基于孟德尔随机化的前列腺癌全转录组关联分析*

南方医科大学公共卫生学院生物统计学系(510515) 朱秋雯 王靖凯 张 顺 麦家豪 周基元△

【提 要】 **目的** 探索前列腺癌的遗传机制,识别前列腺癌的因果基因,为前列腺癌的分子机制提供新见解。**方法** 采用基于孟德尔随机化的IVW方法和PMR-Egger模型实施全转录组关联分析,从基因表达和可变剪切两个角度识别对前列腺癌具有因果效应的基因,随后对两种方法所涉及的公共因果基因进行基因本体富集分析及蛋白质互作网络分析。**结果** 采用IVW方法和PMR-Egger方法对前列腺癌数据进行分析,分别识别出788个和495个基因的表达具有显著因果效应以及1054个和865个基因的可变剪切具有显著因果效应。富集分析表明,部分因果基因显著富集于MHC信号通路及肿瘤免疫反应相关的生物过程,PPI网络分析展示了与前列腺癌发病相关的核心蛋白,如MYC等。**结论** 本研究提供了前列腺癌的部分潜在分子遗传学特征及可能的治疗靶点,为前列腺癌的实验研究、预防策略和临床诊疗提供了重要参考依据。

【关键词】 前列腺癌 孟德尔随机化 全转录组关联分析 富集分析 PPI网络分析
【中图分类号】 R737.25 【文献标志码】 A **DOI** 10.11783/j.issn.1002-3674.2026.04.005

Transcriptome-wide Association Study of Prostate Cancer based on Mendelian Randomization

Zhu Qiuwen, Wang Jingkai, Zhang Shun, et al (*Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou* 510515)

【**Abstract**】 **Objective** To explore the genetic mechanisms of prostate cancer and identify causal genes for prostate cancer, thereby providing new insights into its molecular mechanisms. **Methods** We implemented transcriptome-wide association analysis using the IVW method and the PMR-Egger model based on Mendelian randomization to identify the genes with causal effects on prostate cancer from the perspectives of gene expression and alternative splicing. Subsequently, we performed gene ontology enrichment analysis and protein-protein interaction network analysis on the common causal genes involved in both of the methods. **Results** The IVW method and PMR-Egger method were used to analyze prostate cancer data, identifying 788 and 495 genes with statistically significant causal effects in expression, and 1054 and 865 genes with statistically significant causal effects in alternative splicing, respectively. Enrichment analysis indicated that some causal genes were significantly enriched in MHC signaling pathways and tumor immune response-related biological processes. Protein-protein interaction network analysis revealed core proteins associated with prostate cancer development, such as MYC. **Conclusion** This study provides potential molecular genetic characteristics and possible therapeutic targets for prostate cancer, offering important references for experimental research, prevention strategies, and clinical diagnosis and treatment of prostate cancer.

【**Key words**】 Prostate cancer; Mendelian randomization; Transcriptome-wide association study; Enrichment analysis; PPI network analysis

前列腺癌是男性泌尿系统最常见的恶性肿瘤,具有较强的扩散能力,扩散后死亡率显著增加。其发病率近年来逐渐上升,已成为男性癌症相关死亡的重要原因之一[1]。研究表明,多个基因参与前列腺癌的生物学过程[2],如*PCA3*基因通过调控雄激素受体信号通路及相关基因的表达,影响癌细胞存活[3]。Knipe等人[4]通过全基因组关联分析(genome-wide association studies,GWAS)发现了多个与前列腺癌相关的单核苷酸多态性(single nucleotide polymorphism,SNP)位点。但其具体生物途径尚不明确。为此,Mancuso等人[5]使用全转录组关联分析(transcriptome-wide association studies,TWAS)整合GWAS与基因表达数量性状位点(expression trait quantitative loci,eQTL)数据,更好地揭示了基因表达对前列腺癌的影响。除基因表达外,不同的可变剪切(alternative splicing,AS)过程可生成多种成熟mRNA,导致不同的宏观表型[6]。基于GWAS和剪切数量性状位点(splicing quantitative trait loci,sQTL)数据的TWAS分析有望识别与疾病关联的AS,揭示新的疾病因果机制。为方便讨论,本文将基于GWAS与eQTL数据的TWAS称为eTWAS,基于GWAS与sQTL数据的TWAS称为sTWAS。然而,目前前列腺癌的TWAS研究主要使用传统模型(如PrediXcan[7]、BSLMM[8]等),这些模型不能排除基因表达与结局之间混杂因素的影响,因此无法确定TWAS的关联性是来源于因果效应还是混杂效应。并且,前列腺癌因果性AS的研究仍较为缺乏。为此,本研究引入基于孟德尔随机化的IVW方法[9]和PMR-Egger方法[10],以更严谨的方式探索前列腺癌的因果基因表达

* 基金项目:国家自然科学基金(82574202);广东省基础与应用基础研究基金(2023A1515011242);南方医科大学2023年国家级大学生创新训练计划项目(202312121036)
△通信作者:周基元,E-mail:zhoujiyuan5460@hotmail.com

及AS。其中，IVW通过加权回归整合多个工具变量（instrumental variables，IVs）效应，提升了统计效力；PMR-Egger基于概率框架，允许IVs之间存在相关性，并通过截距项直接检验和校正水平多效性，从而有效提升了因果推断的稳健性。

资料与方法

1. 数据来源

2023年，Wang等[11]采用meta分析整合了来源于英国生物银行等数据库的93项欧洲人群的前列腺癌GWAS汇总数据，总样本量达726828例（病例组122188，对照组604640），并以二分类变量表征前列腺癌发病状态。Genotype Tissue Expression（GTEx）Project[12]是一项转录组数据研究，包含1641份来自欧美人群的尸检样本，涵盖人体54个部分。本研究采用了GTEx提供的欧洲人群前列腺常染色体eQTL和sQTL汇总数据，数据样本量为181例，且均取自男性样本。

2. 数据预处理

本研究对GWAS及eQTL/sQTL数据实施多阶段质量控制：①GWAS数据删除次等位基因频率小于0.01的SNP，eQTL/sQTL数据额外剔除$P > 0.05$的位点；②通过GTEx查找表对跨数据库中的SNP标识进行匹配；③基于千人基因组计划[13]构建LD参考面板，并按方法差异设置LD剪枝阈值（对于需要独立IVs的IVW阈值设置为0.4，可以包含相关的IVs的PMR-Egger阈值设置为0.8；④将剪枝后的SNP作为IVs。数据预处理采用R 4.3.1完成，LD矩阵计算使用PLINK 1.9 beta软件[14]实现。

3. 分析方法

本研究采用IVW和PMR-Egger开展TWAS。IVW通过LD剪枝筛选SNP，整合其QTL及GWAS效应量与标准误计算因果效应值，并利用估计出来的效应值与标准误的比值构建z检验统计量。PMR-Egger在LD剪枝基础上，采用动态阈值调整策略（逐次折半阈值）优化工具变量筛选，以解决协方差矩阵分解异常的问题。将每种方法分别应用于eTWAS和sTWAS，随后在每种方法下对eTWAS和sTWAS结果取并集，并进一步取IVW和PMR-Egger两种方法结果的交集确定前列腺癌潜在的因果基因。针对此基因集，实施多层次功能解析：①基于PANTHER数据库进行基因本体（gene ontology，GO）富集分析[15]，涵盖生物过程（biological process，BP）、细胞组分（cellular component，CC）及分子功能（molecular function，MF）三类功能层级；②利用STRING在线工具构建蛋白质互作（protein-protein interaction，PPI）网络，揭示基因编码产物的功能协同机制。IVW和PMR-Egger分析在R 4.3.1环境下完成，PMR-Egger建模严格遵循“PMR”包的计算框架。

结　果

1. eTWAS结果

图1A与图1B分别展示了基于IVW与PMR-Egger的eTWAS结果曼哈顿图，横轴标注基因染色体位置，纵轴为因果效应P值的-log10转换值，每个散点代表一个基因。研究共纳入25295个基因表达特征，经Bonferroni校正后显著性阈值为1.98×10^{-6}。IVW识别了788个显著基因，其中*NCOA4*因果效应显著性最高（$P = 1.62\times10^{-188}$）；PMR-Egger识别了495个显著基因，*PCAT1*与*CASC8*显著性最高（$P < 1.00\times10^{-270}$）。上述基因的表达调控异常可能通过因果机制驱动前列腺癌发生。

2. sTWAS结果

图2A与图2B分别展示了基于IVW与PMR-Egger方法的sTWAS结果曼哈顿图，其中每个散点代表一个AS，横纵坐标与图1相同。sTWAS共研究了100934个AS，这些AS可根据物理位置匹配到13483个基因，经Bonferroni校正后显著性阈值为4.95×10^{-7}。IVW方法识别了2293个显著的AS，涉及1054个基因，其中*MRGPRF-AS1*（69015719 bp ~ 69017636 bp）因果效应显著性最高（$P = 5.39\times10^{-167}$）；PMR-Egger方法识别了1836个显著的AS，涉及865个基因，其中8号染色体的*PVT1*（127890998 bp ~ 127939508 bp）和10号染色体的*MSMB*（46033568 bp ~ 46038966 bp）、*TIMM23*（45972730 bp ~ 45975454 bp）、*TIMM23*（45975512 bp ~ 45982523 bp）最显著（$P < 1.00\times10^{-321}$）。由于10号染色体的这三个AS在同一条染色体上且P值接近，所以在图2中这三个AS几乎重合在一个点上。

3. 富集分析和PPI网络分析结果

IVW和PMR-Egger分析结果中有706个公共潜在因果基因，经GO富集分析，显著条目包括BP（126条）、CC（57条）及MF（22条）三类。图3展示了富集倍数（fold enrichment）大于6的GO条目。图3中，纵坐标表示富集条目，横坐标为每个条目经-log10变换后的错误发现率（false discovery rate，FDR）。在所有富集条目中，“MHC蛋白复合物”条目的FDR值最小（$FDR = 3.40\times10^{-11}$），且另有16个MHC相关条目共同暗示MHC信号通路可能是前列腺癌发病的一个重要生物通路。此外，关键条目涵盖肿瘤细胞免疫调控（如对肿瘤细胞的免疫应答，$FDR = 8.80\times10^{-3}$；保护免受自然杀伤细胞介导的细胞毒性，$FDR = 4.56\times10^{-3}$）、凋亡信号（CD95死亡诱导信号复合物，$FDR = 2.86\times10^{-2}$）及代谢（亚精胺跨膜转运，$FDR = 2.36\times10^{-2}$；亚精

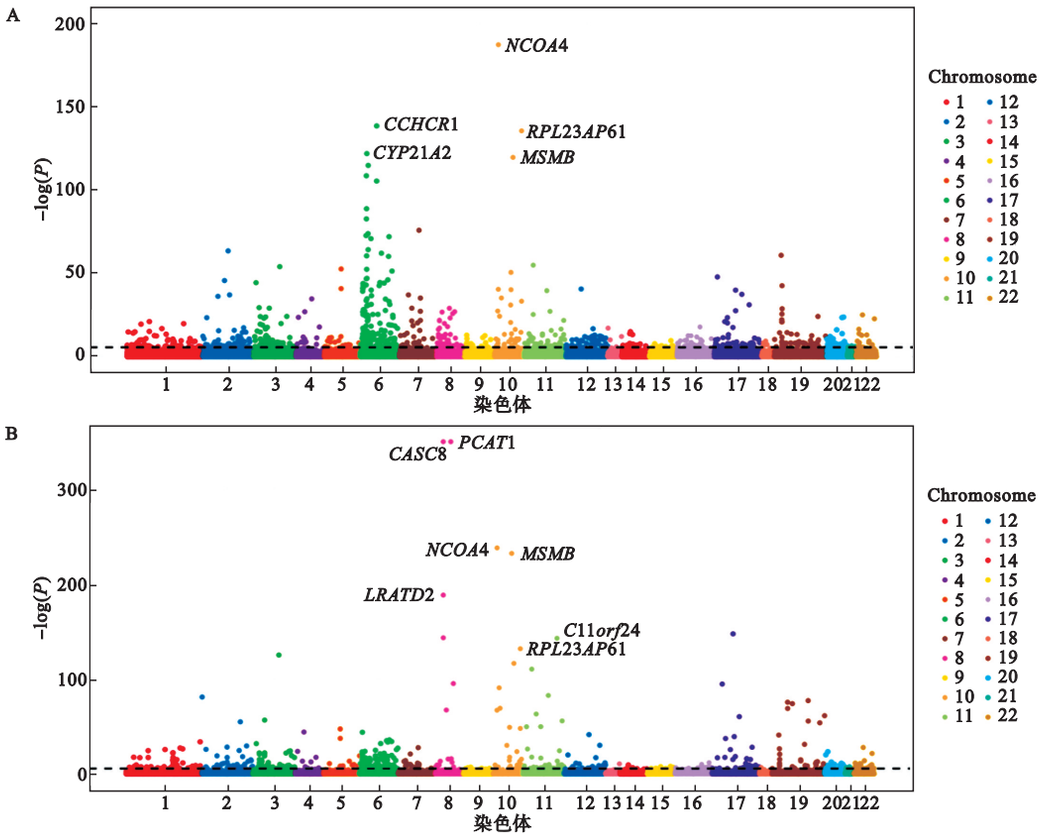


图 1 IVW（A）和 PMR−Egger（B）的 eTWAS 识别结果曼哈顿图

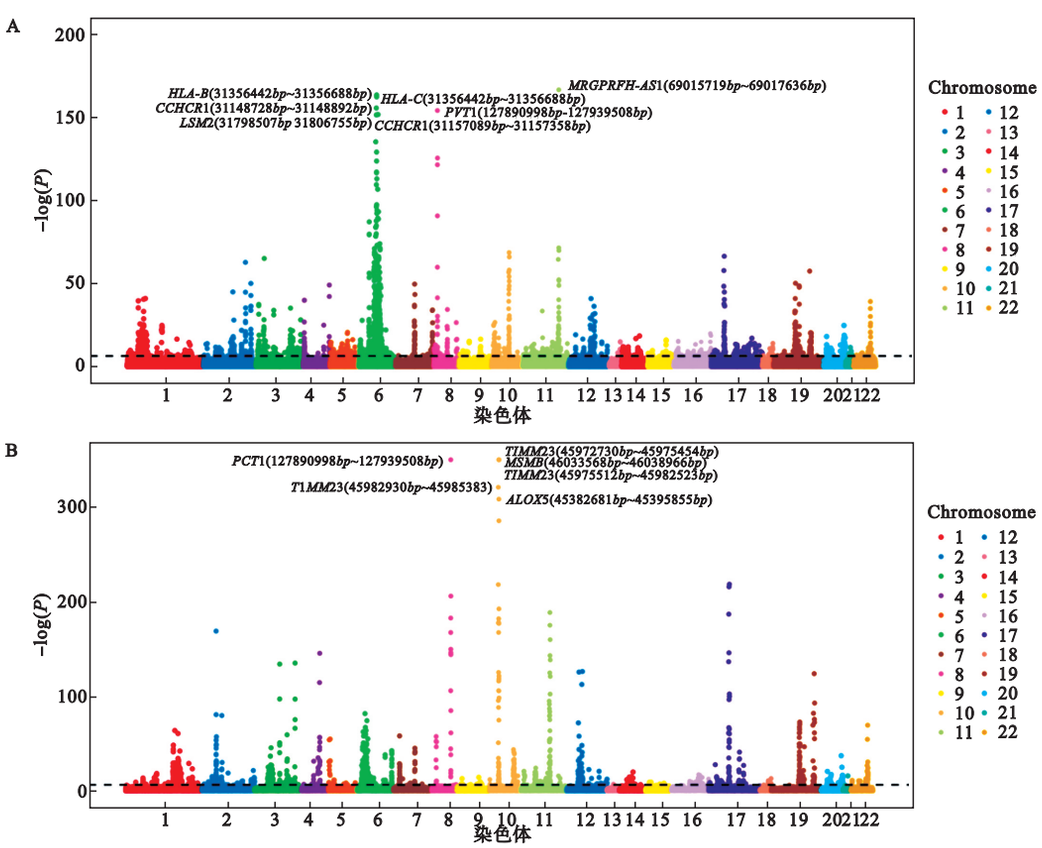


图 2 IVW（A）和 PMR−Egger（B）的 sTWAS 识别结果曼哈顿图

胺跨膜转运蛋白活性，FDR = 3. 61×10^{-2}），提示多维度生物学过程协同驱动前列腺癌的发生发展。

图 4 展示了基于介数中心性（betweenness centrality，BC）构建的前列腺癌潜在因果基因 PPI 网络，节点颜色梯度表征 BC 递增。BC 值为经过该节点的最短路径的数目占网络中最短路径总数目的比例。呈现高 BC 的基因编码蛋白可能通过协调多通路交互在癌变过程中发挥关键调控作用。为突出关键信息，该网

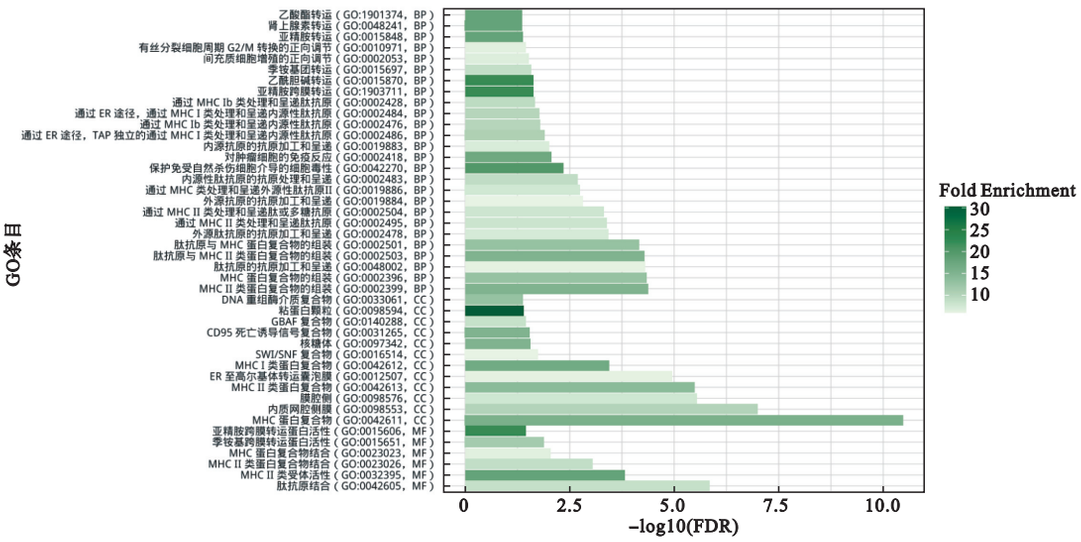


图3　富集倍数大于6的GO富集分析结果

络剔除了合并评分小于0.7的PPI关系和BC小于5000的节点。该PPI网络的外层、中层及核心节点BC分别为5000~8000、8001~20000和>20000。如图4所示，PPI网络最中心的两个蛋白（基因）*MYC*和*RHOA*的BC最高，说明他们在癌变过程中发挥关键调控作用，并且二者之间互作合并评分为0.727，表明二者的转录和翻译可能在前列腺癌的发生机制中发挥相互作用。两种方法识别的显著基因及其交集、GO功能注释和PPI网络分析结果，已全部汇总并作为补充材料上传至GitHub（https://github.com/QiuWenZ/TWAS-of-prostate-cancer）。

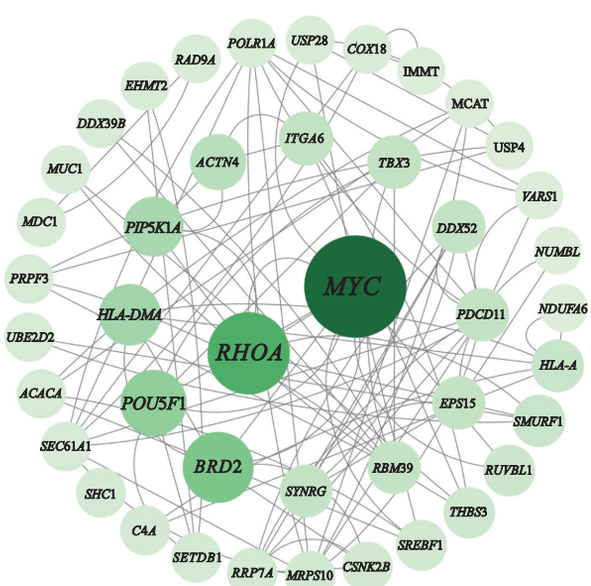


图4　部分节点的PPI网络示意图

讨　论

我们将前列腺癌GWAS数据分别与eQTL和sQTL数据联合分析，采用IVW和PMR-Egger分别识别了788个与495个基因的表达以及1054个与865个基因的可变剪切可能对前列腺癌存在因果效应。我们在每种方法下分别对eTWAS和sTWAS结果取并集，并进一步取IVW和PMR-Egger两种方法结果的交集，此交集的基因包括*NCOA4*、*MSMB*等。我们认为这些基因均是前列腺癌的潜在因果基因。对这些基因进行GO富集分析，发现许多MHC类蛋白的信号通路及与细胞免疫相关的条目显著富集。基于BC构建的PPI网络分析表明，MYC、RHOA等核心蛋白可能通过协调多通路交互在前列腺癌的癌变过程中发挥关键调控作用。

NCOA4、*MSMB*均被本研究的IVW和PMR-Egger方法共同识别为潜在因果基因，这与Smith等[16]的癌症风险位点分析的发现相一致，后者表明抑制*MSMB*表达或*NCOA4*过表达可促进前列腺上皮细胞的锚定非依赖性生长。*PCAT1*基因是PMR-Egger的eTWAS结果中最显著的基因，MYC蛋白是PPI网络中核心的蛋白质。这说明*PCAT1*和*MYC*均可能参与影响前列腺癌的复杂因果机制。事实上，目前已有研究发现*PCAT1*基因能够通过上调cMYC蛋白（由*MYC*基因编码）促进前列腺细胞的增殖。c*MYC*也能够通过拮抗抑制*PCAT1*介导的癌细胞增殖[17]。此外，Bander等[18]的实验研究表明，MHC1常出现在癌细胞表面，协助效应T细胞识别抗原，而MHC2则用于辅助T细胞识别多肽抗原。在本研究中，MHC蛋白在富集分析中的显著性结果同样表明，其可能在前列腺局部组织中由特定基因表达诱导，且与细胞凋亡密切相关。此外，与以往基于转录组的TWAS研究相比，Mancuso等人[5]的研究同样识别了*NCOA4*、*EHBP1*和*USP39*等基因。这些发现验证了本研究的可靠性，并提示未被证实的因果基因和蛋白可能为潜在的前列腺癌标志物。

本研究采用了基于孟德尔随机化的IVW方法和

PMR-Egger方法估计了暴露对前列腺癌的因果效应。同时，将AS引入TWAS框架，识别了潜在的因果性AS及其相关基因。通过富集分析和PPI网络分析，初步揭示了因果基因的潜在作用机制、生物通路及相关蛋白的互作关系。此外，作为一项探索性研究，本研究为进一步揭示前列腺癌致病机制提供了新的方向。在前列腺癌筛查中具有一级和二级预防的潜在价值。这些潜在因果基因有望成为新的诊断标志物或治疗靶点。然而，本研究仅聚焦于转录组学数据的基因表达过程和可变剪切过程，未涵盖所有可能的因果机制。未来，随着数据的不断积累，研究人员可以结合蛋白组、表观遗传组及代谢组的QTL数据更全面地揭示前列腺癌的因果机制。本研究基于生物信息学分析方法探讨了前列腺癌GWAS数据集与QTL数据的关联特征，相关发现需通过分子生物学实验以及临床研究进行功能验证，从而系统阐明候选基因在前列腺癌发生和发展过程中的分子调控机制。跨种族人群前列腺癌的流行病学和基因组学存在异质性[19]。因此，将本文研究结果的欧裔人群直接推广到其他种族群体，尤其是中国人群，可能存在一定的局限性。考虑到中国在前列腺癌防治方面的公共卫生挑战及疾病负担，未来需要开展跨种族的多中心临床试验，以便更全面地了解前列腺癌在不同种族间的流行病学特征及其潜在的生物学机制。此外，系统性地构建多种族队列，深入研究种族特异性的疾病驱动机制，有助于优化临床诊疗策略，促进精准医学的发展，并推动全球健康的实现。

参 考 文 献

[1] Siegel RL, Miller KD, Wagle NS, et al. Cancer Statistics, 2023[J]. CA: A Cancer Journal for Clinicians, 2023, 73(1): 17-48.

[2] Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer: analyses of cohorts of twins from Sweden, Denmark, and Finland[J]. New England Journal of Medicine, 2000, 343(2): 78-85.

[3] Lemos AEG, Matos ADR, Ferreira LB, et al. The long non-coding RNA PCA3: an update of its functions and clinical applications as a biomarker in prostate cancer[J]. Oncotarget, 2019, 10(61): 6589.

[4] Knipe DW, Evans DM, Kemp JP, et al. Genetic variation in prostate-specific antigen-detected prostate cancer and the effect of control selection on genetic association studies[J]. Cancer Epidemiology, Biomarkers & Prevention, 2014, 23(7): 1356-1365.

[5] Mancuso N, Gayther S, Gusev A, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions[J]. Nature Communications, 2018, 9(1): 4079.

[6] Kelemen O, Convertini P, Zhang Z, et al. Function of alternative splicing[J]. Gene, 2013, 514(1): 1-30.

[7] Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data[J]. Nature Genetics, 2015, 47(9): 1091-1098.

[8] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models[J]. PLoS Genetics, 2013, 9(2): e1003264.

[9] Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data[J]. Genetic Epidemiology, 2013, 37(7): 658-665.

[10] Yuan Z, Zhu H, Zeng P, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies[J]. Nature Communications, 2020, 11(1): 3861.

[11] Wang A, Shen J, Rodriguez AA, et al. Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187 novel risk variants[J]. Nature Genetics, 2023, 55(12): 2065-2074.

[12] Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project[J]. Nature Genetics, 2013, 45(6): 580-585.

[13] 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation[J]. Nature, 2015, 526(7571): 68.

[14] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.

[15] Gene OC. The Gene Ontology (GO) database and informatics resource[J]. Nucleic Acids Research, 2004, 32(suppl_1): D258-D261.

[16] Pomerantz MM, Shrestha Y, Flavin RJ, et al. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis[J]. PLoS Genetics, 2010, 6(11): e1001204.

[17] Prensner JR, Chen W, Han S, et al. The long non-coding RNA PCAT-1 promotes prostate cancer cell proliferation through cMyc[J]. Neoplasia, 2014, 16(11): 900-908.

[18] Bander NH, Yao D, Liu H, et al. MHC class I and II expression in prostate carcinoma and modulation by interferon-alpha and gamma[J]. The Prostate, 1997, 33(4): 233-239.

[19] Zhu Y, Mo M, Wei Y, et al. Epidemiology and genomics of prostate cancer in Asian men[J]. Nature Reviews Urology, 2021, 18(5): 282-301.

（责任编辑：邓 妍）