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.

A three-gene co-expression signature for distinguishing head and neck squamous cell carcinoma from non-tumour mucosa: an exploratory diagnostic and prognostic analysis

Peng Chen
First Affiliated Hospital of Gannan Medical University
Xianbin Lan
First Affiliated Hospital of Gannan Medical University
Yingbin Wang
First Affiliated Hospital of Gannan Medical University
Taihai Deng
First Affiliated Hospital of Gannan Medical University
Xuemei Liu
First Affiliated Hospital of Gannan Medical University
Jing Zhou
First Affiliated Hospital of Gannan Medical University
Qiaozhen Fan
First Affiliated Hospital of Gannan Medical University
Feier Guo
First Affiliated Hospital of Gannan Medical University
Gaofeng Zhou
First Affiliated Hospital of Gannan Medical University
Ying Zhou
First Affiliated Hospital of Gannan Medical University
Quanpeng Qiu
First Affiliated Hospital of Gannan Medical University
Riqun Jin
xiaojin99006@163.com

First Affiliated Hospital of Gannan Medical University

Article

Keywords: Head and neck squamous cell carcinoma, Weighted gene co-expression network analysis, Diagnostic model

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11316142/v1

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

Additional Declarations: No competing interests reported.

Abstract

Head and neck squamous cell carcinoma (HNSCC) is usually diagnosed at an advanced stage, and markers for early detection remain limited. We applied weighted gene co-expression network analysis to a discovery cohort of 53 samples and reduced the module most strongly associated with tumour status to a three-gene panel (APOD, CXCL12 and GGTA1). The panel was evaluated by cross-validated discrimination, bootstrap-corrected calibration and recalibrated decision curve analysis, and its expression direction was examined in independent cohorts. Overall survival, KEGG pathway activity, immune composition, mutations, stemness and predicted drug sensitivity were explored as secondary analyses in the discovery cohort, GSE41613 or TCGA-HNSC. The midnightblue module was most negatively correlated with tumour status; all three genes were down-regulated in tumours, and the combined score reached an internal area under the curve of 0.939 with well-calibrated probabilities and net benefit over treat-all. Median-split groups differed in overall survival in GSE41613, the panel reached 0.935 after refitting in TCGA-HNSC, and all three genes were down-regulated in three further cohorts. Tumours showed up-regulated proliferation and down-regulated ribosome, oxidative phosphorylation, drug-metabolism and complement-coagulation programmes; three screened KEGG pathways correlated inversely with the score, and high-risk tumours had lower CD8-positive T cell, M2 macrophage and memory B cell fractions, a higher mutational burden and stemness index, and a distinct predicted drug-sensitivity profile. The signature distinguishes HNSCC from non-tumour tissue and marks loss of a normal-mucosa stromal and lipid-metabolic programme. Its prognostic value was observed in a single cohort, and the panel should therefore be regarded as an exploratory candidate requiring prospective validation.

Introduction

Head and neck squamous cell carcinoma (HNSCC) arises from the mucosal epithelium of the oral cavity, pharynx and larynx; tobacco and alcohol dominate the aetiology of oral and laryngeal tumours, whereas oropharyngeal tumours are increasingly attributable to human papillomavirus (HPV) infection (1). The disease remains among the most common malignancies worldwide, and a substantial part of the global burden of oral cancer is attributable to smokeless tobacco and areca nut consumption (2, 3). HPV-positive and HPV-negative tumours differ in biology, staging, prognosis and treatment selection (4), and HPV-specific stem-like CD8-positive T cells have been identified within HPV-positive tumours (5). Site, stage and HPV status nevertheless explain only part of the heterogeneity in outcome, which has motivated the search for markers that refine risk assessment.

Molecular profiling has clarified this heterogeneity. Proteogenomic and integrative multi-omics analyses have defined HNSCC subtypes with distinct prognoses and therapeutic vulnerabilities, and targeted therapy is being organised around the corresponding pathway dependencies (6–8). Translation into simple, deployable markers has been slow, because single-gene markers show modest effects and large signatures are difficult to implement; compact panels of biologically coherent genes remain attractive, provided that performance is reported with calibration and net benefit rather than discrimination alone

(9). Network approaches that group genes by expression similarity and relate modules to clinical traits are well suited to this problem.

The tumour microenvironment shapes invasion, immune evasion and treatment response in HNSCC (10, 11). Single-cell studies have resolved malignant cell states and immune and non-immune cell interactions (12), stromal subsets that restrict CD8-positive T cell activity have been identified, and immunostimulatory fibroblast subpopulations predict response to immunotherapy (13, 14). B cells and tertiary lymphoid structures are favourable prognostic features, lymphocyte dynamics determine response to checkpoint blockade, and checkpoint inhibitors benefit only a minority of patients (15–18). In this study we used co-expression network analysis to identify the module most closely associated with tumour status and reduced it to a three-gene panel, we characterised its diagnostic performance with explicit calibration and net-benefit analysis, examined the KEGG pathway landscape in relation to the score, and explored the immune, mutational, stemness and pharmacogenomic correlates of the score as secondary analyses. We report both supportive and negative findings because they jointly define the current evidential status of the panel. The panel is not proposed as a clinically ready diagnostic test; rather, we present an exploratory signature whose diagnostic and prognostic performance requires prospective validation prospective validation.

Materials and methods

Human tissue samples

All clinical samples were collected from the First Affiliated Hospital of Gannan Medical University (Ganzhou, Jiangxi, China), In total, 53 specimens were analysed: 44 primary tumours and 9 non-tumour mucosal tissues. The study protocol, including patient recruitment, specimen collection and tissue processing, was approved by the Clinical Research Ethics Committee of the First Affiliated Hospital of Gannan Medical University (approval number: LLSC-2022110901), and all procedures were performed in accordance with the Declaration of Helsinki. Written informed consent for the use of surgical tissue in research was obtained from all participants (or their legally authorised representatives).

RNA extraction and RNA sequencing

Total RNA was extracted from the frozen tissue specimens using TRIzol reagent according to a standard protocol. Sequencing libraries were prepared and sequenced on an Illumina platform in paired-end mode. RNA-seq reads were filtered to remove adaptors and low-quality bases, aligned to the human reference genome GRCh38, and gene-level abundances were quantified. Expression values were expressed as fragments per kilobase of transcript per million mapped reads (FPKM) and log2(FPKM + 1)-transformed.

Discovery cohort

The discovery cohort comprised 53 head and neck squamous cell carcinoma samples (9 non-tumour mucosa, 44 tumour) profiled by RNA sequencing. Expression was expressed as fragments per kilobase of transcript per million mapped reads (FPKM) and transformed to log2(FPKM + 1). Genes were ranked by variance and the 5,000 most variable genes were retained for network construction; sample-level clustering used Euclidean distance and average linkage.

Co-expression network analysis and module enrichment

Co-expression analysis was performed with the WGCNA R package. An unsigned network was chosen because both positively and negatively co-expressed genes are biologically informative. The soft-thresholding power was the smallest power at which the scale-free topology fit reached 0.80, which selected power 22 (signed R-squared = 0.809). The adjacency matrix was converted into topological overlap dissimilarity, modules were detected by dynamic tree cutting, and modules with eigengene similarity above 0.80 were merged. Module-trait relationships were estimated as Pearson correlations between module eigengenes and histological status, and the module with the strongest negative correlation (MEmidnightblue) was carried forward. Its genes were tested for Gene Ontology and KEGG enrichment with clusterProfiler, with Benjamini-Hochberg adjustment.

Signature construction and nomogram

Module genes were used as features in logistic regression penalised by the least absolute shrinkage and selection operator (LASSO) with glmnet; ten-fold cross-validation with the binomial deviance criterion was used to select lambda.min and lambda.1se, and the one-standard-error rule was preferred. The linear predictor defined the risk score: riskScore = a0 + b1 x log2(FPKM APOD + 1) + b2 x log2(FPKM CXCL12 + 1) + b3 x log2(FPKM GGTA1 + 1), with a0 = 2.061747, b1 = -0.0256438, b2 = -0.0375291 and b3 = -0.1688718. The predicted probability of tumour tissue was plogis(riskScore), which reproduced the fitted probabilities to machine precision. Because non-tumour samples were few, complete separation was possible, and Firth-penalised logistic regression (logistf) was used for all probability-based analyses and for the nomogram (rms).

Diagnostic performance, calibration and decision analysis

Discrimination was quantified by the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals (DeLong method), and the operating threshold was chosen by the Youden index. Optimism was estimated by ten-fold cross-validation with refitting in each training partition and out-of-fold probabilities used for the cross-validated AUC, threshold and classification counts. Calibration was assessed with the Hosmer-Lemeshow statistic and with calibration curves; because apparent calibration is optimistic, 1,000 bootstrap resamples were used and the mean optimism was subtracted from the apparent curve to obtain a bias-corrected curve. Clinical usefulness was assessed by decision curve analysis comparing the model with treat-all and treat-none strategies. Because the discovery cohort contains more tumour than non-tumour samples (44 versus 9), decision curves were additionally computed after intercept-only recalibration to target prevalences of 0.20, 0.35 and 0.50 by

shifting the logit of the predicted probability by delta = logit(prevalence) - mean(logit(p)) and weighting tumour samples by w = (n_tumour / n_normal) x (1 - prevalence) / prevalence; the resulting shifts were -3.042, -2.261 and -1.626. Clinical impact curves showed, at each threshold, how many patients were classified as high risk and how many of these had tumour tissue.

Survival analysis

Overall survival was analysed in GSE41613. Patients were dichotomised at the median risk score, groups were compared by the Kaplan-Meier method with the log-rank test, and hazard ratios with 95% confidence intervals were obtained from a Cox proportional-hazards model.

Pathway-level analyses

Genes were ranked by the limma moderated t statistic (tumour versus non-tumour), and genome-wide gene set enrichment analysis (GSEA) was performed with fgsea against KEGG gene sets obtained from the KEGG database(19, 20); positive normalised enrichment scores denote enrichment in tumour tissue. Per-sample pathway activity was estimated by gene set variation analysis (GSVA) using the same KEGG pathway collection and compared between groups with a two-sided t-test. KEGG pathways were screened in four steps: significant in GSEA after Benjamini-Hochberg adjustment (adjusted $p<0.05$); significant in the GSVA comparison ($p<0.05$); exclusion of disease-name, organ-specific and addiction-related gene sets; and correlation with the risk score or with one of the three signature genes ($|\text{Pearson } R| >= 0.5$). Pathways passing all four steps were ranked by thematic coherence with the key module, and the three most coherent were related to the risk score and to the three genes by linear regression. A targeted GSEA of Gene Ontology biological process, cellular component and molecular function gene sets whose names contained extracellular matrix, collagen, fibronectin, integrin, laminin, basement membrane, elastic, matrix, ECM, cell adhesion, cell-matrix, structural constituent or proteoglycan terms was performed on the same ranked list.

Immune, mutational, stemness and drug-sensitivity analyses

Immune cell fractions were estimated in the 44 discovery tumours with CIBERSORT in SVR mode using the LM22 matrix; all 22 cell types were used for group comparisons, and 21 were displayed because gamma-delta T cells were estimated as zero in most samples, with Wilcoxon rank-sum tests and Benjamini-Hochberg correction. A complementary 11-cell analysis (Monaco/ABIS immune cell types) was used to relate gene expression to immune composition. Somatic mutation data for TCGA-HNSC were processed with maftools; oncoprints were generated for each risk group, and tumour mutational burden was calculated as mutations per 30 Mb. The mRNA expression-based stemness index was obtained from the published resource and related to the risk score by the Pearson method. Drug sensitivity was predicted with oncoPredict using the GDSC2 dataset as the training resource, and predicted IC50 values of 198 drugs were compared between risk groups by the Wilcoxon rank-sum test with Benjamini-Hochberg correction; cisplatin, 5-fluorouracil, docetaxel, irinotecan and gefitinib were

selected a priori for display. When more than one sample was available for the same patient, one sample per patient was retained for the mutation, mutational burden and stemness analyses.

Independent evaluation and statistical analysis

TCGA-HNSC was used to refit the three-gene model and to evaluate discrimination, threshold performance and the risk score distribution; because platforms and normalisation differed between cohorts, the model was refitted within each cohort and the results are reported as independent evaluation of the gene panel rather than as direct transfer of the discovery coefficients. Direction of expression change was assessed in GSE30784, GSE59102 and GSE127165 by the Wilcoxon rank-sum test with Benjamini-Hochberg adjustment. All analyses were performed in R 4.6.1; two-sided p values below 0.05 were considered significant, and multiplicity was controlled by the Benjamini-Hochberg procedure where indicated. Public data were obtained from TCGA and from the Gene Expression Omnibus.

Results

Co-expression network construction and identification of the tumour-associated module

The research flow chart depicts the study design **(Figure 1)**.Unsupervised clustering of the 53 discovery samples separated non-tumour mucosa from tumour tissue, with two non-tumour samples clustering among the tumours **(Figure 2A)**. The scale-free topology fit reached the target of 0.80 at soft-thresholding power 22 (signed R-squared 0.809), and higher powers increased the fit only marginally **(Figure 2B)**. Dynamic tree cutting and merging at 0.80 eigengene similarity left seven modules, the largest being the violet, grey and red modules **(Figure 2C)**. Module-trait correlation identified the midnightblue module (61 genes ; hereafter the key module) as the one most negatively correlated with tumour status ($R=-0.582$, $p<0.05$), whereas the lightcyan module was also negatively and the red module positively correlated **(Figure 2D)**. Genes of this module were enriched in extracellular matrix organisation, collagen metabolism, response to reactive oxygen species and leukotriene and icosanoid biosynthesis **(Figure 2E)**, and KEGG enrichment returned arachidonic acid metabolism (adjusted $p<0.05$) and ether lipid metabolism (adjusted $p<0.05$) **(Figure 2F)**.

Construction and internal discrimination of the three-gene score

Univariate logistic regression showed that the three genes subsequently retained were each associated with tumour status (APOD odds ratio OR 0.473; CXCL12 OR 0.381; GGTA1 OR 0.190; adjusted $p<0.05$) **(Figure 3A)**. LASSO coefficient paths shrank as the penalty increased **(Figure 3B)**, and the cross-validated binomial deviance curve was flat between lambda.min and lambda.1se, so the one-standard-error rule was adopted **(Figure 3C)**. In the multivariable Firth-penalised model none of the three genes remained individually significant **(Figure 3D)**. The nomogram converted the three genes into a predicted probability of tumour tissue, spanning approximately 0.70 to 0.86 over the observed range of total points **(Figure 3E)**. All three genes were expressed at lower levels in tumours (APOD median 102.9 versus 1.7

FPKM, p<0.05; CXCL12 23.3 versus 1.1, p<0.05; GGTA1 12.8 versus 1.2, p<0.05) **(Figure 3F)**, and the risk score was correspondingly higher in tumours **(Figure 3G)**.

Calibration, Decision-curve Analysis and Exploratory Survival Validation of the Three-Gene Panel

At the Youden threshold of 0.8245, all nine non-tumour samples were classified as low risk and 37 of the 44 tumours as high risk **(Figure 4A)**. Expression of the three genes decreased monotonically across samples ordered by increasing risk score **(Figure 4B)**. Cross-validated single-gene areas under the curve (AUCs) were 0.841 for APOD, 0.886 for CXCL12 and 0.919 for GGTA1 **(Figure 4C)**, and the combined score reached an apparent AUC of 0.939 (95% CI 0.876-1.000), with a sensitivity of 84.1% and a specificity of 100% at the Youden threshold **(Figure 4D)**. Firth-penalised probabilities were well calibrated (Hosmer-Lemeshow p□0.05; maximum absolute deviation from the ideal line 0.109 apparent and 0.156 after bootstrap correction) **(Figure 4E)**. However, calibration was estimated internally and requires external confirmation. Clinical impact curves showed that a threshold of 0.30 classified nearly all patients as high risk, whereas 0.50 flagged fewer than one fifth **(Figure 4F)**. Because the discovery cohort contains far more tumour than non-tumour samples, decision curves were recalculated after intercept-only recalibration to target prevalences of 0.20, 0.35 and 0.50; net benefit exceeded treat-all for threshold probabilities of approximately 0.09-0.26, 0.18-0.44 and 0.29-0.60, respectively **(Figure 4G)**. Because platforms and normalisation differed between cohorts, the three-gene model was refitted within TCGA-HNSC rather than transferred with the discovery coefficients; the refitted panel reached an AUC of 0.935 (95% CI 0.907–0.962), with single-gene AUCs of 0.858, 0.729 and 0.901 **(Figure S1A)**. the 441 tumours with survival annotation were divided into 220 high-risk and 221 low-risk cases at the median (cut-off 1.8101) **(Figure S1B)**. In TCGA-HNSC, all three genes were expressed at lower levels in tumours than in normal tissue **(Figure S1C)**. The direction of change was consistent in GSE30784, GSE59102 and GSE127165 **(Figure S1D)**. Collectively, these analyses validate the gene panel and the direction of expression change, rather than the exact discovery model.

Pathway Enrichment Analysis Reveals Metabolic and Immune-Related Down-Regulation in Tumours

Genome-wide enrichment analysis identified 49 of 186 KEGG pathways were significant after Benjamini-Hochberg adjustment. cell cycle, spliceosome, homologous recombination and DNA replication were enriched in tumours (normalised enrichment scores 2.13-2.69), whereas ribosome, oxidative phosphorylation, drug metabolism by cytochrome P450, complement and coagulation cascades and branched-chain amino acid degradation were enriched in non-tumour tissue (-1.64 to -2.68) **(Figure 5A)**. A targeted analysis of extracellular matrix-related Gene Ontology gene sets showed lower ECM structural, compression-resistance, elastic and integrin-binding functions in tumours and enrichment of nuclear and chromatin terms, whereas no ECM biological process term reached significance **(Figure 5B)**. Of the pathways passing all screening steps, the three most coherent with the key module were selected **(Figure 5C)**. Their per-sample activity was lower in tumours and correlated inversely with the risk score and positively with the three genes for complement and coagulation cascades **(Figure 5D)**, PPAR signalling **(Figure 5E)** and drug metabolism by cytochrome P450 **(Figure 5F)**.

Immune Microenvironment Profiles of High-Risk Tumours

Deconvolution of the 44 discovery tumours showed that several lymphocyte populations decreased as the risk score increased **(Figure 6A)**, and the stacked composition of the same samples reflected this shift **(Figure 6B)**. After Benjamini-Hochberg adjustment, M2 macrophages ($p<0.05$, adjusted $p<0.05$), CD8-positive T cells ($p<0.05$, adjusted $p<0.05$) and memory B cells ($p<0.05$, adjusted $p<0.05$) were lower in the high-risk group **(Figure 6C)**, and the continuous risk score correlated inversely with all three (M2 macrophages R=-0.725, memory B cells R= -0.574, CD8-positive T cells R=-0.450) **(Figure 6D)**. In the complementary analysis, higher expression of the three genes corresponded to a higher memory T cell fraction (GGTA1 R=0.340; CXCL12 R=0.278) and to lower plasmacytoid dendritic and naive T cell fractions **(Figure 6E)**. These analyses were performed in 44 tumours and are therefore exploratory.

Mutational burden, stemness and predicted drug sensitivity Profiles of High-Risk Tumours

Somatic mutation data were available for 429 of the 441 TCGA-HNSC patients; in the low-risk group TP53 was altered in 144 of 214 tumours (67.3%), followed by TTN, FAT1, MUC16 and CDKN2A **(Figure 7A)**, whereas in the high-risk group TP53 was altered in 155 of 215 tumours (72.1%), followed by TTN, FAT1, CDKN2A and CSMD3 **(Figure 7B)**. Tumour mutational burden was higher in the high-risk group (median 4.23 versus 3.67 mutations per 30 Mb, $p<0.05$) **(Figure 7C)**, and the risk score correlated positively with the mRNA stemness index ($n = 435$; R=0.403, $p<0.05$) **(Figure 7D)**. Predicted half-maximal inhibitory concentrations differed between risk groups for all five agents: 5-fluorouracil (median 110.09 versus 91.17, $p<0.05$), irinotecan (15.58 versus 11.15, $p<0.05$) and cisplatin (24.88 versus 19.45, $p<0.05$) were higher in the high-risk group **(Figure 7E-G)**, whereas gefitinib (19.94 versus 27.63, $p<0.05$) and docetaxel (0.0089 versus 0.0104, $p<0.05$) were lower **(Figure 7I, H)**; the corresponding log2 effect sizes were +0.27, +0.48, +0.36, -0.47 and -0.22 **(Figure 7J)**.

Discussion

We identified the co-expression module most closely associated with tumour status in a head and neck squamous cell carcinoma cohort and reduced it to three transcripts (APOD, CXCL12 and GGTA1) that separated tumour from non-tumour tissue with an internal AUC of 0.939, reproduced in TCGA-HNSC (0.935) and showed consistent down-regulation in three further cohorts. The score supported an interpretable nomogram with well-calibrated probabilities and a defined window of net benefit, and median-split risk groups differed in overall survival in GSE41613. High-risk tumours showed up-regulated proliferation, down-regulated stromal, lipid-metabolic and complement-coagulation programmes, fewer lymphocytes, a higher mutational burden and stemness index, and a distinct predicted drug-sensitivity profile. The panel is an exploratory candidate: its diagnostic performance was estimated internally, its expression direction was reproducible, and its prognostic signal was restricted to a single cohort.

The key module was enriched in extracellular matrix organisation, collagen metabolism and lipid mediator synthesis and contained canonical stromal transcripts such as DCN, LUM, DPT, MFAP4, FBLN5, MMP2 and VIM; stromal programmes of this kind are recurrently identified in expression studies of head

and neck cancer and of other tumours (7, 11). That these transcripts were lower in tumours than in non-tumour mucosa is, however, opposite to the usual interpretation of stroma-rich signatures as markers of aggressive disease. The most plausible explanation is compositional: normal mucosa contains abundant fibroblast-rich lamina propria, whereas tumour samples may be dominated by malignant epithelium, so a bulk comparison can dilute stromal transcripts even when the stroma is activated. The panel is therefore best read as a marker of retained normal-mucosa stroma rather than of stromal activation.

The three genes showed partial agreement with published data. Elevated tumour APOD has been associated with poor outcome in breast cancer and with perineural invasion in salivary adenoid cystic carcinoma (21-23); because APOD has not been characterised in HNSCC, this discrepancy cannot yet be resolved and may reflect tumour-type-specific biology. CXCL12 is produced by cancer-associated fibroblasts in head and neck tumours, where CXCL12-expressing subsets restrict CD8-positive T cell infiltration (13, 24), and the CXCL12-CXCR4/CXCR7 axis is a recognised mechanism of immune resistance (25); our finding that CXCL12 is lower in tumours is therefore consistent with the loss of a fibroblast-derived chemokine programme.

For GGTA1 the comparison is limited, because the cancer literature on this glycosyltransferase is essentially restricted to xenotransplantation and to strategies that exploit the Gal-alpha1-3Gal epitope for antitumour immunity (26), although glycosylation is increasingly recognised as a determinant of tumour immunity (27-29). GGTA1 carried the largest coefficient and the highest single-gene cross-validated AUC, but this observation is hypothesis-generating.

The pathway results agree with the wider literature (7): proliferation programmes are up-regulated in HNSCC, the association of the score with PPAR-dependent lipid metabolism is consistent with reports that lipid reprogramming supports therapy resistance (30), and the positive correlation with the stemness index agrees with the link between dedifferentiation and aggressive tumour programmes (31, 32). One result is discordant within our own data: although the targeted analysis showed lower extracellular matrix structural and mechanical functions in tumours, the canonical KEGG pathways ECM-receptor interaction and focal adhesion were not significant. Large gene sets of this kind are dominated by intracellular signalling and cytoskeletal components that are up-regulated in proliferating tumour cells, so the discrepancy more likely reflects gene-set composition than a biological contradiction.

Taken together, the data describe two coordinated axes: a proliferative axis (cell cycle, DNA replication, mismatch repair and spliceosome programmes up-regulated, a higher stemness index and enrichment of chromatin and nuclear matrix functions) and the loss of a stromal and metabolic axis (lower extracellular matrix structural, elastic and integrin-binding functions and lower complement-coagulation and lipid-metabolic activity, decreasing as the score increases). The key module also contained leukotriene and icosanoid biosynthetic genes, a pathway family implicated in head and neck tumorigenesis (33). The immune correlates place these axes in a microenvironmental context, because high-risk tumours contained fewer CD8-positive T cells, M2 macrophages and memory B cells; B cells

and tertiary lymphoid structures are favourable prognostic features (15), lymphocyte infiltration determines response to checkpoint blockade (16-18), and recurrent tumours frequently adopt an immunosuppressive microenvironment (34). A plausible but unproven link is that the stromal loss captured by the panel may remove chemokine and matrix support required for lymphocyte organisation, leaving a proliferative, immune-cold phenotype with a higher mutational burden and greater predicted resistance to platinum and antimetabolite chemotherapy. Our data are correlative and cannot establish this causal chain.

The main strengths are the explicit reporting of calibration and net benefit rather than discrimination alone; the recalculation of decision curves under prevalences relevant to a screening or triage setting rather than the 83% prevalence of the discovery cohort; the use of a Firth-penalised model to handle the complete separation expected with nine non-tumour samples; and the reporting of both supportive and negative results. We also explicitly distinguish internal performance from independent expression-direction validation and report negative findings.

The discovery cohort is small and contains only nine non-tumour samples, so the apparent AUC and calibration are likely optimistic. The study is retrospective and based on bulk tissue, and external evaluation used refitting within each cohort, so the gene panel rather than the exact coefficients has been validated. Normal controls in GSE59102 were surgical safety margins, and GGTA1 was measured through a pseudogene probe on one platform and was not measurable on another. The survival analysis rests on a single cohort of 97 patients with a wide confidence interval. Immune deconvolution was performed in 44 tumours, which limits precision and precludes adjustment for clinical covariates. The pathway analyses are exploratory: the gene set variation comparisons were not corrected for multiple testing, the KEGG screening was guided by coherence with the key module, and only molecular function terms reached significance in the targeted analysis. The stemness analysis is correlative and, because the index was derived from TCGA data, is not independent of the cohort in which it was evaluated. HPV status, smoking, subsite and treatment were not modelled, and the panel has not been tested in vitro or in vivo. Prospective validation with prespecified thresholds, functional characterisation in cell and animal models, and direct coefficient transfer into a matched independent cohort are therefore required before any clinical claim can be made.

The panel is not yet ready for clinical use. Prospective evaluation in patients with oral potentially malignant lesions or suspicious mucosa (35), external validation with the original coefficients, and functional work to establish whether APOD, CXCL12 and GGTA1 are markers of retained normal mucosa or contribute to the phenotype are the most valuable next steps. Because the score was associated with an immune-cold microenvironment, a higher mutational burden and a distinct predicted drug-response profile, it also provides a testable hypothesis for patient selection in immunotherapy and platinum-based chemotherapy studies (36, 37).

Declarations

Author Contributions: R.J. and Q.Q. conceptualized and designed the study. Y.W. supervised and administrated the study. P.C. performed most of the experiments with F.G.'s, J.Z.'s, Q.F.'s, X.L.'s, G.Z.'s, T.D.'s,and Y.Z.'s help. P.C. and Q.Q. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding Statement: This work was supported by grants from Science and Technology Research Project of the Jiangxi Provincial Department of Education (GJJ2201407) and Jiangxi Province Natural Science Foundation (20262BAC200103).

Ethics approval and consent to participate: The study protocol was approved by the Clinical Research Ethics Committee of the First Affiliated Hospital of Gannan Medical University (approval number: LLSC-2022110901) and was conducted in accordance with the Declaration of Helsinki. Written informed consent for the use of surgical tissue in research was obtained from all participants.

Consent for publication: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The RNA-sequencing data generated in this study have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE353246(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE353246). The record is private until 10 November 2026, when it will become publicly available; reviewer access to the private record is provided to the editorial office. TCGA-HNSC data were obtained from the National Cancer Institute Genomic Data Commons. The public datasets reanalysed in this study are available in GEO under accession numbers GSE41613, GSE30784, GSE59102 and GSE127165. The processed expression matrix, sample annotation and analysis code are provided as Supplementary Data. All other data supporting the findings of this study are available within the article and its Supplementary Information.

Conflicts of Interest: The authors declared no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

References

1. Johnson, D. E. et al. Head and neck squamous cell carcinoma. *Nat. Rev. Dis. Primers*. **6** (1), 92 (2020).
2. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74** (3), 229–263 (2024).
3. Rumgay, H. et al. Global burden of oral cancer in 2022 attributable to smokeless tobacco and areca nut consumption: a population attributable fraction analysis. *Lancet Oncol.* **25** (11), 1413–1423 (2024).
4. Lechner, M., Liu, J., Masterson, L. & Fenton, T. R. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. *Nat. Rev. Clin. Oncol.* **19** (5), 306–327 (2022).

5. Eberhardt, C. S. et al. Functional HPV-specific PD-1(+) stem-like CD8 T cells in head and neck cancer. *Nature* **597** (7875), 279–284 (2021).
6. Huang, C. et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell.* **39** (3), 361–79e16 (2021).
7. Diao, P. et al. Integrative Multiomics Analyses Identify Molecular Subtypes of Head and Neck Squamous Cell Carcinoma with Distinct Therapeutic Vulnerabilities. *Cancer Res.* **84** (18), 3101–3117 (2024).
8. Li, Q., Tie, Y., Alu, A., Ma, X. & Shi, H. Targeted therapy for head and neck cancer: signaling pathways and clinical studies. *Signal. Transduct. Target. Ther.* **8** (1), 31 (2023).
9. Chen, Y., Li, Z. Y., Zhou, G. Q. & Sun, Y. An Immune-Related Gene Prognostic Index for Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.* **27** (1), 330–341 (2021).
10. Ruffin, A. T. et al. Improving head and neck cancer therapies by immunomodulation of the tumour microenvironment. *Nat. Rev. Cancer*. **23** (3), 173–188 (2023).
11. Gong, Y. et al. The tumor ecosystem in head and neck squamous cell carcinoma and advances in ecotherapy. *Mol. Cancer*. **22** (1), 68 (2023).
12. Kürten, C. H. L. et al. Investigating immune and non-immune cell interactions in head and neck tumors by single-cell RNA sequencing. *Nat. Commun.* **12** (1), 7338 (2021).
13. Li, C. et al. Spatial and Single-Cell Transcriptomics Reveal a Cancer-Associated Fibroblast Subset in HNSCC That Restricts Infiltration and Antitumor Activity of CD8 + T Cells. *Cancer Res.* **84** (2), 258–275 (2024).
14. Obradovic, A. et al. Immunostimulatory Cancer-Associated Fibroblast Subpopulations Can Predict Immunotherapy Response in Head and Neck Cancer. *Clin. Cancer Res.* **28** (10), 2094–2109 (2022).
15. Ruffin, A. T. et al. B cell signatures and tertiary lymphoid structures contribute to outcome in head and neck squamous cell carcinoma. *Nat. Commun.* **12** (1), 3349 (2021).
16. Harrington, K. J. et al. Pembrolizumab With or Without Chemotherapy in Recurrent or Metastatic Head and Neck Squamous Cell Carcinoma: Updated Results of the Phase III KEYNOTE-048 Study. *J. Clin. Oncol.* **41** (4), 790–802 (2023).
17. Machiels, J. P. et al. Pembrolizumab plus concurrent chemoradiotherapy versus placebo plus concurrent chemoradiotherapy in patients with locally advanced squamous cell carcinoma of the head and neck (KEYNOTE-412): a randomised, double-blind, phase 3 trial. *Lancet Oncol.* **25** (5), 572–587 (2024).
18. Li, H. et al. Distinct CD8(+) T cell dynamics associate with response to neoadjuvant cancer immunotherapies. *Cancer Cell.* **43** (4), 757–75e8 (2025).
19. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53** (D1), D672–d7 (2025).
20. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** (1), 27–30 (2000).

21. Fyfe-Desmarais, G., Desmarais, F., Rassart, É. & Mounier, C. Apolipoprotein D in Oxidative Stress and Inflammation. *Antioxid. (Basel)* ;**12**(5). (2023).
22. Jankovic-Karasoulos, T. et al. Elevated levels of tumour apolipoprotein D independently predict poor outcome in breast cancer patients. *Histopathology* **76** (7), 976–987 (2020).
23. Huang, G. et al. Apolipoprotein D is crucial for promoting perineural invasion in salivary adenoid cystic carcinoma. *Br. J. Cancer*. **132** (7), 599–610 (2025).
24. Galbo, P. M. Jr., Zang, X. & Zheng, D. Molecular Features of Cancer-associated Fibroblast Subtypes and their Implication on Cancer Pathogenesis, Prognosis, and Immunotherapy Resistance. *Clin. Cancer Res.* **27** (9), 2636–2647 (2021).
25. Daniel, S. K., Seo, Y. D. & Pillarisetty, V. G. The CXCL12-CXCR4/CXCR7 axis as a mechanism of immune resistance in gastrointestinal malignancies. *Semin Cancer Biol.* **65**, 176–188 (2020).
26. Galili, U. & Zhao, Y. Potential induction of protective anti-tumor immune response in cancer patients by oncolytic viruses containing the GGTA1 gene. *Eur. J. Cancer*. **236**, 116251 (2026).
27. He, M., Zhou, X. & Wang, X. Glycosylation: mechanisms, biological functions and clinical implications. *Signal. Transduct. Target. Ther.* **9** (1), 194 (2024).
28. Stanczak, M. A. et al. Targeting cancer glycosylation repolarizes tumor-associated macrophages allowing effective immune checkpoint blockade. *Sci. Transl Med.* **14** (669), eabj1270 (2022).
29. Pinho, S. S., Macauley, M. S. & Läubli, H. Tumor glyco-immunology, glyco-immune checkpoints and immunotherapy. *J. Immunother Cancer* ;**13**(6). (2025).
30. Van den Bossche, V. et al. PPARα-mediated lipid metabolism reprogramming supports anti-EGFR therapy resistance in head and neck squamous cell carcinoma. *Nat. Commun.* **16** (1), 1237 (2025).
31. Malta, T. M. et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173** (2), 338– (2018). – 54.e15.
32. Gomez, K. E. et al. Cancer Cell CD44 Mediates Macrophage/Monocyte-Driven Regulation of Head and Neck Cancer Stem Cells. *Cancer Res.* **80** (19), 4185–4198 (2020).
33. Frejborg, E., Salo, T. & Salem, A. Role of Cyclooxygenase-2 in Head and Neck Tumorigenesis. *Int. J. Mol. Sci.* ;**21**(23). (2020).
34. Watermann, C. et al. Recurrent HNSCC Harbor an Immunosuppressive Tumor Immune Microenvironment Suggesting Successful Tumor Immune Evasion. *Clin. Cancer Res.* **27** (2), 632–644 (2021).
35. Wils, L. J. et al. Elucidating the Genetic Landscape of Oral Leukoplakia to Predict Malignant Transformation. *Clin. Cancer Res.* **29** (3), 602–613 (2023).
36. McGrail, D. J. et al. High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Ann. Oncol.* **32** (5), 661–672 (2021).
37. Gu, Z. et al. Pharmacogenomic landscape of head and neck squamous cell carcinoma informs precision oncology therapy. *Sci. Transl Med.* **14** (661), eabo5987 (2022).

Figures

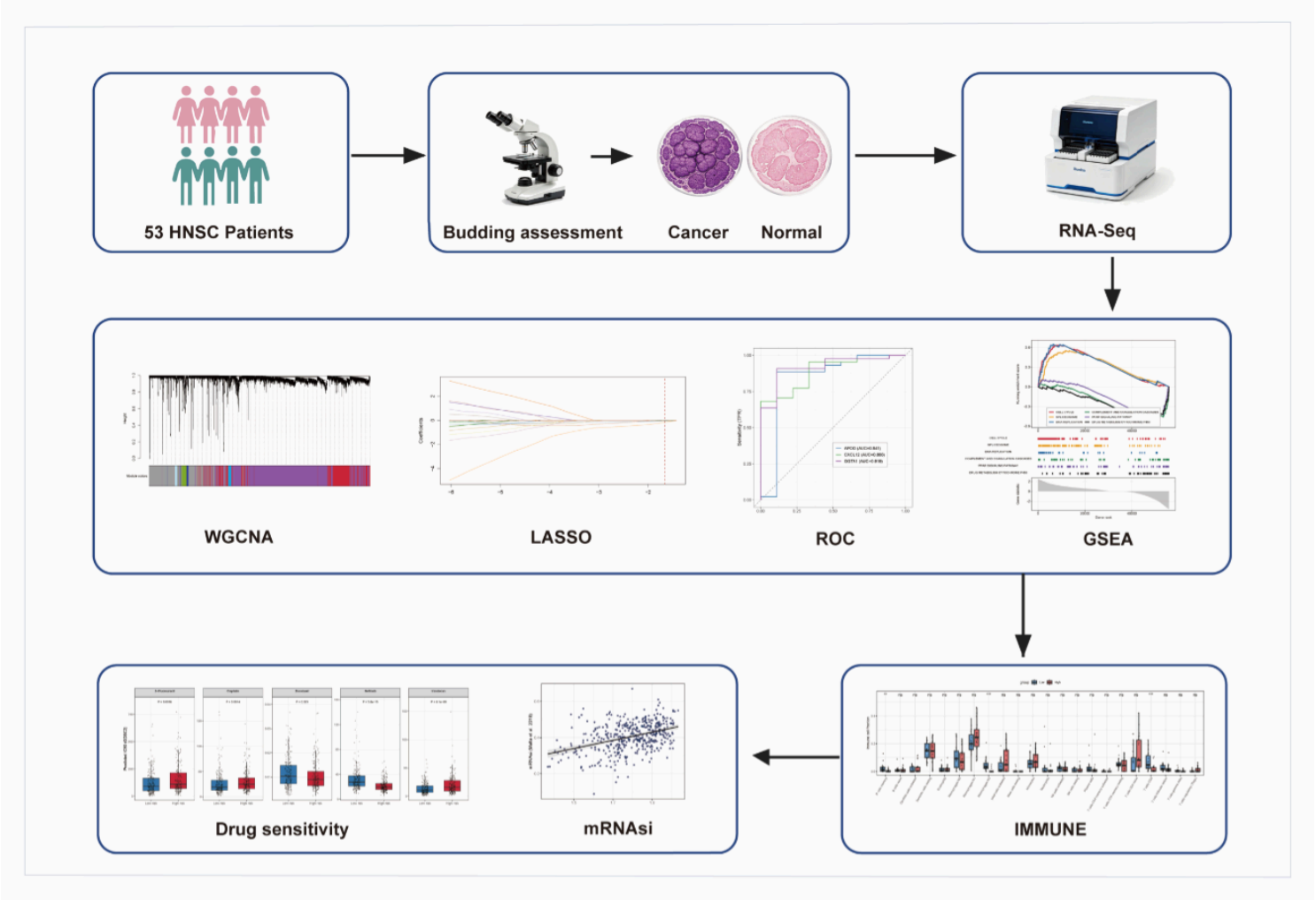

Figure 1

The flow chart of the current study.

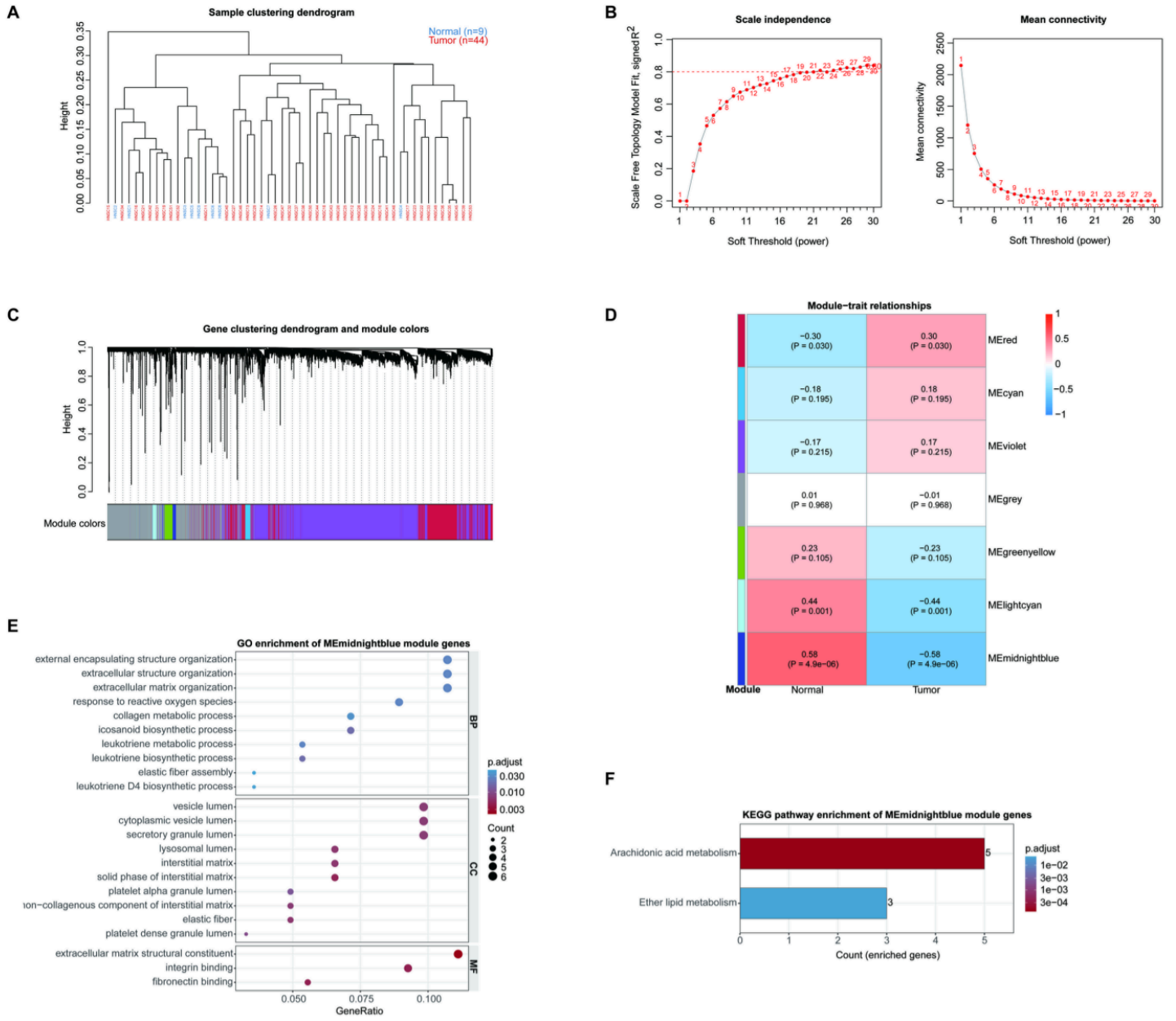

Figure 2

Co-expression network and module enrichment.

(A) Sample clustering dendrogram; leaf labels coloured by sample type. (B) Scale-free fit and mean connectivity versus soft-thresholding power; power 22 selected. (C) Gene dendrogram with module colours; grey, unassigned genes. (D) Module-trait correlations; coefficients and P values in cells. (E) Gene Ontology enrichment of the midnightblue module in three domains; dot size, gene count; colour, adjusted P value. (F) KEGG enrichment of the same module; bar length, gene count. FPKM, fragments per kilobase of transcript per million mapped reads. KEGG enrichment was annotated using the KEGG pathway database(19, 20).

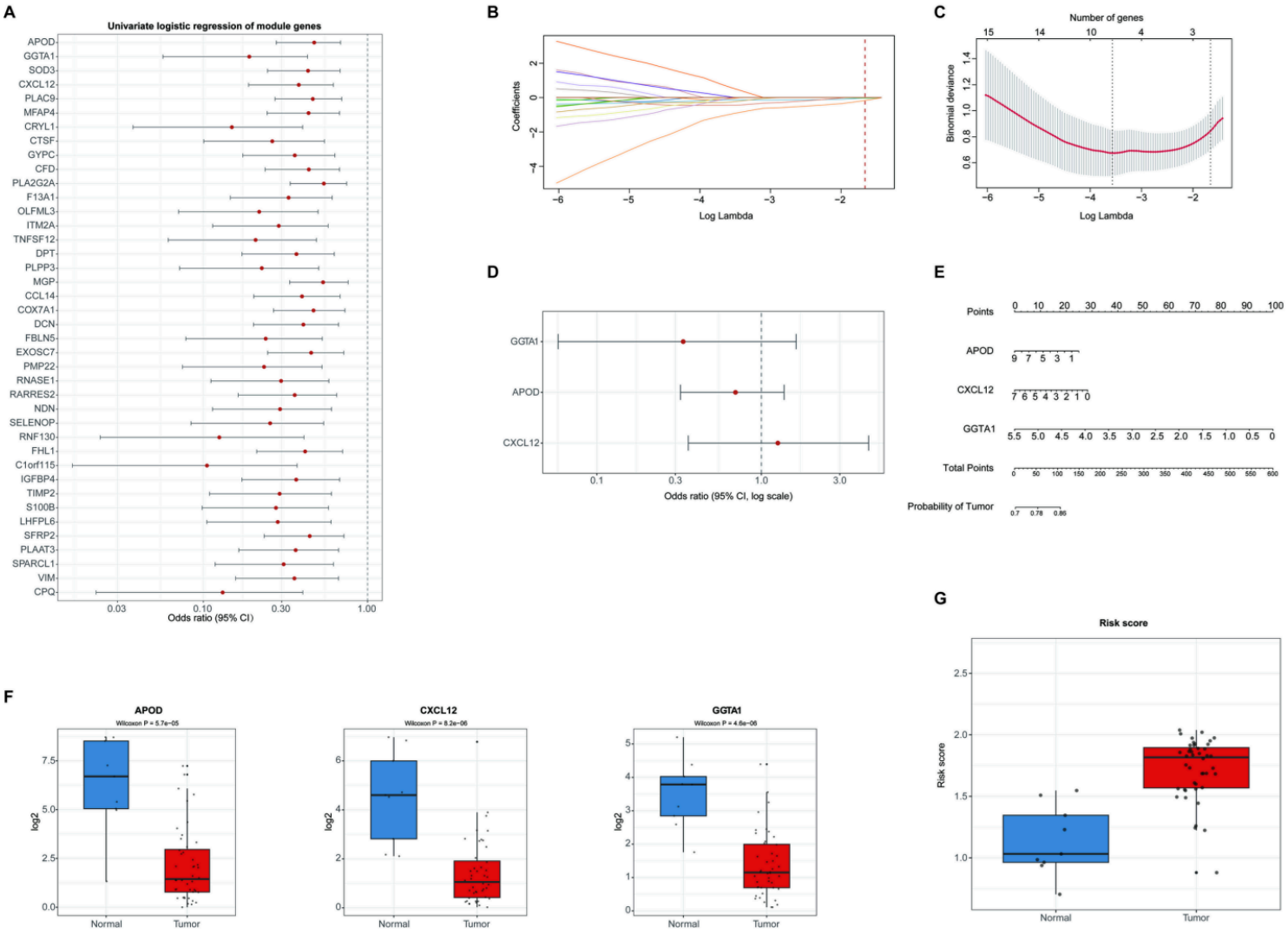

Figure 3

Construction of the three-gene score and nomogram.

(A) Univariate logistic regression of module genes; odds ratios (ORs) with 95% confidence intervals (CIs); red, p<0.05. (B) LASSO coefficient paths. (C) Ten-fold cross-validation curve; dashed lines, lambda.min and lambda.1se. (D) Multivariable Firth-penalised logistic regression of the three genes. (E) Nomogram; total points give the predicted probability of tumour tissue. (F) Expression of APOD, CXCL12 and GGTA1 in non-tumour and tumour tissue; box plots with Wilcoxon rank-sum P values. (G) Risk score in the two groups.

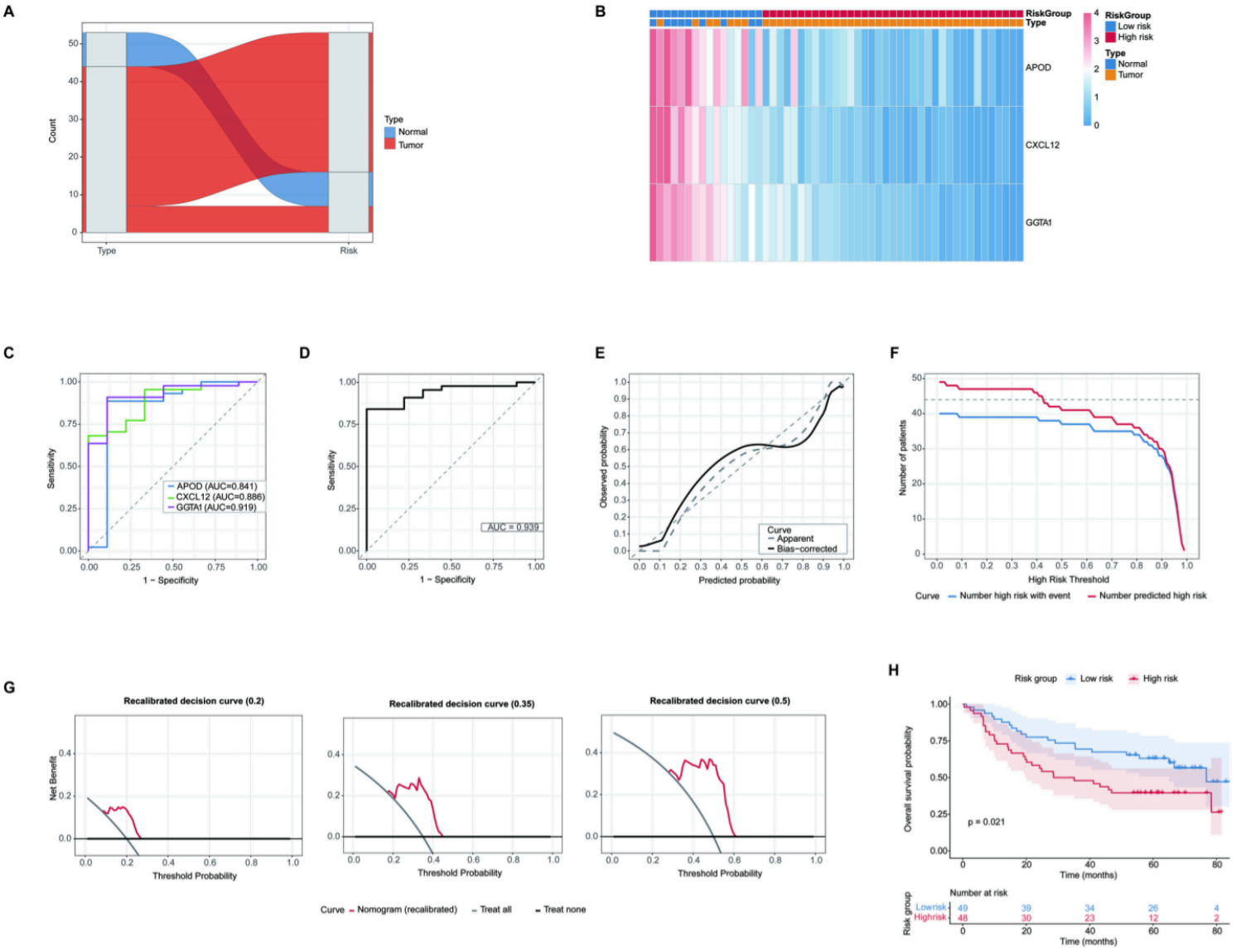

Figure 4

Internal evaluation and overall survival.

(A) Alluvial plot of sample type and risk group. (B) Heatmap of the three genes; samples ordered by risk score. (C) Single-gene ROC curves from cross-validation; AUCs in the legend. (D) ROC curve of the combined model; AUC and 95% CI. (E) Calibration curve; dashed, apparent; solid, bootstrap-corrected; dotted, ideal; H-L, Hosmer-Lemeshow. (F) Clinical impact curve. (G) Decision curves after recalibration to prevalences of 0.20, 0.35 and 0.50; red, nomogram; grey, treat-all; horizontal line, treat-none. (H) Kaplan-Meier overall survival in GSE41613; log-rank P value and numbers at risk. ROC, receiver operating characteristic; AUC, area under the curve.

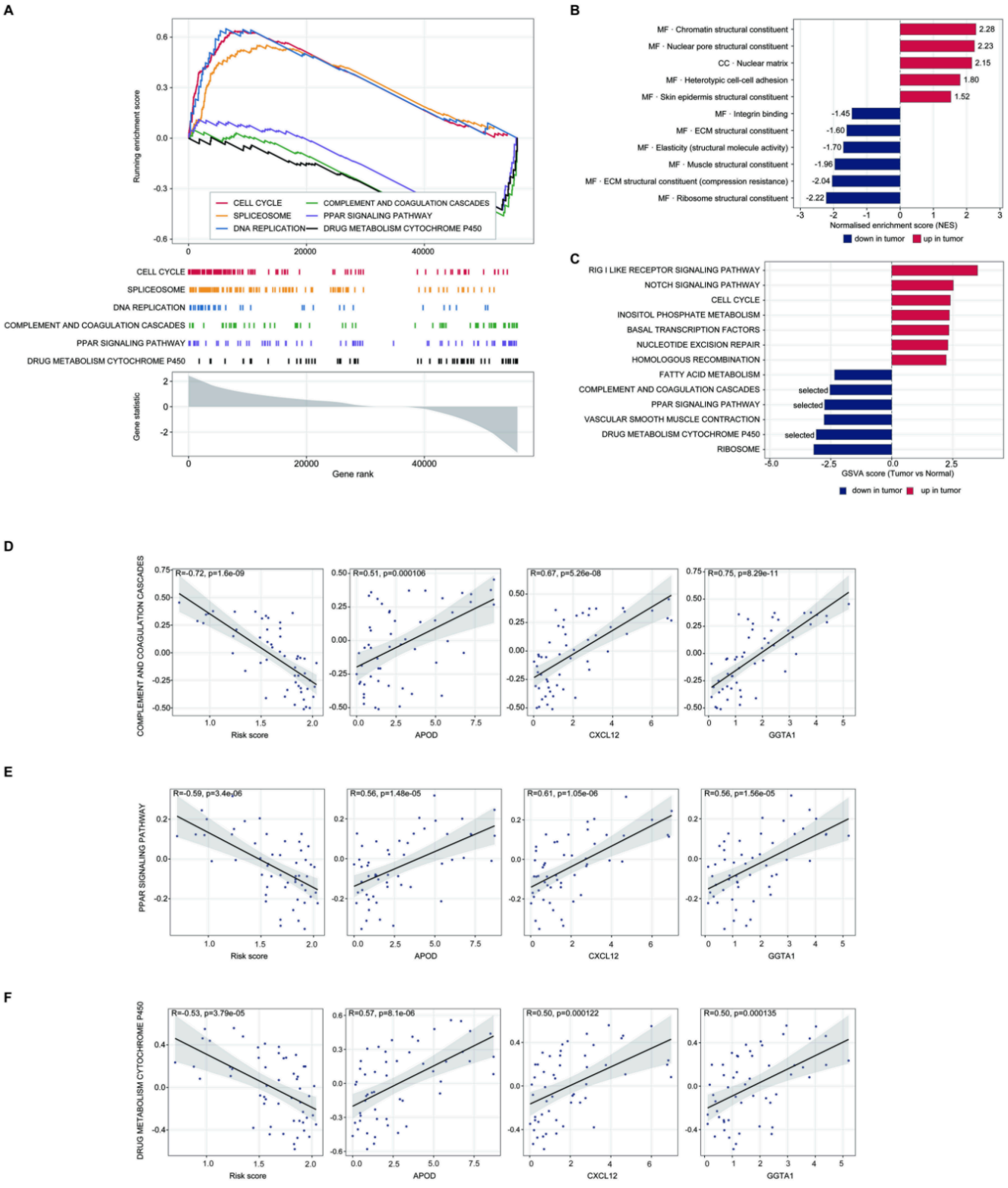

Figure 5

KEGG pathway landscape and its association with the score.

(A) GSEA of KEGG gene sets; top, enrichment score curves; middle, gene positions; bottom, ranked statistic. (B) Targeted GSEA of extracellular matrix (ECM)-related Gene Ontology sets; bars, normalised enrichment score (NES); blue, higher in non-tumour tissue; red, higher in tumour tissue. (C) GSVA score

comparison for screened KEGG pathways; "selected" pathways taken forward. (D-F) GSVA score of complement and coagulation cascades (D), PPAR signalling (E) and cytochrome P450 drug metabolism (F) versus the risk score and the three genes; Pearson coefficients shown. GSEA, gene set enrichment analysis; GSVA, gene set variation analysis. KEGG pathway gene sets were retrieved from the KEGG database (19, 20).

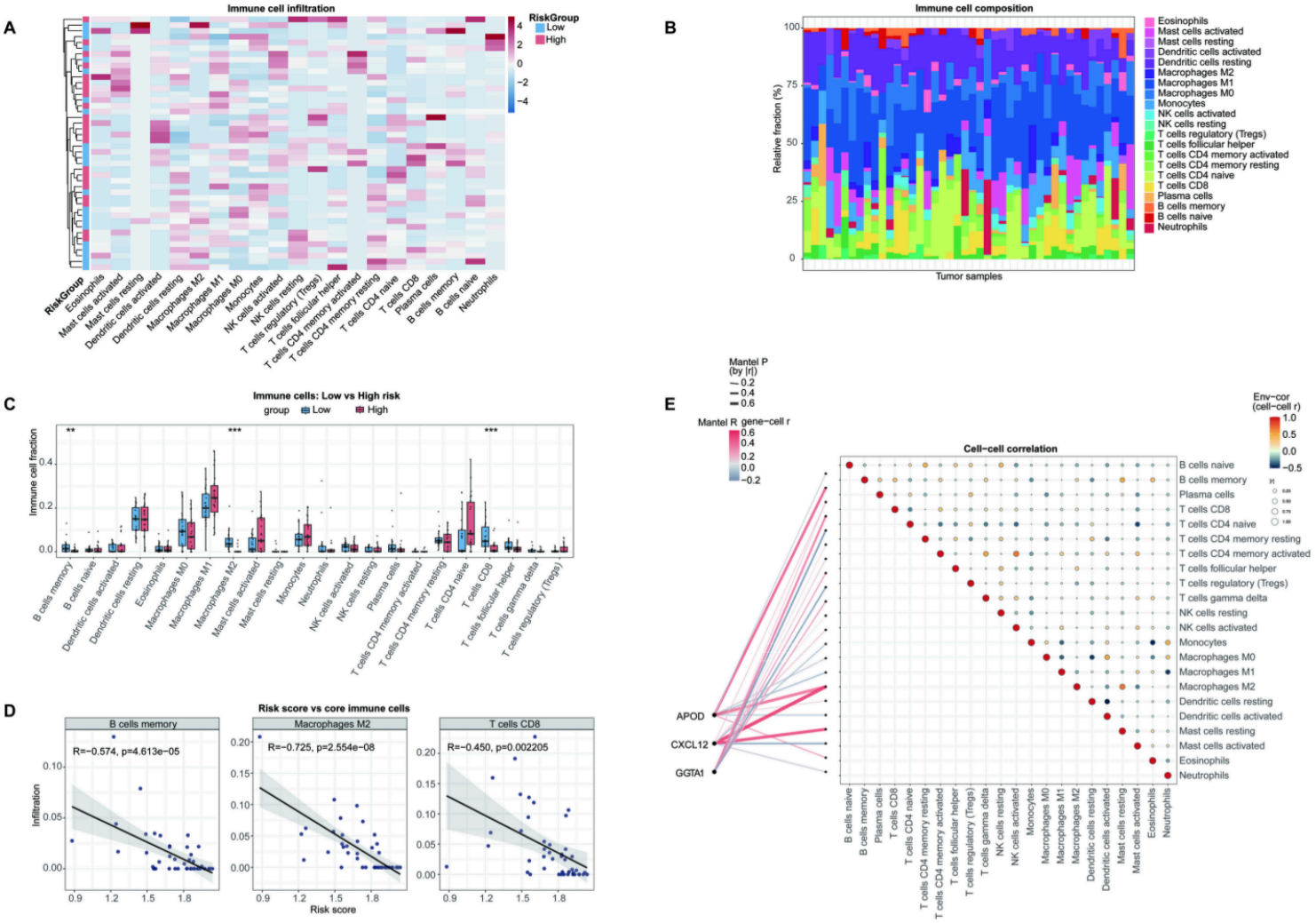

Figure 6

Immune cell composition by risk group.

(A) CIBERSORT fractions (LM22; 21 of 22 cell types are displayed, gamma-delta T cells being zero in most samples); z scores across samples ordered by risk score. (B) Stacked fractions of the same samples. (C) Fractions by risk group; Wilcoxon rank-sum P values with Benjamini-Hochberg adjustment. (D) Risk score versus CD8-positive T cell, M2 macrophage and memory B cell fractions; Pearson coefficients. (E) Gene expression or risk score versus an 11-cell panel (left) and cell-cell correlation matrix (right).

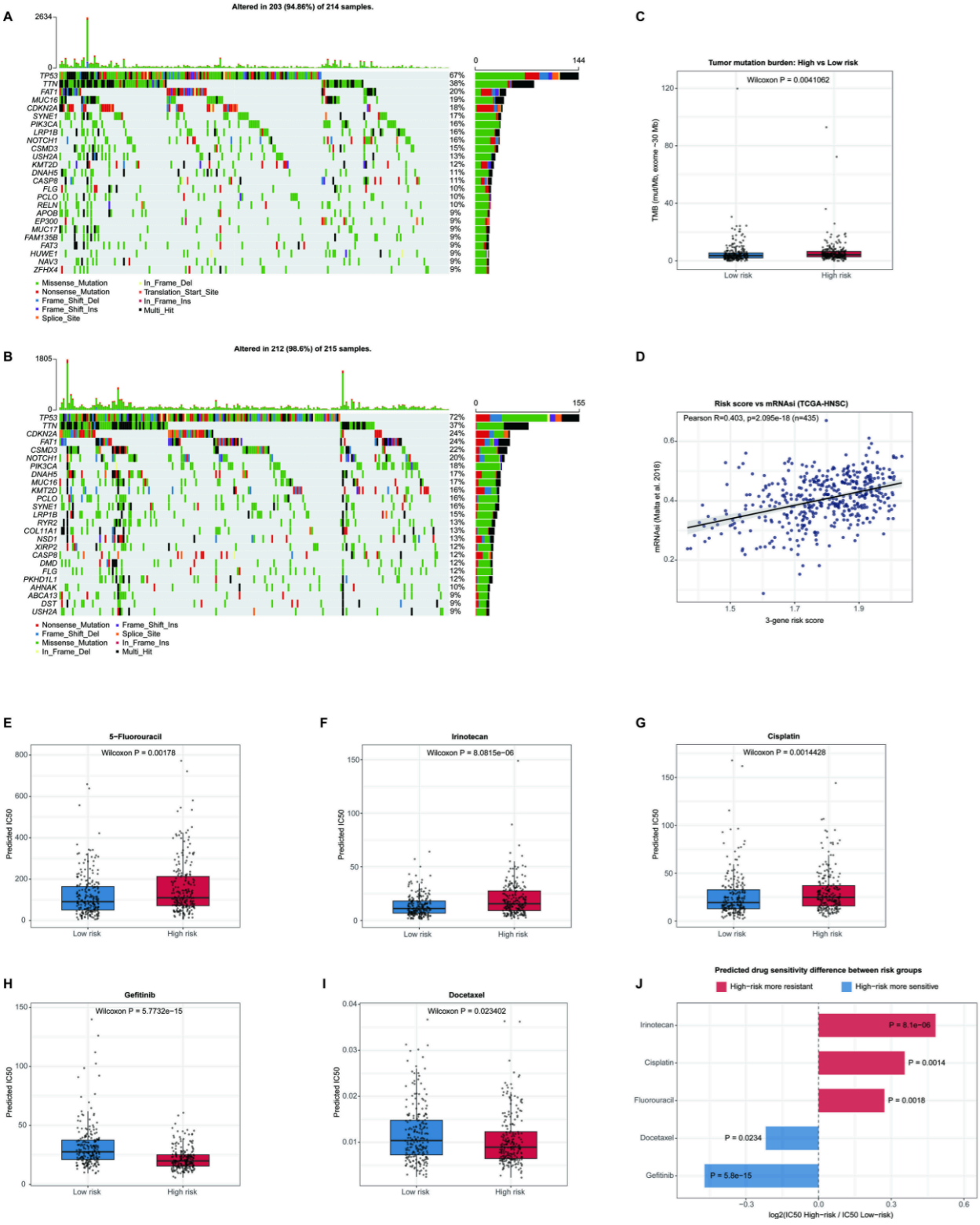

Figure 7

Mutation, mutational burden, stemness and predicted drug sensitivity in TCGA-HNSC.

(A, B) Somatic mutation landscape in low-risk (A) and high-risk (B) groups; rows, genes; colour, variant class. (C) Tumour mutational burden by risk group; Wilcoxon P value. (D) Risk score versus mRNA stemness index (mRNAsi); Pearson coefficient and P value. (E-I) Predicted IC50 of 5-fluorouracil (E),

irinotecan (F), cisplatin (G), gefitinib (H) and docetaxel (I); Wilcoxon P values. (J) log2 ratio of median predicted IC50 (high risk / low risk); red, higher in high risk; blue, lower. IC50, half-maximal inhibitory concentration.

Supplementary Files

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

- floatimage8.png