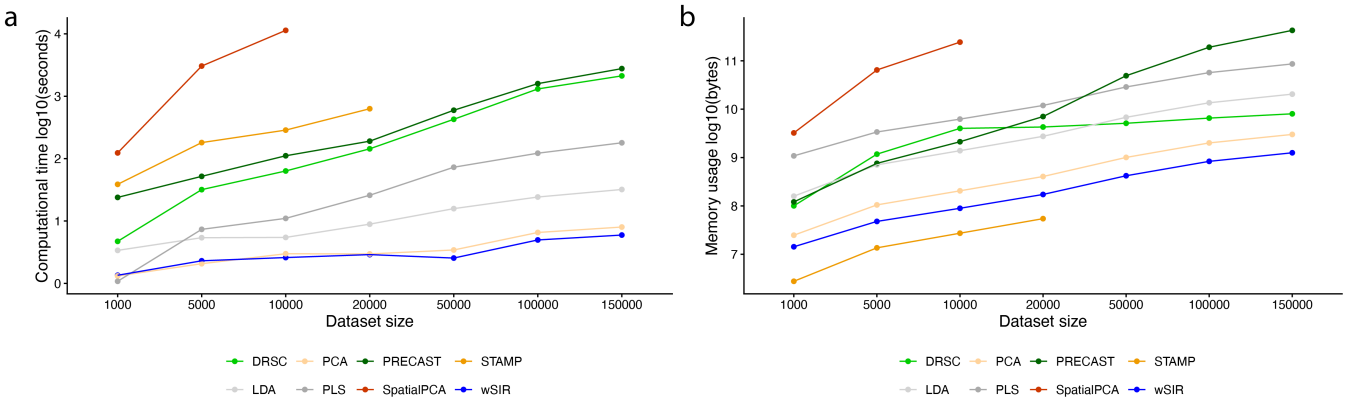
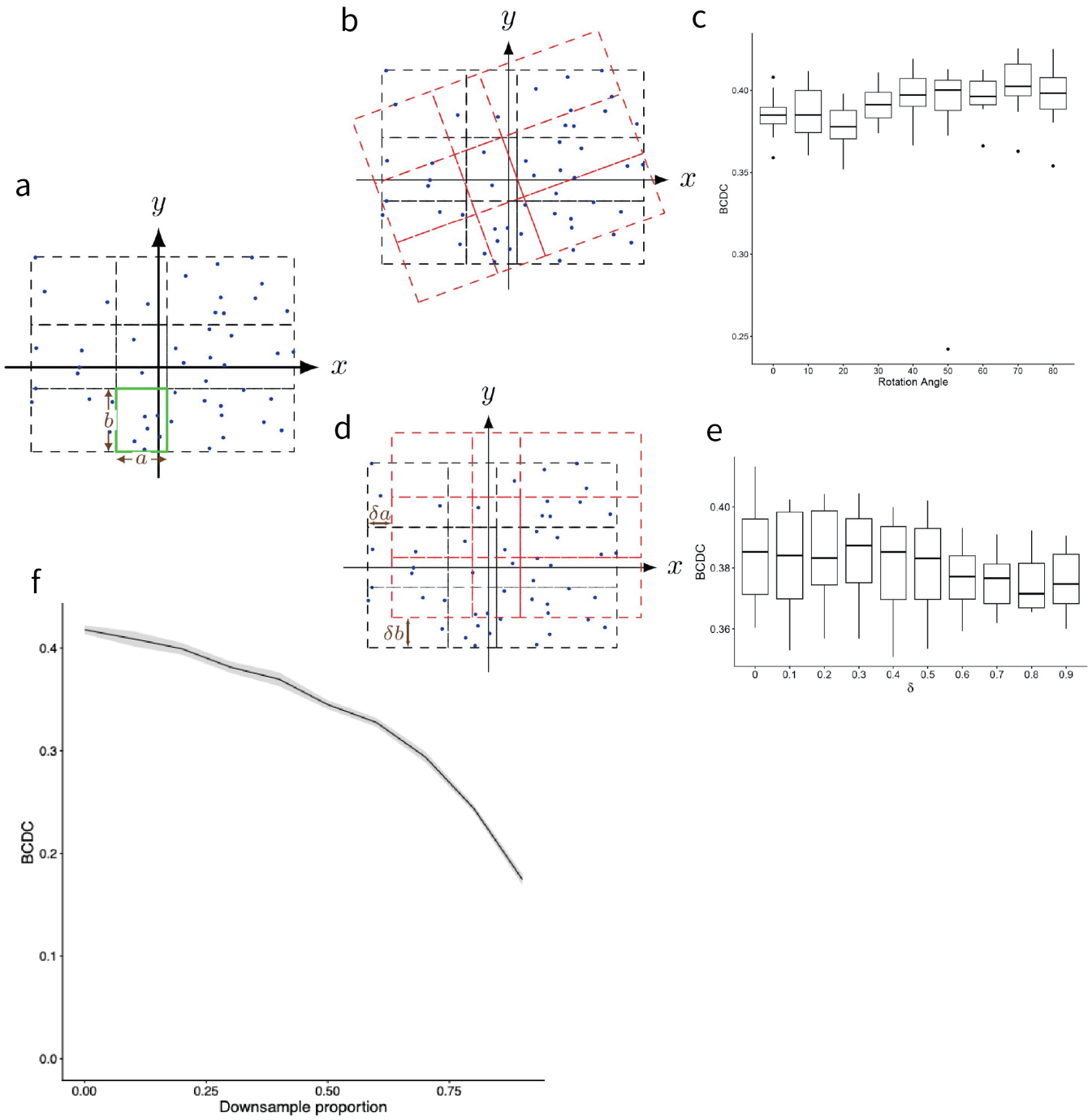
Supplementary Information to "Weighted sliced inverse regression for scalable supervised dimensionality reduction of spatial transcriptomics data"

Maximilian Woollard[1,2,3], Pratibha Panwar[1,2,3], Luke T.G. Harland[4,5,6], Jeremy Mo[7,8,9], Hanyun Zhang[7,8], Alexander Swarbrick[7,8], Berthold Gottgens[4,5], Shila Ghazanfar[*1,2,3], and Linh H. Nghiem[*1,3]

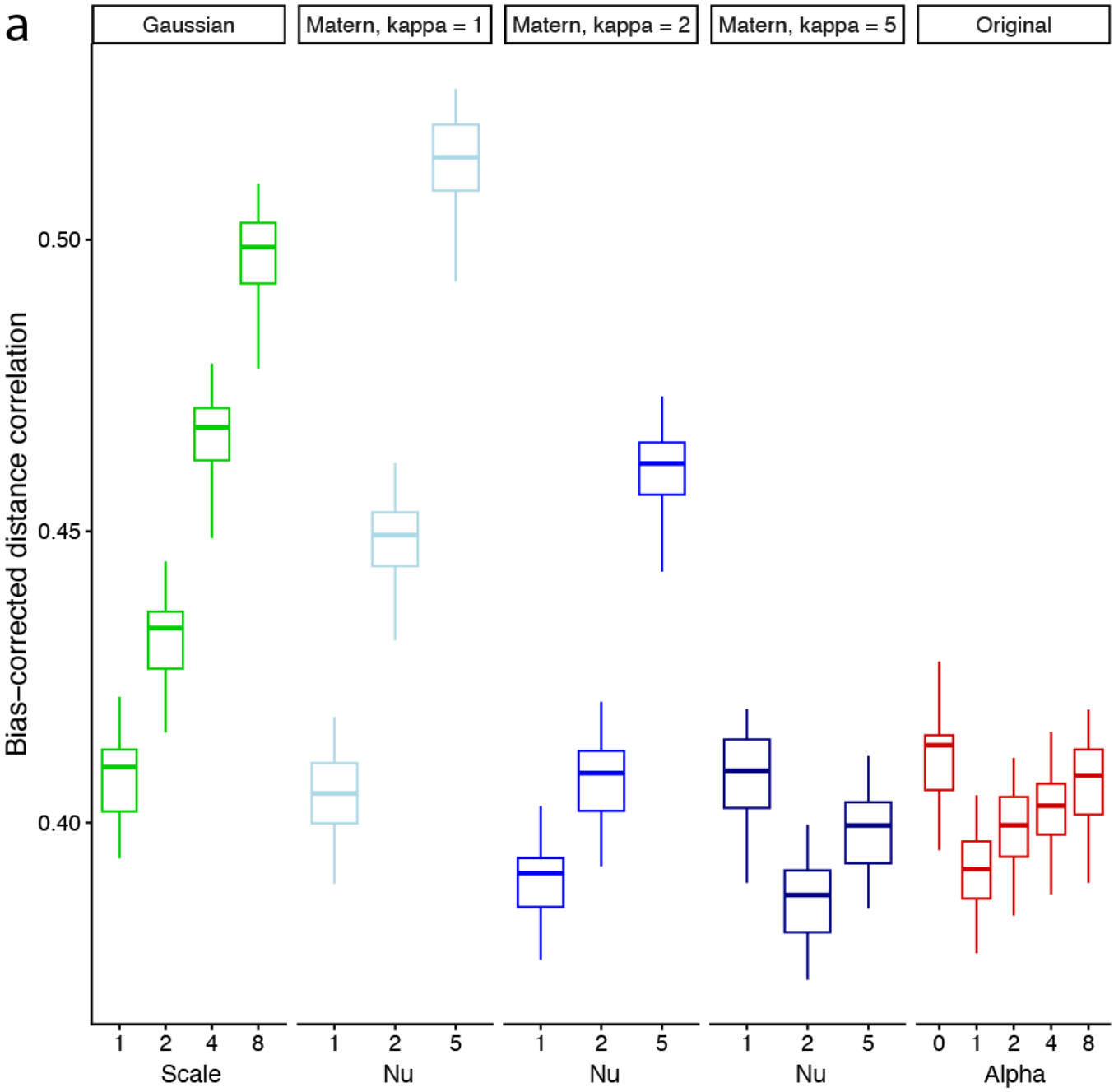
[1]School of Mathematics and Statistics, The University of Sydney, Camperdown, New South Wales, Australia
[2]Charles Perkins Centre, The University of Sydney, Camperdown, New South Wales, Australia.
[3]Sydney Precision Data Science Centre, The University of Sydney, Camperdown, New South Wales, Australia.
[4]Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK.
[5]Department of Haematology, University of Cambridge, Cambridge, UK.
[6]Wolfson College, University of Cambridge, Cambridge, UK.
[7]Cancer Ecosystems Program, Garvan Institute of Medical Research, Darlinghurst, NSW Australia.
[8]School of Clinical Medicine, Faculty of Medicine & Health, UNSW Sydney, Sydney, NSW Australia.
[9]Crown Princess Mary Cancer Centre, Westmead Hospital, Sydney, NSW, Australia.
[*]Co-corresponding authors: Shila Ghazanfar (*shila.ghazanfar@sydney.edu.au*) and Linh H. Nghiem (*linh.nghiem@sydney.edu.au*).



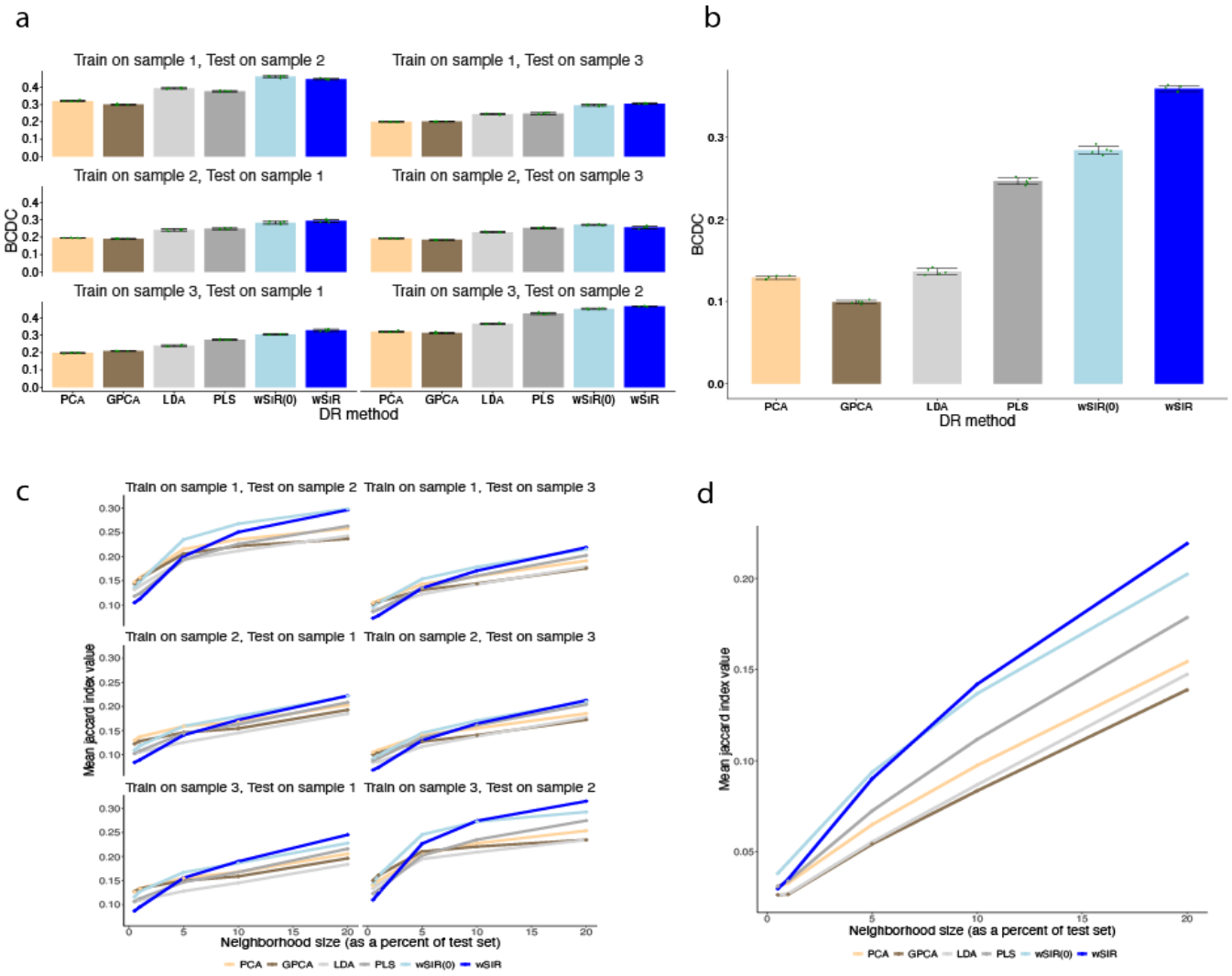
Supplementary Figure 1. Comparison of computational scalability. **a**, Benchmarking of computational time (log-scaled y-axis) of methods with varying dataset sizes (x-axis). SpatialPCA and STAMP failed to run within 1 hour for datasets above 10,000 and 20,000 samples respectively. **b**, As in panel a for memory usage in bytes (log-scaled). Source data are provided as a Source Data file.



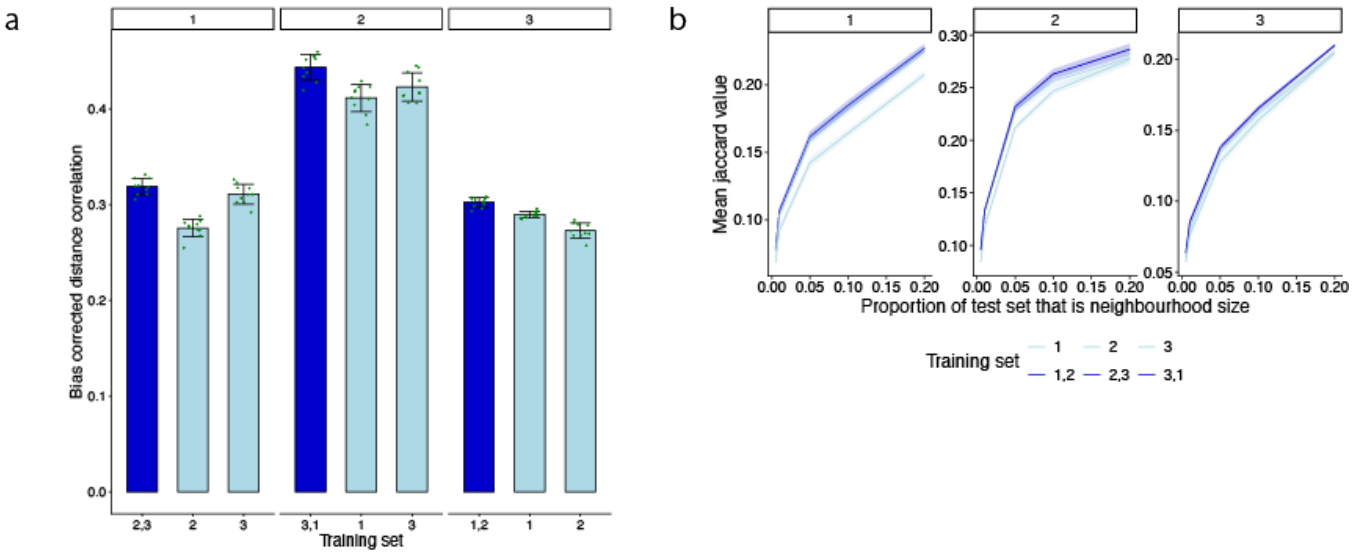
Supplementary Figure 2. Sensitivity analysis of wSIR to slicing mechanisms. **a**, wSIR's default slicing mechanism is constructed on a rectangular grid of 2-D coordinate systems, **b**, a slicing mechanism is formed by rotating the grid in panel a by a certain degree, **c**, boxplots displaying bias-corrected distance correlation (BCDC) between wSIR's loadings and original spatial coordinates for 10 subsamples of the Xenium Breast Cancer dataset as a function of the rotation angle between the original grid and the grid in panel b. **d**, a slicing mechanism is formed by shifting the grid in panel a to the right by δa and upward by δb, where a and b are width and the height of the rectangle with smallest area in the grid of panel a, **e**, same as panel c but for different values of δ for the shift in panel d. **f**, Line plot displaying BCDC between wSIR's loadings and original spatial coordinates for 10 subsamples of mouse embryo dataset when the counts are downsampled by a certain proportion (x axis) to simulate sparsity. Error band represents standard deviation over the 10 subsamples. Source data are provided as a Source Data file.



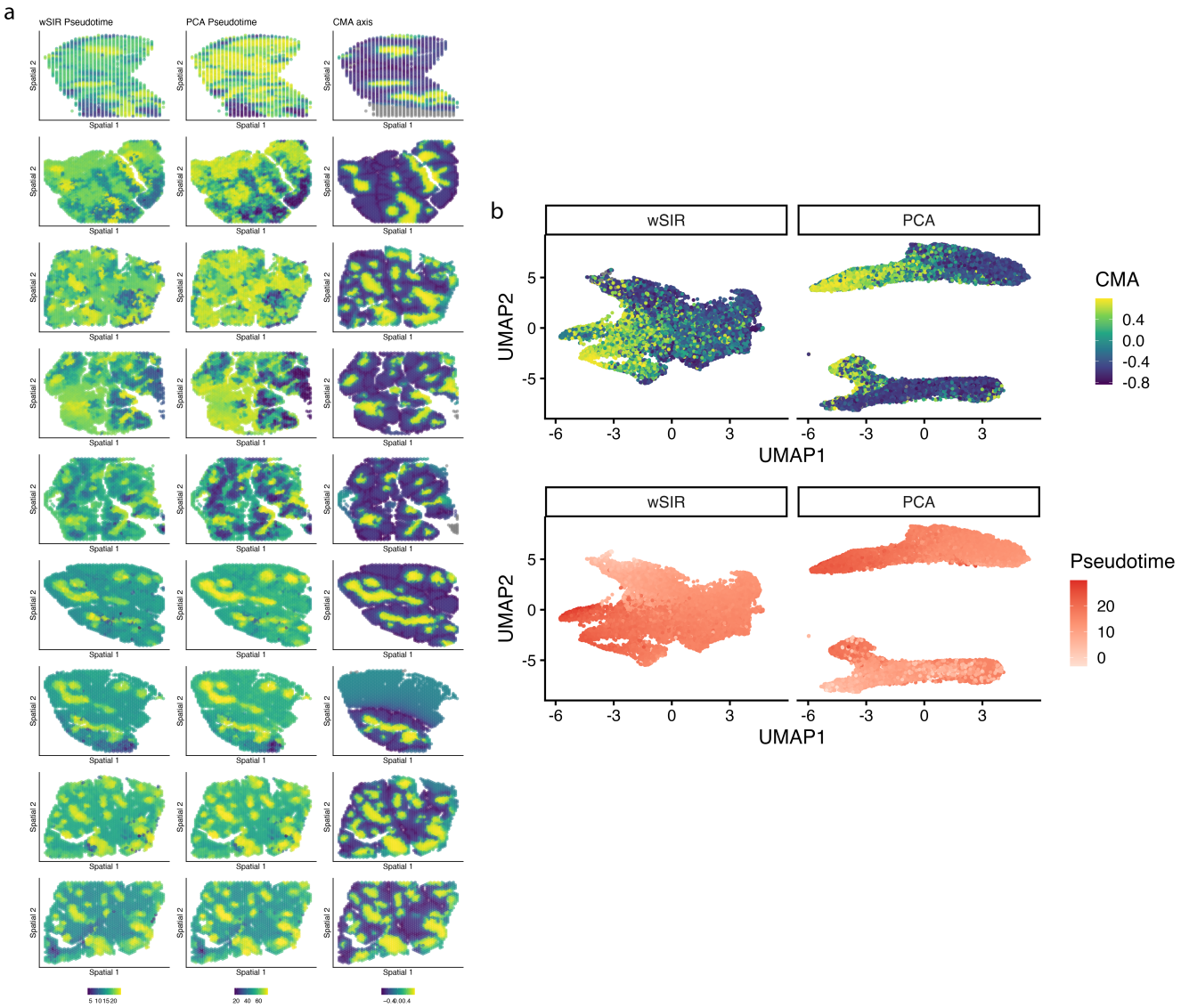
Supplementary Figure 3. Sensitivity of wSIR to the spatial weight-matrix specification. Boxplots show the test-set BCDC performance metric across the 10 folds of cross-validation for the first mouse embryo dataset under different spatial dependence structures for the entries (w_{ij}) of the weight matrix ($\mathbf{W}$). The specifications considered are: a Gaussian kernel, ($w_{ij} = \exp -d_{ij}^2/(2\nu^2)$), with scale parameter $\nu \in 1, 2, 4, 8$; a Matérn kernel with unit variance, scale parameter $\kappa \in 1, 2, 5$, and smoothness parameter $\nu \in 1, 2, 5$; and the default weighting scheme used by wSIR, $w_{ij} = (1 - d_{ij}/\max_{i,j} d_{ij})^{1/\alpha}$), with $\alpha \in 0, 1, 2, 4, 8$ (note $\alpha = 0$ corresponds to regular SIR). All these analyses used $H = 64$ tiles. Source data are provided as a Source Data file.



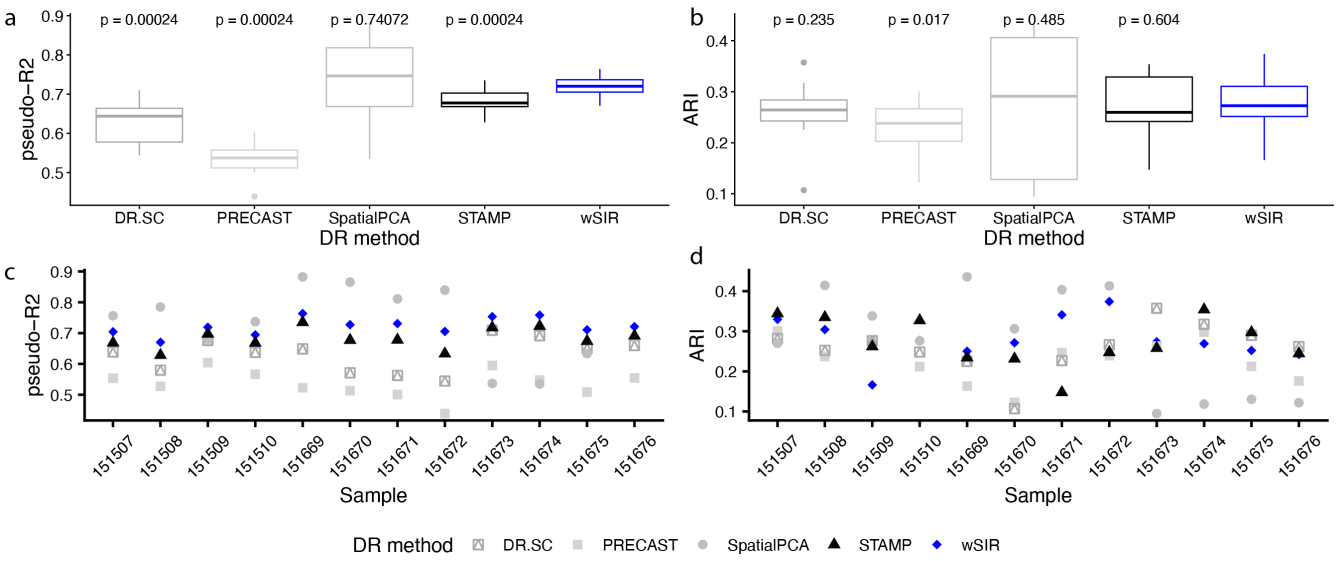
Supplementary Figure 4. Performance of non-spatial dimension reduction methods. **a**, Barplots displaying the BCDC of different dimension reduction (DR) methods for a leave-one-dataset-out evaluation of the three samples of the mouse embryo dataset. wSIR(0) indicates $\alpha = 0$, equivalent to SIR. Error bars indicate standard deviation of BCDC values resulting from 10 bootstrap samples of the training set. **b**, Similar to panel a for Xenium breast cancer dataset. **c**,. Line plots indicating the ability to preserve local neighbourhoods for each method using the same evaluation strategy as panel a. Error bands are computed as the standard deviation over the bootstrap samples like in panel a. **d**, Similar to panel b for Xenium breast cancer dataset. Source data are provided as a Source Data file.



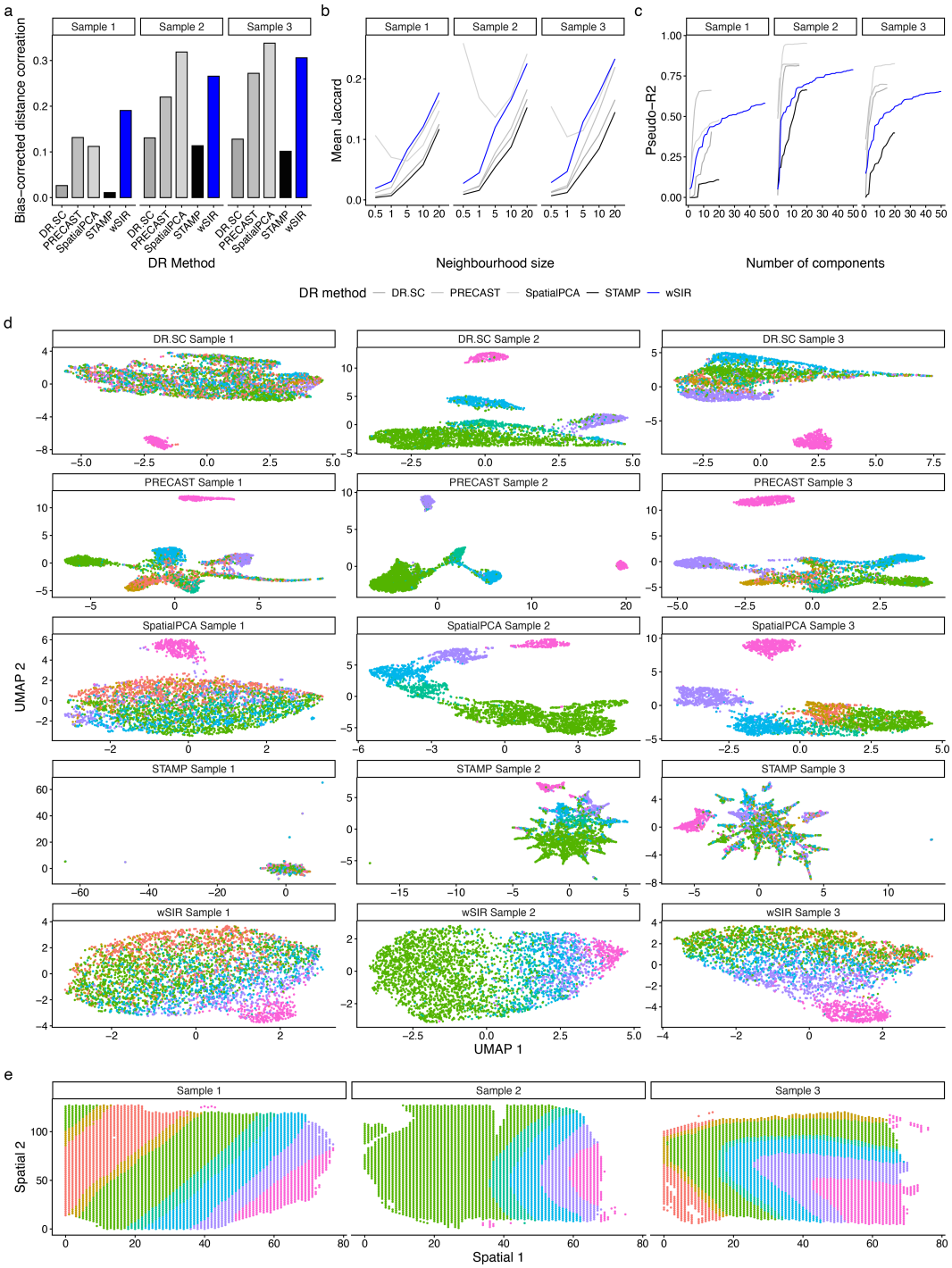
Supplementary Figure 5. Effect of multiple samples on wSIR performance. **a**, BCDC performance metric of wSIR with 50 dimensions when trained on single or multiple samples (x-axis labels) of the mouse embryo dataset. Error bars indicate standard deviation of BCDC values resulting from 10 bootstrap samples of the training set. **b**, Mean Jaccard index value for neighbourhood similarity between wSIR embeddings and spatial coordinates. The label on each panel denotes the sample that is used as the test set, while the line colour indicates the sample(s) used as the training set, darker blue indicates training with two samples. Error band represents standard deviation over the 10 bootstrap samples (. Source data are provided as a Source Data file.



Supplementary Figure 6. Trajectory analysis with wSIR. **a**, Estimated pseudotimes from wSIR and PCA embeddings compared to true CMA axis coloured on spatial plots. Results shown for 9 samples from human thymus data. **b**, UMAP plots on wSIR and PCA embeddings coloured by estimated pseudotime and CMA values on the combined 9 samples. Source data are provided as a Source Data file.



Supplementary Figure 7. Spatial domain detection on DLPFC dataset. **a**, Boxplot of pseudo-R^2 values for spatial domain detection methods across 12 DLPFC samples. P values are from a paired one-sided wilcoxon sign-rank test between each method's metric values and wSIR. **b**, As in panel a for ARI values, p values computed as in panel a. **c**, Dotplot of pseudo-R^2 values arranged by each of the 12 DPLFC samples. **d**, As in panel c for ARI values. Source data are provided as a Source Data file.



Supplementary Figure 8. Comparison on simulated data. **a**, Barplots displaying bias-corrected distance correlation for three samples (facets) from simulated dataset, comparing wSIR against other spatial dimension reduction methods. **b**, Line plots displaying neighbourhood similarity for each sample over a range of neighbourhood sizes (x-axis). **c**, Line plots displaying pseudo-R^2 metric for each sample over different numbers of dimensions for low-dimensional embedding (x-axis). **d**, UMAP visualisation of wSIR and other spatial dimension reduction embeddings, DR.SC, PRECAST, SpatialPCA, and STAMP (rows) for each sample (columns), coloured by true spatial domain. **e**, Plots of spatial coordinates of each sample coloured by true spatial domain. Source data are provided as a Source Data file.