Journal of Clinical Personalized Medicine 临床个性化医学, 2025, 4(2), 1132-1141
Published Online April 2025 in Hans. https://www.hanspub.org/journal/jcpm
https://doi.org/10.12677/jcpm.2025.42285

Hans汉斯

# 基于生物信息学分析EGFR在乳腺癌中的表达和预后意义

**魏扬波**

华北理工大学生命科学学院，河北 唐山

收稿日期：2025年3月14日；录用日期：2025年4月7日；发布日期：2025年4月15日

## 摘　要

**通过生物信息学方法筛选乳腺癌相关基因，探究EGFR作为靶点在乳腺癌分子机制中的作用。分析公共基因芯片数据平台(gene expression omnibus, GEO)的乳腺癌RNA基因芯片数据集，得到差异基因并进行KEGG通路富集分析、蛋白相互作用网络分析得到核心基因，进一步研究其通路及其蛋白性质。结果表明，EGFR在原发性乳腺癌的发生、发展和迁移中起着重要的作用。**

## 关键词

**生物信息学，乳腺癌，EGFR**

# Expression and Prognostic Significance of EGFR in Breast Cancer: A Bioinformatics Analysis

**Yangbo Wei**

College of Life Sciences, North China University of Science and Technology, Tangshan Hebei

Received: Mar. 14th, 2025; accepted: Apr. 7th, 2025; published: Apr. 15th, 2025

## Abstract

**Breast cancer-related genes were screened using bioinformatics methods to explore the role of EGFR as a target in the molecular mechanisms of breast cancer. RNA gene chip datasets of breast cancer from the public gene chip data platform (Gene Expression Omnibus, GEO) were analyzed to obtain differentially expressed genes. Subsequently, KEGG pathway enrichment analysis and protein-protein interaction network analysis were conducted to identify core genes and further**

文章引用: 魏扬波. 基于生物信息学分析 EGFR 在乳腺癌中的表达和预后意义[J]. 临床个性化医学, 2025, 4(2): 1132-1141. DOI: 10.12677/jcpm.2025.42285

**investigate their pathways and protein properties. The results indicate that EGFR plays a significant role in the occurrence, development, and metastasis of primary breast cancer.**

## Keywords

**Bioinformatics, Breast Cancer, EGFR**



## 1. 引言

乳腺癌是女性最常见的恶性肿瘤性疾病，乳腺癌的发病机制尚不明确，其发生特点多因素、多步骤、错综复杂，随着乳腺癌样本数据的产生和计算机技术的快速提高，生物信息学这一新兴学科得到迅速发展。所以通过生物信息学方法找到乳腺癌中与发病机制有关的靶点，无疑是一种快速便捷的方法。通过NCBI 内的 GEO 数据库运用生物信息学方法筛选出与乳腺癌的核心基因，对其中一个核心基因 EGFR 进行生物信息学分析，并进一步深入研究该基因在乳腺癌中产生的通路效应与生物功能。

## 2. 材料及方法

### 2.1. 材料

利用 NCBI 网站提供的 GEO 数据库，选取 GSE124646 数据集，数据样本中包含 10 组(一组十个)不同癌细胞比例的样本，本实验采用了该数据集中两组数据(100%正常细胞与 100%癌细胞)共 20 个数据按照完全正常与完全癌化分为两组[1]。

### 2.2. 方法

#### 2.2.1. 差异基因的筛选

将两组样本数据导入 R 语言(3.6.1 版本)中，确定两组数据无重复性，运用数据包(limma)得到基因表达列表与探针芯片结合产生的基因表达矩阵，以|LogFC| > 2 且 P < 0.05 为筛选标准，运用统计学方法获取差异基因[2]。将上调基因与下调基因分别运用 String (https://string-db.org/)和 Cytoscape (3.7.1 版本)进行网络互作挑选表达最清晰的一组，下调基因表达更完整选择联通度最高的基因 EGFR 对该基因表达的蛋白质进行分析并分析包含该基因的通路[3]。

#### 2.2.2. 蛋白质结构与理化性质分析

对于 EGFR 编码的蛋白使用 ExPASy 在线工具对蛋白质分子量、等电点、疏水性等理化性质分析得到其物理性质(ExPASy, https://web.expasy.org/)，使用 SOPMA 在线工具预测蛋白质二级结构(SOPMA, https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html)，使用 SignalIP 在线工具预测其信号肽及剪切位点证明其是否跨膜(SignalIP, http://www.cbs.dtu.dk/services/SignalP/)，对于蛋白质功能域受体的预测通过 PFAM (https://pfam.xfam.org/)与 SMART (http://smart.embl-heidelberg.de/)两种方法对比详细预测该蛋白的功能域位点[3]。

#### 2.2.3. KEGG 与 GO 通路富集分析

在 R 语言中对挑选出的差异基因进行 KEGG 与 GO 基因富集通路，两种方法得出更详细的基因功能

[2]，运用 Cytoscape 进行通路可视化，寻找关于 EGFR 基因的信号通路，对 EGFR 在这些通路中起到的作用及其功能进行描述，了解差异基因所具有的生物学意义以及参与的重要生物学途径。

#### 2.2.4. 生存曲线分析

使用 Kmplot (http://kmplot.com/analysis/)数据库分析通路中有关基因表达水平与无复发生存率(RFS)的关系，数据以危险比(HR)和 95%可信区间(95%CI)显示这些基因与乳腺癌相关的生存曲线[4]。

## 3. 结果

### 3.1. 差异基因的表达

在 R 语言中导入的两组数据首先进行数据处理证明两组数据不重复，构建实验矩阵并使其实验数据标准化，选用数据包(limma)得到差异基因的统计学数据，选择 GPL570 探针包转换探针名得到基因 id 和基因名称，在热图(见图 1)中显示差异基因得到明显的上调与下调模块，差异基因用火山图(见图 2)表示，显示差异基因上调 128 个(adj.p.val < 0.05 & logFC > 2)，下调 258 个(adj.p.val < 0.05 & logFC < −2)，共 386 个。分别提取出上调与下调基因进行 String 与 Cytoscape 蛋白网络互作分析，将两组基因分别输入到 String 在线软件分析中得到蛋白质的相互作用包括直接物理相互作用和间接的功能相关性，形成蛋白互作网络图，并使用 Cytoscape 进行蛋白质网络可视化处理得到两组蛋白质网络图显示出连通度最高的基因 CDK1，KIAA0101，TOP2A，EGFR，IGF1 等(见表 1)，选择下调基因网络互作图中连通度高的 EGFR 基因进行生物信息学分析。

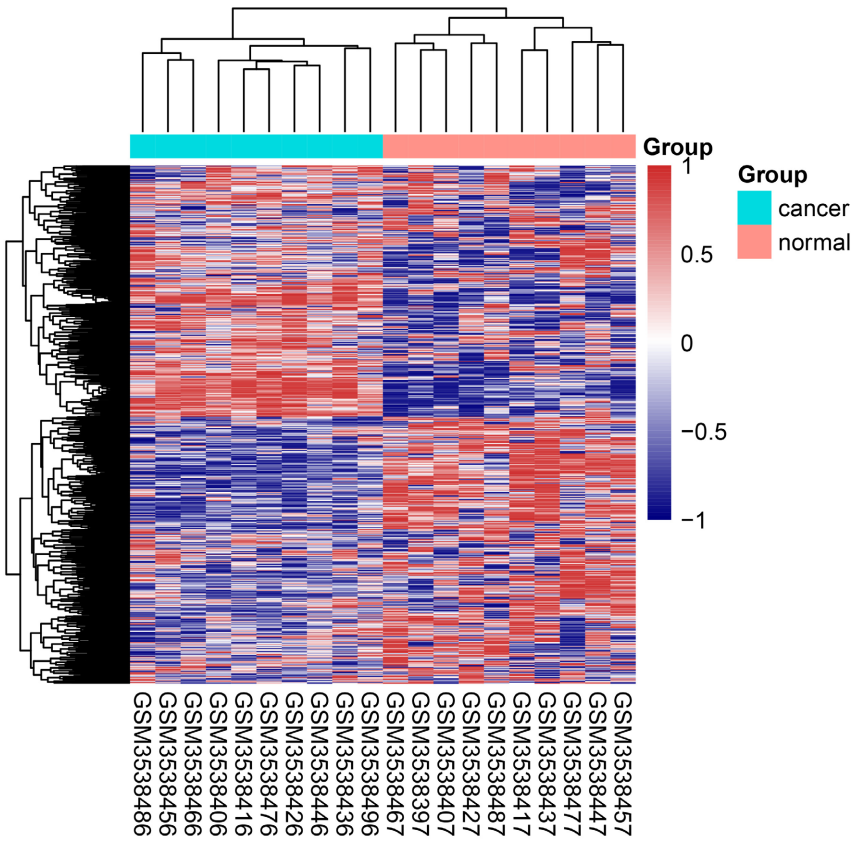



**Figure 1.** Heatmap showing gene comparison and cluster analysis
**图 1.** 热图显示基因对比与聚类分析

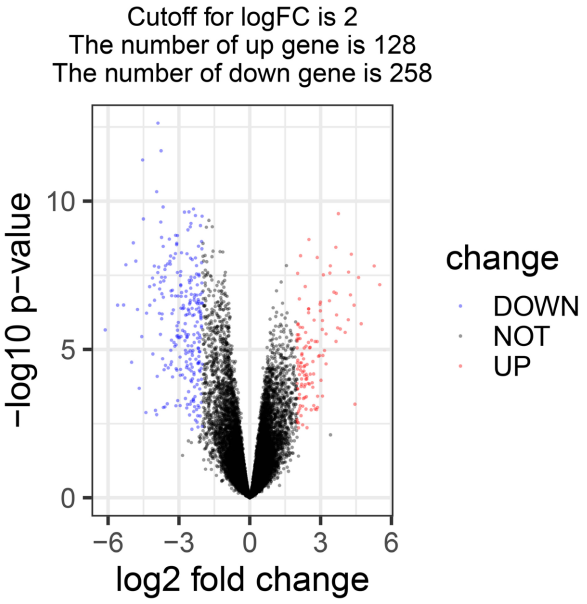



**Figure 2.** Volcano plot showing up-regulated and down-regulated genes
**图 2.** 火山图显示上调与下调基因

**Table 1.** Protein connectivity of gene expression
**表 1.** 基因表达的蛋白连通度

| 上调基因 | | | | 下调基因 | | | |
|---|---|---|---|---|---|---|---|
| CDK1 | 39 | BIRC5 | 35 | EGFR | 38 | FOS | 22 |
| KIAA0101 | 36 | KPNA2 | 35 | IGF1 | 31 | LPL | 20 |
| NUSAP1 | 35 | TOP2A | 35 | LEP | 28 | ADIPOQ | 19 |
| BUB1 | 35 | NURKA | 35 | PPARG | 26 | CXCL12 | 18 |
| CDC20 | 35 | FOXM1 | 35 | JUN | 26 | EGR1 | 17 |

## 3.2. 蛋白质性质

EGFR 基因表达的蛋白质是一种酪氨酸激酶受体(receptor tyrosine kinase, RTK)，该家族成员主要包括 4 种跨膜受体其中 EGFR 属于人表皮生长因子受体 1。EGFR 家族的配体主要有 EGF、TGF$\alpha$、AREG、BTC 等配体家族[5] [6]。

### 3.2.1. 蛋白质的基本理化性质及其疏水性分析[5]

利用 ExPOSy 在线工具对 EGFR 及其蛋白序列分析，其由 1210 个氨基酸组成，相对分子量为 134277.4，等电点为 6.26。该蛋白的半衰期在哺乳动物网织红细胞中约为 30 h；不稳定指数为 44.59，为不稳定蛋白；脂肪指数为 80.74，平均亲水系数为−0.316，预测为疏水蛋白[6]。

### 3.2.2. 蛋白质结构预测

通过 SOPMA 在线工具预测蛋白质二级结构预测发现 $\alpha$ 螺旋占 27.27%，延长链占 15.54%，无规则卷曲占 51.49%。通过 SignalIP 工具预测其信号肽及剪切位点为信号肽的可能性为 99.69%。通过 PFAM 与 SMART 两个工具对比得出其蛋白质功能域共四种：受体 L 域(Recep_L_domain)，类呋喃半胱氨酸富集区(Furin-like)，生长因子受体域 IV (GF_recep_IV)，蛋白质酪氨酸激酶(Pkinase_Tyr) [7]。

### 3.3. 基因通路富集

通过 R 语言分析下调基因的 KEGG 与 GO 富集通路，通过数据包(clusterProfiler)和(org.Hs.eg.db)进行 KEGG 和 GO 通路富集分析，通过(adj.p.val < 0.05)选择出 36 条基因通路(见图 3)，关于 EGFR 基因的通路 KEGG 有 3 条(见图 4)，GO 通路有 11 条(见图 5)，运用 Cytoscape 软件进行通路可视化得到通路富集图。

在 KEGG 通路图中包含 PI3K-Akt signaling pathway (hsa04151)、Focal adhesion (hsa04510)、MAPK signaling pathway (hsa04010)三种通路，关联最高的基因为 EGFR，PDGFD，PDGFRA，IGF1 [4]。

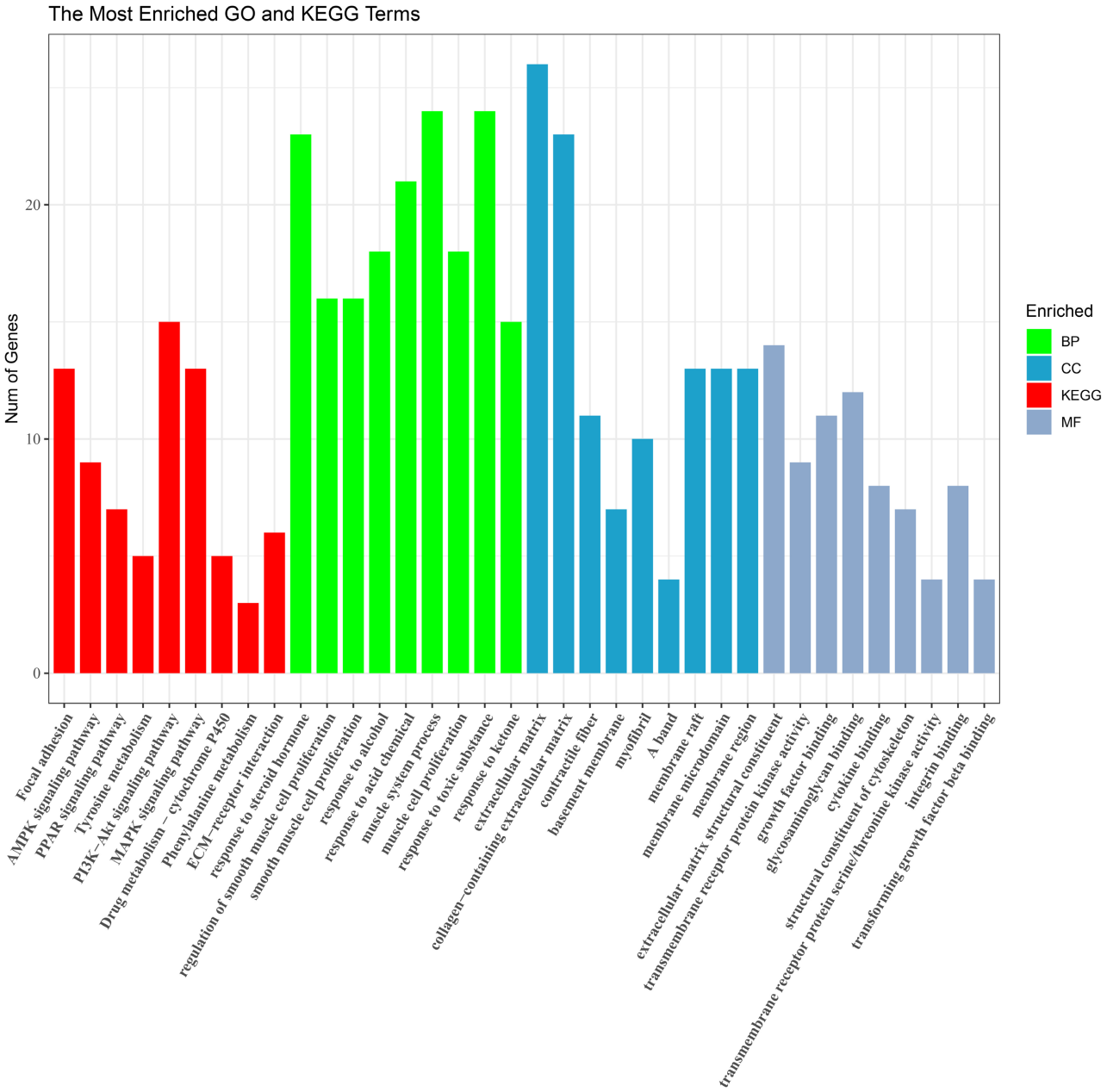



**Figure 3.** KEGG and GO pathway gene enrichment pathways
**图 3.** KEGG 与 GO 通路基因富集通路

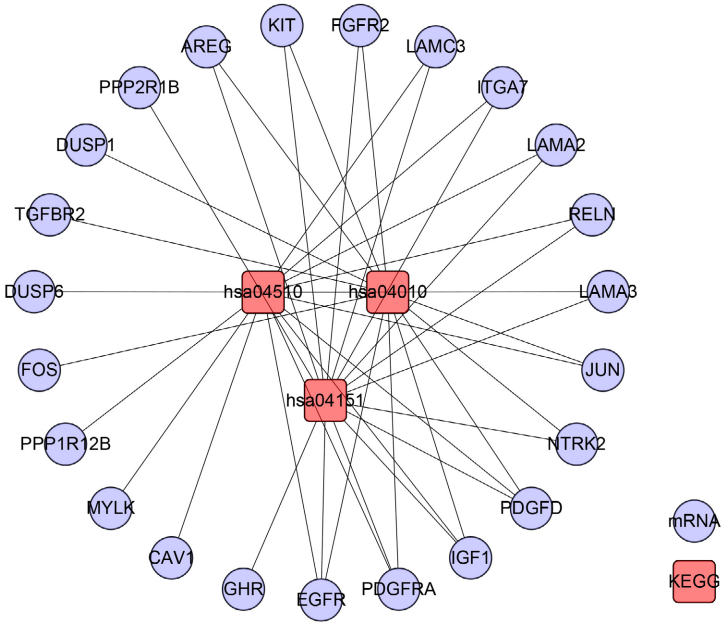



**Figure 4.** KEGG pathway of EGFR gene
**图 4.** EGFR 基因的 KEGG 通路

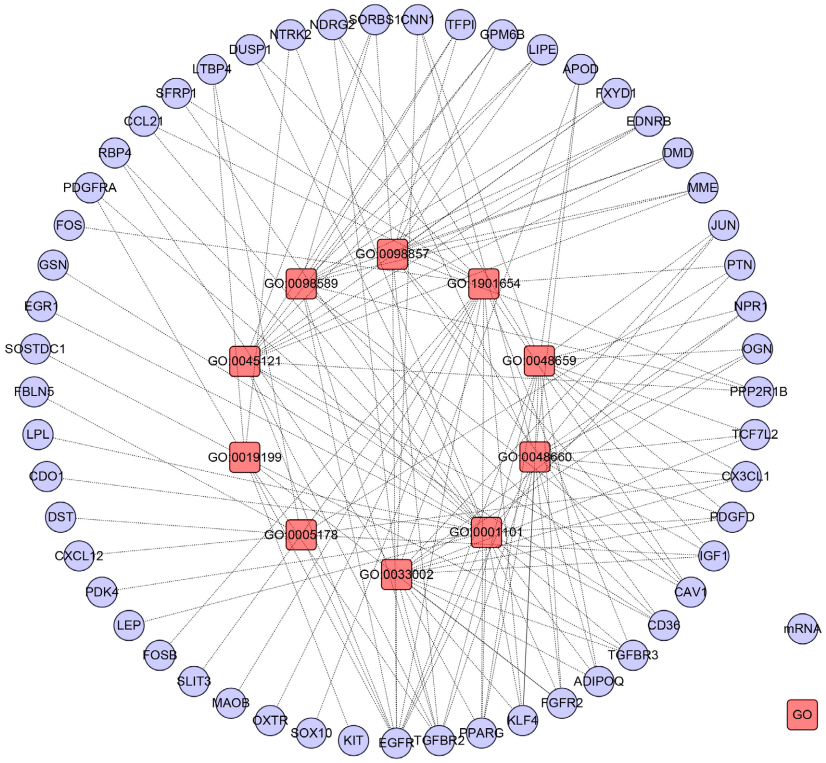



**Figure 5.** GO pathway of EGFR gene
**图 5.** EGFR 基因的 GO 通路

三种通路中 EGFR 等基因会激活生长因子(GF)与生长因子受体(RTK)相结合，该过程在三种通路主要以 RTK-Grb-2-Rass-Raf-ERK 途径影响三种通路的过程。当 RTK 结合时激活酪氨酸激酶，通过衔接蛋

白将信号传递给 Ras 蛋白，使 RAF 被活化在通过其磷酸化激活促分裂原激活的蛋白激酶的激酶(MEK)、促分裂原激活的蛋白激酶(ERK)等，ERK 进一步被转运使 ELK-1、SAP 等转录因子磷酸化调节 SRF 蛋白，参与细胞发育、生长、增殖、分化等多种生理、病理过程[8]。

PI3K-Akt 信号通路(hsa04151)中 GF 与 RTK 结合后会激活磷脂酰肌醇-3-激酶(PI3K)与 PI3K 结合时，会活化蛋白激酶 B (protein kinase B, PKB/Akt)，活化的 Akt 可影响细胞凋亡，细胞周期，糖酵解等过程[9]。两者结果相似会使信号传至细胞核内，使得核内转录因子磷酸化，启动靶基因的转录，最终导致细胞增殖、血管生成、DNA 修复等一系列生物学过程[10]。

MAPK 信号通路(hsa04010)中当 RTK 结合时伴随 Ras 蛋白激活与 MEKK1 产生作用，从而激活 MAP2K 异构体 MKK4，使 c-Jun N 端激酶(JNK)磷酸化。活化后的 JNK 会提高 AP-1 的转录活性，促进 DNA 的表达和蛋白质的合成。JNK 与 p38 可共同使 ELK-1，ATF-2 等转录因子发生磷酸化进而影响 P53 信号通路[8] [11]。

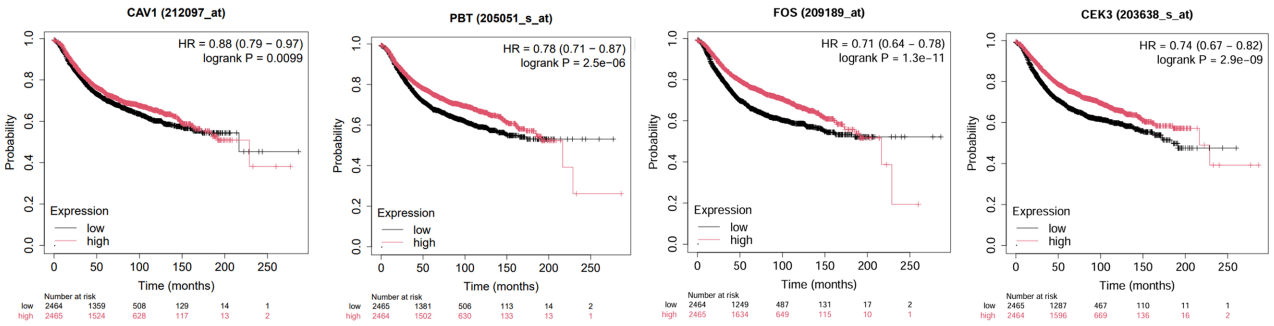
Focal adhesion 通路(hsa04510)中 RTK 结合信号因子时会使粘附斑激酶(FAK)产生一系列的效应，其复合体与衔接子蛋白结合，活化下游 JNK，而 JNK 活化转录因子 C-Jun 从而调控细胞增殖与分化。FAK 磷酸化后与 PI3K 结合，活化下游蛋白激酶 B，促进细胞周期进展及细胞增殖[12]。FAK 信号通路中当 RTK 结合时还会造成多种通路对于肌动蛋白骨架调节通路有一定影响，FAK 通过磷酸化使得 Src、Calpain、paxillin 等蛋白激活，而这些蛋白最后使得 actin 蛋白激活造成应力纤维形成，丝状伪足，板状伪足形成等生物学效应[13]。

在 GO 通路中显示关于细胞组分，生物过程和分子功能三个方面 GO 富集通路，在细胞中主要功能作用于细胞膜区域如：膜筏、膜微结构域、膜区等；其分子功能主要有：跨膜受体蛋白激酶活性，整合素结合，跨膜受体蛋白酪氨酸激酶活性等，生物学过程包括类固醇激素反应，平滑肌细胞增殖及其调节，酸性化学反应，肌肉细胞增殖，关联高的基因有 TGFBR2，KLF4，PPARG，FGFR2，CAV1，CD36，TGFBR3，PDGFD，IGF1，ADIPOQ，CX3CL1 等。

在 KEGG 和 GO 两类通路中挑选出 14 条通路，其中在 KEGG 和 GO 两类都存在的基因有 EGFR，CAV1，PDGFRA，KIT，TGFBR2，JUN，PPP2R1B，PDGFD，IGF1，NTRK2，DUSP1，FGFR2，FOS 共 13 种基因。

### 3.4. 十三种基因在乳腺癌中的生存曲线

在 Kmplot 数据库中检索通路中显示的十三个基因寻找其与 RFS 的关系，十一种显示 3955 例，两种显示 1764 例，分析所有有关数据，选择最佳截止时间从而得到其生存曲线(见图 6)显示出十三种基因表达水平对于乳腺癌患者具有极高的影响[14] (FOS 同为 CAV1；PPP2R1B 同为 PR65B)。其中十三种基因的 log rank P < 0.05 证明十三种基因具有意义。



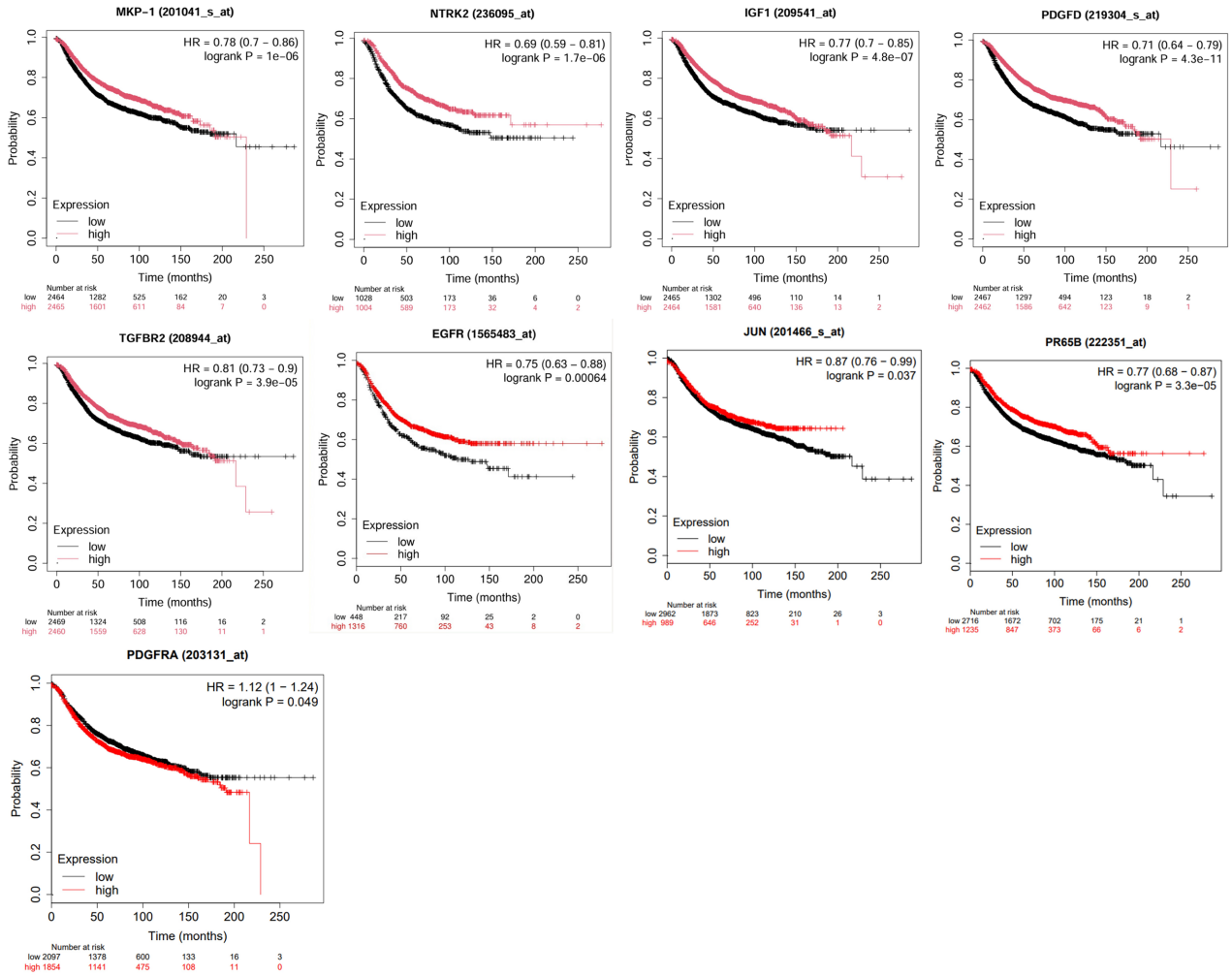



**Figure 6.** Survival curves of thirteen genes
**图 6.** 十三种基因的生存曲线

## 4. 讨论

本实验运用生物信息学方法探索乳腺癌核心基因从而探究该基因在乳腺癌的机制，提高了肿瘤基因定位的准确性，预测肿瘤的转移与否和预后判断有重要意义。通过 GEO 数据库比较乳腺癌肿瘤组织和正常组织对比，寻找肿瘤特异性表达基因，并研究这些基因富集的通路。我们筛选出乳腺癌差异基因 EGFR，并通过生物信息学方法预测了该基因表达的蛋白质基本性质及该基因影响的主要通路及其作用，根据其信号肽预测表现出该蛋白可能是一种跨膜信号蛋白，配合其四种功能域的作用及理化性质分析出该蛋白应该属于受体型，且酪氨酸激酶与磷酸化有相关性反应确定该蛋白为酶偶联型受体。通过检索 NCBI 数据库得出 EGFR 是位于人体 7p13-q22 染色体上的酪氨酸蛋白激酶受体。与通路分析的结果相对应[15]。

表皮生长因子受体(Epidermal Growth Factor Receptor, EGFR)作为受体酪氨酸激酶家族的重要成员，其异常激活与多种肿瘤的发生发展密切相关。在乳腺癌中，EGFR 的过表达或突变通过调控多条信号通路，影响肿瘤细胞的增殖、存活、侵袭和转移[16]。

通过 PI3K-Akt signaling pathway (hsa04151)、Focal adhesion (hsa04510)与 MAPK signaling pathway (hsa04010)三种通路得知关于 EGFR 基因的主要通路为 RTK-Ras-ERK 通路，其在三类通路中主要影响细

胞的增殖与分化等功能[15]。EGFR 基因作为上游基因影响三类通路中绝大多数的信号通路，其在 PI3K 通路和 MAPK 通路中介导的中间产物影响其他通路的反应。例如，在 HER2 阳性乳腺癌中，EGFR 与 HER2 的异源二聚化可激活 PI3K-Akt 通路，导致对曲妥珠单抗等靶向药物的耐药。此外，EGFR 信号还可通过激活 STAT3 通路，促进肿瘤干细胞(CSCs)的自我更新和分化，导致肿瘤的复发和耐药[17]。这三类通路与 GO 分析共同显示出通过与生长因子的结合影响细胞的转移，分化，增殖，周期调控等一系列生物学功能，在某些细胞中应起到协同作用。

通路 P53 signaling pathway (hsa04115)和 cell cycle (hsa04110)是三条通路的进一步结果[18]。三条通路主要调控细胞周期，PI3K 信号通路中其通过影响下游信号因子影响细胞周期的凋亡[19]，MAPK 途径参与细胞周期中的细胞增殖与细胞凋亡发挥重大作用，而黏着斑通路影响 PI3K 与 MAPK 两条通路从而在细胞周期的增长、凋亡的过程中发挥了极为重要的作用[15] [20]。P53 信号通路是通过防止细胞应激或 DNA 损伤引起的突变来保持基因组的稳定性，其与多种信号通路相互作用来稳定基因组，从而调节多种细胞过程，包括凋亡、衰老、细胞周期阻滞、分化、DNA 修复和复制与 cell cycle 信号通路进一步呼应[21]。

通过生存曲线的显示基因表达在乳腺癌的影响巨大，对于肺癌、卵巢癌、宫颈癌、前列腺癌、膀胱癌等其他癌症的研究中发现，因 EGFR 信号通路传导失调所引起的肿瘤占相当大的比重。另外，EGFR 基因的失调可促进肿瘤细胞的增殖、肿瘤血管生成、黏附、侵袭、转移和肿瘤细胞的凋亡等生物学机制，侧面验证了其对于细胞周期调控有关，预测 EGFR 在原发性乳腺癌的发生、发展和迁移中起着重要的作用。EGFR 通过调控多条信号通路，在乳腺癌的发生、发展及耐药过程中发挥关键作用。深入解析 EGFR 的分子机制，可为开发更有效的靶向治疗策略提供理论依据。未来的研究需进一步探讨 EGFR 与其他信号通路的交叉对话，以及其在肿瘤微环境中的动态调控作用。

## 参考文献


[1] Liang, Y., Zhang, H., Song, X., *et al*. (2020) Metastatic Heterogeneity of Breast Cancer: Molecular Mechanism and Potential Therapeutic Targets. *Seminars in Cancer Biology*, **60**, 14-27. https://doi.org/10.1016/j.semcancer.2019.08.012

[2] 朱晓菲, 冯振博, 沈思乔, 等. 基于 GEO 乳腺癌芯片数据的生物信息学分析[J]. 临床合理用药杂志, 2016, 9(10): 21-22+26.

[3] 张莉, 张开炯, 吴立春, 等. 乳腺癌相关长链非编码 RNA 的生物信息学分析[J]. 肿瘤预防与治疗, 2018, 31(5): 305-312.

[4] 冯刚, 李良平, 崔蕾, 等. 乳腺癌脑转移相关基因的生物信息学分析[J]. 巴楚医学, 2019, 2(2): 1-8.

[5] 孟丽荣, 李田, 李小毛. PTEN 和 EGFR 与子宫内膜癌发生和发展研究的最新进展[J]. 肿瘤, 2010, 30(5): 447-449.

[6] 谢萍芳, 赵东怡, 周美容, 等. 乳腺癌中 GFRA1 表达临床意义的生物信息学分析[J]. 中国肿瘤临床, 2018, 45(15): 769-773.

[7] 余海浪. 乳腺癌转移相关基因的生物信息学分析及功能研究[D]: [博士学位论文]. 广州: 南方医科大学, 2012.

[8] Bonin, S., Pracella, D., Barbazza, R., *et al*. (2019) PI3K/Akt Signaling in Breast Cancer Molecular Subtyping and Lymph Node Involvement. *Disease Markers*, **2019**, Article 7832376. https://doi.org/10.1155/2019/7832376

[9] Osaki, M., Oshimura, M. and Ito, H. (2004) PI3K-Akt Pathway: Its Functions and Alterations in Human Cancer. *Apoptosis*: *An International Journal on Programmed Cell Death*, **9**, 667-676. https://doi.org/10.1023/B:APPT.0000045801.15585.dd

[10] Kruger, D.T., Beelen, K.J., Opdam, M., *et al*. (2018) Hierarchical Clustering of Activated Proteins in the PI3K and MAPK Pathways in ER-Positive, HER2-Negative Breast Cancer with Potential Therapeutic Consequences. *British Journal of Cancer*, **119**, 832-839. https://doi.org/10.1038/s41416-018-0221-8

[11] 周海东, 卢姚宏, 胡梁深, 等. 基于信号通路探讨骨肉瘤的发病机制及中药干预研究进展[J/OL]. 中国比较医学杂志, 2025: 1-16. http://kns.cnki.net/kcms/detail/11.4822.R.20250317.1113.008.html, 2025-04-11.

[12] 许明诗. 肿瘤多药耐药中 miR-126-5p 的参与作用和靶向 FAK 的 PROTAC 降解剂逆转多药耐药的作用及机理研

究[D]: [博士学位论文]. 上海: 华东师范大学, 2024.

[13] 白学敏, 邹立考, 金珍木, 等. N-乙酰半胱氨酸通过调控 EGFR/MAPK 信号通路对慢性阻塞性肺疾病气道黏液高分泌的作用研究[J]. 中国临床药理学与治疗学, 2019, 24(10): 1120-1127.

[14] 李树裕, 王志钢. 粘附斑激酶(FAK)及其信号通路研究进展[J]. 生物技术通报, 2009(12): 6-10.

[15] 吴艺舟, 严雨欣, 孙杰. 黏着斑信号通路调控肿瘤上皮间质转化的研究进展[J]. 医学理论与实践, 2015, 28(19): 2601-2604.

[16] 王梦琪, 白雪峰, 赵志英. 人乳腺癌表皮生长因子受体 2 低表达的病理特征及动态变化分析[J]. 包头医学院学报, 2025, 41(2): 69-74.

[17] Brabletz, T., Jung, A., Spaderna, S., *et al*. (2005) Opinion: Migrating Cancer Stem Cells—An Integrated Concept of Malignant Tumour Progression. *Nature Reviews Cancer*, **5**, 744-749. https://doi.org/10.1038/nrc1694

[18] 马可儿. 基于 P53 信号通路的菜籽肽抗肝癌活性研究[D]: [硕士学位论文]. 南京: 南京财经大学, 2022.

[19] 孟艳. PI3K-Akt-mTOR 信号通路参与调控细胞分化[D]: [博士学位论文]. 北京: 中国协和医科大学, 2008.

[20] 吴子鑫, 吴申伟. PI3K/Akt/mTOR 信号通路及其与乳腺癌关系的研究进展[J]. 山东医药, 2020, 60(16): 107-110.

[21] 龚璐. p53 及其异构体△133p53/△113p53 在人类细胞和斑马鱼中的功能及生物学意义研究[D]: [博士学位论文]. 杭州: 浙江大学, 2015.