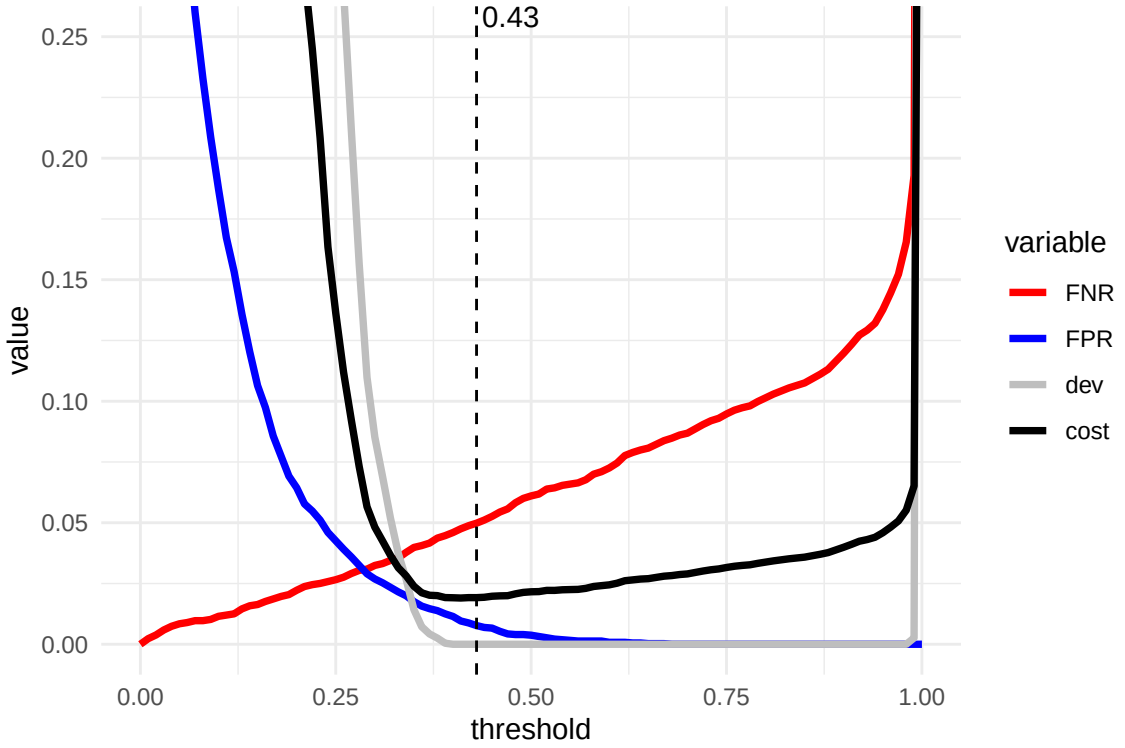
Doublet identification in single-cell sequencing data using scDblFinder

Supplementary Figures

Pierre-Luc Germain

30 August, 2021

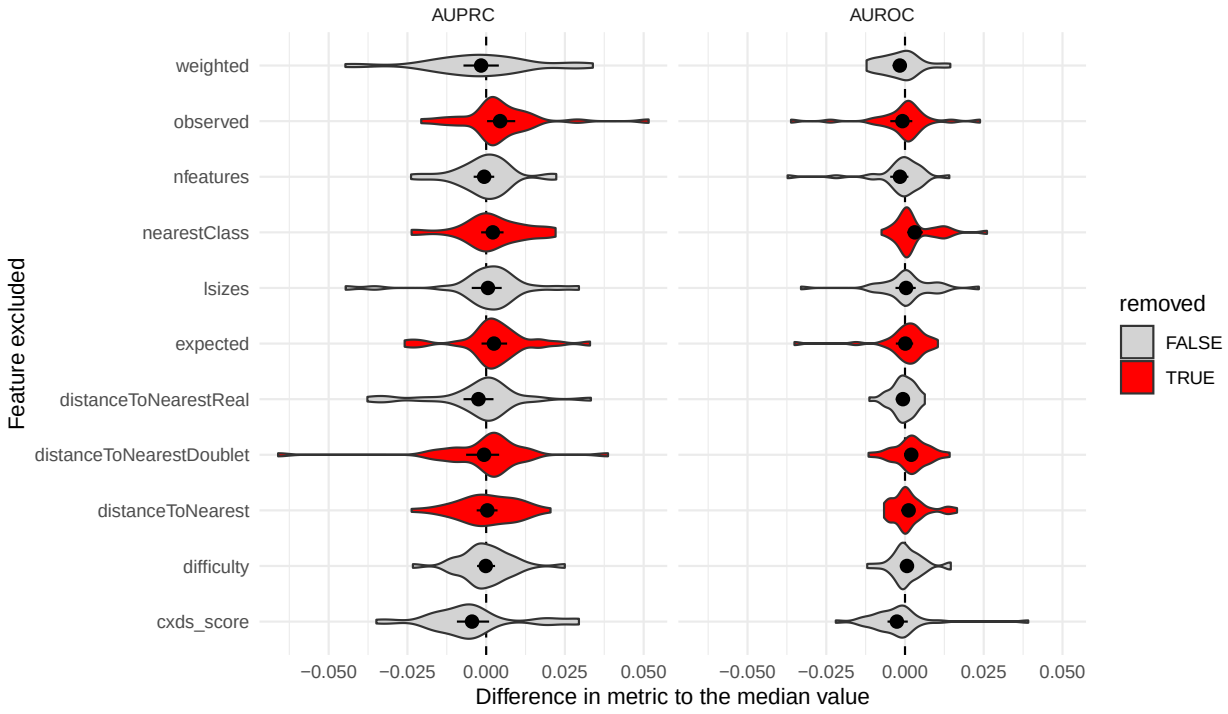
Supplementary Figure 1



Supplementary Figure 1

Example thresholding. Plotted are the false negative rate (FNR; the rate of misclassified artificial doublets), the false positive rate (FPR; the proportion of real cells classified as doublets, excluding those called in previous iterations), the squared proportion deviation from the expected doublet rate (denoted 'dev'), and the integrated cost function to be minimized (mean of the previous). The dashed lined indicates the threshold called.

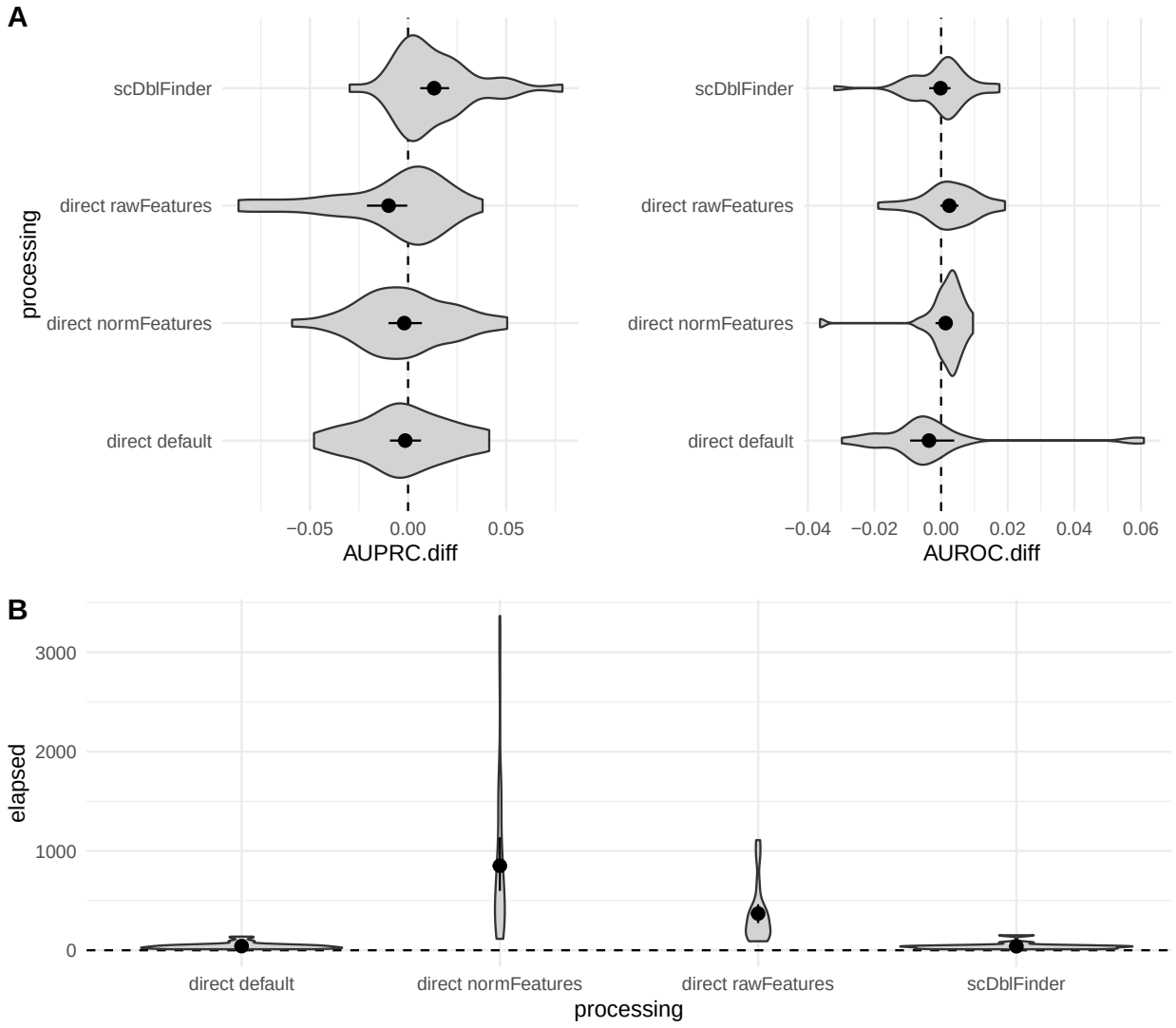
Supplementary Figure 2



Supplementary Figure 2

Effect of removing a feature on the dataset-relative accuracy of doublet prediction. scDblFinder was run across the 16 benchmark datasets removing a given feature, and comparing to the median accuracy. Features in red were then removed from the default settings.

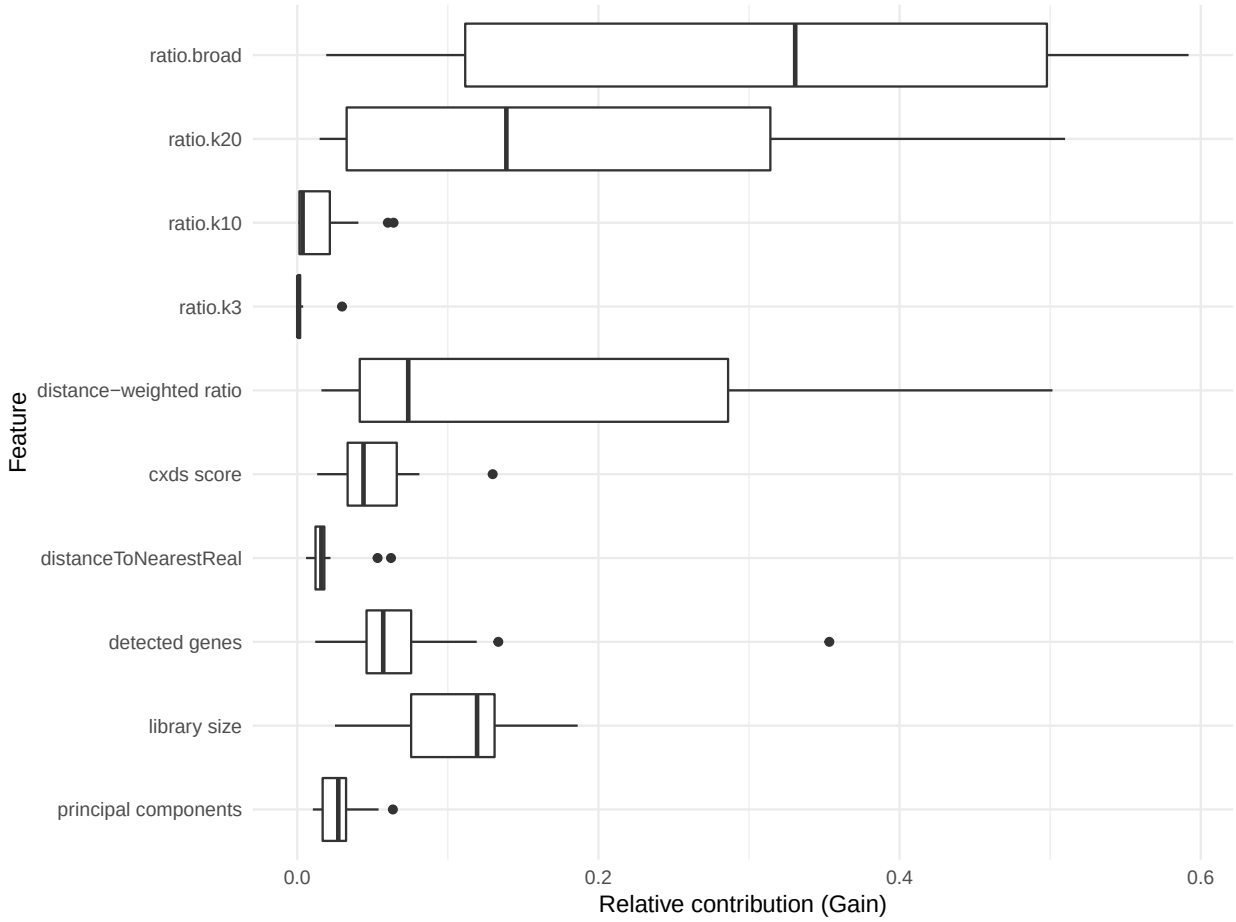
Supplementary Figure 3



Supplementary Figure 3

Direct classification vs classifying on the kNN features. The standard `scDblFinder` method is compared to training a classifier directly on the features (implemented in the package's `directDblClassification` function), either using the PCA ('default'), the normalized ('normFeatures') or the raw counts ('rawFeatures', default). In all cases, the doublet generation, number of features and iterative procedure is the same. `scDblFinder` (i.e. working on the kNN) has a better AUPRC (A, left) at a considerably greater speed than gene-based classifiers (B). Direct classification based on the raw features however had a slightly better AUROC.

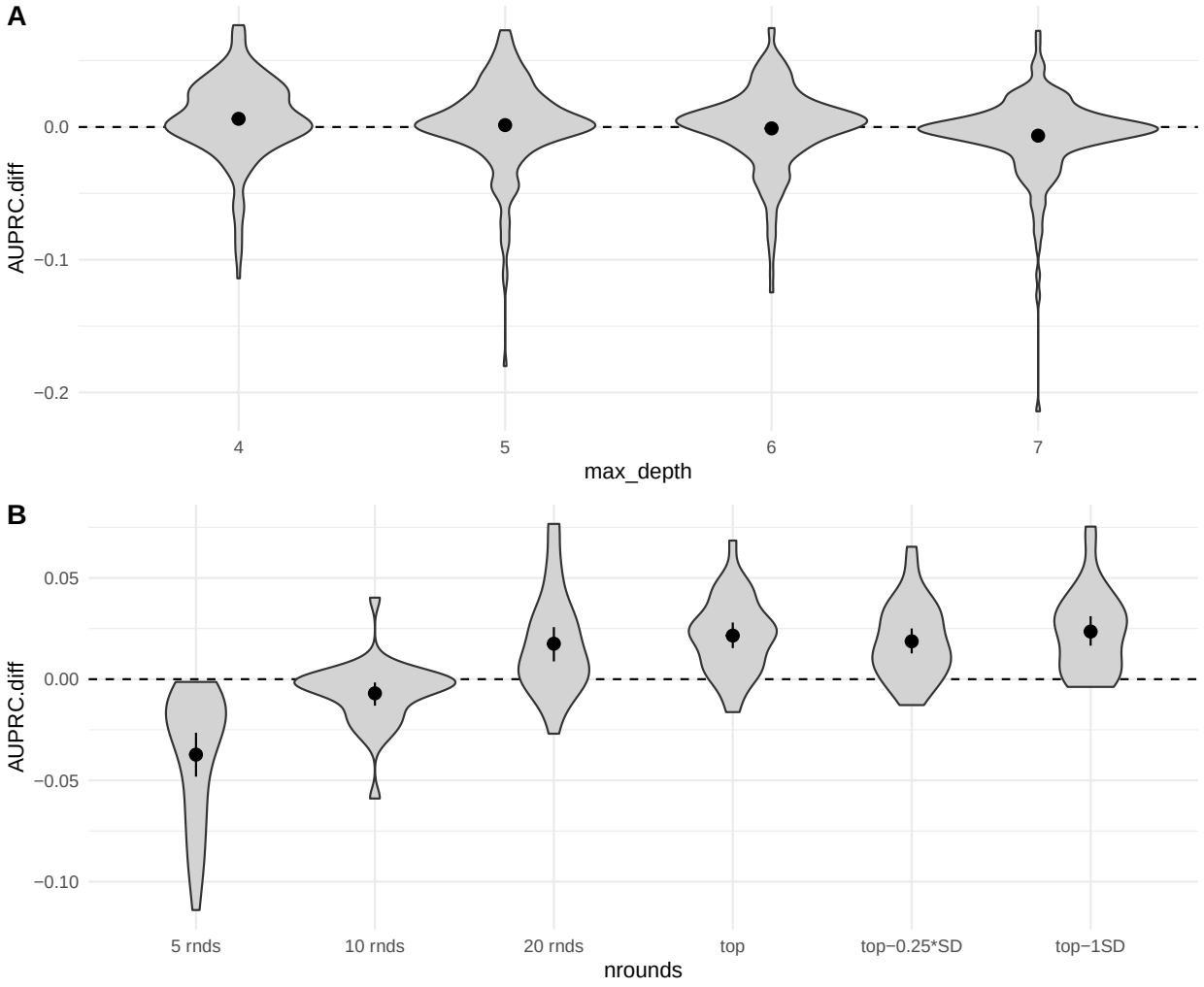
Supplementary Figure 4



Supplementary Figure 4

Variable importance calculated during training. Each dot represents a dataset. For the principal components, the gain of the most informative component per dataset is used. `ratio.broad` refers to the ratio of artificial doublets in the largest neighborhood looked at (which varies across datasets).

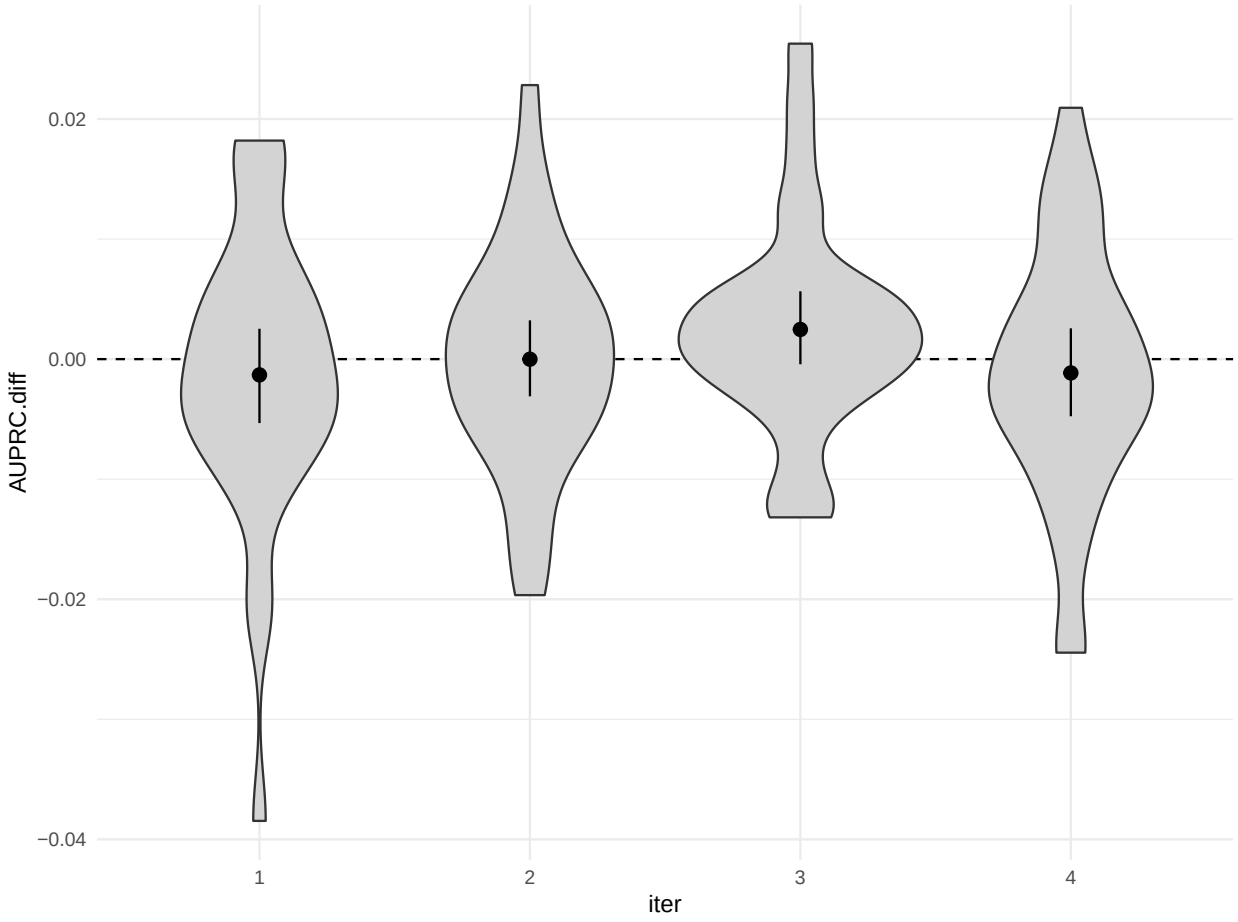
Supplementary Figure 5



Supplementary Figure 5

Hyperparameter optimization: max tree depth (A) and number of boosting rounds (B). 'Top' indicates the optimal number of rounds according to cross-validation logloss in the real vs artificial classification.

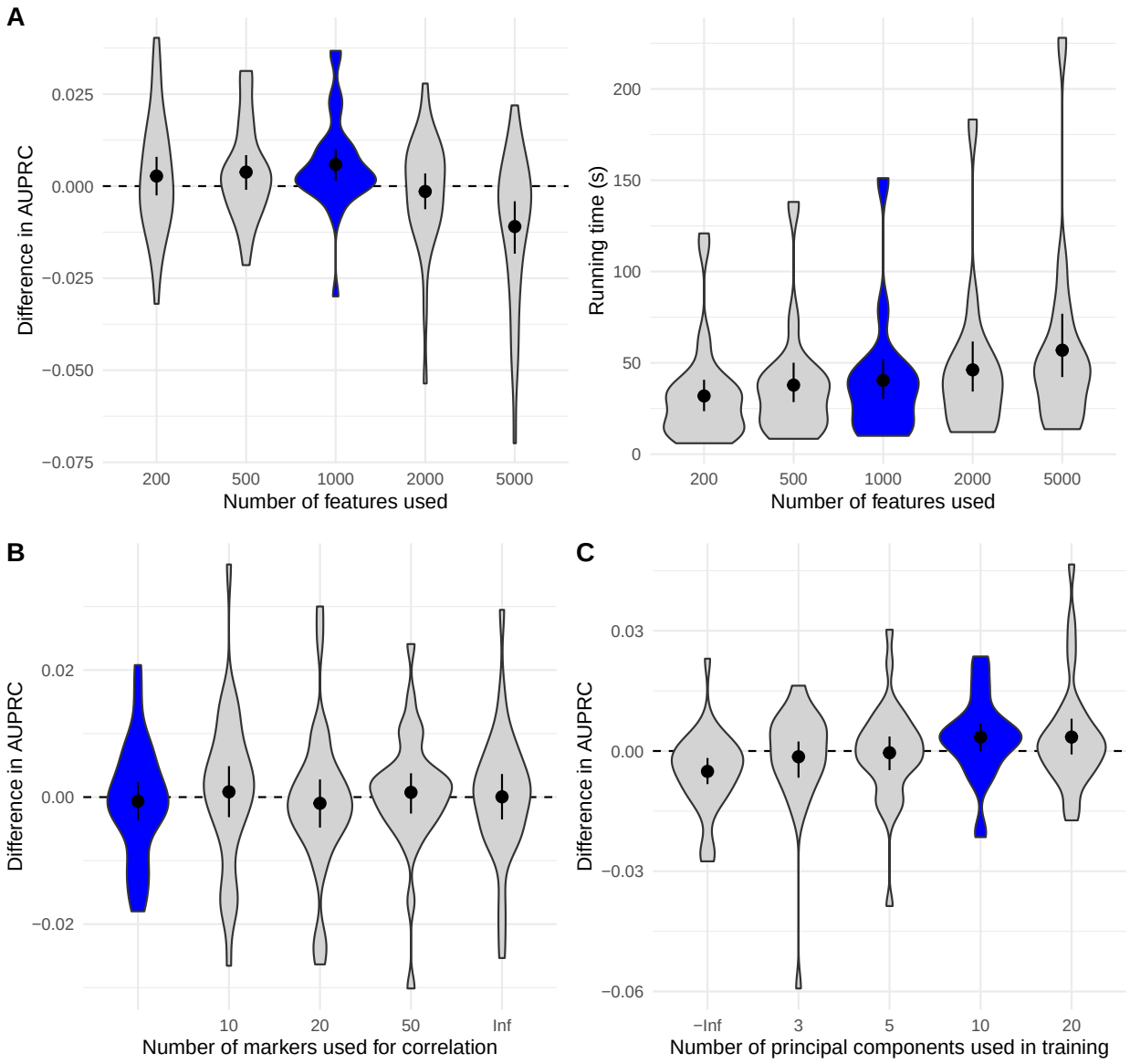
Supplementary Figure 6



Supplementary Figure 6

Number of learning iteration. At each round, the real cells identified as doublets are removed from the training data for the next round.

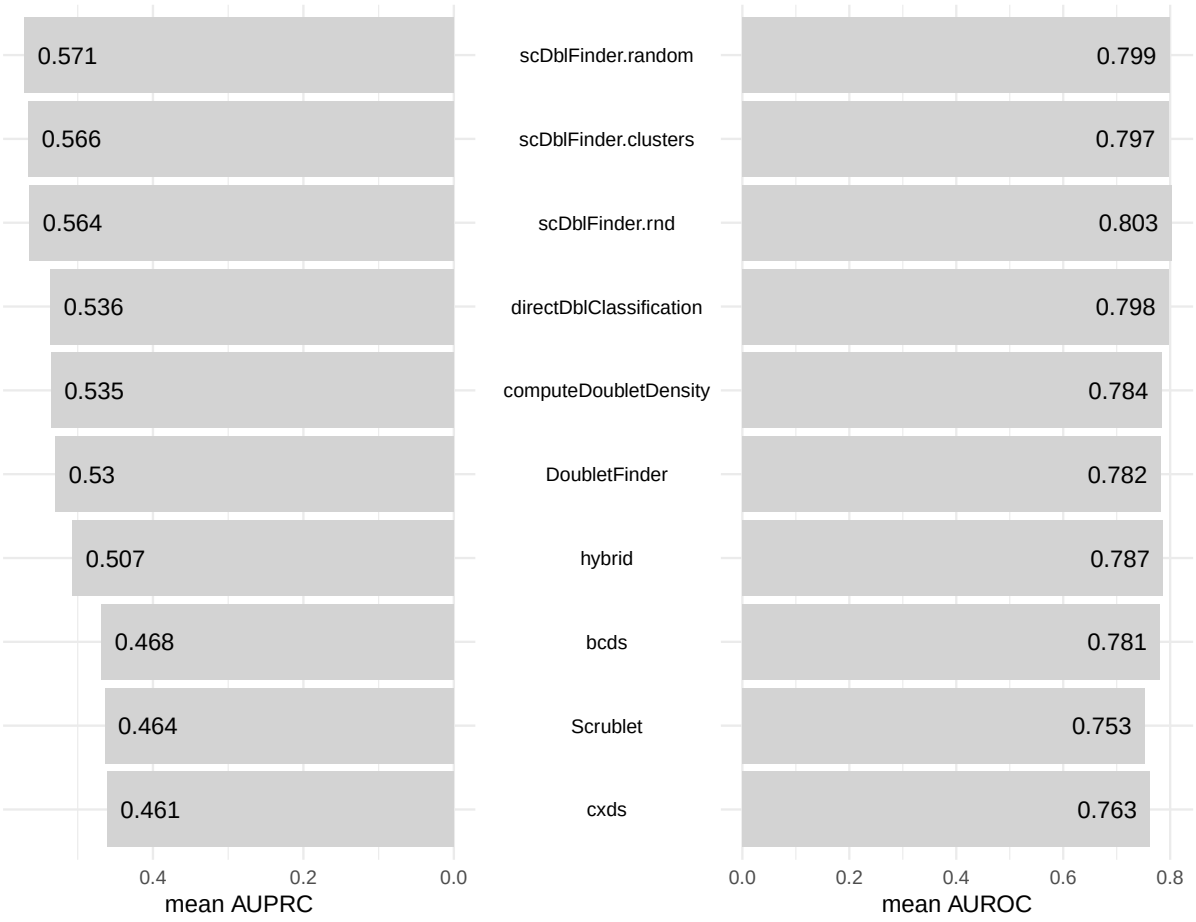
Supplementary Figure 7



Supplementary Figure 7

Effect of number of features, number of components, and marker correlation. The selected default settings are in blue. Using the correlation across cluster-based marker genes increased running time without improving much the accuracy (B).

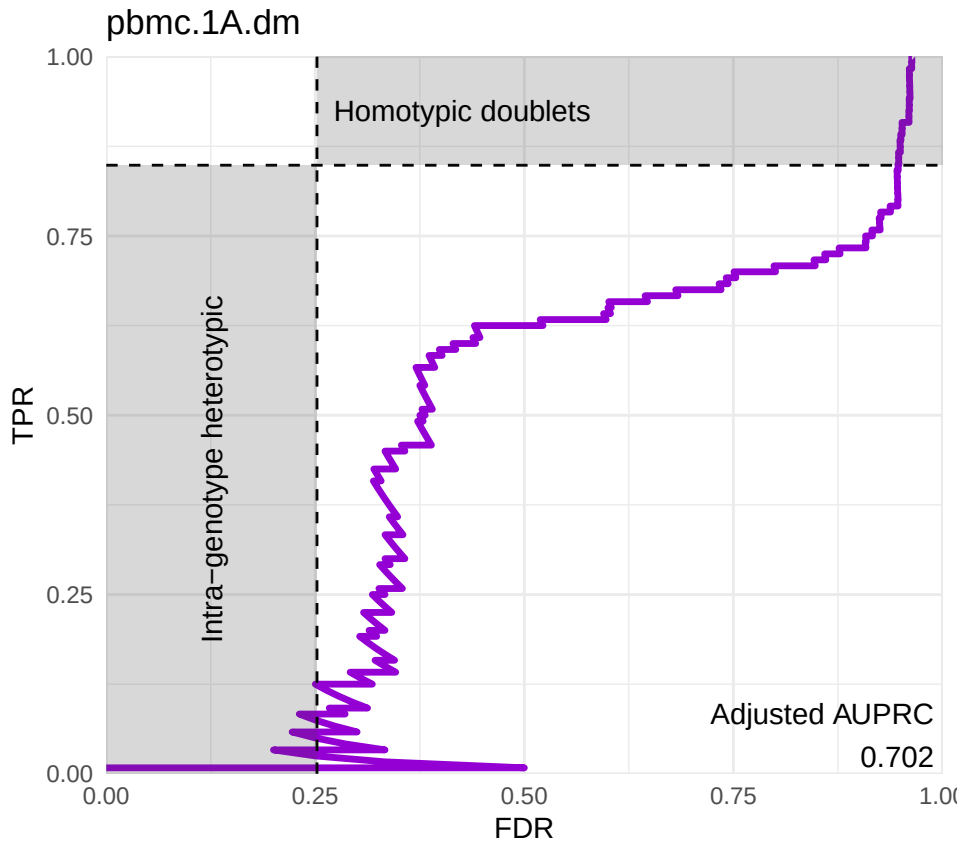
Supplementary Figure 8



Supplementary Figure 8

Average area under the ROC and PR curves across datasets for each method.

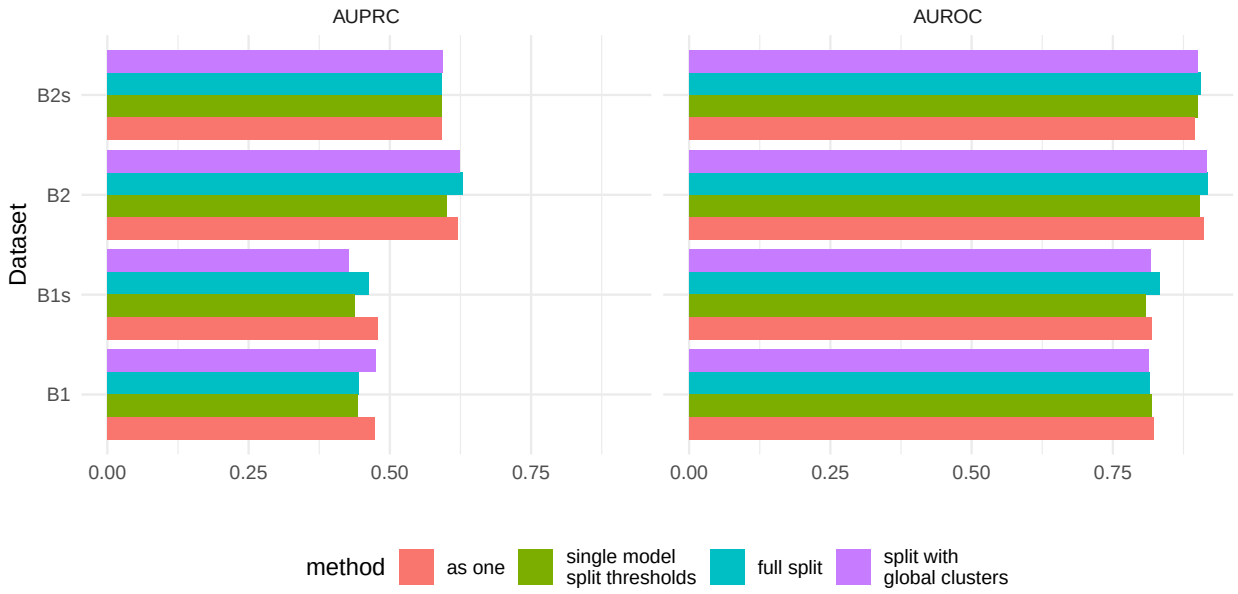
Supplementary Figure 9



Supplementary Figure 9

Estimated accuracy of heterotypic doublet identification. The two shaded areads represent the expected proportion of, respectively, intra-genotype heterotypic doublets (i.e. wrongly labeled as singlets in the truth) and inter-genotype homotypic doublets.

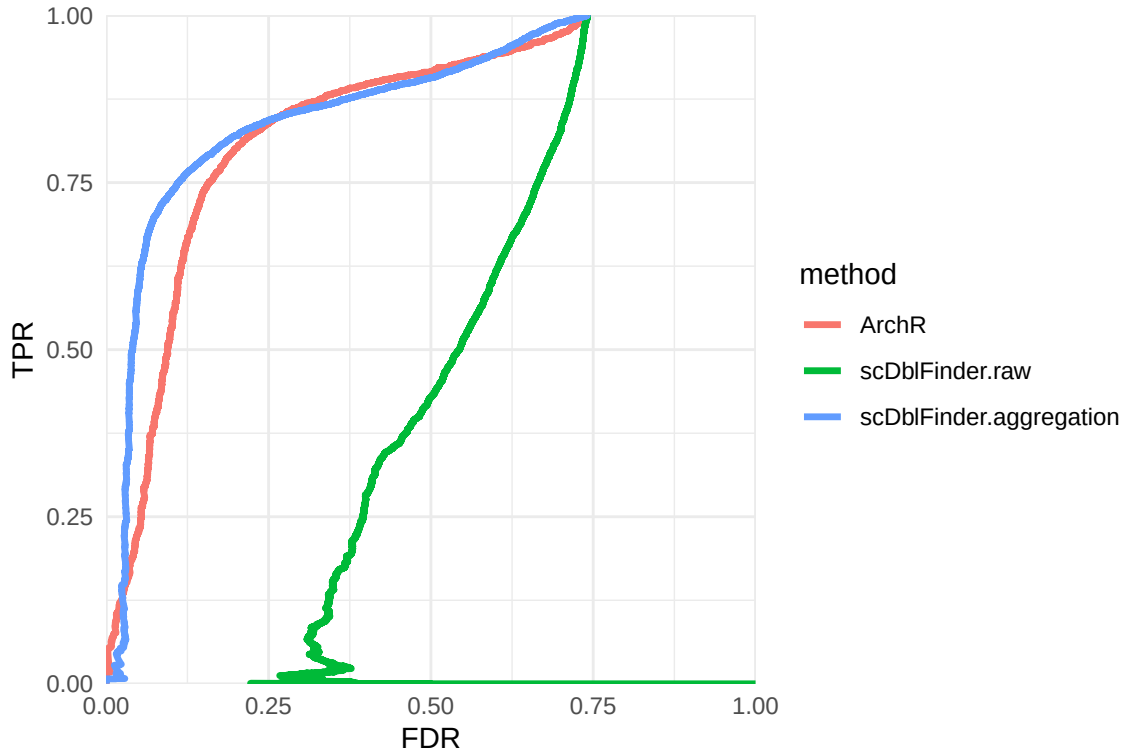
Supplementary Figure 10



Supplementary Figure 10

Comparison of four multi-sample strategies. B1 and B2 the two batches from dataset GSE96583, and contain 3 and 2 captures, respectively. The datasets with the 's' suffix are versions downsampled to 30%. Using doublet detection on each capture separately (full split) was generally comparable to treating the captures as one (and adjusting the doublet rate).

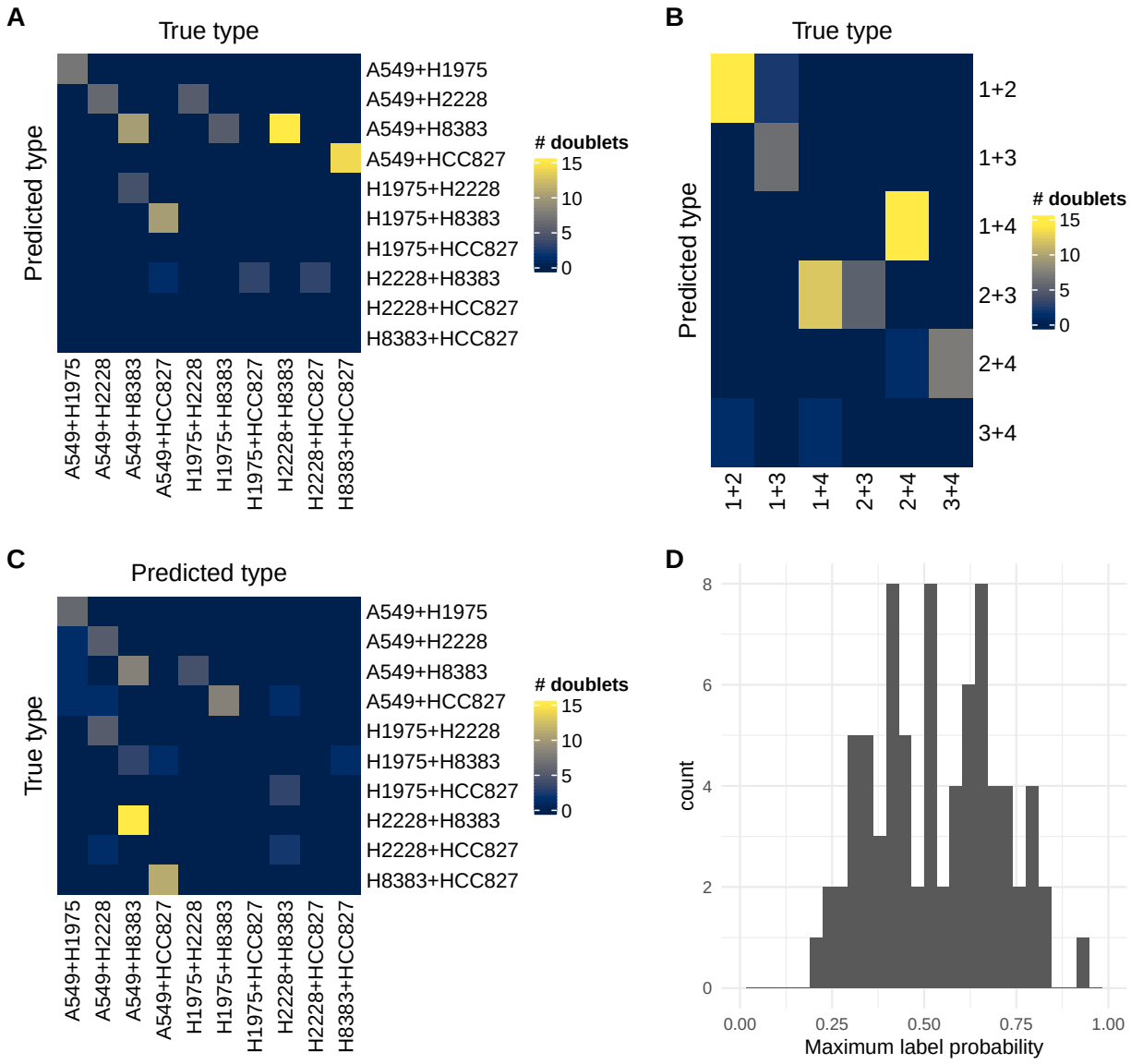
Supplementary Figure 11



Supplementary Figure 11

Doublet identification in single-nucleus ATAC-seq Performance of `scDblFinder` with default (.raw) parameters or on aggregated features (.aggregation) versus ArchR (GSE162690 dataset).

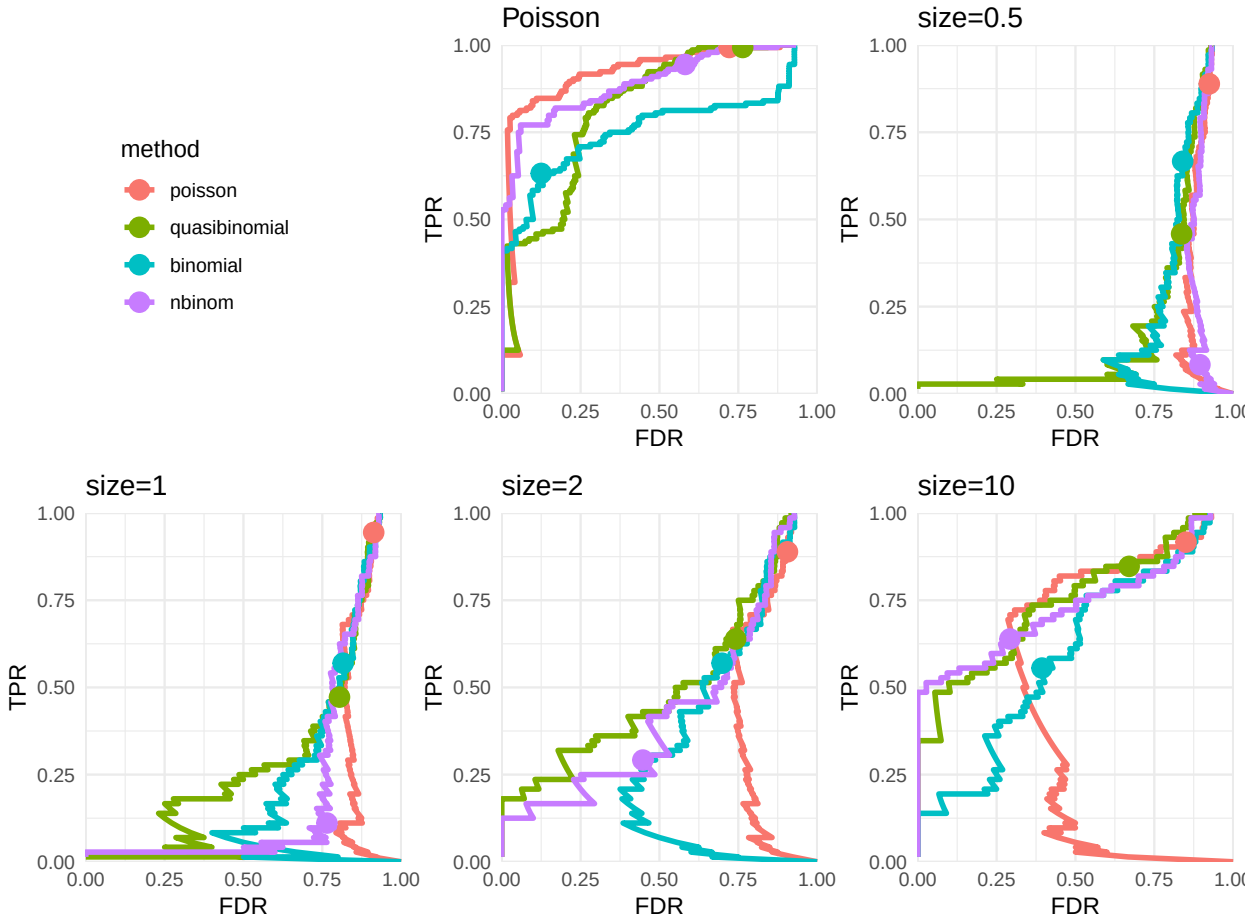
Supplementary Figure 12



Supplementary Figure 12

Failure to recognize doublet types. Confusion matrices of the doublet type (i.e. originating clusters) identification from the nearest artificial doublets on the kNN, for a real dataset (A) and a simple simulation (B), and training a classifier on the problem using artificial doublets (C-D). The maximum label probabilities per doublet (D) of the classifier indicate a low confidence of the predictions.

Supplementary Figure 13



Supplementary Figure 13

FDR of 'cluster stickiness' tests across simulations with different overdispersion parameters.