find_clusters manual

Background

Following cocaine administration neuroses express group of genes called IEGs, which includes also the gene EGR2. The goal of this script is to test whether there is micro organization of cells in a way that neighboring cells will be more likely to express EGR2 following cocaine treatment. This results will suggests a mechanism of clusters of cells responding to a stimuli at the same time.

Data

The data is single-molecule FISH which provides spatial information about gene expression levels. Data was obtained from: *Citri, Ami (2021), "smFISH data of IEG expression in the dorsal striatum after acute, repeated, and challenge cocaine exposures", Mendeley Data, V3, doi: 10.17632/p5tsv2wpmg.3*

The data is smFISH results from striatum sections after 1h stimulation with cocaine.Three gene were tested: Egr2, Arc and Nr4a1. The dataset includes: number of replica, gene expression levels in each cells, x and y coordinates of the cell.

```
##   rep_num Arc Egr2 Nr4a1 center_x center_y
## 1       1  17    6     7 4522.008 23.63286
## 2       1  21   22    24 4620.672 29.46771
## 3       1   7    0     1 4439.245 32.63957
## 4       1   5    1     1 4420.523 78.53510
## 5       1  35   24    16 4597.468 86.32935
## 6       1   0    2    14 5313.910 82.56420
```

Example input

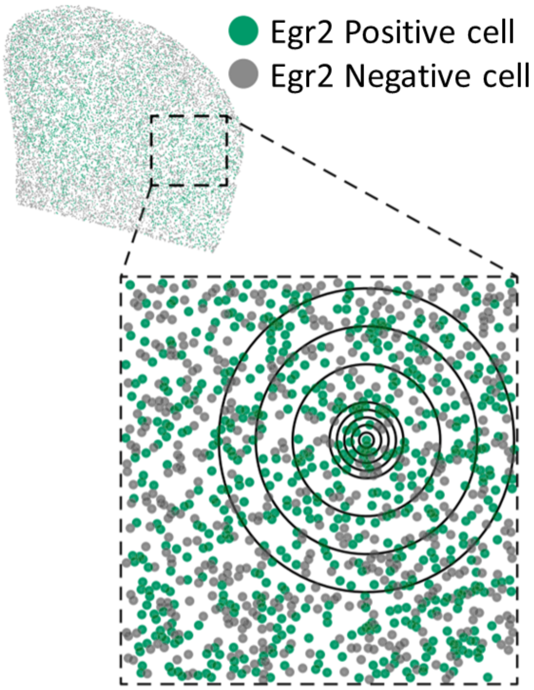
The script running the pipeline is: **Run_find_cluster_pipeline.R** . To use this script one must load files in the format described above. We use the data from the treatment (1h), 3 replicas, and selected the gene Egr2 for analysis. We also need to set the cutoff level of expression to consider cell as Egr2 positive cell (6 in this example)

```
#setwd("/path/to/find_clusters")

input_file = "input_example/_merged_1h_Arc_Nr4a1_.csv"
outDir = "output_example/"
cond = "1h"
myGene = "Egr2"
pos_cut = 6
reps = c(1:3)
```

Analysis guidlines

1. In the first step of analysis we will select each Egr2 positive cells and calculate the fraction of other Egr2 positive cells in its environment. We will repeat this process with an increasing radius distance as demonstrated in the figure bellow.



Once we collect the information fo all the cells, we can calculate the probability of a cells being an Egr2 positive cell as a function of its distance from other Egr2 positive cell. To validate the significance of our results we will repeat this process with shuffled data. The summary of these results, from 3 different replicas will be written to the file **Egr2_distance_probability_plot.pdf** in the output folder. We can see that the shuffled data results with consistence probability (which is equal to the fraction of the positive cells in the data), while in the real data there is a strong dependency to the distance from other positive cell.

2. In the next step of analysis we will try to find “cluster” of positive cells. Meaning, group of Egr2 positive cells with are in a homogenuse environment (>80% of the cells in this environment are also positive). For that goal we will use a recursive functions that starts form a positive cell and adds cells to this cluster as long as they are located less then 20 micron and that at least 80% of the cells in the environment are positive. The Results, for each repica seperatly are in html files at the output folder, for example: **Egr2_1h_1.html**

3. At the last step, we will dived the clusters found to bins by the number of cells, and calculate the mean expression of Egr2 in each cluster group (also in the html files). We can see small, but consistance, correlation between the number of cells in cluster and mean Egr2 epression.