Task1 Write-up (group2)

Anna Cuomo, Caleb Weinreb, Keegan Korthauer, Nils Eling, & Viktor Petukhov

Introduction to the task: detecting 'background cells': For the Human Cell Atlas hackathon, our task was to detect 'background cells', defined (in the context of droplet based single cell RNA seq) as cell barcodes that were not co-encapsulated with a *bona fide* single cell. Here, we describe a computational approach cfor background cell detection called *Wolball*, demonstrate its accuracy to detect simulated background cells, and then apply it a dataset of peripheral blood mononuclear cells (PBMCs).

Motivation for nearest-neighbor detection of empty droplets: We began with the observation that putative background cells tend to connect to each other in nearest neighbor graphs. For example, when viewing force-directed layouts of k-nearest-neighbor (knn) graphs from datasets with simulated background cells, we always observed a distinct cluster of putative background cells (Figure 1). Though cells in the background cluster tended to have lower total counts, they could not be distinguished by total counts alone. The linking of background cells in nearest neighbor graphs could be driven by several phenomena, such as: (i) shared 'expression' of genes that are enriched in the cell buffer (e.g. marker genes of lysis-prone cells); (ii) high distance from bona fide cells, which occupy stereotyped regions of gene expression space.

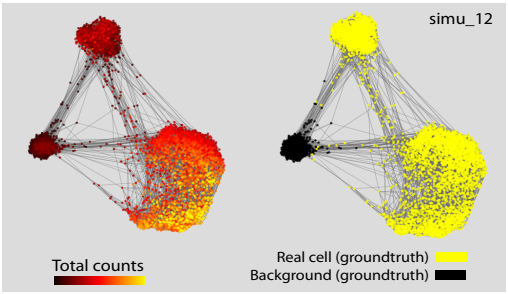

Figure 1: Background cells connect in knn graph. We generated a k-nearest neighbor graph using Euclidian distances in PC space. Simulated background cells (black dots in right –hand plot) form a clear cluster, but this cluster is not clearly distinguished by total counts (right-hand plot). This figure uses "simu_12". The force-directed layout was generated in SPRING.

Wolball identifies background cells using proximity in nearest neighbor graphs: The proximity between background cells in nearest neighbor graphs offers an opportunity to identify cells as background even when their total UMI count is in an ambiguous range. We implemented this approach in an algorithm called Wolball (Figure 2).

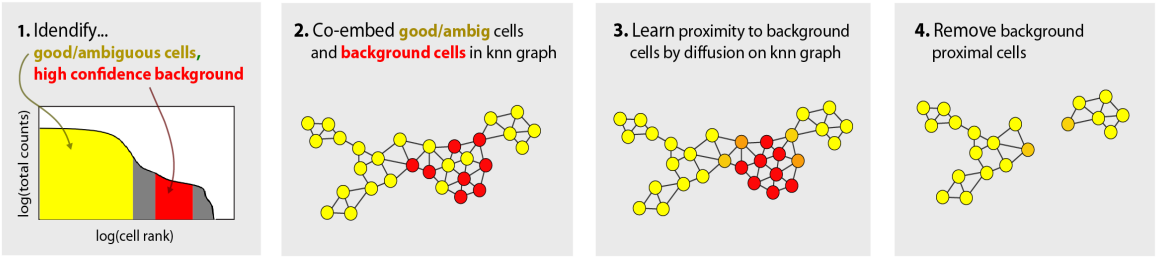

Figure 2: Wolball uses graph proximity to find background cells. Bona fide and ambiguous cells (yellow) are co-embedded with high-confidence background cells (red) in a nearest neighbor graph. Using a graph diffusion process, labels are propagated from high-confidence background cells to the ambiguous cells that link with them in the graph. The diffused labels constitute a 'background score' that can be used to filter the data.

The input to Wolball is two matrices of single cell gene expression counts. The first matrix (A) should contain bona fide cells and also cells that are ambiguous between real and background.

The second matrix (B) should contain expression profiles that can be assigned as background with high confidence. A and B can be chosen using total counts cutoffs or by any other method. The idea is to learn what a background cell looks like using matrix B, and then apply that knowledge to detect the background cells in A. Wolball assigns a background score to each cell in A as follows:

1. Matices A and B are provided input.
2. Expression profiles from A and B are co-embedded in a k-nearest-neighbor graph, where the graph links are formed using Euclidean distances in high dimensional PC space.
3. A binary score vector (1 for B cells, 0 for A cells), is diffused across the graph.
4. Cells with a high diffused-score are likely background and can be removed.

Wolball classifies simulated background cells better than total counts: We applied Wolball to datasets with simulated background cells and compared its accuracythe naïve approach based on total counts alone. In many cases, such as "simu_12" (shown in Figure 1), the Wolball score could discriminate ground-truth background cells much better than total counts could (Figure 3A). In all cases examined, the Wolball score was better than total counts when judging classification accuracy by AUC (Figure 3B).

Wolball filters background cells from PBMC data while retaining rare cell-type: We calculated Wolball scores for 10,000 cells in a dataset of peripheral blood mononuclear cells (PBMCs) provided by 10X Genomics. The 10X in-house pipeline classifies 2,000 of these cells as background on the basis of total UMI count (Figure 3C). The Wolball scores for this dataset mostly agree with the 10X filtering. However, a group of cells expressing megakaryocyte genes that is filtered in the 10X pipeline is assigned a low score by Wolball. Thus Wolball appears to retain certain low-count cell types that would normally be filtered out in standard pipelines.

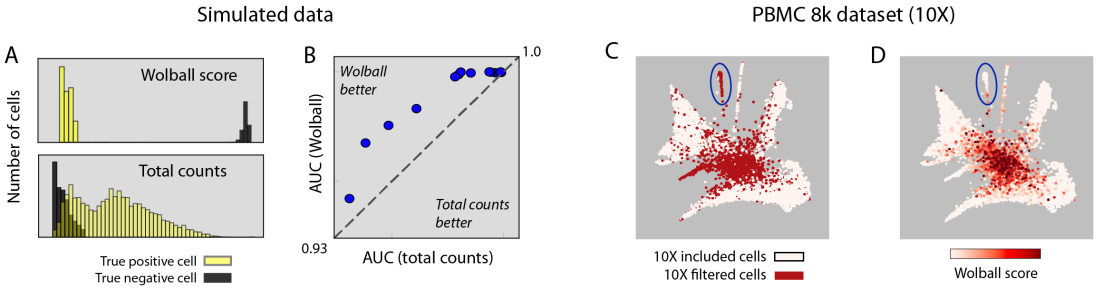

Figure 3: Wolball detects background cells better than total counts alone. (A) Wolball score for "simu_12" compared to total counts. (B) Wolball AUC exceeds naive total counts classification in all simulations. (C) In PBMC data from 10X genomics, the standard filtering pipeline excludes low-counts cell types such as megakaryocytes (blue circle). In contrast, these low-counts cell types were assigned a low score by Wolball (D). Node positions in (C,D) were generated using SPRING.

Discussion: We have described a computational pipeline for filtering single cell expression data using proximity to high-confidence background cells in nearest neighbor graphs. One possible disadvantage of Wolball is that it does not take into account the expected statistical properties of background cells, such as the absence of over-dispersion in their expression counts. However, this model free approach may be useful in situations where 'background cells' do not simply arise from free mRNAs entering empty droplets, but instead include chunks of cellular debris and other biological material that would not be well-described by a simple statistical model.