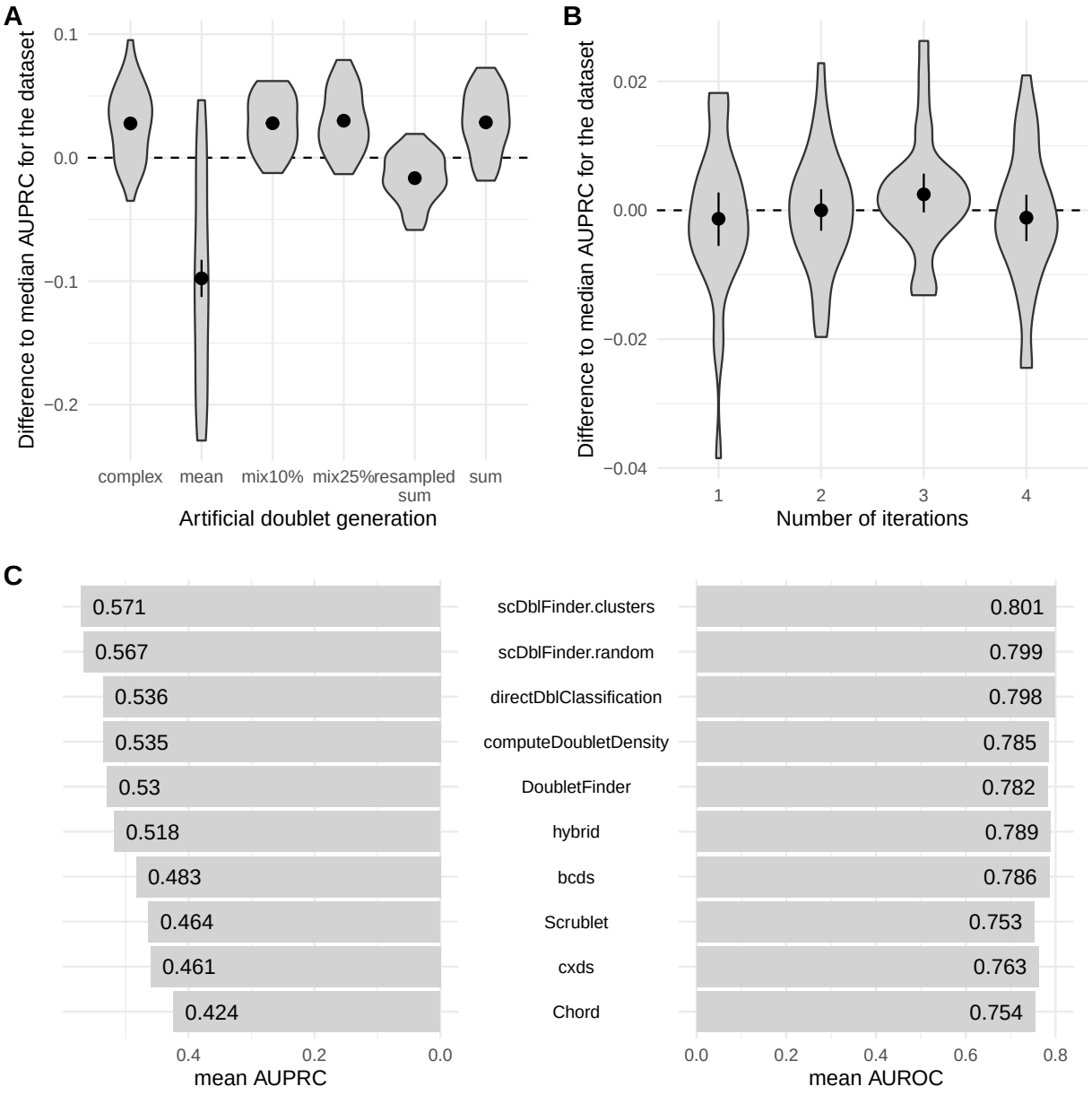
Doublet identification in single-cell sequencing data using scDblFinder

Supplementary Figures

Pierre-Luc Germain

18 April, 2022

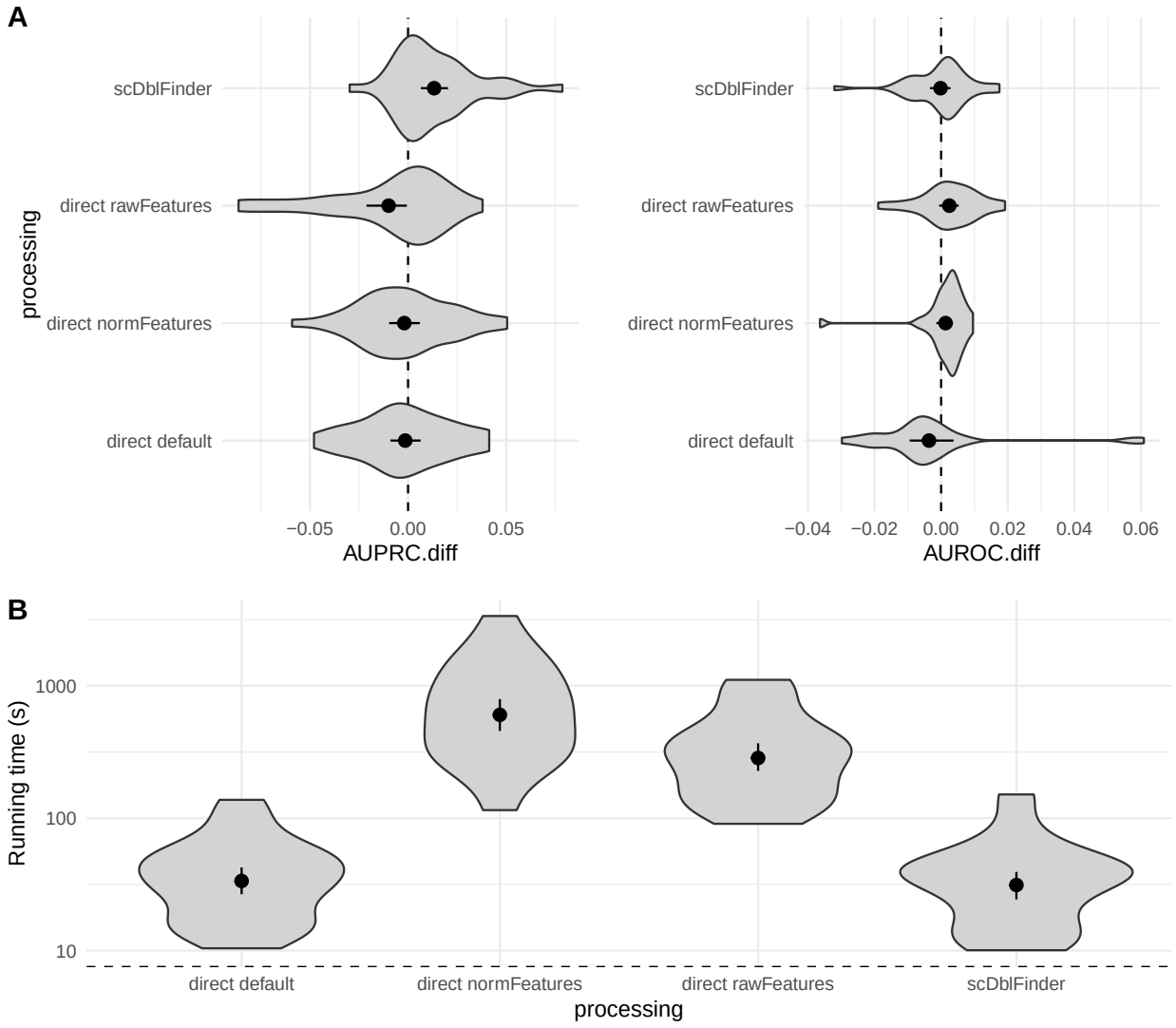
Supplementary Figure 1



Supplementary Figure 1

Artificial doublet generation and iterative classification. A: Effect of different methods of artificial doublet generation on the performance of `scDblFinder` across the 16 benchmark datasets. `sum` and `mean` respectively indicate the sum or mean of the counts of the two cells, `resampled sum` indicates the sum followed by Poisson resampling, and `mix` indicates the mixture of approaches. **B:** Effect of the number of iterations on the accuracy. At each round, the real cells identified as doublets are removed from the training data for the next round. **C:** Average area under the ROC and PR curves across datasets for each method. In **A-B**, the violins represent the distribution (as well as mean and standard error), across benchmark datasets, of differences to the median (across method) AUPRC for the dataset (higher is better).

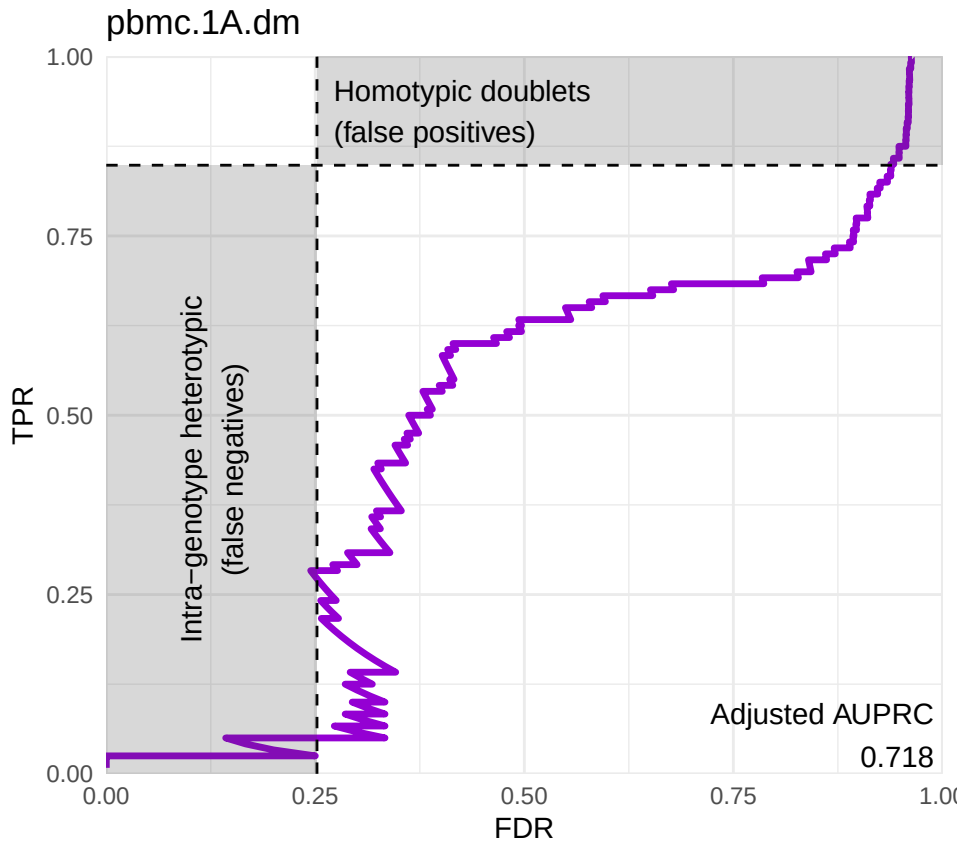
Supplementary Figure 2



Supplementary Figure 2

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.

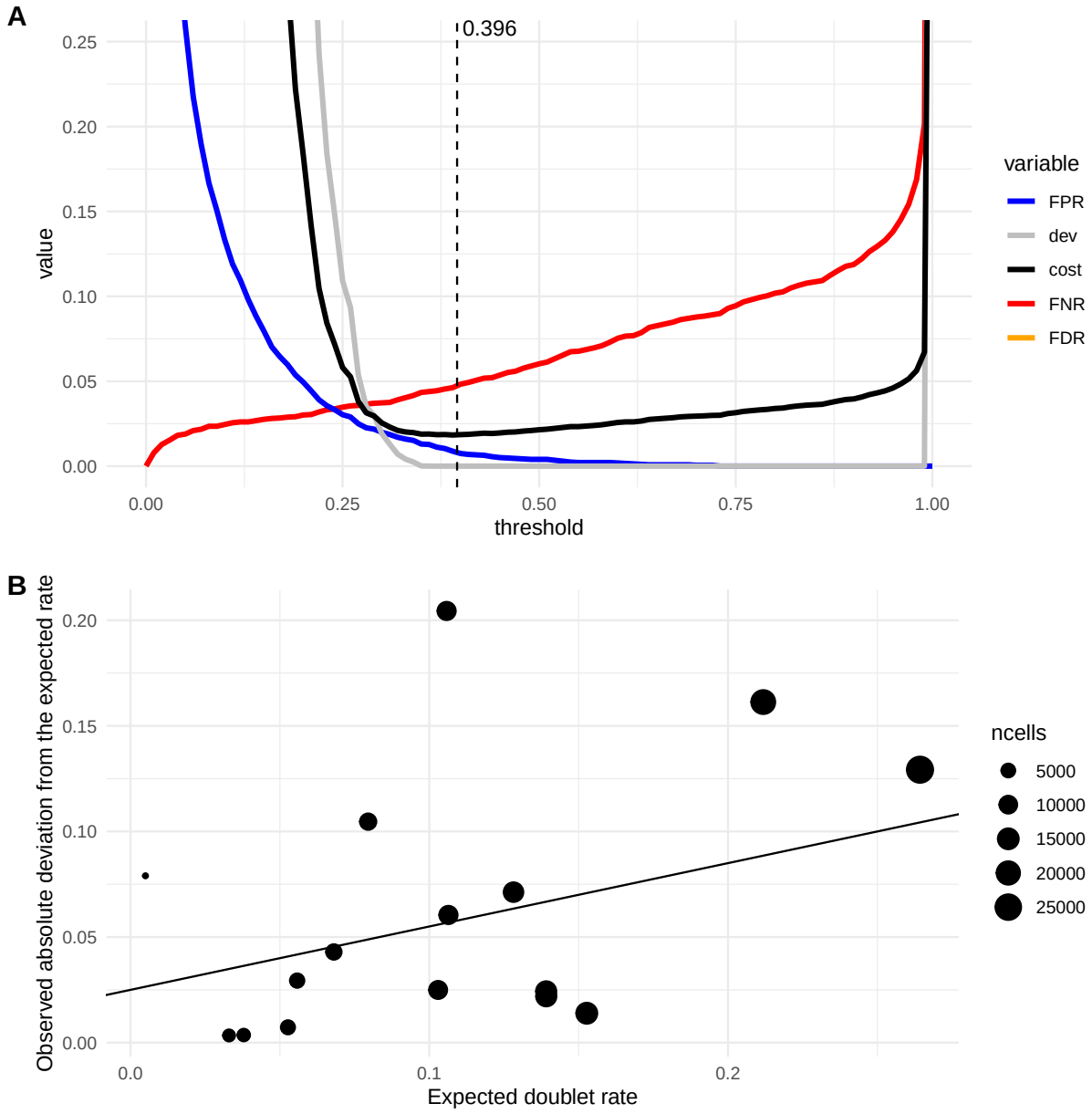
Supplementary Figure 3



Supplementary Figure 3

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

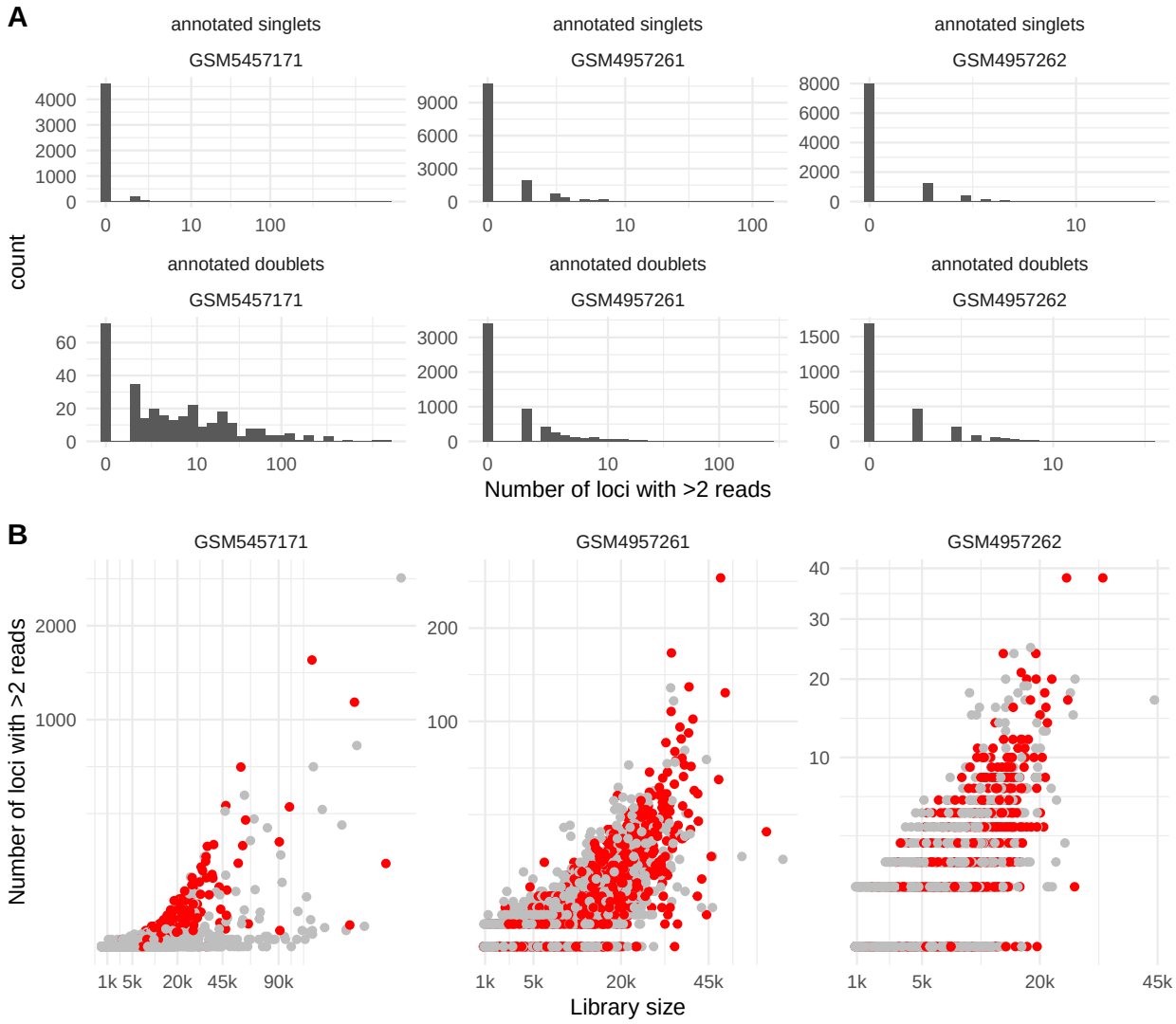
Supplementary Figure 4



Supplementary Figure 4

Combined thresholding. A: Illustration of the cost function to be minimized for thresholding. Plotted are the false negative rate (FNR; the rate of misclassified artificial doublets), the false positive rate (FPR; the proportion of real droplets classified as doublets), the squared proportion deviation from the expected doublet rate (denoted 'dev', accepting an interval range), and the integrated cost function to be minimized (mean of the previous). The dashed lined indicates the threshold called. **B:** By default, the expected doublet rate used for thresholding is a range around the given or calculated rate, whose width (defined by the black line) is roughly based on observed deviation from the expected rate.

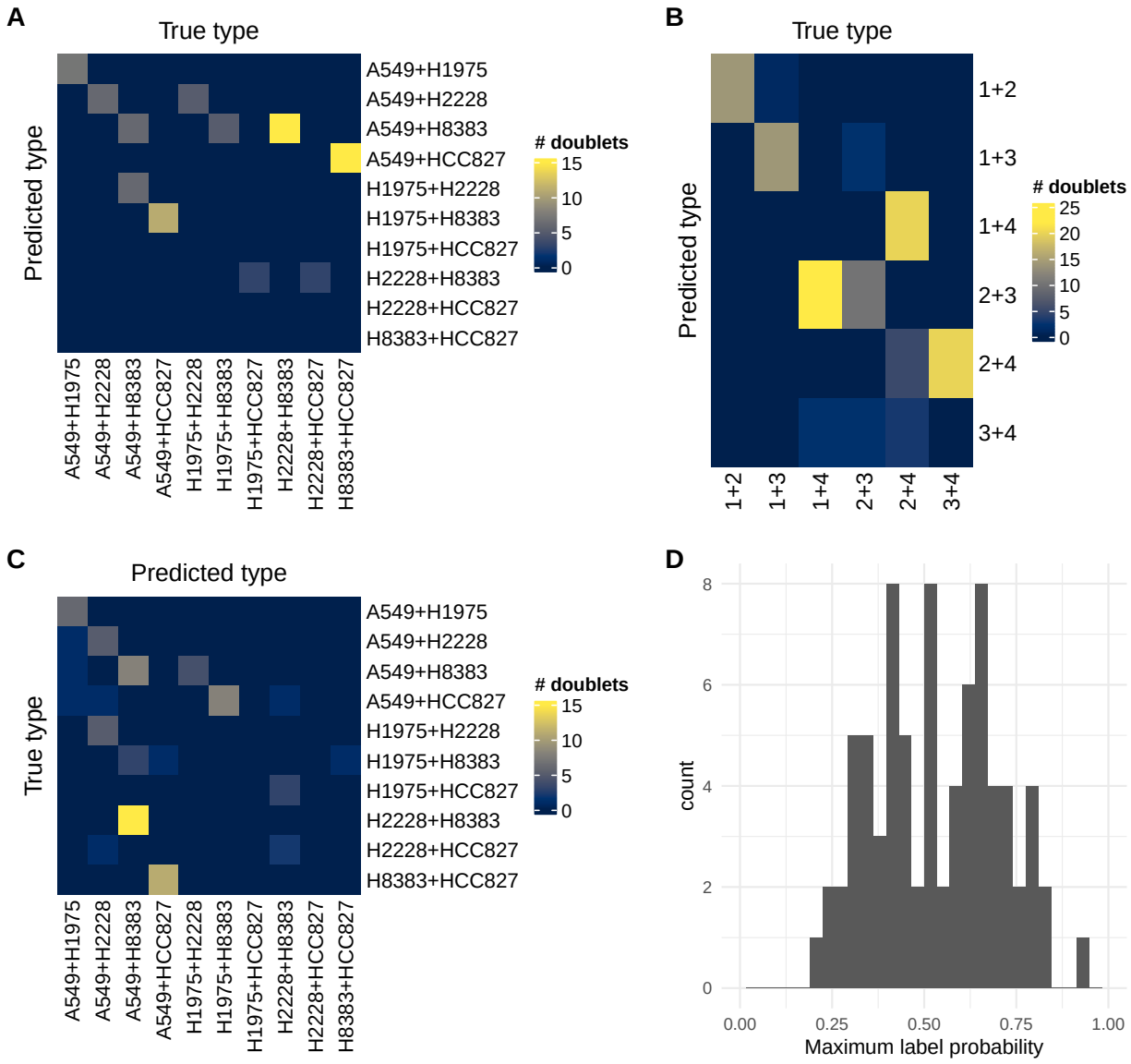
Supplementary Figure 5



Supplementary Figure 5

Doublets and sites covered by more than two reads (scATACseq). A: Number of loci covered by more than two reads in annotated singlets and doublets in each dataset. **B:** Relationship between library size and the number of loci covered by more than two reads. True doublets are plotted in red.

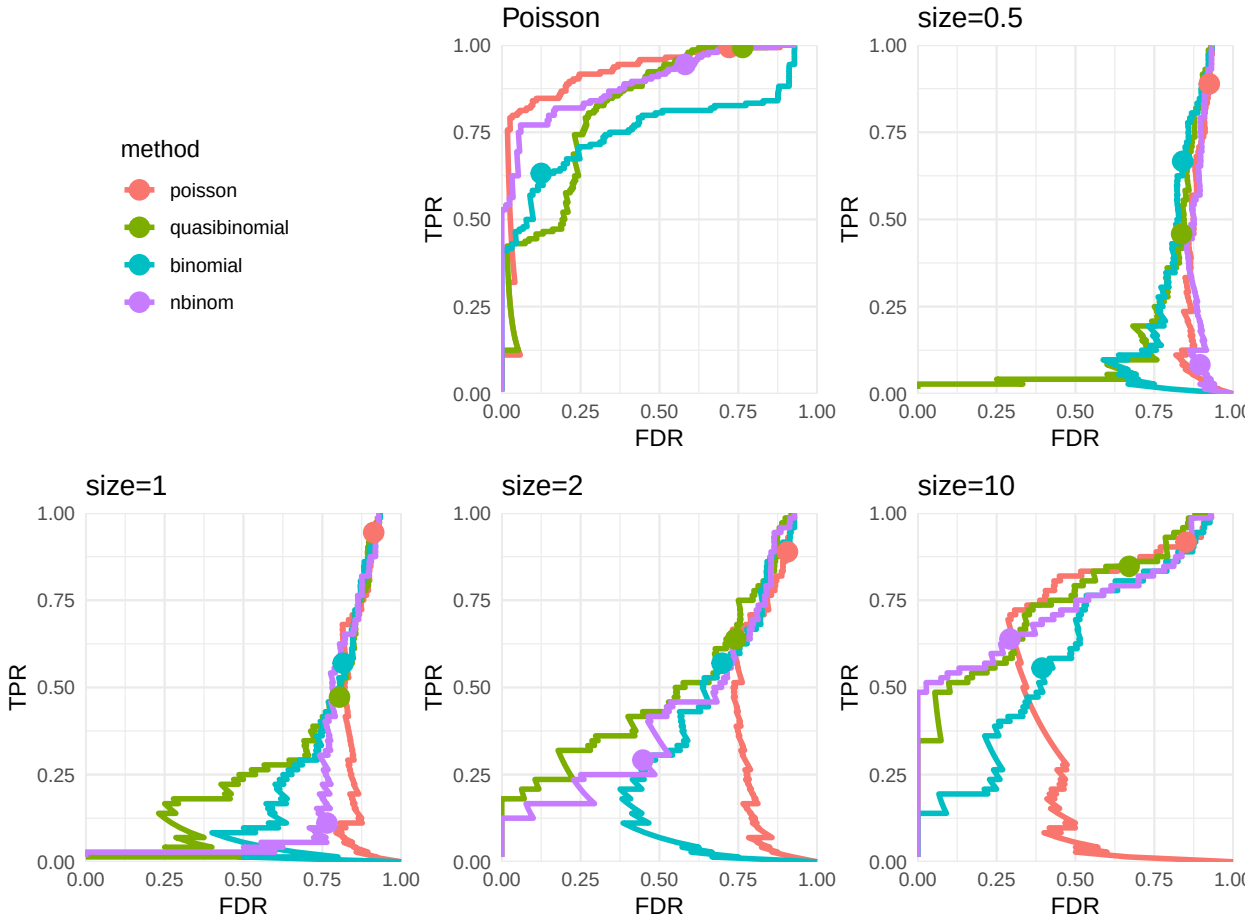
Supplementary Figure 6



Supplementary Figure 6

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.

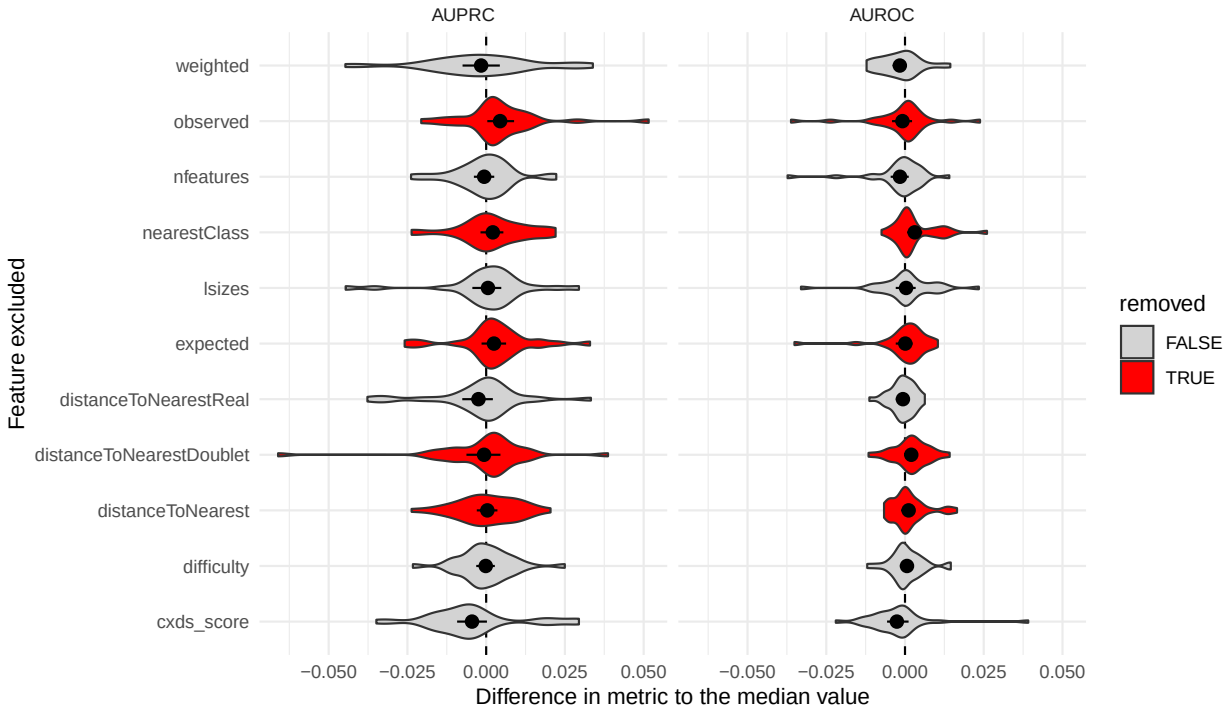
Supplementary Figure 7



Supplementary Figure 7

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

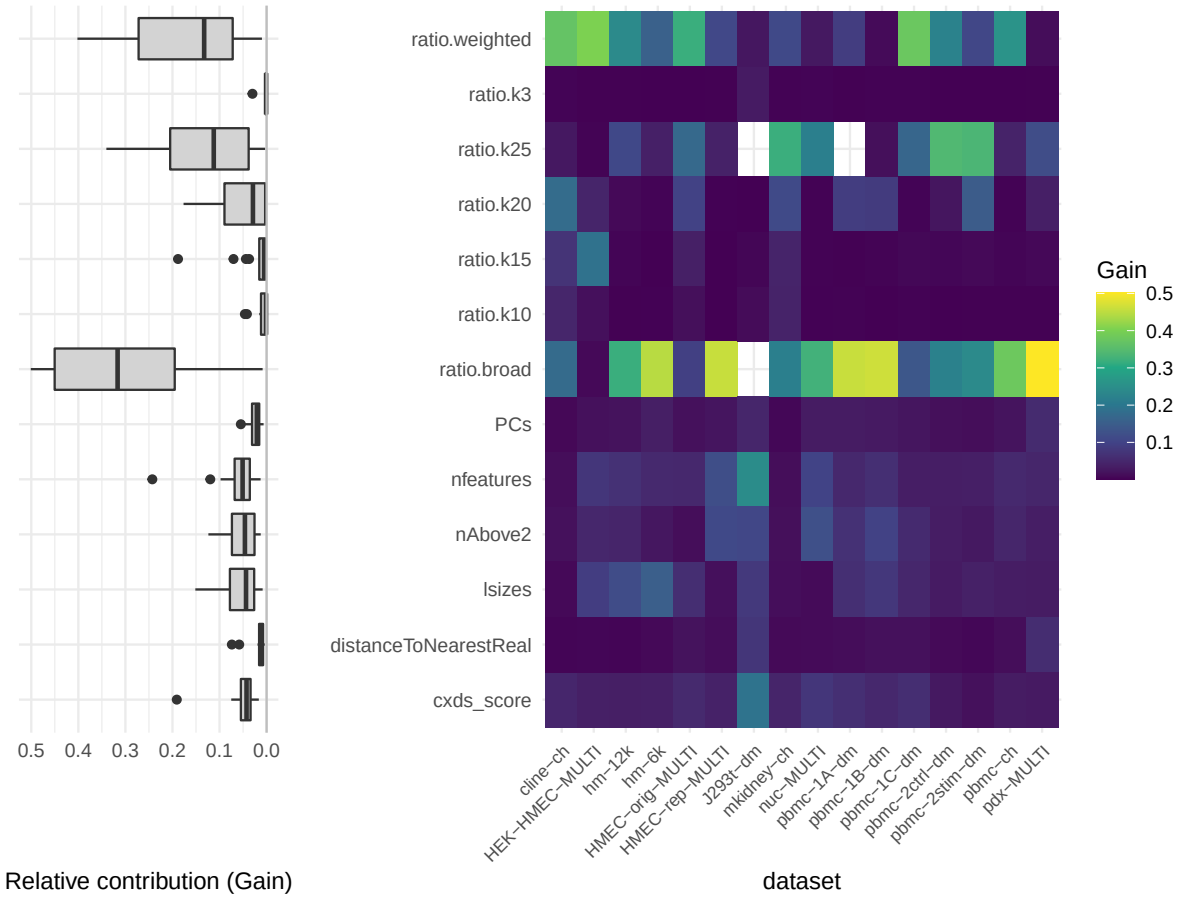
Supplementary Figure 8



Supplementary Figure 8

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.

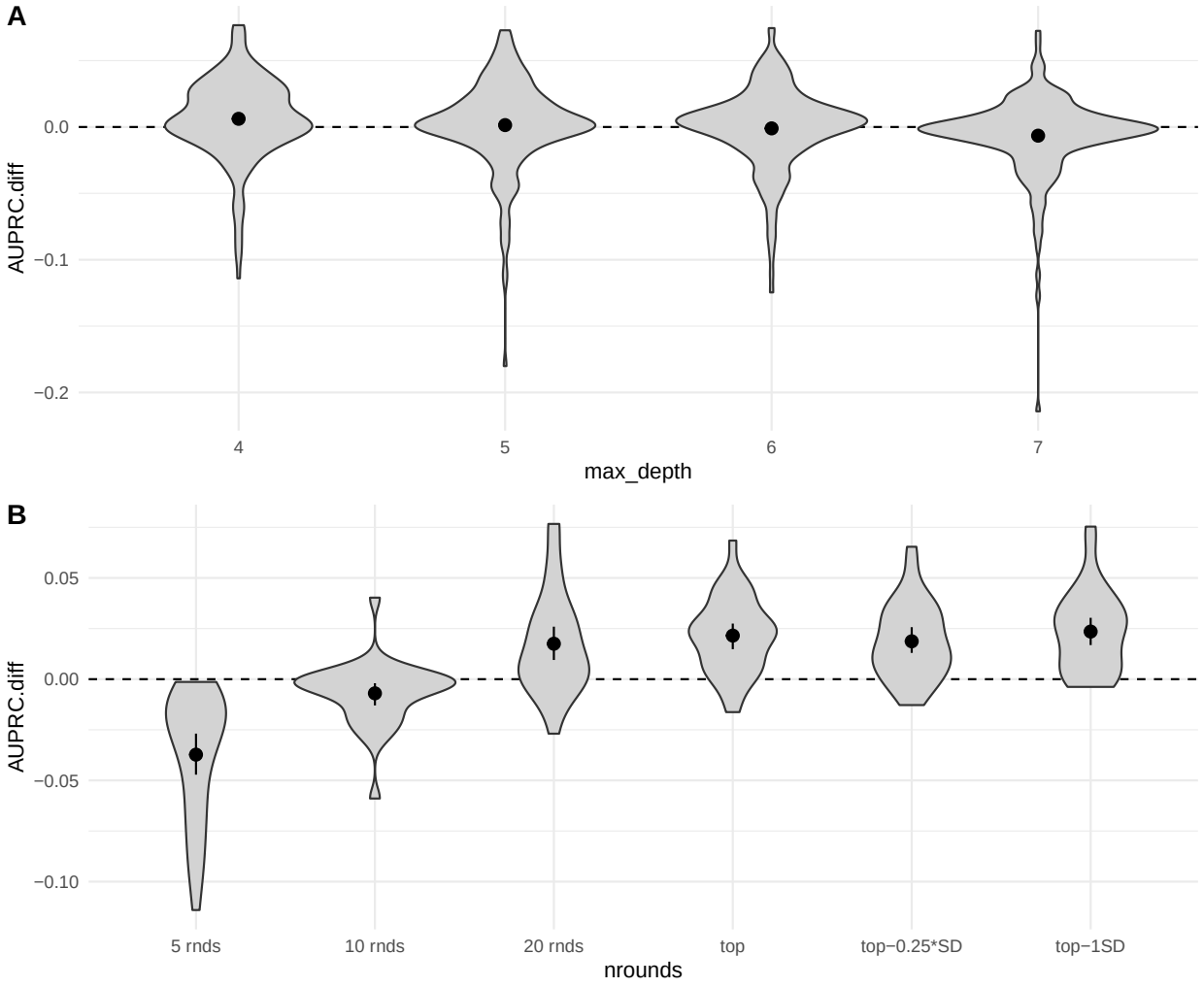
Supplementary Figure 9



Supplementary Figure 9

Variable importance calculated during training. 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, and is not used in small datasets).

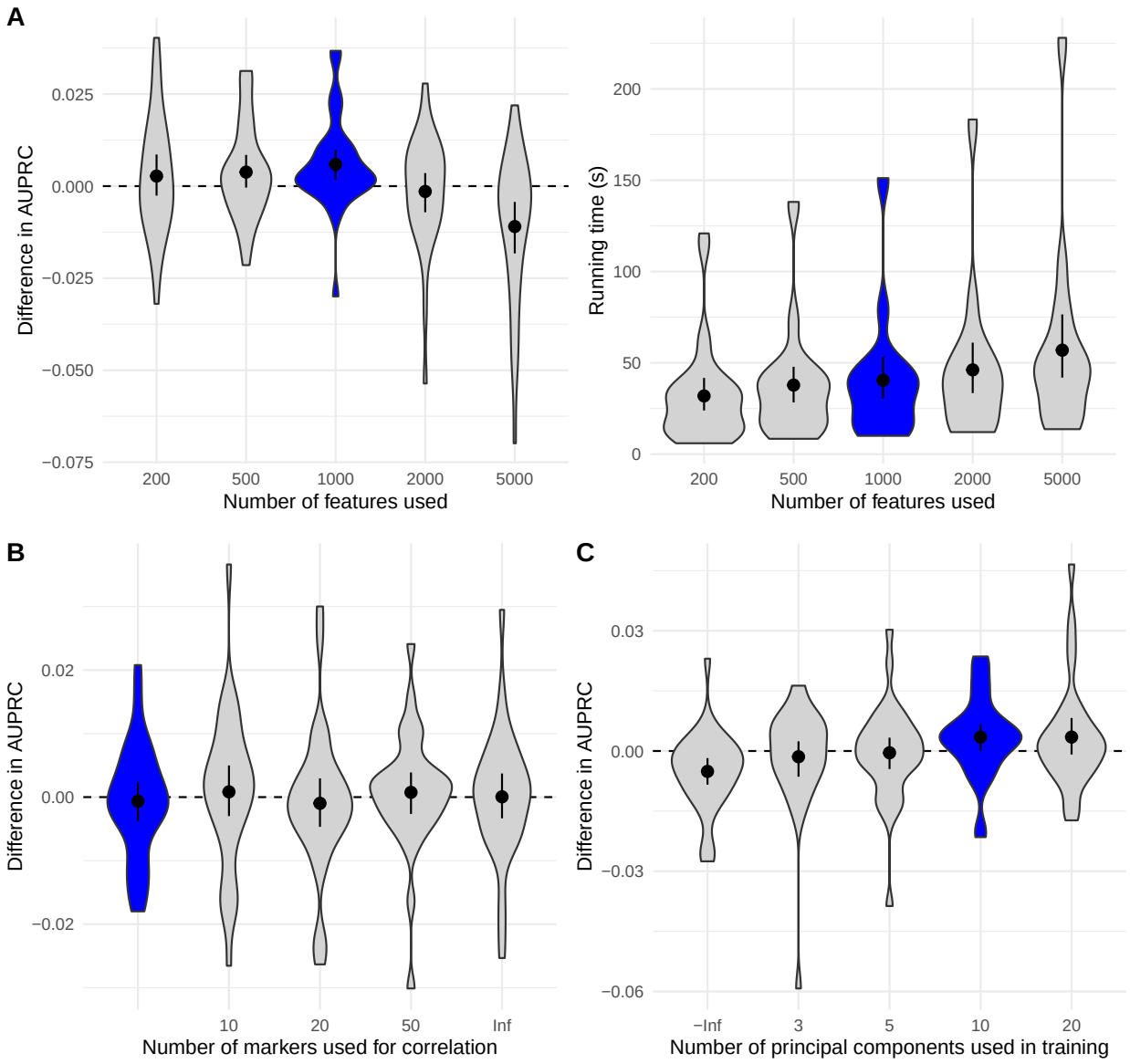
Supplementary Figure 10



Supplementary Figure 10

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 11



Supplementary Figure 11

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).