MOCHA: MULTI-SAMPLE OMICS COHORTS WITH HUMAN ANNOTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In spatially resolved transcriptomics (SRT) research, gene expression profiling with spatial context has enabled spatial domain identification within single tissue samples. Extending these analyses to multiple biological samples presents additional challenges, including cross-sample variability and batch effects. Method development has been limited by the lack of datasets that combine multi-subject cohorts with expert-derived annotations. We present MOCHA (Multi-sample Omics Cohorts with Human Annotation), a curated resource for developing and evaluating multi-sample SRT methods. MOCHA integrates molecular profiles, spatial profiles, and high-resolution Hematoxylin and Eosin (H&E) images across multiple subjects, with each sample paired with domain annotations from expert pathologists. For algorithm development and evaluation, MOCHA provides standardized data organization, efficient storage formats for large-scale processing, and protocols for handling batch effects in multi-sample integration.

1 INTRODUCTION

Spatially resolved transcriptomics (SRT) links gene expression profiles to precise tissue coordinates, enabling quantitative analysis of microanatomy and cellular organization at high resolution. Multiple platforms now make SRT broadly accessible, including sequencing-based assays such as 10x Genomics Visium and Slide-seq (Ståhl et al., 2016; Tian et al., 2023) and imaging-based assays such as MERFISH (Chen et al., 2015) and STARmap (Wang et al., 2018). The resolution of these technologies varies from multi-cellular spots to near single-cell measurements, but all require computational approaches that can identify coherent tissue domains by combining molecular profiles with spatial information.

Several repositories have been developed to organize publicly available datasets, including SORC for cancer research (Zhou et al., 2024), Aquila for cross-disease analyses (Zheng et al., 2023), and others such as SODB (Yuan et al., 2023), STOmicsDB (Xu et al., 2022), and SpatialDB (Fan et al., 2020). Despite this progress, multi-subject datasets with expert-generated spatial annotations remain limited. This gap constrains systematic method development for multi-sample integration—an essential setting for cohort-level studies that must model biological heterogeneity alongside technical variation.

Methodological advances underscore this need. Early work emphasized single-sample domain identification, including Bayesian modeling approaches such as BayesSpace (Zhao et al., 2021) and deep learning methods that integrate histology, including iIMPACT (Jiang et al., 2024). More recent approaches—such as BASS (Li & Zhou, 2022), BayeSmart (Guo et al., 2024), and graph-based methods like STAGATE (Dong & Zhang, 2022)—extend analysis to multiple samples using distinct strategies, from clustering across tissues to learning shared representations. Additional challenges, such as deconvolution of mixed spots (Chen et al., 2022; 2023; Luo et al., 2024) and correction for batch effects, reinforce the importance of datasets that provide aligned molecular, spatial, and histological information together with expert annotations.

We introduce MOCHA, a Multi-sample Omics Cohorts with Human Annotation database for training and evaluation of multi-sample SRT methods. MOCHA aggregates multi-subject datasets that each include a gene expression matrix, spatial coordinates, and a co-registered high-resolution Hematoxylin and Eosin (H&E) image (Chan, 2014). Each sample is accompanied by spatial domain

labels produced by an expert pathologist, enabling evaluation of domain delineation and representation learning in multi-sample contexts. To promote reproducibility and accessibility, MOCHA is released in formats readily usable with Python and R and distributed for integration into existing pipelines.

2 DATASETS

To assemble a resource for multi-sample spatial domain identification, we curated a set of publicly available SRT datasets. Following an approach similar to that in the STimage-1K4M review Chen et al. (2024), we systematically searched repositories including 10x Genomics, Gene Expression Omnibus (GEO), and Spatial Research. Our selection criteria required each study to provide a cell-by-gene expression count matrix, a spatial coordinate matrix, and cellular annotations delineated by a pathologist using the corresponding H&E images.

This search yielded 10 distinct cohorts, summarized in Table 1. Cancer-related datasets include HER2-positive breast cancer (BC_HER2+) (Andersson et al., 2021), high-plasticity subtypes (BC_HP) (Coutant & et al., 2023), recurrent neoplastic heterogeneity (BC_NP) (Wu et al., 2021), triple-negative breast cancer (BC_TNBC) (Wang et al., 2024), colorectal cancer consensus molecular subtypes (CRC_CMS) (Valdeolivas et al., 2024), kidney cancer with tertiary lymphoid structures (KC_TLS) (Dawo et al., 2023), lung cancer with tertiary lymphoid structures (LC_TLS) (Dawo et al., 2023), and renal cell carcinoma with tertiary lymphoid structures (RCC_TLS) (Meylan & et al., 2022), along with human dorsolateral prefrontal cortex (DLPFC) (Maynard et al., 2021) and mouse olfactory bulb (MOB) (Ståhl & et al., 2016).

Table 1: A summary of the SRT datasets. (BC: Breast cancer; CRC: Colorectal cancer; DLPFC: Dorsolateral prefrontal cortex; KC: Kidney cancer; LC: Lung cancer; MOB: Mouse olfactory bulb; RCC: Renal cell carcinoma)

Cohort	Tissue	Technology	Subjects	Samples
BC_HER2+_10x	HER2-positive (HER2+) breast cancer	10x Visium	8	8
BC_HP_10x	High-plasticity (HP) breast cancer subtypes	10x Visium	12	14
BC_NP_10x	Recurrent neoplastic (NP) cell heterogeneity in breast cancer	10x Visium	6	6
BC_TNBC_ST	Triple-negative breast cancer (TNBC)	ST	94	94
CRC_CMS_10x	Colorectal cancer consensus molecular subtypes (CMS)	10x Visium	11	14
DLPFC_10x	Dorsolateral prefrontal cortex	10x Visium	3	12
KC_TLS_10x	Kidney cancer with tertiary lymphoid structures (TLS)	10x Visium	3	3
LC_TLS_10x	Lung cancer with tertiary lymphoid structures (TLS)	10x Visium	5	5
MOB_ST	Mouse olfactory bulb	ST	1	12
RCC_TLS_10x	Tertiary lymphoid structures (TLS) in renal cell carcinoma	10x Visium	23	23

These cohorts span a wide range of tissue types and disease contexts, encompassing both human and mouse studies, and multiple technological platforms (10x Genomics Visium and ST). The scale also varies substantially, with BC_TNBC including 94 subjects and 94 samples, while smaller datasets such as KC_TLS and LC_TLS consist of only three and five samples, respectively. Together, these datasets enable evaluation of multi-sample spatial domain identification methods. A summary of molecular characteristics, including the number of spots, genes, and data sparsity for each cohort, is presented in Figure 1.

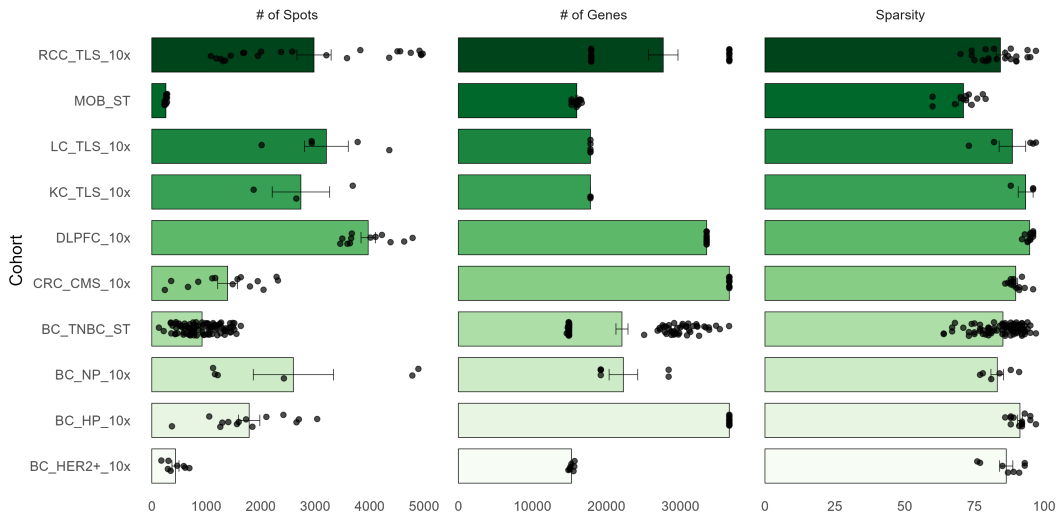

Figure 1: A summary of the molecular profiles for each cohort.

3 Pre-Processing and Batch Effect Correction

A standard pipeline for preprocessing multi-sample SRT data starts by concatenating the raw gene expression matrices from each sample over a set of common genes, followed by library size normalization to correct for variability in sequencing depth. This adjustment can be performed using packages such as `scater` and `scran`, which implement techniques such as the trimmed mean of M-values (TMM), relative log expression (RLE), and upper-quartile scaling (Robinson & Oshlack, 2010; Anders & Huber, 2010; Bullard et al., 2010; McCarthy et al., 2017). Alternatively, frameworks such as `Seurat` and `scanpy` apply a global-scaling approach in which counts for each cell are divided by the total count, rescaled to a fixed scaling factor (e.g., 10,000), and log-transformed to stabilize variance (Hao et al., 2023; Wolf et al., 2018).

Following normalization, dimensionality reduction can be performed through feature selection or projection methods. Feature selection can involve identifying spatially variable genes (SVGs) using methods such as SPARK-X (Zhu et al., 2021; Zhao et al., 2021; Jiang et al., 2024), or highly variable genes (HVGs), which are generally preferred in studies involving multiple subjects to reduce inter-subject variability (Li & Zhou, 2022). Dimensionality reduction can also be achieved by projecting the data into a lower-dimensional space using techniques such as PCA, t-SNE, UMAP, or graph attention autoencoders as implemented in STAGATE (van der Maaten & Hinton, 2008; Becht et al., 2019; Dong & Zhang, 2022).

Batch correction can be subsequently performed to adjust for systematic variation between samples. One common approach is to operate on reduced feature spaces using techniques such as Harmony (Korsunsky et al., 2019; Li & Zhou, 2022; Guo et al., 2024). An overview of this batch effect correction, and feature selection, workflow is demonstrated in Figure 2.

An alternative pipeline for batch correction is implemented in Crescendo, which avoids transformation to a reduced-dimensional space and instead models the raw, integer-valued counts directly (Millard et al., 2025). This approach employs a generalized linear mixed model (GLMM) in which the batch is included as a random effect, preserving the discrete structure of the data. Crescendo can extend single-sample spatial clustering models such as BayesCafe, which relies on the zero-inflated negative binomial (ZINB) distribution (Li et al., 2024), to multi-sample settings by integrating batch correction directly within the generative hierarchy.

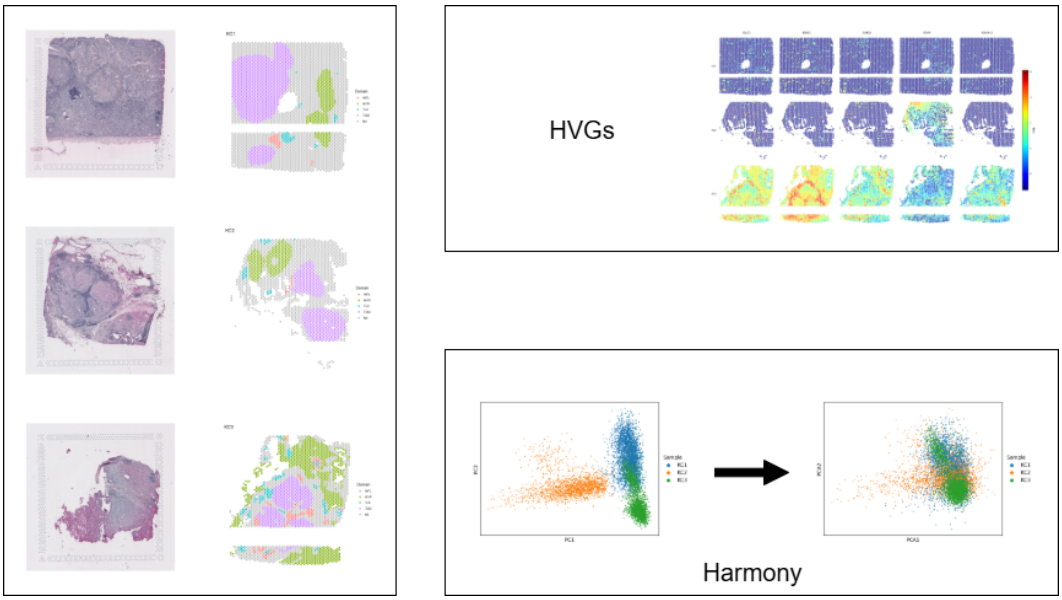

Figure 2: AA standard pipeline for feature selection with HVGs and batch effect correction using Harmony, illustrated with the KC_TLS_10x cohort (Dawo et al., 2023).

4 Multi-Sample Spatial Clustering Methods

Recent advances in computational modeling have led to methods that extend spatial transcriptomics analysis from single-sample to multi-sample settings. These approaches are designed to integrate spatial and molecular information across subjects while accounting for technical and biological variability.

As summarized in Table 2, BayeSMART is a Bayesian framework for multi-sample spatial clustering that integrates reconstructed single-cell information from histology images with spatial gene expression (Guo et al., 2024). BASS is a hierarchical Bayesian model that jointly performs cell type clustering and spatial domain identification across samples (Li & Zhou, 2022). STAGATE is a graph attention autoencoder that generates low-dimensional embeddings by combining spatial neighborhood structure with molecular profiles (Dong & Zhang, 2022).

Table 2: A summary of the existing multi-sample spatial clustering methods. These Bayesian (Bayes) or deep learning (DL) approaches use Principal Component Analysis (PCA) or autoencoders (AE) for dimension reduction. Additionally, BayeSMART integrates information from H&E images.

Method	Dimension reduction	H&E	Approach	Language	Year
BayeSMART	PCA	✓	Bayes	R/C++	2024
BASS	PCA		Bayes	R/C++	2022
STAGATE	AE		DL	Python	2022

In a majority of the cancer studies included in MOCHA, the detailed pathologist annotations can be grouped into four broad categories: immune, stroma, tumor, and normal. These groupings, described in the Supplementary Material, provide a consistent reference structure for applying multi-sample spatial clustering methods while accommodating variability across cohorts.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010. doi: 10.1186/gb-2010-11-10-r106. URL https://doi.org/10.1186/gb-2010-11-10-r106.

A. Andersson, L. Larsson, L. Stenbeck, and et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12:6012, 2021. doi: 10.1038/s41467-021-26271-2.

Etienne Becht, Leland McInnes, John Healy, and et al. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37:38–44, 2019. doi: 10.1038/nbt.4314. URL https://doi.org/10.1038/nbt.4314.

James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(94):1–13, 2010. doi: 10.1186/1471-2105-11-94. URL https://doi.org/10.1186/1471-2105-11-94.

J. K. C. Chan. The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *International Journal of Surgical Pathology*, 22(1):12–32, 2014. doi: 10.1177/1066896913517939. URL https://doi.org/10.1177/1066896913517939.

J. Chen, W. Liu, T. Luo, Z. Yu, M. Jiang, J. Wen, G. P. Gupta, P. Giusti, H. Zhu, Y. Yang, et al. A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Briefings in Bioinformatics*, 23(4):bbac245, 2022. doi: 10.1093/bib/bbac245. URL https://doi.org/10.1093/bib/bbac245.

J. Chen, T. Luo, M. Jiang, J. Liu, G. P. Gupta, and Y. Li. Cell composition inference and identification of layer-specific spatial transcriptional profiles with polaris. *Science Advances*, 9 (9):eadd9818, 2023. doi: 10.1126/sciadv.add9818. URL https://doi.org/10.1126/sciadv.add9818.

J. Chen, M. Zhou, W. Wu, J. Zhang, Y. Li, and D. Li. Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. *arXiv preprint*, jun 2024. URL https://arxiv.org/abs/2406.06393. PMID: 38947920; PMCID: PMC11213178.

K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, apr 2015. doi: 10.1126/science.aaa6090. URL https://doi.org/10.1126/science.aaa6090.

Angèle Coutant and et al. Spatial transcriptomics reveal pitfalls and opportunities for the detection of rare high-plasticity breast cancer subtypes. *Laboratory Investigation*, 103(12):100258, 2023.

Sebastian Dawo, Kalin Nonchev, and Karina Silina. 10x visium spatial transcriptomics dataset: Kidney (3) and lung (5) cancer with tertiary lymphoid structures. https://zenodo.org/records/14620362, 2023.

Ke Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1):1739, Apr 2022. doi: 10.1038/s41467-022-29439-6. URL https://doi.org/10.1038/s41467-022-29439-6.

Zhen Fan, Runsheng Chen, and Xiaowei Chen. Spatialdb: a database for spatially resolved transcriptomes. *Nucleic Acids Research*, 48(D1):D233–D237, jan 2020. doi: 10.1093/nar/gkz934. URL https://doi.org/10.1093/nar/gkz934.

Yanghong Guo, Bencong Zhu, Chen Tang, Ruichen Rong, Ying Ma, Guanghua Xiao, Lin Xu, and Qiwei Li. BayeSMART: Bayesian clustering of multi-sample spatially resolved transcriptomics data. *Briefings in Bioinformatics*, 25(6):bbae524, Nov 2024. doi: 10.1093/bib/bbae524. URL https://doi.org/10.1093/bib/bbae524.

Yuhan Hao, Tim Stuart, Michael H Kowalski, Sagar Choudhary, Paul Hoffman, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023. doi: 10.1038/s41587-023-01767-y. URL https://doi.org/10.1038/s41587-023-01767-y.

Xin Jiang, Shuang Wang, Li Guo, and et al. iimpact: integrating image and molecular profiles for spatial transcriptomics analysis. *Genome Biology*, 25:147, 2024. doi: 10.1186/s13059-024-03289-5. URL https://doi.org/10.1186/s13059-024-03289-5.

Ilya Korsunsky, Neil Millard, Jian Fan, and et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16:1289–1296, 2019. doi: 10.1038/s41592-019-0619-0. URL https://doi.org/10.1038/s41592-019-0619-0.

Huimin Li, Bencong Zhu, Xi Jiang, Lei Guo, Yang Xie, Lin Xu, and Qiwei Li. An interpretable bayesian clustering approach with feature selection for analyzing spatially resolved transcriptomics data. *Biometrics*, 80(3):ujae066, September 2024. doi: 10.1093/biomtc/ujae066. URL https://doi.org/10.1093/biomtc/ujae066.

Zhi Li and Xiang Zhou. Bass: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biology*, 23:168, 2022. doi: 10.1186/s13059-022-02734-7. URL https://doi.org/10.1186/s13059-022-02734-7.

T. Luo, J. Chen, W. Wu, J. Zhao, H. Yao, H. Zhu, and Y. Li. Mast-decon: Smooth cell-type deconvolution method for spatial transcriptomics data. *bioRxiv*, pp. 2024–05, 2024. doi: 10.1101/2024.05.03.592867. URL https://doi.org/10.1101/2024.05.03.592867.

K.R. Maynard, L. Collado-Torres, L.M. Weber, and et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24:425–436, 2021. doi: 10.1038/s41593-020-00787-0.

Davis J. McCarthy, Kieran R. Campbell, Aaron T. L. Lun, and Quin F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33 (8):1179–1186, April 2017. doi: 10.1093/bioinformatics/btw777. URL https://doi.org/10.1093/bioinformatics/btw777.

Maxime Meylan and et al. Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. *Immunity*, 55(3):527–541.e5, 2022.

Neil Millard, Jhen Hwa Chen, Mukta G. Palshikar, and et al. Batch correcting single-cell spatial transcriptomics count data with crescendo improves visualization and detection of spatial gene patterns. *Genome Biology*, 26:36, 2025. doi: 10.1186/s13059-025-03479-9. URL https://doi.org/10.1186/s13059-025-03479-9.

Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3):R25, 2010. doi: 10.1186/gb-2010-11-3-r25. URL https://doi.org/10.1186/gb-2010-11-3-r25.

P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, jul 2016. doi: 10.1126/science.aaf2403. URL https://doi.org/10.1126/science.aaf2403.

Patrik L. Ståhl and et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353:78–82, 2016. doi: 10.1126/science.aaf2403.

L. Tian, F. Chen, and E. Z. Macosko. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6):773–782, jun 2023. doi: 10.1038/s41587-023-01792-0. URL https://doi.org/10.1038/s41587-023-01792-0.

A. Valdeolivas, B. Amberg, N. Giroud, and et al. Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. *npj Precision Oncology*, 8:10, 2024. doi: 10.1038/s41698-023-00488-4.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, jul 2018. doi: 10.1126/science.aat5691. URL https://doi.org/10.1126/science.aat5691.

X. Wang, D. Venet, F. Lifrange, and et al. Spatial transcriptomics reveals substantial heterogeneity in triple-negative breast cancer with potential clinical implications. *Nature Communications*, 15: 10232, 2024. doi: 10.1038/s41467-024-54145-w.

Fabian A Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0. URL https://doi.org/10.1186/s13059-017-1382-0.

S.Z. Wu, G. Al-Eryani, D.L. Roden, and et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53:1334–1347, 2021. doi: 10.1038/s41588-021-00911-1.

Z. Xu, W. Wang, T. Yang, J. Chen, Y. Huang, J. Gould, W. Du, F. Yang, L. Li, T. Lai, et al. Stomicsdb: a database of spatial transcriptomic data. *bioRxiv*, pp. 2022–03, 2022. doi: 10.1101/2022.03.10.483747. URL https://doi.org/10.1101/2022.03.10.483747.

Z. Yuan, W. Pan, X. Zhao, F. Zhao, Z. Xu, X. Li, Y. Zhao, M. Q. Zhang, and J. Yao. Sodb facilitates comprehensive exploration of spatial omics data. *Nature Methods*, 20(3):387–399, mar 2023. doi: 10.1038/s41592-022-01756-4. URL https://doi.org/10.1038/s41592-022-01756-4.

Emma Zhao, Matthew R. Stone, Xin Ren, and et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, 39:1375–1384, 2021. doi: 10.1038/s41587-021-00935-2. URL https://doi.org/10.1038/s41587-021-00935-2.

Y. Zheng, Y. Chen, X. Ding, K. H. Wong, and E. Cheung. Aquila: a spatial omics database and analysis platform. *Nucleic Acids Research*, 51(D1):D827–D834, jan 2023. doi: 10.1093/nar/gkac972. URL https://doi.org/10.1093/nar/gkac874.

W. Zhou, M. Su, T. Jiang, Q. Yang, Q. Sun, K. Xu, J. Shi, C. Yang, N. Ding, Y. Li, et al. Sorc: an integrated spatial omics resource in cancer. *Nucleic Acids Research*, 52(D1):D1429–D1437, jan 2024. doi: 10.1093/nar/gkad892. URL https://doi.org/10.1093/nar/gkad820.

Jingshu Zhu, Shiquan Sun, and Xiang Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22:184, 2021. doi: 10.1186/s13059-021-02404-0. URL https://doi.org/10.1186/s13059-021-02404-0.