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 done 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.

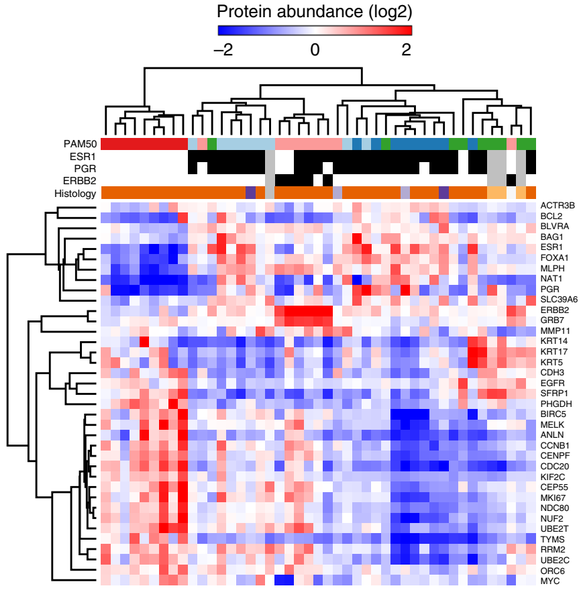

Figure 1: Unsupervised.

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?

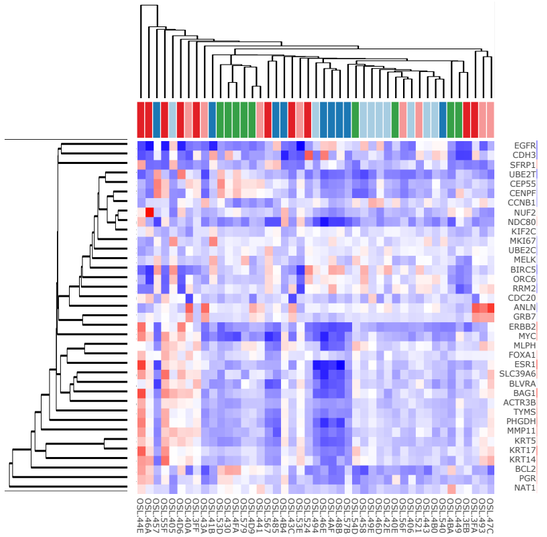

Figure 2: Single.

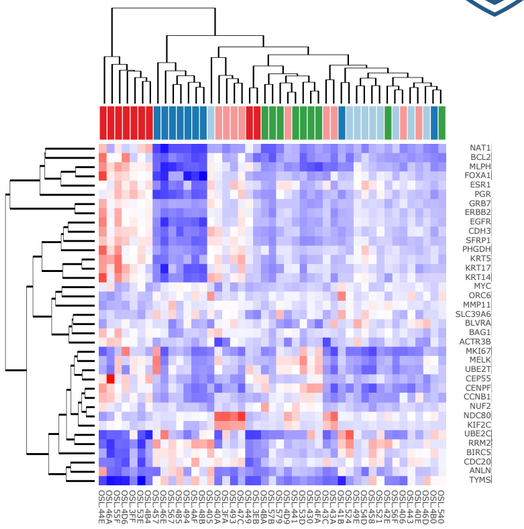

Figure 3: Complete.

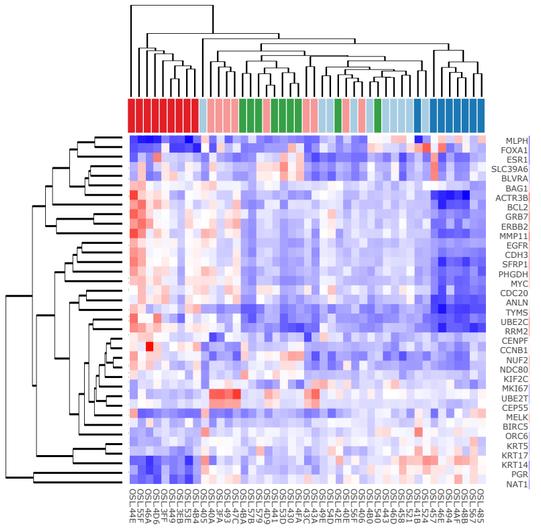

Figure 4: Average.

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

Figure 1: Single hierarchical clustering. Both tumor and gene clusters exhibit typical “chaining” behavior. The more data points in a cluster, the more likely that that cluster is closest to another cluster and will continue to grow. The smaller clusters do a good job of grouping Normal, Luminal A, and Luminal B, respectively. The heatmap has only one clear grouping of lower protein abundant genes in Luminal A tumors.

Ba Ba LA Ba LB Ba HE Ba HE LA No No No
No HE No LA LA Ba HE Ba LB LA LA LA LA
No LB LB LB LB No HE LB HE LB LB LA No
No Ba Ba HE HE

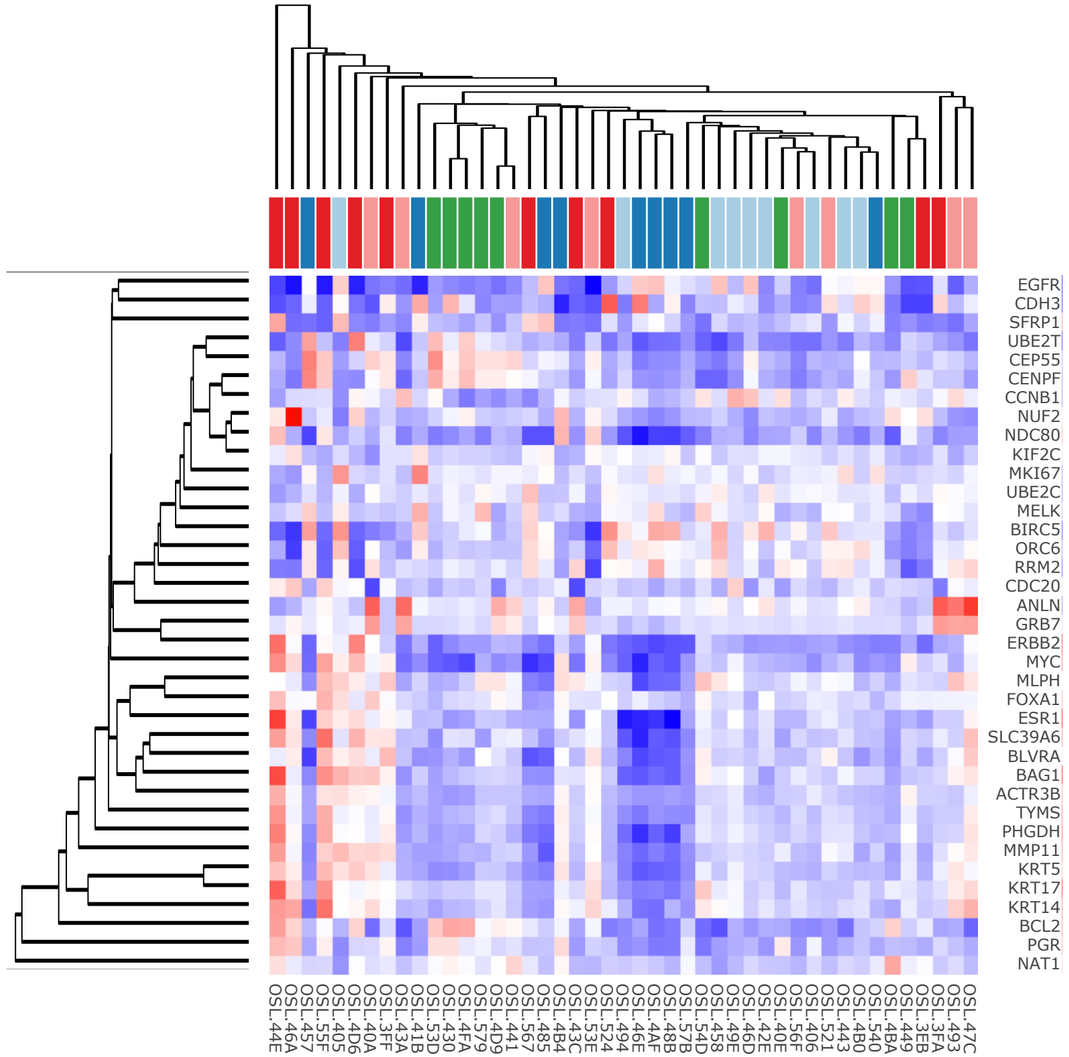

Figure 2: Complete hierarchical clustering. All five subtypes are clustered with the most prominent examples being Basal-like, Luminal A, and Luminal B. The heatmap has two clear lower abundant groupings with respect to Basal-like and Luminal A clusters. The top right corner is predominantly average with a mix of Normal, HER2, and Luminal B.

Ba Ba Ba Ba Ba Ba Ba LA LA LA LA LA
LA LA LB HE HE HE HE Ba Ba No No No
HE No No No No HE HE LA LB LB LB LB
LB No LB HE LB HE LB LA No

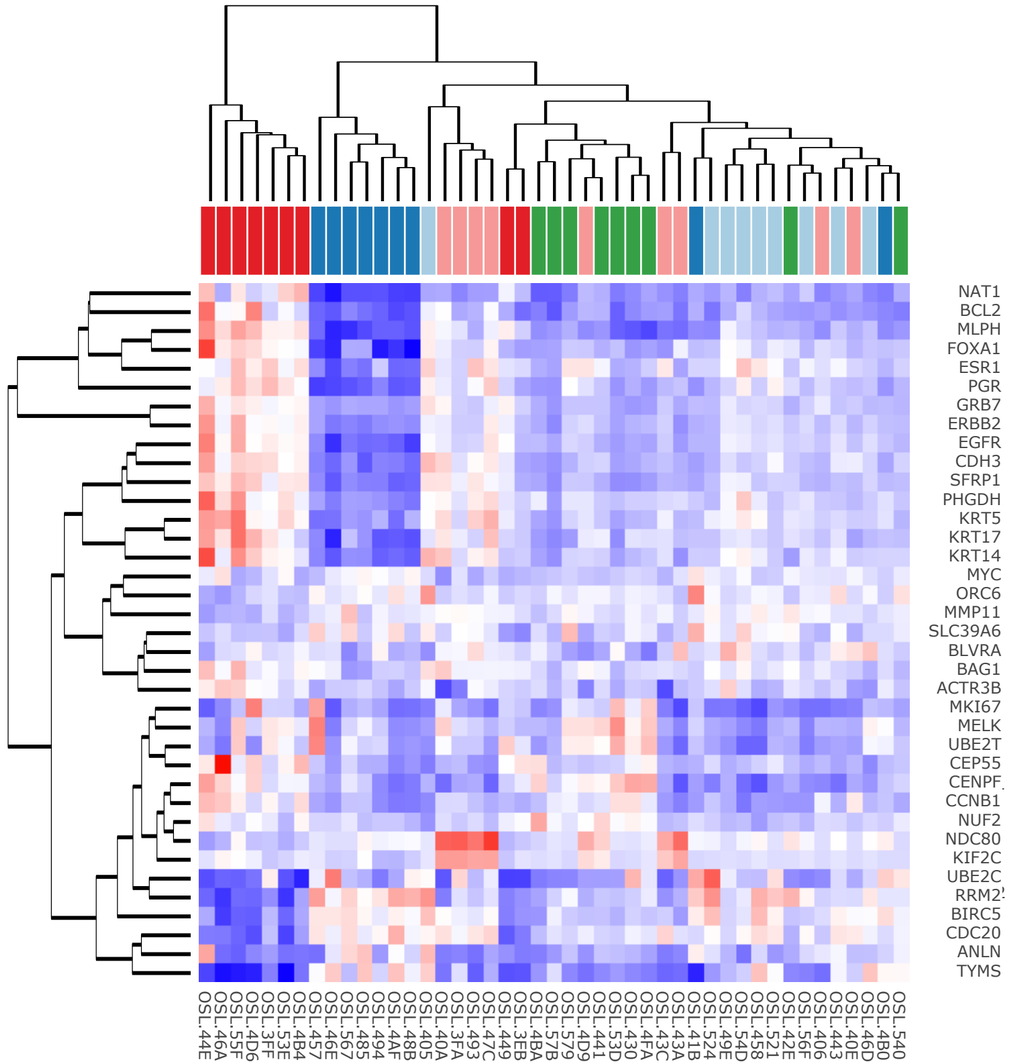

Figure 3: Average hierarchical clustering. Basal-like, Normal, and Luminal A are clustered while HER2 and Luminal B are interspersed. The heat map shows two lower abundant groupings again with respect to Basal-like and Luminal A subtypes. The remainder of the heatmap is not sufficiently grouped to draw meaningful conclusions.

Ba Ba Ba Ba Ba Ba Ba Ba Ba Ba LB HE HE HE HE No No No HE No No No No HE HE LB LB No HE LB HE LA No LB LB LB LB LA LB LA LA LA LA LA LA LA