

• 论 著 •

不同人种乳腺癌的差异表达基因及关键蛋白调控基因预后分析

梁宝宝¹, 黄晶晶², 林 帅¹, 毛国超¹, 周章剑¹, 王亚晨³, 康华峰¹, 张淑群^{1△}

1. 西安交通大学第二附属医院肿瘤科, 陕西西安 710004; 2. 西安交通大学第二附属医院普通外科, 陕西西安 710004; 3. 西安交通大学第二附属医院皮肤科, 陕西西安 710004

摘要:目的 分析不同人种乳腺癌的差异表达基因及关键蛋白调控基因与预后的关系。方法 基于 TCGA 数据库中不同人种乳腺癌转录组表达谱来筛选差异表达基因, 并进行基因本体论注释和京都基因与基因组百科全书(KEGG)富集分析。应用 STRING 数据库构建蛋白质互相作用网络, 并在 Cytoscape 软件中进行可视化和模块挖掘, 获取关键蛋白调控基因。结合对应的临床信息使用 R 软件对这些关键蛋白调控基因绘制生存曲线进行验证。**结果** 共筛选出差异表达基因 894 个, 其中上调基因 875 个, 下调基因 19 个。差异表达基因主要与角化作用、表皮细胞分化等生物学过程有关, 与角质化包膜、中间丝细胞骨架、角蛋白丝等细胞组分有关, 与皮肤表皮的结构成分、葡萄糖醛酸转移酶活性等分子功能有关。KEGG 通路主要与细胞色素 P450 代谢异种生物、戊糖和葡萄糖醛酸转换、化学致癌作用等有关。**结论** 关键蛋白调控基因 CASP14、MYF6、ALB 与不同人种乳腺癌的发生和预后有紧密的关系, 为精准治疗提供了新思路。

关键词:乳腺癌; 差异表达基因; 关键蛋白调控基因; 总生存曲线

DOI:10.3969/j.issn.1673-4130.2021.23.020 **中图法分类号:**R737.9

文章编号:1673-4130(2021)23-2914-05 **文献标志码:**A

Prognostic analysis of key protein regulatory genes and differentially expressed genes in breast cancer of different races

LIANG Baobao¹, HUANG Jingjing², LIN Shuai¹, MAO Guochao¹, ZHOU Zhangjian¹, WANG Yachen³, KANG Huafeng¹, ZHANG Shuqun^{1△}

1. Department of Oncology, the Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi 710004, China; 2. Department of General Surgery, the Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi 710004, China; 3. Department of Dermatology, the Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi 710004, China

Abstract: Objective To analyze the relationship between differentially expressed genes and key protein regulatory genes of different races of breast cancer and the prognosis. **Methods** Based on the transcriptome expression profiles of different races of breast cancer in TCGA database, differentially expressed genes were screened, and gene ontology annotation and Kyoto gene and genome Encyclopedia (KEGG) enrichment analysis were performed. The protein-protein interaction network was constructed by STRING database, and the key protein regulatory genes were obtained by visualization and module mining in Cytoscape software. Combined with the corresponding clinical information, the survival curve of these key protein regulatory genes was drawn and verified by R software. **Results** A total of 894 differentially expressed genes were screened, including 875 up-regulated genes and 19 down-regulated genes. Differentially expressed genes were mainly related to biological pathways such as keratinization and epidermal cell differentiation, related to cellular components such as keratinized envelope, intermediate filament cytoskeleton and keratin filaments, and related to molecular functions such as structural components of skin epidermis and glucuronidase activity. KEGG pathway was mainly related to cytochrome P450 metabolism of xenobiotics, pentose and glucuronic acid conversion, chemical carcinogenesis and so on. **Conclusion** The key protein regulatory genes CASP14, MYF6 and ALB are closely related to the occurrence and prognosis of breast cancer in different races, and provide a new idea for

作者简介:梁宝宝,男,主治医师,主要从事肿瘤方面的研究。△ **通信作者,** E-mail: zhangshuqun1971@aliyun.com。

本文引用格式:梁宝宝,黄晶晶,林帅,等. 不同人种乳腺癌的差异表达基因及关键蛋白调控基因预后分析[J]. 国际检验医学杂志, 2021, 42 (23): 2914-2918.

precise treatment.

Key words: breast cancer; differentially expressed genes; key protein regulatory genes; overall survival curve

乳腺癌是全世界女性最常见的恶性肿瘤之一,也是癌症死亡的主要原因^[1],其危险因素主要有年龄、家族病史、生殖因素、绝经后女性肥胖和激素等^[2]。内源性和外源性雌激素暴露与乳腺癌风险增加有关。为了提供更有效的诊断和治疗策略,持续努力地了解与乳腺癌发病和发展进程有牵连的关键分子十分必要。目前,雌激素受体(ER)、孕激素受体(PR)、Ki67 和人表皮生长因子受体 2(HER2)是可以预示乳腺癌治疗的预后标志物。基于这些基因确定的肿瘤细胞表达,已对乳腺癌的 5 种分子亚型进行了分类^[3],包括以下子类型:lumina A(ER+、PR+、HER- 和 Ki67 < 14%);luminal B 的 HER-2-(ER+、PR+、HER- 和 Ki67 ≥ 14%);luminal B 的 HER-2+(ER+、PR+、HER+、任何数值 Ki67);HER-2 型(ER-、PR-、HER+、任何数值 Ki67);三阴性(ER-、PR-、HER-、CK5/6+ 和(或)egfr+)^[4-6]。有研究表明,乳腺癌具有常见的体细胞拷贝数变化,这些变化因种族而异,并且与生存有关^[7],种族差异存在于不同年龄女性^[8]。黑种人妇女更有可能在较年轻时被确诊,并伴随更多的生物侵略性三阴性亚型^[9]。与白种人妇女比较,年龄小于 40 岁的年轻黑种人妇女乳腺癌病死率更高。在治疗和预后的最新进展中,乳腺癌病死率的种族差距还在逐渐扩大^[10]。种族的生存差异可能是因为在诊断时获得护理时间、肿瘤形态阶段、治疗的质量和合适的辅助治疗等方面有所不同^[11-12]。有研究指出,即使在调整年龄、肿瘤大小、淋巴结状态、激素受体状态和组织学因素后,黑种人妇女在所有疾病阶段的病死率均较高^[13]。因此,本文旨在通过高通量测序的转录组数据,研究不同人种的乳腺癌差异表达基因和其所在的功能通路,并对关键蛋白调控基因进行生存分析,为乳腺癌的临床治疗提供新的见解与方向。

1 资料与方法

1.1 一般资料 本研究使用的不同人种乳腺癌数据来源于 TCGA 数据库(<https://portal.gdc.cancer.gov/>),分别下载了 RNA-seq counts 表达谱数据和临床样本数据。共有 925 例乳腺癌样本,包括 745 例白种人乳腺癌样本和 180 例黑种人或非裔美国人乳腺癌样本。

1.2 方法

1.2.1 转录组数据预处理 将原始测序数据首先通过 STAR 软件^[14]建立索引,把 reads 比对到人类参考基因组 hg38 生存 BAM 文件。之后使用 HTSeq 软件^[15]进行基因表达定量以得到 RNA-seq counts

数据。

1.2.2 差异表达基因筛选 在 TCGA 数据集 745 例白种人乳腺癌样本和 180 例黑种人或非裔美国人乳腺癌样本中使用 R 软件的 DESeq2 程序包对 RNA-seq counts 数据进行归一化,之后再分析差异表达基因。使用 *t* 检验和差异倍数(FC)法筛选差异表达基因,同时采用多重检验校正方法 Benjamini-Hochberg 对 *P* 值进行校正并控制错误发现率(FDR)。本文定义符合 $FDR < 0.05$ 且 $|\log_2 FC| \geq 1$ 的基因为差异表达基因。

1.2.3 功能富集分析 为详细阐明不同人种差异表达基因所在的通路,本文应用 R 软件的 clusterProfiler、org、Hs、eg、db 等程序包进行基因本体论(GO)注释^[16]和京都基因与基因组百科全书(KEGG)富集^[17]分析。GO 注释类型包括:(1)基因参与的生物学过程;(2)基因所处的细胞组分;(3)基因执行的分子功能。同上,采用多重检验校正方法对 *P* 值进行校正,设置满足 $P_{\text{adjust}} < 0.01$ 的 GO 术语和 KEGG 通路具有统计学显著性。

1.2.4 蛋白质相互作用(PPI)网络构建及模块挖掘 采用 STRING 数据库中的 PPI 信息进行网络构建^[18],设置最低相互作用分数为 0.7,并用 Cytoscape 软件进行可视化绘图。使用其中的分子复合物检测(MCODE)插件在构建 PPI 网络中进行模块挖掘。

1.2.5 生存分析 本文标记关键蛋白调控基因中表达高于表达值中位数为高,低于表达值中位数为低,结合 R 语言的 predict 函数对整体关键蛋白调控基因进行风险打分,得到高风险组和低风险组。同时使用临床信息数据绘制生存曲线进行验证。

2 结果

2.1 不同人种乳腺癌差异表达基因分析结果 在白种人乳腺癌组和黑种人或非裔美国人乳腺癌组基因表达谱数据中,使用 *t* 检验和 FC 法进行差异表达基因筛选,设置 $FDR < 0.05$ 且 $|\log_2 FC| \geq 1$ 的阈值,共筛选出 894 个差异表达基因,其中上调基因 875 个,下调基因 19 个。火山图显示白种人乳腺癌组和黑种人或非裔美国人乳腺癌组差异表达基因的表达情况。横坐标 $\log_2$ 表示以 2 为底白种人乳腺癌组基因表达值比黑种人或非裔美国人乳腺癌组基因表达值的对数值,差异越大其绝对值越高;纵坐标 $-\log_{10}$ FDR 表示以 10 为底 FDR 的对数值的相反数,差异越明显其数值越高。图中右虚线右边的点代表差异表达的上调基因,虚线中间的点代表低于阈值($FDR < 0.05$ 且 $|\log_2 FC| \geq 1$)而被认为无差异的基因,左

虚线左边的点表示差异表达下调的基因,差异表达情况见图 1。最大的 5 个上调和 5 个下调基因见表 1。

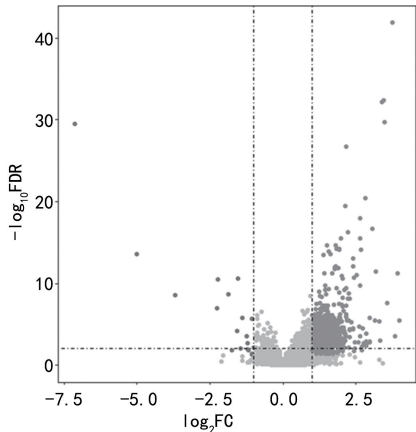


图 1 不同人种乳腺癌差异表达基因火山图

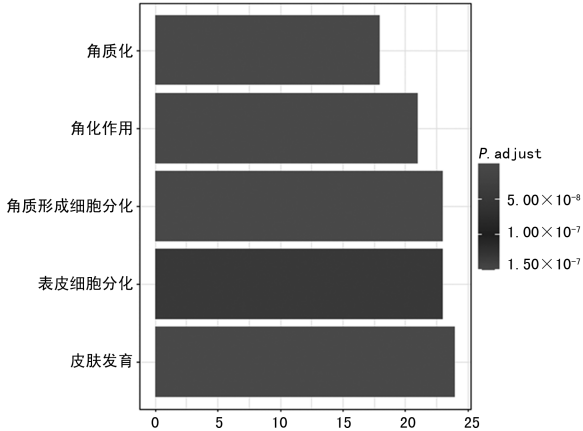
表 1 最大的 5 个上调和 5 个下调基因

基因名称	log ₂ FC	P	P, adjust
上调基因			
RNU1-11P	4.003 8	1.42×10^{-8}	3.83×10^{-6}
MAGEA4	3.925 2	4.55×10^{-15}	5.71×10^{-12}
FTHL17	3.839 9	2.61×10^{-6}	2.53×10^{-4}
SCARNA5	3.759 3	3.33×10^{-47}	1.34×10^{-42}
RNU1-88P	3.565 4	3.47×10^{-11}	2.32×10^{-8}
下调基因			
CSN2	-7.143 3	3.52×10^{-34}	2.83×10^{-30}
LALBA	-5.020 8	1.48×10^{-17}	2.70×10^{-14}
IAPP	-3.685 7	3.14×10^{-12}	2.43×10^{-9}
GABRA1	-2.261 0	2.19×10^{-10}	1.21×10^{-7}
XAGE3	-2.219 4	3.50×10^{-14}	3.61×10^{-11}

2.2 差异表达基因功能富集分析 对差异表达基因进行 GO 注释和 KEGG 富集分析,结果分别富集到 1 857 条 GO 生物学通路,259 条 GO 细胞组分通路,364 条 GO 分子功能通路,137 条 KEGG 通路。符合阈值 $P_{\text{adjust}} < 0.01$ 的生物学术语 12 条,图 2 展示了最显著的 5 条,它们主要与角化作用、表皮细胞分化等有关;细胞组分术语 6 条,图 3 展示了最显著的 5 条,主要与角质化包膜、中间丝细胞骨架、角蛋白丝等有关;分子功能术语 2 条,见图 4,主要与皮肤表皮的结构成分、葡萄糖醛酸转移酶活性有关。符合阈值 $P_{\text{adjust}} < 0.01$ 的 KEGG 通路 10 条,图 5 展示了最显著的 5 条,它们主要与细胞色素 P450 代谢异种生物、戊糖和葡萄糖醛酸转换、化学致癌作用等有关。

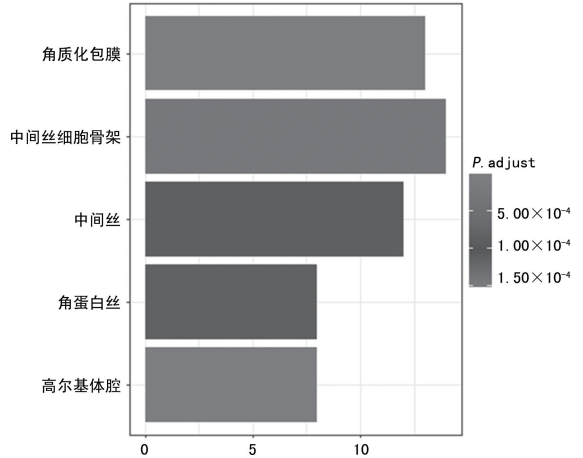
2.3 差异表达基因 PPI 网络构建及模块挖掘 在调节生物学进程中,蛋白与 PPI 起至关重要的作用。对差异表达基因进行 PPI 网络构建可展示这种互相作用关系,本文选择互相作用得分大于 0.7 分的结果,导入 Cytoscape 软件见图 6。在 MCODE 插件中进行模块聚类,可分为 3 个主要模块,见表 2。3 个种子节

点基因分别为 CASP14、MYF6 和 ALB。
2.4 关键蛋白调控基因生存分析 为验证关键的种子节点基因与生存率之间的关系,绘制了 Kaplan-Meier 曲线,见图 7。结果表明,基因 CASP14、MYF6 和 ALB 表达的升高所对应的乳腺癌患者总体生存率明显升高,差异均有统计学意义 ($P < 0.05$)。



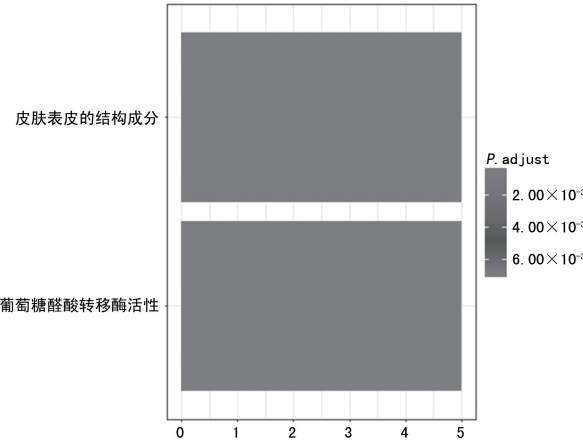
注:纵坐标表示富集到的不同 GO 生物学过程,横坐标表示注释到某条术语的基因数目。

图 2 差异表达基因的 GO 生物学过程注释



注:纵坐标表示富集到的不同 GO 细胞组分,横坐标表示注释到某条术语的基因数目。

图 3 差异表达基因的 GO 细胞组分注释



注:纵坐标表示富集到的不同 GO 分子功能,横坐标表示注释到某条术语的基因数目。

图 4 差异表达基因的 GO 分子功能注释

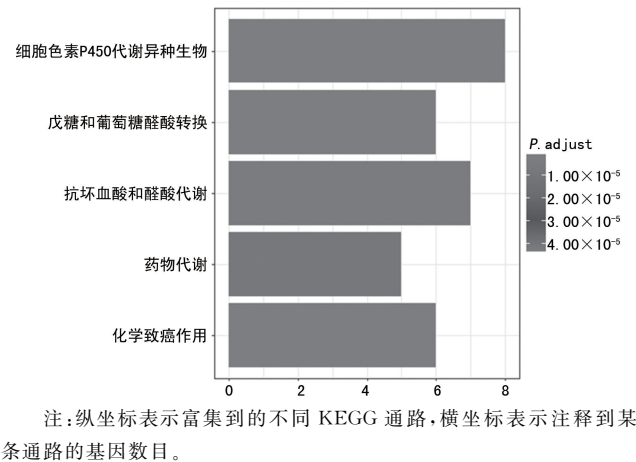


图 5 差异表达基因的 KEGG 富集分析

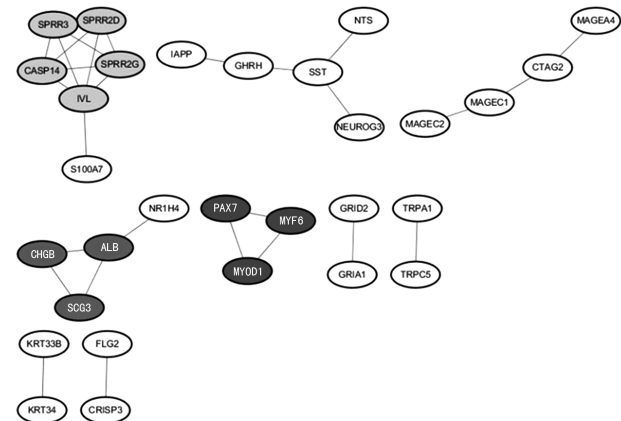


图 6 差异表达基因的 PPI 网络图

表 2 PPI 网络中对应的模块信息

模块分类	节点类型	得分(分)	基因名称
1	种子节点	4	CASP14
1	聚类节点	4	SPRR2G
1	聚类节点	4	SPRR2D
1	聚类节点	4	SPRR3
1	聚类节点	4	IVL
2	种子节点	2	MYF6
2	聚类节点	2	MYOD1
2	聚类节点	2	PAX7
3	种子节点	2	ALB
3	聚类节点	2	SCG3
3	聚类节点	2	CHGB

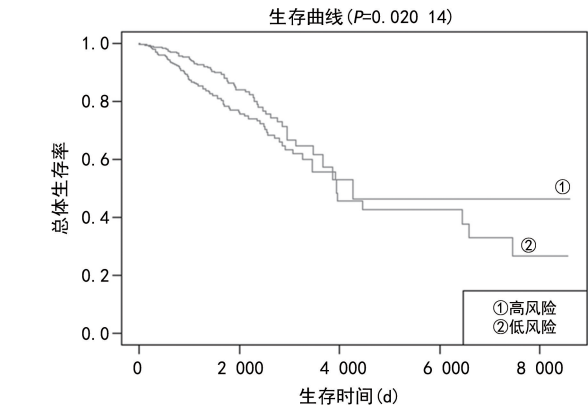


图 7 Kaplan-Meier 生存曲线

3 讨论

乳腺癌是女性最常见的癌症, 占有女性癌症的 30%^[19]。许多患有原发肿瘤或局部晚期乳腺癌的患者接受新辅助化疗, 此方法具有降低肿瘤分期的潜在好处, 同时可减少手术范围, 还可以早期评估全身治疗的疗效。有研究报道了包括乳腺癌在内的多种恶性肿瘤在预后方面的种族差异, 与白种人比较, 非白种人患者生存期较差, 肿瘤中存在的种族差异已归因于许多因素, 包括不同的肿瘤生物学、不积极的治疗及未能寻求医疗护理^[20]。

本研究旨在利用 TCGA 数据库中不同人种乳腺癌的转录组表达谱数据和临床信息数据进行生物信息学分析。鉴定出 894 个差异表达基因, 其中上调基因 875 个, 下调基因 19 个。差异最大的上调基因有 RNU1-11P、MAGEA4、FTHL17、SCARNA5、RNU1-88P, 差异最大的下调基因有 CSN2、LALBA、IAPP、GABRA1、XAGE3。通过 GO 注释和 KEGG 富集分析发现, GO 主要的生物学通路有角化作用、表皮细胞分化等; GO 主要的细胞组分通路有角质化包膜、中间丝细胞骨架、角蛋白丝等; GO 主要的分子功能通路有皮肤表皮的结构成分、葡萄糖醛酸转移酶活性等。KEGG 主要富集通路为细胞色素 P450 代谢异种生物、戊糖和葡萄糖醛酸转换、化学致癌作用等。另外, 对差异表达基因进行 PPI 网络构建和模块分析, 结果显示, 3 个主要模块被挖掘, 其中 CASP14、MYF6 和 ALB 基因被鉴定为关键蛋白调控基因, 并且与乳腺癌患者的生存有明显相关性。

有研究通过评估乳腺癌患者外周血中 2 个 MAGE 转录本 (MAGE-A3、MAGE-A4) 的临床可靠性和准确性发现, MAGE-A3 的存在与淋巴结状态明显相关, MAGE-A4 阳性与组织学分级明显相关, 它们的反转录聚合酶链反应检测可能对预后有预测意义, 有望成为乳腺癌的特异性肿瘤标志物^[21]。对 LALBA 基因与缺氧相关特征的研究表明, 此基因可以作为乳腺癌的潜在预后生物标志物^[22]。差异表达基因, 如 GABRA4、GABRG1 和 GABRA1 被确定为乳腺癌患者的潜在生物标志物^[23]。根据公共数据库, 高 CASP14 表达是与增殖、TNBC 表型和癌干性相关的乳腺癌侵袭性标志物^[24]。雌激素醌的清蛋白和血红蛋白胍合物是早期发现乳腺癌有前途的生物标志物^[25]。

此外, 本研究也存在不足之处。生物信息学方法进行研究的局限性表现在其部分结果需要进一步的功能实验验证。同时, 由于乳腺癌发生机制的个体异质性和研究样本数量的有限性, 本文仅收集到白种人和黑种人或非裔美国人乳腺癌的数据信息, 因此, 研究结果可能只适合于部分乳腺癌患者。

综上所述, 本研究通过分析不同人种乳腺癌转录组层面数据, 寻找相关差异表达基因, 进行 GO 注释

和 KEGG 富集分析,并对关键蛋白调控基因结合临床信息进行生存分析。研究结果可以为更深刻认识不同人种乳腺癌之间的分子机制提供新的见解。

参考文献

- [1] TORRE L A, BRAY F, SIEGEL R L, et al. Global cancer statistics, 2012[J]. *CA Cancer J Clin*, 2015, 65(2): 87-108.
- [2] YOUN H J, HAN W. A review of the epidemiology of breast cancer in asia: focus on risk factors[J]. *Asian Pac J Cancer Prev*, 2020, 21(4): 867-880.
- [3] GOLDBHIRSCH A, WOOD W C, COATES A S, et al. Strategies for subtypes-dealing with the diversity of breast cancer: highlights of the st. gallen international expert consensus on the primary therapy of early breast cancer 2011[J]. *Ann Oncol*, 2011, 22(8): 1736-1747.
- [4] GOLDBHIRSCH A, WINER E P, COATES A S, et al. Personalizing the treatment of women with early breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013[J]. *Ann Oncol*, 2013, 24(9): 2206-2223.
- [5] ARENA V, PENNACCHIA L, VECCHIO F M, et al. ER- /PR+ /HER2- breast cancer type shows the highest proliferative activity among all other combined phenotypes and is more common in young patients: experience with 6643 breast cancer cases[J]. *Breast J*, 2019, 25(3): 381-385.
- [6] PENNACCHIA I, VECCHIO F M, CARBONE A, et al. HER2 immunohistochemical assessment with A0485 polyclonal antibody: is it time to refine the scoring criteria for the "2+" category? [J]. *Appl Immunohistochem Mol Morphol*, 2015, 23(1): 31-35.
- [7] CHEN Y, SADASIVAN S M, SHE R, et al. Breast and prostate cancers harbor common somatic copy number alterations that consistently differ by race and are associated with survival[J]. *BMC Med Genomics*, 2020, 13(1): 116-119.
- [8] RICHARDSON L C, HENLEY S J, MILLER J W, et al. Patterns and trends in age-specific black-white differences in breast cancer incidence and mortality-United States, 1999-2014[J]. *Morb Mortal Wkly Rep*, 2016, 65(40): 1093-1098.
- [9] NEWMAN L A. Breast cancer disparities: high-risk breast cancer and african ancestry[J]. *Surg Oncol Clin N Am*, 2014, 23(3): 579-592.
- [10] DESANTIS C E, SIEGEL R L, SAUER A G, et al. Cancer statistics for African Americans, 2016: progress and opportunities in reducing racial disparities[J]. *CA Cancer J Clin*, 2016, 66(4): 290-308.
- [11] WILLIAMS D R, MOHAMMED S A, SHIELDS A E. Understanding and effectively addressing breast cancer in African American women: Unpacking the social context [J]. *Cancer*, 2016, 122(14): 2138-2149.
- [12] MILLER J W, SMITH J L, RYERSON A B, et al. Disparities in breast cancer survival in the United States (2001-2009): findings from the CONCORD-2 study[J]. *Cancer*, 2017, 123(Suppl 24): 5100-5118.
- [13] CHRISTIANSEN N, CHEN L, GILMORE J, et al. Association between African American race and outcomes in patients with nonmetastatic triple-negative breast cancer: a retrospective analysis by using results from the georgia cancer specialist database[J]. *Clin Breast Cancer*, 2012, 12(4): 270-275.
- [14] DOBIN A, GINGERAS T R. Mapping RNA-seq reads with STAR[J]. *Curr Protoc Bioinformatics*, 2015, 51: 11-14.
- [15] ANDERS S, PYL P T, HUBER W. HTSeq-a python framework to work with high-throughput sequencing data[J]. *Bioinformatics*, 2015, 31(2): 166-169.
- [16] CONSORTIUM G O. Gene ontology consortium: going forward[J]. *Nucleic Acids Res*, 2015, 43: 1049-1056.
- [17] KANEHISA M, GOTO S. KEGG: kyoto encyclopedia of genes and genomes[J]. *Nucleic Acids Res*, 2000, 28(1): 27-30.
- [18] VON MERING C, HUYNEN M, JAEGLI D, et al. STRING: a database of predicted functional associations between proteins[J]. *Nucleic Acids Res*, 2003, 31(1): 258-261.
- [19] SIEGEL R L, MILLER K D, JEMAL A. Cancer Statistics, 2017[J]. *CA Cancer J Clin*, 2017, 67(1): 7-30.
- [20] WARNER E T, TAMIMI R M, HUGHES M E, et al. Racial and ethnic differences in breast cancer survival: mediating effect of tumor characteristics and sociodemographic and treatment factors[J]. *J Clin Oncol*, 2015, 33(20): 2254-2261.
- [21] HUSSEIN Y M, GHARIB A F, ETEWA R L, et al. The melanoma-associated antigen-A3, -A4 genes: relation to the risk and clinicopathological parameters in breast cancer patients[J]. *Mol Cell Biochem*, 2011, 351(1/2): 261-268.
- [22] WANG J, WANG Y, XING P, et al. Development and validation of a hypoxia-related prognostic signature for breast cancer[J]. *Oncol Lett*, 2020, 20(2): 1906-1914.
- [23] WU J, ZHANG Y, LI M. Identification of methylation markers and differentially expressed genes with prognostic value in breast cancer[J]. *J Comput Biol*, 2019, 26(12): 1394-1408.
- [24] HANDA T, KATAYAMA A, YOKOBORI T, et al. Caspase14 expression is associated with triple negative phenotypes and cancer stem cell marker expression in breast cancer patients [J]. *J Surg Oncol*, 2017, 116(6): 706-715.
- [25] LIN P H, YANG H J, HSIEH W C, et al. Albumin and hemoglobin adducts of estrogen quinone as biomarkers for early detection of breast cancer[J]. *PLoS One*, 2018, 13(9): e0201241.