



Cell type-specific inference of differential expression in spatial transcriptomics

Dylan M. Cable^{1,2,3}, Evan Murray², Vignesh Shanmugam^{2,4}, Simon Zhang², Luli S. Zou^{2,3,5}, Michael Diao^{1,2}, Haiqi Chen^{2,6,7}, Evan Z. Macosko^{1,2,8}, Rafael A. Irizarry^{1,3,5,10} and Fei Chen^{1,2,9,10}

A central problem in spatial transcriptomics is detecting differentially expressed (DE) genes within cell types across tissue context. Challenges to learning DE include changing cell type composition across space and measurement pixels detecting transcripts from multiple cell types. Here, we introduce a statistical method, cell type-specific inference of differential expression (C-SIDE), that identifies cell type-specific DE in spatial transcriptomics, accounting for localization of other cell types. We model gene expression as an additive mixture across cell types of log-linear cell type-specific expression functions. C-SIDE's framework applies to many contexts: DE due to pathology, anatomical regions, cell-to-cell interactions and cellular microenvironment. Furthermore, C-SIDE enables statistical inference across multiple/replicates. Simulations and validation experiments on Slide-seq, MERFISH and Visium datasets demonstrate that C-SIDE accurately identifies DE with valid uncertainty quantification. Last, we apply C-SIDE to identify plaque-dependent immune activity in Alzheimer's disease and cellular interactions between tumor and immune cells. We distribute C-SIDE within the R package <https://github.com/dmcable/spacexr>.

Spatial transcriptomics technologies profile gene expression in parallel across hundreds or thousands of genes across spatial measurement units, or pixels^{1–7}. These technologies have the potential to associate gene expression with cellular contexts such as spatial position, proximity to pathology or cell-to-cell interactions. Studying gene expression changes, termed differential expression (DE, that is, differentially expressed genes), within tissue context has the potential to provide insight into principles of organization of complex tissues and disorganization in disease and pathology^{1,8,9}.

Current methods for addressing DE in spatial transcriptomics fall into two categories: nonparametric and parametric methods. Nonparametric DE methods^{10–12} do not use constrained hypotheses about gene expression patterns, but rather fit general smooth spatial patterns of gene expression. Some of these approaches do not take cell types into account¹⁰, while others operate on individual cell types¹². Discovering nonparametric differential gene expression can be advantageous to generate diverse exploratory hypotheses. However, if covariates are available, for example, predefined anatomical regions, parametric approaches increase statistical power substantially and provide directly interpretable parameter estimates. Specific DE problems have been addressed with ad hoc solutions such as detecting gene expression dependent on cell-to-cell colocalization¹³ or anatomical regions¹⁴, but no general parametric framework is currently available. In contrast, general parametric frameworks have been widely applied across bulk and single-cell RNA-sequencing (scRNA-seq) to test for differences in gene expression across cell type, disease state and developmental state, among other problems^{15,16}. Furthermore, although multi-sample, multi-replicate DE methods exist for bulk and scRNA-seq^{15,16}, no statistical framework accounting for technical and biological

variation¹⁷ across samples and replicates has been established for the spatial setting. We refer to samples as spatial transcriptomics experiments that differ in biological conditions (for example, different biological individuals or conditions), whereas replicates are used to describe repeat experiments across identical conditions and biological samples.

An important challenge unaddressed by current spatial transcriptomics DE methods is accounting for observations generated from cell type mixtures. In particular, sequencing-based, RNA-capture spatial transcriptomics technologies, such as Visium⁵, GeoMx⁶ and Slide-seq^{1,2}, can capture multiple cell types on individual measurement pixels. The presence of cell type mixtures complicates the estimation of cell type-specific DE (that is, DE within a cell type of interest) because different cell types have different gene expression profiles, independent of spatial location^{18,19}. Although imaging-based spatial transcriptomics technologies, such as MERFISH³, ExSeq⁷ and STARmap⁴, have the potential to achieve single-cell resolution, these technologies may encounter mixing across cell types due to diffusion or imperfect cellular segmentation²⁰. Several methods^{18,21–23} have been developed to identify cell type proportions in spatial transcriptomics datasets. However, at present no method accounts for cell type proportions in DE analysis. Here, we demonstrate how not accounting for cell type proportions leads to biased estimates of differential gene expression due to cell type proportion changes or contamination from other cell types.

In this work, we introduce cell type-specific inference of DE (C-SIDE), a general parametric statistical method that estimates cell type-specific DE in the context of cell type mixtures. The first step is to estimate cell type proportions on each pixel using a cell

¹Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA. ²Broad Institute of MIT and Harvard, Cambridge, MA, USA.

³Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. ⁴Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. ⁵Department of Biostatistics, Harvard University, Boston, MA, USA. ⁶Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. ⁷Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA. ⁸Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. ⁹Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. ¹⁰These authors contributed equally: Rafael A. Irizarry and Fei Chen.

✉e-mail: rafa@ds.dfcf.harvard.edu; chenf@broadinstitute.org

type-annotated scRNA-seq reference¹⁸. Next, we fit a parametric model, using predefined covariates such as spatial location or cellular microenvironment, that accounts for cell type differences to obtain cell type-specific DE estimates and corresponding standard errors. The model accounts for sampling noise, gene-specific overdispersion, multiple hypothesis testing and platform effects between the scRNA-seq reference and the spatial data. Furthermore, the C-SIDE model permits statistical inference across multiple experimental samples and/or replicates to achieve more stable estimates of population-level differential gene expression.

Using simulated and real spatial transcriptomics data, we show C-SIDE accurately estimates cell type-specific DE while controlling for changes in cell type proportions and contamination from other cell types. We also demonstrate how cell type mixture modeling increases power, especially when single-cell type measurements are rare. Furthermore, on Slide-seq, MERFISH and Visium datasets, we demonstrate how C-SIDE's general parametric framework enables testing differential gene expression for diverse hypotheses including spatial position or anatomical regions²⁴, cell-to-cell interactions, cellular environment or proximity to pathology. By associating gene expression changes with particular cell types, we use C-SIDE to systematically link gene expression changes to cellular context in pathological tissues such as Alzheimer's disease and cancer.

Results

C-SIDE learns cell type-specific DE in spatial transcriptomics. Here, we develop C-SIDE, a statistical method for determining DE in spatial transcriptomics datasets (Fig. 1a). C-SIDE inputs one or more experimental samples of spatial transcriptomics data, consisting of $Y_{i,j,g}$ as the observed RNA counts for pixel i , gene j and experimental sample g . We then assume Poisson sampling so that

$$Y_{i,j,g} | \lambda_{i,j,g} \approx \text{Poisson}(N_{i,g} \lambda_{i,j,g}), \quad (1)$$

with $\lambda_{i,j,g}$ the expected count and $N_{i,g}$ the total transcript count (for example, total unique molecular identifiers, UMIs) for pixel i on experimental sample g . Accounting for platform effects and other sources of technical and natural variability, we assume $\lambda_{i,j,g}$ is a mixture of K cell type expression profiles, defined by,

$$\log(\lambda_{i,j,g}) = \log \left(\sum_{k=1}^K \beta_{i,k,g} \mu_{i,k,j,g} \right) + \gamma_{j,g} + \varepsilon_{i,j,g}, \quad (2)$$

with $\mu_{i,k,j,g}$ the cell type-specific expected gene expression rate for pixel i , gene j , experimental sample g and cell type k ; $\beta_{i,k,g}$ the proportion of cell type k contained in pixel i for experimental sample g ; $\gamma_{j,g}$ a gene-specific random effect that accounts for platform variability and $\varepsilon_{i,j,g}$ a random effect to account for gene-specific overdispersion.

To account for cell type-specific DE, we model across pixel locations the log of the cell type-specific profiles $\mu_{i,k,j,g}$ as a linear combination of L covariates used to explain DE. Specifically, we assume that,

$$\log(\mu_{i,k,j,g}) = \alpha_{0,k,j,g} + \sum_{\ell=1}^L x_{i,\ell,g} \alpha_{\ell,k,j,g}. \quad (3)$$

Here, $\alpha_{0,k,j,g}$ represents the intercept term for gene j and cell type k in sample g , and $x_{i,\ell,g}$ represents the ℓ th covariate, evaluated at pixel i in sample g . Similar to linear and generalized linear models²⁵, x , also called the design matrix, represents predefined covariate(s) that explain DE and the corresponding coefficient(s) $\alpha_{\ell,k,j,g}$ each represent the DE effect size of covariate ℓ for gene j in cell type k for sample g .

With this general framework we can describe any type of DE that can be parameterized with a log-linear model. Examples include (Fig. 1b):

- (1) DE between multiple regions. In this case, the tissue is manually segmented into multiple regions (for example, nodular and anterior cerebellum, Fig. 3). Design matrix x contains discrete categorical indicator variables representing membership in two or greater regions.
- (2) DE due to cellular environment or state (a special case of step (1)). Pixels are discretely classified into local environments based on the surrounding cells (for example, stages in the testes Slide-seq dataset, Fig. 4).
- (3) DE as a function of distance to a specific anatomical feature. In this case, x is defined as the spatial position or distance to some feature (for example, distance to midline in the hypothalamus MERFISH dataset, Fig. 4).
- (4) Cell-to-cell interactions. In this case, we define a cell-to-cell interaction as DE within one cell type (A) due to colocalization with a second cell type (B) (for example, immune cell density in cancer, Fig. 6). For this problem, x is the continuous density of cell type B.
- (5) Proximity to pathology. Similar to step (4), except covariate x represents density of a pathological feature (for example, Alzheimer's Aβ plaque, Fig. 4), rather than cell type density.
- (6) General spatial patterns (termed nonparametric). In this case, we define design matrix x to be smooth basis functions²⁶, where linear combinations of these basis functions represent the overall smooth gene expression function and can accommodate any smooth spatial pattern.

To estimate this complex model with a computationally tractable algorithm, in the first step, we assume $\mu_{i,k,j,g}$ does not vary with i and g and estimate β using a previously published algorithm¹⁸. This assumption does not substantially affect cell type proportion estimates because the gene expression variability across cell types is large relative to the variability across space, for most genes. Some pixels are identified as single cell types while others as mixtures of multiple cell types. Fixing the β estimates, we next use maximum likelihood estimation to estimate the cell type-specific DE coefficients α with corresponding standard errors, allowing for false discovery rate (FDR)-controlled hypotheses testing (Methods). Last, C-SIDE performs statistical inference across multiple replicates and/or samples to estimate consensus population-level DE (Methods and Supplementary Fig. 1).

Because ground truth cell type-specific DE is unknown in spatial transcriptomics data, we first benchmarked C-SIDE's performance on a simulated spatial transcriptomics dataset in which gene expression varied across two regions. Considering the challenging situation where two cell types, termed cell type A and cell type B, are colocalized on pixels within a tissue, we simulated, using a single-nucleus RNA-seq cerebellum dataset, spatial transcriptomics mixture pixels with known proportions of single cells from two cell types known to spatially colocalize²⁷ (Methods and Fig. 2a). Across two spatially defined regions, we varied both the true cell type-specific gene expression of cell types A and B as well as the average cell type proportions of cell types A and B (Fig. 2a and Supplementary Fig. 2). We compared C-SIDE against three alternative methods (Methods): bulk, bulk DE (ignoring cell type); single, single-cell DE that approximates each cell type mixture as a single cell type and decompose¹⁸, a method that decomposes mixtures into single cell types before computing DE. By varying cell type frequencies between the two regions without introducing DE, we observed that C-SIDE correctly attributes gene expression differences across regions to differences in cell type proportions rather than spatial DE (Fig. 2b and Supplementary Fig. 2); in contrast, the bulk method incorrectly predicts spatial DE since it does not control for differences of cell type proportions across regions.

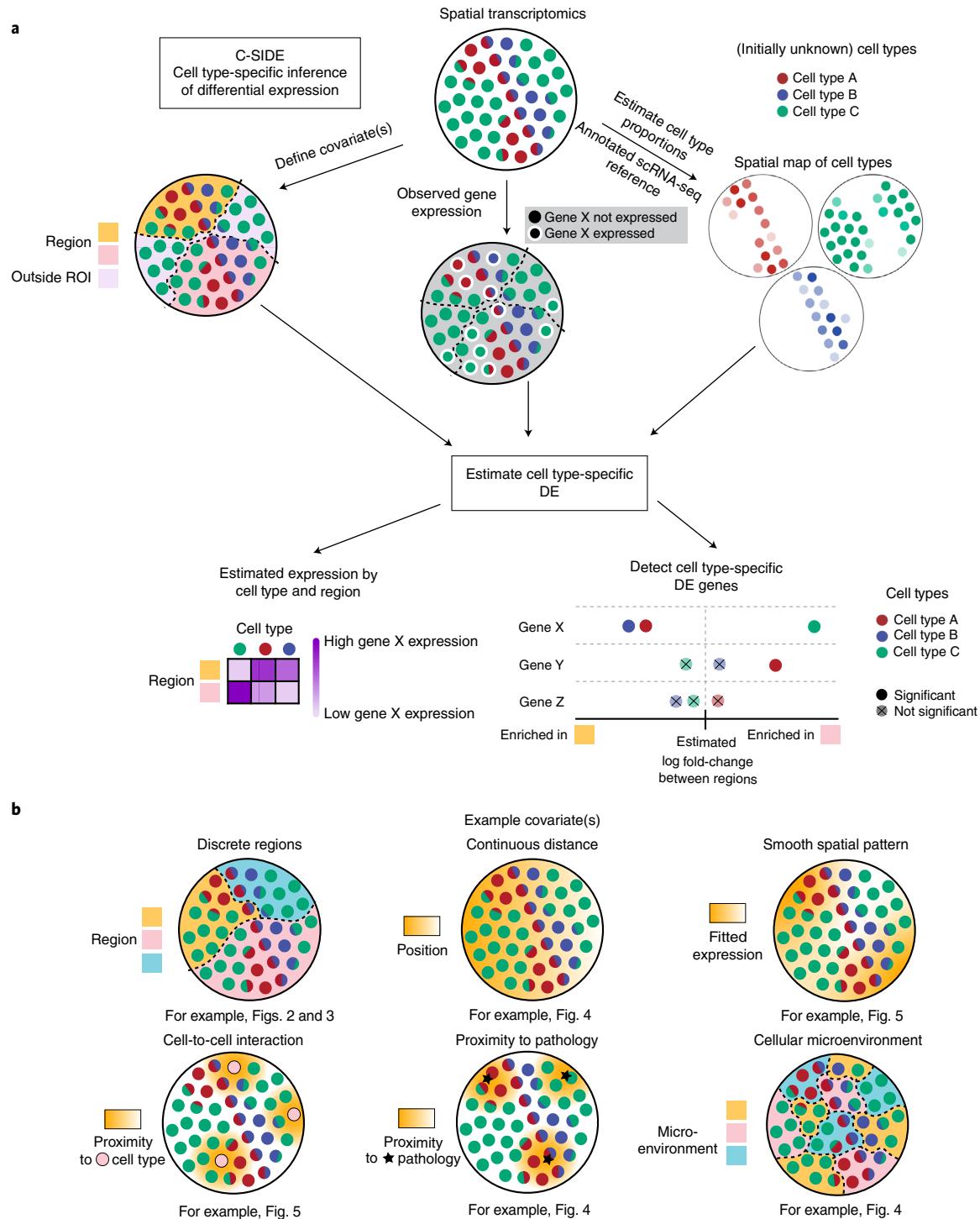


Fig. 1 | C-SIDE learns cell type-specific DE from spatial transcriptomics data. **a**, Schematic of the C-SIDE method. Top, C-SIDE inputs: a spatial transcriptomics dataset with observed gene expression (potentially containing cell type mixtures) and a covariate for DE. Middle, C-SIDE first assigns cell types to the spatial transcriptomics dataset, and covariates are defined. Bottom, C-SIDE estimates cell type-specific gene expression along the covariate axes. **b**, Example covariates for explaining DE with C-SIDE. Top, segmentation into multiple regions, continuous distance from some feature or general smooth spatial patterns (nonparametric). Bottom, density of interaction with another cell type or pathological feature or a discrete covariate representing the cellular microenvironment.

Next, we simulated cell type-specific DE by varying the DE in cell type A while keeping cell type B constant across regions. Background DE in cell type A contaminated estimates of DE in cell type B for all three alternative models: bulk, decompose and single (Fig. 2c and Supplementary Fig. 2). In particular, decompose

assigns gene expression to cell types for each pixel independently but does not have information to distinguish which cell type is responsible for DE in mixture pixels. In contrast, C-SIDE's joint model of cell type mixtures and cell type-specific DE correctly identified DE in cell type A, but not cell type B. Next, we verified that,

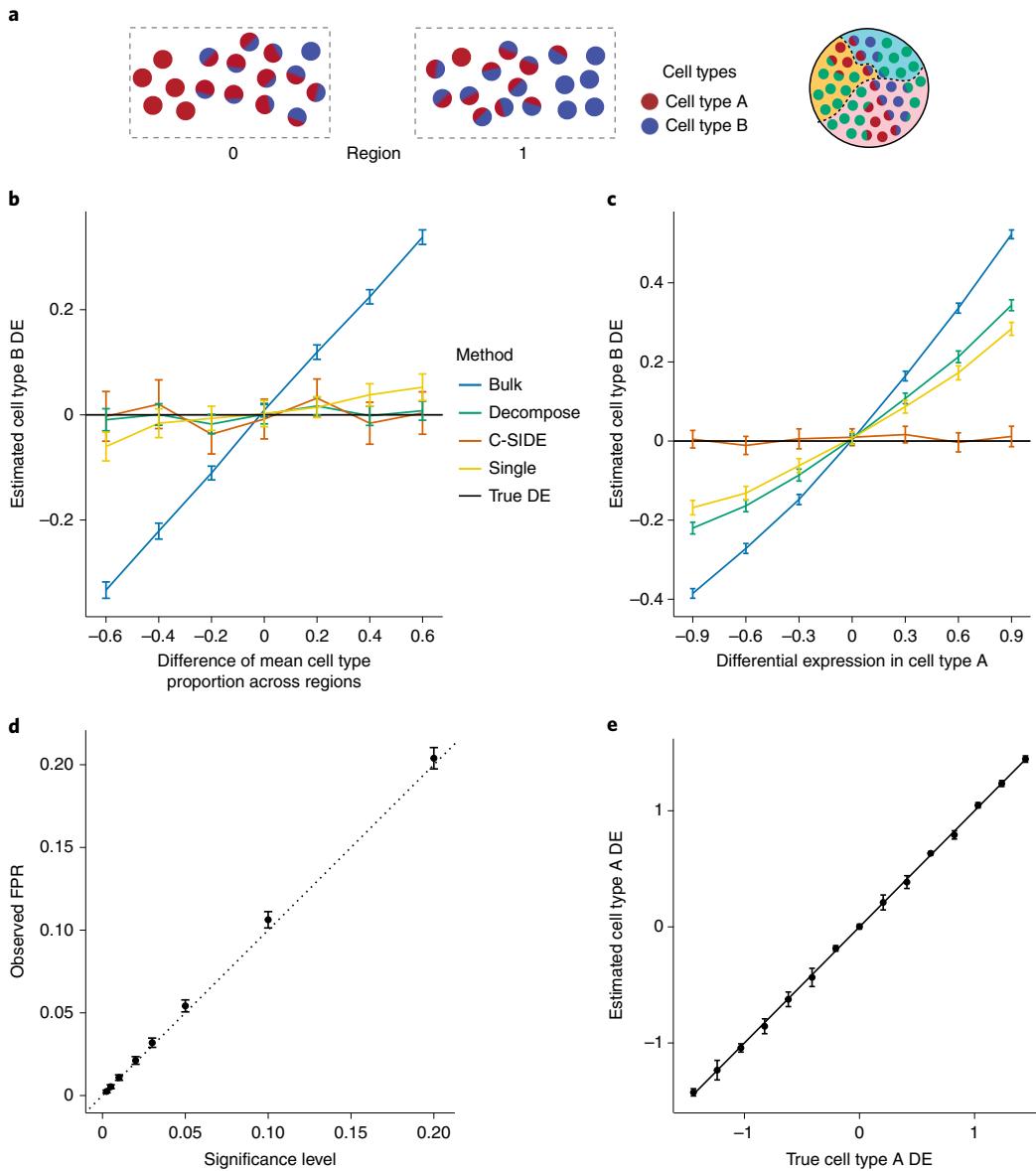


Fig. 2 | C-SIDE provides unbiased estimates of cell type-specific DE in simulated data. All, C-SIDE was tested on a dataset of simulated mixtures of single cells from a single-nucleus RNA-seq cerebellum dataset. DE axes represent DE in log₂ space of region 1 with respect to region 0. **a**, Pixels are grouped into two regions, and genes are simulated with ground truth DE across regions. Each region contains pixels containing mixtures of various proportions between cell type A and cell type B. The difference in average cell type proportion across regions is varied across simulation conditions. **b**, Mean estimated cell type B *Astn2* DE across two regions as a function of the difference in mean cell type proportion across regions. *Astn2* is simulated with ground truth 0 spatial DE, and an average of ($n=100$) estimates is shown, along with standard errors. Black line represents ground truth 0 DE (cell type B). Four methods are shown: bulk, decompose, single and C-SIDE (Methods). **c**, Same as **b** for *Nrxn3* cell type B differential gene expression as a function of DE in cell type A, where *Nrxn3* is simulated to have DE within cell type A but no DE in cell type B. **d**, For each significance level, C-SIDE's false-positive rate (FPR), along with ground truth identity line (s.e. shown, $n=1,500$, 15 genes, 100 replicates per gene). **e**, C-SIDE mean estimated cell type A DE versus true cell type A DE (average over $n=500$ replicates, s.e. shown). Ground truth identity line is shown, and one gene is used for the simulation per DE condition (out of 15 total genes).

under the null hypothesis of zero DE, C-SIDE's false-positive rate was accurately controlled, standard errors were accurately estimated and confidence intervals contained the ground truth DE (Fig. 2d and Supplementary Fig. 2). Finally, when nonzero DE was simulated, C-SIDE achieved unbiased estimation of cell type-specific DE (Fig. 2e). We also found that the power, false-positive rate and true-positive rate of C-SIDE depends on gene expression level, number of cells and DE magnitude (Supplementary Fig. 2). Last, on Slide-seq, MERFISH and Visium spatial transcriptomics data, we verified that C-SIDE's fitted Poisson-lognormal distribution

accurately fits the empirical spatial transcriptomics gene expression distribution (Supplementary Figs. 3–5). Thus, our simulations validate C-SIDE's ability to accurately estimate and test for cell type-specific DE in the cases of asymmetric cell type proportions and contamination from other cell types.

To validate C-SIDE's ability to discover cell type-specific DE on spatial transcriptomics data, we collected Slide-seq V2 data² (including one replicate sourced from a previous study¹⁸) for three cerebellum replicates. We identified a spatial map of cell types (Fig. 3a), previously shown to correspond to known cerebellum

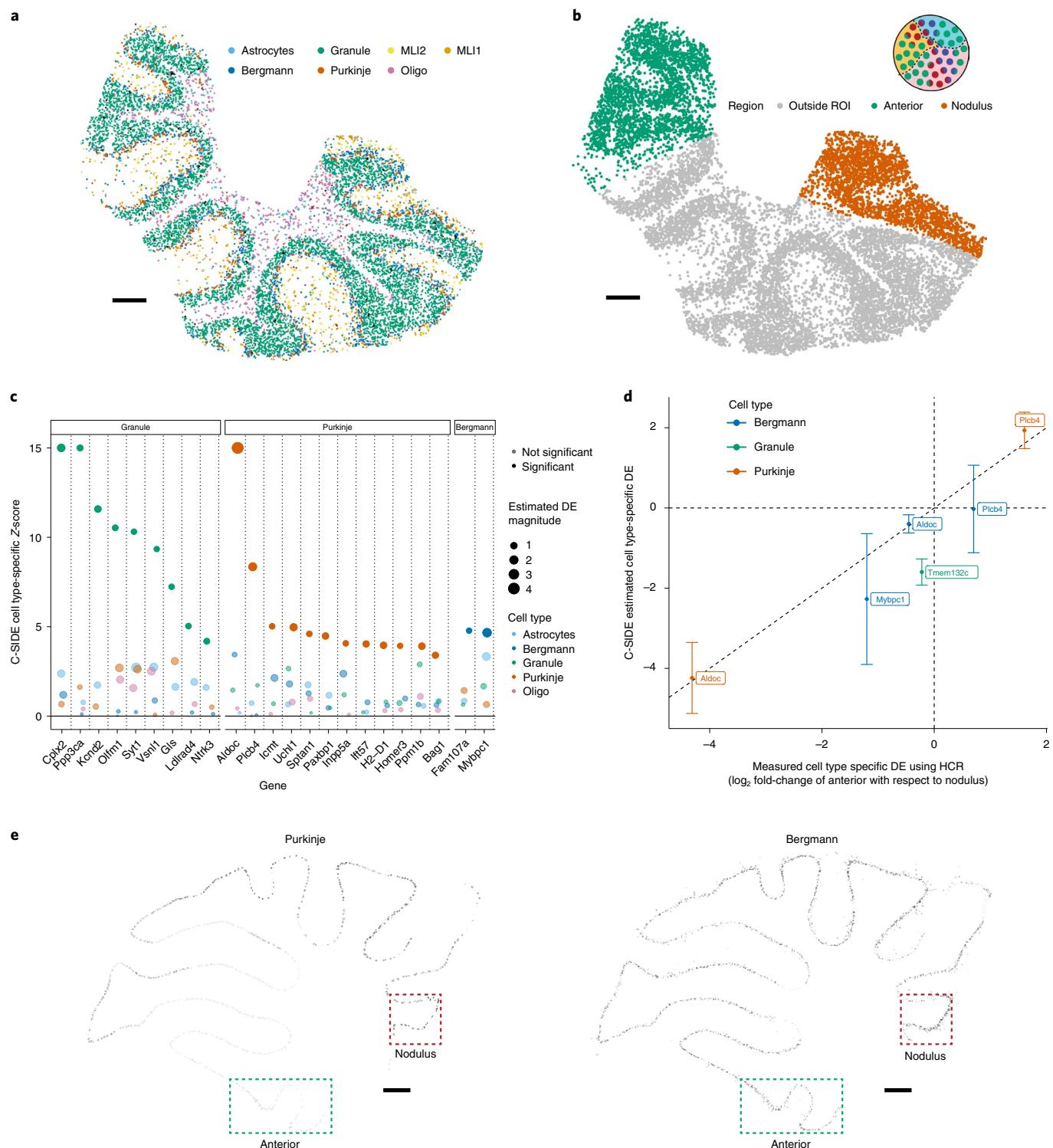


Fig. 3 | C-SIDE's estimated cell type-specific DE is validated by HCR-FISH. C-SIDE ran on ($n=3$) replicates of cerebellum Slide-seq data. **a**, C-SIDE's spatial map of cell type assignments. Out of 19 cell types, the seven most common appear in the legend. Reproduced from ref. ¹⁸. Three total replicates were used to fit C-SIDE. **b**, Covariate used for C-SIDE, representing the anterior lobule region (green) and nodulus (red). Schematic refers to the C-SIDE problem type outlined in Fig. 1b. **c**, C-SIDE Z-score for testing for DE for each gene and for each cell type. Genes are grouped by cell type with maximum estimated DE, and estimated DE magnitude appears as size of the points. Bold genes appear below in HCR validation. **d**, Scatterplot of C-SIDE DE estimates versus HCR measurements for cell type-specific \log_2 DE. Positive values indicate gene expression enrichment in the anterior region. Error bars represent C-SIDE confidence intervals for predicted DE on a new biological replicate. A dotted identity line is shown, and cell types are colored. **e**, HCR images of *Aldoc* continuous gene expression. Only pixels with high cell type marker measurements for Purkinje (left) and Bergmann (right) are shown. ROI of nodulus and anterior regions are outlined in green and red, respectively. All scale bars, 250 μ m.

spatial architecture¹⁸. We used discrete localization in the anterior lobule or nodulus regions (Fig. 3b), a known axis of spatial gene expression variation within the cerebellum²⁷, as a covariate for estimating cell type-specific DE across regions using C-SIDE (Fig. 3c, Supplementary Fig. 6 and Supplementary Table 1). As experimental validation, we performed hybridization chain reaction (HCR) on four genes identified by C-SIDE to be DE in specific cell types, and we observed high correspondence between C-SIDE's estimates of cell type-specific DE and DE measurements from HCR data (Fig. 3d, $R^2=0.89$). For example, we examined *Aldoc* and *Plcb4*, two genes expressed in both Purkinje and Bergmann cell types, which are known to spatially colocalize in the cerebellum and appear as mixtures on Slide-seq pixels¹⁸. C-SIDE determined that both *Aldoc* (\log_2 fold-change -4.24 , $P < 10^{-8}$) and *Plcb4* (\log_2 fold-change 1.93 , $P < 10^{-8}$) were DE in the Purkinje cell type, but not the Bergmann cell type. Similarly, HCR images of *Aldoc* and *Plcb4* showed substantial DE within Purkinje cells across the nodulus and anterior lobule, whereas expression within Bergmann cells was relatively even across regions (Fig. 3d,e). Next, we used C-SIDE to obtain cell subtype-specific DE estimates. Except for one gene, cell type-specific spatial DE did not differ significantly between a cell type and its subtypes. Furthermore, due to reduced sample size, C-SIDE had reduced statistical power to detect subtype-specific DE (Supplementary Fig. 7). We conclude that C-SIDE can successfully identify cell type-specific spatial DE in spatial transcriptomics tissues, even when multiple cell types are spatially colocalized.

C-SIDE solves diverse DE problems in spatial transcriptomics. We next explored the effect of discrete cellular microenvironments on cell type-specific DE in the mouse testes Slide-seq dataset⁹. C-SIDE's testes principal cell type assignments (Fig. 4a) revealed tubular structures corresponding to cross-sectional sampling of seminiferous tubules. Individual tubules have distinct stages of spermatogonia development, grouped into four classes of stages I–III, IV–VI, VII–VIII and IX–XII, which were determined from the previous testes Slide-seq study⁹ (Fig. 4b). We applied C-SIDE to identify genes that were DE, for each cell type, across tubule stages (Supplementary Table 2). C-SIDE identified genes expressed in a single tubule stage within a single cell type (Fig. 4c), which are known drivers of cellular development across stages⁹. For instance, the gene *Tnp1* was identified by C-SIDE as upregulated in the IX–XII stage within the elongating spermatid cell type, in agreement with the known biological role of *Tnp1* in nuclear remodeling of elongating spermatids at the late tubule stage²⁸ (Supplementary Fig. 8). Furthermore, most C-SIDE-identified stage-specific genes followed cyclic patterns, consistent with previously characterized seminiferous epithelial cycle²⁹ (Supplementary Fig. 8).

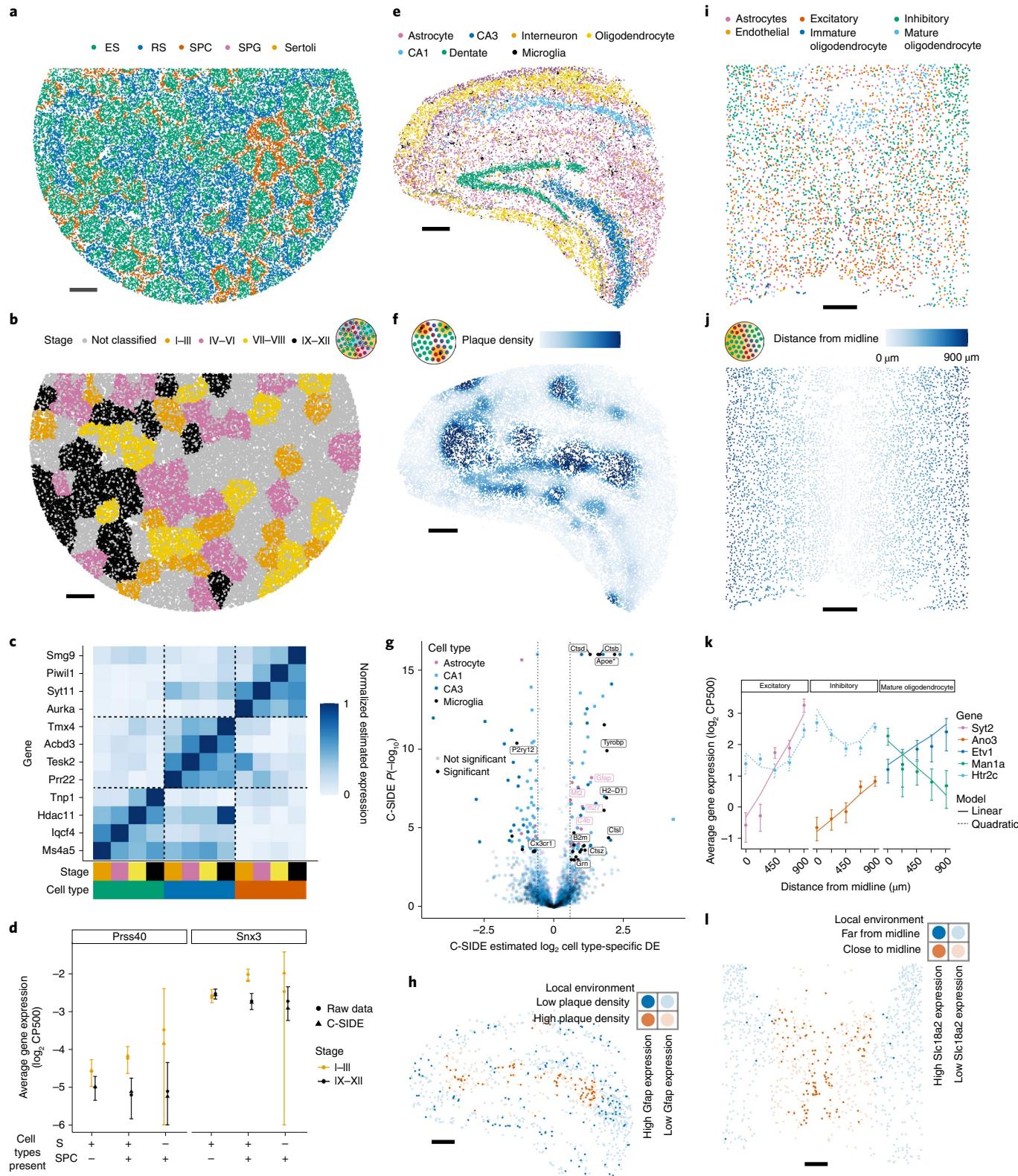
Next, we evaluated C-SIDE's ability to identify DE for cell types that primarily appear as mixtures with other cell types, particularly the spermatocyte (SPC) cell type. According to C-SIDE cell type assignments, SPC frequently comixes with the ES and round spermatid cell types, consistent with previous histological studies³⁰ (Supplementary Fig. 8a). By using cell type mixtures, C-SIDE obtained increased power for identifying DE genes compared to a method only using single-cell type pixels (Supplementary Notes and Supplementary Fig. 8b,c), especially for SPC cell type (217 significant SPC DE genes discovered by C-SIDE versus 1 DE gene for the single-cell method). For most SPC DE genes, including *Prss40* (\log_2 fold-change 1.72 , $P = 8 \times 10^{-5}$) and *Snx3* (\log_2 fold-change 1.17 , $P < 10^{-8}$), pixels containing SPC but not spermatids were too rare to determine DE (Fig. 4d). Instead, C-SIDE determined DE specifically in SPC cells by detecting significant spatial differences among pixels containing both SPC and spermatid cell types, but not within pure spermatid pixels. Therefore, C-SIDE's cell type mixture modeling uniquely enables DE discovery in highly mixed cell types.

A β plaque-dependent DE in Alzheimer's disease. We next explored pathological staining, in particular A β plaques, as a continuous covariate for cell type-specific gene expression changes. We performed Slide-seq V2 on the hippocampal region of a genetic mouse model of amyloidosis in Alzheimer's disease³¹ (J20, $n = 4$ slices, Methods). C-SIDE's cell type assignments (Fig. 4e) were consistent with past characterizations of hippocampus cellular localization¹⁸. We collected paired A β plaque staining images (Anti-Human A β Mouse IgG antibody, Methods) to quantify A β plaque density as a covariate for C-SIDE (Fig. 4f and Supplementary Fig. 10). We then used C-SIDE to identify genes whose expression depended in a cell type-specific manner on plaque density (Fig. 4g and Supplementary Table 6). For instance, in astrocytes colocalizing with A β plaque, C-SIDE detected upregulation of *Gfap* (Fig. 4h and Supplementary Fig. 10, \log_2 fold-change 1.35 , $P < 10^{-8}$), consistent with *Gfap*'s known role in A β plaque attenuation³² and the *C4b* complement gene (\log_2 fold-change $= .85$, $P = 1 \times 10^{-4}$), which is involved in plaque-associated synaptic pruning in Alzheimer's disease^{33–35}. Moreover, several cathepsin proteases including *Ctsb* (\log_2 fold-change 1.65 , $P < 10^{-8}$), *Ctsd* (\log_2 fold-change 1.30 , $P < 10^{-8}$) *Ctsl* (\log_2 fold-change 1.96 , $P = 4 \times 10^{-6}$) and *Ctsz* (\log_2 fold-change 1.11 , $P = 3 \times 10^{-4}$) were determined to be differentially upregulated in microglia around plaque, consistent with the role of cathepsins in amyloid degradation³⁶ (Supplementary Fig. 10). In microglia, we also identified known homeostatic microglia markers^{37,38} including *P2ry12* (\log_2 fold-change -1.33 , $P < 10^{-8}$) and *Cx3cr1* (\log_2 fold-change -0.68 ,

Fig. 4 | C-SIDE discovers cell type-specific DE in a diverse set of problems on testes, Alzheimer's hippocampus and hypothalamus datasets. All panels show results of C-SIDE on the Slide-seqV2 testes (left column), MERFISH hypothalamus (middle column) and Slide-seqV2 Alzheimer's hippocampus (right column). Schematics in **b,f,j** reference C-SIDE problem types (Fig. 1b). **a**, C-SIDE's spatial map of cell type assignments in testes. All cell types are shown with most common in legend. **b**, Covariate used for C-SIDE in testes: four discrete tubule stages. **c**, Cell type and tubule stage-specific genes identified by C-SIDE. C-SIDE estimated expression is standardized between 0 and 1 for each gene. Columns represent C-SIDE estimates for each cell type and tubule stage. **d**, The \log_2 average expression (in counts per 500 (CP500)) of pixels grouped based on tubule stage and presence or absence of spermatid (S) cell types (defined as elongating spermatid (ES) or round spermatid (RS)) and/or SPC cell type. Circles represent raw data averages while triangles represent C-SIDE predictions, and error bars around circular points represent ± 1.96 s.d. ($37 \leq n \leq 2,236$ pixels per group, Supplementary Notes). Genes *Prss40* and *Snx3* are shown on left and right, respectively. **e**, Same as **a** for Alzheimer's hippocampus ($n = 4$ replicates). **f**, Covariate used for C-SIDE in Alzheimer's hippocampus: continuous density of A β plaque. **g**, Volcano plot of C-SIDE DE results in \log_2 space, with positive values corresponding to plaque-upregulated genes. Color represents cell type, and a subset of significant genes is labeled. Dotted lines represent 1.5x fold-change cutoff used for C-SIDE. **h**, Spatial visualization of *Gfap*, identified by C-SIDE as DE in astrocytes. Red/blue represents high/low plaque density areas, respectively. Bold points represent astrocytes expressing *Gfap* at least 1 CP500. **i**, Same as **a** for hypothalamus. **j**, Covariate used for C-SIDE in hypothalamus, midline distance. **k**, The \log_2 average expression (CP500) of C-SIDE significant DE genes for excitatory, inhibitory and mature oligodendrocyte cell types. Single cell type pixels are binned by midline distance, and points represent raw data averages while lines represent C-SIDE predictions and error bars around points represent ± 1.96 s.d. ($34 \leq n \leq 411$ pixels per group) (Supplementary Notes). **l**, Spatial visualization of *Slc18a2*, identified by C-SIDE as DE in inhibitory neurons. Red/blue represents close/far to midline, respectively. Bold points are inhibitory neurons expressing *Slc18a2* at least ten CP500. All scale bars, 250 μ m.

$P=3 \times 10^{-4}$) as downregulated in the presence of plaque. *Apoe*, known to have $A\beta$ plaque-dependent upregulation within microglia³⁹, was also detected (\log_2 fold-change 1.58, $P < 10^{-8}$). Finally, the anti-inflammatory gene *Grn* was upregulated in microglia near plaque (\log_2 fold-change 0.79, $P = 6 \times 10^{-4}$), consistent with previous knowledge⁴⁰.

Spatial gene expression changes in imaging-based transcriptomics. We next applied C-SIDE across different spatial technology length scales, from near single-cell resolution imaging-based spatial transcriptomics (for example, MERFISH) to lower-resolution technologies such as Visium. First, we applied C-SIDE to a MERFISH mouse hypothalamus dataset. During development, hypothalamic



progenitors create radial projections out from the hypothalamic midline, which are used as scaffolds for the migration of differentiating daughter cells⁴¹. Thus, we investigated radial distance to the hypothalamus midline as a predictor of DE in hypothalamus cell types. C-SIDE's assigned cell types were consistent with the previous MERFISH hypothalamus study⁸ (Fig. 4i). Although most pixels were single cell types, a nonnegligible proportion (12.6% double cell type pixels out of $n=3,790$ total pixels) were mixtures of multiple cell types. Using midline distance as a covariate for C-SIDE (Fig. 4j), we detected genes in hypothalamus excitatory, inhibitory and mature oligodendrocyte cell types whose expression depended either linearly or quadratically on distance from the midline (Fig. 4k and Supplementary Tables 3 and 4). For instance, *Slc18a2* (Fig. 4l), identified as upregulated within inhibitory neurons near the midline (\log_2 fold-change 6.14, $P < 10^{-8}$), is required for dopaminergic function in certain inhibitory neuronal subtypes⁴², which are known to localize near the hypothalamus midline⁸. Next, we used C-SIDE to address the known challenging problem of cellular segmentation in imaging transcriptomics²⁰. Since C-SIDE can operate on both cell type mixtures and single cells, we hypothesized that cell segmentation could be skipped altogether and replaced by defining pixels as a fixed grid (Fig. 5a). Indeed, skipping segmentation did not substantially change C-SIDE DE estimates (Fig. 5b) and reduced C-SIDE uncertainty (Supplementary Fig. 9) due to an inclusion of more counts that would be discarded during segmentation.

DE discovery in lower-resolution spatial transcriptomics. We next tested whether C-SIDE on a Visium human lymph node dataset, with spot size 55 μm (ref. ⁴³), resulting in a higher degree of cell type mixtures. By using B cell proportion as the C-SIDE covariate, we tested for gene expression changes within the B cell-rich germinal centers, essential regions for B cell maturation. We note consistency between B cell-rich regions and germinal center morphology from histology (Fig. 5c). C-SIDE identified germinal center-driven DE (Fig. 5d, Supplementary Fig. 9 and Supplementary Table 5), such as correctly identifying germinal center B cell markers, including *RGS13* (\log_2 fold-change 1.26, $P = 2 \times 10^{-5}$) and *STMN1* (\log_2 fold-change 1.07, $P < 10^{-8}$), as upregulated in germinal center B cells⁴⁴. Moreover, C-SIDE detected germinal center-localized follicular dendritic cell (DC) markers *CR2* (\log_2 fold-change 2.40, $P < 10^{-8}$) and *FDCSP* (\log_2 fold-change 1.30, $P < 10^{-8}$) as upregulated in DCs within germinal centers⁴⁵. Despite 3–4 cell types mixing per spot, C-SIDE accurately determined which cell type(s) was responsible for differential gene expression. For instance, the chemokine *CXCL13* was upregulated in B cell-rich regions (Fig. 5e). Despite dendritic cells comprising no more than 3–15% of each spot (Supplementary Fig. 9), C-SIDE was attributed the spatial pattern of *CXCL13* to DCs (\log_2 fold-change 1.84, $P < 10^{-8}$), consistent with the role of DCs in secreting *CXCL13* to attract B cells to germinal centers⁴⁶. C-SIDE assigned *CXCL13* DE to dendritic cells by using that DE between the B cell-rich region and the B cell-poor region increased as a function of DC proportion (Fig. 5f). Thus, C-SIDE can determine cell type-specific DE, even for rare cell types that are consistently mixed with other cells.

Fig. 5 | C-SIDE enables DE discovery on diverse spatial transcriptomics technologies including Visium and MERFISH. All panels show results of C-SIDE on the Visium lymph node (middle and bottom rows) and MERFISH hypothalamus (top row). **a**, C-SIDE's spatial map of cell type assignments in the hypothalamus, where pixels were defined deterministically as squares without segmentation. All cell types are shown, and the most common cell types appear in the legend. **b**, Scatter plot of C-SIDE estimated inhibitory cell type DE with and without cell segmentation. **c**, Covariate used for C-SIDE: discrete region of B cell-rich areas in the lymph node. Overlayed with Visium histology image. **d**, Volcano plot of C-SIDE dendritic cell DE results in \log_2 space, with positive values corresponding to upregulated genes in the B cell regions. A subset of significant genes is labeled (two-sided Z-test with FDR control, Methods). Dotted lines represent 1.5 \times fold-change cutoff used for C-SIDE. **e**, Spatial plot of total expression of the *CXCL13* gene, which was determined by C-SIDE to be DE in dendritic cells. Color represents counts per spot. **f**, Average expression (in counts per 500 (CP500)) of *CXCL13* as a function of dendritic cell proportion and germinal center localization. Points represent raw data averages while lines represent C-SIDE predictions and error bars around points represent ± 1.96 s.d. ($75 \leq n \leq 326$ points per group). All scale bars, 250 μm .

C-SIDE discovers tumor-immune signaling in a mouse tumor model. Finally, we applied C-SIDE to identify genes with cell type-specific spatial DE in a Slide-seq dataset of a *Kras*^{G12D/+}/*Trp53*^{-/-} (KP) mouse tumor model⁴⁷, where we analyzed a metastatic lung adenocarcinoma tumor deposit in the liver. First, C-SIDE found several cell types within the tumor, including both tumor cells and myeloid cells (Fig. 6a). Next, we ran C-SIDE nonparametrically to discover arbitrary smooth gene expression patterns (Supplementary Notes and Supplementary Table 7). We found three categories of genes within tumor cells: genes with variation due to sampling noise, spatial variation (explained by the C-SIDE model) or nonspatial biological variation (Fig. 6b and Supplementary Fig. 11). We then hierarchically clustered the C-SIDE estimated spatial patterns of DE genes into seven clusters (Fig. 6c and Supplementary Fig. 11). Testing for gene set enrichment (Supplementary Notes), we identified the *Myc* targets gene set as enriched in cluster 5 (7 out of 12 genes, $P = 2 \times 10^{-4}$, two-sided binomial test, Supplementary Table 8), a cluster most highly expressed at the tumor boundary (Fig. 6d). High expression of *Myc* target genes potentially indicates an increased rate of proliferation⁴⁸ at the boundary, a proposed correlate of tumor severity⁴⁹. For example, the most DE *Myc* target gene, *Kpnb1* (Supplementary Fig. 11, $P = 1 \times 10^{-5}$), is an oncogene that drives cell proliferation and suppresses apoptosis⁵⁰.

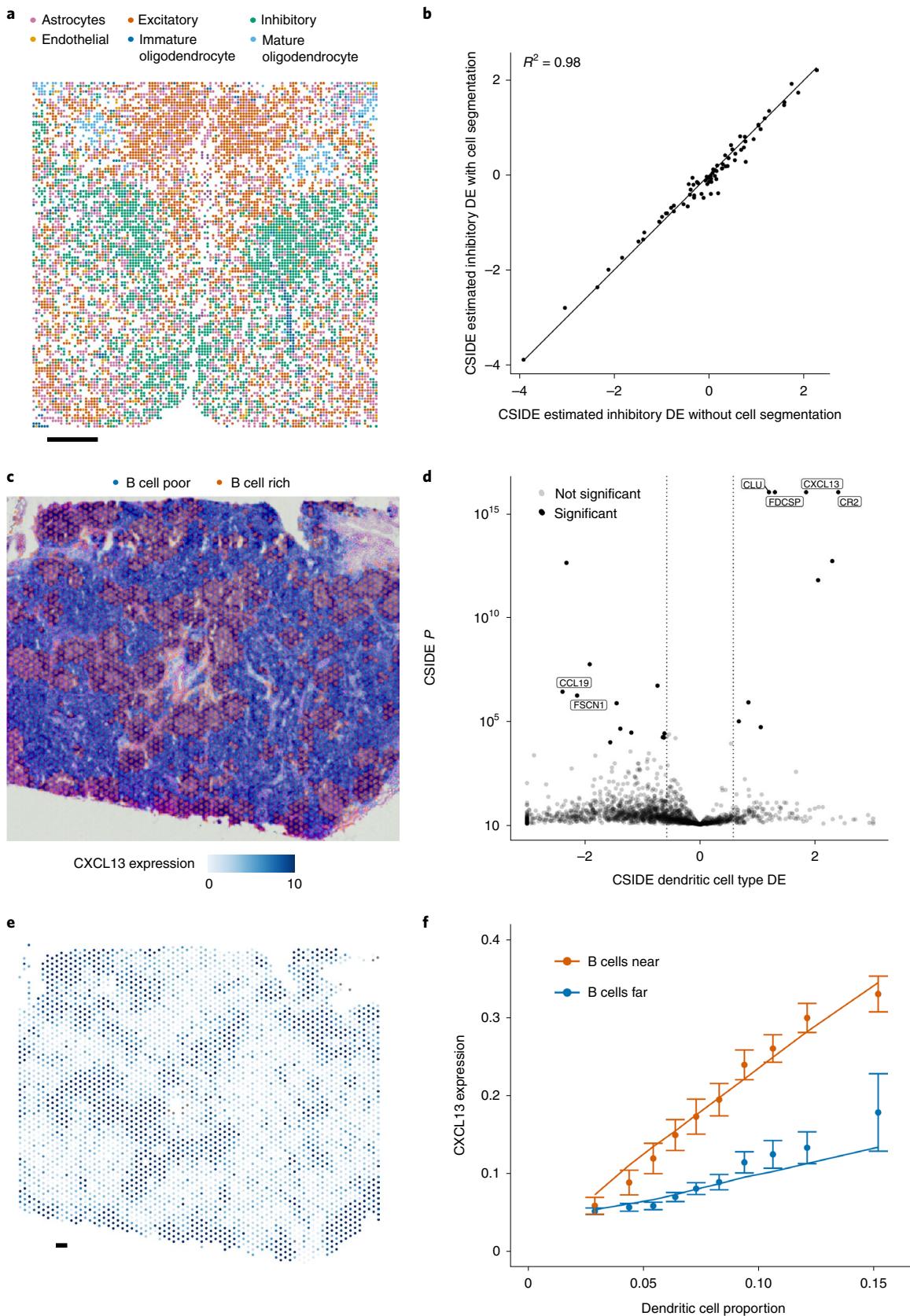
We next used C-SIDE to detect cell-to-cell interactions between tumor cells and immune cells, which are known to influence tumor cell behavior⁵¹. Using myeloid cell type density as the C-SIDE covariate (Fig. 6e), C-SIDE identified genes with immune cell density-dependent DE (Fig. 6f and Supplementary Table 9), including several genes that were also discovered by nonparametric C-SIDE (Supplementary Fig. 11). One of the genes with the largest effects, *Ccl2* (\log_2 fold-change 1.74, $P < 10^{-8}$), is a chemotactic signaling molecule known to attract myeloid cells⁵². Furthermore, we tested C-SIDE's DE gene estimates for aggregate effects across gene sets and found that the epithelial-mesenchymal transition (EMT) pathway was significantly upregulated near immune cells (Fig. 6f and Supplementary Fig. 11, $P = 0.0011$, permutation test (Methods) and Supplementary Table 8). C-SIDE additionally identified EMT-regulator *Nfkbia* as positively DE in tumor cells in immune-rich regions (\log_2 fold-change 1.10, $P = 1 \times 10^{-5}$)⁵³. As validation, we find most tumor cells expressing EMT genes localized to immune-rich regions (Fig. 6g). Furthermore, a hematoxylin and eosin (H&E) tumor stain (Fig. 6h) demonstrated a EMT morphological change in the immune-rich region (spindle-shaped tumor cells) relative to the immune-poor region (polygonal-shaped tumor cells). Thus, morphological and gene expression changes indicate that the immune microenvironment influences EMT in this tumor model⁵⁴.

Discussion

Explaining spatial sources of differential gene expression is a critical challenge for understanding biological mechanisms and disease with spatial transcriptomics. Here we introduced C-SIDE, a statistical method to detect cell type-specific DE in spatial transcriptomics datasets. C-SIDE takes as input one or more biologically relevant

covariates, such as spatial position or cell type colocalization, and identifies genes, for each cell type, that significantly change their expression as a function of these covariates. Tested on simulated spatial transcriptomics data, C-SIDE obtained unbiased estimation

of cell type-specific differential gene expression with a calibrated false-positive rate, while other methods were biased from changes in cell type proportion or contamination from other cell types. In the cerebellum, we additionally used HCR experiments to validate



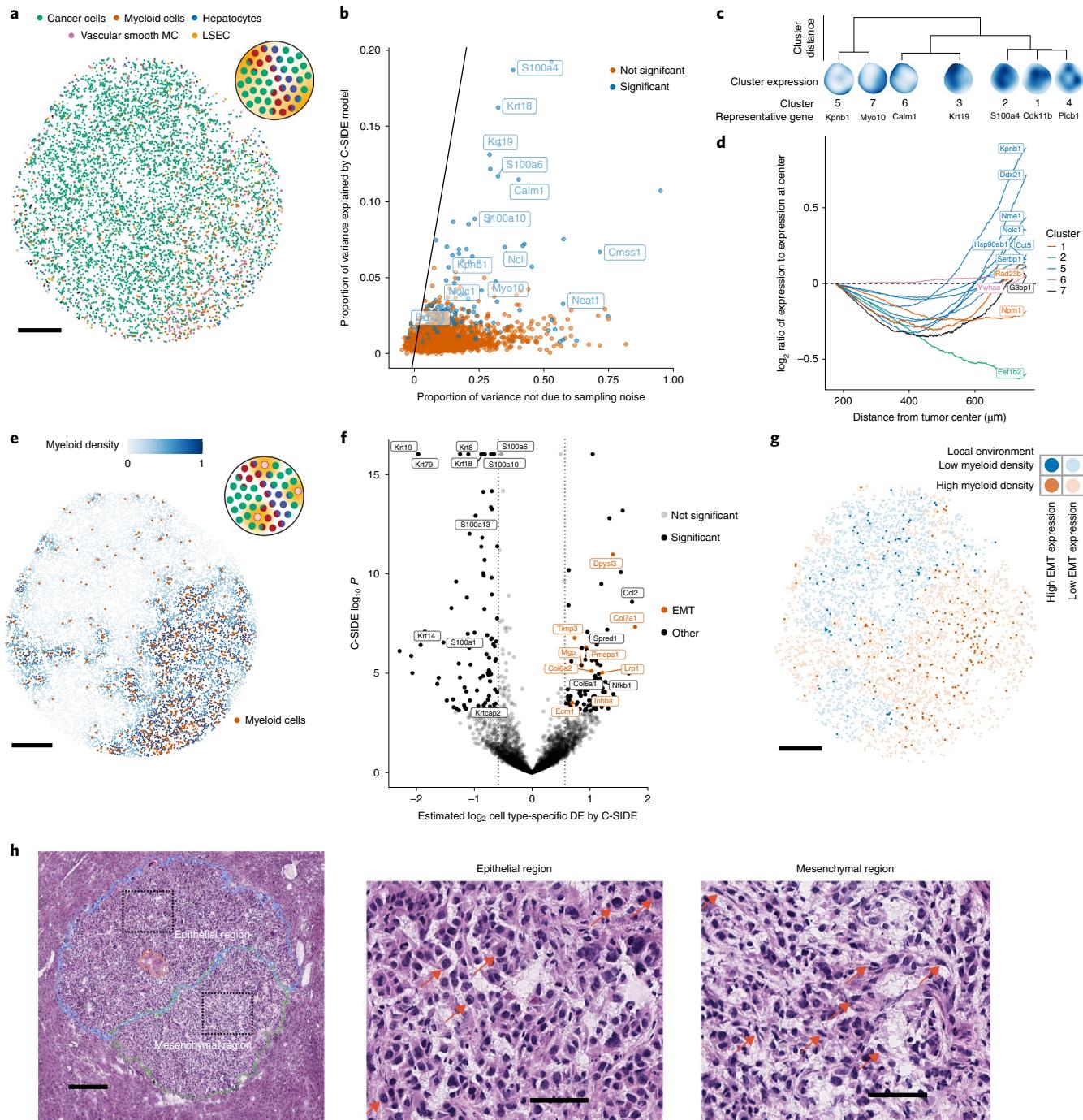


Fig. 6 | C-SIDE enables the discovery of DE pathways in a *Kras*^{G12D/+}*Trp53*^{-/-} (KP) mouse model. All panels: C-SIDE ran on multiple cell types; plots show C-SIDE results on the tumor cell type. Nonparametric/parametric C-SIDE results are shown in **b-d** and **e-h**, respectively. **a**, C-SIDE's spatial map of cell type assignments. Out of 14 cell types, the five most common appear in the legend. LSEC (liver sinusoidal endothelial cells). **b**, Scatter plot of C-SIDE R^2 and overdispersion (defined as proportion of variance not due to sampling noise) for nonparametric C-SIDE results on the tumor cell type. Identity line is shown, representing the maximum possible variance explained. **c**, Dendrogram of hierarchical clustering of ($n=162$ significant genes) C-SIDE's fitted smooth spatial patterns at the resolution of seven clusters. Each spatial plot represents the average fitted gene expression patterns over the genes in each cluster. **d**, Moving average plot of C-SIDE fitted gene expression (normalized to expression at center) as a function of distance from the center of the tumor for 12 genes in the Myc targets pathway identified to be significantly spatially DE by C-SIDE. **e**, Covariate used for parametric C-SIDE: continuous density of myeloid cell types in the tumor. Schematic refers to C-SIDE problem type (Fig. 1b). **f**, Volcano plot of C-SIDE \log_2 DE results ($n=4,201$ pixels) on the tumor cell type with positive values representing upregulation near myeloid immune cells. A subset of significant genes is labeled, and dotted lines represent 1.5 \times fold-change cutoff. **g**, Spatial plot of total expression in tumor cells of the nine DE EMT genes identified by C-SIDE in **f**. Red/blue represents myeloid-dense and myeloid-poor areas, respectively. Bold points represent tumor cells expressing these EMT genes at least 2.5 counts per 500. **h**, H&E image of adjacent section of the tumor ($n=1$ section). Left, mesenchymal (green), necrosis (red) and epithelial (blue) annotated tumor regions, with dotted boxes representing epithelial and mesenchymal areas of focus for the other two panels. Middle/right, enlarged images of epithelial (middle) or mesenchymal (right) regions. Red arrows point to example tumor cells with epithelial (middle) or mesenchymal (right) morphology. 50 μm scale bars (**h**) middle/right. All other scale bars, 250 μm .

C-SIDE's ability to identify cell type-specific DE across regions. We further applied C-SIDE to a detect DE depending on tubular micro-environment in the testes, midline distance in the MERFISH hypothalamus, germinal center localization in Visium lymph node and A β plaque density in the Alzheimer's model hippocampus. Finally, we applied both nonparametric and parametric C-SIDE procedures in a mouse tumor model to discover an increase in tumor cells undergoing EMT transition in immune-rich regions.

Several studies have established the importance of accounting for cell type mixtures in assigning cell types in spatial transcriptomics data^{18,21–23}. However, it remains a challenge to incorporate cell type proportions into models of cell type-specific spatial differential gene expression. C-SIDE enables such cell type-specific DE discovery by creating a statistical model of gene expression in the presence of cell type mixtures. Other potential solutions, such as bulk DE, approximation as single cell types, and decomposition into single cell types can be confounded by cell type proportion changes and contamination from other cell types. C-SIDE solves these issues by controlling for cell type proportions and jointly considering DE within each cell type. Even in imaging-based spatial transcriptomics methods such as MERFISH, we detected some pixels with cell type mixtures, indicating potential diffusion or imperfect cell segmentation²⁰. To control for cell type proportions in DE analysis, C-SIDE can estimate cell types directly or import cell type proportions from any cell type mixture identification method^{18,21–23}.

C-SIDE's parametric mode provides a unified framework for detecting biologically relevant DE in spatial transcriptomics tissues along diverse axes including spatial distance, proximity to pathology, cellular microenvironment and cell-to-cell interactions. In settings without previous biological hypotheses, C-SIDE may be run nonparametrically to discover general cell type-specific spatial gene expression patterns. C-SIDE can also be used to test among multiple models of DE, such as the linear and quadratic models applied to the hypothalamus dataset. To help decide between multiple relationships between covariates and gene expression, C-SIDE generates plots to visualize predicted and observed gene expression as a function of a particular covariate. C-SIDE can also use multiple covariates in a joint model of gene expression, such as spatial position and cell type colocalization, although more complicated models require more data to fit accurately. Beyond individual samples, C-SIDE can also model biological and technical variability in complex multi-sample, multi-replicate experiments. Multi-replicate experiments, although more costly, produce more robust DE estimates by reducing spurious discoveries of DE on single replicates.

One challenge for C-SIDE is obtaining sufficient DE detection statistical power, which can be hindered by low gene expression counts, small pixel number or rare cell types. C-SIDE increases its statistical power by including cell type mixture pixels in its model. Ongoing technical improvements in spatial transcriptomics technologies² such as increased gene expression counts, higher spatial resolution and increased pixel number, will increase the discovery rate of C-SIDE. Another limitation of C-SIDE is the requirement of an annotated single-cell reference for reference-based identification of cell types. Although single-cell atlases are increasingly available, they may contain missing cell types or substantial platform effects¹⁸ and certain spatial transcriptomics tissues may lack a corresponding single-cell reference. An ongoing challenge for spatial transcriptomics is to learn cell type proportions in the absence of an annotated single cell reference.

We envision C-SIDE to be particularly powerful in detecting cell type-specific gene expression changes in pathology. First, previous Alzheimer's disease studies have discovered candidate genes for disease-relevance through genome-wide association studies, bulk RNA, proteomics and scRNA-seq^{34,55}. Here, with C-SIDE using A β plaques as a covariate, we identify many genes previously identified by these methods including *Gfap* in astrocytes³² and *Apoe* in

microglia³⁹; furthermore, we progress a step further toward mechanistic understanding by directly associating spatial plaque localization with cell type-specific DE. For example, previous studies have associated complement pathway activation in plaque-dense areas with synaptic pruning³³ and neuronal degeneration³⁴. Using C-SIDE we specifically assign complement protein *C4b* plaque-localized activation to astrocytes⁵⁶, which could be caused by a plaque-triggered, cytokine-dependent signaling cascade³⁵. Additionally, the downregulation of the homeostatic microglia marker *P2ry12* in Alzheimer's disease is associated with neuronal cell loss³⁷. Using C-SIDE, we further localize this downregulation to plaque-associated microglia, suggesting that plaque-dense areas trigger microglia activation and downregulate homeostatic microglia genes³⁸. Last, the anti-inflammatory gene *granulin* (*Grn*), discovered by C-SIDE as upregulated in microglia near plaques, attenuates microglia activation^{40,57}, potentially mitigating plaque deposition and cognitive pathological decline⁵⁸.

Second, C-SIDE has the potential to explain cellular interactions. For example, recent studies have characterized cell-to-cell interactions of immune cells influencing the behavior of tumor cells⁵¹. Likewise, on a Slide-seq dataset of a mouse tumor model, C-SIDE identified synergistic cell-to-cell signaling between tumor cells and myeloid immune cells. For example, *Ccl2*, upregulated in immune-adjacent tumor cells, chemotactically recruits myeloid cells and induces protumorigenic behavior, including growth, angiogenesis and metastasis, in myeloid cells⁵². Likewise, the EMT pathway, upregulated near myeloid cells, promotes tumor development and metastasis⁵⁴. Although C-SIDE can establish such associations, conclusive establishment of molecular mechanism requires future experimentation. Among other hypotheses, it is plausible that myeloid cells induce tumor cells to undergo the EMT transition, potentially through the *NF- κ B* (also identified as upregulated by C-SIDE) signaling pathway^{51,54}. Therefore, C-SIDE, combined with pathological measurements, can explain cell type-specific responses to disease. We envision C-SIDE as a powerful framework for studying the effects of spatial and environmental context on cellular gene expression in spatial transcriptomics data.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-022-01575-3>.

Received: 18 January 2022; Accepted: 15 July 2022;

Published online: 01 September 2022

References

- Rodrigues, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
- Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
- Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, 412 (2015).
- Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, 380 (2018).
- 10X Genomics: visium spatial gene expression. *10X Genomics* <https://www.10xgenomics.com/solutions/spatial-gene-expression/> (2020).
- Zollinger, D. R., Lingle, S. E., Sorg, K., Beechem, J. M. & Merritt, C. R. GeoMx RNA assay: High multiplex, digital, spatial analysis of RNA in FFPE tissue. In Ian A. Darby & Tim D. Hewitson (eds.) *In Situ Hybridization Protocols*, 331–345 (Springer, 2020).
- Alon, S. et al. Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, 481 (2021).

8. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, 792 (2018).
9. Chen, H. et al. Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell Rep.* **37**, 109915 (2021).
10. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
11. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
12. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* **22**, 184 (2021).
13. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
14. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 12 (2014).
16. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
17. Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
18. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2022).
19. Regev, A. et al. Science forum: the human cell atlas. *eLife* **6**, e27041 (2017).
20. Petukhov, V. et al. Cell segmentation in imaging-based spatial transcriptomics. *Nat. Biotechnol.* **40**, 345–354 (2022).
21. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
22. Dong, R. & Yuan, G. C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **22**, 145 (2021).
23. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
24. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).
25. Hardin, J. W., Hardin, J. W., Hilbe, J. M. & Hilbe, J. *Generalized Linear Models and Extensions* (Stata Press, 2007).
26. Wood, S. & Wood, M. S. Package ‘mgcv’. R package version 1.29 (R Foundation for Statistical Computing, 2015).
27. Kozareva, V. et al. A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature* **598**, 214–219 (2021).
28. Zhao, M., Shirley, C. R., Mounsey, S. & Meistrich, M. L. Nucleoprotein transitions during spermiogenesis in mice with transition nuclear protein Tnp1 and Tnp2 mutations. *Biol. Reproduction* **71**, 1016–1025 (2004).
29. Hasegawa, K. & Saga, Y. Retinoic acid signaling in Sertoli cells regulates organization of the blood-testis barrier through cyclical changes in gene expression. *Development* **139**, 4347–4355 (2012).
30. Xu, J. et al. Computerized spermatogenesis staging (CSS) of mouse testis sections via quantitative histomorphological analysis. *Med. Image Anal.* **70**, 101835 (2021).
31. Mucke, L. et al. High-level neuronal expression of A β 1–42 in wild-type human amyloid protein precursor transgenic mice: Synaptotoxicity without plaque formation. *J. Neurosci.* **20**, 4050–4058 (2000).
32. Kraft, A. W. et al. Attenuating astrocyte activation accelerates plaque pathogenesis in APP/PS1 mice. *FASEB J.* **27**, 187–198 (2013).
33. Hong, S. et al. Complement and microglia mediate early synapse loss in Alzheimer mouse models. *Science* **352**, 712–716 (2016).
34. Zhou, Y. et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer’s disease. *Nat. Med.* **26**, 131–142 (2020).
35. Veerhuis, R. et al. Cytokines associated with amyloid plaques in Alzheimer’s disease brain stimulate human glial and neuronal cell cultures to secrete early complement proteins, but not C1-inhibitor. *Exp. Neurol.* **160**, 289–299 (1999).
36. Bernstein, H. G. & Keilhoff, G. Putative roles of cathepsin B in Alzheimer’s disease pathology: the good, the bad, and the ugly in one? *Neural Regen. Res.* **13**, 2100 (2018).
37. Sobue, A. et al. Microglial gene signature reveals loss of homeostatic microglia associated with neurodegeneration of Alzheimer’s disease. *Acta Neuropathol. Commun.* **9**, 1 (2021).
38. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer’s disease. *Cell* **169**, 1276–1290 (2017).
39. Serrano-Pozo, A., Das, S. & Hyman, B. T. APOE and Alzheimer’s disease: advances in genetics, pathophysiology, and therapeutic approaches. *Lancet Neurol.* **20**, 68–80 (2021).
40. Mendaikhan, A., Tooyama, I. & Walker, D. G. Microglial progranulin: involvement in Alzheimer’s disease and neurodegenerative diseases. *Cells* **8**, 230 (2019).
41. Zhou, X. et al. Cellular and molecular properties of neural progenitors in the developing mammalian hypothalamus. *Nature Commun.* **11**, 4063 (2020).
42. Romanov, R. A. et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**, 176–188 (2017).
43. V1_Human_Lymph_Node—Datasets—Spatial Gene Expression https://support.10xgenomics.com/spatial-geneexpression/datasets/1.1.0/V1_Human_Lymph_Node (10X Genomics, 2020).
44. Milpied, P. et al. Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. *Nat. Immunol.* **19**, 1013–1024 (2018).
45. Abe, Y. et al. A single-cell atlas of non-haematopoietic cells in human lymph nodes and lymphoma reveals a landscape of stromal remodelling. *Nat. Cell Biol.* **24**, 565–578 (2022).
46. Weinstein, A. M. & Storkus, W. J. In Wang, X.-Y. & Fisher, P. B. (eds.) *Immunotherapy of Cancer Vol. 128 Advances in Cancer Research* 197–233 (Academic Press, 2015).
47. Zhao, T. et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).
48. Dang, C. V. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol. Cell. Biol.* **19**, 1 (1999).
49. Jiménez-Sánchez, J. et al. Evolutionary dynamics at the tumor edge reveal metabolic imaging biomarkers. *Proc. Natl Acad. Sci. USA* **118**, 110–118 (2021).
50. Kodama, M. et al. In vivo loss-of-function screens identify KPNB1 as a new druggable oncogene in epithelial ovarian cancer. *Proc. Natl. Acad. Sci. USA* **114**, E7301–E7310 (2017).
51. Chen, D. P. et al. Peritumoral monocytes induce cancer cell autophagy to facilitate the progression of human hepatocellular carcinoma. *Autophagy* **14**, 1335–1346 (2018).
52. Lim, S. Y., Yuzhalin, A. E., Gordon-Weeks, A. N. & Muschel, R. J. Targeting the CCL2-CCR2 signalling axis in cancer metastasis. *Oncotarget* **7**, 28697 (2016).
53. Pires, B. R. et al. NF-kappaB is involved in the regulation of EMT genes in breast cancer cells. *PLoS ONE* **12**, e0169622 (2017).
54. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).
55. Satoh, J.-i. et al. TMEM106B expression is reduced in Alzheimer’s disease brains. *Alzheimer’s Res. Ther.* **6**, 17 (2014).
56. Walker, D. G., Kim, S. U. & McGeer, P. L. Expression of complement C4 and C9 genes by human astrocytes. *Brain Res.* **809**, 31–38 (1998).
57. Götzl, J. K. et al. Opposite microglial activation stages upon loss of PGRN or TREM 2 result in reduced cerebral glucose metabolism. *EMBO Mol. Med.* **11**, e9711 (2019).
58. Minami, S. S. et al. Progranulin protects against amyloid β deposition and toxicity in Alzheimer’s disease mouse models. *Nat. Med.* **20**, 1157–1164 (2014).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Methods

Ethics statement. All procedures involving animals at the Broad Institute were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals under protocol number 0120-09-16. No field samples were collected for this study.

C-SIDE model. Here, we describe C-SIDE, a statistical method for identifying DE in spatial transcriptomics data. We define the C-SIDE model in equations (1)–(3). Before fitting C-SIDE, the design matrix is predefined to contain covariates, variables on which gene expression is hypothesized to depend such as spatial position or cellular microenvironment. Recall that $x_{i,\ell,g}$ represents the ℓ th covariate, evaluated at pixel i in experimental sample g . For each covariate $x_{i,\ell,g}$, there is a corresponding coefficient $\alpha_{\ell,k,j,g}$, representing a gene expression change across pixels per unit change of $x_{i,\ell,g}$ within cell type k of experimental sample g . Next, recall from equation (2) random effects $\gamma_{j,g}$ and $\epsilon_{i,\ell,g}$, which we assume both follow normal distributions with mean 0 and standard deviations $\sigma_{\gamma,g}$ and $\sigma_{\epsilon,g}$, respectively. The overdispersion magnitude, $\sigma_{\epsilon,j,g}$, varies across gene j (Supplementary Fig. 12) and modeling gene-specific overdispersion is necessary for controlling the false-positive rate of C-SIDE.

Due to our finding that genes can exhibit DE in some but not all cell types (Fig. 3c), C-SIDE generally does not assume that genes share DE patterns across cell types, allowing for the discovery of cell type-specific DE. We also developed an option where DE can be assumed to be shared across cell types (Supplementary Notes). C-SIDE can be thought of as a modification of the generalized linear model (GLM)²⁵ in which each cell type follows a cell type-specific log-linear model before an additive mixture of all cell types is observed. See Fitting the C-SIDE model and Hypothesis testing for C-SIDE model fitting and hypothesis testing, respectively.

Parameterization of the design matrix. For specific construction of design matrix x for each dataset, see Cell type proportion estimation and covariate construction. Recall the specific examples of design matrix x presented in Fig. 1b. In general, x can obtain the following forms:

- (1) Indicator variable. In this case, $x_{i,\ell,g}$ is always either 0 or 1, representing DE due to membership within a certain spatially defined pixel set. The coefficient $\alpha_{k,j,g}$ is interpreted as the log-ratio of gene expression between the two sets for cell type k and gene j in experimental sample g .
- (2) Continuous variable. In this case, $x_{i,\ell,g}$ can take on continuous values representing, for example, distance from some feature or density of some element. The coefficient $\alpha_{\ell,k,j,g}$ is interpreted as the log fold-change of gene expression per unit change in $x_{i,\ell,g}$ for cell type k and gene j in sample g .
- (3) Multiple categories. In this case, we use x to encode membership in $L \geq 2$ sets. For each $1 \leq \ell \leq L$, we define $x_{i,\ell,g}$ to be an indicator variable representing membership in set ℓ for sample g . To achieve identifiability, the intercept is removed. The coefficient $\alpha_{\ell,k,j,g}$ is interpreted as the average gene expression in set ℓ for cell type k and gene j . Cell type-specific DE is determined by detecting changes in $\alpha_{\ell,k,j,g}$ across ℓ within cell type k and sample g .
- (4) Nonparametric. In this case, we use x to represent L smooth basis functions, where linear combinations of these basis functions represent the overall smooth gene expression function. By default, we use thin plate spline basis functions, calculated using the mgcv package²⁶.

In all cases, we normalize each $x_{i,\ell,g}$ to range between 0 and 1. The problem is equivalent under linear transformations of x , but this normalization helps with computational performance. The intercept term, when used, is represented in x as a column of 1s.

Fitting the C-SIDE model. C-SIDE estimates the parameters of (1), (2), and (3) via maximum likelihood estimation. First, all parameters are independent across samples, so we fit the model independently for each sample. For population inference across multiple samples, see Statistical inference on multiple samples/replicates. Next, the parameters of $\beta_{i,k}$ and γ_j are estimated by the RCTD algorithm¹⁸. C-SIDE can also optionally import cell type proportions from external cell type proportion identification methods^{21–23}. Here, some pixels are identified as single cell types while others as mixtures of multiple cell types. We can accurately estimate cell type proportions and platform effects without being aware of differential spatial gene expression because differential spatial gene expression (average 0.38 in for example cerebellum Slide-seq data) is smaller than gene expression differences across cell types (average s.d. 1.09 in for example cerebellum Slide-seq data). After determining cell type proportions, C-SIDE estimates gene-specific overdispersion magnitude $\sigma_{\epsilon,j,g}$ for each gene by maximum likelihood estimation (Supplementary Notes). Finally, C-SIDE estimates the DE coefficients α by maximum likelihood estimation. For the final key step of estimating α , we use plugin estimates (denoted by $\hat{\cdot}$) of β , γ and σ_{ϵ} . After we substitute equation (3) into equation (1) and (2), we obtain:

$$Y_{i,j,g} | \epsilon_{i,j,g} \approx \text{Poisson} \left\{ N_{i,g} \exp \left[\log \left(\sum_{k=1}^K \hat{\beta}_{i,k,g} \exp \left(\alpha_{0,k,j,g} + \sum_{\ell=1}^L x_{i,\ell,g} \alpha_{\ell,k,j,g} \right) \right) + \hat{\gamma}_{j,g} + \epsilon_{i,j,g} \right] \right\} \quad (4)$$

$$\epsilon_{i,j,g} \approx \text{Normal}(0, \sigma_{\epsilon,j,g}^2), \quad (5)$$

We provide an algorithm for computing the maximum likelihood estimator of α , presented in the Supplementary Notes. Our likelihood optimization algorithm is a second-order, trust-region²⁹ based optimization (Supplementary Notes). In brief, we iteratively solve quadratic approximations of the log-likelihood, adaptively constraining the maximum parameter change at each step. Critically, the likelihood is independent for each gene j (and sample g) so separate genes are run in parallel, in which case there are $K \times (L+1)\alpha$ parameters per gene and sample.

Hypothesis testing. In addition to estimating the vector $\alpha_{j,g}$ (dimensions $L+1$ by K) for gene j and sample g , we can compute standard errors around $\alpha_{j,g}$. By asymptotic normality³⁰ (Supplementary Notes), we have approximately that (setting n to be the total number of pixels),

$$\sqrt{n}(\hat{\alpha}_{j,g} - \alpha_{j,g}) \approx \text{Normal}(0, I_{\alpha_{j,g}}^{-1}), \quad (6)$$

where $I_{\alpha_{j,g}}$ is the Fisher information of the model in equation (4), which is computed in the Supplementary Notes. Given this result, we can compute standard errors, confidence intervals and hypothesis tests. As a consequence of equation (6), the standard error of $\alpha_{\ell,k,j,g}$, denoted $s_{\ell,k,j,g}$, is $\sqrt{(I_{\alpha_{j,g}}^{-1})_{\ell,k,j,g}/n}$.

First, we consider the case where we are interested in a single parameter, $\alpha_{\ell,k,j,g}$, for ℓ and g fixed and for each cell type k and gene j ; for example, $\alpha_{\ell,k,j,g}$ could represent the log fold-change between two discrete regions. In this case, for each gene j , we compute the z-statistic, $z_{\ell,k,j,g} = \frac{\alpha_{\ell,k,j,g}}{s_{\ell,k,j,g}}$. Using a two-tailed z-test, we compute a P value for the null hypothesis that $\alpha_{\ell,k,j,g} = 0$ as $p_{\ell,k,j,g} = 2 \times F(-|z_{\ell,k,j,g}|)$, where F is the distribution function of the standard normal distribution. Finally, q values are calculated across all genes within a cell type to control the FDR using the Benjamini–Hochberg procedure⁴¹. We used a FDR of 0.01 (0.1 for nonparametric case) and a fold-change cutoff of 1.5 (not applicable for nonparametric case). Additionally, for each cell type, genes were prefiltered so that the expression within the cell type of interest had a total expression of at least 15 UMIs over all pixels and that the mean normalized expression is at least $r\%$ as large as expression within each other cell type. The parameter $r\%$ (default 50%, set to 25% on Alzheimer's dataset) is used to filter out marker genes of other cell types that may contaminate the DE estimates of the cell type of interest. We recommend setting r between 25 and 50%, trading off between increasing DE discoveries and risking false discoveries due to marker gene contamination.

For the multi-region case, we test for differences of pairs of parameters representing the average expression within each region, correcting for multiple hypothesis testing by scaling P values. We select genes that have significant differences between at least one pair of regions. For other cases in which we are interested in multiple parameters, for example the nonparametric case, we test each parameter individually and scale P values due to multiple hypothesis testing.

Statistical inference on multiple samples/replicates. When C-SIDE is run on multiple replicates, we recall α_g and s_g are the DE and standard error for replicate g , where $1 \leq g \leq G$ and $G > 1$ is the total number of replicates. We now consider testing for DE across all replicates for covariate ℓ , cell type k and gene j . In this case, we assume that additional biological or technical variation across samples exists, such that each unknown α_g is normally distributed around a population-level DE A , with standard deviation t :

$$\alpha_{\ell,k,j,g} \stackrel{i.i.d.}{\sim} \text{Normal}(A_{\ell,k,j}, t_{\ell,k,j}^2). \quad (7)$$

Under this assumption, and using equation (6) for the distribution of the observed single-sample estimates $\hat{\alpha}$, we derive the following feasible generalized least squares estimator of A (Supplementary Notes),

$$\hat{A}_{\ell,k,j} := \frac{\sum_{g=1}^G (\hat{\alpha}_{\ell,k,j,g}) / (\hat{t}_{\ell,k,j}^2 + s_{\ell,k,j,g}^2)}{\sum_{g=1}^G 1 / (\hat{t}_{\ell,k,j}^2 + s_{\ell,k,j,g}^2)}. \quad (8)$$

Here, $\hat{\alpha}$ and s are obtained from C-SIDE estimates on individual samples (equation (6)), whereas \hat{t}^2 represents the estimated variance across samples (Supplementary Fig. 12). Please see the Supplementary Notes for additional details such as the method of moments procedure³² for estimating $\hat{t}_{\ell,k,j}^2$ and the standard errors of A . Intuitively, our estimate of the population-level DE is a variance-weighted sum over the DE estimates of individual replicates, similar to meta-analysis methods³². We next use these estimates and standard errors to test the hypothesis that $A_{\ell,k,j} = 0$ as described in Hypothesis testing. For the case of multiple biological samples and multiple replicates within each sample, see the Supplementary Notes. Multiple sample inference with nonparametric C-SIDE requires a common coordinate system, and we create a common spline basis and then test each coefficient across all samples.

Spatial transcriptomics, scRNA-seq, Aβ imaging and HCR data. Using the Slide-seqV2 protocol¹² (Supplementary Notes), we collected four Alzheimer's

Slide-seq mouse hippocampus sections³¹ on a female 8.8-month-old J20 Alzheimer's mouse model³¹ and three Slide-seq mouse cerebellum sections (one from a previous study¹⁸). The Slide-seq mouse testes⁹ and cancer⁴⁷, MERFISH hypothalamus⁸, and Visium lymph node⁴³ datasets were obtained from previous studies. The tumor dataset represented a *Kras*^{G12D/+}/*Trp53*^{-/-} (KP) mouse metastatic lung adenocarcinoma tumor in the liver. We used cell type-annotated scRNA-seq datasets for the testes⁶³, hypothalamus⁸, cerebellum²⁷, cancer⁴⁷, lymph node²³ and Alzheimer's hippocampus datasets⁵⁴.

Slide-seq data was preprocessed using the Slide-seq tools pipeline³. Spatial transcriptomic spots were filtered to a minimum of 100 UMIs, and the region of interest (ROI) was cropped before running C-SIDE using overall anatomical features. For example, in Slide-seq Alzheimer's hippocampus, the somatosensory cortex was cropped out before analysis.

For the Alzheimer's dataset, to define an amyloid plaque C-SIDE covariate, we collected fluorescent images of DAPI (4,6-diamidino-2-phenylindole) and amyloid beta (A β), using IBL America Amyloid Beta (N) (82E1) A β Anti-Human Mouse IgG MoAb on sections adjacent to the Slide-seq data. We coregistered the DAPI image to the adjacent Slide-seq total UMI image using the ManualAlignImages function from the STUtility R package⁴⁵. To calculate plaque density, plaque images were convolved with an exponentially decaying isotropic filter, using a threshold at the 0.9 quantile, normalized to be between 0 and 1, and averaged over two adjacent amyloid sections.

For *in situ* RNA hybridization validation of cerebellum DE results, we collected HCR data on genes *Aldoc*, *Kcnd2*, *Mybpc1*, *Plcb4* and *Tmem132c* (Supplementary Table 10)⁶⁶. We simultaneously collected cell type marker genes²⁷ of Bergmann (*Gdf10*), granule (*Gabra6*) and Purkinje (*Cab1*) cell types. Data from *Kcnd2* were removed due to measuring tissue autofluorescence rather than RNA. ROI of nodular and anterior regions were cropped, and background, defined as median signal, was subtracted. For this data, DE was calculated as the log fold-change, across ROI, of average gene signal over the pixels within the ROI containing cell type markers of a particular cell type. Pixels containing marker genes of multiple cell types were removed. C-SIDE single-sample standard errors in Fig. 3d were calculated by modeling single-sample variance as the sum of the variance across samples and variance representing uncertainty around the population mean.

Cell type proportion estimation and covariate construction. For each dataset, we constructed at least one covariate, an axis along which to test for DE. All covariates were scaled linearly to have minimum 0 and maximum 1. For the cerebellum dataset, the covariate was defined as an indicator variable representing membership within the nodular region (versus the anterior region). For the testes dataset, a discrete covariate represented the cellular microenvironment of tubule stage, labels obtained from tubule-level gene expression clustering (four stages, I–III, IV–VI, VII–VIII and IX–XII) from the previous Slide-seq testes study⁹. For the cancer dataset, the myeloid cell type density covariate was calculated by convolving the cell type locations, weighted by UMI number, with an exponential filter. For this dataset, we also ran C-SIDE nonparametrically. For the Alzheimer's hippocampus dataset, see Spatial transcriptomics, scRNA-seq, A β imaging and HCR data. For the MERFISH hypothalamus dataset, the covariate was linear or quadratic midline distance. For the quadratic MERFISH C-SIDE model, we conducted hypothesis testing on the quadratic coefficient. To estimate platform effects and cell type proportions, RCTD ran with default parameters on full mode for the testes and lymph node datasets and doublet mode for other datasets¹⁸.

Validation with simulated gene expression dataset. We created a ground truth DE simulation, from the cerebellum single-nucleus RNA-seq dataset, to test C-SIDE on mixtures between two cell type layers. We restricted to Purkinje and Bergmann cell types, which are known to spatially colocalize. To simulate a cell type mixture of cell types A (Purkinje) and B (Bergmann), we randomly chose a cell from each cell type, and sampled a predefined number of UMIs from each cell (total 1,000). We defined two discrete spatial regions (Fig. 1a), populated with A/B cell type mixtures. We varied the mean cell type proportion difference across the two regions and also simulated the case of cell type proportions evenly distributed across the two regions. Cell type-specific spatial differential gene expression also was simulated across the two regions. To simulate cell type-specific DE in the gene expression step of the simulation, we multiplicatively scaled the expected gene counts within each cell of each cell type. An indicator variable for the two spatial bins was used as the C-SIDE covariate.

Additional computational analysis. For confidence intervals on data points or groups of data points (Fig. 1d,k), we used the predicted variance of data points from C-SIDE (Supplementary Notes). Likewise, for such analysis we used predicted counts from C-SIDE at each pixel (Supplementary Notes). For the testes dataset, a cell type was considered to be present on a bead if the proportion of that cell type was at least 0.25 (Fig. 4d). Additionally, cell type and stage-specific marker genes were defined as genes that had a fold-change of at least 1.5 within the cell type of interest compared to each other cell type. We also required significant cell type-specific DE between the stage of interest with all other stages (fold-change of at least 1.5, significance at the level of 0.001, Monte Carlo test on Z-scores). Cyclic genes were defined as genes whose minimum expression within a cell type

occurred two tubule stages away from its maximum expression, up to log space error of up to 0.25. For analyzing C-SIDE on cerebellum subtypes, we focused on five granule subtypes since granule was the most commonly occurring cell type, thus yielding the most statistical power (Supplementary Fig. 7). We tested for significant differences in the estimated DE coefficients between the subtype model and the original model without subtypes.

For nonparametric C-SIDE on the tumor dataset, we used hierarchical Ward clustering to cluster quantile-normalized spatial gene expression patterns into seven clusters. For gene set testing on the tumor dataset, we tested the 50 hallmark gene sets from the MSigDB database⁶⁷ for aggregate effects in C-SIDE DE estimates for the tumor cell type. For the nonparametric case, we used a binomial test with multiple hypothesis correction to test for enrichment of any of the seven spatial clusters of C-SIDE-identified significant genes in any of the 50 gene sets. For the parametric case, we used a permutation test on the average value of C-SIDE Z-scores for a gene set. That is, we modified an existing gene set enrichment procedure⁶⁸ by filtering for genes with a fold-change of at least 1.5 and using a two-sided permutation test rather than assuming normality. In both cases, we filtered to gene sets with at least five genes and we used Benjamini–Hochberg procedure across all gene sets to control the FDR at 0.05. The proportion of variance not due to sampling noise (Fig. 6b) was calculated by considering the difference between observed variance on normalized counts and the expected variance due to Poisson sampling noise.

We considered and tested several simple alternative methods to C-SIDE, which represent general classes of approaches. First, we considered a two-sample Z-test on single cells (defined as pixels with cell type proportion at least 0.9). Additionally, we tested bulk DE, which estimated DE as the log-ratio of average normalized gene expression across two regions. The single method of DE rounded cell type mixtures to the nearest single cell type and computed the log-ratio of gene expression of cells in that cell type. Finally, the decompose method of DE used a previously developed method to compute expected gene expression counts for each cell type¹⁸, followed by computing the ratio of cell type-specific gene expression in each region.

Implementation details. C-SIDE is publicly available as part of the R package <https://github.com/dmcable/spacexr>. The quadratic program that arises in the C-SIDE optimization algorithm is solved using the quadprog package in R⁶⁹. Before conducting analysis on C-SIDE output, all ribosomal proteins and mitochondrial genes were filtered out. Additional parameters used for running C-SIDE are shown in Supplementary Table 11. C-SIDE was tested on a Macintosh laptop computer with a 2.4 GHz Intel Core i9 processor and 32 GB of memory (we recommend at least 4 GB of memory to run C-SIDE). For example, we timed C-SIDE with four cores on one of the Slide-seq cerebellum replicates, containing 2,776 pixels across two regions, five cell types and 4,812 genes. Under these conditions, C-SIDE ran in 13 min and 47 s (excluding the cell type assignment step in which computational efficiency has been described previously¹⁸).

C-SIDE can be run using a general covariate matrix with the function run.CSIDE. In most cases, a more specific function exists for specific covariates including run.CSIDE.single (one covariate), run.CSIDE.regions (multiple regions) or run.CSIDE.nonparam (nonparametric). Moreover, to generate covariates for example run.CSIDE.single, the functions exvar.celltoall.interactions and exvar.point.density can be used to calculate covariates for the density of cell types or additional features, respectively. For multiple replicates/samples, the function run.CSIDE.replicates can be used, while the function CSIDE.population.inference performs population inference across these replicates. Multi-replicate analysis requires covariates to be consistently defined, for example, in nonparametric mode the coordinates must exist in a common coordinate system, which can be obtained, for instance, by alignment. Please see <https://github.com/dmcable/spacexr/tree/master/vignettes> for Vignettes for running C-SIDE, <https://github.com/dmcable/spacexr/tree/master/documentation> for additional documentation and Supplementary Software 1 for the manual for the spacexr R package.

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

Data availability

Slide-seq V2 data generated for this study and additional data are available at the Broad Institute Single Cell Portal https://singlecell.broadinstitute.org/single_cell/study/SCP1663. We also used the following publicly available datasets in our study. MERFISH hypothalamus dataset was accessed from Dryad <https://doi.org/10.5061/dryad.8t8s248>. Visium human lymph node is available at <https://www.10xgenomics.com/resources/datasets/human-lymph-node-1-standard-1-1-0>. Testes Slide-seq data can be accessed at https://www.dropbox.com/s/ygpj0d0oh67br0/Testis_Slideseq_Data.zip?dl=0. Cancer Slide-seq data are available at https://singlecell.broadinstitute.org/single_cell/study/SCP1278. Hallmark gene sets were accessed from <https://www.gsea-msigdb.org/>.

Code availability

C-SIDE is implemented in the open-source R package spacexr, with source code freely available at <https://github.com/dmcable/spacexr>. Additional code used for

analysis in this paper is available at <https://github.com/dmcable/spacexr/tree/master/AnalysisCSIDE>.

References

59. Yuan, Y. X. A review of trust region algorithms for optimization. In *Proc. 4th International Congress on Industrial & Applied Mathematics (ICIAM 99)*, Edinburgh 271–282 (Oxford Univ. Press, 2000).
60. Van der Vaart, A. W. *Asymptotic Statistics* Vol. 3 (Cambridge Univ. Press, 2000).
61. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
62. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
63. Green, C. D. et al. A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq. *Dev. Cell* **46**, 651–667 (2018).
64. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
65. Bergensträhle, J., Larsson, L. & Lundberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 482 (2020).
66. Dirks, R. M. & Pierce, N. A. Triggered amplification by hybridization chain reaction. *Proc. Natl Acad. Sci. USA* **101**, 15275–15278 (2004).
67. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
68. Irizarry, R. A., Wang, C., Zhou, Y. & Speed, T. P. Gene set enrichment analysis made simple. *Stat. Methods Med. Res.* **18**, 565–575 (2009).
69. Turlach, B. A. & Weingessel, A. quadprog: Functions to solve quadratic programming problems. R package version 1.5-5 (R Foundation for Statistical Computing, 2013).

Acknowledgements

We thank R. Stickels for providing valuable input on the analysis. We thank T. Zhao and Z. Chiang for generously providing the cancer Slide-seq data. We thank S. Marsh (Harvard Medical School/Boston Children's Hospital) for kindly providing mouse

J20 Alzheimer's model samples. We thank members of the Chen laboratory, Irizarry laboratory and Macosko laboratory including T. Kamath for helpful discussions and feedback. D.M.C. was supported by a Fannie and John Hertz Foundation Fellowship and an National Science Federation Graduate Research Fellowship. This work was supported by an National Institutes of Health (NIH) Early Independence Award (DP5, 1DP5OD024583 to F.C.), the NHGRI (R01, R01HG010647 to F.C. and E.Z.M.), as well as the Burroughs Wellcome Fund, the Searle Scholars Award, and the Merkin Institute to F.C. R.A.I. was supported by NIH grant nos. R35GM131802 and R01HG005220.

Author contributions

D.M.C., R.A.I. and F.C. conceived the study. F.C., E.M., E.Z.M. and D.M.C. designed the Slide-seq, antibody stain and HCR experiments. E.M. generated the Slide-seq, antibody stain and HCR data. D.M.C., R.A.I. and F.C. developed the statistical methods. D.M.C., F.C. and R.A.I. designed the analysis. D.M.C., S.Z., L.S.Z., M.D., R.A.I. and F.C. analyzed the data. D.M.C., F.C., R.A.I., V.S. and H.C. interpreted biological results. V.S. annotated the tumor H&E stain. D.M.C., F.C. and R.A.I. wrote the manuscript and all authors read and approved the final manuscript.

Competing interests

E.Z.M. and F.C. are listed as inventors on a patent application related to Slide-seq. F.C. and E.Z.M. are paid consultants of Atlas Bio. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-022-01575-3>.

Correspondence and requests for materials should be addressed to Rafael A. Irizarry or Fei Chen.

Peer review information *Nature Methods* thanks Pengyi Yang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

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

Statistics

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

n/a Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Slide-seq data was collected and preprocessed using the publicly available slide-seq tools computational pipeline. HCR data and antibody stains were collected using standard protocols that are described in our experimental methods section.

Data analysis CSIDE is implemented as part of the open-source R package spacexr (version 2.0.0), with source code freely available at: <https://github.com/dmcable/spacexr>. Additional code used for analysis in this paper is available at: <https://github.com/dmcable/spacexr/tree/master/AnalysisCSIDE>. Additional software used includes the RCTD algorithm (part of spacexr package)and quadprog R package (version 1.5.8).

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

Slide-seq V2 data generated for this study and additional data are available at the Broad Institute Single Cell Portal https://singlecell.broadinstitute.org/single_cell/study/SCP1663. Hallmark gene set was accessed from <https://www.gsea-msigdb.org/>.

Field-specific reporting

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

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

Sample size	CSIDE was validated across 6 spatial transcriptomics datasets. The sample size (number of pixels) for each dataset was determined by including pixels with at least 100 unique molecular identifiers (UMIs)
Data exclusions	One cerebellum HCR measurement (gene Kcnd2) was removed due to the HCR fluorescent channel failing to localize RNA molecules, but rather reflecting tissue autofluorescence
Replication	CSIDE was validated across 6 spatial transcriptomics datasets. On the J20 Slide-seq data, we ran CSIDE on 4 Slide-seq replicate experiments. On the cerebellum Slide-seq data, we ran CSIDE across 3 Slide-seq replicate datasets.
Randomization	N/A. We didn't have multiple experimental groups across our biological samples.
Blinding	N/A. We didn't have multiple experimental groups across our biological samples.

Reporting for specific materials, systems and methods

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

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies	<input type="checkbox"/>	ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	MRI-based neuroimaging
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		

Antibodies

Antibodies used	To measure A-beta plaque, we used IBL America Amyloid Beta (N) (82E1) A-beta Anti-Human Mouse IgG MoAb. Catalog number 10323. The antibody dilution was 1:100.
Validation	The antibody was validated in: Horikoshi Y, Sakaguchi G, Becker AG, Gray AJ, Duff K, Aisen PS, Yamaguchi H, Maeda M, Kinoshita N, Matsuoka Y. Development of Abeta terminal end-specific antibodies and sensitive ELISA for Abeta variant. Biochem Biophys Res Commun. 2004 Jul 2;319(3):733-7.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Adult, C57BL/6 mouse; female 8.8 month old J20 Alzheimer's mouse. C57BL/6 mice were housed in a 12:12 light-dark cycle with ad libitum access to food and water.
Wild animals	No wild animals were used in this study.
Field-collected samples	No field collected samples were used in this study.
Ethics oversight	All procedures involving animals at the Broad Institute were conducted in accordance with the US

Note that full information on the approval of the study protocol must also be provided in the manuscript.