

# Multi-Omic Integration Reveals a Lung Adenocarcinoma-Specific Tumour-Suppressor Survival Axis Where High-Dimensional Prediction Fails

Thoraia Shinawi

Department of Medical Laboratory Sciences, Faculty of Applied Medical Sciences  
King Abdulaziz University, P.O. Box 80200, Jeddah 21589, Saudi Arabia

\*Corresponding author: tshinawi@kau.edu.sa

## Abstract

Multi-omic profiling is widely expected to improve cancer survival prediction, yet rigorous, transparent comparisons against simple clinical baselines remain scarce. We assembled a seven-modality matrix (gene expression, DNA methylation, somatic mutation, copy-number variation, reverse-phase protein array, clinical variables and protein-interaction context) for 628 patients with lung adenocarcinoma (LUAD) and 497 patients with lung squamous cell carcinoma (LUSC) from The Cancer Genome Atlas. A multimodal variational-autoencoder survival network and a panel of classical machine-learning models were trained and compared for overall-survival discrimination. Deep multimodal survival learning achieved a mean cross-validated concordance index (C-index) of 0.555, whereas a clinical-only random forest attained 0.633; adding high-dimensional omics consistently reduced performance, with the all-feature configuration falling to 0.593. Modality-ablation and nested cross-validation confirmed that this gap reflects the sample-to-feature regime (628 samples versus 1,902 features) rather than model choice. We therefore repurposed the omic data for biological discovery. A six-gene tumour-suppressor panel (KEAP1, STK11, TP53, SMARCA4, NF1 and CDKN2A) added independent prognostic information in LUAD (panel-plus-clinical C-index 0.602; panel hazard ratio 1.16,  $p$

$= 0.025$ ) but did not validate in LUSC (hazard ratio 0.94,  $p = 0.48$ ), indicating histology-specific biology. STK11 mutation was the strongest single-gene signal (log-rank  $p = 0.044$ ). High-risk LUAD tumours showed coordinated up-regulation of cancer-testis antigens (MAGEA family, PRAME;  $p < 10^{-14}$ ), enrichment of lipid and steroid metabolism ( $p = 0.011$ ) and extracellular-matrix degradation ( $p = 0.032$ ), together with reduced KEAP1 protein ( $p = 0.001$ ). Panel-gene recurrence was confirmed across seven independent cBioPortal lung cohorts. These results demonstrate that multi-omic data can yield coherent, mechanistically interpretable biology even when they fail to improve prediction, and define a LUAD-specific metabolic-aggressive survival axis with therapeutic implications.

**Keywords:** lung adenocarcinoma; multi-omic integration; survival analysis; tumour-suppressor genes; metabolic reprogramming; machine learning

## Introduction

Lung cancer remains the leading cause of cancer-related death worldwide, and lung adenocarcinoma (LUAD) is its most common histological subtype (1,2). Despite refinements in anatomical staging, patients within the same stage experience widely divergent outcomes, with five-year survival ranging from roughly 10 % to 70 % (3). This heterogeneity has motivated

intense interest in molecular markers that could refine prognosis and guide therapy beyond conventional clinicopathological variables.

The maturation of large public resources such as The Cancer Genome Atlas (TCGA) has made it possible to profile the same tumours across multiple molecular layers, including the transcriptome, methylome, somatic mutation landscape, copy-number architecture and proteome (4,5). A prevailing expectation is that integrating these complementary layers should improve survival prediction relative to clinical data alone, because each modality captures distinct aspects of tumour biology. A growing body of work has applied multimodal and deep-learning methods to this problem, reporting encouraging discrimination across several cancer types (6,7,8).

However, two recurring problems undermine confidence in this literature. First, high-dimensional omic survival studies frequently omit a fair comparison against a simple clinical baseline, so it is often unclear whether reported performance genuinely exceeds what age, stage and performance status already provide. Second, genomic survival datasets typically pair a modest number of patients (hundreds) with very large feature spaces (thousands to hundreds of thousands), a regime in which the curse of dimensionality erodes generalisation and favours overfitting (9,10). Classical ensembles such as random survival forests are known to perform competitively, and sometimes better than deep networks, when samples are scarce relative to features (11).

These concerns are not merely methodological. If multi-omic data do not improve prediction in realistic sample-size regimes, the field risks overstating clinical readiness. Equally, a negative prediction result need not imply that omic data are uninformative: the same data may still reveal reproducible, mechanistically interpretable biology that single-gene or clinical analyses cannot. This dual perspective — prediction versus interpretation — is rarely examined within a single, transparent study.

Here we address both questions directly. We built a seven-modality survival dataset for LUAD and an aligned LUSC validation cohort, and we systematically compared a multimodal variational-autoencoder survival network against classical machine-learning baselines, including a clinical-only model. Having found that clinical variables dominate prediction and that adding omics degrades it, we explicitly diagnosed why, then repurposed the omic data to characterise the molecular biology of aggressive LUAD. This led to a histology-specific tumour-suppressor survival axis linking *STK11*, *KEAP1*, *SMARCA4* and related alterations to a metabolically reprogrammed, extracellular-matrix-remodelling phenotype with activated cancer-testis antigen expression. We further confirmed the recurrence of these panel genes across seven independent lung cancer cohorts. The study therefore provides both an honest appraisal of multi-omic prediction limits and a coherent biological discovery derived from the same data.

## Methods and Materials

### Cohorts and data sources

Processed multi-omic matrices for TCGA-LUAD ( $n = 628$ ) and TCGA-LUSC ( $n = 497$ ) were obtained from the UCSC Xena platform (5). LUAD served as the primary discovery and modelling cohort, and LUSC as an external lung validation cohort. For independent validation of panel-gene recurrence, seven lung cancer studies were retrieved from cBioPortal (12,13): LUAD CPTAC 2020 ( $n = 110$ ), LUAD MSK-IMPACT 2021 ( $n = 186$ ), LUAD TCGA PanCancer Atlas ( $n = 566$ ), Lung MSK 2017 ( $n = 915$ ), Lung NCI 2022 ( $n = 232$ ), NSCLC MSKCC 2018 ( $n = 75$ ) and NSCLC TRACERx 2017 ( $n = 447$ ). Pathway annotations were obtained from the Reactome knowledgebase (14) and protein-protein interaction context from the STRING database (15). After de-duplication, the assembled inventory comprised ten unique patient cohort groups and 5,707 counted samples or cases, together with non-patient reference resources.

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

### **Multi-omic data processing**

For each TCGA cohort, native feature identifiers were resolved for every matrix and harmonised to gene-level symbols. Six modalities were retained for modelling after feature selection: gene expression (500 features), DNA methylation (500 features), somatic mutation (419 features), copy-number variation (200 features), reverse-phase protein array (RPPA) protein abundance (277 features) and clinical variables (6 features), yielding a combined matrix of 1,902 features per patient. Continuous modalities (expression, methylation, copy number, protein, clinical) were standardised to zero mean and unit variance; the mutation matrix was encoded as binary alteration indicators. The LUSC matrix was aligned to the LUAD-selected feature space to enable external evaluation.

Data completeness was audited for sample counts, feature dimensions, missingness and event rates. After preprocessing, the final LUAD and LUSC matrices contained no missing values. The mutation matrix density was 8.3 % positive in LUAD and 8.8 % in LUSC, consistent with the expected sparsity of somatic alterations.

### **Survival outcome definition**

The primary outcome was overall survival, defined as time from diagnosis to death from any cause, with censoring at last follow-up. Survival durations were recorded in days (LUAD mean 932, standard deviation 891; LUSC mean 975, standard deviation 961). The observed event fraction was 0.381 in LUAD (239 events) and 0.433 in LUSC (215 events).

### **Deep multimodal survival model**

A multimodal survival network was implemented with modality-specific encoders feeding a shared fusion layer. A variational-auto-encoder encoder (16) compressed the concatenated high-dimensional expression and methylation block; a separate encoder processed the mutation block; and a tabular encoder processed the combined clinical, copy-number and

protein block. The three latent representations were combined through a cross-modal attention layer (17) that learned modality weights, followed by two survival heads: a Weibull accelerated-failure-time head and a discrete-time DeepHit-style head (18). Models were trained with the Adam optimiser and L2 weight decay, with early stopping on validation discrimination.

Performance was quantified by Harrell's concordance index (C-index) (19). The discovery cohort was evaluated by five-fold cross-validation; external generalisation was assessed by training on LUAD and evaluating on LUSC. Because the survival-head sign convention was not fixed in this research-grade implementation, the better-oriented C-index (the maximum of survival-score and risk-score orientations) is reported throughout, which is conservative in favour of the deep model. Modality ablation was performed by masking individual modalities under an identical architecture, so reported deltas are directional rather than definitive causal estimates.

### **Classical machine-learning baselines**

Four classical survival-adjacent models were trained for comparison: a ridge-penalised Cox model, an L1-penalised (LASSO) duration model, a random-forest duration model and a gradient-boosting duration model (11,20). Feature configurations included clinical-only (6 features), mutation-only, expression-plus-mutation-plus-clinical, and all-omics-plus-clinical. Survival analyses and model evaluation used the lifelines and scikit-survival libraries in Python (21,22).

### **Selected-omics benchmarking and nested validation**

To test whether targeted feature selection could recover omic signal, features in each modality were ranked by univariate event association and by correlation with survival duration. The top  $k$  features per modality were retained across a sweep of  $k \in \{5, 10, 25, 50, 100, 200\}$  and supplied, together with clinical variables, to random-forest and gradient-boosting mod-

els. To control optimism arising from selecting features on the full dataset, a nested cross-validation procedure was additionally performed in which feature selection was confined to the inner training folds.

### ***Prognostic panel construction and survival analysis***

Candidate genes were prioritised by integrating univariate survival association, effect size, machine-learning importance (ExtraTrees event classification and random-forest duration regression), multi-omic support across modalities, and STRING network degree, with an additional prior for established lung cancer drivers and actionability. From this prioritisation a compact, biologically non-redundant tumour-suppressor panel was defined as KEAP1, STK11, TP53, SMARCA4, NF1 and CDKN2A. A patient panel score was computed as the sum of non-synonymous mutations across the five mutation-defined genes plus an indicator for CDKN2A copy-number loss; patients with a score of two or more were classified as high-risk.

The panel was evaluated by Kaplan-Meier analysis with log-rank testing and by Cox proportional-hazards modelling in three specifications: panel-only, clinical-only and panel-plus-clinical. Discrimination was summarised by the C-index and effect size by the partial hazard ratio (HR). Identical procedures were applied in LUAD and LUSC. Gene-level and co-mutation survival associations were assessed for each panel gene, for clinically relevant co-mutations (including KRAS-context combinations) and for cumulative panel burden; subgroups with fewer than 20 positive cases were flagged as hypothesis-generating.

### ***High-risk molecular signature and pathway enrichment***

LUAD patients were stratified into high-risk ( $n = 261$ ) and low-risk ( $n = 367$ ) groups by panel score. Differential analysis between groups was performed for expression (500 features), methylation (500 features), copy number (200 features) and protein (277 features)

using two-sample t-tests with standardised effect sizes; p-values were adjusted within each modality using the Benjamini-Hochberg false-discovery-rate (FDR) procedure (23). Over-representation analysis of the top up-regulated expression genes against 2,855 Reactome pathways was performed using the hypergeometric test with FDR correction.

### ***External recurrence validation***

For each of the seven cBioPortal cohorts, panel-gene mutation frequencies, copy-number loss and gain frequencies, and co-mutation frequencies were computed from the available molecular profiles. Harmonised overall-survival time-event pairs were not consistently available in the retrieved cBioPortal exports; external analysis was therefore restricted to alteration recurrence rather than survival replication.

### ***Statistical analysis***

All analyses were conducted in Python using NumPy, pandas, SciPy, scikit-learn, lifelines and scikit-survival (21,22). Statistical significance was assessed at a two-sided  $\alpha$  of 0.05. Multiple comparisons within each high-risk signature modality were controlled by the Benjamini-Hochberg FDR.

## ***Results and Discussion***

### ***Cohort composition and data quality***

The discovery LUAD cohort comprised 628 patients with 239 deaths (event fraction 0.381) and the LUSC validation cohort 497 patients with 215 deaths (event fraction 0.433). Both cohorts were represented by the same 1,902-feature multi-omic matrix spanning expression, methylation, mutation, copy number, protein and clinical modalities (Supplementary Table S1). After preprocessing, neither matrix contained missing values, and mutation-matrix density (8.3 % in LUAD, 8.8 % in LUSC) was consistent with expected somatic-alteration sparsity.

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

### **Clinical variables outperform deep multi-modal survival learning**

The multimodal variational-autoencoder survival network achieved a mean five-fold cross-validated C-index of 0.555 (standard deviation 0.031) in LUAD, with fold-level values ranging from 0.526 to 0.598 (Table 1; Figure 1a; Supplementary Table S2). Discrimination was unstable across folds, with the optimal orientation alternating between survival-score and risk-score conventions, indicating a weak and inconsistent signal (Figure 2a). External evaluation on LUSC yielded a C-index of 0.529, below the in-cohort mean and consistent with poor cross-histology generalisation.

In contrast, a clinical-only random forest using just six variables achieved a mean C-index of 0.633 (standard deviation 0.073), substantially exceeding the deep model (Table 1; Figure 1a). The per-fold distributions confirmed that the clinical model was both higher and more stable than the omic-augmented and deep models, whose fold-level performance frequently fell to or below chance (Figure 2a). A clinical-only Cox baseline reached 0.575. Thus a low-dimensional clinical model outperformed a high-dimensional multimodal network trained on the same patients.

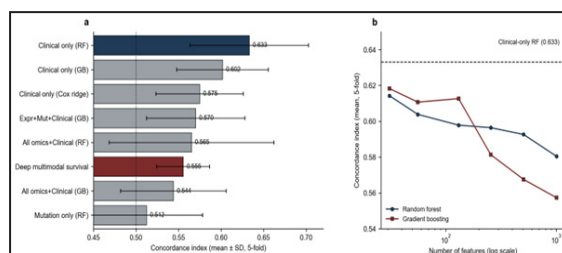


Figure.1: Model performance

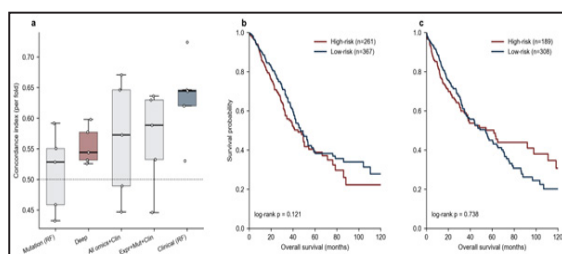


Figure.2: Prediction survival

### **Adding omic modalities degrades prediction**

Modality-ablation analysis clarified why integration did not help (Supplementary Table S3; Figure S1). Under an identical architecture, the all-modality configuration achieved a best-oriented C-index of only 0.502. Removing expression improved discrimination by 0.075, and removing mutation improved it by 0.066, indicating that these high-dimensional blocks introduced more noise than signal when combined. The single best modality in isolation was mutation (0.587), followed by clinical (0.554), while expression alone performed at chance (0.500). No modality combination matched the clinical-only random forest.

### **Selected-omics benchmarking and nested validation confirm the gap**

A systematic feature-selection sweep reinforced these findings (Table 1; Figure 1b; Supplementary Table S5). The clinical-only random forest (six features) remained the top performer at a mean C-index of 0.633. The best omic-augmented configuration reached 0.618, and discrimination declined monotonically as more omic features were added, falling to 0.557 at 1,006 features (Figure 1b). Nested cross-validation, which confines feature selection to inner training folds, yielded best omic performance below the clinical-only random forest, confirming that the modest apparent benefit of aggressive selection reflected optimism rather than reproducible signal.

Taken together, the discovery cohort of 628 patients against a 1,902-feature space placed the analysis firmly in a regime where high-dimensional integration cannot reliably exceed a parsimonious clinical model. We therefore turned to interpretation rather than prediction.

### **A six-gene tumour-suppressor panel adds prognostic information in LUAD but not LUSC**

Gene-level survival analysis identified a compact tumour-suppressor panel (KEAP1, STK11, TP53, SMARCA4, NF1 and CDKN2A).

Table 1. Overall-survival discrimination across feature sets, models and feature dimensionality in LUAD (five-fold cross-validation). The clinical-only random forest is the reference for the performance delta; the deep multimodal network additionally yielded an external LUSC C-index of 0.529. C-index, concordance index; RF, random forest; GB, gradient boosting; Cox, ridge-penalised Cox; LASSO, L1-penalised duration model; SD, standard deviation.

| Feature set                         | Model             | No. features | C-index (mean $\pm$ SD) | Range (min–max) | $\Delta$ vs clinical RF |
|-------------------------------------|-------------------|--------------|-------------------------|-----------------|-------------------------|
| Clinical                            | Random forest     | 6            | 0.633 $\pm$ 0.070       | 0.530–0.724     | reference               |
| Clinical + selected omics (k = 5)   | Gradient boosting | 31           | 0.618 $\pm$ 0.083       | 0.485–0.688     | –0.015                  |
| Clinical + selected omics (k = 5)   | Random forest     | 31           | 0.614 $\pm$ 0.071       | 0.525–0.676     | –0.019                  |
| Clinical                            | Gradient boosting | 6            | 0.602 $\pm$ 0.054       | 0.523–0.658     | –0.031                  |
| Clinical + selected omics (k = 100) | Random forest     | 506          | 0.593 $\pm$ 0.084       | 0.503–0.682     | –0.040                  |
| Clinical                            | Cox ridge         | 6            | 0.575 $\pm$ 0.051       | 0.524–0.660     | –0.058                  |
| Expression + mutation + clinical    | Gradient boosting | 925          | 0.570 $\pm$ 0.058       | 0.489–0.640     | –0.063                  |
| Expression + mutation + clinical    | Random forest     | 925          | 0.567 $\pm$ 0.079       | 0.446–0.636     | –0.066                  |
| All omics + clinical                | Random forest     | 1,902        | 0.565 $\pm$ 0.097       | 0.447–0.671     | –0.068                  |
| Deep multimodal (VAE + attention)   | Weibull + DeepHit | 1,902        | 0.555 $\pm$ 0.031       | 0.526–0.598     | –0.078                  |
| All omics + clinical                | Gradient boosting | 1,902        | 0.544 $\pm$ 0.062       | 0.448–0.609     | –0.089                  |
| Mutation                            | Random forest     | 419          | 0.512 $\pm$ 0.066       | 0.432–0.592     | –0.121                  |

In LUAD, the panel score added independent prognostic information beyond clinical covariates: the panel-plus-clinical Cox model reached a C-index of 0.602 with a significant panel hazard ratio of 1.16 ( $p = 0.025$ ), compared with 0.595 for clinical-only and 0.548 for panel-only specifications (Table 2; Supplementary Table S4). In LUSC, the same panel failed to validate, with a panel hazard ratio below unity (0.94,  $p = 0.48$ ) and no improvement over the clinical model, indicating that the prognostic signal is specific to adenocarcinoma rather than common to lung cancer.

Kaplan-Meier stratification was concordant with the Cox results. In LUAD, high-risk patients ( $n = 261$ ) showed a higher event rate (39.5 %) than low-risk patients ( $n = 367$ , 37.1 %), with marginal separation (log-rank  $p = 0.121$ ; Figure 2b), whereas in LUSC the curves were essen-

tially superimposed (log-rank  $p = 0.738$ ; Figure 2c). Ranking LUAD patients by panel score showed that deaths were distributed across the burden spectrum but were enriched among tumours above the high-risk threshold (Figure 5c).

Principal-component analysis of the panel-gene and clinical feature space showed broadly overlapping distributions for LUAD and LUSC and for high- versus low-risk LUAD tumours (Figure 5a,b), consistent with the modest discrimination of the panel and with the absence of a single dominant axis of separation. Co-occurrence networks of panel alterations were, however, differently wired between histologies: the densely co-altered TP53–CDKN2A–KEAP1 structure seen in LUAD was reconfigured in LUSC, where TP53 dominated and STK11 was nearly absent (Figure 4a,b), reinforcing the histology specificity of the axis.

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

**STK11 is the strongest single-gene signal and KRAS is context-dependent**

Among individual alterations, STK11 mutation provided the clearest survival association in LUAD (log-rank  $p = 0.044$ , HR = 1.40; Table 2, Figure 3a). CDKN2A copy-number loss, present in 65.9 % of tumours, showed a borderline association ( $p = 0.065$ ). Notably, KRAS mutation alone was not prognostic ( $p = 0.601$ , HR = 1.08) despite a frequency of 23 %, whereas the KRAS-plus-SMARCA4 co-mutation defined a small but markedly high-risk subgroup ( $n = 8$ , HR = 2.20,  $p = 0.021$ ) and the STK11-plus-KEAP1 combination was suggestive ( $n = 23$ , HR = 1.61,  $p = 0.106$ ). A co-mutation hazard network placed KRAS as a central hub whose pairwise combinations carried the elevated hazards, while the corresponding co-oc-

Table 2. Genomic frequency, survival association, external recurrence and therapeutic context of panel genes and key co-mutations in LUAD. Frequencies are within the TCGA-LUAD discovery cohort ( $n = 628$ ); event rates compare altered versus wild-type tumours; hazard ratios (HR) with 95 % confidence intervals and log-rank  $p$ -values are from univariate analysis with clinical covariates. External mean frequency is the mean mutation frequency across seven cBioPortal lung cohorts. CNA, copy-number alteration; n.a., not applicable.

currence/mutual-exclusivity network showed that the highest-hazard partners (SMARCA4, KEAP1) tended to co-occur rather than exclude one another (Figure 4c,d). These small subgroups are hypothesis-generating, but together they indicate that KRAS confers risk chiefly in specific tumour-suppressor contexts rather than in isolation. Panel-gene alterations were recurrent across seven independent external lung cohorts, with mean mutation frequencies closely matching the discovery cohort (Table 2, rightmost column; Figure S2), confirming that the panel captures recurrent rather than cohort-specific biology; harmonised survival fields were unavailable in those exports, so external analysis establishes recurrence but not survival replication.

| Alteration              | Pathway / therapeutic context                                  | LUAD n (%) | Event rate altered vs WT (%) | Log-rank p | HR (95% CI)      | External mean freq (%) |
|-------------------------|----------------------------------------------------------------|------------|------------------------------|------------|------------------|------------------------|
| STK11 mutation          | LKB1–AMPK metabolic checkpoint; AMPK activators, metformin     | 69 (11.0)  | 43.5 vs 37.4                 | 0.044      | 1.40 (0.98–1.99) | 10.7                   |
| KEAP1 mutation          | KEAP1–NRF2 oxidative-stress axis; NRF2/ glutaminase inhibitors | 90 (14.3)  | 43.3 vs 37.2                 | 0.687      | 1.07 (0.78–1.47) | 10.5                   |
| TP53 mutation           | p53 tumour suppressor; not directly drug-gable                 | 254 (40.4) | 37.0 vs 38.8                 | 0.277      | 1.15 (0.91–1.45) | 50.9                   |
| SMARCA4 mutation        | SWI/SNF chromatin remodelling; BRD4/ SMARCA2 dependency        | 41 (6.5)   | 46.3 vs 37.5                 | 0.127      | 1.32 (0.85–2.05) | 7.3                    |
| NF1 mutation            | RAS GAP; MEK inhibition                                        | 58 (9.2)   | 36.2 vs 38.2                 | 0.800      | 1.01 (0.67–1.50) | 6.5                    |
| KRAS mutation           | MAPK oncogene; KRAS-G12C inhibitors                            | 147 (23.4) | 34.7 vs 39.1                 | 0.601      | 1.08 (0.82–1.42) | 21.2                   |
| CDKN2A copy-number loss | CDK4/6–RB control; CDK4/6 inhibitors                           | 414 (65.9) | 42.0 vs 30.4                 | 0.065      | 1.21 (0.94–1.55) | 4.2                    |

|                               |                                     |          |               |       |                  |      |
|-------------------------------|-------------------------------------|----------|---------------|-------|------------------|------|
| KRAS + SMARCA4 (co-mutation)  | Convergent MAPK + SWI/SNF loss      | 8 (1.3)  | 62.5 vs 37.7  | 0.021 | 2.20 (0.91–5.31) | n.a. |
| STK11 + KEAP1 (co-mutation)   | Convergent metabolic–oxidative axis | 23 (3.7) | 47.8 vs 37.7  | 0.106 | 1.61 (0.91–2.87) | n.a. |
| KRAS + TP53 (co-mutation)     | MAPK with p53 loss                  | 51 (8.1) | 37.3 vs 38.1  | 0.122 | 1.41 (0.91–2.19) | n.a. |
| Panel score + clinical (LUAD) | Cumulative tumour-suppressor burden | 628      | C-index 0.602 | n.a.  | 1.16 (1.02–1.32) | n.a. |
| Panel score + clinical (LUSC) | Same panel, squamous histology      | 497      | C-index 0.587 | n.a.  | 0.94 (0.78–1.13) | n.a. |

### **High-risk LUAD displays cancer-testis antigen, metabolic and matrix-remodelling programmes**

Differential analysis between high-risk and low-risk LUAD revealed a coherent, multi-layer signature (Table 3; Figure 3b,c). At the transcriptional level the most strongly up-regulated genes were cancer-testis antigens (MAGEA3, MAGEA6, MAGEA2, MAGEA12 and PRAME;  $p < 10^{-14}$ ) together with aldo-keto reductase and cytochrome P450 family members implicating lipid and steroid metabolism, while differentiation- and lineage-associated transcripts (HHIP, surfactant genes) were down-reg-

ulated. Reactome over-representation analysis against 2,855 pathways confirmed enrichment of lipid metabolism ( $p = 0.011$ ), collagen degradation ( $p = 0.032$ ), steroid metabolism ( $p = 0.032$ ) and broader extracellular-matrix programmes. Protein-level analysis was concordant, with elevated PAI-1, G6PD, Chk1 and Cyclin B1 indicating active glucose metabolism and cell-cycle progression, and significantly reduced KEAP1 protein ( $p = 0.001$ ) consistent with de-repression of the NRF2 programme, alongside reduced TTF-1 lineage marker (full lists in Supplementary Tables S6 and S7).

Table 3. Integrated multi-omic molecular signature distinguishing high-risk from low-risk LUAD, organised by data layer. Standardised effect is the high-minus-low risk group difference (transcript/protein layers); overlap is the number of signature genes in the pathway (pathway layer). Up, higher in high-risk tumours; down, lower in high-risk tumours.

| Data layer    | Feature(s)                       | Direction | Standardised effect / overlap | p-value               | Biological programme                   |
|---------------|----------------------------------|-----------|-------------------------------|-----------------------|----------------------------------------|
| Transcriptome | MAGEA3, MAGEA6, MAGEA2, MAGEA12  | Up        | +0.70 to +0.79                | $<10^{-14}$           | Cancer-testis antigen / immune evasion |
| Transcriptome | PRAME                            | Up        | +0.72                         | $2.4 \times 10^{-17}$ | Cancer-testis antigen                  |
| Transcriptome | AKR1B15, AKR1C1, AKR1C2, CYP24A1 | Up        | +0.74 (max)                   | $1.6 \times 10^{-17}$ | Lipid / steroid metabolism             |

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

|                    |                                         |      |                |                       |                                   |
|--------------------|-----------------------------------------|------|----------------|-----------------------|-----------------------------------|
| Transcriptome      | HHIP, SFTPA1, SFT-PA2                   | Down | -0.66 to -0.68 | <10 <sup>-15</sup>    | Loss of differentiation / lineage |
| Proteome (RPPA)    | PAI-1, G6PD                             | Up   | +0.37 to +0.44 | <10 <sup>-5</sup>     | Glucose metabolism / EMT          |
| Proteome (RPPA)    | Chk1, Cyclin B1                         | Up   | +0.32 to +0.34 | <10 <sup>-4</sup>     | Cell-cycle progression            |
| Proteome (RPPA)    | KEAP1, TTF-1                            | Down | -0.27          | ≤1.4×10 <sup>-3</sup> | NRF2 de-repression / lineage loss |
| Pathway (Reactome) | Metabolism of lipids                    | Up   | 10 genes       | 0.011                 | Metabolic reprogramming           |
| Pathway (Reactome) | Collagen degradation                    | Up   | 5 genes        | 0.032                 | Extracellular-matrix remodelling  |
| Pathway (Reactome) | Metabolism of steroids                  | Up   | 5 genes        | 0.032                 | Metabolic reprogramming           |
| Pathway (Reactome) | Degradation of the extracellular matrix | Up   | 6 genes        | 0.059                 | Invasion / matrix remodelling     |

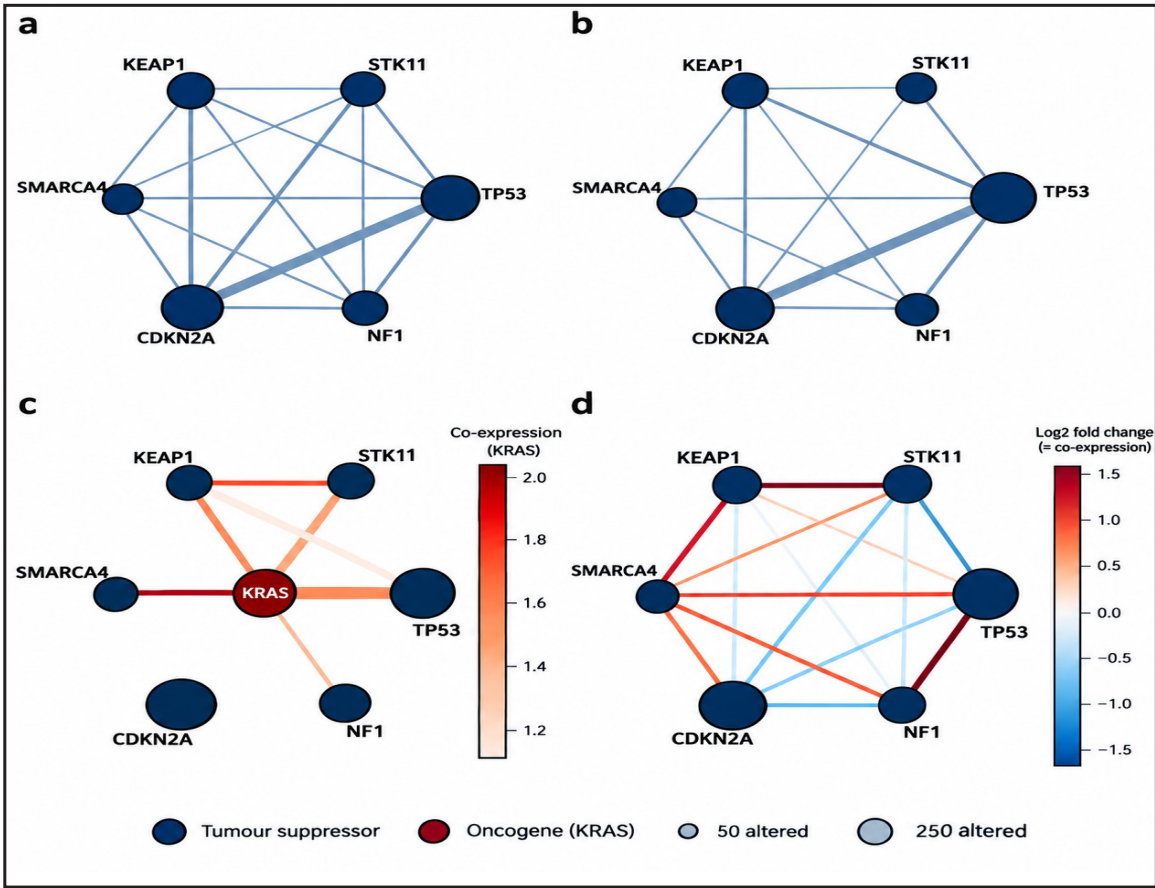


Figure.3: Associations signature

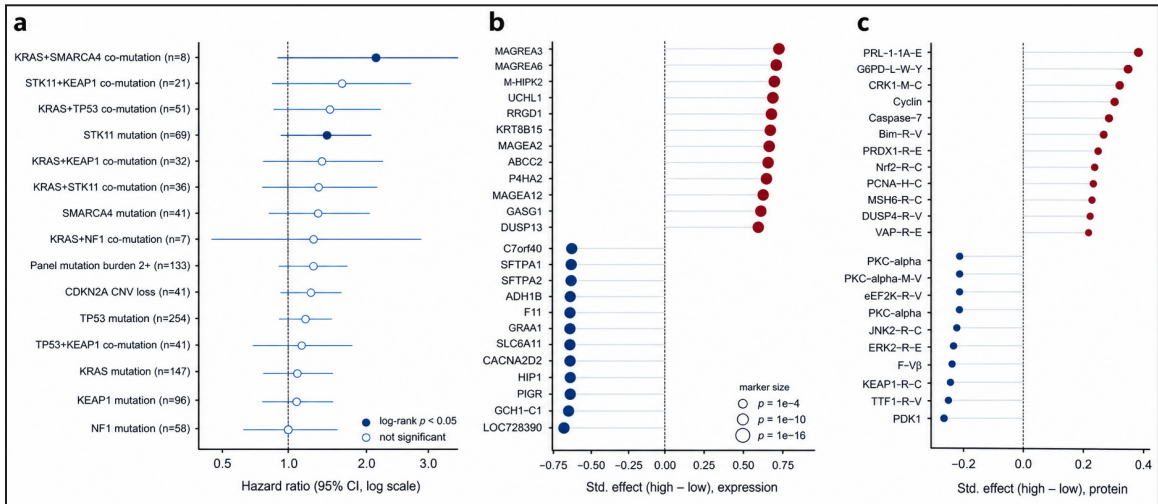


Figure.4: Networks

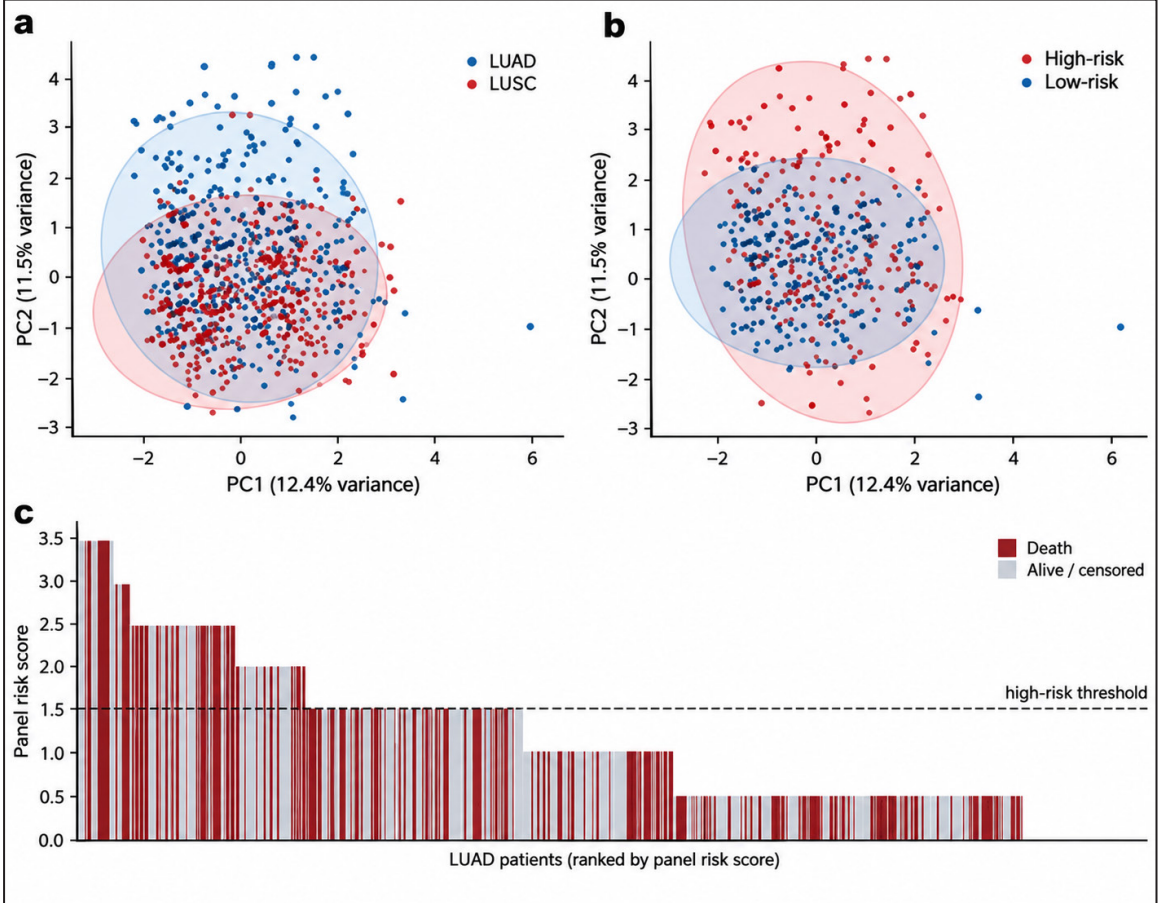


Figure.5: PCA waterfall

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

This study set out to test, transparently, whether multi-omic integration improves survival prediction in lung adenocarcinoma, and then to ask what biology those same data reveal. The central prediction result is unambiguous and, we argue, instructive: a clinical-only random forest using six variables (C-index 0.633) outperformed a multimodal deep survival network trained on 1,902 features (0.555), and adding omic modalities consistently degraded performance. Rather than concealing this outcome, we diagnosed it and then extracted reproducible biology from the same dataset.

Three convergent lines of evidence indicate that the limitation is one of statistical regime rather than implementation. First, modality ablation showed that removing the highest-dimensional blocks (expression, mutation) improved discrimination, the signature of noise accumulation. Second, the selected-omics sweep showed monotonic degradation as feature count increased, with the best configuration using only five features per modality. Third, nested cross-validation eliminated even that modest advantage, exposing it as selection optimism. With 628 patients and 1,902 features, the sample-to-feature ratio is far below what stable high-dimensional survival modelling requires (9,10,11). Clinical variables, by contrast, are compact and act as direct proxies for survival, encoding age, stage and performance status that already capture much of the prognostic variance. This result echoes broader observations that classical ensembles remain competitive with, or superior to, deep networks on tabular biomedical data of limited size, and it underscores the importance of always reporting a fair clinical baseline in multi-omic survival studies.

Repurposing the omic data for interpretation yielded a coherent biological narrative. The six-gene panel added independent prognostic value in LUAD but reversed direction and lost significance in LUSC, a clear statement that this axis is histology-specific. The strongest single-gene contributor was STK11, encoding the

LKB1 kinase, a master regulator of the AMPK energy-sensing pathway whose loss releases anabolic metabolism and promotes aggressive lung tumour behaviour (24,25,26). KEAP1, the second axis member, is the principal negative regulator of NRF2; its loss activates an antioxidant and detoxification programme that supports tumour survival under metabolic stress and is associated with therapy resistance (27,28). The observation of reduced KEAP1 protein in high-risk tumours provides orthogonal, protein-level support for NRF2 pathway activation in this group. SMARCA4 loss disrupts the SWI/SNF chromatin-remodelling complex, which can de-repress normally silenced loci, plausibly contributing to the cancer-testis antigen activation observed here (29).

The functional reading of the high-risk signature reinforces this axis. Up-regulation of lipid and steroid metabolism is the expected downstream consequence of LKB1/AMPK loss, while coordinated extracellular-matrix degradation through matrix metalloproteinases (MMP1, MMP11, MMP12) and elevated PAI-1 protein point to enhanced invasive capacity (30,31). The simultaneous activation of cancer-testis antigens (MAGEA family, PRAME), among the most statistically robust findings in the dataset, links aggressive biology to a recognised mechanism of immune modulation and to a class of established immunotherapy targets (32). The down-regulation of the lineage transcription factor TTF-1 at the protein level is consistent with the loss of differentiation that accompanies aggressive adenocarcinoma (33).

A noteworthy secondary finding is that KRAS mutation alone carried no prognostic signal, whereas KRAS combined with SMARCA4 defined the highest-risk subgroup observed. Although the co-mutation group was small and must be regarded as hypothesis-generating, this pattern is consistent with the established view that co-occurring alterations, rather than KRAS status in isolation, define clinically and biologically distinct subsets of KRAS-mutant lung adenocarcinoma (26,28). The same logic

applies to the STK11-plus-KEAP1 combination, which trended toward higher risk and unites two arms of the metabolic axis.

While the panel does not improve prediction beyond clinical variables enough to warrant standalone clinical use, the biology it organises is therapeutically relevant. STK11-deficient tumours are candidates for strategies that exploit AMPK pathway dependency; KEAP1-mutant or NRF2-activated tumours suggest vulnerability to glutaminolysis or oxidative-stress-directed approaches; SMARCA4-deficient tumours have been linked to dependency on residual chromatin-remodelling activity; and the strong cancer-testis antigen expression in high-risk LUAD nominates this subgroup for antigen-directed immunotherapy. The histology specificity of the axis argues against one-size-fits-all pan-lung biomarkers and in favour of subtype-resolved molecular stratification.

Several limitations should be acknowledged. The discovery and primary validation were performed within TCGA cohorts; prospective validation in independent LUAD series with clean survival fields is needed. The LUSC result, while informative, restricts generalisability to adenocarcinoma. The high-risk signature is correlative and requires functional confirmation, for example by perturbation of STK11, KEAP1 or SMARCA4 in LUAD models with metabolic and invasion read-outs. External validation in cBioPortal cohorts established alteration recurrence but not survival replication, because harmonised time-event fields were not consistently available. Finally, several co-mutation subgroups were small, and the deep survival model, although carefully implemented, was not exhaustively architecture-optimised; nonetheless, the consistency of the clinical-baseline advantage across independent analyses makes it unlikely that a different architecture would overturn the central conclusion in this sample-size regime.

## Conclusions

We performed a transparent, seven-modality survival analysis of lung adeno-

carcinoma and found that a compact clinical model outperformed multimodal deep learning, with omic integration degrading prediction in a sample-to-feature regime unfavourable to high-dimensional modelling (628 patients, 1,902 features; clinical C-index 0.633 versus deep-model 0.555). Rather than discard the omic data, we used them to define a LUAD-specific tumour-suppressor survival axis. A six-gene panel (KEAP1, STK11, TP53, SMARCA4, NF1, CDKN2A) added independent prognostic information in LUAD (panel-plus-clinical C-index 0.602; HR 1.16,  $p = 0.025$ ) but not in LUSC, with STK11 as the strongest single-gene signal ( $p = 0.044$ ). High-risk tumours displayed coordinated cancer-testis antigen activation, lipid and steroid metabolic reprogramming, and extracellular-matrix remodelling, with protein-level reduction of KEAP1 supporting NRF2 activation. Panel-gene recurrence was confirmed across seven independent lung cohorts. The work demonstrates that multi-omic data can deliver coherent, mechanistically interpretable biology even when they do not improve prediction, and it provides a histology-specific framework for understanding aggressive LUAD. Prospective validation and functional dissection of the STK11-KEAP1-SMARCA4 axis are the logical next steps.

## Declarations

Declaration of competing interests. The authors declare no competing interests.

## Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Data availability

All primary data are publicly available: TCGA-LUAD and TCGA-LUSC matrices via the UCSC Xena platform; external cohorts via cBioPortal; pathway annotations via Reactome; and protein-interaction data via STRING. Processed matrices and analysis outputs can be regenerated from the documented workflow.

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

# Author contributions

Conceptualization, methodology, formal analysis, investigation, writing — original draft, and writing — review and editing. All authors have approved the final article.

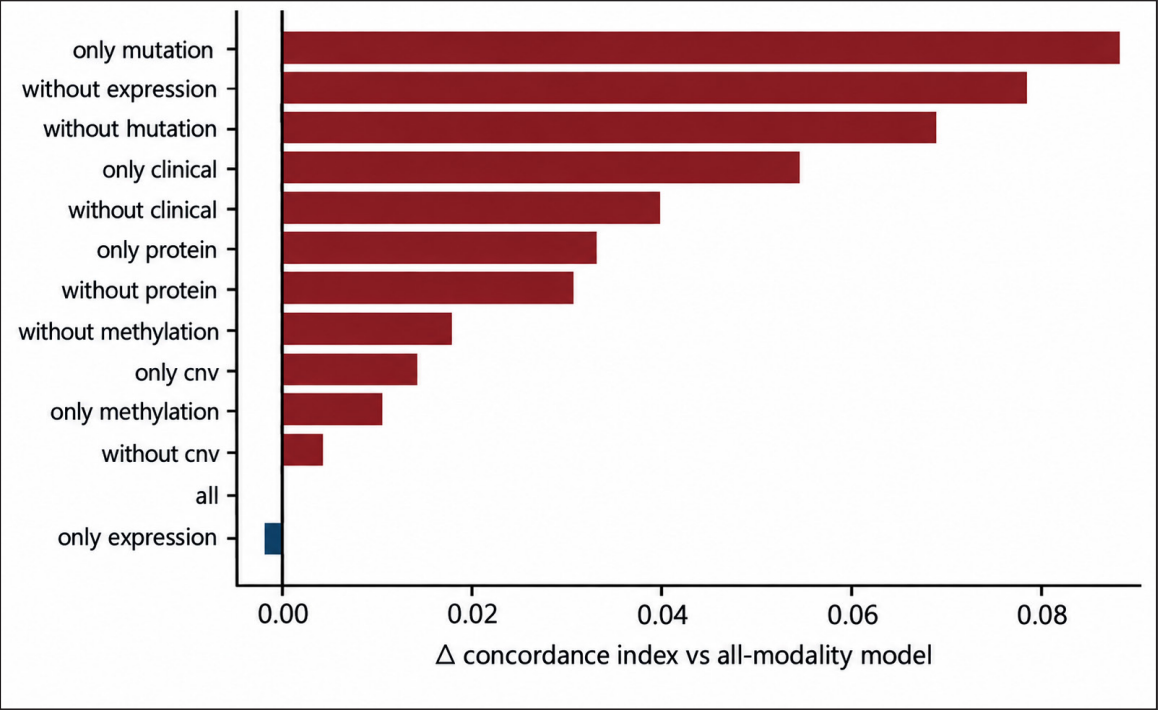
# References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209-249.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*. 2023;73(1):17-48.
3. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. *Journal of Thoracic Oncology*. 2016;11(1):39-51.
4. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-550.
5. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*. 2020;38(6):675-678.
6. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*. 2015;13:8-17.
7. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancreatic prognosis prediction. *Bioinformatics*. 2019;35(14):i446-i454.
8. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*. 2018;24(6):1248-1259.
9. Johnstone IM, Titterton DM. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A*. 2009;367(1906):4237-4253.
10. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*. 2021;11(1):13505.
11. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics*. 2008;2(3):841-860.
12. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*. 2012;2(5):401-404.
13. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*. 2013;6(269):pl1.
14. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*. 2020;48(D1):D498-D503.
15. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 2019;47(D1):D607-D613.
16. Kingma DP, Welling M. Auto-encoding variational Bayes. In: *Proceedings of the*

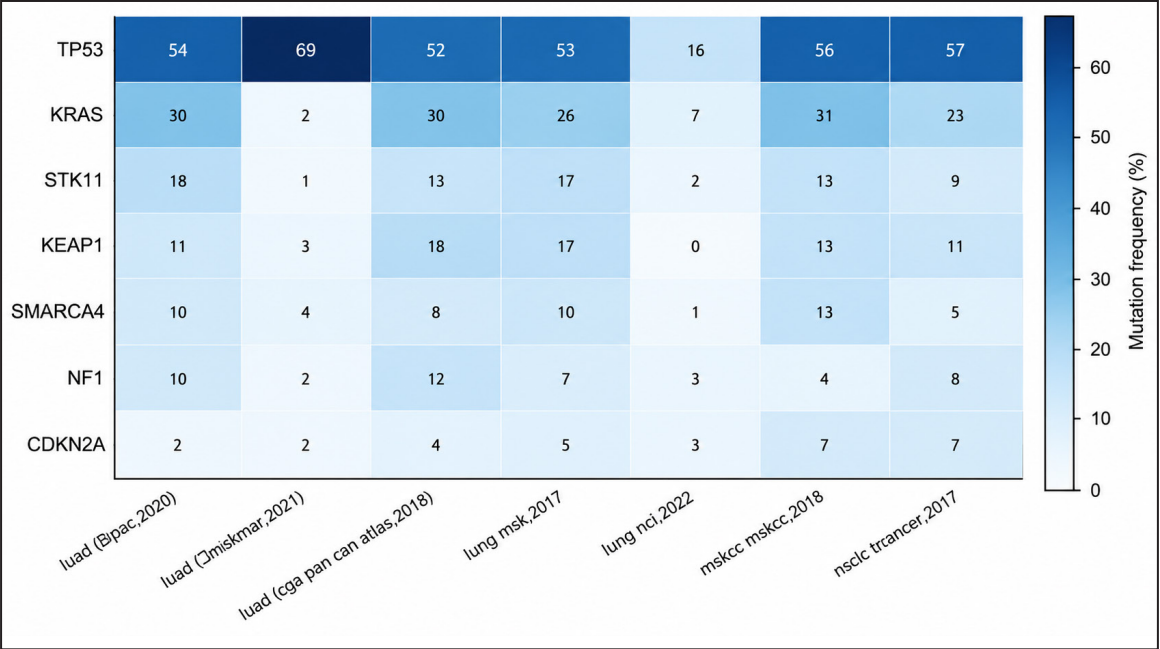
- International Conference on Learning Representations (ICLR). 2014.
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30:5998-6008.
18. Lee C, Zame WR, Yoon J, van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1):2314-2321.
19. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543-2546.
20. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*. 2011;39(5):1-13.
21. Davidson-Pilon C. lifelines: survival analysis in Python. *Journal of Open Source Software*. 2019;4(40):1317.
22. Pölsterl S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*. 2020;21(212):1-6.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*. 1995;57(1):289-300.
24. Ji H, Ramsey MR, Hayes DN, Fan C, McNamara K, Kozlowski P, et al. LKB1 modulates lung cancer differentiation and metastasis. *Nature*. 2007;448(7155):807-810.
25. Shackelford DB, Shaw RJ. The LKB1-AMPK pathway: metabolism and growth control in tumour suppression. *Nature Reviews Cancer*. 2009;9(8):563-575.
26. Skoulidis F, Byers LA, Diao L, Papadimitrakopoulou VA, Tong P, Izzo J, et al. Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. *Cancer Discovery*. 2015;5(8):860-877.
27. Singh A, Misra V, Thimmulappa RK, Lee H, Ames S, Hoque MO, et al. Dysfunctional KEAP1-NRF2 interaction in non-small-cell lung cancer. *PLoS Medicine*. 2006;3(10):e420.
28. Skoulidis F, Heymach JV. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nature Reviews Cancer*. 2019;19(9):495-509.
29. Romero OA, Torres-Diz M, Pros E, Savola S, Gomez A, Moran S, et al. MAX inactivation in small cell lung cancer disrupts the MYC-SWI/SNF programs and is synthetic lethal with BRG1. *Cancer Discovery*. 2014;4(3):292-303.
30. Egeblad M, Werb Z. New functions for the matrix metalloproteinases in cancer progression. *Nature Reviews Cancer*. 2002;2(3):161-174.
31. Pavlova NN, Thompson CB. The emerging hallmarks of cancer metabolism. *Cell Metabolism*. 2016;23(1):27-47.
32. Simpson AJG, Caballero OL, Jungbluth A, Chen YT, Old LJ. Cancer/testis antigens, gametogenesis and cancer. *Nature Reviews Cancer*. 2005;5(8):615-625.
33. Winslow MM, Dayton TL, Verhaak RGW, Kim-Kiselak C, Snyder EL, Feldser DM, et al. Suppression of lung adenocarcinoma progression by Nkx2-1. *Nature*. 2011;473(7345):101-104.
34. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discovery*. 2022;12(1):31-46.

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

Supplementary Information:



Supplementary figure.1: Modality ablation



Supplementary figure.2: External frequencies

Supplementary Table.1: Cohort characteristics

| Characteristic               | LUAD  | LUSC  |
|------------------------------|-------|-------|
| Patients                     | 628   | 497   |
| Deaths (events)              | 239   | 215   |
| Event fraction               | 0.381 | 0.433 |
| Mean overall survival (days) | 932   | 975   |
| Expression features          | 500   | 500   |
| Methylation features         | 500   | 500   |
| Mutation features            | 419   | 419   |
| Copy-number features         | 200   | 200   |
| Protein (RPPA) features      | 277   | 277   |
| Clinical features            | 6     | 6     |
| Total features               | 1902  | 1902  |
| Missing values               | 0     | 0     |

Supplementary Table.2: Model benchmark

| Feature set                  | Model                      | Features | Mean C-index | SD C-index |
|------------------------------|----------------------------|----------|--------------|------------|
| clinical                     | random_forest_duration     | 6        | 0.6331       | 0.0696     |
| clinical                     | gradient_boosting_duration | 6        | 0.6017       | 0.0539     |
| clinical                     | cox_ridge                  | 6        | 0.5748       | 0.0514     |
| expression_mutation_clinical | gradient_boosting_duration | 925      | 0.5701       | 0.0579     |
| expression_mutation_clinical | random_forest_duration     | 925      | 0.5667       | 0.0791     |
| all_omics_clinical           | random_forest_duration     | 1902     | 0.5652       | 0.0968     |
| all_omics_clinical           | gradient_boosting_duration | 1902     | 0.5439       | 0.0621     |
| mutation                     | random_forest_duration     | 419      | 0.5124       | 0.0658     |
| mutation                     | gradient_boosting_duration | 419      | 0.5077       | 0.0683     |
| expression_mutation_clinical | lasso_duration             | 925      | 0.4968       | 0.0937     |
| mutation                     | lasso_duration             | 419      | 0.4943       | 0.0106     |
| all_omics_clinical           | lasso_duration             | 1902     | 0.4902       | 0.04       |
| clinical                     | lasso_duration             | 6        | 0.4891       | 0.015      |

Supplementary table.3: Top DE genes

| gene      | mean_high_risk | mean_low_risk | standardized_effect | ttest_p                | mannwhitney_p          |
|-----------|----------------|---------------|---------------------|------------------------|------------------------|
| MAGEA3    | 0.45           | -0.32         | 0.794               | 6.605480249851526e-19  | 4.8862633774196535e-14 |
| LOC723809 | -0.421         | 0.3           | -0.776              | 1.79605064012914e-20   | 2.3191162389878642e-18 |
| MAGEA6    | 0.439          | -0.312        | 0.772               | 4.227392133645647e-18  | 2.09313943673288e-13   |
| HHIPL2    | 0.423          | -0.301        | 0.761               | 9.88620522126813e-19   | 7.335621637088481e-17  |
| UCHL1     | 0.425          | -0.302        | 0.761               | 1.7290091055101376e-18 | 3.027681219652319e-17  |

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

|           |        |        |        |                        |                        |
|-----------|--------|--------|--------|------------------------|------------------------|
| NR0B1     | 0.43   | -0.306 | 0.745  | 1.0838937333786955e-16 | 1.5935996012378384e-12 |
| AKR1B15   | 0.417  | -0.296 | 0.74   | 1.5903437322671064e-17 | 2.0505070201827651e-16 |
| MAGEA2    | 0.425  | -0.302 | 0.739  | 1.383323519827786e-16  | 1.3840009079575729e-15 |
| ABCC2     | 0.424  | -0.302 | 0.733  | 3.5824879986272334e-16 | 1.0537113542818943e-10 |
| PRAME     | 0.401  | -0.285 | 0.721  | 2.421511686244885e-17  | 6.425020190631727e-16  |
| GGTLC1    | -0.394 | 0.28   | -0.703 | 1.861900708446712e-16  | 4.416328540076946e-17  |
| MAGEA12   | 0.403  | -0.287 | 0.697  | 4.1785473600111305e-15 | 3.599100878437511e-11  |
| PIGR      | -0.384 | 0.273  | -0.683 | 1.0640986925153709e-15 | 5.218154412493998e-17  |
| CSAG1     | 0.397  | -0.282 | 0.683  | 1.5544535195708212e-14 | 3.248242563813619e-09  |
| HHIP      | -0.377 | 0.268  | -0.682 | 2.863096299079255e-16  | 7.519397769912596e-15  |
| CACNA2D2  | -0.38  | 0.27   | -0.679 | 1.0140644563210592e-15 | 2.714105321239592e-15  |
| SLC5A1    | -0.379 | 0.27   | -0.678 | 1.1309640027672174e-15 | 1.2957115933708236e-16 |
| GRIA1     | -0.367 | 0.261  | -0.676 | 1.0945016960774227e-16 | 2.690880829790687e-13  |
| F11       | -0.37  | 0.263  | -0.676 | 2.46592810876543e-16   | 1.0651928473249866e-15 |
| ADH1B     | -0.372 | 0.265  | -0.671 | 8.610155024497783e-16  | 1.989180220144998e-15  |
| SFTPA2    | -0.374 | 0.266  | -0.668 | 2.3388324561562576e-15 | 1.2356382962158742e-16 |
| SFTPA1    | -0.372 | 0.265  | -0.664 | 3.921580857220225e-15  | 1.1792293430000863e-16 |
| DUSP13    | 0.375  | -0.266 | 0.662  | 8.759925802044512e-15  | 6.318966075218035e-14  |
| C2orf40   | -0.364 | 0.259  | -0.661 | 1.0395540070445462e-15 | 6.256962585852664e-14  |
| MAGEA4    | 0.388  | -0.276 | 0.658  | 2.3354010240429773e-13 | 1.1151310723347267e-14 |
| GABRA3    | 0.379  | -0.269 | 0.657  | 4.9748349022086045e-14 | 5.660921260641771e-07  |
| PAH       | 0.376  | -0.267 | 0.65   | 9.429952072025421e-14  | 3.6280256576759484e-10 |
| CD300LG   | -0.35  | 0.249  | -0.65  | 4.285977448683899e-16  | 4.45630945339337e-12   |
| KRT81     | 0.374  | -0.266 | 0.649  | 9.05884931954801e-14   | 9.013597770187072e-15  |
| CLDN18    | -0.358 | 0.255  | -0.648 | 4.9544265750807535e-15 | 7.969609130617505e-13  |
| IRX1      | -0.358 | 0.255  | -0.647 | 5.543537566379646e-15  | 2.4180913619099162e-14 |
| C19orf59  | -0.353 | 0.251  | -0.643 | 3.8399649975875364e-15 | 8.506085375989404e-13  |
| MAGEA1    | 0.374  | -0.266 | 0.642  | 3.039652134596144e-13  | 8.152971373497439e-17  |
| FAM83A    | 0.352  | -0.25  | 0.642  | 3.891817813519603e-15  | 3.1527064352178044e-14 |
| RPL13AP17 | -0.349 | 0.248  | -0.64  | 2.788752429056982e-15  | 7.721462358389886e-12  |
| YBX2      | 0.358  | -0.255 | 0.639  | 2.8088088563468956e-14 | 1.0249731395388631e-14 |
| SLC6A20   | -0.359 | 0.255  | -0.637 | 4.6667486905014814e-14 | 9.076960916839376e-15  |
| PPP1R1B   | -0.356 | 0.254  | -0.623 | 3.508066504388091e-13  | 2.32921277751319e-12   |
| CYP4B1    | -0.35  | 0.249  | -0.622 | 1.40625195949734e-13   | 4.1787025364470674e-14 |
| AGER      | -0.34  | 0.242  | -0.622 | 1.617977771889694e-14  | 2.4482516014501243e-10 |
| KIAA0408  | -0.347 | 0.246  | -0.621 | 8.124586886506385e-14  | 2.570004484832994e-13  |
| VSIG2     | -0.347 | 0.247  | -0.616 | 2.268870719556934e-13  | 2.5627754736743226e-15 |
| PEBP4     | -0.345 | 0.246  | -0.613 | 2.6470321702783133e-13 | 1.7656314778291549e-12 |

|         |        |        |        |                        |                        |
|---------|--------|--------|--------|------------------------|------------------------|
| CSAG3   | 0.359  | -0.255 | 0.612  | 3.651003544630216e-12  | 7.981636008480564e-05  |
| INSL4   | 0.358  | -0.254 | 0.611  | 3.302776238264804e-12  | 2.166665284686342e-12  |
| INHA    | 0.349  | -0.248 | 0.604  | 2.7646801948472133e-12 | 1.394874077353947e-10  |
| PRG4    | -0.335 | 0.238  | -0.602 | 2.818486004481369e-13  | 1.981955181570849e-12  |
| FIGF    | -0.334 | 0.237  | -0.602 | 2.1404570010961218e-13 | 3.0872835538042764e-11 |
| RETN    | -0.333 | 0.237  | -0.601 | 2.291963548204518e-13  | 6.551937811738462e-12  |
| MAGEA9B | 0.352  | -0.25  | 0.596  | 1.5572543254904758e-11 | 1.144317659292765e-13  |

Supplementary table.4: Reactome enrichment

| pathway                                      | over-<br>lap_n | pathway_<br>genes_in_<br>background | p_value | fdr_bh | overlap_genes                                                                                                                |
|----------------------------------------------|----------------|-------------------------------------|---------|--------|------------------------------------------------------------------------------------------------------------------------------|
| Metabolism of lipids                         | 10             | 24                                  | 0.011   | 1.0    | AKR1B15,AKR1C1,AKR1C2,CYP24A1,CYP-4F11,CYP4F3,GPX2,SLC44A5,SLCO1B3,UGT1A9                                                    |
| Collagen degradation                         | 5              | 10                                  | 0.0319  | 1.0    | COL11A1,COL2A1,MMP1,MMP11,MMP12                                                                                              |
| Metabolism of steroids                       | 5              | 10                                  | 0.0319  | 1.0    | AKR1B15,AKR1C1,AKR1C2,CYP24A1,SLCO1B3                                                                                        |
| Metabolism                                   | 19             | 64                                  | 0.0331  | 0.794  | ABCC2,AKR1B10,AKR1B15,AKR1C1,AKR1C2,AKR7A3,BHMT2,CPS1,CYP24A1,CYP4F11,CYP4F3,GNG4,GPX2,HGD,PAH,SLC44A5,SLCO1B3,UGT1A6,UGT1A9 |
| Cytochrome P450 - arranged by substrate type | 3              | 5                                   | 0.0576  | 1.0    | CYP24A1,CYP4F11,CYP4F3                                                                                                       |
| Aspirin ADME                                 | 3              | 5                                   | 0.0576  | 0.922  | ABCC2,UGT1A6,UGT1A9                                                                                                          |
| Bile acid and bile salt metabolism           | 3              | 5                                   | 0.0576  | 0.79   | AKR1C1,AKR1C2,SLCO1B3                                                                                                        |
| Arachidonate metabolism                      | 3              | 5                                   | 0.0576  | 0.691  | CYP4F11,CYP4F3,GPX2                                                                                                          |
| Degradation of the extracellular matrix      | 6              | 15                                  | 0.0592  | 0.631  | COL11A1,COL2A1,MMP1,MMP11,MMP12,SPP1                                                                                         |
| Extracellular matrix organization            | 7              | 19                                  | 0.065   | 0.624  | COL11A1,COL2A1,FGB,MMP1,MMP11,MMP12,SPP1                                                                                     |
| Integrin cell surface interactions           | 3              | 6                                   | 0.0986  | 0.861  | COL2A1,FGB,SPP1                                                                                                              |
| Fatty acid metabolism                        | 3              | 6                                   | 0.0986  | 0.789  | CYP4F11,CYP4F3,GPX2                                                                                                          |

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails

|                                                                           |    |    |        |       |                                                                               |
|---------------------------------------------------------------------------|----|----|--------|-------|-------------------------------------------------------------------------------|
| Surfactant metabolism                                                     | 7  | 7  | 0.0    | 0.001 | DMBT1,NAPSA,SFTPA1,SFTPA2,SFTPB,S-FTPC,SFTPD                                  |
| Defective CSF2RB causes SMDP5                                             | 5  | 5  | 0.0003 | 0.013 | SFTPA1,SFTPA2,SFTPB,SFTPC,SFTPD                                               |
| Defective CSF2RA causes SMDP4                                             | 5  | 5  | 0.0003 | 0.009 | SFTPA1,SFTPA2,SFTPB,SFTPC,SFTPD                                               |
| Diseases associated with surfactant metabolism                            | 5  | 5  | 0.0003 | 0.007 | SFTPA1,SFTPA2,SFTPB,SFTPC,SFTPD                                               |
| Vesicle-mediated transport                                                | 8  | 20 | 0.03   | 0.522 | CFTR,GRIA1,HBA1,HBB,LRP2,SCARA5,SCGB3A2,TUBA4B                                |
| Disease                                                                   | 13 | 42 | 0.0555 | 0.805 | C4BPA,CFTR,F11,HBA1,HBB,SFTPA1,SFTPA2,SFTPB,SFTPC,SFTPD,SLC5A1,SLC6A20,TUBA4B |
| Binding and Uptake of Ligands by Scavenger Receptors                      | 4  | 8  | 0.0555 | 0.69  | HBA1,HBB,SCARA5,SCGB3A2                                                       |
| Toll Like Receptor TLR1:TLR2 Cascade                                      | 4  | 8  | 0.0555 | 0.604 | AGER,SFTPA1,SFTPA2,SFTPD                                                      |
| Toll Like Receptor 2 (TLR2) Cascade                                       | 4  | 8  | 0.0555 | 0.537 | AGER,SFTPA1,SFTPA2,SFTPD                                                      |
| Immuno-regulatory interactions between a Lymphoid and a non-Lymphoid cell | 3  | 5  | 0.0576 | 0.501 | CD1A,CD300LG,SFTPD                                                            |
| Toll Like Receptor 4 (TLR4) Cascade                                       | 4  | 9  | 0.0848 | 0.671 | AGER,SFTPA1,SFTPA2,SFTPD                                                      |
| Toll-like Receptor Cascades                                               | 4  | 9  | 0.0848 | 0.615 | AGER,SFTPA1,SFTPA2,SFTPD                                                      |
| Infectious disease                                                        | 5  | 13 | 0.0977 | 0.654 | C4BPA,HBA1,HBB,SFTPD,TUBA4B                                                   |

|                                  |   |    |        |       |                                 |
|----------------------------------|---|----|--------|-------|---------------------------------|
| Adaptive Immune System           | 5 | 13 | 0.0977 | 0.607 | CD1A,CD207,CD300LG,SFTPD,TUBA4B |
| Complement cascade               | 3 | 6  | 0.0986 | 0.572 | C4BPA,C6,CPB2                   |
| Regulation of Complement cascade | 3 | 6  | 0.0986 | 0.536 | C4BPA,C6,CPB2                   |

Supplementary table.8: Panel COX

| cohort | model               | concordance_index | panel_hr | panel_p | n   | events |
|--------|---------------------|-------------------|----------|---------|-----|--------|
| LUAD   | panel_only          | 0.5478            | 1.1416   | 0.042   | 628 | 239    |
| LUAD   | clinical_only       | 0.5952            |          |         | 628 | 239    |
| LUAD   | panel_plus_clinical | 0.6021            | 1.1611   | 0.0248  | 628 | 239    |
| LUSC   | panel_only          | 0.5006            | 0.9228   | 0.3816  | 497 | 215    |
| LUSC   | clinical_only       | 0.5876            |          |         | 497 | 215    |
| LUSC   | panel_plus_clinical | 0.5867            | 0.9366   | 0.4825  | 497 | 215    |

Multi-omic integration reveals a lung adenocarcinoma specific tumor suppressor survival axis where high dimensional prediction fails