**Filtering Amino Acid Coding Changing SNPs**

We use the Bioconductor package VariantAnnotation [1] to filter amino acid coding change SNPs. Variants that are 1) annotated as SNP, 2) pass the quality control filter, 3) happen on at least one allele and 4) causing amino sequence change are kept. Across all 645 patients, we identified 179,744 unique amino acid coding change SNPs.

**Partitioning Around Medoids (PAM)**

PAM is a way of implementing k-medoids clustering, a more robust version of k-means clustering [2]. To get the best clustering scheme, we use silhouette score [3] to evaluate the clustering performance under each k value (from 2 to 20). Result shows silhouette score increase as k goes up, reaches maximum value (0.22) when k ranges from 8 to 11and decreases as k is greater than 12, indicating the optimal classification scheme lies within k=8, 9, 10 and 11.

**Consensus Clustering**

Consensus clustering can 1) determine the number of clusters and 2) assess the stability of the discovered clusters by evaluating the consensus across multiple runs of a clustering algorithm (in our case PAM clustering) [4]. Result shows a general trend that as k increases, the PAC (proportion of ambiguous clustering) score decreases. Also, no significant decrease of PAC is observed if k goes beyond 7, indicating the optimal classification scheme lies within k greater than 7.

**Affinity Propagation**

Affinity propagation determines heterogeneities within data by exchanging messages between data points. Such process is repeated until a high-quality set of exemplars and corresponding clusters gradually emerges [5]. Affinity propagation gives clusters with few patients, and we consider those as non-representative. After removing clusters with less than 10 patients, we have 12 representative clusters (negative distance as pairwise similarity, clustering scheme slightly variates when different pairwise similarity measurements methods are used).

**Bipartite Network Modularity**

The relationship between SNPs and patients can be modeled with a bipartite network [6]. It has been reported that the heterogeneity information with the data can be reflected by the network [7]. Based on the constructed bipartite network, we measure modularity using method developed by Newman [8]. This is an especially powerful method compared to the above mentioned ones, because the cluster specific SNPs are also highlighted.

**References**

1. <http://bioconductor.org/packages/release/bioc/manuals/VariantAnnotation/man/VariantAnnotation.pdf>
2. Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2), 3336-3341.
3. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.
4. Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, 52(1-2), 91-118.
5. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. science, 315(5814), 972-976.
6. Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nature genetics, 45(10), 1127-1133.
7. Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.
8. Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69(6), 066133.