Clustering the data

Do tissue types match up with gene expression?

Reporter: Hsin Jou Yang
Partner: Huixin Fan

Introduction

Result and Discussion

K – mean (this sub tittle will be deleted.)

K – mean clustering was done by MatLab programing with different K input ranging from 40 to 10 clusters (Fig. S1). In general, most of the cell lines are very distributed in to clusters and we did not see a well match between clusters and tissue types. In k =15 cluster, cells lines in each tissue type seems to be highly overlap into the same clusters. This may because of too less cluster comparing to the sample type. K mean around 25 is very similar to result k = 31, so will not be discussed. Since result from too large cluster numbers (ex: k = 40) contains too separated tissue type cluster, nor will it be discussed. Since there are 31 tissue types, results from k = 31 will be mainly focused and k =20 will also be addressed.

In 31 clustering colormap (fig. 2), although most of the tissue type were not distinctly clustered (dark blue color), there are still some interesting data could be potentially pointed out. In cluster 16, only fibroblast cells lines were defined in this group. In addition to this, cells lines from breast, pancreas and

ovary are grouped into cluster 7,8,9, 10, 22, 24 and 25 in a very similar extent. Since the clustering is based on mRNA difference, this result indicates that their genetic expression could be similar. In some of the tissue type such as adrenal cortex, salivary gland, small intestine and so on, there data did not seem to be clustered in to any set. This is due to the very low cell line numbers (1 or 2 data available) are available for analysis. On the opposite hand, lung data with 198 cell lines data available which is much higher than most tissue type showed a very high cell number (yellow colored) in many clustered groups.

In 20 clustering colormap (fig. 1), shows a very similar pattern with k mean cluster 31 data. Fibroblast has are distinctly clustered. Breast, Ovary and cancer still have a very similar genetic expression.

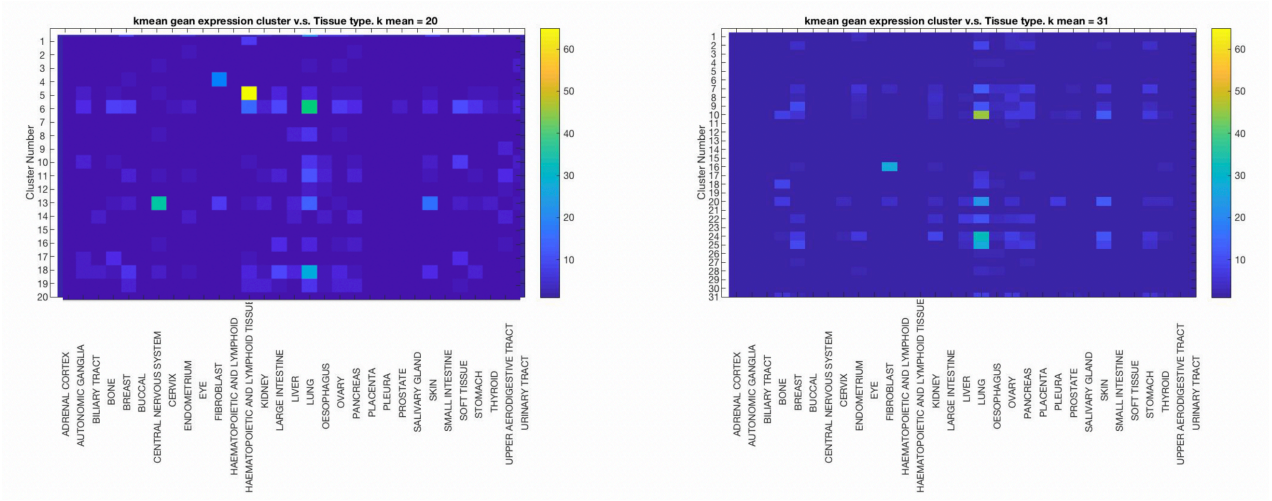

*Figure 1 &2 | **K mean clustering result with k = 20 & 31***

The numbers on y- axis represent the 20 or 31 cluster numbers, and the labeling on x - axis represent the 31-tissue type. Each small block contains a color represents the number of cell lines that are clustered in

to same group from one tissue. This color rank from dark blue (small number of cell lines) to light yellow (large number of cell lines).

*Figure 2 | **K mean clustering result with k = 20***

Method

- K- mean clustering algorithm

K – mean method was programmed on MatLab (full code in the supplemented Matlab K_mean.m file.) The MatLab function "`idx = kmeans(X,k)`" was the main function used here with the filtered data input as "x" and the number of clusters as "k". By default, this function uses squared Euclidean algorithm and the k – mean ++ algorithm [1]. With this function, K mean algorithm is compute through following steps [2]:

1. Randomly generate k numbers of means (centroids) on the data set.
2. Compute the distance from each observation to centroids through squared Euclidean algorithm.
3. K cluster are created by association every association with the nearest centroids.
4. New centroid of each K clusters becomes the new mean.
5. Repeated step 2 to 4 until convergence is reached.

Here, K input ranging from 50 to 10 were tested with our code, and then represented through colormap function in MatLab.

Reference

1. k-means clustering - MATLAB kmeans. Mathworks.com. https://www.mathworks.com/help/stats/kmeans.html. Published 2018.
2. Hamerly, Greg and Charles Elkan. "Alternatives to the k-means algorithm that find better clusterings." *CIKM*(2002)..

Supplementary

Project location	https://github.com/hsinjou0714/PHGY425final_project_cell_clustering
K_mean.m file	https://github.com/hsinjou0714/PHGY425final_project_cell_clustering/blob/master/K_mean.m

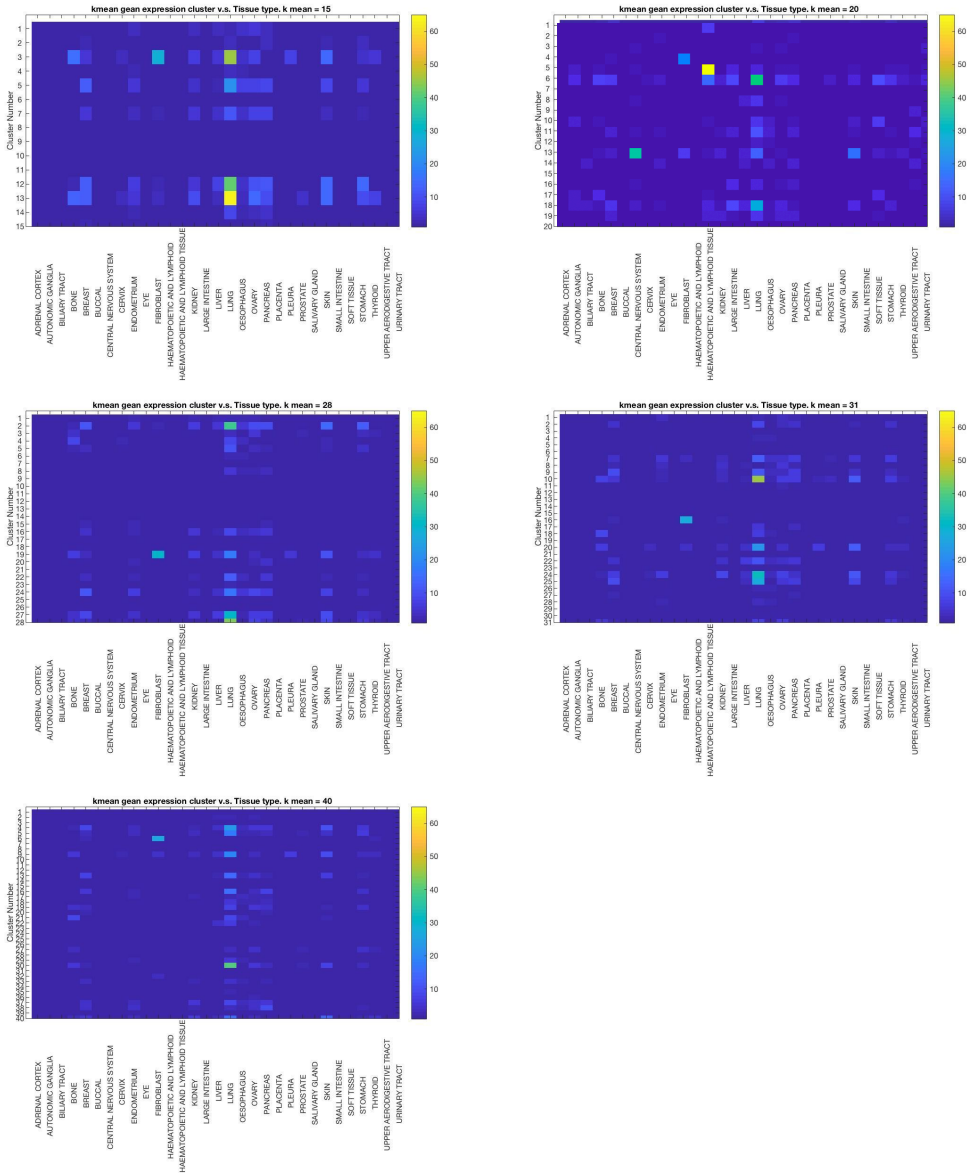

Figure S1 | **Complete K mean clustering result**

This diagram shows the color mapping of k = 15, 20, 28, 31 and 40 independently. The numbers on y - axis represent the cluster numbers, and the labeling on x - axis represent the 31-tissue type.