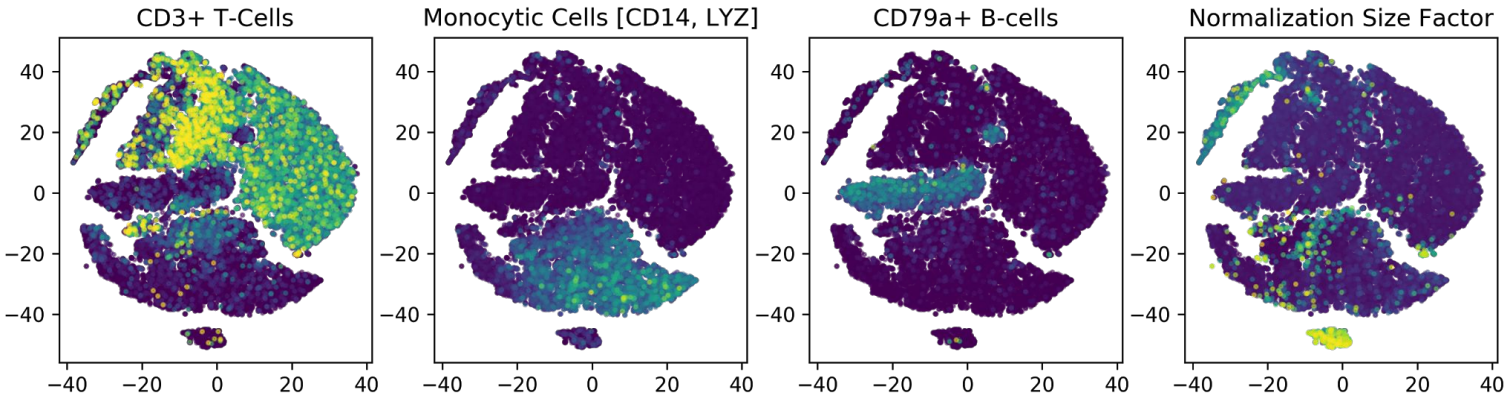
Task 2: Doublets

Ambrose Carr, John Marioni, Dana Pe’er

PBMC Dataset Contains Different Cell Types



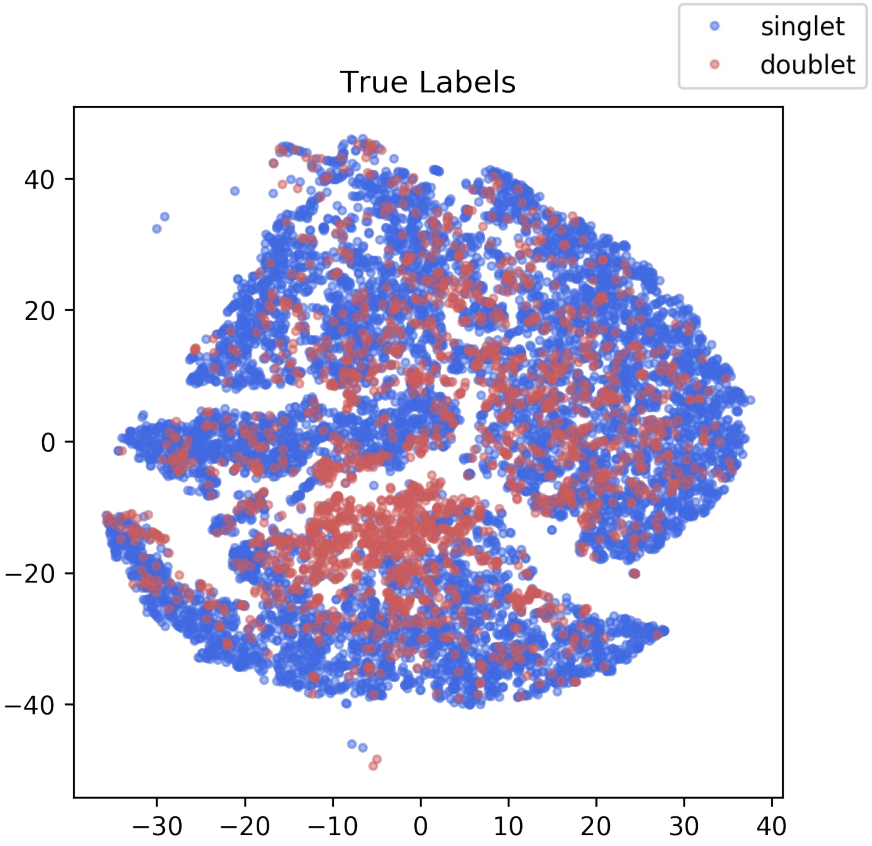
Test Dataset

11010 total cells in the test set

2143 assigned doublets (based on genotype)

8867 assigned singlets

Reasonable dispersed across sub-populations



Classifier Summary

Group	Classifier	Salient Features
5	Logistic Regression	Entropy over LDA posteriors
6	Logistic Regression	Saturation Features
7	Random Forest	
8	KNN	
9.1	Random Forest	
9.2	Random Forest	
9.3/4	GMM / Clustering	

Classifier Summaries

- Group 5:
  - Logistic Regression - 6-7 features (# genes, # umis, Entropy measure (variance of posterior probability within LDA assignments -- doublets have lower variance, singlets dominated by lower cell type)
- Group 6:
  - Logistic regression:
  - Saturation curve for each cell -- estimate #genes and UMIs in a cell -- estimate expected #total reads
  - Lots of saturation statistics for genes and reads (correlated) # proportion of redundant UMIs, # reads / umi, fraction of UMIs with same sequence
  - Coexpression marker scores
- Group 8:
  - Simulated doublets KNN classifier (gene expression, total counts, highly variable genes, z-score normalize at gene level, run PCA)
  - Total counts + GLM
- Group 7:
  - Random Forest
  - Technical features: UMIs, # genes, # fraction of genes with single UMI, Std deviation inside one cell, PCA in gene expression (size norm, top 100 overdispersed genes, scaled, and centered 20 PCs)
- Group 9.1
  - Random forest, 400 variable genes 301 trees # genes, # umis, entropy
- Group 9.2
  - Random forest: Set of low-variance genes (from GTex) added bam file metrics (7+), all genes (z-scores) total UMIs (random forest >> other classifiers for ROC)
- Group 9 Ch1 & 2:
  - Expression - Clustered cells, fit gaussian mixture module w/ equal variance on # genes detected (within cluster). - any cell assigned to larger mean is a doublet.

Performance!

FN: assigned doublet, called as a singlet

FP: assigned singlet, called as a doublet (could include two (or more) cells with the same genetic background

	TN	FN	FP	TP	TPR	TNR
5.0	0.75	0.07	0.06	0.13	0.66	0.93
6.0	0.74	0.06	0.06	0.14	0.71	0.92
7.0	0.74	0.02	0.07	0.17	0.87	0.92
8.0	0.74	0.03	0.07	0.16	0.82	0.91
9.1	0.79	0.06	0.02	0.13	0.68	0.98
9.2	0.77	0.08	0.03	0.11	0.57	0.96
9.3	0.74	0.10	0.06	0.10	0.50	0.92
9.4	0.70	0.09	0.11	0.11	0.56	0.86

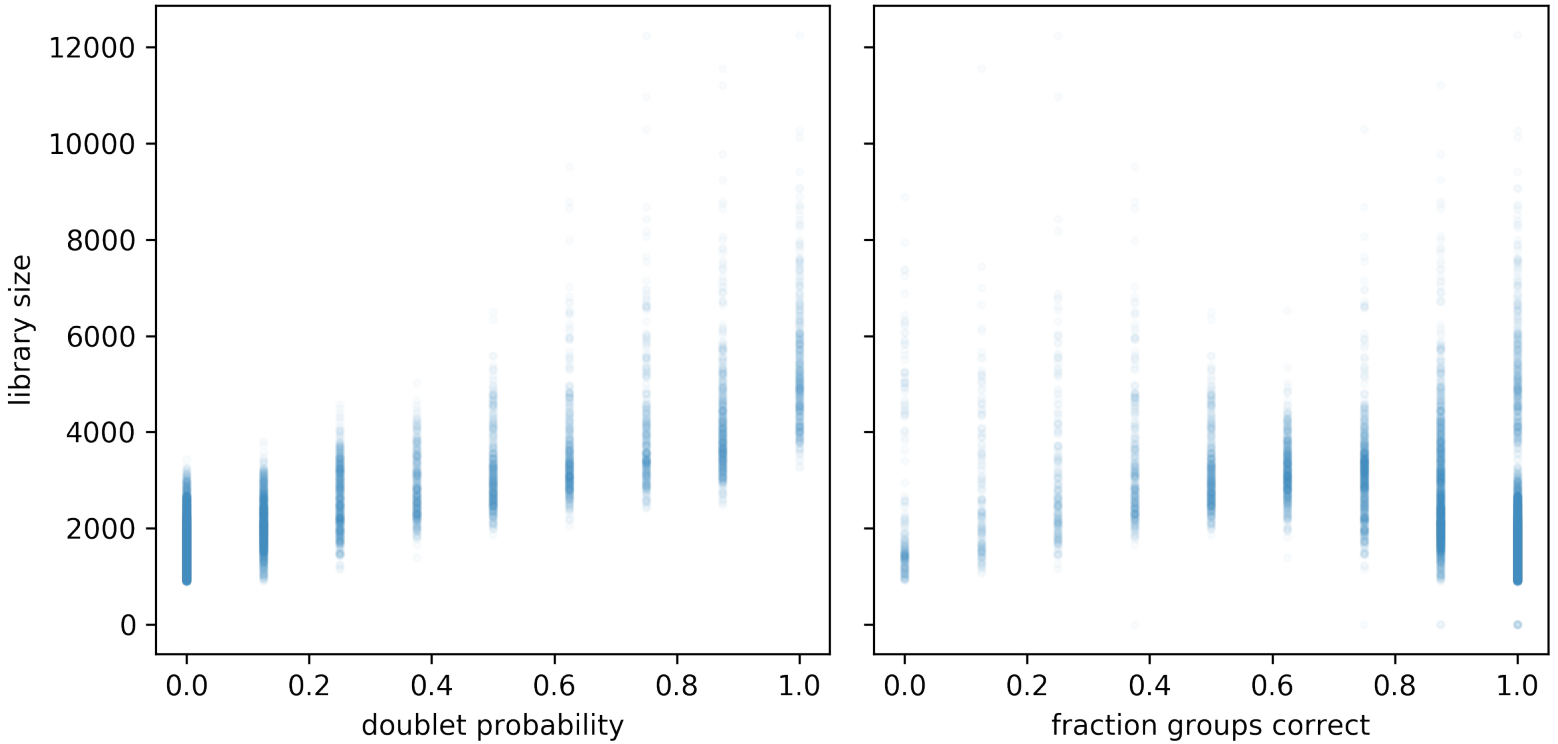
Performance!

Most specific, but most false negatives

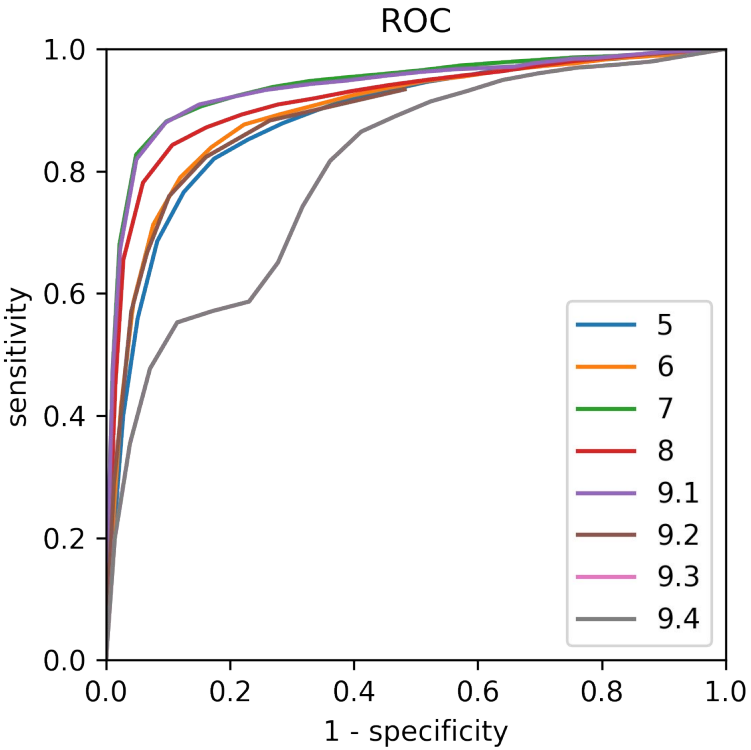
Most sensitive

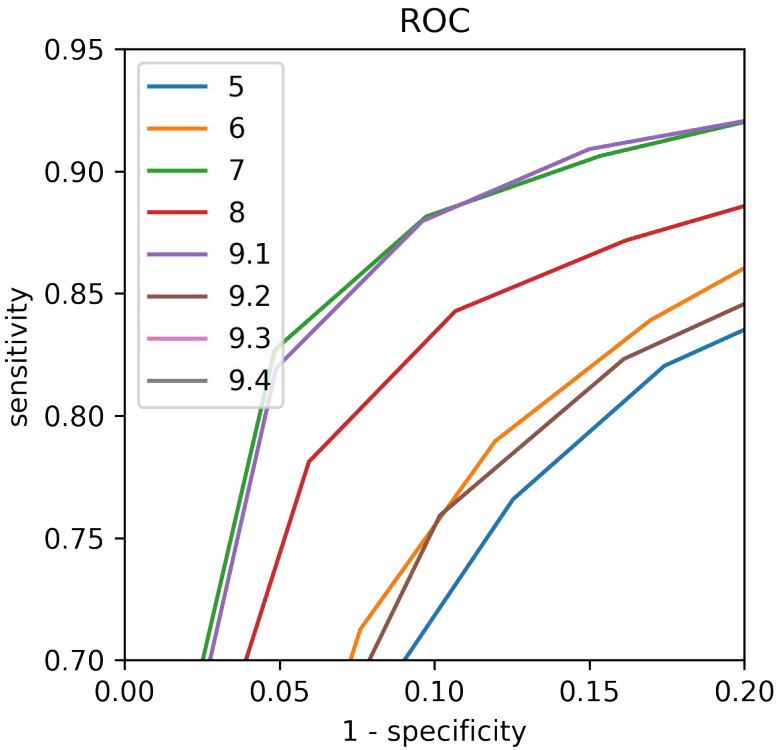
	TN	FN	FP	TP	TPR	TNR
5.0	0.75	0.07	0.06	0.13	0.66	0.93
6.0	0.74	0.06	0.06	0.14	0.71	0.92
7.0	0.74	0.02	0.07	0.17	0.87	0.92
8.0	0.74	0.03	0.07	0.16	0.82	0.91
9.1	0.79	0.06	0.02	0.13	0.68	0.98
9.2	0.77	0.08	0.03	0.11	0.57	0.96
9.3	0.74	0.10	0.06	0.10	0.50	0.92
9.4	0.70	0.09	0.11	0.11	0.56	0.86

Library Size Effect

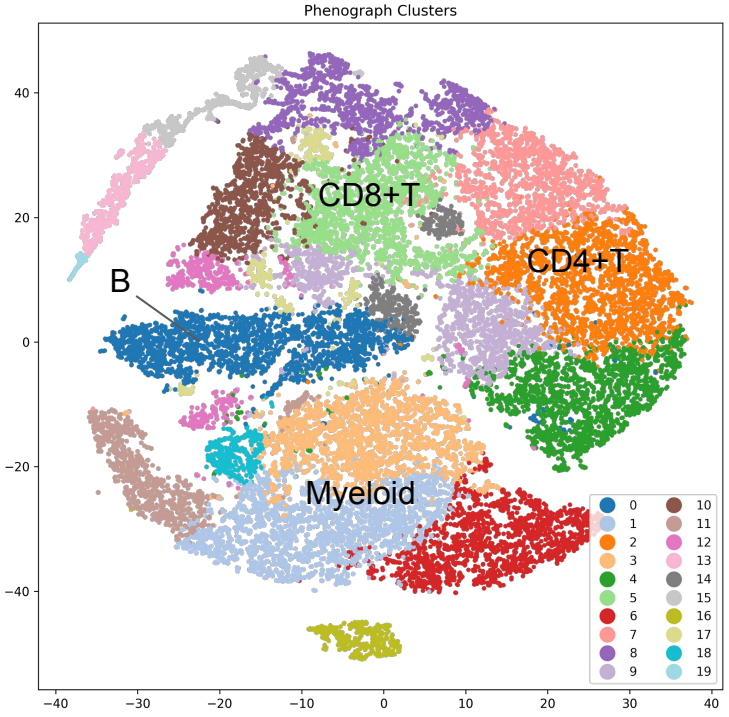


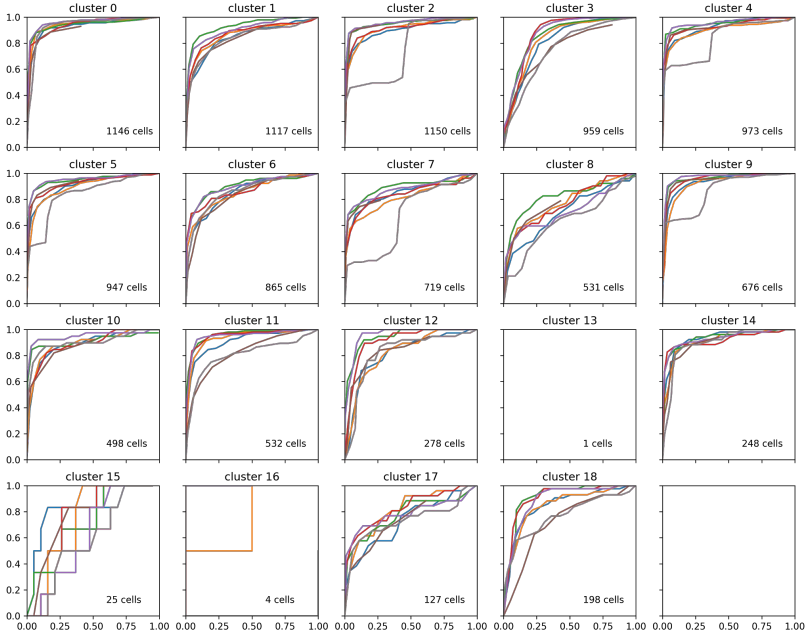
Performance!





Test Dataset by Cluster





	TNR	TPR
0	0.97	0.64
1	0.92	0.62
2	0.97	0.65
3	0.73	0.76
4	0.94	0.74
5	0.97	0.60
6	0.95	0.53
7	0.98	0.45
8	0.98	0.24
9	0.93	0.78
10	0.99	0.49
11	0.76	0.83
12	0.84	0.74
13	1.00	NaN
14	0.96	0.54
15	0.92	0.08
16	0.88	0.19
17	0.97	0.31
18	0.40	0.87

Summary information

7058 cells were called singlets by all methods

- Of real singlets, 154 were called as doublets by all methods

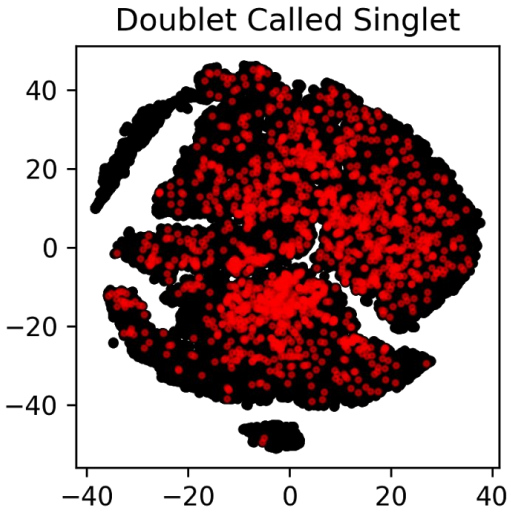
4108 were called as a doublet by at least one method

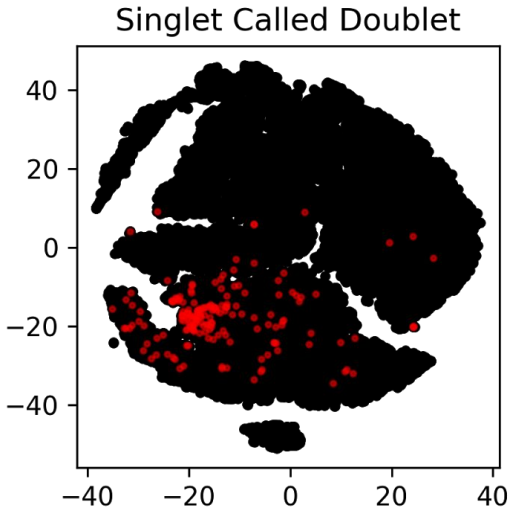
- Of these, 500 were called as doublets by all methods
  - Of these 500, 71 were defined (by the genotypes) as singlets

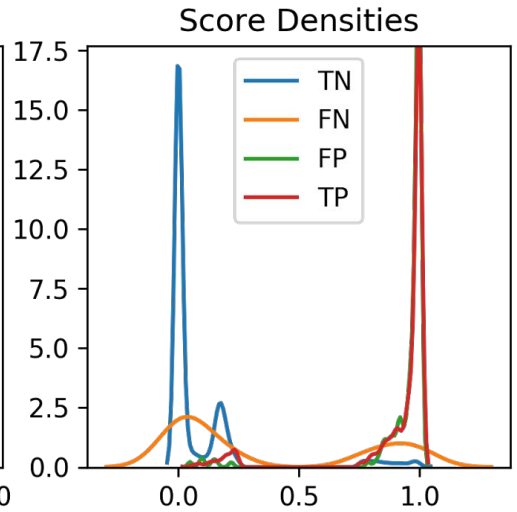
Top genes for predicting (random forest) the doublets that all methods called singlets

MT-CO3
ISCU
POLDIP3
ITCH
WRN
STK17B
TMEM189
VIM-AS1
PCBP1
RPS3
PBRM1
HLA-A
RPL7P23
RPL23A
EIF3M
GLIPR2
NOP10
HLA-C
PPP4R3A

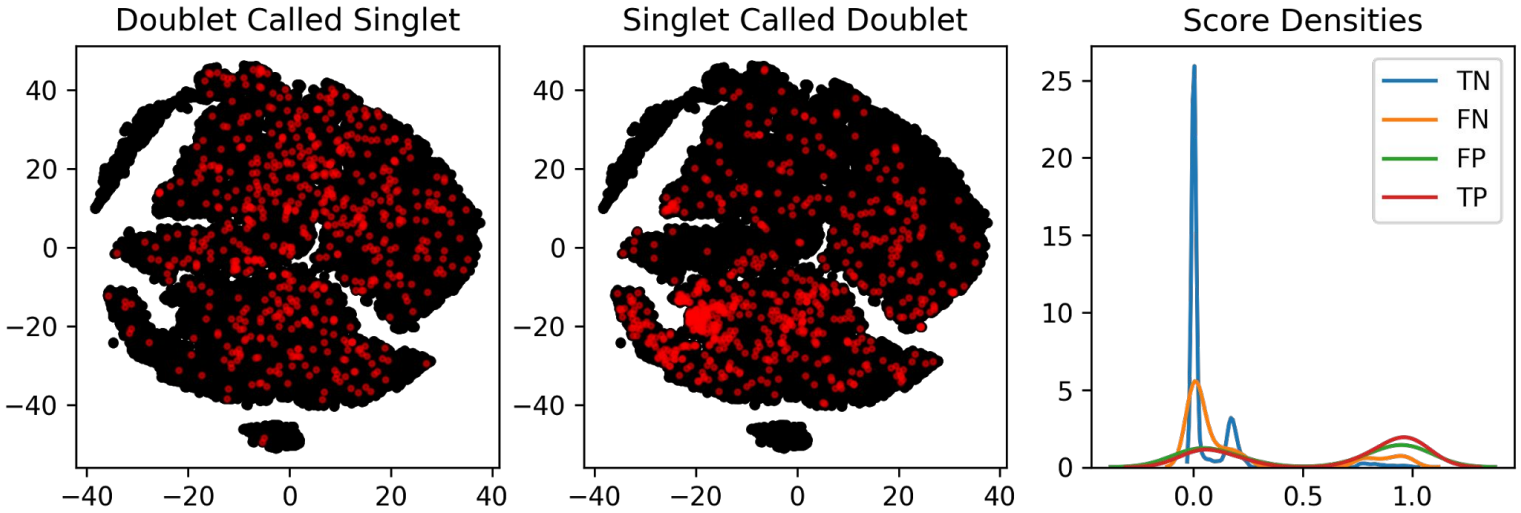
Group 5


Doublet Called Singlet
40
20
0
−20
−40
−40
−20
0
20
40

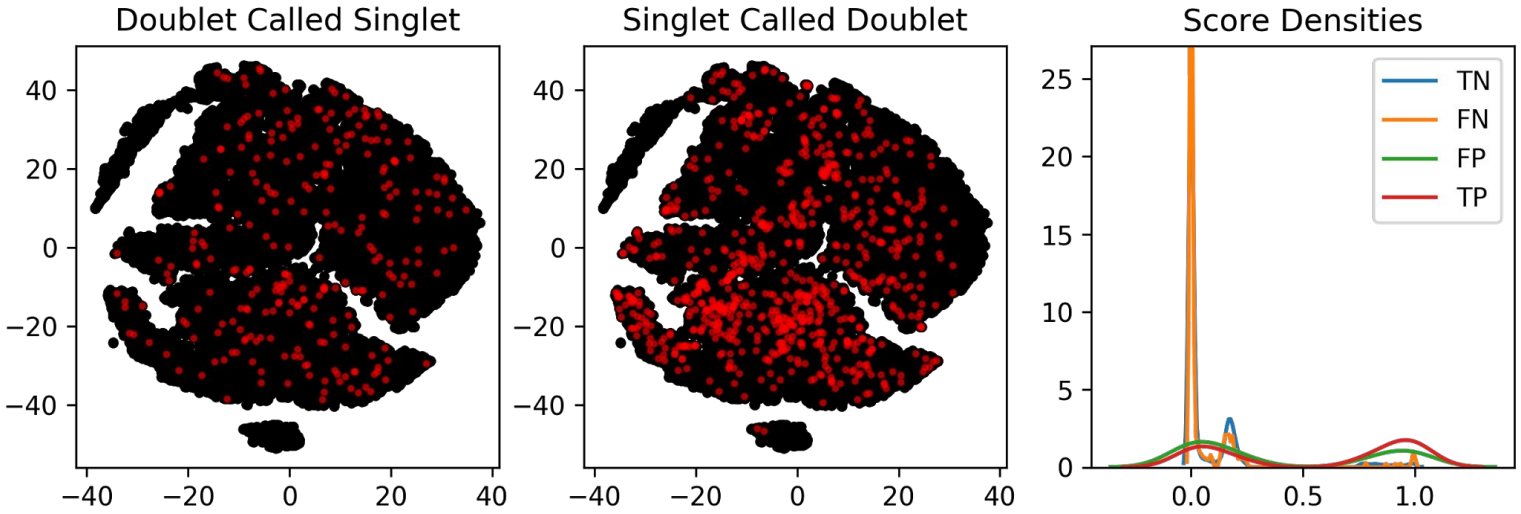

Singlet Called Doublet
40
20
0
−20
−40
−40
−20
0
20
40


Score Densities
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.5
1.0
TN
FN
FP
TP

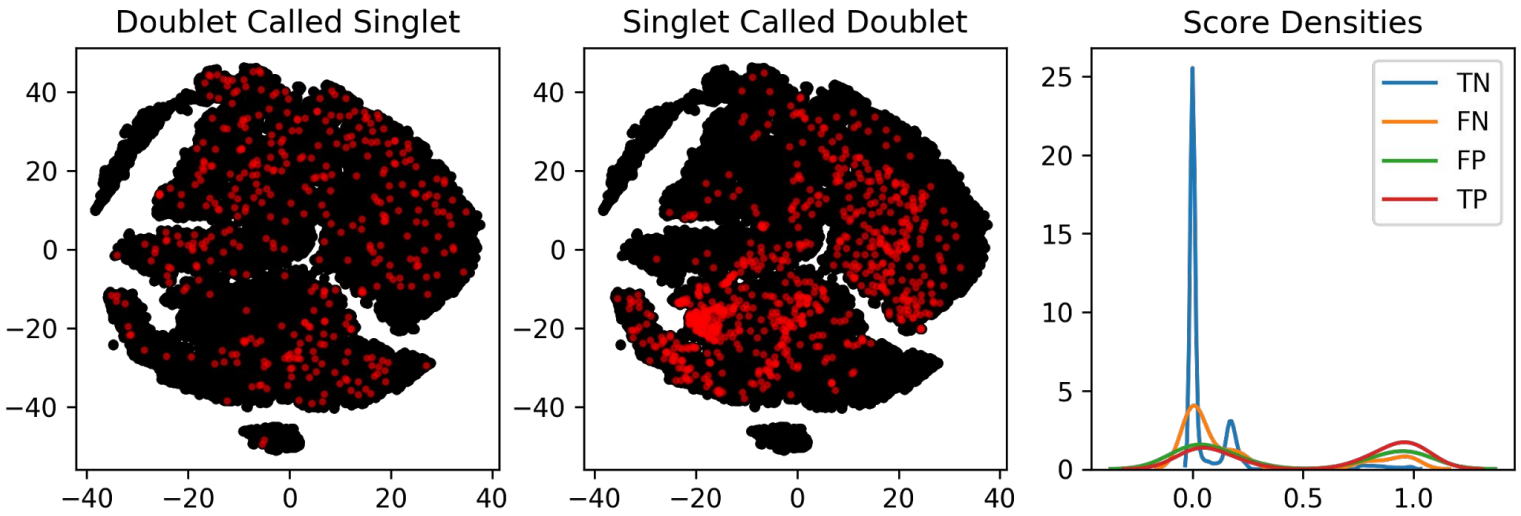
Group 6


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
25
20
15
10
5
0
0.0
0.5
1.0

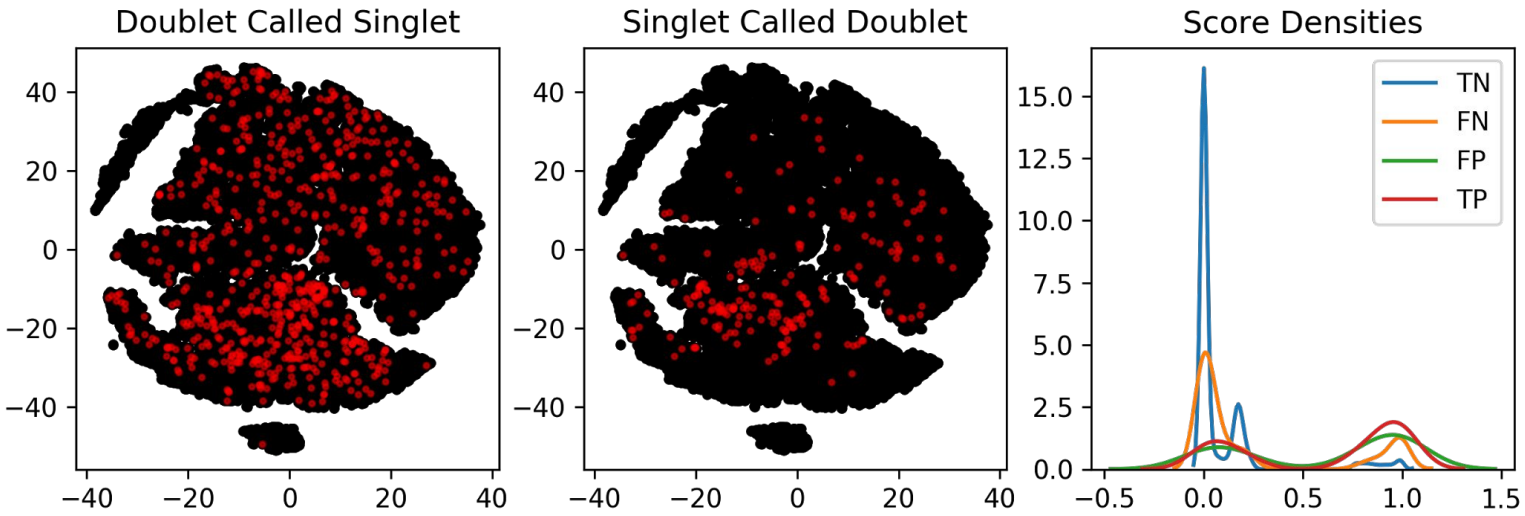
Group 7


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
25
20
15
10
5
0
0.0
0.5
1.0

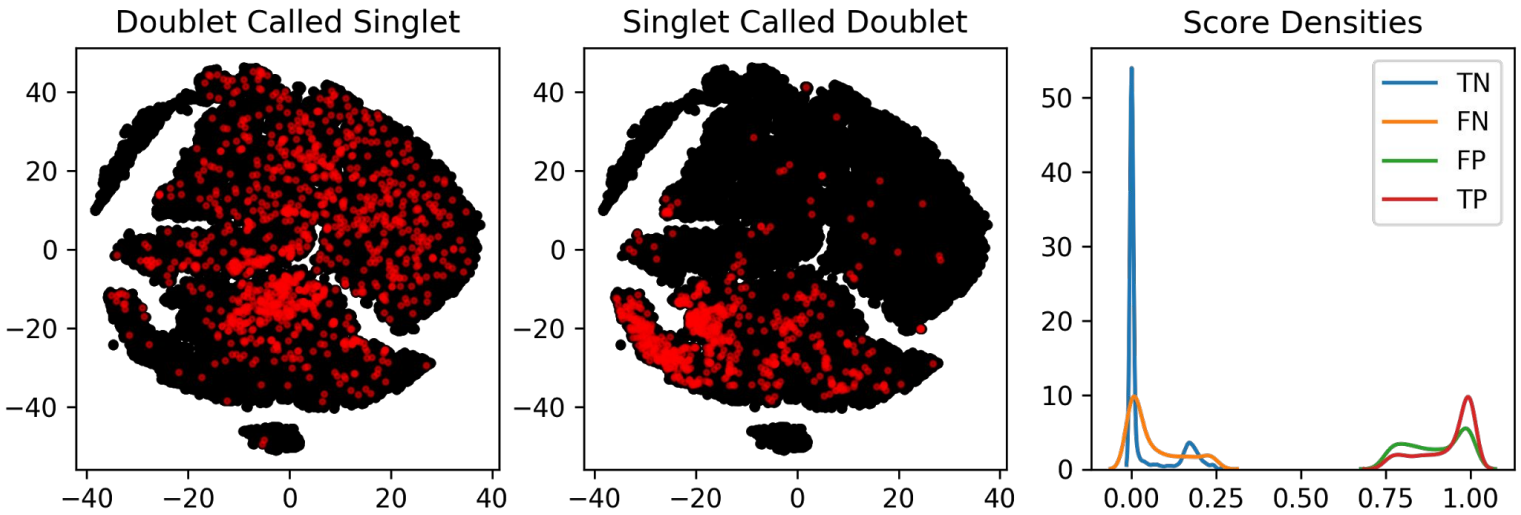
Group 8


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
25
20
15
10
5
0
0.0
0.5
1.0

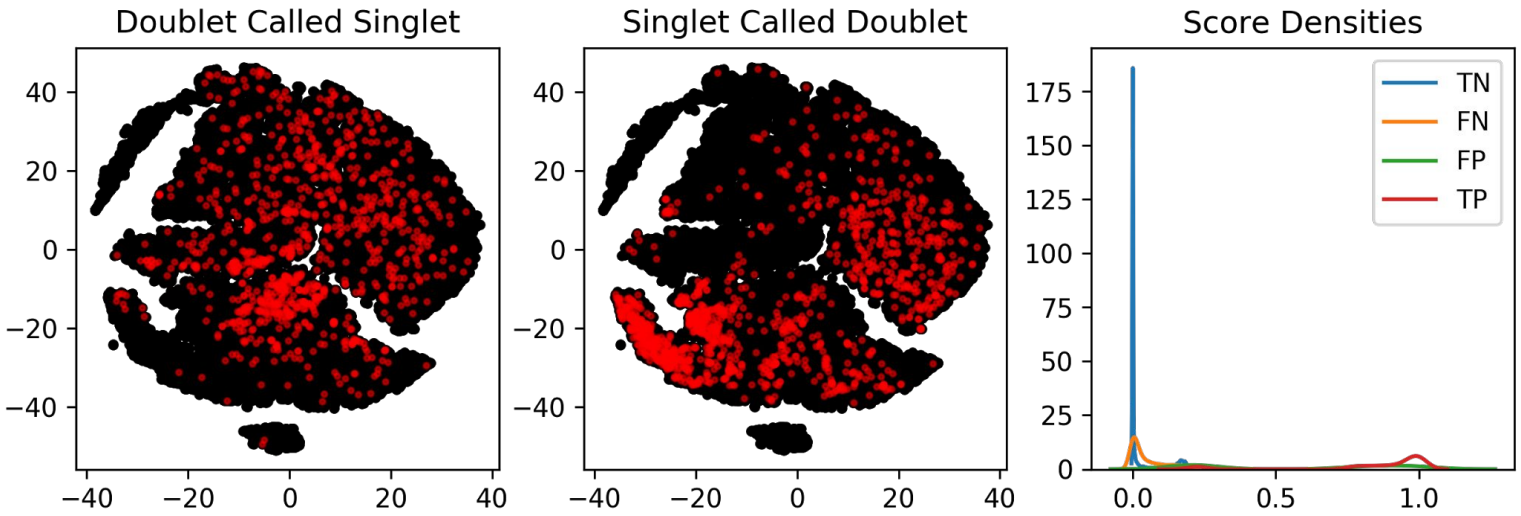
Group 9


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−0.5
0.0
0.5
1.0
1.5

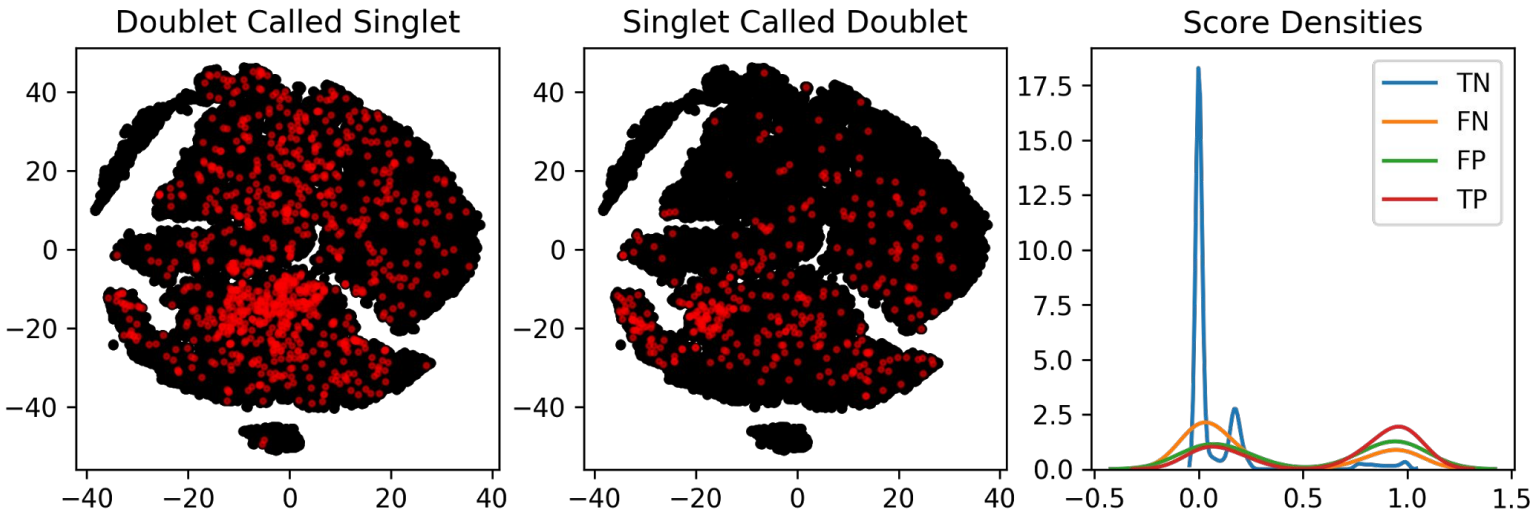
Group 9.2


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
50
40
30
20
10
0
0.00
0.25
0.50
0.75
1.00

Group 9.3


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
175
150
125
100
75
50
25
0
0.0
0.5
1.0

Group 9.4


Doublet Called Singlet
Singlet Called Doublet
Score Densities
TN
FN
FP
TP
40
20
0
−20
−40
−40
−20
0
20
40
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−0.5
0.0
0.5
1.0
1.5