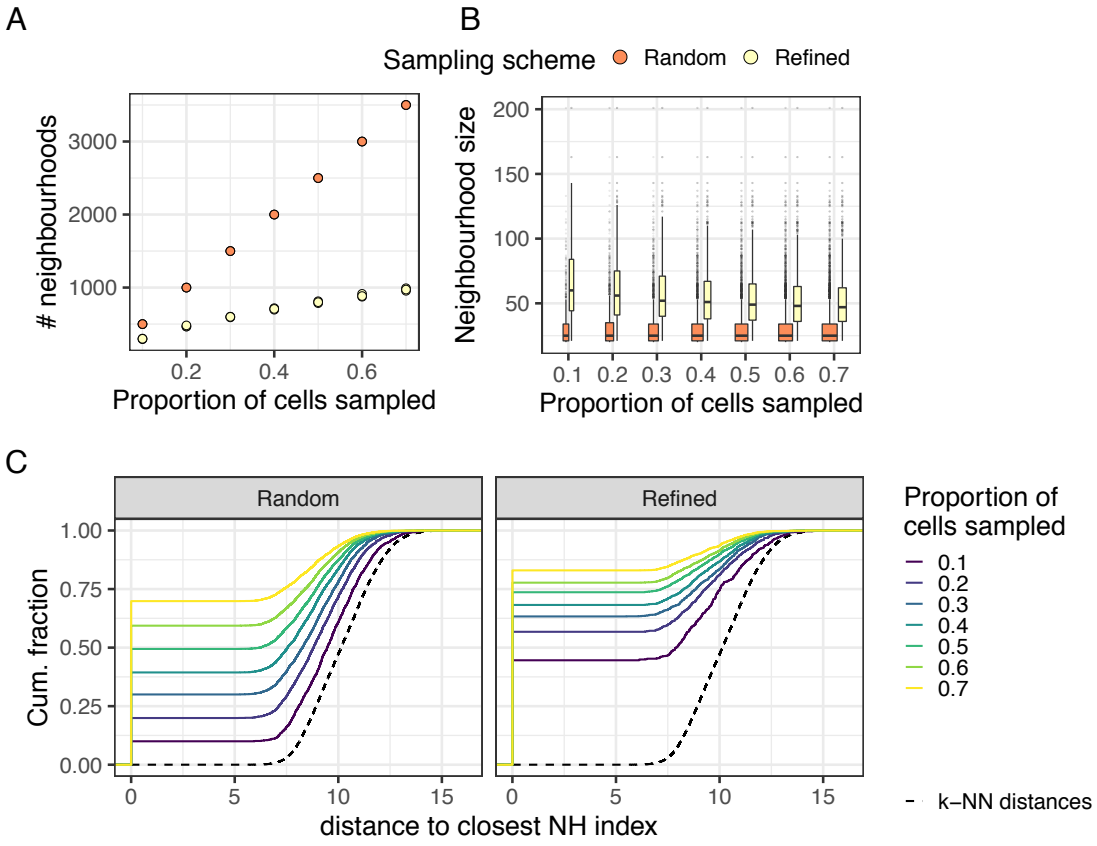
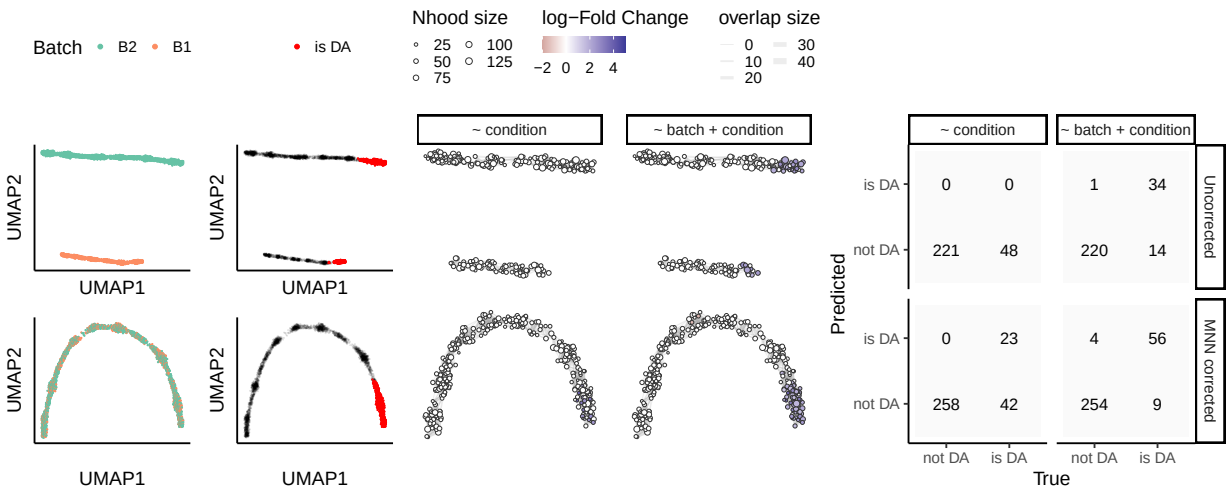
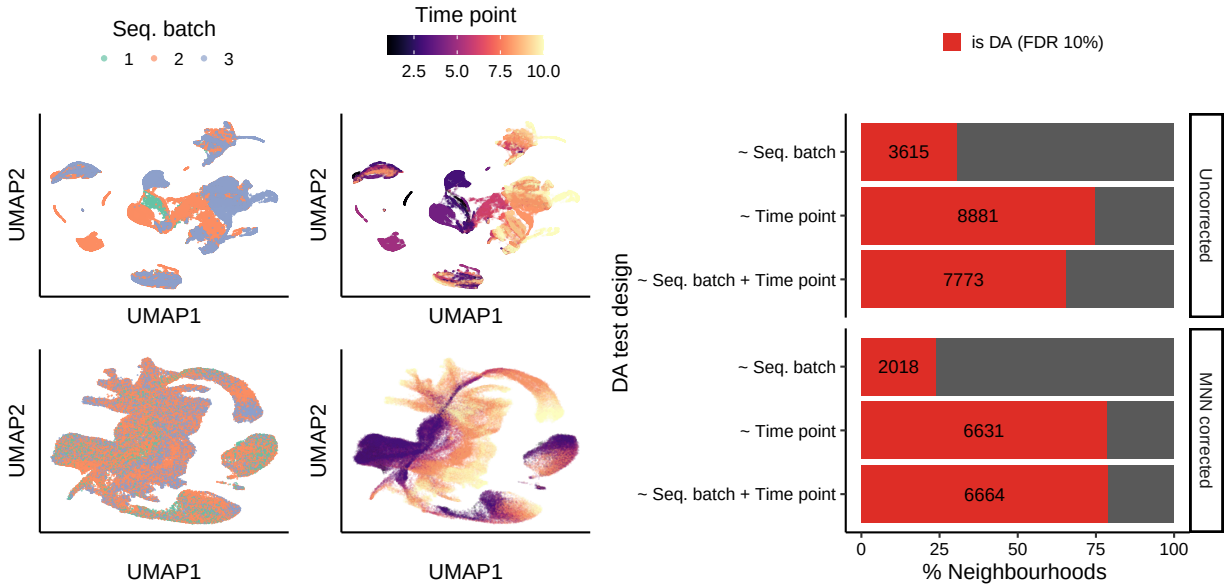


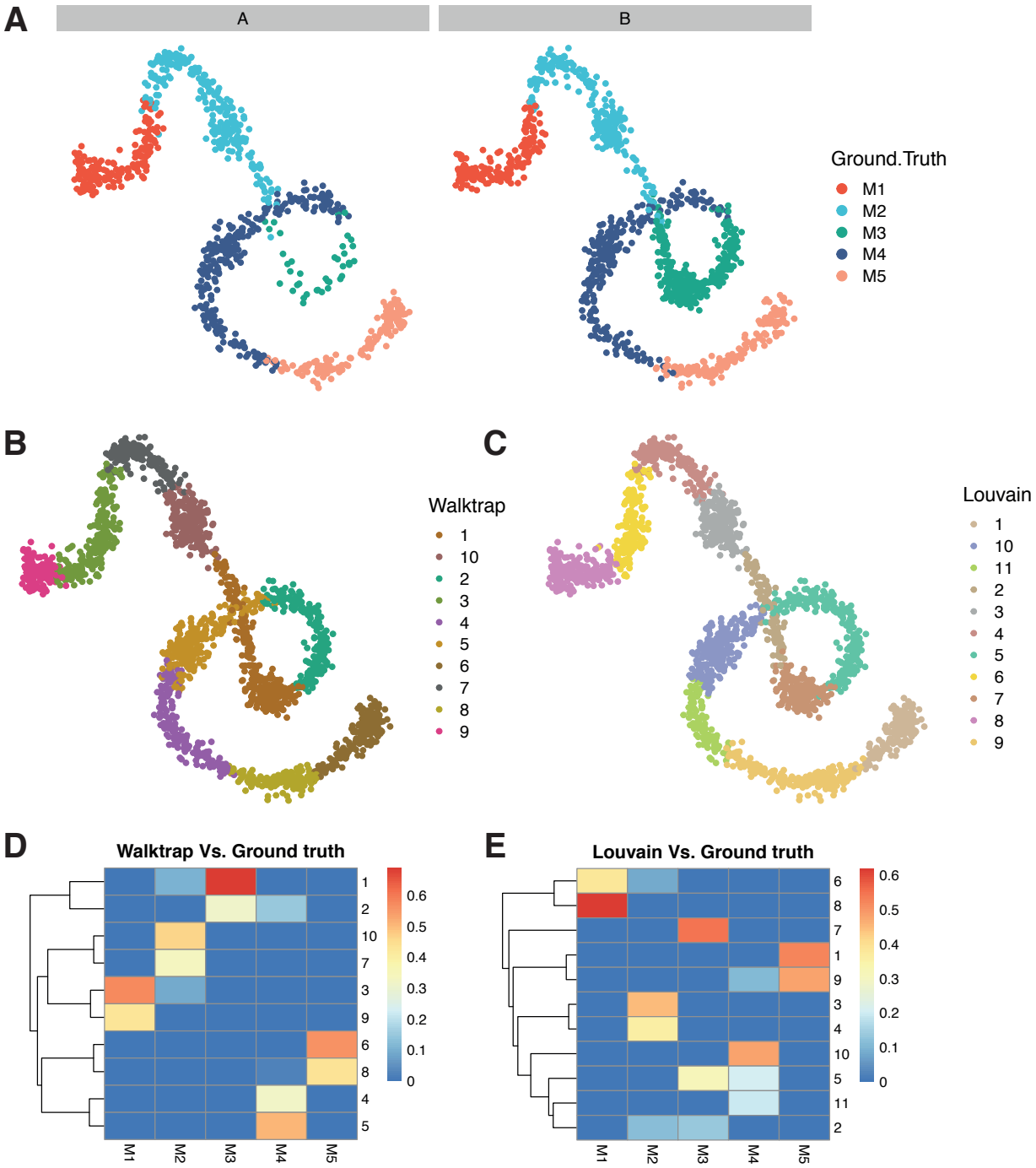

Supplementary Figure 1: **Random sampling of k-NN graph vertices is suboptimal compared to sampling with refinement.** (A) Sampling with refinement leads to selection of fewer neighbourhoods (B) Sampling with refinement leads to selection of bigger neighbourhoods for DA testing, independently of the 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.



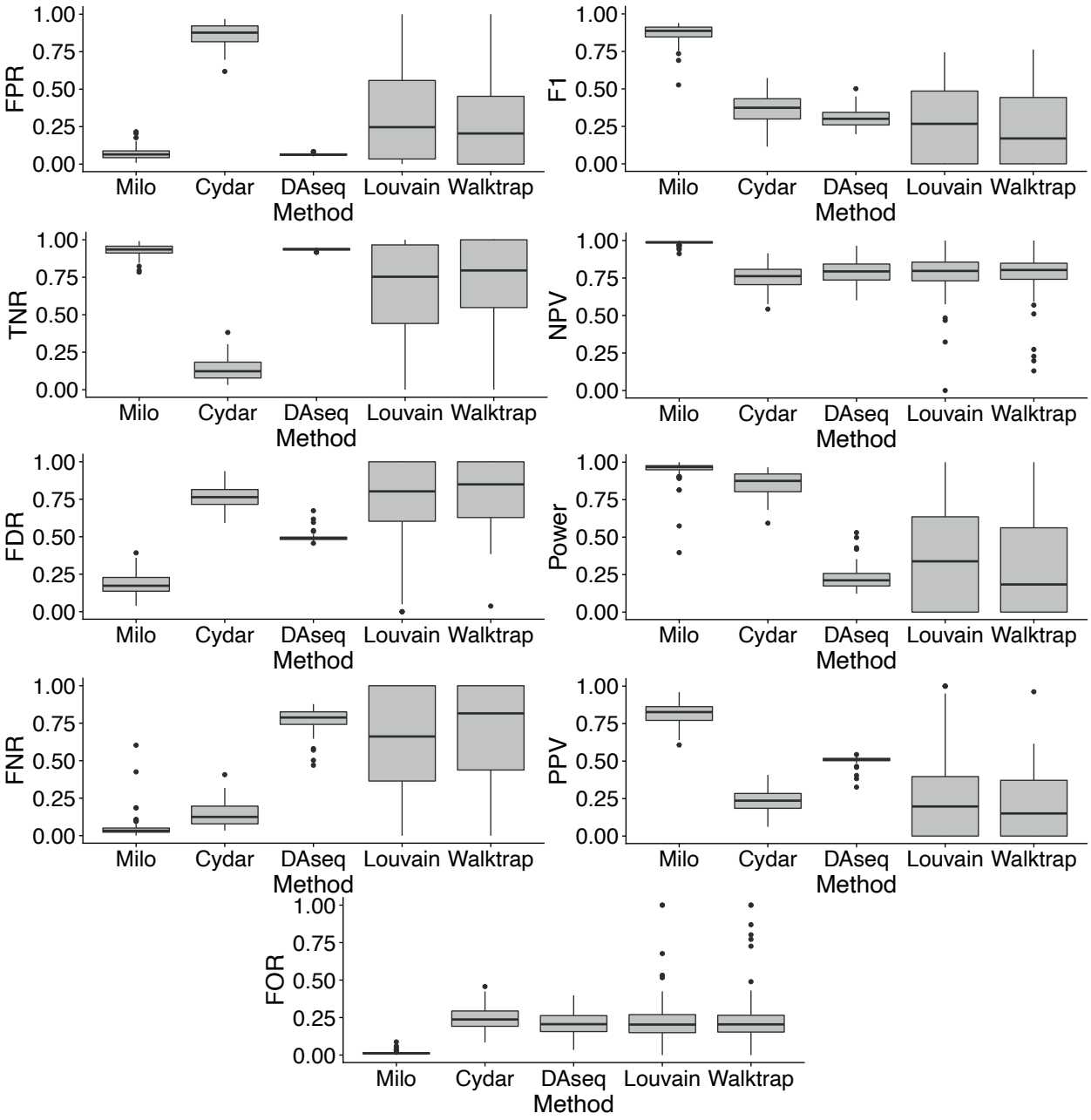

Supplementary Figure 2: **Tolerance of Milo to uncorrected batch effects in simulated data** (A-B) UMAP embeddings of simulated scRNA-seq data containing a batch effect, before batch correction (top row) and after correction with fastMNN (bottom row) (5000 cells). Cells are colored by simulated batch (A) and by presence of differential abundance between 2 simulated conditions (20% cells in condition 'A', 80% cells in condition 'B') (B). (C) A graph representation of the results from Milo differential abundance testing. Neighbourhoods were tested for DA between conditions, with (~ batch + condition) or without (~ condition) accounting for the simulated batch. Nodes are neighbourhoods, coloured by their log fold change between conditions. Non-DA neighbourhoods (FDR 10%). Node sizes correspond to the number of cells in a neighbourhood. Graph edges depict the number of cells shared between adjacent neighbourhoods. (D) Confusion matrices comparing the number of true and predicted DA neighbourhoods, with different batch effect correction (rows) and different testing design (columns).



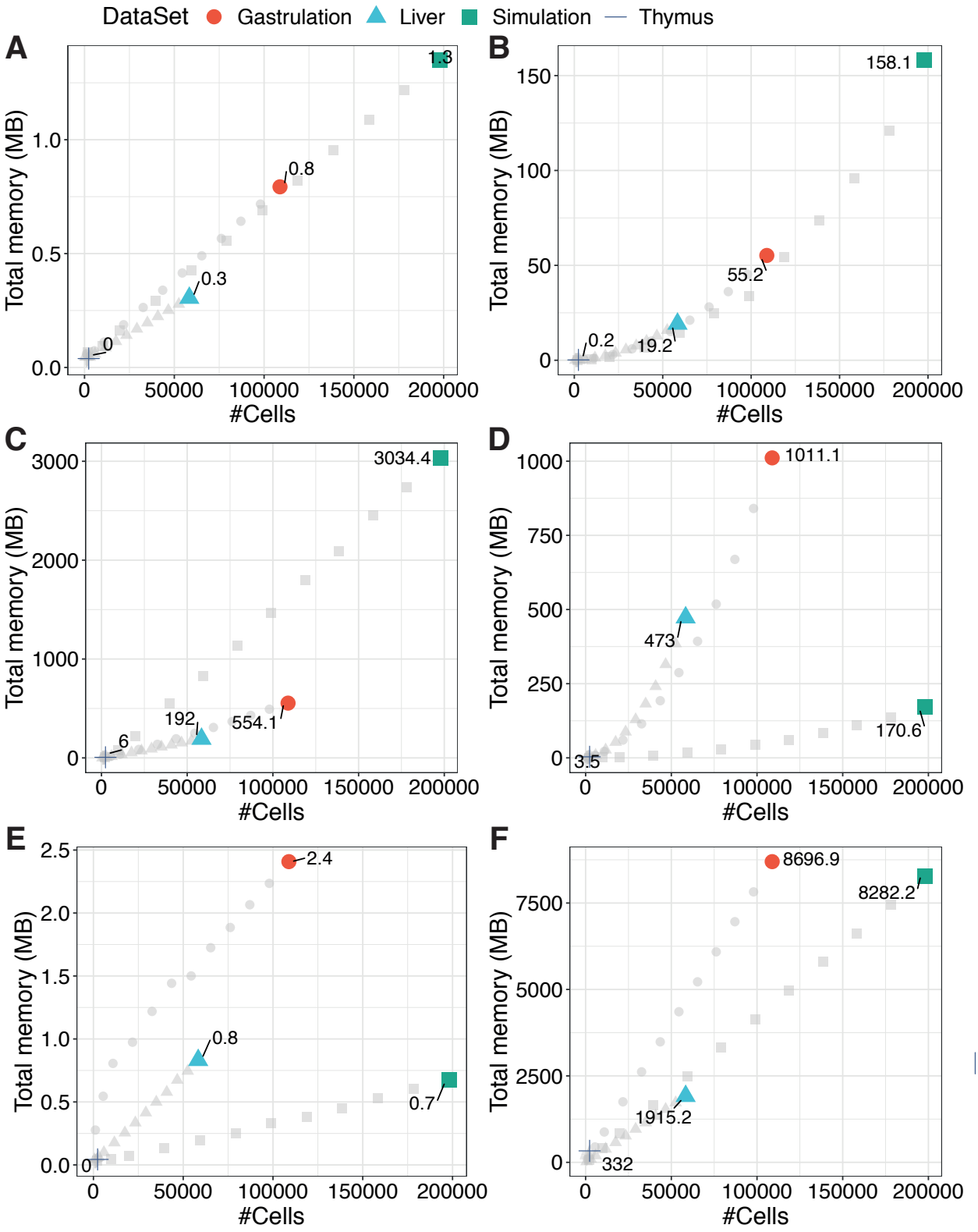

Supplementary Figure 3: **Tolerance of Milo to uncorrected batch effects in mouse gastrulation atlas** (A-B) UMAP embedding of mouse gastrulation atlas before batch correction (top row) and after correction with fastMNN (bottom row). Cells are colored by sequencing batch (A) and developmental time point (B). (C) Barplot depicting the percentage of DA neighbourhoods at FDR 10%, testing for different experimental covariates: DA between sequencing batches (~ Seq. batch), DA across developmental time points (~ Time points), DA across developmental time points accounting for the sequencing batch (~ Seq. batch + time points). The total number of DA neighbourhoods is shown in each bar.



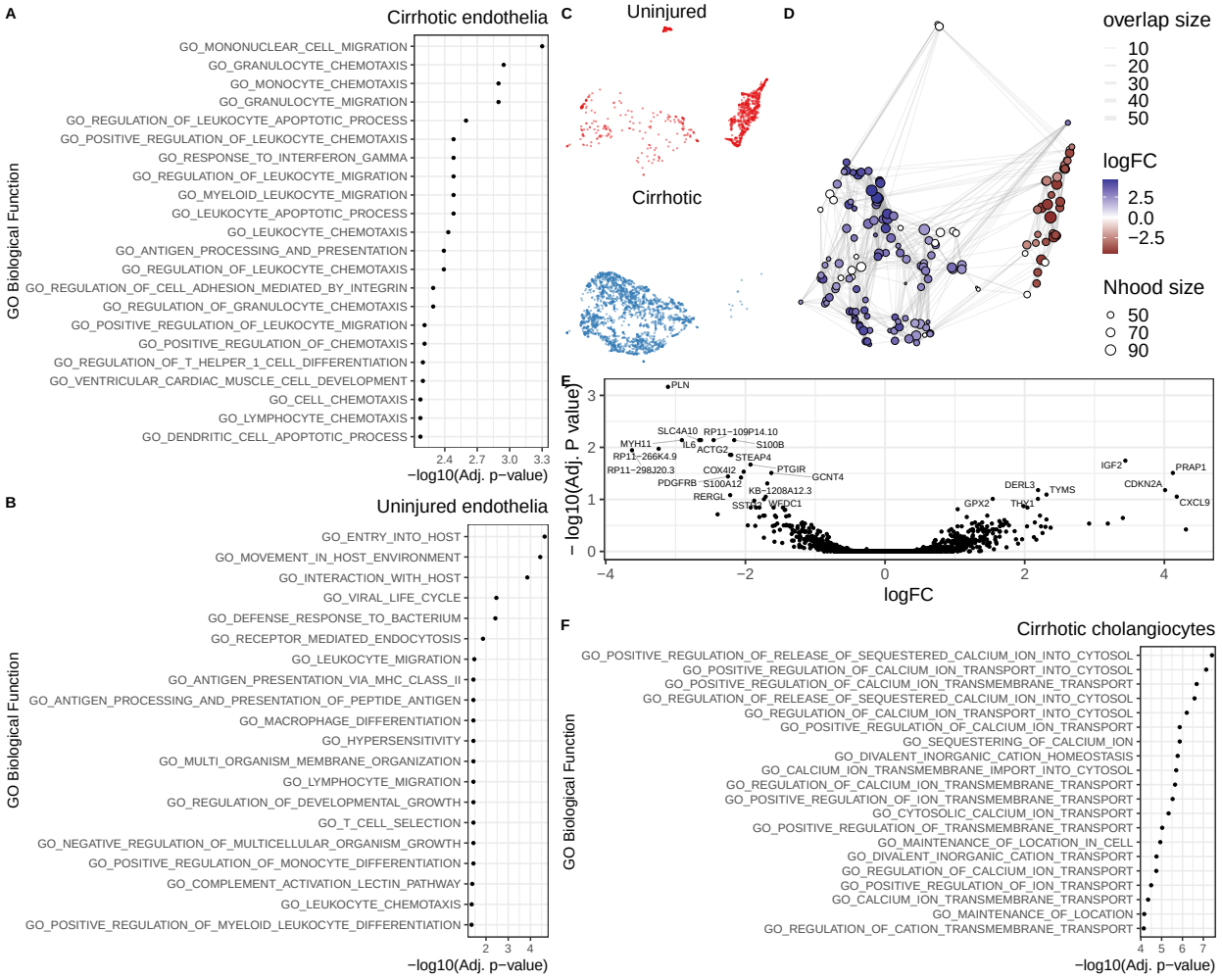

Supplementary Figure 4: **Graph-clustering does not faithfully capture simulated groups and differentially abundant subpopulations in a simulated continuous trajectory.** (A) A simulated linear trajectory of 2000 single-cells generated from 5 different groups, with cells assigned to either condition '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 5: **Comparison of Milo to alternative differential abundance methods** Each panel shows a measure of method performance computed across 100 independent simulations. Boxplots denote the median and interquartile range (IQR), with whiskers extending 1.5 x IQR; outliers are shown as individual points. Each analysis on each independent simulation used the same parameter values in Supplementary Table 1.




Supplementary Figure 6: **Memory usage across the Milo analysis workflow** Total memory usage across the steps of the Milo analysis workflow in 4 data sets containing different numbers of cells (Gastrulation: circles, Liver: triangles, Thymus: crosses, Simulation: squares). Grey points denoted 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 data-sets for graph building and neighbourhood construction (k=30).




Supplementary Figure 7: **Downstream analysis of disease-specific subpopulations in liver cirrhosis** (A) GO term enrichment analysis on marker genes of cirrhosis-enriched endothelia. The top 20 significant terms are shown. (B) GO term enrichment analysis on marker genes of healthy-enriched endothelia. The top 20 significant terms are shown. (C) UMAP embedding and graph representation of neighbourhoods of 3369 cells from cholangiocytes lineage. (D) 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. (E) GO term enrichment analysis on marker genes of cirrhosis-enriched cholangiocytes. The top 20 significant terms are shown.

# 1 Supplementary tables

| Method | Key Parameters | Values | Hypothesis testing |
|---|---|---|---|
| Milo | K | 10 | Negative binomial GLM, 10% FDR |
| | d | 15 | Negative binomial GLM, 10% FDR |
| Cydar | r | 2 | Negative binomial GLM, 10% FDR |
| DAseq | k.vector | 5-500, steps of 50 | Logistic classifier prediction, top 10% |
| Louvain + edgeR | K | 10 | Negative binomial GLM, 10% FDR |
| | d | 15 | |
| Walktrap + edgeR | K | 10 | Negative binomial GLM, 10% FDR |
| | d | 15 | |

Supplementary Table 1: **Method comparison parameter values.**