sets for each cell type and each species by performing differential gene expression analysis on the corresponding reference scRNA-seq dataset using Seurat V3[9]. Next, we used AUCell[23] to score the activity of glutamatergic, GABAergic, and non-neuronal gene sets at each capture location based on marker gene expressions. Capture location clusters in the sample can then be classified as neuronal or non-neuronal according to the sum of AUCell scores across capture locations[23]. Finally, we used SingleR[20] to annotate the neuronal clusters as glutamatergic or GABAergic based on a filtered version of the reference dataset that only contained neuronal cells.

Website Development

SOAR is a comprehensive and user-friendly database that aids the exploration and analysis of spatial transcriptomics datasets. SOAR was implemented using the R Shiny framework (R v4.0.5) on an Apache2 HTTP server and is compatible with smartphones and tablets. The website consists of five functional components, “Home”, “Data Browser”, “Explore Gene”, “Download”, and “Help” (Supplementary Fig. 1a). The “Home” module includes an overview of SOAR, and users can search for a gene of interest in this module. In the “Data

Browser" module, users can identify a sample of interest using different filters and visualize its spatial gene expressions (Supplementary Fig. 1b). Upon searching for a gene on the homepage, users will land in the "Explore Gene" module, which enables users to evaluate the spatial variability of genes in different tissues and assess possible cell-cell interactions (Supplementary Fig. 1c). All the results and visualizations from user-performed analyses are downloadable. In the "Download" module, users can download all the curated gene expression data, coordinate data, metadata, and sample-wise analysis results. The "Help" page documents the website and includes a tutorial with step-by-step instructions for using the database. SOAR is free and open to all users at https://soar.fsm.northwestern.edu/ and there is no login requirement.

Data Browser Module

To aid user conducted analysis, we constructed a comprehensive data browser that is part of SOAR and contains the meta-data for all included spatial transcriptomics datasets. For each dataset, detailed information includes the hyperlink to the corresponding publication, the spatial transcriptomics technology used, and sample information including the number of samples, the species, organ, tissue, and the disease state of the sample. Furthermore, we document the average number of capture locations and genes in each sample after QC. Our data browser allows users to quickly select samples of interest to visualize spatial gene expressions and perform spatial variability analysis via interactive figures and tables. All the generated figures and tables are easily downloadable to support customized and large-scale research projects.

Explore Gene Module

The gene search bar on the homepage of SOAR allows users to query the results of these analyses for a specific gene of interest. Upon searching for a gene, SOAR directs the user to the "Explore Gene" tab, which subsequently prompts the user to narrow down the list of

datasets by tissue type and species. SOAR allows users to perform three types of analyses – spatial variability, neighborhood-based cell-cell interaction, and distance-based cell-cell interaction. All the p-values were adjusted for multiple testing using the false discovery rate (FDR) approach, and we assume statistical significance at an adjusted p-value of $q < 0.05$.

Spatial Variability

Studying the spatial variation of gene expression is helpful for understanding cell migration and cell-cell interaction[24, 25]. To facilitate the characterization of the functional architecture of complex tissues, we identified genes with spatial patterns of significant expression variation using SPARK-X, a non-parametric method for detecting spatially variable genes[26]. Spatial variability analyses were conducted across the whole tissue and in different cell types on all the samples.

Neighborhood-based Cell-Cell Interaction

Cells of different cell types may interact through cell-cell contact[27]. Spatial transcriptomics enables us to study cell-cell interactions by investigating whether a gene's expression in one cell type appears to be promoted or inhibited when in another cell type's neighborhood. To identify possible cell-cell interactions, we investigated whether the gene expression levels in a query cell type (CT_Q) are influenced by its neighboring interacting cell type (CT_I).

Denote the cell type CT_Q's capture locations as CL_Q and the cell type CT_I's capture locations as CL_I. In order to evaluate neighbouring interactions, we performed Wilcoxon rank-sum tests to test if genes are differentially expressed in CL_Q adjacent and non-adjacent to CL_I using the FindAllMarkers function in Seurat V3[9]. If a gene G is more highly expressed in CL_Q adjacent to CL_I, CL_I may promote the expression of G in CL_Q.

Distance-based Cell-Cell Interaction

Denote the Cartesian coordinates of the ith capture location CL_Q as $\left(x_Q^i, y_Q^i\right)$ and the Cartesian coordinates of the jth CL_I as $\left(x_I^j, y_I^j\right)$. The Euclidean distance between the ith CL_Q and the jth CL_I $(d_{i,j})$ can be calculated as

$$d_{i,j} = \sqrt{\left(x_Q^i - x_I^j\right)^2 + \left(y_Q^i - y_I^j\right)^2}$$

In addition to direct contact, cells can interact over a distance through signalling involving the secretion of cytokines[27]. To assess cell-cell interactions over distance, in a given sample, we modelled the regulation effect of an interacting cell type CT_I on the expression of gene G in a query cell type CT_Q using the following model.

$$\boldsymbol{y} = ZX\boldsymbol{a} + \boldsymbol{c}$$

where:

$$Z_{i,j} = \exp\left(-d_{i,j}^2/2b^2\right)$$

The notations are:

- N – the number of CL_Q in the given sample.
- M – the number of CL_I in the given sample.
- D – the total number of genes in the given sample.
- $\boldsymbol{y}$ – the expression of G in CL_Q. Dimensions: $N \times 1$.
- Z – a function of $d_{i,j}$, the Euclidean distance between the ith CL_Q and the jth CL_I. Dimensions: $N \times M$.
- X – the expression of all the genes in CL_I. Dimensions: $M \times D$.
- $\boldsymbol{a}$ – a parameter vector whose elements are of the same value a. Dimensions: $D \times 1$.
- b – a scalar parameter.

- $\boldsymbol{c}$ – a parameter vector whose elements are of the same value c. Dimensions: $N \times 1$.

Given CT_Q, CT_I, and G, to quantify the overall influence of $d_{i,j}$ on $\boldsymbol{y}$ in a certain sample, we can calculate md (mean derivative), the average of all the derivatives $\frac{\partial y_i}{\partial d_{i,j}}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, M$.

$$md = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial y_i}{\partial d_{i,j}}$$

where:

$$\frac{\partial y_i}{\partial d_{i,j}} = -\frac{a}{b^2} \cdot d_{i,j} \cdot \sum_{k=1}^{D} \exp\left(-\frac{d_{i,j}^2}{2b^2}\right) \cdot X_{j,k}$$

The magnitude of md reflects how strongly $d_{i,j}$ affects $\boldsymbol{y}$. Note that the sign of md is the opposite from the sign of a. Intuitively, the sign of a reflects the direction of the association between the gene expression in CL_Q and the distance between CL_Q and CL_I. Therefore, a negative md (positive a) may reflect CL_I's positive regulation effect on the expression of G in CL_Q, whereas when $md > 0$ ($a < 0$), CL_I capture locations may play an inhibitory role.

References

1. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nature Methods*, 1-13 (2022).
2. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89-93 (2019).
3. Maynard, K.R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* **24**, 425-436 (2021).
4. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810-826. e823 (2021).
5. Ståhl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).
6. Liu, Y. et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665-1681. e1618 (2020).
7. Srivatsan, S.R. et al. Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111-117 (2021).

8. Qian, J. et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell research* **30**, 745-762 (2020).
9. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
10. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
11. Moffitt, J.R. et al. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proceedings of the National Academy of Sciences* **113**, 14456-14461 (2016).
12. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature methods* **15**, 932-935 (2018).
13. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature methods* **11**, 360-361 (2014).
14. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342-357 (2016).
15. Eng, C.-H.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235-239 (2019).
16. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
17. Stickels, R.R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology* **39**, 313-319 (2021).
18. Fan, Z., Chen, R. & Chen, X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Research* **48**, D233-D237 (2020).
19. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome biology* **20**, 1-15 (2019).
20. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* **20**, 163-172 (2019).
21. Sun, D. et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic acids research* **49**, D1420-D1430 (2021).
22. Miller, J.A. et al. Common cell type nomenclature for the mammalian brain. *Elife* **9**, e59928 (2020).
23. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **14**, 1083-1086 (2017).
24. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome biology* **23**, 1-23 (2022).
25. Svensson, V., Teichmann, S.A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nature methods* **15**, 343-346 (2018).
26. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology* **22**, 1-25 (2021).
27. Armingol, E., Officer, A., Harismendy, O. & Lewis, N.E. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics* **22**, 71-88 (2021).

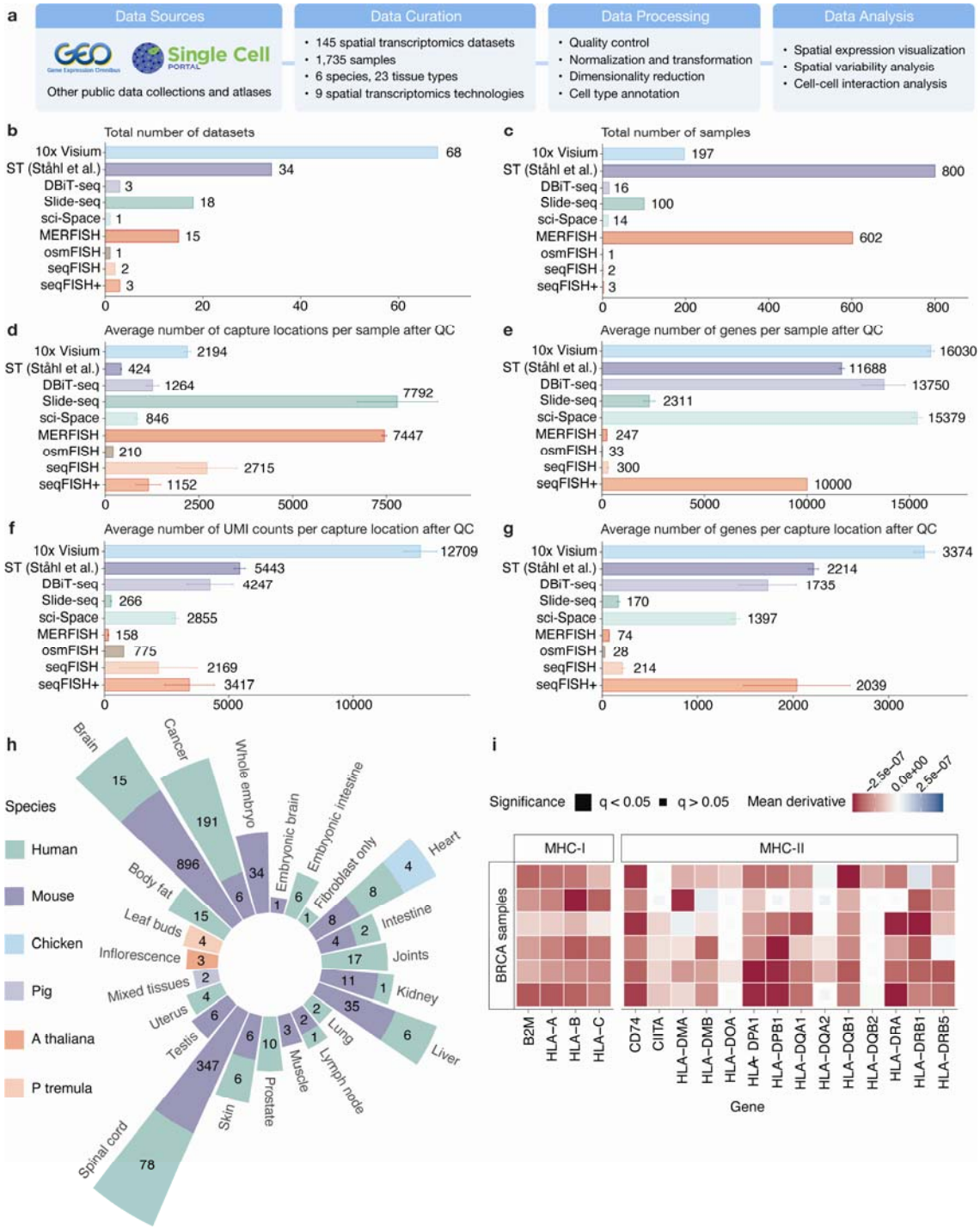

Fig. 1: Overview of SOAR, summary statistics of the hosted datasets, and illustration of SOAR's functionality. a, Datasets hosted on SOAR were curated from public domains and processed using a standardized workflow. SOAR also provides interactive interfaces for users to visualize spatial gene expressions, evaluate the spatial variability of genes, and assess cell-cell interactions. **b-g**, Statistics of data from different spatial transcriptomics technologies in SOAR. The 95% confidence intervals for the means

are plotted as error bars. **h**, Number of samples of each species and tissue type combination. SOAR contains 1,735 spatial transcriptomics samples from 6 different species across 23 tissue types. **i**, The relationship between the expression of MHC genes in malignant capture locations and their distances from CD8-infiltrated capture locations. Each row corresponds to the analysis results of one cancer sample. SOAR reveals that malignant capture locations have higher expression levels of MHC genes when they are closer to CD8-infiltrated capture locations. BRCA, breast cancer; Log2FC, log-fold changes; MHC, major histocompatibility complex; q, false-discovery-rate-adjusted p-values; QC, quality control; UMI, unique molecular identifier.

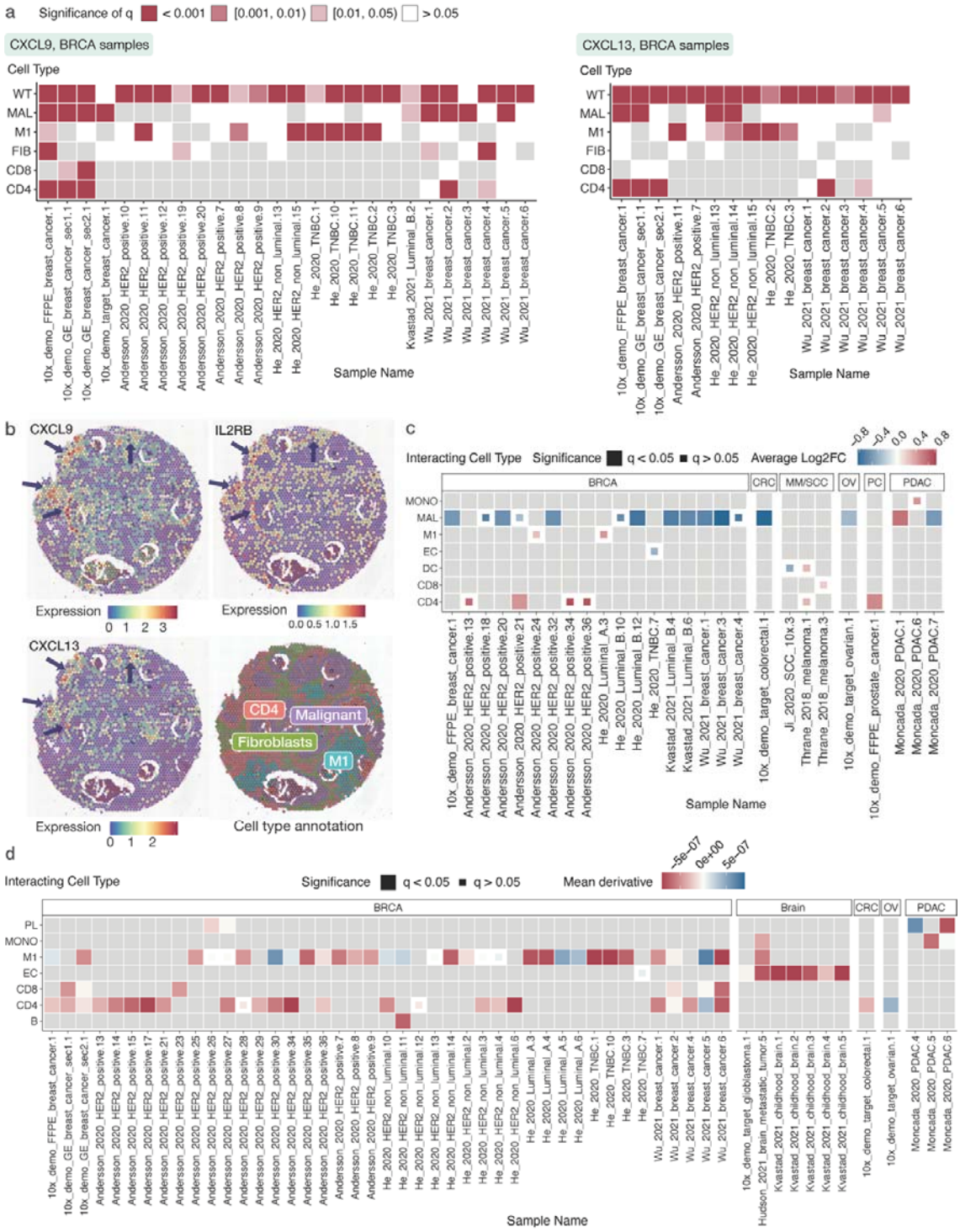

Fig. 2: Using SOAR to explore spatially variable genes and cell-cell interactions in cancer spatial transcriptomics datasets. Users can use SOAR to visualize (**a**) the significance of a gene's spatial variability in different cell types and (**b**) spatial gene expressions in a sample of interest. To study cell-cell interactions, users can explore (**c**) neighborhood-based and (**d**) distance-based cell-cell interactions. **a**, The

spatial variability of CXCL9 and CXCL13 in breast cancer datasets. p-values are adjusted using the false discovery rate approach. **b**, Spatial gene expressions of CXCL9, CXCL13, and IL2RB in a breast cancer sample. Black arrows indicate distinct regions of high expression of the genes. **c**, C3's differential expression between fibroblast capture locations adjacent and non-adjacent to other cell types. Each tile in the heatmap is coloured by the log-fold change in gene expression and sized according to its statistical significance after false discovery rate adjustment. **d**, The relationship between B2M's expression in malignant capture locations and their distances from other cell types. Positive (blue) mean derivatives indicate inhibitory gene regulatory effects, and negative (red) mean derivatives suggest promotional effects. BRCA, breast cancer; CD4, CD4-infiltrated; CD8, CD8-infiltrated; CRC, colorectal cancer; DC, dendritic cells; EC, endothelial cells; FIB, fibroblasts; Log2FC, log-fold changes; MAL, malignant cells; MM, melanoma; MONO, monocytes; OV, ovarian cancer; PC, prostate cancer; PDAC, pancreatic ductal adenocarcinoma; PL, plasma cells; q, false-discovery-rate-adjusted p-value; SCC, squamous cell carcinoma; WT, whole tissue.