Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 8 May 2017
Application Note

OXFORD

Computational Genomics

DoubletDetection: Identifying Technical Error in Single-cell RNA-sequencing Data

Adam Gayoso ,[1] Jonathan Shor [1] and Ryan Brand [1]

[1] Department of Computer Science, Columbia University, New York, NY, 10027, USA

Abstract

Motivation: Modern Single-cell RNA-sequencing techniques commonly include noise caused by multiplets. This undermines any single-cell analysis from such data, and to date is an open problem.
Results: We build a software tool, DoubletDetection, that reliably identifies doublets (multiplets of two cells) from different cell types, the most damaging source of commonly occurring multiplet noise.
Availability: Software available at `https://github.com/adamgayoso/DoubletDetection`
Contact: rmb2208@columbia.edu, ajg2188@columbia.edu, J.Shor@columbia.edu

1 Introduction

Full understanding of cellular function requires knowledge of the specific genes each cell of interest expresses. Single-cell RNA-sequencing (scRNA-seq) enables the characterization of cells in a heterogeneous population at single cell resolution. The process of cell isolation often involves a microfluidic device that encapsulates the cell in a droplet containing one cellular barcode for demultiplexing the resulting reads (Zilionis *et al.*, 2017; Klein *et al.*, 2015). As experimenters increase the quantity of cells to be sequenced, the occurrence of multiplets, technical errors in which two or more cells are encapsulated in the same droplet and receive the same cellular barcode, also increases. Multiplet rates typically range from 4% to 10% (Zheng *et al.*, 2017). Such errors add noise to the resultant sequencing data as transcripts from multiple cells are erroneously tagged as a single cell's transcripts.

We present DoubletDetection, a software tool designed to detect doublets, the specific case of a multiplet droplet containing exactly two cells. Current methods of doublet detection involve filming the cell encapsulation process to establish an inferred doublet rate (Zheng *et al.*, 2017), or introducing staining reagents into the cell population sample (Zunder *et al.*, 2015). Our tool is, to the best of our knowledge, the first of its kind that detects doublets in downstream count tables without imposing any particular experimental setup constraints.

2 Approach

`doubletdetection` is a python module that treats doublets in scRNA-seq data as outliers, and detects them by boosting the signal of doublets in the dataset with synthetic doublets. First, it creates synthetic doublets and appends them to the original dataset. Next, it clusters the new dataset with PhenoGraph, an existing clustering algorithm for scRNA-seq data (Levine *et al.*, 2015). A score is assigned to each cell based on the proportion of synthetic doublets in its PhenoGraph cluster. We select a suggested cutoff score by a heuristic, and all cells with a score greater than or equal to the cutoff are identified as doublets. Overall, this works well in finding doublets whose parent cells are from distinct cell types. Because doublets with parents of the same cell type express a similar set of genes as a singlet from this cell type, they are indistinguishable from singlets of this cell type using our approach.

While all multiplets are deleterious, previous work (Zheng *et al.*, 2017) has shown that the occurrence of multiplets for a given experiment follows a Poisson distribution. The distribution's parameter is dependent on the exact experimental setup, but is expected to always be well below 1.0. With doublets expected to be at least an order of magnitude more likely than greater multiplets, we focus our efforts accordingly.

3 Methods

3.1 Creation of synthetic doublets

We represent each cell as a row vector, $c \in \mathbb{R}^d$, with each dimension representing the expression of a particular gene, and typically $d \in [15\,000 - 30\,000]$. We refer to the l_1 norm of c as the library size. We create a synthetic doublet by randomly choosing two cells in the dataset (c_1, c_2), summing them $(c_1 + c_2 = c_{12})$, and then downsampling so that the library size of the synthetic doublet is equal to the largest library size of the two randomly selected cells. Downsampling works by treating the value for each gene divided by the library size as a probability in a multinomial distribution, and then sampling genes until we reach the desired target library size of the synthetic doublet. This method reflects the way in which counts are processed by the sequencer. The number of synthetic doublets created is dictated by the parameter `boost_rate`.

3.2 Clustering with PhenoGraph

We cluster the augmented dataset with PhenoGraph, which works by creating a k-nearest-neighbor graph of the cells in the data, weighting edges following the Jaccard similarity index, and then assigning community labels using the Louvain method(Levine *et al.*, 2015; Blondel *et al.*, 2008). The input to PhenoGraph is the augmented dataset normalized by dividing each cell by its library size and then multiplying all cells by the median library size, replacing zeros with 0.1, and then taking the natural log of the dataset. We then reduce the normalized data to 30 dimensions using Principle Components Analysis before inputing to Phenograph.

The output from PhenoGraph is a community assignment, representing an approximation of an optimal assignment of the augmented dataset into communities using a modularity metric. PhenoGraph is typically used to cluster cell types for further biological analysis, but in this case, we leverage the technique to identify those clusters to which many of our synthetic doublets are assigned.

3.3 Cell scoring

We assign a score to each cell by calculating, for each PhenoGraph community, the fraction of synthetic doublets clustered within that community. All of the original samples in a given community receive the same score. Since cells with a high score cluster with many synthetic doublets, they are the most likely to be doublets themselves.

The score value for each cell is heavily reliant on the `boost_rate`, and its meaning is derived by looking at the distribution across all cells. Many clusters have low scores, while a small number have relatively higher scores. We set a suggested cutoff score, for which cells with score $\geq$ cutoff are classified as doublets. We do so using a heuristic that sets the suggested cutoff score within the largest gap between adjacent scores in sorted order.

4 Case Study

We demonstrate functionality and performance of `doubletdetection` on a dataset provided by 10x which contains scRNA-seq counts for two distinct cell lines: a male (Jurkat) and a female (293T) which preferentially express the genes CD3D and XIST, respectively. Zheng *et al.* use this as a validation set to calculate an inferred doublet rate. The underlying assumption for this dataset is that any true doublet with parents from distinct cell types should be composed of one Jurkat and one 293T and should therefore express both genes. Furthermore, we know from the outset that only 812 of all 3388 cells express both genes at any level above zero and we hypothesize that doublets will constitute a subset of these 812 cells that express both genes at more significant levels.

Figure 1 provides a visualization of the application of DoubletDetection to the 10x dataset with `boost_rate = 0.25` and using the 30 nearest neighbors of each cell for clustering with PhenoGraph (`knn=30`). In this particular example, we identify 61 cells in the original data as doublets, and all 61 express both the XIST and CD3D gene. The two main cell lines manifest as two distinct groups of cells that only vary significantly along the first dimension as computed using Principal Components Analysis, which we know is driven by the difference in expression of the aforementioned genes. As expected, each of the identified doublets lies in the open space that exists between the two primary groups of cells along the first dimension, since significant enough expression of both genes leads to simultaneous differentiation from both of the predominant cell lines. Repeating this process 15 times yields the following results: on average, 60.87 cells are identified as doublets and 100% of those doublets express both the XIST and CD3D genes. 65 unique cells are identified across all trials, generating a distribution of identification for these cells with a mean of 14.12 and a standard deviation of 3.25. This provides support for the method outlined above, which assumes that a doublet can be identified by a unique pattern of gene expression that differs substantially from that of any individual cell in a given scRNA-seq dataset.

5 Discussion

We interpret the success of our algorithm to suggest an empirical definition of cell type as a group of cells that produce synthetic doublets indistinguishable from the group with high probability. While clearly a data-focused definition, it ties to the intuitive idea of a cell type as a group of cells exhibiting a consistent phenotype while allowing greater granularity on what exactly constitutes consistent through the use of different similarity metrics. Where we used the modularity optimized Jaccard similarity metric of PhenoGraph that considers all genes as equally important, and therefore equally discriminative in defining cell types, a study of sub populations within a sample of known origin might use a metric that considers only specific genes of interest.

In our experience, a high proportion of identified doublets form their own clusters when PhenoGraph processes the original data. This can be used in tandem with prior knowledge of the sample to validate that identified doublets are unlike any expected cell type in the data.

6 Conclusion

Our algorithm finds that doublets often cluster together and that those clusters are distinct from "real" cell type clusters. This confirms that such doublets add noise to the original data and may masquerade as single cell types without detection.

Our method reliably identifies doublets consisting of cells from different cell types. By providing scores for each cell, users are able to select a cutoff appropriate to their data, or use our suggested cutoff. We anticipate that this tool will enable existing scRNA-seq pipelines to automatically improve their analyses with only one-time initial tuning.

Acknowledgements

We would like to thank Itsik Pe'er, Dana Pe'er, David van Dijk, and Ambrose J. Carr.

References

Blondel, V. D. *et al.* (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), P10008.

Klein, A. M. *et al.* (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**(5), 1187–1201.

Levine, J. H. *et al.* (2015). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**(1), 184–197.

Zheng, G. X. Y. *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**, 14049 EP –.

Zilionis, R. *et al.* (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protocols*, **12**(1), 44–73.

Zunder, E. R. *et al.* (2015). Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protocols*, **10**(2), 316–333.

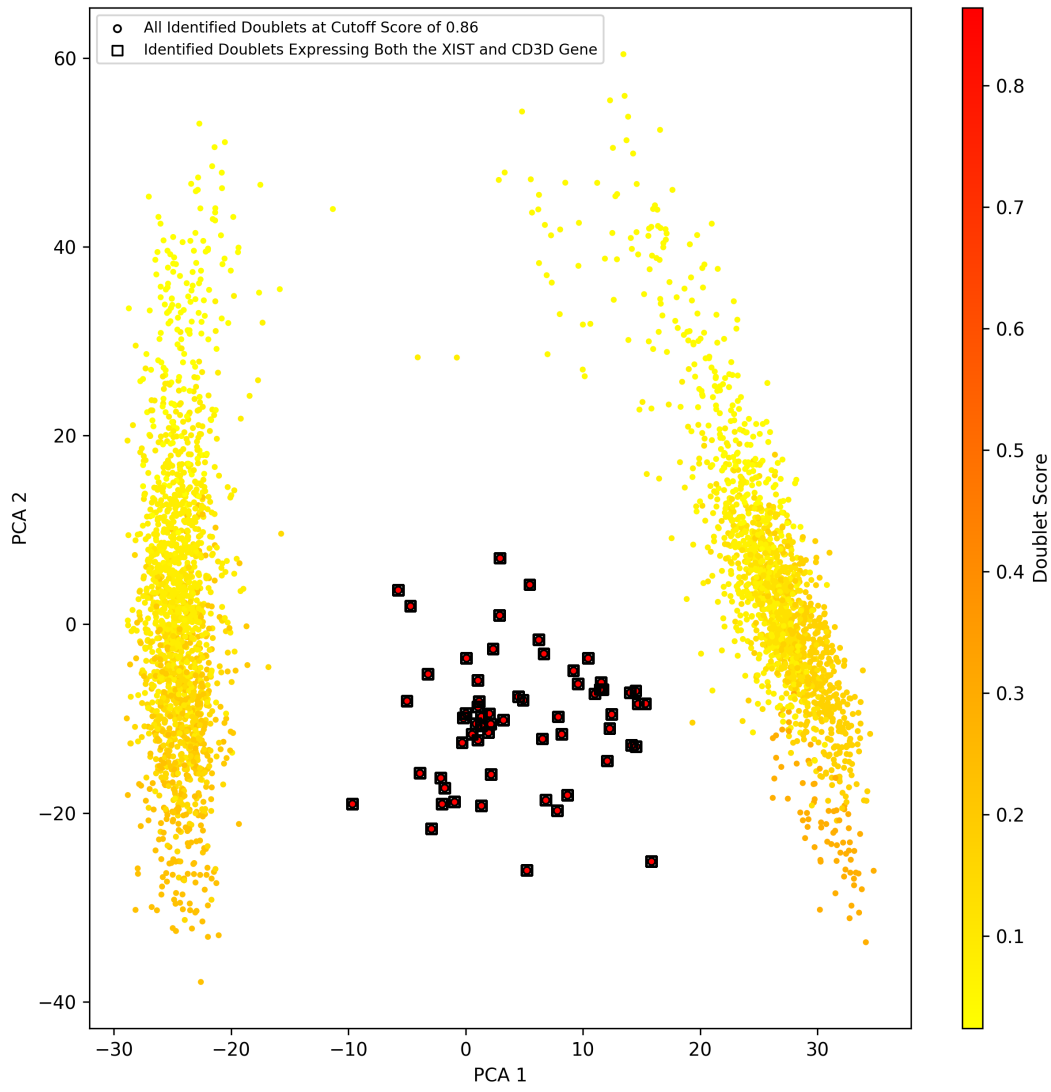

Fig. 1. Scoring for 10x dataset as computed by DoubletDetection.