

0 Introduction

Doublets are a characteristic error source in droplet-based single-cell sequencing data where two cells are encapsulated in the same oil emulsion and are tagged with the same cell barcode. Across-type doublets manifest as fictitious phenotypes that can be incorrectly interpreted as novel cell types. We present a novel, fast, unsupervised classifier to detect across-type doublets in single-cell RNA-sequencing data that operates on a count matrix and imposes no experimental constraints. This classifier leverages the creation of in silico synthetic doublets to determine which cells in the input count matrix have gene expression that is best explained by the combination of distinct cell types in the matrix.

In the next section, we invite you to explore the pseudocode of our method. A bioRxiv submission is currently in the works.

1 Algorithms

Algorithm 1.1 Classify cells as doublets or singlets. *counts* is an N by D count matrix, where N is the number of cells and D is the number of genes.

Precondition: *ITERS* and *BOOSTRATE* have been set.

```
1: function DOUBLETDETECTION(counts)
2:    $N, D \leftarrow \text{shape}(\text{counts})$  ▷ Dimensions of counts.
3:   doublet  $\leftarrow$  empty ITERS by  $N$  boolean matrix
4:   for  $i = 1, \dots, \text{ITERS}$  do
5:     raw_synths  $\leftarrow$  CreateDoublets(counts, BOOSTRATE)
6:     augmented  $\leftarrow$  NormalizeCounts(append(counts, raw_synths))
7:     reduced  $\leftarrow$  PCADimReduction(augmented)
8:     clusters  $\leftarrow$  PhenographCluster(reduced)
9:      $P \leftarrow$  AssignP-Values( $N$ , length(raw_synths), clusters)
10:    for  $j = 1, \dots, N$  do
11:      doublet[ $i, j$ ]  $\leftarrow P[j] \leq 0.01$  ▷ Call cell  $j$  a doublet this run if  $\leq 0.01$ .
12:    end for
13:  end for
14:  labels  $\leftarrow$   $N$  length array of zeros
15:  for  $j = 1, \dots, N$  do
16:    if [CountTrue(doublet[:,  $j$ ])/length(doublet[:,  $j$ ])]  $\geq 0.9$  then
17:      labels[ $j$ ]  $\leftarrow 1$  ▷ Cell  $j$  was called doublet on at least 90% of runs.
18:    end if
19:  end for
20:  return labels
21: end function
```

Algorithm 1.2 Cluster p-value Assignment via Hypergeometric Test. N is the number of cells in the augmented dataset. K is the number of synthetic doublets in the augmented dataset. $clusters$ is the set of clusters for the augmented dataset, where each cluster contains its member cells.

```

1: function ASSIGNP-VALUES( $N, K, clusters$ )
2:    $P \leftarrow$  empty  $N$  length array
3:   for  $c$  in  $clusters$  do
4:      $n \leftarrow |c|$ 
5:      $k \leftarrow$  number of synthetic doublets in  $c$ 
6:      $p \leftarrow 1 - \text{hypergeom.cdf}(N, K, n, k)$ 
7:     for  $cell$  in  $c$  do
8:        $P[\text{IndexOf}(cell)] \leftarrow p$ 
9:     end for
10:  end for
11:  return  $P$ 
12: end function

```
