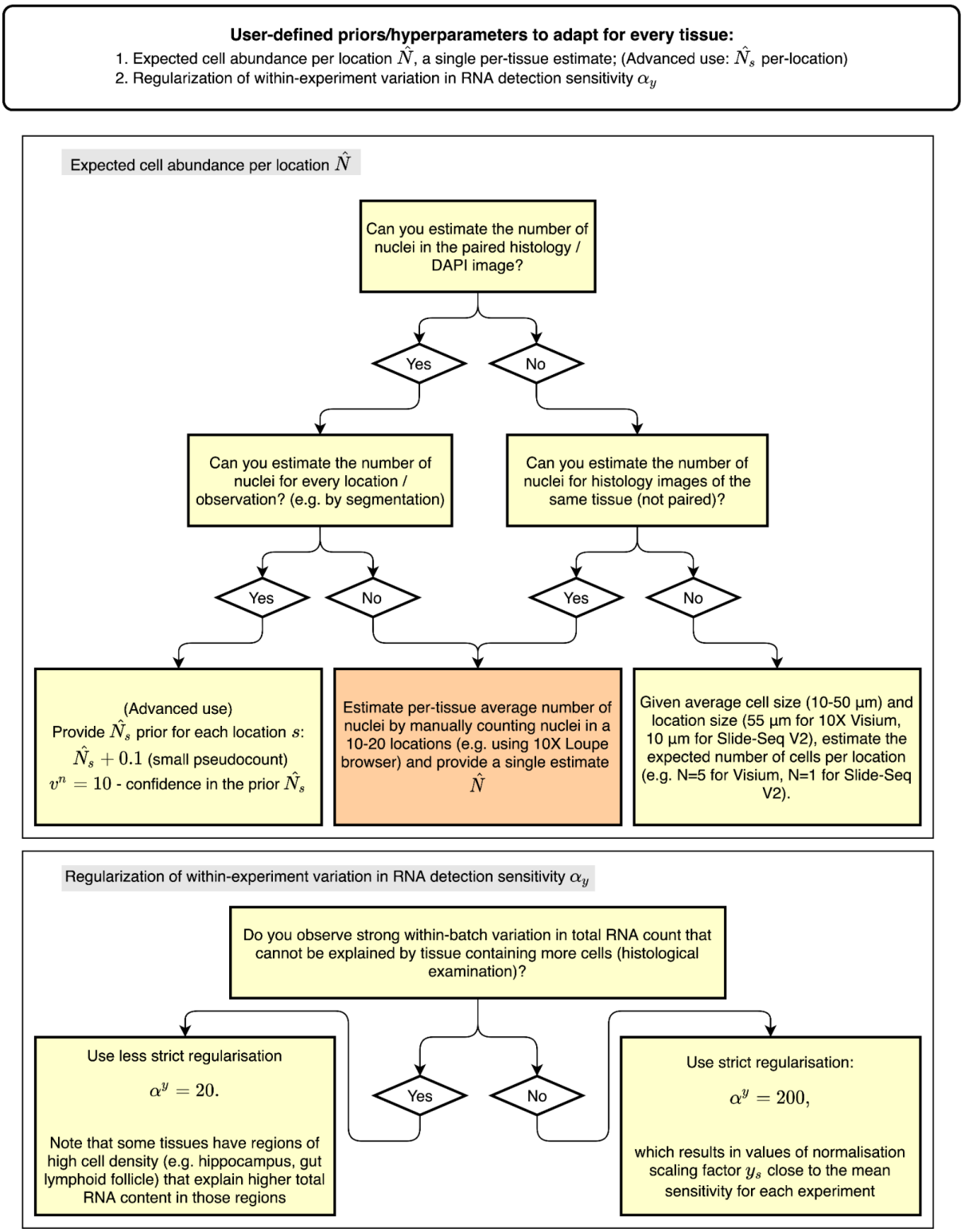
Note on determining model hyperparameters.



Flow diagram to choose cell2location hyperparameters. Cell2location has 2 hyperparameters that can be set by the users (see a separate box for each hyperparameter). The following flowchart can be used to determine these hyperparameters in the light of experimental or biological prior information. An appropriate setting of the expected cell abundance $\hat{N}$ is essential to inform the estimation of

absolute cell type abundance values. Term 'location' refers to 10X Visium spots, Slide-Seq V2 beads, WTA Nanostring regions-of-interest (ROI) and other spatially defined voxels where transcriptomic measurements are taken. The choice of expected cell abundance $\hat{N}$ highlighted in orange is applicable to most 10X Visium users.

Note on determining model hyperparameters. The cell2location model has 2 parameters that should be adjusted by the user, taking known experimental and biological characteristics of a given dataset into consideration:

1) Expected cell abundance per location $\hat{N}$
2) Regularization of within-experiment variation in RNA detection sensitivity α^y

The Fig S27 provides a flowchart to guide the user to determine appropriate settings for these hyper-priors, which was used to determine the model parameters for all presented results.

The expected cell abundance $\hat{N}$ per location is provided as a single tissue-level 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 10-20 locations in the histology image (e.g. using 10X Loupe browser), 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. 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 $\hat{N}$ (Slide-Seq V2, Fig S27). 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.

Hyperparameter α^y for regularising within-experiment variation in RNA detection sensitivity. α^y is the shape parameter of the prior Gamma distribution over y_s that represents the extent to which RNA detection sensitivity for individual locations deviates from the mean sensitivity of each batch. Large values of α^y correspond to small within-experiment variability (Suppl. Methods). By default, we assume small within-experiment variability in RNA detection sensitivity across locations ($\alpha^y = 200$), which results in values of y_s close to the mean sensitivity for each experiment. However, when strong gradients in mRNA detection sensitivity are observed, a common issue in adult human 10X Visium data, we recommend less strict regularisation ($\alpha^y = 20$).

While good choices of these hyper-parameters can have a positive impact on accuracy and sensitivity, the estimates are robust to a range of suboptimal choices. The estimation of absolute cell abundance requires appropriate settings of $\hat{N}$ and μ_m in particular.

Brief description of the cell2location model

For a complete derivation of the cell2location model, please see supplementary methods. Briefly, cell2location is a Bayesian model, which estimates the absolute abundance of cell types at each location by decomposing the spatial expression count matrix into a predefined set of reference cell type signatures. The model takes an untransformed spatial

expression count matrix $d_{s,g}$ of genes $g = \{1, .., G\}$ at locations $s = \{1, .., S\}$as input, as for example obtained from the 10X SpaceRanger software (10X Visium data). Cell2location allows to jointly model multiple spatial data sets $e = \{1, .., E\}$, aka batch integration, where e denote individual experiments obtained using one or multiple technologies. The second input is a matrix of reference cell type signatures $g_{f,g}$, which correspond to the expected mRNA count of genes g for each cell type $f = \{1, .., F\}$. By default, this matrix is estimated using a negative binomial regression model, which allows for combining data across batches and technologies (see below and Suppl. methods).

Cell2location models the elements of the spatial expression count matrix $d_{s,g}$ as Negative Binomial (NB) distributed, given an unobserved gene expression level (rate) $\mu_{s,g}$ and gene- and batch-specific over-dispersion $\alpha_{e,g}$:

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

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

$$\mu_{s,g} = \left(\underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f} g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Here, $w_{s,f}$ denotes regression weight of each reference cell type f at location s, which can be interpreted as the abundance of cells expressing reference cell-type signature f at location s; m_g is a gene-specific scaling parameter, which adjusts for differences in sensitivity between technologies; $s_{e,g}$ captures gene-specific additive shift, which allows the model to account for contaminating RNA; y_s is a location-specific scaling parameter, which models variation in RNA detection sensitivity across locations s across one or multiple batches e (by use of hierarchical priors; Suppl. Methods).

Because of the organization of tissues into zones the cell type abundance estimates $w_{s,f}$ are not independent across locations. In order to borrow statistical strength and leverage information sharing across locations, $w_{s,f}$ is modelled using a hierarchical decomposition prior (factorization), assuming $r = \{1, .., R\}$ groups of cell types that partially share a common cell abundance profile (e.g. due to tissue zones; Suppl. Methods). Unless stated otherwise, R is set to 50.

Approximate Variational Inference is used to estimate all model parameters, implemented in scvi-tools framework[19], building on pyro and pymc3 probabilistic programming frameworks as backend[51,52], which supports GPU acceleration. For full details see Suppl. Methods.

Note on the construction of reference cell type signatures.

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. This can be achieved by generating a paired snRNA-seq reference from the same tissue sample. However, imperfect matching of cell populations is often acceptable.

Based on one or multiple snRNA-seq input datasets, reference cell type signatures (gene expression) are estimated based on a defined set of cell populations, which in turn can be identified using conventional clustering workflows. By default, the cell2location employs a Negative Binomial regression to estimate reference cell type signatures. This model-based approach allows robustly combining data across technologies and batches, which results in improved spatial mapping accuracy. Alternatively, if batch effects are not a concern, a hard coded method that estimates per-cluster average mRNA counts without any model-based strategy can be sufficient. This strategy is simple and fast and yields similar levels of accuracy if batch effects are small. The hard-coded method is also recommended for non-UMI technologies for which the negative binomial count model is often not appropriate, such as Smart-Seq2.