Heliyon 10 (2024) e29736

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Cross-platform gene expression profiling of breast cancer: Exploring the relationship between breast cancer grades and gene expression pattern

Shamim Sarhadi [a], Arta Armani [b], Davoud Jafari-Gharabaghlou [c], Somayeh Sadeghi [d], Nosratollah Zarghami [c,e,*]

[a] *Institute of Clinical Chemistry and Pathobiochemistry, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Germany*
[b] *Department of Medical Biology and Genetic, Faculty of Medicine, Istanbul Aydin University, Istanbul, Turkey*
[c] *Department of Clinical Biochemistry and Laboratory Medicine, Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran*
[d] *Department of Immunology, Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran*
[e] *Department of Medical Biochemistry, Faculty of Medicine, Istanbul Aydin University, Istanbul, Turkey*

ARTICLE INFO

Keywords:
Gene expression profiling
Transcriptome
Grade classification
Systems biology

ABSTRACT

Gene expression profiling is a powerful tool that has been extensively used to investigate the underlying biology and etiology of diseases, including cancer. Microarray gene expression analysis enables simultaneous measurement of thousands of mRNA levels. Sophisticated computational approaches have evolved in parallel with the rapid progress in bioassay technologies, enabling more effective analysis of the large and complex datasets that these technologies produce.

In this study, we utilized systems biology approaches to examine gene expression profiles across different grades of breast cancer progression. We conducted a meta-analysis of publicly available microarray data to elucidate the molecular mechanisms underlying breast cancer grade classification.

Our results suggest that while grade index is commonly used for evaluating cancer progression status in the clinic, the complexity of molecular mechanisms, histological characteristics, and other factors related to patient outcomes raises doubts about the utility of breast cancer grades as a foundation for formulating treatment protocols.

Our study underscores the importance of advancing personalized strategies for breast cancer classification and management. More research is crucial to refine diagnostic tools and treatment modalities, aiming for greater precision and tailored care in patient outcomes.

1. Introduction

With the advent of high-throughput sequencing and other bioassay platforms, researchers can now generate massive amounts of biological data from cells in various states. To stay at the forefront of biology research, many investigators are turning to integrative approaches for analyzing this data. Such approaches involve both horizontal integration, which entails merging data from a single omics level, and vertical integration, which involves integrating data from two or more omics levels [1,2]. These high-dimensional

* Corresponding author. Professor of Clinical Biochemistry and Laboratory Medicine Faculty of Medicine of Medical Sciences Tabriz, Iran.
E-mail address: zarghamin@gmail.com (N. Zarghami).

https://doi.org/10.1016/j.heliyon.2024.e29736
Received 12 May 2023; Received in revised form 12 April 2024; Accepted 15 April 2024
Available online 16 April 2024

techniques are powerful tools that can shed light on fundamental concepts of cell biology and be used in clinical applications.

Systems biology, an interdisciplinary field that combines statistics, computer science, engineering, and mathematics, has emerged as a powerful tool for deciphering the complexity of biological data. Over the past few decades, these fields have contributed to the development of systems biology approaches [3]. This comprehensive approach, combined with computational methods, holds promise in comprehending the origins of complex diseases like cancer. Researchers in this field believe that embracing a systemic approach offers a broader perspective, leading to crucial insights into the development of cancer [4].

Cancer represents a significant burden on healthcare systems worldwide, and the number of cancer cases is projected to increase to 29.5 million by 2040, according to reports from the World Health Organization (WHO) [5]. Among cancers, breast cancer is the most common malignancy in American women, according to the American Cancer Society Statistics Center. Although fewer people have been dying from breast cancer since around 1975 because we find it earlier, know more about it, and have better treatments, there are still a lot of physical and emotional problems that stay [6]. To address these challenges, numerous studies have sought to identify clinical and molecular features with prognostic value [7]. Various histological and molecular signatures have been established, some of which play a critical role in clinical applications such as cancer classification, response to specific drugs or treatments, prediction of overall survival among patients with different phenotypes, risk assessment, disease classification, and treatment planning for subgroups of patients with desirable outcomes [8].

Tailored medicine uses cell profiling to guide treatment decisions, mostly based on genome analysis [9]. The accurate classification of patients is a critical step in clinical treatment, but discrepancies between classification methods can present both advantages and disadvantages. Therefore, understanding the relationship between molecular-based classification systems and clinical characteristics is essential [10]. Various classification methods are employed in breast cancer, encompassing histological grading, TNM staging, and classification derived from histological, anatomical, and molecular properties. While each of these methods has a different approach, it is important to consider all the information derived from these classification systems in clinical decision-making [11]. Histopathological tumor grading, which measures cellular differentiation and proliferation potential, is commonly used as a histological signature for chemotherapy decision-making. The Elston and Ellis modified Scarff-Bloom-Richardson (SBR) system, also known as the Nottingham grading system, is the most widely accepted method for grade classification. This index is based on the microscopic evaluation of morphological and cytological features of tumor cells, including the degree of tubule formation, nuclear pleomorphism, and mitotic count, which are used to stratify tumors into grade 1 (well-differentiated and slow proliferative), grade 2 (moderately differentiated), and grade 3 (poorly differentiated and highly proliferative) [12].

Gene expression signatures have emerged as a powerful tool for predicting the prognosis and therapeutic response in various diseases, including cancer. These signatures represent specific gene expression alterations that have prognostic or predictive value for a particular phenotype or condition [13]. The first gene expression signature was published in the late 1990s, and since then, numerous studies have evaluated their strengths and weaknesses in practice. While some signatures, such as MammaPrint and Oncotype DX, have been approved for clinical use, many others require further validation in independent studies to gain wider acceptance among the scientific community [14]. MammaPrint, a 70-gene signature, identifies breast cancer patients who may safely avoid chemotherapy, while Oncotype DX, a 21-gene RT-PCR assay, helps to determine patients who are likely to benefit from treatment and avoid unnecessary chemotherapy. However, with the multitude of reported gene expression signatures for various phenotypes and conditions, it is critical to validate their efficacy and reliability in independent studies before their clinical application [14–16].

2. Methods

2.1. Data collection

The data were collected from Gene Expression Omnibus (GEO). Data collection was followed by “Preferred Reporting Items for Systematic Reviews and Meta-Analyses” (PRISMA) criteria for meta-analysis of gene expression data [17,18]. Since the aim of this study is to provide an insight into grade classification molecular profile, we excluded noise sources for example by removing samples treated by chemical and physical agents and eliminating samples with NA clinical annotation.

2.2. Preprocessing of the data

The raw microarray data underwent normalization and preprocessing using the Frozen Robust Multi-Array Analysis (fRMA) method [19]. This method was implemented using the ’fRMA’ package in R statistical software. To accomplish fRMA normalization, the raw data were background-corrected, and quantile normalized. Then fRMA’s pre-computed parameters from a reference database were applied for batch effect correction and robust summarization of probe-level intensities. No specific adjustments or modifications to default fRMA parameters were made during the normalization process.

2.3. Merging data and removing batch to batch variation

Combining data from the different batches for increasing statistical power of gene expression analysis was done with “Combined Association Test” (COMBAT) method [20]. This method was employed to mitigate batch-related variations and harmonize the merged dataset. This methodology is specifically designed to mitigate technical discrepancies between batches, such as variations arising from different sample processing times, platforms, or experimental conditions. COMBAT applies an empirical Bayes framework to adjust for batch effects without compromising the biological signal inherent in the data. The implementation of COMBAT involved estimating and modeling batch effects present in the merged dataset. Then employing empirical Bayes frameworks to adjust gene expression values, ensuring the normalization of data across batches while preserving biological differences. Ultimately, addressing and eliminating batch-related variations in gene expression profiles without introducing bias remains crucial.

2.4. Finding differentially expression genes (DEGs)

The identification of differentially expressed genes (DEGs) involved conducting analyses across three distinct conditions: control versus grade 1, control versus grade 2, and control versus grade 3. These comparative analyses were executed utilizing the Limma package, renowned for its proficiency in employing linear models for microarray data analysis [21]. To ensure robustness and stringent criteria for DEG selection, thresholds were set at a minimum log fold change of >0.5 and an adjusted p-value of <0.05. These criteria were implemented to focus on genes demonstrating a substantial level of differential expression while controlling for false positives, thereby narrowing down the list of genes considered significantly altered across the specified conditions.

2.5. Gene set enrichment and pathway perturbation analysis

The GSEA desktop application was utilized to analyze the differentially expressed genes (DEGs) obtained from the top table function within the limma package, individually for each grade comparison. GSEA was executed with default settings [22]. This approach facilitated the exploration of coordinated gene expression changes within predefined gene sets or pathways, offering valuable insights into their potential roles across different grades. Additionally, the Signaling Pathway Impact Analysis (SPIA) method was implemented to assess pathway perturbation by contrasting its impact against mere over-representation [23]. SPIA provides a deeper understanding of how biological pathways are perturbed beyond simple enrichment. It computationally evaluates the significance of pathway alterations, distinguishing between pathway dysregulation and random gene set over-representation. This approach enables a more nuanced interpretation of pathway changes within the context of the experimental conditions or grades under investigation.

2.6. Survival analysis

To comprehensively assess the prognostic strength, distinctiveness, and practical applicability of the top 150 markers identified for the grade 3 signature, the SigCheck package was employed. This analysis entailed comparing the grade 3 signature against 48 well-established cancer signatures from previous studies. For this evaluation, independent data from the NKI (Netherlands Cancer Institute) dataset, available within the breast Cancer NKI package [24], was utilized. By leveraging this external dataset, the aim was to gauge the performance and uniqueness of the identified grade 3 signature markers in a distinct cohort. This comparative analysis allowed for an in-depth examination of how these top-ranked markers stood against known cancer signatures, shedding light on their potential prognostic significance and specificity within the context of breast cancer. The utilization of the NKI dataset in this comparative assessment facilitated an understanding of the grade 3 signature's predictive power and its potential clinical relevance, providing valuable insights into its utility as a prognostic tool within the broader landscape of established cancer signatures.

Table 1
Characteristics of the individual studies.

GEO ID	Platform	Sample Count (Case-Control)	NA Clinical Data Removed
GSE26639	GPL570 [HG-U133_Plus_2]	226	7
GSE18864	GPL570 [HG-U133_Plus_2]	84	24
GSE21653	GPL570 [HG-U133_Plus_2]	266	7
GSE36771	GPL570 [HG-U133_Plus_2]	107	0
GSE23177	GPL570 [HG-U133_Plus_2]	116	0
GSE10810	GPL570 [HG-U133_Plus_2]	58	9
GSE42568	GPL570 [HG-U133_Plus_2]	121	0
GSE23593	GPL570 [HG-U133_Plus_2]	50	0
GSE17907	GPL570 [HG-U133_Plus_2]	55	4
GSE11001	GPL570 [HG-U133_Plus_2]	30	0
GSE29431	GPL570 [HG-U133_Plus_2]	66	15
GSE20711	GPL570 [HG-U133_Plus_2]	90	0
GSE11121	GPL96 [HG-U133A]	200	0
GSE2990	GPL96 [HG-U133A]	189	81
GSE4922	GPL96 [HG-U133A]	289	0
GSE15852	GPL96 [HG-U133A]	86	2
GSE23988	GPL96 [HG-U133A]	61	4
GSE22597	GPL96 [HG-U133A]	82	0
GSE1456	GPL96 [HG-U133A]	159	12
GSE7390	GPL96 [HG-U133A]	198	2
		2533	167

NA stands for not available.

2.7. Network analysis

For a comprehensive overview of the protein–protein interactions (PPI) in each grade, PPI networks were constructed using the top 150 ranked genes alongside their log-fold changes. The NetworkAnalyst module detector was utilized to screen for functional modules within the PPI networks and identify molecular pathways primarily affected by the signatures of each grade [1,2] and enrichment in KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database. In order to construct and compare networks that were generated with each signature, genes of each gene signature mapped on NetworkAnalyst PPI database. We set ≥ 15 as a cut-off for the degree of nodes to construct sub-networks with hub nodes as a network level signature for each grade.

3. Results

3.1. Overview of the studies included

In this study, 20 microarray datasets were collected. Genome-wide expression profiling of all samples was done using Affymetrix GeneChip Human GenomeU133A and U133 Plus 2.0 Arrays. After merging all datasets, the merged meta data has 105, 337, 887, and 1059 samples, related to control, grade 1, grade 2, and grade 3 respectively. Details of the individual studies are summarized in Table 1.

3.2. The merged dataset quality assessment

To fix differences caused by data batches, we used the COMBAT method. Then, we checked the combined dataset's quality using an MDS plot. This plot helps see how well the batches were fixed (Fig. 1).

3.3. Differentially expression genes analysis

Driver genes of each grade that lead to the development of breast cancer in grade framework found via limma package. There are 763 probe-IDs showing differential expression, summarized to 566 gene symbols associated with grade 1, 830 DE probe-IDs summarized to 625 gene symbols related to grade 2, and 1045 probe-IDs summarized to 795 gene symbols related to grade 3. Fig. 2 shows an expression portrait of 100 top-ranked genes of each grade in randomly selected 30 samples from each grade along with normal samples from the merged file. We employed the Affinity propagation (AP) clustering method from apcluster package in R [25,26] to estimate the best number of k for k-mean clustering of genes (k = 6, 10 and 7 respectively for 100 top-ranked genes of grade 1, grade 2, and grade 3). K-mean clustering was done and visualized with k-mean clustering and Heatmap viewer modules on the GenePattern desktop [27] (Fig. 2).

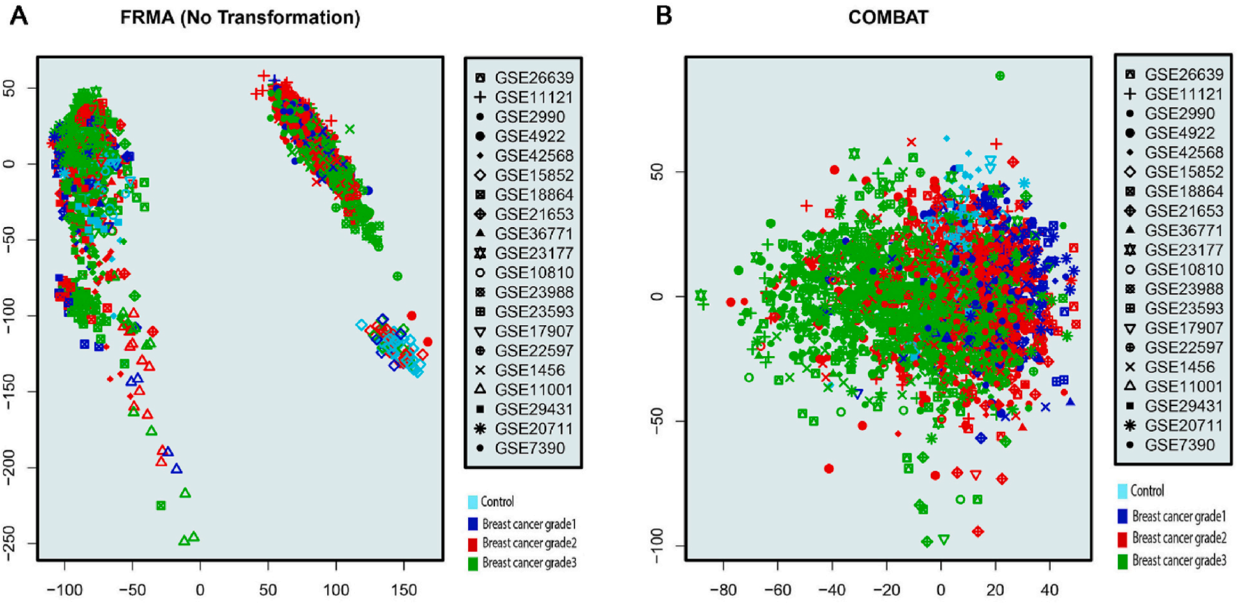

Fig. 1. Visual inspection of merged data sets without any transformation on the left and with performing COMBAT method on the right. Samples are labeled by color based on the biological phenotype of interest and are labeled with a symbol based on a study that they obtained from. As it shown, on the left part of the figure, samples are more clustered by study rather than biological variables that represent unacceptable bias but on the right part of the figure, it is intuitively obvious that samples are clustered by phenotype variables.

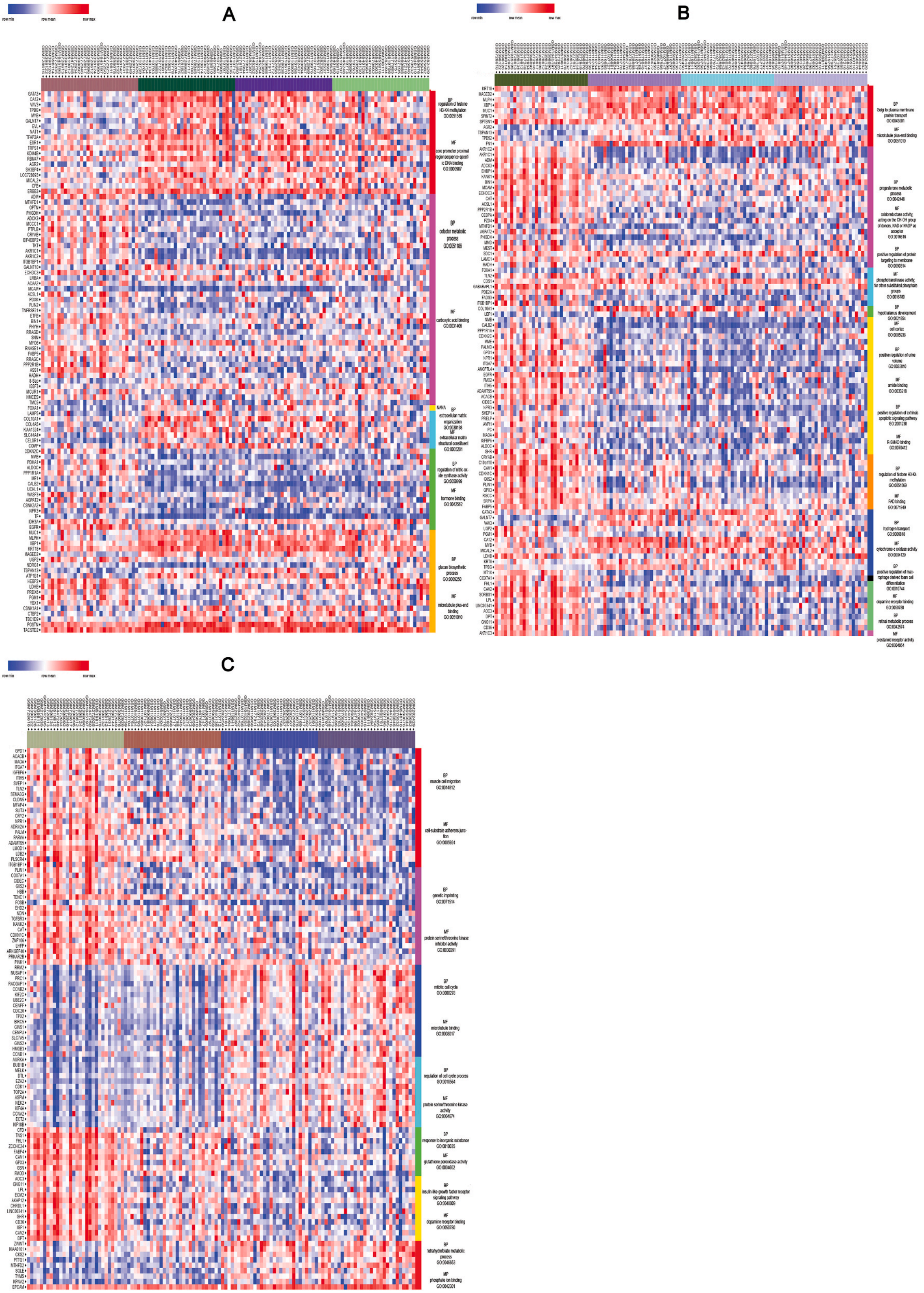

Fig. 2. Expression portrait of 100 top ranked driver genes of each grade in 30 randomly selected samples for each grade. Each cluster enriched in gene ontology biological process and molecular function with Enrichr [28]. Each cluster of columns from left to right represent, normal, grade 1, grade 2 and grade 3 samples, respectively.

3.4. Survival analysis

A Cox model multigene assay was established to evaluate the grade 3 signature's effectiveness in predicting overall survival. We hypothesized that the grade 3 signature would demonstrate superior resolution in distinguishing patients with good and poor prognoses. Fig. 3 presents the Kaplan-Meier curves related to this analysis. (Fig. 3A–C).

Given the ongoing debate about the applicability and utility of gene signatures [29] we evaluated the uniqueness of grade 3 signature by NKI dataset. In this step the data was divided into two data sets, one was the trained data (Fig. 3 A) and the other was the test data (Fig. 3B). As shown in Fig. 3 the signature shows the strong prognostic power to distinguish poor and good prognosis samples in both train and test data ($p = 0.001$). In addition, to compare the performance of this signature with previously published signatures, we test the performance of grade 3 signature with 48 known signatures (Fig. 3C).

3.5. Comparative analysis of grade 3 signature across intrinsic subtypes and clinical variables

Gene expression associated with the Grade 3 signature extracted from GSE7390 dataset. We investigated the relationship between this signature and patient prognosis along with the intrinsic subtypes of breast cancer. Upon visual inspection, there was no apparent correlation observed between high-grade breast cancer and patients with poor prognoses, which contrasts with our initial expectations. Moreover, within each intrinsic subtype, we observed instances of both poor and good prognosis patients. In addition, it represents that the distribution of ER-negative and grade 3 sample is in concordance with basal-like and HER2+ and luminal B samples. Grade 2 samples are in concordance with luminal A and normal-like samples, and to some extent, grade 1 samples. Additionally, ER-negative samples appear connected to basal-like, HER2+, and normal-like breast cancer molecular subtypes (Fig. 4).

3.6. Gene set enrichment analysis (GSEA)

The GSEA is a computational method that assesses whether a prior defined set of genes related to a special phenotype shows statistically significant correlation. To explore the difference and molecular etiology of each grade, gene set enrichment analysis was done with more reproducible probe-IDs of each grade including 5814, 6091, and 6578 related to grade 1, grade 2, and grade 3 respectively that have q-value less than 0.05. Enrichment analysis results from GSEA results enable us to track the events that result in cancer progression from the beginning stages to advance stages. Overview of the grade 1 results represent changes in the extracellular structure of cells and tissue development (Fig. 5A) and also events that involved estrogen-receptor (Fig. 5B). Specifically, the enrichment of grade 1 markers shows change in the tight junction. Also, results from grade 2 enrichment analysis again represent events related to the extracellular matrix and changes related to cell skeletal and mitotic organization and modifications on chromosome structural that start a process that eventuates to the cell cycle.

Gene enrichment analyses of grade 2 markers point to the M phase of the mitotic cell cycle (q-value <0.05), cell cycle process (q-value <0.05) and deposition of new CENPA (Centromere protein A) containing nucleosomes at the centromere (q-value = 0.013) events. Deposition of new CENPA (a unique form of histone H3) refers to action during the late telophase/early G1 phase related to the

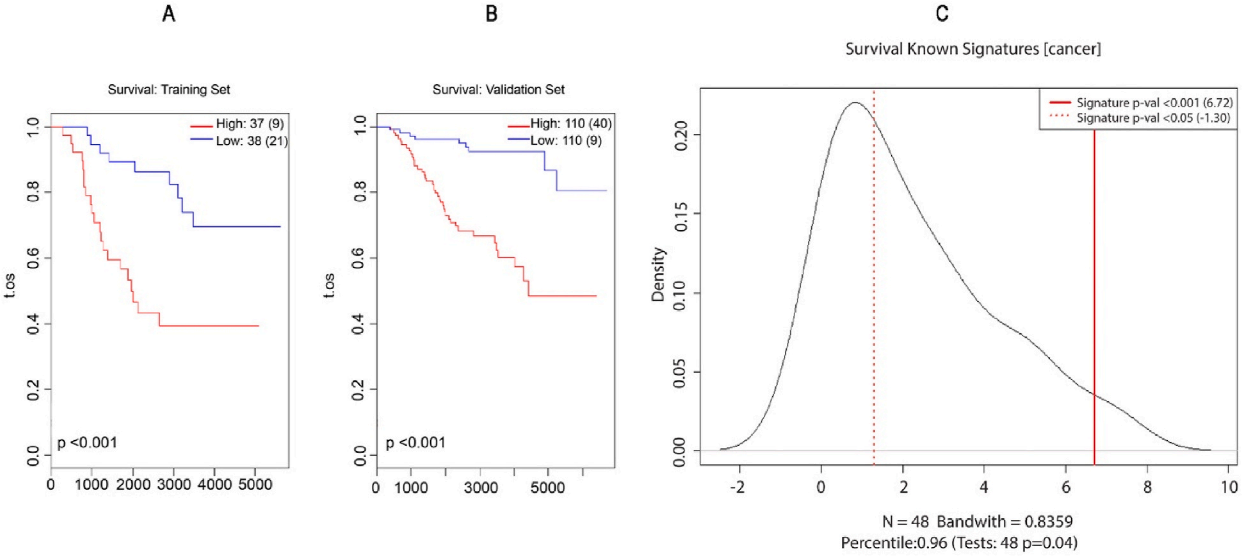

Fig. 3. Assessment the prognostic performance of grade 3 signature. NKI dataset split into two datasets (train and test (A, B)). As it seems this signature has a robust performance to separate good and poor prognostic samples in both train and test dataset (with overall survival as variable). (C) Comparison between grade 3 signature with 48 previously known signatures ($p = 0.001$, 0.05) (15 genes from signature were missing in the NKI dataset).

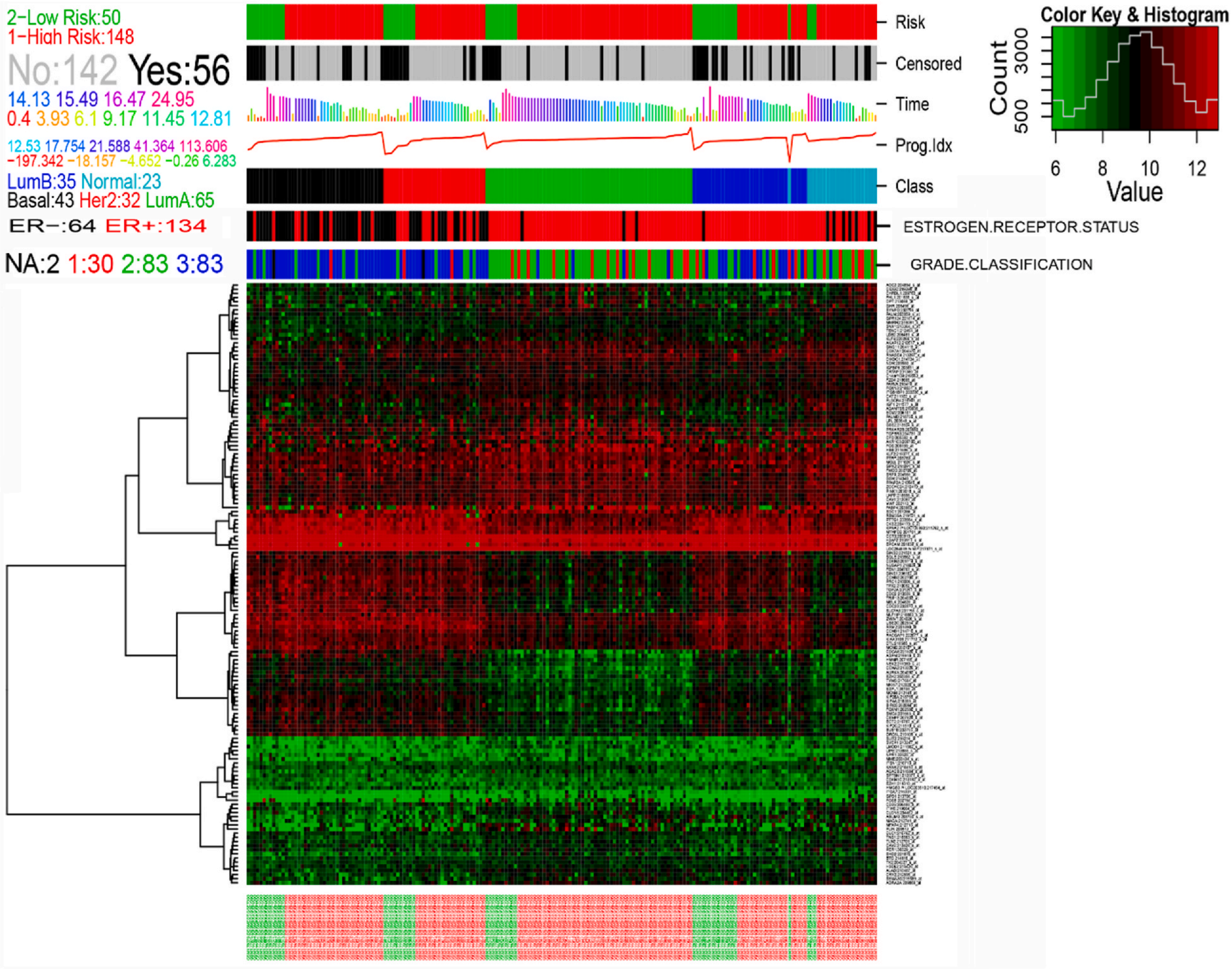

Fig. 4. Gene expression pattern of grade 3 signature in GSE7390. There is unobvious disagreement in gene expression of grade 3 signature in intrinsic subtypes for each gene clusters. Top bars represent other clinical variables that represent correlation and difference between them.

formation of centromeric chromatin (Fig. 5C–F). Furthermore, GSEA for grade 3 markers highlight events related to the cell cycle and the un-differentiation situation of cancer cells. The Grade 3 markers show correlation with M phase of the mitotic cell cycle (q-value <0.05), rosty_cervical_cancer_proliferation_cluster (q-value <0.05) that represents the un-differentiation situation of cancer cells, cell cycle and DNA metabolism ($p < 0.05$) (Fig. 5G–J).

3.7. *Pathway perturbation analysis*

Finding mostly affected pathways that have significant contributions to developing the disease phenotype is one of the important topics of systems biology [30]. In this study impact analysis of pathway in each grade was done with SPIA. Impact analysis identifies significantly impacted pathways that involved two types of evidence, over-representation and accumulated perturbation analysis (perturbation of a pathway that is computed by propagating the measured expression changes across the pathway topology) [23]. This approach offers a more comprehensive understanding of the molecular mechanisms that underlie the classification of breast cancer grades. The most impacted pathway in grade 1 breast cancer cells is ECM-receptor interaction ($p = 0.02$), which represents events that lead to extracellular matrix (ECM) structural abnormalities (Fig. 6). ECM serves an important role in tissue and organ development, and maintenance of cell and tissue structure and also has prominent functions in the direct and indirect control of cellular activity such as migration, adhesion, apoptosis, and proliferation via interactions that mediated with transmembrane molecules like integrins, CD36 and other cell surface components [30,31]. The results suggest a connection between events that weaken tissue structure and drive cells towards a less differentiated state that aligns with grade classification framework. In grade 2 and grade 3, the most impacted pathway is focal adhesion with $p = 0.005$ and $p = 0.007$ for grade 2 and grade 3 respectively. The focal adhesion is the specialized structures at the cell-extracellular matrix contact point. Some of these structure components participate in linking between membrane receptor and actin cytoskeleton, while others are signaling molecules including protein kinases, phosphatases, and a wide variety of

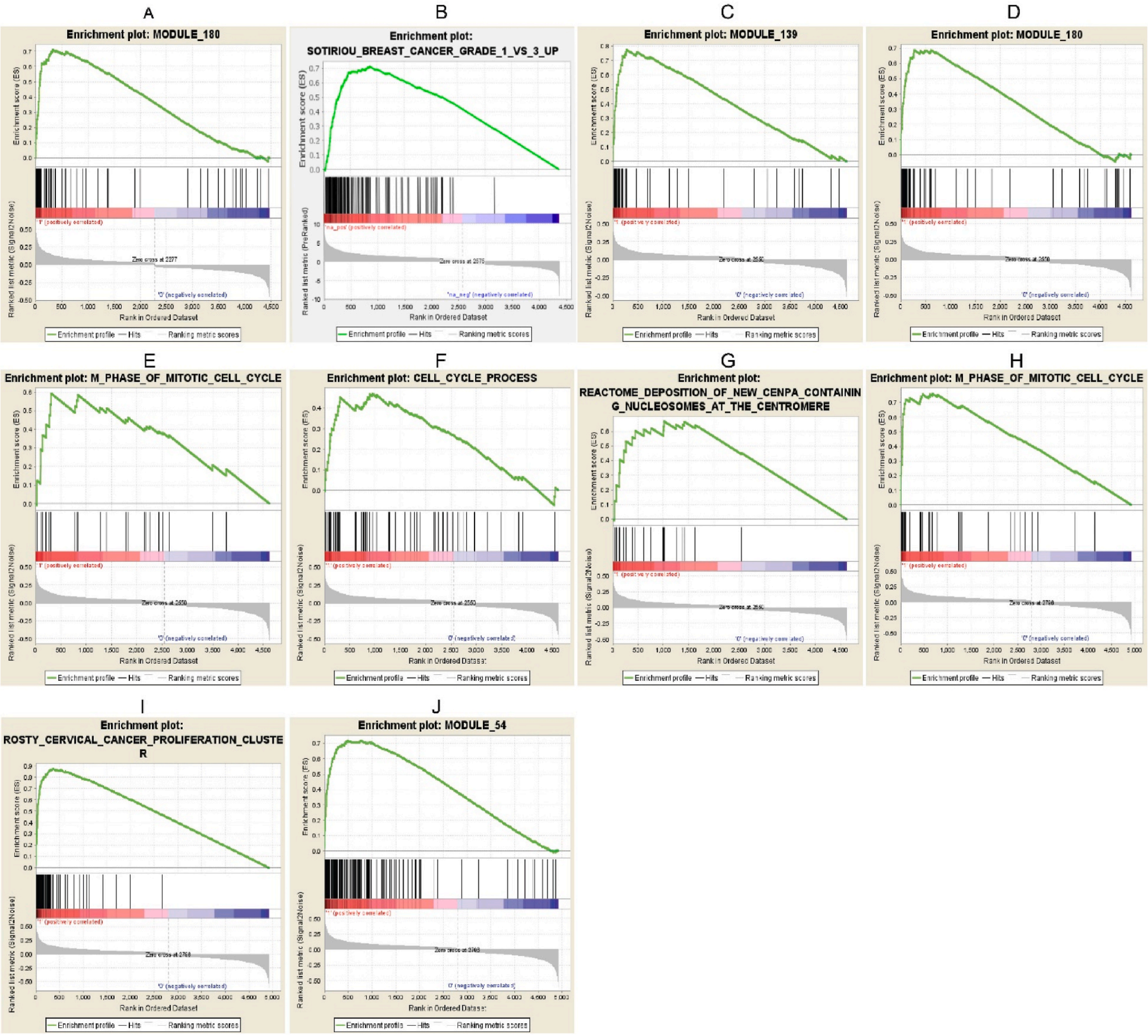

Fig. 5. GSEA results. GSEA of grade 1 markers against computational gene sets and cancer module (A–B). GSEA of grade 2 markers against computational gene sets, cancer module, GO biological process gene set and Reactome gene set (C–F). GSEA of grade 3 markers against Reactome gene set, GO biological process gene set, C2 curated gene sets and cancer module (G–J). The peak of each plot represents highest enrichment score from leading age analysis of GSEA desktop application.

adapter proteins [30,31]. The comparison of affected focal adhesion pathways in grade 2 and grade 3 highlights greater perturbation for FAK and Src genes in grade 3. These proteins are linked to integrin signaling, acting as non-receptor tyrosine kinases along with their adapter proteins to initiate downstream signaling events, especially accentuated in grade 3 compared to grade 2. Given that cancer disrupts the equilibrium between two crucial states—cell growth and cell death—the overexpression of bcl-2, serving as an anti-apoptotic factor, signifies one of the aggressive properties exhibited by cancer cells [32]. Additionally, the number of differentially expressed genes in the focal adhesion pathway is more in grade 3 than grade 2 and this pathway is more perturbed in grade 3 rather than grade 2.

3.8. *Network analysis and enrichment of top-ranked modules*

Tree PPI networks were constructed named grade 1, grade 2, and grade 3 networks. The grade 1 network was constructed with 135 seed proteins and has 306 nodes and 992 edges, grade 2 includes 309 nodes, and 887 edges with 135 seed proteins and the grade 3 network includes 288 nodes, 872 edges, and 133 seed proteins. In grade 1 network UBC, ESR1, FN1, EGFR, IKBKB, in grade 2 UBC, EGFR, FN1, CAV1, KRT18 and in grade 3 network UBC, KIAA0101, FOS, CDK1, CCNA2 are 5 top-ranked hub nodes. We employed walktrap Algorithm [1,2] to find modules in each network. Table [2,3] shows the KEGG pathway enrichment analyses results of each

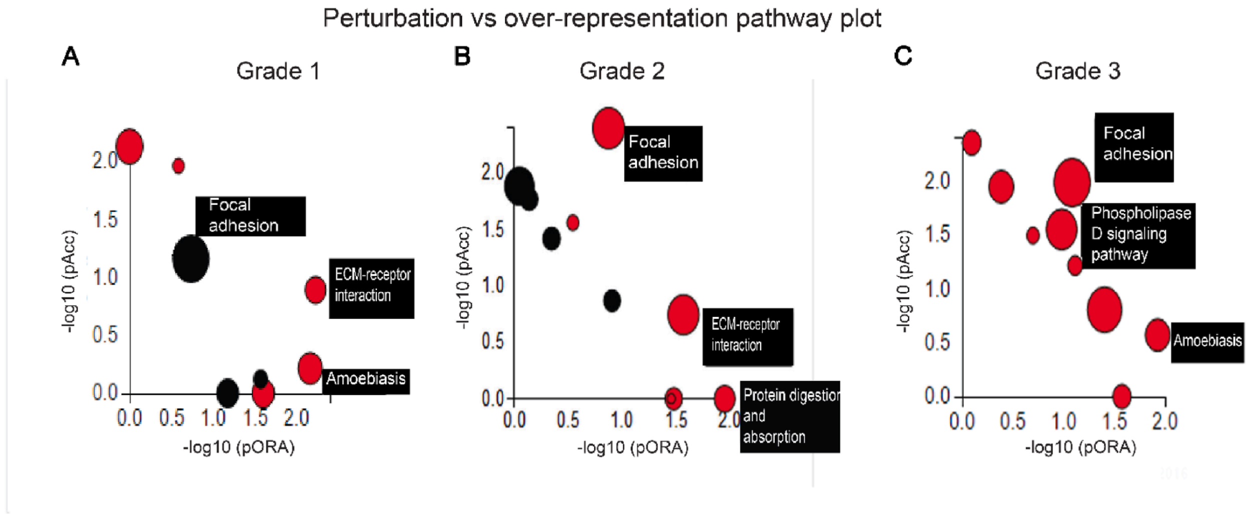

Fig. 6. Perturbation.vs over-representation: From left to right each plot shows the most impacted pathway in grade 1, 2, and 3 respectively. Using negative log of the perturbation and overrepresentation p-values, pathways in red significantly were based on the combined uncorrected p-value, while the ones in black were not.

grade network and top-ranked modules selected by walktrap Algorithm (see Table 2 and 3).

While creating the sub-networks for each grade under the same condition (Degree ≥ 15) as displayed in Fig. 7 (A-C) and conducting enrichment analysis on these sub-networks (as outlined in Table 4), the findings indicate a higher number of ongoing events in grade 2 compared to other grades. This observation might be attributed to this state resembling a transitional phase between aggressive and non-aggressive states in breast cancer, similar to what Anna V. Ivshina et al. suggest in their paper regarding the division of grade 2 breast cancer into grade 2a and grade 2b [33]. Additionally, alongside events aimed at meeting the metabolic demands of cancer cells, crucial for their accelerated growth, there are notable occurrences associated with the cell cycle and metastasis that drive the development of an aggressive state in cancer cells. Fig. 7 illustrates the larger network for grade 2, encompassing hub nodes linked to divergent cancer events in grade 2.

4. Discussion

The study of gene expression data has grown exponentially in recent years, generating vast amounts of data that can provide insights into the underlying mechanisms of complex biological systems, such as cells. To make knowledge from this huge data, large-scale analyses are required to uncover hidden patterns and relationships within the data.

Meta-analysis combines the results of individual studies. However, this approach may suffer from variability in experimental

Table 2
KEGG pathway enrichment analyses of grade 1-3 networks.

Grade 1			**Grade 2**			**Grade 3**		
Name	Hits	p-value	Name	Hits	p-value	Name	Hits	p-value
Pathways in cancer	37	1.79e-11	Pathways in cancer	44	2.18e-14	Cell cycle	35	2.01e-24
Hepatitis C	21	1.83e-11	Focal adhesion	30	1.39e-10	Pathways in cancer	36	7.77e-12
Chronic myeloid leukemia	17	3.11e-10	Chronic myeloid leukemia	18	2.39e-10	Prostate cancer	19	2.28e-11
Prostate cancer	18	7.28e-10	Acute myeloid leukemia	16	3.07e-10	HTLV-I infection	28	2.53e-11
Epstein-Barr virus infection	18	1.58e-9	ErbB signaling pathway	19	7.03e-10	Chronic myeloid leukemia	16	9.08e-10
Chagas disease (American trypanosomiasis)	15	3.66e-7	Hepatitis C	20	1.29e-9	Focal adhesion	23	9.56e-8
Acute myeloid leukemia	12	4.71e-7	Adipocytokine signaling pathway	15	1.37e-8	Epstein-Barr virus infection	15	1.84e-7
Small cell lung cancer	14	5.67e-7	Insulin signaling pathway	22	1.4e-8	Oocyte meiosis	16	3.24e-7
Adherens junction	13	7.19e-7	Prostate cancer	17	3.47e-8	Glioma	12	9.37e-7
Adipocytokine signaling pathway	12	0.000001	Pancreatic cancer	15	5.14e-8	p53 signaling pathway	12	0.00000156

Table 3
KEGG pathway enrichment analysis of grade networks.

Grade 1		**Grade 2**	**Grade 3**
Name	KEGG enrichment	KEGG enrichment	KEGG enrichment
Module1	PPAR signaling pathway, p = 0.00145 Sphingolipid metabolism, p = 0.0117 Adipocytokine signaling pathway,p = 0.0213	PPAR signaling pathway, p = 0.013 Glycerophospholipid metabolism, p = 0.0209 Fat digestion and absorption,p = 0.0247	Biotin metabolism, p = 0.00342 Porphyrin and chlorophyll metabolism, p = 0.06 Staphylococcus aureus infection, p = 0.0664
Module2	Small cell lung cancer, p = 0.00122 Pathways in cancer, p = 0.00837 Leishmaniasis, p = 0.00841	Bacterial invasion of epithelial cells, p = 1.44e-7 Pathways in cancer, p = 0.00000504 Thyroid cancer,p = 0.000274	Cell cycle, p = 0.00000273 Oocyte meiosis, p = 0.0000506 HTLV-I infection, p = 0.000544
Module3	Cytosolic DNA-sensing pathway, p = 0.0000519 Epithelial cell signaling in Helicobacter pylori infection, p = 0.000182 Shigellosis, p = 0.000294	One carbon pool by folate, p = 0.0241 Maturity onset diabetes of the young, p = 0.0304 Neuroactive ligand-receptor interaction, p = 0.0341	Chagas disease (American trypanosomiasis), p = 0.00345 Hepatitis C, p = 0.00433 Transcriptional misregulation in cancer, p = 0.0201
Module4	Cell cycle, p = 0.00000139 Epstein-Barr virus infection, p = 0.000015 Hepatitis C, p = 0.000724	Acute myeloid leukemia, p = 0.00000238 Pathways in cancer, p = 0.000125 Measles, p = 0.000768	Cell cycle, p = 3.18e-13 p53 signaling pathway, p = 5.86e-10 Oocyte meiosis, p = 0.0000149 (This module was not significant)

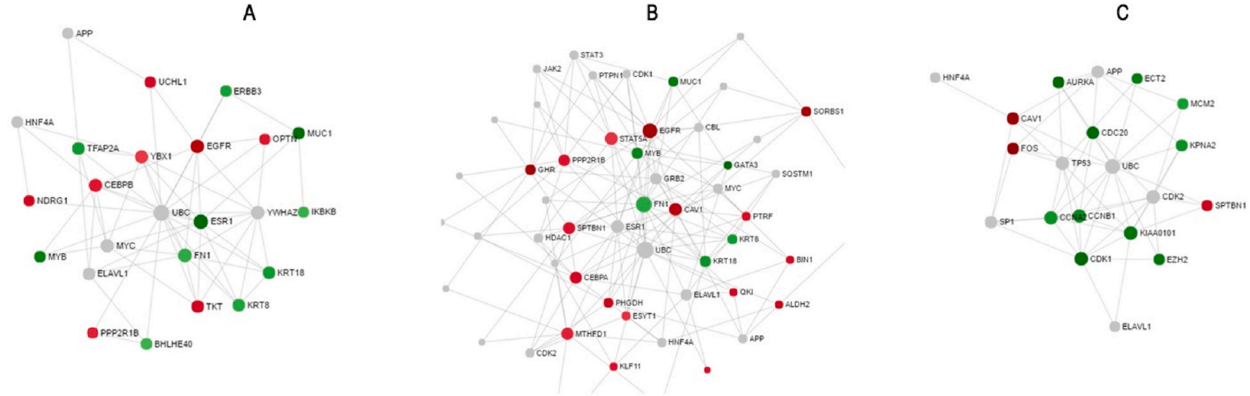

Fig. 7. Tree subnetworks were constructed with hub nodes of each grade (degree ≤ 15). The color of each node represents down (green) and up (red) expression of nodes. Also, the size of nodes represents the measure of changes in expression of nodes. In grade 2 subnetwork (B) the complexity of network is more than grade 1 (A) and grade 3 (C), which shows the situation that a greater number of functional gene clusters related to different interwoven pathways are activated in this grade.

Table 4
GO biological process of sub-networks.

Grade 1/Sub-network			Grade 2/Sub-network			Grade 3/Sub-network		
Name	**Hits**	**p-value**	**Name**	**hits**	**p-value**	**name**	**hits**	**p-value**
regulation of apoptotic process	12	0.00000178	enzyme linked receptor protein signaling pathway	23	3.12e-12	regulation of cell cycle	13	9.5e-12
regulation of programmed cell death	12	0.00000202	transmembrane receptor protein tyrosine kinase signaling pathway	19	1.18e-11	interphase of mitotic cell cycle	10	8.69e-11
negative regulation of apoptotic process	8	0.00000932	positive regulation of metabolic process	32	1.28e-11	interphase	10	1.04e-10

protocols and data quality between studies. An alternative approach for analyzing of gene expression data is data integration, which involves combining the intensity of each feature (prob-IDs) after transforming expression values into numerically comparable measures. This approach provides a more reliable and powerful analysis due to its ability to reduce false discovery rates and increase statistical power.

Several methods have developed for data integration, including DWD, COMBAT, and XPN. However, no single method works perfectly in all situations and with all technologies. In this study, the COMBAT method was chosen due to its simplicity and flexibility, making it suitable for handling different batch sizes and similarities between batches. The use of the COMBAT method successfully removed batch effects and transformed the clustering of data from study-based to phenotype-based.

After identifying differentially expressed genes (DEGs), the grade 3 signature was screened for survival analysis. The results suggested that events related to cancer progression are interconnected and nested, making it difficult to make reliable decisions on treatment protocols only based on grade classifications. This highlights the need for a more comprehensive and personalized approach to cancer treatment, considering the complexity of biological systems.

In addition, network analysis approaches have been proposed as a powerful tool for understanding complex cellular interactions. By systematically representing all molecules and their interactions, researchers can identify important hub nodes in the network, providing insights into the underlying mechanisms of biological systems. Although it may take time for network-derived signatures to be applied in clinical settings, this approach can provide researchers with a more comprehensive understanding of complex systems.

Overall, the results of this study suggest that there is no comprehensive and precise concordance between each grade and the biological mechanisms and survival outcomes of patients in each grade. Therefore, it may not be reasonable to separate patients based solely on a histological and visual classification. A more personalized approach that considers the complexity of biological systems and their interactions may provide a more effective and precise approach to cancer treatment. This study benefitted from a substantial number of samples, enhancing statistical power and reliability of the findings. By incorporating two microarray platforms, this study provided a more comprehensive assessment. A novel aspect of this research was its exploration of the relationship between grade classifications and gene expression patterns, shedding light on potential associations between tumor grading and molecular signatures. While utilizing microarrays, the study was unable to incorporate newer and more advanced technologies like RNA-seq, which could have offered a higher resolution and expanded insights into gene expression profiles of breast cancer grade classification. Exploring more diverse datasets and integrating advanced technological platforms in future studies could significantly enhance the depth and accuracy of understanding the molecular landscape of cancer progression.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Consent to participate

Not applicable.

Data availability statement

The data utilized in this study are sourced from openly accessible databases, specifically from the GEO database, with all GSE accession numbers reported in the text.

CRediT authorship contribution statement

Shamim Sarhadi: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Arta Armani:** Writing – review & editing, Software. **Davoud Jafari-Gharabaghlou:** Writing – review & editing, Software, Methodology. **Somayeh Sadeghi:** Writing – review & editing, Conceptualization. **Nosratollah Zarghami:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

References

[1] J. Xia, M.J. Benner, R.E.W. Hancock, NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration, Nucleic Acids Res. 42 (W1) (2014 Jul 1) W167–W174.

[2] J. Xia, E.E. Gill, R.E.W. Hancock, NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data, Nat. Protoc. 10 (6) (2015 Jun 7) 823–844.

[3] H. Kitano, Computational systems biology, Nature 420 (6912) (2002 Nov) 206–210.

[4] Y. Suhail, M.P. Cain, K. Vanaja, P.A. Kurywchak, A. Levchenko, R. Kalluri, et al., Systems biology of cancer metastasis, Cell Syst 9 (2) (2019 Aug) 109–127.

[5] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 71 (3) (2021 May 4) 209–249.

[6] S.A. Narod, J. Iqbal, A.B. Miller, Why have breast cancer mortality rates declined? J Cancer Policy 5 (2015 Sep) 8–17.

[7] L. Hartwell, D. Mankoff, A. Paulovich, S. Ramsey, E. Swisher, Cancer biomarkers: a systems approach, Nat. Biotechnol. 24 (8) (2006 Aug) 905–908.

[8] R. Aguirre-Gamboa, H. Gomez-Rueda, E. Martínez-Ledesma, A. Martínez-Torteya, R. Chacolla-Huaringa, A. Rodriguez-Barrientos, et al., SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis, PLoS One 8 (9) (2013 Sep 16) e74250.

[9] S.J. Ruberg, L. Shen, Personalized medicine: four perspectives of tailored medicine, Stat. Biopharm. Res. 7 (3) (2015 Jul 3) 214–229.

[10] S. Chen, Y. Li, L. Qian, S. Deng, L. Liu, W. Xiao, et al., A review of the clinical characteristics and novel molecular subtypes of endometrioid ovarian cancer, Front. Oncol. 11 (2021 Jun 3).

[11] N. Min, Y. Wei, Y. Zheng, X. Li, Advancement of prognostic models in breast cancer: a narrative review, Gland Surg. 10 (9) (2021 Sep) 2815–2831.

[12] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, et al., Breast cancer intrinsic subtype classification, clinical use and future trends, Am. J. Cancer Res. 5 (10) (2015) 2929–2943.

[13] F. Chibon, Cancer gene expression signatures – the rise and fall? Eur. J. Cancer 49 (8) (2013 May) 2000–2009.

[14] C.L. Sawyers, The cancer biomarker problem, Nature 452 (7187) (2008 Apr 2) 548–552.

[15] J.S. Reis-Filho, L. Pusztai, Gene expression profiling in breast cancer: classification, prognostication, and prediction, Lancet 378 (9805) (2011 Nov) 1812–1823.

[16] A. Taherian-Fard, S. Srihari, M.A. Ragan, Breast cancer classification: linking molecular mechanisms to disease prognosis, Brief Bioinform 16 (3) (2015 May 1) 461–474.

[17] D. Moher, Preferred reporting Items for systematic reviews and meta-analyses: the PRISMA statement, Ann. Intern. Med. 151 (4) (2009 Aug 18) 264.

[18] A. Ramasamy, A. Mondry, C.C. Holmes, D.G. Altman, Key issues in conducting a meta-analysis of gene expression microarray datasets, PLoS Med. 5 (9) (2008 Sep 2) e184.

[19] M.N. McCall, R.A. Irizarry, Thawing frozen robust multi-array analysis (fRMA), BMC Bioinf. 12 (1) (2011 Dec 16) 369.

[20] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (1) (2007 Jan 1) 118–127.

[21] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Apr 20, Nucleic Acids Res. 43 (7) (2015) e47. e47.

[22] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (43) (2005 Oct 25) 15545–15550.

[23] A.L. Tarca, S. Draghici, P. Khatri, S.S. Hassan, P. Mittal, J sun Kim, et al., A novel signaling pathway impact analysis, Bioinformatics 25 (1) (2009 Jan 1) 75–82.

[24] R. Stark, Checking gene expression signatures against random and known signatures with SigCheck, Bioconductor http://www.bioconductor.org/packages/release/bioc/html/SigCheck.html.2014.Bioconductor, , 2015.

[25] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science (1979) 315 (5814) (2007 Feb 16) 972–976.

[26] U. Bodenhofer, A. Kothmeier, S. Hochreiter, APCluster: an R package for affinity propagation clustering, Bioinformatics 27 (17) (2011 Sep 1) 2463–2464.

[27] H. Kuehn, A. Liberzon, M. Reich, J.P. Mesirov, Using GenePattern for gene expression analysis, Curr Protoc Bioinformatics 22 (1) (2008 Jun).

[28] M.V. Kuleshov, M.R. Jones, A.D. Rouillard, N.F. Fernandez, Q. Duan, Z. Wang, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucleic Acids Res. 44 (W1) (2016 Jul 8) W90–W97.

[29] D. Venet, J.E. Dumont, V. Detours, Most random gene expression signatures are significantly associated with breast cancer outcome, PLoS Comput. Biol. 7 (10) (2011 Oct 20) e1002240.

[30] F.T. Bosman, I. Stamenkovic, Functional structure and composition of the extracellular matrix, J. Pathol. 200 (4) (2003 Jul) 423–428.

[31] E.K. Paluch, I.M. Aspalter, M. Sixt, Focal adhesion–independent cell migration, Annu. Rev. Cell Dev. Biol. 32 (1) (2016 Oct 6) 469–490.

[32] A.T. Ibrahiem, A.K. Makhdoom, K.S. Alanazi, A.M. Alanazi, A.M. Mukhlef, S.H. Elshafey, et al., Analysis of anti-apoptotic PVT1 oncogene and apoptosis-related proteins (p53, Bcl2, PD-1, and PD-L1) expression in thyroid carcinoma, J. Clin. Lab. Anal. 36 (5) (2022 May 7).

[33] A.V. Ivshina, J. George, O. Senko, B. Mow, T.C. Putti, J. Smeds, et al., Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer, Cancer Res. 66 (21) (2006 Nov 1) 10292–10301.