Appendix 1

1. Study design and datasets

This was a multi-stage, purely in-silico study aimed at identifying and validating subtype-specific biomarkers for endometriosis-associated ovarian cancer (EAOC). Public transcriptomic cohorts from GEO were curated into discovery and validation sets. Discovery comprised GSE226575, GSE157153 and GSE230956 (EAOC cases versus non-malignant tissues, including endometrioma and/or normal adjacent). Validation cohorts comprised an EAOC-enriched survival cohort (GSE65986) and a pure-EAOC cohort for immune deconvolution (GSE226870), with histotype-aware benchmarking in OCCC and ENOC where applicable. Additionally, an independent single-cell RNA-seq OCCC cohort (GSE224334) was utilized to validate the cellular origin of our target gene. No patient-identifiable information was accessed.

2. Software environment and reproducibility

All analyses were performed using R (4.5.2) and Python (3.11). Key packages included GEOquery (≥2.70), limma (≥3.58), fgsea (≥1.24), msigdbr (≥7.5), data.table (≥1.15), ggplot2 (≥3.5), sva (≥3.48) for ComBat, glmnet (≥4.1–7), metafor (≥4.6), Seurat (≥4.3.0), numpy (≥1.26), pandas (≥2.2), scikit-learn (≥1.3), and matplotlib (≥3.8). Random seeds were fixed at 11 in R and 42 in scikit-learn to support reproducibility. The exact scripts used to run GSEA (R/fgsea), LASSO (Python/scikit-learn; logistic L1 with cross-validation), and SVM-RFE (Python/scikit-learn), together with minimal runner commands and environment files (Appendix 2 available at https://cdn.amegroups.cn/static/public/tcr-2025-aw-2458-1.pdf).

3. Preprocessing and probe-to-gene mapping

Series matrices and platform annotations were retrieved using GEOquery with GSEMatrix and GPL downloads enabled. Where submissions contained more than one ExpressionSet, the matrix with the greatest sample count was used. Intensity distributions were inspected using the 1st and 99th percentiles; where the upper percentile exceeded 50, data were transformed with log2(x+1) to ensure an approximate log scale. Probes were mapped to HGNC symbols using platform-specific annotation fields; when multiple probes mapped to the same symbol, the probe with the highest mean expression across samples was retained, yielding a unique gene-by-sample matrix per dataset.

4. Group assignment and design matrices

To harmonize heterogeneous sample annotations, a rule-based detector scanned phenotype fields for plausible grouping columns with two to four unique levels and token patterns indicative of disease versus control. In the absence of a suitable column, the lower-cased sample title field was used as a fallback. Samples matching 'tumor/cancer/carcinoma/endometriosis/disease/lesion' were assigned to the case group, and those matching 'control/normal/benign/adjacent/healthy' to the control group. Unas-signed samples triggered a stop for manual review. A two-level factor (control, case) was encoded and a design matrix with an intercept and the case-versus-control contrast was constructed for limma.

5. Differential expression analysis

Within each discovery dataset, differential expression was performed using limma with empirical Bayes moderation. Genes meeting absolute log2 fold-change ≥2 and Benjamini-Hochberg FDR <0.05 were considered differentially expressed. Volcano plots were generated per dataset. Consensus differentially expressed genes (DEGs) were de-fined as the intersection across discovery datasets; directionality was reported per dataset, but intersection was direction-agnostic.

6. Functional enrichment and PPI

Functional characterization used gene ontology, biological process, and cellular component terms, and KEGG pathways, tested with a hypergeometric framework and BH FDR control; the gene universe comprised all successfully mapped symbols in the relevant dataset. Protein–protein interaction analysis was performed via Metascape with default settings to identify interaction-dense modules consistent with the observed enrichment of mitotic and checkpoint programmes.

7. Gene set enrichment analysis

Hallmark gene sets (MSigDB Category H; human) were retrieved via msigdbr. Genes were ranked by limma moderated t-statistics (positive towards case) and passed to fgsea with 10,000 permutations. Results are reported as normalized enrichment scores and BH-adjusted FDR. For each dataset, enrichment plots were produced for the top positively and negatively enriched Hallmark sets among those with FDR <0.25; an overview bubble plot summarized significant terms. The exact R code used for data retrieval, symbol collapsing, group detection, limma modeling and fgsea visualization is provided as a separate Supplementary Files (Appendix 3 available at https://cdn.amegroups.cn/static/public/tcr-2025-aw-2458-2.pdf).

https://dx.doi.org/10.21037/tcr-2025-aw-2458

8. Machine-learning feature selection

Feature selection combined L1-regularized logistic regression with cross-validation and SVM-RFE, implemented in Python (scikit-learn). For the L1 model, we used a pipeline of median imputation and z-scaling followed by LogisticRegressionCV with an L1 penalty (solver = 'liblinear'), a logarithmic grid of 60 C values spanning 10-4 to 104, five-fold stratified cross-validation with shuffling, max_iter = 20,000, and random_state = 42. The primary scoring mirrored deviance minimization (neg_log_loss), with optional alternatives (roc_auc, f1, accuracy). Labels from the Excel input were coerced to 0/1 as needed. The selected gene set comprised features with non-zero coefficients at the best C, and outputs included: (i) a full coefficient table, (ii) the non-zero (selected) coefficients, and (iii) a cross-validation curve (mean score versus log10(C)) saved as a PDF.

For SVM-RFE, we used a pipeline of median imputation and z-scaling followed by a linear SVM; RFECV with five-fold StratifiedKFold (shuffle = True, random_state = 42) and F1-macro scoring iteratively removed features, selecting the subset that maximized the mean cross-validated score. For very high-dimensional inputs, a smaller elimination step (e.g., steP=0.05) can be used. Both routines accept an Excel input with columns [id, disease, $gene_1$…$gene_n$]; outputs include per-method reports and plots, as documented in the separate Supplementary Files (Appendix 4 available at https://cdn.amegroups.cn/static/public/tcr-2025-aw-2458-3.pdf).

9. Hub-gene consolidation

Candidate hub genes were defined by intersecting LASSO-selected genes, SVM-RFE-selected genes and KEGG-anchored pathway members arising from enrichment analyses. This yielded five high-confidence genes (B4GALNT3, CLDN4, MARVELD2, OCLN and SGPP2) for downstream validation.

10. Survival analysis and evidence synthesis

Pan-ovarian prognostic evidence was assembled from TCGA-OV, ICGC-OV and KM-Plotter. TCGA counts underwent upper-quartile normalization; each gene was modelled as a continuous predictor scaled to one standard deviation in a univariable Cox model for overall survival. For ICGC and KM-Plotter, hazard ratios were extracted from high- versus low-expression contrasts. Study-level log hazard ratios and standard errors were combined using a DerSimonian–Laird random-effects model in metafor, with pooled estimates, confidence intervals and heterogeneity metrics (I^2 and τ^2) reported. Subtype specificity was examined in GSE65986 using Cox models with Wald inference and proportional hazards checks based on Schoenfeld residuals.

11. Diagnostic performance

To strictly evaluate the robustness of the diagnostic biomarker and rule out dataset-specific bias, a leave-one-dataset-out cross-validation strategy was employed. Instead of pooling datasets with batch correction, expression values within each dataset were independently log-transformed (where applicable) and Z-score normalized to ensure comparability without introducing data leakage. We iteratively trained the logistic regression model on two of the three discovery datasets and validated it on the remaining independent dataset. This process was repeated for all three combinations. An overall combined ROC curve was generated based on the prediction probabilities across all folds, and the AUC with its 95% CI was calculated to quantify diagnostic performance.

12. Immune-cell deconvolution and associations

The tumor immune microenvironment was profiled using CIBERSORTx with the LM22 signature matrix. Absolute mode and B-mode batch correction were enabled; quantile normalization was disabled for RNA-seq; 500 permutations were used for significance. Pre-specified QC thresholds were applied and only samples meeting deconvolution $P < 0.05$, reconstruction correlation ≥ 0.80 and RMSE ≤ 0.30 were retained. Associations between B4GALNT3 expression and immune-cell fractions were assessed using two-sided Spearman's rank correlation with BH FDR control across 22 cell types. Identical settings and QC criteria were applied in OCCC and ENOC to gauge histotype-aware concordance.

13. Single-Cell RNA-Seq Analysis

To validate the cellular origin of B4GALNT3, we analysed an independent, publicly available single-cell RNA-seq (scRNA-seq) dataset of Ovarian Clear Cell Carcinoma (OCCC; GSE224334) (37). This cohort served as a relevant proxy given that OCCC is a predominant EAOC histotype and our bulk-data immune correlations were directionally concordant in the OCCC subgroup. Analysis was performed using the Seurat package in R. Following standard quality control filtering based on gene counts and mitochondrial read percentage, data were normalized (LogNormalize), scaled, and principal component analysis was performed. Cells were clustered using graph-based clustering (Find-Neighbors, FindClusters) and

https://dx.doi.org/10.21037/tcr-2025-aw-2458

visualized via t-distributed Stochastic Neighbor Embed-ding (t-SNE). Major cell lineages (including malignant, epithelial, CD4T, CD8T, mono/macro, fibroblasts) were annotated based on the expression of canonical marker genes. To determine the primary cellular source, normalized B4GALNT3 expression was visualized across clusters, and the mean expression (via AverageExpression function) was formally calculated and compared across all major annotated lineages.

14. Multiple testing, reporting and exports

False discovery rate was controlled by Benjamini–Hochberg for gene-level, enrichment and correlation analyses. Unless otherwise stated, tests were two-sided and effect sizes are reported together with confidence intervals or adjusted q-values, as ap-propriate. Figures were exported as vector PDFs with embedded fonts; when raster formats were required, images were written at approximately 300 dpi at their final print size, with line widths exceeding 0.25 pt to ensure legibility at the journal's standard dimensions.

15. Sensitivity checks

Robustness was examined by inspecting PCA/UMAP before and after ComBat, comparing LASSO lambda.min versus the 1-SE rule, using alternative RFECV settings and class weighting for imbalanced subsets, and re-ranking genes for GSEA by signed –log10 P versus moderated t. Wait-list or alternative cross-validation strategies were also considered to ensure stability. Qualitative conclusions remained stable across these perturbations.

16. Limitations

All analyses rely on public transcriptomes and available metadata; platform het-erogeneity and unobserved confounders may remain. The prognostic validation cohort is enriched for EAOC-related histotype but is not a strictly pure EAOC cohort; further protein-level and spatial validation is warranted.

https://dx.doi.org/10.21037/tcr-2025-aw-2458