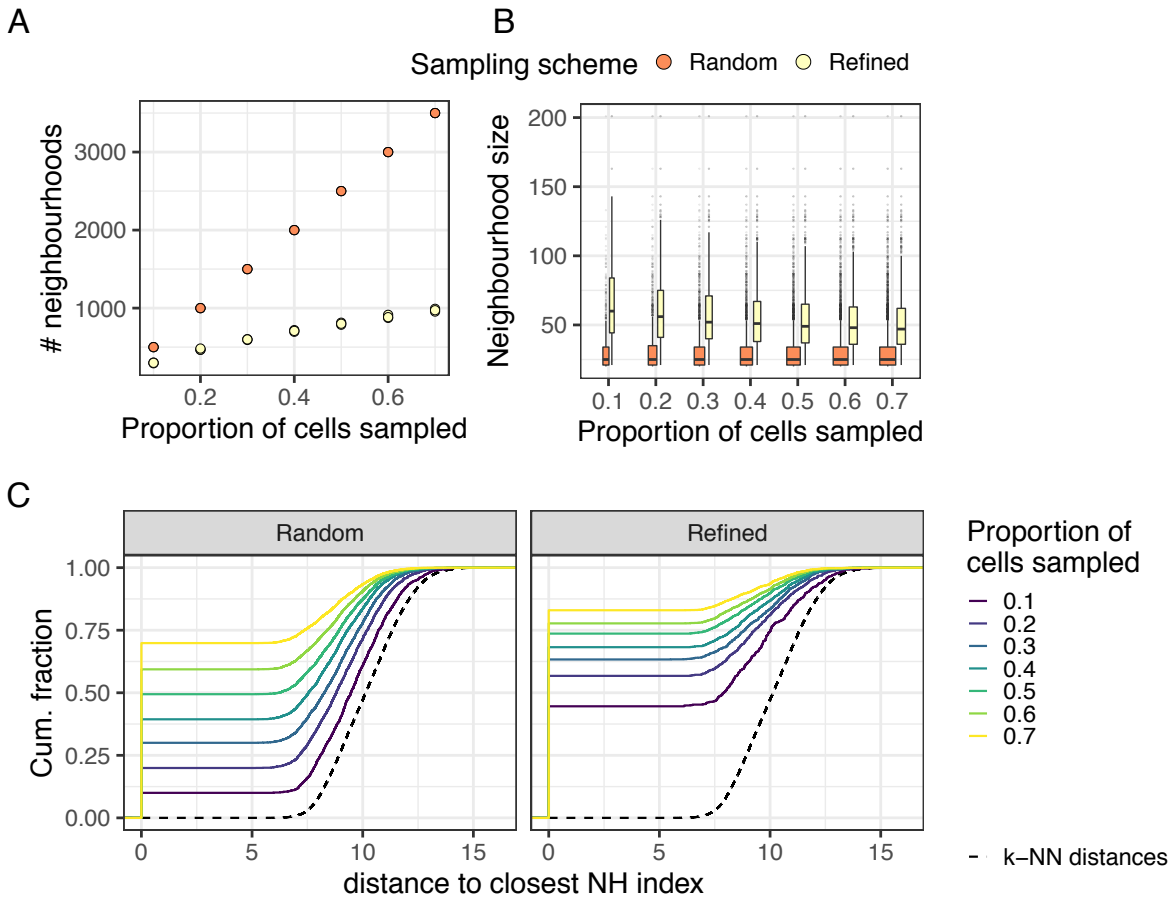
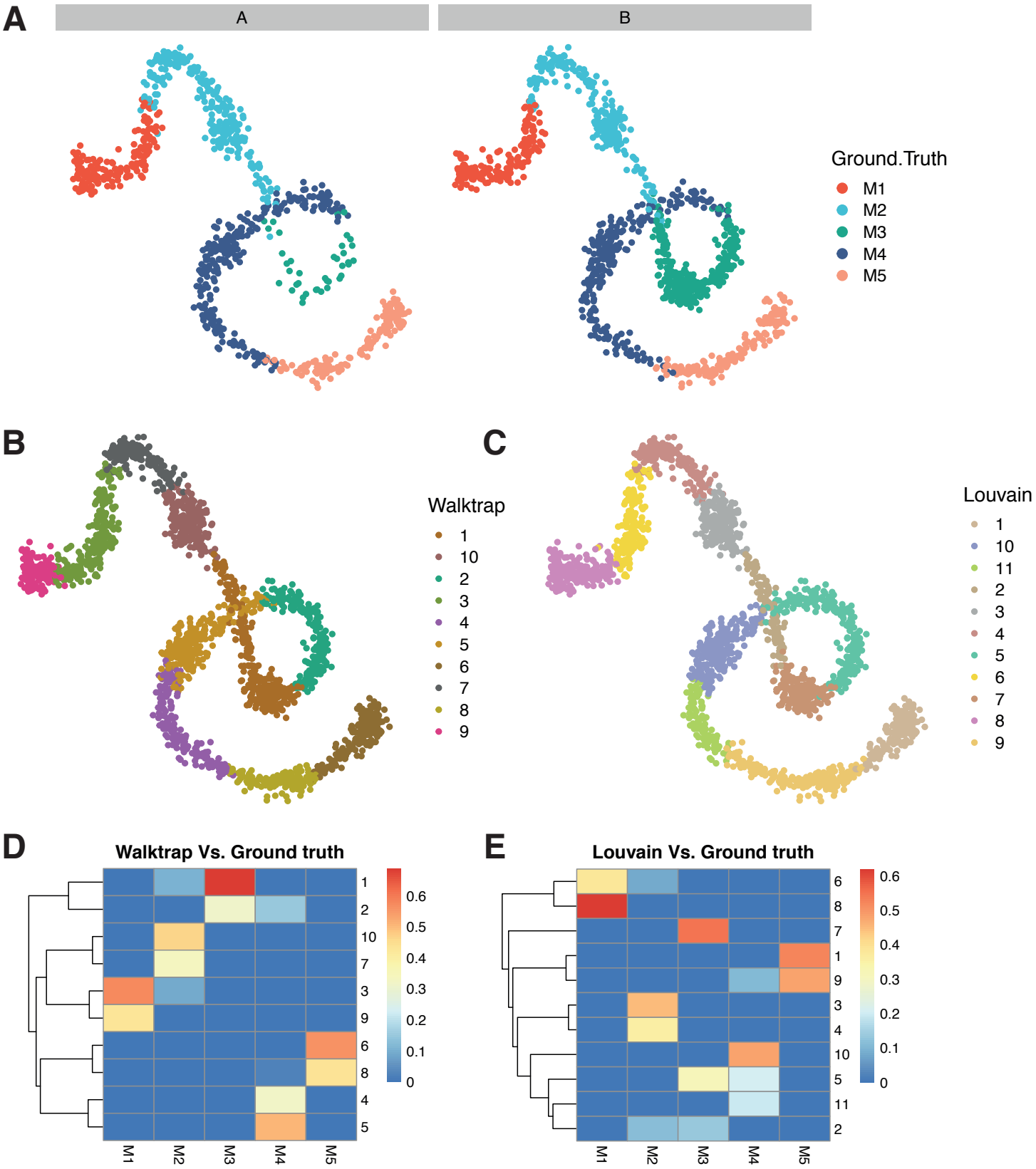
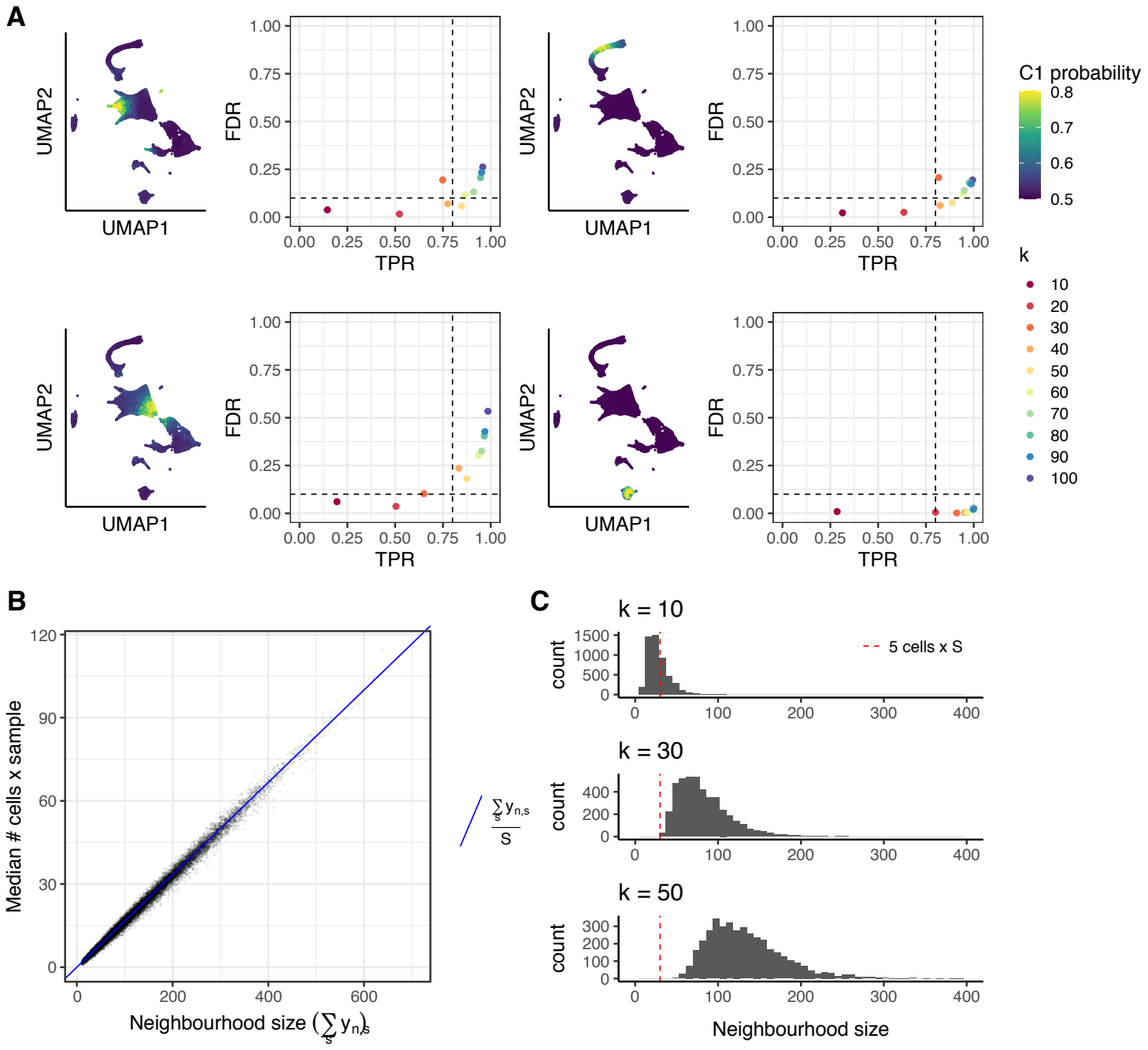


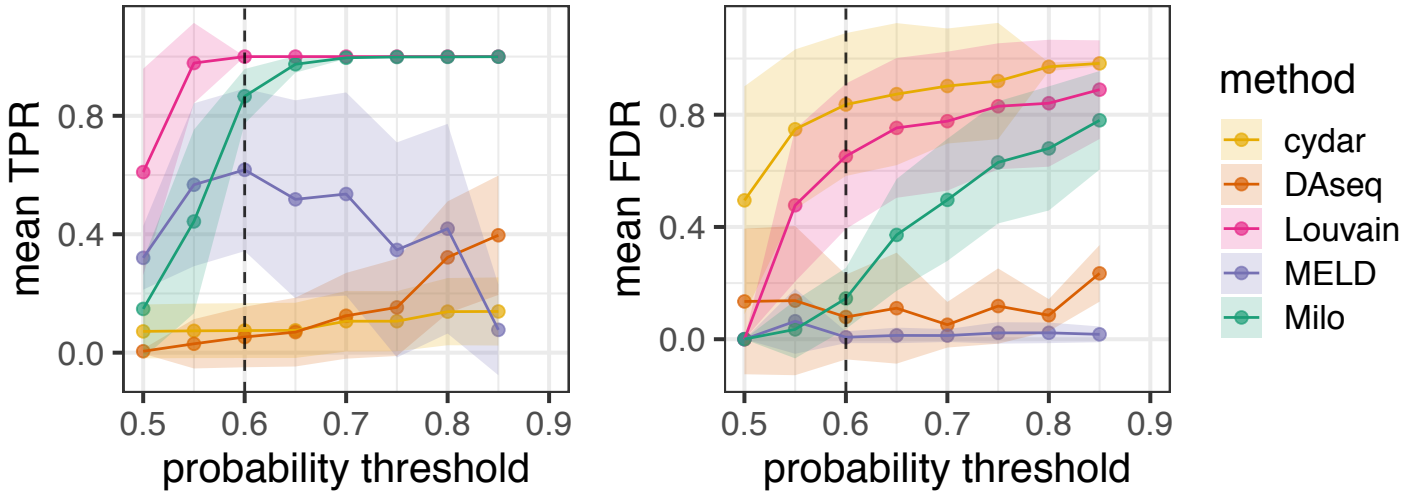
Supplementary Figure 1: **Random sampling of k-NN graph vertices is suboptimal compared to sampling with refinement.** (A) Sampling with refinement leads to selection of fewer neighbourhoods (B) Sampling with refinement leads to selection of bigger neighbourhoods for DA testing, independently of the initial proportion of cells sampled (C) Sampling with refinement generates robust neighbourhoods across initializations: for each index cell we calculate the distance from the closest index in a sampling with different initialization. The cumulative distribution of distances to the closest index is shown. The black dotted line denotes the distribution of distances between k nearest neighbors in the dataset (k=30) (NH: neighbourhood). Neighbourhood statistics were calculated using a simulated trajectory dataset of 5000 cells. All plots show results from three sampling initializations for each proportion.



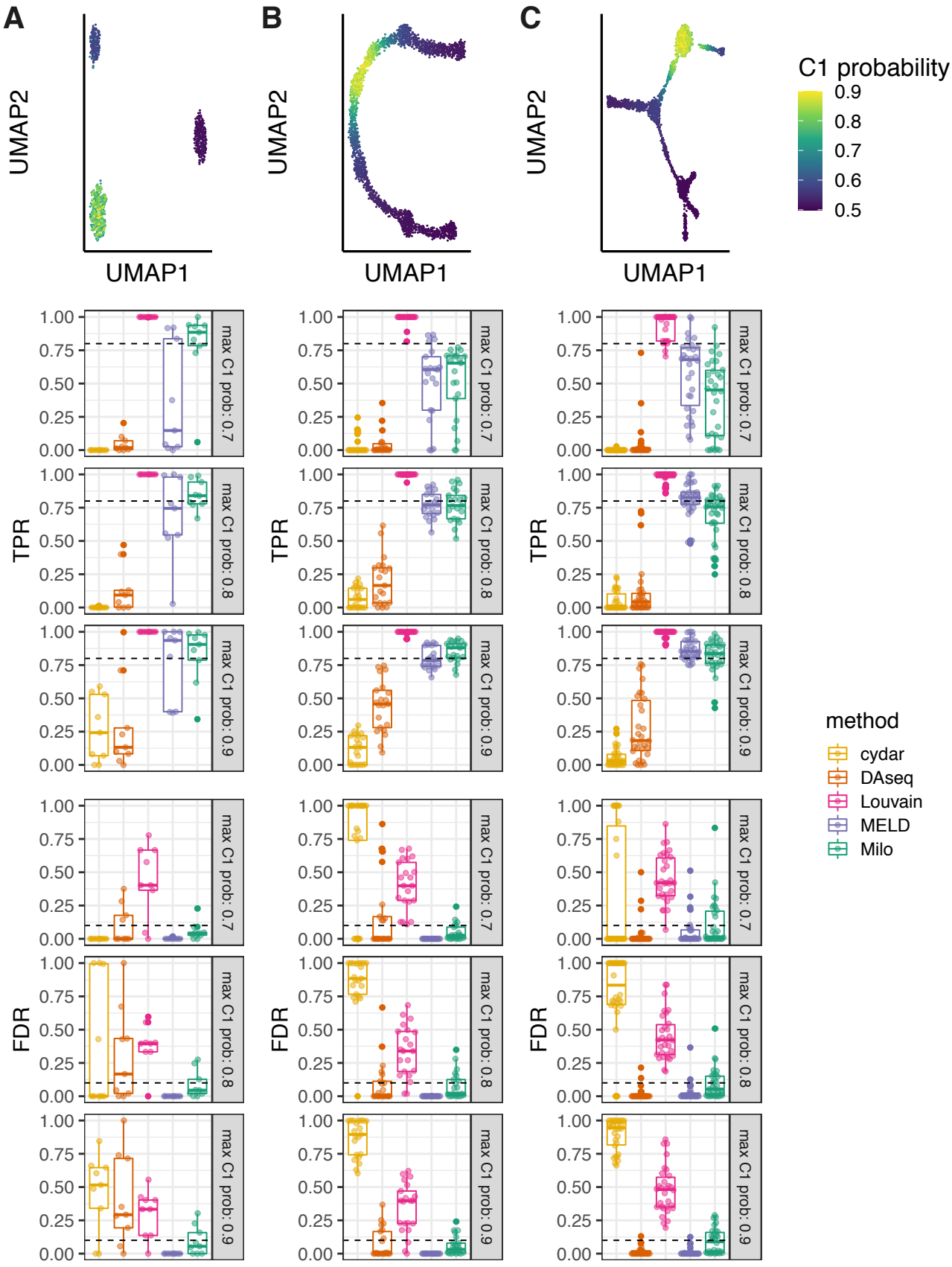
Supplementary Figure 2: **Graph-clustering does not faithfully capture simulated groups and differentially abundant subpopulations in a simulated continuous trajectory.** (A) A simulated linear trajectory of 2000 single-cells generated from 5 different groups, with cells assigned to either condition 'A' (left) or condition 'B' (right). (B) A Walktrap clustering of the data in (A) using the same k-NN graph. Cells are coloured by Walktrap cluster identity. (C) A Louvain clustering of the data in (A) using the same k-NN graph. Cells are coloured by the Louvain clustering identity. (D-E) Heatmaps comparing the numbers of cells in each cluster with respect to the ground truth groups in (A). Each cell is coloured by the proportion of cells from the column groups (ground truth) that are assigned to the respective cluster.



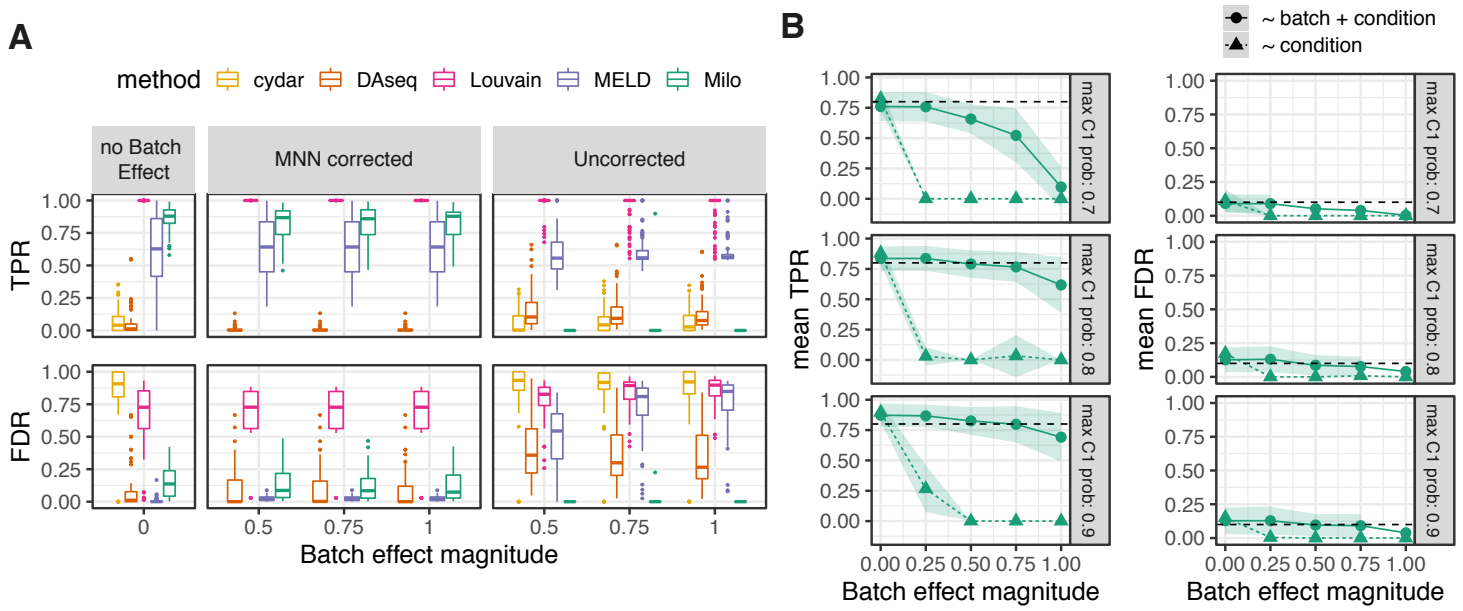
Supplementary Figure 3: **Selection of K parameter** (A) Illustrative examples of TPR-FDR tradeoff for increasing values of k used for k-NN graph building on simulated DA on 4 regions. Dotted lines highlight TPR=0.8 and FDR=0.1 thresholds. (B) The median number of cells per experimental sample is a function of the neighbourhood size snsdivided by the total number of samples (S). (C) Histogram of neighbourhood sizes for different choices of k. The red dotted line denotes the minimum neighbourhood size to obtain 5 cells per sample on average.



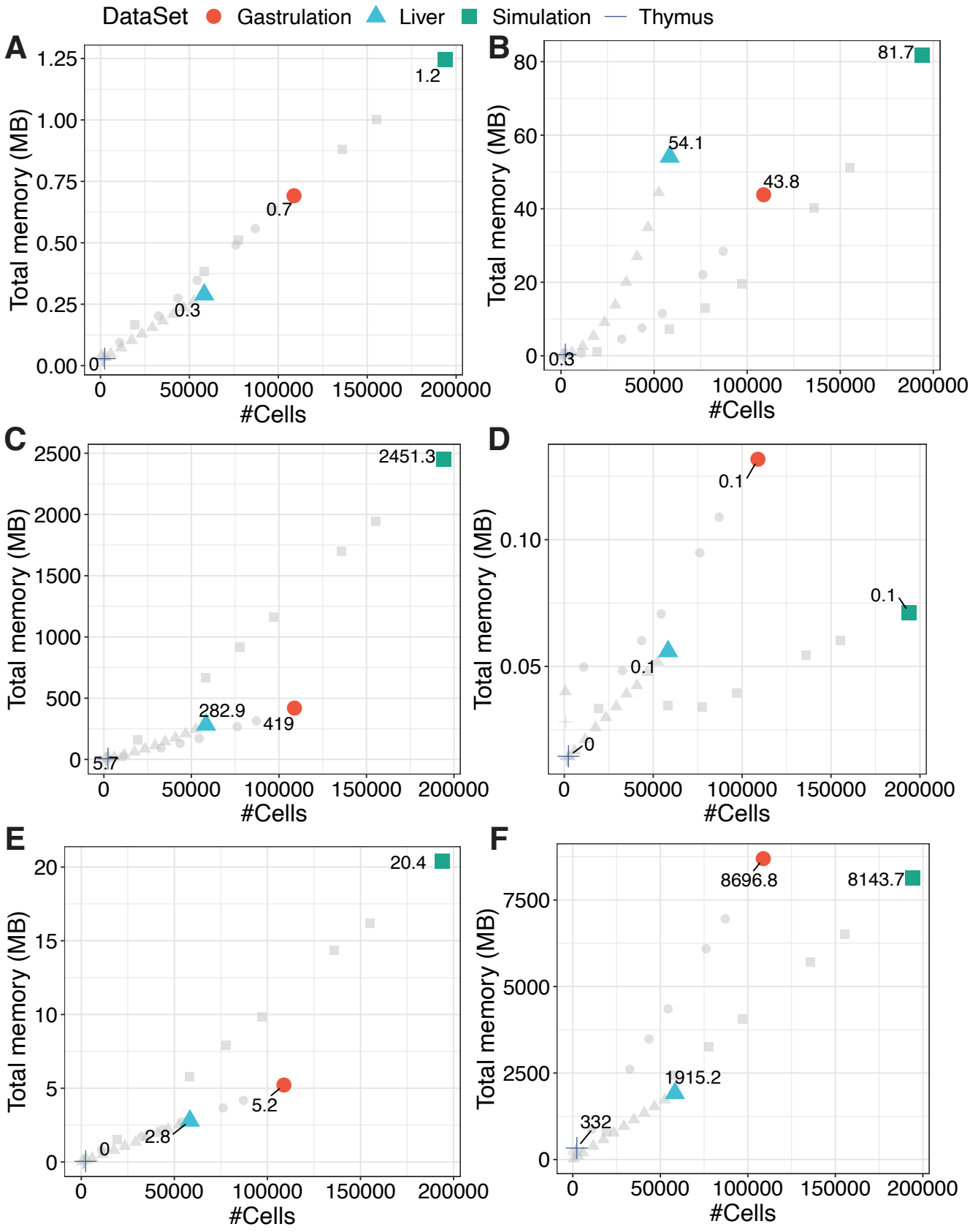
Supplementary Figure 4: **Selection of probability threshold for DA benchmarking.** Mean True Positive Rate (left) and False Discovery Rate (right) for recovery of cells in simulated DA regions as a function of probability threshold t picked to define true DA. The dashed line indicates $t = 0.6$, that was selected for benchmarking analyses. The mean is calculated over 81 simulations on 9 populations. Line shading indicates the standard deviation of the mean.



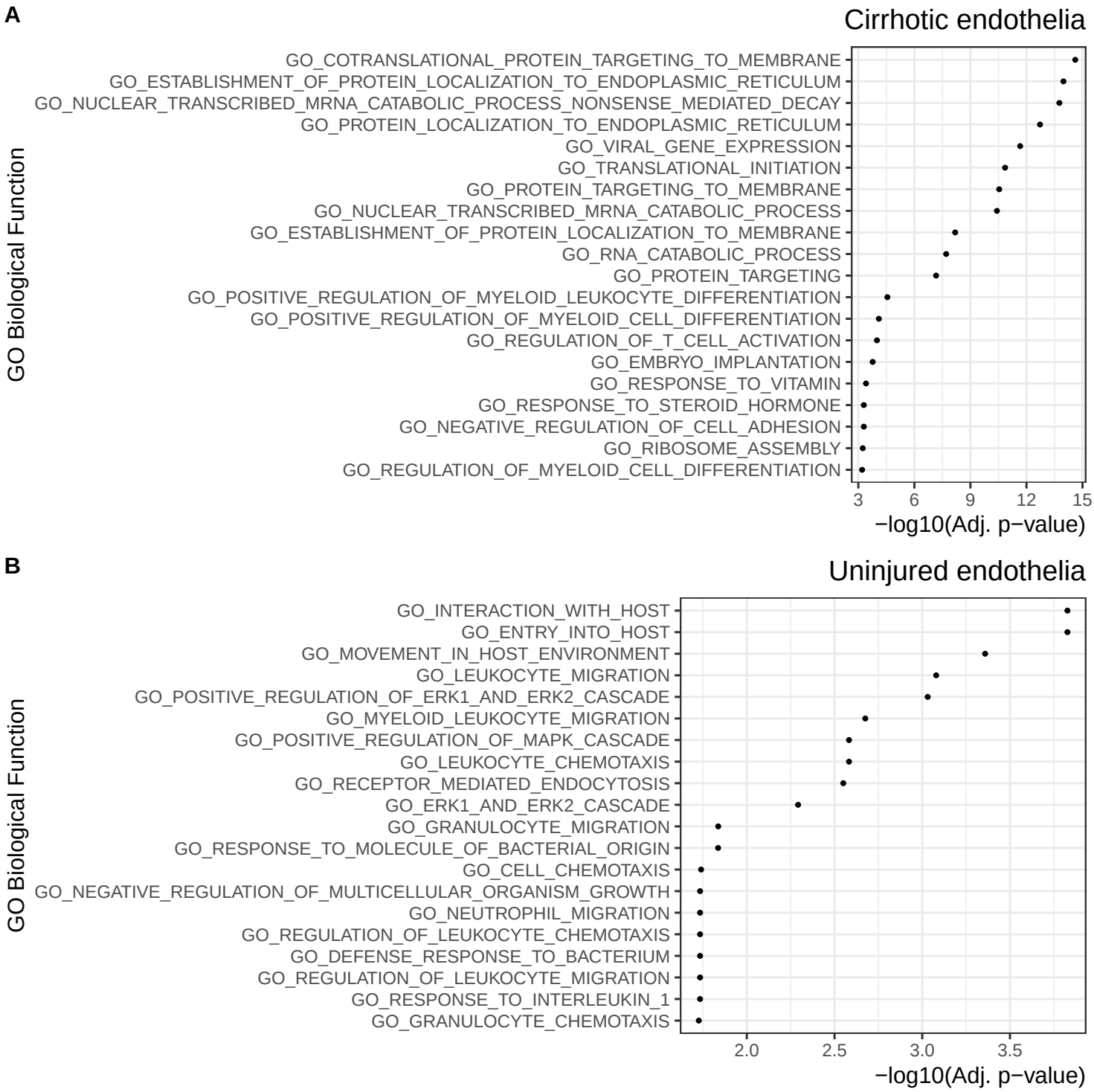
Supplementary Figure 5: **Benchmarking DA methods on simulated data.** DA analysis performance on KNN graphs from simulated datasets of different topologies: (A) discrete clusters (900 cells, 3 populations); (B) 1-D linear trajectory (2500 cells, 7 populations); (C) Branching trajectory (2500 cells, 10 populations).



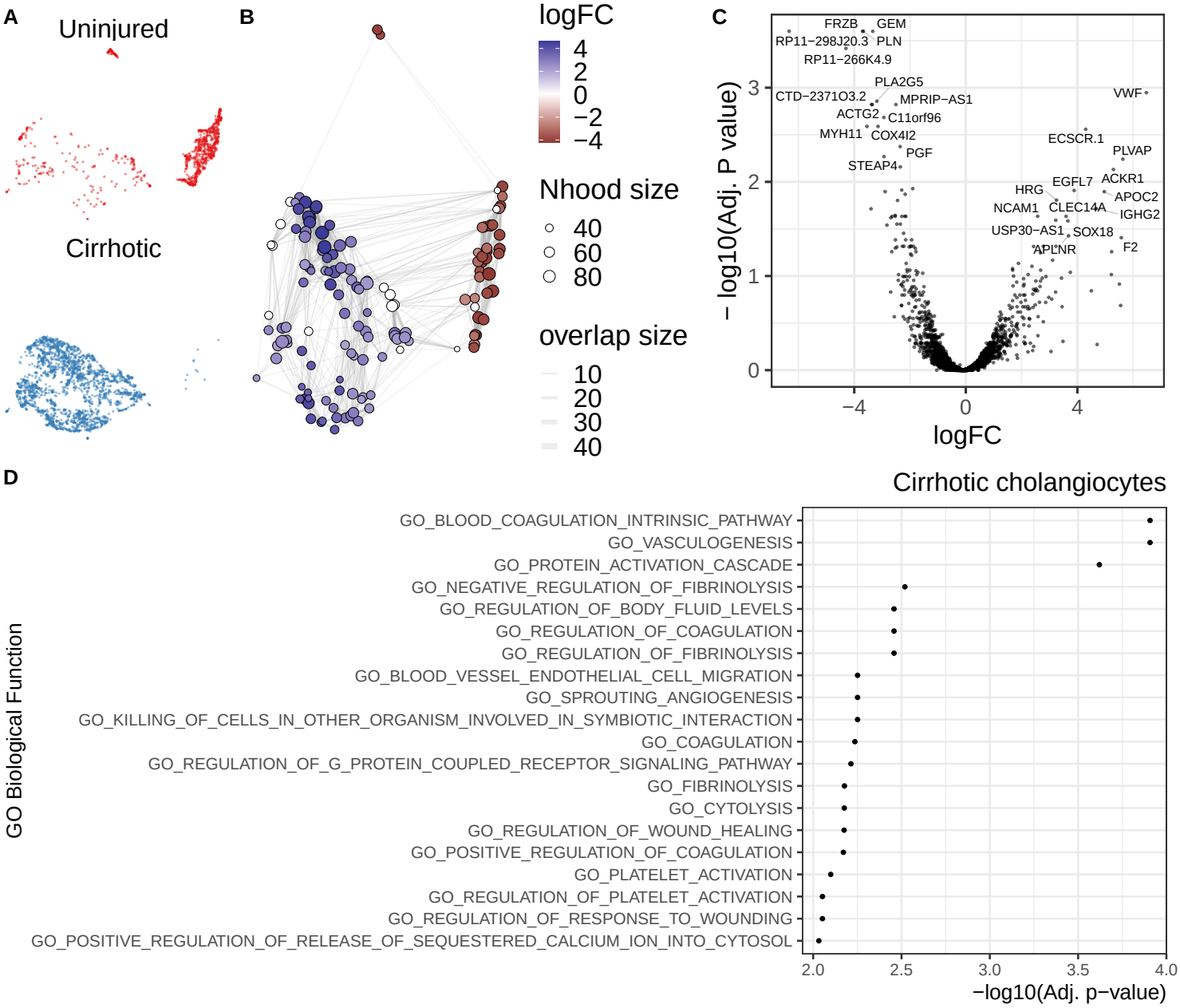
Supplementary Figure 6: **Tolerance of DA analysis to batch effects.** (A) Comparison of performance of DA methods with no batch effect, with batch effects of increasing magnitude corrected with MNN, and uncorrected batch effects. (B) Comparison of Milo performance with (~ batch + condition) or without (~ condition) accounting for the simulated batch in the GLM. Points denote the Mean True Positive Rate (left) and False Discovery Rate (right) for recovery of cells in simulated DA regions with simulated batch effects of increasing magnitude. The mean is calculated over 27 simulations on 9 populations. Line shading indicates the standard deviation of the mean. Each panel represents a different DA effect size.



Supplementary Figure 7: **Memory usage across the Milo analysis workflow** otal memory usage across the steps of the Milo analysis workflow in 4 datasets containing different numbers of cells (Gastrulation: circles, Liver: triangles, Thymus: crosses, Simulation: squares). Grey pointed denote down-sampled datasets of the corresponding type. Coloured points denote the total number of cells for the respective dataset. Total memory usage (y-axis) is shown in megabytes (MB). (A) k-NN graph building, (B) neighbourhood sampling and construction, (C) within-neighbourhood distance calculation, (D) cell counting in neighbourhoods according to the input experimental design, (E) differential abundance testing, (F) total in memory R object size. A fixed value was used in all datasets for graph building and neighbourhood construction (k=30).



Supplementary Figure 8: **Downstream analysis of disease-specific endothelial subpopulations in liver cirrhosis** (A) GO term enrichment analysis on marker genes of cirrhosis-enriched endothelia. (B) GO term enrichment analysis on marker genes of healthy-enriched endothelia. The top 20 significant terms are shown.



Supplementary Figure 9: **Downstream analysis of disease-specific cholangiocyte subpopulations in liver cirrhosis** Downstream analysis of disease-specific cholangiocyte subpopulations in liver cirrhosis (A-B) UMAP embedding (A) and graph representation (B) of neighbourhoods of 3369 cells from cholangiocyte lineage. (C) Volcano plot for DGE test on cholangiocytes DA subpopulations: the x-axis shows the log-fold change between expression in cirrhotic and healthy cholangiocytes. The y-axis shows the adjusted p-value. (D) GO term enrichment analysis on marker genes of cirrhosis-enriched cholangiocytes. The top 20 significant terms are shown.