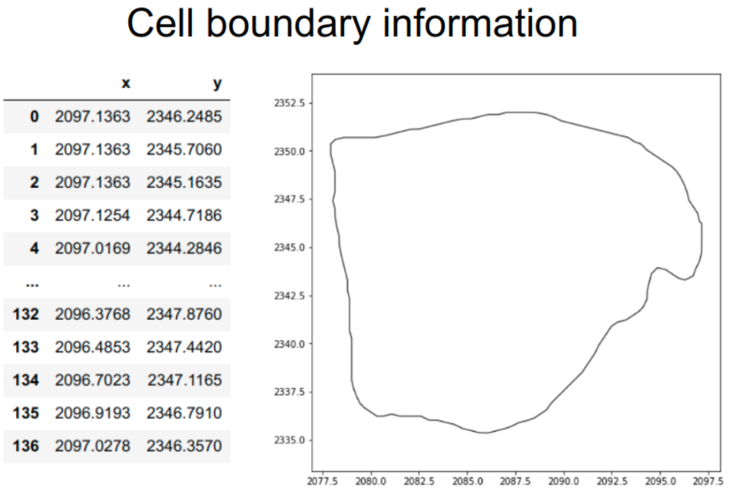
1 Introduction

Mathematical formulations for the subcellular metrics computed on MERFISH datasets

2 Informal description of available data

MERFISH data from the preprint here has cell-boundary and RNA spot information available for $238,000$ cells with 252 different genes. Not all cells have RNA spots detected for all genes.

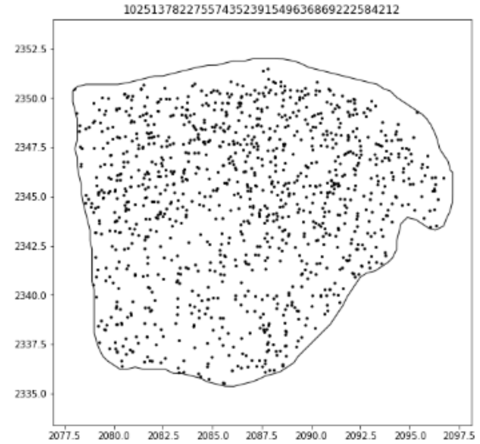
Cell-boundaries take the form of a list of (x,y) positions for each cell. Drawing a straight line between each position in this list results in a closed polygon that represents the outside boundary of the cell

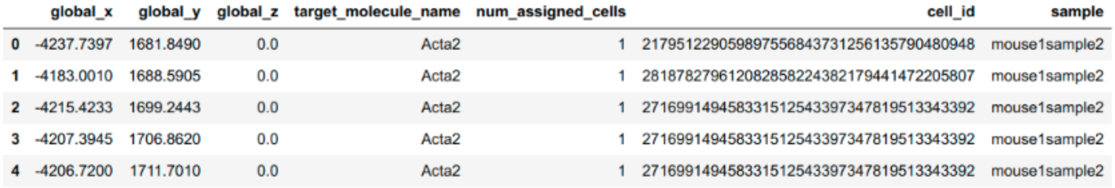


Cell boundary information

	x	y
0	2097.1363	2346.2485
1	2097.1363	2345.7060
2	2097.1363	2345.1635
3	2097.1254	2344.7186
4	2097.0169	2344.2846
...	...	...
132	2096.3768	2347.8760
133	2096.4853	2347.4420
134	2096.7023	2347.1165
135	2096.9193	2346.7910
136	2097.0278	2346.3570

RNA-spots are the MERFISH imaging output and contain information for both x,y location of the spot, and the RNA identity at that location. I've identified which RNA spots are within each cell boundary.





	global_x	global_y	global_z	target_molecule_name	num_assigned_cells	cell_id	sample
0	-4237.7397	1681.8490	0.0	Acta2	1	217951229059897556843731256135790480948	mouse1sample2
1	-4183.0010	1688.5905	0.0	Acta2	1	281878279612082858224382179441472205807	mouse1sample2
2	-4215.4233	1699.2443	0.0	Acta2	1	271699149458331512543397347819513343392	mouse1sample2
3	-4207.3945	1706.8620	0.0	Acta2	1	271699149458331512543397347819513343392	mouse1sample2
4	-4206.7200	1711.7010	0.0	Acta2	1	271699149458331512543397347819513343392	mouse1sample2

3 Definitions

There are n spots, c cells, and g unique genes in the total dataset

- Let P be a matrix of positions of each RNA spot. P has dimensions of n by 2 for the x and y coordinates
- Let I be an binary matrix of size n by c, where $I_{i,j} = 1$ indicates that RNA spot i is within cell j, and 0 otherwise. Row sums are 1.
- Let G be a binary matrix of size n by g where $G_{i,j} = 1$ if RNA spot n is of gene type G and 0 otherwise. Row sums are 1.
- Let X be the number of RNA spots per gene per cell, defined by $I^T G$. $X_{i,j}$ is the number of RNA spots of gene j in cell i. X has dimensions of c by g.
- Let Y be a binary matrix of size c by g where $Y_{i,j} = 1$ if cell i has at least one RNA spot of gene g and 0 otherwise.

From these definitions we can formulate all calculations below

4 Common calculations

Count the number of all RNA spots, regardless of gene-type, in cell c

$$\sum_{i=1}^{n} I_{i,c} \tag{1}$$

Count the number of RNA spots of type g in cell c

$$\sum_{i=1}^{n} G_{i,g} I_{i,c} \tag{2}$$

Count the number of unique genes in cell c

$$\sum_{i=1}^{g} Y_{c,i} \tag{3}$$

Get the average x-position of all RNA spots of gene g in cell c

$$\frac{\sum_{i=1}^{n} I_{i,c} G_{i,g} P_{i,1}}{X_{c,g}} \tag{4}$$

Get the average y-position of all RNA spots of gene g in cell c

$$\frac{\sum_{i=1}^{n} I_{i,c} G_{i,g} P_{i,2}}{X_{c,g}} \tag{5}$$

5 Metric definition

All metrics are functions that are applied to each cell/gene pair and result in a scalar value.

Let M be a metric such that $M(c,g)$ is a real number for the result of the metric calculation on gene g in cell c. $M(c,g)$ is only defined if there is at least one RNA spot of gene g in cell c ($Y_{c,g} = 1$)

In order to compare the results of metric M between cells with different sizes and shapes, I am Z-score normalizing M over all genes within each cell.

The Z-score normalization for gene g in cell c is defined to be $M(c,g)$ minus the mean of M for cell c over all genes, divided by the variance of M for cell c over all genes.

The mean of M over all genes in cell c is defined as:

$$Mean(M(c)) = \frac{\sum_{i=1}^{g} M(c,i)}{\sum_{i=1}^{g} Y_{c,i}} \tag{6}$$

The var of M over all genes in cell c is defined as:

$$Var(M(c)) = \frac{\sum_{i=1}^{g}(M(c,i) - Mean(M(c)))^2}{(\sum_{i=1}^{g} Y_{c,i}) - 1} \tag{7}$$

Then finally we have the definition for the Z-normalized metric on gene g for cell c

$$Z(c,g) = \frac{M(c,g) - Mean(M(c))}{Var(M(c))} \tag{8}$$

6 Centrality metric

This metric calculates the mean distance of all the RNA spots of gene g in cell c from the cell centroid.

For this metric we will make a new definition for the cell centroids

- Let C be cell centroids where $C_{i,1}$ is the x-centroid location of cell i. $C_{i,2}$ is the y-centroid location. C has shape c by 2

Then lets define a distance function $D(i,j)$ to be the L2 distance between the centroid of cell i and the position of RNA spot j

$$D(i,j) = \sqrt{(C_{i,1} - P_{j,1})^2 + (C_{i,2} - P_{j,2})^2} \tag{9}$$

$$M_{centrality}(c,g) = \frac{\sum_{i=1}^{n} D(c,i) I_{i,c}}{X_{c,g}} \tag{10}$$

7 Polarity metric

8 Periphery metric