

骨质疏松症关键基因的筛选及生物信息学分析

许航 崔宇韬 任广凯 刘贺 王雁冰 彭传刚 吴丹凯

【摘要】 目的 通过生物信息学的方式筛选骨质疏松症关键基因,并进一步分析其与骨质疏松症的联系及作用机制。**方法** 从公共基因表达数据库下载基因表达谱数据集,通过R软件筛选差异表达基因并进行功能与通路分析。使用在线工具String构建蛋白质互作网络,导入cytoscape软件筛选关键基因并构建集簇模块。**结果** 共筛选出1 334个差异表达基因,其中上调基因722个,下调基因612个。GO分析显示功能主要富集在细胞外基质结构成分,信号受体激活剂活性,跨膜转运蛋白结合及细胞因子结合等方面。KEGG通路富集中显示差异基因主要参与PI3K-Akt信号通路、MAPK信号通路、Rap1信号通路以及Ras信号通路等通路。根据蛋白质互作网络筛选出AKT1、EGF、VEGFA、PROM1、TP53、NES、CD21、SNAI1、FGF13、LIF共十个关键基因,以及一个集簇模块。**结论** 筛选并分析了关键基因与集簇模块的功能、作用及其与骨质疏松可能存在的联系,为揭示骨质疏松症潜在的分子机制和药物靶点提供新的思路。

【关键词】 骨质疏松症; 差异基因表达; 关键基因筛选; 生物信息学

Screening and Bioinformatics Analysis of Core Genes of Osteoporosis Xu Hang, Cui Yutao, Ren Guang-kai, Liu He, Wang Yanbing, Peng Chuangang, Wu Dankai. Orthopaedic Medical Center, The Second Hospital of Jilin University, ChangChun 130000, China

Corresponding author: Wu Dankai, Email: wudk@jlu.edu.cn

【Abstract】 Objective In this study, we screened the hub genes of osteoporosis by bioinformatics, and further analyzed their relationship with osteoporosis and their mechanism of action. **Methods** Firstly, the gene expression profile data set was downloaded from the public gene expression database, and the differentially expressed genes were screened by R software, and their functions and pathways were analyzed. Then, the online tool String was used to construct the protein interaction network, and the software cytoscape was imported to screen the core genes and construct the clustering module. **Results** A total of 1334 differentially expressed genes were screened, including 722 up-regulated genes and 612 down-regulated genes. GO analysis shows that the functions are mainly concentrated in extracellular matrix structural constituent, signal receptor activator activity, transmembrane transporter binding and cytokine binding. The enrichment of KEGG pathway shows that the differential genes are mainly involved in PI3K- Akt signaling pathway, MAPK signaling pathway, Rap1 signaling pathway and Ras signaling pathway. Ten key genes, namely AKT1, EGF, VEGFA, PROM1, TP53, NES, CD21, SNAL1, FGF13, LIF, and one clustering module were selected according to protein interaction network. **Conclusion** The functions and roles of key genes and clustering modules and their possible relationship with osteoporosis were screened and analyzed, which provided new ideas for revealing the potential molecular mechanism and drug targets of osteoporosis.

【Key words】 Osteoporosis; Differential gene expression; Core gene screening; Bioinformatics

骨质疏松症是一种常见的骨代谢疾病,其特征为骨密度降低和骨微结构恶化,进而导致脆性增加和对骨折的易感性升高^[1]。骨质疏松性骨折是一种低能量、非暴力的骨折,是指日常活动中受到不明

显外力或者本不能引起骨折的创伤而发生的骨折^[2]。骨质疏松性骨折作为骨质疏松症患者最常见和最严重的并发症,在55岁后的女性和65岁后的男性中越来越常见,已成为一个世界性的问题^[3]。在我国约

有1.4亿的骨质疏松患者,男性和女性的10年骨质疏松症增长率分别约为15%和20%,40岁以上人群骨质疏松症发病率约为25%^[4]。

生物信息学是一门生物学与信息学交叉的新兴学科,为生物医学大数据转化为有价值的知识提供了无限可能^[5]。基因测序和微阵列分析技术的出现,为了解相关基因及通路对骨质疏松症的影响提供了有力支持。基因关联研究表明,骨质疏松症是一种受多基因和多种信号通路调节的疾病,其易感性已被定位到多个位点^[6]。然而,目前对骨质疏松症患者的药物靶点和基因通路注释尚不够明确,针对骨质疏松患者的治疗药物双膦酸盐、核因子 κ B受体活化因子配体(receptor activator of nuclear factor- κ B ligand, RANKL)抑制性抗体和特立帕肽等由于不良反应和长期效益等原因不能完全满足临床需要^[7]。因此,本论文旨在通过对基因芯片数据进行分析,筛选出骨质疏松症相关的差异表达基因、关键基因及关键网络,为揭示骨质疏松症潜在的分子机制和药物靶点提供可能。

材料和方法

一、研究资料

公共基因表达数据库(gene expression omnibus, GEO)(<https://www.ncbi.nlm.nih.gov/geo/>)是一个公共功能基因组数据库,从GEO下载GSE35955、GSE35956以及GSE35957的基因表达谱数据集。三个数据集均由GPL570[Hg-U133plus 2]Affymetrix人类基因组U133Plus2.0阵列平台(Affymetrix公司,美国加利福尼亚州)注释,共包括19例老年骨质疏松症患者和健康老年的mRNA微阵列数据。

二、芯片质量分析与数据预处理

首先,通过R软件(4.1.2)对芯片数据进行相对数表达、相对标准差和RNA降解等分析,选择合适的样本进行下一步分析。然后,我们使用Affy软件包将CEL文件中的探针数据转换为探针表达式矩阵,并通过RMA算法对不同的芯片数据进行标准化与归一化。最后,通过数据集芯片的平台注释包将矩阵中的探针序列号转换为基因名称。我们将所有探针表达值的平均值作为单个基因的表达值,并使用最近邻居法(K-NearestNeighbor, KNN)来补充缺失值。

三、差异表达基因的筛选

我们使用R软件中的limma软件包筛选差异基

因,并通过火山图可视化差异基因,筛选标准为: $P < 0.05$ 和 $|\log FC| > 1$ 。为了确定筛选的差异基因在患者和正常对照组之间的表达是否不同,我们对差异基因进行了聚类分析,用热图进行可视化。

四、差异基因的GO与KEGG富集途径

为了进一步了解骨质疏松患者差异表达基因的细胞组分、分子功能、生物过程及相关通路,本研究使用在线分析工具DAVID和R软件中的ClusterProfiler程序包共同对差异基因进行基因本体(gene ontology, GO)与京都基因和基因组百科全书(Kyoto Encyclopedia of Genes and Genomes, KEGG)通路富集分析,筛选标准为 $P < 0.05$ 。

五、PPI网络构建与关键基因、模块分析

本研究使用在线分析工具STRING(<https://cn.string-db.org/>)用于分析蛋白质相互作用(protein-protein interaction, PPI)网络,选择置信度得分为高置信度(> 0.700)的点与线构成蛋白质网络。然后,我们将相关数据导入Cytoscape软件,使用其内置的软件包cytoHubba,通过MMC算法筛选骨质疏松患者的10个关键基因。最后,我们使用Cytoscape软件内置MCODE软件包筛选和构建核心模块,并对网络评分最高的核心模块内的基因进行进一步的分析。

结 果

一、差异表达基因的筛选

对三个数据集中的数据进行处理,构建相对对数表达图、相对标准差图及RNA降解图进行芯片质量分析(图1~3)。我们发现编号为GSM878105的样本一致性较低且RNA降解较多,因此我们在随后的分析中剔除了该样本。用R软件中的limma软件包对剩余18个样本筛选差异基因,取值 $P < 0.05$ 和 $|\log FC| > 1$,共得到1334个差异表达基因,其中上调基因722个,下调基因612个。绘制火山图对差异基因进行可视化,红色代表上调基因,绿色代表下调基因(图4)。我们使用热图对前50个差异基因进行聚类及可视化,热图的颜色代表基因表达水平,红色为升高,蓝色为降低,颜色越亮代表基因升高及降低的程度越高(图5)。

二、GO与KEGG富集分析

对差异基因进行GO富集分析,可知差异基因的功能主要富集在细胞外基质结构成分(extracellular matrix structural constituent)跨膜转运蛋白结合

(transmembrane transporter binding)、细胞因子结合(cytokine binding)、受体配体活性(receptor ligand activity)、肌动蛋白结合(actin binding)以及信号受体激活剂活性(signaling receptor activator activity)等方面(图6)。在KEGG通路富集中,我们对 P 值 <0.05 的通路按照基因富集数排序,得到了PI3K-Akt信号通路、MAPK信号通路、Rap1信号通路以及Ras信号通路等通路的信息(表1)。

三、关键基因筛选和集簇模块的构建

使用在线分析工具STRING构建差异表达基因蛋白质相互作用网络,筛选标准为高信服度(>0.700)最终得到由964个点与2 796条线构成的PPI网络(图7)。将PPI网络导入Cytoscape,使用内置的软件包CytoHubba,通过MMC算法筛选骨质疏松患者的10个关键基因: AKT1、EGF、VEGFA、PROM1、TP53、NES、CD24、SNAI1、FGF13、LIF, AKT1、VEGFA、TP53、NES、CD24、SNAI1、LIF 为上调基

因,EGF、PROM1、FGF13为下调基因(图8)。最后,我们通过MCODE插件筛选并构建了网络评分为11.000的模块,见图9。对模块中的相关基因进行GO分析,发现模块里面的基因生物过程主要富集在MAPK的正向调节上,而功能主要富集在生长因子活性之上(表2),这与以往对骨质疏松的研究相符,进一步证明了通过以上方法探究骨质疏松的分子机制与通路是可行的。

讨 论

骨质疏松症是全球最常见的骨骼疾病之一,由于骨质疏松症引发的骨折已成为世界性的难题^[8]。虽然已经证实骨矿物质密度,骨质疏松症和骨质疏松性骨折具有高度遗传性,与特定基因的表达有关,但想要阐明遗传结构,特别是潜在的基因组和分子机制仍然具有极大的挑战性^[9]。

在本研究中,通过对GSE35955、GSE35956以及GSE35957的基因表达谱芯片进行分析,获得1 334个差异表达基因,并对其进行GO与KEGG富集分析。GO功能主要富集在细胞外基质结构成分,信号受体激活剂活性,跨膜转运蛋白结合及细胞因子结合等方面。因此,我们推测差异基因可能通过降低受体配体结合及跨膜转运蛋白结合等结合能

表1 差异表达基因的KEGG富集分析结果

通路	基因数目
hsa04151: PI3K-Akt signaling pathway	23
hsa04010: MAPK signaling pathway	19
hsa04015: Rap1 signaling pathway	12
hsa04014: Ras signaling pathway	11

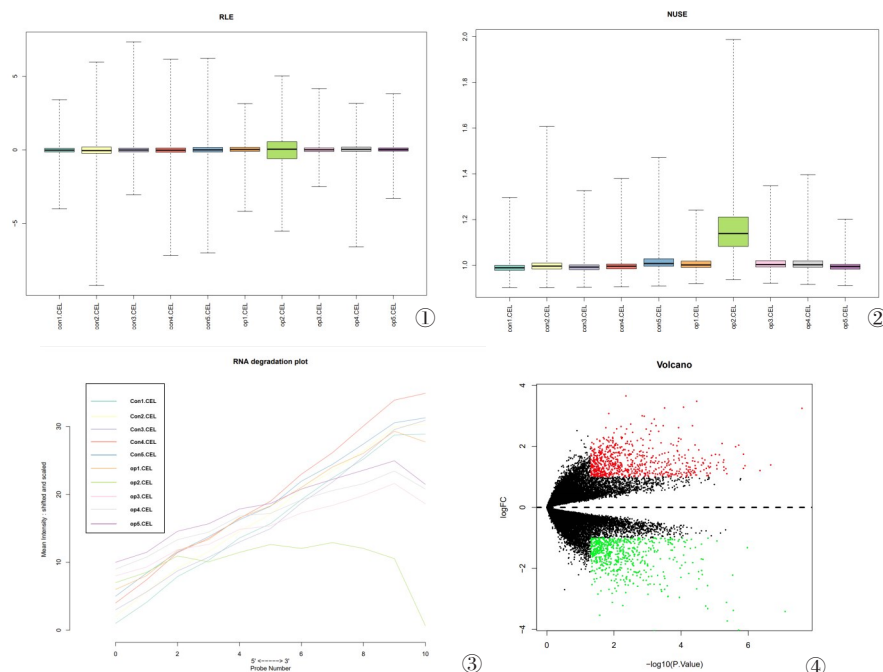


图1 GSE35956芯片表达数据的相对对数表达图 图2 GSE35956芯片表达数据的相对标准差图 图3 GSE35956芯片表达数据的RNA降解图 图4 差异表达基因的火山图

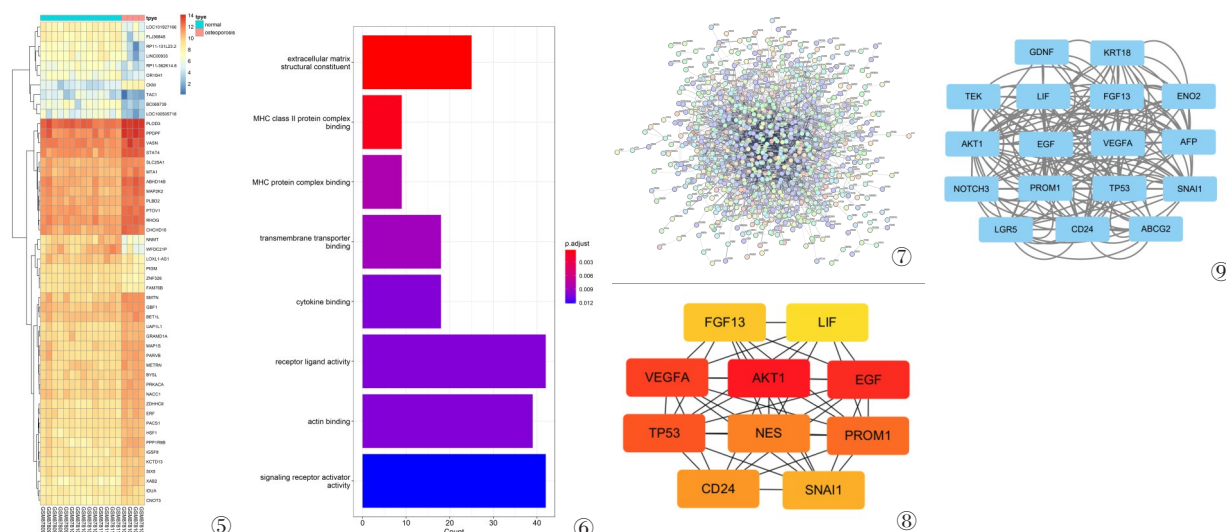


图5 前50个差异表达基因的热图 图6 差异表达基因的GO富集分析结果 图7 差异表达基因的蛋白质互作网络 图8 PPI网络中前10个关键基因,颜色深度代表了结点在网络中的重要性 图9 PPI网络中的关键集簇模块

表2 模块中基因的GO富集分析结果

名称	类别	描述	基因
GO:0008083	分子功能	生长因子活性	GDNF, EGF, LIF, FGF13, VEGFA
GO:0043410	生物过程	MAPK 级联的正向调节	EGF, LIF, TEK, VEGFA
GO:0005576	细胞组分	细胞外区域	NOTCH3, GDNF, EGF, LIF, TEK, FGF13, VEGFA

力,影响相关信号通路,进一步影响成骨细胞的相关功能,导致骨质疏松症的发生。KEGG 通路分析表明,骨质疏松症与多条信号通路有关,包括:PI3K-Akt 信号通路、MAPK 信号通路、Rap1 信号通路以及 Ras 信号通路等通路,这些通路可能成为骨质疏松研究的新思路。

本研究通过构建 PPI 网络,应用 MCC 算法进一步筛选出 10 个关键基因及其编码的蛋白质。Akt1 编码丝氨酸/苏氨酸蛋白激酶,可通过磷脂酰肌醇 3 激酶(phosphatidylinositol-3-kinase, PI3K)依赖的机制被胞外信号激活。已有研究表明, Akt1 为成骨细胞中独特的信号中间体,可以控制成骨细胞和破骨细胞分化^[10]。Akt1 的上调表达会降低成骨细胞分化,增强破骨细胞分化能力,导致骨质疏松症的发生与发展。表皮细胞生长因子(epidermal growth factor, EGF)属于表皮生长因子家族,其作用已被确定: EGF 增强骨髓基质细胞的增殖和迁移,诱导成骨细胞,软骨细胞和脂肪细胞的生成,是间充质组织再生的来源^[11]。因此,骨质疏松症患者 EGF 基因表达下调,导致骨量降低,骨微环境恶化。血管内皮生长因子 A (vascular endothelial growth factor A, VEGFA) 与骨质疏松症具有显著关联,并可作为支持向量机

(support vector machine, SVM)法筛选骨质疏松症的指标之一^[12]。然而,对于其影响骨质疏松症发展的具体机制仍需进一步实验探究。PROM1 是一种编码跨膜蛋白 CD133 的基因,该蛋白可作为重要的细胞表面标志物。同时,使用 CD133 处理后的细胞具有更好的血管生成能力与矿化能力,这与我们的结果相一致^[13]。TP53 基因在血清中的表达水平在骨质疏松症患者和 小鼠骨质疏松模型中上调,升高的 p53 水平与骨量的减少有关,可以通过敲低 p53 来实现骨量的部分逆转^[14]。NES 基因编码的巢蛋白是一种中间纤维蛋白,最初被用作神经上皮干细胞的标记蛋白,进一步研究发现,巢蛋白可以作为成骨细胞的有用标记蛋白^[15]。CD24 基因编码的蛋白被称为热稳定抗原,是一种高度糖基化的糖基磷脂酰肌醇锚定表面蛋白。研究表明,表面糖蛋白 CD24 阳性的骨髓间充质干细胞增殖和细胞外基质矿化能力降低^[16]。由此可知,上调表达的 CD24 会降低骨骼的再生潜能和细胞外基质的矿化能力,进而导致骨质疏松症的产生。SNAIL 基因编码锌指转录抑制因子,一项功能获得研究表明,小鼠长骨中 SNAIL 蛋白活性的上调会导致骨长度减少,对小鼠软骨合成也有显著影响。因此,我们推测该基因的上调对

人类骨质疏松症的产生具有相同的机制与作用效果。成纤维细胞生长因子13(fibroblast growth factor 13, FGF13)基因在急性髓系白血病中异常表达,与骨髓微环境密切相关^[17]。骨髓微环境的变化是否会导致骨密度降低和骨微结构变化仍需要更多的临床研究证明。LIF 基因编码白血病抑制因子,该因子是一种多效性细胞因子,与各种病理状况有关,如类风湿性关节炎和骨质疏松症^[18]。根据目前研究进展,该基因导致骨质疏松症的具体原因以及能否作为骨质疏松易感人群的遗传标记仍需要探索与阐明。

我们通过Cytoscape的MCODE插件对PPI网络进行处理,总共筛选构建了22个集簇模块。根据网络评分排序,我们选择评分最高的模块进行研究。该模块由17个相关基因与176条线组成,评分为11.000。对该模块进行GO分析,发现分子功能主要富集于生长因子活性,生物过程主要富集在MAPK级联的正向调节,而其作用的细胞组分为细胞外基质,这与上述对十个关键基因的讨论与猜想相符。

综上所述,本文通过对10个关键基因研究现状的检索分析,发现大部分基因在骨代谢领域已被初步阐明,但多数缺乏临床与实验的验证。同时,对上述模块的功能分析不仅证明了本研究的合理性,同时为进一步揭示骨质疏松症潜在的分子机制和药物靶点提供新的思路。然而,本研究得出的结论仍需进一步的具体实验来探究其可行性。

参 考 文 献

- 1 Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis [J]. Am J Med, 1993, 94(6): 646-50.
- 2 Siris ES, Adler R, Bilezikian J, et al. The clinical diagnosis of osteoporosis: a position statement from the National Bone Health Alliance Working Group [J]. Osteoporos Int, 2014, 25(5): 1439-1443.
- 3 Palacios Mateos JM. Osteoporosis [J]. Revista clinica espanola, 1959, 74: 215-22.
- 4 张智海, 张智若, 刘忠厚, 等. 中国大陆地区以-2.0SD为诊断标准的骨质疏松症发病率回顾性研究 [J]. 中国骨质疏松杂志, 2016, 22(01): 1-8.
- 5 Min S, Lee B, Yoon S. Deep learning in bioinformatics [J]. Brief Bioinform, 2017, 18(5): 851-869.
- 6 Ma WL, Xin K, Chen K, et al. Relationship of common variants in VEGFA gene with osteonecrosis of the femoral head: A Han Chinese population based association study [J]. Sci Rep, 2018, 8(1): 16221.
- 7 李沫, 王海阳, 王丽娟, 等. 老年骨质疏松常用治疗药物的研究现状及展望 [J]. 中华老年骨科与康复电子杂志, 2019, 5(06): 365-369.
- 8 Compston J, Cooper A, Cooper C, et al. UK clinical guideline for the prevention and treatment of osteoporosis [J]. Arch Osteoporos, 2017, 12(1): 43.
- 9 Yang TL, Shen H, Liu AQ, et al. A road map for understanding molecular and genetic determinants of osteoporosis [J]. Nat Rev Endocrinol, 2020, 16(2): 91-103.
- 10 Mukherjee A, Rotwein P. Selective signaling by Akt1 controls osteoblast differentiation and Osteoblast-Mediated osteoclast development [J]. Mol Cell Biol, 2012, 32(2): 490-500.
- 11 Muller-Deubert S, Seefried L, Krug M, et al. Epidermal growth factor as a mechanosensitizer in human bone marrow stromal cells [J]. Stem Cell Research, 2017, 24: 69-76.
- 12 Lv NN, Zhou ZZ, He SJ, et al. Identification of osteoporosis based on gene biomarkers using support vector machine [J]. Open Medicine, 2022, 17(1): 1216-1227.
- 13 Preininger B, Duda G, Gerigk H, et al. CD133: Enhancement of Bone Healing by Local Transplantation of Peripheral Blood Cells in a Biologically Delayed Rat Osteotomy Model [J]. Plos One, 2013, 8(2): 52650.
- 14 Yu T, You XM, Zhou HC, et al. p53 plays a central role in the development of osteoporosis [J]. Aging, 2020, 12(11): 10473-10487.
- 15 Su XL, Yu M, Qiu GX, et al. Evaluation of nestin or osterix promoter-driven cre/lox system in studying the biological functions of murine osteoblastic cells [J]. Am J Transl Res, 2016, 8(3): 1447-1459.
- 16 Van DJ, Schaaf GJ, Matos AA, et al. Cell surface glycoprotein CD24 marks bone Marrow-Derived human mesenchymal stem/stromal cells with reduced proliferative and differentiation capacity in vitro [J]. Stem Cells Dev, 2021, 30(6): 325-336.
- 17 Li R, Xue K, Li JM. FGF13 suppresses acute myeloid leukemia by regulating bone marrow niches [J]. Front Med, 2022: Online ahead of print.
- 18 Ishida R, Ezura Y, Iwasaki H, et al. Linkage disequilibrium and haplotype analysis among four novel single-nucleotide polymorphisms in the human leukemia inhibitory factor(LIF)gene [J]. J Hum Genet, 2001, 46(10): 557-559.

(收稿日期:2022-06-27)

(本文编辑:吕红芝)

许航, 崔宇韬, 任广凯, 等. 骨质疏松症关键基因的筛选及生物信息学分析 [J/CD]. 中华老年骨科与康复电子杂志, 2023, 9(1): 18-22.