创新团队成果专栏

癌睾丸基因的基因表达程序分析

侯宗良，杨　琴，李少白，雷　鸣
上海交通大学医学院附属第九人民医院上海精准医学研究院，上海 200125

［摘要］**目的**·基于睾丸单细胞转录组数据，鉴定精子发生过程中癌睾丸基因（cancer-testis gene，CTG）的基因表达程序（gene expression program，GEP），并探究其与肿瘤患者预后的关系。**方法**·从GTEx数据库和TCGA数据库获取正常组织和肿瘤组织的表达谱，筛选CTG。基于睾丸单细胞转录组，使用leiden聚类算法鉴定出CTG在精子发生过程中的GEP。使用DecoupleR评估GEP的活跃程度，以确定每个GEP活跃的细胞类型和精子发生时期。利用DecoupleR评估GEP在肿瘤组织中的活跃程度，并分析GEP与肿瘤患者生存的相关性。**结果**·基于GTEx和TCGA数据库中正常组织和肿瘤组织的基因表达谱，筛选到917个CTG。利用CTG在睾丸单细胞转录组中的表达情况，通过聚类算法鉴定出7个GEP。GEP活性分析结果表明，GEP5活跃于精子发生前期，包括精原干细胞、分化中的精原细胞和早期初级精母细胞等细胞类型。统计其在染色体上的分布发现，GEP5包含的基因主要分布于X染色体上。生存分析结果表明GEP5在多种肿瘤类型中的活跃程度与患者的生存情况呈负相关。**结论**·在精子发生过程中，GEP5活跃于精子发生过程的前期，其包含的基因主要分布于X染色体上。在多种肿瘤类型中，GEP5的活跃程度与患者的预后密切相关。

［关键词］癌睾丸基因；基因表达程序；单细胞基因表达分析
［DOI］10.3969/j.issn.1674-8115.2023.08.001　［中图分类号］R730.7；R318.04　［文献标志码］A

Gene expression program analysis of cancer-testis genes

HOU Zongliang, YANG Qin, LI Shaobai, LEI Ming
Shanghai Institute of Precision Medicine, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200125, China

[Abstract] **Objective**·To identify the gene expression program (GEP) of cancer-testis genes (CTGs) during spermatogenesis based on single-cell transcriptome data from the testis and investigate their association with the prognosis of cancer patients. **Methods**·Expression profiles of normal and tumor tissues were obtained from the GTEx and TCGA databases to screen CTGs. The GEP of CTGs during spermatogenesis was identified by applying the leiden clustering algorithm to testicular single-cell transcriptome data. DecoupleR was used to evaluate the activity levels of GEP and determine the cell types and stages of spermatogenesis where each GEP was active. Subsequently, DecoupleR was used to evaluate the activity levels of GEP in tumor tissues and analyze the correlation between GEP and cancer patient survival. **Results**·Based on the expression profiles of normal and tumor tissues from the GTEx and TCGA databases, 917 CTGs were identified. By using the expression patterns of CTGs in the testicular single-cell transcriptome data, seven GEPs were identified through the clustering algorithm. Activity level analysis revealed that GEP5 was active in the early stages of spermatogenesis, including spermatogonia stem cells, differentiating spermatogonia, and early primary spermatocytes. The distribution of GEP5-associated genes was predominantly found on the X chromosome. Additionally, survival analysis demonstrated a statistically significant negative correlation between GEP5 activity levels and patient survival in various tumors. **Conclusion**·During spermatogenesis, GEP5 is active in early stages, and its associated genes are primarily located on the X chromosome. In multiple tumor types, the activity level of GEP5 is closely related to patient prognosis.

[Key words] cancer-testis genes; gene expression program; single-cell gene expression analysis

癌睾丸抗原（cancer-testis antigen，CTA）是一类特异性表达于睾丸和肿瘤中的抗原，已应用于肿瘤治疗[1-2]。由于CTA具有局限表达于睾丸且在肿瘤中高表达特征，有研究者采用生物信息分析等方法鉴定

［基金项目］国家重点研发计划（2018YFA0107004）。
［作者简介］侯宗良（1998—），男，硕士生；电子信箱：EnderZ@sjtu.edu.cn。
［通信作者］雷　鸣，电子信箱：leim@shsmu.edu.cn。
[Funding Information] National Key Research and Development Program of China (2018YFA0107004).
[Corresponding Author] LEI Ming, E-mail: leim@shsmu.edu.cn.

癌睾丸基因（cancer-testis gene，CTG）[3-4]。这些技术不能确定基因表达的蛋白质是否具有免疫原性，故这些基因连同之前的CTA合称为CTG。在2009年，研究者分析了已鉴定的CTG，将其信息整理成数据库CTdatabase[3]。CTdatabase数据库中所有内容都经过了相关领域内专家的注释，包括基因名称、基因组位置等。此外，研究者根据CTG是否分布在X染色体上，将CTG分成2类：分布在X染色体上的CTG（CT-X）和分布在常染色体上的CTG（Non CT-X）[5-6]。

肿瘤相关基因特异性表达的研究，使人们注意到精子和肿瘤这2种生理与病理上截然不同的组织，其发生过程可能存在相似之处[5,7]。研究者认为，肿瘤中CTG的异常表达，可能反映了体细胞中本该沉默的代表了精子发生过程的基因表达程序（gene expression program，GEP）被再次激活，并且这些GEP可能参与肿瘤的发生和发展[7]。然而，CTG在精子发生和肿瘤发展过程中的GEP特征，及其与肿瘤患者预后的关系等问题，尚未得到解决。

本研究采用生物信息学技术，研究CTG的分布特征和表达规律。通过多个数据库的联合分析，重新对CTG进行筛选；分析睾丸精子发生过程中的基因表达谱，鉴定出与CTG相关的GEP，并分析其在各种肿瘤类型中的活跃程度及其与患者预后的关联性。

1 材料与方法

1.1 CTG的筛选

1.1.1 GTEx转录组数据的获取 GTEx（Genotpye-Tissue Expression）项目致力于提供给研究者用于研究遗传变异与基因表达调控的资源。本研究中使用的是移除个体水平身份信息的基因表达数据，即正常组织的转录组数据（https://www.gtexportal.org/home/datasets）。其发布的时间是2019年8月20日，版本号是V8。共有54种组织类型，包含睾丸组织。

1.1.2 HBM转录组数据的获取 Illumina Human Body Map（HBM）项目包含16种组织类型的转录组数据。为了与GTEx数据库的转录组数据相匹配，从HBM下载原始的测序短读测序片段（reads）数据（FASTQ格式存储），并用GTEx数据库的分析流程对HBM的数据进行比对和基因定量。HBM的FASTQ原始数据存储于EMBL-EBI数据库，访问号为E-MTAB-315。

1.1.3 TCGA转录组数据的获取 通过TCGAbiolink[8]检索TCGA数据库中33种癌症类型样本的转录组，并将用于gdc-client下载的索引保存至gdc_manifest.txt文件中。之后，用gdc-client下载33种癌症类型样本的转录组。本研究中使用的数据版本为v33.1，发布时间为2022年5月31日。

1.1.4 基因表达特异性分析 使用基因特异性度量（specific measure，SPM）衡量基因在睾丸组织中的特异表达情况[4]。具体定义如下：基因在不同组织中的表达矩阵表示为$X \in \mathbb{R}^{m \times n}$，共有$m$个基因，$n$个不同的组织。基因$i$在不同组织中的表达量可以表示为向量$x_i \in \mathbb{R}^n$，如式1：

$$x_i = \left(x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{in}\right) \quad (1)。$$

其中，i表示第i个基因，x_{ij}表示基因i在j组织中的表达量，共有n个不同的组织类型。

基因特异性表达是指基因仅在某个组织中表达，而不在其他组织中表达。因此当基因i特异地在组织t中表达，那么基因i在不同组织中表达量的向量可以表示为式2：

$$x_i^{(t)} = \left(0, 0, \cdots, x_{it}, \cdots, 0\right) \quad (2)。$$

根据式1和式2，可以定义基因i在t组织中的特异表达程度，即基因i在不同组织中的真实表达情况x_i和基因i特异表达于组织t时在不同组织中的理想表达情况$x_i^{(t)}$之间的cos相似度，$\mathrm{SPM}_i^{(t)}$。其定义如式3所示：

$$\begin{aligned}\mathrm{SPM}_i^{(t)} &= \frac{x_i x_i^{(t)}}{\left|x_i\right| \cdot \left|x_i^{(t)}\right|} \\ &= \frac{\left|x_i^{(t)}\right|}{\left|x_i\right|} \\ &= \frac{x_{it}}{\sqrt{\sum_{j=0}^{m} x_{ij}^2}}\end{aligned} \quad (3)。$$

1.1.5 筛选CTG CTG的主要特征是在正常的睾丸组织中高表达，并且在肿瘤中也高表达。因此，可以分成两步筛选：先筛选睾丸特异表达基因（testis specific gene，TSG），再基于TSGs在肿瘤中的表达情况筛选出CTG。

使用正常转录组数据（GTEx数据库和HBM数据库）和肿瘤转录组数据（TCGA数据库）筛选CTG。首先，筛选TSG，具体步骤如下：①计算GTEx数据库中每个基因在不同组织类型中的平均表达量。②计算GTEx数据库中每个基因在睾丸组织中的特异性，即$\mathrm{SPM}_{m \times 1}^{(\mathrm{Testis},\ \mathrm{GTEx})}$。③计算HBM数据库中

每个基因在不同组织类型中的平均表达量。④计算HBM数据库中每个基因在睾丸组织中的特异性，即$SPM_{m\times 1}^{(Testis,\ HBM)}$。⑤筛选出满足$SPM_{m\times 1}^{(Testis,\ GTEx)}>0.9$和$SPM_{m\times 1}^{(Testis,\ HBM)}>0.9$，并且在睾丸组织中的平均表达量高于1 TPM的蛋白编码基因。其次，CTG还需要TSG满足在肿瘤中高表达（至少在1种肿瘤的1%样本中表达，表达的阈值为5 TPM）。

1.2　GEP的鉴定

精子发生过程中的GEP是一组表达模式相似的基因构成的基因集[9]。因此，为了鉴定精子发生过程中CTG所包含的GEP，对CTG在睾丸单细胞的表达谱进行聚类分析。聚类结果可以反映CTG在精子发生过程中的表达模式。

1.2.1　睾丸单细胞转录组原始数据的获取　睾丸单细胞转录组数据来自2018—2019年发表的3篇睾丸单细胞转录组研究文献，其数据在GEO数据库中的标识号分别为GSE109037[10]、GSE112013[9]和GSE124263[11]。这3套睾丸的单细胞转录组均是10X平台产生的数据。由于选用的参考基因组的版本不同，导致3套数据在基因名称等方面不兼容，不利于后续分析。因此，从原始的测序reads出发，重新处理这3套数据的原始数据，比对到统一的参考基因组上，有利于整合这3组数据。

1.2.2　睾丸单细胞转录组测序分析　将下载的原始测序reads转换成FASTQ格式，并将文件名命名为cellranger可以接受的格式，最后使用cellranger对基因的表达水平进行定量。下游分析使用Python语言的scanpy库进行[12]。定量后的数据，经过质控、过滤、合并，使用scvi-tools进行去批次效应和过滤双细胞[13-14]。随后，使用leiden图聚类算法对细胞进行聚类[15]。leiden算法中影响聚类的参数主要是resolution。为了确定最优的聚类参数，在一系列resolution下使用leiden对细胞进行聚类，并用轮廓系数（silhouette score）评估聚类的效果。Silhouette score是一个经典的用于评估聚类结果的算法，数值越大说明聚类的结果越好。Silhouette score是每个样本单独计算的，由2个部分组成：a表示样本与同一类中其他点之间的平均距离，b为样本与次近类中其他点之间的平均距离。单个样本的silhouette score（s），可以表示为式4：

$$s=\frac{b-a}{\max(a,b)} \quad (4)。$$

s的值域为$s\in[-1,\ +1]$，−1表示该样本分类错误，+1表示该样本分类良好，s位于0附近表示该样本的分类不明确。为了衡量数据集（K个样本）的整体聚类情况，使用数据集中每个样本的silhouette score的平均值，记为S，表示为式5：

$$S=\frac{\sum_{i=0}^{K}s_i}{K} \quad (5)。$$

S值越大，表示数据集中整体样本的分类情况越好。选择一系列resolution参数，根据每个参数的聚类结果计算出S。最终选择S最大时对应的resolution作为聚类的参数，并以此参数确定细胞类群。随后，通过查阅文献获取之前研究[9-11,16-17]中睾丸组织不同细胞类型的标志基因（表1），并根据这些基因在不同细胞类群中的表达情况对所有细胞类群进行注释。

表1　睾丸细胞类型、缩写以及其对应的标志基因和细胞数

Tab 1　Testicular cell types, abbreviations, and their corresponding marker genes and cell numbers

Cell type	Abbreviation	Marker gene	Cell number/*n*
Spermatogonia stem cell	SSC	*UTF1, DMRT1*	5 822
Differentiating spermatogonia	Differentiating SPG	*DMRT1*	2 006
Early primary spermatocyte	Early primary SPC	*DMC1*	2 085
Late primary spermatocyte	Late primary SPC	*ZPBP, SPAG6*	4 484
Round spermatid	Round ST	*ZPBP, SPAG6, ACR*	5 653
Elongating spermatid	Elongating ST	*TNP1*	3 541
Elongated spermatid	Elongated ST	*TNP2*	2 511
Sertoli cell	SC	*SOX9*	1 368
Peritubular cell	PC	*MYH11*	3 740
Leydig cell	LC	*DLK1*	4 485
Smooth muscle cell	SMC	*MYH11, NOTCH3*	672
Epithelial cell	EC	*VWF*	1 792
Testis macrophage	TM	*CD14*	1 442

1.2.3 GEP鉴定过程 为了鉴定精子发生过程中CTG包含的GEP，首先查看所有CTG在睾丸单细胞转录组中的表达情况。由于CTG主要表达于生殖系细胞（见结果部分），随后选择生殖系细胞的表达谱用于鉴定CTG包含的GEP，具体步骤如下：①提取CTG在生殖系细胞中的表达矩阵。②使用leiden图聚类算法对CTG进行聚类。③类似于细胞聚类，选定一系列resolution，并用silhouette score评估聚类结果，选择最终的resolution。④使用最优的聚类参数对CTG进行聚类，得到表达模式相似的CTG组成的基因集。⑤使用热图展示CTG的表达情况，观察不同基因集表达模式的异同。

这些基因集包含不同的基因表达模式，每个基因集表示一个GEP，并且这些GEP之间互不重叠。每个GEP表示为G，其包含的基因数为N。在GEP内部，以每个基因与该基因集中心的斯皮尔曼相关系数表示该基因的表达模式与其所属GEP的相似程度，表示为ρ_g，则ρ_g可以表示为式6：

$$\rho_g = \text{spearmanr}\left(x_g, x_{\bar{g}}\right), g \in G \quad (6)。$$

式6中，spearmanr表示斯皮尔曼相关系数的计算函数，由SciPy库提供。x_g表示GEP中的一个基因在所有细胞中的表达量组成的向量。$x_{\bar{g}}$表示基因集的中心，定义为式7：

$$x_{\bar{g}} = \frac{1}{N}\sum_{g}^{g \in G} x_g \quad (7)。$$

1.3 GEP的活跃程度及其与患者预后的关系

1.3.1 GEP活性分析 为了定量表示GEP在每个细胞（scRNA-seq）或者样本（Bulk RNA-seq）中的活跃程度，使用DecoupleR包中的wmean函数计算每个GEP的活跃程度[18]，表示为a_G。通过wmean先计算出GEP内基因表达量的权重平均值，再根据权重平均值的零分布标准化前面计算出来的权重平均值。标准化后的权重平均值定义为GEP的活跃程度。其中权重为前文中定义的基因表达模式与其所属GEP的相似程度，即ρ_g。GEP内基因表达量的权重平均值定义为式8，y_g表示一个基因在一个细胞/样本中的表达量：

$$\text{wm}_G = \frac{1}{N}\sum_{g}^{g \in G} y_g \rho_g \quad (8)。$$

对一个细胞中所有基因，按照每个GEP对应的基因数重复抽样1 000次，并计算其权重平均值，得到该GEP权重平均值的零分布。零分布的均值和方差表示为μ_G和σ_G^2。由此，GEP在一个细胞/样本中的活跃程度可以表示为式9：

$$a_G = \frac{\text{wm}_G - \mu_G}{\sigma_G} \quad (9)。$$

1.3.2 生存分析 生存分析用于探索GEP的活跃程度与患者预后的关系，使用R语言中的survival包实现。首先，通过DecoupleR包的wmean函数计算TCGA中每例患者GEP的活跃程度；再将患者分为GEP活跃程度高组和GEP活跃程度低组，比较2组生存曲线；同时，基于患者的GEP活跃程度与患者的生存情况，进行Cox回归分析。

2 结果

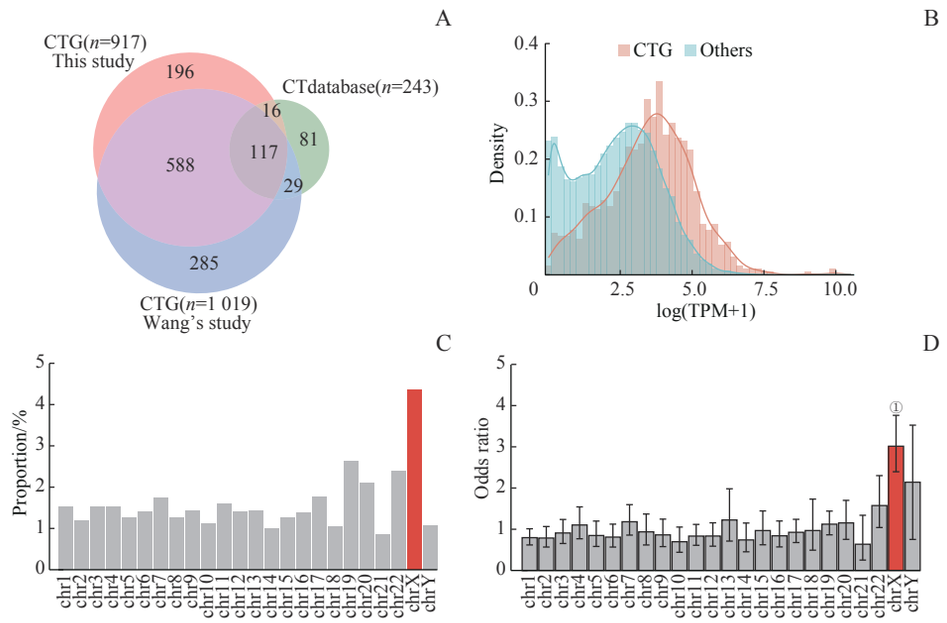
2.1 CTG的筛选

共筛选出1 271个TSG，进一步筛选出917个CTG。其中，鉴定到的CTG与WANG等[4]鉴定到的CTG有705个重合，其中收录于CTdatabase数据库[3]的有117个。新鉴定到212个CTG，其中收录于CTdatabase数据库的有16个（图1A）。

图1B比较了CTG和非CTG在睾丸组织中的表达情况，结果表明CTG整体上的表达水平高于非CTG的表达水平。图1C、D统计了CTG在染色体上的分布情况，其中CTG在X染色体上显著富集。图1C表明CTG在X染色体上的占比远大于其他染色体。通过Fisher's检验计算CTG在每条染色体上的富集情况（图1D），结果显示CTG在X染色体上的比值比（odds ratio，OR）高于其他染色体，且比值比显著大于1（P=0.000），说明CTG显著富集于X染色体上。

2.2 GEP的鉴定

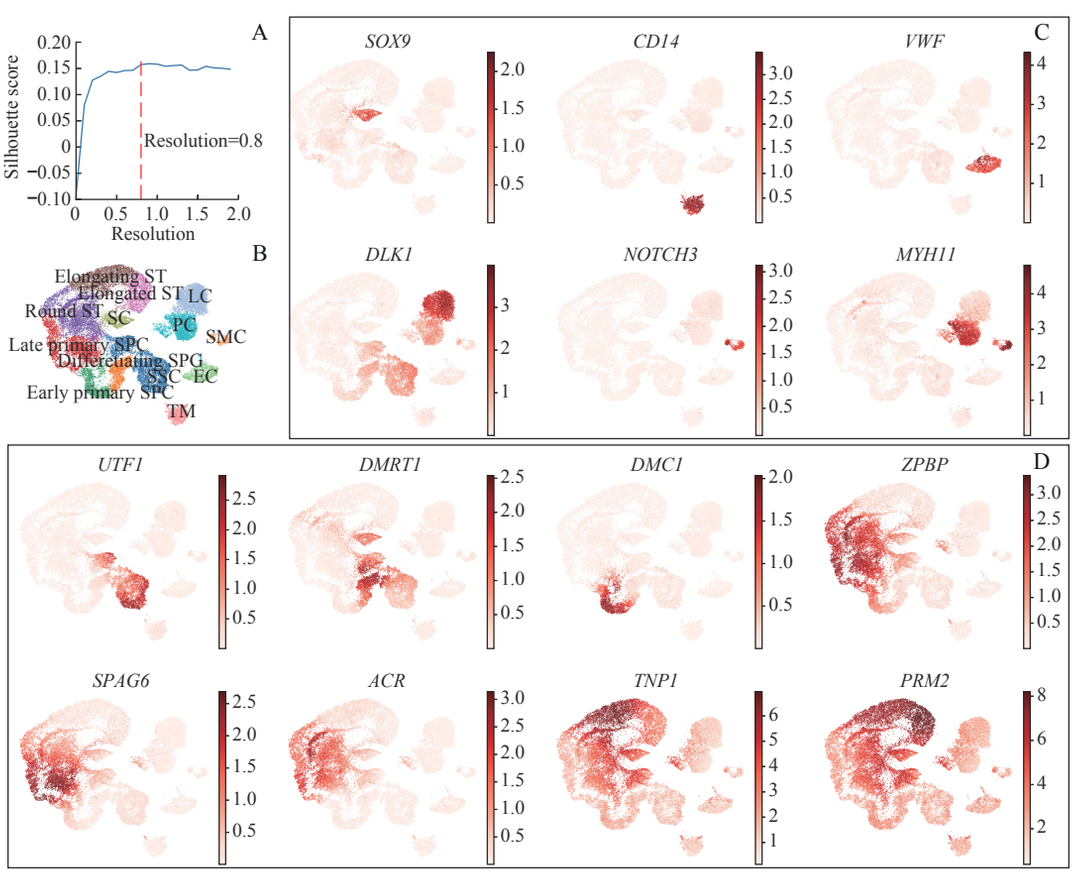
为了得到睾丸单细胞转录图谱，统一定量3套数据的原始测序数据，并进行质控和过滤。采用scvi-tools的整合算法将3组数据整合，共得到39 601个细胞。对整合后的睾丸组织单细胞转录组进行细胞聚类分析（图2）。结果显示，在一系列resolution参数中，当silhouette score取最大值时，对应的resolution为0.8（图2A）。基于此resolution值将细胞聚类为22个类。图2C、D分别展示了睾丸组织中体细胞和生殖系细胞的标志基因表达情况。基于此标注每个聚类类别的细胞类型，最终得到睾丸组织的单细胞转录组图谱（图2B）。共有13种细胞类型，每种细胞类型的全称、数量和标志基因详见表1。



Note：A. Comparison between CTGs identified in this study and previous studies. B. Distribution of CTGs and non-CTGs in testis. C. The proportion of CTGs identified in different chromosomes. D. Odds ratio of CTGs on chromosomes. ①*P*=0.000, compared with other chromosomes.

图1　CTG的筛选

Fig 1　Screening of CTGs



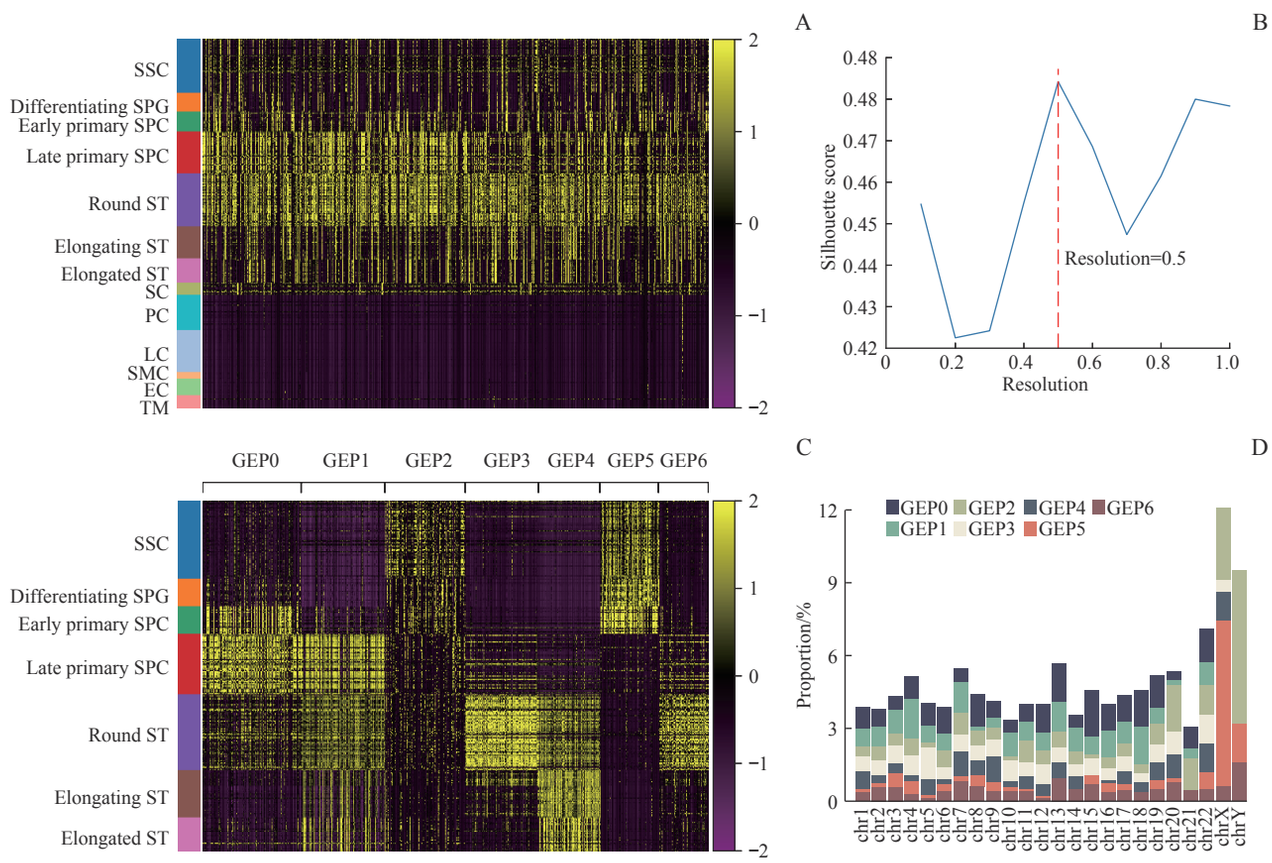
Note：A. Relationship between resolution parameters in cell clustering and the silhouette score. The red dotted line in the panel indicates the optimal resolution parameter in clustering. B. Single-cell transcriptome map of the testis. C. Expression of marker genes of somatic cell types (*SOX9, CD14, VWF, DLK1, NOTCH3* and *MYH11*) in testis. D. Expression of marker genes of germline cell types (*UTF1, DMRT1, DMC1, ZPBP, SPAG6, ACR, TNP1* and *PRM2*) in testis. The intensity of color in the C/D panels represents the level of gene expression after normalization.

图2　睾丸单细胞转录组聚类分析和细胞类型标注

Fig 2　Cluster analysis of testicular single-cell transcriptome and cell types labeling

CTG主要在生殖系细胞中表达，而在睾丸体细胞中不表达或低表达（图3A）。因此，CTG中包含的GEP主要是精子发生过程中的GEP。根据GEP鉴定的方法，在不同的resolution参数下，计算CTG聚类后的silhouette score。结果表明，在resolution = 0.5时，silhouette score达到最大值（图3B）。因此，以此参数对CTG进行聚类，将其分成7类，即7个GEP。每个GEP偏好表达于特定的细胞类型和精子发生时期（图3C）。

从染色体分布来看，GEP5包含的CTG绝大部分位于X染色体上（图3D）。这表明文献中提到的CT-X基因，在GEP的层面上主要集中在GEP5中。此外，图3C表明GEP5包含的基因在精子发生过程的早期阶段表达水平较高。



Note: A. Expression of CTGs in the entire testicular single-cell transcriptome. B. The relationship between resolution and silhouette score in CTGs clustering. The red dotted line in the panel indicates the optimal position of resolution. C. Expression of CTGs in testicular germ line cells after clustering. D. The proportion of CTGs contained in each GEP on the chromosomes. The values in the heat map of A/B panels were converted to *z*-scores by the gene. The closer the color is to yellow, the higher the gene expression is; the closer the color is to purple, the lower the gene expression is.

图3 鉴定CTG包含的精子发生过程中的GEP

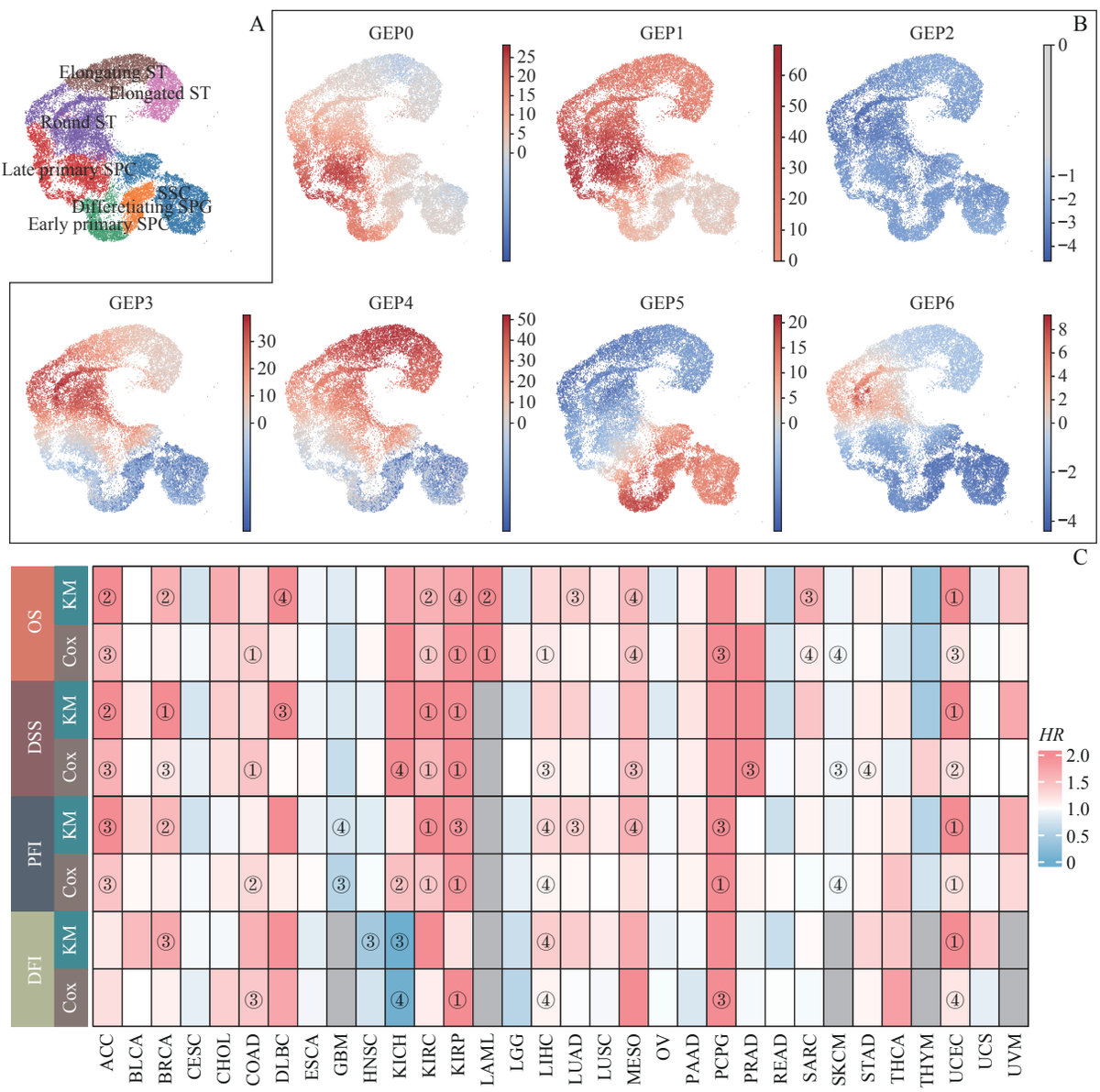
Fig 3 Identification of GEP contained in CTGs during spermatogenesis

2.3 GEP的活跃程度及其与肿瘤患者生存的关系

对7个GEP在每个细胞中的活跃程度进行定量计算，并用UMAP降维分析展示每个GEP在每个细胞中的活跃程度（图4A、B）。结果显示：GEP0主要活跃于Early primary SPC、Late primary SPC、Round ST中；GEP1几乎在每个细胞类型中都是活跃的，但在SSC、Differentiating SPG、Early primary SPC中的活跃程度较低；GEP2几乎在每个细胞类型中都是不活跃的；GEP3和GEP4主要活跃于Round ST、Elongating ST和Elongated ST中，GEP3在Round ST中最活跃，GEP4在Elongating ST中最活跃；GEP5主要活跃于SSC、Differentiating SPG、Early primary SPC；GEP6仅在Round ST中活跃。细胞类型缩写及含义见表1。

生存分析的结果汇总如图4C所示，整体来看，GEP5的活跃程度与患者的生存期有一定的关系。在

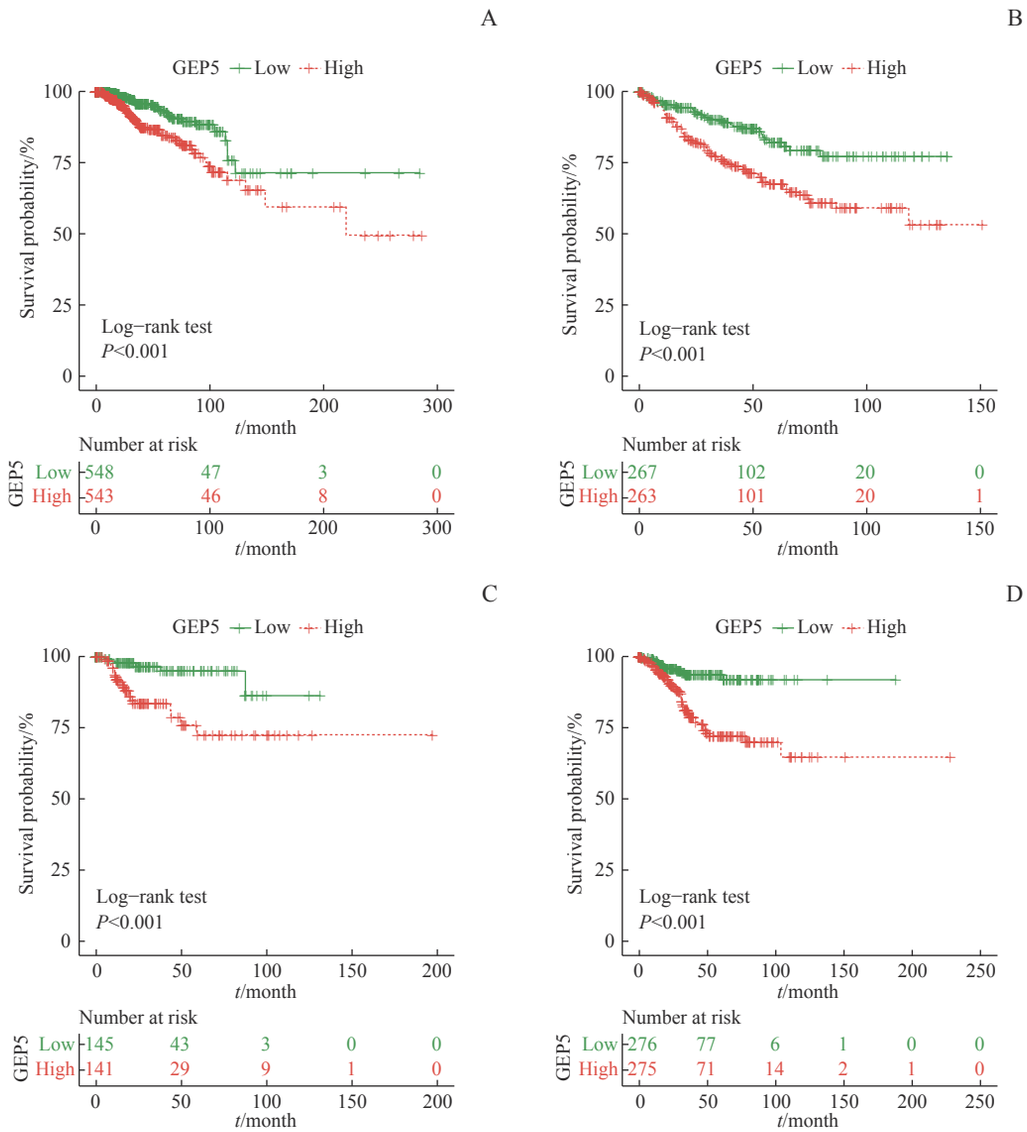
大部分的肿瘤类型中，GEP5活跃程度高与患者的生存差有显著的关联，其中ACC（adrenocortical carcinoma）、BRCA（breast invasive carcinoma）、COAD（colon adenocarcinoma）、KIRC（kidney renal clear cell carcinoma）、KIRP（kidney renal papillary cell carcinoma）和UCEC（uterine corpus endometrial carcinoma）最显著（至少有4个风险比显著大于1）。BRCA、KIRC、KIRP、UCEC的疾病特异性生存（disease specific survival，DSS）曲线如图5所示，GEP5活跃程度高组的疾病特异性生存期与GEP5活跃程度低组比较，显著缩短。以上这些数据说明GEP5在多种肿瘤类型中的活跃程度与患者的生存期呈负相关。



Note: A. Single-cell map of testis germ line cells. B. The activity of GEP in testis germ line cells. The colors in the panel represent the degree of activity of each GEP. The closer the GEP is to red, the more active it is, and the closer the GEP is to blue, the more inhibited it is. C. Relationship between GEP5 activity and prognosis in different cancer types. OS—overall survival; PFI—progression free interval; DFI—disease free interval. The colors represent the hazard ratio (*HR*) of Kaplan-Meier (KM) survival analysis and Cox regression analysis. White means *HR*=1; pink means *HR*=2; light blue means *HR*=0; gray indicates missing data. ①*P*<0.001; ②*P*<0.01; ③*P*<0.05; ④*P*<0.10.

图4　GEP的活跃情况以及GEP5与肿瘤患者生存的关系

Fig 4　Activity of GEP and the relationship between GEP5 and the survival of cancer patients



Note: A. DSS curve of BRCA.B. DSS curve of KIRC. C. DSS curve of KIRP. D. DSS curve of UCEC. The upper panel shows Kaplan-Meier survival curves; the lower panel presents the number of people at risk (disease) at different time points.

图5　GEP5在不同癌症类型中的活跃程度与DSS曲线

Fig 5　GEP5 activity in different cancer types and DSS curves

3　讨论

CTG具有局限表达于睾丸且在肿瘤中高表达的特征，可作为肿瘤的治疗靶点或诊断标志物。目前，很多这方面的研究聚焦于单基因分析，缺乏对基因间协同作用的研究。本研究利用生物信息学计算分析技术，探索CTG的协同表达规律。

本研究通过基因表达特异性分析，筛选出1 271个TSG，结合TCGA数据库进一步确认917个CTG。在睾丸组织中，CTG的表达水平整体上高于其他基因的表达水平。收录在CTdatabase而没有鉴定到的CTG有110个，CTdatabase根据CTG的表达特征将其分为睾丸限定的（testis-restricted）、睾丸/大脑限定的（testis/brain-restricted）、睾丸选择性的（testis-selective）3类[3]。我们推测，这110个CTG在本研究分析的睾丸组织表达特异性较差，故造成此差异。由于GTEx数据库和TCGA数据库更新，收录了更多的数据，本研究用这2个数据库的所有编码基因作为初始候选基因，而CTdatabase的初始候选基因是文献中记录的CTG[3]。因此，本研究筛选到的CTG远多于CTdatabase中收录的CTG。

为了鉴定精子发生过程中CTG包含的GEP，本研究重新整理了睾丸单细胞转录组数据，并标注了细胞类型。通过睾丸单细胞转录组数据的分析，发现

CTG主要表达于生殖系细胞中。睾丸的体细胞类型有SC、LC、PC、SMC、EC、TM。除了SC之外，其他的细胞类型或多或少都会在其他成体组织或器官中存在，故特异表达于睾丸的CTG不会表达于这些细胞中。

根据睾丸生殖系细胞的基因表达谱，对CTG进行聚类，鉴定出7个精子发生过程中的GEP。每个GEP有其偏好表达的细胞类型和精子发生时期。对GEP的活跃程度定量计算后，发现GEP5活跃于SSC、Differentiating SPG、Early primary SPC细胞。此外，研究还发现GEP5中包含了大量的CT-X基因[5-6]。这说明在精子发生过程的前期，大部分的CT-X基因需要活跃的转录。X染色体上的基因受到减数分裂性染色体失活调控（meiotic sex chromosome inactivation，MSCI），其在减数分裂过程的前期活跃转录，在完成染色体联会后转录活性快速受到抑制，并形成XY体[19]。这或许意味着在精子发生过程中，GEP5的活跃程度受到MSCI的调控。

GEP5活跃的细胞类型是SSC、Differentiating SPG、Early primary SPC，即精原细胞增殖期和减数分裂早期。之前的研究结果表明，减数分裂早期相关的CTG对于肺腺癌的发生和发展有重要作用[4]。因此，本研究将GEP5在肿瘤中的活跃情况与患者的预后进行关联分析，以探究GEP5的活跃程度与患者预后之间的关系。结果显示，GEP5在ACC、BRCA、COAD、KIRC、KIRP和UCEC中的活跃程度高与患者的生存状况差有显著的关联。这说明GEP5在多种肿瘤类型中的活跃程度与患者的预后有关，具有成为多种肿瘤类型预后判断标志物的潜力。

综上所述，本研究通过多数据库联合分析，对CTG进行重新筛选和分类，并鉴定出7个GEP。其中，GEP5活跃于精子发生过程的前期，且富集CT-X基因。此外，在肿瘤中，GEP5的活跃程度越高，患者的生存状况越差，有望支持多种肿瘤类型的预后判断。

利益冲突声明/Conflict of Interests

所有作者声明不存在利益冲突。
All authors disclose no relevant conflict of interests.

作者贡献/Authors′ Contributions

雷鸣设计并指导了整个课题的研究；侯宗良主要完成数据收集、数据分析及处理、文章撰写；杨琴和李少白主要负责数据分析指导、算法分析指导以及整篇论文的修改。所有作者均阅读并同意最终稿件的提交。
LEI Ming designed and guided the whole research project. HOU Zongliang mainly completed the work of data collection, data analysis and processing and article writing. YANG Qin and LI Shaobai were mainly responsible for the guidance of data analysis and algorithm analysis, as well as the revision of the whole paper. All authors read the final manuscript and approved the submission.

· Received: 2023-03-29
· Accepted: 2023-05-18
· Published online: 2023-08-28

参·考·文·献

[1] SCANLAN M J, GORDON C M, WILLIAMSON B, et al. Identification of cancer/testis genes by database mining and mRNA expression analysis[J]. Int J Cancer, 2002, 98(4): 485-492.
[2] HUBBARD J M, AHN D H, JONES J C, et al. Trial in progress: a phase Ⅱ, multicenter, open-label study of PolyPEPI1018 in combination with atezolizumab in participants with relapsed or refractory microsatellite-stable metastatic colorectal (MSS mCRC) cancer (Oberto-301)[J]. J Clin Oncol, 2023, 41(4 suppl): TPS283-TPS283.
[3] ALMEIDA L G, SAKABE N J, DEOLIVEIRA A R, et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens[J]. Nucleic Acids Res., 2009, 37(suppl_1): D816-D819.
[4] WANG C, GU Y Y, ZHANG K, et al. Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types[J]. Nat Commun, 2016, 7: 10499.
[5] GORDEEVA O. Cancer-testis antigens: unique cancer stem cell biomarkers and targets for cancer therapy[J]. Semin Cancer Biol, 2018, 53: 75-89.
[6] MENG X Y, SUN X Q, LIU Z L, et al. A novel era of cancer/testis antigen in cancer immunotherapy[J]. Int Immunopharmacol, 2021, 98: 107889.
[7] SIMPSON A J G, CABALLERO O L, JUNGBLUTH A, et al. Cancer/testis antigens, gametogenesis and cancer[J]. Nat Rev Cancer, 2005, 5(8): 615-625.
[8] MOUNIR M, LUCCHETTA M, SILVA T C, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx[J]. PLoS Comput Biol, 2019, 15(3): e1006701.
[9] GUO J T, GROW E J, MLCOCHOVA H, et al. The adult human testis transcriptional cell atlas[J]. Cell Res, 2018, 28(12): 1141-1157.
[10] HERMANN B P, CHENG K R, SINGH A, et al. The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids[J]. Cell Rep, 2018, 25(6): 1650-1667. e8.
[11] SOHNI A, TAN K, SONG H W, et al. The neonatal and adult human testis defined at the single-cell level[J]. Cell Rep, 2019, 26(6): 1501-1517. e4.
[12] ALEXANDER WOLF F, ANGERER P, THEIS F J. SCANPY: large-scale single-cell gene expression data analysis[J]. Genome Biol, 2018, 19(1): 15.

[13] LOPEZ R, REGIER J, COLE M B, et al. Deep generative modeling for single-cell transcriptomics[J]. Nat Methods, 2018, 15(12): 1053-1058.

[14] BERNSTEIN N J, FONG N L, LAM I, et al. Solo: doublet identification in single-cell RNA-seq *via* semi-supervised deep learning[J]. Cell Syst, 2020, 11(1): 95-101. e5.

[15] TRAAG V A, WALTMAN L, VAN ECK N J. From Louvain to Leiden: guaranteeing well-connected communities[J]. Sci Rep, 2019, 9(1): 5233.

[16] WANG M, LIU X X, CHANG G, et al. Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis[J]. Cell Stem Cell, 2018, 23(4): 599-614. e4.

[17] FAYOMI A P, ORWIG K E. Spermatogonial stem cells and spermatogenesis in mice, monkeys and men[J]. Stem Cell Res, 2018, 29: 207-214.

[18] BADIA-I-MOMPEL P, VÉLEZ SANTIAGO J, BRAUNGER J, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data[J]. Bioinformatics Adv, 2022, 2(1): vbac016.

[19] LIU W S. Mammalian sex chromosome structure, gene content, and function in male fertility[J]. Annu Rev Anim Biosci, 2019, 7(1): 103-124.

［本文编辑］吴 洋

学术快讯

上海交通大学医学院范先群院士团队在《柳叶刀》子刊发表视网膜母细胞瘤治疗新成果

上海交通大学医学院附属第九人民医院眼科范先群院士牵头，联合首都医科大学附属北京儿童医院、广州市妇女儿童医疗中心、上海交通大学医学院附属新华医院、中国人民解放军总医院第三医学中心和中南大学附属湘雅医院等全国视网膜母细胞瘤诊疗中心，开展全球首个视网膜母细胞瘤眼动脉介入化学治疗（化疗）的多中心前瞻性随机对照研究。研究结果于2023年7月31日以*Intravenous versus super-selected intra-arterial chemotherapy in children with advanced unilateral retinoblastoma: an open-label, multicentre, randomised trial*为题目发表于《柳叶刀》子刊*The Lancet Child & Adolescent Health*。该研究表明，眼动脉介入化疗在不影响总体生存率的前提下，较之静脉化疗能够显著提高晚期视网膜母细胞瘤患儿的保眼率，并显著降低全身并发症的发生率。研究提示，眼动脉介入化疗可作为单侧晚期视网膜母细胞瘤患儿的首选治疗方案。