

Sex-Associated Molecular Signature in Bladder Cancer: An Exploratory Multi-Cohort Analysis

Estratificação Molecular Associada ao Sexo no Cancro de Bexiga: Uma Análise Exploratória de Múltiplas Coortes

 Rita Ferreira¹,  Paula Oliveira²,  Francisco Amado³,  Lúcio Lara Santos⁴,  Rui Vitorino³

1. LAQV-REQUIMTE, Department of Chemistry, University of Aveiro, Aveiro, Portugal
2. Centre for Research and Technology of Agro Environmental and Biological Sciences (CITAB), Inov4Agro, University of Trás-os-Montes and Alto Douro (UTAD), Vila Real, Portugal
3. iBiMED, Department of Medical Sciences, University of Aveiro, Aveiro, Portugal
4. Pathology Department, Portuguese Oncology Institute of Porto, Porto, Portugal

Corresponding Author/Autor Correspondente:

Rui Vitorino [rvitorino@ua.pt]

Department of Medical Sciences, University of Aveiro, 3810-193 Aveiro, Portugal

<https://doi.org/10.34635/rpc.1202>

ABSTRACT

Introduction: Bladder cancer (BCa) exhibits marked differences between females and males in incidence, presentation, and clinical outcomes. However, the regulatory molecular networks underlying these sex-specific differences remain incompletely understood.

Methods: Publicly available transcriptomic datasets from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) were analyzed to identify sex-biased gene expression patterns associated with BCa progression. Sex-annotated GEO datasets were analyzed using cohort-specific and integrated differential expression models. Tumour-by-sex interaction analysis was performed using available tumour and non-tumour bladder specimens. Analyses were complemented by female-versus-male comparisons performed within non-muscle invasive BCa (NMIBC) and muscle invasive BCa (MIBC), exploratory survival analysis in TCGA-BLCA, cross-cancer expression contextualization and exploratory associations with model-predicted drug sensitivity.

Results: *XIST* emerged as a consistent female-associated marker across BCa subtypes, while *RPS4Y1*, *EIF1AY*, *KDM5D*, and *NLGN4Y* constituted the most reproducible Y-linked transcripts across male tumors. Female MIBC was further characterized

Received/Recebido: 01/07/2026 Accepted/Aceite: 03/07/2026 Published online/Publicado online: 00/00/000 Published/Publicado: 30/09/2026

© Portuguese Society of Surgery 2026. Re-use permitted under CC BY-NC. No commercial re-use.

© Sociedade Portuguesa de Cirurgia 2026. Reutilização permitida de acordo com CC BY-NC. Nenhuma reutilização comercial.

by the upregulation of *KRT5*, *S100A8*, *CALML3*, *PI3*, and *CEACAM6*, pointing to stage-specific differences spanning epithelial differentiation, immune signaling and cell adhesion.

Conclusion: Our study reveals distinct sex-specific transcriptional programs in BCa, involving *XIST*-related and Y chromosome-linked gene activity, with potential implications for disease stage and therapeutic response.

Keywords: Biomarkers; Gene Expression; Gene Expression Profiling; Sex Differences; Sex Factors; Urinary Bladder Neoplasms

RESUMO

Introdução: O cancro da bexiga (CB) apresenta diferenças marcadas entre mulheres e homens relativamente à incidência, apresentação clínica e resultados clínicos. No entanto, as redes moleculares reguladoras subjacentes a estas diferenças específicas do sexo permanecem ainda pouco compreendidas.

Métodos: Foram analisados conjuntos de dados de transcriptómica disponíveis publicamente no *Gene Expression Omnibus* (GEO) e no *The Cancer Genome Atlas* (TCGA), com o objetivo de identificar padrões de expressão génica associados ao sexo e relacionados com a progressão do CB. Os conjuntos de dados GEO com informação sobre o sexo foram analisados através de modelos de expressão diferencial específicos de cada coorte e de modelos integrados. Foi realizada uma análise da interação tumor × sexo, utilizando amostras disponíveis de tecido vesical tumoral e não tumoral. As análises foram complementadas por comparações entre mulheres e homens, realizadas separadamente nos casos de cancro da bexiga não músculo-invasivo (NMIBC) e cancro da bexiga músculo-invasivo (MIBC), por uma análise exploratória da sobrevivência no TCGA-BLCA, pela contextualização da expressão génica em diferentes tipos de cancro e por associações exploratórias com a sensibilidade a fármacos prevista por modelos.

Resultados: O *XIST* emergiu como um marcador consistentemente associado ao sexo feminino nos diferentes subtipos de CB, enquanto *RPS4Y1*, *EIF1AY*, *KDM5D* e *NLGN4Y* constituíram os transcritos ligados ao cromossoma Y mais consistentemente identificados nos tumores masculinos. O MIBC feminino caracterizou-se adicionalmente pela sobre-expressão de *KRT5*, *S100A8*, *CALML3*, *PI3* e *CEACAM6*, apontando para diferenças específicas do estágio que envolvem a diferenciação epitelial, a sinalização imunitária e a adesão celular.

Conclusão: O nosso estudo revela programas transcricionais distintos associados ao sexo no CB, envolvendo a atividade relacionada com *XIST* e com genes ligados ao cromossoma Y, com potenciais implicações na progressão da doença e na resposta terapêutica.

Palavras-chave: Biomarcadores; Diferenças entre Sexos; Expressão Génica; Fatores Sexuais; Neoplasias da Bexiga Urinária

INTRODUCTION

Bladder cancer (BCa) represents a significant global health burden, ranking as the ninth most commonly diagnosed cancer worldwide. In 2022, an estimated 614 298 new cases and 220 596 deaths occurred globally, with approximately 75% of the burden observed in males. Its incidence is influenced by demographic trends, most notably population ageing, and exposure to risk factors, particularly tobacco smoking, posing substantial challenges for healthcare systems worldwide.¹ Most bladder tumors are urothelial, and approximately 75% are non-muscle invasive (NMIBC) at diagnosis,² a stage associated with high recurrence rates and significant healthcare costs.

Although diagnosis of BCa occurs three to four times more frequently in males, female patients tend to present at more advanced stages and experience worse outcomes, including a statistically significant 21% increase in mortality

over two years and a greater incidence of disease recurrence, even at comparable disease stages.³⁻⁵ These disparities have historically been attributed to delays in diagnosis, as initial hematuria symptoms in women are frequently misattributed to urinary tract infections, leading to multiple rounds of antibiotics before urological investigation, as well as to differences in tobacco exposure patterns. However, such factors do not fully account for the persistent survival gap between sexes,^{6,7} suggesting that hormonal and non-hormonal biological factors intrinsic to sex play a fundamental role in shaping tumor biology, immune microenvironment and treatment response.

Growing evidence implicates sex hormones, sex chromosomes, and sex-specific epigenetic mechanisms as key modulators of BCa risk and behavior. Androgens and the androgen receptor (AR) have been shown to promote BCa development and invasion,⁸ whilst the X chromosome confers

a degree of protection in females, partly through KDM6A-dependent epigenetic regulation of the p53 pathway.⁹ Loss of the Y chromosome has been associated with immune evasion and worse oncological outcomes.¹⁰ Furthermore, sex differences in the tumour immune microenvironment, including distinct CD8⁺ T cell programming, may contribute to divergent outcomes between sexes.¹¹ Interest in understanding the role of sex differences in cancer incidence, prevalence, and severity is growing, with sex as a biological variable increasingly recognized as deserving of being incorporated in research efforts in BCa.⁶

To address this gap and provide further insights into the impact of biological sex on BCa development and disease stage, we performed a multi-cohort transcriptomic and systems-level analysis, integrating sex-annotated data from GEO BCa cohorts with TCGA-BLCA data, complemented by *in silico* drug sensitivity prediction. Our findings identify a panel of overexpressed genes that may underlie the sex-related differences observed in BCa, offering preliminary insights that could inform future thinking on molecular stratification and therapeutic approaches.

METHODS

1. DATA ACQUISITION AND COHORT SELECTION

This study was designed as a sex-stratified transcriptomic and clinicogenomic analysis of BCa. Publicly available gene-expression datasets were retrieved from the Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>), The Cancer Genome Atlas Urothelial Bladder Carcinoma project (TCGA-BLCA), accessed through the National Cancer Institute Genomic Data Commons portal (<https://portal.gdc.cancer.gov/>), using the TCGAbiolinks R package, and

Human Protein Atlas (HPA; <https://www.proteinatlas.org/>). GEO datasets were selected if they included human BCa samples, transcriptomic profiling data, available biological sex annotation, and accessible clinical or pathological metadata. Datasets based exclusively on cell lines, animal models, normal tissue, treatment-only samples, duplicated samples, or lacking sex annotation were excluded. Fig. 1 provides an overview of the analytical pipeline implemented in the present study.

The GEO cohorts retained for analysis were GSE32894, GSE48075, and GSE13507 (Table 1), retrieved from the studies.¹²⁻¹⁴ After exclusion of samples without confirmed sex annotation, the harmonized GEO dataset included 542 tumors, comprising 126 female and 416 male cases. None of these studies considered the effect of sex as a variable, which may partially explain the imbalance in the number of male and female samples across the cohorts. Across the harmonized GEO datasets, 17 405 shared genes were retained for cross-cohort analysis. GEO data were downloaded in R using the GEOquery package version 2.81.31.

2. METADATA HARMONIZATION AND DISEASE-CONTEXT CLASSIFICATION

Clinical metadata were manually curated to harmonize biological sex and tumor stage across datasets. Sex labels were standardized as “Female” or “Male”. Samples without reliable sex annotation were excluded. Tumor stage information was extracted from available metadata fields and classified as non-muscle-invasive bladder cancer (NMIBC) when compatible with Ta, T1, Tis, superficial, or non-muscle-invasive disease, and as muscle-invasive bladder cancer (MIBC) when compatible with T2–T4 or explicit muscle-invasive disease. Ambiguous cases were excluded from stage-stratified meta-cohort analyses.

Table 1: Overview of the gene expression datasets included in the study. For each cohort, the total number of tumor samples, sex distribution, patient age (when available), number of genes analyzed, platforms used for transcriptomics, and analytical role are reported.

Cohort	Tumor samples available, n	Tumor samples included, n	Male, n (%)	Female, n (%)	Age	NMIBC/MIBC, n (%)	Number of genes analysed
GSE32894	308	308	228 (74.0%)	80 (26.0%)	71 (20–96)	215 (69.8%)/ 93 (30.2%)	24 402
GSE48075	142	69 with complete annotation	52 (75.0%)	17 (25.0%)	–	6 (100%)/ 0 (0%)	48 803
GSE13507	256	165 primary tumors	136 (82.4%)	29 (17.6%)	–	103 (62.4%)/ 62 (37.6%)	43 148

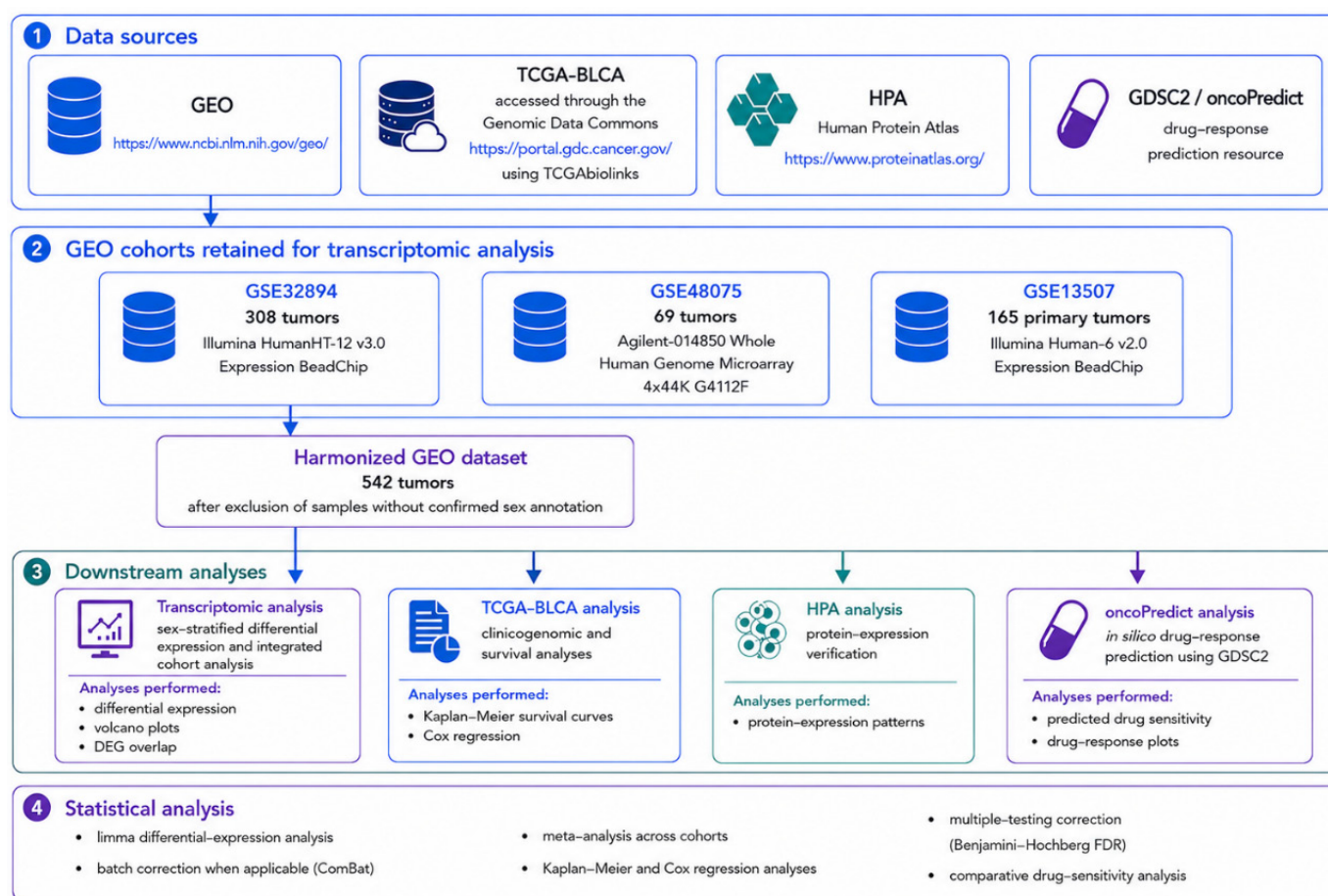


Figure 1: Overview of the study flow performed, including data sources, cohort selection, and the analyses performed.

The availability and completeness of potential confounding variables were evaluated separately for each cohort. Variables considered included age, disease stage, tumour grade, histological variant, molecular subtype, smoking exposure and history of treatment. Covariates were incorporated into adjusted models only when they were sufficiently complete, consistently defined and represented in both sexes. Smoking, treatment and histological-variant variables were not incorporated into the cross-cohort models because their annotation was unavailable, inconsistent or insufficiently complete across the included GEO datasets. Molecular-subtype adjustment was evaluated separately in TCGA-BLCA. Covariate availability, missingness and analytical use are summarized in Supplementary Table S1.

3. GENE-EXPRESSION PREPROCESSING AND DIFFERENTIAL EXPRESSION ANALYSIS

Expression matrices were extracted from GEO objects and log2-transformed when required. Probe annotations were mapped to gene symbols using platform annotation files. We removed probes without valid gene-symbol annotation and

collapsed multiple probes mapping to the same gene using limma's *avereps* function (Linear Models for Microarray Data), a Bioconductor R package for gene expression analysis.

Differential-expression analyses were performed separately in GSE32894, GSE48075 and GSE13507 using moderated linear models implemented in the limma R package. Biological sex was coded with male patients as the reference category. Consequently, the reported sex coefficient and log2 fold change represent female minus male expression: positive values indicate higher expression in females and negative values indicate higher expression in males. Genes were considered differentially expressed in each cohort using a combined threshold of Benjamini-Hochberg-adjusted $p < 0.05$ ($FDR < 0.05$) and $|\log_2FC| > 0.5$. Differentially expressed genes (DEG) intersections across the three cohorts were visualized using UpSetR.

For the integrated analysis of tumor versus normal tissue tumor *versus* normal, genes shared across the three datasets were merged for the same gene-symbol and batch effects

were corrected using ComBat (sva R package), with cohort identity specified as the batch variable. The design matrix was specified as: `mod = model.matrix(~sex + sample_type, data = metadata)`, where `sample_type` distinguishes tumor *versus* normal.

For the integrated NMIBC vs. MIBC analyses, the same merging and ComBat correction procedure was applied, with the design matrix specified as: `mod = model.matrix(~sex + disease_stage, data = metadata)`, where `disease_stage` distinguishes NMIBC from MIBC. Differential expression analysis was performed using limma, with biological sex as the primary variable of interest. Graphical thresholds for volcano plots were set at $FDR < 0.05$ and $|\log_2FC| > 1.0$.

Meta-analysis was performed to evaluate sex-biased DEG across the three independent GEO datasets. Differential expression models were fitted independently for each cohort using linear models in limma. For each gene, the pooled $\log_2$ fold change, 95% confidence interval, nominal p -value, Benjamini–Hochberg-adjusted p -value (FDR), Cochran's Q , I^2 , τ^2 and the direction of the cohort-specific effects were assessed. Because only three cohorts contributed to the analysis, estimates of between-study heterogeneity were interpreted cautiously; genes with consistent directions and low heterogeneity ($I^2 < 50\%$) were distinguished from those showing substantial heterogeneity or cohort-level directional discordance.

4. TCGA-BLCA SURVIVAL ANALYSIS

TCGA-BLCA RNA-seq data and clinical information were retrieved from the Genomic Data Commons using TCGAblinks. Only primary tumor samples were retained. Gene identifiers were mapped to gene symbols, and expression values were normalized as counts per million when pre-normalized assays were not available.

Overall survival was defined using days to death for deceased patients and days to last follow-up for censored patients. Patients with missing or non-positive follow-up time were excluded. Sex-restricted survival analyses were performed. High- and low-expression groups were defined using the median expression value for each gene.

Kaplan–Meier curves were generated after median dichotomization of expression and were used for descriptive visualization. High- and low-expression groups were compared using two-sided log-rank tests. Log-rank P -values were adjusted using the Benjamini–Hochberg procedure across the six prespecified analyses.

Cox proportional-hazards regression was used as the principal survival modelling approach. Gene expression was $\log_2$ -transformed where required and standardized to a mean of zero and standard deviation of one. Hazard ratios therefore represent the change in mortality hazard associated with a one-standard-deviation increase in expression. Each gene was evaluated in a separate sex-restricted model.

Tumour grade was assessed in a sensitivity model where sufficient variation and complete annotation were available. Smoking and treatment variables were not included because they were incomplete or inconsistently defined. Complete-case analysis was used, and missing clinical covariates were not imputed.

Hazard ratios, 95% confidence intervals, nominal p -values and Benjamini–Hochberg-adjusted P -values were reported. Multiple-testing correction was applied separately across the univariable expression coefficients and the multivariable expression coefficients. The proportional-hazards assumption was evaluated using scaled Schoenfeld residuals for each covariate and globally.

5. HUMAN PROTEIN ATLAS EXPLORATION

Human Protein Atlas (version 23.0, accessed on February 2026) RNA and immunohistochemistry data were used as orthogonal, tissue-level references for the selected sex-chromosome gene panel, independent of the sex-stratified analysis performed in the GEO and TCGA cohorts. RNA expression was assessed using HPA consensus transcript expression values, whereas protein expression was evaluated descriptively using available immunohistochemistry staining categories (High/Medium/Low/Not detected). For each gene/antibody entry, we recorded the corresponding Ensembl/HPA identifiers, the antibody staining reliability score (Approved/Supported), the tissue category examined (normal urothelium *versus* urothelial carcinoma), the total core count, and any available demographic attributes. As HPA specimens are not annotated by donor sex, these data do not permit direct assessment of sex-associated differences and were used solely to contextualize whether the panel genes show detectable tissue- and protein-level expression in the relevant tissue context.

6. IN SILICO CHEMOSENSITIVITY AND TARGETED THERAPY PREDICTION

To evaluate the therapeutic vulnerabilities of the stratified patient cohorts, *in silico* drug sensitivity profiling was performed using the oncoPredict R package (version 0.2). Ridge

regression models were constructed utilizing high-throughput baseline transcriptomic and screening data from the Genomics of Drug Sensitivity in Cancer (GDSC2) database as the training dataset. Specific drug identifiers evaluated include cisplatin (Drug ID: 1005), gemcitabine (Drug ID: 119), docetaxel (Drug ID: 1007), and erdafitinib (Drug ID: 1928). Training matrices comprised GDSC2 pan-cancer cell line baselines ($n=808$ cell lines), while the primary female target matrix comprised $n=71$ female TCGA-BLCA patients (with $n=267$ total patients in the combined sex matrix). After matching gene identifiers and removing duplicated or non-shared genes, n common genes were retained for model fitting and prediction. Candidate-gene expression and model-imputed $\ln(\text{IC}_{50})$ scores were examined within the female MIBC subset because the preceding differential-expression analysis identified five genes as female-associated within MIBC. Pearson correlations were calculated for five genes, and four drugs, producing 20 prespecified comparisons. Two-sided nominal p -values were adjusted across the 20 comparisons using the Benjamini–Hochberg procedure. Significance levels were categorized according to standard thresholds: $*p < 0.05$; $p < 0.01$; $***p < 0.001$; ns (non-significant, $p \geq 0.05$). All computational pipelines, downstream matrix transformations, and multi-panel stratification dashboards were implemented in R (v4.3) using the tidyverse framework, with visualizations generated via highly customized ggplot2 geom layers.

7. STATISTICAL ANALYSIS

All analyses were performed in R (version 4.3). Differential expression analysis used moderated statistics implemented in limma. Tests were two-sided. Benjamini–Hochberg correction was applied separately within each prespecified family of analyses. Transcriptome-wide DEGs were defined by an adjusted p -value below 0.05 and an absolute $\log_2$ fold change above 0.5. Nominal p -values, adjusted p -values, effect sizes and 95% confidence intervals were reported where applicable. Associations meeting only nominal significance were described as exploratory and were not classified as statistically supported findings.

RESULTS

We first asked whether BCa is associated with distinct transcriptional programs by integrating the three independent GEO cohorts into an expression matrix containing 17 405 unique genes from more than 542 samples. The analyzed cohorts showed the expected male predominance, with male patients representing approximately 74%–82% of sex-annotated tumor samples across the major datasets. This imbalance was explicitly accounted for by treating biological

sex as the primary variable of interest and by using cohort-level and covariate-adjusted models rather than pooled unadjusted comparisons. Comparison of DEGs across the three datasets showed that GSE13507 had the largest number of DEGs (266 genes), whereas eight genes were shared by all three datasets (Fig. 2A). Volcano plot analysis of the combined datasets, stratified by sex, identified *XIST* (X-inactive specific transcript) as the strongest female-enriched transcript, with a pronounced positive $\log_2$ fold-change and very high statistical significance. In contrast, canonical Y-linked genes, including *RPS4Y1* (ribosomal protein S4, Y-linked 1), *EIF1AY* (eukaryotic translation initiation factor 1A, Y-linked), *CYORF15A* (chromosome Y open reading frame 15A, also referred to as *TXLNGY*), *NLGN4Y* (neuroligin 4, Y-linked), *TMSB4Y* (thymosin beta 4, Y-linked) and *DDX3Y* (DEAD-box helicase 3, Y-linked), were strongly shifted toward the male-enriched side of the plot (Fig. 2B).

Human Protein Atlas RNA and immunohistochemistry data were used to assess whether the selected Y-linked transcripts were supported at the tissue-expression level. As expected, *XIST* was not detected in the protein dataset because it encodes a long non-coding RNA with no translated protein product. Therefore, HPA immunohistochemistry data cannot provide tissue-level contextualization of *XIST*-related transcript. The RNA profiles showed that *DDX3Y*, *EIF1AY*, *RPS4Y1*, *CYORF15A* and *TMSB4Y* were expressed across multiple malignancies, including BCa (Fig. 2C), indicating that these genes should not be interpreted as bladder-specific markers but rather reflect broader male chromosomal dosage across tumor types. At the protein level, however, the Y-linked genes displayed marked differences in BCa (Fig. 2D). *TMSB4Y* showed the strongest representation, with a substantial proportion of tumors exhibiting medium or high staining. *EIF1AY* and *DDX3Y* were also broadly detectable, mainly in the medium-staining range, whereas *RPS4Y1* was more often detected at low or medium intensity. In contrast, *NLGN4Y* and *CYORF15A* were frequently classified as low or not detected, indicating that their biological relevance is more evident at the transcript level than at the immunohistochemistry level.

To determine whether sex-dependent transcriptional variation was associated with clinical outcome, exploratory Kaplan–Meier and Cox proportional-hazards survival analysis were performed in TCGA-BLCA patients. This sex-restricted design was essential because Y-linked genes are intrinsically male-associated and *XIST* is female-biased. Five Y-linked genes were evaluated in the final TCGA survival analysis:

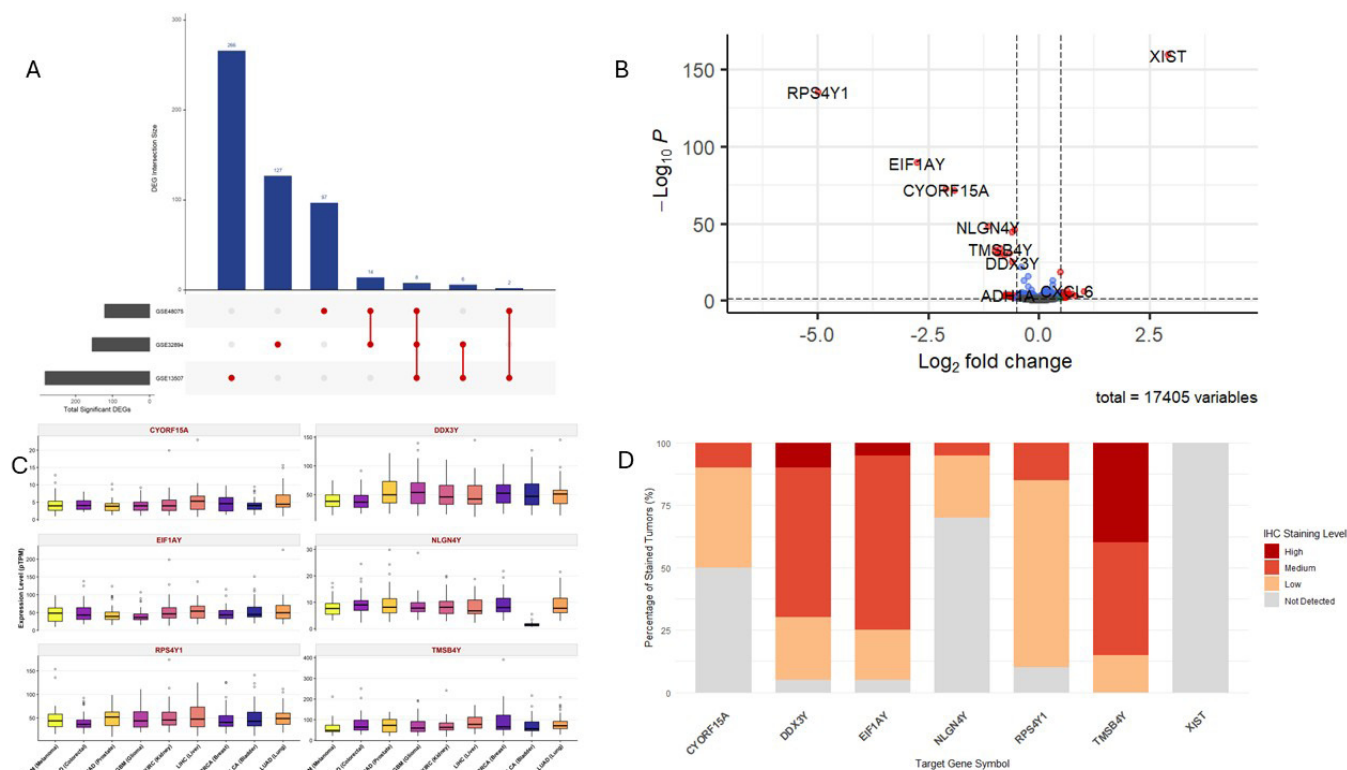


Figure 2: Comparison of differentially expressed genes (DEGs) between women and men bladder tissue. (A) UpSet plot depicting the intersection of DEGs across GEO cohorts. Vertical bars represent the number of DEGs in each intersection, with the dot matrix below indicating which cohorts contribute to each intersection. Single filled dots denote cohort-specific DEGs, while connected dots indicate DEGs shared between the respective cohorts. Horizontal bars on the left represent the total number of significant DEGs identified in each cohort. (B) Volcano plot showing sex-associated DEGs across the combined BCa datasets. The x-axis represents log2 fold change and the y-axis represents $-\log_{10} P$. Each point corresponds to one of 17 405 analyzed genes, colored according to significance criteria: non-significant (NS, gray), log2 fold change only (green), p value only (blue), or both p value and log2 fold change (red). (C) Human Protein Atlas multi-cohort transcript abundance profiles for selected Y-linked genes across TCGA malignancies. Boxplots show consensus RNA-seq transcript abundance (pTPM) across multiple cancer cohorts. Boxplots display median, interquartile range, whiskers, and outliers, with cohorts sorted by median expression. (D) Consensus protein expression profiles across target BCa cohorts for selected genes. Stacked bar plots show the percentage of tumors classified as high, medium, low, or not detected by immunohistochemistry across the analyzed BCa cohorts.

RPS4Y1, *EIF1AY*, *NLGN4Y*, *TMSB4Y* and *DDX3Y* (Fig. 3). Patients were divided into high- and low-expression groups using median expression thresholds, generating balanced strata of 149 male patients per group. Most Y-linked genes did not significantly stratify overall survival. *NLGN4Y* showed the strongest survival trend among the Y-linked panel (log-rank p -value = 0.086). After Benjamini-Hochberg correction across the comparisons, no Y-linked transcript remained significant (*NLGN4Y* adjusted p = 0.279; all others FDR $\geq$ 0.400). The log-rank p -value was 0.093 for *XIST* in females, indicating a non-significant trend. Median expression dichotomization yielded balanced groups (high *XIST*, n = 52; low *XIST*, n = 53). Kaplan-Meier analysis suggested poorer overall survival in the high-*XIST* group,

although the difference was not significant (log-rank p -value = 0.093). Consistently, multivariable Cox models adjusted for age and disease stage showed no significant association for *XIST* in women or *NLGN4Y* in men. These findings should be considered exploratory given the absence of significance after multiple-testing correction (Supplementary Table S10). Nevertheless, the most biologically informative trend is the early survival divergence observed for *NLGN4Y*.

To complement the transcriptomic findings with functional context, a literature-informed conceptual model was constructed to summarize the putative biological roles of the key sex chromosome-associated genes identified in this study (Fig. 4). *XIST* may function as a competing endogenous

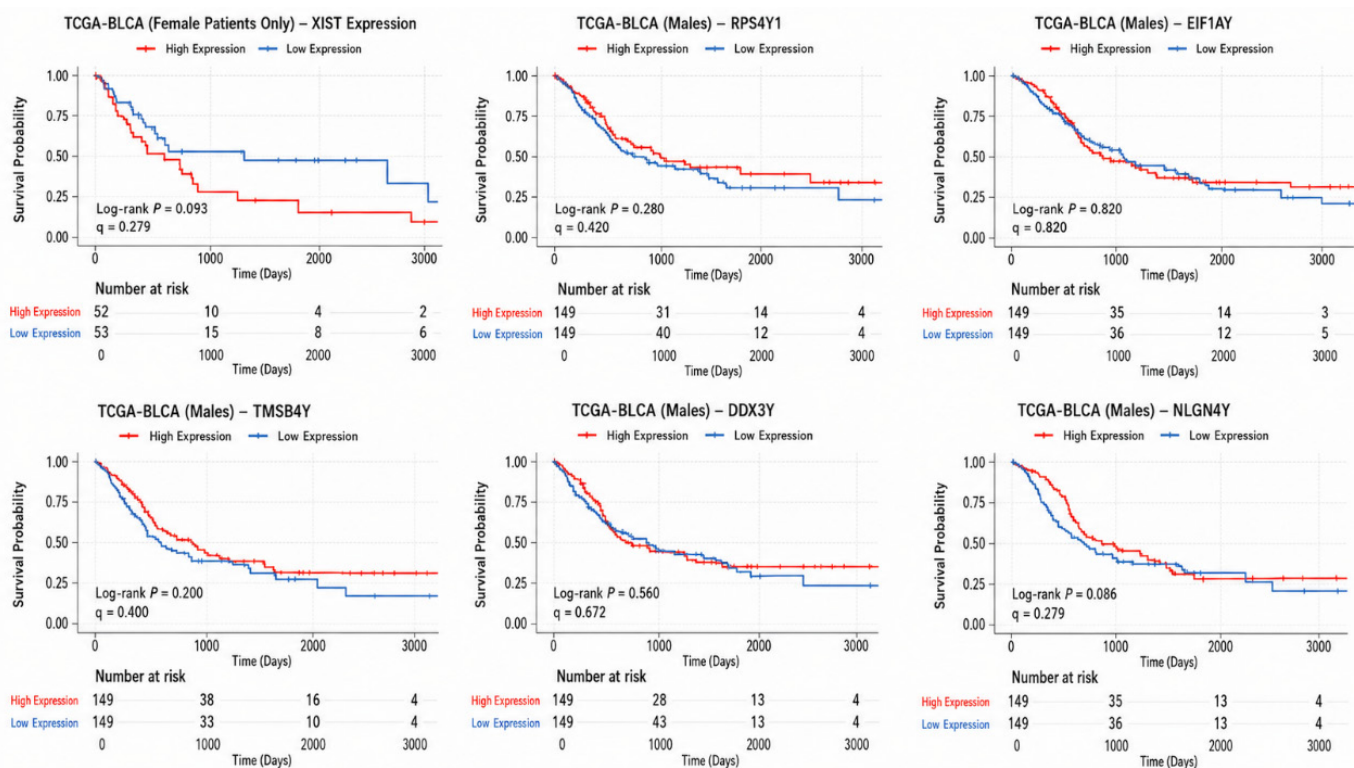


Figure 3: Kaplan–Meier survival curves of sex chromosome–related genes in *TCGA-BLCA*. Overall survival curves stratified by median gene expression for *XIST* in 105 female patients and five Y-linked transcripts (*RPS4Y1*, *EIF1AY*, *TMSB4Y*, *DDX3Y*, and *NLGN4Y*) in 298 male patients. Nominal log-rank p -values are displayed on each panel. Corresponding risk tables report the number of patients at risk across 1000-day intervals. None of the comparisons retained statistical significance after Benjamini–Hochberg multiple-testing correction ($FDR > 0.05$).

RNA (ceRNA), interacting with several microRNAs, including miR-124, miR-129-5p, miR-133a, miR-139-5p, miR-200c, miR-335, and miR-15a-5p, thereby influencing pathways involved in proliferation, epithelial–mesenchymal transition (EMT), invasion, metastasis, stemness, apoptosis resistance, and epigenetic regulation (Fig. 4A). In male BCa patients, the Y chromosome–linked genes identified in our analysis were associated with distinct cellular functions (Fig. 4B). *RPS4Y1* was linked to ribosomal structure and protein synthesis, *EIF1AY* to translation initiation, *TMSB4Y* to cytoskeletal organization, *NLGN4Y* to cell adhesion and membrane signaling, and *DDX3Y* to RNA metabolism and helicase activity. Although these genes were not significantly associated with overall survival in the TCGA cohort, their functional annotations suggest potential contributions to sex-specific biological processes in BCa. Nevertheless, this model was not derived from a formal network-inference pipeline (e.g., co-expression or interaction-database analysis); the associations depicted should therefore be interpreted as literature-derived hypotheses rather.

To assess whether this sex-associated expression profile was dependent on disease stage, volcano plot analysis was performed separately for NMIBC and MIBC, stratified by biological sex. In the aggregated NMIBC meta-cohort, female-versus-male differential expression analysis revealed a dominant sex-chromosome-linked structure. *XIST* was the strongest female-biased transcript, occupying the far positive side of the volcano plot with very high statistical significance, whereas *RPS4Y1*, *EIF1AY*, *KDM5D* and *NLGN4Y* were clearly male-biased (Fig. 5A). The MIBC meta-cohort reproduced the main sex-chromosome pattern observed in NMIBC (Fig. 5B). The recurrence of this pattern in MIBC indicates a conserved *XIST*/Y-linked dosage axis feature across BCa contexts.

The MIBC volcano plot also showed that sex-associated differences extended beyond sex chromosomes. Female-biased autosomal genes included *KRT5* (Keratin 5), *S100A8* (S100 calcium binding protein A8), *CALML3* (calmodulin-like protein 3), *PI3* (peptide inhibitor 3) and *CEACAM6*

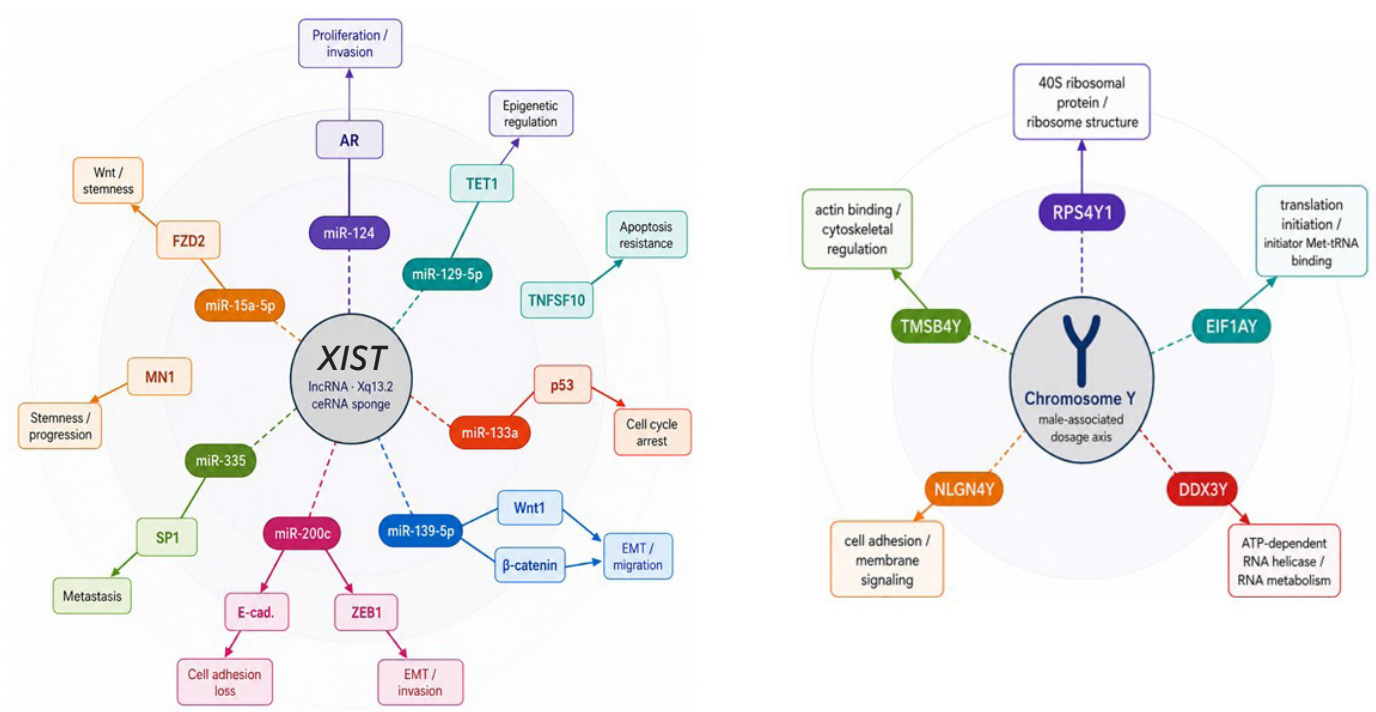


Figure 4: Literature-informed sex-specific regulatory networks potentially relevant to bladder cancer, highlighting (A) the *XIST*-centered hub as a lncRNA acting as a ceRNA sponge for the indicated miRNAs, based on interactions reported in ⁽¹⁵⁾; downstream targets and associated biological processes reflect functional annotations reported in the same line. Dashed lines indicate miRNA-sponging relationships; solid lines indicate downstream regulatory or signaling relationships; arrows indicate the direction of the proposed regulatory effect. Colors group genes/miRNAs by the biological process indicated in the outer nodes and (B) Y chromosome-associated genes and their canonical molecular functions, based on Uniprot annotation (accessed in May 2026); these represent general gene function rather than BCa-specific experimental evidence. Associations depicted in this figure should be interpreted as literature-derived hypotheses rather than validated bladder-cancer-specific interaction networks.

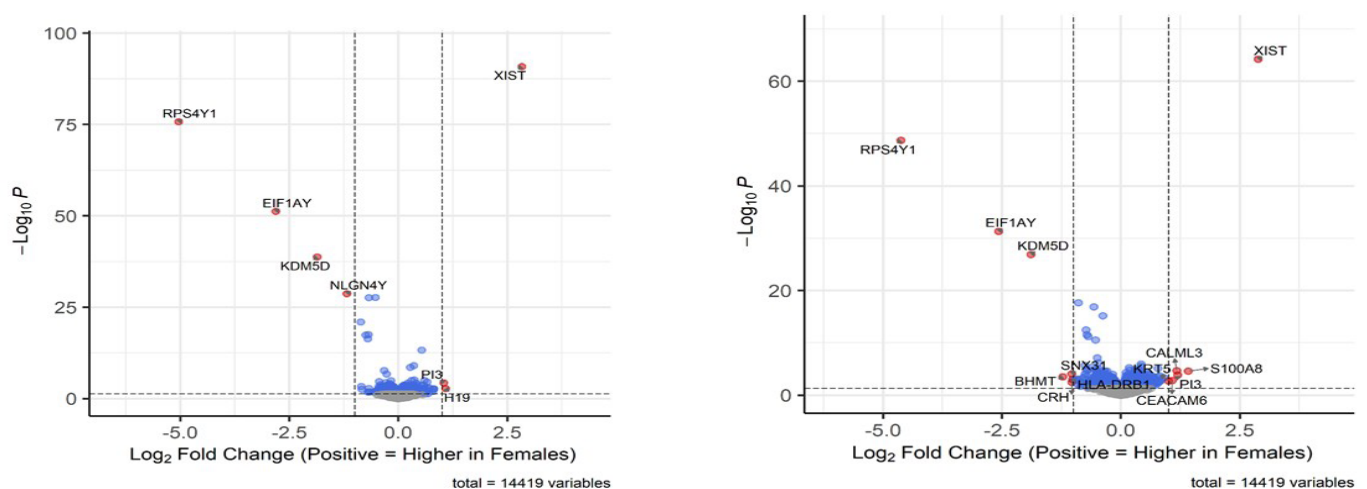


Figure 5: Volcano plot showing sex-associated DEGs across NMIBC (A) and MIBC (B). The x-axis represents $\log_2$ fold change and the y-axis represents $-\log_{10} P$. Each point corresponds to one of 14,419 analyzed genes, colored according to significance criteria: non-significant (NS, gray), $\log_2$ fold change only (green), p value only (blue), or both p value and $\log_2$ fold change (red).

(carcinoembryonic antigen-related cell adhesion molecule 6), suggesting enrichment of basal/squamous, inflammatory and epithelial-remodelling programs in a subset of female tumors. Thus, the MIBC sex signal was not limited to *XIST* and Y-linked genes; rather, sex-chromosome polarity appeared to be embedded within broader tumor-lineage and immune-epithelial transcriptional programs.

To explore BCa-specific therapeutic vulnerabilities associated with sex-linked genes, we mapped the putative biomarker panel against imputed drug sensitivity models for standard chemotherapies and targeted agents. Evaluation of the male-specific Y-linked panel revealed no linear, dosage-dependent correlations with cytotoxic chemotherapy $\ln(\text{IC}_{50})$ scores ($p > 0.05$). Notably, elevated expression of the canonical basal marker KRT5 demonstrated negative correlations with imputed $\ln(\text{IC}_{50})$ scores for cisplatin

($R = -0.728$, $p = 6.30 \times 10^{-13}$), docetaxel ($R = -0.529$, $p = 2.15 \times 10^{-6}$), and gemcitabine ($R = -0.434$, $p = 1.57 \times 10^{-4}$) (Fig. 6). This chemosensitive profile was further supported under cisplatin treatment by significant negative correlations with the secondary basal and inflammatory markers CALML3 ($R = -0.442$), PI3 ($R = -0.354$), and S100A8 ($R = -0.312$). Intriguingly, this therapeutic paradigm completely inverted when modeling the targeted FGFR inhibitor, erdafitinib. In this context, the luminal progression marker CEACAM6 emerged as a predictor of high drug sensitivity ($R = -0.691$, $p = 2.63 \times 10^{-11}$), whereas KRT5 expression strongly tracked with *in silico* resistance ($R = +0.314$, $p = 7.65 \times 10^{-3}$).

DISCUSSION

Biological sex is recognized as a determinant of BCa incidence and malignancy, yet the molecular mechanisms underlying this dimorphism remain poorly understood. Our study identified

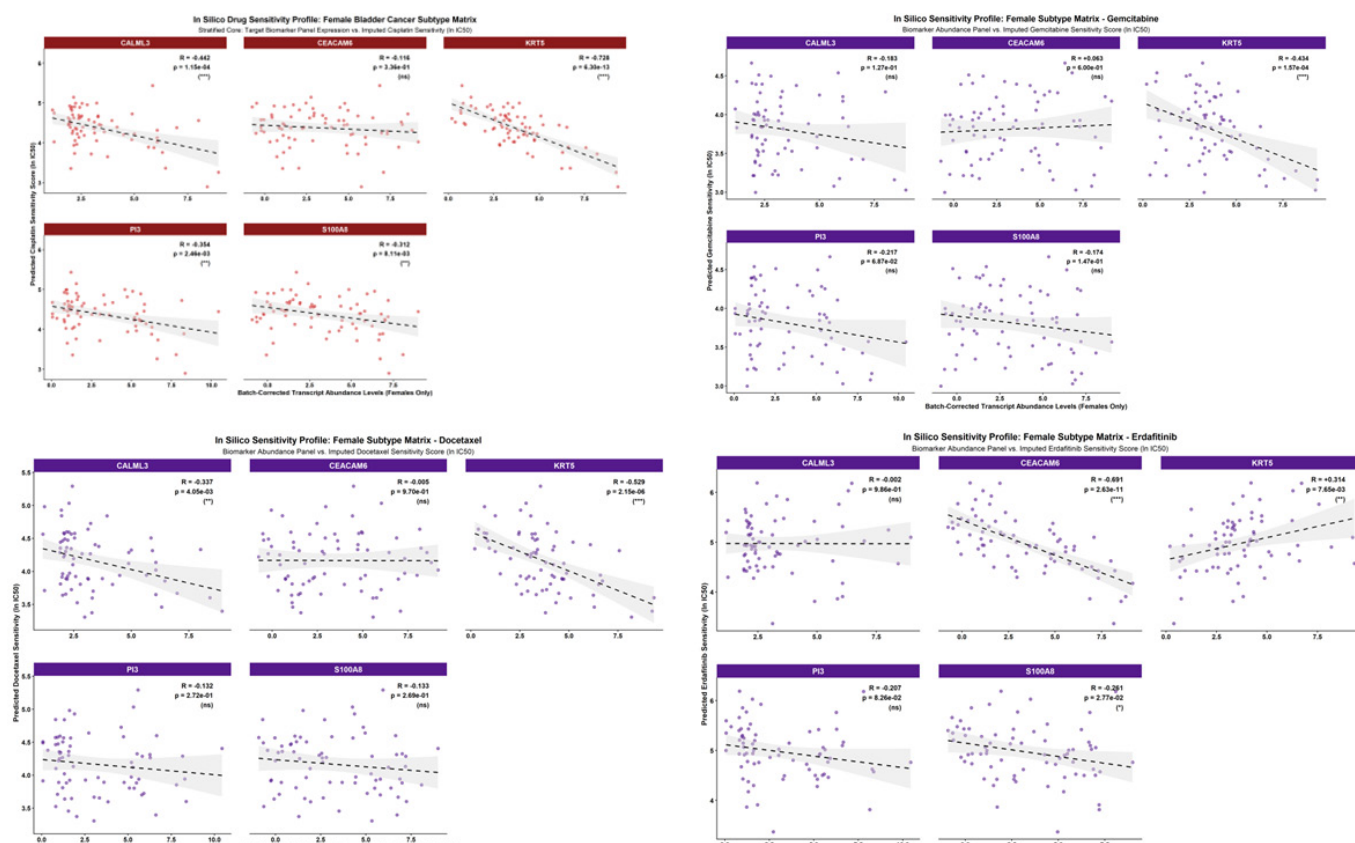


Figure 6: Association between biomarker expression and predicted therapeutic response in female muscle-invasive bladder cancer (n=71). Scatter plots show the relationships between batch-corrected transcript abundance levels of CALML3, CEACAM6, KRT5, PI3, and S100A8 and oncoPredict-derived cisplatin, gemcitabine, docetaxel and erdafitinib response scores in the female muscle-invasive bladder cancer meta-cohort. Predicted response is expressed as $\ln(\text{IC}_{50})$, with lower values indicating greater predicted drug sensitivity. Dashed lines represent linear regression fits, and shaded areas indicate 95% confidence intervals. Pearson correlation coefficients and two-sided p -values are shown in each panel. Significance notation: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns, $p \geq 0.05$.

a panel of DEGs that may contribute to sex disparities in BCa, including the female-biased *XIST* and the male-biased *RPS4Y1*, *EIF1AY*, *CYORF15A/TXLNGY*, *NLGN4Y*, *TMSB4Y*, and *DDX3Y*.

Our findings regarding the female-specific upregulation of *XIST* in BCa tissues are in agreement with Cai et al,¹⁶ who associated such overexpression with poor prognosis in female patients, an association that was also preserved when MIBC and NMIBC were considered in isolation (Fig. 5). *XIST* is a long non-coding RNA of approximately 19 kb, located at the X chromosome inactivation center of chromosome Xq13.2. This gene is essential for X chromosome inactivation (XCI) in female mammals to ensure equivalent expression levels of X-linked genes between sexes.¹⁷ Beyond its canonical role in dosage compensation, *XIST* has emerged as a multifunctional oncogenic driver of BCa through ceRNA-mediated mechanisms, acting as a molecular sponge for multiple tumor-suppressive miRs including miR-129-5p, miR-124, miR-139-5p, miR-200c, miR-133a, and miR-335, thereby derepressing downstream oncogenic effectors such as AR, Wnt1, TET1, and p53.^{15,18} More recently, the *XIST*/miR-15a-5p/MN1/FZD2 axis was shown to be overactivated specifically in female BCa, adding further mechanistic depth to the female-specific oncogenic role of *XIST*.¹⁶ Consistent with this biological framework, Kaplan-Meier survival analysis of TCGA-BLCA female patients (n=105) revealed a trend whereby high *XIST* expression was associated with consistently worse overall survival compared to low *XIST* expression, with survival curves diverging progressively from approximately day 500 onwards and becoming particularly pronounced between days 1000 and 3000, where low *XIST* patients maintained a survival probability of approximately 50% and high *XIST* patients declined to approximately 25%. Nevertheless, this trend did not reach statistical significance (log-rank $p=0.093$), possibly reflecting the small number of female BCa patients in TCGA-BLCA, a well-recognized limitation of sex-stratified survival analyses in datasets where female patients are substantially underrepresented. In the broader context of lncRNA-based immunotherapy, the feedback loop between lncRNAs, miRs, and their targets has been proposed as a targetable network for BCa therapy, given that immunotherapy approaches including immune checkpoint blockade have demonstrated off-target effects and limited clinical outcomes when acting on single targets.¹⁹ Collectively, these findings position *XIST* not only as a prognostic biomarker but as a therapeutically relevant molecular target whose silencing, through RNA interference or antisense oligonucleotide strategies, may offer novel sex-informed therapeutic avenues specifically for female BCa patients with high *XIST* expression.

In MIBC, other upregulated genes were observed in female patients (Fig. 4) and its biologically significant warrants careful interpretation in the context of tumor biology and potential clinical implications. KRT5, a canonical basal cell marker, defines the basal-like MIBC molecular subtype, which is associated with aggressive tumor behavior but paradoxically greater sensitivity to cisplatin-based neoadjuvant chemotherapy compared to luminal tumors, positioning KRT5 as a predictive marker with direct therapeutic implications.^{20,21} Our *in silico* drug sensitivity prediction supports that MIBC tumors with high KRT5 expression in female patients are preferentially sensitive to cisplatin, docetaxel, and gemcitabine (Fig. 6). CALML3, a calmodulin-like protein expressed in differentiating epithelia, has been reported as downregulated during malignant transformation in other cancer types.²² Its relative overexpression in female BCa samples may reflect retention of a more differentiated epithelial state in females, which warrants further investigation in the context of sex-associated differences in tumor differentiation. CEACAM6 has previously been associated with a malignant program in BCa.²³ In other solid tumors, as in gastric cancer, CEACAM6 drives EMT, invasion, and chemoresistance through PI3K/AKT signaling and MMP-9 upregulation.²⁴ Beyond its tumor-intrinsic oncogenic roles, this gene was reported to suppress cytotoxic CD8+ T cell activity, with anti-CEACAM6 monoclonal antibodies showing promise in restoring antitumor immune responses.²⁵ Whether these mechanisms operate equivalently in MIBC warrants dedicated experimental investigation. In the context of drug sensitivity prediction in MIBC, CEACAM6 emerged as a predictor of response to erdafitinib, an *FGFR1-4* inhibitor approved for locally advanced or metastatic urothelial carcinoma harboring *FGFR2/3* alterations.²⁶ Since *FGFR3* mutations occur in only 10%-15% of MIBC, and routine molecular testing of FGFR status remains costly and time consuming,²⁷ identifying complementary, more broadly expressed predictors of erdafitinib sensitivity, such as CEACAM6, could help extend patient stratification beyond FGFR genotyping alone. This positions CEACAM6 as a candidate biomarker of potential value in MIBC.

The dysregulation and functional implications of Y chromosome genes in BCa remain poorly understood. In BCa, Y-chromosome loss has been reported as a relatively early event, although its role in tumor progression remains unclear. More recently, experimental Y-chromosome deletion has been shown to alter T cell function, promoting T cell exhaustion and increasing sensitivity to PD-1-targeted immunotherapy, highlighting a potential link between Y-chromosome status and immune regulation in cancer¹⁰. Our data do not indicate

Y-chromosome loss, instead, they show higher expression of the Y-linked genes *DDX3Y*, *EIF1AY*, *RPS4Y1*, *TMSB4Y*, *NLGN4Y* and *CYORF15A* in men, compared with women. This pattern most likely reflects a sex-biased transcriptomic signature driven by Y-chromosome dosage, rather than a tumor-specific alteration, and may contribute to the observed molecular dimorphism between male and female bladder tissue.²⁸

This study is limited by the retrospective nature of GEO and TCGA analyses, which involve heterogeneous platforms, processing workflows, and clinical annotations. The marked underrepresentation of women in BCa datasets further limits statistical power for sex-specific comparisons, may mask female-associated molecular features, and also affect the reliability of batch correction procedures, which assume reasonably balanced group sizes. Because the multi-cohort analysis included only three cohorts, heterogeneity estimates were interpreted cautiously and should not be over-generalized. In addition, incomplete clinical metadata (including limited information on smoking history, treatment regimens, and other potential confounders) combined with the absence of functional validation constrain interpretation, as transcriptomic differences cannot be assumed to reflect causal biological mechanisms. Similarly, the *in silico* drug sensitivity predictions are computational estimates rather than measured clinical or experimental drug responses and require independent confirmation. The HPA contextualization is based on non-sex-stratified specimens and therefore cannot directly validate sex-associated differences. Therefore, the findings should be considered hypothesis-generating and require confirmation in independent cohorts and dedicated experimental studies.

ETHICAL DISCLOSURES

Conflicts of Interest: The authors have no conflicts of interest to declare.

Financing Support: This work has not received any contribution, grant or scholarship

Confidentiality of Data: The authors declare that they have followed the protocols of their work center on the publication of patient data.

Protection of Human and Animal Subjects: The authors declare that the procedures followed were in accordance with the regulations of the relevant clinical research ethics committee and those of the Code of Ethics of the World Medical Association (Declaration of Helsinki as revised in 2024).

Provenance and Peer Review: Not commissioned; externally peer-reviewed.

RESPONSABILIDADES ÉTICAS

Conflitos de Interesse: Os autores declaram a inexistência de conflitos de interesse na realização do presente trabalho.

Fontes de Financiamento: Não existiram fontes externas de financiamento para a realização deste artigo.

CONCLUSION

Our results reveal a sex-dependent transcriptional signature in BCa, marked by *XIST* overexpression in female patients, increased expression of *KRT5*, *S100A8*, *CALML3*, *PI3*, and *CEACAM6* in MIBC among women relative to men, and the overexpression of Y chromosome-linked genes in men. These findings suggest that sex shapes not only chromosomal dosage but also the molecular features associated with disease stage, supporting a biologically distinct course of BCa in women and men. Moreover, the female-enriched panel identified here points to potential clinically actionable markers of therapy sensitivity, with implications for sex-informed treatment stratification in BCa.

ACKNOWLEDGMENTS

This work was financed by national funds through FCT, Fundação para a Ciência e Tecnologia, I.P., RISE (LA/P/0053/2020), Institute of Biomedicine (iBiMED) (UIDB/04501/2020, <https://doi.org/10.54499/UID/04501/2025>), LAQV-REQUIMTE (UID/50006/2025-Laboratório Associado para a Química Verde – Tecnologias e Processos Limpos), Centre for the Research and Technology of Agro-Environmental and Biological Sciences (CITAB) UID/04033/2025 (10.54499/UID/04033/2025), and LA/P/0126/2020 (10.54499/LA/P/0126/2020), Institute for Innovation, Capacity Building and Sustainability of Agri-Food Production (Inov4Agro), and by FEDER – Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2030 in the framework of the project COMPETE2030-FEDER 00733600 (SEXFIT2CARE). For the purpose of Open Access, the author applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

Confidencialidade dos Dados: Os autores declaram ter seguido os protocolos da sua instituição acerca da publicação dos dados de doentes.

Proteção de Pessoas e Animais: Os autores declaram que os procedimentos seguidos estavam de acordo com os regulamentos estabelecidos pela Comissão de Ética responsável e de acordo com a Declaração de Helsínquia revista em 2024 e da Associação Médica Mundial.

Proveniência e Revisão por Pares: Não comissionado; revisão externa por pares.

CONTRIBUTORSHIP STATEMENT

RF: Write, data collection, interpretation.

PO and FA: Write, interpret.

LLS: Clinical data validation, write.

RV: Write, data collection, interpretation.

All authors approved the final version of the manuscript for publication and assume responsibility for all aspects of the work, ensuring the accuracy and integrity of the data presented.

DECLARAÇÃO DE CONTRIBUIÇÃO

RF: Redação, recolha de dados, interpretação.

PO e FA: Redação, interpretação.

LLS: Validação dos dados clínicos, redação.

RV: Redação, recolha de dados, interpretação.

Todos os autores aprovaram a versão final do manuscrito para publicação e assumem responsabilidade por todos os aspetos do trabalho, garantindo a exatidão e a integridade dos dados apresentados.

REFERENCES

1. Filho AM, Briganti A, Jemal A, Bray F. Bladder Cancer Incidence and Mortality: A Global Overview and Recent Trends. *Eur Urol*. 2026;89:426-436. doi: 10.1016/j.eururo.2025.12.011
2. Radkiewicz C, Edgren G, Johansson AL, Jahnson S, Häggström C, Akre O, et al. Sex Differences in Urothelial Bladder Cancer Survival. *Clin Genitourin Cancer*. 2020;18:26-34.e26. doi: 10.1016/j.clgc.2019.10.020
3. Isali I, Khooblal P, Helstrom E, Bukavina L. Targeting bladder cancer: A sex sensitive perspective in mutations and outcomes. *Urol Oncol*. 2025;43:673-679. doi: 10.1016/j.urolonc.2023.05.008
4. de Jong JJ, Boormans JL, van Rhijn BW, Seiler R, Boorjian SA, Konety B, et al. Distribution of Molecular Subtypes in Muscle-invasive Bladder Cancer Is Driven by Sex-specific Differences. *Eur Urol Oncol*. 2020;3:420-3. doi: 10.1016/j.euo.2020.02.010
5. Chenard S, Jackson C, Vidotto T, Chen L, Hardy C, Jamaspishvili T, et al. Sexual Dimorphism in Outcomes of Non-muscle-invasive Bladder Cancer: A Role of CD163+ Macrophages, B cells, and PD-L1 Immune Checkpoint. *Eur Urol Open Sci*. 2021;29:50-58. doi: 10.1016/j.euros.2021.05.002
6. Theodorescu D, Li Z, Li X. Sex differences in bladder cancer: emerging data and call to action. *Nat Rev Urol*. 2022;19:447-9. doi: 10.1038/s41585-022-00591-4
7. Toren P, Wilkins A, Patel K, Burley A, Gris T, Kockelbergh R, et al. The sex gap in bladder cancer survival – a missing link in bladder cancer care? *Nat Rev Urol*. 2024;21:181-192. doi: 10.1038/s41585-023-00806-2
8. Li P, Chen J, Miyamoto H. Androgen Receptor Signaling in Bladder Cancer. *Cancers*. 2017;9. doi: 10.3390/cancers9020020
9. Kaneko S, Li X. X chromosome protects against bladder cancer in females via a KDM6A-dependent epigenetic mechanism. *Sci Adv*. 2018;4:eaar5598. doi: 10.1126/sciadv.aar5598
10. Abdel-Hafiz HA, Schafer JM, Chen X, Xiao T, Gauntner TD, Li Z, et al. Y chromosome loss in cancer drives growth by evasion of adaptive immunity. *Nature*. 2023;619:624-31. doi: 10.1038/s41586-023-06234-x
11. Doshi B, Athans SR, Woloszynska A. Biological differences underlying sex and gender disparities in bladder cancer: current synopsis and future directions. *Oncogenesis*. 2023;12:44. doi: 10.1038/s41389-023-00489-9
12. Sjödal G, Lauss M, Lövgren K, Chebil G, Gudjonsson S, Veerla S, et al. A molecular taxonomy for urothelial carcinoma. *Clin Cancer Res*. 2012;18:3377-86. doi: 10.1158/1078-0432.Ccr-12-0077-t
13. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell*. 2014;25:152-65. doi: 10.1016/j.ccr.2014.01.009
14. 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;9:3. doi: 10.1186/1476-4598-9-3



15. Wang W, Min L, Qiu X, Wu X, Liu C, Ma J, et al. Biological Function of Long Non-coding RNA (LncRNA) Xist. *Front Cell Dev Biol.* 2021;9. doi: 10.3389/fcell.2021.645647
16. Cai F, Xu S, Li Y, He Q, Su Q, Chen H, et al. The role of the LncRNA XIST/miR-15a-5p/MN1 signaling axis in gender disparities in bladder cancer prognosis. *Front Immunol.* 2025;16:1554829. doi: 10.3389/fimmu.2025.1554829
17. Brockdorff N, Bowness JS, Wei G. Progress toward understanding chromosome silencing by Xist RNA. *Genes Dev.* 2020;34:733-44. doi: 10.1101/gad.337196.120
18. Kong YL, Wang HD, Gao M, Rong SZ, Li XX. LncRNA XIST promotes bladder cancer progression by modulating miR-129-5p/TNFSF10 axis. *Discov Oncol.* 2024;15:65. doi: 10.1007/s12672-024-00910-8
19. Li K, Yao T, Wang Z. LncRNA-mediated ceRNA network in bladder cancer. *Non-coding RNA Res.* 2023;8:135-145. doi: 10.1016/j.ncrna.2022.12.002
20. Eckstein M, Wirtz RM, Gross-Weege M, Breyer J, Otto W, Stoehr R, et al. mRNA-Expression of KRT5 and KRT20 Defines Distinct Prognostic Subgroups of Muscle-Invasive Urothelial Bladder Cancer Correlating with Histological Variants. *Int J Mol Sci.* 2018;19:3396. doi: 10.3390/ijms19113396
21. Sjö Dahl G, Abrahamsson J, Bernardo C, Eriksson P, Höglund M, Liedberg F. Molecular subtypes as a basis for stratified use of neoadjuvant chemotherapy for muscle-invasive bladder cancer: a narrative review. *Cancers.* 2022;14. doi: 10.3390/cancers14071692
22. Brooks MD, Bennett RD, Weaver AL, Sebo TJ, Eckert SE, Strehler EE, et al. Human Calmodulin-Like Protein CALML3: A Novel Marker for Normal Oral Squamous Mucosa That Is Downregulated in Malignant Transformation. *Int J Dent.* 2013;2013:592843. doi: 10.1155/2013/592843
23. Wang P, Li C, Xu Y, Altuwajiri S, Niu Y, Zhao Y, et al. CEACAM6 regulates glycolytic metabolism in bladder cancer cell by controlling ENO1 stability. *Eur J Med Res.* 2025;30:1233. doi: 10.1186/s40001-025-03659-0
24. Zang M, Zhang B, Zhang Y, Li J, Su L, Zhu Z, et al. CEACAM6 promotes gastric cancer invasion and metastasis by inducing epithelial-mesenchymal transition via PI3K/AKT signaling pathway. *PLoS One.* 2014;9:e112908. doi: 10.1371/journal.pone.0112908
25. Burgos M, Cavero-Redondo I, Álvarez-Bueno C, Galán-Moya EM, Pandiella A, Amir E, et al. Prognostic value of the immune target CEACAM6 in cancer: a meta-analysis. *Ther Adv Med Oncol.* 2022;14:17588359211072621. doi: 10.1177/17588359211072621
26. Loriot Y, Matsubara N, Park SH, Huddart RA, Burgess EF, Houede N, et al. Erdafitinib or Chemotherapy in Advanced or Metastatic Urothelial Carcinoma. *N Engl J Med.* 2023;389:1961-71. doi: 10.1056/NEJMoa2308849
27. Bannier PA, Saillard C, Mann P, Touzot M, Maussion C, Matek C, et al. AI allows pre-screening of FGFR3 mutational status using routine histology slides of muscle-invasive bladder cancer. *Nat Commun.* 2024;15:10914. doi: 10.1038/s41467-024-55331-6
28. Sauter G, Moch H, Wagner U, Novotna H, Gasser TC, Mattarelli G, et al. Y chromosome loss detected by FISH in bladder cancer. *Cancer Genet Cytogenet.* 1995;82:163-9. doi: 10.1016/0165-4608(95)00030-s

