Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and External Validation of a Transcriptomic Prognostic Signature in Triple-Negative Breast Cancer Treated with Neoadjuvant Chemotherapy

Grifton Tafadzwa Muchovu

20063791@studenti.uniupo.it

Universita degli Studi del Piemonte Orientale Amedeo Avogadro https://orcid.org/0009-0001-4465-4116

Meron Abraha Gebremaryam

Universita degli Studi del Piemonte Orientale Amedeo Avogadro

Concorde Niyigaba Isange

Universita degli Studi del Piemonte Orientale Amedeo Avogadro

Laiba Tahir

Universita degli Studi del Piemonte Orientale Amedeo Avogadro

Research Article

Keywords: Triple-negative breast cancer, Prognostic signature, Gene expression profiling, LASSO Cox regression, Survival analysis.

Posted Date: March 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8702492/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Purpose

Triple-negative breast cancer (TNBC) subtype is the most aggressive representative of breast cancer; it does not have any reliable prognostic biomarkers for clinical stratification of the risk associated with the disease. This study is designed for the development and external validation of a transcriptome-based prognostic gene signature which can be utilised for stratifying TNBC patients in accordance with the survival risk.

Methods

The clinical data, along with the gene expression gathered from the GEO dataset GSE25066, were utilised as an initial “discovery” cohort for building a prognostic model. The univariate Cox proportional hazards regression method was utilised for gene mapping from a probe; low-variability genes were filtered along with the identification of survival-associated genes. Moreover, this feature selection was further refined for the construction of a multigene prognostic model utilising the LASSO-penalised Cox regression method. These 17 genes were analysed, and the risk score was calculated likewise. This risk score was used to categorise the patients into high-risk groups and low-risk groups. Finally, external validation of this prognostic model was carried out in (GSE58812), which is an independent GEO dataset utilising risk cut-off and identical gene coefficients.

Results

This study depicted that in the discovery cohort, disease-free survival was poorer in high-risk patients, as compared to low-risk patients. Gene signature in the validation cohort successfully predicted metastasis-free survival, which is an identifiable indicator of reproducibility in the independent datasets. Upon functional enrichment analysis of the signature genes, it was indicated that these genes are key determinants of cellular response to stress, regulation of metabolism, repairing DNA, along with major transcriptionally controlled biological processes in TNBC metastatic progression and chemotherapy resistance.

Conclusion

This study suggests a well-developed transcriptomic framework which provides a baseline for further clinical investigations and futuristic translational research. It also identifies and highlights a prognostic signature of 17 genes, which is reproducible and biologically relevant.

Introduction

TNBC (Triple-negative breast cancer), an aggressive and most intensive subtype of breast cancer, is actually caused by the absence of some hormonal receptors, i.e., estrogen, progesterone and human epidermal growth factor (Receptor Type-II expression). Amongst all the cancers, it accounts for 15–20% of the cases, with the highest recurrence and worst survival rates in comparison with other subtypes of cancer [1]. Cytotoxic chemotherapy is the treatment option for TNBC, and other options are negligible. At first, most of the patients show significant improvement with the symptoms, but substantial metastasis or relapses occur, depicting the intense biology of TNBC even with the standard treatment protocol.

TNBC patients are currently evaluated primarily with clinicopathological indicators, i.e., size of the tumour, status of lymph nodes, and histological grading. Although much of the useful clinical information can be gathered from these variables but it is still not sufficient for adequate assessment and evaluation of molecular heterogeneity, which determines TNBC. TNBC is a multi-molecular-subtype with variable survival outcomes and biological features as depicted by transcriptomic-molecular-heterogeneity [2]. This heterogeneity most probably contributes to the broader term variables when it comes to treatment response and long-term prognosis observed amongst TNBC patients treated with chemotherapy.

Several studies and research on TNBC have tried to assess transcriptomic data (high throughput) for identifying “prognostic gene expression signatures” and determining molecular subtypes and pathway-level variations. The most recent studies, which are carried out, suggested and highlighted the importance of multigene prognostic models utilising machine learning approaches and survival analysis [3, 4]. However, some of the signatures depicted encouraging and strong prognostic behavior, although most of them were isolated from relatively smaller cohorts, which lacked sufficient independent external validation, limiting clinical applicability and reproducibility. Moreover, data pre-processing heterogeneity and strategies of feature selection have significantly contributed towards inconsistent outcomes and findings across the study design.

Approaches related to penalised regression, specifically LASSO-penalised (Cox regression), are significantly highlighted and have gained a lot of attention for survival analysis/assessment of high-dimensional transcriptomic data. Coefficient-shrinkage and simultaneous feature-selection can be carried out utilising LASSO, which thereby devises more concise interpretable prognostic models via reduced overfitting [6].

Fewer prognostic studies carried out on TNBC have adapted this modelling based upon LASSO (despite all the above-mentioned advantages), within a strict discovery to validation window of fixed model coefficients before external validation. Additionally, validation throughout independent cohorts devised upon variable microarray platforms had limited exposure to the existing literature [7].

Biological interpretation data is very limited for these identified gene-signatures, although statistical performance is satisfactory. Usually, predictive accuracy is emphasised; a small number of studies focus on examining biological processes influenced by TNBC, I.e., resistance to treatment regimen, progression of disease and metastasis. Important processes like transcriptional regulation, response to

stress at the cellular level, reprogramming of metabolism, and repairing DNA have been depicted in chemotherapeutic resistance and intensive tumour attitude in TNBC [1, 4], suggesting poor clinical outcomes despite utilising the available treatment option, which is the impact of interlinking transcriptomic programs and biological pathways.

Based upon these observations and findings, we hypothesise that transcriptional programs related to the above-mentioned biological process contribute significantly to poor response related to prognosis in TNBC patients treated with available chemotherapeutic options.

To validate the effectiveness of this study, the targeted goal of this study is the presentation of a transcriptomic prognostic signature which is more robust and depicts biological pathways interpretation utilising available datasets. Dataset GSE25066 was utilised for analysis of the discovery phase, which comprised TNBC patients who adapted a non-adjuvant chemotherapeutic regimen. Survival-associated genes, which were highlighted via univariate Cox regression [5], were refined subsequently utilising LASSO (penalised Cox modelling) [6] for deriving a set of compact genes. It also involved the construction of a prognostic risk score for patient stratification into distinct risk groups. Most important is the fact that the model obtained as a result was validated externally in GSE58812, which is an independent GEO cohort, emphasising generalizability without re-estimation of coefficients. To test its effectiveness in biological processes, functional enrichment analysis was carried out [8]. Adoption of this approach helped to identify the statistical risks as the underlying biological indicators of poor treatment outcomes.

Methodology

Study Design and Objective

This transcriptomic analysis in the study was designed at a discovery-level targeting the development of a prognostic signature (gene expression-based) for TNBC patients receiving Neo-adjuvant chemotherapy. The main objective of this study was to develop a clinically meaningful and significant gene signature which has the capacity for patient stratification into distinguished risk groups having variable outcomes related to survival under a chemotherapeutic treatment regimen because of TNBC molecular heterogeneity and lack of trusted prognostic markers.

This study comprises two two-stage frameworks: (i) Discovery-level cohort in model development, (ii) External Validation (independently) in a distinguished/separate dataset without any alteration or modifications in model coefficients. The focus of this study, which is carried out at the discovery level, was not to attain clinical independence for already existing prognostic variables (proposed for future investigations) rather was identification of a transcriptomic (robust) prognostic signal.

An overview of the study design and analytical workflow is shown in Fig. 1.

Data Sources and Dataset Selection

Clinical data related to gene expressions were extracted from the GEO database (Gene Expression Omnibus). The basic dataset GSE25066 was opted as the “discovery-cohort” reason being it encompasses pre-treatment transcriptomic profiles of the breast cancer patients following Neo-adjuvant chemotherapeutic treatment regimen with DRFS (disease-free relapse-free survival information) and follow-ups. Treatment-targeted survival data make this available dataset the best option for the evaluation of prognosis with reference to chemotherapy response. GPL96 (Affymetrix Human Genome U133A microarray platform) was used for generating GSE25066. Meanwhile, External validation (independent) was carried out utilising the GSE58812 GEO dataset, generated on the U133 Plus 2.0 microarray platform (Affymetrix Human Genome). These specified data sets comprised tumour samples TNBC-specific with annotated MFS (metastasis-free survival) outcomes. Although a great alteration was seen in treatment regimens across available datasets, the great advantage of this cohort is independent reproducibility related to the prognostic signature across populations and platforms. No sample overlap was observed in the validation and discovery cohorts, which enabled a completely independent assessment related to model performance.

Importantly, rather than highlighting the natural history of TNBC, which is left untreated, survival risk is identified for patients under treatment (treated with neoadjuvant chemotherapy).

Inclusion and Exclusion Criteria

For the discovery cohort (GSE25066), patients were included if they had histologically confirmed breast cancer, triple-negative receptor status (ER-negative, PR-negative, HER2-negative), available pretreatment gene expression data, and valid DRFS time and event information. Patients were excluded if survival time or event status was missing or invalid, or if receptor status could not be determined. After filtering, 178 TNBC patients were retained for downstream analyses.

For the external validation cohort (GSE58812), samples were included if gene expression data and corresponding MFS information could be reliably extracted from GEO annotations and matched at the sample level. Only samples with complete expression–clinical linkage were included in validation analyses.

Gene Expression Preprocessing

For both cohorts, probe-level expression data were obtained from GEO series matrix files. Probe identifiers were mapped to gene symbols using the corresponding platform annotation files (GPL96 for GSE25066 and GPL570 for GSE58812). Probes which lack valid gene symbol annotations were significantly removed. For gene representation via multiple probes, expression values were taken to a “single gene-level” utilising median-expression (across probes). This approach significantly affected technical redundancy (reduced it) while keeping robust gene-level signals intact.

Low-Variance Gene Filtering

Filtration of genes in the discovery cohort was carried out for the reduction of noise and removal of features with no information (with a variance threshold of ≤ 0.01). This step significantly improved stability in survival modelling, a reduction in the feature space. No variance filtering was carried out in the external validation cohort to avoid leakage of important information, as feature selection was highly restricted to only the discovery phase.

Outcome Definition

In this discovery-cohort, DRFS (disease-free relapse-free survival). Survival time was in years, and “event status” was noted as 1 for status of recurrence and 0 was assigned for observations which were censored.

In the external validation process of the identified cohort, MFS (metastasis-free survival) was utilised as the endpoint of validation. MFS was defined as the time required from the diagnostic process to the occurrence of metastasis (or last follow-up). Survival times were likely converted from days to year to ensure consistency in cohorts.

Univariate Cox Regression Analysis

Each gene in this process was tested independently for survival utilising univariate Cox proportional hazards regression [5]. Genes retained had a p-value < 0.05 and were “candidate prognostic features”. This threshold was opted to balance out sensitivity and reduction of features in this setting of a discovery phase, with the fact that subsequent penalised modelling would significantly lead to a reduction in false-positive associations.

LASSO-Penalised Cox Regression

LASSO was also applied to the Candidate gene set [6] for the prevention of multicollinearity and overfitting. Moreover, the 10-fold cross-validation was carried out for the optimal regularisation parameter (λ). LASSO was opted because of its performance on shrinkage of simultaneous coefficients and selection of features, resulting in a model which appropriate for downstream validation.

Risk Score Construction and Stratification

LASSO coefficients were utilised for the calculation of the risk score (as a weighted linear combination) for each participant. Via median risk-score (cutoff), patient stratification into high and low risk groups was carried out. The major benefit of this strategy is better optimisation of data with balanced-sized groups.

Survival and Model Performance Analysis

Survival outcomes (in risk groups) were compared by generating survival curves (Kaplan-Meier); on the other hand, a long-rank test was conducted to test statistical significance. Hazard regression (Cox proportional) was performed for the Estimation of hazard ratios (95% C.I.). Similarly, the C-Index (concordance index) was utilised for model discrimination.

Functional Enrichment Analysis

Functional enrichment analysis (Enrichr via the gseapy Python package) of the selected genes was utilised for testing the biological relevance of the selected genes. Functional enrichment analysis was performed using [8]. Gene Ontology biological processes, KEGG pathways, Reactome pathways, and MSigDB Hallmark gene sets were evaluated. Pathways with adjusted p-values < 0.05 were considered statistically significant.

Statistical Analysis

Version 3.12 of Python was used for analysis. Major data sources involve gseapy, matplotib, lifelines, pandas, scikit-learn and numpy. Two-sided statistical tests were performed with $p < 0.05$ (statistically significant).

Results

Patient Characteristics

Inclusion criteria involved 178 patients from the GSE25066 datasets who were involved in “discovery analysis “. Patients were distributed evenly into two groups, i.e., the low-risk group as well as the high-risk group, following the baseline of median-based stratification. Additionally, all the participants had “pre-treatment gene expression profiling with completed DRFS information. For validating externally, GSE58812 MFS datasets having complete expression were inculcated.

Gene Filtering and Feature Selection

Mapping (Probe-to-gene) resulted in about 16,355 genes in the discovery cohort. After filtering variance, survival analysis was carried out on 9,579 genes. 801 genes associated with DRFs ($p < 0.05$) were identified via univariate Cox regression. This set was further reduced to a 17-gene prognostic signature dataset by applying LASSO (penalised Cox Regression), indicating a significant reduction of dimensionality while keeping the prognostic relevance intact.

The gene filtering and feature selection process is summarized in Fig. 2.

Genes significantly associated with disease-free relapse-free survival in univariate Cox regression are listed in **Supplementary Table S1**.

The final 17-gene prognostic signature selected by LASSO-penalized Cox regression, along with corresponding gene coefficients, is provided in **Supplementary Table S2**.

Risk Stratification and Survival Differences

Low-risk and high-risk groups were devised utilising LASSO driving risk scores. A clearly marked statistically appropriate separation was demonstrated via Kaplan-Meier survival analysis in these two classified groups, i.e. high-risk and low-risk. The high-risk group is more prone to recurrence and other threatening outcomes with worse survival conditions as compared to patients categorised into the low-risk group. This depicts that the 17-G signature promotes stratification of chemotherapeutically treated patients of TNBC in accordance with the risk of recurrence.

Kaplan–Meier analysis demonstrated a clear separation between risk groups in the discovery cohort Fig. 3.

Cox Regression Analysis

This 17-gene signature depicted that clinically significant information was gathered in patients with TNBC who were treated via Neo-adjuvant chemotherapy. A significant interlinking is highlighted via univariate Cox regression utilising a continuous risk score in the high-risk group and low-risk group (elevated hazard and recurrence in high-risk group participants and vice versa).

External Validation

The performance of this 17-gene signature was tested and evaluated I'm GSE58812 cohort keeping the median-based cut-off as well as gene coefficients alike. This external and independent validation was backed by Kaplan-Meir analysis as well, which depicted that the rate of metastasis-free survival is high on low-risk patients and vice versa. The above-mentioned validation was further authenticated via Cox regression, and it still indicated that the "risk-score" is the staunch indicator of metastasis-free survival (hazard ratio > 1, C.I = 95% excluding unity). Validation cohort retained model discrimination, which is further reflected via "concordance-index" in accordance with prognostic utility.

Composition and Directionality of the 17-Gene Prognostic Signature

The finalised prognostic model comprised 17 genes, and the selection was made via LASSO (penalised Cox regression). Elevated recurrence was observed in genes having positive coefficients and vice versa (genes with negative coefficients were more protective).

Functional annotation in the light of published literature depicted that the genes involved in the signature influence biological processes (tumour metabolism, stress adaptation, transcriptional regulation, genome stability, and cellular homeostasis).

The coefficients and directionality of the genes included in the final prognostic signature are shown in Fig. 4.

The genes included in the final 17-gene prognostic signature and their reported biological characteristics are summarized in Table 1.

Table 1
Genes Included in the 17-Gene Prognostic Signature and Their Biological Characteristics

Gene	Prognostic Role	Reported Biological Function
AK1	Risk Promoting	Regulates adenine nucleotide balance and cellular energy metabolism, supporting high metabolic demand in proliferating cells [3].
CYP7B1	Risk Promoting	Involved in steroid and lipid metabolism; altered lipid metabolism has been linked to aggressive tumor behavior [23].
GSPT1	Risk Promoting	Regulates translation termination and cell cycle progression; overexpression has been associated with tumor proliferation [10].
MAG	Risk Promoting	Mediates cell–cell interactions and adhesion-related signaling, which can influence tumor invasion [18].
PRKAR1B	Risk Promoting	Component of cAMP signaling involved cellular stress responses and survival signaling [21].
SLC13A3	Risk Promoting	Sodium-dependent transporter involved in metabolite uptake and metabolic adaptation [16].
ANG	Protective	Regulates angiogenesis and cellular stress responses; dysregulation has been linked to tumor progression [22].
ATP6V1B1	Protective	Regulates vesicular acidification and intracellular pH homeostasis, influencing stress adaptation [14].
BLOC1S1 / RDH5	Protective	Involved in vesicle trafficking and retinoid metabolism, processes linked to cellular differentiation [12].
HOXA7	Protective	Homeobox transcription factor involved in differentiation and developmental gene regulation [20].
MAOB	Protective	Regulates oxidative metabolism and reactive oxygen species balance (9).
PCDHGA3	Protective	Protocadherin involved in cell–cell adhesion and tissue architecture maintenance [11].
REPS1	Protective	Involved in endocytosis and receptor signaling regulation [15].
SERP1	Protective	Plays a role in endoplasmic reticulum stress response and protein folding [24].
SLC9A2	Protective	Sodium–hydrogen exchanger involved in intracellular pH regulation [17].
STAG1	Protective	Component of the cohesin complex required for chromatin organization and genome stability [19].
ZNF222	Protective	Zinc finger transcription factor involved in gene regulation [25].

Biological Characteristics of the 17-Gene Signature

Functional annotation of the 17-gene prognostic signature revealed enrichment of biological processes related to cellular metabolism, stress response, transcriptional regulation, ion transport, and genome stability. Several risk-promoting genes (e.g., AK1, CYP7B1, GSPT1, SLC13A3) are involved in metabolic regulation and protein synthesis—processes commonly upregulated in aggressive tumours and linked to chemotherapy resistance in TNBC [23]. The presence of these genes suggests that metabolic adaptation and enhanced cellular stress tolerance may contribute to poor outcomes in chemotherapy-treated patients.

In contrast, many protective genes in the signature are associated with maintenance of cellular homeostasis and genomic integrity. Genes such as STAG1 and HOXA7 are involved in chromatin organisation and transcriptional control [19, 20], while SERP1, ATP6V1B1, and SLC9A2 are linked to endoplasmic reticulum stress regulation and intracellular pH balance [14, 17, 24]. These processes might provide a foundation for controlled responses at the cellular level to chemotherapeutic-induced stress, thereby significantly reducing the chances of recurrence.

Cumulatively, the signature's bidirectional nature suggests that TNBC prognosis depends widely upon the balance in aggressive-metabolic program and proliferative-program, adaptation to stress, protective pathways, process of transcriptional regulation, and genomic stability likewise. Thus, it depicts that the relevance of the identified signature (17-Gene) is beyond statistical computation and aligns well with already existing mechanisms of chemotherapeutic resistance and metastasis in TNBC.

Functional Annotation of the 17-Gene Signature

Functional enrichment analysis was performed on this 17-gene prognostic signature, which suggested over-representation of specific biological processes with reference to metabolism at the cellular level, response to stress, transcriptional-regulation, transport of ions, and genomic stability. Several risk-associated genes were associated with metabolic regulation and stress adaptation at the cellular level, while on the other hand, protective genes were significantly inculcated in transcriptional control, homeostasis maintenance at the cellular level and chromatin organisation. Most of these biological processes have a huge influence on how chemotherapy responses, the aggressiveness of tumour and most importantly progression of metastasis in TNBC (Triple negative breast cancer).

Detailed functional enrichment results from Gene Ontology biological processes, KEGG pathways, and Reactome pathway analyses are summarized in **Supplementary Tables S3–S5**.

Functional enrichment results for the 17-gene signature are illustrated in Fig. 5.

Summary of Key Findings

In summary, Survival-Modelling-Framework (LASSO-based) was utilised for core identification of "17-gene transcriptomic prognostic signature". Successful stratification of Triple-negative breast cancer patients was carried out via "signature" into high-risk groups as well as into low-risk groups (in the discovery cohort) and depicted reproducibility in prognostic performance while assessed in an

independent external dataset. Overfitting was minimized via feature selection, while biological relevance was preserved. It successfully established a validation framework (via robust discovery) for triple-negative breast cancer (TNBC) prognostic biomarker development.

Discussion

An aggressive and intensive disease pattern was identified in TNBC via the 17-gene prognostic signature, highlighted in this study. It also highlights the importance of key biological and metabolic processes (metabolic regulation,stress response at the cellular level, control of the translation process, channelling pathways for survival of the tumour in worse conditions) influenced by specified genes of high-risk scores identified via functional annotation. On the other hand, transcriptional regulation, homeostatic maintenance and genomic stability are channelled via protective genes. Moreover, Neo-adjuvant chemotherapy was received by patients categorised into discovery-cohort; therefore, residual risk (under treatment) was captured by “Signature” rather than the history (natural) of untreated disease. Thus, it suggests that the tumours which have biological features (intrinsic) are capable of progression and survival despite wide exposure to this cytotoxic therapy. These findings align well with the predetermined traits of TNBC, involving genomic instability, elevated stress tolerance and metabolic plasticity implicated widely in metastatic progression and chemotherapeutic resistance [1, 4]. Although there is a limitation in overlapping direct gene with TNBC signatures reported previously (Convergence of biological processes highlighted by signatures and Pathways of intensive TNBC pathways) [2, 3]. Overall, these findings support the fact that adaptation to stress, metabolic alterations, and control of transcriptional processes highlight patients (the TNBC patient subset) who are given standard chemotherapeutic treatment but still exhibit poor outcomes. It depicts that the findings have great potential for improvement in prognostic stratification and futuristic research in the translational process.

Another Foremost aspect of this study is the utilisation of “independent-external-validation” across variable microarray platforms. Most of the prognostic studies depend wholly upon internal validation, which can lead towards overly optimistic performance estimation. This study also suggests that the “fixed gene signature” and coefficients are validated in a GEO dataset (independent) generated on a variable platform. This study also provides staunch evidence in support of model reproducibility and significantly reduces the risk of “overfitting-related bias” [7]. Similarly, Functional-enrichment-analysis backend the biological plausibility of this identified signature [1, 23]. Although not all enrichment outcomes remained “significant” after conducting multiple testing corrections, these results are expected, given that the relatively smaller size of the gene set has been reported in very similar “transcriptomic prognostic studies” [3, 7].

The findings of this study suggested that a carefully devised and validated (externally) transcriptomic signature can provide productive prognostic information in triple-negative breast cancer (TNBC). The suggested framework may serve as a baseline for future research targeting refining of prognostic tools and the integration of molecular risk stratification in the clinical decision-making process.

Limitations

Major aspects of this study need to be interpreted according to its novelty as well as its scope. This analysis was carried out utilising publicly available retrospective datasets, which highlight independent cohorts while enabling the discovery of focused prognostic signals. The major drawback of this evaluation is “variability” in terms of sample processing.

Secondly, this study is a discovery-level analysis designed intentionally to evaluate risk-stratification in TNBC patients receiving chemotherapy. The aim was to highlight outcomes of under-treated patients and the association of intrinsic molecular programs with worsening of the condition. This intervention proposed a promising foundation for future clinical studies in developing integrated models.

Thirdly, this study is performed utilising “microarray-based gene expression data”, which made it easy to validate across external platforms. No doubt, for the assessment of generalizability, validating RNA sequencing-based datasets will always remain important. The major limitation of this study is the validation of identified genes via statistical and computational modelling; for clarity in therapeutic and functional roles, experimental validation is essential, which is not in the scope of this study.

Conclusion

In conclusion, this study helped in the identification of an externally validated 17-gene (transcriptomic) prognostic signature for TNBC (triple-negative breast cancer) utilising a robust “survival- modelling-framework”. This signature efficiently and effectively stratified the patients into high-risk groups as well as the low-risk groups, in both the discovery and the independent “validation-cohorts”, depicting reproducibility in prognostic performance. By successful integration of rigorous-feature-selection, external validation performed independently, and functional interpretation, this study helps in building up a reliable, trusted and biologically interpretable “prognostic model” for triple-negative breast cancer (TNBC). Future studies in prospective cohorts incorporation, experimental validation and clinical covariates promise further assistance in the clinical utilisation of this prognostic signature and its significant role in “personalised risk-stratification” in patients with TNBC.

Declarations

Conflict of Interest

The authors declare that they have no competing interests.

References

1. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L (2016) Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol* 13:674–690.

https://doi.org/10.1038/nrclinonc.2016.66

2. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* 121:2750–2767. https://doi.org/10.1172/JCI45014
3. Liu J, et al (2018) A prognostic gene signature for triple-negative breast cancer. *Oncotarget* 9:7205–7216. https://doi.org/10.18632/oncotarget.23931
4. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, Díez M, Viladot M, Arance A, Muñoz M (2015) Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* 24(Suppl 2):S26–S35. https://doi.org/10.1016/j.breast.2015.07.008
5. Cox DR (1972) Regression models and life-tables. *J R Stat Soc Series B* 34:187–220
6. Tibshirani R (1997) The lasso method for variable selection in the Cox model. *Stat Med* 16:385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3
7. Wu J, et al (2020) Identification of a prognostic gene signature in triple-negative breast cancer. *Front Oncol* 10:870. https://doi.org/10.3389/fonc.2020.00870
8. Kuleshov MV, et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server. *Nucleic Acids Res* 44:W90–W97. https://doi.org/10.1093/nar/gkw377
9. Bortolato M, et al (2008) Monoamine oxidase A and B in neurological disorders. *CNS Drugs* 22:189–210
10. Chen C, et al (2021) GSPT1 promotes cell proliferation and tumor progression. *Cell Death Dis* 12:728. https://doi.org/10.1038/s41419-021-04066-3
11. Chen WV, Maniatis T (2013) Protocadherin-mediated cell recognition. *Cell* 154:269–281. https://doi.org/10.1016/j.cell.2013.06.018
12. Dell'Angelica EC (2004) The BLOC complexes and organelle biogenesis. *Trends Cell Biol* 14:512–518. https://doi.org/10.1016/j.tcb.2004.08.006
13. Dzeja PP, Terzic A (2009) Adenylate kinase and energy metabolism. *J Exp Biol* 212:1–9. https://doi.org/10.1242/jeb.023275
14. Forgac M (2007) Vacuolar ATPases: rotary proton pumps in physiology and pathophysiology. *Nat Rev Mol Cell Biol* 8:917–929. https://doi.org/10.1038/nrm2272
15. Ikeda M, et al (2001) Regulation of receptor endocytosis by REPS1. *Mol Biol Cell* 12:2159–2170. https://doi.org/10.1091/mbc.12.7.2159
16. Markovich D, Murer H (2004) The SLC13 gene family of sodium sulphate/carboxylate cotransporters. *Pflugers Arch* 447:594–602. https://doi.org/10.1007/s00424-003-1090-7
17. Orlowski J, Grinstein S (2004) Diversity of the mammalian sodium/proton exchanger SLC9 gene family. *Am J Physiol Cell Physiol* 286:C2–C12. https://doi.org/10.1152/ajpcell.00321.2003
18. Quarles RH (2007) Myelin-associated glycoprotein (MAG): past, present and beyond. *J Neurochem* 100:1431–1448. https://doi.org/10.1111/j.1471-4159.2006.04319.x

19. Remeseiro S, et al (2012) Cohesin and chromatin organization. *Genes Dev* 26:952–967. https://doi.org/10.1101/gad.193540.112
20. Shah N, Sukumar S (2010) The Hox genes and their roles in oncogenesis. *Nat Rev Cancer* 10:361–371. https://doi.org/10.1038/nrc2826
21. Taylor SS, et al (2013) PKA signaling specificity. *Biochim Biophys Acta* 1834:1271–1278. https://doi.org/10.1016/j.bbapap.2013.01.019
22. Tello-Montoliu A, et al (2006) Angiogenin and tumor angiogenesis. *Cancer Res* 66:4473–4480. https://doi.org/10.1158/0008-5472.CAN-05-3474
23. Wu X, et al (2019) Lipid metabolism reprogramming in cancer. *Front Oncol* 9:125. https://doi.org/10.3389/fonc.2019.00125
24. Yamaguchi A, et al (1999) SERP1, a novel endoplasmic reticulum stress-associated protein. *J Biol Chem* 274:3753–3758. https://doi.org/10.1074/jbc.274.6.3753
25. Liu Y, et al (2014) Zinc finger proteins in transcriptional regulation. *J Mol Biol* 426:818–832. https://doi.org/10.1016/j.jmb.2013.11.012.

Figures

Figure 1

Study design and analysis workflow

Schematic overview of the study design and analytical workflow. Gene expression and clinical survival data were obtained from the GEO dataset GSE25066. Triple-negative breast cancer (TNBC) samples were selected based on ER-negative, PR-negative, and HER2-negative status. Probe-level microarray data were mapped to gene symbols, followed by variance filtering to remove low-variability genes. Prognostic genes were screened using univariate Cox proportional hazards regression, followed by feature selection using LASSO-penalised Cox regression to derive a parsimonious prognostic gene signature. A multigene risk score was constructed, patients were stratified into high- and low-risk groups, and survival

differences were evaluated using Kaplan–Meier analysis. Functional annotation was performed to assess the biological relevance of the selected genes.



Figure 2

Gene filtering and feature selection workflow

Gene filtering and feature selection across the analytical pipeline. Probe-to-gene mapping resulted in approximately 16,355 gene-level features. Variance filtering reduced the feature space to 9,579 genes. Univariate Cox regression identified 801 genes significantly associated with disease-free survival ($p < 0.05$). Subsequent LASSO-penalised Cox regression reduced this set to a 17-gene prognostic signature.

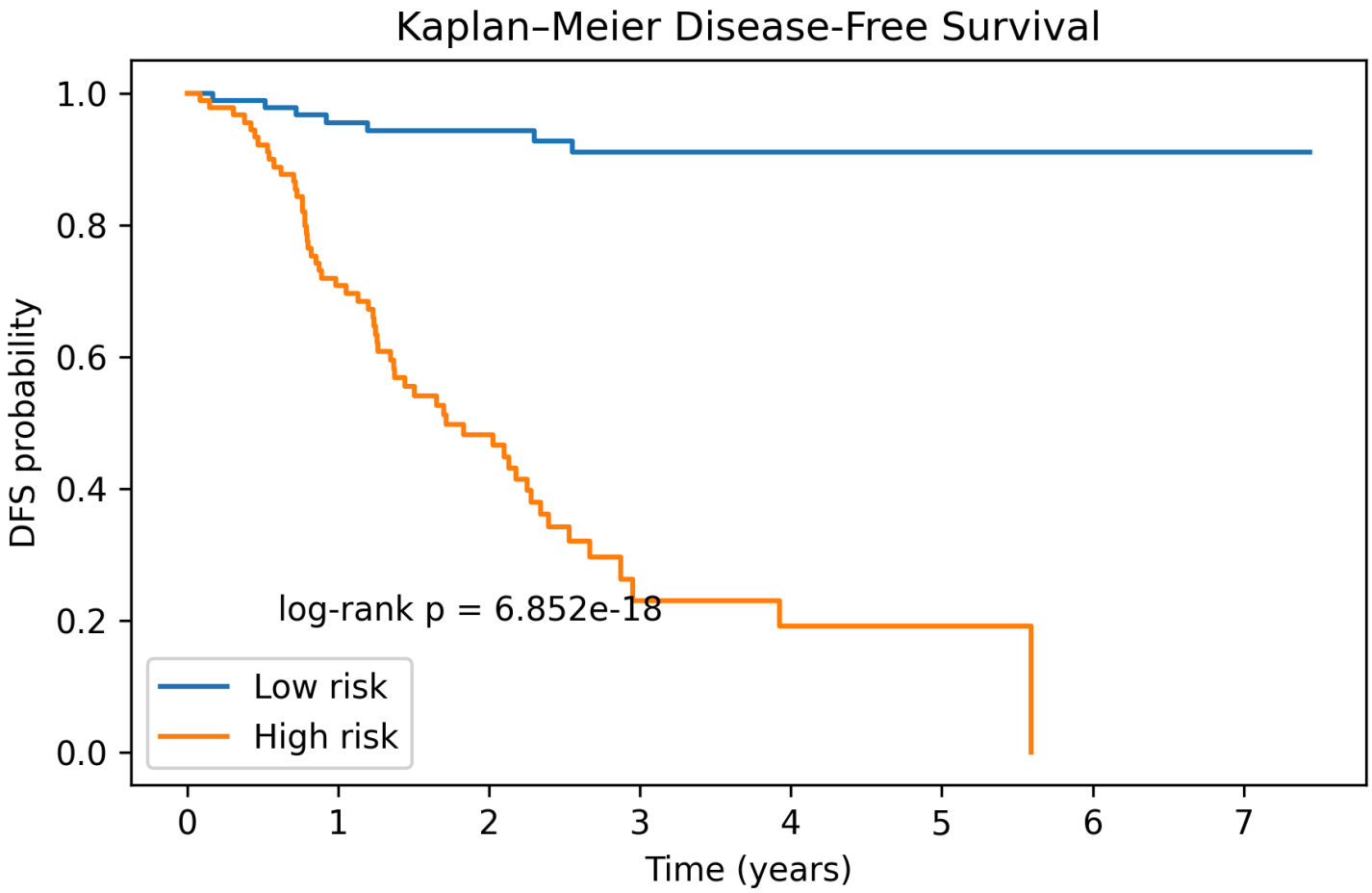


Figure 3

Kaplan–Meier disease-free survival analysis

Kaplan–Meier disease-free survival curves for TNBC patients stratified by the LASSO-derived prognostic risk score in the discovery cohort. Patients were divided into high-risk and low-risk groups using the median risk score as the cutoff. Survival differences were assessed using the log-rank test.

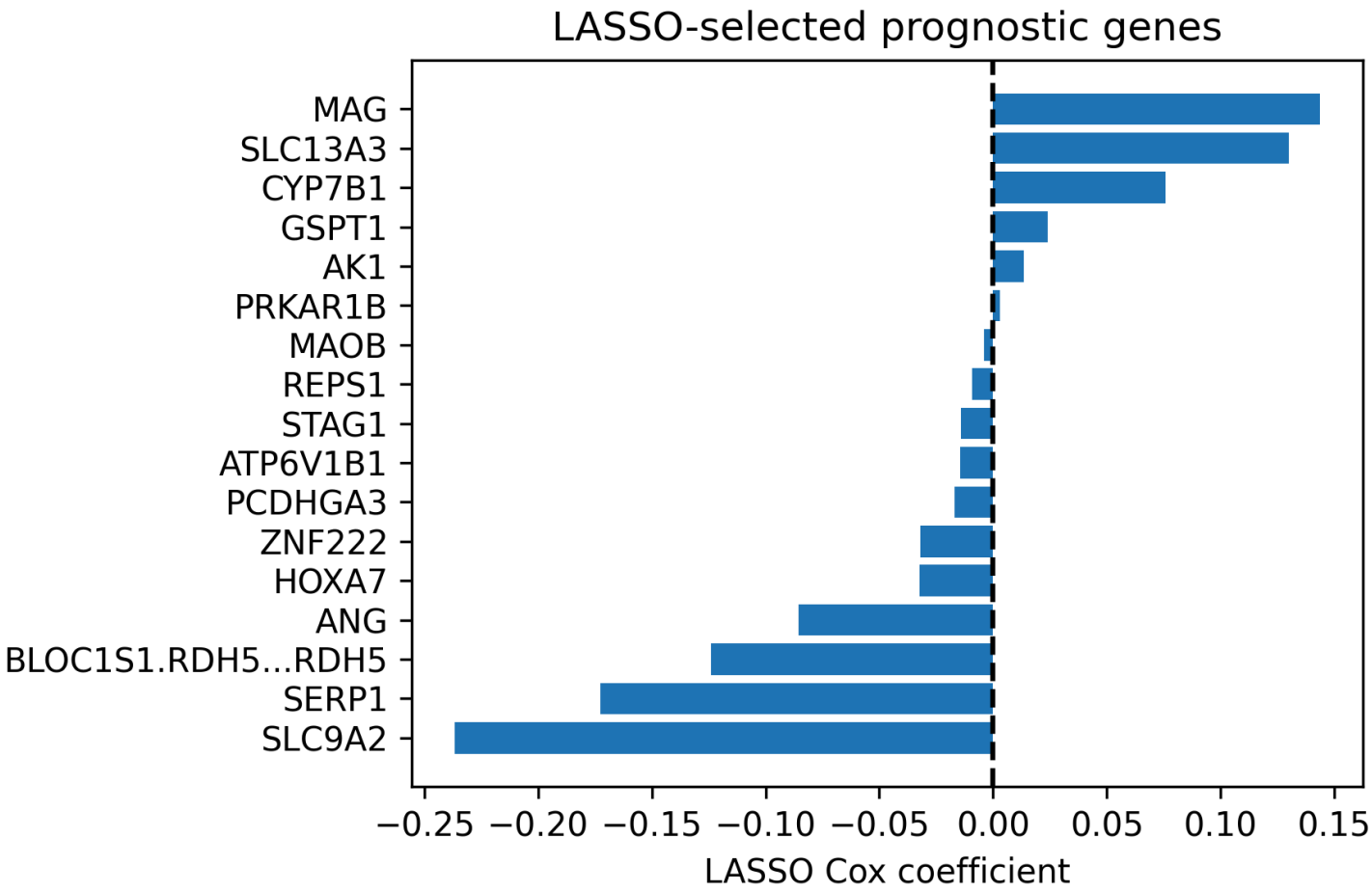


Figure 4

LASSO Cox–selected prognostic gene signature

Coefficients of the 17 genes selected by LASSO-penalised Cox regression. Bars represent the magnitude and direction of each gene's contribution to the prognostic risk score, with positive coefficients indicating increased risk and negative coefficients indicating protective effects.

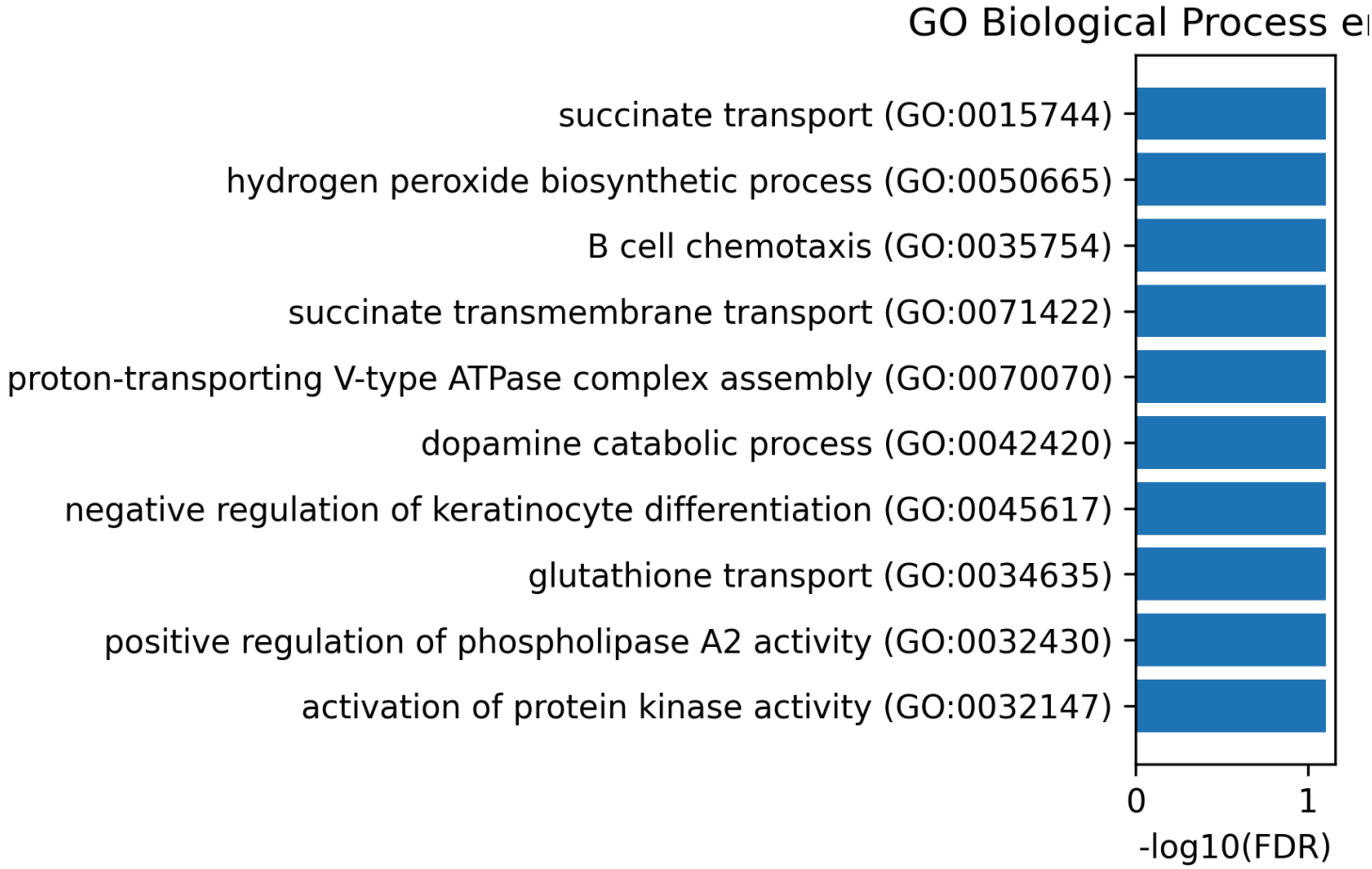


Figure 5

Functional enrichment analysis of the LASSO-derived gene signature

Gene Ontology (GO) biological process enrichment analysis of the 17-gene LASSO-derived prognostic signature. Enriched terms are ranked by statistical significance (−log10 adjusted p-value).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.csv
- SupplementaryTableS2.csv
- SupplementaryTableS3.csv
- SupplementaryTableS4.csv
- SupplementaryTableS5.csv