中山大学硕士学位论文

构建肿瘤假基因-miRNA-mRNA CeRNA 网络与泛癌分析

Construction of Tumor Pseudogene-miRNA-mRNA CeRNA Network and Pan-Cancer Analysis

专业名称: 生物信息学

申请人:

指导教师:

答辩委员会（签名）

主席：

成员：

广州　中山大学　生命科学学院

2020 年 5 月 25 日

论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：2020 年 5 月 25 日

学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版；有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅；有权将学位论文的内容编入有关数据库进行检索；可以采用复印、缩印或其他方法保存学位论文；可以为建立了馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

保密论文保密期满后，适用本声明。

学位论文作者签名：　　导师签名：

日期：2020 年 5 月 25 日　　日期：2020 年 5 月 25 日

构建肿瘤假基因-miRNA-mRNA ceRNA 网络与泛癌分析

专业名称：

申请人：

指导教师：

摘要

假基因是一类与基因相似，但有残缺的基因组序列。自从 1977 年被定义起，假基因长期被科学界认定是没有功能的，是进化上的残缺品。然而近年陆续有研究发现假基因在基因表达调控上具有不可忽视的功能。ceRNA 假说提出假基因可能提供 miRNA 应答元件，与细胞内其他基因竞争 miRNA 从而影响基因表达。

本研究基于假基因特异 exon 注释，对 TCGA 数据库乳腺浸润癌、肾嫌色细胞癌、肾透明细胞癌、肾乳头状细胞癌、胰腺癌、前列腺癌一共 2977 个 RNA-seq 数据，GTEx 数据库乳房组织、肾组织、胰腺组织、前列腺组织共 589 个 RNA-seq 数据进行了假基因表达谱分析，构建 ceRNA 数据库。根据 ceRNA 解说，对六种癌症分别构建 ceRNA 网络，重点讨论了假基因通过竞争性结合 miRNA 调控与假基因相似度较低的蛋白编码 RNA，分析了受调控 mRNA 的功能。在乳腺癌 ceRNA 网络中锌指蛋白 ZNF623 与锌指蛋白假基因 ZNF256P 基因序列相似度较低，但是具有 10 个相同的 miRNA 结合位点。高表达 ZNF256P 和 ZNF623 的样品对应较差的预后（P=0.07，P=0.0005）。基于假基因表达水平，分别对六种癌症进行肿瘤分型。在 KIRC、BRCA、PAAD、PRAD 四种癌症的肿瘤亚型中均存在一个以多个免疫球蛋白假基因作为特征基因的特殊亚型。通过分析四个特殊亚型的特征基因发 IGLC4 和 IGKV2-29 出现在这四个亚型中。分析肿瘤亚型之间免疫细胞浸润水平，这些特殊亚型在各自的癌症中与其他亚型相比有较多的差异。结果暗示免疫球蛋白假基因 IGLC4 和 IGKV2-29 在肿瘤中可能参与免疫调节。基于假基因表达水平，使用 LASSO COX 多因素回归分析构建肿瘤预后模型。KICH 预后模型特征基因 GOLGA2P8 在 ceRNA 网络中竞争性结合 3 种 miRNA，

调控 25 种 mRNA。

本研究构建了六种癌症的 ceRNA 网络，讨论了假基因通过 ceRNA 网络对非相似的蛋白编码基因进行调控；基于假基因差异表达基因的表达水平进行了肿瘤分型及构建预后模型，获得了与肿瘤亚型相关和预后相关的特征假基因。这项工作可能促进癌症的研究，作为未来深入理解假基因的作用的基础，帮助开发新的生物标记和提高对肿瘤生物学的认识。

关键词：假基因，ceRNA 网络，泛癌分析

Construction of Tumor Pseudogene-miRNA-mRNA CeRNA Network and Pan-Cancer Analysis

Major:

Name:

Supervisor:

ABSTRACT

Pseudogenes are a group of genomic sequences that are similar to genes but have incomplete sequences. Since being defined in 1977, pseudogenes have long been considered by the scientific community to be functionally disabled and evolutionally disabled. However, recent studies have found that pseudogenes play an important role in the regulation of gene expression. The ceRNA hypothesis proposes that pseudogenes may provide miRNA response elements that compete with other genes in the cell to influence gene expression.

This study constructed the pseudogene annotation to TCGA database breast ductal carcinoma, and renal cell carcinoma, renal clear cell carcinoma, papillary carcinoma, pancreatic cancer, prostate cancer, a total of 2977 RNA - seq data GTEx database breast tissue, kidney, pancreas, prostate tissue RNA-seq data for a total of 589 RNA gene expression spectrum analysis. Follow-up analysis was conducted based on the expression profile of differentially expressed pseudogenes. According to the explanation of ceRNA, ceRNA networks were constructed for six types of cancer, focusing on the competitive binding of pseudogenes with miRNA-regulated protein-coding RNAs with low similarity of pseudogenes, and the functions of regulated mRNAs were analyzed. In the ceRNA network of breast cancer, the sequence similarity of ZNF623 of zinc finger protein and ZNF256P of zinc finger protein pseudogene is low, but there are 10 identical miRNA binding sites. High

expression of ZNF256P and ZNF623 was associated with poor prognosis (P=0.07, P=0.0005). Based on the expression level of pseudogenes, tumor types were performed for each of the six cancers. In the tumor subtypes of KIRC, BRCA, PAAD and PRAD, there exists a special subtype with multiple immunoglobulin pseudogenes as the characteristic genes. By analyzing the characteristic genes of four special subtypes, IGLC4 and IGKV2-29 appeared in these four subtypes. To analyze the level of immune cell infiltration among tumor subtypes, these specific subtypes were more different in their respective cancers than other subtypes. The results suggest that IGLC4 and IGKV2-29, pseudogenes of immunoglobulin, may be involved in immune regulation in tumors. Based on the expression level of pseudogenes, a tumor prognosis model was constructed using LASSO COX multivariate regression analysis. The prognostic model was successfully constructed on KICH and PRAD.GOLGA2P8, a KICH prognostic model characteristic gene, competitively binds 3 miRNAs and regulates 25 mRNAs in the ceRNA network.

In this study, ceRNA networks of six cancers were constructed, and the regulation of non-similar protein-coding genes by pseudogenes through ceRNA networks was discussed. Based on the expression level of the differentially expressed pseudogenes, tumor types and prognostic models were constructed, and the characteristic pseudogenes related to tumor subtypes and prognosis were obtained. The work could advance cancer research as a basis for a deeper understanding of the role of pseudogenes in the future, helping to develop new biomarkers and improve understanding of tumor biology.

Keywords: Pseudogenes, ceRNA networks, Pan-cancer analysis

目 录

第一章 前言

1.1 假基因

假基因是一类与某的基因相似，但有残缺的基因组序列。自从 1977 年被定义起，假基因长期被科学界认定是进化上的残缺品，自身不具有功能。然而近年陆续有研究发现假基因在基因表达调控上扮演了不可忽视的角色[1]。

1.1.1 假基因来源与分类

根据不同假基因的来源，目前假基因有以下分类[2]。从已加工 mRNA 发生逆转录转座产生的假基因，称为已加工假基因（Processed pseudogenes）。由于基因组片段重复产生的假基因，被称为未加工假基因（Unprocessed pseudogenes）。原始基因发生突变而失活产生的假基因被称为单一假基因（Unitary pseudogenes）。多态假基因（Polymorphic pseudogenes）是指一些假基因在某些个体上是正常基因，而在某些个体上存在失活突变，变成假基因。

1.1.2 假基因的功能

目前转录的假基因被证实有以下功能：假基因可以被翻译成与亲本功能基因相似的完整的蛋白[3, 4]或者翻译成多肽片段通过特定的功能结构域发挥作用[5, 6]。假基因被转录为亲本功能基因的正义或反义产物，以调节功能基因的表达[7]。假基因将 dsRNAs 与亲本基因结合，生成内源性 siRNAs，并通过 RNAi 途径进一步抑制基因表达或者影响蛋白编码 mRNA 的稳定性[8]。假基因作为 ceRNA，与亲本功能基因竞争 miRNA[9, 10]。

1.2 竞争性内源 RNA 假说

1.2.1 竞争性内源 RNA 假说内容

竞争性内源 RNA（ceRNA，competing endogenous RNA）假说是 2011 年哈佛大学医学院著名学者 Pandolfi 等提出[9]。该假说系统地阐述了信使 RNA、转录假基因和长非编码 RNA 如何通过 miRNA 应答元件（MRE，miRNA response elements）相互调节。

根据 Pandolfi 的 ceRNA 假说，在 ceRNA 网络中 mRNA 和 microRNAs 之间的关系可能存在相反的关系，导致其中一种 mRNA 的表达水平影响另一种 mRNA 的表达水平和活性。因此 RNA 分子可以通过 microRNA 和 MREs 进行相互沟通。

RNA 分子之间共享的 MRE 数量越多，共调控的能力越强。RNA 分子 3'UTR 区域存在 MREs。传统认为 MREs 可以作为一种顺式调节因子调控 RNA 分子自身的功能，而在 ceRNA 假说中认为这些 3 UTR MREs 也可以通过反式调控改变相应的 miRNA 和其他 RNA 分子表达水平。

1.2.2 miRNA 简介

miRNA 是细胞生物过程中一种重要的调节因子。miRNA 是一种短的非编码 RNA，约为 22 个碱基长度[11]。目前 miRNA 权威数据库 miRbase 收录了 2656 种人类成熟 miRNA 序列、1917 种 RNA 前体序列[12]。

miRNA 主要通过 AGO 蛋白行驶功能。miRNA 引导 AGO（Argonaute protein，AGO）蛋白结合到 mRNA 3'UTR（Untranslated region，UTR）区域，使得 mRNA

沉默[13]。装载了miRNA的AGO蛋白作为miRNA诱导沉默复合物（miRNA-induced silencing complex，miRISC）的靶向模块，促进靶向mRNA的降解并抑制mRNA翻译[14]。超过60%人类蛋白编码基因可以被检测到miRNA结合位点。miRNA功能失调会导致多种疾病，特别是癌症。

miRNA5’端有一段2-8的碱基序列，称为种子序列，是miRNA与mRNA结合的区域。根据其种子序列相似性可以分类成不同的miRNA家族。

1.2.4 假基因与ceRNA

类似于蛋白编码基因，假基因上有着大量miRNA 结合位点。 Starbase数据库[15]对108个CLIP-seq数据分析，鉴定了16000个miRNA-假基因调控关系。

在ceRNA假说中，假基因是一种典型的ceRNA。它可以通过与其他基因竞争性结合miRNA，从而调节靶基因的功能。

2010年Pliseno 等的研究首次显示假基因PTENP1和KRASP1分别与其对应的亲本功能基因PTEN和KRAS通过竞争共同的miRNA调控亲本功能基因的功能，影响肿瘤的生长与发展[1]。Karreth等报道了原癌基因BRAF与其假基因BRAFP1之间的ceRNA调控关系，以及两者在小鼠上的同源基因Braf 和Braf-rs1的ceRNA调控关系[16]。最近有研究发现Forkhead转录因子家族成员Foxo3 受到环化的Foxo3和Foxo3 假基因Foxo3P的调控，三者具有相同的八个miRNA竞争位点。作者发现异位表达Foxo3、环状Foxo3，Foxo3P均会影响细胞增殖与生存[17]。在前列腺癌中FTH1基因受到多种假基因通过竞争miRNA调控，导致肿瘤发生[18]。

但是假基因在癌症种对亲本功能基因的影响仍然不确定。PTEN是一种抑癌基因。目前研究普遍认为同源假基因PTENP1可以通过与PTEN竞争miRNA调节PTEN表达水平[1]。然而最近有研究发现在前列腺癌细胞系DU145中删除PTEN不影响PTENP1表达，删除PTENP1对PTEN mRNA水平也没有影响 [19]。但是删除PTEN或PTENP1会减少PTEN蛋白翻译，且提高细胞增殖[1, 19]。

1.3 假基因对癌症的影响

目前已有相当多在癌症上的研究发现假基因在癌症上扮演了不可忽视的角色。

在目前研究中，假基因一般作为一种 miRNA 海绵体调控亲本功能基因的表达与肿瘤发生。因此假基因的作用一般取决于其调控的功能基因的作用。根据其对癌症的作用效果可以分为致癌假基因和抑癌假基因。

致癌假基因典型的例子是 KRASP1 和 BRAFP1。BRAFP1 可以与 BRAF 竞争性结合多种 miRNA，使得 BRAF 表达上调，从而激活 MAPK 通路并诱发淋巴瘤[16]。

抑癌假基因典型例子有 PTENP1、TUSC2P、TIMP2 和 TIMP3。PTENP1 与 PTEN 竞争性结合 miR-17、miR19、miR20a 和 miR-21。在透明细胞癌[20]和口腔鳞癌[21]的研究中真实了存在有功能的 PTENP1-miR-21-PTEN 调控轴。在胃癌的研究中也发现具有肿瘤抑制作用的的 PTENP1/miR-106b/miR-93/PTEN 调控轴[22]。

1.4 高通量测序与癌症假基因 ceRNA 网络研究

1.4.1 高通量测序项目数据库

TCGA 项目介绍。肿瘤基因组图谱（The Cancer Genome Atlas，TCGA）是 2006 年由美国国家癌症研究所和美国国家人类基因组研究所联合启动的项目。目前 TCGA 一共收录了 33 种癌症，11,315 个案例。数据类型包括基因分型矩阵、甲基化矩阵、RNA 测序、全外显子组数据、诊断图片、组织图片、单细胞 ATAC-Seq

数据。

GTEX 项目。组织基因型表达计划（Genotype-Tissue Expression，GTEx）启动于 2010 年。该计划提供了数据资源和组织数据库，用于研究在组织间和个体间遗传变异和基因表达的关系。目前 GTEx 数据库包括了来自 714 个供体和约 11688 个 RNA 测序数据，覆盖率 53 种组织和 2 种细胞系。

1.4.2 使用大规模测序数据研究假基因

目前 TCGA 和 GTEX 数据库一共记录了超过两万个 RNA 测序样本，已有许多研究基于这些测序样本研究假基因。

Wei 等基于 ceRNA 假说在 TCGA 肺腺癌中鉴定了 33 种竞争性内源假基因。结合 DNA 甲基化分析，其中 21 个假基因与其竞争性 mRNA 发生共甲基化。在共甲基化网络中，研究人员发现了 6 种差异表达假基因，称为潜在 LUAD 相关假基因[23]。Liu 等通过 LASSO-COX 建模筛选低级别胶质瘤(LGG)的预后假基因，并探索其潜在的分子机制[24]。在头颈部鳞状细胞癌的研究[25]中发现 PTTG1、PTTG2 基因和伪基因 PTTG3P 的致癌作用。在乳腺浸润性癌的研究[26]正面临转录的假基因表现出亚型特异性表达和 ceRNA 潜能

除了研究假基因在单个癌症中的功能，研究人员还对大规模数据进行泛癌分析。GEPIA 数据库[27]综合整理了 TCGA 和 GTEX 数据库中的表达数据，其中包括了对假基因表达谱的分析。Liang 等分析了 7 种癌症的假基因表达水平，基于假基因表达水平进行肿瘤分型。Dreambase 数据库分析了 TCGA 和 GTEx 数据库的 RNA 测序数据中有表达的假基因的 DNA 修饰、RNA 调控和蛋白质结合事件。

1.5 研究目的与方法

目前已经有许多研究证实假基因在细胞活动中可以被转录为 RNA，但是功

能尚不明确。随着二代测序不断成熟，已经积累了大量高通量 RNA 测序数据。在许多基于 RNA seq 数据的研究中，可以检测到假基因的表达。因此可以利用高通量测序数据研究假基因的功能。

要从 RNA 测序数据中研究假基因，第一个关键的步骤是对假基因进行注释。目前对假基因注释数据库主要有 GENCODE 数据库、假基因装饰资源（Pseudogene Decoration Resource ，psiDR）、ENSEMBL 数据库。这些数据库的注释文件被广泛应用于各种测序数据分析。如 TCGA 数据使用 GENCODE v22 版本的注释文件对 RNA seq bam 文件进行定量。Dreambase 数据库[28]对假基因的注释参考了耶鲁假基因数据库（现已整合到 GENCODE 数据库）、psiDR、ENSEMBL 数据库和 GENCODE 数据库的注释，取同时被两个数据库检测到的假基因注释进行研究。

然而，由于假基因通常是从基因组中其他功能基因产生，因此假基因注释往往与蛋白编码基因注释有交叉，导致难以区分 reads 的真正来源。Liang 等[29]的研究中过滤了假基因注释文件中与蛋白编码基因交叉的外显子和可比对性（Mappability）差的外显子，构建了一个包括 9925 个假基因的注释文件。这种方法获得的假基因在基因组中高度特异，保证了假基因定量的准确性，但是也过滤了一部分假基因。这种分析流程中被过滤的这部分假基因在基因组中与蛋白编码基因交叉，或者具有相似的基因。而保留下来的注释在基因组中缺少与之相似的基因组序列。但是这些在基因组上高度特异的假基因不一定是没有功能的。在 TCGA 肺腺癌的研究中[23]，研究人员过滤了与蛋白质重叠的假基因序列，并构建 ceRNA 网络，甲基化测序的结果侧面验证了假基因与其调控的蛋白编码基因的关系。

目前基于假基因表达水平的功能研究更多关注假基因是如何调控其亲本功能基因或者相似的基因。Dreambase 数据库[28]使用 TCGA 数据库和 GTEx 数据约 18000 个 RNAseq 数据，研究了假基因与其亲本功能基因的共表达图谱，构建 ceRNA 网络讨论了假基因如何通过竞争 miRNA 调控亲本功能基因的表达。在本研究中希望通过使用 liang 的方法构建假基因注释文件，研究假基因在多种癌症中对基因组中相似度低的功能基因的调控功能。[1]

基于以上目的本研究主要包括以下内容：

（1）参考前人研究的方法构建假基因注释文件，并进行定量和差异表达分析。因为该研究发表于 2014 年，而目前基因组注释已经更新了 15 个版本。且该注释流程使用的软件也有了更新的替代。另外相比 2014 年，TCGA 数据库数据量也增加多了。本研究中还纳入了 GTEx 数据库 RNA 测序数据作为正常样品的补充。因此数据量的扩充也有望得到更加准确的结果。在本研究中还额外分析了 4 种癌症。

（2）构建 ceRNA 网络。假基因被广泛认为在转录水平可以通过竞争 miRNA 调节靶基因的 RNA 水平。在本研究中使用的假基因注释在基因组中高度特异，希望通过这种注释分析假基因对非相似基因的调控作用。

（3）基于假基因表达水平进行肿瘤分型。同一癌症不同亚型往往具有不同的生物学机制。因此不同亚型的特征基因往往参与了不同的功能。本研究希望通过肿瘤分型获得特征假基因，并讨论其生物学意义。

（4）基于假基因表达水平进行预后模型的构建。通过构建预后模型，鉴定与病人生存风险相关的假基因。

第二章 材料与方法

2.1 数据来源与分析工具

本研究使用 TCGA 数据库癌症项目乳腺癌[30, 31]（Breast invasive carcinoma，BRCA）、肾透明细胞癌[32]（Kidney renal clear cell carcinoma ，KIRC）、肾乳头状细胞癌[33]（Kidney renal papillary cell carcinoma ，KIRP）、肾嫌色细胞癌[34]（Kidney Chromophobe ，KICH）、胰腺癌[35]（Pancreatic adenocarcinoma ，PAAD）、前列腺癌[36]（Prostate adenocarcinoma ，PRAD）RNA 测序数据及表达谱数据、miRNA 表达谱数据和对应的 GTEX 数据库乳房、肾、胰腺、前列腺正常组织 RNA 测序数据[37]进行 ceRNA 网络构建和后续分析。

分析流程使用的数据来源如表 2-1：

表 2-1 数据来源

Table 2-1 Data source

数据	来源	链接
TCGA RNA 测序数据（Bam 文件）	TCGA	https://portal.gdc.cancer.gov/
TCGA 表达谱数据（FPKM）	TCGA	https://portal.gdc.cancer.gov/
TCGA 表达谱数据（HtseqCounts）	TCGA	https://portal.gdc.cancer.gov/
TCGA miRNA 表达谱数据（miRNA Isoform）	TCGA	https://portal.gdc.cancer.gov/
GTEx RNA 测序数据	GTEx	https://www.gtexportal.org/home/datas

（fastq 文件）		ets
GTEx 表达谱数据（Reads Counts）	GTEx	https://www.gtexportal.org/home/datasets
人类基因组参考文件 Hg38.fa	UCSC	http://genome.ucsc.edu/
GENCODE v34 人类基因注释	GENCODE	ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34/gencode.v34.annotation.gff3.gz
GENCODE v34 假基因注释	GENCODE	ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34/gencode.v34.2wayconspseudos.gff3.gz
miRNA 成熟体序列	miRbase release 21	http://www.mirbase.org/
蛋白编码基因 miRNA 结合位点信息	starbase 2.0	http://genome.ucsc.edu/
免疫细胞浸润评分	TIMER 2.0	http://timer.cistrome.org/infiltration_estimation_for_tcga.csv.gz

本研究主要使用的分析工具和版本如下表 2-2:

表 2-2 分析工具和版本

Table 2-2 Analysis tools and versions

工具	版本
Fastq-dump	2.8.2
Trimmomatic	3.8
STAR	2.7.5c
GenMap	1.2.0
Samtools	1.7

Bedtools	v2.29.2
featureCounts	v2.0.1
MiRanda	v3.3a
Python	2.7
R	4.02
R 包 ggplot2	3.3.2
R 包 survival	3.2.7
R 包 survminer	0.4.8
R 包 glmnet	4.0-2
R 包 NMF	0.23.0
R 包 RCy3	2.9.4
R 包 org.Hs.eg.db	3.12.0
clusterProfiler	3.17.5
Cytoscape	3.8

2.2 数据分析

2.2.1 构建假基因注释文件

从 GENCODE 下载 v34 基因组注释 gencode.v34.annotation.gff3 文件和 gencode.v34.2wayconspseudo.gff3 假基因注释文件。提取假基因外显子注释。

GENCODE 数据库[38]是 ENCODE 计划的成果，由 Sanger 研究所负责维护。该数据库致力于基于生物学证据提供人类和小鼠所有基因组特征的注释。基因组注释 gencode.v34.annotation.gff3 记录了参考染色体上的综合基因注释。根据注释证据的来源把注释的可靠性分为三个等级。Level 1 表示为已验证注释。假基因需要同时通过 Yale Pseudopipe 、UCSC Retrofinder pipeline 和 Havana manual

annotation， 方可注释为 Level 1。其他基因需要通过 GENCODE 实验流程的 RT-PCR 和测序实验验证方可注释为 Level 1。Level 2 表示为人工注释，即通过 Havana manual annotation 验证注释。Level 3 为自动注释，即与 Havana 注释不同或没有 Havana 注释的 Ensembl 位点。假基因注释 gencode.v34.2wayconspseudo.gff3 为耶鲁大学和加利福尼亚大学圣克鲁兹分校假基因分析流程共同注释到的假基因。

为了得到能准确定量的假基因外显子序列，对假基因注释进行了如下过滤：

（1）使用 bedtools 移除假基因外显子中与编码蛋白质基因的外显子重叠的序列。2wayconspseudo 假基因注释有部分注释和 GENCODE v34 基因组注释的假基因有重叠。考虑到 GENCODE 的注释有更充足的证据，因此 gencode.v34.2wayconspseudo.gff3 假基因注释中与重叠的假基因移除。

（2）过滤在基因组可比对性低的外显子序列。使用 GenMap 软件[39]计算上一步保留的各个外显子平均可比对性评分（alignable score）。在 kmer=75，允许两个错配的情况下，平均可比对性评分大于 0.95 的外显子定义为在基因组上唯一存在的序列。

把上述流程获得的外显子注释整理成 gtf 文件用作后续分析。

2.2.2 RNA 测序数据预处理

从 TCGA 数据库下载 RNA-seq Bam 文件。TCGA RNA seq BAM 文件是由 STAR 2-PASS 流程生成。TCGA 数据库提供了详细的流程详细参数：https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#rna-seq-alignment-command-line-parameters。

从 GTEx 数据库下载 SRA 文件，使用 fastq-dump 工具对数据进行解压，Trimmomatic 进行数据清洗。对成对的 fastq 文件进行比对。为了与 TCGA 数据尽可能保持一致，这里参考 TCGA 分析流程进行 STAR 2-PASS 比对。Bam 文件过小（<1G）的 GTEX 样本从本研究移除。

得到 TCGA 的 Bam 文件和 GTEX 数据的 BAM 文件后，使用 samtools view -F

256 去除多重比对结果。然后使用 samtools sort -n 进行排序。

2.2.3 基因表达水平分析

利用前面得到的假基因注释.gtf 文件对 BAM 文件进行假基因表达定量。使用 featureCounts 计算各个样本在假基因上的 reads 数。

使用上一步得到的假基因注释文件 R 脚本计算假基因 FPKM 值。FPKM 的计算如下

$$FPKM = \frac{10^6 \times n_f}{L \times N} \quad （公式 2-1）$$

L 为过滤后假基因注释长度，n_f 为比对上过滤后假基因注释上的片段数，N 为总片段数。同一个样本重复测序的结果取平均值合并为一个结果。为了保证结果的可靠性，在本研究中，每种癌症仅保留在肿瘤样本中平均 FPKM 大于 0.3 的假基因进行后续分析。

在本研究中把 TCGA 肿瘤组织作为肿瘤组、TCGA 癌旁组织和 GTEx 组织作为正常组。

使用 edgeR R 包进行差异表达计算。假基因差异表达使用前述步骤中 featureCounts 的结果。基因差异表达使用 TCGA 数据库提供的 htseq 数据和 Xena 提供的 GTEX reads Counts 数据。miRNA 差异表达使用 TCGA 数据库提供的 ISOform 定量结果。把 Log2FC >1,FDR <0.05 的基因定义为差异表达基因。

2.2.4 构建 ceRNA 网络

使用肿瘤组和正常组之间差异表达假基因、蛋白编码基因和 miRNA 构建 ceRNA 网络。

根据 ceRNA 假说，构建 ceRNA 网络包括以下步骤：

1 获得 mRNA 和假基因上 miRNA 结合位点。从 dreambase 数据库下载位于

mRNA 的 miRNA 结合位点信息。使用 miRanda 工具预测位于假基因的 miRNA 结合位点信息。为了与 TCGA ISOform miRNA 表达谱数据保持一致，miRanda 工具使用 miRbase 版本 21 的 miRNA 参考序列（目前为版本 22）。

2 提取差异表达 miRNA 的假基因-miRNA、mRNA-miRNA。

3 移除具有相同的表达趋势的 假基因-miRNA、mRNA-miRNA。即保留皮尔森相关系数 cor<0 p<0.05 的假基因-miRNA、mRNA-miRNA。

4 计算竞争相同 miRNA 的假基因和 mRNA，假基因-mRNA 表达水平 FPKM 值的皮尔森相关系数，满足 cor >0.7，P<0.05 的 mRNA-假基因对保留。

5 计算竞争相同 miRNA 的假基因和 mRNA，假基因-mRNA 表达水平比值。在本文中竞争相同 miRNA 的假基因和 mRNA 表达水平相差不超过 100 倍。

6 如果筛选后的 ceRNA 网络中出现 miRNA 所结合的 mRNA 和假基因全被移除，则该 miRNA 也剔除。

上述过程使用 shell 脚本和 python 脚本实现。皮尔森相关系数使用 python math 模块计算。获得假基因-miRNA-mRNA 网络后使用 RCy3 和 Cytoscape 进行数据可视化。

2.2.5 建立生存风险评分模型

为了了解假基因对病人预后的影响，本研究使用差异表达假基因的表达谱数据分别对各种癌症构建预后模型。具体流程可以分为以下两步：

第一步使用 Cox 比例风险模型分别对各个癌症种差异表达的基因进行标准单变量回归分析，将其中表达水平和病人生存存在显著关联（p<0.05）的基因作为候选基因。

第二步使用 LASSO 算法（Least absolute shrinkage and selection operator）进一步筛选关键的基因，然后进行多因素 COX 建模。目前 LASSO 已广泛应用于具有高维数据的基于 Cox 比例风险回归模型的生存分析[40-42]。LASSO 使用 L1 正则化使 Cox 回归中的部分变量的回归系数可压缩为 0，通过调整正则参数 λ的值，我们能够控制最终模型中变量的数目[43]。通过使用该模型，我们进一

步从候选基因中筛选出最有用的预后相关特征。

最后根据模型得到的系数计算每个样本的生存风险评分（Risk Score）。生存风险评分计算使用以下公式：

$$\text{RiskScore}=\sum_{i=1}^{n} wixi \quad \text{公式（2-2）}$$

对生存风险评分绘制 3-5 年多个时间点的 ROC 曲线，评价模型的准确性。

LASSO 回归使用 R 语言 glmnet 包实现，COX 回归模型和生存分析使用 R 语言 survival 包、survminer 包实现。使用 R 语言基础作图工具和 ggplot2 包绘制模型的森林图。

2.2.6 基于差异表达假基因表达水平分析肿瘤亚型

为了使用假基因表达水平分别对各个癌症的肿瘤样本进行分类，选择各个癌症的差异表达的假基因的表达水平分析。这里使用非负矩阵分解法[44]（NMF）进行聚类，并提取各个簇的特征基因。

在本研究中使用 NMF R 包实现。第一步确定分解等级（factorization rank ,r）。确定分解等级一般策略为在多个不同 r 值下进行 NMF，然后计算某些衡量结果质量的指标，最后选择质量指标最优的 r 值。在本研究中使用同一个癌症的原发性肿瘤样本和继发性肿瘤样本差异表达基因的表达水平数据作为输入矩阵，分别在 r 值为 2、3、4、5、6 进行 50 次非负矩阵分解。选择同表象系数（cophenetic coefficient）曲线的拐点处的 r 值进行下一步分析。第二步使用该 r 值进行 NMF 分析，nrun 设置为 200 次。

2.2.7 免疫微环境

本文使用 TIMER 2.0 数据库提供的 TCGA 数据库各个癌症的 TIMER 值作为衡量肿瘤组织中免疫细胞浸润程度的指标。TIMER 2.0 提供了六种 B 细胞, CD4+ T 细胞, CD8+ T 细胞, 中性粒细胞, 巨噬细胞和树突细胞数据。这些数据用于分

析通过非负矩阵分解法获得的不同的假基因亚型之间免疫微环境的差异。使用 wilcox.test 对各个亚型的免疫细胞浸润得分进行两两比较，取 P<0.05 为差异显著。

2.2.8 基因功能注释

使用基因本体论（Gene ontology，GO）富集分析进行基因功能注释。使用 R 包 clusterProfiler 实现，参数选择 ont = 'ALL', pAdjustMethod = 'fdr', pvalueCutoff = 0.05, qvalueCutoff = 0.2。

2.2.9 生存分析

使用 R 的 survival 包进行生存分析、使用 survminer 包实现可视化。

2.2.10 数据统计与可视化

在本研究中统计分析和作图如无特殊说明均为使用 R 语言实现。作图使用 R 包 ggplot2。

第三章 结果与讨论

3.1 假基因注释文件构建

从 GENCODE 网站下载 V34 版本的基因注释文件和 2wayconspseudos 假基因注释。GENCODE V34 基因注释文件中包括 15265 个假基因，44668 个外显子。2wayconspseudos 假基因注释提供了 9020 个假基因转录本，其中 8565 个转录本与 GENCODE V34 假基因注释重合度大于 90%、5667 个转录本完全重合、377 个转录本完全不重合。因为 GENCODE 的基因注释拥有更多证据，因此仅把 2wayconspseudos 假基因注释中 377 个 GENCODE V34 假基因注释纳入到假基因注释文件构建。

过滤中与蛋白编码基因重叠的外显子序列，得到 43941 个外显子 。使用 GenMap （参数 -K 75 -E 2）计算候选假基因注释各个位点在 hg38 上的可比对性。计算各个外显子平均可比对性，平均可比对性低于<0.95 的序列移除。最后得到的假基因注释文件包括来自 9455 个假基因的 23832 个外显子。如图 3-1。

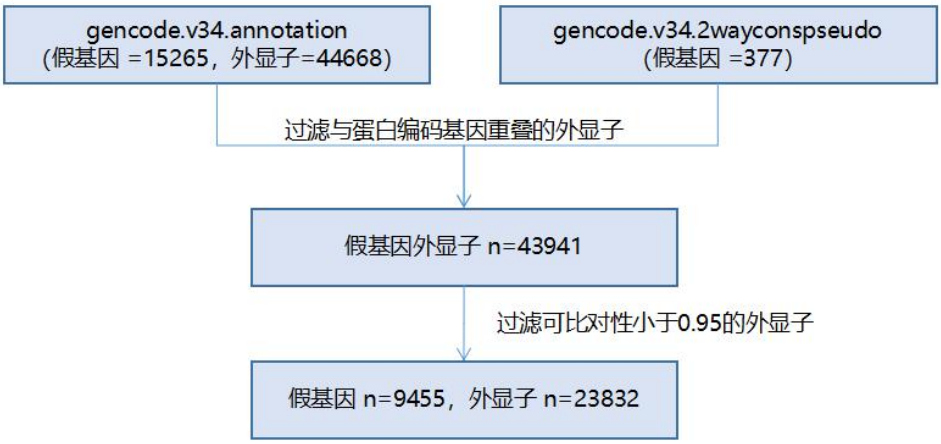


图 3-1 假基因注释文件构建过程
Fig 3-1 Pseudogene annotation file construction process

3.2 计算假基因表达水平

基于上一步获得的假基因注释文件，对 TCGA 六种癌症及其在 GTEX 数据库对应的组织的 RNA seq 数据进行定量。本研究中选择的六种癌症为：乳腺癌（Breast invasive carcinoma， BRCA）、肾透明细胞癌（Kidney renal clear cell carcinoma ，KIRC）、肾乳头状细胞癌（Kidney renal papillary cell carcinoma ，KIRP）、肾嫌色细胞癌（Kidney Chromophobe ，KICH）、胰腺癌（Pancreatic adenocarcinoma ，PAAD）、前列腺癌（Prostate adenocarcinoma ，PRAD）。

首先下载 TCGA 数据的 bam 文件和 GTEx 数据库的 fastq 文件。然后对下载到的 GTEX fastq 文件参考 TCGA 数据库运行 STAR 2 PASS 比对流程。参考前人研究去除 TCGA 和 GTEx 的 bam 文件中多重比对的序列并排序。最后使用 featureCounts 计算落在注释文件上的 reads 数，根据公式（2-1）计算 FPKM 值。

在本研究中，定义假基因在各个癌症的 TCGA 样本中平均 FPKM >0.3 为可检测假基因。在后续分析中只可检测假基因作为研究对象。

统计结果显示假基因在六种癌症中广泛表达。如图 3-2 （A），本文研究的六种癌症中表达的可检测假基因数量为 600~700 个。其中 410 个假基因在六种癌症中均有表达，288 个假基因仅在一种癌症上表达，结果如图 3-2 （B）。在本研究中，在 BRCA 和 KIRC 中检测到的假基因数量分别为 726 和 687。在前人使用类似的假基因注释文件的研究中， BRCA 检测到 747 个假基因、KIRC 检测到 712 个假基因。这种差异可能是多方面结果造成的，如原始基因组注释文件更新、使用了不同的可比对性计算工具、使用了 GTEx 数据等。

表 3-1 TCGA 和 GTEX 样本数统计

Table 3-1 TCGA and GTEX sample statistics

癌症	TCGA 项目	TCGA 样本数	GTEX 组织	GTEX 样本数
乳腺癌	BRCA	1217	乳房	217
肾透明细胞癌	KIRC	607	肾	36
肾乳头状细胞癌	KIRP	321	肾	36
肾嫌色细胞癌	KICH	89	肾	36
胰腺癌	PAAD	182	胰腺	196
前列腺癌	PRAD	552	前列腺	125

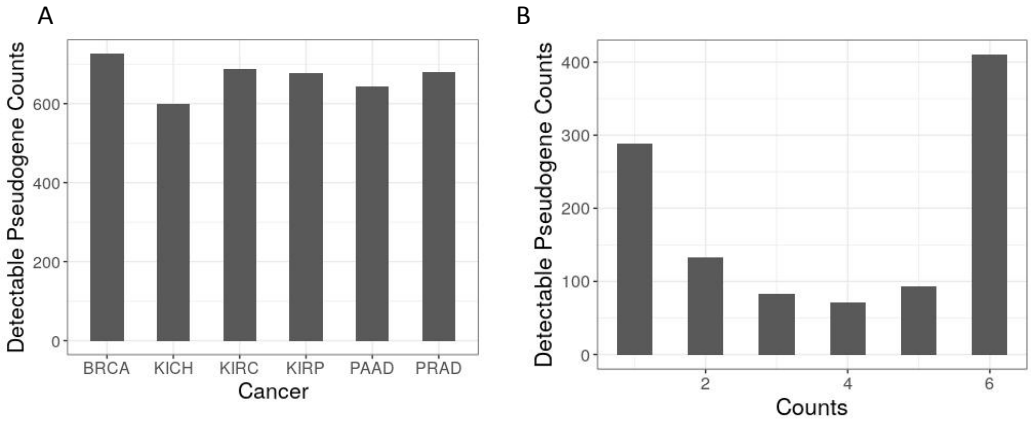


图 3 - 2 可检测假基因在六种癌症的表达。（A）六种癌症检测到的假基因数量。（B）假基因在六种癌症中分布情况

Fig 3-2 The expression of detectable pseudogenes in six cancers. (A) The number of pseudogenes detected in six cancers (B) The distribution of pseudogenes in six cancers

3.3 基因差异表达分析

对肿瘤组和正常组的假基因、蛋白编码 mRNA、miRNA 进行差异表达分析，获得与肿瘤相关的差异表达基因。在本研究中使用 R 包 edgeR 实现，把结果中 Log2FC >1，P<0.05 的基因视为差异表达基因。

在本研究中把 TCGA 原发性肿瘤样本、继发性肿瘤样本作为肿瘤组，把 TCGA 项目癌旁样本、GTEx 样本作为正常组。

假基因差异表达分析使用 mRNA 差异表达分析采用 TCGA 下载的 HtseqCounts 数据和 GTEx 数据库下载的 Counts 数据。miRNA 差异表达分析采用 TCGA 下载的成熟体 counts 数据。

表 3-2 为各个癌症假基因、mRNA、miRNA 差异表达结果。PAAD miRNA 差异表达结果由于缺少正常样本导致差异表达 miRNA 较少，只有 9 个。

表 3-2 六种癌症差异表达假基因、mRNA、miRNA 数

Table 3-2 Differentially expressed pseudogene, mRNA and miRNA counts of six cancers

癌症	TCGA 项目	假基因	mRNA	miRNA
乳腺癌	BRCA	280	443	4377
肾透明细胞癌	KIRC	170	3225	296
肾乳头状细胞癌	KIRP	139	3057	427
肾嫌色细胞癌	KICH	249	4391	380
胰腺癌	PAAD	297	5375	9
前列腺癌	PRAD	232	3177	189

3.4 ceRNA 网络

对差异表达的结果进行筛选，6 种癌症 ceRNA 网络假基因、mRNA、miRNA 节点数如表 3-3。根据筛选条件，在本研究构建的 ceRNA 网络中假基因-mRNA 具有相同的 miRNA 结合位点且表达水平呈显著正相关关系（cor>0.7，P<0.05）。假基因-miRNA，mRNA-miRNA 呈负相关关系（cor<0，P<0.05）。

表 3-3 六种癌症 ceRNA 网络假基因、mRNA、miRNA 节点

Table 3-3 Nodes of pseudogene，mRNA、miRNA in ceRNA network of six cancers

癌症	TCGA 项目	假基因	mRNA	miRNA
乳腺癌	BRCA	16	58	56
肾透明细胞癌	KIRC	20	165	49
肾乳头状细胞癌	KIRP	11	69	24
肾嫌色细胞癌	KICH	76	925	68
胰腺癌	PAAD	1	5	1
前列腺癌	PRAD	19	92	30

图 3-2 为乳腺癌 ceRNA 网络。该 ceRNA 网络包含 16 个假基因节点、58 个 mRNA 节点、56 个 miRNA 节点，89 对假基因-miRNA、198 对 mRNA-miRNA。6 个假基因拥有超过 10 个 miRNA 竞争位点 AC241952（n=10），AHSA2P（n=10），AP000347（n=10），GGTA1P（n=10），ZNF252P（n=10），TSSC2（n=11）。其中 ZNF252P 与 ZNF623 竞争 10 种 miRNA。分别按各种在肿瘤组织的表达水平中位数把 ZNF252P 和 ZNF623 分为高表达水平组和低表达水平组。图 B 和图 C 使用生存分析比较了 ZNF252P 和 ZNF623 在高低表达水平时病人预后情况。结果显示在 ZNF252P 高表达时，病人预后较差（P=0.076），在 ZNF623 高表达

时，病人生存率显著下降（P=0.005）。ZNF252P 基因长度为 29311 位于 8 号染色体 145002899 - 144973589、ZNF263 长度为 17712，位于 8 号染色体 143636019 -143653730。通过 blastn 比对两者序列，ZNF252P 与 ZNF263 比对最大得分为 364、总体得分为 5226、覆盖率为 21%、E 值为 0.07。因此两者在 DNA 层面相似度较低。这一结果说明了假基因对非亲本功能基因仍然可能行使调控功能，并可能通过这一调控影响肿瘤的发展。

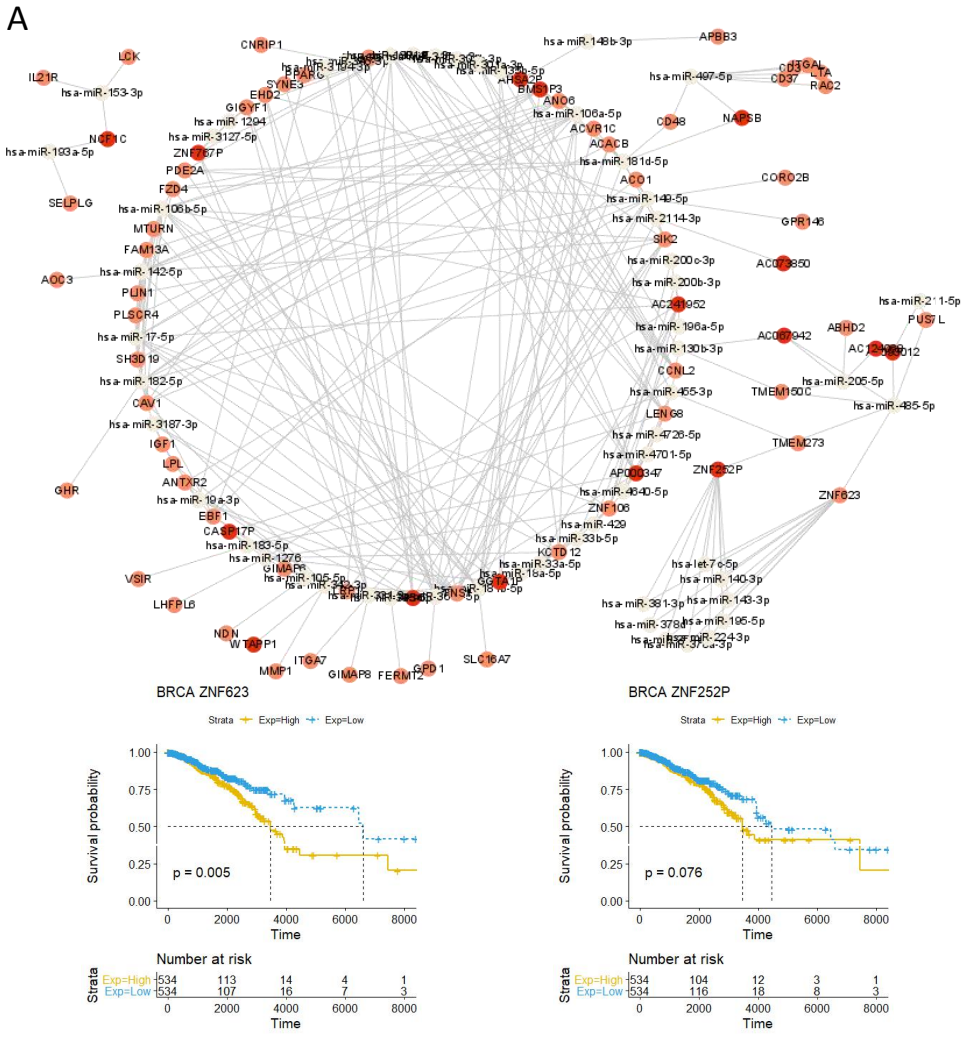


图 3-2 （A）乳腺癌 ceRNA 网络；基于表达水平对乳腺癌 ceRNA 网络中 ceRNA （B）ZNF252P 和（C）ZNF623 进行生存分析

Fig 3-2 （A）ceRNA network of Breast Cancer. Survival analysis of（B） ZN252P and （C）ZNF623 base in their expression level.

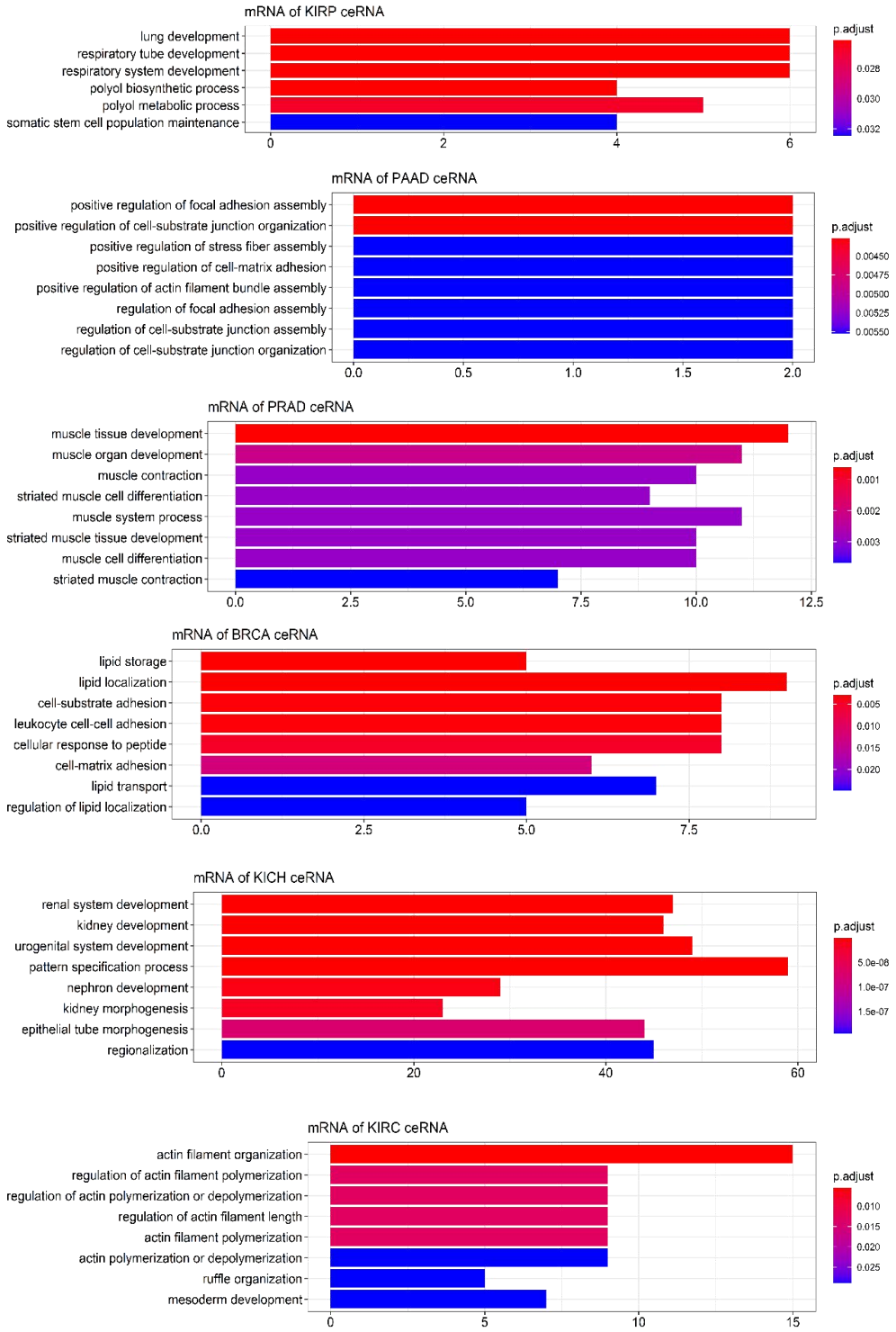


图 3-3 mRNA GO 富集分析

Fig 3-3 mRNA GO enrichment analysis

由于大量假基因缺少功能注释，因此本研究对各个 ceRNA 网络中 mRNA 进行基因功能富集分析，以此说明网络功能。图 3-3 为 6 种癌症 ceRNA 网络 GO 富集分析结果。

在 GO 富集分析中，PRAD ceRNA 网络的 mRNA 富集到与肌肉发育相关的基因。BRCA ceRNA 网络的 mRNA 富集到与脂肪代谢、细胞连接等相关基因。KICH 主要富集到与肾脏发育相关的基因。KIRC 主要富集到与肌动蛋白功能相关的基因。根据上述结果，假基因调控的 ceRNA 网络对细胞正常的功能具有一定的调控作用。

为了进一步了解 ceRNA 网络中假基因对病人预后的影响，根据各个假基因在肿瘤细胞中的中位表达水平把样本分为高低表达水平组，然后对两组病人进行生存分析。在 KIRC ceRNA 网络中有 20 个假基因，其中 9 个假基因与生存显著相关。这 9 个与生存相关的假基因有 5 个有超过一个 miRNA 结合位点，平均 3.44 个 miRNA 结合位点。它们调控了 142 个 mRNA。对这些 mRNA 进行 GO 富集分析如图 3-4（A）显示这些 mRNA 主要与细胞连接相关。

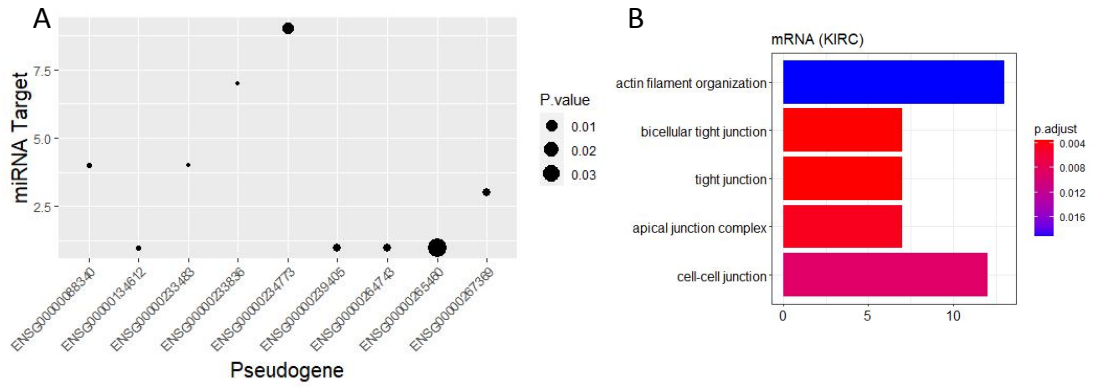


图 3-4 （A）KIRC 中 与生存相关的假基因。横坐标为基因表达水平与生存显著相关的假基因，纵坐标为假基因上 miRNA 结合位点数量，黑点大大小表达 P 值，黑点越小 P 值越显著。（B）对与生存相关的假基因调控的 mRNA 进行 GO 富集分析。

Fig 3-4 （A） The pseudogenes associated with survival in KIRC.The horizontal coordinate is the pseudogenes whose gene expression level is significantly related to survival, and the vertical coordinate is the number of miRNA binding sites on the

pseudogenes. The larger the black point is, the more significant the P value is.（B）GO enrichment analysis was performed on the mRNA regulated by the survival related pseudogenes.

3.5 基于差异表达假基因肿瘤分型

癌症是一类复杂疾病，涉及机体多个层面的异常。因此难以用某一种分子定义肿瘤的亚型。随着多组学的发展已经有多种分子数据如 mRNA 表达水平、miRNA 表达水平、DNA 甲基化、体细胞突变和蛋白质表达水平用于肿瘤分型。前人研究使用假基因表达水平进行肿瘤分型，假基因分型结果与其它分子分型结果在多种癌症中显示广泛的相似性[29]。在本研究中重复对 KIRC 和 BRCA 的分型，并对 KIRP、PAAD、PRAD、KICH 进行分型。

参考前人研究使用非负矩阵分解法[29, 44, 45]，基于差异表达假基因表达水平对肿瘤组织进行分型。图 3-5 为各个癌症同表象相关系数曲线，展示了分解等级 r 值为 2、3、4、5、6 时同表象相关系数值。在各个癌症中选择曲线拐点作为 r 值[44]。因此 KIRP、PAAD、PRAD、BRCA、KICH、KIRC 分别可以分为 3、3、3、3、4、2 类。表格 3-3 统计了各个癌症亚型数及各个亚型特征基因数。附表为各癌症 NMF 假基因分型中各分型的特征基因。KIRC 和 BRCA 的分型结果与前人研究一致[29]。

为了分析比较本研究基于假基因表达水平的 NMF 分型与已有成熟的肿瘤分型的关系，比较了 NMF 分型和 PAM50 分型和三阴乳腺癌（TNBC）的关系。图 3-6 为 BRCA 差异表达假基因表达水平热图，纵轴为样本，横轴为假基因。表达水平为 log2（FPKM+0.001），并按列进行标准。横轴和纵轴均进行了聚类，使用的默认方法为“complete”。如图所示，NMF 分型 cluster3 与基底型乳腺癌相近。

在不同的肿瘤亚型中具有不同的生物学过程。通过研究不同亚型中特异的分子可以进一步了解不同亚型之间肿瘤发展的差异，从而为肿瘤的精准治疗提供更多候选基因。本文研究的假基因缺少基因功能注释，因此在这里尝试从假基因的 GENCODE 注释、免疫细胞浸润得分理解这些特征假基因的功能。

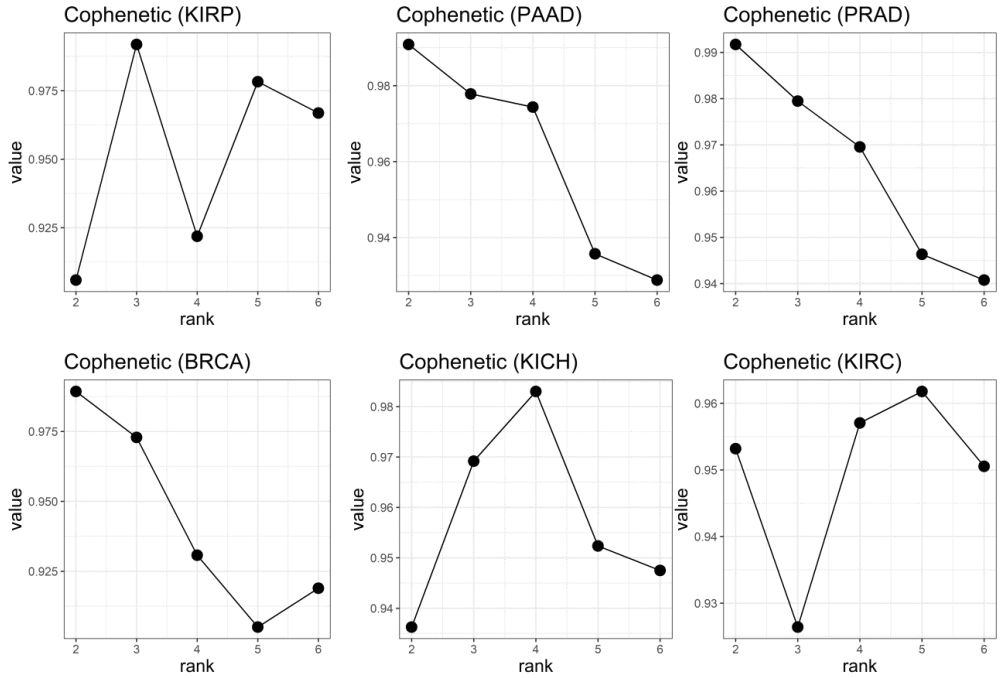


图 3-5 同表象相关系数 r =2-6

Fig 3-5 Cophenetic correlation coefficient for r = 2–6 is shown.

表 3-4 六种癌症 NMF 分型结果及特征基因数统计

Fig 3-3 The results of six cancer NMF types and the number of feature genes count

癌症项目	亚型数	亚型及特征基因数
BRCA	3	Cluster1（n=13）Cluster2（n=15）Cluster3（n=21）
PAAD	3	Cluster1（n=2）Cluster2（n=27）Cluster3（n=5）
PRAD	3	Cluster1（n=3）Cluster2（n=1）Cluster3（n=1）
KIRC	2	Cluster1（n=38）Cluster2（n=5）
KICH	4	Cluster1（n=10）Cluster2（n=2）
		Cluster3（n=4）Cluster4（5）
KIRP	3	Cluster1（n=6）Cluster2（n=3）Cluster3（n=3）

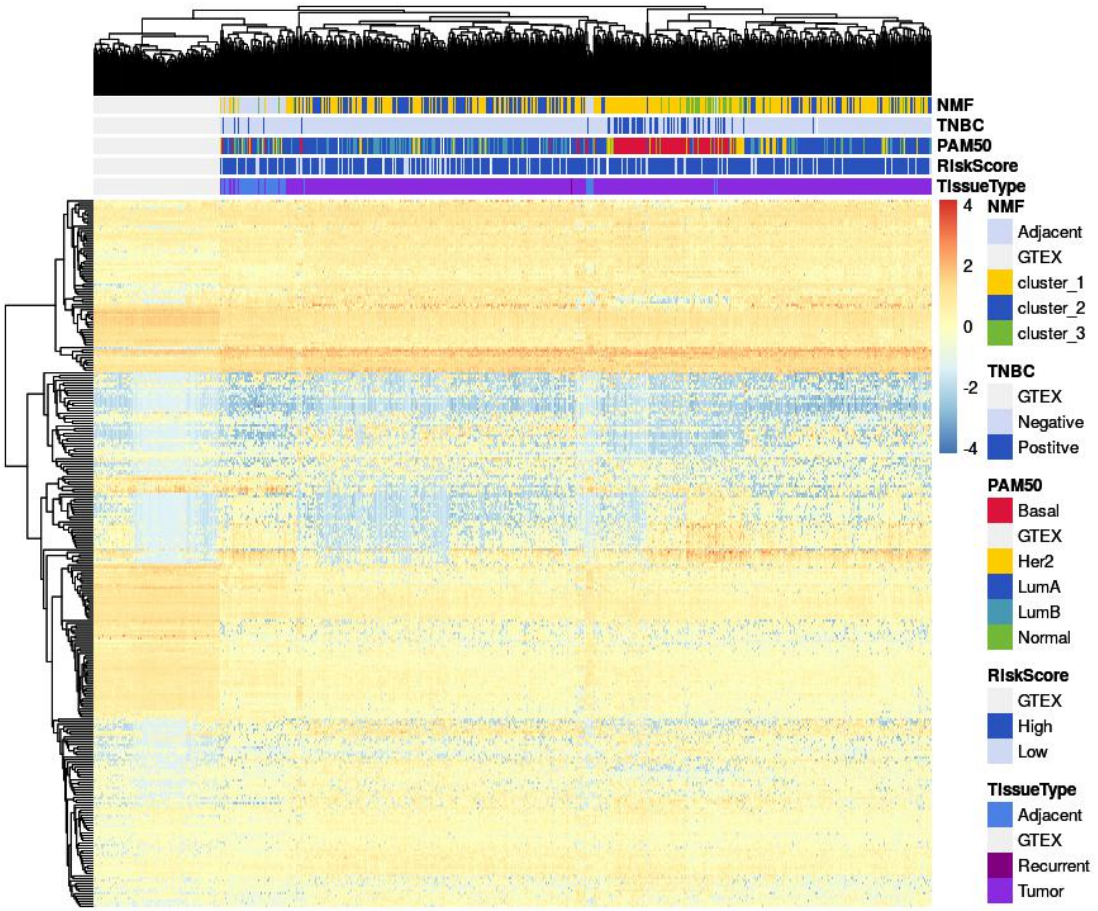


图 3-6 比较 BRCA NMF 亚型和其他肿瘤亚型。NMF 为本研究中基于假基因表达水平进行的 NMF 分型，TNBC 即三阴乳腺癌，PAM50 分型系统，RiskScore 为本研究构建的生存风险评分。TissueType 标记了 GTEX 样本、癌旁样本、复发性肿瘤和原发性肿瘤。

Fig 3-6 Comparison of pseudogene expression subtypes with other tumor subtypes. NMF is the NMF typing based on the expression level of pseudogenes in this study, TNBC, namely three-negative breast cancer, PAM50 typing system, and RiskScore is the survival RiskScore constructed in this study.TissueType labeled GTEX samples, paracancer samples, recurrent tumors, and primary tumors.

四种癌症 NMF 分型均有一个分型的出现过半数特征基因为免疫球蛋白假基因，即失活的免疫球蛋白假基因：KIRC Cluster2（3/5）、BRCA Cluster3（19/21）、PAAD Cluster3（3/5）、PRAD Cluster1（2/3）。ENSG00000254029（IGLC4）和 ENSG00000253998（IGKV2-29）作为特征基因存在在在这四种癌症的 Cluster 中。这些特征基因在分型中的聚集暗示这些分型中中的肿瘤样本在免疫调节方面可能与其它 cluster 有差异。

为了了解不同 NMF 分型下肿瘤微环境的差异，从 TIMER 2.0 数据库下载 TCGA 样本 B 细胞, CD4+ T 细胞, CD8+ T 细胞, 中性粒细胞, 巨噬细胞和树突细胞 TIMER 得分进行分析。图 3-7 为在各个癌症中，在不同分型之间进行两两比较，检验免疫细胞浸润水平是否有差异（wilcox.test，$P<0.05$）。

在 KIRC 中 Cluster2 在 T 细胞、巨噬细胞、树突细胞均与 Cluster1 有差异

在 BRCA 3 种 NMF 分型中除巨噬细胞外免疫细胞浸润水平具有显著差异。相较于其它两组分型，Cluster3 的肿瘤细胞免疫微环境中 B 细胞, CD4+ T 细胞, CD8+ T 细胞, 中性粒细胞,和树突细胞浸润程度更高。暗示失活的免疫球蛋白假基因可能以某种方式影响细胞免疫。图 3-7 还展示了其他 5 种癌症 NMF 分型之间存在不同程度的免疫微环境差异。

在 PAAD 中 Cluster3 六种免疫细胞浸润得分与 Cluster1 有不同程度的差异。

在 PRAD 中 Cluster1 的 CD4 T 细胞和巨噬细胞浸润性得分显著低于 Cluster3 和 Cluster3，而在 CD8 细胞上浸润性得分均低于 Cluster3 和 Cluster2.

综合上述结果，这四种癌症种聚集了免疫球蛋白假基因的 Cluster 在六种免疫细胞 TIMER 免疫浸润性得分的比较中，总是差异比较多的 Cluster。结果提示这些亚型在免疫调节过程中免疫球蛋白假基因可能发挥着特殊的功能。

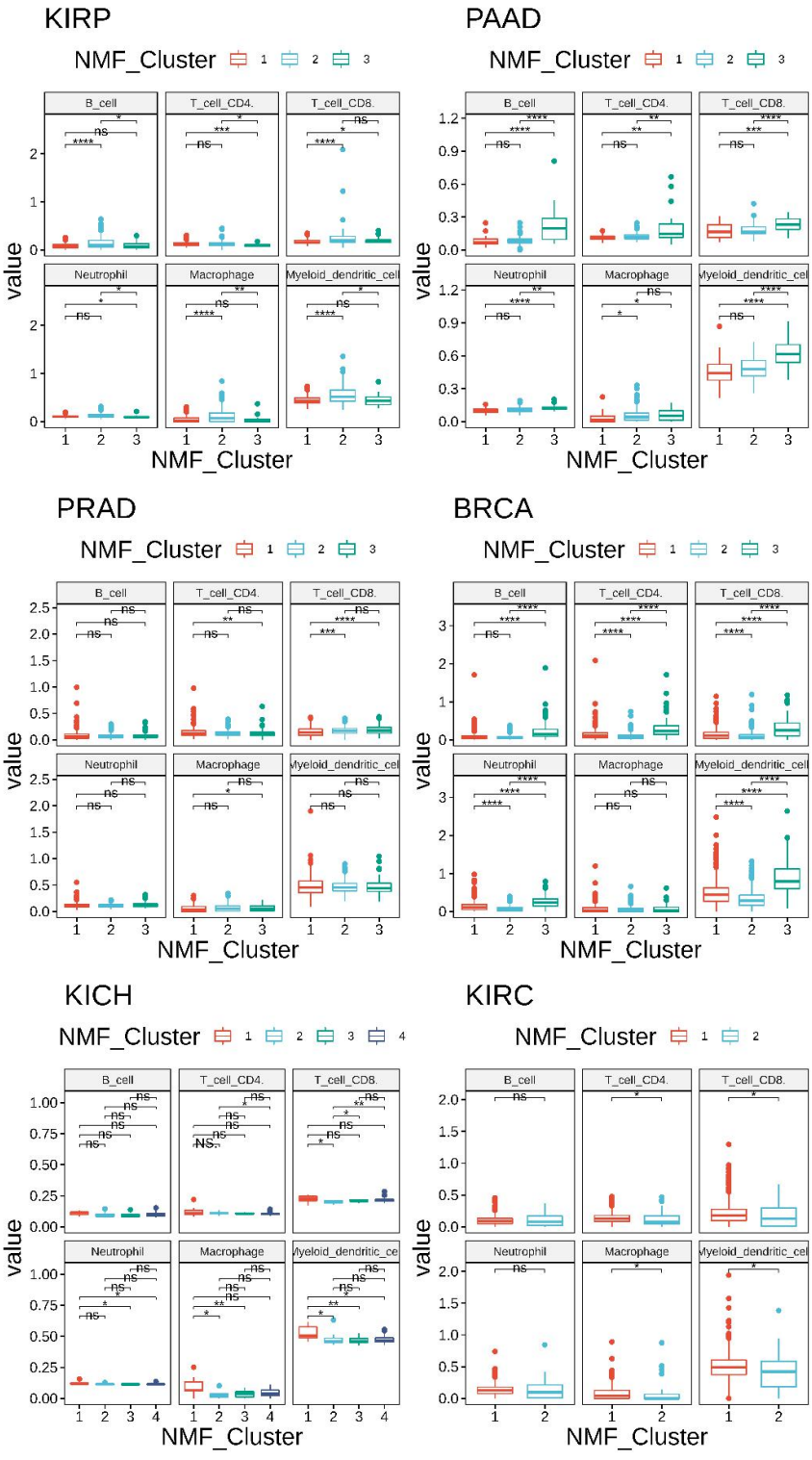


图 3-7 NMF 分型免疫浸润得分

Fig 3-7 NMF type immune infiltration score

3.6 构建生存风险评分模型

为了描述假基因对病人预后的影响，本文使用两步法对各个癌症构建一个预后模型，并根据模型的系数和特征基因的表达水平计算生存风险评分用于预测样本的生存风险。通过建立有效的预后模型来说明模型中的特征假基因在病人生存中具有影响。在本研究中首先使用单因素 COX 回归分析筛选与生存相关的基因，然后使用 LASSO Cox 回归建立预后模型。基于模型中的基因的表达水平及其系数，计算生存风险评分。

使用交叉验证的方法在各癌症 LASSO 回归模型中进行变量选择如图（图 3-8）。在本研究中选择交叉验证过程中模型误差最小处 lamda 值。

使用多因素 COX 回归建立预后模型，根据模型绘制森林图（图 3-9）。各因子的危险 HR、置信区间、模型的 AIC 值和模型一致性系数（concordance index）。模型的一致性系数 C-index 反映了模型的预测能力。C-index 等于 0.5 认为完全随机，0.5-0.7 为低、0.7-0.9 为中等、0.9 以上为高，1 为完全一致。PRAD C-index 高达 0.94、KICH 为 1 。KIRC C-index 为 0.69，预测能力低。其余预测能力为中等（附图 1-4）。

根据公式（2-2）计算生存风险评分，绘制 ROC 曲线（图 3-10）。ROC 曲线用于衡量预测模型的准确性。ROC 曲线面积越大、ROC 曲线越接近左上角，表示预测效果越好。一般认为 AUC>0.7 为预测准确性较好。图 3-11 反映 KICH 和 PRAD 均具有良好的预测效果。

选择 AUC 最大的时间点 ROC 曲线进行生存分析。使用该 ROC 曲线转折点作为风险评分的阈值，分为高风险组和低风险组。然后进行生存分析。生存分析结果如图 3-11。KICH、PRAD、BRCA、KIRP、KIRC 均成功通过生存风险评分区分出高低生存组（$P<0.05$）。

在这部分结果中 C-index、ROC 曲线、生存分析等结果均能说明 KICH 和 PRAD 两种癌症构建的预后模型具有较好的预测能力。两个模型的特征假基因如森林图所示。这些特征假基因在癌症中可能有重要的作用。

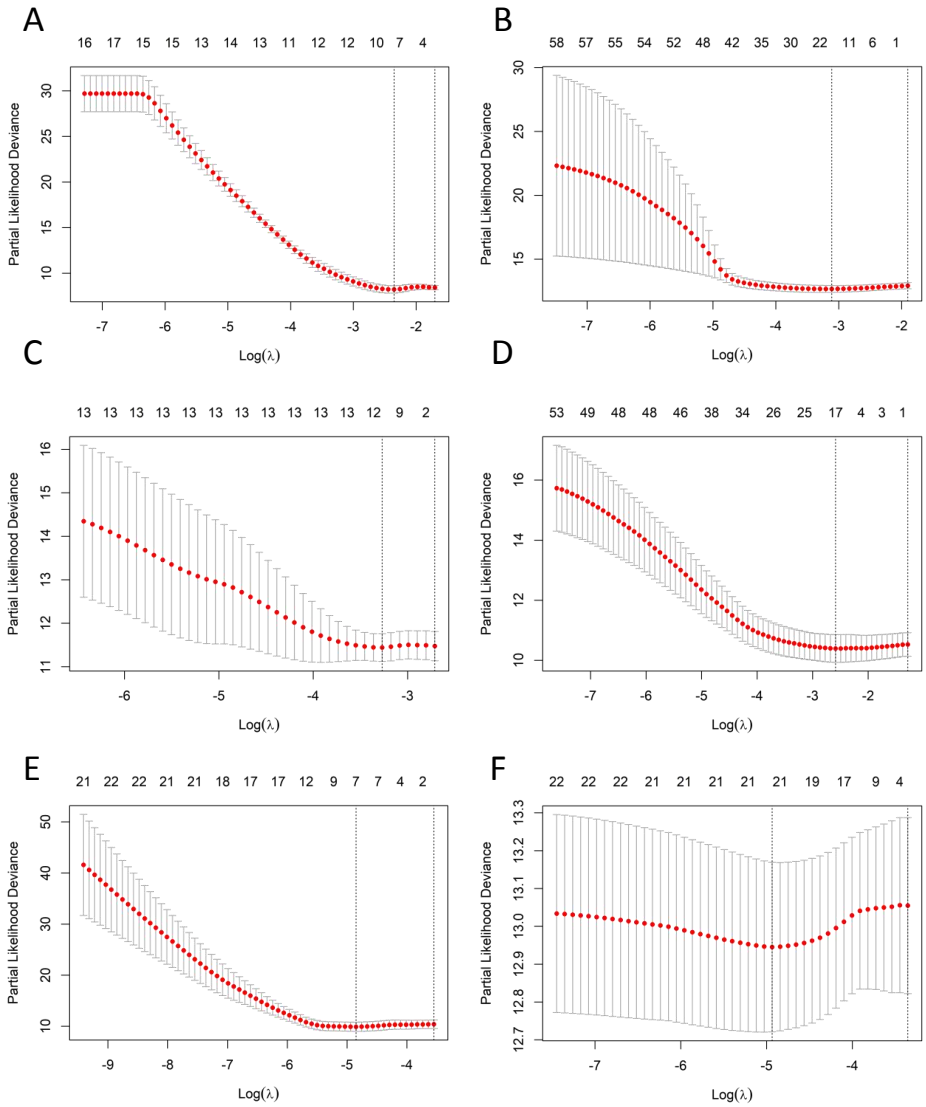


图 3-8 各癌症 LASSO 回归模型中基于交叉验证进行变量选择（A）KICH（B）KIRC（C）KIRP（D）PAAD（E）PRAD（F）BRCA。横轴是 lambda 值的对数，纵轴是模型误差。

Fig 3-8 In each cancer LASSO regression model, variables were selected based on cross-validation (A) KICH (B) KIRC (C) KIRP (D) PAAD (E) PRAD (F) BRCA.The horizontal axis is the logarithm of the lambda value, and the vertical axis is the model error

在 KICH 预后模型的特征基因中 ENSG00000214433 （GOLGA2P8）是一种转录未加工假基因（Transcribed unprocessed pseudogene）。在本文构建的 KICH ceRNA 网络中显示 GOLGA2P8 与 25 种蛋白编码基因竞争 3 种 miRNA，分别是 hsa-miR-214-3p、hsa-miR-455-3p、hsa-miR-199a-5p。

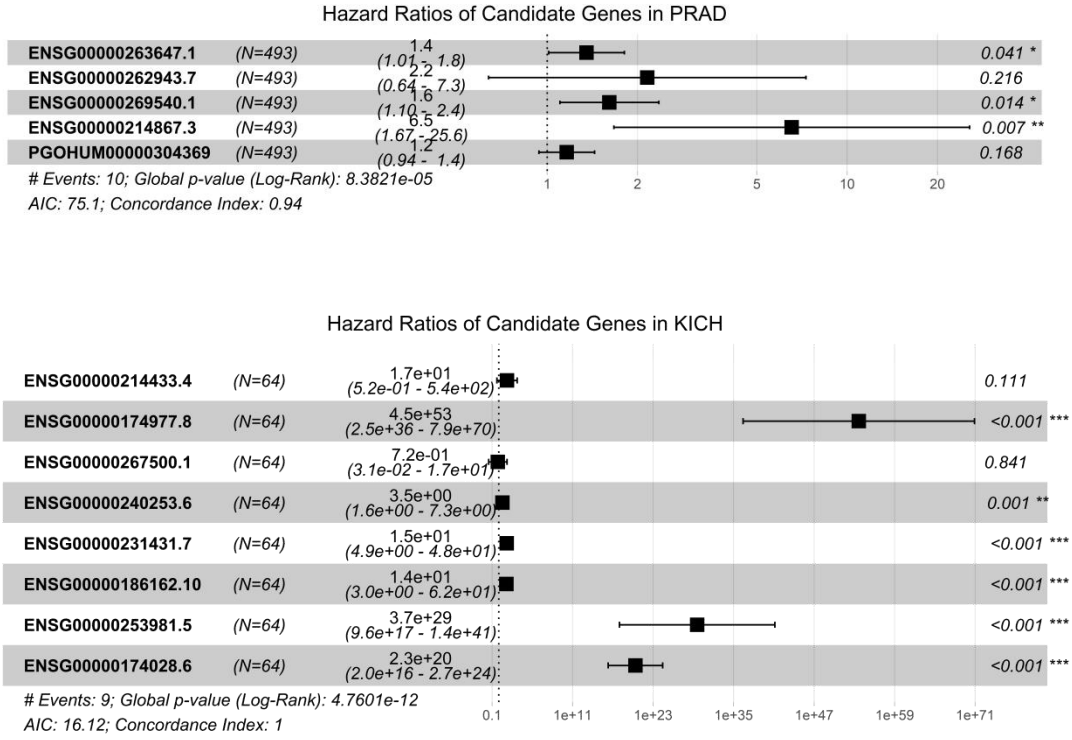


图 3-9 多变量 COX 回归分析森林图

Fig 3-9 Forest maps multivariate COX regression analysis

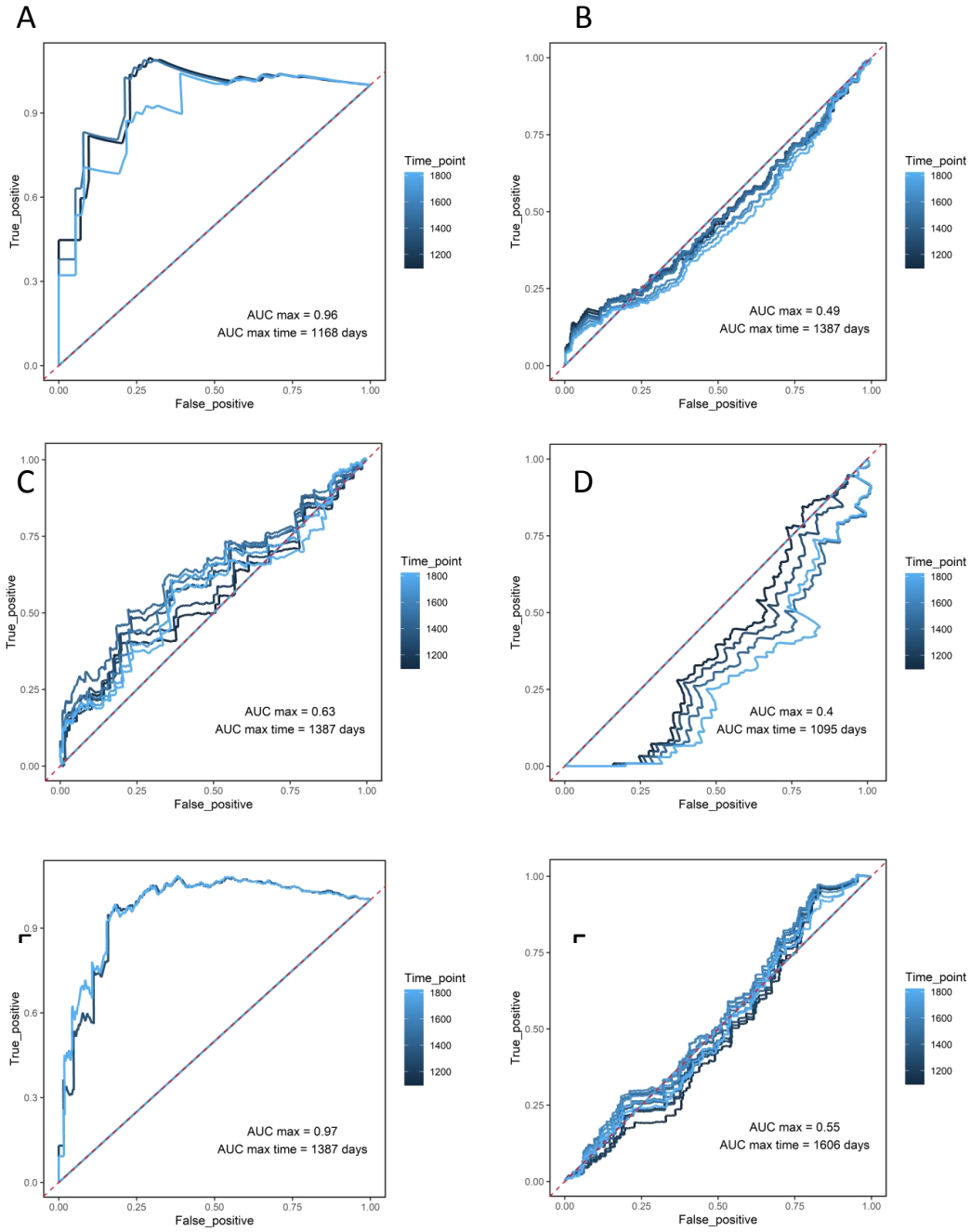


图 3-10 ROC 曲线（A）KICH（B）KIRC（C）KIRP（D）PAAD（E）PRAD（F）BRCA

Fig 3-10 ROC curves（A）KICH（B）KIRC（C）KIRP（D）PAAD（E）PRAD（F）BRCA

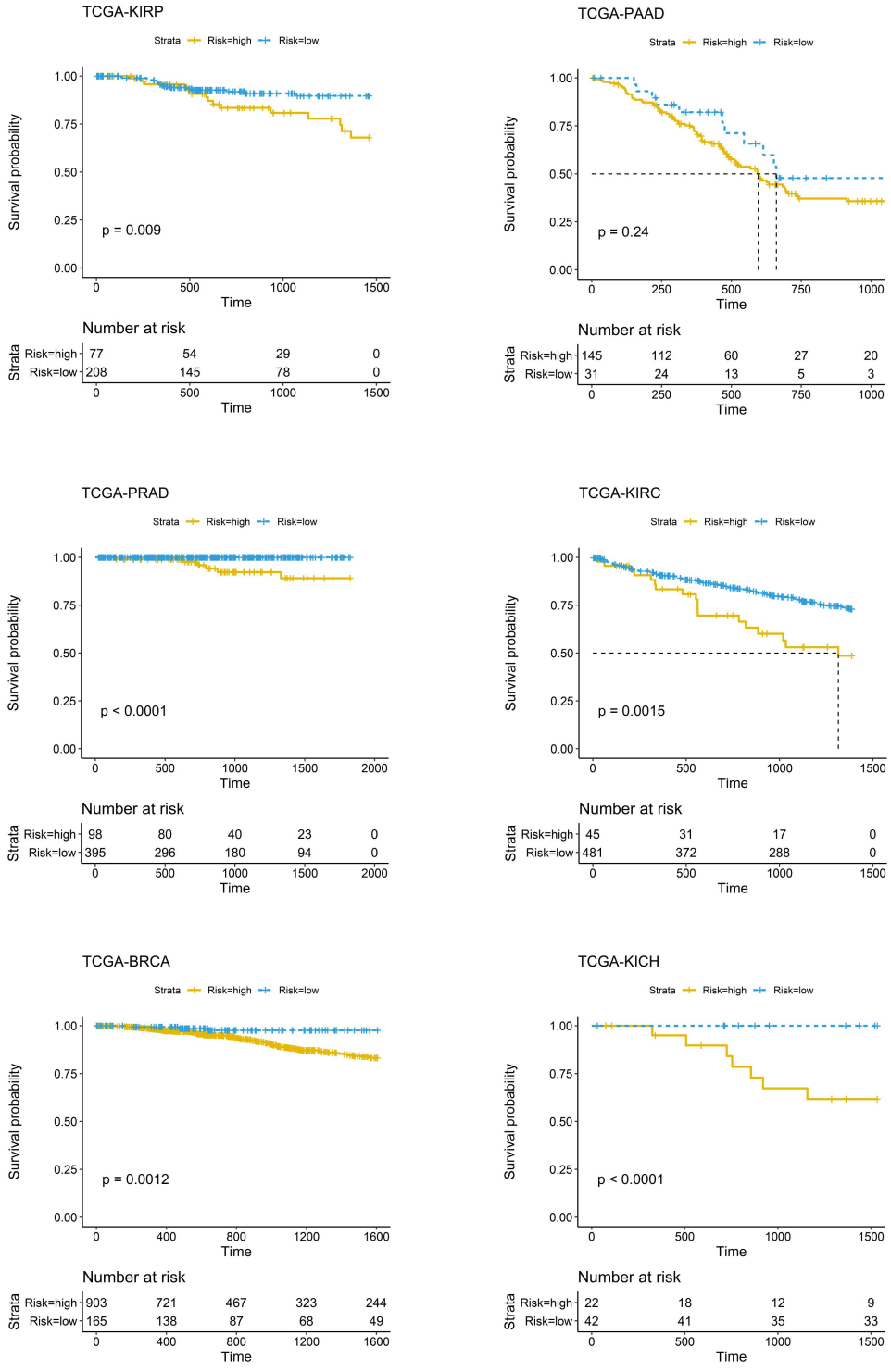


图 3-11 使用生存风险评分构建 Kaplan-Meier 生存曲线

Fig 3-11 The Kaplan-Meier survival curve was constructed using the survival risk score

总结与展望

4.1 总结

假基因在正常组织和肿瘤中广泛表达，但是具体的功能尚不清晰。本文使用生物信息学的方法分析了六种癌症假基因表达谱，并基于差异表达假基因构建ceRNA 网络、进行肿瘤分型、构建预后模型。试图通过上述分析理解假基因在肿瘤 RNA 调控中发挥的功能，找到在肿瘤中发挥重要作用的假基因。

目前对于假基因作为 ceRNA 的研究关注重点主要是假基因对其亲本功能基因的研究。如 PTENP1/PTEN、Foxo3P/Foxo3P。对于假基因对非亲本功能基因的讨论较少。本文筛选在基因组中高度特异的假基因构建注释文件。试图解析这部分在基因组缺乏相似基因的假基因在转录组水平扮演的角色。为了提高本研究的准确性，本文结合了 TCGA 数据库和 GTEX 数据库的测序数据进行分析。把 TCGA 原发性肿瘤样本和继发性肿瘤样本作为肿瘤组、把 TCGA 项目癌旁样本和对应的 GTEX 正常组织样本作为正常组。通过纳入 GTEx 数据库数据作为正常样本，弥补了 TCGA 数据库正常样本的样本量不足的缺点。本研究基于上述的假基因注释文件和测序数据，分析六种癌症的可检测假基因表达谱。本文从三个角度解释假基因参与的调控功能、筛选具有重要调控的假基因。

首先，通过构建假基因 ceRNA 网络研究假基因与蛋白编码基因的调控关系。尽管本研究针对的这部分假基因在基因组上缺少相似基因，但是仍然可以构建出包含许多节点的 ceRNA 网络（表 3-2）。不少假基因可以通过 miRNA 同时调控多个蛋白编码基因。例如在乳腺癌的 ceRNA 网络（图 3-2）上可以看到 ZNF252P 和 ZNF623 共同竞争了 10 个 miRNA，但是通过 blast 比对这两种 RNA 分子相似度较低。对 ZNF252P 和 ZNF623 根据表达水平中位数对样本分组进行生存分析，发现这两种基因的高表达的样本具有更高的死亡率。

为了系统讨论 ceRNA 网络的生物学意义，本研究使用 GO 富集分析和生存分析评价 ceRNA 网络的功能。但是假基因缺少功能注释，因此在这里我们对 ceRNA 网络中 mRNA 进行 GO 富集分析定义 ceRNA 网络的功能。结果显示各

个癌症 ceRNA 网络中的 mRNA 能富集到不同的功能，推测假基因可能通过竞争性结合 miRNA 调控这些功能基因的表达。分析 ceRNA 网络中假基因在高表达与低表达样品之间生存状况，讨论假基因节点的表达对病人生存的影响。在 KICH 样本中，20 个假基因节点有 9 个假基因与病人生存显著相关。这 9 个假基因在 ceRNA 网络中调控了 142 个基因。以上结果为假基因对非相似的基因的调控作用提供了更多的证据。

然后使用假基因表达谱对肿瘤进行分型。肿瘤的发生与发展机制非常复杂，同一种不同亚型之间往往具有不同的生物学过程。对肿瘤进行分型可以为肿瘤的个性化治疗提供更加精确的分子信息，指导制定个性化治疗方案和进行疗效监测。但是由于肿瘤的复杂性，使用单一分子并不能准确分型。目前多种组学的发展已经为肿瘤的分型提供了许多分子数据，包括 mRNA 表达水平、miRNA 表达水平、DNA 甲基化、体细胞突变、蛋白质表达水平和假基因表达水平。在本文中试图使用综合了 TCGA 和 GTEX 测序数据得到的差异表达假基因的表达谱进行分型，并分析不同的癌症样本。在本研究中成功地使用非负矩阵分解法对六种癌症进行分型。KIRC 和 BRCA 分型结果与前人研究相符。有趣的是四种癌症 KIRC、BRCA 、PAAD 、PRAD 中均可以获得一个特征基因聚集了免疫球蛋白假基因的分型。这些聚集了免疫球蛋白假基因分型的特征基因中都存在 IGLC4 和 IGKV2-2 两种免疫球蛋白假基因。分析肿瘤样本六种免疫细胞浸润性得分，发现这些聚集了免疫球蛋白假基因的分型免疫细胞浸润程度与其它分型差异较大。这些结果提示了 IGLC4 和 IGKV2-2 免疫球蛋白假基因可能并没有失活，并且对免疫系统有一定的调节功能。

最后构建预后模型。通过预后模型获得影响病人生存的特征假基因，并使用假基因及其模型的系数计算生存风险评分用于预测病人预后。在本研究中对六种癌症进行 LASSO-COX 多因素回归建模。根据 C-index 评分、ROC 曲线结果评价模型的预测能力，认为成功构建了 KICH 和 PRAD 两种癌症的假基因表达水平预后模型。这些特征假基因对癌症的发生和发展可能有重要的调节作用。如 KICH 模型中的特征基因 GOLGA2P8 在 ceRNA 网络中显示与 25 种蛋白编码基因竞争 3 种 miRNA。

通过以上的分析本研究构建了六种癌症的 ceRNA 网络，讨论了假基因通过

ceRNA 网络对非相似的蛋白编码基因进行调控；基于假基因差异表达基因的表达水平进行了肿瘤分型及构建预后模型，筛选到与肿瘤亚型以及预后相关的特征假基因。

本研究可以为假基因的研究提供更多证据，作为未来深入理解假基因在癌症种形式的的功能提供理论基础，帮助开发新的生物标记和提高对肿瘤生物学的认识。

4.2 展望

本研究仍有许多不完备的地方。

在 ceRNA 网络构建方面。本文简单粗暴地把 miRNA 在不同 RNA 上结合位点视为完全相同的结合位点。然而事实上同一个 miRNA 在不同分子的结合位点序列可能会有细微的差异[9]。而这些细微的差异放在细胞 ceRNA 调控网络可能会带来显著的变化。也有研究表明尽管 miRNA 可以与多个 RNA 结合，但是在实际的结合中 miRNA 对 RNA 的选择有主次之分[46]。因此未来可以进一步完善这部分的工作。

TCGA 数据库提供了 33 种癌症 RNA 测序数据。然而本文只研究了其中六种，未来可以继续补充其他癌症假基因分析。

参考文献

[1] Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465(7301):P.1033-8.

[2] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nature Reviews Genetics. 2019;21(3):1-11.

[3] Mccarrey JR, Thomas K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature. 1987;326(6112):501-5.

[4] McCarrey, John R. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. Gene. 1987;61(3):291-8.

[5] Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, et al. Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. Cell. 2018;173(6):1370-84.e16.

[6] Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, et al. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell. 2018;173(6):1356-69.e22.

[7] Zhou BS, Beidler DR, Cheng YC. Identification of Antisense RNA Transcripts from a Human DNA Topoisomerase I Pseudogene. Cancer Research. 1992;52(15):4280.

[8] Chan WL, Chung-Yee Y, Yang WK, Shih-Ya H, Chang YS, Chien-Chih C, et al. Transcribed pseudogene ψPPM1K generates endogenous siRNA to suppress oncogenic cell growth in hepatocellular carcinoma. Nucleic Acids Research. 2013(6):3734-47.

[9] Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi Pier P. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? Cell. 2011.

[10] Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. Nature Reviews Genetics. 2016.

[11] Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. Nat Rev Mol Cell Biol. 2019;20(1):21-37.

[12] Ana K, Maria B, Sam GJ. miRBase: from microRNA sequences to function. Nucleic Acids Research. 2018(D1):D1.

[13] Daan, Swarts, Kira, Makarova, Yanli, Wang, et al. The evolutionary journey of Argonaute proteins. Nature Structural & Molecular Biology. 2014.

[14] Stefanie, Jonas, Elisa, Izaurralde. Towards a molecular understanding of microRNA-mediated gene silencing. Nature Reviews Genetics. 2015.

[15] Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014;42(Database issue):D92-7.

[16] The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell. 2015;161(2):319-32.

[17] Yang W, Du WW, Li X, Yee AJ, Yang BB. Foxo3 activity promoted by non-coding effects of circular RNA and Foxo3 pseudogene in the inhibition of tumor growth and angiogenesis. Oncogene. 2016;35(30):3919-31.

[18] Jia CJ, Hao KZ, Hong CX, Bin Z, Chao L, Wah ST, et al. A FTH1 gene:pseudogene:microRNA network regulates tumorigenesis in prostate cancer. Nuclc Acids Research. 2018(4):4.

[19] Kerwin J, Khan I, Iorns E, Tsui R, Denis A, Perfito N, et al. Replication Study: A coding-independent

function of gene and pseudogene mRNAs regulates tumour biology. Elife. 2020;9.
[20] Gan, Yu, Weimin, Yao, Kiranmai, Gumireddy, et al. Pseudogene PTENP1 functions as a competing endogenous RNA to suppress clear-cell renal cell carcinoma progression. Molecular Cancer Therapeutics. 2014.
[21] Ling, Gao, Wenhao, Ren, Linmei, Zhang, et al. PTENp1, a natural sponge of miR-21, mediates PTEN expression to inhibit the proliferation of oral squamous cell carcinoma. Molecular Carcinogenesis. 2017.
[22] Zhang R, Guo Y, Ma Z, Ma G, Xue Q, Li F, et al. Long non-coding RNA PTENP1 functions as a ceRNA to modulate PTEN level by decoying miR-106b and miR-93 in gastric cancer. Oncotarget. 2017.
[23] Wei Y, Chang Z, Wu C, Zhu Y, Xu Y. Identification of potential cancer-related pseudogenes in lung adenocarcinoma based on ceRNA hypothesis. Oncotarget. 2017;8(35):59036-47.
[24] Liu B, Liu J, Liu K, Huang H, Cheng Q. A prognostic signature of five pseudogenes for predicting lower-grade gliomas. Biomedicine & Pharmacotherapy. 2019;117:109116-.
[25] Grzechowiak I, Gra J, Szymańska D, Biernacka M, Kolenda T. The Oncogenic Roles of PTTG1 and PTTG2 Genes and Pseudogene PTTG3P in Head and Neck Squamous Cell Carcinomas. Diagnostics. 2020.
[26] Welch JD, Baran-Gale J, Perou CM, Sethupathy P, Prins JF. Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. BMC Genomics. 2015;16(1):113.
[27] Zefang T, Chenwei L, Boxi K, Ge G, Cheng L, Zemin Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nuclc Acids Research. 2017(W1):W1.
[28] Zheng LL, Zhou KR, Liu S, Zhang DY, Wang ZL, Chen ZR, et al. dreamBase: DNA modification, RNA regulation and protein binding of expressed pseudogenes in human health and disease. Nucleic Acids Res. 2018;46(D1):D85-D91.
[29] Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, et al. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nat Commun. 2014;5:3963.
[30] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.
[31] Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015;163(2):506-19.
[32] Creighton CJ, Morgan M, Gunaratne PH, Wheeler DA, Gibbs RA, Gordon Robertson A, et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43-9.
[33] Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med. 2016;374(2):135-45.
[34] Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014;26(3):319-30.
[35] Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell. 2017;32(2):185-203.e13.
[36] The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163(4):1011-25.
[37] Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank. 2015;13(5):307-8.
[38] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-74.
[39] Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. GenMap: ultra-fast computation of genome

mappability. Bioinformatics. 2020;36(12):3687-92.
[40] Zhu X, Tian X, Sun T, Yu C, Cao Y, Yan T, et al. GeneExpressScore Signature: a robust prognostic and predictive classifier in gastric cancer. Mol Oncol. 2018;12(11):1871-83.
[41] Liu F, Liao Z, Song J, Yuan C, Liu Y, Zhang H, et al. Genome-wide screening diagnostic biomarkers and the construction of prognostic model of hepatocellular carcinoma. J Cell Biochem. 2020;121(3):2582-94.
[42] Huang R, Meng T, Chen R, Yan P, Zhang J, Hu P, et al. The construction and analysis of tumor-infiltrating immune cell and ceRNA networks in recurrent soft tissue sarcoma. Aging (Albany NY). 2019;11(22):10116-43.
[43] Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. Biom J. 2010;52(1):70-84.
[44] Brunet, Jean-Philippe, Tamayo, Pablo, Golub, Todd R, et al. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America. 2004.
[45] Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;531(7592):47-52.
[46] Seitz H. Redefining microRNA targets. Curr Biol. 2009;19(10):870-3.

附录

附表 3-1 BRCA NMF 亚型特征假基因

Appendix table 3-1 BRCA NMF subtype features pseudogenes

NMF_Cluster	Gene_Type	Gene_Symbol
Cluster1	transcribed_unprocessed_pseudogene	SULT1C2P1
Cluster1	transcribed_unprocessed_pseudogene	BPIFA4P
Cluster1	unprocessed_pseudogene	AC245100.3
Cluster1	processed_pseudogene	AC016885.3
Cluster1	unprocessed_pseudogene	AC108729.3
Cluster1	transcribed_unprocessed_pseudogene	KRT223P
Cluster1	transcribed_processed_pseudogene	TPRXL
Cluster1	processed_pseudogene	SCML2P1
Cluster1	unprocessed_pseudogene	BX119927.1
Cluster1	unprocessed_pseudogene	GLYATL1P1
Cluster1	processed_pseudogene	AL450996.1
Cluster1	transcribed_unitary_pseudogene	WFDC21P
Cluster1	unprocessed_pseudogene	AC005682.1
Cluster2	unprocessed_pseudogene	CYP2A7P2
Cluster2	unprocessed_pseudogene	MED15P4
Cluster2	transcribed_unitary_pseudogene	CYP2T1P
Cluster2	processed_pseudogene	AC025423.3
Cluster2	transcribed_processed_pseudogene	HNRNPA1P57
Cluster2	processed_pseudogene	AC107890.1
Cluster2	unprocessed_pseudogene	CYP2T3P
Cluster2	transcribed_processed_pseudogene	ACTG1P22
Cluster2	transcribed_unprocessed_pseudogene	SLC25A24P1
Cluster2	unprocessed_pseudogene	SLC25A24P2
Cluster2	processed_pseudogene	AC073869.6
Cluster2	unprocessed_pseudogene	AC073508.1
Cluster2	transcribed_processed_pseudogene	AC067942.1
Cluster2	transcribed_unprocessed_pseudogene	CYP2G1P
Cluster2	unprocessed_pseudogene	KRT41P
Cluster3	IG_V_pseudogene	IGKV1D-27
Cluster3	IG_C_pseudogene	IGLC4
Cluster3	transcribed_unitary_pseudogene	CSN1S2AP
Cluster3	IG_V_pseudogene	IGKV2-38
Cluster3	IG_V_pseudogene	IGKV2-29

Cluster3	IG_V_pseudogene	IGKV3-25
Cluster3	IG_V_pseudogene	IGHV1OR15-6
Cluster3	IG_V_pseudogene	IGLVI-70
Cluster3	IG_V_pseudogene	IGLV3-15
Cluster3	IG_V_pseudogene	IGKV2-10
Cluster3	IG_V_pseudogene	IGLV3-29
Cluster3	IG_V_pseudogene	IGKV1D-35
Cluster3	IG_V_pseudogene	IGKV2-36
Cluster3	IG_V_pseudogene	IGKV3-34
Cluster3	IG_V_pseudogene	IGKV2-23
Cluster3	IG_V_pseudogene	IGKV1-32
Cluster3	IG_V_pseudogene	IGLV2-28
Cluster3	IG_V_pseudogene	IGKV7-3
Cluster3	IG_V_pseudogene	IGLV3-13
Cluster3	unprocessed_pseudogene	AC245028.2
Cluster3	IG_V_pseudogene	IGKV1-35

附表 3-2 KIRP NMF 亚型特征假基因

Appendix table 3-2 KIRC NMF subtype features pseudogenes

NMF_Cluster	Gene_Type	Gene_Symbol
Cluster1	processed_pseudogene	BASP1P1
Cluster1	transcribed_unprocessed_pseudogene	MT1JP
Cluster1	transcribed_unprocessed_pseudogene	AL049767.1
Cluster1	processed_pseudogene	ATP5MC1P3
Cluster1	processed_pseudogene	CRB3P1
Cluster1	transcribed_unprocessed_pseudogene	CTSL3P
Cluster2	transcribed_processed_pseudogene	WTAPP1
Cluster2	transcribed_unprocessed_pseudogene	SAA3P
Cluster2	IG_V_pseudogene	IGLV2-28
Cluster3	transcribed_processed_pseudogene	AC015910.1
Cluster3	unprocessed_pseudogene	AC074101.1
Cluster3	processed_pseudogene	MTND4LP30

附表 3-3 KIRC NMF 亚型特征假基因

Appendix table 3-3 KIRC NMF subtype features pseudogenes

NMF_Cluster	Gene_Type	Gene_Symbol
Cluster1	transcribed_unprocessed_pseudogene	MST1L
Cluster1	processed_pseudogene	SRP14P4
Cluster1	processed_pseudogene	Z98751.2
Cluster1	transcribed_processed_pseudogene	AC098934.2
Cluster1	transcribed_unitary_pseudogene	BX571818.1
Cluster1	processed_pseudogene	AC103591.2
Cluster1	unprocessed_pseudogene	ZNF37CP
Cluster1	transcribed_unprocessed_pseudogene	CCNYL2
Cluster1	processed_pseudogene	AP000445.1
Cluster1	transcribed_unprocessed_pseudogene	OR7E47P
Cluster1	processed_pseudogene	BASP1P1
Cluster1	processed_pseudogene	AC051619.2
Cluster1	processed_pseudogene	HMGN2P47
Cluster1	transcribed_unprocessed_pseudogene	CSPG4P13
Cluster1	processed_pseudogene	AC087468.1
Cluster1	transcribed_unprocessed_pseudogene	SLC47A1P2
Cluster1	transcribed_processed_pseudogene	LRRC37A7P
Cluster1	processed_pseudogene	AC017100.2
Cluster1	transcribed_unprocessed_pseudogene	ZNF826P
Cluster1	unprocessed_pseudogene	AC010487.2
Cluster1	unprocessed_pseudogene	AC018804.1
Cluster1	processed_pseudogene	AC015977.1
Cluster1	processed_pseudogene	MRPL36P1
Cluster1	transcribed_unitary_pseudogene	AC063952.1
Cluster1	transcribed_processed_pseudogene	RPL23AP49
Cluster1	transcribed_processed_pseudogene	MTHFD2P1
Cluster1	processed_pseudogene	ORAOV1P1
Cluster1	unprocessed_pseudogene	UGT2B27P
Cluster1	unprocessed_pseudogene	AC111000.2
Cluster1	processed_pseudogene	HMGN1P13
Cluster1	processed_pseudogene	MTND4LP30
Cluster1	transcribed_unprocessed_pseudogene	ZNF300P1
Cluster1	transcribed_processed_pseudogene	H2AC3P
Cluster1	transcribed_unprocessed_pseudogene	DPY19L2P2
Cluster1	unprocessed_pseudogene	AC005682.1

Cluster1	processed_pseudogene	TTC4P1
Cluster1	processed_pseudogene	AC073850.1
Cluster1	processed_pseudogene	AL008707.1
Cluster2	transcribed_processed_pseudogene	AC015910.1
Cluster2	IG_V_pseudogene	IGKV2-10
Cluster2	IG_V_pseudogene	IGKV2-29
Cluster2	unprocessed_pseudogene	AC245028.3
Cluster2	IG_C_pseudogene	IGLC4

附表 3-4 KICH NMF 亚型特征假基因

Appendix table 3-4 KICH NMF subtype features pseudogenes

NMF_Cluster	Gene_Type	Gene_Symbol
Cluster1	transcribed_unprocessed_pseudogene	MST1L
Cluster1	transcribed_unitary_pseudogene	BX571818.1
Cluster1	transcribed_processed_pseudogene	WTAPP1
Cluster1	transcribed_unprocessed_pseudogene	MT1JP
Cluster1	processed_pseudogene	AP001542.1
Cluster1	processed_pseudogene	SCML2P1
Cluster1	rRNA_pseudogene	RNA5SP107
Cluster1	IG_V_pseudogene	IGKV2-29
Cluster1	unprocessed_pseudogene	AC111000.2
Cluster1	transcribed_unprocessed_pseudogene	AL158819.1
Cluster2	transcribed_unprocessed_pseudogene	MT1L
Cluster2	processed_pseudogene	MTCO2P22
Cluster3	unprocessed_pseudogene	AL138900.2
Cluster3	processed_pseudogene	AC073465.1
Cluster3	transcribed_processed_pseudogene	RPL32P9
Cluster3	unprocessed_pseudogene	BX842568.4
Cluster4	transcribed_processed_pseudogene	LRRC37A6P
Cluster4	transcribed_processed_pseudogene	AL355376.2
Cluster4	unprocessed_pseudogene	AC108729.3
Cluster4	transcribed_unprocessed_pseudogene	AC005077.2
Cluster4	transcribed_processed_pseudogene	AL390726.4

附表 3-5 PAAD NMF 亚型特征假基因

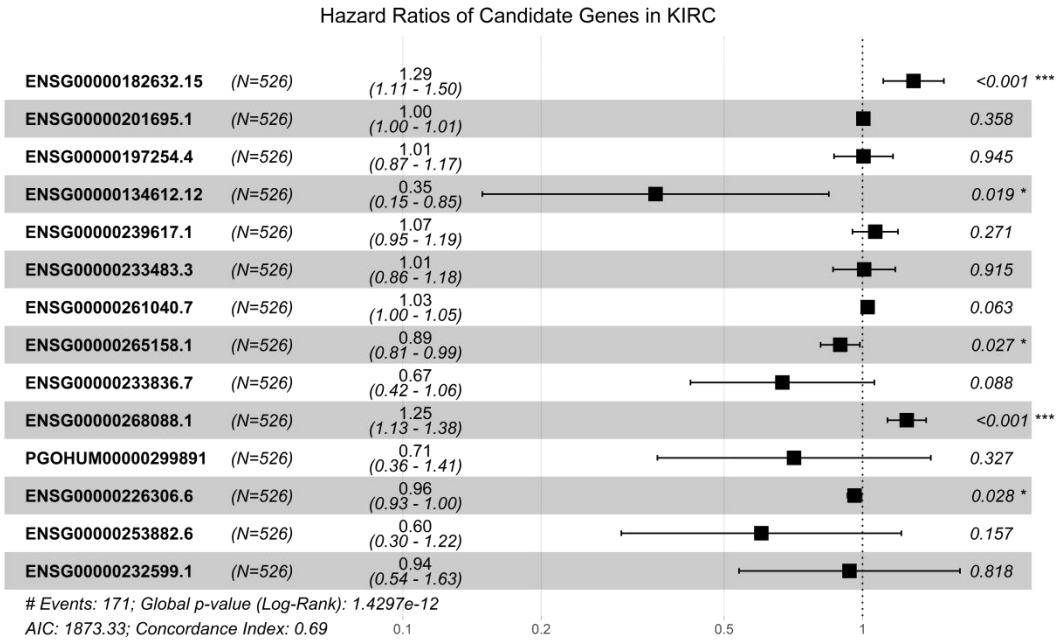
Appendix table 3-5 PAAD NMF subtype features pseudogenes

NMF_Cluster	Gene_Type	Gene_Symbol
Cluster1	polymorphic_pseudogene	PNLIPRP2
Cluster1	unprocessed_pseudogene	AL034369.1
Cluster2	processed_pseudogene	RPS15AP30
Cluster2	transcribed_unitary_pseudogene	SLC22A20P
Cluster2	processed_pseudogene	AC092851.1
Cluster2	processed_pseudogene	PPIAP45
Cluster2	transcribed_processed_pseudogene	AC124947.2
Cluster2	transcribed_unprocessed_pseudogene	LINC00431
Cluster2	transcribed_unitary_pseudogene	GUCY1B2
Cluster2	IG_V_pseudogene	IGHV4-80
Cluster2	processed_pseudogene	AC092868.2
Cluster2	transcribed_unprocessed_pseudogene	AC126323.1
Cluster2	unprocessed_pseudogene	TPSP2
Cluster2	unprocessed_pseudogene	AC145285.4
Cluster2	transcribed_unprocessed_pseudogene	KRT16P6
Cluster2	unprocessed_pseudogene	AC132812.1
Cluster2	processed_pseudogene	AP001542.1
Cluster2	processed_pseudogene	SCML2P1
Cluster2	unprocessed_pseudogene	IGFL1P1
Cluster2	transcribed_unitary_pseudogene	FER1L4
Cluster2	transcribed_unprocessed_pseudogene	AL049767.1
Cluster2	processed_pseudogene	AL121776.1
Cluster2	unprocessed_pseudogene	AC003072.1
Cluster2	transcribed_unitary_pseudogene	CPHL1P
Cluster2	processed_pseudogene	MTCO2P22
Cluster2	rRNA_pseudogene	RNA5SP236
Cluster2	transcribed_unitary_pseudogene	MYH16
Cluster2	transcribed_processed_pseudogene	AC103705.1
Cluster2	transcribed_processed_pseudogene	ZNF736P9Y
Cluster3	processed_pseudogene	AC098935.2
Cluster3	IG_V_pseudogene	IGKV3-25
Cluster3	IG_V_pseudogene	IGKV2-29
Cluster3	unprocessed_pseudogene	AC103563.3
Cluster3	IG_C_pseudogene	IGLC4

附表 3-6 PRAD NMF 亚型特征假基因

Appendix table 3-6 PRAD NMF subtype features pseudogenes

NMF_Cluster	Gene_Type	Gene_Symbol
Cluster1	IG_V_pseudogene	IGKV3-25
Cluster1	IG_C_pseudogene	IGLC4
Cluster1	processed_pseudogene	PPIAP29
Cluster2	processed_pseudogene	SCML2P1
Cluster3	unprocessed_pseudogene	FRG2JP



附图 3-1 多变量 COX 回归分析森林图（KIRC）

Appendix Fig 3-1 Forest maps multivariate COX regression analysis（KIRC）

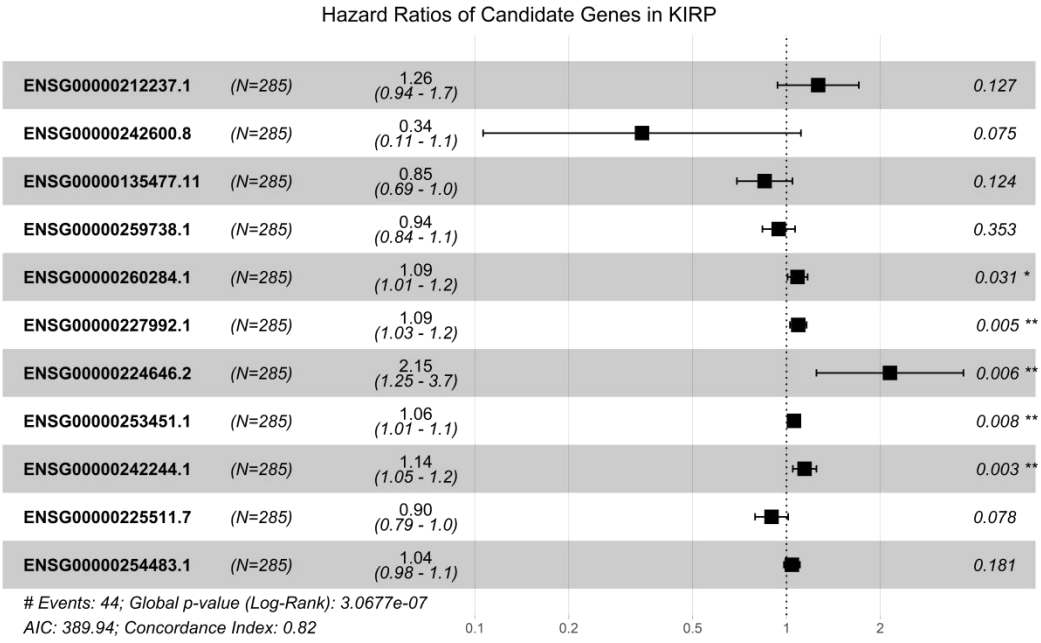


图 3-2 多变量 COX 回归分析森林图（KIRP）

Appendix Fig 3-2 Forest maps multivariate COX regression analysis（KIRP）

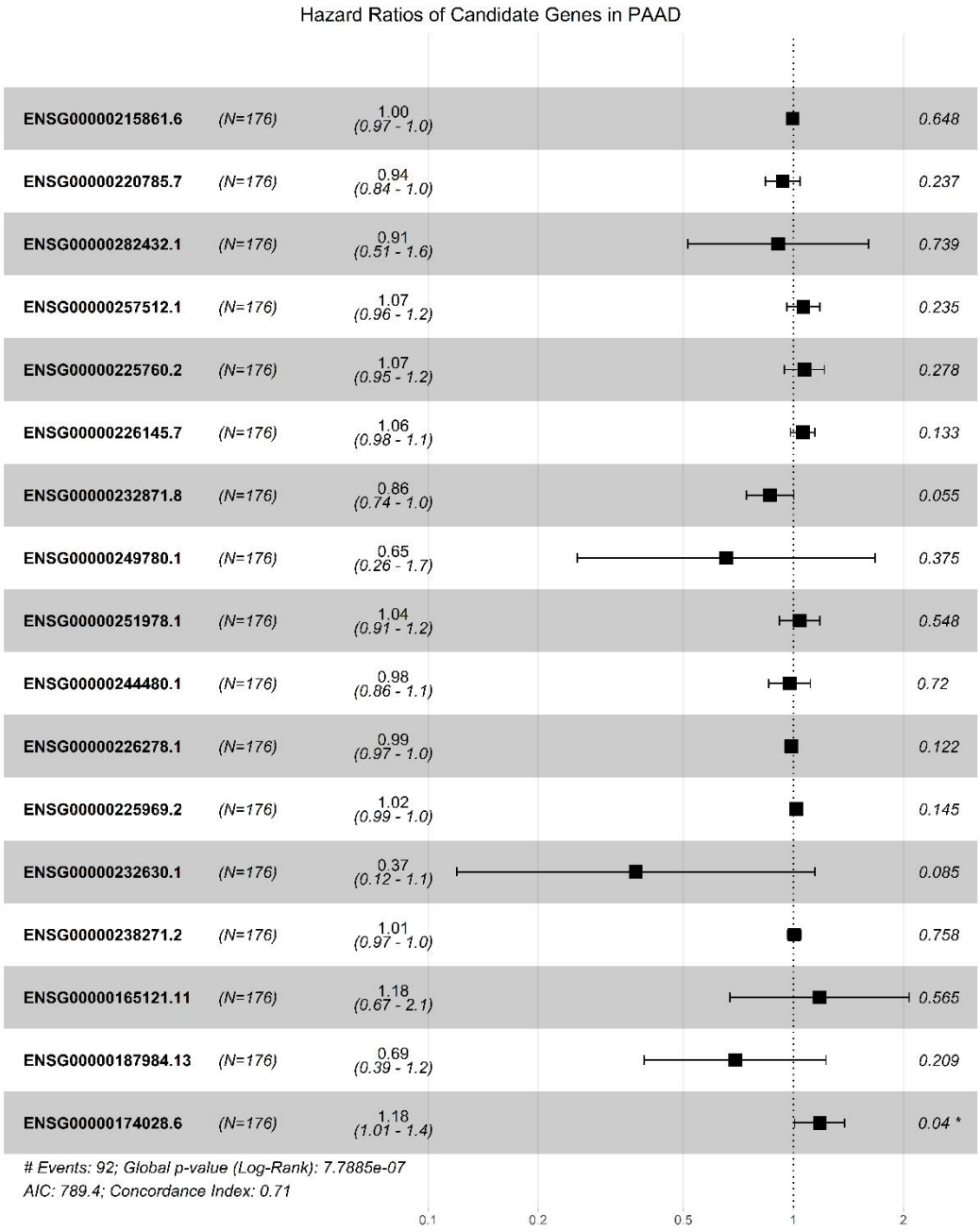


图 3-3 多变量 COX 回归分析森林图（PAAD）

Appendix Fig 3-3 Forest maps multivariate COX regression analysis（PAAD）

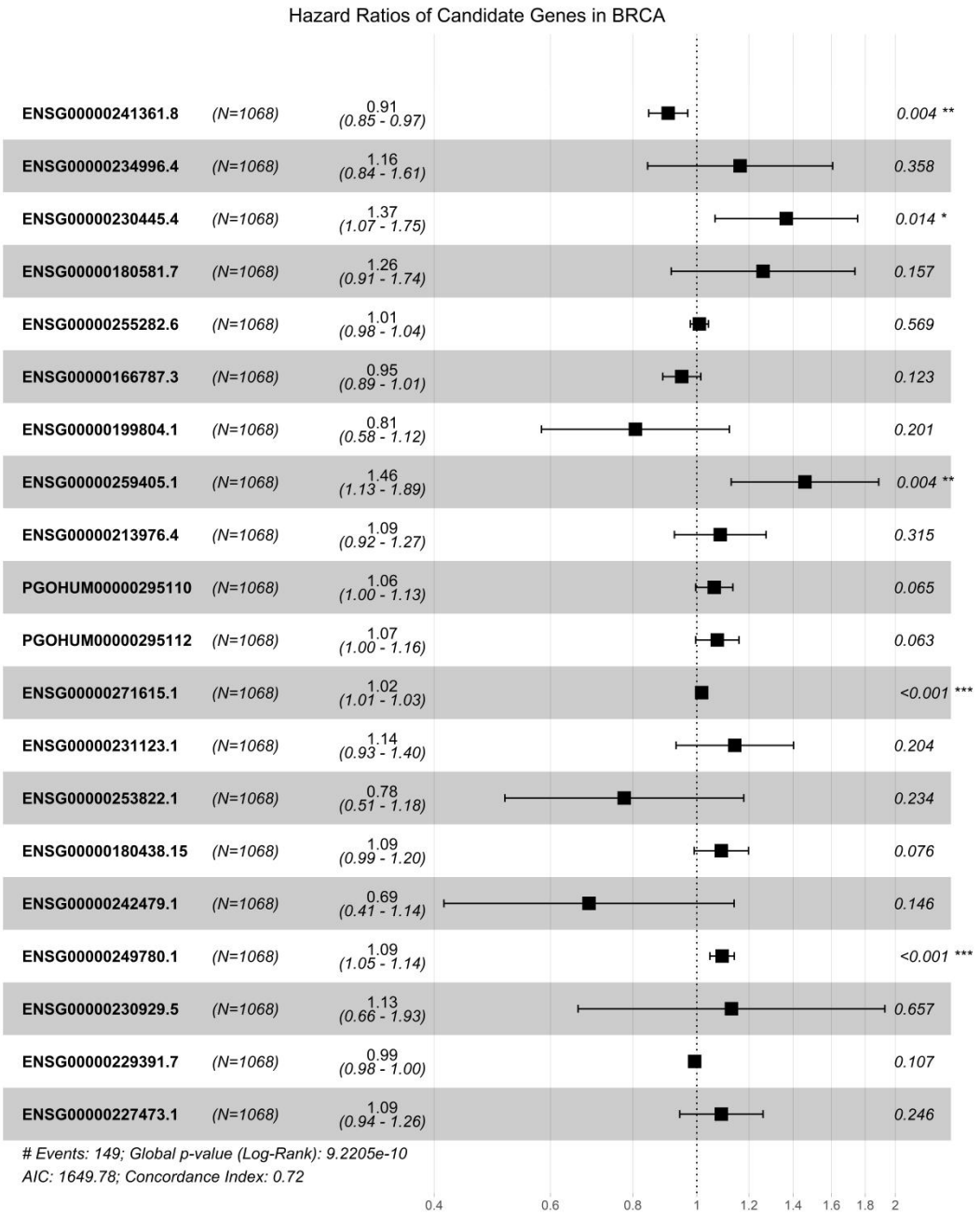


图 3-4 多变量 COX 回归分析森林图（BRCA）

Appendix Fig 3-4 Forest maps multivariate COX regression analysis（BRCA）