Research Square

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

SpaCoDA: fast compositional deconvolution maps recurrent cellular niches across hundreds of spatial transcriptomics sections

Xiaokai Fan

hsiaokaifan@gmail.com

Fudan University

Lun Liang

First Affiliated Hospital of GuangXi Medical University

Research Article

Keywords: spatial transcriptomics, cell-type deconvolution, compositional data, tissue niche, meta-niche, breast cancer, glioblastoma

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11482224/v1

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

Additional Declarations: No competing interests reported.

SpaCoDA: fast compositional deconvolution maps recurrent cellular niches across hundreds of spatial transcriptomics sections

Xiaokai Fan[1,2]*, Lun Liang[3]*

1. School of Life Sciences, Fudan University, Shanghai, China
2. Center for Intelligent Medicine, Greater Bay Area Institute of Precision Medicine (Guangzhou), Guangzhou, China
3. Department of Neurosurgery, The First Affiliated Hospital, Guangxi Medical University, Nanning, China

*Correspondence: hsiaokaifan@gmail.com

Abstract

Background: Spot-based spatial transcriptomics such as 10x Visium measures a mixture of cells in each spot. Estimating the cell-type composition of each spot against a single-cell reference is the first step towards cell-level maps of tissue microenvironments, but the most accurate deconvolution methods are slow and memory-intensive at the scale of current atlases, and there still need an established way to make tissue niches comparable across hundreds of sections.

Results: We present SpaCoDA, which fits each spot by variance-weighted non-negative least squares against a reference signature, re-estimates a gene-level platform factor between iterations, and converts RNA shares into cell fractions using cell-size and marker-capture factors. Niches are identified within each section from fused compositional (Aitchison distances) and spatial distances, and their prototypes are integrated across sections into meta-niches. Across 123 simulated configurations under three reference-mismatch scenarios, SpaCoDA was the only one of nine methods ranked within the top two on all four evaluation metrics in every scenario, second to RCTD in mean rank. On 12 imaging-derived datasets with known composition (six Xenium datasets, each evaluated at two annotation levels, and six CosMx/Xenium datasets from three patients), it achieved the best mean rank in every summary, while running 4 to about 1,800-fold faster than the other eight methods (233-fold faster than RCTD by median runtime) with the lowest peak memory. Applied to 254 breast cancer and 181 glioblastoma sections (about 1.18 million spots), SpaCoDA identified 29 and 31 meta-niches, respectively. In breast cancer, the expression signature of a B-cell and plasma-cell meta-niche was associated with longer overall survival in TCGA-BRCA and METABRIC and was higher in tumors with a pathological complete response (pCR) to neoadjuvant therapy in I-SPY2. In glioblastoma, signatures of a perivascular antigen-presenting myeloid meta-niche and of a proliferative, hypoxic meta-niche were associated with shorter survival in TCGA-GBM and three CGGA glioma cohorts.

Conclusions: SpaCoDA combines competitive accuracy with orders-of-magnitude lower computational cost, making composition and niche analysis practical at cohort scale. It provides a practical framework for such analyses, enabling disease subtyping and the identification of novel therapeutic targets, and shows promise for clinical translation.

Keywords: spatial transcriptomics; cell-type deconvolution; compositional data; tissue niche; meta-niche; breast cancer; glioblastoma

Introduction

Spatially resolved transcriptomics is becoming a central technology for relating the cellular composition and organization of tissues to their function and clinical behavior [1-3]. Current platforms fall broadly into two classes. Spot-based assays such as 10x Visium capture the whole transcriptome over large tissue areas, but each spot measures the pooled RNA of several cells. Imaging-based assays such as 10x Xenium and CosMx Spatial Molecular Imager (SMI) reach single-cell resolution but measure predefined gene panels, currently of up to about 5,000 (Xenium) and 6,000 (CosMx) genes; whole-transcriptome imaging like 10x Atera, CosMx Whole Transcriptome (WTX) assay is emerging, but such data are still scarce. Because of their coverage, cost and the volume of existing data, spot-based measurements remain the main source for large spatial cohorts.

For spot-based data, deconvolution which estimating the proportion of each cell type in every spot from a single-cell reference can turn mixed measurements into maps of where each cell type is and how abundant it is. These maps are the starting point for describing cellular neighborhoods, or niches, much as an ecological survey records which species live together. Many deconvolution methods exist: RCTD models spot counts with a Poisson likelihood and platform effects [4]; CARD combines non-negative matrix factorization with a conditional autoregressive spatial prior [5]; SPOTlight uses seeded non-negative matrix factorization followed by non-negative least squares [6]; SpatialDWLS selects candidate cell types by enrichment and then applies dampened weighted least squares [7]; Tangram learns a mapping between single cells and spots by deep learning [8]; Seurat transfers labels through anchors [9]; SMART is a marker-assisted topic model [10]; and SpaTrio aligns single-cell multi-omics with spatial data probabilistically [11]. Benchmarks have shown that accuracy and cost vary widely between methods [12, 13].

Two problems become pressing as spatial cohorts grow to hundreds of sections and hundreds of thousands of spots. First, the methods that are most accurate in benchmarks typically need hours of computation and several to tens of gigabytes of memory per section. Second, clinically motivated questions — which niches recur across patients, and which are associated with outcome or treatment response — require niches that are comparable across sections and datasets. Methods that identify spatial niches across samples, such as CellCharter and scNiche [14, 15], specifically targets single-cell-resolution data; comparing niches defined from deconvolved spot compositions across a whole cohort has received less attention.

Here we present SpaCoDA, a deconvolution estimator designed around the ways in which real references differ from their targets, together with a two-level niche framework that makes the resulting compositions comparable across a cohort. We benchmarked SpaCoDA against eight established methods on simulated data under three reference-mismatch mechanisms of matched magnitude and on imaging-derived data with known composition, and then applied it to 254 breast cancer and 181 glioblastoma sections, yielding 29 and 31 meta-niches. In breast cancer, the signature of a B-cell and plasma-cell meta-niche was associated with longer survival and with pathological complete response in I-SPY2; in glioblastoma, signatures of two meta-niches were associated with shorter survival. These associations are exploratory and generate hypotheses for future validation.

Results

SpaCoDA estimates cell-type composition and identifies recurrent tissue niches

SpaCoDA combines fast spot-level deconvolution with a two-level niche analysis (Fig. 1). From an annotated single-cell reference, it scales each cell to the same library size, averages the profiles of each cell type and normalizes each type of profile to sum to one, giving a signature matrix (Fig. 1A). Marker genes are ranked within each cell type by specificity multiplied by log expression, and a cell-size factor records the average RNA content of each type. Each spot is then fitted by variance-weighted non-negative least squares (NNLS) against the signature restricted to the marker genes; between iterations, a gene-level platform factor that rescales the signature towards the spatial data is re-estimated. The fitted RNA shares are converted into cell fractions by dividing by the cell-size factor and by the fraction of each type's RNA captured by the marker genes.

Within each section, spots are compared by the Aitchison distance between their compositions and by their Euclidean distance in space; the two distances are scaled to a common range and combined, embedded by principal coordinate analysis (PCoA), and clustered on a shared-nearest-neighbor (SNN) graph with the Louvain algorithm, choosing the resolution with the highest modularity (Fig. 1B). The mean composition of each local niche serves as its prototype. Prototypes from all sections are then normalized, reduced by principal component analysis (PCA), integrated across sections with Harmony [16] and clustered jointly, so that each resulting meta-niche is by construction comparable across the cohort; meta-niche labels are mapped back to every spot (Fig. 1C).

Accuracy on simulated data under three reference-mismatch scenarios

To evaluate deconvolution under controlled conditions, we built a simulator that generates a single-cell reference with families of correlated cell types, rare types and cell-to-cell heterogeneity; places compositions on spatial patterns; and generates spot counts by explicitly sampling cells, with negative-binomial expression, ambient RNA, dropout, variable sequencing depth and a restricted gene panel (Fig. 2A; Methods). We considered three relationships between reference and target. In scenario I, independent Gaussian noise proportional to expression was added to every gene–cell-type entry of the reference, a strong and unstructured divergence. In scenario II, the reference differed from the target by a multiplicative factor per gene (a platform effect) plus type-specific perturbations on a subset of genes, calibrated to the same overall divergence as scenario I. In scenario III, the reference was re-estimated from a random 60% of the cells, so that reference and target differed only by sampling noise.

The design comprised 123 configurations, each analyzed under all three scenarios: 27 main configurations which include three difficulty tiers, three spatial patterns and three seeds, 84 single-factor perturbations, and 12 configurations varying the number of cell types. Within a configuration, the three scenarios shared the same target data. We compared SpaCoDA with RCTD, CARD, Seurat anchor transfer (Seurat anchor), Tangram, SpatialDWLS, SMART, SpaTrio and SPOTlight using Spearman and Pearson correlation, root-mean-square error (RMSE) and Jensen-Shannon divergence (JSD).

On the main configurations, RCTD achieved the highest Pearson correlation and the lowest RMSE and JSD in all three scenarios, and SpatialDWLS the highest Spearman correlation (Fig. 2B; Supplementary Table S1). SpaCoDA ranked second on each of the four metrics in every scenario and was the only method within the top two on all metrics; RCTD ranked third on Spearman correlation in each scenario. For example, under the platform-effect scenario (II), SpaCoDA reached a mean Pearson correlation of 0.957 and an RMSE of 0.040, compared with 0.976 and 0.031 for RCTD. Across the single-factor perturbations

platform-effect strength, reference divergence, gene-panel coverage, cell-type collinearity, expression noise, ambient RNA and cells per spot, which the relative ordering of the methods was broadly preserved (Fig. 2C).

Accuracy on pseudo-spots built from single-cell-resolution data

Fully synthetic data cannot reproduce which cells are actually neighbors in a tissue. We therefore built Visium-like pseudo-spots from six Xenium datasets with different gene panels — breast cancer (313 genes and 5K), colorectal cancer (480 genes), lymph node (377 genes and 5K) and renal cell carcinoma (377 genes) — by aggregating segmented cells within circular capture areas on a regular grid, retaining spots with at least three cells (25,293 pseudo-spots in total; Fig. 3A). The cell-type counts within each spot give its true composition. As references we used up to 5,000 sketched cells per dataset, annotated at a fine level (20–48 cell type labels) and a major-type level (8–9 cell type labels). Because the reference cells were drawn from the same section and were not excluded from the aggregated pool, this benchmark has almost no platform effect and represents a favourable, same-section setting.

Averaged over the six datasets, SpaCoDA had the highest Spearman (0.604) and Pearson (0.693) correlations and the lowest JSD (0.145) at the fine level, with an RMSE (0.059) close to the lowest (CARD, 0.058); at the major-type level it had the highest Spearman (0.746) and Pearson (0.819) correlations, with RMSE and JSD within 0.001 of the best methods (Fig. 3B; Supplementary Table S2). SpaCoDA had the best mean rank across the four metrics at both annotation levels. SPOTlight failed on four of the 12 dataset–level combinations.

Cross-platform accuracy and computational cost

To test performance when the reference comes from a different technology, we used the benchmark of Ren et al., in which colorectal (COAD), hepatocellular (HCC) and ovarian (OV) carcinomas were each profiled by CosMx 6K and Xenium 5K on adjacent sections together with matched scRNA-seq [17]. We aggregated each imaging dataset into pseudo-spots and deconvolved them against the scRNA-seq reference (Fig. 4A). By median over datasets, SpaCoDA had the highest Spearman correlation (0.453) and the lowest RMSE (0.063) and JSD (0.079), whereas RCTD had the highest Pearson correlation (0.636 versus 0.579 for SpaCoDA) (Fig. 4B; Supplementary Table S3). SpaCoDA again had the best mean rank. Its advantage in rank correlation held across the range of cell-type abundance, and its RMSE and conditional JSD were among the lowest across abundance levels (Fig. 4C), composition maps of SpaCoDA, RCTD and CARD reproduced the annotated tissue architecture (Fig. 4D).

SpaCoDA was the fastest method, with a median runtime of 10.8 s per dataset, and had the lowest peak memory (1.69 ± 0.27 GB) (Fig. 4E; Supplementary Table S4). By median runtime, the other methods were 4.0-fold (CARD) to 1,818-fold (SMART) slower; RCTD took a median of 2,517 s, 233-fold longer than SpaCoDA (197-fold by the geometric mean shown in Fig. 4E). Peak memory of the other methods ranged from 2.94 GB (RCTD) to 15.3 GB (SpaTrio). This cost profile makes cohort-scale deconvolution of hundreds of sections feasible on standard hardware.

A breast cancer atlas resolves recurrent malignant, stromal and immune environments

We applied SpaCoDA to breast cancer spatial transcriptomics data from seven sources, comprising 254 section identifiers and 594,341 spots with assigned labels, using the single-cell atlas of Wu et al. with 49 cell-type and cell-state labels as reference [18]. The 2,221 local niche prototypes were grouped into 29 meta-niches (Fig. 5A). Serial sections from the same block shared meta-niches with consistent spatial layouts (Fig. 5C, I), even though the atlas combines data from several studies and platforms. Meta-niches distinguished luminal, HER2-associated, basal-like and cycling malignant epithelial states, as well as environments enriched for myoepithelial cells, fibroblasts, vascular and perivascular cells, T cells, myeloid cells or B-cell lineages (Fig. 5B). Marker expression agreed with these compositions: basal-like and myoepithelial environments expressed KRT5, KRT14 and KRT17, stromal environments expressed collagen genes, and B-cell and plasma-cell environments showed immunoglobulin-rich profiles (Fig. 5D). On the basis of their dominant cell types and marker genes, the 29 meta-niches were annotated as 25 niche types and grouped into five regions: tumor core, tumor–immune–stromal interface, stromal region, immune region, and basal and myoepithelial region (Fig. 5E).

We next asked whether the expression signatures of the meta-niches were associated with outcome in bulk-expression cohorts. A signature of meta-niche 23, a B-cell and plasma-cell environment, was associated with longer overall survival. In TCGA-BRCA [19], scores above the median were associated with a lower hazard of death (n = 1,068; 144 deaths; hazard ratio [HR] 0.62, 95% confidence interval [CI] 0.44–0.86; log-rank P = 0.004), and the association was in the same direction in METABRIC [20] (n = 1,980; 1,143 deaths; HR 0.79, 95% CI 0.70–0.89; log-rank $P = 7.8 \times 10^{-5}$) (Fig. 5F; Supplementary Table S5). Genes of this signature were enriched for B-cell receptor signaling, immunoglobulin-mediated immune responses, complement-dependent cytotoxicity and humoral immunity (Fig. 5H).

In the pre-treatment expression data of the I-SPY2 trial [21] (988 evaluable records, 319 with pathological complete response [pCR]), meta-niche 23 scores were higher in records with pCR than without (median 0.684 versus 0.647; two-sided Wilcoxon $P = 1.3 \times 10^{-8}$). Within the paclitaxel plus pembrolizumab arm (69 records, 31 with pCR), the corresponding medians were 0.706 and 0.595 (two-sided Wilcoxon $P = 2.2 \times 10^{-4}$) (Fig. 5G). Similar associations were present in several other arms, including paclitaxel plus AMG 386 (OR 2.03) and paclitaxel plus ABT-888 and carboplatin (OR 2.63), but not in the trastuzumab or pertuzumab containing arms (Supplementary Table S6); the association therefore reflects response to neoadjuvant therapy in general rather than a pembrolizumab-specific effect. These observations agree with reports linking B cells and tertiary lymphoid structures to favorable outcome and response to therapy [22-25], and suggest that this niche is worth investigating as a predictive marker.

Glioblastoma meta-niches capture neural, malignant and microenvironmental programs

We assembled 181 glioblastoma section identifiers from seven sources with 585,556 spots and deconvolved them against an integrated glioblastoma single-cell reference with 25 cell-type and cell-state labels from Lin et al. scRNA-seq dataset [26]. The 2,283 local niche prototypes formed 31 meta-niches (Fig. 6A). They captured neural, malignant, myeloid, vascular and stromal programmers: oligodendrocyte-associated environments expressed MBP, PLP1 and MOBP; neuronal environments expressed SYT1, NRGN and SNAP25; mesenchymal-like and hypoxic environments expressed CHI3L1, VEGFA and SERPINE1; and immune-associated groups expressed the antigen-presentation genes CD74, HLA-DRA and B2M (Fig. 6B, C). Programmers of this kind, and their organization around hypoxic regions, have been described by integrative spatial analyses of glioblastoma [27]. On the basis of composition and marker genes, the 31 meta-niches were annotated as 24 niche types — including reactive-astrocyte, proliferative,

proliferative-hypoxic, proliferative-OPC, perivascular, perivascular-myeloid, perivascular-MES, oligodendrocyte, oligodendrocyte-immune, neuronal, MES-myeloid, MES-inflammatory, interferon-high (OPC, AC-OPC and AC), inflammatory-perivascular, inflammatory-hemorrhagic, hypoxic (proliferative, glycolytic and MES), fibrotic (stromal and MES), astrocyte-perivascular and AC-OPC tumor niches — and grouped into eight regions: tumor core, proliferative, perivascular, oligodendrocyte, neuronal, mesenchymal–stromal, immune and hypoxic (Fig. 6D).

We then examined the meta-niche signatures in TCGA-GBM [28] and three cohorts of the Chinese Glioma Genome Atlas (CGGA) [29]. Signatures of meta-niche 1, a perivascular environment with myeloid and antigen-presentation features, and of meta-niche 29, a proliferative environment with hypoxia and glycolysis-related expression, were associated with shorter overall survival (Fig. 6E; Supplementary Table S5). In TCGA-GBM (n = 169), the HRs for scores above versus below the median were 1.49 (95% CI 1.06–2.09) for meta-niche 1 and 1.53 (1.09–2.16) for meta-niche 29. In the CGGA cohorts, the HRs ranged from 1.88 to 3.15 for meta-niche 1 and from 2.93 to 4.68 for meta-niche 29. Gene Ontology analysis linked meta-niche 1 to antigen processing and presentation, and meta-niche 29 to glycolytic processes, the latter supported by only two or three genes per term (Fig. 6F). Meta-niches 1 and 29 are therefore candidate markers of aggressive glioma biology that warrant validated analyses.

Discussion

SpaCoDA was developed to make composition and niche analysis practical for spatial transcriptomics cohorts of hundreds of sections. Across simulated data with three types of reference mismatch and two imaging-derived benchmarks, it was consistently among the most accurate of nine methods, ranking second to RCTD by mean rank in simulation and first on the imaging-derived data, while running about 230-fold faster than RCTD (median 10.8 s per dataset) with the lowest peak memory. This allowed us to deconvolve about 1.18 million spots from 254 breast cancer and 181 glioblastoma sections and to integrate them into 29 and 31 meta-niches that recovered recognizable malignant, stromal, immune and neural environments. The contribution of SpaCoDA is therefore a favorable balance between accuracy and computational cost, combined with a niche representation that is comparable across a cohort, rather than uniformly superior accuracy.

RCTD, which models spot counts with a Poisson likelihood and explicit platform effects [4], achieved the highest Pearson correlation and the lowest RMSE and JSD in simulation, whereas SpaCoDA led mainly through rank correlation on the imaging-derived benchmarks. SpaCoDA keeps the idea of platform correction but replaces the count likelihood with variance-weighted non-negative least squares, an objective related to the regression steps of SPOTlight and SpatialDWLS [6, 7] that can be solved exactly for all spots at once. Rank correlation reflects how well spots are ordered by the abundance of each cell type, which is what spatial maps and niche boundaries depend on, whereas Pearson correlation and RMSE are more sensitive to the absolute scale of abundant types; count-based likelihoods may therefore remain preferable when accurate absolute proportions are required. That no method led on every metric is consistent with previous benchmarks [12, 13].

Two design choices have broader methodological implications. First, calibrating independent and platform-type reference perturbations to the same overall divergence allowed methods to be compared across mismatch structures at matched magnitude. The leading methods kept their relative order across scenarios, indicating that SpaCoDA's ranking does not depend solely on the structure that its gene-level factor is designed to absorb, although the contributions of individual model components have not been tested by ablation. Second, whereas CellCharter and scNiche identify niches in single-cell-resolution data [14, 15], SpaCoDA defines niches from spot compositions in Aitchison geometry and integrates only local prototypes, reducing about 1.18 million spots to about 4,500 objects for integration with Harmony [16]; consistent meta-niches across serial sections support its reproducibility within a tissue block. A meta-niche is, however, a recurrent estimated composition state rather than evidence of a shared cellular arrangement or interaction, and it depends on the handling of zeros, the granularity of the reference, the clustering resolution and the integration step, which may remove biological differences correlated with section identity.

In breast cancer, the signature of the B-cell and plasma-cell meta-niche 23 was associated with longer survival in TCGA-BRCA and METABRIC and was higher in I-SPY2 tumors with pCR, consistent with reports linking B cells and tertiary lymphoid structures to favorable outcome and immunotherapy response [23-25]. In glioblastoma, signatures of the perivascular antigen-presenting myeloid meta-niche 1 and the proliferative, hypoxic meta-niche 29 were associated with shorter survival, in line with the organization of glioblastoma around hypoxic regions [27] and with reports linking hypoxic and mesenchymal-like programs to aggressive disease.

Although SpaCoDA performs relatively well in some scenarios, several limitations remain. SpaCoDA cannot correct for cell types missing from the reference, misannotation or state-specific expression changes,

and fitting on the proportion scale discards sequencing depth. Within these bounds, SpaCoDA makes composition estimation and niche comparison practical at the scale of current spatial atlases and provides a reproducible starting point for asking which cellular environments recur across patients and how they relate to the course of disease. SpaCoDA offers a practical framework for such studies, enabling disease subtyping and the identification of novel therapeutic targets, and shows promise for clinical translation.

Methods

The principle of SpaCoDA

Reference signature and marker genes

Let X be the cells × genes count matrix of the reference, with cell-type labels $l_c \in \{1, \dots, K\}$. Each cell is scaled to 10^4 counts, and the scaled profiles are averaged within each type:

$$\tilde{x}_{cg} = \frac{10^4 x_{cg}}{\sum_{g'} x_{cg'}}, \quad B_{gk} = \frac{1}{n_k} \sum_{c:l_c=k} \tilde{x}_{cg} \tag{1}$$

where n_k is the number of cells of type k. The cell-size factor s_k is the mean raw library size of type k, $s_k = \frac{1}{n_k} \sum_{c:l_c=k} \sum_g x_{cg}$. The columns of B are then normalised to sum to one, $B_{gk} \leftarrow \frac{B_{gk}}{\sum_{g'} B_{g'k}}$, so that each type profile describes relative rather than absolute expression.

Marker genes were ranked within each type by

$$\text{score}_{gk} = \text{Spec}_{gk} \cdot \log\left(1 + 10^4 B_{gk}\right), \quad \text{Spec}_{gk} = \frac{B_{gk}}{\sum_{k'} B_{gk'}} \tag{2}$$

considering genes with $\text{Spec}_{gk} \geq 0.15$ and B_{gk} at or above the median of the positive entries of type k. The top-ranked genes of every type were combined into one marker set. The number of markers per type followed the run settings: up to 30 in the imaging-derived benchmarks, and $\max(30, [G/K])$ in the simulations and atlases.

Variance-weighted non-negative least squares

Let y_i be the marker-gene counts of spot i scaled to sum to one, and B the signature restricted to the marker genes with its columns renormalized. SpaCoDA estimates the RNA-share vector $a_i \geq 0$ of every spot by

$$\hat{a}_i = \text{argmin}_{a \geq 0} \|y_i - Ba\|^2_{V_i} = \text{argmin}_{a \geq 0} \sum_g v_{ig} \left(y_{ig} - \sum_k B_{gk} a_k\right)^2 \tag{3}$$

with gene weights derived from the negative-binomial mean–variance relation at the current fit, $v_{ig} \propto \frac{1}{\mu_{ig} + \varphi \mu_{ig}^2}$, $\mu_{ig} = (B\hat{a}_i)_g$, $\varphi = 0.1$; weights are capped at 10^4 times the smallest weight of the spot and rescaled to mean one within each spot. Because y_i is on the proportion scale, μ_{ig} is small and these weights behave close to Poisson weights. Expanding the weighted norm gives a non-negative quadratic programme per spot,

$$\min_{a \geq 0} a^T Q_i a - 2 c_i^T a, \quad Q_i = B^T V_i B, \quad c_i = B^T V_i y_i \tag{4}$$

All $K \times K$ matrices Q_i are computed in a single matrix product by stacking the K^2 element-wise products of the columns of B. With the Cholesky factorisation $Q_i = R_i^T R_i$ and $R_i^T z_i = c_i$, each problem becomes a standard NNLS problem $\min_{a \geq 0} \|R_i a - z_i\|^2$, solved exactly with the Lawson–Hanson active-set algorithm.

Platform correction and iteration

The fit is initialised with unweighted NNLS and run for three rounds. From the second round on, a gene-level platform factor is estimated by comparing the mean observed spatial profile $\bar{Y}_g = \frac{1}{N}\sum_i y_{ig}$ with the profile predicted from the mean composition $\bar{a} = \frac{1}{N}\sum_i \hat{a}_i$, $\quad \bar{\mu}_g = \sum_k B_{gk}\bar{a}_k$, both normalised to sum to one:

$$g_g = \mathrm{clip}\left(\exp\left(0.5 \cdot \left[\log\frac{\bar{Y}_g}{\bar{\mu}_g} - \mathrm{median}_{g'} \log\frac{\bar{Y}_{g'}}{\bar{\mu}_{g'}}\right]\right), 0.2, 5\right) \tag{5}$$

after which the signature is updated as $B_{gk} \leftarrow B_{gk} g_g$ and its columns are renormalized. Weights are recomputed at every round.

From RNA shares to cell fractions

The fitted vectors are RNA shares on the marker genes. Restricting B to the marker set and renormalising rescales column k by $1/f_k$, where $f_k = \sum_{g \in \mathrm{markers}} B_{gk}$ is the fraction of type k's RNA captured by the markers. Writing A_{ik} for the fitted RNA share of type k in spot i, cell fractions are therefore obtained by dividing by the product of cell size and capture and renormalising:

$$W_{ik} = \frac{A_{ik}/(s_k f_k)}{\sum_t (A_{it}/(s_t f_t))} \tag{6}$$

Local niches within a section

Compositions are compared on the simplex with the Aitchison distance [17], the Euclidean distance between centred log-ratio (CLR) transforms, after adding a pseudocount $\varepsilon = 10^{-6}$ and re-closing:

$$\mathrm{clr}(w_i)_k = \log w_{ik} - \frac{1}{K}\sum_l \log w_{il}, \quad d_{\mathrm{Ait}}(i,j) = \left\|\mathrm{clr}(w_i) - \mathrm{clr}(w_j)\right\|_2 \tag{7}$$

Cell types that are absent from the whole section are removed first. The spatial distance $d_{\mathrm{spat}}(i,j)$ is the Euclidean distance between spot coordinates (a Gaussian-kernel distance is available but was not used). Each distance matrix is divided by its maximum and the two are combined,

$$D(i,j) = \alpha \cdot \frac{d_{\mathrm{Ait}}(i,j)}{\max d_{\mathrm{Ait}}} + (1-\alpha) \cdot \frac{d_{\mathrm{spat}}(i,j)}{\max d_{\mathrm{spat}}}, \quad \alpha = 0.7 \tag{8}$$

The fused matrix is embedded by classical PCoA with up to 30 dimensions, and a shared-nearest-neighbour (SNN) graph ($k = 20$) is built on the embedding. Louvain clustering is run at resolutions from 0.3 to 3.0 in steps of 0.1, and the partition with the highest modularity is retained. Sections with fewer than two cell types present, fewer than ten spots or a constant fused distance are returned as a single niche. The prototype of niche j is its mean composition, $P_{jk} = \frac{1}{|S_j|}\sum_{i \in S_j} W_{ik}$.

Meta-niches across sections

The prototypes of all sections are stacked into a prototype × cell-type matrix (cell types absent from a section are set to zero), log-normalised (scale factor 10^4) and scaled, and reduced to 15 principal components. Harmony corrects the components for the section of origin, a SNN graph is built ($k = 20$, capped at one third of the number of prototypes), and Louvain clustering assigns each prototype to a meta-niche. Each spot inherits the meta-niche of its local niche. The atlases comprised 2,221 (breast cancer) and 2,283 (glioblastoma) prototypes.

Simulation framework

The simulator has four layers, which include a single-cell reference, spatial patterns, a reference–target mismatch and spot expression, all driven by fixed random seeds.

Reference. Global abundances of K cell types, a fraction of which are rare, were drawn as $\pi_k = \exp\big(N(\mu, \sigma^2)\big)$ with $\mu = -2.6$ for main and -4.5 for rare types (about 5–10% and 0.1–2%), with a floor of 0.001 so that rare types retain about six cells among 6,000, and renormalised. Types were grouped into families; family and type profiles were drawn from $\mathrm{NB}(\mu = 0.5, \mathrm{size} = 0.5)$ and mixed as $\mu_{gk} = \rho P_{g,f(k)} + (1-\rho) O_{gk}$, with collinearity ρ ($\rho = 0.5$ is close to the within-family correlation of real tissues). Each family received shared markers (40% of the marker set) and each type its own markers, with increments drawn from $N(2, 0.5^2)$. Cells were assigned types from π and expression $X_{gj} \sim \mathrm{NB}\big(\mu_{gk} c_j, \mathrm{size} = 0.3\big)$, with cell-size factors $c_j = \exp\big(N(0, 0.4^2)\big)$.

Spatial patterns. Compositions were generated on coordinates in $[0,1]^2$ as gradients (a direction per type), boundaries (nearest of K centres, with a purity parameter), patches ($3K$ centres, Gaussian kernels of width 0.15) or mixed hotspots, in which $w_i \sim \mathrm{Dirichlet}(\alpha_i)$ with $\alpha_{ik} = \alpha_0 + 1.5 \cdot \frac{\exp\left(-\|x_i - c_k\|^2/(2\sigma^2)\right)}{\sum_{k'} \exp\left(-\|x_i - c_{k'}\|^2/(2\sigma^2)\right)}$, $\alpha_0 = 0.3$ and $\sigma = 0.25$.

Reference–target mismatch. Scenario I added independent noise, $B_{gk}^{\mathrm{ref}} = \max\Big(0, B_{gk} + N\Big(0, \delta^2\big(B_{gk} + 0.1\big)^2\Big)\Big)$. Scenario II applied a gene factor shared across types plus type-specific perturbations on 10% of genes, $B_{gk}^{\mathrm{ref}} = B_{gk} \gamma_g \varepsilon_{gk}$ with $\gamma_g = \exp\Big(N\big(0, \sigma_{\mathrm{plat}}^2\big)\Big)$, where σ_{plat} was calibrated by bisection so that the realised log divergence matched scenario I at the same δ. Scenario III re-estimated the signature from a random 60% of the reference cells, so that marker selection could not read the true signature. An optional state shift multiplied 10% of genes by $\exp\big(N(0, \tau^2)\big)$.

Spot expression. Half of the genes formed the spatial panel. Spatial gene efficiencies $\gamma_g = \exp\Big(N\big(0, \sigma_{\mathrm{plat}}^2\big)\Big)$ were applied independently of the reference. The number of cells per spot was $n_i = \mathrm{clamp}\big(\mathrm{round}\big(\mathrm{LogNormal}(\log(8), 0.6^2)\big), 1, 25\big)$. Expected compositions combined the spatial pattern with the global abundances, $\alpha_{ik} = 2.5[0.35 w_{ik} + 0.65 \pi_k] + 0.03$, with 2% near-pure spots (dominant concentration 15); realised cell counts were drawn from a multinomial, and the realised fractions served as truth. Each cell's expression was drawn as $\mathrm{NB}\Big(B_{g,l}^{\mathrm{spot}} c_j, \mathrm{size} = 0.3\Big)$ with $c_j = \exp\big(N(0, 0.3^2)\big)$ and summed within the spot. Technical noise was then added in order: dropout with retention probability $1 - \exp(-Y/1.2)$, log-normal depth factors ($\sigma = 0.8$), ambient RNA drawn from the pooled profile at 8% of each spot's counts, and log-normal noise ($\sigma = 0.1$), after which values were rounded.

Imaging-derived pseudo-spots

Segmented cells were aggregated within circular capture areas on a regular square grid. The grid covered the bounding box of all cells extended by one spot diameter; each cell was assigned to its nearest grid centre (kd-tree search) and retained only if it lay within the capture radius, so that no retained cell contributed to more than one spot. For the same-section benchmark the capture diameter and centre spacing were 55 and 100 coordinate units, and spots with at least three cells were kept. Expression was aggregated with a sparse cells × spots indicator matrix, $Y = XS$, and cell-type counts as $C = S^T T$, giving the true

fractions $w_{nk} = \frac{C_{nk}}{\sum_k C_{nk}}$. When coordinate units or cell densities gave too few or too many cells per spot, the diameter and spacing were rescaled together by the square root of the ratio between the target and the current median number of cells, with a density-based initialisation $d = 2\sqrt{\frac{m_{\mathrm{tgt}}}{\rho\pi}}$ for extreme units. This geometry resembles Visium but does not reproduce its hexagonal layout or the capture and diffusion of real spots. The cross-platform benchmark used pseudo-spots provided with the platform-comparison study [19]; their coordinate units and whether diameter and spacing were scaled by cell occupancy are not fully documented. In the same-section benchmark, reference cells were sampled from the section that was aggregated and were not excluded from the target pool, so reference and target overlap. Truth is conditional on the segmentation and annotation used.

Single cell RNA-seq, spatial transcriptomics, and bulk RNA-seq datasets

The same-section benchmark used six Xenium datasets: breast cancer with the 313-gene and 5K panels, colorectal cancer with a 480-gene panel, lymph node with the 377-gene and 5K panels, and renal cell carcinoma with a combined gene and protein panel. The first breast cancer dataset corresponds to GSM7780153 and the integrated study of Janesick et al.; the others are public 10x Genomics releases, analyzed with their accompanying preprocessing and annotation tables. Fine annotations contained 20–48 labels per dataset and major-type annotations 8 or 9. The cross-platform benchmark used the colorectal (COAD), hepatocellular (HCC) and ovarian carcinoma (OV) data of Ren et al., with CosMx SMI 6K, Xenium 5K and matched scRNA-seq [19]; each tumor type corresponds to one patient, profiled on adjacent sections.

The breast cancer atlas included public 10x Genomics data and the studies of Wu et al., Andersson et al., Bassiouni et al., Mo et al., Wang et al. and Li et al.[18, 30-34], with the 49 cell-types and cell-state labels of the Wu breast cancer atlas as reference. The brain tumor atlas included public 10x Genomics data and the studies of Ravi et al., Mei et al., Ren et al., Lv et al. and Lin et al.[26, 35-38], together with the GBMSpace resource described by de Jong and colleagues [39]; the reference was the integrated glioblastoma single-cell object associated with the Lin et al. study and Zenodo17117252. GSE253080 corresponds to the Lv et al. dataset with 16 sections from three patients. (Supplementary Table S4).

External expression cohorts were TCGA-BRCA, METABRIC, TCGA-GBM, three CGGA series and the pre-treatment I-SPY2 dataset GSE194040. Complete-case survival analyses included 1,068 TCGA-BRCA, 1,980 METABRIC and 169 TCGA-GBM records, and 313, 285 and 657 records of CGGA mRNA-seq 325, microarray 301 and mRNA-seq 693. The I-SPY2 response table contained 988 evaluable records.

Comparator methods and missing results

The main figures compare SpaCoDA, RCTD, CARD, Seurat label transfer, Tangram, SpatialDWLS, SMART, SpaTrio and SPOTlight. We used RCTD in full mode, SCT-based Seurat anchor transfer and Tangram in cluster mode. Feature selection, normalization, cell-number inputs and initialization followed the analysis wrapper of each method; for example, SMART markers in the simulations came from the mean matrix used to generate the reference, whereas SpaCoDA selected markers from reference cells, and some mapping methods received pseudo-spot cell numbers while SpatialDWLS received their median. The comparison therefore evaluates analysis pipelines with method-specific inputs. Exploratory analyses of cell2location, Redeconve and CytoSPACE were not included in the rankings, and no comparative conclusions are drawn for them. Failed runs were kept as missing values: RCTD failed on two high-difficulty main configurations in each scenario (25 of 27 available; 121 of 123 in the full design), and

SPOTlight failed on three of the 12 same-section dataset–level combinations and on the Xenium ovarian dataset.

Accuracy metrics

For every aligned cell type, Pearson and Spearman correlations were computed across spots and macro-averaged over the types for which they are defined (types with zero variance were excluded). RMSE was pooled over all aligned spot × type entries:

$$\mathrm{RMSE} = \sqrt{\left(\frac{1}{NK}\sum_i \sum_k (\hat{p}_{ik} - p_{ik})^2\right)} \tag{9}$$

With missing values, the mean was taken over available squared errors. JSD compares complete composition vectors within a spot and was averaged over spots [46]:

$$\mathrm{JSD}(p,q) = \frac{1}{2}\mathrm{KL}(p\|m) + \frac{1}{2}\mathrm{KL}(q\|m), \quad m = \frac{p+q}{2}, \quad \mathrm{KL}_b(p\|m) = \sum_k p_k \log_b \frac{p_k}{m_k} \tag{10}$$

with $0\log(0/m_k) = 0$ (the imaging benchmarks added 10^{-12} to each component before closure). The divergence itself, not its square root, is reported.

Runtime and memory

Runtime was summarized from per-dataset run logs. Supplementary Table S3 reports median runtimes and the ratio of each method's median to that of SpaCoDA; the fold labels in Fig. 4E use geometric means. Failed runs were not imputed. Memory was the peak proportional set size of the process tree minus the data-loading baseline, including tracked R and Python child processes, summarized as mean ± standard deviation; accelerator memory was not measured. The environment was a 32-core Linux host with 384 GiB of memory and R 4.4.1.

Marker genes and functional annotation

Cell-type enrichment of meta-niches was evaluated on local prototypes grouped by meta-niche. Marker genes were identified from the spatial expression matrix and spot-level meta-niche labels with the Wilcoxon test and area under the curve (AUC) statistics of presto. In the breast cancer analysis, genes were processed in blocks of 2,000 and the adjusted P values returned by each block were kept. Because spots within a section are correlated, these tests are descriptive and do not constitute differential expression with patients as replicates. For functional analysis, genes with adjusted $P < 0.05$, a mean difference of at least 0.5 and an AUC of at least 0.6 were ranked by effect size, and up to 20 genes per meta-niche were selected; Gene Ontology biological-process enrichment used clusterProfiler and org.Hs.eg.db with Benjamini–Hochberg correction[40].

Clinical association analyses

For each meta-niche, the selected marker genes defined a signature that was scored in each bulk-expression sample by using ssGSEA (single-sample GSEA)[41]. For survival, complete cases were dichotomized at the cohort median (scores strictly above the median formed the high group), and Kaplan–Meier curves, log-rank tests and Cox proportional-hazards models were fitted. Intervals are 95% Wald confidence intervals and P values are two-sided and unadjusted for the wider screen of clinical features. For I-SPY2, scores were compared between records with and without pCR by two-sided Wilcoxon rank-sum tests.

Data availability

All data analysed are publicly available: Xenium datasets from 10x Genomics and GEO (GSM7780153); the platform-comparison data of Ren et al.; breast cancer and glioblastoma spatial datasets from the original publications and repositories listed in Supplementary Table S4, including GSE253080 and the GBMSpace resource [43]; the glioblastoma single-cell reference from Zenodo record 17117252; TCGA-BRCA and TCGA-GBM from UCSC Xena, METABRIC from cBioPortal (study brca_metabric); CGGA bulk RNA-seq data were obtained from the CGGA database (http://www.cgga.org.cn); and I-SPY2 expression and associated clinical annotations were downloaded from GEO under accession GSE194040.

Code availability

SpaCoDA is an open-source R package, available at https://github.com/EvoSpatialTool/SpaCoDA.

Acknowledgements

We thank everyone who has publicly shared their spatial transcriptomics data, all the data source providers whose data we used, and all the patients and researchers involved in contributing the data.

Funding

Not applicable.

Author contributions

XK.F. conceived and designed the study, developed the algorithm, conducted the analysis, drafted the manuscript and secured supporting. L.L. revised the manuscript and provided extensive advice and guidance on niche identification and analysis.

Competing interests

The authors declare no potential conflicts of interest.

References

1. Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al: **Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.** *Science* 2016, **353:**78–82.
2. Janesick A, Shelansky R, Gottscho AD, Wagner F, Williams SR, Rouault M, Beliakoff G, Morrison CA, Oliveira MF, Sicherman JT, et al: **High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis.** *Nat Commun* 2023, **14:**8353.
3. He S, Bhatt R, Brown C, Brown EA, Buhr DL, Chantranuvatana K, Danaher P, Dunaway D, Garrison RG, Geiss G, et al: **High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging.** *Nat Biotechnol* 2022, **40:**1794–1806.
4. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, Irizarry RA: **Robust decomposition of cell type mixtures in spatial transcriptomics.** *Nat Biotechnol* 2022, **40:**517–526.
5. Ma Y, Zhou X: **Spatially informed cell-type deconvolution for spatial transcriptomics.** *Nat Biotechnol* 2022, **40:**1349–1359.
6. Elosua-Bayes M, Nieto P, Mereu E, Gut I, Heyn H: **SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.** *Nucleic Acids Res* 2021, **49:**e50.
7. Dong R, Yuan GC: **SpatialDWLS: accurate deconvolution of spatial transcriptomic data.** *Genome Biol* 2021, **22:**145.
8. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe A, Zhang M, et al: **Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram.** *Nat Methods* 2021, **18:**1352–1362.
9. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R: **Comprehensive Integration of Single-Cell Data.** *Cell* 2019, **177:**1888–1902 e1821.
10. Yang CX, Sin DD, Ng RT: **SMART: spatial transcriptomics deconvolution using marker-gene-assisted topic model.** *Genome Biol* 2024, **25:**304.
11. Yang P, Jin L, Liao J, Jin K, Shao X, Li C, Qian J, Cheng J, Yu D, Guo R, et al: **Revealing spatial multimodal heterogeneity in tissues with SpaTrio.** *Cell Genom* 2023, **3:**100446.
12. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, Hu Y, Zhang X, Yao X, Tang M, et al: **Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution.** *Nat Methods* 2022, **19:**662–670.
13. Sang-Aram C, Browaeys R, Seurinck R, Saeys Y: **Spotless, a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics.** *Elife* 2024, **12**.
14. Varrone M, Tavernari D, Santamaria-Martinez A, Walsh LA, Ciriello G: **CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity.** *Nat Genet* 2024, **56:**74–84.
15. Qian J, Shao X, Bao H, Fang Y, Guo W, Li C, Li A, Hua H, Fan X: **Identification and characterization of cell niches in tissue from spatial omics data at single-cell resolution.** *Nat Commun* 2025, **16:**1693.
16. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S: **Fast, sensitive and accurate integration of single-cell data with Harmony.** *Nat Methods* 2019, **16:**1289–1296.
17. Ren P, Zhang R, Wang Y, Zhang P, Luo C, Wang S, Li X, Zhang Z, Zhao Y, He Y, et al: **Systematic benchmarking of high-throughput subcellular spatial transcriptomics platforms across human tumors.** *Nat Commun* 2025, **16:**9232.
18. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, et al: **A single-cell and spatially resolved atlas of human breast cancers.** *Nat Genet* 2021, **53:**1334–1347.
19. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490:**61–70.
20. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG,

Samarajiwa S, Yuan Y, et al: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486:**346–352.
21. Nanda R, Liu MC, Yau C, Shatsky R, Pusztai L, Wallace A, Chien AJ, Forero-Torres A, Ellis E, Han H, et al: **Effect of Pembrolizumab Plus Neoadjuvant Chemotherapy on Pathologic Complete Response in Women With Early-Stage Breast Cancer: An Analysis of the Ongoing Phase 2 Adaptively Randomized I-SPY2 Trial.** *JAMA Oncol* 2020, **6:**676–684.
22. Wolf DM, Yau C, Wulfkuhle J, Brown-Swigart L, Gallagher RI, Lee PRE, Zhu Z, Magbanua MJ, Sayaman R, O'Grady N, et al: **Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies.** *Cancer Cell* 2022, **40:**609–623 e606.
23. Helmink BA, Reddy SM, Gao J, Zhang S, Basar R, Thakur R, Yizhak K, Sade-Feldman M, Blando J, Han G, et al: **B cells and tertiary lymphoid structures promote immunotherapy response.** *Nature* 2020, **577:**549–555.
24. Cabrita R, Lauss M, Sanna A, Donia M, Skaarup Larsen M, Mitra S, Johansson I, Phung B, Harbst K, Vallon-Christersson J, et al: **Tertiary lymphoid structures improve immunotherapy and survival in melanoma.** *Nature* 2020, **577:**561–565.
25. Petitprez F, de Reynies A, Keung EZ, Chen TW, Sun CM, Calderaro J, Jeng YM, Hsiao LP, Lacroix L, Bougouin A, et al: **B cells are associated with survival and immunotherapy response in sarcoma.** *Nature* 2020, **577:**556–560.
26. Lin J, Chen C, Li S, Chen L, Fang M, Fu S, Chen K, Xu H, Zhang F, Wang R, et al: **Spatial and single-cell characterization of human glioblastoma tumor microenvironment reveals malignant cellular communities.** *Nat Neurosci* 2026, **29:**1342–1354.
27. Greenwald AC, Darnell NG, Hoefflin R, Simkin D, Mount CW, Gonzalez Castro LN, Harnik Y, Dumont S, Hirsch D, Nomura M, et al: **Integrative spatial analysis reveals a multi-layered organization of glioblastoma.** *Cell* 2024, **187:**2485–2501 e2426.
28. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, et al: **The somatic genomic landscape of glioblastoma.** *Cell* 2013, **155:**462–477.
29. Zhao Z, Zhang KN, Wang Q, Li G, Zeng F, Zhang Y, Wu F, Chai R, Wang Z, Zhang C, et al: **Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients.** *Genomics Proteomics Bioinformatics* 2021, **19:**1–12.
30. Andersson A, Larsson L, Stenbeck L, Salmen F, Ehinger A, Wu SZ, Al-Eryani G, Roden D, Swarbrick A, Borg A, et al: **Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions.** *Nat Commun* 2021, **12:**6012.
31. Bassiouni R, Idowu MO, Gibbs LD, Robila V, Grizzard PJ, Webb MG, Song J, Noriega A, Craig DW, Carpten JD: **Spatial Transcriptomic Analysis of a Diverse Patient Cohort Reveals a Conserved Architecture in Triple-Negative Breast Cancer.** *Cancer Res* 2023, **83:**34–48.
32. Mo CK, Liu J, Chen S, Storrs E, Targino da Costa ALN, Houston A, Wendl MC, Jayasinghe RG, Iglesia MD, Ma C, et al: **Tumour evolution and microenvironment interactions in 2D and 3D space.** *Nature* 2024, **634:**1178–1186.
33. Wang X, Venet D, Lifrange F, Larsimont D, Rediti M, Stenbeck L, Dupont F, Rouas G, Garcia AJ, Craciun L, et al: **Spatial transcriptomics reveals substantial heterogeneity in triple-negative breast cancer with potential clinical implications.** *Nat Commun* 2024, **15:**10232.
34. Li T, Yang Q, Acs B, Sifakis EG, Toosi H, Engblom C, Thrane K, Lin Q, Mold JE, Sun W, et al: **Computational pathology annotation enhances the resolution and interpretation of breast cancer spatial transcriptomics data.** *NPJ Precis Oncol* 2025, **9:**310.
35. Ravi VM, Will P, Kueckelhaus J, Sun N, Joseph K, Salie H, Vollmer L, Kuliesiute U, von Ehr J, Benotmane JK, et al: **Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma.** *Cancer Cell* 2022, **40:**639–655 e613.
36. Mei Y, Wang X, Zhang J, Liu D, He J, Huang C, Liao J, Wang Y, Feng Y, Li H, et al: **Siglec-9 acts as an immune-checkpoint molecule on macrophages in glioblastoma, restricting T-cell priming and immunotherapy response.** *Nat Cancer* 2023, **4:**1273–1291.
37. Ren Y, Huang Z, Zhou L, Xiao P, Song J, He P, Xie C, Zhou R, Li M, Dong X, et al: **Spatial transcriptomics reveals niche-specific enrichment and vulnerabilities of radial glial stem-like cells in malignant gliomas.** *Nat Commun* 2023, **14:**1028.

38. Lv X, Wang B, Liu K, Li MJ, Yi X, Wu X: **Decoding heterogeneous and coordinated tissue architecture in glioblastoma using spatial transcriptomics.** *iScience* 2024, **27:**110064.
39. Grant de Jong FM, Tannia Gracia, Olga Lazareva, Oliver Gould, Alexander Aivazidis, Manas Dave, Qianqian Zhang, Melanie Jensen, Ahmet Sureyya Rifaioglu, Joao D. Barros-Silva, Sabine Eckert, Di Zhou, Yvette Wood, Elizabeth Tuck, Sezgin Er, Henry Marshall, Kenny Roberts, Andrew L. Trinh, Shreya Rai, Tyler Shaw, Agnes Oszlanczi, Hayden Powell, Robert Petryszak, Zoi Katsirea, Irfan Mamun, Ilaria Mulas, Annelies Quaegebeur, Mayen Briggs, Stanislaw Makarchuk, Jessica Cox, Jimmy Tsz Hang Lee, Laura Rueda, Manu Saraswat, Harry Bulstrode, Adam Young, Minal Patel, Tarryn Porter, Elena Prigmore, Moritz Mall, Julio Saez-Rodriguez, James Briscoe, David H. Rowitch, Richard Mair, Sam Behjati, Oliver Stegle, Omer Ali Bayraktar: **A spatiotemporal cancer cell trajectory underlies glioblastoma heterogeneity.** *biorxiv* 2025.
40. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *OMICS* 2012, **16:**284–287.
41. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data.** *BMC Bioinformatics* 2013, **14:**7.

Figures

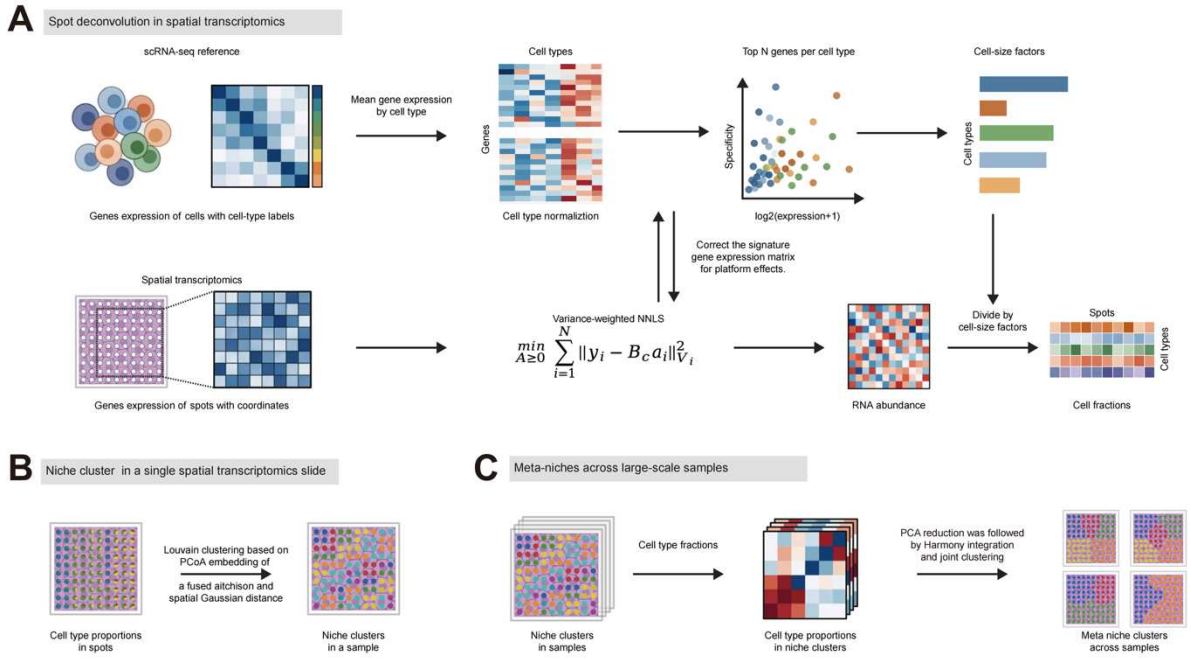


Fig. 1 | SpaCoDA deconvolution and two-level niche identification. (A) Annotated reference cells are used to build the cell-type signature, the marker-gene set and the effective cell-size factors. Iterative variance-weighted non-negative regression with gene-level correction estimates RNA contributions, which are converted to cell fractions. (B) Within each section, the Aitchison compositional distance is combined with the normalized Euclidean distance between spot coordinates, embedded by principal coordinate analysis (PCoA), and used to build a shared-nearest-neighbor (SNN) graph, among candidate resolutions, the Louvain partition with the highest modularity is selected. (C) The mean compositions of local niches are normalized, reduced by principal component analysis (PCA), integrated with Harmony and clustered jointly; meta-niche labels are mapped back to member spots. Created in BioRender. Fan, H. (2027) https://BioRender.com/acn70zj.

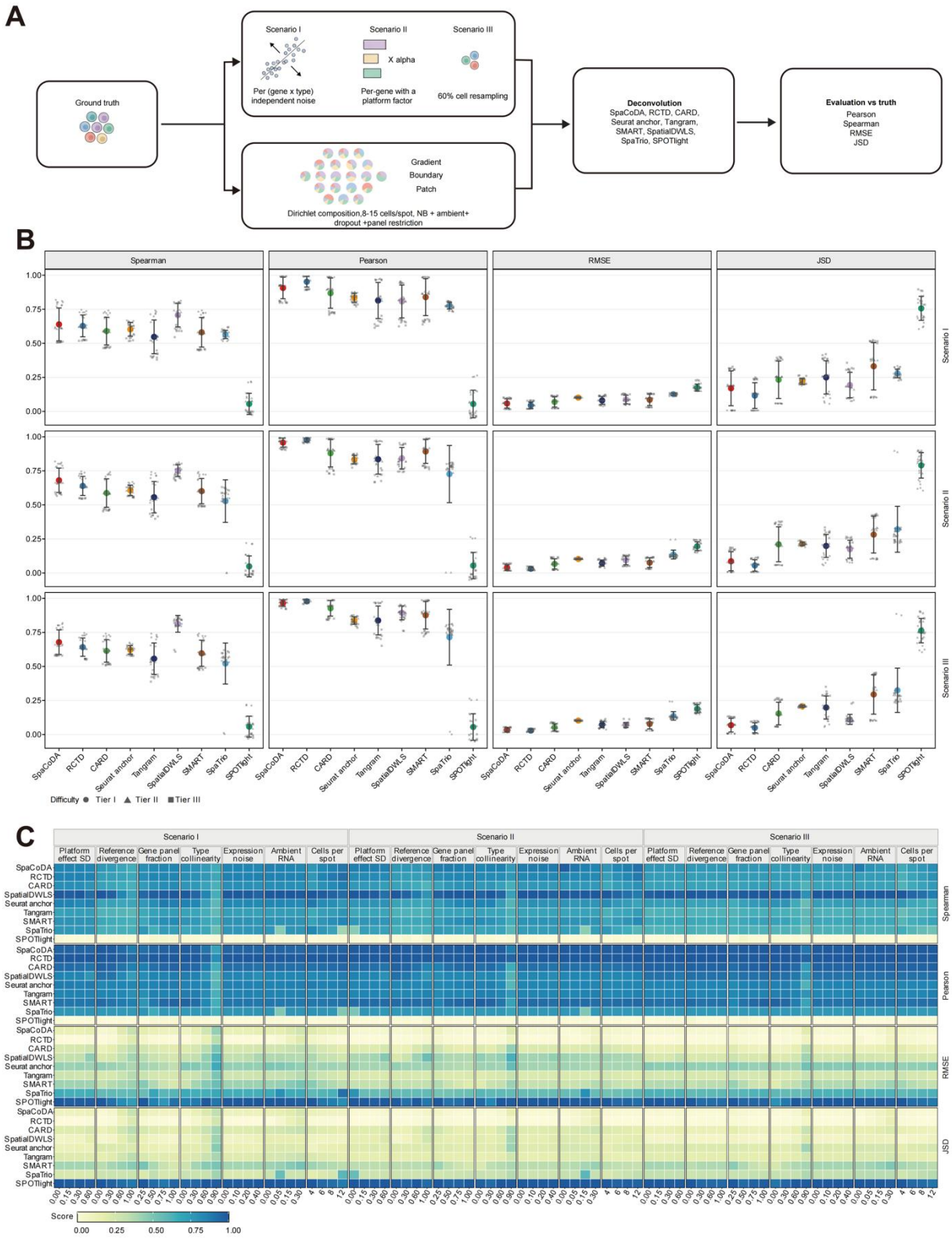


Fig. 2 | Simulation design and sensitivity to reference mismatch. (A) Simulated targets combine spatial composition patterns, sampled cell numbers, negative-binomial expression and technical perturbations. Each target is paired with a reference perturbed independently per gene and cell type (scenario I), by gene-level platform factors (scenario II) or by resampling 60% of the reference cells (scenario III). The overall divergence of scenarios I and II is matched; the resampled reference diverges less. Created in BioRender. Fan, H. (2027) https://BioRender.com/6zwq6hx. (B) Metrics for the 27 main configuration–seed combinations in each scenario (RCTD: 25 successful runs; all other methods: 27). Boxes show the median and interquartile range, whiskers extend to observations within 1.5 times the interquartile range, points

show runs and point shape indicates difficulty tier. Higher Spearman and Pearson correlations and lower RMSE and JSD indicate closer agreement with the realized cell-count fractions. (C) Single-factor perturbations; each cell summarizes three seeds. Colors are scaled to the raw values within each metric × scenario panel, so a darker color means a larger correlation, RMSE or JSD — better for correlations but worse for errors.

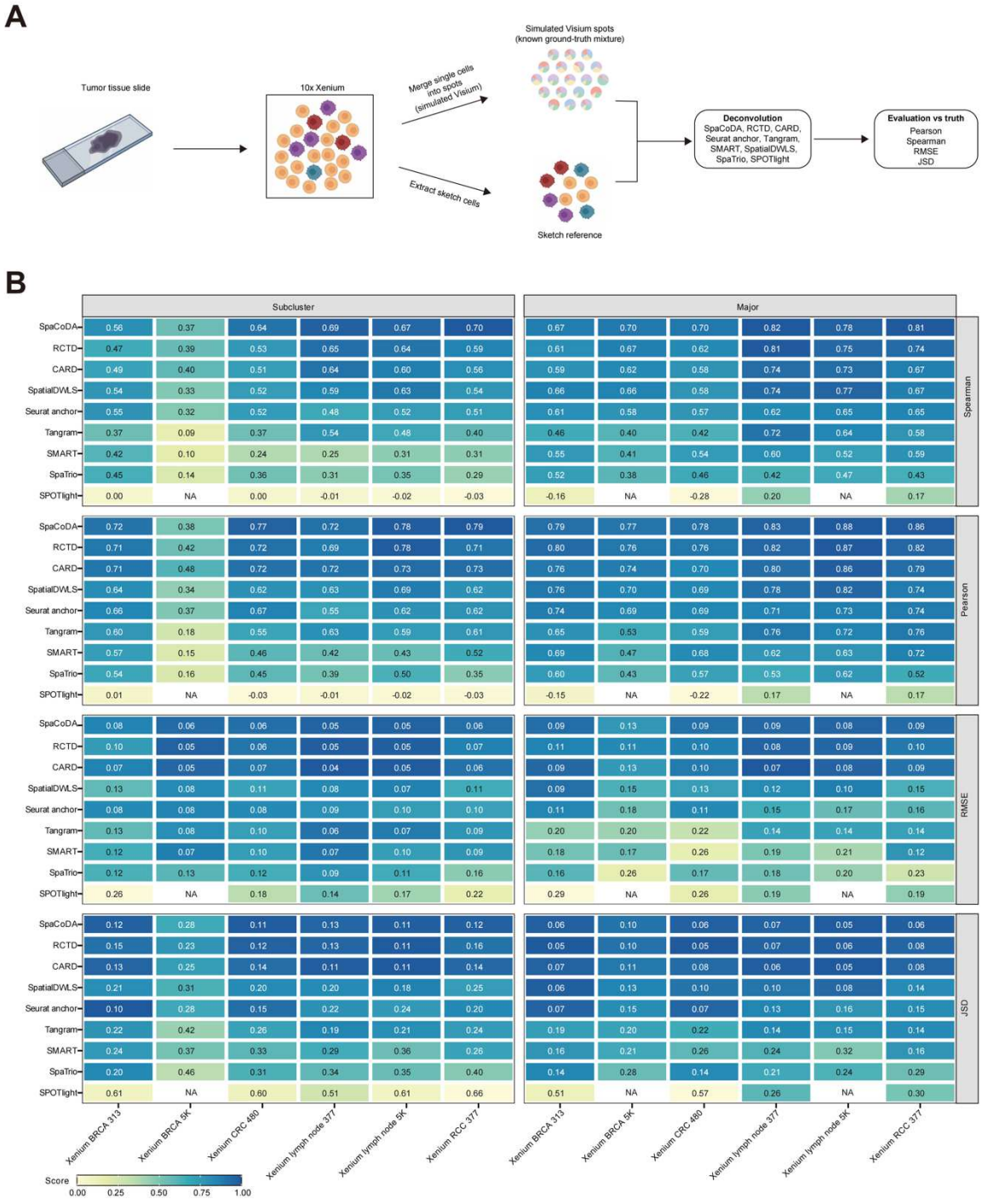


Fig. 3 | Same-section Xenium benchmark at two annotation levels. (A) Segmented Xenium cells were aggregated into pseudo-spots by capture area, and up to 5,000 sketched cells from the same section were used as reference; these cells remained in the target pool. (B) Spearman and Pearson correlation, pooled RMSE and JSD for six datasets at two annotation levels. Correlations are averaged over evaluable types and JSD over spotsl. Numbers are metric values, darker shading indicates better performance within each metric. NA, failed run. The six datasets contain 25,293 pseudo-spots, truth is defined by the original cell segmentation and annotation.

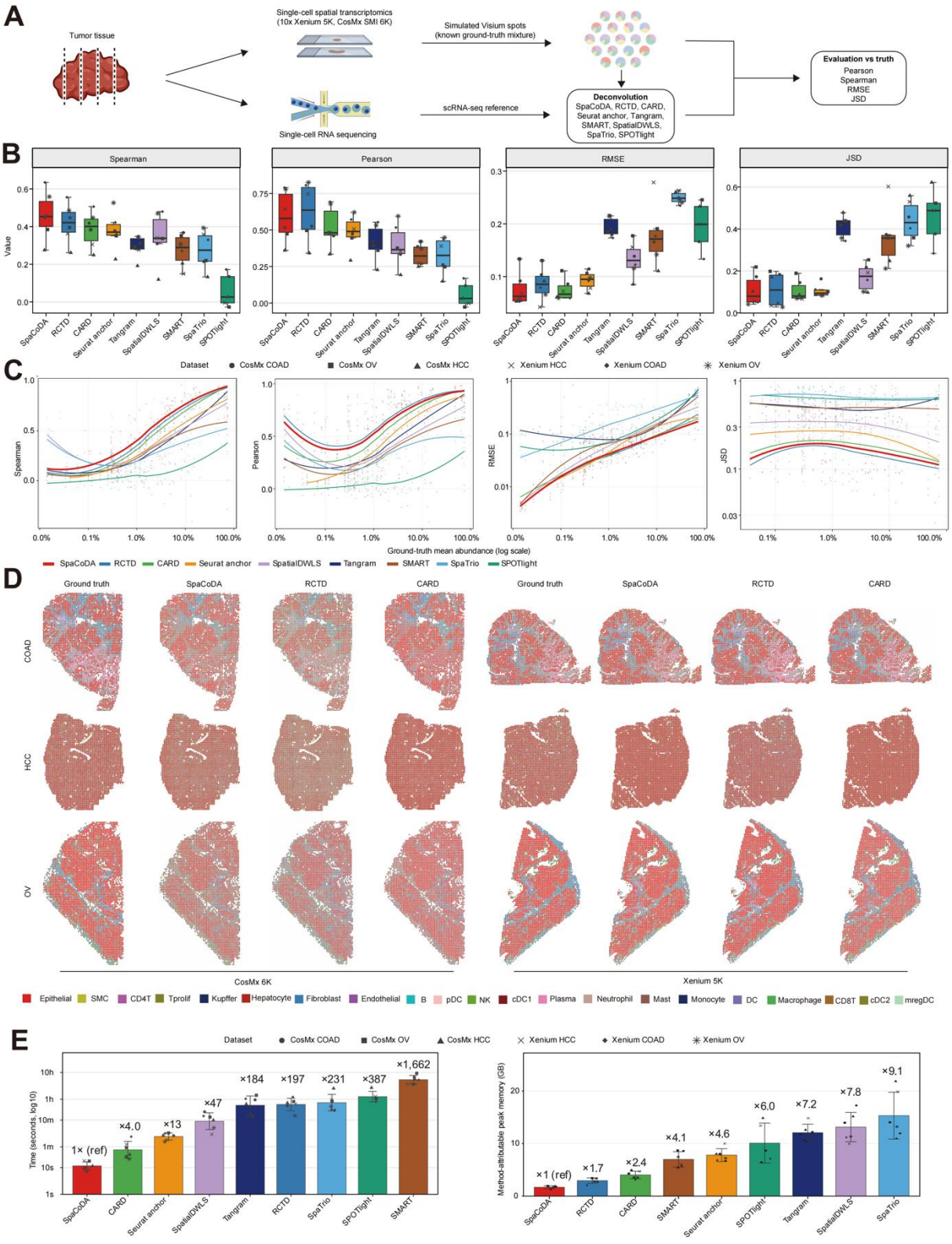


Fig. 4 | Cross-platform accuracy and computational cost. (A) CosMx 6K and Xenium 5K measurements were aggregated into pseudo-spots and deconvolved with the matched scRNA-seq reference. The six platform × tissue datasets come from three patients with colorectal (COAD), hepatocellular (HCC) or ovarian (OV) carcinoma. (B) Dataset-level accuracy. Boxes show the median and interquartile range, whiskers extend to observations within 1.5 times the interquartile range; symbols distinguish datasets (n = 6 for most methods, 5 for SPOTlight). (C) Cell-type accuracy against mean true abundance, with descriptive smooth fits; the conditional JSD is the mean complete-vector divergence over spots containing the type. (D) Composition maps of the annotated truth, SpaCoDA, RCTD and CARD; sectors show cell

fractions and colours cell lineages. (E) Runtime and incremental peak process-tree memory. Runtime fold labels are geometric means. Memory bars show mean ± s.d. of six measurements (five for SPOTlight).

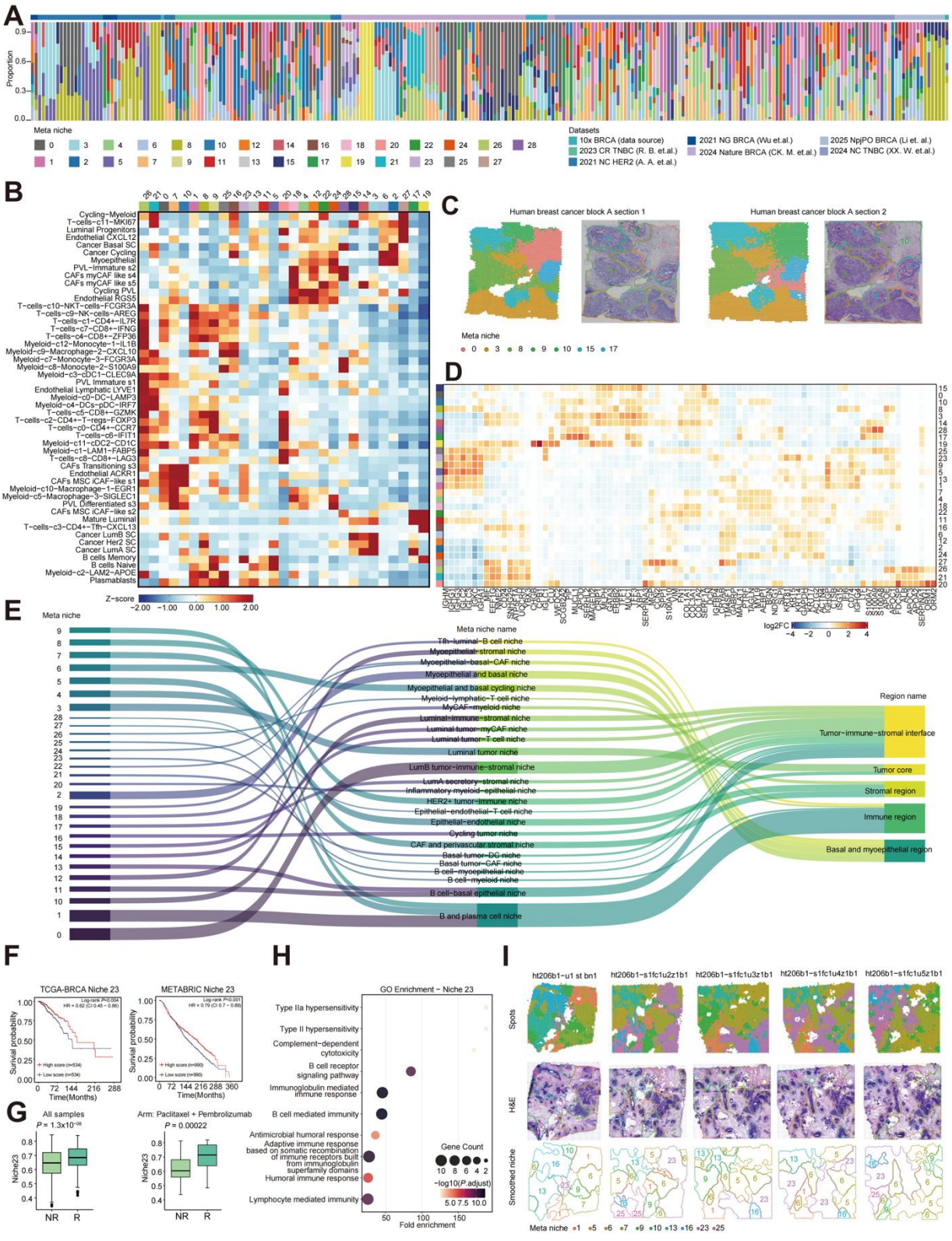


Fig. 5 | Breast cancer meta-niche atlas and clinical associations of the expression signatures. (A) Meta-niche composition of each section. The full analysis set contains 254 section identifiers from seven sources, 29 meta-niches and 594,341 labelled spots. (B) The heatmap shows the cell types predominantly enriched in each meta-niche. (C) Spot labels and H&E images of representative Human breast cancer block A section1 and section2. (D) Highly expressed genes of the meta-niche. (E) Meta-niche annotations and their grouping into five regions. (F) Overall survival by meta-niche 23 expression score in TCGA-BRCA and METABRIC. (G) Scores by pCR status in 988 I-SPY2 records and in 69 records of the paclitaxel plus pembrolizumab arm, two-sided Wilcoxon tests. (H) Gene ontology biological process

enrichment of the meta-niche 23 genes, point size shows gene count and color the adjusted P value. (I) Spot labels, H&E overlays and smoothed outlines of five HT206B1 sections.

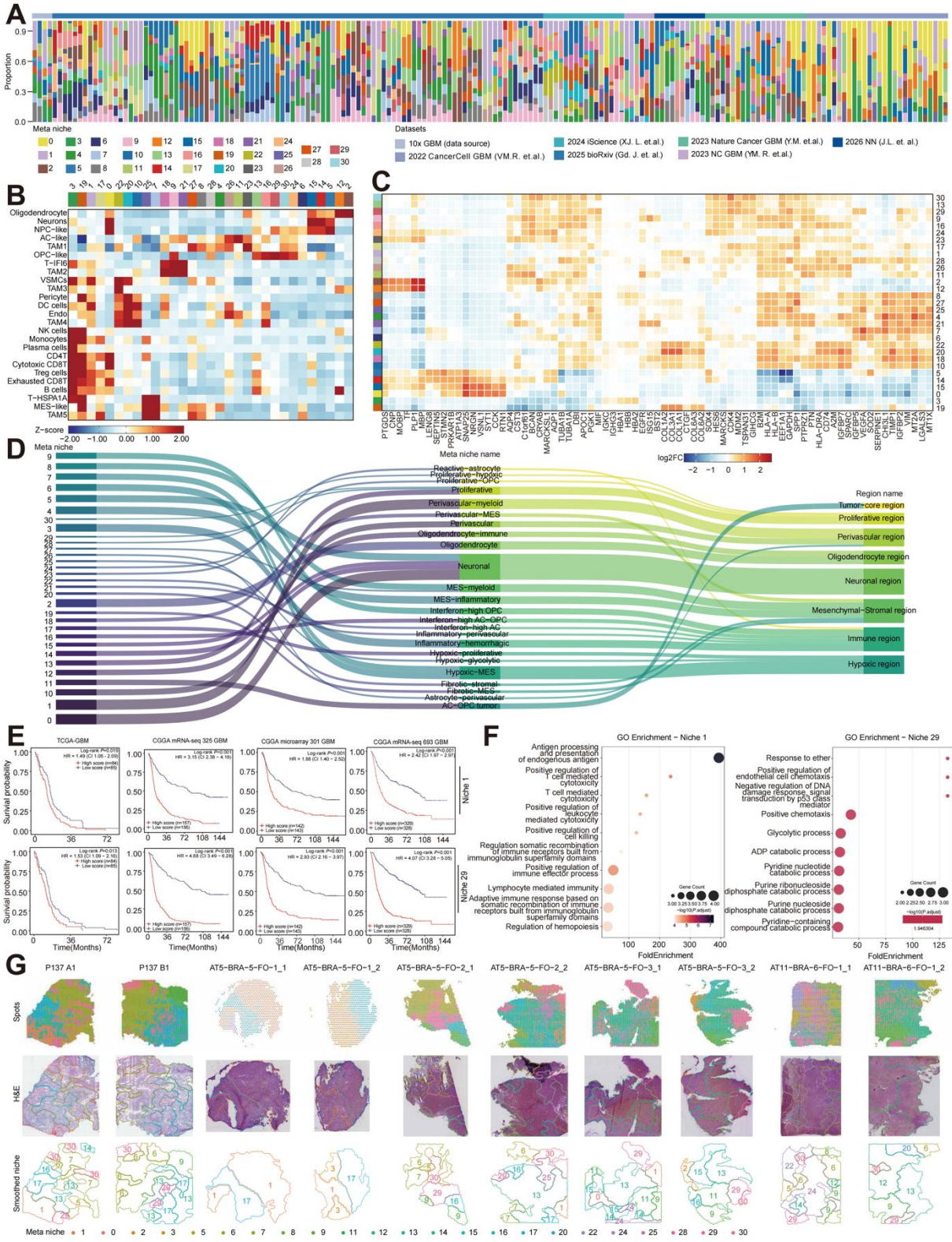


Fig. 6 | Glioblastoma meta-niche atlas and survival associations. (A) Proportions of 31 meta-niches in 181 section identifiers from seven sources, with 585,556 labelled spots. (B) The heatmap shows the cell types predominantly enriched in each meta-niche. (C) Highly expressed genes of the meta-niche. (D) Annotation into eight tissue regions on the basis of composition and expression. (E) Overall survival by meta-niche 1 and 29 expression scores in TCGA-GBM, CGGA mRNA-seq 325, CGGA microarray 301and CGGA mRNA-seq 693. (F) Gene ontology biological-process enrichment. Meta-niche 1 includes antigen-presentation terms; the meta-niche 29 terms shown have adjusted P values of about 0.011, each supported by two or three mapped genes. (G) Spot labels, H&E overlays and smoothed outlines (for

display) of ten representative sections.

Supplementary tables

Supplementary Table S1 | Performance on the main simulated configurations.

Scenario I: independent perturbation

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	27	0.6389	0.9064	0.0582	0.1697
RCTD	25	0.6278	0.9521	0.0445	0.1162
CARD	27	0.5898	0.8682	0.0693	0.2323
Seurat anchor	27	0.6024	0.8337	0.1012	0.2204
Tangram	27	0.5474	0.8145	0.0803	0.2498
SpatialDWLS	27	0.7067	0.8075	0.0871	0.1934
SMART	27	0.5808	0.8390	0.0851	0.3313
SpaTrio	27	0.5649	0.7769	0.1263	0.2789
SPOTlight	27	0.0552	0.0544	0.1746	0.7566

Scenario II: gene-level platform perturbation

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	27	0.6806	0.9573	0.0398	0.0860
RCTD	25	0.6393	0.9762	0.0307	0.0550
CARD	27	0.5870	0.8792	0.0666	0.2104
Seurat anchor	27	0.6069	0.8323	0.1027	0.2135
Tangram	27	0.5563	0.8351	0.0725	0.1987
SpatialDWLS	27	0.7511	0.8425	0.0926	0.1743
SMART	27	0.6010	0.8923	0.0760	0.2818
SpaTrio	27	0.5279	0.7263	0.1352	0.3209
SPOTlight	27	0.0484	0.0542	0.1962	0.7903

Scenario III: reference resampling

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	27	0.6782	0.9658	0.0342	0.0686
RCTD	25	0.6418	0.9786	0.0288	0.0496
CARD	27	0.6141	0.9274	0.0521	0.1544
Seurat anchor	27	0.6219	0.8374	0.1024	0.2077
Tangram	27	0.5570	0.8375	0.0725	0.1983
SpatialDWLS	27	0.8131	0.8936	0.0702	0.1114
SMART	27	0.5968	0.8759	0.0782	0.2943
SpaTrio	27	0.5214	0.7149	0.1368	0.3249
SPOTlight	27	0.0581	0.0554	0.1891	0.7628

Supplementary Table S2 | Performance on the imaging-derived benchmarks.

Same-section Xenium, fine annotation,subclusters (mean)

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	6	0.6039	0.6931	0.0588	0.1446
RCTD	6	0.5439	0.6706	0.0634	0.1482
CARD	6	0.5346	0.6823	0.0580	0.1463
Seurat anchor	6	0.4827	0.5808	0.0874	0.1972
Tangram	6	0.3740	0.5264	0.0901	0.2575
SpatialDWLS	6	0.5227	0.5891	0.0981	0.2253
SMART	6	0.2725	0.4261	0.0906	0.3077
SpaTrio	6	0.3157	0.4005	0.1205	0.3412
SPOTlight	5	−0.0116	−0.0153	0.1934	0.6005

Same-section Xenium, major cell types (mean)

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	6	0.7461	0.8190	0.0948	0.0690
RCTD	6	0.6997	0.8045	0.0987	0.0682
CARD	6	0.6566	0.7727	0.0944	0.0741
Seurat anchor	6	0.6144	0.7165	0.1462	0.1228
Tangram	6	0.5337	0.6663	0.1718	0.1728
SpatialDWLS	6	0.6789	0.7510	0.1253	0.1029
SMART	6	0.5407	0.6396	0.1845	0.2209
SpaTrio	6	0.4490	0.5460	0.1992	0.2162
SPOTlight	3	0.0691	0.0645	0.2265	0.3567

Cross-platform (median)

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	6	0.4533	0.5791	0.0625	0.0793
RCTD	6	0.4209	0.6359	0.0853	0.1087
CARD	6	0.4009	0.4804	0.0665	0.0800
Seurat anchor	6	0.3705	0.4895	0.0948	0.0938
Tangram	6	0.3013	0.4079	0.1877	0.4270
SpatialDWLS	6	0.3378	0.3651	0.1306	0.1744
SMART	6	0.2890	0.3222	0.1713	0.3563
SpaTrio	6	0.2748	0.3260	0.2485	0.4303
SPOTlight	5	0.0258	0.0313	0.1991	0.4876

Cross-platform (mean)

Method	n	Spearman	Pearson	RMSE	JSD
SpaCoDA	6	0.4600	0.5934	0.0760	0.1070
RCTD	6	0.4177	0.6236	0.0854	0.1101
CARD	6	0.3848	0.5210	0.0748	0.1050
Seurat anchor	6	0.3779	0.4848	0.0927	0.1049
Tangram	6	0.2939	0.4148	0.1930	0.4121
SpatialDWLS	6	0.3422	0.3946	0.1325	0.1695
SMART	6	0.2754	0.3283	0.1774	0.3579
SpaTrio	6	0.2747	0.3223	0.2493	0.4378
SPOTlight	5	0.0600	0.0578	0.1954	0.4585

Supplementary Table S3 | Computational cost on the cross-platform series.

Method	n (time)	Median s	Fold	Memory (GB)	n (memory)
SpaCoDA	6	10.8	1.0	1.69 ± 0.27	6
RCTD	6	2,516.8	233.1	2.94 ± 0.52	6
CARD	6	43.0	4.0	4.02 ± 0.72	6
Seurat anchor	6	158.0	14.6	7.80 ± 1.22	6
Tangram	6	2,892.3	267.8	12.06 ± 1.61	6
SpatialDWLS	6	683.8	63.3	13.13 ± 2.79	6
SMART	6	19,630.0	1,817.8	6.98 ± 1.46	6
SpaTrio	6	2,502.5	231.7	15.31 ± 4.46	6
SPOTlight	5	3,786.2	350.6	10.08 ± 3.80	5

Supplementary Table S4 | Sources and size of the spatial atlases.

Breast cancer

Source	Sections	Labelled spots
10x Genomics	8	30,079
Andersson et al. 2021	36	13,530
Wu et al. 2021	6	15,601
Bassiouni et al. 2023	43	56,567
Wang et al. 2024	94	270,310
Mo et al. 2024 (HTAN)	53	171,048
Li et al2025	14	37,206
Total	254	594,341

Glioblastoma

Source	Sections	Labelled spots
10x Genomics	4	23,570
Ravi et al. 2022	28	88,789
Mei et al. 2023	20	44,994
Ren et al.2023	6	16,344
Lv et al. 2024 (GSE253080)	16	44,272
de Jong et al. 2025 (GBMSpace)	97	343,744
Lin et al. 2026	10	23,843
Total	181	585,556

Supplementary Table S5 | Exploratory survival associations of selected signatures.

Cohort	Meta-niche	n / deaths	HR high vs low	Log-rank P
TCGA-BRCA	23	1,068 / 144	0.62 (0.44–0.86)	0.004
METABRIC	23	1,980 / 1,143	0.79 (0.70–0.89)	7.8×10^{-5}
TCGA-GBM	1	169 / 137	1.49 (1.06–2.09)	0.019
TCGA-GBM	29	169 / 137	1.53 (1.09–2.16)	0.013
CGGA mRNA-seq 325	1	313 / 218	3.15 (2.38–4.18)	6.6×10^{-17}
CGGA mRNA-seq 325	29	313 / 218	4.68 (3.49–6.28)	1.6×10^{-28}
CGGA microarray 301	1	285 / 185	1.88 (1.40–2.52)	1.8×10^{-5}
CGGA microarray 301	29	285 / 185	2.93 (2.16–3.97)	4.0×10^{-13}
CGGA mRNA-seq 693	1	657 / 394	2.42 (1.97–2.97)	3.5×10^{-18}
CGGA mRNA-seq 693	29	657 / 394	4.07 (3.28–5.05)	2.1×10^{-42}

Supplementary Table S6 | Exploratory associations with pathological complete response in I-SPY2.

Arm	n	pCR	Median pCR / non-pCR	Wilcoxon P
All records	988	319	0.684 / 0.647	1.3×10^{-8}
Paclitaxel + ganitumab	107	24	0.717 / 0.655	0.018
Paclitaxel + ganetespib	93	26	0.694 / 0.665	0.066
Paclitaxel + MK-2206	60	18	0.698 / 0.660	0.294
Paclitaxel + pembrolizumab	69	31	0.706 / 0.595	2.2×10^{-4}
Paclitaxel	179	31	0.707 / 0.662	0.119
Paclitaxel + AMG 386	114	33	0.701 / 0.627	0.003
T-DM1 + pertuzumab	52	30	0.658 / 0.671	0.491
Paclitaxel + trastuzumab	31	8	0.668 / 0.670	0.842
Paclitaxel + pertuzumab + trastuzumab	44	26	0.648 / 0.617	0.547
Paclitaxel + MK-2206 + trastuzumab	34	17	0.645 / 0.593	0.433
Paclitaxel + AMG 386 + trastuzumab	19	6	0.611 / 0.620	0.966
Paclitaxel + neratinib	114	42	0.675 / 0.639	0.015
Paclitaxel + ABT-888 + carboplatin	71	27	0.689 / 0.623	8.0×10^{-4}