PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Sparse deconvolution of cell type medleys in spatial transcriptomics

Nuray Sogunmez Erdogan[1*], **Deniz Eroglu**[1,2]

1 Faculty of Natural Sciences and Engineering, Kadir Has University, Istanbul, Turkiye, **2** Department of Mathematics, Imperial College London, London, United Kingdom

* nuray.erdogan@khas.edu.tr

OPEN ACCESS

Citation: Erdogan NS, Eroglu D (2025) Sparse deconvolution of cell type medleys in spatial transcriptomics. PLoS Comput Biol 21(6): e1013169. https://doi.org/10.1371/journal.pcbi.1013169

Editor: Ilya Ioshikhes, CANADA

Received: April 24, 2025

Accepted: May 27, 2025

Published: June 12, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1013169

Data availability statement: The scRNA-seq and spatial transcriptomics data for human developing heart data can be obtained from https://www.spatialresearch.org/resources-published-datasets/doi-10-1016-j-cell-2019-11-025/. The Visium data for the mouse brain can

Abstract

Mapping cell distributions across spatial locations with whole-genome coverage is essential for understanding cellular responses and signaling However, current deconvolution models aim to estimate the proportions of distinct cell types in each spatial transcriptomics spot by integrating reference single-cell data. These models often assume strong overlap between the reference and spatial datasets, neglecting biology-grounded constraints such as sparsity and cell-type variations, as well as technical sparsity. As a result, these methods rely on over-permissive algorithms that ignore given constraints leading to inaccurate predictions, particularly in heterogeneous or unmatched datasets. We introduce Weight-Induced Sparse Regression (WISpR), a machine learning algorithm that integrates spot-specific hyperparameters and sparsity-driven modeling. Unlike conventional approaches that neglect biology-grounded constraints, WISpR accurately predicts cell-type distributions while preserving biological coherence, i.e., spatially and functionally consistent cell-type localization, even in unmatched datasets. Benchmarking against five alternative methods across ten datasets, WISpR consistently outperformed competitors and predicted cellular landscapes in both normal and cancerous tissues. By leveraging sparse cell-type arrangements, WISpR provides biologically informed, high-resolution cellular maps. Its ability to decode tissue organization in both healthy and diseased states highlights WISpR's practical utility for spatial transcriptomics, particularly in challenging settings involving noise, sparsity, or reference mismatches.

Author summary

Life is like a puzzle: Each cell or gene is a piece that contributes to the big picture. To truly understand health and disease, we must not only know which components are present but also how they are spatially organized and interact. Tissues are composed of diverse cell types with distinct roles, and spatial transcriptomics allows us to map gene activity across tissue sections. However, due to limited resolution, each measurement often contains signals from multiple cell types. In contrast, single-cell RNA sequencing (scRNA-Seq) offers cell-level resolution, but lacks spatial context, as having puzzle pieces without the guiding image. To bridge this gap, deconvolution methods aim to

be obtained from https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Adult_Mouse_Brain. The scRNA-Seq data for mouse brain L1 hippocampus can be obtained from https://storage.googleapis.com/linnarsson-lab-loom/l1_hippocampus.loom and the cells in each cell type are filtered as cells ≤ 25 and cells ≥ 250. The scRNA-Seq for breast cancer patient data can be downloaded from GSE176078 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176078 and the matching spatial data can be downloaded from https://zenodo.org/record/4739739. The implementation of the WISpR method and its applications are publicly accessible on Zenodo under the DOI: 10.5281/zenodo.11109636.

Funding: Work by N.S.E was supported by TUBITAK (Grant No. 222S096) and TUSEB (Grant No. 40026). Work by D.E. was partially supported by TUBITAK (Grant No. 118C236), UKRI (EP/Z002656/1), and the BAGEP Award of the Science Academy. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

infer which cell types are present at each spatial location. Yet, using cross-study scRNA-Seq references can introduce mismatches due to batch effects, incomplete annotations, or biological variability. We developed WISpR (Weight-Induced Sparse Regression), a mathematically robust deconvolution framework that selectively extracts only the most informative cell types at each spot. WISpR is resilient to noisy, sparse, or mismatched data, outperforming existing tools. Applied to simulated and real datasets, including developing heart, brain, and tumors, WISpR reveals meaningful biological patterns and enhances our understanding of tissue architecture and cellular heterogeneity in health and disease.

Introduction

Biological coherence, the harmonious integration of cellular components and their interactions, underpins the complex organization and functionality of living systems. In nature, species form organizations that serve to perpetuate their lineage or enhance the efficiency of community tasks while minimizing energy consumption at different levels of life. Similarly, at the genetic level, coherence in region-specific gene expression allows cells to perform specialized roles, contributing to tissue complexity and organismal physiology [1–6]. This coherence is fundamental to ensuring that cells operate in concert rather than in isolation. Disruptions in coherence can lead to pathological states, emphasizing its role as a key organizing principle in both health and disease. Despite this, variability within cell types introduces *cellular heterogeneity*, a key feature of complex tissues. Our understanding of these complex biological systems depends on our ability to accurately perform deconvolution, the estimations of underlying cell-type compositions, from observations. Yet, this coherence and heterogeneity dilemma becomes even more noticeable when the datasets originate from different sources. In such cases, the reference profiles i.e., transcriptomic profiles of cell types may not adequately represent the cell types present in the spatial transcriptomics data, as the profiles of similar cell types can vary due to biological or temporal changes. However, the global goal is to establish a comprehensive collection of cellular transcriptomics profiles that can be used to map all possible cells in tissues and accurately detect disease states and locations. Therefore, accurately deconvoluting cell types using large-scale single-cell reference datasets derived from diverse studies and understanding their interactions and organizations in spatial contexts is crucial for unraveling cellular heterogeneity and tissue complexity. Consequently, understanding the cellular landscapes and interactions opens up possibilities for deciphering complex biological processes such as cellular development, metabolism, and signal transduction [7–9] and diseases such as cancer, neurological disorders, and others [10–12], which are manifestations of cellular responses and are orchestrated through intricate signaling pathways.

Spatial transcriptomics, refers to technologies that measure gene expression in spatial context by assigning expression profiles to defined regions on a tissue section, known as capture spots. Each spot captures transcripts from multiple nearby cells, allowing the identification of tissue regions with distinct transcriptional signatures. However, capture spots may contain multiple cell types or fractions of cells (Visium, Visium-HD, Tomo-Seq, Slide-Seq) [13–16], thus require deconvolution, to recover the original cells that were "mixed" together, to estimate cell-type proportions. On the other hand, high resolution spatial expression data with high dropout rates, prior requirement of gene panels, producing sparse count matrices (as seen in MERFISH, Stereo-Seq, Seq-Scope, PIXEL-Seq) [17–20], still lack direct cell-type identity and rely on segmentation-based or reverse deconvolution methods for detecting true cell-type identification. Consequently, directly extracting the whole genome coverage at single-cell

resolution is impossible with current spatial transcriptomics technologies. On the other hand, *single-cell RNA sequencing* (scRNA-Seq) technology, which profiles gene expression at single-cell resolution through running high-throughput sequencing on isolated single cells, has revolutionized the study of cellular heterogeneity and gene expression at single-cell resolution, allowing examination of cell-specific transcriptomes within tissues. Despite potential confounders and biological variations [21–23], scRNA-Seq has revealed previously inaccessible insights into cellular diversity and function [24–27]. However, scRNA-Seq lacks spatial information on the original tissue locations of distinct cell types, defined here as groups of cells with shared gene expression profiles and functional roles. Hence, a combinatorial deconvolution approach using the cell-type information from scRNA-Seq and the spatial information from spatial transcriptomics is a natural solution to the problem.

Accurate predictions of cell types from spatial capture spots require methods that incorporate constraints derived from biological facts, such as the sparsity of cell-type distributions and the non-negative nature of cell presence. Therefore, utilization of scRNA-Seq data as a reference for deconvolution, merely applying it without enforcing biologically grounded constraints—such as spot-specific sparsity, cell-type rejection, or non-negativity—often leads to biologically implausible cell distributions. Existing deconvolution tools aim to integrate scRNA-Seq and spatial transcriptomics to create tissue atlases [28–33].

However, many existing methods either neglect biology-grounded constraints, often assigning non-zero contributions to most cell types across spatial locations or either consider broad, tissue-level sparsity or not consider it at all, which overlooks spot-specific variation and fail to capture localized cellular exclusivity as well as dilute meaningful signals [28,32,34]. This can lead to biologically implausible interpretations, obscuring spatial organization and hindering the discovery of disease-relevant cellular interactions. Therefore, finding an optimal and biologically reliable deconvolution algorithm for accurate mapping of cell types in spatial transcriptomics data remains an open problem.

The challenge lies in developing a computational tool that effectively aligns spatial transcriptomics with scRNA-Seq data while accounting for biology-grounded constraints to filter out spurious cell-type contributions during deconvolution. Deconvolution, a mathematical technique used to recover the original signal or components that were "mixed" together in an observed measurement, can be approached using regression methods to solve the linear equation $\boldsymbol{y} = \boldsymbol{Ax}$, where $\boldsymbol{y}$ is the gene expression vector from spatial data (capture spot), $\boldsymbol{A}$ is the gene expression matrix for cell types, and $\boldsymbol{x}$ is the cell-type coefficients predicted by solving this equation [31,32,35,36]. However, the high coherence of gene expressions among different cell types can pose difficulties for traditional regression approaches, as the column vectors in matrix $\boldsymbol{A}$ may be very similar. Attempts are made to overcome the problem by relying on correlation-based thresholding to select cell-type marker genes, assuming that these gene sets are highly specific to certain cell types [37].

Yet, in spatial transcriptomics deconvolution, the linear system $\boldsymbol{y} = \boldsymbol{Ax}$ is inherently ill-posed due to strong correlations among the columns of $\boldsymbol{A}$ (cell-type-specific gene expression profiles), noise in both the reference and spatial data, and limited gene coverage (i.e., being an overdetermined system). These challenges prevent the derivation of a unique and stable solution without additional constraints. Given the nature of our problem, the solution is expected to identify only relevant and distinguishable cell types, which are a few in reality, per spot. Therefore, sparse recovery methods (also known as compressed sensing or basis pursuit) emerge as clear candidates.

Sparse deconvolution aims to identify the most relevant cell types contributing to each spatial transcriptomics spot by eliminating those with weak or biologically unsupported signals. This approach reflects experimental evidence from solid tissues, where spatially confined

regions typically harbor only a few distinct cell types. As such, sparse deconvolution is especially well-suited for structured tissues with region-specific cellular architecture, enabling more interpretable and biologically coherent results.

Pitfalls in reference-based deconvolution

Accurate prediction of cell types in spatial transcriptomics data (deconvolution) is essential to understand the spatial organization of cells that govern tissue functions. Today's technology provides a widely adopted and transformative technique called scRNA-Seq, which profiles cellular gene expressions and opens the possibility of applying deconvolution from scRNA-Seq.

Deconvolution in cell-type mapping task is commonly formulated as a regression problem, where the gene expression profile of each spatial spot is modeled as a weighted combination of reference expression profiles from each cell type, with the weights representing estimated cell-type proportions [38].

Since scRNA-Seq data provides gene expression profiles at the single-cell level for all cells detected, these cells must first be clustered into distinct cell types [39–41] and their corresponding marker genes identified. Clustering approaches explore similarities and differences in multivariate data and draw inferences from unlabeled data, resulting in cell populations sharing similar characteristics compared to other data members. For example, 3,717 cells from scRNA-Seq data of a developing human heart were clustered into 15 cell types [42] and visualized in a 2D space with uniform manifold approximation and projection (UMAP) [43], which is a nonlinear dimensionality reduction technique that preserves local and global structure to reveal meaningful patterns in high-dimensional data (Fig 1A). However, deconvolution of these cell types is non-trivial, since the standard regression methods are insufficient for the desired reconstruction due to the following two main facts:

- *Cellular heterogeneity within a single cell type*. Cell-to-cell variability within the same cell type reflects how biological systems are regulated, respond to external influences, and develop over time. The heterogeneous nature of cells in a single cluster is illustrated for Cluster-0 (Fig 1B). The distance from the centroid (stars) to the maximum distance to the convex hull of the cluster (black line) illustrates the cellular dispersion σ_0. Thus, the identification of DEGs, the expression values of which represent all cells found in Cluster-0, is not trivial to select accurate representative gene expressions (Fig 1B).
- *Genetic coherence in differently annotated cell types (i.e., cell-type similarity)*. Along with cellular heterogeneity in clusters, it is likely to observe coherent gene expressions in separately annotated cell types due to their biological or functional similarities (i.e., cellular subtypes of the same tissue). For instance, distinct cell types, such as Clusters-2, 3, and 4, are nested in the 2D UMAP projection due to cellular heterogeneity (Fig 1C). In other words, while σ_i, where $i = 2, 3$, and 4, represents the heterogeneity for Clusters-2, 3, and 4, the overlapping circles with radius σ_is illustrate the similarities. A pairwise Pearson correlation between the expression vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\rho_{\boldsymbol{x},\boldsymbol{y}}$ (see the Methods section for the Pearson correlation formula), quantitatively measures the similarities between DEGs (Fig 1D). This result reveals how the transcript levels of the expression vectors for different cell types, such as cluster 3 and cluster 4 ($\rho_{3,4} = 0.91$), or cluster 1 and cluster 12 ($\rho_{1,12} = 0.94$), strongly correlate with each other, declaring possible reasons for the failure of deconvolution.

Therefore, both intra-cluster heterogeneity and inter-cluster coherence can yield computationally valid but biologically misleading regression fits.

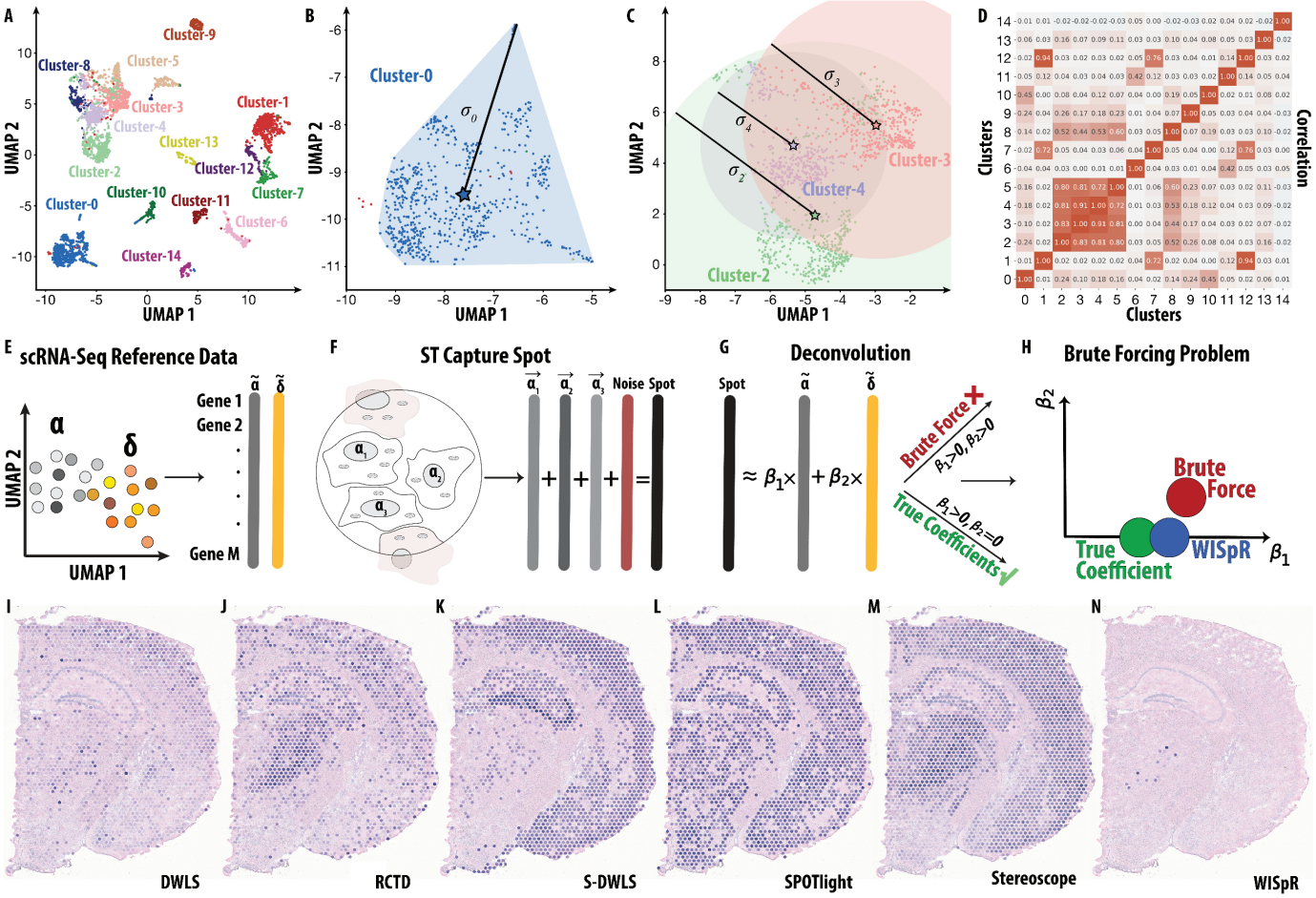

Fig 1. Deconvolution challenges due to coherence and heterogeneity. **A**, Two-dimensional UMAP representation of scRNA-Seq data from the developing human heart. Each color denotes cell types (clusters), revealing local and global structures based on their similarity. **B**, Illustration of high cellular heterogeneity within Cluster-0, represented by the line connecting the centroid (star) of Cluster-0 to the cell at the maximum distance. **C**, Overlapping clusters (Cluster-2 in green, Cluster-3 in orange, and Cluster-4 in blue) due to cell type similarity, with cluster radii calculated using centroid distances and cell locations at maximum distances. Large intersecting areas indicate similarity between cell types. **D**, Heatmap of Pearson's correlations among clusters, with increased red intensity signifying higher correlations between functionally coherent cell types. The correlation values between clusters are displayed. **E**, The 2D UMAP plot of scRNA-Seq reference data showing cellular clusters from two cell types α (grey) and δ (orange) as separate clusters with their DEG vectors $\tilde{\alpha}$ and $\tilde{\delta}$. **F**, A spatial transcriptomics capture spot containing multiple cells from one cell type α_1, α_2, and α_3 (grey) and background noise (brick-red), illustrating the complexity of the spot's transcriptomic profile. **G**, General deconvolution strategy using scRNA-Seq reference vectors to infer cell-type contributions in capture spots. Conventional over-permissive methods assign non-negative coefficients, β_1 and β_2, to both cell types α and δ even when only α is present, leading to inaccurate, non-sparse solutions. **H**, Illustration of noise tolerance in deconvolution. Over-permissive tools distribute coefficients across both cell types to account for noise, while WISpR uses thresholding to eliminate the contribution of δ and recover the true sparse coefficient for α, enhancing deconvolution accuracy. **I–N**, The reference scRNA-Seq data from the developing human embryonic heart is used to assess deconvolution approaches for distinguishing human heart-specific cell types from a mouse brain spatial transcriptomics dataset. Model predictions for atrial cardiomyocyte cell types in the mouse brain are presented. **I**, In DWLS prediction, human atrial cardiomyocytes show low percentages in mouse brain spots, with occurrences observed in cortical layers and thalamus. **J**, RCTD prediction reveals high percentages of cells in spots located in hippocampal, thalamic, and cortical regions. **K**, S-DWLS prediction, similar to RCTD, displays high percentages of cells in spots concentrated in cortical and hippocampal regions. **L**, SPOTlight and **M**, Stereoscope predicts a high abundance of cells across various brain zones, while **N**, WISpR predicts a minimal number of spots in the thalamus, supporting its more accurate prediction on mismatched datasets.

https://doi.org/10.1371/journal.pcbi.1013169.g001

Contemporary deconvolution techniques aiming to overcome these challenges can be broadly categorized as model-based or model-free, depending on how they approach the estimation of underlying components. Model-based methods rely on predefined assumptions or frameworks that describe how the observed data is generated, often incorporating prior

knowledge or reference profiles. In contrast, model-free methods make minimal assumptions about the data and estimate underlying components directly using optimization or statistical constraints, offering greater flexibility in diverse or noisy settings.

The SPOTlight algorithm [32] is a prominent model-based approach that uses seeded non-negative matrix factorization (NMF) to derive topic-like components from the reference data, followed by non-negative least squares regression to project them onto spatial locations. This strategy, while powerful, tends to distribute all components across all spots, even in regions where certain cell types are biologically implausible. Similarly, CARD [34] employs NMF combined with spatial kernel smoothing, which reinforces spatial continuity but can result in over-smoothing across anatomical boundaries. This can be problematic especially in tissues with sharp spatial transitions and oversee the cellular heterogeneity which is generally observed in immune cells and diseases as cancer. STRIDE [44] follows a related logic using latent Dirichlet allocation (LDA) instead of NMF, modeling cell types as latent topics. Although this introduces greater flexibility, it also reduces transparency, making the direct interpretation of results more difficult. Across these approaches, the reliance on latent decomposition techniques, whether NMF or LDA, can obscure biological interpretability and enforce unrealistic uniformity across space.

Probabilistic generative models like RCTD [28] and Stereoscope [30] represent another group of model-based methods. RCTD models cell-type mixtures using a Poisson likelihood, incorporating platform variability and overdispersion, while Stereoscope applies a negative binomial framework, estimating gene- and cell-type–specific parameters to infer relative frequencies. These models can be powerful when assumptions hold, but are vulnerable to reference mismatches, noise, and overfitting, especially in heterogeneous tissues. Cell2location [33] expands on this paradigm by embedding a Bayesian hierarchical framework, incorporating multiple priors to improve robustness to noise. Yet, like RCTD and Stereoscope, its performance still heavily depends on the quality and completeness of the reference, and it lacks mechanisms to discard irrelevant cell types. Redeconve [36] takes a semi-supervised approach by clustering individual reference cells into meta-groups to manage uncertainty of annotations and noise in reference data. However, this strategy can obscure fine-grained cell-type distinctions and reduce resolution, similar to latent component methods such as SPOTlight, STRIDE, and Cell2location. Like most deconvolution approaches, Redeconve lacks a mechanism for spot-specific rejection of irrelevant cell types, often leading to diffuse or biologically implausible predictions in heterogeneous tissues, similar to the over-smoothing observed in CARD and SpatialDWLS [31].

Among model-free methods, Dampened Weighted Least Squares (DWLS) [29] estimates cell-type proportions by weighted least-squares regression, introducing dampening to reduce the impact of dominant cell types. However, its weight formulation may introduce biologically implausible behavior (e.g., negative squared terms), and it was originally designed for bulk RNA-seq, making it computationally inefficient for large-scale spatial data. An extension of this method, SpatialDWLS [31], incorporates cell-type enrichment and spatial smoothness into the regression, but, like CARD, it may oversmooth biologically sharp features and, like most of the above methods, lacks any formal mechanism to reject unsupported or irrelevant cell types.

While these model-based and model-free approaches vary in formulation, ranging from matrix factorization (SPOTlight, CARD), topic models (STRIDE), and Bayesian generative models (RCTD, Stereoscope, Cell2location) to semi-supervised clustering (Redeconve) and constrained regression (DWLS, SpatialDWLS), they share key limitations: they tend to force-fit all reference cell types across all spatial locations, assume global applicability of the reference, and often smooth over true biological heterogeneity. These limitations manifest

most clearly in tissues with sharp anatomical transitions or high heterogeneity, where over-permissive regression or probabilistic inference can result in biologically implausible cell-type predictions. Even when methods incorporate spatial priors or smoothing, such as in CARD or SpatialDWLS, they rarely support spot-level rejection of irrelevant reference types. This highlights a general gap in current tools, the lack of built-in mechanisms to preserve spatial specificity, reject noise, and avoid reference overfitting.

This article presents the implementation of a machine learning-based algorithm called *Weight-Induced Sparse Regression (WISpR)*, which incorporates an intelligently configured sparsity property to eliminate irrelevant cells and spurious expressions as required. WISpR is specifically designed to estimate the most biologically plausible subset of cell types (i.e., the optimal sparse arrangement) present in each spatial transcriptomics spot, promoting biologically coherent sparsity by systematically rejecting unsupported cell types. In this optimal sparse arrangement, only cell types with meaningful transcriptomic support are retained, while spurious contributions are suppressed through a spot-wise thresholding parameter. This threshold is tuned via cross-validation to balance interpretability and prediction accuracy, avoiding both overfitting and underestimation. This improves specificity and helps prevent signal dilution, indirectly reducing false negatives. These tradeoffs are captured through F1 score evaluation, balancing precision and recall. (see the Materials and Methods section for details of the WISpR Model.)

WISpR addresses the limitations of existing methods by incorporating sparsity constraints and spot-specific hyperparameter and gene weight optimization, leading to predictions that align with biologically plausible cell-type distributions, even under conditions of noise or reference-target mismatches. To ensure a comprehensive and representative comparison, five deconvolution methods were used in this paper that exemplify the major algorithmic strategies used in spatial transcriptomics. SPOTlight [32] was included as an early and influential matrix factorization–based method that set the foundation for many NMF-driven approaches. DWLS [29] represents a canonical model-free regression-based technique known for its simplicity and general applicability, while SpatialDWLS [31] extends this paradigm by incorporating spatial priors and has been frequently cited as a benchmark for spatially aware deconvolution. RCTD [28] and Stereoscope [30] were chosen as representative probabilistic model-based frameworks, one relying on Poisson mixtures, and the other on a negative binomial generative model, both of which are commonly used for their robustness and statistical depth. Together, these methods span the dominant methodological families in spatial deconvolution (regression-based, matrix factorization, and generative modeling), enabling a fair and informative contextualization of performance and innovations of our newly developed model.

In evaluations, WISpR demonstrated superior accuracy and sensitivity in four matched scenarios (where the reference scRNA-Seq and target spatial transcriptomics datasets originate from the same source) and six unmatched scenarios (where the reference and target datasets originate from different sources), using synthetic data derived from two distinct organisms and tissue types. Furthermore, WISpR provided biologically meaningful information by accurately mapping cell-type proportions and localizations in diverse tissues, including the developing human heart and a coronal mouse brain section. Its ability to resolve abundant and rare cell populations demonstrates its utility in studying tissue complexity and cellular organization. Moreover, WISpR accurately mapped cell states in breast cancer tissues, revealing regional cancer subtypes and clarifying the location of LumB traits in the DCIS regions. This precision highlights the potential of WISpR to uncover tumor heterogeneity, advancing our understanding of cancer progression and resistance mechanisms. By excelling in both synthetic benchmarks and real biological applications, WISpR sets a new standard

for deconvolution, providing a versatile and reliable tool for studying tissue architecture and cellular dynamics in diverse contexts.

Application-driven hurdles in reference-based deconvolution

The challenges associated with spatial transcriptomics deconvolution when using the scRNA-Seq data as a reference is illustrated with a conceptual workflow in Fig 1E–1H. The 2D UMAP representation of the reference data (Fig 1E) highlights the heterogeneous cluster of cell types annotated as α and δ as separate colors, and their DEG vectors $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\delta}}$ with M number of genes, and N and L number of cells, where $\tilde{\boldsymbol{\alpha}} = f(\boldsymbol{\alpha_i})$ for $i = 1, 2, .., N$, and $\tilde{\boldsymbol{\delta}} = f(\boldsymbol{\delta_i})$ for $i = 1, 2, .., L$, respectively. The function f is an operation to find the representative vectors $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\delta}}$, which is the mean value of each gene type within the cluster for our study. Furthermore, an example of a capture spot in spatial transcriptomics is given in Fig 1F, where the spot is composed of three cells of a single cell type α along with contributions from background noise. The transcriptomic profile with M genes of the spot is thus a mixture of cells $\boldsymbol{\alpha_i}$, where $i = 1, 2, 3$ in this example, and noise, representing the measurement errors and fraction of cells partially contributing to the transcriptome profile of the spot. The general deconvolution strategy uses representative vectors, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\delta}}$, from the scRNA-Seq data to infer the cellular composition of the spatial capture spot. This approach relies on over-permissive methods, assigning non-negative coefficients β_1 and β_2, to both cell types α and δ, respectively, even when only one cell type, α, is present in reality (Fig 1G). This results in inaccurate predictions, as these methods fail to account for the sparsity of the true cellular composition. As a result, the concept of noise tolerance in deconvolution is required to predict the true coefficients.

Over-permissive tools attempt to explain the noise by indiscriminately assigning coefficients for almost all cell types given in reference data, which incorrectly represent different cell types, thereby introducing biologically implausible predictions that compromise interpretability. In contrast, WISpR is tailored to employ a thresholding mechanism to eliminate irrelevant cell types by assigning $\beta_2 = 0$, and accurately assign sparse coefficients, $\beta_1 > 0$, thus improving the precision of deconvolution in spatial data (Fig 1H).

The real-life task is given to predict human heart cell types in a mouse brain tissue dataset. This setup, while representing an extreme case of dataset mismatch, highlights a fundamental criterion for algorithmic reliability: the ability to identify relevant cell types while discarding irrelevant ones. Here, probing atrial cardiomyocytes in a developing human heart (Cluster-7 in Fig 1A) tested with alternative deconvolution methods (Fig 1I–1M, Fig A in S1 Text), along with our model (Fig 1N), to reveal incorrectly reconstructed mouse brain cells (10X Visium array) using human heart data [42], highlighting potential inaccuracies in disease studies where precise cell localization is critical. This outcome underscores a critical limitation: when tasked with even a straightforward dataset separation, these algorithms fail to exclude cell types absent from the target dataset.

While the given conceptual framework highlights the importance of accurate deconvolution, it initially assumes a highly non-ideal scenario where cell-type references have almost no match with the target tissue. This deliberate challenge was designed to test the robustness of alternative models under extreme conditions, using datasets from different organs and organisms. Although such mismatches are rare in biological practice, these tests are crucial for assessing model limitations. To complement this, we adopted a more biologically relevant approach, using adult mouse hippocampal scRNA-Seq data as a reference ([45]) (Fig 2A–2C) for deconvoluting hippocampus-only cells in postnatal day 8 (P8) [46] and adult mouse brain (10X Visium array).

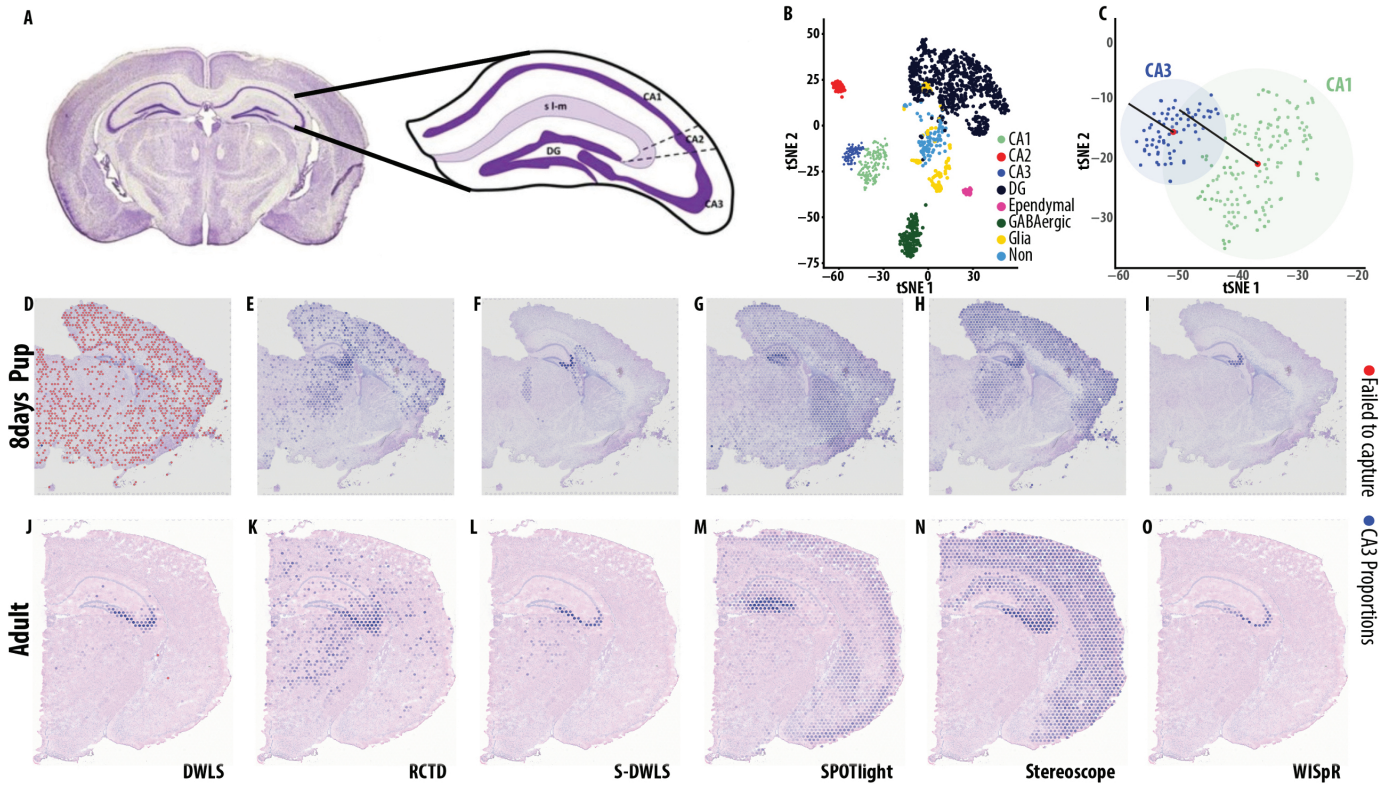

Fig 2. Overcoming limitations in sparse cell-type predictions for a biologically reliable task. A, Illustration of mouse hippocampus on mouse brain coronal section. **B**, 2D UMAP representation of the 8 HPF cell types. **C**, Overlapping CA1 and CA3 due to cell type similarity, with cluster radii calculated using centroid distances and cell locations at maximum distances. **D-I**, Model predictions for CA3 in the P8 and **J-O** in adult mouse brain are presented. **D, J**, In DWLS prediction, CA3 cells are slightly over-represented in mouse brain spots (blue shade), with frequent spot prediction failures (red). **E, K**, RCTD prediction reveals high percentages of cells in spots located in hippocampal, thalamic and cortical regions. **F, L**, S-DWLS prediction displays confined but still disproportionate of cells in spots concentrated in stratum oriens, stratum pyramidale and stratum lucidum and thalamus regions. **G, M**, SPOTlight predicts a high abundance of CA3 cells in DG and across various brain zones including cortex. **H, N**, Stereoscope predicts CA3 in almost all hippocampus, thalamus and cortical regions, while **I, O** WISpR predicts CA3 cells in exactly CA3 morphological region in both P8 and adult mice, supporting its more accurate prediction on mismatched datasets.

https://doi.org/10.1371/journal.pcbi.1013169.g002

The objective is to ensure that reference hippocampal cells are located exclusively within the hippocampus, with no presence detected in other regions. This dual approach allowed us to rigorously evaluate the performance of the model in both extreme and realistic biological scenarios, consistently demonstrating the limitations of alternative models. An example is given here by using CA3 cells (Fig 2B,2C). Similar to previous case, all tested alternative deconvolution methods (Fig 2D–2H for P8 mouse, Fig 2J–2N for adult mouse, and Fig B in S1 text), but WISpR (Fig 2I for P8 mouse and Fig 2O for adult mouse), over-reconstructed mouse brain cells using mouse hippocampus data. This outcome reveals that these algorithms are ineffective in removing absent cell types from the target data and tend to exaggerate the abundance of those present. Consequently, this limitation underscores the reliability of alternative tools in real-world applications where the ground truth may be unknown, and their tendency to over-represent present cell types while failing to exclude irrelevant ones could compromise accurate deconvolution. Therefore, methods that inherently sense biological phenomena, such as sparsity, intrinsically are needed to ensure deconvolution accuracy and to produce highly accurate high-resolution organ maps.

Materials and methods

Similarity analysis

Similarity scores via linear correlations involve calculating the pairwise Pearson's correlation coefficient between gene expression profiles of all cell types. The Pearson correlation coefficient ($\rho_{x,y}$) of gene expression profile vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is determined by the formula:

$$\rho_{x,y} = \frac{n\sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{\sqrt{n\sum_i^n x_i{}^2 - (\sum_i^n x_i)^2}\sqrt{n\sum_i^n y_i{}^2 - (\sum_i^n y_i)^2}} \tag{1}$$

where $\boldsymbol{x} = (x_i), \boldsymbol{y} = (y_i) \in \mathbb{N}^n$ are n-dimensional vectors, and $i = 1, \ldots, n$. The Pearson correlation yields a number from -1 (indicating a perfect negative correlation) to 1 (indicating a perfect positive correlation), with 0 denoting no correlation.

Spatial clustering of zones

To define transcriptionally coherent spatial regions (zones), an unsupervised clustering procedure was applied to the spatial transcriptomics expression matrix. Let $\boldsymbol{y} \in \mathbb{R}^{MXL}$ denote the raw gene expression matrix, where M is the number of genes and L is the number of spatial spots. The matrix $\boldsymbol{y}$ was normalized using the normalizeGiotto() function from the Giotto package [47] to correct for technical variation. Highly variable genes (HVGs) were selected from the normalized matrix by computing gene-wise variance across all spots using calculateHVG(). A subset $G_{HVG} \subseteq \{1, \ldots, M\}$ was retained for downstream dimensionality reduction. Principal Component Analysis (PCA) was performed that yielded a low-dimensional embedding $Z \in \mathbb{R}^{L\times d}$, with $d = 10$ components.

A k-nearest neighbor (k-NN) graph $\mathcal{N}$ was constructed using the Euclidean distances in $\mathcal{Z}$, where $k = 10$. Community detection was subsequently performed on $\mathcal{N}$ using the Leiden algorithm using a resolution parameter of 0.2 and 1000 iterations, as implemented in doLeidenCluster(). This resulted in a discrete partitioning of spatial spots into K transcriptionally coherent zones, denoted $\{\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_K\}$.

These spatial zones were then used to compute zone-specific DEGs across the clusters. The sets of resulting genes were integrated with the specific DEGs of cell types of scRNA-Seq to construct the reference matrix $\mathbf{X} \in \mathbb{R}^{M\times Q}$ for deconvolution.

Selection of differentially expressed genes

Differentially Expressed Genes are identified using an algorithm based on the Gini index [47–49], whose robustness to gene selection has been demonstrated [50]. In this context, we refer to the Gini index as the overall method for evaluating expression inequality, and to the Gini coefficient as the specific numerical value computed for each gene. The Gini coefficient G calculated as the mean absolute divergence among all pairs of data points within the provided dataset. In cases of consistent measurements, the Gini coefficient approaches its minimum value of 0. In heterogeneous cases, the index tends towards the theoretical maximum of 1, indicating the highest inequality [51,52]. The Gini coefficient for a gene is calculated as:

$$G_\gamma = \frac{\sum_{i=0}^n \sum_{j=0}^n |\gamma_i - \gamma_j|}{2n^2\bar{\gamma}} \tag{2}$$

where γ_i and γ_j are the average expression values of gene γ in cell types i and j, n is the number of cell types, and $\bar{\gamma}$ is the overall mean expression of gene γ across all cell types.

To identify robust DEGs, the algorithm first computes the average expression values and detection rates (i.e., the proportion of cells where a gene is expressed) of each gene within each cell type. A lower detection threshold of 0.2 is applied, meaning a gene must be detected in at least 20% of the cells of a given type to be considered further. The Gini coefficients for both average expression and detection rate are then computed for each gene across all cell types to quantify inequality in their distributions.

Next, all genes are independently ranked according to their average expression levels and detection rates across cell types. Specifically:

- Expr. Rank(γ) reflects the rank of gene γ among all genes based on its average expression values across cell types.
- Detect. Rank(γ) reflects the rank of gene γ based on its distribution of detection rate (i.e., the percentage of cells that the gene γ is detected in) across cell types.

These rankings are then combined with the corresponding Gini coefficients to calculate an integrated score (IS) for each gene:

$$\text{IS}(\gamma) = \text{Detect. Rank}(\gamma) \times \text{Expr. Rank}(\gamma) \times \text{Expr. GC}(\gamma) \times \text{Detect. GC}(\gamma) \tag{3}$$

where Expr. GC(γ) and Detect. GC(γ) are the Gini coefficients for average expression and detection, respectively.

Genes within each cell type γ_i for $i = 1, \ldots, n$, where n is the number of cell types, ranked according to their gene-specific expression and detection ranking, and following the approach of Dries et al., (2021) [47] the top 100 genes with the highest values specific to each cell types are used to construct the DEG matrix used for downstream deconvolution.

WISpR model

WISpR effectively obtains a reliable solution to a complex challenge of deconvolution. This problem is formally expressed through a linear equation,

$$\boldsymbol{y} = \boldsymbol{X}\beta, \tag{4}$$

where $\boldsymbol{X} \in \mathbb{N}^{M\times Q}$, $\boldsymbol{y} \in \mathbb{N}^{M\times L}$ and $\beta \in \mathbb{N}^{Q\times L}$, where M is the number of genes, Q is the number of cell types, L is the number of capture spots in the spatial transcriptomics data and $\mathbb{N}$ is the set of natural numbers. Within this equation, $\boldsymbol{y}$ is the gene expressions matrix, and each column of $\boldsymbol{y}$ represents a particular spatial transcriptomic location (that is, a spot), while $\boldsymbol{X}$ encompasses gene expressions derived from diverse cell types as column vectors, obtained from scRNA-Seq data. The purpose of β is to capture coefficients that clarify the extent to which cell types within $\boldsymbol{X}$ influence the spatial transcriptomics data represented by $\boldsymbol{y}$. Since the total number of DEGs used in the deconvolution exceeds the total number of cell types identified from the scRNA-Seq data ($M > Q$), the deconvolution problem exhibits non-uniqueness in its solutions. In other words, there are infinitely many solutions for Eq 4. Hence, the challenge lies in discerning the correct solution from the multitude of possibilities.

The fundamental assumption in the deconvolution approach is that one has access to all the necessary scRNA-Seq data for reconstructing spatial transcriptomics. Suppose that spatial transcriptomics data can be represented as linear combinations of a given library $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_Q\}$ of scRNA-Seq data, with $\boldsymbol{X}_i \in \mathbb{N}^M$. For each spot i, the objective is to determine a coefficient vector $\beta_i = (\beta_{1,i}, \ldots, \beta_{Q,i})$ that can express the i-th component of spatial data $\boldsymbol{y}_i$ as a linear combination of $\boldsymbol{X}$, such that $\boldsymbol{y}_i \approx \sum_j \beta_{j,i}\boldsymbol{X}_j$. In a matrix representation, we are seeking a coefficient matrix β so that $\boldsymbol{y} \approx \boldsymbol{X}\beta$, where:

$$\boldsymbol{y} = \begin{pmatrix} \mathrm{DEG}_{1,1} & \dots & \mathrm{DEG}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathrm{DEG}_{M,1} & \dots & \mathrm{DEG}_{M,L} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} X_{1,1} & \dots & X_{1,Q} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \dots & X_{M,Q} \end{pmatrix}, \beta = \begin{pmatrix} \beta_{1,1} & \dots & \beta_{1,L} \\ \vdots & \ddots & \vdots \\ \beta_{Q,1} & \dots & \beta_{Q,L} \end{pmatrix}.$$

Deconvolution problems are concerned with situations where the number of DEGs, denoted as M, exceeds the number of cell types, represented by Q, resulting in an "over-determined" linear equation $\boldsymbol{y} = \boldsymbol{X}\beta$. However, considering the inherent sparsity of spatial transcriptomics data, our actual objective is to find a solution for the following minimization problem.

$$\min_{\beta \in \mathbb{N}^{Q \times L}} ||\beta||_0 \quad \text{subject to} \quad \boldsymbol{w}||\boldsymbol{y} - \boldsymbol{X}\beta||_2 - \lambda^T||\beta||_2, \tag{5}$$

where $||\beta||_0 := \#\{(i,j) : \beta_{i,j} \neq 0\}$, counts the number of non-zero elements in β. In this context, $\lambda \in \mathbb{R}$ is a vector comprising penalty terms, each associated with a specific spot denoted by i. These terms play a crucial role in controlling the model overfit by penalizing the coefficients within $\boldsymbol{\beta}$. The notation $||.||_\ell$ represents a norm used to measure distances in space $\mathcal{L}_\ell$, i.e., $\mathcal{L}_2$ is the Euclidean distance.

For each spot i, we optimize the gene-specific weights indicated as $\boldsymbol{w} = (w_1, \dots, w_Q)$, with $w_i \in \mathbb{R}^+$. These weights are primarily fine-tuned to mitigate errors that may arise due to the rarity of cell types and gene expressions. To calculate these weights, we employ the multiplicative inverse of the values obtained from $\boldsymbol{y}_i$ and use them in sequentially thresholded least squares regressions [53]. It is worth noting that these weights are constrained to be non-negative, effectively ensuring that negative rows are zeroed out, resulting in a weight vector $w_i \geq 0$, as exemplified in Algorithm 1.

Algorithm 1. WISpR algorithm to predict the spatial cell localizations.

Input: scRNA-Seq matrix $\boldsymbol{X}$ (library matrix), spatial transcriptomics vector $\boldsymbol{y_i}$ of spot i, weight vector w_i, thresholding parameter τ_i, penalty parameter λ_i
Output: Sparse coefficient vector $\boldsymbol{\beta_i}$ for spot i
Require: $\beta_i \geq 0$, $w_i \geq 0$, $\tau_i > 0$
  function WEIGHT$(\boldsymbol{X}, \boldsymbol{y}_i)$
    Solve: $\min_{\beta_i \geq 0} ||\boldsymbol{y}_i - \boldsymbol{X}\beta_i||_2^2 + \lambda||\beta_i||_2^2$
    Use solution β_i to calculate weights $w_i = 1/(\boldsymbol{X}\beta_i)$
    return w_i
  end function
  function GRIDSEARCHCV.FIT$(\boldsymbol{X}, \boldsymbol{y}_i, w_i)$
    Define grid: $(\tau, \lambda) \in \mathcal{G}$
    Objective: minimize cross-validated NMAE$(\boldsymbol{y}_i, \boldsymbol{X}\beta_i)$
    Model: $\beta_i = \arg\min_{\beta_i \geq 0} w_i||\boldsymbol{y}_i - \boldsymbol{X}\beta_i||_2^2 + \lambda||\beta_i||_2^2$
    return optimal (τ_i, λ_i) from $\mathcal{G}$
  end function
  function WISpR$(\boldsymbol{X}, \boldsymbol{y}_i, \tau_i, \lambda_i, w_i)$ ▷ Find the best sparse solution
    $\min_{\beta_i \in \mathbb{N}^Q} ||\beta_i||_0$ subject to $w_i||\boldsymbol{y}_i - \boldsymbol{X}\beta_i||_2 - \lambda_i||\beta_i||_2$ ▷ Initial prediction
    while not converged **do**
      $k = k + 1$
      if $\beta^k{}_i \leq \tau_i$ **then** $I \leftarrow \arg(\beta^k{}_i \leq \tau_i)$
        fix $\beta^k_{i,I} = 0$
        $\min_{\beta^k_{i,I} \in \mathbb{N}^Q} ||\beta^k_{i,I}||_0$ subject to $w_i||\boldsymbol{y}_i - \boldsymbol{X}\beta^k_{i,I}||_2 - \lambda_i||\beta^k_{i,I}||_2$ ▷ The best sparse prediction
      end if
    end while
    return $\boldsymbol{\beta}^k_{i,I}$
  end function

Throughout each iteration, WISpR computes spot-specific thresholding parameters represented as (τ_i) in Algorithm 1. Coefficients falling below this threshold, indicating nonsignificant coefficients, are eliminated and the regression is conducted sequentially until significant cell-type coefficients are determined.

The sequential thresholding algorithm introduces a specific requirement for $\boldsymbol{X}$ to guarantee a "unique sparse solution." This requirement is based on the idea that the columns of $\boldsymbol{X}$ should exhibit a high degree of orthogonality. This characteristic is of paramount importance, serving as a critical design criterion for our library $\boldsymbol{X}$, thereby ensuring the uniqueness of our learning objective within a sparse context.

Hyperparameter tuning

The aim of hyperparameter selection is to optimize model performance and accuracy by adjusting parameters related to complexity and sparsity regularization. In WISpR, three hyperparameter tuning approaches were employed to estimate the optimal three spot-specific parameters, targeting the sparsity parameter (λ_i) to control the sparsity of the model, the weight parameter (w_i) to scale the abundance of cells and gene expression levels, and the thresholding parameter (τ_i) to remove nonsignificant coefficients where $i = 1, \ldots, L$ is the ID of the spot and L is the number of spots (Fig 3).

The parameter w_i is optimized for each individual capture spot using non-negative regression, employing the Limited-memory Broyden, Fletcher, Goldfarb, and Shannon (L-BFGS) algorithm [54], with temporary λ set to 0.1. The convergence is achieved when the support of the coefficient vector no longer changes, or after reaching the maximum number of iterations (*max_iter* = 1000). Additionally, a threshold tolerance parameter (`tol`, default = 10^{-5}) is used to prune coefficients below this value during each iteration. The multiplicative inverse of this parameter is then used to predict the optimal hyperparameters λ_i and τ_i. GridSearch, using the precalculated w_i parameter, is employed to fine-tune the hyperparameters λ_i and τ_i, with a parameter grid covering a range of values for both τ_i and λ_i, including normalization and intercept point values. The evaluation is based on the Negative Mean Absolute Error, and cross-validation is conducted with 5 splits and repetitions for robust assessment.

Data-driven synthetic spatial transcriptomics reconstruction from scRNA-Seq

Due to the inherent absence of cellular-level gene expression information in real spatial transcriptomics data, evaluating the accuracy of the deconvolution model requires synthetic datasets. These synthetic datasets are expected to mirror the fundamental characteristics of spatial transcriptomics data to simulate real-life deconvolution procedures, encompassing a maximal number of cells and cell types within a synthetic spot. For performance comparisons of deconvolution models, semi-synthetic spatial transcriptomics datasets were prepared using the count matrices and the cell-type annotations given in the scRNA-Seq data based on the methodology explained in Andersson *et al.* (2020) [30].

The synthetic data generation algorithm utilizes scRNA-Seq data, which is represented as a cell × gene matrix where the cell type information for each cell is known. Suppose that the number of available cell types is N for the scRNA-Seq data, and the boundaries for the number of cell types ($ct_{\min}$ and $ct_{\max}$) and the number of cells ($c_{\min}$ and $c_{\max}$) per spot are set. In this work, we choose $ct_{\min} = 1$ and $ct_{\max} = 5$. The selections for $c_{\min}$ and $c_{\max}$ can vary depending on the sequencing technology used.

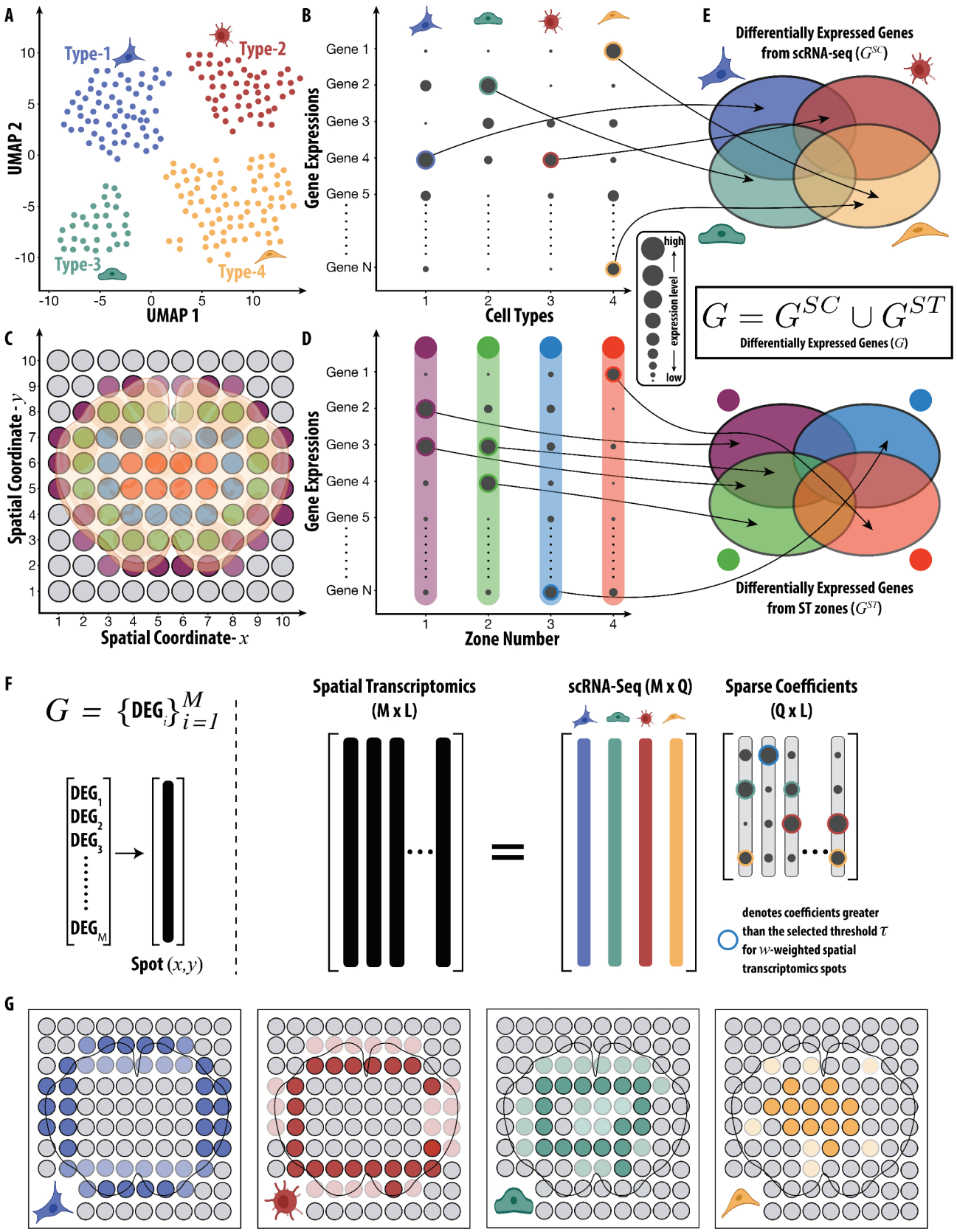

Fig 3. WISpR spatial deconvolution workflow overview. A, UMAP visualization of scRNA-Seq data showcasing four distinct cell types, each depicted in a unique color. **B**, Gene expression profiles for individual cell types derived from the intersection of genes (N) in scRNA-Seq and spatial

transcriptomics datasets. **C**, Representative *x*- and *y*-coordinates of spots in the spatial transcriptomics dataset, with each color indicating a zone profile identified through clustering analysis. **D**, Gene expression profiles for individual spatial zones based on a selected gene set from the intersection of scRNA-Seq and spatial transcriptomics datasets. **E**, Identification of differentially expressed genes specific to cell types (G^{SC}) and zones (G^{ST}), defining cluster differentiation strength. The union of G^{SC} and G^{ST} forms the reference scRNA-Seq and spatial transcriptomics datasets (G) for WISpR deconvolution. **F**, WISpR algorithm summary. The spatial transcriptomics matrix (MxL) and reference scRNA-Seq matrix (MxQ) guide the prediction of the best sparse coefficient matrix (QxL), representing the number of genes (M) in G for each spot, number of spots (L), and number of reference cell types (Q). WISpR iteratively thresholds coefficients based on spot-specific weights (w) and thresholds (τ). **G**, Spatial deconvolution results reveal estimated abundance patterns, locations, and co-occurring cell type compositions for each reference cell type using the WISpR algorithm.

https://doi.org/10.1371/journal.pcbi.1013169.g003

First, the number of cell types is drawn uniformly (without replacement) from the N available cell types within the range of $ct_{\min}$ and $ct_{\max}$. Then, the Dirichlet distribution (with a concentration parameter of 1) is employed to probabilistically determine the abundance of cells per cell type in the spot within the selected boundaries of the number of cells ($c_{\min}$ and $c_{\max}$). These determined numbers are real numbers, so the relative proportions of the cells are rounded to the nearest integers to maintain a biologically meaningful cell distribution. Finally, the maximum number of cells per cell type is randomly sampled from the available cells. For each sampled cell, the gene expression values are multiplied by a ratio factor (ratio factor =0.1 [30]) and added to the spot. Using this algorithm as a foundation, synthetic datasets are prepared to simulate the possible scenarios introduced in the following sections.

Scenarios for technologies and boundaries

This section will explore four distinct scenarios related to potential challenges in real RNA sequencing technologies, emphasizing their properties and limitations.

Scenario 1: Spatial transcriptomics. The in-situ barcoding-based spatial transcriptomics technique was introduced in 2016 [14]. Since its inception, researchers have rapidly adopted and extensively applied this technique in various studies, including transcriptomics analyses of plant species, developmental studies, cancer research, and investigations into neurodegenerative diseases [55–58].

The spatial transcriptomics slides contain capture spots, each featuring approximately 200 million polydT–oligo-based capture probes. These probes bind to the existing RNAs present in the corresponding tissue position, with each capture probe having a diameter of about 100 μm. The number of cells within a single capture spot can range from 10 to 30, depending on tissue type and location of the section. Thus, it can be regarded as a *bulk RNA* sample on a smaller scale. Our synthetic spatial transcriptomics dataset was created following the same principles as the original data, maintaining a total cell count per spot ranging from 10 to 30 and including approximately 1 to 5 cell types selected from available scRNA-Seq datasets [42]. Three independent synthetic spatial transcriptomics datasets, each containing 500 spots and 15,323 genes, were prepared according to Data-Driven Synthetic Spatial Transcriptomics Reconstruction from scRNA-Seq section to evaluate the prediction accuracy of the WISpR model.

Scenario 2: 10X genomics visium. Advancements in in-situ barcoding-based spatial transcriptomics techniques have led to the development of the 10X Visium platform [16]. This platform allows for spatially resolved gene expression analysis across tissue sections with higher resolution. The reduction in the area of the capture spot from 100 μm to 55 μm has enabled the capture of gene expression profiles from a smaller number of cells per capture spot, with the total number of cells typically ranging from 1 to 10.

To simulate data similar to those generated by the 10X Visium platform, the synthetic dataset for this scenario was created by sampling the total number of cells per capture

spot, which varied between 1 and 10. Additionally, the number of cell types per spot was randomly selected to range from 1 to 5 as explained in the Data-Driven Synthetic Spatial Transcriptomics Reconstruction from scRNA-Seq section. Three independent synthetic spatial transcriptomics datasets were generated to evaluate the prediction accuracy. Each dataset consisted of 500 spots and 15,323 genes, ensuring a robust assessment of the WISpR model's prediction accuracy.

Scenario 3: Unmatched Cell Types Cases. Comparing the potential cell numbers analyzed by spatial transcriptomics and scRNA-Seq reveals that spatial transcriptomics surpasses scRNA-Seq in both the quantity and variety of cells. This difference can lead to the inclusion of additional cell types in spatial transcriptomics, depending on tissue type, which may not be present in the reference scRNA-Seq data. As a result, these distinct cell types in spatial transcriptomics remain unidentified in the reference data, contributing to noise in the spatial dataset.

To simulate a noisy dataset, synthetic spatial datasets were generated according to Data-Driven Synthetic Spatial Transcriptomics Reconstruction from scRNA-Seq section using all 15 cell types in the reference scRNA-Seq data, resulting in three independent datasets, each comprising 500 spots and 15,323 genes. Subsequently, the atrial cardiomyocyte cell type ranked as the sixth most populous cell type with 152 cells among the fifteen identified clusters, was excluded from the reference scRNA-Seq sequencing dataset to introduce noise.

Scenario 4: Mislabeled Cell Types. Another possible and realistic scenario involves cases where cell types are mislabeled in the metadata.

To simulate this scenario, we initially generated synthetic data following the approach outlined in Scenario 1, utilizing 15 cell types. We produced three distinct synthetic datasets, each consisting of 500 spots and 15,323 genes. Subsequently, all 462 cells in cell type 2 were relabeled as cell type 3 in the modified scRNA-Seq metadata. Consequently, the metadata included 14 distinct cell-type labels designated for use in the WISpR deconvolution process.

Blended data

The accuracy of the deconvolution algorithms in handling unmatched cell types was assessed by controlled simulations. Initially, a synthetic spatial dataset with 1000 spots was generated, incorporating 2923 mouse brain cells from various cell types: Astrocytes_14 ($n = 250$), Astrocytes_40 ($n = 250$), Astrocytes_41 ($n = 31$), Blood_73 ($n = 46$), Ependymal_47 ($n = 27$), Excluded_38 ($n = 25$), Immune_32 ($n = 131$), Immune_34 ($n = 250$), Immune_35 ($n = 38$), Neurons_25 ($n = 250$), Neurons_26 ($n = 250$), Neurons_27 ($n = 250$), Neurons_63 ($n = 250$), Oligos_0 ($n = 250$), Oligos_1 ($n = 158$), Oligos_14 ($n = 101$), Vascular_14 ($n = 75$), Vascular_67 ($n = 250$), Vascular_69 ($n = 41$).

The cells and cell types utilized in the synthetic data were consistent with those found in the scRNA-Seq dataset, representing 50% of the total scRNA-Seq dataset. To introduce unmatched cells, the remaining 50% consisted of a random selection of cells from the entire human developing heart dataset and non-overlapping cell types from the mouse brain. This selection included Astrocytes 42, Oligos 5, Neurons 48, Oligos 53, Vascular 68, Neurons 12, Neurons 14, Neurons 21, Neurons 52, Neurons 51, Neurons 18, Neurons 15, Neurons 11, Neurons 24, and Neurons 23 as outlined in Table 1.

State-of-the-art deconvolution approaches

To evaluate the predictive performance of the WISpR method, each synthetically generated spatial transcriptomics scenario and the developing human embryonic heart data were deconvoluted and compared with state-of-the-art techniques, including Dampened Weighted Least

Table 1. Generated blended data using mouse brain cells and human embryonic heart cells, labeled as *Mix*[*a*,*b*,*c*], indicates the proportion of mouse brain cells found in both the reference scRNA-Seq and synthetic spatial datasets (*a*), human embryonic heart data (*b*), and mouse brain cells absent in the synthetic spatial dataset within the reference data (*c*).

Condition	ST(+)	Human Heart	ST(–)
Mix[50,50,0]	2923 (50%)	2923 (50%)	0 (0%)
Mix[50,40,10]	2923 (50%)	2338 (40%)	585 (10%)
Mix[50,30,20]	2923 (50%)	1754 (30%)	1169 (20%)
Mix[50,20,30]	2923 (50%)	1169 (20%))	1754 (30%)
Mix[50,10,40]	2923 (50%)	585 (10%)	2338 (40%)
Mix[50,0,50]	2923 (50%)	0 (0%)	2923 (50%)

https://doi.org/10.1371/journal.pcbi.1013169.t001

Square (DWLS) [29], Spatial DWLS [31], Robust Cell Type Decomposition (RCTD) [28], Stereoscope [30], and SPOTlight [32] models.

DWLS

The Dampened Weighted Least Squares (DWLS) [29] model employs vector decomposition-based optimization to address the deconvolution problem, particularly tailored for bulk RNA-seq datasets. The objective is to solve a linear regression equation $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{x}$, where $\boldsymbol{y}$ represents the gene expression values of the RNA-Seq data in general. The scRNA-Seq signature genes are predicted using the hurdle model implemented in the MAST R package [59], with an absolute log2 mean fold change >0.5, to identify the best cell types $\boldsymbol{x}$ based on the scRNA-Seq signature gene matrix $\boldsymbol{S}$ and gene expression vector $\boldsymbol{y}$. The scRNA-Seq static gene signature matrix $\boldsymbol{S}$ encompasses the gene expression vector of selected marker genes for corresponding cell types. A weighted error function is utilized to determine the optimal $\boldsymbol{x}$, which is particularly crucial when dealing with datasets containing rare cell types to avoid overlooking low-expressed informative genes during predictions. To perform the analysis, the publicly available DWLS source code with default parameters is used from the following data repository: https://github.com/dtsoucas/DWLS.

Spatial DWLS

The Spatial DWLS method [31] integrates a Parametric Analysis of Gene Set Enrichment (PAGE) step prior to the DWLS deconvolution process. Furthermore, the Giotto model [47] is utilized to generate the signature gene matrix. The deconvolution procedure in this study was carried out using the code available on GitHub: https://github.com/rdong08/spatialDWLS_dataset.

RCTD

The Robust Cell Type Deconvolution (RCTD) method [28], which employs a model-based approach to estimate the relative abundance of cells at capture points. Initially, the average mRNA count per cluster is computed. Then, a Maximum Likelihood Estimation (MLE)-based approach is employed to predict cell types. The code and hyperparameters were selected based on the recommendations provided in the GitHub tutorial: https://github.com/dmcable/spacexr.

Stereoscope

The Stereoscope method [30] utilizes a model-based negative binomial deconvolution approach to predict cell-type abundance in spots and identify differentially expressed genes

per cell type. Following the official documentation of Stereoscope, the top 5,000 genes that were expressed highest in single-cell data were selected. The algorithm was run with 50,000 epochs, while the remaining parameters were kept as default, as outlined in the repository: https://github.com/almaan/stereoscope.

SPOTlight

The SPOTlight method [32] combines non-negative matrix factorization (NMF) and non-negative least squares (NNLS) for the analysis of transcriptomic data. It operates in three main steps: First, the scRNA-Seq data matrix is decomposed into lower-dimensional matrices, capturing cell-type-specific gene expression patterns and their presence in each cell, called topics. Initialization involves seeding cell-type-specific gene expression patterns with unique marker genes for each cell type and assigning binary values to topics based on cell-type membership. Secondly, it employs NNLS regression to map the transcriptomic data to the specific patterns of the cell type in topics, generating the distributions of the topic profile for each capture location. (referred to as NMFreg). Lastly, it derives consensus cell-type-specific signatures from topics and determines the weights of each cell type present at capture locations using NNLS. In this study, the default parameters recommended for implementation were followed and applied, as indicated in the associated code repository: https://marcelosua.github.io/SPOTlight/.

Performance comparison tools

The performance of the 6 deconvolution models was evaluated by comparing their accuracy in estimating each cell type per spatial capture spot using the generated synthetic spatial datasets.

Root mean squared error

The performance of the method was compared by calculating the *Root Mean Squared Error* (RMSE), When the total number of spots was given as L, for each spot ℓ,

$$RMSE_\ell = \frac{1}{N}\sqrt{\sum_{j=1}^{N} ({}^{\ell}r_j - {}^{\ell}p_j)^2} \tag{6}$$

where $({}^{\ell}r_j)$ is the ground truth of cell type proportions and $({}^{\ell}p_j)$ is the predicted number from the deconvolution models. Here, $j = 1, \ldots, N$ is the index of cell types, and N is the total number of cell types.

To statistically analyze the RMSE scores of the deconvolution models, the *Wilcoxon rank-sum test* [60,61] was employed. This widely recognized statistical hypothesis test is utilized to compare the distributions of two datasets by employing two matched samples. In this work, the built-in test function in the R programming language [62] was utilized with the following parameters:

$$wilcox.test(x, y, alternative = c(\text{"less"}), mu = 0, paired = TRUE, conf.int = TRUE)$$

where x represents the RMSE values of WISpR at all capture spots, and y denotes the RMSE values of the considered alternative methods at all capture spots. The parameter *"alternative = c("less")"* is used to statistically analyse whether the RMSE distribution of the WISpR method is closer to zero compared to others.

Total variational distance

The difference between false positive predictions and ground truth values for each spot x for cell types b and c in $Mix[a, b, c]$ is calculated as;

$$D_{TV}(P, Q) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)| \tag{7}$$

where $P(x)$ denotes the predicted proportion of cell type b or c in spot x and $Q(x)$ is the ground truth (zero value) for these cell types in all spots, respectively over all spots Ω making the TVD a measure of aggregate false-positive predictions for mismatched cell types.

F1 score

Precision is a metric that is used to evaluate the performance of the prediction model. Measures the proportion of true positive predictions (correctly predicted positive instances) of all positive predictions that the model makes. In other words, precision measures the accuracy of the model's positive predictions given as

$$Precision = \frac{True\ Positives}{True\ Positives\ + False\ Positives} \tag{8}$$

where *True Positives* represent the number of correctly predicted positive instances and *False Positives* represent the number of incorrectly predicted positive instances.

Recall, also known as the sensitivity rate, quantifies the proportion of true positive predictions out of all actual positive instances. It is calculated using the equation:

$$Recall = \frac{True\ Positives}{True\ Positives\ + False\ Negatives} \tag{9}$$

where *False Negatives* represents the number of positive instances incorrectly classified as negative.

F1 Score, which combines precision and recall, offers a comprehensive evaluation of the performance of the classification or prediction model. It serves as a balance between precision and recall, which is particularly beneficial for unbalanced data. The F1 Score is calculated using the following formula:

$$F1\ Score\ = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

The F1 score ranges from 0 to 1, with 1 being the best possible score. A higher F1 score indicates better model performance in terms of precision and recall.

Results

WISpR: weight-induced sparse regression

Weight-Induced Sparse Regression (WISpR) offers a targeted approach to deconvolution, particularly suited for high-resolution datasets such as scRNA-Seq. By integrating gene expression profiles with sparse regression and adaptive, spot-specific thresholding, WISpR addresses limitations observed in existing methods, such as the inability to suppress irrelevant cell types. To handle challenges as biological multicollinearity, gene expression imbalance, and overfitting, WISpR dynamically calculates a sparsity threshold, penalty term, and gene expression

weights for each spot. These parameters are optimized through a grid search procedure combined with cross-validation and negative mean absolute error (NMAE) minimization. The best-fitting parameters are then used in an iterative inference framework to estimate cell-type coefficients. As demonstrated in our results, this enables more accurate and interpretable identification of spatially relevant cell types across diverse biological scenarios.

Sparsity is a critical factor in cell-type deconvolution, since tissue regions typically consist of only a limited number of cell types. WISpR addresses this by implementing an adaptive, spot-specific mechanism to eliminate spurious or unsupported cell-type contributions (noise). This approach draws conceptual inspiration from the Sparse Identification of Nonlinear Dynamics (SINDy) framework [53], which seeks parsimonious models by iteratively thresholding small coefficients in a candidate function library to retain only the most significant components. Similarly, WISpR applies a thresholding procedure on regression coefficients, after cross-validation and error minimization, to remove cell types whose contributions fall below a learned spot-specific threshold τ. This threshold τ acts as a cutoff point for biological relevance: If the predicted contribution of a cell type at a given location is lower than τ, it is set to zero. Unlike global L1 regularization, which uniformly shrinks coefficients, this adaptive thresholding promotes interpretable, context-specific sparsity aligned with local biological composition.

WISpR excels at solving the challenging inverse deconvolution problem with a strong emphasis on achieving sparsity. This problem is represented by $\boldsymbol{y} = \boldsymbol{X}\beta$, where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_Q)$ is an $M \times Q$ matrix of gene expression profiles from Q cell types across M genes and $\boldsymbol{X}_i$ is an M-dimensional column vector (gene expression vector for each cell type, Q_i)), $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L)$ is an $M \times L$-dimensional matrix of gene expression profiles from L number of capture spots across M genes and $\boldsymbol{y}_i$ is an M dimensional column vector (gene expression vector for each capture spot, L_i), and $\beta = (\beta_1, \ldots\ \beta_L)$ is a $Q \times L$-dimensional matrix of coefficients and β_i is a Q-dimensional column vector. In this equation, $\boldsymbol{y}_i$ represents gene expressions in spatial transcriptomics spot-i, while $\boldsymbol{X}$ consists of gene expressions from various cell types obtained by scRNA-Seq data, organized as column vectors. The role of β_i is to capture the coefficients that elucidate how the cell types within $\boldsymbol{X}$ contribute to the spatial transcriptomics data $\boldsymbol{y}_i$. The process of combining cell-type vectors in $\boldsymbol{X}$ with their corresponding coefficients in β results in the spatial transcriptomics matrix $\boldsymbol{y}$. Given that $\boldsymbol{y}$ is already known and does not require further manipulation, and the regression's primary task is to determine the elements of β, WISpR's initial step focuses on creating a suitable matrix $\boldsymbol{X}$. This is achieved by leveraging differentially expressed genes (DEGs) from scRNA-Seq and spatial transcriptomics data, underscoring the importance of sparsity.

A DEG shows unique expression or read counts relative to other genes within specific cell types or conditions, serving as a cell type indicator. The presence of DEGs with observed expression levels distinguishes these cell types from others. However, capturing DEGs specific to cell types is a formidable task due to a range of challenges, including biological factors such as biology-grounded constraints and annotation, and methodological issues such as the absence of a gold standard scRNA-Seq dataset, high heterogeneity, dropout rates, potential errors in preprocessing, and the absence of biological ground truth, as depicted in Fig 1A–1D. Various methodological studies have been developed to accurately identify differentially expressed genes (DEGs) [47,49,63–65]. In the WISpR pipeline, we use a Gini coefficient–based DEG selection algorithm [47,49] to identify cell-type–specific genes that are informative for constructing the reference matrix used in deconvolution. This algorithm assigns a score to each gene for each cell type based on the expression inequality in the dataset. For details of the DEG identification step, please refer to the Methods section.

Cellular heterogeneity within a specific cell type can lead to an inadequate representation of DEGs from scRNA-Seq in spatial data. To address this, WISpR separates DEGs specific to cell types from scRNA-Seq data (Fig 3A,3B) and zone-specific DEGs from spatial data (Fig 3C,3D). Zones refer to spatially coherent regions identified by clustering capture spots with similar gene expression profiles. These zones are first defined using unsupervised clustering of spatial expression data, and DEGs are then calculated across zones to identify genes that distinguish transcriptionally distinct regions. The details of the analysis is given in the Methods section.

In contrast to S-DWLS, which assesses the enrichment of scRNA-Seq DEGs on spots, WISpR combines scRNA-Seq DEGs with zone-specific DEGs (Fig 3E). Consequently, WISpR infers DEGs and generates DEG vectors ($\boldsymbol{X}_i$) for each cell type i. These vectors $\boldsymbol{X}_i$ are assembled into a matrix, forming $\boldsymbol{X}$ for deconvolution (Fig 3F) and are used to deconvolute the abundance and composition of cell types per spot (Fig 3G).

In practice, WISpR is a weight-induced sparse regression method that seeks the best and most sparse fit between scRNA-Seq and spatial transcriptomics. WISpR achieves this by iteratively eliminating spurious or unsupported contributions of scRNA-Seq vectors $\boldsymbol{X}_i$ in the basis matrix $\boldsymbol{X}$, setting $\beta_{i,j} = 0$. The algorithm incorporates three crucial hyperparameters to account for biological heterogeneity in both the number of cell types and transcript levels in tissues: (*i*) spot-specific bias terms, (*ii*) spot-specific threshold parameters, and (*iii*) gene-specific weights per spot. Ridge regression, employing the Euclidean (L_2) norm with penalization by adding a *bias term*, is used to reduce the standard error and improve the reliability of the regression. This discarding process iterates through $\beta_{i,j}$ values smaller than the specified *threshold*. Low-contributing vectors $\boldsymbol{X}_i$ are removed by fixing $\beta_{i,j} = 0$, and regression continues with an updated coefficient vector β_i until no removable vectors remain in $\boldsymbol{X}$. However, unconditional removal of vectors can lead to undesirable outcomes, as the rarity of some cell types and expressed genes should be preserved to prevent false negative results. To address this, spot-specific *weight* parameters are introduced, derived from the initial prediction's non-negative solution with the multiplicative inverse as the weight vector. Spot-specific bias, threshold, and weight parameters are optimized through an exhaustive search within a hyperparameter subset and through cross-validation using synthetic datasets with known ground truth. This process allows us to evaluate the impact of different parameter values based on both the RMSE and F1 scores (see Materials and Methods Section for hyperparameter tuning and score calculations). Biological priors, such as the approximate number of cells per Visium spot ([16]), guide the plausible range for these parameters. We prevent overfitting by validating on held-out data and ensuring that no test information is leaked into the training process. This search algorithm simplifies WISpR's usability, reducing the number of parameters while still identifying optimal parameters for each spot.

In summary, WISpR performs targeted deconvolution, retaining significant predictors while eliminating spurious or unsupported cell-type contributions The enhanced cell-type sparsity achieved through this technology contributes to improved prediction accuracy in our model compared to alternative methods. A comprehensive description of mathematical methodology with a generalization to deconvolute all spatial spots in a matrix form (Fig 3F) can be found in the Methods section.

Applications: Benchmarking

Semi-synthetic data. Deconvolution in spatial transcriptomics typically assumes significant overlap between cell types in reference and spatial datasets. However, mismatched and unlabeled cell types often arise due to differences in data preparation, sequencing platforms, and protocols. Rare or closely related cell types in scRNA-Seq may remain indistinguishable

from spatial transcriptomics data, adding technical variability. Additionally, spatial transcriptomics covers larger tissue sections, resulting in the inclusion of more cell types than are typically found in scRNA-Seq references. These challenges necessitate rigorous benchmarking of deconvolution models.

To evaluate WISpR's robustness against these challenges, we tested it on semi-synthetic spatial transcriptomics datasets generated using two distinct methodologies. Datasets were created following a Dirichlet distribution, simulating diverse cell ratios and capture location scenarios (as detailed in Andersson et al. (2020) [30] and the Methods section).

The first dataset was generated using human embryonic heart scRNA-Seq data [42], covering four scenarios: (1) dense sampling with 10 to 30 cells per capture spot, (2) sparse sampling (1 to 10 cells per spot, mimicking 10X Visium), (3) added noise, and (4) mislabeled data, each comprising 1,500 spots.

In the second approach, we combined cells from two biologically distinct sources—human embryonic heart [42] and mouse brain [66]—to evaluate the impact of gene expression disparities and reference mismatches on deconvolution performance. This scenario simulates common challenges encountered in practical applications, where scRNA-Seq references may include cell types that are biologically irrelevant or absent from the spatial dataset, or may originate from different tissues, species, or developmental stages. A blended reference dataset, denoted as $Mix[a, b, c]$, was constructed to systematically test the robustness of the method. In this formulation, a represents the types of mouse brain cells present in both the reference and the synthetic spatial data, b represents the types of human heart cells (biologically distinct), and c represents mouse brain cell types that are included in the reference but absent from the spatial data. The proportion a was fixed at 50%, while $b + c = 50\%$ was varied across conditions to simulate different noise characteristics. In a reliable deconvolution framework, contributions from cell types b and c should be suppressed, as they do not reflect the true cellular composition of the spatial dataset. This setup allows us to assess the ability of each method to distinguish the true signal from the structured noise, a critical capability in real-world analyses that involve reference mismatch.

We compared WISpR to DWLS, RCTD, S-DWLS, Stereoscope, and SPOTlight using Root Mean Squared Error (RMSE) and F1 scores (detailed in the Methods section). Across the four scenarios, WISpR consistently achieved the lowest RMSE in Scenarios 1, 2, and 4, with values of 0.032 ± 0.023, 0.050 ± 0.049, and 0.057 ± 0.054, respectively. In Scenario 3, WISpR performed comparably to DWLS and RCTD (0.071 ± 0.075, 0.071 ± 0.072, and 0.068 ± 0.063, respectively) (Fig 4A, 4B). Separate analysis of RMSE for a, b, and c in $Mix[a, b, c]$ reveals the lowest errors in predicting true cells, along with minimal false positive rates for unmatched cells (Figs C and D in S1 Text). Statistical analysis using the Wilcoxon rank-sum test revealed that WISpR significantly outperformed competing models in Scenarios 1, 2, and 4 ($p<0.01$) and did not show significant differences in RMSE distributions in Scenario 3 between DWLS ($p = 0.258$), RCTD ($p = 0.740$), and S-DWLS ($p = 0.152$) (Tables A and B, and Fig E in S1 Text).

F1 scores further highlighted the superior predictive performance of WISpR. The F1 score, defined as the harmonic mean of precision and recall, provides a balanced metric that is especially informative in imbalanced or noisy classification settings. It consistently achieved scores exceeding 0.84 in all scenarios, demonstrating robust performance even under noisy and mislabeled conditions (Scenario 3, $F1 = 0.843 \pm 0.225$, Scenario 4, $F1 = 0.892 \pm 0.165$) (Fig 4C, 4D). While RCTD performed comparably to WISpR in Scenario 3 with no statistically significant difference in RMSE ($p = 0.740$), its F1 scores remained consistently lower across all scenarios, indicating reduced recall and/or precision in distinguishing relevant cell types. WISpR, in contrast, maintained high F1 scores even under noisy or mismatched conditions, suggesting stronger robustness and discriminatory power.

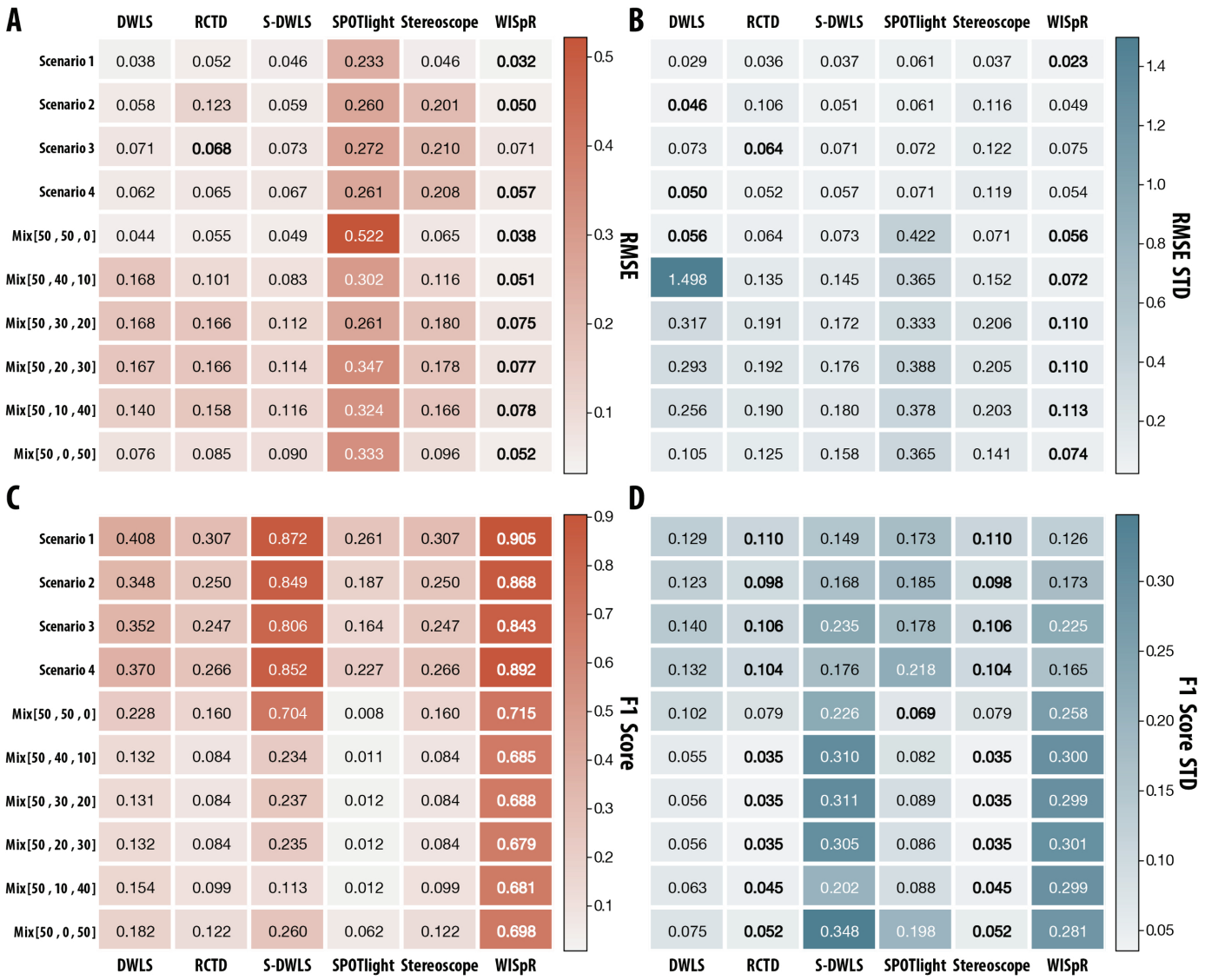

Fig 4. Overview of the comparative analysis of deconvolution model performances using synthetic spatial data in four different scenarios and six different mixture percentages. Four benchmark scenarios are generated by selected cells from 15 cell types in the developing human embryonic heart scRNA-Seq dataset. 6 different mixtures are generated using human heart cells and mouse brain cells. Tested deconvolution algorithms are DWLS, RCTD, S-DWLS, SPOTlight, Stereoscope, and WISpR. **A**, A mean Root Mean Squared Error (RMSE) assessment of predictive performances of six methods' performances. Except for Scenario 3, WISpR outperforms the other five models. In Scenario 3, a slightly better mean predictive performance was observed in RCTD model, where the distribution of errors between RCTD and WISpR displayed no statistical significance. **B**, The mean standard deviations of RMSE scores. WISpR emerged as low error prone especially in blended data (Mix[50,50,0], Mix[50,40,10], Mix[50,30,20], Mix[50,20,30], Mix[50,10,40], Mix[50,0,50]). **C**, The mean F1 scores and **D**, their mean standard deviations in all scenarios. When there are highly matching cell types between scRNA-Seq and synthetic data, WISpR's and S-DWLS's F1 scores are the highest; however, when unmatched cell types are introduced, the S-DWLS score decreases dramatically. The low standard deviations in DWLS, RCTD and SPOTlight indicate their ubiquitous false predictions of the non-existing cell types.

https://doi.org/10.1371/journal.pcbi.1013169.g004

In blended datasets *Mix*[*a*, *b*, *c*], WISpR demonstrated outstanding predictive capacity, achieving the lowest RMSE values across all blends, with values ranging from 0.038 ± 0.056 for *Mix*[50, 50, 0] to 0.078 ± 0.113 for *Mix*[50, 0, 50], all of which are significantly lower ($p < 0.01$) than the alternative model predictions (Fig 4A, 4B, Table B, and Fig F in S1 text). These results highlight WISpR's robustness across varying levels of mismatch and noise, where competing methods showed significant performance deterioration. Notably, S-DWLS, a widely used method, performed well in scenarios involving distinct cell types (i.e.,

Mix[50, 50, 0]), but struggled when similar cell types absent in the synthetic spatial dataset were introduced into the reference dataset. This difficulty underscores the challenges existing methods face when handling biologically similar cell types, particularly when some were shared between datasets and others were absent in the spatial data.

Furthermore, WISpR maintained high predictive performance in all tested configurations, with F1 scores consistently exceeding 0.68, even under challenging mismatch conditions. This performance reflects WISpR's ability to accurately localize biologically relevant cell types while minimizing false-positive predictions. For example, in *Mix*[50, 50, 0], where 50% of the cells in the dataset were mouse brain cells and the remaining 50% were human heart cells, WISpR excelled at identifying the correct cell types with minimal errors (F1 = 0.715). Notably, S-DWLS also performed well in this configuration, achieving an F1 score of 0.704. However, its performance was dramatically reduced to 0.113 when similar mouse brain cells, absent in the synthetic spatial dataset, were intentionally introduced into other blended datasets. This highlights the lower sensitivity of S-DWLS to the incorporation of biologically similar cell types, which led to a weaker precision and recall in distinguishing between cell types from different tissues, particularly in the presence of noise or additional reference mismatches (Fig 4C,4D).

When comparing F1 values across all models (DWLS, RCTD, S-DWLS, SPOTlight, Stereoscope, and WISpR), WISpR consistently outperforms the alternatives, achieving substantial improvements in prediction performance. In all datasets, WISpR achieves F1 scores above 0.68, with the highest score of 0.715, while competing methods such as SPOTlight, Stereoscope and S-DWLS produce significantly lower scores in most cases, often below 0.26. For instance, compared to S-DWLS, which performs well in *Mix*[50,50,0], but struggles in other configurations, WISpR demonstrates an improvement of up to 502% (*Mix*[50,10,40]). Similarly, compared to DWLS, which shows a maximum F1 score of 0.228, WISpR offers an enhancement of approximately 213%. Even Stereoscope, the closest competitor in certain configurations, lags significantly behind, highlighting WISpR's unparalleled robustness and predictive capacity (Fig 4C, 4D). Individual F1 scores for *a* and total variational distances for *b* and *c* further exhibit robust prediction of WISpR independent of noise (Fig D in S1 text). Notably, models such as RCTD and Stereoscope, while exhibiting low standard deviations, displayed consistently lower F1 scores (highest F1 scores are 0.122 and 0.160, respectively) indicating weaker recall and precision under reference noise. In contrast, the higher F1 scores of WISpR underscore its ability to resolve subtle cellular patterns while minimizing false positives.

WISpR's performance in blended datasets demonstrates its capability to account for biological sparsity and natural coherence, even in cross-tissue or cross-species analyses. The method's resilience to noise and its ability to avoid overestimating cell diversity make it particularly suited for applications involving heterogeneous datasets, such as mapping human disease signatures to model organisms or analyzing archival datasets with incomplete annotations. These attributes underscore the potential of WISpR to bridge gaps in spatial transcriptomics and scRNA-Seq integration, providing reliable insights into tissue heterogeneity and cellular interactions.

The broader applicability of WISpR extends to various biological contexts, including the analysis of tissues with mixed origins or unmatched conditions. Its ability to handle challenging scenarios with mislabeled or absent cell types reinforces its utility in uncovering complex tissue architectures and identifying disease-specific cellular signatures. These findings highlight the robustness and reliability of WISpR as a tool for high-resolution mapping of cellular compositions in diverse datasets, advancing our understanding of tissue organization and disease mechanisms.

Developing human embryonic heart

Previous studies extensively analyzed the human embryonic heart at 6.5 postconceptional weeks (PCW), demonstrating the utility of datasets from this stage for high-resolution tissue mapping through integration and deconvolution model [30,44,67]. The original study included spatial transcriptomics, scRNA-Seq, and in situ sequencing analysis of developing human embryonic heart tissue at three stages (4.5-5.0 PCW, 6.5-7.0 PCW and 9.0 PCW), in order to understand cell-type interactions in human organogenesis and obtain a 3D representation of heart development [42].

For this study, scRNA-Seq data from 6.5–7.0 PCW human embryonic heart tissue [42] served as the reference, encompassing 15 cell types from 3,717 cells and 15,323 genes. Spatial transcriptomics data, comprising 1,515 capture spots from the same developmental stage, were deconvoluted to spatially localize these 15 cell types. Fig 5 highlights the mapping performance for two cell types: ventricular cardiomyocytes (Fig 5A–5F) and fibroblast-like cells associated with vascular development in the outflow tract (Fig 5G– 5L). WISpR's comprehensive predictions for all 15 cell types are provided in Fig G in S1 text.

Although earlier efforts have advanced characterization of cell types and their spatial distribution, limitations in deconvolution accuracy persist, particularly in accounting for the biological sparsity and heterogeneity inherent to spatial transcriptomics. Existing methods such as DWLS, RCTD, and Stereoscope often produced biologically implausible results, including negative cell-type proportions and spatial overextension into unrelated tissue regions. These problems were most apparent in the context of spatially restricted or rare cell types, such as ventricular cardiomyocytes and fibroblast-like cells in the outflow tract, which are critical for modeling heart development.

For example, ventricular cardiomyocytes, which are typically confined to the ventricular regions, where ventricles are started to be developed during weeks 4.5–5.0 post-conception and fully developed at 6.5 PCW (as confirmed by TNNT2 immunohistochemistry [42]), were incorrectly predicted by several methods. The heart ventricle anatomy is depicted in yellow. DWLS and S-DWLS produced slightly negative predictions in the ventricular zones (Fig 5A,5C) and have false-positive predictions on the atria, while RCTD, SPOTlight, and Stereoscope failed to localize these cells accurately, instead assigning them broadly in the atrial and ventricular regions as well as outflow tract (Fig 5B,5D,5E). In contrast, WISpR provided spatially consistent and biologically relevant predictions, accurately localizing ventricular cardiomyocytes to the correct anatomical regions and reflecting their known proportions (Fig 5F).

A similar pattern was observed in the predictions for fibroblast-like cells involved in vascular development (e.g., aorta and pulmonary artery formation), as identified by ACTA2 staining at 6.5 PCW [42]. The outflow tract anatomy is depicted in yellow. DWLS, again, produced negative values (Fig 5G), while RCTD, S-DWLS, and Stereoscope misassigned these cells to inappropriate locations, including the ventricular and septal regions (Fig 5H,5I,5K). SPOTlight's predictions were poorly structured, appearing largely random with minimal biological relevance (Fig 5J). However, WISpR accurately mapped fibroblast-like cells to the outflow tract, aligning closely with the known anatomy and minimizing false-positive assignments (Fig 5L).

The number of cell types per capture spot in spatial transcriptomics varies depending on tissue heterogeneity and anatomical structure. Given that each ST and Visium capture spot typically contains between 1-30 and 1-10 cells [14,16], respectively, and that the number of distinct cell types is physiologically expected to be lower than the number of cells, the overall cell-type diversity per spot is anticipated to be relatively limited. In our benchmark (Fig 5M),

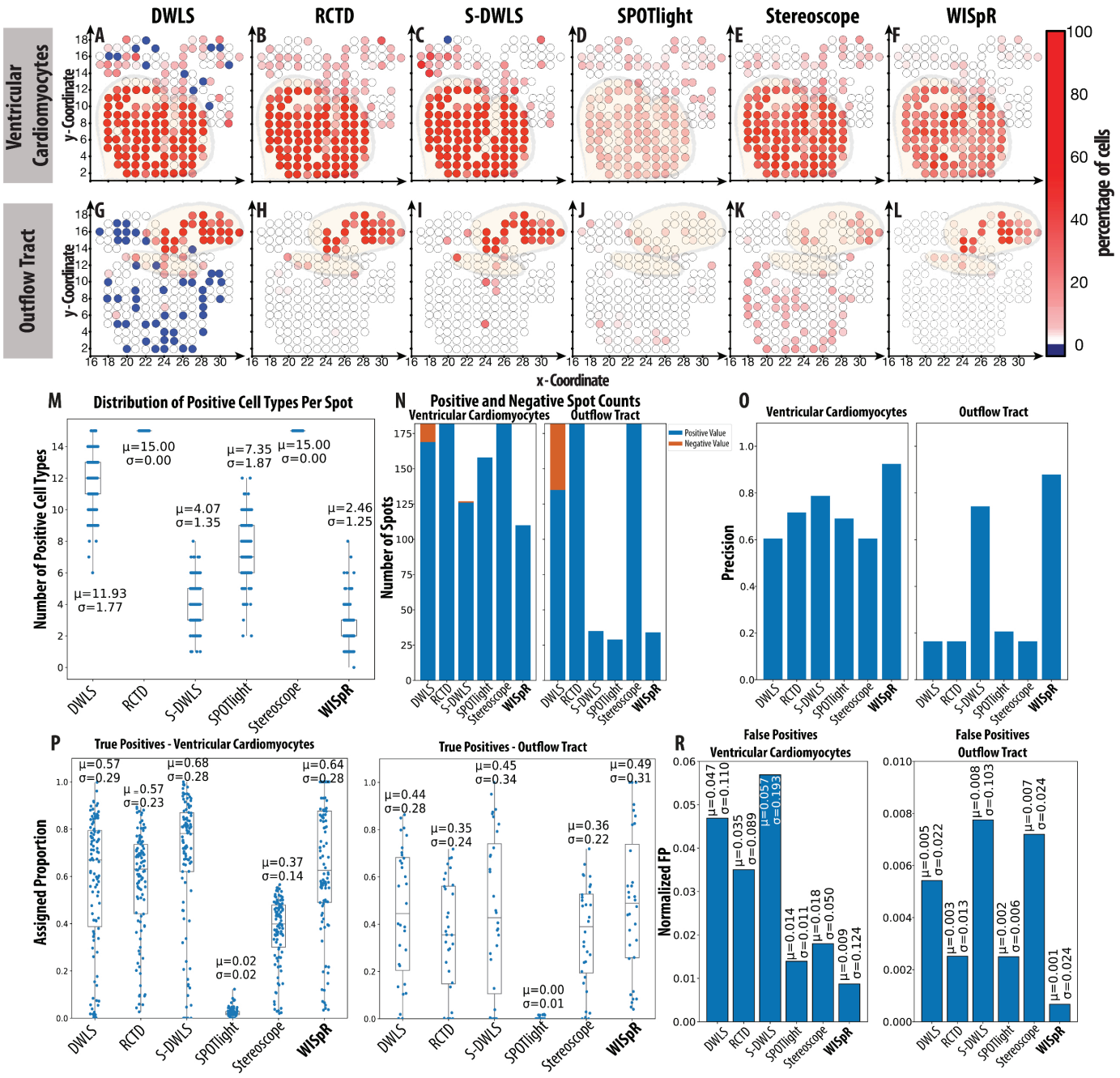

Fig 5. Predictive performance evaluation on real heart data with selected cell types. Top: Predictions for ventricular cardiomyocytes. Bottom: Predictions for outflow tract cells. Blue denotes unrealistic negative predictive coefficients, while the red gradient represents the cell type percentages in the corresponding capture spots. **A** and **G**, DWLS predictions reveal numerous spots with unrealistic negative coefficients. Biologically incorrect positive coefficients are also evident, particularly around the atria and outflow tract, especially for ventricular cardiomyocyte cell prediction. **B** and **H**, RCTD shows a high abundance of false-positive predictions for ventricular cardiomyocyte cells. For the outflow tract, four spots represent false-positive predictions at a low percentage. **C** and **I**, S-DWLS exhibits improved predictions; however, false-positive predictions persist for both ventricular cardiomyocytes and the outflow tract. One spot shows a negative coefficient in **C**. **D** and **J**, SPOTlight predictions indicate overpredicted capture spots in the corresponding tissue. **E** and **K**, Stereoscope also overpredicts cells belonging to ventricular cardiomyocytes. Moreover, cells for the outflow tract are incorrectly predicted in the epicardial zone of the heart. **F** and **L**, WISpR predictions accurately emphasize abundant ventricular cardiomyocytes and outflow tract cells in associated spots. **M**, WISpR estimates an average of $\mu = 2.46 \pm 1.25$ cell types per spot, aligning with expected spatial complexity. **N–O**, WISpR has the lowest positive spots for ventricular cardiomyocytes ($n = 119$) and one of the lowest for outflow tract ($n = 33$), and achieves the highest precision (0.92 and 0.88) across validated regions. **P**, WISpR and S-DWLS show the most highest proportion estimates for dominant cell types in biologically validated region. **R**, WISpR demonstrates the lowest off-target assignment rates, confirming its sparsity-aware design minimizes overfitting and enhances biological fidelity.

https://doi.org/10.1371/journal.pcbi.1013169.g005

WISpR predicted an average of $\mu = 2.46 \pm 1.25$ cell types per spot, consistent with empirical expectations. WISpR also localized ventricular cardiomyocytes and outflow tract cells to biologically plausible regions validated by immunohistochemistry [42], with high precision (0.92 and 0.88, respectively; Fig 5N–Fig 5O) and low false positive rates ($\mu = 0.009$ and 0.001), confirming that WISpR's sparsity-based regularization effectively prevents erroneous cell-type placement. The biological fidelity within the regions were assessed with the distribution of predicted cell-type proportions (Fig 5P). WISpR and S-DWLS gave the most consistent assignments, with WISpR achieving $\mu = 0.64 \pm 0.28$ for ventricular cardiomyocytes and $\mu = 0.49 \pm 0.31$ for cells of the outflow tract, closely aligned with the expected tissue composition. Although absolute abundances per spot are not directly measurable, these values support the idea that WISpR preserves the biologically dominant cell types in each region. Furthermore, WISpR produced the lowest false positive rates ($\mu = 0.009 \pm 0.124$ for ventricular cardiomyocytes, 0.001 ± 0.024 for the outflow tract), indicating that its sparsity-aware formulation effectively limits overfitting and avoids incorrect cell-type assignments beyond biologically plausible domains (Fig 5R).

These observations validate that WISpR consistently outperforms existing methods by generating biologically coherent, sparse, and interpretable spatial predictions across diverse cell types and anatomical contexts. Its ability to precisely localize rare cell types, while avoiding overprediction, was particularly evident in complex tissue regions of the developing human heart. This makes WISpR a powerful tool for uncovering subtle, spatially organized cell populations and deepening our understanding of cellular interactions during organogenesis. To complement these results, spatial predictions for all cell types across nine heart sections are included in Supplementary Fig G in S1 text.

Applications: WISpR at work

Mouse brain cellular maps. After benchmarking WISpR's performance on synthetic and human heart datasets, we next evaluated its predictive power on an independent, biologically complex system: the adult mouse brain. Here, we aimed to assess whether WISpR could produce spatial cell-type distributions that are consistent with prior anatomical knowledge, and whether those distributions reflect biologically meaningful organization. The analysis focused on the use of different datasets from different studies, where spatial data was obtained from the 10X Visium database [68] and scRNA-Seq data was obtained from the Mouse Brain Atlas [66,69].

The spatial transcriptomics dataset of the adult mouse brain, includes 2,698 capture spots, each 55 μm in diameter [68]. 10 clusters, given in the dataset, were used to generate spatially coherent regions. The resulting clusters were annotated as CTX+AMYG (violet), HY+cerebral peduncle (blue), olfactory sensory neurons (yellow), ventral posterolateral TH (dark green), Ndnf+ neurons (cyan), fiber tracts (magenta), HPF and DG (brown), Npy+ neurons (light green), ventricles (white), and stria terminalis (grey), as shown in Fig 6A and 6B. These regional annotations served as reference zones to assess spatial resolution and biological plausibility of predicted cell-type localizations across deconvolution methods.

For cell-type annotation, the scRNA-Seq data from the Mouse Brain Atlas [66,69] was preprocessed using the descriptive metadata file provided in the original dataset. Specifically, in this metadata file, "Class" and "ClusterName" columns, annotated by the authors of the dataset, were aggregated to assign unified cell-type labels for downstream deconvolution analysis. Rigorous cell-type filtering was performed that adhered to the criteria outlined in the study by Andersson *et al.* (2020) [30], selecting cell types with a population size ranging from a minimum of 25 to a maximum of 250 cells.

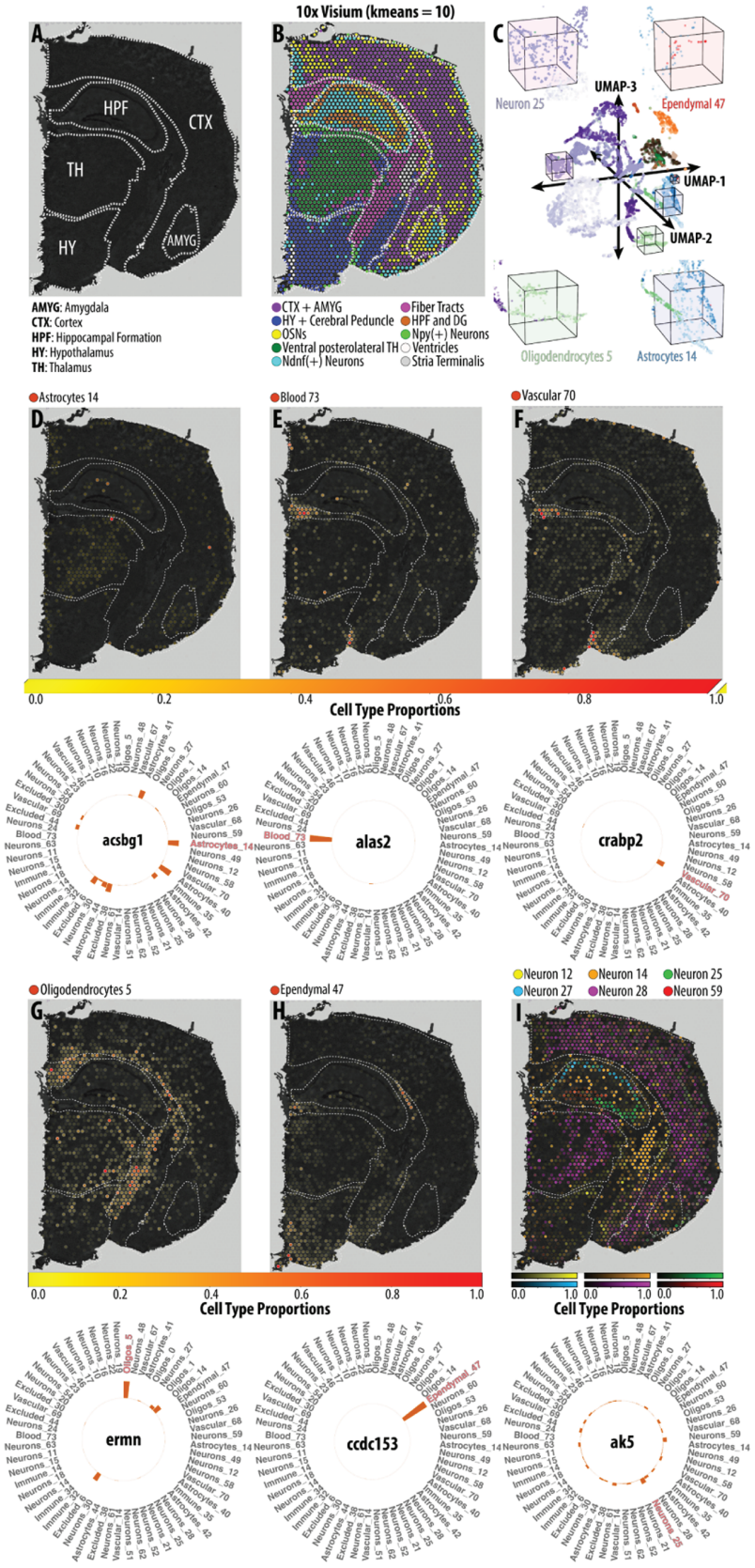

Fig 6. Precise spatial delineation of brain cell types in adult mouse brain. A, Illustration of the right hemisphere coronal section of an adult mouse brain, outlining major brain regions with dashed lines. **B**, Identification of

10 spatial clusters, extracted from the 10X Genomics database, representing CTX+AMYG (violet), HY+cerebral peduncle (blue), OSNs (yellow), ventral posterolateral TH (dark green), Ndnf+ neurons (cyan), fiber tracts (magenta), HPF and DG (brown), Npy+ neurons (light green), ventricles (white), and stria terminalis (grey). **C**, 3D UMAP illustration of scRNA-Seq [69], representing major cell types with distinct colors and cellular subtypes as shades of the major color. For simplicity, Neuron 25 (violet), Ependymal 47 (orange), Oligodendrocytes 5 (light green), and Astrocytes 14 (light blue) are depicted. **D-I** Spatial localization of some selected cell types and their selected marker genes are given below. **D**, Spatial localization of Astrocytes 14 on AMYG, CA1, CA3, DG, and outer layers of CTX, with expression levels of a critical astrocyte-specific DEG, *Acsbg1*. **E** and **F**, Coinciding spatial locations of Blood 73 cells and Vascular 70 cells, respectively, on the third ventricle, the ventral intersection point of the cerebral peduncle and the molecular layer of the DG. The expression of *Alas2* and *Crabp2*, which are an erythrocyte and a fibroblast marker, respectively, are enriched. **G**, Cells of the Oligodendrocytes 5 localized in the corpus callosum, fimbria, stria terminalis, and cerebral peduncle, enriched in *Ermn* **K** (left), specifically expressed in myelinating oligodendrocytes. **H**, Ependymal 47 cells predominantly located in the lateral and third ventricles expressing *Ccdc153*, an ependymal cell marker gene. **I**, Representation of six distinctly located neuronal subtypes, Neuron 12 (yellow), Neuron 14 (orange), Neuron 25 (green), Neuron 27 (blue), Neuron 28 (magenta), and Neuron 59 (red). The brain-specific *Ak5*, is differentially expressed within Neuron Cluster 25 but also found in other neuronal cell types and a few excluded cells, indicating their possible neuronal characteristics.

https://doi.org/10.1371/journal.pcbi.1013169.g006

The final reference scRNA-Seq dataset comprised 8,449 cells assigned to 56 distinct cell-type clusters. These clusters were grouped into 8 major categories based on biological annotations: 5 astrocyte subtypes, 1 blood cell type, 1 ependymal cell type, 4 excluded clusters (due to ambiguous annotations), 4 immune subtypes, 30 neuronal subtypes, 5 oligodendrocyte subtypes, and 6 vascular subtypes. Although predominant cell populations clustered clearly on the UMAP plot (highlighting localized differences; (S1 Movie)), cellular subtypes showed similarities in gene expression profiles, highlighting heterogeneity between cell types (Fig 1C, Fig 6C, and S2 Movie and Fig I in S1 text).

The analysis focused on representative cell types known to exhibit spatial specificity: astrocytes (Cluster 14), oligodendrocytes (Cluster 5), ependymal cells (Cluster 47), spatially distinct neuronal subtypes (Clusters 12, 14, 25, 27, 28, 59), and vascular cells (Cluster 70). These predictions are shown in Fig 6D–6I, where individual capture spots represent estimated cell-type proportions across the tissue section.

To validate that WISpR's predictions were not only spatially coherent but also biologically plausible, we performed gene set enrichment analysis on the selected cell-types' DEGs using the Molecular Signatures Database (MSigDB v7.5.1; Category M8), which contains curated gene sets from single-cell studies in mouse tissues [70] (see Fig J in S1 text). The DEG signatures used for the spatial prediction of WISpR matched well with these established markers, describes the functional roles of the unknown subclusters of these cell types, further supporting the biological relevance of the predicted distributions.

Among the selected spatial specific cell types, Astrocytes (Cluster 14) localized predominantly to the TH and AMYG, and expressed markers such as *Acsbg1* (Fig 6D). Vascular and blood cell types were concentrated near the third ventricle and the cerebral peduncle, characterized by expression of *Crabp2* and *Alas2*, respectively (Fig 6E, 6F). Oligodendrocyte Cluster 5 was enriched in myelin-related markers (*Ptgds*, *Ermn*) and spatially aligned with corpus callosum and fimbria (Fig 6G). Ependymal cells (Cluster 47) expressing *Ccdc153* localized to the lateral and third ventricles (Fig 6H). Despite transcriptomic similarities among neuronal subtypes, WISpR successfully distinguished fine spatial patterns across CTX, HPF, and TH, with neuronal Cluster 25 showing strong enrichment in HPF and expression of *Ak5* (Fig 6I). Together, these predictions demonstrate WISpR's ability to produce spatially and biologically grounded deconvolution results in a real, complex brain dataset. (Detailed cell-type predictions and enrichment results are provided in Figs H and J in S1 text.)

While this section does not directly compare WISpR to alternative methods—such comparisons were addressed extensively in the benchmarking sections—it demonstrates WISpR's capacity to generate biologically grounded spatial predictions in complex, real-world datasets. Using WISpR to integrate spatial transcriptomics with scRNA-Seq data, this study provides a comprehensive high-resolution atlas of cellular heterogeneity and spatial organization in the adult mouse brain. WISpR accurately mapped distinct astrocyte, oligodendrocyte, neuronal, blood, vascular, and ependymal cell populations, revealing their spatial distributions and functional annotations. By delineating region-specific functions, such as the role of astrocytes in neural plasticity, oligodendrocytes in myelination, neurons in circuit-specific connectivity, and vascular cells in structural and metabolic support, WISpR uncovers critical insights into intricate neural circuitry and brain function. These findings underscore WISpR's potential as a transformative tool for exploratory neuroscience and for advancing our understanding of brain organization, health, and disease.

Cancer origins in human breast cancer. Cancer fundamentally alters cellular characteristics, gene expression profiles, and physiological features [71], driving the need for bottom-up genetic approaches that underpin modern precision oncology [72]. Transcriptomics has become central to cancer research, providing comprehensive insights into the intricate gene expression patterns of cancer cells and their microenvironment [72,73].

Therefore, deconvolution of cellular properties from spatial cancer data can clarify tumor-stroma-immune interactions, and enables precise estimation of cell-type co-localizations that aligns spatial data with observed expression patterns.

This study used six primary human breast cancer samples—four triple negative breast cancer (TNBC) and two positive estrogen receptors (ER +)—from publicly available data repositories [74]. Spatial datasets were generated using the 10X Visium assay and preprocessed using the Space Ranger software v 1.1.0 protocol [75].

scRNA-Seq datasets consist of 100,064 cells (TNBC: 42,512 cells, HER2+: 19,311 cells, and ER+: 38,241 cells) and 29,733 genes. This dataset encompasses nine major cell types [74], which were further clustered into 29 minor cellular subtypes that are used for cell-type deconvolution.

WISpR's spot-specific optimization addresses the challenges of subtype overlaps and rare cell detection. During this process, hyperparameters are optimized independently for each Visium capture spot using grid search. For Visium datasets, the penalty parameter (α) was searched within the range [0.0–0.3] and the thresholding parameter (τ) within [0.001–0.03]. The optimal values for each spot were selected based on minimizing the Negative Mean Absolute Error (NMAE) through GridSearch algorithm. The six spatial transcriptomics cancer samples were deconvolved using WISpR, with the results for two samples shown in Fig 7 and the remaining results provided in Fig K in S1 text.

Fig 7 presents spatial transcriptomics data from the CID4535 and CID44971 patient samples, along with a subset of their corresponding cell types as predicted by WISpR deconvolution. In CID4535, an ER + cancer sample, the capture spots were grouped in the original paper into eight zones: invasive cancer areas (yellow, cyan, dark blue), stroma (dark green), lymphoid tissue (red), and noncancerous adipose tissue (brown), which occupied a small portion of the sample (Fig 7A). Additionally, 22 spots were originally identified as artifacts (magenta).

Notably, the spatial clustering alone resulted in three distinct zones being annotated as "invasive cancer", each exhibiting different expression profiles despite originating from the same ER+ cancer type. This suggests potential intra-tumoral heterogeneity, which may arise

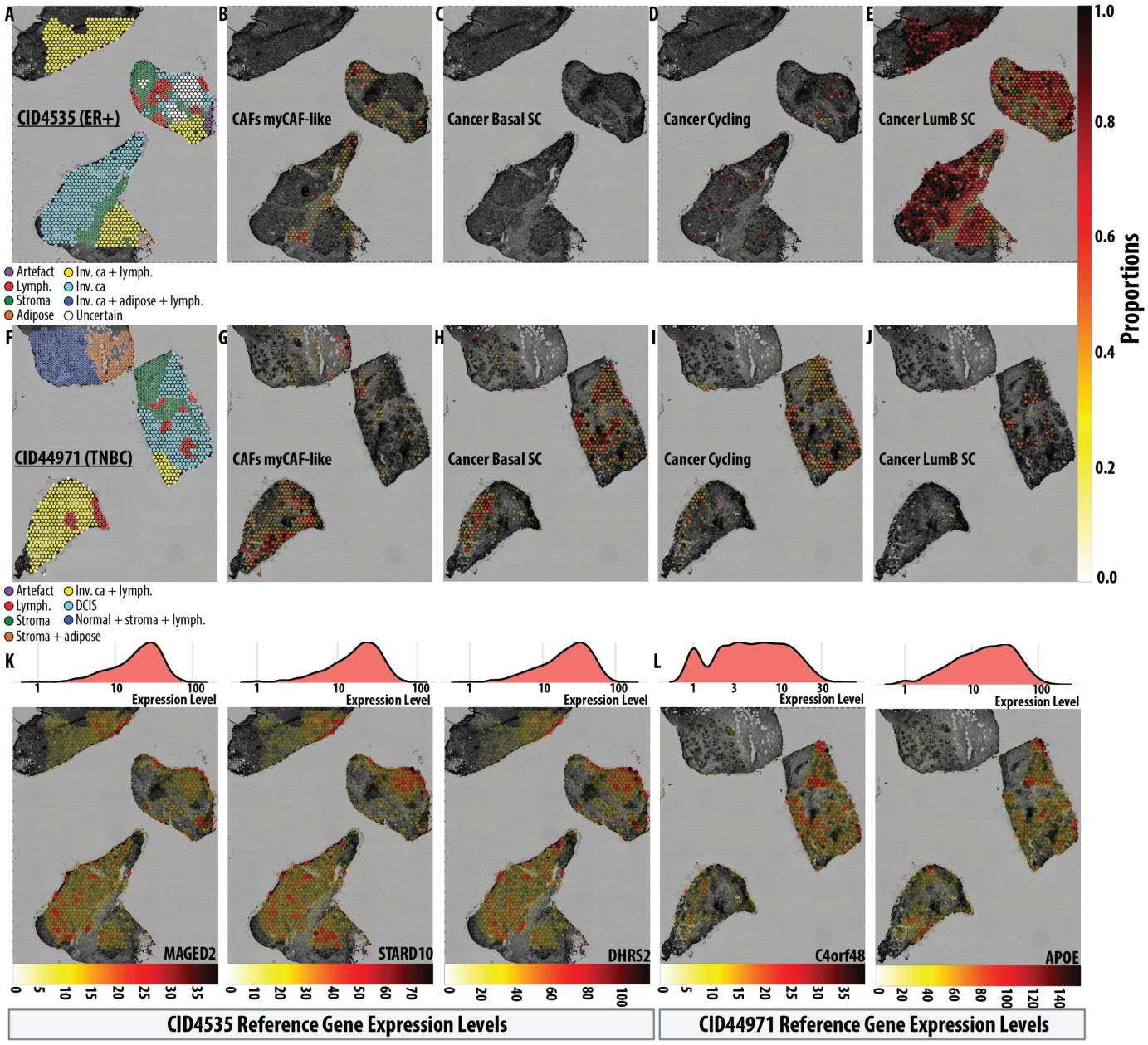

Fig 7. Spatial characterization of cancer-associated cell types in human breast cancer malignancies. A, 8 clusters of ER+ breast cancer (Patient ID: CID4535). Invasive cancer-affected tissues were given by shades of cyan, yellow (cancer within lymphatic tissues), and blue (cancer in adipose and lymphatic tissues). Notably, stromal regions (dark-green) and lymphocytes (red) formed distinct clusters without cancerous involvement. Adipose tissue (brown) exhibited minimal representation within the dataset. A cluster labeled as "Uncertain" (white) was identified, alongside sparse occurrences of artifact spots (magenta) [74]. **B**, Localization of CAFs myCAF-like cells, with their abundance proportions in the stroma (less abundance:white, complete abundance:black). **C**, Basal cells were not detected in ER+ tissue sample. **D**, Cycling cells were localized in the "Uncertain" zone. **E**, LumB SCs were highly concentrated in zones annotated as invasive cancer while absent in stromal and lymphatic zones. **F**, 7 clusters of TNBC breast cancer (Patient ID: CID44971). Invasive cancer and lymphatic zones (yellow), and DCIS (cyan). Normal tissue, including lymphocytes and stroma, were represented in blue, with stroma+adipose tissue in brown. Stromal regions were clustered in dark-green. Healthy lymphocytes were distributed across both invasive cancer and DCIS regions but distinctly clustered in red. A few magenta spots were artifacts [74]. **G**, CAFs myCAF-like cells were predominantly predicted in the invasive cancer and stromal+adipose tissue, representing their cancer-prone characteristics. **H**, Basal SCs and **I**, cycling cells were confined on the DCIS and a bounded region of the invasive cancer spatial cluster, representing similar gene expression patterns in distinct zones at high resolution. **J**, The DCIS had the capacity to diversify its cancer microenvironment based on the existence of LumB SCs. **K**, Expression profiles of representative genes, namely *MAGED2*, *STARD10*, and *DHRS2*, for TNBC, and **L**, *C4orf48* and *APOE*, for ER+ cancer, identified in "SCSubtypes" algorithm [74]. (Top) The ridge plot represents the distribution of the gene expression levels and (bottom) the spatial distribution of the expression of the genes.

https://doi.org/10.1371/journal.pcbi.1013169.g007

from differences in underlying cell-type composition, tumor microenvironment, or transcriptional states, and highlights the limitations of relying solely on spatial clustering to capture cancer identity and prognosis. Moreover, two of the three invasive cancer zones are located within regions identified as adipose tissue (blue) and lymphoid tissue (yellow and blue), where the original tissue composition is not clearly distinguishable. Also, a cluster named "Uncertain" exhibited the limitation of cancer profiling in spatial transcriptomics data. Therefore, a high-resolution mapping is needed to clarify the distribution of cancer within tissue sections.

Fig 7B–7G demonstrate WISpR's ability to resolve specific subtypes within the stromal zones. In particular, myofibroblastic CAF cells (CAFs myCAF-like) are predicted to be exclusively enriched in the stromal and adipose tissue in both cancer samples, consistent with previous findings [74,76] (Fig 7B). WISpR deconvolution revealed no basal cancer stem cells in the ER + sample (CID4535), consistent with their reported prevalence in TNBCs and not in ER+ [74] (Fig 7C). Interestingly, WISpR identified cycling cancer cell–positive spots within spatial regions labeled as "Uncertain" in the original dataset. Some of these spots showed high expression of genes such as *SCGB2A2* — a known biomarker of metastatic breast cancer [77] — highlighting WISpR's ability to detect biologically meaningful signals even in ambiguous regions (Fig 7D). WISpR also identified high LumB stem cell positivity in nonstromal regions, confirming the LumB molecular subtype of this cancer (Fig 7E). This level of resolution, unattainable with clustering-based spatial transcriptomics alone [74], highlights the strength of WISpR to refine broad clusters into biologically meaningful sub-populations, offering critical insights into cellular composition and tumor specialization.

Fig 7F depicts a TNBC (patient ID: CID44971). This sample was clustered into seven zones in the original paper as yellow indicating invasive cancer and lymphocytes, and cyan representing ductal carcinoma in situ (DCIS), a premalignant breast tissue lesion [78].

A distinct group of lymphocytes is observed in the neighborhood of DCIS and invasive cancer (red). Stromal tissue is localized at the border of DCIS (green). One of the tissues comprises stroma, adipose tissue, lymphocytes, and normal tissue (brown and blue). Only one spot was identified as an artifact (magenta).

WISpR accurately localized myCAF-like CAFs within lymphoid regions of invasive cancer clusters as well as basal cancer stem cells and cycling cancer cells within the confines of DCIS (Fig 7G, 7I), revealing spatial patterns indicative of differential proliferative activity between DCIS and invasive cancer zones. Interestingly, both basal cancer stem cells and cycling cancer cells are detected within the upper left segment of the spatial cluster comprising invasive cancer and lymphocytes, potentially indicating variances in proliferative activity compared to other invasive cancer zones.

Interestingly, the expression of the DCIS genes has been reported to reveal its intrinsic subtypes, with a significant percentage (20%) that does not match any of these subtypes, indicating a heterogeneous composition of DCIS [79]. Similarly, the TNBC sample CID44971 with DCIS contained Luminal B cancer stem cells within the DCIS areas, indicating heterogeneity and dual tumorigenesis in this sample (Fig 7J).

Patients were stratified into subtypes using the PAM50 classification [80,81], adapted for single-cell data to create "pseudobulk" profiles per tumor. This process also involved pairing tumor cells and analyzing markers to identify four distinct gene sets, termed "SCSubtypes" [74], to define single cell-derived molecular subtypes, which were used for external validation of WISpR's cell-type predictions.

Of the 65 LumB SC markers, 60 were present in CID4535, and of the 89 Basal SC markers, 82 were present in the spatial data of CID44971, where each of their expression levels are

given in Fig L in S1 text. Among these marker genes, the expression profiles for the 3 representative genes, namely *MAGED2*, *STARD10* and *DHRS2*, for ER+ breast cancer and 2 representative genes, namely *C4orf48* (also annotated as *NICOL1* according to HUGO Gene Nomenclature Committee (HGNC) [82]) and *APOE*, for TNBC were given in Fig 7K,7L, respectively. Interestingly, a significant subset of these genes shows elevated expressions, with spatial expression patterns aligning closely with the predicted localizations of tumor-associated cell types (Figs M and N in S1 text). Specifically, *MAGED2*, *STARD10*, and *DHRS2* show predominant expression in regions initially identified as invasive cancer sites [74], as well as in spots predicted by WISpR to be LumB stem cell territories.

In particular, these three genes show low or negligible expression in the stromal and lymphatic regions, underscoring their strong associations with ER+ breast cancer (Fig 7K). Additionally, *C4orf48* and *APOE* are prominently expressed in DCIS and invasive cancer sites, as well as in spots where WISpR predicts the presence of CAFs myCAF-like and basal stem cell phenotypes (Fig 7L). However, no expression was detected in normal tissues such as stroma, adipose tissue, and lymphatic tissues, strengthening the link between these two genes and invasive breast cancer, including DCIS.

Consequently, WISpR's advanced deconvolution capabilities provide a paradigm shifting strategy to understand the spatial heterogeneity and molecular complexity of breast cancer. By integrating scRNA-Seq data with spatial transcriptomics, WISpR accurately resolved cellular subtypes, including invasive cancer cells, basal and cycling cancer stem cells, and myCAF-like CAFs, across diverse tumor microenvironments. This study underscores the precision of WISpR in mapping the dynamics of the tumor stroma, identifying spatially distinct cancer subpopulations, and clarifying ambiguous zones such as DCIS. Importantly, external validation of WISpR predictions using spatial expression patterns of established cancer marker genes—such as *MAGED2*, *STARD10*, and *C4orf48*—demonstrated strong alignment with predicted tumor subpopulations and their microenvironmental contexts, reinforcing the reliability of the model. The ability of WISpR to uncover functional subpopulations and their spatial relationships holds significant promise for advancing cancer research, enabling deeper insights into tumor progression, therapeutic resistance, and the development of precision oncology strategies.

Discussion

Advances in next-generation sequencing have enabled high-resolution scRNA-Seq and spatial transcriptomics datasets, offering profound biological insights when integrated. Extensive research in transcriptomics has led to the development of efficient techniques to understand not only the spatial location of cells but also their spatial communication and interactions. Understanding the spatial organization of cells is fundamental for decoding tissue function, disease progression, and cellular communication. While scRNA-Seq captures cell-specific gene expression profiles, it lacks spatial context. Conversely, spatial transcriptomics preserves tissue architecture but lacks single-cell resolution and often captures mixed signals from multiple cells per spot. Integrating these complementary technologies requires deconvolution methods capable of resolving underlying cell-type compositions. However, challenges arise from biological variability, such as intra-cell-type heterogeneity and inter-cell-type gene expression similarity, which can confound standard regression-based approaches. Additionally, reference and spatial datasets often differ due to batch effects or tissue-specific divergence, leading to mismatches that exacerbate prediction errors. Existing methods frequently fail to impose biologically grounded constraints such as non-negativity or localized sparsity, resulting in unrealistic or overly dense predictions.

To address these limitations, we introduce WISpR, a robust deconvolution method for integrating scRNA-Seq and spatial transcriptomics data, tailored to both matched and mismatched datasets. WISpR builds on a biologically constrained, sparse regression framework that enforces cell-type sparsity and non-negativity, addressing the key limitations of current methods. Unlike traditional approaches, WISpR adapts to each spatial spot by optimizing hyperparameters and gene weights locally, while systematically rejecting unsupported or spurious cell types through a dynamic thresholding mechanism. By leveraging biological sparsity and spot-specific optimization, WISpR achieves unparalleled precision in resolving spatial cell-type composition, offering a transformative approach to understand tissue architecture. This breakthrough not only advances computational deconvolution but also opens avenues for deeper biological exploration, such as the identification of region-specific cellular interactions in health and disease.

The false positive predictions of alternative deconvolution models (DWLS, RCTD, S-DWLS, SPOTlight, and Stereoscope) are emphasized, particularly in scenarios involving the stress test for deconvolution of human heart cells within mouse brain tissue and the biological validation test for human brain HPF cells within whole mouse brain tissue. In addition, almost all models generated divergent spatial patterns for each cell type, which complicates the evaluation of their reliability. Notably, only WISpR accurately represents mouse brain data with biologically acceptable signals, reflecting substantial tissue and organismal, as well as regional differences between scRNA-Seq and spatial data. This accurately reflects the attempt to deconvolute human heart cells in this specific scenario.

When benchmarked across both synthetic and real datasets, including extreme mismatch conditions and biologically relevant tissues, WISpR consistently outperforms five state-of-the-art methods by up to 86% in RMSE reduction and over 400% in F1 score improvement. Its robustness to noise, ability to eliminate false-positive predictions, and precision in localizing both abundant and rare cell types make it uniquely effective in deciphering spatial cell-type architectures. These results underscore the translational value of WISpR in building highly resolving, biologically interpretable tissue maps, with strong implications for understanding developmental biology, disease mechanisms, and therapeutic resistance in complex tissues.

WISpR successfully deconvoluted spatial transcriptomics data in both mouse brain and human breast cancer tissues, demonstrating its ability to resolve fine-grained cellular landscapes. In the mouse brain, WISpR reconstructed the spatial distribution of eight major cell types and 56 subtypes, despite the lack of subtype annotations, accurately mapping them to biologically meaningful locations. This revealed unannotated subtypes and their functional context, with GSEA validation confirming the spatial organization of astrocytes, neurons, oligodendrocytes, and vascular cells. In six human ER+ and TNBC samples, WISpR generated a high-resolution cancer cell atlas, identifying cancer subtypes and clarifying previously uncertain regions. It revealed LumB-like traits in DCIS and accurately localized myCAF-like fibroblasts in tumor-adjacent stroma. Predictions were validated by spatial cancer subtype-specific marker gene expression profiles from SCSubtypes, confirming subtype-specific localization. Together, these findings highlight WISpR's power to resolve tissue heterogeneity, uncover spatial cell interactions, and inform future therapeutic strategies in both neuroscience and oncology.

The systematic deactivation of weakly contributing cell types in WISpR is particularly beneficial in structured tissues with well-defined spatial compartments. In these contexts, each region typically harbors only a limited subset of cell types, and the built-in thresholding mechanism of WISpR helps eliminate low-contribution noise, such as transcripts originating from incomplete cells, ambient RNA, or spillover between spots, phenomena especially prevalent in spatial technologies such as Visium (as illustrated in Fig 1F). The threshold

range is systematically optimized per spot via GridSearch, ensuring that only signals, which are way below than signals from full cells, that exceed this threshold are retained and iteratively refined via L2 regularization. This intrinsic filtering improves spatial sharpness and biological coherence by suppressing weak, spurious profiles unlikely to reflect the complete or meaningful cellular presence.

The sparsity threshold can be tuned to be more permissive in specific biological contexts where the true signal from cell types is expected to be low but biologically relevant in WISpR, such as very low-input spatial data, where sequencing depth is shallow, and the signal strength is generally reduced in all cell types. In such cases, making the sparsity threshold more permissive allows WISpR to retain subtle but biologically meaningful signals that may otherwise be missed. This flexibility makes WISpR adaptable across tissue types, resolution scales, and biological questions, from robust signal sharpening in well-compartmentalized organs to sensitive detection in noisy or transitional tissues.

Conclusion

Creating a comprehensive cell atlas for healthy and diseased states requires integrating horizontal (same omics) and vertical (different omics) tools. Sparse reconstruction approaches address challenges from cell state variability, technological differences, and dynamic cellular environments by leveraging the inherent sparsity of biological systems to recover cell-type information. Transcriptomics deconvolution methods that incorporate sparse reconstruction must balance maintaining coherence with accounting for biological complexity and variability, a balance crucial for advancing our understanding of tissue composition and function in health and disease. By redefining deconvolution methodologies through sparse regression, WISpR establishes a scalable framework for integrating multi-modal data, paving the way for future innovations in spatial omics. Future work could extend the WISpR application to temporal data integration, enabling to model dynamic changes in cell-type composition and gene expression modules across developmental time, while jointly incorporating spatial continuity and symbolic rules to reveal interpretable, causal relationships underlying spatiotemporal tissue organization. Furthermore, its robustness in noisy datasets suggests potential in clinical settings, such as cancer heterogeneity analysis and prediction of response to treatment.

These results highlight the superior ability of WISpR to uncover spatially distinct cell states, establishing it as an indispensable tool to investigate tissue microenvironments and unravel cellular heterogeneity in complex biological systems. WISpR emerges as a potent instrument for dissecting the biologically sparse makeup of cell types and spatial configurations within intricate tissues. WISpR transcends the limitations of traditional methods, offering not just incremental improvements but a paradigm shift in how spatial transcriptomics data is analyzed. Its potential to uncover new biological insights into health and disease underscores its value as a cornerstone tool in the era of precision medicine.

Supporting information

S1 Text. Supplementary figures and tables. This supporting file includes figures and tables that provide detailed benchmarking, evaluation, and biological validation of the WISpR model. Supplementary Figs A–F present comparative analyses of prediction errors and false positives across multiple deconvolution methods. Supplementary Figs G–H show spatial prediction results for human heart and mouse brain cell types across various tissue sections. Supplementary Figs I and J highlight correlation patterns and enrichment analyses for selected

cell types. Supplementary Figs K–N illustrate spatial predictions and marker gene expression profiles for six breast cancer patient samples. The two supplementary tables A and B summarize key performance metrics and gene sets used in the analysis.
(PDF)

S1 Movie. Localized differences between major mouse brain cell types.
(MP4)

S2 Movie. Cellular heterogeneities between mouse brain cell subtypes.
(MP4)

Acknowledgments

We thank A.O. Argunsah for his helpful comments on mouse brain analysis, and G. Korkmaz for valuable discussions with her on human breast cancer results. We also want to thank I. Topal Kement for her valuable insights and for inspiring our work.

Ethics approval and consent to participate

Not Applicable.

Consent for publication

All authors reviewed and approved the final manuscript.

Author contributions

Conceptualization: Nuray Sogunmez Erdogan, Deniz Eroglu.

Data curation: Nuray Sogunmez Erdogan.

Formal analysis: Nuray Sogunmez Erdogan.

Methodology: Nuray Sogunmez Erdogan, Deniz Eroglu.

Validation: Deniz Eroglu.

Writing – original draft: Nuray Sogunmez Erdogan, Deniz Eroglu.

Writing – review & editing: Nuray Sogunmez Erdogan, Deniz Eroglu.

References

1. 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. 2016;92(2):342–57. https://doi.org/10.1016/j.neuron.2016.10.001 PMID: 27764670
2. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601. https://doi.org/10.1126/science.1257601 PMID: 25700523
3. Akula N, Baranova A, Seto D, Solka J, Nalls MA, Singleton A, et al. A network-based approach to prioritize results from genome-wide association studies. PLoS One. 2011;6(9):e24220. https://doi.org/10.1371/journal.pone.0024220 PMID: 21915301
4. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21(7):1109–21. https://doi.org/10.1101/gr.118992.110 PMID: 21536720
5. Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. Genome Res. 2008;18(7):1150–62. https://doi.org/10.1101/gr.075622.107 PMID: 18417725

6. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021;596(7871):211–20. https://doi.org/10.1038/s41586-021-03634-9 PMID: 34381231
7. Bäckdahl J, Franzén L, Massier L, Li Q, Jalkanen J, Gao H, et al. Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. Cell Metab. 2021;33(9):1869-1882.e6. https://doi.org/10.1016/j.cmet.2021.07.018 PMID: 34380013
8. Olaniru OE, Kadolsky U, Kannambath S, Vaikkinen H, Fung K, Dhami P, et al. Single-cell transcriptomic and spatial landscapes of the developing human pancreas. Cell Metab. 2023;35(1):184-199.e5. https://doi.org/10.1016/j.cmet.2022.11.009 PMID: 36513063
9. Melo Ferreira R, Freije BJ, Eadon MT. Deconvolution tactics and normalization in renal spatial transcriptomics. Front Physiol. 2022;12:812947. https://doi.org/10.3389/fphys.2021.812947 PMID: 35095570
10. Andersson A, Larsson L, Stenbeck L, Salmén F, Ehinger A, Wu SZ, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. Nat Commun. 2021;12(1):6012. https://doi.org/10.1038/s41467-021-26271-2 PMID: 34650042
11. Rájová J, Davidsson M, Avallone M, Hartnor M, Aldrin-Kirk P, Cardoso T, et al. Deconvolution of spatial sequencing provides accurate characterization of hesc-derived da transplants in vivo. Molecul Therapy-Methods Clin Developm. 2023;29:381–94.
12. Piwecka M, Rajewsky N, Rybak-Wolf A. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. Nat Rev Neurol. 2023;1–17.
13. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363(6434):1463–7.
14. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78–82.
15. Kruse F, Junker JP, van Oudenaarden A, Bakkers J. Tomo-seq: a method to obtain genome-wide expression data with spatial resolution. Methods Cell Biol. 2016;135:299–307. https://doi.org/10.1016/bs.mcb.2016.01.006 PMID: 27443932
16. Visium Spatial Gene Expression User Guide. 2020. [cited 2024 Jan 27]. https://support.10xgenomics.com/spatial-gene-expression
17. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell. 2022;185(10):1777-1792.e21. https://doi.org/10.1016/j.cell.2022.04.003 PMID: 35512705
18. Fu X, Sun L, Chen J, Dong R, Lin Y, Palmiter R, et al. Continuous polony gels for tissue mapping with high resolution and RNA capture efficiency. bioRxiv. 2021. https://www.biorxiv.org/content/early/2021/03/17/2021.03.17.435795
19. Cho C-S, Xi J, Si Y, Park S-R, Hsu J-E, Kim M, et al. Microscopic examination of spatial transcriptome using Seq-Scope. Cell. 2021;184(13):3559-3572.e22. https://doi.org/10.1016/j.cell.2021.05.010 PMID: 34115981
20. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):aaa6090. https://doi.org/10.1126/science.aaa6090 PMID: 25858977
21. Hu Z, Ahmed AA, Yau C. CIDER: an interpretable meta-clustering framework for single-cell RNA-seq data integration and evaluation. Genome Biol. 2021;22(1):337. https://doi.org/10.1186/s13059-021-02561-2 PMID: 34903266
22. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093–5. https://doi.org/10.1038/nmeth.2645 PMID: 24056876
23. Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity. Genome Biol. 2019;20(1):110. https://doi.org/10.1186/s13059-019-1713-4 PMID: 31159854
24. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. https://doi.org/10.1038/nmeth.2967 PMID: 24836921
25. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:75. https://doi.org/10.1186/s13059-016-0947-7 PMID: 27122128
26. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11(6):637–40. https://doi.org/10.1038/nmeth.2930 PMID: 24747814
27. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat

Commun. 2015;6:8687. https://doi.org/10.1038/ncomms9687 PMID: 26489834

28. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40(4):517–26.
29. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. Accurate estimation of cell-type composition from gene expression data. Nat Commun. 2019;10(1):2975. https://doi.org/10.1038/s41467-019-10802-z PMID: 31278265
30. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Commun Biol. 2020;3(1):565. https://doi.org/10.1038/s42003-020-01247-y PMID: 33037292
31. Dong R, Yuan G-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. Genome Biol. 2021;22(1):145. https://doi.org/10.1186/s13059-021-02362-7 PMID: 33971932
32. Elosua-Bayes M, Nieto P, Mereu E, Gut I, Heyn H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucl Acids Res. 2021;49(9):e50. https://doi.org/10.1093/nar/gkab043 PMID: 33544846
33. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40(5):661–71. https://doi.org/10.1038/s41587-021-01139-4 PMID: 35027729
34. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. Nat Biotechnol. 2022;40(9):1349–59. https://doi.org/10.1038/s41587-022-01273-7 PMID: 35501392
35. Tu J-J, Yan H, Zhang X-F, Lin Z. Precise gene expression deconvolution in spatial transcriptomics with STged. Nucleic Acids Res. 2025;53(4):gkaf087. https://doi.org/10.1093/nar/gkaf087 PMID: 39970279
36. Zhou Z, Zhong Y, Zhang Z, Ren X. Spatial transcriptomics deconvolution at single-cell resolution using Redeconve. Nat Commun. 2023;14(1):7930. https://doi.org/10.1038/s41467-023-43600-9 PMID: 38040768
37. Geras A, Darvish Shafighi S, Domżał K, Filipiuk I, Rączkowska A, Szymczak P, et al. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. Genome Biol. 2023;24(1):120. https://doi.org/10.1186/s13059-023-02951-8 PMID: 37198601
38. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 6th ed. Hoboken, NJ: Wiley; 2021.
39. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6. https://doi.org/10.1038/nmeth.4236 PMID: 28346451
40. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048 PMID: 34062119
41. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. PLoS Comput Biol. 2015;11(11):e1004575. https://doi.org/10.1371/journal.pcbi.1004575 PMID: 26600239
42. Asp M, Giacomello S, Larsson L, Wu C, Fürth D, Qian X, et al. A spatiotemporal organ-wide gene expression and cell 7atlas of the developing human heart. Cell. 2019;179(7):1647-1660.e19. https://doi.org/10.1016/j.cell.2019.11.025 PMID: 31835037
43. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. [Preprint]. 2018 [Cited 2025 Apr 20]. https://arxiv.org/abs/1802.03426
44. Sun D, Liu Z, Li T, Wu Q, Wang C. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. Nucleic Acids Res. 2022;50(7):e42. https://doi.org/10.1093/nar/gkac150 PMID: 35253896
45. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, et al. Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science. 2016;353(6302):925–8. https://doi.org/10.1126/science.aad7038 PMID: 27471252
46. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat Commun. 2021;12(1):463. https://doi.org/10.1038/s41467-020-20343-5 PMID: 33469025
47. Dries R, Zhu Q, Dong R, Eng C-HL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. 2021;22(1):78. https://doi.org/10.1186/s13059-021-02286-2 PMID: 33685491
48. Gini C. Memorie di metodologia statistica. Padova, Italy: Libreria Goliardica. 1955.
49. Jiang L, Chen H, Pinello L, Yuan GC. Giniclust: detecting rare cell types from single-cell gene expression data with gini index. Genome Biol. 2016;17(1):1–13.

50. Wright Muelas M, Mughal F, O'Hagan S, Day PJ, Kell DB. The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. Sci Rep. 2019;9(1):17960. https://doi.org/10.1038/s41598-019-54288-7 PMID: 31784565
51. Kendall M, Stuart A, Ord JK. Distribution theory. London: Arnold. 1994.
52. Sen A. Poverty, inequality and unemployment: some conceptual issues in measurement. Economic and Political Weekly. 1973. p. 1457–64.
53. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci U S A. 2016;113(15):3932–7. https://doi.org/10.1073/pnas.1517384113 PMID: 27035946
54. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program. 1989;45(1–3):503–28. https://doi.org/10.1007/bf01589116
55. Giacomello S, Salmén F, Terebieniec BK, Vickovic S, Navarro JF, Alexeyenko A, et al. Spatially resolved transcriptome profiling in model plant species. Nat Plants. 2017;3:17061. https://doi.org/10.1038/nplants.2017.61 PMID: 28481330
56. Asp M, Salmén F, Ståhl PL, Vickovic S, Felldin U, Löfling M, et al. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. Sci Rep. 2017;7(1):12941. https://doi.org/10.1038/s41598-017-13462-5 PMID: 29021611
57. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol. 2020;38(3):333–42. https://doi.org/10.1038/s41587-019-0392-8 PMID: 31932730
58. Maniatis S, Äijö T, Vickovic S, Braine C, Kang K, Mollbrink A, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science. 2019;364(6435):89–93. https://doi.org/10.1126/science.aav9776 PMID: 30948552
59. McDavid A, Finak G, Yajima M. Mast: model-based analysis of single cell transcriptomics. Genome Biol. 2015;16(278):10–1186.
60. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. 3rd ed. Hoboken, NJ: Wiley; 2013.
61. Wilcoxon F. Individual comparisons by ranking methods. Breakthroughs in statistics: methodology and distribution. New York, NY: Springer. 1992. p. 196–202.
62. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2021.
63. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502. https://doi.org/10.1038/nbt.3192 PMID: 25867923
64. Mallick H, Chatterjee S, Chowdhury S, Chatterjee S, Rahnavard A, Hicks SC. Differential expression of single-cell rna-seq data using tweedie models. Statist Med. 2022;41(18):3492–510.
65. Shi Y, Lee J-H, Kang H, Jiang H. A two-part mixed model for differential expression analysis in single-cell high-throughput gene expression data. Genes (Basel). 2022;13(2):377. https://doi.org/10.3390/genes13020377 PMID: 35205420
66. Cell Scatterplot, Hippocampus. [cited 2022 Nov 21]. http://loom.linnarssonlab.org/datet/cells/Mousebrain.org.level1/L1_Hippocampus.loom
67. Zhuang X. Spatially resolved single-cell genomics and transcriptomics by imaging. Nat Methods. 2021;18(1):18–22. https://doi.org/10.1038/s41592-020-01037-8 PMID: 33408406
68. Mouse Brain Visium. [cited 2023 Nov 6]. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Adult_Mouse_Brain
69. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, et al. Molecular architecture of the mouse nervous system. Cell. 2018;174(4):999-1014.e22. https://doi.org/10.1016/j.cell.2018.06.021 PMID: 30096314
70. Castanza AS, Recla JM, Eby D, Thorvaldsdóttir H, Bult CJ, Mesirov JP. Extending support for mouse data in the Molecular Signatures Database (MSigDB). Nat Methods. 2023;20(11):1619–20. https://doi.org/10.1038/s41592-023-02014-7 PMID: 37704782
71. Yuan S, Norgard RJ, Stanger BZ. Cellular plasticity in cancer. Cancer Discov. 2019;9(7):837–51.
72. Griewank KG, Scolyer RA, Thompson JF, Flaherty KT, Schadendorf D, Murali R. Genetic alterations and personalized medicine in melanoma: progress and future prospects. J Natl Cancer Inst. 2014;106(2):djt435. https://doi.org/10.1093/jnci/djt435 PMID: 24511108
73. Valdes-Mora F, Handler K, Law AMK, Salomon R, Oakes SR, Ormandy CJ, et al. Single-cell transcriptomics in cancer immunobiology: the future of precision oncology. Front Immunol. 2018;9:2582. https://doi.org/10.3389/fimmu.2018.02582 PMID: 30483257

74. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53(9):1334–47. https://doi.org/10.1038/s41588-021-00911-1 PMID: 34493872

75. Cancer Spatial. https://zenodo.org/record/4739739. Accessed 2023 June 11.

76. Liu L, Liu L, Yao HH, Zhu ZQ, Ning ZL, Huang Q. Stromal myofibroblasts are associated with poor prognosis in solid cancers: a meta-analysis of published studies. PLoS One. 2016;11(7):e0159947. https://doi.org/10.1371/journal.pone.0159947 PMID: 27459365

77. Galvis-Jiménez JM, Curtidor H, Patarroyo MA, Monterrey P, Ramírez-Clavijo SR. Mammaglobin peptide as a novel biomarker for breast cancer detection. Cancer Biol Ther. 2013;14(4):327–32. https://doi.org/10.4161/cbt.23614 PMID: 23358476

78. Coleman WB. Breast ductal carcinoma in situ: precursor to invasive breast cancer. Am J Pathol. 2019;189(5):942–5. https://doi.org/10.1016/j.ajpath.2019.03.002 PMID: 31029232

79. Allred DC, Wu Y, Mao S, Nagtegaal ID, Lee S, Perou CM, et al. Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. Clin Cancer Res. 2008;14(2):370–8. https://doi.org/10.1158/1078-0432.CCR-07-1127 PMID: 18223211

80. Kim HK, Park KH, Kim Y, Park SE, Lee HS, Lim SW, et al. Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: potential implication of genomic alterations of discordance. Cancer Res Treat. 2019;51(2):737–47. https://doi.org/10.4143/crt.2018.342 PMID: 30189722

81. Picornell AC, Echavarria I, Alvarez E, López-Tarruella S, Jerez Y, Hoadley K, et al. Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. BMC Genomics. 2019;20(1):452. https://doi.org/10.1186/s12864-019-5849-0 PMID: 31159741

82. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). Hum Genet. 2001;109(6):678–80.