Hands-On #4

Clustering and BiClustering

1. Hierarchical Clustering:
    a. Log2-transform `all_aml_mll.mas5.gct` using PreprocessDataSet.
    b. Filter genes by taking the top 300 genes using MAD to generate `all_aml_mll.mas5.top300mad.gct` (as in HO #1) using VariationFiltering.
    c. Cluster using average linkage using the HierarchicalClustering module.
    d. Use Pearson correlation as the distance metric for genes and Euclidean for samples. Center (subtract mean) and normalize the data before clustering. Cluster again using single linkage. Compare the results.
2. NMF:
    a. Go back to the raw values by inversing the log2-transformation to ensure positive values only (use PreprocessData module) and save as `all_aml_mll.mas5.top300mad.raw.gct`
    b. Apply NMF with k=3 to the top300 genes. Cluster the samples to 3 groups by identifying the component with maximal value for each sample (use HeatMapViewer to view H.gct). Compare clusters to those obtained in 1.
    c. KEEP FOR THE END: Use Excel to generate 3 GCT files which represent the decomposition of the matrix – one for each. Save them as tab-delimited files named `all_aml_mll.mas5.top300mad.nmf#.gct`, where # is replaced with 1,2,3. View them with HeatMapViewer. Use global color scheme.
    Notice that there are many zeros in the matrices.
3. Consensus Clustering:
    a. Run ConsensusClustering on `all_aml_mll.mas5.top300mad.gct` to determine the number of clusters using hierarchical clustering, normalizing both rows and columns, use 10 iterations. After how many clusters the values in the consensus matrix are already bi-modal (look at the statistics.pdf file)?
    b. Look at the gif files that are produced or load the corresponding GCT file in HeatMapViewer.
    c. Rerun on `all_aml_mll.mas5.bottom300mad.gct` and compare the results (generate the bottom300 has in HO #1).
    d. Run NMFConsensus to determine the number of clusters with NMF. Run k from 2 to 5. View the cophenetic score as a function of the number of clusters. What is the optimal number of clusters? Are these the same clusters obtained by hierarchical clustering? (Warning: takes a long time. continue to the next step and return when finished).
4. "Manual" Bi-clustering (Coupled Two-Way Clustering):
    a. The files `cluster1.txt` and `cluster2.txt` contain lists of genes which form distinct clusters when running average linkage. Use HeatMapViewer on the results of hierarchical clustering using average linkage of the top300mad genes. Load the two files as 'feature lists' and identify the two clusters.

b. Filter the GCT using SelectFeaturesRows or ‘save dataset option in HeatMapViewer and save as two separate GCT files: all_aml_mll.top300.mad.c{1,2}.gct.
c. Apply average linkage to the samples (using Euclidean as a distance metric between columns and Pearson correlation between genes) in each of these datasets. Which samples cluster together in the two runs?