Research Square

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

External Transportability and Tissue Context of Parsimonious Breast Tumour-Normal Transcriptomic Score

Subhajit Ghosh

subhajit.ghosh.academic@gmail.com

Siliguri, West Bengal https://orcid.org/0000-0002-2897-5956

Abhirup Saha

National Centre for Disease Control

Vetrivel Baskar

National Centre for Disease Control

Research Article

Keywords: Breast cancer, transcriptomics, external validation, transportability, tissue composition, elastic net, public microarray data

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11148420/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Bulk tumour-versus-normal breast transcriptomic classifiers commonly show high internal discrimination, but a high area under the curve may reflect the biology of tissue contrast, cohort construction or technical separation as much as a portable malignant-cell program. We asked whether a pre-specified, interpretable score trained in one public cohort could retain rank discrimination in a locked cohort while its underlying biology was examined in parallel.

Methods

Processed Affymetrix GPL570 data from GSE42568 were used for discovery and nested model development. GSE7904 was held as an external cohort. Within-study normalization, probe-to-gene collapse and gene intersection produced 22,737 shared genes. Differential expression and enrichment were conducted. The elastic-net logistic score was the primary model; feature selection and tuning were contained in discovery-cohort nested cross-validation.

Results

The discovery cohort had 1,213 upregulated and 1,369 downregulated genes at false-discovery rate < 0.05 and log2 fold change ≥ 1. Of 2,582 discovery differentially expressed genes, 1,729 (67.0%) had a concordant external direction; 563 met the stringent criterion with the same direction in both cohorts. Cell-cycle/mitotic programmes dominated the upregulated signal, while PPAR, lipid-metabolism and adipocyte-lipolysis terms dominated the downregulated signal. The 20-feature elastic-net score produced a nested discovery AUC of 0.933 (95% CI 0.830–0.997) and a locked external AUC of 0.970 (95% CI 0.900-1.000). External basal-like and non-BLC AUCs were 0.968 and 0.986, respectively; the BRCA1-associated subgroup was too small for credible estimation.

Conclusion

The score transports as a rank discriminator, but its uniformly negative coefficients and lipid/adipose-associated features indicate that it likely captures loss of normal breast-tissue programmes in addition to tumour biology. The evidence supports a contrast-aware transportability signal. Prospective cohorts with representative benign controls, unified preprocessing and formal calibration assessment are required in future works.

1. Introduction

Breast cancer is not a single transcriptomic entity, rather a molecularly and biologically heterogeneous disease. Its molecular state can be influenced by several factors, such as hormone receptor and HER2 status, underlying somatic drivers, patterns of copy-number variation, immune and stromal composition, and prior or ongoing treatment exposure (Harbeck et al., 2019). This heterogeneity is clinically important because the same expression pattern may mean different things in a resected tumour, a biopsy with benign tissue, or a population-level screening sample (Sung et al., 2021; Waks & Winer, 2019). The immediate attraction of a tumour-normal classifier is obvious, but its apparent accuracy is not automatically a measure of diagnostic validity, prognosis, therapeutic benefit, or patient-level risk.

Foundational expression profiling showed why this distinction matters. Perou et al. (2000) used cDNA microarrays to describe reproducible molecular portraits of breast tumours, and Sørlie et al. (2001) extended the approach by relating transcriptional subclasses to clinical outcome. Later integrated datasets confirmed that intrinsic subtype, copy-number change, and microenvironmental programmes coexist in the same specimen (Burstein et al., 2015; Cancer Genome Atlas Network, 2012; Curtis et al., 2012; Lehmann et al., 2011). Thus, bulk transcriptomics remains highly informative, but it is not a pure reading of malignant epithelial cells.

The clinical translation of breast-cancer expression signatures has followed a much more demanding route than demonstrating a high classifier area under the curve. The early 70-gene studies by van 't Veer et al. (2002) and van de Vijver et al. (2002), followed by the 21-gene assay work of Paik et al. (2004), linked a defined assay to a prognostic construct. Prospective and prospective-retrospective programmes then tested analytical performance, outcome association and treatment decision consequences (Bueno-de-Mesquita et al., 2007; Cardoso et al., 2016; Goldstein et al., 2008; Kalinsky et al., 2021; Simon et al., 2009; Sparano et al., 2015b; Sparano et al., 2018). Studies of late recurrence and formal biomarker-validation frameworks further show that a multigene signal becomes clinically useful only after its intended use, population, and decision context are fixed (Duffy et al., 2015; Hayes, 2015; Sestak & Cuzick, 2015; Sgroi et al., 2013).

A particular vulnerability of small, high-dimensional transcriptomic studies is that model development can reuse the same limited information at several stages. Ambroise and McLachlan (2002) demonstrated how selecting genes before performance estimation creates optimistic error rates, while Varma and Simon (2006) showed that cross-validation used for both selection and evaluation can be biased. More recent discussions of external validation make same point in a broader setting: a stable-looking model may still fail when sampling, measurement or biological context changes (Ho et al., 2020).

Bulk expression is a mixture measurement. Changes in immune cells, fibroblasts, vascular cells and adipose-rich normal tissue can alter gene abundance even when the transcriptional state of tumour cells has not changed equivalently. Newman et al. (2015) introduced CIBERSORT to estimate leukocyte fractions from bulk profiles; Aran et al. (2017) and Becht et al. (2016) later developed complementary cell-enrichment and abundance-estimation strategies. Similarly, Yoshihara et al. (2013) used expression-derived stromal and immune scores to characterise tumour purity and admixture.

The present study tries to addresses this gap in a deliberately narrow public-data setting. GSE42568 was used as the discovery cohort and GSE7904 as a locked external cohort, and normal-organelle samples that were biologically non-comparable were removed prior to modeling. These GEO datasets and their study context have been described previously (Clarke et al., 2013; Edgar et al., 2002; Richardson et al., 2006). We examined gene-direction replication, direction-specific biology, external rank discrimination, and the likely tissue provenance of final score features in parallel.

Our conceptual contribution is a contrast-aware transportability reading. Rather than treating an external AUC as a complete validation result, it asks three linked questions: does the sample ranking transport, does a non-model biological direction reproduce, and do the selected features have a plausible cellular or tissue context? A fourth question then follows: does the evaluated contrast resemble the intended clinical use? This is a conservative interpretive framework, and it aligns with the emphasis on transparency, risk of bias, and applicability in prediction-model research (Collins et al., 2015; Wolff et al., 2019).

2. Materials and Methods

2.1 Study design and cohorts

This was a secondary analysis of public, de-identified Gene Expression Omnibus datasets (Table 1). GSE42568 was fixed as the discovery/training cohort (104 primary tumours and 17 normal-breast samples), and GSE7904 was reserved as the locked external cohort (43 primary tumours and 7 eligible normal-breast controls). Twelve GSE7904 specimens annotated as normal organelle were excluded before model fitting because they are biologically non-comparable to normal breast. Both studies used the Affymetrix Human Genome U133 Plus 2.0 platform (GPL570). GEO was established by Edgar et al. (2002) as a public repository for the functional genomics data; the cohort-level source publications were used to check study context (Clarke et al., 2013; Richardson et al., 2006).

Table 1
Eligible samples and analysis roles. The external cohort was not utilized for discovery-stage feature selection, tuning, or model selection.

Cohort	Role	Tumours	Eligible normal breast	Excluded	Platform
GSE42568	Discovery / training	104	17	0	GPL570
GSE7904	Locked external test	43	7	12 normal-organelle	GPL570
Combined eligible analysis	Gene-direction context	147	24	12	22,737 shared genes

2.2 Preprocessing and quality assessment

Series-matrix expression values were inspected for scale and processed within study. Probe identifiers were mapped to gene symbols and multiple probes were collapsed by the median, yielding a common 22,737-gene matrix. We did not jointly normalise the cohorts or use cross-cohort batch correction for the locked evaluation. This choice preserves the external-domain shift rather than allowing information from the test cohort to reshape the training representation. Quantile normalisation was treated as a within-cohort distributional procedure, not as raw-CEL RMA reprocessing. The rationale follows the array-preprocessing work of Bolstad et al. (2003) and Irizarry et al. (2003); batch adjustment methods such as ComBat are valuable for pooled inference but were not used to improve external test performance (Johnson et al., 2007; Leek & Storey, 2007).

Sample-wise boxplots, within-study Pearson correlation heatmaps, and principal component analysis (PCA) were examined prior to any cross-study batch adjustment as quality and domain-shift diagnostics. Correlation heatmaps were not interpreted as a representation of gene networks, likewise, PCA was not considered a biological separation analysis. Principal components provide a compact representation of the dominant variance structure rather than a causal decomposition (Jolliffe & Cadima, 2016). The study-level PCA therefore only indicates that the cohort origin remains a major source of variation.

2.3 Differential expression and enrichment

Within each cohort, Welch tests compared eligible tumours with eligible normal-breast controls. The unequal-variance test was selected because the group sizes were strongly unbalanced (Welch, 1947). Benjamini-Hochberg false-discovery adjustment was applied (Benjamini & Hochberg, 1995), and exploratory differentially expressed genes were defined using adjusted $P < 0.05$ and $|\log2 \text{ fold change}| \geq 1$. Cross-cohort replication was summarised by unthresholded sign agreement, sign agreement among discovery DEGs, and stringent two-cohort concordance. This was not a formal meta-analysis, and the external cohort was not used to choose gene thresholds. The differential-expression framing is consistent with established limma guidance for microarray-scale data (Ritchie et al., 2015).

Direction-specific gene lists were explored with Enrichr. Interpretation was organised using Gene Ontology, KEGG and MSigDB resources. Gene Ontology was developed by Ashburner et al. (2000) to standardise functional annotation, whereas Kanehisa and Goto (2000) established KEGG pathway maps. Subramanian et al. (2005) introduced gene-set enrichment analysis; MSigDB collections were subsequently expanded by Liberzon et al. (2011) and Liberzon et al. (2015). Enrichr and clusterProfiler provide reproducible interfaces for such over-representation analyses (Kuleshov et al., 2016; Yu et al., 2012). These analyses remain hypothesis-generating because enrichment may reflect changing cellular composition as well as tumour-intrinsic transcription.

2.4 Model development and locked evaluation

The primary model was a logistic elastic-net score. In the discovery cohort, imputation, standardisation, univariate feature filtering and hyperparameter choice were contained inside nested stratified cross-validation. The final primary model used $C = 0.01$, l1 ratio = 0.25 and retained 20 features. Zou and Hastie (2005) described the elastic-net penalty as a combined L1/L2 regulariser that can stabilise selection

among correlated predictors. Standard logistic regression, radial-basis support-vector machine, random forest, extra trees, gradient boosting and Gaussian naïve Bayes were pre-specified sensitivity architectures, following the corresponding algorithmic formulations (Breiman, 2001; Cortes & Vapnik, 1995; Friedman, 2001; Geurts et al., 2006; Hand & Yu, 2001). Containment of all learning steps was used because feature selection performed before cross-validation can produce optimistic error estimates (Ambroise & McLachlan, 2002; Varma & Simon, 2006).

The final discovery-fitted score was evaluated once in the locked cohort. ROC AUC and average precision were estimated with 2,000 stratified bootstrap resamples. AUC summarises rank discrimination across thresholds (Hanley & McNeil, 1982), whereas precision-recall plots can be especially sensitive to a case-enriched sample fraction (Saito & Rehmsmeier, 2015); bootstrap resampling was used to represent sampling uncertainty (Efron & Tibshirani, 1994). We did not report a clinical calibration curve, decision curve, probability threshold, positive predictive value or screening sensitivity/specificity. The external series is a retrospective, deliberately case-enriched cohort with 43 cases and 7 controls, so its observed case fraction is not an intended-use prevalence. Reporting and bias considerations were guided by TRIPOD, PROBAST and external-validation literature (Collins et al., 2015; Ho et al., 2020; Wolff et al., 2019).

2.5 External subtype checks

The locked score was applied descriptively to basal-like (n = 18) and non-BLC (n = 21) external tumour subsets using the same seven controls. Four BRCA1-associated tumours were available but were not assigned a subgroup AUC. Subgroup confidence intervals are correlated because the normal controls are shared. Figure 1 depicting the complete study design.

3. Results

3.1 Quality assessment shows within-study consistency and cross-study separation of domains

The distributions of the processed probe-expression values were broadly comparable within each study, although differences in their centers and spreads were observed (Fig. 2A). Principal components computed prior to any batch correction revealed a strong separation between GSE42568 and GSE7904 (Fig. 2B). Such separation is expected for independently generated public cohorts and is not itself a defect; however, it confirms that study origin is a transportability challenge. This is why the external data were not pooled into discovery-stage preprocessing (Bolstad et al., 2003; Johnson et al., 2007; Leek & Storey, 2007). The correlation structure within each study is reported in Supplementary Figures S1-S2.

3.2 Tumour-normal changes show partial cross-cohort direction replication

The discovery cohort produced 2,582 DEGs: 1,213 upregulated and 1,369 downregulated. Under the same threshold, the external cohort produced 570 upregulated and 1,289 downregulated genes (Fig. 3; Table 2). Across 22,875 discovery genes, 11,893 (52.0%) had the same unthresholded external direction. Among the discovery DEGs, 1,729/2,582 (67.0%) had the same external direction; 563 met the stringent criteria with concordant direction in both cohorts (162 upregulated and 401 downregulated). Across 22,586 finite paired fold changes, the unselected Pearson correlation was r = 0.268. Detailed results were provided as supplementary file 1 and 2.

Table 2
Summary of the cross-cohort direction-replication study.

Measure	Result	Importance
Discovery DEGs	1,213 up; 1,369 down	FDR < 0.05 and log2FC ≥ 1
External DEGs	570 up; 1,289 down	FDR < 0.05 and log2FC ≥ 1
All-gene sign agreement	11,893 / 22,875 (52.0%)	Unthresholded direction comparison
Discovery-DEG sign agreement	1,729 / 2,582 (67.0%)	Post hoc directional replication
Stringent concordant DEGs	563 (162 up; 401 down)	Significant and same direction in both cohorts
Unselected effect correlation	Pearson r = 0.268	22,586 paired finite fold changes

3.3 Directional biology is compatible with proliferation and loss of normal metabolic tissue programmes

Discovery upregulated genes were enriched for mitotic sister-chromatid segregation (adjusted P = 1.90×10^{-7}; odds ratio 8.86), mitotic spindle organisation (adjusted P = 1.90×10^{-7}; odds ratio 10.71) and cell-cycle terms. Downregulated genes were enriched for broad metabolism (adjusted P = 5.21×10^{-17}; odds ratio 3.12), PPAR signalling (adjusted P = 2.77×10^{-13}; odds ratio 16.33), regulation of lipolysis in adipocytes (adjusted P = 1.40×10^{-8}; odds ratio 14.54) and lipid metabolism. This directional pattern is compatible with proliferating tumour tissue plus relative depletion of normal breast/adipose-associated transcription (Becht et al., 2016; Keren et al., 2018; Newman et al., 2015; Ståhl et al., 2016). Detailed results were provided as supplementary file 3 and 4.

The observed biology fits a general cancer framework in which proliferative capacity, metabolic plasticity, cellular differentiation and microenvironmental organisation all contribute to the bulk transcriptome (Hanahan & Weinberg, 2011; Hanahan, 2022; Pavlova & Thompson, 2016). In the present contrast, normal-tissue programmes may be as informative as tumour programmes.

3.4 The primary elastic-net score preserves rank discrimination in the locked cohort

The pre-specified 20-feature elastic-net model achieved a nested discovery AUC of 0.933 (95% CI 0.830–0.997) and a locked external AUC of 0.970 (95% CI 0.900-1.000) (Table 3). External average precision was 0.995 (95% CI 0.982-1.000), but this is a rank-based metric in a case-enriched cohort (case fraction 0.86) and is not a clinical positive predictive value. Logistic regression showed external AUC 0.963 and RBF SVM 0.970, whereas gradient boosting transported poorly (AUC 0.593) despite a strong nested discovery AUC (0.965). The contrast illustrates why discrimination, calibration and clinical utility must be separated when evaluating a prediction model (Collins et al., 2015; Hanley & McNeil, 1982; Saito & Rehmsmeier, 2015; Wolff et al., 2019). Detailed results were provided as supplementary file 5 and 6.

Table 3
Discrimination results. Values are AUC or average precision with bootstrap 95% confidence intervals. The sensitivity models are not competing clinical candidates; they serve as a transport stress test.

Model	Nested CV AUC	Locked external AUC	External average precision
Elastic-net logistic (primary)	0.933 (0.830–0.997)	0.970 (0.900-1.000)	0.995 (0.982-1.000)
RBF support-vector machine	0.959 (0.912–0.995)	0.970 (0.900-1.000)	0.995 (0.983-1.000)
Logistic regression	0.913 (0.807–0.999)	0.963 (0.880-1.000)	0.993 (0.979-1.000)
Extra trees	0.972 (0.934–0.999)	0.917 (0.731-1.000)	0.981 (0.940-1.000)
Random forest	0.979 (0.950–0.999)	0.900 (0.711-1.000)	0.976 (0.931-1.000)
Gaussian naïve Bayes	0.891 (0.746–0.994)	0.844 (0.661–0.967)	0.964 (0.921–0.993)
Gradient boosting	0.965 (0.913–0.999)	0.593 (0.535–0.651)	0.886 (0.870–0.902)

3.5 Score composition and external subtype checks

Every coefficient in the fitted primary score was negative for the tumour class after standardisation. The largest absolute coefficients were IGSF10 (− 0.082), TMEM132C (− 0.069), ALDH1L1 (− 0.061), LOC284825 (− 0.058) and ACVR1C (− 0.056). The presence of ACSM5, GPD1, CIDEA, PDE3B, LYVE1 and ACADL is consistent with the downregulated metabolic/adipose/tissue-context signal. Elastic-net coefficients are conditional effects within a regularised multivariable model; they neither rank causal importance nor establish independent biomarker utility, particularly when correlated features can act as

proxies for tissue composition (Ambroise & McLachlan, 2002; Varma & Simon, 2006; Zou & Hastie, 2005).

Among external tumours, the primary score had AUC 0.968 (95% CI 0.881-1.000) in 18 basal-like tumours and AUC 0.986 (95% CI 0.939-1.000) in 21 non-BLC tumours, both against the same seven normal controls (Table 4). Four BRCA1-associated tumours were available; no AUC was estimated. These point estimates are encouraging, but their precision is limited and the shared controls prevent independent subgroup interpretation.

Table 4
Exploratory external subtype checks. Each comparison reuses the same seven normal controls; confidence intervals are therefore not independent.

External subgroup	Tumours	Controls	External AUC (95% CI)
Basal-like	18	7 shared	0.968 (0.881-1.000)
non-BLC	21	7 shared	0.986 (0.939-1.000)
BRCA1-associated	4	7 shared	-

4. Discussion

This study shows that an intentionally small elastic-net score can retain high rank discrimination when transported from GSE42568 to the eligible GSE7904 tumour-normal contrast. The result is informative, but only if it is read together with the score's biological composition and the sampling structure of the external cohort. A model with strong internal discrimination was not automatically externally robust: gradient boosting fell to modest discrimination, while the more regularised elastic-net and RBF SVM retained higher AUC. This discordance is a practical reminder that algorithm ranking inside one cohort cannot certify portability (Ambroise & McLachlan, 2002; Ho et al., 2020; Varma & Simon, 2006; Zou & Hastie, 2005).

The biological qualifier is that the replicated transcriptome was biphasic. Mitotic programmes increased in tumour samples, whereas PPAR, lipolysis and lipid-metabolism programmes were reduced. In a bulk breast specimen, this is compatible with malignant proliferation together with loss of normal breast, adipose and stromal representation. It does not distinguish a tumour-cell metabolic switch from a changing cellular mixture. Deconvolution work, multiplexed imaging and breast-cancer single-cell atlases have made this point increasingly clear: composition is both biology and a possible statistical shortcut (Aran et al., 2017; Azizi et al., 2018; Becht et al., 2016; Bindea et al., 2013; Keren et al., 2018; Newman et al., 2015; Newman et al., 2019; Savas et al., 2018; Thorsson et al., 2018; Wagner et al., 2019; Wu et al., 2021; Yoshihara et al., 2013).

The all-negative score supports the same interpretation. Conditional on the other retained features, lower expression of all 20 genes increased the model's tumour-class score. Several genes have adipose, metabolic, lymphatic or broad tissue-maintenance context. The current result is therefore best named a breast-tumour versus available-normal-breast transcriptomic contrast, rather than a twenty-gene tumour-cell diagnostic mechanism. This is not a semantic distinction: it determines whether the next experiment should prioritise tumour-cell perturbation, cell-type deconvolution, spatial localisation, representative benign controls or all four.

Table 5 (A, B, C) places each final score feature beside conservative earlier cancer evidence. It is not a systematic gene-by-gene review and does not validate an individual biomarker. The table separates direct breast-cancer evidence from cross-cancer mechanistic context and from genes for which no sufficiently direct evidence was retained. This avoids a common overreach in transcriptomic modelling-equating feature selection with biomarker discovery. In particular, ALDH1L1 has a stronger mechanistic cancer literature than several other features, while FHL1, PGM5-AS1 and LYVE1 have prior cancer reports that do not demonstrate the same biology in the present tumour-normal comparison.

Table 5

A. Final elastic-net score features as candidate biomarkers with prior cancer evidence. Feature selection does not establish individual diagnostic or therapeutic utility.

Candidate score feature	Coefficient	Earlier study and cancer setting	Earlier reports
IGSF10	−0.0822	Ewing sarcoma	A recent functional study identified IGSF10 as a RET antagonist in Ewing sarcoma (Jayabal et al., 2025).
TMEM132C	−0.0694	-	--
ALDH1L1	−0.0606	Pan-cancer review and liver-carcinogenesis model	ALDH1L1 loss has been linked to folate-metabolic rewiring and tumour growth in cancer models (Krupenko & Krupenko, 2018; Krupenko & Horita, 2019; Krupenko & Krupenko, 2019; Krupenko et al., 2021; Tsybovsky et al., 2022).
LOC284825	−0.0580	-	-
ACVR1C	−0.0563	-	-
HSPB7	−0.0490	-	-
PGM5-AS1	−0.0481	Prostate, colorectal and bladder cancer	Independent lncRNA studies reported tumour-suppressive associations in prostate, colorectal and bladder cancer models (Du & Gao, 2021; Duan et al., 2022; Zhou et al., 2020).

Table 5

B. Candidate score features with prior cancer evidence continued.

Candidate score feature	Coefficient	Earlier study and cancer setting	What earlier work reported
STX11	−0.0473	-	-
FHL1	−0.0456	Breast cancer cell model	FHL1 was reported to interact with oestrogen receptors and regulate breast-cancer cell growth (Ding et al., 2011).
ACSM5	−0.0419	-	-
GPD1	−0.0412	Castration-resistant prostate cancer	GPD1 has been linked to lipid-droplet and ferroptosis-resistance context in prostate cancer (Liu et al., 2024).
CIDEA	−0.0408	-	-
ASPA	−0.0360	Gastric cancer	ASPA expression was associated with tumour development and immune infiltrates in gastric cancer (Han et al., 2023).
ADH1C	−0.0358	Breast-cancer susceptibility meta-analysis	A pooled study found no association between ADH1C genotype and breast-cancer susceptibility in a Caucasian population (Wang et al., 2012).

Table 5

C. Candidate score features with prior evidence of cancer continued.

Candidate score feature	Coefficient	Earlier study and cancer setting	What earlier work reported
AQPEP	−0.0317	-	-
GLYAT	−0.0282	-	-
PDE3B	−0.0275	cAMP and adipocyte-lipolysis context	The gene fits the enriched lipolysis context, but pathway enrichment does not establish PDE3B causality.
LYVE1	−0.0247	Murine breast cancer and mammary stroma	LYVE1-positive macrophages were implicated in chemotherapy response in murine breast cancer and in mammary stromal regulation of tumour growth (Anstee et al., 2023; Elfstrum et al., 2024).
ACADL	−0.0227	Hepatocellular carcinoma	Bulk and single-cell work linked extracellular-matrix mechanics to lipid reprogramming through a YAP/TEAD4/ACADL axis in hepatocellular carcinoma (Cai et al., 2023).
SLC19A3	−0.0088	-	-

The study also offers a contrast-aware transportability framework to interpret public-cohort classifiers (Table 6). First, external rank discrimination asks if the ordering is preserved outside the training sample. Second, direction replication asks whether a non-model gene-level biological signal recurs. Third, feature provenance asks whether selected genes plausibly represent tumour cells, normal tissue, immune/stromal cells or mixtures. A fourth, non-scoring axis is intended-use alignment: does the evaluated contrast resemble the clinical or research decision being contemplated? Thus the claim should advance only when these dimensions are coherent (Collins et al., 2015; Ho et al., 2020; Wolff et al., 2019).

Table 6
Contrast-aware transportability framework. This conceptual synthesis supports interpretation of public-cohort classifier reports; it is not a clinical decision rule.

Dimension	Question	What this study supports	What is still required
External discrimination	Does sample ordering transport?	High external AUC for the primary score in 43 tumours and 7 controls.	Larger independent cohorts; pre-specified intended use; prospective timing.
Gene-direction replication	Does bulk biology recur?	Partial sign agreement and 563 stringent concordant DEGs.	Meta-analysis, raw-data harmonisation and covariate-aware models.
Feature provenance	Which cells or tissues could generate the signal?	Lipid/adipose/lymphatic context is plausible.	Single-cell or spatial validation, deconvolution, histology and tumour-purity analysis.
Intended-use alignment	Does the tested contrast match a real decision context?	Not evaluated; the test set is retrospective and case enriched.	Representative benign controls, prevalence, calibration, utility and prospective validation.

This framing connects the present findings to the evidentiary standard reached by established prognostic assays. Those assays moved through analytical validation, outcome association, prospective-retrospective studies, treatment-benefit assessment and decision trials rather than relying on a single retrospective external AUC. The clinical literature includes assay-development and community-feasibility studies, prospective validation, randomised decision trials, biomarker-guideline work and long-term endocrine-therapy evidence (Bartlett et al., 2013; Davies et al., 2013; Drukker et al., 2013; Early Breast Cancer Trialists' Collaborative Group (EBCTCG) et al., 2011; Early Breast Cancer Trialists' Collaborative Group (EBCTCG) et al., 2012; Hammond et al., 2010; Hayes & Padnos, 2011; Munoz et al., 2014; Siegel et al., 2014; Simon et al., 2009; Sparano & Paik, 2008; Sparano et al., 2015a; Tang et al., 2011; Wolff et al., 2013; Wong et al., 2012).

However, the present study has none of the prospective outcome, treatment-benefit or representative-control evidence required for that level of translation. A stronger next study would recruit consecutive participants under a declared use case, such as research triage or pathology support, and include

normal breast, benign disease, inflammatory lesions and relevant tumour subtypes (Collins et al., 2015; Vickers & Elkin, 2006; Wolff et al., 2019). The molecular workflow should be fixed before testing, ideally with raw-data harmonisation or a clinically deployable assay. Cell-type provenance needs to be measured with histology, spatial transcriptomics or matched single-cell references (Keren et al., 2018; Ståhl et al., 2016; Wagner et al., 2019; Wu et al., 2021).

5. Conclusions

A 20-feature elastic-net score trained in GSE42568 retained high rank discrimination in the eligible GSE7904 tumour-normal comparison, including the analysable basal-like and non-BLC subgroups. Its supporting transcriptome combined proliferative programmes with reduced normal metabolic or adipose-associated expression. Therefore, the result is plausible as a transportability pilot. The immediate scientific value lies in the contrast-aware interpretation: external AUC should be read together with replication biology and tissue-composition sensitivity.

Declarations

6.2 Generative AI statement

Generative artificial intelligence was used to assist with language editing, improving sentence clarity, organising the manuscript structure, and preparing preliminary drafts of selected text sections. All scientific content, literature selection, interpretation, data presentation, and final revisions were independently reviewed, verified, and approved by the authors. Generative AI was not used to generate, manipulate, or analyse research data, images, or results.

6.4 Declaration of competing interests

The authors have no competing interests.

6.5 CRediT authorship contribution statement

SG: Conceptualization, methodology, literature review, investigation, data curation, writing - original draft, writing - review and editing, visualization, supervision, project administration, and correspondence. AS: Literature review, data curation, writing - original draft, writing - review and editing, and visualization. VB: Literature review, data curation, writing - original draft, writing - review and editing, and visualization.

6.7 Ethics statement

This secondary analysis used public, de-identified datasets. No new human-participant recruitment, intervention or identifiable-data collection was undertaken.

6.8 Consent to publish

All authors have reviewed and approved the final version of this manuscript and consent to its publication.

6.3 Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

6.9 Acknowledgments

The authors would like to acknowledge their respective institutions and colleagues for their support and encouragement during the preparation of this manuscript. The authors also acknowledge the scientific literature and publicly available databases that contributed to the development of this work.

6.1 Data availability

All analysed expression data are publicly available through GEO under GSE42568 and GSE7904. The Python workflow, cleaned expression matrices, sample metadata, differential-expression outputs, model benchmark and score coefficients are available in the accompanying files.

All data supporting the findings of this study are available on paper.

References

1. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 99(10):6562–6566. https://doi.org/10.1073/pnas.102102699
2. Anstee JE, Feehan KT, Opzoomer JW, Dean I, Muller HP, Bahri M, Cheung TS, Liakath-Ali K, Liu Z, Choy D, Caron J, Sosnowska D, Beatson R, Muliaditan T, An Z, Gillett CE, Lan G, Zou X, Watt FM, Arnold JN (2023) LYVE-1 + macrophages form a collaborative CCR5-dependent perivascular niche that influences chemotherapy responses in murine breast cancer. Dev Cell 58(17):1548–1561e10. https://doi.org/10.1016/j.devcel.2023.06.006
3. Aran D, Hu Z, Butte AJ (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 18(1):220. https://doi.org/10.1186/s13059-017-1349-1
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE,

Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Gene Ontology Consortium Nat Genet 25(1):25–29. https://doi.org/10.1038/75556

5. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M, Choi K, Fromme RM, Dao P, McKenney PT, Wasti RC, Kadaveru K, Mazutis L, Rudensky AY (2018) & Pe'er, D. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell, 174(5), 1293–1308.e36. https://doi.org/10.1016/j.cell.2018.05.060
6. Bartlett J, Canney P, Campbell A, Cameron D, Donovan J, Dunn J, Earl H, Francis A, Hall P, Harmer V, Higgins H, Hillier L, Hulme C, Hughes-Davies L, Makris A, Morgan A, McCabe C, Pinder S, Poole C, Stein R (2013) Selecting breast cancer patients for chemotherapy: the opening of the UK OPTIMA trial. Clinical oncology (Royal College of Radiologists. (Great Britain)) 25(2):109–116. https://doi.org/10.1016/j.clon.2012.10.005
7. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, de Reyniès A (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol 17(1):218. https://doi.org/10.1186/s13059-016-1070-5
8. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc Ser B Methodological 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
9. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A, Bruneval P, Fridman WH, Becker C, Pagès F, Speicher MR, Trajanoski Z, Galon J (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity 39(4):782–795. https://doi.org/10.1016/j.immuni.2013.10.003
10. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) Bioinf (Oxford England) 19(2):185–193. https://doi.org/10.1093/bioinformatics/19.2.185. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias
11. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
12. Bueno-de-Mesquita JM, van Harten WH, Retel VP, van 't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, Wu TS, Atsma D, Rutgers EJ, Brink G, Floore AN, Glas AM, Roumen RM, Bellot FE, van Krimpen C, Linn SC (2007) Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). Lancet Oncol 8(12):1079–1087. https://doi.org/10.1016/S1470-2045(07)70346-7
13. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SAW, Savage MI, Osborne CK, Hilsenbeck SG, Chang JC, Mills GB, Lau CC, Brown PH (2015) Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin cancer research: official J Am Association Cancer Res 21(7):1688–1698. https://doi.org/10.1158/1078-0432.CCR-14-0432
14. Cai J, Chen T, Jiang Z, Yan J, Ye Z, Ruan Y, Tao L, Shen Z, Liang X, Wang Y, Xu J, Cai X (2023) Bulk and single-cell transcriptome profiling reveal extracellular matrix mechanical regulation of lipid

metabolism reprograming through YAP/TEAD4/ACADL axis in hepatocellular carcinoma. Int J Biol Sci 19(7):2114–2131. https://doi.org/10.7150/ijbs.82177

15. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61–70. https://doi.org/10.1038/nature11412
16. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, Pierga JY, Brain E, Causeret S, DeLorenzi M, Glas AM, Golfinopoulos V, Goulioti T, Knox S, Matos E, Meulemans B, Neijenhuis PA, Nitz U, Passalacqua R (2016) … MINDACT Investigators 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. The New England journal of medicine, 375(8), 717 – 29. https://doi.org/10.1056/NEJMoa1602253
17. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, Gallagher WM, Hennessy BT, Moriarty M, Crown J, Kennedy S, Clynes M (2013) Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis 34(10):2300–2308. https://doi.org/10.1093/carcin/bgt208
18. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ (Clinical research ed. 350:g7594. https://doi.org/10.1136/bmj.g7594
19. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
20. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group METABRIC, Langerød A, Green A, Aparicio S (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486(7403) 346 – 52. https://doi.org/10.1038/nature10983
21. Davies C, Pan H, Godwin J, Gray R, Arriagada R, Raina V, Abraham M, Medeiros Alencar VH, Badran A, Bonfill X, Bradbury J, Clarke M, Collins R, Davis SR, Delmestri A, Forbes JF, Haddad P, Hou MF, Inbar M … Adjuvant Tamoxifen: Longer Against Shorter (ATLAS) Collaborative Group (2013). Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. Lancet (London England), 381(9869), 805 – 16. https://doi.org/10.1016/S0140-6736(12)61963-1
22. Ding L, Niu C, Zheng Y, Xiong Z, Liu Y, Lin J, Sun H, Huang K, Yang W, Li X, Ye Q (2011) FHL1 interacts with oestrogen receptors and regulates breast cancer cell growth. J Cell Mol Med 15(1):72–85. https://doi.org/10.1111/j.1582-4934.2009.00938.x
23. Drukker CA, Bueno-de-Mesquita JM, Retèl VP, van Harten WH, van Tinteren H, Wesseling J, Roumen RMH, Knauer M, van 't Veer LJ, Sonke GS, Rutgers EJT, van de Vijver MJ, Linn SC (2013) A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. Int J Cancer 133(4) 929 – 36. https://doi.org/10.1002/ijc.28082
24. Du L, Gao Y (2021) PGM5-AS1 impairs miR-587-mediated GDF10 inhibition and abrogates progression of prostate cancer. J translational Med 19(1):12. https://doi.org/10.1186/s12967-020-02572-w

25. Duan X, Wang L, Wang Z, Wei W, Wang M, Ding D (2022) LncRNA PGM5-AS1 Inhibits the Progression of Bladder Cancer by Regulating miR-587/SLIT3 Axis. Crit Rev Eukaryot Gene Expr 32(8):9–22. https://doi.org/10.1615/CritRevEukaryotGeneExpr.2022042376
26. Duffy MJ, Sturgeon CM, Sölétormos G, Barak V, Molina R, Hayes DF, Diamandis EP, Bossuyt PMM (2015) Validation of new cancer biomarkers: a position statement from the European group on tumor markers. Clin Chem 61(6):809–820. https://doi.org/10.1373/clinchem.2015.239863
27. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Davies C, Godwin J, Gray R, Clarke M, Cutter D, Darby S, McGale P, Pan HC, Taylor C, Wang YC, Dowsett M, Ingle J, Peto R (2011) Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet (London England) 378(9793):771–784. https://doi.org/10.1016/S0140-6736(11)60993-8
28. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Peto R, Davies C, Godwin J, Gray R, Pan HC, Clarke M, Cutter D, Darby S, McGale P, Taylor C, Wang YC, Bergh J, Di Leo A, Albain K, Swain S, Piccart M, Pritchard K (2012) Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. Lancet (London England) 379(9814) 432 – 44. https://doi.org/10.1016/S0140-6736(11)61625-5
29. Edgar R, Domrachev M, Lash AE (2002) Nucleic Acids Res 30(1):207–210. https://doi.org/10.1093/nar/30.1.207. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository
30. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman and Hall/CRC
31. Elfstrum AK, Rumahorbo AH, Reese LE, Nelson EV, McCluskey BM, Schwertfeger KL (2024) LYVE-1-expressing Macrophages Modulate the Hyaluronan-containing Extracellular Matrix in the Mammary Stroma and Contribute to Mammary Tumor Growth. Cancer Res Commun 4(5):1380–1397. https://doi.org/10.1158/2767-9764.CRC-24-0205
32. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
33. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. https://doi.org/10.1007/s10994-006-6226-1
34. Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, Shak S, Baehner FL, Ravdin PM, Davidson NE, Sledge GW, Perez EA, Shulman LN, Martino S, Sparano JA (2008) Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features. J Clin oncology: official J Am Soc Clin Oncol 26(25):4063–4071. https://doi.org/10.1200/JCO.2007.14.4501
35. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, Fitzgibbons PL, Francis G, Goldstein NS, Hayes M, Hicks DG, Lester S, Love R, Mangu PB, McShane L, Miller K, Osborne CK, Paik S, Perlmutter J, Wolff AC (2010) American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and

progesterone receptors in breast cancer. J Clin oncology: official J Am Soc Clin Oncol 28(16):2784–2795. https://doi.org/10.1200/JCO.2009.25.6529

36. Han Y, Wang X, Xu M, Teng Z, Qin R, Tan G, Li P, Sun P, Liu H, Chen L, Jia B (2023) Aspartoacylase promotes the process of tumour development and is associated with immune infiltrates in gastric cancer. BMC Cancer 23(1):604. https://doi.org/10.1186/s12885-023-11088-7
37. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5) 646 – 74. https://doi.org/10.1016/j.cell.2011.02.013
38. Hanahan D (2022) Hallmarks of Cancer: New Dimensions. Cancer Discov 12(1):31–46. https://doi.org/10.1158/2159-8290.CD-21-1059
39. Hand DJ, Yu K (2001) Idiot's Bayes-not so stupid after all? Int Stat Rev 69(3):385–398. https://doi.org/10.1111/j.1751-5823.2001.tb00465.x
40. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic ROC curve. Radiology 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747
41. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, Ruddy K, Tsang J, Cardoso F (2019) Breast cancer. Nat reviews Disease primers 5(1):66. https://doi.org/10.1038/s41572-019-0111-2
42. Hayes D, Padnos SB (2011) Predictive and prognostic markers in cancer. Clin Adv Hematol oncology: H&O 9(2):130–132
43. Hayes DF (2015) Biomarker validation and testing. Mol Oncol 9(5):960–966. https://doi.org/10.1016/j.molonc.2014.10.004
44. Ho SY, Phua K, Wong L, Goh B, W. W (2020) Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns, vol 1. New York, p 100129. 8https://doi.org/10.1016/j.patter.2020.100129
45. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford England) 4(2):249–264. https://doi.org/10.1093/biostatistics/4.2.249
46. Jayabal P, Ma X, Akram M, Bow H, Mandal P, Mansano N, Chen Y, Weintraub ST, Houghton PJ, Wray S, Shiio Y (2025) IGSF10 is a RET antagonist regulating Ewing sarcoma growth and GnRH neuron migration. Cell Rep 44(8):116025. https://doi.org/10.1016/j.celrep.2025.116025
47. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford England) 8(1):118–127. https://doi.org/10.1093/biostatistics/kxj037
48. Jolliffe IT, Cadima J (2016) Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
49. Kalinsky K, Barlow WE, Gralow JR, Meric-Bernstam F, Albain KS, Hayes DF, Lin NU, Perez EA, Goldstein LJ, Chia SKL, Dhesy-Thind S, Rastogi P, Alba E, Delaloge S, Martin M, Kelly CM, Ruiz-

Borrego M, Gil-Gil M, Arce-Salinas CH, Hortobagyi GN (2021) 21-Gene Assay to Inform Chemotherapy Benefit in Node-Positive Breast Cancer. N Engl J Med 385(25):2336–2347. https://doi.org/10.1056/NEJMoa2108873

50. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30. https://doi.org/10.1093/nar/28.1.27
51. Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, Yang SR, Kurian A, Van Valen D, West R, Bendall SC, Angelo M (2018) A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. Cell 174(6):1373–1387e19. https://doi.org/10.1016/j.cell.2018.08.039
52. Krupenko SA, Krupenko NI (2018) ALDH1L1 and ALDH1L2 Folate Regulatory Enzymes in Cancer. Adv Exp Med Biol 1032:127–143. https://doi.org/10.1007/978-3-319-98788-0_10
53. Krupenko SA, Horita DA (2019) The Role of Single-Nucleotide Polymorphisms in the Function of Candidate Tumor Suppressor ALDH1L1. Front Genet 10:1013. https://doi.org/10.3389/fgene.2019.01013
54. Krupenko SA, Krupenko NI (2019) Loss of ALDH1L1 folate enzyme confers a selective metabolic advantage for tumor progression. Chemico-Biol Interact 302:149–155. https://doi.org/10.1016/j.cbi.2019.02.013
55. Krupenko NI, Sharma J, Fogle HM, Pediaditakis P, Strickland KC, Du X, Helke KL, Sumner S, Krupenko SA (2021) Knockout of Putative Tumor Suppressor Aldh1l1 in Mice Reprograms Metabolism to Accelerate Growth of Tumors in a Diethylnitrosamine (DEN) Model of Liver Carcinogenesis. Cancers 13(13):3219. https://doi.org/10.3390/cancers13133219
56. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44(W1) W90-7. https://doi.org/10.1093/nar/gkw377. 'ayan
57. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3(9):1724–1735. https://doi.org/10.1371/journal.pgen.0030161
58. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Investig 121(7):2750–2767. https://doi.org/10.1172/JCI45014
59. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics (Oxford, England), 27(12), 1739-40. https://doi.org/10.1093/bioinformatics/btr260
60. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1(6):417–425. https://doi.org/10.1016/j.cels.2015.12.004
61. Liu R, Zou Z, Zhang Z, He H, Xi M, Liang Y, Ye J, Dai Q, Wu Y, Tan H, Zhong W, Wang Z, Liang Y (2024) Evaluation of glucocorticoid-related genes reveals GPD1 as a therapeutic target and regulator of

sphingosine 1-phosphate metabolism in CRPC. Cancer Lett 605:217286. https://doi.org/10.1016/j.canlet.2024.217286

62. Munoz D, Near AM, van Ravesteyn NT, Lee SJ, Schechter CB, Alagoz O, Berry DA, Burnside ES, Chang Y, Chisholm G, de Koning HJ, Ergun A, Heijnsdijk M, Huang EAM, Stout H, Sprague NK, Trentham-Dietz BL, Mandelblatt A, J. S., Plevritis SK (2014) Effects of screening and systemic adjuvant therapy on ER-specific US breast cancer mortality. J Natl Cancer Inst 106(11):dju289. https://doi.org/10.1093/jnci/dju289
63. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA (2015) Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12(5):453–457. https://doi.org/10.1038/nmeth.3337
64. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M, Alizadeh AA (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 37(7):773–782. https://doi.org/10.1038/s41587-019-0114-2
65. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817–2826. https://doi.org/10.1056/NEJMoa041588
66. Pavlova NN, Thompson CB (2016) The emerging hallmarks of cancer metabolism. Cell Metabol 23(1):27–47. https://doi.org/10.1016/j.cmet.2015.12.006
67. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. Nature 406(6797) 747 – 52. https://doi.org/10.1038/35021093
68. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S (2006) X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 9(2):121–132. https://doi.org/10.1016/j.ccr.2006.01.013
69. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47. https://doi.org/10.1093/nar/gkv007
70. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432
71. Savas P, Virassamy B, Ye C, Salim A, Mintoff CP, Caramia F, Salgado R, Byrne DJ, Teo ZL, Dushyanthen S, Byrne A, Wein L, Luen SJ, Poliness C, Nightingale SS, Skandarajah AS, Gyorki DE, Thornton CM, Beavis PA, Loi S (2018) Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat Med 24(7):986–993. https://doi.org/10.1038/s41591-018-0078-7

72. Sestak I, Cuzick J (2015) Markers for the identification of late breast cancer recurrence. Breast cancer research: BCR 17(1):10. https://doi.org/10.1186/s13058-015-0516-0
73. Sgroi DC, Sestak I, Cuzick J, Zhang Y, Schnabel CA, Schroeder B, Erlander MG, Dunbier A, Sidhu K, Lopez-Knowles E, Goss PE, Dowsett M (2013) Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population. Lancet Oncol 14(11):1067–1076. https://doi.org/10.1016/S1470-2045(13)70387-5
74. Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics, 2014. Cancer J Clin 64(1):9–29. https://doi.org/10.3322/caac.21208
75. Simon RM, Paik S, Hayes DF (2009) Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst 101(21):1446–1452. https://doi.org/10.1093/jnci/djp335
76. Sparano JA, Paik S (2008) Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin oncology: official J Am Soc Clin Oncol 26(5):721–728. https://doi.org/10.1200/JCO.2007.15.1068
77. Sparano JA, Zhao F, Martino S, Ligibel JA, Perez EA, Saphner T, Wolff AC, Sledge GW, Wood WC, Davidson NE (2015a) Long-Term Follow-Up of the E1199 Phase III Trial Evaluating the Role of Taxane and Schedule in Operable Breast Cancer. J Clin oncology: official J Am Soc Clin Oncol 33(21):2353–2360. https://doi.org/10.1200/JCO.2015.60.9271
78. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, Geyer CE, Dees EC, Perez EA, Olson JA, Zujewski J, Lively T, Badve SS, Saphner TJ, Wagner LI, Whelan TJ, Ellis MJ, Paik S, Wood WC, Sledge GW (2015b) Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N Engl J Med 373(21):2005–2014. https://doi.org/10.1056/NEJMoa1510764
79. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, Geyer CE, Dees EC, Goetz MP, Olson JA, Lively T, Badve SS, Saphner TJ, Wagner LI, Whelan TJ, Ellis MJ, Paik S, Wood WC, Ravdin PM, Sledge GW (2018) Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N Engl J Med 379(2):111–121. https://doi.org/10.1056/NEJMoa1804710
80. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Frisén J (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics, vol 353. Science, New York, N.Y.), pp 78–82. 6294https://doi.org/10.1126/science.aaf2403
81. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102
82. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. Cancer J Clin 71(3):209–249. https://doi.org/10.3322/caac.21660

83. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98(19):10869–10874. https://doi.org/10.1073/pnas.191367098
84. Tang G, Shak S, Paik S, Anderson SJ, Costantino JP, Geyer CE, Mamounas EP, Wickerham DL, Wolmark N (2011) Comparison of the prognostic and predictive utilities of the 21-gene Recurrence Score assay and Adjuvant! for women with node-negative, ER-positive breast cancer: results from NSABP B-14 and NSABP B-20. Breast Cancer Res Treat 127(1):133–142. https://doi.org/10.1007/s10549-010-1331-z
85. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang O, Porta-Pardo TH, Gao E, Plaisier GF, Eddy CL, Ziv JA, Culhane E, Paull AC, Sivakumar EO, Gentles IKA, Malhotra AJ, Farshidfar R, Colaprico F, Parker A, Shmulevich JS, I (2018) The Immune Landscape of Cancer. Immunity 48(4):812–830e14. https://doi.org/10.1016/j.immuni.2018.03.023
86. Tsybovsky Y, Sereda V, Golczak M, Krupenko NI, Krupenko SA (2022) Structure of putative tumor suppressor ALDH1L1. Commun biology 5(1):3. https://doi.org/10.1038/s42003-021-02963-9
87. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536. https://doi.org/10.1038/415530a
88. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999–2009. https://doi.org/10.1056/NEJMoa021967
89. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91. https://doi.org/10.1186/1471-2105-7-91
90. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis making: Int J Soc Med Decis Mak 26(6):565–574. https://doi.org/10.1177/0272989X06295361
91. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, Rees M, Ramaswamy A, Muenst S, Soysal SD, Jacobs A, Windhager J, Silina K, van den Broek M, Dedes KJ, Rodríguez Martínez M, Weber WP, Bodenmiller B (2019) A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. Cell 177(5):1330–1345e18. https://doi.org/10.1016/j.cell.2019.03.005
92. Waks AG, Winer EP (2019) Breast Cancer Treatment: A Review. JAMA 321(3):288–300. https://doi.org/10.1001/jama.2018.19323

93. Wang L, Zhang Y, Ding D, He X, Zhu Z (2012) Lack of association of ADH1C genotype with breast cancer susceptibility in Caucasian population: a pooled analysis of case-control studies. Breast (Edinburgh Scotland) 21(4):435–439. https://doi.org/10.1016/j.breast.2012.01.007
94. Welch BL (1947) The generalization of Student's problem when several different population variances are involved. Biometrika 34(1–2):28–35. https://doi.org/10.1093/biomet/34.1-2.28
95. Wolff AC, Hammond MEH, Hicks DG, Dowsett M, McShane LM, Allison KH, Allred DC, Bartlett JMS, Bilous M, Fitzgibbons P, Hanna W, Jenkins RB, Mangu PB, Paik S, Perez EA, Press MF, Spears PA, Vance GH, Viale G (2013) … College of American Pathologists Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 31(31), 3997–4013. https://doi.org/10.1200/JCO.2013.50.9984
96. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, PROBAST Group† (2019) PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 170(1):51–58. https://doi.org/10.7326/M18-1376
97. Wong WB, Ramsey SD, Barlow WE, Garrison LP, Veenstra DL (2012) The value of comparative effectiveness research: projected return on investment of the RxPONDER trial (SWOG S1007). Contemp Clin Trials 33(6):1117–1123. https://doi.org/10.1016/j.cct.2012.08.006
98. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, Wang T, Larsson L, Kaczorowski D, Weisenfeld NI, Uytingco CR, Chew JG, Bent ZW, Chan CL, Gnanasambandapillai V, Swarbrick A (2021) A single-cell and spatially resolved atlas of human breast cancers. Nat Genet 53(9):1334–1347. https://doi.org/10.1038/s41588-021-00911-1
99. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RGW (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 4:2612. https://doi.org/10.1038/ncomms3612
100. Yu G, Wang LG, Han Y, He QY (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16(5):284–287. https://doi.org/10.1089/omi.2011.0118
101. Zhou B, Yi F, Chen Y, Li CH, Cheng YS, Yang K (2020) Reduced long noncoding RNA PGM5-AS1 facilitated proliferation and invasion of colorectal cancer through sponging miR-100-5p. Eur Rev Med Pharmacol Sci 24(15):7972–7981. https://doi.org/10.26355/eurrev_202008_22480
102. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B Stat Methodol 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Figures

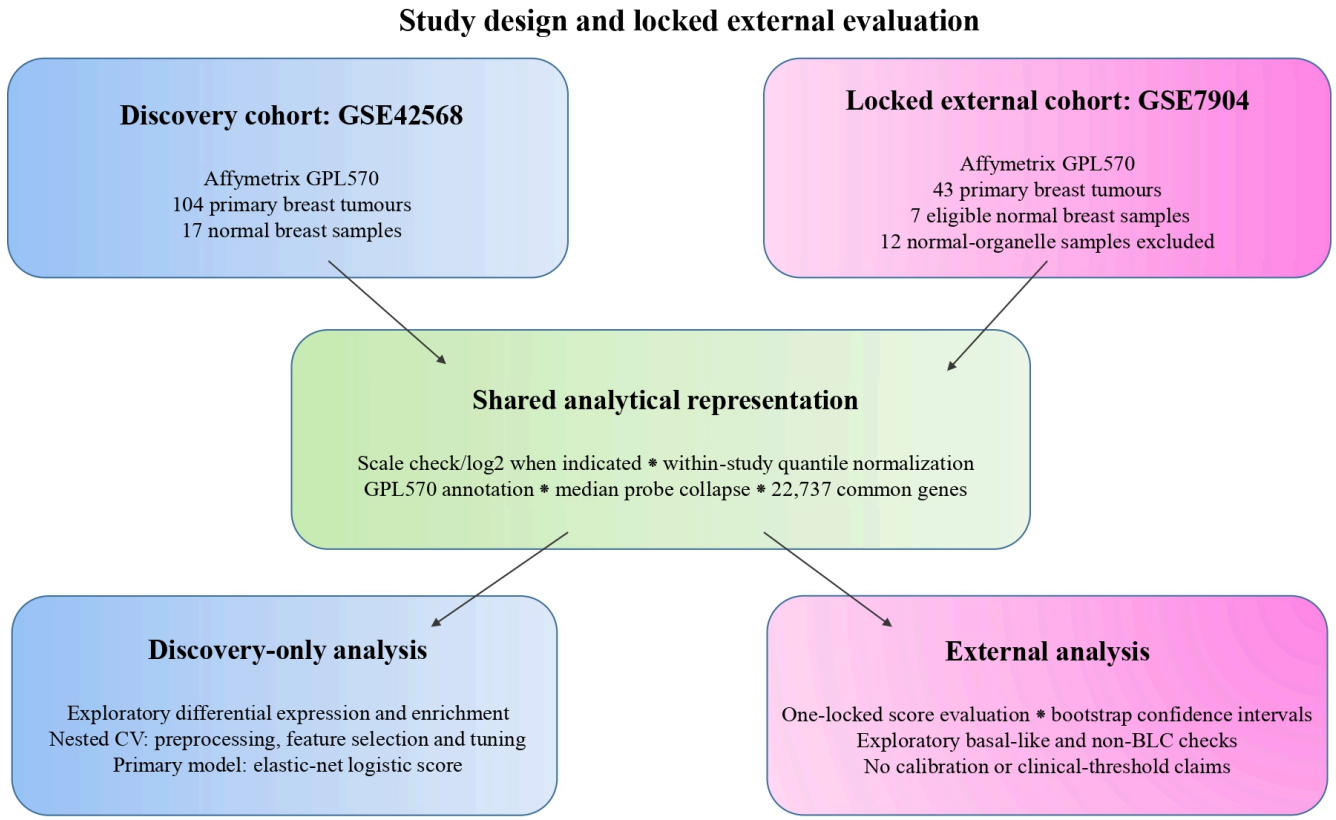


Figure 1

The complete study design. GSE42568 was used for the discovery dataset and for nested model development within the training framework, whereas GSE7904 was maintained independently as a locked external cohort.

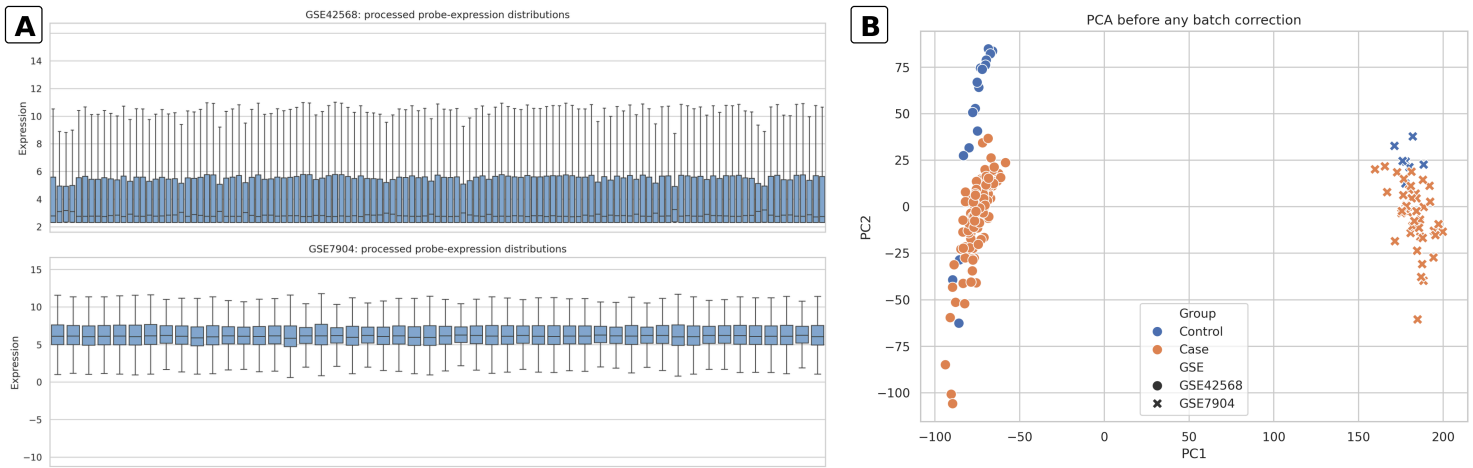


Figure 2

Quality-control and domain-shift diagnostics. (A) Processed probe-expression distributions in GSE42568 and GSE7904. (B) Principal component analysis before any cross-study batch correction, coloured by case/control class and marked by cohort.

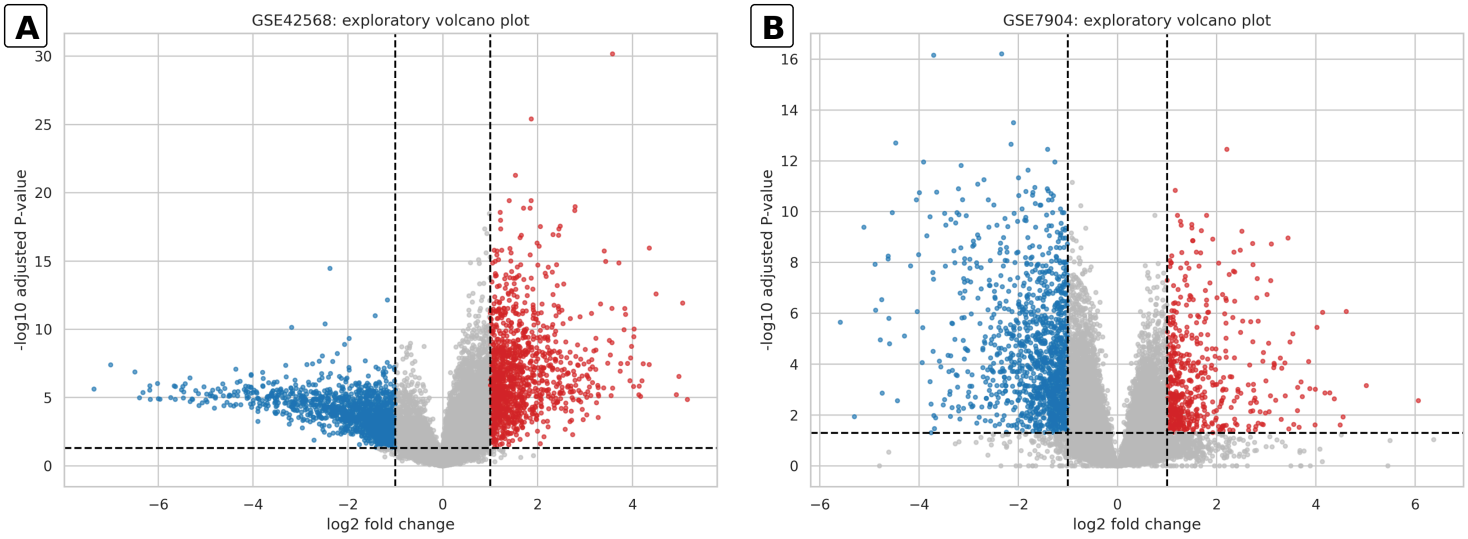


Figure 3

Differential-expression volcano plots. (A) GSE42568 discovery and (B) GSE7904 locked external cohort. The dashed lines represent the exploratory thresholds (adjusted P <0.05 and log2 fold change ≥1). The plots demonstrate the distribution of differential cohort-specific effect.

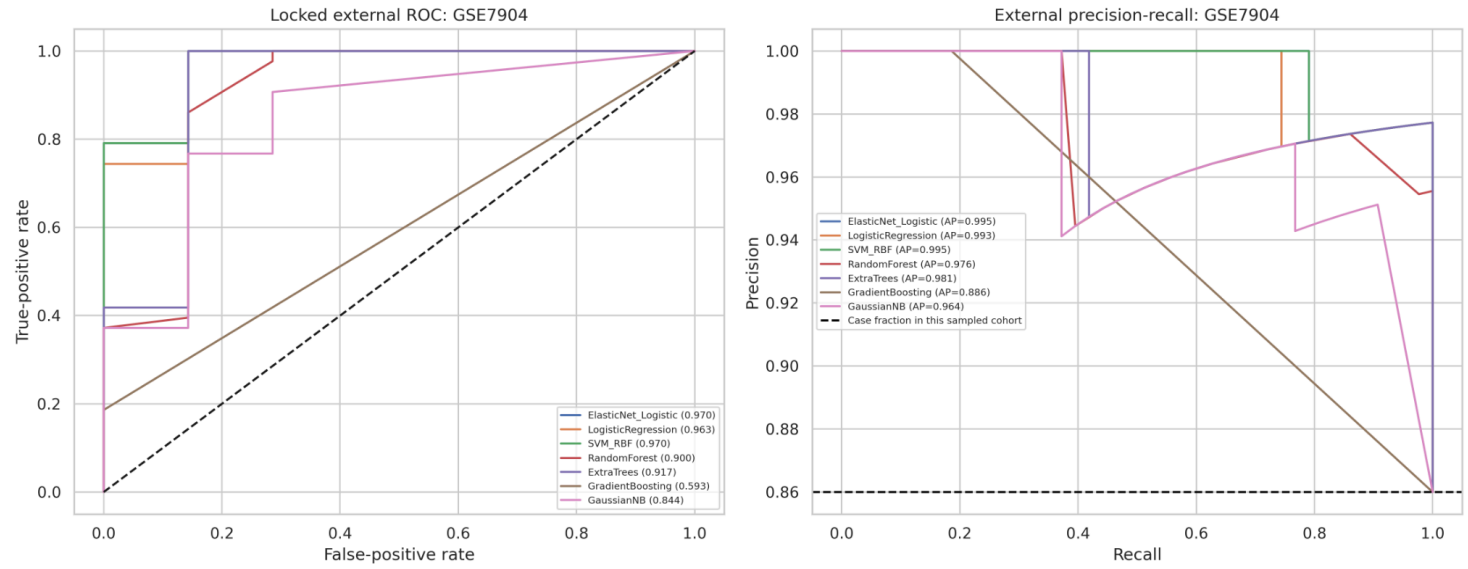


Figure 4

Locked external ROC and precision-recall curves in GSE7904. Curves are shown for the primary elastic-net logistic model and sensitivity classifiers.

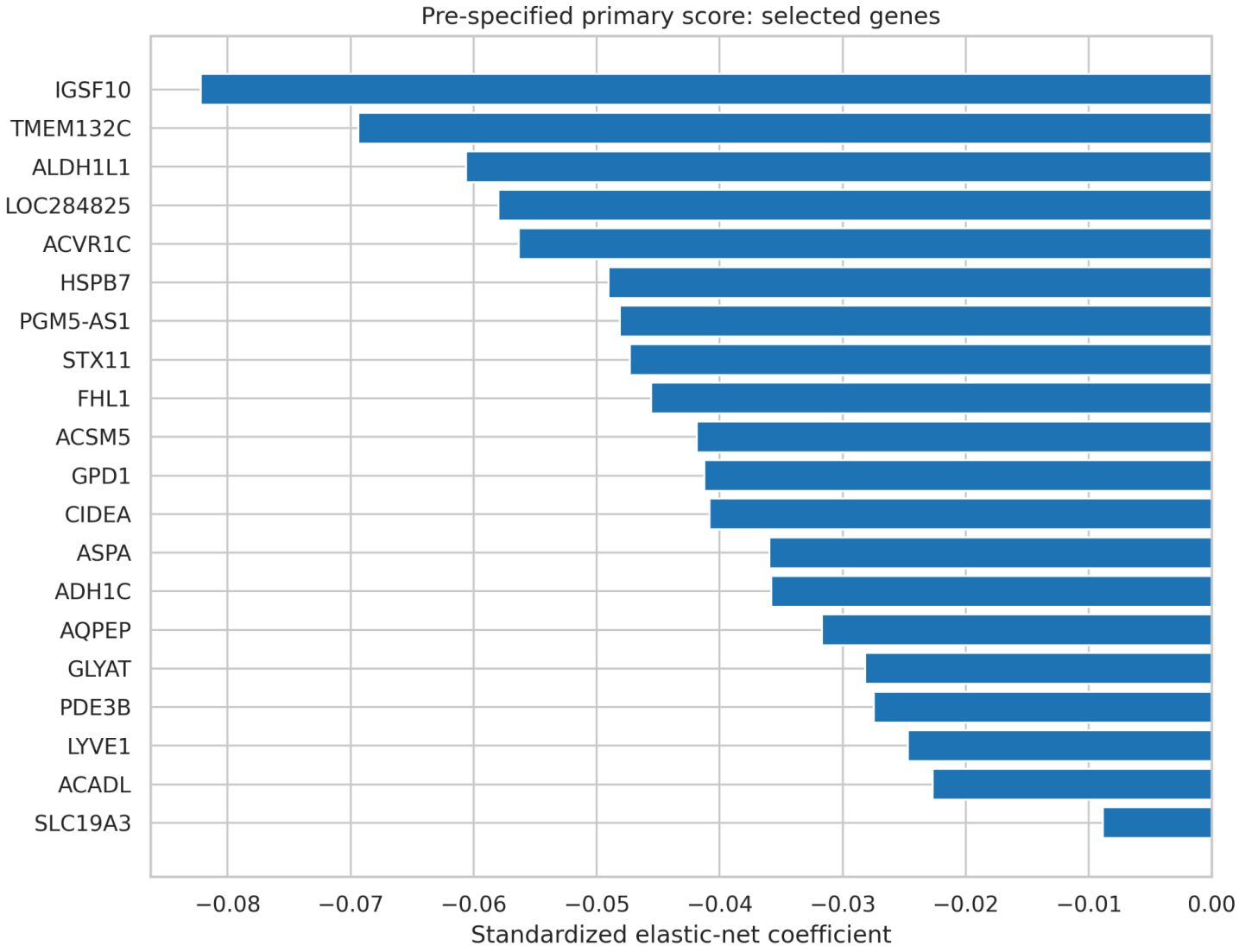


Figure 5

Coefficients in the pre-specified primary elastic-net score. All 20 retained features have negative coefficients for the tumour class.

Supplementary Files

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

- S1.xls
- S2.xls
- S3.xls
- S4.xls
- S5.xls

- S6.xls
- SF1JPED.jpg
- SF2JPEG.jpg