［4］ KENJI S，MAMI O，KEIKO A，et al. Randomized，prospective assessment of moisturizer efficacy for the treatment of radiation dermatitis following radiotherapy after breast-conserving surgery［J］. Japanese Journal of Clinical Oncology，2015，45（12）：1146-1153.
［5］ 张静，粱婷婷. 维生素 B12 联合用药治疗 2 级放射性皮炎的护理观察［J］. 中国实用医药，2016，11（4）：198-199.
［6］ 刘猛，贾立群. 李佩文教授中医外治肿瘤并发症的临证经验初探［J］. 中国中西医结合杂志，2014，34（11）：1390-1391.
［7］ 刘柳红. 白杨素对大鼠胺碘酮外渗性皮肤损伤影响的实验研究［D］. 广州：广东药科大学，2021.
［8］ 王青，岳佳佳，王珊，等. 中药外治法在放射性皮炎中的应用现状［J］. 江苏中医药，2020，52（1）：87-90.

［文章编号］ 1007-0893(2024)02-0004-07 DOI：10.16458/j.cnki.1007-0893.2024.02.002

基于生物信息学和机器学习筛选阴虚质绝经后骨质疏松症关键基因的研究

陈 刚 李 峰※ 郭珈宜 高泉阳 郭艳幸 郭艳锦

（洛阳正骨医院 河南省骨科医院，河南 洛阳 471002）

［摘 要］ **目的**：应用生物信息学和机器学习的方法筛选阴虚质绝经后骨质疏松症（PMOP）中的关键基因。**方法**：从基因表达数据库（GEO）数据库下载 GSE87474 中平和质和阴虚体质人群的外周血单核细胞数据，下载与 PMOP 相关的数据集 GSE56116 和 GSE100609，并用 SVA 包进行合并。其次，利用加权基因共表达网络分析筛选与中医阴虚质密切相关的基因；利用 limma 包筛选合并数据集中的差异表达基因；并将两者取交集获取与阴虚质 PMOP 相关的基因。然后，通过最小绝对值收敛和选择算子（LASSO）、支持向量机–递归特征消除（SVM–RFE）和随机森林（RF）算法，识别阴虚质 PMOP 过程中潜在的关键基因，并评价其诊断效能。最后，对关键基因开展单基因的基因集富集分析（GSEA）。**结果**：共获得 46 个与阴虚质 PMOP 相关的基因，利用 LASSO、SVM–RFE 和 RF 算法从其中筛选出 1 个关键基因即溶质载体家族 39 成员 8（SLC39A8）。单基因 GSEA 显示，SLC39A8 高表达组富集到了 HIPPO、白细胞介素 –12、P38MAPK、P53、细胞凋亡等 104 个信号通路。**结论**：本研究发现 SLC39A8 可能为阴虚质 PMOP 的关键基因，结果可为进一步阐释阴虚质 PMOP 分子机制及从中医体质学说论治 PMOP 提供新的思路和切入点。

［关键词］ 绝经后骨质疏松症；阴虚质；生物信息学；机器学习；生物标志物
［中图分类号］ R 681 ［文献标识码］ B

Screening of Key Genes for Oosteoporosis in Postmenopausal Osteoporosis with Yin Deficiency Based on Bioinformatics and Machine Learning

CHEN Gang, LI Feng※, GUO Jiayi, GAO Quanyang, GUO Yanxing, GUO Yanjin

(Luoyang Bone-setting Hospital, Henan Orthopaedic Hospital, Henan Luoyang 471002)

[Abstract] Objective To apply bioinformatics and machine learning to screen for key genes in postmenopausal osteoporosis (PMOP) with Yin deficiency. Methods The peripheral blood mononuclear cell data of GSE87474 people with moderate

［收稿日期］ 2023 - 11 - 21
［基金项目］ 河南省中医药科学研究专项课题（20–21ZY2239）
［作者简介］ 陈刚，男，主治医师，主要研究方向是中医药防治骨关节炎。
［※ 通信作者］ 李峰（E-mail：545459315@qq.com）

constitution and yin deficiency constitution were downloaded from GEO database, and the data sets GSE56116 and GSE100609 related to PMOP were downloaded and combined with SVA package. Secondly, weighted gene co-expression network was used to screen genes closely related to Yin deficiency in traditional Chinese medicine (TCM). limma package was used to screen the differentially expressed genes in the combined data set. They were intersected to obtain genes associated with Yin deficiency PMOP. Then, the minimum absolute convergence and selection operator (LASSO), support vector machine-recursive feature elimination (SVM-RFE) and random forest (RF) algorithms were used to identify potential key genes in the process of PMOP in Yin deficiency and evaluate their diagnostic effectiveness. Finally, single gene set enrichment analysis (GSEA) was performed for key genes. Results A total of 46 genes related to PMOP with Yin-deficiency constitution were obtained, and one key gene, solute carrier family 39 member 8 (SLC39A8), was screened out by using LASSO, SVM-RFE and RF algorithms. Single-gene GSEA showed that the SLC39A8 high expression group was enriched in 104 signaling pathways including HIPPO, interleukin-12, P38MAPK, P53, and apoptosis. Conclusion This study found that SLC39A8 may be the key gene of PMOP in Yin deficiency, and the results can provide a new idea and breakthrough point for further elucidation of the molecular mechanism of PMOP in Yin deficiency and the treatment of PMOP from the constitution theory of TCM.

[Keywords] Postmenopausal osteoporosis; Constitution with Yin-deficiency; Bioinformatics; Machine learning; Biomarkers

骨质疏松症（osteoporosis，OP）是一种以骨量减少、骨密度降低、骨显微结构退化以至骨脆性增加的一类代谢性骨病[1]，而绝经后骨质疏松症（postmenopausal osteoporosis，PMOP）主要发生在绝经后女性，由于雌激素水平缺乏导致体内发生的一种慢性代谢性疾病[2-3]，为原发性 OP 的主要类型，表现为患者的骨骼变薄，骨量减少，易伴发骨折风险，主要累及患者的上肢肩关节、腕关节，下肢的髋关节以及脊柱的腰椎部位等[4]。随着我国步入老龄化社会，据统计，绝经后妇女发生 PMOP 的患病率更是达到了 50 % 以上[5]，因 PMOP 发病缓慢，相当一部分患者早期临床症状不明显而易被忽视，以致影响早期辨识和早期诊治，多数患者只有在骨折发生了以后，才明确患 PMOP，故 PMOP 亦被称为“沉默的杀手”[6]。每年因 PMOP 引发的骨折甚至残疾并由此产生的高昂治疗费用，给个体、家庭和社会带来沉重的负担[7]。因此，PMOP 的早期防治已成为当务之急。

近年来，中医药在防治 OP 方面表现出独特优势和显著疗效，从中医体质角度开始探讨防治 PMOP 正逐步成为新的研究热点[8-9]。中医传统体质学说注重个体差异性，认为体质与疾病的发生、发展及预后关系密切[10-11]。多项研究表明[12-13]，阴虚质类型是 PMOP 患者的高危偏颇体质，阴虚质人群 OP 的发病概率明显高于其他体质人群，为此，研讨阴虚质 PMOP 的相关病理过程及分子机制，可为其针对性地防治及患者的预后和生活质量等提供一定的科学依据，具有重要的意义。

当前生物信息学方法被广泛应用于研发与诊断或预后的潜在生物标志物。加权基因共表达网络分析（weighted gene co–expression network analysis，WGCNA）是生物信息学中常用的方法，已经被诸多研究成功运用于探索疾病数据集中基因集的内连性和基因集与表型之间的关联，挖掘潜在的生物标记基因或治疗靶点[14-15]。机器学习作为人工智能的重要组成部分，已经应用于各种医疗领域，在骨科疾病方面，机器学习已应用于诊断疾病和探索生物标志物，检测或预测骨折，以及寻找潜在的治疗靶标等多个方面[16-17]。

基于以上认识，本研究一方面通过 limma 包筛选 PMOP 患者外周血单核细胞（peripheral blood mononuclear cells，PBMCs）中的差异表达基因（differentially expressed genes，DEGs）获取与 PMOP 相关的基因；另一方面，WGCNA 筛选与中医阴虚质密切相关的基因；同时将两者取交集获取与阴虚质 PMOP 相关的基因；然后，再利用最小绝对值收敛和选择算子（least absolute shrinkage and selection operator，LASSO）、支持向量机–递归特征消除（support vector machine–recursive feature elimination，SVM–RFE）和随机森林（random forest，RF）算法 3 种经典的机器学习算法，识别阴虚质 PMOP 过程中潜在的关键基因，并评价其诊断效能；最后，开展关键基因的基因集富集分析（gene set enrichment analysis，GSEA）来探索关键基因参与阴虚质 PMOP 病理过程可能的分子机制。总之，本研究利用生物信息学和机器学习相关结合的方法，筛选阴虚质 PMOP 过程中潜在的关键基因及分子机制，以期为其后续的基础和临床研究提供新的研究思路和切入点。

1 资料与方法

1.1 数据的来源和预处理

从基因表达数据库（gene expression omnibus，GEO）（链接：http://www.ncbi.nlm.nih.gov/geo/）中检索样本来源为 PBMCs 的与中医阴虚质和 PMOP 相关的数据集；共检索出与 PMOP 相关的数据集为 GSE56116（control 组 3 个样本，PMOP 组 10 个样本，检测方法为 Array）和 GSE100609（control 组和 PMOP 组分别为 4 个样本，检测方法为 RNA–seq），与中医阴虚质相关的数据集为 GSE87474［平和质（balanced constitution，BC）组 8 个样本，

阴虚质组 12 个样本，检测方法为 Array〕；并下载其矩阵和平台文件。依据相应数据集的平台文件将数据集中的探针名转化为基因名称；若存在基因对应多个探针者，使用 avereps 函数计算这些探针所对应样本的表达量的均值，并将其作为该基因的表达量。对于 Array 检测的数据集 GSE56116 和 GSE87474 需用 normalizeBetweenArrays 函数对数据进行矫正，若基因表达数值较大，则对其取 log2 处理。对于 RNA–seq 检测的数据集 GSE100609 则需将其 FPKM 格式的数据矩阵转化为 TPM 格式。最后，利用 SVA 包对预处理后的 GSE56116 和 GSE100609 数据集进行合并，保存为 “merge.normalize.txt” 文件。

1.2　筛选 DEGs

利用 limma 包，以 |log2（fold change，FC）| ＞ 0.585 和 P ＜ 0.05 为阈值条件对 “merge.normalize.txt” 文件的表达数据进行差异分析，筛选 DEGs，并进行可视化。

1.3　WGCNA

利用 WGCNA 包寻找最佳软阈值，将预处理后的 GSE87474 数据矩阵逐一转化为邻接和拓扑重叠矩阵，再进行基因聚类；将模块基因数目设为 30，通过动态剪切树的方法对相似模块进行合并，将合并后的基因模块与 BC 组和 YDC 组别信息进行关联，计算其相关系数和 P 值[18-19]。选择与 YDC 组正相关性最强的基因模块中的基因作为 YDC 相关性基因进行后续分析。

1.4　识别阴虚质 PMOP 的关键基因

将 1.2 中挖掘出的 DEGs 与 1.3 中 WGCNA 鉴定出的与 YDC 组正相关性最强的基因模块中的基因进行交集，得到与阴虚质 PMOP 相关的基因。然后，分别利用 LASSO、SVM–RFE 和 RF 算法从阴虚质 PMOP 相关基因中筛选重要基因，再利用 venn 包绘制维恩图，3 种机器学习算法分别识别出的重要基因的重叠基因即为关键基因[20]。同时，基于 GSE87474 和 “merge.normalize.txt” 文件数据，利用 pROC 包绘制关键基因的受试者工作特征（receiver operating characteristic curve，ROC）曲线，计算 ROC 的曲线下面积（area under curve，AUC）值，以评价关键基因的诊断效能。

1.5　关键基因的 GSEA

计算关键基因在 “merge.normalize.txt” 文件中所有样本表达量的中位数。依据关键基因的中位数将所有基因分为关键基因的低、高表达两组。然后，利用 GSEA4.1.0 软件，分别以 c2.cp.kegg.v2022.1.Hs.symbols.gmt、c2.cp.reactome.v2022.1.Hs.symbols.gmt、c2.cp.wikipathways.v2022.1.Hs.symbols.gmt 为参考基因集开展 GSEA。以标准化富集得分（normalized enrichment score, NES）的绝对数＞ 1、名义 P 值（nominal P–value）＜ 0.05 和错误发现率（false discovery rate，FDR）＜ 0.25 为阈值筛选阳性基因集[19]。

2　结　果

2.1　DEGs 的筛选结果

如插页 1 图 1A、图 1B 所示，SVA 包合并 GSE56116 和 GSE100609 后，数据间的批次效应已不明显。插页 1 图 1C、图 1D 差异分析结果显示，在 control 组和 PMOP 组间，共筛选出 196 个 DEGs（上调 69 个，下调 127 个）。

2.2　WGCNA 结果

插页 1 图 2A、图 2B 网络拓扑分析结果显示，当软阈值为 3 时能够使邻接函数较好满足无尺度条件且 R^2 ＞ 0.9。动态剪切树法获得了 15 个基因模块（插页 1 图 2C）。插页 1 图 2D、图 2E 展示了 15 个基因模块与组别之间的相关系数及 P 值，其中 turquoise 模块（共有 5498 个基因）与 YDC 组的相关性最高。因此，本研究将选择 turquoise 模块中的基因开展后续研究。

2.3　阴虚质 PMOP 关键基因的筛选

如插页 2 图 3A 所示，196 个 DEGs 与 turquoise 模块中的 5498 个基因进行交集，共获得 46 个与阴虚质 PMOP 相关的基因。LASSO 算法从 46 个基因中筛选出了 10 个重要基因（插页 2 图 3B），SVM–RFE 算法筛选出了 24 个重要基因（插页 2 图 3C），RF 算法筛选出了 10 个重要基因（插页 2 图 3D、图 3E），三者取交集后鉴定出溶质载体家族 39 成员 8（solute carrier family 39 member 8，SLC39A8）为关键基因（插页 2 图 3F）。为了进一步评价 SLC39A8 作为关键基因诊断阴虚质和 PMOP 的效能，本研究基于 GSE87474 和 “merge.normalize.txt” 文件数据，绘制了 SLC39A8 的 ROC 曲线，结果表明，SLC39A8 在两个数据集的 AUC 值分别达到了 0.729 和 0.867，提示 SLC39A8 表现出了良好的诊断性能（插页 2 图 3G、图 3H）。

2.4　单基因 GSEA

为了进一步探索 SLC39A8 参与阴虚质 PMOP 病理过程可能的分子机制，本研究对 SLC39A8 进行了单基因 GSEA。结果显示，按照筛选条件 SLC39A8 低表达组在 3 个参考基因集中均未富集到阳性基因集；SLC39A8 高表达组在 c2.cp.kegg.v2022.1.Hs.symbols.gmt 中富集到了 4 个阳性基因集，在 c2.cp.reactome.v2022.1.Hs.symbols.gmt 中富集到了 88 个阳性基因集，在 c2.cp.wikipathways.v2022.1.Hs.symbols.gmt 中富集到了 12 个阳性基因集，见表 1。

表 1　SLC39A8 单基因 GSEA 的结果

NAME	NES	P-val	FDR
KEGG_BASE_EXCISION_REPAIR	1.8490031	0.004056795	0.062983505
KEGG_DNA_REPLICATION	1.8174908	0.024489796	0.046171430
KEGG_NUCLEOTIDE_EXCISION_REPAIR	1.7257938	0.020618556	0.102996916
KEGG_PROTEIN_EXPORT	1.6716709	0.006036217	0.148634800
REACTOME_POLYMERASE_SWITCHING_ON_THE_C_STRAND_OF_THE_TELOMERE	1.8547775	0.002012072	0.227326260
REACTOME_BASE_EXCISION_REPAIR	1.8529524	0.012121212	0.174865860
REACTOME_REGULATION_OF_PLK1_ACTIVITY_AT_G2_M_TRANSITION	1.8392426	0.008196721	0.161725610
REACTOME_RECOGNITION_OF_DNA_DAMAGE_BY_PCNA_CONTAINING_ REPLICATION_COMPLEX	1.7520595	0.014830508	0.235935580
REACTOME_TRANSCRIPTION_COUPLED_NUCLEOTIDE_EXCISION_REPAIR_TC_NER	1.6922401	0.030303031	0.247914720
REACTOME_TRANSCRIPTIONAL_REGULATION_BY_TP53	1.6883153	0.025263159	0.244517490
REACTOME_REGULATION_OF_TNFR1_SIGNALING	1.6856986	0.033970278	0.237990250
REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX	1.6799879	0.028985508	0.240951820
REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_C_STRAND_OF_THE_TELOMERE	1.6732710	0.030991735	0.246285570
REACTOME_RESOLUTION_OF_ABASIC_SITES_AP_SITES	1.6688272	0.039256200	0.246234940
REACTOME_PCNA_DEPENDENT_LONG_PATCH_BASE_EXCISION_REPAIR	1.6658382	0.029227557	0.243466230
REACTOME_MRNA_CAPPING	1.6653711	0.032520324	0.234649600
REACTOME_FORMATION_OF_TC_NER_PRE_INCISION_COMPLEX	1.6622212	0.043568466	0.231742590
REACTOME_FORMATION_OF_THE_EARLY_ELONGATION_COMPLEX	1.6511936	0.029239766	0.247228370
REACTOME_TRANSLATION	1.6181413	0.020449897	0.243931870
REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_ LOADING_OF_CLASS_I_MHC	1.6134753	0.013916501	0.234648960
REACTOME_INHIBITION_OF_DNA_RECOMBINATION_AT_TELOMERE	1.6130574	0.026530612	0.229914110
REACTOME_ACTIVATED_TAK1_MEDIATES_P38_MAPK_ACTIVATION	1.6119530	0.025210084	0.227177650
REACTOME_DISEASES_OF_MITOTIC_CELL_CYCLE	1.6118803	0.030800821	0.221983430
REACTOME_MATURATION_OF_SARSCOV2_NUCLEOPROTEIN	1.6078640	0.010416667	0.224229570
REACTOME_MICRORNA_MIRNA_BIOGENESIS	1.6063981	0.014799154	0.216876460
REACTOME_BASE_EXCISION_REPAIR_AP_SITE_FORMATION	1.6055129	0.033195022	0.214020730
REACTOME_TICAM1_RIP1_MEDIATED_IKK_COMPLEX_RECRUITMENT	1.6052123	0.010660981	0.210262130
REACTOME_LYSOSOME_VESICLE_BIOGENESIS	1.6016428	0.033898305	0.204676760
REACTOME_ABORTIVE_ELONGATION_OF_HIV_1_TRANSCRIPT_IN_THE_ABSENCE_ OF_TAT	1.5947343	0.036960986	0.212866000
REACTOME_NEF_MEDIATES_DOWN_MODULATION_OF_CELL_SURFACE_ RECEPTORS_BY_RECRUITING_THEM_TO_CLATHRIN_ADAPTERS	1.5891697	0.039920160	0.214927760
REACTOME_NUCLEAR_IMPORT_OF_REV_PROTEIN	1.5743084	0.046092186	0.219975170
REACTOME_GAP_FILLING_DNA_REPAIR_SYNTHESIS_AND_LIGATION_IN_GG_NER	1.5738022	0.041928720	0.217210640
REACTOME_RIPK1_MEDIATED_REGULATED_NECROSIS	1.5624053	0.043298970	0.223755430
REACTOME_INTERACTIONS_OF_REV_WITH_HOST_CELLULAR_PROTEINS	1.5547717	0.049079753	0.227691740
REACTOME_TRANSCRIPTIONAL_REGULATION_BY_VENTX	1.5526295	0.018480493	0.224799830
REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	1.5517828	0.040169135	0.217054350
REACTOME_RHOBTB2_GTPASE_CYCLE	1.5366800	0.043298970	0.217850790
REACTOME_SUMOYLATION_OF_SUMOYLATION_PROTEINS	1.5299032	0.046370970	0.221014870
REACTOME_DDX58_IFIH1_MEDIATED_INDUCTION_OF_INTERFERON_ALPHA_BETA	1.5298977	0.047422680	0.218444930
REACTOME_ORGANELLE_BIOGENESIS_AND_MAINTENANCE	1.5248682	0.040160640	0.221633660
REACTOME_INTERLEUKIN_12_SIGNALING	1.5230297	0.042510122	0.220105800
WP_LEUKOCYTEINTRINSIC_HIPPO_PATHWAY_FUNCTIONS	1.8110938	0.002136752	0.148861780
WP_COHESIN_COMPLEX_CORNELIA_DE_LANGE_SYNDROME	1.7869320	0.022088353	0.136800930
WP_BASE_EXCISION_REPAIR	1.7565134	0.010204081	0.150981470
WP_MRNA_PROC			