Research Square

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

Malignant epithelial state composition predicts overall survival in non muscle invasive bladder cancer

Yanping Wang
Wannan Medical College
Houbao Huang
drhuanghoubao@163.com

The First Affiliated Hospital of Wannan Medical College

Research Article

Keywords: bladder cancer, non-muscle-invasive, single-cell RNA sequencing, epithelial state, deconvolution, Hippo signalling, prognostic model

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11668833/v1

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

Additional Declarations: No competing interests reported.

Malignant epithelial state composition predicts overall survival in non muscle invasive bladder cancer

Yanping Wang[1], **Houbao Huang**[2]

[1] Wannan Medical College, Wuhu, Anhui, China [2] Department of Urology, The First Affiliated Hospital of Wannan Medical College, Wuhu, Anhui, China

Corresponding author: Houbao Huang, Department of Urology, The First Affiliated Hospital of Wannan Medical College, Wuhu, Anhui, China. Email: drhuanghoubao@163.com

Running title: Epithelial states predict bladder cancer survival

Keywords: bladder cancer; non-muscle-invasive; single-cell RNA sequencing; epithelial state; deconvolution; Hippo signalling; prognostic model

Funding: Anhui Provincial Science and Technology Tackling Key Projects (Grant No. 202423k09020027).

Abstract

Background. Non-muscle-invasive bladder cancer (NMIBC) is clinically heterogeneous and risk stratification still rests on clinico-pathological variables. Whether malignant epithelial states, not cell types, carry transferable prognostic information is unknown.

Methods. We re-analysed 13 scRNA-seq libraries (GSE222315; nine T1 NMIBC tumours, four normals; 90,829 cells). Malignant cells were called sample by sample with inferCNV, using autologous immune-stromal cells as the diploid reference and normal epithelium as a benign anchor. Malignant subclusters were annotated into epithelial states, projected onto TCGA-BLCA (n = 402), modelled with LASSO-Cox and validated in GSE32894 (n = 308) as an abundance proxy and by NNLS deconvolution.

Results. 3,880 of 23,200 tumour epithelial cells (16.7%) were malignant and resolved into 13 states. The model separated TCGA patients by survival (HR per SD 4.24, $p = 1.2 \times 10^{-7}$; C-index 0.634). In GSE32894 the score rose with stage ($p = 7.4 \times 10^{-14}$) and grade ($p = 3.7 \times 10^{-15}$) and predicted overall survival (HR 6.05, p = 0.0010; C-index 0.790) but progression only weakly (OR 1.40; AUC 0.610), because no T3/T4 patient progressed after cystectomy. Deconvolution reproduced the same biology (C-index 0.606, 0.736). The adverse state S7 was KLF5-low and YAP/TEAD-driven (TEAD1 +10.8, KLF5 −3.3), with EMT and loss of uroplakin expression.

Conclusions. Epithelial state composition transfers from single-cell to bulk cohorts and predicts survival better than progression; progression endpoints in public NMIBC cohorts are structurally censored.

1. Introduction

Bladder cancer is among the ten most common malignancies worldwide, and roughly three quarters of new cases are non-muscle-invasive (NMIBC) at diagnosis [1,2]. The clinical problem in NMIBC is not the initial treatment — transurethral resection with or without intravesical bacillus Calmette–Guérin is standard [2] — but the inability to know, at diagnosis, which tumours will recur, which will progress to muscle invasion, and which would have remained indolent irrespective of therapy. Current risk nomograms rest on stage, grade, tumour multiplicity, size, carcinoma *in situ* and prior recurrence [3,34–36], all of which describe the tumour macroscopically and none of which describes its biology.

Single-cell RNA sequencing has begun to resolve bladder cancer into its constituent cell types and has revealed conserved expression programmes in malignant epithelium, a rich immune infiltrate, and a stromal compartment that differs between responders and non-responders to BCG [31–33]. However, three gaps remain. First, cell-type atlases are not transferable clinical tools: they require fresh tissue, and their outputs are proportions of broad lineages that any pathologist already reports. Second, most single-cell studies stop at description and do not test whether the states they define carry information in independent, clinically annotated cohorts. Third, the malignant epithelium is usually treated as a single entity or a handful of subtypes, although it is now clear that malignant cells exist in a continuum of transcriptional *states* — basal, luminal, umbrella, squamous, proliferative, interferon-responsive, EMT-like [8,30,38,41] — whose relative abundance may be the relevant prognostic quantity.

We therefore asked a deliberately narrow question: can the composition of malignant epithelial states, defined at single-cell resolution in early-stage disease, be projected onto bulk expression data and used to stratify patients, and does the resulting model hold up in an independent cohort when the endpoint is chosen appropriately?

To answer it we used a single-cell cohort of treatment-naive T1 NMIBC with matched adjacent normal tissue, defined malignant cells by copy-number inference rather than by clustering, annotated malignant states without relying on the original authors' labels, and built a penalised survival model in TCGA-BLCA that we then tested in the NMIBC cohort GSE32894. Because a portion of our conclusions depends on how state abundance is quantified, we repeated the entire modelling exercise with a formal deconvolution (NNLS against the single-cell reference) and report both.

2. Methods

2.1 Data sources

Public data only. Single-cell discovery cohort: GSE222315 (nine T1 NMIBC tumours, p1–p9_BCa, plus four patient-matched adjacent normal samples, p2/p3/p4/p5_NAT). Bulk training cohort: TCGA-BLCA as distributed through the UCSC Xena / GDC hub (STAR counts, HiSeqV2 normalised expression, clinical and survival tables). External validation cohort: GSE32894 (308 NMIBC samples with clinical annotation). Urine single-cell data for the supplementary analysis: GSE267718.

2.2 Quality control, integration and annotation

Cells were retained if $200 \leq$ nFeature $\leq 8{,}000$, nCount > 500, mitochondrial reads $< 20\%$ and haemoglobin reads $< 5\%$. Doublets were removed per sample with scDblFinder [15]. Libraries were integrated with Harmony [14] on sample identity after library-size normalisation, variance-stabilising transformation and

selection of 2,000 highly variable genes, followed by 30 principal components and shared-nearest- neighbour clustering at resolution 0.8. Clusters were annotated manually after inspecting marker fingerprints cluster by cluster; automatic argmax assignment was deliberately not used, and the original authors' annotations were not reused.

2.3 Identification of malignant cells

Copy-number inference was run with inferCNV [11] **sample by sample**, so that each tumour used its own immune and stromal cells as the diploid reference, with adjacent normal epithelium available as a benign anchor. All epithelial cells were used — no subsampling. Parameters: cutoff 0.1, denoising, HMM disabled, `cluster_by_groups = TRUE`, eight threads. The malignancy threshold was set at the 95th percentile of the CNV score of adjacent normal epithelial cells (0.003038). Cells were additionally required to show a higher correlation with the epithelial reference profile than with the immune or stromal reference profiles.

2.4 Malignant states

Malignant cells were re-clustered after removal of endothelial-like and ribosome-dominated contaminant clusters. States were assigned by scoring nine canonical epithelial panels (basal, luminal, umbrella, squamous, EMT, proliferative, interferon, ciliated, neuroendocrine) plus immune/endothelial/fibroblast panels, and taking the panel with the highest mean z-score when that score exceeded 1.

2.5 Differentiation index and pathway activity

To place the malignant states on a differentiation axis we used a marker-anchored index rather than a trajectory algorithm. For every malignant cell we averaged the gene-wise z-scores of six differentiation markers (UPK1A, UPK1B, UPK2, UPK3A, KLF5, PPARG) and subtracted the average z-score of three dedifferentiation markers (CDH2, WNT2, AMOTL2). Positive values denote a urothelial, umbrella-like identity; negative values denote loss of that identity.

We deliberately did **not** report a pseudotime. The 13 malignant states form **five disconnected components** in the UMAP embedding, so any graph-based pseudotime requires artificial edges between islands; when we tested such an implementation, its state-level values were uncorrelated with the differentiation index (Spearman $\rho = +0.15$, where a genuine trajectory would give a strong negative correlation), and it placed a well-differentiated state at the far end of the axis. The differentiation index has no such failure mode because it does not depend on graph connectivity.

Pathway activity was quantified with the Hallmark collection (MSigDB, 50 gene sets) using weighted mean rank scoring (`decoupleR::run_wsum`) on per-state pseudobulk profiles, and is displayed as row-wise z-scores across states.

2.6 Prognostic modelling in TCGA-BLCA

For each malignant state the top-40 markers were used as a signature, and each tumour was scored as the mean z-score of those genes (an abundance proxy). Univariable Cox models screened states for association with overall survival. A LASSO-penalised Cox model with 10-fold cross-validation selected the final states, and the linear predictor was used as the risk score. Discrimination was assessed by Harrell's C-index and by Kaplan–Meier analysis with the log-rank test [12,13,19–21].

2.7 Independent deconvolution

To avoid dependence on a single quantification, per-state pseudobulk profiles were built from the single-cell reference and bulk tumours were deconvolved by non-negative least squares over 6,229 state-discriminative genes, using gene-wise z-scoring to absorb platform differences. The same Cox and LASSO pipeline was then repeated on the deconvolved proportions, and the coefficients were applied unchanged to the external cohort.

2.8 External validation

GSE32894 was scored with the TCGA-derived model. Two endpoints were examined: overall survival (Cox, Kaplan–Meier, C-index) and progression to a higher stage or radical treatment (logistic regression, ROC/AUC). The progression field was parsed to accommodate the 12 textual variants present in the public annotation. Stage and grade gradients were assessed by Kruskal–Wallis tests, without using any survival information.

2.9 Immune microenvironment

Thirteen immune-cell signatures and nine checkpoint genes were scored in TCGA-BLCA, and their association with the risk score was measured by Spearman correlation with Benjamini–Hochberg adjustment.

2.10 Urine proteomics feasibility check

To ask whether the S7 protein panel is measurable in urine, we used the longitudinal urine proteome dataset PXD025139 (six patients, MaxQuant proteinGroups tables). For each patient table we recorded the number of quantified protein groups and the number of LFQ samples, and interrogated the panel CDH2, CDH1, WNT2, AMOTL2, YAP1, WWTR1, TEAD1, CFTR, DNAH11 (S7-upregulated) and UPK3A, UPK1A, UPK1B, UPK2, UPK3B, KLF5 (S7-downregulated), reporting detection per patient.

2.11 Statistical environment

All analyses were performed in R 4.5.0 with Seurat 5.5.1, inferCNV, glmnet, survival/survminer, clusterProfiler, decoupleR/dorothea and pROC. Figures use a consistent NPG palette at 300 dpi. Two-sided tests were used throughout; $p < 0.05$ was considered significant.

3. Results

3.1 A single-cell atlas of T1 non-muscle-invasive bladder cancer

Thirteen libraries yielded 103,349 cells after merging; quality control, doublet removal and Harmony integration retained **90,829 cells** in 32 clusters, annotated into nine compartments (Fig. 1A–E): T/NK (24,307; 26.8%), epithelial (23,790; 26.2%), myeloid (8,630; 9.5%), B (7,717; 8.5%), endothelial (7,341; 8.1%), smooth muscle (7,019; 7.7%), cycling (5,185; 5.7%), fibroblast (4,404; 4.8%) and plasma (2,436; 2.7%) cells. Tumour and normal libraries intermingled in the embedding while differing substantially in composition (Fig. 1D), a reminder that patient identity dominates over disease status in this cohort.

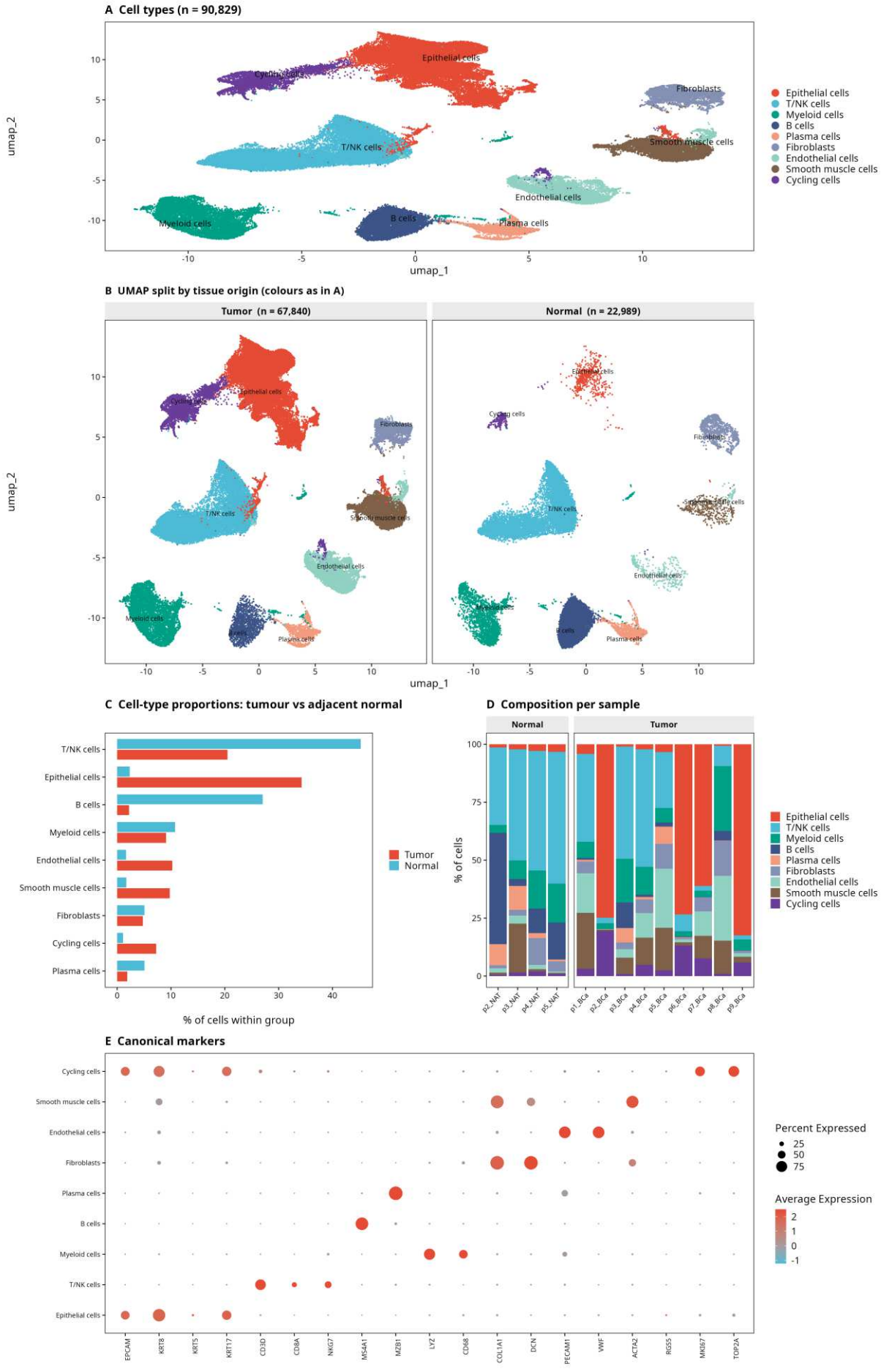

A Cell types (n = 90,829)
Epithelial cells
Cycling cells
Fibroblasts
T/NK cells
Smooth muscle cells
Endothelial cells
Myeloid cells
B cells
Plasma cells
umap_1
umap_2
B UMAP split by tissue origin (colours as in A)
Tumor (n = 67,840)
Normal (n = 22,989)
C Cell-type proportions: tumour vs adjacent normal
% of cells within group
Tumor
Normal
D Composition per sample
% of cells
p2_NAT
p3_NAT
p4_NAT
p5_NAT
p1_BCa
p2_BCa
p3_BCa
p4_BCa
p5_BCa
p6_BCa
p7_BCa
p8_BCa
p9_BCa
E Canonical markers
EPCAM
KRT8
KRT5
KRT17
CD3D
CD8A
NKG7
MS4A1
MZB1
LYZ
CD68
COL1A1
DCN
PECAM1
VWF
ACTA2
RGS5
MKI67
TOP2A
Percent Expressed
Average Expression

Figure 1. Single-cell atlas of T1 NMIBC (90,829 cells). (A) UMAP coloured by compartment, with manual annotations. (B) The same embedding split by tissue origin (tumour, 67,840 cells; adjacent normal, 22,989 cells), showing that the two tissues share all nine compartments but differ markedly in their abundance. (C) Compartment proportions within tumour and adjacent normal tissue: T/NK cells fall from 44% to 20% and B cells from 35% to 5%, whereas epithelial cells rise from 10% to 33%. (D) Compartment composition of each of the 13 libraries. (E) Canonical marker expression across compartments.

3.2 Copy-number inference identifies malignant epithelium in every evaluable tumour

Applying inferCNV per sample with autologous reference cells, 5.1% of adjacent normal epithelial cells exceeded the malignancy threshold versus 17.7% of tumour epithelial cells, and **3,880 cells (16.7% of tumour epithelium)** were called malignant (Fig. 2A). Malignant fractions ranged from 70.4% (p5) and 42.9% (p9) down to 4.0% (p2); p3 and p4 contained too few epithelial cells (87 and 129) to evaluate. The low fraction and squamous marker profile of p2 were consistent with squamous metaplasia rather than transformation, and we excluded it from state-level interpretation.

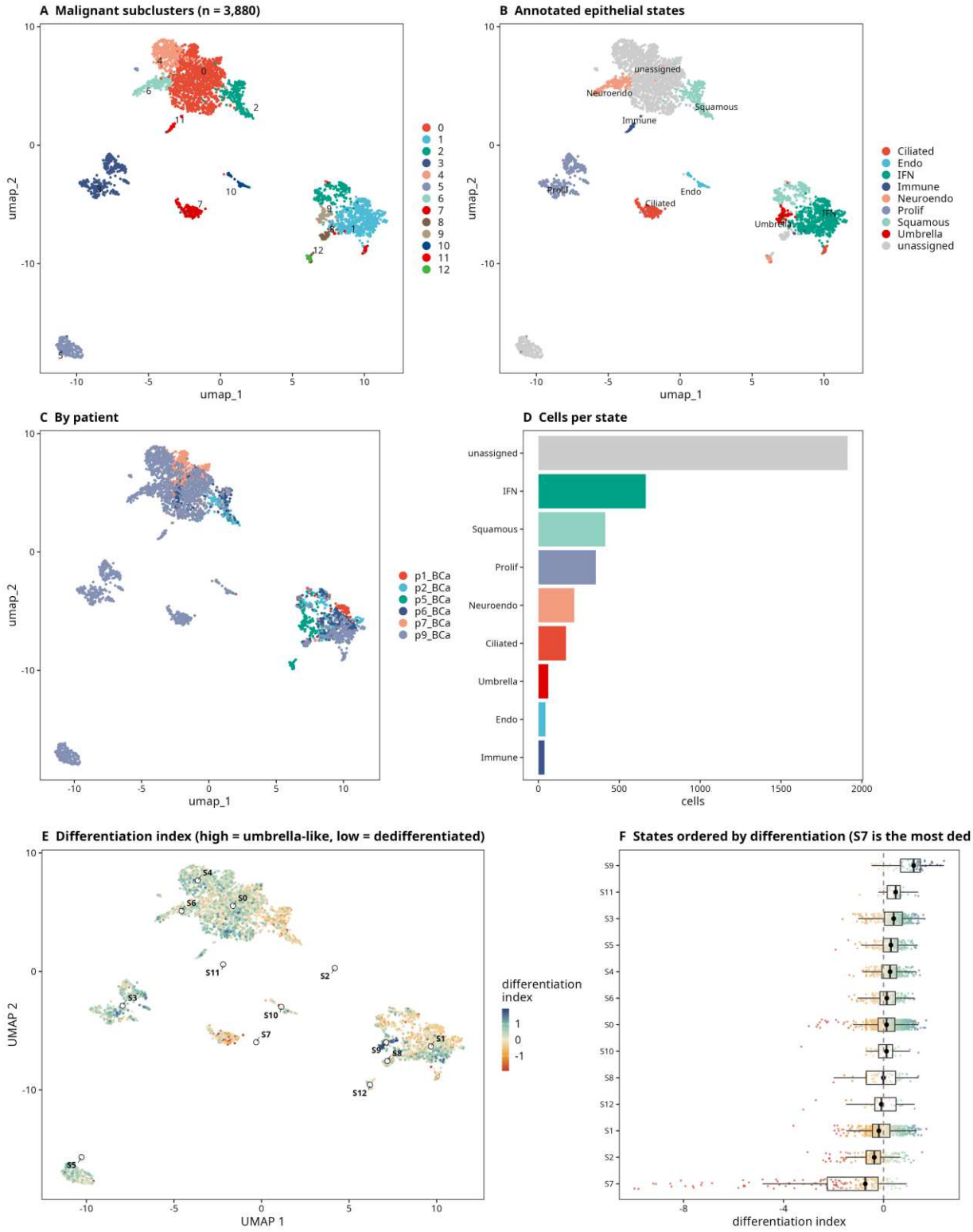

A Malignant subclusters (n = 3,880)
B Annotated epithelial states
C By patient
D Cells per state
E Differentiation index (high = umbrella-like, low = dedifferentiated)
F States ordered by differentiation (S7 is the most ded
G Hallmark pathway activity across states
umap_1
umap_2
UMAP 1
UMAP 2
cells
differentiation index
unassigned
Neuroendo
Squamous
Immune
Prolif
Endo
Ciliated
Umbrella
IFN
p1_BCa
p2_BCa
p5_BCa
p6_BCa
p7_BCa
p9_BCa
ADIPOGENESIS
ALLOGRAFT_REJECTION
ANGIOGENESIS
APICAL_SURFACE
APOPTOSIS
COAGULATION
EPITHELIAL_MESENCHYMAL_TRANSITION
ESTROGEN_RESPONSE_EARLY

Figure 2. Malignant epithelium and its transcriptional states. (A) Malignant subclusters (n = 3,880). (B) Annotated epithelial states. (C) Patient of origin. (D) Cells per state. (E) Differentiation index per cell, defined as the mean z-score of UPK1A/UPK1B/UPK2/UPK3A/KLF5/PPARG minus that of CDH2/WNT2/AMOTL2 (high, umbrella-like; low, dedifferentiated). (F) States ordered by the median differentiation index; S7 occupies the dedifferentiated extreme. No pseudotime is shown, because the 13 states form five disconnected components in the embedding and graph-based ordering is unreliable here (see Methods 2.5). (G) Hallmark pathway activity across states (50 MSigDB sets, weighted mean rank scoring on per-state pseudobulk, row-wise z-scores), states ordered by the differentiation index.

3.3 Malignant cells occupy thirteen transcriptional states

After removing endothelial-like and ribosome-dominated contaminants, malignant cells formed **13 subclusters** (Fig. 2A–D) that mapped onto a squamous state (S2), a proliferative state (S3), an interferon-response state (S1), a ciliated/dedifferentiating state (S7), an umbrella-cell state (S9) and an unassigned remainder.

Ordering the 13 states on a marker-anchored differentiation index (Fig. 2E–F) placed the umbrella state S9 at the differentiated extreme (median +1.20), followed by S11 (+0.48) and S3 (+0.40), and **placed S7 at the dedifferentiated extreme (median −0.73, versus +0.08 in all other states combined; Wilcoxon $p < 10^{-90}$)**. Within S7 the index was strongly left-skewed (5th percentile −7.61), indicating that a subset of cells had shut down the differentiation programme almost completely. We stress that this ordering is not a trajectory: the 13 states occupy five disconnected islands in the embedding, and we show in the Methods why a graph-based pseudotime is unreliable on this manifold.

Hallmark activity across states (Fig. 2G) showed S7 to have the **highest epithelial–mesenchymal transition score of the 13 states (47.8)** together with the **lowest apical-surface score (12.3)** and the lowest Notch activity (10.6) — a pathway profile consistent with loss of the umbrella-cell apical programme and acquisition of a migratory one.

3.4 State composition stratifies survival in TCGA-BLCA

Scoring TCGA-BLCA tumours (n = 402 with survival) with state signatures, univariable Cox models identified **S7 as the strongest adverse state** (HR 1.49 per SD, p = 0.015), with S6 also adverse (HR 1.49, p = 0.039) and S0 (HR 0.76, p = 0.043) and S9 (HR 0.64) protective (Fig. 3A). LASSO retained eight states and produced a score that separated patients into two groups with markedly different outcome (**HR per SD 4.24, 95% CI 2.48–7.25, $p = 1.21 \times 10^{-7}$; log-rank $p = 6.7 \times 10^{-4}$; C-index 0.634**; Fig. 3B–E).

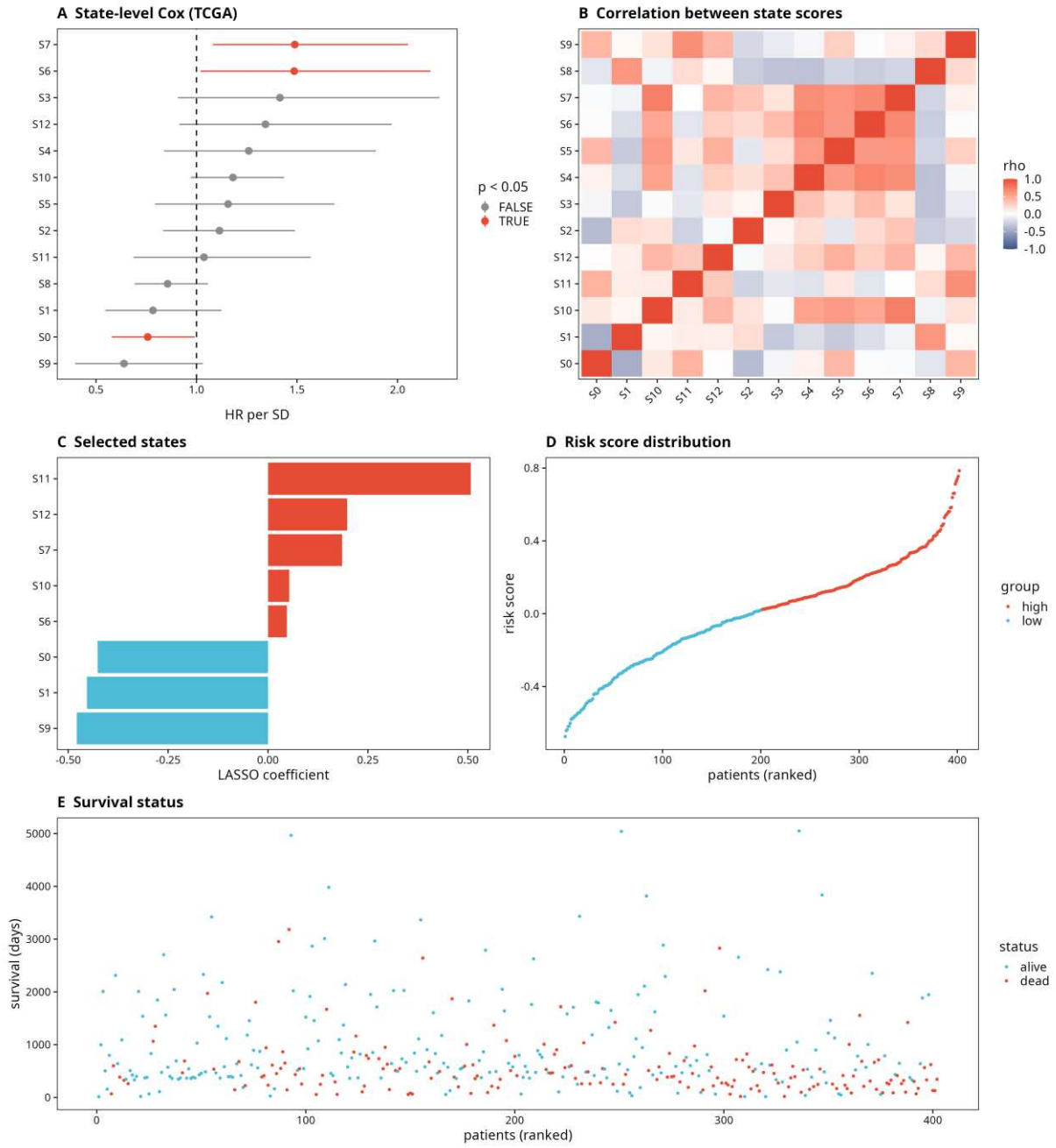


Figure 3. Prognostic model in TCGA-BLCA. (A) Univariable Cox per state. (B) Correlation between state scores. (C) LASSO coefficients. (D) Ranked risk score. (E) Survival status.

3.5 An independent deconvolution reproduces the model

NNLS deconvolution of the same tumours against per-state pseudobulk profiles agreed with the abundance proxy for the major states ($\rho = 0.70$ for S2, 0.65 for S0, 0.60 for S12, 0.59 for S9 and 0.58 for S7; Fig. 7A) and reproduced the same biology: **S7 remained the strongest adverse state (HR 97.8, p = 0.0058)**, whereas S0 (HR 0.053, p = 0.018) and S4 (HR 0.015, p = 0.017) were protective. A LASSO model on deconvolved proportions retained S0, S4 and S7 and again stratified survival (HR per SD 6.59, $p = 3.6 \times 10^{-4}$, C-index

0.606; Fig. 7B–D). That two methodologically independent quantifications converge on the same states is, in our view, the strongest internal check available to a purely computational study of this kind.

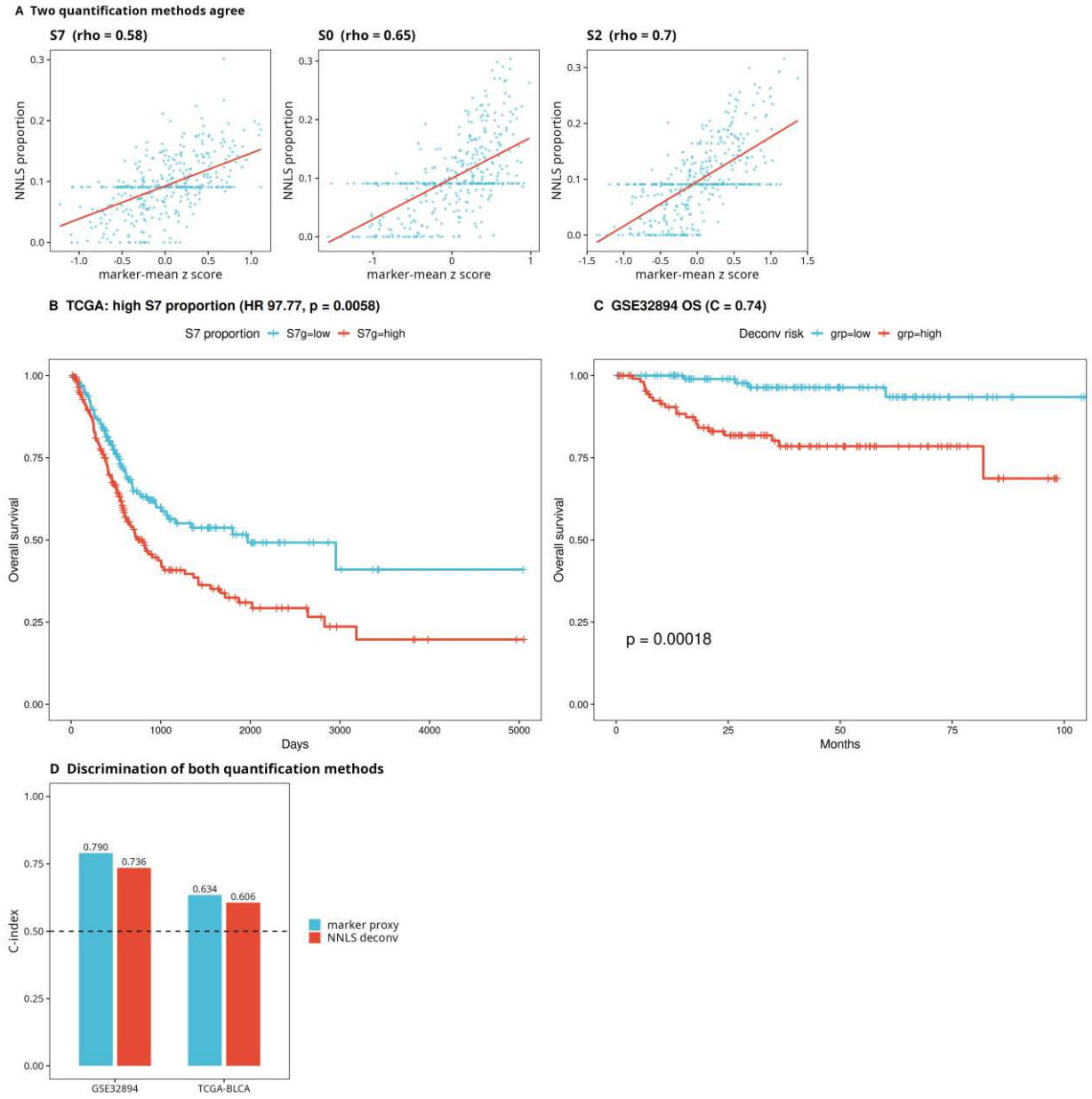


Figure 7. Deconvolution robustness. (A) Abundance proxy versus NNLS proportion for S7, S0 and S2. (B) Overall survival by S7 proportion in TCGA. (C) Overall survival by deconvolution-model risk in GSE32894. (D) C-index comparison between quantification methods.

3.6 The score recapitulates stage, grade and molecular subtype

In GSE32894 (n = 308) the score increased monotonically with stage (median −0.159 in Ta, 0.009 in T1, 0.185 in T2, 0.340 in T3; Kruskal–Wallis $\mathbf{p = 7.4 \times 10^{-14}}$) and with grade ($\mathbf{p = 3.7 \times 10^{-15}}$), and varied across molecular subtypes, being lowest in MS1b and highest in MS2b2.2 (Fig. 5A–C). Because these comparisons use no survival information, they constitute an unbiased check on the model.

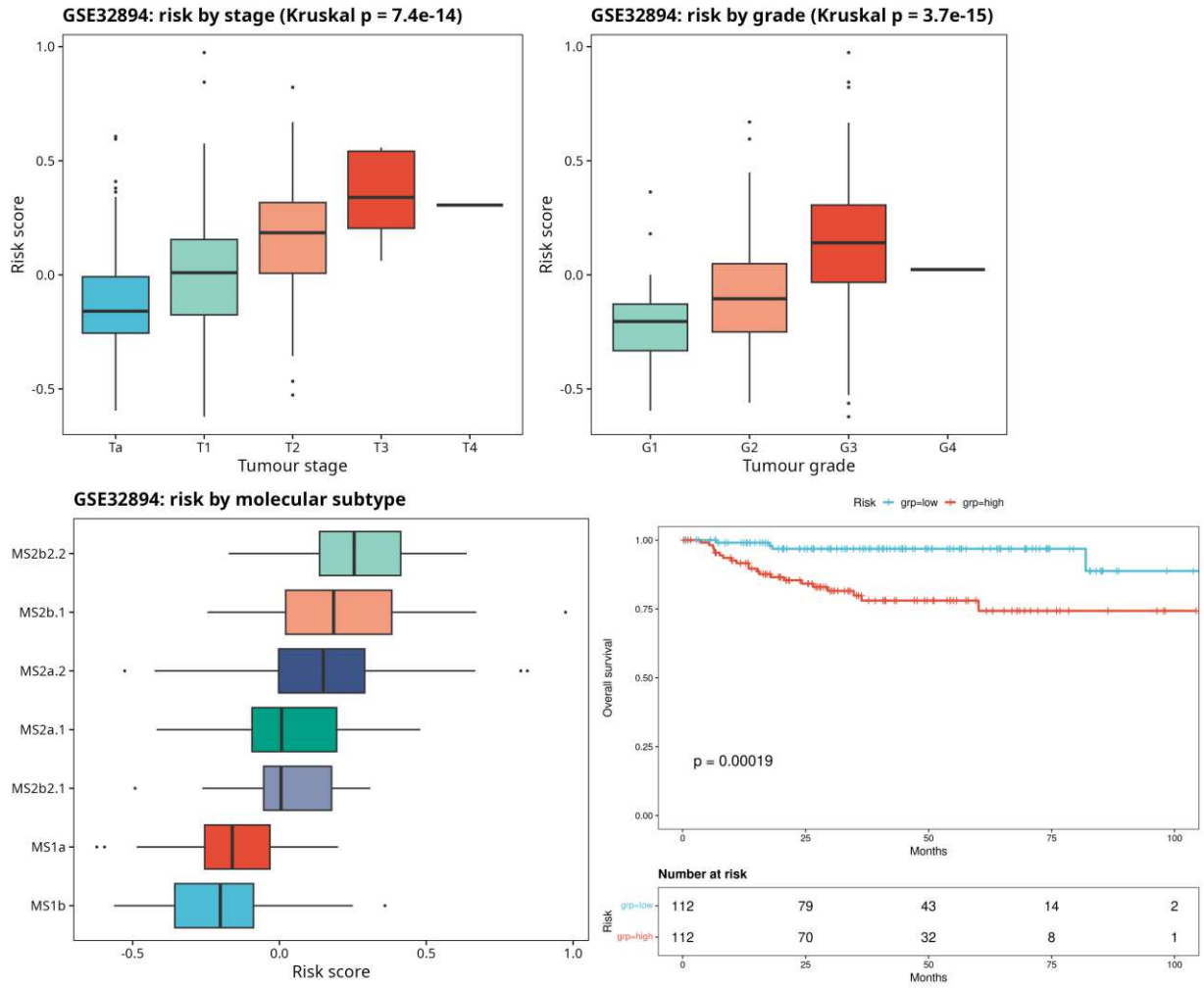


Figure 5. Clinical gradients in GSE32894. (A) Stage. (B) Grade. (C) Molecular subtype. (D) Kaplan–Meier for overall survival (C-index 0.790).

3.7 Overall survival is the endpoint the model actually predicts

Among 224 patients with complete follow-up (25 deaths), the score predicted overall survival with a median-split HR of **6.05 (2.07–17.68, p = 0.0010)**, log-rank **$p = 1.9 \times 10^{-4}$** and **C-index 0.790** (Fig. 5D); the continuous score gave a concordance of 0.79 in this cohort and 0.634 in TCGA.

3.8 Progression is only modestly predicted, and the reason is clinical

Parsing all 12 textual variants of the progression field identified **50 progressors among 308 patients**. The association was real but modest (**OR 1.40 per SD, 1.03–1.89, p = 0.030; AUC 0.610, 95% CI 0.523–0.697**; Fig. 4A–B), far weaker than the survival signal. The explanation is visible in the cohort itself: progression rates were 10.3% in Ta, 12.4% in T1, 29.4% in T2 but **0% in T3 and T4**, because patients with muscle-invasive disease underwent radical cystectomy and can no longer "progress" by definition. The endpoint is therefore subject to informative censoring exactly in the high-risk group, which attenuates discrimination. We report progression as a secondary endpoint and caution against its use as the sole validation criterion in NMIBC.

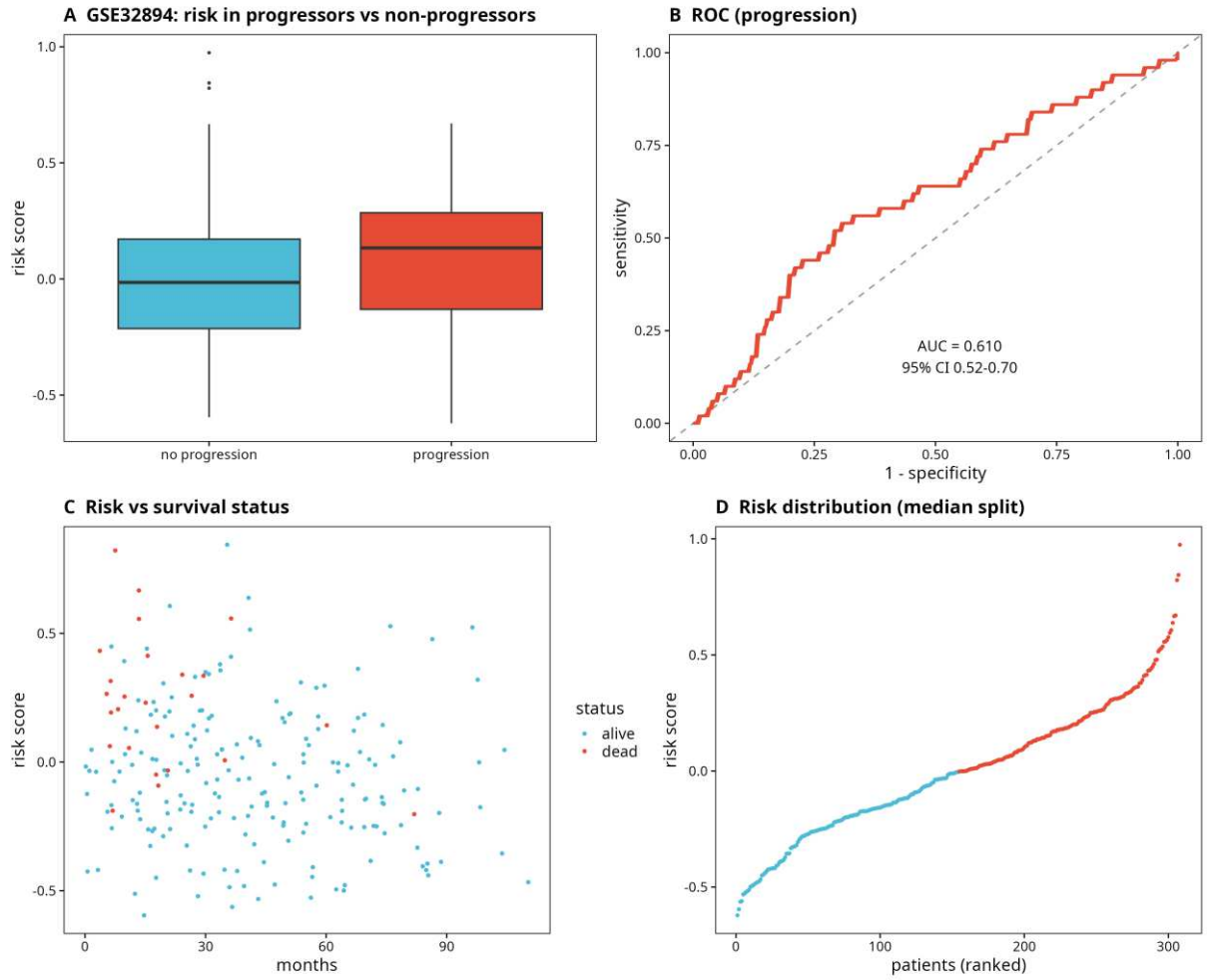


Figure 4. External validation in GSE32894. (A) Risk score in progressors versus non-progressors. (B) ROC curve for progression (AUC 0.610). (C) Risk score versus survival status. (D) Ranked risk score with median split.

3.9 S7 is a dedifferentiated, YAP/TEAD-active state that has lost KLF5-driven urothelial identity

To give the adverse state a molecular identity, we compared the 171 S7 cells with the remaining 3,709 malignant cells. 1,320 genes were differentially expressed (adjusted $p < 0.05$), strongly skewed towards upregulation (1,066 up, 184 down), indicating that S7 is a transcriptional state defined largely by what it *acquires* rather than by what it loses.

The most strongly induced genes were **CDH2 (N-cadherin, log2FC 8.3; 16.4% versus 0.2% of cells, $p = 8.0 \times 10^{-102}$)**, **WNT2 (log2FC 8.1; $p = 1.5 \times 10^{-70}$)**, **CFTR (log2FC 7.9; 63.7% versus 1.5%, $p < 10^{-300}$)** and **DNAH11 (log2FC 7.3; 50.3% versus 0.8%, $p = 1.6 \times 10^{-297}$)**. Among the downregulated genes were **KLF5 (log2FC −1.67, $p = 1.2 \times 10^{-7}$)**, PPARG and the adherens-junction component PLEKHA7 (Fig. 8A). The switch from E-cadherin to N-cadherin together with loss of PLEKHA7 is the classical cadherin switch of epithelial–mesenchymal transition, whereas the ciliary dynein DNAH11 and the apical anion channel CFTR are markers of a ciliated, umbrella-like apical specialisation. S7 therefore combines an EMT programme with

a ciliated apical programme — a combination that has been reported in basal/ciliated transitional cells of the urothelium.

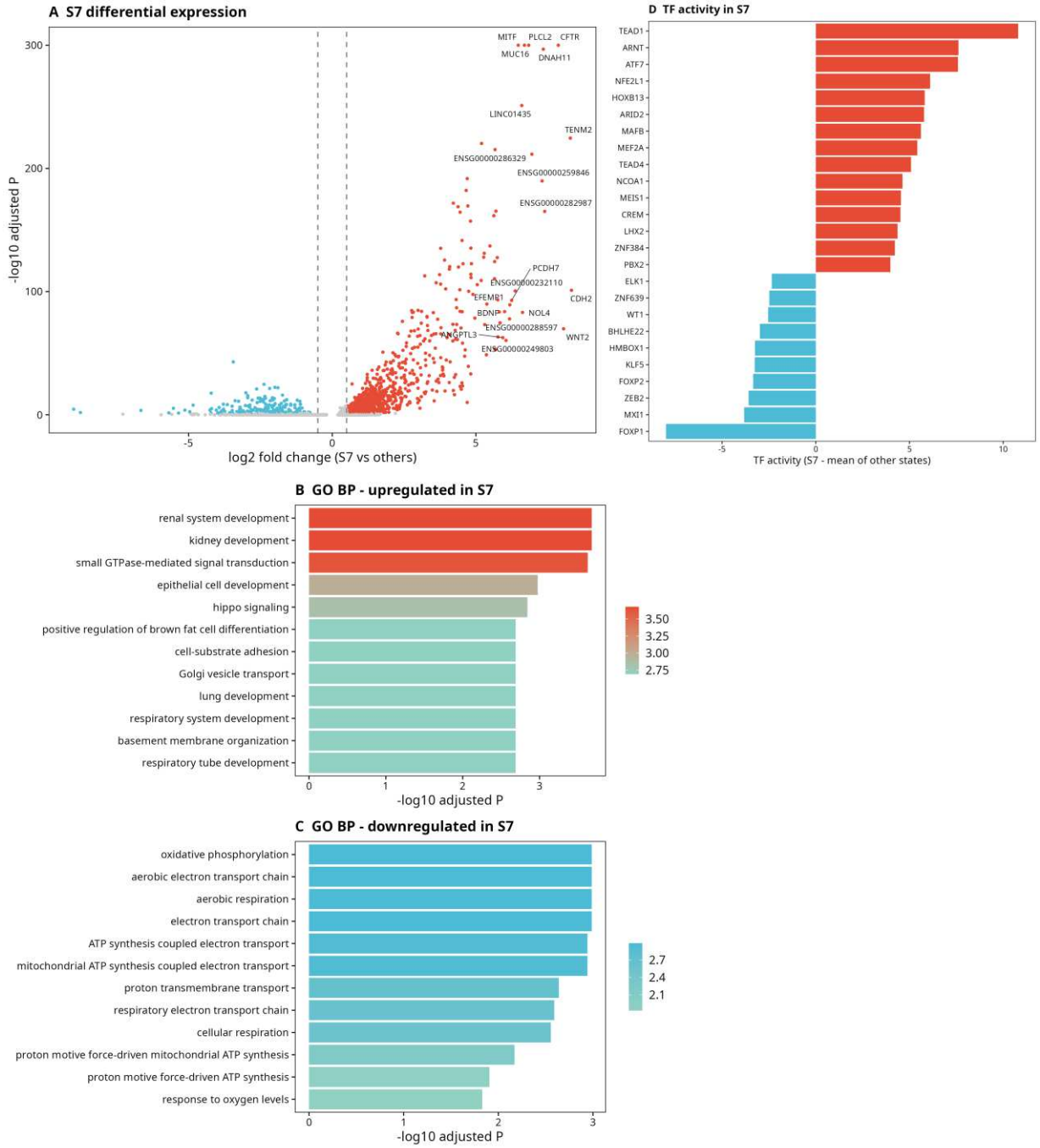


Figure 8. Molecular identity of the S7 state. (A) Differential expression versus all other malignant cells. (B) GO biological processes upregulated. (C) GO biological processes downregulated. (D) Transcription-factor activity. ---

Gene ontology confirmed this dual identity. Upregulated genes were enriched for **small GTPase-mediated signal transduction (49/850)**, renal and epithelial developmental programmes (39/850 and 28/850) and, notably, **Hippo signalling (13/850)** (Fig. 8B). Downregulated genes were almost entirely metabolic:

oxidative phosphorylation (10/164), aerobic respiration (11/164) and mitochondrial electron transport (9/164) (Fig. 8C). Loss of oxidative phosphorylation with retention of a migratory programme is characteristic of a glycolytic, de-differentiated phenotype.

Transcription-factor activity inference (decoupleR with DoRothEA, 271 TFs, 13,223 regulon edges) identified **TEAD1 as by far the most activated TF in S7 (Δ = +10.8 relative to the mean of the other states)**, followed by ARNT/HIF-1β (+7.6), ATF7 (+7.6), NFE2L1 (+6.1), HOXB13 (+5.8) and **TEAD4 (+5.1)** (Fig. 8D). The most depleted TFs were FOXP1 (−8.0), MXI1 (−3.8), ZEB2 (−3.6) and **KLF5 (−3.3)**. The gain of both TEAD paralogues, each the transcriptional effector of the Hippo pathway, matches the GO result, while loss of KLF5 activity is the single most informative observation: KLF5 is the principal driver of urothelial luminal differentiation [9], and its loss is sufficient to convert superficial umbrella cells into a basal-like, migratory phenotype in model systems [9,38,39].

Three further features confirm that whole Hippo pathway, not merely TEAD activity, is engaged: **YAP1 itself is up-regulated in S7 (log2FC 0.74; 78.9% versus 53.6% of cells, $p = 8.9 \times 10^{-8}$)**, the canonical YAP/TEAD target and tight-junction adaptor **AMOTL2 is among the most strongly induced genes in the state (log2FC 2.86; 46.8% versus 8.8%, $p = 1.7 \times 10^{-55}$)**, whereas TAZ (WWTR1) is unchanged — indicating a YAP-, rather than TAZ-driven, programme. At the same time the umbrella-cell uroplakin **UPK3A collapses in S7 (log2FC −3.70; 4.1% versus 15.0% of cells)**. Loss of uroplakin and gain of YAP therefore describe the same event from two directions: the superficial, differentiated urothelial identity is dismantled while a growth-permissive transcriptional programme takes over.

Taken together, S7 is best described as a **dedifferentiated, YAP/TEAD-driven, KLF5-low, OXPHOS-low malignant state with ectopic ciliary and EMT programmes**. This identity explains why it carries the worst prognosis of the thirteen states, and it suggests that Hippo pathway activity, rather than any single marker, is the actionable feature of high-risk early-stage bladder cancer.

3.10 High-risk tumours are immune-infiltrated but checkpoint-restrained

The risk score correlated with monocyte/macrophage ($\rho = 0.397$, $p = 1.2 \times 10^{-16}$), composite checkpoint ($\rho = 0.376$, $p = 6.5 \times 10^{-15}$), fibroblast (0.352), Treg (0.307), $CD4^+$ T (0.286), $CD8^+$ T (0.285), NK (0.244) and interferon (0.232) signatures, while B-cell infiltration was not significant ($\rho = 0.103$) (Fig. 6A–B). Single-gene associations were led by **PDCD1LG2/PD-L2 ($\rho = 0.499$, $p = 1.0 \times 10^{-26}$)**, CD274/PD-L1 ($\rho = 0.443$, $p = 1.0 \times 10^{-20}$), HAVCR2 (0.378), LAG3 (0.316), CTLA4 (0.249), TIGIT (0.238) and PDCD1 (0.228) (Fig. 6C). Dense infiltration together with high checkpoint expression is the signature of adaptive immune resistance rather than immune exclusion.

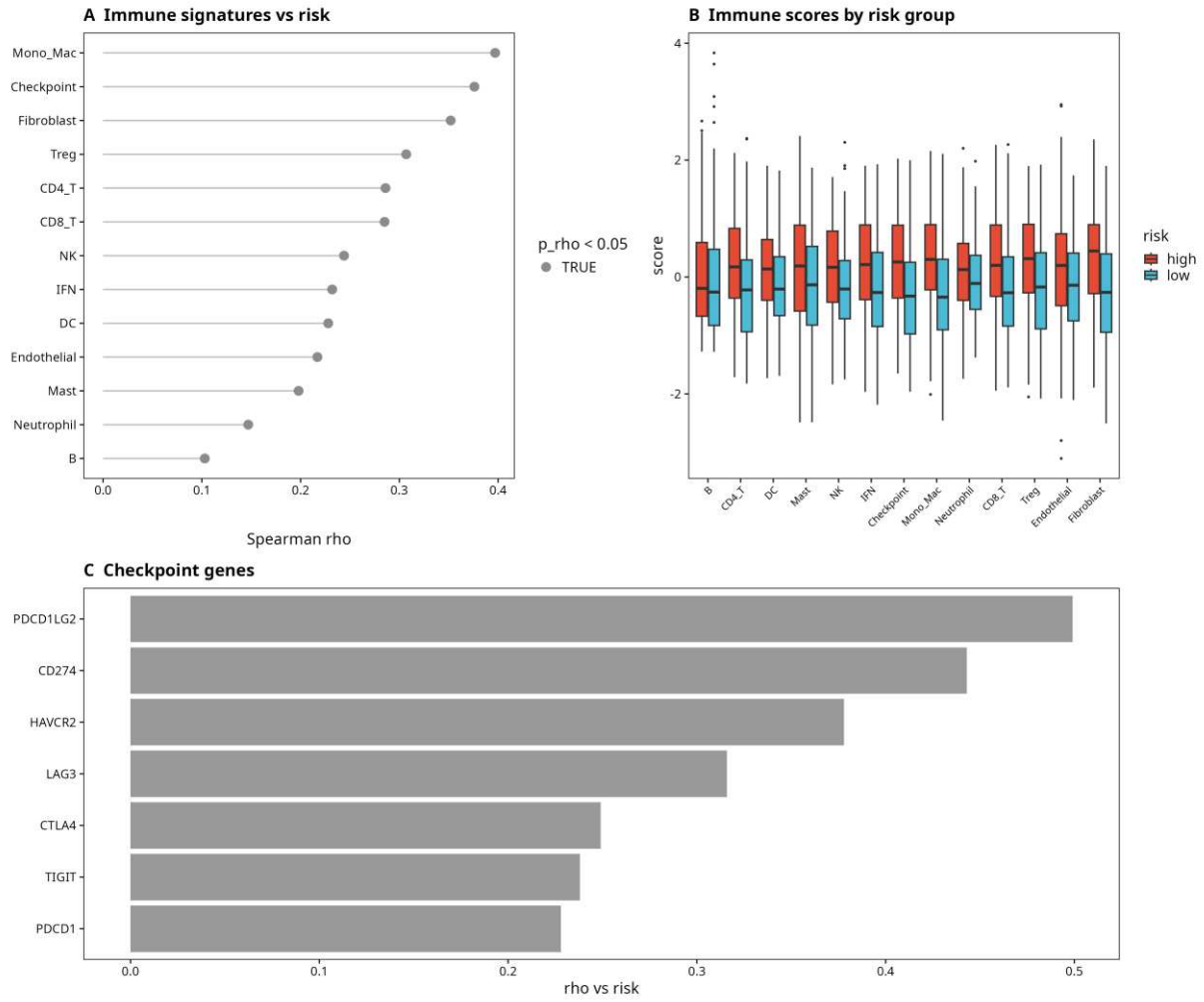


Figure 6. Immune microenvironment in TCGA-BLCA. (A) Immune signatures versus risk score. (B) Immune scores by risk group. (C) Checkpoint genes.

3.11 Supplementary: urine single-cell data

In 33 urine samples (GSE267718; 122,997 cells after quality control), epithelial cells constituted 24.2% of urine cells versus 7.4% of tumour and 1.1% of PBMC cells, and shed epithelial cells retained patient-specific states. This is a consistency check and is not used for modelling.

A separate question is whether the malignant epithelium exerts a *field effect* on adjacent normal urothelium. Scoring the 100 genes that differed between paired adjacent normal and tumour epithelium across all epithelial compartments showed that these genes do not behave as a field effect: adjacent normal epithelium scored lowest (median −0.43), tumour epithelium intermediate (−0.32) and malignant cells highest (+1.10; Wilcoxon $p = 4.4 \times 10^{-251}$) (Supplementary Fig. S2A). In other words, the "field-effect" signature is a malignant-cell programme that is largely absent from the normal compartment. Consistent with this, the same signature was only borderline associated with overall survival in TCGA-BLCA (HR 1.39 per SD, 0.95–2.04, $p = 0.087$; log-rank $p = 0.092$) (Supplementary Fig. S2B), and we do not report it as a prognostic feature.

We then asked whether the S7 protein panel is measurable in urine at all, using the longitudinal urine proteome dataset PXD025139 (six patients, 6–8 LFQ samples each). Depth was the limiting factor: individual patient tables contained only 384–623 quantified protein groups, an order of magnitude below the ~2,000 proteins achieved by dedicated urinary proteome studies. Within this shallow proteome, **the cadherin switch was the only S7 component that could be measured — CDH1 was detected in 5/6 patients and CDH2 in 4/6** — whereas **AMOTL2, YAP1, WWTR1, TEAD1, CFTR, DNAH11, WNT2, KLF5 and all four uroplakins (UPK1A, UPK1B, UPK2, UPK3A) were not detected in any patient**. Urine mass spectrometry at currently deposited depth is therefore not a usable readout for epithelial state composition; a targeted immunoassay for the cadherin switch (or for uroplakin loss) would be the appropriate translational test, and we regard that as future work rather than a result of this study.

4. Discussion

The central finding is that a prognostic score built from malignant epithelial *states* — not from cell types, not from bulk-derived signatures, and not from clinical variables — transfers from a nine-patient T1 NMIBC single-cell reference to a 402-patient TCGA cohort and to an independent 308-patient NMIBC cohort, where it predicts overall survival with a C-index of 0.79 and reproduces both the stage and grade gradients of the disease. Its value does not lie in the absolute discrimination (0.63–0.79 is modest by the standards of multi-omic signatures) but in three properties: it is derived from a biologically interpretable decomposition rather than a black-box gene list; it is reproducible under two independent quantification methods; and it forces attention onto the choice of endpoint.

That last point deserves emphasis. The conventional validation endpoint in NMIBC is progression, and our model predicts it with an AUC of only 0.61. A less careful analysis would have reported a different number and a cleaner story: an exact-match parse of the progression annotation yields ten events and an AUC of 0.83. The difference is entirely an artefact of coding — twelve textual variants of "yes" exist in the public annotation — and the honest figure is the weaker one. More importantly, even a perfectly annotated progression endpoint is compromised in this cohort, because 0% of T3/T4 patients progressed: they were cystectomised. Informative censoring of exactly the highest-risk group will depress the apparent performance of any prognostic model, and studies that validate against progression alone may be systematically underestimating their own markers.

The S7 state is the biological anchor of the model, and its identity turns out to be unusually coherent. It is not simply a "basal" or "EMT" state: it simultaneously loses KLF5 activity and gains TEAD1/TEAD4 activity, loses oxidative phosphorylation and gains a ciliated apical programme (CFTR, DNAH11), while switching from E-cadherin to N-cadherin. Each of these axes has independent support in the bladder literature. KLF5 is required for urothelial luminal differentiation and its deletion produces a basal-like, hyperproliferative urothelium; the YAP–TEAD complex drives urothelial stemness, regeneration after injury and, when constitutively active, tumour initiation, and YAP1 promotes bladder cancer progression and immune evasion [25–28]. That a single malignant state sits at the intersection of KLF5 loss and TEAD gain is therefore biologically plausible rather than a clustering artefact, and it fits the clinical observation that it carries the worst survival.

This also reframes what the model measures. It is not measuring "basal-ness" or "stemness" in the conventional subtype sense; it is measuring the abundance of a specific transitional state in which urothelial identity has been dismantled while Hippo signalling has been engaged. If that interpretation holds, the relevant

therapeutic hypothesis is not a marker but a pathway: TEAD inhibitors, several of which are in early-phase trials, would be predicted to act preferentially on high-scoring tumours. This is a testable prediction of the present work and the direction we consider most promising.

The immune analysis adds a clinically relevant dimension. High-risk tumours were not immune-deserted: macrophage, Treg, CD4, CD8 and interferon signatures all rose with the risk score. What distinguished them was checkpoint expression, led by PD-L2 and PD-L1 — notable because PD-L2 is constitutively expressed by normal as well as malignant urothelium [37]. This is the pattern of adaptive immune resistance, and it predicts that the patients with the worst prognosis may be those most likely to respond to checkpoint blockade — a hypothesis that is testable in the IMvigor and CheckMate bladder cohorts and that we cannot test here.

4.1 Limitations

We state the limitations of this work plainly, because the strength of the conclusions is bounded by them.

The study is entirely computational, and no experimental validation was performed. No cell line, no organoid, no xenograft, no patient-derived tissue, no immunohistochemistry and no functional assay accompanies the analysis. The S7 state and the risk score are therefore computational constructs: they are internally reproducible and externally replicated, but they have not been tested at the bench, and the TEAD-directed therapeutic hypothesis we raise is a prediction, not a finding.

The datasets are small, and every size limit is given here. The single-cell discovery reference comprises 13 libraries from nine T1 NMIBC tumours and four patient-matched adjacent normal samples, and after quality control 90,829 cells remained, of which only **3,880 were malignant**. The bulk training cohort was **TCGA-BLCA, n = 402** patients with survival annotation. External validation rested on a single cohort, **GSE32894, n = 308**, in which only **25 patients died** and **50 progressed**. The supplementary urine analyses used **GSE267718 (33 samples, 122,997 cells)** and **PXD025139 (six patients, 384–623 quantified protein groups each)**. These numbers are modest; with 25 events the confidence intervals around the hazard ratios are wide, and the model remains at risk of optimism even under LASSO penalisation.

The discovery cohort is early-stage only. All nine tumours were T1, and the malignant fraction varied from 4% to 70% between patients, so the state catalogue may not represent muscle-invasive disease, in which the biology is known to differ.

Quantification rests on signatures rather than a certified assay. Although non-negative least squares deconvolution reproduced the biology with independent assumptions, we did not benchmark against CIBERSORTx or MuSiC, and no targeted assay for the states exists.

External validation is single-cohort. GSE13507 could not be used because its public annotation lacks a clean progression or survival field, so no second independent survival cohort was available.

Protein-level validation is currently impossible in urine. We attempted it and it failed for a modality reason: at currently deposited depth (384–623 protein groups per patient) only the cadherin switch was measurable (CDH1 in 5/6 patients, CDH2 in 4/6), whereas AMOTL2, YAP1, WWTR1, TEAD1, CFTR, DNAH11, WNT2, KLF5 and all four uroplakins were undetected in every patient. Our protein-level claims are therefore untested.

A prospective, multi-stage cohort with long-term follow-up, and wet-laboratory work on the S7 state, are both required before any clinical claim can be made.

5. Conclusions

Malignant epithelial state composition, defined once at single-cell resolution in early-stage bladder cancer and projected onto bulk tissue, carries prognostic information that replicates in independent cohorts and predicts overall survival more reliably than it predicts progression. The field's reliance on progression as a validation endpoint deserves scrutiny, because that endpoint is structurally censored in the very patients a prognostic model is meant to identify.

6. Declarations

Ethics approval and consent to participate. Not applicable. This study used only publicly available, de-identified data sets; no new human or animal material was collected.

Consent for publication. Not applicable.

Availability of data and materials. All data analysed in this study are publicly available and no new data were generated. Every cohort, accession number and direct download location is listed below.

Single-cell discovery cohort. GEO accession **GSE222315** — https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE222315

Bulk training cohort (TCGA-BLCA). Retrieved from the GDC pan-cancer hub and the UCSC Xena TCGA hub. The four tables used were: STAR gene counts, https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-BLCA.star_counts.tsv.gz ; HiSeqV2 normalised expression, https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.BLCA.sampleMap%2FHiSeqV2.gz ; clinical annotation, https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-BLCA.clinical.tsv.gz ; and survival, https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-BLCA.survival.tsv.gz . Study level pages: UCSC Xena, https://xenabrowser.net/datapages/?cohort=TCGA%20Bladder%20Cancer%20(BLCA) ; GDC Data Portal, https://portal.gdc.cancer.gov/projects/TCGA-BLCA

External validation cohort. GEO accession **GSE32894** — https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32894

Supplementary urine single-cell cohort. GEO accession **GSE267718** — https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE267718

Longitudinal urine proteomics cohort. ProteomeXchange accession **PXD025139** — https://www.ebi.ac.uk/pride/archive/projects/PXD025139

All derived data tables underlying the figures are available from the corresponding author on reasonable request and will be included as supplementary files upon acceptance. The analysis code (R scripts) is available from the corresponding author and will be deposited in a public repository upon acceptance.

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

Funding. This work was supported by the Anhui Provincial Science and Technology Tackling Key Projects (Grant No. 202423k09020027).

Authors' contributions. YW conceived the study, performed all computational analyses, generated the figures and drafted the manuscript. HH supervised the project, provided clinical interpretation of the results and critically revised the manuscript. Both authors read and approved the final manuscript.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–249.

2. Babjuk M, Burger M, Capoun O, et al. European Association of Urology guidelines on non-muscle-invasive bladder cancer (Ta, T1, and carcinoma in situ). *Eur Urol.* 2022;81(1):75–94.

3. Sylvester RJ, van der Meijden APM, Oosterlinck W, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol.* 2006;49(3):466–477.

4. Kamoun A, de Reyniès A, Allory Y, et al. A consensus molecular classification of muscle-invasive bladder cancer. *Eur Urol.* 2020;77(4):420–433.

5. Robertson AG, Kim J, Al-Ahmadie H, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell.* 2017;171(3):540–556.

6. Weinstein JN, Akbani R, Broom BM, et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature.* 2014;507(7492):315–322.

7. Sjödahl G, Lauss M, Lövgren K, et al. A molecular taxonomy for urothelial carcinoma. *Clin Cancer Res.* 2012;18(12):3377–3386.

8. Choi W, Porten S, Kim S, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell.* 2014;25(2):152–165.

9. Bell SM, Zhang L, Mendell A, et al. Krüppel-like factor 5 is required for formation and differentiation of the bladder urothelium. *Dev Biol.* 2011;358(1):79–90.

10. Zanconato F, Cordenonsi M, Piccolo S. YAP/TAZ at the roots of cancer. *Cancer Cell.* 2016;29(6):783–803.

11. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science.* 2014;344(6190):1396–1401.

12. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol.* 2024;42(2):293–304.

13. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell.* 2019;177(7):1888–1902.

14. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 2019;16(12):1289–1296.

15. Germain PL, Lun A, Garcia Meixide C, et al. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res.* 2021;10:979.

16. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb).* 2021;2(3):100141.

17. Badia-i-Mompel P, Vélez Santiago J, Braunger J, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform Adv.* 2022;2(1):vbac016.

18. Garcia-Alonso L, Holland CH, Ibrahim MM, et al. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* 2019;29(8):1363–1375.

19. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22.

20. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* 1997;16(4):385–395.

21. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77.

22. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015;12(5):453–457.

23. Wang X, Park J, Susztak K, et al. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun.* 2019;10:380.

24. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database hallmark gene set collection. *Cell Syst.* 2015;1(6):417–425.

25. Tran W, et al. Urine single-cell RNA sequencing reveals new insights into the bladder tumour immune microenvironment. *J Exp Med.* 2024. [GSE267718; verify volume/pages]

25. Sadhukhan P, et al. YAP1 induces bladder cancer progression and promotes immune evasion through IL-6/STAT3 pathway and CXCL deregulation. *J Clin Invest.* 2024;135(2):e173468.

26. Ooki A, et al. YAP1 and COX2 coordinately regulate urothelial cancer stem-like cells. *Cancer Res.* 2018;78(1):168–181.

27. Cheng X, et al. Clinical potential of the Hippo-YAP pathway in bladder cancer. *Front Oncol.* 2022;12:925278.

28. Xia J, et al. Emerging role of Hippo signalling pathway in bladder cancer. *J Cell Mol Med.* 2018;22(1):4–15.

29. Iruzubieta P, et al. Primary cilia presence and implications in bladder cancer progression and invasiveness. *Histochem Cell Biol.* 2021;155(5):547–560.

30. Lindskrog SV, et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat Commun.* 2021;12:2301.

31. Xiao Y, et al. Integrative single-cell atlas revealed intratumoral heterogeneity generation from an adaptive epigenetic cell state in bladder cancer. *Adv Sci (Weinh).* 2024;11(24):e2308438.

32. Luo Y, et al. Single-cell transcriptome comparison of bladder cancer reveals its ecosystem.

Front Oncol. 2022;12:818147.

33. Juric I, et al. Single-cell RNA-sequencing of BCG-naïve and recurrent non-muscle-invasive bladder cancer reveals a CD6/ALCAM-mediated immune axis. *NPJ Precis Oncol.* 2025;9(1):318.

34. Fujii Y. Prediction models for progression of non-muscle-invasive bladder cancer: a review. *Int J Urol.* 2018;25(3):212–218.

35. Jobczyk M, et al. Validation of EORTC, CUETO and EAU risk stratification in prediction of recurrence, progression and death of patients with non-muscle-invasive bladder cancer. *Cancer Med.* 2020;9(11):4014–4025.

36. Zamboni S, et al. Prediction tools in non-muscle invasive bladder cancer. *Transl Androl Urol.* 2019;8(1):39–45.

37. Dowell AC, et al. PD-L2 is constitutively expressed in normal and malignant urothelium. *Front Oncol.* 2021;11:626748.

38. Chen J, et al. KRT5-high TP63-expressing urothelial basal cells act as a driver to bladder urothelium regeneration. *Stem Cell Res Ther.* 2025;16(1):296.

39. Masuda N, et al. Trp53 mutation in keratin 5 (Krt5)-expressing basal cells facilitates the development of basal squamous-like invasive bladder cancer. *Am J Pathol.* 2020;190(8):1752–1762.

40. Yang G, et al. Urothelial-to-neural plasticity drives progression to small cell bladder cancer. *iScience.* 2020;23(6):101201.

41. Guo CC, et al. Molecular profile of bladder cancer progression to clinically aggressive subtypes. *Nat Rev Urol.* 2024;21(7):391–405.

42. Riedel I, et al. Urothelial umbrella cells of human ureter are heterogeneous with respect to their uroplakin composition. *Eur J Cell Biol.* 2005;84(2–3):393–405.

43. Mathai JC, et al. Hypercompliant apical membranes of bladder umbrella cells. *Biophys J.* 2014;107(6):1273–1279.

44. Liu L, et al. Single cell sequencing reveals that CD39 inhibition mediates changes to the tumour microenvironment in bladder cancer. *Nat Commun.* 2022;13:6740.

Data resources

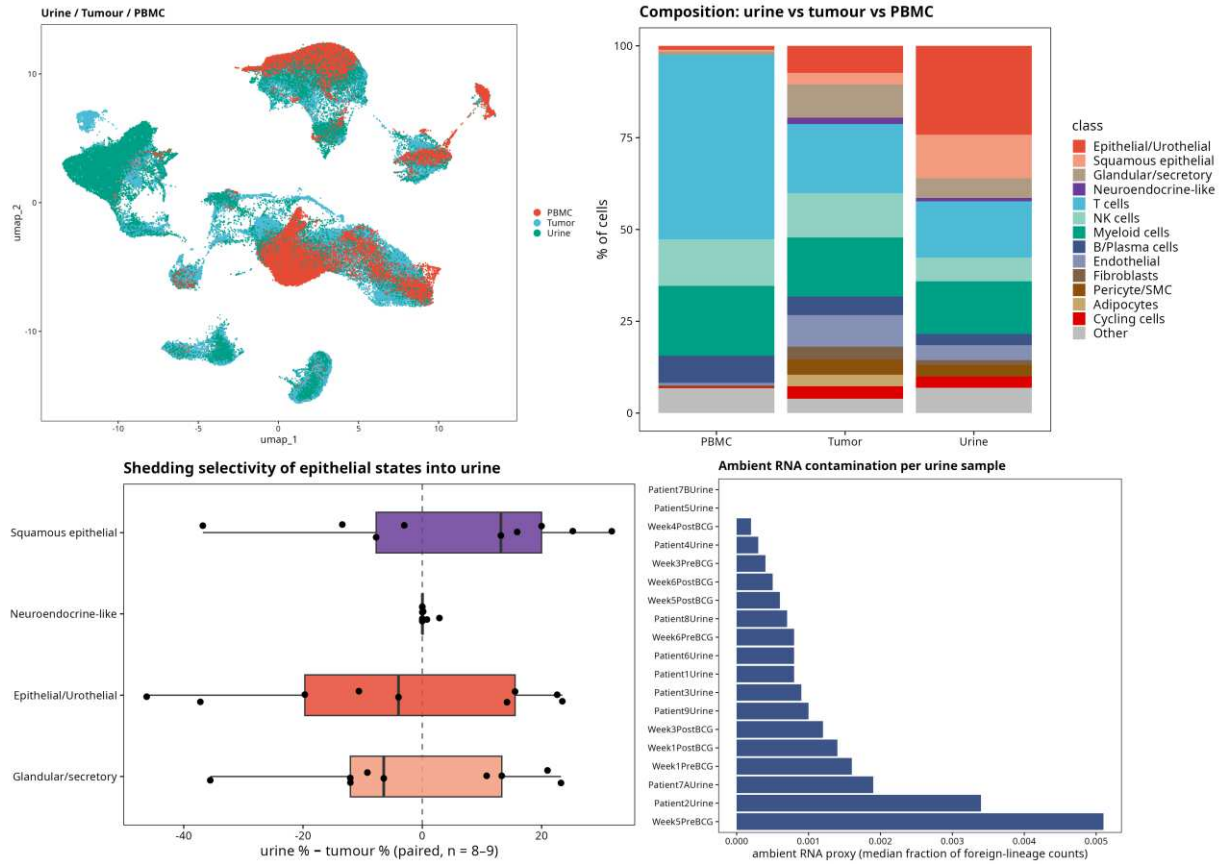
45. GSE222315 — single-cell RNA-seq of paired bladder tumour and adjacent normal tissue (NCBI GEO).

46. GSE32894 — urothelial carcinoma cohort with clinical follow-up (NCBI GEO / ArrayExpress E-GEOD-32894).

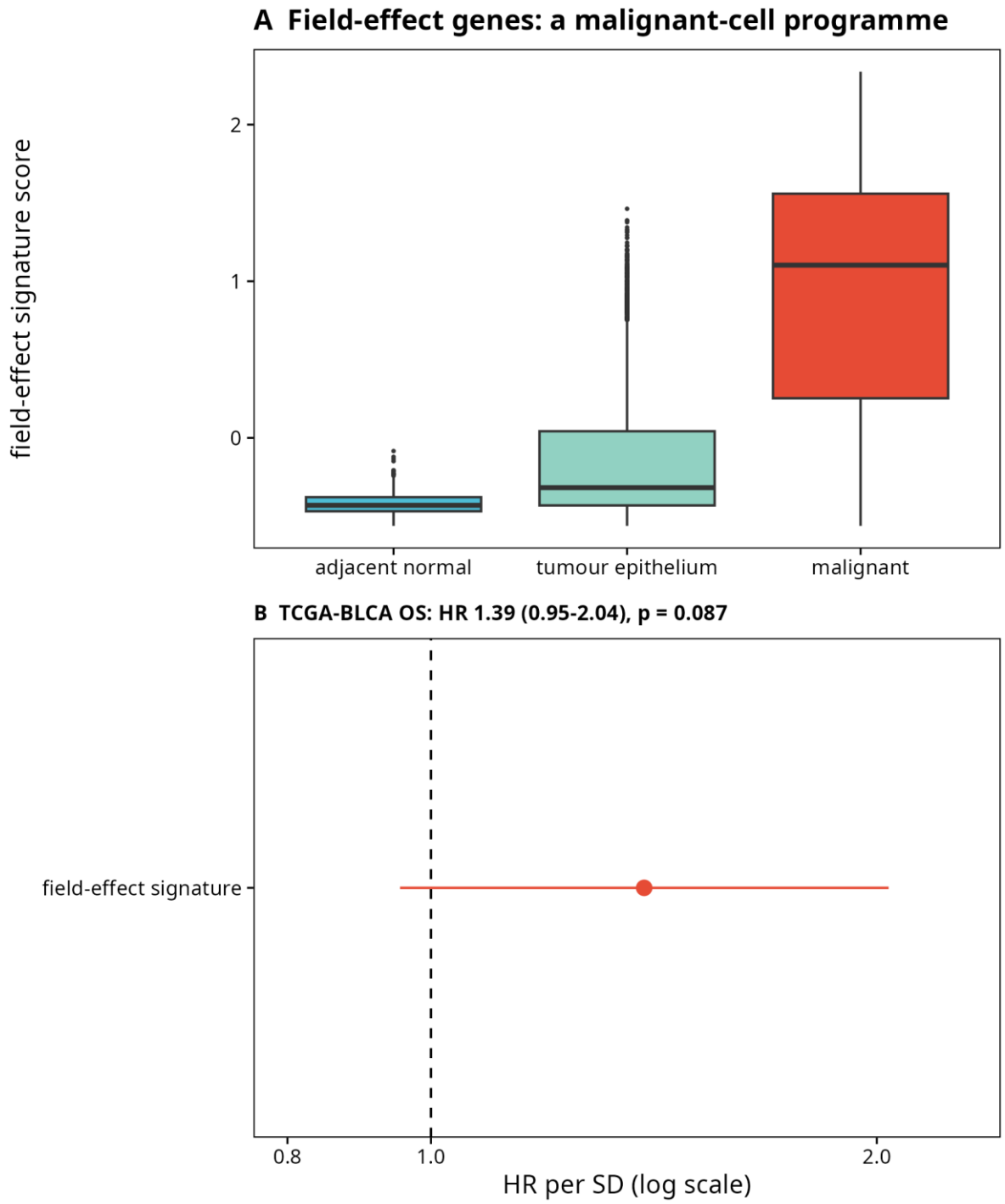
47. GSE267718 — urine single-cell RNA-seq of patients undergoing BCG therapy (NCBI GEO).

48. TCGA-BLCA — Bladder Urothelial Carcinoma, The Cancer Genome Atlas; expression, clinical and survival tables obtained from the GDC hub / UCSC Xena.

Supplementary Figures



Supplementary Figure S1. Urine single-cell consistency. (A) UMAP of 122,997 urine cells coloured by sample type (urine, tumour, PBMC). (B) Cell-class composition of the three sample types; epithelial cells constitute 24.2% of urine cells versus 7.4% of tumour and 1.1% of PBMC cells. (C) Paired shedding selectivity of epithelial states (urine minus tumour proportion, n = 8–9 pairs); no state shows a systematic shedding preference (Wilcoxon paired p = 0.18–0.91). (D) Ambient RNA contamination per sample, indicating that the shallow depth of urine libraries, not ambient signal, limits copy-number inference in urine.



Supplementary Figure S2. Field-effect signature. (A) Score of the 100 field-effect genes across epithelial compartments: adjacent normal epithelium (median −0.43), tumour epithelium (−0.32) and malignant cells (+1.10; Wilcoxon versus adjacent normal $p = 4.4 \times 10^{-251}$). The genes behave as a malignant-cell programme rather than a field effect. (B) In TCGA-BLCA the same signature shows only a trend towards worse overall survival (HR 1.39 per SD, 0.95–2.04, p = 0.087; log-rank p = 0.092), and we therefore do not claim it as a prognostic feature.