Assignment 7

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

Bio5488
3/1/2024
Assignment 7: Demultiplexing problem

• Input
  • call.matrix.csv

• Goals & output
  • sort calls based on barcode reads
  • count cells per condition
Input data: call.matrix.csv

CSV = Comma Separated Values

“GGTACTG—NNNNNNN--GAATTC”.

7-nucleotide CellTags:

<table>
<thead>
<tr>
<th>cell.barcode</th>
<th>ATGGTTGC</th>
<th>GATTACA</th>
<th>ATGTAGC</th>
<th>GTGTAGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAGCTACAAACGTGG</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAACGGGGACTTAAAGC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AGCCTAAAGCGTCTAT</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAACGGGCATCTATGG</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CATCAAGCACCAGATT</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

For simplicity, in this assignment:
- CellTag expression per cell is binarized in this problem;
- Only focus on 50 cells;
- Only focus on 4 oligo tags
TODO 1: Write a Python script:
Determine number of cells with tags for control, treatment, and non-determined groups.

<table>
<thead>
<tr>
<th>cellbarcode</th>
<th>ATGTTGC</th>
<th>GATTACA</th>
<th>ATGTAGC</th>
<th>GTGTAGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAGCTACAAACGTGG</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAACGGGCACCTTAAGC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AGCCTAAAGCGTCTAT</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAACGGGGCATCTATGG</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CATCAAGCACCAGATT</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Control (ATGTTGC) = ? cells
- Treatment (GATTACA) = ? cells
- Non-determined = ? cells

Non-determined defined as:
- either not expressing CellTags,
- or expressing CellTags that are not an exact match to the above sequences
TODO 2: Focusing on the non-determined cell group, modifying your Python script: allowing an edit distance of 1, re-count cells in control, treatment, and non-determined groups.

**Edit distance:** the number of characters which differ between two strings

Problem CellTag Examples:

- Edit distance = 1
  - TATGC×GT
  - TAGGCGT

- Edit distance = 2
  - TATGC×GT
  - TAGCCGT
TODO 2: Focusing on the non-determined cell group, modifying your Python script:
allowing an edit distance of 1, re-count cells in control, treatment, and non-determined groups.

**Edit distance**: the number of characters which differ between two strings

Problem CellTag Examples:

- **Edit distance = 1**
  
  TATGCGT
  TAGGCCG
  TA   GCGT
  TANGCGT

- **Edit distance = 2**
  
  TATGCGT
  TAGCCCGT

No need to consider in this simplified question
TODO 3: Suggest three different strategies (either computational or experimental) to maximize the proportion of cells that can be assigned to Control and Treatment groups.
Assignment 7: requirements

- Due March 22th, week after Spring Break! (HW 8 due same day though)
- What to turn in:
  ✔ demultiplex.py
  ✔ README.txt

- REMINDER:
  Don’t update your submission folder after due time!