Supplementary Materials for *Frequent Subgraph Mining of Personalized Signaling Pathway Networks Groups Patients with Frequently Dysregulated Disease Pathways and Predicts Prognosis*

Pacific Symposium of Biocomputing (PSB) 2017

Arda Durmaz[*], Tim A. D. Henderson[*], Douglas Brubaker, Gurkan Bebek[†]

1 Method Overview

The method proposed [8] aims to integrate gene expression data and protein-protein interaction data to identify dysregulated modules given in a sample set using a novel approach. We hypothesize that complex machinery in the cell and dysregulation of the crucial mechanisms leading to diseases such as cancer are regulated by relatively small core modules in contrast to pathways consisting of 50 - 200 genes. The approach is to mine a large network of protein-protein interactions weighted by gene expression profiles to identify network modules of size 4 to 8. Identified modules are then quantified for dysregulation and samples are clustered using Non-negative matrix factorization (NMF) to reveal distinct groups. In this study we focused on identifying subtypes of Glioblastoma Multiforme (GBM) and Breast Cancer (BRCA).

$$ES_{i,j} = \sqrt{G_i^2 + G_j^2} \tag{1}$$

$$DS_d = \frac{1}{n} \sum_{e=1}^{n} ES_e \tag{2}$$

1.1 Dysregulated Signaling Pathways

Dysregulated Signaling Pathways are labeled graphs (Section 2.2) where vertices represent proteins and edges represent dysregulated activation/inhibition interactions. They are constructed from mRNA expression data (Section 3) and known PPI data [6, 20].

[*]Co-first Author
[†]Corresponding Author; gurkan.bebek@case.edu

Dysregulation is computed by constructing a matrix $\mathbf{P}$, where $\mathbf{P}_{i,a}$ is the standard score of expression level of gene a for patient i. Then an *interaction matrix* $\mathbf{S}$ constructed from $\mathbf{P}$ in Equation 3. In Equation 3 (ab) represents two genes a and b such that the protein encoded by a interacts with the protein encoded by b. The variable i represents a particular patient.

$$\mathbf{S}_{(ab),i} = \sqrt{\mathbf{P}_{i,a}^2 + \mathbf{P}_{i,b}^2} \tag{3}$$

To determine if the relationship between two genes a and b is dysregulated for patient i the *z-score* for each interaction is computed. In the equation below, $\mu(\mathbf{S}_{(ab),\cdot})$ and $\sigma(\mathbf{S}_{(ab),\cdot})$ respectively refer to the mean and standard deviation of the dysregulation scores for genes a and b.

$$Z(\mathbf{S})_{(ab),i} = \frac{\mathbf{S}_{(ab),i} - \mu(\mathbf{S}_{(ab),\cdot})}{\sigma(\mathbf{S}_{(ab),\cdot})} \tag{4}$$

If $Z(\mathbf{S})_{(ab),i} > c$ then an edge $a \rightarrow b$ is included in the graph for patient i indicating a and b are dysregulated. In Section 3 the constant c, the z-score threshold, was set to 2 to mine for top %2.5 dysregulation.

1.2 Frequent Subgraph Mining

The *Frequent Subgraph Mining* problem was originally posed in the year 2000 by Inokuchi et al. [14] as an extension of frequent subset mining. Generally, algorithms works on graphs with labels on both vertices and edges. Algorithms proceed by enumerating subgraphs and counting the number appearances of each subgraph in the database by using subgraph isomorphism tests or embedding lists and graph isomorphism tests.

AGM [14], gSpan [27], MoFa [2], FFSM [12], and GASTON [21] demonstrate a variety of strategies for enumerating frequent subgraphs. SPIN [13] is an example of algorithm which finds just the maximal (that is largest) frequent subgraphs. For many real world datasets, enumerating all or even the just the maximal frequent subgraphs may be impractical.

All algorithms, whether explicitly or not, operate on a search space defined by the *connected subgraph lattice* of the graph database. This lattice is also known as the *partial order graph* of the *connected subgraphs* of the graph database. All of the connected subgraphs of the graph at the bottom are arranged as a lattice. The vertices of the lattice are subgraphs and the edges represent adding to the subgraph at the tail of the edge to obtain the subgraph at the head. Since the vertex at the tail of an edge is a subgraph of the vertex at the head the lattice represents a partial order on the subgraph relationship.

Unfortunately, even finding just the maximal frequent subgraphs often finds many more subgraphs than are useful. This is do to large amounts of overlap between many of the maximal frequent subgraphs. For instance, it is not uncommon for to maximal frequent subgraphs to differ in only one or two edges. Which if both are included causes the subgraph to no longer be frequent but if just one (or the other) is included the subgraph is frequent. Furthermore, finding every maximal frequent subgraph is an expensive operation even with the best algorithms as the problem is inherently exponential in nature.

```
1  # param start: frequent single vertex subgraphs
2  # param score: a function to score queue items
3  # param max_size: the max size of the queue
4  # param min_sup: int, amount of support
5  # returns: a generator of frequent subgraphs
6  def qsplor(start, score, min_sup):
7      while not start.empty():
8          queue = [ start.pop() ]
9          while not queue.empty()
10             lattice_node = take(queue, score)
11             kids = lattice_node.extend(min_sup)
12             for ext in kids: add(queue, score, ext, max_size)
13             yield subgraph
14 def add(queue, score, item, max_size):
15     queue.append(item)
16     while len(queue) >= max_size:
17         i = argmin(score(idx, queue) for idx in sample(10, len(queue)))
18         queue.drop(i)
19 def take(queue, score):
20     i = argmax(score(idx, queue) for idx in sample(10, len(queue)))
21     return queue.take(i)
```

Figure S1: QSPLOR: a new algorithm for mining a subset of frequent subgraphs.

1.3 QSPLOR: Mining a Subset of Frequent Subgraphs

Figure S1 shows pseudo code for QSPLOR a new algorithm to mine a subset of frequent subgraphs. It proceeds as a graph traversal of k-$\mathcal{L}_{\mathcal{D}}$ (the k frequent connected subgraph lattice of the graph database). It begins the traversal at each lattice node representing a frequent subgraph containing only one vertex. At each outer step it initializes a queue with one of the starting lattice nodes. Then in each inner step it removes an item of the queue. The `take` function removes one item from a uniform sample of the queue such that a user supplied scoring function is maximized.

On line 11, the lattice node is extended. This involves finding all possible one edge extensions to the subgraph represented by the lattice node. The ones that are frequent are returned by the extend method. After the extensions are found they are added to the queue with the `add` method. If the queue is at the maximal size after the addition, one item from the queue is dropped. The dropped item is from a uniform sample of the queue and minimizes the user supplied score function. After all extensions have been processed the subgraph is output with the `yield` statement.

The key to our algorithm is the user supplied scoring function which guides the traversal. The simplest scoring function simply returns a uniform random number. This will cause the traversal to be unguided. Complex scoring functions can prioritize certain labels or structures. The best general scoring functions are those that prioritize *queue diversity* such that traversal is encouraged to explore as much of the lattice as possible.

One such function which encourages queue diversity is based on a graph walk kernel. Let the graphs be represented as an adjacency matrix $\mathbf{E}$ with the labels of the vertices represented by a labeling matrix $\mathbf{L}$ (constructed from the set of labels L labeling function l). Equation 5 defines a distance function between two graphs which incorporates both

structural differences and labeling differences.

$$\text{walkd}(a,b) = \left\| \left(\mathbf{L} \cdot \sum_{i=1}^{|\mathbf{E}_a|} \mathbf{E}_a^i \cdot \mathbf{L}^T \right) - \left(\mathbf{L} \cdot \sum_{i=1}^{|\mathbf{E}_b|} \mathbf{E}_b^i \cdot \mathbf{L}^T \right) \right\|_2 \tag{5}$$

A scoring function can then be easily constructed from such a distance function by having the function maximum be the graph which is most distant from all other graphs in the queue. This ensures that when a graph must be skipped by the `add` function, it is the graph that is most similar to the graphs in the queue. Conversely, when an item is taken from the queue for processing it is the graph which is most dissimilar from all other graphs in the queue.

To find a precise computational bound of QSPLOR is complex as it involves characterizing the behavior of the queue with respect to a particular scoring function. That behavior is in part driven by underlying structure of the frequent connected subgraph lattice. There are two special cases which are easy to analyze: when the queue has a max size of 1 and when the queue is unbounded. For a queue size of 1, the number of steps is bounded from above by $\mathcal{O}(\frac{g^{h+1}-1}{g-1})$ where g is the size of the graph and h is the size of the subgraph. For queue of unbounded size the complexity is the same as complete frequent subgraph mining: $\mathcal{O}(2^g g^h)$. Finding a closed expression for other queue sizes is difficult but it is guaranteed to fall between these two bounds.

1.4 Non-Negative Matrix Factorization

Clustering via Non-Negative Matrix Factorization (NMF) is used to partition patients into subgroups. Section 3 shows that the partitions are prognostically discriminative between the patient subgroups. NMF method was first proposed by Lee and Seung [16] with the aim of decomposing images into explanatory basis vectors. NMF has also been used on gene expression data [3].

To apply NMF, first the frequent sub-pathways identified by QSPLOR in Section 1.3 are quantified by the amount of dysregulation in each patient. A score for each frequent sub-pathways $H = (V_H, E_H)$ and each patient x is calculated and stored in matrix $\mathbf{V}$ by Equation 6. Rows of matrix $\mathbf{V}$ correspond to the frequent sub-pathways and columns correspond to patients. The matrix $\mathbf{S}$ in Equation 6 is constructed in Equation 3.

$$\mathbf{V}_{H,x} = \frac{1}{|E_H|} \sum_{(a,b) \in E_H} \mathbf{S}_{(ab),x} \tag{6}$$

The matrix $\mathbf{V}$ is then input to NMF for clustering. NMF decomposes $\mathbf{V}$ into 2 components with non-negative entries, $\mathbf{V} \approx \mathbf{W} \times \mathbf{H}$. $\mathbf{W}$ is called the basis matrix and $\mathbf{H}$ is the coefficient matrix, Rows of matrix $\mathbf{W}$ (size $\mathcal{F} \times n$) correspond to frequent sub-graphs and columns correspond to basis vectors. Entries of matrix $\mathbf{H}_{i,j}$ (size $n \times P$) represent the coefficient of basis vector i for patient j. Given the factorization, samples are clustered into n groups based on the coefficients. NMF requires number of groups n to be predetermined. We utilized both consensus clustering and other metrics for selecting the optimal number of clusters.

Multiple refinements have been made to NMF [3, 17, 15, 22]. To achieve more localized and compact clusters, we utilized nsNMF [22] which uses a non-smoothness constraint [9].

1.5 Consensus Clustering

The NMF method does not converge to same clustering with each run hence consensus clustering is applied with 150 runs and random seeding. For NMF method we have used Non-Smooth NMF which is best suited for more localized pattern identification [22]. Consensus clustering combines multiple runs with consensus matrix C. Consensus matrix incorporates the number of co-occurence of two samples. Normalized consensus matrix is then used as a distance matrix by $1 - C$ for hierarchical clustering to determine group membership.

1.6 Pathway & Transcription Factor Enrichment

In order to determine biological impact of the findings, we have conducted pathway enrichment analysis based on Reactome pathways using available online tool on Reactome website. Transcription factor enrichment is provided on EnrichR website [5]. We utilized transcription factors as terms and proteins interacting as variable sets hence allowing to assess transcriptional dysregulation in the data.

1.7 Coverage

In order to determine a reasonable threshold for FSM method we have calculated *coverage* for each dataset. Coverage is the total sample count achievable given the z-score threshold. More specifically given the threshold for edge scores, how many patients are represented overall. The aim is to both maximize sample count and threshold for relevant stratification. We have applied a thresold of 2 standard deviations since increasing the threshold further would decrease the required information in the dataset meaning that the threshold would be too strict.

1.8 Transcription Factor PPI

We compared the proteins from identified dysregulated sub-pathways against a transcription factor library using the online tool EnrichR [5]. The library contains transcription factors as terms and related genes as sets.

1.8.1 Enrichment in Glioblastoma multiforme

Enrichment for transcription factors using genes from the short survivor group in [25] found 52 genes that are significantly overrepresented in interacting proteins including; *Tp53, Foxm1, Rad21, Bmi1, Myc.* These proteins play crucial roles in cancer and GBM progression. Literature suggests *FOXM1* as a potential drug target for glioma patients [18]. *MYC* and *BMI1* are proto-onco genes which sustain stem cell renewal in GBM patients by repressing tumor suppressor pathways [1, 24, 4]. *CCNE1* and *CCND1* cyclin proteins are also overrepresented and are associated with poor survival [26].

Eight proteins in the long survivor group in the Verhaak *et al.* data are significant; *Stat3, Stat5A, Stat5B, Dlx4, Stat1, Smad4, Ctnnb1, Klf5.* The Stat proteins are important elements of signal transduction processes and are crucial elements for GBM proliferation. Inhibition of Stat3 is suggested to positively correlate with inhibition of cell proliferation in GBM stem

cells [23]. Furthermore Stat3 inhibition is also associated with decrease in Temozolomide resistance suggesting possible markers for future therapies. Also *Smad4*, another important signal transduction element, with reduced expression is suggested to correlate with survival of GBM patients [11].

1.8.2 Enrichment in Breast Cancer

Running enrichment tests for breast cancer data for microarray and RNA-Seq data revealed *Ilf3, Ilf2, Nacc1, Hdac2, Smarcc1, Ccnd1, Brca1* in short survivor group. In contrast long survivor groups were enriched in *Stat1, Stat3, Stat5a, Bcl3, Ctnnb1, Notch1, Htt, Myb*. Transcription factors have shown to be crucial for development and progression of cancer hence comparison of TFs in short and long survivors might reveal functional and possible therapeutic applications.

1.9 Rank Survey

Rank survey is used to determine the number of clusters based on the 'quality' of the consensus clustering. Multiple metrics are used to assess the quality of the clustering; Cophenetic correlation coefficient, residuals, explained variance, average silhouette, sparseness and dispersion. Additionally visual inspection is conducted on the consensus matrices. Cophenetic correlation coefficient is measures as the correlation between the distance matrix and the resulting clustering given the distance matrix. Sparseness is the average scores of the basis matrices, silhouette scores are average number of patients in the clustered groups. To summarize we require the metrics to be above the randomized results. For the given plots, dashed lines represent randomized results obtained by randomizing the features of the original matrix.

1.10 Survival

We have utilized log-Rank test for estimating significance of survival curves. Additionally since the dataset contained death events below 30 we have removed those data points instead of imputation. The reason for removing the points is to reduce noise incorporated by events that do not represent survival profile related to disease at hand. Figure S4 shows the survival curves of patients stratified according to intrinsic subtypes defined previously.

2 Validation

We compared our method against 2 recently published work integrating PPI and pathway information: *Pathifier* and *NCIS*. Pathifier quantifies dysregulation of a given pathway based on gene expression data and principal curve analysis [7]. To separate the samples, hierarchical clustering with average linkage is applied. *NCIS* [19] integrates gene expression and network data with the clustering process to identify clinically significant groupings of samples and genes. Prior to clustering genes are weighted according to their *'importance'* in the network and weighted co-clustering algorithm based on semi-nonnegative matrix tri-factorization developed by the same authors is used. The algorithm groups both genes and samples. Group

numbers are selected based on multiple runs of the algorithm with different parameters and results of the consensus matrices are evaluated based on cophenetic correlation coefficient.

2.1 Pathifier

Pathifier identified pathways related to survival in a GBM microarray dataset from TCGA [7]. Pathifier's survival analysis is supervised in comparison to our unsupervised approach. In the analysis, the groups are manually selected based on observations of a scoring matrix constructed from large pathways. This leads to failure to stratify patients that are dysregulated with different machinery in the same pathway. We used the Pathifier package from Bioconductor [10] on the TCGA GBM dataset with the KEGG pathways. The package requires normal samples 5 of which were downloaded from TCGA using the Cancer Genome Browser.

Pathifier identified 6 groups with significant differences in survival (Figure S14a). The number of samples in each group does not suggest biologically relevant clustering (n = 6, and the larger clusters are not significant in terms of survival). The separation distances between groups are not robust, as shown in the heatmap in Figure S14b. Cophenetic correlation coefficient quantifies clustering robustness and defined as the correlation between dissimilarities of each pair and the corresponding cophenetic distances. A coefficient of 0.61 of the Pathifier method does not suggest a robust clustering. Finally, Pathifier suggests which pathways contain dysregulation but makes it difficult to pinpoint the dysregulated interactions due to the large number of interactions in the pathways used in the scoring matrix.

2.2 NCIS

NCIS [19] identified 4 previously established subtypes in the GBM microarray dataset in conjunction with a curated PPI network. The network was constructed by the authors from Reactome, NCI-Nature Curated PID, and KEGG. It consists of 11,648 genes, 211,794 interactions matching 7,183 genes in the GBM dataset. The identified subtypes are similar to established subtypes and have significant differences in survival. However, it is not clear how the patients are clustered since previously identified subtypes do not provide overall significant survival difference.

We compared our method to the NCIS results and found 5 clusters (based on the clustering metrics) which show separation of survival curves (Figure S15a). We were also able to cluster previously proposed mesenchymal and proneural subtypes with further stratification of mesenchymal group (Figure S15b). Based on the survival analysis, proneural clustered groups show the longest survival curves in agreement with previous findings (Figure S15a).

The longest survivor group was highly enriched in axon guidance, collagen degradation, and extracellular matrix organization similar to our previous analysis. The long survivors were also enriched in VEGA-VEGFR2 pathways which was found to be highly associated with cancer. The short survivor group 2 was enriched in RNA Polymerase processes including transcription-coupled NER process (Nucleotide Excision Repair) and mRNA splicing. As previously identified, GPVI-mediated activation cascade, RHO-GTPase, and immune system processes are also enriched.

Our method performed better than the NCIS algorithm in terms of significance of survival stratification and relevance of the identified genes and pathways which can be used as precursor targets for future therapeutic studies.

2.3 Cross-Platform

We have applied the digraphs mined using breast cancer microarray data to rna-seq data to assess whether the method is eligible for cross platform application. The identified subgraphs were informative such that 5 clusters are identified with significance $p - value < 0.05$. However subgraphs identified with rna-seq data did not provide informative clustering in microarray data. This might be due to the heterogeneous structure of the data since no filtering were applied to TCGA rna-seq dataset.

2.4 Multiple Runs

To further assess the power of frequent subgraph mining we have evaluated the mined networks separately using bootstrap approach. Multiple sets of runs were combined using hierarchical clustering on consensus matrix. We were able to obtain silhouette value of 0.72 which suggest occurrence of a strong pattern.(Figure S17)

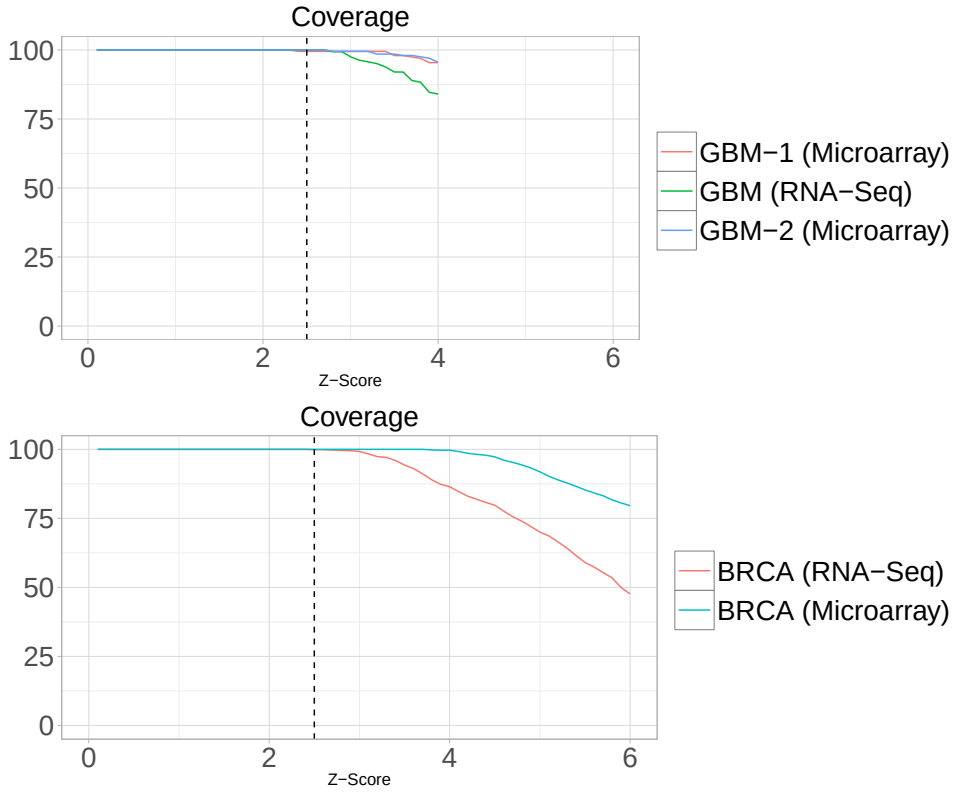


Figure S2: Coverage plot for GBM and BRCA datasets. GBM-1 data is from Verhaak study and GBM-2 data is from NCIS study. Y axis represents percentage of samples and x-axis represents z-scores. Z-Score 2.5 is the point where decrease in sample coverage begins for GBM datasets. For BRCA z-scores 3 and 4 are decreasing points however since we do not want to apply a strict threshold, 2 standard deviation is chosen as threshold for all datasets

Figure S3: Transcription factor protein - protein interaction network enrichment. (a,b) represents breast cancer microarray data short and long survivor groups respectively. (c,d) represents breast cancer RNA-Seq data short and long survivor groups respectively. (e) represents glioblastoma multiforme short survivor group. Figures (f,g) represents short and long survivors using subgraphs obtained from breast cancer microarray data

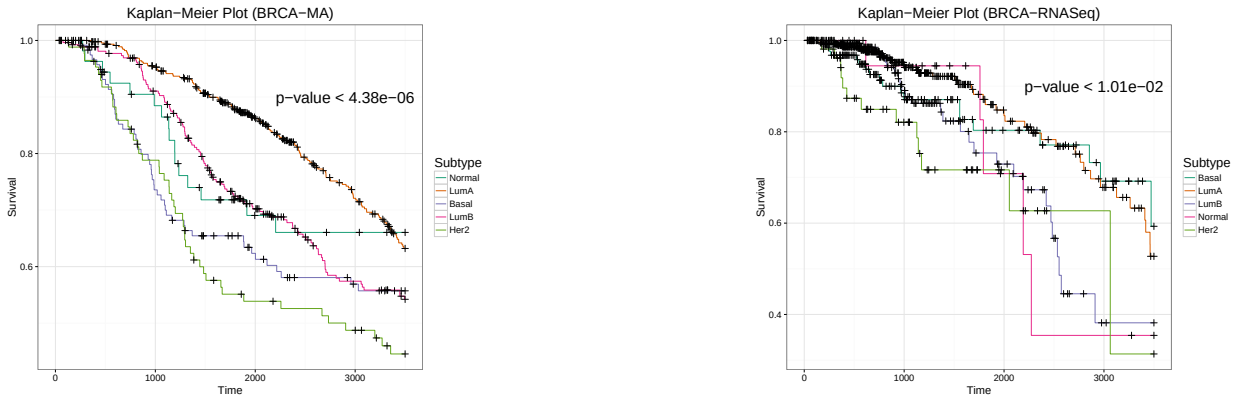


Figure S4: Kaplan-Meier curves for breast cancer and glioblastoma multiforme data comparing previously identified subtypes (intrinsic).

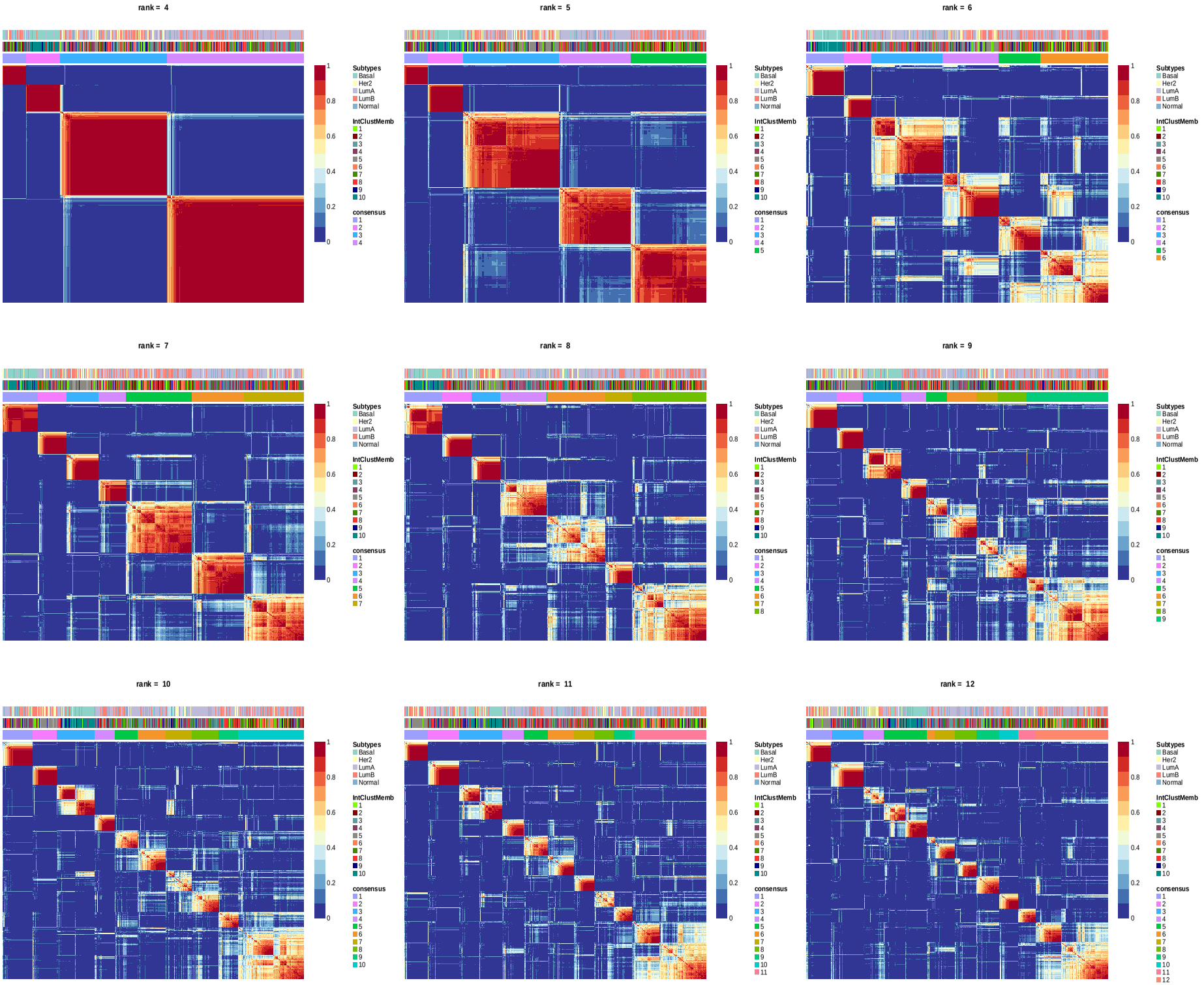


Figure S5: Consensus clustering results of breast cancer microarray data

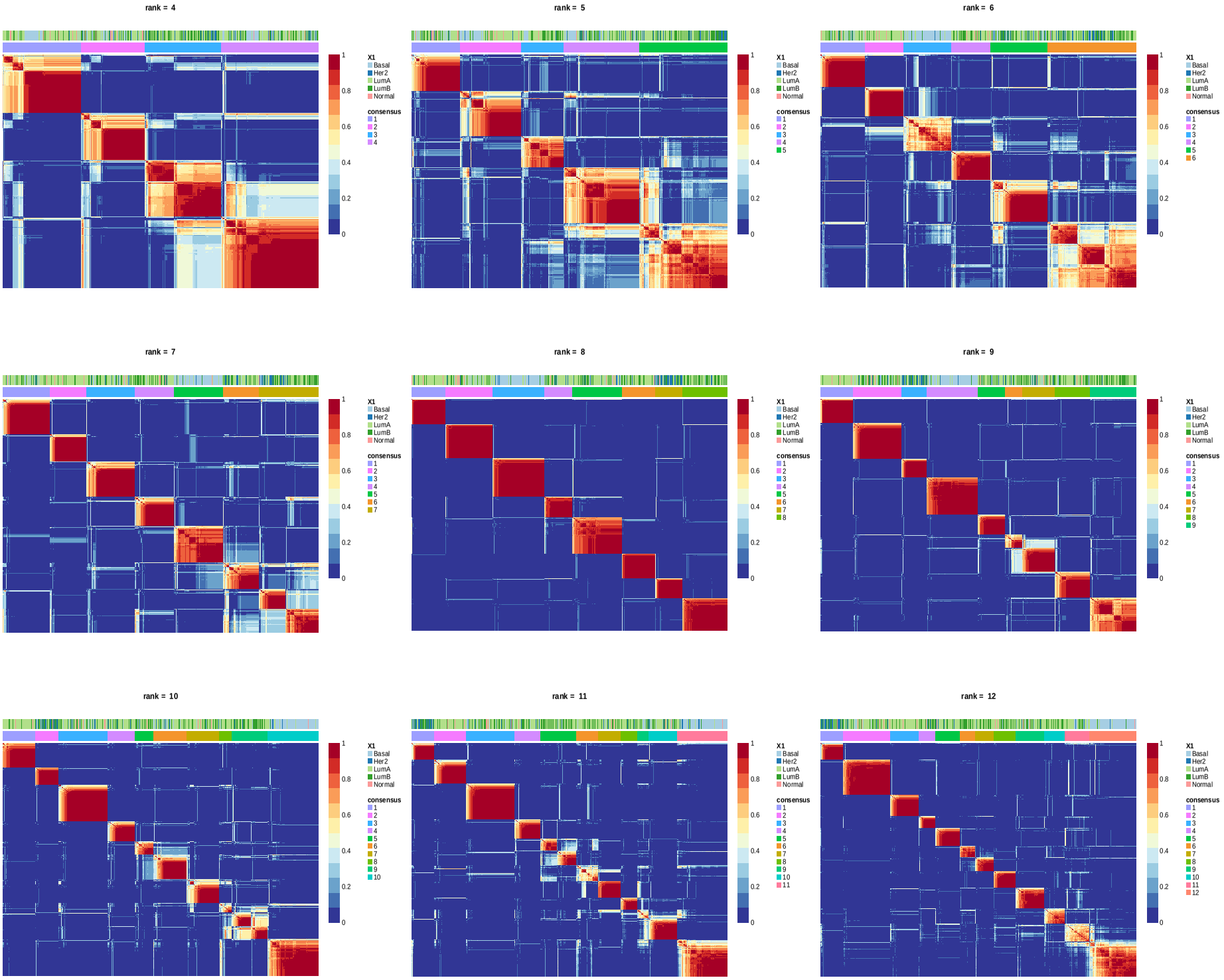


Figure S6: Consensus clustering results of breast cancer RNA-Seq data

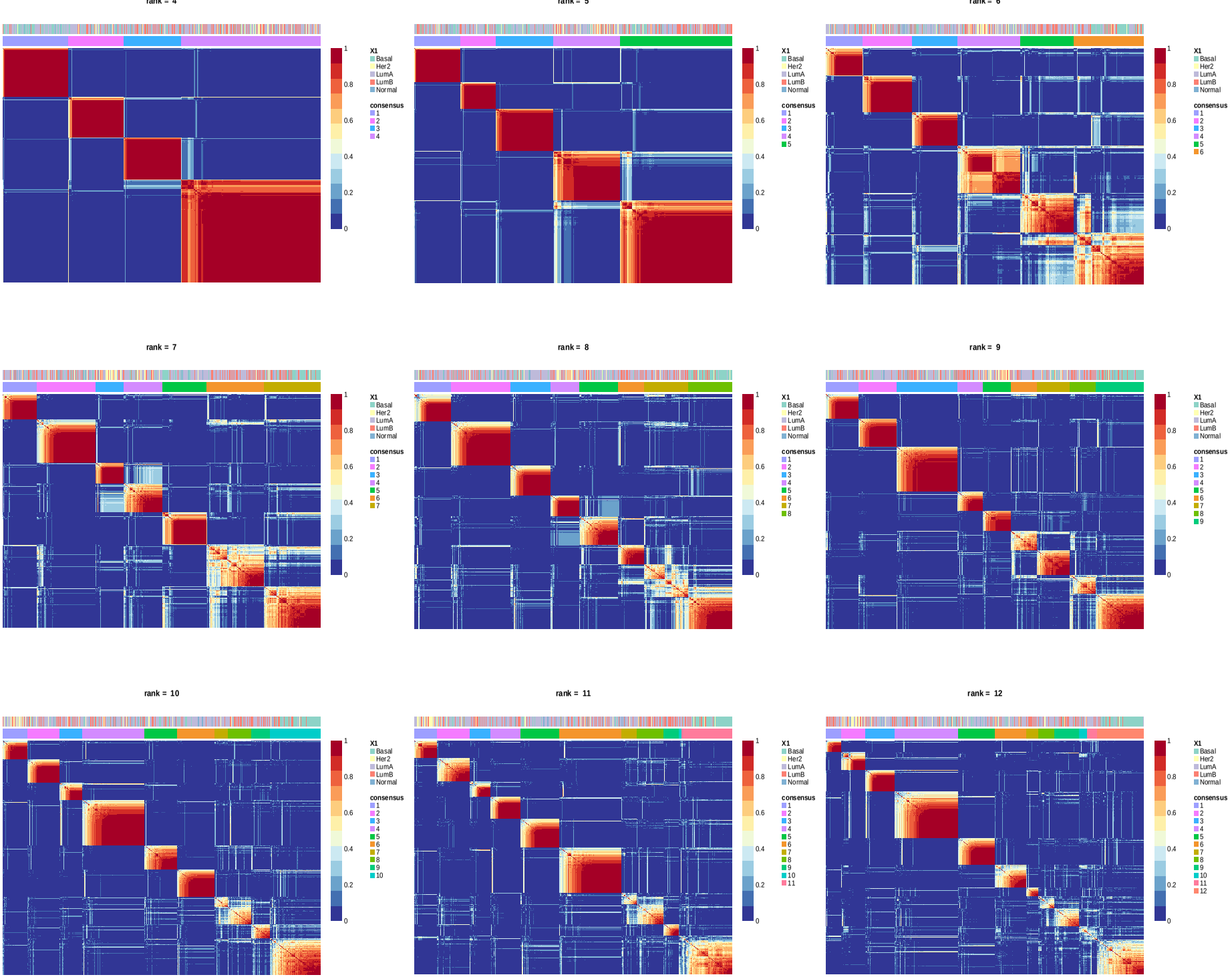


Figure S7: Consensus clustering results of Breast Cancer RNA-Seq data using digraphs mined using Breast Cancer microarray data

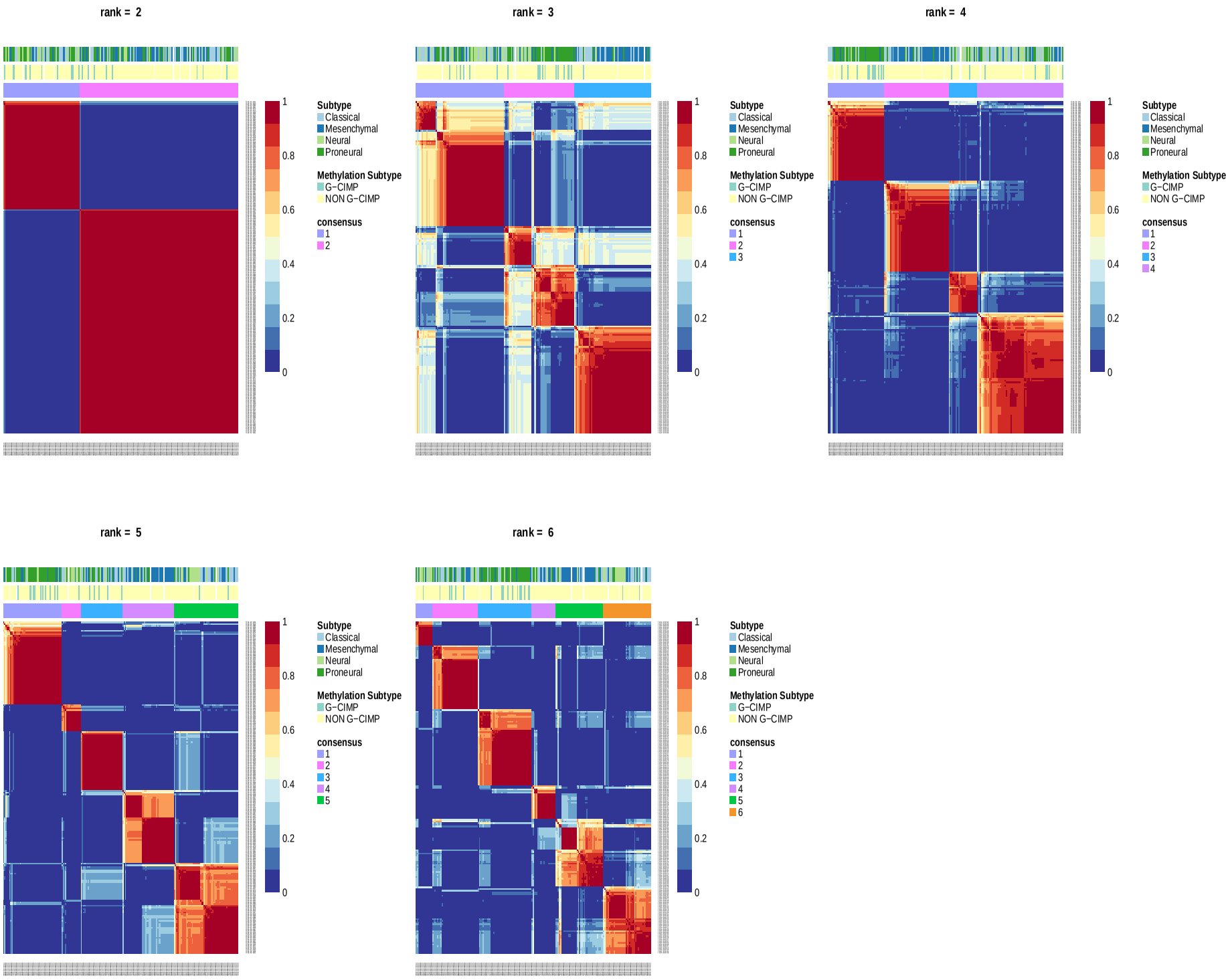


Figure S8: Consensus clustering results of GBM microarray data

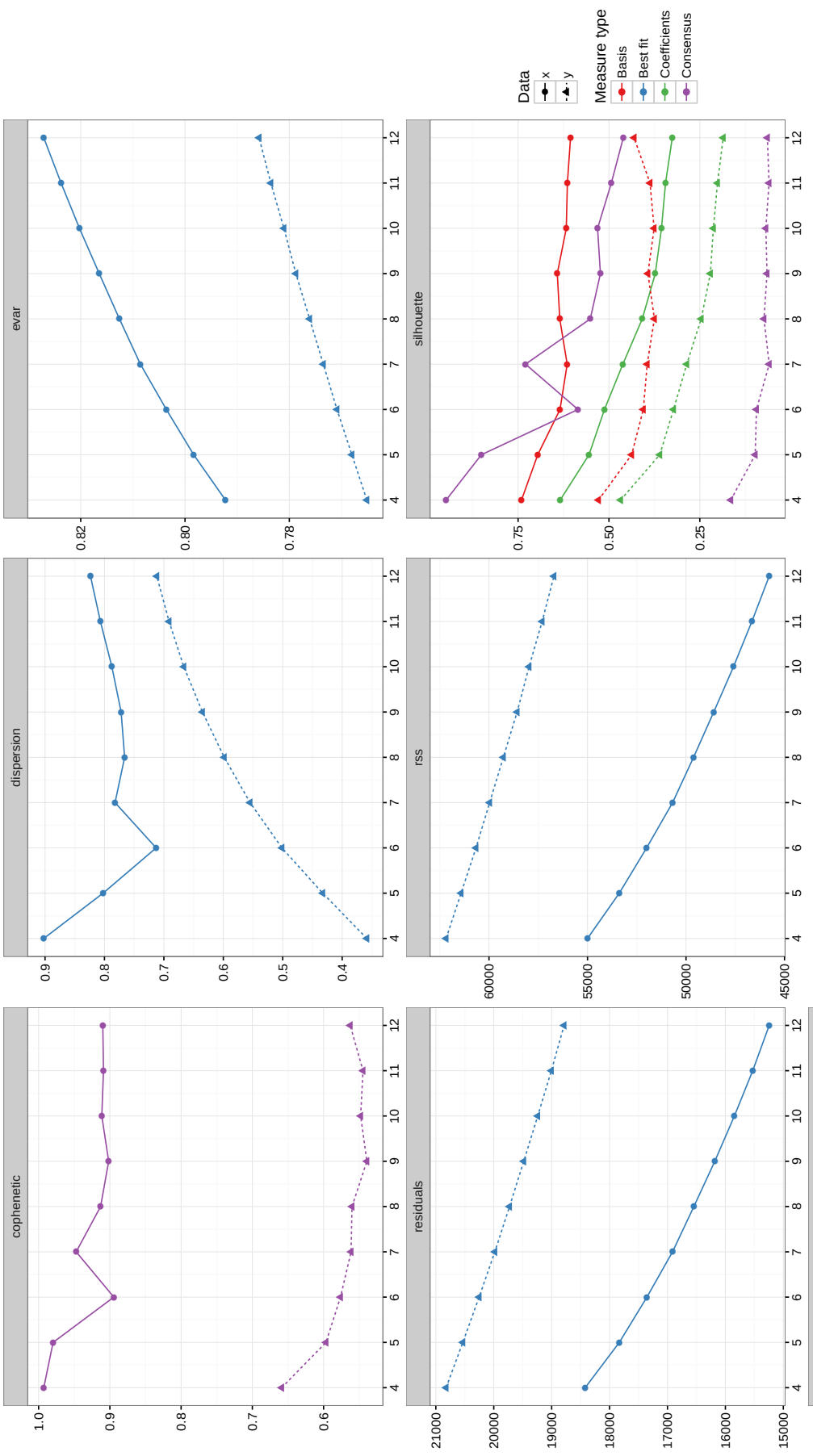


Figure S9: Rank survey of consensus clustering of breast cancer microarray data

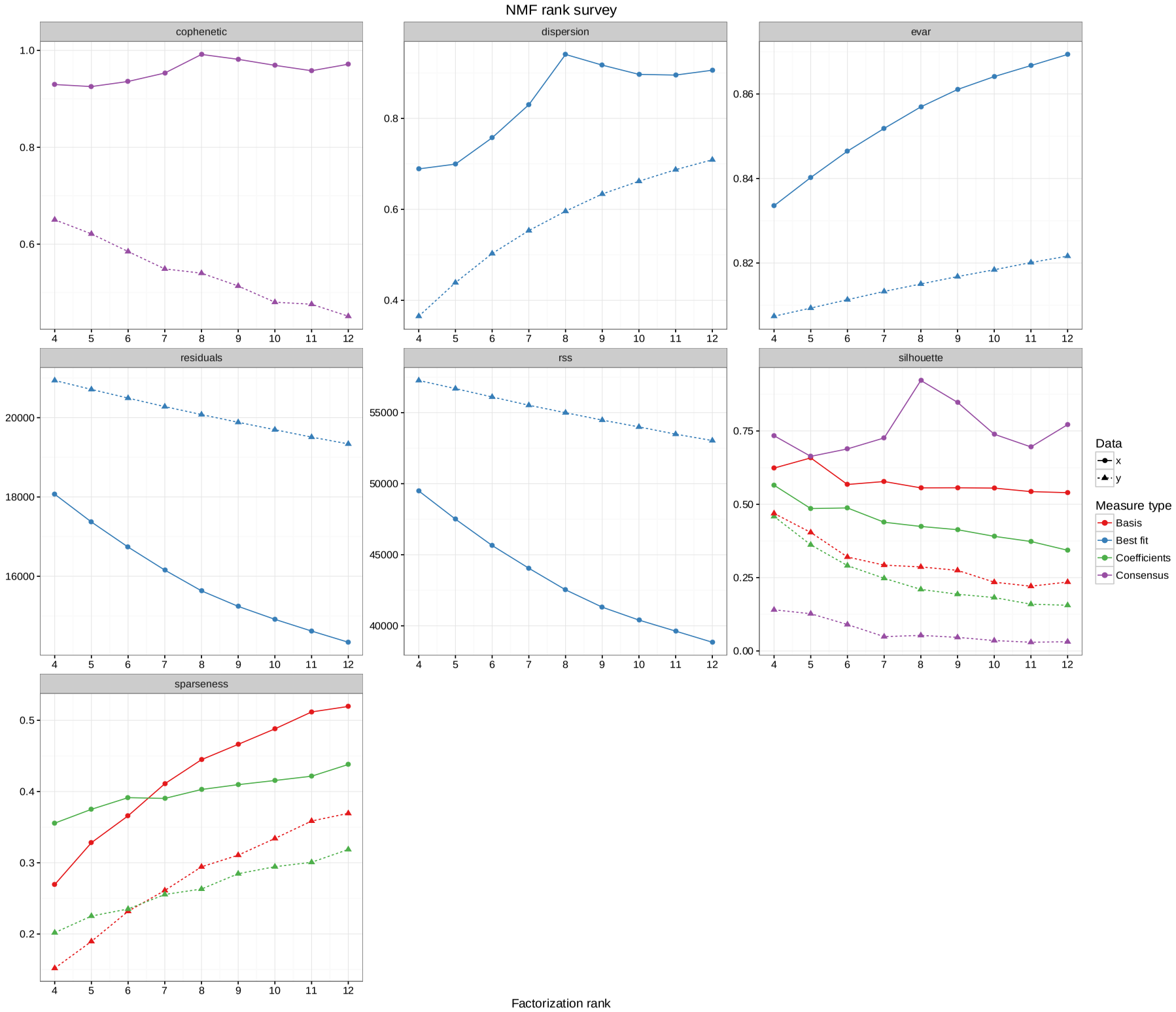


Figure S10: Rank survey of consensus clustering of breast cancer RNA-Seq data

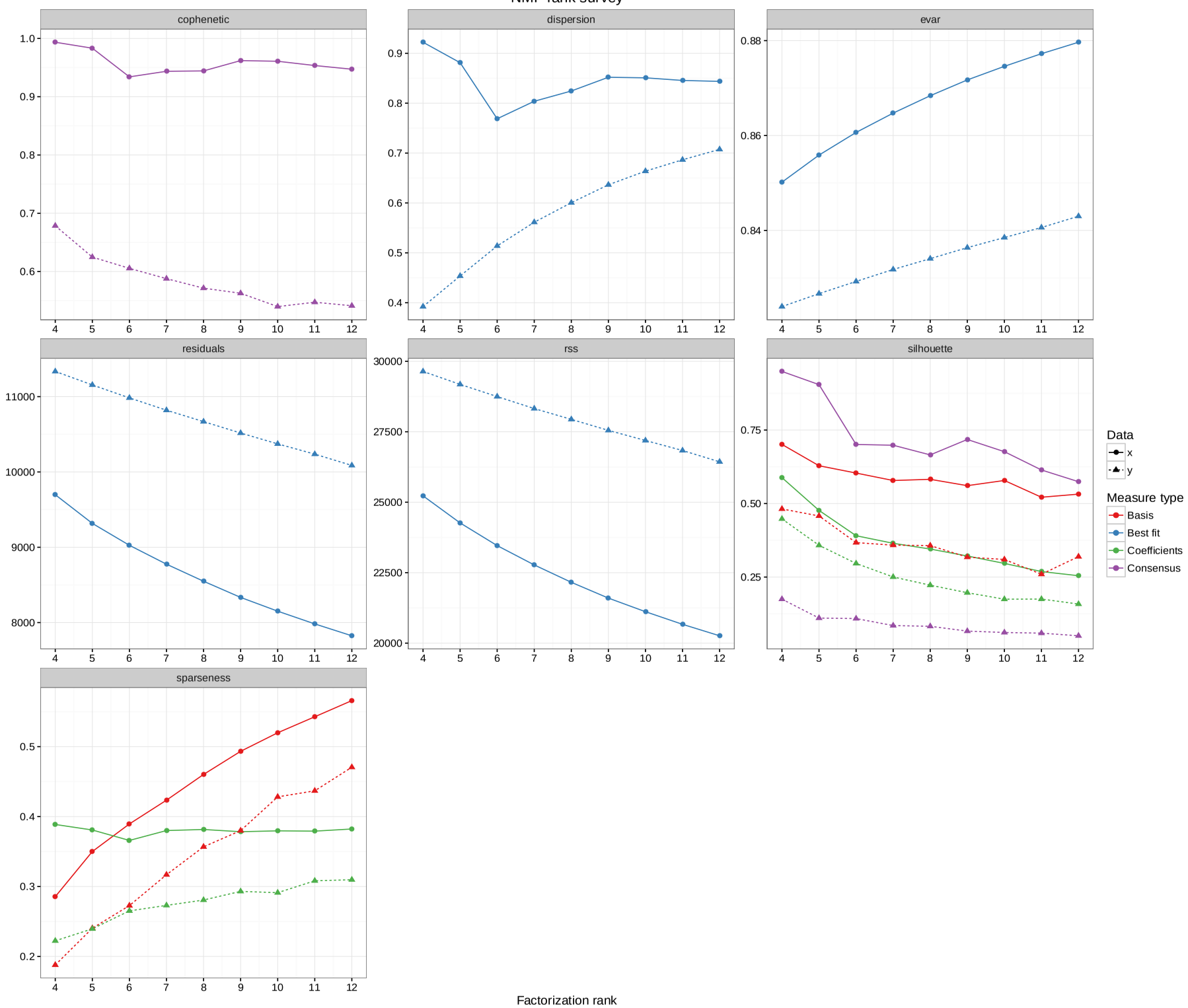


Figure S11: Rank survey of consensus clustering of breast cancer RNA-Seq data using digraphs obtained from mining breast cancer microarray data

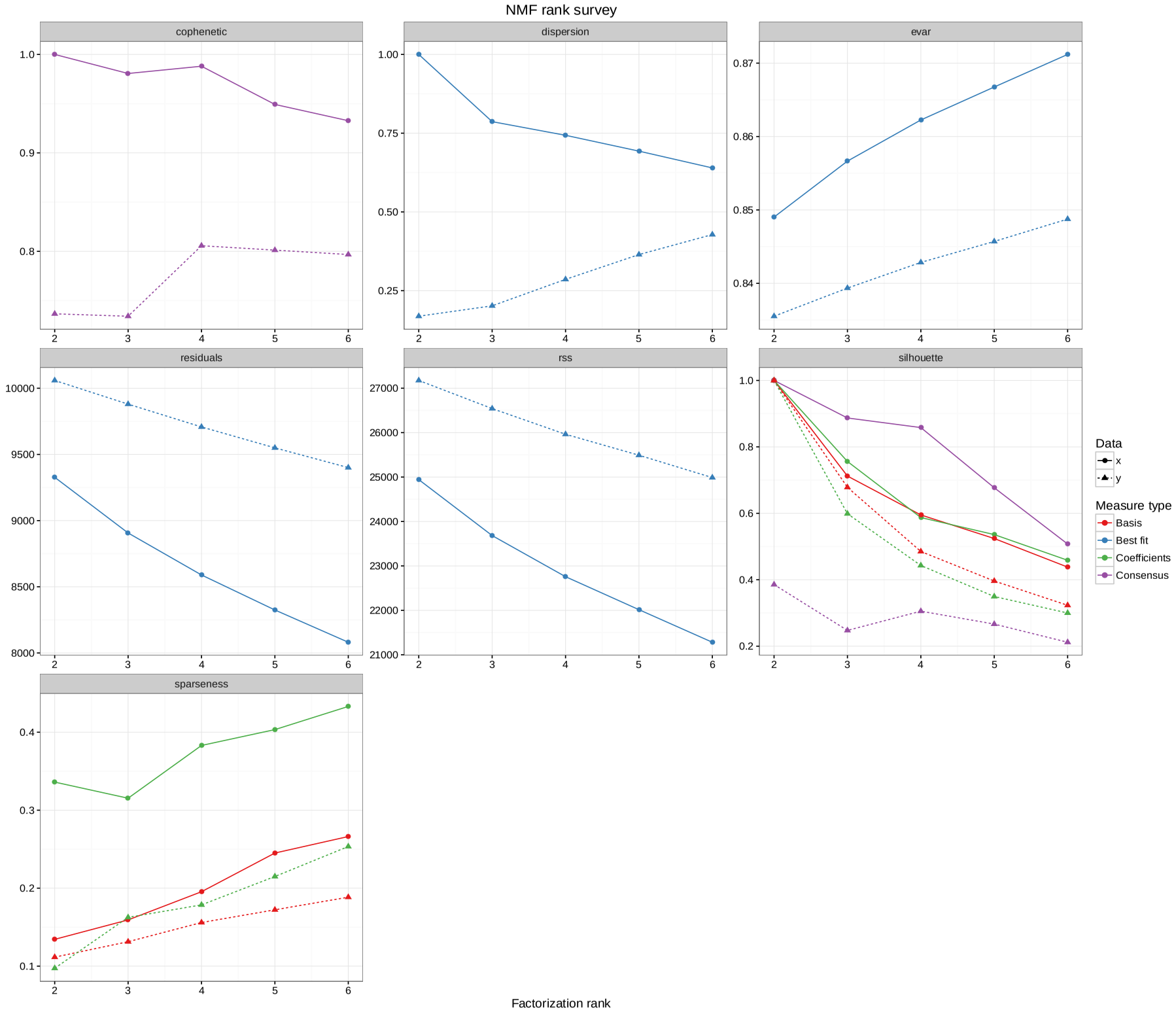


Figure S12: Rank survey of consensus clustering of glioblastoma multiforme RNA-Seq data

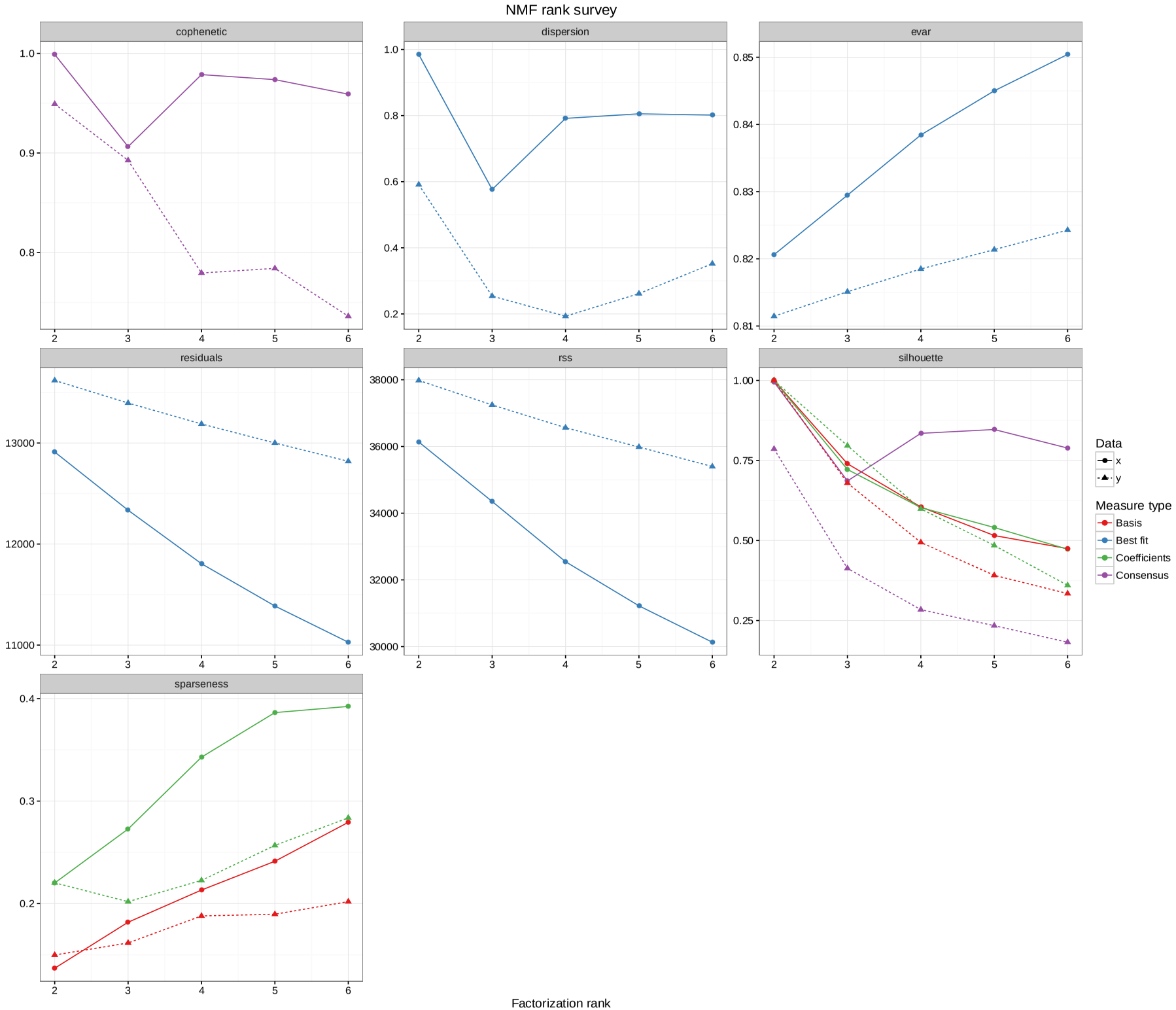
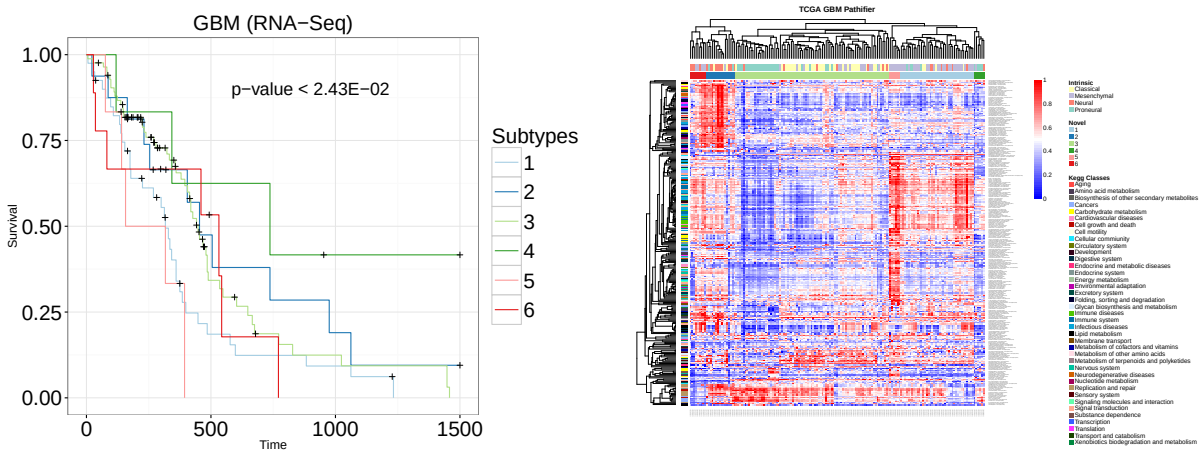
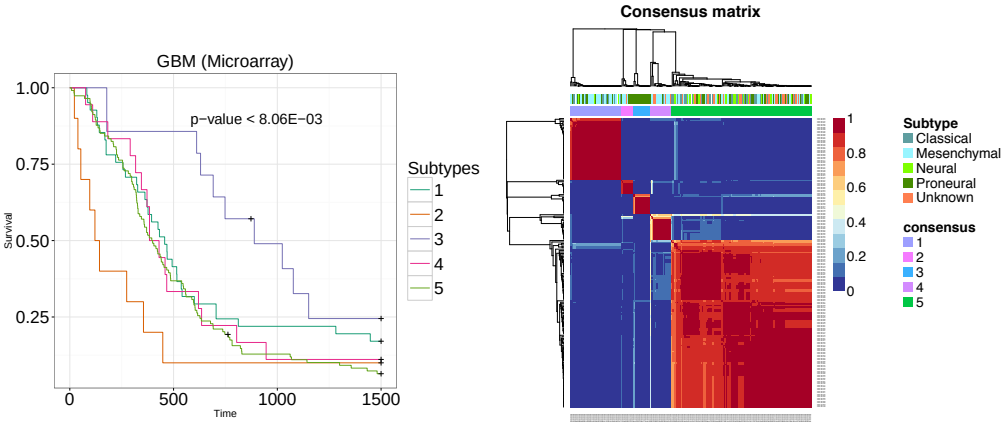


Figure S13: Rank survey of consensus clustering of glioblastoma multiforme microarray data



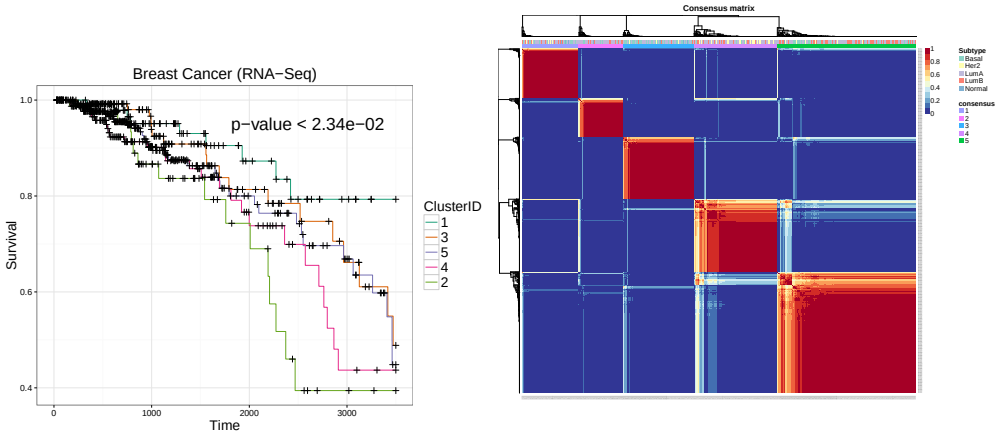
(a) (b)

Figure S14: (a) Kaplan-Meier curves and (b) heatmap of GBM data from TCGA using Pathifier algorithm [7]. Survival curves are significantly different (Log-rank $p - value < 2.43E - 02$). However note that there are 6 biologically irrelevant groups. Figure b shows the heatmap of pathway dysregulation scores (PDS) obtained by the Pathifier algorithm. Rows correspond to Kegg pathways, columns correspond to samples. Top bar shows intrinsic subtypes defined previously. Row annotation shows classes for each KEGG pathway. Furthermore, cophenetic correlation coefficient is 0.61 suggesting non-robust clustering based on the PDS.



(a) (b)

Figure S15: (a) Kaplan-Meier and (b) consensus clustering for GBM data used in NCIS study. Figure a shows highly significant separation of novel groups ($p - value < 8E - 03$). Figure b shows the clustering of patients with top bar representing intrinsic subtypes. Note that the data used in the NCIS study is from Verhaak *et al.* [25], i.e. same samples with different pre-processing and PPI network integration.



(a) (b)

Figure S16: Kaplan-Meier and consensus cluster results for breast cancer rna-seq data using subgraphs obtained from microarray data. The significant survival stratification suggests the functional importance of the networks identified in microarray data.

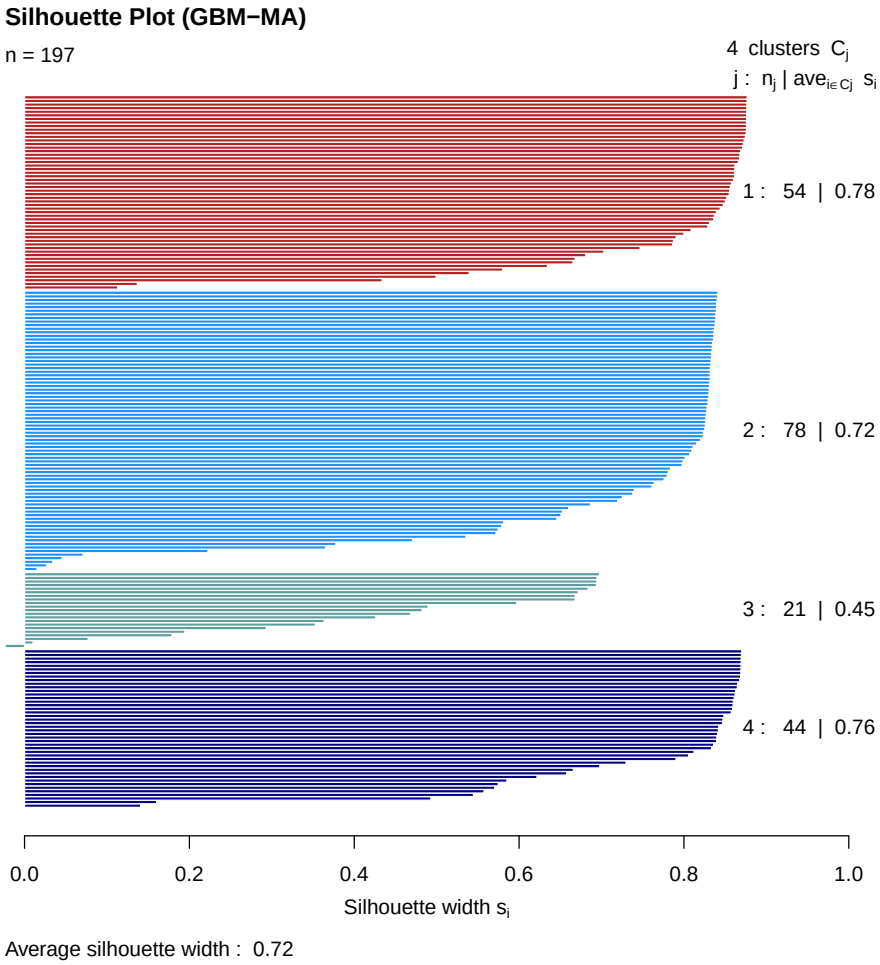


Figure S17: Sihouette plot of GBM microarray data obtained from bootstrap runs. Using different sets of mined subgraphs we show that the clustering of patients is stable which also supports the functional relevance of the dysregulated networks.

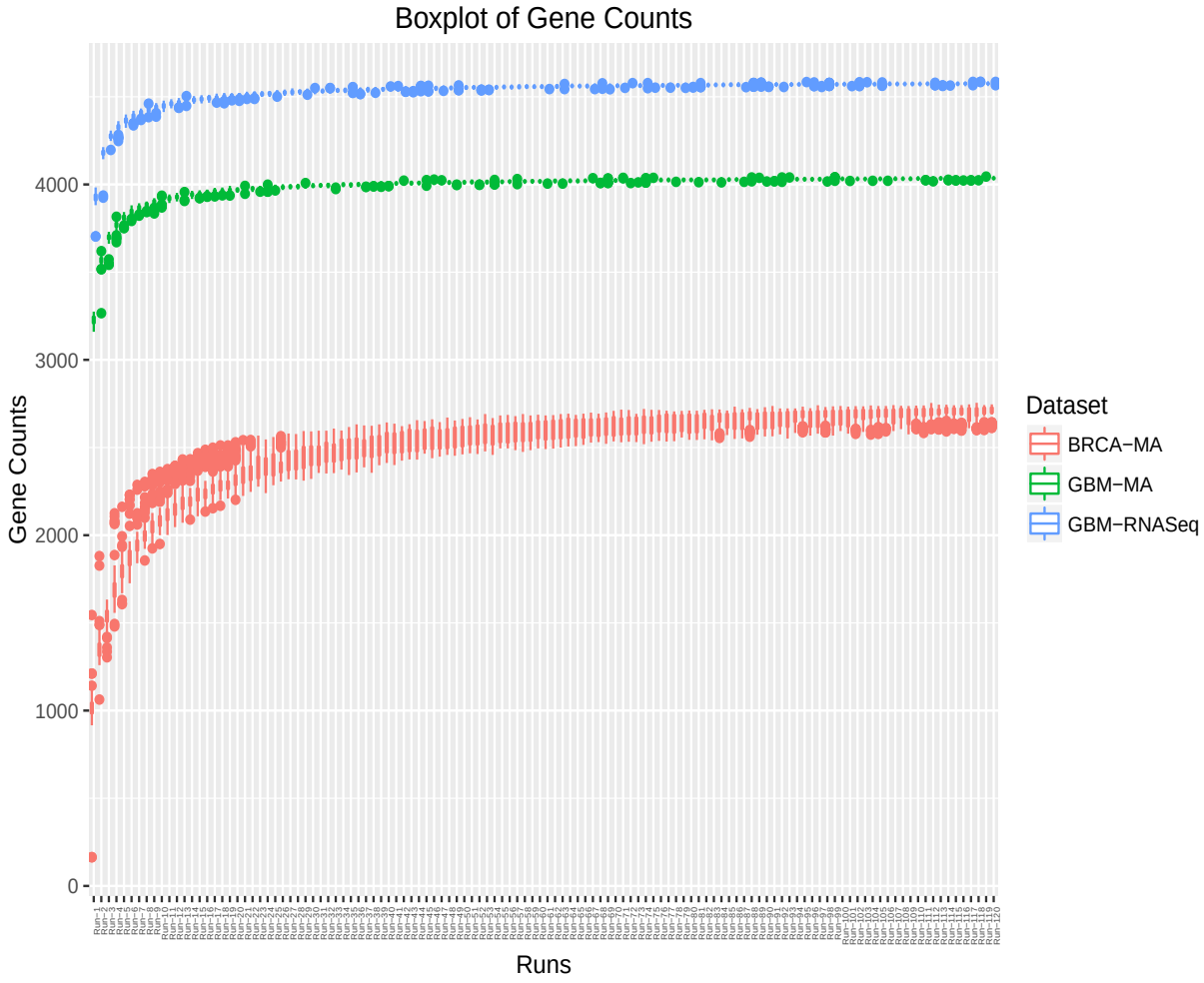


Figure S18: Gene Counts obtained from sampling varying set sizes from multiple qsplor runs. The plot provides a threshold for minimum required runs to obtain maximum informative sets. Y-axis represents the count of unique genes. X-axis represents the set size.

Pathway identifier	Pathway name	Entities FDR
R-HSA-69278	Cell Cycle, Mitotic	4.24e-14
R-HSA-1640170	Cell Cycle	4.24e-14
R-HSA-69242	S Phase	5.38e-13
R-HSA-69206	G1/S Transition	2.93e-12
R-HSA-453279	Mitotic G1-G1/S phases	4.09e-12
R-HSA-176187	Activation of ATR in response to replication stress	4.54e-10
R-HSA-69481	G2/M Checkpoints	1.28e-08
R-HSA-69620	Cell Cycle Checkpoints	2.33e-08
R-HSA-69306	DNA Replication	3.51e-08
R-HSA-69239	Synthesis of DNA	1.18e-07
R-HSA-1442490	Collagen degradation	1.22e-06
R-HSA-69202	Cyclin E associated events during G1/S transition	1.89e-06
R-HSA-69656	Cyclin A:Cdk2-associated events at S phase entry	1.97e-06
R-HSA-1474228	Degradation of the extracellular matrix	1.97e-06
R-HSA-187577	SCF(Skp2)-mediated degradation of p27/p21	4.48e-06
R-HSA-68874	M/G1 Transition	6.49e-06
R-HSA-69002	DNA Replication Pre-Initiation	6.49e-06
R-HSA-453276	Regulation of mitotic cell cycle	6.95e-06
R-HSA-174143	APC/C-mediated degradation of cell cycle proteins	6.95e-06
R-HSA-113510	E2F mediated regulation of DNA replication	6.95e-06
R-HSA-6804756	Regulation of TP53 Activity through Phosphorylation	9.83e-06
R-HSA-69300	Removal of licensing factors from origins	1.32e-05
R-HSA-1474244	Extracellular matrix organization	1.32e-05
R-HSA-204005	COPII (Coat Protein 2) Mediated Vesicle Transport	1.32e-05
R-HSA-69190	DNA strand elongation	1.32e-05
R-HSA-69304	Regulation of DNA replication	1.47e-05
R-HSA-69091	Polymerase switching	3.75e-05
R-HSA-69109	Leading Strand Synthesis	3.75e-05
R-HSA-68867	Assembly of the pre-replicative complex	4.41e-05
R-HSA-174411	Polymerase switching on the C-strand of the telomere	4.50e-05

Table S1: Breast Cancer Microarray Pathway Enrichment Cluster 1

Pathway identifier	Pathway name	Entities FDR
R-HSA-927802	Nonsense-Mediated Decay (NMD)	4.03e-14
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	4.03e-14
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	4.35e-11
R-HSA-156902	Peptide chain elongation	4.35e-11
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	5.50e-11
R-HSA-156842	Eukaryotic Translation Elongation	5.53e-11
R-HSA-72689	Formation of a pool of free 40S subunits	5.87e-11
R-HSA-72764	Eukaryotic Translation Termination	6.27e-11
R-HSA-192823	Viral mRNA Translation	1.19e-10
R-HSA-2408557	Selenocysteine synthesis	1.40e-10
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	1.54e-10
R-HSA-452723	Transcriptional regulation of pluripotent stem cells	1.72e-10
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	1.83e-10
R-HSA-72613	Eukaryotic Translation Initiation	4.32e-10
R-HSA-72737	Cap-dependent Translation Initiation	4.32e-10
R-HSA-1266738	Developmental Biology	1.79e-09
R-HSA-72766	Translation	1.79e-09
R-HSA-168273	Influenza Viral RNA Transcription and Replication	1.13e-08
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	1.51e-08
R-HSA-168255	Influenza Life Cycle	2.56e-08
R-HSA-2408522	Selenoamino acid metabolism	3.45e-08
R-HSA-168254	Influenza Infection	5.23e-08
R-HSA-8868773	rRNA processing in the nucleus and cytosol	6.76e-08
R-HSA-72312	rRNA processing	5.96e-07
R-HSA-74160	Gene Expression	1.18e-06
R-HSA-5368286	Mitochondrial translation initiation	4.43e-06
R-HSA-195721	Signaling by Wnt	4.43e-06
R-HSA-5419276	Mitochondrial translation termination	1.64e-05
R-HSA-5389840	Mitochondrial translation elongation	1.64e-05
R-HSA-5663205	Infectious disease	4.39e-05

Table S2: Breast Cancer Microarray Pathway Enrichment Cluster 2

Pathway identifier	Pathway name	Entities FDR
R-HSA-114604	GPVI-mediated activation cascade	1.34e-14
R-HSA-418594	G alpha (i) signalling events	1.34e-14
R-HSA-168249	Innate Immune System	1.34e-14
R-HSA-168256	Immune System	1.34e-14
R-HSA-1280218	Adaptive Immune System	1.34e-14
R-HSA-388396	GPCR downstream signaling	1.34e-14
R-HSA-162582	Signal Transduction	1.34e-14
R-HSA-420499	Class C/3 (Metabotropic glutamate/pheromone receptors)	1.34e-14
R-HSA-372790	Signaling by GPCR	2.40e-14
R-HSA-500792	GPCR ligand binding	2.50e-12
R-HSA-76002	Platelet activation, signaling and aggregation	1.02e-11
R-HSA-2424491	DAP12 signaling	2.66e-11
R-HSA-2172127	DAP12 interactions	6.66e-11
R-HSA-168898	Toll-Like Receptors Cascades	1.47e-10
R-HSA-983705	Signaling by the B Cell Receptor (BCR)	2.23e-10
R-HSA-166016	Toll Like Receptor 4 (TLR4) Cascade	1.15e-09
R-HSA-109582	Hemostasis	1.38e-09
R-HSA-388841	Costimulation by the CD28 family	2.01e-09
R-HSA-1280215	Cytokine Signaling in Immune system	6.50e-09
R-HSA-177929	Signaling by EGFR	7.85e-09
R-HSA-1433557	Signaling by SCF-KIT	1.41e-08
R-HSA-186763	Downstream signal transduction	2.87e-08
R-HSA-166054	Activated TLR4 signalling	2.87e-08
R-HSA-983695	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	4.69e-08
R-HSA-2454202	Fc epsilon receptor (FCERI) signaling	4.73e-08
R-HSA-5654708	Downstream signaling of activated FGFR3	7.74e-08
R-HSA-5654716	Downstream signaling of activated FGFR4	7.74e-08
R-HSA-5654696	Downstream signaling of activated FGFR2	7.74e-08
R-HSA-186797	Signaling by PDGF	8.17e-08
R-HSA-5654743	Signaling by FGFR4	8.17e-08

Table S3: Breast Cancer Microarray Pathway Enrichment Cluster 3

Pathway identifier	Pathway name	Entities FDR
R-HSA-112315	Transmission across Chemical Synapses	1.30e-13
R-HSA-112316	Neuronal System	6.07e-11
R-HSA-112314	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	2.04e-09
R-HSA-162582	Signal Transduction	4.81e-09
R-HSA-977443	GABA receptor activation	5.39e-08
R-HSA-881907	Gastrin-CREB signalling pathway via PKC and MAPK	7.56e-08
R-HSA-212676	Dopamine Neurotransmitter Release Cycle	9.45e-08
R-HSA-372790	Signaling by GPCR	1.91e-07
R-HSA-1855204	Synthesis of IP3 and IP4 in the cytosol	1.26e-06
R-HSA-422475	Axon guidance	2.00e-06
R-HSA-1266738	Developmental Biology	3.14e-06
R-HSA-166520	Signalling by NGF	3.14e-06
R-HSA-375165	NCAM signaling for neurite out-growth	1.22e-05
R-HSA-5654696	Downstream signaling of activated FGFR2	1.22e-05
R-HSA-5654716	Downstream signaling of activated FGFR4	1.22e-05
R-HSA-5654708	Downstream signaling of activated FGFR3	1.22e-05
R-HSA-5654738	Signaling by FGFR2	1.22e-05
R-HSA-5654687	Downstream signaling of activated FGFR1	1.22e-05
R-HSA-5654743	Signaling by FGFR4	1.22e-05
R-HSA-5654741	Signaling by FGFR3	1.28e-05
R-HSA-186797	Signaling by PDGF	1.28e-05
R-HSA-5654736	Signaling by FGFR1	1.28e-05
R-HSA-190236	Signaling by FGFR	1.28e-05
R-HSA-186763	Downstream signal transduction	1.41e-05
R-HSA-112310	Neurotransmitter Release Cycle	1.44e-05
R-HSA-187037	NGF signalling via TRKA from the plasma membrane	1.44e-05
R-HSA-2424491	DAP12 signaling	1.44e-05
R-HSA-177929	Signaling by EGFR	1.81e-05
R-HSA-977444	GABA B receptor activation	1.81e-05
R-HSA-991365	Activation of GABAB receptors	1.81e-05

Table S4: Breast Cancer Microarray Pathway Enrichment Cluster 4

Pathway identifier	Pathway name	Entities FDR
R-HSA-162582	Signal Transduction	5.03e-10
R-HSA-195721	Signaling by Wnt	5.03e-10
R-HSA-1266738	Developmental Biology	1.06e-07
R-HSA-500792	GPCR ligand binding	1.42e-07
R-HSA-373080	Class B/2 (Secretin family receptors)	4.10e-07
R-HSA-3238698	WNT ligand biogenesis and trafficking	4.10e-07
R-HSA-4086400	PCP/CE pathway	7.91e-07
R-HSA-1638074	Keratan sulfate/keratin metabolism	9.59e-07
R-HSA-3000171	Non-integrin membrane-ECM interactions	2.87e-06
R-HSA-5610787	Hedgehog 'off' state	3.84e-06
R-HSA-3858494	Beta-catenin independent WNT signaling	5.14e-06
R-HSA-983231	Factors involved in megakaryocyte development and platelet production	8.54e-06
R-HSA-372790	Signaling by GPCR	1.03e-05
R-HSA-216083	Integrin cell surface interactions	2.65e-05
R-HSA-109582	Hemostasis	3.41e-05
R-HSA-5358351	Signaling by Hedgehog	3.71e-05
R-HSA-201681	TCF dependent signaling in response to WNT	5.03e-05
R-HSA-381340	Transcriptional regulation of white adipocyte differentiation	7.14e-05
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	1.02e-04
R-HSA-3000178	ECM proteoglycans	1.14e-04
R-HSA-3000170	Syndecan interactions	1.33e-04
R-HSA-1474244	Extracellular matrix organization	2.82e-04
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	2.82e-04
R-HSA-2022854	Keratan sulfate biosynthesis	3.76e-04
R-HSA-4608870	Asymmetric localization of PCP proteins	4.27e-04
R-HSA-1630316	Glycosaminoglycan metabolism	4.27e-04
R-HSA-422475	Axon guidance	4.41e-04
R-HSA-1474290	Collagen formation	4.49e-04
R-HSA-390522	Striated Muscle Contraction	4.51e-04
R-HSA-1442490	Collagen degradation	4.73e-04

Table S5: Breast Cancer Microarray Pathway Enrichment Cluster 5

Pathway identifier	Pathway name	Entities FDR
R-HSA-194840	Rho GTPase cycle	4.39e-12
R-HSA-162582	Signal Transduction	6.99e-09
R-HSA-194315	Signaling by Rho GTPases	7.10e-08
R-HSA-194441	Metabolism of non-coding RNA	1.20e-07
R-HSA-191859	snRNP Assembly	1.20e-07
R-HSA-74752	Signaling by Insulin receptor	1.21e-07
R-HSA-1433557	Signaling by SCF-KIT	1.47e-06
R-HSA-5683057	MAPK family signaling cascades	1.05e-05
R-HSA-186763	Downstream signal transduction	1.14e-05
R-HSA-2424491	DAP12 signaling	1.14e-05
R-HSA-372790	Signaling by GPCR	1.14e-05
R-HSA-177929	Signaling by EGFR	1.14e-05
R-HSA-451927	Interleukin-2 signaling	1.14e-05
R-HSA-2172127	DAP12 interactions	1.14e-05
R-HSA-186797	Signaling by PDGF	1.14e-05
R-HSA-4420097	VEGFA-VEGFR2 Pathway	1.14e-05
R-HSA-5654716	Downstream signaling of activated FGFR4	1.14e-05
R-HSA-5654708	Downstream signaling of activated FGFR3	1.14e-05
R-HSA-5654696	Downstream signaling of activated FGFR2	1.14e-05
R-HSA-194138	Signaling by VEGF	1.14e-05
R-HSA-5654743	Signaling by FGFR4	1.14e-05
R-HSA-5654687	Downstream signaling of activated FGFR1	1.14e-05
R-HSA-5654741	Signaling by FGFR3	1.14e-05
R-HSA-1250347	SHC1 events in ERBB4 signaling	1.14e-05
R-HSA-179812	GRB2 events in EGFR signaling	1.14e-05
R-HSA-112412	SOS-mediated signalling	1.14e-05
R-HSA-5673001	RAF/MAP kinase cascade	1.14e-05
R-HSA-180336	SHC1 events in EGFR signaling	1.14e-05
R-HSA-5654736	Signaling by FGFR1	1.14e-05
R-HSA-5654712	FRS-mediated FGFR4 signaling	1.14e-05

Table S6: Breast Cancer RNA-Seq Pathway Enrichment Cluster 1 (Microarray Subgraphs)

Pathway identifier	Pathway name	Entities FDR
R-HSA-1638074	Keratan sulfate/keratin metabolism	4.02e-06
R-HSA-194315	Signaling by Rho GTPases	1.20e-04
R-HSA-114604	GPVI-mediated activation cascade	1.20e-04
R-HSA-162582	Signal Transduction	1.83e-04
R-HSA-69278	Cell Cycle, Mitotic	1.97e-04
R-HSA-3238698	WNT ligand biogenesis and trafficking	2.36e-04
R-HSA-76002	Platelet activation, signaling and aggregation	2.36e-04
R-HSA-195258	RHO GTPase Effectors	2.36e-04
R-HSA-1640170	Cell Cycle	2.89e-04
R-HSA-373760	L1CAM interactions	2.89e-04
R-HSA-2022854	Keratan sulfate biosynthesis	5.21e-04
R-HSA-176187	Activation of ATR in response to replication stress	5.60e-04
R-HSA-1855204	Synthesis of IP3 and IP4 in the cytosol	5.60e-04
R-HSA-109582	Hemostasis	9.66e-04
R-HSA-2022857	Keratan sulfate degradation	9.86e-04
R-HSA-422475	Axon guidance	1.21e-03
R-HSA-195721	Signaling by Wnt	1.21e-03
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	1.70e-03
R-HSA-389957	Prefoldin mediated transfer of substrate to CCT/TriC	2.00e-03
R-HSA-1630316	Glycosaminoglycan metabolism	2.12e-03
R-HSA-68877	Mitotic Prometaphase	2.17e-03
R-HSA-380287	Centrosome maturation	2.42e-03
R-HSA-380270	Recruitment of mitotic centrosome proteins and complexes	2.42e-03
R-HSA-380320	Recruitment of NuMA to mitotic centrosomes	2.93e-03
R-HSA-68962	Activation of the pre-replicative complex	3.23e-03
R-HSA-389958	Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding	3.43e-03
R-HSA-881907	Gastrin-CREB signalling pathway via PKC and MAPK	7.22e-03
R-HSA-1483249	Inositol phosphate metabolism	8.23e-03
R-HSA-416476	G alpha (q) signalling events	8.32e-03
R-HSA-1266738	Developmental Biology	8.41e-03

Table S7: Breast Cancer RNA-Seq Pathway Enrichment Cluster 2 (Microarray Subgraphs)

Pathway identifier	Pathway name	Entities FDR
R-HSA-1592389	Activation of Matrix Metalloproteinases	1.01e-14
R-HSA-1474228	Degradation of the extracellular matrix	1.01e-14
R-HSA-1442490	Collagen degradation	1.01e-14
R-HSA-1474244	Extracellular matrix organization	1.01e-14
R-HSA-420499	Class C/3 (Metabotropic glutamate/pheromone receptors)	3.20e-14
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	2.21e-12
R-HSA-500792	GPCR ligand binding	2.85e-12
R-HSA-1474290	Collagen formation	3.78e-11
R-HSA-418594	G alpha (i) signalling events	1.03e-09
R-HSA-388396	GPCR downstream signaling	5.88e-07
R-HSA-5368286	Mitochondrial translation initiation	1.48e-06
R-HSA-5419276	Mitochondrial translation termination	5.28e-06
R-HSA-5389840	Mitochondrial translation elongation	5.28e-06
R-HSA-372790	Signaling by GPCR	9.53e-06
R-HSA-3000171	Non-integrin membrane-ECM interactions	9.53e-06
R-HSA-418555	G alpha (s) signalling events	1.04e-05
R-HSA-5368287	Mitochondrial translation	2.42e-05
R-HSA-446107	Type I hemidesmosome assembly	3.14e-05
R-HSA-3000170	Syndecan interactions	5.27e-05
R-HSA-5601884	PIWI-interacting RNA (piRNA) biogenesis	9.32e-05
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	3.77e-04
R-HSA-381426	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by IGFBPs	3.77e-04
R-HSA-3000178	ECM proteoglycans	3.96e-04
R-HSA-216083	Integrin cell surface interactions	7.23e-04
R-HSA-399719	Trafficking of AMPA receptors	2.43e-03
R-HSA-2214320	Anchoring fibril formation	2.75e-03
R-HSA-399721	Glutamate Binding, Activation of AMPA Receptors and Synaptic Plasticity	2.75e-03
R-HSA-2142688	Synthesis of 5-eicosatetraenoic acids	5.17e-03
R-HSA-2142700	Synthesis of Lipoxins (LX)	7.29e-03
R-HSA-162582	Signal Transduction	1.57e-02

Table S8: Breast Cancer RNA-Seq Pathway Enrichment Cluster 3 (Microarray Subgraphs)

Pathway identifier	Pathway name	Entities FDR
R-HSA-5610787	Hedgehog 'off' state	1.92e-09
R-HSA-187577	SCF(Skp2)-mediated degradation of p27/p21	4.86e-09
R-HSA-69202	Cyclin E associated events during G1/S transition	1.33e-08
R-HSA-5358351	Signaling by Hedgehog	1.33e-08
R-HSA-69656	Cyclin A:Cdk2-associated events at S phase entry	1.33e-08
R-HSA-5610783	Degradation of GLI2 by the proteasome	5.65e-08
R-HSA-204005	COPII (Coat Protein 2) Mediated Vesicle Transport	2.54e-07
R-HSA-174113	SCF-beta-TrCP mediated degradation of Emi1	3.75e-07
R-HSA-69206	G1/S Transition	6.06e-07
R-HSA-5676590	NIK–>noncanonical NF-kB signaling	6.06e-07
R-HSA-5610780	Degradation of GLI1 by the proteasome	6.06e-07
R-HSA-5610785	GLI3 is processed to GLI3R by the proteasome	6.06e-07
R-HSA-453276	Regulation of mitotic cell cycle	6.06e-07
R-HSA-174143	APC/C-mediated degradation of cell cycle proteins	6.06e-07
R-HSA-202424	Downstream TCR signaling	6.31e-07
R-HSA-5607761	Dectin-1 mediated noncanonical NF-kB signaling	8.49e-07
R-HSA-1474244	Extracellular matrix organization	1.27e-06
R-HSA-1169091	Activation of NF-kappaB in B cells	1.48e-06
R-HSA-69242	S Phase	1.81e-06
R-HSA-453279	Mitotic G1-G1/S phases	1.81e-06
R-HSA-202403	TCR signaling	2.36e-06
R-HSA-180534	Vpu mediated degradation of CD4	2.92e-06
R-HSA-176408	Regulation of APC/C activators between G1/S and early anaphase	3.13e-06
R-HSA-195253	Degradation of beta-catenin by the destruction complex	5.64e-06
R-HSA-5668541	TNFR2 non-canonical NF-kB pathway	1.31e-05
R-HSA-69231	Cyclin D associated events in G1	1.32e-05
R-HSA-69236	G1 Phase	1.32e-05
R-HSA-174178	APC/C:Cdh1 mediated degradation of Cdc20 and other targeting proteins in late mitosis/early G1	1.75e-05
R-HSA-199977	ER to Golgi Anterograde Transport	3.22e-05
R-HSA-211733	Regulation of activated PAK-2p34 by proteasome mediated degradation	3.22e-05

Table S9: Breast Cancer RNA-Seq Pathway Enrichment Cluster 4 (Microarray Subgraphs)

Pathway identifier	Pathway name	Entities FDR
R-HSA-4086400	PCP/CE pathway	1.00e-07
R-HSA-3858494	Beta-catenin independent WNT signaling	1.23e-06
R-HSA-69242	S Phase	1.53e-06
R-HSA-69278	Cell Cycle, Mitotic	1.53e-06
R-HSA-69620	Cell Cycle Checkpoints	2.63e-06
R-HSA-174178	APC/C:Cdh1 mediated degradation of Cdc20 and other targeting proteins in late mitosis/early G1	3.04e-06
R-HSA-195721	Signaling by Wnt	4.29e-06
R-HSA-1640170	Cell Cycle	4.29e-06
R-HSA-69239	Synthesis of DNA	6.07e-06
R-HSA-174143	APC/C-mediated degradation of cell cycle proteins	9.31e-06
R-HSA-453276	Regulation of mitotic cell cycle	9.31e-06
R-HSA-69306	DNA Replication	9.31e-06
R-HSA-69481	G2/M Checkpoints	9.31e-06
R-HSA-6804756	Regulation of TP53 Activity through Phosphorylation	1.14e-05
R-HSA-69109	Leading Strand Synthesis	1.65e-05
R-HSA-69091	Polymerase switching	1.65e-05
R-HSA-174411	Polymerase switching on the C-strand of the telomere	2.12e-05
R-HSA-110312	Translesion synthesis by REV1	3.38e-05
R-HSA-4641257	Degradation of AXIN	3.38e-05
R-HSA-5656121	Translesion synthesis by POLI	4.10e-05
R-HSA-5655862	Translesion synthesis by POLK	5.06e-05
R-HSA-5610780	Degradation of GLI1 by the proteasome	5.08e-05
R-HSA-5632684	Hedgehog 'on' state	5.08e-05
R-HSA-69275	G2/M Transition	5.36e-05
R-HSA-453274	Mitotic G2-G2/M phases	5.71e-05
R-HSA-156711	Polo-like kinase mediated events	6.51e-05
R-HSA-4608870	Asymmetric localization of PCP proteins	6.51e-05
R-HSA-110320	Translesion Synthesis by POLH	7.45e-05
R-HSA-69615	G1/S DNA Damage Checkpoints	9.21e-05
R-HSA-69186	Lagging Strand Synthesis	1.01e-04

Table S10: Breast Cancer RNA-Seq Pathway Enrichment Cluster 5 (Microarray Subgraphs)

Pathway identifier	Pathway name	Entities FDR
R-HSA-202430	Translocation of ZAP-70 to Immunological synapse	3.66e-15
R-HSA-202433	Generation of second messenger molecules	3.66e-15
R-HSA-983170	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	3.66e-15
R-HSA-2132295	MHC class II antigen presentation	3.66e-15
R-HSA-909733	Interferon alpha/beta signaling	3.66e-15
R-HSA-202403	TCR signaling	3.66e-15
R-HSA-202424	Downstream TCR signaling	3.66e-15
R-HSA-1236974	ER-Phagosome pathway	3.66e-15
R-HSA-1236977	Endosomal/Vacuolar pathway	3.66e-15
R-HSA-389948	PD-1 signaling	3.66e-15
R-HSA-202427	Phosphorylation of CD3 and TCR zeta chains	3.66e-15
R-HSA-388841	Costimulation by the CD28 family	3.66e-15
R-HSA-1280218	Adaptive Immune System	3.66e-15
R-HSA-983169	Class I MHC mediated antigen processing & presentation	3.66e-15
R-HSA-1236975	Antigen processing-Cross presentation	3.66e-15
R-HSA-168256	Immune System	3.66e-15
R-HSA-913531	Interferon Signaling	3.66e-15
R-HSA-198933	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	3.66e-15
R-HSA-1280215	Cytokine Signaling in Immune system	3.66e-15
R-HSA-877300	Interferon gamma signaling	3.66e-15
R-HSA-114604	GPVI-mediated activation cascade	6.09e-13
R-HSA-168249	Innate Immune System	1.94e-11
R-HSA-375276	Peptide ligand-binding receptors	5.33e-07
R-HSA-380108	Chemokine receptors bind chemokines	5.95e-07
R-HSA-76002	Platelet activation, signaling and aggregation	1.03e-06
R-HSA-194840	Rho GTPase cycle	1.11e-06
R-HSA-373076	Class A/1 (Rhodopsin-like receptors)	2.10e-06
R-HSA-392451	G beta:gamma signalling through PI3Kgamma	2.30e-06
R-HSA-418594	G alpha (i) signalling events	2.69e-06
R-HSA-397795	G-protein beta:gamma signalling	5.07e-06

Table S11: Breast Cancer RNA-Seq Pathway Enrichment Cluster 1

Pathway identifier	Pathway name	Entities FDR
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	8.10e-15
R-HSA-1474290	Collagen formation	8.10e-15
R-HSA-3000178	ECM proteoglycans	8.10e-15
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	8.10e-15
R-HSA-1474228	Degradation of the extracellular matrix	8.10e-15
R-HSA-1442490	Collagen degradation	8.10e-15
R-HSA-1474244	Extracellular matrix organization	8.10e-15
R-HSA-3000171	Non-integrin membrane-ECM interactions	8.10e-15
R-HSA-216083	Integrin cell surface interactions	8.10e-15
R-HSA-162582	Signal Transduction	4.57e-11
R-HSA-3000157	Laminin interactions	6.59e-10
R-HSA-419037	NCAM1 interactions	8.14e-10
R-HSA-422475	Axon guidance	3.11e-09
R-HSA-2129379	Molecules associated with elastic fibres	3.67e-09
R-HSA-1630316	Glycosaminoglycan metabolism	8.88e-09
R-HSA-186797	Signaling by PDGF	1.67e-08
R-HSA-1566948	Elastic fibre formation	1.89e-08
R-HSA-3560782	Diseases associated with glycosaminoglycan metabolism	5.55e-07
R-HSA-3781865	Diseases of glycosylation	7.37e-07
R-HSA-418555	G alpha (s) signalling events	1.01e-06
R-HSA-1793185	Chondroitin sulfate/dermatan sulfate metabolism	1.15e-06
R-HSA-2022870	Chondroitin sulfate biosynthesis	1.15e-06
R-HSA-2214320	Anchoring fibril formation	1.30e-06
R-HSA-375165	NCAM signaling for neurite out-growth	1.39e-06
R-HSA-5173105	O-linked glycosylation	2.26e-06
R-HSA-3000170	Syndecan interactions	2.59e-06
R-HSA-3560801	Defective B3GAT3 causes JDSSDHD	7.22e-06
R-HSA-4420332	Defective B3GALT6 causes EDSP2 and SEMDJL1	7.22e-06
R-HSA-3560783	Defective B4GALT7 causes EDS, progeroid type	7.22e-06
R-HSA-1266738	Developmental Biology	9.91e-06

Table S12: Breast Cancer RNA-Seq Pathway Enrichment Cluster 2

Pathway identifier	Pathway name	Entities FDR
R-HSA-68877	Mitotic Prometaphase	4.22e-15
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	4.22e-15
R-HSA-2467813	Separation of Sister Chromatids	4.22e-15
R-HSA-68874	M/G1 Transition	4.22e-15
R-HSA-69002	DNA Replication Pre-Initiation	4.22e-15
R-HSA-69278	Cell Cycle, Mitotic	4.22e-15
R-HSA-69306	DNA Replication	4.22e-15
R-HSA-2555396	Mitotic Metaphase and Anaphase	4.22e-15
R-HSA-453279	Mitotic G1-G1/S phases	4.22e-15
R-HSA-68882	Mitotic Anaphase	4.22e-15
R-HSA-1640170	Cell Cycle	4.22e-15
R-HSA-68886	M Phase	4.22e-15
R-HSA-69206	G1/S Transition	4.22e-15
R-HSA-69481	G2/M Checkpoints	4.22e-15
R-HSA-69620	Cell Cycle Checkpoints	4.22e-15
R-HSA-5663220	RHO GTPases Activate Formins	4.22e-15
R-HSA-195258	RHO GTPase Effectors	4.22e-15
R-HSA-176187	Activation of ATR in response to replication stress	4.11e-14
R-HSA-69242	S Phase	1.87e-13
R-HSA-156711	Polo-like kinase mediated events	3.33e-13
R-HSA-194315	Signaling by Rho GTPases	2.57e-12
R-HSA-453274	Mitotic G2-G2/M phases	2.79e-12
R-HSA-73894	DNA Repair	3.34e-12
R-HSA-68962	Activation of the pre-replicative complex	1.61e-11
R-HSA-69239	Synthesis of DNA	1.74e-11
R-HSA-174143	APC/C-mediated degradation of cell cycle proteins	2.18e-11
R-HSA-453276	Regulation of mitotic cell cycle	2.18e-11
R-HSA-69304	Regulation of DNA replication	7.58e-11
R-HSA-69275	G2/M Transition	2.56e-10
R-HSA-68867	Assembly of the pre-replicative complex	3.53e-10

Table S13: Breast Cancer RNA-Seq Pathway Enrichment Cluster 3

Pathway identifier	Pathway name	Entities FDR
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	2.11e-15
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	2.11e-15
R-HSA-72689	Formation of a pool of free 40S subunits	2.11e-15
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	2.11e-15
R-HSA-927802	Nonsense-Mediated Decay (NMD)	2.11e-15
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	2.11e-15
R-HSA-192823	Viral mRNA Translation	2.11e-15
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	2.11e-15
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	2.11e-15
R-HSA-6799198	Complex I biogenesis	2.11e-15
R-HSA-72613	Eukaryotic Translation Initiation	2.11e-15
R-HSA-156902	Peptide chain elongation	2.11e-15
R-HSA-72764	Eukaryotic Translation Termination	2.11e-15
R-HSA-72766	Translation	2.11e-15
R-HSA-72737	Cap-dependent Translation Initiation	2.11e-15
R-HSA-611105	Respiratory electron transport	2.11e-15
R-HSA-156842	Eukaryotic Translation Elongation	2.11e-15
R-HSA-163200	Respiratory electron transport, ATP synthesis by chemiosmotic coupling.	2.11e-15
R-HSA-8868773	rRNA processing in the nucleus and cytosol	2.11e-15
R-HSA-168273	Influenza Viral RNA Transcription and Replication	2.11e-15
R-HSA-2408557	Selenocysteine synthesis	2.11e-15
R-HSA-1428517	The citric acid (TCA) cycle and respiratory electron transport	2.11e-15
R-HSA-74160	Gene Expression	2.11e-15
R-HSA-5663205	Infectious disease	2.11e-15
R-HSA-168254	Influenza Infection	2.11e-15
R-HSA-168255	Influenza Life Cycle	2.11e-15
R-HSA-2408522	Selenoamino acid metabolism	2.11e-15
R-HSA-162599	Late Phase of HIV Life Cycle	6.00e-15
R-HSA-72312	rRNA processing	2.40e-14
R-HSA-162587	HIV Life Cycle	2.45e-14

Table S14: Breast Cancer RNA-Seq Pathway Enrichment Cluster 4

Pathway identifier	Pathway name	Entities FDR
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	4.33e-15
R-HSA-156902	Peptide chain elongation	4.33e-15
R-HSA-927802	Nonsense-Mediated Decay (NMD)	4.33e-15
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	4.33e-15
R-HSA-72689	Formation of a pool of free 40S subunits	4.33e-15
R-HSA-192823	Viral mRNA Translation	4.33e-15
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	4.33e-15
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	4.33e-15
R-HSA-72764	Eukaryotic Translation Termination	4.33e-15
R-HSA-72766	Translation	4.33e-15
R-HSA-72737	Cap-dependent Translation Initiation	4.33e-15
R-HSA-72613	Eukaryotic Translation Initiation	4.33e-15
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	4.33e-15
R-HSA-156842	Eukaryotic Translation Elongation	4.33e-15
R-HSA-2408557	Selenocysteine synthesis	4.33e-15
R-HSA-74160	Gene Expression	4.33e-15
R-HSA-5663205	Infectious disease	4.33e-15
R-HSA-168255	Influenza Life Cycle	3.38e-14
R-HSA-168273	Influenza Viral RNA Transcription and Replication	7.99e-14
R-HSA-168254	Influenza Infection	1.24e-13
R-HSA-4086400	PCP/CE pathway	6.64e-13
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	1.19e-12
R-HSA-450531	Regulation of mRNA stability by proteins that bind AU-rich elements	2.23e-12
R-HSA-3858494	Beta-catenin independent WNT signaling	2.77e-12
R-HSA-195721	Signaling by Wnt	3.89e-12
R-HSA-2408522	Selenoamino acid metabolism	4.29e-12
R-HSA-201681	TCF dependent signaling in response to WNT	6.95e-12
R-HSA-8868773	rRNA processing in the nucleus and cytosol	1.26e-11
R-HSA-174178	APC/C:Cdh1 mediated degradation of Cdc20 and other targeting proteins in late mitosis/early G1	1.26e-11
R-HSA-174113	SCF-beta-TrCP mediated degradation of Emi1	9.75e-11

Table S15: Breast Cancer RNA-Seq Pathway Enrichment Cluster 5

Pathway identifier	Pathway name	Entities FDR
R-HSA-5617833	Assembly of the primary cilium	1.63e-13
R-HSA-1280215	Cytokine Signaling in Immune system	6.96e-11
R-HSA-881907	Gastrin-CREB signalling pathway via PKC and MAPK	7.11e-11
R-HSA-4420097	VEGFA-VEGFR2 Pathway	7.11e-11
R-HSA-8853659	RET signaling	7.63e-11
R-HSA-5683057	MAPK family signaling cascades	9.41e-11
R-HSA-194138	Signaling by VEGF	9.41e-11
R-HSA-512988	Interleukin-3, 5 and GM-CSF signaling	1.94e-10
R-HSA-5218921	VEGFR2 mediated cell proliferation	1.94e-10
R-HSA-167044	Signalling to RAS	1.94e-10
R-HSA-5620912	Anchoring of the basal body to the plasma membrane	1.94e-10
R-HSA-69275	G2/M Transition	1.94e-10
R-HSA-180336	SHC1 events in EGFR signaling	1.94e-10
R-HSA-1250347	SHC1 events in ERBB4 signaling	1.94e-10
R-HSA-112412	SOS-mediated signalling	1.94e-10
R-HSA-5673001	RAF/MAP kinase cascade	1.94e-10
R-HSA-179812	GRB2 events in EGFR signaling	1.94e-10
R-HSA-5654700	FRS-mediated FGFR2 signaling	1.95e-10
R-HSA-5654712	FRS-mediated FGFR4 signaling	1.95e-10
R-HSA-5654693	FRS-mediated FGFR1 signaling	1.95e-10
R-HSA-5654706	FRS-mediated FGFR3 signaling	1.95e-10
R-HSA-1852241	Organelle biogenesis and maintenance	1.95e-10
R-HSA-453274	Mitotic G2-G2/M phases	1.95e-10
R-HSA-187706	Signalling to p38 via RIT and RIN	2.71e-10
R-HSA-170984	ARMS-mediated activation	2.71e-10
R-HSA-187687	Signalling to ERKs	3.11e-10
R-HSA-5684996	MAPK1/MAPK3 signaling	3.22e-10
R-HSA-170968	Frs2-mediated activation	3.22e-10
R-HSA-2586552	Signaling by Leptin	3.44e-10
R-HSA-169893	Prolonged ERK activation events	3.81e-10

Table S16: Glioblastoma Multiforme RNA-Seq Pathway Enrichment Cluster 1

Pathway identifier	Pathway name	Entities FDR
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	8.44e-15
R-HSA-1474290	Collagen formation	8.44e-15
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	8.44e-15
R-HSA-216083	Integrin cell surface interactions	8.44e-15
R-HSA-1474244	Extracellular matrix organization	8.44e-15
R-HSA-3000171	Non-integrin membrane-ECM interactions	8.44e-15
R-HSA-3000178	ECM proteoglycans	8.44e-15
R-HSA-1442490	Collagen degradation	8.44e-15
R-HSA-1474228	Degradation of the extracellular matrix	8.44e-15
R-HSA-397014	Muscle contraction	8.44e-15
R-HSA-3000157	Laminin interactions	4.60e-14
R-HSA-3000170	Syndecan interactions	4.19e-13
R-HSA-445355	Smooth Muscle Contraction	9.98e-12
R-HSA-2129379	Molecules associated with elastic fibres	1.30e-11
R-HSA-1566948	Elastic fibre formation	1.53e-10
R-HSA-1793185	Chondroitin sulfate/dermatan sulfate metabolism	5.10e-10
R-HSA-422475	Axon guidance	2.67e-09
R-HSA-2022870	Chondroitin sulfate biosynthesis	1.26e-08
R-HSA-1266738	Developmental Biology	2.85e-08
R-HSA-390522	Striated Muscle Contraction	7.12e-08
R-HSA-2214320	Anchoring fibril formation	7.28e-08
R-HSA-1630316	Glycosaminoglycan metabolism	1.72e-07
R-HSA-3560782	Diseases associated with glycosaminoglycan metabolism	4.86e-07
R-HSA-419037	NCAM1 interactions	1.92e-06
R-HSA-375165	NCAM signaling for neurite out-growth	6.88e-06
R-HSA-162582	Signal Transduction	1.20e-05
R-HSA-4420332	Defective B3GALT6 causes EDSP2 and SEMDJL1	1.20e-05
R-HSA-3560783	Defective B4GALT7 causes EDS, progeroid type	1.20e-05
R-HSA-3560801	Defective B3GAT3 causes JDSSDHD	1.20e-05
R-HSA-1971475	A tetrasaccharide linker sequence is required for GAG synthesis	1.20e-05

Table S17: Glioblastoma Multiforme RNA-Seq Pathway Enrichment Cluster 2

Pathway identifier	Pathway name	Entities FDR
R-HSA-68877	Mitotic Prometaphase	5.33e-15
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	5.33e-15
R-HSA-2467813	Separation of Sister Chromatids	5.33e-15
R-HSA-69306	DNA Replication	5.33e-15
R-HSA-73886	Chromosome Maintenance	5.33e-15
R-HSA-69275	G2/M Transition	5.33e-15
R-HSA-113510	E2F mediated regulation of DNA replication	5.33e-15
R-HSA-2555396	Mitotic Metaphase and Anaphase	5.33e-15
R-HSA-68882	Mitotic Anaphase	5.33e-15
R-HSA-68886	M Phase	5.33e-15
R-HSA-69481	G2/M Checkpoints	5.33e-15
R-HSA-69278	Cell Cycle, Mitotic	5.33e-15
R-HSA-1640170	Cell Cycle	5.33e-15
R-HSA-69242	S Phase	5.33e-15
R-HSA-453279	Mitotic G1-G1/S phases	5.33e-15
R-HSA-69206	G1/S Transition	5.33e-15
R-HSA-69620	Cell Cycle Checkpoints	5.33e-15
R-HSA-194315	Signaling by Rho GTPases	5.33e-15
R-HSA-195258	RHO GTPase Effectors	5.33e-15
R-HSA-5663220	RHO GTPases Activate Formins	5.33e-15
R-HSA-453274	Mitotic G2-G2/M phases	1.02e-14
R-HSA-69239	Synthesis of DNA	1.40e-14
R-HSA-69190	DNA strand elongation	4.33e-14
R-HSA-5685942	HDR through Homologous Recombination (HRR)	1.36e-13
R-HSA-68962	Activation of the pre-replicative complex	2.89e-13
R-HSA-176187	Activation of ATR in response to replication stress	1.19e-12
R-HSA-68874	M/G1 Transition	2.93e-12
R-HSA-69002	DNA Replication Pre-Initiation	2.93e-12
R-HSA-5693537	Resolution of D-Loop Structures	1.09e-11
R-HSA-2132295	MHC class II antigen presentation	1.12e-11

Table S18: Glioblastoma Multiforme RNA-Seq Pathway Enrichment Cluster 3

Pathway identifier	Pathway name	Entities FDR
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	3.77e-15
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	3.77e-15
R-HSA-72689	Formation of a pool of free 40S subunits	3.77e-15
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	3.77e-15
R-HSA-927802	Nonsense-Mediated Decay (NMD)	3.77e-15
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	3.77e-15
R-HSA-72764	Eukaryotic Translation Termination	3.77e-15
R-HSA-192823	Viral mRNA Translation	3.77e-15
R-HSA-73857	RNA Polymerase II Transcription	3.77e-15
R-HSA-72737	Cap-dependent Translation Initiation	3.77e-15
R-HSA-72766	Translation	3.77e-15
R-HSA-72613	Eukaryotic Translation Initiation	3.77e-15
R-HSA-611105	Respiratory electron transport	3.77e-15
R-HSA-156842	Eukaryotic Translation Elongation	3.77e-15
R-HSA-163200	Respiratory electron transport, ATP synthesis by chemiosmotic coupling.	3.77e-15
R-HSA-156902	Peptide chain elongation	3.77e-15
R-HSA-162599	Late Phase of HIV Life Cycle	3.77e-15
R-HSA-1852241	Organelle biogenesis and maintenance	3.77e-15
R-HSA-162587	HIV Life Cycle	3.77e-15
R-HSA-162906	HIV Infection	3.77e-15
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	3.77e-15
R-HSA-1428517	The citric acid (TCA) cycle and respiratory electron transport	3.77e-15
R-HSA-74160	Gene Expression	3.77e-15
R-HSA-5663205	Infectious disease	3.77e-15
R-HSA-392499	Metabolism of proteins	3.77e-15
R-HSA-2408557	Selenocysteine synthesis	3.77e-15
R-HSA-168273	Influenza Viral RNA Transcription and Replication	3.77e-15
R-HSA-3700989	Transcriptional Regulation by TP53	3.77e-15
R-HSA-168254	Influenza Infection	3.77e-15
R-HSA-168255	Influenza Life Cycle	3.77e-15

Table S19: Glioblastoma Multiforme RNA-Seq Pathway Enrichment Cluster 4

Pathway identifier	Pathway name	Entities FDR
R-HSA-1474290	Collagen formation	9.88e-15
R-HSA-380108	Chemokine receptors bind chemokines	9.88e-15
R-HSA-1474228	Degradation of the extracellular matrix	9.88e-15
R-HSA-1474244	Extracellular matrix organization	9.88e-15
R-HSA-1442490	Collagen degradation	9.88e-15
R-HSA-422475	Axon guidance	9.88e-15
R-HSA-216083	Integrin cell surface interactions	9.88e-15
R-HSA-168256	Immune System	9.88e-15
R-HSA-168249	Innate Immune System	9.88e-15
R-HSA-109582	Hemostasis	9.88e-15
R-HSA-375276	Peptide ligand-binding receptors	9.88e-15
R-HSA-162582	Signal Transduction	9.88e-15
R-HSA-500792	GPCR ligand binding	9.88e-15
R-HSA-373076	Class A/1 (Rhodopsin-like receptors)	9.88e-15
R-HSA-372790	Signaling by GPCR	1.84e-14
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	2.60e-14
R-HSA-1266738	Developmental Biology	4.86e-14
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	7.66e-14
R-HSA-186797	Signaling by PDGF	2.53e-13
R-HSA-1280218	Adaptive Immune System	3.72e-13
R-HSA-1592389	Activation of Matrix Metalloproteinases	7.77e-13
R-HSA-114604	GPVI-mediated activation cascade	7.77e-13
R-HSA-1280215	Cytokine Signaling in Immune system	1.24e-12
R-HSA-418594	G alpha (i) signalling events	1.33e-12
R-HSA-881907	Gastrin-CREB signalling pathway via PKC and MAPK	2.77e-11
R-HSA-166520	Signalling by NGF	2.55e-10
R-HSA-877300	Interferon gamma signaling	3.17e-10
R-HSA-4420097	VEGFA-VEGFR2 Pathway	4.24e-10
R-HSA-1236975	Antigen processing-Cross presentation	7.56e-10
R-HSA-1236977	Endosomal/Vacuolar pathway	7.94e-10

Table S20: Glioblastoma Multiforme Microarray Pathway Enrichment Cluster 1

Pathway identifier	Pathway name	Entities FDR
R-HSA-112314	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	2.72e-14
R-HSA-112315	Transmission across Chemical Synapses	2.72e-14
R-HSA-372790	Signaling by GPCR	2.72e-14
R-HSA-112316	Neuronal System	2.72e-14
R-HSA-162582	Signal Transduction	2.72e-14
R-HSA-500792	GPCR ligand binding	2.72e-14
R-HSA-881907	Gastrin-CREB signalling pathway via PKC and MAPK	3.96e-13
R-HSA-1266738	Developmental Biology	2.97e-12
R-HSA-977443	GABA receptor activation	1.21e-11
R-HSA-166520	Signalling by NGF	1.21e-11
R-HSA-422475	Axon guidance	1.21e-11
R-HSA-388396	GPCR downstream signaling	1.81e-11
R-HSA-418555	G alpha (s) signalling events	1.96e-11
R-HSA-416476	G alpha (q) signalling events	2.84e-11
R-HSA-1296071	Potassium Channels	3.47e-10
R-HSA-187037	NGF signalling via TRKA from the plasma membrane	4.98e-10
R-HSA-422356	Regulation of insulin secretion	2.00e-09
R-HSA-210500	Glutamate Neurotransmitter Release Cycle	2.35e-09
R-HSA-373076	Class A/1 (Rhodopsin-like receptors)	1.03e-08
R-HSA-373760	L1CAM interactions	1.30e-08
R-HSA-1296072	Voltage gated Potassium channels	1.86e-08
R-HSA-375276	Peptide ligand-binding receptors	2.20e-08
R-HSA-163685	Integration of energy metabolism	2.75e-08
R-HSA-195721	Signaling by Wnt	3.79e-08
R-HSA-381676	Glucagon-like Peptide-1 (GLP1) regulates insulin secretion	5.09e-08
R-HSA-111885	Opioid Signalling	5.78e-08
R-HSA-373080	Class B/2 (Secretin family receptors)	1.72e-07
R-HSA-397014	Muscle contraction	3.22e-07
R-HSA-3858494	Beta-catenin independent WNT signaling	3.36e-07
R-HSA-177929	Signaling by EGFR	3.81e-07

Table S21: Glioblastoma Multiforme Microarray Pathway Enrichment Cluster 2

Pathway identifier	Pathway name	Entities FDR
R-HSA-68962	Activation of the pre-replicative complex	4.44e-15
R-HSA-68877	Mitotic Prometaphase	4.44e-15
R-HSA-69206	G1/S Transition	4.44e-15
R-HSA-453279	Mitotic G1-G1/S phases	4.44e-15
R-HSA-5693607	Processing of DNA double-strand break ends	4.44e-15
R-HSA-69242	S Phase	4.44e-15
R-HSA-69306	DNA Replication	4.44e-15
R-HSA-176187	Activation of ATR in response to replication stress	4.44e-15
R-HSA-5693538	Homology Directed Repair	4.44e-15
R-HSA-5693579	Homologous DNA Pairing and Strand Exchange	4.44e-15
R-HSA-69190	DNA strand elongation	4.44e-15
R-HSA-5693567	HDR through Homologous Recombination (HR) or Single Strand Annealing (SSA)	4.44e-15
R-HSA-69239	Synthesis of DNA	4.44e-15
R-HSA-5685942	HDR through Homologous Recombination (HRR)	4.44e-15
R-HSA-5693616	Presynaptic phase of homologous DNA pairing and strand exchange	4.44e-15
R-HSA-5693532	DNA Double-Strand Break Repair	4.44e-15
R-HSA-69278	Cell Cycle, Mitotic	4.44e-15
R-HSA-1640170	Cell Cycle	4.44e-15
R-HSA-68886	M Phase	4.44e-15
R-HSA-5685938	HDR through Single Strand Annealing (SSA)	4.44e-15
R-HSA-69481	G2/M Checkpoints	4.44e-15
R-HSA-73894	DNA Repair	4.44e-15
R-HSA-69473	G2/M DNA damage checkpoint	4.44e-15
R-HSA-69620	Cell Cycle Checkpoints	4.44e-15
R-HSA-453274	Mitotic G2-G2/M phases	2.60e-14
R-HSA-5620912	Anchoring of the basal body to the plasma membrane	3.70e-14
R-HSA-3700989	Transcriptional Regulation by TP53	3.12e-13
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	8.55e-13
R-HSA-69275	G2/M Transition	8.61e-13
R-HSA-380284	Loss of proteins required for interphase microtubule organization from the centrosome	9.31e-13

Table S22: Glioblastoma Multiforme Microarray Pathway Enrichment Cluster 3

Pathway identifier	Pathway name	Entities FDR
R-HSA-72695	Formation of the ternary complex, and subsequently, the 43S complex	2.78e-15
R-HSA-72702	Ribosomal scanning and start codon recognition	2.78e-15
R-HSA-72649	Translation initiation complex formation	2.78e-15
R-HSA-72689	Formation of a pool of free 40S subunits	2.78e-15
R-HSA-72662	mRNA activation upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	2.78e-15
R-HSA-5389840	Mitochondrial translation elongation	2.78e-15
R-HSA-72613	Eukaryotic Translation Initiation	2.78e-15
R-HSA-72737	Cap-dependent Translation Initiation	2.78e-15
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	2.78e-15
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	2.78e-15
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	2.78e-15
R-HSA-5419276	Mitochondrial translation termination	2.78e-15
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	2.78e-15
R-HSA-927802	Nonsense-Mediated Decay (NMD)	2.78e-15
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	2.78e-15
R-HSA-192823	Viral mRNA Translation	2.78e-15
R-HSA-611105	Respiratory electron transport	2.78e-15
R-HSA-5368287	Mitochondrial translation	2.78e-15
R-HSA-72766	Translation	2.78e-15
R-HSA-163200	Respiratory electron transport, ATP synthesis by chemiosmotic coupling.	2.78e-15
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	2.78e-15
R-HSA-156902	Peptide chain elongation	2.78e-15
R-HSA-72764	Eukaryotic Translation Termination	2.78e-15
R-HSA-156842	Eukaryotic Translation Elongation	2.78e-15
R-HSA-5368286	Mitochondrial translation initiation	2.78e-15
R-HSA-162906	HIV Infection	2.78e-15
R-HSA-1428517	The citric acid (TCA) cycle and respiratory electron transport	2.78e-15
R-HSA-1852241	Organelle biogenesis and maintenance	2.78e-15
R-HSA-8868773	rRNA processing in the nucleus and cytosol	2.78e-15
R-HSA-5663205	Infectious disease	2.78e-15

Table S23: Glioblastoma Multiforme Microarray Pathway Enrichment Cluster 4

References

[1] Mohamed Abdouh, Sabrina Facchino, Wassim Chatoo, Vijayabalan Balasingam, José Ferreira, and Gilbert Bernier. Bmi1 sustains human glioblastoma multiforme stem cell renewal. *The Journal of Neuroscience*, 29(28):8884–8896, 2009.

[2] C Borgelt and M R Berthold. Mining molecular fragments: finding relevant substructures of molecules. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 51–58, 2002.

[3] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

[4] Tonia Cenci, Maurizio Martini, Nicola Montano, Quintino G D'Alessandris, Maria Laura Falchetti, Daniela Annibali, Mauro Savino, Federico Bianchi, Francesco Pierconti, Sergio Nasi, et al. Prognostic relevance of c-myc and bmi1 expression in patients with glioblastoma. *American journal of clinical pathology*, 138(3):390–396, 2012.

[5] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

[6] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2014.

[7] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

[8] Arda Durmaz, Tim A.D. Henderson, Douglas Brubaker, and Gurkan Bebek. Frequent subgraph mining of personalized signaling pathway networks groups patients with frequently dysregulated disease pathways and predicts prognosis. *Pac Symp Biocomput*, 2017 (in press).

[9] Renaud Gaujoux and Cathal Seoighe. A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):1, 2010.

[10] Robert C Gentleman et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[11] Shi-ming He, Zhen-wei Zhao, Yuan Wang, Ji-pei Zhao, Liang Wang, Fang Hou, and Guo-dong Gao. Reduced expression of smad4 in gliomas correlates with progression and survival of patients. *Journal of Experimental & Clinical Cancer Research*, 30(1):1, 2011.

[12] Jun Huan, Wei Wang, and Jan Prins. Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 549—-, Washington, DC, USA, 2003. IEEE Computer Society.

[13] Jun Huan, Wei Wang, Jan Prins, and Jiong Yang. SPIN: Mining Maximal Frequent Subgraphs from Graph Databases. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, number 1, page 581, New York, New York, USA, 2004. ACM Press.

[14] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In Djamel A. Zighed, Jan Komorowski, and Jan Żytkow, editors, *Principles of Data Mining and Knowledge Discovery - PKDD*, volume 1910 of *Lecture Notes in Computer Science*, pages 13–23. Springer Berlin Heidelberg, Berlin, Heidelberg, jul 2000.

[15] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[16] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[17] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[18] Yu Li, Sicong Zhang, and Suyun Huang. Foxm1: a potential drug target for glioma. *Future Oncology*, 8(3):223–226, 2012.

[19] Yiyi Liu, Quanquan Gu, Jack P Hou, Jiawei Han, and Jian Ma. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC bioinformatics*, 15(1):1, 2014.

[20] Marija Milacic, Robin Haw, Karen Rothfels, Guanming Wu, David Croft, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, 4(4):1180–1211, 2012.

[21] Siegfried Nijssen and Joost N Kok. A Quickstart in Frequent Structure Mining Can Make a Difference. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 647–652, New York, NY, USA, 2004. ACM.

[22] Alberto Pascual-Montano, Jose Maria Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):403–415, 2006.

[23] Maureen M Sherry, Andrew Reeves, Julian K Wu, and Brent H Cochran. Stat3 is required for proliferation and maintenance of multipotency in glioblastoma stem cells. *Stem cells*, 27(10):2383–2392, 2009.

[24] Chitra Venugopal, Na Li, Xin Wang, Branavan Manoranjan, Cynthia Hawkins, Thorsteinn Gunnarsson, Robert Hollenberg, Paula Klurfan, Naresh Murty, Jacek Kwiecien, et al. Bmi1 marks intermediate precursors during differentiation of human brain tumor initiating cells. *Stem cell research*, 8(2):141–153, 2012.

[25] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

[26] JunYu Wang, Qi Wang, Yong Cui, Zhen Yang Liu, Wei Zhao, Chun Lin Wang, Yan Dong, LiJun Hou, GuoHan Hu, Chun Luo, et al. Knockdown of cyclin d1 inhibits proliferation, induces apoptosis, and attenuates the invasive capacity of human glioblastoma cells. *Journal of Neuro-oncology*, 106(3):473–484, 2012.

[27] Xifeng Yan and Jiawei Han. gSpan: graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724. IEEE Comput. Soc, 2002.