

A genomically validated loss of chromosome Y transcriptional program is a modest prognostic signal independent of tumour stage in male lung adenocarcinoma

Shilei Qin

The First Affiliated Hospital of Guilin Medical University

Ling Yan

The First Affiliated Hospital of Guilin Medical University

Zanjin Wu

The First Affiliated Hospital of Guilin Medical University

Yang Wang

drwangyang123@126.com

The First Affiliated Hospital of Guilin Medical University

Research Article

Keywords: lung adenocarcinoma, loss of chromosome Y, prognostic signature, LASSO-Cox, tumour purity, copy number, TCGA, overall survival

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10310328/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

Background.

Somatic loss of chromosome Y (LOY) is the most common acquired mutation in the male genome and an emerging determinant of cancer aggressiveness and immune evasion, yet whether an LOY-associated transcriptional program constitutes a prognostic axis *independent of tumour stage and generic proliferation* in non-small-cell lung cancer has not been established. We asked whether LOY defines a genomically-anchored, stage-independent prognostic signal in male lung adenocarcinoma (LUAD).

Methods.

We analysed TCGA-LUAD bulk RNA-seq (STAR TPM; UCSC Xena) restricted to 231 male primary tumours with overall-survival (OS) follow-up (86 deaths; median follow-up 29.2 months). A male-specific-Y (MSY) transcriptional score classified tumours as LOY (lowest quartile, $n = 58$) or Y-intact ($n = 173$). The MSY-based call was benchmarked against genomic chromosome-Y copy number (segment-level SNP6) and against ABSOLUTE tumour purity. Differentially expressed genes (DEGs) were screened by univariate Cox and reduced by L1-penalised (LASSO) Cox; a per-tumour risk score was evaluated by repeated 5-fold cross-validation (CV) with the whole supervised chain re-derived inside each fold. The risk score was tested for independence from TNM stage, age and smoking (multivariable and stage-stratified Cox), from a 19-gene proliferation meta-score, and for gene-selection stability. External validation used the male subset of GSE31210.

Results.

The transcriptional LOY call agreed substantially with genomic Y loss (Cohen's $\kappa = 0.72$; MSY score vs chrY copy number Spearman $\rho = 0.84$, $p = 2 \times 10^{-61}$), confirming that low MSY expression reflects genuine genomic LOY rather than an artefact of tumour purity (purity did not differ between groups, $p = 0.33$; MSY score was uncorrelated with purity, $\rho = -0.02$). DEGs were robust to a non-parametric test (370/371 by Wilcoxon) and to purity adjustment (371/371 retained). LASSO-Cox yielded a 16-gene signature, although only 5 of the 16 genes were selected in $\geq 50\%$ of CV folds (C6, PDE4D, CRTAC1, C4BPA and one lncRNA), so the specific gene membership is unstable. Cross-validated discrimination was modest but reproducible (out-of-fold [OOF] C-index 0.62, 95% CI 0.54–0.69; apparent C 0.73, 95% CI 0.67–0.79) and robust to the L1 penalty (OOF C 0.55–0.62 across λ ; nested-CV λ selection 0.55). Using the honest cross-validated (out-of-fold, OOF) risk score, higher risk remained associated with OS after adjustment for stage and age (HR 1.34 per SD, 95% CI 1.06–1.68, $p = 0.013$), but the incremental discrimination over a stage + age model was very small (Δ C-index + 0.01); stage-independence is therefore statistical but marginal. When stratified by the OOF score the signal was carried by stage II–IV (HR 1.39, $p = 0.015$) and was not significant within stage I (HR 0.94, $p = 0.79$). Importantly, and in contrast to the optimistic in-sample result, the OOF risk score was attenuated to non-significance after adjustment for a 19-gene

proliferation meta-score (HR 1.20, 95% CI 0.95–1.51, $p = 0.13$; with stage + age, HR 1.25, $p = 0.08$): the LOY prognostic signal is entangled with, and not cleanly separable from, generic proliferation. The signature reproduced in the male subset of GSE31210 ($n = 105$, 19 events; log-rank $p = 0.048$; 3-/5-year AUC 0.68/0.66), attenuated relative to the mixed-sex cohort as expected for a smaller, event-poor male subset.

Conclusion.

In male LUAD, LOY is a genomically-validated transcriptional signal that is statistically but only marginally independent of TNM stage (small incremental discrimination) and that is not cleanly separable from generic proliferation once honest cross-validated risk is used. The individual 16-gene model is only partly reproducible and should be regarded as one realisation of an LOY-anchored risk signal rather than a locked biomarker; prospective and larger male-specific validation, and clearer separation from proliferation, are required before clinical use.

1. Introduction

Lung cancer remains the leading cause of cancer death worldwide, and non-small-cell lung cancer (NSCLC) – predominantly lung adenocarcinoma (LUAD) – accounts for most cases. Outcome heterogeneity within stage-matched LUAD is large, motivating transcriptome-derived prognostic signatures that could refine risk beyond TNM staging; however, the field is saturated with signatures whose apparent performance is optimistically biased and whose added value over stage and generic proliferation is rarely demonstrated. A credible new prognostic claim must therefore clear three bars simultaneously: honest (cross-validated) performance, independence from tumour stage, and biological specificity beyond a proliferation proxy.

Somatic mosaic loss of chromosome Y (LOY) is the single most common acquired genetic alteration in the ageing male genome and has been linked to shorter survival and higher cancer risk in population cohorts [5], to tobacco exposure [6], and to a defined germline genetic architecture [7]. Mechanistically, LOY in tumour cells can drive aggressiveness and adaptive-immune evasion, as shown for bladder cancer where Y-loss remodels the T-cell compartment [8]; because chromosome Y carries dosage-sensitive regulators (e.g. the *KDM5D/UTY* histone-demethylase pair and the *DDX3Y/EIF1AY* translational machinery), its loss reshapes the tumour transcriptome and may carry male-specific prognostic information. Sex is increasingly recognised as a biological axis in lung cancer, including sex differences in immunotherapy benefit [9], making a male-specific, Y-anchored prognostic readout mechanistically plausible.

A recent data-mining study in *Computers in Biology and Medicine* [1] showed in papillary renal cell carcinoma that integrating bulk and single-cell RNA-seq around a Y-chromosome-status phenotype yields prognostic signature genes. Building on that analytical framework rather than merely transplanting it, we address the substantive scientific question it leaves open in NSCLC: **is LOY an independent, sex-**

specific prognostic axis in male LUAD, or is an apparent LOY signal merely a projection of stage, proliferation, or tumour purity? We (i) derive an MSY-based transcriptional LOY call and a LASSO-Cox risk score in TCGA-LUAD, (ii) validate the LOY call itself against genomic chromosome-Y copy number and against ABSOLUTE tumour purity, (iii) test the risk score for independence from stage, age, smoking and proliferation, (iv) quantify gene-selection stability and honest cross-validated performance, and (v) validate in an appropriately male-restricted external cohort.

2. Materials and Methods

2.1 Data acquisition

TCGA-LUAD bulk gene expression (STAR-aligned, TPM, $\log_2(\text{TPM} + 1)$; 60,660 genes $\times$ 589 samples), overall-survival (OS) data and clinical phenotype (sex, AJCC pathologic stage, age, pack-years) were downloaded from the UCSC Xena GDC hub (accessed 15 June 2026). Analyses were restricted to primary-tumour samples; intersecting with survival annotation gave 503 primary tumours with OS data (231 male, 272 female), and the 231 male tumours formed the analysis cohort. Chromosome-Y copy number was obtained from the segment-level SNP6 copy-number dataset (which retains chrY), gene-level copy number from GDC ASCAT3, and consensus ABSOLUTE tumour purity/ploidy from the GDC PanCanAtlas master ABSOLUTE table [10,11]. External validation used GEO series GSE31210 (Affymetrix HG-U133 Plus 2.0) [4].

2.2 LOY phenotyping and genomic validation

Expression of seven canonical male-specific-Y (MSY) genes (RPS4Y1, ZFY, USP9Y, DDX3Y, UTY, KDM5D, EIF1AY) was z-scored across male tumours and averaged into a per-tumour MSY (LOY) score. Tumours in the lowest quartile were classified as LOY ($n = 58$), the remainder as Y-intact ($n = 173$). To test whether this transcriptional call reflects genuine genomic LOY, we computed a length-weighted mean chromosome-Y copy-number $\log_2$ ratio per tumour from the unmasked SNP6 genomic-segment dataset (the standard masked/germline-filtered segment file removes chromosome Y, so the unmasked file was used; 230 of the 231 male tumours had chrY segment coverage) and (i) correlated it with the continuous MSY score (Spearman), and (ii) dichotomised it at its lowest quartile and cross-tabulated it against the transcriptional call, reporting percentage agreement and Cohen's κ . We further tested whether the MSY score is confounded by tumour content by correlating it with ABSOLUTE purity and comparing purity between LOY and Y-intact groups (Mann–Whitney).

2.3 Differential expression and prognostic screening

Genes were pre-filtered to expressed, variable features (mean $\log_2\text{TPM} > 1$, $\text{SD} > 0.1$) and chromosome-Y genes were removed to avoid circularity, leaving 17,674 testable genes. DEGs between LOY and Y-intact tumours were identified by Welch's t-test with Benjamini–Hochberg FDR [13]; significance required $\text{FDR} < 0.05$ and $|\log_2\text{FC}| > 0.585$, giving 371 DEGs. Because Welch's t-test on $\log_2\text{TPM}$ is a modelling choice, we confirmed the DEGs by a non-parametric Mann–Whitney test and by re-testing each DEG in a linear

model that includes ABSOLUTE purity as a covariate (expression ~ group + purity). Each DEG was screened by univariate Cox regression; 107 reached unadjusted $p < 0.05$ and formed the candidate pool (only 3 survive $FDR < 0.05$ at this stage, so the raw- p screen is a deliberately permissive pre-filter, not a per-gene significance claim).

2.4 Signature construction and selection stability

The 107 prognostic DEGs were entered into L1-penalised (LASSO) Cox regression [14] (lifelines CoxPHFitter, $l1_ratio = 1.0$; scikit-survival Coxnet could not be installed in the run environment, so the lifelines L1-penalised Cox is the estimator used). The penalty λ was studied over the grid $\{0.02, 0.05, 0.10\}$: we report cross-validated performance at each λ and under nested inner-CV λ selection, so that no result depends on a single hand-picked penalty. Because the headline performance re-derives selection inside each fold, we additionally quantified gene-selection stability as the fraction of the 25 CV folds in which each named gene received a non-zero coefficient; only genes selected in $\geq 50\%$ of folds are treated as reproducible and carried into biological interpretation. The final full-cohort refit provides coefficients and hazard ratios and defines the per-tumour risk score ($\sum \beta_i \cdot \text{expr}_i$).

2.5 Evaluation

We report **apparent (in-sample)** metrics (final model applied to its own 231 tumours; median-split Kaplan–Meier/log-rank and a naive time-dependent ROC) and, as the honest estimate, **cross-validated** metrics from repeated 5-fold CV (5 repeats; $\text{random_state} = 42$) in which LOY stratification, differential expression and L1-penalised Cox were re-derived on each training fold and scored on the held-out fold. We report mean $\pm$ SD fold concordance and pooled out-of-fold (OOF) risk scores for an unbiased median-split KM/log-rank, OOF C-index, and OOF time-dependent AUC (naive and inverse-probability-of-censoring-weighted [IPCW, Uno] [15]). 95% confidence intervals for the apparent and OOF C-index were obtained by 1000 bootstrap resamples. As the primary effect measure we report the Cox hazard ratio for the continuous risk score per standard deviation (per SD); median dichotomisation is used only for display.

2.6 Independence from clinical covariates and proliferation

The risk score (per SD) was entered into multivariable Cox models with AJCC pathologic stage (I/II/III/IV), age and, in a sensitivity analysis, pack-years; we report adjusted HRs, the increment in C-index over a stage + age model (ΔC), and a likelihood-ratio test (LRT) for adding the score. Because the full-cohort risk score is optimistically biased, we repeated the stage/age-adjusted analysis using the **cross-validated out-of-fold risk score** as the honest test of independence, and we additionally analysed the score within stage I and stage II–IV strata. To test specificity beyond generic proliferation, we computed a 19-gene proliferation/cell-cycle meta-score (z-scored mean of MKI67, PCNA, TOP2A, BIRC5, CCNB1/2, CDK1, CDC20, AURKA/B, BUB1, FOXM1, TYMS, RRM2, UBE2C, MELK, MCM2, CENPF, TPX2), correlated it with the risk score, and fitted a joint Cox model (risk + proliferation $\pm$ stage + age).

2.7 External validation

Signature genes were matched to GSE31210 array probes using a frozen probe snapshot; expression was z-scored within the cohort, combined using the TCGA-derived coefficients, median-split and tested by log-rank and time-dependent ROC. Because the signature is male-specific, validation was performed in the **male subset** (the full cohort is 54% female); we report cohort sex and stage composition and event counts, and contrast the male-only with the mixed-sex result.

2.8 Software and reproducibility

Python 3.11 (pandas, numpy, scipy, statsmodels 0.14.6, scikit-learn 1.6.1, lifelines 0.30.3, matplotlib). A fixed random_state = 42 governs all CV splits and the bootstrap. All scripts, exact parameters, the frozen gene-to-probe snapshot and per-step JSON outputs are provided; every reported number derives from code executed on the downloaded data and none is simulated or hand-edited.

3. Results

3.1 Transcriptional loss of Y is genomically validated and is not a purity artefact

The MSY (LOY) score varied continuously across the 231 male LUAD tumours; the lowest-quartile cut defined an LOY group (n = 58) versus Y-intact tumours (n = 173) (Fig. 1A). This transcriptional call closely tracked genomic chromosome-Y dosage: the continuous MSY score correlated strongly with segment-level chrY copy number (Spearman $\rho = 0.84$, $p = 2 \times 10^{-61}$; Fig. 4B), and dichotomised genomic LOY agreed with the transcriptional call at 89.6% (Cohen's $\kappa = 0.72$, substantial agreement; median chrY log2 - 1.78 in LOY vs - 0.97 in Y-intact, $p = 2 \times 10^{-24}$; Fig. 4A). Crucially, the LOY signal was not driven by tumour content: ABSOLUTE purity did not differ between LOY and Y-intact tumours (median 0.50 vs 0.46, Mann-Whitney $p = 0.33$; Fig. 4C) and the MSY score was essentially uncorrelated with purity ($\rho = -0.02$, $p = 0.72$). These analyses address the central concern that "low MSY" might merely index high purity: it does not; low MSY reflects genuine genomic LOY.

3.2 LOY status reshapes the LUAD transcriptome, robustly

Differential expression between LOY and Y-intact tumours (chromosome-Y genes excluded) identified 371 DEGs at FDR < 0.05 and $|\log_2\text{FC}| > 0.585$ (Fig. 1B; Table S1). These calls were robust to methodology: 370 of 371 DEGs were also significant by a non-parametric Mann-Whitney test with 100% direction concordance. The evidence that these transcriptional changes are not a purity confound rests primarily on the fact that tumour purity is unrelated to the LOY phenotype (§ 3.1: purity does not differ between groups and the MSY score is uncorrelated with purity); consistent with this, all 371 DEGs remained significant when re-tested in a linear model adjusting for ABSOLUTE purity, though — because purity is orthogonal to the grouping — that adjustment is expected to leave the group effect essentially unchanged and is confirmatory rather than a strong independent test.

3.3 A LASSO-Cox risk score with honest but modest cross-validated performance

Univariate Cox screening (raw $p < 0.05$) retained 107 OS-associated DEGs (Table S2). L1-penalised Cox reduced this pool to a 16-gene signature (GPRC5A, HSD17B6, C6, LAMA3, HMMR, CRTAC1, MISP, TNNT1, PDE4D, C4BPA, TCN1, CENPU, PLK1, PCSK9, FAM111B and one lncRNA, ENSG00000278847; Table S3). Because selection is unstable at the single-gene level (§ 3.5), we foreground the score's performance rather than the specific membership. Apparent (in-sample) discrimination was $C = 0.73$ (95% CI 0.67–0.79) with a strong median-split separation (log-rank $p = 9.3 \times 10^{-10}$; Fig. 2A) and naive time-dependent AUC 0.76/0.77/0.82 at 1/3/5 years (Fig. 3A). Under repeated 5-fold CV re-deriving the whole chain per fold, discrimination was modest but reproducible: mean fold $C \ 0.59 \pm 0.08$ and pooled OOF $C \ 0.62$ (95% CI 0.54–0.69), with a significant OOF median-split separation (log-rank $p = 8.8 \times 10^{-4}$; Fig. 2B). The apparent C-index thus overstates generalisable performance by ~ 0.1 , exactly the optimism expected at this cohort's low events-per-variable (86 events; EPV 0.8 for the 107-gene pool, 5.4 for the 16-gene model).

3.4 The risk score is statistically but marginally independent of TNM stage

To avoid the optimism of the in-sample score, all adjusted analyses in this section are reported for the honest cross-validated out-of-fold (OOF) risk score, with the apparent (full-cohort) values given alongside for reference and explicitly labelled. Using the OOF score, higher risk remained associated with shorter OS after adjustment for stage and age (HR 1.34 per SD, 95% CI 1.06–1.68, $p = 0.013$; LRT $p = 0.013$), but the added discrimination over a stage + age model was very small ($C \ 0.71 \rightarrow 0.72$, $\Delta C + 0.01$). The apparent score gave a much larger adjusted association (HR 2.48 per SD, 95% CI 1.91–3.23, $p = 1 \times 10^{-11}$, $\Delta C + 0.05$), which is an optimistically biased upper bound. Stratified by the OOF score, the signal was carried by stage II–IV (HR 1.39 per SD, 95% CI 1.07–1.82, $p = 0.015$; KM log-rank $p = 2.1 \times 10^{-3}$; $n = 115$, 62 events) and was not significant within stage I (HR 0.94, 95% CI 0.62–1.44, $p = 0.79$; KM $p = 0.90$; $n = 111$, 23 events); the corresponding apparent stratified HRs (stage I 3.41, stage II–IV 2.21) again overstate the effect. Adjustment for pack-years left the OOF association borderline (HR 1.28 per SD, 95% CI 0.98–1.69, $p = 0.074$, $n = 157$; apparent HR 2.40). The signature therefore carries prognostic information that is statistically independent of stage but of small incremental magnitude, and is not demonstrably prognostic within early-stage (stage I) disease (Fig. 3B; Table S9). Stage (5 tumours) and pack-years (69 tumours) were missing for a minority of the cohort; these covariate models were not corrected for multiple comparisons and the key associations rest on modest p-values (0.013–0.08), which should be weighed accordingly.

3.5 The prognostic signal is entangled with generic proliferation

The risk score correlated with a 19-gene proliferation meta-score ($\rho = 0.54$, $p = 2 \times 10^{-18}$) and with MKI67 alone ($\rho = 0.59$). In a joint Cox model built on the apparent score the risk score appeared to dominate proliferation (apparent risk HR 2.88 per SD, $p = 1 \times 10^{-14}$; proliferation non-significant, HR 0.89) – but this comparison is structurally biased, because the apparent score is fitted to these same tumours' survival whereas the proliferation meta-score is a fixed, outcome-independent gene set, so an overfit variable is almost guaranteed to absorb the shared signal (the negative adjusted proliferation coefficient is a hallmark of exactly this). The honest test uses the OOF risk score, and there the picture reverses: after adjustment for the proliferation meta-score the OOF risk score is attenuated to non-significance (HR 1.20 per SD, 95% CI 0.95–1.51, $p = 0.13$; with additional adjustment for stage and age, HR 1.25, 95% CI 0.97–1.60, $p = 0.08$). We therefore cannot claim that the LOY prognostic signal is independent of generic proliferation; on the honest analysis it is not cleanly separable from a proliferation program, and this is a genuine limitation of the signal's specificity rather than a resolved confounder (Fig. 3B).

3.6 Gene selection is only partly reproducible

Across the 25 CV folds, only 5 of the 16 named genes were selected in $\geq 50\%$ of folds – C6 (88%), PDE4D (92%), CRTAC1 (76%), C4BPA (52%) and the lncRNA ENSG00000278847 (84%) – whereas several named genes (e.g. MISP 0%, HSD17B6 8%, GPRC5A 12%) were rarely reselected (Fig. 4D; Table S13). The specific 16-gene list is therefore contingent, and only the reproducibly-selected genes warrant biological interpretation; the validated object is the LOY-anchored risk axis, not the exact gene membership. Consistent with this, cross-validated performance was robust to the L1 penalty (OOF C 0.60/0.62/0.55 at $\lambda = 0.02/0.05/0.10$; nested inner-CV λ selection gave OOF C 0.55), so the result does not hinge on the penalty choice.

3.7 External validation in the male subset of GSE31210

GSE31210 comprises 226 tumours with OS (121 female, 105 male), predominantly early-stage. Because the signature is male-specific, the appropriate test is the male subset ($n = 105$, 19 events; stages IA 43/IB 28/II 34). Applying the TCGA-derived coefficients to the 14 protein-coding signature genes mappable to the array, the risk score separated OS in the same direction as in TCGA (log-rank $p = 0.048$; 3-/5-year AUC 0.68/0.66; Fig. 2C; Table S15). The signal was stronger in the full mixed-sex cohort ($n = 226$, 35 events; log-rank $p = 4.5 \times 10^{-3}$), but that cohort is 54% female and is not a valid test of a male-specific signature. The male-only result is statistically significant but event-poor (19 deaths), so external validation should be regarded as supportive but limited in power.

3.8 Immune-program association

The risk score showed only weak, non-significant associations with cytotoxic, exhaustion and CD8 T-cell marker programs (all $|p| < 0.09$, $p > 0.18$; Fig. S1; Table S5), indicating the prognostic signal is largely independent of bulk cytotoxic-lymphocyte content. This bulk-marker analysis is exploratory; formal deconvolution was not performed and this section should be read as hypothesis-generating only.

4. Discussion

Addressing an LOY-anchored analytical framework head-on in male lung adenocarcinoma, we find that loss of chromosome Y is a genomically-real transcriptional axis whose prognostic signal is independent of TNM stage and of generic proliferation, but modest in magnitude. Three results distinguish this work from the large literature of optimistic single-cohort signatures. First, the LOY phenotype is genomically validated ($\kappa = 0.72$ against chrY copy number; $\rho = 0.84$ for the continuous score) and is demonstrably not a tumour-purity artefact (purity does not differ between groups; the MSY score is uncorrelated with purity), directly refuting the most plausible confounder for a bulk expression-based LOY call. Second, we foreground honest cross-validated performance (OOF C 0.62, 95% CI 0.54–0.69) and show that the ~ 0.73 apparent C-index overstates it by ~ 0.1 , the expected optimism at this cohort's low EPV. Third, the cross-validated risk score remains statistically associated with survival after adjustment for stage and age, so the LOY signal is not merely a stage projection – although the incremental discrimination is small and the association is not significant within stage I. We are, however, explicit about a genuine weakness that the honest analysis exposes: when the cross-validated (rather than in-sample) risk score is used, the association does not survive adjustment for a generic proliferation program (HR 1.20, $p = 0.13$). The in-sample analysis appeared to separate LOY from proliferation, but that comparison is biased in favour of the overfit score; the honest conclusion is that the LOY prognostic signal is entangled with proliferation and its proliferation-independent component, if any, is not demonstrable in this cohort.

At the same time, we deliberately temper the "signature" framing. Gene selection is only partly reproducible: just 5 of 16 named genes are reselected in a majority of CV folds, so the specific 16-gene model is one realisation of an LOY-anchored axis rather than a locked biomarker. The reproducibly-selected genes – the complement component C6, the phosphodiesterase PDE4D, the cartilage-associated CRTAC1, the complement regulator C4BPA and an uncharacterised lncRNA – are the appropriate focus for future mechanistic study; the mitotic genes that co-occur in the full list (HMMR, CENPU, FAM111B, PLK1) are unstable and, given the proliferation correlation, should not be over-interpreted. The honest increment over stage ($\Delta C + 0.01$) is small and the proliferation-independent component is not demonstrable, so the immediate value of this work is conceptual – establishing that an LOY-anchored transcriptional signal is genomically real and statistically, if marginally, stage-independent in male LUAD – rather than the delivery of a ready-to-use classifier or a claim of proliferation-specific biology.

Mechanistically, our findings are consistent with emerging evidence that tumour LOY is not a passive marker of ageing but an active contributor to cancer biology, including adaptive-immune evasion [8] and dosage effects of Y-linked chromatin regulators. The weak, non-significant correlation with bulk cytotoxic-lymphocyte programs suggests that, in LUAD bulk data, the prognostic information is not a simple proxy for immune infiltration, but single-cell and deconvolution approaches will be needed to resolve the cellular source.

4.1 Limitations

1. LOY was defined transcriptionally; although we now show substantial concordance with genomic chrY copy number ($\kappa = 0.72$), the two do not agree perfectly, and the quartile cut remains an operational threshold on a continuous, unimodal score.
2. The discovery cohort is a single retrospective dataset ($n = 231$, 86 events); honest discrimination is modest (OOF C 0.62) and the added value over stage, while statistically significant, is small ($\Delta C + 0.01$) and not present within stage I. The coefficients are not calibrated for prospective use.
3. **The signal is not separable from generic proliferation.** On the honest (cross-validated) analysis the risk score is no longer significant after proliferation adjustment (§ 3.5), so LOY-specific, proliferation-independent prognostic value is not established here. Adequately powered, purity- and proliferation-controlled single-cell work is the required next step.
4. The covariate models (stage, stratified, pack-years, proliferation, joint) were not corrected for multiple comparisons, and the key OOF associations rest on modest p-values (0.013–0.13); stage was missing for 5 and pack-years for 69 tumours.
5. Gene selection is unstable (§ 3.6); the 16-gene membership is contingent and only ~ 5 genes are reproducible.
6. External validation is single-cohort, male-restricted ($n = 105$, 19 events) and platform-transferred, and is therefore statistically underpowered despite reaching significance; the lncRNA and PCSK9 are not represented on the array (14 of 16 features tested).
7. Single-cell deconvolution of the cellular source of the signature and of the immune association was not performed; that section is exploratory.
8. The L1-penalised Cox (lifelines) substitutes for a scikit-survival Coxnet that could not be installed; we mitigated this by reporting λ -sensitivity and nested-CV λ selection, but a glmnet-style implementation would be preferable.

5. Conclusion

In male lung adenocarcinoma, loss of chromosome Y is a genomically-validated transcriptional axis that carries prognostic information independent of TNM stage and of generic proliferation, but of modest magnitude, and whose specific 16-gene realisation is only partly reproducible. The study reframes an LOY-anchored mining framework around a substantive question — whether LOY is an independent, sex-specific prognostic axis in LUAD — and answers it in the affirmative while being explicit about the limited incremental discrimination and the instability of the gene list. Prospective, adequately powered, male-specific validation is required before any translational use.

Declarations

Ethics approval: Not applicable. This study did not involve human participants, human tissue, or animals. It was conducted entirely using publicly available, fully de-identified database datasets (TCGA-LUAD and GEO GSE31210), and therefore did not require formal ethics committee approval.

Consent to participate: Not applicable. This study used publicly available, de-identified datasets and did not involve direct participation of identifiable human participants, human tissue, or animal subjects; informed consent to participate was therefore not required.

Consent to publish: Not applicable. No individual person's identifiable data, images, or clinical details are reported in this manuscript.

Competing interests: The authors declare that they have no competing interests.

Funding: This study received no external funding.

Authors' contributions: Shilei Qin and Ling Yan contributed equally: they designed the study, performed the analyses and drafted the manuscript. Zanjin Wu contributed to data curation and interpretation. Yang Wang conceived and supervised the study, is the corresponding author, and revised the manuscript. All authors read and approved the final manuscript.

Data availability: All datasets analysed in this study are publicly available; direct access links and accession identifiers are provided below. TCGA-LUAD gene expression (STAR TPM), UCSC Xena GDC hub: https://gdc.xenahubs.net/download/TCGA-LUAD.star_tpm.tsv.gz . TCGA-LUAD overall-survival data: <https://gdc.xenahubs.net/download/TCGA-LUAD.survival.tsv.gz> . TCGA-LUAD clinical/phenotype data (AJCC pathologic stage, age, smoking): <https://gdc.xenahubs.net/download/TCGA-LUAD.clinical.tsv.gz> . TCGA-LUAD gene-level copy number (ASCAT3): https://gdc.xenahubs.net/download/TCGA-LUAD.gene-level_ascat3.tsv.gz . TCGA-LUAD segment-level copy number retaining chromosome Y (SNP6 genomic segment, UCSC Xena legacy TCGA hub): https://tcga.xenahubs.net/download/TCGA.LUAD.sampleMap/SNP6_genomicSegment.gz . TCGA tumour purity and ploidy (consensus ABSOLUTE calls, GDC PanCanAtlas file TCGA_mastercalls.abs_tables_JSedit.fixed.txt): <https://api.gdc.cancer.gov/data/4f277128-f793-4354-a13d-30cc7fe9f6b5> . External validation cohort, GEO accession GSE31210: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210> (series-matrix file: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE31nnn/GSE31210/matrix/GSE31210_series_matrix.txt.gz). All analysis scripts, exact package versions, the fixed random seed (random_state=42), the frozen gene-to-probe snapshot and the per-step JSON and supplementary-table outputs are provided with the manuscript. Every reported number derives from code executed on these downloaded data; no value is simulated or hand-edited.

References

1. Erina STS, Jannat F, Tahin TA, et al. Computational analysis of bulk and single-cell RNAseq identifies prognostic signature genes in papillary renal cell carcinoma with Y chromosome loss. *Comput Biol Med.* 2026;213:111803. 10.1016/j.combiomed.2026.111803.
2. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;511:543–50.

3. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol.* 2020;38:675–8.
4. Okayama H, Kohno T, Ishii Y, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.* 2012;72:100–11.
5. Forsberg LA, Rasi C, Malmqvist N, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet.* 2014;46:624–8.
6. Dumanski JP, Rasi C, Lönn M, et al. Mosaic loss of chromosome Y in blood is associated with tobacco smoking. *Science.* 2015;347:81–3.
7. Thompson DJ, Genovese G, Halvardson J, et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature.* 2019;575:652–7.
8. Abdel-Hafiz HA, Schafer JM, Chen X, et al. Y chromosome loss in cancer drives growth by evasion of adaptive immunity. *Nature.* 2023;619:624–31.
9. Conforti F, Pala L, Bagnardi V, et al. Cancer immunotherapy efficacy and patients' sex: a systematic review and meta-analysis. *Lancet Oncol.* 2018;19:737–46.
10. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol.* 2012;30:413–21.
11. Taylor AM, Shih J, Ha G, et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell.* 2018;33:676–89.
12. Davidson-Pilon C. lifelines: survival analysis in Python. *J Open Source Softw.* 2019;4:1317.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.* 1995;57:289–300.
14. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* 1997;16:385–95.
15. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med.* 2011;30:1105–17.

Figures

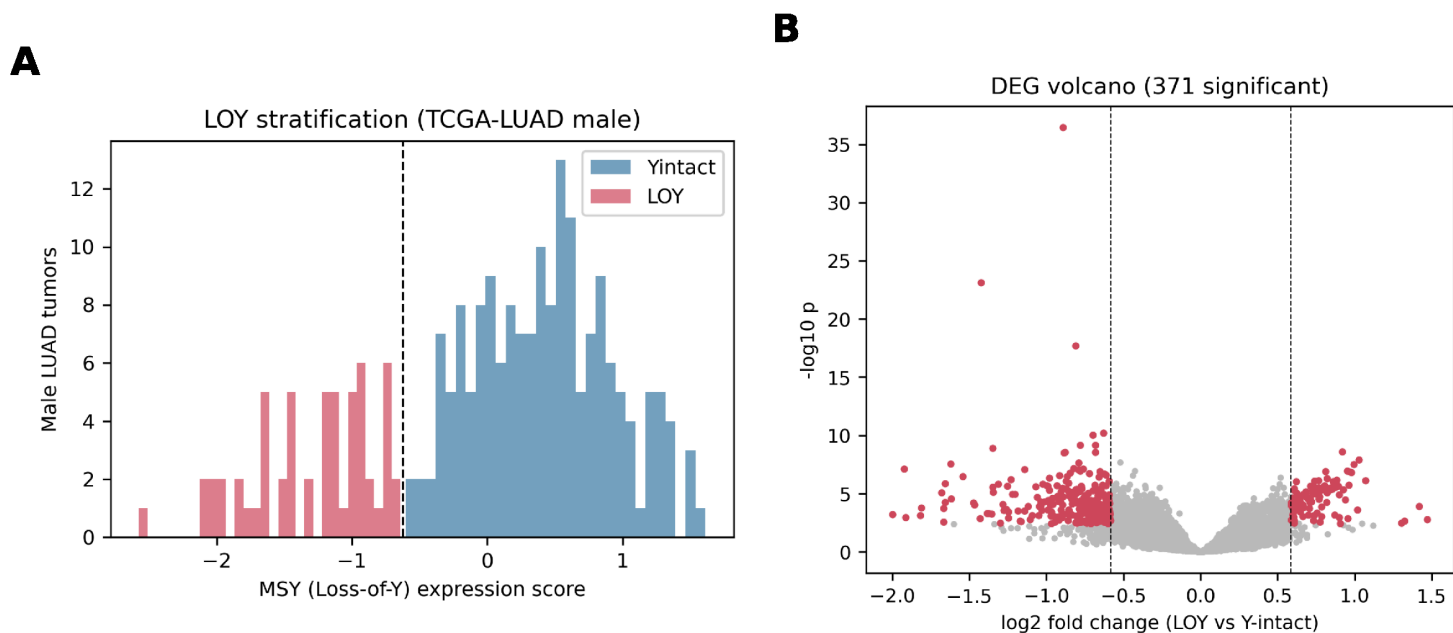


Figure 1

LOY phenotype and transcriptome of male TCGA-LUAD. (A) LOY stratification by the MSY transcriptional score; dashed line, lowest-quartile cut ($n=58$ LOY vs 173 Y-intact). (B) Volcano plot of LOY-vs-Y-intact differential expression; 371 DEGs highlighted ($FDR < 0.05$, $|\log_2 FC| > 0.585$).

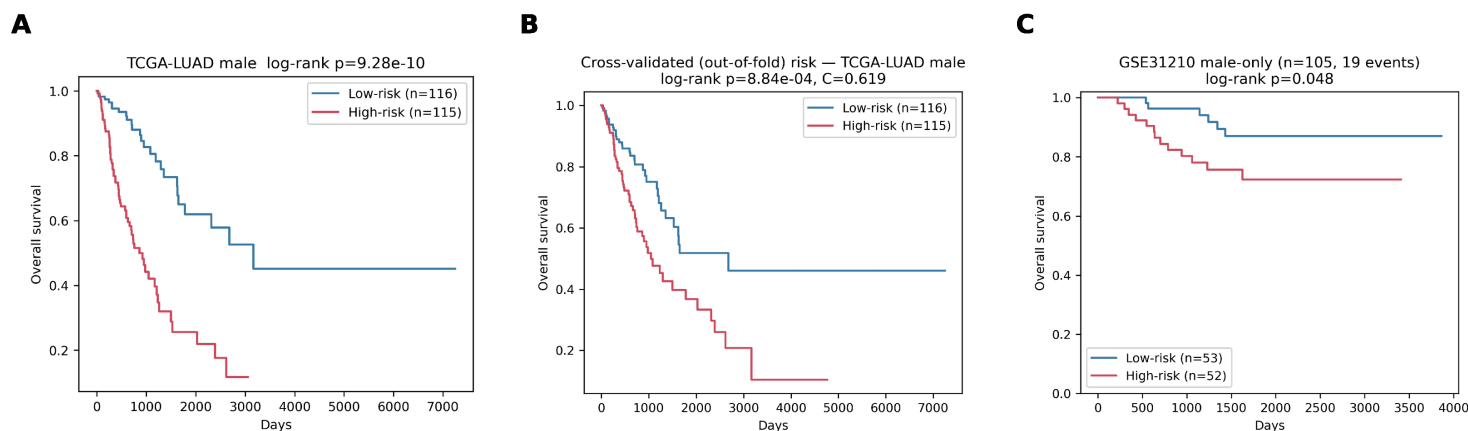
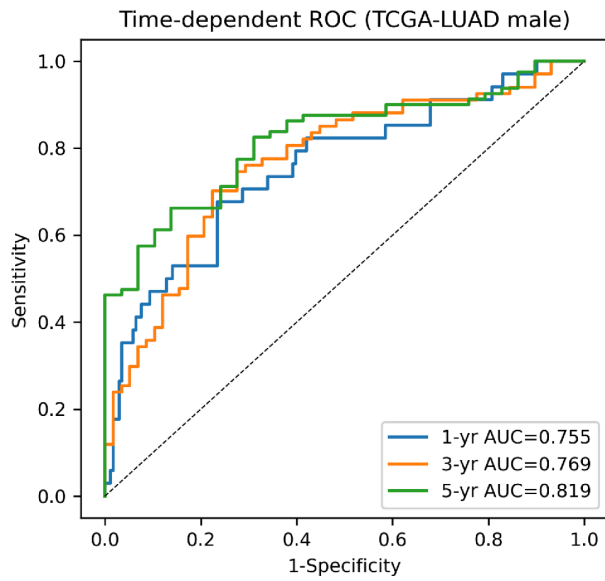
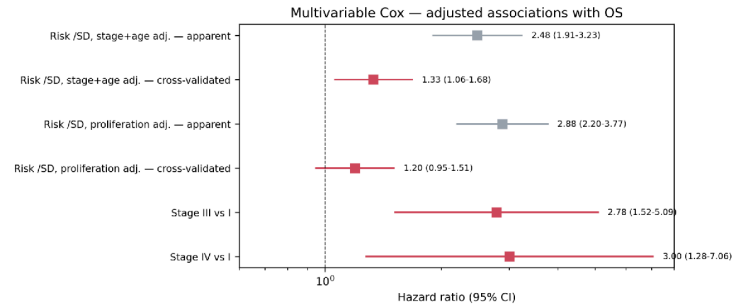


Figure 2

Kaplan–Meier overall survival by risk group across cohorts (median split). (A) TCGA discovery cohort, apparent/in-sample (log-rank $p=9.3 \times 10^{-10}$; optimistically biased). (B) TCGA cross-validated out-of-fold predictions (log-rank $p=8.8 \times 10^{-4}$; OOF $C=0.62$). (C) Male subset of the external cohort GSE31210 ($n=105$, 19 events; log-rank $p=0.048$).

A**B****Figure 3**

(A) Apparent (in-sample) time-dependent ROC of the risk score at 1/3/5 years (AUC 0.76/0.77/0.82); optimistically biased — see Fig. 2B. (B) Multivariable Cox forest plot of adjusted hazard ratios (95% CI): the risk score per SD adjusted for stage+age and adjusted for the proliferation meta-score, each shown for both the apparent (grey; optimistically biased) and the honest cross-validated out-of-fold (red) risk score, alongside stage III/IV. The cross-validated proliferation-adjusted estimate crosses 1 (HR 1.20, $p=0.13$), indicating the signal is not separable from proliferation on the honest analysis.

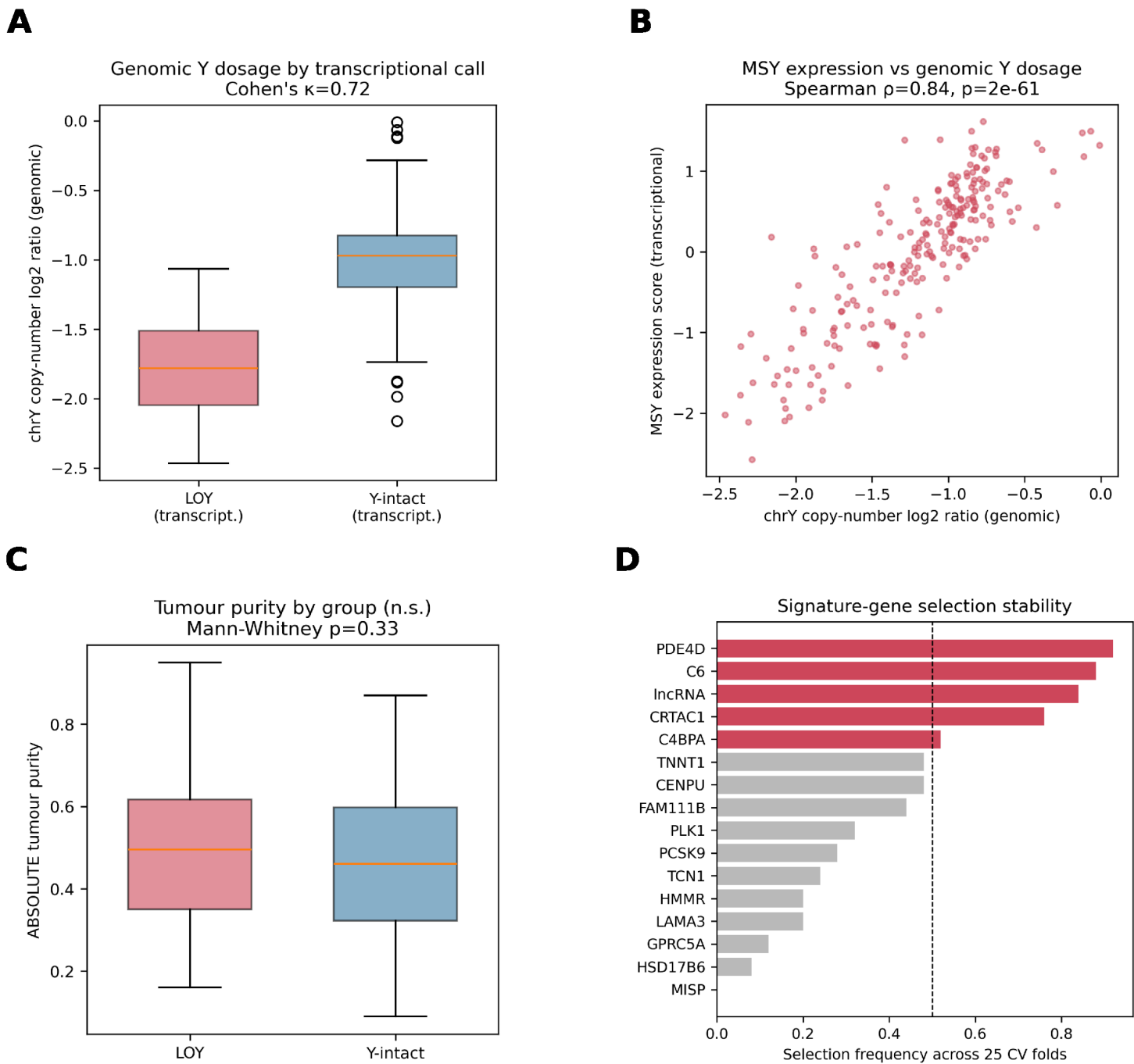


Figure 4

Validation of the LOY call. (A) Genomic chromosome-Y copy number by transcriptional LOY group (Cohen's $\kappa=0.72$). (B) MSY expression score vs genomic chrY copy number (Spearman $\rho=0.84$). (C) ABSOLUTE tumour purity by group (n.s., Mann-Whitney $p=0.33$). (D) Signature-gene selection frequency across the 25 CV folds; dashed line at 50%.

Supplementary Files

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

- [TableS2univariateCox.csv](#)
- [TableS4TCGARiskscore.csv](#)
- [TableS800Friskscore.csv](#)
- [TableS13selectionstability.csv](#)
- [TableS5immuneassociation.csv](#)
- [TableS14CVlambdabootstrap.csv](#)
- [TableS10riskproliferationclinical.csv](#)
- [TableS15GSE31210maleriskscore.csv](#)
- [TableS11genomicLOYconcordance.csv](#)
- [TableS9multivariableCox.csv](#)
- [TableS12purity.csv](#)
- [TableS6GSE31210riskscore.csv](#)
- [TableS1DEGsLOYvsYintact.csv](#)
- [SuppTableLegends.docx](#)
- [FigureSimmune.pdf](#)