DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors

Christopher S. McGinnis[1], Lyndsay M. Murrow[1] and Zev J. Gartner[1,2,3,4]*
[1]Department of Pharmaceutical Chemistry, University of California, San Francisco
[2]Chan Zuckerbeg Biohub, University of California, San Francisco
[3]Center for Cellular Construction, University of California, San Francisco, San Francisco, CA 94158, USA
[4]Lead Contact
*Correspondence: zev.gartner@ucsf.edu

SUMMARY

Single-cell RNA sequencing (scRNA-seq) using droplet microfluidics occasionally produces transcriptome data representing more than one cell. These technical artifacts are caused by cell doublets formed during cell capture and occur at a frequency proportional to the total number of sequenced cells. The presence of doublets can lead to spurious biological conclusions, which justifies the practice of sequencing fewer cells to limit doublet formation rates. Here, we present a computational doublet detection tool – DoubletFinder – that identifies doublets based solely on gene expression features. DoubletFinder infers the putative gene expression profile of real doublets by generating artificial doublets from existing scRNA-seq data. Neighborhood detection in gene expression space then identifies sequenced cells with increased probability of being doublets based on their proximity to artificial doublets. DoubletFinder robustly identifies doublets across scRNA-seq datasets with variable numbers of cells and sequencing depth, and predicts false-negative and false-positive doublets defined using conventional barcoding approaches. We anticipate that DoubletFinder will aid in scRNA-seq data analysis and will increase the throughput and accuracy of scRNA-seq experiments.

INTRODUCTION

Since its introduction nearly a decade ago, scRNA-seq has been used to elucidate previously unknown cell types and reconstruct developmental dynamics among heterogeneous cell populations (Human Cell Atlas Consortium, 2017). At first, scRNA-seq workflows were

limited to tens to hundreds of cells which hindered data interpretation due to batch effects and low statistical power (Stegle et al., 2016). Today, sequencing thousands to hundreds of thousands of cells is routine due to the advent of droplet microfluidics and nanowell-based sequencing strategies (Macosko et al., 2015; Klein et al., 2015; Zheng et al., 2017; Gierahn et al., 2017; Takara Bio USA, 2018). These techniques rely on a Poisson loading strategy to compartmentalize individual cells and mRNA capture beads before cell lysis, mRNA capture, and transcript barcoding via reverse transcription. Since cells are captured randomly, the proportion of droplets containing >1 cell – known as doublets – scales linearly across an experimentally-relevant range of input cell concentrations (10X Genomics, 2017), justifying the practice of limiting the number of sequenced cells to minimize doublet formation rates.

The confounding effects of doublets in scRNA-seq data are well-appreciated (Ilicic et al., 2016). However, genomic and cellular barcoding techniques for identifying doublets have only recently been developed (Stoeckius et al., 2017; Kang et al., 2018; Gehring et al., 2018; Guo et al., 2018; Rosenberg et al., 2018). In one such strategy, distinct samples receive unique oligonucleotide barcodes delivered by conjugation to antibodies targeting broadly expressed cell-surface antigens. When the barcoded pools are combined and sequenced, doublets can be identified according to the co-occurrence of orthogonal cell 'hashtags' (Stoeckius et al., 2017). In a second strategy, doublets in a pooled population of cells from different individuals are identified by a computational pipeline, Demuxlet, which facilitates doublet inference based on the co-occurrence of mutually-exclusive SNP profiles (Kang et al., 2018).

By detecting doublets, both Demuxlet and Cell Hashing minimize technical artifacts while enabling users to "superload" droplet microfluidics devices for increased scRNA-seq throughput. However, both methods have limitations. First, neither method can identify doublets formed from identically-barcoded cells. Second, neither method is universally applicable across experimental

54 systems, since Demuxlet requires genetically distinct samples and Cell Hashing requires unique
55 antibody-oligonucleotide conjugate panels for the cell types and species of interest. Third,
56 neither method can be used to analyze existing scRNA-seq datasets. For these reasons,
57 computational methods for defining doublets based on gene expression patterns alone are
58 highly desirable.

59 Here, we present DoubletFinder, a computational doublet detection tool that relies solely
60 on gene expression data. Beginning with the observation that doublets cluster separately from
61 singlets in high-dimensional gene expression space (Stoeckius et al., 2017; Kang et al., 2018),
62 we reasoned that real doublets would cluster together with synthetic doublets formed by
63 averaging the expression data of two real cells. By merging artificial doublets with existing
64 scRNA-seq data, we can distinguish doublets from singlets according to the proportion of
65 artificial nearest neighbors (pANN) for each real cell in gene expression space. Thresholding the
66 resulting pANN distribution to match the expected number of doublets provides an accurate
67 metric for doublet prediction that can be applied to any scRNA-seq dataset.

68

69 **RESULTS**

70 <u>DoubletFinder predicts doublets more accurately than nUMIs</u>: Existing strategies for identifying
71 doublets using gene expression features primarily rely on two sources of information. First, since
72 the total number of captured mRNA molecules is expected to be greater for doublets than
73 singlets, doublets are commonly excluded by thresholding cells with high numbers of unique
74 molecular identifiers (nUMIs; Islam et al., 2014; Ziegenhain et al., 2017). While intuitively
75 appealing, technical variability in mRNA capture efficiency and biological variability in mRNA
76 content limits the utility of nUMI-based doublet predictions (Stoeckius et al., 2017). In a second
77 strategy, doublets are removed from scRNA-seq data by identifying groups of cells exhibiting

78 the co-expression of genes with non-overlapping expression patterns *in vivo* (Rosenberg et al.,
79 2018). This strategy cannot be applied to biological systems where such marker genes are
80 unknown, undetected, or unavailable. Moreover, such a strategy could theoretically lead to the
81 erroneous removal of new cell types or developmental states with intermediate expression
82 profiles (Fig. 1A). Given these shortcomings, new methods for predicting doublets using gene
83 expression features alone would greatly benefit the single-cell genomics field.

84 DoubletFinder predicts doublets in a fashion agnostic to nUMIs, marker gene expression,
85 genetic background or exogenous barcodes and can be split into four distinct steps: (1) Generate
86 artificial doublets, (2) Merge real and artificial data and reduce dimensionality with principal
87 component analysis (PCA), (3) Define the nearest neighbors for every real cell in PC space, and
88 (4) Compute and threshold the proportion of artificial nearest neighbors (pANN; Fig. 1B). We
89 tested the efficacy of DoubletFinder against scRNA-seq datasets where doublets are empirically-
90 defined: The publically-available Cell Hashing and Demuxlet datasets comprised of 15,178 and
91 35,524 peripheral blood mononuclear cells (PBMCs), respectively. Demuxlet PBMCs were
92 derived from 8 genetically-distinct human sources while Cell Hashing PBMCs were barcoded
93 with 8 distinct antibody-oligonucleotide conjugate panels. Using optimized input parameters
94 (Supplementary Materials, Fig. S1), we tested whether pANN outperforms nUMIs as a doublet-
95 prediction feature by using receiver operating curve (ROC) analysis to compare logistic
96 regression models trained on the Cell Hashing data using pANN alone, nUMIs alone, or both
97 (Fig. 1C). ROC analysis demonstrates that pANN predicts doublets more accurately than nUMIs.
98 Moreover, the model trained with both features performed nearly indistinguishably to the pANN-
99 alone model, suggesting that DoubletFinder captures all of the doublet-specific information
100 inherent to nUMIs.
101

102 DoubletFinder predicts Cell Hashing doublets: To make specific doublet predictions for each
103 cell, DoubletFinder rank-orders cells by their pANN values and thresholds this list according to
104 the number of expected doublets. To test the robustness of pANN thresholding, the number of
105 expected doublets was determined using two different strategies (Fig. 1D). First, since 8 samples
106 were multiplexed in the Cell Hashing and Demuxlet studies, we reasoned that 1/8 of the true
107 doublets were undetected because they were formed from genetically-identical or identically-
108 barcoded cells. Thus, we thresholded pANN according to the number of detected doublets with
109 an assumed 12.5% false negative rate (β). Second, since the doublet formation rate can be
110 accurately estimated by applying Poisson statistics to the number of cells loaded into the droplet
111 microfluidics device (10X Genomics, 2017), we thresholded pANN according to this rate. For
112 standard scRNA-seq experiments where doublets are not empirically-defined, pANN can only
113 be thresholded using the Poisson strategy.

114 Depending on the threshold used, DoubletFinder predicted 2680 or 2155 doublets when
115 applied to the full Cell Hashing dataset. Single-cell gene expression data was visualized using
116 t-stochastic neighborhood embedding (t-SNE; van der Maaten and Hinton, 2008) and cells were
117 colored according to their real and predicted doublet status (Fig. 1E). Visual comparison of
118 doublets in t-SNE space illustrates that DoubletFinder predictions closely track Cell Hashing
119 results. This result is further supported by the observation that the frequency of DoubletFinder
120 predictions is highly enriched in Cell Hashing-defined doublet groups relative to singlet groups
121 (Fig. 1F), regardless of the thresholding strategy used.

122

123 DoubletFinder predicts Cell Hashing false-negatives: DoubletFinder predictions exhibit a less
124 'speckled' appearance in t-SNE space relative to the Cell Hashing results (Fig. 2A, insets).
125 Considering that doublets often cluster separately from singlets in gene expression space

126 (Stoeckius et al., 2017; Kang et al., 2018), we reasoned that cells called as singlets via Cell
127 Hashing, that nonetheless co-cluster with high-confidence doublets, are actually false-negatives
128 derived from identically-barcoded cells. Two main predictions follow from this line of reasoning.
129 First, if the putative false negatives are truly doublets, then they should exhibit gene expression
130 patterns associated with distinct cell types. In line with this prediction, Cell Hashing-defined
131 doublets and singlets in the highlighted region express marker genes for both B cells and NK
132 cells (Fig. 2B) – hematopoietic cell types that do not share a common progenitor in peripheral
133 blood. Second, since false negative cells would be associated with the combined barcodes of
134 two cells, the nUMI counts for the most abundant barcode should be significantly higher in false
135 negatives than high-confidence doublets. Moreover, since false negatives would not be
136 associated with high levels of multiple barcodes, the second most abundant barcode should be
137 similar to high-confidence singlets and significantly lower relative to high-confidence doublets.
138 Statistical analysis supports these predictions (Wilcoxon rank sum test, $p < 10^{-13}$; Fig. 2C).
139 Collectively, these results demonstrate that DoubletFinder robustly recapitulates doublet
140 assignments and accurately predicts Cell Hashing false-negatives.
141
142 <u>DoubletFinder predicts Demuxlet doublets and identifies putative false-positives</u>: To test whether
143 DoubletFinder performance is sensitive to changes in the number of sequenced cells and
144 sequencing depth, we applied DoubletFinder to the Demuxlet dataset. In addition to having more
145 cells than the Cell Hashing data, the average number of UMIs (2408 vs 676) and genes (837 vs
146 376) per cell is also greater in the Demuxlet data. In line with our previous results, visual
147 comparison of real and predicted doublets using t-SNE illustrates that DoubletFinder
148 successfully identifies all doublet-enriched regions in gene expression space (Fig. 2D).

149 As with our Cell Hashing comparison, there were a number of regions in gene expression
150 space where DoubletFinder predictions differed from Demuxlet classifications. Specifically, there
151 were many DoubletFinder-defined doublets called as singlets by Demuxlet that give doublet-
152 enriched clusters the 'speckled' appearance discussed above. Moreover, in contrast to the Cell
153 Hashing comparison, there was a subset of cells classified as doublets by Demuxlet and singlets
154 using DoubletFinder (Fig. 2D, insets). Interestingly, the majority of these discordant calls are
155 scattered amongst high-confidence singlet clusters in gene expression space. This observation
156 can be explained by two alterative models. In one model, these discordant calls are caused by
157 homotypic doublets – i.e., doublets formed from cells of the same type – which presumably have
158 a similar transcriptional profile to singlets and, thus, would be more difficult for DoubletFinder to
159 detect relative to heterotypic doublets. Alternatively, the discordant calls are due to false-positive
160 Demuxlet classifications.

161 If these cells were in fact homotypic doublets left undetected by DoubletFinder, then one
162 would expect that DoubletFinder was insensitive to homotypic doublets throughout the Demuxlet
163 dataset. To test this possibility, we tracked the cell types comprising each artificial doublet and
164 deconvolved pANN values into homotypic and heterotypic components. Visualization of cells
165 with majority homotypic or heterotypic nearest neighbors highlights a region of homotypic
166 doublets formed from $CD4^+$ T-cells (Fig. 2E; Supplementary Materials, Fig. S2), which suggests
167 that DoubletFinder has the sensitivity to detected certain classes of homotypic doublets.
168 Moreover, the scattering of discordant doublet classifications amongst high-confidence singlet
169 clusters in gene expression space is also evident in Demuxlet classifications of the Cell Hashing
170 data (Supplementary Materials, Fig. S2). Cell Hashing is sensitive to homotypic doublets, which
171 suggests that the discordant calls are not a consequence of homotypic doublets missed by
172 DoubletFinder. Finally, if the putative false-positive calls are homotypic doublets, one would

173 expect the number of RNA UMIs to approximate levels observed for high-confidence doublets.
174 Interestingly, the RNA nUMI distribution for putative false-positives is nearly indistinguishable to
175 high-confidence singlets (Wilcoxon rank sum test, $p = 0.34$), while high-confidence doublets and
176 putative false-negatives are both significantly enriched for RNA nUMIs (Wilcoxon rank sum test,
177 $p < 10^{-15}$; Fig. 2F). While it is difficult to definitively ascertain the ground truth for these discordant
178 calls, these results collectively demonstrate that DoubletFinder is robust across a range of cell
179 numbers and sequencing depths and prospectively predicts Demuxlet false-positives.
180

181 **DISCUSSION**

182 High-throughput scRNA-seq suffers from the formation of doublets due to the inherent
183 nature of Poisson cell loading. Doublets can lead to spurious conclusions during analysis when
184 left unidentified because the resulting artefactual expression data may be interpreted as
185 previously-undescribed cell types, developmental intermediates, or disease states. As a result,
186 it has become common practice to minimize the doublet formation rate by minimizing the ratio
187 of sequenced cells to mRNA capture beads. Although recent advances in direct doublet
188 detection methodologies have proven to be effective, they are not universally or retroactively
189 applicable. For this reason, complementary techniques for predicting doublets based only on
190 gene expression data have the potential to further increase scRNA-seq throughput while
191 removing technical artifacts.

192 Towards this goal, DoubletFinder accurately identifies doublets in scRNA-seq data by
193 integrating artificial doublets into real data and computing the pANN for every real cell. We have
194 shown that DoubletFinder distinguishes real doublets from singlets better than nUMIs in the Cell
195 Hashing dataset. Moreover, we demonstrate that DoubletFinder accurately predicts doublets for
196 two independent PBMC scRNA-seq datasets of different sizes and sequencing depths. As these

197 are the only publically available data with empirically-defined doublets, it is unclear whether
198 DoubletFinder will require further optimization for scRNA-seq datasets describing different
199 tissues or biological systems. We have also shown that DoubletFinder identifies false-negative
200 and putative false-positive doublet classifications present in these datasets, which supports the
201 use of DoubletFinder in concert with Cell Hashing, Demuxlet, and other barcoding approaches.
202 Finally, we demonstrate that DoubletFinder performs robustly with pANN thresholding strategies
203 that differed by >5000 cells (Fig. 1D). This suggests that DoubletFinder can be applied in
204 experimental contexts where doublet formation rates differ significantly from industry estimates
205 – e.g., clumpy single-cell suspensions or especially cohesive cell types. Collectively,
206 DoubletFinder represents a fast, easy-to-use doublet detection strategy that will aid the single-
207 cell genomics community in data analysis and enable high-throughput scRNA-seq technologies
208 to be utilized to their fullest potential.
209

210 **MATERIALS & METHODS**

211 DoubletFinder Overview: Artificial doublets were generated from raw UMI count matrices via
212 random sampling of cell expression profiles without replacement before pre-processing using
213 the 'Seurat' R package, as described previously (Butler et al., 2018). Notably, no sources of
214 variation were regressed out of the merged data before PCA, and the top 10 PCs – chosen via
215 inflection point estimation on the corresponding elbow plot – were used to define the Euclidean
216 distance matrix using the 'dist' R function.
217

218 ROC Analysis: ROC analysis-based model comparisons were performed using the 'ROCR'
219 (Sing et al., 2005) and 'pROC' (Robin et al., 2011) R packages. Briefly, logistic regression
220 models were defined on a training set comprising half the total data using the 'glm' R function

221 with the link argument set to 'logit'. These models were then used to create a vector describing
222 each cell's doublet probability with the 'predict' R function. ROC analysis was then performed by
223 calculating the sensitivity and specificity of doublet predictions based on the aforementioned
224 probability vector at varying probability thresholds. The AUC was then calculated for the resulting
225 curve, and AUC was used as a proxy for doublet detection model performance.
226
227 Statistical Analysis: Statistically-significant differences between UMI levels were defined using
228 the Wilcoxon rank sum test implemented with the 'pairwisewilcox.test' R function. Multiple
229 comparison correction was performed using the Benjamini-Hochberg procedure.
230
231 Data Availability: Cell Hashing (GEO: GSE108313) and Demuxlet (GEO: GSE96583) UMI count
232 matrices were downloaded from the Gene Expression Omnibus. DoubletFinder is implemented
233 as a fast, easy-to-use R package that interfaces with Seurat version 2.0 and higher.
234 DoubletFinder can be downloaded from GitHub (github.com/chris-mcginnis-ucsf/DoubletFinder).
235

236 **ACKNOWLEDGMENTS**

237 We thank Matt Thomson (California Institute of Technology) for helpful discussion, as well as
238 Jimmie Ye (UCSF), Marlon Stoeckius and Shiwei Zheng (New York Genome Center) for
239 providing data access.
240

241 **AUTHOR CONTRIBUTIONS**

242 C.S.M., L.M.M., and Z.J.G conceptualized the method and wrote the manuscript. C.S.M. wrote
243 the software and performed all bioinformatics analyses.
244

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

FUNDING

This research was supported in part by grants from the Department of Defense Breast Cancer Research Program (W81XWH-10-1-1023 and W81XWH-13-1-0221), the NIH Common Fund (DP2 HD080351-01), the NSF (MCB-1330864), and the UCSF Center for Cellular Construction (DBI-1548297), an NSF Science and Technology Center. Z.J.G is a Chan-Zuckerberg Biohub Investigator. L.M.M is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-2239-15).

REFERENCES

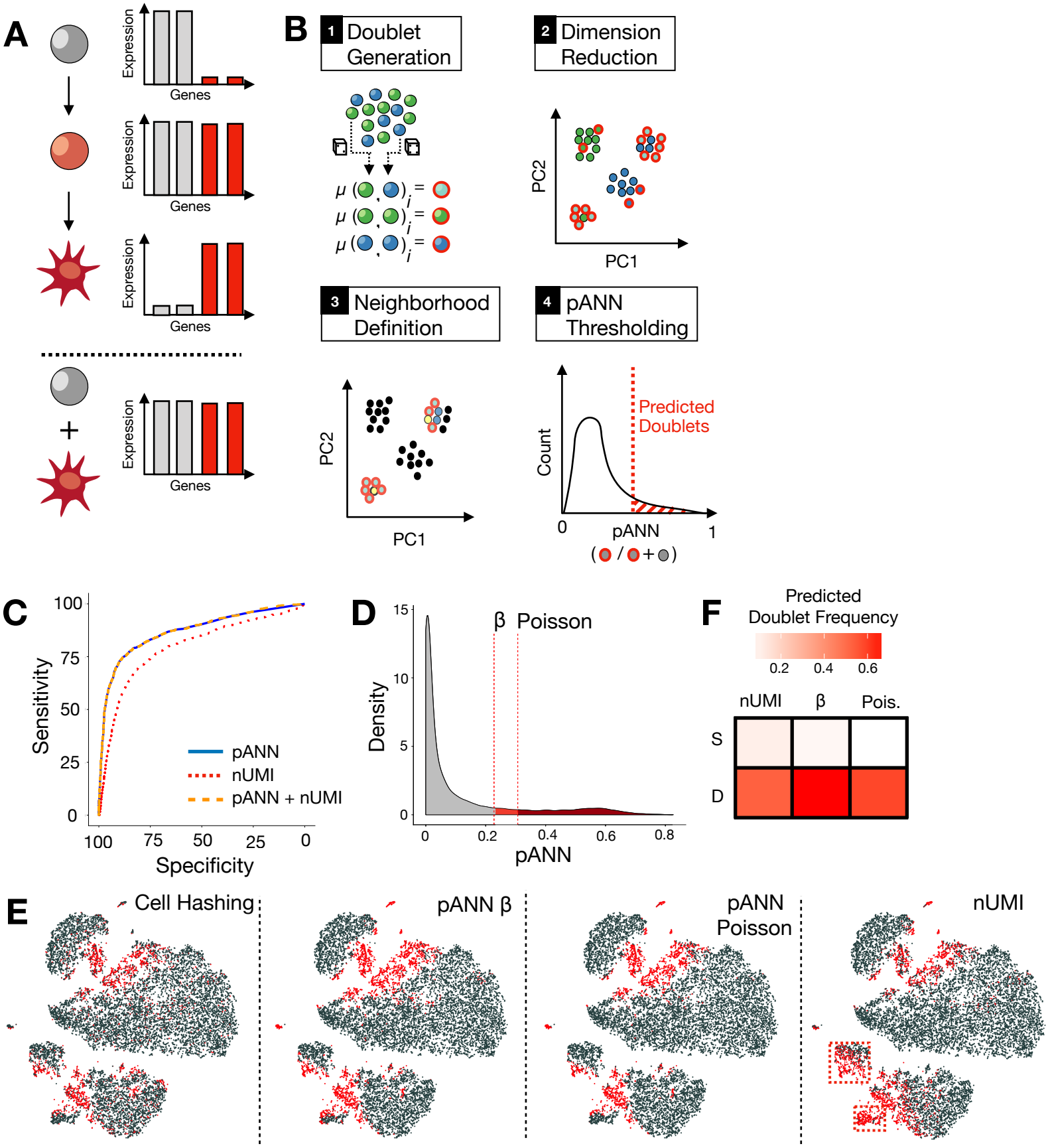
1. The HCA Consortium. The Human Cell Atlas White Paper. 2017. Retrieved from: humancellatlas.org/files/HCA_WhitePaper_18Oct2017.pdf
2. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015; 16(3):133-45.
3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. 2015. Cell; 161(5):1202-1214.
4. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. 2015. Cell; 161(5):1187-1201.
5. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8:14049.
6. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, Satija R, *et al*. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017; 14(4):395-398.
7. Takara Bio USA. Full-length, single-cell RNA-seq with the SMARTer™ ICELL8® Single-Cell System (Application note 633878). Mountain View, CA: Takara Bio. 2018. Available from: clontech.com/US/Products/cDNA_Synthesis_and_Library_Construction/NGS_Automation/Single-Cell_Automation/Single-Cell_Automation_Overview?sitex=10020:22372:US
8. 10X Genomics. ChromiumTM Single Cell 3' Reagent Kits v2 User Guide. Pleasanton, CA: 10X Genomics. 2017.

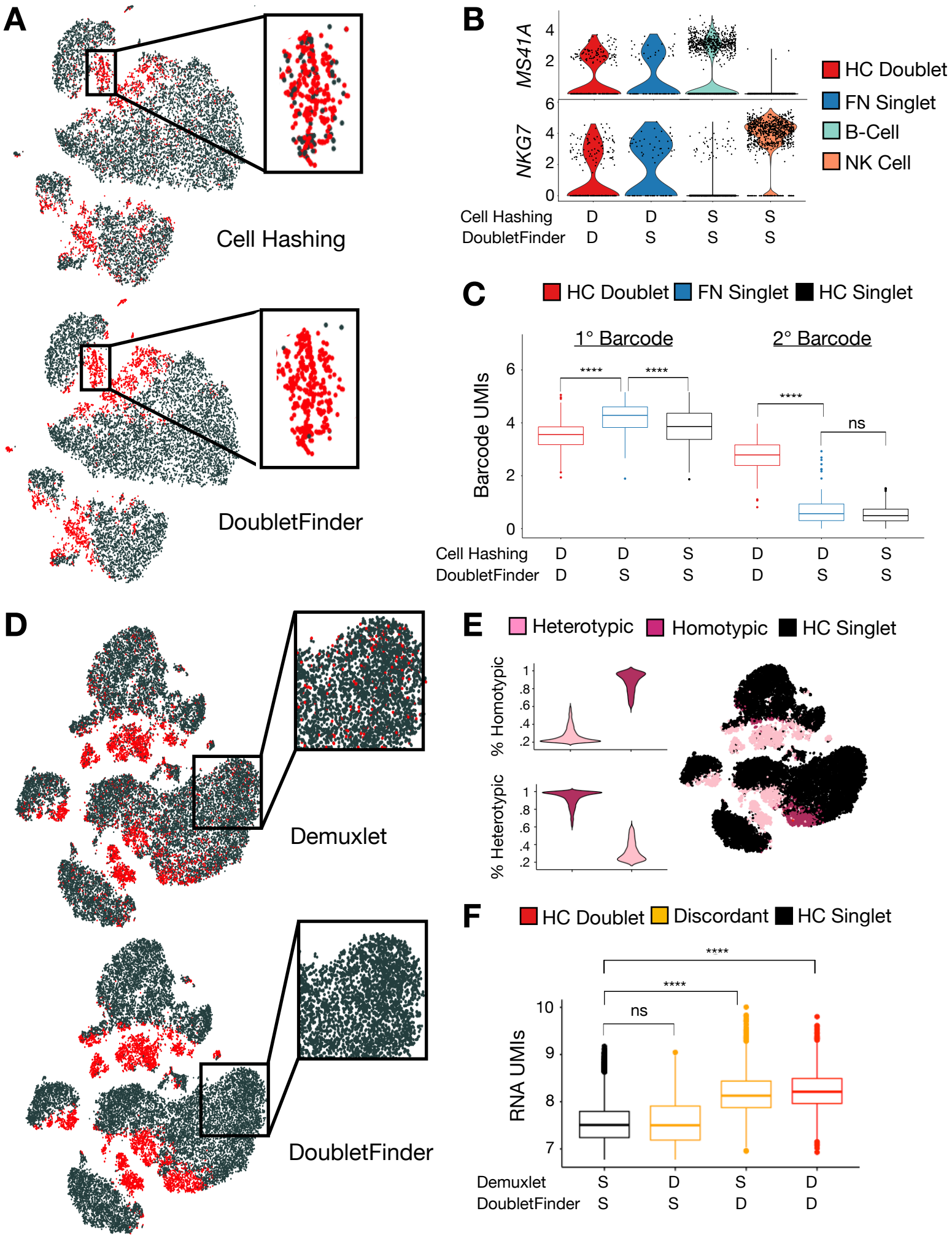
9. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 17:29.
10. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Smibert P, Satija R. Cell "hashing" with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. 2017. Preprint. bioRxiv doi: 10.1101/237693.
11. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018; 36(1):89-94.
12. Gehring J, Park JH, Chen S, Thomson M, Pachter L. Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. 2018. Preprint: bioRxiv doi: 10.1101/315333.
13. Guo C, Biddy BA, Kamimoto K, Kong W, Morris SA. CellTag Indexing: a genetic barcode-based multiplexing tool for single- cell technologies. 2018. Preprint: bioRxiv doi: 10.1101/335547.
14. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018; 360(6385):176-182.
15. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014; 11(2):163-6.
16. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017. 65(4):631-643.e4.
17. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018; doi: 10.1038/nbt.4096.
18. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 2008; 9: 2579–2605.
19. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21(20): 3940-1.
20. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 2011; 12: 77.

FIGURE LEGENDS

Figure 1: DoubletFinder robustly predicts Cell Hashing doublets and outperforms nUMI thresholding. (A) Schematic describing importance of doublet detection. Developmental intermediates (light red) can express genes associated with both progenitor (grey) and differentiated (dark red) cell types. Doublets formed from progenitor and mature cells may mimic the expression profile of intermediate cell states, and thereby confound analysis. (B) Schematic of DoubletFinder workflow. After artificial doublet (red outline) generation, the proportion of artificial nearest neighbors is defined for every real cell (examples highlighted yellow). These results are thresholded to define doublet predictions. (C) Density plot of pANN values with red dotted lines denoting expected doublet thresholds. Histogram is colored according to whether the cells were called as singlets (grey) or doublets by one (light red) or both thresholding strategies (dark red). (D) t-SNE visualization of real and predicted Cell Hashing doublets (red) and singlets (grey), DoubletFinder predictions, and nUMI thresholding predictions. nUMI predictions deviate significantly from Cell Hashing results (red dashed boxes). (E) Heat map showing the proportion of DoubletFinder and nUMI-predicted doublets present in Cell Hashing singlet (S) and doublet (D) groups. (F) ROC analysis of logistic regression models trained using nUMIs alone (dotted red), pANN alone (solid blue) and both nUMIs and pANN (dashed orange) as features.

Figure 2: DoubletFinder detects false-negative and false-positive doublet predictions in Cell Hashing and Demuxlet datasets. (A) t-SNE visualization of real and predicted Cell Hashing doublets (red) and singlets (grey) highlighting a doublet-enriched, 'speckled' region. (B) Violin plots describing the distribution of marker gene expression in high-confidence doublets (red), putative false-negatives (blue), and singlet B-cells (teal) and NK cells (orange). (C) Barcode UMI box plots for the 1st and 2nd most abundant barcodes in high-confidence singlets (black) and doublets (red), as well as false-negative singlets (blue). (D) t-SNE visualization of real and predicted Demuxlet doublets (red) and singlets (grey) highlighting putative false-positives. (E) Violin plots describing the contributions of homotypic and heterotypic doublets to each DoubletFinder-defined doublet's nearest neighborhood. t-SNE visualization of putative homotypic and heterotypic doublet clusters in gene expression space. (F) RNA UMI box plots for high- confidence singlets (black) and doublets (red) as well as discordant doublet predictions between Demuxlet and DoubletFinder (orange).


A
Expression
Genes
B
1 Doublet Generation
2 Dimension Reduction
PC2
PC1
3 Neighborhood Definition
4 pANN Thresholding
Count
Predicted Doublets
0
pANN
1
C
Sensitivity
Specificity
pANN
nUMI
pANN + nUMI
D
Density
β
Poisson
pANN
F
Predicted Doublet Frequency
nUMI
β
Pois.
S
D
E
Cell Hashing
pANN β
pANN Poisson
nUMI


A
Cell Hashing
DoubletFinder
B
MS41A
NKG7
HC Doublet
FN Singlet
B-Cell
NK Cell
Cell Hashing D D S S
DoubletFinder D S S S
C
HC Doublet FN Singlet HC Singlet
1° Barcode
2° Barcode
Barcode UMIs
****
****
****
ns
Cell Hashing D D S D D S
DoubletFinder D S S D S S
D
Demuxlet
DoubletFinder
E
Heterotypic Homotypic HC Singlet
% Homotypic
% Heterotypic
F
HC Doublet Discordant HC Singlet
RNA UMIs
****
****
ns
Demuxlet S D S D
DoubletFinder S S D D