An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types

Sunho Park[1], Seung-Jun Kim[2], Donghyeon Yu[1], Samuel Pena-Llopis[3,4], Jianjiong Gao[5], Jin Suk Park[1], Hao Tang[1], Beibei Chen[1], Jiwoong Kim[1], Jessie Norris[1], Xinlei Wang[6], Min Chen[7], Minsoo Kim[1], Jeongsik Yong[8], Zabi WarDak[4,9], Kevin S Choe[4,9], Michael Story[4,9], Timothy K. Starr[10,11], Jaeho Cheong[†12], Tae Hyun Hwang[†1,4]

[1]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

[2]Department of Computer Science and Electrical Engineering, University of Maryland at Baltimore, Baltimore County, Maryland, United States of America

[3]Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

[4]Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas, USA.

[5]Center for Molecular Biology, Memorial Sloan Kettering Cancer Center, New York, New York, 10065, United States of America

[6]Department of Statistical Science, Southern Methodist University, Dallas, Texas, United States of America

[7]Department of Mathematical Sciences, University of Texas at Dallas, Dallas, Texas, United States of America

[8]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States of America

[9]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

[10]Masonic Cancer Center, University of Minnesota – Twin Cities, Minneapolis, Minnesota, Unites States of America

[11]Department of Obstetrics, Gynecology & Women's Health and Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, Minnesota, United States of America

[12]Department of Surgery, Yonsei University College of Medicine, Seoul, Korea

[†]Corresponding author

Email addresses:

SP: sunho.park@utsouthwestern.edu

SK: sjkim@umbc.edu

DY: donghyeon.yu@utsouthwestern.edu

SPL: samuel.pena-llopis@utsouthwestern.edu

JG: jgao@cbio.mskcc.org

JP:jinsuk.park@utsouthwestern.edu

HT: hao.tang@utsouthwestern.edu

BC: beibei.chen@utsouthwestern.edu

JK: jiwoong.kim@utsouthwestern.edu

JN: jessie.norris@utsouthwestern.edu

XW: swang@mail.smu.edu

MC: mchen@utdallas.edu

MK: mins.kim@utsouthwestern.edu

JY:jyong@umn.edu

ZW:zabi.wardak@utsouthwestern.edu

KC: kevin.choe@utsouthwestern.edu

MS: michael.story@utsouthwestern.edu

TS: star0044@umn.edu

JC: jhcheong@yuhs.ac

THH: taehyun.hwang@utsouthwestern.edu

Abstract

Identification of altered pathways that are clinically relevant across human cancers is a key challenge in cancer genomics. We developed a network-based algorithm to integrate somatic mutation data with gene networks and pathways, in order to identify pathways altered by somatic mutations across cancers. We applied our approach to The Cancer Genome Atlas (TCGA) dataset of somatic mutations in 4,790 cancer patients with 19 different types of malignancies. Our analysis identified cancer-type-specific altered pathways enriched with known cancer-relevant genes and drug targets. Consensus clustering using gene expression datasets that included 4,870 patients from TCGA and multiple independent cohorts confirmed that the altered pathways could be used to stratify patients into subgroups with significantly different clinical outcomes. Of particular significance, certain patient subpopulations with poor prognosis were identified because they had specific altered pathways for which there are available targeted therapies. These findings could be used to tailor and intensify therapy in these patients, for whom current therapy is suboptimal.

Background

In the last few years, studies using high-throughput technologies have highlighted the fact that the development and progression of cancer hinges on somatic alterations. These somatic alterations may disrupt gene function, such as activating oncogenes or inactivating tumor suppressor genes, and thus, dysregulate critical pathways contributing to tumorigenesis. Therefore, precise identification and understanding of disrupted pathways may provide insights into therapeutic strategies and the development of novel agents. Many large-scale cancer genomics studies, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have performed an integrated analysis to draft an overview of somatic alterations in the cancer genome[1-4]. Many of these studies have reported novel

candidate cancer genes mutated at high and intermediate frequencies in a specific cancer as well as across many cancer types[4]. However, it is a still challenge to translate somatic mutations in tumors into the pathway model to accurately predict patient clinical outcomes [5, 6]. Recently, in order to improve the clinical relevance and utility of somatic mutation analyses, Hopfree et al[7] proposed integrating somatic mutation data with molecular interaction networks for patient stratification. They demonstrated that inclusion of prior knowledge, captured in molecular interaction networks, could improve identification of patient subgroups with significantly different histological, pathological or clinical outcomes and discover novel cancer-related pathways or subnetworks. In a similar manner, other network-based methods have demonstrated that incorporating molecular networks and/or biological pathways can improve accuracy in identifying cancer-related pathways[8-11].

One limitation of these network-based methods is that they are not designed to fully utilize large-scale somatic mutation data from multiple cancer types to determine which particular pathways are altered by somatic mutations across a range of human cancers. In addition, due to the incomplete knowledge of existing gene set and/or pathway database, these methods are limited to detect pathways based on a number of altered genes annotated in existing gene set and pathway databases. Alternatively, the methods that build pathways *de novo* without incorporating biological prior knowledge can be applicable to detect altered pathways, but these methods were not designed to detect cancer-type specific or commonly altered pathways either.

To address these, we developed an algorithm named NTriPath (Network regularized non-negative TRI matrix factorization for PATHway identification) to integrate somatic mutation, gene-gene interaction networks and gene set or pathway databases to discover pathways altered by somatic mutations in 4,790 cancer patients with 19

different types of cancers. Incorporating existing gene set or pathway databases enables NTriPath to report a list of altered pathways across cancers, and make easy to determine/compare which particular pathways are altered in a particular cancer type(s). In particular, the use of large-scale genome-wide somatic mutations from 4,790 cancer patients and gene-gene interaction networks enables NTriPath to classify genes, which were not annotated in existing gene set or pathway databases, as new member genes of the identified altered pathways based on modular structures of mutational data within a cancer type and/or across multiple cancer types (using matrix factorization) and connectivity in the gene-gene interaction networks. The questions that we investigate here are: first, whether large-scale integrative somatic mutation analysis can reliably identify cancer-type-specific or commonly altered pathways by somatic mutations across cancers; second, whether the identified pathways can be used as a prognostic biomarker for patient stratification - with the assumption that the altered pathways contribute to cancer development and progression and, thus, impact survival.

In these experiments, we demonstrated that the cancer-type-specific and commonly altered pathways identified by NTriPath are biologically relevant to the corresponding cancer type and are associated with patient survival outcomes. We also showed that cancer-specific altered pathways are enriched with many known cancer-relevant genes and targets of available drugs including those already FDA-approved. These results imply that the cancer-specific altered pathways can guide therapeutic strategy to target the altered pathways that are pivotal in each cancer type.

Results

NTriPath: An integrative somatic mutation analysis for discovering pathways altered by somatic mutations across multiple cancer types

NTriPath integrates somatic mutations with gene-gene interaction networks and a pathway database to discover altered pathways across human cancers. We collected somatic mutation data from TCGA for 4,790 patients and 19 different cancer types (Table 1). A diagrammatic description of our algorithm is depicted in Figure 1. Four types of data were used as input for our algorithm. First, we generated a binary matrix ($\boldsymbol{X}$) of patients x genes, with '1' indicating a mutation and '0' no mutation. Second, we constructed gene-gene interaction networks ($\boldsymbol{A}$). Third, we incorporated a pathway database ($\boldsymbol{V_0}$) (e.g., conserved 4,620 subnetworks across species[12]). Fourth, we included clinical data on the patient's tumor type ($\boldsymbol{U}$). NTriPath produces two matrices as output; 1) altered pathways by mutated genes ($\boldsymbol{V}$) and 2) altered pathways by cancer type matrix ($\boldsymbol{S}$). The use of both large-scale somatic mutation profiles and gene-gene interaction networks enabled NTriPath to identify cancer-related pathways containing known cancer genes mutated at different frequencies across cancers with newly added member genes according to high network connectivity ($\boldsymbol{V}$) [8, 13]. Finally we use the altered pathways by cancer type matrix ($\boldsymbol{S}$) to identify altered pathways that are specific for each cancer type. For further details, please see the Materials and Methods section. Our method is also available at www.taehyunlab.org.

NTriPath identifies cancer-type-specific altered pathways that are biologically and clinically relevant

In each cancer type, we selected the top 3 ranked altered pathways by statistical significance from NTriPath with the 4,620 subnetwork modules to generate cancer-type-specific altered pathways (See Material and Method section and Supplementary Table 1). Interestingly, NTriPath was able to find altered pathways containing not only genes that were frequently mutated but also genes that were

mutated in a small subset of patients in each cancer type (Supplementary Table 2 and Supplementary Figure 1). Gene set enrichment analysis using the genes from the top 3 altered pathways showed that the altered pathways are significantly enriched with well-known cancer-related genes from COSMIC database [14] and known drug target genes as well as cancer-relevant biological processes (Supplementary Table 3 and 4).

Focusing on kidney renal clear cell carcinoma (KIRC) as a proof of concept, NTriPath identified the pathway consisting of *VHL, USP33, DIO2*, *TCEB1* and *TCEB2* as the top-ranked altered pathway in KIRC (Figure 2A). The *VHL* (*von-Hippel Lindau*) gene is a well-known tumor suppressor associated with KIRC, and is frequently mutated in patients with KIRC[15-18]. *VHL* was the most frequently mutated gene in TCGA KIRC with 55.7% of patients harboring mutations in the gene. *TCEB1* is mutated at very low frequency in TCGA KIRC cohort. A recent study found that *TCEB1* is mutated in about 3% of the KIRC patients without *VHL* inactivation, and found *TCEB1* preventing the binding of Elongin C to *VHL,* which inactivates the *VHL* pathway[16]. The second highest ranked pathway contained *EP300* and *TP53*. *EP300* and *TP53* were mutated in 8.1% and 5.2% of patients, respectively. *EP300* has been identified as a co-activator of hypoxia-inducible factor 1 alpha (*HIF1*α), whose activation is a hallmark of KIRC tumors. *TP53* was previously found to be associated with poor outcome in TCGA KIRC[19]. The third highest ranked pathway contains *LRP1* and matrix metalloproteinases (*MMP1*, *MMP7*, *MMP9*, *MMP26*). *LRP1* is mutated in 10% of TCGA KIRC cohort, but matrix metalloproteinases (*MMPs*) were not mutated in TCGA KIRC cohort. Biological and clinical relevance of *LRP1* mutation in KIRC has not been previously reported. *MMPs* have been implicated in different types of cancer progression including the acquisition of invasive and metastatic properties in many cancer types. The aberrant expression of *MMPs* has

been associated with poor patient survival and prognosis in KIRC patients [8, 20]. Interestingly, recent studies suggested that *LRP1* induces the expression of matrix metalloproteinase (MMPs) and thus promotes cancer cell invasion and metastasis in many cancers including KIRC[16,21-23].

NTriPath identified many new member genes in the top ranked pathways including *TCEB2*, *JUN*, and *SP1* as well as other tumor suppressors such as *CREBBP*, *SMAD3*, *BRCA1* and *RB1*. These newly identified member genes by NTriPath were mutated at a very low frequency or not mutated at all in TCGA KIRC patients. Instead, these genes interacted with many frequently mutated genes in the networks and were often dysregulated at the mRNA and protein levels in many KIRC patients (Figure 2B). For example, *TCEB2*, *SP1* and *JUN* were not mutated but yet their expression was dysregulated in 7%, 10% and 2% of TCGA KIRC patients, respectively. Previous studies have shown that dysregulation in *TCEB2* is expected to disrupt the protein complex that ubiquitinates *HIF1α,* resulting in the same phenotype as *VHL* inactivation by mutation or promoter hypermethylation[24-26]. In addition, *SP1* and *JUN* were previously identified as major transcriptional regulators associated with signaling circuit to promote tumor growth and invasion in KIRC [18].

Taken together, these results demonstrate that NTriPath is an effective tool to accurately identify cancer-specific altered pathways including known cancer genes mutated at a high or intermediate frequency in the patients, as well as genes mutated at a very low frequency or not mutated at all yet may be fundamental role in development and/or progression of KIRC.

Cancer-type-specific altered pathways across cancer types correlate with patient survival outcomes

We hypothesized that cancer-type-specific altered pathways reflect the molecular basis underlying the patient clinical outcomes. This would allow us to use member

genes in the altered pathways as gene signatures to stratify patients into subgroups with different clinical outcomes for each type of cancer. We first collected a dataset consisting of gene expression profiles from 3,656 patients with their survival information from TCGA cohorts. We then used member genes in the top 3 ranked cancer-type-specific altered pathways to perform consensus clustering for each cancer type (see Material and Method section). We generated Kaplan-Meier (KM) curves based on the groups produced by consensus clustering and found that patient survival was significantly different among the groups (Figure 3 and Supplementary Figure 2). In TCGA KIRC, we found three patient subgroups (A, B and C), with the Group C having the poorest survival. A log-rank test indicated that Groups A and C had significantly different survival outcomes (Log-rank test p-value = 1.840e-08, Hazard ratio = 2.94) with median survival times of 41.9 months for group A compared to 30.8 months for group C (Figure 3A). Other examples are Bladder Urothelial Carcinoma (BLCA), Head and Neck squamous cell carcinoma (HNSC), and Skin Cutaneous Melanoma (SKCM) patient subgroups identified by NTriPath pathway signatures. While the molecular classification of clinically relevant subtypes of these cancers is still challenging, we found patient subgroups having significantly different survival in these cancers (Log-rank test p-value= 0.0086, 0.0010, and 0.0210, repectively) (Figure 3B, 3C and 3D).

Experiments with other TCGA datasets, including those for Breast invasive carcinoma (BRCA), Glioblastoma Multiforme (GBM), Lung adenocarcinoma (LUAD), and Ovarian serous cystadenocarcinoma (OV) consistently showed that the use of member genes in cancer-type-specific altered pathways could serve as a prognostic biomarker for patient stratification (Supplementary Figure 2 and 3). For comparison, we also attempted to cluster patients using significant frequently mutated

genes previously identified by the TCGA Pan-Cancer study[1]. The results of consensus clustering using the NTriPath-derived pathway signatures and the TCGA Pan-Cancer-derived mutated gene signatures showed that the results from NTriPath-derived pathway signatures had higher significance levels for BLAC, BRCA, and KIRC, and comparable results for the GBM, HNSC, and LUAD cancer types (Figure 4). These findings suggested that NTriPath-derived altered pathways could be used as prognostic biomarkers for better patient stratification.

Independent cohorts for the validation of the cancer-type-specific altered pathways

We performed multiple validations to evaluate the robustness and the reproducibility of NTriPath. First, we evaluated the robustness of the cancer-type-specific altered pathways identified in the TCGA cohort for prognostic stratification. We generated gene expression profiles of 102 HNSC patients from our institution and used the member genes of the top 3 HNSC cancer-type-specific altered pathways in the TCGA cohort for patient stratification. In addition, we also used publically available gene expression data from two ovarian cancer datasets, one lung cancer dataset, two colon cancer datatsets for a total of 1,112 patients, and used the top 3 cancer-type specific altered pathways for corresponding cancer type for independent validation. In the HNSC cohorts, we found six patient subgroups (A through F), with the group F patients having the poorest survival times (Figure 5A). A log-rank test indicated that groups A and F had significantly different survival outcomes (p-value = 0.038, hazard ratio = 1.88) with median survival times of 78.1 months for group A and 26.7 months for group F. Similarly, we found patient subgroups having significantly different survival outcomes in lung cancer, ovarian cancer, and colorectal cancer (Figure 5B-D and Supplementary Figure 3). Secondly, we verified the reproducibility of NTriPath for the identification of the cancer-type-specific altered pathways. We collected the

level 2 somatic mutation data from 19 human cancer types those were updated after we collected initial dataset used in the original experiments from the TCGA data portal. We found that there are 1891 newly updated patients' mutation data from 15 cancer types (see Supplementary Table 5). We re-ran NTriPath to identify cancer-type-specific pathways across 19 cancers using 6681 patients' somatic mutation data including those of newly updated patients' mutation data. Interestingly, we found that many top ranked pathways identified by NTriPath in the original experiments were consistently highly ranked in the new experiments (see Supplementary Table 6).

These results reassure that NTriPath is a robust tool to detect the altered pathways across cancers, and the altered pathways identified by NTriPath can serve as robust prognostic signatures for identifying patient subgroups with different survival outcomes across multiple cancer types.

NTriPath identified potential therapeutic targets in poor prognosis patient subgroups

We further investigated whether we could identify potential targets or the therapy for the identified poor prognosis patient subgroups. Interestingly, we found that many known drug targets in the cancer-type specific altered pathways are often up-regulated in poor prognosis patient subgroups across cancers (see Method section and Supplementary 7). For example, in TCGA KIRC cohort, *LRP1* and *MMP9,* targets of FDA-approved drugs Tenecteplase and Captopril, were significantly up-regulated in poor prognosis group compared to good prognosis group (FDR-adjusted p-value < 0.05 with t-test). Tenecteplase binds to *LRP1* and induces both *LRP1* and *MMP9* expression, and Captoprilare inhibits *MMP9* expression. Thus, combinatorial therapy of these drugs can be beneficial for the KIRC patients with high *LRP1* and *MMP9* expression [27-44]. Another notable example includes DNA Topoisomerase I (*TOP1),* a target of well-known FDA-approved anticancer drugs such as Irinotecan and

Topotecan, identified by NTriPath as a new member gene into the cancer-type-specific altered pathways across many cancers including HNSC. Interestingly, we found that *TOP1* was up-regulated in poor prognosis subgroups in HNSC from both TCGA and UTSW cohorts (Group E and F in Figure 3C and 5A, respectively). In addition, we found that some patients with overexpression of *TOP1* in TCGA HNSC poor prognosis subgroup have developed therapy resistance against single chemotherapeutic agent such as Cisplatin. Interestingly, there is an ongoing trial in advanced HNSC showing efficacy of *TOP1* inhibitor Irinotecan with Cisplatin in a poor prognosis patient subgroup [45]. These observations may suggest that *TOP1* inhibitors-based combinations might offer an effective treatment option for HNSC patients with overexpression of *TOP1*. Taken together, these findings suggested that the use of NTriPath-derived altered pathways containing available drug targets may allow for the development of more tailored therapeutics.

Discussion

Systematic understanding of how somatic mutations influence clinical outcomes is essential for the development and application of personalized therapies. Especially organizing alterations at the individual gene level and in the molecular pathways can correlate altered pathways and vulnerabilities with specific genetic lesions, and provide novel insights into cancer biology, biomarkers for patient stratification in clinical trials, and potential targeted drug development[46]. Here, we systematically identified biological and clinical relevant cancer-type-specific altered cross multiple cancer types. In particular, the integration of somatic mutation with biological prior knowledge led to the identification of altered pathways that contain recurrently mutated genes as a hallmark of specific cancer types. Interestingly, we found that several genes, while not frequently mutated or not mutated at all in patients, were part

of cancer-type-specific altered pathways that have been causally implicated in the development of corresponding cancer types, and expressions of those genes are significantly associated with clinical outcomes (Supplementary Figure 4). For example, no mutation of *MMP7* has been reported, but high expression of *MMP7* (p = 0.00191, HR=1.7 (95%CI,1.21−2.38)) is significantly associated with poor survival in TCGA KIRC patients. Other examples include *CABLES1* (p = 0.00272, HR=0.486 (95%CI,0.301−0.787)) in TCGA HNSC and LUAD, and *GCH1* (p = 0.0000528, HR=0.52 (95%CI,0.367−0.763)) in TCGA SKCM are not frequently or not mutated but low or high expression of those genes are significantly associated with poor survival. In addition, we found that known drug targets are not frequently mutated but often up-regulated in poor prognosis patient subgroups across many cancers. These results further corroborate that the integrative analysis of somatic mutations with additional biological prior knowledge may elucidate potential candidate genes associated with clinical outcomes and could be potentially used to design targeted therapy, which cannot be readily identified by somatic mutation analysis alone.

In our analysis, we did not remove synonymous mutations or further select a shorter list of recurrent mutated genes in cohorts with stringent criteria either [3, 47]. However, Hopfree et al[7] showed that filtering synonymous mutations resulted in a decreased ability to detect patient subgroups with different survival outcomes. Another recent study also showed that synonymous mutations could affect functions of oncogene and tumor suppressors [48]. In addition, our experimental results for patient stratification in comparison with recurrent mutated gene signatures identified by the TCGA Pan-Cancer[1] indicated that the use of NTriPath-derived pathways showed a comparable or better performance in discovering patient subgroups with different survival outcomes across cancers. To evaluate the impact of different network resources, we used

networks from the HPRD [49] and Rossin, E.J. *et al*[50] and repeated experiments. We summarize the results of altered pathways and patient stratification using different network resources and provide on our supplement website.

Lastly, NTriPath is a general computational algorithm and can be applied to other data types such as gene expression, copy number alteration, and methylation to identify altered pathways by different types of genomic aberrations. NTriPath can also be used to find altered pathways across associated with other cancer-related phenotypes (e.g., patient groups having therapy resistance vs. sensitivity, metastatic vs. non-metastatic).

Conclusions

We have described an integrative somatic mutation analysis for discovering altered pathways in human cancers. NTriPath integrates somatic mutation data and prior biological knowledge from the pathway database and molecular networks to identify significantly altered pathways and their associations with specific cancer types. Specifically, NTriPath effectively utilizes mutation patterns that exist in only a subset of samples (or specific cancer types), thus revealing pathways altered by complex mutation patterns across cancer types. Furthermore, the use of gene-gene interaction networks and the pathway database provides the potential to identify altered pathways enriched with genes harboring mutations at high/intermediate frequencies, as well as those not mutated per se but nevertheless playing critical roles in tumorigenesis in network and pathway contexts. Thus, NTriPath is uniquely suited to provide a global analysis of altered pathways by somatic mutation across cancer types.

We applied NTriPath to somatic mutation data from 19 types of cancers, and discovered cancer-type-specific altered pathways based on these mutations in human cancers. Functional enrichment analysis of cancer-type-specific pathways demonstrated that the identified cancer-type-specific altered pathways are biologically

meaningful to each cancer type. It also provided unique pathway views of key biological processes underlying each cancer type. Of particular significance, we identified a patient subgroup with poor survival by cancer-type-specific altered pathway signatures from TCGA cohorts, which in independent cohorts. These results implied the potential utility of cancer-type-specific altered pathway signatures to serve as a guide to tailored treatment in a patient subgroup.

Materials and Methods

Somatic mutation, human gene-gene interaction networks, and pathway data

The level 2 somatic mutation data from 19 human cancer types were collected from the TCGA data portal on May 19th 2013[51]. We constructed a gene-gene interaction network by combining networks from Zhang, S. *et al* [52], the Human Protein Reference Database (Dec. 2013)[53] and Rossin, E.J. *et al*[50]..Four sets of pathways were used in the analysis: 1) 4,620 conserved subnetworks from the human gene-gene interaction network[12], 2) KEGG, 3) Biocarta, and 4) Reactome gene sets from MsigDB (Sept. 2010)[54].

Algorithm

The algorithm identifies pathways disrupted by mutated genes. Disrupted pathways are found based on the factorization results from the network regularized nonnegative tri-matrix factorization.

1. Notations

We construct a binary data matrix $\boldsymbol{X} \in R^{N \times M}$ from the mutation data, where N is the number of patients, M is the number of genes and the $(i,j)^{th}$ element of the matrix $\boldsymbol{X}$, $[\boldsymbol{X}]_{ij}$, is '1' if the ith patient has a mutation on the jth gene, '0' otherwise. We derive the adjacency matrix from the human gene-gene interaction networks and denote it as

$\boldsymbol{A}$, where $[\boldsymbol{A}]_{ij}$='1' if the ith gene is interacting with the jth genes in the networks and '0' otherwise. We define the graph Laplacian matrix by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, where each diagonal element in the diagonal matrix $\boldsymbol{D}$ is given by $[\boldsymbol{D}]_{ii} = \sum_{ji}[\boldsymbol{A}]_{ij}$. We construct a binary matrix $\boldsymbol{U} \in R^{N\times K_1}$ denoting patient cluster, where K_1 indicates the number of cancer types and $[\boldsymbol{U}]_{ij}$=1 indicates the ith patient has jth cancer type. We construct a binary matrix $\boldsymbol{V}_0 \in R^{M\times K_2}$ from the specific pathway database denoting pathway information, where K_2 is the number of pathways and $[\boldsymbol{V}_0]_{ij}$=1 if the ith gene is annotated in jth pathway as a member in the pathway database, otherwise 0. Since current pathway database annotation is still incomplete, we define a matrix $\boldsymbol{V} \in R^{M\times K_2}$ denoting newly updated pathway information including newly added member genes by NTriPath. We define a matrix $\boldsymbol{S} \in R^{K_1\times K_2}$ denoting cancer type and pathway associations, where each element of $[\boldsymbol{S}]_{ij}$ represents associations between ith cancer type with j th pathway. Higher values of elements indicate stronger associations between cancer types and pathways. Since $\boldsymbol{V}$ and $\boldsymbol{S}$ are unknown, we need to learn about those matrices during optimization (see below section for details)

2. Network regularized non-negative tri-matrix factorization

The network regularized nonnegative tri-matrix factorization is an extension of Nonnegative Tri Matrix Factorization (NTMF); in this work, somatic mutation data $\boldsymbol{X}$ is factorized as the products of three element-wise non-negative matrices $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$ denoting patient's cancer type, cancer-type and pathway associations, and cancer-related pathways, respectively. We here consider a weighted loss function to deal with the sparseness of the data matrix $\boldsymbol{X}$ (More than 98% of entries are zero). It enables us to focus on an approximation error at nonzero entries, which correspond to mutated genes. In addition, to incorporate the prior knowledge from human gene-gene interaction networks and pathway datasets into factorizations, we enforce constraints

on parameters, which involve the graph Laplaican $\boldsymbol{L}$ and the pathway information $\boldsymbol{V}_0$. All these ideas are accomplished by minimizing the following objective function

$$\min_{S,V\geq 0} \| \boldsymbol{W} \circ \left(\boldsymbol{X} - \boldsymbol{USV^T}\right) \|_F^2 + \lambda_S \| \boldsymbol{S} \|_1^2 + \lambda_V \| \boldsymbol{V} \|_1^2 + \lambda_0 \| \boldsymbol{V} - \boldsymbol{V}_0 \|_F^2 + \lambda_L tr\{\boldsymbol{V^T LV}\},$$

where $\boldsymbol{W} \in R^{N\times M}$ is a weight matrix where $[\boldsymbol{W}]_{ij} = 1$ if $[\boldsymbol{X}]_{ij} > 0$ otherwise 0, and the operator $\circ$ represents the element-wise multiplication. Here, we are only interested in learning of $\boldsymbol{S}$ and $\boldsymbol{V}$ among three factor matrices, since factor $\boldsymbol{U}$ can be obtained from the patient's clinical information.

To solve our minimization problem, we adapt the multiplicative update method for NTMF proposed in a recent study[55], which contains a routine for avoiding '*inadmissible zeros problem'* where the solution of multiplicative update rules is stuck at zero when an entry in the factor becomes.

Step 1: Initialization

Initialize the factor matrices $\boldsymbol{S} = \boldsymbol{1}$ and $\boldsymbol{V} = \boldsymbol{V_0}$, where $\boldsymbol{1} \in R^{K_1\times K_2}$ is a matrix whose elements are all one. Set the regularization parameters $\lambda_S = \lambda_V = \lambda_L = 1$ and λ_0=0.1. Set the user specified parameters for avoiding the inadmissible zeros problem, $\kappa_{tol} = 10^{-10}$, $\kappa = 10^{-6}$ and $\epsilon = 10^{-10}$.

Step 2: Iteration

Iterate until it converges or reaches the maximum number of iterations:

$$[\boldsymbol{S}]_{ij} \leftarrow ([\boldsymbol{S}]_{ij} + \kappa_{ij}^S)\tau_{ij}^S$$

$$[\boldsymbol{V}]_{ij} \leftarrow ([\boldsymbol{V}]_{ij} + \kappa_{ij}^V)\tau_{ij}^S$$

where κ_{ij}^M is set to κ if $[\boldsymbol{M}]_{ij} \geq \kappa_{tol}$ *and* $\tau_{ij}^M > 1$, otherwise 0, and

$$\tau_{ij}^S = \frac{[\boldsymbol{U^T XV}]_{ij}}{[\boldsymbol{U^T}\left(\boldsymbol{W} \circ (\boldsymbol{USV^T})\right)\boldsymbol{V}]_{ij} + \lambda_S \| \boldsymbol{S} \|_1 + \epsilon},$$

$$\tau_{ij}^V = \frac{[\boldsymbol{X^T US} + \lambda_L \boldsymbol{AV} + \lambda_0 \boldsymbol{V_0}]_{ij}}{[(\boldsymbol{W} \circ (\boldsymbol{USV^T}))^T \boldsymbol{US} + \lambda_0 \boldsymbol{V} + \lambda_L \boldsymbol{DV}]_{ij} + \lambda_V \| \boldsymbol{V} \|_1 + \epsilon}.$$

Empirically, the algorithm converges fast within 50 iterations in the experiments.

3. **Identification of cancer-type-specific altered pathways**

Once the above optimization problem is solved, we used $\boldsymbol{S}$ matrix to identify cancer-type-specific altered pathways across cancers. Specifically, we ranked pathways based on values of elements of the $\boldsymbol{S}$ matrix for each cancer type (e.g., rank pathways based on all values of ith row which indicate association scores between ith cancer type and all pathways). In addition, to measure statistical significance of cancer-type and pathway associations, we performed a permutation test (e.g., we randomly permuted somatic mutation data and repeated experiments 5,000 times to calculate empirical p-values) and defined cancer-type-specific altered pathways based on the following strict criteria: 1) Pathways must be ranked within the top $\boldsymbol{K}$th compared to other pathways in each cancer type based on their association scores in matrix $\boldsymbol{S}$. 2) Pathways must have significant BH-adjusted p-values (Benjamini-Hochberg adjusted p-values using a false discovery rate cutoff of 0.1) (See Supplementary X). In this work, we selected the top 3 ranked pathways having significant BH- adjusted p-values per each cancer type. Top ranked pathways for KICH, KIRP, and THCA were excluded for further analysis, due to the insignificant BH-adjusted p-values.

Gene expression data and clustering

We collected RNA-seq data for TCGA BLCA, BRCA, HNSC, KIRC, LAML, LUAD, SKCM, STAD, UCEC from cBioPortal[56-58] using CGDS MATLAB toolbox with RNA Seq V2 RSEM option. We collected microarray gene expression profiles for TCGA GBM from the TCGA dataportal[51] and TCGA OV and two others from Zhang, W. *et al*. [59]. We collected colon cancer data from GSE39582 and lung cancer data from Shedden, K. *et al.* [60]. RNA-seq data were z-transformed while other expression data were quantile normalized, log transformed, and expression values

were median centered. To perform consensus clustering, we used Matlab K-means clustering and used two-way hierarchical clustering.

Authors' contributions

SP and THH designed the study. SP and THH performed the analysis. THH supervised the project. SP, JN, SL, and THH wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank to Quantitative Biomedical Research Center to allow us to use their computational resources.

Figures

Figure 1. Overview of NTriPath

This figure describes steps to discover altered pathways across multiple cancer types. The aim of the approach is to integrate the somatic mutation data with gene-gene interaction networks and a pathway database for discovering altered pathways across cancers.

Figure 2. KIRC-specific altered pathways

(A) Diagrams of the top three ranked altered pathways in patients with KIRC. Red color indicates genes that are frequently mutated. A circular-shaped node represents the original member genes annotated in the pathway database, and a diamond-shaped node represents newly identified members genes of the pathways by NTriPath (B) Protein and mRNA expression and mutation status for all genes identified in the top three KIRC altered pathways. Each row represents a member gene in the TCGA

KIRC-specific altered pathway, and each column represents a patient sample in TCGA KIRC cohort.

Figure 3. Cancer-type-specific altered pathways across cancers correlate with survival outcomes

Kaplan-Meier survival plots based on patient subgroups defined by consensus clustering using genes from the top 3 altered pathways for (a) kidney renal cell carcinoma (KIRC), (b) bladder urothelial carcinoma (BLCA), (c) head and neck squamous carcinoma (HNSC), and (d) skin cutaneous melanoma (SKCM).

Figure 4. Comparing NTriPath-derived signatures with mutation-frequency-based signatures

This figure describes comparisons of patient stratification using signatures derived from NTriPath and mutation frequency reported in Kandoth, C. *et al.* [1].

Figure 5. Validation in independent cohort

This figure describes Kaplan-Meier survival plots for patient subgroups from (a) UTSW HNSC (b) Lung adenocarcinoma (c) Colon cancer, (d) Ovarian cancer.

	Cancer Type	Number of Patients
1	Acute Myeloid Leukemia (LAML)	75
2	Bladder Urothelial Carcinoma (BLCA)	136
3	Brain Lower Grade Glioma (LGG),	217
4	Breast Invasive Carcinoma (BRCA)	772
5	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC)	41
6	Colon Adenocarcinoma (COAD)	270
7	Glioblastoma Multiforme (GBM)	290
8	Head and Neck Squamous Cell Carcinoma (HNSC)	323

9	Kidney Chromophobe Renal Cell Carcinoma (KICH)	65
10	Kidney Renal Clear Cell Carcinoma (KIRC)	210
11	Kidney Renal Papillary Cell Carcinoma (KIRP)	111
12	Lung Adenocarcinoma (LUAD)	380
13	Ovarian Serous Cystadenocarcinoma (OV)	463
14	Prostate Adenocarcinoma (PRAD)	171
15	Rectum Adenocarcinoma (READ)	116
16	Skin Cutaneous Melanoma (SKCM)	269
17	Stomach Adenocarcinoma (STAD)	264
18	Thyroid Carcinoma (THCA)	369
19	Uterine Corpus Endometrioid Carcinoma	248

Table 1. A list of cancer types used in the analysis.

References

1. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333-339 (2013).
2. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* 3 (2013).
3. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-218 (2013).
4. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495-501 (2014).
5. Osmanbeyoglu, H.U., Pelossof, R., Bromberg, J.F. & Leslie, C.S. Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res* 24, 1869-1880 (2014).
6. Baselga, J. Targeting the phosphoinositide-3 (PI3) kinase pathway in breast cancer. *The oncologist* 16 Suppl 1, 12-19 (2011).
7. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Meth* 10, 1108-1115 (2013).
8. Hwang, T. *et al.* Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers. *BMC Genomics* 14, 1-13 (2013).
9. Vaske, C.J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26, i237-i245 (2010).
10. Cerami, E., Demir, E., Schultz, N., Taylor, B.S. & Sander, C. Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE* 5, e8918 (2010).

11. Vandin, F., Upfal, E. & Raphael, B. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology* 18, 507-522 (2011).
12. Suthram, S. *et al.* Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Comput Biol* 6, e1000662 (2010).
13. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature methods* 10, 1108-1115 (2013).
14. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* 43, D805-811 (2015).
15. Pena-Llopis, S. *et al.* BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet* 44, 751-759 (2012).
16. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* 45, 860-867 (2013).
17. Peña-Llopis, S. & Brugarolas, J. Simultaneous isolation of high-quality DNA, RNA, miRNA and proteins from tissues for genomic applications. *Nat. Protocols* 8, 2240-2255 (2013).
18. Nickerson, M.L. *et al.* Improved Identification of von Hippel-Lindau Gene Alterations in Clear Cell Renal Tumors. *Clinical Cancer Research* 14, 4726-4734 (2008).
19. The Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43-49 (2013).
20. Gialeli, C., Theocharis, A.D. & Karamanos, N.K. Roles of matrix metalloproteinases in cancer progression and their pharmacological targeting. *The FEBS journal* 278, 16-27 (2011).
21. Staudt, N.D. *et al.* Myeloid Cell Receptor LRP1/CD91 Regulates Monocyte Recruitment and Angiogenesis in Tumors. *Cancer Research* 73, 3902-3912 (2013).
22. Langlois, B. *et al.* LRP-1 Promotes Cancer Cell Invasion by Supporting ERK and Inhibiting JNK Signaling Pathways. *PLoS ONE* 5, e11584 (2010).
23. Song, H., Li, Y., Lee, J., Schwartz, A.L. & Bu, G. Low-density lipoprotein receptor-related protein 1 promotes cancer cell migration and invasion by inducing the expression of matrix metalloproteinases 2 and 9. *Cancer Res* 69, 879-886 (2009).
24. Duan, D. *et al.* Inhibition of transcription elongation by the VHL tumor suppressor protein. *Science* 269, 1402-1406 (1995).
25. Ohh, M. *et al.* Ubiquitination of hypoxia-inducible factor requires direct binding to the [bgr]-domain of the von Hippel-Lindau protein. *Nat Cell Biol* 2, 423-427 (2000).
26. Tanimoto, K., Makino, Y., Pereira, T. & Poellinger, L. Mechanism of regulation of the hypoxia-inducible factor-1a by the von Hippel-Lindau tumor suppressor protein. *The EMBO Journal* 19, 4298-4309 (2000).
27. Yamamoto, D., Takai, S. & Miyazaki, M. Inhibitory profiles of captopril on matrix metalloproteinase-9 activity. *European journal of pharmacology* 588, 277-279 (2008).
28. Williams, R.N., Parsons, S.L., Morris, T.M., Rowlands, B.J. & Watson, S.A. Inhibition of matrix metalloproteinase activity and growth of gastric adenocarcinoma cells by an angiotensin converting enzyme inhibitor in in vitro and murine models. *European journal of surgical oncology : the journal*

of the European Society of Surgical Oncology and the British Association of Surgical Oncology 31, 1042-1050 (2005).

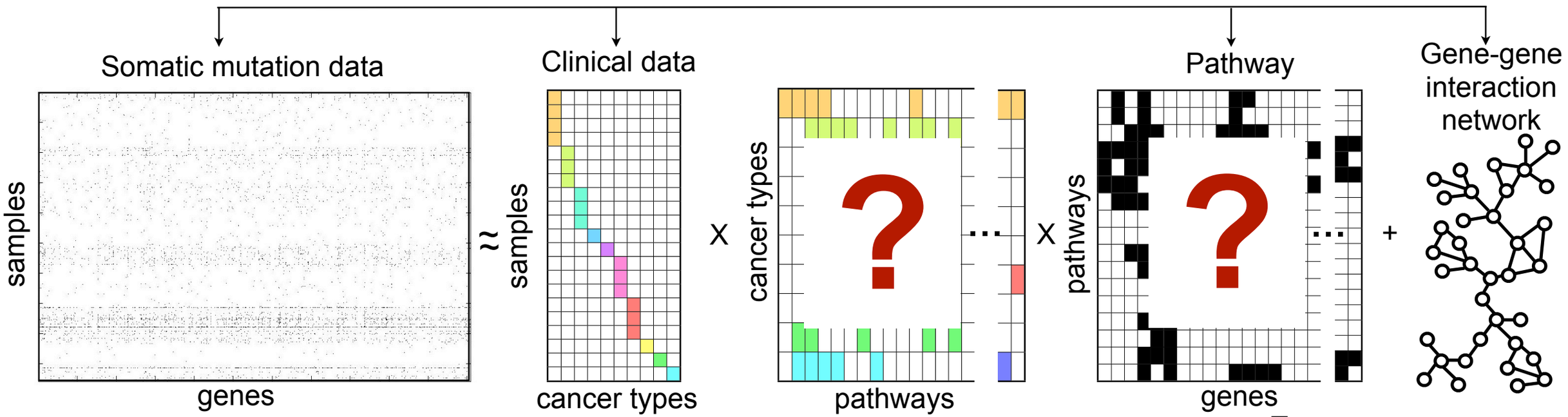
29. Underwood, C.K., Min, D., Lyons, J.G. & Hambley, T.W. The interaction of metal ions and Marimastat with matrix metalloproteinase 9. *Journal of inorganic biochemistry* 95, 165-170 (2003).
30. Reinhardt, D. *et al.* Cardiac remodelling in end stage heart failure: upregulation of matrix metalloproteinase (MMP) irrespective of the underlying disease, and evidence for a direct inhibitory effect of ACE inhibitors on MMP. *Heart* 88, 525-530 (2002).
31. Rask-Andersen, M., Almen, M.S. & Schioth, H.B. Trends in the exploitation of novel drug targets. *Nature reviews. Drug discovery* 10, 579-590 (2011).
32. Rajapakse, N., Mendis, E., Kim, M.M. & Kim, S.K. Sulfated glucosamine inhibits MMP-2 and MMP-9 expressions in human fibrosarcoma cells. *Bioorganic & medicinal chemistry* 15, 4891-4896 (2007).
33. Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. How many drug targets are there? *Nature reviews. Drug discovery* 5, 993-996 (2006).
34. Okada, M. *et al.* Captopril attenuates matrix metalloproteinase-2 and -9 in monocrotaline-induced right ventricular hypertrophy in rats. *Journal of pharmacological sciences* 108, 487-494 (2008).
35. Nenan, S. *et al.* Metalloelastase (MMP-12) induced inflammatory response in mice airways: effects of dexamethasone, rolipram and marimastat. *European journal of pharmacology* 559, 75-81 (2007).
36. Mendis, E., Kim, M.M., Rajapakse, N. & Kim, S.K. Carboxy derivatized glucosamine is a potent inhibitor of matrix metalloproteinase-9 in HT1080 cells. *Bioorganic & medicinal chemistry letters* 16, 3105-3110 (2006).
37. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nature reviews. Drug discovery* 5, 821-834 (2006).
38. Fenton, J.I., Chlebek-Brown, K.A., Caron, J.P. & Orth, M.W. Effect of glucosamine on interleukin-1-conditioned articular cartilage. *Equine veterinary journal. Supplement*, 219-223 (2002).
39. Dodge, G.R. & Jimenez, S.A. Glucosamine sulfate modulates the levels of aggrecan and matrix metalloproteinase-3 synthesized by cultured human osteoarthritis articular chondrocytes. *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society* 11, 424-432 (2003).
40. Chu, S.C. *et al.* Glucosamine sulfate suppresses the expressions of urokinase plasminogen activator and inhibitor and gelatinases during the early stage of osteoarthritis. *Clinica chimica acta; international journal of clinical chemistry* 372, 167-172 (2006).
41. Chen, X., Ji, Z.L. & Chen, Y.Z. TTD: Therapeutic Target Database. *Nucleic acids research* 30, 412-415 (2002).
42. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic acids research* 28, 235-242 (2000).
43. Sato, A. *et al.* Inhibition of MMP-9 using a pyrrole-imidazole polyamide reduces cell invasion in renal cell carcinoma. *International journal of oncology* 43, 1441-1446 (2013).
44. Hu, K. *et al.* Tissue-type plasminogen activator acts as a cytokine that triggers intracellular signal transduction and induces matrix metalloproteinase-9 gene expression. *The Journal of biological chemistry* 281, 2120-2127 (2006).

45. Gilbert, J. *et al.* Phase II trial of irinotecan plus cisplatin in patients with recurrent or metastatic squamous carcinoma of the head and neck. *Cancer* 113, 186-192 (2008).
46. Garraway, Levi A. & Lander, Eric S. Lessons from the Cancer Genome. *Cell* 153, 17-37 (2013).
47. Dees, N.D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Research* 22, 1589-1598 (2012).
48. Supek, F., Minana, B., Valcarcel, J., Gabaldon, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156, 1324-1335 (2014).
49. Keshava Prasad, T.S. *et al.* Human Protein Reference Database--2009 update. *Nucleic acids research* 37, D767-772 (2009).
50. Rossin, E.J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet* 7, e1001273 (2011).
51. strel'tsov, S.A., Mikheikin, A.L. & Nechipurenko Iu, D. [Interaction of topotecan--a DNA topoisomerase I inhibitor--with dual-stranded polydeoxyribonucleotides. II. Formation of a complex containing several DNA molecules in the presence of topotecan]. *Molekuliarnaia biologiia* 35, 442-450 (2001).
52. Zhang, S., Li, Q., Liu, J. & Zhou, X.J. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* 27, i401-i409 (2011).
53. Keshava Prasad, T.S. *et al.* Human Protein Reference Database—2009 update. *Nucleic acids research* 37, D767-D772 (2009).
54. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102, 15545-15550 (2005).
55. Seung-Jun, K., TaeHyun, H. & Giannakis, G.B. in Cognitive Information Processing (CIP), 2012 3rd International Workshop on 1-6 (2012).
56. Teicher, B.A. Next generation topoisomerase I inhibitors: Rationale and biomarker strategies. *Biochemical pharmacology* 75, 1262-1271 (2008).
57. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* 2, 401-404 (2012).
58. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* 6, pl1- (2013).
59. Zhang, W. *et al.* Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment. *PLoS Comput Biol* 9, e1002975 (2013).
60. Shedden, K. *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 14, 822-827 (2008).

Fig. 1

Input:

1) Somatic mutation
2) Clinical data (e.g., patient's cancer type)
3) Pathway information
4) Gene-gene interaction networks



$$\min_{S,V} \; ||X-USV^T||_2^F + ||S||_1^2 + ||V-V_0||_2^F + \mathrm{tr}(V^T(D-A)V)$$

X: Somatic mutation data (#patients X #genes)
U: Patient cluster (#patients X #cancer types)
S: Cancer type-pathway association matrix (#cancer types X #pathways)
$\mathbf{V_0}$: Initial pathway information (#pathways X #genes)
V: Newly updated pathway information (#pathways X #genes)
A: Gene-gene interaction network (#genes X #genes)

Output:

1) Newly updated pathway information (V)
2) Cancer type-pathway association (S)

Newly updated member genes in the pathways

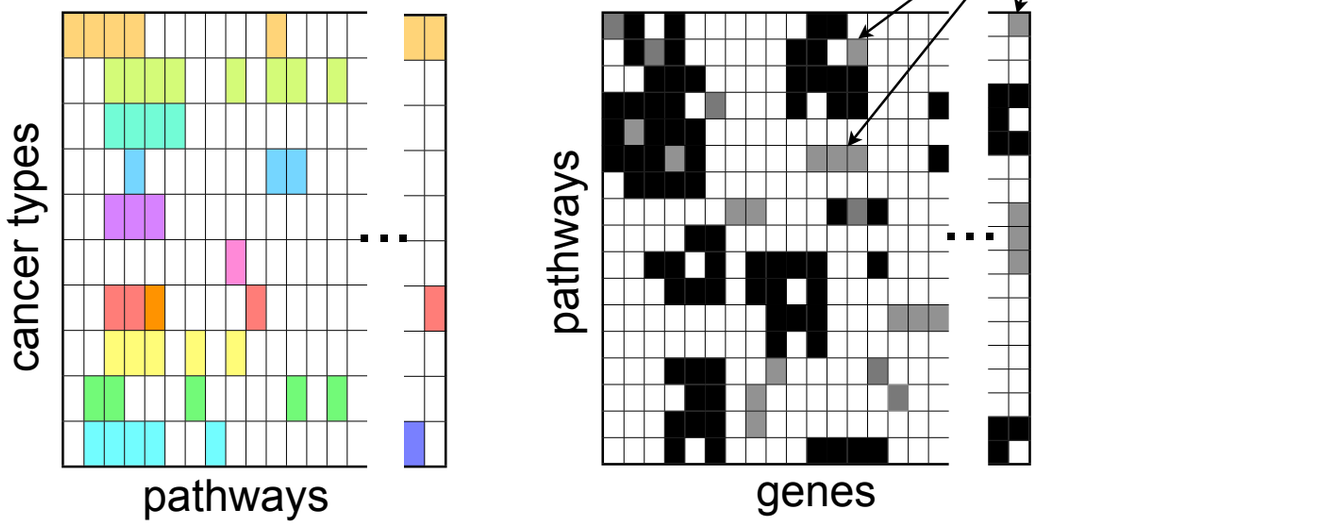
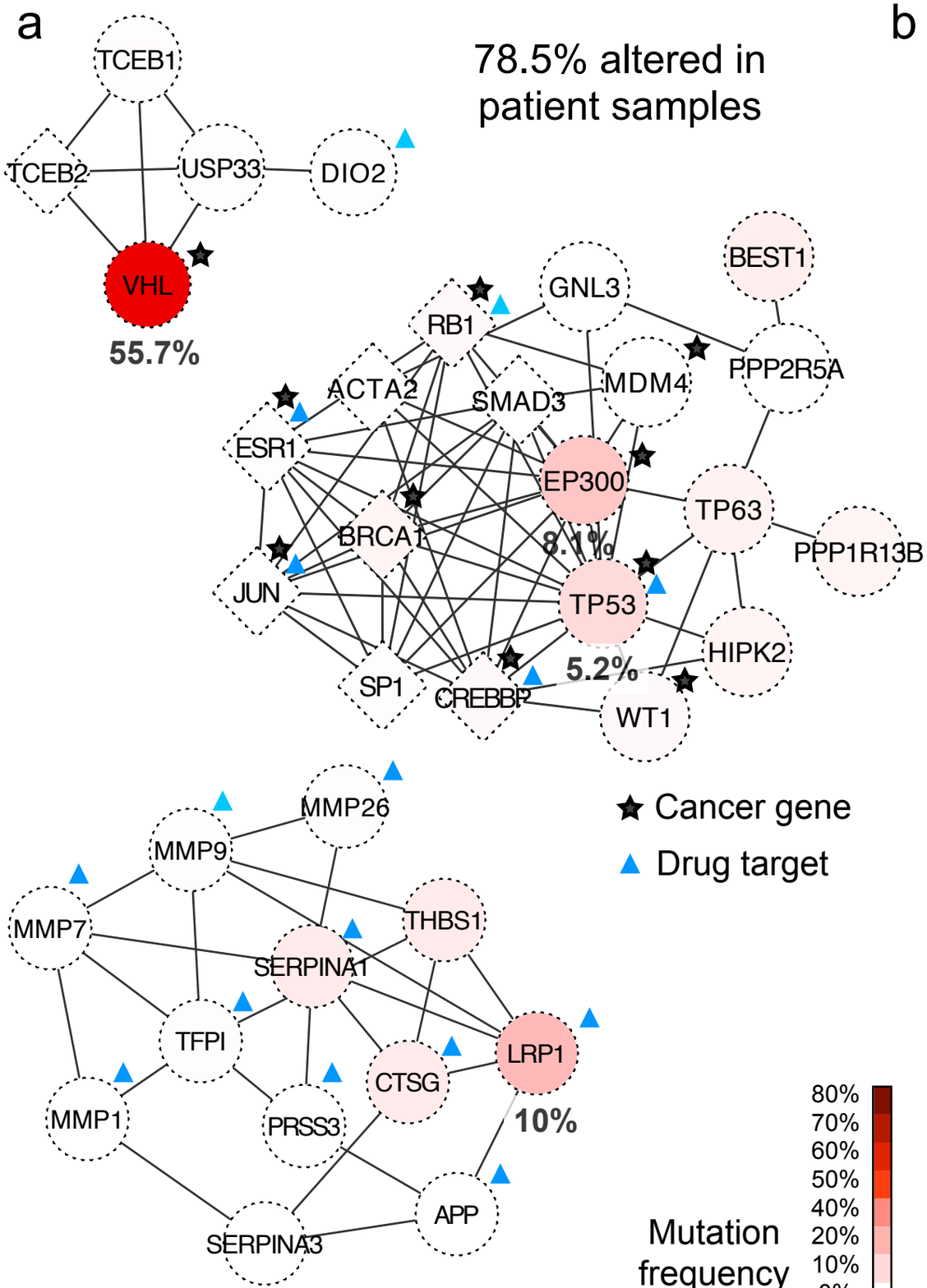


Fig. 2


a
78.5% altered in patient samples
TCEB1
TCEB2
USP33
DIO2
VHL
55.7%
BEST1
GNL3
RB1
PPP2R5A
MDM4
ACTA2
SMAD3
ESR1
EP300
TP63
PPP1R13B
BRCA1
8.1%
JUN
TP53
HIPK2
5.2%
SP1
CREBBP
WT1
MMP26
MMP9
MMP7
THBS1
SERPINA1
TFPI
LRP1
CTSG
10%
MMP1
PRSS3
APP
SERPINA3
Cancer gene
Drug target
Mutation frequency
80%
70%
60%
50%
40%
20%
10%
0%

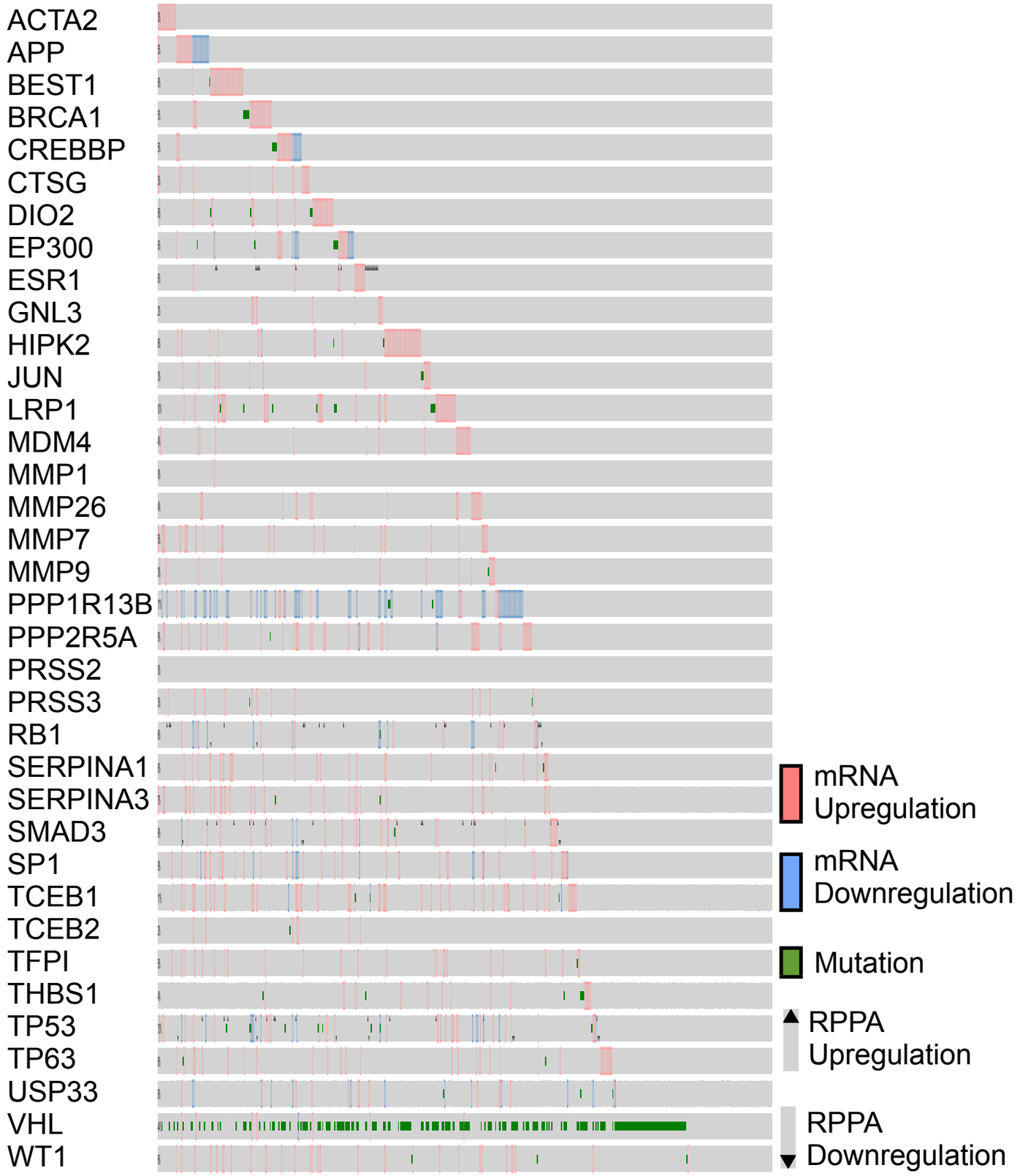
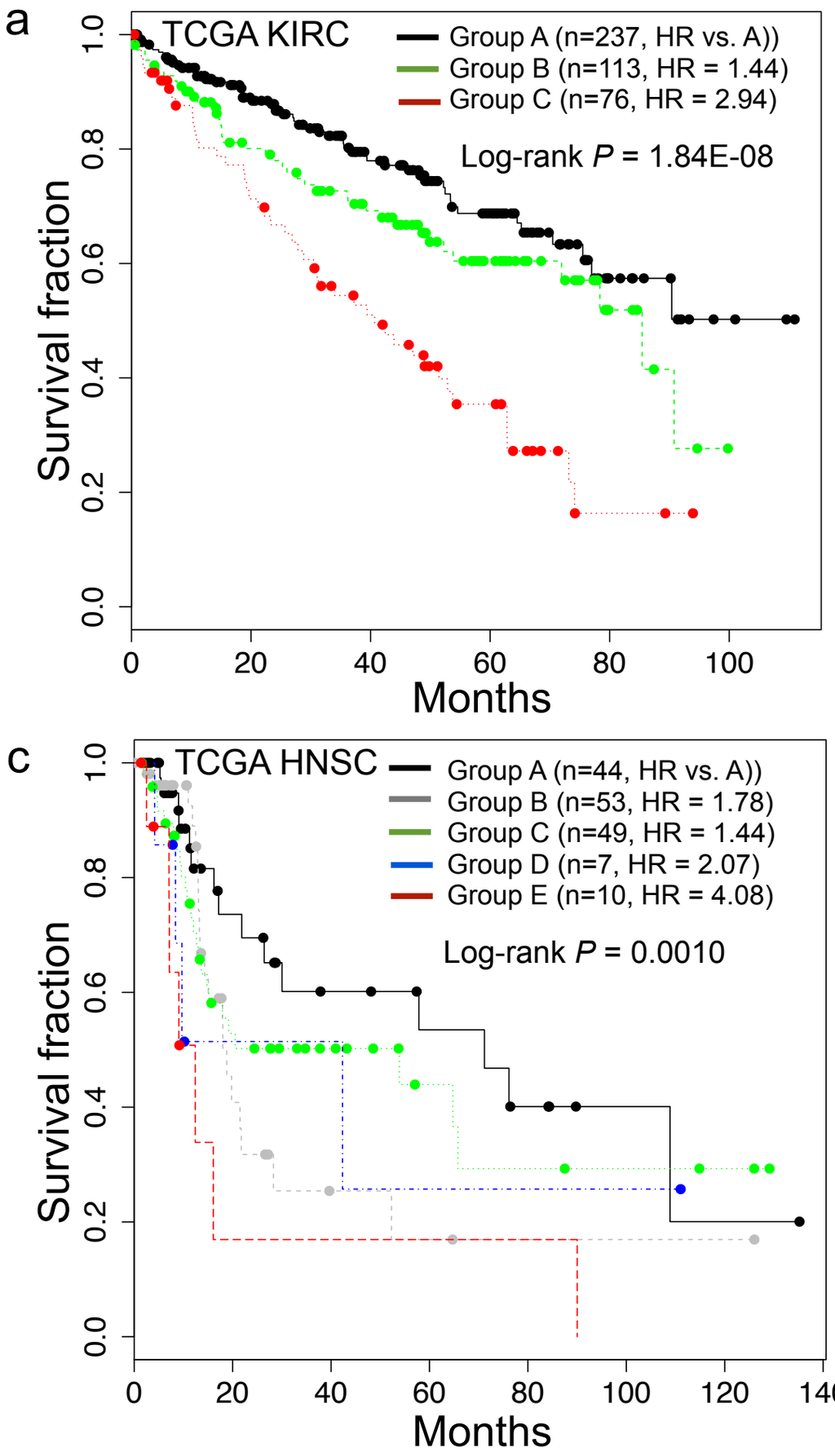
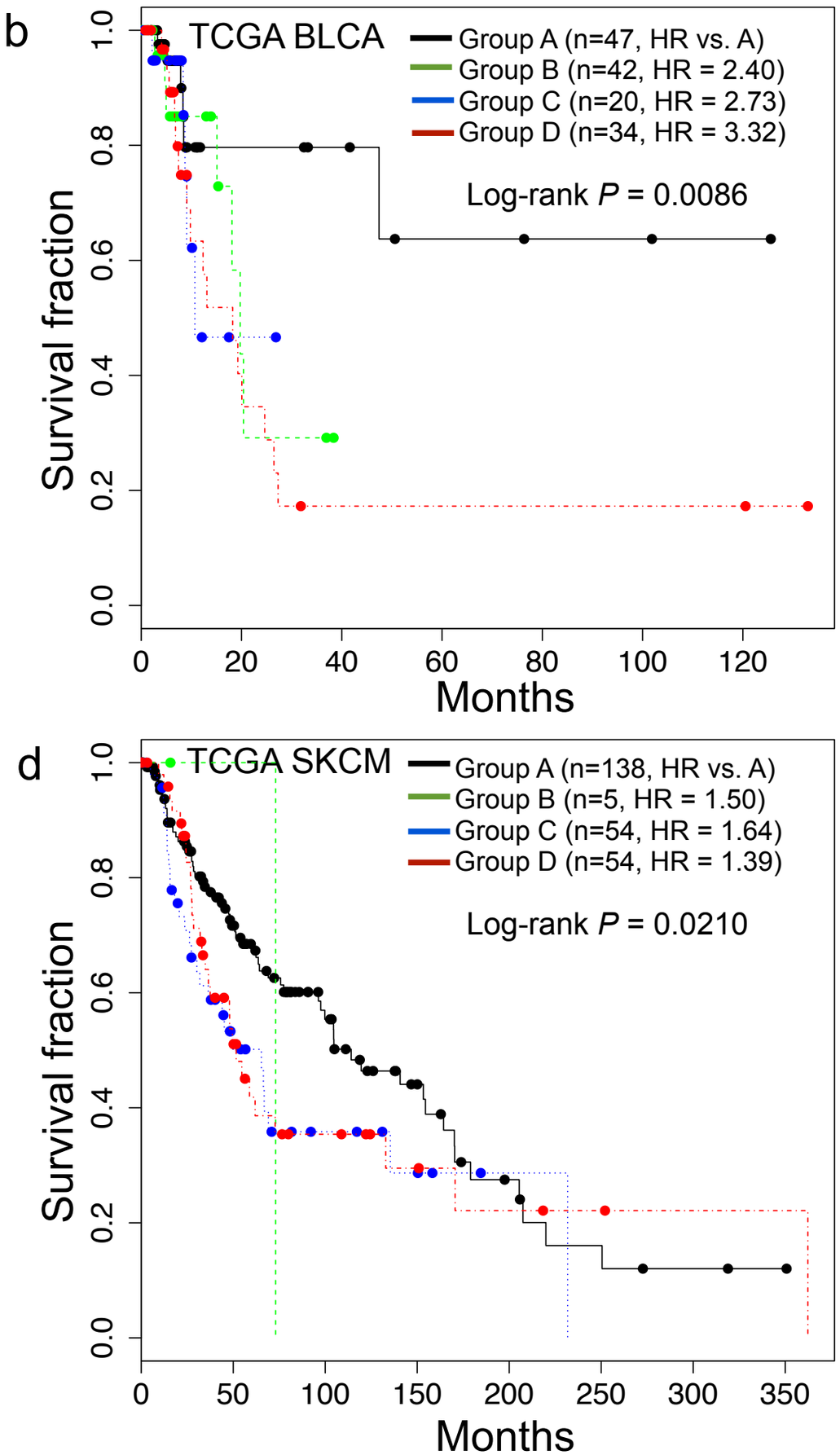

b
ACTA2
APP
BEST1
BRCA1
CREBBP
CTSG
DIO2
EP300
ESR1
GNL3
HIPK2
JUN
LRP1
MDM4
MMP1
MMP26
MMP7
MMP9
PPP1R13B
PPP2R5A
PRSS2
PRSS3
RB1
SERPINA1
SERPINA3
SMAD3
SP1
TCEB1
TCEB2
TFPI
THBS1
TP53
TP63
USP33
VHL
WT1
mRNA Upregulation
mRNA Downregulation
Mutation
RPPA Upregulation
RPPA Downregulation

Fig. 3

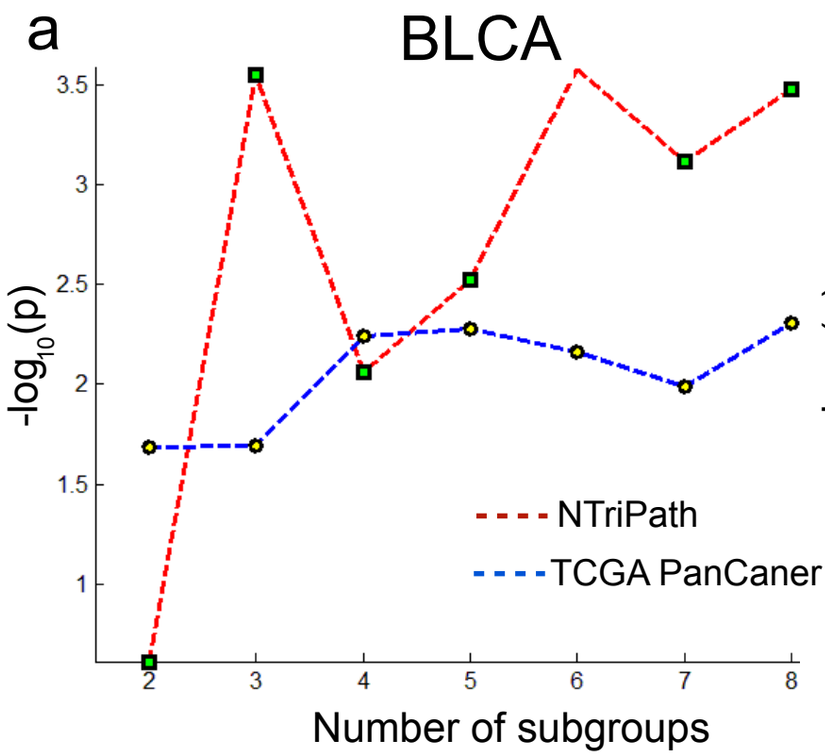

a
TCGA KIRC
Group A (n=237, HR vs. A))
Group B (n=113, HR = 1.44)
Group C (n=76, HR = 2.94)
Log-rank P = 1.84E-08
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Months

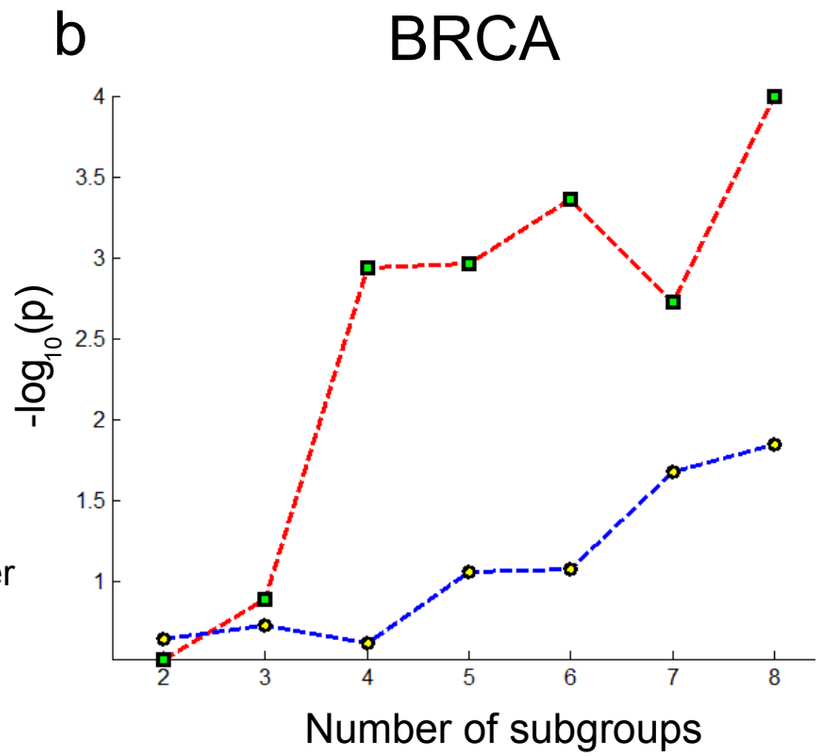

b
TCGA BLCA
Group A (n=47, HR vs. A)
Group B (n=42, HR = 2.40)
Group C (n=20, HR = 2.73)
Group D (n=34, HR = 3.32)
Log-rank P = 0.0086
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Months

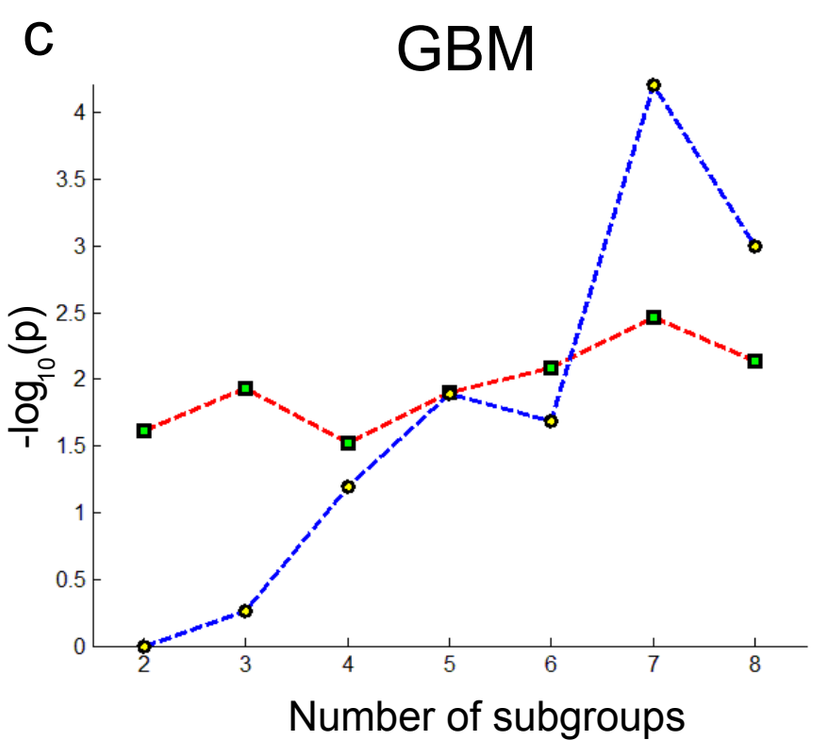
c
TCGA HNSC
Group A (n=44, HR vs. A))
Group B (n=53, HR = 1.78)
Group C (n=49, HR = 1.44)
Group D (n=7, HR = 2.07)
Group E (n=10, HR = 4.08)
Log-rank P = 0.0010
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
140
Months

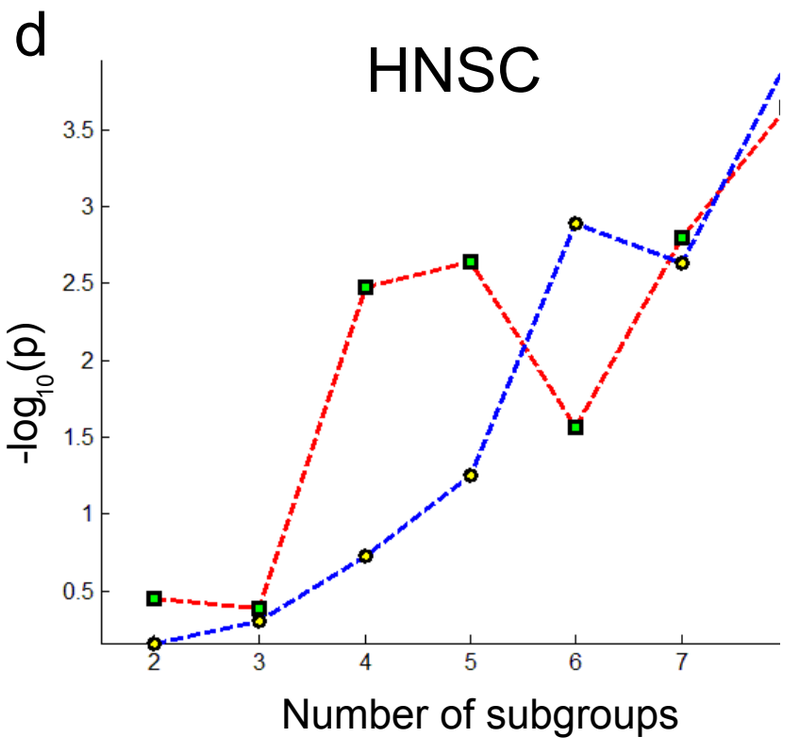
d
TCGA SKCM
Group A (n=138, HR vs. A)
Group B (n=5, HR = 1.50)
Group C (n=54, HR = 1.64)
Group D (n=54, HR = 1.39)
Log-rank P = 0.0210
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Months

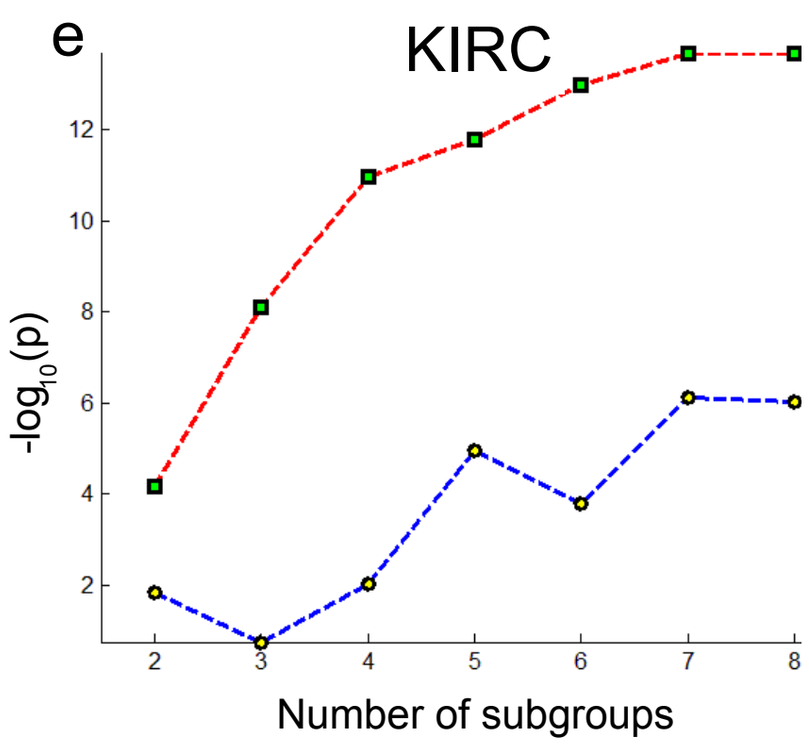
Fig. 4


a
BLCA
-log10(p)
Number of subgroups
NTriPath
TCGA PanCaner


b
BRCA
-log10(p)
Number of subgroups


c
GBM
-log10(p)
Number of subgroups


d
HNSC
-log10(p)
Number of subgroups


e
KIRC
-log10(p)
Number of subgroups

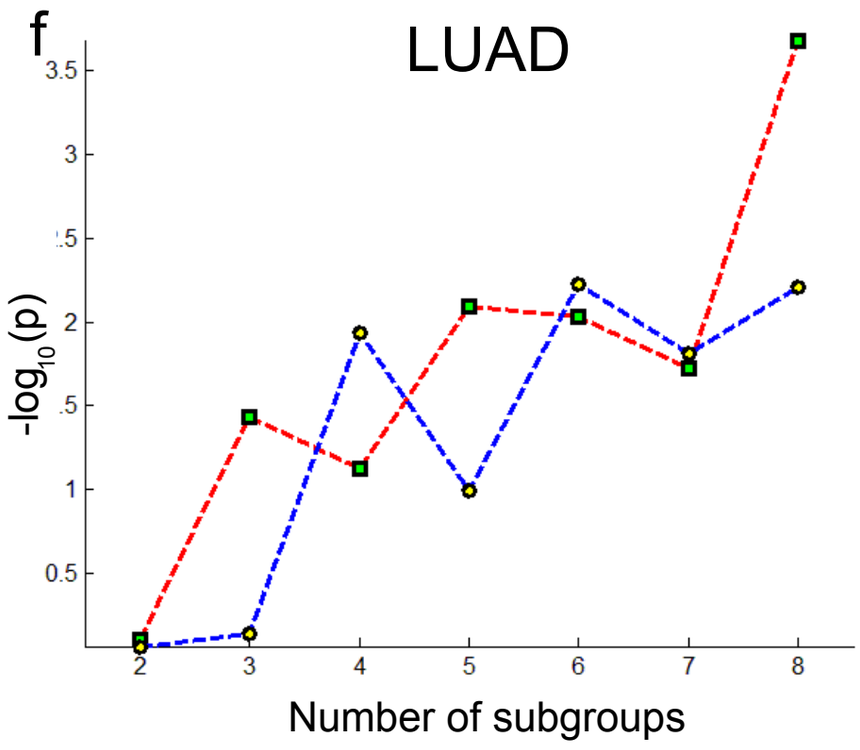
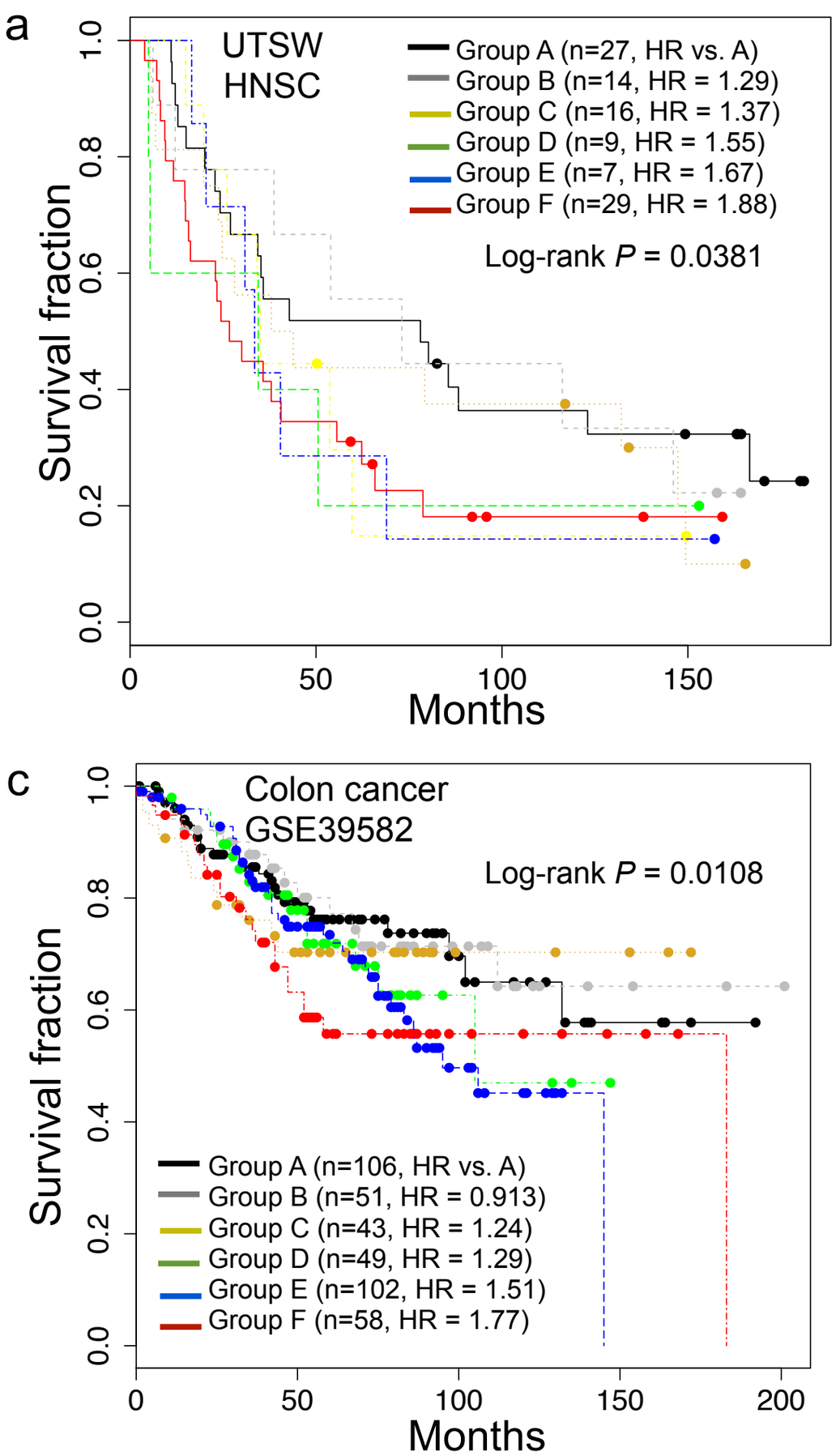
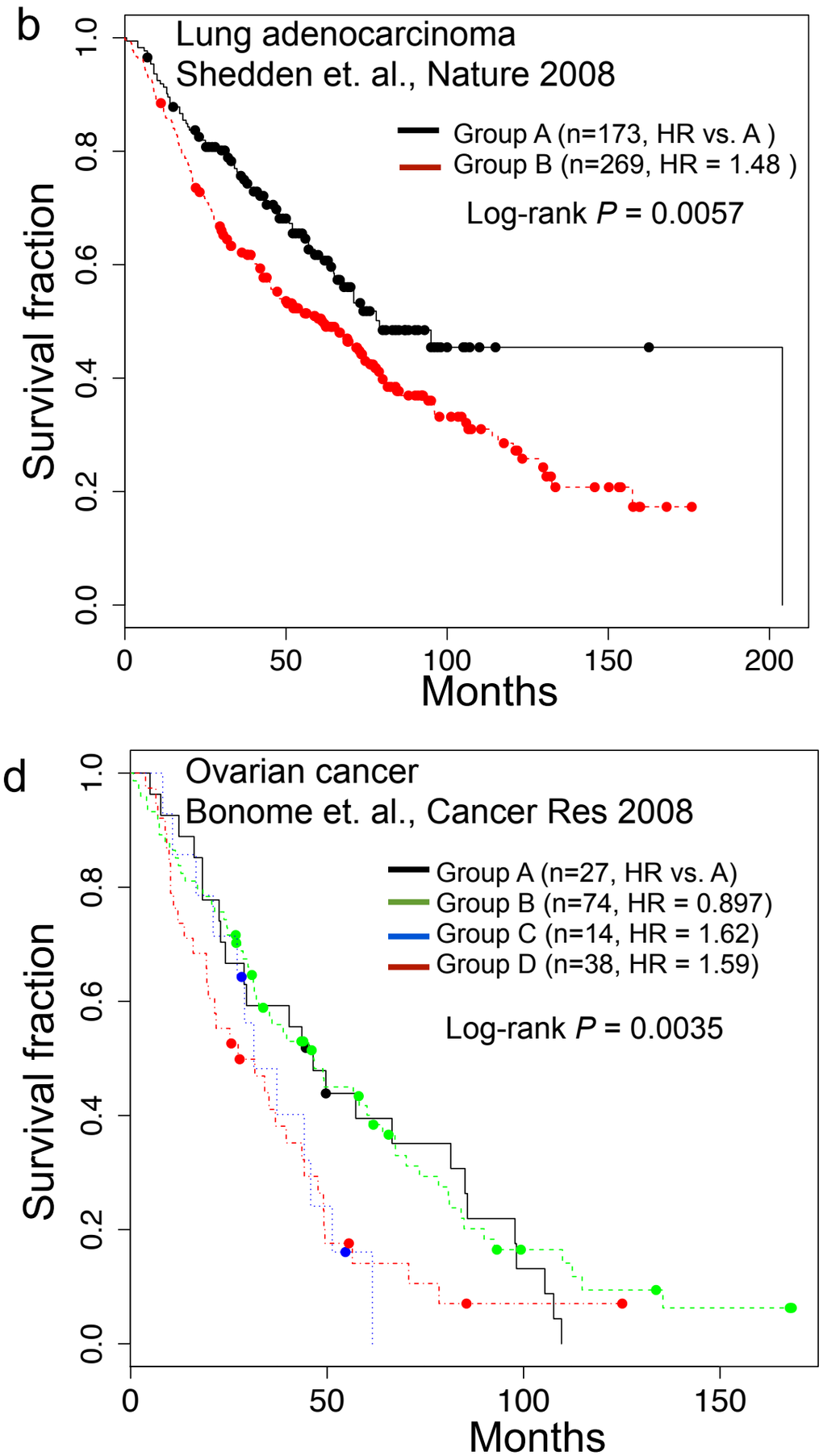

f
LUAD
-log10(p)
Number of subgroups

Fig. 5


a
UTSW
HNSC
Group A (n=27, HR vs. A)
Group B (n=14, HR = 1.29)
Group C (n=16, HR = 1.37)
Group D (n=9, HR = 1.55)
Group E (n=7, HR = 1.67)
Group F (n=29, HR = 1.88)
Log-rank P = 0.0381
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Months

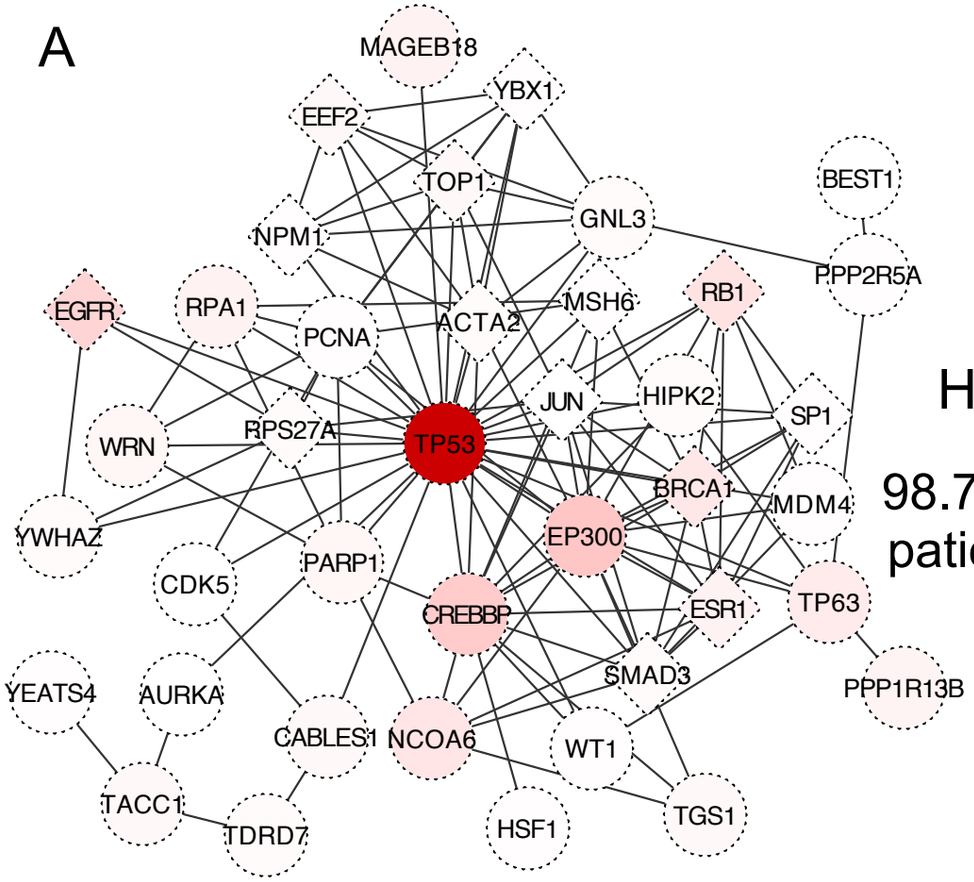

b
Lung adenocarcinoma
Shedden et. al., Nature 2008
Group A (n=173, HR vs. A)
Group B (n=269, HR = 1.48)
Log-rank P = 0.0057
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Months

c
Colon cancer
GSE39582
Log-rank P = 0.0108
Group A (n=106, HR vs. A)
Group B (n=51, HR = 0.913)
Group C (n=43, HR = 1.24)
Group D (n=49, HR = 1.29)
Group E (n=102, HR = 1.51)
Group F (n=58, HR = 1.77)
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Months

d
Ovarian cancer
Bonome et. al., Cancer Res 2008
Group A (n=27, HR vs. A)
Group B (n=74, HR = 0.897)
Group C (n=14, HR = 1.62)
Group D (n=38, HR = 1.59)
Log-rank P = 0.0035
Survival fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Months

Supplementary Figure 1

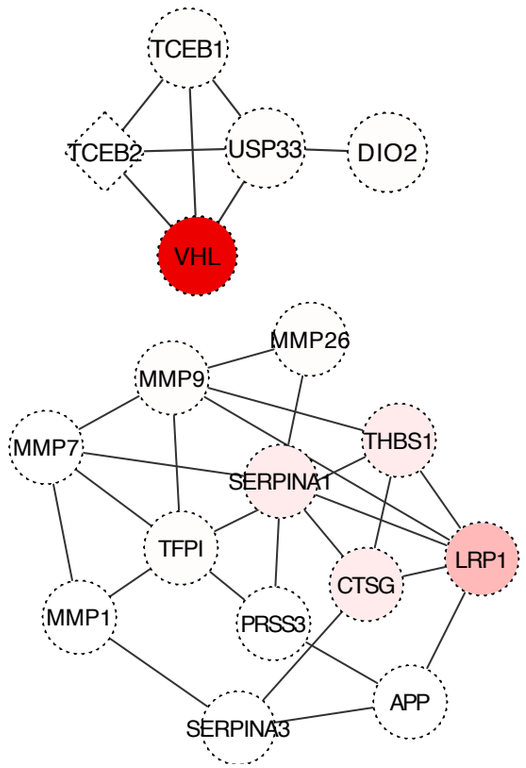
A

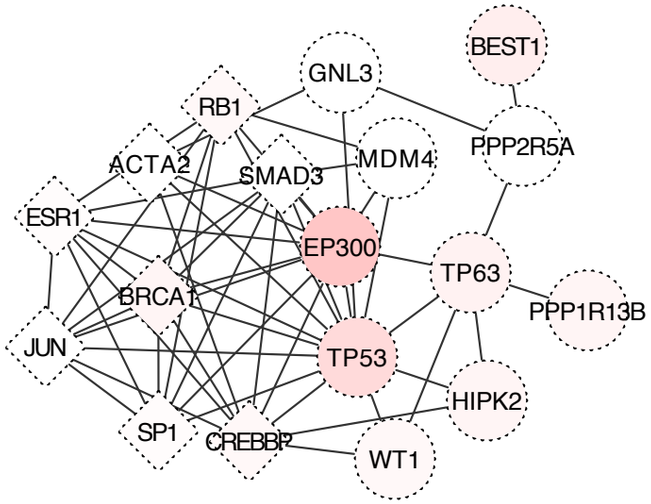


HNSC

98.7% altered in patient samples

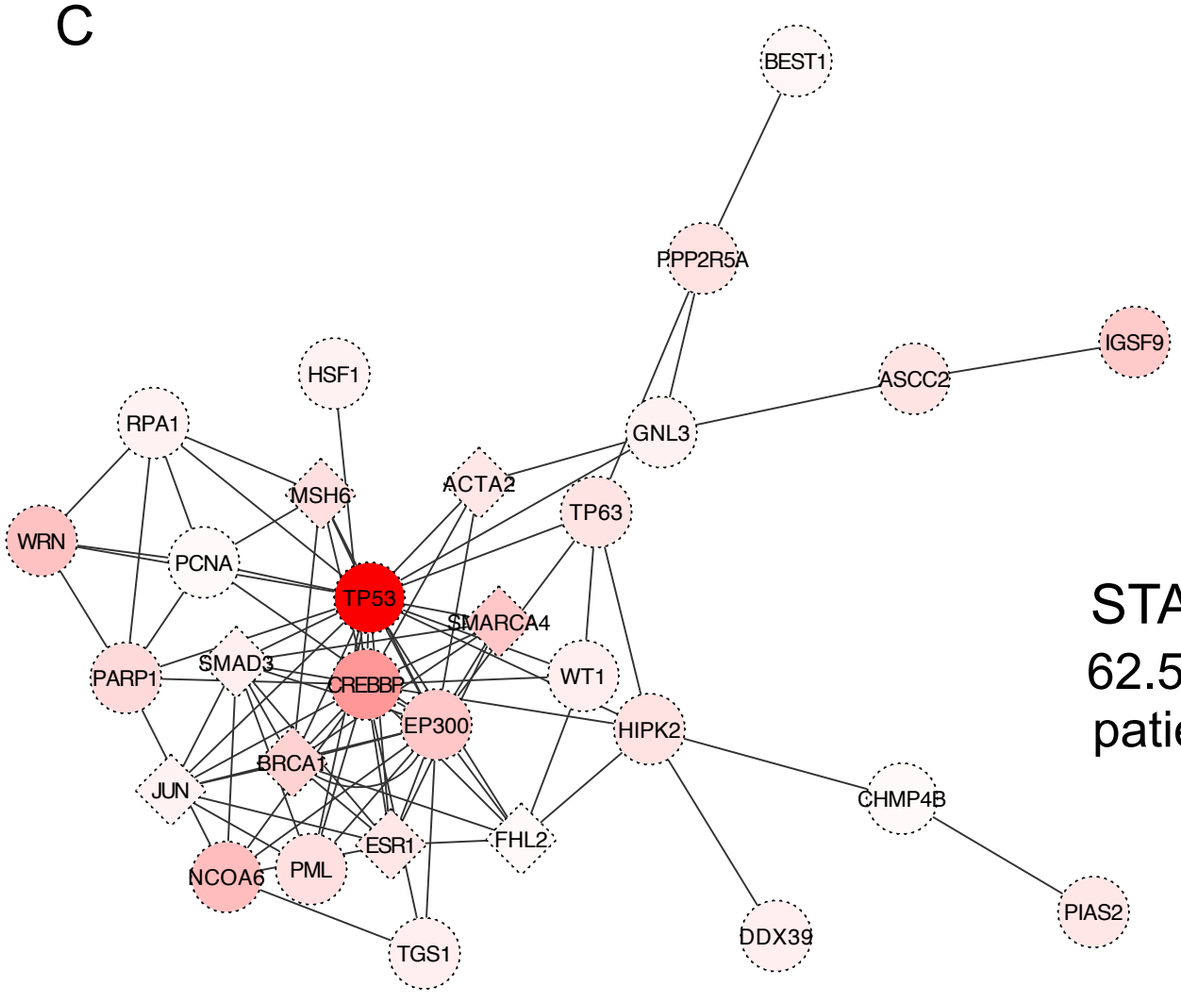
B





KIRC 78.5% altered in patient samples

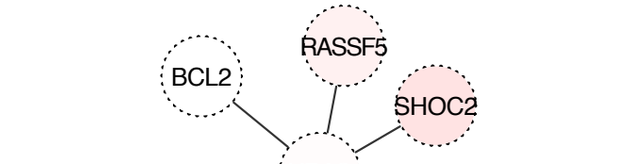
C



STAD

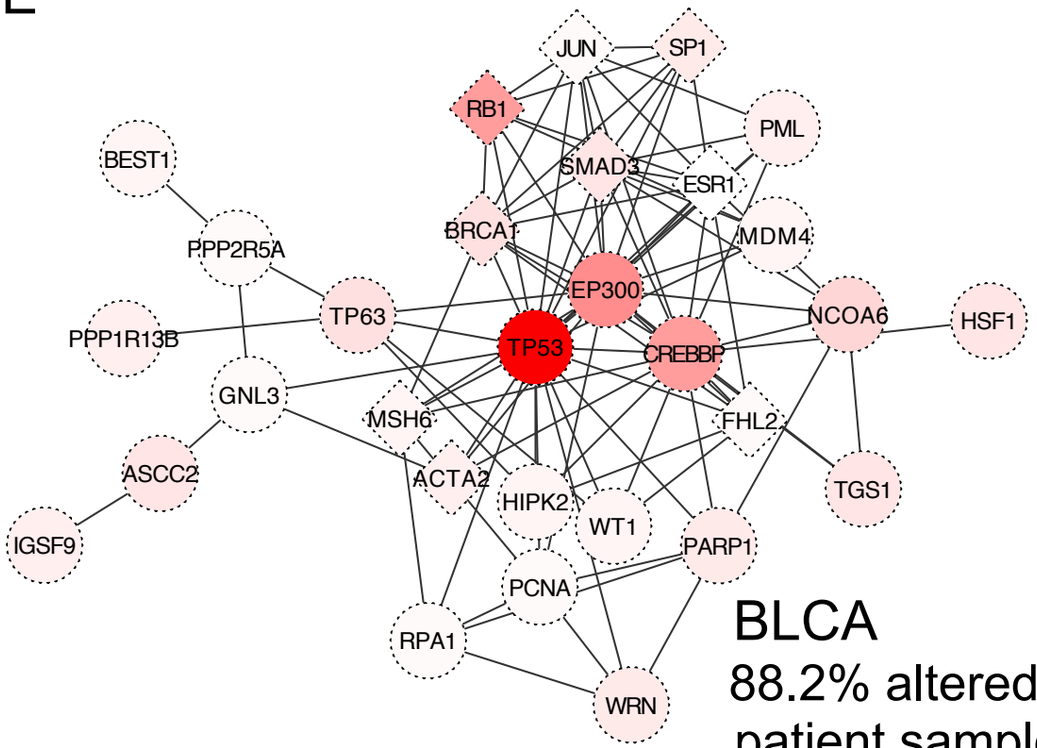
62.5% altered in patient samples

D



UCEC 99.2% altered in patient samples

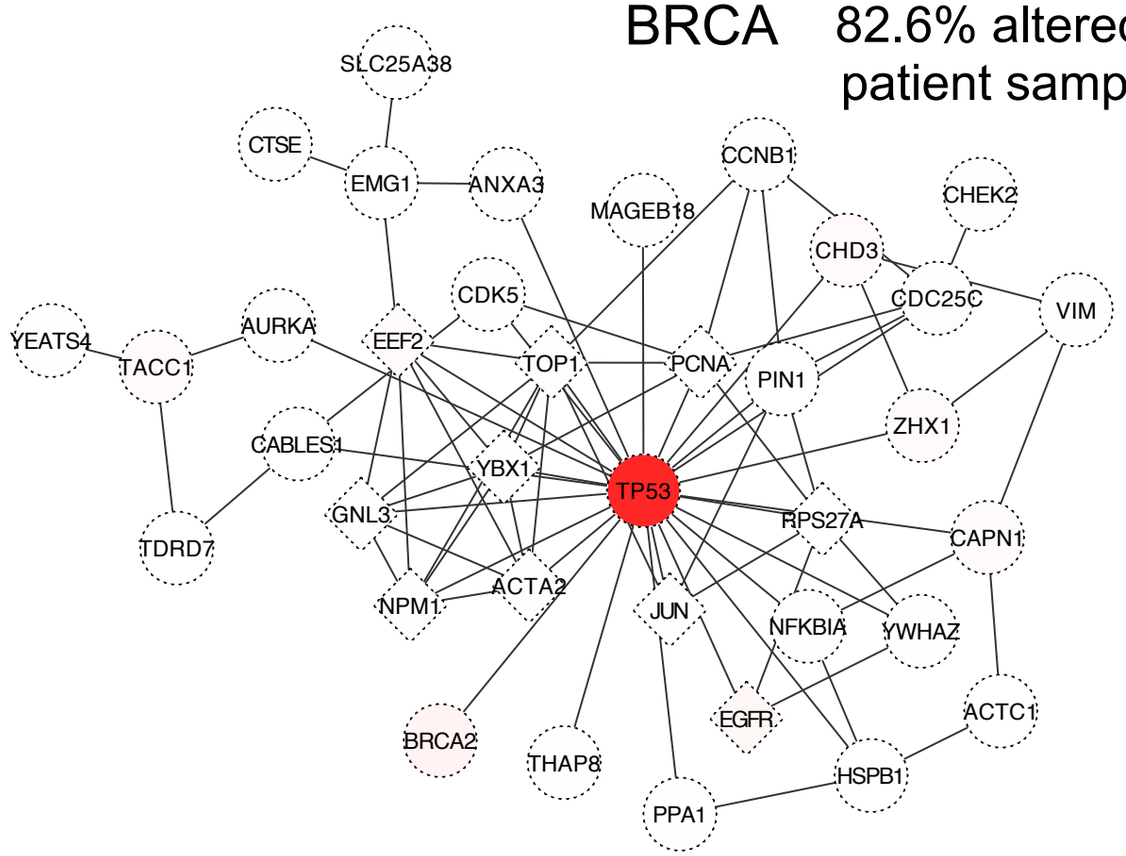
E



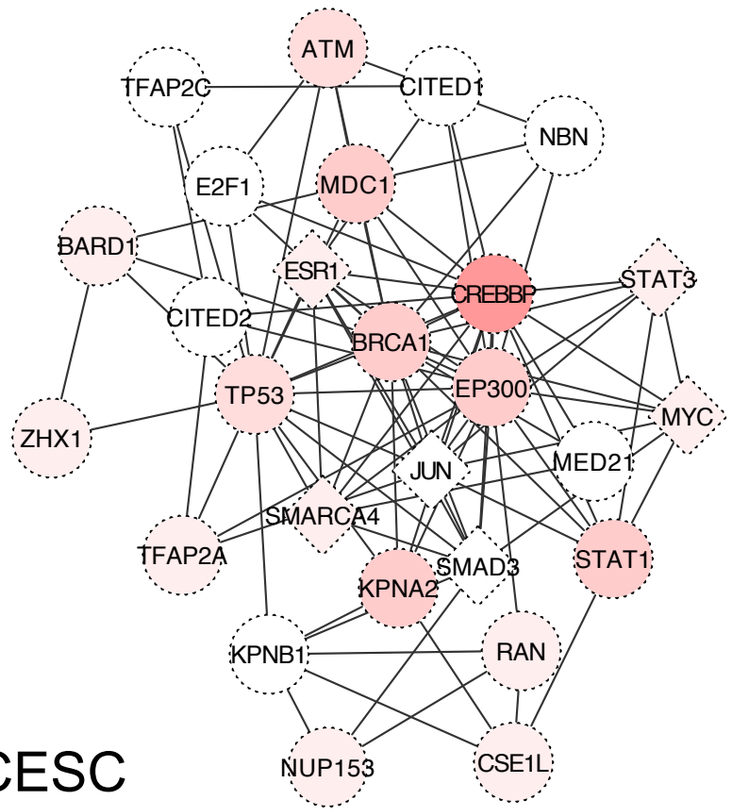
BLCA

88.2% altered in patient samples

F



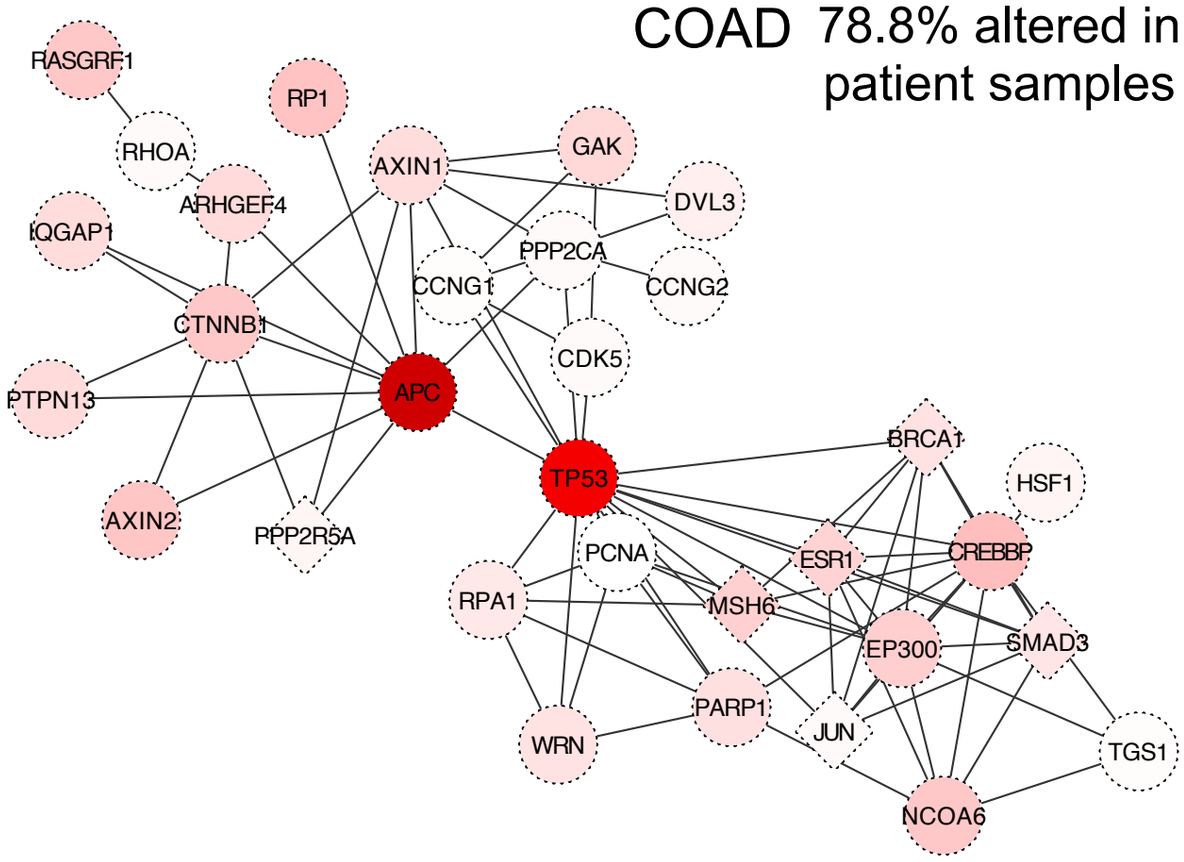
G



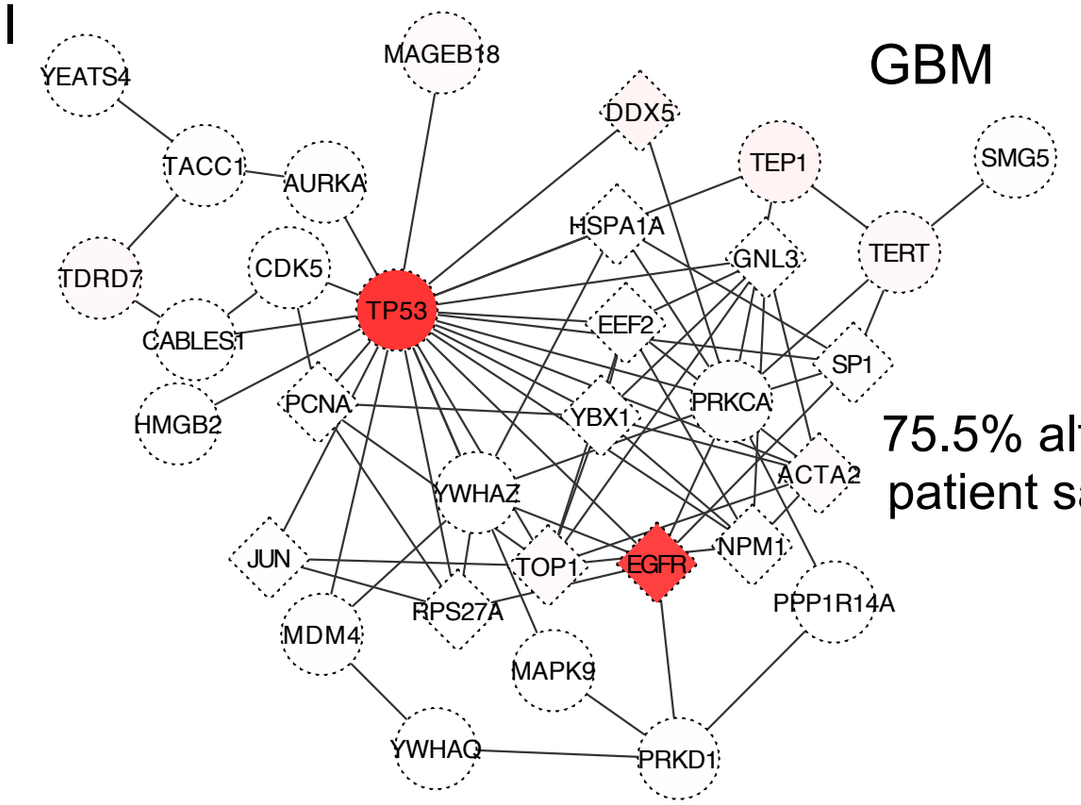
CESC

85.4% altered in patient samples

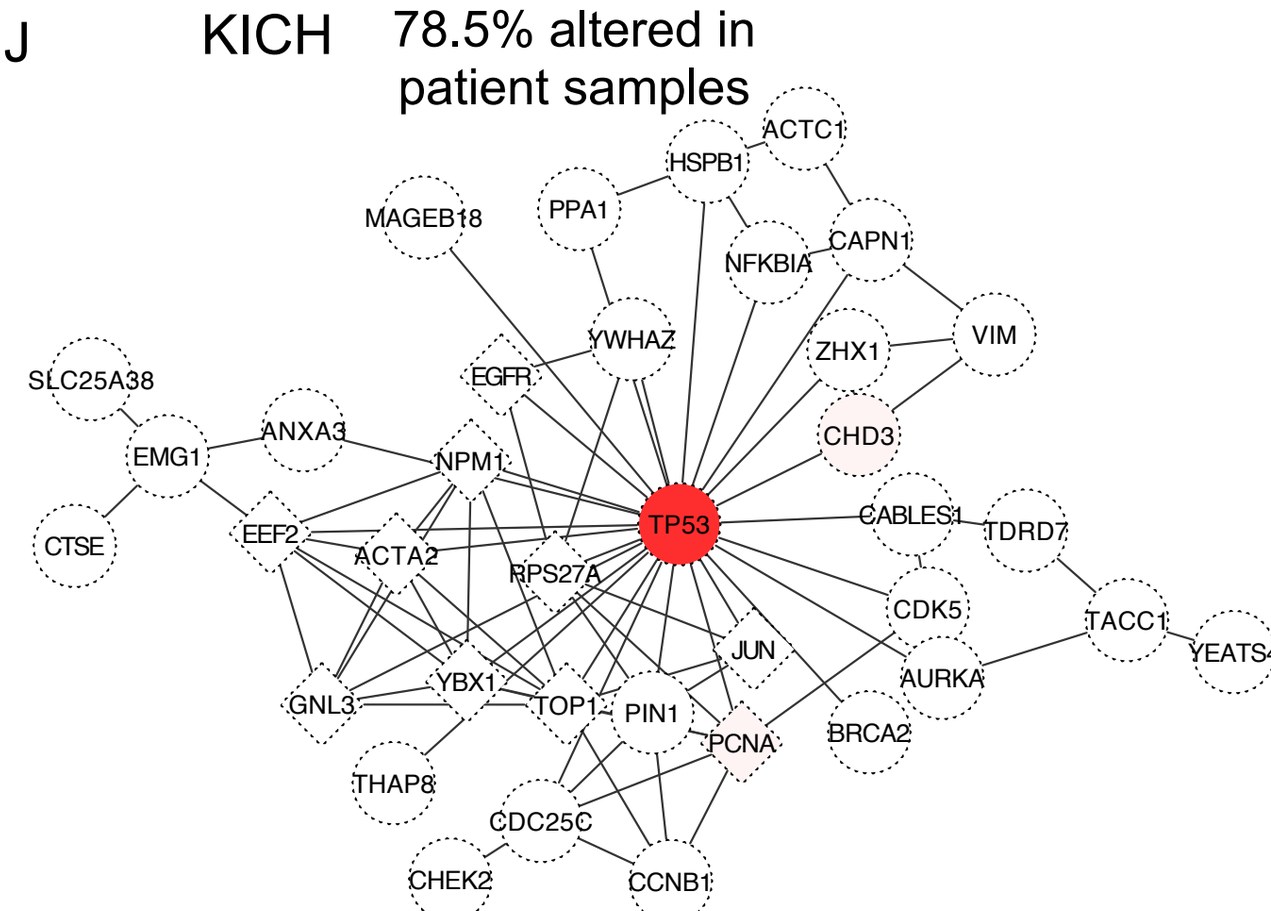
H



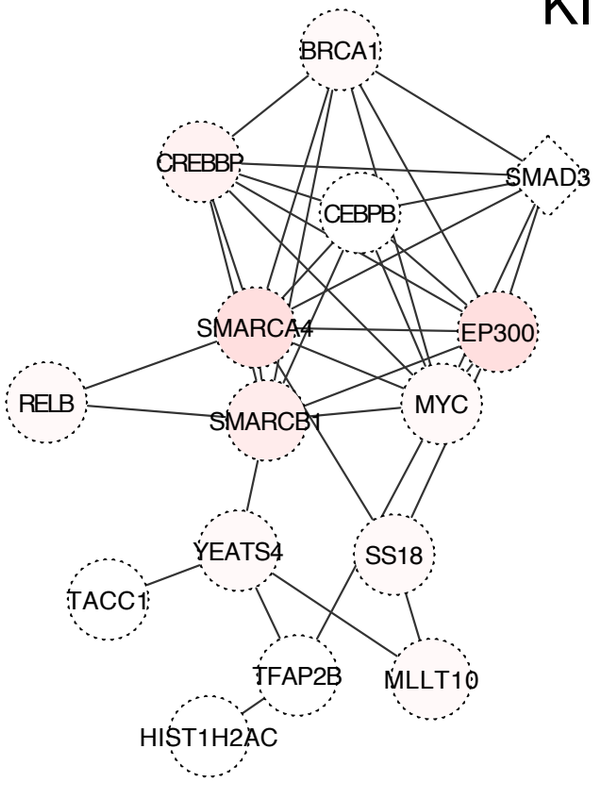
I


GBM
75.5% altered in patient samples
YEATS4
MAGEB18
DDX5
TEP1
SMG5
TACC1
AURKA
HSPA1A
GNL3
TERT
TDRD7
CDK5
TP53
EEF2
CABLES1
SP1
HMGB2
PCNA
PRKCA
YBX1
ACTA2
YWHAZ
JUN
TOP1
EGFR
NPM1
RPS27A
PPP1R14A
MDM4
MAPK9
YWHAQ
PRKD1

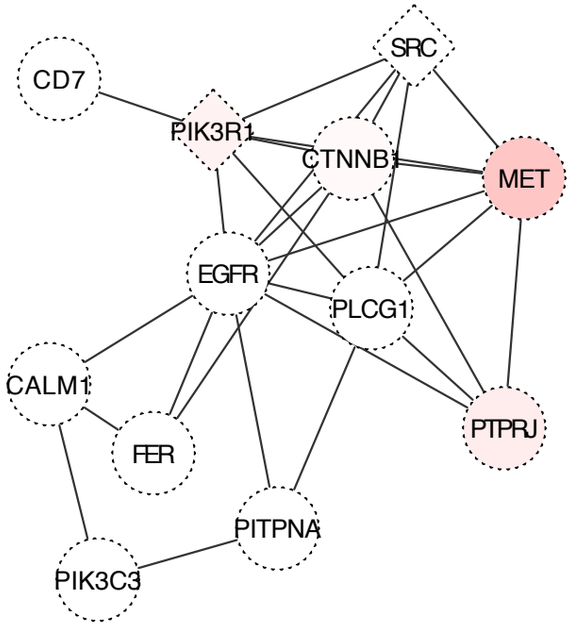
J

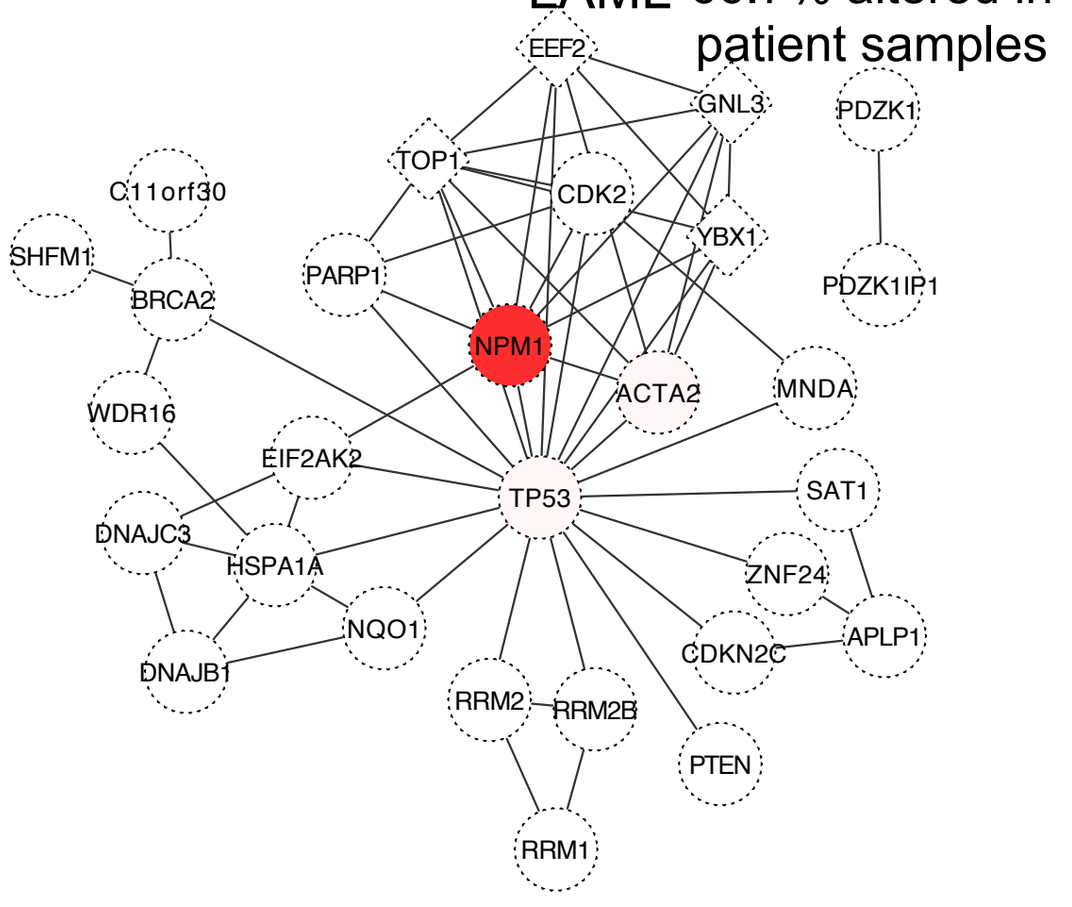

KICH
78.5% altered in patient samples
ACTC1
HSPB1
MAGEB18
PPA1
CAPN1
NFKBIA
YWHAZ
ZHX1
VIM
SLC25A38
EGFR
CHD3
EMG1
ANXA3
NPM1
CTSE
EEF2
TP53
CABLES1
TDRD7
ACTA2
RPS27A
CDK5
TACC1
YEATS4
JUN
AURKA
GNL3
YBX1
TOP1
PIN1
BRCA2
PCNA
THAP8
CDC25C
CHEK2
CCNB1

K

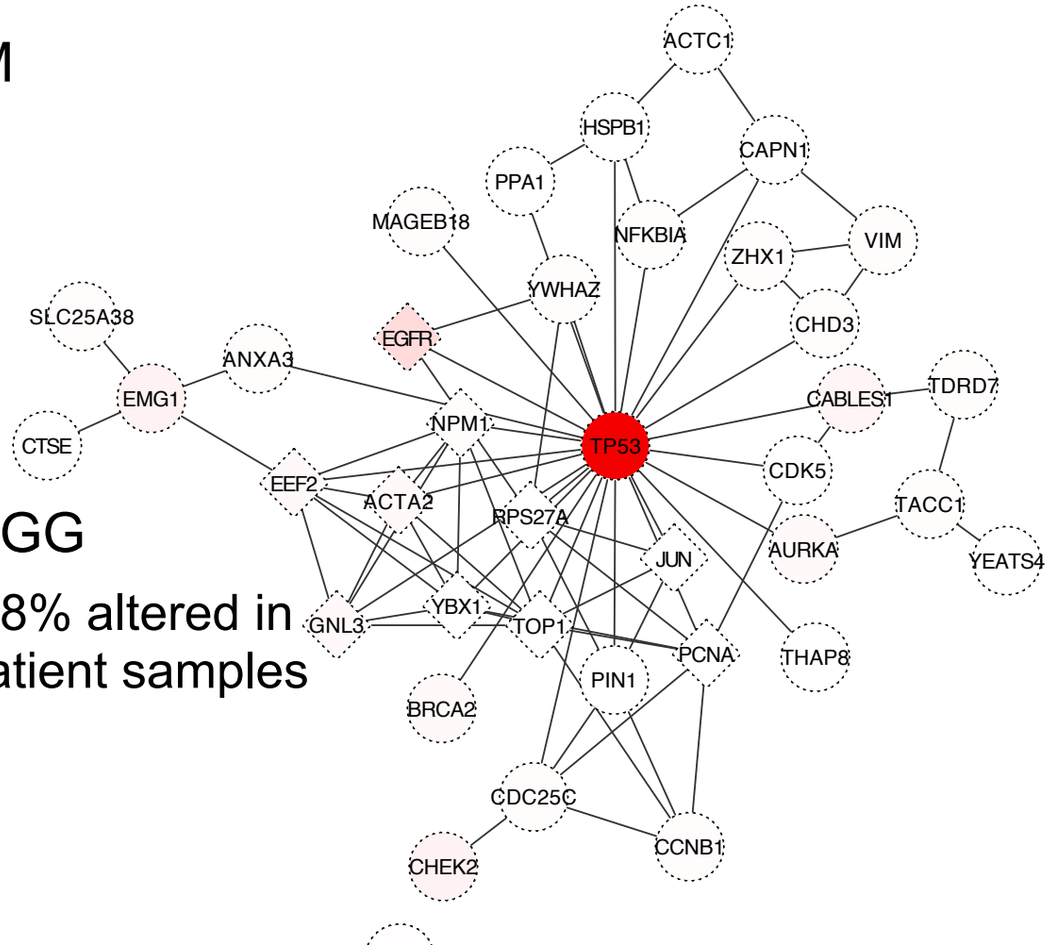

KIRP 65.8% altered in patient samples
BRCA1
CREBBP
SMAD3
CEBPB
SMARCA4
EP300
RELB
SMARCB1
MYC
YEATS4
SS18
TACC1
TFAP2B
MLLT10
HIST1H2AC

L


CD7
SRC
PIK3R1
CTNNB1
MET
EGFR
PLCG1
CALM1
PTPRJ
FER
PITPNA
PIK3C3


LAML 66.7% altered in patient samples
EEF2
GNL3
PDZK1
TOP1
CDK2
C11orf30
YBX1
SHFM1
PARP1
PDZK1IP1
BRCA2
NPM1
ACTA2
MNDA
WDR16
EIF2AK2
TP53
SAT1
DNAJC3
HSPA1A
ZNF24
NQO1
APLP1
DNAJB1
CDKN2C
RRM2
RRM2B
PTEN
RRM1

M



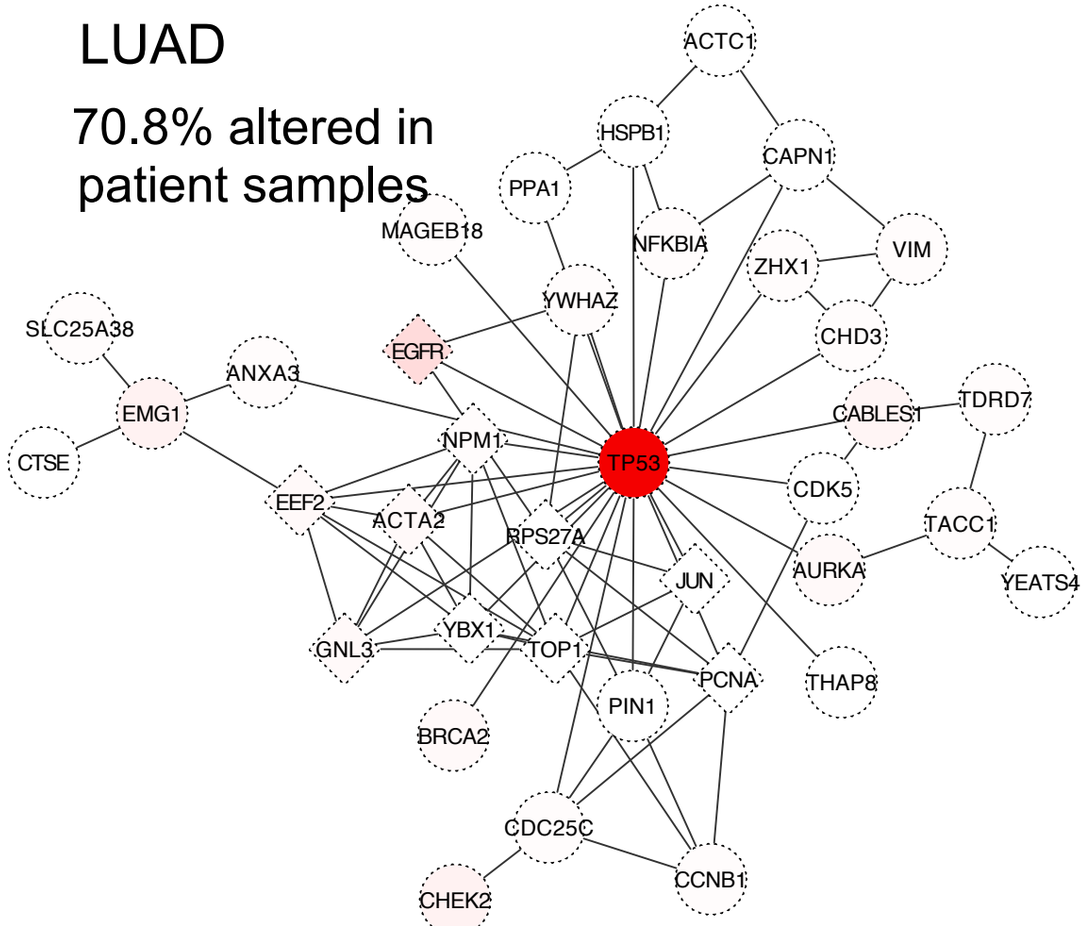
LGG

88% altered in patient samples

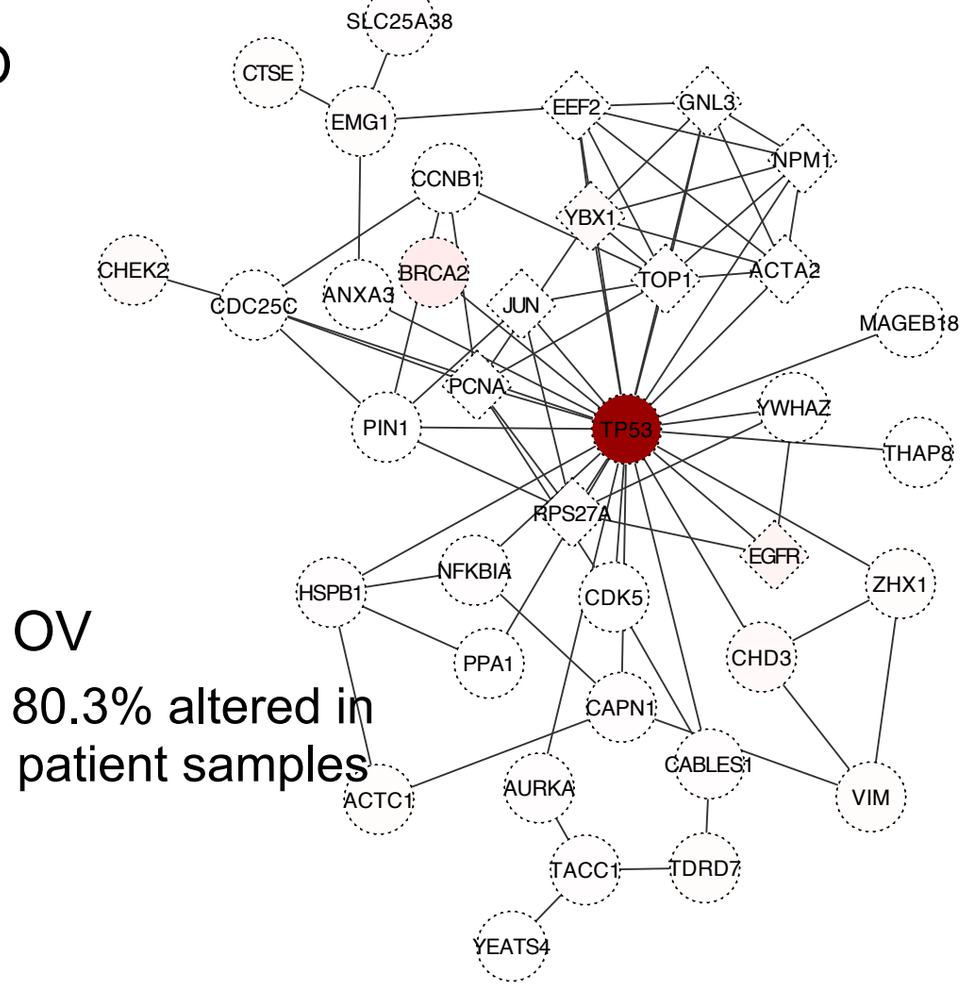
N

LUAD

70.8% altered in patient samples



O



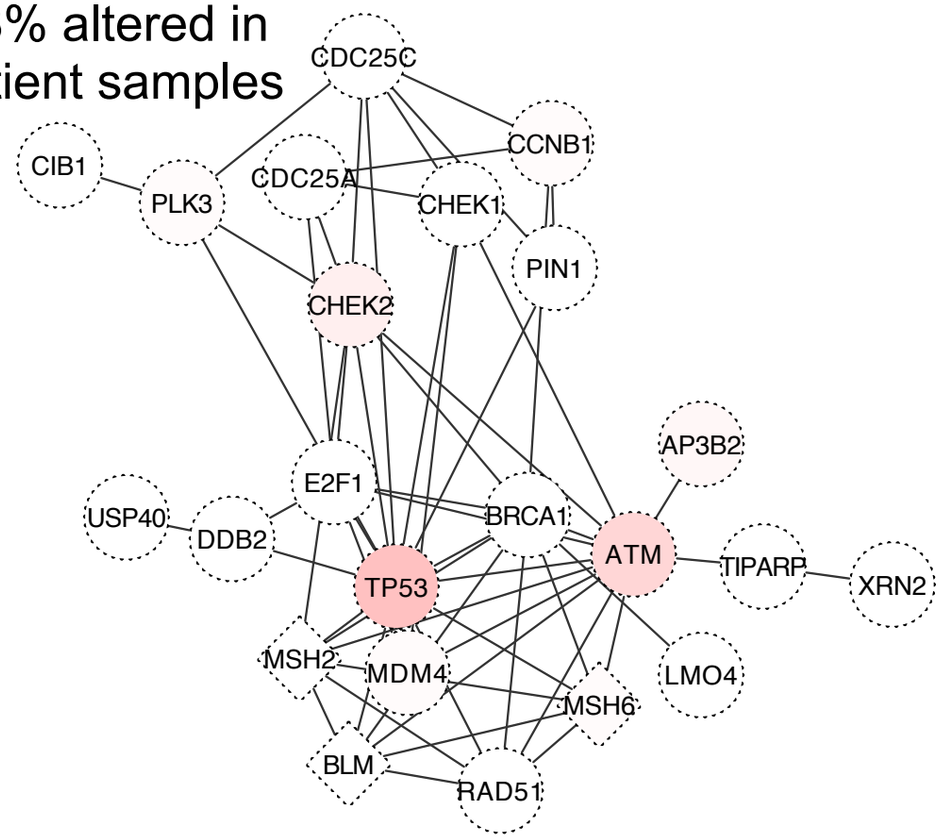
OV

80.3% altered in patient samples

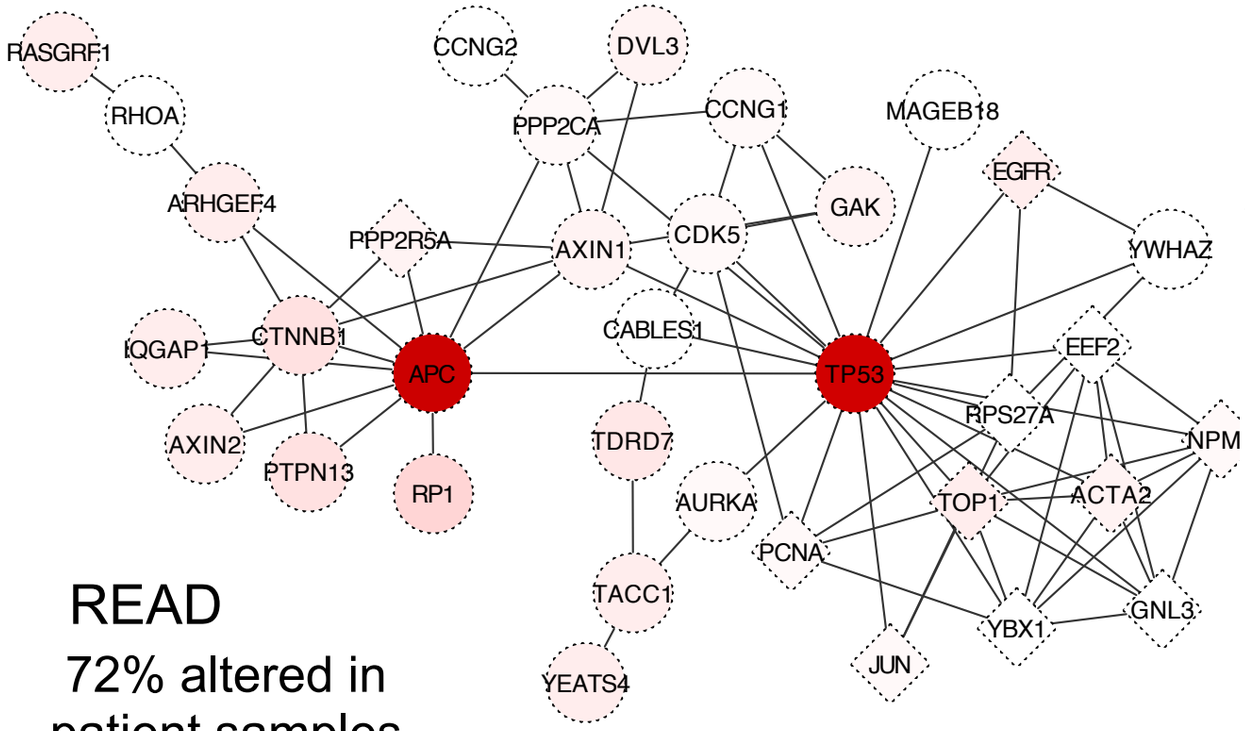
P

PRAD

48% altered in patient samples



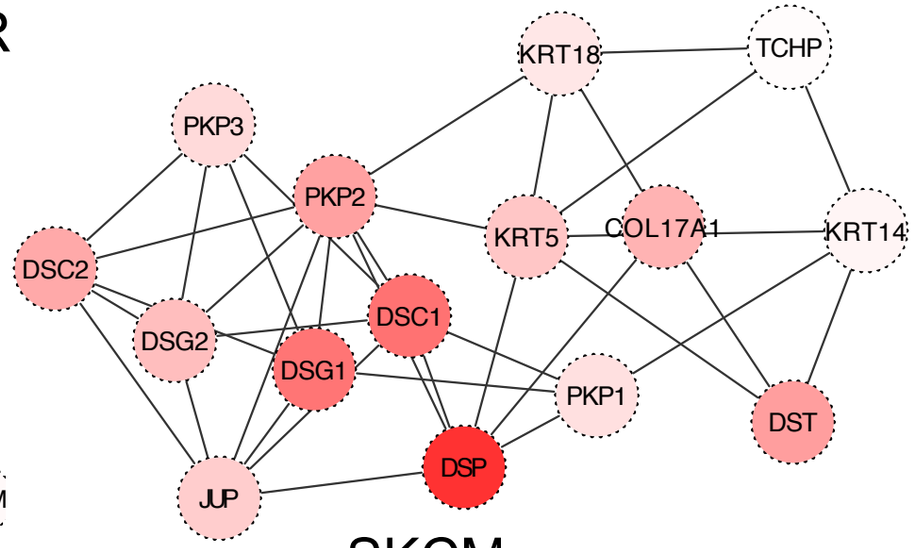
Q



READ

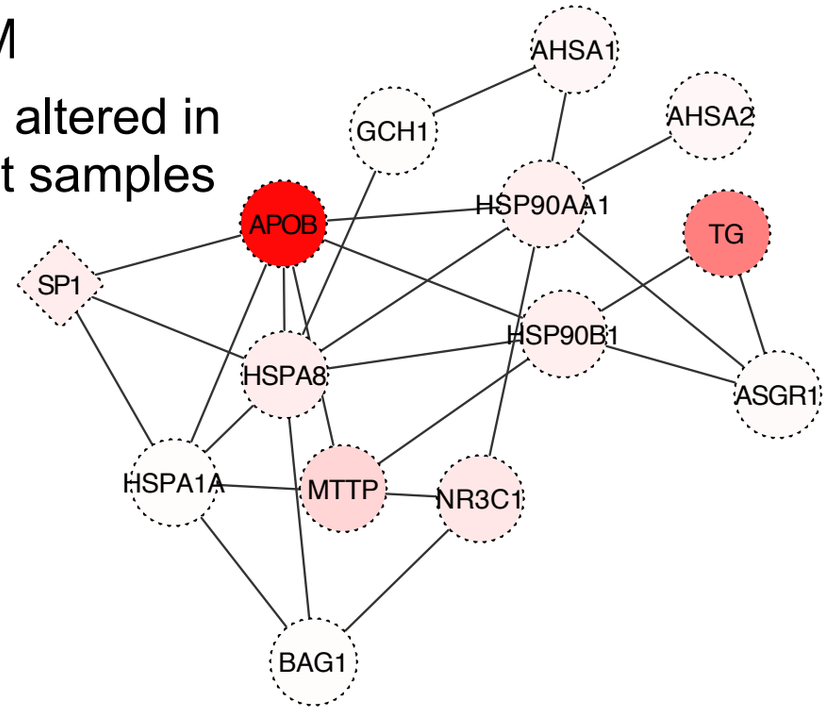
72% altered in patient samples

R

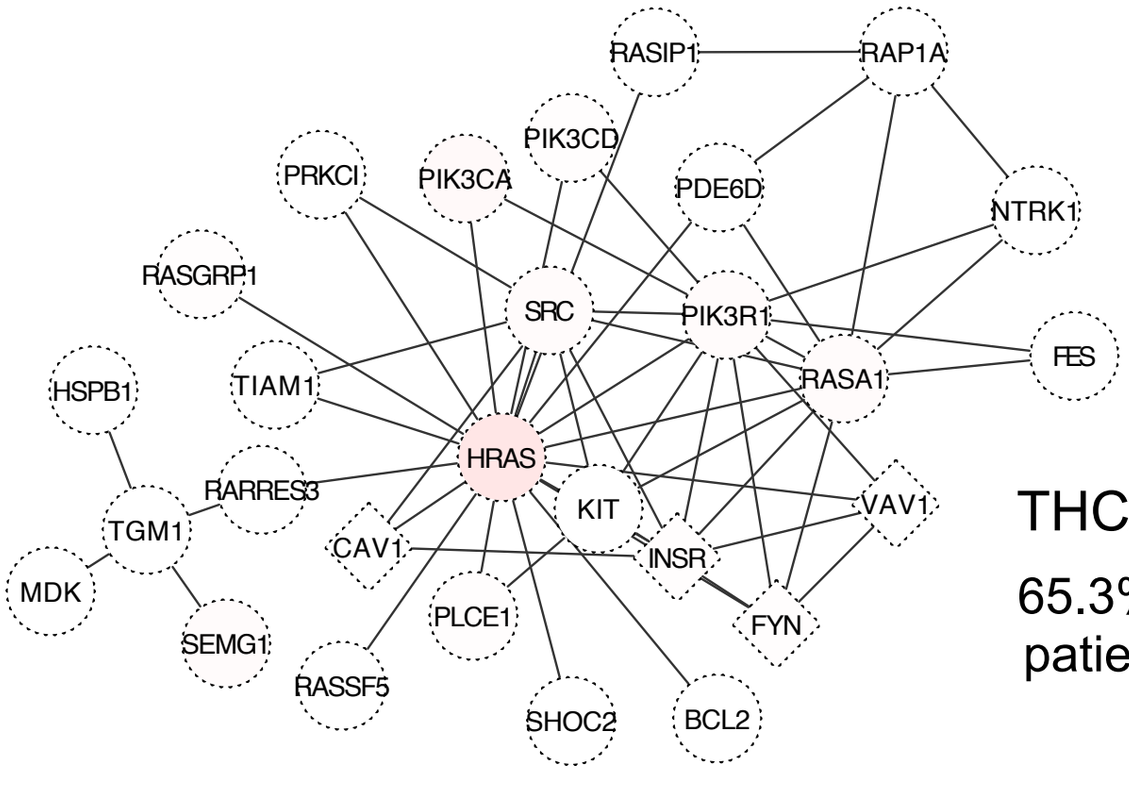


SKCM

86.2% altered in patient samples



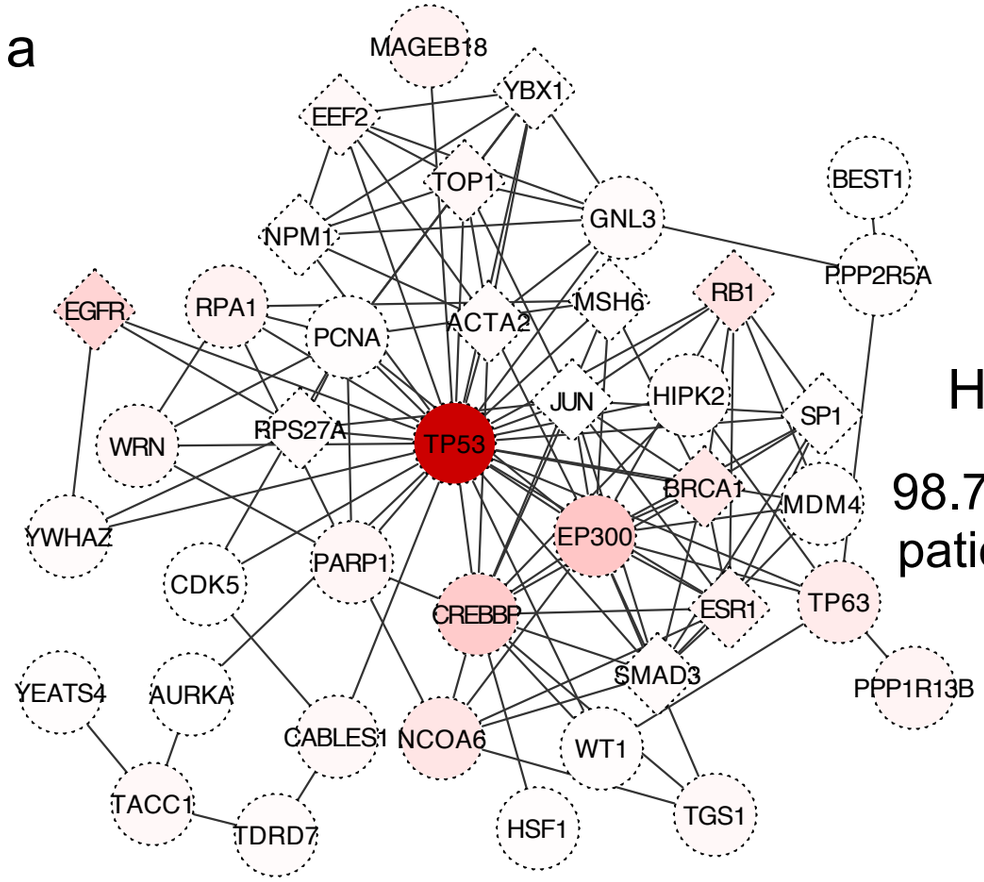
S



THCA

65.3% altered in patient samples

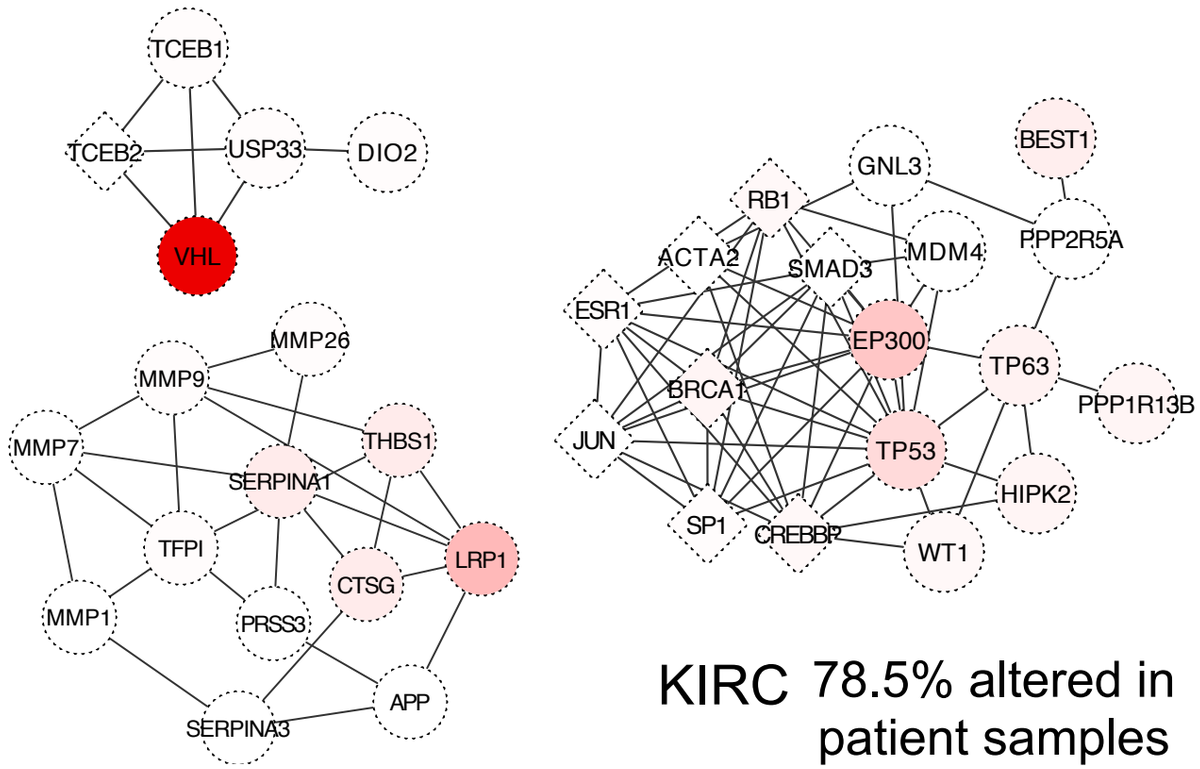
a



HNSC

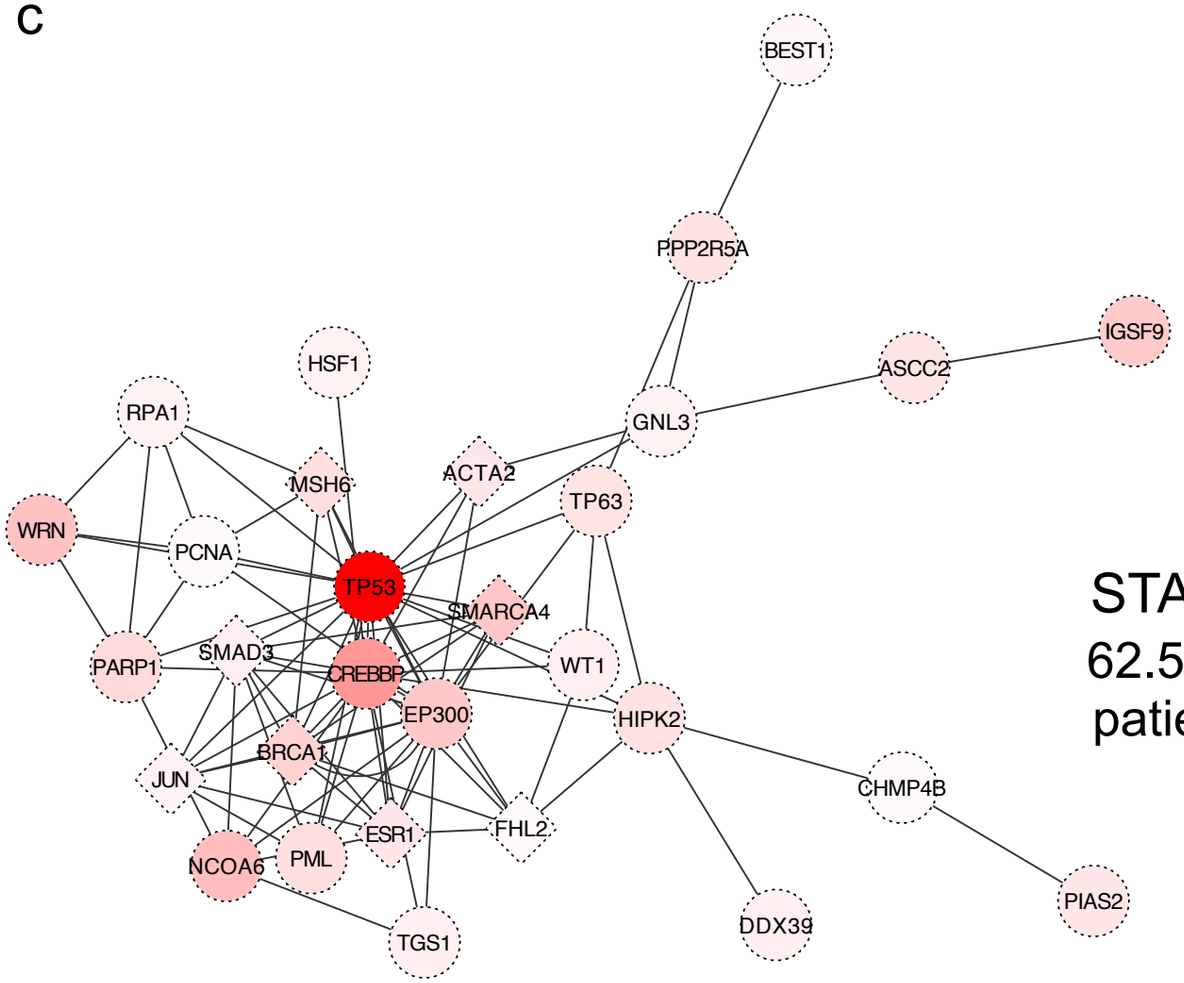
98.7% altered in patient samples

b



KIRC 78.5% altered in patient samples

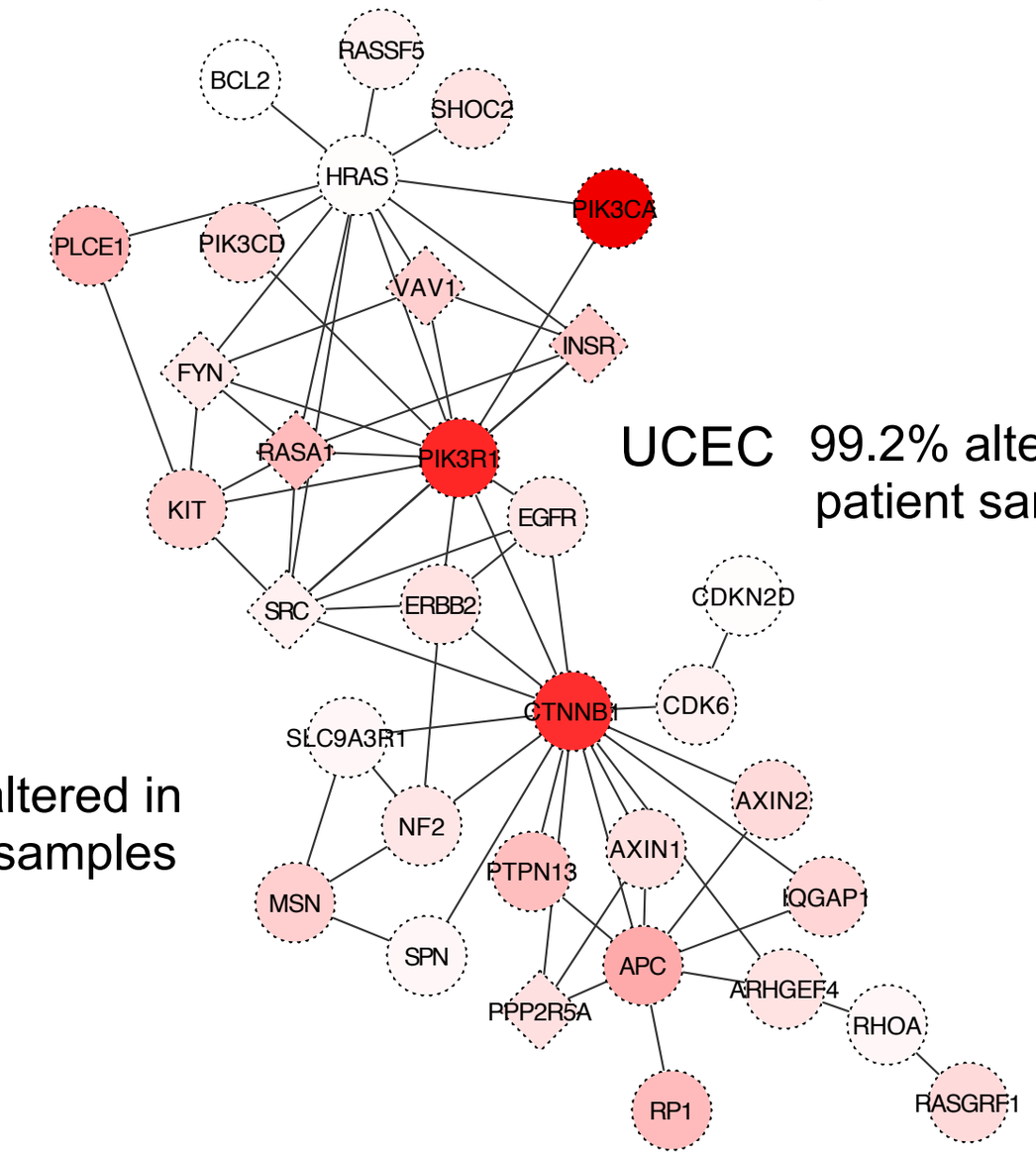
c



STAD

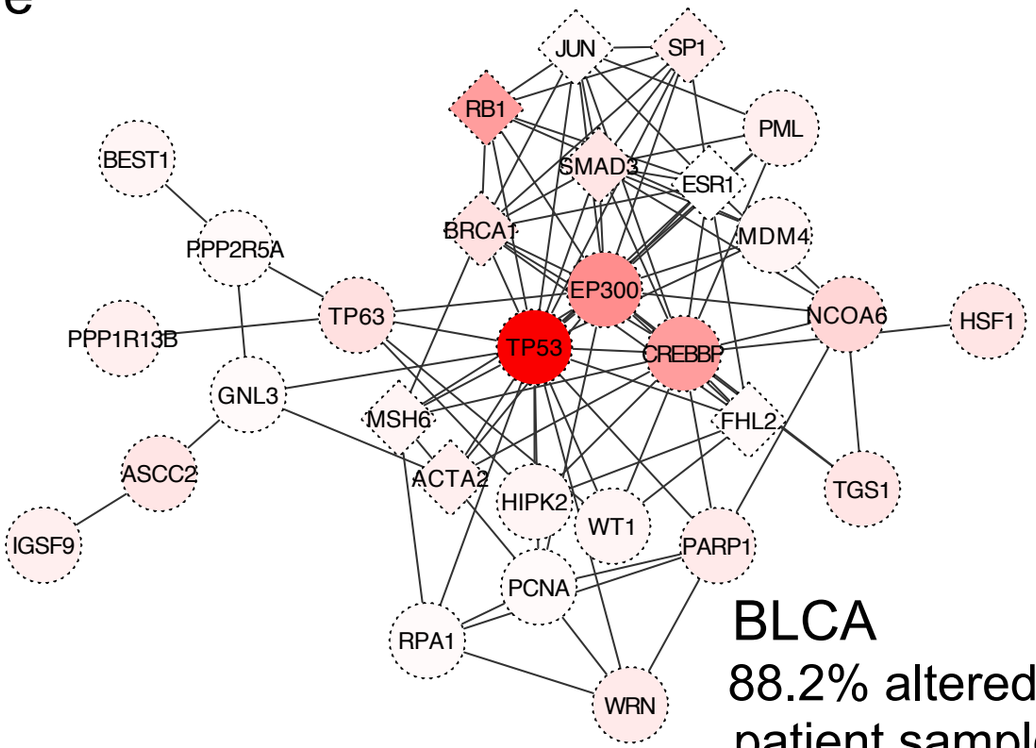
62.5% altered in patient samples

d



UCEC 99.2% altered in patient samples

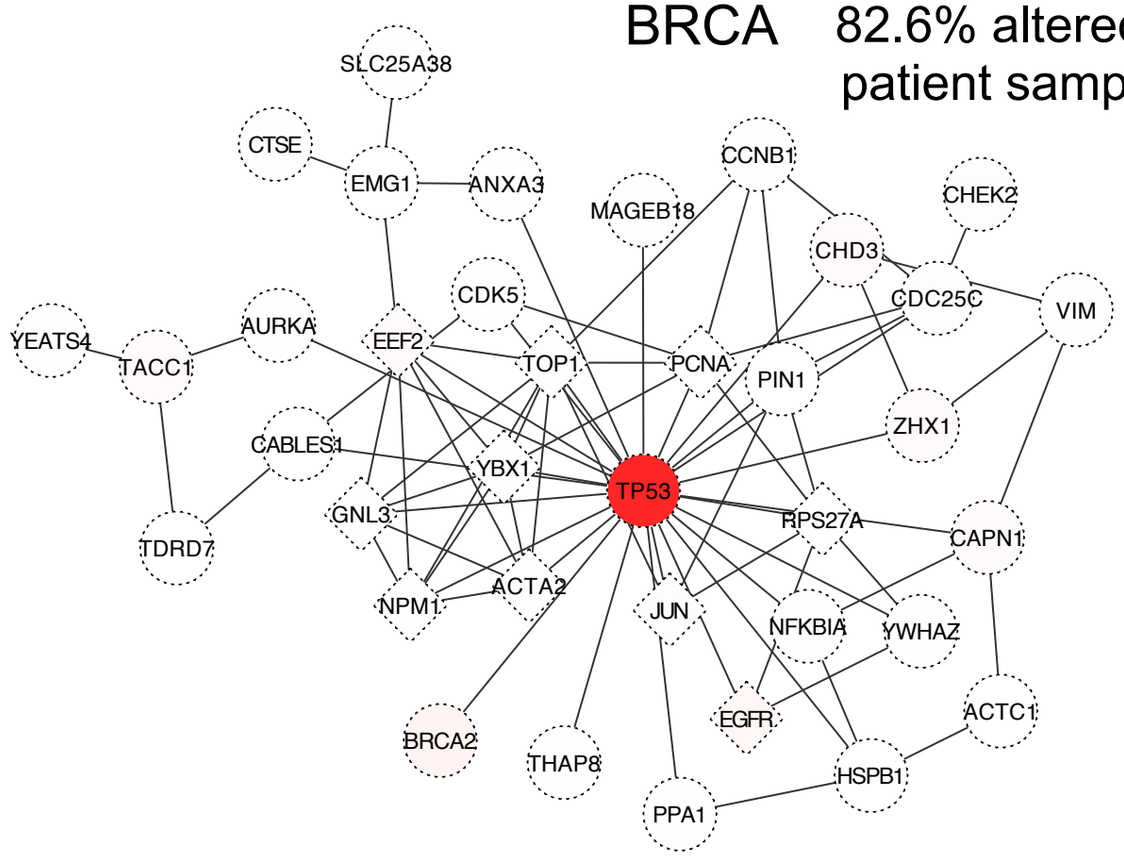
e



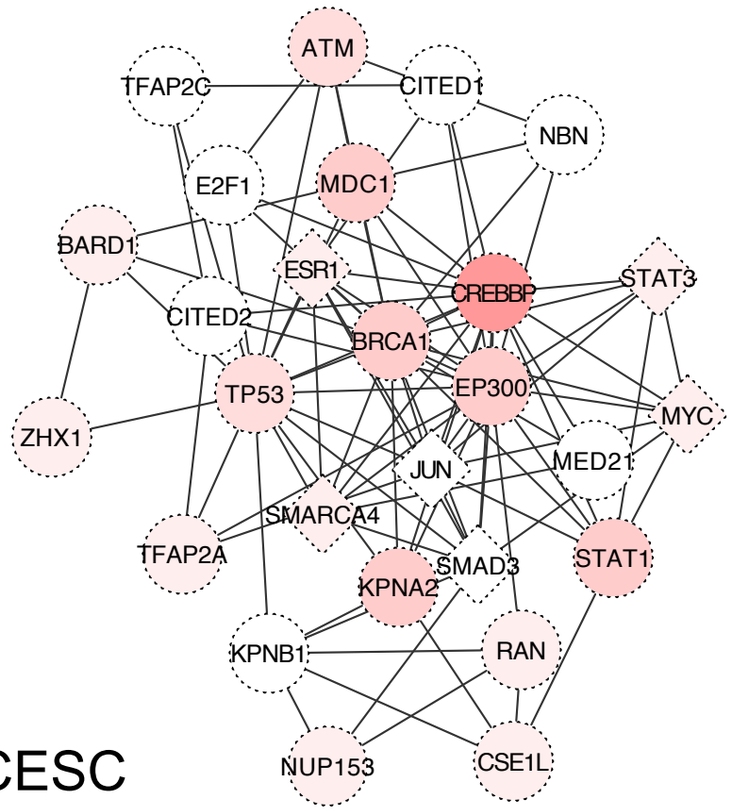
BLCA

88.2% altered in patient samples

f



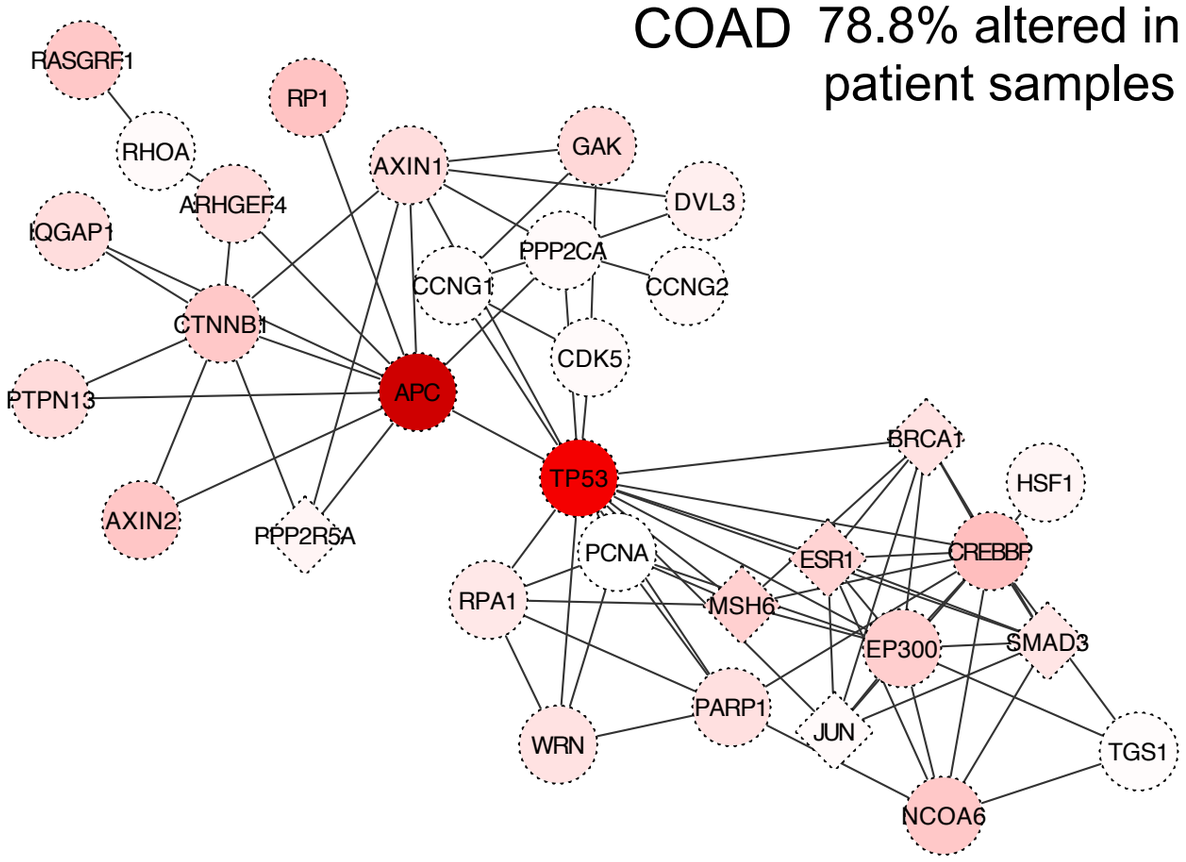
g



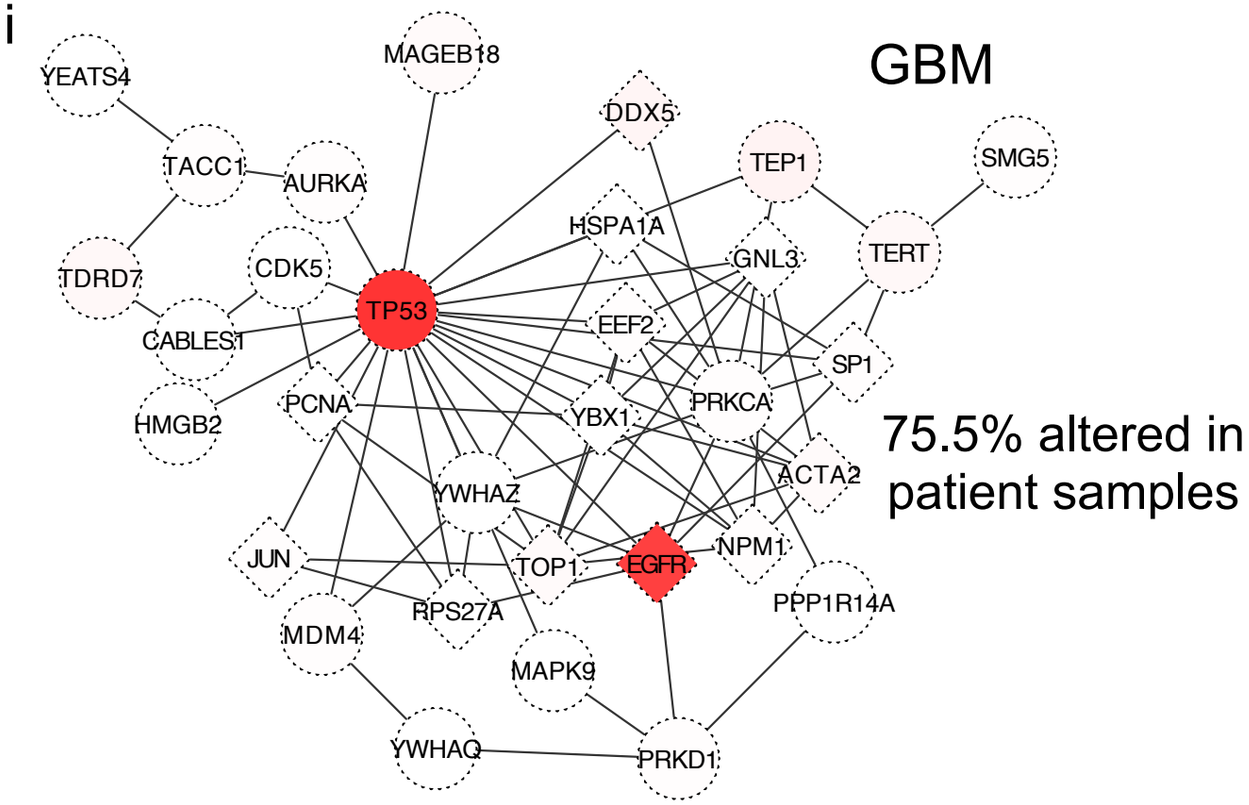
CESC

85.4% altered in patient samples

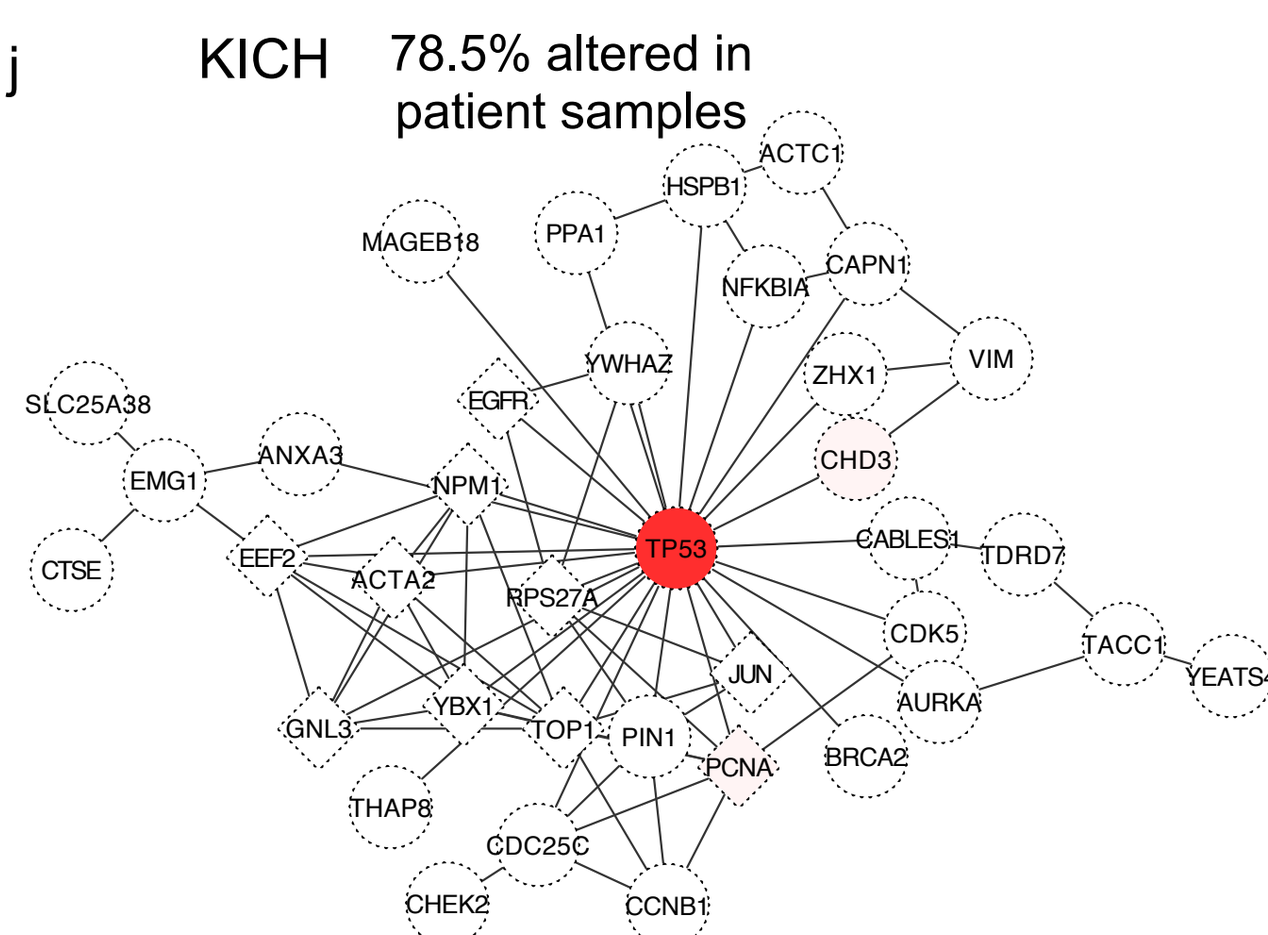
h



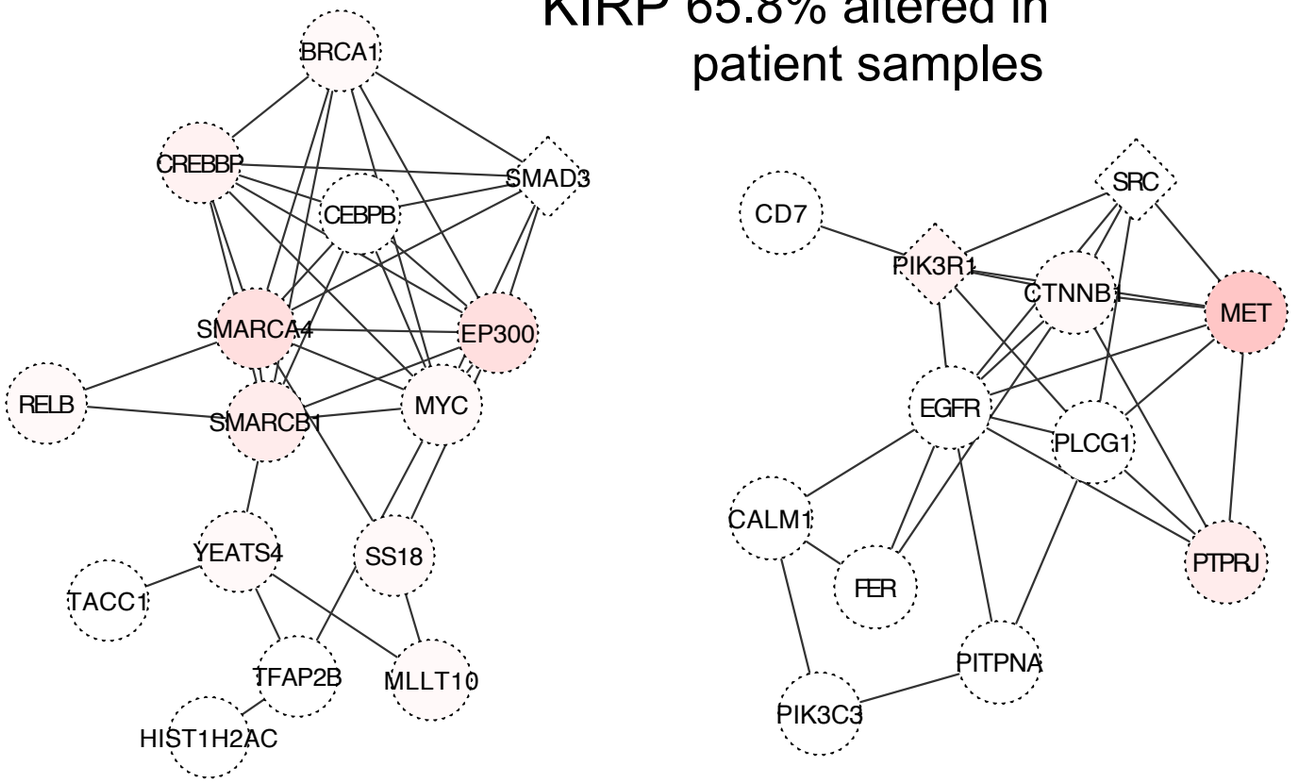
i


GBM
75.5% altered in patient samples
YEATS4
MAGEB18
DDX5
TEP1
SMG5
TACC1
AURKA
HSPA1A
GNL3
TERT
TDRD7
CDK5
TP53
EEF2
CABLES1
SP1
HMGB2
PCNA
PRKCA
YBX1
ACTA2
YWHAZ
JUN
TOP1
EGFR
NPM1
RPS27A
PPP1R14A
MDM4
MAPK9
YWHAQ
PRKD1

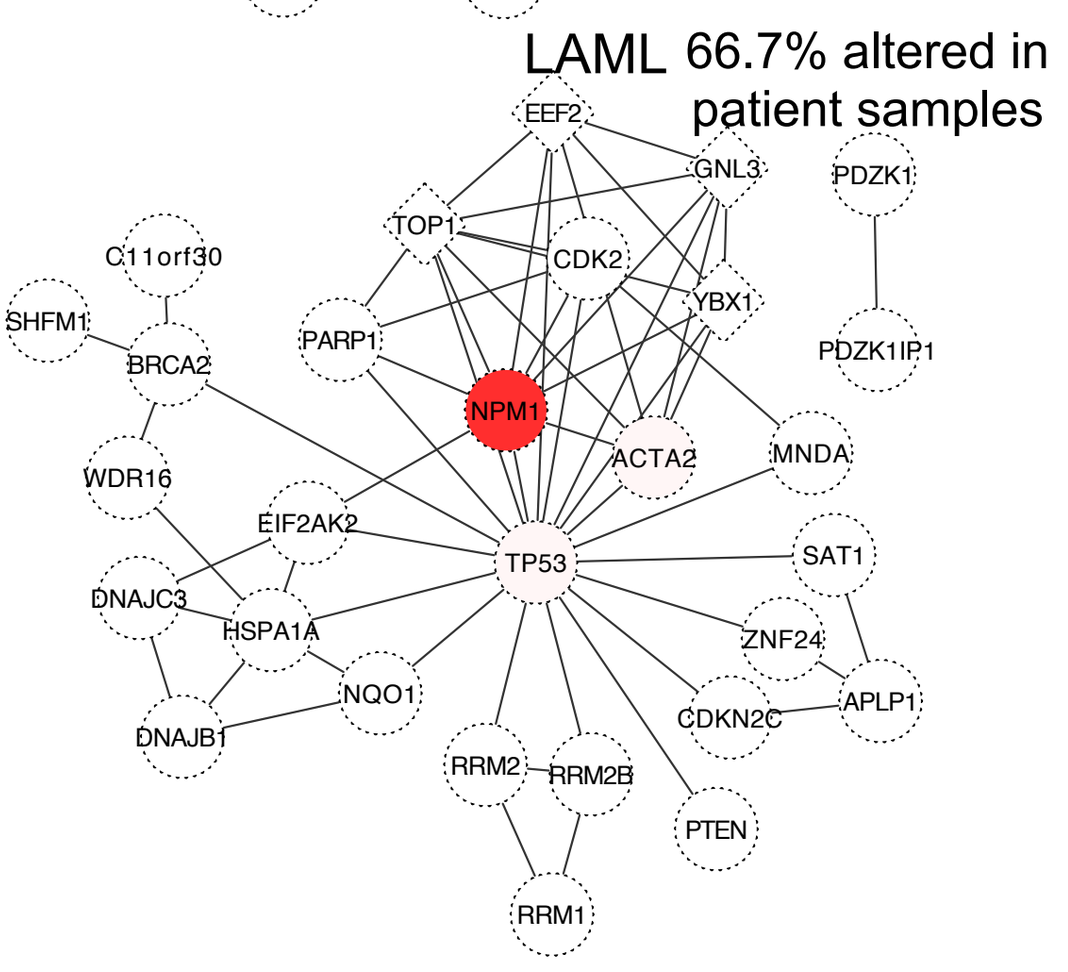
j


KICH
78.5% altered in patient samples
ACTC1
HSPB1
MAGEB18
PPA1
CAPN1
NFKBIA
YWHAZ
ZHX1
VIM
SLC25A38
EGFR
CHD3
EMG1
ANXA3
NPM1
CTSE
EEF2
TP53
CABLES1
TDRD7
ACTA2
RPS27A
CDK5
TACC1
JUN
YEATS4
AURKA
YBX1
GNL3
TOP1
PIN1
PCNA
BRCA2
THAP8
CDC25C
CHEK2
CCNB1

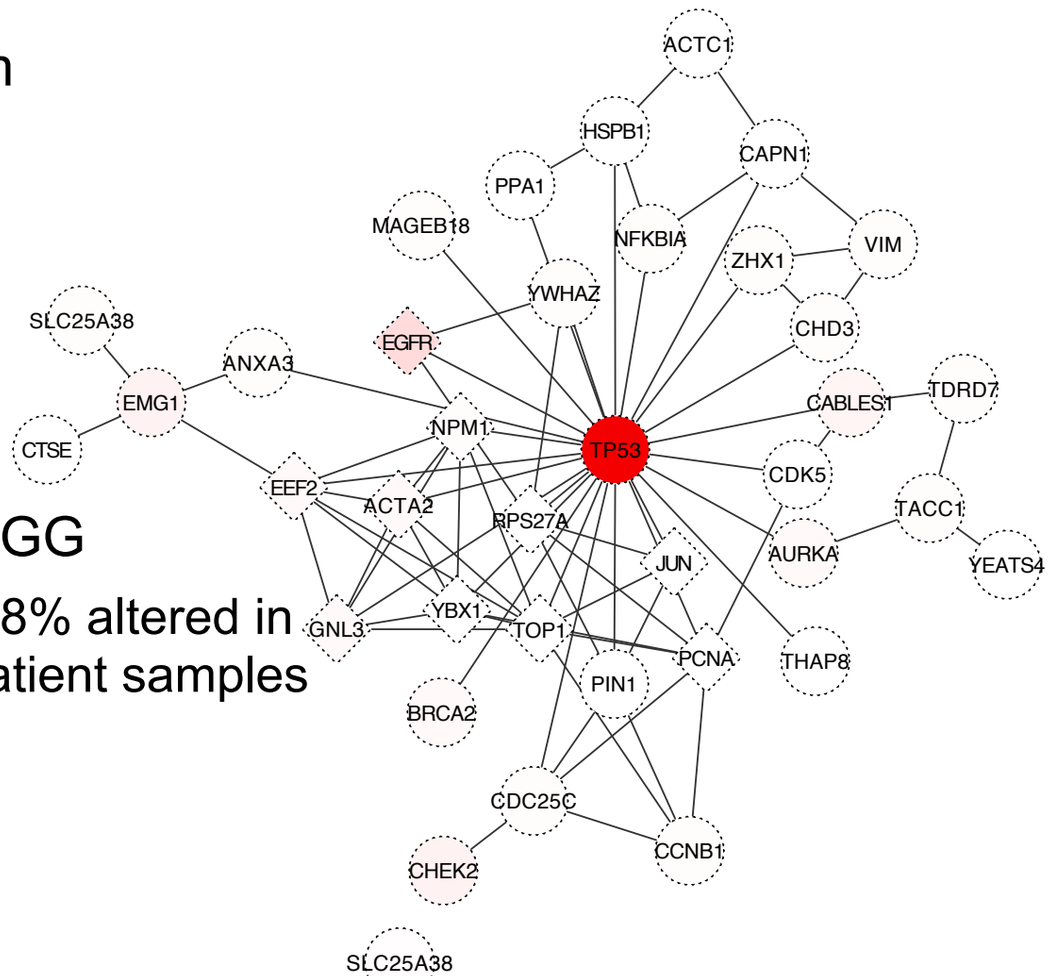
k


KIRP 65.8% altered in patient samples
BRCA1
CREBBP
SMAD3
CEBPB
SMARCA4
EP300
RELB
SMARCB1
MYC
YEATS4
SS18
TACC1
TFAP2B
MLLT10
HIST1H2AC
CD7
SRC
PIK3R1
CTNNB1
MET
EGFR
PLCG1
CALM1
FER
PTPRJ
PITPNA
PIK3C3

l


LAML 66.7% altered in patient samples
EEF2
GNL3
PDZK1
TOP1
CDK2
C11orf30
YBX1
SHFM1
PARP1
PDZK1IP1
BRCA2
NPM1
ACTA2
MNDA
WDR16
EIF2AK2
TP53
SAT1
DNAJC3
HSPA1A
ZNF24
NQO1
APLP1
DNAJB1
CDKN2C
RRM2
RRM2B
PTEN
RRM1

m



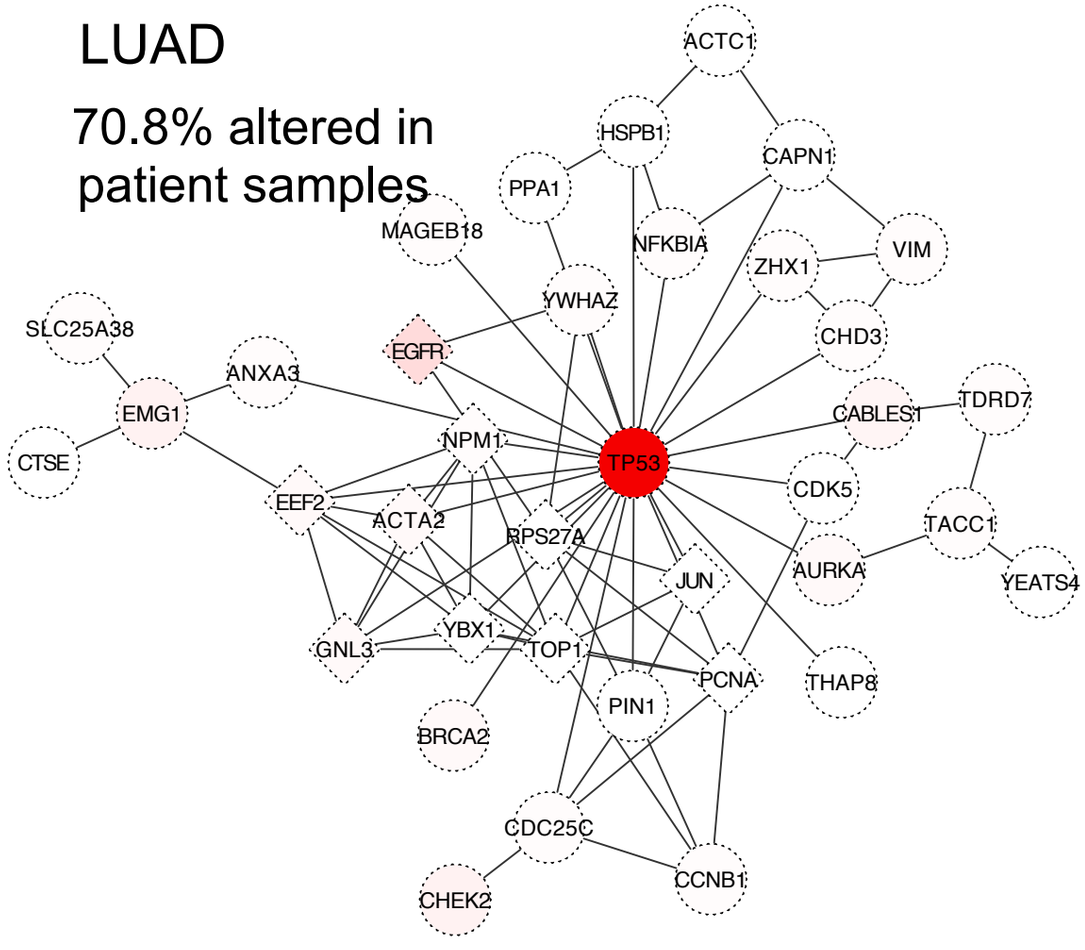
LGG

88% altered in patient samples

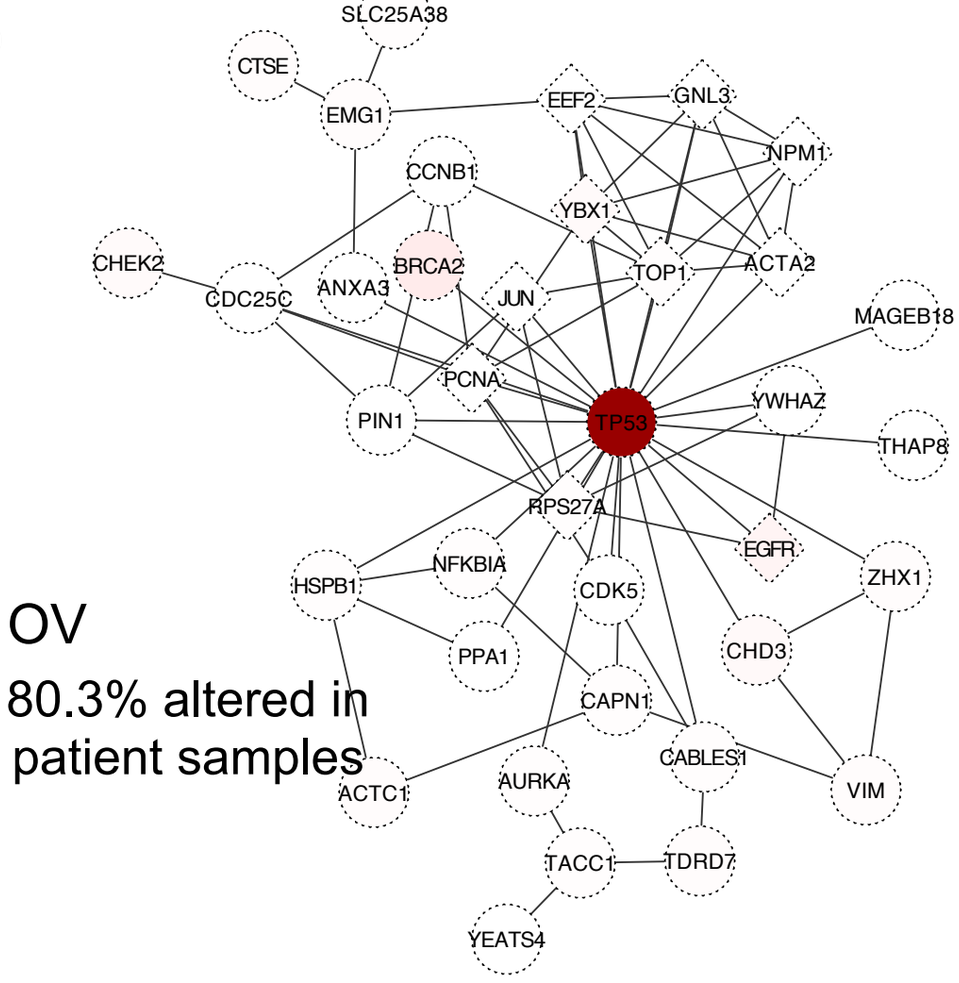
n

LUAD

70.8% altered in patient samples



o



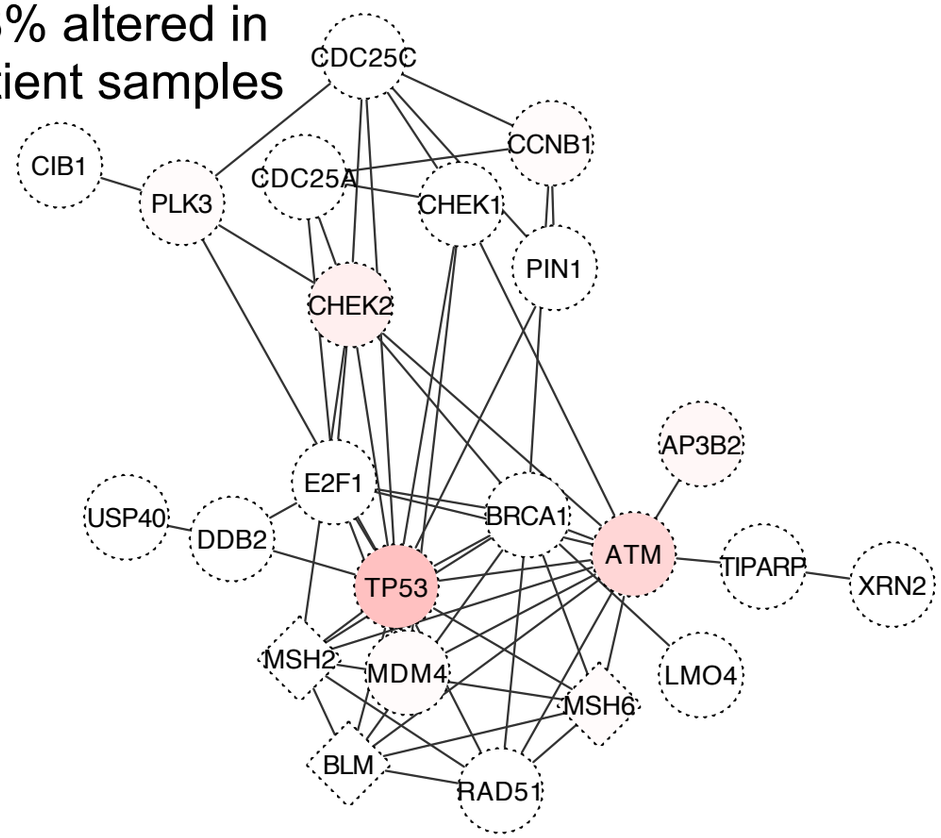
OV

80.3% altered in patient samples

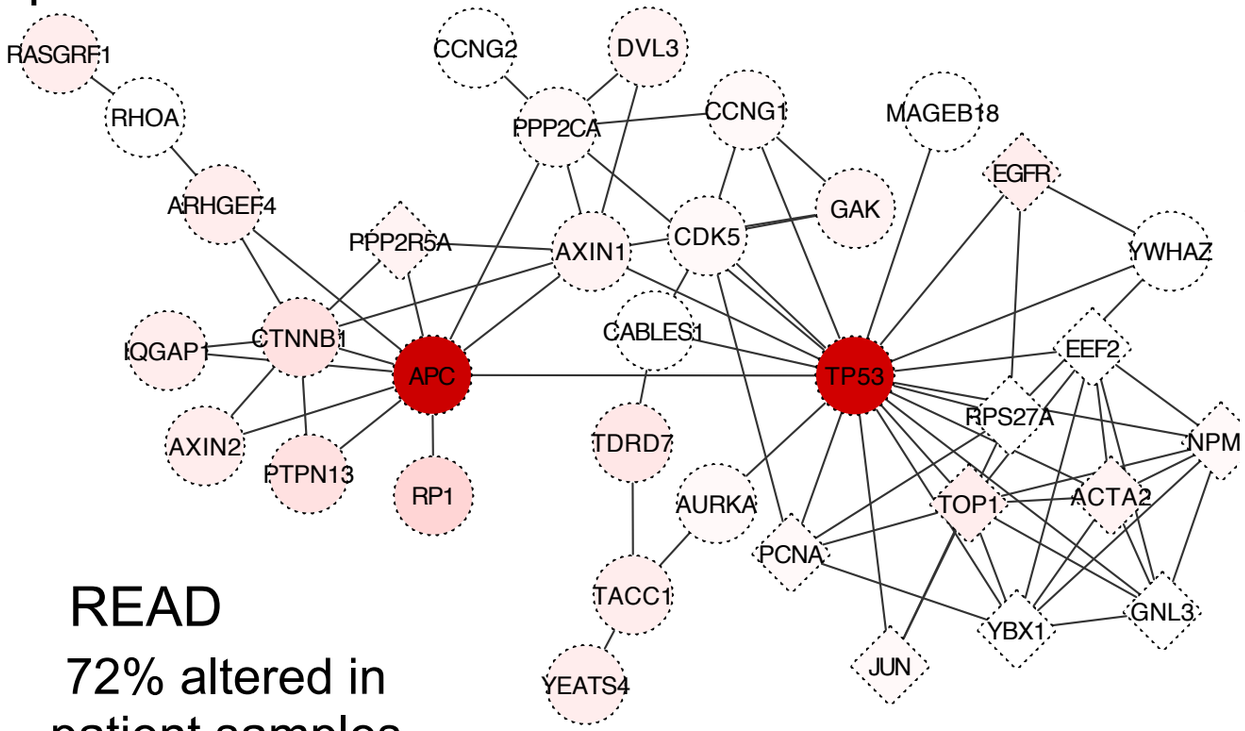
p

PRAD

48% altered in patient samples



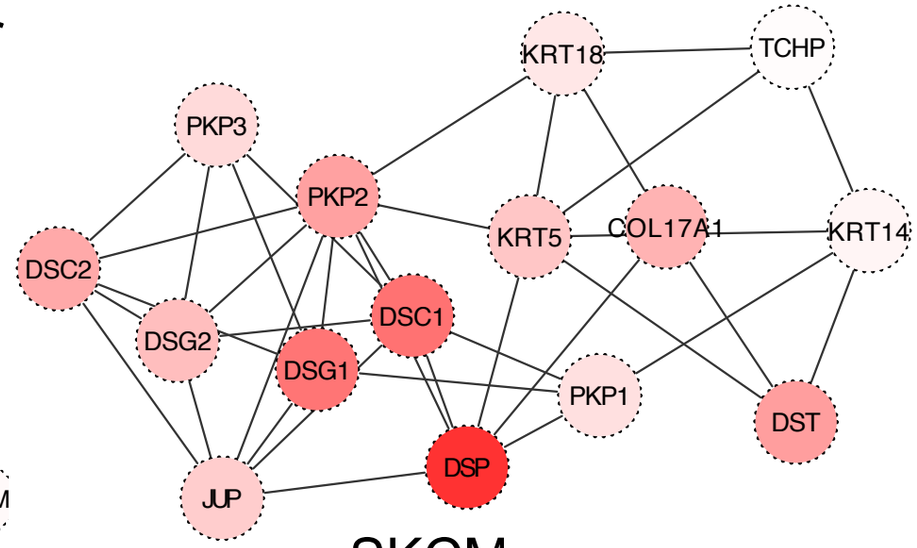
q



READ

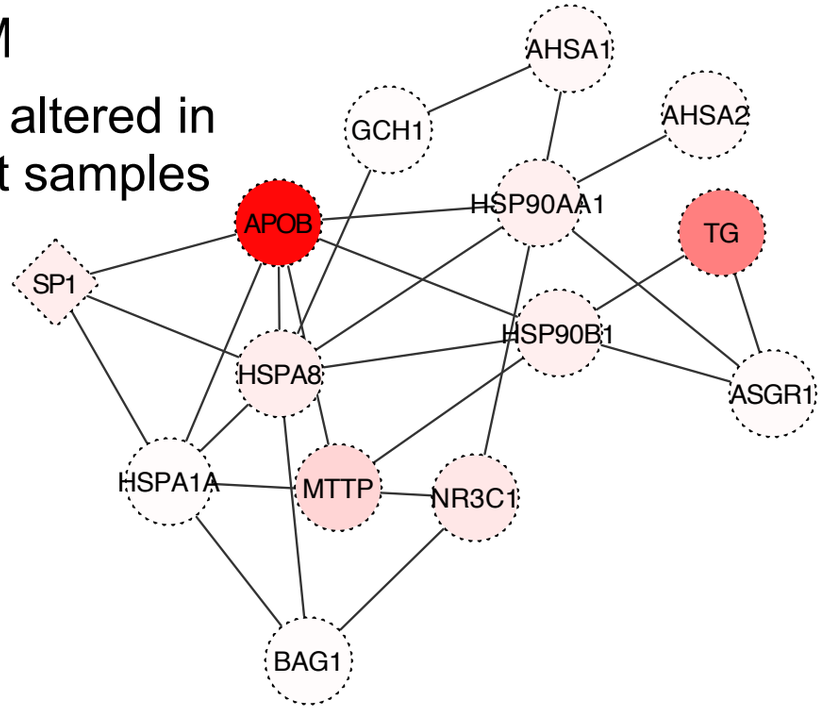
72% altered in patient samples

r

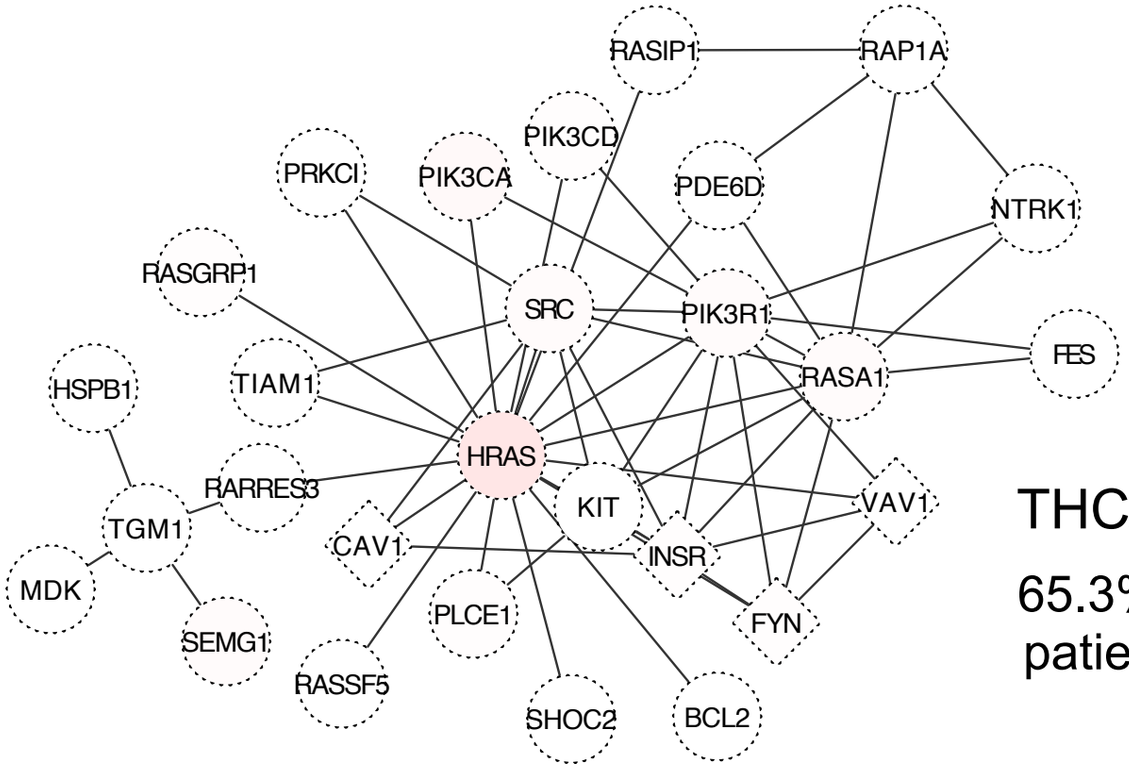


SKCM

86.2% altered in patient samples



s



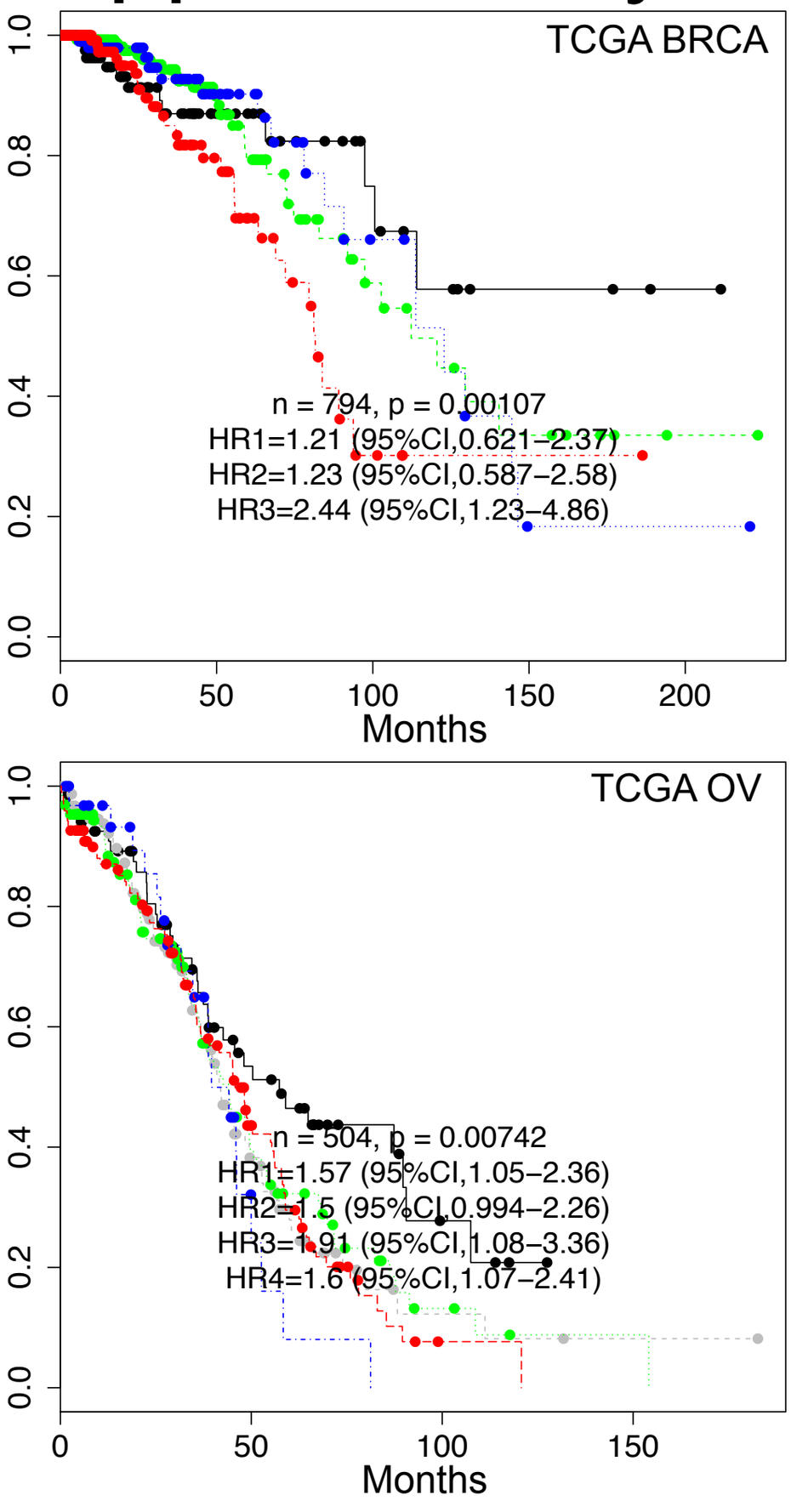
THCA

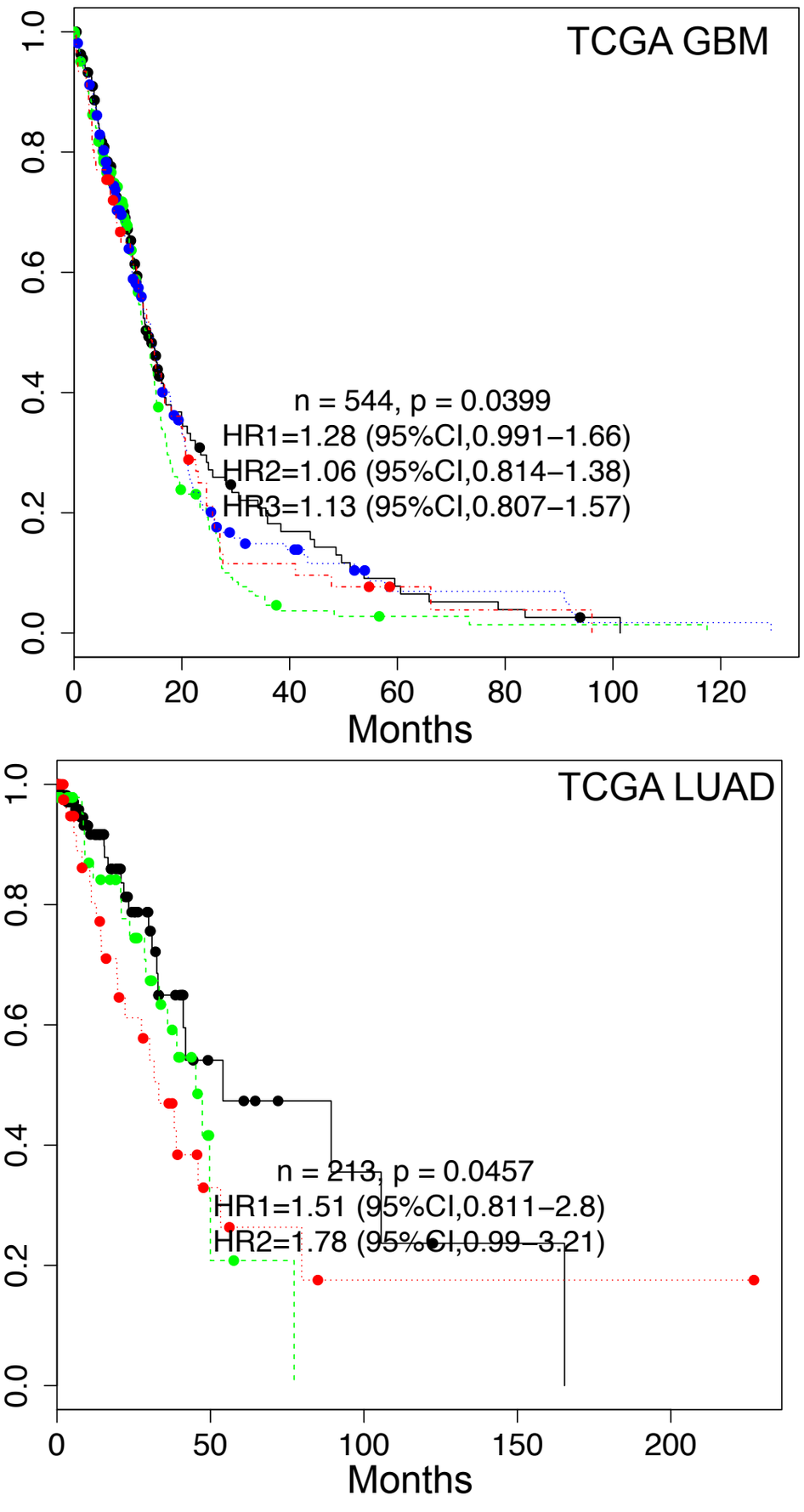
65.3% altered in patient samples

Mutation frequency

80%
70%
60%
50%
40%
20%
10%
0%

Supplementary Figure 2

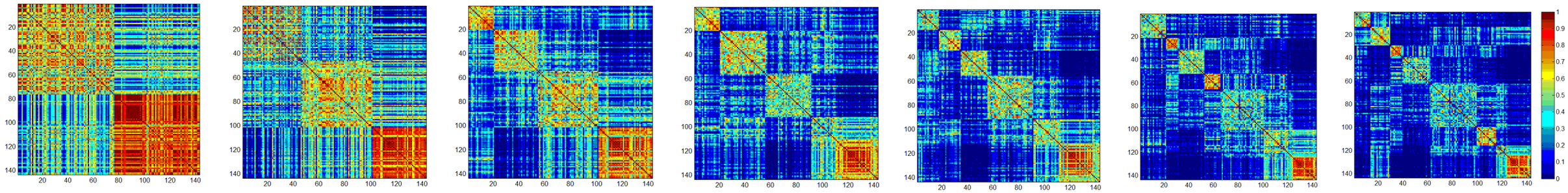




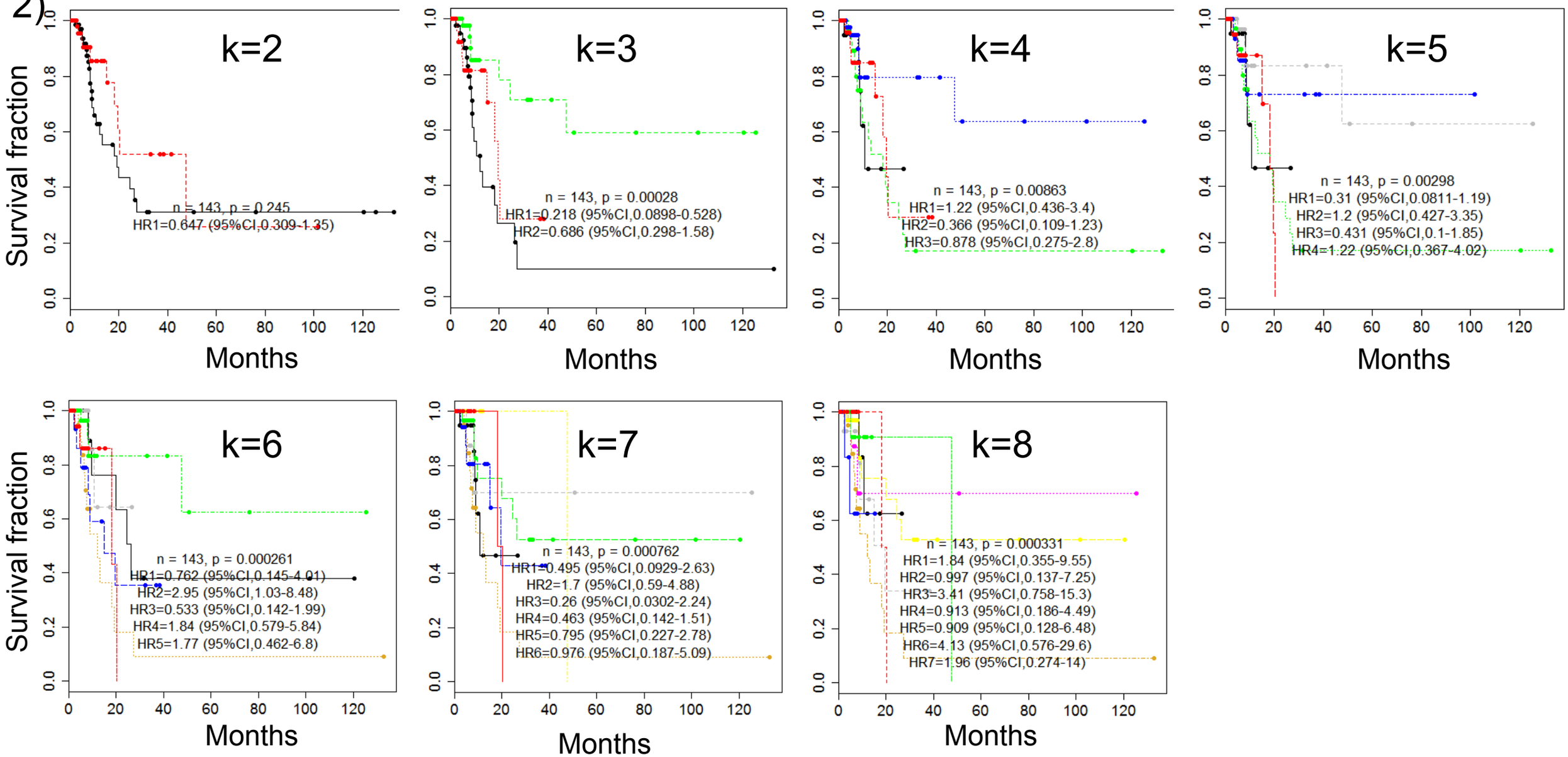
Supplementary Figure 3
KM plots

a. TCGA BLCA

1) k=2 k=3 k=4 k=5 k=6 k=7 k=8

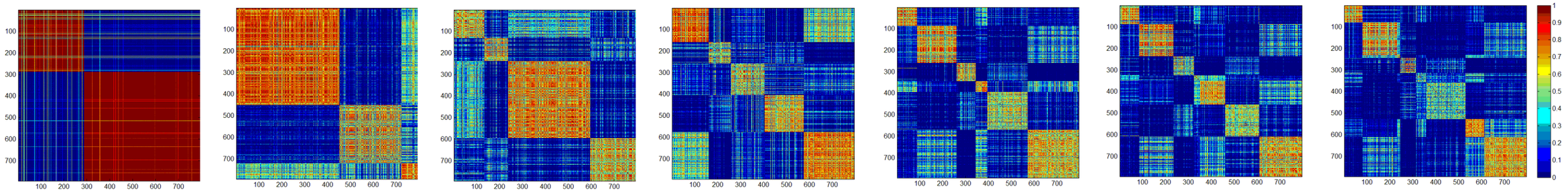


2)

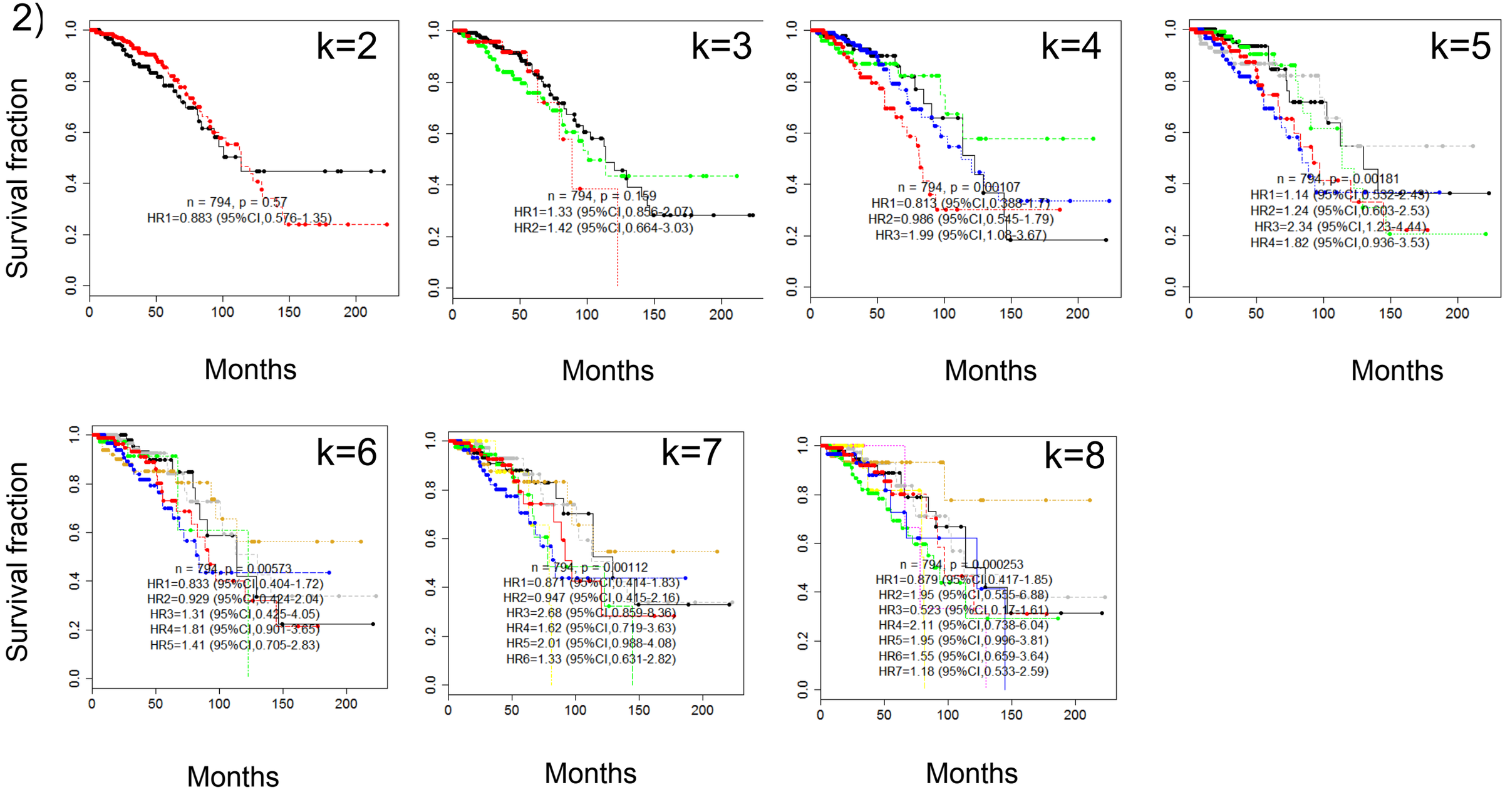


b. TCGA BRCA

1)

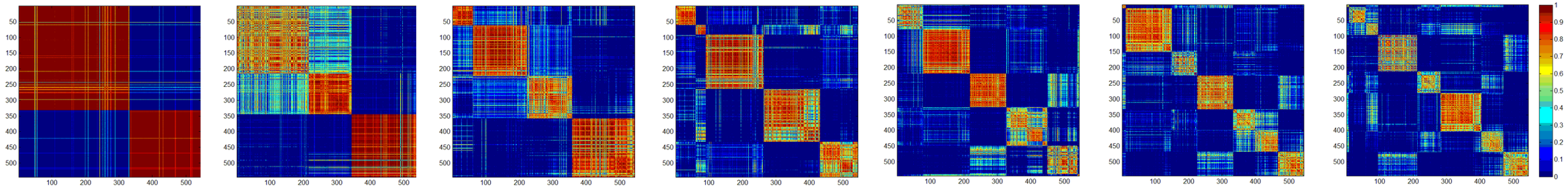


2)

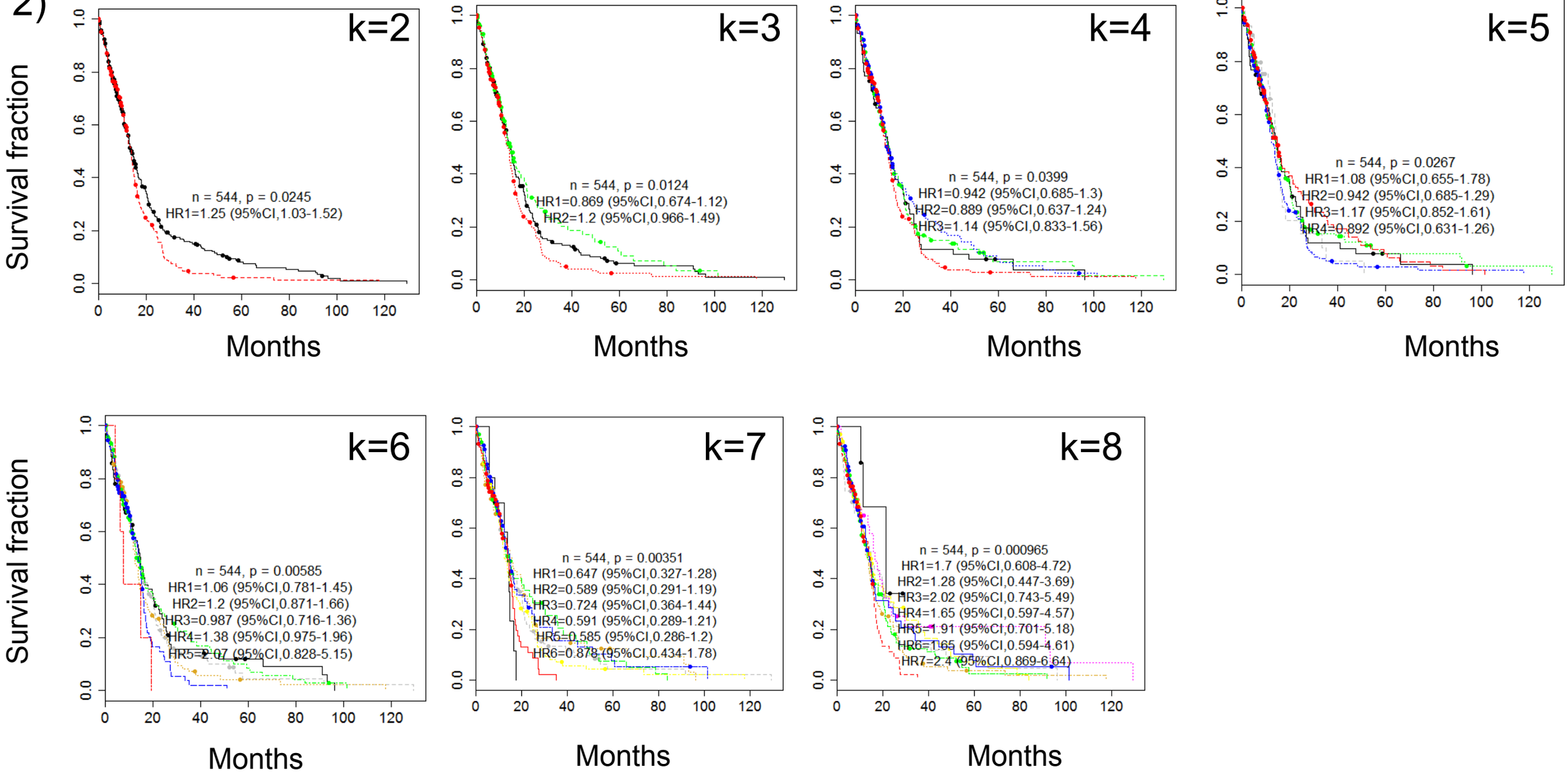


c. TCGA GBM

1) k=2 k=3 k=4 k=5 k=6 k=7 k=8

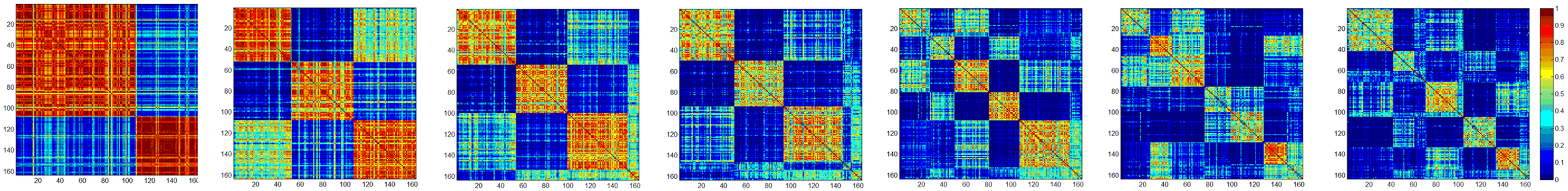


2)

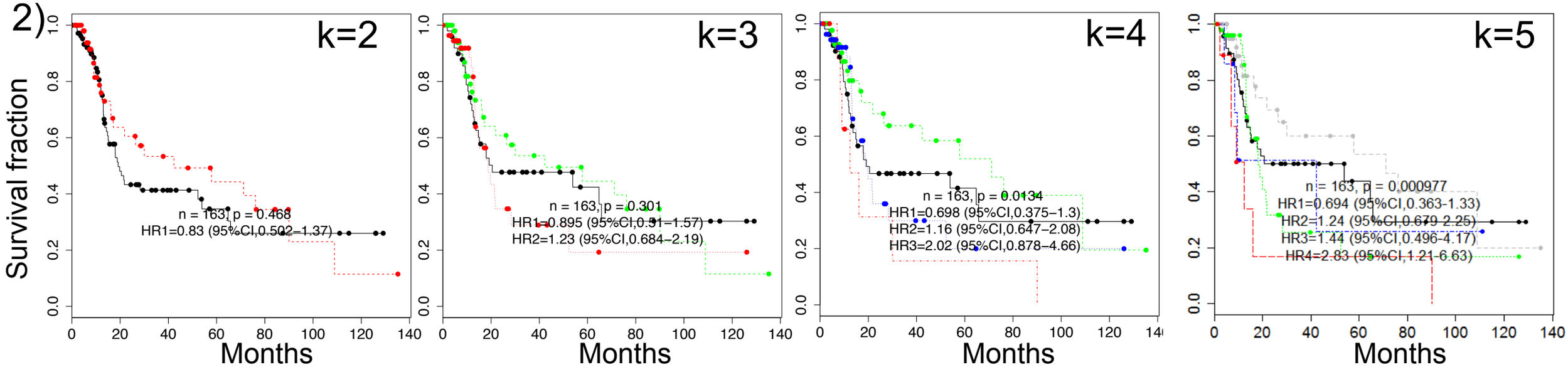


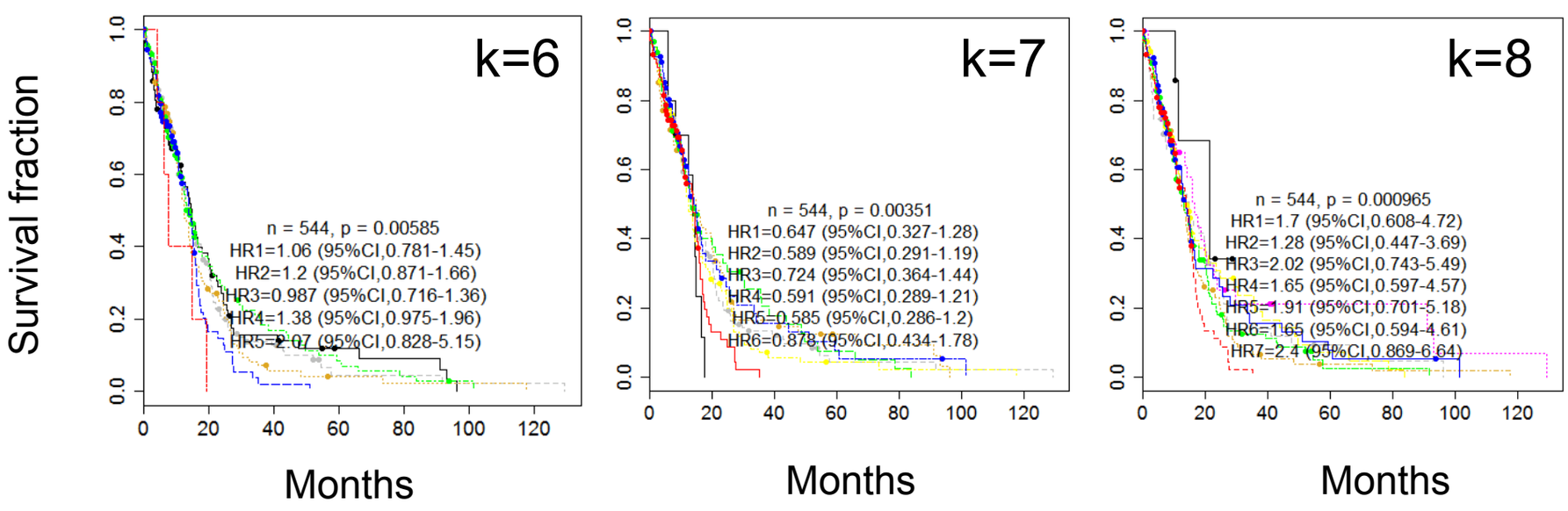
d.TCGA HNSC

1) k=2 k=3 k=4 k=5 k=6 k=7 k=8



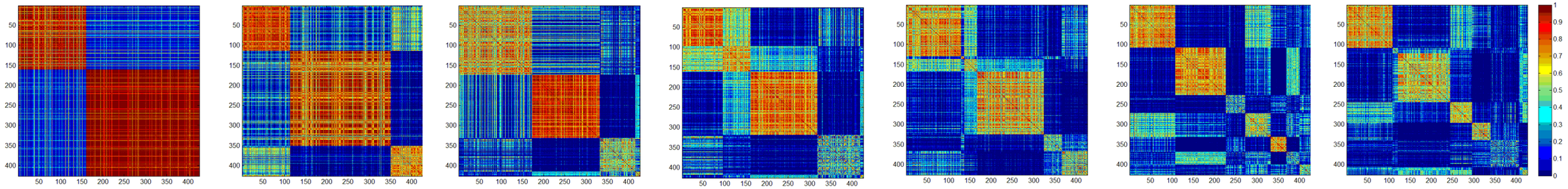
2)



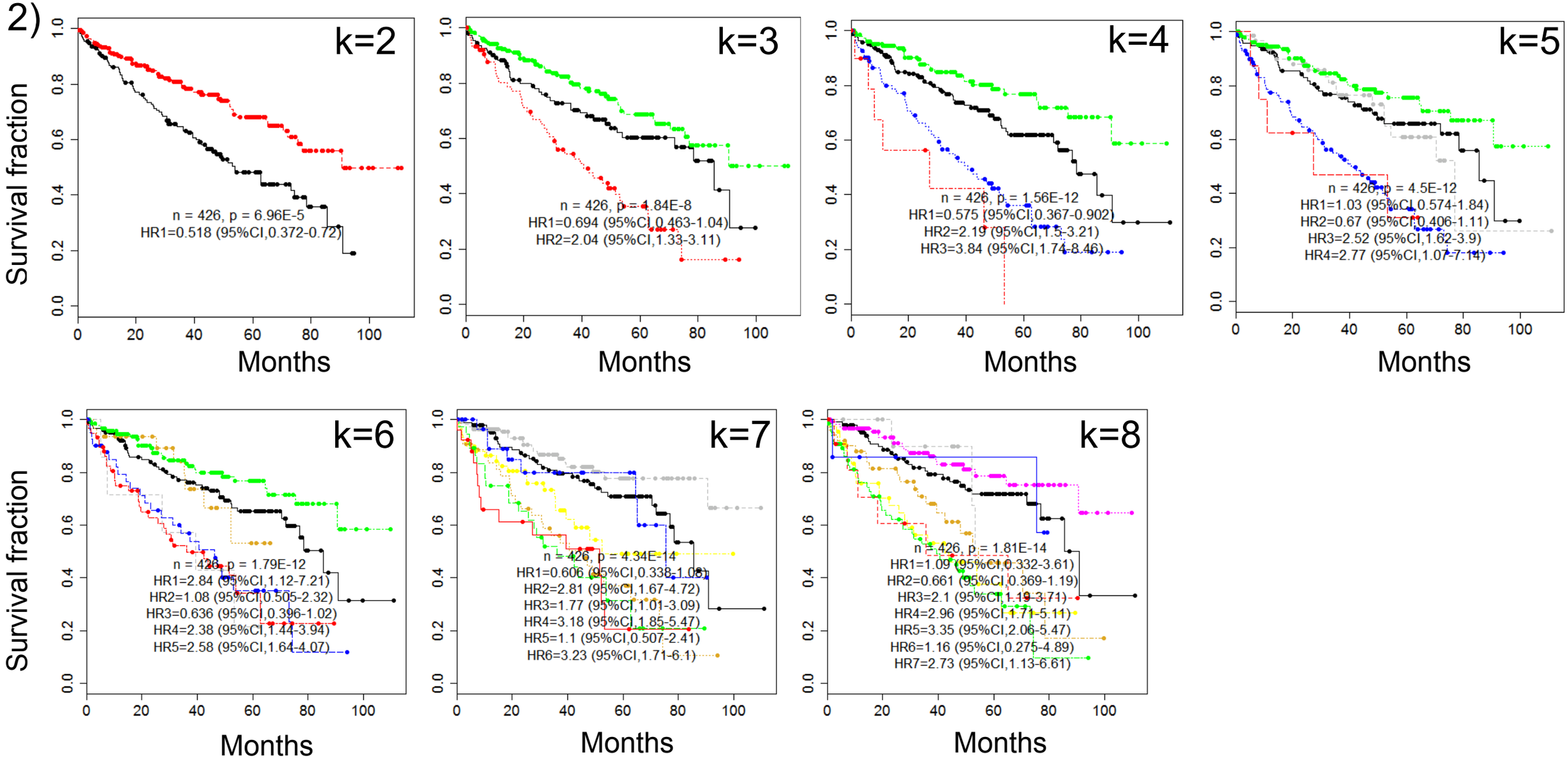


e. TCGA KIRC

1)

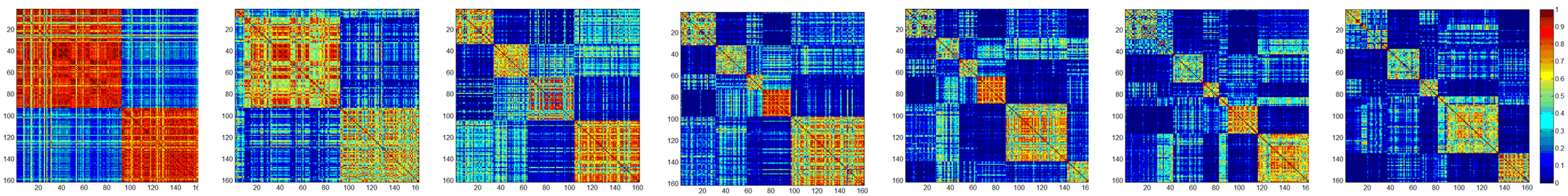


2)

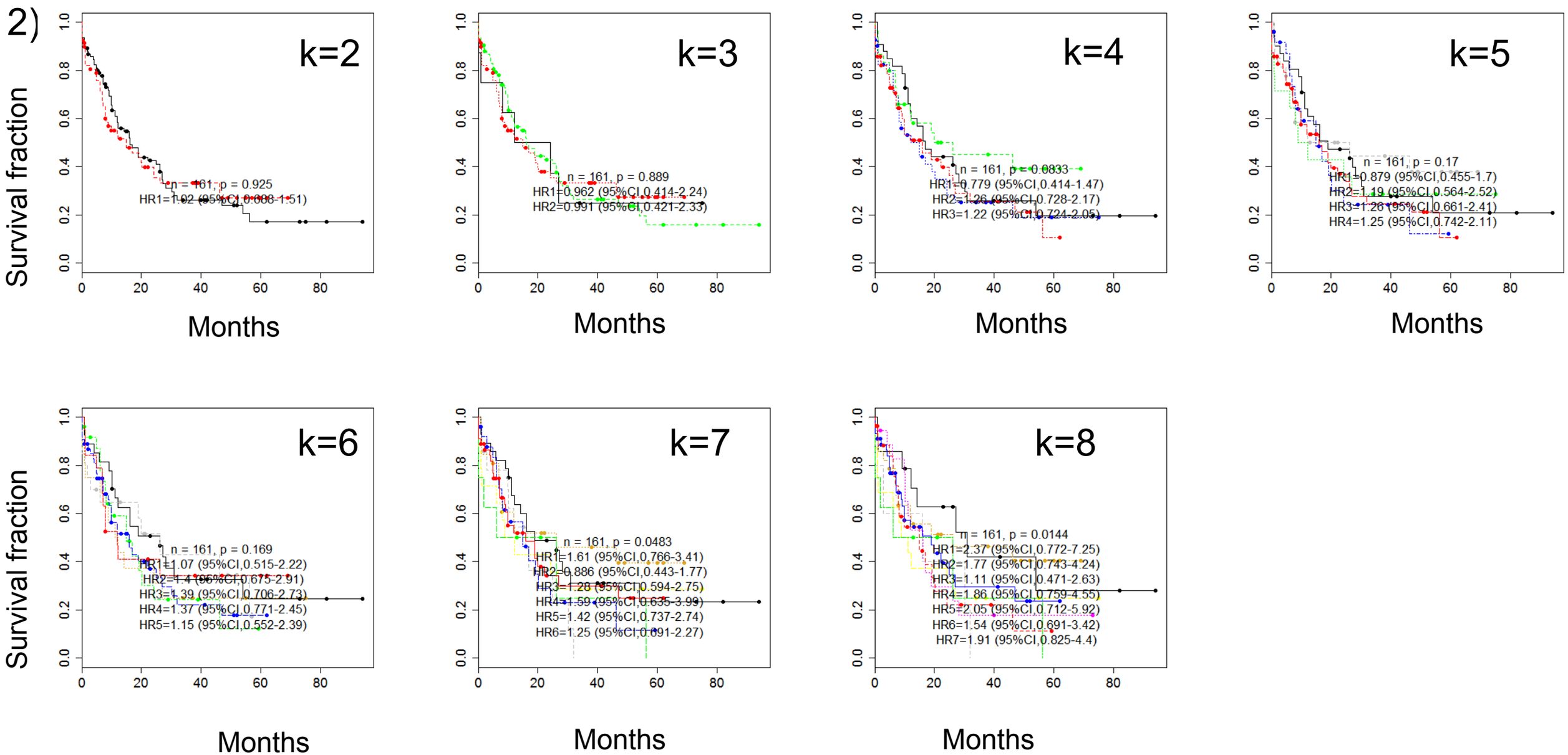


f. TCGA LAML

1)

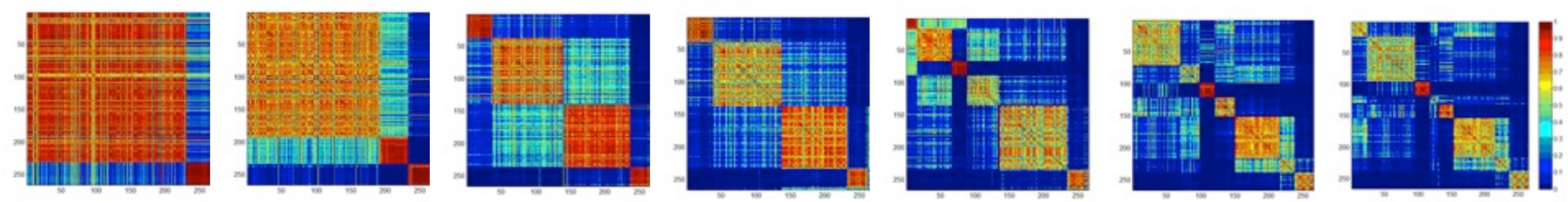


2)

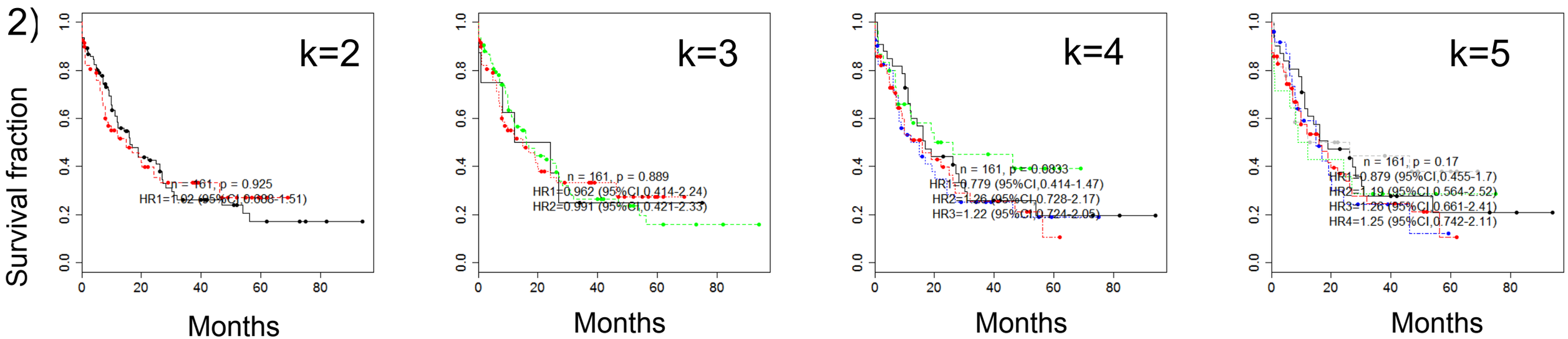


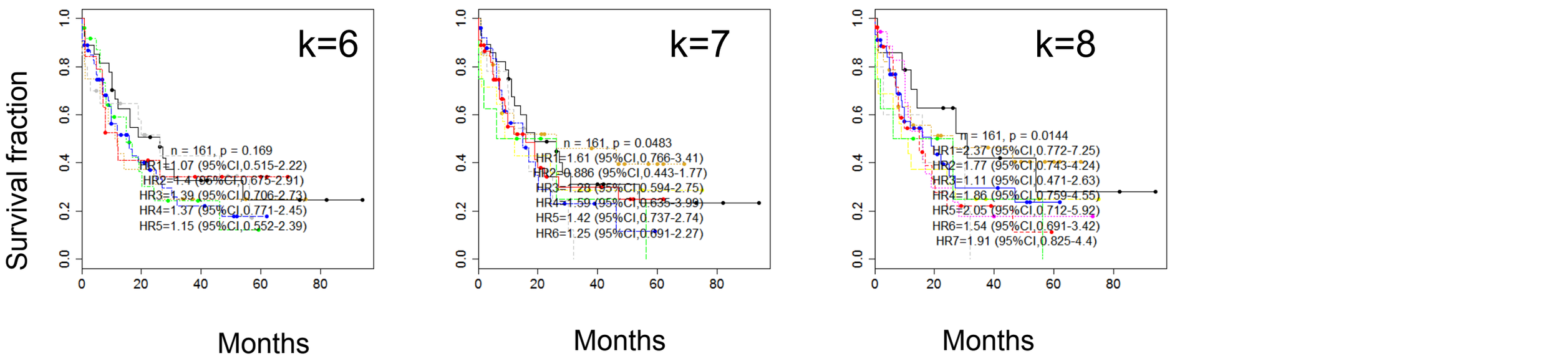
g. TCGA LGG

1) k=2 k=3 k=4 k=5 k=6 k=7 k=8



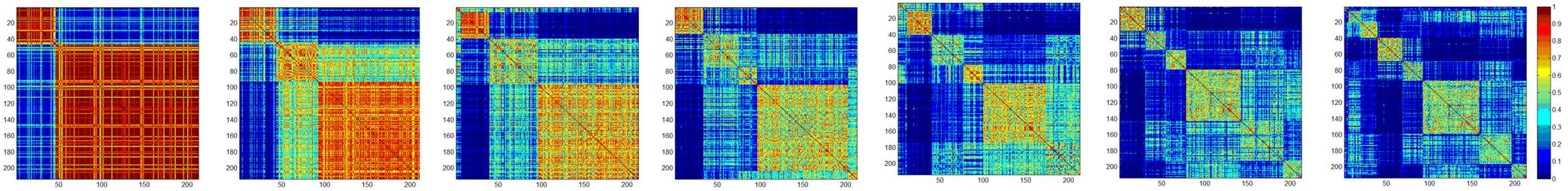
2)



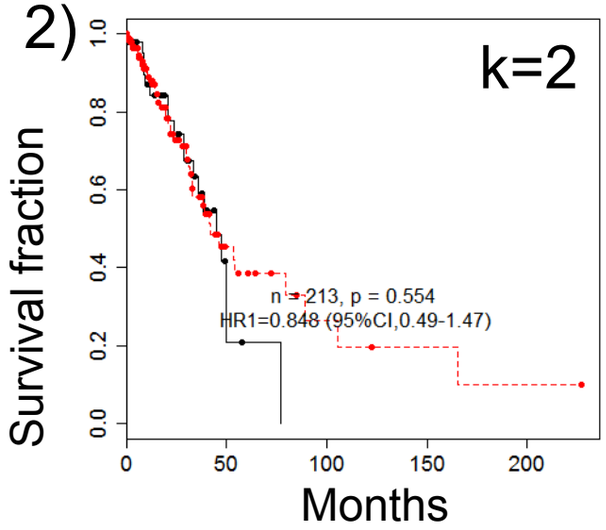


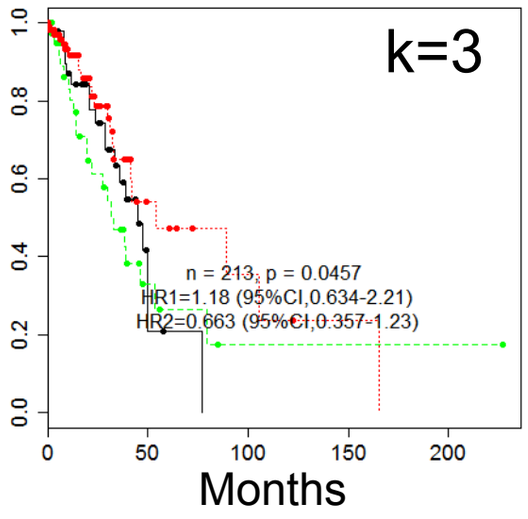
h.TCGA LUAD

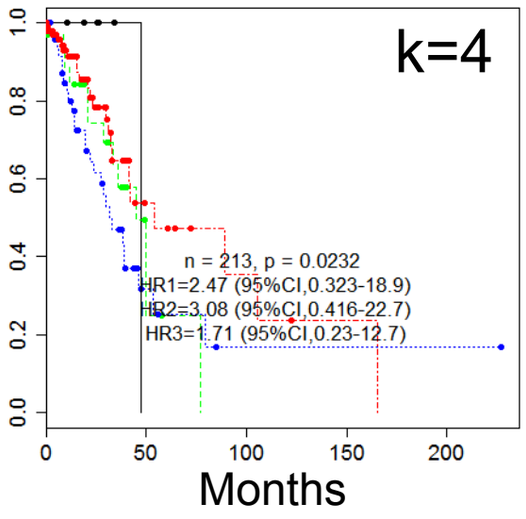
1) k=2 k=3 k=4 k=5 k=6 k=7 k=8

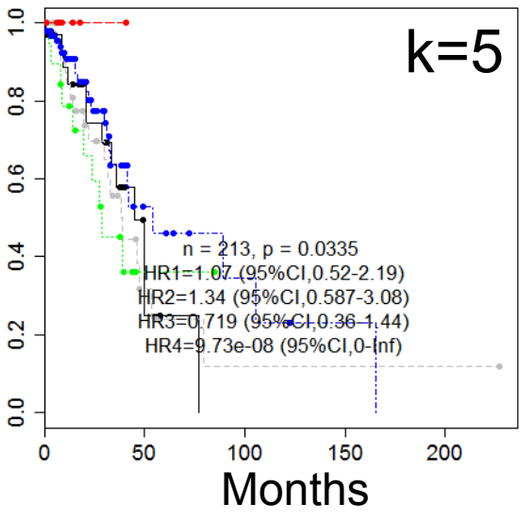


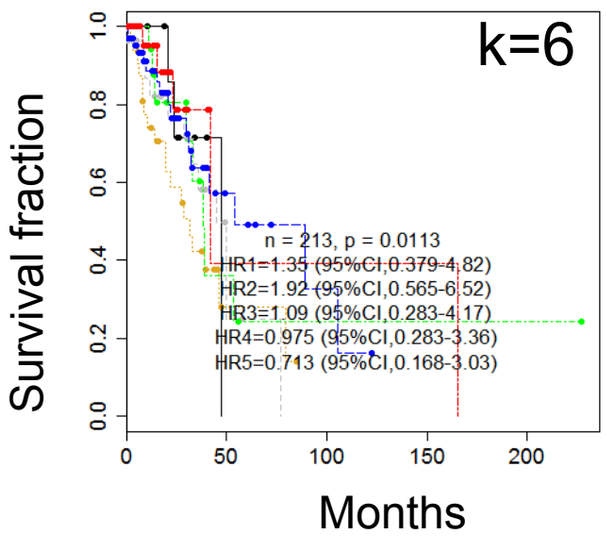
2)

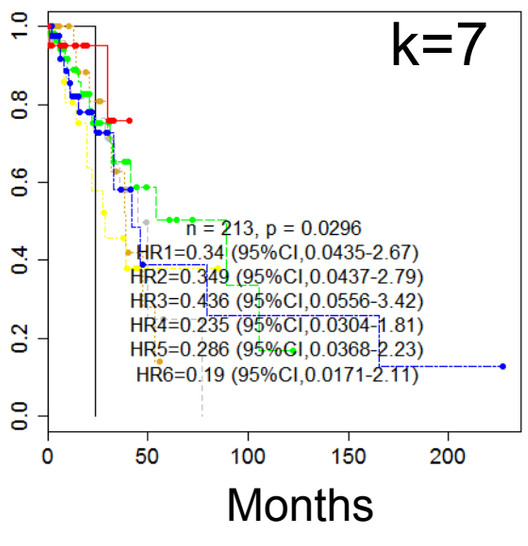


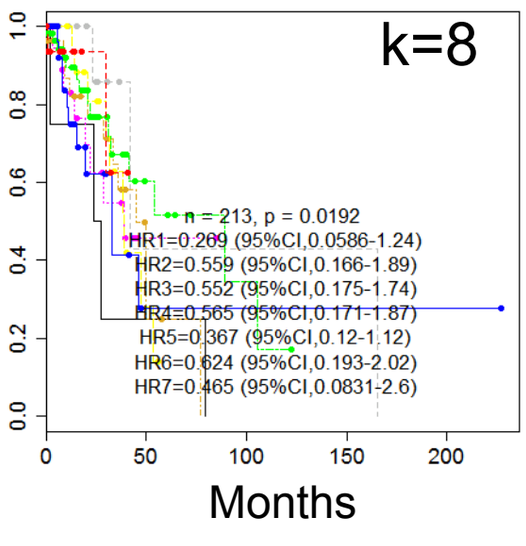






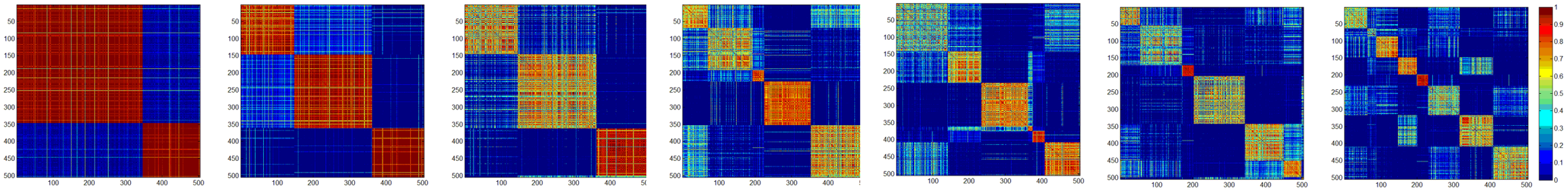




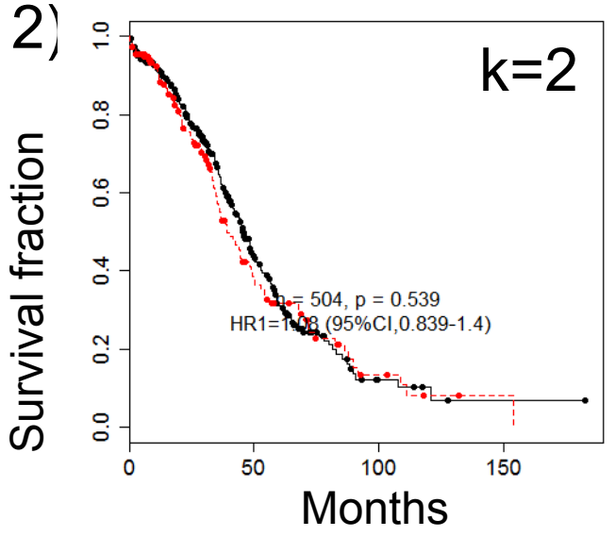


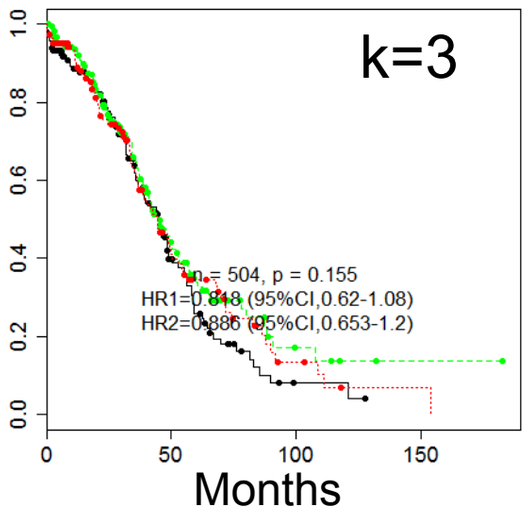
i. TCGA OV

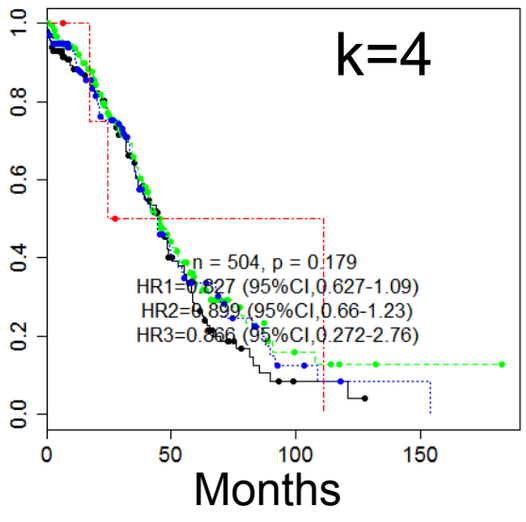
1) k=2 k=3 k=4 k=5 k=6 k=7 k=8

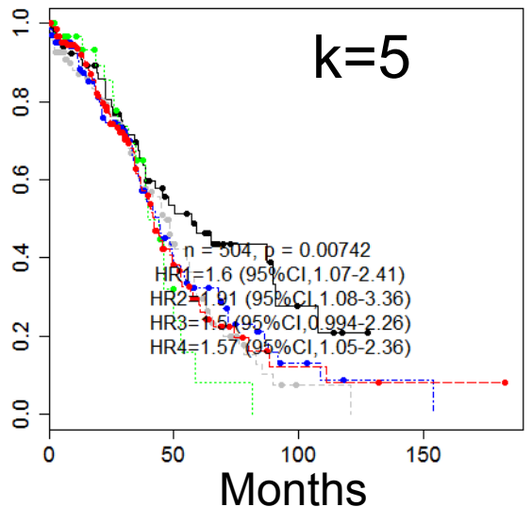


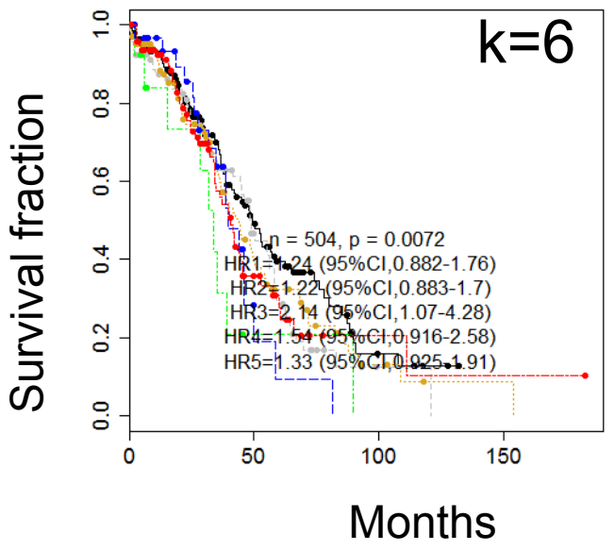
2)

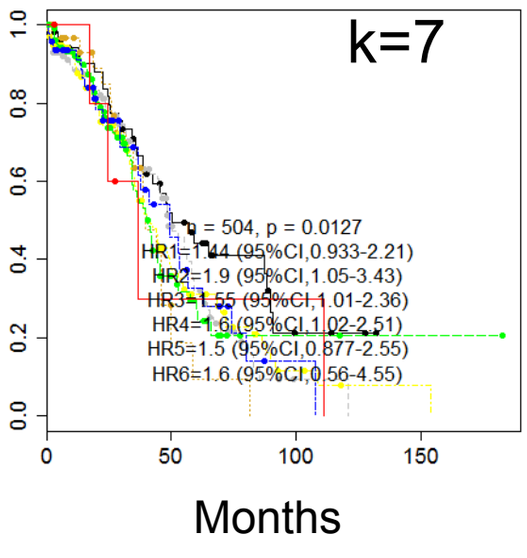


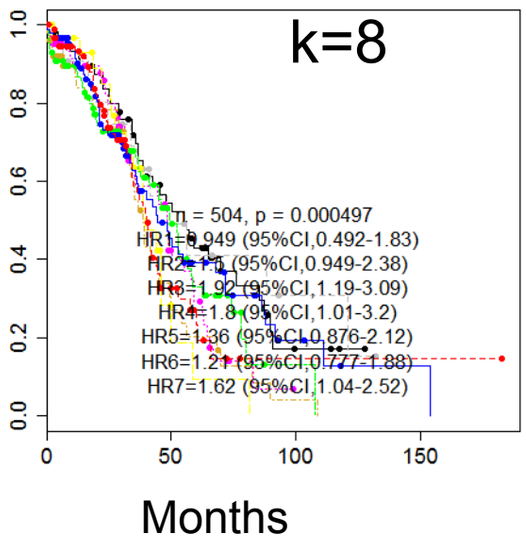






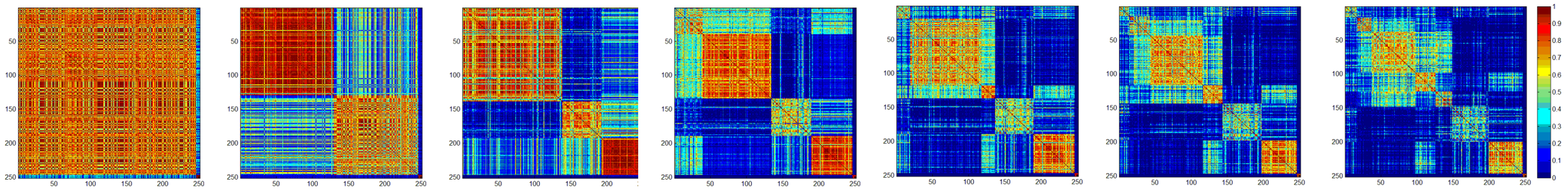




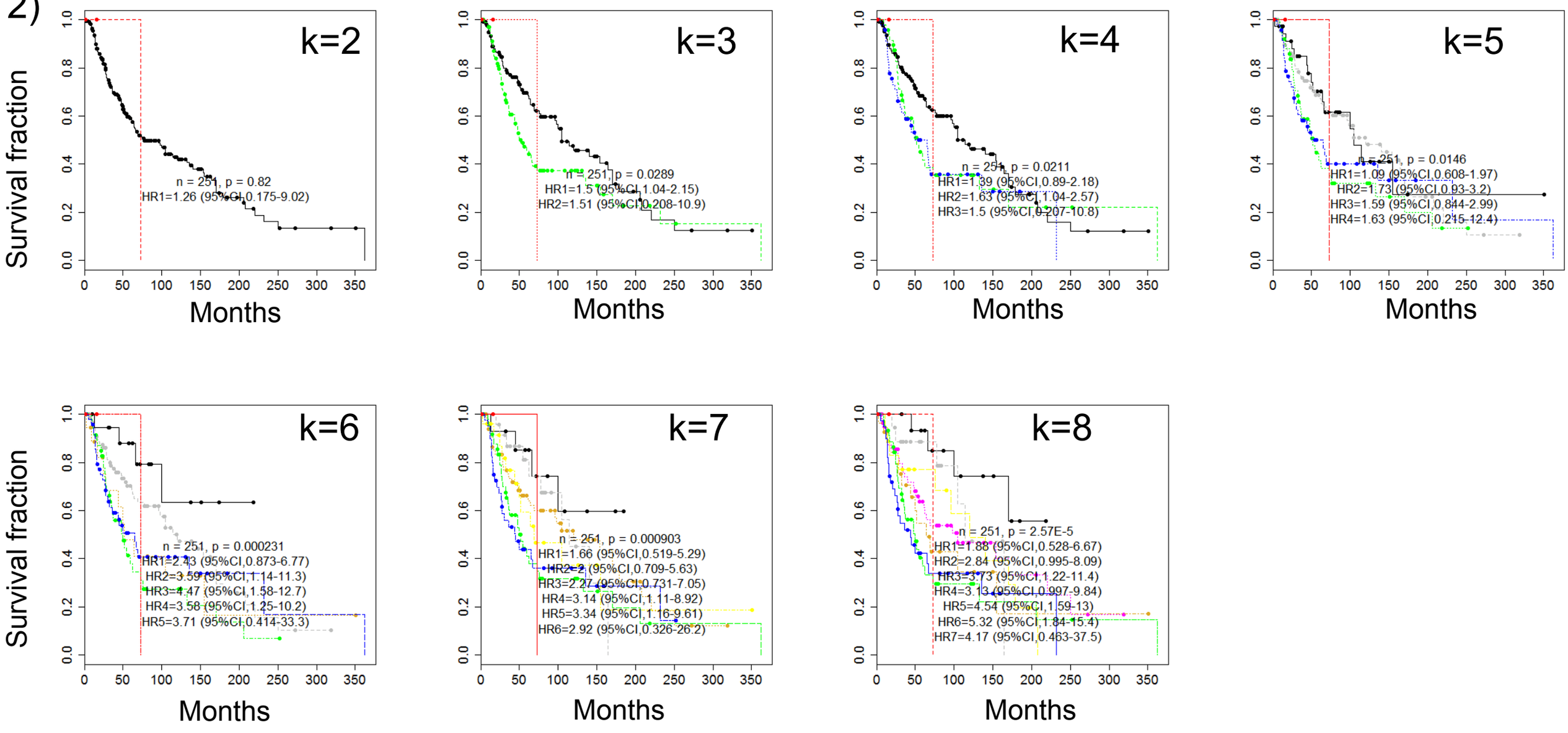


j.TCGA SKCM

1)



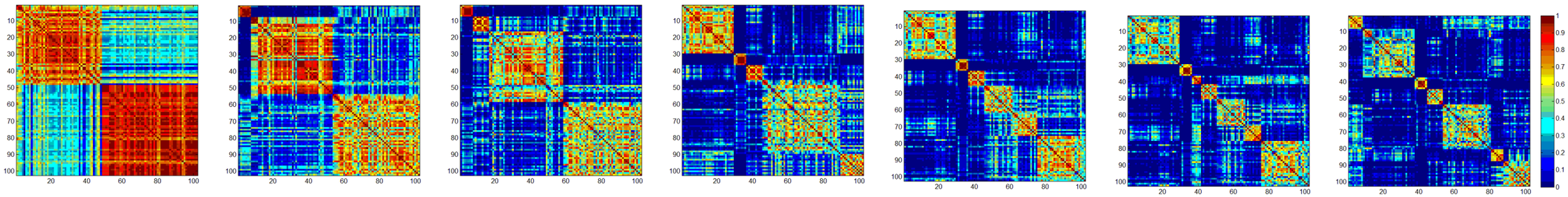
2)



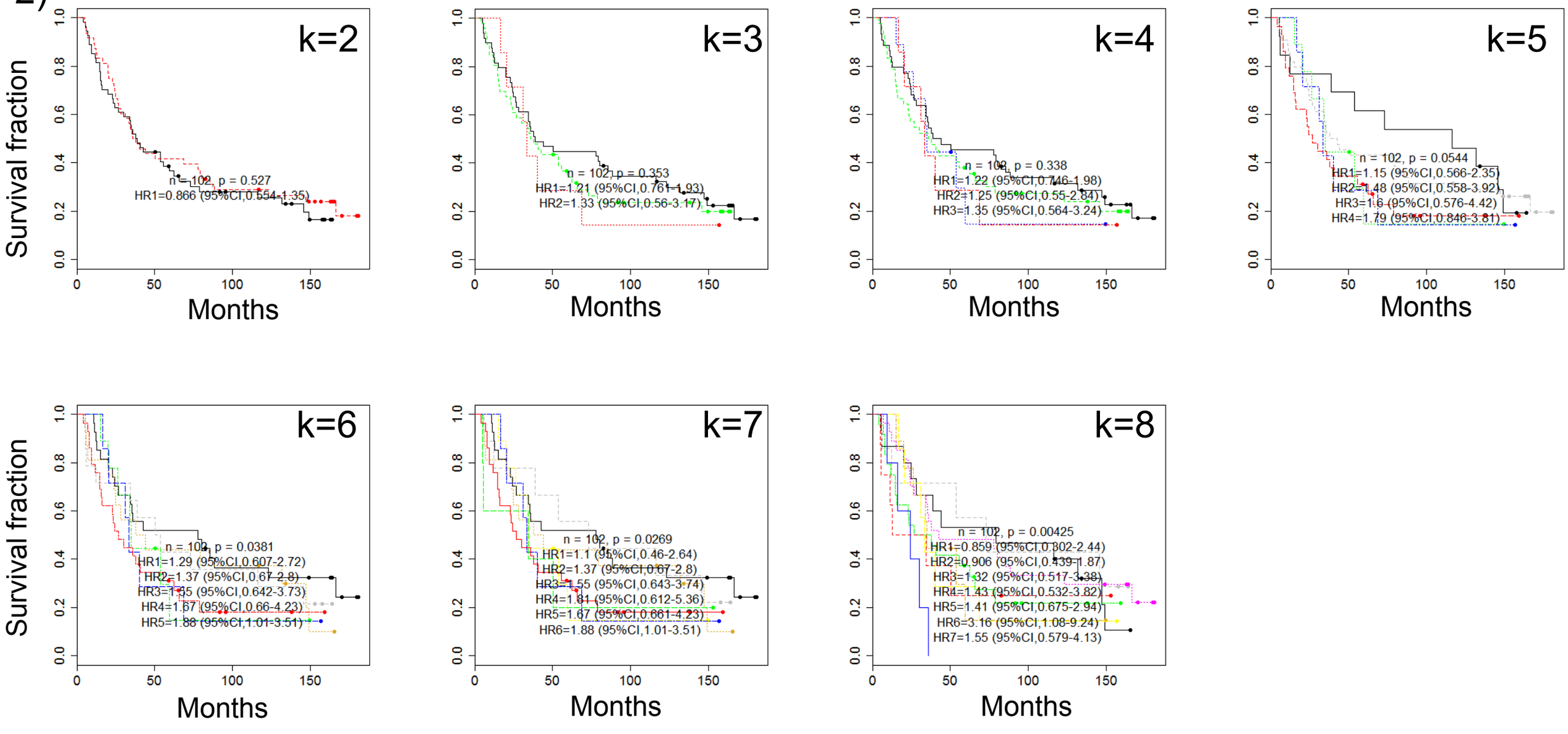
Supplementary Figure 4
Independent dataset

a. UTSW HNSC

1)

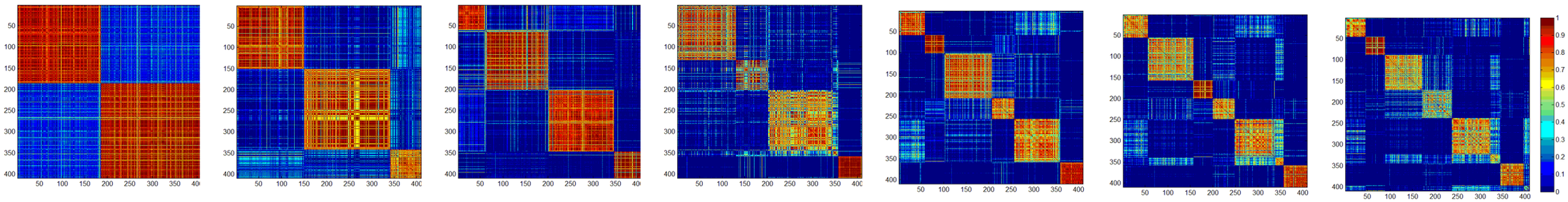


2)

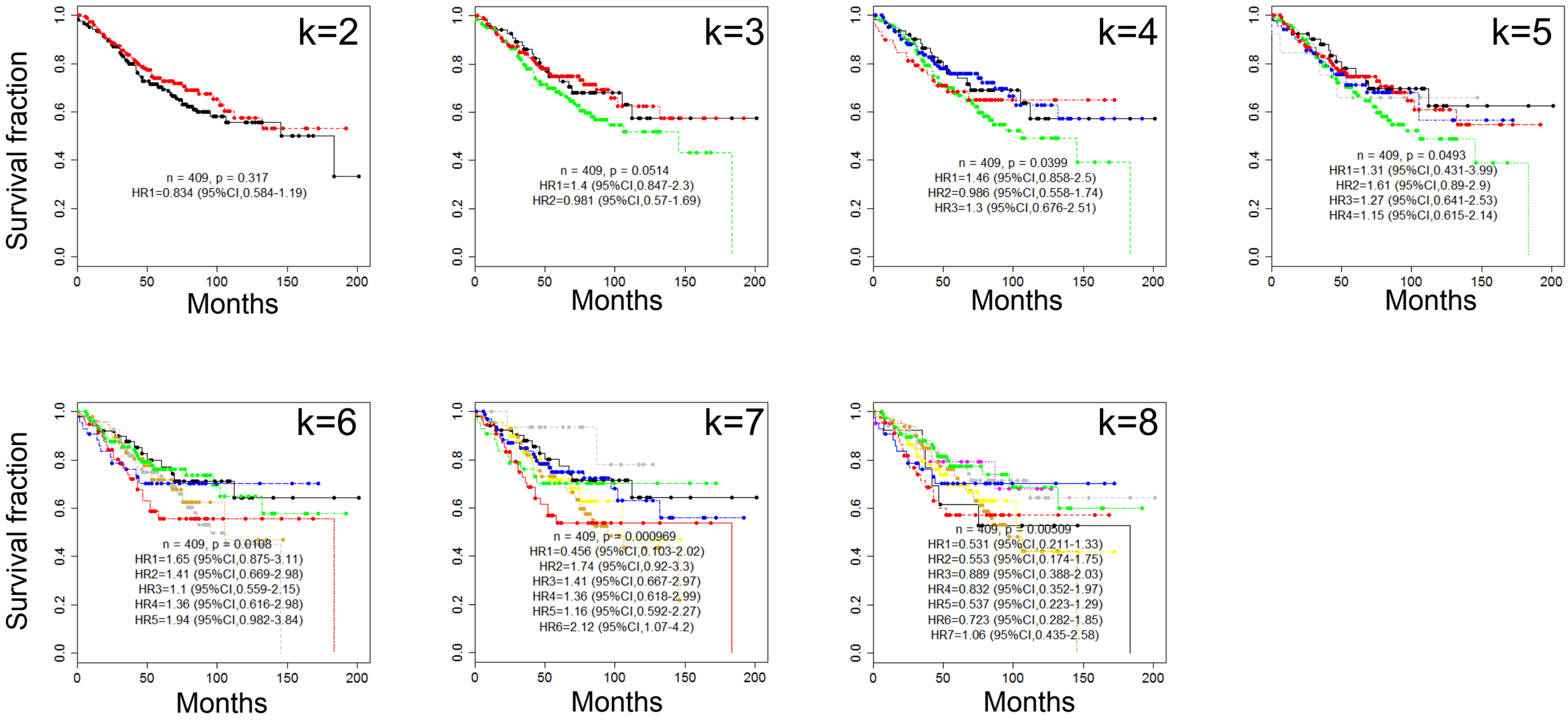


b. Colon cancer (GSE39582)

a

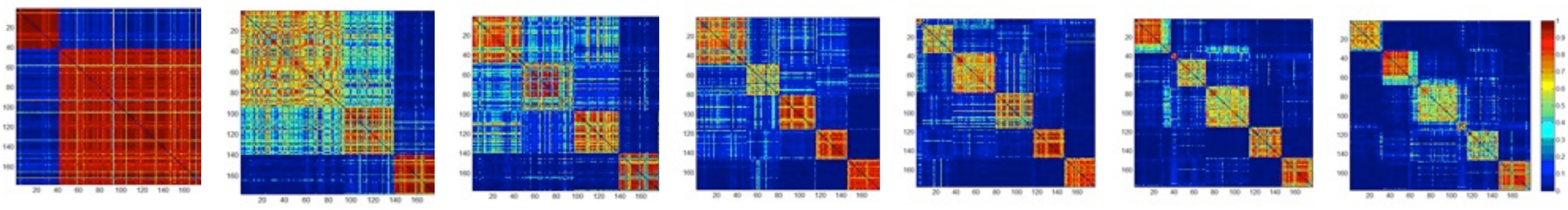


b

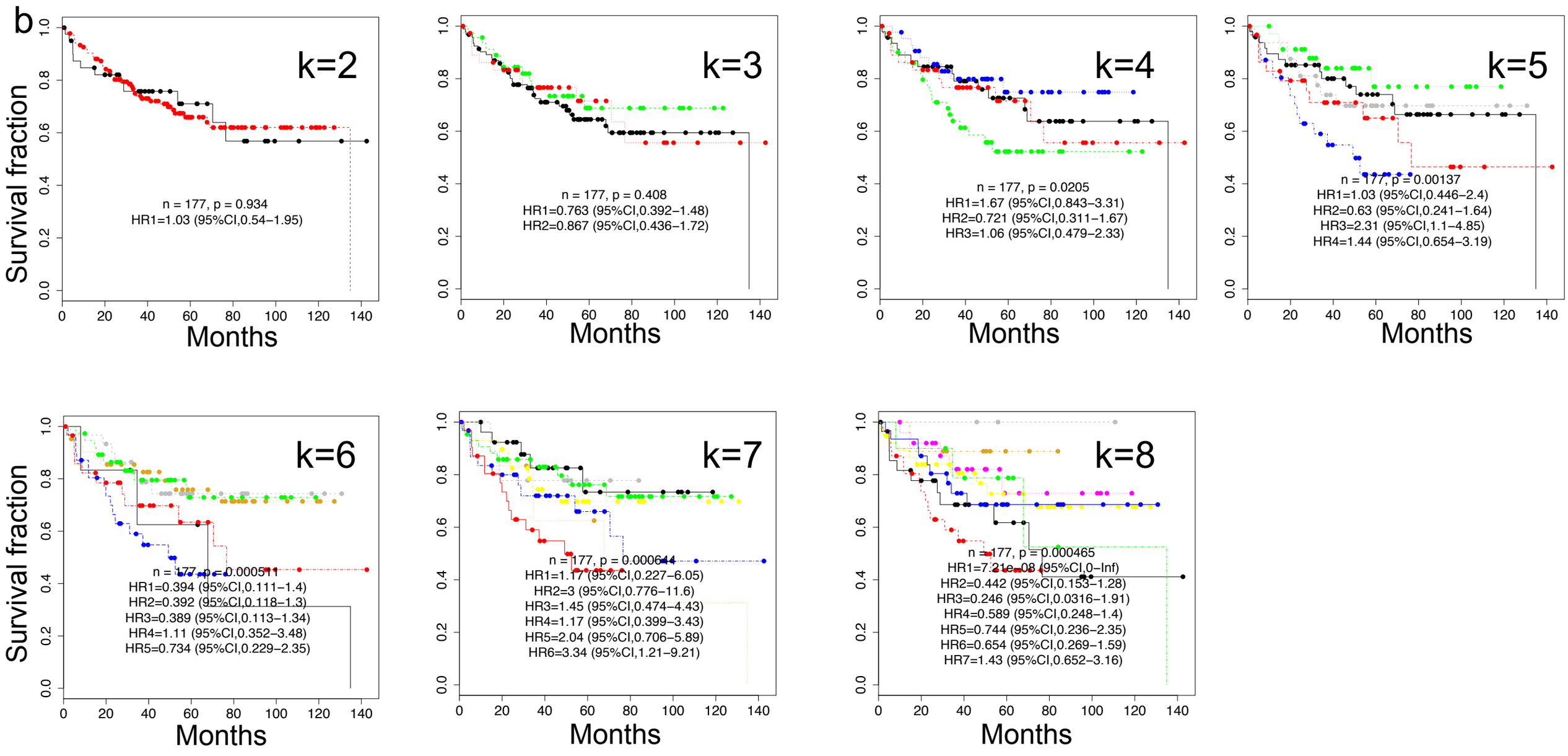


b. Colon cancer (GSE17538)

a

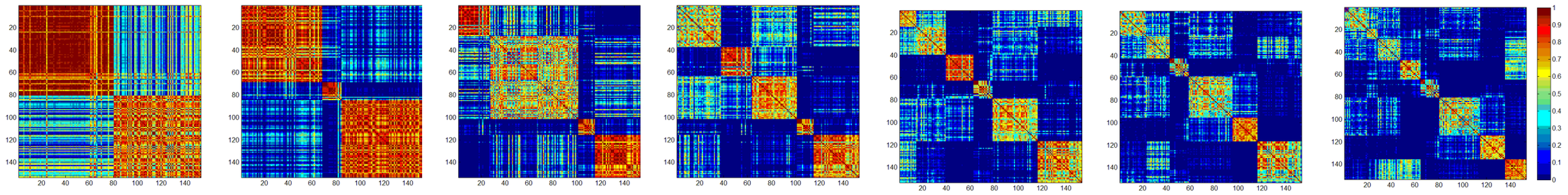


b

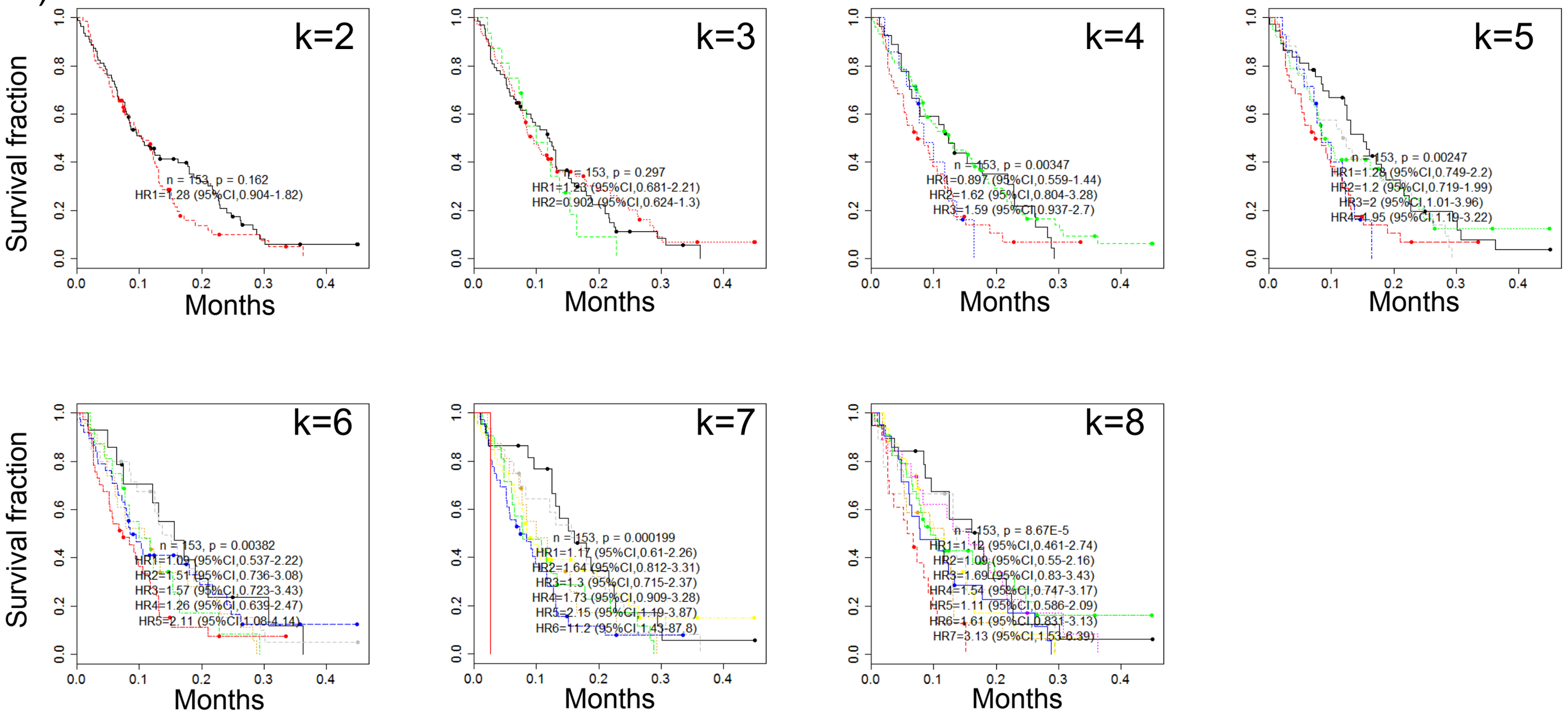


c. OV (Bonome et. al.,Cancer Research 2008,)

1) k=2 k=3 k=4 k=5 k=6 k=7 k=8

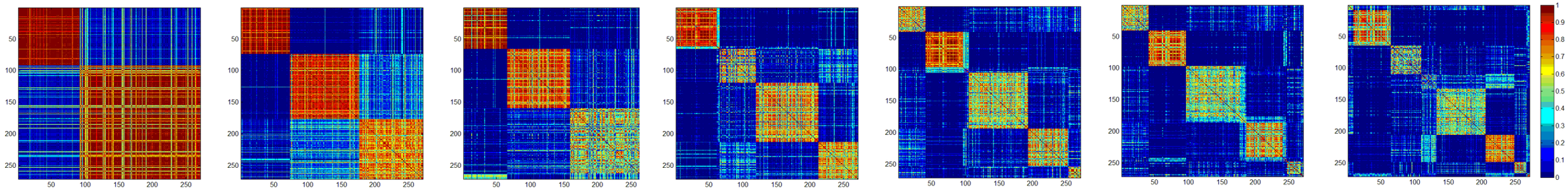


2)

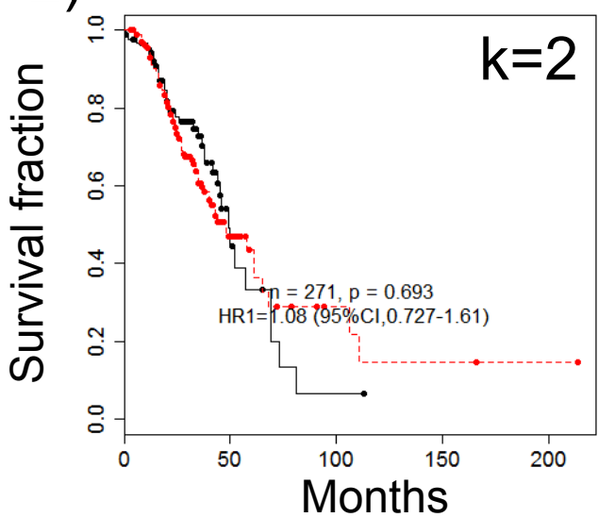


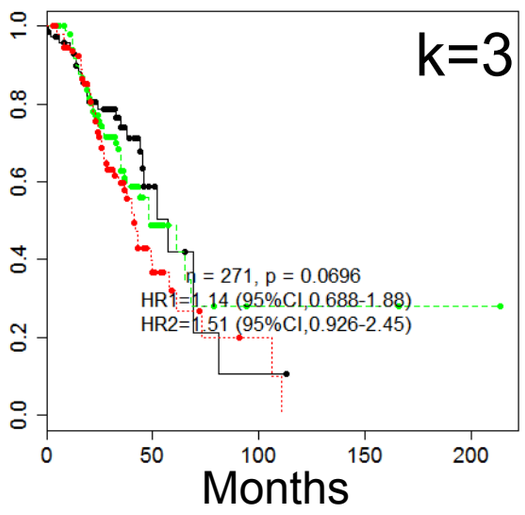
d. OV (Tothill et. al., Clinical Cancer Research, 2008)

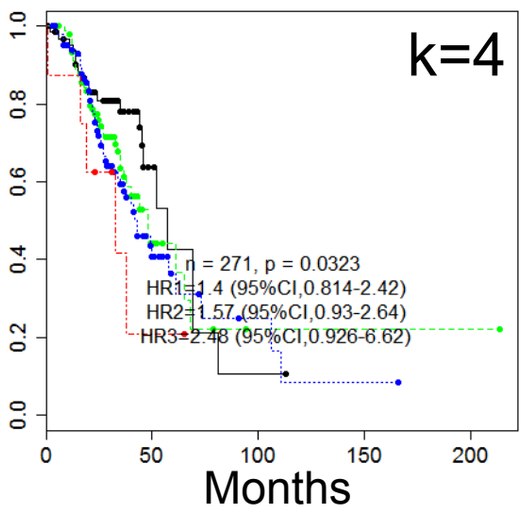
1)

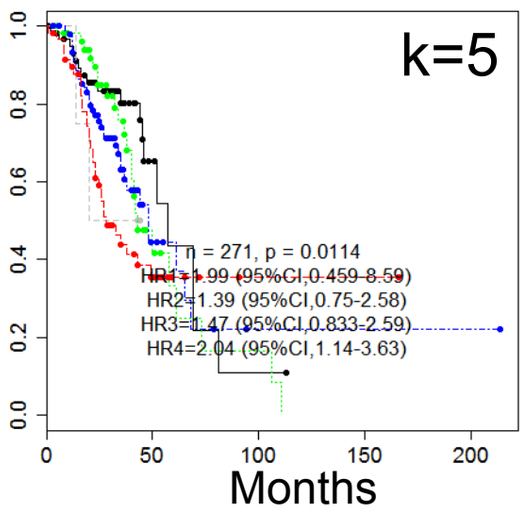


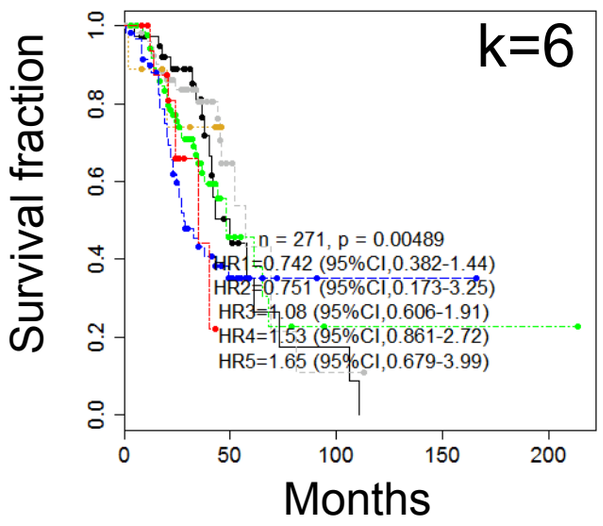
2)

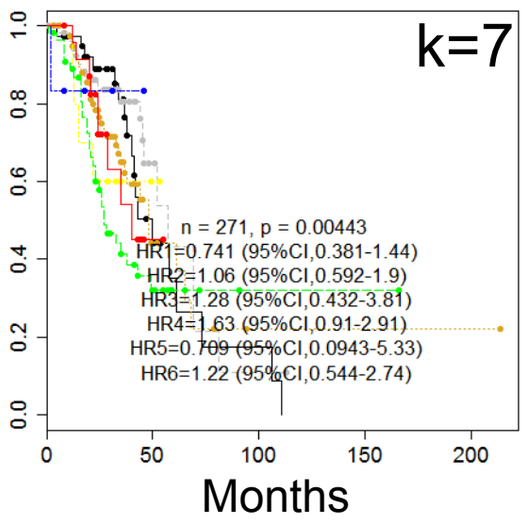


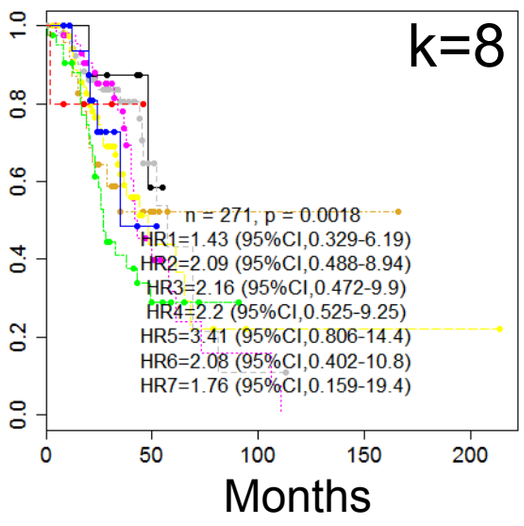






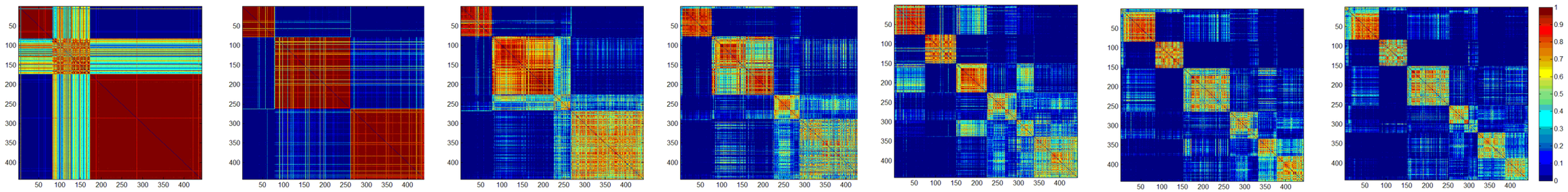




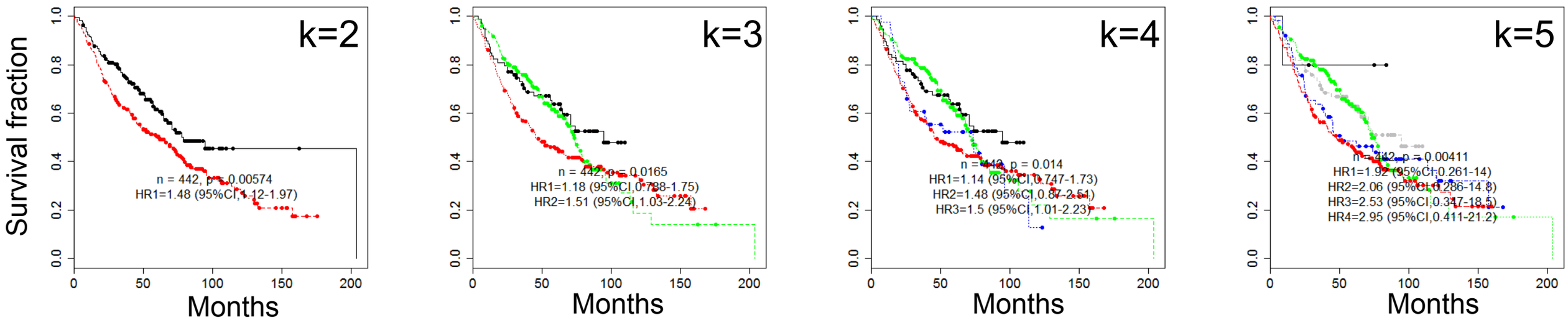


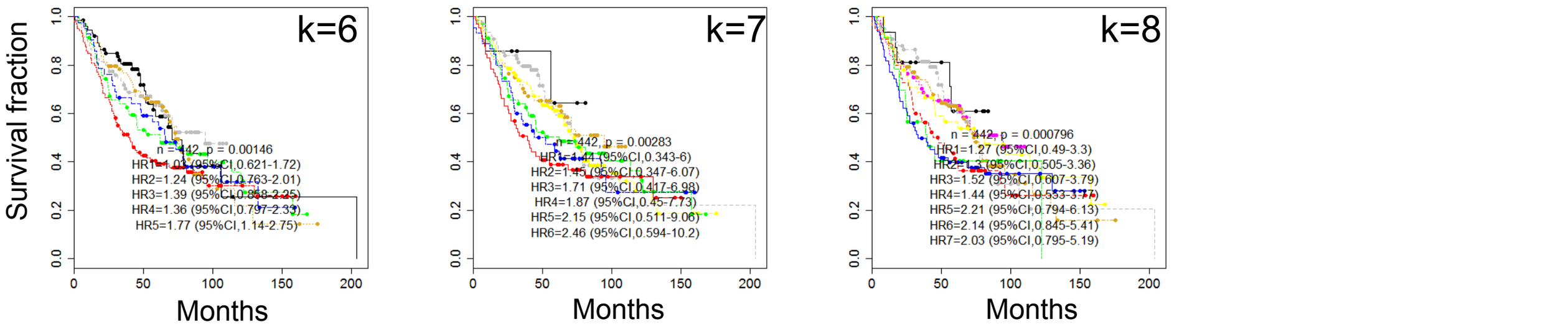
e. Lung adenocarcinoma (Shedden et. al., Nature Methods, 2008)

1) k=2 k=3 k=4 k=5 k=6 k=7 k=8



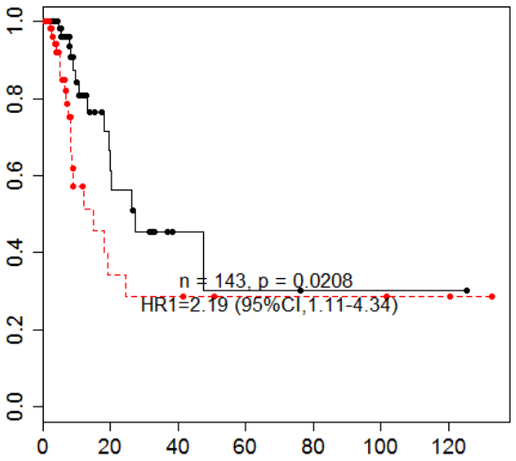
2)



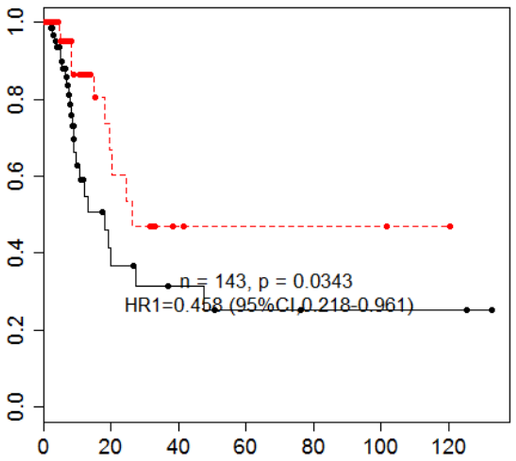


Supplementary Figure 5
Single gene analysis

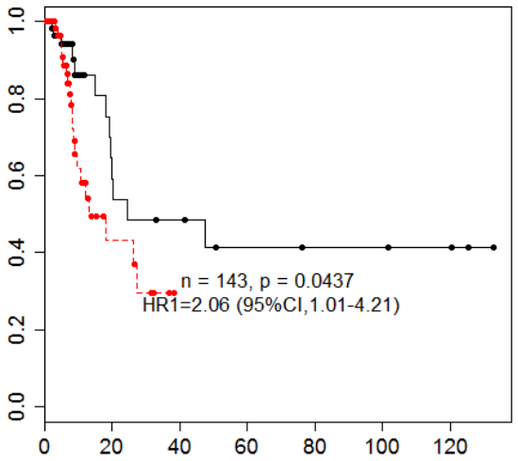
a. TCGA BLCA



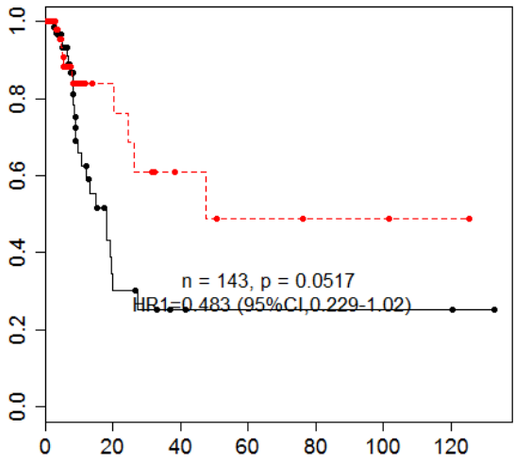
TP53



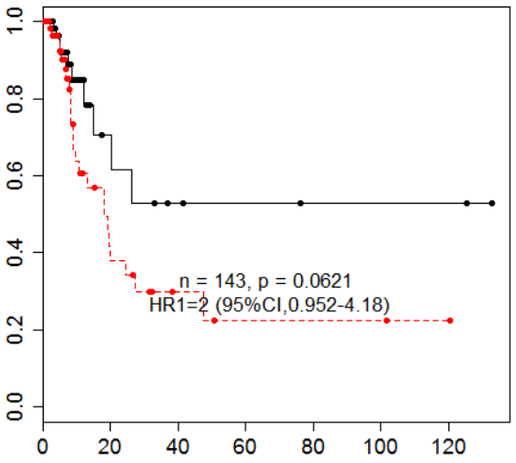
ASCC2



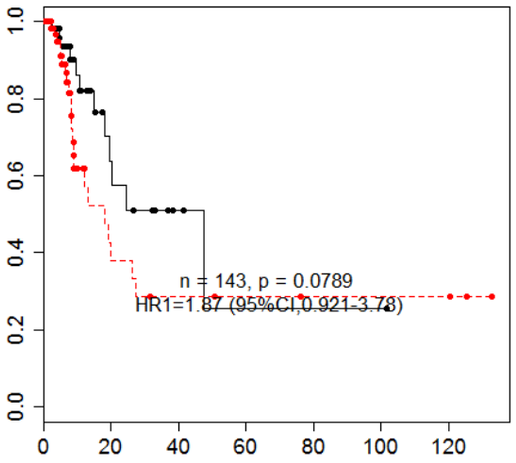
WRN



FHL2

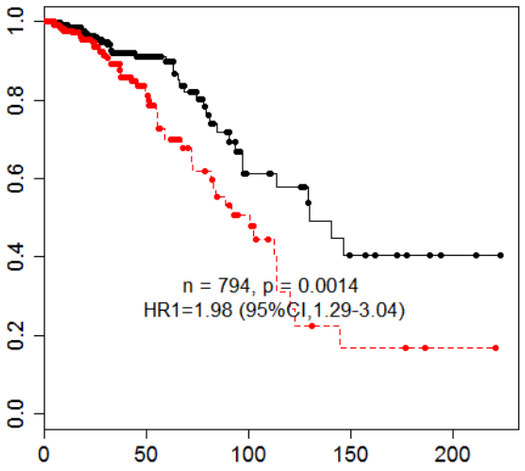


NCOA6

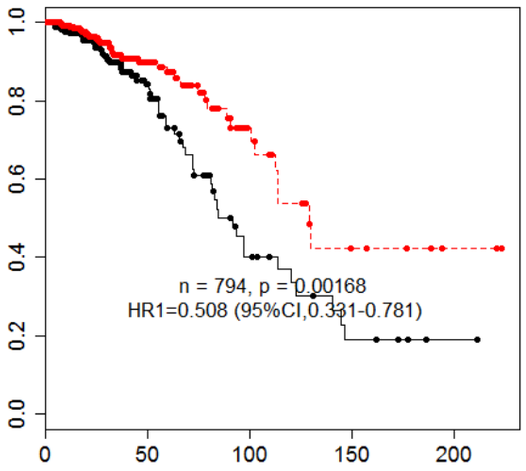


MSH6

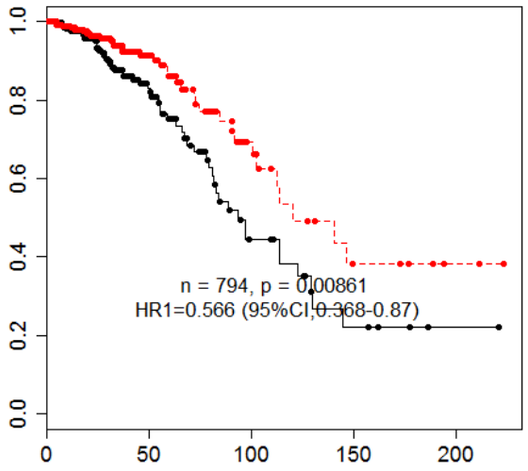
b. TCGA BRCA



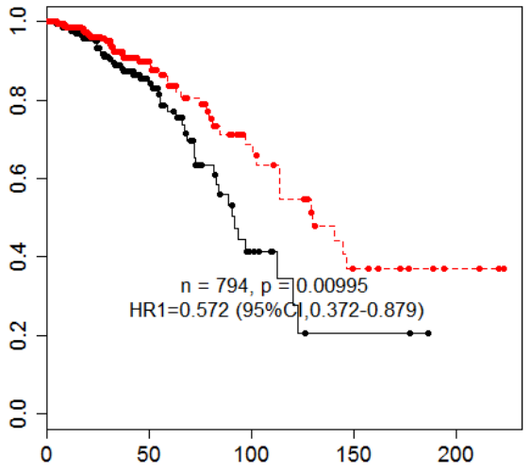
ZHX1



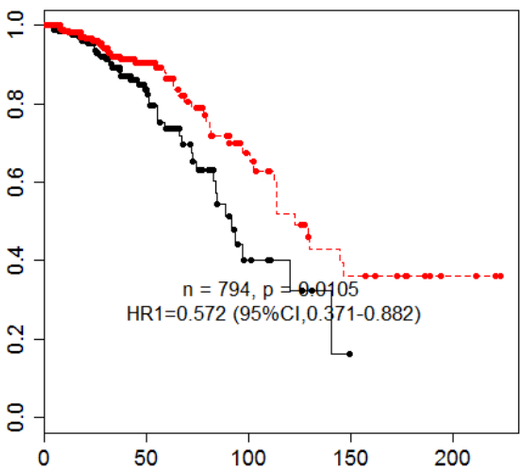
NFKBIA



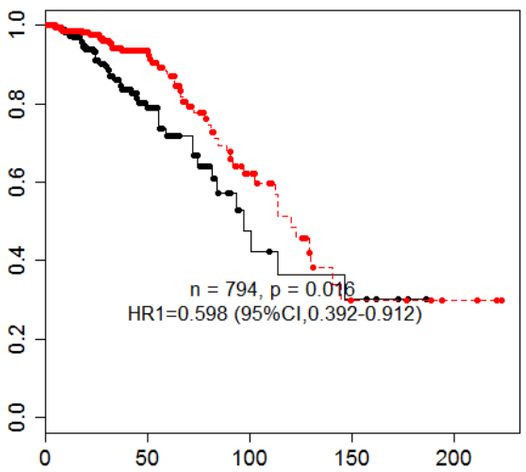
ACTA2



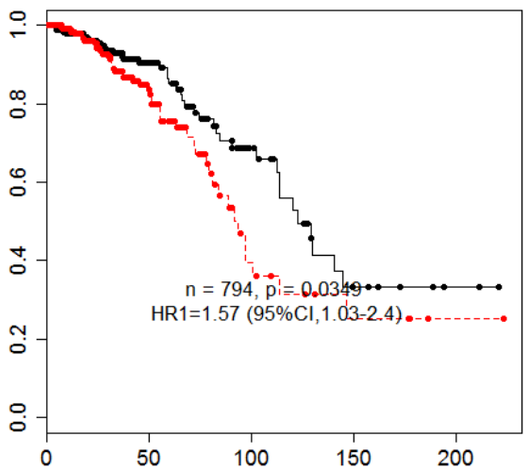
PIN1



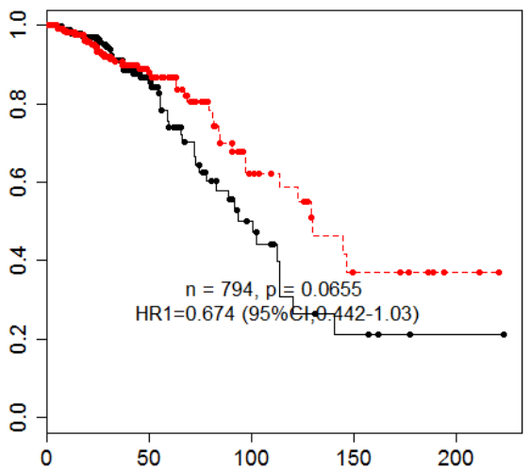
THAP8



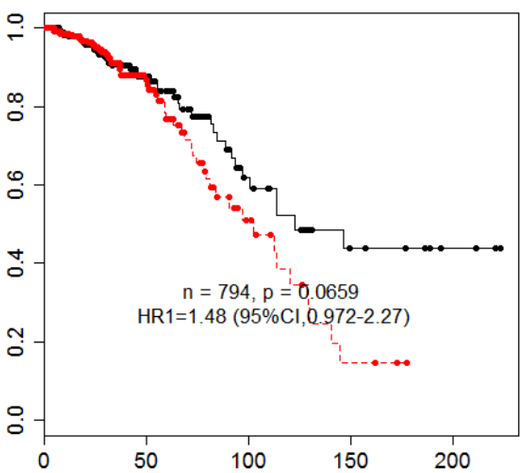
NAT6



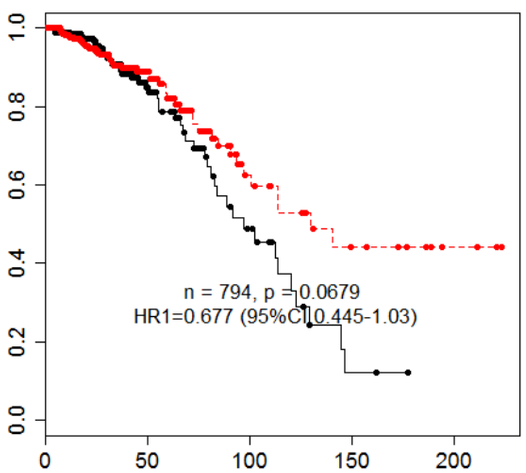
BRCA2



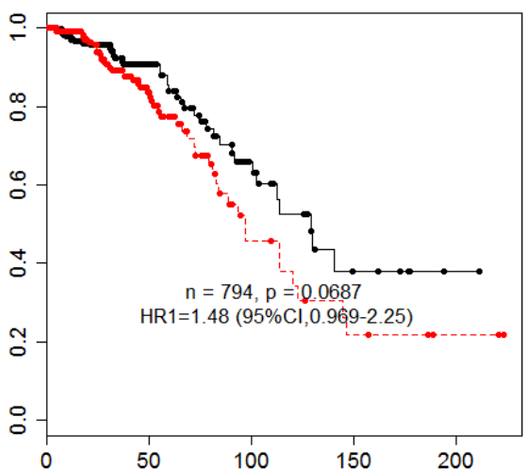
CDK5



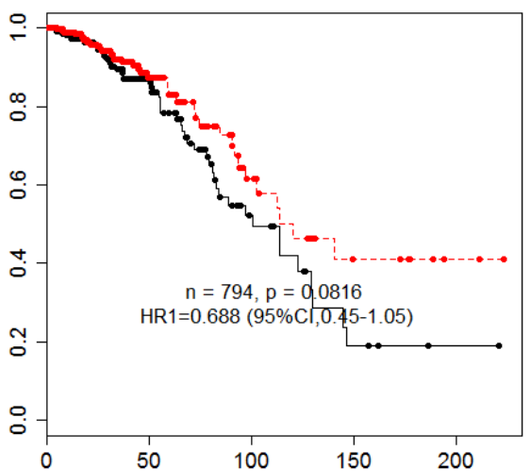
CHD3



YBX1

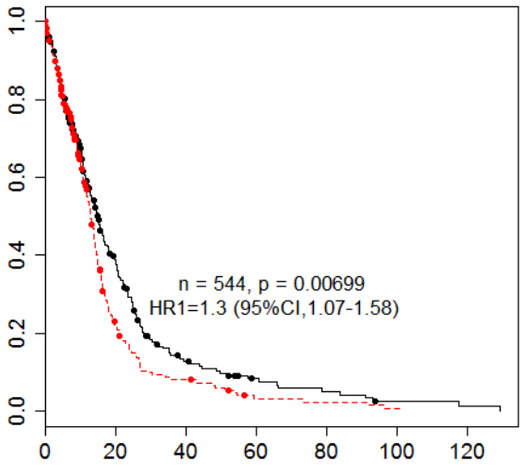


TDRD7

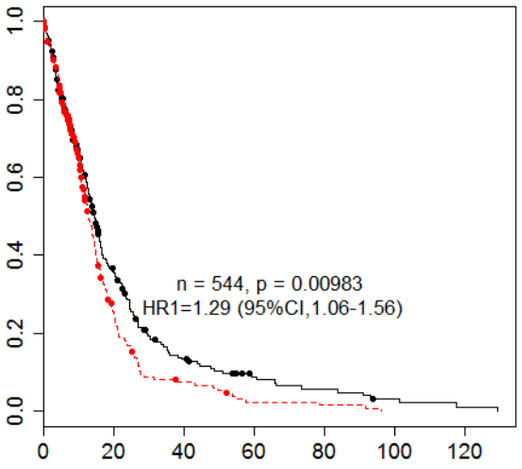


VIM

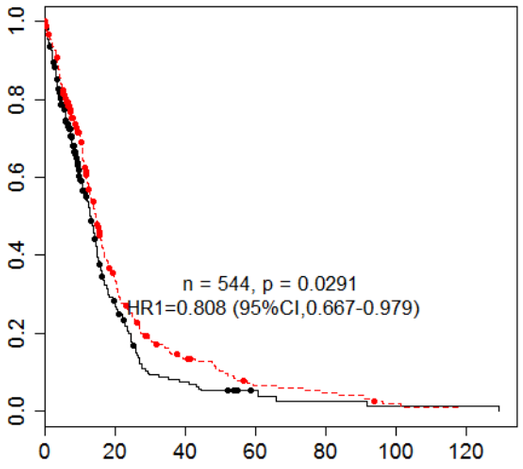
c. TCGA GBM



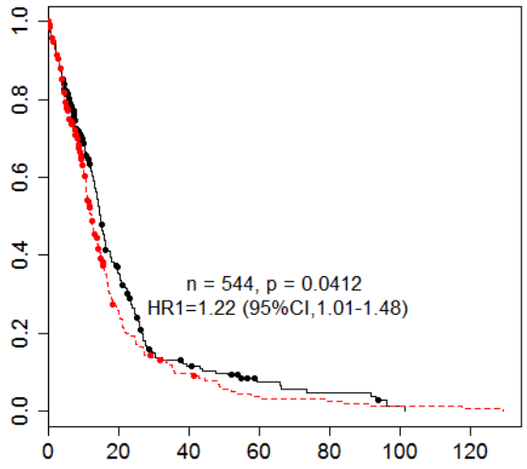
TEP1



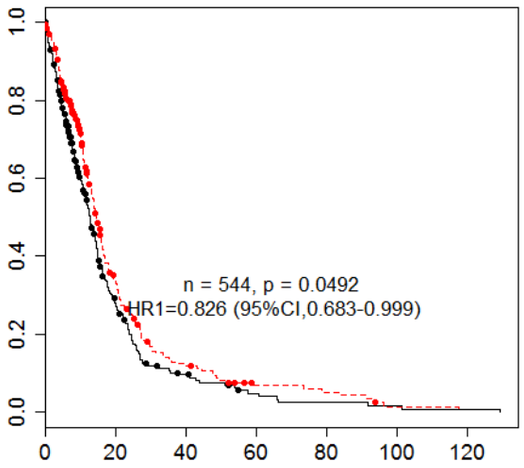
PTRF



EEF2

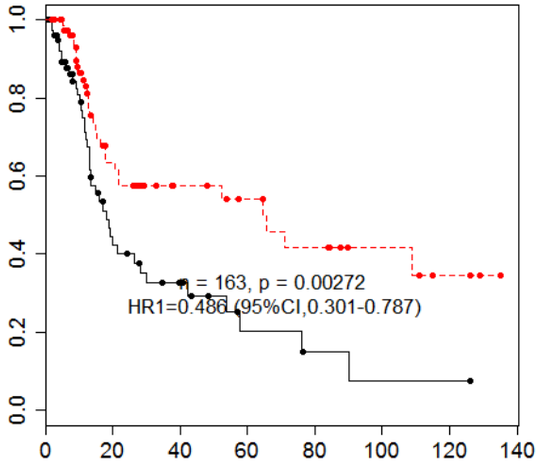


GNL3

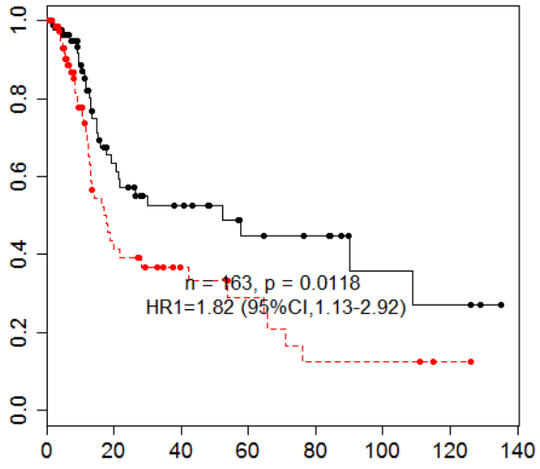


TOP1

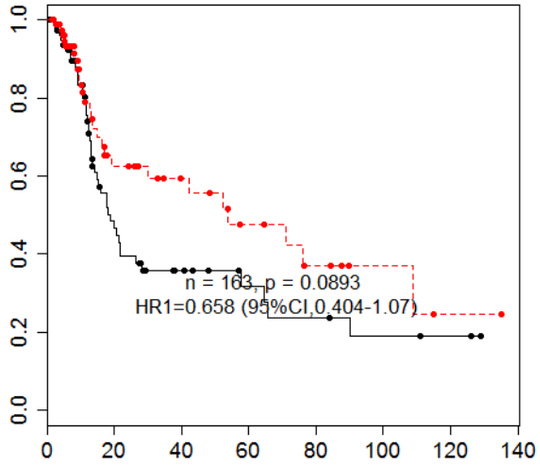
d. TCGA HNSC



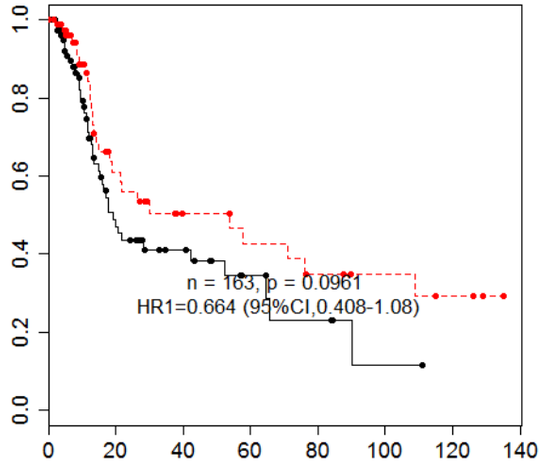
CABLES1



TOP1

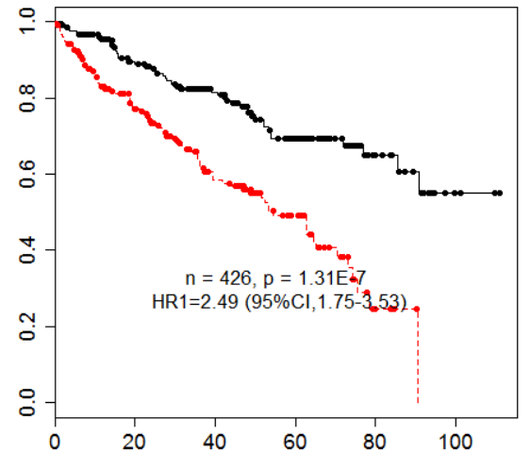


EEF2

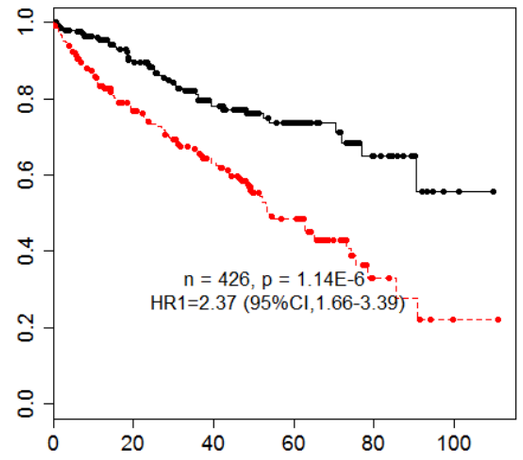


HIPK2

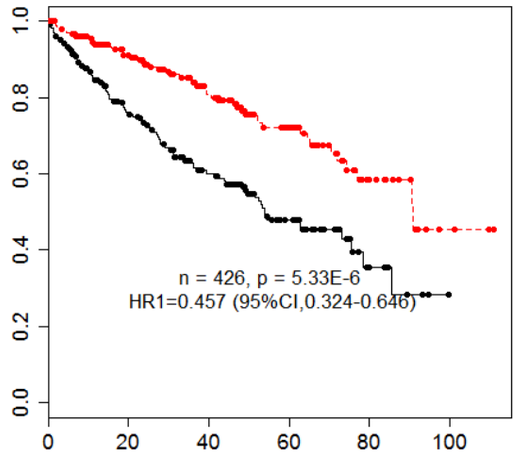
e. TCGA KIRC



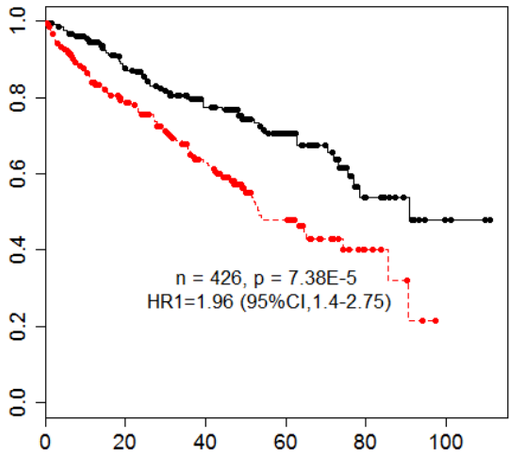
GNL3



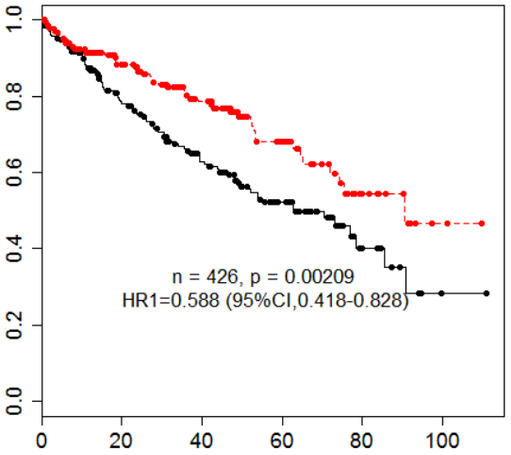
SERPINA3



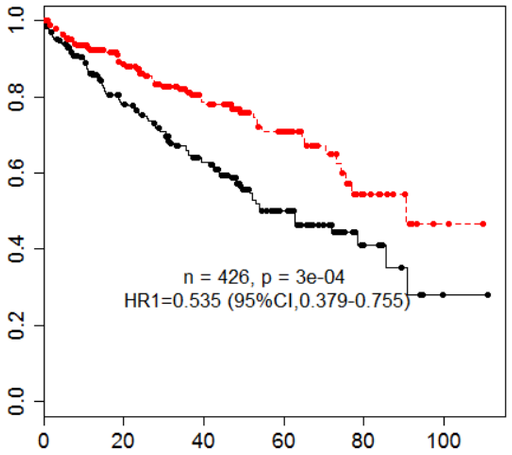
HIPK2



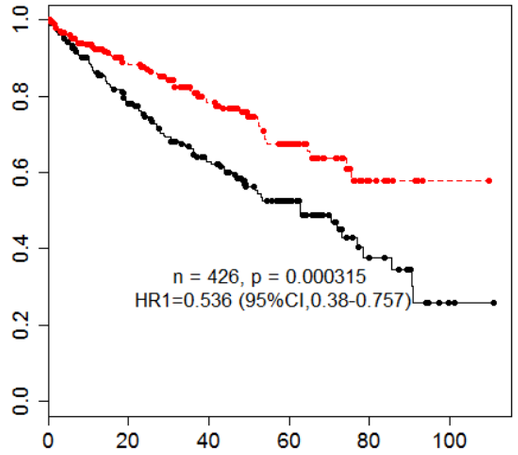
BEST1



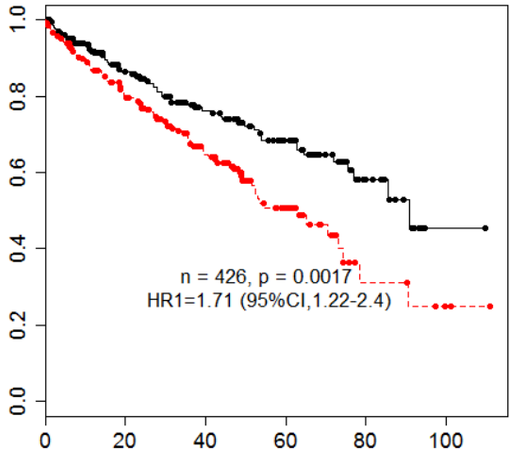
CREBBP



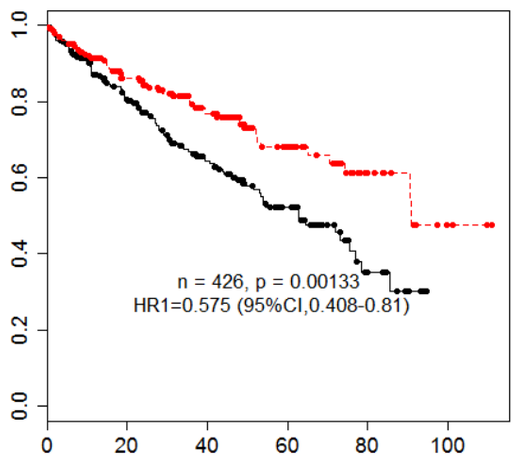
EP300



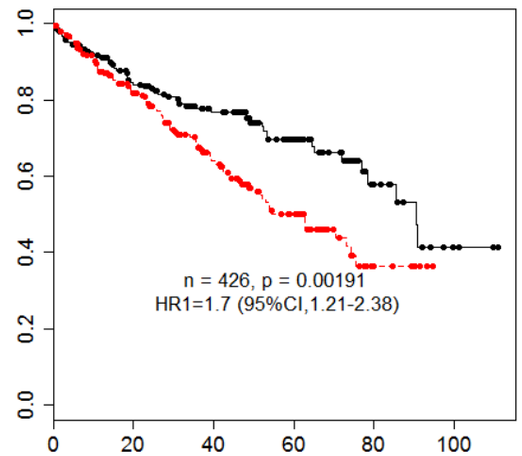
APP



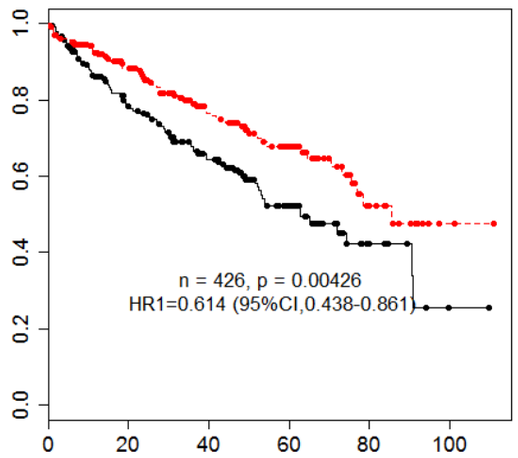
MDM4



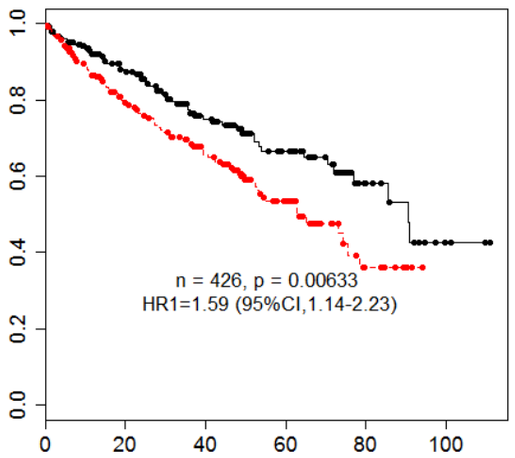
PPP1R13B



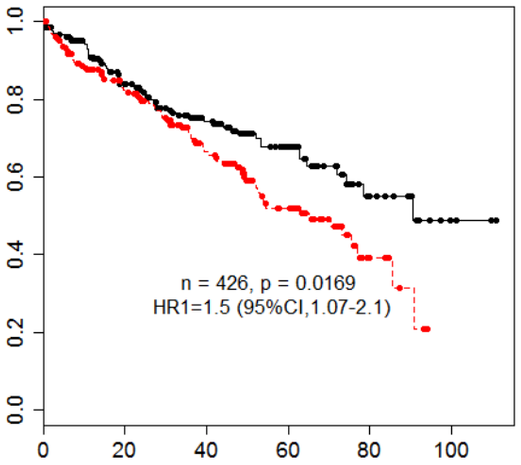
MMP7



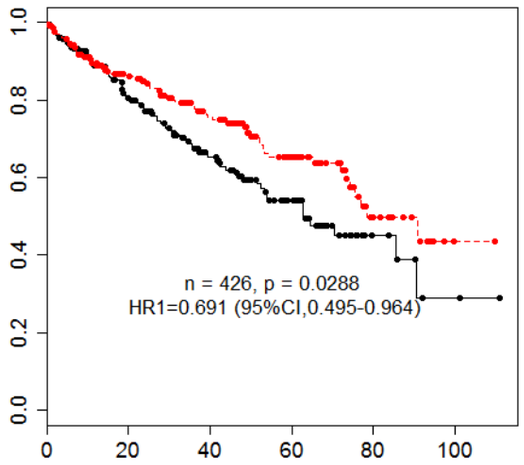
USP33



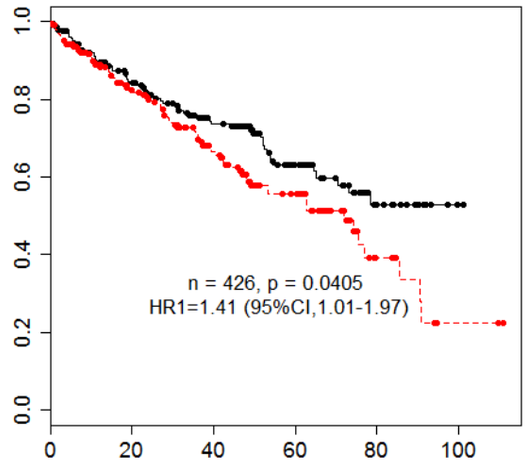
MMP9



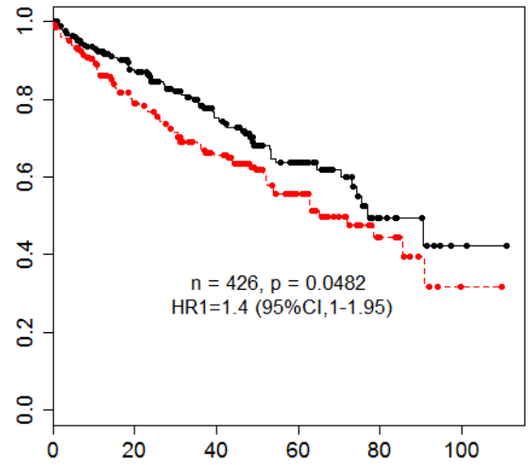
TP53



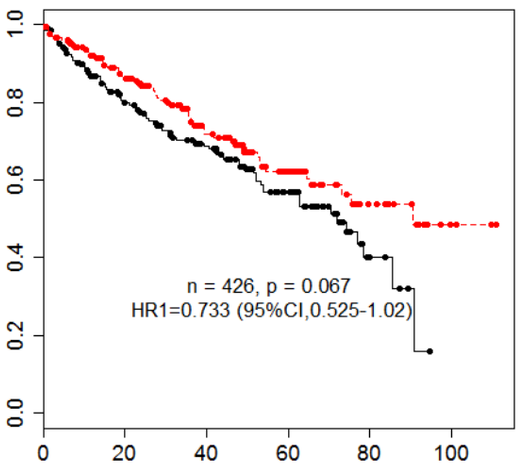
CTSG



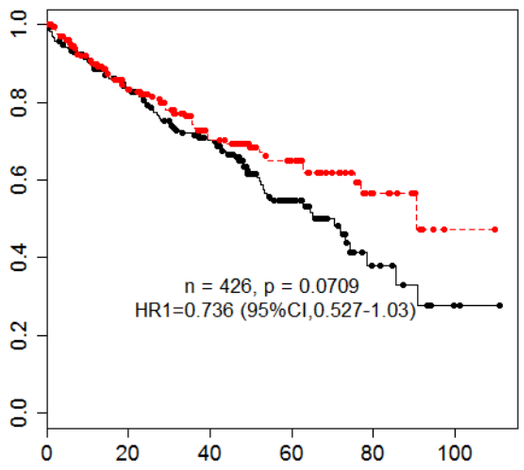
SERPINA1



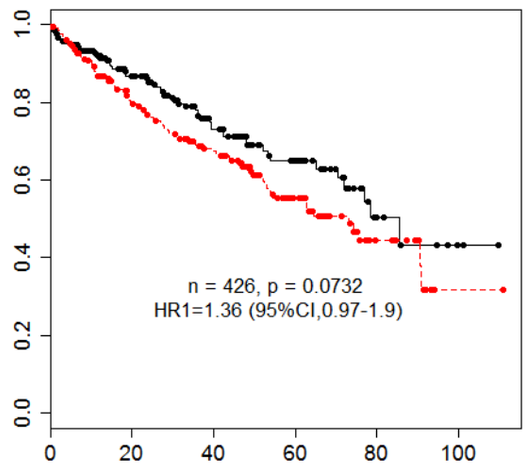
TCEB2



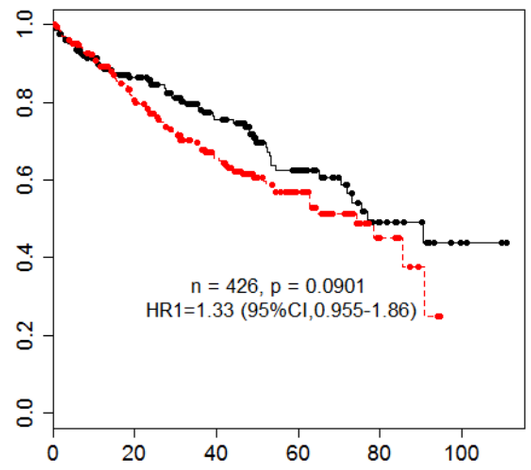
SP1



THBS1

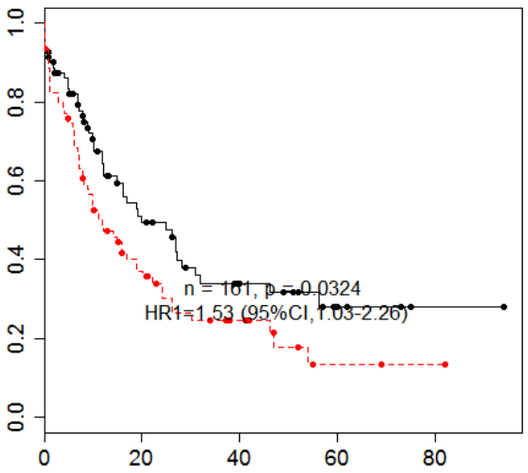


WT1

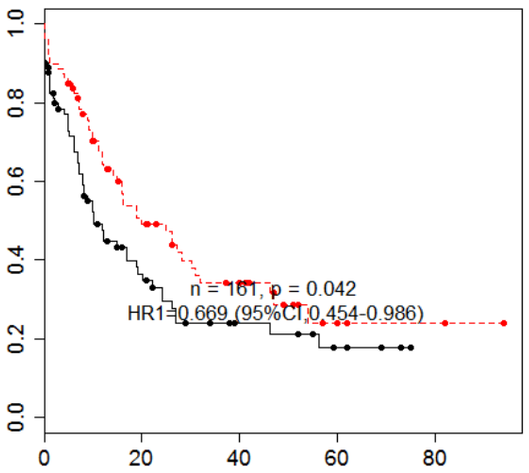


TCEB1

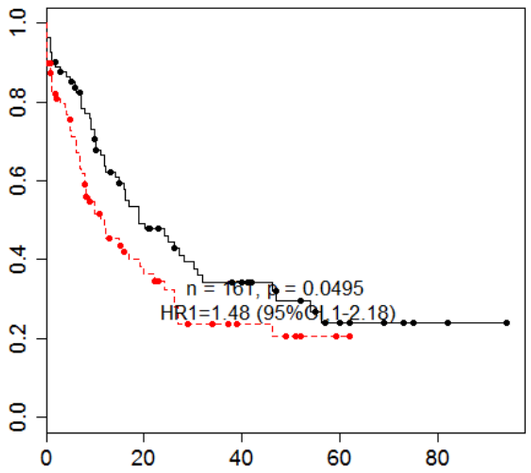
f. TCGA LAML



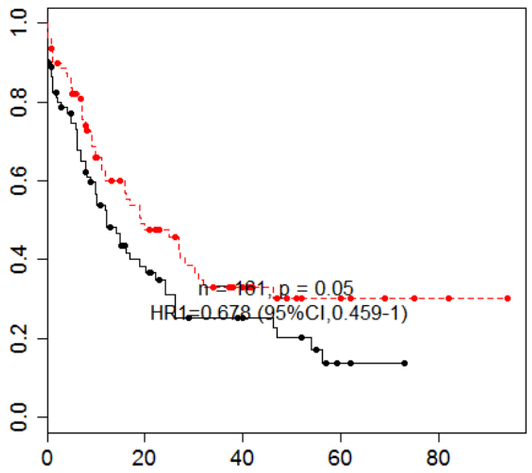
SHFM1



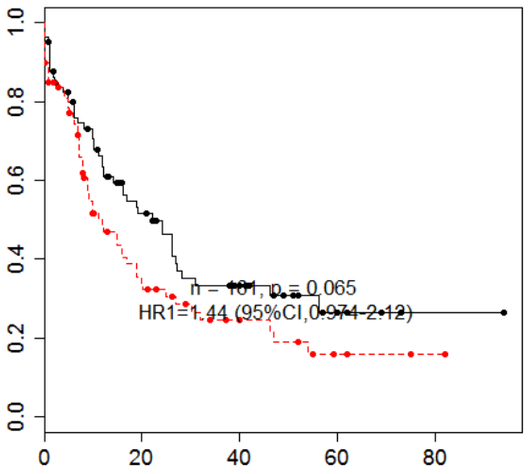
NPM1



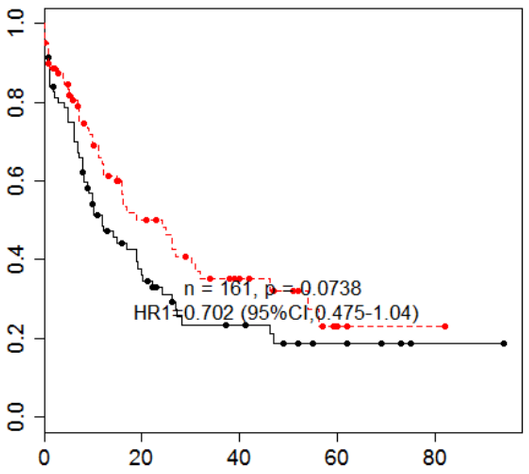
HSPA1A



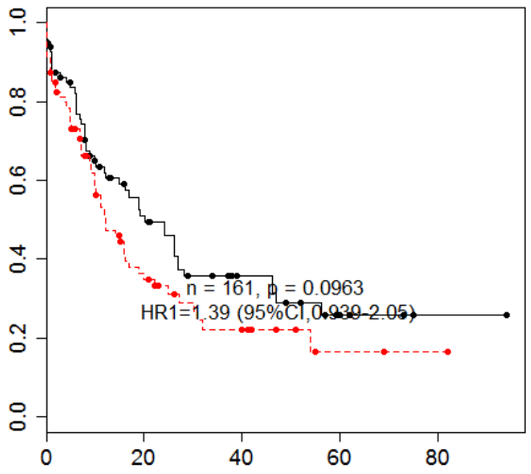
TP53



NQO1

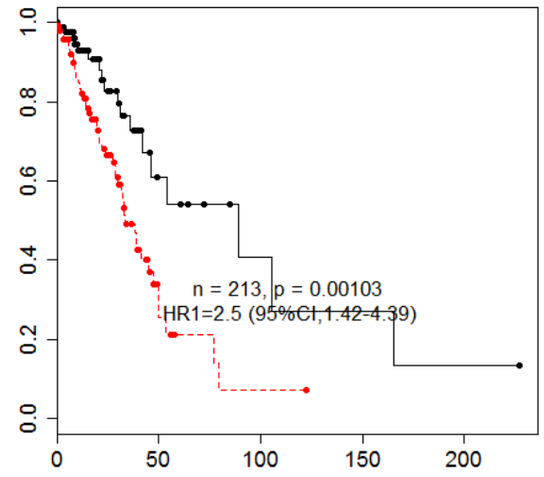


WDR16

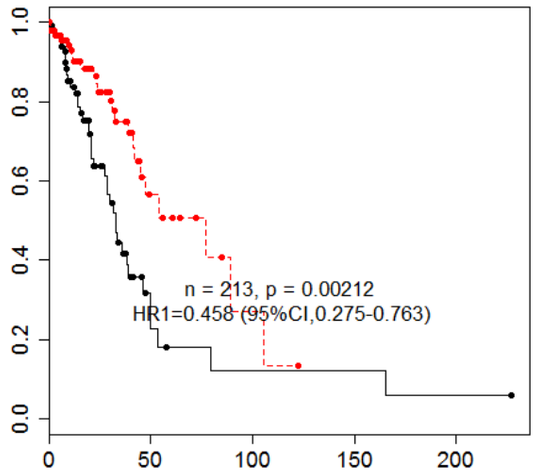


PARP1

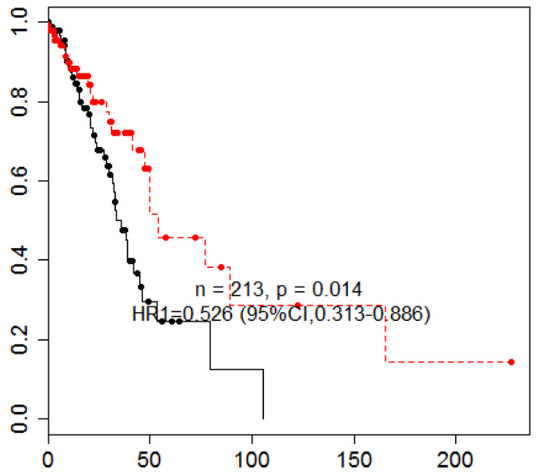
g. TCGA LUAD



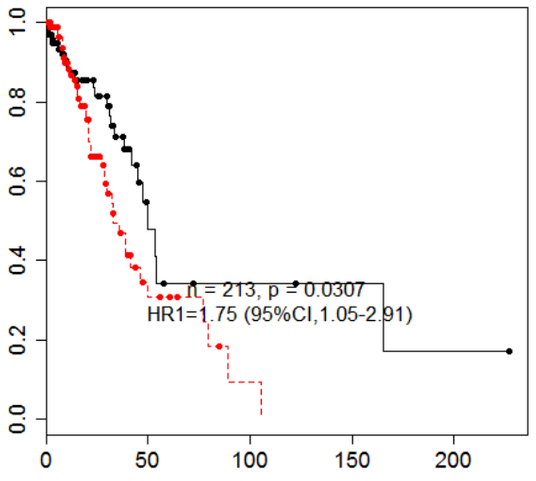
YBX1



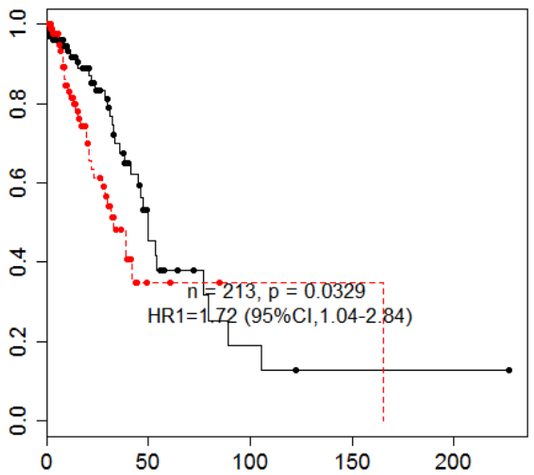
CABLES1



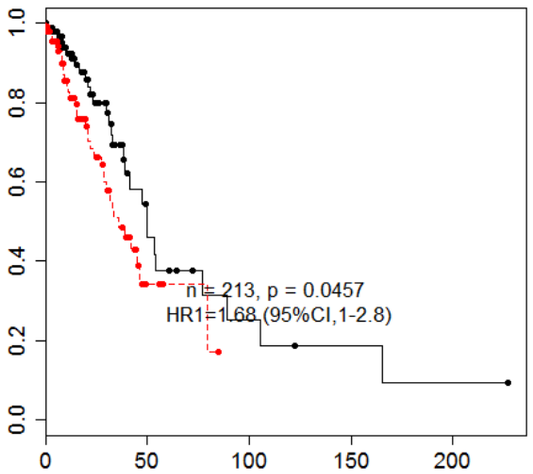
DDX5



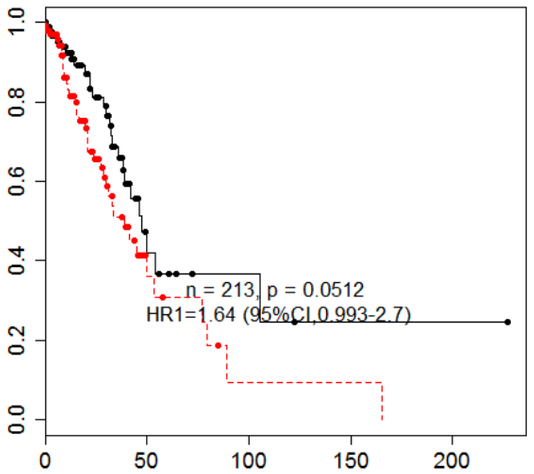
HSPA1A



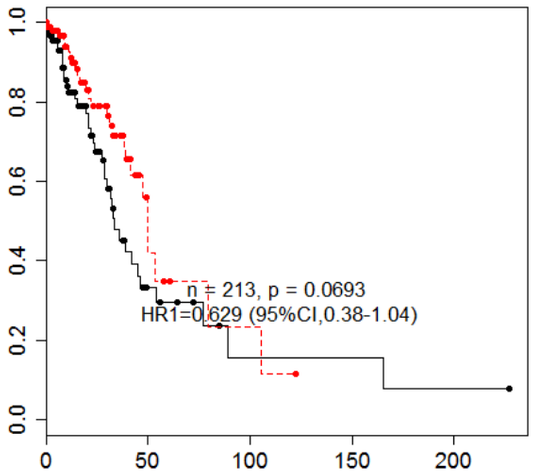
UPF3B



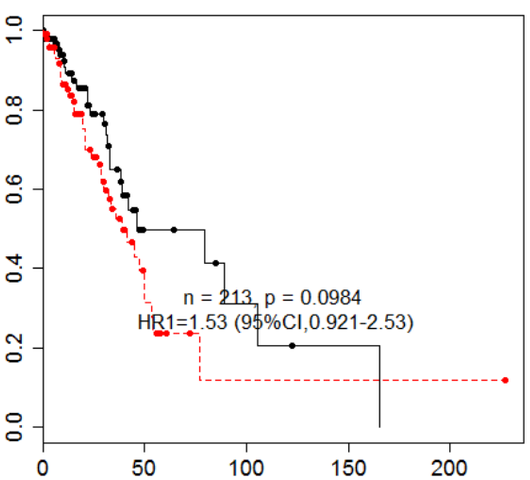
AURKA



YWHAZ

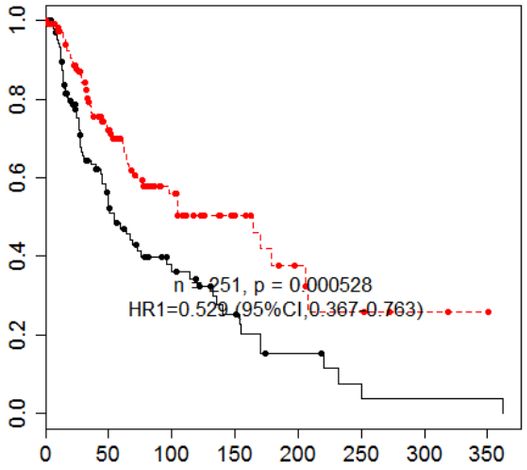


ACTC1

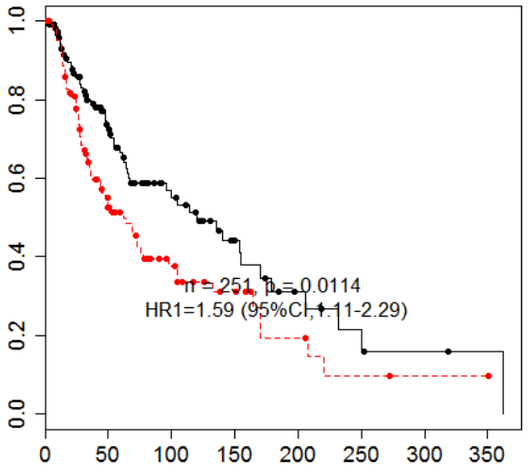


NPM1

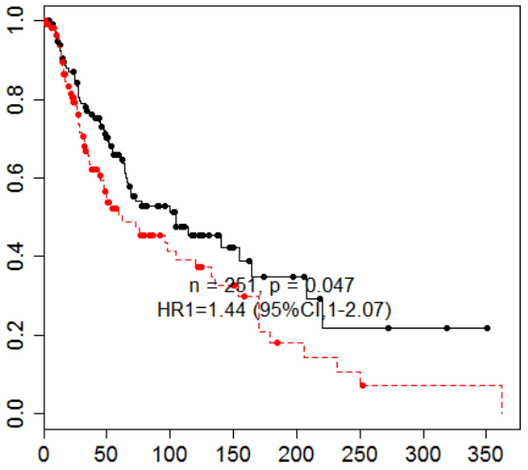
h.TCGA SKCM



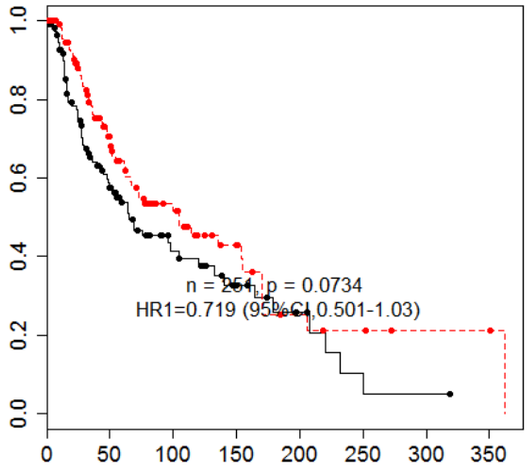
GCH1



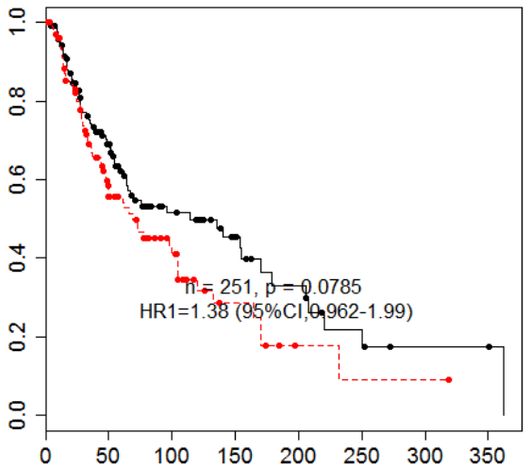
DSC1



AHSA1

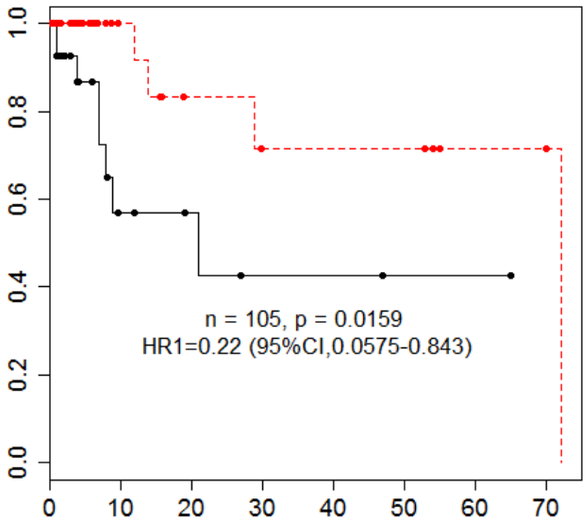


KRT18

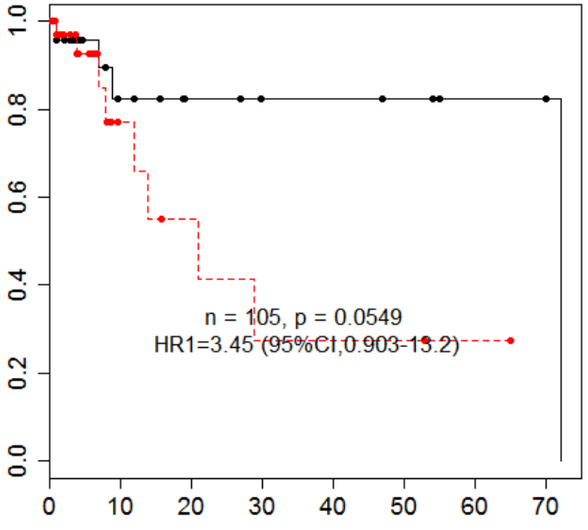


COL17A1

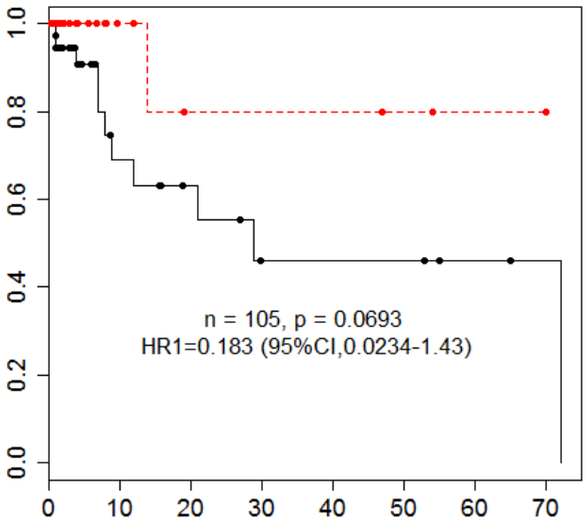
i. TCGA STAD



HSF1

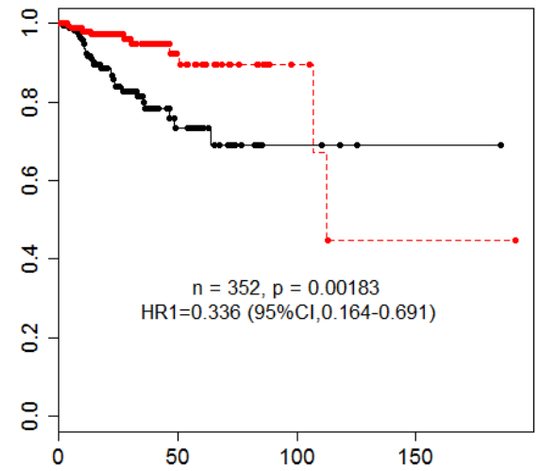


ASCC2

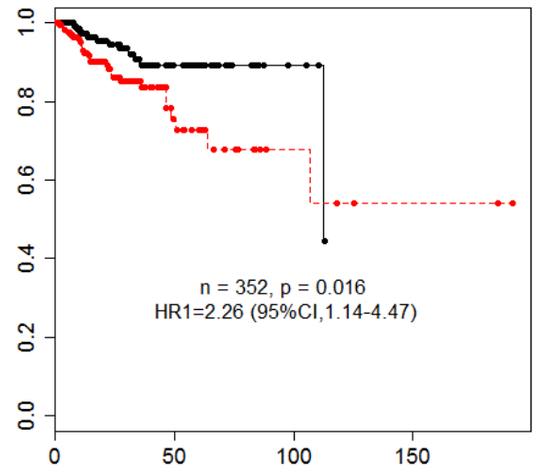


ACTA2

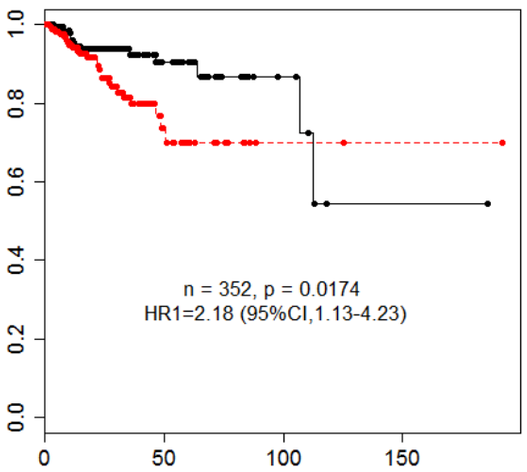
j. TCGA UCEC



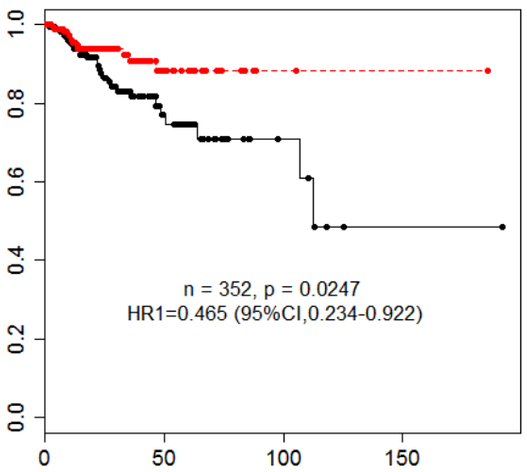
SPN



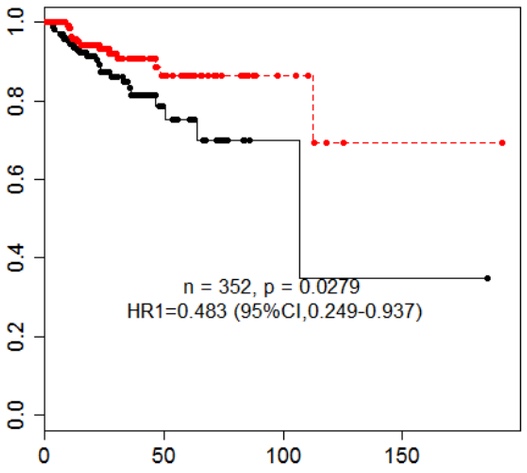
CDK6



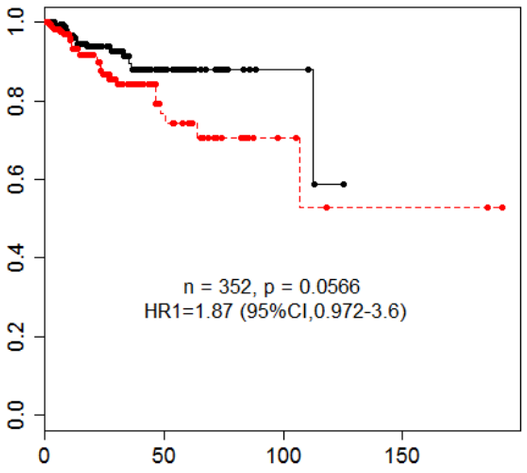
PIK3CA



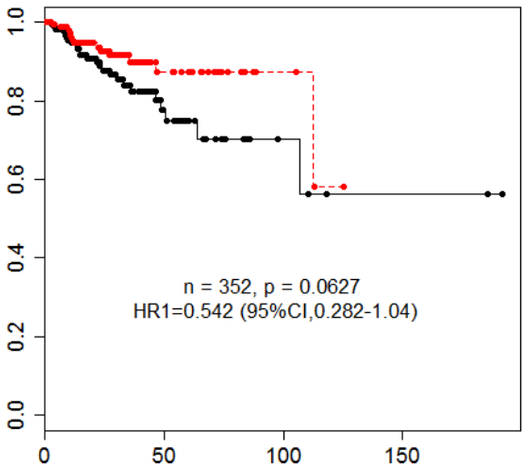
PIK3R1



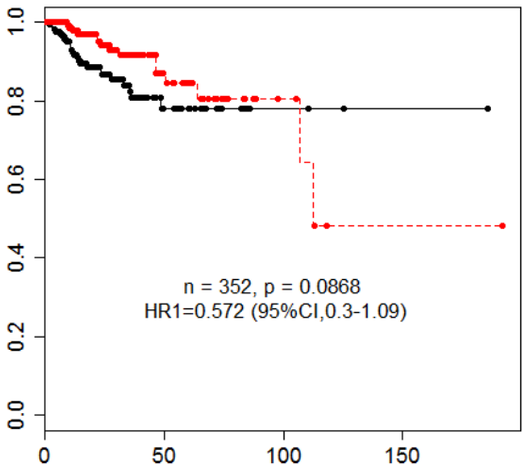
CDKN2D



FYN



RP1



PIK3CD