Supplementary Document for Frequent Subgraph Mining of Functional Interaction Patterns Across Multiple Cancers

Arda Durmaz, Tim A. D. Henderson*, and Gurkan Bebek**

Case Western Reserve University, Cleveland OH 44106, USA
{arda.durmaz2,tadh,gurkan.bebek}@case.edu

Supplementary Methods

FSG Clustering

Since the number of frequent subgraphs are relatively large, to evaluate biological significance we have clustered the protein-protein interaction network filtered to include only the genes found in frequent subgraphs using a greedy approach. Basically each gene starts as its own cluster and iteratively genes are merged together in a greedy fashion based on modularity metric.

Pathway Enrichment

Pathway enrichment for Reactome data set is done with filtering pathways that do not have at least 8 genes. For hypothesis correction bonferroni method is applied with p-value threshold 0.1. For enrichment of clusters identified using PAM on UMAP reduced dimensions, we have selected top 5% of genes with highest counts across different frequent subgraphs. More specifically for each cluster enrichment we try to identify frequent subgraphs that show significant dysregulation for samples in the given cluster compared to other samples with t-test and bonferroni correction with p-value threshold 0.05. Identified frequent subgraphs are then merged and top 5% genes with highest number of counts are selected for enrichment tests.

UMAP

UMAP is used for dimensionality reduction in the FSG space. Since the number of FSG mined is 135k clustering on the high dimension presents various challenges hence we initially applied UMAP which uses local manifold approximations to generate a reduced dimensional set on 2 dimensions to apply clustering. Simply the dysregulation matrix where columns represent frequent subgraphs and rows represent samples and the dysregulation score is the euclidean norm of standardized expression values per gene.

* Now at Google (tadh@google.com)
** Correspondence should be addressed to gurkan.bebek@case.edu.

Supplementary Figures

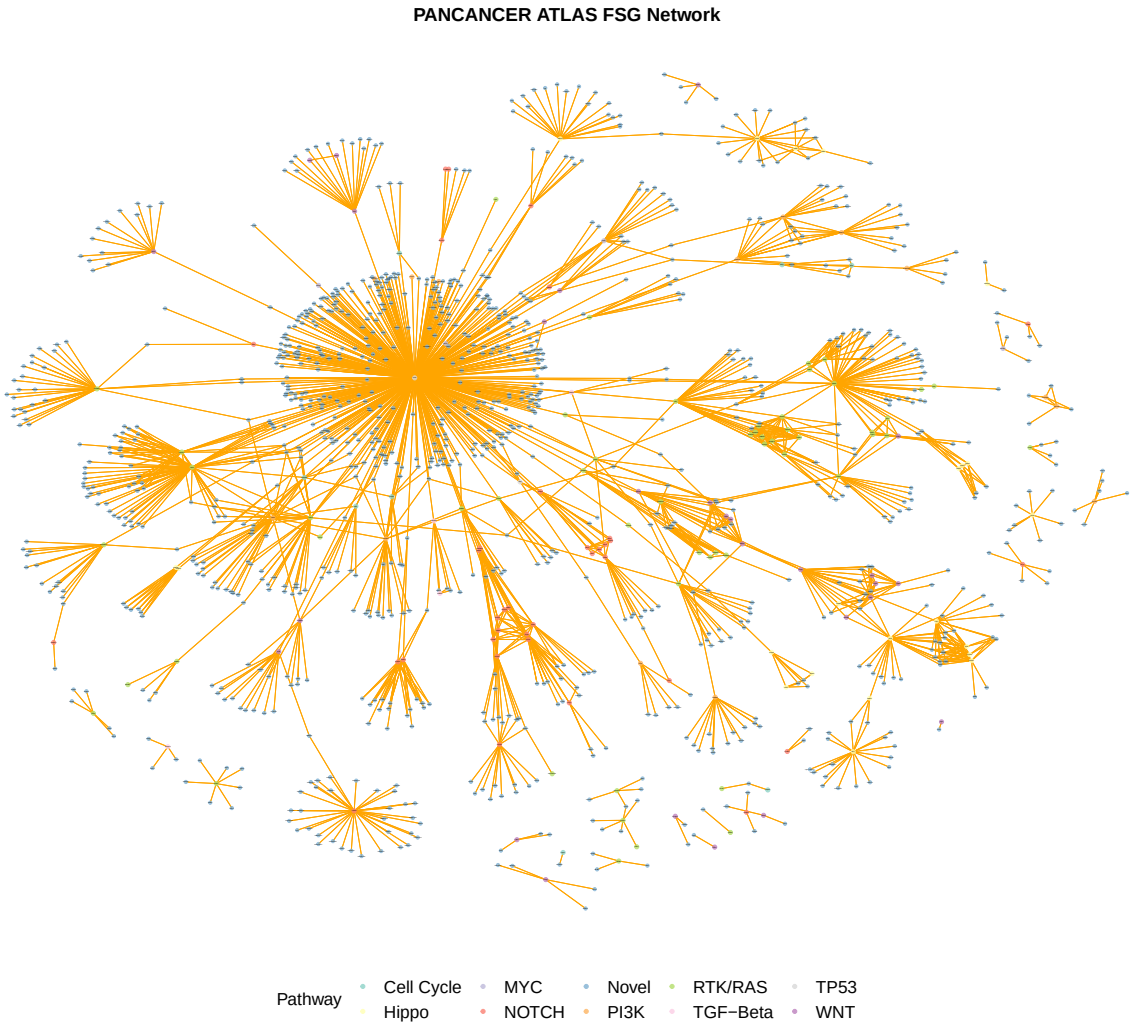

Fig. S1. Network representation for identified frequent subgraphs showing overlap between expert-curated pathways. Nodes are colored based on the overlap with the pathway and novel nodes (not found in the expert curated pathway list) are colored differently. Note that additional interactions that do not interact with any gene in the expert curated pathways are not shown.

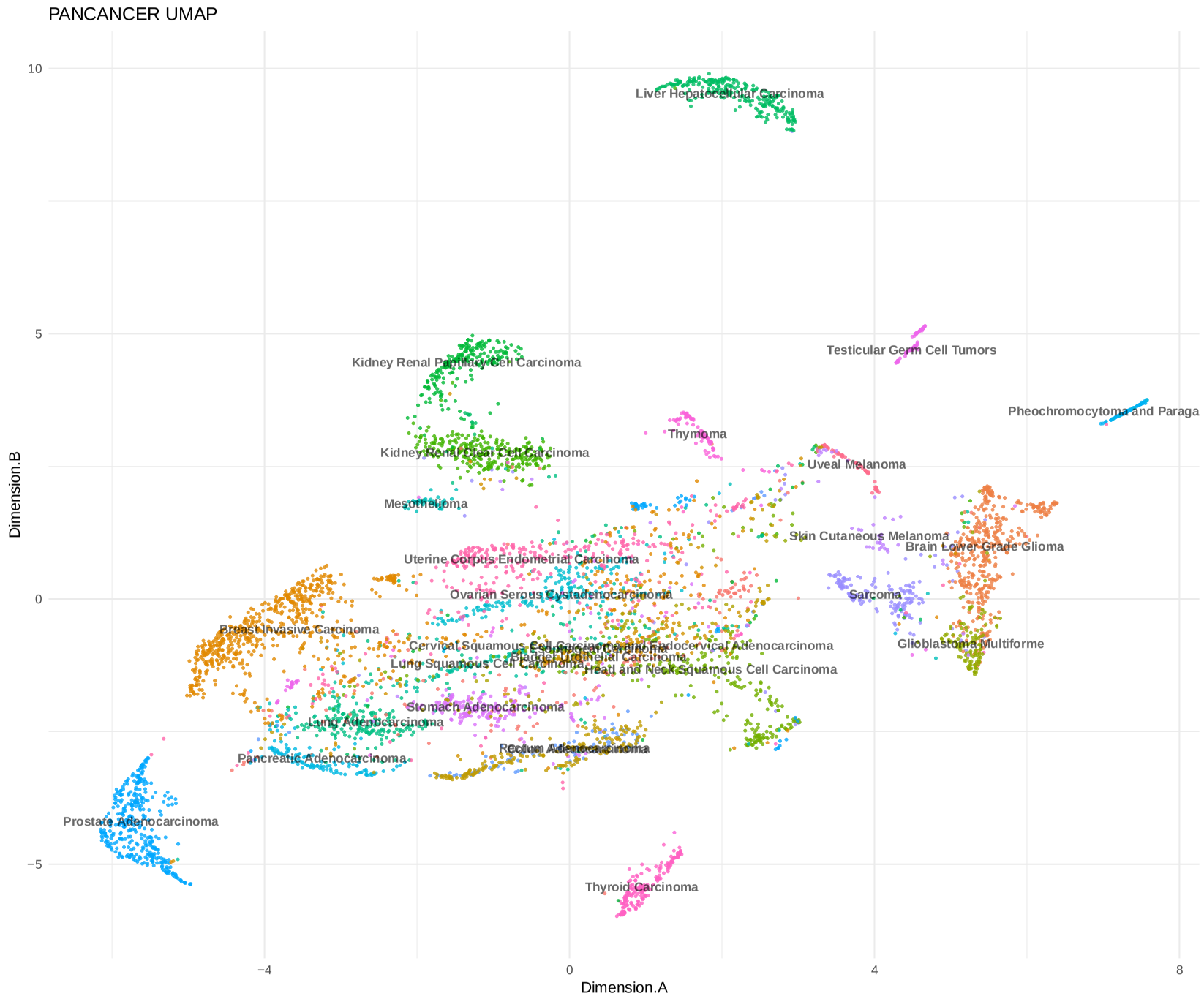

Fig. S2. UMAP dimensionality reduction on scored frequent subgraph matrix with support 8 (larger version of Figure in paper)

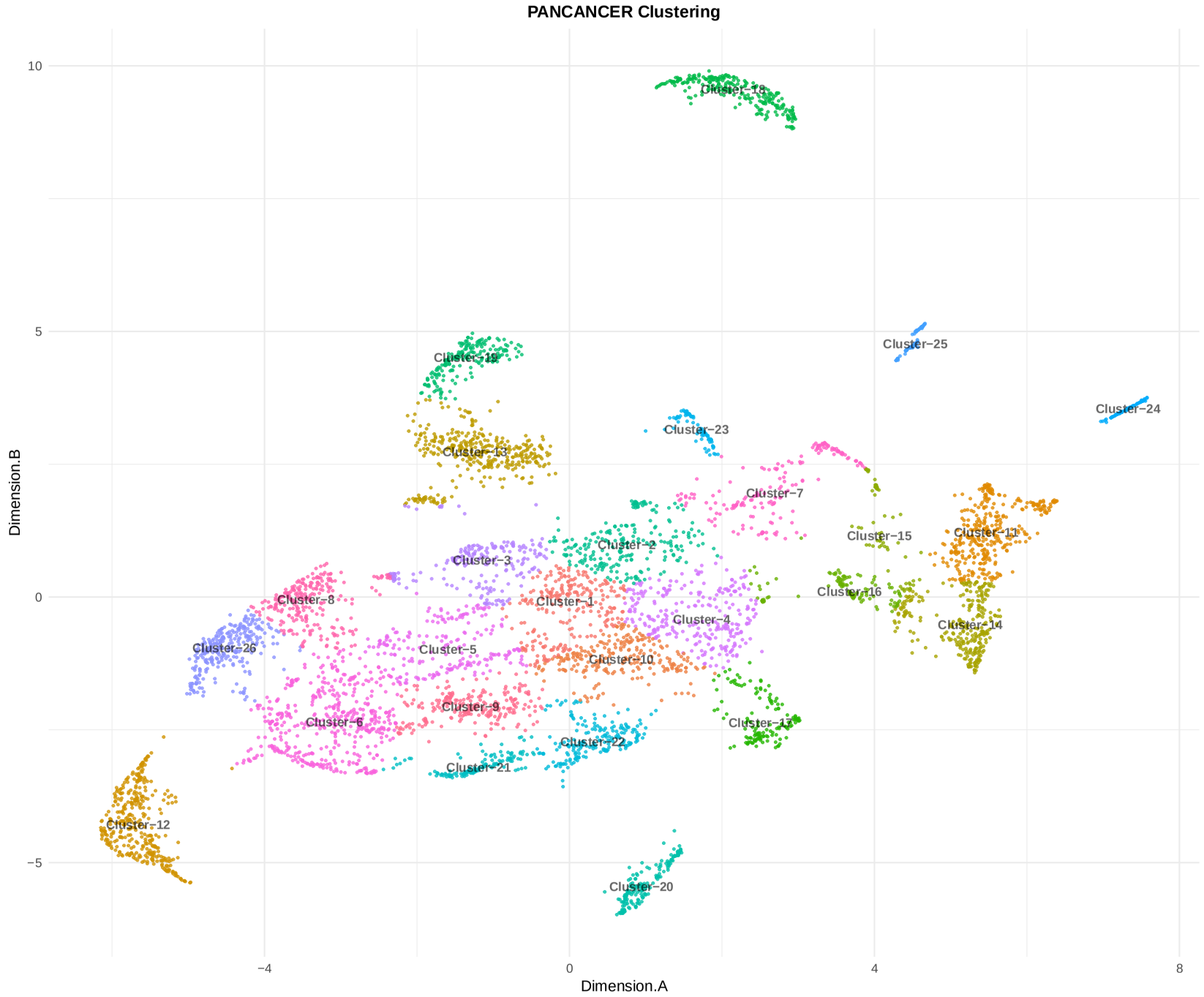

Fig. S3. UMAP dimensionality reduction on scored frequent subgraph matrix with support 8 with PAM clustering labels.

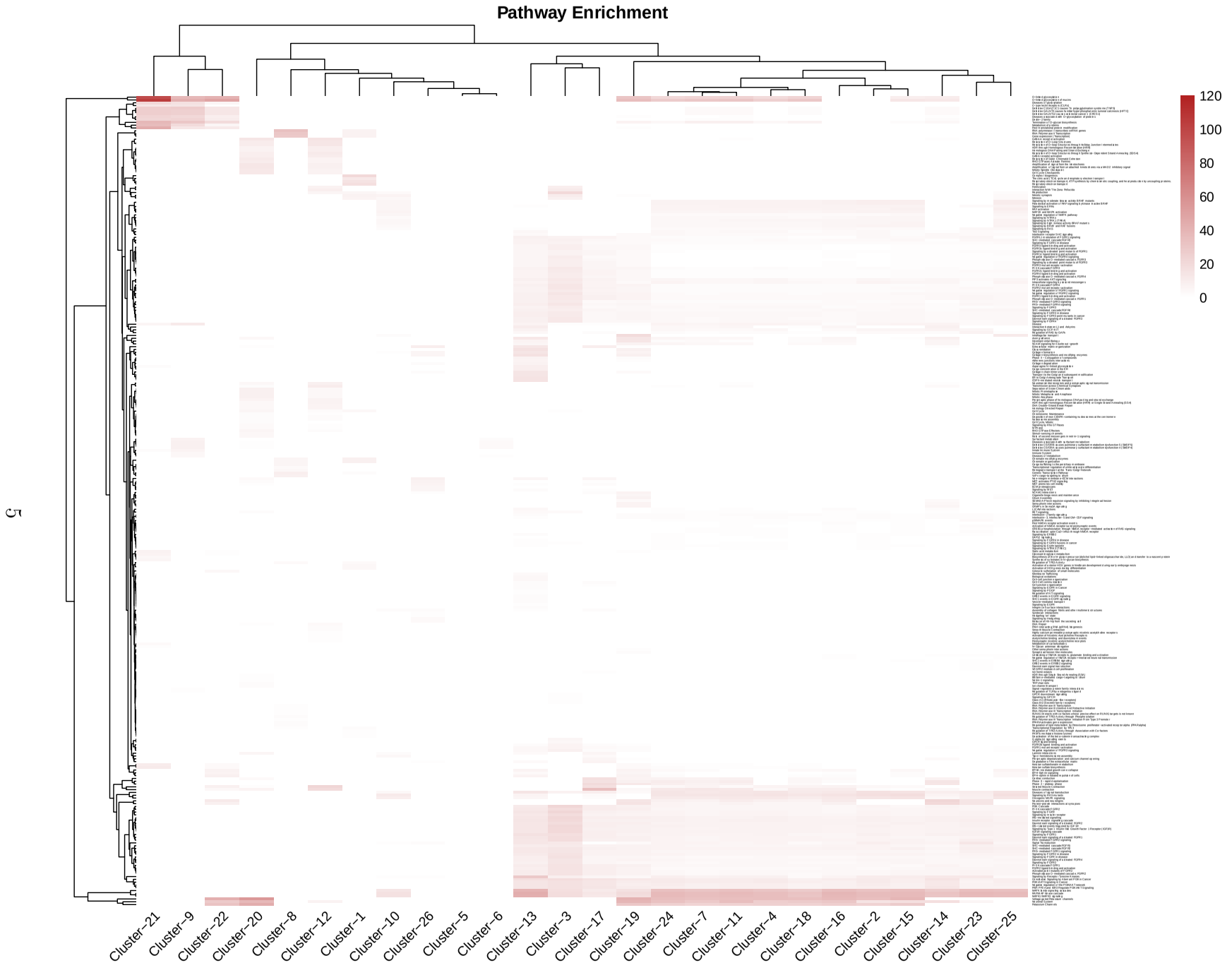

Fig. S4. Reactome pathway enrichment for significant subgraphs between PAM clusters.

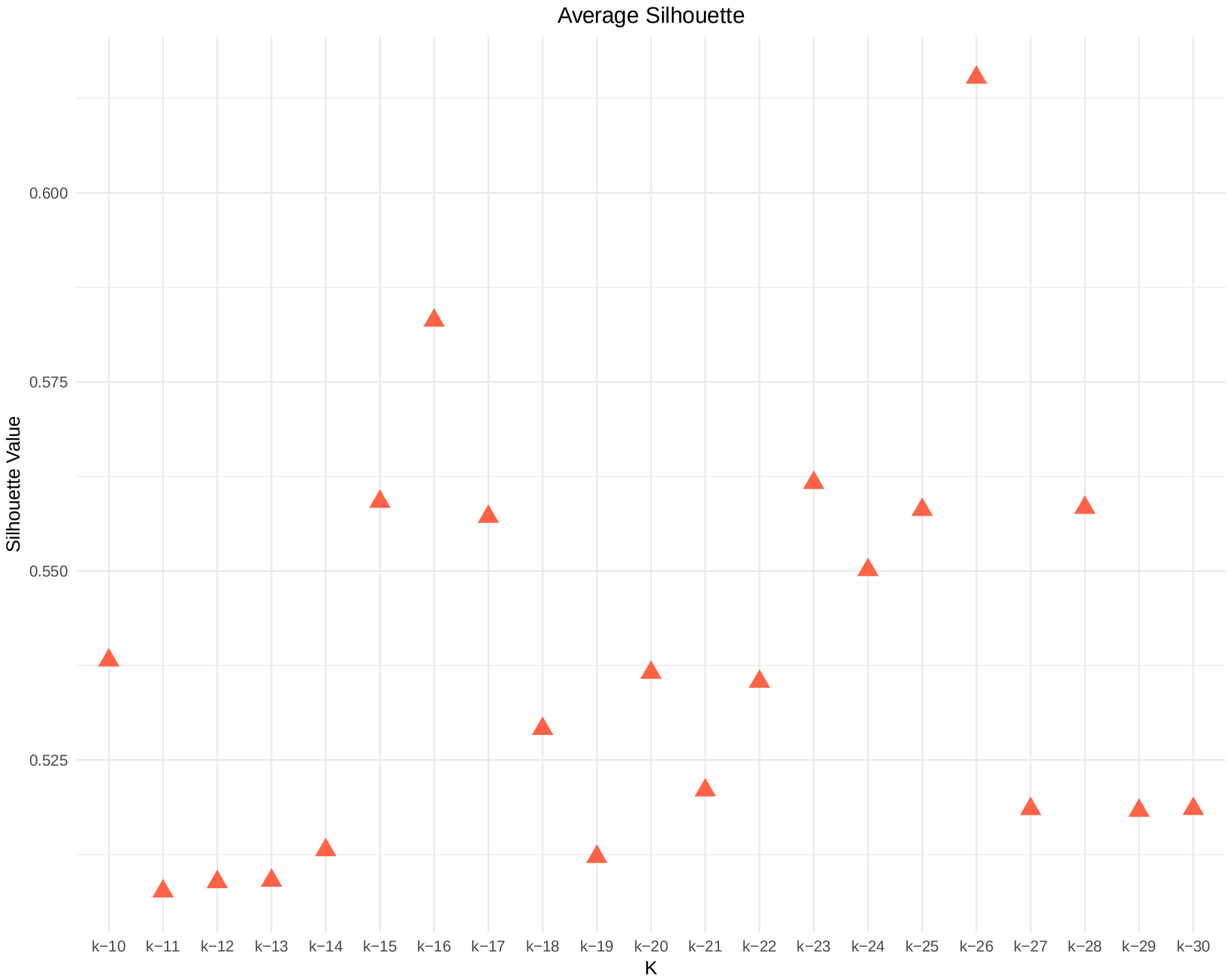

Fig. S5. Average silhouette values for number of cluster selection showing k-26 as best.

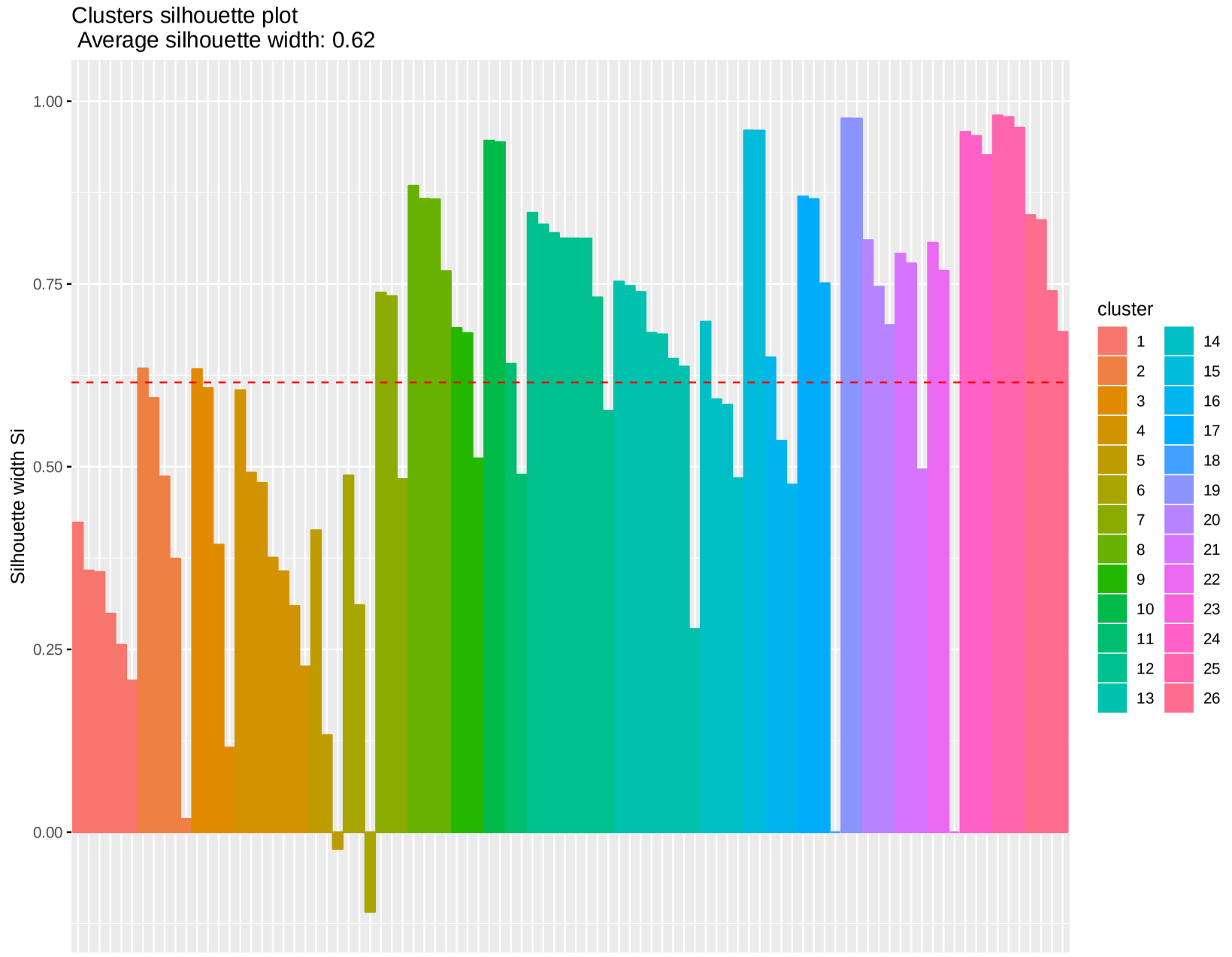

Fig. S6. Silhouette plot for given k-26 number of clusters with average silhouette value 0.62

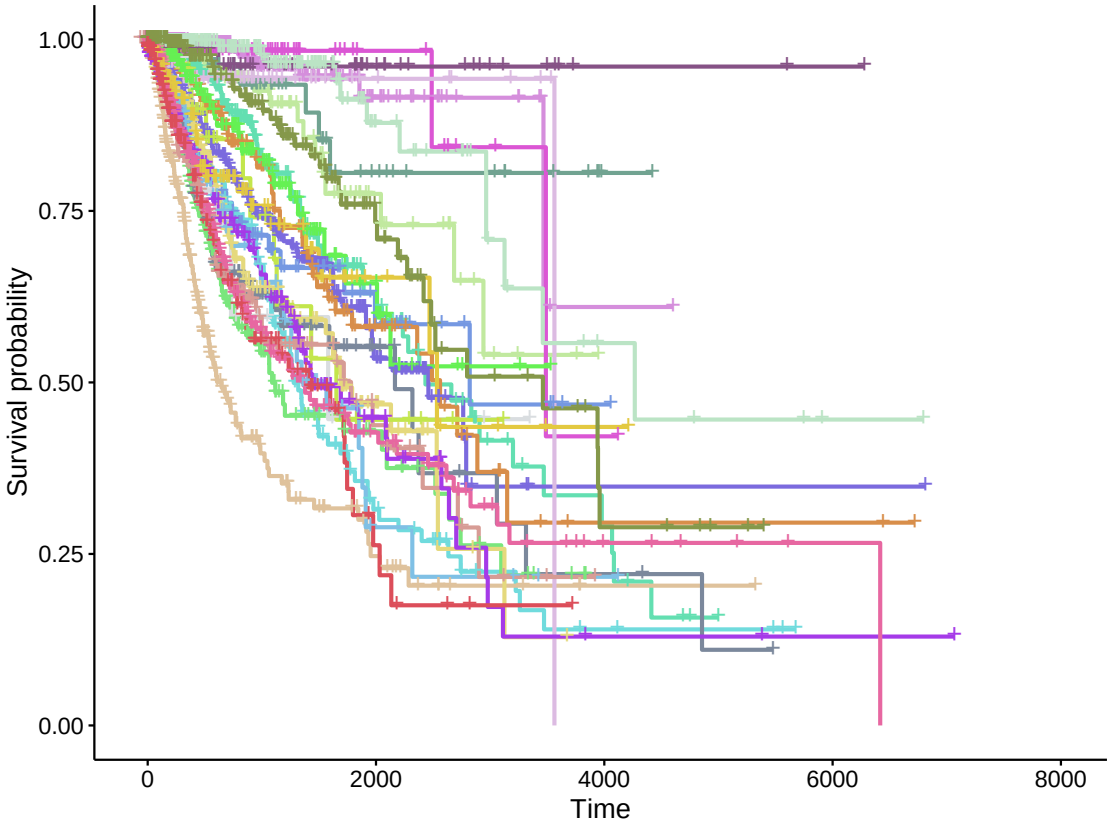

Fig. S7. Kaplan-Meier curves and estimates of survival for patient samples that correspond to the FSM clusters are shown

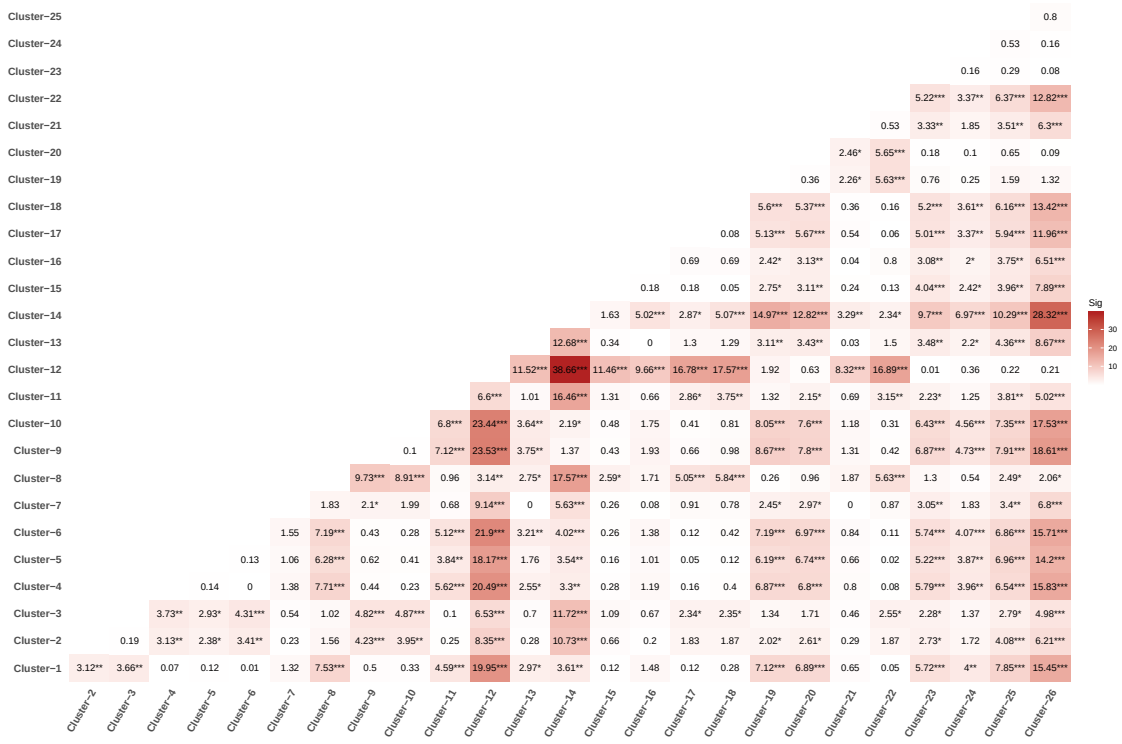

Fig. S8. Pairwise survival differences using log-rank test are shown for FSM patient clusters.

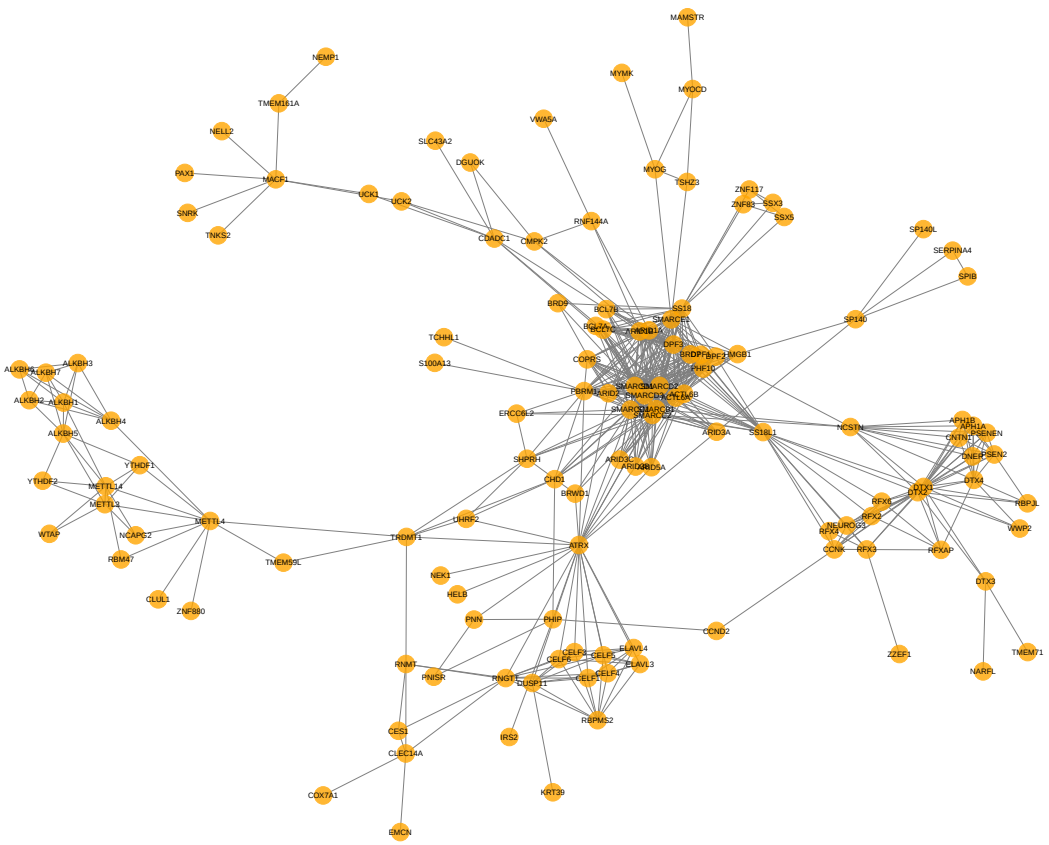

Fig. S9. Cluster 63 network identified using greedy clustering of the frequent subgraphs

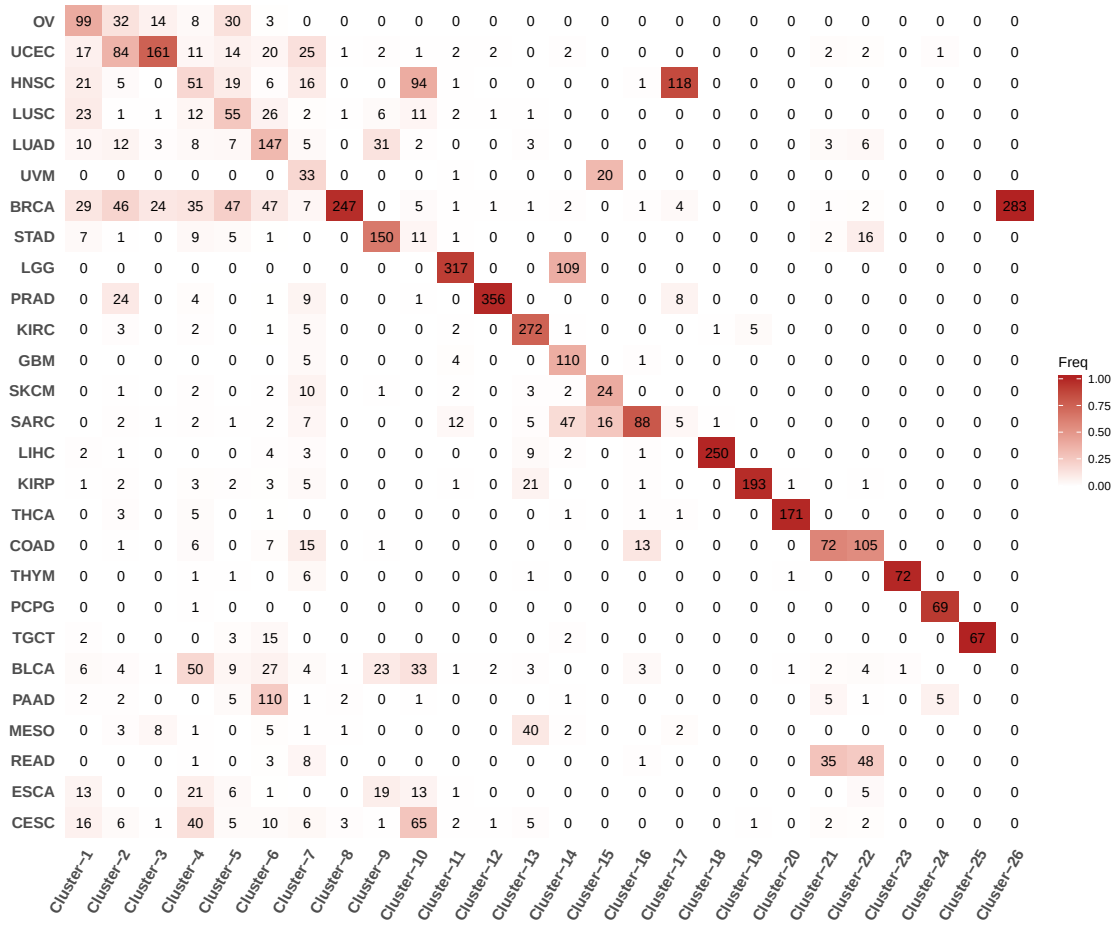

Fig. S10. Disease representation per cluster. Values represent the number of patients belong to a given disease found in the cluster. Color coding represents the frequency in column-wise order.