# The workflow:

## In Python

Import data
- Read csv files
- Join spread sheets
- Average out PxC5,PxC6

Cleaning data
- Set gene as index
- Identify gene repeats
- Drop some columns

Preprocessing
- Normalization
- PCA

Build visualization tool

Clustering
- k-mean
- t-SNE + k-mean*
- MeanShift*
- AgglomerativeClustering
- SpectralClustering
- DBSCAN*
- Brich

* Not work well with this dataset

Comparing and combing clustering results
- Collective Plotting
- Build ensemble

## In R

Plate map

Dendrogram*

Circular dendrogram

* Not work well

# Data. Drop columns which end by '_cnt','_norm' or start with 'Max_','Min_'.

29 features to use:

['Green_5+foci_ratio', '1R1G_nuc_ratio', 'Nofoci_cell_ratio', 'Avg_green_periphery_std_dist', 'Avg_All_std_dist', 'Large_nuc_ratio', 'Red_1foci_ratio', 'Green_3_foci_ratio', 'Avg_RG_mean_dist', 'Red_2foci_ratio', 'Red_4foci_ratio', 'Avg_Red_mean_dist', 'Avg_All_mean_dist', 'Avg_green_periphery_mean_dist', 'Red_far_foci_ratio', 'Avg_Red_std_dist', 'Green_1foci_ratio', 'Avg_Green_std_dist', 'Green_far_foci_ratio', 'Avg_Green_mean_dist', 'Avg_RG_std_dist', 'Green_4foci_ratio', 'Green_2foci_ratio', 'Red_3_foci_ratio', 'Total_far_foci_ratio', 'Avg_red_periphery_std_dist', '1R1G_touch_nuc_ratio', 'Avg_red_periphery_mean_dist', 'Red_5+foci_ratio']

46 features dropped:

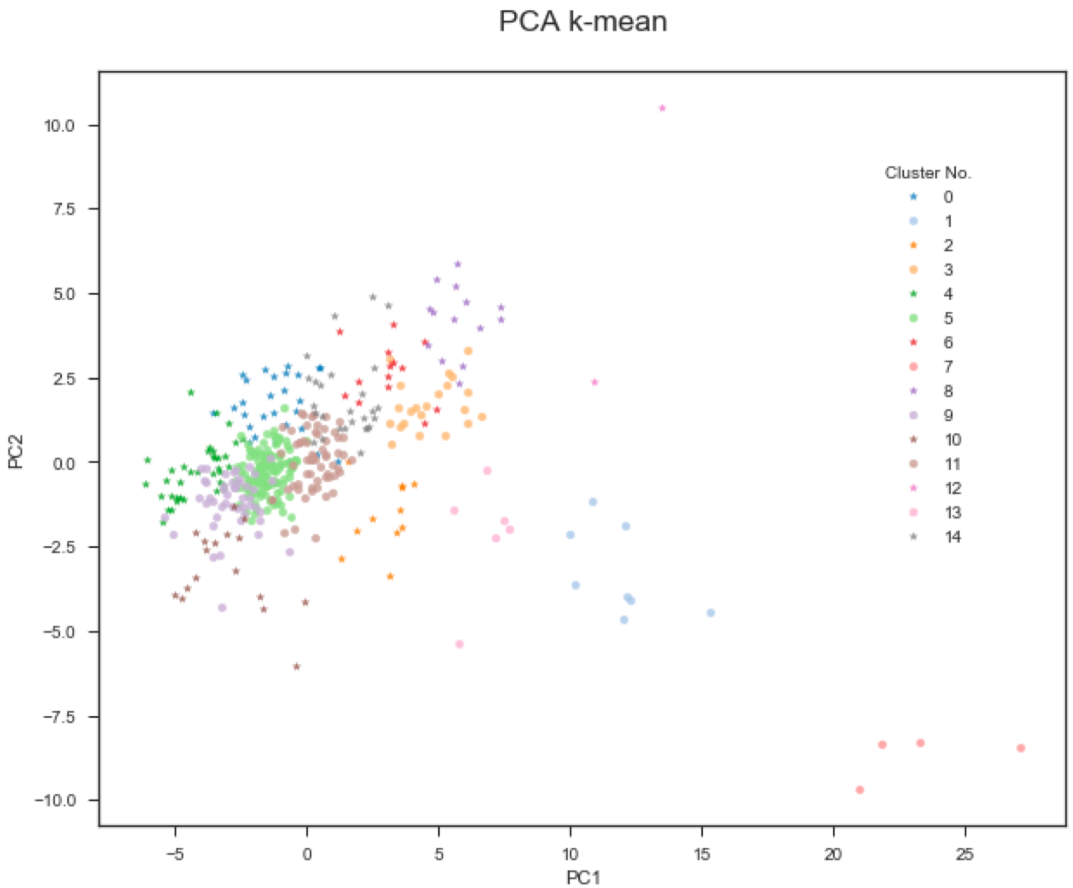
['Cell_cnt', 'Nofoci_cell_cnt', 'Green_cell_cnt', 'Red_cell_cnt', 'Mean_green_foci_cnt', 'Mean_red_foci_cnt', 'Max_Green_mean_dist', 'Max_Green_std_dist', 'Max_Red_mean_dist', 'Max_Red_std_dist', 'Max_All_mean_dist', 'Max_All_std_dist', 'Min_RG_mean_dist', 'Min_RG_std_dist', 'Max_RG_mean_dist', 'Max_RG_std_dist', 'Min_green_periphery_mean_dist', 'Min_green_periphery_std_dist', 'Min_red_periphery_mean_dist', 'Min_red_periphery_std_dist', 'Avg_Green_mean_dist_norm', 'Avg_Green_std_dist_norm', 'Max_Green_mean_dist_norm', 'Max_Green_std_dist_norm', 'Avg_Red_mean_dist_norm', 'Avg_Red_std_dist_norm', 'Max_Red_mean_dist_norm', 'Max_Red_std_dist_norm', 'Avg_All_mean_dist_norm', 'Avg_All_std_dist_norm', 'Max_All_mean_dist_norm', 'Max_All_std_dist_norm', 'Avg_RG_mean_dist_norm', 'Avg_RG_std_dist_norm', 'Min_RG_mean_dist_norm', 'Min_RG_std_dist_norm', 'Max_RG_mean_dist_norm', 'Max_RG_std_dist_norm', 'Avg_green_periphery_mean_dist_norm', 'Avg_green_periphery_std_dist_norm', 'Min_green_periphery_mean_dist_norm', 'Min_green_periphery_std_dist_norm', 'Avg_red_periphery_mean_dist_norm', 'Avg_red_periphery_std_dist_norm', 'Min_red_periphery_mean_dist_norm', 'Min_red_periphery_std_dist_norm']

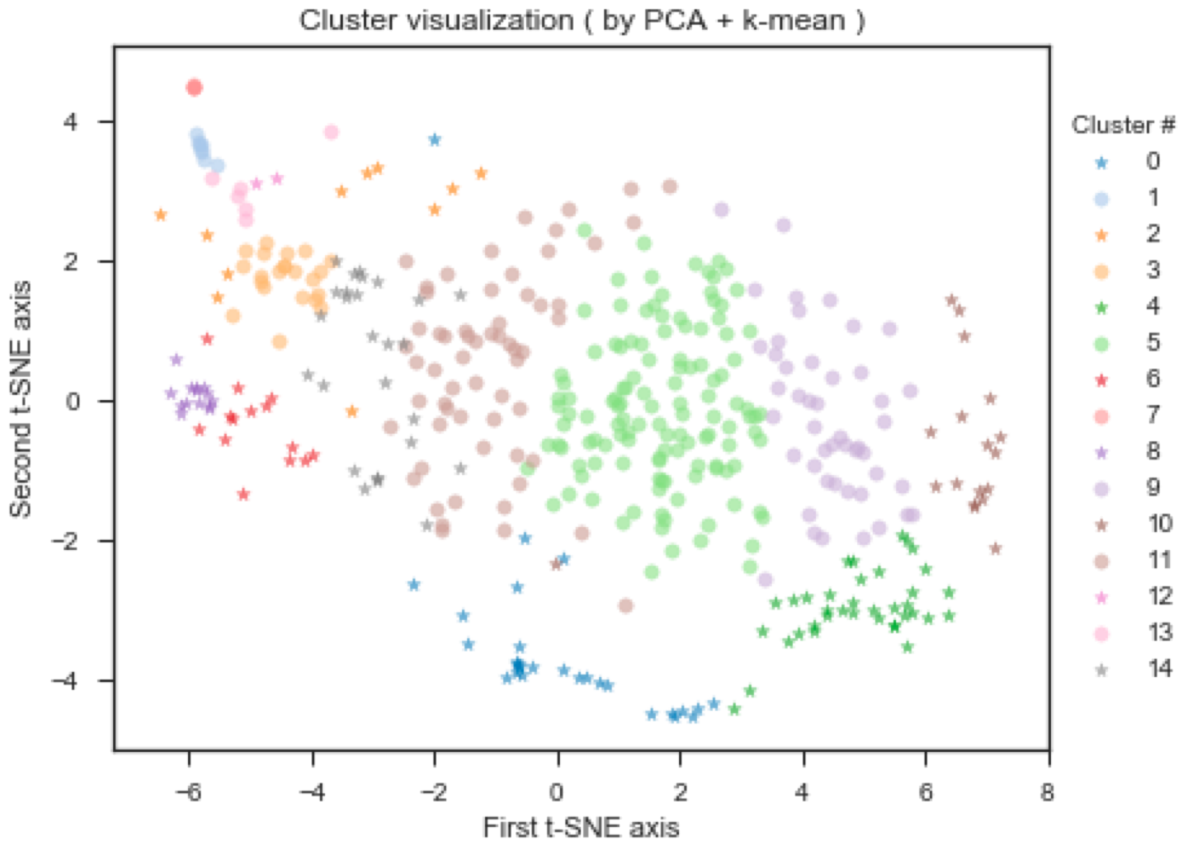
61 genes have repeats:

{'Gdh ', 'abba ', 'TfIIFbeta ', 'RpL22 ', 'CG34180 ', 'retinin ', 'Ggamma30A ', 'CG5343 ', 'CG14414 ', 'CG4673 ', 'CG3534 ', 'Aats-arg ', 'Pros28.1A ', 'Det ', 'rap ', 'ben ', 'sphinx ', 'CG34175 ', 'CG4933 ', 'CG42288 ', 'mars ', 'CG31742 ', 'ade5 ', 'Orc1 ', 'borr ', 'Nlp ', 'Vm32E ', 'Rrp40 ', 'CG7236 ', 'CG15390 ', 'Tom40 ', 'eIF2B-delta ', 'RpII18 ', 'CG13742 ', 'CG3246 ', 'CG16986 ', 'Spc25 ', 'Irc ', 'l(1)dd4 ', 'CG11905 ', 'Bsg ', 'CG8290 ', 'CD98hc ', 'CG34350 ', 'CG42336 ', 'Arc1 ', 'CR32661 ', 'CG3817 ', 'CCKLR-17D3 ', 'CG17612 ', 'Nc73EF ', 'dgt4 ', 'sw ', 'CG13482 ', 'scra ', 'cal1 ', 'CG14463 ', 'nAcRalpha-96Ab ', 'RpL10Aa ', 'CG3162 ', 'CG17737 '}

# PCA. PCA reduces data size from (424, 29) to (424, 10) and preserves 95% of variance.

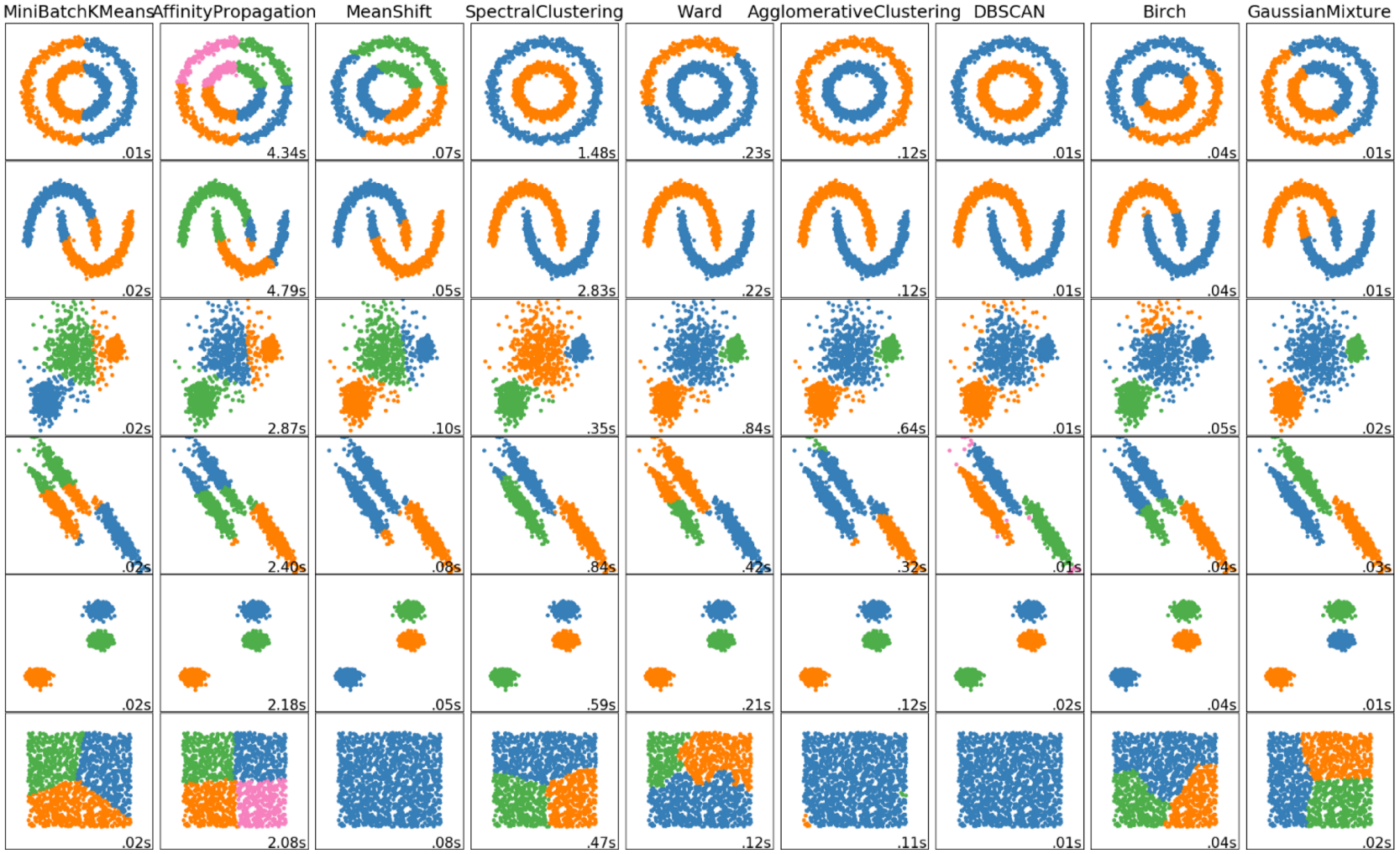
Although distorting the space, t-SNE helps visualizing the relative relationship among instances



# Comparing different clustering algorithms on toy datasets

http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html




Check file ‘Summary.pdf’ to see how these genes get clustered differently.

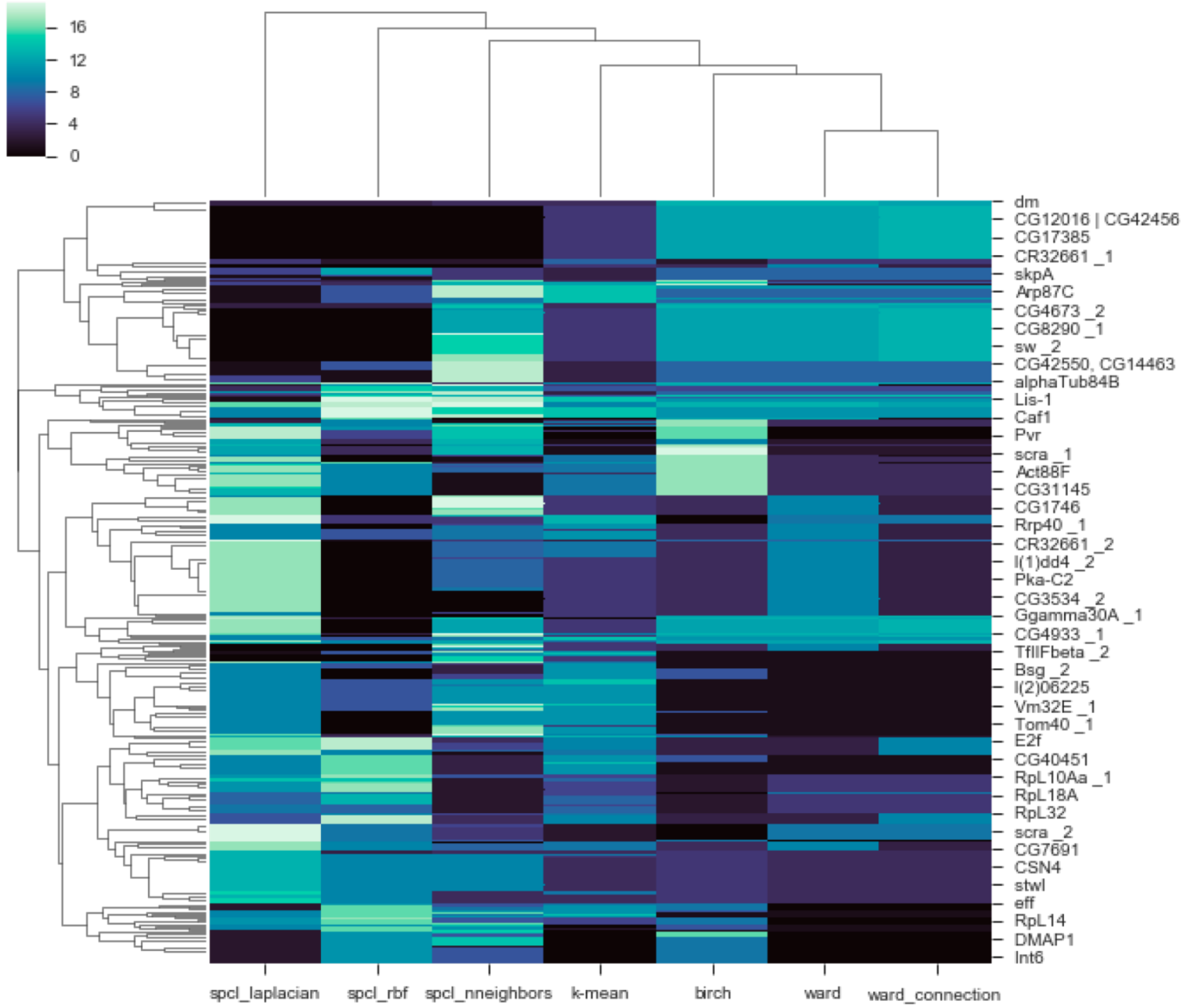
For more plots see

Cluster_visualization_by_PCA + k-mean.pdf
Cluster_visualization_by_PCA+birch.pdf
Cluster_visualization_by_PCA+k-mean.pdf
Cluster_visualization_by_PCA+spcl_laplacian.pdf
Cluster_visualization_by_PCA+spcl_nneighbors.pdf
Cluster_visualization_by_PCA+spcl_rbf.pdf
Cluster_visualization_by_PCA+ward.pdf
Cluster_visualization_by_PCA+ward_connection.pdf

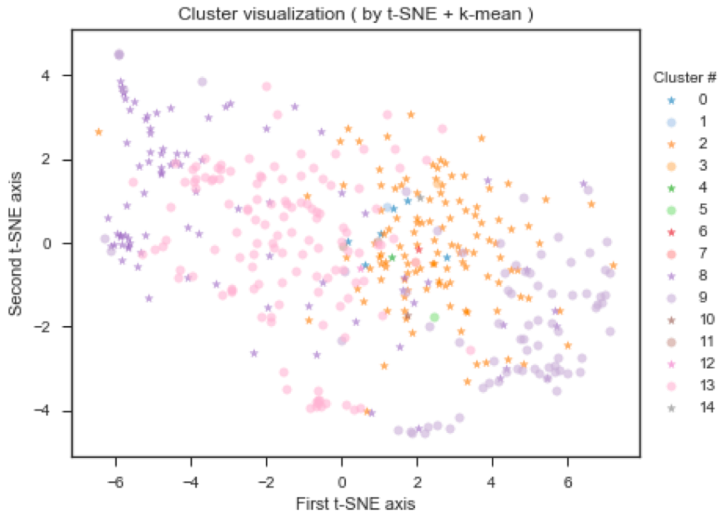
# Ensemble:

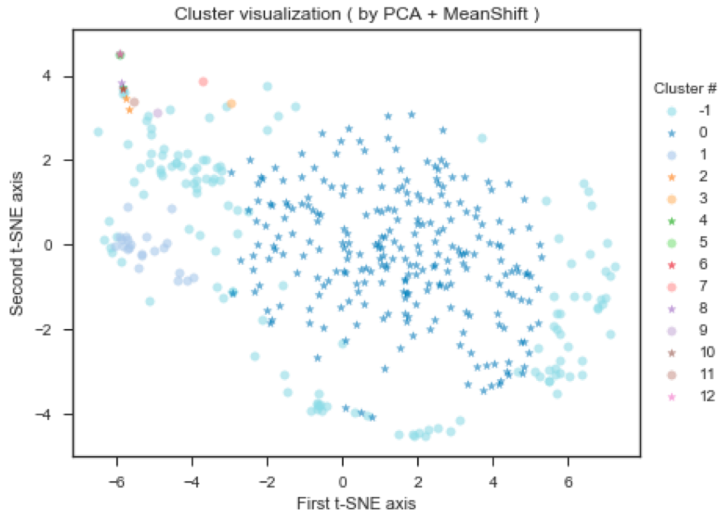
Check file ‘Ensemble.pdf’ and ‘Circle_dendrogram.pdf’ for dendrogram showing the final clustering result.

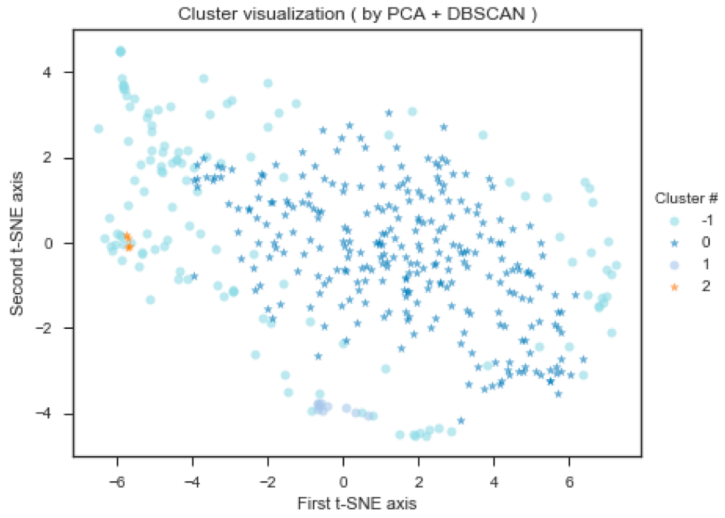
Try eye-balling those 61 genes with repeated instances to see where those twins gets mapped in either the t-SNE plot or dendrogram.
Ideally repeats of same gene would not separate much far away from each other if Hi-Fish and data analysis being both steady and robust.
However, I see not a few of them assigned into different clusters.



# Failed trials




- T-SNE and density based algorisms (MeanShift and DBSCAN) do not generate appropriate clusters here.
- Trying to comparing two dendrograms, only got a mess.