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 Downsampling of two parent gene-vectors. *cell*1 and *cell*2 are gene-vectors represented as arrays of gene indices. Each element of the array is an index corresponding to one transcript of that gene in the gene-vector. If a gene-vector has, for example, 15 counts of a particular gene, the corresponding index will be present 15 times in the array.

```
function DOWNSAMPLECELLPAIR($cell1, cell2$)
    $newsize \leftarrow$ max(length($cell1$), length($cell2$))
    $summed \leftarrow$ append($cell1, cell2$)
    $shuffled \leftarrow$ RandomPermutation($summed$)          ▷ Shuffle position of elements.
    $synthetic \leftarrow shuffled[: newsize]$
    return $synthetic$
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.

```
function ASSIGNP-VALUES($N, K, clusters$)
    $P \leftarrow$ empty $N$ length array
    for $c$ in $clusters$ do
        $n \leftarrow |c|$
        $k \leftarrow$ number of synthetic doublets in $c$
        $p \leftarrow$ hypergeom.cdf($N, K, n, k$)
        for $cell$ in $c$ do
            $P$[IndexOf($cell$)] $\leftarrow p$
        end for
    end for
    return $P$
end function
```

Algorithm 1.3 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.

```
function DoubletDetection($counts$)
    $N, D \leftarrow$ shape($counts$)                                ▷ Dimensions of $counts$.
    $doublet \leftarrow$ empty $ITERS$ by $N$ boolean matrix
    for $i = 1, \dots, ITERS$ do
        $raw\_synths \leftarrow$ CreateDoublets($counts, BOOSTRATE$)
        $augmented \leftarrow$ NormalizeCounts(append($counts$, $raw\_synths$))
        $reduced \leftarrow$ PCADimReduction($augmented$)
        $clusters \leftarrow$ PhenographCluster($reduced$)
        $P \leftarrow$ AssignP-Values($N$, length($raw\_synths$), $clusters$)
        for $j = 1, \dots, N$ do
            $doublet[i, j] \leftarrow P[j] \geq 0.99$         ▷ Call cell $j$ a doublet this run if $\geq 0.99$.
        end for
    end for
    $labels \leftarrow N$ length array of zeros
    for $j = 1, \dots, N$ do
        if [CountTrue($doublet[:, j]$)/length($doublet[:, j]$)] $\geq 0.9$ then
            $labels[j] \leftarrow 1$                ▷ Cell $j$ was called doublet on at least 90% of runs.
        end if
    end for
    return $labels$
end function
```