UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea Magistrale in Molecular Biology

TESI DI LAUREA

Integrated Multi-omics Survival Analysis of Gynecologic and Breast Cancers

Relatore: Prof. Chiara Romualdi
Dipartimento di Biologia

Controrelatore: Prof. Luca Pagani
Dipartimento di Biologia

Laureando: Amin Zolfaghari

ANNO ACCADEMICO 2022/2023

Table of Contents

ABSTRACT

This study utilized a multi-omics approach, incorporating gene expression, methylation, copy number variation, and mutation data, to analyze the survival of patients with Breast Carcinoma and Gynecologic Cancers (Ovarian Serous Cystadenocarcinoma, Uterine Corpus Endometrial Carcinoma, Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma, and Uterine Carcinosarcoma). The goal was to identify pathways and specific genes within those pathways that were significantly associated with patient survival. The MOSClip R package, a topological pathway analysis tool, was utilized to identify significant pathways, modules, and genes in survival analysis. This tool was chosen for its unique ability to perform survival analysis using multi-omics data while accounting for interactions among genes. Then, Cytoscape was used to visualize the topology of significant genes within each module. Through this analysis, 33 genes were identified as being common among different types of cancers. Afterwards, a comprehensive literature review was conducted to compare our findings with those of other studies. Then, heatmaps were created for each cancer type to illustrate the effect of significant genes on patient survival. Subsequently, Kaplan-Meier plots were compared among different types of cancers to provide valuable insights into the survival rates. Finally, an additional test was performed to assess the accuracy of survival prediction.

1. INTRODUCTION

1.1 Gynecologic and Breast Cancer

Cancer has been among the top causes of death in the world for many years, and unfortunately, the number of patients suffering from this disease is escalating annually (Łapińska et al., 2022). Types of cancers that can be formed and developed in women's reproductive organs are known as gynecologic cancer, which usually occurs between the age of 35 to 70. Among the known gynecologic cancers, ovarian, cervical, and endometrial neoplasms are the most aggressive types.

In this study, breast cancer was included in the analysis alongside gynecologic cancers due to the observed molecular similarities between these malignancies (Hoadley et al., 2018). Furthermore, referring to the WHO (World Health Organization) statistics, breast cancer was the cause of almost 700,000 deaths among women in 2020, with nearly 2.3 million new cases globally (Łapińska et al., 2022). So, statistical analysis has revealed a high mortality rate among patients diagnosed with breast and gynecologic cancers.

1.2 Survival Analysis

Optimal outcomes for cancer patients require the use of tailored prognostic, diagnostic, and therapeutic methods based on the specific type of malignancy (Łapińska et al., 2022). Despite significant research efforts to develop effective treatment techniques, cancer remains incurable. Thus, further efforts are necessary to improve patient survival rates.

Survival analysis is a commonly used tool in cancer biology research to compare the efficacy of different therapies and obtain statistics on tumor progression (Liu et al., 2018). One endpoint used in survival analysis is progression-free survival (PFS), which assigns each patient a status of either zero or one. A status of one indicates an event, such as tumor progression or death, while censored patients are assigned a status of zero. In this study, a multi-omics analysis approach was employed to identify variables that impact patient survival.

A multi-omics approach was chosen for the survival analysis to provide a more comprehensive understanding of patient survival compared to a single omic approach (Martini et al., 2019). By incorporating multiple molecular features into survival predictions, their accuracy can be improved. Furthermore, advances in technology and sequencing methods have made various omics data readily available on public databases such as The Cancer Genome Atlas (TCGA). Moreover, cancer is not solely the result of defects in gene expression but can also arise from abnormal changes in methylation sites, copy number variation, or mutation. Therefore, multi-dimensional analysis is necessary to enhance our understanding of cancer biology and compare the efficacy of different treatments.

The Cancer Genome Atlas (TCGA) database was established to provide researchers with molecular information on various types of cancer (Liu et al., 2018). Between 2006 and 2015, TCGA characterized the molecular features of 11,160 patients spanning a total of 33 different types of cancer. The availability of high-quality molecular data on the TCGA database has facilitated a significant amount of research, leading to advances in our understanding of cancer biology. The necessary omics data were obtained from TCGA using the curatedTCGAData R package.

The curatedTCGAData package, implemented in R, provides users with access to multi-omics molecular data from the TCGA database (Ramos et al., 2020). This package streamlines access to TCGA data by presenting complex information in the MultiAssayExperiment data structure. This data structure facilitates integrative analysis by researchers, who can easily navigate complex classes and data formats.

The MultiAssayExperiment data structure, implemented in R, provides bioinformaticians and data scientists with a practical means of representing complex data such as multi-omics reports (Ramos et al., 2017). This data structure simplifies the representation, visualization, and statistical analysis of genomics and transcriptomics data.

1.3 Ovarian Serous Cystadenocarcinoma

Among females, ovarian cancer is a prevalent malignancy within the reproductive system, comprising 2.5% of all such cases (Tong et al., 2023). The prognosis of ovarian cancer is not very promising, as the chances of surviving for 5 years are only 47.6%. This is largely due to the fact that many cases are diagnosed at a late stage.

Statistically, ovarian cancer ranks as the eleventh most prevalent neoplasm in the female population, and it is ranked as the fifth highest cause of mortality among women (Stewart et al., 2019). It is also the deadliest gynecologic cancer.

Annually, within the United States of America, over 22,000 new cases of ovarian cancer are diagnosed, and 14,000 deaths are attributed to this disease (Stewart et al., 2019). Subsequent to the whites, the highest occurrence of ovarian cancer per ethnicity is observed among Hispanic, Asian/Pacific Islander, African American, and American Indian/Alaska Native populations, with respective rates of 9.8, 9.0, 8.5, and 7.9 per 100,000 individuals.

Upon diagnosing the cancer condition, approximately 70% of ovarian cancer patients have advanced cancer (Shi et al., 2021). This is because ovarian cancer often does not present with clear and noticeable signs in its early stages. The unfavorable prognosis associated with ovarian cancer can largely be attributed to the lack of efficient techniques for the prompt detection, diagnosis, and the absence of reliable predictive markers.

Serous ovarian cystadenocarcinomas make up two-thirds of all malignant epithelial ovarian tumors (Mallick et al., 2018). When this cancer spreads, it commonly metastasizes to areas such as the bladder, lungs, liver, lymph nodes, peritoneum, and intestinal surfaces. On the other hand, metastasis from ovarian cancer to the cervix, vagina or vulva is very infrequent. However, in some cases of ovarian malignancy and cancers of the gall bladder, lung, breast, stomach, and pancreas, tumor emboli can spread through the bloodstream and lead to cervical metastasis. When cervical metastasis occurs in the previously referenced cancerous conditions, it is an indication that the neoplasm has progressed to a late stage and has spread to multiple organs, which suggests an unfavorable prognosis. Surprisingly, evidence indicates that many ovarian cancers actually originate in the fallopian tube rather than the ovaries (Stewart et al., 2019).

1.4 Uterine Corpus Endometrial Carcinoma

In high- and middle-income countries, the most frequently occurring gynecological malignancy is endometrial cancer (Koskas et al., 2021). Despite the fact that the overall prognosis is fairly positive, there is a tendency for high-grade endometrial cancers to recur. Preventing recurrence is crucial because the prognosis for endometrial cancer that has returned is very poor. The use of molecular factors to determine prognosis and treatment has been on the rise since The Cancer Genome Atlas defined four molecular subgroups of endometrial cancers. The conventional therapeutic approach involves a hysterectomy and the removal of both fallopian tubes and ovaries (bilateral salpingo-oophorectomy). The identification of patients with positive lymph nodes who require additional treatment, such as radiotherapy and chemotherapy, is made possible through lymphadenectomy and more frequently, sentinel node biopsy. For Stage I-II patients with high-risk factors and Stage III patients, adjuvant therapy is utilized.

Non-endometrioid cancers and those in the copy-number high molecular group characterized by a TP53 mutation are particularly treated with chemotherapy (Koskas et al., 2021). The optimal outcome for advanced disease is achieved through a combination of surgery to remove all visible disease and chemotherapy, incorporating or excluding the utilization of radiotherapy. Only patients exhibiting favorable performance status and an extended disease-free period are recommended for surgery for recurrent illness.

Globally, endometrial cancer is ranked as the sixth most prevalent neoplastic disease (Koskas et al., 2021). In 2018, approximately a total of 382,000 novel instances of this neoplastic condition were identified.

The prevalence of endometrial cancer is greater among high-income nations, 11.1 per 100,000 females, in comparison to countries with limited resources countries, 3.3 per 100,000 females (Koskas et al., 2021). The elevated occurrence of endometrial cancer among high-income nations may be attributed to the widespread presence of obesity and sedentary lifestyles, in addition to an aging demographic. Elevated concentrations of estrogen are postulated to be the principal factor contributing to the augmented risk of endometrial cancer among postmenopausal females who are obese. Participation in regular physical exercise and the prolonged administration of continuous combined estrogen-progestin therapy may reduce the likelihood of developing endometrial cancer. Furthermore, obesity is linked to a younger age of diagnosis of the endometrioid-type endometrial cancers. Endometrial cancer is most common in North America and Europe, where it represents the most commonly diagnosed neoplastic disease of the female reproductive system, and it is ranked as the fifth most prevalent neoplastic disease among females after breast, lung, colorectal, and non-basal skin cancer.

1.5 Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma

In 2018, uterine cervical cancer was the fourth most prevalent cancer among women, with over 560,000 new diagnoses and more than 310,000 fatalities globally (Dejima et al., 2020). Cervical cancer is unique in that it disproportionately affects younger women, despite cancer-related fatalities is more frequently observed among the geriatric population. Among women aged 20 to 39, cervical cancer ranks as the second highest cause of death due to cancer. Squamous cell carcinoma accounts for approximately 70% of cervical cancer cases. In excess of 95% of such cases can be attributed to an infection with the human papillomavirus (HPV).

It has been found that nearly all cases of cervical cancer can be attributed to infection with high-risk strains of the human papillomavirus (Cohen et al., 2019). Successful measures for prevention encompass programs for HPV screening and vaccination.

The two most common subtypes of cervical cancer are squamous cell carcinoma, accounting for approximately 70% of cases as mentioned before, and adenocarcinoma, accounting for around 25% of cases (Cohen et al., 2019).

Due to considerable geographical and worldwide differences in cervical cancer results, global gynecological oncology organizations have published guidelines for managing the disease based on evidence (Cohen et al., 2019). The objective of these guidelines is to improve patient care, even with advancements in prevention, screening, diagnosis, and treatment over the past ten years.

A cervical biopsy specimen is subjected to histopathological analysis in order to confirm the diagnosis of cervical cancer (Cohen et al., 2019). A pelvic examination, cervical cytology, and visual inspection of the cervix and vaginal mucosa are required for women displaying symptoms indicative of cervical cancer. An examination utilizing a speculum should be conducted to visually assess the cervix and vaginal mucosa. In instances of microinvasive illness or when the malignancy is situated in the endocervical canal, the cervix may seem unremarkable. Through the lymphatic vessels, lymph nodes located in the pelvic, para-aortic, mediastinal, supraclavicular, and inguinal areas can be metastasized to by cervical cancer. In the later stages of cervical cancer, it may be possible to detect swollen and hardened lymph nodes in the inguinal and supraclavicular areas through physical examination. In cases where malignancy is thought to be present based on clinical examination or cervical cytology results but has not been verified through the examination of cervical biopsy tissue samples using histopathological techniques, a cone biopsy must be performed to obtain a more definitive diagnosis.

1.6 Uterine Carcinosarcoma

Uterine serous carcinoma (USC) is a type of uterine cancer that is both uncommon and highly malignant in nature, making it a particularly dangerous form of the disease (Bloom et al., 2023). After undergoing surgery to treat uterine cancer, it is common for patients to receive additional treatment in the form of chemotherapy and/or radiation therapy. These treatments are known as adjuvant therapy and are intended to help prevent the cancer from returning. However, it is important to note that despite their use, these treatments may have limited effectiveness in reducing the likelihood of a high recurrence rate.

Despite the fact that uterine serous carcinomas accounts for only 10% of all uterine carcinoma cases, it is responsible for nearly 40% of fatalities resulting from the disease (Bloom et al., 2023). This highlights the aggressive nature of this type of cancer and the importance of early detection and treatment.

An in-depth and comprehensive understanding of the molecular intricacies of these malignant growths is of utmost importance for the effective implementation of targeted therapeutic interventions (Bloom et al., 2023). Such knowledge can significantly improve the likelihood of achieving favorable outcomes for patients suffering from these conditions. In contrast to the 16% of patients who are diagnosed with the more commonly occurring endometrioid histological subtype of uterine cancer, approximately 38% of individuals affected by the highly malignant uterine serous carcinoma are frequently identified at a progressed phase of the illness. This disparity in diagnosis can be attributed to the particularly virulent nature of uterine serous carcinoma.

Individuals identified with a progressed phase of the illness have a significantly elevated rate of recurrence, reaching up to 90% (Bloom et al., 2023). For individuals afflicted with uterine serous carcinoma, who constitute a demographic with a significant need for improved clinical outcomes, the presently accepted course of treatment involves an initial surgical procedure to determine the extent of the disease, followed by supplementary interventions such as chemotherapy and/or radiation therapy.
The PD-1 inhibitors pembrolizumab and dostarlimab have recently received approval for administration in patients with recurrent uterine cancer and microsatellite unstable tumors (Bloom et al., 2023). The combination of pembrolizumab with lenvatinib has been approved for the treatment of recurrent microsatellite stable uterine cancer. Additionally, the use of trastuzumab has resulted in improved survival outcomes for patients with advanced and recurrent uterine serous cancer exhibiting HER2 overexpression.

1.7 Breast Carcinoma

Among women worldwide, the most detected type of malignant tumor is breast cancer (Smolarz et al., 2022). It poses a significant health risk and represents the foremost cause of mortality resulting from malignant tumors. Despite progress in early detection and treatment, breast cancer continues to be a major source of death among women. Globally, the occurrence of breast cancer is on the rise, with an upward trend observed in all regions. This increase in prevalence is alarming and underscores the importance of ongoing research and efforts in prevention and treatment.
Although there have been advancements in identifying and diagnosing the disease, leading to a decrease in the number of deaths caused by the disease, it is still essential to explore innovative therapeutic approaches and indicators that can predict and provide insight into the likely course of the disease (Smolarz et al., 2022).
The selection of therapeutic interventions is dependent on the specific molecular characteristics of the disease, which means that doctors and researchers must carefully analyze the underlying causes and mechanisms of the disease in order to determine the most effective course of treatment (Smolarz et al., 2022). This approach allows for personalized medicine that is tailored to the individual needs of each patient. The management of breast cancer involves a collaborative approach that draws on the expertise of multiple medical disciplines and includes both locoregional interventions such as surgery and radiation therapy, as well as systemic therapy. Systemic therapies for the management of breast cancer encompass a range of interventions including hormone therapy for hormone-positive cases, chemotherapy, anti-HER2 therapy for HER2-positive cases, and lately, the use of immunotherapy.

According to recent studies and statistics, it has been observed that triple negative breast cancer, a particularly aggressive form of the disease, accounts for approximately 15-20% of all diagnosed cases of breast cancer worldwide (Smolarz et al., 2022). Triple negative breast cancer presents significant difficulties in treatment, primarily as a result of its low responsiveness to conventional treatments and its tendency to aggressively invade surrounding tissues. These particular challenges have garnered a considerable amount of attention from members of the scientific community. In the field of breast cancer treatment, there is a growing emphasis on the development of personalized therapeutic approaches that aim to tailor treatment to the individual patient. This involves adjusting the intensity of treatment based on factors such as the specific biology of the cancer and the patient's initial reaction to therapy.

1.8 Aims of the work

This study aimed to conduct a multi-omics survival analysis, incorporating gene expression, methylation, copy number variation (CNV), and mutation data, on Breast Carcinoma (BRCA) and Gynecologic Cancers (Ovarian Serous Cystadenocarcinoma (OV), Uterine Corpus Endometrial Carcinoma (UCEC), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), and Uterine Carcinosarcoma (UCS)). The initial objective was to identify pathways significantly associated with patients' survival. Subsequent analyses sought to determine which modules within each significant pathway and which genes within each significant module were significantly associated with survival. The MOSClip R package was used to perform these analyses.

MOSClip is an R package that facilitates integrated multi-omic survival analysis through the use of multivariate models and dimensionality reduction of multi-omics data (Martini et al., 2019). This topological pathway analysis tool can identify significant pathways, modules, and genes in survival analysis. MOSClip was chosen for this study because it is currently the only available tool capable of performing survival analysis using multi-omics data while accounting for interactions among genes. To be more specific, MOSClip is a versatile approach that employs omic-specific dimensionality reduction techniques to conduct multi-omic survival tests on individual pathways or modules derived from a pathway's graph structure. These tests can be performed independently at the pathway or module level using a multivariate survival model to identify associations with patient survival.

These analyses aimed to improve prognosis and treatment for specific cancer types by examining survival rates and relevant data, providing valuable insights into the nature of these diseases and how they can be effectively managed.

2. MATERIALS AND METHODS

2.1 Retrieving Cancer Data from the TCGA Database using the curatedTCGAData Package

Gene expression, methylation, CNV, and mutation data (version 2.0.1) for five types of cancers (OV, UCEC, UCS, CESC, and BRCA) were retrieved from the TCGA database using version 1.18.0 of the curatedTCGAData package (Ramos et al., 2020) on April 18th, 2023. A MultiAssayExperiment object was obtained for each cancer type, containing different assays corresponding to various omics data types.

2.2 Focusing on Primary Tumors and Female Patients in TCGA Data Analysis

After retrieving the data, we made the decision to focus our attention solely on primary tumors. This decision was made due to the fact that the TCGA database primarily contains molecular data for primary tumors, which have not yet metastasized. As a result of this, the majority of the data that was downloaded for each cancer type fell into this particular tumor category.
To isolate primary tumors, we utilized the TCGAutils R package (Ramos et al., 2020). Following this step, our analysis focused exclusively on multi-omics data for primary tumors, with the aim of ensuring homogeneity in our results. As a consequence of this focus, when interpreting our findings, we can be confident that any significant genes that are identified in a specific cancer type, such as BRCA, are specific to primary BRCA tumor cells and not metastasized or normal cells.
Another modification that we made to the data was the removal of molecular data for male patients from the BRCA dataset. This was due to the fact that it was the only cancer type that was shared between males and females. As a result of this modification, 13 male patients were excluded from further analysis. From this point forward, our analysis focused solely on multi-omics data for five types of primary tumors in female patients.

2.3 Modifying CNV Data for Analysis with MOSClip

The CNV assay provided as a MultiAssayExperiment object for each cancer type was deemed unsuitable for further analysis. As indicated in Table 1, the CNV matrix lacked gene names corresponding to each row of the data frame. Additionally, the numeric values in each cell represented GISTIC2 copy number values with a noise cutoff of 0.3 (GDC Docs, n.d.). In light of this, we made the decision to convert these values to -1, 0, and 1 in order to represent loss of copy number, no change in copy number, and gain of copy number, respectively.

TCGA.3C.AAAU.01A.11D.A41E.01	TCGA.3C.AALI.01A.11D.A41E.01	TCGA.3C.AALJ.01A.31D.A41E.01
0.069	−1.008	−0.330
0.249	−0.440	−0.324
0.867	−0.440	−0.324
0.216	−0.440	−0.324
3.657	0.387	0.083
0.071	0.387	0.083
0.707	0.387	0.421
0.075	0.387	0.421
1.205	0.387	0.421
1.233	0.387	0.421
0.023	0.387	0.421
1.262	0.387	0.421
0.167	0.387	0.421
−0.487	0.387	0.421
0.192	0.387	0.421
−0.454	0.387	0.421
0.121	0.387	0.421
0.804	0.387	0.421
−0.488	0.387	0.421
−1.195	0.387	0.421
−0.451	0.387	0.421
−0.439	0.387	0.421
0.203	0.387	0.421
0.187	0.412	0.412
−0.447	0.434	−0.000
−0.421	0.416	0.374
0.353	0.000	−0.381

Table 1. ***Subset of Initial CNV Data Frame for BRCA*****.** As demonstrated, each row of the initial CNV data frame represents a single gene, but the names of the genes were not indicated in this primary matrix. Additionally, GISTIC2 copy number values are displayed for each patient. This table serves as an example of the CNV matrix structure prior to subsequent modifications. The same structure was present for the other four cancer types.

To address the issue of adding gene names corresponding to each row of the CNV data frame, we utilized the “rowData” of the MultiAssayExperiment object. As gene names were provided as rowData, we simply added this information to the existing matrix to create a properly formatted data frame with gene names as row names and patient codes as column names (Table 2).

The second modification involved converting GISTIC2 copy number values to -1, 0, and 1. According to the TCGA database documentation (GDC Docs, n.d.), all GISTIC2 copy number values between -0.3 and 0.3, indicating no change in copy number, were converted to 0. For values below -0.3 and above 0.3, -1 and 1 were assigned to represent loss and gain of copy number, respectively (Table 2).

Additional modifications included converting gene names from Symbols and Ensembl to Entrez id to ensure compatibility with the MOSClip package (Martini et al., 2019), and formatting patient names to match the patient ids in the survival data frame (Table 10).
To implement the aforementioned modifications to the initial CNV data frame (Table 1), we developed a function in R that produced an output matrix suitable for use with MOSClip (Table 2).

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:100874253	0	0	1
ENTREZID:10158	0	0	1
ENTREZID:6886	0	0	1
ENTREZID:6491	0	0	1
ENTREZID:51727	0	0	1
ENTREZID:2301	0	0	1
ENTREZID:2306	0	0	1
ENTREZID:388630	0	0	1
ENTREZID:200010	0	0	1
ENTREZID:54558	0	0	1
ENTREZID:84871	0	0	1
ENTREZID:79656	0	0	1
ENTREZID:1996	0	-1	1
ENTREZID:63950	0	0	1
ENTREZID:11124	0	0	1
ENTREZID:1031	0	0	1

Table 2. ***Subset of Final CNV Data Frame for BRCA****.* As shown, the final matrix features gene Entrez ids as row names and patient codes as column names. Each cell of the data frame can have one of three values (-1, 0, or 1), representing loss, no change, or gain of copy number, respectively. This matrix is a subset of the BRCA copy number variation data frame. Similar data frames were created for the other four cancer types.

In conclusion, to complete the modification of the CNV data frame, we utilized the same function implemented in R to create a properly formatted data frame for each cancer type. This allowed us to use MOSClip to analyze the data and draw meaningful conclusions from our findings. By ensuring that the data frames for each cancer type were appropriately structured and formatted, we were able to conduct a thorough and accurate analysis of the multi-omics data.

2.4 Modifying Gene Expression Data for Analysis with MOSClip

The gene expression assay was modified in a manner similar to the CNV assay. As shown in Table 3, the raw gene expression data for BRCA consisted of a data frame with gene symbols as row names and patient IDs as column names. To avoid repetition, only the data frame for BRCA gene expression is shown here; however, the structure and modifications were identical for all five cancer types. The modifications aimed to create an interpretable gene expression matrix for further analysis.

First, the patient names were standardized to match the patient IDs in the survival matrix (Table 10). Second, the gene names were converted from Symbols to Entrez IDs to ensure compatibility with the MOSClip package (Martini et al., 2019). As shown in the final gene expression data frame (Table 4), the table now consists of Entrez IDs as row names and appropriate patient IDs as column names.

In order to mitigate the issue of data sparsity, the TPM gene expression values were subjected to a filtering process as described by Martini et al. (2019). Only those genes that exhibited a TPM value of at least 100 among all patients were retained for further analysis. Conversely, any genes that did not meet this threshold, i.e., had TPM values less than 100 among all patients, were subsequently discarded as outlined in Table 4. This approach ensures that the data used in subsequent analyses is of high quality and reduces the potential impact of data sparsity on the results.

In order to ensure accurate analysis, several steps were taken to normalize the gene expression data. The gene expression values were normalized using “between lane normalization” with the “upper quartile” method. This approach normalized the gene expression values between samples, while TPM values were already normalized within patients (thanks to TPM). By performing between lane normalization, a normalized data set was created that accounted for variation both within and between samples. The normalized gene expression values were then converted to a logarithmic scale to facilitate further analysis and visualization (Table 4).

To complete the modifications to the gene expression matrix, an R function was implemented to perform all necessary changes. This function takes an initial MultiAssayExperiment object and the assay name (gene expression) as input and produces a final data frame ready for use with MOSClip (Martini et al., 2019) for further analysis (Table 4).

	TCGA.A1.A0SB.01A.11R.A144.07	TCGA.A1.A0SD.01A.11R.A115.07	TCGA.A1.A0SE.01A.11R.A084.07
A1BG	164	546	1341
A1CF	0	0	0
A2BP1	22	1	2
A2LD1	127	331	498
A2ML1	94	144	114
A2M	102123	107181	101192
A4GALT	890	1409	1711
A4GNT	6	5	2
AAA1	0	0	1
AAAS	2139	2219	4294
AACSL	2930	3	4
AACS	6533	3102	5271
AADACL2	0	0	1
AADACL3	0	0	0

Table 3. ***Subset of Initial gene expression Data Frame for BRCA.*** As shown, each row corresponds to a gene symbol and each column represents a patient code. Each cell of the data frame contains a TPM (transcripts per million) gene expression value. This table illustrates the structure of the gene expression data frame prior to any modifications. The same structure was present for the other four cancer types.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:1	7.965784	9.948367	10.324181
ENTREZID:144568	7.169925	6.409391	6.894818
ENTREZID:2	17.243974	16.184391	16.234368
ENTREZID:53947	10.403012	10.299208	10.773963
ENTREZID:100329167	0.000000	1.000000	0.000000
ENTREZID:8086	11.667112	11.626165	12.191368
ENTREZID:65985	13.277723	11.921841	10.901621
ENTREZID:344752	0.000000	1.000000	0.000000
ENTREZID:126767	0.000000	0.000000	0.000000
ENTREZID:13	6.022368	3.459432	2.807355
ENTREZID:51166	10.044394	7.209453	6.700440
ENTREZID:79719	12.112765	12.911205	12.332596
ENTREZID:22848	12.500344	12.323336	11.808964

Table 4. ***Subset of Final Gene Expression Data Frame for BRCA.*** The final gene expression data frame consists of an Entrez gene ID for each row and a patient code for each column. The gene expression values are TPM counts normalized using “between lane normalization” with the “upper quartile” method and then converted to a logarithmic scale.

2.5 Modifying Methylation Data for Analysis with MOSClip

Similar to previous omics, the methylation data (Table 5) required modification. To achieve this, a function was implemented in R to perform the necessary modifications. This function takes as input a MultiAssayExperiment object downloaded using CuratedTCGAData (Ramos et al., 2020) and produces an appropriate matrix for further analysis. For the methylation assay, “Probe-level methylation β-values from Infinium HumanMethylation 450K BeadChip” was used. However, this assay only had methylation data for 10 patients in OV. Therefore, for OV, “Probe-level methylation β-values from Illumina HumanMethylation 27K BeadChip” was used instead, providing methylation data for 575 patients.
The initial methylation data frame (Table 5) contained the names of the Illumina probes as row names and the patient IDs as column names. Each probe corresponded to a CpG island. However, there were two considerations that needed to be addressed. First, multiple CpG islands could be associated with a single gene. Since our project aimed to compare methylation levels among different genes rather than different CpG islands, we collapsed multiple CpG islands corresponding to a single gene, calculated the average methylation value for these sites, and assigned it to the related gene. Second, two different genes could share a CpG island due to their proximity on the chromosome. In such cases, we considered the methylation value for both genes (Table 6).

	TCGA.3C.AAAU.01A.11D.A41Q.05	TCGA.3C.AALI.01A.11D.A41Q.05	TCGA.3C.AALJ.01A.31D.A41Q.05
cg00000029	0.10362281	0.14125995	0.14674178
cg00000108	*NA*	*NA*	*NA*
cg00000109	*NA*	*NA*	*NA*
cg00000165	0.09736179	0.74387661	0.70057546
cg00000236	0.87820501	0.87990070	0.76554616
cg00000289	0.38562216	0.50921010	0.46636368
cg00000292	0.67848346	0.26104505	0.85062771
cg00000321	0.38603727	0.29728965	0.60842645
cg00000363	0.21931279	0.43003277	0.76810877
cg00000622	0.01629138	0.01640277	0.01750924
cg00000658	0.87140921	0.89519018	0.92390651
cg00000714	0.09512542	0.12939231	0.10329124

Table 5. ***Subset of Initial Methylation Data Frame for BRCA.*** As illustrated, each row corresponds to an Illumina probe that can bind to a CpG island, providing information about that CpG island. In contrast, the columns represent patient names, similar to previous data frames for other types of omics. Each cell in the table displays the methylation beta value of a specific CpG island in a particular patient. These values range from 0 to 1, indicating the level of methylation at that site.

The function that was implemented made a number of significant modifications to the initial matrix with the ultimate goal of producing an appropriate data frame. One decision that had to be made involved choosing between two options: removing probes that had NA values for some patients or imputing the NA values with the median methylation value for that specific patient. Since removing probes that had NA values could potentially result in the loss of a considerable amount of valuable information, we ultimately decided to use the median beta value of all probes in each patient to impute NA values. This approach allowed us to retain all information without introducing strong biases, as we used the median value to replace NAs.

In the subsequent step of the process, the rowData of the methylation assay was utilized in order to determine which specific gene each individual CpG island corresponded to. Following the completion of this step, the names of the genes were added to the initial data frame with the ultimate goal of being used in place of the probe names. At the same time, CpG islands that corresponded to more than one gene were identified through careful analysis. Once these CpG islands were identified, new rows representing the second gene associated with that particular CpG island were added to the matrix in order to ensure that all relevant information was accurately represented.

Furthermore, by determining the names of the genes corresponding to each individual CpG island, we were able to assign the mean beta value of different CpG islands that were associated with a single gene to that specific gene. This process involved calculating the mean beta value of all CpG islands associated with a single gene and assigning that value to the gene in question. As a result of this process, instead of having multiple methylation beta values for different CpG islands corresponding to a single gene, we were able to represent each gene with a single methylation value. This approach allowed us to more accurately represent the methylation levels of each gene and facilitated further analysis.

In the final step of the process, several important modifications were made to the data. First, the gene symbols were converted to gene Entrez IDs. This involved replacing the gene symbols with their corresponding Entrez IDs, which are unique identifiers assigned to each gene by the National Center for Biotechnology Information (NCBI). This conversion ensured that the data was accurately represented using a standardized nomenclature. Second, the patient IDs were modified to match the patient codes in the survival data frame (Table 10). This modification ensured consistency between the two data frames and facilitated further analysis by allowing for easy comparison and integration of data from both sources.

In conclusion, the R function that was implemented successfully modified the initial data frame to produce the final matrix, as shown in Table 6.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:5934	0.13222470	0.15496310	0.16117266
ENTREZID:64778	0.42246049	0.56557881	0.53156675
ENTREZID:7419	0.22706916	0.28177166	0.26004421
ENTREZID:87	0.39107283	0.52199149	0.49191216
ENTREZID:487	0.56322880	0.64470095	0.61950580
ENTREZID:6422	0.36971157	0.31078254	0.42297270
ENTREZID:81614	0.24357044	0.27883399	0.26604676
ENTREZID:11253	0.44800478	0.54414507	0.52283215
ENTREZID:79042	0.04469041	0.05782955	0.04703624
ENTREZID:7555	0.14153272	0.20467049	0.19203970
ENTREZID:57696	0.13396774	0.19352094	0.17965353
ENTREZID:114818	0.65428079	0.64375933	0.67019831
ENTREZID:7097	0.28344690	0.31626283	0.31015606
ENTREZID:145773	0.21052457	0.36933046	0.31707182
ENTREZID:3784	0.54740705	0.65125275	0.61592121
ENTREZID:261734	0.68501979	0.75987277	0.74390786

Table 6. ***Subset of Final Methylation Data Frame for BRCA.*** As demonstrated, the final methylation data frame for BRCA, analogous to other cancer types investigated in this study, comprises Entrez gene IDs as row names and corrected patient IDs as column names. Furthermore, each element of this matrix represents the degree of methylation for a specific gene in a particular patient.

2.6 Modifying Mutation Data for Analysis with MOSClip

The mutation data frame required significant modifications. Initially, we had a MultiAssayExperiment containing various types of assays. Two of these assays, namely “Hugo_Symbol” and “Variant_Classification,” were utilized in our study. The former, “Hugo_Symbol,” was a matrix with patient IDs as column names and chromosome positions as row names (Table 7). This assay contained gene symbols wherever a mutation occurred at a specific position on a patient’s chromosome and NA values to indicate the absence of a mutation. Conversely, the latter assay, “Variant_Classification,” comprised the type of mutation in each matrix cell where a mutation was present and an NA value for each cell without a mutation (Table 8). The row and column names of both assays were identical.

As previously stated, the “Variant_Classification” assay contained the types of mutations for each position in each patient. To elaborate, if a mutation occurred at a specific position on a patient’s chromosome, the type of mutation would be indicated within this matrix. Consequently, for each non-NA cell, the value could represent mutations in regulatory regions, nonsense, missense, silent, splice site, frameshift, or indel mutations (National Cancer Institute, n.d.).

	TCGA.A1.A0SB.01A.11D.A142.09	TCGA.A1.A0SD.01A.11D.A10Y.09	TCGA.A1.A0SE.01A.11D.A099.09
X16.88790292..	PIEZO1	*NA*	*NA*
X1.44476442..	SLC6A9	*NA*	*NA*
X17.7491739..	SOX15	*NA*	*NA*
X14.65266493..	SPTB	*NA*	*NA*
X2.46707888..	TMEM247	*NA*	*NA*
X19.36940860..	ZNF566	*NA*	*NA*
X19.42585066..	ZNF574	*NA*	*NA*
X7.149129243..	ZNF777	*NA*	*NA*
X10.61834851..	*NA*	ANK3	*NA*
X.41379771..	*NA*	CASK	*NA*
X9.34564705..	*NA*	CNTFR	*NA*
X4.15067943..	*NA*	CPEB2	*NA*
X20.23420938..	*NA*	CSTL1	*NA*
X.23956735..	*NA*	CXorf58	*NA*
X1.160210108..	*NA*	DCAF8	*NA*
X20.25755889..	*NA*	FAM182B	*NA*
X10.8115955.8115956..	*NA*	GATA3	*NA*
X12.112622620..	*NA*	HECTD4	*NA*
X1.158054332..	*NA*	KIRREL	*NA*
X19.51329938..	*NA*	KLK15	*NA*
X.153132281..	*NA*	L1CAM	*NA*
X4.151821369..	*NA*	LRBA	*NA*
X2.170070247..	*NA*	LRP2	*NA*
X5.140209165..	*NA*	PCDHA6	*NA*
X11.117063930..	*NA*	SIDT2	*NA*

Table 7. ***Subset of the “Hugo_Symbol” Mutation Assay for BRCA.*** As illustrated, each row represents a specific position on the chromosome where a mutation is present, and each column corresponds to an individual patient. As previously stated, each cell of this data frame may contain either an NA value or a gene symbol. The presence of a gene name in a cell indicates that a mutation occurred at that specific position on the patient’s chromosome, while an NA value signifies the absence of a mutation.

Our analysis revealed that a mutation at a given position could only occur in a single patient. Although different patients may have mutations in the same gene, there were no instances of mutations occurring at the same position among different patients in our data. This is logical, as mutations are rare events and the likelihood of multiple patients having mutations at the exact same position on their chromosomes is exceedingly low.

It should be noted that the preceding table (Table 7), as well as the subsequent table (Table 8), depict only a subset of the BRCA mutation data. However, the same structure is present for all five cancer types investigated in this study.

	TCGA.A1.A0SB.01A.11D.A142.09	TCGA.A1.A0SD.01A.11D.A10Y.09	TCGA.A1.A0SE.01A.11D.A099.09
X16.74425902..	Missense_Mutation	*NA*	*NA*
X22.16449539..	Missense_Mutation	*NA*	*NA*
X20.16730581..	Missense_Mutation	*NA*	*NA*
X.78216689..	Silent	*NA*	*NA*
X16.88790292..	Missense_Mutation	*NA*	*NA*
X1.44476442..	Missense_Mutation	*NA*	*NA*
X17.7491739..	Missense_Mutation	*NA*	*NA*
X14.65266493..	Missense_Mutation	*NA*	*NA*
X2.46707888..	Frame_Shift_Del	*NA*	*NA*
X19.36940860..	Frame_Shift_Del	*NA*	*NA*
X19.42585066..	Missense_Mutation	*NA*	*NA*
X7.149129243..	Missense_Mutation	*NA*	*NA*
X10.61834851..	*NA*	Missense_Mutation	*NA*
X.41379771..	*NA*	Missense_Mutation	*NA*
X9.34564705..	*NA*	Missense_Mutation	*NA*
X4.15067943..	*NA*	Missense_Mutation	*NA*
X20.23420938..	*NA*	Silent	*NA*
X.23956735..	*NA*	Missense_Mutation	*NA*
X1.160210108..	*NA*	Silent	*NA*
X20.25755889..	*NA*	Missense_Mutation	*NA*
X10.8115955.8115956..	*NA*	Frame_Shift_Ins	*NA*
X12.112622620..	*NA*	Missense_Mutation	*NA*
X1.158054332..	*NA*	Missense_Mutation	*NA*
X19.51329938..	*NA*	Missense_Mutation	*NA*
X.153132281..	*NA*	Missense_Mutation	*NA*
X4.151821369..	*NA*	Missense_Mutation	*NA*
X2.170070247..	*NA*	Missense_Mutation	*NA*
X5.140209165..	*NA*	Missense_Mutation	*NA*
X11.117063930..	*NA*	Missense_Mutation	*NA*
X1.95293148..	*NA*	Missense_Mutation	*NA*
X17.33884688..	*NA*	Missense_Mutation	*NA*
X20.17928176.17928178..	*NA*	In_Frame_Del	*NA*

Table 8. ***Subset of the "Variant_Classification" Mutation Assay for BRCA.*** As depicted, in this subset of the matrix from the "Variant_Classification" assay of BRCA, the row and column names are identical to those in the "Hugo_Symbol" assay. The only distinction between these assays lies in the values of each cell. While each cell of the "Hugo_Symbol" assay represents the gene name if a mutation is present at that specific position on the patient's chromosome, the "Variant_Classification" assay employs the type of mutation to indicate the presence of a mutation. In both assays, the absence of a mutation is denoted by an NA value.

Similar to the approach used for other omics data, another R function was created to prepare the data for further analysis. The first modification performed by this function involved utilizing the data contained within the "Variant_Classification" assay to identify and remove silent mutations. The rationale behind this decision was to improve the overall quality and accuracy of our data, which ultimately led to more reliable results. This is because silent mutations do not alter the final proteins encoded by a given gene. Thus, by leveraging the information contained within the "Variant_Classification" assay, we were able to accurately identify and exclude silent mutations from our subsequent analyses.

The subsequent modification performed by the R function involved utilizing the information contained within the "Hugo_Symbol" assay to accurately determine the gene symbol corresponding to each specific position on a patient's chromosome.

Subsequently, the R function performed an additional modification to the data frame by eliminating rows representing chromosomal positions where no mutations were present among the entire patient population. Following this step, the data structure was further modified by converting the use of gene symbols and NA values to represent the presence or absence of a mutation, respectively, to a binary representation using 1 and 0. In this new representation, if a mutation was present at a specific position on a patient's chromosome, the corresponding cell in the data frame would contain a value of 1. Conversely, if no mutation was present at that position, the cell would contain a value of 0.

Furthermore, it should be noted that the specific position of a mutation on a patient's chromosome was not required for our analyses. Instead, our focus was solely on determining whether or not a given gene was mutated in a particular patient. Consequently, the subsequent modification performed by our implemented R function involved collapsing multiple positions within a single gene to produce a binary representation indicating the presence or absence of a mutation. In this representation, if a mutation was present within a given gene in a specific patient, the corresponding cell in the data frame would contain a value of 1. Conversely, if no mutations were present within that gene in the corresponding patient, the cell would contain a value of 0.

The final two modifications performed by the R function were analogous to those applied to the previous omics data. Specifically, we needed to convert the gene symbols to their corresponding Entrez IDs and modify the patient IDs to ensure that they were consistent with the structure of patient names used in both the mutation and survival data frames (Table 10).

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:3983	1	0	0
ENTREZID:80070	1	0	0
ENTREZID:1838	1	0	0
ENTREZID:4437	1	0	0
ENTREZID:4602	1	0	0
ENTREZID:81061	1	0	0
ENTREZID:56914	1	0	0
ENTREZID:9780	1	0	0
ENTREZID:6536	1	0	0
ENTREZID:6665	1	0	0
ENTREZID:6710	1	0	0

Table 9. ***Subset of Final Mutation Data Frame for BRCA.*** As depicted in this table, each row of the final data frame represents a single gene, identified by its Entrez ID, while each column corresponds to an individual patient. Thus, for each gene, the value in the corresponding cell would be 1 if a mutation was present in that patient or 0 if no mutation was present. It should be noted that this table represents only a subset of the mutation matrix for BRCA and that data frames with an identical structure were created for the other cancer types investigated in this study.

2.7 Modifying Survival Data for Analysis with MOSClip

For the purposes of our survival analysis, it was deemed necessary to utilize survival data in order to accurately assess the outcomes of patients with different cancer types. To this end, we elected to use progression-free survival (PFS) events as our primary measure of survival outcomes (Liu et al., 2018). More specifically, the occurrence of an event in this context could indicate a number of different outcomes, including recurrence of the cancer, the development of a new primary tumor, distant metastasis, progression of the tumor, or death. Conversely, the absence of an event would represent a censored patient who had discontinued follow-up for any number of reasons. In such cases, no further information is available about the patient's status or condition.

It is important to note that censored patients, who have discontinued follow-up, may be either alive or deceased. In such cases, no further information is available about their status or condition. As a result, it is not possible to determine with certainty whether or not these patients have experienced an event such as recurrence, the development of a new primary tumor, distant metastasis, progression of the tumor, or death. In light of this uncertainty, we utilized the survival data provided by Liu et al., 2018 to extract

progression-free survival (PFS) information for the five cancer types of interest in our study.
Furthermore, in order to facilitate our survival analyses, we utilized our implemented R function to create a survival data frame for each cancer type under investigation. This data frame consisted of a matrix representing the status of each individual patient and the number of days that had elapsed from the start of follow-up until either the occurrence of an event or the last follow-up for censored patients. In cases where an event occurred, the status of the corresponding patient would be represented by a value of 1. Conversely, if a patient was censored and discontinued follow-up, their status would be represented by a value of 0. Additionally, the data frame also indicated the number of days that had elapsed from the start of follow-up until either the occurrence of an event or the last follow-up for censored patients (Table 10).

	status	days
TCGA.A1.A0SB	0	259
TCGA.A1.A0SE	0	1321
TCGA.A1.A0SF	0	1463
TCGA.A1.A0SG	0	434
TCGA.A1.A0SH	0	1437
TCGA.A1.A0SI	0	635
TCGA.A1.A0SJ	0	416
TCGA.A1.A0SK	1	967
TCGA.A1.A0SN	0	1196
TCGA.A1.A0SP	0	584
TCGA.A1.A0SQ	0	554
TCGA.A2.A04R	0	3709

Table 10. ***Subset of Final Survival Data Frame for BRCA.*** As depicted, each row corresponds to an individual patient and contains two types of information: status and days. The status column indicates whether or not an event occurred, with values of 1 and 0 representing the presence or absence of an event, respectively. The days column, on the other hand, represents the number of days that elapsed from the start of follow-up until either the occurrence of an event or the last follow-up for censored patients.

In conclusion, the PFS survival data for each patient was extracted from the survival data provided by Liu et al. (2018) for the five selected cancer types. The input for our implemented function is the Excel file created by Liu et al. (2018) and the specific cancer type for which we wish to extract survival data. The output is a cleaned survival data frame for each cancer type, with

NA values removed and patient IDs modified by replacing “-” signs with “.” signs.

2.8 Patient Selection and Pathway Retrieval from Reactome Database

The subsequent modification involved selecting only those patients for whom all omics data were available. This ensured that further analyses and comparisons were more accurate by including only patients with available expression, CNV, methylation, mutation, and survival data.
Subsequently, Homo Sapiens pathways were retrieved from the Reactome database. Gene modifiers were then converted to Entrez IDs for further analysis. Specifically, the downloaded Reactome pathways contained 2439 entries and were retrieved on April 18, 2023.

2.9 MOSClip Survival Analysis

After acquiring and modifying the multi-omics data for five cancer types, the subsequent step involved utilizing the MOSClip package to conduct survival tests on pathways and modules (Martini et al., 2019).

2.9.1 MOSClip Pathway Analysis

To execute MOSClip’s multi-omics survival pathway test, it was necessary to specify the modified multi-omics matrix, the dimensionality reduction strategy for each omic type, the survival data for each cancer type, and the Reactome pathways for pathway analysis. Given that Reactome pathways can comprise numerous pathways with few nodes (genes), we filtered the downloaded pathways from the Reactome database based on pathway size to expedite and enhance analysis accuracy. Consequently, only Reactome pathways with a minimum of ten nodes were used, while those with fewer than ten nodes were discarded. Specifically, for filtering Reactome pathways for each cancer type, only genes with available expression data were utilized. Thus, the filtered Reactome for each cancer type comprised pathways with at least ten genes with available expression data for that specific cancer type.
As previously stated, MOSClip conducts survival tests on dimensionality-reduced data (Martini et al., 2019). A detailed explanation of how MOSClip performs its analysis can be found in Figure 1. Consequently, it was necessary to select a dimensionality reduction strategy for each omic type. In accordance with Martini et al., 2019, we opted for the “summarize With Pca” method for gene expression, the “summarize In Cluster” method for methylation data, and the “summarize To Binary Events” method for both mutation and CNV data.
In the “Summarize with PCA” method, gene expression data for a specific pathway topology is modeled using a directed acyclic graph (DAG) and a

graphical model with a normal distribution and a concentration matrix (Martini et al., 2019). The maximum likelihood estimate of the covariance matrix is obtained using the Iterative Proportional Scaling (IPS) algorithm. Principal component analysis is then performed using the spectral decomposition of the IPS-estimated covariance matrix. For small sample sizes, a shrinkage approach is used to estimate the sample covariance matrix. The number of principal components to be selected is estimated using a cross-validation approach. When dimension reduction is performed on modules of the graph, a sparse PCA is implemented.

Hierarchical cluster analysis, “summarize In Cluster”, was used to identify the optimal number of clusters using the NbClust R package (Martini et al., 2019). The Silhouette index was used by default, but other validity measures can be selected. Patients were then classified into groups based on the optimal number of clusters. The numerical matrix was summarized with a vector reporting the assigned cluster for each patient.

In the “summarize To Binary Events” method, the binary matrix was summarized with a sample binary vector (Martini et al., 2019). A value of 1 was assigned if at least one gene in the pathway or module is mutated, amplified, or deleted, and 0 otherwise. This strategy is used for both pathway and module analyses.

The remaining MOSClip survival test settings were retained as default. With the requisite data for various omics, survival data, and dimensionality reduction strategy in place, the final variable to be determined for MOSClip’s survival pathway test was the genes to be analyzed. Consequently, only genes with available expression data were input into the MOSClip survival test function.

The MOSClip survival pathway test yielded a matrix with pathway names as row names and p-values as column names. Each row of the output matrix corresponded to a single pathway input into the MOSClip function as a filtered Reactome pathway file, while each column indicated the statistical significance of each individual omic. Specifically, for each pathway, an overall p-value denoted the general significance level of that specific pathway, while the remaining columns indicated the p-value of each dimensionality-reduced omic. A significant overall p-value for a single pathway denoted a significant association between that pathway and patient survival for that specific cancer. Additionally, the p-value of each omic enabled the determination of which omic types were significantly associated with patient survival. To illustrate the structure of the aforementioned matrix, a subset of the data frame corresponding to running the survival pathway test on BRCA is presented below (Table 11).

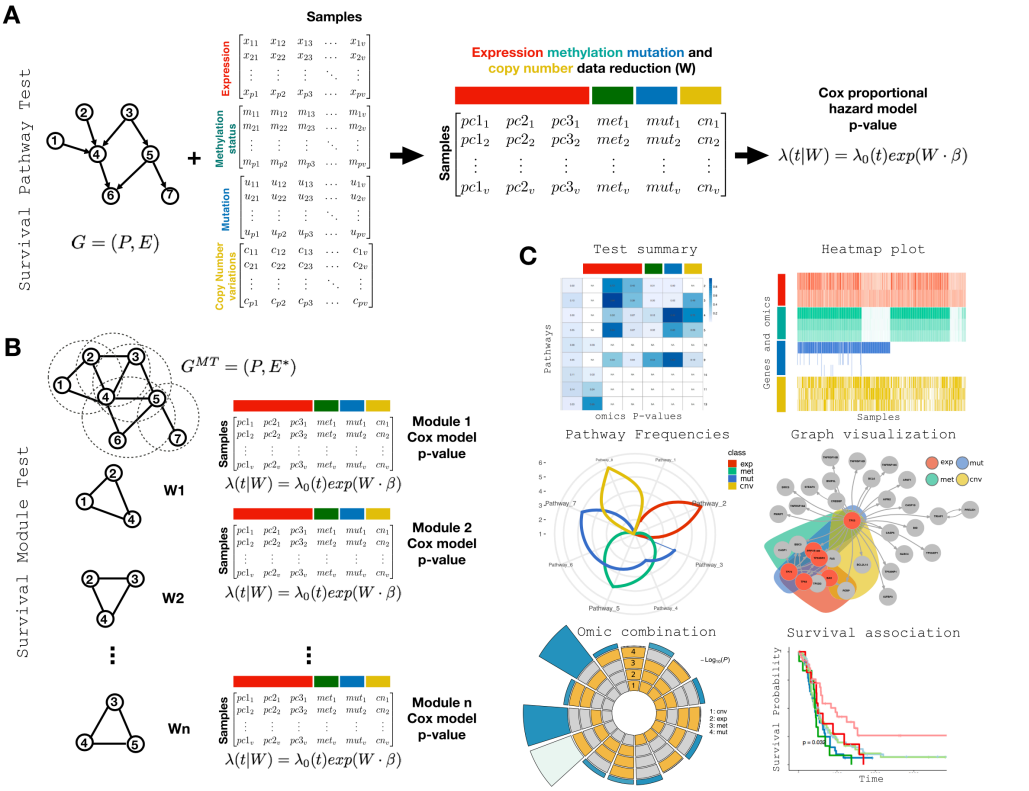

Figure 1. MOSClip Analysis Process. This figure illustrates the process by which MOSClip performs its analysis. It was reproduced from the MOSClip paper (Martini et al., 2019).

As illustrated in Figure 1, part (A) illustrates the survival pathway analysis. By utilizing the graph topology G and matrices for gene expression, methylation, copy number, and mutations, matrix W is generated through the application of dimension reduction. This matrix consists of reduced omic vectors with patient classes derived from PCA, hierarchical clustering, and binary/vote counting techniques. Subsequently, a multivariate Cox proportional hazard model is employed with matrix W serving as covariates. The resulting P-value for the full model is then provided. Then, part (B) demonstrates the survival module analysis. Following the moralization and, if required, triangularization of the graph, modules (or maximal cliques) are determined. The analysis outlined in panel (A) is then applied to each individual module. Finally, part (C) shows an overview of MOSClip graphical tools. The MOSClip package offers a range of tools including module/pathway ranking, heatmaps, radial plots, graph visualizations, omic combination summaries, and Kaplan-Meier curves with log-rank tests.

	pvalue	cnvTRUE	expPC1	expPC2	expPC3	met2k2	met3k2	met3k3	mutTRUE	resamplingCount
NOTCH3 Intracellular Domain Regulates Transcription	4.055173e-06	0.003511592	2.606325e-07	8.600566e-01	1.019357e-01	0.0006342617	NA	NA	3.172034e-02	10
Activation of the mRNA upon binding of the cap-bind...	3.668970e-05	0.702733871	2.628617e-02	3.518204e-01	4.415768e-03	0.0072511336	NA	NA	2.091701e-03	10
Other interleukin signaling	6.519542e-05	0.048819856	2.099935e-06	8.635563e-01	1.595681e-02	0.0851773753	NA	NA	2.762590e-01	10
TBC/RABGAPs	1.073932e-04	0.149712008	2.111115e-04	8.340975e-01	8.033086e-03	0.7143289342	NA	NA	6.651644e-03	10
DAP12 interactions	1.116292e-04	0.013587051	3.176905e-02	7.416000e-05	7.694236e-01	0.1815341186	NA	NA	3.105144e-02	10
Chemokine receptors bind chemokines	1.701839e-04	0.062084249	7.325588e-01	7.328992e-06	6.653649e-01	0.0611226649	NA	NA	5.792617e-01	10
Diseases associated with glycosaminoglycan metaboli...	1.902181e-04	0.045087002	1.143952e-04	6.672415e-05	1.212137e-03	0.1010718593	NA	NA	2.240139e-01	10
TP53 Regulates Transcription of Cell Death Genes	1.963234e-04	0.008510527	1.128509e-02	1.250746e-02	5.306772e-01	0.0288128514	NA	NA	1.619885e-01	10
TNFs bind their physiological receptors	1.986151e-04	0.276078485	4.709726e-05	1.646293e-01	3.284011e-02	0.1240606625	NA	NA	3.609472e-01	10
Role of phospholipids in phagocytosis	2.023004e-04	0.032310171	6.405904e-01	5.853827e-03	8.811449e-04	0.6077173023	NA	NA	3.942196e-02	10
FLT3 Signaling	2.071095e-04	0.052490473	2.832552e-05	1.051402e-02	6.952504e-01	0.6993136095	NA	NA	8.698067e-02	10
Cellular responses to stress	2.359428e-04	0.787810242	5.072161e-02	4.977270e-03	3.325808e-02	0.0046524452	NA	NA	9.842259e-01	10
Repression of WNT target genes	2.420620e-04	0.143700335	8.846401e-02	9.501050e-05	7.069929e-01	0.0051106534	NA	NA	5.309391e-02	10
Nucleotide salvage	2.680410e-04	0.025934878	2.040300e-05	3.806882e-02	1.970499e-01	0.6056282986	NA	NA	8.066063e-01	10
Antigen activates B Cell Receptor (BCR) leading to gen...	2.865518e-04	0.821099571	1.273792e-04	1.120937e-02	4.948043e-01	0.2533954671	NA	NA	9.965165e-01	10
Biosynthesis of DHA-derived SPMs	2.958945e-04	0.005899951	1.019755e-03	2.988358e-01	9.500322e-02	0.2339541902	NA	NA	4.036546e-02	10
Regulation of FZD by ubiquitination	4.038142e-04	0.376819370	4.816133e-01	9.725092e-02	3.342424e-05	0.2407509310	NA	NA	5.172678e-02	10
Intrinsic Pathway for Apoptosis	4.415739e-04	0.061527894	4.484388e-01	5.884835e-04	2.348406e-01	0.0426731630	NA	NA	3.227337e-01	10
RND2 GTPase cycle	5.012961e-04	0.049895591	1.334230e-01	6.513905e-04	5.641190e-01	0.0006899736	NA	NA	6.608140e-01	10
TP53 regulates transcription of several additional cell ...	5.218224e-04	0.002373538	1.165495e-02	8.830338e-03	4.769351e-01	0.7647940402	NA	NA	5.235211e-02	10
Interleukin-10 signaling	6.074701e-04	0.470317688	1.217373e-02	1.186261e-04	8.491654e-01	0.7449717996	NA	NA	1.883891e-01	10
Cellular responses to stimuli	6.306671e-04	0.776641023	3.321802e-02	3.800105e-03	3.185254e-02	0.0160926346	NA	NA	8.336208e-01	10
Interleukin-2 family signaling	6.857227e-04	0.134485504	1.206963e-02	9.290783e-04	4.545426e-01	0.6563845621	NA	NA	6.153843e-03	10
LXRs regulate gene expression linked to cholesterol tr...	6.955058e-04	0.395895359	4.771977e-04	8.261769e-01	8.134072e-03	0.0433052268	NA	NA	4.044801e-01	10
Interleukin-2 signaling	7.060885e-04	0.103977165	2.678841e-02	4.455415e-03	4.982827e-01	0.2072486237	NA	NA	5.602157e-03	10
Death Receptor Signalling	7.251175e-04	0.348379979	9.317611e-01	5.223301e-01	7.852393e-04	0.0026722187	NA	NA	3.168249e-01	10
Hedgehog ligand biogenesis	7.562900e-04	0.186277689	2.195781e-02	9.489180e-03	7.336048e-04	0.5043025564	NA	NA	7.256091e-01	10
FLT3 signaling in disease	7.897954e-04	0.929787773	9.312908e-05	4.453407e-02	1.650390e-01	0.4094960107	NA	NA	2.489468e-01	10

Table 11. ***Subset of MOSClip Multi-Omics Survival Pathway Test Output for BRCA*****.** As illustrated, the MOSClip package's survival pathway test reveals the association between each pathway and omic and patient survival. In the given data, each row represents a pathway. The first column indicates the overall p-value for each pathway. Subsequent columns show the p-values for each omic type, indicating the significance of correlation between alteration in different omics and patient survival. These include p-values for alteration in CNV, principal component 1-3 of expression, different methylation clusters (e.g., met2K2, met3K2, met3K3), mutation, and resampling score. This matrix was generated using the survival test on BRCA, while data frames for other cancer types exhibit a similar structure.

Due to considerable gene redundancy within pathways and modules, the independence of obtained P-values is compromised (Martini et al., 2019). To address this issue, enhance pathway and module selection reliability, and effectively control false positives, and False Discovery Rate (FDR) calculation a resampling strategy was implemented. Specifically, another MOSClip function was utilized to execute this resampling approach to augment analysis significance. A subset of the previous matrix, filtered to include only pathways with an overall p-value of 0.05 or less, was input into the MOSClip resampling function. After ten permutation cycles, only pathways significant in eight or more permutations were selected. Consequently, another matrix with the same format and structure as the previous one (Table 11) was generated, containing only pathways with an overall p-value of 0.05 or less and a resampling score of at least 8.

2.9.2 MOSClip Module Analysis

The MOSClip survival module test closely resembled the survival pathway test in many ways. A module can be defined as a connected component of the pathway graph chosen automatically by specific algorithms. Specifically, the survival module test utilized multi-omics data, survival data, dimensionality reduction techniques for each omic type (employing the same strategies as those used for pathway analysis), and Reactome pathways. However, despite these similarities, there were three crucial steps that differentiated the survival module test from the survival pathway test.

The first crucial step that differentiated the survival module test from the survival pathway test involved the selection of pathways for analysis. Instead of conducting survival analysis on entire Reactome pathways with at least 10 genes with available expression data, we took a more targeted approach. Specifically, we performed survival analysis only on pathways that were significantly associated with patient survival based on the output of the survival pathway test. As a result, the survival module test was executed solely on pathways that had been selected as significantly associated with patient survival in the previous step.

The second crucial step that differentiated the survival module test from the survival pathway test involved the selection of dimensionality reduction techniques. As recommended by Martini et al. (2019), we employed the "summarize With Pca" strategy without covariance matrix shrinkage and using the "topological" PCA method. This approach contrasted with the methodology utilized for the survival pathway test, where we employed the sparse PCA method with covariance matrix shrinkage as suggested by Martini et al. (2019).

Finally, after executing the resampling step and selecting modules with an overall p-value of 0.05 or less and a resampling score of at least 8, we applied the Holm method to correct p-values to augment test significance. For all cancers except UCS, a significant p-value cut-off of 0.1 was chosen after correction, while for the UCS group, a cut-off value of 0.15 was selected.

It is important to note that all other factors and steps involved in the execution of the survival module test were identical to those utilized in the survival pathway test. In other words, apart from the aforementioned differences, the methodology employed in the survival module test mirrored that of the survival pathway test. Furthermore, it is worth mentioning that a subset of the final matrix generated after performing the survival module test on BRCA is presented in Table 12 for reference.

	pathway	module	pvalue	cnvTRUE	expPC1	expPC2	expPC3	met2k2	met3k2	met3k3	mutTRUE	resamplingCount
Amino acid and derivative metabolism.95	Amino acid and derivative metabolism	95	0.0007540064	0.011404558	8.549297e-04	NA	NA	3.936877e-07	NA	NA	NA	10
Vesicle-mediated transport.44	Vesicle-mediated transport	44	0.0102876803	0.120645299	2.910385e-01	0.06448534	1.181921e-03	4.392970e-06	NA	NA	0.128495997	10
Activation of the mRNA upon binding of the cap-bind...	Activation of the mRNA upon binding of the cap-bind...	1	0.0277346523	0.617952579	9.389055e-03	0.16683879	4.069335e-03	2.533109e-03	NA	NA	0.002115334	10
Metabolism.177	Metabolism	177	0.0320636751	0.317630157	2.895083e-01	0.02265052	8.973200e-01	1.473336e-06	NA	NA	0.188660465	10
Eukaryotic Translation Initiation.3	Eukaryotic Translation Initiation	3	0.0338873606	0.778305025	2.551388e-03	0.09180220	1.519412e-02	3.579050e-03	NA	NA	0.004720828	10
Signaling Pathways.91	Signaling Pathways	91	0.0339164111	0.239293707	1.013728e-04	0.48649773	1.973846e-02	2.452289e-02	NA	NA	0.696235017	10
Metabolism.152	Metabolism	152	0.0486727446	0.030949490	2.837267e-01	0.20347087	5.519425e-05	5.044914e-02	NA	NA	0.030250701	10
Signaling Pathways.239	Signaling Pathways	239	0.0681469220	0.001668627	3.514633e-07	0.33642600	8.057731e-01	5.858204e-01	NA	NA	0.630524660	10
Signaling by GPCR.14	Signaling by GPCR	14	0.0681469220	0.001668627	3.514633e-07	0.33642600	8.057731e-01	5.858204e-01	NA	NA	0.630524660	10
Nervous system development.58	Nervous system development	58	0.0746393913	0.039681013	1.570345e-06	0.06455248	3.285133e-04	8.704270e-02	NA	NA	0.314027109	10
Axon guidance.58	Axon guidance	58	0.0746393913	0.039681013	1.570345e-06	0.06455248	3.285133e-04	8.704270e-02	NA	NA	0.314027109	10
Cap-dependent Translation Initiation.3	Cap-dependent Translation Initiation	3	0.0775960217	0.890990312	3.630133e-03	0.14959574	7.146032e-03	7.239875e-03	NA	NA	0.002325357	10
Nervous system development.70	Nervous system development	70	0.0817285244	0.103753764	2.436284e-06	0.97081153	2.274416e-02	1.506387e-02	NA	NA	0.283224239	10
Axon guidance.68	Axon guidance	68	0.0817285244	0.103753764	2.436284e-06	0.97081153	2.274416e-02	1.506387e-02	NA	NA	0.283224239	10
Metabolism.151	Metabolism	151	0.0844381142	0.027306496	6.588593e-01	0.20568402	4.218217e-05	6.070633e-02	NA	NA	0.040480298	10
Signaling by FLT3 ITD and TKD mutants.4	Signaling by FLT3 ITD and TKD mutants	4	0.0925669787	0.868743349	4.713110e-04	NA	NA	5.499278e-03	NA	NA	NA	10
GPCR downstream signalling.9	GPCR downstream signalling	9	0.0949728246	0.227470943	1.075955e-04	0.40476311	1.453112e-01	2.229121e-03	NA	NA	0.985535004	10
Extracellular matrix organization.48	Extracellular matrix organization	48	0.0990356132	0.703568652	1.915564e-03	0.04834151	3.406001e-05	5.320126e-02	NA	NA	0.017045262	10

Table 12. *Subset of MOSClip Multi-Omics Survival Module Test Output for BRCA*. Each row of the data frame represents a module denoted by the pathway name and module number. The first two columns indicate the pathway name and module number, while the next column displays the overall p-value for each module. Subsequent columns show the p-values for each omic type, indicating the significance of correlation between alteration in different omics and patient survival. These include p-values for alteration in CNV, principal component 1-3 of expression, different methylation clusters (e.g., met2K2, met3K2, met3K3), mutation, and resampling score.

In summary, we began with a data frame for each omic type of each cancer, in addition to information regarding the survival of each patient with a specific cancer. Utilizing the MOSClip survival pathway test, we identified pathways significantly associated with patient survival for a specific cancer. Subsequently, another MOSClip function was employed to conduct a survival module test on the significant pathways selected in the previous step. This enabled us to determine which modules of a pathway were significantly associated with patient survival for a specific cancer. From this point forward, we endeavored to analyze and visualize these outputs in a practical manner.

2.10 Constructing an Optimal Data Frame for Network Analysis using Cytoscape

After identifying significant pathways and modules using MOSClip, we expanded our analysis to examine the genes within each significant module. Specifically, we aimed to identify genes significantly associated with patient survival in each type of cancer. To achieve this, we developed a function in R that takes the output of the “module survival test” and Reactome database pathways for Homo sapiens as input and generates a data frame that can be visualized in Cytoscape to display the significant genes and their interactions.

The function we implemented performs several steps to generate a data frame suitable for visualization in Cytoscape. First, it creates a “graphNEL” object for each significant pathway of a specific cancer type to represent the pathway’s topology. This “graphNEL” object is then converted to an “igraph” object to improve the representation of edges between different genes. Next, genes significantly associated with patient survival are selected based

on the MOSClip survival test output and extracted from the “igraph” object. The gene IDs are then converted from Entrez to symbol and the type of multi-omics associated with patient survival is added to the data frame. Finally, to eliminate modules that are too dense we decided to define a cut-off threshold. So, we filter modules based on the average ratio of edges to nodes in cases where the number of edges per node is excessively high. While this results in some loss of information, it enables practical visualization in Cytoscape.

Thus, to determine an unbiased cut-off for the threshold of the average number of edges per node, we performed several analyses. We implemented a function in R to visualize the average number of edges per node in each “subgraph” generated based on the pathway topology and significant genes. As shown in Figure 2 for BRCA cancer, most modules have fewer than 40 edges per node on average, with only a small portion having more than 40 edges per node on average. Based on the thorough and comprehensive analysis that has been conducted, it would be reasonable and logical to conclude that an appropriate cut-off point for BRCA cancer would be to remove any module that have an average of more than 40 edges per node, as this would provide the most accurate and reliable results.

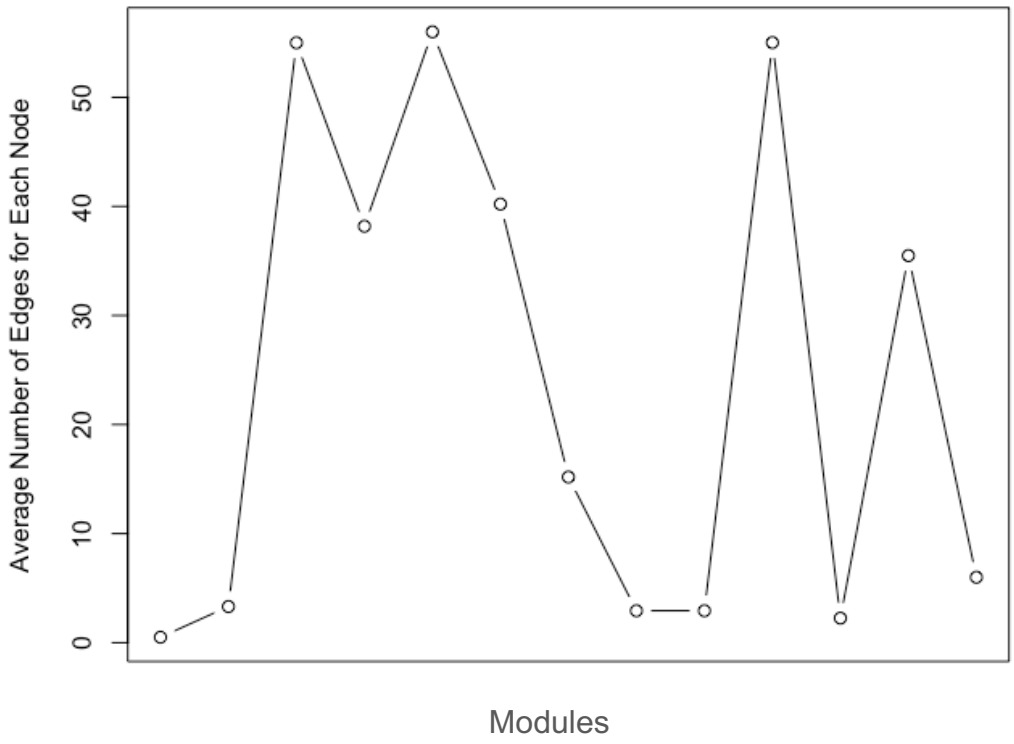

Figure 2. ***Distribution of Average Edges per Node in BRCA Modules.***

As illustrated in the figure, the majority of the modules have fewer than 40 edges per node on average. Based on this observation, we decided to exclude all modules with an average of more than 40 edges per node.
To further examine the distribution of the ratio of edges to nodes, we analyzed the box plot of the average ratio (Figure 3).

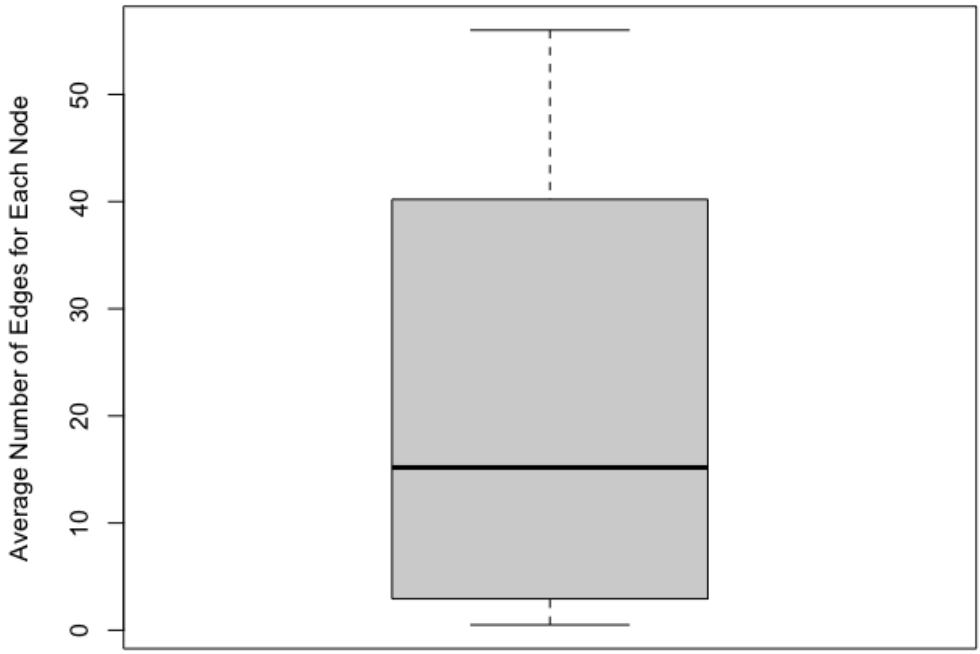

Figure 3. ***Box Plot of Average Edges per Node in BRCA Modules.*** As demonstrated in the previous figure, this plot confirms that the majority of modules have fewer than 40 edges per node on average. Thus, based on our analysis of the BRCA group, 40 is the optimal cut-off.

After determining the optimal cut-off for the average number of edges per node in BRCA cancer modules, we expanded our analysis to include four other types of cancer. Our goal was to assess whether a cut-off of 40 would also be suitable for these additional cancer types. To do this, we examined plots of the average number of edges per node for each type of cancer (Figure 4, Figure 5, Figure 6, Figure 7). These plots suggested that a cut-off of 40 would indeed be an excellent choice for our analysis. Furthermore, using this cut-off value would not result in any loss of information for the other types of cancer we examined. Only a few BRCA modules would be excluded from visualization in Cytoscape.
In conclusion, as is often the case in data analysis and visualization, we were faced with a trade-off between the amount of information we could represent and the practicality and usefulness of the resulting visualization. After careful consideration and analysis of the data, we ultimately made the decision to sacrifice some of the BRCA modules in order to create a more effective and informative Cytoscape visualization.

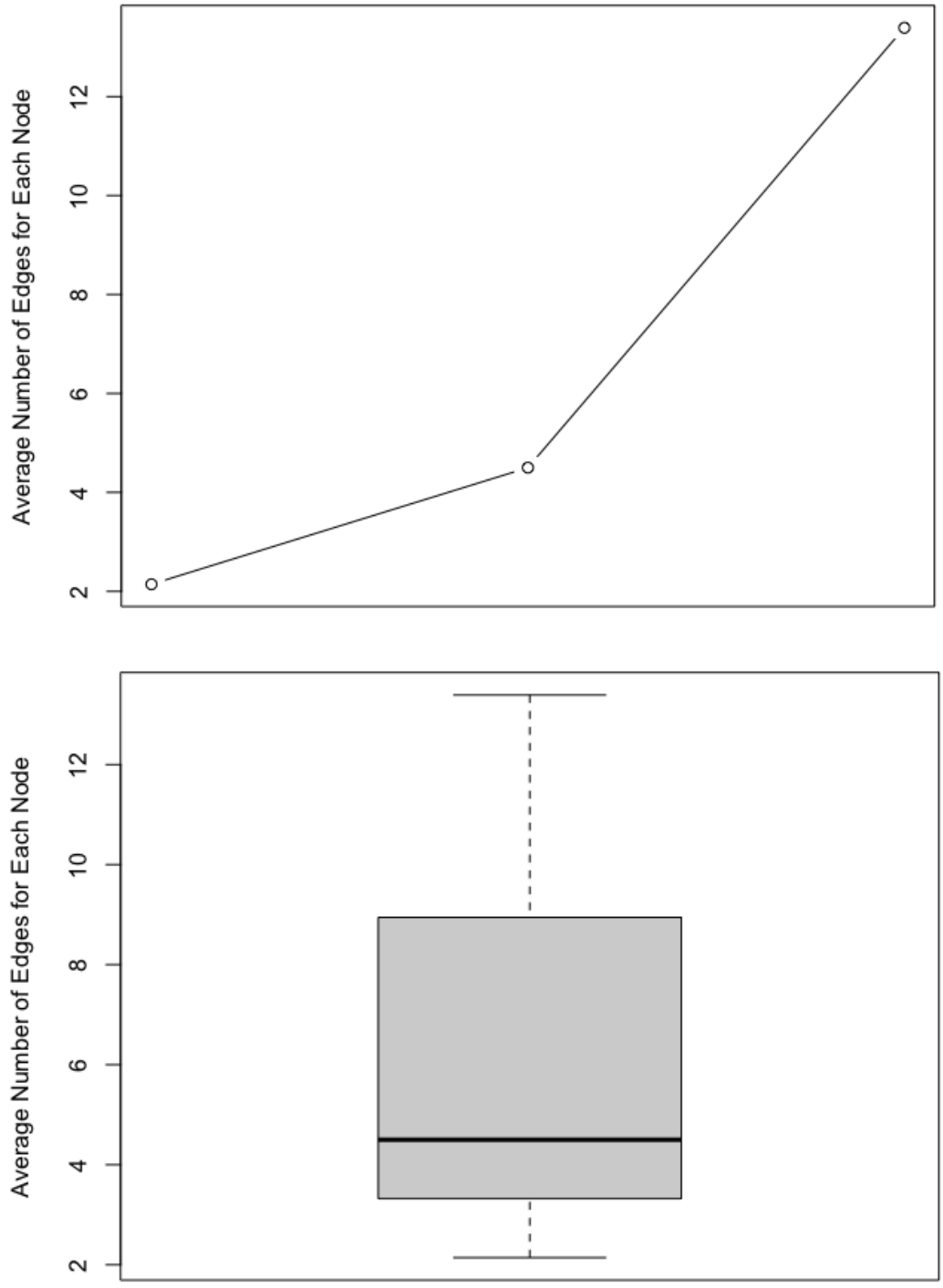

Figure 4. ***Distribution of Average Edges per Node in CESC Modules.*** As shown in the figure, all three CESC cancer modules have fewer than 40 edges per node on average. Therefore, setting the cut-off to 40 would not affect the inclusion of CESC modules of significant genes.

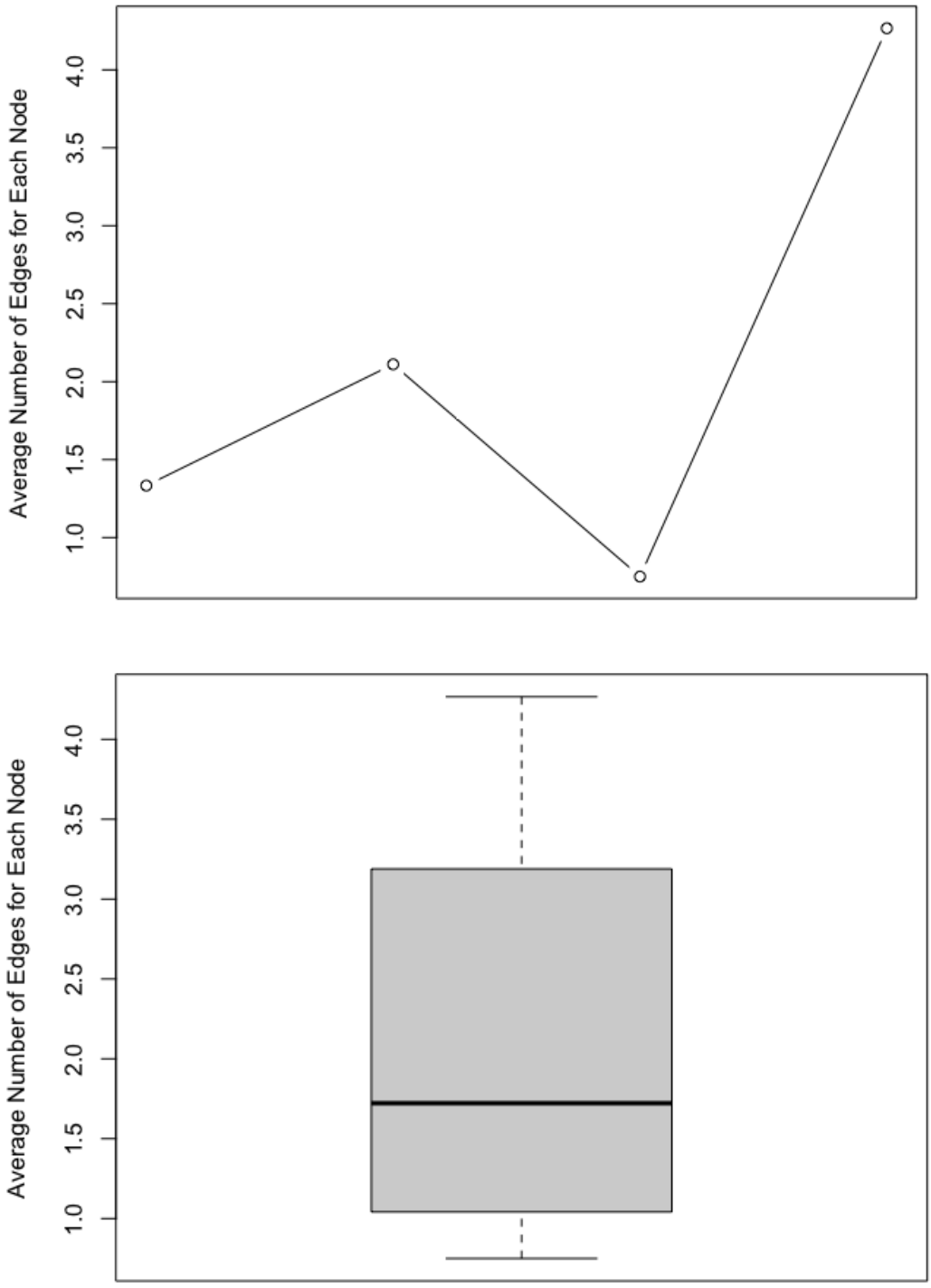

Figure 5. ***Distribution of Average Edges per Node in OV Modules.*** As shown in the figure, a cut-off threshold of 40 would not affect the inclusion of OV modules. The range of the average number of edges per node in OV cancer is acceptable and no changes are required.

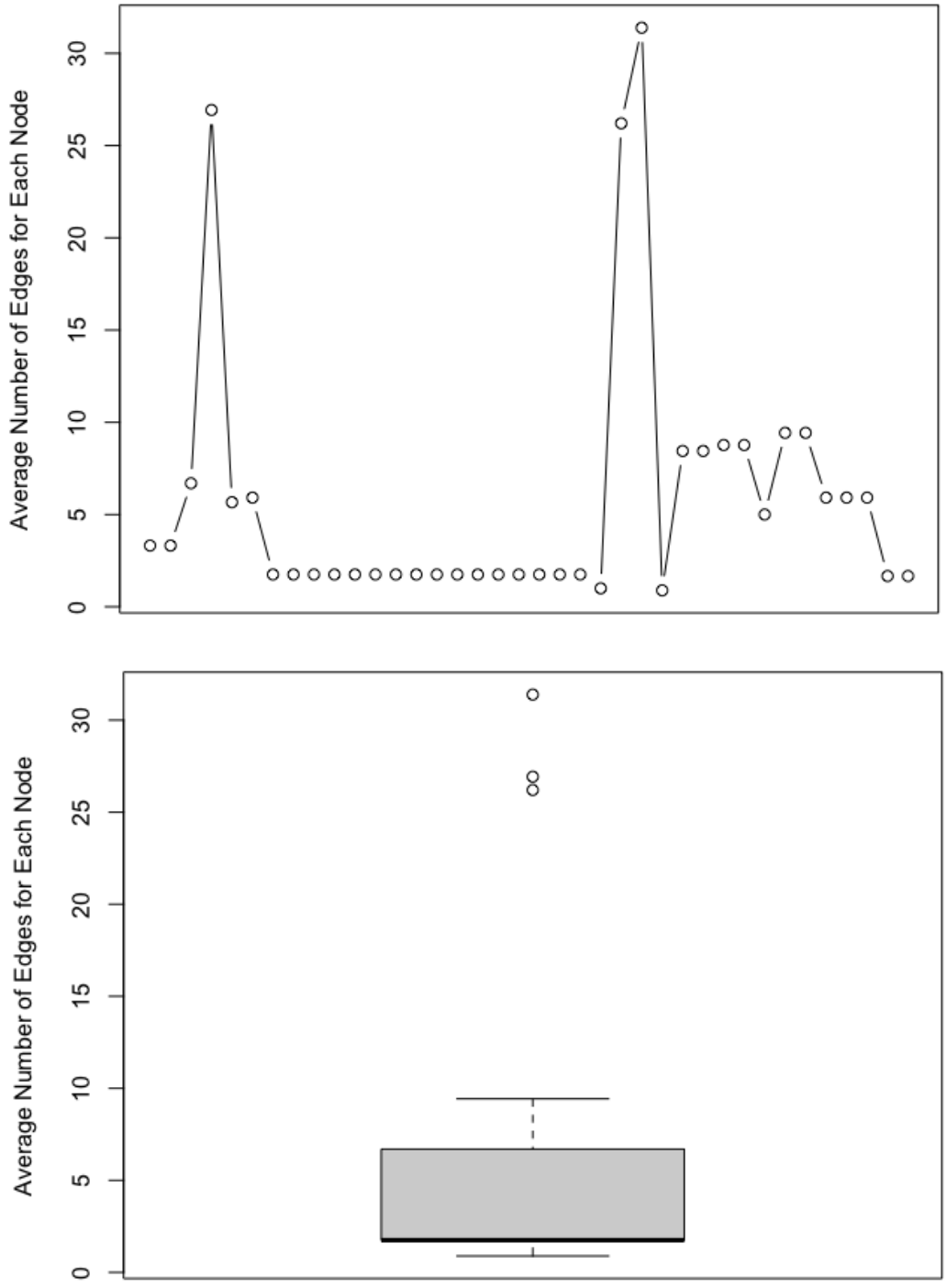

Figure 6. ***Distribution of Average Edges per Node in UCEC Modules.*** As demonstrated in the figure, all UCEC group modules have fewer than 40 edges per node on average. Therefore, setting the cut-off to 40 would not exclude any UCEC modules.

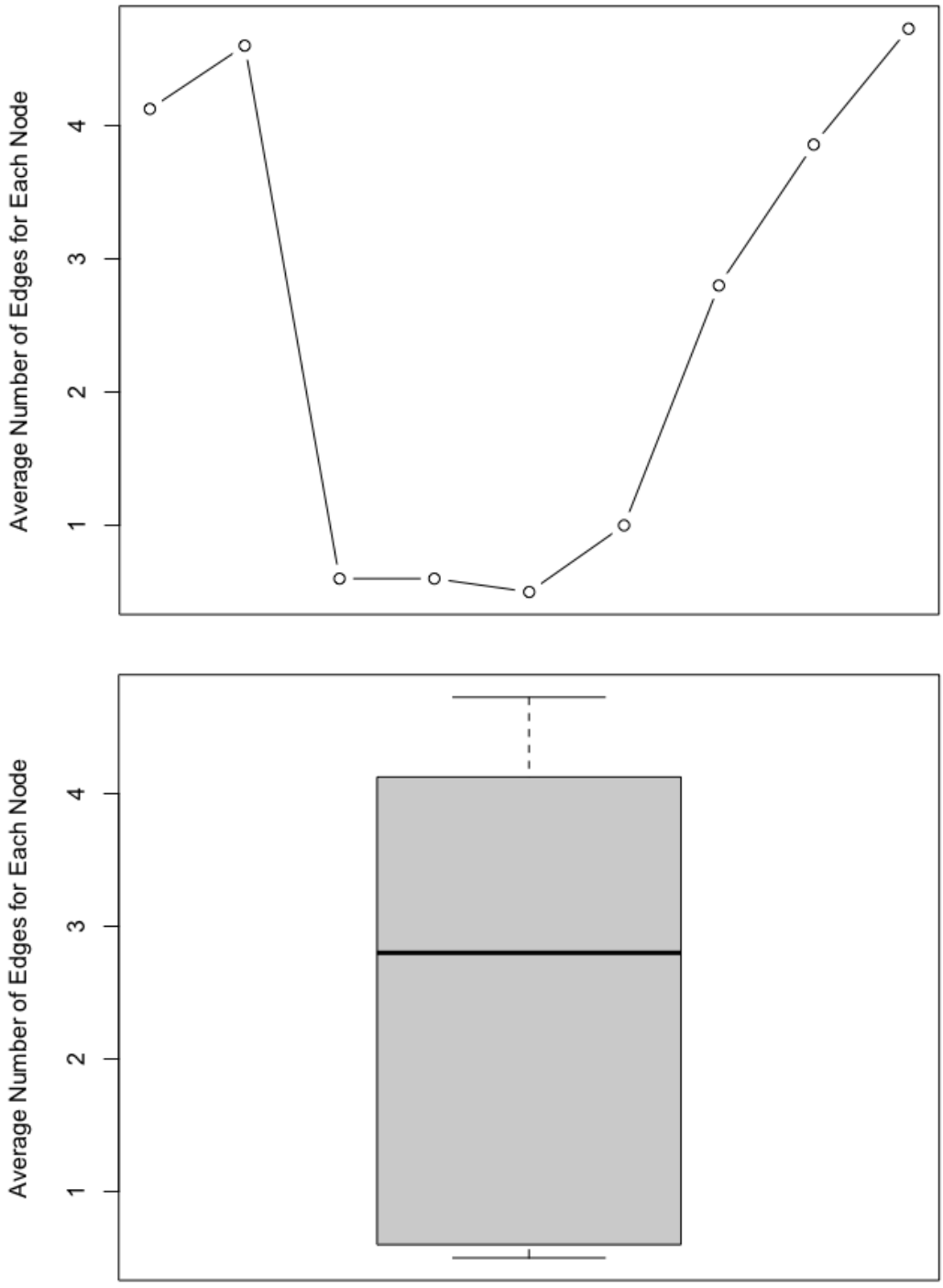

Figure 7. ***Distribution of Average Edges per Node in UCS Modules.*** As shown in the figure, the final type of cancer we examined, UCS, would not be affected by a cut-off threshold of 40 for module size. This is consistent with our findings for the previous three cancer types (CESC, OV, and UCEC).

In conclusion, the R function we implemented generates a data frame suitable for visualization in Cytoscape by performing the necessary modifications and filtering, as previously described in detail. With this data frame (Table 13), we can now visualize the genes significantly associated with patient survival in different types of cancer using Cytoscape.

	X1	X2	cancer	pathway	module	omics
17123	L1CAM	ITGB1	BRCA	Axon guidance	58	cnv-exp
17124	ITGA10	ITGB1	BRCA	Axon guidance	58	cnv-exp
17125	ANK1	L1CAM	BRCA	Axon guidance	58	cnv-exp
17126	ITGA5	L1CAM	BRCA	Axon guidance	58	cnv-exp
17127	ITGA9	L1CAM	BRCA	Axon guidance	58	cnv-exp
17128	ITGAV	L1CAM	BRCA	Axon guidance	58	cnv-exp
17129	ITGB1	L1CAM	BRCA	Axon guidance	58	cnv-exp
17130	NRP1	L1CAM	BRCA	Axon guidance	58	cnv-exp
17131	CHL1	ITGA10	BRCA	Axon guidance	58	cnv-exp
17132	ITGB1	ITGA10	BRCA	Axon guidance	58	cnv-exp
17133	CHL1	NRP1	BRCA	Axon guidance	58	cnv-exp
17134	L1CAM	NRP1	BRCA	Axon guidance	58	cnv-exp
17135	EZR	DCC	BRCA	Nervous system development	70	exp-met
17136	EZR	L1CAM	BRCA	Nervous system development	70	exp-met
17137	EZR	PRKCQ	BRCA	Nervous system development	70	exp-met
17138	DCC	EZR	BRCA	Nervous system development	70	exp-met
17139	L1CAM	EZR	BRCA	Nervous system development	70	exp-met
17140	EZR	DCC	BRCA	Axon guidance	68	exp-met
17141	EZR	L1CAM	BRCA	Axon guidance	68	exp-met
17142	EZR	PRKCQ	BRCA	Axon guidance	68	exp-met
17143	DCC	EZR	BRCA	Axon guidance	68	exp-met
17144	L1CAM	EZR	BRCA	Axon guidance	68	exp-met
17145	GLUL	ADA	BRCA	Metabolism	151	cnv-exp-mut
17146	GLUL	GNPDA1	BRCA	Metabolism	151	cnv-exp-mut
17147	FPGS	AASS	BRCA	Metabolism	151	cnv-exp-mut
17148	GLS2	AASS	BRCA	Metabolism	151	cnv-exp-mut
17149	GLS	AASS	BRCA	Metabolism	151	cnv-exp-mut

Table 13. ***BRCA Data Frame Subset for Cytoscape Visualization.*** Since the data frames for all types of cancer have the same general structure, we have only illustrated a subset of the BRCA matrix. As shown in the table, the columns represent the source node, target node, cancer type, name of the pathway, module number, and the type of omics significant for that pair of genes, respectively.

2.11 Generation of Survival Heatmap Matrix

First, we extracted all significant genes within significant modules for each cancer type. For each cancer, we created a list of genes significantly associated with patient survival. We then used the cleaned version of multi-omics data downloaded from "curatedTCGAData" and extracted the significant genes identified in the first step from each omics data frame. For example, in the expression data frame, each row corresponds to a single gene and each column to a patient (Table 4). We selected only rows containing genes significantly associated with patient survival and added the "exp" tag to the end of each gene name using a dot as a separator (Table 14). This procedure was repeated for different omics and cancer types, resulting in four data frames for each cancer type, one for each omics type. These data frames have row names consisting of gene names plus dot plus omics type and column names representing patients. A subset of different omics data frames for BRCA is shown in Table 14, Table 15, Table 16, and Table 17. Since the structure for other cancer types is identical, we have only illustrated BRCA here.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:55349.exp	12.807556	12.553149	10.195987
ENTREZID:23446.exp	12.978710	14.779668	13.728452
ENTREZID:208.exp	13.178509	13.646109	13.947363
ENTREZID:57706.exp	11.172428	11.965424	11.982281
ENTREZID:163486.exp	9.865733	9.787903	8.348728
ENTREZID:11021.exp	12.062383	11.673751	11.701740
ENTREZID:7531.exp	15.831085	15.287893	15.332841
ENTREZID:4012.exp	10.941048	11.048487	9.726218
ENTREZID:23216.exp	13.964251	12.288578	11.465566
ENTREZID:9882.exp	13.820079	12.345128	11.615170
ENTREZID:4644.exp	12.037890	11.988330	11.171802
ENTREZID:805.exp	15.502987	15.888315	16.125514
ENTREZID:5898.exp	12.644082	12.870557	12.892163

Table 14. ***BRCA Expression Data Frame Subset.*** As shown in the table, each column represents a patient and each row represents a gene and its associated omics type, in this case, "expression". The cells of this data frame contain the TPM (transcripts per million) gene expression values.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:55349.cnv	0	0	-1
ENTREZID:23446.cnv	0	0	-1
ENTREZID:208.cnv	0	0	1
ENTREZID:57706.cnv	0	0	0
ENTREZID:163486.cnv	0	1	1
ENTREZID:11021.cnv	0	-1	-1
ENTREZID:7531.cnv	0	-1	-1
ENTREZID:4012.cnv	0	0	0
ENTREZID:23216.cnv	0	0	0
ENTREZID:9882.cnv	0	0	0

Table 15. ***BRCA CNV Data Frame Subset.*** As demonstrated in the table, each column represents a patient and each row represents a gene and its associated omics type, in this case, copy number variation. Each cell of this data frame can have one of three values: -1, 0, or 1, indicating loss, no change, or gain of copy number, respectively.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:55349.met	0.46663081	0.56250890	0.55901840
ENTREZID:23446.met	0.27725302	0.28615744	0.27406130
ENTREZID:208.met	0.26714752	0.34731513	0.31742549
ENTREZID:57706.met	0.57091779	0.71273313	0.68038230
ENTREZID:163486.met	0.37561015	0.48104487	0.46497371
ENTREZID:11021.met	0.53695343	0.62332366	0.60958716
ENTREZID:7531.met	0.20846821	0.23201662	0.22023713
ENTREZID:4012.met	0.22942379	0.26698875	0.24696011
ENTREZID:23216.met	0.52062438	0.65219301	0.62551315
ENTREZID:9882.met	0.34934033	0.41670780	0.41313767
ENTREZID:4644.met	0.31872279	0.38875710	0.36276393

Table 16. ***BRCA Methylation Data Frame Subset.*** As shown in the table, each column represents a patient and each row represents a gene and its associated omics type, in this case, methylation. Each cell of this matrix displays the methylation level for a specific gene and patient.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:23446.mut	0	0	0
ENTREZID:208.mut	0	0	0
ENTREZID:57706.mut	0	0	0
ENTREZID:163486.mut	0	0	0
ENTREZID:11021.mut	0	0	0
ENTREZID:4012.mut	0	0	0
ENTREZID:23216.mut	0	0	0
ENTREZID:9882.mut	0	0	0
ENTREZID:4644.mut	0	0	0
ENTREZID:805.mut	0	0	0
ENTREZID:5898.mut	0	0	0
ENTREZID:57186.mut	0	0	0

Table 17. ***BRCA Mutation Data Frame Subset.*** As depicted, each column represents an individual patient, while each row represents a specific gene and the corresponding omic type, which in this case is mutation. The value of each cell within this data frame is determined by the presence or absence of a mutation in the patient; a value of one indicates the presence of a mutation, while a value of zero indicates its absence.

After identifying the genes that were significantly associated with patient survival across various cancers, it was determined that multi-omics information would be utilized to ascertain whether a particular gene played a role in an individual patient's survival outcome.
To be more precise, a Cox Proportional Hazards model was employed in conjunction with a penalized function in R to shrink the coefficients towards zero and determine which genes had non-zero coefficients for each patient. Additionally, a function was implemented in R for the expression and methylation data frames of five cancer types. The optimal lambda value for shrinking the coefficients using the penalized function was determined by comparing various lambda values and selecting the one that yielded the most favorable coefficients. Subsequently, with the optimal lambda value established, non-zero coefficient values were calculated for significant genes.
Afterward, the sign of each gene's coefficient was analyzed to determine its potential impact on patient survival. The median gene expression and methylation values were calculated for each gene across all patients. For gene expression, a positive coefficient sign indicated that over-expression of the gene was correlated with poorer survival outcomes, while a negative sign indicated the opposite. In cases where the coefficient was positive, genes with expression and methylation values exceeding the calculated median were assigned the absolute value of their coefficient, while those

with values at or below the median were assigned a value of zero. The inverse approach was applied to genes with negative coefficient values.
The implemented function takes as input the survival data frame for each patient, a matrix with gene names as row names and patient names as column names containing expression or methylation values for each cell, and the chosen lambda value. The output is a new matrix containing only genes whose coefficients were non-zero after shrinking.
The following table (Table 18) displays a subset of the final expression data frames for BRCA cancer generated by the aforementioned function. The data frames for methylation and other cancer types share the same structure, but only the expression data frame is shown here for brevity.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:163486.exp	0.0000000000	0.000000000	0.000000000
ENTREZID:9882.exp	0.0000000000	0.000000000	1.095690495
ENTREZID:6187.exp	0.0000000000	0.000000000	0.000000000
ENTREZID:6192.exp	0.0000000000	0.000000000	0.000000000
ENTREZID:1591.exp	0.0751634627	0.000000000	0.075163463
ENTREZID:240.exp	0.0000000000	0.000000000	0.000000000
ENTREZID:9915.exp	0.0000000000	0.000000000	0.000000000
ENTREZID:1586.exp	0.4238516456	0.000000000	0.000000000
ENTREZID:340665.exp	0.0000000000	0.000000000	0.036589268
ENTREZID:1594.exp	0.2109891690	0.210989169	0.210989169
ENTREZID:1553.exp	0.0000000000	0.000000000	0.161122599
ENTREZID:1558.exp	0.3856704320	0.385670432	0.000000000
ENTREZID:1571.exp	0.1656137514	0.165613751	0.000000000
ENTREZID:1573.exp	0.0000000000	0.521022603	0.000000000
ENTREZID:126410.exp	0.0000000000	0.005483104	0.005483104
ENTREZID:8529.exp	0.0000000000	0.000000000	0.000000000
ENTREZID:4051.exp	0.0537115394	0.000000000	0.000000000

Table 18. ***Subset of BRCA Expression Data Frame After Use of Penalized Function.*** As depicted in the illustration, the output of the aforementioned function would resemble this table for all five cancer types. Specifically, each cell displays the coefficient of the corresponding gene for a particular patient. The lambda value utilized for the penalized function for BRCA was 3.4.

Subsequently, it was necessary to implement an additional function to handle copy number variation (CNV) data. Unlike gene expression and methylation data frames, which contained measured values for gene expression and methylation respectively, CNV data frames did not contain measured values for the copy number of each gene. As previously mentioned, CNV data frames could only contain values of -1, 0, or 1 to represent loss, no change, or gain of copy number. As such, the function implemented for gene expression and methylation could not be applied to CNV data.

A new function was implemented in R that takes as input a CNV matrix and a survival data frame for patients with available CNV information, as well as a lambda value to determine the strength of coefficient shrinking using the penalized function of the penalized package. The function shrinks all coefficients towards zero and retains only those genes whose coefficients were non-zero after shrinking.

Two scenarios could arise: a negative coefficient sign for a gene or a positive coefficient sign. In the first scenario, for each patient, CNV values of -1 were converted to the absolute value of the corresponding gene coefficient, while 0 and 1 values were converted to 0. In the second scenario, CNV values of +1 were converted to the absolute value of the coefficient, while -1 and 0 values were converted to 0.

In addition to the aforementioned steps, it is important to note that prior to running the function designed to handle copy number variation (CNV) data, the procedure that was previously described for determining the optimal value for lambda in the gene expression and methylation function was also applied in this instance. As a result of this process, the output generated by the CNV function would take the form of another matrix, with gene names and the corresponding type of omic data included as row names, and individual patients represented by the column names (as shown in Table 19).

Afterwards, given that the mutation matrix contained only 0 and 1 values to represent the absence or presence of a mutation, respectively, it was determined that another function should be implemented in R to handle this final omic data type.

The implemented function operates similarly to the one designed for CNV data, but with a unique distinction due to the fact that mutation values cannot be negative. In cases where a coefficient had a negative sign, all values of 1 were converted to 0 since mutation values could not be negative. Conversely, in cases where a coefficient had a positive sign, values of 0 remained unchanged while values of 1 were converted to the absolute value of the coefficient. The final output of this function shared the same structure as the CNV data frame depicted in Table 19.

In summary, five data frames were generated for our five cancer types, representing whether an alteration in a gene's omic data could negatively impact a patient's survival outcome for a specific cancer. Specifically, a value of 0 for a gene indicated that it did not adversely affect patient survival, while a value greater than zero indicated that it could worsen the patient's survival outcome. It was determined that the coefficients of a gene from its Cox Proportional Hazards model would be used to indicate its negative effect on patient survival. This approach enabled comparison of the negative impact of each gene on patient survival across different cancers. Since a single lambda value was used for each cancer type to generate these coefficients, it can be inferred that the larger the coefficient, the greater the gene's impact. These data frames were subsequently used to create heatmaps for each cancer.

	TCGA.A1.A0SB	TCGA.A1.A0SE	TCGA.A1.A0SF
ENTREZID:7531.cnv	0	0.09090046	0.09090046
ENTREZID:57186.cnv	0	0.00000000	0.29632026
ENTREZID:57148.cnv	0	0.00000000	0.00000000
ENTREZID:8661.cnv	0	0.00000000	0.00000000
ENTREZID:1978.cnv	0	0.31061318	0.31061318
ENTREZID:1981.cnv	0	0.00000000	0.00000000
ENTREZID:196.cnv	0	0.00000000	0.04350394
ENTREZID:240.cnv	0	0.00000000	0.16433402
ENTREZID:1557.cnv	0	0.00000000	0.00000000
ENTREZID:1582.cnv	0	0.00000000	0.09155797
ENTREZID:8564.cnv	0	0.00000000	0.00000000
ENTREZID:2181.cnv	0	0.00000000	0.00000000
ENTREZID:51074.cnv	0	0.00000000	0.00000000
ENTREZID:3613.cnv	0	0.00000000	0.00000000
ENTREZID:66036.cnv	0	0.00000000	0.00000000
ENTREZID:8396.cnv	0	0.00000000	0.00000000

Table 19. ***Subset of BRCA CNV Data Frame After Use of Penalized Function.*** As depicted in the illustration, each row represents a single gene and its corresponding omic type, which in this case is copy number variation. Each column represents an individual patient. The values within each cell of this data frame represent the coefficient of each gene for a particular patient. The lambda value utilized to generate this matrix was 3.4.

3. RESULTS and DISCUSSION

3.1 Preliminary Analysis of MultiAssayExperiment Data

The clinical dataset for breast carcinoma (BRCA) encompasses information on 1,098 patients diagnosed between 1988 and 2013, ranging in age from 26 to 90 years old. Of these patients, survival data (days to death) was available for only 151 individuals. The data indicates that the lifespan of patients with BRCA varied between about 3.86 and 248.46 months. Notably, BRCA is the only cancer type in this study that includes both male and female patients. However, of the total patient population, only 13 were male, while the remainder were either female or their gender was not specified. Consequently, male patients were excluded from further analysis to ensure a more homogeneous sample.

The clinical dataset for cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) encompasses information on 307 patients diagnosed between 1994 and 2013, ranging in age from 20 to 88 years old. Of these patients, survival data (days to death) was available for 72 individuals. The data indicates that the lifespan of patients with CESC varied between 0.46 and 136.2 months.

The clinical dataset for ovarian serous cystadenocarcinoma (OV) encompasses information on 588 patients diagnosed between 1992 and 2013, ranging in age from 26 to 89 years old. Of these patients, survival data (days to death) was available for 343 individuals. The data indicates that the lifespan of patients with OV varied between 0.26 and 154.13 months.

The clinical dataset for uterine corpus endometrial carcinoma (UCEC) encompasses information on 547 patients diagnosed between 1995 and 2013. However, age data was not available for this cancer type. Of these patients, survival data (days to death) was available for 91 individuals. The data indicates that the lifespan of patients with UCEC varied between 1.66 and 114.1 months.

The clinical dataset for uterine carcinosarcoma (UCS) encompasses information on 57 patients diagnosed between 2002 and 2012, ranging in age from 51 to 90 years old. Of these patients, survival data (days to death) was available for 35 individuals. The data indicates that the lifespan of patients with UCS varied between 0.26 and 103.83 months.

Table 20 presents a comprehensive overview of the clinical data for the five cancer types under investigation in this study. This table provides a detailed summary of the key information and statistics for each cancer type, including the number of patients, their age range, diagnosis dates, and survival data. By examining this table, readers can gain a better understanding of the clinical characteristics of these cancer types and the patient populations affected by them.

	Patients	*Diagnosing_period*	*Age*	*Patients_with_survival_data*	*Days_of_survival*
BRAC	*1098*	*1988-2013*	*26-90*	*151*	*116-7454*
CESC	*307*	*1994-2013*	*20-88*	*72*	*14-4086*
OV	*588*	*1992-2013*	*26-89*	*343*	*8-4624*
UCEC	*547*	*1995-2013*	*NA*	*91*	*50-3423*
UCS	*57*	*2002-2013*	*51-90*	*35*	*8-3115*

Table 20. ***Overview of Clinical Data for Five Cancer Types.***

As shown in the previous table, the highest number of patients is observed in the BRCA dataset, while the lowest number is found in the UCS dataset. However, when considering only patients for whom survival data is available, the OV dataset contains the highest number of patients among these five cancer types.

3.2 Overview of Multi-Omics Data for Each Cancer Type

As outlined in the materials and methods section, our analysis included only patients for whom complete multi-omics data was available. After filtering the patient population based on this criterion, the number of patients for BRCA, CESC, OV, UCEC, and UCS decreased to 458, 190, 183, 108, and 56, respectively. Despite this reduction in patient numbers, BRCA remained the largest group, while CESC became the second largest group, surpassing UCEC. Furthermore, OV became the second smallest group after UCS, which remained the smallest group both before and after filtering (Table 21).

Table 21 provides a detailed comparative analysis of the number of interpreted genes among five different cancer types. This table offers valuable insights into the distribution of gene expression, copy number variation (CNV), methylation, and mutation data across these cancer types and allows for a more nuanced understanding of their molecular characteristics.

As shown in the table, the highest number of genes in the expression matrix is observed in the BRCA group, followed by OV. This indicates that these two cancer types exhibit a greater degree of complexity and diversity in their gene expression profiles. In contrast, for CNV data, the number of interpreted genes is consistent across all five cancer types. This suggests that the CNV data frame contains precisely the same genes for each cancer type, providing a common basis for comparison and analysis. Additionally, the methylation data frame contains the same genes for all cancer types except OV. In contrast to the other data frames, the mutation data frame exhibits significant variation in the number of genes among different cancer types. UCEC has the highest number of genes, while UCS has the lowest.

	Patients	*Expression*	*CNV*	*Methylation*	*Mutation*
BRAC	*458*	*16088*	*21952*	*17874*	*15377*
CESC	*190*	*15253*	*21952*	*17874*	*12759*
OV	*183*	*15890*	*21952*	*13044*	*7737*
UCEC	*108*	*15005*	*21952*	*17874*	*16467*
UCS	*56*	*14892*	*21952*	*17874*	*5445*

Table 21. ***Comparison of Genes and Patients Across Different Cancer Types.***

The table above presents the final statistics for the number of patients and genes included in the expression, copy number variation (CNV), methylation, and mutation data frames for the five cancer types under investigation. These statistics reflect the results of all previous data processing and filtering steps and were used as the basis for further analysis.

3.3 Pathway Analyses

After conducting pathway and module survival tests using MOSClip, we sought to identify pathways significantly associated with patient survival that were shared among different types of cancer. Since the pathways significantly associated with the survival of patients had been determined, it was not challenging to perform more analysis regarding these pathways (A comprehensive list of all significant pathways, modules, and all the other data that has been produced in this research can be accessed via the following GitHub link: https://github.com/Amin-Zlf/Integrated-Multi-omics-Survival-Analysis-of-Gynecologic-and-Breast-Cancers). To this end, we created a Venn diagram. It illustrates the number of unique and common pathways among different types of cancer (Figure 8).
Upon careful examination of our data, it became apparent that there were no pathways common to all five types of cancer. In light of this finding, we made the decision to shift our focus to those pathways that were shared among the greatest number of cancers. As can be seen in Figure 8, our analysis revealed the presence of several pathways that were shared among three cancers at most. Given the significance of these shared pathways, we determined that it would be prudent to select them for further analysis in order to gain a deeper understanding of their role in cancer survival.

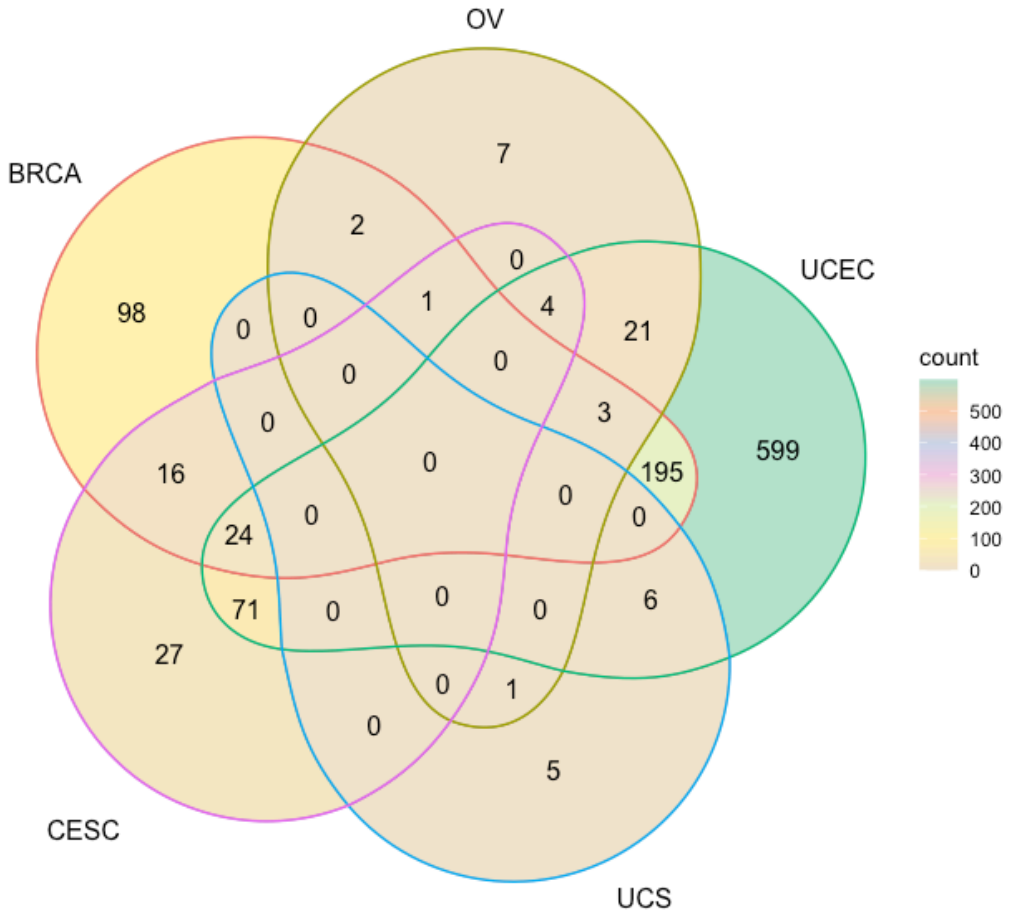

Figure 8. ***Shared Pathways Among Five Cancer Types.*** As our analysis shows, there are no pathways shared among all five types of cancer. However, we did identify several pathways shared among specific combinations of cancers. For instance, 4 pathways were shared among OV-UCEC-CESC, 3 pathways were shared among BRCA-UCEC-OV, 1 pathway was shared among CESC-OV-BRCA, and 24 pathways were shared among BRCA-CESC-UCEC.

As depicted in Figure 8, we elected to concentrate our analysis on those sections of the Venn diagram containing significant pathways shared among three types of cancer. The first section corresponded to the OV, CESC, and UCEC cancers and contained four pathways: “Signaling by NOTCH2,” “GRB2:SOS provides linkage to MAPK signaling for Integrins,” “Integrin signaling,” and “Platelet Aggregation (Plug Formation).” The second section selected corresponded to the BRCA, UCEC, and OV cancers and included the “TNFs bind their physiological receptors,” “Purine salvage,” and “TNFR2 non-canonical NF-kB pathway” pathways. The third section chosen was shared among the CESC, OV, and BRCA cancers and contained only one pathway: “TP53 Regulates Metabolic Genes.” Finally, the last section selected was shared among the BRCA, CESC, and UCEC cancers and contained 24 pathways. The names of these 24 pathways can be found in Figure 9.

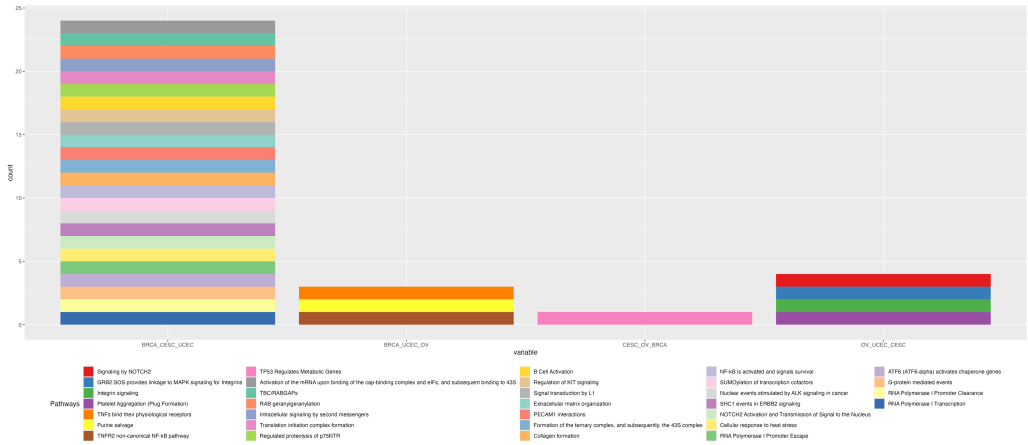

Figure 9. ***Comparative Analysis of Pathways Shared Among Different Cancer Types.*** As our analysis shows, among the selected group of cancers, the BRCA-CESC-UCEC combination has the greatest number of shared pathways, with a total of 24. The OV-UCEC-CESC combination has the second highest number of shared pathways, with 4. The BRCA-UCEC-OV and CESC-OV-BRCA combinations follow, with 3 and 1 shared pathways, respectively.

Thus, Figure 9 displays the names of the shared pathways for each group of cancers. It is important to note that these pathways were identified as being significantly associated with patient survival based on our previous statistical analyses. This suggests that alterations in these pathways can have a significant impact on patient outcomes.

In order to gain a more comprehensive understanding of the distribution of pathways among the five types of cancer, we conducted an analysis of the percentage of pathways for each cancer type, whether unique or shared. Our findings revealed that UCEC had the highest proportion of unique pathways among the five cancer types, at 55.5%. This was followed by BRCA, CESC, OV, and UCS, with 9.1%, 2.5%, 0.6%, and 0.5%, respectively. In addition to these unique pathways, our analysis also showed that 18.1% of pathways significantly associated with patient survival were shared between BRCA and UCEC. These results provide valuable insights into the distribution of significant pathways among different cancer types.

Upon identifying the shared pathways among different types of cancer, we turned our attention to the question of whether the p-values of the same omics were significant for all cancers within a given group. In other words, we sought to determine whether a given omic was consistently associated with patient survival across all cancer types that included a specific pathway. For example, we investigated whether gene expression in a single pathway was significantly associated with patient survival in all cancer types that included that pathway. To facilitate our analysis, we created Figure 10,

which provides a visual representation of our findings. As can be seen in this figure, gene expression emerged as the only omic common to the shared pathways of different cancer types. This finding has important implications for our understanding of the role of gene expression in cancer survival.

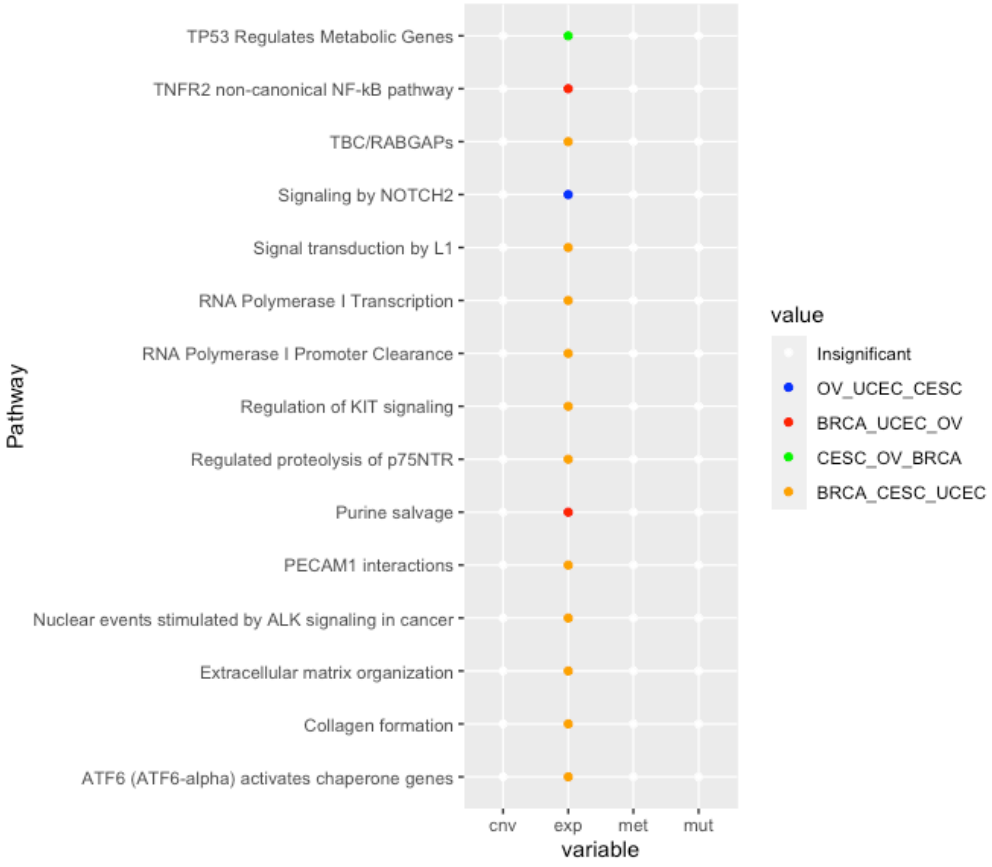

Figure 10. ***A Comprehensive Comparative Analysis of the Significance of Different Omics Across Multiple Cancer Types.*** In the course of our analysis, we discovered that gene expression was the only omic that exhibited a significant correlation with patient survival across all cancer types within each group. In order to provide a more detailed and comprehensive representation of our findings, we organized our data into a table format. In this table, each row corresponds to a specific pathway, while each column represents a different type of omic, including Copy number variation (cnv), gene expression (exp), methylation (met), and mutation (mut). The color of each cell within the table conveys two important pieces of information. First and foremost, it indicates whether the omic for the corresponding pathway is significantly associated with survival in all cancer types within the group. If this is the case, the cell is colored; if not, it remains white. Second, if the cell is colored, its hue provides an indication of which group it is associated with. This color-coding system allows for easy visualization and interpretation of our results.

3.4 Network Analysis of Pathways Using Cytoscape

Following the pathway analyses, it was determined that further investigations into the genes within the significant modules were necessary. As such, Cytoscape was utilized to visualize the topology of the genes within each module that exhibited a significant association with patient survival. Subsequently, we will commence with an interpretation of the graphs for each cancer type individually before proceeding to an interpretation of all cancers collectively.

The first cancer visualized using Cytoscape was BRCA (Figure 11). To enhance the quality of the visualization, it was determined that only genes with 1 to 100 edges would be selected and filtered. As indicated in the legend of Figure 11, a single guidance legend was created for all types of Cytoscape graphs presented in this research. Each cancer type was represented by a specific color, while different omics were denoted by distinct node shapes. Consequently, there were five distinct cancer types and 11 unique cancer groups: BRCA, BRCA-UCEC, BRCA-OV, UCS, UCEC-UCS, OV, BRCA-UCS, UCEC, CESC, CESC-UCEC, and BRCA-CESC. Additionally, there were 11 unique omic types: cnv-exp-met, cnv-exp-mut, cnv-exp-met-mut, exp-met, exp-mut, cnv-met, cnv-exp, cnv, exp, met and exp-met-mut. To illustrate further, a light-green gene in the shape of a small circle signifies that an alteration in the expression of that specific gene is significantly associated with the survival of BRCA patients.

In addition to the aforementioned details, it is important to note that other pertinent information was also available. This included, for example, the name of the specific pathway in which each gene performs its designated activity. As a result of this availability of information, it was possible to further classify the genes based on several distinct criteria. These criteria included the type of cancer with which the gene was associated, the specific pathway in which the gene played a role, and the type of significant omic that was relevant to the gene in question.

In the subsequent pages, the graphs for each individual cancer type, as well as a composite graph representing all cancers collectively, are presented. As illustrated in Figure 11, the first graph corresponds to BRCA. As can be observed, there were eight pathways that exhibited a significant association with the survival of BRCA patients. These pathways included GPCR downstream signaling, Axon guidance, Amino acid and derivative metabolism, Signaling by FLT3 ITD and TKD mutants, Signaling by GPCR, Vesicle-mediated transport, Extracellular matrix organization, and Metabolism. Additionally, this graph provides information regarding which specific genes within these pathways were significantly associated with survival and which type or types of omic alteration could be responsible for the survival of BRCA patients.

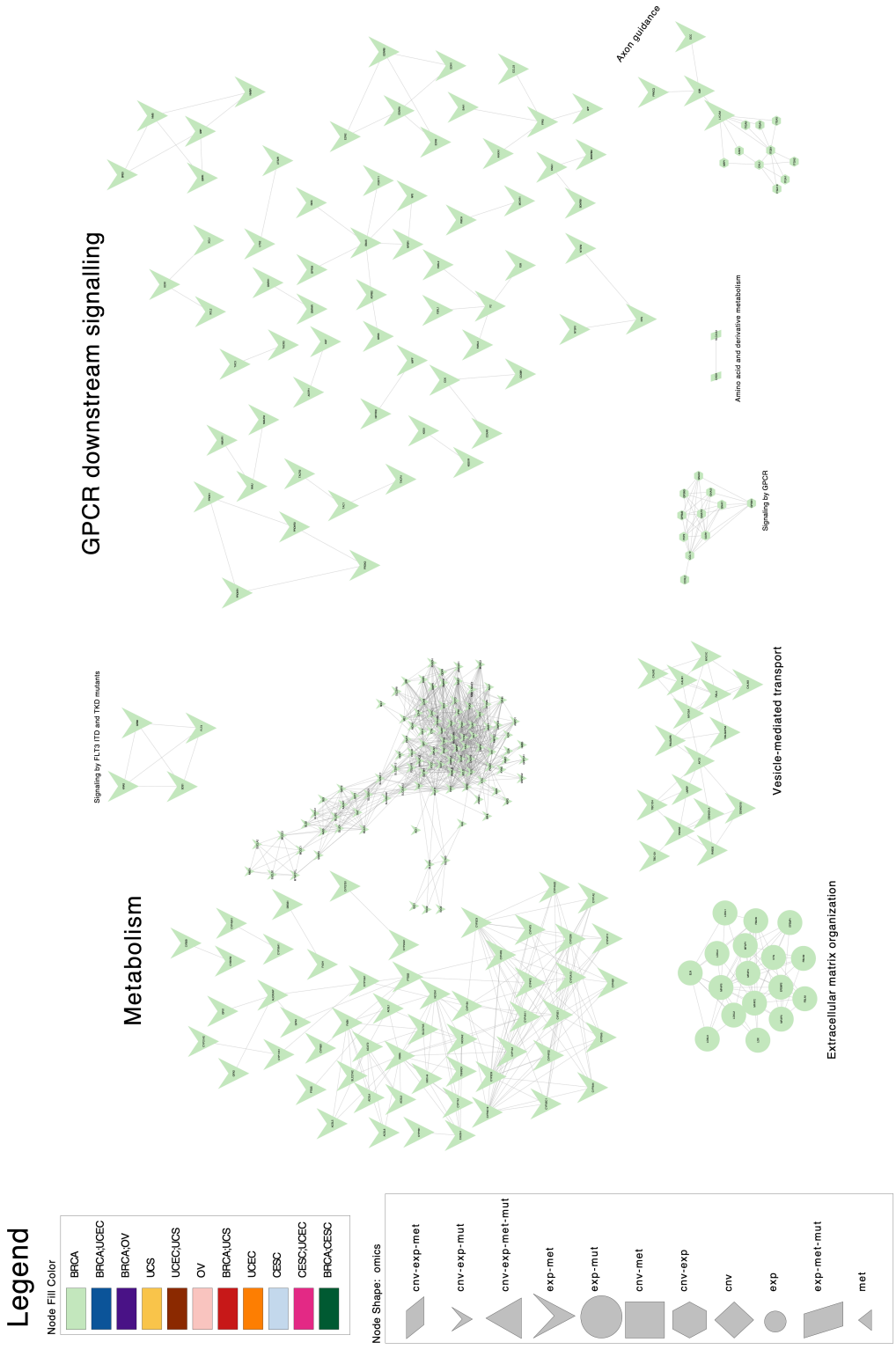

Figure 11. ***Cytoscape Graph Depicting Significant Genes in BRCA.*** This graph depicts the genes, pathways, and types of omics that exhibited a significant association with the survival of BRCA patients.

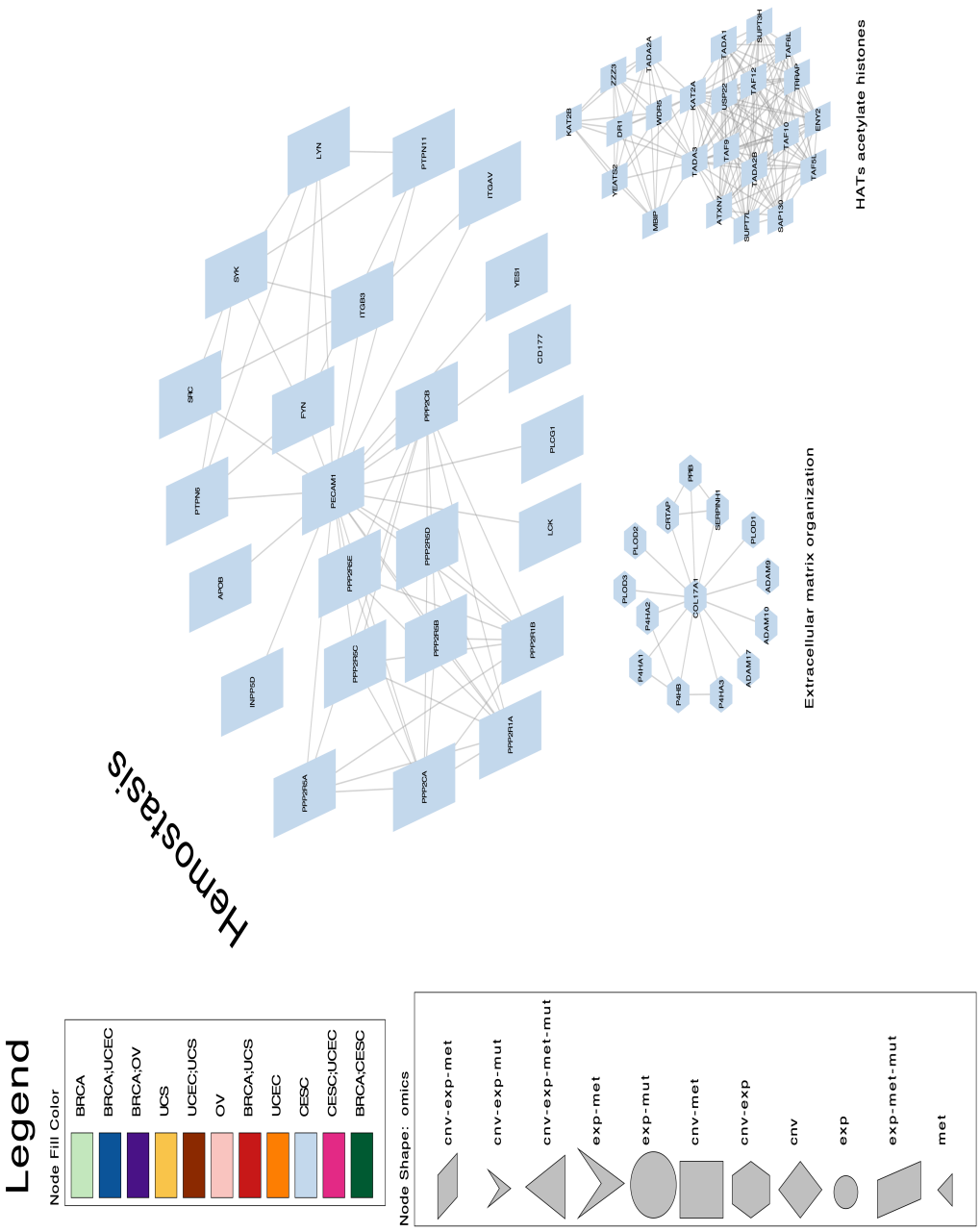

Figure 12. ***Cytoscape Graph Depicting Significant Genes in CESC.*** This graph provides a detailed and comprehensive visual representation of the various genes, pathways, and types of omics that have been demonstrated to exhibit a significant association with the survival of CESC patients. Through this graphical representation, it is possible to gain a deeper understanding of the complex relationships between these various factors and their impact on patient survival.

As illustrated in Figure 12, there were three distinct pathways that exhibited a significant association with the survival of CESC patients. These pathways were identified through careful analysis and examination of the available data and were found to play a crucial role in the survival of CESC patients. These pathways were identified as HATs acetylate histones, Extracellular matrix organization, and Hemostasis. In addition to this information, the graph also provides a visual representation of which specific genes within these pathways were significantly associated with survival and which type or types of omic alteration could potentially be responsible for the survival of CESC patients.

With regard to OV cancer, Figure 13 illustrates that there were four distinct pathways that exhibited a significant association with the survival of OV patients. These pathways were identified as TP53 Regulates Metabolic Genes, NGF-stimulated transcription, Activation of SMO, and Hh mutants abrogate ligand secretion. In addition to this information, the graph also provides a visual representation of which specific genes within these pathways were significantly associated with survival and which type or types of omic alteration could potentially be responsible for the survival of OV patients.

As depicted in Figure 14, there were 11 distinct pathways that exhibited a significant association with the survival of UCEC patients. These pathways were identified as Termination of O-glycan biosynthesis, Membrane Trafficking, Cytokine Signaling in Immune system, Signaling Pathways, mRNA Capping, RNA Polymerase III Transcription Initiation, RHO GTPases Activate Formins, Interferon alpha/beta signaling, Gluconeogenesis, Developmental Biology, and Cellular responses to stress. In addition to this information, the graph also provides a visual representation of which specific genes within these pathways were significantly associated with survival and which type or types of omic alteration could potentially be responsible for the survival of UCEC patients.

With regard to UCS cancer, it was determined that there were eight distinct pathways that exhibited a significant association with the survival of UCS patients. These pathways were identified as Synthesis of bile acids and bile salts via 27-hydroxycholesterol, Caspase-mediated cleavage of cytoskeletal proteins, Apoptotic cleavage of cellular proteins, GPCR ligand binding, Vitamin D (calciferol) metabolism, CaMK IV-mediated phosphorylation of CREB, RIP-mediated NFkB activation via ZBP1, and TICAM1,TRAF6-dependent induction of TAK1 complex. Similar to the previous graphs, this graph (Figure 15) also provides a visual representation of which specific genes within these pathways were significantly associated with survival and which type or types of omic alteration could potentially be responsible for the survival of UCS patients.

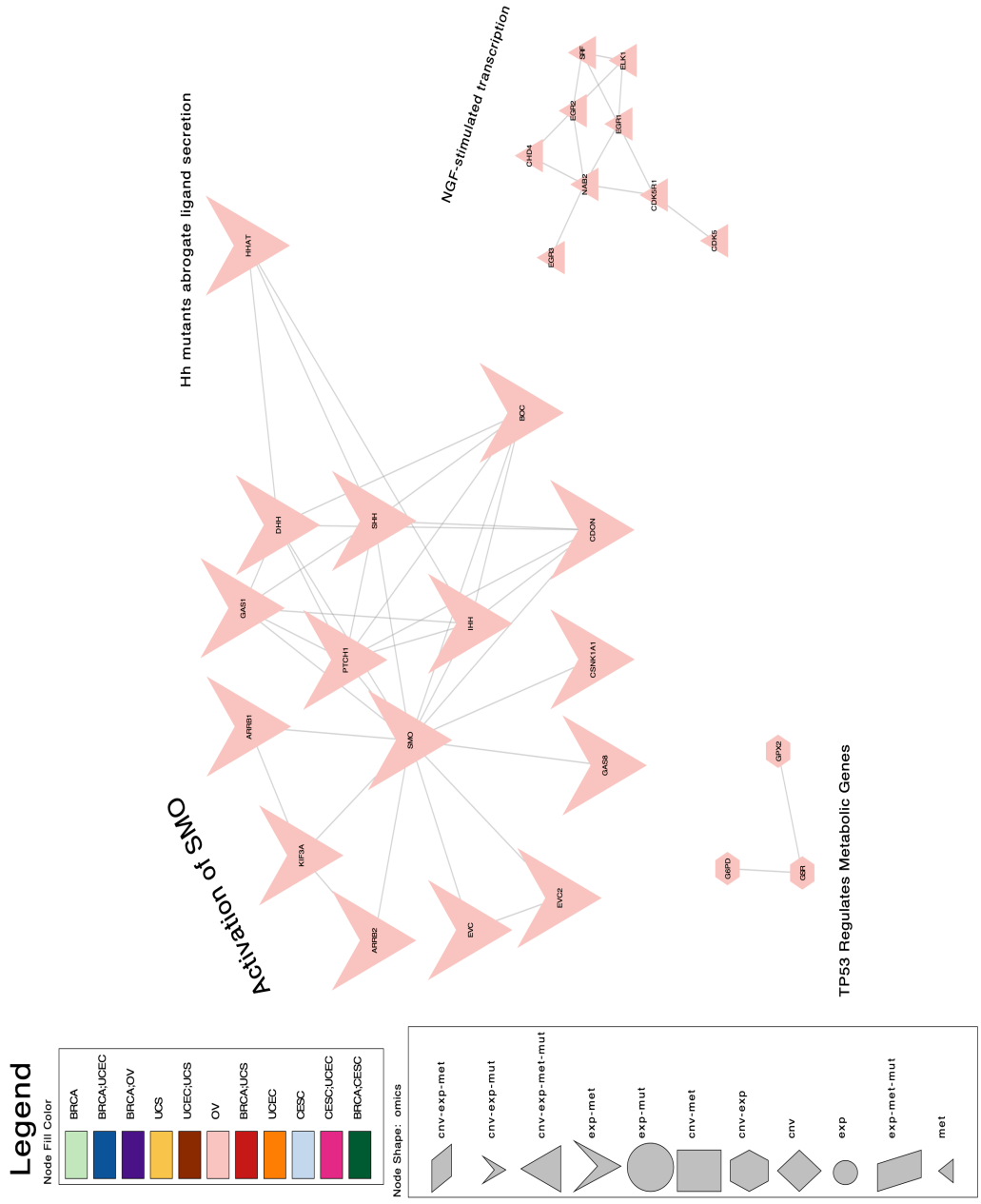

Figure 13. ***Cytoscape Graph Depicting Significant Genes in OV.*** This graph provides a detailed and comprehensive visual representation of the various genes, pathways, and types of omics that have been demonstrated to exhibit a significant association with the survival of OV patients. Through this graphical representation, it is possible to gain a deeper understanding of the complex relationships between these various factors and their impact on patient survival.

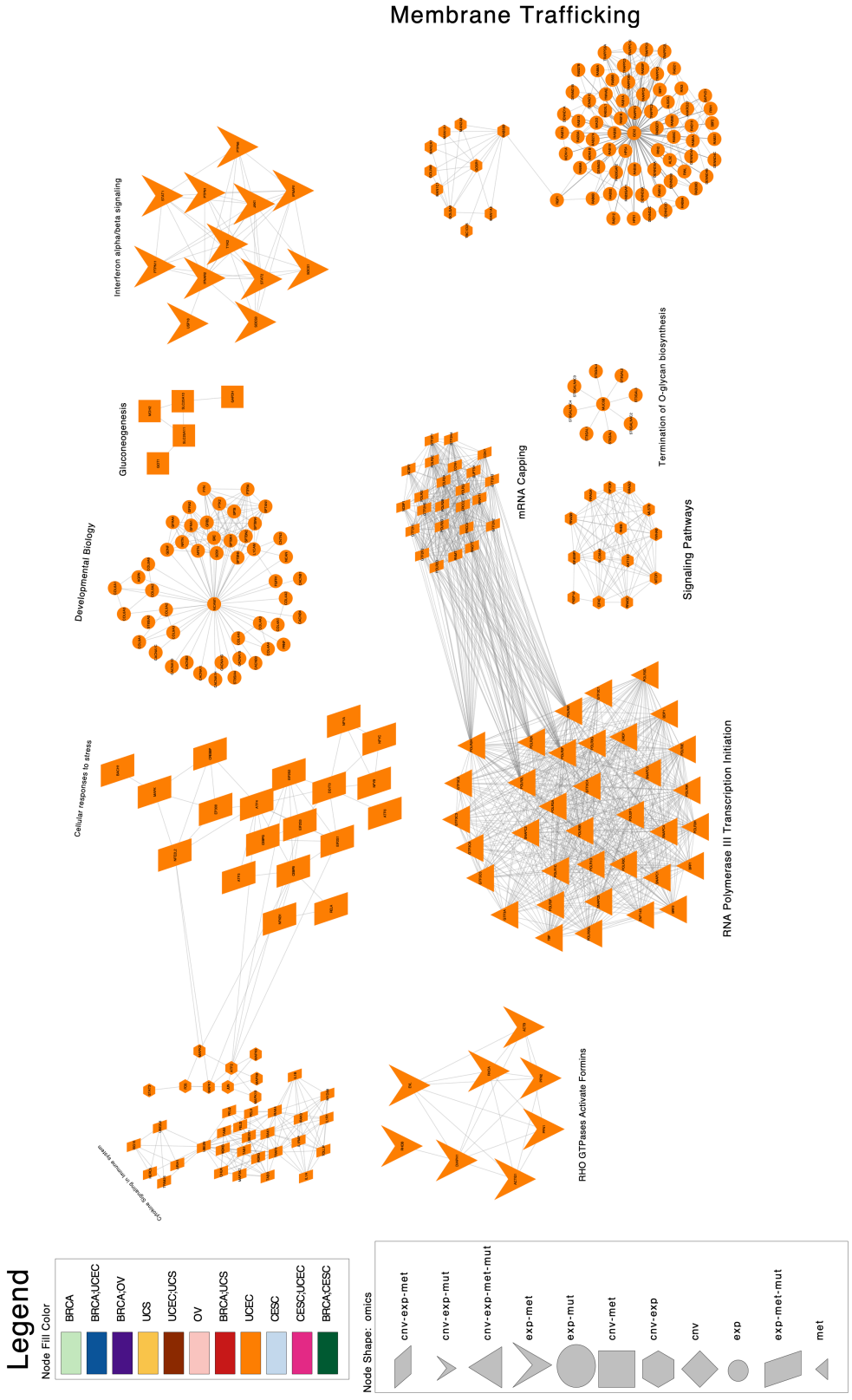

Figure 14. ***Cytoscape Graph Depicting Significant Genes in UCEC.*** This graph shows genes, pathways, and omics associated with UCEC patient survival and their complex interrelationships.

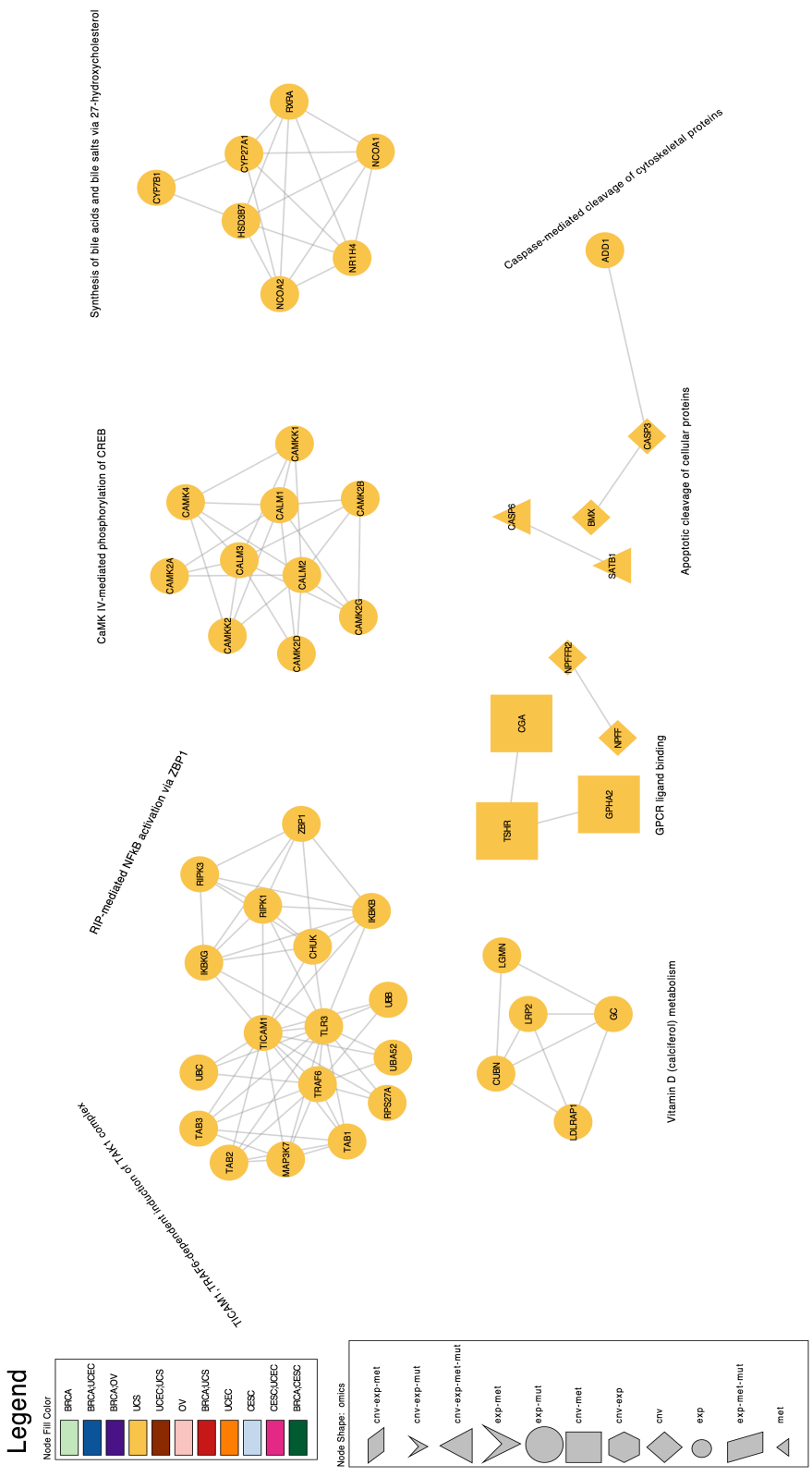

Figure 15. ***Cytoscape Graph Depicting Significant Genes in UCS.*** This graph provides a detailed and comprehensive visual representation of the various genes, pathways, and omics that have been demonstrated to have a significant association with the survival of UCEC patients.

For the final graph, it was determined that consolidating all previous information into a single visual representation would facilitate a more effective comparison of genes and pathways across different types of cancers. It is important to note that due to the vast number of multi-omics combinations, it was decided to display genes only based on cancer type and the specific pathway in which each gene plays a role, in order to maintain clarity and comprehensibility.

As depicted in Figure 16, there exist notable biological similarities among the five types of cancers under investigation. For instance, the L1CAM gene was found to have a significant association with the survival of both BRCA and UCEC patients. Specifically, alterations in both the expression and methylation of L1CAM within the "Axon guidance" pathway were significantly associated with BRCA patient survival, while only the expression of L1CAM within the "Developmental Biology" pathway was significantly associated with UCEC patient survival.

Additionally, the expression of the GRB2 gene within the "Developmental Biology" pathway was significantly associated with UCEC patient survival. This gene, like L1CAM, was common to both BRCA and UCEC. However, the expression and methylation of GRB2 within the "Signaling by FLT3 ITD and TKD mutants" pathway were significantly associated with BRCA patient survival.

Another gene, GPX2, which plays a role within the "Metabolism" pathway, had its expression and methylation significantly associated with BRCA patient survival. Moreover, alterations in the copy number variation and expression of the GPX2 gene within the "TP53 Regulates Metabolic Genes" pathway were found to be significantly associated with the survival of OV patients.

Other genes that exhibited a significant association with the survival of UCEC and UCS patients include MAP3K7 and CHUK. Specifically, alterations in the copy number variation, expression, and methylation of MAP3K7 and CHUK within the "Cytokine Signaling in Immune system" pathway were significantly associated with UCEC patient survival.

Additionally, the expression of MAP3K7 within the "TICAM1, TRAF6-dependent induction of TAK1 complex" pathway and CHUK within the "RIP-mediated NFkB activation via ZBP1" pathway were significantly associated with UCS patient survival. These are just a few examples of the insights that can be gleaned from the graphs generated by Cytoscape in this research. To avoid repetition and facilitate a better understanding of genes shared among different types of cancers, Table 22 was created.

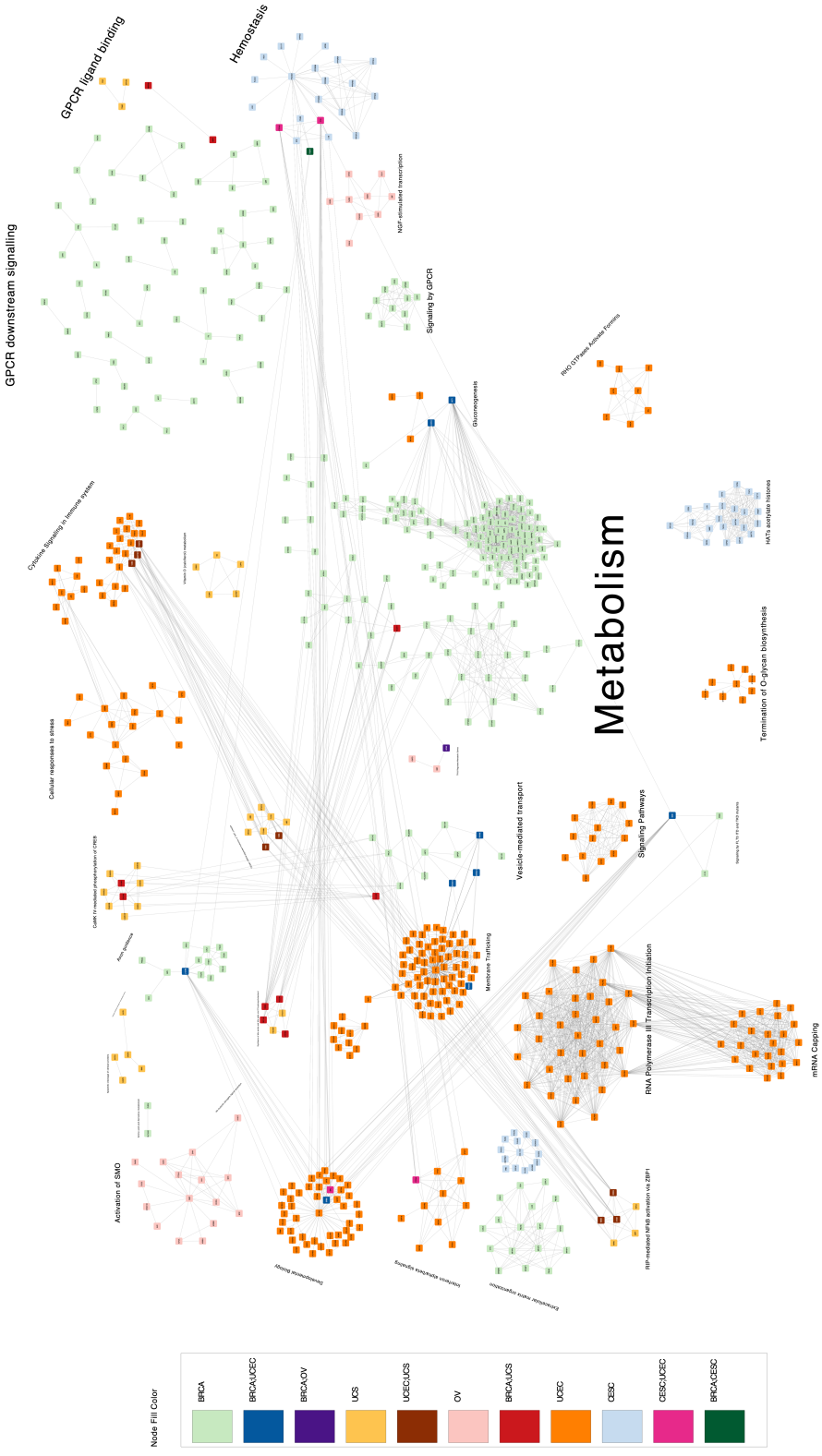

Figure 16. ***Visualization of Significant Genes and Pathways Across All Cancer Types.*** This graph illustrates the genes and pathways that have a significant association with the survival of BRCA, CESC, OV, UCEC, and UCS patients.

	Gene_Name	Cancer_1	Pathway_1	Omic_1	Cancer_2	Pathway_2	Omic_2
1	L1CAM	BRCA	Axon guidance	exp-met	UCEC	Developmental Biology	exp
2	SLC25A11	BRCA	Metabolism	cnv-exp-mut	UCEC	Gluconeogenesis	cnv-met
3	DENND1B	BRCA	Vesicle-mediated transport	exp-met	UCEC	Membrane Trafficking	exp
4	DENND1A	BRCA	Vesicle-mediated transport	exp-met	UCEC	Membrane Trafficking	exp
5	GRB2	BRCA	Signaling by FLT3 ITD and TKD mutants	exp-met	UCEC	Developmental Biology	exp
6	GOT1	BRCA	Metabolism	cnv-exp-mut	UCEC	Gluconeogenesis	cnv-met
7	SOS1	BRCA	Signaling by FLT3 ITD and TKD mutants	exp-met	UCEC	Developmental Biology	exp
8	RAB35	BRCA	Vesicle-mediated transport	exp-met	UCEC	Membrane Trafficking	exp
9	YWHAE	BRCA	Vesicle-mediated transport	exp-met	UCEC	Membrane Trafficking	exp
10	GPX2	BRCA	Metabolism	exp-met	OV	TP53 Regulates Metabolic Genes	cnv-exp
11	TRAF6	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	TICAM1,TRAF6-dependent induction of TAK1 complex	exp
12	MAP3K7	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	TICAM1,TRAF6-dependent induction of TAK1 complex	exp
13	IKBKB	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	RIP-mediated NFkB activation via ZBP1	exp
14	TAB1	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	TICAM1,TRAF6-dependent induction of TAK1 complex	exp
15	TAB2	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	TICAM1,TRAF6-dependent induction of TAK1 complex	exp
16	IKBKG	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	RIP-mediated NFkB activation via ZBP1	exp
17	TAB3	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	TICAM1,TRAF6-dependent induction of TAK1 complex	exp
18	CHUK	UCEC	Cytokine Signaling in Immune system	cnv-exp-met	UCS	RIP-mediated NFkB activation via ZBP1	exp
19	RXRA	BRCA	Metabolism	exp-met	UCS	Synthesis of bile acids and bile salts via 27-hydroxycholesterol	exp
20	NR1H4	BRCA	Metabolism	exp-met	UCS	Synthesis of bile acids and bile salts via 27-hydroxycholesterol	exp
21	CALM3	BRCA	Vesicle-mediated transport	exp-met	UCS	CaMK IV-mediated phosphorylation of CREB	exp
22	CALM2	BRCA	Vesicle-mediated transport	exp-met	UCS	CaMK IV-mediated phosphorylation of CREB	exp
23	NPFFR2	BRCA	GPCR downstream signalling	exp-met	UCS	GPCR ligand binding	cnv
24	NCOA2	BRCA	Metabolism	exp-met	UCS	Synthesis of bile acids and bile salts via 27-hydroxycholesterol	exp
25	CYP7B1	BRCA	Metabolism	exp-met	UCS	Synthesis of bile acids and bile salts via 27-hydroxycholesterol	exp
26	CALM1	BRCA	Vesicle-mediated transport	exp-met	UCS	CaMK IV-mediated phosphorylation of CREB	exp
27	NCOA1	BRCA	Metabolism	exp-met	UCS	Synthesis of bile acids and bile salts via 27-hydroxycholesterol	exp
28	NPFF	BRCA	GPCR downstream signalling	exp-met	UCS	GPCR ligand binding	cnv
29	SRC	CESC	Hemostasis	exp-met-mut	UCEC	Developmental Biology	exp
30	PTPN6	CESC	Hemostasis	exp-met-mut	UCEC	Interferon alpha/beta signaling	exp-met
31	PTPN11	CESC	Hemostasis	exp-met-mut	UCEC	Interferon alpha/beta signaling	exp-met
32	FYN	CESC	Hemostasis	exp-met-mut	UCEC	Developmental Biology	exp
33	ITGAV	BRCA	Axon guidance	cnv-exp	CESC	Hemostasis	exp-met-mut

Table 22. ***Common Genes Across Multiple Cancer Types.*** This table provides a comprehensive and detailed visual representation of the genes that are shared among different types of cancers. It includes information on the specific cancers with which each gene is associated, the types of omics for which alterations have been shown to have a significant association with patient survival, and the pathways in which each gene plays a role.

3.5 Comparative Analysis of Common Cancer Genes: A Literature Review

After identifying the common genes among various types of cancers, a comprehensive literature review was conducted to compare our findings with those of other studies. In the following section, we will focus on the genes that were discovered through our analysis.

The first gene identified in our analysis was L1CAM, which has been shown to have a significant association between its expression and methylation alterations and the survival of BRCA patients. Moreover, the expression of this gene was found to be associated with the survival of UCEC patients. Previous studies have demonstrated that L1CAM expression is correlated with the aggressiveness and size of BRCA tumors (Moisini et al., 2021). As such, it has been linked to larger tumor size and earlier recurrence in BRCA cancer. Additionally, research has indicated that this gene may have a significant impact on UCEC from an immunological perspective (Fang et al., 2022).

The subsequent gene identified in our analysis was SLC7A11, which demonstrated a significant association with the survival of patients diagnosed with BRCA and UCEC cancers. Research has revealed that an increased expression of SLC7A11 correlates with improved prognosis for individuals diagnosed with UCEC (Fang et al., 2023). Furthermore, research has revealed that this gene is instrumental in regulating cell death induced by glucose starvation in various cancers, including breast cancer (Koppula et al., 2018).

In our analysis, GRB2 and SOS1 were identified as the subsequent genes, exhibiting a significant association with the survival of BRCA and UCEC patients. Numerous studies have shown that inhibiting the interaction between GRB2 and SOS1 can have a significant impact on the treatment of breast cancer (Yu et al., 2017).

The subsequent gene identified in our analysis was GOT1, which demonstrated a significant association with the survival of patients diagnosed with BRCA and UCEC cancers. Additional research has corroborated the link between this gene and the distinct metabolism of BRCA tumor cells (An et al., 2022).

In this study, the genes SOS1 and YWHAE were found to be significantly associated with the survival of patients with BRCA and UCEC. This finding is supported by previous research, which has demonstrated that the SOS1 gene plays a role in the migration of breast cancer cells (Zhu et al., 2013). The YWHAE gene has been identified as a key factor in the growth and spread of breast cancer cells, as well as their ability to resist the effects of chemotherapy (Yang et al., 2019). This gene plays a crucial role in promoting cell proliferation, facilitating metastasis, and increasing resistance to chemotherapy in breast cancer cells.

Another gene identified in this study as being associated with the survival of patients with breast and ovarian cancer is GPX2. Research has shown that the expression of this gene can serve as a prognostic marker in breast cancer patients, providing valuable information for predicting disease outcomes and guiding treatment decisions (Esworthy et al., 2022).
In this study, the CHUK gene was also identified as being significantly associated with the survival of patients with UCEC and UCS. Recent research has shown that this gene acts as a tumor suppressor in various types of tumors, suggesting that it may play a crucial role in preventing the development and progression of cancer (Li & Hu, 2021).
In our analysis, RXRA, CALM2, NCOA2, and NCOA1 genes were found to be significantly associated with the survival of patients with BRCA and UCS. Research has demonstrated that RXRA plays a crucial role in the prognosis of breast cancer (Pande et al., 2013). Furthermore, research has shown that the CALM2 gene is significantly associated with breast cancer tumors (Haddad et al., 2015). The NCOA2 gene has been also identified as playing a key role in the development of tumors (Cai et al., 2019). By inducing cell cycle arrest and apoptosis, this gene is able to suppress cell proliferation and inhibit the growth and spread of breast cancer cells. The latest gene, NCOA1 has been demonstrated to be involved in breast cancer metastasis, resistance to endocrine therapy, and recurrence (Qin et al., 2014).
Furthermore, in our analysis, the PTPN6 and SRC genes were found to be significantly associated with the survival of patients with CESC and UCEC. Previous research has shown that this gene is associated with the metastasis of cervical cancer (Eswaran et al., 2022). Moreover, overexpression of PTPN6 has been associated with reduced overall survival in patients with cervical cancer, suggesting that this gene may play a crucial role in the development and progression of the disease. Additionally, research has demonstrated that SRC proteins play crucial roles in cell migration, adhesion, and proliferation in patients with UCEC (Akin & Özkan, 2023).
Finally, the ITGAV gene was identified as being significantly associated with the survival of patients with BRCA and CESC. Studies have indicated that the ITGAV gene has the potential to be a therapeutic target in the treatment of breast cancer (Cheuk et al., 2020). This presents a promising opportunity for the development of innovative treatment approaches.
As demonstrated by the findings of Berger et al., 2018, it can be confirmed that, although there are similarities among the various types of breast and gynecologic cancers, they exhibit a range of significant distinctions, each with its own unique characteristics and behaviors.

3.6 Survival Heatmap

After identifying the genes that were common among different types of cancers with respect to patient survival, an analysis was conducted to determine which specific genes could negatively impact the survival of individual patients and to what extent. To this end, a heatmap was created for each cancer type, illustrating the effect of significant genes on the survival of each patient. In these heatmaps, each row represents a gene symbol and the associated type of omic, separated by a dot. The columns of the heatmaps correspond to the IDs of patients with that specific cancer. Furthermore, each heatmap is divided into two sections, with the top section displaying a score for each patient and the bottom section representing the magnitude of the effect of each gene/omic on each patient. The score is calculated by summing the coefficients for each patient, with higher scores indicating a greater risk of death. Since the coefficients for each patient are calculated using the same parameters (such as lambda), it is possible to compare the coefficients of patients with a single type of cancer.

Only genes that had a negative effect on patient survival were selected for inclusion in the heatmap, with all genes represented having a negative effect on at least one patient. However, the magnitude of the negative effect of each gene/omic may vary, and these variations are represented by different colors, ranging from white to red. The redder a spot on the heatmap, the greater the negative effect of an alteration in that omic for that specific gene on that specific patient.

Moreover, In the heatmaps presented in this research, the genes are carefully ordered based on their relative importance in determining patient survival. This ordering is achieved by calculating the sum of coefficients for each gene/omic and arranging them in descending order. The score values for each cancer type are then meticulously divided into three distinct ranges, representing the low, medium, or high risk of death for each individual patient. These ranges are determined based on a thorough analysis of the quantiles of scores for each cancer type.

Patients with scores falling within the first 25% of the quantile (including the 25th percentile) are assigned to the low-risk group, which is visually represented by a light blue color. Patients with scores falling between the 25th and 75th percentiles (including the 75th percentile) are classified as medium risk and are represented by an orange color. Patients with scores falling within the top 75% of the quantile are considered to be at high risk and are represented by a dark red color. The heatmaps for each cancer type, which provide a detailed and comprehensive visual representation of these findings, are presented below (Figure 17, Figure 18, Figure 19, Figure 20, Figure 21).

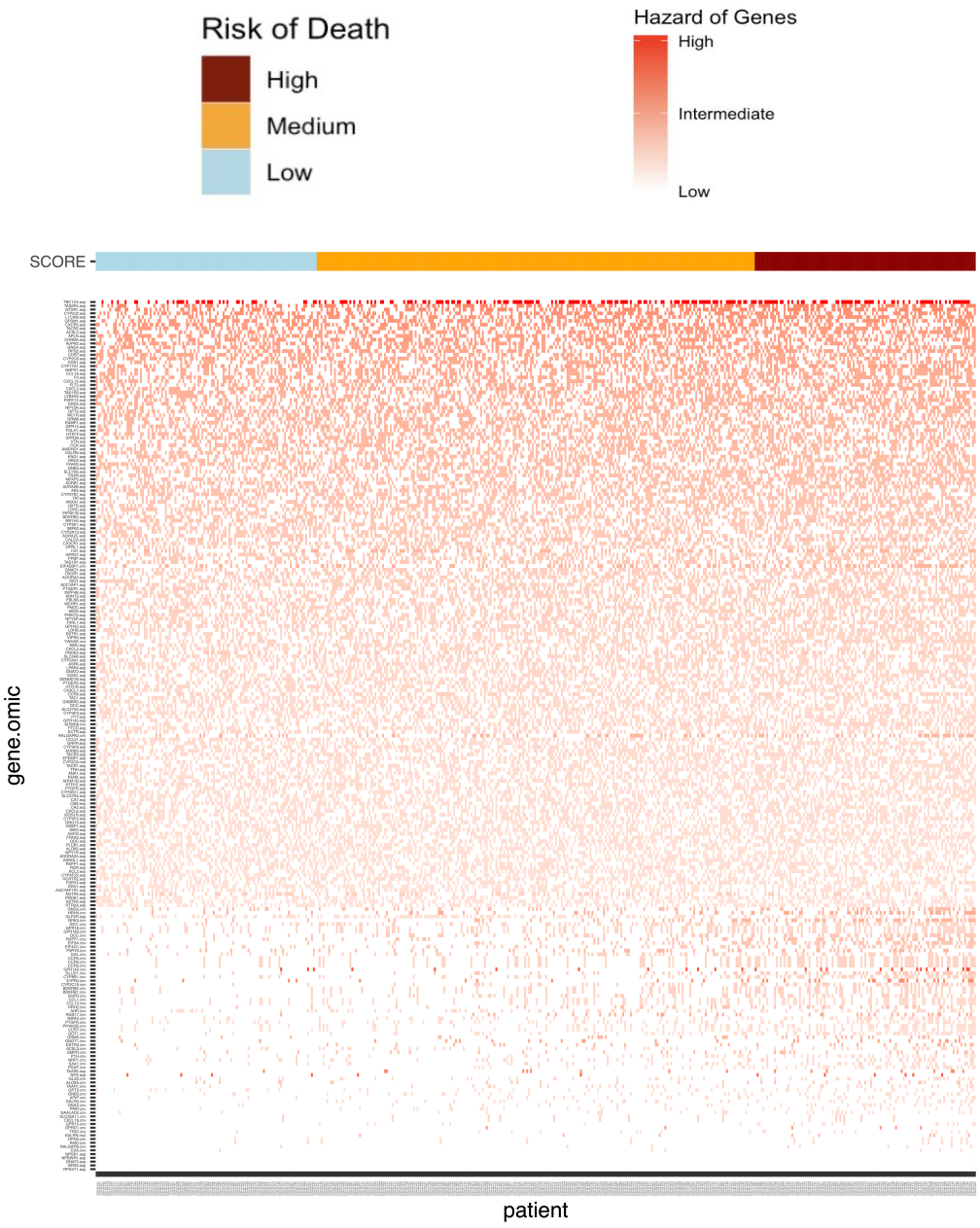

Figure 17. ***BRCA Survival Heatmap.*** As depicted in the heatmaps, the patients are carefully separated into three distinct groups based on their relative risk of death. These groups are visually represented by different colors, with light blue indicating low risk, orange indicating medium risk, and dark red indicating high risk. The genes, on the other hand, are separated by different shades of color ranging from white to red. The intensity of the red color corresponds to the magnitude of the effect that a particular gene/omic can have on patient survival. The darker the shade of red, the greater the potential impact of that gene/omic on survival.

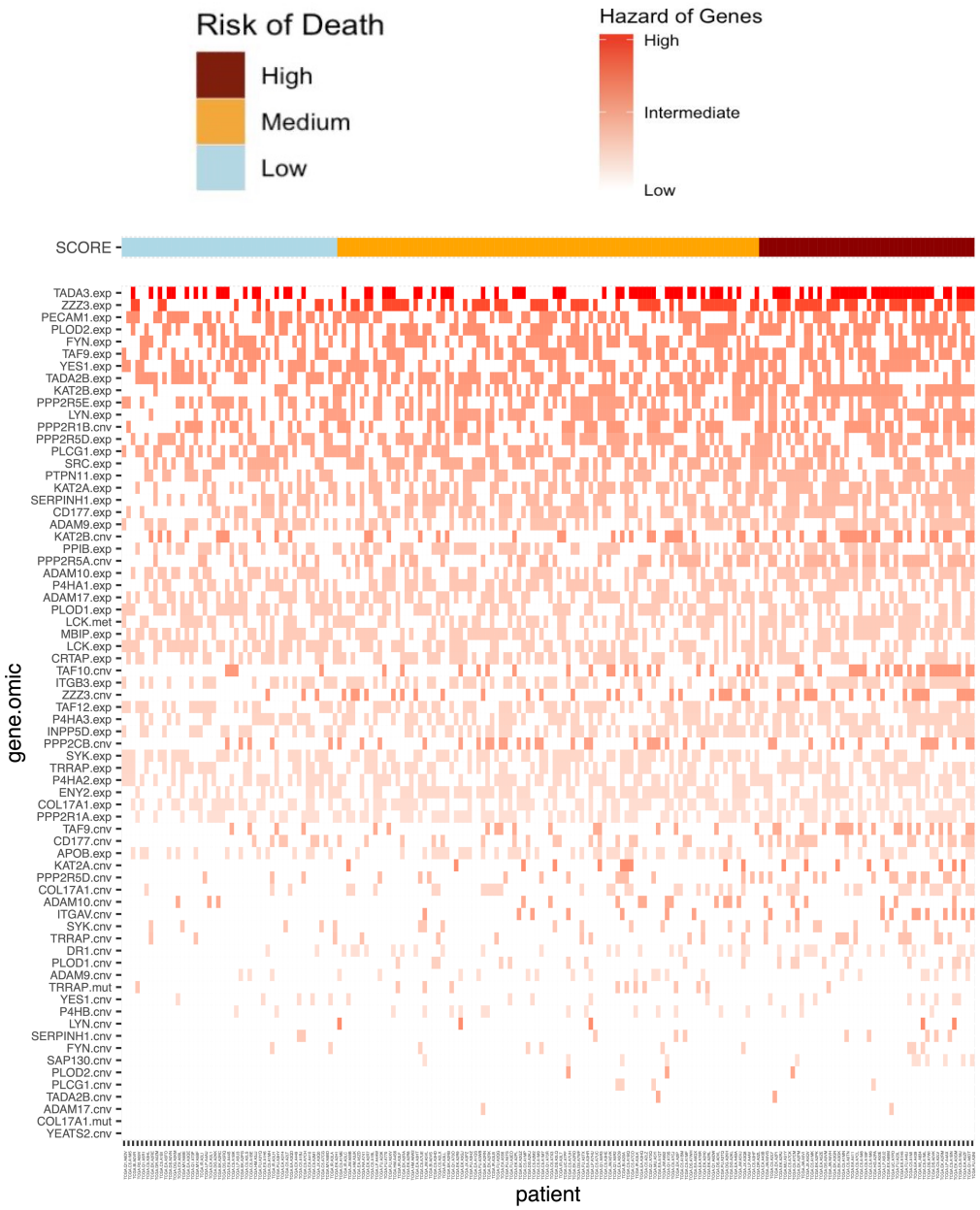

Figure 18. ***CESC Survival Heatmap.*** As clearly illustrated in the meticulously constructed heatmaps, patients are carefully divided into three distinct groups based on their relative risk of death. These groups are visually represented by different colors, with light blue indicating low risk, orange indicating medium risk, and dark red indicating high risk. The genes, on the other hand, are separated by different shades of color ranging from white to red. The intensity of the red color corresponds to the magnitude of the effect that a particular gene/omic can have on patient survival. The darker the shade of red, the greater the potential impact of that gene/omic on survival.

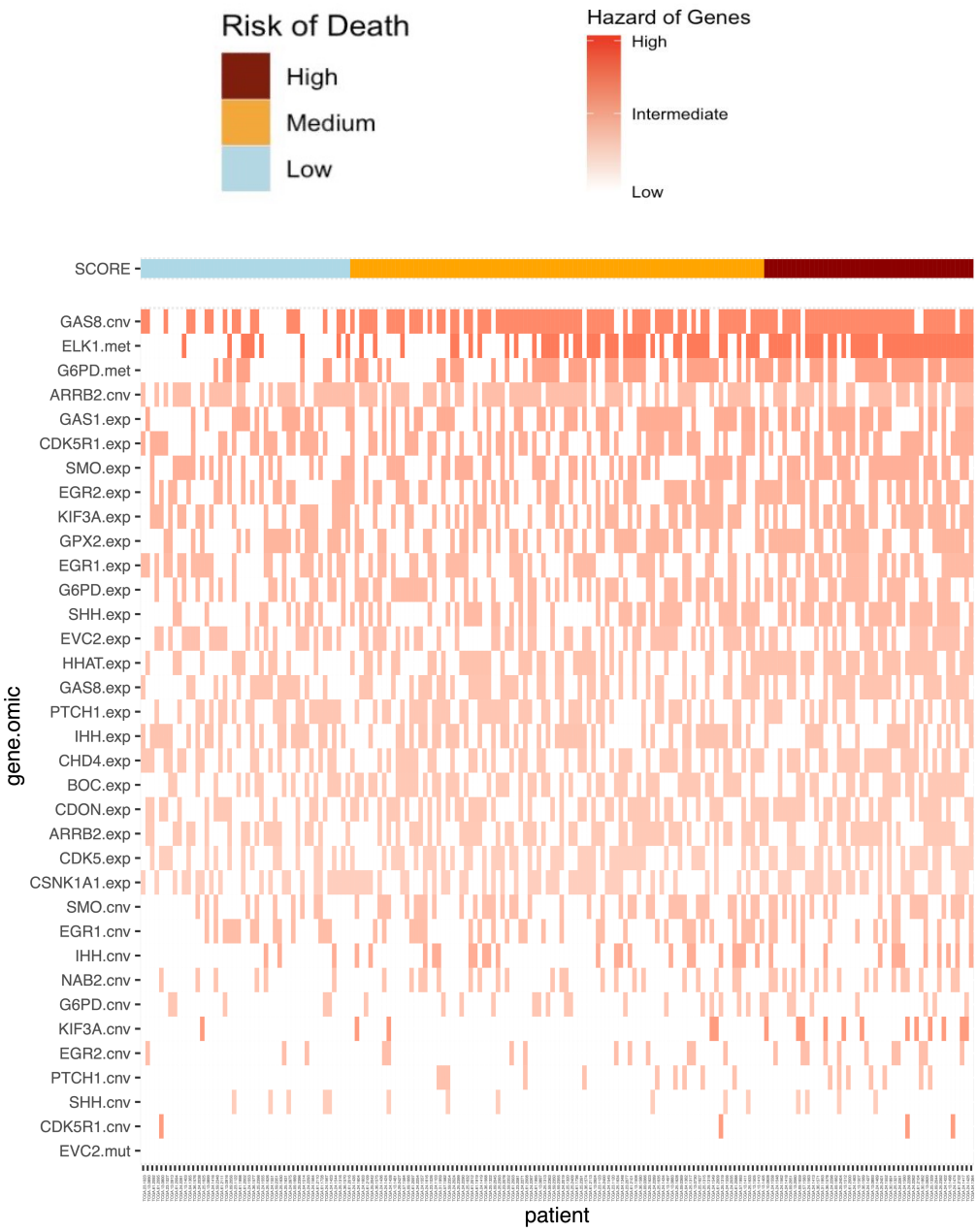

Figure 19. ***OV Survival Heatmap.*** The heatmaps provide a detailed and comprehensive visual representation of the patients, who are meticulously categorized into three groups according to their death risk. These groups are differentiated by distinct colors, with light blue representing low risk, orange representing medium risk, and dark red representing high risk. The genes are distinguished by varying shades from white to red. The more significant the impact of a gene/omic on survival, the more intense its red hue will be.

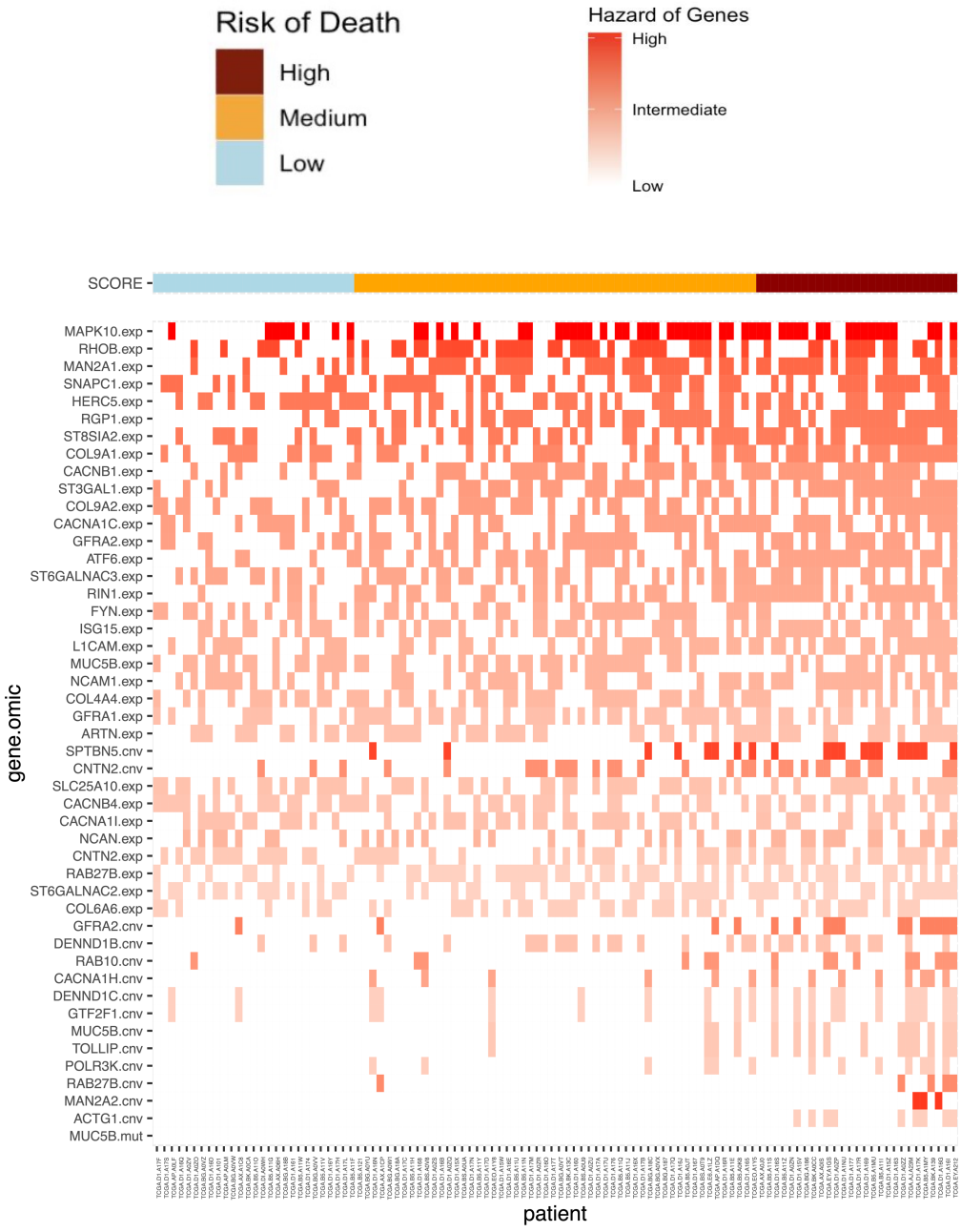

Figure 20. ***UCEC Survival Heatmap.*** As shown in the carefully crafted heatmaps, patients are classified into three well-defined groups based on their likelihood of death. These groups are visually differentiated by different colors, with light blue denoting low risk, orange denoting medium risk, and dark red denoting high risk. The genes are differentiated by shades of color from white to red. The greater the influence of a gene/omic on survival, the deeper its shade of red will be.

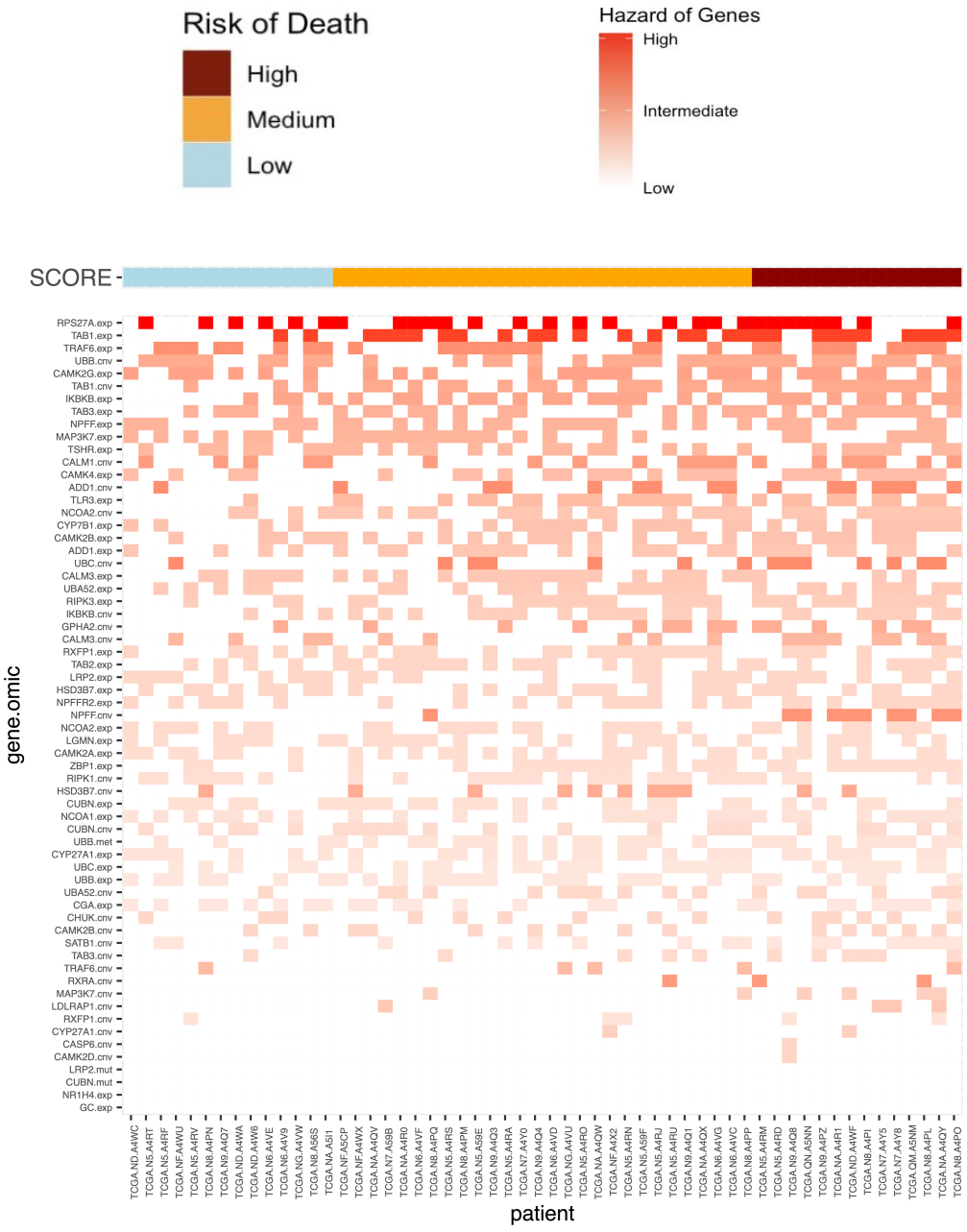

Figure 21. ***UCS Survival Heatmap.*** The heatmaps demonstrate with great clarity that patients are sorted into three distinct groups according to their risk of death. These groups are represented by different colors, with light blue signifying low risk, orange signifying medium risk, and dark red signifying high risk. The genes are separated by different shades from white to red. The more substantial the effect a gene/omic has on survival, the more vivid its red coloration will be.

3.7 Kaplan Meyer Plots

After conducting a meticulous and comprehensive analysis of the heatmaps, it was determined that a comparison of Kaplan-Meier plots among different types of cancers would provide valuable insights.
In order to create the Kaplan-Meier plot for each cancer type, the range of risk for death that had been used in the heatmap plots was utilized once again. Specifically, since patients had been carefully separated based on their sum of coefficients, which was visually represented by the "SCORE" label in the heatmap, this information was extracted for each individual patient.
Thus, a new data frame was then created that combined the previous survival data with the group into which each patient had been categorized. As previously mentioned in the heatmap section, there were three distinct groups of patients. The first group consisted of patients with scores in the top 75% quantile of patient scores, who were classified as high risk. The second group was the intermediate group, consisting of patients with scores between 25% and 75%. The final group was the low-risk group, consisting of patients with scores in the first 25% of scores.
By utilizing the grouping tags of High, Intermediate, and Low for each patient, it was possible to determine the group into which each patient fell prior to creating the Kaplan-Meier plot. Specifically, these three groups were passed to the Kaplan-Meier plot creation function, and a separate Kaplan-Meier plot was generated for each cancer type. As expected, each plot contained three groups of patients: the Low-risk group represented by a light blue color, the Intermediate group represented by an orange color, and the High-risk group represented by a dark red color.
The underlying rationale for this approach was that if the previous tests and analyses had been performed correctly, the resulting Kaplan-Meier plot for each cancer type should accurately represent the relative survival rates of the different risk groups. Specifically, the Kaplan-Meier plot should show the lowest survival rate for the light blue group, which was classified as low risk, an intermediate survival rate for the orange intermediate group, and the worst survival rate for the dark red group, which was classified as high risk. As depicted in Figure 22, Figure 23, Figure 24, Figure 25, and Figure 26, the separation of patients into different risk groups based on their survival rates was almost perfectly achieved for all types of cancer. For all five cancer types, the high-risk group generally exhibited the worst survival rates, while the low-risk group exhibited the best survival rates. Furthermore, the survival rates of the intermediate group were typically situated between those of the low and high-risk groups in most parts of the Kaplan-Meier plot for all cancer types, which is entirely consistent with expectations.

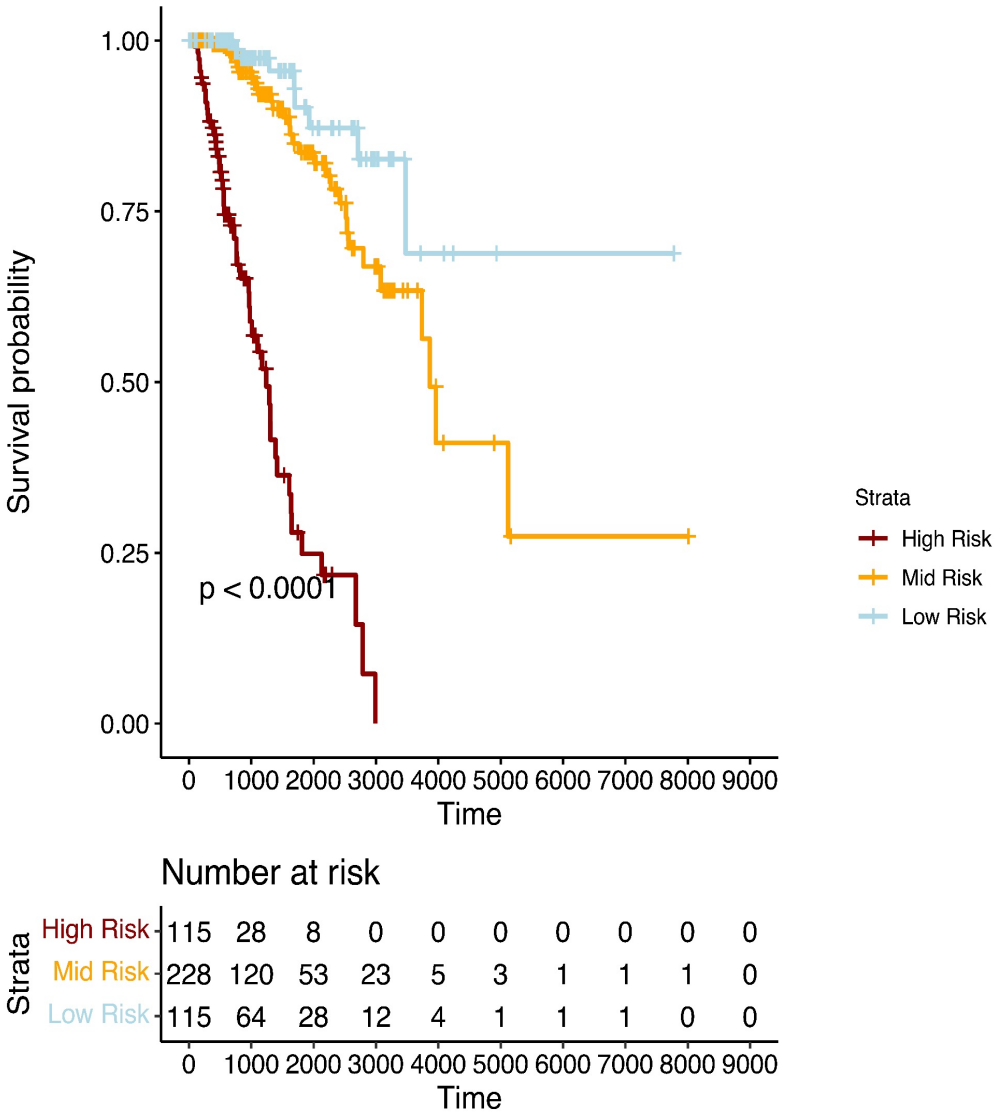

Figure 22. ***Kaplan-Meyer Plot for BRCA.*** As depicted in the Kaplan-Meier plot for BRCA, the high-risk group exhibits the lowest survival rate. This group is characterized by a higher likelihood of experiencing adverse outcomes and a shorter overall survival time. In contrast, the mid-risk group displays an intermediate survival rate, indicating a moderate likelihood of experiencing adverse outcomes and a longer overall survival time than the high-risk group. Finally, the low-risk group demonstrates the highest survival rate, indicating a lower likelihood of experiencing adverse outcomes and the longest overall survival time among the three groups. These findings highlight the importance of accurately identifying and stratifying patients into risk groups to inform treatment decisions and improve patient outcomes.

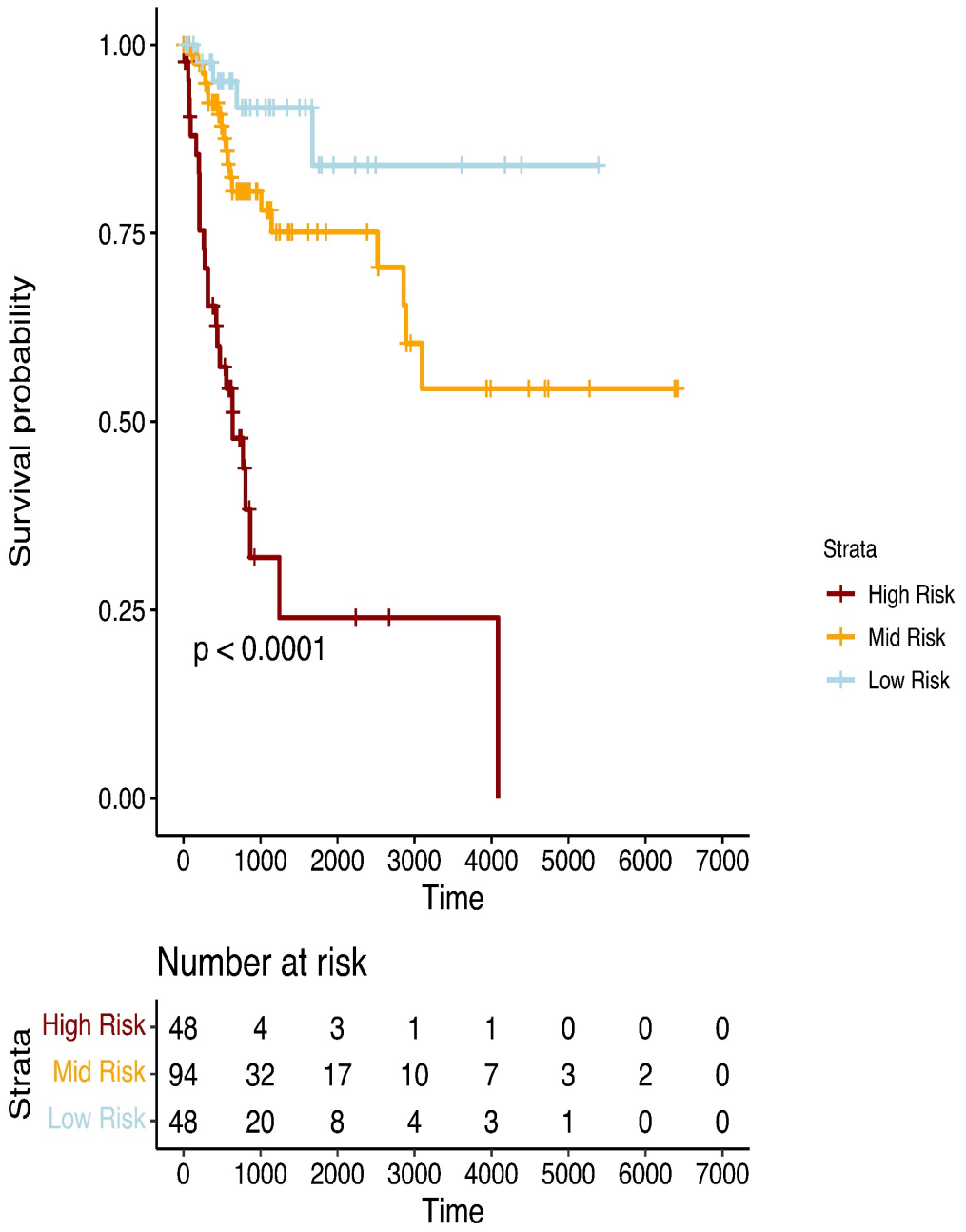

Figure 23. ***Kaplan-Meyer Plot for CESC.*** The Kaplan-Meier plot for CESC reveals that the high-risk group has the lowest survival rate, while the mid-risk group has an intermediate survival rate and the low-risk group has the highest survival rate. These results underscore the importance of accurately stratifying patients into risk groups to inform treatment decisions and improve patient outcomes.

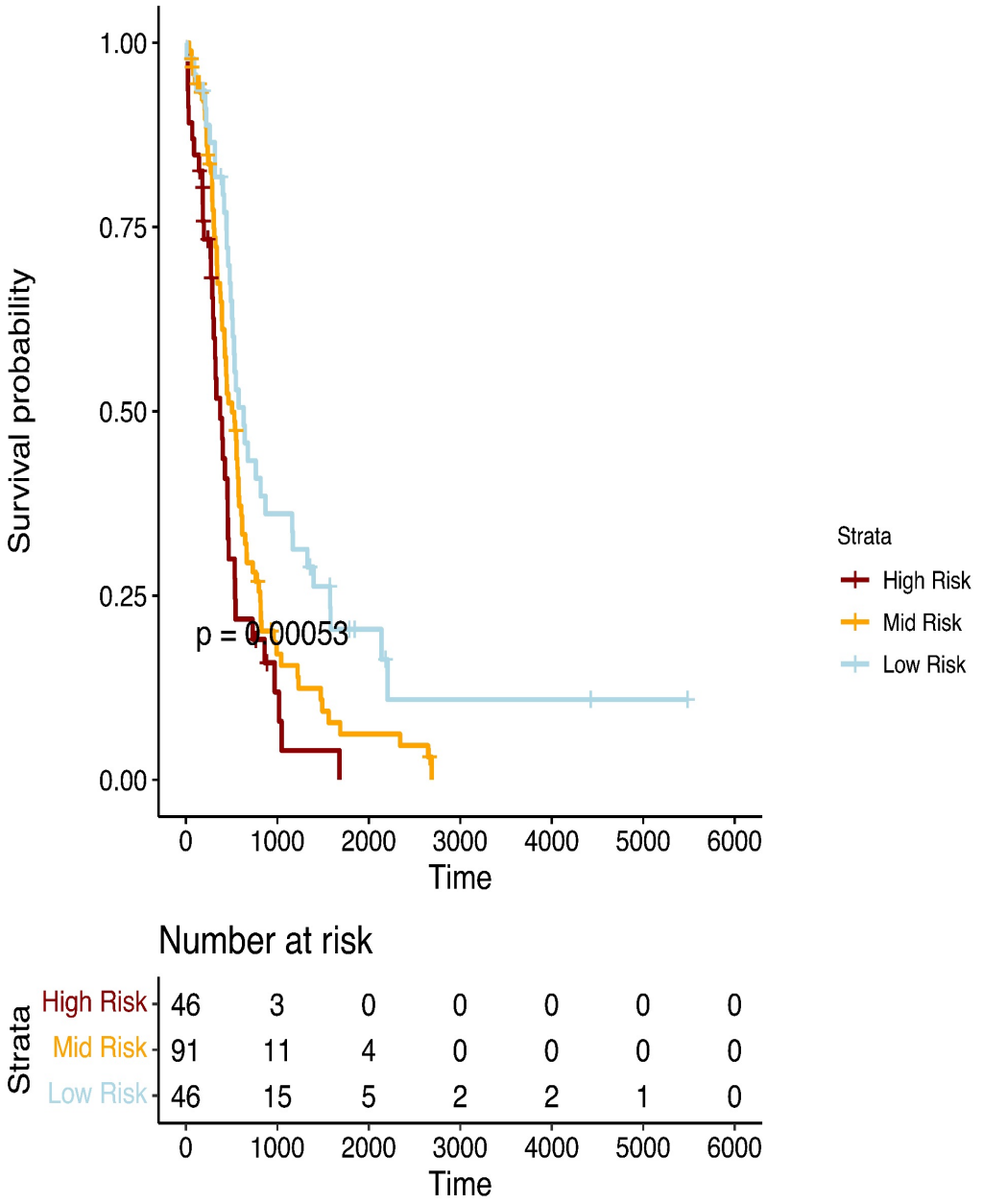

Figure 24. ***Kaplan-Meyer Plot for OV.*** According to the Kaplan-Meier plot for OV, the high-risk group experiences the lowest survival rate, whereas the mid-risk group has a moderate survival rate and the low-risk group has the highest survival rate. These findings emphasize the significance of correctly categorizing patients into risk groups to guide treatment choices and enhance patient outcomes.

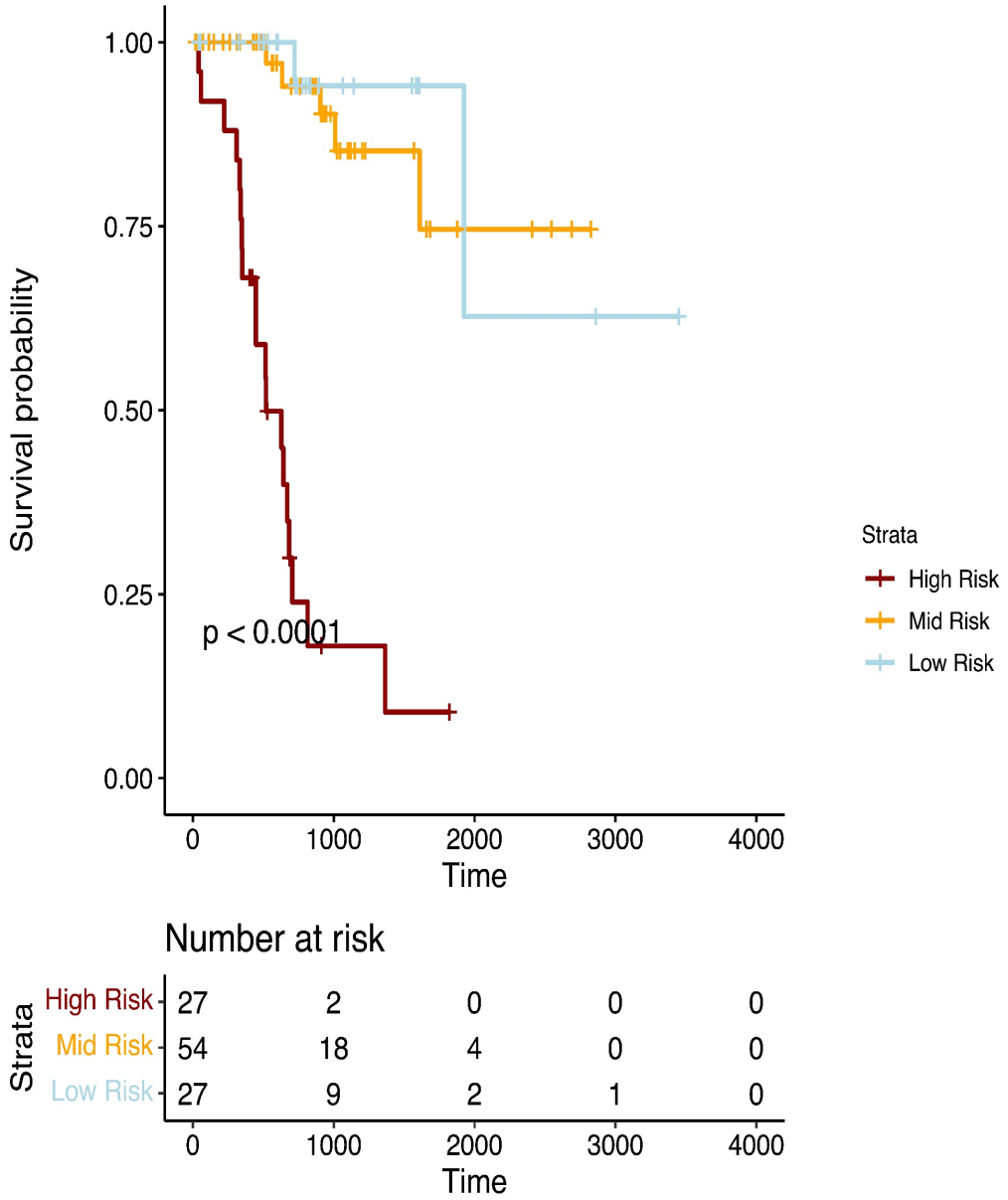

Figure 25. ***Kaplan-Meyer Plot for UCEC.*** As shown in the Kaplan-Meier plot for UCEC, the high-risk group exhibits the poorest survival rate, while the mid-risk group displays a moderate survival rate and the low-risk group demonstrates the best survival rate. These results highlight the value of properly classifying patients into risk groups to inform treatment options and improve patient outcomes.

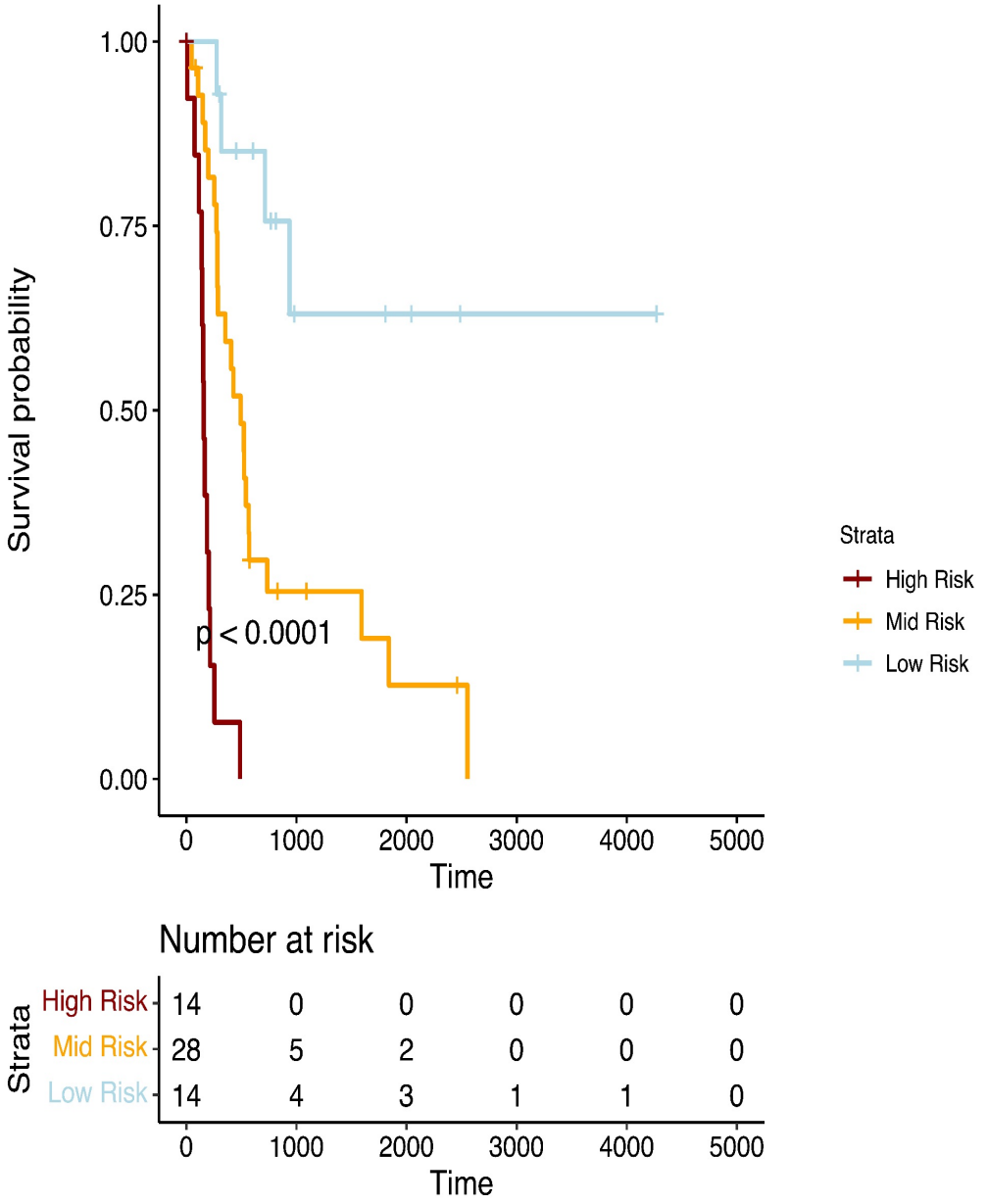

Figure 26. ***Kaplan-Meyer Plot for UCS.*** As demonstrated in the Kaplan-Meier plot for UCS, the high-risk group has the least favorable survival rate, while the mid-risk group has a relatively better survival rate and the low-risk group has the most favorable survival rate. These findings stress the importance of accurately assigning patients to risk groups to guide treatment decisions and enhance patient outcomes.

3.8 Training and testing set

Following the completion of in-depth analyses aimed at identifying which omics and specific genes could negatively impact the survival of patients with a particular type of cancer, it was determined that an additional test should be conducted. The purpose of this test was to assess the accuracy with which the survival prediction could be made.

In the previous analyses, all available patient data was utilized to divide the patients into three distinct risk groups: High, Mid, and Low. However, we were interested in determining whether it would be possible to accurately predict the same results for patients who were not included in the training set. To this end, we randomly excluded 10% of our patients for each type of cancer and proceeded to identify the non-zero coefficients for genes based on the data of the remaining 90% of patients.

The process of reducing the coefficient towards zero and identifying the non-zero coefficients was carried out in the same manner as previously described. However, in this instance, instead of utilizing the entirety of the data to determine the coefficients, only a subset comprising 90% of the patients was employed for this purpose.

The goal of this analysis was to determine whether the 10% of patients who were excluded from our analysis would exhibit the same survival risk as when they were included in the data used to determine the coefficients. To this end, after identifying the coefficients based on 90% of the patients, the non-zero coefficients for each gene were utilized to divide this 10% of patients into three distinct risk groups: Low, Mid, and High. Subsequently, an assessment was conducted to determine how many of the patients' survival risks were accurately predicted in the same manner as before.

For patients with BRCA cancer, the survival risk was accurately predicted 69.56522% of the time. This means that in nearly 70 percent of cases, the correct survival risk of patients could be determined without including them in the data training process. In the case of CESC cancer, the prediction accuracy was significantly higher, with a true prediction rate of 94.73684%. The next cancer type analyzed was OV, which yielded a promising accuracy rate of 94.44444%. The final two cancer types, UCEC and UCS, had accuracy rates of 81.81818% and 66.66667%, respectively.

4. CONCLUSION

In conclusion, this study represents a significant contribution to the field of cancer research by successfully employing a multi-omics approach to analyze the survival of patients with Breast Carcinoma and various Gynecologic Cancers. By incorporating gene expression, methylation, copy number variation, and mutation data, the study was able to identify pathways and specific genes within those pathways that were significantly associated with patient survival.

The MOSClip R package, a topological pathway analysis tool, was utilized to identify significant pathways, modules, and genes in survival analysis. This tool was chosen for its unique ability to perform survival analysis using multi-omics data while accounting for interactions among genes.

Cytoscape was then used to visualize the topology of significant genes within each module. Through this analysis, 33 genes were identified as being common among different types of cancers. A comprehensive literature review was conducted to compare our findings with those of other studies. This review confirmed that, although there are similarities among the various types of breast and gynecologic cancers, they exhibit a range of significant distinctions, each with its own unique characteristics and behaviors.

Afterwards, heatmaps were created for each cancer type to illustrate the effect of significant genes on patient survival. In these heatmaps, the genes are carefully ordered based on their relative importance in determining patient survival. The score values for each cancer type are then meticulously divided into three distinct ranges, representing the low, medium, or high risk of death for each individual patient.

Then, Kaplan-Meier plots were compared among different types of cancers to provide valuable insights into the survival rates and differences among cancer types. The underlying rationale for this approach was that if the previous tests and analyses had been performed correctly, the resulting Kaplan-Meier plot for each cancer type should accurately represent the relative survival rates of the different risk groups. The Kaplan-Meier plots for all five cancer types demonstrated that the separation of patients based on their death risk was performed correctly.

An additional test was performed to assess the accuracy of survival prediction, with rates of 69.56522% for BRCA, 94.73684% for CESC, 94.44444% for OV, 81.81818% for UCEC, and 66.66667% for UCS.

In summary, this research provides valuable insights that may inform future research and treatment strategies. We hope that these findings will serve as a foundation for further investigation into the underlying mechanisms and potential therapeutic targets for these cancers.

5. SUPPLEMENTARY

In the supplementary materials section of this thesis, readers will find a wealth of additional information and resources to support their understanding of the research presented in this study. These materials are available via the GitHub link provided here: https://github.com/Amin-Zlf/Integrated-Multi-omics-Survival-Analysis-of-Gynecologic-and-Breast-Cancers. Among these materials is the R script used to perform the analyses described in the main text. This script provides a detailed and transparent overview of the methods used to generate our results and may be useful for those interested in replicating or building upon our findings.
In addition to the R script, the supplementary materials also include all important tables and figures referenced in the main text. These materials are presented in high resolution and quality, ensuring that readers can easily access and interpret the data presented in this study. Overall, the supplementary materials provide a comprehensive and valuable resource for readers interested in delving deeper into the methods and results of this research.

6. ACKNOWLEDGMENT

I would like to express my deepest gratitude to my advisor, Prof. Chiara Romualdi, who has been a constant source of guidance and support throughout this journey. Her expertise and insights have been invaluable, and I am grateful for the countless hours she has spent helping me navigate the complexities of my research. I would also like to thank my controrelatore, Prof. Luca Pagani, for his valuable feedback and guidance. I am grateful for his contributions to my work. I would also like to thank my family and friends for their unwavering encouragement and for always being there for me, even when I was buried under a mountain of books and papers. And last but not least, I would like to thank coffee, without which this thesis would not have been possible.

7. REFERENCES

1. Akin, D. F., & Özkan, D. (2023). Molecular profiling of TAM tyrosine kinase receptors and ligands in endometrial carcinoma: An in silico-study. *Taiwanese Journal of Obstetrics and Gynecology*, *62*(2), 311-324. https://doi.org/10.1016/j.tjog.2022.09.010
2. An, R., Yu, H., Wang, Y., Lu, J., Gao, Y., Xie, X., & Zhang, J. (2022). Integrative analysis of plasma metabolomics and proteomics reveals the metabolic landscape of breast cancer. *Cancer & Metabolism*, *10*(1), 13. https://doi.org/10.1186/s40170-022-00289-6
3. Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., ... & Moore, R. A. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, *33*(4), 690-705. https://doi.org/10.1016/j.ccell.2018.03.014
4. Bloom, E. A., Peters, P. N., Whitaker, R., Russell, S., Albright, B., Cummings, S., ... & Previs, R. A. (2023). Association of Genomic Instability Score, Tumor Mutational Burden, and Tumor-Infiltrating Lymphocytes as Biomarkers in Uterine Serous Carcinoma. *Cancers*, *15*(2), 528. https://doi.org/10.3390/cancers15020528
5. Cai, M., Liang, X., Sun, X., Chen, H., Dong, Y., Wu, L., ... & Han, S. (2019). Nuclear receptor coactivator 2 promotes human breast cancer cell growth by positively regulating the MAPK/ERK pathway. *Frontiers in oncology*, *9*, 164. https://doi.org/10.3389/fonc.2019.00164
6. Cheuk, I. W. Y., Siu, M. T., Ho, J. C. W., Chen, J., Shin, V. Y., & Kwong, A. (2020). ITGAV targeting as a therapeutic approach for treatment of metastatic breast cancer. *American journal of cancer research*, *10*(1), 211. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7017729/
7. Cohen, P. A., Jhingran, A., Oaknin, A., & Denny, L. (2019). Cervical cancer. In *www.thelancet.com* (Vol. 393). http://clinicaltrials.gov
8. Dejima, M., Hashimoto, H., Sasajima, Y., Nomura, N., Sugita, M., & Morikawa, T. (2020). Uterine cervical squamous cell carcinoma with reactive multinucleated giant cells expressing cluster of differentiation 204: A case report and literature review. *Journal of Obstetrics and Gynaecology Research*, *46*(10), 2174–2178. https://doi.org/10.1111/jog.14404
9. Eswaran, S., Adiga, D., Khan, N., Sriharikrishnaa, S., & Kabekkodu, S. P. (2022). Comprehensive analysis of the exocytosis pathway genes in cervical cancer. *The American Journal of the Medical Sciences*, *363*(6), 526-537. https://doi.org/10.1016/j.amjms.2021.12.008

10. Esworthy, R. S., Doroshow, J. H., & Chu, F. F. (2022). The beginning of GPX2 and 30 years later. *Free Radical Biology and Medicine*, *188*, 419-433. https://doi.org/10.1016/j.freeradbiomed.2022.06.232
11. Fang, F., Wang, P., Huang, H., Ye, M., Liu, X., & Li, Q. (2022). m6A RNA methylation regulator-based signature for prognostic prediction and its potential immunological role in uterine corpus endometrial carcinoma. *BMC cancer*, *22*(1), 1-14. https://doi.org/10.1186/s12885-022-10490-x
12. Fang, X., Zhang, T., & Chen, Z. (2023). Solute Carrier Family 7 Member 11 (SLC7A11) is a Potential Prognostic Biomarker in Uterine Corpus Endometrial Carcinoma. *International Journal of General Medicine*, 481-497. https://doi.org/10.2147/IJGM.S398351
13. GDC Docs. (n.d.). Bioinformatics Pipeline: Copy Number Variation Analysis. Retrieved from https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/#copy-number-estimation_1
14. Haddad, S. A., Lunetta, K. L., Ruiz-Narváez, E. A., Bensen, J. T., Hong, C. C., Sucheston-Campbell, L. E., ... & Palmer, J. R. (2015). Hormone-related pathways and risk of breast cancer subtypes in African American women. *Breast cancer research and treatment*, *154*, 145-154. https://link.springer.com/article/10.1007/s10549-015-3594-x
15. Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., Van't Veer, L. J., Lopez-Bigas, N., Laird, P. W., … Stuart, J. M. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, *158*(4), 929−944. https://doi.org/10.1016/j.cell.2014.06.049
16. Koppula, P., Zhang, Y., Zhuang, L., & Gan, B. (2018). Amino acid transporter SLC7A11/xCT at the crossroads of regulating redox homeostasis and nutrient dependency of cancer. *Cancer Communications*, *38*, 1-13. https://doi.org/10.1186/s40880-018-0288-x
17. Koskas, M., Amant, F., Mirza, M. R., & Creutzberg, C. L. (2021). Cancer of the corpus uteri: 2021 update.*International Journal of Gynecology and Obstetrics*, *155*(S1), 45–60. https://doi.org/10.1002/ijgo.13866
18. Łapińska, Z., Szwedowicz, U., Choromańska, A., & Saczko, J. (2022). Electroporation and electrochemotherapy in gynecological and breast cancer treatment. *Molecules*, *27*(8), 2476. https://doi.org/10.3390/molecules27082476

19. Li, X., & Hu, Y. (2021). Attribution of NF-κB activity to CHUK/IKKα-involved carcinogenesis. *Cancers*, *13*(6), 1411. https://doi.org/10.3390/cancers13061411
20. Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Cancer Genome Atlas Research Network, & Hu, H. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052
21. Mallick, A., Jena, S. K., & Kuanar, D. (2018). Bilateral ovarian serous cystadenocarcinoma metastasizing to cervix: a rare case report. *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, *7*(11), 4789. https://doi.org/10.18203/2320-1770.ijrcog20184550
22. Martini, P., Chiogna, M., Calura, E., & Romualdi, C. (2019). MOSClip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic acids research*, *47*(14), e80. https://doi.org/10.1093/nar/gkz324
23. Moisini, I., Zhang, H., D'Aguiar, M., Hicks, D. G., & Turner, B. M. (2021). L1CAM expression in recurrent estrogen positive/HER2 negative breast cancer: A novel biomarker worth considering. *Applied Immunohistochemistry & Molecular Morphology*, *29*(4), 287-292. https://doi.org/10.1097/PAI.0000000000000909
24. National Cancer Institute. (n.d.). TCGA VCF 1.1v2 - GDC Docs. GDC Docs. Retrieved from https://docs.gdc.cancer.gov/Encyclopedia/pages/Mutation_Annotation_Format_TCGAv2/#table-1-file-column-headers
25. Pande, M., Thompson, P. A., Do, K. A., Sahin, A. A., Amos, C. I., Frazier, M. L., ... & Brewster, A. M. (2013). Genetic variants in the vitamin D pathway and breast cancer disease-free survival. *Carcinogenesis*, *34*(3), 587-594. https://doi.org/10.1093/carcin/bgs369
26. Qin, L., Wu, Y. L., Toneff, M. J., Li, D., Liao, L., Gao, X., ... & Xu, J. (2014). NCOA1 Directly Targets M-CSF1 Expression to Promote Breast Cancer MetastasisNCOA1, CSF1, and Breast Cancer Metastasis. *Cancer research*, *74*(13), 3477-3488. https://doi.org/10.1158/0008-5472.CAN-13-2639

27. Ramos, M., Geistlinger, L., Oh, S., Schiffer, L., Azhar, R., Kodali, H., de Bruijn, I., Gao, J., Carey, V. J., Morgan, M., & Waldron, L. (2020). Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO clinical cancer informatics*, *4*, 958–971. https://doi.org/10.1200/CCI.19.00119
28. Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., Chan, T., Chapman, P., Davis, S. R., Gomez-Cabrero, D., Culhane, A. C., Haibe-Kains, B., Hansen, K. D., Kodali, H., Louis, M. S., Mer, A. S., Riester, M., Morgan, M., Carey, V., & Waldron, L. (2017). Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer research*, *77*(21), e39–e42. https://doi.org/10.1158/0008-5472.CAN-17-0344
29. Shi, Z., Zhao, Q., Lv, B., Qu, X., Han, X., Wang, H., ... & Hua, K. (2021). Identification of biomarkers complementary to homologous recombination deficiency for improving the clinical outcome of ovarian serous cystadenocarcinoma. *Clinical and Translational Medicine*, *11*(5), e399. https://doi.org/10.1002/ctm2.399
30. Smolarz B, Nowak AZ, Romanowicz H. Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). *Cancers*. 2022; 14(10):2569. https://doi.org/10.3390/cancers14102569
31. Stewart, C., Ralyea, C., & Lockwood, S. (2019). Ovarian Cancer: An Integrated Review. In *Seminars in Oncology Nursing* (Vol. 35, Issue 2, pp. 151–156). W.B. Saunders. https://doi.org/10.1016/j.soncn.2019.02.001
32. Tong, A., Di, X., Zhao, X., & Liang, X. (2023). Review the progression of ovarian clear cell carcinoma from the perspective of genomics and epigenomics. *Frontiers in genetics*, *14*, 952379. https://doi.org/10.3389/fgene.2023.952379
33. Yang, Y. F., Lee, Y. C., Wang, Y. Y., Wang, C. H., Hou, M. F., & Yuan, S. S. F. (2019). YWHAE promotes proliferation, metastasis, and chemoresistance in breast cancer cells. *The Kaohsiung journal of medical sciences*, *35*(7), 408-416. https://doi.org/10.1002/kjm2.12075
34. Yu, Y., Nie, Y., Feng, Q., Qu, J., Wang, R., Bian, L., & Xia, J. (2017). Targeted covalent inhibition of Grb2–Sos1 interaction through proximity-induced conjugation in breast cancer cells. *Molecular pharmaceutics*, *14*(5), 1548-1557. https://doi.org/10.1021/acs.molpharmaceut.6b00952
35. Zhu, Y., Shen, T., Liu, J., Zheng, J., Zhang, Y., Xu, R., ... & Gu, L. (2013). Rab35 is required for Wnt5a/Dvl2-induced Rac1 activation and cell migration in MCF-7 breast cancer cells. *Cellular signalling*, *25*(5), 1075-1085. http://dx.doi.org/10.1016/j.cellsig.2013.01.015