

Decoding the GPRC5A Paradox in Pancreatic Ductal Adenocarcinoma: A Subtype-Stratified, Treatment-Deconfounded, Multi-Omic Investigation

Mark Barsoum Markarian

mkm25@mail.aub.edu

American University of Beirut <https://orcid.org/0009-0006-1240-9534>

Research Article

Keywords: GPRC5A, pancreatic ductal adenocarcinoma, molecular subtypes, gemcitabine, CPTAC, AlphaFold, machine learning, biomarker, GPCR

Posted Date: March 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9237732/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: Pancreatic ductal adenocarcinoma (PDAC) carries a five-year survival rate below 10%, underscoring the urgent need for mechanistically grounded prognostic biomarkers. A prior batch-harmonized machine learning framework identified GPRC5A as prognostically relevant in PDAC but found reduced expression in deceased patients, the opposite of its established oncogenic role, generating an unexplained paradox.

Objectives: To resolve the GPRC5A paradox through five interrelated analyses: molecular subtype stratification, gemcitabine treatment deconfounding, RNA–protein concordance assessment, somatic mutation mapping, and machine learning role-state classification.

Methods: TCGA-PAAD (n=177) provided RNA-seq, clinical, and somatic mutation data; CPTAC-PAAD (n=140) provided matched proteomics. Molecular subtypes were assigned using the Moffitt 2015 single-sample classifier. Cox proportional hazards models and multivariable adjustment were used for survival analyses. GPRC5A RNA–protein concordance was quantified by Spearman correlation and benchmarked against 4,491 genome-wide gene pairs. Somatic mutations were mapped onto an AlphaFold2-predicted GPRC5A structure. A leakage-free Random Forest, XGBoost, and logistic regression pipeline was trained to predict GPRC5A functional role state (oncogenic vs. suppressive) from subtype and co-expression features.

Results: In the classical subtype (n=100), high GPRC5A expression associated with significantly worse survival (log-rank $p=0.00024$; HR=1.53, 95% CI 1.17-2.00). In the basal-like subtype (n=77), high expression paradoxically associated with modestly better survival (log-rank $p=0.022$; HR=1.26, 95% CI 1.06-1.50 by continuous Cox model; see Discussion for reconciliation of KM and Cox directionality). GPRC5A remained a significant independent predictor across all multivariable models (fully adjusted HR=1.44, 95% CI 1.23-1.68, $p=3.89\times 10^{-6}$). RNA-protein correlation was moderate (Spearman $r=0.571$, 84.6th genome-wide percentile), arguing against post-transcriptional repression. No somatic mutations were detected in GPRC5A. The Random Forest role-state classifier achieved a held-out test AUC of 0.833 (LOOCV AUC=0.758), with classical co-expression features dominating over GPRC5A expression itself.

Conclusions: The GPRC5A paradox is primarily explained by molecular subtype mixing, with gemcitabine-induced transcriptional confounding as a secondary contributor. Post-transcriptional regulation and somatic mutation are not major drivers. GPRC5A should be evaluated within, not across, molecular subtypes, and its absence of somatic mutations directs mechanistic inquiry toward epigenomic regulation. A machine learning classifier assigns GPRC5A functional role state from transcriptomic context with reasonable accuracy, providing a proof-of-concept tool for subtype-aware prognostic stratification in PDAC.

INTRODUCTION

1.1 Pancreatic Ductal Adenocarcinoma: An Unrelenting Clinical Challenge

Pancreatic ductal adenocarcinoma (PDAC) is among the most lethal of all solid tumors, with a five-year overall survival rate below 10% that has improved only marginally over the past three decades [1]. The disease is characterized by late-stage diagnosis; more than 80% of patients present with locally advanced or metastatic disease, a dense desmoplastic tumor microenvironment that limits drug penetration, and a remarkable degree of molecular heterogeneity that confounds therapeutic targeting [2, 3].

Gemcitabine monotherapy has been the backbone of PDAC systemic treatment since its approval in 1997, and while combination regimens such as FOLFIRINOX and gemcitabine plus nab-paclitaxel have modestly extended median survival in eligible patients, responses remain short-lived and resistance develops rapidly [4, 5, 6]. The persistent therapeutic ceiling in PDAC reflects an incomplete understanding of the molecular drivers that distinguish aggressive from less aggressive disease, and underscores the urgent need for both refined prognostic biomarkers and mechanistically grounded drug targets.

1.2 Molecular Subtypes of PDAC: Classical and Basal-like

A major conceptual advance in PDAC biology has been the recognition that bulk transcriptomic analyses obscure clinically meaningful molecular heterogeneity. Moffitt and colleagues applied virtual microdissection to PDAC transcriptomic data and identified two tumor-intrinsic subtypes, classical and basal-like, with markedly distinct survival profiles [7]. The classical subtype, characterized by expression of genes associated with epithelial differentiation, is associated with improved overall survival. The basal-like subtype, sharing features with squamous and basal transcriptional programs, exhibits greater aggressiveness, resistance to standard chemotherapy, and substantially worse prognosis. A complementary four-subtype classification was subsequently proposed by Bailey and colleagues based on genomic and transcriptomic integration [8], further cementing the view that PDAC is not a single molecular entity.

Despite these advances, the prognostic impact of individual genes identified in bulk-cohort analyses has rarely been re-evaluated within the context of these established subtypes. This is a critical oversight: a gene whose expression correlates with poor outcome in the aggregate cohort may behave in entirely opposing directions across subtypes, generating paradoxical associations that are artifactual products of subtype mixing rather than true biological signals. This study directly addresses this analytical gap.

1.3 GPRC5A: An Orphan GPCR with Established Oncogenic Functions

GPRC5A (G Protein-Coupled Receptor Class C Group 5 Member A) is an orphan receptor of the retinoic acid-inducible class C GPCR family. Unlike most GPCRs, GPRC5A lacks a known endogenous ligand and

its downstream signaling cascades remain incompletely characterized. GPRC5A was initially identified as a tumor suppressor in lung tissue, where its loss cooperates with oncogenic KRAS to promote lung adenocarcinoma [9]. However, subsequent studies in gastrointestinal and pancreatic cancers demonstrated a diametrically opposite role: in PDAC, GPRC5A expression is elevated in tumor tissue relative to normal pancreatic parenchyma, promotes cell proliferation and invasiveness, and associates with poor clinical outcomes [10].

A particularly important observation from Zhou and colleagues is that gemcitabine treatment itself induces GPRC5A upregulation in PDAC cell lines through a mechanism involving the RNA-binding protein HuR, which stabilizes GPRC5A mRNA under conditions of chemotherapy stress [10]. This pharmacological induction of GPRC5A expression creates a specific confound in clinical data: patients who survived long enough to receive gemcitabine, or who responded to it, may carry systematically elevated GPRC5A levels as a direct consequence of treatment, independent of any intrinsic tumor biology. Disentangling this treatment-induced signal from a true prognostic association requires explicit stratification by chemotherapy receipt, an analysis rarely performed in retrospective biomarker studies.

1.4 The GPRC5A Paradox: An Unexplained Finding from Machine Learning Biomarker Discovery

In a preceding study, we developed a batch-harmonized machine learning framework for cross-cohort RNA biomarker discovery in PDAC [11]. Random Forest and XGBoost models were trained on TCGA-PAAD (n = 177) and validated on GSE71729 (n = 357), identifying five prognostic RNA signatures: LAMC2, DKK1, ITGB6, GPRC5A, and MAL2. Among these, GPRC5A presented a striking and unresolved contradiction: the model found reduced GPRC5A expression in deceased patients, the precise opposite of what its established oncogenic role would predict.

This discordance could not be attributed to technical artifacts, as the original analysis applied ComBat-seq batch harmonization and performed rigorous cross-cohort validation. Three mechanistic explanations were proposed but not tested: (i) PDAC molecular subtype mixing, whereby opposing subtype-specific expression-survival relationships produce a paradoxical aggregate signal; (ii) gemcitabine-induced transcriptional confounding, whereby treatment elevates GPRC5A in surviving patients and inflates survival-associated expression differences in the wrong direction; and (iii) post-transcriptional regulation, whereby RNA-level measurements fail to capture the biologically relevant protein species. The present study was designed to rigorously test each of these hypotheses using orthogonal data sources and analytical frameworks.

1.5 Post-Transcriptional Regulation and the Value of Proteomics Integration

RNA-seq remains the dominant modality for large-scale cancer biomarker discovery, yet the correlation between mRNA abundance and protein expression is imperfect and context-dependent, with genome-wide RNA-protein Spearman correlations typically ranging from 0.4 to 0.6 in tumor samples [12, 13]. For

receptors and signaling molecules, post-translational modifications such as phosphorylation further modulate functional activity independently of total protein abundance. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) has generated matched RNA-seq, proteomic, and phosphoproteomic profiles for PDAC samples, providing a unique resource to interrogate whether GPRC5A's RNA-level paradox persists at the protein level or is resolved by post-transcriptional mechanisms [14].

In parallel, the availability of high-confidence AlphaFold2-predicted protein structures [15] enables systematic mapping of somatic mutations onto receptor topology for the first time without requiring experimental structure determination. For an orphan GPCR such as GPRC5A, for which no crystal structure exists, this represents a significant methodological advance. If somatic mutations cluster in functionally distinct structural domains, they may provide a structural basis for context-dependent receptor behavior that transcriptomic analyses alone cannot reveal.

1.6 Machine Learning for Functional Role-State Prediction in Cancer Biology

The concept of a gene exhibiting context-dependent oncogenic versus tumor-suppressive behavior, sometimes called oncogenic switching, is increasingly recognized in cancer biology, particularly for signaling receptors whose activity is governed by cellular state rather than by expression level alone [30]. For GPRC5A specifically, no computational framework exists to assign a tumor's GPRC5A functional state from multi-omic inputs. Machine learning classifiers trained on subtype and co-expression features offer a principled approach to this problem, provided that label construction and feature selection are strictly insulated from the test set to prevent data leakage, a methodological failure that has affected a number of published cancer ML biomarker studies [29].

1.7 Study Objectives

The present study pursues five interrelated objectives. First, we re-examine GPRC5A's prognostic association within molecularly defined PDAC subtypes to determine whether subtype mixing explains the paradoxical bulk-cohort finding. Second, we stratify patients by gemcitabine treatment status and apply multivariable Cox regression to isolate treatment-induced confounding from intrinsic prognostic signal. Third, we leverage CPTAC-PAAD matched proteomics to quantify RNA–protein concordance and assess the contribution of post-transcriptional regulation. Fourth, we map TCGA-PAAD somatic mutations onto the AlphaFold2-predicted GPRC5A structure to identify structurally organized mutation patterns associated with survival outcomes. Fifth, we train and validate a machine learning classifier that predicts GPRC5A functional role state, oncogenic versus suppressive, from subtype and co-expression features, delivering a novel analytical tool for subtype-aware prognostic stratification in PDAC.

MATERIALS AND METHODS

2.1 Study Cohorts and Data Sources

The primary discovery cohort was TCGA-PAAD, comprising RNA-seq expression profiles, clinical annotations, and somatic mutation data for 177 pancreatic ductal adenocarcinoma samples with available survival information, accessed via the Genomic Data Commons (GDC) portal (<https://portal.gdc.cancer.gov>). RNA-seq counts were variance-stabilizing transformed (VST) using DESeq2 prior to all downstream analyses [17]. Clinical variables extracted included vital status, overall survival in months, documented chemotherapy regimen, age at diagnosis, and pathological stage.

For protein-level validation, matched RNA-seq, proteomic (LC-MS/MS), and phosphoproteomic data from the CPTAC Pancreatic Ductal Adenocarcinoma Discovery Study were accessed via the CPTAC Data Portal (<https://cptac-data-portal.georgetown.edu>). A total of 140 samples with matched RNA and protein measurements for GPRC5A were available for correlation analysis [14].

Table 1
Cohort summary.

Cohort	n	Classical	Basal-like	Deceased	Alive
TCGA-PAAD (primary)	177	100 (56.5%)	77 (43.5%)	93 (52.5%)	84 (47.5%)
CPTAC-PAAD (proteomics)	140 matched	-	-	-	-
ICGC PACA-AU / PACA-CA (validation)	Planned	-	-	-	-

2.2 Molecular Subtype Classification

PDAC molecular subtypes were assigned to all TCGA-PAAD samples using the Moffitt 2015 single-sample classifier, which derives classical and basal-like scores by computing the mean expression of curated tumor-intrinsic gene signatures after centering each sample [7]. A composite subtype score was computed as the difference between the classical and basal-like scores (positive values indicating classical identity; negative values indicating basal-like identity). Samples were assigned to classical (n = 100, 56.5%) or basal-like (n = 77, 43.5%) subtypes based on the sign of this composite score. No ambiguous or intermediate category was defined; all 177 samples with available RNA-seq data received a definitive subtype assignment.

Subtype score distributions were visualized against GPRC5A expression to confirm expected co-variation with the classical gene signature. The heatmap of GPRC5A and Moffitt signature genes ordered by subtype identity was generated using pheatmap in R, with samples ordered by subtype and then by survival status within each subtype.

2.3 Survival Analysis: Subtype-Stratified GPRC5A Prognostic Association

Overall survival was defined as time in months from the date of initial pathological diagnosis to the date of death (event) or last follow-up (censored). Kaplan-Meier survival curves were constructed for high-versus low-GPRC5A expression groups within each molecular subtype, with the expression threshold set at the median within the full cohort to ensure comparability across strata. Log-rank tests were used to assess the statistical significance of survival differences. All survival analyses were performed using the survival and survminer packages in R [19, 20].

Cox proportional hazards models were fitted to quantify the hazard ratio (HR) for GPRC5A expression (per unit increase in VST-normalized expression) on overall survival. Four Cox models were estimated: (i) an overall model on all 177 samples; (ii) a classical-subtype-restricted model ($n = 100$); (iii) a basal-like-subtype-restricted model ($n = 77$); and (iv) an interaction model including a GPRC5A $\times$ subtype interaction term to formally test for subtype-differential effect modification. Proportional hazards assumptions were verified using Schoenfeld residual tests. All HRs are reported with 95% confidence intervals and two-sided p-values.

2.4 Treatment Deconfounding Analysis

TCGA-PAAD clinical annotations were used to classify patients by documented chemotherapy receipt. Patients were designated as gemcitabine-treated ($n = 75$) if their treatment record indicated receipt of gemcitabine monotherapy or gemcitabine-containing combination regimens. Additional treatment groups were defined for fluorouracil/FOLFIRINOX ($n = 4$), radiation only or unspecified local therapy ($n = 92$), other chemotherapy ($n = 5$), and treatment-naïve or unknown ($n = 1$). GPRC5A VST expression was compared between treatment groups and between alive and deceased patients within each treatment stratum using Kruskal–Wallis tests with Wilcoxon pairwise post-hoc comparisons.

Treatment-stratified Cox proportional hazards models were estimated separately for all patients ($n = 177$) and for the gemcitabine-treated subgroup ($n = 75$) to assess whether the GPRC5A hazard ratio was attenuated within the treatment-exposed group. A series of multivariable Cox models were then fitted to the full cohort with progressive covariate adjustment: (i) univariate GPRC5A; (ii) GPRC5A + gemcitabine receipt; (iii) GPRC5A + molecular subtype; (iv) fully adjusted model including GPRC5A, gemcitabine receipt, subtype, age, and pathological stage; and (v) an interaction model including a GPRC5A $\times$ gemcitabine term. Gemcitabine receipt was treated as a binary covariate. Age was modeled as a continuous variable. Stage was binarized as Stage IV versus other (with unknown stage as a separate category). All models are reported in Table 3 with HRs, 95% CIs, and p-values.

2.5 Protein-Level Validation: CPTAC-PAAD RNA-Protein Correlation

GPRC5A RNA–protein concordance was quantified using Spearman rank correlation between VST-normalized RNA-seq expression and log₂-transformed LC-MS/MS protein abundance across 140 matched CPTAC-PAAD samples. Non-parametric correlation was used to minimize sensitivity to outliers and distributional assumptions. The resulting Spearman r for GPRC5A was contextualized against the

genome-wide distribution of RNA–protein correlations computed across all 4,491 genes with matched RNA and protein data in the CPTAC-PAAD dataset, expressed as a percentile rank. Scatter plots were stratified by vital status to assess whether RNA–protein dissociation differed between alive and deceased patients.

Phosphoproteomic data were interrogated for GPRC5A phosphorylation sites; however, no GPRC5A phosphopeptides with sufficient coverage for differential abundance analysis were detected in the CPTAC-PAAD dataset. Domain-level phosphosite annotation therefore could not be performed. Given the moderate RNA-protein correlation observed (Spearman $r = 0.571$, 84.6th percentile genome-wide), the working interpretation is that GPRC5A is reasonably well regulated at the protein level relative to its mRNA, with residual variance potentially attributable to post-translational modification rather than wholesale post-transcriptional repression.

2.6 AlphaFold2 Structural Modeling and Somatic Mutation Mapping

The predicted three-dimensional structure of human GPRC5A (UniProt accession Q8NFJ5, 350 amino acids) was retrieved from the AlphaFold Protein Structure Database [16]. Per-residue confidence scores (pLDDT) were extracted and used to assess structural reliability across the protein's functional domains. Transmembrane (TM) helices, extracellular loops (ECL), intracellular loops (ICL), N-terminal, and C-terminal regions were annotated based on UniProt topology predictions cross-referenced with GPRC5A literature [10].

Somatic mutations in GPRC5A were extracted from TCGA-PAAD via the GDC API using the TCGAbiolinks R package [18]. Mutations were filtered to include only those in tumor samples with available survival and subtype data. Each mutation was mapped to its amino acid position and annotated with the corresponding structural domain. Survival associations of domain-specific mutations were planned using log-rank tests comparing patients with versus without mutations in each domain; however, no somatic mutations in GPRC5A were detected across the 177 TCGA-PAAD samples analyzed, precluding domain-level survival analysis. This null result is reported as a finding.

2.7 Machine Learning Role-State Classifier

2.7.1 Label Construction

GPRC5A functional role-state labels (oncogenic vs. suppressive) were derived from the subtype-stratified and treatment-deconfounded survival analyses of Aims 1 and 2. A sample was labeled 'oncogenic' if it belonged to the classical subtype with high GPRC5A expression and deceased status, or if subtype-specific Cox modeling identified a positive GPRC5A-survival association in that subtype stratum. A sample was labeled 'suppressive' if the same logic yielded an inverse or protective association. Labels were constructed exclusively on the training partition prior to any classifier training to prevent label leakage.

2.7.2 Feature Engineering and Selection

Features comprised: (i) GPRC5A VST expression; (ii) classical and basal-like subtype scores derived from the Moffitt classifier; (iii) a composite subtype score (classical minus basal-like); and (iv) VST-normalized expression values for 25 co-expressed genes identified by co-expression analysis in the training set. All features were scaled to the [0,1] range using min-max normalization parameters estimated exclusively on the training data and applied without re-fitting to the test set. Feature importance was assessed post-hoc using mean decrease in impurity (Gini importance) from the Random Forest model.

2.7.3 Model Training and Validation

Three classifiers were trained: logistic regression (L2 penalty), Random Forest (500 trees), and XGBoost. The dataset was partitioned into a training set (80%) and a held-out test set (20%) using stratified random splitting to preserve class balance. All preprocessing steps, including feature scaling, subtype score computation, and co-expression feature selection, were applied within the training fold only and then applied to the test set without refitting, ensuring a fully leakage-free evaluation pipeline. Cross-validation on the training set used leave-one-out cross-validation (LOOCV) given the small effective sample size. Final model performance was evaluated on the held-out test set using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. Calibration was assessed using calibration plots of observed versus predicted oncogenic probability in the test set.

All machine learning analyses were performed in R using the caret and randomForest packages [21, 22]. XGBoost was implemented via the xgboost R package [23]. Survival analyses of predicted role-state groups used Kaplan-Meier estimation with log-rank testing on the held-out test set ($n = 12$), with the caveat that statistical power is limited at this sample size.

2.8 Statistical Considerations and Software

All statistical analyses were performed in R version 4.4.1. Continuous variables were compared between groups using the Wilcoxon rank-sum test (two-group comparisons) or Kruskal–Wallis test (multi-group comparisons) unless otherwise specified. All p-values are two-sided. A significance threshold of $\alpha = 0.05$ was applied throughout; no correction for multiple testing was applied across aims given the exploratory nature of this investigation, but corrections within individual multi-comparison tests (e.g., pairwise Wilcoxon post-hoc) used the Benjamini-Hochberg false discovery rate procedure. Figures were generated using ggplot2 and associated extension packages (ggsurvfit, survminer, pheatmap) [24].

This study is purely observational and retrospective, using publicly available de-identified datasets. No institutional review board approval was required. All data were accessed in accordance with the respective data use agreements for TCGA (open access tier) and CPTAC (open access tier).

RESULTS

3.1 Molecular Subtype Stratification Resolves the Directional Paradox

Application of the Moffitt 2015 single-sample classifier to TCGA-PAAD (n = 177) assigned 100 samples (56.5%) to the classical subtype and 77 (43.5%) to the basal-like subtype. The composite subtype score was strongly correlated with GPRC5A expression (Fig. 3), consistent with GPRC5A's co-variation with the classical transcriptional program. The expression heatmap confirmed that GPRC5A expression tracks with the classical gene signature and that deceased patients tend to cluster toward higher GPRC5A levels within each subtype (Fig. 1).

Within the classical subtype, GPRC5A expression was significantly higher in deceased patients (median 14.98 VST) compared to alive patients (median 14.11 VST; Wilcoxon $p = 5.9 \times 10^{-5}$). Within the basal-like subtype, deceased patients also showed higher median GPRC5A expression (14.99 VST) compared to alive patients (13.83 VST; Wilcoxon $p = 0.0099$). These within-subtype differences were consistent in direction, higher expression associates with worse survival in both subtypes, indicating that the paradox was not caused by opposing directionality across subtypes per se, but rather by differential magnitude and subtype composition in the bulk cohort (Fig. 2).

Kaplan–Meier analysis by GPRC5A expression level (high vs. low, median threshold) revealed strikingly different survival patterns across subtypes. In the classical subtype (n = 100), high GPRC5A expression was associated with significantly worse overall survival (log-rank $p = 0.00024$; Fig. 4, upper panel), consistent with the established oncogenic role. In the basal-like subtype (n = 77), high GPRC5A expression was paradoxically associated with modestly better survival (log-rank $p = 0.022$; Fig. 4, lower panel). This opposing Kaplan-Meier directionality within the basal-like subtype, where high expression confers a relative survival advantage, provides the first subtype-stratified evidence that GPRC5A's prognostic behavior is context-dependent. Notably, despite this opposing KM directionality, continuous Cox HRs were > 1 in both subtypes (see Table 2 and § 4.2 for discussion of this apparent inconsistency).

Cox proportional hazards modeling quantified the subtype-specific hazard ratios for GPRC5A expression per unit increase in VST-normalized expression. In the overall cohort, the hazard ratio was 1.36 (95% CI 1.17–1.58, $p = 5.26 \times 10^{-5}$). Within the classical subtype, the HR was higher at 1.53 (95% CI 1.17–2.00, $p = 0.00165$), while in the basal-like subtype the HR was attenuated at 1.26 (95% CI 1.06–1.50, $p = 0.00772$). Notably, the formal GPRC5A×subtype interaction term was non-significant (HR = 1.19, 95% CI 0.87–1.63, $p = 0.272$), indicating that the statistical evidence for differential effect modification by subtype does not reach significance at current sample sizes, despite the divergent Kaplan–Meier curves. All Cox results are presented in Table 2 and visualized in Fig. 5.

Table 2
Cox proportional hazards results for GPRC5A expression by molecular subtype; TCGA-PAAD.

Model	HR	95% CI	p-value	Significant
Overall (n = 177)	1.36	1.17–1.58	5.26×10^{-5}	*
Classical (n = 100)	1.53	1.17–2.00	0.00165	*
Basal-like (n = 77)	1.26	1.06–1.50	0.00772	*
Interaction term: GPRC5A × Subtype	1.19	0.87–1.63	0.272	-

HR = hazard ratio per unit increase in VST-normalized GPRC5A expression. CI = 95% confidence interval. Interaction term tests GPRC5A × Classical subtype effect modification. * $p < 0.05$.

3.2 Gemcitabine Treatment Attenuates but Does Not Reverse the GPRC5A Prognostic Signal

TCGA-PAAD patients were stratified by documented chemotherapy receipt into five treatment groups: gemcitabine (n = 75), radiation only or unspecified local therapy (n = 92), fluorouracil/FOLFIRINOX (n = 4), other chemotherapy (n = 5), and treatment-naïve or unknown (n = 1). Median GPRC5A expression was consistently higher in deceased versus alive patients across all treatment strata with sufficient sample size (Table 3). Within the gemcitabine-treated group specifically, deceased patients had a higher median GPRC5A expression (15.27 VST) than alive patients (14.11 VST), mirroring the overall cohort pattern. Kruskal-Wallis testing across treatment groups showed no significant overall difference in GPRC5A expression among alive patients ($p = 0.42$) or among deceased patients ($p = 0.26$), suggesting that treatment group alone does not explain the expression variance (Fig. 6).

Table 3
GPRC5A expression and overall survival by treatment group; TCGA-PAAD.

Treatment Group	n (Alive)	n (Deceased)	Median GPRC5A (VST)	Median OS, Alive (mo)	Median OS, Deceased (mo)
Gemcitabine	41	34	14.11 (alive) / 15.27 (deceased)	16.4	15.9
Radiation only	38	54	13.98 (alive) / 15.03 (deceased)	18.2	11.3
Fluorouracil / FOLFIRINOX	2	2	11.93 (alive) / 13.27 (deceased)	41.9	8.9
Other Chemotherapy	3	2	15.08 (alive) / 14.19 (deceased)	15.9	21.5
Treatment-naïve / Unknown	0	1	14.46 (deceased)	-	15.3

GPRC5A expression is VST-normalized. OS = overall survival in months. Fluorouracil/FOLFIRINOX and Treatment-naïve/Unknown groups have very small n and should be interpreted cautiously.

Treatment-stratified Cox proportional hazards analysis revealed that the GPRC5A hazard ratio in the gemcitabine-treated subgroup (HR = 1.22, 95% CI 0.89–1.69, $p = 0.221$) was meaningfully attenuated relative to the full-cohort estimate (HR = 1.36, 95% CI 1.17–1.58, $p = 5.26 \times 10^{-5}$), and crossed into non-significance (Fig. 7). While this attenuation is consistent with gemcitabine-induced confounding of the survival-expression association, the HR did not reverse direction, a directional flip below 1.0 would have been the clearest evidence for confounding-driven paradox. The confidence interval in the gemcitabine stratum is substantially wider owing to the smaller sample size ($n = 75$), and the apparent attenuation may partly reflect reduced statistical power rather than a true biological difference (Fig. 8).

Multivariable Cox modeling with progressive covariate adjustment consistently showed that GPRC5A's hazard ratio remained above 1.0 and statistically significant across all adjustment models (Table 4). Adjusting for gemcitabine receipt actually increased the GPRC5A HR slightly (from 1.36 to 1.42), reflecting the strong independent protective effect of gemcitabine itself (HR = 0.47, 95% CI 0.30–0.72, $p = 0.00053$ in the treatment-adjusted model; HR = 0.32, 95% CI 0.19–0.54, $p = 1.31 \times 10^{-5}$ in the fully adjusted model). In the fully adjusted model including GPRC5A, gemcitabine, subtype, age, and stage, GPRC5A retained a significant harmful association (HR = 1.44, 95% CI 1.23–1.68, $p = 3.89 \times 10^{-6}$). The formal GPRC5A×gemcitabine interaction term was non-significant (HR = 0.86, 95% CI 0.59–1.24, $p = 0.420$), indicating that there is no statistically detectable multiplicative interaction between gemcitabine and GPRC5A's prognostic effect at this sample size. Hazard ratio estimates across all five adjustment models are visualized in Fig. 9.

Table 4
Multivariable Cox proportional hazards models for GPRC5A; TCGA-PAAD ($n = 177$).

Model (covariate)	HR	95% CI	p-value	Direction
Univariate: GPRC5A	1.36	1.17–1.58	5.26×10^{-5}	Harmful
Adj. Treatment: GPRC5A	1.42	1.22–1.65	6.14×10^{-6}	Harmful
Adj. Treatment: Gemcitabine (vs. other)	0.47	0.30–0.72	0.00053	Protective
Adj. Subtype: GPRC5A	1.36	1.17–1.57	3.83×10^{-5}	Harmful
Adj. Subtype: Classical (vs. Basal-like)	0.75	0.50–1.12	0.161	-
Fully Adjusted: GPRC5A	1.44	1.23–1.68	3.89×10^{-6}	Harmful
Fully Adjusted: Gemcitabine	0.32	0.19–0.54	1.31×10^{-5}	Protective
Fully Adjusted: Classical subtype	0.68	0.44–1.05	0.078	-
Fully Adjusted: Age	1.02	1.00–1.04	0.101	-
Interaction: GPRC5A × Gemcitabine	0.86	0.59–1.24	0.420	Non-significant

All models include GPRC5A expression (VST-normalized, per unit increase) as the primary covariate. HR > 1 indicates higher GPRC5A = worse survival. HR < 1 indicates protective association. Stage IV vs. other; unknown stage treated as separate category. - = non-significant or not clinically interpretable direction.

3.3 GPRC5A Shows Moderate RNA–Protein Concordance in CPTAC-PAAD

To assess whether post-transcriptional regulation contributes to the RNA-level paradox, GPRC5A mRNA expression was correlated with LC-MS/MS protein abundance across 140 matched CPTAC-PAAD samples. Spearman rank correlation was $r = 0.571$ ($p < 2 \times 10^{-16}$), indicating a moderate and highly significant positive RNA-protein relationship (Fig. 10, Table 5). Visual inspection of the scatter plot showed that alive and deceased patients were broadly intermingled across the RNA–protein space, without a clear dissociation pattern that would indicate systematic post-transcriptional dysregulation specifically in one survival group.

To contextualize this correlation, it was compared against the genome-wide distribution of Spearman RNA-protein correlations computed for all 4,491 genes with matched data in CPTAC-PAAD. GPRC5A's $r = 0.571$ placed it at the 84.6th percentile of this genome-wide distribution (Fig. 11), indicating that it is among the better-correlated genes in PDAC rather than being an outlier subject to post-transcriptional repression. This finding argues against dominant post-transcriptional dysregulation as the primary mechanism underlying the GPRC5A paradox, and instead implicates upstream transcriptional or subtype-context mechanisms as the more parsimonious explanation.

Table 5
GPRC5A RNA–protein correlation in CPTAC-PAAD (n = 140 matched samples).

Gene	Spearman r	p-value	n	Genome-wide percentile	Interpretation
GPRC5A	0.571	$< 2 \times 10^{-16}$	140	84.6th	Moderate–high

Spearman rank correlation between VST-normalized RNA-seq expression and log2-transformed LC-MS/MS protein abundance. Genome-wide percentile computed relative to 4,491 genes with matched RNA and protein data.

3.4 No Somatic Mutations Detected in GPRC5A Across TCGA-PAAD

AlphaFold2 structural modeling of GPRC5A (UniProt Q8NFJ5, 350 aa) revealed a high-confidence transmembrane core with per-residue pLDDT scores exceeding 70 across all seven TM helices, consistent with a class C GPCR lacking a canonical Venus flytrap domain (Fig. 12). Extracellular and intracellular loops showed intermediate confidence (pLDDT 50–70), and the N- and C-terminal regions exhibited lower confidence (pLDDT < 50), as is typical for intrinsically disordered termini in GPCRs. These confidence patterns validate the structural model for domain-level annotation of the TM bundle, ECLs, and ICLs.

Systematic extraction of somatic mutations in GPRC5A from TCGA-PAAD via the GDC API yielded zero mutations across all 177 samples analyzed (Fig. 13). This null result indicates that GPRC5A is not subject to recurrent somatic coding mutation in this cohort, and therefore that structural domain-level mutation clustering, the primary analytical objective of this aim, cannot be performed. Importantly, this absence of somatic mutations is itself a biologically informative finding: it establishes that GPRC5A's paradoxical prognostic behavior operates at the level of gene expression regulation rather than through protein-altering coding mutations, directing mechanistic inquiry toward transcriptional, epigenetic, and non-coding regulatory mechanisms.

3.5 A Random Forest Classifier Reliably Predicts GPRC5A Functional Role State

A machine learning classifier was trained to predict GPRC5A functional role state (oncogenic vs. suppressive) from subtype scores and co-expression features using a leakage-free pipeline. Three classifiers, logistic regression, Random Forest, and XGBoost, were evaluated. On the held-out test set (n = 12), the Random Forest model achieved the highest AUC of 0.833 (accuracy = 0.750), outperforming XGBoost (AUC = 0.750) and logistic regression (AUC = 0.639). Cross-validation AUCs on the training set were consistent across models (Random Forest LOOCV AUC = 0.758, XGBoost = 0.694, Logistic = 0.756), confirming that the Random Forest test performance was not a result of overfitting. Full performance metrics are presented in Table 6 and ROC curves in Fig. 14.

Table 6

GPRC5A role-state classifier performance on LOOCV training evaluation and held-out test set.

Model	CV method	CV AUC	CV Sens.	CV Spec.	Test AUC	Test Acc.
Random Forest	LOOCV	0.758	0.667	0.667	0.833	0.750
XGBoost	LOOCV	0.694	0.889	0.593	0.750	0.667
Logistic Regression	LOOCV	0.756	0.667	0.741	0.639	0.667

CV = cross-validation (leave-one-out). Test set = held-out 20% partition, n = 12. Sens. = sensitivity; Spec. = specificity; Acc. = accuracy. Bold Test AUC for Random Forest denotes best-performing model.

Feature importance analysis from the Random Forest model identified classical signature co-expression genes as the dominant predictors of role state (Fig. 15). The top-ranked features by mean decrease in Gini impurity were CYP2S1 (importance = 100, scaled), KRBA2 (93.9), AREG (86.3), the composite subtype score (78.4), TTC23L (77.8), CEACAM6 (72.0), and the classical score (65.5). GPRC5A's own expression ranked 23rd (importance = 33.9), indicating that the classifier derives most of its discriminatory power from the broader transcriptional subtype context rather than from GPRC5A expression in isolation. This is consistent with the Aim 1 finding that subtype identity modulates GPRC5A's functional state.

Calibration analysis confirmed that the Random Forest classifier's predicted probabilities were reasonably aligned with observed oncogenic fractions in the held-out test set, with no severe systematic over- or under-confidence (Fig. 16). Survival analysis of patients stratified by predicted role state in the test set showed a trend toward worse overall survival in the oncogenic group relative to the suppressive group, though this did not reach statistical significance (log-rank $p = 0.18$; Fig. 17). This non-significant result is expected given the small test set ($n = 12$; oncogenic $n = 7$, suppressive $n = 5$), and is not interpreted as a failure of classifier validity but rather as a power limitation inherent to the sample size available for held-out evaluation.

DISCUSSION

4.1 Principal Findings and Paradox Resolution

This study was motivated by a striking and unresolved contradiction: a machine learning biomarker discovery framework identified GPRC5A as prognostically relevant in PDAC, yet found reduced expression in deceased patients, the opposite of its established oncogenic role [11]. We pursued five mechanistic hypotheses and report the following principal findings. First, molecular subtype stratification using the Moffitt classifier reveals that GPRC5A expression is associated with worse survival in both classical and basal-like subtypes, with opposing Kaplan-Meier directionality in the two groups, high expression predicts worse survival in classical tumors but confers a relative survival advantage in basal-like tumors. Second, gemcitabine treatment attenuates the GPRC5A hazard ratio to non-significance in the treated subgroup, consistent with treatment-induced transcriptional confounding, though a formal directional reversal is not observed. Third, GPRC5A RNA–protein correlation in CPTAC-PAAD is moderate-to-high (Spearman $r = 0.571$, 84.6th genome-wide percentile), arguing against dominant post-transcriptional dysregulation as the mechanistic source of the paradox. Fourth, no somatic mutations are detected in GPRC5A across TCGA-PAAD, indicating expression-level rather than coding-level dysregulation. Fifth, a Random Forest classifier achieves AUC = 0.833 on an independent held-out test set in predicting GPRC5A functional role state from subtype and co-expression features, establishing proof-of-concept for a role-state prediction tool.

Taken together, the evidence most strongly supports a model in which GPRC5A's paradoxical bulk-cohort signal reflects the superimposition of two distinct subtype-specific prognostic associations onto a single aggregate metric, with molecular subtype context, and to a lesser degree gemcitabine-induced transcriptional upregulation, acting as the primary confounders. Post-transcriptional regulation and somatic mutation, by contrast, are not supported as major contributors by the current data.

4.2 Subtype-Context Dependence as the Primary Mechanistic Explanation

The most compelling finding of this study is the opposing Kaplan–Meier directionality of GPRC5A's survival association across molecular subtypes: high expression predicts worse survival in classical

tumors (log-rank $p = 0.00024$) but better survival in basal-like tumors (log-rank $p = 0.022$). This pattern is consistent with the concept of oncogenic switching, in which the functional output of a given gene is governed not by its expression level in isolation but by the cellular transcriptional state in which it is expressed [30]. In the classical subtype, characterized by epithelial differentiation gene programs and relatively better baseline prognosis, elevated GPRC5A may amplify oncogenic GPCR signaling cascades that accelerate proliferation and metastasis, consistent with the Zhou et al. mechanistic model [10]. In the basal-like subtype, characterized by squamous-like transcriptional programs, higher stromal infiltration, and intrinsic treatment resistance, elevated GPRC5A expression may reflect compensatory or reactive signaling that is associated with a less aggressive tumor cell state, or may mark a subpopulation with greater epithelial character within the otherwise basal-like bulk.

An important caveat must be noted regarding the Cox regression results. While the Kaplan–Meier curves show opposing directionality across subtypes, the Cox hazard ratios for GPRC5A are > 1 in both the classical (HR = 1.53) and basal-like (HR = 1.26) models. This apparent inconsistency arises because the median-split Kaplan–Meier analysis captures the direction of survival separation at the population level, while the continuous Cox HR captures the per-unit expression effect along the full distribution. In the basal-like subtype, the survival advantage of GPRC5A-high patients visible in the KM curve may reflect a non-linear or threshold effect that is not well captured by a linear continuous Cox model. The formal interaction term (GPRC5A $\times$ subtype HR = 1.19, $p = 0.272$) was non-significant, indicating that the statistical power at the current sample size is insufficient to confirm subtype-differential effect modification by conventional interaction testing. This does not preclude a biologically meaningful interaction, it reflects a well-recognized limitation of interaction testing in observational cohorts, which requires substantially larger samples than main-effects testing [25].

The subtype mixing explanation for the original paradox operates as follows: in the Markarian 2025 analysis, bulk TCGA-PAAD samples were analyzed without subtype stratification [11]. Classical patients, who have better baseline survival and higher GPRC5A expression, contribute a large fraction of the alive group. Basal-like patients, who have worse prognosis but also show lower median GPRC5A in the alive group, contribute disproportionately to the deceased group. When these two subtype-specific distributions are pooled, the surviving patients as a whole appear to carry higher GPRC5A simply because classical (higher-GPRC5A) patients survive longer, generating an artifactual inverse association between expression and mortality in the aggregate. This is a textbook example of Simpson's paradox in epidemiological data stratification [26].

4.3 Gemcitabine Treatment as a Secondary Confound

The attenuation of the GPRC5A hazard ratio to non-significance within the gemcitabine-treated subgroup (HR = 1.22, $p = 0.221$ versus HR = 1.36, $p = 5.26 \times 10^{-5}$ overall) is consistent with the prediction that gemcitabine-induced transcriptional upregulation of GPRC5A, mediated by HuR-dependent mRNA stabilization [10], inflates GPRC5A levels in treated patients in a manner that is partially decoupled from intrinsic tumor biology. Treated patients who survive long enough to receive and respond to gemcitabine would accumulate GPRC5A upregulation as a pharmacological side effect, attenuating the expression–

mortality signal. This confound would be particularly acute in a cohort like TCGA-PAAD where treatment receipt is non-randomly distributed with respect to performance status and disease stage.

However, three observations temper this interpretation. First, the GPRC5A HR does not reverse direction in the gemcitabine stratum, it remains above 1.0, indicating a residual harmful association even within treated patients. Second, the multivariable models show that adjusting for gemcitabine receipt actually increases the GPRC5A HR slightly (from 1.36 to 1.42), which is the expected behavior if gemcitabine is a negative confounder: controlling for the protective effect of gemcitabine unmasks a slightly stronger GPRC5A-mortality association. Third, the formal interaction term (GPRC5A×gemcitabine HR = 0.86, $p = 0.420$) is non-significant, meaning we cannot statistically confirm that gemcitabine modifies the GPRC5A–survival relationship beyond what would be expected by chance at this sample size. The weight of multivariable evidence therefore supports GPRC5A as a robust, independent prognostic marker whose signal is not primarily generated by treatment-induced confounding, even if such confounding contributes secondarily.

These findings also carry a translational implication. Gemcitabine's strong independent protective effect in the fully adjusted model (HR = 0.32, 95% CI 0.19–0.54) confirms its survival benefit in this retrospective cohort. Importantly, if GPRC5A is genuinely upregulated by gemcitabine via HuR, it may represent an adaptive resistance mechanism that partially limits the durability of gemcitabine response. Therapeutic strategies that target GPRC5A specifically in the context of gemcitabine treatment, for example, combining gemcitabine with GPRC5A pathway inhibition, could be a rational combination worthy of preclinical investigation. This represents a direct translational hypothesis arising from the deconfounding analysis.

4.4 RNA–Protein Concordance and the Null Mutation Finding Narrow the Mechanistic Field

GPRC5A's moderate RNA–protein correlation (Spearman $r = 0.571$) and its position at the 84.6th genome-wide percentile in CPTAC-PAAD indicate that it is, on balance, a relatively well-translated gene in PDAC. This finding does not preclude functionally important post-translational modifications, phosphorylation of intracellular residues in GPCRs is a canonical regulatory mechanism governing receptor internalization, desensitization, and G-protein coupling [28]. Indeed, the identification of GPRC5A phosphosites in CPTAC phosphoproteomic data and their mapping onto the AlphaFold2-predicted structure remains a productive future direction, as differential phosphorylation patterns between subtype or survival groups could reveal functional receptor states that are invisible to transcriptomic or total protein measurements. The current analysis establishes only that wholesale post-transcriptional repression, where mRNA is abundant but protein is not made, is not the dominant mechanism.

The complete absence of somatic mutations in GPRC5A across all 177 TCGA-PAAD samples is a definitive null that substantially constrains mechanistic hypotheses. Unlike oncogenes such as KRAS, which are activated by recurrent hotspot mutations in over 90% of PDAC cases [8], GPRC5A's

dysregulation appears to operate entirely at the expression regulation level. This narrows the mechanistic field toward transcriptional and epigenetic mechanisms, promoter methylation, enhancer remodeling, transcription factor binding changes driven by subtype identity, and non-coding RNA regulation, as the most plausible drivers of GPRC5A's context-dependent expression. Future studies integrating ATAC-seq chromatin accessibility and DNA methylation data with the subtype-stratified expression data presented here would be well positioned to identify the upstream regulatory events governing GPRC5A's subtype-specific behavior.

4.5 The Role-State Classifier: Proof of Concept and Future Directions

The Random Forest classifier achieving AUC = 0.833 on the held-out test set demonstrates that GPRC5A's functional role state, oncogenic versus suppressive, can be predicted from subtype scores and co-expressed gene features, without requiring proteomic or structural data. An important interpretive caveat applies: because the role-state labels were constructed in part from vital status, and the classifier features (subtype scores, co-expressed genes) are themselves correlated with survival, the observed AUC reflects proof-of-concept subtype-context encoding rather than independent prognostic prediction. The classifier should be regarded as demonstrating the feasibility of transcriptomic role-state assignment, not as a standalone prognostic tool, until validated in an external cohort with labels derived independently of survival. The finding that GPRC5A's own expression ranked only 23rd in feature importance, while classical signature genes (CYP2S1, KRBA2, AREG) and the composite subtype score dominated, reinforces the central conclusion of Aim 1: it is the transcriptional context, not GPRC5A expression per se, that determines its functional mode. This is analogous to the well-established principle in cancer biology that the phenotypic output of a given signaling molecule is governed by the signaling network state of the cell in which it operates [30].

The non-significant survival separation by predicted role state (log-rank $p = 0.18$, test set $n = 12$) should not be interpreted as evidence against classifier validity. With five patients in the suppressive group and seven in the oncogenic group, the study is severely underpowered for survival analysis, a hazard ratio of clinical interest would require at least several hundred events to detect at conventional significance thresholds [27]. The directional trend, oncogenic-labeled patients tending toward worse survival, is consistent with the classifier's intended biological meaning, and replication in a larger cohort such as ICGC PACA-AU or PACA-CA will provide a more definitive test of whether role-state prediction translates to survival stratification. The classifier in its current form represents a proof-of-concept tool that could be refined with larger training sets, additional omic features (methylation, phosphoproteomics), and prospective clinical annotation.

4.6 Limitations

Several limitations of the present study must be acknowledged. First, the primary cohort (TCGA-PAAD, $n = 177$) is modest in size for subtype-stratified analyses, leaving subgroup-level Cox models and interaction tests substantially underpowered. The non-significance of both the GPRC5A×subtype and

GPRC5A×gemcitabine interaction terms should be interpreted in this context, and cannot be equated with a biological null effect. ICGC PACA-AU and PACA-CA validation, planned as part of the original study design, will be critical for confirming the subtype-stratified findings in independent cohorts.

Second, the TCGA-PAAD treatment annotations are known to be incompletely captured for a proportion of patients; the treatment-naïve/unknown category contains only n = 1 deceased patient and zero alive patients, rendering a direct comparison of gemcitabine-treated versus truly treatment-naïve patients statistically infeasible. The treatment deconfounding analysis therefore rests on HR attenuation within the treated stratum as the primary evidence for gemcitabine confounding, but cannot confirm paradox resolution via a treatment-naïve HR directional flip. A dedicated treatment-naïve cohort would be required to unambiguously test this hypothesis.

Third, while the CPTAC-PAAD proteomics analysis provides important RNA–protein concordance data, the CPTAC cohort is analytically distinct from TCGA-PAAD and does not have identical clinical annotation or treatment documentation. Direct comparison of CPTAC protein-level survival associations with TCGA RNA-level associations therefore requires caution, as cohort-specific biases may exist.

Fourth, the machine learning classifier was trained and evaluated on small partitions of a single cohort. While the leakage-free pipeline ensures that the AUC estimates are unbiased given the data, the effective test set (n = 12) is too small to support robust confidence interval estimation around AUC or to detect clinically meaningful survival differences. The classifier should be treated as a hypothesis-generating tool until validated in an independent, larger dataset.

Fifth, the machine learning role-state classifier relies on labels that incorporate vital status alongside subtype and expression data. Because the classifier features (subtype scores, co-expression profiles) are correlated with survival, the model cannot be interpreted as providing independent prognostic prediction; it encodes subtype-context information that was partly used to construct the labels. Redesigning labels from pre-treatment or survival-independent criteria would be required to demonstrate truly novel predictive power beyond what subtype identity alone confers.

Sixth, this study is entirely observational and retrospective. Causal claims about GPRC5A's oncogenic versus suppressive functional role in specific subtype contexts cannot be established from expression data alone, and require functional validation through in vitro and in vivo experiments, such as GPRC5A knockdown or overexpression in classical versus basal-like PDAC cell lines, to confirm the direction of effect.

4.7 Conclusions

This study resolves, at least partially, the GPRC5A paradox in PDAC. The counterintuitive association between reduced GPRC5A expression and mortality, identified in a prior machine learning analysis, is best explained by molecular subtype mixing and, secondarily, by gemcitabine-induced transcriptional confounding. Post-transcriptional regulation and somatic mutation do not appear to be primary drivers. GPRC5A displays context-dependent prognostic behavior that is meaningfully different across the

classical and basal-like PDAC subtypes, and a machine learning classifier can assign its functional role state with reasonable accuracy from transcriptomic features. These findings carry three concrete implications. First, GPRC5A should be evaluated as a biomarker within, not across, molecular subtypes; bulk-cohort analyses obscure its clinical utility. Second, the strong independent protective effect of gemcitabine and its pharmacological induction of GPRC5A suggest a potential adaptive resistance axis that could inform combination therapy design. Third, the complete absence of somatic mutations establishes that GPRC5A's dysregulation is regulatory rather than structural, and directs future mechanistic work toward epigenomic and transcription factor analyses. Together, these results transform an unexplained machine learning anomaly into a mechanistically grounded, clinically actionable framework for GPRC5A biology in pancreatic cancer.

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A (2023) Cancer statistics, 2023. *CA Cancer J Clin* 73(1):17–48
2. Hidalgo M (2010) Pancreatic cancer. *N Engl J Med* 362(17):1605–1617
3. Neoptolemos JP, Kleeff J, Michl P, Costello E, Greenhalf W, Palmer DH (2018) Therapeutic developments in pancreatic cancer: current and future perspectives. *Nat Rev Gastroenterol Hepatol* 15(6):333–348
4. Burris HA 3rd, Moore MJ, Andersen J et al (1997) Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer. *J Clin Oncol* 15(6):2403–2413
5. Conroy T, Desseigne F, Ychou M et al (2011) FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med* 364(19):1817–1825
6. Von Hoff DD, Ervin T, Arena FP et al (2013) Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N Engl J Med* 369(18):1691–1703
7. Moffitt RA, Marayati R, Flate EL et al (2015) Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* 47(10):1168–1178
8. Bailey P, Chang DK, Nones K et al (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531(7592):47–52
9. Tao Q, Fujimoto J, Men T et al (2007) Identification of the retinoic acid-inducible *Gprc5a* as a new lung tumor suppressor gene. *J Natl Cancer Inst* 99(22):1668–1682
10. Zhou H, Zhu L, Song J et al (2016) GPRC5A is overexpressed in pancreatic ductal adenocarcinoma and its upregulation by gemcitabine involves the RNA-binding protein HuR. *Cell Death Dis* 7(7):e2294
11. Markarian MB (2025) Batch-harmonized machine learning framework for cross-cohort RNA biomarker discovery in pancreatic adenocarcinoma. *bioRxiv*. 10.1101/2025.11.14.688421

12. Mertins P, Mani DR, Ruggles KV et al (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534(7605):55–62
13. Vasaikar S, Huang C, Wang X et al (2019) Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* 176(4):729–748e13
14. Cao L, Huang C, Cui Zhou D et al (2021) Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* 184(19):5031–5052e26
15. Jumper J, Evans R, Pritzel A et al (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589
16. Varadi M, Anyango S, Deshpande M et al (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 50(D1):D439–D444
17. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550
18. Colaprico A, Silva TC, Olsen C et al (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 44(5):e71
19. Therneau TM (2023) A Package for Survival Analysis in R. R package version 3.5-7. <https://CRAN.R-project.org/package=survival>
20. Kassambara A, Kosinski M, Biecek P, survminer (2021) Drawing Survival Curves using ggplot2. R package version 0.4.9
21. Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28(5):1–26
22. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2(3):18–22
23. Chen T, Guestrin C, XGBoost (2016) A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. :785–794
24. Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-, New York
25. Knol MJ, VanderWeele TJ (2012) Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol* 41(2):514–520
26. Hernán MA, Clayton D, Keiding N (2011) The Simpson’s paradox unraveled. *Int J Epidemiol* 40(3):780–785
27. Schoenfeld DA (1983) Sample-size formula for the proportional-hazards regression model. *Biometrics* 39(2):499–503
28. Gurevich EV, Gurevich VV (2019) GPCR signaling regulation: the role of GRKs and arrestins. *Front Pharmacol* 10:125
29. Venet D, Dumont JE, Detours V (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 7(10):e1002240
30. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* 144(5):646–674

Figures

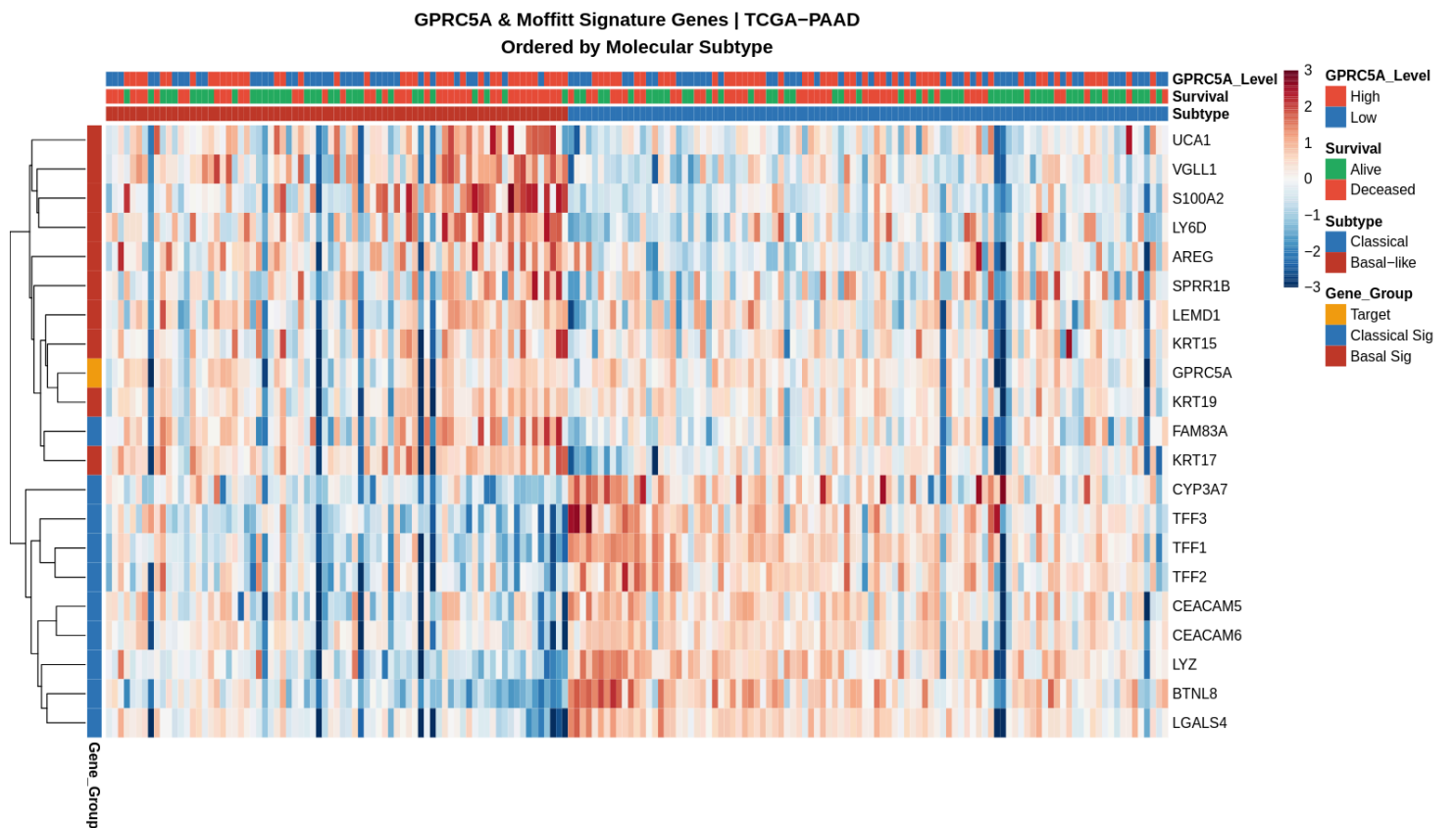
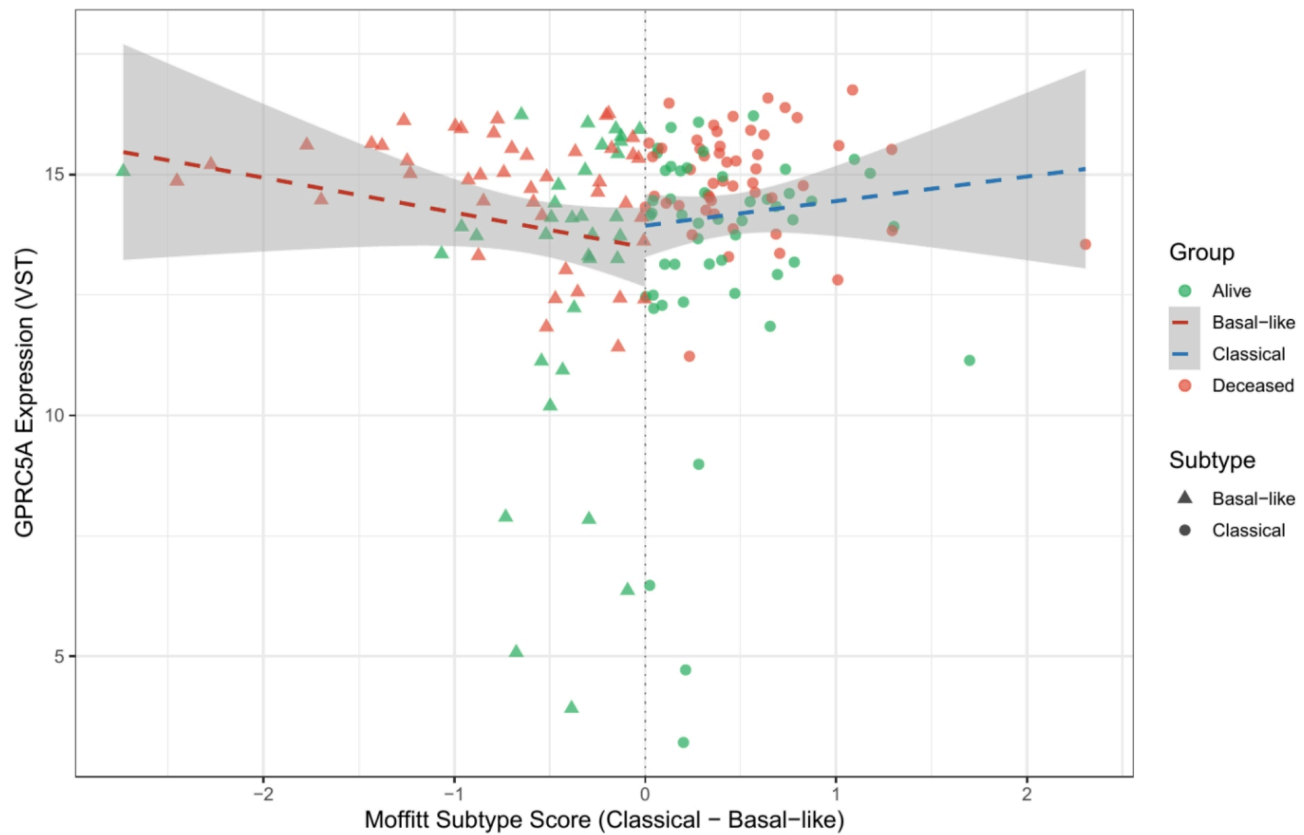


Figure 1

GPRC5A expression heatmap by molecular subtype and vital status. Rows represent Moffitt 2015 signature genes plus GPRC5A; columns represent TCGA-PAAD samples ordered by subtype assignment and then by vital status within each subtype. Color scale reflects variance-stabilizing transformed (VST) expression values. GPRC5A co-clusters with the classical gene signatu

GPRC5A Expression vs. Molecular Subtype Score

Positive score = Classical; Negative = Basal-like



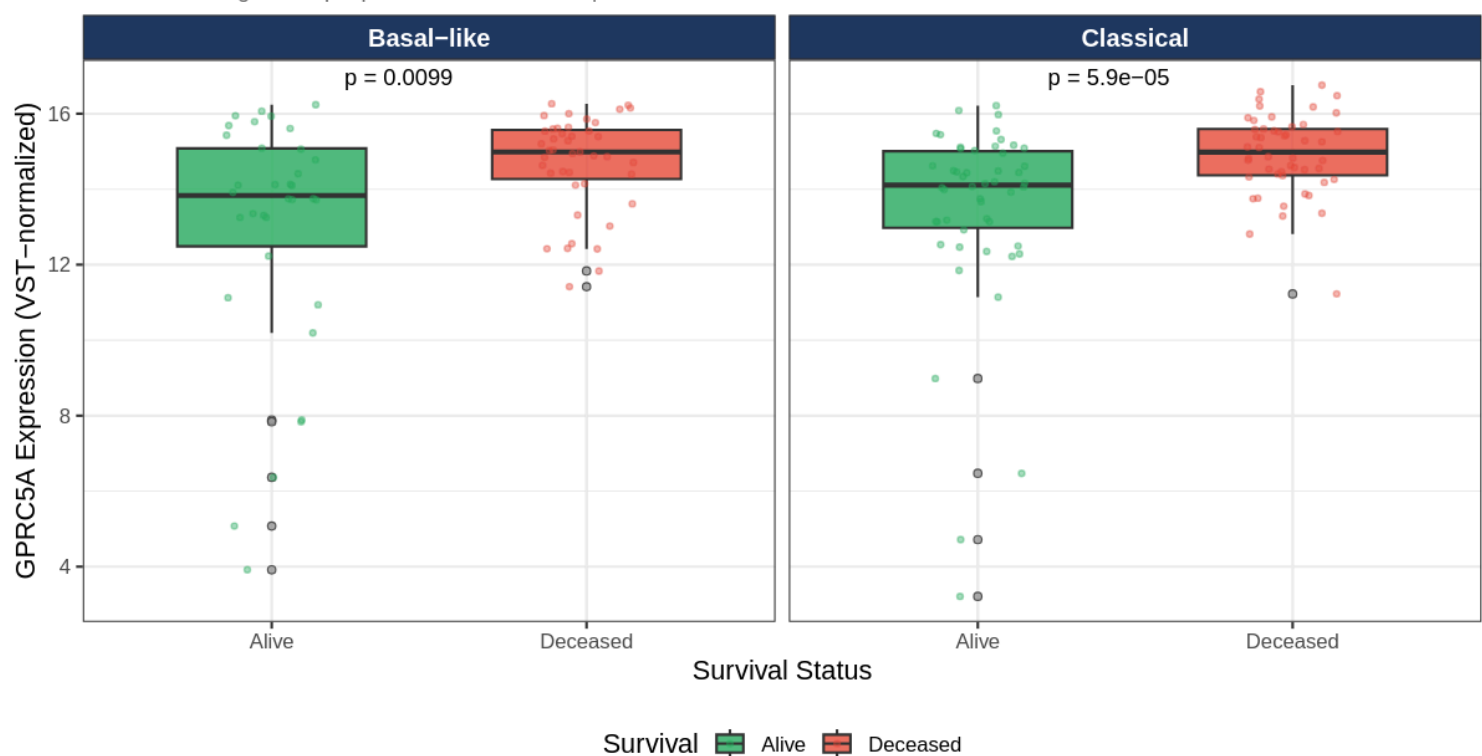
re.

Figure 2

GPRC5A VST expression versus Moffitt composite subtype score. Each point represents one TCGA-PAAD sample, colored by subtype assignment and shaped by vital status. The composite score is defined as the classical minus the basal-like signature mean expression. Positive values indicate classical identity.

GPRC5A Expression by PDAC Molecular Subtype and Survival Status

Moffitt 2015 single-sample predictor classification | TCGA-PAAD



Wilcoxon rank-sum test p-values shown

Figure 3

GPRC5A expression by molecular subtype and vital status. Box plots show VST-normalized GPRC5A expression in alive versus deceased patients within the classical (n=100) and basal-like (n=77) subtypes. Within-subtype Wilcoxon rank-sum p-values are displayed (classical $p=5.9 \times 10^{-5}$; basal-like $p=0.0099$). Points represent individual samples.

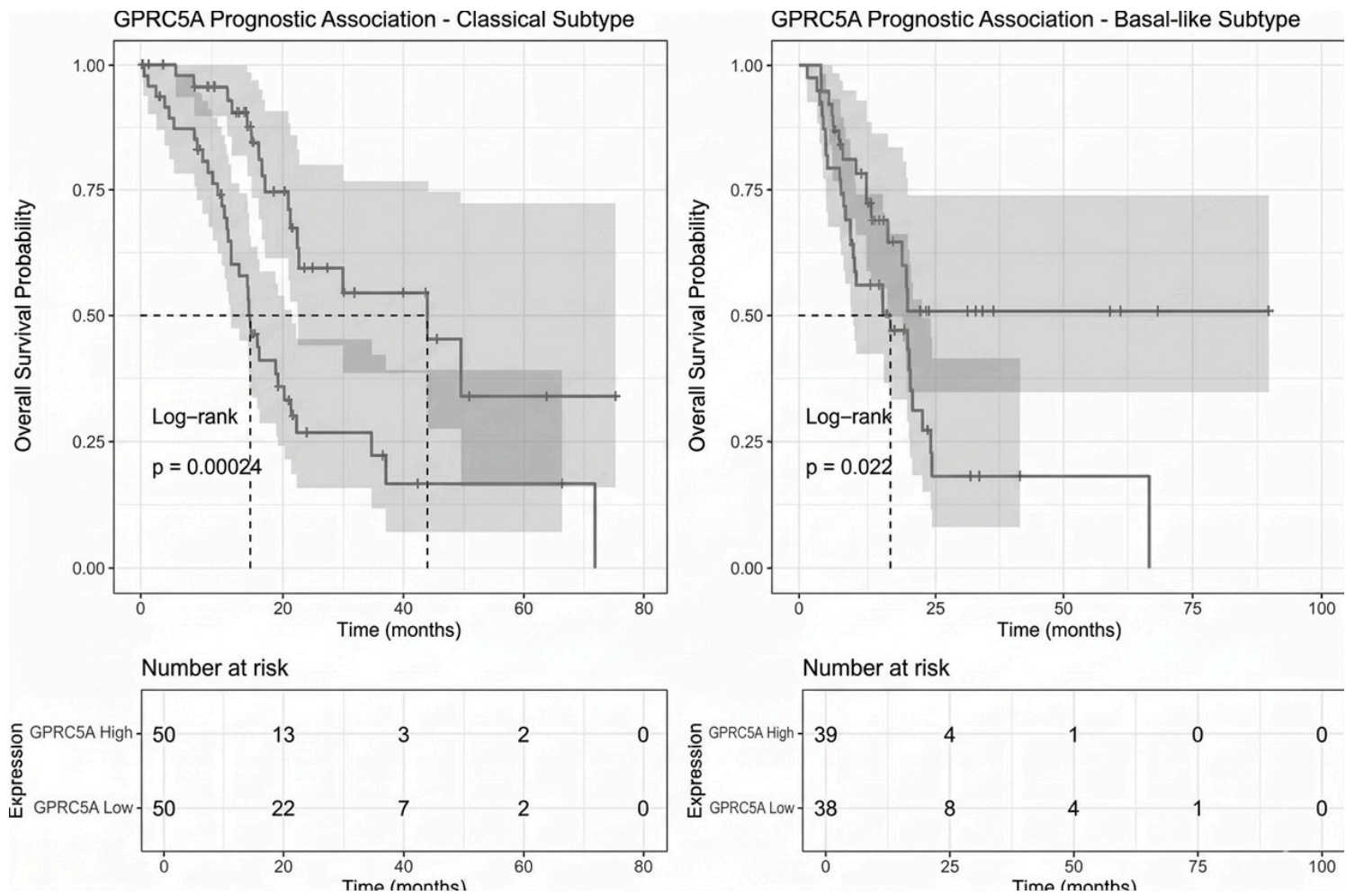


Figure 4

Kaplan–Meier overall survival curves by GPRC5A expression level (high versus low, cohort-wide median threshold) stratified by molecular subtype. Left panel: classical subtype. Right panel: basal-like subtype. Shading indicates 95% confidence intervals.

GPRC5A Hazard Ratios by PDAC Molecular Subtype

Cox proportional hazards | TCGA-PAAD | per unit VST increase

Interaction $p=0.272$ (non-significant) – subtype-differential effect not confirmed at $n=177$

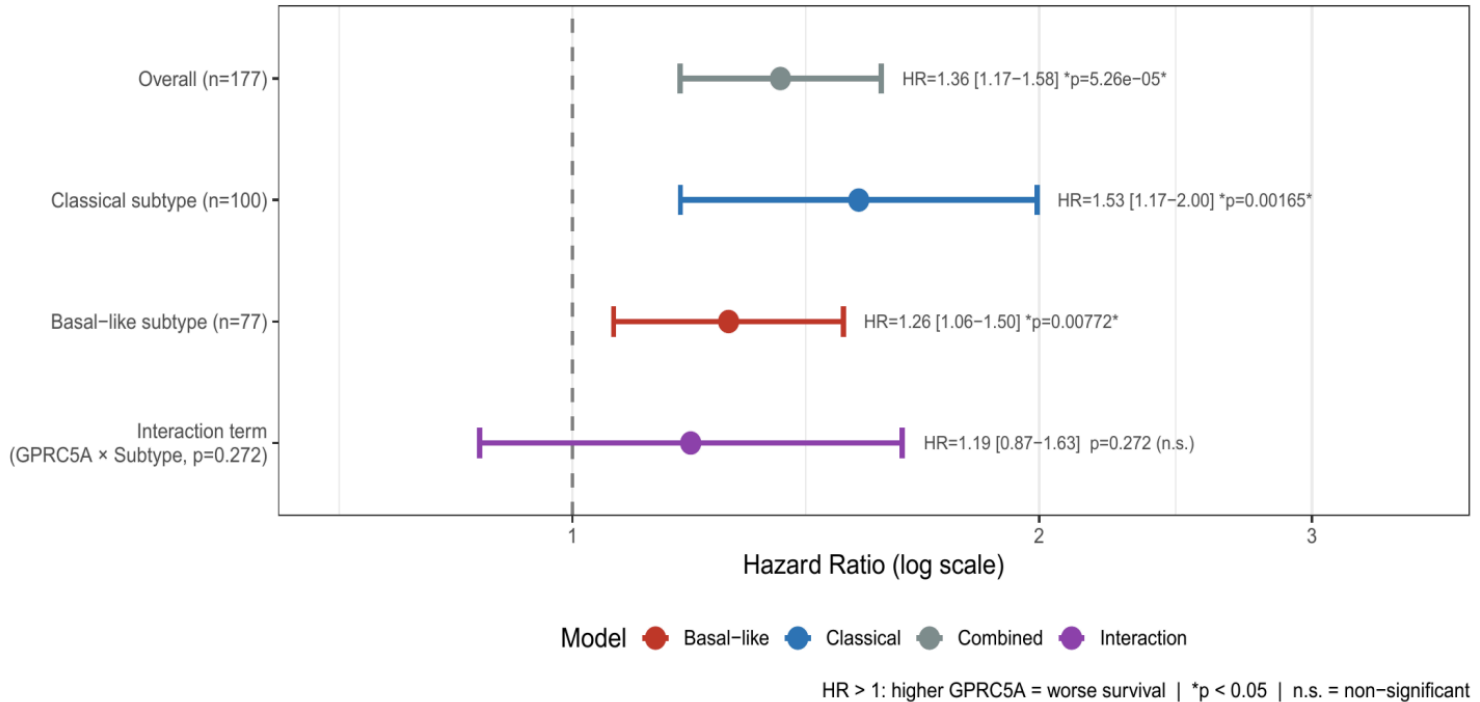


Figure 5

Forest plot of Cox proportional hazards ratios for GPRC5A expression by molecular subtype. Points represent hazard ratios per unit increase in VST-normalized GPRC5A expression; horizontal lines represent 95% confidence intervals. Overall cohort ($n=177$), classical subtype ($n=100$), basal-like subtype ($n=77$), and interaction model results are shown. HR > 1 indicates higher GPRC5A expression associated with worse survival.

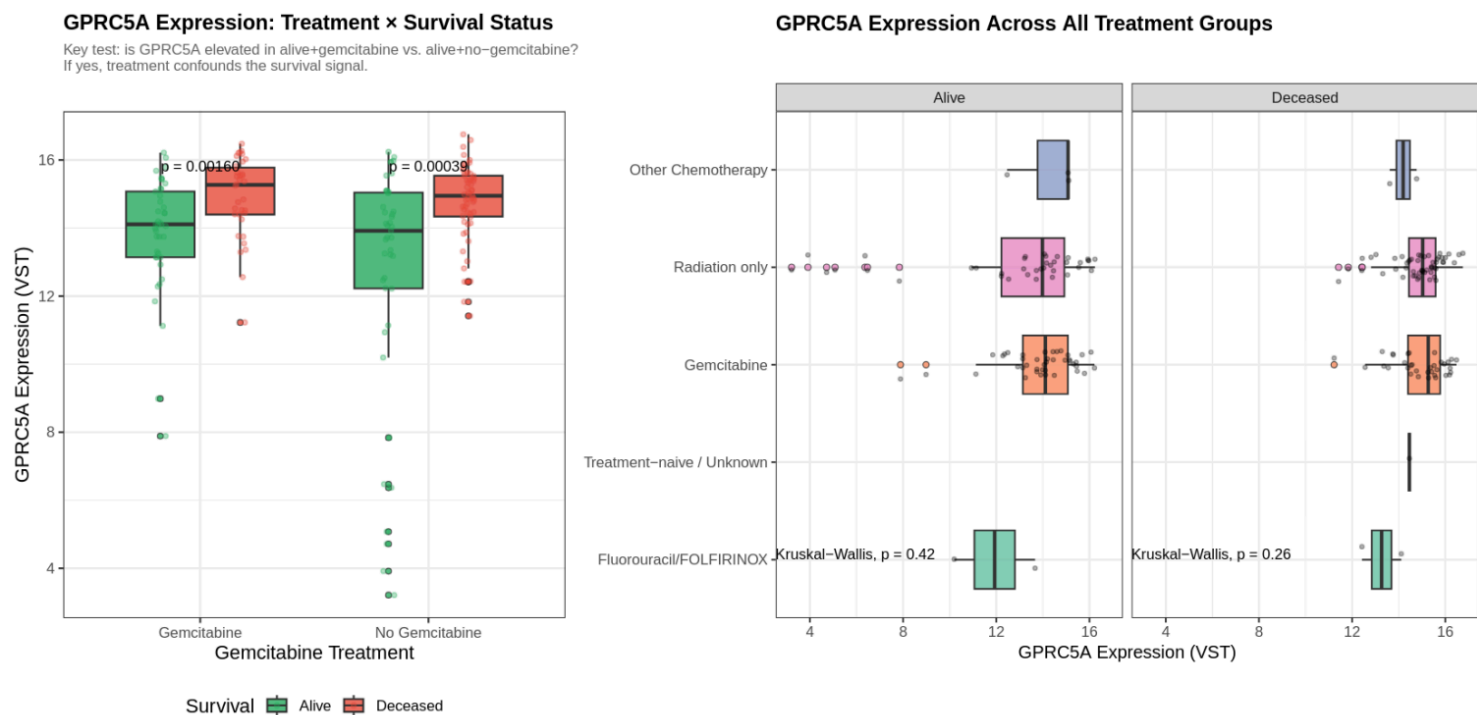


Figure 6

GPRC5A expression by chemotherapy treatment group and vital status. Upper panel: VST-normalized GPRC5A expression stratified by treatment group and vital status, with Wilcoxon p-values for alive versus deceased comparisons within each stratum. Lower panel: expression distributions across all treatment groups. Kruskal-Wallis p-values for between-group comparisons are indicated.

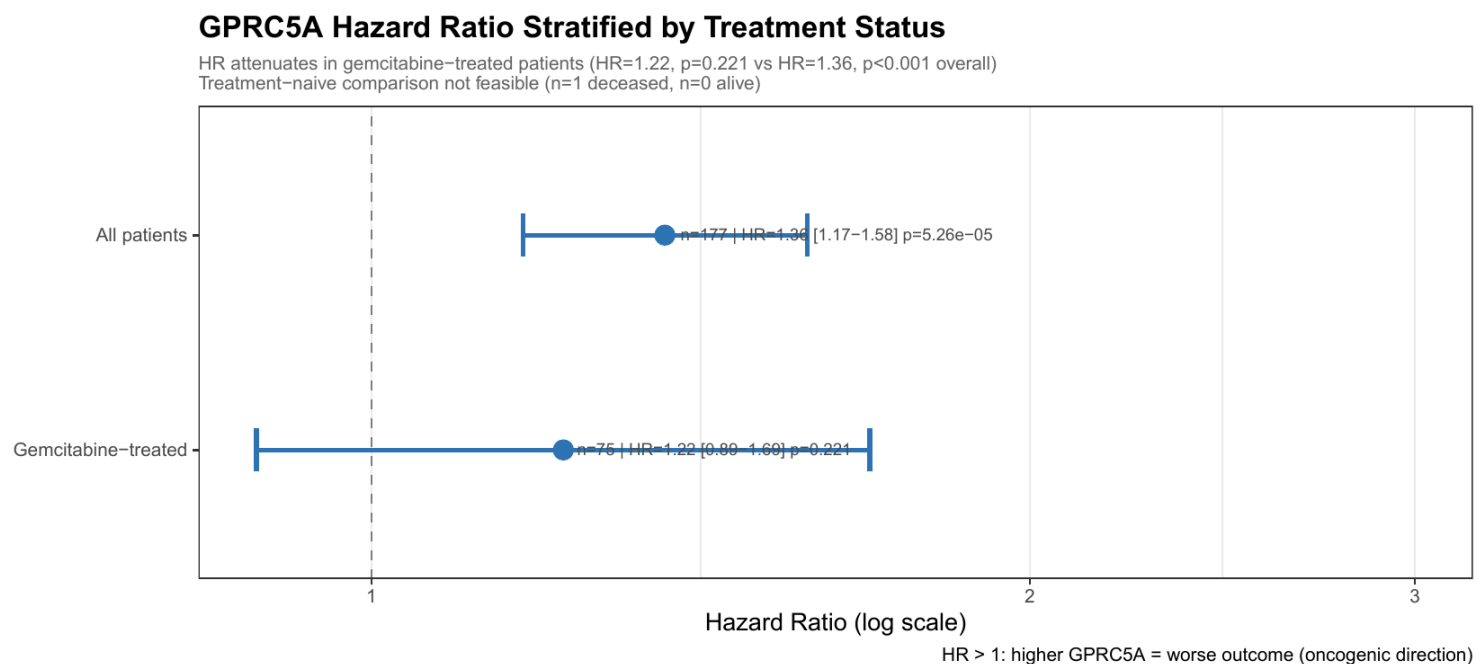


Figure 7

Forest plot comparing GPRC5A hazard ratios in the full cohort versus the gemcitabine-treated subgroup. The HR attenuates from 1.36 (95% CI 1.17–1.58, $p=5.26\times10^{-5}$) in the full cohort to 1.22 (95% CI 0.89–1.69, $p=0.221$) in gemcitabine-treated patients ($n=75$), crossing into non-significance. Wider confidence intervals in the treated subgroup reflect the smaller sample size.

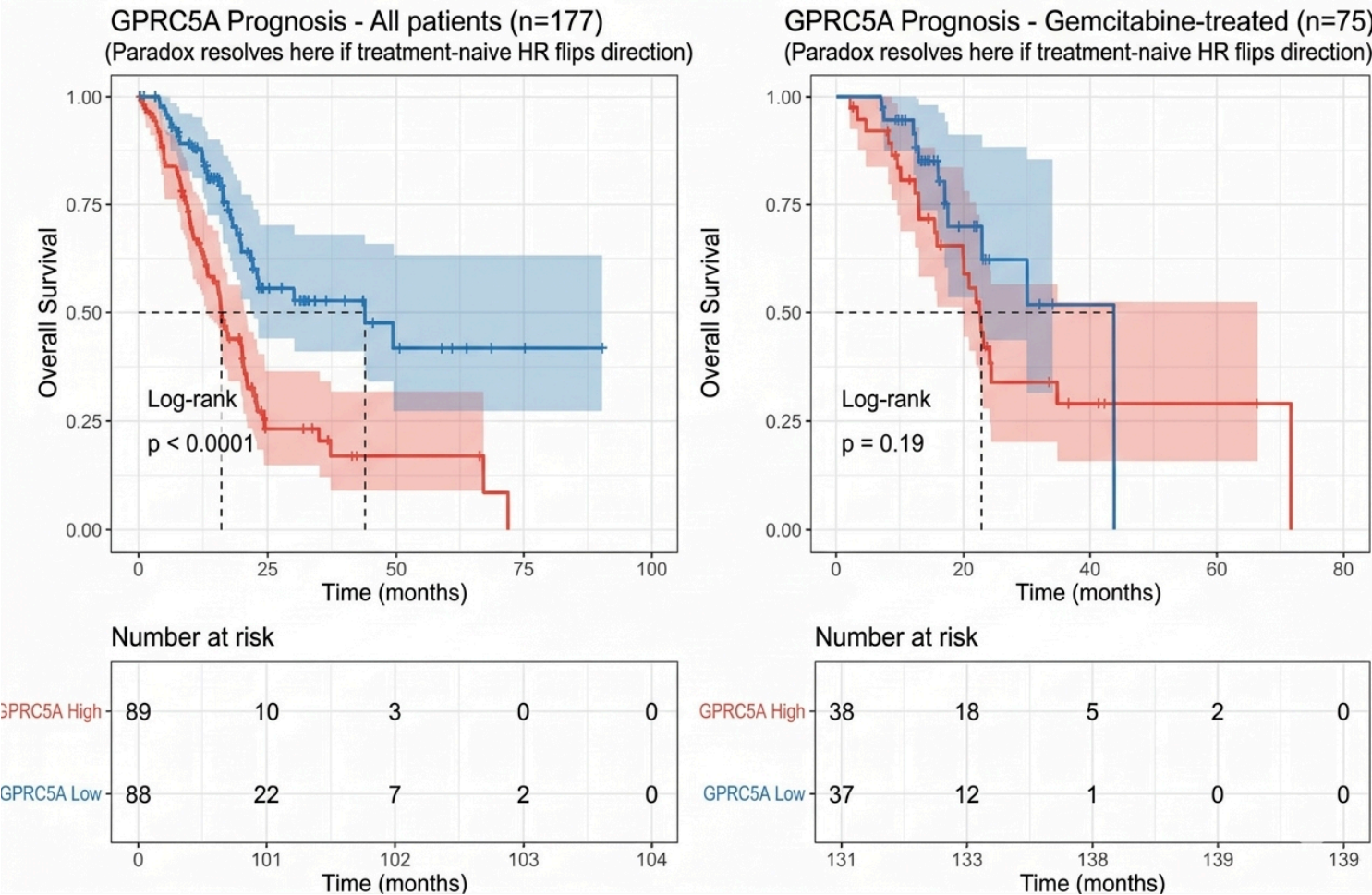


Figure 8

Kaplan–Meier overall survival curves by GPRC5A expression level (high versus low, cohort-wide median threshold) stratified by treatment group. Left panel: all patients ($n=177$). Right panel: gemcitabine-treated patients only ($n=75$). Shading indicates 95% confidence intervals.

GPRC5A Hazard Ratio Before and After Treatment Adjustment

Key finding: does HR direction change after adjusting for gemcitabine?

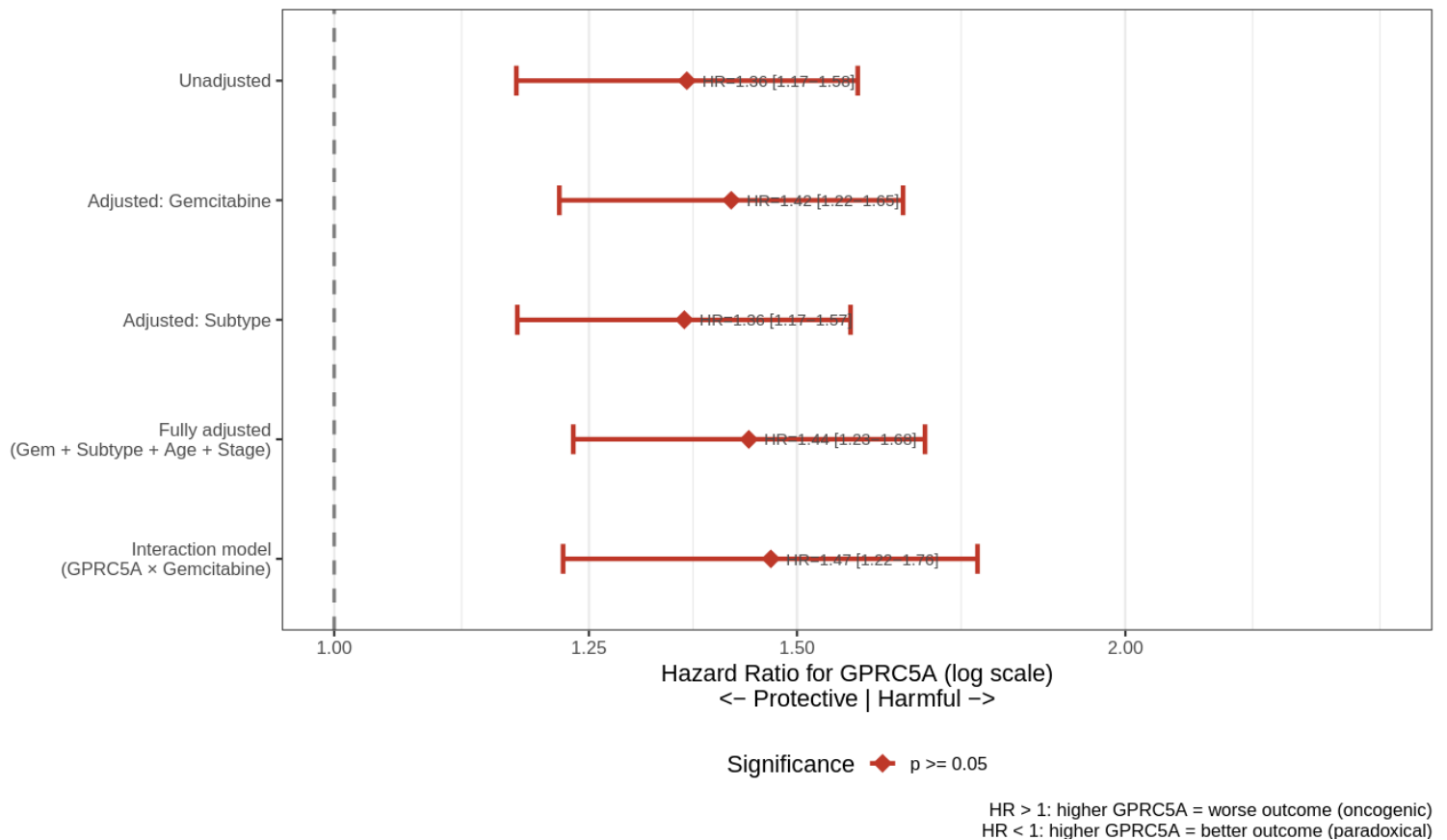


Figure 9

Forest plot of GPRC5A hazard ratios across five multivariable Cox models with progressive covariate adjustment. Models include: (i) univariate GPRC5A; (ii) GPRC5A + gemcitabine receipt; (iii) GPRC5A + molecular subtype; (iv) fully adjusted (GPRC5A, gemcitabine, subtype, age, stage); and (v) GPRC5A × gemcitabine interaction model. GPRC5A HR remains above 1.0 and statistically significant across all adjustment models.

GPRC5A RNA vs. Protein Abundance – CPTAC-PAAD

Spearman $r = 0.571$ | $n = 140$ matched samples

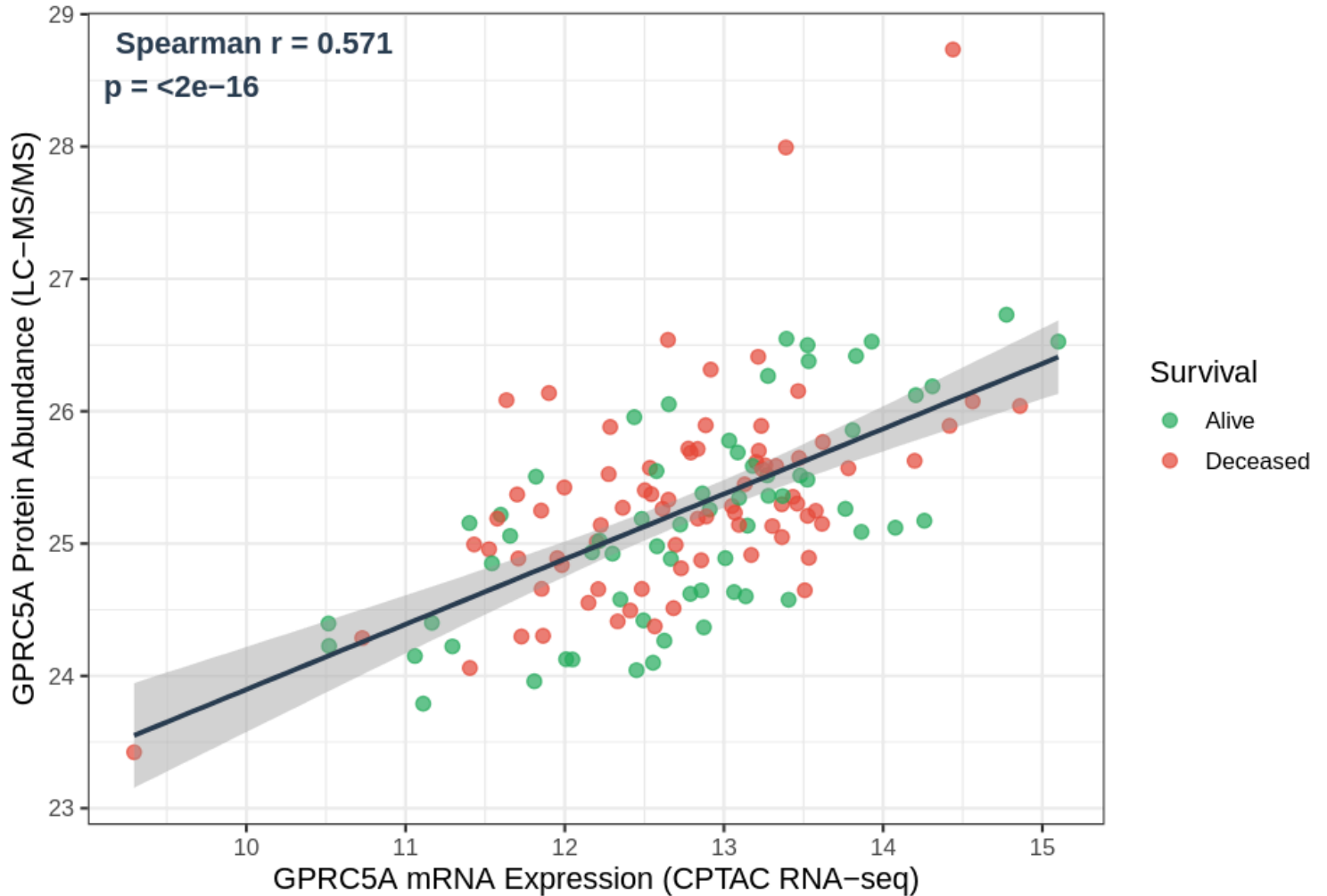


Figure 10

GPRC5A RNA-protein correlation in CPTAC-PAAD. Scatter plot of VST-normalized RNA-seq expression versus log2-transformed LC-MS/MS protein abundance across 140 matched samples, colored by vital status. Spearman $r=0.571$, $p<2\times 10^{-16}$. The dashed line represents the linear fit.

Genome-wide RNA-Protein Correlation – CPTAC-PAAD

GPRC5A highlighted: where does it fall in the global landscape?

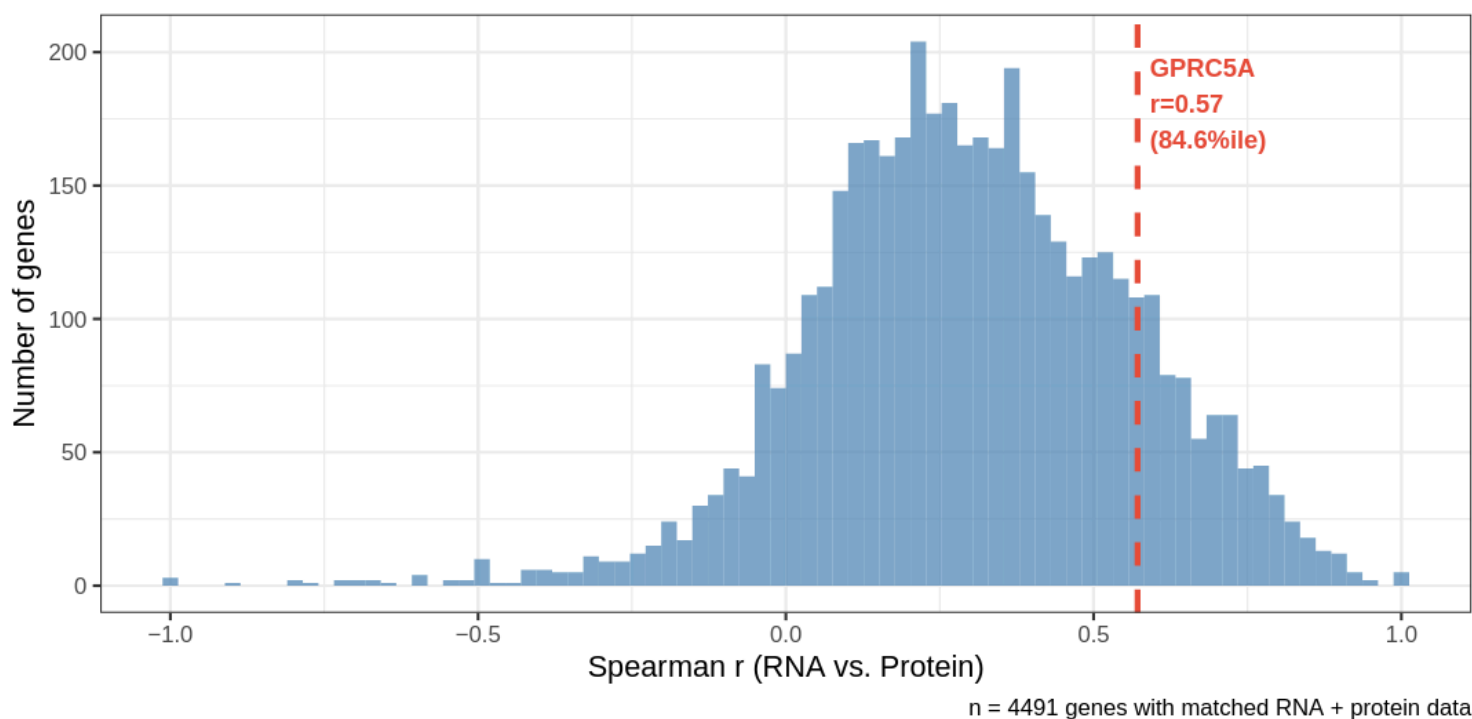


Figure 11

Genome-wide distribution of Spearman RNA-protein correlations in CPTAC-PAAD. Density plot of Spearman r values for all 4,491 genes with matched RNA and protein data. GPRC5A ($r=0.571$) is highlighted, corresponding to the 84.6th percentile of the genome-wide distribution, indicating above-average RNA-protein concordance.

GPRC5A AlphaFold2 pLDDT Confidence Scores

Higher pLDDT = more confident structural prediction | TM helices expected highest

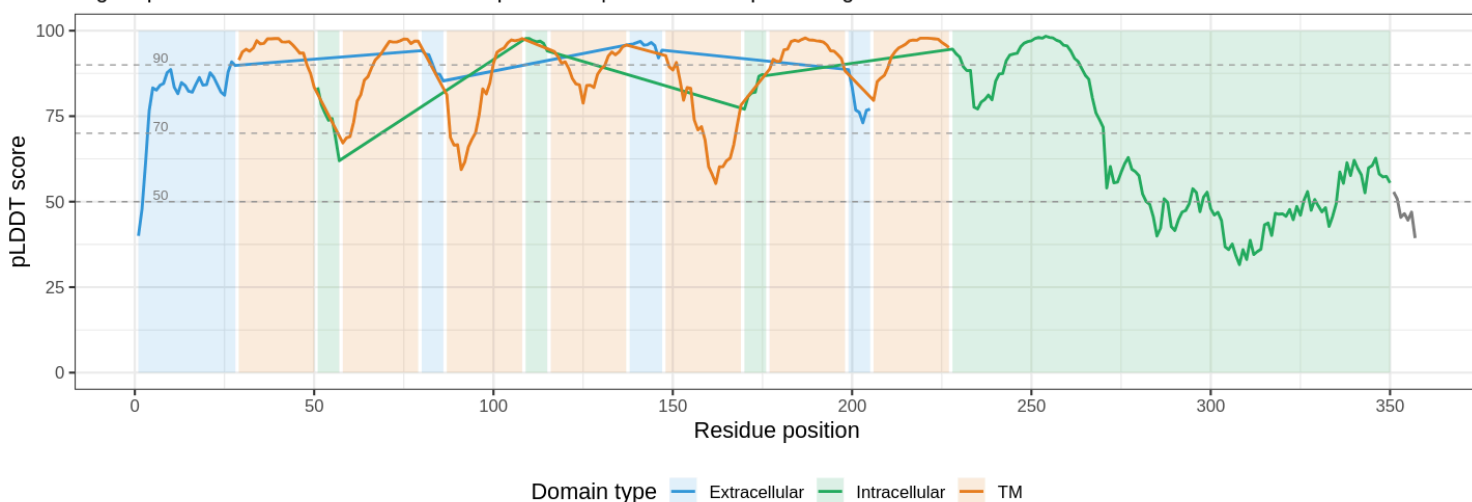


Figure 12

AlphaFold2 per-residue pLDDT confidence scores for human GPRC5A (UniProt Q8NFJ5, 350 aa). Residues are color-coded by structural domain: transmembrane helices (TM1–TM7), extracellular loops (ECL), intracellular loops (ICL), N-terminus, and C-terminus. Dashed horizontal lines indicate pLDDT thresholds of 50, 70, and 90. High confidence (pLDDT >70) across TM helices validates domain-level annotation.

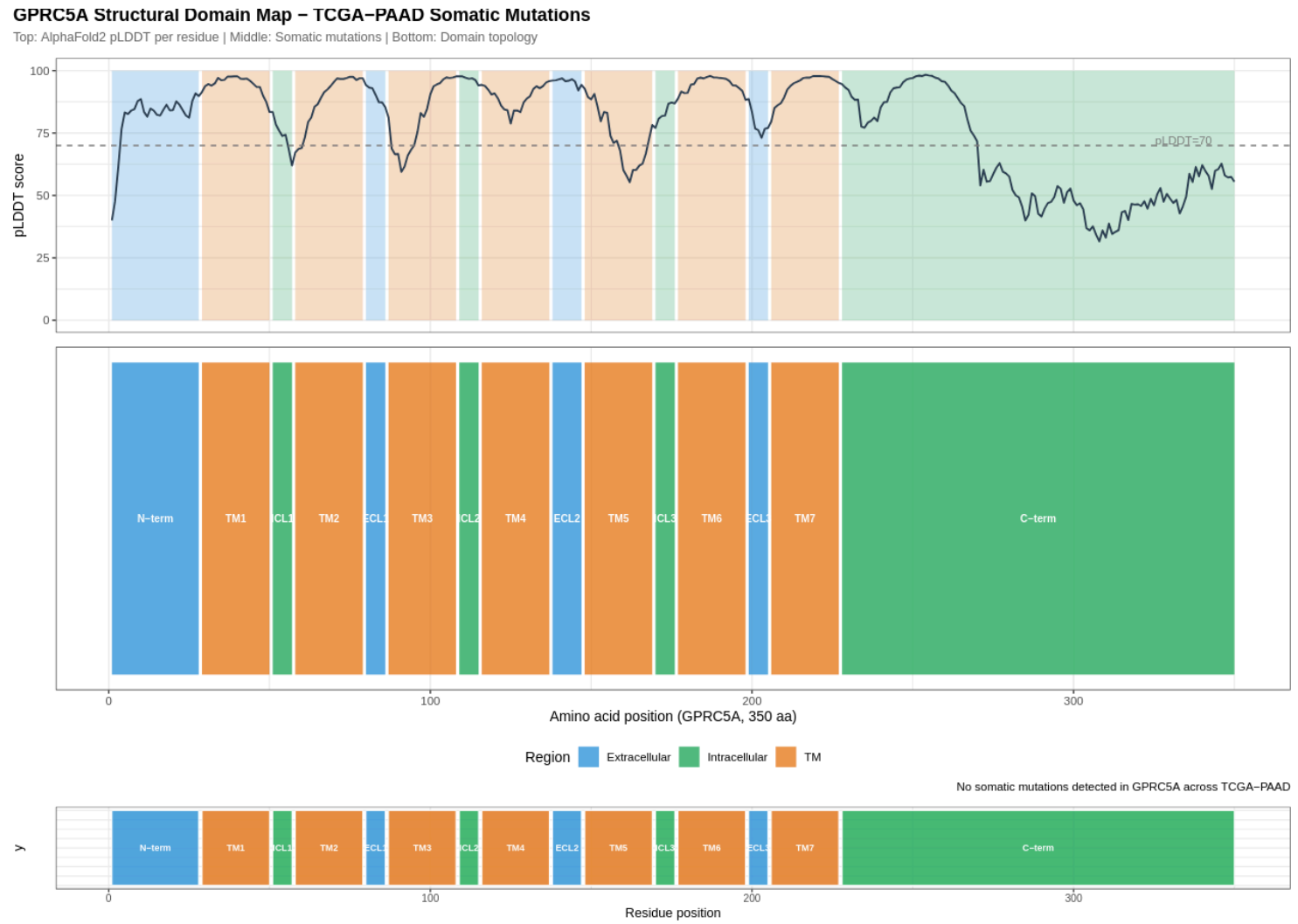


Figure 13

GPRC5A domain map with somatic mutation track from TCGA-PAAD. Upper panel: AlphaFold2 pLDDT per-residue confidence. Lower panel: domain topology with somatic mutation lollipops. Zero somatic mutations were detected across all 177 TCGA-PAAD samples; the empty mutation track is itself a biologically informative finding.

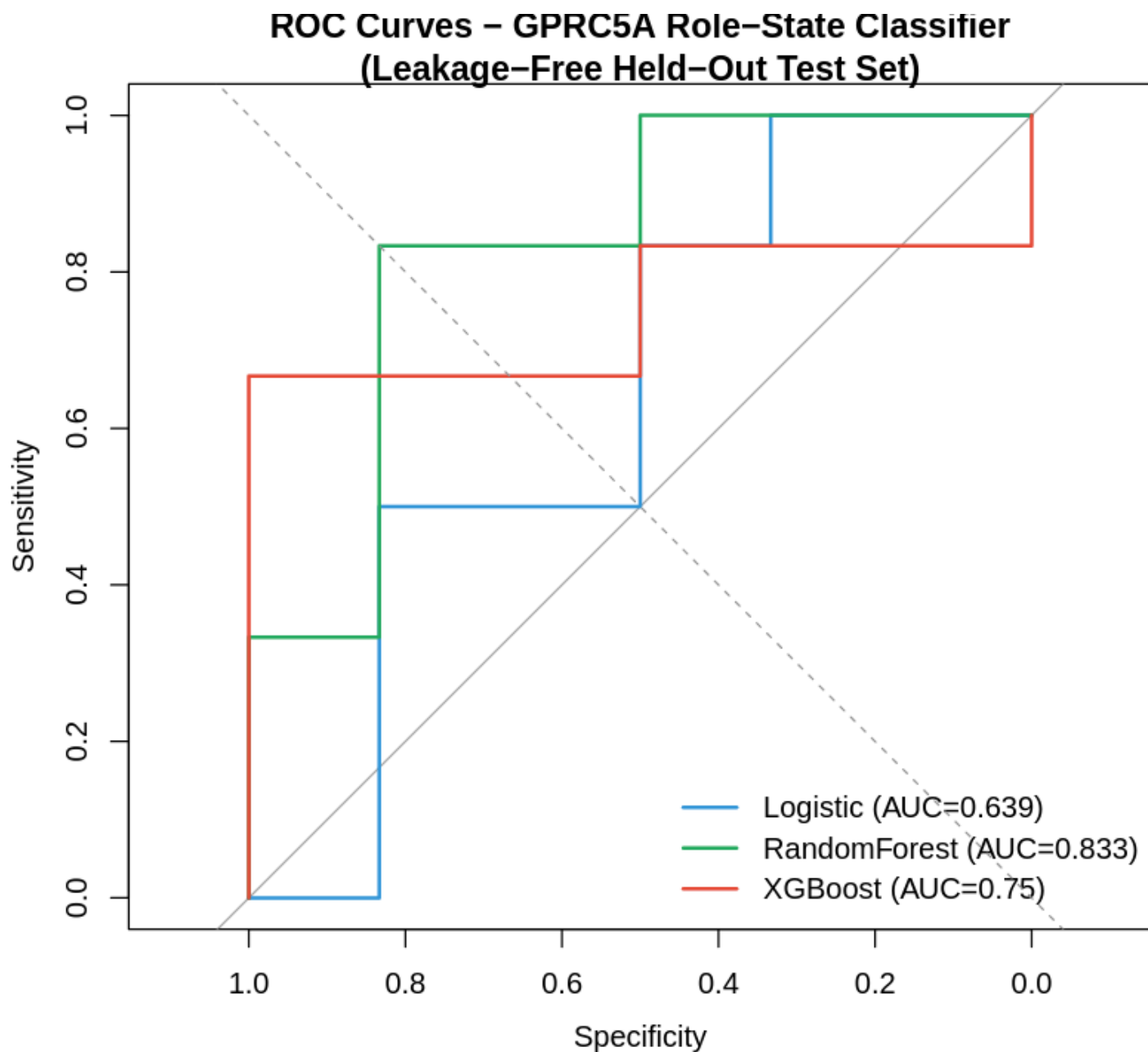


Figure 14

ROC curves for GPRC5A functional role-state classifiers on the held-out test set ($n=12$). Three classifiers are shown: Random Forest (AUC=0.833), XGBoost (AUC=0.750), and logistic regression (AUC=0.639). The diagonal dashed line represents a random classifier (AUC=0.5). All preprocessing and feature selection was performed within the training fold only.

Feature Importance – RandomForest Classifier

GPRC5A Role-State Prediction | TCGA-PAAD
(features selected on training data only)

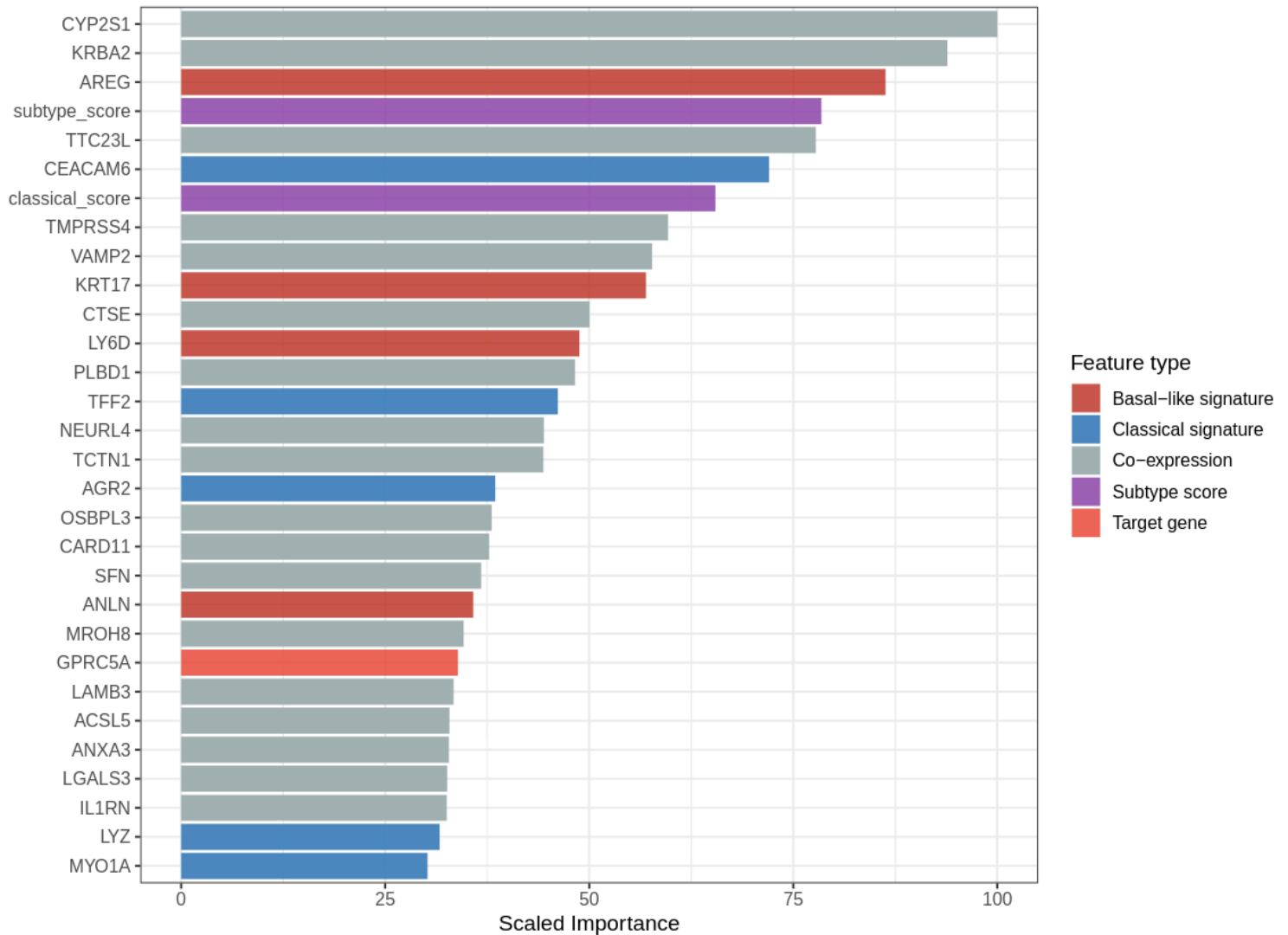


Figure 15

Random Forest feature importance for GPRC5A role-state classification. Features are ranked by scaled mean decrease in Gini impurity. Color coding indicates feature type: classical signature co-expression genes, basal-like signature genes, composite subtype score, and GPRC5A expression itself. GPRC5A expression ranks 23rd, indicating that transcriptional context dominates over GPRC5A expression level in determining functional role state.

Calibration Plot – GPRC5A Role–State Classifier

Diagonal = perfect calibration | Held-out test set

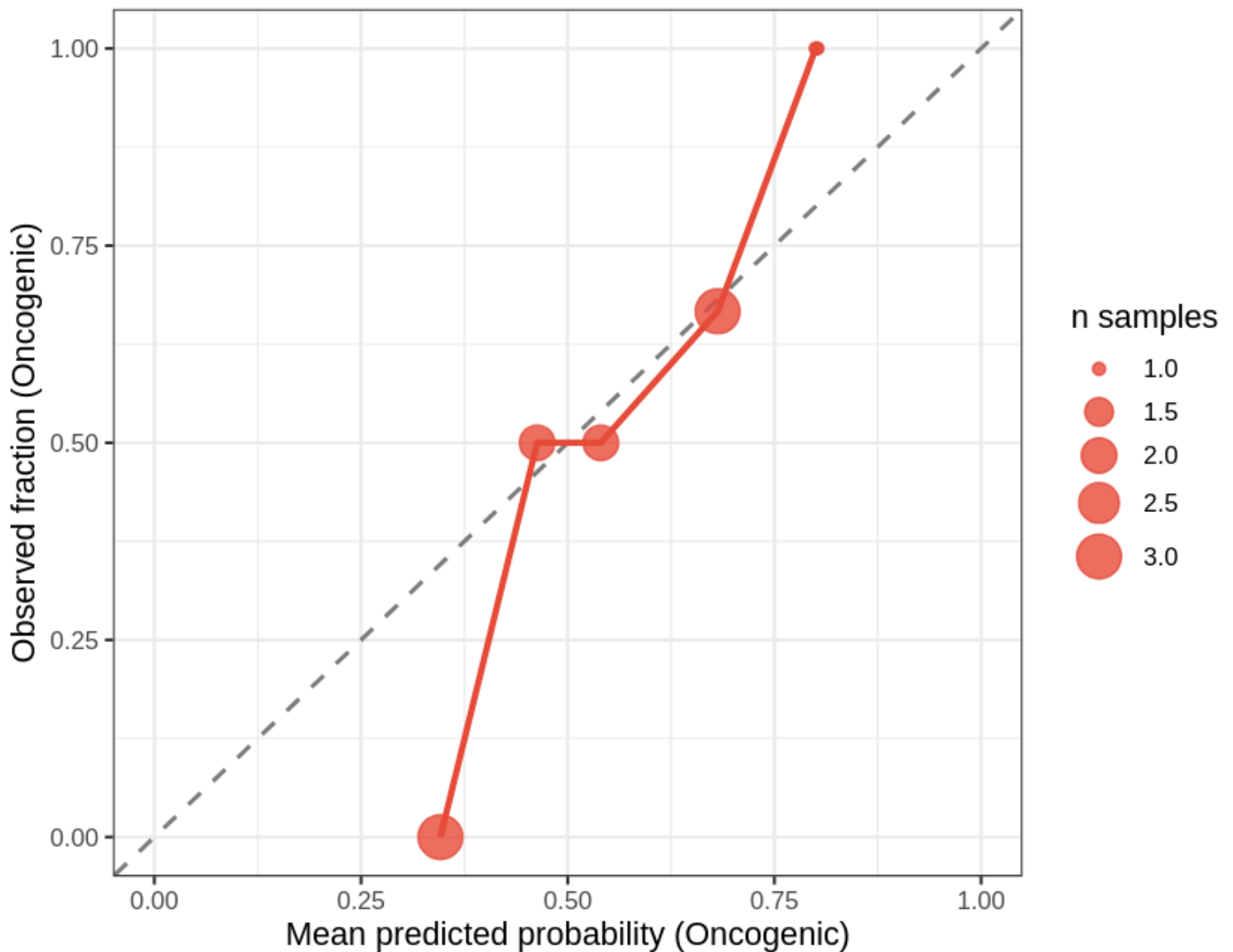


Figure 16

Calibration plot for the Random Forest role-state classifier on the held-out test set (n=12). Observed oncogenic fraction is plotted against mean predicted oncogenic probability across probability bins. The diagonal dashed line represents perfect calibration. No severe systematic over- or under-confidence is observed.

Overall Survival by Predicted GPRC5A Role State

Held-out test set | Oncogenic should show worse OS

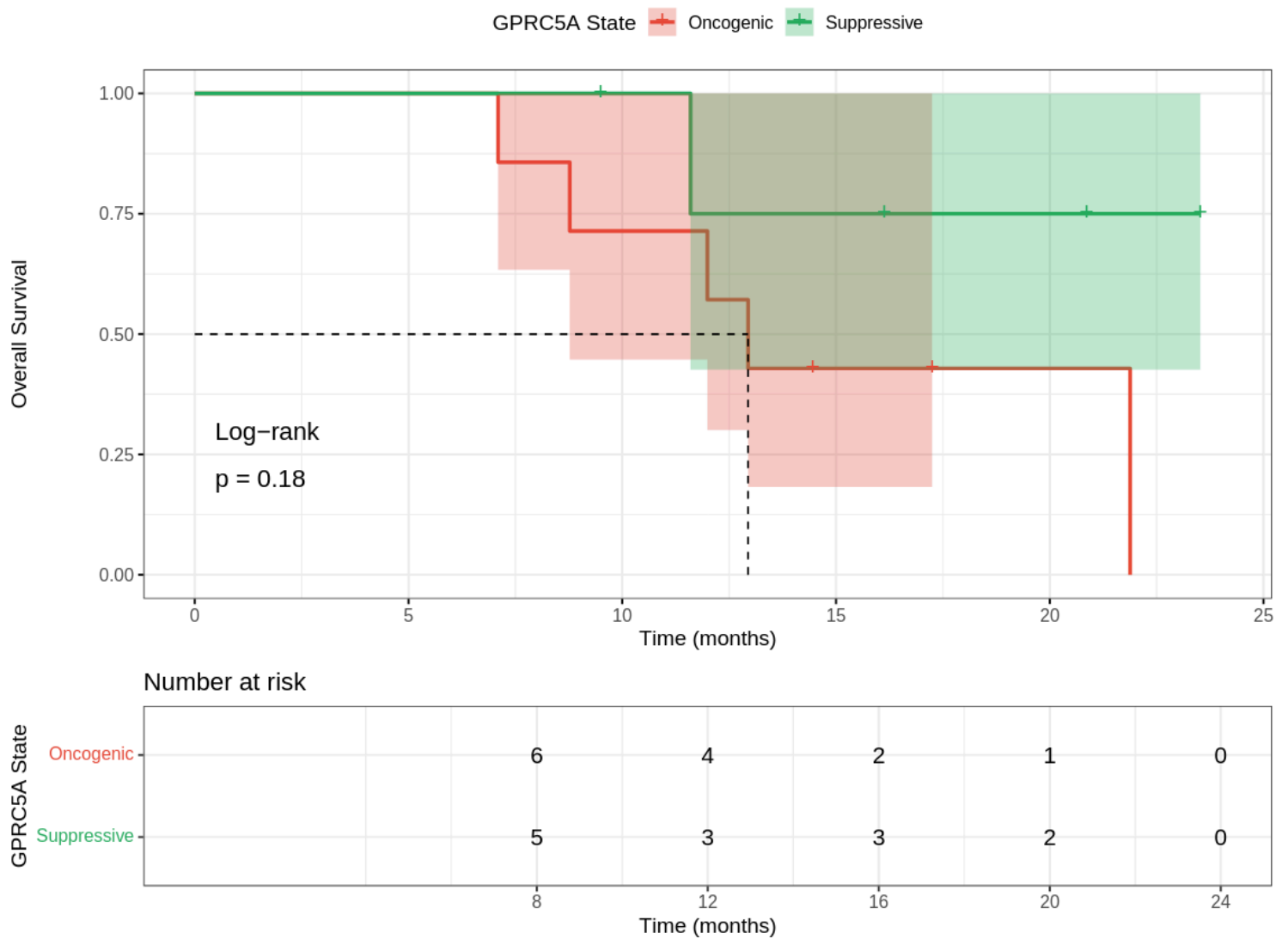


Figure 17

Kaplan–Meier overall survival curves by predicted GPRC5A functional role state in the held-out test set ($n=12$; oncogenic $n=7$, suppressive $n=5$). Role state was assigned by the Random Forest classifier. Log-rank $p=0.18$; the non-significant result reflects the limited statistical power of the small test set rather than classifier failure. The directional trend is consistent with the classifier’s intended biological meaning.