Hierarchical Clustering of Cancerous Tumors

R. **Teal** Witter ’20

Middlebury College

Goal

Our goal is to replicate the hierarchical clustering analysis from Johansson et al.† in Figure 1.

Background

Many cancer treatments have been specific to the affected organ rather than the genomic make-up of the cancerous tumors. However, tumors often behave according to their gene expressions. With the goal of developing organ-specific treatments, significant research has been tone to identify which are relevant to the likelihood that a tumor will metastasize (spread to other organs).

PAM50, a result of extensive research, is a set of 50 genes that characterize five subtypes of breast cancer tumors: Basal-like, HER2, Luminal A, Luminal B, and Normal-like. Johansson et al.† considered gene expression for 45 tumors from the Oslo2 study set. They used unsupervised hierarchical clustering on the available 37 (of 50) PAM50 genes to characterize subtypes.

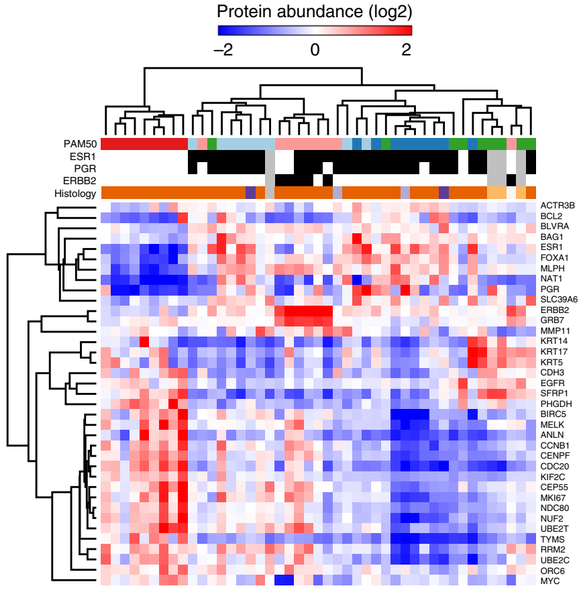

Figure 1: Unsupervised.

Algorithm

HierarchicalClustering(data, n, method)
  clusters = n lists each of a single data point
  while |clusters| > 1
    Ci, Cj = closest clusters using method
    remove Ci, Cj from clusters
    add Ci + Cj to clusters
  return clusters

Experiment

We use hierarchical clustering on the 37 available genes. Whereas Johansson et al.† utilized unsupervised clustering, we compare and contrast single, complete, and average clustering techniques.

The two-way clustering analysis naturally lends itself to two measures of success:

PAM50 subtype
- Basal-like
- HER2
- Luminal A
- Luminal B
- Normal-like

How closely do tumor clusters match PAM50 subtypes?

Are there patterns in the heatmap?

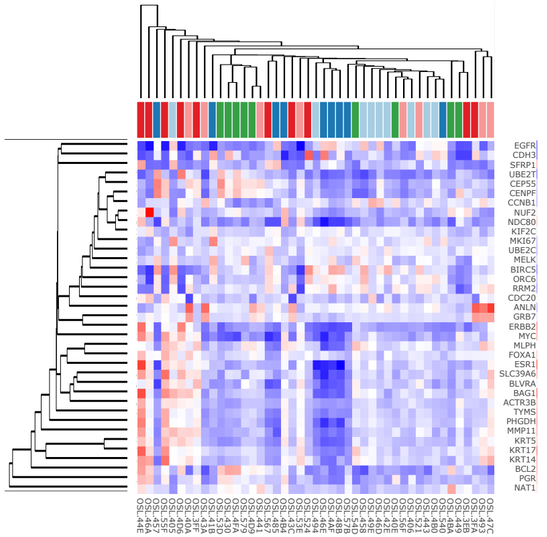

Figure 2: Single.

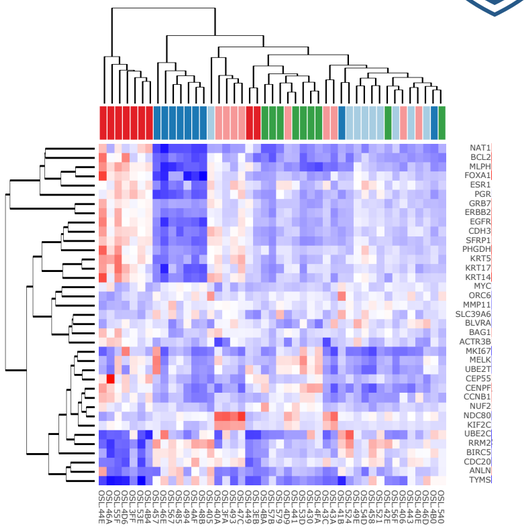

Figure 3: Complete.

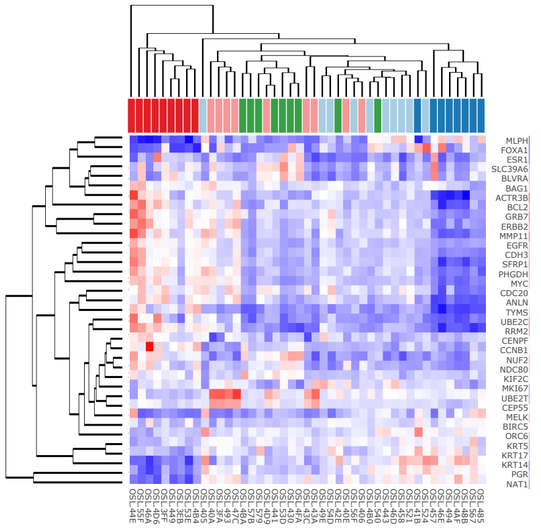

Figure 4: Average.

† Johansson, Henrik J., et al. "Breast Cancer Quantitative Proteome and Proteogenomic Landscape." *Nature Communications* 10.1 (2019): 1600.