Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Genomic AI-Guided Discovery, Systems-Biology Characterization and Independent Multi-Cohort Validation of Stable Prognostic Biomarkers in Breast Cancer

Dilip Sisodia[a,b] , Virendra Kumar Sharma[c]
[a] Research Scholar, Department of Computer Science & Engineering, Bhagwant University, Ajmer, Rajasthan, India
[b] Assistant Professor, Department of Computer Science & Engineering, Engineering College, Ajmer, Rajasthan, India
[c] Vice Chancellor, Bhagwant University, Ajmer, Rajasthan, India

Abstract*—Computational biomarker studies are vulnerable to circularity when recurrence, biological annotation, survival association, and external testing repeatedly influence the same candidate set. We therefore used a staged breast-cancer workflow in which genomic candidate stabilization, systems-biology characterization, survival discovery, and external evaluation were analytically separated. The sequencing discovery set comprised 75 samples from 50 patient groups, including 25 matched groups. Candidate status combined patient-group recurrence, paired-role enrichment, TRAIN/VALIDATION-only attribution support, family-specific Benjamini-Hochberg control, and mandatory stability across 1,000 deterministic patient-group bootstrap iterations; locked test predictions and test explainability were excluded. The resulting gene set was characterized with Reactome v97, Ensembl VEP release 115, and STRING v12, then tested for overall-survival association in TCGA-BRCA. Prognostic families were fixed before evaluation in METABRIC, SCAN-B/GSE96058, and GSE20685. Of 5,119 consensus candidates derived from 267,647 recurrent variants and 21,456 recurrent genes, 398 met the internal-stability criteria (386 genes and 12 variants). Twenty-four Reactome pathways were significant under both declared backgrounds. STRING mapped all 386 stable genes, and network coherence exceeded size-matched null expectations; an exact-degree-matched sensitivity analysis also retained significant excess internal connectivity. TCGA-BRCA identified HSP90AA1, SEC63, WT1, WWOX, and ZP1 as the five primary FDR-significant prognostic genes. Strict replication was 2/5 in METABRIC, 5/5 in SCAN-B, and 0/5 in GSE20685, although all five GSE20685 effects retained the TCGA-BRCA risk direction. The adjusted 17-gene family replicated 6/16, 13/17, and 3/17, respectively. These findings support an evidence-narrowing framework that distinguishes genomic stability, expression-associated prognosis, directional agreement, and strict statistical replication. The five-gene family remains hypothesis-generating rather than a causal or deployment-ready clinical signature.

Keywords—breast cancer; prognostic biomarkers; genomic AI; biomarker stability; Reactome; STRING; TCGA-BRCA; METABRIC; SCAN-B; GSE20685; survival analysis; external validation.

I. INTRODUCTION

The present biomarker study follows a previously developed leakage-controlled hierarchical evaluation framework for breast-cancer variant pathogenicity prediction with locked-test validation [5]. Breast cancer is not a single molecular disease. Large-scale genomic and transcriptomic studies have shown that tumors with similar clinical presentations may differ substantially in driver alterations, transcriptional programs, copy-number states, and pathway activity [6,7]. This heterogeneity makes biomarker discovery attractive but also statistically fragile: a signal may appear important because it recurs within a particular cohort, because it is connected to a well-annotated pathway, or because it happens to separate survival in a single dataset. None of those observations alone demonstrates a transportable prognostic biomarker. A second challenge is that modern computational studies often combine several evidence types in one pipeline. Predictive models, feature attributions, recurrence statistics, pathway enrichment, interaction networks, and survival analyses answer different scientific questions. When the same observations are allowed to influence every stage, a seemingly rich evidence chain can become a sequence of post-hoc selections. Reporting recommendations for prognostic tumor markers therefore emphasize prespecified hypotheses, transparent multiplicity control, complete reporting of negative or partial replication, and validation in independent cohorts [18]. Similar principles underlie contemporary guidance for prediction-model studies that use machine learning [19]. Our earlier work addressed NGS processing, mutation prediction, multi-omics risk analysis, and explainable genomic AI [1-4]. Here the question is narrower: can a broad genomic candidate space be reduced to a stable gene family and then evaluated for prognosis without allowing later biological annotation or favorable outcomes to feed back

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

into selection? The term "genomic AI-guided" refers to this analytical provenance. Model-development TRAIN/VALIDATION attribution intensity contributed one predefined support axis, but it could not determine stable status by itself; locked test predictions, labels, and test SHAP/LIME were excluded. Stable status also required recurrence-based evidence and mandatory bootstrap persistence under the multi-axis consensus described below. The sequencing and prognostic cohorts therefore serve different purposes. The former supplies recurrent genomic evidence, whereas TCGA-BRCA provides the primary overall-survival discovery setting for the fixed stable-gene family [6]. METABRIC, SCAN-B/GSE96058, and GSE20685 then test prognostic transportability under different sample sizes, event rates, clinical compositions, and expression platforms [7-9]. External evaluation was intended to assess preservation of effect direction and prespecified statistical evidence, not numerical duplication of TCGA hazard ratios. The analysis followed a one-way evidence hierarchy: patient-group recurrence and bootstrap consensus; assignment of stable gene and variant status; pathway, domain, and network characterization; TCGA-BRCA survival testing with family-wide FDR control; independent external evaluation of the fixed prognostic families; and, finally, a secondary quantitative synthesis of completed external effects. Partial and failed replication were retained as observed, and neither biological annotation nor external outcomes could retrospectively alter biomarker membership.

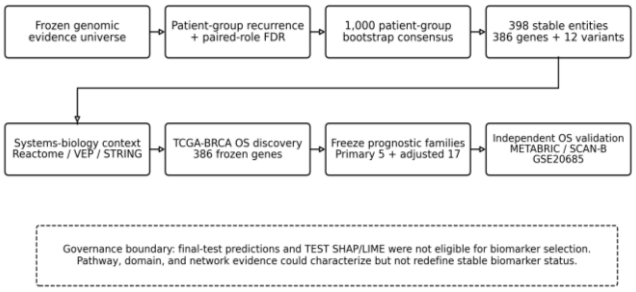

Fig. 1. Ordered evidence hierarchy for the prognostic-biomarker study. Biomarker stability was determined independently of final-test predictions and test explainability; biological resources supplied supportive context only, and prognostic families were fixed before external cohort outcomes were analyzed.

II. MATERIALS AND METHODS

A. *Study design and analytical boundaries*

Each evidence layer had a predefined analytical role. Upstream genomic processing and predictive development had been versioned and locked before the biomarker analyses reported here [5]. This manuscript begins at candidate stabilization. Reproducing the downstream study does not require re-running the earlier classifier; the relevant inherited safeguards are patient-group-separated development, validation-based model and threshold locking, single-use final testing, exclusion of test explainability from biomarker nomination, and immutable upstream output identities. The candidate-stabilization, biological-characterization, survival, multiplicity, and external-replication rules needed to interpret the present work are specified directly below. External survival cohorts evaluate expression-associated prognostic hypotheses derived from the stable-gene family; they do not validate the earlier variant-pathogenicity classifier. ClinVar was used only as an upstream clinical-variant annotation resource [10], not as prognostic outcome evidence. The cross-modal bridge was deliberately gene-identity based. Genomic recurrence/stability nominated a fixed gene set, after which TCGA-BRCA and the external expression cohorts independently asked whether expression of those same gene identities was associated with overall survival. No patient-level genomic-to-expression join was performed between the sequencing discovery cohort and TCGA, and the analysis does not assume that a recurrent variant causes an expression change or that expression mediates the genomic association. Genomic stability and expression-associated prognosis are therefore separate evidence dimensions linked only by gene identity. The discovery branch was intentionally one-directional. Reactome, VEP, and STRING could describe the fixed biomarker family but could not promote an unstable gene or variant. TCGA-BRCA survival analysis could define prognostic subsets only within the 386-gene family, and external cohorts could not replace genes, optimize outcome-specific cut-points, select probes by survival performance, or revise the prior statistical families. Candidate stabilization, biological coherence, prognostic discovery, and transportability were therefore treated as distinct inferential tasks.

B. *Sequencing discovery cohort and candidate generation*

The genomic discovery cohort comprised 75 breast-cancer sequencing samples resolved to 50 patient groups. Primary recurrence calculations used patient_group_id rather than raw sample count so that repeated or related observations within a biological unit did not receive disproportionate weight. Twenty-five matched groups supported a restricted tumor-versus-adjacent-normal paired-role analysis. This paired comparison was treated as enrichment evidence and was not interpreted as a replacement for a dedicated somatic-calling design. The recurrence space contained 267,647 recurrent variants and 21,456 recurrent genes. Entry-level recurrence required observation in at least 2 of the 50 patient groups. Candidate families were evaluated separately at variant and gene levels. Paired-role enrichment used only the 25 matched patient groups and an exact discordant-pair binomial/McNemar formulation; unmatched tumor-only groups did not enter that test. BH-FDR was applied separately to the declared variant- and gene-level paired-role families at $\alpha = 0.05$. The candidate universe was capped at 5,000 variants and 2,000 genes according to the predefined priority hierarchy. Deterministic patient-group bootstrapping used 1,000 iterations (seed 49001), a within-bootstrap minimum recurrence rate of 0.05, and a stability-probability threshold of 0.80. Consensus support axes were PRIORITY, PAIRED_ROLE_ENRICHMENT, MODEL_CONSISTENCY, and BOOTSTRAP_STABILITY. Paired-role support used $q \leq 0.10$; model-consistency support used model-development TRAIN/VALIDATION row attributions, defined as the sum of absolute SHAP values across the 26 original predictors and thresholded at the 75th percentile. Internally stable status required bootstrap stability plus at least two total support axes; final test attribution was prohibited. These rules were fixed before manuscript preparation.

For a feature f across G patient groups, I_{jf} was an indicator equal to 1 when feature f occurred in patient group

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

j and 0 otherwise. The recurrence count C_f and recurrence proportion R_f were defined as:

$$C_f = \sum_{j=1}^{G} I_{jf}\,,\ R_f = \frac{C_f}{G} \quad (1)$$

Here, $j = 1, \ldots, G$; C_f is the number of patient groups containing f, and R_f is the corresponding patient-group recurrence proportion.

For the m raw hypothesis-test p values ordered as $p_{(1)} \leq \cdots \leq p_{(m)}$, the Benjamini-Hochberg adjusted value for rank i was [17]:

$$q_i = \min_{j \geq i}\left\{\min\left[1, \frac{m}{j} p_{(j)}\right]\right\} \quad (2)$$

Here, m is the number of hypotheses, i is the ordered rank, and q_i is the multiplicity-adjusted value used for FDR decisions.

For B deterministic patient-group bootstrap iterations, I_{fb}^{sel} was 1 when feature f met the selection rule in bootstrap iteration b and 0 otherwise. Bootstrap stability S_f was:

$$S_f = \frac{1}{B}\sum_{b=1}^{B} I_{fb}^{\text{sel}} \quad (3)$$

In this study, $B = 1000$, $b = 1, \ldots, B$, and stable status required $S_f \geq 0.80$ together with the declared multi-axis consensus rule.

C. Stability governance and reproducibility

Biomarker discovery was deliberately independent of final-test model behavior. No model retraining, inference, threshold modification, access to the locked independent TEST dataset, access to locked final TEST predictions/results, or TEST explainability was permitted in candidate selection. The only model-derived support allowed was the predeclared TRAIN/VALIDATION attribution-intensity axis described above, and it could not substitute for the mandatory bootstrap criterion. An independent read-only audit reconstructed recurrence topology, paired-role FDR calculations, deterministic bootstrap behavior, support-axis consensus, and final counts. The audit passed 95/95 declared checks. These checks document computational reproducibility and policy adherence; they are not additional biological observations.

D. Reactome pathway analysis

Pathway context was evaluated only after biomarker membership had been fixed. Reactome v97 was locally versioned and contained 2,868 pathways [11]. Of 386 stable genes, 304 mapped to Reactome. The primary background comprised the 2,000 gene candidates (1,353 mapped), whereas the sensitivity background comprised all 21,456 recurrent genes (10,915 mapped). One-sided over-representation analysis with Fisher/hypergeometric consistency checks was performed under each background. Pathways were eligible when the background intersection contained 5-500 genes, and BH-FDR correction was applied separately across all eligible pathways, including eligible zero-hit pathways. A pathway was called background-robust only when it met FDR significance under both the primary and sensitivity backgrounds. Because Reactome contains nested and overlapping pathways, robust pathways were interpreted as coherent biological context rather than independent mechanisms. For a declared background of N genes, a pathway/domain containing K background genes, a foreground of n genes, and observed overlap k, the one-sided upper-tail hypergeometric probability was:

$$P(X \geq k) = \sum_{x=k}^{\min(n,K)} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad (4)$$

Here, X is the hypergeometric overlap variable; N, K, n, and k denote background size, annotated-set size, foreground size, and observed overlap, respectively.

For pathway p, let $q_{p,\text{primary}}$ and $q_{p,\text{sensitivity}}$ denote the BH-adjusted values under the two declared backgrounds. The background-robustness indicator R_p^{rob} was defined as:

$$R_p^{\text{rob}} = I\left(q_{p,\text{primary}} \leq 0.05\right) I\left(q_{p,\text{sensitivity}} \leq 0.05\right) \quad (5)$$

Thus, $R_p^{\text{rob}} = 1$ only when the same pathway satisfied $q \leq 0.05$ under both backgrounds.

E. VEP protein-domain characterization

Protein-domain context was reconstructed with Ensembl Variant Effect Predictor release 115 on GRCh38 [12]. All 3,119 variant candidates were evaluated against the selected-transcript authority; alternative-transcript annotations were retained as sensitivity context but could not override the selected transcript. Domain and functional-region associations were tested with the predeclared Fisher/BH-FDR procedure. Gene-provenance discordance between the biomarker source and selected transcript was retained explicitly rather than silently reassigned. The same upper-tail Fisher/hypergeometric formulation in Eq. (4) was used for the declared domain and functional-region families before BH-FDR correction. Domain findings were considered supportive rather than mechanistic validation. This restriction is important because multiple domain database records may arise from the same small number of stable transcript observations. Spatial clustering within protein domains was not assessed because a suitable locked coordinate resource was not part of the analysis contract.

F. STRING interaction-network analysis

STRING v12 was used to map the 2,000 candidate genes and the 386 stable genes into functional and physical association networks [13]. Connected components were calculated, and Louvain community detection was applied to the functional network [14]. Network centrality, community membership, or connectivity was not permitted to alter stable biomarker status. The prespecified convergence question was narrower: whether genes belonging to the background-robust Reactome foreground were represented within the principal functional network component. Network coherence was assessed with a predefined permutation analysis. In each network, the observed 386-gene stable set was compared with 10,000 random gene sets of the same size drawn from the 1,965 STRING-mapped candidate-gene universe, without adding external STRING neighbors. Four statistics were

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

specified in advance: internal-edge count, mean degree, largest-connected-component membership, and hub count. Functional and physical networks used deterministic seeds 52001 and 52002, respectively. One-sided empirical p values were corrected across the eight tests with BH-FDR. The analysis quantified whether the already selected gene set was more coherent than size-matched candidate sets; network results had no role in biomarker selection.

Let S denote the stable-gene set, $|S|$ its size, $A = [A_{ij}]$ the network adjacency matrix, and d_i the full-network degree of node i. The four prespecified coherence statistics were:

$$E_S = \frac{1}{2}\sum_{i\in S}\sum_{j\in S} A_{ij}\,, \qquad \bar{d}_S = \frac{2E_S}{|S|}$$

$$L_S = |\mathrm{LCC}(G[S])|,\ H_S = \textstyle\sum_{i\in S} I\,(d_i \geq d_{\mathrm{hub}}) \quad (6)$$

Here, E_S is the number of stable-stable edges, $\bar{d}_S$ is mean internal degree, $G[S]$ is the subgraph induced by S, $\mathrm{LCC}(H)$ denotes the largest connected component of a graph H, $L_S = |\mathrm{LCC}(G[S])|$, H_S is the hub count, and d_{hub} is the predefined 95th-percentile full-network degree threshold.

For Louvain community detection, the modularity objective Q_L was [14]:

$$Q_L = \tfrac{1}{2M}\textstyle\sum_{i,j}\left(A_{ij} - \frac{d_i d_j}{2M}\right) I(c_i = c_j) \quad (7)$$

In Eq. (7), M is total edge weight, A_{ij} is the edge weight between nodes i and j, d_i and d_j are their degrees, c_i and c_j are community labels, and $I(c_i = c_j)$ is an indicator for common community membership.

For a prespecified network statistic T, let T_{obs} be the observed value and T_b the value from matched random set b, for $B = 10000$ permutations. The empirical upper-tail p value and fold enrichment were:

$$p_{\mathrm{perm}} = \frac{1+\sum_{b=1}^{B} I(T_b \geq T_{\mathrm{obs}})}{B+1},\ \mathrm{FE} = \frac{T_{\mathrm{obs}}}{B^{-1}\sum_{b=1}^{B} T_b} \quad (8)$$

Here, p_{perm} is the permutation p value and FE is the observed statistic divided by the mean null statistic.

Because size-matched sampling can remain sensitive to the degree distribution of well-studied STRING genes, a secondary sensitivity analysis was added without altering the primary network result or biomarker status. Separately for the functional and physical networks, 10,000 random 386-gene sets were drawn from the same 1,965 mapped candidate universe while preserving the exact full-network degree-frequency distribution of the observed stable set. Internal stable-stable edge count was the only tested statistic because mean degree and hub count are fixed by exact-degree matching. Deterministic seeds were 52101 (functional) and 52102 (physical); the two one-sided empirical p values were BH-corrected as one two-test sensitivity family. This analysis addresses degree-composition bias but cannot remove STRING knowledge-density bias.

G. TCGA-BRCA overall-survival analysis

The complete 386-gene family was tested for overall-survival association in TCGA-BRCA [6]. The RNA-evaluable primary cohort contained 1,092 patients and 151 deaths. Cox models used continuous $x_{ig} = \log_2(\mathrm{TPM}_{ig} + 1)$ expression; no standardized z-scale model was refitted for the primary analysis. For comparable reporting across cohorts, the native TCGA coefficient and confidence limits were algebraically rescaled by the gene-specific expression standard deviation to obtain HRs per 1 SD. BH-FDR was applied across all 386 genes. Median-expression Kaplan-Meier curves were descriptive only, with no outcome-optimized cut-point search. A secondary age- and pathologic-stage-adjusted complete-case analysis contained 1,048 patients and 138 deaths. The prespecified adjusted TCGA model included expression, age per 10 years, and AJCC stage II/III/IV indicator terms with stage I as reference. The adjusted model changes the estimand because it conditions on clinical covariates and excludes incomplete cases; it was therefore reported as a complementary robustness analysis rather than a replacement for the primary model. Proportional-hazards assumptions were evaluated using Schoenfeld rank-time residual diagnostics with separate BH control across each declared 386-gene family. The complete 386-gene primary and adjusted Cox/PH output, including coefficients, standard errors, raw Wald p values, BH q values, HR/CI per one gene-specific expression SD, and PH q values, is retained as a machine-readable reproducibility artifact and can be supplied to editors/reviewers or deposited with the journal if requested.

For sample i and gene g, let x_{ig} denote native TCGA expression, μ_g the cohort mean, σ_g the cohort standard deviation, z_{ig} standardized expression, and $\beta_{g,\mathrm{native}}$ the native-scale Cox coefficient. The reporting-scale transformation was:

$$x_{ig} = \log_2(\mathrm{TPM}_{ig} + 1), \qquad z_{ig} = \frac{x_{ig} - \mu_g}{\sigma_g}$$

$$\beta_{g,\mathrm{SD}} = \beta_{g,\mathrm{native}}\,\sigma_g \quad (9)$$

Thus, $\beta_{g,\mathrm{SD}}$ is a per-one-standard-deviation rescaling of the fitted TCGA coefficient; no survival model was refitted to obtain it.

For the single-gene Cox proportional-hazards model [15], $h_i(t)$ is the hazard for patient i at time t, $h_0(t)$ is the baseline hazard, β_g is the log-hazard coefficient, and $\mathrm{SE}(\beta_g)$ is its standard error:

$$h_i(t) = h_0(t)\exp(\beta_g z_{ig}), \qquad \mathrm{HR}_g = \exp(\beta_g)$$

$$\mathrm{CI}_{95\%} = \exp[\beta_g \pm 1.96\ \mathrm{SE}(\beta_g)] \quad (10)$$

Here, HR_g is the hazard ratio for a one-unit increase on the modeled expression scale; the confidence interval uses the usual Wald approximation.

For the clinically adjusted model, R is the number of clinical covariates, C_{ir} is covariate r for patient i, and γ_r is its coefficient:

$$h_i(t) = h_0(t)\exp\left(\beta_g z_{ig} + \textstyle\sum_{r=1}^{R}\gamma_r\, C_{ir}\right) \quad (11)$$

For the descriptive Kaplan-Meier estimator [16], at ordered event time t_j, d_j is the number of events and n_j is the number at risk immediately before t_j:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \quad (12)$$

For the proportional-hazards diagnostic [29], z_{gj} is the observed value of covariate g for the subject failing at t_j, and $\bar{z}_g(t_j)$ is its risk-set weighted expectation under the fitted model:

$$r_{gj} = z_{gj} - \bar{z}_g(t_j) \quad (13)$$

H. Prespecification of prognostic families

Genes passing BH-FDR in the primary TCGA-BRCA analysis formed the primary prognostic family; genes significant in the adjusted analysis formed a separately labeled adjusted family. Both lists were fixed before any METABRIC, SCAN-B/GSE96058, or GSE20685 outcome analysis, so external results could not change family membership.

I. Prespecification of prognostic families

Three independent breast-cancer expression cohorts were used. METABRIC contributed 1,980 patients with 1,143 events to the primary analysis and 1,454 patients with 821 events to the adjusted analysis [7]. SCAN-B/GSE96058 contributed 3,273 patients with 336 events and 3,149 patients with 316 adjusted-analysis events [8]. GSE20685/GPL570 contributed 327 patients with 83 events and 325 patients with 83 events in its adjusted complete-case analysis [9]. The cohorts differ substantially in platform, treatment era, case composition, follow-up, and event fraction; these differences were treated as transportability stress tests rather than nuisance differences to be post-hoc harmonized away. External preprocessing and covariates were specified separately for each cohort before outcome interpretation. METABRIC used authoritative Entrez mapping, 374/386 assay-evaluable genes, z-scores over all 1,980 expression samples, and adjustment for age/10 years plus stage II/III/IV (stage I reference); stage 0 or missing stage was excluded only from adjusted models. SCAN-B used the supplied transformed matrix without a second log transform, exact/current gene symbols plus a predefined legacy bridge, exclusion of 136 technical replicates, and z-scores over 3,273 independent tumors; adjusted covariates were age/10 years, node positivity, and tumor size/10 mm. GSE20685 used the processed GPL570 matrix, locked exact/current-or-legacy symbol mapping, sample-wise median aggregation across all eligible probes without best-probe selection, and z-scores over all 327 patients (ddof=1); adjusted covariates were age/10 years and ordinal T and N stage. These cohort-specific estimands were retained rather than harmonized after seeing survival results. Strict external replication required two conditions: the external Cox coefficient had to preserve the TCGA-BRCA direction and the gene had to satisfy BH $q \le 0.05$ within the prespecified external family. Directional concordance was reported separately because a lower-powered cohort can retain a coherent effect orientation without crossing an FDR threshold. External data were not allowed to replace genes, optimize probes according to survival outcome, change expression cut-points, redefine covariates opportunistically, or retrain the upstream classifier. In METABRIC, EIF2B3 was structurally non-evaluable in the adjusted 17-gene family and was not replaced, leaving 16 evaluable genes. For gene g in external cohort c, let β_{gc} be the external Cox coefficient, $\beta_{g,\mathrm{TCGA}}$ the discovery-cohort coefficient, and q_{gc} the BH-adjusted value within the prespecified external family. The strict-replication indicator $\mathcal{R}_{gc}$ was:

$$\mathcal{R}_{gc} = I\left[\mathrm{sign}(\beta_{gc}) = \mathrm{sign}(\beta_{g,\mathrm{TCGA}})\right] I(q_{gc} \le 0.05) \quad (14)$$

Accordingly, $\mathcal{R}_{gc} = 1$ only when the coefficient direction is preserved and $q_{gc} \le 0.05$; direction concordance without FDR significance was reported separately.

TABLE I. COHORTS AND ANALYTICAL ROLES.

Dataset	Role	Primary N / events	Adjusted N / events	Key constraint
Sequencing discovery cohort	Recurrence and biomarker stability	75 samples / 50 patient groups	25 matched groups for paired-role analysis	No TEST predictions or TEST XAI used for biomarker selection
TCGA-BRCA	Primary prognostic discovery	1,092 / 151	1,048 / 138	386-gene family tested; FDR across declared family
METABRIC	Independent prognostic validation	1,980 / 1,143	1,454 / 821	No gene/probe/cut-point reselection; EIF2B3 non-evaluable in adjusted family
SCAN-B / GSE96058	Independent prognostic validation	3,273 / 336	3,149 / 316	Fixed family and direction + within-family FDR rule
GSE20685 / GPL570	Independent cross-platform validation	327 / 83	325 / 83	Low-event microarray stress test; strict and direction-only results reported separately

J. Secondary cross-cohort quantitative synthesis

After the three external analyses were complete, a secondary meta-analysis summarized transportability without changing the prespecified replication rule. Cohort-specific log-HRs and standard errors for the five primary genes were read directly from completed external outputs on the within-cohort standardized-expression scale. TCGA-BRCA was excluded from the principal pooled analysis because it was the discovery cohort. Random-effects models used REML for between-cohort variance and t-based Hartung-Knapp inference with the variance scale constrained not to fall below 1 [26,27]. Heterogeneity was summarized by τ^2, Cochran Q, and I^2, with BH-FDR across the five pooled tests. A secondary four-cohort synthesis included TCGA-BRCA, and leave-one-external-cohort-out estimates assessed dependence on individual validation cohorts. No pooled result could revise family membership, cohort-level replication status, probes, covariates, expression scaling, or upstream results.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

For external cohort c, HR_c is the cohort-specific hazard ratio, y_c its log-hazard-ratio effect, and v_c its sampling variance:

$$y_c = \ln(\mathrm{HR}_c),\ v_c = [\mathrm{SE}(y_c)]^2 \quad (15)$$

For a candidate between-cohort variance τ^2, define $w_c(\tau^2) = 1/(v_c + \tau^2)$ and $\mu(\tau^2) = \sum_c w_c\, y_c / \sum_c w_c$. The REML estimate of τ^2 was obtained from the restricted-likelihood objective [26]:

$$\hat{\tau}^2_{\mathrm{REML}} = \underset{\tau^2 \geq 0}{\operatorname{argmin}}\, \ell_R(\tau^2)$$

$$\ell_R(\tau^2) = \tfrac{1}{2}[\textstyle\sum_c \ln(v_c + \tau^2) + \ln(\sum_c w_c) + \sum_c w_c\,(y_c - \mu)^2] \quad (16)$$

Using $\hat{\tau}^2_{\mathrm{REML}}$, the random-effects weight w_c, pooled log-HR $\hat{\mu}$, and pooled hazard ratio were:

$$w_c = \frac{1}{v_c + \hat{\tau}^2_{\mathrm{REML}}},\ \hat{\mu} = \frac{\sum_c w_c y_c}{\sum_c w_c},\ \mathrm{HR}_{\mathrm{pooled}} = e^{\hat{\mu}} \quad (17)$$

For heterogeneity [30], k is the number of cohorts and $\hat{\mu}_F$ is the fixed-effect pooled estimate. Cochran's Q and I^2 were:

$$Q = \textstyle\sum_c v_c^{-1}\,(y_c - \hat{\mu}_F)^2,\ I^2 = \max\left(0, \frac{Q-(k-1)}{Q}\right) \times 100\% \quad (18)$$

For modified Hartung-Knapp inference [27], q_{HK} is the variance inflation factor constrained to be at least 1, $\mathrm{SE}_{\mathrm{HK}}$ is the corresponding pooled standard error, and $t_{k-1,0.975}$ is the two-sided 95% critical value with $k-1$ degrees of freedom:

$$q_{\mathrm{HK}} = \max\left[1, \frac{\sum_c w_c\,(y_c - \hat{\mu})^2}{k-1}\right]$$

$$\mathrm{SE}_{\mathrm{HK}} = \sqrt{\frac{q_{\mathrm{HK}}}{\sum_c w_c}},\ \mathrm{CI}_{95\%} = \exp\left[\hat{\mu} \pm t_{k-1,0.975}\,\mathrm{SE}_{\mathrm{HK}}\right] \quad (19)$$

BH-FDR in Eq. (2) was applied separately to each declared family, including the 386-gene TCGA survival family, each external replication family, the network permutation family, and the five pooled gene tests. Effect sizes are reported as HRs with 95% confidence intervals where available. Independent audits reconstructed cohort mappings, model specifications, FDR families, direction-concordance rules, and output identities without scientific reselection. Reporting was designed to remain consistent with REMARK principles for prognostic marker studies and broader transparency principles for machine-learning research [18,19].

III. RESULTS

A. *Candidate stabilization reduced a large recurrent space to 398 stable entities*

The discovery cohort yielded 267,647 recurrent variants and 21,456 recurrent genes. Consensus generation reduced this space to 5,119 candidates (2,000 genes and 3,119 variants), of which 398 met the internal-stability criteria: 386 genes and 12 variants. The other 4,721 candidates remained explicitly exploratory. The reduction was especially strong at the variant level. Stable genes represented 19.3% of the 2,000-gene candidate family, whereas stable variants represented only 0.38% of the 3,119-variant candidate family. This pattern demonstrates why recurrence alone was not sufficient for promotion to stable biomarker status. Importantly, the independent biomarker audit reproduced the candidate counts and stability logic without loading the predictive model or accessing TEST outputs.

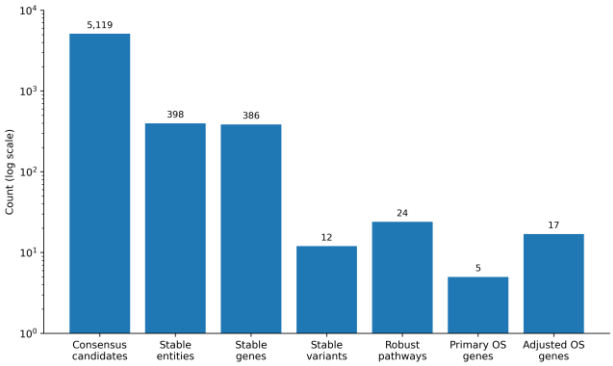

Fig. 2. Ordered evidence hierarchy for the prognostic-biomarker study. Biomarker stability was determined independently of final-test predictions and test explainability; biological resources supplied supportive context only, and prognostic families were fixed before external cohort outcomes were analyzed.

B. *Reactome enrichment was robust to background definition*

Of the 386 stable genes, 304 mapped to Reactome. The primary mapped background produced 840 eligible pathways, 36 of which passed BH-FDR; the broader recurrent-gene background produced 2,125 eligible pathways and 44 FDR-significant results. Twenty-four pathways were significant under both backgrounds. The independent pathway audit reproduced all 231 declared checks. The strongest robust signals were dominated by DNA-repair biology. "Diseases of DNA repair" had primary-background q=0.000728, while DNA Repair, Homologous DNA Pairing and Strand Exchange, the presynaptic phase of homologous pairing, TP53-regulated transcription of DNA-repair genes, diseases of DNA double-strand-break repair, and defective homologous-recombination repair due to BRCA2 loss all had q=0.001298. Other robust signals included homologous recombination repair, TP53 transcriptional regulation, BRCA1/PALB2-associated repair defects, PIP3/AKT signaling, and related growth-factor signaling. Because several Reactome terms share genes and hierarchy, the 24 pathways were interpreted as overlapping evidence for repair and signaling programs rather than 24 independent biological discoveries.

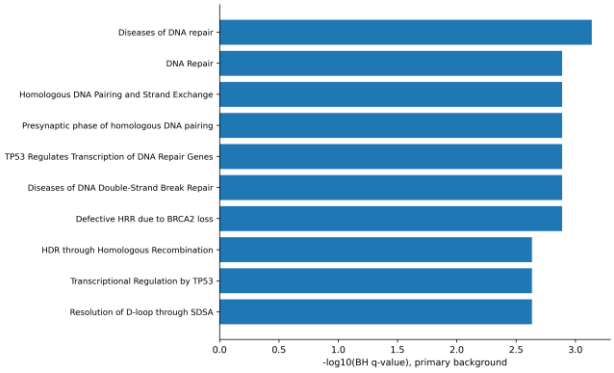

Fig. 3. Ten of the most significant background-robust Reactome pathways under the primary background. All plotted pathways also passed FDR under the broader recurrent-gene sensitivity background.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

C. *Protein-domain results supported a small subset of stable variants*

The selected transcript was recovered for 3,116 of 3,119 variant candidates. Domain annotation was present for 1,379 candidates; 1,190 mapped to a domain on the selected transcript and 460 had selected-transcript Pfam annotation. Seven of the 12 stable variants mapped to a selected-transcript domain and six to Pfam. Gene-provenance discordance was retained for 150 candidates, including two stable variants. Across 464 primary domain tests, three annotations passed FDR correction: PRINTS PR00889, PROSITE profile PS51122, and Pfam PF00402. None of 14 functional-region tests passed FDR. The three significant domain rows were largely driven by the same two stable transcript observations and were therefore not counted as three independent mechanisms. The independent domain audit passed 299/299 checks.

D. *Stable genes occupied a coherent STRING interaction structure*

STRING v12 mapped 1,965 of 2,000 candidate genes and all 386 stable genes. The functional network contained 5,236 unique edges and a largest connected component (LCC) of 1,385 genes; the physical-sensitivity network contained 1,601 unique edges. Louvain analysis identified 21 functional communities. All 71 genes contributing to the background-robust Reactome foreground lay in the full functional LCC. Their induced subnetwork contained a 62-gene, 285-edge LCC. The independent network audit reproduced 8,239 declared checks.

Connectivity exceeded the size-matched null in both network definitions. In the functional network, the stable set had 607 internal edges (null mean 201.27), mean degree 8.42 (5.33), LCC membership 338 (272.13), and 38 hubs (20.97). The corresponding physical-network values were 195 edges (61.55), mean degree 2.62 (1.63), LCC membership 179 (123.58), and 36 hubs (19.66). All eight one-sided permutation tests had p=0.0001 and BH q=0.0001.

Exact-degree matching reduced the apparent edge enrichment but did not eliminate it. Functional stable-stable edges were 607 versus a null mean of 537.66 (1.129-fold; p=0.0005; BH q=0.0010), and physical-network edges were 195 versus 167.67 (1.163-fold; p=0.0037; BH q=0.0037). Thus, internal connectivity remained greater than expected after the random sets were constrained to reproduce the observed full-network degree distribution.

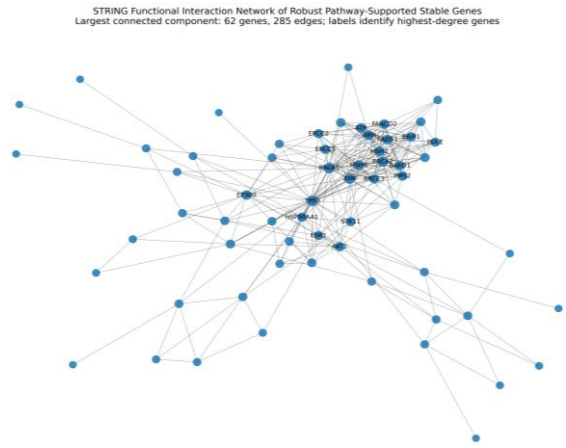

Fig. 4. STRING functional interaction network of background-robust pathway-supported stable genes. The displayed induced-subnetwork LCC contains 62 genes and 285 functional edges; all 71 robust-pathway genes were members of the broader 1,385-gene functional LCC.

TABLE II. BIOLOGICAL EVIDENCE SUMMARY AFTER BIOMARKER STABILIZATION.

Evidence layer	Denominator	Observed result	Interpretive role
Stable candidates	5,119 consensus candidates	398 stable = 386 genes + 12 variants	Internal stability, not external validation
Reactome primary	840 eligible pathways	36 FDR-significant	Primary pathway context
Reactome sensitivity	2,125 eligible pathways	44 FDR-significant	Background sensitivity
Robust pathways	Cross-background overlap	24 significant in both	Background-robust supportive context
VEP selected transcript	3,119 variants	3,116 matched; 1,190 selected-domain	Transcript-resolved variant context
Stable variant domains	12 stable variants	7 selected-domain; 6 Pfam	Supportive domain evidence
STRING functional	2,000 candidate genes	1,965 mapped; 5,236 edges; LCC=1,385	Functional interaction context
STRING physical	Mapped network	1,601 unique edges	Physical interaction sensitivity
Louvain communities	Functional network	21 communities	Community structure
Robust-pathway genes	71 genes	71/71 in full functional LCC; induced LCC 62 genes / 285 edges	Pathway-network convergence without biomarker reselection
STRING functional permutation	386 stable vs 1,965 mapped candidates; 10,000 null sets	Edges 607 vs 201.27 (3.02x); degree 8.42 vs 5.33; LCC 338 vs 272.13; hubs 38 vs 20.97; all 4 BH q=0.0001	Permutation-supported network coherence; supportive context only
STRING physical permutation	386 stable vs 1,965 mapped candidates; 10,000 null sets	Edges 195 vs 61.55 (3.17x); degree 2.62 vs 1.63; LCC 179 vs 123.58; hubs 36 vs 19.66; all 4 BH q=0.0001	Cross-network sensitivity for non-random coherence
STRING functional degree-matched sensitivity	386 stable vs exact-degree-matched candidate null; 10,000 sets	Internal edges 607 vs 537.66 (1.129x); p=0.0005; BH q=0.0010	Sensitivity to degree-composition bias
STRING physical degree-matched sensitivity	386 stable vs exact-degree-matched candidate null; 10,000 sets	Internal edges 195 vs 167.67 (1.163x); p=0.0037; BH q=0.0037	Cross-network degree-matched sensitivity

E. *TCGA-BRCA survival analysis identified a five-gene primary prognostic family*

TCGA-BRCA contributed 1,092 RNA-evaluable patients and 151 deaths to the primary analysis; the adjusted complete-case analysis included 1,048 patients and 138 deaths. Testing all 386 genes identified five primary FDR-significant associations: HSP90AA1, SEC63, WT1, WWOX, and ZP1. HRs per 1 expression SD ranged from 1.239 to 1.321, and all 95% CIs excluded 1. Each gene had primary-family BH

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

q=0.0488. Age/stage adjustment identified a 17-gene FDR-significant family: BMS1, BRCA2, DIAPH3, EIF2B3, EIF4G1, HSP90AA1, KNL1, LRP1B, MCM4, PDSS1, PDZD2, RAD51, SEC63, SPECC1, VPS35, WT1, and ZP1. Among the primary five, HSP90AA1, SEC63, WT1, and ZP1 remained significant; WWOX retained the risk direction with BH q approximately 0.073. The primary and adjusted families were fixed before external outcomes were examined. Proportional-hazards diagnostics did not identify a family-level violation among the five primary genes. Only AFF3 and GALC were flagged after FDR correction across the 386-gene primary family, and neither belongs to the five-gene prognostic family; no adjusted-family gene was flagged. These diagnostics reduce concern about a detected gross proportional-hazards violation but do not prove that biological effects are perfectly time-invariant. The five primary TCGA-BRCA discoveries all had BH q=0.0488. Their raw Wald p values, effect estimates, confidence intervals, and proportional-hazards diagnostics were calculated within the complete declared 386-gene family rather than within a post-selected subset. The full 386-gene primary and adjusted Cox/PH table is retained as a machine-readable reproducibility artifact so that the near-threshold discoveries and all non-selected genes can be audited if requested by an editor or reviewer.

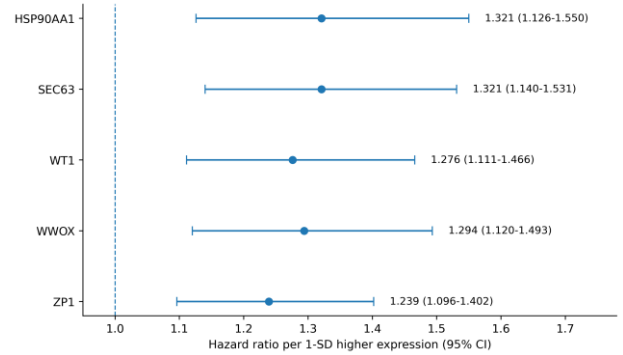

Fig. 5. TCGA-BRCA primary Cox estimates for the five FDR-significant prognostic genes. Hazard ratios are expressed per one standard deviation higher within-cohort gene expression.

TABLE III. TCGA-BRCA PRIMARY, AGE/STAGE-ADJUSTED, AND PROPORTIONAL-HAZARDS DIAGNOSTIC EVIDENCE FOR THE FIVE-GENE PROGNOSTIC FAMILY.

Gene	Primary HR per 1 SD (95% CI)	Primary BH q	Adjusted HR per 1 SD (95% CI)	Adjusted BH q	Primary PH BH q
HSP90AA1	1.321 (1.126-1.550)	0.0488	1.318 (1.114-1.560)	0.0391	0.9793
SEC63	1.321 (1.140-1.531)	0.0488	1.416 (1.205-1.665)	0.0049	0.8942
WT1	1.276 (1.111-1.466)	0.0488	1.292 (1.120-1.491)	0.0280	0.5596
WWOX	1.294 (1.120-1.493)	0.0488	1.252 (1.075-1.459)	0.0732	0.5405
ZP1	1.239 (1.096-1.402)	0.0488	1.345 (1.188-1.523)	0.0011	0.2064

F. Gene-level external transportability was strongest in SCAN-B

SCAN-B provided the strongest gene-level replication. All five primary genes retained HR>1 and met the within-family FDR criterion: HSP90AA1 HR 1.234 (q=0.0001), SEC63 1.148 (q=0.0066), WT1 1.239 (q<0.0001), WWOX 1.180 (q=0.0015), and ZP1 1.140 (q=0.0066).

METABRIC strictly replicated WT1 (HR 1.071, q=0.0320) and WWOX (HR 1.130, q<0.0001); the other three genes remained direction-concordant. In GSE20685, all five HRs were above 1.0 but none passed within-family FDR. With 83 events on a legacy microarray platform, GSE20685 therefore contributed directional cross-platform evidence rather than strict primary replication.

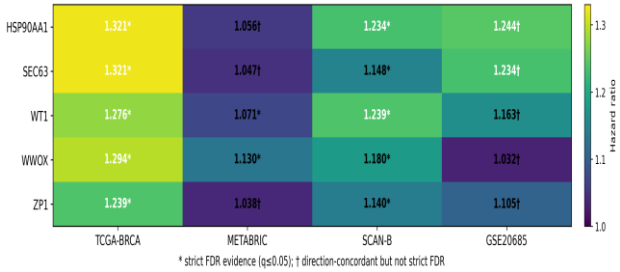

Fig. 6. Gene-level transportability of the five-gene primary family across TCGA-BRCA and three independent cohorts. An asterisk denotes strict within-family FDR evidence (q<=0.05); a dagger denotes direction concordance without strict FDR significance.

TABLE IV. GENE-LEVEL TRANSPORTABILITY OF THE PRIMARY FAMILY.

Gene	TCGA-BRCA	METABRIC	SCAN-B / GSE96058	GSE20685
HSP90AA1	HR 1.321; q=0.0488	HR 1.056; family q=0.0888; direction only	HR 1.234; family q=0.0001; replicated	HR 1.244; family q=0.1208; direction only
SEC63	HR 1.321; q=0.0488	HR 1.047; family q=0.1213; direction only	HR 1.148; family q=0.0066; replicated	HR 1.234; family q=0.1208; direction only
WT1	HR 1.276; q=0.0488	HR 1.071; family q=0.0320; replicated	HR 1.239; family q<0.0001; replicated	HR 1.163; family q=0.2480; direction only
WWOX	HR 1.294; q=0.0488	HR 1.130; family q<0.0001; replicated	HR 1.180; family q=0.0015; replicated	HR 1.032; family q=0.7761; direction only
ZP1	HR 1.239; q=0.0488	HR 1.038; family q=0.2324; direction only	HR 1.140; family q=0.0066; replicated	HR 1.105; family q=0.4758; direction only

G. External-only synthesis quantified pooled direction and heterogeneity

Across METABRIC, SCAN-B, and GSE20685, pooled random-effects HRs were above 1.0 for all five genes: HSP90AA1 1.153 (95% CI 0.887-1.500), SEC63 1.106 (0.913-1.340), WT1 1.146 (0.910-1.444), WWOX 1.136 (1.029-1.254), and ZP1 1.081 (0.915-1.276). Heterogeneity was gene-specific (I2: 76.2%, 53.1%, 72.1%, 0%, and 27.9%, respectively). WWOX had an unadjusted Hartung-Knapp p=0.0309, but no pooled test survived BH-FDR across the five-gene family (minimum q=0.1547).

Including TCGA-BRCA in a four-cohort sensitivity analysis again produced pooled HRs above 1 for all five genes (range 1.119-1.188), without five-test FDR significance (minimum q=0.0813). Every leave-one-external-cohort-out point estimate also remained above 1.0 (range 1.042-1.236), indicating stable direction but cohort-dependent magnitude and precision.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

TABLE V. EXTERNAL-ONLY RANDOM-EFFECTS SYNTHESIS OF THE FIVE-GENE PRIMARY FAMILY.

Gene	Pooled HR (95% CI)	I2 (%)	tau2	Q p	BH q
HSP90AA1	1.153 (0.887-1.500)	76.2	0.0076	0.0149	0.1828
SEC63	1.106 (0.913-1.340)	53.1	0.0031	0.1184	0.1828
WT1	1.146 (0.910-1.444)	72.1	0.0055	0.0278	0.1828
WWOX	1.136 (1.029-1.254)	0.0	0.0000	0.5060	0.1547
ZP1	1.081 (0.915-1.276)	27.9	0.0018	0.2498	0.1828

Note: Principal synthesis includes METABRIC, SCAN-B/GSE96058, and GSE20685 only; TCGA-BRCA is excluded because it was the discovery cohort. BH q values correct the five pooled gene tests.

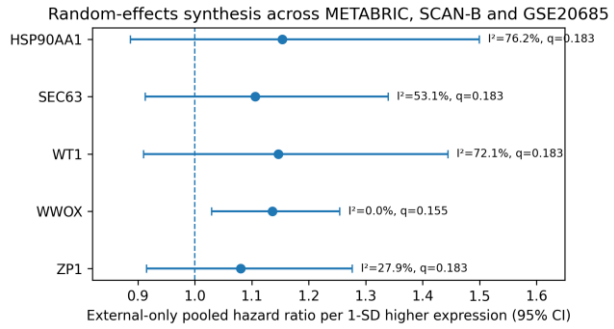

Fig. 7. External-only random-effects synthesis of the primary five-gene family across METABRIC, SCAN-B/GSE96058, and GSE20685. Points show pooled hazard ratios per one within-cohort expression standard deviation with 95% Hartung-Knapp intervals; I2 and five-gene-family BH q values are shown for context. This secondary synthesis did not alter prespecified cohort-level replication calls.

H. Family-level external replication was heterogeneous but reproducible

Family-level results extended the gene-level pattern. METABRIC replicated 2/5 primary genes, 0/5 after primary-family adjustment, and 6/16 assay-evaluable genes from the adjusted 17-gene family; 82/374 genes were significant in the broader supportive analysis, and the structurally non-evaluable EIF2B3 was not replaced. SCAN-B replicated 5/5 primary genes, 3/5 after adjustment, and 13/17 genes in the adjusted family, with 219/385 supportive associations. GSE20685 replicated 0/5 primary genes but preserved direction for 5/5; its adjusted family replicated 3/17 with direction preserved for 17/17, and 7/376 genes passed the supportive FDR analysis. No cohort was repaired by post-hoc gene, probe, covariate, or cut-point changes.

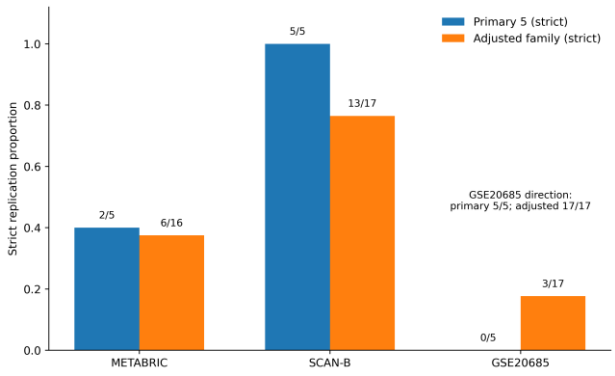

Fig. 8. Strict family-level replication proportions in the three independent cohorts. GSE20685 additionally retained complete direction concordance for the primary five-gene family (5/5) and adjusted 17-gene family (17/17).

TABLE VI. INDEPENDENT EXTERNAL OVERALL-SURVIVAL VALIDATION AND REPRODUCIBILITY AUDIT.

Cohort	Locked cohort	Strict replication summary	Direction/supportive evidence	Audit
METABRIC	Primary 1,980/1,143 events; adjusted 1,454/821	Primary 2/5; adjusted-primary 0/5; adjusted family 6/16	Supportive FDR 82/374; all primary HRs >1	12,488/12,488 PASS
SCAN-B / GSE96058	Primary 3,273/336; adjusted 3,149/316	Primary 5/5; adjusted-primary 3/5; adjusted family 13/17	Supportive FDR 219/385; all primary HRs >1	12,798/12,798 PASS
GSE20685 / GPL570	Primary 327/83; adjusted 325/83	Primary 0/5; adjusted family 3/17	Primary direction 5/5; adjusted direction 17/17; supportive FDR 7/376	15,387/15,387 PASS

I. Independent audits reproduced the evidence without scientific reselection

Independent audits reproduced the declared computational checks: 95/95 for biomarker discovery, 231/231 for Reactome, 299/299 for protein-domain analysis, 8,239/8,239 for STRING, 256/256 for TCGA-BRCA survival, 12,488/12,488 for METABRIC, 12,798/12,798 for SCAN-B, and 15,387/15,387 for GSE20685. These are provenance and calculation checks, not additional biological observations or replication cohorts.

IV. DISCUSSION

A. Principal findings

The main contribution is the separation of claims that are often collapsed in high-dimensional biomarker studies. Recurrence was treated as candidate evidence, not as proof of biomarker status; biological annotation was used to interpret a fixed set rather than select it; survival association was tested only after stabilization; and external cohorts were allowed to disagree. This ordering yielded a compact internally stable genomic set and a five-gene expression-associated prognostic family while preserving the substantial heterogeneity seen across cohorts. The external results are informative precisely because they are uneven. SCAN-B supplied strong strict replication, METABRIC supported a smaller subset, and GSE20685 preserved direction without meeting the primary FDR criterion. Taken together, these cohorts support transportability of effect orientation more strongly than they support a single common effect size. Keeping direction concordance separate from strict replication makes that distinction explicit rather than forcing all cohorts into a binary success/failure label.

B. Stability filtering is more defensible than recurrence alone

The reduction from 5,119 consensus candidates to 398 internally stable entities illustrates why recurrence alone was insufficient. Patient-group analysis limits the influence of repeated observations, and deterministic bootstrap persistence asks whether support survives perturbation of the biological sampling unit. Stability here is therefore an internal robustness designation, not an estimate of population prevalence. Selection governance is equally important. TRAIN/VALIDATION attribution intensity was only one support axis; bootstrap stability was mandatory, at least two support axes were required, and final test predictions and explanations were excluded. Reactome, VEP, STRING, and survival outcomes entered only after biomarker membership had been set. This prevents attractive downstream biology or prognosis from becoming a retrospective selection criterion.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

C. *Pathway, domain, and network results provide convergence rather than independent validatione*

Reactome consistently pointed to DNA-repair and related signaling programs across two declared backgrounds. The value of this result lies less in the number of significant terms than in their persistence when the enrichment universe changes. Because many Reactome pathways are nested and share genes, the 24 cross-background results should be read as overlapping biological themes rather than independent mechanisms. STRING added a complementary structural view: the robust pathway-supported genes were concentrated in the main functional component, and the induced subnetwork remained densely connected. Network evidence was deliberately downstream of biomarker selection, so connectivity could contextualize the stable set but could not create it. The protein-domain findings reinforce the same need for restraint. Three database annotations passed domain-level FDR, yet they were largely driven by the same two stable transcript observations. Treating these as overlapping annotations, rather than three separate mechanisms, better reflects the information content of the data. Permutation testing strengthened the network interpretation without converting it into mechanistic evidence. The initial size-matched null showed marked excess connectivity; exact-degree matching reduced that contrast, as expected, but significant internal-edge enrichment persisted in both network definitions. Degree matching addresses one important source of network bias, although STRING knowledge density and literature coverage remain potential contributors.

D. *Interpretation of the five-gene primary family*

The five genes do not form a single canonical pathway signature. HSP90AA1 and WT1 have prior breast-cancer prognostic support [20,21], whereas WWOX is more commonly discussed as a tumor suppressor at the FRA16D fragile site [22,23]. The positive transcript-level WWOX association observed here should therefore be interpreted from the cohort model itself rather than inferred from its canonical gene label. SEC63 and ZP1 are more exploratory in this setting. SEC63 regulates endoplasmic-reticulum protein translocation, and an IRE1alpha-SEC63-ACLY axis has been linked to stress adaptation, lipid-metabolic reprogramming, metastasis, and adverse prognosis in hepatocellular carcinoma [24]. This offers a plausible cancer-stress context, not a breast-cancer mechanism. ZP1 is best known as a zona-pellucida structural glycoprotein [25], although a 2026 triple-negative breast-cancer study included ZP1 in a three-gene prognostic model derived from TCGA/GEO data [28]. That precedent is useful but remains computational and subtype-limited. Both genes therefore warrant biological follow-up rather than mechanistic inference from association alone.

E. *Why heterogeneous external replication is scientifically useful*

External cohorts differed in the information available for survival estimation: METABRIC had many more deaths than SCAN-B, whereas GSE20685 had only 83 events and used a legacy microarray platform. Within-cohort standardization puts gene effects on a comparable per-SD scale, but it cannot remove differences in assay behavior, case mix, treatment era, follow-up, or endpoint capture. The high I2 values for HSP90AA1 and WT1 are therefore evidence against assuming one precise common effect across settings. Molecular-subtype composition, stage, treatment exposure, and measurement platform are plausible contributors, but post-hoc subtype interaction tests or meta-regression were not introduced after observing heterogeneity. Resolving those sources will require harmonized covariates and more independent cohorts. For this reason, external validation was judged by the prespecified combination of direction and within-family FDR rather than by numerical agreement with the TCGA hazard ratio. GSE20685's complete directional concordance is weaker evidence than strict replication, but retaining it as a separate evidence layer is more informative than either discarding the cohort or modifying the hypothesis until significance appears.

F. *Secondary pooled synthesis complements rather than replaces replication*

The random-effects synthesis adds a useful summary but does not replace cohort-specific validation. Pooled effects remained on the TCGA risk side of one, yet none survived five-gene FDR correction and heterogeneity was substantial for some genes. With only three external cohorts, both tau2/I2 and Hartung-Knapp intervals are imprecise. The meta-analysis is therefore best read as a sensitivity analysis of transportability, not as a new tier of confirmation.

G. *Limitations and interpretation boundaries*

The main limitation is the size and composition of the sequencing discovery cohort: 50 patient groups are sufficient for an internally auditable stability analysis but not for population-wide prevalence claims. Bootstrap persistence cannot substitute for independent genomic recurrence in a larger sequencing cohort, and the optional TRAIN/VALIDATION attribution-support axis may favor genes linked to rows the upstream model considered influential. The paired-role analysis is also not a dedicated tumor-normal somatic-calling design. In addition, transferring gene identity from genomic recurrence to expression-based survival analysis does not demonstrate mutation-expression mediation. Reactome, VEP, and STRING are version- and knowledge-density dependent; exact-degree matching reduces degree-composition bias but cannot remove annotation bias. The prognostic evidence is observational and cohort-dependent. TCGA-BRCA served as discovery, while the external cohorts differed in expression platform, covariate availability, event count, treatment era, and endpoint capture; adjusted models therefore do not represent perfectly identical estimands. GSE20685 is particularly limited by its event count. With only three independent external cohorts, heterogeneity estimates remain uncertain and formal publication-bias assessment would not be informative. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

five-gene family should consequently be regarded as a prognostic hypothesis set rather than a deployable signature. The study includes no prospective validation, treatment-interaction analysis, decision-impact evaluation, or functional experiment, and it does not test transportability of the upstream variant-level classifier. Those questions require separate data and study designs.

H. Implications for biomarker research

Methodologically, the study argues for treating biomarker development as a sequence of distinct claims rather than a single score of 'evidence.' Internal stability, biological coherence, prognostic association, direction across cohorts, strict statistical replication, and clinical usefulness require different data and safeguards. A practical next step is therefore orthogonal genomic confirmation of the stable set together with prospective evaluation of the prognostic family, ideally using harmonized molecular-subtype and treatment information.

V. CONCLUSION

This study establishes a traceable path from recurrent genomic evidence to an externally tested expression-associated prognostic hypothesis. Patient-group stabilization reduced a large candidate space to 398 internally stable entities, while downstream pathway, domain, and network analyses provided biological context without feeding back into selection. TCGA-BRCA identified a five-gene primary prognostic family—HSP90AA1, SEC63, WT1, WWOX, and ZP1—which replicated most strongly in SCAN-B, partially in METABRIC, and directionally in GSE20685. Network sensitivity analysis and cross-cohort meta-analysis supported robustness while also exposing residual bias and heterogeneity. The appropriate conclusion is therefore not that a clinical signature has been validated, but that a prespecified evidence-governance strategy produced a reproducible, testable prognostic hypothesis that merits independent genomic, functional, and prospective evaluation.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Computation Laboratory of the Department of Mechanical Engineering, Engineering College, Ajmer for providing the high-end computational facility used in this research, procured under the World Bank-supported TEQIP-III project. The authors also acknowledge the Department of Computer Science & Engineering, Bhagwant University, Ajmer, for academic support; the investigators and participants who contributed data to the public sequencing cohort, TCGA-BRCA, METABRIC, SCAN-B/GSE96058, and GSE20685; and the maintainers of Reactome, STRING, Ensembl/VEP, ClinVar, and associated public genomic repositories and open-source computational tools.

REFERENCES

[1] D. Sisodia, V. K. Sharma, R. Joshi, H. Arya, and T. K. Bhatt, "Developing an automated computational genomics pipeline for breast cancer detection using NGS," Research Plateau Current Trends in Engineering and Technology, vol. 4, pp. 36–45, 2025, presented at the 3rd International Conference on Recent Trends in Materials Science & Devices (ICRTMD-2025). Online Available at: https://researchplateau.com/uploads/reasearchpapers/1751596826.pdf

[2] D. Sisodia and V. K. Sharma, "A comprehensive literature survey on machine learning-based mutation prediction in breast cancer using next-generation sequencing data," International Education and Research Journal (IERJ), vol. 12, no. 04, 2026. doi: 10.5281/zenodo.20021195.

[3] D. Sisodia and V. K. Sharma, "A Hybrid Explainable Multi-Omics Machine Learning Framework For Breast Cancer Mutation Prediction And Clinical Risk Stratification Using Tcga-Brca Data". International Education and Research Journal (IERJ), 12(05), 299–313. https://doi.org/10.5281/zenodo.20484976

[4] D. Sisodia and V. K. Sharma, "Computational formulation of explainable genomic AI for breast cancer risk prediction," in Programme and Abstracts, 6th IEEE-Sponsored International Conference on Emerging Trends in Networks and Computer Communications (ETNCC 2026), Windhoek, Namibia, Aug. 4–6, 2026, Proceedings publication pending, Online Abstract Available: https://etncc.nust.na/sites/default/files/2026-08/ETNCC-2026-Progrm-4-6-August-2026-V5_0.pdf , Paper ID 516, p. 95.

[5] D. Sisodia and V. K. Sharma, "A Leakage-Controlled Hierarchical Evaluation Framework for Breast Cancer Variant Pathogenicity Prediction with Locked-Test Validation", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 15, Issue 09 , September – 2026, doi: 10.5281/zenodo.22654096

[6] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61-70. doi:10.1038/nature11412.

[7] C. Curtis et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346-352. doi:10.1038/nature10983.

[8] C. Brueffer et al. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network-Breast initiative. JCO Precision Oncology. 2018;2:1-18. doi:10.1200/PO.17.00135.

[9] K.-J. Kao, K.-M. Chang, H.-C. Hsu, and A. T. Huang. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer. 2011;11:143. doi:10.1186/1471-2407-11-143.

[10] M. J. Landrum et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Research. 2016;44(D1):D862-D868. doi:10.1093/nar/gkv1222.

[11] P. Milacic et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Research. 2024;52(D1):D672-D678. doi:10.1093/nar/gkad1025.

[12] W. McLaren et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17:122. doi:10.1186/s13059-016-0974-4.

[13] D. Szklarczyk et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research. 2023;51(D1):D638-D646. doi:10.1093/nar/gkac1000.

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008.

[15] D. R. Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B. 1972;34(2):187-220.

[16] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457-481.

[17] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B. 1995;57(1):289-300.

[18] L. M. McShane et al. Reporting recommendations for tumor marker prognostic studies (REMARK). Journal of the National Cancer Institute. 2005;97(16):1180-1184. doi:10.1093/jnci/dji237.

[19] G. S. Collins et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

[20] S. A. Alsaeed et al. Prognostic significance of heat shock protein 90AA1 (HSP90alpha) in invasive breast cancer. Journal of Clinical Pathology. 2022;75(4):263-269. doi:10.1136/jclinpath-2020-207106.

[21] Y. Miyoshi et al. High expression of Wilms tumor suppressor gene predicts poor prognosis in breast cancer patients. Clinical Cancer Research. 2002;8(5):1167-1171.

[22] K. Pospiech, E. Pluciennik, and A. K. Bednarek. WWOX tumor suppressor gene in breast cancer, a historical perspective and future directions. Frontiers in Oncology. 2018;8:345. doi:10.3389/fonc.2018.00345.

[23] F. Ge et al. WWOX suppresses KLF5 expression and breast cancer cell growth. Chinese Journal of Cancer Research. 2014;26(5):511-516. doi:10.3978/j.issn.1000-9604.2014.09.03.

[24] C. Hu et al. Activation of ACLY by SEC63 deploys metabolic reprogramming to facilitate hepatocellular carcinoma metastasis upon endoplasmic reticulum stress. Journal of Experimental & Clinical Cancer Research. 2023;42:108. doi:10.1186/s13046-023-02656-7.

[25] S. V. Prasad, S. M. Skinner, C. Carino, N. Wang, J. Cartwright, and B. S. Dunbar. Structure and function of the proteins of the mammalian zona pellucida. Cells Tissues Organs. 2000;166(2):148-164. doi:10.1159/000016730.

[26] W. Viechtbauer. Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics. 2005;30(3):261-293. doi:10.3102/10769986030003261.

[27] J. IntHout, J. P. A. Ioannidis, and G. F. Borm. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Medical Research Methodology. 2014;14:25. doi:10.1186/1471-2288-14-25.

[28] P. Zou, H. Xu, Y. Gao, W. Bai, X. Zhao, and S. Shao. A SUMOylation/immune-related gene signature predicts the prognosis and immunotherapy efficacy of patients with triple-negative breast cancer. PeerJ. 2026;14:e21139. doi:10.7717/peerj.21139.

[29] D. Schoenfeld. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239-241. doi:10.1093/biomet/69.1.239.

[30] J. P. T. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21(11):1539-1558. doi:10.1002/sim.1186.