Cell Systems, Volume *8*

Supplemental Information

DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors

Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner

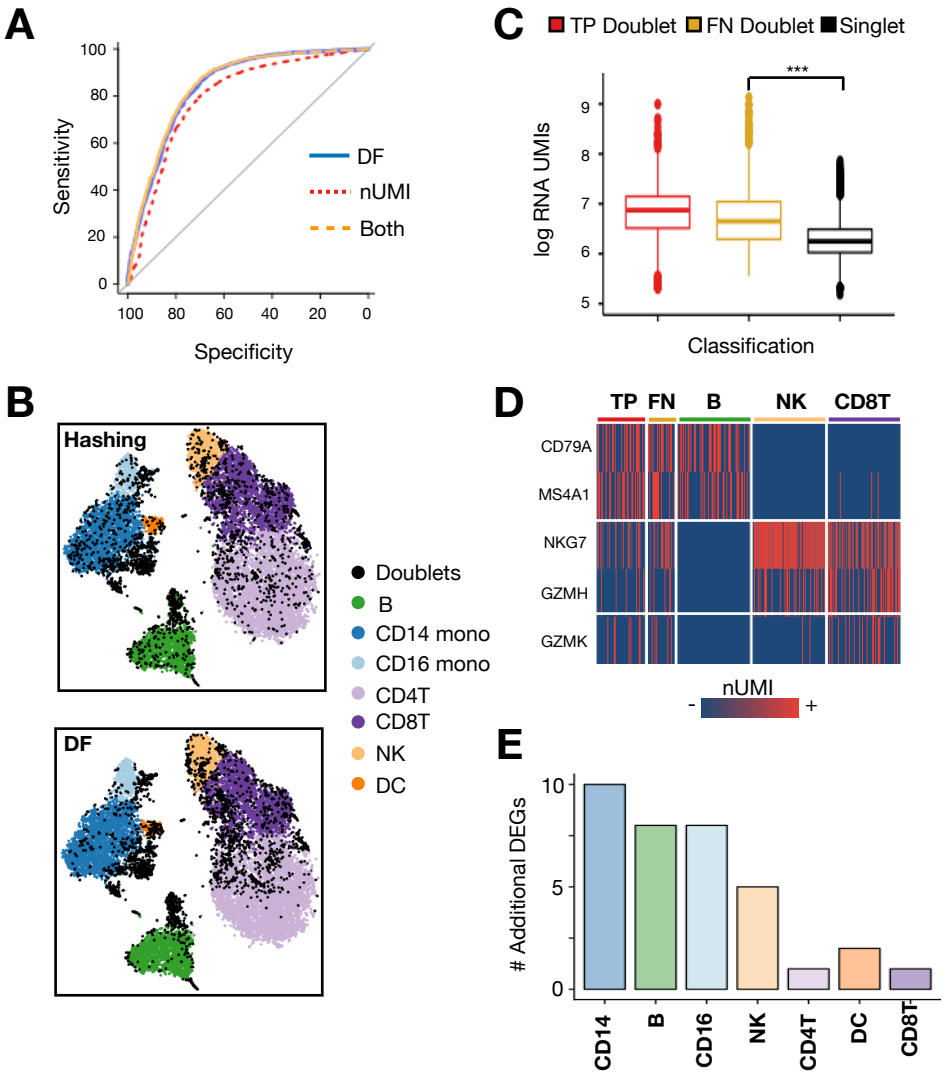


Figure S1: Application of DoubletFinder to Cell Hashing dataset. Related to Figure 1.
(A) ROC analysis of logistic regression models trained using DoubletFinder alone (blue), nUMIs alone (red), and both nUMIs and DoubletFinder (orange).
(B) t-SNE visualizations of Cell Hashing and DoubletFinder doublets (black) amongst PBMC cell types.
(C) RNA UMI box plots for true positive doublets (red), putative false negative doublets (gold), and singlets (black). Data are represented as mean ± SEM. *** = statistically-significant ($p < 2e\text{-}16$).
(D) Marker gene heat maps for true positive doublets, false negative doublets, B cells, NK cells, and CD8+ T-cells.
(E) Bar chart describing the number of additional differentially-expressed genes identified following doublet removal.

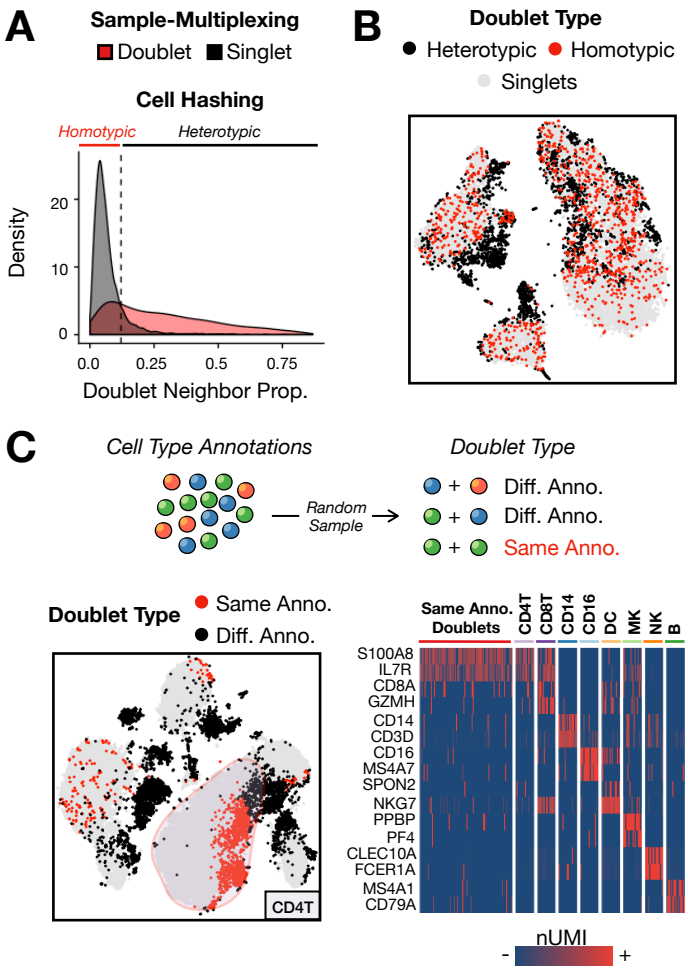


Figure S2: DoubletFinder is insensitive to homotypic doublets reflecting transcriptional divergence but not cell type nomenclature. Related to Figure 1.

(A) Density plots of Cell Hashing data describing the proportion of ground-truth doublet neighbors in gene expression space. Singlets (grey) have low doublet neighbor proportions, whereas doublets (red) have variable doublet neighbor proportions. Homotypic and heterotypic doublets were thresholded at the intersection of single and doublet densities (black dotted line).

(B) t-SNE visualization of Cell Hashing data demonstrates that homotypic doublets (red) localize amongst singlets (grey), unlike heterotypic doublets (black).

(C) Demuxlet cell type annotations were tracked during DoubletFinder artificial doublet generation (top), resulting in artificial doublets formed from cells with the same or different cell type annotations. We predicted which real doublets were made from cells of the same type as doublets where 50% of their nearest neighbors were artificial doublets made from cells with the same cell type annotation. Comparing the location of these doublets (red) versus doublets with artificial nearest neighbors made from cells with different cell type annotations (black) reveals that DoubletFinder detects many CD4T-CD4T doublets. These doublets localize with CD4+ T-cells in gene expression space (bottom left) and exclusively exhibit CD4+ T-cell markers (bottom right).

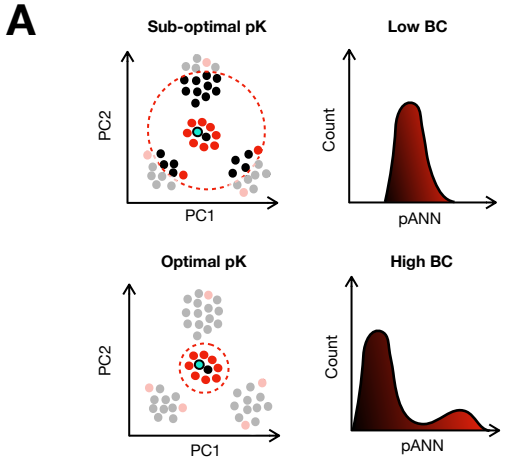
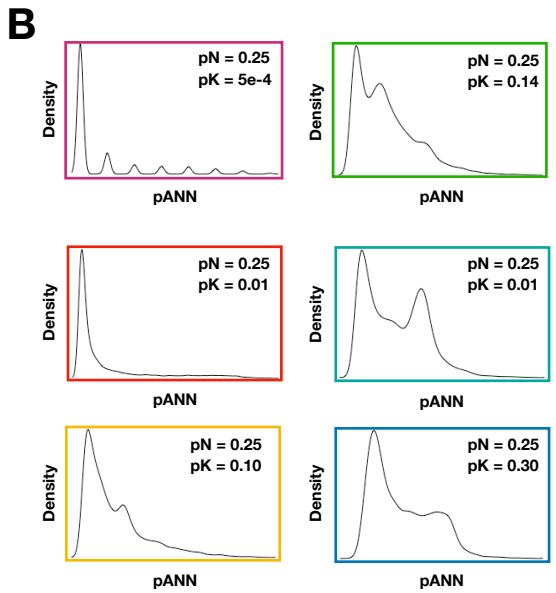


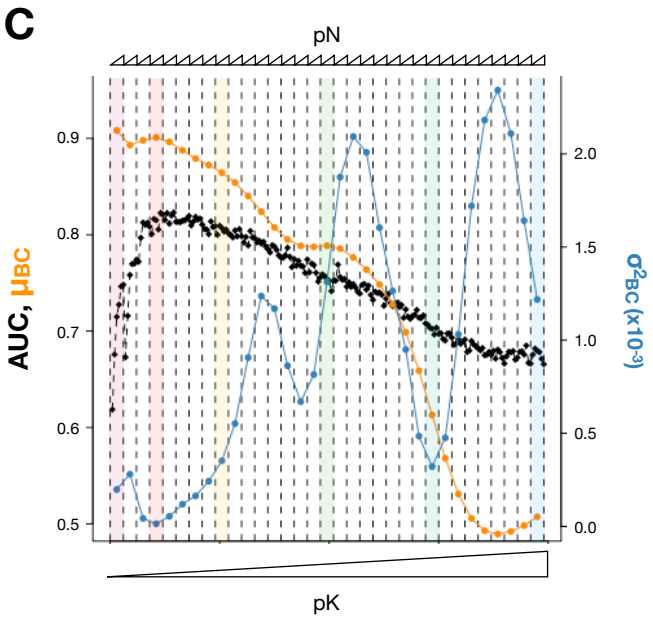
Figure S3: BC_{MVN} overview. Related to Figure 2.

(A) Schematic overview of relationships between neighborhood size (pK) and bimodality coefficient (BC) of pANN distributions. As neighborhood sizes become too large, pANN for real doublets and singlets become more similar, resulting in unimodal pANN distributions with low BC (top). Neighborhood sizes that reflect the structure of clusters in gene expression space correspond to distinct pANN regimes for real singlets and doublets, resulting in non-unimodal pANN distributions with high BC (bottom). pANN gradients correspond to singlets (black) and doublets (red).

(B) Representative pANN distributions across the Cell Hashing pN-pK parameter sweep. Border colors correspond to pK bins highlighted in Fig. S3C.

(C) Maximizing BC mean (μ_{BC}, orange) while minimizing BC variance (σ^2_{BC}, blue) enables identification of the optimal DoubletFinder pK value for Cell Hashing data, as measured using AUC from ROC analysis (black). Highlighted pK bins correspond to pANN distribution borders in Fig. S3B. pK values are separated into pN bins by the black dashed lines.





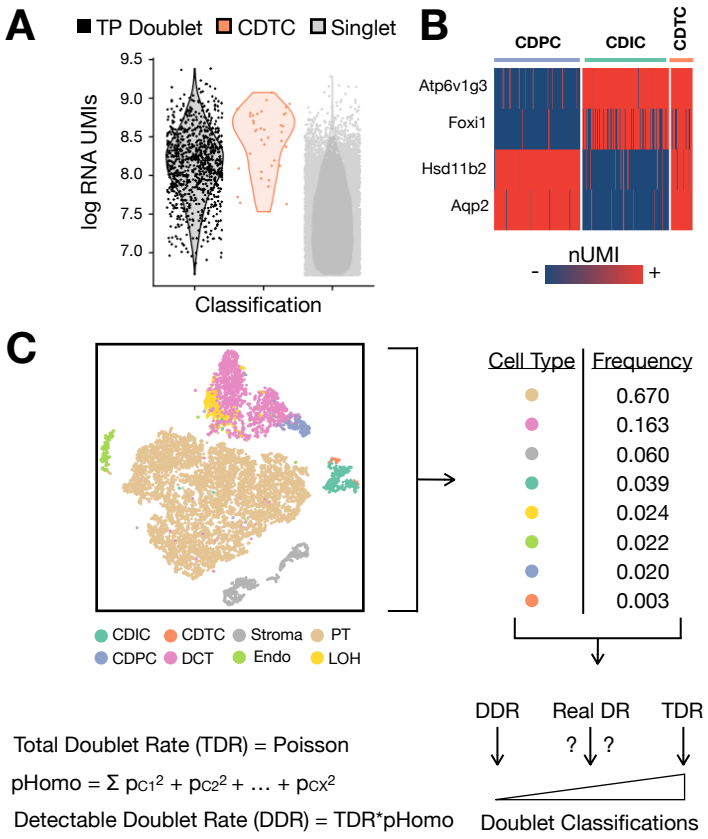


Figure S4: DoubletFinder preservation of 'hybrid' CDTCs enhanced following pANN thresholding adjustment. Related to Figure 2.
(A) Mouse kidney CDTCs (peach) exhibit elevated nUMIs similar to doublets (black) and distinct from singlets (grey).
(B) Mouse kidney CDTCs (peach) co-express marker genes associated with CDPCs (blue) and CDICs (turquoise).
(C) Schematic describing strategy for pANN threshold adjustment to account for homotypic doublets using baseline cell type frequencies. pANN threshold is adjusted to match the detectable doublet rate (DDR), which is computed by multiplying the proportion of homotypic doublets (pHomo) by the total doublet rate (TDR; i.e., Poisson doublet formation rate).