分类号：
学　号：20232114205

密　　级：公开
单位代码：10759

石河子大学

硕士学位论文

整合单细胞空间转录组及全外显子测序初探膀胱癌枢纽基因及上皮细胞异质性

学位申请人　李琳
指导教师　李前跃教授
申请学位类别　专业硕士
专业名称　临床医学
研究领域　外科学
所在学院　临床医学院

中国·新疆·石河子
2026 年 5 月

分类号：　　密　　级：公开
学　号：20232114205　　单位代码：10759

石河子大学

硕 士 学 位 论 文

整合单细胞空间转录组及全外显子测序初探膀胱癌枢纽基因及上皮细胞异质性

学 位 申 请 人　李琳
指 导 教 师　李前跃教授
申请学位类别　专业硕士
专 业 名 称　临床医学
研 究 领 域　外科学
所 在 学 院　临床医学院

中国·新疆·石河子

2026 年 05 月

Integration of Single-Cell and Spatial Transcriptomics with Whole-Exome Sequencing to Explore Hub Genes and Epithelial Cell Heterogeneity in Bladder Cancer

A Dissertation Submitted to

Shihezi University
In Partial Fulfillment of the Requirements

for the Degree of

Master of Medicine

By

LiLin
(Surgery)

Dissertation Supervisor: Prof. LI Qianyue

May, 2026

石河子大学学位论文独创性声明及使用授权声明

学位论文独创性声明

本人所呈交的学位论文是在我导师的指导下进行的研究工作及取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示谢意。

研究生签名：  时间：2026 年 5 月 21 日

使用授权声明

本人完全了解石河子大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或指定机构送交论文的电子版和纸质版。有权将学位论文在学校图书馆保存并允许被查阅。有权自行或许可他人将学位论文编入有关数据库提供检索服务。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

研究生签名：  时间：2026 年 5 月 21 日

导师签名：  时间：2026 年 5 月 21 日

摘要

目的：膀胱癌（Bladder Cancer, BLCA）是一种起源于尿路上皮的恶性肿瘤，其发生发展伴随着肿瘤微环境（Tumor Microenvironment, TME）的动态重塑及肿瘤突变负荷（Tumor Mutational Burden, TMB）的异常积累。而癌相关上皮细胞（Cancer-associated Epithelial Cells, EpiCs）作为膀胱癌的重要组成部分，在肿瘤进展过程中表现出明显的异质性，不同功能状态的 EpiCs 可能在微环境互作及疾病演进中发挥不同作用。因此，本研究系统性地分析 EpiCs 的异质性及功能特征，初步探讨其与 TME 和 TMB 的关系，并筛选出与膀胱癌进展相关的枢纽基因，为 TMB 与 TME 在膀胱癌中的后续研究提供初步理论基础。

方法：本研究整合 13 例单细胞 RNA 测序（Single-cell RNA sequencing, scRNA-seq）样本、514 例批量转录组测序样本以及 30 例全外显子测序（Whole-Exome Sequencing, WES）样本，对 EpiC 亚型进行系统性分析。采用统一流形近似与投影（Uniform Manifold Approximation and Projection, UMAP）进行非线性降维分析，结合聚类算法识别主要上皮细胞亚群，随后进行二次聚类分析以精细化分群。基于自测的 WES 数据计算出的 TMB 值，将其整合至单细胞活性评分算法（single-cell Activity-Based scoring, scAB）及观察值/期望值比值算法（Ratio of Observed to Expected, Ro/e）中，用于筛选与 TMB 显著相关的上皮细胞亚群，最终确定关键细胞簇 Epi14。对 Epi14 的差异表达基因（Differentially Expressed Genes, DEGs）进行筛选与功能分析，并利用细胞通讯推断细胞间信号通讯网络。应用细胞分化潜能预测算法及单细胞轨迹分析评估细胞干性潜能与分化轨迹，结合随机生存森林模型（Random Survival Forest, RSF）与 DEGs 分析结果筛选关键枢纽基因。根据筛选的枢纽基因进行免疫浸润、药物敏感性预测及功能通路富集等分析，并使用 Pathway RespOnsive GENes、CellChat 构建细胞间配体—受体通讯网络与通路活性分析，观察不同区域中信号通路的活跃程度及其潜在调控关系。最后选取候选枢纽基因，在临床收集的膀胱癌组织及对应癌旁组织中开展实时荧光定量聚合酶链式反应（Quantitative Real-Time Polymerase Chain Reaction，qPCR）和蛋白质免疫印迹实验（Western blot，WB）以验证其表达差异。

结果：本研究共纳入77,263个细胞，并筛选得到3,000个高变异基因。经过降维和聚类分析后，共识别出32个注释较为明确的细胞簇。经对EpiCs二次聚类划分出14个亚群，结合scAB、Ro/e等算法和TMB，确定了Epi14为关键亚群。通过DEGs分析、RSF模型、多队列数据、WB以及PCR等的验证，最终筛选出了ABRACL和ARPC3为潜在的核心枢纽基因。

结论：本研究整合多组数据，对膀胱癌 EpiCs 的异质性特征进行了初步分析。结果显示：ABRACL 与 ARPC3 在 EpiCs 中表现出与 TMB 相关的表达特征，可能参与膀胱癌相关的分子调控过程。为进一步探讨膀胱癌的分子机制提供了新的研究线索，也为相关靶点的后续研究提供了初步参考。

关键词：膀胱癌；癌相关上皮细胞；单细胞测序；肿瘤突变负荷；肿瘤微环境

Abstract

Objective: Bladder cancer is a malignant tumor originating from the urothelium. And Bladder cancer is also characterized by a complicated tumor microenvironment and a prominent tumor mutational burden exists.,whch changed in the tumor microenvironment and the gradual accumulation of tumor mutational burden, are widely regarded as extremely important factors influencing the unfolding and progression of bladder cancer. During tumor evolution, cancer-associated epithelial cells display notable heterogeneity. Epithelial cells in different functional states, which may participate in microenvironmental interactions in distinct ways and also contribute differently to disease progression. A closer examination of epithelial cell diversity and their functional characteristics, that along with the identification of key hub genes within these cells, help clarify the molecular mechanisms underlying bladder cancer. It also provide useful clues for subsequent studies.

Methods: This study, integrated several types of omics data, explore epithelial cell heterogeneity in bladder cancer including 13 single-cell RNA sequencing samples, 514 transcriptome sequencing samples, and 30 whole-exome sequencing samples. Single-cell data were processed using Uniform Manifold Approximation and Projection for nonlinear dimensionality reduction, and then followed by clustering to recognition the major epithelial cell populations. To obtain a detailed classification, Which means epithelial cells were extracted for secondary classification to further divide the epithelial subgroups. Tumor mutational burden values were calculated from the whole-exome sequencing data generated in this study. These values were combined with single-cell transcriptomic profiles using the single-cell activity-based scoring method and the ratio of observed to expected algorithm to estimate the association between individual cells and TMB status. Through this process, the epithelial cluster Epi14 was identified as a subgroup closely related to TMB. Genes specifically expressed in Epi14 were then screened and subjected to functional analysis. Cell–cell communication patterns were explored to examine potential signaling interactions among different cell populations. The differentiation potential of epithelial cells and their developmental trajectories were evaluated using CytoTRACE together with single-cell trajectory analysis. Candidate hub genes were further selected by integrating differential gene expression results with random survival forest modeling. Immune cell infiltration patterns, drug sensitivity prediction, and pathway enrichment analyses were also carried out. Spatial transcriptomic data were analyzed using deconvolution methods, CellChat, and the pathway activity inference approach PROGENy to examine the spatial distribution of cell populations and their signaling activity. Finally, quantitative real-time polymerase chain reaction and Western blot experiments were performed to examine the expression levels of the selected hub genes in tumor tissues and matched adjacent normal tissues.

Results: This study included a total of 77,263 cells and identified 3,000 highly variable genes. After dimensionality reduction and clustering analysis, 32 well-annotated cell clusters were recognized. Secondary clustering of EpiCs revealed 14 subpopulations, and by integrating scAB, Ro/e algorithms, and TMB analysis, Epi14 was determined as the key subpopulation. Through differential gene expression (DEG) analysis, RSF modeling, multi-cohort data validation, Western blotting (WB), and PCR experiments, ABRACL and ARPC3 were ultimately identified as potential core hub genes.

Conclusion: By integrating multi-omics and spatial transcriptomic data, this study preliminarily elucidates the heterogeneity of epithelial cells in bladder cancer and identifies ABRACL and ARPC3 as key tumor mutational burden-associated hub genes. These findings provide an important theoretical foundation and potential therapeutic targets for future mechanistic studies and clinical translation.

Key words: Bladder cancer; Cancer-associated epithelial cells; Single-cell sequencing; Tumor mutational burden; Tumor microenvironment.

目录

主要符号对照表

中文全称	英文缩写	英文全称
膀胱癌	BLCA	Bladder Cancer
肿瘤突变负荷	TMB	Tumor Mutational Burden
肿瘤微环境	TME	Tumor Microenvironment
单细胞 RNA 测序	scRNA-seq	Single-cell RNA sequencing
空间转录组	ST	Spatial Transcriptomics
全外显子测序	WES	Whole-Exome Sequencing
癌相关上皮细胞	EpiCs	Cancer-associated Epithelial Cells
主成分分析	PCA	Principal Component Analysis
统一流形近似与投影	UMAP	Uniform Manifold Approximation and Projection
单细胞活性评分算法	scAB	Single-cell Association with Bulk phenotype
观察值/期望值比值	Ro/e	Ratio of Observed to Expected
差异表达基因	DEGs	Differentially Expressed Genes
随机生存森林	RSF	Random Survival Forest
淋巴结转移	LNM	Lymph node metastasis
基因集富集分析	GSEA	Gene Set Enrichment Analysis
基因集变异分析	GSVA	Gene Set Variation Analysis
决策曲线分析	DCA	Decision Curve Analysis
受试者工作特征曲线	ROC	Receiver Operating Characteristic Curve
曲线下面积	AUC	Area Under the Curve
程序性死亡配体 1	PD-L1	Programmed Death-Ligand 1
癌症药物敏感性基因组	GDSC	Genomics of Drug Sensitivity in Cancer
癌症基因组图谱	TCGA	The Cancer Genome Atlas
基因表达综合数据库	GEO	Gene Expression Omnibus
蛋白质免疫印迹	WB	Western Blot
实时荧光定量聚合酶链式反应	RT-qPCR	Real-Time Quantitative Polymerase Chain Reaction

第 1 章 前言

膀胱癌是泌尿系统中最常见的恶性肿瘤之一，起源于尿路上皮细胞[1]。其发病率和死亡率近年来持续处于较高水平，并呈逐步上升趋势。根据 2022 年全球癌症统计数据，全球癌症新发病例已接近 2000 万例，死亡病例超过 970 万例，而膀胱癌分别占全球恶性肿瘤的 3.1%和 2.3%[2-3]。以 2020 年的统计数据为例，当年全球新发膀胱癌病例约 50 万例，占全部新发癌症的约 3%，发病率在所有癌症中位居第 10 位；同期死亡病例约为 20 万例[4]。与 2022 年全球数据相比，膀胱癌的发病率和死亡率仍呈上升趋势，说明其总体疾病负担依然在增加。随着人口老龄化程度不断加深、膀胱癌患病的危险因素仍持续存在，其流行情况仍然较为严峻，也对医疗资源配置和公共卫生体系带来了持续压力[5]。尽管随着膀胱癌早期筛查水平的提高以及影像学诊断技术的不断发展，手术治疗、化疗、靶向治疗和免疫治疗等多种治疗手段也在不断改善。但从临床实际情况来看，膀胱癌患者的长期预后改善仍然有限。复发率较高、疾病进展较快以及远处转移等问题依然较为常见，这些因素在一定程度上影响了整体治疗效果[6-7]。从全球范围来看，上述情况也在一定程度上反映出膀胱癌带来的疾病负担仍然较重，其在公共卫生领域中的关注度也随之不断提高[8-10]。

当前膀胱癌的整体治疗效果仍存在一定的局限，这与其复杂的肿瘤生物学特征有关。膀胱癌并不是一种由单一因素驱动的恶性肿瘤，而是在长期发展过程中逐渐形成的一类具有明显分子差异的疾病[11]。在不同患者之间，甚至同一肿瘤内部的不同区域，肿瘤的基因突变谱、信号通路活性以及细胞组成等方面都可能存在很大差异，而这些差异大大的增加了疾病诊断分型和治疗决策的难度。同时，肿瘤微环境（Tumor Microenvironment, TME）在肿瘤发生发展过程中也发挥着重要作用[12-13]。其由多种细胞成分以及细胞外基质共同构成，包括免疫细胞、成纤维细胞、内皮细胞和巨噬细胞等。不同细胞之间可以通过细胞因子、趋化因子以及多种信号通路相互作用，逐渐形成复杂的调控网络，并持续影响肿瘤细胞的生长、侵袭以及对治疗的反应[14-17]。而肿瘤突变负荷（Tumor Mutational Burden, TMB）常被用来反映肿瘤体细胞突变的总体水平。在不同患者之间的 TMB 水平差异较为明显，而较高的 TMB 往往会产生更多新抗原，从而增强肿瘤细胞对免疫检查点抑制剂的反应[18-21]。在包括膀胱癌在内的多种实体瘤中，例如黑色素瘤、肺癌和肝癌等，较高的 TMB 水平通常与更好的免疫治疗效果有关[22-23]。但 TMB 主要反映的是整体基因组层面的突变情况，对于不同细胞亚群之间的转录差异以及在组织中的空间分布情况仍然难以全面反映，因此仅依赖突变数量往往不足以解释患者之间在治疗反应和预后方面存在的明显差异。而膀胱癌起源于尿路上皮细胞，在肿瘤发生发展的过程中，上皮

细胞逐渐获得异常的增殖能力和侵袭能力，并进一步转变为癌相关上皮细胞。这些细胞不仅参与肿瘤形成，还可以通过分泌多种调控因子以及改变细胞外基质结构，参与肿瘤微环境的形成和维持[24-26]。不同上皮细胞亚群在基因表达、代谢状态以及信号通路活性方面往往存在明显差异，这种细胞层面的异质性被认为与肿瘤进展、免疫逃逸以及治疗耐药等过程有关[27-29]。因此，从细胞亚群的角度对上皮细胞进行讨论分析，并进一步探索其与 TMB 之间的关系，有助于从不同角度去理解膀胱癌分子演变的过程。

随着测序技术和生物信息学方法的不断发展，多组学数据逐渐被应用于肿瘤研究中，也为从不同层面理解肿瘤生物学特征提供了新的研究思路。单细胞测序技术能够在单细胞水平上分析细胞的转录表达情况，从而更深度的观察细胞群体之间的差异情况，识别出与传统总体转录组分析中较难发现的细胞亚群及其状态变化[30-31]。但单细胞测序本身缺乏空间信息，因此很难反映细胞在组织结构中的真实分布情况。而空间转录组技术在保留组织结构的同时，会将基因表达信息与空间位置进行对应，从而能够观察到不同细胞群体在肿瘤中的分布情况和其可能的相互作用[32-33]。而全外显子测序可用于分析体细胞突变谱并计算 TMB，而机器学习方法则能够在大量数据中筛选出与疾病相关的关键变量，用于构建预测模型并开展生存分析或风险评估[34-35]。将上述分析的数据进行整合，从单细胞水平、基因组方面以及空间表达维度对肿瘤进行综合分析，从而得到更全面地理解肿瘤内部的异质性。基于这一思路，本研究利用 scAB 和 Ro/e 等算法将膀胱癌突变负荷相关信息映射至特定的细胞亚群，并结合空间转录组数据进行分析，来探索其在膀胱癌中可能参与的关键细胞亚群[36]。目前已有研究从基因突变或转录表达等角度对膀胱癌进行了较多探索，但将 TMB 与上皮细胞亚群特异性表达特征结合起来进行系统分析的研究仍然相对较少[37-39]。因此，本研究以上皮细胞异质性与肿瘤突变负荷相关机制为切入点，整合单细胞转录组、空间转录组、全外显子测序及总体转录组等多组学数据，构建多维度联合分析框架。并使用单细胞分析探索上皮细胞的亚群分型，将 WES 计算得到的 TMB 与转录特征相结合，映射至不同上皮细胞亚群，以识别与突变负荷显著相关的特异性细胞状态；随后在此基础上开展差异表达分析并引入生存模型筛选潜在关键枢纽基因。并进一步在多队列数据中对候选基因进行表达一致性与预后相关性验证，结合空间转录组对其组织内分布特征及功能关联进行交叉验证，从细胞亚群、分子特征及空间定位等多个角度系统评估关键分子的潜在生物学意义与临床价值，为以后后续研究提供初步基础理论。

综上，本研究通过整合多组学数据，对膀胱癌上皮细胞的异质性特征及其与肿瘤突变负荷之间的关系进行了初步探索，希望能够为进一步理解膀胱癌的发生发展机制提供新的研究线索。

第 2 章 资料与方法

2.1 BLCA 枢纽基因多组学分析

2.1.1 研究对象

本研究所使用的临床样本均来源于新疆生产建设兵团医院泌尿外科。收集自 2024 年 9 月至 2025 年 9 月，共 30 对膀胱癌患者的肿瘤组织及对应癌旁正常组织样本。所有患者均经术后病理学确诊为膀胱癌，且在手术前未接受放疗、化疗或免疫治疗等抗肿瘤干预措施。组织样本均由至少 2 名以上具有丰富经验的病理科医师进行确认，确保肿瘤组织中癌细胞含量符合后续分子检测要求，同时癌旁组织取材距离肿瘤边缘足够安全，以排除肿瘤细胞浸润的干扰。在所收集的 30 对样本中，其中 15 对样本送至第三方公司进行转录组测序及全外显子测序，获得测序的原始数据，以用于后续转录组表达谱构建及体细胞突变谱分析；剩余 15 对组织样本在离体后立即置于液氮中快速冷冻保存，并转运至实验室进行后续分子生物学实验及验证分析，以确保 RNA 及蛋白质的完整性与稳定性。本研究方案经新疆生产建设兵团医院伦理委员会审批通过，所有入组患者在术前均签署书面知情同意书。本次研究的全过程严格遵循《赫尔辛基宣言》的相关伦理原则。

单细胞测序数据来源于 Gene Expression Omnibus（GEO）数据库。选取数据集 GSE222315，该数据集包含 13 例具有完整单细胞转录组数据的样本，其中包括 4 例正常膀胱组织样本和 9 例膀胱癌肿瘤组织样本。转录组数据来源于 GSE236932 数据集（平台编号 GPL24676），共纳入 63 例样本，其中正常组织 25 例，肿瘤组织 38 例。空间转录组数据下载自 GSE246011 数据集，该数据集包含 2 例膀胱癌样本的完整空间转录组测序信息，可用于构建空间维度上的细胞分布图谱及信号通路活性网络。此外，为进一步扩大样本量并增强研究结果的稳定性，本研究还从 The Cancer Genome Atlas（TCGA）数据库中下载 421 例膀胱癌患者的转录组数据，其中包括 19 例正常组织样本及 402 例肿瘤组织样本。

2.1.2 TMB 值的计算

将 15 对 BLCA 肿瘤样本及其对应的相邻组织样本使用 Agilent SureSelect 人类全外显子 V6 试剂盒（靶区域约 38 Mb）进行了 WES 测序。在获得原始数据后，我们进行了体细胞突变分析，使用 ANNOVAR 对测序公司提供的突变文件进行注释。为避免因多

个转录本注释导致的重复计数，本研究根据染色体（Chr）、位置（Pos）、参考等位基因（Ref）和替代等位基因（Alt）为每个位点生成唯一的突变标识符，并对重复记录进行去除。保留非同义外显子突变（包括错义突变、无义突变、剪接突变、移码插入/删除以及非移码插入/删除变异），同时排除同义和非编码突变。通过将过滤后的非同义突变总数除以 Agilent V6 采样区域的大小（38 Mb）来计算每个样本的 TMB，并以突变/兆碱基为单位表示。鉴于膀胱癌的总体突变负荷水平通常低于黑色素瘤及肺癌等高突变负荷肿瘤类型，若采用固定阈值（如 10 mut/Mb）进行分组，可能在中小样本队列中造成样本分布不均衡，从而影响后续统计分析结果。因此，本研究未采用统一固定阈值进行高低分组，而是依据本队列的 TMB 中位数将样本划分为 TMB-high 组与 TMB-low 组，以减少样本分布不均对后续统计分析的影响。

2.1.3 单细胞 RNA 测序数据质控与细胞注释

使用 Seurat 包进行数据处理，导入原始表达矩阵并构建 Seurat 对象，对每个细胞的分子标识符总数、检测到的基因数量以及线粒体基因表达比例进行统计。异常值定义为偏离中位数超过 3 倍中位绝对偏差（median absolute deviations, MADs）的细胞。根据 MAD 分布结果设定筛选阈值，保留检测基因数介于 200 至 4612 之间、UMI 总数小于 13000.00 且线粒体基因比例低于 5%的细胞用于后续分析。数据标准化采用 LogNormalize 方法，将每个细胞的表达值缩放至 10,000 个 UMI 后进行对数转换。使用 CellCycleScoring 函数计算细胞周期评分，利用 FindVariableFeatures 筛选高变异基因。随后通过 ScaleData 函数对线粒体基因比例、核糖体基因比例及细胞周期评分进行回归处理，以降低技术因素对表达矩阵的影响。降维分析采用 RunPCA 进行主成分分析，在此基础上使用 Harmony 算法进行批次效应校正。校正后数据通过 RunUMAP 进行非线性降维并生成 UMAP 嵌入结果，用于后续聚类分析及可视化展示。细胞类型注释基于已发表文献报道的经典标志基因及 CellMarker 数据库进行初步判定，同时结合 SingleR 进行自动化预测。最终结合各聚类的基因表达特征进行人工手动校正，确定各细胞簇的类型，并筛选代表性标志亚群用于后续分析。

2.1.4 基于 scAB 与 Ro/e 算法鉴定 TMB 相关上皮细胞亚群

提取注释后的 EpiCs 进行二次分析。将 EpiCs 重新构建 Seurat 对象，进行标准化与归一化处理后，依次实施 PCA 和 UMAP 非线性降维，以获得更加精细的 EpiCs 亚群结构。并基于 WES 计算得到的 15 例患者 TMB 数值，以其中位数为界将样本划分为 TMB-high 组和 TMB-low 组，用于后续表型分析。

因本次研究单细胞数据与自测的 TMB 数据不是匹配数据，所以此次研究在单细胞

与样本层面表型整合过程中，并非直接将 TMB 数值简单赋予单细胞数据，而是基于单细胞表达谱与自测 bulk RNA-Seq 转录组数据表达模式之间的关联关系进行整合分析。scAB 等算法通过联合单细胞转录组数据与 bulk RNA-Seq 表达矩阵，在计算每个单细胞与各 bulk RNA-Seq 样本表达相似性的基础上，将样本层面的 TMB 表型信息量化映射至单细胞层面。该方法综合考虑单细胞与不同 bulk RNA-Seq 样本之间的表达相关性，并结合对应样本的 TMB 分组信息，为每个细胞计算表型关联概率值，从而实现样本 TMB 表型向单细胞层级的统计学映射，而非直接强制赋值。scAB 最终为每个细胞生成一个与 TMB-high 表型相关的概率评分，用于衡量该细胞转录特征与高 TMB 相关表达模式的一致程度。该概率值越高，表示该细胞在表达模式上越接近 TMB-high 样本的转录特征。

为进一步评估不同上皮细胞亚群中 TMB 相关细胞的富集程度，采用 Ro/e 方法对各亚群进行统计分析。通过比较每个亚群中 scAB 鉴定细胞的实际观察数量与在随机分布假设下的期望数量，计算其比值。Ro/e 值大于 1 表示该亚群中 TMB 相关细胞数量高于随机期望水平。通过 scAB 概率评分与 Ro/e 富集分析的联合筛选，最终鉴定 Epi14 亚群为与 TMB-high 表型显著相关的上皮细胞亚群。

2.1.5 差异表达基因筛选与功能富集分析

提取与 Epi14 亚群显著相关的基因，在 scAB 鉴定细胞与 Other 细胞之间进行差异表达分析。差异表达基因的筛选标准设定为$|\log2FC| > 0.585$ 且调整后 P 值< 0.05。获得的差异表达基因随后用于功能富集分析。并对筛选得到的基因集进行注释与功能分类分析，包括 Gene Ontology（GO）功能注释和 Kyoto Encyclopedia of Genes and Genomes（KEGG）通路分析。通过对生物过程、分子功能、细胞组分及信号通路的系统富集分析，识别差异表达基因所涉及的潜在生物学功能及相关调控通路。

2.1.6 配体—受体互作分析及细胞分化与拟时序分析

为评估不同 TMB 水平下 EpiCs 在细胞间通讯及分化状态方面的差异，依据前述 TMB 相关分析结果，将上皮细胞划分为 TMB-Epi 组与 Other epithelial 组。基于标准化后的单细胞表达矩阵，采用 CellChat R 包构建两组细胞之间的配体—受体相互作用网络。根据内置的配体—受体数据库筛选潜在相互作用，计算各配体—受体对在不同细胞群体之间的通信概率，并进一步整合为信号通路层面的通讯强度评分与数量评分。在此基础上，对两组之间的信号通路活跃程度进行比较，识别显著富集的细胞通讯模式及潜在关键调控通路，以定量不同 TMB 状态下上皮细胞的通讯特征。

在细胞分化潜能分析方面，本研究基于单细胞转录复杂度信息，采用 CytoTRACE

算法对每个细胞的干性水平进行推断。该算法通过整合基因表达多样性与表达谱复杂度，计算细胞的干性评分。根据评分结果对细胞的发育成熟度及潜在分化方向进行评估，并比较不同亚群之间的干性分布差异，以分析 TMB 相关细胞群体在分化状态上的特征。

为进一步分析 EpiCs 的动态分化过程，采用 Monocle 2 软件包进行拟时序分析。筛选用于构建轨迹的高变基因并进行降维处理，随后基于最小生成树算法重建细胞发育轨迹结构，推断潜在的分化路径及谱系分支。根据轨迹排序结果，对关键基因在拟时序过程中的表达变化趋势进行分析，识别沿分化进程呈现动态调控特征的基因表达模式，为后续功能关联分析提供依据。

2.1.7 机器学习联合全外显子明确关键基因

采用随机生存森林算法的机器学习方法筛选 EpiCs 中的具有预后意义的枢纽基因。利用 randomForestSRC 包对前述 EpiCs 中筛选得到的差异表达基因进行随机生存森林分析，以总体生存数据作为结局变量构建模型，并计算各基因的相对重要性指标。设定相对重要性值> 0.2 作为筛选标准，确定最终特征基因。随后，基于 TCGA-BLCA 队列的总体生存数据对候选基因进行生存分析。总体生存定义为自初次诊断至任意原因死亡或末次随访时间。未发生终点事件的患者在末次随访时进行删失处理。对筛选得到的候选基因，在自建数据集、TCGA 队列及 GSE236932 数据集中比较肿瘤组织与癌旁正常组织之间的表达差异，以验证其表达模式的一致性。进一步结合全外显子测序数据，分析候选基因表达水平与 TMB 之间的相关性，以筛选与肿瘤细胞 TMB 状态密切相关的核心枢纽基因。

2.1.8 列线图模型构建与药物敏感性分析

基于膀胱癌患者的临床资料进行数据整理，剔除存在关键信息缺失的样本后纳入后续分析。采用回归分析构建预测 1 年、3 年及 5 年总体生存率的列线图模型。模型的区分能力通过受试者工作特征曲线进行评估，并计算曲线下面积以衡量预测准确性。采用决策曲线分析评价模型在不同风险阈值下的临床应用价值。在药物敏感性分析方面，本研究采用 oncoPredict 包，结合其内置的回归预测框架，对膀胱癌常用化疗药物的半数抑制浓度（half maximal inhibitory concentration，IC50）进行预测，从而评估枢纽基因表达水平与药物反应之间的潜在关联。以癌症药物敏感性基因组数据库（Genomics of Drug Sensitivity in Cancer，GDSC） 中的药物反应数据及对应的基因表达谱作为训练数据集。基于 GDSC 数据构建药物敏感性预测模型，通过将样本的基因表达信息与已知的药物 IC50 值进行关联建模，从而建立基因表达模式与药物反应之间的预测关系。最后采用 10 折交叉验证方法对模型进行训练与验证。