Journal of Ovarian Research

RESEARCH

Open Access

Development and validation of gene expression-based signature for high-grade serous ovarian cancer

Ieva Vaicekauskaitė[1,2], Julius Juodakis[3], Paulina Kazlauskaitė[1], Rūta Čiurlienė[4], Giedrė Smailytė[1], Juozas Rimantas Lazutka[2] and Rasa Sabaliauskaitė[1,2*]

Abstract

Background High-grade serous ovarian cancer (HGSOC) is the second most lethal gynecologic malignancy, often diagnosed at a late stage due to the lack of reliable early detection strategies. Currently, there are no specific diagnostic or prognostic biomarkers for ovarian cancer (OC). Thus, there is a great need for novel validated biomarkers for OC diagnosis.

Methods A two-step machine learning approach was employed to identify potential HGSOC biomarkers in The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression project (GTEx) ovarian cohorts. The selected genes were then validated in an external, clinically annotated tissue cohort of 65 samples from OC patients, treated in Lithuania, to assess biomarker performance in separating HGSOC from benign gynecologic conditions and predict overall survival.

Results A ten-gene signature (*EXO1*, *RAD50*, *PPT2*, *LUC7L2*, *PKP3*, *CDCA5*, *ZFPL1*, *VPS33B*, *GRB7*, and *TCEAL4*) was selected for further analysis in the external cohort. Expression of each of the ten genes was highly indicative of HGSOC compared to benign gynecologic conditions ($p \leq 0.030$) and separated these groups with *GRB7* expression reaching the highest area under the ROC curve (AUC) of 0.986. *RAD50*, *VPS33B*, and *ZFPL1* expression also correlated with stage in HGSOC cases ($p < 0.042$), while *TCEAL4* expression was associated with tumor grade ($p = 0.038$). The 10-gene signature was also predictive of 5-year survival in the OC tissue cohort (AUC = 0.815).

Conclusions The ten selected gene expression biomarkers could be useful for HGSOC diagnosis and prognosis; however, further investigations in their prediction of OC patients' survival are still required.

Keywords High-grade serous ovarian carcinoma, Ovarian cancer, Gene expression, Diagnostic model

*Correspondence:
Rasa Sabaliauskaitė
rasa.sabaliauskaite@gmc.vu.lt
[1]National Cancer Institute, Vilnius, Lithuania
[2]Institute of Biosciences, Life Sciences Center, Vilnius University, Vilnius, Lithuania
[3]Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[4]Vilnius University Hospital Santaros Clinics, National Cancer Centre, Vilnius, Lithuania

Introduction

Ovarian cancer (OC) is the second deadliest and third most common gynecologic malignancy in the world. About half (48%) of OC cases are high-grade serous ovarian carcinomas (HGSOCs), which majority of deaths are attributable to, as it is primarily detected at stage III (51% of cases) or IV (29%) [1]. Currently, the late diagnosis and high death rate are linked to a lack of specific diagnostic and prognostic biomarkers for OC. The main OC biomarkers currently used in clinical practice are CA125 and HE4; however, both biomarkers are serum proteins that lack specificity and are not recommended for use for diagnostics or prognostics [2]. Attempts to increase the accuracy of protein biomarkers by combining them with clinical features also do not provide the desired sensitivity or specificity for a screening assay [3], thus, new biomarkers for OC clinical care are in high demand.

Genetic biomarkers can offer a more precise and personalized approach to the OC detection and clinical management as compared to traditional biomarkers, such as CA125 and HE4, as genetic changes can be identified earlier and thus improve diagnostic accuracy and predict treatment outcomes. Despite the lack of diagnostic tests based on genetic changes, some genetic alterations such as mutations are currently used for OC risk assessment and treatment prediction as patients with family history of breast or ovarian cancers can undergo hereditary cancer genetic testing, stratifying patients with increased risk of cancer based on mutations in DNA repair genes such as *BRCA1*, *BRCA2*, and to lesser extent *RAD51C* or *RAD51*, *BRIP1*, *PALB2*, coding for homologous recombination repair proteins, as well as mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6* and *PMS2*. The benefit of genetic testing is not only risk analysis, but also treatment sensitivity prediction, as particularly homologous recombination-deficient cancers can be treated with poly (ADP-ribose) polymerase (PARP) inhibitors [4]. Based on the success of mutation biomarkers, it is not unreasonable to assume that other genetic biomarkers, such as gene expression, could be useful for OC diagnosis and prognosis, as gene expression often reflects the downstream effects of both gene mutations and other molecular changes, providing insight into tumorigenesis.

Gene expression profiling is a valuable technique used for identification of promising prognostic biomarkers and their combinations. Currently, The Cancer Genome Atlas (TCGA) has one of the largest datasets of OC tissue gene expression and clinical information datasets. The initial genetic analysis of the TCGA stratified OC cases via clustering algorithms into mesenchymal, immunoreactive, differentiated and proliferative subtypes, however this stratification did not offer insights into patients' survival and was not intended for diagnostic and predictive purposes [5]. A few other studies have attempted to develop gene signatures for survival prediction, often focusing on cancer mechanisms, such as cell death [6], hypoxia [7], and epithelial-mesenchymal transition [8]. However, OC is a highly heterogeneous disease, and gene expression regulation is a complex process with multiple interacting genes, thus a single gene or even a single cellular mechanism is unlikely to predict every OC case or its outcome [8]. Moreover, biomarkers discovered in TCGA and similar large databases are rarely validated in external cohorts, limiting their adoption in real-world applications.

Herein, we propose a diverse 10-gene panel aimed at OC diagnosis and prognosis. The mRNA biomarkers were selected using elastic net and LASSO-Cox (proportional hazards) regression models, then the biomarker expression was validated in an external ovarian tissue cohort using RT-qPCR. We observed that the markers were able to predict OC and patient survival.

Methods

TCGA and GTEx cohort analysis

The ovarian cancer transcriptomic dataset was obtained from TCGA. The TCGA ovarian cancer (TCGA-OV) dataset (RNA-seq STAR-counts data), which included 416 tumor samples together with clinical data, was downloaded using the TCGAbiolinks R package (version 2.29.6), while clinical data was downloaded via UCSCXenaTools (version 1.4.8). The normal ovarian tissue transcriptomic dataset was obtained from the Genotype-Tissue Expression (GTEx) portal, which included data from 180 normal ovarian tissue samples. The GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2) RNA-seq gene read counts were downloaded from https://www.gtexportal.org/home/datasets accessed on 2023-08-21. After combining the datasets, expression data were available for 56,156 genes; from these, only the protein-coding genes were selected for further analysis (19,197 genes) via biomaRt (version 2.56.1) package. The raw RNA counts were normalized via GDCRNATools (version 1.20.1) and voom normalization from the limma package (version 3.56.2). Genes present in only one of the datasets were excluded in this step, leaving 13,681 genes suitable for further selection. During the normalization step, one TCGA case was removed due to outlier values. The full dataset was then split into training data (489 samples, of them 153 GTEx and 336 TCGA) and test data (106 samples, of them 27 GTEx and 79 TCGA) using an 80:20 split ratio.

To find candidate OC biomarkers, the training data was first analyzed using elastic net logistic regression from the glmnet (version 4.1-8) package with alpha = 0.5. Data source (GTEx vs. TCGA-OV) was the outcome, and the normalized expression levels were the predictors. Prediction error was evaluated using internal cross-validation,

and the penalty that leads to the lowest error ("lambda.min" in glmnet) was chosen, resulting in 214 genes selected. To gain insight into which biological processes the selected genes were involved in, we performed gene ontology (GO) enrichment analysis using clusterProfiler (v4.8.3), selecting enriched terms with adjusted p-values < 0.05.

To further narrow down the candidate set, a second, survival, regression model was applied to the training data from the TCGA-OV cohort. The outcome was time from diagnosis to death, in days; patients lost to follow-up were censored at the time of last observation. The predictors were the normalized counts of the 214 genes. A LASSO-regularized Cox regression was applied to this, again implemented in the glmnet package. Various values of the penalty (lambda) were tested, and the value that resulted in a manageable set of candidates was selected (namely, lambda = 0.088 leading to 10 genes). These ten genes were selected for further experimental validation.

A polygenic expression risk model was created using formula: $\text{Risk score} = \sum_i^{10} gene_i * \beta_i$, where $gene_i$ indicates the normalized expression of each of the 10 selected genes, and β represents the corresponding coefficients derived from the LASSO-Cox model. All data analysis and visualizations were performed in R (version 4.3.1, R Foundation for Statistical Computing, Vienna, Austria). All code used in the analysis is available at https://github.com/ieva-vaic/TCGA-OV-RISK-PROJECT.

Ovarian tissue samples

The selected candidate genes were validated on a cohort of 65 patients with suspected OC who underwent the removal of ovaries and fallopian tubes at the National Cancer Institute of Lithuania between 2018 and 2023. The regional bioethics committee approved the study (No. 158200–18/5–988–539 amendment No. 2). All patients have given informed consent. Samples were obtained from the removed ovarian tissues during the procedure and immediately preserved at − 80 °C for future analysis.

Of the 65 patient samples, nine were gynecologic tissues samples with benign conditions (benign cysts, endometriosis and one case of preventative ovary and fallopian tube removal due to *BRCA2* germline mutation), while 56 were gynecologic malignant tumors. The malignant gynecologic tumor groups were made up of 42 HGSOC samples and 14 other, non-HGSOC, gynecologic tumors (two mucinous type ovarian cancers, one clear cell ovarian carcinoma, one low-grade serous ovarian carcinoma, one endometrioid type ovarian carcinoma, one granulosa cell tumor of the ovary, three synchronous primary endometrioid endometrial and ovarian cancer cases, and five cases with borderline ovarian tumors). Clinical features are described in detail in Table 1.

Validation of gene expression biomarkers in tissue samples

RNA from ovarian tissue samples was extracted with TRIzol reagent (Invitrogen, TFS, Carlsbad, CA, USA) using the manufacturer's instructions. The final RNA was air-dried and dissolved in nuclease-free water (Thermo scientific, Vilnius, Lithuania). The purity and quantity of RNA determined using Nanodrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The nucleic acid samples were stored at − 80 °C until the cDNA synthesis step.

The tissue RNA samples were used for cDNA synthesis with Maxima First Strand cDNA Synthesis Kit for RT-qPCR with dsDNase (ThermoScientific, TFS, Vilnius, Lithuania) and a ProFlex PCR System (Applied Biosystems, TFS, Singapore), following the manufacturer's instructions. Expression of the 10 gene transcripts, selected in the TCGA models, was determined by quantitative PCR (qPCR), which was performed using Maxima SYBR Green qPCR Master Mix (2X) kit (ThermoScientific, TFS, Vilnius, Lithuania) and Metabion primers (Metabion, Planegg, Germany), on a QuantStudio 5 Real-Time PCR System (Applied Biosystems, TFS, Singapore) following the manufacturer's protocols. Primer sequences are provided in Supplementary Table 1. The initial Ct values were collected using QuantStudio Design & Analysis Software v1.4.3 (Applied Biosystems). Gene expression was normalized to *GAPDH* expression with ΔCt method.

Statistical analysis

When analyzing gene expression associations with OC or clinical features, Mann-Whitney, Student's t, or Welch's t tests were applied as appropriate. Associations between three or more groups were analyzed via ANOVA or Kruskal-Wallis tests with the post-hoc analysis of either Tukey HSD or Dunn tests as appropriate. Receiver operating characteristic (ROC) tests from the pROC package (version 1.18.5) were applied to analyze the performance of biomarkers or their combinations. Multiple gene expression biomarkers were combined using logistic regression or Cox regression for the survival analysis using the glmnet package (version 4.1-8). Kaplan-Meier survival curves and Cox regression from the survival (version 3.5.7) package were used to estimate the biomarker ability to predict survival time. Time-dependent ROC curves at 5 years were generated to estimate the biomarker's predictive power with the survivalROC (version 1.0.3.1) package. The results were considered statistically significant when adjusted $p \leq 0.05$.

Results

Selection of gene expression biomarkers

Based on the public-data analysis of TCGA-OV cases and GTEx controls, an elastic-net model identified 214 genes

Table 1 Clinical features of the ovarian tissue study cohort

Clinical features	Ovarian cancer	Benign ovarian tumor tissues	All ovarian tumors	*p* value
n	56	9	65	
Histological group				
Type II OC (HGSOC)	42 (75.00%)		42 (64.62%)	
Other OC	14 (25.00%)		14 (21.54%)	
Benign		9 (100.00%)	9 (13.85%)	
Average age at diagnosis, years (± SD)	59.16 (± 9.69)	53.67 (± 9.33)	58.40 (± 9.76)	0.13, t test
CA125 concentration at diagnosis				
Norm (< 35 U/mL)	1 (1.79%)	4 (44.44%)	5 (7.69%)	< 0.001, Fisher's test
CA125 increase (> 35 U/mL)	47 (83.93%)	3 (33.33%)	50 (76.92%)	
NA[1]	8 (14.29%)	2 (22.22%)	10 (15.38%)	
Grade group				
G1	6 (10.71%)		6 (9.23%)	
G3	42 (75.00%)		42 (64.62%)	
NA	8 (14.29%)	9 (100%)	17 (26.15%)	
FIGO stage				
I	9 (16.07%)		9 (13.85%)	
II	3 (5.36%)		3 (4.62%)	
III	30 (53.57%)		30 (46.15%)	
IV	14 (25.00%)		14 (21.54%)	
NA		9 (100.00%)	9 (13.85%)	
Median overall survival, months (min–max)	47 (1 – 82)	70 (59 – 70)	47 (1 – 82)	0.14, t test
Survival status at the time of study				
Deceased	19 (33.93%)		19 (29.23%)	0.54, Fisher's test
Alive	36 (64.28%)	3 (33.33%)	39 (60.00%)	
NA	1 (1.79%)	6 (66.67%)	7 (10.77%)	

[1] NA – data not available

associated with case status. In the following step, these gene expression biomarkers were narrowed down to 10 genes associated with survival. The full list of selected genes is available in Supplementary Fig. 1. The selected genes are primarily involved in mitosis and cell-cell junction organization, according to a GO enrichment analysis (Supplementary Fig. 2).

The training dataset was filtered for the selected biomarkers and TCGA-OV samples that had survival data, then a LASSO-Cox model was applied to find genes that could predict the patient survival. To end up with a short list of biomarkers for experimental validation, the LASSO penalty value was chosen to select the top 10 genes (Fig. 1A). The final list of selected biomarkers and their functions (retrieved from the HUGO gene nomenclature committee (HGNC) database) is listed in Table 2. The selected genes are involved in various essential cell mechanisms such as DNA reparation (*EXO1* and *RAD50*), cell replication (*CDCA5*), lysosome activity (*VPS33B* and *PPT2*), gene expression regulation (*LUC7L2*, *ZFPL1*, *TCEAL4*), cell-to-cell signal transduction (*GRB7*, *PKP3*), and all involved in cancer development and progression.

All selected genes showed significant changes ($p < 0.001$) in mRNA expression when compared to normal ovarian tissues in both training and test cohorts. Notably, the same pattern of change – 4 genes upregulated and 6 genes downregulated in OC cases – was replicated in both training (Fig. 1B) and test (Fig. 2) cohorts. The greatest increase in expression was observed in *PKP3* expression (train cohort $\log_2FC = 7.63$, test cohort $\log_2FC = 7.33$), while the greatest downregulation was observed in *RAD50* (train cohort $\log_2FC = -5.03$, test cohort $\log_2FC = -5.11$) expression (Fig. 1B and Fig. 2). The final 10 genes were also predictive of the overall survival rate, with *RAD50*, *PKP3* and *GRB7* gene expression showing positive coefficients reflecting predictive impact on shorter overall survival, and the rest showing positive impact on overall survival (Fig. 1C).

Combining gene expression and coefficients into a single risk score showed significant prediction ($p < 0.001$) of overall survival in the train dataset (Fig. 1D). Although a combination of the 10 gene expression in the train cohort showed great correlation with overall survival, same correlation was not found in the smaller test cohort ($p > 0.050$). The 5-year overall survival prognosis of the 10 gene combination did reach an AUC of 0.68 which outperformed the best single prognostic biomarker, *GRB7* with an AUC of 0.61. However, in the test cohort, the 10-gene combination did not outperform the single biomarkers, and the best prediction of the 5-year survival was achieved by *ZFPL1* gene expression (AUC = 0.64) (Supplementary Fig. 3).

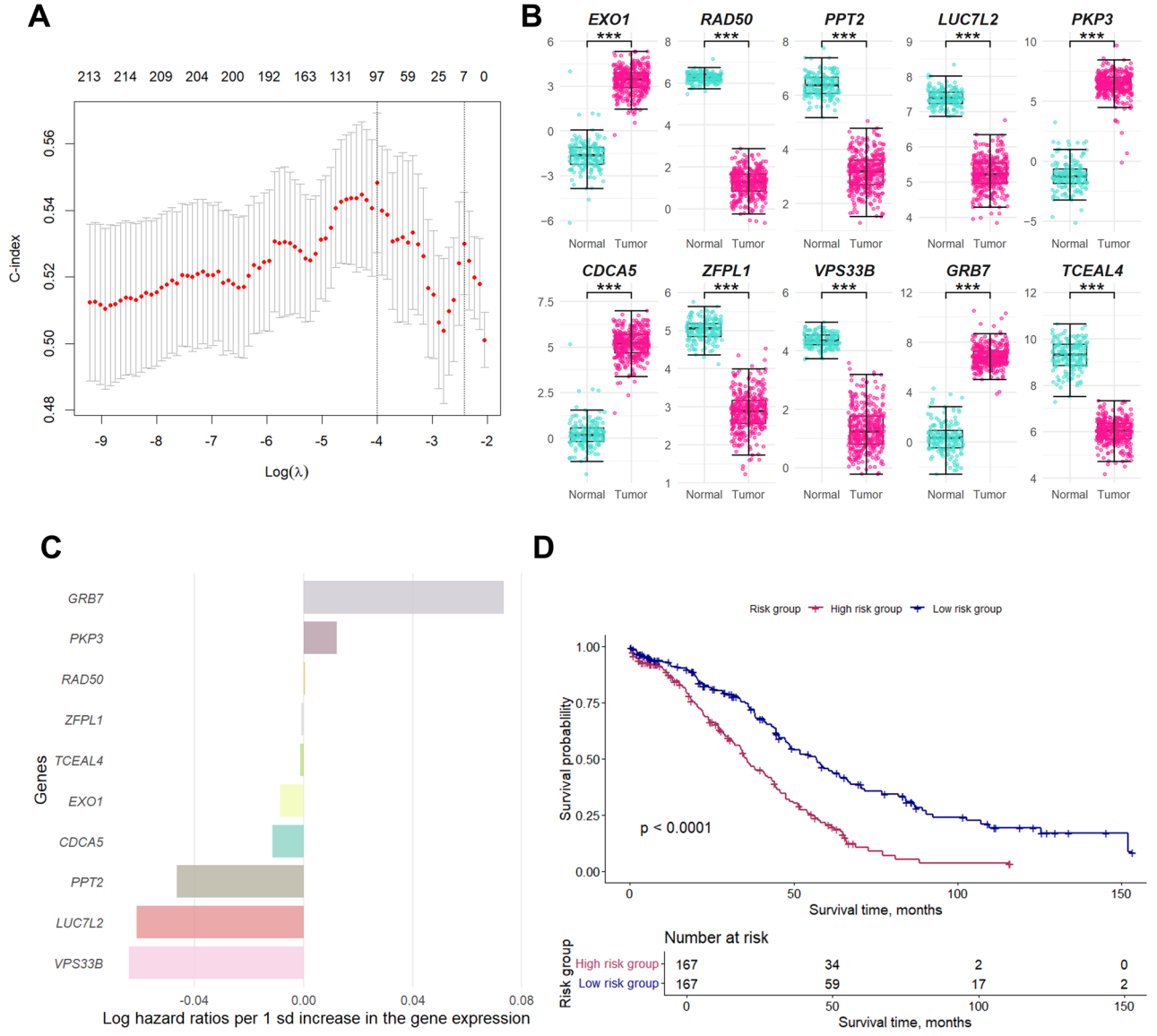


Fig. 1 Selection of the genes in the training data (153 normal (GTEx) and 336 tumor (TCGA-OV) samples). **A** Cross-validation plot from the LASSO-Cox regression. Y-axis shows the accuracy of the classification as mean +/- SE, given different values of the penalty parameter (lambda). Two optimal values of lambda are marked with dashed lines (see the function cv.glmnet documentation for details). **B** Expression of the 10 selected genes in GTEx and TCGA-OV train cohorts. **C** Coefficient estimates for the 10 selected genes in Cox model (i.e. log hazard ratios per 1 SD increase in the gene expression). **D** Risk model – combination of selected genes and coefficients in predicting overall survival

Performance of selected gene expression biomarkers in ovarian tissues

The selected genes were then examined in an external tissue cohort comprised of 42 HGSOC, 14 other OC, and 9 benign ovarian tissues using RT-qPCR. Despite the small number of benign samples, all 10 gene expressions were significantly altered in HGSOC cases compared to benign tissues ($p \leq 0.030$) (Fig. 3). Importantly, the dysregulation of each gene expression matched directions with the TCGA and GTEx test and train cohorts; for example, *EXO1* expression was significantly increased in HGSOC when compared to benign ovarian tumors, both in our validation cohort and in the public dataset. The greatest difference in gene expression between HGSOC and benign tumor tissues was found for *TCEAL4* expression ($p < 0.001$, $\log_2 FC = -4.24$). Significant difference was also observed in *TCEAL4* gene expression in HGSOC cases compared to other malignant gynecologic tumors ($p < 0.001$, $\log_2 FC = -0.75$). Half of the selected genes also had significantly altered expression in other malignant gynecologic tumors compared to benign ovarian tissues (*GRB7*, *PKP3*, *RAD50*, *TCEAL4*, and *ZFPL1*, $p \leq 0.014$), again matching the dysregulation directions from the train and testing cohorts (Fig. 3).

Table 2 Names and functions of the 10 genes selected as candidate ovarian cancer biomarkers

Gene	Gene name	Gene function
EXO1	Exonuclease 1	5'->3' exonuclease and endonuclease, involved in DNA replication and reparation
RAD50	RAD50 double strand break repair protein	Double strand break repair protein
PPT2	Palmitoyl-protein thioesterase 2	Lysosome thioesterase
LUC7L2	LUC7 like 2, pre-mRNA splicing factor	Part of the spliceosome
PKP3	Plakophilin 3	Involved with connecting cadherins to cytoskeleton
CDCA5	Cell division cycle associated 5	Soronin, involved in sister chromatid cohesion
ZFPL1	Zinc finger protein like 1	Transcription factor
VPS33B	VPS33B late endosome and lysosome associated	Involved in protein sorting
GRB7	Growth factor receptor bound protein 7	Adaptor protein that interacts with receptor tyrosine kinases
TCEAL4	Transcription elongation factor A like 4	Transcription elongation protein

Comparing the selected gene expression to the state-of-art OC biomarker, CA125 status, we found that two of the selected gene expression biomarkers, *LUC7L2* and *TCEAL4* were also downregulated in OC cases with increased (above the clinical threshold of 35 U/mL) CA125 serum biomarker concentrations at diagnosis (*LUC7L2* $p = 0.03$, $\log_2\text{FC} = -1.08$, *TCEAL4* $p = 0.05$, $\log_2\text{FC} = -3.36$) (Supplementary Fig. 4).

In order to analyze the diagnostic potential of the selected genes, ROC analysis was applied. All 10 gene expression levels showed a good separation of the HGSOC and benign ovarian tumor groups, with the lowest AUC = 0.795 for *EXO1*, and the highest AUC value achieved by *GRB7* gene expression (AUC = 0.986, sensitivity = 0.946, and specificity of 1.00). *TCEAL4* expression also showed high separation of HGSOC and benign cases (AUC = 0.984, sensitivity = 0.952, and specificity of 1.00), with slightly higher sensitivity than *GRB7* (Fig. 4).

When comparing the ability to predict HGSOC cases between the proposed expression markers and the clinical biomarker CA125, all of the gene expression biomarkers achieved higher AUCs; however, due to the small sample size, limiting statistical power, *TCEAL4* and *GRB7* were the only two genes for which the improvement was statistically significant ($p < 0.050$) when comparing the ROC curves (Fig. 4).

In order to see if the selected genes could also differentiate OC histotypes, ROC analysis between HGSOC and other OC groups was performed. All of the selected genes showed higher AUC values when separating HGSOC from benign ovarian tumors, rather than when separating HSGOC from other types of OC. *TCEAL4* was the best predictor of HGSOC cases vs. other OC group (AUC = 0.799, sensitivity = 0.690, specificity = 0.857), with no other genes reaching AUC of 0.800 (Supplementary Fig. 5).

We next combined the 10 gene expression biomarkers together to see if that led to better prediction. When combining biomarkers together, all possible combinations of the 10 selected genes were explored. Many combinations were able to perfectly separate (AUC = 1) the benign cases from HGSOC, including 8 different combinations of gene pairs (6 of them with *GRB7*), and 53 combinations of gene trios. Similarly, separation of all cancer cases (both HGSOC and other types of OC) from benign tumors of AUC = 1.00 was achieved by 7 gene pairs and 40 trios, showing strong diagnostic power of the biomarker combinations; however, given the rather small sample size, these results should be regarded with caution and further validated in larger external cohorts. The best separation of HGSOC vs. other gynecologic cancer samples was achieved by combining *RAD50*, *PKP3*, *CDCA5*, *ZFPL1*, *VPS33B* and *TCEAL4* expression reaching AUC = 0.935, sensitivity = 0.778 and specificity of 1.000, outperforming other smaller or larger gene panels or combination of all 10 biomarkers which reached AUC = 0.869, sensitivity = 0.656 and specificity of 1.000 (Supplementary file 2), indicating that larger models do not necessarily improve prediction.

Selected gene association with clinical features

We investigated the associations between selected gene expression and clinical features to better understand their significance as clinical biomarkers. *RAD50*, *VPS33B* and *GRB7* expression were predictive of FIGO stage in the HGSOC subgroup ($p \leq 0.006$), with *GRB7* significantly decreased in stage III and stage IV cases compared to stage II ($p = 0.026$, $\log_2\text{FC} = -2.08$ and $p = 0.004$, $\log_2\text{FC} = -2.77$ respectively), while *RAD50* and *VPS33B* showed negative correlation between all three stages ($p \leq 0.040$) and *ZFPL1* expression also showed a tendency of reduced expression in stage IV cases compared to stage II ($p = 0.066$, $\log_2\text{FC} = -1.20$) (Fig. 5). *TCEAL4* expression was significantly reduced in grade 3 cases compared to grade 1 ($p = 0.038$, $\log_2\text{FC} = -1.42$) (Supplementary Fig. 6), and *LUC7L2* expression showed low correlation with age ($r = -0.26$, $p = 0.047$) (Supplementary Fig. 7), showing that *RAD50*, *VPS33B*, *GRB7*, *TCEAL4* and *LUC7L2* were not only predictive of the OC state, but also significantly associated with clinical and demographical features.

Selected gene association with overall survival

To see if the selected gene expression levels could also serve as prognostic biomarkers, we tested gene expression association with overall survival of the OC patients. In the OC group, high gene expression was associated

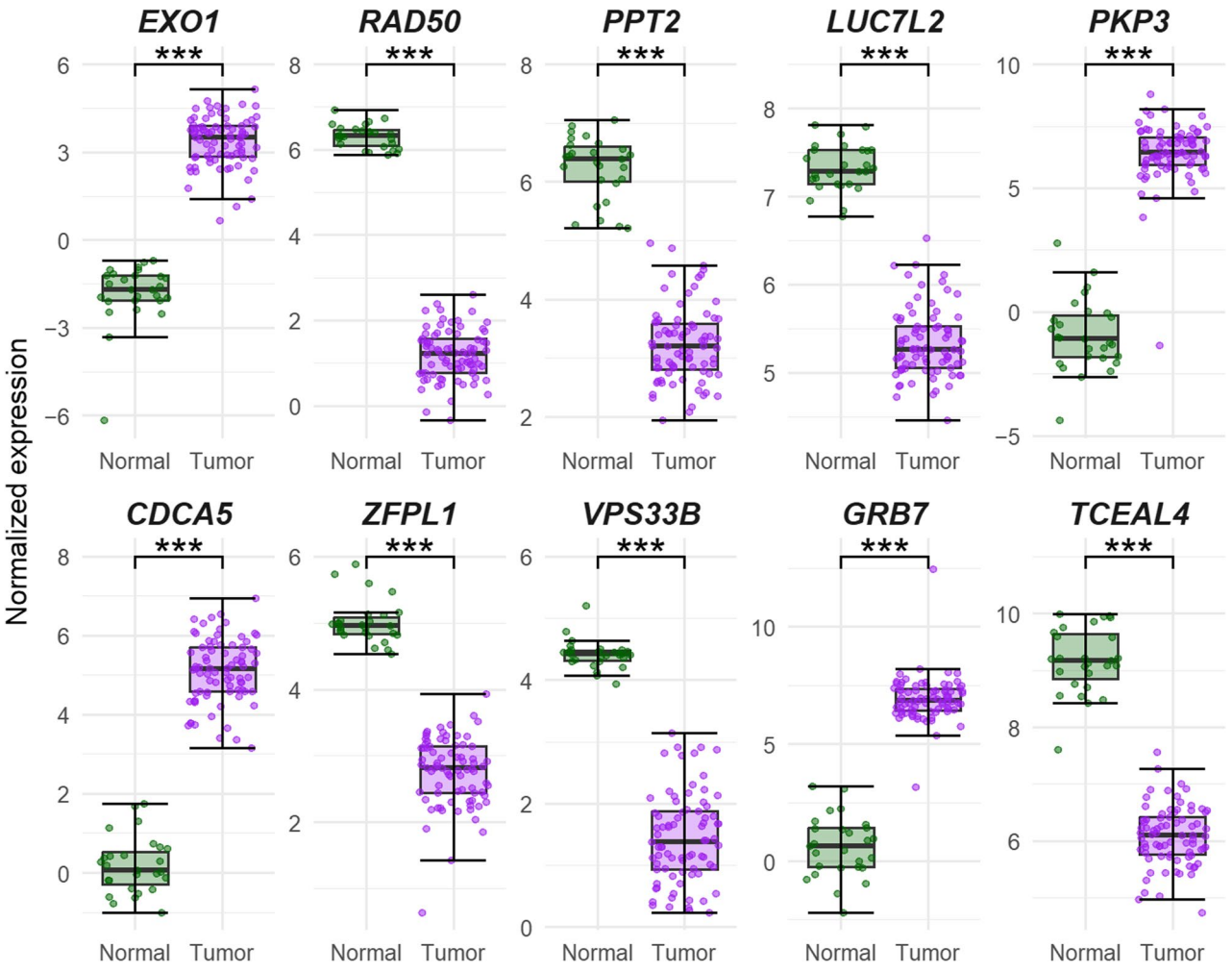


Fig. 2 Selected biomarkers in the test dataset (27 normal (GTEx) and 79 tumor (TCGA-OV) samples): boxplots depicting normalized gene expression in GTEx and TCGA samples

with improved overall survival (HR < 1), with the exception of *PKP3* expression (HR = 1.19); however, none of the associations showed statistical significance (Supplementary Fig. 8). Meanwhile, the combination of all 10 gene expressions into a single risk score did significantly correlate ($p = 0.01$) with longer OC patients' survival (HR = 0.24, 95% CI: 0.07–0.81, adjusted by age and CA125 U/mL at diagnosis: HR = 0.21, 95% CI: 0.05–0.91) (Fig. 6). However, given that the same 10-gene combination predicted survival only in the training, but not the testing cohort, further validation in a larger external cohort is necessary to confirm the finding. Nevertheless, the risk score was able to predict 5-year survival with the AUC of 0.816, sensitivity = 0.677, specificity = 1.00, outperforming all single biomarkers, of which the best AUC was achieved by *GRB7* expression with AUC of 0.556, sensitivity = 0.920, specificity 0.273 (Supplementary Fig. 9).

Discussion

In the present study, we identified and validated a panel of potential diagnostic and prognostic gene expression biomarkers for OC. The analysis selected a set of 10 biomarkers with the greatest association with OC patients' diagnosis and overall survival, which were then validated using RT-qPCR on an external gynecologic tissue cohort. All 10 gene expression biomarkers showed consistent changes in regulation across training, testing, and external validation cohorts. The analysis of the external cohort distinguished *TCEAL4* and *GRB7* as the two biomarkers with the highest diagnostic power, as these biomarkers outperformed the current clinically used biomarker CA125 in separating HGSOC from benign cases, and their combination was able to completely separate these cases. The combination of the 10 gene expression into a single risk score also showed significant prediction of 5-year survival in OC cases, demonstrating diagnostic and prognostic value in biomarker combinations.

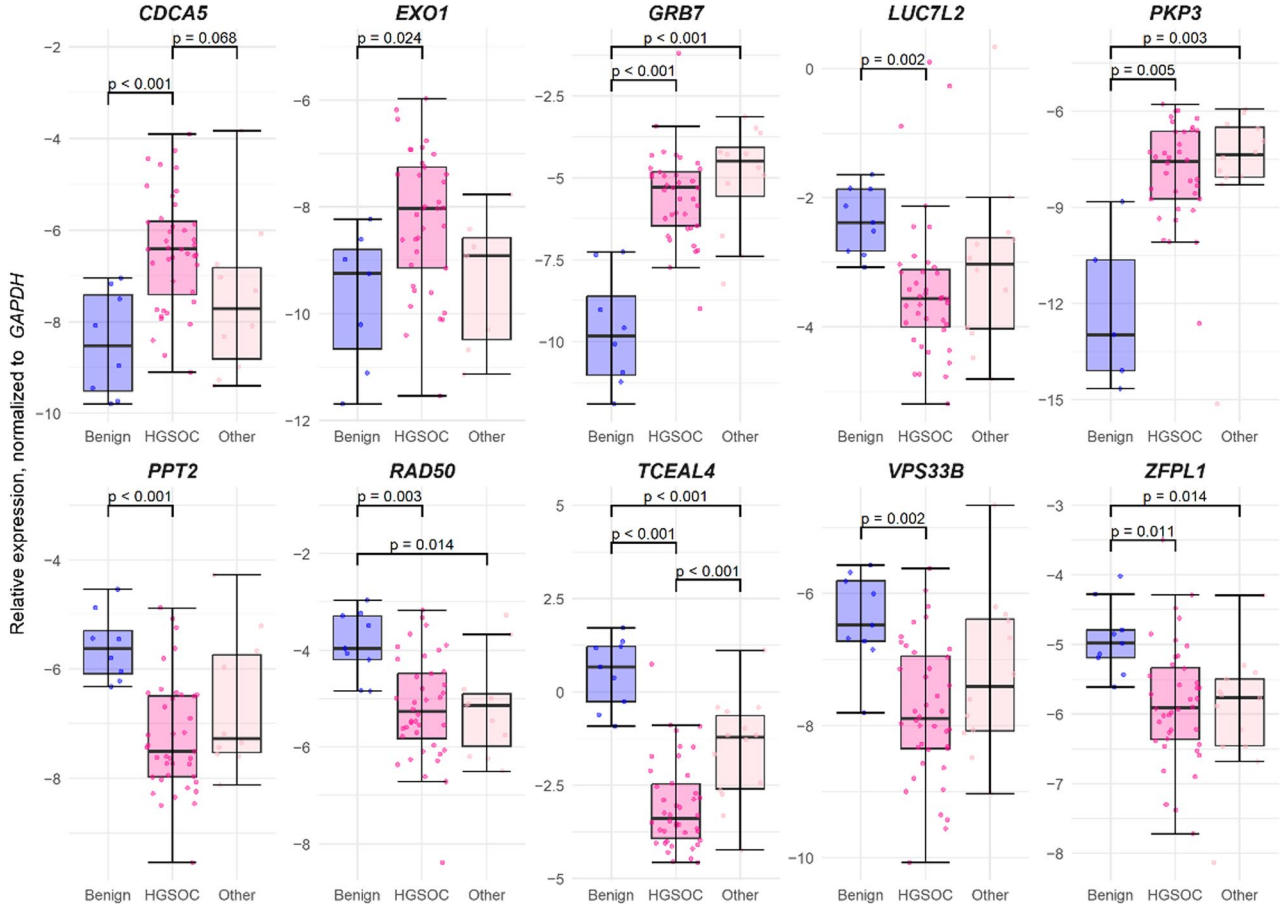
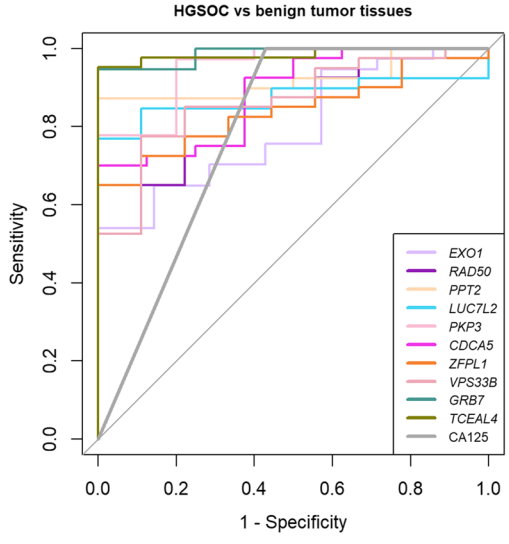


Fig. 3 Gene expression in HGSOC ($n=42$), benign gynecologic tissues ($n=9$) and other malignant gynecologic tumors tissues ($n=14$)



Predictor	AUC	threshold	accuracy	sensitivity	specificity	precision	npv	tpr	fpr
EXO1	0.795	−8.186	0.614	0.541	1.000	1.000	0.292	0.541	0.000
RAD50	0.847	−4.866	0.714	0.650	1.000	1.000	0.391	0.650	0.000
PPT2	0.920	−6.344	0.894	0.872	1.000	1.000	0.615	0.872	0.000
LUC7L2	0.875	−3.078	0.812	0.769	1.000	1.000	0.500	0.769	0.000
PKP3	0.950	−8.791	0.805	0.778	1.000	1.000	0.385	0.778	0.000
CDCA5	0.884	−6.898	0.750	0.700	1.000	1.000	0.400	0.700	0.000
ZFPL1	0.839	−5.617	0.714	0.650	1.000	1.000	0.391	0.650	0.000
VPS33B	0.864	−6.875	0.796	0.775	0.889	0.969	0.471	0.775	0.111
GRB7	0.986	−7.253	0.956	0.946	1.000	1.000	0.800	0.946	0.000
TCEAL4	0.984	−0.927	0.961	0.952	1.000	1.000	0.818	0.952	0.000
CA125	0.786	0.500	0.932	1.000	0.571	0.925	1.000	1.000	0.429

Fig. 4 ROC curves of selected genes for separation of HGSOC ($n=42$), and benign tumors ($n=9$). The ROC measures selected via threshold value determined by Youden index. Npv – negative predictive value, tpr – true positive rate, fpr – false positive rate

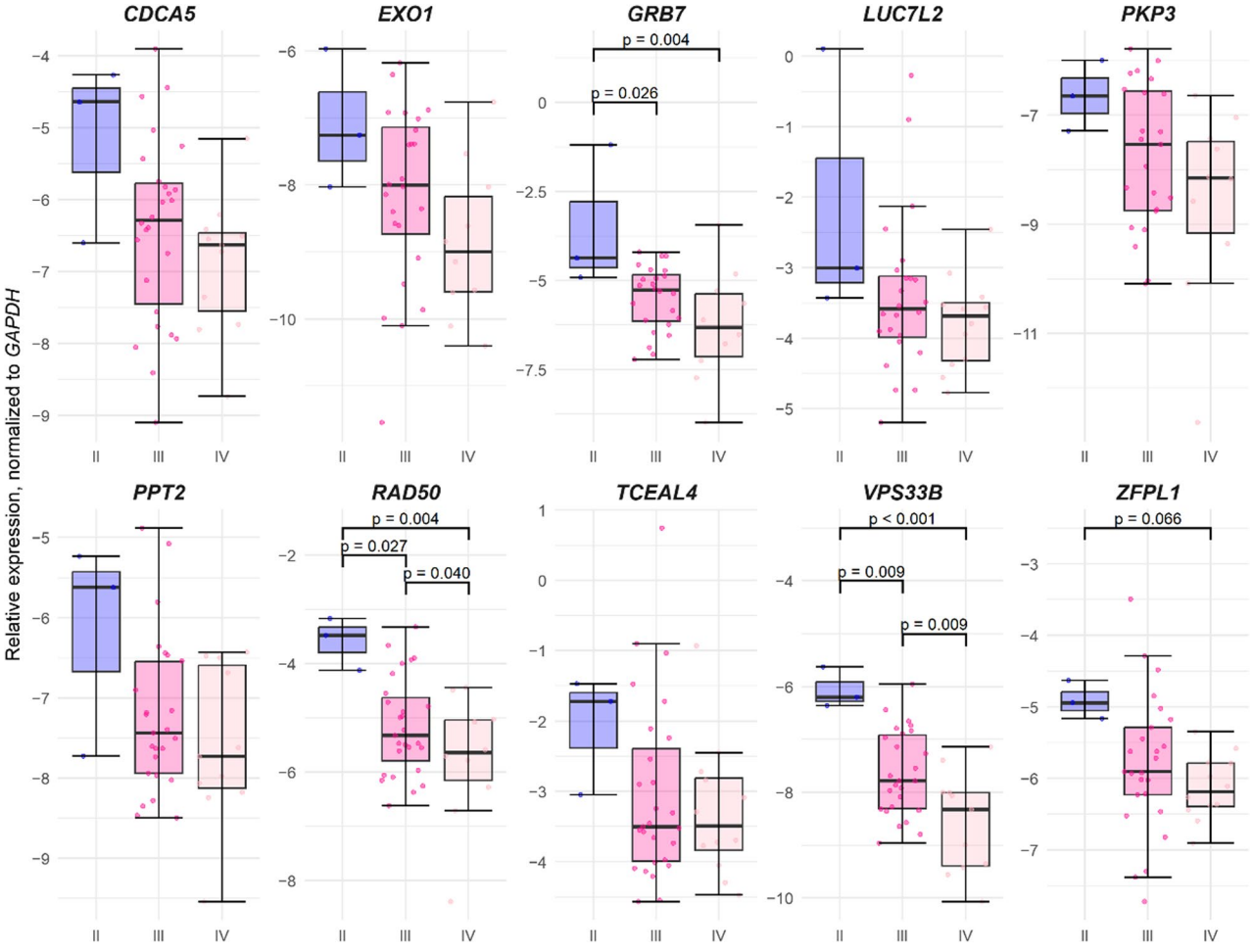


Fig. 5 Boxplots of gene expression in HGSOC samples in relation to FIGO stage (stage II $n=3$, stage III $n=27$, stage IV $n=12$)

The changes in *PPT2* and *GRB7* expression were also found in other TCGA/GTEx studies and some GEO ovarian tissue cohorts as well [9, 10], while *CDCA5* [11] and *PPT2* [9] expression changes were also seen in additional ovarian tissue cohorts using RT-qPCR, showing consistent dysregulation of the selected biomarkers in OC tissues.

Some of the selected biomarkers may have predictive uses as well as diagnostic and prognostic uses. For instance, the *EXO1*, required for single-stranded DNR repair in *BRCA1*-deficient ovarian cells, is overexpressed in *BRCA1*-mutated tumors, thus may serve as a therapeutic target [12]. About 18% of OC patients without *BRCA1* alterations exhibit *RAD50* deletion, associated with better overall survival and sensitivity to olaparib and cisplatin [13]. Another selected gene, *GRB7*, is a potential modulator of immunotherapy response [10] and angiogenesis [14], supporting its potential as a prognostic and therapeutic target.

Despite the limited size of the external OC cohort, limiting the interpretability of the results, gene expression dysregulation patterns were consistent across training, testing, and external datasets. Larger studies using non-invasive samples are necessary to confirm whether the proposed panel can achieve the 99.6% specificity and 75% sensitivity required for screening tests [15]. Study limitations include a limited sample size, incomplete survival data due to some patients not reaching the 5-year follow-up period, and the use of benign tumors instead of normal tissues as controls, which limits comparability with TCGA and GTEx; nonetheless, a high degree of replication was observed. Moreover, using benign samples as a control provides a more clinically relevant cohort and enables evaluation of the ability of the selected biomarkers to distinguish between the benign and malignant tumors. Further validation is warranted to confirm the diagnostic and prognostic potential of the proposed gene panel.

Our study selected candidate OC biomarkers from publicly accessible datasets. Variable selection, such as used in our study, is inherently unstable and must be examined with caution [16]; however, we observed

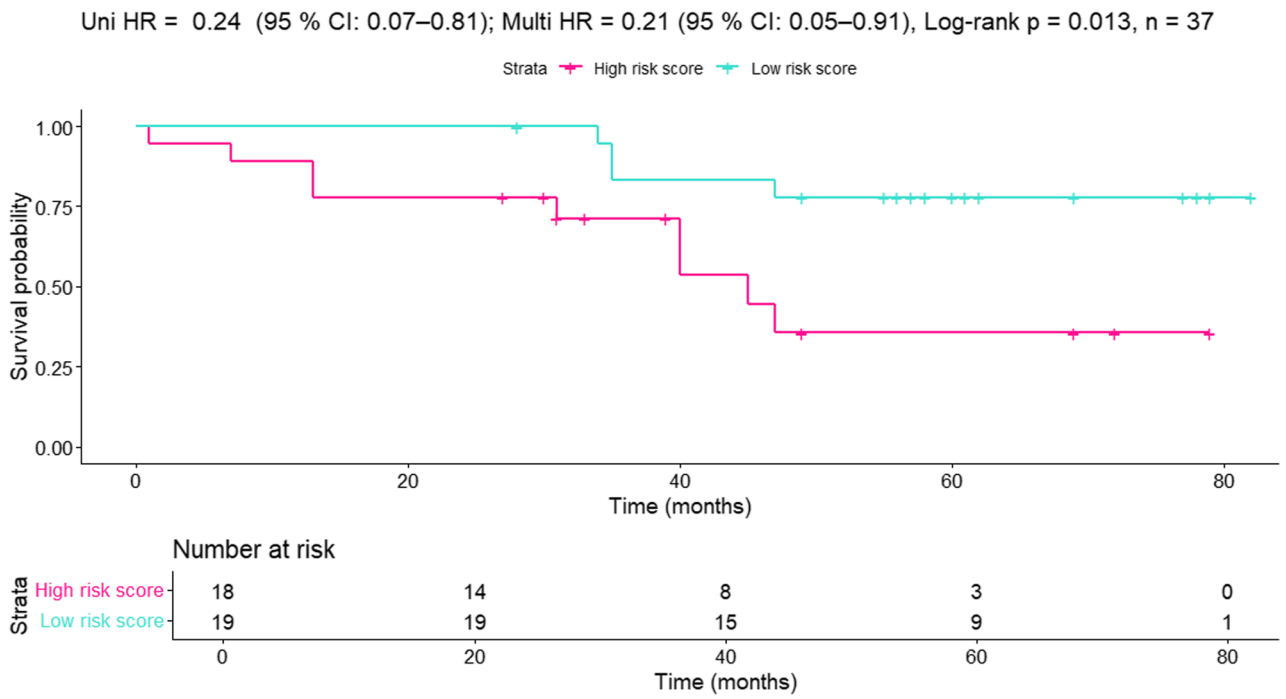


Fig. 6 Gene expression combination (risk score) association with overall survival in the OC cases ($n = 37$, other data excluded due to missingness). Uni HR = univariable cox regression hazard ratio, Multi HR = multivariable cox regression adjusted for age and CA125 concentration at diagnosis

consistent gene expression changes between test, train, and external cohorts, with the selected biomarkers demonstrating promising diagnostic and prognostic potential. Further validation using larger and ideally non-invasive cohorts is still essential to confirm the selected biomarker panel's clinical utility.

Conclusion

The present study shows the ability of transcriptional biomarkers to differentiate between HGSOC and benign or non-serous gynecologic tumors, and suggests the potential utility of biomarker combinations in predicting OC patients' overall survival. By applying machine-learning algorithms to large public datasets, we determined the transcriptomic biomarkers that could exhibit acceptable diagnostic and prognostic accuracy and validated these biomarkers in an external tissue cohort. While further validation in larger and less invasive sample cohorts is necessary for developing a feasible screening strategy for OC, the study provides the groundwork for building transcriptomic diagnostic and prognostic tests for OC.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13048-026-01989-z.

Supplementary Material 1.

Supplementary Material 2.

Institutional review board statement

The study was conducted in accordance with the Declaration of Helsinki. The study was approved by the Regional Bioethics Committee No. 158200–18/5–988–539 amendment No. 2.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Code availability

https://github.com/ieva-vaic/TCGA-OV-RISK-PROJECT.

Clinical trial number

Not applicable.

Authors' contributions

Conceptualization – Rasa Sabaliauskaitė, Julius Juodakis; Data Curation – Rasa Sabaliauskaitė, Ieva Vaicekauskaitė, Rūta Čiurlienė; Giedrė Smailytė; Formal Analysis – Ieva Vaicekauskaitė, Julius Juodakis; Funding Acquisition – Rasa Sabaliauskaitė, Ieva Vaicekauskaitė; Investigation – Ieva Vaicekauskaitė, Paulina Kazlauskaitė; Methodology – Rasa Sabaliauskaitė, Julius Juodakis, Ieva Vaicekauskaitė; Project Administration – Rasa Sabaliauskaitė, Juozas Rimantas Lazutka; Resources – Rasa Sabaliauskaitė; Supervision – Rasa Sabaliauskaitė, Juozas Rimantas Lazutka, Julius Juodakis, Visualization – Ieva Vaicekauskaitė; Writing – Original Draft Preparation - Ieva Vaicekauskaitė; Writing – Review & Editing – all authors.

Funding

This research was funded by the Science Foundation from the National Cancer Institute, Lithuania and doctoral studies fund of Vilnius University.

Data availability

The public datasets analyzed in this study are available through the TCGA (http://portal.gdc.cancer.gov/) and GTEx (www.gtexportal.org) databases. The external validation dataset used in this study is available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 18 November 2025 / Accepted: 16 January 2026
Published online: 27 January 2026

References

1. Torre LA, et al. Ovarian cancer statistics, 2018. CA Cancer J Clin. 2018;68:284–96.
2. Ledermann JA et al. ESGO–ESMO–ESP consensus conference recommendations on ovarian cancer: pathology and molecular biology and early, advanced and recurrent disease. in Annals of Oncology. 2024;35:248–266.
3. Zhang M, Cheng S, Jin Y, Zhao Y, Wang Y. Roles of CA125 in diagnosis, prediction, and oncogenesis of ovarian cancer. Biochim Biophys Acta Rev Cancervol. 2021;1875:188503.
4. Fostira F, Papadimitriou M, Papadimitriou C. Current practices on genetic testing in ovarian cancer. Ann Transl Med. 2020;8:1703.
5. Bell D, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.
6. Cai X, et al. A novel TCGA-validated programmed cell-death-related signature of ovarian cancer. BMC Cancer. 2024;24:1–17.
7. Huang Y, Zhou Y, Zhang M. Identification of seven hypoxia-related genes signature and risk score models for predicting prognosis for ovarian cancer. Funct Integr Genomics. 2023;23:1–13.
8. Pan X, Ma XA. Novel Six-Gene signature for prognosis prediction in ovarian cancer. Front Genet. 2020;11:540331.
9. Xu H, et al. Investigating PPT2's role in ovarian cancer prognosis and immunotherapy outcomes. J Ovarian Res. 2024;17:198.
10. Wen L, et al. GRB7 plays a vital role in promoting the progression and mediating immune evasion of ovarian cancer. Pharmaceuticals. 2024;17:1043.
11. Chen X, Zhou M, Ma S, Wu H, Cai H. KLF5-mediated CDCA5 expression promotes tumor development and progression of epithelial ovarian carcinoma. Exp Cell Res. 2023;429:113645.
12. van de Kooij B, et al. EXO1 protects BRCA1-deficient cells against toxic DNA lesions. Mol Cell. 2024;84:659–e6747.
13. Zhang M, et al. Copy number deletion of RAD50 as predictive marker of BRCAness and PARP inhibitor response in BRCA wild type ovarian cancer. Gynecol Oncol. 2016;141:57–64.
14. Xu Q, et al. Knockdown of growth factor receptor bound protein 7 suppresses angiogenesis by inhibiting the secretion of vascular endothelial growth factor A in ovarian cancer cells. Bioengineered. 2021;12:12179–90.
15. Clarke-Pearson DL. Clinical practice. Screening for ovarian cancer. N Engl J Med. 2009;361:170–7.
16. Cadiou S, Slama R. Instability of Variable-selection algorithms used to identify true predictors of an outcome in Intermediate-dimension epidemiologic studies. Epidemiology. 2021;32:402–11.

Publisher's Note