Welcome to the documentation of Singlet Code!

Singlet code (link) uses cell barcoding technology to identify **real singlets** in scRNAseq data, benchmarks existing doublet detection methods and provides a *classifier* to detect singlets in *unbarcoded* datasets as well of the same cell type.

Singlet Code

This section contains information about the need for doublet detection in RNAseq analysis, singlet code detects singlets, how is this data used for benchmarking other doublet detection methods and the command line interface to

TO-DO: Add information about command-line interface and figures.

Doublets in RNA sequencing data

How scRNA-seq technologies work?
scRNA-seq technologies rely on distributing individual cells from a suspension into individual reactions, each labeled with a unique "ID", usually in the form of a reaction-specific sequence barcode.

Since each cell has an unique ID, why are there doublets?
Despite numerous technological optimizations, multiple cells can occasionally be encapsulated in a single reaction, resulting in doublets or multiplets where two or more cells are assigned the same reaction ID. The percentage of doublets in a given experiment depends on several factors, including the features of the sample and throughput, and can be as high as 40-50% [B1] [B3]. In turn, such artifacts affect the downstream analysis [B2].

How does Singlet Code detect singlets?

Singlet Code uses the fact that the lineage barcodes are within the cell before oil encapsulation for sequencing. Therefore, a cell would have one lineage barcode or a result of multiplicity, more than barcode but with restrictions. To be labelled a singlet, a barcoded cell needs to satisfy one of these three conditions:

1. a single barcode identified per cell ID.
2. multiple barcodes per cell but the same barcode combination is found in other cells in the same samples.
3. multiple barcodes per cell but the same barcode combination is found in other cells across samples within the same experimental design (common for barcoding studies).

Creating Benchmark datasets from Singlet Code results

Typical benchmarking for different doublet detection methods are created by simulating doublets and using the methods to identify these cells when mixed with singlet cells in varying percentages. However, until now, there really was no ground truth for singlets in experimental data.

Singlet Code provides singlets that are closest possible to ground truth and can be used to simulate doublets for the benchmarking datasets.

To simulate the doublets, we randomly selected the count data from two cells we identify as true singlets. We then averaged the counts from these two cells to generate simulated doublets as performed previously [B1]. The final scRNA-seq datasets are generated by adding such simulated doublets into the datasets at different percentages (5-25%).

Classifier

This page will contain a brief introduction to the classifier and links to 4 pages: how does it work with the figure, maybe explain what is AUPRC, how does it compare to other methods, how can people integrate this into their work

Doublet detection methods

This section will contain information about the doublet detection methods, maybe their main figure along with the link to the main paper.

Using the dataset we created using the results from Singlet code as described in mainSingletCode/benchmarking, we benchmark four doublet detection methods - scDblFinder, DoubletFinder, Scrublet and Hybrid.

In this section, you can find more information about each of these methods and how they perform on the benchmark.

Doublet Detection Methods

Benchmark

AUPRC vs AUROC

Datasets used in Singlet Code

This section will contain all the information about the different datasets in different tabs: the big table containin gall the information about the datasets, maybe one tab with cell types in the datasets

Datasets used

SingletCode was tested on datasets with many different barcoding techniques, cell types and sequencing methods. Comprehensive details about the datasets used are described in the table.

References

References in the website

[B1] Nicholas J. Bernstein, Nicole L. Fong, Irene Lam, Margaret A. Roy, David G. Hendrickson, and David R. Kelley. Solo: doublet identification in single-cell rna-seq via semi-supervised deep learning. *Cell Systems*, 11(1):95–101.e5, July 2020. doi:10.1016/j.cels.2020.05.010.

[B2] Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. doi:10.15252/msb.20188746.

[B3] Nan Miles Xi and Jingyi Jessica Li. Benchmarking computational doublet-detection methods for single-cell rna sequencing data. *Cell Systems*, 12(2):176–194.e6, February 2021. doi:10.1016/j.cels.2020.11.008.

[B4] Ziyang Zhang, Madeline E. Melzer, Karun Kiani, and Yogesh Goyal. Singletcode: synthetic barcodes identify singlets in scrna-seq datasets and evaluate doublet algorithms. *bioRxiv*, 2023. URL: https://www.biorxiv.org/content/early/2023/08/07/2023.08.04.552078, doi:10.1101/2023.08.04.552078.

Also, this website was built with Sphinx and hosted on Read the Docs. We would like to also acknowledge the TomoBank as the template for building this site.

FAQs

This page will contain FAQs about the paper, the classifier and about whom to contact. Example:

1. What to do if we do not use barcoded cells in our lab?
2. Where can I access the code in the paper?
3. Does the singlet detection using barcode depend on cell type?