Demultiplexing single-cell RNA-seq based on expressed cell barcodes

Overview

In any sequencing-based analysis strategies, errors are introduced via PCR during library preparation and subsequent sequencing. This is particularly true in single-cell RNA-sequencing (scRNA-seq) where initial cDNA libraries undergo multiple rounds of PCR amplification. In this problem, we will demultiplex a mixture of single cells tagged with two distinct ‘CellTag’ barcodes, introduced via lentivirus and expressed as transcripts that are detected in each single-cell transcriptome (see references[1,2] for more background). You will initially assign cells to their sample group based on their CellTag expression. You will then attempt to assign any initially ambiguous ‘non determined’ cells which have errors in their CellTag sequence.

In this problem, cells express one of two 8-nucleotide CellTags:
Control: ATGTTGC
Treatment: GATTACA

Notes:
Generation of the dataset (call.matrix.csv) used for this problem.
- scRNA-seq, followed by extraction of CellTag expression in each cell, based on the CellTag motif, 'GGT[ACTG]{8}GAATTC'
- For CellTag expression to be reported, must be represented by 2 or more transcripts (represented by independent UMIs)
Example:

cell.barcode	ATGTTGC	GATTACA	ATGTAGC	GTGTAGC
ACAGCTACAAACGTGG	52	0	1	0
AAACGGGCACTTAAGC	1	45	0	0
AGCCTAAAGCGTCTAT	63	0	0	0
AAACGGGCATCTATGG	0	74	0	0
CATCAAGCACCAGATT	0	23	0	0

- CellTag expression per cell is then binarized in this problem, for simplicity
Example:

cell.barcode	ATGTTGC	GATTACA	ATGTAGC	GTGTAGC
ACAGCTACAAACGTGG	1	0	0	0
AAACGGGCACTTAAGC	0	1	0	0
AGCCTAAAGCGTCTAT	1	0	0	0
AAACGGGCATCTATGG	0	1	0	0
CATCAAGCACCAGATT	0	1	0	0

In a typical scRNA-seq dataset, we would be analyzing 10,000's of single cells. For simplicity, in this problem we are focusing on a small subset of 50 cells.

1. Write a Python script to assign cells to one of three groups: Control, Treatment, and non-determined (defined as either not expressing CellTags, or expressing CellTags that are not an exact match to the above sequences). How many cells are assigned to each group?

2. Focusing on the non-determined cell group, write a Python script, incorporating an edit distance of 1, to assign cells to one of three groups: Control, Treatment, and non-determined. How many cells are assigned to each group?

3. Suggest three different strategies (either computational or experimental) to maximize the proportion of cells that can be assigned to Control and Treatment groups.

References

1. Guo, C., Biddy, B. A., Kamimoto, K., Kong, W. & Morris, S. A. CellTag Indexing: a genetic barcode-based multiplexing tool for single-cell technologies. *bioRxiv* 335547 (2018). doi:10.1101/335547
2. Biddy, B. A. *et al.* Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564,** 219–224 (2018).