Estimation of cell type reference signatures from scRNA-seq. Given cell type annotation for each cell, the corresponding reference cell type signatures $g_{f,g}$, which represent the average mRNA count of each gene g in each cell type $f = \{1, .., F\}$, can be estimated using a negative binomial regression model, which allows for combining data across batches and technologies (see below and Suppl. methods).

Cell2location model. An untransformed spatial expression count matrix $d_{s,g}$ is used for input, as obtained from the 10X SpaceRanger software (10X Visium data). Cell2location models the elements of $d_{s,g}$ as Negative Binomial (NB) distributed, given an unobserved gene expression level (rate) $\mu_{s,g}$ and a gene-specific over-dispersion α_g:

$$d_{s,g} \sim NB(\mu_{s,g}, \alpha_g).$$

The expression level of genes $\mu_{s,g}$ in the mRNA count space is modelled as a linear function of expression signatures of reference cell types $g_{f,g}$:

$$\mu_{s,g} = \underbrace{m_g}_{technology\ sensitivity} \cdot \underbrace{\left(\sum_f w_{s,f}\ g_{f,g}\right)}_{cell\ type\ contributions} + \underbrace{l_s + s_g}_{additive\ shift},$$

where, $w_{s,f}$ denotes regression weight of each reference signature f at location s, which can be interpreted as the number of cells at location s that express reference signature f; m_g is a gene-specific scaling parameter, which adjusts for global differences in sensitivity between technologies; l_s and s_g are additive variables that account for gene- and location-specific shift, such as due to contaminating or free-floating RNA.

To account for the similarity of location patterns across cell types, $w_{s,f}$ is modelled using another layer of decomposition (factorization) using $r = \{1, .., R\}$ groups of cell types, that can be interpreted as cellular compartments or tissue zones (Suppl. Methods). Unless stated otherwise, R is set to 50.

Approximate Variational Inference is used to estimate all model parameters, implemented in the pymc3 framework [52], which supports GPU acceleration. For full details see Supp. Methods.

Note on selecting scRNA-seq profiles for constructing reference cell type data. It is important to aim for a comprehensive and detailed cell-type reference, which includes as many of the cell types and subpopulations that are present *in-situ* as possible, for example, by generating a paired snRNA-seq reference from the same tissue sample. However, imperfect matching of cell populations is often acceptable (see Fig 4, Fig S4D). In such instances, the stability of the model fit, which can be assessed using multiple random restarts, can serve as diagnostic criteria (see Supp. Methods).

Note on selecting the method for estimating reference signatures of cell types. The first step of our model is to estimate reference cell type signatures from sc/snRNA-seq profiles, by providing the model with annotated cell type and subpopulation labels for each cell. The cell2location software comes with two implementations for this estimation step: 1) a statistical method based on Negative Binomial regression and 2) hard-coded computation of per-cluster average mRNA counts for individual genes. We generally recommend using NB regression, which allows to robustly combine data across technologies and batches (Fig S23),

which results in improved spatial mapping accuracy (Fig S22B). However, when the batch effects are small a faster hard-coded method of computing per cluster averages provides similarly high accuracy (Fig S22A). We also recommend the hard-coded method for non-UMI technologies such as Smart-Seq 2.

Hyperparameter selection. The cell2locaiton model has 4 hyper-priors, which can be set by the user taking known experimental and biological characteristics of a given dataset into consideration:

1) Expected number of cells per location $\widehat{N}$
2) Expected number of cell types per location $\hat{Y}$
3) Expected number co-abundance cell type groups per location $\hat{A}$
4) Expected mean of gene-specific technology sensitivity parameter μ_m

The Fig S24 provides a flowchart of how the values of these hyper-priors can be determined.

Expected cell abundance $\widehat{N}$ per location is a tissue-level global estimate, which can be derived from histology images (H&E or DAPI), ideally paired to the spatial expression data or at least representing the same tissue type. This parameter can be estimated by manually counting nuclei in a 10-20 locations in the histology image (e.g. using 10X Loupe browser, Fig S8), and computing the average cell abundance. An appropriate setting of this prior is essential to inform the estimation of absolute cell type abundance values, however, the model is robust to a range of similar values (Fig S5). In settings where suitable histology images are not available, the size of capture regions relative to the expected size of cells can be used to estimate $\widehat{N}$ (Slide-Seq V2, Fig S24). For all analysis in this manuscript, a single tissue-level estimate was used, however, as an advanced feature, cell2location can utilise the per-location number of cells.

Expected number of cell types per location $\hat{Y}$ and expected number co-abundance cell type groups per location $\hat{A}$. The value of these hyper-priors has minimal effect on model accuracy (Fig S5). Consequently, we recommend setting their values to 7, a single global estimate.

The difference in technology sensitivity mean μ_m and variance ${\sigma_m}^2$ parameters can be chosen by comparing the average total number of mRNA per cell in the reference cell type data to the average total number of mRNA per location in the spatial data divided by $\widehat{N}$ (Fig S24).

While good choices of these hyper-parameters can have a positive impact on accuracy, the estimate of relative cell abundance is robust to a range of suboptimal choices (Fig S5). The estimation of absolute cell abundance requires appropriate settings of $\widehat{N}$ and μ_m in particular.

Constructing a synthetic spatial transcriptomics data set

Simulated spatial transcriptomics data were generated by combining expression profiles of cells drawn from each one of 49 cell types in the mouse brain snRNA-seq reference data (see below), to generate abundance profiles at 2,500 locations. snRNA data from the two most homogenous mouse brain snRNA-seq samples were split into one dataset used to generate the synthetic data (50% of cells) and a second dataset used to evaluate cell2location and alternative approaches (50% of cells), similarly to the strategy proposed by Andersson et al[3]. Hyperparameters for data simulation were chosen to mimic (the typically low) cell counts observed for cell types in real tissues, additionally matching sparsity profiles as observed in real data. Cell type abundances were simulated according to either a spatially ubiquitous

pattern (8 cell types), or a regional pattern (41 cell types). Regional patterns are represented by 12 tissue zones defined by co-located cell types that mimic the organisation of real tissues. The assignment of 41 regional cell types to the 12 tissue zones is shown in Fig S2, with each cell type belonging to 1-3 tissue zones and each tissue zone containing 2-8 cell types. The number of cell types present at each location, as well as the absolute abundance (the number of cells per location), were simulated according to either low or high average cell type abundance (Fig 1B), stratified by ubiquitous and regional location pattern (see below). The mathematical description and the step-by-step procedure to simulate abundance of cell types across locations and to generate multi-cell mRNA counts is described in detail below in three sections: 1) generating abundance of cell types across locations, 2) generating expected multi-cell mRNA expression of genes across locations, 3) generating multi-cell mRNA integer counts weighted by technology difference effect.

First, follow this step-by-step procedure to generate ground truth spatial abundance $w_{s,f}$, integer cell count $count_{s,f}$ and fraction of mRNA captured $fraq_{s,f}$ for cell types f across locations s:

1. Assign cell types to ubiquitous (n=8) and regional (n=41) abundance patterns (denoted as r).
2. Perform binary assignment of 41 sparse cell types to 12 tissue zones and 8 ubiquitous cell types to 8 ubiquitous patterns (total n=20), shown in Fig S2A and denoted as $x_{r,f}$.
3. Stratified by location pattern, randomly assign up to 20% of cell types to high abundance groups and all other cell types to low abundance groups. Generate per cell type average abundance d_f, which is different for 4 groups shown in Fig 1B:

   a) Ubiquitous and low density: 5 cell types present in most locations at density:

   $$d_f \sim Gamma(\mu = 1.0, \sigma^2 = \mu / 5).$$

   b) Ubiquitous and high density: 3 cell types present in most locations at density:

   $$d_f \sim Gamma(\mu = 2.8, \sigma^2 = \mu / 5).$$

   c) Sparse and low density: 32 cell types present in sparse tissue zones at density:

   $$d_f \sim Gamma(\mu = 1.0, \sigma^2 = \mu / 5).$$

   d) Sparse and high density: 9 cell types present in sparse tissue zones at density:

   $$d_f \sim Gamma(\mu = 2.8, \sigma^2 = \mu / 5).$$

   By following this procedure, sparsity and density parameters for each cell type were generated that produced an average total number of cells per location close to 10, mimicking cell count observed by nuclear segmentation of the mouse brain histology images (Fig S8, Suppl Methods).
   Per cell type maximum abundance d_f was used to scale $x_{r,f} = x_{r,f} * d_f$, thus defining the average abundance of each cell type across patterns r.
4. Generate spatial abundance $z_{s,r}$ for locations s for 20 location patterns (denoted as r) representing 12 tissue zones and 8 ubiquitous cell types. Gaussian Process in 50x50 grid of locations was used with randomly generated bandwidth parameters:

   $$bw \sim Gamma(\mu = 8.0, \sigma^2 = \mu / 1.2)$$

   a) 8 ubiquitous patterns with for non-zero density in most locations:

$$z_{s,r} \sim GP(\mu = 0, eta = 0.5, bw = bw)$$

b) 12 tissue zones with $z_{s,r} \sim GP(\mu = 0, eta = 1.5)$ for sparse locations

$$z_{s,r} \sim GP(\mu = 0, eta = 1.5, bw = bw)$$

To ensure positive scale, cell abundance for each pattern $z_{s,r}$ were softmax-transformed $z_{s,r} = exp(z_{s,r}) / \sum_r exp(z_{s,r})$. Next, to ensure that maximum for each location pattern r is equal to 1 further normalisation was applied: $z_{s,r} = z_{s,r} / \sum_s z_{s,r}$, which is needed to use abundances established in step 3 as average value for each cell type.

5. Per cell type abundance for each location s was generated as $w_{s,f} = (\sum_r z_{s,r} x_{r,f}) q_{s,f}$ (shown in Fig S2B), where $log(q_{s,f}) \sim Normal(\mu = 0, \sigma = 0.35)$ introduces randomness to abundances of individual cell types. This additional variability mimics the observation that co-located cell types in real tissues do not have perfectly correlated abundance within tissue zones.
6. Cell abundance $w_{s,f}$ was used to generate integer cell count $count_{s,f}$ and fraction of mRNA captured $fraq_{s,f}$ for each location and cell type as follows:
   a) generate $count_{s,f}$ by rounding $w_{s,f}$ to the smallest integer, such that $count_{s,f} >= w_{s,f} (\lceil w_{s,f} \rceil)$.
   b) Compute the fraction of mRNA captured as $fraq_{s,f} = w_{s,f} / count_{s,f}$.

Second, follow this step-by-step procedure to use 1) the integer cell count $count_{s,f}$ and 2) the fraction of mRNA captured $fraq_{s,f}$ for each cell type f in a given location s to generate expected multi-cell mRNA count profiles $ed_{s,g}$ for every gene g in a given location s by combining cells c drawn from reference cell types f in the snRNA-seq data $j_{c,g}$ as follows:

1. Randomly select indices of cells $c \in f$ that form a subset $p \subset c$ containing $n = count_{s,f}$ cells.
2. Construct per cell type expected mRNA abundance profiles for a given location and cell type: $ed_{s,f,g} = (\sum_{c \in p} j_{c,g}) fraq_{s,f}$
3. Construct multi-cell expected mRNA abundance profiles by adding mRNA across all cell types:

$$ed_{s,g} = \sum_f ed_{s,f,g}$$

Third, follow this step-by-step procedure to generate multi-cell mRNA integer counts d_{sg}. In this step, gene-specific scaling was applied, denoted as m_g, to mimic the difference in sensitivity between technologies and counts were samples from Poisson distribution:

$$d_{s,g} \sim Poisson(ed_{s,g} m_g),$$

where m_g characterises the difference between the mouse brain Visium data and single nucleus RNA-seq reference (Fig S2C, estimated by cell2location). Using these values makes the simulated data representative for mapping single nucleus RNA-seq derived reference cell types.

Under this simulation, the total number of mRNA per location mimics that observed in mouse brain data (Fig S2D).