find_hostspots manual

Background

In some application such as single-molecule FISH we obtain a spatial information about gene expression levels. The goal of this script is to identify a patten of expression in 2D space, explore the reproducibility between biological replicates, and compare between few genes tested in the same experiment.

Data

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 before and 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   8    4    15 3019.079  45.86756
## 2       1   1    0     0 2631.921  62.09142
## 3       1  14    8    10 3773.377 101.17299
## 4       1   4    1     1 4018.893  93.77170
## 5       1  27    1     7 3709.124 104.29448
## 6       1   4    1     0 2538.018  89.10385
```

Analysis guidelines

1. Find the high expressors cells based on the control samples.
2. Apply 2D kernel density estimation with contours on the highly expressors cells data
3. Select the highest contours that include at least 20% of the cells.
4. Apply the analysis separately for each replica. Overlap the results from all the replicas to see if we the selected contours are overlapping.
5. Calculate the average expression levels in the selected contours to gain information about the expression level of the genes at the hotspot.
6. Perform the same analysis with shuffled data. The expression levels of the original data are shuffled, while the x and y coordinates remained the same. This allows as to evaluate the specificity of your results.

Example input

To use this script one must load files in the format described above. Two files are needed: one for the control (0h) and one from the treatment (1h). In addition, the list of genes to be analyzed should be included.

```
#setwd("/path/to/find_hotspot/dir")

#for control
path_data_control = "input_example/_merged_0h_Arc_Nr4a1_.csv"

#for treatment
path_data_treat = "input_example/_merged_1h_Arc_Nr4a1_.csv"

list_genes = c("Arc","Nr4a1","Egr2")

output_file_name = "output_example/hotspot_res.pdf"

shuffle_data = TRUE

shuffle_file_name = "output_example/hotspot_shuffled.pdf"
```

Results

The file **hotspot_res.pdf** includes three parts:

1. General view of the x and y coordinates of the cells.
2. Plot for each gene in each condition separately. In these plots each replica in marked with different color so we can test the reproducibility between replicas.
3. One plot for all the data together. Each panel include the data for one gene. In this plot the color transpose indicates on average expression levels within the selected areas.

The file **hotspot_shuffled.pdf** includes the same analysis as above, only with shuffled data.

We can see that there is a clear pattern of spacial expression in **hotspot_res.pdf** but not in the shuffled data. The pattern of expression, as well as the levels of genes expression are changing between the control and treatment.