BMC Cancer

RESEARCH

Open Access

Novel Gene expression-based Risk Stratification tool predicts recurrence in Non-muscle invasive Bladder cancer

Srivatsa N[1†], Hari PS[2†], Rahul P[1], Lista Paul[3], Durgadevi Veeraiyan[2], Ambili Narikot[2], Vidya Veldore[4], Nishtha Tanwar[4], Peddagangannagari Sreekanthreddy[4], Hitesh Goswami[4], Rekha V. Kumar[5], B S Srinath[3] and Aruna Korlimarla[2*]

Abstract

Background Bladder cancer represents a heterogeneous disease with distinct clinical challenges. Non-muscle invasive bladder cancer (NMIBC) typically presents as indolent and slow-growing, yet a critical clinical challenge remains: identifying which patients will progress to muscle-invasive disease requiring radical interventions. Early detection of progression propensity is essential, as once muscle invasion occurs, the risk of distant metastasis increases substantially, and treatment shifts from conservative TURBT (Transurethral Resection of Bladder Tumor) to aggressive surgical interventions with significant morbidity. Current risk stratification methods fail to adequately predict this transition in approximately 30% of cases, highlighting the urgent need for more accurate prognostic tools.

Objective This retrospective study aimed to develop and validate a transcriptomics-based mRNA score for predicting early NMIBC recurrence, comparing its performance against traditional risk stratification methods.

Methods We analyzed mRNA expression profiles from primary retrospective NMIBC tumor specimens ($n=25$) collected between [2018–2022]. Traditional risk stratification tools, including EORTC scoring, were applied alongside our novel mRNA-based risk score to evaluate predictive accuracy for recurrence.

Results The transcriptomics-based mRNA score demonstrated a median prediction accuracy of 90% across 10,000 resampling iterations for predicting early NMIBC recurrence, significantly outperforming traditional EORTC risk scores. Our comprehensive gene set identified 435 differentially expressed genes associated with recurrence. Kaplan–Meier analysis showed significantly different recurrence-free survival between high and low mRNA risk score groups (Bonferroni corrected p-value < 0.0001).

Conclusions This retrospective analysis confirms that mRNA expression-based risk stratification provides superior predictive accuracy compared to conventional clinicopathologic risk tools. Implementation of this gene signature could potentially reduce over-investigation and improve surveillance cost-effectiveness after TURBT in patients with primary high-risk NMIBC. These findings may transform the clinical management paradigm by enabling more personalized follow-up protocols based on molecular risk assessment.

Keywords Bladder Cancer, Targeted therapy, Classifier Signature, Subtypes, EORTC

†Srivatsa N and Hari PS contributed equally to this work.

*Correspondence:
Aruna Korlimarla
aruna.k@ssnccpr.org
Full list of author information is available at the end of the article

Introduction

Bladder cancer presents a significant clinical challenge, with non-muscle invasive bladder cancer (NMIBC) accounting for approximately 75% of cases [1]. Despite its generally favourable prognosis compared to muscle-invasive disease, NMIBC demonstrates remarkable heterogeneity in clinical outcomes. The recurrence rates of 40–60% and progression rates of 10–20% to muscle-invasive disease highlight the critical need for improved risk stratification tools [2, 3].

Even after complete transurethral resection (TUR) of exophytic lesions, NMIBC can recur in up to 80% of cases during long-term follow-up [4]. This high recurrence rate necessitates effective adjuvant treatments and surveillance protocols to prevent both recurrence and the potentially life-threatening progression to muscle-invasive disease. A personalized, risk-based therapeutic approach is essential for optimizing treatment efficacy while maintaining quality of life [3].

Current risk stratification methods rely primarily on clinicopathological parameters, which often fall short in capturing the molecular diversity underlying NMIBC's variable clinical behavior. This limitation underscores the urgent need for molecular biomarkers that can more accurately predict recurrence and progression, enabling truly personalized treatment decisions and improving patient outcomes.

The conventional stratification of NMIBC patients based on clinicopathological parameters such as tumor stage, grade, and presence of carcinoma in situ (CIS) has been the cornerstone of treatment decision-making. The European Organization for Research and Treatment of Cancer (EORTC) scoring system was introduced into the EAU guidelines in the year 2008 and American Urology Association (AUA) guidelines (2016) which included both clinical and pathological factors to predict the recurrence and progression patterns of each patient [4, 5]. However, this approach lacks precision in predicting individual patient outcomes and guiding personalized therapeutic interventions. Consequently, there has been a growing interest in unravelling the molecular landscape of NMIBC to identify robust biomarkers that can refine risk stratification and prognostication [3, 5].

These guidelines are currently the most widely used and validated prediction model in NMIBC [4]. These guidelines stratify NMIBCs into low, intermediate, and high risk for treatment options. The scoring system is based on clinical and pathological factors that are commonly assessed and have been found to be of prognostic importance in previous publications. Positive predictive value of score is about 70% across all studies. The EORTC risk calculator, derived from pooled data sets of patients involved in European clinical trials, may not accurately represent non-European patients or real-world scenarios [3, 4, 6]. Undertreatment of the patients based on the current scoring systems leads to disease recurrence and overtreatment leads to notable adverse effects and reduction of quality of life of patients. This necessitates a scoring system with better recurrence predictivity.

Recent advancements in high-throughput genomic technologies have revolutionized our ability to characterize the genomic alterations driving tumorigenesis. These studies categorized Western patients and predicted progression with an accuracy of around 80–85% when combined with EORTC score. Large-scale genomic profiling studies have uncovered distinct molecular subtypes within NMIBC, each associated with unique biological features and clinical behaviours. These molecular subtypes offer a promising avenue for refining NMIBC classification beyond traditional histopathological criteria [7, 8, 9]. In this study, we conducted RNA sequencing on tumors from Indian patients with NMIBC, developed machine learning models, and compared them with the EORTC and AUA scores to assess their predictive utility with respect to early recurrence within one year. The signature was correlated with recurrence status in public omics datasets of patients with NMIBC.

Materials and methods

Sample selection, library preparation, and sequencing

All patients who had confirmed diagnosis of NMIBC diagnosed at the Uro Oncology department (2018–2022) of the tertiary cancer care center that caters predominantly to patients of south Indian states were included in the study. Surgically excised specimens of NMIBC, in the form of formalin fixed paraffin embedded blocks (FFPE) were collected retrospectively. Informed consent for use of the material for research was obtained by the Institutional ethics committee. Only tumour blocks with >50% tumor content as estimated by a pathologist were chosen for analysis. Patients were followed up for one and half year period after completion of treatment. Overall methodology workflow is provided in supplementary Fig. 1.

Tumor RNA extraction from FFPE blocks was performed using the All Prep FFPE DNA/RNA Kit (Cat. No. 80234, Qiagen, Valencia, CA, USA). RNA samples passing quality control (QC) were further processed for library construction, involving fragmentation, adapter ligation, and amplification. The Agilent RNA Prep with Enrichment Kit (Cat. No. 5191–6874) [10] was used for RNA-Seq library preparation. Prepared libraries were assessed for fragment size and concentration via QC analysis. The prepared libraries underwent QC analysis for the detection of library fragment size and concentration. A qualified NGS library had at least 10-nM concentration with a single distinct peak of

approximately 300 bp. Paired-end sequencing (2 × 150 read length) was conducted on the Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA) to achieve a median coverage of 200X. Quality of the NGS libraries was confirmed using Qubit and Bioanalyzer.

Alignment of reads to reference genome, normalisation and differential expression analysis

Quality check of raw Fastq files were performed using FastQC [11]. Trimmomatic was used to remove bad quality reads [12]. The processed fastq files then were mapped to GRCh38. HISAT2 [13] was used for alignment and the read counts of each gene were obtained using Feature Counts [14]. The counts were normalised using DESeq2 [15]. Further, the genes with greater than 5 counts across all samples were retained. The differentially expressed genes between recurrent and non-recurrent patients were selected based on the p-value < 0.05 and log2 FC greater than 1 or log2 FC less than −1 cut offs.

Machine learning model building and comparison of model accuracy with EORTC and AUA scores

Partial Least Squares Discriminant Analysis (PLS-DA) was used to build a machine learning model using mixOmics package [16] and Prediction Analysis of Microarrays (PAM) model was built using pamr package [17]. Differentially expressed gene matrix was used as independent variable and patient recurrence status was the outcome variable. Fivefold cross validation was applied with both models. For PLS-DA, to determine the optimal number of latent components, we employed fivefold cross-validation using the perf() function. Model performance was assessed based on overall classification error rate and balanced error rate (BER). Error rates with different distance measures were also calculated. To assess model stability, PLS-DA model with 10 components was run 10,000 times using random seeds. For each run, 60% of samples were randomly selected for training and the remaining for testing. Prediction accuracy was calculated as the proportion of correct predictions. Classification accuracy of both models was compared against EORTC and AUA scores using confusion matrices and fisher's exact test. Average expression of upregulated genes was taken across samples to determine the optimal cutpoint using the maximally selected rank statistics from the 'maxstat' R package [18]. This cut point was used for Kaplan–Meier survival analysis. log-rank p value was used for significance analysis. The same analysis was performed with average expression of downregulated genes.

Table 1 Clinical Characteristics of entire cohort patients with NMIBC

Clinical Characteristics *N* = 32	Category	N (%)
Age (Median)		65
BCG Treatment		27 (84.37)
Tumor Stage	T1 Stage	19 (59.37)
	TA Stage	13 (40.63)
Recurrence Status	Recurrent	16 (50)
	Non-Recurrent	16 (50)
EORTC Risk Score	RS-High	4 (12.5)
	RS-Intermediate	24 (75)
	RS-Low	4 (12.5)
AUA Risk Score	AUA-High	11(34.37)
	AUA-Intermediate	16(50)
	AUA-Low	5(15.63)
Grade	G1	22(68.75)
	G3	10 (31.25)
Cis	Present	3(9.37)
Focality	Unifocal	15(47)
	Multifocal	17(53)

Gene expression signature correlation analysis with publicly available NMIBC datasets

Gene expression signature obtained was validated using publicly available datasets including and EGAS00001004693 [7], GSE13507[8] and GSE154261 [9]. Normalised expression data was downloaded from respective datasets from gene expression ominibus (GEO) and European Genome-phenome Archive (EGA) and only differentially expressed genes of our signature was looked at. Clinical data of these datasets were also downloaded from the same site. Average expression of upregulated and downregulated genes of our sample set were compared with clinicopathological parameters of those cohorts. Wilcoxon Rank-Sum Test was used to identify significant association with clinicopathological parameters.

Pathway Enrichment Analysis

Pathway analysis of differentially expressed genes was performed using Enrichr [19]. Significant pathways of Reactome, KEGG and WikiPathways were visualised using Enrichr - KG.

Results

Samples and clinicopathological data

Of the 32 patients, data from RNA sequencing was obtained from 25 patients; the remaining did not meet the criteria pertaining to tumor percentage, quality

of the RNA, and the library concentrations. Various QC parameters used have been explained in the methods section. Clinicopathological data of entire clinical set is provided in Table 1

The differential expression analysis revealed distinct patterns, and the prediction models achieved an accuracy of 90%

Four hundred thirty-five genes were differentially expressed between patients with recurrence and those without. (p-value <0.05 & log2 FC >1 | log2 FC <-1). Of these, 73 genes were upregulated, and 362 genes were downregulated in patients with recurrence. To visualise the genes, differentiate between recurrent and non-recurrent patients, heatmap was constructed on scaled expression values of 435 genes. Scale ranged from -2 to $+2$. -2 being lowest expressed and $+2$ being highest expressed in relation to across samples. Patients who recurred show consistent high expression of 73 genes, while patients with non-recurrent tumours show down regulation of these genes. Non- recurrent patients show high expression of 362 genes, while these genes were downregulated in recurrent patients. Volcano plot was drawn to visualise the distribution of p-values and fold changes of genes between recurrent and non-recurrent cases (Fig. 1a,b). Figure 1a represents the heatmap for Z-score of 435 genes differentiating recurrent and non-recurrent patients. Figure 1b is volcano plot showing log2

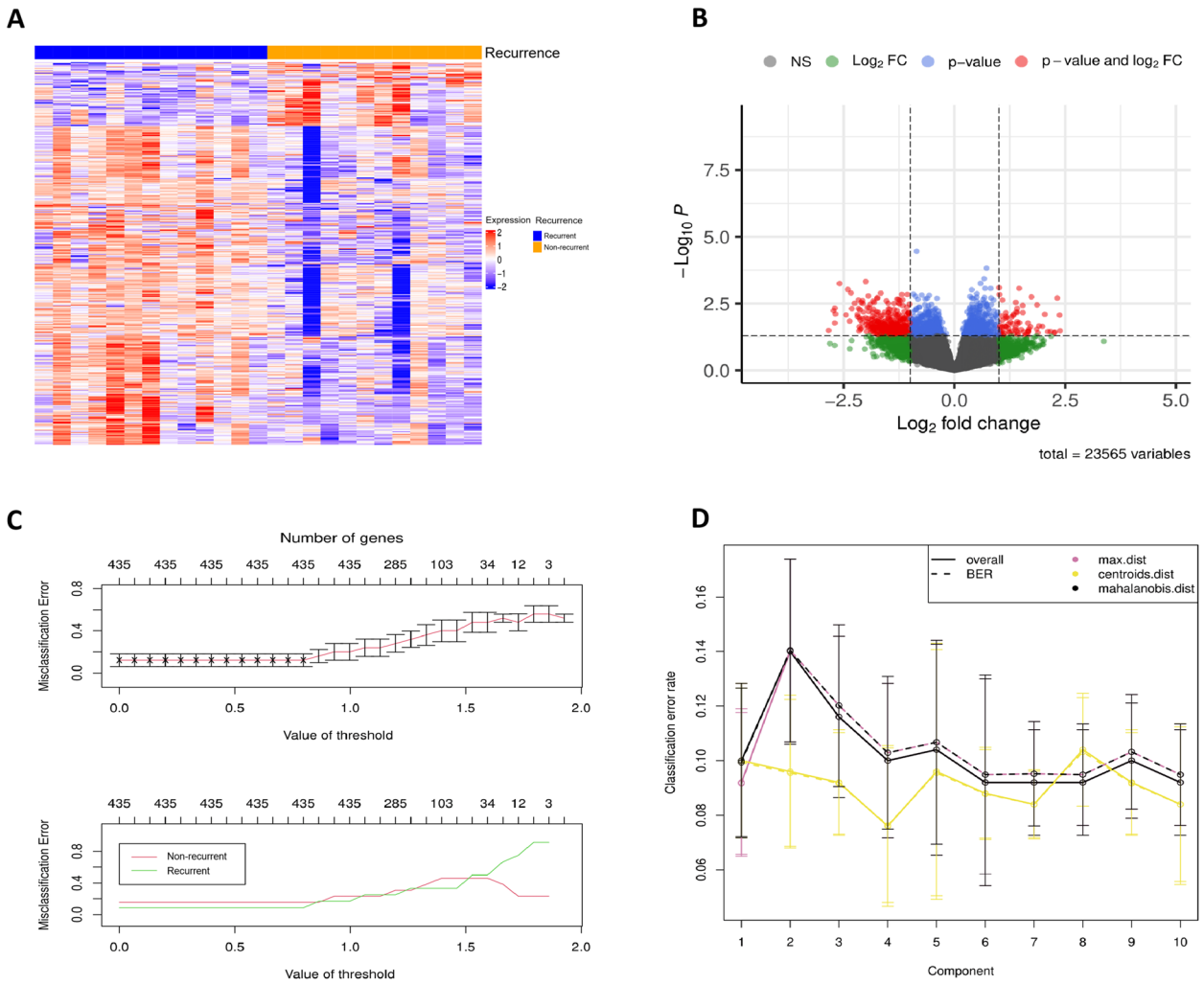

Fig. 1 Stratification of NMIBC patients. a: Heatmap representing Z-score of differentially expressed genes of recurrent and non-recurrent NMIBC tumors. b: Volcano plot illustrating the transcriptome-wide distribution of $\log_2$ fold changes and $-\log_{10}$ p-values between recurrent versus non-recurrent patient groups. c: PAM cross validation error plot. d: PLS-DA cross-validation error plot. Error rate is shown across the components with different distance methods. The overall error rate (OER) and balanced error rate (BER) for the three different distance metrics across the first ten components are depicted

fold change and p value difference between recurrent and non-recurrent patients.

Two prediction methods including PAM and PLS-DA were used to check the ability of differentially expressed genes to predict the subtypes. Here we have used 435 differentially expressed genes as input for PAM and found that these genes could differentiate recurrent and non-recurrent patients with overall 90% accuracy of recurrence prediction via fivefold cross validation (Fig. 1c). For PLS-DA as also we used 435 differentially expressed genes as input and checked different distance measures while using fivefold cross validation analysis. Misclassification error plot yielded 92% accuracy of recurrence prediction by centroid distance method via fivefold cross validation with 4 components (Fig. 1d). The model demonstrated a median prediction accuracy of 90% across 10,000 resampling iterations of training and test sets (Supplementary Fig. 2). Model calibration parameters and performance matrices are given in supplementary Table 1.

Pathway analysis of differentially expressed genes

To identify functional aspect of the signature, pathway analysis was performed. Significant upregulated KEGG pathways in patients with recurrence include "Systemic Lupus Erythematosus", "Alcoholism", "Neutrophil Extracellular Trap Formation", "Transcriptional Misregulation in Cancer", "Shigellosis", "Nuclear Receptor Meta-Pathway" and "Fatty Acid Metabolism" (Fig. 2a). Green bubbles indicate genes, pink indicates WikiPathways, purple indicates Reactome pathways and grey indicates the KEGG pathways. Many histone proteins were common across various pathways including "Systemic Lupus Erythematosus", "Alcoholism", "Neutrophil Extracellular Trap Formation", "Transcriptional Misregulation in Cancer" and "Shigellosis". As annotated by KEGG: H3C8, H3C13, H3C10, H3C3 and H3C4 genes were associated with "Shigellosis" as well as "Transcriptional Misregulation in Cancer". H3C8, H2BC6, H2AC4, H2BC3, H2AB1, H2AC16, H2AC11, H2C13, H2BC11, H3C10, H3C3, H2C2 and H3C4 were associated with "Neutrophil Extracellular Trap Formation" and "Systemic Lupus Erythematosus Pathways". Interestingly, H3C8, H2BC6, H2AC4, H2BC3, H2AB1, H2AC16, H2AC11, H2C13, H2BC11, H3C10, H3C3, H2C2 and H3C4 were also associated with alcoholism along with GRIN2D. Similarly, Reactome pathways show: H3C13, H2AC4, H2BC3, H2BC11 and H2AB1 were associated with "DNA Methylation", "Meiotic Recombination", "RNA Polymerase I Promoter Opening", "SIRT1 Negatively Regulates rRNA Expression" pathways. ARTN was another gene associated with "Meiotic Recombination". Wicki pathways show above histone modifiers as well as THBD, SLC7 A5, SLC6 A8, SCD and FASN genes associated with "Nuclear Receptor Meta Pathway". SCD and FASN genes associated with "Fatty Acid Biosynthesis".

Interestingly we found that significantly downregulated pathways in patients with recurrence include "Cell Adhesion Molecules", "Calcium Signaling Pathway", "Extracellular Matrix Organisation", "PIK3 K-AKT Signaling Pathway", "Phospholipase D Signaling Pathway", "MicroRNAs in Cancer", "Focal Adhesion", Vascular Smooth Muscle Contraction", "Glycine, serine and threonine metabolism" and "RAP1 signaling pathway" (Fig. 2b). The genes involved in pathway annotated as "Cell Adhesion Molecules" by KEGG

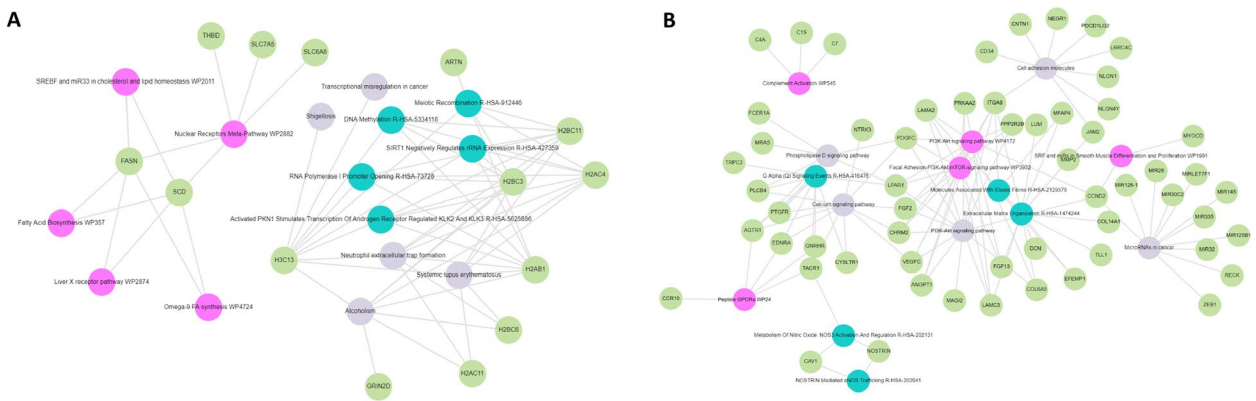

Fig. 2 Pathway analysis of differentially expressed genes. a: Upregulated genes in recurrent NMIBC tumors in enrichr-KG network. Green bubbles indicate genes, pink bubbles indicate WikiPathways, purple bubbles indicate Reactome pathways and grey indicate the KEGG pathways. b: Downregulated genes in non-recurrent tumours in enrich-KG network. Genes were labelled as green color and connecting pathways from different sources are labelled with different colors. Green bubbles indicate genes, pink bubbles indicate WikiPathways, purple indicate Reactome pathways and grey indicate the KEGG pathways

include NLGN4Y, NLGN1, NEGR1, CNTN1, ITGA8, PDCD1LG2, LRRC4 C, CD34 and JAM2. The genes involved in pathway annotated as "Calcium Signaling pathway" by KEGG include CHRM2, PTGFR, EDNRA, CYSLTR1, PLCB4, PDGFC, AGTR1, VEGFC, TACR1 and FGF2. The genes involved in pathway annotated as "PIK3 K-AKT Signaling Pathway" by KEGG include CHRM2, PRKAA2, LAMA2, ANGPT1, LAMC3, MAGI2, LPAR1, VEGFC, FGF2, CCND2, PPP2R2B, PDGFC, ITGA8 and COL6 A5. The genes involved in pathway annotated as "Extracellular Matrix Organisation" by Reactome include LAMA2, COL14 A1, LAMC3, LUM, MMP2, DCN, MFAP4, EFEMP1, ITGA8, TLL1, COL6 A5, FGF13 and JAM2. The genes involved in pathway annotated as "G Alpha (Q) Signaling Events" by Reactome include PTGFR, EDNRA, CYSLTR1, PLCB4, TRPC3, AGTR1, GNRHR and TACR1. The genes involved in pathway annotated as "Molecules Associated with Elastic Fibres" by Reactome include MFAP4, EFEMP1 and ITGA8. The genes involved in pathway annotated as "Complement activation" by WikiPathways include C4A, C1S and C7.

Gene expression analysis predicts recurrence with a greater accuracy as compared to EORTC and AUA score

EORTC risk stratification classifies 83% who were recurrent and 85% who were non recurrent into intermediate risk group. 15% of low-risk group of patients by EORTC did not have recurrence and 17% of high-risk group had recurrence (Fig. 3a). AUA risk stratification classifies 33% of patients who were recurrent and 69% of patients who were non-recurrent into intermediate risk group (Fig. 3b). 23% of low-risk group by AUA fall into non-recurrent group, while 67% of high-risk patients fall in recurrent class. 8% of high-risk patients had no recurrence.

The analysis above necessitates a binary classifier as most of the intermediate group fall in to either high or low group. PAM analysis and PLS-DA analysis showed superior performance compared to EORTC and AUA and classified patients with a median accuracy of 90%. PAM fivefold cross validation model classifies patients with 90% accuracy for their recurrence status (Fig. 3c, Fig. 1c, d, supplementary Fig. 2). Fisher's exact test p-value for PAM prediction with recurrence status was less than 0.0001. Survival analysis showed significant bonferroni corrected p-value with higher average expression

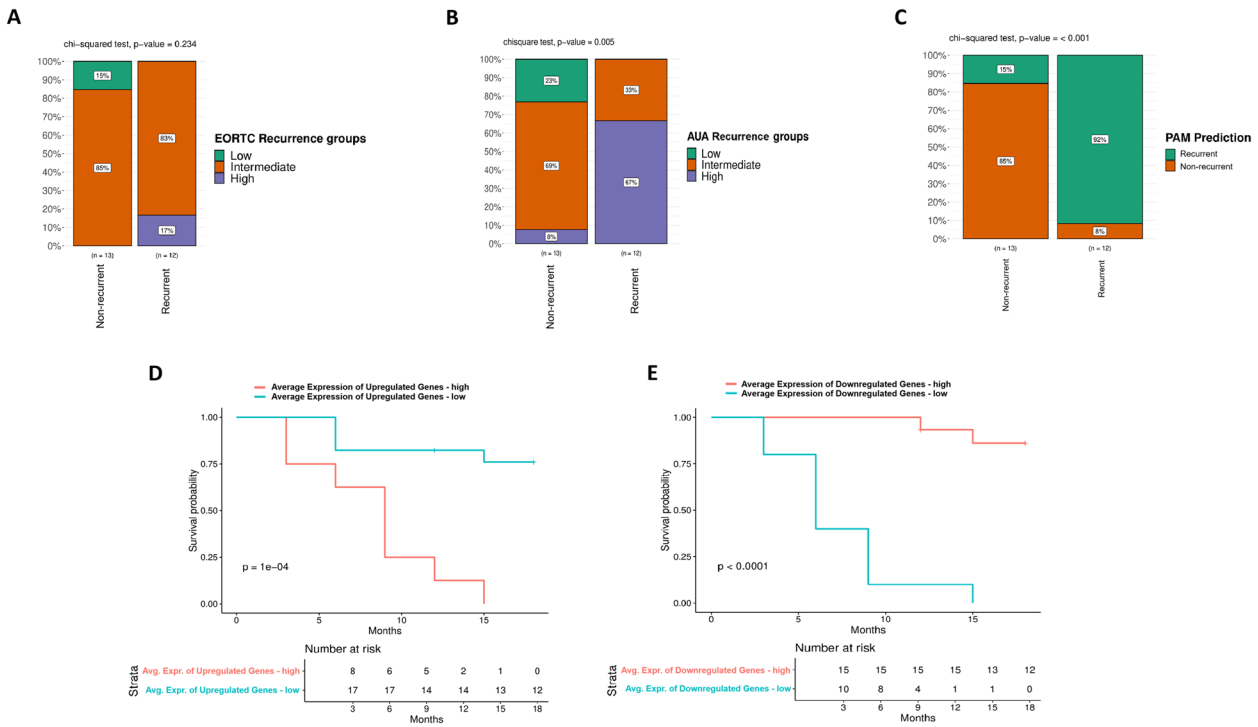

Fig. 3 Comparison of genes expression with EORTC, AUA and clinical parameters. a: Comparison of EORTC risk groups and recurrence status of the patients. b: Comparison of AUA risk groups and recurrence of the patients. c: Comparison of gene expression signature prediction (PAM) and recurrence status of the patients. d: Kaplan–Meier curve stratifying patients based on average expression of upregulated genes. e: Kaplan–Meier curve stratifying patients based on average expression of down-regulated genes

of upregulated genes resulting in poor survival of the patients (Fig. 3d), while higher average expression of downregulated genes resulting in good the survival of the patients (Fig. 3e). Recurrence status was also correlated with T stage, Grade and CIS status of the patient. TA and T1 stage were not associated with the recurrence with p-value of 0.22 (Supplementary Fig. 3). 75% of patients who recurred were T1, while rest 25% were TA. 46% of patients who were non recurrent were T1, while 54% were T2. Grade showed association with recurrence status ($p = 0.01$) (Supplementary Fig. 4). 92% of patients who were non-recurrent were G1, while remaining 8% were G3. 56% of recurrent patients were G3, while 44% were G1. CIS status was not associated with recurrence of the patient ($p = 0.09$) (Supplementary Fig. 5). 25% were CIS positive.

Association of gene expression signature and recurrence status of publicly available NMIBC datasets

In order to see association of gene expression signature and recurrence in a larger external cohort, we have made use of available information from publicly available NMIBC datasets. Our gene expression signature positively correlated with patient recurrence in publicly available NMIBC datasets. (GSE13507, GSE154261 and EGAS00001004693) (Fig. 4a,b,c,d,e) [7, 8, 9]. GSE13507 includes transcriptome expression from 165 primary bladder cancer samples, 23 recurrent non-muscle invasive tumour tissues, 58 normal looking bladder mucosae surrounding cancer and 10 normal bladder mucosae. It was found that normal samples got the lowest average expression of the upregulated genes and then comes adjacent normal and then primary bladder cancer samples. Highest expression was found in recurrent NMIBC (Fig. 4a). GSE154261 includes transcriptome expression from primary samples of 99 NMIBC patients with T1 status. We have taken normalised expression of 73 patients of discovery cohort having recurrence status and it was found that average expression of upregulated genes from our cohort was significantly higher in recurrent patients compared to non-recurrent patients (Fig. 4b). EGAS00001004693 includes patients enrolled in the UROMOL project, a European multicenter prospective study of NMIBC. Normalised transcriptome data from 535 NMIBIC patients were analysed. Stage included both T1 and TA. CIS status, cystectomy status, grade and tumor size were included. It was found that average expression of upregulated genes of patients from our patient set was significantly higher in patients who progressed to T2 compared to those who did not progress (Fig. 4c). Average expression of upregulated genes was also significantly higher with higher grade and higher tumor size compared to lower grade and lower tumor size (Figs. 4d, e). Similar trend was observed with downregulated genes in the above cohorts. Highest expression of downregulated genes was observed in normal samples compared to tumor samples as wells as non-recurrent patients compared to recurrent patients (data not

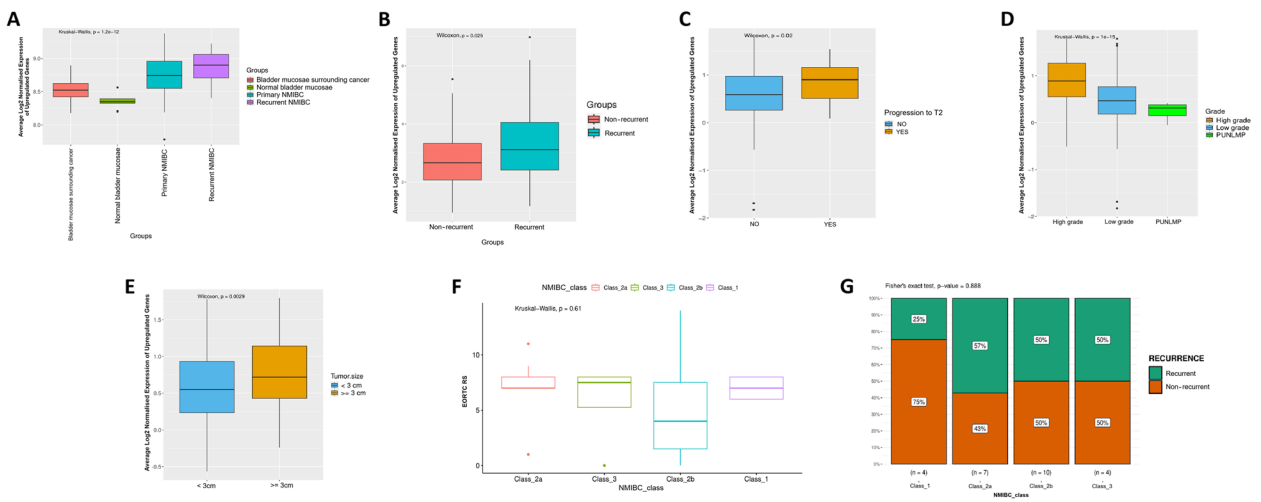

Fig. 4 Correlation of gene expression signature with publicly available expression datasets of patients with NMIBC. a: Average expression of upregulated genes in NMIBC cohort of GSE13507 with sample annotations. b: Average expression of upregulated genes in NMIBC cohort of GSE154261 with sample annotations. c: Average expression of upregulated genes in NMIBC data set of EGAS00001004693 with sample annotations. d: Average expression of upregulated genes in NMIBC data set of EGAS00001004693 was significantly higher in high grade compared to lower grades. e: Average expression of upregulated genes in NMIBC data set of EGAS00001004693 with sample annotations. f: UROMOL class prediction for our sample set and association with EORTC risk score. g: UROMOL class prediction for our sample set and association with recurrence status

shown). We have also checked whether the classifier that assigns class labels to single samples according to the four transcriptomic UROMOL classes of NMIBC [7]: class 1, class 2a, class 2b and class 3. This was not significantly associated with recurrence status and EORTC RS of our set of patients. (Figs. 4f, g).

Discussion

Evidence is clear that NMIBC and MIBC are heterogenous groups of tumors and TNM classification system is not sufficient for the characterization of this heterogeneity. Recent evidence suggests that molecular classification of bladder cancer captures underlying biological differences and tumour behaviours that may provide more precise stratification that could impact treatment and patient survival. This information consequently leads to differences in aggressiveness, prognosis, and progression. Therefore, risk stratification and prognostic models are of great importance because they enable standardisation of the treatment and follow-up, and data comparison.

Available scoring systems should be updated to match current standards of treatment. The low overall performance of the existing models reflects the unmet need for accurate biomarkers that measure the inherent biological potential of the tumours in the context of the microenvironment and host factors in general. Risk stratification and prognosis estimation should be performed when NMIBC is diagnosed. At present, scoring models use clinical and pathological variables that are known at the time of diagnosis and have been found to be of prognostic importance. Evaluation of some of these variables is subjective. Estimation of tumour size during TURBT is inaccurate, as well as determining the number of tumours in patients with diffuse lesions. Tumour stage and histological grade are associated with high observer variabilities [20, 21]. All these inaccuracies may lead to an incorrect tumour classification, which makes the process of validation complicated.

In addition to clinical prognostic factors molecular parameters would potentially be beneficial; however, no molecular markers have currently been recommended for widespread use in routine clinical practice. Use of genomics in risk stratification of bladder cancer patients is one of the promising future perspectives. Information from the Cancer Genome Atlas–MIBC project, which produced a comprehensive, open-access catalogue of DNA alterations, enables grouping of tumours into distinct molecular subtypes with different prognoses [22]. However, most sequencing efforts have focused on MIBC, and a significant unmet need is to translate this knowledge to NMIBC recurrence. Regarding NMIBC, the potential benefit and clinical utility of molecular subtyping may be in the more-accurate prognostication of recurrence and progression.

In our current study of retrospective set of 25 patients, we have performed transcriptomic analysis and showed that disease aggressiveness is associated with immune cell infiltration, genomic modifications, Nuclear Receptor Meta-Pathway and fatty acid metabolism and transcriptomic classes. In our approach we generated a score-based method, with two prediction methods including PAM and PLS-DA for checking differential genes ability to predict the subtypes and we could also predict recurrence under one year with an overall median accuracy of 90%. The signature was validated in other public datasets with similar accuracy in prediction. We have also observed that no significant association of UROMOL groups with recurrence status and EORTC RS was found for our set of patients. This may be attributable to the fact that, the EORTC risk calculator as discussed earlier may not accurately represent Indian scenario and early recurrence status of the patients.

Few study limitations that were identified by us are—the retrospective nature of our analysis and our small cohort size ($n = 25$). While our 1.5-year follow-up period was sufficient to detect early recurrences, it may not capture later recurrence events or disease progression patterns that could emerge with longer observation.

Finally, our findings, while validated in external datasets, require prospective validation in a larger, ethnically diverse cohorts before they can be considered for clinical implementation. The molecular signature we identified may perform differently in various patient populations with different genetic backgrounds and environmental exposures.

Despite these limitations, our study provides valuable preliminary evidence for the potential utility of transcriptomic profiling in risk stratification of NMIBC, particularly in the Indian population that has been underrepresented in previous molecular studies. Taken together, we conclude that classification based on gene expression enhances understanding of tumor behavior and may aid in tailoring treatments, thus improving the prognosis of NMIBC patients. Achievement of classification consensus could pave the way for well-designed prospective clinical trials that include molecular subtyping, which could change current guidelines and treatment approaches for NMIBC patients.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14273-y.

Supplementary Material 1: Supplementary Table 1: Model performance matrices and calibration parameters.

Supplementary Material 2: Supplementary Figure 1: Overall methodology Workflow.

Supplementary Material 3: Supplementary Figure 2: PLS-DA performance evaluation with test set of samples.

Supplementary Material 4: Supplementary Figure3: Comparison of patient recurrence status with T stage.

Supplementary Material 5: Supplementary Figure 4: Comparison of patient recurrence status with Grade.

Supplementary Material 6: Supplementary Figure 5: Comparison of patient recurrence status with CIS status.

Acknowledgements
We thank Prof P K Kondaiah and Dr Savitha Sharma for critical review and editing of the manuscript.

Authors' contributions
SN: Formal analysis, Investigation, Data curation; Writing – review & editing. HPS: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. RP: Investigation, Data curation. LP: Resources, Data curation, Investigation. DV and AN:Formal molecular analysis. VV,NT,PS and HG: Investigation. RVK: Pathology Investigation, Review. BSS: Resources, Investigation. AK: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, – review & editing, Supervision, Project administration.

Funding
We thank Sri Shankara Cancer Foundation for their generous intramural funding that made the study possible.

Data availability
The datasets generated and/or analysed during the current study are available in the Gene Expression Omnibus (GEO) repository, under the accession number GSE295809.

Declarations

Ethics approval and consent to participate
The study was conducted as per Helsinki guidelines, approved by the Institutional Ethics Committee of Sri Shankara Cancer Hospital and Research Centre, Bangalore (with Identification number- ECR/1715/Inst/KA/2022, affiliated with, Ministry of Health & Family Welfare, Government of India), which waived the requirement for individual patient informed consent due to the retrospective nature of the study using anonymized archival tissue specimens with anonymized data analysis.

Consent for publication
The study was conducted as per Helsinki guidelines, approved by the Institutional Ethics Committee of Sri Shankara Cancer Hospital and Research Centre, Bangalore (with Identification number- ECR/1715/Inst/KA/2022, affiliated with, Ministry of Health & Family Welfare, Government of India), which waived the requirement for individual patient informed consent due to the retrospective nature of the study using anonymized archival tissue specimens with anonymized data analysis and consent for publishing research findings. The authors affirm that the content does not contain identifiable personal information, and all reasonable measures have been taken to maintain privacy and confidentiality.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Uro Oncology, Sri Shankara Cancer Hospital and Research Center, Sri Shankara Cancer Foundation, Bangalore, India. [2]Department of Molecular Oncology, Sri Shankara National Center for Cancer Prevention and Research, Sri Shankara Cancer Foundation, Bangalore, India. [3]Department of Surgical Oncology, Sri Shankara Cancer Hospital and Research Center, Sri Shankara Cancer Foundation, Bangalore, India. [4]BaseCare Pvt Ltd, Bangalore, India. [5]Department of Pathology, Sri Shankara Cancer Hospital and Research Center, Sri Shankara Cancer Foundation, Bangalore, India.

Received: 26 October 2024 Accepted: 5 May 2025
Published online: 22 May 2025

References

1. Kamat AM, Hahn NM, Efstathiou JA, Lerner SP, Malmström PU, Choi W, et al. Bladder cancer. Lancet. 2016;388(10061):2796–810.
2. Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder Cancer Incidence and Mortality: A Global Overview and Recent Trends. Eur Urol. 2017;71(1):96–108.
3. Babjuk M, Burger M, Compérat EM, Gontero P, Mostafid AH, Palou J, et al. European Association of Urology Guidelines on Non-muscle-invasive Bladder Cancer (TaT1 and Carcinoma In Situ) - 2019 Update. Eur Urol. 2019;76(5):639–57.
4. Sylvester RJ, van der Meijden APM, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, 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):466–5 discussion 475-7.
5. Chang SS, Boorjian SA, Chou R, Clark PE, Daneshmand S, Konety BR, et al. Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline. J Urol. 2016;196(4):1021–9.
6. Cambier S, Sylvester RJ, Collette L, Gontero P, Brausi MA, van Andel G, et al. EORTC Nomograms and Risk Groups for Predicting Recurrence, Progression, and Disease-specific and Overall Survival in Non-Muscle-invasive Stage Ta-T1 Urothelial Bladder Cancer Patients Treated with 1–3 Years of Maintenance Bacillus Calmette-Guérin. Eur Urol. 2016;69(1):60–9.
7. Lindskrog SV, Prip F, Lamy P, Taber A, Groeneveld CS, Birkenkamp-Demtröder K, et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. Nat Commun. 2021;12(1):2301.
8. Kim WJ, Kim EJ, Kim SK, Kim YJ, Ha YS, Jeong P, et al. Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. Mol Cancer. 2010;8(9):3.
9. Robertson AG, Groeneveld CS, Jordan B, Lin X, McLaughlin KA, Das A, et al. Corrigendum to "Identification of Differential Tumor Subtypes of T1 Bladder Cancer" [Eur. Urol. 78 (2020) 533–537]. Eur Urol. 2022;81(2):e53.
10. Malhotra R, Javle V, Tanwar N, Gowda P, Varghese L, K A, et al. An absolute approach to using whole exome DNA and RNA workflow for cancer biomarker testing. Front Oncol. 2023;13:1002792.
11. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010.
12. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
13. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
14. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
16. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752.
17. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A. 2002;99(10):6567–72.
18. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. Stat Med. 2004;23(1):77–91.
19. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;15(14):128.
20. Cowherd SM. Tumor staging and grading: a primer. Methods Mol Biol. 2012;823:1–18.

21. Adamczyk P, Pobłocki P, Kadlubowski M, Ostrowski A, Wróbel A, Mikołajczak W, et al. A Comprehensive Approach to Clinical Staging of Bladder Cancer. J Clin Med. 2022;11(3):761.
22. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315–22.

Publisher's Note